UBMK'21

Deep Learning Based Topic Classification for Sensitivity Assignment to Personal Data

- Dr Apdullah Yayık*, Huawei Turkey R&D Center
- Hasan Apik, Mobildev
- Dr Ayşe Tosun, Istanbul Technical University

* former Mobildev employee

Introduction

- Article 6 in PDPR defines a **special category** of personal data that reveal the tendency of a **religion**, a belief, a sexual choice, the state of **health**, or **convictions** for any **crime**.
- By this definition, topic of a textual content directly affects **severity of** personal data in its context.
- We aim to model non-linear relations in Turkish textual contents through an expanded dataset to predict their topic.
- The model is deployed inside our existing personal data exploration tool, Datamin.

Datamin

- AI-driven solution to provide PDPR compliance.
- Detection of 77 number personal entity types through regex, look-up tables and Bi-LSTM-based named entity recognition on over 30 file extensions.
- Person-based grouping that intends to associate personal data entities with their belongings.

finance
- IBAN no
- tax no
- bank credit card number
- bank credit number
- bank customer number
- bank account number
- policy number
- bank name
- insurance company
- central registry system no
- credit number
  - bank name
- customer number
  - bank name
- policy number
  - insurance company
- bank account number
  - bank branch name
  - bank branch code
  - bank name

health
- blood group
- drug
- disease
- organization health

contact
- phone number home
- phone number mobile
- e-mail
- IMEI number
- tel number (fax)
- address
- address no

location
- location

cyber operation security
- ip address (v4)
- ip address (v6)

profession
- university name
- primary school
- high school
- diploma number
- graduated program
- education type
- class number
- studying program

id
- id number
- human name surname
- human name
- surname
- father name
- mother name
- date of birth
- place of birth
- registered province
- registered distinct
- sub-distinct
- driving license
  - document no
  - paper no
  - register no
  - class
- passport no
- mother maiden name
- volume no
- household no
- id card serial no
- gender
- age

religious
- religious tendency

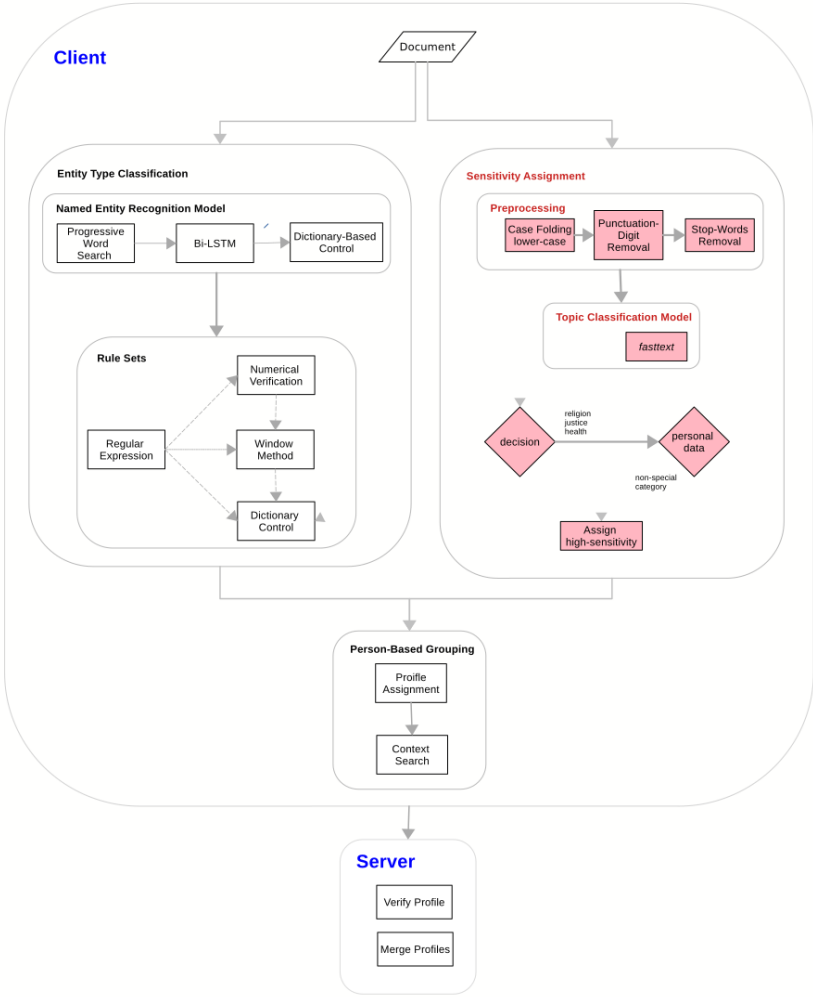
other
- vehicle plate no
- organisation
  - municipality
  - institution
  - industry
  - service
- military status
  - exemption
  - postpone
  - demobilization
- marriage date
- divorce date

marketing
- organisation
  - accommodation
  - restaurant
  - travel
  - mall

criminal offence
- record
- achieve record

Extending Datamin

- Sensitivity assignment module, developed in this study, assigns a high-sensitivity to the personal data entities only supposed they appear in the textual context whose topic is religion, justice, or health.



Dataset and Cleaning Operations

- The dataset released by Yıldırım et al. was expended by adding instances of health, religion, and justice.
- 31K instances each of which are formed by paragraphs or sentences, and contains a total of 6M words 30K of which are unique.

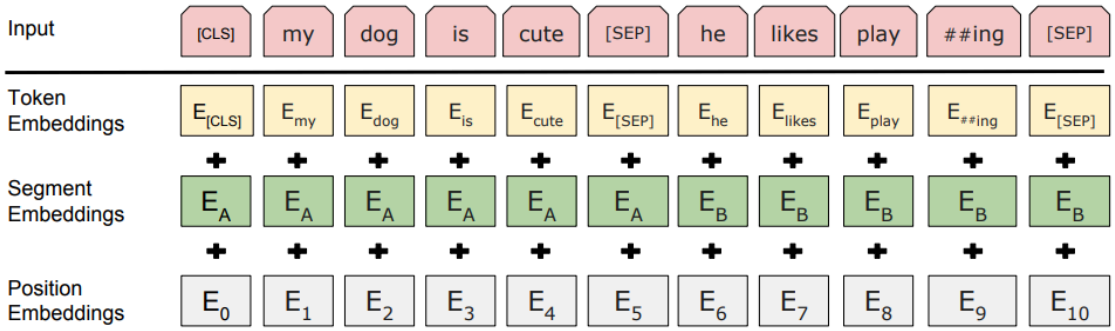
Class	number of instances	words		
		number of total	number of unique	mean ± ste
Politics	2549	652K	2522	255.67±4.06
Economy	3962	912K	3704	230.20±2.83
Culture	1852	378K	1551	204.10±4.15
Health	3078	540K	2892	175.20±2.82
Sport	10679	1.6M	10149	157.16±1.30
Technology	1471	246K	1339	167.00±3.24
Religion	4914	68K	4645	13.80±0.12
Justice	2547	50K	2417	20.00±0.33

Models, Fasttext

- Trains a logistic regression with minimizing negative likelihood through the generic stochastic gradient descent algorithm (lr=0.7)
- Represents the inputs as the average of the n-gram word vectors (100 dimension)
- Has hierarchical softmax layers.

Models, BERT and BERTurk

- Trained by concurrently optimizing 2 loss functions for the following tasks:
  - predict masked language model
  - predict consecutive sentences in the given textual content.
- Represents words
  - Segmentation, position and token embeddings
- Has 12 transformer blocks in 12 attention heads, and 768 hidden sizes.

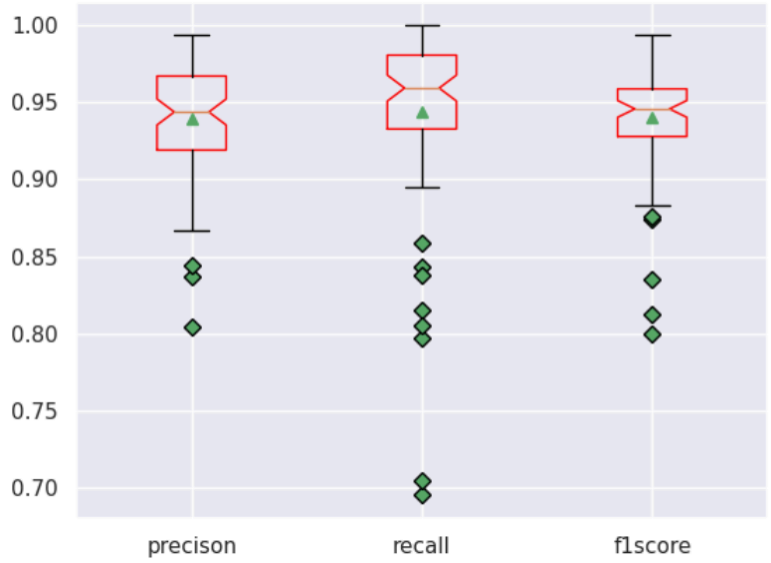


Evaluation - 1

- 10% test (left), 20% test (right)

Model	Precision	Recall	F1 Score
fasttext	94.33 ± 0.99	95.20 ± 0.88	94.73 ± 0.67
BERT-based	93.25 ± 1.73	93.59 ± 1.96	93.29 ± 1.37
BERTurk-based	97.09 ± 0.96	97.61 ± 0.75	97.33 ± 0.77

Model	Precision	Recall	F1 Measure
fasttext	93.47 ± 1.50	95.57 ± 2.26	94.37 ± 1.40
BERT-based	93.85 ± 1.82	95.02 ± 1.27	94.31 ± 1.00
BERTurk-based	97.27 ± 1.29	98.56 ± 0.28	97.88 ± 0.68



- median and average values of the performance measures are around 95%
- a recall rate between 70% and 85%
- a precision and F1-measure rate between 80% and 87%

Evaluation - 2

- Performance measures of the models are closer.
- Nemenyi and Mann-Whitnet post-hoc testa are employed on model predictions
- BERT and fasttext models acts similarly at identifying the topic of a document.

Ratio	Post-Hoc	Model	*fasttext*	Bb	BTb
10 %	Nemenyi	*fasttext*	1		
		Bb	0.900	1	
		BTb	0.286	0.381	1
	Mann-Whitney	*fasttext*	1		
		Bb	0.606	1	
		BTb	0.046	0.1365	1
20 %	Nemenyi	*fasttext*	1		
		Bb	0.517	1	
		BTb	0.411	0.900	1
	Mann-Whitney	*fasttext*	1		
		Bb	0.170	1	
		BTb	0.126	0.859	1

Train and Inference Resources

- Fasttext model
  - has almost 10X less memory-usage,
  - can be trained in almost 250X faster,
  - can make inference in almost 350X faster

		fasttext	BERT-based	BERTurk-based
Model Train	**duration**	1 minute 35 sec	7 hour 45 minutes 56 sec	6 hours 15 minutes 32 sec
	memory usage	1 GB	14 GB	9 GB
	hardware	GPU Nvidia Tesla V4, 256 GB RAM		
Model Test Execution	**duration of an instance**	1.6 ms	569 ms	615 ms
	hardware	CPU 8 Core, 16 GB RAM		
	model size	6.4 MB	438 MB	442 MB
	memory usage	12 MB	987 MB	1.1 GB
	CPU level	1	46	52

Discussion and Future Work

- Fasttext model is integrated to Datamin product since low resource usage in every steps of its ML life-cycle.

References

[1] M. Amasyalı and T. Yıldırım, “Otomatik haber metinleri sınıflandırma,” *SIU 2004*, pp. 224–226, 2004.

[2] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” *arXiv preprint arXiv:1810.04805*, 2018.

[3] A. Joulin, E. Grave, P. Bojanowski, and T. Mikolov, “Bag of tricks for efficient text classification,” *arXiv preprint arXiv:1607.01759*, 2016.