ATATÜRK ÜNİVERSİTESİ

Mühendislik Fakültesi

Kümeleme Tabanlı Metin Sınıflandırma: Cinsiyet Çıkarımı Örneği

Cinsiyet Bilgisi İçeren Metinlerin Kümeleme Algoritmalarıyla Sınıflandırılması

Hazırlayan: Ensar Şehitoğlu

12.12.2024

Kümeleme Tabanlı Metin Sınıflandırma: Cinsiyet Çıkarımı Örneği

Bu sunum, kümeleme tabanlı metin sınıflandırma tekniklerini kullanarak cinsiyet çıkarımını nasıl gerçekleştirebileceğimizi ele almaktadır. Bu yöntem, doğal dil işleme ve makine öğrenimi alanlarında giderek daha önemli bir rol oynamaktadır.

Giriş

Amaç:

- Metinlerden cinsiyet çıkarımı yapmak ve bunu kümeleme teknikleriyle değerlendirmek.
- Cinsiyet çıkarımı için metinlerin önce kümeleme algoritmaları ile gruplandırılması ve ardından sınıflandırma algoritmaları ile etiketlenmesi.

- **Makine Öğreniminde Sınıflandırma ve Kümeleme:**
  - Sınıflandırma: Etiketli verilerle çalışır (örneğin, "erkek" veya "kadın").
  - Kümeleme: Etiketsiz verilerle gruplar oluşturur.
- **Uygulama Alanları:**
  - Doğal dil işleme (NLP) ve sosyal medya analizinde cinsiyet çıkarımı.
  - Daha önce etiketi olmayan verilerden bilgi çıkarma.

Araştırma Sorusu ve Yöntem

Araştırma Sorusu:

- Metinlerdeki cinsiyet bilgisi kümeleme ile başarılı bir şekilde tespit edilebilir mi?

Yöntem:

- Kümeleme algoritmaları ile metin sınıflandırma yapılması
- Farklı kelime gömme (word embedding) teknikleri ile metinlerin temsil edilmesi

Metin Sınıflandırma: Genel Bakış

Metin sınıflandırma, makine öğrenimindeki temel bir görevdir. Metin verilerini anlamlı kategorilere ayırmak için çeşitli yöntemler kullanır.

Denetimli Öğrenme

Denetimli öğrenme, önceden etiketlenmiş verileri kullanarak bir model eğitir. Model, etiketlere dayalı olarak yeni verileri sınıflandırır.

Denetimsiz Öğrenme

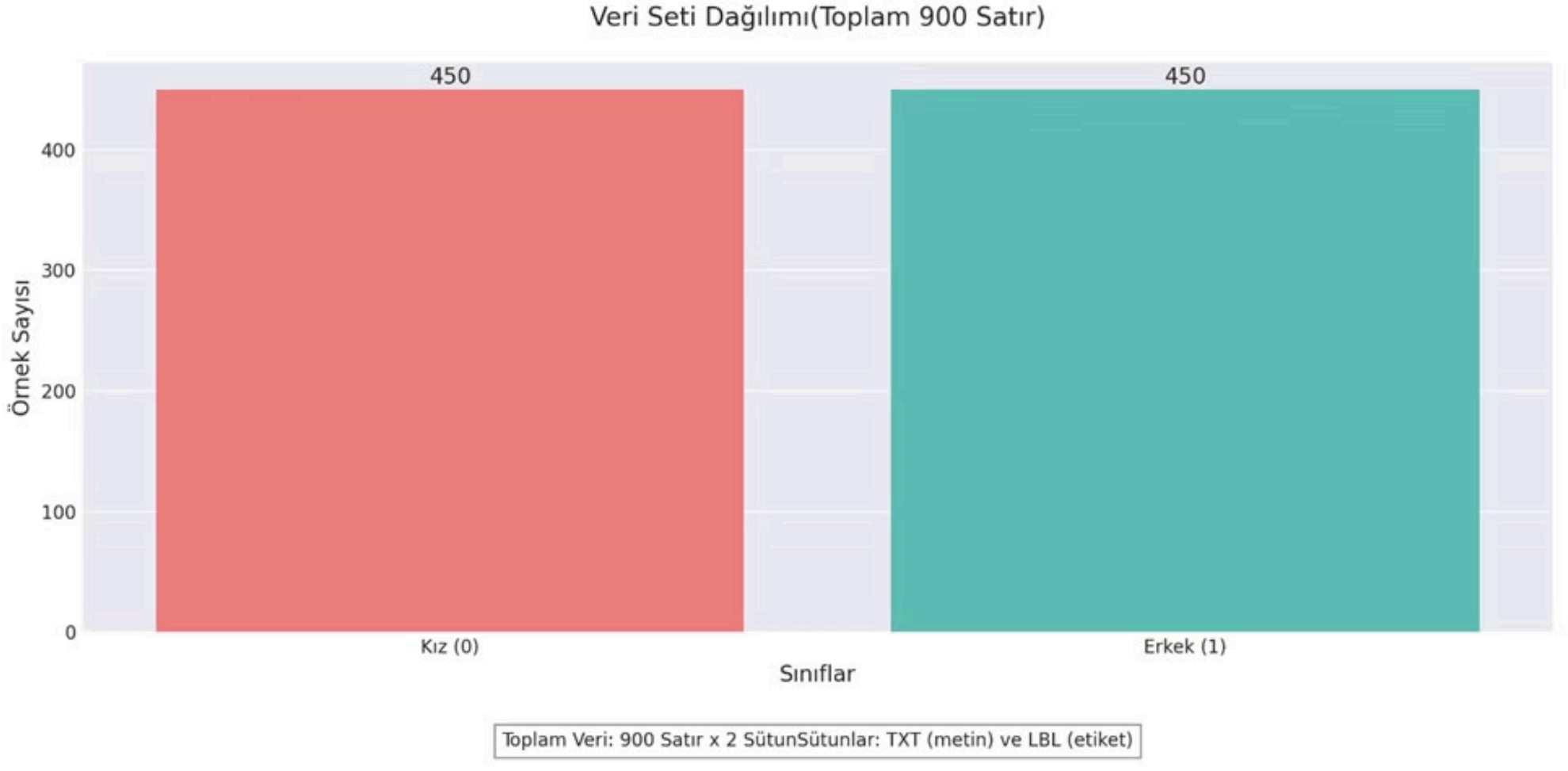
Denetimsiz öğrenme, etiketlenmemiş verileri kullanarak kalıpları ve ilişkileri keşfeder. Kümeleme, denetimsiz öğrenmenin yaygın bir yöntemidir.

Yarı Denetimli Öğrenme

Yarı denetimli öğrenme, hem etiketlenmiş hem de etiketlenmemiş verileri kullanır. Etiketlenmiş verilerden bir model eğitir ve daha sonra etiketlenmemiş verileri kullanarak modeli iyileştirir.

Veri Kümesi

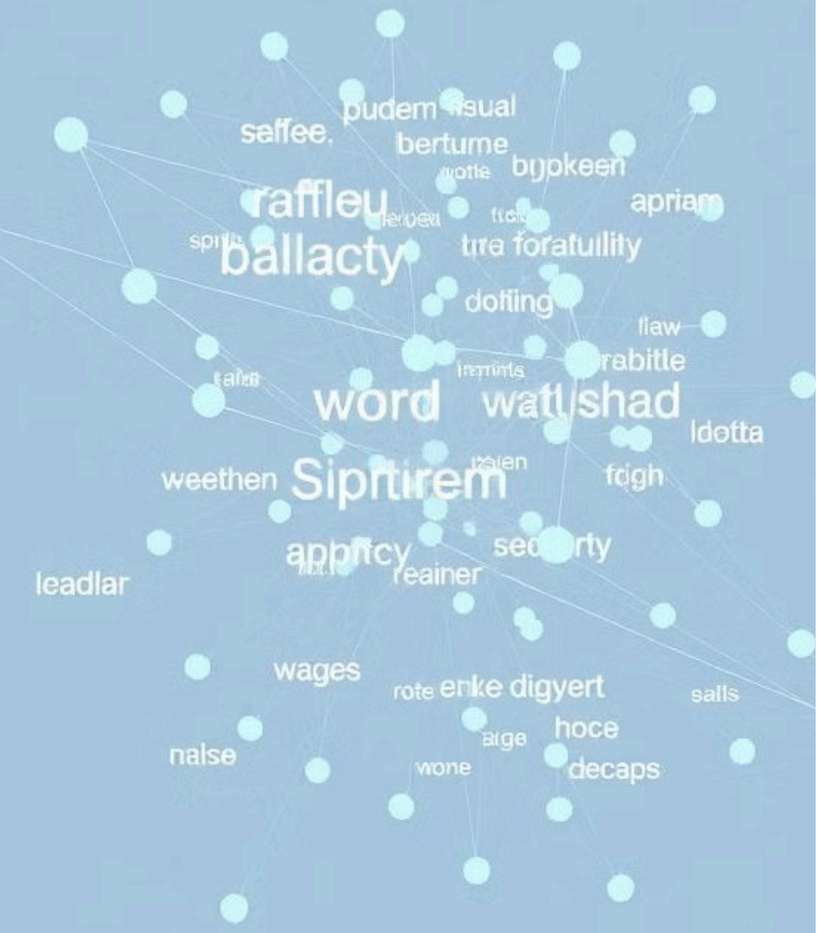
Bu çalışmada kullanılan veri kümesi, Twitter'dan alınmış mesajları içeren Türkçe ve İngilizce metinlerden oluşmaktadır. Her satırda bir metin ve ona ait cinsiyet etiketi yer almaktadır. Türkçe veri kümesi 900 adet ingilizce veri kümesi ise 5000 adet metinden oluşmaktadır.



Veri Seti ve Ön İşleme

Cinsiyet çıkarımı için metin verisi toplanır ve ön işleme yapılır. Bu adım, verileri modele uygun hale getirmeyi içerir.

- **Küçük Harfe Çevir**: Tüm harfleri küçük harfe dönüştürme.
- **Boşlukları Temizle**: Fazla boşlukları ve gereksiz boşlukları kaldırma.
- **Emojileri Kaldır**: Tüm emoji ve simgeleri metinden çıkarma.
- **URL'leri Kaldır**: Metindeki URL bağlantılarını temizleme.
- **Sayıları Kaldır**: Tüm rakamları metinden silme.
- **Alfabetik Olmayan Karakterleri Kaldır**: Harf dışındaki karakterleri (özel karakterler, semboller) temizleme.
- **Noktalama İşaretlerini Kaldır**: Nokta, virgül, ünlem gibi tüm noktalama işaretlerini silme.
- **Fazla Boşlukları Kaldır**: Ardışık boşlukları tek boşluğa indirme.
- **Stop Kelimeleri Kaldır**: Sık kullanılan anlamsız bağlaç ve zamirleri (stop kelimeleri) temizleme.
- **Tekrar Eden Karakterleri Kaldır**: Art arda gelen aynı harfleri sadeleştirme (örneğin, "harrrika" → "harika").
- **Kısa Kelimeleri Kaldır**: Belirli bir harf sayısından daha kısa kelimeleri silme (örneğin, 2 veya 3 harfli kelimeler).
- **Kök Bulma (Stemming)**: Kelimelerin köklerine indirilmesi (örneğin, "koşuyorum" → "koş").
- **Sabit Önek Kök Bulma**: Kelimeleri sabit bir önek üzerinden köklerine ayırma (örneğin, "anlamış" → "anla").



Özellik Çıkarma Yöntemleri

Özellik çıkarma, metin verisinden anlamlı özellikleri ayıklama sürecidir. Bu özellikler, sınıflandırma modeli tarafından kullanılır.

1

Kelime Frekansı: TF (Term Frequency)

Her kelimenin metin içindeki frekansı hesaplanır. Yaygın olarak kullanılan kelimeler, cinsiyet çıkarımı için önemli ipuçları sağlayabilir.

2

TF-IDF

Terim Frekansı-Ters Belge Frekansı (TF-IDF), bir kelimenin bir belgede ne kadar önemli olduğunu ölçer.

3

Kelime Çantası (BoW - Bag of Words)

Metin veri kümesindeki tüm kelimeleri dikkate alarak, her bir belgeyi kelime frekanslarıyla temsil eder.

4

n-gram

"n" bir sayı ile ifade edilir ve bu sayı, bir dizi ardışık kelimenin uzunluğunu belirtir.

(3-gram): Üç ardışık kelime grubu. Örneğin, "bu bir örnek" bir trigramdır.

Gömme Yöntemleri

Gömme, metin verilerini daha iyi temsil etmek için kelimeleri vektörlere dönüştürür.

1

Word2Vec

Kelimeleri, komşu kelimelerin bağlamına dayalı olarak vektörlere dönüştürür.

2

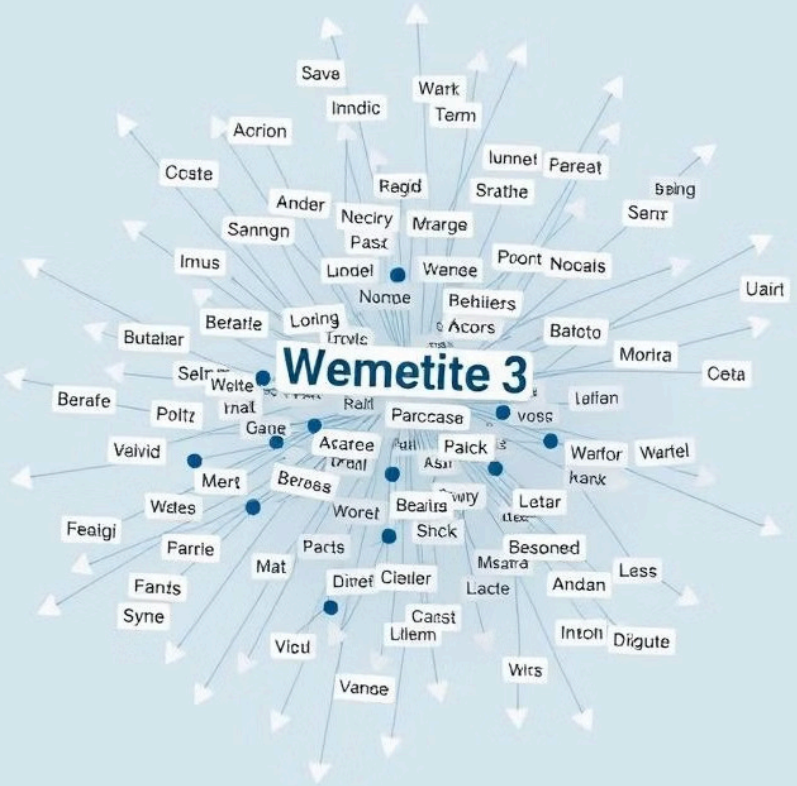
GloVe

Global Vektörler için Küresel Sözlük (GloVe), kelimelerin eşzamanlı olarak tüm kelimelerin frekanslarını dikkate alarak vektörler oluşturur.

3

FastText

Gömme oluşturmak için karakter seviyesinde bilgi kullanır, bu da nadir veya çoklu dil kelimeleri için daha iyi performans sağlar.

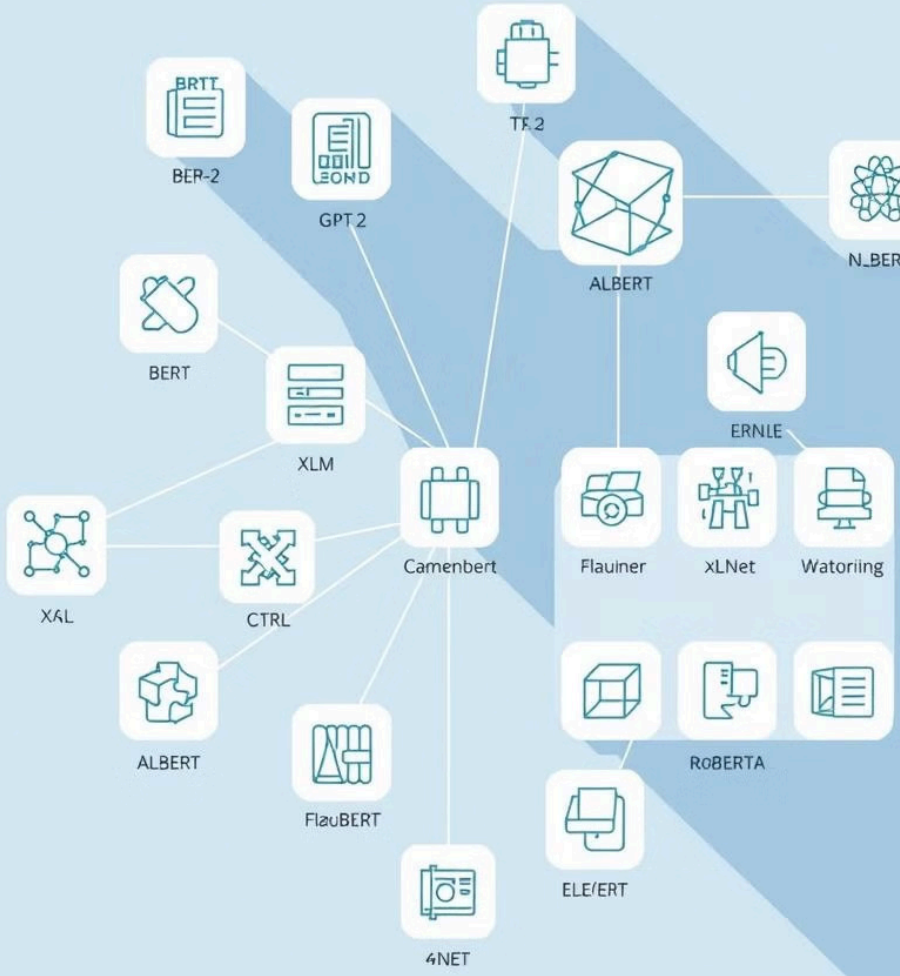


Dil Modelleri

Doğal dil işleme (NLP) alanında önemli ilerlemeler sağlayan çeşitli dil modelleri mevcuttur. Bu modeller, dil anlama, metin üretimi, çeviri ve birçok başka görevde kullanılmıştır.

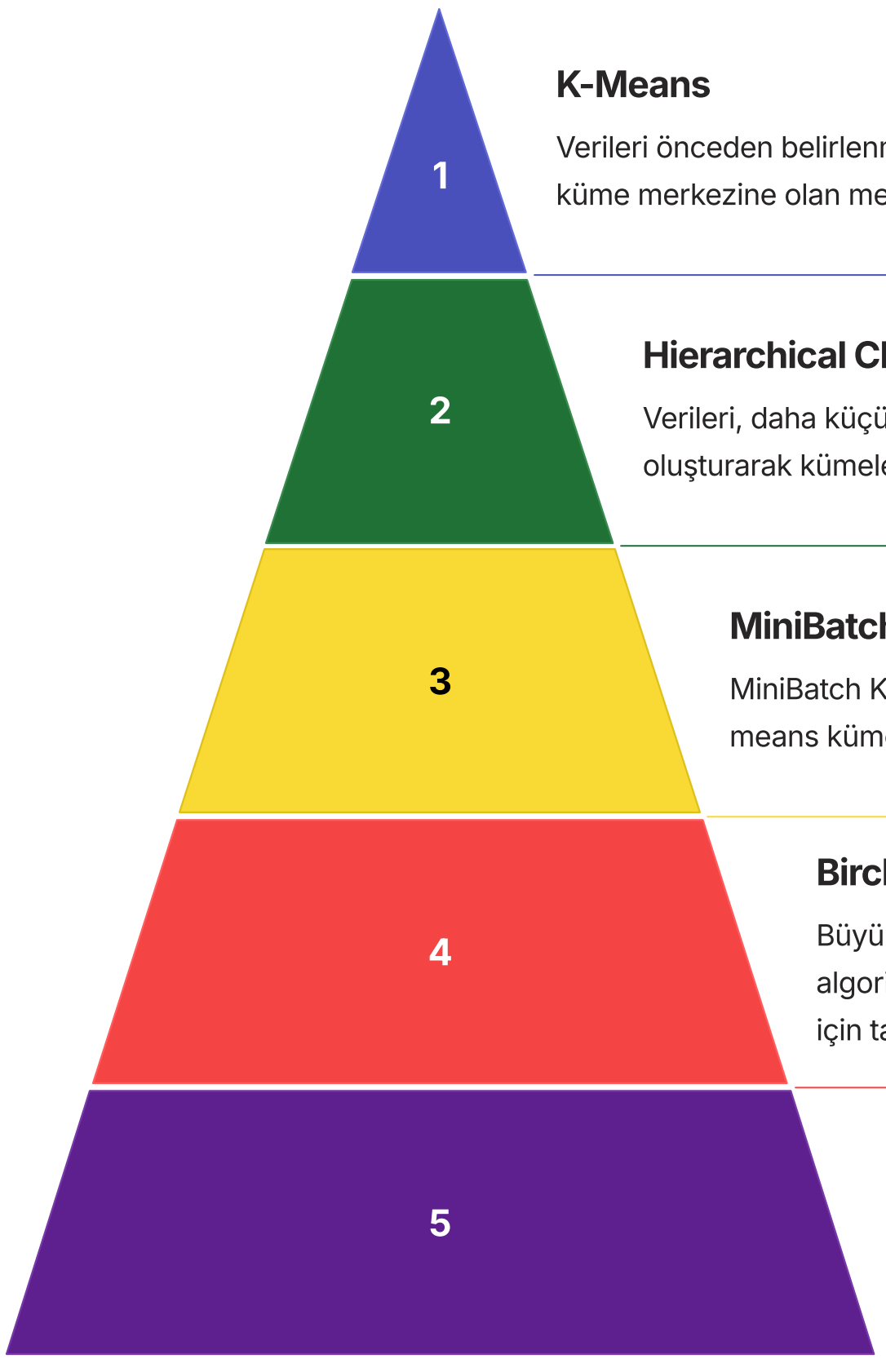
1

- BERT, GPT-2, MarianMT, T5, CamemBERT, XLM, CTRL, Reformer, XLNet, ALBERT, ERNIE, FlauBERT, ELECTRA, RoBERTa



Kümeleme Algoritmaları

Kümeleme, etiketlenmemiş verileri benzerliklerine göre gruplara ayırmayı amaçlar.



K-Means

Verileri önceden belirlenmiş K sayıda kümeye böler, böylece her kümedeki noktaların küme merkezine olan mesafesi en aza indirilir.

Hierarchical Clustering

Verileri, daha küçük kümelere bölünerek veya birleştirilerek bir hiyerarşik yapı oluşturarak kümeler.

MiniBatch K-Means

MiniBatch K-Means, büyük veri kümeleri için kullanılan hızlı ve verimli bir k-means kümeleme algoritmasıdır.

Birch

Büyük veri kümeleriyle çalışırken kullanılan hiyerarşik bir kümeleme algoritmasıdır. Özellikle büyük ve bellek kullanımı kısıtlı veri setleri için tasarlanmıştır.

Agglomerative Clustering

Hiyerarşik kümeleme yöntemlerinden biridir ve veri noktalarını bir hiyerarşi içinde kümeler

Sınıflandırma Süreci

Kümelenen metin verileri, cinsiyet çıkarımı için kullanılır.

1

Küme Etiketleme

Her kümenin cinsiyetle ilişkilendirilmesi gerekir (örneğin, küme 1 "erkek", küme 2 "kadın").

2

Sınıflandırma

Yeni bir metin verisi, en benzer kümeye atanarak sınıflandırılır.

Klasik Makine Öğrenmesi Modelleri:

- Naive Bayes (MNB)
- Lojistik Regresyon (LR)
- Destek Vektör Makineleri (SVM)
- Karar Ağaçları (DT)
- Rastgele Ormanlar (RF)

3

Cinsiyet Çıkarımı

Yeni metin verisinin cinsiyeti, atandığı kümenin cinsiyet etiketiyle belirlenir.

Değerlendirme Metrikleri

Modelin performansını değerlendirmek için çeşitli metrikler kullanılır.

Sınıflandırma Metrikleri:

- Doğruluk (Accuracy)
- F1-Makro Skoru (F1-Macro)
- Kesinlik (Precision), Geri Çağırma (Recall)

Kümeleme Metrikleri:

- Silhouette Skoru
- Davies-Bouldin Skoru
- Calinski-Harabasz Skoru

Sonuçlar ve Tartışma

Kümeleme tabanlı metin sınıflandırma, cinsiyet çıkarımı için etkili bir yöntem olabilir.

- Cinsiyet bilgisi metinlerden kümeleme ile çıkarılabilir.
- Farklı embedding ve dil modellerinin performansa etkisi gözlemlenmiştir.
- Küme sayısının sınıflandırma başarısına etkisi gözlemlenmiştir.
- Dil modellerinin geleneksel yöntemlere göre avantajları ve dezavantajları.

thank you
danke
рахмат
謝謝
ngiyabonga
teşekkür ederim
Баярлалаа
spas
mersi
kia ora
barka
welalin
tack
спасибо
faafetai lava
vinaka
dank je
misaotra
paldies
grazzi
mahalo
tapadh leat
blagodaram
спасибі
matondo
gracias
kiitos
dankie
nanni
nandri
dhanyavad
mauruuru
хвала
asante
manana
bedankt
köszönöm
obrigada
tenki
bayarlalaa
gràçie
hvala
dankon
aĉiŭ
murakoze
enkosi
akun
djiere dieuf
tau
mochchakkeram
dziękuję
дякую
mamnun
chnorakaloutioun
gratias ago
gràcies
sulpáy
go raibh maith agat
chokrane
sobodi
dĕkuji
sagolun
sukriya
kop khun krap
taiku
obrigado
arigatô
takk
dakujem
trugarez
mèsi
didi madloba
ありがとう
tanemirt
grazie
najis tuke
terima kasih
rahmet
diolch
dhanyavadagalu
shukriya
mercé
мерси
kam sah hamnida
rahmat
তোমাকে ধন্যবাদ
감사합니다
xièxie
ευχαριστώ
merci