Delivery 3 - DL-based Text Classification

CSE4095 - Special Topics in Comp. Eng. III

Ali Reza Ibrahimzada
Muhammed Fatih Öztel
Basil Osama Hamid Elkhalifa
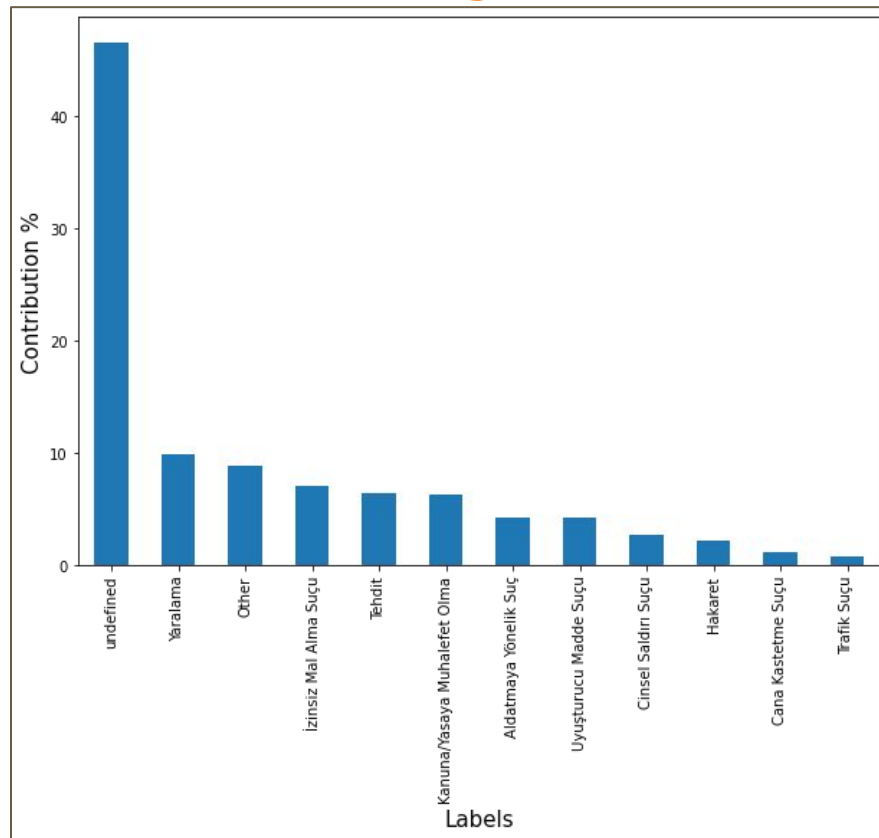
Huzeyfe Ayaz
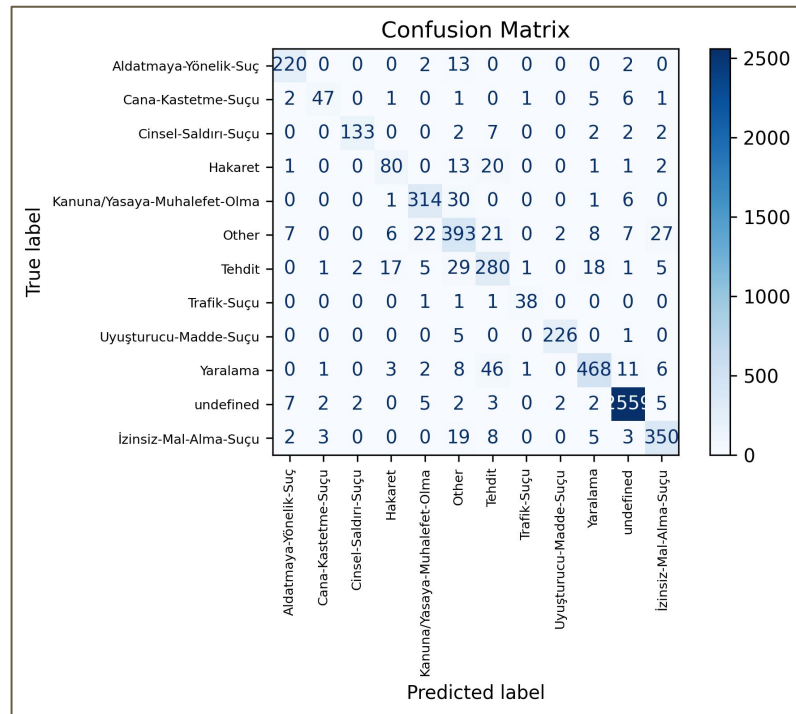Yunus Ahmed Stahlschmidt
Sameeh Kunbargi

Content

Data Cleaning and Preprocessing

- Cleaned Data from Delivery 1

- Less Frequent Labels in "Other"

- Empty Labels in "Undefined"

- Training (80%) and Test (20%) in Stratified Manner

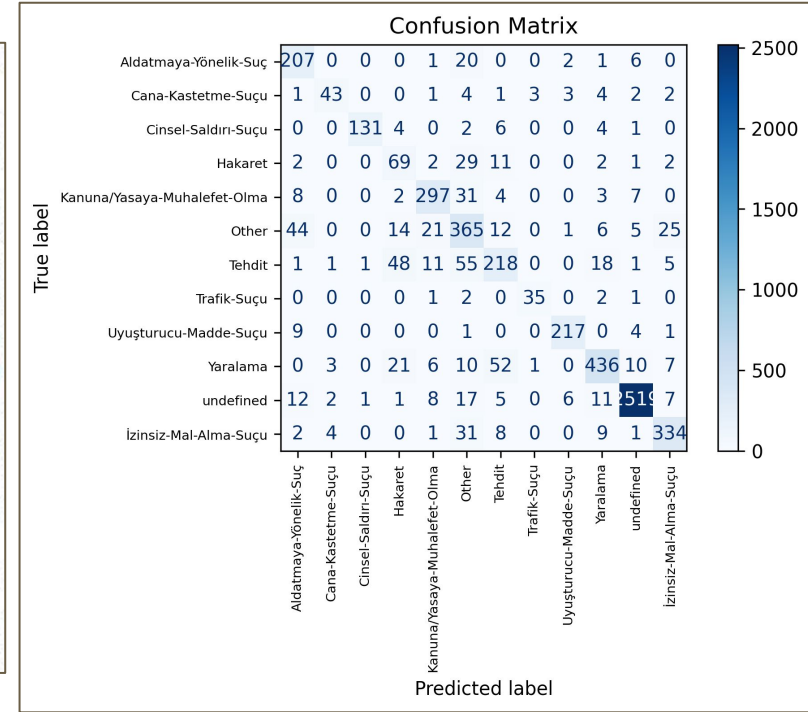- Same split as in Delivery 2 - fair comparison

FastText

	precision	recall	f1-score	support
Aldatmaya-Yönelik-Suç	0.92	0.93	0.92	237
Cana-Kastetme-Suçu	0.87	0.73	0.80	64
Cinsel-Saldırı-Suçu	0.97	0.90	0.93	148
Hakaret	0.74	0.68	0.71	118
Kanuna/Yasaya-Muhalefet-Olma	0.89	0.89	0.89	352
Other	0.76	0.80	0.78	493
Tehdit	0.73	0.78	0.75	359
Trafik-Suçu	0.93	0.93	0.93	41
Uyuşturucu-Madde-Suçu	0.98	0.97	0.98	232
Yaralama	0.92	0.86	0.89	546
undefined	0.98	0.99	0.99	2589
İzinsiz-Mal-Alma-Suçu	0.88	0.90	0.89	390
accuracy			0.92	5569
macro avg	0.88	0.86	0.87	5569
weighted avg	0.92	0.92	0.92	5569

Confusion Matrix

LSTM

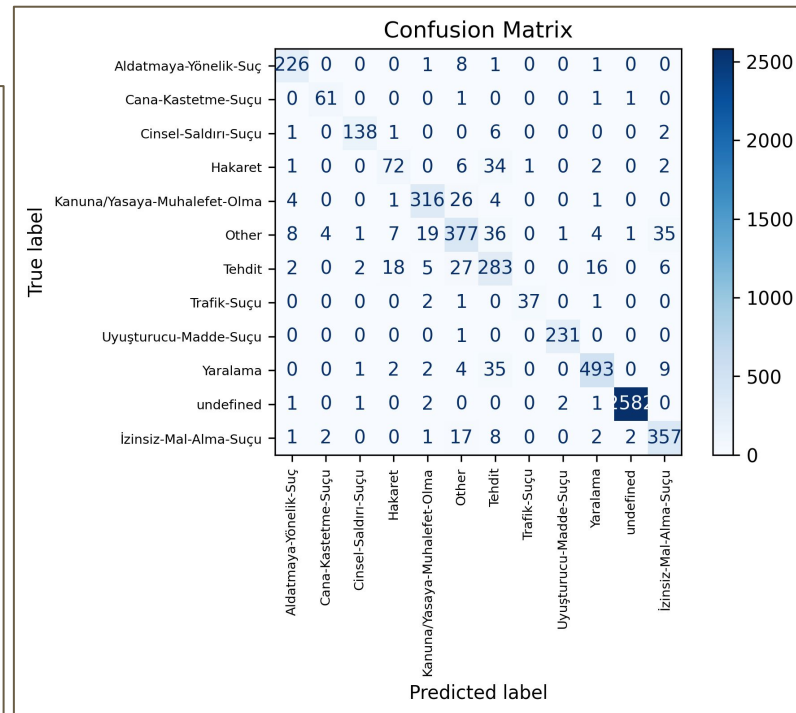|	precision	recall	f1-score	support
Aldatmaya-Yönelik-Suç	0.72	0.87	0.79	237
Cana-Kastetme-Suçu	0.81	0.67	0.74	64
Cinsel-Saldırı-Suçu	0.98	0.89	0.93	148
Hakaret	0.43	0.58	0.50	118
Kanuna/Yasaya-Muhalefet-Olma	0.85	0.84	0.85	352
Other	0.64	0.74	0.69	493
Tehdit	0.69	0.61	0.64	359
Trafik-Suçu	0.90	0.85	0.88	41
Uyuşturucu-Madde-Suçu	0.95	0.94	0.94	232
Yaralama	0.88	0.80	0.84	546
undefined	0.98	0.97	0.98	2589
İzinsiz-Mal-Alma-Suçu	0.87	0.86	0.86	390
accuracy			0.87	5569
macro avg	0.81	0.80	0.80	5569
weighted avg	0.88	0.87	0.88	5569

Confusion Matrix

WHAT IF I TOLD YOU
ATTENTION IS ALL YOU NEED
imgflip.com

Transformer-based Model

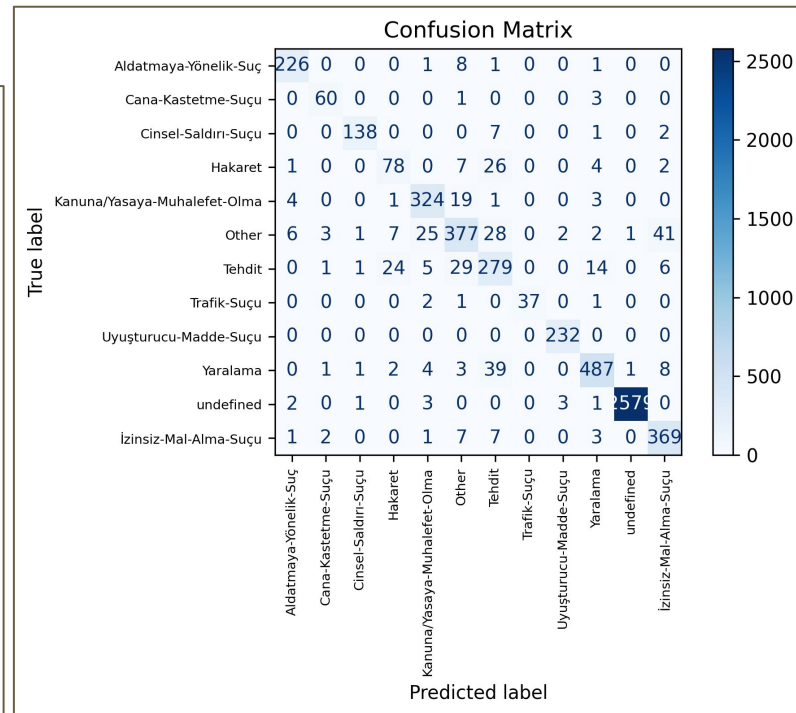savasy/bert-base-turkish-sentiment-cased from 🤗 - 10 epochs

	precision	recall	f1-score	support
Aldatmaya-Yönelik-Suç	0.93	0.95	0.94	237
Cana-Kastetme-Suçu	0.91	0.95	0.93	64
Cinsel-Saldırı-Suçu	0.97	0.93	0.95	148
Hakaret	0.71	0.61	0.66	118
Kanuna/Yasaya-Muhalefet-Olma	0.91	0.90	0.90	352
Other	0.81	0.76	0.78	493
Tehdit	0.70	0.79	0.74	359
Trafik-Suçu	0.97	0.90	0.94	41
Uyuşturucu-Madde-Suçu	0.99	1.00	0.99	232
Yaralama	0.94	0.90	0.92	546
undefined	1.00	1.00	1.00	2589
İzinsiz-Mal-Alma-Suçu	0.87	0.92	0.89	390
accuracy			0.93	5569
macro avg	0.89	0.88	0.89	5569
weighted avg	0.93	0.93	0.93	5569

Confusion Matrix

True label \ Predicted	Aldatmaya-Yönelik-Suç	Cana-Kastetme-Suçu	Cinsel-Saldırı-Suçu	Hakaret	Kanuna/Yasaya-Muhalefet-Olma	Other	Tehdit	Trafik-Suçu	Uyuşturucu-Madde-Suçu	Yaralama	undefined	İzinsiz-Mal-Alma-Suçu
Aldatmaya-Yönelik-Suç	226	0	0	0	1	8	1	0	0	1	0	0
Cana-Kastetme-Suçu	0	61	0	0	0	1	0	0	0	1	1	0
Cinsel-Saldırı-Suçu	1	0	138	1	0	0	6	0	0	0	0	2
Hakaret	1	0	0	72	0	6	34	1	0	2	0	2
Kanuna/Yasaya-Muhalefet-Olma	4	0	0	1	316	26	4	0	0	1	0	0
Other	8	4	1	7	19	377	36	0	1	4	1	35
Tehdit	2	0	2	18	5	27	283	0	0	16	0	6
Trafik-Suçu	0	0	0	0	2	1	0	37	0	1	0	0
Uyuşturucu-Madde-Suçu	0	0	0	0	0	1	0	0	231	0	0	0
Yaralama	0	0	1	2	2	4	35	0	0	493	0	9
undefined	1	0	1	0	2	0	0	0	2	1	2582	0
İzinsiz-Mal-Alma-Suçu	1	2	0	0	1	17	8	0	0	2	2	357

Transformer-based Model

savasy/bert-turkish-text-classification from 🤗 - 10 epochs

	precision	recall	f1-score	support
Aldatmaya-Yönelik-Suç	0.94	0.95	0.95	237
Cana-Kastetme-Suçu	0.90	0.94	0.92	64
Cinsel-Saldırı-Suçu	0.97	0.93	0.95	148
Hakaret	0.70	0.66	0.68	118
Kanuna/Yasaya-Muhalefet-Olma	0.89	0.92	0.90	352
Other	0.83	0.76	0.80	493
Tehdit	0.72	0.78	0.75	359
Trafik-Suçu	1.00	0.90	0.95	41
Uyuşturucu-Madde-Suçu	0.98	1.00	0.99	232
Yaralama	0.94	0.89	0.91	546
undefined	1.00	1.00	1.00	2589
İzinsiz-Mal-Alma-Suçu	0.86	0.95	0.90	390
accuracy			0.93	5569
macro avg	0.89	0.89	0.89	5569
weighted avg	0.93	0.93	0.93	5569

Confusion Matrix

Comparison

Model	Accuracy	F1-Macro	F1-Weighted	Precision-Macro	Recall-Macro
LR	0.91	0.87	0.91	0.88	0.86
Majority Voting	0.92	0.87	0.92	0.88	0.87
FastText	0.92	0.87	0.92	0.88	0.86
LSTM	0.87	0.80	0.88	0.81	0.80
BERT-TR-Sent.	**0.93**	**0.89**	**0.93**	**0.89**	**0.88**
BERT-TR-Clf.	**0.93**	**0.89**	**0.93**	**0.89**	**0.89**

THANK YOU FOR YOUR ATTENTION!!!

A Work By Members of Group 1