Word Embedding ve Word embeddingsiz NLP problemi çözümü

Berke Abik-220609019

Muhammed Serdar Meşe-220609039

Sadık Ahmet Karabulut-220609018

Word Embedding nedir ?

Word Embedding, doğal dil işleme (NLP) alanında kullanılan bir tekniktir. Bu yöntem, kelimeleri sayısal vektörlere (veya yoğun vektörler) dönüştürür. Word embedding'ler, kelimelerin anlamını daha iyi temsil etmek için kullanılır ve kelimeler arasındaki semantik ilişkileri (benzerlikler ve farklar) yakalamaya çalışır.

1. Word2Vec

- **Word2Vec**, kelimeleri vektörlere dönüştüren popüler bir modeldir.
- **Skip-Gram** ve **CBOW** (Continuous Bag of Words) olmak üzere iki ana yaklaşımı vardır.
- Kelimeler arasındaki semantik ilişkiyi öğrenir ve benzer anlamdaki kelimeleri benzer vektörlerle temsil eder.

2. GloVe (Global Vectors for Word Representation)

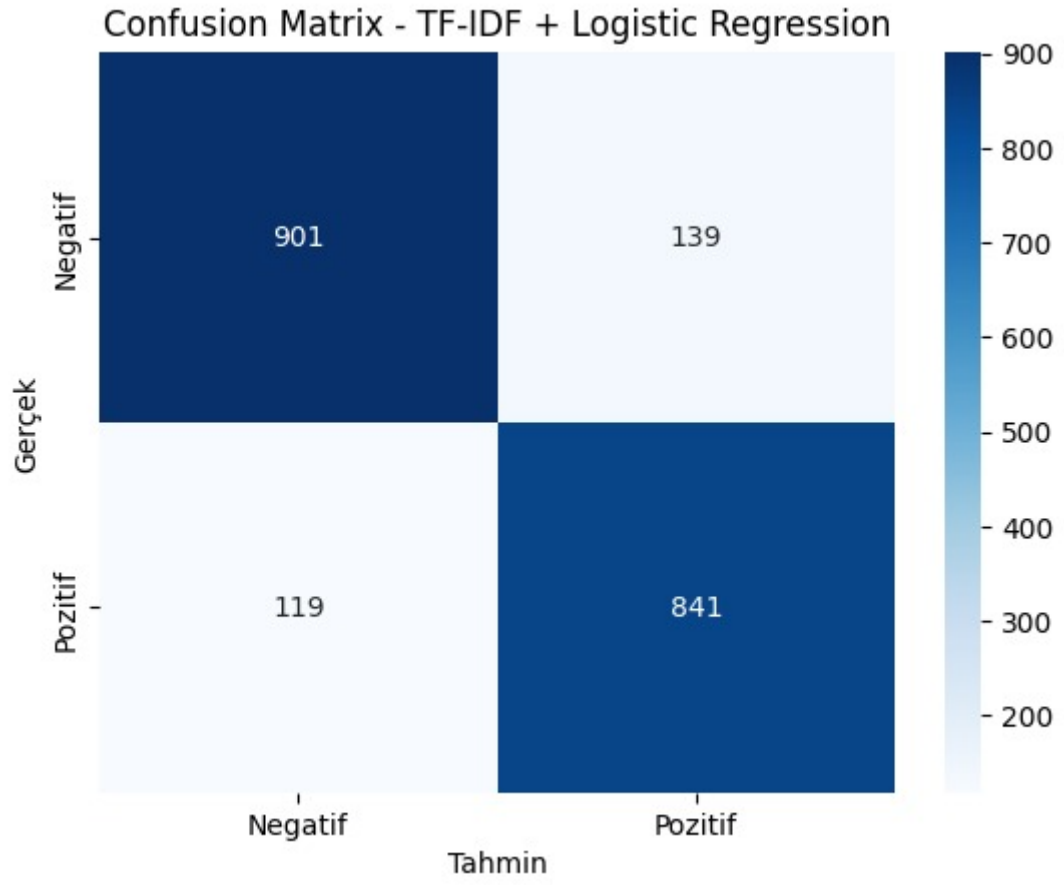
- GloVe, kelimeleri **global** istatistiklere dayanarak vektörlere dönüştürür.
- Kelimelerin birlikte görülme sıklığını dikkate alır ve bu istatistiklerle kelimeler arasındaki anlamlı ilişkileri öğrenir.

3. FastText

- **FastText**, kelimeleri **karakter n-gramları** kullanarak temsil eder.
- Bu sayede nadir kelimeleri ve bileşik kelimeleri daha doğru modelleyebilir.
- Kelimeleri sadece kelime bazında değil, aynı zamanda içerdikleri alt parçalara (n-gramlara) dayanarak işler.

IMD film eleştirisi veriseti

Word Embedding olmadan çözüm



Satırlar (Gerçek Değerler):

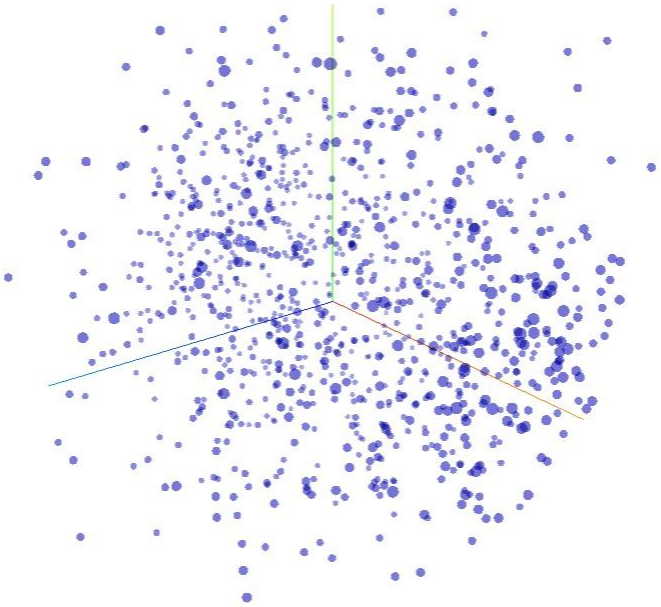
- **Negatif:** Gerçekte olumsuz olan film eleştirilerini temsil eder.
- **Pozitif:** Gerçekte olumlu olan film eleştirilerini temsil eder.

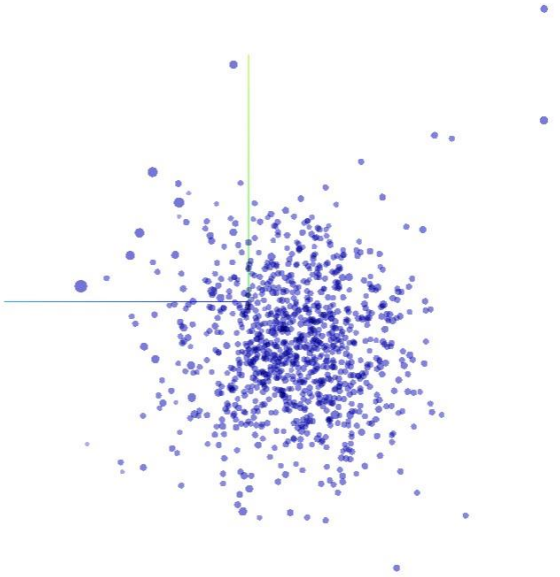
Sütunlar (Tahmin Edilen Değerler):

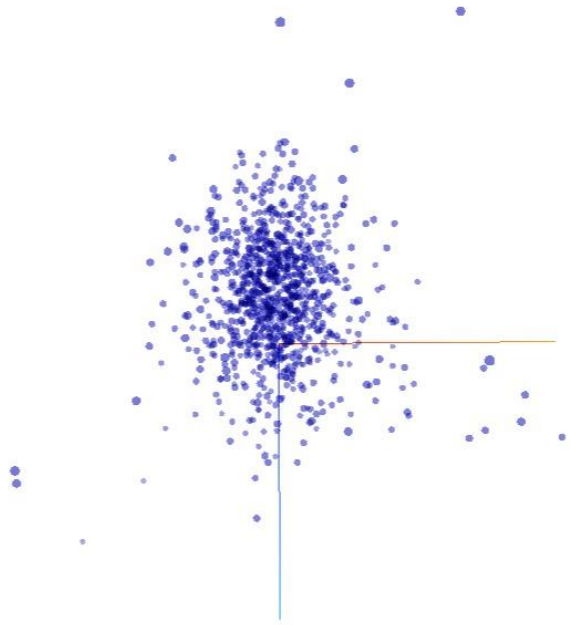
- **Negatif:** Modelin olumsuz olarak tahmin ettiği film eleştirilerini temsil eder.
- **Pozitif:** Modelin olumlu olarak tahmin ettiği film eleştirilerini temsil eder.

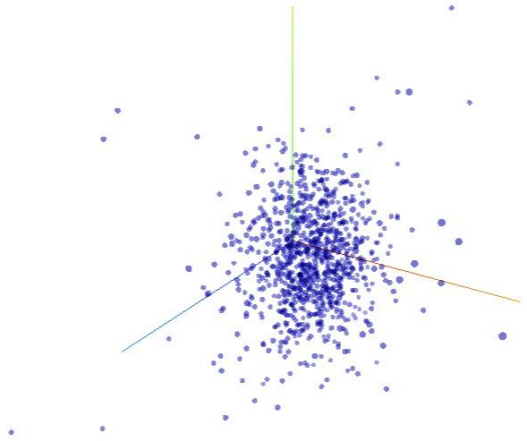
Model genel olarak iyi bir performans sergilemiş gibi görünüyor çünkü doğru tahmin sayıları (901 ve 841) yanlış tahmin sayılarından (139 ve 119) oldukça yüksek. Ancak, hala iyileştirilebilecek noktalar mevcut. Özellikle, gerçekte olumsuz olan bazı eleştirilerin (139 adet) yanlışlıkla olumlu olarak sınıflandırılması ve gerçekte olumlu olan bazı eleştirilerin (119 adet) yanlışlıkla olumsuz olarak sınıflandırılması söz konusu.

Word Embedding ile çözüm









Problem Tanımı:

Bu projede IMDb film yorumları veri seti kullanılarak bir *duygu analizi (sentiment analysis)* problemi çözülmüştür. Amaç, verilen film yorumunun olumlu (positive) veya olumsuz (negative) olduğunu otomatik olarak sınıflandırmaktır.

Çözüm 1 – Word Embedding Olmadan (TF-IDF + Logistic Regression):

İlk çözümde, yorumlar metin haline getirilip TF-IDF (Term Frequency–Inverse Document Frequency) yöntemiyle sayısal vektörlere dönüştürüldü. Bu vektörler kullanılarak klasik bir makine öğrenmesi modeli olan *Lojistik Regresyon* eğitildi.

- **Avantajı**: Hızlı eğitim, anlaşılabilirlik.
- **Dezavantajı**: Kelimeler arası bağlamı ve sıralamayı anlamaz.

Çözüm 2 – Word Embedding ile (Embedding + LSTM Modeli):

İkinci çözümde, TensorFlow'un `Embedding` katmanı ile her kelimeyi çok boyutlu bir vektörle temsil eden bir *Word Embedding* modeli kuruldu. Ardından sıralı veriler için uygun olan bir *LSTM (Long Short-Term Memory)* ağı kullanılarak duygu tahmini yapıldı.

- **Avantajı**: Kelimeler arası ilişkileri ve sıralamayı öğrenebilir.
- **Dezavantajı**: Daha fazla eğitim süresi ve veri gerektirir.

Grup Üyeleri Katkıları

Berke Abik-Word Embeddingsiz Çözüm

Serdar Mese-Word Embeddingli Çözüm

Sadık Karabulut-Rapor ve dökümantasyon

Kaynakça

1. **IMDB Veri Seti:**
   - **Hugging Face Datasets – IMDB Movie Reviews: https://huggingface.co/datasets/imdb**
2. **TF-IDF Açıklaması:**
   - **scikit-learn Documentation – TfidfVectorizer**
3. **Word Embedding Açıklamaları:**

- Mikolov, T. et al. (2013). "Efficient Estimation of Word Representations in Vector Space". arXiv preprint. https://arxiv.org/abs/1301.3781
- Pennington, J., Socher, R., Manning, C.D. (2014). "GloVe: Global Vectors for Word Representation". https://nlp.stanford.edu/projects/glove/

4. TensorFlow Embedding Layer Dokümantasyonu:
   - https://www.tensorflow.org/api_docs/python/tf/keras/layers/Embedding