Türkçe Haber Derlemi ve Haberlerin Kategorik Sınıflandırılması

Ömer Gözüaçık

Biz Kimiz, Ekip Bilgisi

- Ömer Gözüaçık (Bireysel Katılım)
  - Bilkent Üniversitesi Elektrik-Elektronik Mühendisliği 2017 mezunudur. Mezuniyeti takip eden yıldan itibaren aynı üniversitede Bilgisayar Mühendisliğinde yüksek lisans yapmaktadır. Birden fazla lisans ve yüksek lisans dersine asistanlık tecrübesi bulunmaktadır. Yaklaşık üç yıldır veri akışları üzerinde makine öğrenmesi üzerine çalışmaktadır.
  - Github: ogozuacik
  - LinkedIn: link

TÜRKÇE HABER DERLEMİ

Haberlerin Kategorik Sınıflandırması

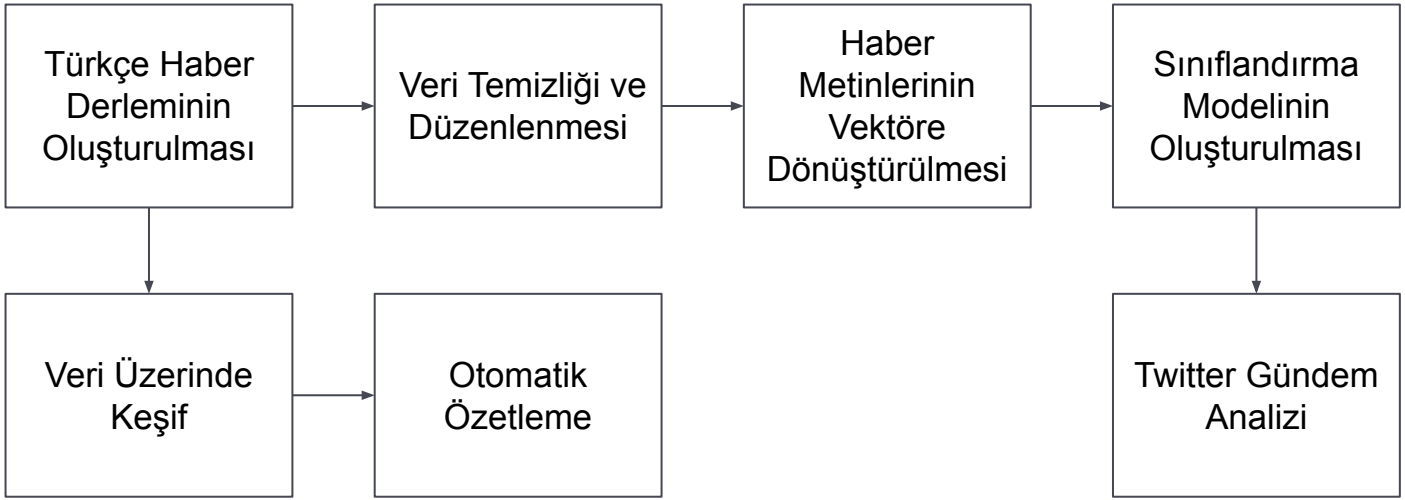
Problem Nedir?

- Türkçe işaretli veri setlerinin yetersizliği
  - Örnek sayılarının azlığı
  - Açık kaynak erişim imkanının kısıtlı oluşu (çoğu veri seti için izin alınması ya da üyelik gibi gereksinimler var)
- Türkçe haberlerin sınıflandırılmasında genel kategoriler:
  - Haberler türlerinden bağımsız olarak "Gündem" vb. konu başlıkları altında paylaşılıyor.
  - Haber metinlerinin internetten çekilmesi ile oluşturulan veri setlerinde bu tip kategorilerden gelmiş örnekler model üretiminde kullanılamıyor.

Çözüm Nedir?

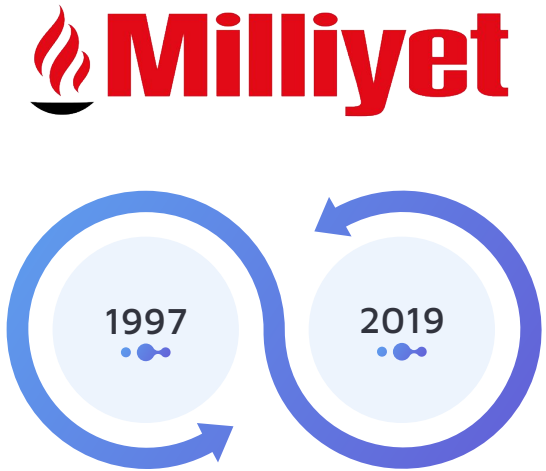
- Proje başvurusunda ana hedef “Haber başlıkları üzerinden konu (kategori) tahmini” olarak tanımlanmıştır.
    - Milliyet.com üzerinden 1997-2019 yılları arası çıkan haberler kullanılmıştır (kendi oluşturduğumuz). Derlem açık kaynak olarak paylaşılmıştır.
    - Haber başlıklarına ek olarak haber metni ve özeti üzerinden de konu (kategorinin) tahmin edilmesi sağlanmıştır.
- Oluşturulan derlem ve sınıflandırma modeli ile yapılabilecekler:
    - Yeni işaretli veri setlerinin oluşturulması, genel kategori sınıfına sahip haberlerin (gündem vs.) ilgili oldukları kategorilerdeki sınıflara (ekonomi, siyaset) atanması.
    - Twitter üzerinde @nedenttoldu gibi hesaplar üzerinden günlük ülke gündeminin takibi.
    - Türkiye’nin son 23 yılına haberler üzerinden bakış.

İş Akışı


Türkçe Haber Derleminin Oluşturulması
Veri Temizliği ve Düzenlenmesi
Haber Metinlerinin Vektöre Dönüştürülmesi
Sınıflandırma Modelinin Oluşturulması
Veri Üzerinde Keşif
Otomatik Özetleme
Twitter Gündem Analizi

Türkçe Haber Derleminin Oluşturulması

- Milliyet.com adresinden 1997-2019 yılları arasında çıkan, haberler büyük oranda çekilmiştir.
- Her haber, internet sayfasında bulunduğu gibi kaydedilmiştir.
- Başlık, özet, haber metni, kategori ve link
    - Özet bazı haberlerde bulunmamaktadır.
- Temizlenmemiş hali ile 116.068 örnek bulunmaktadır. (403.9Mb)
- Benzer bir veri seti olan TTC 3600, 3600 örnek içermektedir. (link)



Veri Temizliği ve Düzenlenmesi

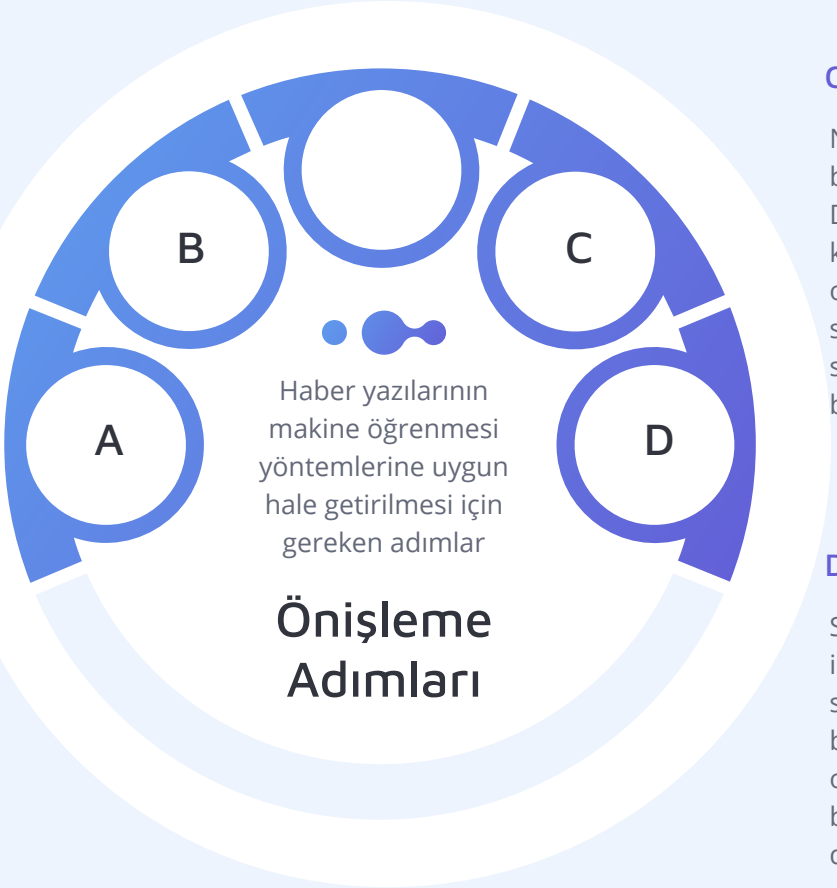
- Toplam 199 kategori bulunmaktadır. Bu kategorilerden bazıları köşe yazarlarıdır. Bazı kategorilerde 10'dan az haber bulunmaktadır.
- Kategorilerin birleştirilmesi ve silinmesi:
  - Futbol, Skorer, Basketbol, Fenerbahçe, Galatasaray, Beşiktaş... gibi spor ile alakalı haberlere genel bir kategori oluşturulup "Spor" adı verillmiştir.
  - Teknoloji_Bilim, Magazin, Yaşam ve Dünya kategorileri de yukarıdaki mantık ile birden fazla kategorinin birleştirilmesi ile oluşturulmuştur.
  - Kategorisi köşe yazarları olan örnekler silinmiştir.
- Haberlerin orijinal (milliyet sitesinde bulundukları) kategori derlemde kategori_yedek sütununda tutulmuştur.

B. Sayılar

Bütün sayılar kategori tahmininde işe yaramayacağı düşünülerek boşluk ile değiştirilmiştir.

C. Noktalama İşaretleri

Noktalama işaretlerinin yerine boşluk getirilmiştir.
Doğrudan silmek yerine boşluk konulması haberlerde büyük oranda bulunan virgülden sonra atlanan boşluk hatası sonucu kelimelerin birleşmesini önüne geçmektir.



A. Genel yazım hataları

Verilerde genel olarak bulunan yazım hataları düzeltilmiştir.
Örneğin:
... kazandıTürkiye... -->
...kazandı Türkiye...

D. Fazla boşluklar

Sayıların, noktalama işaretlerinin temizlenmesi sırasında iki kelime arası birden fazla boşluk oluşmuştur. Bu boşluklar tek boşluk olucak şekilde değiştirilnmiştir.

Örnek Veri ve Derlem Hakkında Bilgiler

	news_title	summary	category	date	link	news	category_backup
0	mili piyango yılbaşı çekilişi sıralı tam liste...	her yıl büyük bir heyecana sahne olan ve milyo...	Gündem	2019-12-31	/gundem/mili-piyango-yilbasi-cekilisi-sirali-t...	insanların en büyük hayallerinden biri zengin...	Gündem
1	mili piyango sıralı tam listesi aralık çekiliş...	her yıl büyük bir heyecana sahne olan milli pi...	Gündem	2019-12-31	/gundem/mili-piyango-sirali-tam-listesi-31-ara...	insanların en büyük hayallerinden biri zengin...	Gündem
2	mpi bilet amorti ikramiye sonuç sorgulama ekr...	milli piyango yılbaşı özel çekilişi aralık tar...	Gündem	2020-01-01	/gundem/mpi-bilet-amorti-ikramiye-sonuc-sorgul...	milli piyango yılbaşı özel çekilişinin ardında...	Gündem
3	yılbaşı mesajları ve sözleri sevdiklerinize gö...	google yılbaşı gününe özel doodle yayımladı yı...	Gündem	2019-12-31	/gundem/yilbasi-mesajlari-sosyal-medya-ve-tele...	yeni yıl heyecanı tüm yurdumuzu sardı bu akşam...	Gündem
4	yeni yıl mesajları yılbaşı mesajları kısa uzun...	bugün günlerden aralık senenin son günü yılbaş...	Gündem	2019-12-31	/gundem/yilbasi-kutlama-mesajlari-2020-yeni-yi...	bugün en özel günlerden biri yeni yıl öncesi k...	Gündem

- **news_title:** haber başlığı
- **summary:** haber özeti
- **category:** haberin kategorisi (filtrelenmiş)
- **date:** haberin paylaşıldığı gün (bazı haberlerde bu tarihte sapmalar bulunmaktadır)
- **link:** haberin çekildiği bağlantı adresi
- **news:** haber metni
- **category_backup:** haberin orijinal (milliyet sitesinde bulunduğu) kategori

Haber Metinlerinin Vektöre Dönüştürülmesi

- Sınıflandırma işlemi öncesi sklearn kütüphanesi üzerinden CountVectorizer fonksiyonu ile haber metinleri vektöre çevrilmiştir.
    - BoW (Bag of words) yöntemi kullanılmıştır. Vektörde her sütun bir kelimeyi, her satır ise haberde o kelimenin kaç defa geçtiğini temsil etmektedir. Tf–idf yöntemi de denenmiştir fakat performansı BoW'e göre geride kaldığı için analize eklenmemiştir.
    - Türkçe Dolgu Sözcükleri (stop words) sık kullanılan, fakat metinden çıkarıldıklarında cümlenin anlamında önemli değişiklikler oluşturmayan sözcüklerdir. Necmettin Çarkacı'nın GitHub hesabında paylaşılan dolgu sözcükleri listesine göre vektöre dönüştürülme esnasında metinlerden bu sözcükler çıkarılmıştır. (link)

Sınıflandırma Modelinin Oluşturulması

- Model olarak Çokterimli Naive Bayes (Multinomial Naive Bayes) sınıflandırıcı kullanılmıştır.
  - Çeşitli parametreler ile çapraz geçerleme (cross validation) yapılmış ve en uygun (optimize) şekle getirilmiştir. [%70 train (eğitim), %30 test (deney)]
  - SVM, Random Forests, XGBoost, Yapay sinir ağları gibi daha karmaşık algoritmalar da denenmiştir fakat performans olarak Çokterimli Naive Bayes'e göre geride kaldıkları için analize eklenmemişlerdir.

Sonuçlar ve Analiz

- Gündem, Dünya, Cumartesi ve Pazar kategorilerindeki haberler genel konular hakkında olduğu için, Ege kategorisindeki haberler yerel haberler oldukları için analizin dışında tutulmuşlardır.
- 4-kategori ve 5-kategori içeren iki tür veri üzerinden model oluşturulmuştur. Farklı modeller haber başlığı, özeti ve metni üzerine eğitilmiştir. (Örneğin başlık üzerine eğitilen modelde eğitim ve sınıflandırma esnasında sadece haber başlıkları kullanılmıştır)
- 4-kategori: Ekonomi, Siyaset, Spor, Teknoloji-Bilim
- 5-kategori: Ekonomi, Siyaset, Spor, Teknoloji-Bilim, Diğer (Kültür-Sanat, Magazin, Yaşam)

	Başlık	Özet	Haber Metni
4-Kategori	%71.3	%82.2	%85.5
5-Kategori	%66.1	%79.3	%80.9

Tablo: Veri türüne ve kategori sayısına göre sınıflandırma performansları (doğruluk)

Sonuçlar ve Analiz

- Sonuçlara detaylı olarak GitHub proje sayfasından erişebilirsiniz.
- Spor ve Teknoloji-Bilim haberleri kullanılan veri türünden bağımsız olarak (başlık, özet, haber metni) %90 civarı oranda doğru tahmin edilmiştir.
- Siyaset ve Ekonomi haberleri birbiri ile karışmaktadır.

	precision	recall	f1-score	support
Diğer	0.70	0.79	0.74	1735
Ekonomi	0.80	0.71	0.75	2791
Siyaset	0.78	0.80	0.79	3633
Spor	0.91	0.91	0.91	2675
Teknoloji_Bilim	0.92	0.93	0.92	776
accuracy			0.81	11610
macro avg	0.82	0.83	0.82	11610
weighted avg	0.81	0.81	0.81	11610

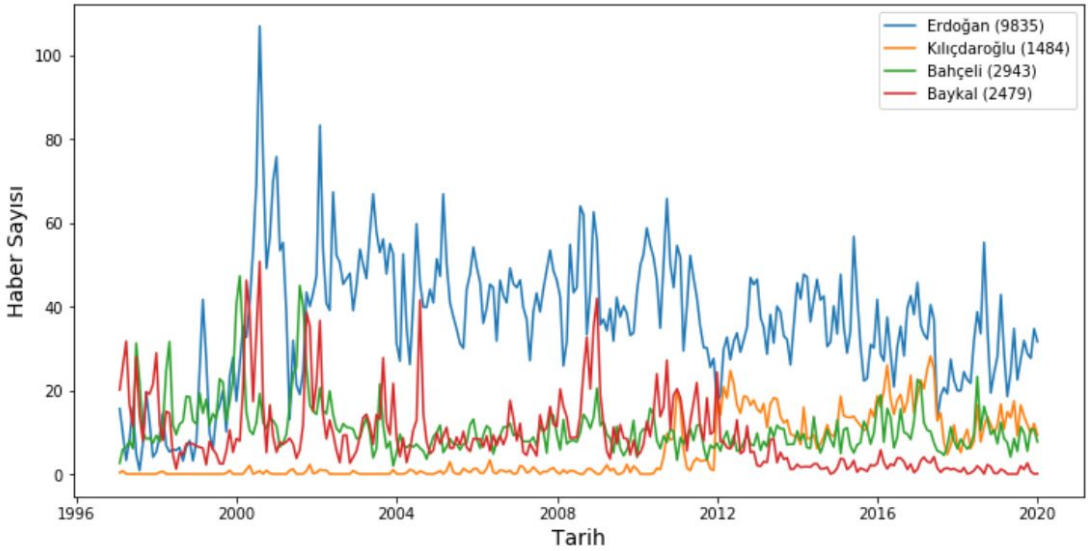
Tablo: 5-kategori ve haber metni üzerine geliştirilen modelin sınıflandırma raporu

Kategorilere göre en çok geçen kelimeler

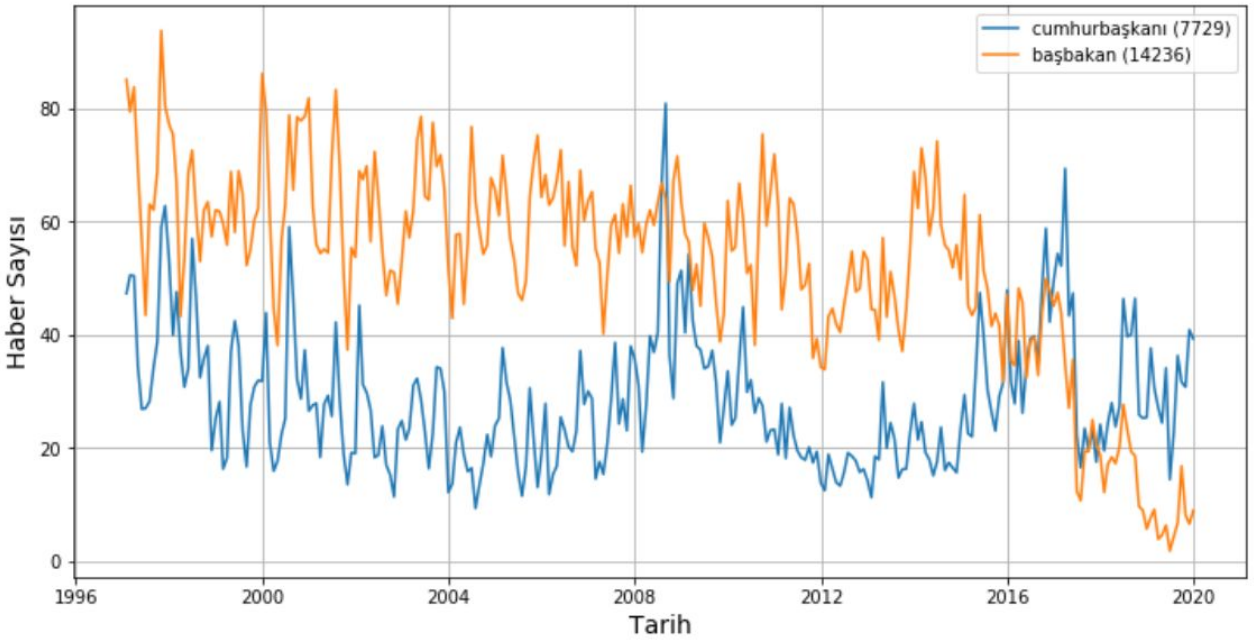
- **Ekonomi:** 'önemli', 'türk', 'dolar', 'yıl', 'yüzde'
- **Siyaset:** 'parti', 'başbakan', 'başkanı', 'erdoğan', 'chp'
- **Spor:** 'takım', 'futbol', 'beşiktaş', 'galatasaray', 'fenerbahçe'
- **Teknoloji_Bilim:** 'akıllı', 'samsung', 'galaxy', 'apple', 'yeni'

Veri Üzerinde Keşif

- Haberlerin paylaşıldığı tarih bilgisini kullanarak kelimelerin aya ve yıla göre haberlerde bulunma durumlarını inceleyebiliriz. Bu sayede kişilerin, kurumların, vs. 1997-2019 yılları arasında medyada ne kadar yer kapladıkları gözlemlenebilmekte ve bunun üzerine çıkarımlar yapılabilmektedir.

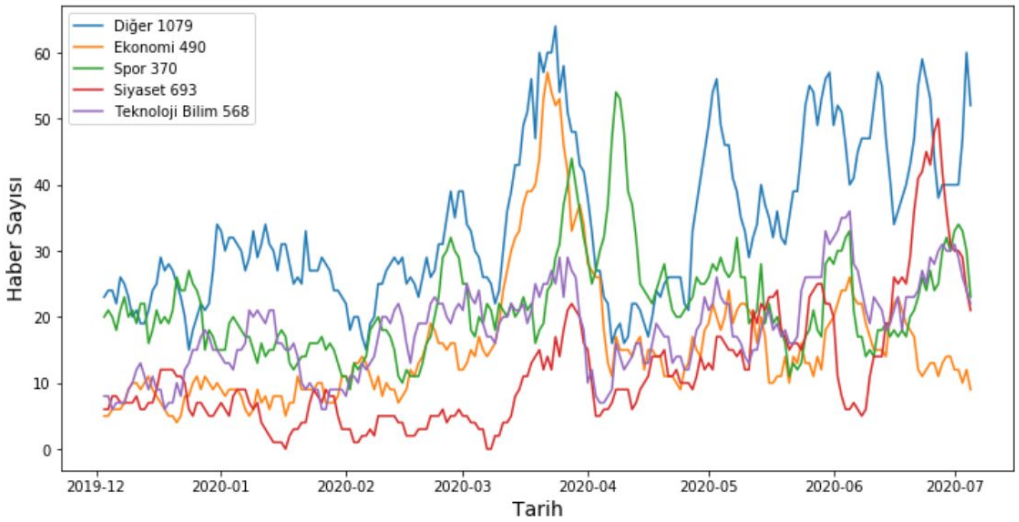


Veri Üzerinde Keşif



Twitter Gündem Takibi

- Twitter'da @nedenttoldu gibi birçok Twitter gündemi anlık olarak yazan hesaplar bulunmaktadır.
- Bu tarz hesapların incelenmesi ile Twitter gündemini gün ve ay olmak üzere inceleyebiliriz.
- 11 Temmuz 2020 ile 3 Aralık 2019 tarihleri arasında @nedenttoldu hesabının attığı bütün tweetler çekilmiştir.



Otomatik Özetleme (Ekstra)

- Proje kapsamına ek olarak Text Rank algoritması kullanılarak metin özetleyici ve önemli noktaları bulan bir model geliştirilmiştir.
    - Metinlerde çıkarım-bazlı özetleme (extractive summarization) günümüzde dikkat çeken konulardan biri olmuştur. Açık Hack yarışmalarında (2019-2020) bu konu ile ilgili birden fazla proje gözlemlenebilmektedir.
- Detaylar için proje GitHub sayfasına bakabilirsiniz.

Demo

- Link

Kapanış

- Proje kapsamıda açık kaynak paylaşılanlar (link):
  - 3 farklı haber derlemi paylaşılmıştır. (116.068 örnek)
    - **milliyet_derlem.csv.gz:** Haberler milliyet.com'dan çekildiği gibi saklanmıştır.
    - **temizlenmis_derlerm.csv.gz**: Bazı kategorilerde bulunan haberler atılmıştır. Benzer kategorideki haber türleri birleştirilmiştir.
    - **filtrelenmis_temizlenmis_derlem.csv.gz:** Ön-işleme sonucu filtrelenmiş haberleri içerir.
  - 4 ve 5 kategorili sınıflandırıcı.
  - Haberler üzerinde keşif için yardımcı program.
  - Otomatik özetleme, önemli noktalar için yardımcı program.