YILDIZ TEKNOLOJİ TRANSFER OFİSİ A.Ş.

PROJE SONLANDIRMA RAPORU

PROJE ADI : HAVADİS

HAVADİS Projesi İle İlgili Olarak Çalışmalarımız 7 Mart 2019 Tarihinde Başarıyla sonuçlanmıştır.

Proje Bitiminde Aşağıdaki Sonuçlar Elde Edilmiştir :

Havadis Projesi 3 iş paketinden oluşmaktadır. Bunlar Duygu Analizi, Varlık İsmi Tanıma ve Anomali Tespiti'dir.

1. Birinci İş Paketi : Duygu Analizi

Büyük miktarda görüş belirten metni almak ve sınıflandırmak için otomatik duygu analizi sistemlerine ihtiyaç vardır. Makine öğrenme yaklaşımını ele alan mevcut araştırmalar çeşitli öznitelik üretme ve sınıflandırma teknikleri üretmiştir. Bu çalışmada terim-doküman indeksi veya kelime temsilleri ile öznitelik üretimi ile klasik makine öğrenme yöntemleri ve yapay sinir ağı modellerin Twitter mesajlarından oluşan veri kümesi üzerinde sınıflandırma performansı rapor edilmiştir.

Tablo 1. 1 Veri Kümesi

	Olumlu	Nötr	Olumsuz	Toplam
Eğitim	3663	4658	5511	13832
Test	916	1164	1377	3457

Raporun devamı iki ana bölüme ayrılmıştır: Bölüm 1'de kullanılan modeller, optimize edilmiş parametreler ve başarı sonuçları verilmiş, Bölüm 2'de ise sonuçlar özetlenmiştir.

1.1 Modeller ve Test Sonuçları

Sınıflandırmada kullanılan temel modeller Naive Bayes ve XGBoost, yapay sinir ağı modeller için ise basit derin yapay sinir ağı, LSTM-RNN ve CNN mimarileri analiz edilmiştir. Hiperparametere optimizasyonunda temel modeller için Grid Search, yapay sinir ağı modeller için ise Bayesian Optimization kullanılmıştır.

1.1.1. Temel Modeller

Temel modeller için terim doküman (TF) ve ters-terim doküman (IDF) öznitelikleri ile kelime (word-gram) ve karakter tabanlı birimler (character-gram) test edilmiştir. Arama uzayında elde edilen en iyi sonuçlar tablolarda verilmiştir.

1. Naive Bayes: *n-gram* öznitelik çıkarma işlemi sırasında kullanılan birim aralığını, *IDF* ters terim frekansının kullanılıp kullanılmadığını, *norm* n-gram normalizasyon metodunu, *smoothing* ise Naive Bayes yönteminin sparse terimler için kullandığı düzeltme katsayısını belirtir.

2. XGBoost: *n-gram* öznitelik çıkarma işlemi sırasında kullanılan birim aralığını, *IDF* ters terim frekansının kullanılıp kullanılmadığını, *learnign rate* ve *max depth* karar ağaçları parametreleri ve *number of classifiers* sınıflandırıcı sayısını belirtir.

Tablo 1. 2 Optimiza Naive Bayes Sonuçları-1

Naive Bayes, Word-Gram Model (n-gram:(1,2); IDF: False; Norm.: L2; Smoothing: 0.1)				
SINIF	**precision**	**recall**	**f1-score**	**support**
notr	0.7087	0.5814	0.6387	1180
olumlu	0.7115	0.5637	0.6290	919
olumsuz	0.6538	0.8477	0.7382	1359
weighted average	0.6879	0.6813	0.6753	3458
Test Accuracy: 0.6813				

Tablo 1. 3 Optimiza Naive Bayes Sonuçları-2

Naive Bayes, Character-Gram Model (n-gram:(1,5); IDF: True; Norm.: L2; Smoothing: 0.01)				
SINIF	**precision**	**recall**	**f1-score**	**support**
notr	0.7056	0.5890	0.6420	1180
olumlu	0.6919	0.5963	0.6406	919
olumsuz	0.6692	0.8278	0.7401	1359
weighted average	0.6877	0.6848	0.6802	3458
Test Accuracy: 0.6848				

Tablo 1. 4 Optimiza XGBoost Sonuçları-1

XGBoost, Word-Gram Model (n-gram:(1,1); IDF: True; Norm.: L2; Learning rate: 0.3; Max depth: 5; number of classifiers: 15;)				
SINIF	**precision**	**recall**	**f1-score**	**support**
notr	0.5397	0.4492	0.4903	1180
olumlu	0.6546	0.3547	0.4601	919
olumsuz	0.5243	0.7631	0.6215	1359
weighted average	0.5642	0.5474	0.5338	3458
Test Accuracy: 0.5474				

Tablo 1. 5 Optimiza XGBoost Sonuçları-2

XGBoost, Character-Gram Model (n-gram:(1,4); IDF: False; Norm.: L1; Learning rate: 0.3; Max depth: 5; **number of classifiers: 1000**[1];)				
SINIF	**precision**	**recall**	**f1-score**	**support**
notr	0.6350	0.6975	0.6901	1180
olumlu	0.6458	0.5822	0.7572	919
olumsuz	0.6403	0.6346	0.7221	1359
weighted average	0.6733	0.6726	0.6710	3458
Test Accuracy: 0.6726				

[1] Başarıyı ve eğitim süresini en çok etkileyen parametredir.

1.1.2. Yapay Sinir Ağı Modeller

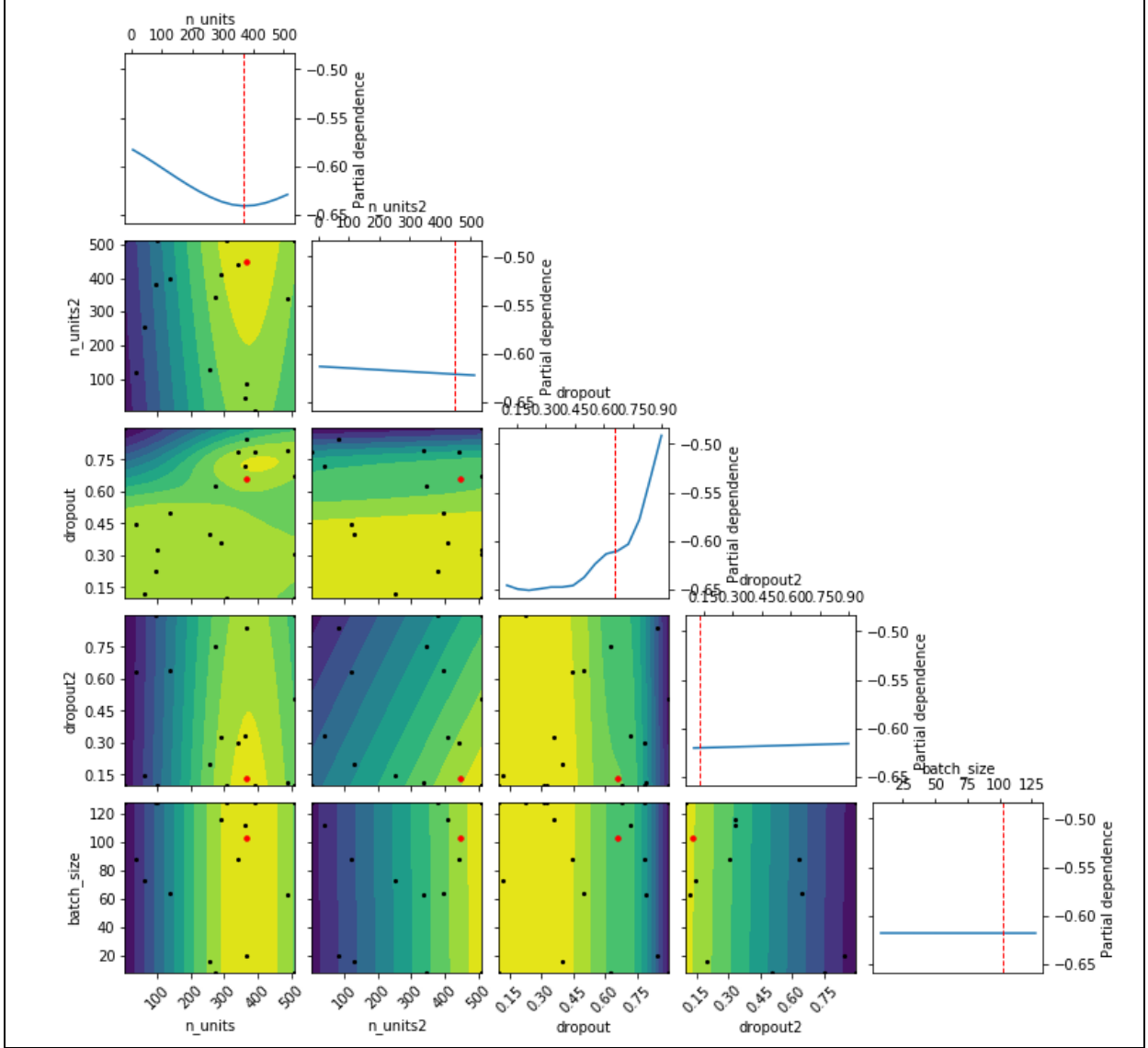
Yapılan testlerde, başka veri kümelerinden öğrenilen kelime temsillerinin performansı olumlu etkilemediği görülmüştür. Basit derin ağda bag-of-words yöntemi ile öznitelikler elde edilirken, LSTM ve CNN ağlarında eğitim kümesinden kelime temsilleri eğitim sırasında öğrenilmiştir.

1. Basit Derin Ağ: İki saklı katmanlı, her saklı katmanda belli bir dropout uygulanan modeldir. Yığın büyüklüğü (batch size), saklı katmanlardaki nöron sayısı (n_units) ve dropout oranları optimize edilmiş modele ilişkin sonuçlar ve arama uzayını gösteren grafik aşağıdadır.

Tablo 1. 6 Optimize Basit Derin Ağ **Sonuçları**

Basit Derin Ağ Modeli

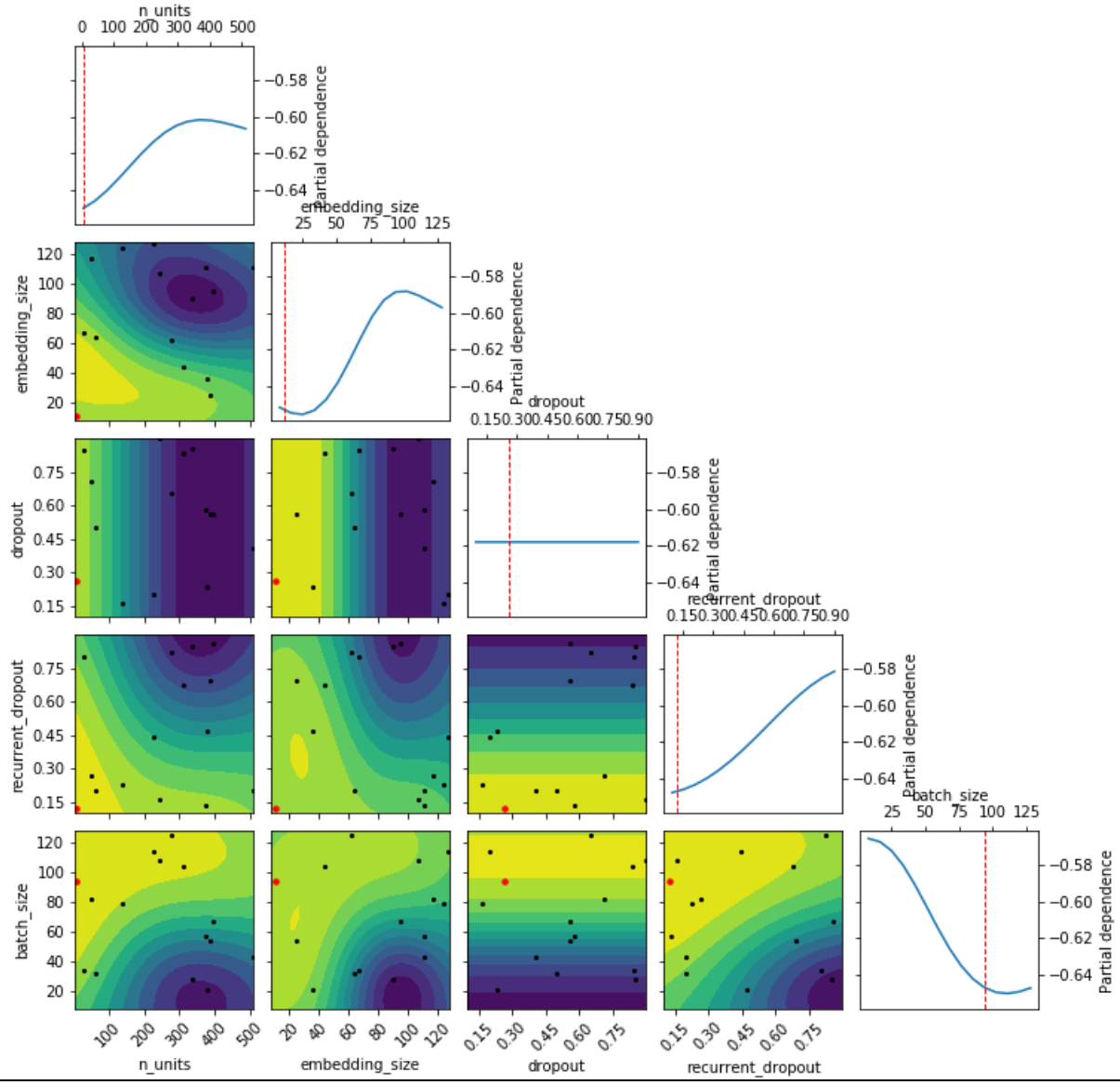
[n_units: 368, n_units2: 447, dropout:0.6581, dropout2:0.1334, batch_size:103]



Test Accuracy: 0.6873

2. LSTM Ağ: LSTM hücrelerinde oluşan RNN ağıdır. Yığın büyüklüğü (batch size), kelime temsili uzunluğu (embedding_size), saklı katmanlardaki LSTM hücre sayısı (n_units), katman dropout ve recurrent dropout oranları optimize edilmiş modele ilişkin sonuçlar ve arama uzayını gösteren grafik aşağıdadır.

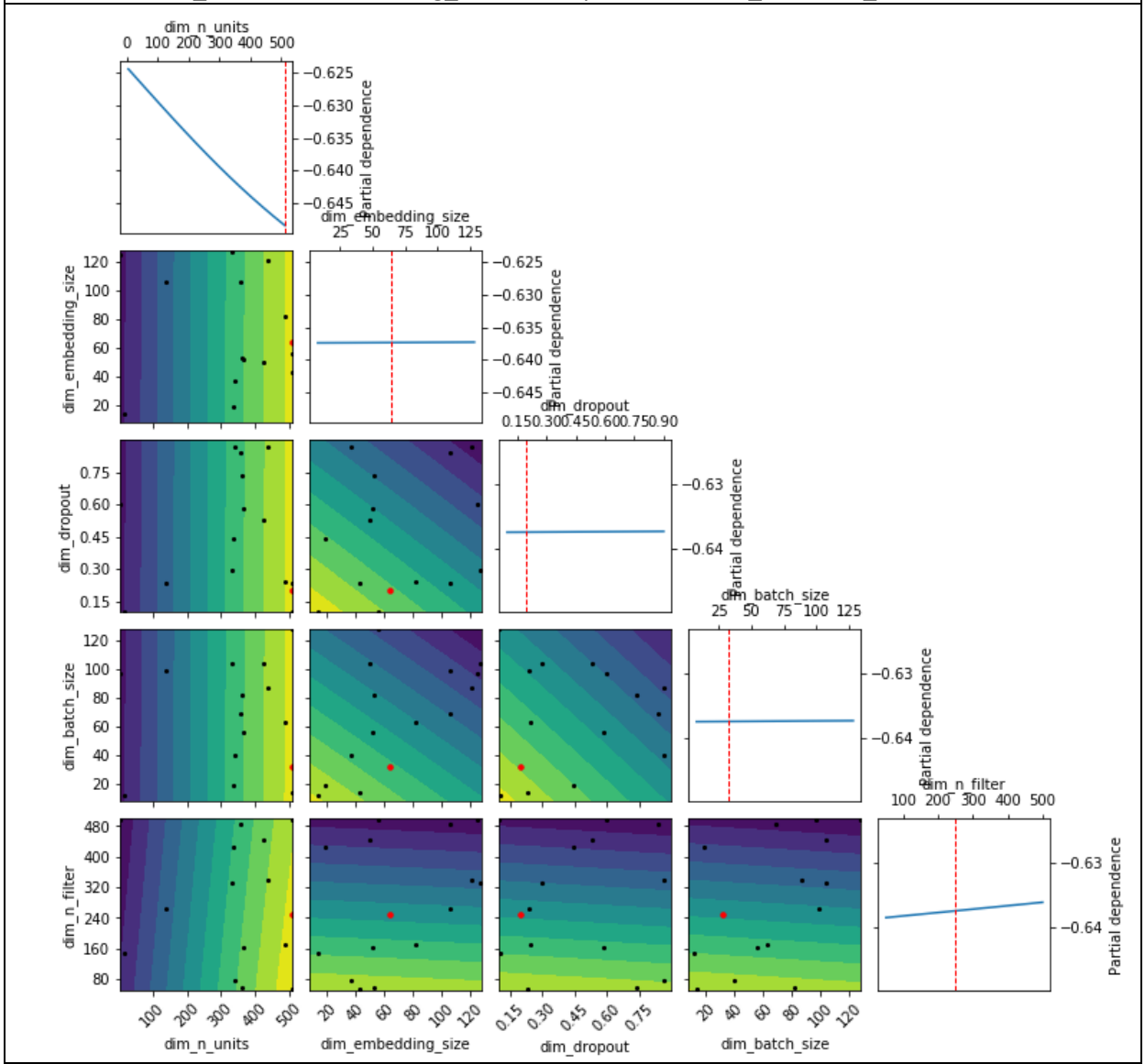
Tablo 1. 7 Optimize LSTM Ağ **Sonuçları**

LSTM Ağ [n_units:7, embedding_size:11, dropout:0.2613, recurrent_dropout:0.1231, batch_size:94]

Test Accuracy: 0.6611

3. CNN Ağ: Yığın büyüklüğü (batch size), kelime temsili uzunluğu (embedding_size), saklı katmanlardaki hücre sayısı (n_units), katman dropout oranı ve filtre sayısı optimize edilmiş modele ilişkin sonuçlar ve arama uzayını gösteren grafik aşağıdadır.

4.

Tablo 1. 8 Optimize CNN **Sonuçları**

CNN Ağ [n_units:512, embedding_size:64, dropout:0.2, batch_size:32, n_filter:250]

Test Accuracy: 0.6539

1.2 Sonuç

Özet sonuçlar tabloda verilmiştir.

Tablo 1. 9. Model Başarı Özeti

METOT	BAŞARI
Naive Bayes, Character-Gram Model	0.6848
XGBoost, Character-Gram Model	0.6726
Basit Derin Ağ	**0.6873**
LSTM Ağ	0.6611
CNN Ağ	0.6539

17.000 twitter mesajı içeren küme üzerinde yapılan eğitim ve test işlemleri sonucunda en yüksek başarı 0.6873 olarak optimize edilmiş Basit Derin Ağ Modelinde elde edilmiştir. Temel metotlardan Naive Bayes 0.6848 ile buna en yakın model olarak tespit edilmiştir. Literatürde iyi sonuçlar elde edildiği raporlanan XGBoost yöntemi ancak çok yüksek sayıda sınıflandırıcı kullanıldığında, dolayısıyla eğitim süresi uzatıldığında, yüksek başarıya ulaşmaktadır. LSTM ve CNN ağlarda Bayesian optimizasyon yapıldığı halde Basit Derin Ağ modeline yakın başarı sonucu elde edilememiştir.

2. İkinci İş Paketi : Varlık İsmi Tanıma

Bu dokümanda birçok NLP uygulamasına önişlem olabilen varlık-isim tespitine dair deney sonuçları yer almaktadır. Deneyler iki farklı veri kümesinde gerçekleştirilmiştir. İzleyen tablolarda veri kümelerine ait etiket dağılımları verilmiştir.

Tablo 2. 10. "7 Sınıflı" Veri Kümesi Kelime Tabanlı Dağılım

	Eğitim	Test
Date	2424	260
Location	11115	731
Money	1804	119
Organization	15192	824
Percent	1329	122
Person	21524	1019
Time	291	32
N/A	379361	20859
TOPLAM	**433080**	**23966**

Tablo 2. 11 "Havadis" Veri Kümesi Kelime Tabanlı Dağılım

	Eğitim	Test
Aitlik	730	325
Bina	1168	325
Kişi	9658	2906
Kurum	7308	2360
Miktar	2977	1509
Nitelik	4671	1249
Olay	720	157
Para	2138	1764
Plaka	107	54
Tarih	4820	2414
Tel	12	3
URL	147	74
Ürün	275	77
Yer	7563	3343
Yüzde	1901	1194
Zaman	274	229
N/A	245657	108516
TOPLAM	**290146**	**126439**

Etiketlemede kullanılan yöntemler CRF, Kelime Bi-LSTM Model ve Karakter-CNN Bi-LSTM Model'dir. Raporun devamı iki ana bölüme ayrılmıştır: Bölüm 1'de kullanılan modeller ve başarı sonuçları verilmiş, Bölüm 2'de ise sonuçlar özetlenmiştir. Tablo ağırlıklı bu raporda genel resmi daha iyi görebilmek için Bölüm 2'nin incelenmesi önerilir.

2.1 Modeller ve Test Sonuçları

2.1.1. Temel Model: CRF

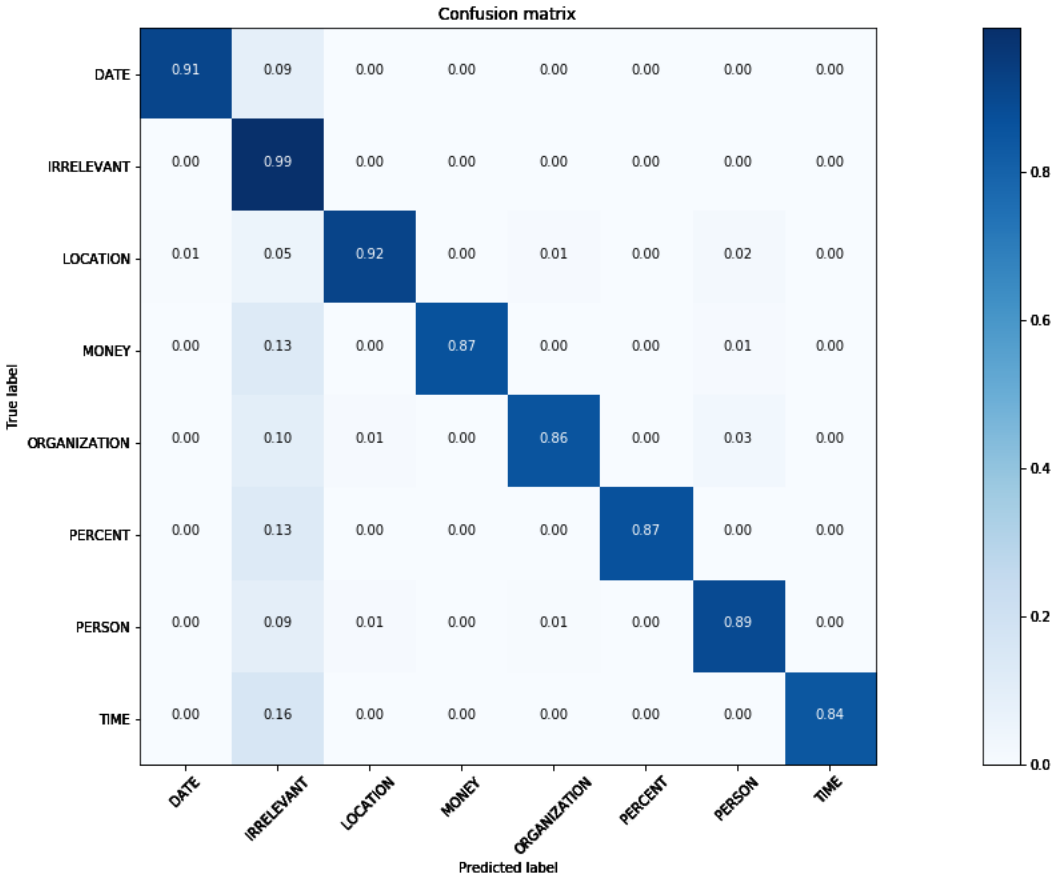
Deneye ait testler her iki veri kümesi için de etiket başarı oranları ve normalize edilmiş karışıklık matrisi verilmiştir. Kullanılan CRF öznitelikleri şunlardır:

- Kelime küçük harflerden oluşur: (True/False)
- 3-gram
- Kelime büyük harflerden oluşur: (True/False)
- Kelimenin yalnız ilk harfi büyük harftir: (True/False)
- Kelime rakamlardan oluşur: (True/False)

Tablo 2. 12. "7 Sınıflı" Veri Kümesi CRF Sonuçları

	Precison	Recall	F1	Support
Date	0.89	0.91	0.90	260
Location	0.96	0.92	0.94	731
Money	0.93	0.87	0.90	119
Organization	0.96	0.86	0.91	824
Percent	0.95	0.87	0.91	122
Person	0.90	0.89	0.90	1019
Time	0.96	0.84	0.90	32
N/A (irrelevant)	0.99	0.99	0.99	20859

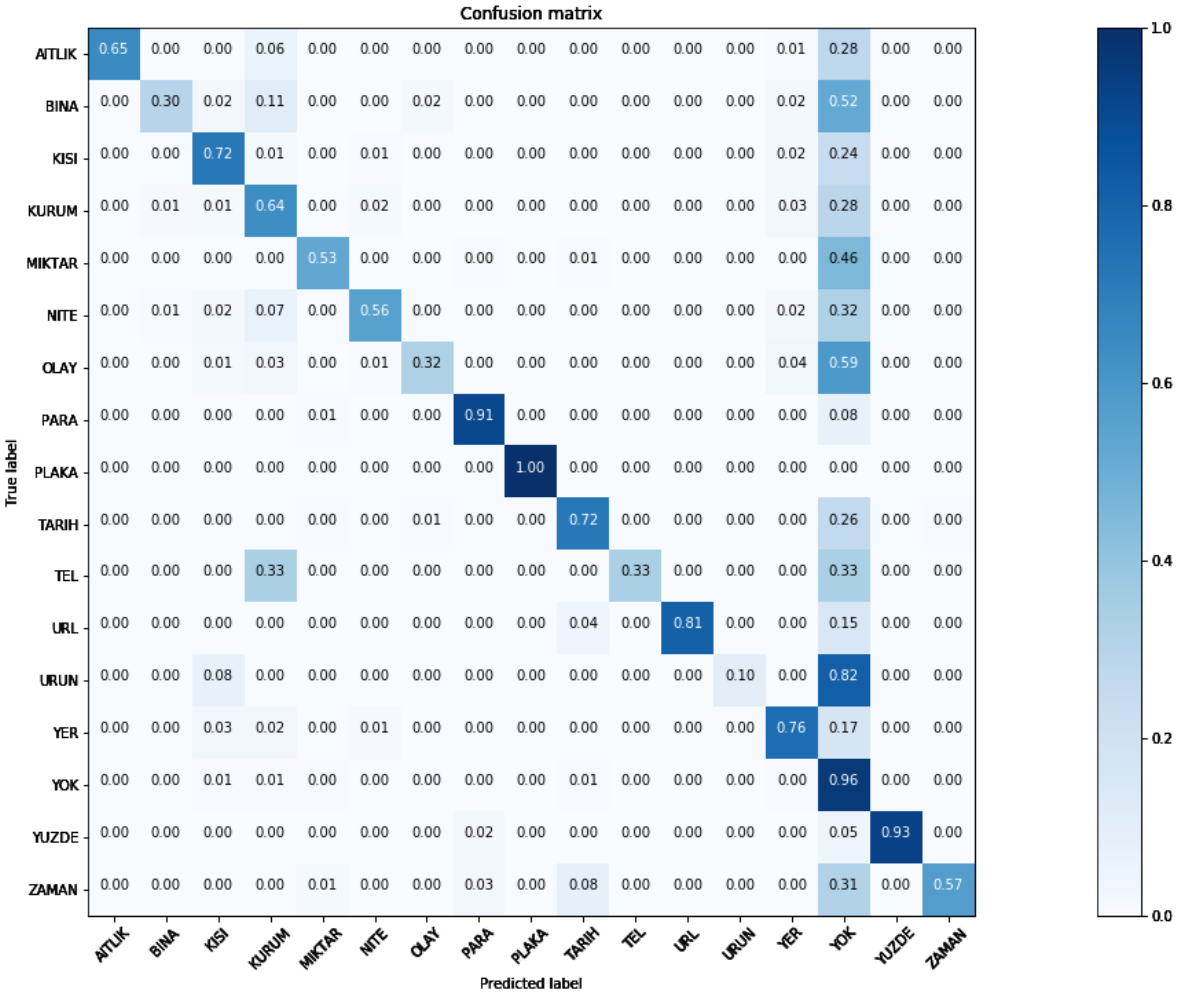
Tablo 2. 13. "7 Sınıflı" Veri Kümesi CRF Normalize EdilmişKarışıklık Matrisi



Tablo 2. 14. "Havadis" Veri Kümesi CRF Sonuçları

	Precision	Recall	F1	Support
Aitlik	0.74	0.65	0.69	265
Bina	0.43	0.30	0.35	325
Kişi	0.69	0.72	0.71	2906
Kurum	0.61	0.64	0.63	2363
Miktar	0.61	0.53	0.57	1572
Nitelik	0.60	0.56	0.58	1259
Olay	0.39	0.32	0.35	162
Para	0.81	0.91	0.86	1764
Plaka	0.66	1.00	0.79	54
Tarih	0.65	0.72	0.69	2414
Tel	0.50	0.33	0.40	3
URL	0.94	0.81	0.87	74
Ürün	0.24	0.10	0.15	77
Yer	0.78	0.76	0.77	3343
Yüzde	0.85	0.93	0.89	1210
Zaman	0.65	0.57	0.61	229
N/A (Yok)	0.96	0.96	0.96	113426

Tablo 1. 15. "Havadis" Veri Kümesi Normalize EdiilmişKarışıklık Matrsisi



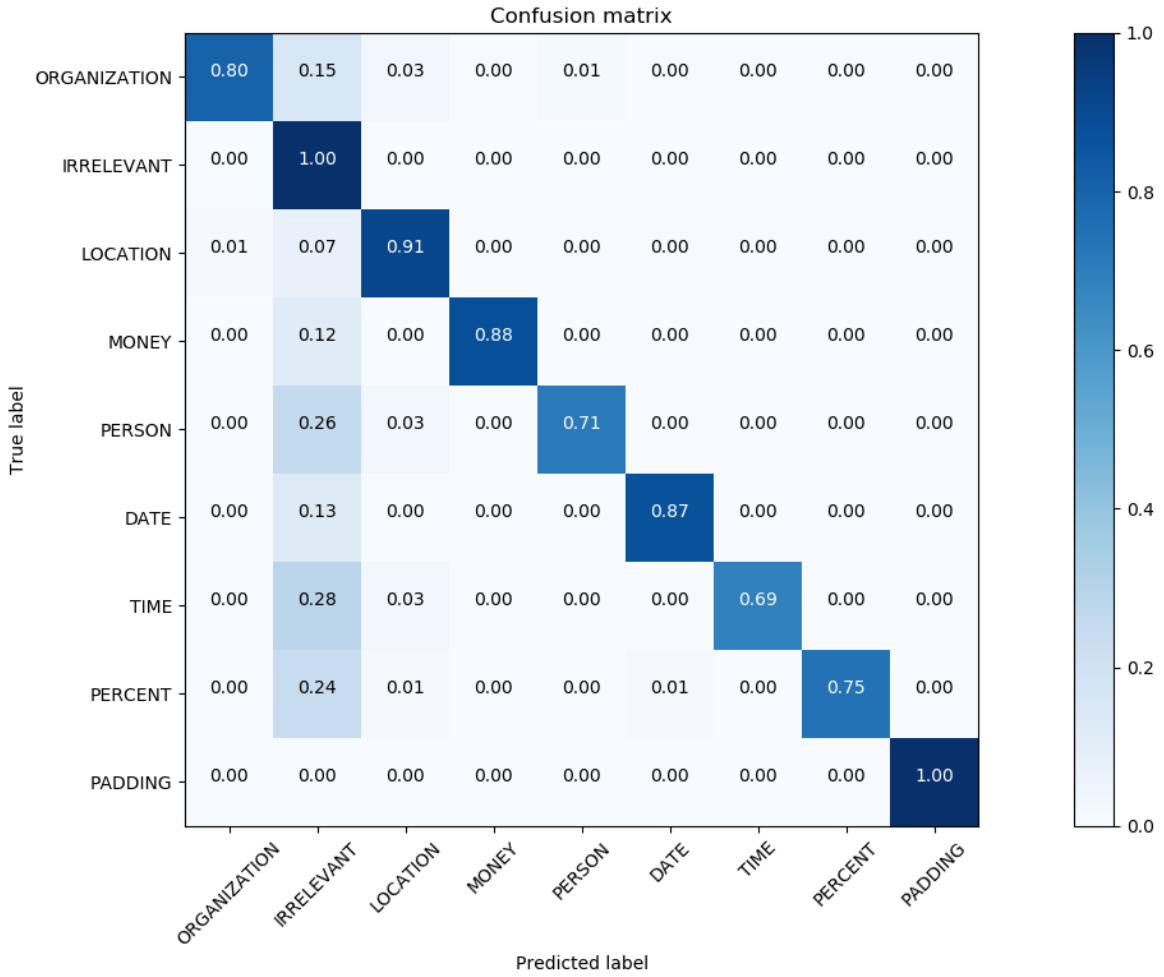
2.2 Yapay Sinir Ağı Modelleri

2.2.1 Word-Level Bi-LSTM:

Tablo 2. 16. "7 Sınıflı" Veri Kümesi Bi-LSTM Sonuçları

	Precison	Recall	F1
Date	0.85	0.87	0.86
Location	0.91	0.91	0.91
Money	0.88	0.88	0.88
Organization	0.96	0.80	0.88
Percent	0.93	0.75	0.83
Person	0.97	0.71	0.82
Time	0.92	0.69	0.79
N/A (irrelevant)	0.98	1.00	0.99

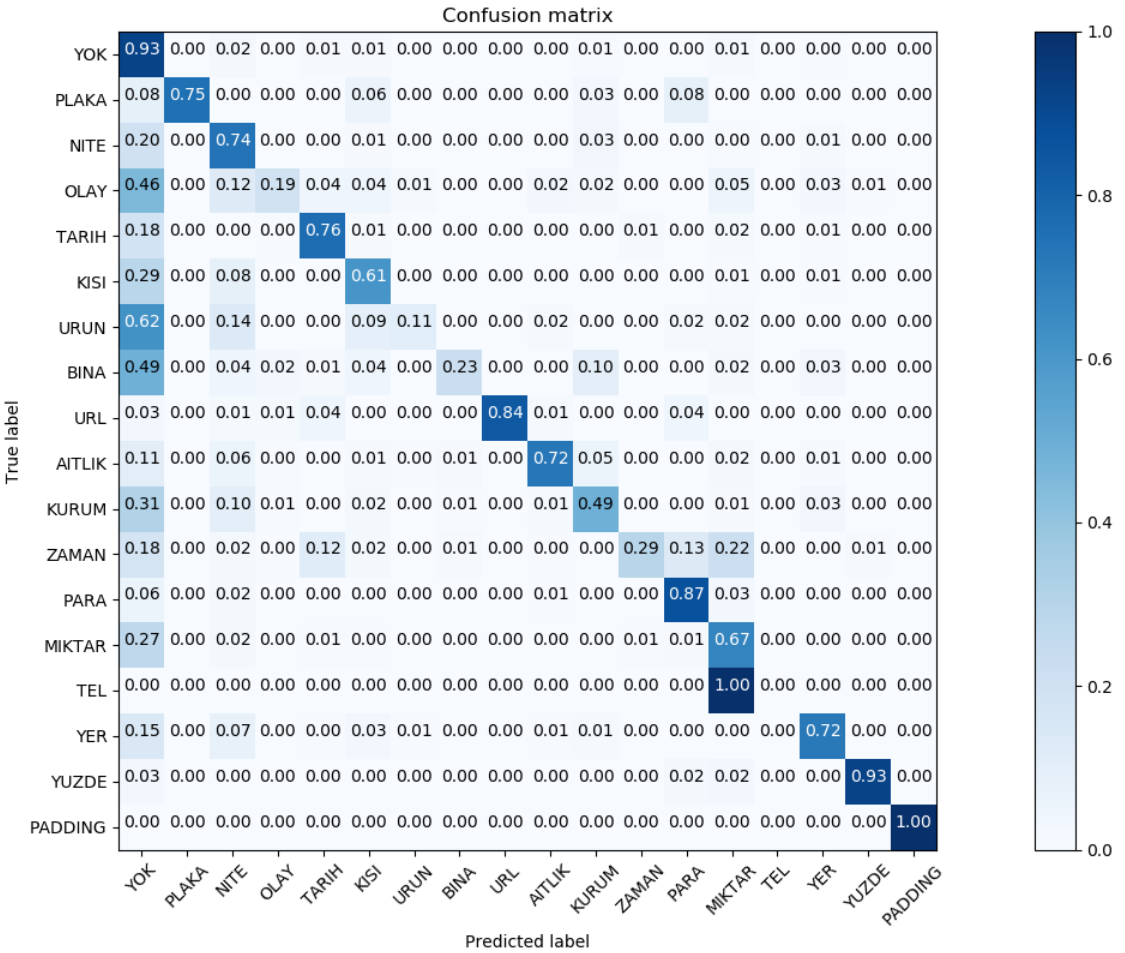
Tablo 2. 17. "7 Sınıflı" Veri Kümesi Bi-LSTM Normalize Edilmiş Karışıklık Matrisi



Tablo 2. 18. "Havadis" Veri Kümesi Bi-LSTM Sonuçları

	Precision	Recall	F1
Aitlik	0.45	0.72	0.55
Bina	0.41	0.23	0.29
Kişi	0.60	0.61	0.60
Kurum	0.58	0.49	0.53
Miktar	0.39	0.67	0.49
Nitelik	0.23	0.74	0.35
Olay	0.14	0.19	0.16
Para	0.77	0.87	0.82
Plaka	0.54	0.75	0.63
Tarih	0.59	0.76	0.67
Tel	0.0	0.0	0.0
URL	0.83	0.84	0.84
Ürün	0.05	0.11	0.07
Yer	0.77	0.72	0.74
Yüzde	0.84	0.93	0.88
Zaman	0.29	0.29	0.29
N/A (Yok)	0.97	0.93	0.95

Tablo 2. 19. "Havadis" Veri Kümesi Bi-LSTM Normalize Karışıklık Matrisi

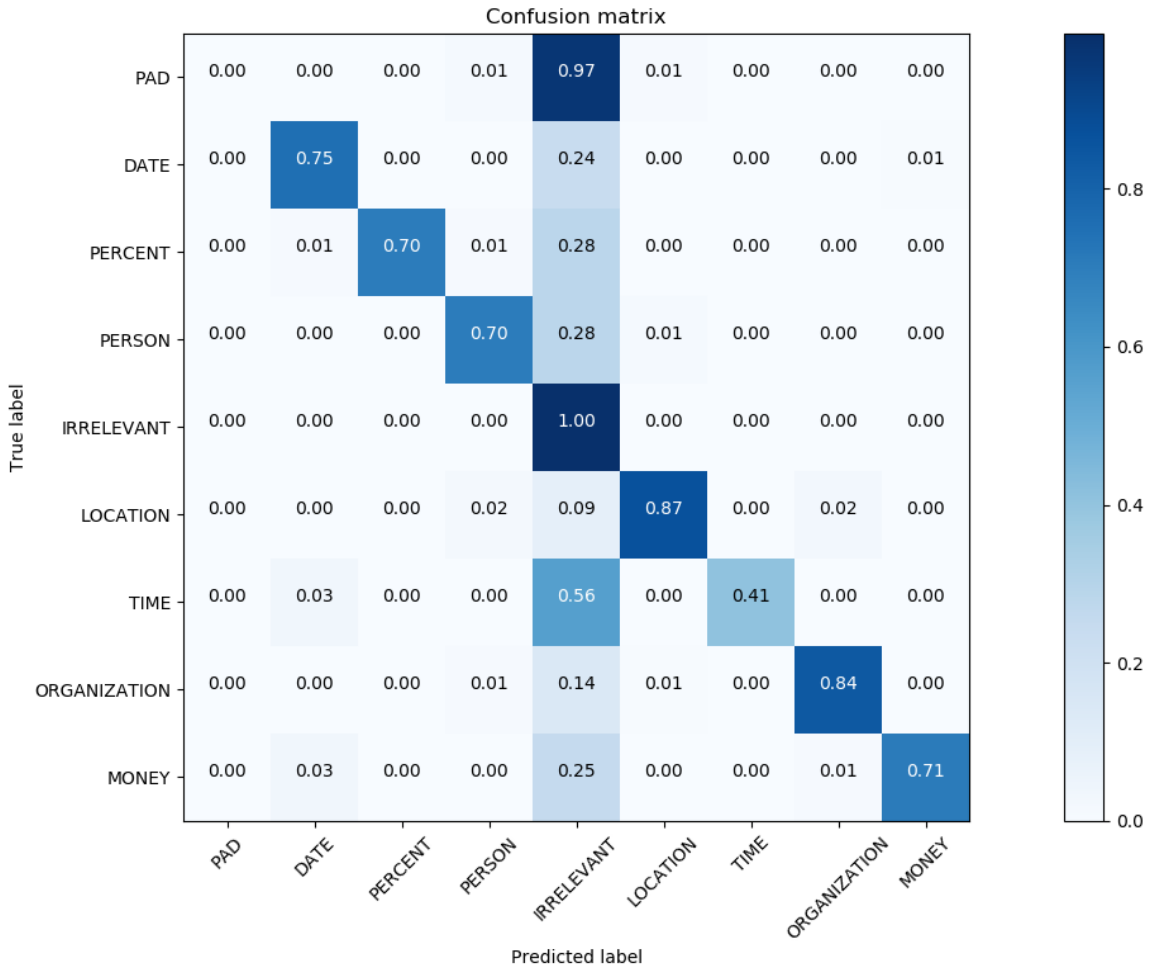


2.2.2 Character-CNN Bi-LSTM Model

Tablo 2. 20. "7 Sınıflı" Veri Kümesi Char-CNN Bi-LSTM Sonuçları

	Precison	Recall	F1
Date	0.88	0.75	0.81
Location	0.19	0.87	0.31
Money	0.10	0.71	0.17
Organization	0.40	0.84	0.54
Percent	0.93	0.70	0.80
Person	0.15	0.70	0.25
Time	1.00	0.41	0.58
N/A (irrelevant)	0.07	1.00	0.13

Tablo 2. 21. "7 Sınıflı" Veri Kümesi Char-CNN Bi-LSTM Normalize Karışıklı Matrisi[2]

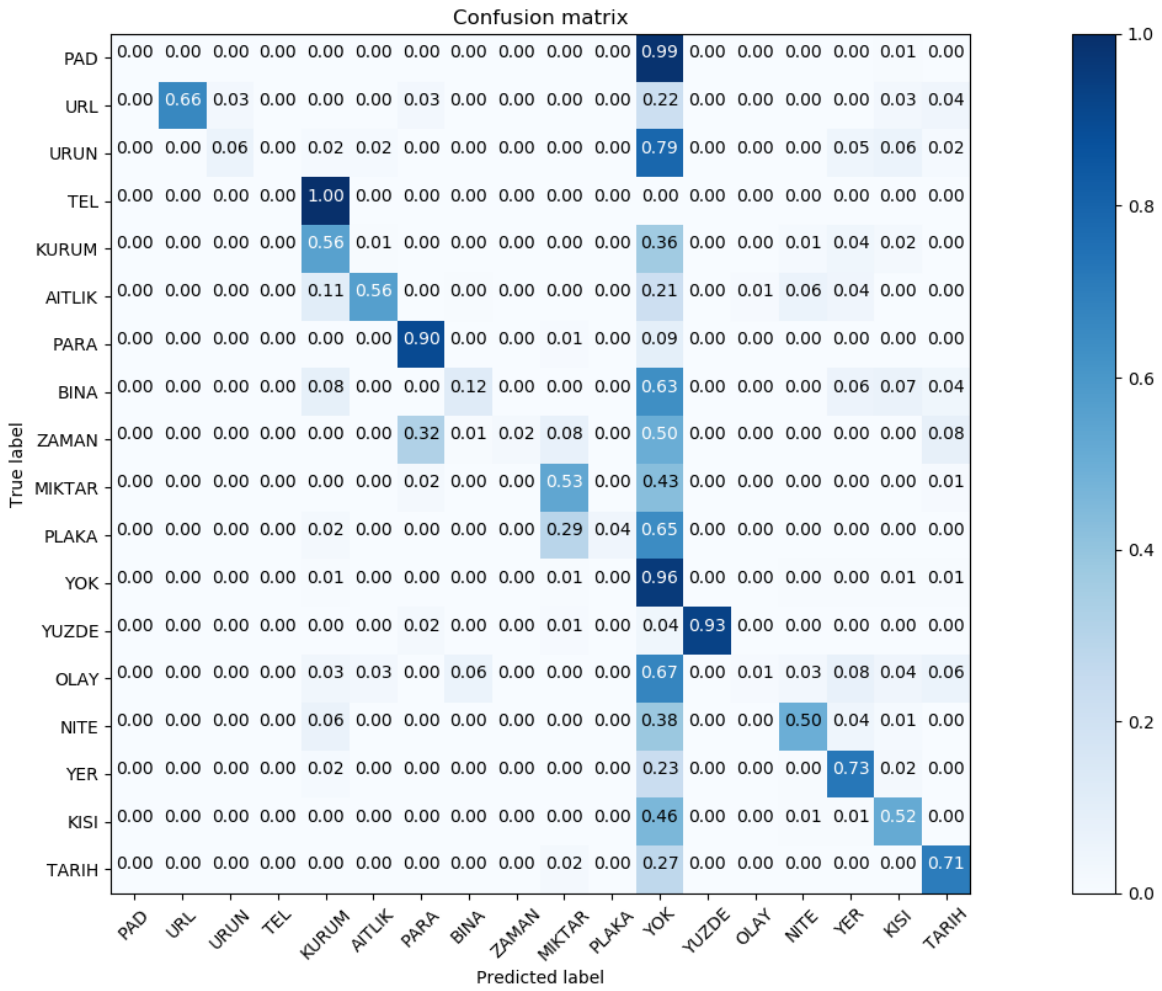


[2] Burada görünen PAD etiketi yalnız eğitim kümesinde olduğu için test aşamasında incelenirken IRRELEVANT olarak düşünülmeli.

Tablo 2. 22. "Havadis" Veri Kümesi Char-CNN Bi-LSTM Sonuçları

	Precision	Recall	F1
Aitlik	0.66	0.58	0.62
Bina	0.41	0.12	0.18
Kişi	0.43	0.54	0.47
Kurum	0.53	0.58	0.55
Miktar	0.41	0.12	0.18
Nitelik	0.45	0.52	0.48
Olay	0.41	0.12	0.18
Para	0.76	0.82	0.79
Plaka	0.0	0.0	0.0
Tarih	0.67	0.69	0.68
Tel	0.0	0.0	0.0
URL	0.95	0.54	0.69
Ürün	0.43	0.05	0.08
Yer	0.78	0.67	0.72
Yüzde	0.84	0.86	0.85
Zaman	0.40	0.10	0.16
N/A (Yok)	0.47	0.95	0.63

Tablo 2. 23. "Havadis" Veri Kümesi Char-CNN Bi-LSTM Normalize Karışıkık Matrisi



2.3 Sonuç

Yöntemlerin F1 skor karşılaştırması tabloda verilmiş, "Havadis" tablosunda, "7 Sınıflı" veri kümesi ile ortak etiketler sarı işaretlenmiştir.

Tablo 2. 24. "7 Sınıflı" Veri Kümesi Karşılaştırmalı F1 Skoları

	CRF	Word Bi-LSTM	Char CNN Bi-LSTM
Date	0.90	0.86	0.81
Location	0.94	0.91	0.31
Money	0.90	0.88	0.17
Organization	0.91	0.88	0.54
Percent	0.91	0.83	0.80
Person	0.90	0.82	0.25
Time	0.90	0.79	0.58
N/A (irrelevant)	0.99	0.99	0.13

Tablo 2. 25. "Havadis" Veri Kümesi Karşılaştırmalı F1 Skoları

	CRF	Word Bi-LSTM	Char CNN Bi-LSTM
Aitlik	0.69	0.55	**0.62**
Bina	0.35	0.29	0.18
Kişi	0.71	0.60	0.47
Kurum	0.63	0.53	0.55
Miktar	0.57	0.49	0.18
Nitelik	0.58	0.35	**0.48**
Olay	0.35	0.16	0.18
Para	0.86	0.82	0.79
Plaka	0.79	0.63	0.0
Tarih	0.69	0.67	**0.68**
Tel	0.40	0.0	0.0
URL	0.87	0.84	0.69
Ürün	0.15	0.07	0.08
Yer	0.77	0.74	0.72
Yüzde	0.89	0.88	0.85
Zaman	0.61	0.29	0.16
N/A (Yok)	0.96	0.95	0.63

Her iki veri kümesinde de CRF performansı daha iyidir. Genel olarak TARİH (DATE), YÜZDE (PERCENT) ve PARA (MONEY) etiketleri kolay ayırt edilmiştir. Char-CNN Bi-LSTM modeli iyi ayırt edilebilen etiketlerde en azından Word Bi-LSTM'e yakın (hatta Havadis için TARİH ve NİTELİKTE daha iyi) sonuçlar göstermiş, ancak genel olarak üç model içinde en kötü performansı vermiştir. Rakam içeren kelimelerin etiketlenmesi CRF'de önemli ölçüde daha üstündür. Optimize edilmiş Word Bi-LSTM modeli CRF'e yakın veya daha üstün sonuçlara götürebileceği düşünülmektedir.

3. Üçüncü İş Paketi : Anomali Tespiti

Yapılan tüm işlemler Python dilinde farklı kütüphaneler kullanılarak yapılmıştır.

- Tüm tweet veri kümesini csv formatından istenilen json formatına çevrildi.
- Tüm tweetler için tarih, Türkiye saatine göre ayarlandı.
- Tüm tweetler için;
    - o Stopwordleri çıkarıldı.
    - o Harf, rakam dışındaki tüm diğer karakterleri silindi.
    - o # veya @ ile başlayan tüm kelimeler silindi.
    - o Stopword listesinde olmayabilecek 3 harften daha kısa tüm kelimeleri atıldı.

o Metinler içerisinden emoji, url link, vs. fazlalıklar da çıkarıldı.
o Bütün kelimelerden Türkçe karakterler çıkarılarak, yerine İngilizce alfabeden karşılıkları getirildi.

3.1 Modeller ve Test Sonuçları

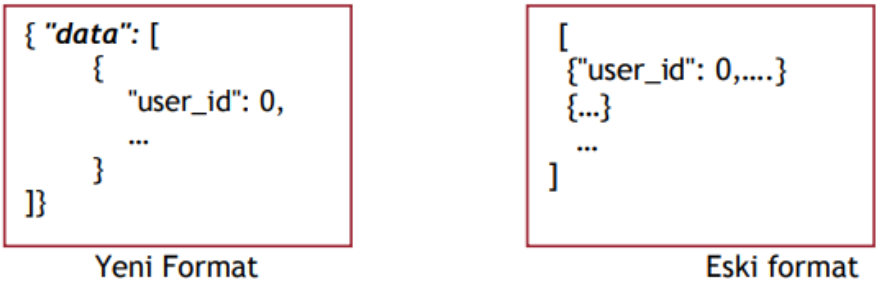
3.1.1. İzlenen Yöntem ve Parametreler

Gerçek zamanlı bir sistem olması açısından basit ve hızlı bir yöntem önerilmiştir. Literatür taramasında da çalışmaların karmaşık modeller yerine basit özellikler kullanılarak yapıldığı görülmüştür. Twitter verisi özelinde yapılan çalışmalarda, tf/idf değerleri, mention, hashtag gibi özel kelime grupları anomali olayların tespitinde kullanılmıştır.

• Bu çalışmada da frekans tabanlı bir yöntem geliştirilmiştir.
• Günlük olarak ayrılan tweetlerden kelime frekansları bulundu.
• Bu frekans listesinden yola çıkarak iki seçenek belirlendi. (percentFlag ile seçilebilir) o Günlük atılan toplam twit sayısının belli bir yüzdesi alınarak bir eşik değeri hesaplandı. (percent) o Günlük atılan toplam twitten bağımsız olarak, sabit bir eşik değeri belirlendi. (minWordFreq)
• Belirlenen değerin üstünde frekansa sahip kelimeler günlük olan ayrılmış listelerde tutulur.
• Aranacak anomalilerin en fazla kaç kelimeden oluşacağını belirlemek için ngramsize değeri set edilir. Bu değer default 2 olarak belirlenmiştir.
• Bir kelime grubunun en az kaç süreyle devam ettiğini belirlemek için minDayForAnomaly parametresi kullanılır. Default olarak bu değer 7 olarak belirlenmiştir. (bir hafta)
• Girdi olarak verilen json dosyası için bin klasörü içine eklenecektir. Dosya adı filename adında bir değişkende tutulur.
• Bu kelimelerden/kelime gruplarından en az bir hafta boyunca aralıksız olarak listede bulunanlar anomali olayı olarak işaretlendi.
• Önceki aşamadan alınan anomaliler ve tarih değerleri alınarak, her anomali için başladığı ve bittiği gündeki twitler incelenerek zaman değeri de elde edilmiştir.
• Olayla ilgili ilk ve son tweet zamanı istenilen formatta bir json dosyasına kaydedildi.
• Çıktı dosyası "dataset374K-minFreq %1-minDay7-2gram-v2" isminde oluşmaktadır. Burada belirtilen ifadeler, verisetinin 1000 ölçeğinde büyüklüğü (dataset374K), % kaçlık frekans eşiği (minFreq %1), en az kaç süren olay arandığı (mindDay5) ve kelime gruplarındaki en fazla kelime sayısını (2gram) göstermektedir.
• Eğer yüzdelik yerine doğrudan eşik değeri kullanılırsa "dataset374K-minFreq 10-minDay7-2gramv2" şeklinde bir çıktı dosyası oluşacaktır.

3.2 Sonuç

370.000 tweet bulunan bir veri kümesinde yapılan çalışma yaklaşık 3-4 dakika çalışma süresine sahiptir. Yaklaşık 125 sn json dosyasının okunması için harcanmıştır. Girdi formatında verilen json dosyasının nested olması durumu bu süreyi normalde 3 kat daha yavaşlatmıştır.



Şekil 3.1 Yeni ve Eski Json Formatı

Yaklaşık 60 sn günlük olarak kelime frekansı bulmak için harcanmıştır. Yaklaşık 0,4 sn 7 gün süren anomali olaylarının bulunması için harcanmıştır. Yaklaşık 7 sn bulunan anomalilerin başlama-bitme zamanlarının tespiti ve json dosyasına yazılması için harcanmıştır.

Sonuçlarda "adalet, chp, akp, teror orgutu, abd, …" gibi kelime ve kelime gruplarının çıktığı görülmüştür. Süre azaltıldığında daha fazla anomalinin yakalandığı da tespit edilmiştir. Özellikle sosyal medyada bir olayın 1 hafta sürede gündemde kalması zaten onun TT (trend topic)listesinde bulunduğunu göstermektedir. Bu bakımdan bu sürenin daha kısa tutulması gerektiği düşünülmektedir. Kendi veri setimiz özelinde, kelime gruplarının büyüklüğü incelendiğinde 2 kelimelik grupların ideal olduğu görülmüştür. Daha fazla sayıda kelimeye bakıldığında benzer olayların geldiği tespit edilmiştir. Bu durum başka bir veri setinde farklı parametre değeri ile denenebilir. Veri seti oluşturulurken belli anahtar kelimelere göre twitler çekilmiştir. Bu kelimelerin tweet çekilen anahtar kelimelerin içinde olduğu düşünülmektedir. Testlerin yapıldığı veri setinin yukarıda belirtilen problemi sebebiyle parametreler değiştirilebilir olarak ayarlanmıştır. Bu bağlamda SSM tarafından yapılacak test için parametrelerin fine-tuning yapılması gerekebilir (frekans eşik değer, yüzdelik değer, minimum gün sayısı, n-gram büyüklüğü parametreleri için de farklı denemeler yapılabilir).

Proje Yöneticisinin ,

Adı /Soyadı : Banu DİRİ
Unvan : Prof. Dr.
Üniversite : Yıldız Teknik Üniversitesi
Fakülte : Elektrik Elektronik Fakültesi