PRACE SHAPE MOBILDEV – DATAMIN PROJESI

Hazirlayan : Filiz Aksoy

MOBILDEV

- 18 yillik deneyim
- 250 is ortagi
- 10 servis saglayici
- Digital pazarlama cozumleri
- Urunler
  - IVT (Izinli veri tabani),
  - Toplu SMS (Bulk SMS),
  - Toplu Email,
  - Anket Yonetimi vs
- Ar-ge : Datamin (TÜBİTAK- AGY100 - 03)

DATAMIN

- Kisilsel verilerin kategorize edilmesi
  - standart formatli dokumanlar

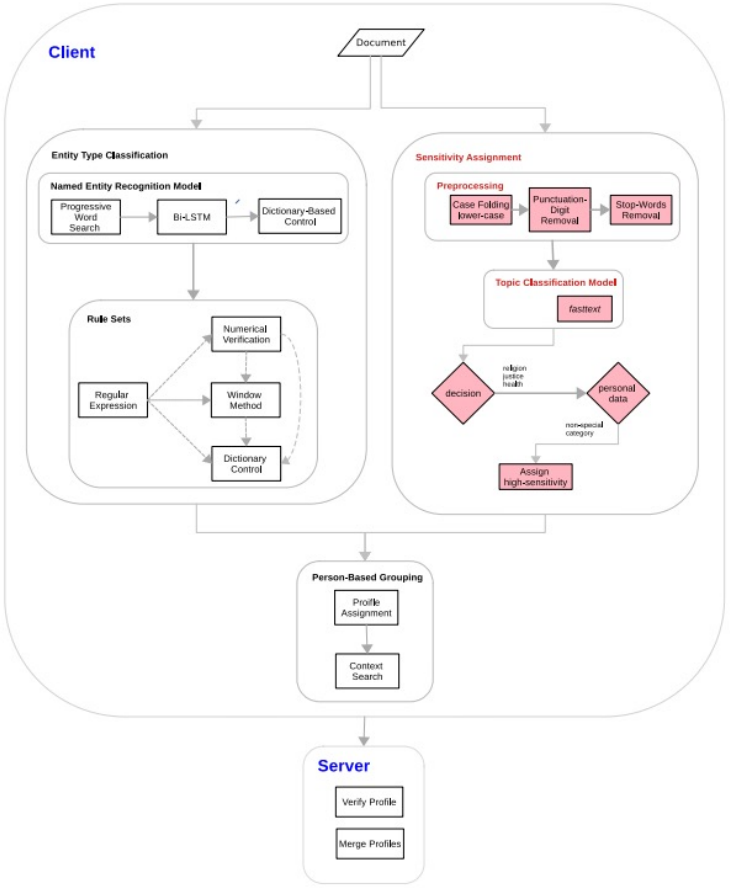
    * Kimlik, ehliyet, pasaport, askerlik belgesi, tapu, sicil kaydi....
  - Standart olmayan dokumanlar

    * text, excel, pdf, word vs...

YONTEMLER

- NER (Named Entity Recognition)
- LSTM (Long-Short Term Memory)
- Reg-ex (Regular expression)
- CNN
- Look-up table
- Dictionary
- Fasttext vektor uzayi & Bidirectional Encoder Representations Transformers (BERT)
- OCR - OpenCv

GENEL MIMARI YAPI



- Sistemdeki her düğüm, NVLink aracılığıyla bağlanan dört NVIDIA Tesla V100 Genel Amaçlı Grafik İşlem Birimi (GPGPU) içeriyor.
- Her GPU'nun 32 GB belleği ve her düğümün toplam 384 GB sistem belleği var.
- Düğümler arasındaki ara bağlantı ağı, Gelişmiş Veri Hızı InfiniBand aracılığıyla sağlanır. Donanım platformu ağırlıklı olarak fasttext ve BERT tabanlı modellerin eğitim aşaması için kullanılmıştır.

DATASET KATEGORI

Class	Number of instances
Politics	2549
Economy	3962
Culture	1852
Health	3078
Sport	10679
Technology	1471
Religion	4914
Justice	2574

MODEL TRAINING & MODEL TEST EXECUTION

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

RESULT

Class	Model	Precision (%)	Recall (%)	F-1 Measure (%)
	fasttext	91.88	99.32	95.45
Technology	BERT-based	92.16	95.27	93.69
	fasttext	94.87	100.00	97.36
Culture	BERT-based	94.36	99.46	96.84
	fasttext	87.85	87.85	87.85
Justice	BERT-based	96.58	91.50	93.97
	fasttext	94.69	98.74	96.67
Economy	BERT-based	95.06	92.19	93.61
	fasttext	87.14	99.03	92.71
Health	BERT-based	83.01	96.75	89.36
	fasttext	96.45	83.10	89.28
Religion	BERT-based	97.99	89.41	93.50
	fasttext	99.43	97.28	98.34
Sport	BERT-based	99.81	98.69	99.25
	fasttext	95.47	99.22	97.31
Politics	BERT-based	91.82	96.86	94.27
	fasttext	93.47 ± 1.50	95.57 ± 2.26	94.37 ± 1.40
Macro-avg $\pm$ ste	BERT-based	93.85 ± 1.82	95.02 ± 1.27	94.31 ± 1.00