TWİTTER VERİLERİ ÜZERİNDEN DUYGU ANALİZİ

Muhammet Koç
muhammetkoc2019@gtu.edu.tr
İsmail Akkoyun
i.akkoyun2019@gtu.ed.tr
[2]Elektronik Mühendisliği Bölümü, GTÜ, Kocaeli, Türkiye

I. GİRİŞ

Duygu Analizi, verilen bir metnin duygu durumunu sınıflandırma görevidir. Örneğin, metin tabanlı bir tweet, "olumlu", "olumsuz" veya "tarafsız" olarak kategorize edilebilir. Metin ve eşlik eden etiketler verildiğinde, bir model doğru duygu durumunu tahmin etmek üzere eğitilebilir.

Duygu Analizi teknikleri, makine öğrenimi yaklaşımları, sözlük tabanlı yaklaşımlar ve hibrit yöntemler olmak üzere çeşitli kategorilere ayrılabilir. Duygu analizi araştırmasının alt kategorileri arasında multimodal duygu analizi, yön bazlı duygu analizi, ince detaylı görüş analizi, dil özel duygu analizi bulunmaktadır.

1) *Makine Öğrenimi Tabanlı Yöntemler:*
   - Doğal Dil İşleme (NLP) Modelleri (CNN, LSTM)
   - Destek Vektör Makineleri (SVM)
   - Karar Ağaçları ve Random Forest

2) *Sözlük Tabanlı Yöntemler:*
   - Duygu Sözlükleri
   - Duygu Puanlama Yöntemleri.

3) *Hibrit Yöntemler:*
   - Makine Öğrenimi ve Sözlük Birleşimi

Ayrıca yöntemleri gözetimli gözetimsiz olarak listelersek:

1. *Gözetimli Modeller:*
   - Sınıflandırma Modelleri: Bu modeller, belirli bir metni önceden tanımlanmış duygu kategorilerine sınıflandırmak için kullanılır. Örneğin, olumlu, olumsuz veya tarafsız.

     Örnek algoritmalar: Destek Vektör Makineleri (SVM), Karar Ağaçları, Derin Sinir Ağları (CNN, LSTM).
   - Regresyon Modelleri: Bu modeller, metindeki duygusal yoğunluğu belirlemek için kullanılır. Metnin içerdiği duygu düzeyini sürekli bir değer olarak tahmin eder.

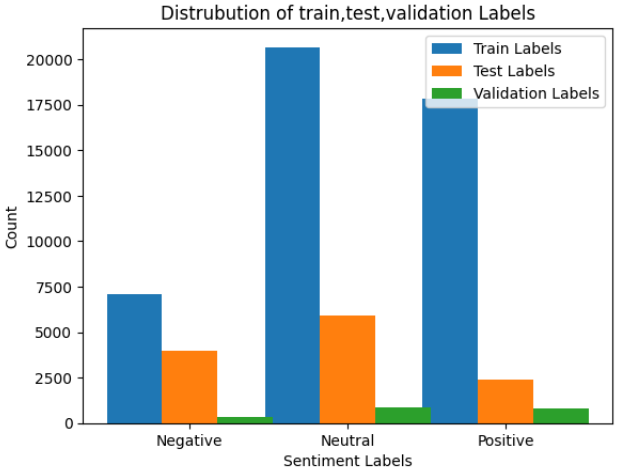
     Örnek algoritmalar: Lineer Regresyon, Destek Vektör Regresyonu.

Gözetimli öğrenme yöntemleri kullanılarak uygulama gerçekleştirilmiştir. Kullanılan algoritmalar SVM, K-Nearest-Neighbor ve LSTM ile sonuçlar elde edilmiştir. K-means algoritması ile duygu analizi yapılamaz. K-Means Algoritması gözetimsiz bir yöntemdir. K-means ile yazılan makaleler bulunsa da itibar edilmemesi gereken bir yaklaşımdır.

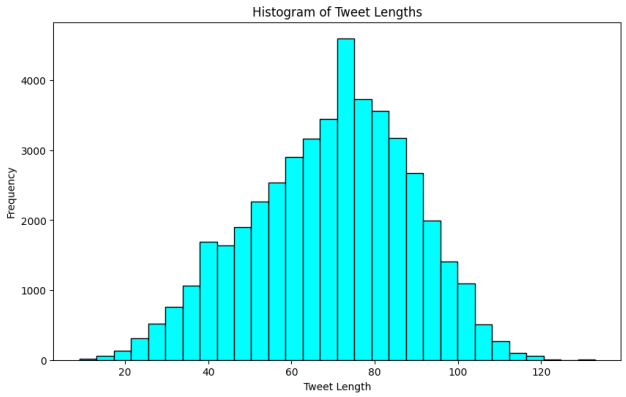
II. METEDOLOJİ

A. *Veriye ön işlem uygulamak*

Kullanılan veri seti TweetEval veri setinin Sentiment alt setidir. Veriyi yükleyip, Metin içinde sayı filtresi, noktalama işaretleri filtresi, N-karakter filtresi (N karakterden küçük tokenlar), durak kelimeleri(stop words) filtresi uygulanmıştır. Tokenization gerekli olduğu yerlerde Bert modelinin tokenizer modülü kullanılmıştır. Şekil 1'de veri seti içinde verilerin dağılımı gösterilmektedir. Train veri seti içinde 45615 etiketli tweet bulunmaktadır.



Şekil 1:verilerin dağılımı



Şekil 2: Tweetlerin sahip olduğu karakterlere sayılarına göre dağılımı

Şekil 2'de Tweetlerin sahip olduğu karakterlere sayılarına göre dağılımı gösterilmektedir.

Şekil 3: Veri seti için kelime bulutu

Şekil 3'te veri seti içinde bulunan kelimelerin sıklıklarına göre oluşturulan kelime bulutu gösterilmektedir. Buradan en sık geçen kelimenin 'user' olduğu anlaşılmaktadır.

B. Özellik Çıkarımı(Feature Extraction)

Özellik çıkarımı için Metinleri numerik olarak ifade etme yöntemlerinden TfİDF, Word2vec ve Countvector yöntemleri yanı sıra Sentiwordnet gibi sözlüklerden oluşturulan tokenlardan elde skorlar ile özellikler çıkarılmıştır.

C. SVM Algoritmasının uygulanması

SVM (Support Vector Machine), özellikle sınıflandırma ve regresyon problemleri için kullanılan bir makine öğrenimi algoritmasıdır. SVM, veri noktalarını sınıflarına ayırmak için bir hiper düzlem oluşturur. Bu hiper düzlem, veri noktalarını sınıflarına göre iki farklı tarafına böler.

SVM'nin temelde iki ana türü vardır: doğrusal SVM ve çekirdek SVM. Doğrusal SVM, veri noktalarını doğrusal bir hiper düzlemle ayırırken, çekirdek SVM, veriler arasında daha karmaşık ilişkileri modellemek için bir çekirdek fonksiyonu kullanır. Uygulamada Doğrusal SVM kullanılmıştır. Şekil 4'te SVM sonuçları gösterilmiştir

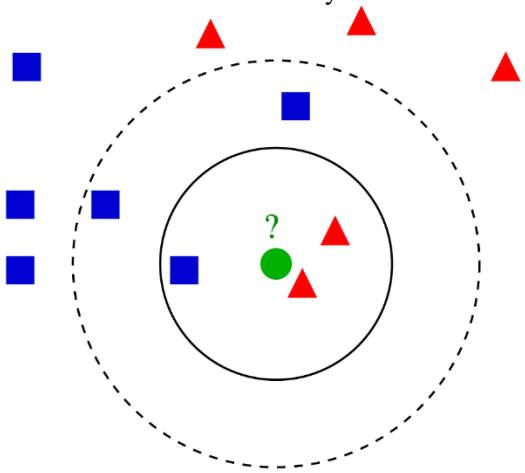
```
Accuracy: 0.5752197981113644
Classification Report:
              precision    recall  f1-score   support

           0       0.64      0.33      0.43      3972
           1       0.57      0.77      0.66      5937
           2       0.53      0.50      0.51      2375

    accuracy                           0.58     12284
   macro avg       0.58      0.53      0.53     12284
weighted avg       0.59      0.58      0.56     12284
```

Şekil 4: SVM sonuçları

D. KNN Algoritmasının uygulanması

KNN (K-Nearest Neighbors) Algoritması iki temel değer üzerinden tahmin yapar;

- Distance (Uzaklık): Tahmin edilecek noktanın diğer noktalara uzaklığı hesaplanır. Bunun için Minkowski uzaklık hesaplama fonksiyonu kullanılır.
- K (komuşuluk sayısı): En yakın kaç komşu üzerinden hesaplama yapılacağını belirtir. K değeri sonucu direkt etkileyecektir. K=1 olursa overfit etme olasılığı çok yüksek olacaktır. Çok büyük olursa da çok genel sonuçlar verecektir. Bu sebeple optimum K değerini tahmin etmek problemin asıl konusu olarak karşımızda durmaktadır. K değerinin önemini aşağıdaki Şekil 5 güzel bir şekilde göstermektedir. Eğer K=3 (düz çizginin olduğu yer) seçersek sınıflandırma algoritması ? işareti ile gösterilen noktayı, kırmızı üçgen sınıfı olarak tanımlayacaktır. Fakat K=5 (kesikli çizginin olduğu alan) seçersek sınıflandırma algoritması, aynı noktayı mavi kare sınıfı olarak tanımlayacaktır.



Şekil 5:KNN k parametresi için gösterim

KNN (K-Nearest Neighbors) Algoritması ile üretilmiş bir modelin başarımını ölçmek için genel olarak kullanılan 3 adet indikatör vardır.

- **Jaccard Index**: Doğru tahmin kümesi ile gerçek değer kümesinin kesişim kümesinin bunların birleşim kümesine oranıdır.1 ile 0 arası değer alır. 1 en iyi başarım anlamına gelir.
- **F1-Score**: Confusion Matriks üzerinden hesaplanan Precission ve Recall değerlerinden hesaplanır. Pre=TP/(TP+FP) Rec=TP/(TP+FN) F1-Score= 2((PreRec)/(Pre+Rec)) 1 ile 0 arası değer alır. 1 en iyi başarım anlamına gelir.
- **LogLoss**: Logistic Regresyon sonunda tahminlerin olasılıkları üzerinden LogLoss değeri hesaplanır. 1 ile 0 arası değer alır. Yukarıdaki iki değerden farklı olarak 0 en iyi başarım anlamına gelir.

Sonuçlarda F1 skor parametresi kullanılmıştır.

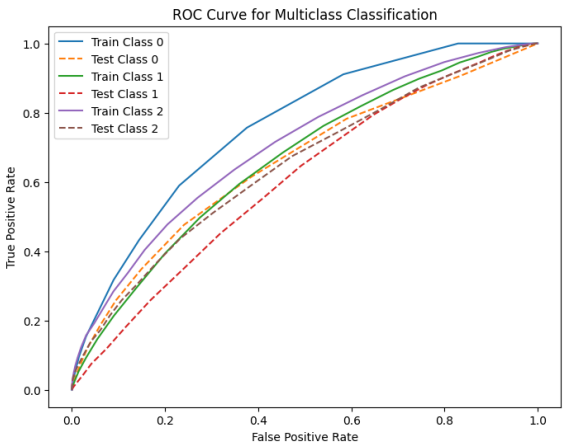
Uygulama yapılırken Feature vektörü için hem TFIDF hemde CountVector yöntemleri ile vektörler elde edilmiştir.

```
Accuracy: 0.48787040052100294
Classification Report:
              precision    recall  f1-score   support

           0       0.53      0.09      0.15      3972
           1       0.50      0.84      0.63      5937
           2       0.39      0.27      0.32      2375

    accuracy                           0.49     12284
   macro avg       0.48      0.40      0.37     12284
weighted avg       0.49      0.49      0.41     12284
```

Şekil 6: TFIDF feature vektörü ile elde edilen KNN sonuçları

```
Best Parameters: {'n_neighbors': 40, 'metric': 'manhattan'}
Best Accuracy: 0.4992475588459236
```

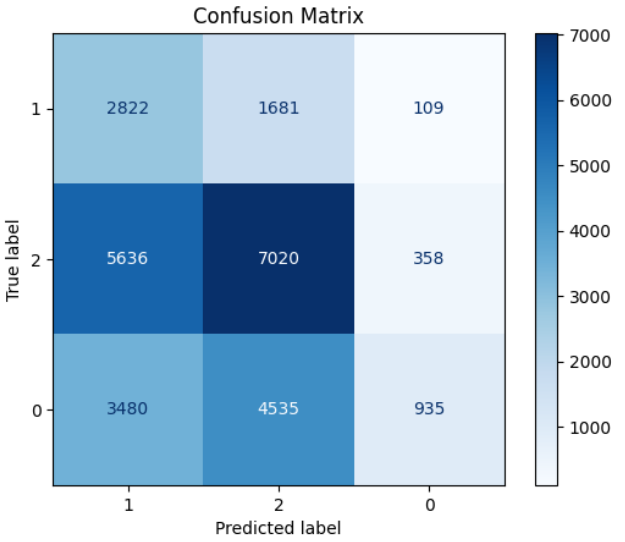
Şekil 7: Countvector ile elde edilen KNN algoritması sonuçları



Şekil 8:Countvector ile KNN Roc Curve

E. Sentiwordnet Duygu Sözlüğü Kullanımı

Sentiwordnet ile tokenların duygu skoruna göre pozitifskor pozitif, negatif skor negatif olarak değerlendirildiğinde; sentiwordnet ile doğruluk Oranı: 0.41



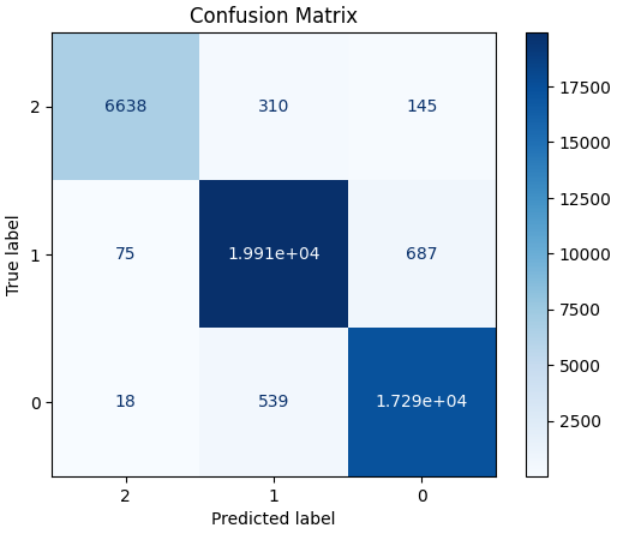
Şekil 9: sentiwordnet ile elde edilen matris

F. LSTM Algoritmasının Uygulanması

LSTM: Eğitim sırasında model, metin verilerindeki karmaşık ilişkileri öğrenir.

LSTM metin verilerindeki belirgin özellikleri öğrenme yeteneği ile bilinir ve duygu analizi için genellikle iyi performans gösterir.

LSTM ile Doğruluk Oranı: 0.96



Şekil 10:LSTM etiketleri ile elde edilen etiketlerin karmaşıklık matrisi

ELM472 Makine Öğrenmesinin Temelleri

III. SONUÇLAR

A. *Küme Sayısı*

Duygu analizi için kümeler pozitif, negatif ve nötr olarak seçildiği için küme sayısı parametresi k=3 olarak kullanılmıştır.

B. *Küme görselleştirmesi*

Train veri setinde bulunan gerçek etiketler Şekil 2'de gösterilmiştir. Veri setinde 45615 tweet bulunmaktadır.

C. *Modellerin Karşılaştırılması*

Duygu analizi için kullanılabilecek birçok yöntem bulunmaktadır. En iyi model olarak söylenebilecek bir yöntem yoktur. Uygulamanın gereksinimlerine göre modeller seçilmelidir.
Ayrıca K-means gibi gözetimsiz modeller duygu analizi için kullanılmaz

KAYNAKÇA

[1] E. Alpaydin, Introduction to Machine Learning, 3. bs. Cambridge, MA, USA: MIT Press, 20[1] E. Alpaydin, Introduction to Machine Learning, 3. bs. Cambridge, MA, USA: MIT Press, 2014.14.