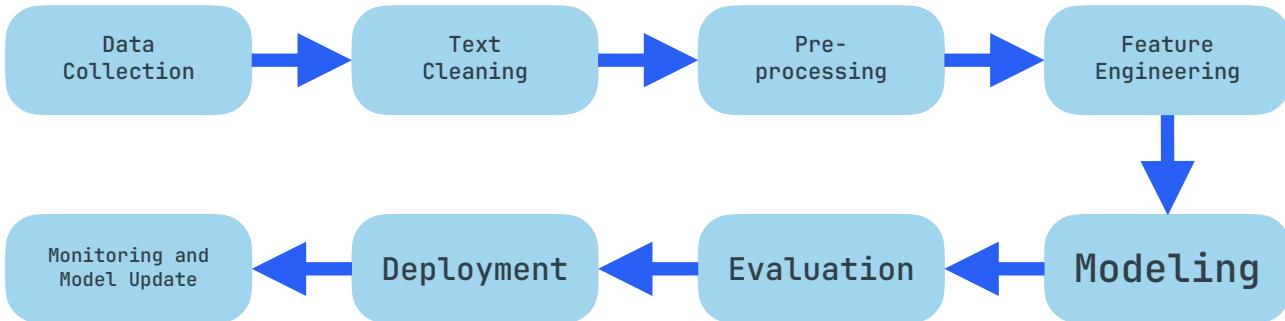


DOĞAL DİL İŞLEME (NLP) NOTLARI

Doğal Dil İşleme Akış Şeması

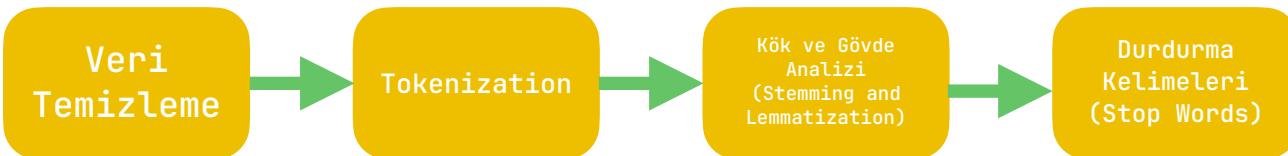


Metin Ön İşleme

→ Ham metin verilerini analiz ve işlem için daha uygun bir formata dönüştürme sürecidir.

Neden Yapılmalıdır?

- Veri kalitesinin artırır
- Model performansını artırır
- Veri boyutunu yönetir



Veri Temizleme

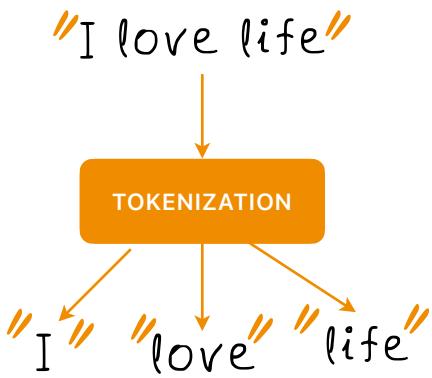
→ Veri temizleme, metin verilerini analiz edilebilir hale getirmek için yapılan bir dizi işlemi içerir.

→ Bu adımda, metinlerde bulunan hatalı, gereksiz, düzensiz veya model performansını olumsuz etkileyebilecek öğeler kaldırılır veya düzelttilir.

→ Veri Temizleme Adımları:

- * Boşlukların Temizlenmesi
- * Büyük-Küçük Harf Dönüşümleri
- * Noktalama işaretlerinin Kaldırılması
- * Özel Karakterlerin Kaldırılması
- * Yazım Hatalarının Düzeltilmesi
- * HTML ve URL Temizleme

Tokenizasyon (Tokenization)



→ Tokenization, bir metni daha küçük parçalara ayırma işlemidir.

→ Bu küçük parçalar genellikle "token" olarak adlandırılır.

→ Tokenler, kelimeler, cümleler veya hatta karakterler olabilir.

Kök ve Gövde Analizi

→ Stemming (Kök Bulma)

* Stemming, kelimelerin kök formunu (yani temel anlamını) bulmak için kelimenin sonundaki eklerin (suffix) çıkarılması işlemidir.

* Stemming işlemi, kelimenin anlamını tamamen doğru bir şekilde elde etmeyi amaçlamaz; daha ziyade, kelimenin en basit formunu bulmaya odaklanır.

* Örnek:

- "koşuyor", "koştı", "koşmak" → "koş"
- "evde", "evler", "evimiz" → "ev"

→ Lemmatization (Gövdeleme)

* Lemmatization, kelimeleri sözlükteki temel formlarına (lemma) dönüştürme işlemidir.

* Lemmatization, kelimenin anlamını ve dilbilgisel yapısını dikkate alarak doğru bir kök bulmaya çalışır.

* Bu nedenle, lemmatization sonrası elde edilen kelime dilbilgisel olarak anlamlı ve sözlükte yer alan bir kelime olur.

* Örnek:

- "koşuyor", "koştı", "koşmak" → "koşmak"
- "evde", "evler", "evimiz" → "ev"

Durdurma Kelimeleri (Stop Words)

→ Durdurma kelimeleri (stop words), metinlerde genellikle anlamı çok az olan veya metnin analizi sırasında çok faydalı olmayan kelimelerdir.

→ Bu kelimeler genellikle bağlaçlar, edatlar, zamirler ve diğer bilgisel işlevi olan kelimelerdir.

→ Metin işleme süreçlerinde bu kelimeleri kaldırmak, analizlerin doğruluğunu artırabilir ve metin üzerinde daha anlamlı sonuçlar elde edilmesine yardımcı olabilir.

→ Örnek:

- Türkçe Stop Words: ve, bir, bu, ile, da, de, mi, o, çok, gibi
- İngilizce StopWords: and, the, is, in, to, of, it, that

Metin Temsili

→ Metin temsili, bir metni sayısal veya başka türde bir formatta temsil etme işlemidir.

Neden Yapılmalıdır?

- Bilgisayarların anlayabilmesi,
- Öz nitelik çıkartma,
- Model Eğitimi

Yöntemleri?

- Bag of Words (BoW)
- TF-IDF (Term Frequency-Inverse Document Frequency)
- N-Gram Modelleri
- Word Embeddings
- Transformers Tabanlı Metin Temsili

Bag of Words Nedir?

→ Bag of Words (Bow), doğal dil işleme (NLP) ve metin madenciliğinde kullanılan temel bir metin temsili yöntemidir.

→ BoW, metinlerdeki kelimeleri sayısal verilere dönüştürür ve metinlerin analizini sağlar.

BoW Yönteminin İşleyışı

→ Kelime kümesi oluşturma

Metin: Kedi evde

Kelime Kümesi: ["Kedi", "evde", "bahçede"]

Bugün güzel bir gün.

→ Kelime frekansı hesaplama

Metin1: "Kedi evde"

"Kedi": 1

"evde": 1

"bahçede": 0

Bugün güneşli bir gün.

→ Vektör temsili

Metin1: [1, 1, 0]



TF-IDF (Term Frequency-Inverse Document Frequency)

→ TF-IDF, metin madenciliğinde ve bilgi erişiminde sıkça kullanılan bir özellik çıkarım yöntemidir.

→ TF-IDF, kelimelerin belgeler içinde ne kadar önemli olduğunu belirlemek için kullanılır.

• **Term Frequency (TF):** Bir kelimenin bir belgede ne kadar sık geçtiğini ölçer.

• **Inverse Document Frequency (IDF):** Bir kelimenin tüm belgelerdeki yaygınlığını ölçer. Bir kelime çok belgede geçiyorsa, o kelime çok fazla bilgi saklamaz.

$$TF(t, d) = \frac{\text{Sayac}(t, d)}{\text{Toplam Kelime Sayısı}(d)}$$

Burada, Sayac(t, d) kelimenin t bir belgede d sayısıdır.

$$IDF(t, D) = \log \left(\frac{\text{Toplam Belgeler Sayısı}(D)}{1 + \text{Belgelerde Geçen Sayısı}(t)} \right)$$

Burada, Belgelerde Geçen Sayısı(t) kelimenin t belgelerdeki sayısıdır.

$$\text{TF-IDF}(t, d, D) = \text{TF}(t, d) \times \text{IDF}(t, D)$$

N-Gram Modelleri

- Bir dil modelinde kullanılan kelime veya karakter dizisinin uzunluğunu belirten bir terimdir.
- N-Gram modelleri, metinleri n kelimelik veya n karakterlik kısımlara bölgerek analiz eder.



- "Bu bir örnek metindir"

→ Unigram (n=1)

- ['Bu', 'bir', 'örnek', 'metindir']

→ Bigram (n=2)

- ['Bu bir', 'bir örnek', 'örnek metindir']

→ Trigram (n=3)

- ['Bu bir örnek', 'bir örnek metindir']

Kullanım Alanları

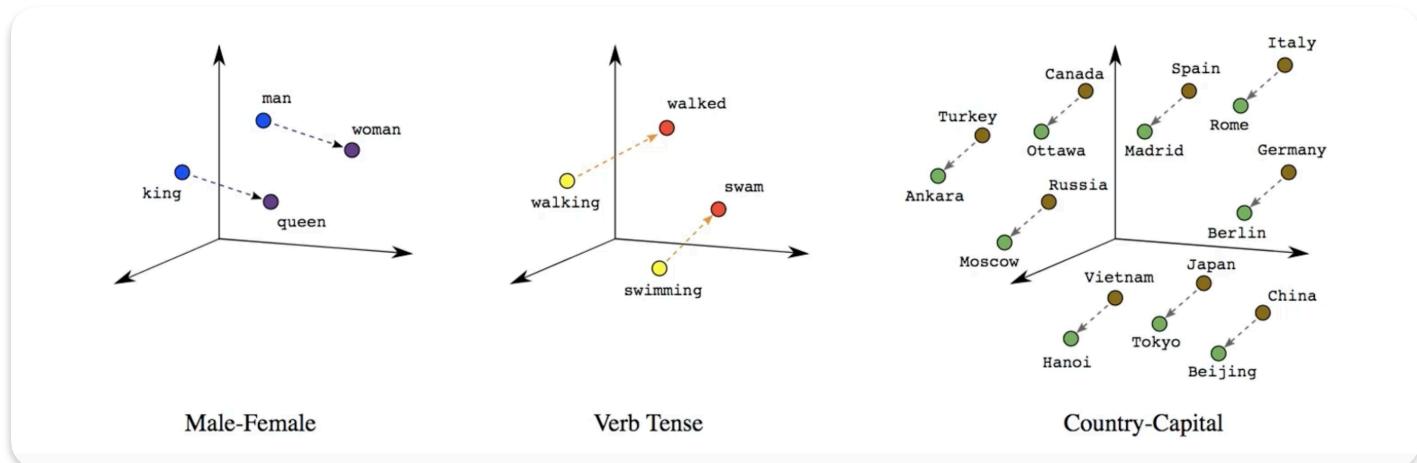
- Metin Modelleme
- Metin Sınıflandırma
- Metin Üretimi
- Metin Benzerliği

Word Embeddings

→ Word Embeddings (Kelime Gömme), doğal dil işleme (NLP) ve makine öğreniminde kullanılan bir tekniktir.

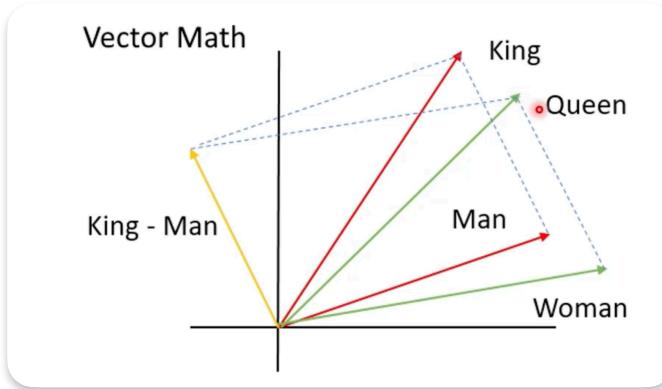
→ Kelimeleri, genellikle sürekli bir vektör uzayında anlamlı temsil edecek şekilde sayısal vektörlere dönüştürülür.

→ Bu temsiller, kelimeler arasındaki anlamsal ve dilbilgisel ilişkiler yakalamayı hedefler.



Word Embeddings Özellikleri

- Anlamsal benzerlik
 - * Örneğin, "king" ve "queen" kelimeleri benzer vektörler alabilir.
- Matematiksel İşlemler
 - * Örneğin, "king" - "man" + "woman" = "queen"
- kapsamlılık

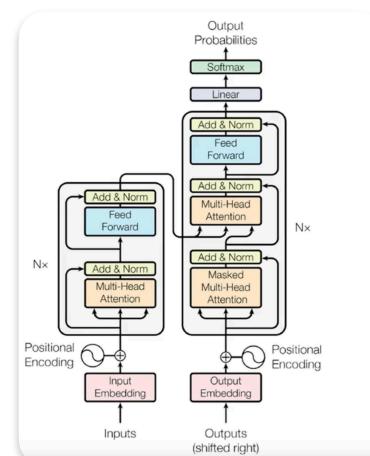


Word Embeddings Modelleri

- **Word2Vec**: Google tarafından geliştirilen, kelimeleri vektörlere dönüştüren ve bu vektörleri dildeki ilişkileri yakalayacak şekilde eğiten bir modeldir.
- **GloVe (Global Vectors for Word Representation)**: Stanford Üniversitesi tarafından geliştirilen, kelime gömme temsillerini kelime ortaklıklarını yakalayacak şekilde hesaplayan bir modeldir.
- **FastText**: Facebook tarafından geliştirilen ve kelime gömme temsillerini kelime alt-birimlerini de dikkate alarak hesaplayan bir modeldir.

Transformers Tabanlı Metin Temsili

- Transformers, doğal dil işleme (NLP) ve diğer yapay zeka alanlarında son yıllarda devrim niteliğinde yenilikler getiren bir mimaridir.
- İlk olarak 2017 yılında Google tarafından yayınlanan "Attention is All You Need" adlı makalede tanıtılmıştır.
- Neden Transformers
 - * Bağlamı daha iyi anlama
 - * Paralel işleme yeteneği
 - * Çeşitli NLP görevlerinde kullanım
 - * Önceden eğitilmiş modellerin yeniden kullanımı (Fine-Tuning)
- En bilindik transformers modelleri
 - * BERT (Bidirectional Encoder Representations from Transformers)
 - * GPT (Generative Pre-trained Transformers)



→ **Attention**, modelin bilirli girdi parçalarına farklı derecelerde dikkat göstermesine olanak tanır.

→ Özellikle, bir kelimenin diğer kelimelerle olan ilişkisini anlamak için kullanılır.

→ Örneğin, "Kedi hızlıdır" cümlesinde, "hızlıdır" kelimesi "Kedi" kelimesine olan dikkat skorlarını hesaplar:

- * **Sorgu (Query) ve Anahtar (Key) Çarpımı**: "Kedi" kelimesinin "hızlıdır" kelimesine ile olan ilişki skoru hesaplanır.

- * **Dikkat Skoru**: Bu skora göre "Kedi" kelimesinin temsilini günceller.

→ **Input Embedding**, girdi verilerini modelin işleyebilecegi bir formata dönüştürmek için kullanılan bir tekniktir.

→ **Örnek**: Bir cümle düşünelim: "Kedi hızlıdır."

→ **Kelime Vektörleri**: Bu cümledeki kelimeler, Word2Vec, GloVe, veya BERT gibi bir embedding tekniği kullanılarak sayısal vektörlere dönüştürülür.

- * "Kedi" → [0.21, -0.32, 0.87, ...]

- * "hızlıdır" → [-0.13, 0.45, -0.20, ...]

→ **Multi-Head Attention**, attention mekanizmasının birden fazla başlıkla (head) çalıştığı bir tekniktir.

→ Bir cümledeki her kelime, diğer kelimelerle olan ilişkilerini farklı açılardan öğrenmek isteyebilir.

→ Örneğin, "Kedi" kelimesinin "hızlıdır" kelimesiyle ilişkisini anlamak için birden fazla dikkat başlığı kullanır.

- * **Başlık 1**: "Kedi" ve "hızlıdır" arasındaki anlam ilişkisini öğrenir.

- * **Başlık 2**: "Kedi" ve "hızlıdır" arasındaki gramatik ilişkileri öğrenir.

- * **Başlık 3**: "Kedi" kelimesinin cümledeki konumunu öğrenir.

→ **Masked Multi-Head Attention**, modelin gelecekteki kelimeleri görmesini engeller, yani model sadece geçmiş bilgileri kullanarak tahminde bulunur.

→ **Örneğin**, "Kedi _____ hızlıdır" cümlesinde, model "Kedi" ve "hızlıdır" arasındaki ilişkilere dayanarak boşluğa "oldukça" kelimesini tahmin eder.

→ **Add & Norm**, bir katman çıktı ile girdi arasındaki kısa yolu (residual connection) ekleyip ardından layer normalization uygulayan bir adımdır.

→ **Feed-Forward Network**, her encoder ve decoder katmanında bulunan bir açıdır.

→ **Output Embedding**, modelin çıktısını temsil eden ve genellikle bir dil modelinde kullanılan bir tekniktir.

Metin Temsili Yöntemlerinin Karşılaştırılması

Yöntem	Temel Özellikler	Kullanım Kolaylığı	Sonuçların Başarı Durumu
Bag of Words (BoW)	Kelime freksanslarına dayalı, sıkılık matrisleri oluşturur.	Basit, doğrudan uygulanabilir.	Genellikle düşük, bağlam bilgisinden yoksundur.
TF-IDF	Kelime sıklığına ek olarak, kelimenin belgelerdeki önemini ölçer.	Kolay, standart kütüphaneler mevcut.	Orta, bağlam bilgisi kısıtlıdır ama bilgiye değer katar.
N-grams	Kelime ya da karakter n-gramlarını kullanarak bağlamı yakalar.	Orta, işlem gücü ve bellek kullanımı artar.	Orta, bağlam bilgisi artırılabilir ancak model karmaşıklığı artar.
Word Embeddings (GloVe, Word2Vec, FastText)	Kelimeleri vektörlere dönüştürür, anlam ilişkilerini yakalar.	Orta, önceden eğitilmiş modeller mevcut.	Yüksek, bağlamı daha iyi anlar, semantik ilişkiler sağlar.
Transformers (BERT, GPT-3, vb.)	Derin öğrenme temelli, bağlamı dikkat mekanizması ile yakalar.	Orta-ileri, genellikle yüksek hesaplama gücü gerektirir.	Çok yüksek, bağlamı derinlemesine anlar, çeşitli NLP görevlerinde başarılı.

Olasılıksal Dil Modelleri

→ olasılıksal dil modelleri, dilin yapısını ve düzenini anlamak ve modellemek için kullanılan istatistiksel yöntemlerdir.

- N-Gram Modelleri
- Gizli Markov Modelleri (Hidden Markov Models - HMM)
- Maximum Entropy Modelleri (MaxEnt)

N-Gram Modelleri

→ N-Gram modelleri, bir dizideki (genellikle bir cümledeki) ardışık kelime veya karakter gruplarının olasılıksal tahmin eder.

→ Burada "n", bu gruptaki öge sayısını belirtir.

→ Örneğin, bir **unigram** modelinde tek kelimeler, **bigram** modelinde iki kelimelik gruplar, **trigram** modelinde ise üç kelimelik gruplar dikkate alınır.

• Kullanım

- N-Gram'lar, bir cümlede bir kelimeden hangi olasılıkla diğer bir kelimeden sonra geleceğini tahmin etmek için kullanılır.

• Avantajlar

- Basit ve hızlı
- Yerel bağlantıları iyi

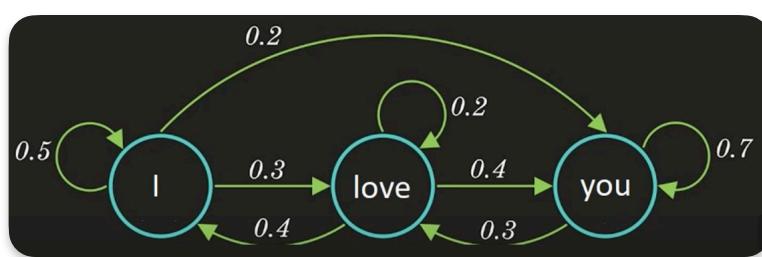
• Dezavantajlar

- Bağlam Sınırımları
- Veri Gereksinimi

Gizli Markov Modelleri (Hidden Markov Models)

→ Gizli Markov Modelleri, bir dizi gözlemin (kelimeler veya karakterler gibi) arkasındaki gizli bir durum dizisinin (örneğin, dil bilgisel kategoriler) olduğu varsayıma dayanır.

→ HMM, her bir durumun belirli bir olasılıkla başka bir duruma geçeceğini ve her durumun belirli bir gözlemi üreteceğini varsayar.



• Kullanım

- Konuşma tanıma
- Dil modellem
- Parça etiketleme (Part-of-Speech Tagging)

• Avantajlar

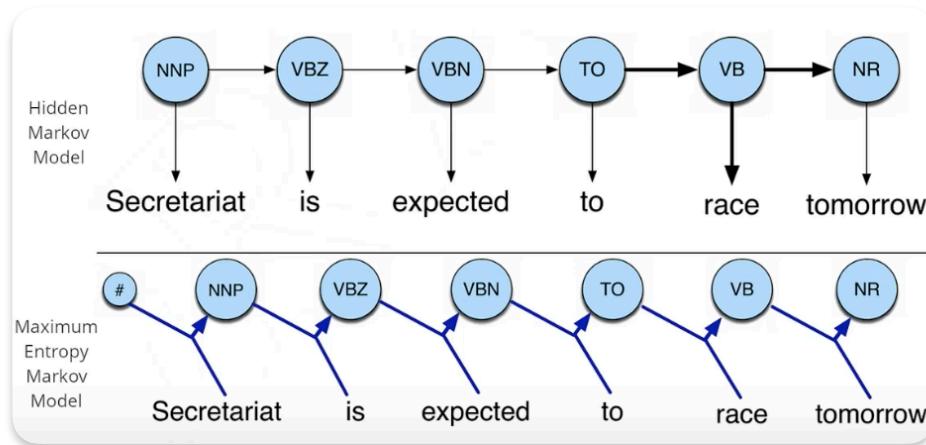
- Bağlam Modelleme
- Verimli Algoritmalar

• Dezavantajlar

- Basitleştirici Varsayımlar
- Eğitim Zorluğu

Maximum Entropy Modelleri (MaxEnt)

→ Maximum Entropy (Maksimum Entropi) Modelleri, bir olasılık dağılımını tahmin ederken mümkün olduğunda az varsayımda bulunmayı hedefler.



→ Kullanım

- * MaxEnt modelleri, özellikle sınıflandırma görevlerinde kullanılır.
- * Örneğin, bir cümlenin belirli bir sınıfı (pozitif, negatif duygusal gibi) ait olma olasılığını tahmin edebilir.

→ Avantajlar

- * Esneklik
- * iyi Genelleme

→ Dezavantajlar

- * Hesaplama Maliyetleri
- * Özellik Mühendisliği

Derin Öğrenme Tabanlı Modern Dil Modelleri

Word Embeddings (Kelimelerin Vektör Temsili)

→ Neden kelimeleri vektörlerle temsil etmeliyiz?

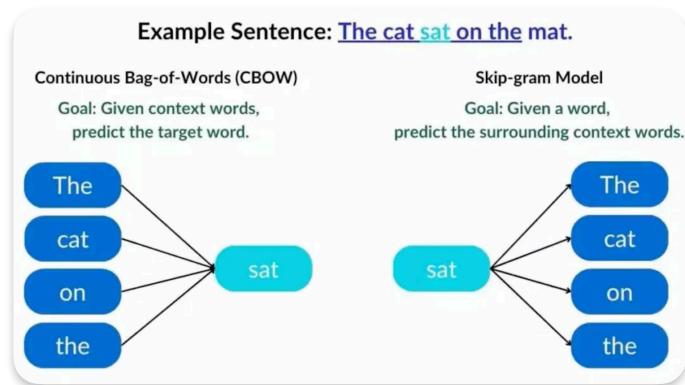
- Dilsel Anlamı Yakalama
- Matematiksel İşlemler
- Verimli Temsil

→ Word2Vec Nedir?

- Özellikleri ve Avantajları
 - Anlamsal yakınlıklarını kelime komşuluklarına göre öğrenir.
 - Kelimeler arasında anlamlı matematiksel işlemler yapmayı sağlar.
- Kim Gelişirdi?
 - Google, Tomas Mikolov ve ekibi.
- Kullanılan Veri Seti: Google News corpus.
 - Kaynak: Haber metinleri
 - Büyüklük: Yaklaşık 100 milyar kelime
 - Kapsam: Siyaset, ekonomi, spor, teknoloji, vb.

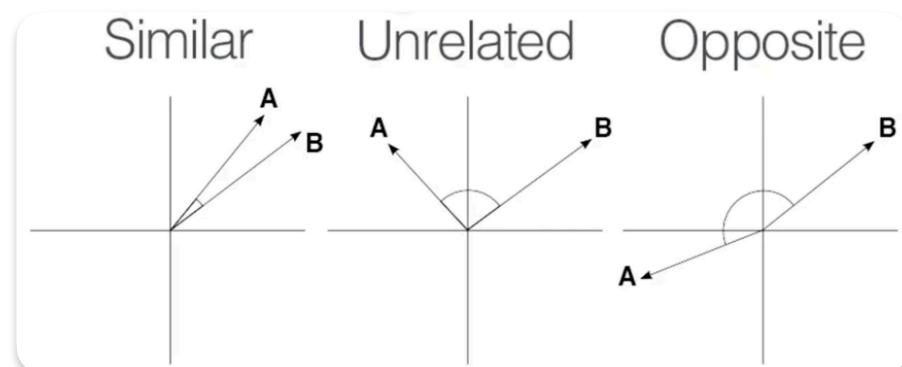
→ Word2Vec Nedir?

- Temel Modelleri: CBOW (Continuous Bag of Words) ve Skip-Gram.



→ Vektör Uzayı ve Anlam Yakınlığı Nedir?

- Kelimelerin vektör temsilleri, anlamlarına göre vektör uzayında konumlandırılır.
- Anlamca benzer kelimeler (örneğin, "kedi" ve "köpek") vektör uzayında birbirine yakın olurken, farklı anlamdaki (örneğin, "kedi" ve "araba") uzakta yer alır.
- Yakınlık ve uzaklık arımı Kosinüs Benzerliği ile yapılır.



→ Benzerliklerin Görselleştirilmesi (t-SNE, PCA)



Recurrent Neural Network (RNN)

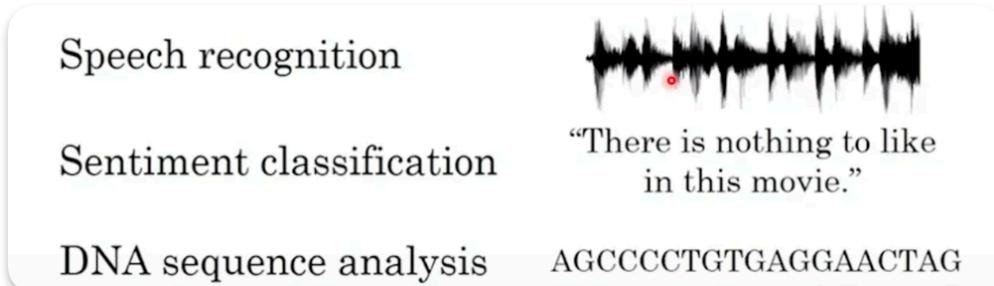
→ Zaman Serisi ve Sekans Verisi Nedir?

- **Zaman Serisi Verisi**

- Zaman içerisinde ardışık olarak kaydedilen verilerdir.
- Örneğin; finans, iklim, sensör, vb.

- **Sekans Verisi (Sıralı Veri)**

- Doğal Dil
- DNA Dizileri

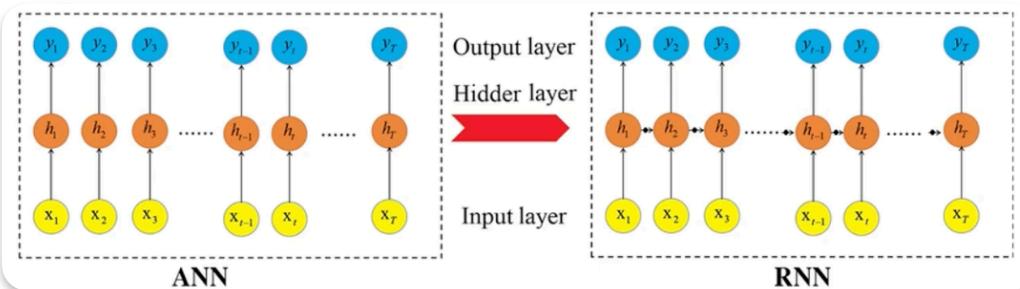


→ Dil ve Zaman Serisi Verilerinde Dizisel Bağımlılık Nedir?

- **Dizisel Bağımlılık:** Zaman serisi ve dil verilerinde, her bir öğe (kelime veya veri noktası) sırasıyla önceki öğeler bağımlıdır.
- **Doğal Dil Örneği:** Cümlede bir kelimenin anlamı, önceki kelimelerle ilişkili olabilir. Örneğin, "Ben kahve içiyorum." cümlesiinde "icıyorum" kelimesinin anlamı, önceki kelimelerden etkilenir.
- **Zaman Serisi Örneği:** Hava sıcaklığının yarın ne olacağı, bugünkü ve önceki günlerin sıcaklığına bağlı olabilir.

→ Standart Sinir Ağları (Örneğin, ANN) Sekans Verilerinde Neden Yetersizdir?

- Sabit Girdi/Çıktı
- Zaman Bağımlılığı
- Geçmiş Bilgiyi Kaydetme



→ RNN'ler, sekans verilerini işlemek için özel olarak tasarlanmış sinir ağlarıdır.

→ Her zaman adımda, önceki zaman adımdındaki bilgiyi saklayarak ve sonraki adımlarda bu bilgiyi güncelliyorlar.

→ RNN'in Temel Özellikleri

- Zaman Boyutunda Tekrar
- Sekans Verisi için Uygun

RNN Mimarisi Nedir ve Nasıl Çalışır?

- Ağ Yapısı
- Zaman Boyutunda Tekrar Yapısı

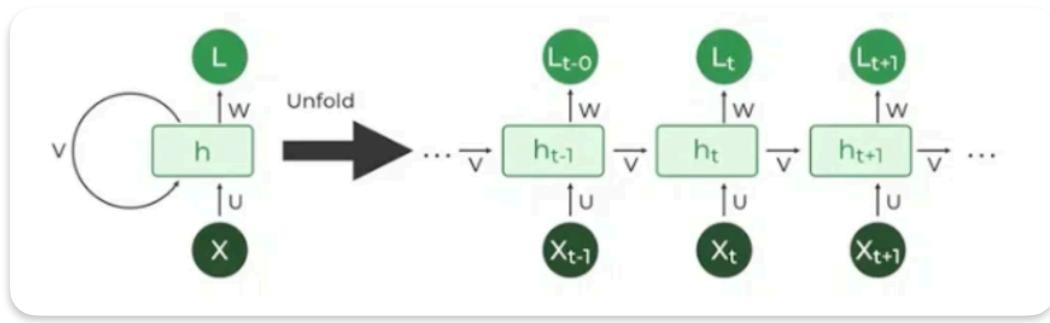
Adımlar:

1. **Girdi:** Dizideki bir öğe (örneğin, bir kelime).
2. **Gizli Durum:** Önceki adımda üretilen bilgi.
3. **Çıktı:** Girdi ve gizli duruma dayanarak üretilen çıktı.

Gizli Durum Denklemi:

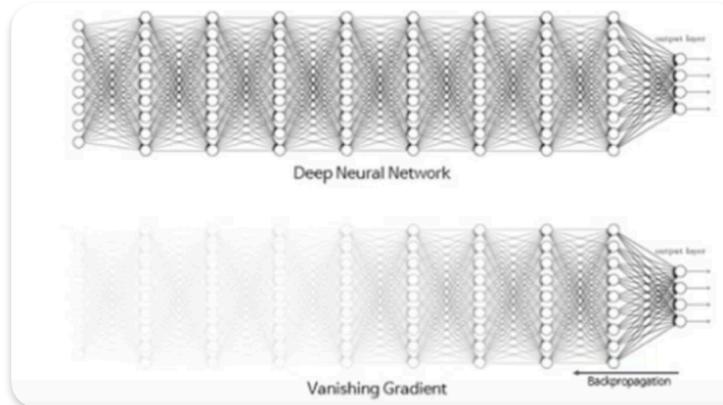
- $h_t = f(W \cdot x_t + U \cdot h_{t-1} + b)$
- $y_t = g(V \cdot h_t)$

h_t : Şu andaki gizli durum
 x_t : Şu andaki giriş
 h_{t-1} : Önceki gizli durum
 y_t : Çıktı



Vanishing Gradient Sorunu Nedir?

- Vanishing Gradient Sorunu, RNN'lerde eğitim sırasında ortaya çıkan bir problemdir.
- Geriye dönük hata yayılımı (backpropagation) sırasında, gradyanlar çok küçük hale gelir ve bu, uzun süreli bağımlılıkların öğrenilmesini zorlaştırır.
- Neden Oluşur?
 - Her zaman adımda zincirleme türev alınır.
 - Derin ağlarda bu türevler zaman içinde küçülebilir ve neredeyse sıfıra yaklaşır.
 - Bu durumda, önceki adımlardaki bilginin etkisi kaybolur.
- Sonuç
 - Kısa dönem bağımlılıkları öğrenir, ancak uzun dönem bağımlılıkları öğrenmekte zorlanır.

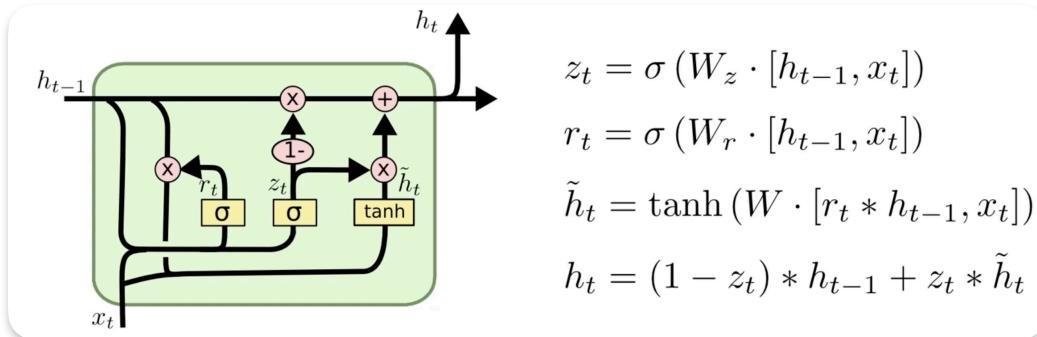


RNN ile NLP Alanında Yapılan Uygulamalar

- Dil Modelleme
- Makine Çevirisi
- Duygu Analizi
- Konuşma Tanıma
- Metin Üretimi

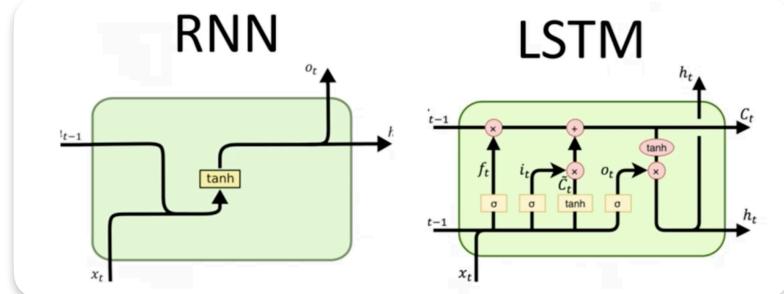
Long Short-Term Memory (LSTM)

- Long Short-Term Memory (LSTM), Recurrent Neural Network (RNN) türlerinden biridir.
- Uzun vadeli bağımlılıkları öğrenemek için özel olarak tasarlanmıştır.
- LSTM'in Ana Amacı
 - Zaman bağımlı verilerde uzun dönem bağımlılıkları öğrenmek.
 - Geleneksel RNN'lerin yaşadığı **vanishing gradient** sorununu çözmek için geliştirilmiştir.



LSTM'in RNN Üzerindeki İyileştirmesi Nedir?

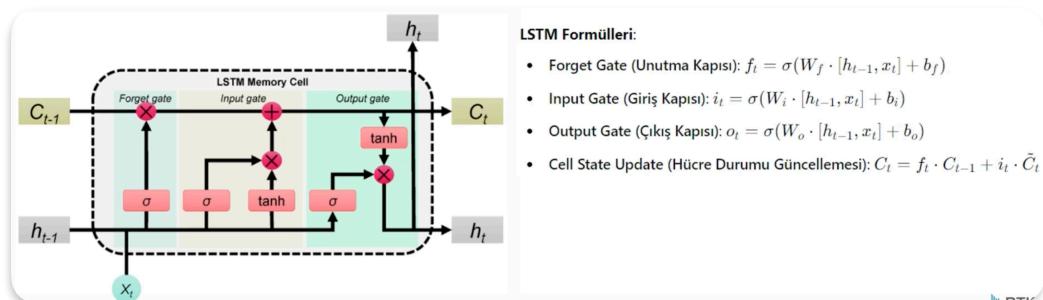
- RNN'lerde Yaşanan Sorunlar
 - Vanishing Gradient Sorunu
 - Uzun Vadeli Bağımlılıkları Öğrenememe
- LSTM'in Çözümü
 - Hücre Durumu (Cell State)
 - Kapılar (Gates)



LSTM Mimarisi, Bileşenleri ve İşleyışı Nedir?

- LSTM'in Bileşenleri
 - Hücre Durumu (Cell State)
 - Giriş Kapısı (Input Gates)
 - Unutma Kapısı (Forget Gate)
 - Çıkış Kapısı (Output Gates)

- LSTM'in İşleyışı
 - Unutma Kapısı
 - Giriş Kapısı
 - Hücre Durumunun Güncellenmesi



LSTM Kullanım Alanları

- Doğal Dil İşleme (NLP)
- Konuşma Tanıma
- Zaman Serisi Tahmini
- Müzik ve Metin Üretimi
- Video İşleme

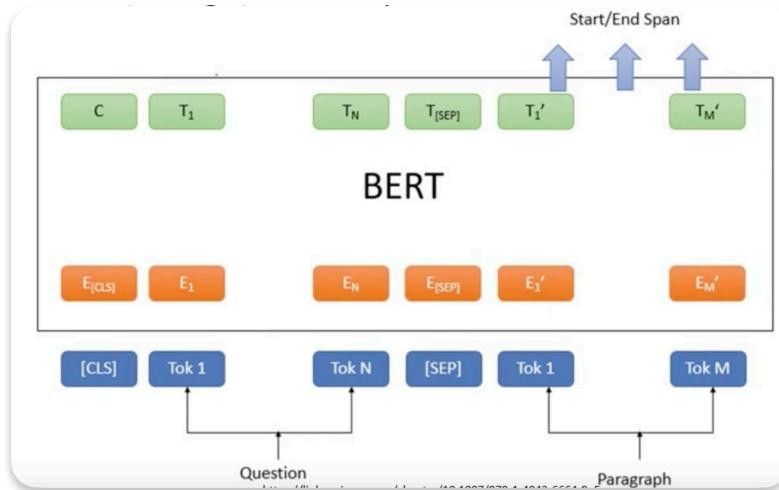
Özellik	RNN	LSTM
Uzun Dönem Bağımlılık	Uzun vadeli bağımlılıkları öğrenmede zayıf	Uzun vadeli bağımlılıkları etkili bir şekilde öğrenir
Vanishing Gradient	Vanishing Gradient sorunu yaşar	Bu sorunu çözer, uzun vadeli bağımlılıkları tutar
Kapılar (Gates)	Kapılar yok, yalnızca basit bir yapı	Giriş, unutma ve çıkış kapıları ile bilgi akışını kontrol eder
Hafıza Yönetimi	Hafızayı verimli bir şekilde yönetemez	Hücre durumu sayesinde hafızayı verimli bir şekilde yönetir
Kullanım Alanları	Kısa dizelerde veya küçük veri setlerinde daha hızlı olabilir	Uzun dizelerde daha etkili ve geniş çapta kullanılır

Transformers Modelleri: BERT

→ BERT, dil anlaması ve metin işleme görevleri için kullanılan bir dil modelidir.

→ BERT, metni hem soldan sağa hem de sağdan sola okuyarak bağlamı anlamaya çalışır. (hem geçmişe hem de geleceğe bakar)

→ Metni anlamak için kullanılır, örneğin; soru-cevaplama ve metni sınıflandırma.



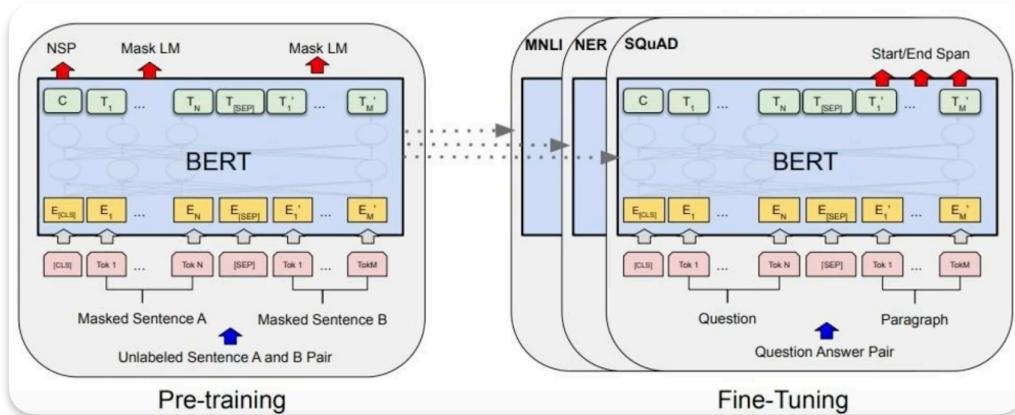
Transformers Modelleri: BERT Mimarisi

→ Transformers Encoder

- BERT, transformer mimarisinin sadece **encoder** kısmını kullanır.
- Transformer, dikkat(attention) mekanizmasına dayanan bir modeldir.

→ İki Aşamalı Eğitim

- Ön Eğitim (Pre-trained)
 - Masked Language Modeling (MLM)
 - Next Sentence Prediction (NSP)
- İnce Ayar (Fine-Tuning)



Transformers Modelleri: BERT Özellikleri

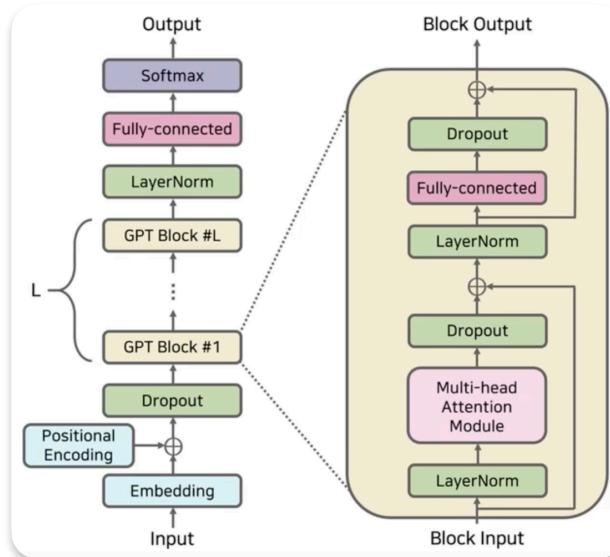
→ Çift Yönlü (Bidirectional)

→ Transfer Öğrenme

→ Transformer Encoder Kullanımı

Transformers Modelleri: GPT

- GPT, metin üretme ve dil modelleme için kullanılan bir dil modelidir.
- GPT, metni sadece soldan sağa okur ve bir kelimeyi tahmin etmek için önceki kelimelere dayanır.
- Metin üretme, öykü yazma, yaratıcı içerik oluşturma gibi görevlerde kullanılır.



→ Transformer Decoder

→ Otokorelasyonlu (Autoregressive) Yaklaşım: Her bir tokenın yalnızca önceki tokenlara dayanarak tahmininde bulunmasını sağlar.

→ Tek Aşamalı Eğitim

Transformers Modelleri: GPT Özellikleri

- Tek Yönlü (Unidirectional)
- Metin Üretimi
- Transfer Öğrenme
- Çok Büyük Modeller

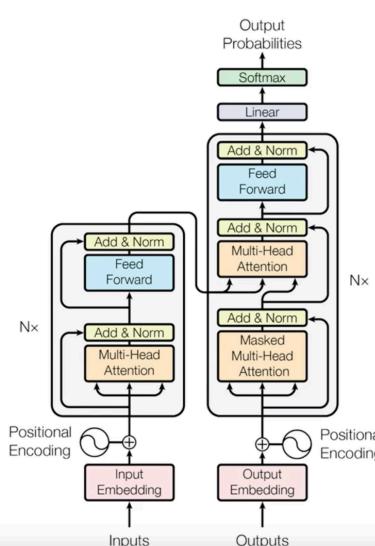
BERT vs GPT

BERT

Encoder

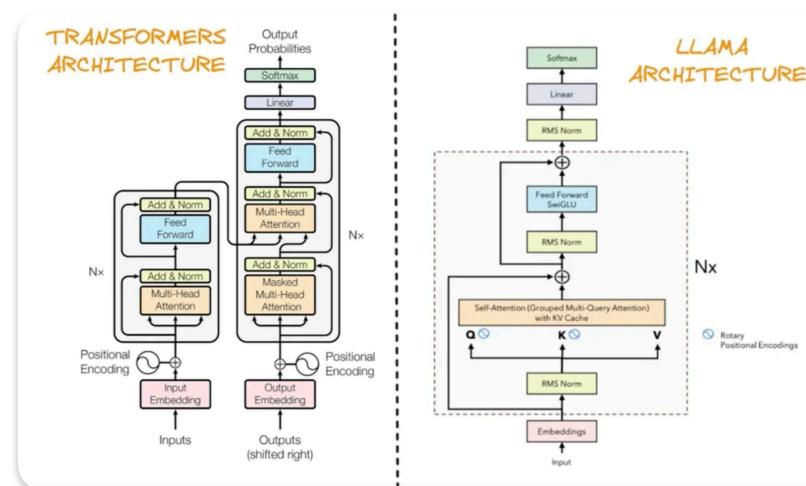
GPT

Decoder



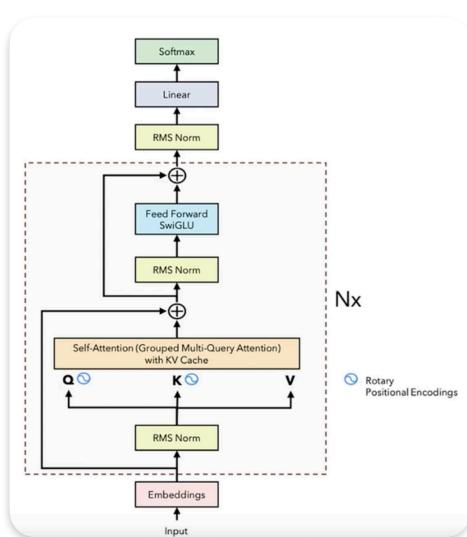
Transformers Modelleri: LLaMA

- LLaMA, META tarafından geliştirilen büyük dil modelleri ailesidir.
- LLaMA, büyük veri kümeleri üzerinde eğitilmiş olup, GPT-3 gibi modellerle kıyaslandığında daha az parametre ile benzer performans sağlamayı hedefler.



Transformers Modelleri: LLaMA Mimarisi

- Transformer Tabanlı
- Hafif ve Verimli
- Daha Küçük ve Daha Hızlı



Transformers Modelleri: LLaMA Özellikleri

- Verimli Eğitim
- Araştırmacı Odaklı
- Model Boyutları: [405B, 70B, 8B]

Özellik	BERT	GPT	LLaMA
Eğitim Yönü	Çift yönlü (bidirectional)	Tek yönlü (unidirectional)	Çift yönlü
Kullanılan Transformer	Encoder	Decoder	Encoder + Decoder
Ana Görev	Metin anlama ve sınıflandırma	Metin üretimi ve dil modelleme	Hem metin üretimi hem metin anlama
Eğitim Görevleri	Masked Language Modeling, NSP	Language Modeling	Language Modeling
Kullanım Alanları	Soru-cevap, duygusal analizi, NER	Metin üretimi, hikaye yazma, sohbet	NLP araştırmaları, düşük kaynaklı cihazlarda kullanım
Öne Çıkan Özellik	Çift yönlü bağlam öğrenme	Büyük metin üretimi, otokorelasyonlu	Verimlilik ve parametre açısından optimize edilmiş

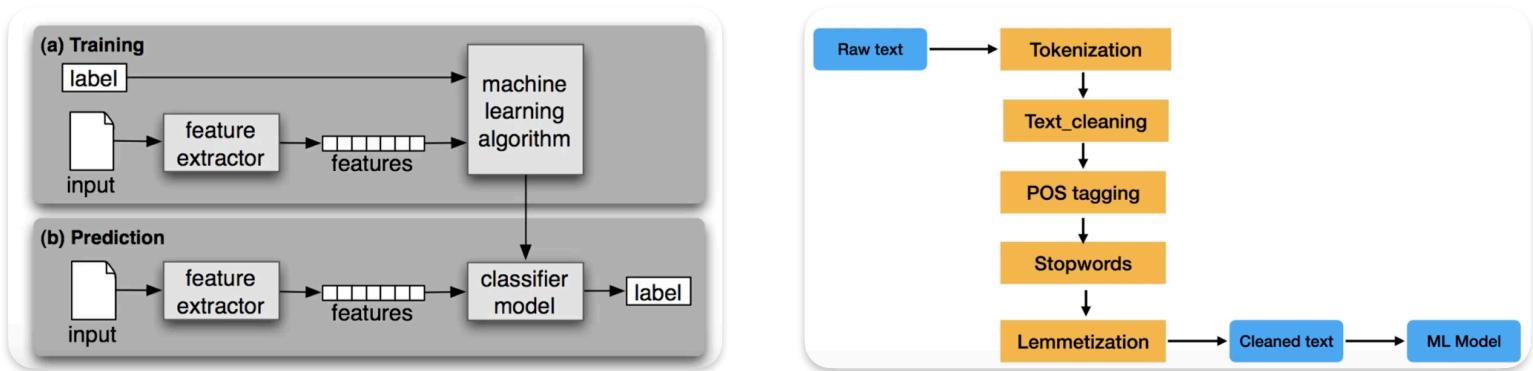
Temel NLP Görevleri

Metin Sınıflandırma (Text Classification)

→ Metin sınıflandırma, metinlerin otomatik olarak belirli kategorilere ayrılmalarını sağlar.

- E-Posta Filtreleme
- Döküman Yönetimi
- Sosyal Medya İzleme

Metin Sınıflandırma Akış Şeması



Varlık İsmi Tanıma

→ Named Entity Recognition (NER), metin içerisindeki kişi, yer, organizasyon gibi özel isimleri tanıyararak yapılandırılmamış veriden anlamlı bilgiler çıkarmayı sağlar.

- Bilgi Çıkarımı
- Otomatik Özetteleme
- Müşteri İlişkileri

Morfolojik Analiz (Morphological Analysis)

→ Morfolojik Analiz, kelimelerin yapısını inceleyerek dilbilgisel özelliklerini belirler.

→ Kullanım Alanları:

- Dil Öğrenme Araçları
- Doğal Dil İşleme
- Otomatik Çeviri

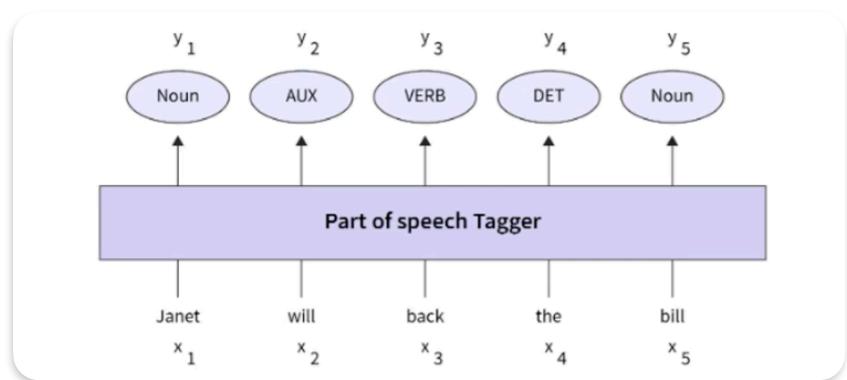
→ Örnek: Türkçe'de "kitaplar" kelimesinin kökünün "kitap" ve ekinin "-lar" olduğunu belirleyerek kelimenin çoğul olduğunu tespit etmek.

Metin Parçası Etiketleme

→ Part-of-Speech (POS) Tagging, bir metindeki her kelimenin dilbilgisel kategorisini belirleyerek cümlelerin dil yapısını analiz etmeyi sağlar.

→ Kullanım Alanları

- Doğal Dil İşleme
- Makine Çevirisi
- Sözcük Türü Analizi



Kelime Anlamı Belirsizliği Giderme

→ Word Sense Disambiguation, bir kelimenin farklı anamları arasından doğru olanı bağlama göre seçme işlemidir.

→ Kullanım Alanları:

- Makine Çevirisi
- Arama Motorları
- Doğal Dil İşleme

ÇAY
İçeceğ <-----| |-----> Akarsu

Duygu Analizi (Sentiment Analysis)

- Müşteri Geri Bildirimleri
- Sosyal Medya Analizi
- Pazarlama Stratejisi

Gelişmiş NLP Görevleri

Soru Cevaplama

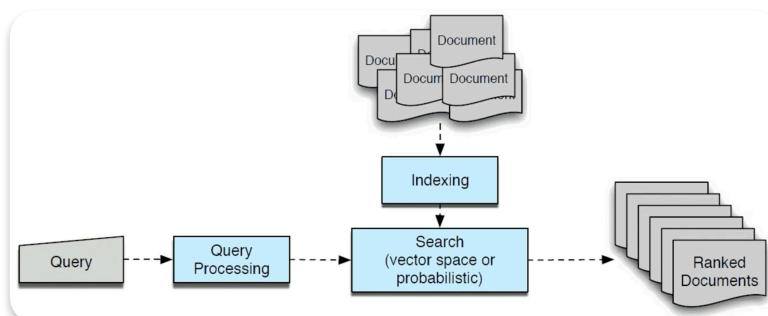
→ Soru Cevaplama (QA), bir sistemin doğal dilde sorulan sorulara, bir bilgi kaynağına dayalı olarak doğru cevaplar vermesini sağlayan bir doğal dil işleme (NLP) görevidir.

→ QA sistemler, tıbbi sorular, bilimsel araştırmalar, hukuk belgeleri veya kullanıcı destek sistemleri gibi daha karmaşık senaryolarda da kullanılabilir.

Bilgi Getirme (Information Retrieval)

→ Bilgi getirme, kullanıcının sorularına (yani tarama terimlerine) en uygun belgeleri veya verileri bulmak için kullanılan bir tekniktir.

Örnekler: Arama motorları ve kütüphane sistemleri

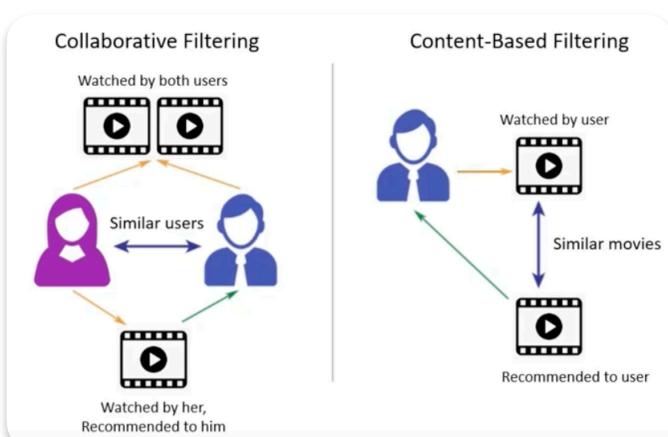
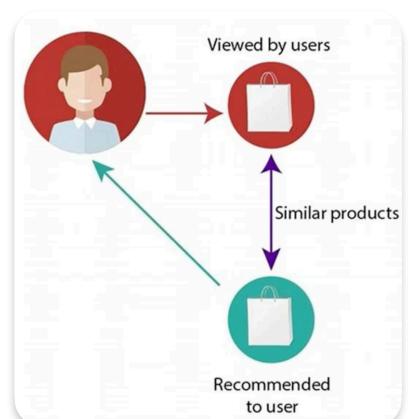


Öneri Sistemleri (Recommendation System)

→ Öneri sistemleri (Recommendation Systems), kullanıcılarla ilgilerini çekebilecek ürünler, hizmetler veya içerikler önermek için kullanılan sistemlerdir.

→ Bu sistemler, kullanıcıların geçmiş davranışlarına, tercihlerine ve diğer veriye dayalı analizlere dayanarak önerilerde bulunur.

- İçerik Tabanlı Öneri Sistemleri
 - Kullanıcının geçmişteki tercihlerine dayalı olarak benzer özelliklere sahip ürünleri önerir.
- İşbirlikçi Filtreleme (Collaborative Filtering)
 - Kullanıcıların geçmişteki davranışları ve diğer kullanıcılarla olan benzerliklere dayalı önerilerde bulunur.
 - Kullanıcı Temelli İşbirlikçi Filtreleme
 - Öğe Temelli İşbirlikçi Filtreleme



Makine Çevirisi (Machine Translation)

- Makine çevirisi (Machine Translation - MT), bir dilde yazılmış bir metin, bilgisayar programları ve algoritmalar aracılığıyla başka bir dile otomatik olarak çevrilmesi işlemidir.
- Çalışma yöntemleri
 - Kural Tabanlı Çeviri
 - İstatistiksel Makine Çevirisi
 - Nöral Makine Çevirisi
 - Seq2Seq Modelleri
 - Transformers

Metin Öztleme (Text Summarization)

- Metin öztleme (text summarization), uzun metinlerin daha kısa ve öz biçimde özetenmesini amaçlayan bir doğal dil işleme (NLP) görevidir.

- Yaklaşımları
 - Özelleştirilmiş (Extractive) Öztleme
 - Özgün (Abstractive) Öztleme

