DETECTION AND MANAGEMENT OF PERSONAL DATA WITH AN ONTOLOGY-BASED APPROACH

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES
OF
MIDDLE EAST TECHNICAL UNIVERSITY

BY

PINAR BIL ATASOY

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF MASTER OF SCIENCE
IN
COMPUTER ENGINEERING

MAY 2022

Approval of the thesis:

DETECTION AND MANAGEMENT OF PERSONAL DATA WITH AN ONTOLOGY-BASED APPROACH

submitted by **PINAR BIL ATASOY** in partial fulfillment of the requirements for the degree of **Master of Science in Computer Engineering Department, Middle East Technical University** by,

Prof. Dr. Halil Kalıpçılar
Dean, Graduate School of **Natural and Applied Sciences** _____

Prof. Dr. Halit Oğuztüzün
Head of Department, **Computer Engineering** _____

Prof. Dr. Ferda Nur Alpaslan
Supervisor, **Computer Engineering, METU** _____

Examining Committee Members:

Prof. Dr. Nihan Kesim Çiçekli
Computer Engineering, METU _____

Prof. Dr. Ferda Nur Alpaslan
Computer Engineering, METU _____

Assoc. Prof. Dr. Çiğdem Turhan
Computer Engineering, Atılım University _____

Date: 09.05.2022

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Surname: Pınar Bil Atasoy

Signature :

ABSTRACT

DETECTION AND MANAGEMENT OF PERSONAL DATA WITH AN ONTOLOGY-BASED APPROACH

Atasoy, Pınar Bil
M.S., Department of Computer Engineering
Supervisor: Prof. Dr. Ferda Nur Alpaslan

May 2022, 44 pages

According to paragraph 3 of Article 20 of the Constitution of the Republic of Turkey has the right to the protection of personal data concerning himself with each individual. This law of protection of personal data is provided by the Law No. 6698 on the Personal Data Protection Law. When technology is intertwined with our lives, it has become very important to protect the privacy of the personal data such as the identity, health, communication and financial information of the individual and religious belief. Amount of data is increased exponentially with the increase in the place of technology in our lives and the manual detection and management of personal data requires great effort. This thesis aims to define personal data and manage it quickly and efficiently with an ontology-based approach. Spelling mistakes and errors in the text are removed by pre-processing the text with natural language processing tools and methods. We also propose an ontology in the domain of Personal Data Protection. In this way, using domain-specific information extraction and inference rules, personal data is determined and managed accurately and unerringly.

Keywords: Ontology, Personal Data, Inference

ÖZ

KİŞİSEL VERİLERİN ONTOLOJİ TABANLI YAKLAŞIM İLE TESPİTİ VE YÖNETİMİNİN SAĞLANMASI

Atasoy, Pınar Bil
Yüksek Lisans, Bilgisayar Mühendisliği Bölümü
Tez Yöneticisi: Prof. Dr. Ferda Nur Alpaslan

Mayıs 2022 , 44 sayfa

Türkiye Cumhuriyeti Anayasası'nın 20. Maddesinin 3.fıkrasına göre her birey kendisi ile ilgili kişisel verilerin korunmasını isteme hakkına sahiptir. Kişisel verilere ilişkin bu koruma 6698 sayılı Kişisel Verilerin Korunması Kanunu ile sağlanmıştır. Teknolojinin hayatlarımızla iç içe olduğu bu dönemde, bireyin kimlik, sağlık, iletişim ve mali bilgileri, dini inancı gibi kişisel verilerin mahremiyetini korumak çok önemli hale gelmiştir. Hayatımızda teknolojinin yerinin artmasıyla beraber katlanarak artan veride, kişisel verilerin el ile tespiti ve yönetimi büyük bir çaba gerektirmektedir. Bu tez ile kişisel verilerin hızlı ve verimli bir şekilde tanımlanması ve sonrasında ontoloji tabanlı yaklaşım ile istenen şekilde yönetilmesi sağlanacaktır. Doğal dil işleme ile metin ön işleme yapılarak metin içerisindeki yazım yanlışları ve hatalar ayıklanacak, kişisel veri alanına uygun olarak yarattığımız ontoloji ve anlamsal sorgulama ile çıkarımlar yapmamızı sağlayan kurallar sayesinde yazılan kişisel veriler doğru ve eksiksiz olarak tespit edilerek yönetimi sağlanacaktır.

Anahtar Kelimeler: Ontoloji, Kişisel Veri, Anlamsal Çıkarım

To my family

ACKNOWLEDGMENTS

I would like to express my gratitude and give my warmest thanks to my supervisor Prof. Dr. Ferda Nur Alpaslan for all of her patience, advice, criticism, encouragements and insight throughout the research. Without her guidance and wisdom, I could not accomplish the task.

I would like to thank my committee members Prof. Dr. Nihan Kesim Çiçekli and Assoc. Prof. Dr. Çiğdem Turhan for taking the time to review my thesis. Their advice helped me finalize my thesis.

The technical assistance and guidance of Osman Tufan Doğan, Hakan Merdanoğlu, Samet Akpınar, Elif Tansu Sunar and Alper Karamanlıoğlu is greatly appreciated.

I would like to thank my husband İsmet for his positivity, constant support, help, encouragement, love and comforting. But most of all, thank you for being my best friend and for being a caring and trustworthy partner.

I would like to express my love and gratitude to my dearest mother, Fatma and my dearest father, Erdoğan. I am lucky to have their endless support and encouragements. I feel blessed to have them by my side from miles away.

I would like to express my sincere thanks to Atasoy family. Ercemal and Aynur Atasoy, who have been a second family to me. I always feel their love and support.

I give thanks to my friends Cansu Alptekin Gökbender, Emel Varol, Erbil Yakışkan, Meliz Ersoy, Uğur Orhan, Esra Berktaş and Furkan Ulugün. They have always supported and encouraged me to finish this study.

TABLE OF CONTENTS

LIST OF TABLES

TABLES

LIST OF FIGURES

FIGURES

LIST OF ABBREVIATIONS

GDPR	General Data Protection Regulation
PDPL	Personal Data Protection Law
PDPA	Personal Data Protection Authority
IoT	Internet of Things
IoMT	Internet of Medical Things
SPARQL	SPARQL Protocol and RDF Query Language
SQWRL	Semantic Query-Enhanced Web Rule Language
OWL	Ontology Web Language
SWRL	Semantic Web Rule Language
NLP	Natural Language Processing
NER	Named Entity Recognition
CRF	Conditional Random Fields
OBIE	Ontology-based Information Extraction

CHAPTER 1

INTRODUCTION

1.1 Motivation and Problem Definition

When technology is intertwined with our lives, it has become very important to protect the privacy of personal data such as the identity, health, communication and financial information of the individual, and religious belief. Data is increased exponentially with the increase in the place of technology in our lives and the manual detection and management of personal data requires great effort.

Various regulations have been made in both Turkish law and European Union law regarding the protection of personal data. To elaborate on these; according to paragraph 3 of Article 20 of the Constitution of the Republic of Turkey every individual has the right to the protect the personal data concerning himself.

This law of protection of personal data [3] is provided by the Law No. 6698 on the Personal Data Protection Law (PDPL). This law is a law that summarizes the process of protecting personal data in EU countries before the General Data Protection Regulation (GDPR) [4], dated 25 May 2018, applied by the European Union (EU) countries. With the GDPR and PDPL numbered 6698, the obligations of data controllers who process data have increased and many legal sanctions have been applied. The obligations imposed on the data controllers in Turkey being controlled and by the Personal Data Protection Authority (PDPA) was established on January 30, 2017. Institutions do not comply with the obligations imposed sanctions on the scope PDPL are applied by PDPA. Violations regarding the protection of personal data are severely penalized by the Personal Data Protection Authority in Turkey. Although these penalties are usually based on complaints, companies and institutions that process personal

data are obliged to register in a system called Data Controllers Registry Information System (VERBIS). Legal sanctions against those responsible for violating this registration obligation have been actively implemented since the beginning of 2022.

As mentioned before, the protection of personal data and the prevention of violations have become essential. Data controllers, who are required to keep personal data in order to continue their commercial activities, will be able to benefit from this study. Nowadays, there is almost no legal entity that does not have responsibility for personal data. Therefore, any legal entity that has access to personal data which belongs to real persons, within the scope of its commercial activities, will be able to benefit from this thesis. These legal entities can be listed as e-commerce companies, stock companies and limited liability companies. This study is of great importance for data controllers such as hospitals, clinics, pharmacies, aesthetic centers and psychologists who have responsibility for sensitive personal data. According to the law, the processing of sensitive personal data is subject to more severe conditions than the processing of non-sensitive personal data. The Personal Data Protection Authority (Kişisel Verileri Koruma Kurumu) also has decisions that impose heavier fines in violation of sensitive personal data. In this context, The Personal Data Protection Authority has decided [2] that taking fingerprint data of individuals in order to keep track of check-in and check-out times to the workplace using fingerprints is not proportional depending on the purpose of processing. PDPA has decided that the consent obtained is invalid due to the fact that the biometric data in this case is sensitive personal data. In this decision, the fine imposed on the data controller for the violation of sensitive personal data is 4 times the fine imposed for the violation of non-sensitive quality data.

With this thesis, personal data is to be identified and managed quickly and efficiently with the ontology-based system where data samples are obtained from different data sources. Spelling mistakes and errors in the text will be removed by pre-processing the text with natural language processing tools. Because personal data violations are now actively sanctioned, it is of great importance to prevent these violations.

1.2 Proposed Method

In this thesis, personal data detection and management system is proposed with an ontology-based approach. Formal documents such as Supreme Court decisions containing personal data are used as input in their raw form. Afterwards, various NLP operations are carried out on these documents and they are made suitable for finding the personal data they contain. Then, the personal data contained in these text documents and the information of which person they belong to are detected and printed to an output file in a structured format.

The personal data protection law mentions the terms of processing of personal data, the characteristics of sensitive personal data, the obligations of data controllers and the rights of the data subject. Personal data ontology is designed and created with the analysis made in line with this law. This ontology is populated with the personal data obtained in structured format, and inference rules are used for extraction of new personal data hidden in the text.

In order to detect personal data in the first stage, the outputs obtained by using specified NLP methods and the dependencies between these outputs are used in this method. For the personal data obtained in the second stage, inference rules are used on the populated ontology.

1.3 Contributions of This Thesis

Ontologies are designed to allow for the modeling of knowledge about a specified domain, as well as the reuse and interchange of individual modules and also they are open to development by adding new classes and features into them. Ontology designed in the personal data domain is one of the greatest contributions of this thesis. This ontology classifies personal data with a distinction in accordance with the definition of sensitive personal data (Özel nitelikli kişisel veri) in the law and keeps their sensitivity levels. With the development of technology and science, new types of personal data will be added to this ontology, which will be able to grow and meet current needs.

Nowadays, almost all official documents contain personal data and most legal entities have responsibilities regarding the storage and processing of personal data. This study provides the detection and management of personal data in Turkish texts. With the inferences rules developed in this study, the detection of the personal data hidden in the sentences is also provided.

1.4 The Outline of the Thesis

The rest of the thesis is organized as follows :

Chapter 2 introduces the related work on the subject under three main sections: ontology-based approaches, named entity recognition studies in Turkish and studies in protection of personal data.

Chapter 3 contains design of the ontology and programs that are used to create and work on this ontology.

Chapter 4 shows the system design and explains how the system works in detail.

Chapter 5 evaluates the overall results of this ontology-based system and explains what can be done to improve the results.

Chapter 6 contains the conclusion and future work that mentions what can be done with this study in the future, how the results obtained in this study can be improved.

CHAPTER 2

RELATED WORK

Related work on this subject have been analyzed in 3 sections: ontology-based approaches, named entity recognition studies in Turkish, and studies in protection of personal data.

2.1 Ontology-based Approaches

An ontology is a formal specification of a certain domain. Since ontologies represent a subject area formally and support reusability, their use in this study has been deemed appropriate.

Although machine learning may seem to be a complex concept for non-expert users, the ontology-based interface designed in this article [7] is aimed to enable these people to understand and influence a machine learning system. The mentioned ontology-based system works by dividing the work between the user and the system. While the system generates a classification model from statistical data, the user will help the system with decision-making mechanism and domain knowledge.

This study [31] has an ontology-based approach to resolving the ambiguities in the sentence. With the difficulty of Turkish being a free word order language, they present a method that parses Turkish sentences.This method is for finding the argument structure in the sentence using the constraints on the structures defined in the ontology.

2.1.1 Ontology-based Approaches in Healthcare System

In Medical Ontology Study [36] an ontology is proposed in the domain of medical services which focused on knowledge sharing between devices and actors during the diagnostic process in the emergency domain. To create an ontology, firstly they extracted entities and created a concept dictionary table that shows names, characteristics of them and relationships among these entities. Then, they developed a taxonomy to provide a structure for the ontology and a binary relationship table for indicating the interaction among the concepts. Finally, they develop axioms for adding logical expressions to ontology. After creating the ontology, they used reasoning tools to evaluate the consistency of the ontology. In the case of emergency cases, ontology-based systems provide the benefit of flexibility and semantic interoperability for knowledge transfer.

This study [19] focused on healthcare system based on ontology method. Ontology decision making process is used to assist doctors in emergency situations and to provide quick and easy solutions to doctors about diseases and specific treatments. In this system, user use Internet of Things (IoT) platform to manage data and use queries to retrieve information from the database.

This study [30] outlines a methodology for developing and implementing a centralized semantic knowledge base for India's healthcare system. This methodology is extracting data from structured and semantic heterogeneous health records in RDF format associating RDF triples with the appropriate domain ontology concepts using Oracle native inference engine using OWLPrime rules.

This study [28] focused on two main parts. The first one is proposing a HealthIoT Ontology by using semantic representation of the medical connected objects and their data. The second part focuses on an IoT Medicare System that integrates this ontology. On the first part, collected data is separated into two categories: data concerning the connected objects and data concerning the patient's state. After pre-processing, files contains these data converted to a unified format called 'SNON (Sensor Network Object Notation). During semantic knowledge modeling step, they propose the HealthIoT ontology that presents harmonization between the IoMT (Internet of Med-

ical Things) domain knowledge and the healthcare domain knowledge. They used an inference engine that uses their ontology and SWRL rules to perform reasoning tasks. The second part is integrated with query and inference engine deals with SWRL rules and SQWRL and SPARQL queries.

2.1.2 Ontology-based Information Extraction

Ontology-based information extraction (OBIE) is a subfield of information extraction. (Wimalasuriya Dou, 2010) [33] reviewed the studies in this field and tried to determine a common architecture that would enable them to classify these studies. They also compares how these systems are implemented, which tools are used to implement these systems, and the metrics used to measure their performance.

One of the oldest studies in this field [14] presents a method that can automatically create domain-specific ontologies from documents. Ontologies created with this method are dynamic, so they are continuously updated as new documents are added throughout the process. One of the other earlier works, the TEXT-TO-ONTO Ontology Learning Environment [23] offers an architecture that extracts ontology from text, which includes taxonomical and non-taxonomical relationships.

In this study [34], Kylin system creates ontology by extracting information from info-boxes in Wikipedia. Maedche et al. [22] shows an approach for an ontology-based knowledge extraction system. They used machine learning techniques for this domain-specific system. Both of these systems extract class names and the taxonomy as well as data type properties during the ontology creation.

This study [15] is an ontology-based information extraction and retrieval system that looks at usability, scalability, and retrieval performance as three concerns in semantic search. They present keyword-based semantic retrieval approach and domain-specific information extraction, inference, and rules significantly increase its performance. Pellet is used for inference module and Jena Rules is used for generating domain-specific rules. Scalability is achieved by the use of a semantic indexing strategy. This study also covers this systems application to soccer domain. This application has a rule that allows you to deduce the player who made an assist to the goal.

2.2 Named Entity Recognition in Turkish Language

Turkish has complex morphological structures and is an agglutinative language, consequently using only word-level modelling is inadequate for complex problems. Yeniterzi initially used word level modelling and then morpheme-level modelling with various features in order to analyze the impact of morphology in named-entity-recognition (NER) for Turkish.

There are two tokenization methods in this paper [35]. In Word-Level tokenization, each word is associated with a single state. Yeniterzi developed several feature sets which enables the word to be represented with a greater amount of information. In lexical model, the word tokens is used in its simplest form. In Root Feature, root forms of the words are used so to decrease the high variation of words in Turkish. They also used a Turkish morphological analyzer (Of lazer, 1994) that breaks down words into their roots and morphological characteristics. These tags are referred to as Part-of-Speech (POS) features as they are also used as extra features. In Proper-Noun Feature, the proper name database in the morphological analyzer is used to tag Turkish person, location and organizations as proper nouns. The orthographic case information of the words is used as the final feature. Capitalization of initial letter of the word is used to detect the named entities except the first word of the sentence. Case Feature is the name of this feature. In Morpheme-Level tokenization word is represented in multiple states, one for each morphological feature and one for a root. In this model, Root-Morph Feature is used to assign the value “morph” or “root” for the states containing a root or morpheme. In addition, the previously mentioned features were used in this model. Experiments and results indicate that all of these features dramatically enhance the system. On the other hand, the new method they used for Morpheme-level tokenization yield no dramatic improvement. The reason for this may be morphemes in Turkish may not be functional in NER or their representation of words in multiple states may not be the best possible option for using morphemes.

Zemberek is a open-source project that aims to present NLP solutions for Turkic Languages [6] . Although there have been some studies on this subject in Turkic languages before, these are generally studies on a single language, such as only Turk-

ish. This application offers solutions to many Turkic languages and the most important difference from other existing solutions is that it is completely open-source. Zemberek framework offers NLP operations such as spell checking, morphological parsing, stemming and word construction.

The library consists of two parts as language structure information and NLP operations. In order for the library to work in accordance with the desired language, the grammar and other language rules of that language must be provided to the library and each language added must comply with the rules of the previous languages. Generally, NLP operations adapts to the language loaded and provides easy access to the software by using its information. Although there are many studies on named entity recognition in different languages, there are fewer studies on Turkish. But in recent years, the studies on NER in Turkish have increased. The Turkish language is distinguished from other languages by vowel harmony and comprehensive agglutination. Proper nouns are easy to distinguish in written official texts because the first letters of proper nouns are capitalized and suffixes are separated by an apostrophe. However, these spelling rules are not followed on social media. In addition, many proper nouns in Turkish are also common nouns and the only way to separate them is spelling rules. This distinction is very difficult in situations that do not follow the spelling rules.

In this paper, they created three new data sets focused on social media and colloquial language data which is collected from different domains. Person, location and organization names are used for comparison.

[9] has a new NER approach similar to the cited previous work of (Şeker et al. 2012) which tokenizes the data before using features obtained from a morphological analysis process and predefined features such as POS-tags, proper noun tags, lower/upper case status and some features indicating the beginning of a sentence. However, since this approach is designed to be applied on formal written texts, the fact that the data sets used in this study contain informal words poses a problem. To solve this problem, they create a text normalizer that used for slang words, repeated characters, hash tags, smiley icons and capitalization. As the machine learning algorithm they used Conditional Random Fields (CRF).

Another study [29] aims to present a suitable model for Turkish NER by working

on the news domain and using conditional random fields trained with morphological and lexical features. In this study, the best result is obtained on name entities of ENAMEX (person, location and organization names) type, but it is also desired to work on TIMEX (date and time entities) and NUMEX (numerical expressions like money and percentages) in the future. Firstly, the data is tokenized in order to represent each word as a token. Then, a morphological analyzer (Oflazer, 1994) and morphological disambiguator is used to obtain most probable analysis for each word. Two types of gazetteers are used in this study : base gazetteers is the one likely to contain words occurring in a named entity, while generator gazetteers are smaller gazetteers that contain the stems of the some named entity generator words.

In data preparation stage, raw tags like PERSON, ORGANIZATION, LOCATION and OTHER are used to prepare future vectors. Conditional Random Fields (CRFs) are a discriminative framework for sequence prediction and CRF++ is used as an open source implementation of CRFs in this study. Although this study tries to be compared with previous studies in the literature, it is not possible as different types of NE types and different evaluation metrics are used. As future work, they aim to adding new generator gazetteers and TIMEX NUMEX entity types.

[13] present a platform which puts up many advanced natural language processing tools such as preprocessing, morphological analysis, and named entity recognition that even people without a computer background can easily use. ITU NLP Platform can be used in a variety of ways, including using it on the website, by uploading files, and using provided the WEB APIs for building higher level applications.

Several recent studies indicate that when named entity recognition systems created for well-formed text types are applied to text on social media such as tweets, their performance suffers dramatically. In this study [18] , they conduct comparative named entity recognition experiments using a rule-based multilingual named entity recognition system customized to Turkish. Based on the analysis and experimenting results, they propose system features needed to improve named entity recognition results for social media platforms such as Twitter.

2.3 Protection of Personal Data

The protection of personal data is provided with certain measures in both the European Union directive and the Personal Data Protection Law (PDPL). Therefore, the processing of a personal data depends on the person's freely given consent. This declaration of consent does not have a general validity and is taken separately according to the subject and purpose of processing the personal data [21].The fact that a person does not know where and how their personal data is processed does not only limit the person, but also has the potential to negatively affect the society [8]. Besides all this, in this study [16], it is mentioned that the definition of personal data and the sensitivity levels of this data may vary depending on countries. However, some data that the legislator does not define as non-sensitive personal data can be defined as sensitive personal data by data owners. It is the duty of the courts to solve the problems that may arise in this regard [10]. As a matter of fact, the European Court of Human Rights has evaluated the protection of personal data within the scope of the right to respect private life in many of its decisions and has imposed active obligations on states in this regard [12].

The processing of personal data must be lawful, fair and transparent. However, using these principles with machine learning can be challenging. In the article [20], it is discussed at what level it is possible to adapt the prevailing principles in the processing of personal data with machine learning. As a result, it is argued that the application of some human concepts such as fairness and transparency together with machine learning may lead to certain biased and troublesome decisions, but the use of machine learning in the processing of personal data can reach an optimal point after undergoing a human audit in the end.

In today's applications and services, Machine Learning techniques are commonly used for Data Analytics. However, as importance of personal data develops, it becomes a problematic social issue to keep personal data private while still benefiting from the data processing power of machine learning methods. Many existing solutions to this problem concentrate on improving the already prevalent cloud-based analytics approach, but this solution has some drawbacks such as single point failure and increased communication costs. Jianxin et al 2018 [37] has a new solution which is

to deploy Machine Learning services on edge devices. However, their solution brings along some issues such as a lack of appropriate user models and difficulty in implementing services for users. To solve these problems, they designed a system named Zoo which enables construction, composition and deployment of machine learning models on edge devices. Basically, there are two parts of this system: Development and Deployment.

In development, user designs the desired service by composing one or more existing ML services. Then, Zoo checks if models of these services are held in cache. If not, required ML services will be pulled from GitHub Gist.

Service interface and location definition are dealt with in deployment. Unlike development stage, deployment is not restricted to edge devices. As a result of this design, service can be easily shared between other devices. Also, the system outperforms current deep learning platforms and cloud-based services according to the article.

Bringing machine learning services and data analytics to the edge devices can properly protect privacy and personal data. But this system has some vulnerabilities such that hackers may gain access to personal devices or they may work together to infer data about users from the shared model. These topics need further investigations and improvements.

Consent is required for the processing of personal data under GDPR. Obtaining this consent to be valid under the GDPR, some conditions are stipulated. In this article [27], the analyzed information is related with the concept of consent to present the ontology called GConsent. Ontology is basically based on the validity or invalidity of consent. While evaluating this, examples are produced whether consent was given explicitly or implicitly.

This study [11] aims to detect personal data in semi-structured or non structured documents with machine learning algorithms. An algorithm and an architecture to form this algorithm is presented with increased flexibility and robustness, and results are very promising.

The use of personal data is subject to the explicit consent of the person. According to the law, express consent must be obtained in a way that leaves no room for doubt. In

this study [26], terms that can be found in consent management stages were extracted by data mining method and a consent ontology was developed that holds the rules for accessing personal data. The level of importance and sensitivity of personal data is determined by the individual. While personal data owners can keep a personal data they deem important at a high or critical sensitivity level, they can keep an information they deem unimportant at a low or medium sensitivity level.

CHAPTER 3

ONTOLOGY DESIGN

Ontologies are descriptions of concepts and their relationships. Ontologies are created with the goal of allowing the modeling of information about a real or imaginary domain and they encourage the reuse and interoperability of individual modules.

In this thesis, Protégé Desktop [5] is used for creation and editing on the ontology. Protégé is the most popular tool for creating and maintaining ontologies [24]. Protégé has complete support for OWL 2 Web Ontology Language and direct connectivity to description logic reasoners like HermiT. HermiT [1] is an ontology reasoner that uses the Web Ontology Language (OWL) and it can analyze an OWL file to see if the ontology is consistent and find subsumption relationships between classes. HermiT supports Semantic Web Rule Language (SWRL) rules and is used as a reasoner in this work.

Personal data refers to all kinds of data relating to identified or identifiable real persons. Personal data is divided into two; sensitive and non-sensitive personal data. Sensitive personal data is more delicate and can simply be counted as the person's race, political opinion, ethnic origin, religion, health information, sexual life or sexual preferences etc. Non-sensitive personal data can be listed as the person's name, surname, place of birth, age, education information, and marital status etc. The ontology which is created in this thesis is also based on this dual distinction.

The legal knowledge and definitions required while creating and designing this ontology are obtained from the Personal Data Protection Law (Kişisel Verilerin Korunması Kanunu). In the literature, there are previous studies on creating ontology based on laws such as POWER [32] that develop an ontology of legislation for Dutch Tax and

Customs Administration (DTCA).

All the classes created in the ontology and their interdependence and hierarchies can be seen in Figure 3.1

The person and several types of personal data with different levels of sensitivity were used to design the ontology. Ontology is created under 3 main classes as person, personal data and sensitivity level. Sensitive and non-sensitive personal data are the subclasses of personal data as seen in Figure 3.2

Personal data has certain levels of sensitivity. These sensitivity levels are classified as low, medium, high and critical in the ontology we created. For example, a person's name has a low sensitivity level, while a person's health information has a critical sensitivity level. While determining these sensitivity levels, the possible social effects of these data on the person's life were taken into account. It is obvious that the social impact of a person's disclosure of birth place and sexual preference will not be the same.

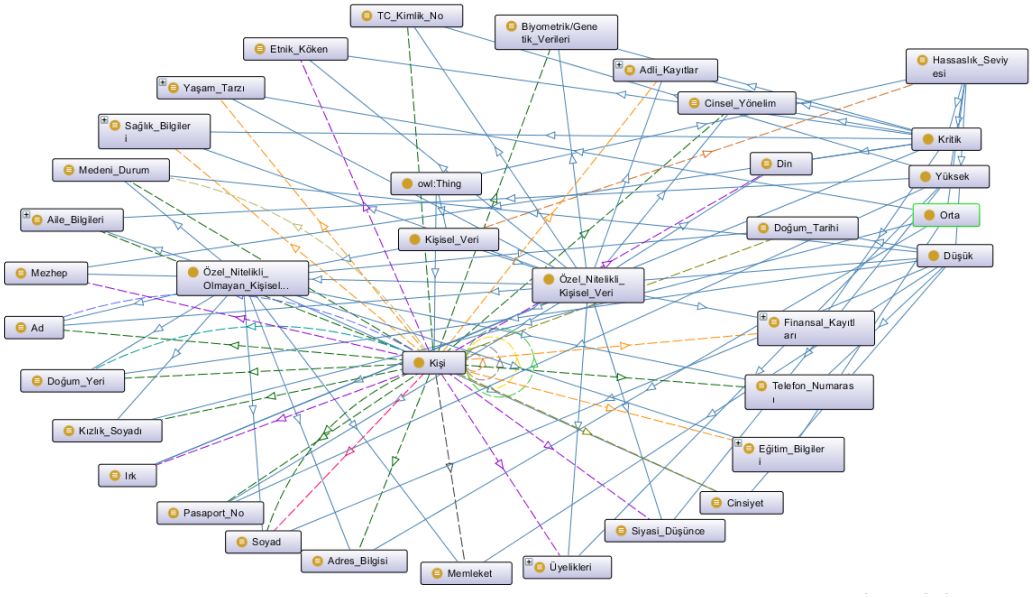

Figure 3.1: Personal Data Ontology

Hierarchy of ontology classes and which personal data are sensitive and which are non-sensitive are seen in the Figure 3.2 The variety of personal data seen here can be easily customized .

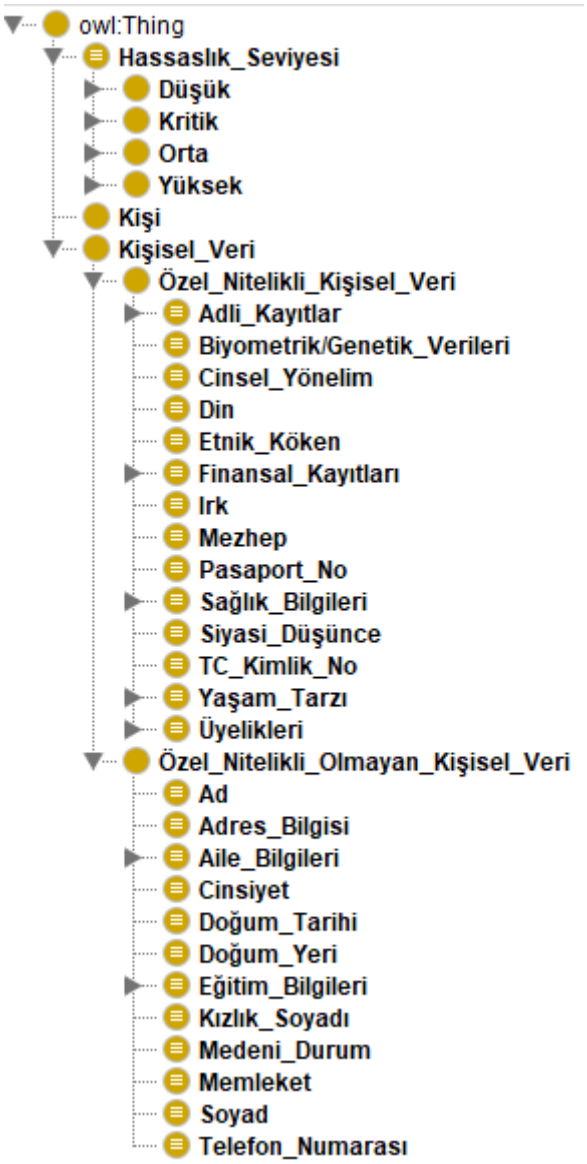

Figure 3.2: Sensitivity Levels of the Ontology

Each personal data class created under the sensitive and non-sensitive personal data classes is also defined to have a subclass of a sensitivity level class. Which personal data has which level of sensitivity is shown in Figure 3.3.

One of the most important benefits of designing and using an ontology for holding and managing personal data is the tendency of this structure to add new personal data. A new personal data extracted for a different purpose can be easily added to this ontology and its sensitivity level can be specified.

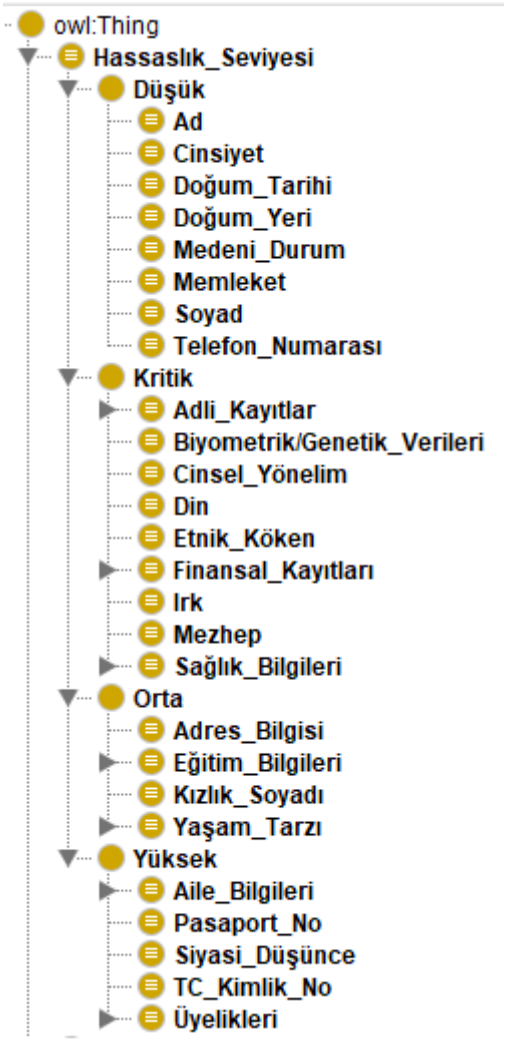

Figure 3.3: Sensitivity Levels of the Personal Data

Each personal data class created in the ontology is created and defined only once and it will be the same for new classes that will be added in the future. While this personal data class is defined under the classes that specify whether it is sensitive or non-sensitive, it is also defined as a subclass of the classes that specify what sensitivity level it has. For example, when a personal data is defined as Name (Ad), it will be

a subclass of the non-sensitive personal data class and will have a low sensitivity level.

CHAPTER 4

SYSTEM DESIGN

We have developed an ontology-based application that receives input documents in raw form, then generates output by detecting the personal data in these documents and populates ontology with these structured output. As the last step, extracts new personal data with Semantic Web Rule Language (SWRL) rules on the populated ontology that is created for personal data domain. Figure 4.1 shows the overall structure of our system.

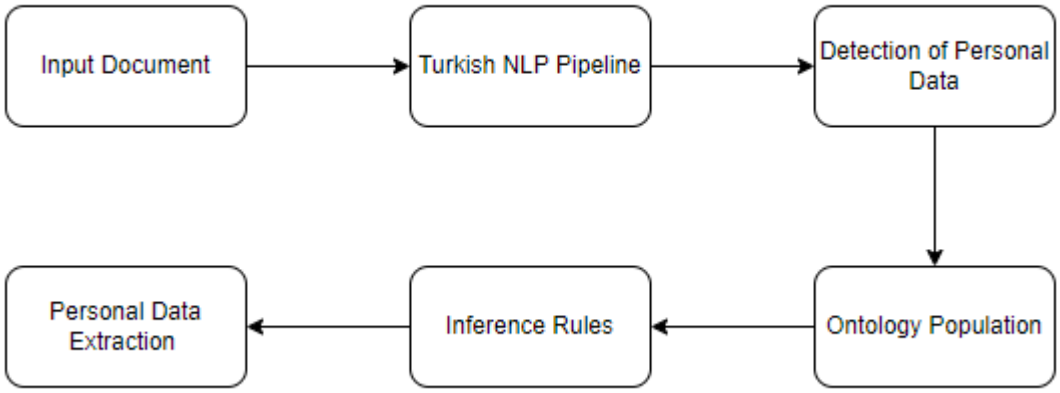

Figure 4.1: System Design Diagram

4.1 Overall Process

We find it helpful to initially see the overall flow of the system before going into detail about each module.

1. Supreme Court decisions were taken on the website of the Yargıtay[1] as an input document. Spelling mistakes and errors were removed by using Zemberek library on these Supreme Court decisions and made ready for use for the next stages.

2. Tokenizer, Normalizer, Morphological Tagger, Named Entity Recognizer, and Dependency Parser are used in the ITU Turkish Natural Language Processing Pipeline to do preprocessing on the input documents.

3. After the pre-processing is finished, each token and named-entity are examined and the personal data in the input documents are extracted by making use of the dependencies issued by the dependency-parser, and these personal data are printed to the output file with a structured format.

4. The ontology is populated with personal data retrieved in structured format from Supreme Court decisions.

5. SWRL inference rules are written in accordance with the personal data desired to be obtained on the populated ontology.

6. By running the reasoner on the populated OWL files, new personal data is extracted from the input documents with the written inference rules.

[1] https://karararama.yargitay.gov.tr/YargitayBilgiBankasiIstemciWeb/ (last visited on 05/05/2022)

4.2 Input Document

The Supreme Court examines appeals in the judicial field and makes the final verdict that are called Supreme Court decisions. We used Supreme Court Decisions as input documents because they contain a wide range of information and personal data. Also, being shared with the public and having a certain format make them suitable for use to detect personal data.

4.3 Turkish NLP Pipeline

Using the ITU Turkish Natural Language Processing Pipeline, which was mentioned in the related work section, the sentences in the input documents is preprocessed. There are multiple ways to use ITU Turkish Natural Language Processing Pipeline, such as using it on the website or using it by uploading files. We used the provided WEB APIs to be able to call it from within our own application. This Turkish NLP pipeline includes: Tokenizer, Normalizer, Morphological Tagger, Named Entity Recognizer and Dependency Parser.

Tokenizer splits sentences into smaller units which are called tokens. This step is important so that we can further analyze the tokens. Normalizer converts this tokens into their canonical form. This step is necessary to improve the processes in the next stages for the token. Morphological Tagger assigns labels to the tokens that morphologically describes them. Named Entity Recognizer identifies the key elements in a text. ITU Named Entity Recognizer supports seven different types of entities such as person, organization, location, date, time, money, percentage. Dependency parser evaluates the dependencies between the phrases of a sentence in order to determine its grammatical structure. The direct relationship that each linguistic unit has is called dependency.

4.4 Detection of Personal Data

After preprocessing of the input document, application analyzes each token and named entities respectively. Named entities holding the person information and their indexes in the dependency list are stored in a map. Then, other named entities are analyzed to determine which personal data they correspond to. The person to which these named entities are directly or indirectly connected is determined in the dependency list. Afterwards, this personal data is added to our list that holds personal data, along with the information of who it belongs to. Also, suffixes or prefixes in the named entities were detected and were erased before being added to the list.

```
Sentence: Dosya kapsamına göre, kayden 15.11.1999 doğumlu olup,
suçun işlendiği 20.08.2014 tarihinde 15 yaşını
ikmal etmediği anlaşılan suça sürüklenen Veli Karadag
hakkında tayin olunan cezanın bozulması lüzumunun ihbar olunduğu anlaşıldı.
BIRTH DATE: [OUT 15.11.1999,1]
PERSON: [PERSON Veli Karadag,1]
```

Figure 4.2: Personal Data Detection Example

In Figure 4.2 shows an example sentence from a Supreme Court decision, person and birth date information are detected in this sentence. Since the dependency index of the person is 1 and the birth date information belongs to this person, it is seen that the dependency index of the birth date information is also 1.

Personal data that cannot be captured with named entity recognition is detected by analysis of normalized versions of tokens. For example, the date of birth information, which we frequently see in many Supreme Court decisions, could not be captured as a named entity due to the format they are in or for other reasons. However, they are detected because the normalized versions of this date of birth information is "DD.MM.YYYY". Afterwards, it is checked whether the words they refer to directly or indirectly are in the dependency list. For example, a text containing a word that means “birth” (Doğum in Turkish) is a personal data whereas other dates are not.

4.5 Ontology Population

The ontology is populated by creating an individual in our ontology for each person information in the list where personal data is kept, and adding the personal information of this person to this individual. Also, object properties shown in Figure 4.3 are defined and used to help create the inference rules.

Figure 4.3: Object Properties

The domain and range of each of the object properties seen in Figure 4.3 refer to a personal data in the ontology. For example, in “babası” (isFather) domain and range are both “Kişi” (Person), in “memleketiOlmak” (isHometown) domain is “Kişi” (Person) and range is “Memleket” (Hometown).

Ontology population examples are shown in Figure 4.4 and Figure 4.5. In the first example, there are two persons (Kişi1, Kişi2). In addition to the information such as name, surname, hometown belonging to Person1 and Person2, object properties that are detected are also seen in this example.

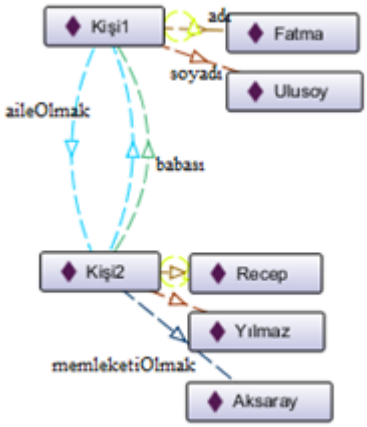

Figure 4.4: Ontology Population Example 1

In the second example, there are three persons (Kişi3, Kişi4, Kişi5). Also, in this example, personal data and object property information of these persons can be seen.

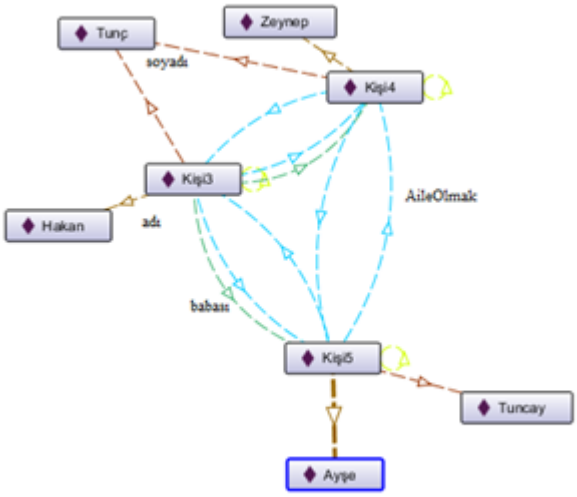

Figure 4.5: Ontology Population Example 2

4.5.1 Automatic Ontology Population

To populate the ontology, the ontology we created is saved in the Turtle Syntax Format. When we save the ontology in this format, we can see how each individual created is actually stored. We have written a population module that converts the format in which we keep the personal data we detected in the previous stage, to the format in which individuals are kept in ontology. While this module creates an auto-creation code for each individual, it also generates a code that creates a Person (Kişi) individual for each person and inserts the personal data within it. After running this module for all identified persons and their personal data, we add the automatic individual creator codes into the ontology saved in specified format. In this way, we automatically populate our ontology instead of manually creating an individual one by one.

For example, in Figure 4.6, name, surname and hometown information of a person can be seen. Name (Ahmet), Surname (Atasoy) and hometown (Kastamonu) information are defined as individuals. Afterwards, a person named Person6 (Kişi6) is created and personal data of Person6 is inserted in this individual by using object properties. When we insert the code seen in Figure 4.6 into the ontology file, these individuals are created in the ontology and the population is finished for this person and personal data.

```
###  http://www.semanticweb.org/pinar/ontologies/2021/3/kisiselveri.owl#Ahmet
kisiselveri:Ahmet rdf:type owl:NamedIndividual .

###  http://www.semanticweb.org/pinar/ontologies/2021/3/kisiselveri.owl#Atasoy
kisiselveri:Atasoy rdf:type owl:NamedIndividual .

###  http://www.semanticweb.org/pinar/ontologies/2021/3/kisiselveri.owl#Kastamonu
kisiselveri:Kastamonu rdf:type owl:NamedIndividual .

###  http://www.semanticweb.org/pinar/ontologies/2021/3/kisiselveri.owl#Kişi6
kisiselveri:Kişi6 rdf:type owl:NamedIndividual ,
                          kisiselveri:Kişi ;
                  kisiselveri:adı kisiselveri:Ahmet ;
                  kisiselveri:kisi kisiselveri:Kişi6 ;
                  kisiselveri:memleketiOlmak kisiselveri:Kastamonu ;
                  kisiselveri:soyadı kisiselveri:Atasoy .
```

Figure 4.6: Individual Examples in Turtle Syntax Format

4.6 Inference Rules

Rules have been defined in Semantic Web Rule Language (SWRL) in order to make inferences of new personal data using the personal data populated in the ontology. In SWRL, every rule is written in terms of OWL concepts such as classes, properties and individuals. These rules can collaborate with reasoners and Hermit is used as a reasoner in this work.

In Figure 4.7, the inference rule is used to infer the person's hometown information. In this rule, we check that these persons are related or not (aileOlmak). If these two people are in the same family, then we check if one of them has hometown information. If any, then, we can infer that the other person from the same family has the same hometown information.

```
kisiselveri:AileOlmak(?kisi_1, ?kisi_2) ^ kisiselveri:kisi(?kisi_1, ?kisi_1) ^ kisiselveri:kisi(?kisi_2, ?kisi_2) ^
kisiselveri:memleketiOlmak(?kisi_1, ?memleket) -> kisiselveri:memleketiOlmak(?kisi_2, ?memleket)
```

Figure 4.7: Rule Example 1

Rules shown in Figure 4.8 and 4.9 infer the marital status of a person. Both rules check the same conditions such as one person is the father of the other, except for the condition which compares their surnames. Rule example 2 checks if the last names of these people are different and rule example 3 checks if they are the same. Using this information, rule 2 deduces the information that the person is married and Rule 3 deduces that the person is single (if the person is female). Here, in order to understand whether the person is male or female, it is necessary to use or create a tool that can distinguish between Turkish names by gender. This can be considered as future work.

```
differentFrom(?soyad1, ?soyad2) ^ kisiselveri:babası(?kisi_1, ?kisi_2) ^ kisiselveri:soyadı(?kisi_1, ?soyad1) ^
kisiselveri:soyadı(?kisi_2, ?soyad2) ^ kisiselveri:kisi(?kisi_1, ?kisi_1) ^ kisiselveri:kisi(?kisi_2, ?kisi_2) ->
kisiselveri:evli_olmak(?kisi_2, kisiselveri:Evli)
```

Figure 4.8: Rule Example 2

So, rule example 2 detects if the person is married, rule example 3 detects if the person is single.

```
sameAs(?soyad_1, ?soyad_2) ^ kisiselveri:babası(?kisi_1, ?kisi_2) ^ kisiselveri:soyadı(?kisi_1, ?soyad_1) ^
kisiselveri:soyadı(?kisi_2, ?soyad_2) ^ kisiselveri:kisi(?kisi_1, ?kisi_1) ^ kisiselveri:kisi(?kisi_2, ?kisi_2) ->
kisiselveri:evli_olmak(?kisi_2, kisiselveri:Bekar)
```

Figure 4.9: Rule Example 3

Scientific research shows that psychological disorders are inherited and transferred to lower generations. For example, while the lifetime prevalence of schizophrenia is 1%, relatives of people with this disease have a higher risk of schizophrenia [25]. Also, in borderline personality disorder, genetic transmission can be up to 40% effective [17]. A person's health information is personal data with a critical sensitivity level, and this information is sensitive for the rest of the family. Especially for genetically inherited psychological diseases, the fact that a member of the family has this disease may indicate that the remaining members are also likely to have this psychological disease. Therefore, with the rules we have written, it is determined that people with such genetically inherited psychological disorders have a risk of having these diseases in their families. In the inference rule example seen in Figure 4.10, we look at whether two people in the same family have borderline personality disorder, and if there is, we determine the risk of this disease for the other person and add this information to the person's personal data.

```
kisiselveri:kisi(?kisi_1, ?kisi_1) ^ kisiselveri:kisi(?kisi_2, ?kisi_2) ^ kisiselveri:AileOlmak(?kisi_1, ?kisi_2) ^
sameAs(?val, kisiselveri:Borderline_disorder) ^ kisiselveri:Olmak(?kisi_1, ?val) -> kisiselveri:Olmak(?kisi_2,
kisiselveri:Borderline_disorder_riski)
```

Figure 4.10: Rule Example 4

In Figure 4.10, rule example 4 is written to infer the risk of having borderline personality disorder, the same rule has been applied to other genetically transmitted mental illnesses such as schizophrenia, and new rules may be added for different diseases in the future.

4.7 Personal Data Extraction

When the rules are run with the help of reasoner, it can make inferences of new personal data that are not directly obtained from the input document. Here, in Figure 4.11, person's hometown information is extracted using Rule Example 1. Also, for all three people, marital status information is extracted.

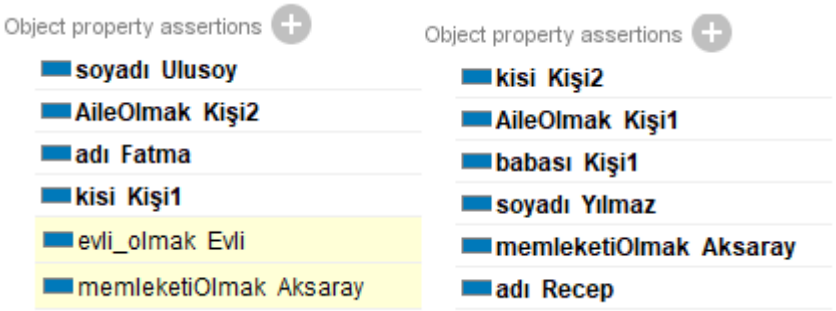

Figure 4.11: Personal Data Extraction Example 1

In Figure 4.11, two different rules are applied. Person2 (Recep Yılmaz) is the father of Person1 (Fatma Ulusoy) and their last names are different, then Rule Example 2 infers that Person1's marital status is married. Also, Rule Example 1 extracts that Person1's hometown information because Person1 and Person2 are from the same family.

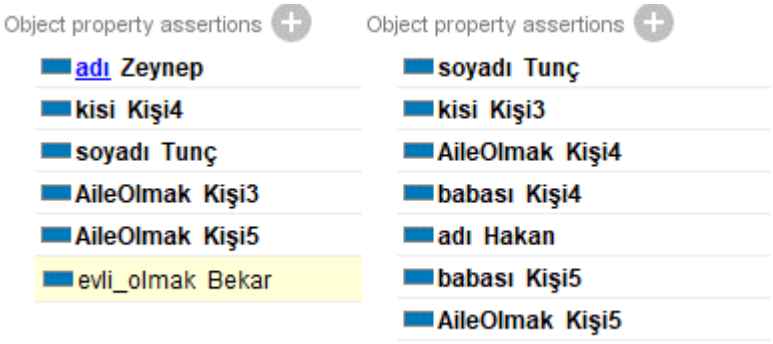

Figure 4.12: Personal Data Extraction Example 2

In Figure 4.12, Person3 (Hakan Tunç) is the father of Person4 (Zeynep Tunç) and their last names are the same, then Rule Example 3 works in this situation and infers that Person4's marital status is single.

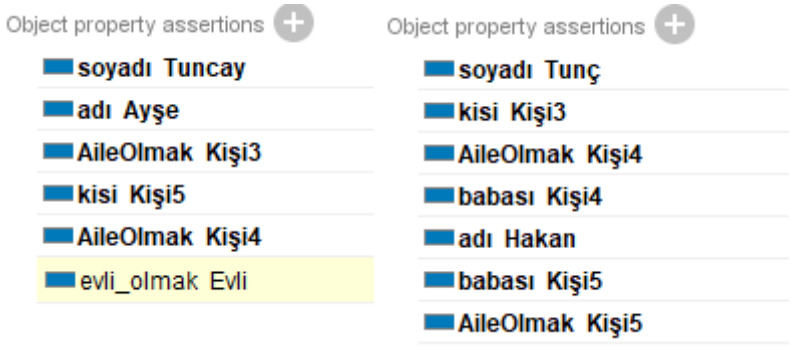

Figure 4.13: Personal Data Extraction Example 3

In Figure 4.13 Person3 (Hakan Tunç) is the father of Person5 (Ayşe Tuncay) and their last names are different, then Rule Example 2 infers that Person5's marital status is married (Evli).

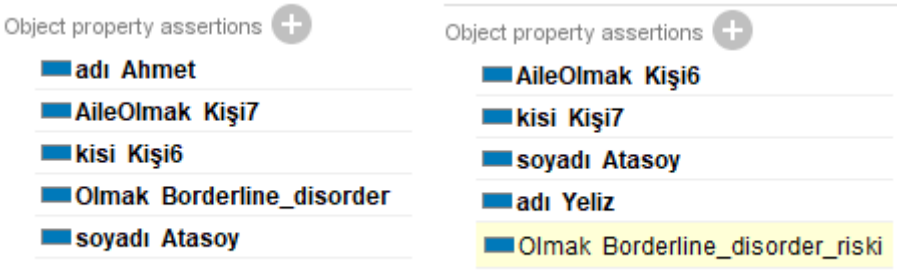

Figure 4.14: Personal Data Extraction Example 4

In Figure 4.14 Person7 (Ahmet Atasoy) is the father of Person6 (Yeliz Atasoy) and he has borderline personality disorder. Since people with a family history of these genetically inherited psychological diseases have a higher risk than others, we can deduce that Person6 has a risk of having borderline disorder. Rule Example 4, which we mentioned in the previous section, can also produce this result for Person6.

CHAPTER 5

EVALUATION

5.1 Test Data

We have used 84 supreme court decisions with a total of 512 personal data to evaluate the performance of our ontology-based approach in detecting personal data. Our personal data detection system was able to recognize 484 of these personal data. The majority of Supreme Court decisions contains the same type of personal data, because of this, name and surname, Republic of Turkey identity number, date of birth, place of birth, telephone number, family relationship, organization and health information could be obtained from these input documents. The personal data detected in the input documents was used to populate the ontology and the results of the inference rules run in the populated ontology were checked to evaluate the inference performance. There were 54 cases where personal data could be extracted with inference rules and 4 of them are for extracting marital status of the person, 5 of them are place of birth information and other 45 of them are for detecting disease risks.

12 of the Supreme Court decisions include cases where one person has bipolar affective disorder, 24 cases with borderline personality disorder and 28 cases with schizophrenia. There was no personal data other than the defendant's name and surname in almost all of these Supreme Court decisions. However, for the rule of inference that determines whether people are at risk of genetic psychological diseases, people from the same family as the person with this disease should also be mentioned in this text. Therefore, new documents were created by blending these Supreme Court decisions with other Supreme Court decisions containing family information. Additions containing family information were made to 8 of the Supreme Court decisions

of people with bipolar affective disorder, to 16 of the Supreme Court decisions of people with borderline personality disorder, and to 21 of the Supreme Court decisions of people with schizophrenia. In this way, it allowed to have a lot of input documents to test the disease risk rules. In addition, it was ensured that the number of personal data that can be detected in Supreme Court decisions increased.

5.2 Evaluation Process

When the Supreme Court decisions are examined, it is seen that most of the decisions contain the same type of personal data. For this reason, with this input document selection, certain types of personal data could be determined and with the rules written using this detected personal data, the inference of a certain number of personal data can be tested. Name and surname, Republic of Turkey identity number, date of birth, place of birth, telephone number, family relationship, organization and health information could be tested. With the inference rules, five rules that extract the marital status, hometown and potential disease risks that I explained above in Chapter 4.5 could be tested.

5.3 Analysis of Results

We analyze the evaluation results in detail in this section. First we analyzed the personal data detection results, then we analyzed the results obtained with the inference rules.

5.3.1 Basic Extraction

The performance of basic extraction of personal data is important because it shows how well the pre-processing process of our application results, the results we obtained from NLP tools and the final detection results we obtained using our own detection methods on them. Also, the personal data obtained at this stage are of great importance as they are used to find the personal data to be obtained with the inference rules.

As we can see in table 5.1, the name and surname information is found correctly with a high rate of 99.37%. TC Identity Number, Phone Number and Organization information was found to be correct with 100 percent, but the effect of the low number of sentences containing this data also has an effect on this result. While family information and place of birth data were detected with high success rates such as 91.84% and 86.67%, date of birth and health information were detected with relatively lower success rates such as 77.78% and 78.13%, respectively.

5.3.2 Extraction with Inference Rules

The performance of Extraction with Inference Rules is important because it shows how successfully the inference rules we have written work with the data we have and reveals the direction in which we should write new inference rules. For example, by looking at the success of the place of birth rule, rules can be written about other personal data expected to be common in the same family, but if the performance of this rule is low, we can predict that we need to write different rules from a new perspective.

Table 5.2 shows that scores of personal data extraction with inference rules have more than 50 percent success in all of them. The family and health information obtained in the previous stage has a great impact on the success rate of the rules that determine the risk of disease.

5.4 Worst Case

In all of our rules that extract personal data, there is a check on whether people are in the same family. If they are not in the same family, all the rules will not be applied and a new personal data will not be extracted. In addition, the fact that the personal data found by basic extraction always consists of the same type of personal data can be seen as a worst case, as it will make it difficult, perhaps even impossible, to produce new rules. Therefore, choosing the right input document type is very important.

5.5 Summary of Evaluation

Our evaluation results of basic extraction of personal data shows that success rate of our system that works before the inference rules is high. Name and surname has only 1 false detection among 158 occurrences and the reason for this one mistake is that the word “Aydın” is used for both the surname of a person and the city where the incident took place in the same sentence. TC Identity Number, Phone Number and Organization information has been found to be completely correct. Unfortunately, this personal data is not used very often in Supreme Court decisions, and this may have an effect on the results being so high. We had a high success rate in detecting family information because most of the Supreme Court decisions similarly stated that the person was the son or daughter of another. Date of birth was more difficult to detect than other personal data. The main reason why the date of birth is difficult to find is that it can be used in many different places of the sentence, completely different from the person who has it, and in some cases it is even indicated only in parentheses (born in 1967). This situation has also made it difficult to follow up on the dependency list and in some cases it has become impossible to go to the relevant person from this personal data. There were two data that we could not detect for place of birth. The main reason why is that it is hard to decide the location information found is the place where the person was born and to distinguish it from the place where the events described may have taken place.

Table 5.2 illustrates that the success of Personal Data Inference Scores actually depends on the success of Personal Data Detection Scores. The correct operation of the rule that deduces the place of birth described in rule example 2 in section 4 actually depends on the correct determination of the person’s hometown and the family information about these persons. In the absence of either of these two information, the rule will not work and will not yield any results. Similarly, in order for the rules that make marital status inference to work, family information and name-surname information must be found completely. The only case where it is not possible to determine whether the person is married or single is the case where the child with whom a family relationship is established is a male. This situation was determined because the expression "son" (oğlu) was used, and in this case, the information showing that the

people belong to the same family was added to the ontology, but the "father"(babası) information that we used for the rule that allows us to extract the marital status was not added. For this reason, the inference rule did not work for this case and could not find any results. It is also important to correctly extract health information and family information for the rules that determine the risk of disease.

When we look at the success of the first part of the system, we see that 484 out of 512 personal data have been identified successfully. So the overall success of the first stage is 94.53%.

When we analyze the overall results of the second section, 40 out of 54 inferences could be made successfully and the success score was found to be 74.07%.

5.6 Discussion

In this thesis, we have presented the results of detecting and managing the personal data in the text using an ontology-based approach. The success rate of the basic extraction results was quite high, but different results can be obtained with different input documents containing more diverse personal data. Ontology-based systems are flexible and open to development, so it will be very easy to adapt this system to the work to be done for different input documents to be used in the future.

Although the score is high in the results obtained using the inference rules, it is predicted that even better results will be obtained when the diversity of personal data is increased and new rules can be written. Since the score of the extraction results obtained with the inference rules is affected by the success rate of the results obtained in the first stage, future studies to increase the score in this section will indirectly affect the results in the second section positively.

Table 5.1: Personal Data Detection Scores

Personal Data	**True**	**False**	**Score (%)**
Name	157	1	99.37
Surname	157	1	99.37
Date of Birth	21	6	77.78
TC Identity Number	36	0	100
Phone Number	3	0	100
Place of Birth	13	2	86.67
Organization	2	0	100
Family Information	45	4	91.84
Health Information	50	14	78.13

Table 5.2: Personal Data Inference Scores

Inferred Extractions of Personal Data	**True**	**False**	**Score (%)**
Place of Birth	3	2	60
Marital Status	3	1	75
Borderline P. Disorder Risk	13	3	81.25
Schizophrenia Risk	16	5	76.19
Bipolar A. Disorder Risk	5	3	62.5

CHAPTER 6

CONCLUSION AND FUTURE WORK

With the development of technology, the use and processing of personal data has become quite common. Such widespread use of personal data brings major risks, as well as providing many conveniences. Risks such as identity theft, discrimination and damage to the reputation of people whose data has been breached are among the most common. In this respect, the protection of personal data is at least as important as the personal data itself.

Extraction of new personal data is possible with the help of system developed in this thesis. Inference rules are created and used in the ontology which is populated from the official documents by identifying the personal data in these documents.

The diversity of personal data, which is the most important factor that can reduce the success rate of the system. When this diversity of personal data is increased, it will be possible to extract much more personal data with different inference rules. Overall success of the inference system with SWRL rules is 74.07% and overall success of the detection of personal data in the first stage is 94.53%.

With the advantage of being a system that includes an ontology, different types of personal data to be added in the future can be easily integrated into the system, and it will be possible to make different personal data inferences by writing new rules based on new requirements. Also, the rule that infers the risks of genetic psychological diseases can be extended with new genetic diseases to be added.

The main reason for choosing the use of ontology is its structure that is suitable for development and change. In the future, by expanding the ontology to cover more types of personal data and writing new inference rules, extraction of new personal data

for different purposes can be ensured. In this study, Supreme Court decisions were preferred due to its wide range of information, being publicly available and being written in a certain format. But, the most of Supreme Court decisions contain the same kind of personal data. In the future, useful results can be obtained by working on different official documents such as health documents, property deeds and identity papers. Since the success of the inference rules depends on the success of the personal data that can be detected in the first stage, it is important to increase the score in the first part in the future and to increase the success rate above 90 percent.

The dependency of the system developed in this study on the document format is very low, and thus it will give good results on different documents containing different personal data. In the future, in order to measure the dependence of this system on the document format, the results can be analyzed by experimenting with different types of documents.

This study may also enable to obtain statistical information about which official documents usually contain which personal data. Future research on this could help make an inference about which documents are more likely to contain sensitive data and what level of sensitivity these personal data have.

REFERENCES

[1] HermiT OWL Reasoner. 2017. http://www.hermit-reasoner.com/ last visited on 15.05.2022.

[2] K.2020/404,T.20.05.2020. https://kvkk.gov.tr/Icerik/6913/2020-404 last visited on 15.05.2022.

[3] Kisisel Verilerin Korunmasi Kanunu. No 6698. 2016. Turkiye https://www.mevzuat.gov.tr/mevzuatmetin/1.5.6698.pdf last visited on 15.05.2022.

[4] The EU General Data Protection Regulation (GDPR). 2016. https://gdpr-info.eu/ last visited on 15.05.2022.

[5] Protégé Ontology Editor. 2019. https://protege.stanford.edu/ last visited on 15.05.2022.

[6] A. A. Akın and M. D. Akın. Zemberek, an open source nlp framework for turkic languages. *Structure*, 10(2007):1–5, 2007.

[7] M. Bauer and S. Baldes. An ontology-based interface for machine learning. In *Proceedings of the 10th international conference on Intelligent user interfaces*, pages 314–316, 2005.

[8] M. Çekin. An analyze of the new turkish code on the protection of personal data nr. 6698 regarding big data and freedom of will. *Journal of Istanbul University Law Faculty*, 74(2):629–644, 2016.

[9] G. Çelikkaya, D. Torunoğlu, and G. Eryiğit. Named entity recognition on real data: a preliminary investigation for turkish. In *2013 7th International Conference on Application of Information and Communication Technologies*, pages 1–5. IEEE, 2013.

[10] K. Cemil. Avrupa birliği veri koruma direktifi ekseninde hassas (kişisel) veriler ve işlenmesi. *Journal of Istanbul University Law Faculty*, 69(1-2):317–334, 2011.

[11] F. Di Cerbo and S. Trabelsi. Towards personal data identification and anonymization using machine learning techniques. In *European Conference on Advances in Databases and Information Systems*, pages 118–126. Springer, 2018.

[12] M. V. Dülger. Avrupa birliği genel veri koruma tüzüğü bağlaminda kişisel verilerin korunmasi. *Yaşar Hukuk Dergisi*, 1(2):71–174, 2019.

[13] G. Eryiğit. Itu turkish nlp web service. In *Proceedings of the Demonstrations at the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1–4, 2014.

[14] C. H. Hwang. Incompletely and imprecisely speaking: using dynamic ontologies for representing and retrieving information. In *KRDB*, volume 21, pages 14–20. Citeseer, 1999.

[15] S. Kara, Ö. Alan, O. Sabuncu, S. Akpınar, N. K. Cicekli, and F. N. Alpaslan. An ontology-based retrieval system using semantic indexing. *Information Systems*, 37(4):294–305, 2012.

[16] M. T. Kartal. Kişisel verilerin korunması: Türk bankacılık sektörü üzerine kavramsal bir değerlendirme. *Uluslararası Ekonomi ve Yenilik Dergisi*, 4(1):1–18, 2018.

[17] S. S. Köse and O. Erbaş. Personality disorders diagnosis, causes, and treatments. *Demiroglu Science University Florence Nightingale Journal of Transplantation*, 5(2):022–031, 2020.

[18] D. Küçük, G. Jacquet, and R. Steinberger. Named entity recognition on turkish tweets. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 450–454, 2014.

[19] V. Kumar et al. Ontology based public healthcare system in internet of things (iot). *Procedia Computer Science*, 50:99–102, 2015.

[20] C. Kuner, D. J. B. Svantesson, F. H. Cate, O. Lynskey, and C. Millard. Machine learning with personal data: is data protection law smart enough to meet the challenge? *International Data Privacy Law*, 7(1):1–2, 2017.

[21] D. Kılınç. Anayasal bir hak olarak kişisel verilerin korunması. *Ankara Üniversitesi Hukuk Fakültesi Dergisi*, 61(3):1089–1172, 2012.

[22] A. Maedche, G. Neumann, and S. Staab. Bootstrapping an ontology-based information extraction system. In *Intelligent exploration of the web*, pages 345–359. Springer, 2003.

[23] A. Maedche and S. Staab. The text-to-onto ontology learning environment. In *Software Demonstration at ICCS-2000-Eight International Conference on Conceptual Structures*, volume 38, pages 890–930. sn, 2000.

[24] M. A. Musen. The protégé project: a look back and a look forward. *AI matters*, 1(4):4–12, 2015.

[25] N. Norton, H. J. Williams, and M. J. Owen. An update on the genetics of schizophrenia. *Current opinion in psychiatry*, 19(2):158–164, 2006.

[26] C. Özgü and O. Emre. Kişisel verilerin korunması kanunu için onam ontolojisi geliştirimi. *Dokuz Eylül Üniversitesi Mühendislik Fakültesi Fen ve Mühendislik Dergisi*, 21(62):559–575.

[27] H. J. Pandit, C. Debruyne, D. O'Sullivan, and D. Lewis. Gconsent-a consent ontology based on the gdpr. In *European Semantic Web Conference*, pages 270–282. Springer, 2019.

[28] A. Rhayem, M. B. A. Mhiri, M. B. Salah, and F. Gargouri. Ontology-based system for patient monitoring with connected objects. *Procedia computer science*, 112:683–692, 2017.

[29] G. A. Şeker and G. Eryiğit. Initial explorations on using crfs for turkish named entity recognition. In *Proceedings of COLING 2012*, pages 2459–2474, 2012.

[30] A. Sunitha and G. S. Babu. Ontology-driven knowledge-based health-care system, an emerging area-challenges and opportunities-indian scenario. *The Inter-*

national Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences, 40(8):239, 2014.

[31] M. Temizsoy and I. Cicekli. An ontology-based approach to parsing turkish sentences. In *Conference of the Association for Machine Translation in the Americas*, pages 124–135. Springer, 1998.

[32] T. M. Van Engers, P. J. Kordelaar, J. den Hartog, and E. Glassée. Power: Programme for an ontology based working environment for modeling and use of regulations and legislation. In *Proceedings 11th International Workshop on Database and Expert Systems Applications*, pages 327–334. IEEE, 2000.

[33] D. C. Wimalasuriya and D. Dou. Ontology-based information extraction: An introduction and a survey of current approaches, 2010.

[34] F. Wu, R. Hoffmann, and D. S. Weld. Information extraction from wikipedia: Moving down the long tail. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 731–739, 2008.

[35] R. Yeniterzi. Exploiting morphology in turkish named entity recognition system. In *Proceedings of the ACL 2011 Student Session*, pages 105–110, 2011.

[36] F. Zeshan and R. Mohamad. Medical ontology in the dynamic healthcare environment. *Procedia Computer Science*, 10:340–348, 2012.

[37] J. Zhao, R. Mortier, J. Crowcroft, and L. Wang. Privacy-preserving machine learning based data analytics on edge devices. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 341–346, 2018.