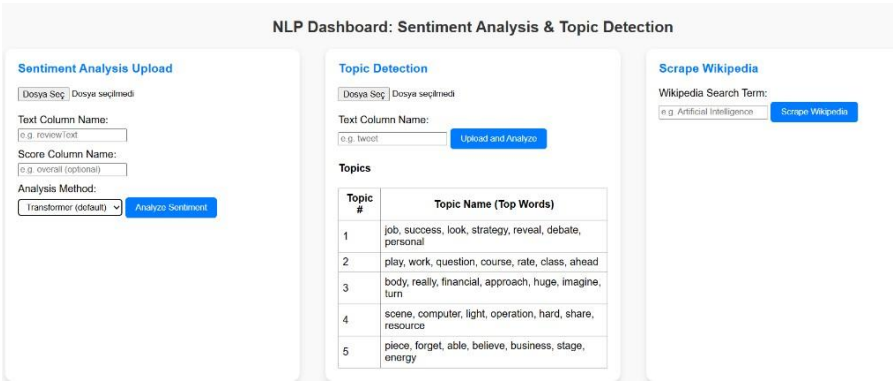
Samir Ganbarli

Veb Madenciliği için Çok Modüllü Bir Python/Django Uygulaması: Duygu Analizi, Konu Belirleme ve Wikipedia'dan Veri Çekme

- Projenin amacı ve Motivasyonu

Bu proje, web madenciliği kapsamında **duygu analizi**, **konu belirleme** ve **Wikipedia'dan veri çekme (scraping)** olmak üzere üç temel modül içermektedir. Duygu analizi modülü, sosyal medya veya kullanıcı yorumları gibi metinlerdeki olumlu, olumsuz ve nötr duyguları otomatik olarak tespit ederek, toplumsal eğilimlerin veya müşteri memnuniyetinin ölçülmesine olanak tanır. Proje kapsamındaki konu belirleme modülü ise büyük metin yığınlarında öne çıkan ana temaları ve konuları otomatik olarak çıkararak, veri içerisindeki bilgi yoğunluğunu anlamayı ve özetlemeyi kolaylaştırır. Wikipedia scraping modülü ise, güvenilir ve güncel bilgiye hızlıca erişim sağlayarak analizlerde kullanılacak ek veri kaynağı sunar. Bu üç modül, web üzerindeki büyük ve çeşitli veri kaynaklarından anlamlı bilgi çıkarımı yapılmasını sağlayarak, web madenciliği projelerinde veri toplama, ön işleme ve analiz süreçlerinin bütünleşik bir şekilde yürütülmesine katkı sağlar.



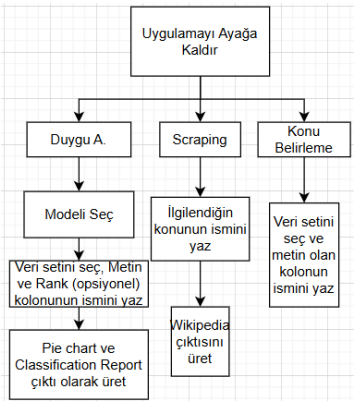
Şekil 1. Uygulamanın önyüzü

- Veri Setleri

Bu projede kullanılan veri setleri, **Kaggle** platformundan temin edilmiştir. Duygu analizi (opinion mining) ve konu kümeleme modülleri için, özellikle **Amazon**, **Starbucks** gibi popüler uygulamalara ait kullanıcı yorumları içeren veri setleri tercih edilmiştir. Projede, veri setlerinin makine öğrenmesi ve doğal dil işleme modellerinde kullanılabilir hale gelmesi için çeşitli **ön işleme (pre-processing)** adımları uygulanmıştır. Bu adımlar arasında: Boşlukların ve eksik değerlerin temizlenmesi, Metinlerin standartlaştırılması (Tüm metinler küçük harfe çevrilmiş ve gereksiz karakterler temizlenmiştir), TF-IDF (Term Frequency-Inverse Document Frequency – Naive Bayes ve Clustering için) vektörleştirme, Gereksiz boşlukların ve sadece boşluklardan oluşan satırların çıkarılması gibi yöntemler kullanılmıştır (Duygu analizi için dönüştürücü tabanlı modelde kelime gömme modelin kendi içinde yapılmıştır).

- Sistemin Yapısı

Duygu analizi modülünde, klasik yöntem olarak Naive Bayes sınıflandırıcısı ve clustering (kümeleme) için K-Means algoritması uygulanmıştır. Ayrıca, daha yüksek doğruluk ve bağlamsal analiz için dönüştürücü (transformer) tabanlı bir model olan CardiffNLP/twitter-roberta-base-sentiment modeli kullanılmıştır. Konu modelleme (topic modeling) aşamasında ise, metinlerdeki ana temaları ortaya çıkarmak amacıyla NMF (Non-negative Matrix Factorization) algoritmasından yararlanılmıştır. Uygulamanın arka planında, modüllerin entegrasyonu ve veri akışının yönetimi için Django web çatısı tercih edilmiştir. Kullanıcı arayüzü ise Django'nun DTL (Django Template Language) tabanlı şablon sistemi ile geliştirilmiş, böylece kullanıcıların veri yükleme, analiz ve sonuçları görselleştirme işlemlerini kolayca gerçekleştirmesi sağlanmıştır.

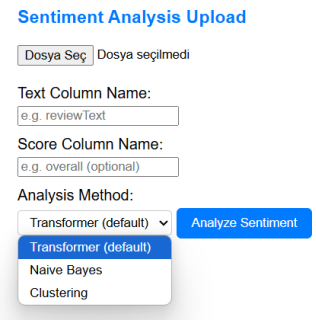


Şekil 2. Uygulama Basit Akış Diyagramı

- Kullanım Senaryosu, Algoritmalar

Duygu Analizi Kullanım Senaryosu

Duygu analizi modülünde kullanıcıya üç farklı model seçeneği sunulmaktadır: Transformer tabanlı model, Naive Bayes ve Clustering (kümeleme). Kullanıcı, analiz yapmak istediği modeli açılır menüden seçtikten sonra, analiz etmek istediği veri setini yükler. Ardından, veri setinde yorumların bulunduğu **kolonun adını** ilgili alana girer. Eğer kullanıcı, modelin başarımını ölçmek istiyorsa ve veri setinde gerçek etiketleri (örneğin puan/rank) içeren bir kolon varsa, bu kolonun adını da girer. Tüm bilgiler girildikten sonra "Analyze Sentiment" butonuna tıklanır. Analiz tamamlandığında, kullanıcıya tahmin edilen duygu dağılımını gösteren bir pie chart ve eğer başarı ölçümü için etiketli veri sağlanmışsa, modelin performansını gösteren bir classification report sunulur.



Konu Modelleme Kullanım Senaryosu:

Konu modelleme modülünde kullanıcıdan, analiz etmek istediği veri setini yüklemesi ve metinlerin (örneğin tweet veya yorumların) bulunduğu kolonun adını belirtmesi istenir. Kullanıcı bu bilgileri girdikten sonra "Upload and Analyze" butonuna tıklar. Sistem, yüklenen metinler üzerinde **NMF (Non-negative Matrix Factorization)** algoritmasını kullanarak ana konuları otomatik olarak belirler ve her bir konuyu temsil eden en önemli anahtar kelimeleri tablo halinde kullanıcıya sunar.

Wikipedia Scraping Kullanım Senaryosu: Wikipedia scraping modülünde kullanıcı, bilgi almak istediği konuyu veya anahtar kelimeyi arama kutusuna yazar ve "Scrape Wikipedia" butonuna tıklar. Sistem, **wikipedia** Python kütüphanesini kullanarak ilgili başlığın özet bilgisini çeker ve kullanıcıya sunar.

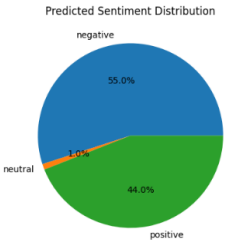
- Karşılaşılan Problem Örneği

Farklı kaynaklardan alınan veri setlerinde kolon isimleri ve veri yapısı değişiklik gösterebiliyor. Kullanıcıdan kolon adını manuel olarak istemek bu sorunu büyük ölçüde çözse de, yanlış kolon adı girilmesi durumunda hata mesajı gösterilmesi gerekebiliyor.

- Gelicekte Yapılması Planlanan

Scraping ile çekilen faydalı verilerin belirli bir formatta indirilmesi opsiyonun temin edilmesi.

Sentiment Distribution



Classification Report

	precision	recall	f1-score	support
negative	0.73	1.00	0.84	40
neutral	1.00	0.05	0.10	20
positive	0.89	0.97	0.93	40
accuracy			0.80	100
macro avg	0.87	0.67	0.62	100
weighted avg	0.85	0.80	0.73	100

- Duygu analizi Naïve Bayes ile örnek