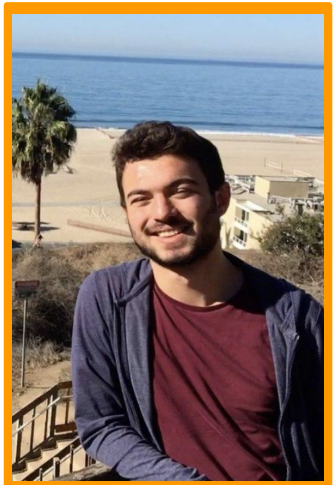
MANDALİNA

# Türkçe Duygu Analizi: Takip Sistemi

# Ekibimiz

**Abdullatif Köksal**

Araştırmacı



**Erkam Şeker**

Araştırmacı

Technical University of Munich

**Alperen Yakut**

Yazılım Mühendisi

Google

**Melih Mutlu**

Yazılım Mühendisi

Microsoft

# İş Dağılımı

**Tahminleyici:**

**Abdullatif**
*Model oluşturma/eğitme*

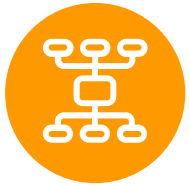
**Alperen**
*Veri seti oluşturma/işaretleme*

**Platform:**

**Erkam**
*Ön yüz geliştirme*

**Melih**
*Sunucu uygulaması geliştirme*

# Önceki Projelerimiz & Deneyimlerimiz



## Türkçe Doğal Dil İşleme

- ➔ Bağlılık Ayrıştırıcısı
- ➔ Duygu Analizi
- ➔ Word2Vec
- ➔ Lemmatizer
- ➔ İlişki Sınıflandırma

## Hackathon Dereceleri

- ➔ Kuveyt Türk · 3.lük
- ➔ Garanti · Jüri Özel
- ➔ Hürriyet · 2.lik

# Ne Yaptık?

Proje ve Problem Tanımı

# Türkçe Duygu Analizi

**Duygu Analizi**, verilen bir yazı verisinin içeriğinin anlamsal olarak pozitif mi yoksa negatif mi olduğunu tespit etme işlemidir.

**Problemimiz** Türkçe veriler üzerinde, güncel ve özgün yöntemler aracılığıyla işlevsel bir duygu analizi sistemi oluşturabilmek.

**Projemizin** hedefi Türkçe duygu analizlerinin yapılıp, sonuçlarının kullanıcıyla paylaşılacağı bir ortam oluşturmak.

Kısaca gelişim süreci basamaklarımız:

- ➔ Türkçe bir  veri setinin oluşturulması
- ➔ Modelin yaratılması ve eğitilmesi
- ➔ Son kullanıcıya sunulabilecek platformun geliştirilmesi

# Sentmon

## Duygu Analiz Sistemi

# Sentmon Neler Yapabilir?

## Duygusal Yönelim Takibi

Belli bir zaman aralığı içerisindeki tweetler üzerinden duygusal değişimi takip etmenizi sağlar.

## Özelleştirilen Sorgular

Belirlediğiniz sorgular hakkında duygu analizinde bulunur.

## Detaylı Analiz

Zamana yayılmış duygu analizi sonuçlarını ve örnek tweetleri size sunar.

# Nasıl Yaptık?

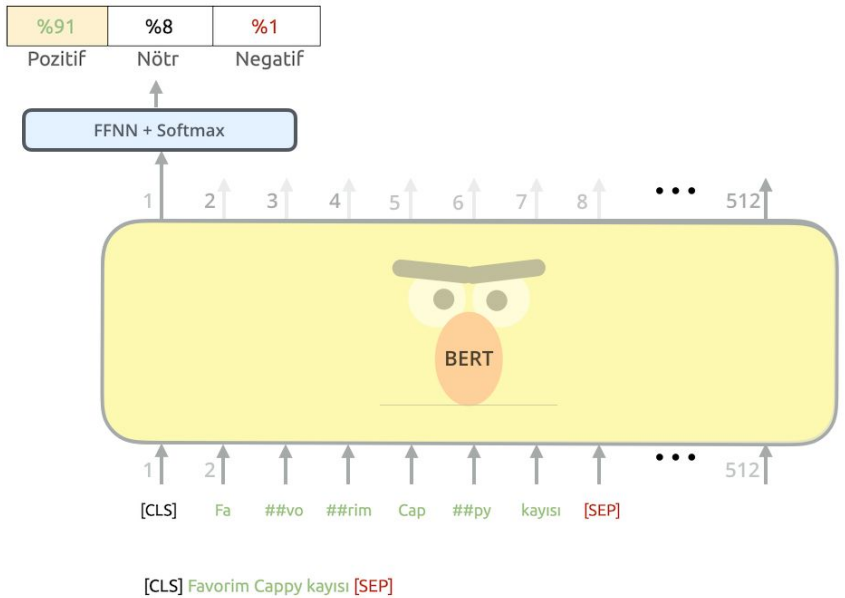
Kullandığımız Yöntemler ve Detayları

# BERT Tabanlı Çözüm

BERT gibi transformers temelli modeller son zamanlarda çeşitli NLP problemlerinde başarılar gösterdi.

Bu sebeple transformer temelli bir çözüm arayışı öneriyoruz.

- **Multilingual BERT (MBERT)**: Google tarafından 104 dil için üretilmiş bir model.
- **BerTurk**: 35 GB'lık Türkçe kaynaklardan eğitilmiş MBERT ile aynı büyüklüğe sahip bir model.
- **Damıtılmış (Distilled) BerTurk:** Distillation [1] tekniği ile BerTurk'un başarımı korunarak küçültülmüş hali.





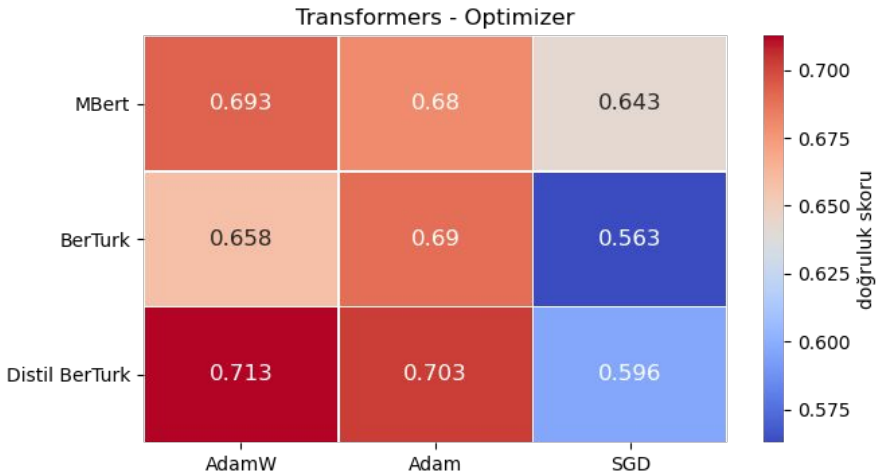
1. Victor Sanh, Lysandre Debut, Julien Chaumond, Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108.

# BERT Tabanlı Çözüm

Eğitim ve test verisi olarak BOUN 2018 Twitter Verisini kullandık.

**İnce ayar** (fine-tuning) için çeşitli modellerle farklı optimizerları, weight decay ve learning rate oranlarını kıyasladık.

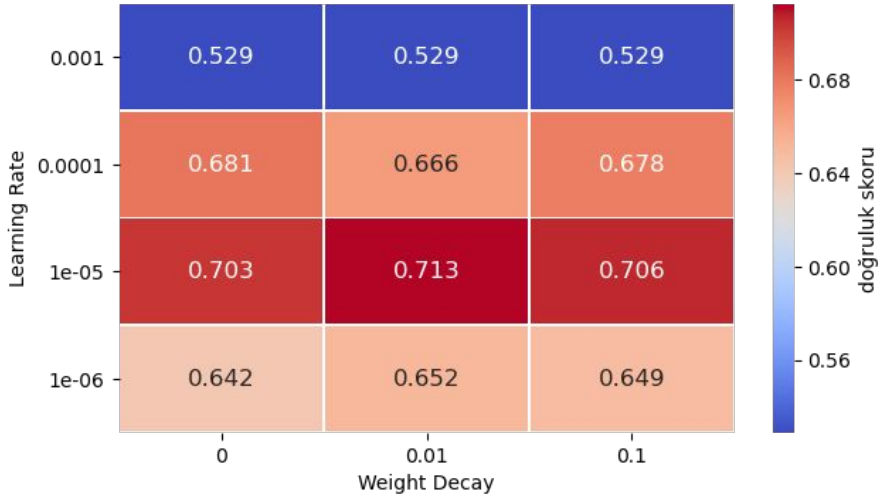
BERT'lerde oluşan varyans probleminin önüne geçmek için de her bir değer için 3 farklı çalıştırmanın ortalamasını aldık.



# BERT Tabanlı Çözüm

En iyi model'in detaylı analizi.

- Transformer: **Distil BerTurk**
- Optimizer: **AdamW**
- Learning Rate: **0.00001**
- Weight Decay: **0.01**
- Ön işlem: **Url Silmek**

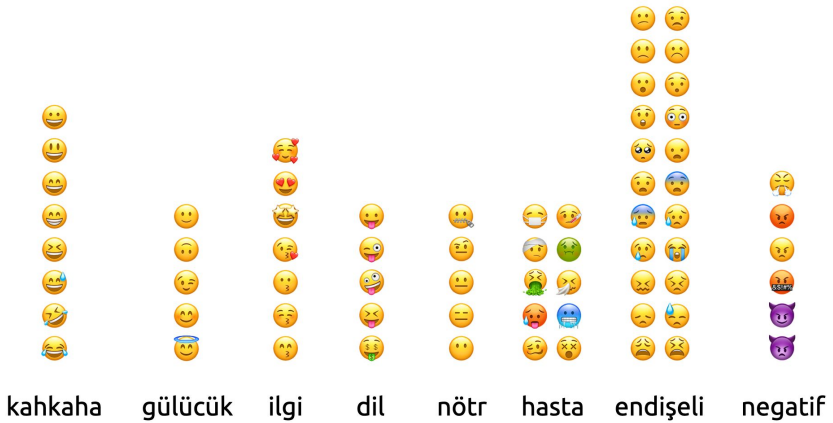


# Mandalina Emoji Veri Seti

Transformers tabanlı modellerin çok büyük verilerle ön eğitiminin (pretraining) genel başarıdaki artışı çeşitli çalışmalarda gösterilmiştir.

Bu sebeple Türkçe **Mandalina Emoji Veri Seti**'ni sunuyoruz:

- **767.197** twit
- **67** emoji
- **8** kategori

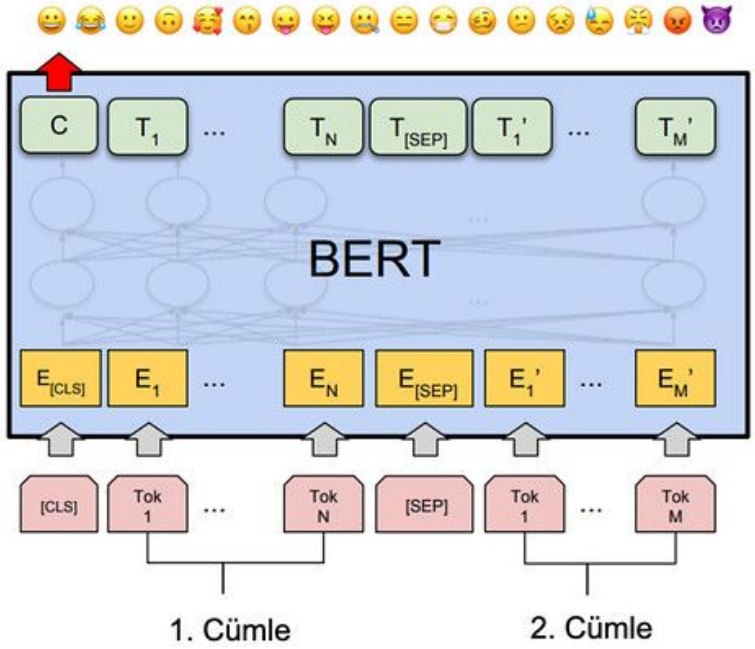


# Öneğitim*

Distil BerTurk, emoji verisi ile iki farklı şekilde öneğitime sokuldu. Bu öneğitimlerin ikisinde de BERT modelininde **maskelenmiş dil modeli de eğitildi**.

**Öneğitim #1 - Sınıflandırma**
Mandalina veri setindeki twitlerin emojileri silindikten sonra içerdiği emojinin kategorisinin tahmini.

**Öneğitim #2 - Eşli Sınıflandırma**
Mandalina veri setinden emojileri silinmiş, rastgele seçilmiş iki tane twitin aynı kategoriye sahip olup olmadığının tahmini. [2]




2. Soares et. al. 2019. Matching the Blanks: Distributional Similarity for Relation Learning. arXiv preprint arXiv:1906.03158.
* Bu iki öneğitimde de istatistiksel olarak önemli bir artış elde edilemedi. Fakat gözle görülen artış sebebiyle notebooklar yorum da eklenerek paylaşıldı.

# Model Hızlandırma

Halihazırda Distil BerTurk kullandığımız için model diğer BERT modellerine göre hızlı çalışıyor.

Hız konusunda yapacağımız tüm geliştirmeler çekeceğimiz tweet sayısının artmasına ve doğal olarak daha iyi analiz etmemize sebebiyet verecektir.

**ONNX** kullanarak en iyi modelimizin başarısını koruyarak %45'e varan hız artışı elde ettik.


| | 1000 Tweet için Duygu Analizi Hızı |
|---|---|
| PyTorch | 99.86 sn. |
| PyTorch No Grad | 79.65 sn. |
| ONNX Çıkarım Modu | **55.73 sn.** |

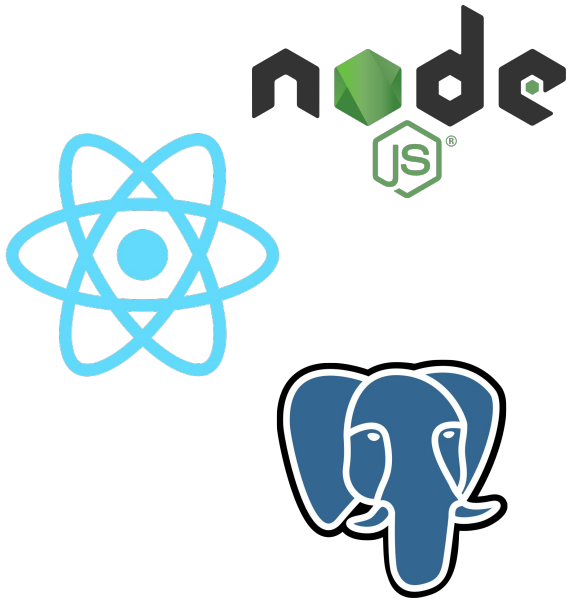
# Sentmon Platform

Modelimizin son kullanıcıya sunulması amacıyla bir platform geliştirdik.

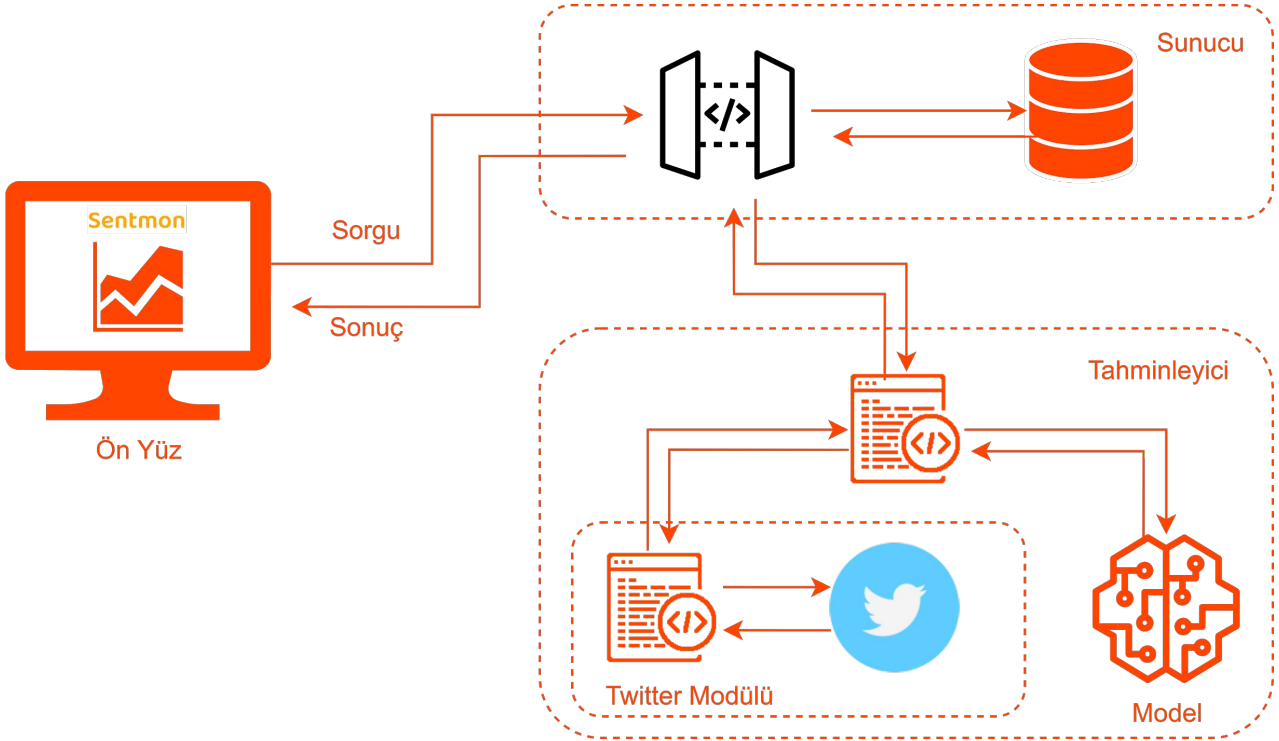
Sunucu tarafında **Nodejs** ve **Postgresql** tabanlı uygulamamız;

- ➔ verilen sorguyla ilgili tweetleri çeker,
- ➔ bu verilerle modelimizi çalıştırır,
- ➔ sonuçları veritabanına kaydeder.

Sorgu girişi ve sonuçların gösterilmesi için de **React** kullanarak bir kullanıcı arayüzü oluşturduk.



# Sistem İş Akışı



# Neler Çıktı?

Sonuçlar ve Demo

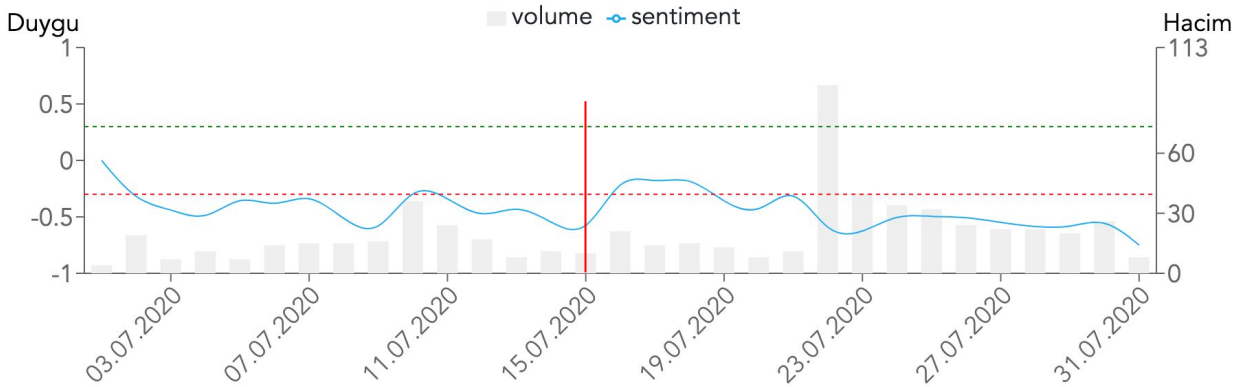
# Demo: Video

https://drive.google.com/file/d/1ZeNRqsaEq3_aWTphKKVqbi9bxZ3IfCUK/view

# Demo: Cappy

Cappy canavarı henüz yavruyken karşımıza çıktı, bir kısmını da döktük. Bugüne kadar inanmazdım taa ki karşılaşana kadar. İnsanlara neler içiriyorsunuz, tebrikler.

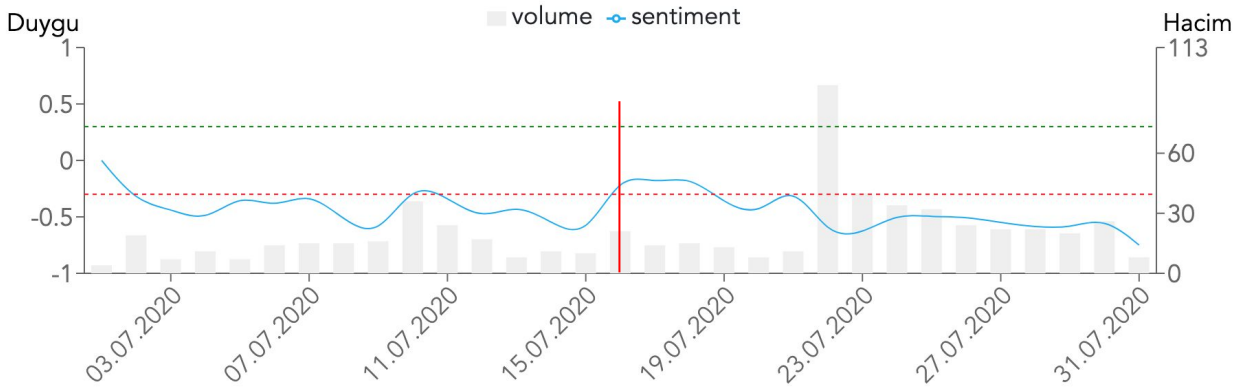
20



# Demo: Cappy

1 tabak noodle bi bardak cappy portakal ziyade olsun

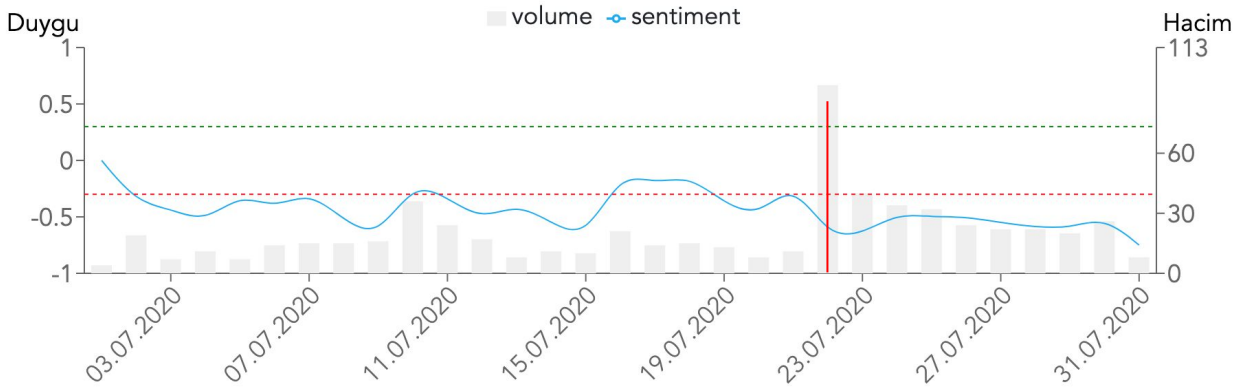
14



# Demo: Cappy

almayın işte artık şunu ya ben almıyorum mesela her sene yeni bir cappy canavarı görüyoruz

94



# Çıktılar

1. Duygu analizi için **geniş kapsamlı** bir **ince ayar** çalışması ve detaylı incelenmesi.
2. **Büyük ölçekli Türkçe emoji veri seti**nin sunulması, alakalı öneğitim notebooklarının dökümantasyonu ve paylaşılması.
3. Transformer tabanlı, damıtılmış ve **ONNX'e çevrilmiş model**in ve bu model üzerinden **duygu analizi için gerekli kodlar**ın paylaşılması.
4. Markalar, kurumlar ve ünlüler hakkında analizlerin yapıldığı ve tweetlerin anlamlandırıldığı **açık kaynaklı ve Türkçe ilk duygu analizi takip sistemi** sitesi:

**Sentmon**

**Not:** *Bu çalışmanın cross-lingual şeklinde uzatılmış hali **CoLING 2020 - Peoples** workshop'ına çalışma olarak gönderilecektir.*