TÜBİTAK

Turkish Journal of Electrical Engineering & Computer Sciences

http://journals.tubitak.gov.tr/elektrik/

Research Article

Turk J Elec Eng & Comp Sci
(2022) 30: 167 – 183
© TÜBİTAK
doi:10.3906/elk-2102-54

Deep learning-aided automated personal data discovery and profiling

Apdullah YAYIK[1,*], **Vedat AYBAR**[2], **Hasan APİK**[2], **Sevcan İÇÖZ**[2]
Bekir BAKAR[2], **Tunga GÜNGÖR**[3]
[1]Huawei R&D Center, Istanbul, Turkey
[2]Mobildev, Istanbul, Turkey
[3]Computer Engineering, Boğaziçi University, Istanbul, Turkey

Received: 11.02.2021 • **Accepted/Published Online:** 09.08.2021 • **Final Version:** 19.01.2022

Abstract: In Turkey, Turkish Personal Data Protection Rule (PDPR) No. 6698, in force since 2016, provides protection to citizens for the legal existence of their personal data. Although the law provides excellent guidance, companies currently face challenges in complying with its regulations in terms of storing, sharing, or monitoring personal data. Since any specially designed software with wide industrial usage is not on the market, almost all of the companies have no other choice but to take expensive and error-prone operations manually to ensure their compliance. In this paper, we present an automated solution to facilitate and accelerate PDPR compliance. In a structured or unstructured document, words or phrases that could include personal data entities are tagged with the help of a Bi-LSTM based named entity recognition model and a rule-based matching component that employs contextual analysis. To find associations in personal data and obtain individual personal profiles, these entities are divided into categories according to their confidence levels. Personal profiles are constructed using an approach similar to clustering. It treats the personal data categories with high identification levels as separate clusters and finds related personal data entities at the left and/or right of its contexts. We evaluated the system on a data set formed of 70 documents of different types and personal data entities. We obtained 91.76 % micro-averaged F1-measure for personal data entity classification and 72.46 % accuracy for profile extraction. We also performed experiments related to the performance of the named entity recognition and to the time complexity of the overall system on a data set formed of 33K documents.

Key words: Turkish Personal Data Protection Rule, named entity recognition, rule-based matching, personal data associations, relation extraction

1. Introduction

Due to concerns in providing data privacy and safety, the evaluation of compliance with the applicable regulations is becoming increasingly important. The Turkish Personal Data Protection Rule (PDPR)[1] has been developed to regulate personal data privacy across Turkey to offer greater safeguards and capabilities for individuals to track their personal data usage in the light of emerging technological advances. While beyond any doubt it is useful to people in several ways, the fact with the PDPR is that organizations are struggling with understanding what compliance is required and how it needs to be implemented legally with ease and efficiency. Organizations need to create simple solutions for customers to know what their rights are in terms of their

*Correspondence: apdullah.yayik@huawei.com

[1] Personal Data Protection Rule No. 6698 [online]. Website https://www.mevzuat.gov.tr/MevzuatMetin/1.5.6698.pdf [accessed 02 February 2021]

personal data, and to respond within 30 days (under Article 13 at PDPR) to any request submitted when any incident related to data leakage occurs. It is additionally important to take preincident measures such as data consenting or anonymization. Under the terms of the PDPR, processing personal data such as storing, sharing, or monitoring is allowed only if there is explicit consent provided on a legal basis. Such consent is considered explicit only if it is completely transparent, freely given, specific, informed, and unambiguous. In order to be able to obtain consent for a piece of personal data, first the data should be ”discovered”. This necessitates answering the following key questions, which enable extracting the data and identifying the owner: what type of personal data is used (severity), where is it located, how is it associated with other data, and which personal profile does it belong to? Besides, these questions should be answered over time repeatedly to keep up with the changes in the data sources.

Well-organized personal data discovery solutions may be used at a data center where structured data sources are located. It is a fact that personal data types and severity even in structured data sources preserve ambiguity. This ambiguity exponentially grows when it comes to unstructured data sources. For instance, old e-mail attachments often contain personal data and are easily forgotten about, and are usually ignored by data controllers who are responsible for data under the PDPR. This common reason and many similar ones are frequently encountered and that is why organizations are having difficulty in compliance with PDPR. To the best of our knowledge, there is a lack of an automated solution to address such difficulties on the market in Turkey. This explicit gap will be much more salient when individuals begin to request their respective rights under the PDPR.

This paper proposes automated data mapping and discovery techniques with the help of state-of-the-art machine learning approaches and rule sets. For this aim, an effective discussion among academics, industrial IT practitioners, and lawyers about the roles of machine learning and rule sets in PDPR compliance has been made. As a result, a methodology was proposed and software was developed in line with the joint decisions. To discover personal data, a bidirectional long short-term memory network (Bi-LSTM) based named entity recognition (NER) system alongside rule sets relying on pattern matching and contextual analysis was designed. The robustness of the NER model was increased using dictionary-based controls and was quantized using an integer weight representation with linear normalization. Each personal data was categorized based on predefined confidence levels and relations between data entities were employed to create clusters and personal profiles. The proposed hybrid system was tested in terms of execution time and performance on extracting personal data together with personal profiles. We observed that promising results were achieved that can be used for PDPR compliance.

The main contributions of this study are highlighted as follows: (1) proposing a model for extracting 77 types of personal data entities in 11 categories within the scope of the PDPR, (2) designing a hybrid approach that integrates the machine learning model with a rule-based model and dictionaries, and (3) using a heuristic method that is based on assigning confidence levels to personal data entities and employing associating and profiling of personal data.

In Section 2, basic concepts of the PDPR that comprise the rights of data subjects are overviewed. In Section 3, entity type classification and NER researches for the Turkish language are given. In Section 4.1 confidence levels and types of personal data are defined. In Section 4.2, the NER model and the rule sets used to extract personal data entities are described. Section 4.3 explains data associating and profiling. Section 5 explains the experiments and gives the test results, which is followed by a discussion of the results in Section 6. Section 7 is for the conclusions.

2. Turkish Personal Data Protection Rule (PDPR)

The PDPR is a legislation comprised of 33 articles divided into 7 chapters. It is the primary law in Turkey that regulates the privacy of personal data to protect the rights and freedoms of persons. It specifies the core principles and procedures that not only private or state-based companies but also individuals processing personal data have to comply with. Accordingly, in Article 3 personal data coverage is defined in a broad sense in that any data is accepted as personal if it can be associated with an identifiable person. A data controller is a person who is responsible for correctly setting-up and managing the data filing system. The law defines in Article 3 data subject as a natural person whose personal data is wholly or partially being processed. Under Article 11 of the law, the data subject is entitled to apply to the data controller to be informed about whether his/her personal data is being processed without any consent. The data subject has even the rights to request the wipe-out of his/her personal data if any legal consent does not exist. If a satisfactory response is not given by the data controller to the data subject within 30 days, the data subject has the right to make a written complaint to Personal Data Protection Board under Article 13. By the Board, if it is determined that the data controller has a fault and the applicant's request is not being responded in an appropriate way, heavy financial sanctions can be applied.

3. Related works

This section provides related works about entity type classification (ETC), named entity recognition (NER), and entity profiling (grouping).

3.1. Entity type classification

In 2012, Ling and Weld proposed [1] a fine-grained ETC approach using a conditional random forest model for the English language. They achieved tagging 112 types of entities. In 2012, Yosef et al. [2] proposed an ETC approach for 505 entity types which was employed as a hierarchical method. These studies aimed at more informative tagging for the English language compared to the coarse-grained NER researches. In 2016, Kalender and Korkmaz [3] implemented linear classifier models for the classification of 100 types of entities for the Turkish language. In our study, we designed a Turkish entity type classifier with a combination of a Bi-LSTM based NER model and a rule sets approach for 77 types of entities.

3.2. Named entity recognition

Named entity recognition (NER) studies face challenges for the Turkish language due to its agglutinative morphology. It is possible to create a valid Turkish word by appending multiple suffixes to a root. For instance, the word ”düşünemediklerimizdendir” *(it is one of those that we could not think)* is a valid surface form derived from the verbal root ”düşün” *(think)* with appending seven suffixes. This agglutinative nature makes the job of NER taggers difficult due to the nonexistence of a canonical form for most of the named entities.

In 2003, Gökhan et al. [4] released the first data set for the Turkish NER task that includes almost 30K sentences, 500K words of which 66K are unique, and three entity types with the following distribution: 24K person, 16K organization, and 13K location. From 2003 to 2010, mostly rule-based Turkish NER models appeared. In 2008, Bayraktar and Temizel [5] proposed a local grammar-based method that relies on constructing polylexical parts having frozen properties on Turkish financial texts and achieved an accuracy of 81.9 %. In 2009, Küçük and Yazıcı [6] built a NER system, exploring numerical and temporal patterns together with verifying via dictionaries. They achieved an accuracy of 75.4 %, however the rules explicitly induced performance degradation. From 2010 to 2014, mostly statistical machine learning models such as conditional random fields (CRF) and hidden markov models (HMM) were employed to create adaptive learning rules instead of hand-crafted rules designed by domain experts. In 2011, Tatar and Çiçekli [7] proposed an automated rule designing approach that relies on CRF and achieved an accuracy of 91.8 %. In 2012, Şeker and Eryiğit [8] proposed a hybrid approach that also relies on CRF and hand-crafted morphology dependent features and reached an accuracy of 91.94 %.

From 2014 onwards, various architectural types of neural networks from classical to modern deep learning models have been designed. In 2014, Demir and Özgür [9] trained averaged perceptron with the word embeddings and features that are independent of the language itself. As a result, they achieved an accuracy of 91.8%. In 2017, Şeker and Eryiğit [10] presented a CRF based model with enriched lexical and morphological features and reached an F1-measure of 92 %. In 2018, Güngör et al. [11] built a stacked recurrent neural network and CRF model fed with word embeddings and reached an F1-measure of 93.4 %. In 2018, Akkaya [12] proposed a transfer learning model that uses fasttext, morph2vec, and character-level embeddings and relies on a stacked Bi-LSTM and CRF model for noisy textual content. In 2019, Güngör et al. [13] proposed an approach that relies on a 3-layer Bi-LSTM and a CRF model with word, character, and morphological embeddings together. This study achieved an accuracy of 92.93 %. The contribution was that it showed performance improvement by using embeddings related to morphological information.

In 2020, Schweter [15] proposed the BERTurk model and achieved an accuracy of 95.4 %. Yamada et al. [14] proposed entity-aware self-attention mechanism for BERT and achieved high accuracies on several benchmark data sets for English language. In 2020, Yu et al. [15] utilized contextual embeddings and biaffine model with multi-layer Bi-LSTM and reached promising results. Luoma and Pyysalo [16] proposed a BERT model to capture cross-sentence relations. They proved on several benchmark datasets that cross-sentence information improves the performance of the NER model. In our study, we designed Bi-LSTM models with several different architectures. Fasttext embeddings [17], where each word is represented as the sum of its character n-gram embeddings, were used as word representations.

3.3. Personal data discovery

Contrary to the former regulations in Turkey and across the Europe, PDPR and General Data Protection Rule (GDPR)[2] enforce protecting personal data in both structured and unstructured data sources. We have encountered few researches that aim to accomplish the task of discovering personal data through building a system that provides compliance with GDPR. In 2018, Olby and Thomander [18] investigated the usage of a NER model to provide GDPR compliance by discovering personal data in e-mails. However, this study declared inadequate performance measures. In the same year, Dasgupta et al. [19] successfully created a model that relies on NER to discover 134 types of personal data entities in 10 categories to provide GDPR compliance for English language. They also released an annotated data set to help researchers make studies to improve their results. In 2020, Mariana et al. [20] employed a hybrid approach that consists of rule sets, lexical based models, and machine learning algorithms to detect and make notifications about personal data for Portuguese language. This study extracts only 11 types of personal data entities in 3 categories. To the best of our knowledge, our

[2] General Data Protection Rule [online]. Website https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32016R0679 [accessed 02 February 2021]

study is the first attempt to construct a personal data discovery system to comply with PDPR for the Turkish language.

4. Methodology

The general system architecture is shown in Figure 1. As shown in the figure, the system is built as a pipeline that retrieves a document, employs entity type classification (ETC) using NER and rule matching, and groups extracted entities to obtain individual personal profiles through data associating and profiling. In the following subsection, confidence levels and types of entities are defined concerning the PDPR. Then, the modules shown in Figure 1 are described in detail.

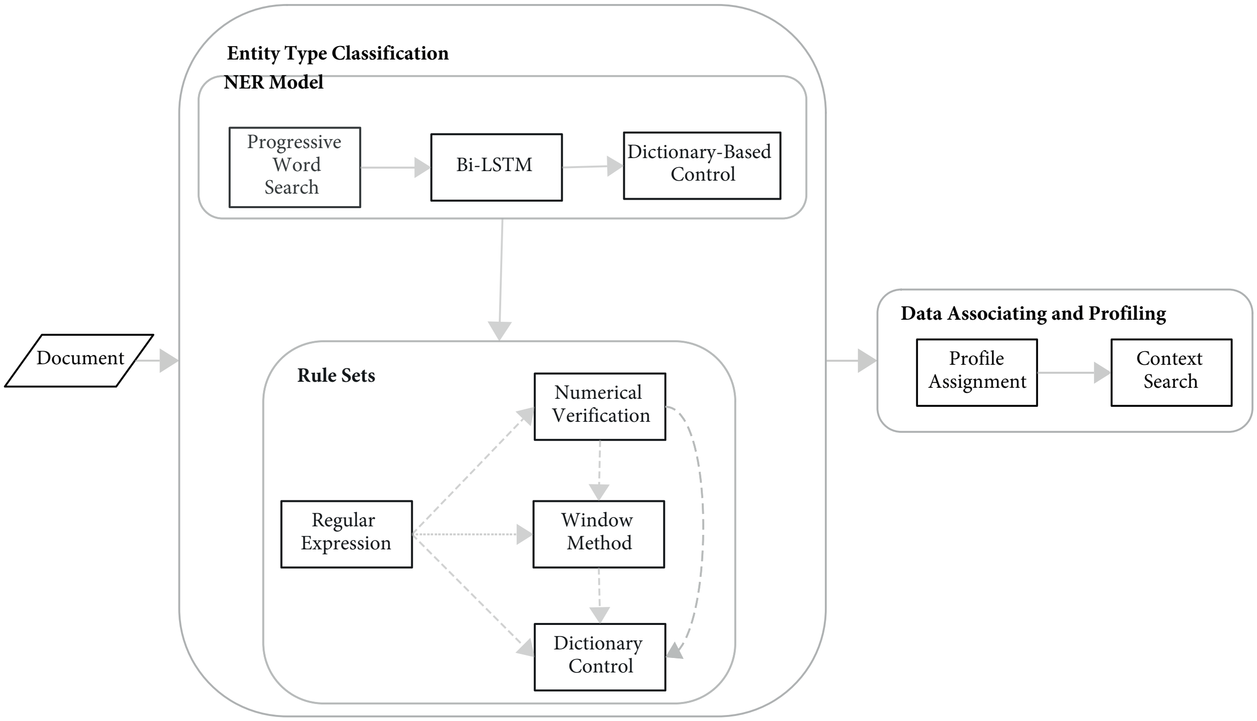


Figure 1. General system architecture.

4.1. Defining confidence levels and types of entities

In this study, it is aimed to construct personal profiles by grouping the personal data extracted in different documents on an individual basis. For this reason, we expand the personal data definition (under Article 3 at PDPR) by adding the concept of confidence level which indicates the capability in, directly or indirectly, identifying a natural person. The confidence level is expressed by using color codes for personal data such that different colors denote different types of entities.

Definition 1 (Red personal data) *Red personal data represents personal information that can identify the person directly without any further reference; such as the identification number, e-mail, IBAN of that person.*

Definition 2 (Green personal data) *Green personal data represents personal information that can identify* *the person directly together with at least one reference in the same context; such as student number with the university name, bank account number with bank name and bank branch code of that person.*

Definition 3 (Orange personal data) *Orange personal data represents personal information that can identify a person indirectly by association with red or green personal data semantically; such as name, blood group, location, criminal record, the sexual or religious tendency of that person.*

By examining the categories determined in the PDPR, we introduce color levels for 77 types of personal data entities in 11 categories as shown in Table 1.

Table 1. Personal data entity types. In PDPR health, religion, and criminal offence are evaluated as special personal data types, but the rest as normal personal data.

finance	**contact**	**id**	**other**
IBAN no	phone number home	id number	vehicle plate no
tax no	phone number mobile	human name surname	organisation
bank credit card number	e-mail	human name	municipality
bank credit number	IMEI number	surname	institution
bank customer number	tel number (fax)	father name	industry
bank account number	address	mother name	service
policy number	address no	date of birth	military status
bank name		place of birth	exemption
insurance company	**location**	registered province	postpone
central registry system no	location	registered distinct	demobilization
credit number		sub-distinct	marriage date
bank name	**cyber operation security**	driving license	divorce date
customer number	ip address (v4)	document no	
bank name	ip address (v6)	paper no	**marketing**
policy number		register no	organisation
insurance company	**profession**	class	accommodation
bank account number	university name	passport no	restaurant
bank branch name	primary school	mother maiden name	travel
bank branch code	high school	volume no	mall
bank name	diploma number	household no	
	graduated program	id card serial no	**criminal offence**
health	education type	gender	record
blood group	class number	age	achieve record
drug	studying program		
disease		**religious**	
organization health		religious tendency	

4.2. Entity type classification

ETC is the task of assigning categories to words or phrases in a document. Generally, ETC reveals what personal data is used (types and severity) and where it is located at. In this study, ETC in documents was employed using the NER model and the rule sets approach together. Due to not having certain patterns and common homonym usage, researchers mostly face challenges to tag entities such as person, organisation, and location.

The usage of machine learning (ML) based approaches is a requirement for such ambiguous entities. For this reason, recognition of these entities is addressed as a ML problem as described in Section 4.2.1. Following this, the rule sets approach used is explained in Section 4.2.2.

4.2.1. Named entity recognition

Named entity recognition (NER) is a labeling operation that aims to detect special types of atomic elements (words or phrases) and to assign them to predefined entity types [21]. NER is referred as a basic task before implementing more complex natural language understanding tasks. This task was extensively studied and evaluated as a successfully handled problem for the English language. Unfortunately, when it comes to morphologically rich languages like Turkish, additional researches are still required due to the complicated structure of the language and few language processing tools and data sets available. The motivation of the NER is the need to detect whether an entity in a given document is a person's name, location, organization, or none of these, which can not be easily detected. In this study, these entities may have orange or red confidence levels. Every single entity labeled as a named entity is first evaluated as having orange confidence. These entities may be part of a relevant personal profile. When two consecutive entities of person name type are detected, it is checked whether they form a name and surname pair, in which case their confidence levels are changed from orange to red. In this way, entities labeled via NER may have orange or red confidence regarding these cases. That is why, in our study, NER is a crucial step in that it defines the starting point of creating personal profiles.

One of the challenges in the NER and similar tasks is the unknown words problem. In Turkish, the unknown word problem is more severe than languages like English since several different surface forms may be obtained from a single word that does not exist in the word embedding repository. In our study, to avoid such out of vocabulary cases, a progressive search method shown in Figure 2 is employed.

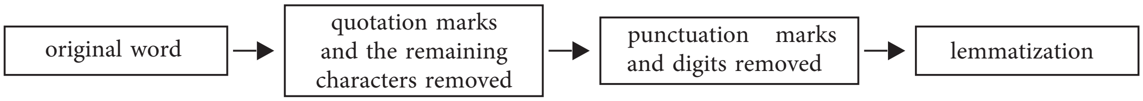

original word
quotation marks and the remaining characters removed
punctuation marks and digits removed
lemmatization

Figure 2. Steps of the proposed progressive search method for unknown words. Beginning from the left, the embedding of the first word form found in the embedding database is retrieved. The following examples provide information on how to read the algorithm. If the original word ”İstanbul'dan” (*from Istanbul*) is searched but not found, the quotation marks and remaining characters are removed, so the word ”İstanbul” is retrieved instead of the original word. If ”Okuldakilerdenmiş” (*would be one of those at school*) is searched but not found, the quotation marks and the remaining characters, punctuation marks, and digits are removed, and then lemmatized, so ”Okul” (*school*) is retrieved instead of the original word.

It aims to find a word vector semantically closer to the unknown word in the word embedding lexicon. The intuition is that it is better to use a related word rather than treating it as an unknown word. The words are represented using pretrained word embeddings and fed into the Bi-LSTM network. The main difference from previous NER studies that are based on LSTM models comes from the representation scheme used for unknown words as explained above. The dropout layer [22] is stacked at the next layer to regularize weights, 10 % of them are randomly selected and values are set to zero to avoid possible overfitting. Besides, a classical fully-connected layer (FCL) and a softmax layer are employed on the top of the LSTM layers. Figure 3 shows the architecture of the NER model.

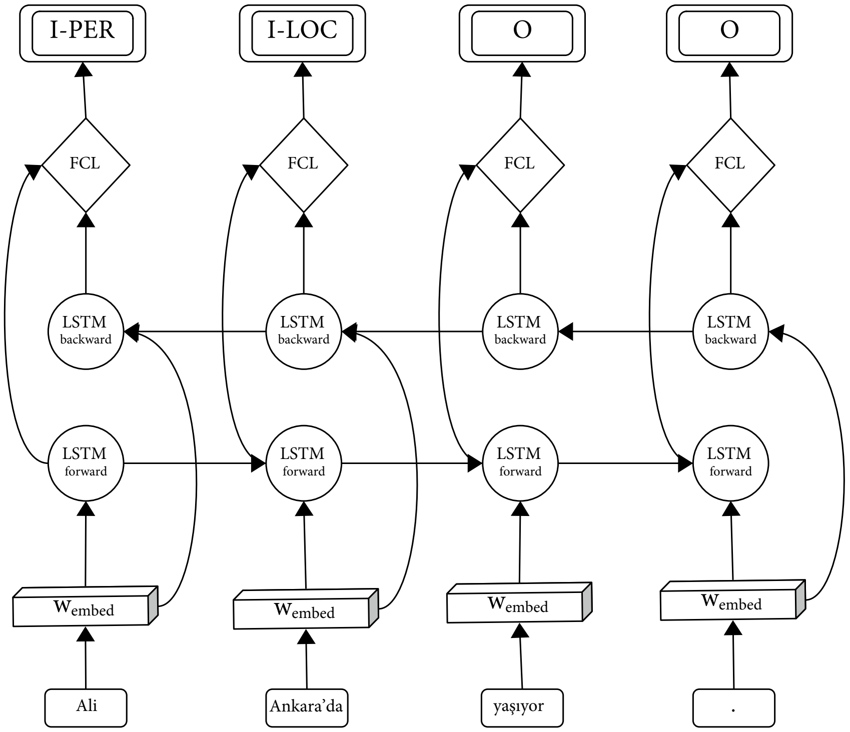


Figure 3. Bi-LSTM architecture.

4.2.2. Rule sets

In this section, the rule sets used for personal data extraction based on pattern matching and contextual analysis are explained. Mainly, personal data entities are detected by regular expressions and confirmed by using one or more strategies as shown in Table 2.

(1) Numerical verification

The personal data entities identity number[3], IMEI number [23], tax number[3], IBAN no[4], IP address version 4 [24] and 6 [25], and bank credit card number [23] are verified using the corresponding numerical techniques given in the stated references.

Table 2. Personal data entity detection and verification approaches. For each approach, example entities are shown. For instance, the entity type "university name" can be assigned to a phrase only if the corresponding regular expression matches with the phrase, window method detects the proper words in context, and dictionary control verifies the phrase.

Regular expression + numerical verification	Regular expression + window method	Regular expression + window method + dictionary control
id number IMEI number IBAN number	vehicle plate student number blood group	university name bank name organisation military

[3] TC Kimlik Numarası ve Vergi Numarası Doğrulama [online]. Website https://www.simlict.com [accessed 02 February 2021]

[4] ISO 13616-1:2020 Financial services International bank account number, Part 1: Structure of the IBAN [online]. Website https://www.iso.org/standard/81090.html [accessed 02 February 2021]

(2) Window method

The window method is used when a personal data entity can only be verified by looking at its context. The entity is verified using a predefined window size for its left and/or right context (words or phrases) in the paragraph where it is located at. For instance, although vehicle plates in Turkey share a unique pattern (province code - letters - digits), verification is needed to avoid false positives. A naive solution for this task may be storing millions of valid vehicle plates which is an expensive and infeasible solution. Using the window method, predefined keywords related to the entity that can semantically verify it are searched case-insensitively in the left and/or right contexts of the entity. The window size for each personal data entity was empirically determined.

The keywords used to verify vehicle plates are ”plaka” (*plate*), ”araç” (*vehicle*), and ”sürücü” (*driver*). In addition to vehicle plate, student number (keywords ”öğrenci no” (*student no*), ”öğrenci numarası” (*student number*)), blood group (keywords ”rhesus”, ”trombosit”, ”kan bağışı” (blood donation), ”kan grubu” (blood group)), central registry system no (keyword ”mersis”), credit number (keywords ”kredi no” (*credit no*), ”nolu kredi” (*no of credit*), ”kredi numaralı” (*credit number*), ”numaralı kredi” (*number of credit*)), and account number (keywords ”hesap no” (*account number*), ”hesap numaralı” (*no of account*), ”hesap numara” (*account number*), ”numaralı hesap” (*number of account*)) are verified using the window method.

(3) Dictionary control

Dictionary control is used in conjunction with the window method to avoid false positives. For instance, to extract entities from the sentence ”Kuveyt Türk Katılım Bankası hesabından EFT ..” (*EFT from the account of Kuveyt Türk Katılım Bank ..*), the left context of the word ”bankası” with a margin of maximum 10 words is searched from a dictionary of bank names. If a group of consecutive words in the left context appear as an entry in the dictionary, then it is accepted as a valid bank name. In addition to bank names, universities, insurance companies, hotels, hospitals, municipalities, military units, high-schools, etc. are classified likewise with keywords and dictionary controls of the left and/or right context using the related dictionaries.

Below, two example sentences that are tagged using the NER model and the rule sets approach together are shown. For each, the entity type and the confidence levels (colors) are indicated.

Example 1: Gazi Hastanesinde covid-19 tedavisine alınan Hasan Yıldırım eski Mobildev çalışanıydı (*Hasan Yıldırım, who was taken to covid-19 treatment at Gazi Hospital, was a former Mobildev employee*)

Gazi Hastanesinde	covid-19	tedavisine	alınan	Hasan Yıldırım	eski	Mobildev	çalışanıydı
organisation health	disease	O	O	name surname	O	organisation	O

Example 2: 3869187513 vergi numaralı Ali Yılmaz’ın Are Elektrik Üretimde Müdür olduğu bildirildi (*It was reported that Ali Yılmaz whose tax number is 3869187513 is a manager at Are Elektrik Üretim*)

3869187513	vergi	numaralı	Ali Yılmaz’ın	Are Elektrik Üretimde	Müdür	olduğu	bildirildi
tax number	O	O	name surname	organisation	O	O	O

4.3. Data associating and profiling

After the personal data extraction process explained in the previous section, each document D is represented as a list of entities $[e_1, e_2, \ldots, e_n]$ where e_i, $1 \leq i \leq n$, is a positionally tagged personal data entity with information of color, type, and content. Now we need to group the entities that belong to the same person to obtain the profile of that person. The profiling algorithm is shown in Algorithm 1. The profiling module AssignProfiles is called with a document D which is a list of extracted entities as stated above. The algorithm returns P which is a list of personal profiles $[p_1, p_2, \ldots, p_m]$ where each p_i, $1 \leq i \leq m$, consists of personal data entities about a person. P is initialized as empty. The entities in the document are processed in a loop. Whenever a red or green entity is found, the existing profiles are scanned to see whether this entity has already been inserted into a profile before. If no, a new profile is created and the entity is put into that profile. In either case, a module named ContextSearch is called to check for related information in the left and/or right context of the entity. In this study, the left and right relation window sizes are specific to each entity type. Thus, the left and right margins are first assigned accordingly in the module. It is assumed that all the entities within this range belong to the same profile provided that there is no hot border entity. We use the term hot border to denote an entity type that is unique for each person, such as id number. While scanning the left/right context beginning from the current entity, if a hot border entity is encountered, the module does not continue further in that direction.

5. Experiments and results

5.1. Experimental setup

(A) Data set

The data set used in this study [4] is a collection of sentences from a Turkish newspaper Milliyet, covering a period of almost two years. The data set consists of Turkish sentences labeled with the IOB scheme [26]. As an example, "Saffet Sancaklı ile anlaşan Konyaspor, Erkan için de kolları sıvadı" (*Konyaspor signed with Saffet Sancaklı and rolled up its sleeves for Erkan*) is labeled as follows:

Saffet	Sancaklı	ile	anlaşan	Konyaspor	,	Erkan	için	de	kolları	sıvadı
B-PER	I-PER	O	O	I-ORG	O	I-PER	O	O	O	O

The data set was split into a training set of 29,915 sentences, 15 % of which is used as the development set, and a test set of 2668 sentences.

(B) Word representation

In this study, words are represented with 300-dimensional fasttext embeddings trained with the continuous bag of words (CBOW) model [17] with position weights using the Common Crawl and Wikipedia dumps. Embedding vectors for 2M unique Turkish words were obtained from the official fasttext release [5]. However, we realized that the embeddings need to be subjected to both noise removal and size reduction before using in production. In terms of storage size (larger than 6 gigabytes), keeping word embeddings on disk or RAM would be highly expensive, hence we tried to minimize its size with minimum loss of accuracy. Reducing word embedding vector dimension below 300 was not accepted due to high loss in accuracy observed empirically. To detect noise, the frequency distribution of words in the fasttext corpus was analyzed. We observed that the total frequency of the 100K most frequently used words was over 99 % of all the words in the corpus. In other words, 1 % of the words in the corpus have relatively lower frequency when compared with the rest. So, removing embeddings of such rarely used words corresponds to 20 times cost reduction with 1 % loss of accuracy. Based on this trade-off

[5] Fasttext Turkish word embeddings [online]. Website https://dl.fbaipublicfiles.com/fasttext/vectors-crawl/cc.tr.300.bin.gz [accessed 02 February 2021]

between cost and accuracy, it was decided to preserve only 100K word embeddings (with almost 130 MB storage size). To denoise the word embeddings, words longer than 50 characters and including more than 2 digits or punctuation marks were removed.

(C) The NER model

NER is considered as a labeling task over a given sentence that consists of a sequence of words [21], thus the NER model was designed to work on a sentence basis. We analysed the data set and concluded that almost 99 % of the sentences in the data set have less than 80 words. Sentences having more than 80 words which correspond to outliers were removed as they pose difficulty in the learning process.

5.2. Controls and quantization

The maximum probability in the softmax layer for a given word was used to decide whether to tag as a named entity or not. If the probability of a named entity label (among three possible labels) is at least 80 %, it is accepted as the tag of the word. If the probability is between 30 % and 80 %, additional dictionary-based controls over sets of almost 10K Turkish human names, 2K location names, and 3K organisation names are employed to avoid tagging the word erroneously. In the case that the probability is below 30 %, the word is labeled as O, indicating that it is not part of a named entity. In this way, the robustness of the NER tagging process used in this study was increased using a combination of machine learning-based and dictionary-based approaches. In addition, the NER model was quantized by representing the learned weights as integers instead of expensive float numbers using linear normalization with minimum accuracy loss as described in a pioneering paper [27]. With this quantization, the NER model size was decreased from 6.5 MB to 0.83 MB that can be regarded as a plausible storage size for deploying into production.

5.3. Results

Several Bi-LSTM models having different hidden vector dimensions and unknown word strategies were trained using the root mean square propagation (RMSProp) optimizer [28] with an initial learning rate of 1e-02 and epsilon of 1e-07. Each model was set to train for 1000 epochs with early stopping criterion of 10 consecutive increments of loss on the evaluation data set. We designed six different setups, each of which differs in hidden vector dimension and word embedding search approach. The results are shown in Table 3. We observe that Bi-LSTM with 100 dimensional hidden vector size gives the best results using the progressive word search method. It is seen that increasing the dimension of the Bi-LSTM model does not improve the accuracy. However, accuracy is increased regardless of the dimension when the progressive word search approach is employed. After observing the performance of the NER approach, two more experiments were designed to evaluate the performance of the overall system in two different aspects.

The first one aims to test the effect of the file extension and size on the execution time of personal profile extraction, whereas the second one aims to check the accuracy of both labelling the entity types and the personal profiles extracted.

For each experiment, a custom test set was created. For the first experiment, a test set including almost 33K files was created. The numbers of the files used in this test set by extension and size are shown in Table 4. For the second experiment, a relatively smaller test set was created and annotated to be used as the gold standard data set. In these test sets, we employed text recognition operation for images with a publicly available optical character recognition engine [6] to obtain textual contents. Table 5 shows the number of files for each

[6] Tesseract [online]. Website https://github.com/tesseract-ocr/tesseract [accessed 02 May 2021]

Algorithm 1 Entity profiling

```
$H \leftarrow [id\_number]$                                             ▷ Hot borders
procedure AssignProfiles($D$)                  ▷ Document, $D = [e_1, e_2, \ldots, e_n]$
    $c \leftarrow 0$                                                    ▷ Cursor
    $P \leftarrow \emptyset$                                            ▷ Profiles
    for $i \leftarrow 0;\ i < len(D);\ i++$ do            ▷ Scan entities in the document
        if $i < c$ then        ▷ Move cursor to the position of the last processed entity
            continue
        end if
        if $D[i].color$ in $[red, green]$ then   ▷ Profile starting point: red and green
            for $j \leftarrow 0;\ j < len(P);\ j++$ do         ▷ Check existing profiles
                if $D[i].content$ in $P[j]$ then        ▷ If entity exists in the profile
                    $c, P \leftarrow$ ContextSearch($D, i, P, j$)  ▷ Check left/right context
                    continue
                end if
                $P[j].insert(D[i])$                              ▷ Create a new profile
                $c, P \leftarrow$ ContextSearch($D, i, P, j$)      ▷ Check existing profiles
            end for
        end if
    end for
    return $P$                                           ▷ Return profiles extracted
end procedure

procedure ContextSearch($D,\ i,\ P,\ j$)
    $left, right \leftarrow margins(D_i.type)$         ▷ Window size for the entity type
    $RE \leftarrow D[i - left : i - 1], D[i + 1 : i + right]$   ▷ Entities in the window
    for $k \leftarrow 0;\ k < len(RE);\ k++$ do         ▷ Scan the entities in the window
        if $RE[k].type$ not in $H$ and $RE[k].content$ not in $P[j]$ then   ▷ Check for hot border
            if $RE[k].content$ not in $P[j]$ then
                $P[j].insert(RE[k])$                          ▷ Add entity to the profile
            end if
        else
            return $max(i, i - left + k), P$   ▷ Break loop when hot border entity is encountered
        end if
    end for
    return $i + right, P$
end procedure
```

extension and the number of personal data entities in those files. The personal data entities in the second test set were annotated using an XML schema[7] that is specific to this study to indicate both personal data entities and personal profiles. Personal data entity types and their frequencies in this test set are shown in Table 6.

Figure 4 shows the results of the first experiment. The figure is divided into three parts concerning the sizes of the files analyzed, to be able to observe the differences between the execution time and the profile numbers. In the second experiment, we measured the performance of personal data extraction and profiling. For personal data extraction, we obtained micro and macro-averaged F1-measures of 91.76 % and 89.30 %, respectively.

[7] <annotation><index>entityType</index><profile>profileNumber</profile></annotation>, where *index* stores a string that indicates the entity type and *profile* stores an integer that indicates the profile number of the related personal data entity in the file.

Table 3. Performance scores on test set. The word search column shows the effect of the proposed progressive word search method (PS) compared to the method where the words that do not exist in the embedding lexicon are accepted as unknown (NS – no search).

Model / Dimension	Word Search	Label	Precision	Recall	F1-Measure	Overall
Bi-LSTM / 100	PS	Location	94.28	95.47	94.87	**94.05**
		Organisation	92.96	91.84	92.40	
		Person	94.35	95.39	94.87	
	NS	Location	92.16	90.65	91.40	91.10
		Organisation	86.78	89.12	87.93	
		Person	95.32	92.59	93.94	
Bi-LSTM / 200	PS	Location	96.25	93.77	94.99	93.66
		Organisation	89.14	92.28	90.68	
		Person	95.54	95.01	95.27	
	NS	Location	90.11	89.75	89.93	91.04
		Organisation	90.66	90.13	90.39	
		Person	92.46	93.11	92.78	
Bi-LSTM / 300	PS	Location	95.14	93.55	94.34	93.61
		Organisation	92.78	89.95	91.34	
		Person	95.48	94.76	95.12	
	NS	Location	92.09	90.57	91.32	91.17
		Organisation	89.78	89.15	89.46	
		Person	92.89	92.56	92.72	

Table 4. The number of files by extension in the test set that is used in the first experiment. Since the files were collected from an office worker's computer, there are few files for underused extensions.

Extension	**jpeg**	**png**	**jpg**	**pdf**	**docx**	**doc**	**xlsx**	**pptx**	**bmp**	**ppt**	**xls**
0 –200 KB	12601	1139	2502	1144	223	153	122	10	4	7	3
200 KB – 1 MB	94	2103	2943	1192	15	10	7	6	5	3	1
Over 1 MB	27	5492	1444	1058	7	1	7	28	4	1	1

Table 5. The number of files and personal data entities by extension in the test set that is used in the second experiment.

Extension	**pdf**	**docx**	**doc**	**xls**	**xlsx**	**csv**	**ppt**	**pptx**
number of files	10	10	10	10	10	10	5	5
number of entities	87	104	77	335	340	330	17	17

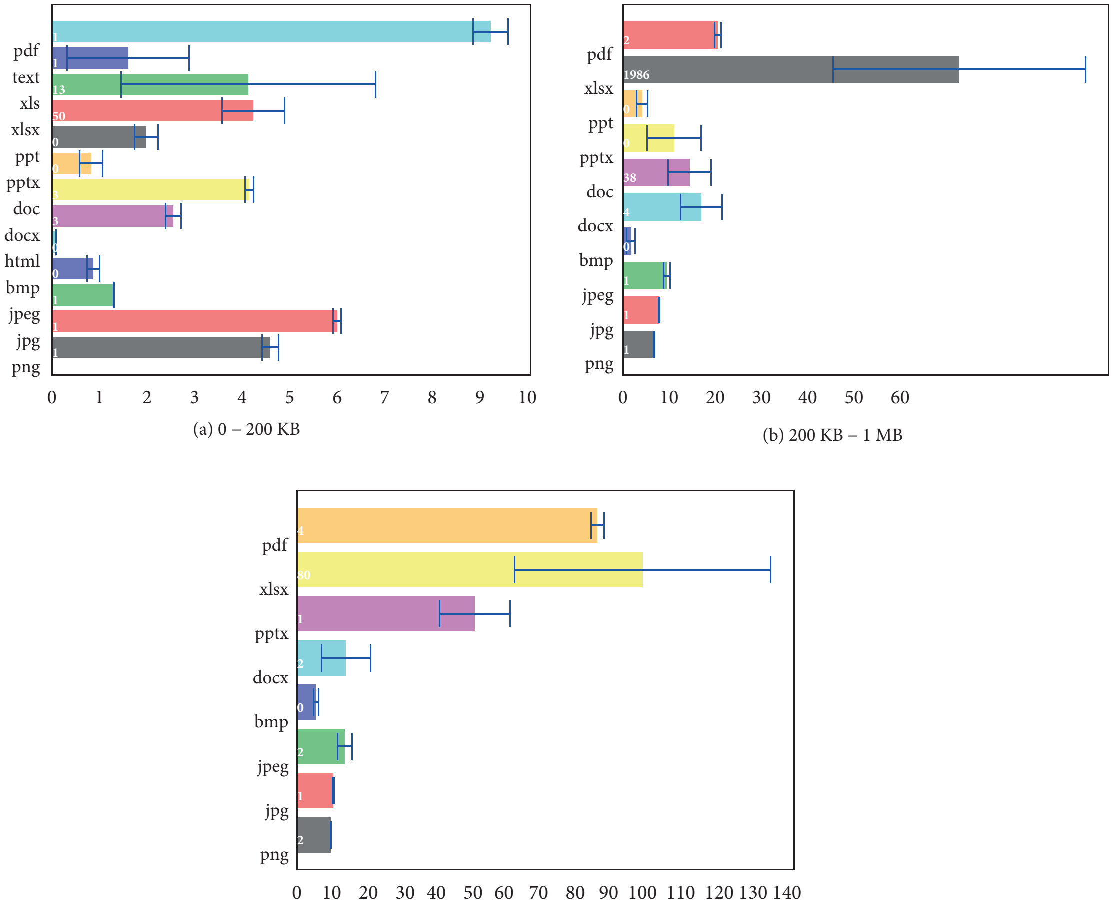
In the case of profiling, the percentage of the number of personal data entities assigned to the correct profiles was measured. We obtained an accuracy of 72.46 % in this test.

6. Discussion

For the last 10 years, NER studies have been carried out with machine learning methods and different types of learning approaches have been used. Since 2015 deep recurrent neural networks have begun to be utilized, which can model the temporal relationships in a sequence.

Table 6. The distribution of the entities by their types in the test set that is used in the second experiment. Freq stands for frequency.

Entity type	**Freq**	**Entity type**	**Freq**	**Entity type**	**Freq**
phone number mobile	107	date of birth	1	location name	72
organisation mall	75	organisation restaurant	57	organisation name	37
human name	30	tr vehicle plate no	42	e-mail	147
bank name	24	tr id number	57	organisation institution	8
human name surname	150	organisation municipality	6	university name	18
phone number home	48	tr tax no	24	iban no	45
organisation military	16	disease	21	blood group	4
insurance company	21	drug	42	other	67
date	42	organisation travel	4		



(a) 0 – 200 KB

(b) 200 KB – 1 MB

(c) over 1 MB

Figure 4. Test results on ETC and entity profiling. x and y axes represent duration (s) and file extensions, respectively. Length of the bars on the charts show the average time with the error margin for each file extension. Values on the bars indicate the number of personal profiles created.

The rising popularity of attention-based networks in NLP recently directed the researchers to focus on fine-tuning pretrained general-purpose transformer models such as bidirectional encoder representations from transformers (BERT) [29]. However, the depth of the BERT model not only leads to high-memory consumption but also high latency at inference. Considering these constraints in our study, we have experimented with several Bi-LSTM models and decided to deploy the one both having the least number of parameters and reaching the optimal accuracy.

Two test scenarios were conducted and promising results were achieved. The result of the first test experiment showed that the proposed model can process the files in an acceptable time, employ ETC, and extract the profiles successfully. The results of the second test experiment indicate that the proposed model can perform ETC and personal data profiling at a high accuracy level. ETC was performed with a micro-averaged F1-measure of 91.76 %. The reason behind macro-averaged F1-measure being slightly lower than micro-averaged F1-measure is that the precision of human name entity detection is affected by the false-positive decisions for human name surname. This situation shows that Bi-LSTM performs at a high level, however, the system should be improved in making decisions whether to assign consecutive human name entity pairs as human name surname entity or not. In the second experiment, the accuracy of 72.46 % in extracting personal profiles indicates that the developed system can provide a group of personal data in response to a person who requests to be informed whether his/her personal data is processed or not under the rights specified in Article 11 of the PDPR.

This study can be evaluated as a fine-grained entity extraction model that extracts 77 types of entities, three of which rely on machine learning methods and the others make use of rules and heuristics. It differs from the other fine-grained NER studies [1–3] in that it only focuses on personal data. In this regard, this study is the first attempt to construct a personal data extraction system with a hybrid approach.

7. Conclusion

In this paper, we proposed a hybrid model to facilitate and accelerate PDPR compliance in Turkey. We proposed an automated data mapping and discovery system for Turkish personal data extraction and profiling with the help of a Bi-LSTM based NER model and a rule sets approach. Our best result in the NER model is slightly higher than the studies that use the same data set [7–11, 13], whose success rates were given in Section 3.2. The progressive word search method proposed in this study is considered to have a positive effect on these results.

The performance measures obtained in this work show that the proposed approach and the system developed can enable the data controllers to gain insights about what and how personal data is currently being processed. The research in this study has been further enhanced with the deployed Turkish document-type detection [30] and topic detection [31] models. By the beginning of 2021, this study has been productized under the name Datamin and commercialized in the market[8].

As future work, due to the effect of morphological embeddings in agglutinative languages, we plan to incorporate those embeddings into the model and compare them with the approach proposed in this paper.

Acknowledgments

This work was supported by the Scientific and Technological Research Council of Turkey (TÜBİTAK) under Grant No. 3190083 with the cooperation of Mobildev and Boğaziçi University. We would like to thank Filiz Aksoy for her contribution to the management of this project.

[8] Datamin [online]. Website https://www.mobildev.com/datamin [accessed 02 May 2021]

References

[1] Ling X,Weld D. Fine-grained entity recognition. In *Proceedings of the Association for the Advancement of Artificial Intelligence Conference on Artificial Intelligence*, volume 26, 2012.

[2] Yosef MA, Bauer S, Hoffart J, Spaniol M, Weikum G. Hyena: Hierarchical type classification for entity names. In *Proceedings of Conference on Computational Linguistics*, pages 1361–1370, 2012.

[3] Kalender M, Korkmaz E. Turkish entity discovery with word embeddings. *Turkish Journal of Electrical Engineering & Computer Sciences*, 25 (3):2388–2398, 2017.

[4] Tür G. *A statistical information extraction system for Turkish.* PhD thesis, Bilkent University, 2000.

[5] Bayraktar Ö , Temizel TT. Person name extraction from Turkish financial news text using local grammar-based approach. In *23rd International Symposium on Computer and Information Sciences*, pages 1–4, 2008.

[6] Küçük D,Yazıcı A. Rule-based named entity recognition from turkish texts. In *Proceedings of the International Symposium on Innovations in Intelligent Systems and Applications*, pages 456–460, 2009.

[7] Tatar S,Çiçekli I. Automatic rule learning exploiting morphological features for named entity recognition in turkish. *Journal of Information Science*, 37 (2):137–151, 2011.

[8] GA Şeker,Eryiğit G. Initial explorations on using crfs for turkish named entity recognition. In *Proceedings of Conference on Computational Linguistics*, pages 2459–2474, 2012.

[9] Demir H, Özgür A. Improving named entity recognition for morphologically rich languages using word embeddings. In *13th International Conference on Machine Learning and Applications*, pages 117–122, 2014.

[10] Şeker GA, Eryiğit G. Extending a crf-based named entity recognition model for turkish well formed text and user generated content 1. *Semantic Web*, 2017;8 (5):625–642.

[11] Güngör O, Üsküdarlı S,Güngör T. Recurrent neural networks for turkish named entity recognition. In *26th Signal Processing and Communications Applications Conference*, pages 1–4, 2018.

[12] Akkaya EK. Deep neural networks for named entity recognition on social media. Master's thesis, Hacettepe University, Institute of Natural Sciences, 2018.

[13] Güngör O, Güngör T, Üsküdarli S. The effect of morphology in named entity recognition with sequence tagging. *Natural Language Engineering*, 2019;25 (1):147–169.

[14] Yamada I, Asai A, Shindo H, Takeda H, Matsumoto Y. Luke: Deep contextualized entity representations with entity-aware self-attention. *arXiv preprint*, 2020. arXiv: 2010.01057.

[15] Luoma J,Pyysalo S. Exploring cross-sentence contexts for named entity recognition with bert. *arXiv preprint*, 2020. arXiv:2006.01563.

[16] Yu J,Bohnet B,Poesio M. Named entity recognition as dependency parsing. *arXiv preprint*, 2020. arXiv:2005.07150.

[17] Bojanowski P, Grave E, Joulin A, Mikolov T. Enriching word vectors with subword information. *arXiv preprint*, 2016. arXiv:1607.04606.

[18] Olby L, Thomander I. A step toward gdpr compliance: Processing of personal data in email. 2018.

[19] Dasgupta R, Ganesan B, Kannan A, Reinwald B, Kumar A. Fine grained classification of personal data entities. *arXiv preprint*, 2018. arXiv:1811.09368.

[20] Dias M, Boné J, Ferreira J, Ribeiro R,Maia R. Named entity recognition for sensitive data discovery in portuguese. *Applied Sciences*, 2020;10 (7):2303.

[21] Jurafsky D. *Speech & Language Processing.* Pearson Education India, 2000.

[22] Srivastava N,Hinton G, Krizhevsky A,Sutskever I,Salakhutdinov R. Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research* 2014;15 (1):1929–1958.

[23] Luhn HP. Computer for verifying numbers, 1960.

[24] Postel J. Dod standard internet protocol. *Association for Computing Machinery Special Interest Group on Data Communications Computer Communication Review* 1980;10 (4):12–51.

[25] Deering S et al. Internet protocol, version 6 (ipv6) specification, 1998.

[26] Ramshaw L,Marcus M. Text chunking using transformation-based learning. In *Third Workshop on Very Large Corpora*, 1995.

[27] Jacob B, Kligys S,Chen B, Zhu M, Tang M et al. Quantization and training of neural networks for efficient integer-arithmetic-only inference. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2704–2713, 2018.

[28] Tieleman T,Hinton G. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *Coursera: Neural Networks For Machine Learning*2012; 4 (2):26–31.

[29] Devlin J, Chang MW,Lee K,Toutanova K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint*, 2018. arXiv:1810.04805.

[30] Bakar B, Aksoy F, Yayık A, İçöz S,Aybar V et al. Turkish rule-based official document type detection. In *28th Signal Processing and Communications Applications Conference*, pages 1–4, 2020.

[31] Yayık A,Apik H, Tosun A, Ozdemir E. Deep learning based topic classification for sensitivity assignment to personal data. Technical report, Partnership for Advanced Computing in Europe (EU Union Horizon 2020 Research and Innovation Program), 2021.