**Doğal Dil İşleme (NLP) Uygulaması: Word Embedding Tabanlı ve Taban Olmayan Duygu Analizi Karşılaştırması**

**1. Giriş ve Problem Tanımı**

Bu projede amaç, doğal dil işleme (NLP) tekniklerini kullanarak kullanıcı yorumları üzerinden duygu (sentiment) analizinin gerçekleştirilmesidir. Bu bağlamda hem geleneksel yöntemler hem de kelime gömme (word embedding) teknikleri kullanılarak iki ayrı model geliştirilmiş, performansları karşılaştırılmıştır.

**2. Kullanılan Veri Seti**

Proje kapsamında kullanılan veri seti, IMDb Movie Reviews Dataset olarak bilinen 50.000 film yorumundan oluşan etiketli bir veri kümesidir. Her yorum pozitif ya da negatif bir duygu içerdiği şekilde sınıflandırılmıştır. Veri seti Kaggle üzerinden temin edilmiştir.

**Veri kümesi adı:** IMDB Movie Reviews Dataset (veya senin kullandığın isim)

**Kaynak:** https://www.kaggle.com/datasets/lakshmi25npathi/imdb-dataset-of-50k-movie-reviews

**İçeriği:** Film yorumları (text) ve bu yorumlara ait duygu etiketi (pozitif/negatif)

**Toplam veri sayısı:** 50.000 örnek

**Etiket dağılımı:** %50 pozitif – %50 negatif

**Veri setindeki başlıca sütunlar:**

* review: Film yorumu metni
* sentiment: Pozitif veya negatif olarak duygu etiketi

**3. Word Embedding Nedir?**

Word Embedding, kelimeleri çok boyutlu sayısal vektörlere dönüştüren, kelimeler arasındaki anlamsal ilişkileri bu vektörler aracılığıyla modellemeye yarayan bir temsildir. Bu teknik sayesinde benzer anlamlara sahip kelimeler, vektör uzayında birbirine yakın konumlanır.

**Word Embedding Kullanım Amaçları**

* Anlamsal benzerliği öğrenmek
* Kelime bağlamını hesaba katmak
* Metin sınıflandırma, çeviri, özetleme gibi NLP problemlerinde performansı artırmak

**4. Word Embedding Yöntemleri ve Çalışma Mantıkları**

**4.1 Word2Vec**

* Google tarafından geliştirilen bir Word embedding modelidir.
* CBOW (Continuous Bag of Words) ve Skip-Gram mimarilerini kullanır.
* Sözcüklerin bağlamlarına göre vektörler üretir.

**4.2 GloVe (Global Vectors for Word Representation)**

* Stanford tarafından geliştirilmiştir.
* Sözcük eşliği (co-occurrence) istatistiklerini temel alır.
* Global bilgiye dayalı daha stabil vektörler üretir.

**4.3 FastText**

* Facebook tarafından geliştirilmiştir.
* Kelime parçacıklarını (subword) dikkate alır.
* Duyulmamış (out-of-vocabulary) kelimeler için bile anlamlı vektörler üretebilir.

**5. Uygulama Adımları**

**Bu adımlar Colab dosyasında detaylı şekilde açıklanmıştır.**

**5.1 Word Embedding KULLANMADAN Çözüm**

* Metinler, temel önişlem adımlarından geçirildi (küçük harf, noktalama işaretlerinin silinmesi, stopword kaldırma vs.)
* TF-IDF vektörleştirme yöntemi kullanıldı.
* Logistic Regression sınıflandırıcısı ile model kuruldu.
* Elde Edilen Doğruluk Oranı: **%89.36**

**5.2 Word Embedding KULLANARAK Çözüm**

* FastText modeliyle kelime vektörleri çıkarıldı.
* Her cümle için ortalama vektör alınarak öznitelik çıkarımı yapıldı.
* Logistic Regression modeli ile sınıflandırma gerçekleştirildi.
* Hyperparameter tuning ile en iyi sonuç elde edildi.
* Elde Edilen Doğruluk Oranı (optimize edilmiş): **%88.03**

**6. Görselleştirme (projector.tensorflow.org)**

* FastText ile oluşturulan vektörler. tsv formatında kaydedildi.
* Bu vektörler projector.tensorflow.org sitesine yüklenerek görselleştirildi.

Bu bölümde, FastText modelimiz tarafından öğrenilen kelime vektörlerinin yüksek boyutlu uzayını iki farklı boyut indirgeme yöntemi kullanarak görselleştirdik: Temel Bileşenler Analizi (PCA) ve t-Dağıtılmış Stokastik Komşu Gömme (t-SNE). Amacımız, modelimizin kelimeler arasındaki anlamsal ilişkileri nasıl yakaladığını ve bu ilişkilerin görsel olarak nasıl temsil edildiğini incelemektir.

ekran görüntüsü içeren bir resim

Yapay zeka tarafından oluşturulan içerik yanlış olabilir.

ekran görüntüsü, metin içeren bir resim

Yapay zeka tarafından oluşturulan içerik yanlış olabilir.metin, ekran görüntüsü içeren bir resim

Yapay zeka tarafından oluşturulan içerik yanlış olabilir.metin, ekran görüntüsü içeren bir resim

Yapay zeka tarafından oluşturulan içerik yanlış olabilir.

PCA görselleştirmesi, kelime vektörlerinin genel dağılımını ve ana farklılıklarını anlamamıza yardımcı olurken, t-SNE görselleştirmesi kelimeler arasındaki daha ince anlamsal ilişkileri ve kümeleri ortaya koymada daha etkili görünmektedir. t-SNE'nin doğrusal olmayan boyut indirgeme yaklaşımı, yüksek boyutlu uzaydaki karmaşık ilişkileri düşük boyuta indirgerken daha başarılı olabilir. Özellikle duygu analizi gibi anlamsal benzerliklerin önemli olduğu görevlerde, t-SNE görselleştirmeleri daha anlamlı içgörüler sunabilir.

Sonuç olarak, her iki görselleştirme yöntemi de modelimizin kelime vektörlerini farklı açılardan incelememize olanak sağlamıştır. PCA genel bir bakış sunarken, t-SNE anlamsal ilişkilere daha derinlemesine bakmamızı sağlamıştır. Bu görseller, modelimizin kelime anlamlarını ve bağlamlarını öğrenme sürecini görsel olarak desteklemektedir.

metin, ekran görüntüsü, diyagram içeren bir resim

Yapay zeka tarafından oluşturulan içerik yanlış olabilir.

Bu görselde, Projector aracını kullanarak kelime vektörlerimiz arasında belirli bir kelimeyi, "love" kelimesini aradık. Arama sonucunda, "love" kelimesi görsel üzerinde kırmızı renkte vurgulanmıştır. Ayrıca, sol taraftaki arama sonuçları panelinde "love" kelimesine benzer yazılışa sahip veya potansiyel olarak ilişkili olabilecek diğer kelimeler de listelenmektedir ("lovecraft", "loved", "lovers", "beloved", "lover", "slovenia", "mirror").

Görseldeki "love" kelimesinin konumu, bu kelimenin modelimizin öğrendiği vektör uzayında diğer kelimelerle olan ilişkisini temsil etmektedir.

metin, ekran görüntüsü, diyagram içeren bir resim

Yapay zeka tarafından oluşturulan içerik yanlış olabilir.

Bu görselde, "friendship" (arkadaşlık) kelimesinin vektör uzayındaki en yakın komşularını kosinüs benzerliği kullanarak inceledik. Sağ taraftaki panelde listelenen kelimeler, modelimizin "friendship" kelimesiyle en çok ilişkilendirdiği veya benzer bağlamlarda kullandığı kelimelerdir.

Listenin en üstünde yer alan "friend" (arkadaş) ve "relationship" (ilişki) kelimeleri, "friendship" kavramıyla doğrudan ve güçlü bir anlamsal bağa sahip olmaları beklendiği üzere, en yakın komşular arasında yer almaktadır. "Ties" (bağlar), "collaboration" (iş birliği), "companionship" (yoldaşlık), "alliance" (ittifak), "fidelity" (sadakat), "mutual" (karşılıklı), "lifelong" (ömür boyu), "ally" (müttefik), "treaty" (antlaşma), "affinity" (yakınlık), "partnership" (ortaklık), "correspondence" (yazışma) gibi kelimeler de arkadaşlığın farklı yönlerini, niteliklerini veya arkadaşlıkla sıkça ilişkilendirilen durumları temsil etmektedir.

"Concepts" (kavramlar) kelimesinin de listede yer alması, "friendship" kelimesinin daha soyut bir düzeyde ele alındığını veya modelin "friendship"i bir kavram olarak diğer kavramlarla ilişkilendirdiğini düşündürmektedir.

Genel olarak, "friendship" kelimesinin en yakın komşuları incelendiğinde, modelin bu kavramla tutarlı ve anlamlı ilişkiler kurduğu görülmektedir. Listede yer alan kelimelerin çoğu, arkadaşlığın farklı boyutlarını ve özelliklerini yansıtmaktadır. Bu görselleştirme, modelimizin "friendship" kavramını nasıl anladığına ve hangi diğer kelimelerle ilişkilendirdiğine dair değerli bir sonuç sunmaktadır.

metin, ekran görüntüsü, diyagram, öykü gelişim çizgisi; kumpas; grafiğini çıkarma içeren bir resim

Yapay zeka tarafından oluşturulan içerik yanlış olabilir.

Bu görselde, modelimizin "happy" (mutlu) kelimesini vektör uzayında hangi diğer kelimelerle en yakın ilişki içinde gördüğünü kosinüs benzerliği metriği kullanarak inceledik. Sağ tarafta sıralanan kelimeler, "happy" kavramıyla en sık birlikte kullanılmış veya model tarafından benzer anlamsal bağlamlarda değerlendirilmiş olabilecek kelimelerdir.

Listenin başında yer alan "glad" (memnun) ve "funny" (komik) kelimeleri, mutluluk duygusuyla sıklıkla ilişkilendirilen veya mutluluğa yol açabilen duygusal durumları ve nitelikleri temsil etmektedir. "You" (sen) ve "love" (sevgi) gibi kelimelerin yakınlığı, mutluluğun kişiler arası ilişkiler ve sevgi duygusuyla olan bağlantısını yansıtabilir. "Remember" (hatırlamak) ve "king" (kral) gibi kelimelerin listede yer alması, mutluluğun anılarla veya olumlu figürlerle olan çağrışımlarından kaynaklanabilir.

"Lucky" (şanslı), "kids" (çocuklar), "baby" (bebek), "young" (genç), "beautiful" (güzel), "want" (istemek), "girl" (kız), "couple" (çift), "proud" (gururlu), "friend" (arkadaş) gibi kelimelerin de listede yer alması, mutluluğun çeşitli yaşam deneyimleri, ilişkiler ve olumlu niteliklerle olan yaygın bağlantısını göstermektedir. Genel olarak, "happy" kelimesinin en yakın komşuları incelendiğinde, modelin bu temel pozitif duyguyla tutarlı ve çeşitli olumlu kavramları, ilişkileri ve deneyimleri ilişkilendirdiği görülmektedir.

**KARŞILAŞTIRMA**

**Klasik yöntemler (TF-IDF)**, küçük veri kümelerinde hızlı ve etkili sonuçlar verebilir.

**FastText embedding’leri**, metinlerin anlamsal bilgisini daha iyi yakalasa da daha fazla eğitim süresi ve ince ayar gerektirmiştir.

En iyi doğruluk oranı klasik yöntemde elde edilse de, **FastText modeli daha esnek, genelleştirilebilir ve gelecekte başka veri kümelerine daha iyi adapte olabilecek** yapıdadır.

Performans farkı küçük olsa da embedding yönteminin daha modern ve ileri seviye NLP problemlerinde kullanılabilir olması, tercih sebebi olabilir.

Projemizin farklı aşamalarında, IMDB duygu analizi veri setini sınıflandırmak için iki temel yaklaşım izledik: geleneksel özellik mühendisliği (TF-IDF) ve kelime gömütmeleri (word embeddings).

**Embedding Kullanmadan Elde Edilen Sonuç (%89.36 Doğruluk):**

Metin verilerini sayısal özelliklere dönüştürmek için TF-IDF (Term Frequency-Inverse Document Frequency) yöntemini kullandığımızda, Lojistik Regresyon sınıflandırıcısı ile %89.36'lık bir doğruluk oranı elde ettik. Bu sonuç, kelime frekanslarının ve ters belge frekanslarının, bu özel duygu analizi görevinde oldukça güçlü bir sinyal sağladığını göstermektedir. Model, kelimelerin doküman içindeki önemini ve genel yaygınlığını dikkate alarak, duygusal polariteyi etkili bir şekilde ayırt edebilmiştir. Bu yaklaşım, basit ve yorumlanabilir özellikler sunması açısından avantajlıdır.

**Embedding Kullanarak Elde Edilen Sonuçlar (%87.82- GridSearchCV, %88.03 - RandomizedSearchCV):**

Word embeddings yaklaşımında ise, FastText modelini kullanarak kelimeler için yoğun, düşük boyutlu vektörler öğrendik. Bu vektörler, kelimelerin anlamsal ilişkilerini yakalamayı amaçlamaktadır. Bu embedding'leri kullanarak eğittiğimiz Lojistik Regresyon modelinde, hiperparametre optimizasyonu ile farklı sonuçlar elde ettik:

GridSearchCV ile en iyi doğruluk oranı %87.82 olarak gerçekleşti. RandomizedSearchCV ile ise %88.03'lük bir doğruluk oranına ulaştık.

**Karşılaştırma ve Değerlendirme:**

Elde ettiğimiz sonuçlara göre, geleneksel TF-IDF yaklaşımının (%89.36), word embedding yaklaşımının en iyi sonuçlarına (%88.03) kıyasla bu özel duygu analizi görevinde biraz daha yüksek bir performans gösterdiği görülmektedir. Aradaki yaklaşık %1.33'lük fark, birkaç olası nedenden kaynaklanabilir (veri seti özellikleri, basit cümle temsili, model karmaşıklığı, FastText eğitimi gibi).

Hiperparametre optimizasyonu (özellikle RandomizedSearchCV ile) word embedding kullanılarak elde edilen performansı artırmış olsa da TF-IDF'in basit ama etkili yaklaşımı bu veri setinde daha iyi sonuç vermiştir.

**Sonuç:**

Bu karşılaştırma, farklı özellik çıkarma yöntemlerinin farklı NLP problemleri üzerindeki etkisini göstermektedir. Bu özel duygu analizi görevinde, TF-IDF'in güçlü bir temel çizgi oluşturduğu görülmüştür. Word Embedding ise, anlamsal bilgiyi yakalama potansiyeline sahip olmasına rağmen, bu ilk denemelerde TF-IDF kadar yüksek bir doğruluk sağlayamamıştır. Ancak, daha gelişmiş embedding teknikleri ve sınıflandırma modelleri kullanılarak bu tür görevlerde daha iyi performans gösterme potansiyeli bulunmaktadır. Bu proje, farklı yaklaşımların güçlü ve zayıf yönlerini anlamamız açısından değerli bir deneyim olmuştur.

**Grup Üyelerinin Katkısı**

|  |  |  |
| --- | --- | --- |
| **Grup Üyesi** | **Görevler** | **Katkı Oranı** |
| Hatice OGUR | FastText modeli, optimizasyon, rapor yazımı ve görselleştirme | %33.33 |
| Fatma Nur AKBAŞ | Veri temizleme, TF-IDF modeli | %33.33 |
| Nisa Sümeyye ALBAYRAK | FastText modeli, optimizasyon farklı parametrelerle deneme | %33.33 |

**GitHub Linki**

<https://github.com/HaticeOgur/word-embedding-nlp-proje>

**Kaynakça**

NLTK, Scikit-learn, Pandas, NumPy dökümantasyonları

IMDB Dataset – https://www.kaggle.com/datasets/lakshmi25npathi/imdb-dataset-of-50k-movie-reviews

Google Colaboratory Documentation - https://colab.research.google.com/