YAPAY ZEKA BİTİRME PROJESİ OTOMOBİL YORUM ANALİZİ
TOLGA TARHAN
2102131031

1. GİRİŞ
1.1 Ödevin Amacı

Bu çalışmanın temel amacı, Türkçe metinler arasında anlam temelli benzerlik ölçümü yapabilen yöntemlerin karşılaştırmalı analizini gerçekleştirmektir. Günümüzde dijital ortamda üretilen metin sayısının hızla artmasıyla birlikte, metin madenciliği ve bilgi çıkarımı tekniklerine olan ihtiyaç artmıştır. Bu bağlamda, farklı metinler arasında benzerlik tespiti; öneri sistemleri, arama motorları, belge sınıflandırma ve anlam kümeleme gibi birçok doğal dil işleme (NLP) görevinde kritik rol oynamaktadır.

Bu projenin özel hedefi; araç alım-satımına ilişkin Türkçe içeriklerin anlam düzeyinde benzerliğini ölçmek, farklı vektörleme tekniklerini uygulayarak bu içerikler arasından giriş cümlesine en yakın olanları tespit etmek ve sonuçları karşılaştırmalı olarak değerlendirmektir.

1.2 Kullanılan Veri Seti

Veri seti, otomobil alım-satım süreçlerine dair kullanıcı yorumlarını içeren Türkçe cümlelerden oluşmaktadır. Bu metinler forum yazışmaları, kullanıcı soruları ve tavsiye yazılarından alınmış olup, gerçek hayat diliyle yazıldığından dolayı çeşitli yazım bozuklukları, argo ifadeler ve dilbilgisi hataları içermektedir. Bu da modelleme sürecine doğal zorluklar eklemiştir.

Veri ön işleme aşamasında iki ayrı metin seti oluşturulmuştur:

Lemmatized Set (Sözlük tabanlı kökler): Kelimeler TDK tabanlı sözcük formlarına dönüştürülerek bağlamsal bütünlük korunmuştur.
Stemmed Set (Yüzeysel kök alma): Simple Turkish Stemmer gibi araçlarla kelimelerin gövdeleri çıkarılmıştır; fakat bu yaklaşım anlam bütünlüğünü zayıflatabilir.
Bu iki farklı yaklaşımın performanslarını kıyaslayabilmek adına tüm deneyler her iki veri seti için de tekrarlanmıştır.

2. YÖNTEM
2.1 Benzerlik Ölçme Yöntemleri

Benzerlik hesaplamalarında iki temel vektörleme ve karşılaştırma yöntemi kullanılmıştır:

1. TF-IDF (Term Frequency - Inverse Document Frequency)
Her kelimenin bir cümledeki önemini yansıtan bu yöntem, metinleri sabit boyutlu vektörlere çevirir. Benzerlik skoru, iki vektör arasındaki kozünüs benzerliği hesaplanarak belirlenmiştir. Avantajı hız ve sadeliktir; dezavantajı ise bağlamı dikkate almamasıdır.

2. Word2Vec

Gensim kütüphanesi kullanılarak kelimeler bağlamlarıyla birlikte vektörleştirilmiştir. CBOW ve Skip-gram mimarileri, farklı pencere boyutları (2, 4) ve vektör boyutları (100, 300) ile eğitilmiş toplam 16 model kullanılmıştır. Her cümle, içerdiği kelimelerin vektörlerinin ortalaması alınarak temsil edilmiştir.

Ayrıca her modelin sonuçlarının benzerliği Jaccard benzerlik skoru ile karşılaştırılmıştır. Bu analiz, hangi modellerin aynı veya benzer cümleleri önerdiğini ölçmek için yapılmıştır.

2.2 Uygulanan Parametreler

Word2Vec modelleri aşağıdaki konfigürasyonlarla eğitilmiştir:

Model Tipi	Pencere Boyutu	Vektör Boyutu
CBOW	2, 4	100, 300
Skip-gram	2, 4	100, 300

Bu 8 kombinasyon, hem lemmatized hem stemmed veriler için ayrı ayrı uygulanmıştır. Böylece toplam 16 model oluşturulmuştur. Bu modellerin her biri için giriş cümlesine benzer ilk 5 cümle elde edilmiştir.

3. SONUÇLAR VE DEĞERLENDİRME
3.1 Benzerlik Sonuçlarının Puanlanması

Her bir modelin ürettiği beş benzer cümle, içerik açısından manuel olarak 1 ile 5 arasında puanlanmıştır. Bu değerlendirme, modelin anlamsal yakınlık başarısını ölçmek için yapılmıştır. Puanlama sonuçları, “model_bazli_benzerlik_puanlari.xlsx” dosyasında detaylı şekilde yer almaktadır.

Yapılan gözlem ve puanlamalara göre:

Skip-gram modelleri, CBOW modellerine göre daha iyi sonuçlar üretmiştir.
Lemmatized veri kullanılarak eğitilen modeller, stemmed veriye göre daha başarılıdır.
En başarılı modeller, skip-gram + window 4 + vector_size 300 yapılandırmalarıdır.
3.2 Jaccard Benzerlik Analizi

“model_model_jaccard_matrisi.xlsx” dosyasındaki analiz, farklı modellerin önerdiği cümle kümelerinin ne kadar örtüştüğünü göstermektedir. Bu analiz, modellerin çeşitlilik açısından değerlendirilmesine yardımcı olur.

Jaccard skorlarının genel olarak yüksek olması, modellerin benzer cümleleri önerdiğini göstermektedir.
Bunun olası nedeni, veri kümesinin sınırlılığı ve bazı cümlelerin diğerlerinden çok daha belirgin şekilde giriş cümlesine yakın olmasıdır.
3.3 TF-IDF vs Word2Vec Karşılaştırması

TF-IDF yöntemi hızlı ve hesaplama açısından verimli olsa da, anlam bütünlüğünü göz ardı ettiği için önerdiği cümleler çoğunlukla kelime düzeyinde benzerlik taşımaktadır.

Word2Vec modelleri ise bağlamsal benzerliği daha doğru şekilde yakalayarak anlamlı sonuçlar üretmiştir.
En yüksek puanlı Word2Vec modelinde önerilen cümleler, yalnızca aynı kelimeleri değil, aynı niyet ve bağlamı paylaşan ifadeleri içermektedir.

4. GENEL DEĞERLENDİRME VE ÖNERİLER

4.1 Genel Çıkarımlar

Word2Vec, Türkçe gibi dilbilgisel olarak zengin dillerde bağlamsal anlamı yakalamada daha başarılıdır.
Lemmatization, kök alımına kıyasla daha doğru bağlam yakalanmasına olanak tanımıştır.
En başarılı yapılandırma: Skip-gram + lemmatized + window=4 + dim=300
TF-IDF, kaynak sınırlı sistemler veya hızlı uygulamalar için hâlâ makul bir alternatiftir.

4.2 Gelecek Çalışmalar İçin Öneriler

Veri seti genişletilmeli, farklı senaryolarda testler yapılmalıdır.
Derin öğrenme tabanlı modeller (BERT, RoBERTa vb.) ile karşılaştırmalar yapılabilir.
Anlam benzerliği kadar duygu benzerliği, kullanım amacı benzerliği gibi kriterlerle çoklu değerlendirme yapılabilir.
Benzerlik analizleri, otomatik öneri sistemlerinde, belge kümeleme ve chatbotlarda entegre biçimde kullanılabilir.