Hukuk Projesi: Tanık İfadesi - Dava Metni Eşleştirme

1. Giriş

Bu projede, tanık ifadeleri ile dava metinleri arasındaki anlamsal benzerliklerin ölçülmesi hedeflenmiştir. Doğal dil işleme (NLP) teknikleri kullanılarak metinler temizlenmiş, çeşitli vektörleştirme yöntemleriyle sayısallaştırılmış ve benzerlikleri çeşitli metriklerle hesaplanmıştır.

2. Veri İşleme

- **Metin Temizleme:** Noktalama işaretleri ve durak kelimeler (stopwords) çıkarılmış, metinler küçük harfe dönüştürülmüştür.
- **Tokenization:** Metinler kelime bazında ayrıştırılmıştır.
- **Lemmatization & Stemming:** Her iki yöntemle metinler ayrı ayrı işlenmiş, farklı varyasyonlarla modeller oluşturulmuştur.

3. TF-IDF Analizi

- **Yöntem:** Tanık ifadeleri ve dava metinleri TF-IDF yöntemi ile vektörleştirilmiştir.
- **Benzerlik:** Cosine similarity ile karşılaştırılmıştır.
- **Sonuç:** TF-IDF yönteminin ortalama Jaccard skoru düşüktür (~0.111), bu da bağlamsal anlamı sınırlı yakalayabildiğini göstermektedir.

4. Word2Vec Modelleme

Farklı yapılandırmalarda CBOW ve SkipGram modelleri eğitilmiştir:

Kullanılan Modeller:

- CBOW vs SkipGram
- Window Size: 2 ve 4
- Vektör Boyutu: 100 ve 300
- Her bir kombinasyon için hem **lemmatized** hem **stemmed** versiyonları oluşturulmuştur.

5. Model Yapılandırma Yorumları

Model başarımı, kullanılan mimari (CBOW veya SkipGram), pencere genişliği (window size) ve vektör boyutu (dimension) gibi parametrelerden önemli ölçüde etkilenmiştir:

- **CBOW modelleri**, genel olarak SkipGram'a kıyasla daha yüksek başarı göstermiştir.
- **Pencere genişliği 2 olan modeller**, pencere 4 olanlara göre daha başarılı sonuçlar vermiştir.
- **100 boyutlu vektörler**, 300 boyutlu modellere göre genellikle daha iyi sonuç vermiştir.

Bu nedenle, `CBOW_win2_dim100_lemma` ve `CBOW_win2_dim100_stem` modelleri en başarılı sonuçları üretmiş ve ortalama Jaccard skorları 0.267 olarak gözlemlenmiştir.

6. Jaccard Benzerliği

Her modelin ilk 5 örnek kelimesi kullanılarak Jaccard benzerlik matrisi oluşturulmuş ve ortalamalar hesaplanmıştır. Bu benzerlik, kavramsal anlam yakınlıklarını gözlemlemek için kullanılmıştır.

6.1 Jaccard Benzerlik Skorları – Detaylı Tablo

Model	Jaccard Skor
CBOW_win2_dim100_lemma	**0.267**
CBOW_win2_dim100_stem	**0.267**
CBOW_win2_dim300_lemma	0.214
CBOW_win2_dim300_stem	0.214
SkipGram_win4_dim300_stem	0.198
SkipGram_win4_dim300_lemma	0.198
CBOW_win4_dim300_lemma	0.172
CBOW_win4_dim300_stem	0.172
CBOW_win4_dim100_lemma	0.170
CBOW_win4_dim100_stem	0.170
SkipGram_win4_dim100_stem	0.162
SkipGram_win4_dim100_lemma	0.162
SkipGram_win2_dim100_stem	0.140
SkipGram_win2_dim100_lemma	0.140
TFIDF_lemma	0.111
TFIDF_stem	0.111
SkipGram_win2_dim300_stem	0.102
SkipGram_win2_dim300_lemma	0.102

7. Sonuç ve Öneriler

Bu çalışmada tanık ifadeleri ile dava metinlerinin benzerliği çeşitli vektörleme ve benzerlik ölçüm yöntemleriyle değerlendirilmiştir. Özellikle Word2Vec tabanlı modellerin, TF-IDF gibi geleneksel yöntemlere göre daha başarılı olduğu görülmüştür.

- **CBOW modeli**, dar pencere genişliği ve düşük boyutlu vektörlerle birlikte kullanıldığında en yüksek başarıyı göstermiştir.
- **TF-IDF yöntemleri**, temel benzerlik hesapları için uygun olsa da anlamsal bağlamı yakalamada yetersiz kalmıştır.
- **Jaccard metrikleri**, benzerlik ölçümünde sınırlı kalabilse de modeller arasında göreli kıyaslama için yeterlidir.