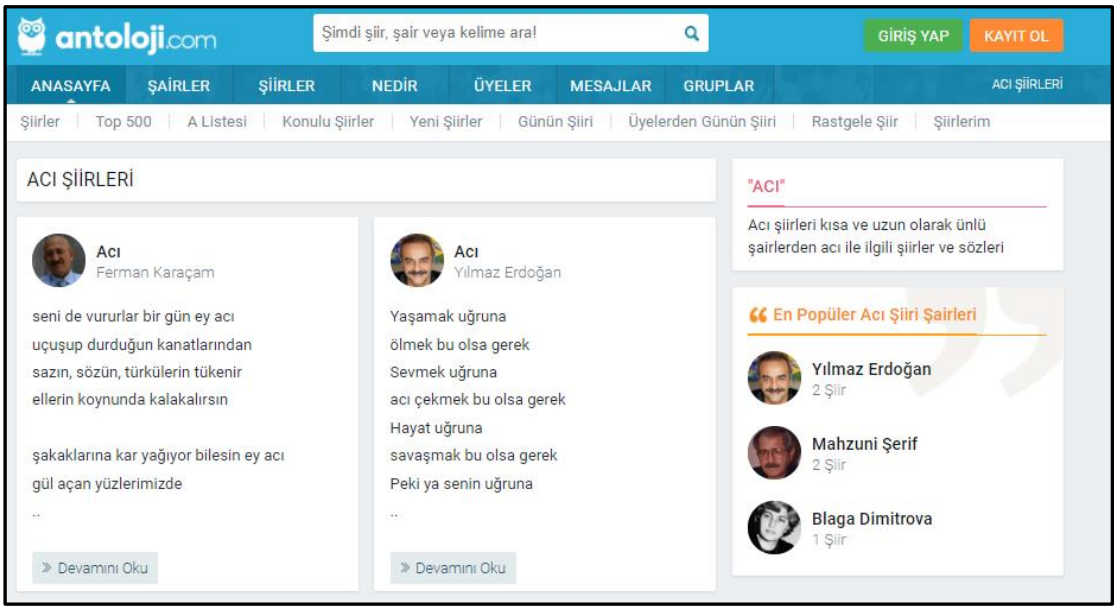
Konu : Türkçe Şiirlerde Konu Tespiti

İrem Atılgan (Ç17061036), Hüma Bilgin (18011087)

Veriler Hakkında

Veriler, antoloji.com sitesinde yer alan şiirlerden alınmıştır. Burada yaklaşık 140-150 konu üzerinden, 10 temel şiir konusu seçilmiş ve BeautifulSoup, Selenium kütüphaneleri kullanılarak Python üzerinden crawl edilmiştir.

Şiir konuları "Acı","Neşe","Özlem","Aşk","Hayat","Hüzün","Doğa","Spor","Vatan","İnanç" olarak seçilmiştir.



Oluşturulan verisetinde şiir içeriği, konusu ve şairin ismi yer almaktadır. Her konudan 200 şiir olmak üzere toplamda 2000 şiir verisi elde edilmiştir.

	topic	text	author	label
0	Acı	seni de vururlar bir gün ey acı uçuşup durduğun kanatlarından sazın sözür	Ferman Karaçam	0
1	Acı	Yaşamak uğruna ölmek bu olsa gerek Sevmek uğruna acı çekmek bu olsa	Yılmaz Erdoğan	0
2	Acı	Ve bir kadın Bize acıdan bahset dedi Ve o cevap verdi Acınız anlayışınızı s	Halil Cibran	0
3	Acı	Bir tek dileğim var mutlu ol yeter sözünün bir kamyon yükü anlam taşıdığ	Yılmaz Erdoğan	0
4	Acı	Çilelerim köprü oldu Tuna ya Dilimden anlamaz kulun Almanya Döneceği	Ozan Arif	0
5	Acı	anlatmak istedikçe herseyi birden yitiriyorum bir kutupyildizi bir ben bir d	Hasan Hüseyin Korkmazgil	0
6	Acı	Beni böyle deli eden Yarin açı sözü imiş Sırat sırat dedikleri Bir çift ela göz	Aşık Sefai	0

Verilerin İşlenmesi

Siteden elde edilen şiirler üzerindeki HTML tag'leri, özel karakterler ('\n', '\b', '\r' vb.) ve noktalama işaretleri çıkarılmıştır.

Konular LabelEncoder ile 0-9 aralığında numaralandırılacak şekilde düzenlenmiştir.

Random Forest, Naïve Bayes ve Support Vector Machine yöntemleri için şiirler içerisindeki stop words'ler çıkarılmıştır. Bu yöntemler üzerinde eğitim gerçekleştirilirken metinin encode edilmesinde CountVectorizer ve TF-IDF yöntemleri denenmiştir.

2000 Şiir verisinin %80'i eğitim (1600) , %20'si test (400) için kullanılmıştır.

SONUÇLAR

Metin sınıflandırmada en sık kullanılan 4 yönteme (Random Forest, Naïve Bayes, Support Vector Machine, BERT) yer verilmiş ve sonuçları incelenmiştir. Bununla birlikte ilk üç yöntem, metinlerin encode edilme yöntemine bağlı olarak iki farklı sonuçla ele alınmıştır.

1. RANDOM FOREST

Random forest algoritması, özniteliklerim önem derecesini tahmin eder. Classification algoritmaları arasında en iyilerden biri olarak belirtilmektedir.

Problemimizin çözümünde ise Random Forest yöntemi için ağaç sayısı 1000 olarak belirlenmiştir (num_estimators).

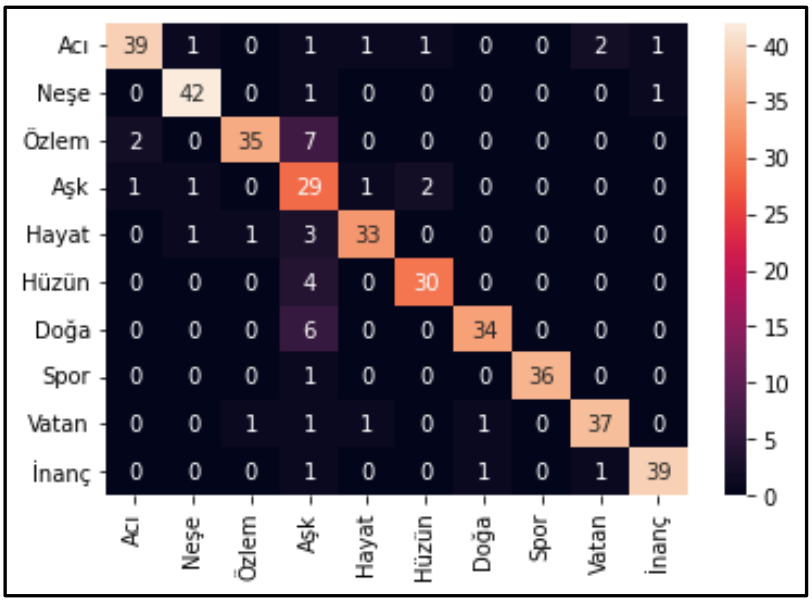
CountVectorizer

Accuracy = 0.885

Classification Report

	precision	recall	f1-score	support
0	0.93	0.85	0.89	46
1	0.93	0.95	0.94	44
2	0.95	0.80	0.86	44
3	0.54	0.85	0.66	34
4	0.92	0.87	0.89	38
5	0.91	0.88	0.90	34
6	0.94	0.85	0.89	40
7	1.00	0.97	0.99	37
8	0.93	0.90	0.91	41
9	0.95	0.93	0.94	42
accuracy			0.89	400
macro avg	0.90	0.89	0.89	400
weighted avg	0.91	0.89	0.89	400

Confusion Matrix



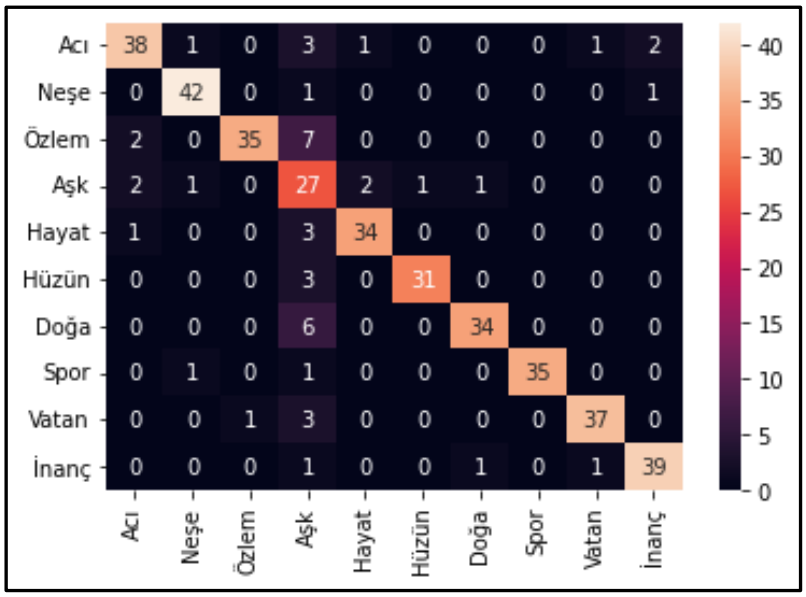
TF-IDF

Accuracy = 0.88

Classification Report

	precision	recall	f1-score	support
0	0.88	0.83	0.85	46
1	0.93	0.95	0.94	44
2	0.97	0.80	0.88	44
3	0.49	0.79	0.61	34
4	0.92	0.89	0.91	38
5	0.97	0.91	0.94	34
6	0.94	0.85	0.89	40
7	1.00	0.95	0.97	37
8	0.95	0.90	0.92	41
9	0.93	0.93	0.93	42
accuracy			0.88	400
macro avg	0.90	0.88	0.88	400
weighted avg	0.90	0.88	0.89	400

Confusion Matrix



2. NAÏVE BAYES

Multinomial Naïve Bayes, verilen bir metnin ait olduğu etiketin bulunmasında kullanılan yöntemlerden biridir. Verilen etiketler arasından, ait olma ihtimali en yüksek olan etiketi seçer. Bu yüzden, konumuz olan şiir sınıflandırması için uygun bir yöntem olarak belirlenmiştir.

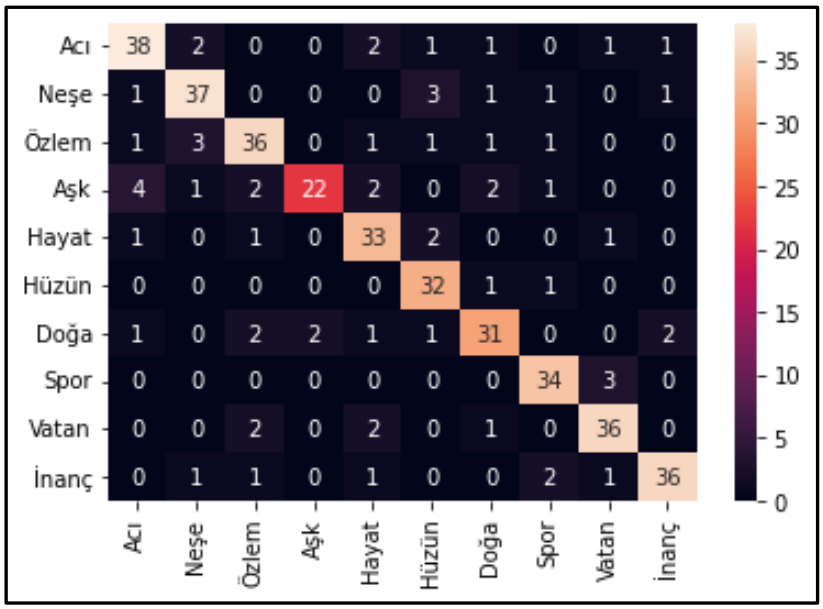
CountVectorizer

Accuracy = 0.8375

Classification Report

	precision	recall	f1-score	support
0	0.83	0.83	0.83	46
1	0.84	0.84	0.84	44
2	0.82	0.82	0.82	44
3	0.92	0.65	0.76	34
4	0.79	0.87	0.82	38
5	0.80	0.94	0.86	34
6	0.82	0.78	0.79	40
7	0.85	0.92	0.88	37
8	0.86	0.88	0.87	41
9	0.90	0.86	0.88	42
accuracy			0.84	400
macro avg	0.84	0.84	0.84	400
weighted avg	0.84	0.84	0.84	400

Confusion Matrix



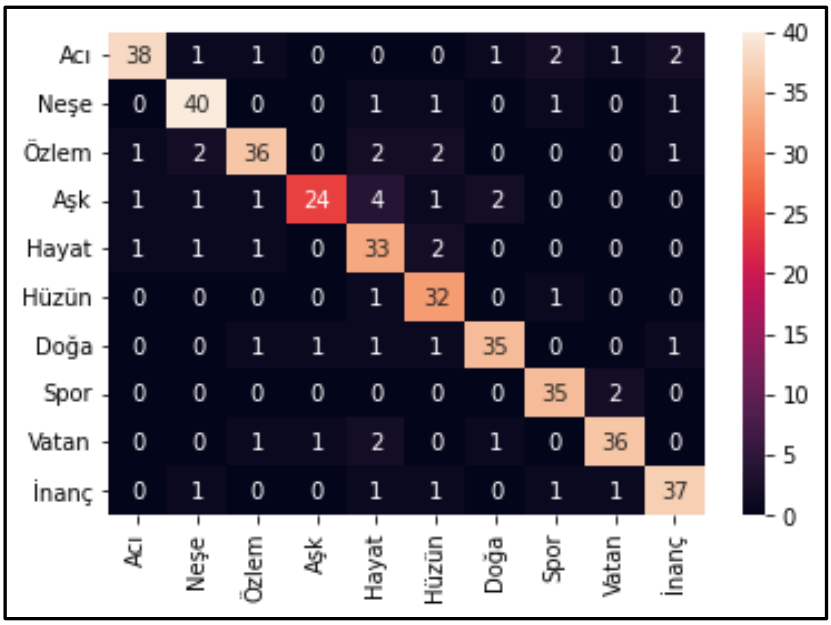
TF-IDF

Accuracy = 0.865

Classification Report

	precision	recall	f1-score	support
0	0.93	0.83	0.87	46
1	0.87	0.91	0.89	44
2	0.88	0.82	0.85	44
3	0.92	0.71	0.80	34
4	0.73	0.87	0.80	38
5	0.80	0.94	0.86	34
6	0.90	0.88	0.89	40
7	0.88	0.95	0.91	37
8	0.90	0.88	0.89	41
9	0.88	0.88	0.88	42
accuracy			0.86	400
macro avg	0.87	0.86	0.86	400
weighted avg	0.87	0.86	0.86	400

Confusion Matrix



3. SUPPORT VECTOR MACHINE (CLASSIFIER)

Yine sıkça kullanılan yöntemlerden biri olan SVM'ler için, bazı durumlarda Naive Bayes yöntemine göre daha hızlı olduğu, az bellek kullandığı ve daha doğru sonuçlar verebildiği belirtilmektedir. Bu yüzden projede kullanılacak yöntemlerden biri olarak seçilmiştir.

CountVectorizer

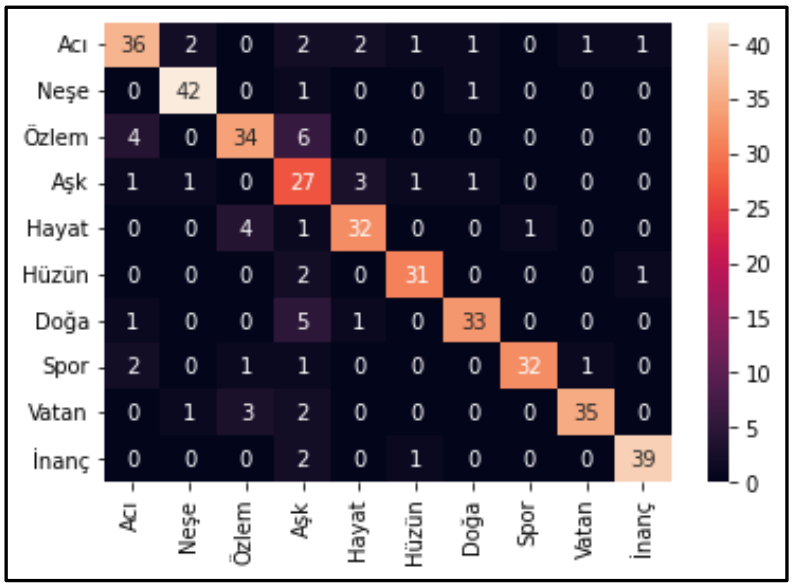
Karar Destek Sınıflandırıcısı, regülarizasyon parametresi 1, kernel 'lineer' ve derecesi 3 olacak şekilde oluşturulmuştur.

Accuracy = 0.8525

Classification Report

	precision	recall	f1-score	support
0	0.82	0.78	0.80	46
1	0.91	0.95	0.93	44
2	0.81	0.77	0.79	44
3	0.55	0.79	0.65	34
4	0.84	0.84	0.84	38
5	0.91	0.91	0.91	34
6	0.92	0.82	0.87	40
7	0.97	0.86	0.91	37
8	0.95	0.85	0.90	41
9	0.95	0.93	0.94	42
accuracy			0.85	400
macro avg	0.86	0.85	0.85	400
weighted avg	0.87	0.85	0.86	400

Confusion Matrix



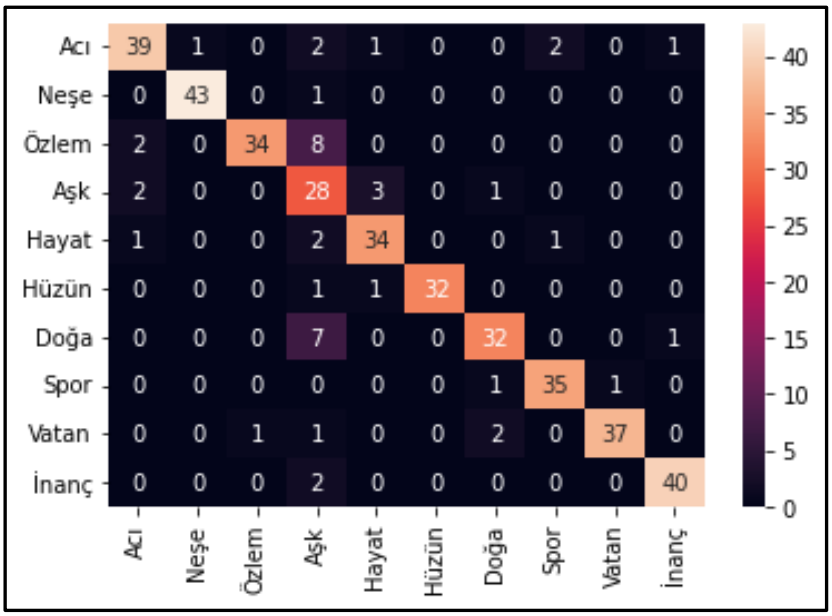
TF-IDF

Accuracy = 0.885

Classification Report

	precision	recall	f1-score	support
0	0.89	0.85	0.87	46
1	0.98	0.98	0.98	44
2	0.97	0.77	0.86	44
3	0.54	0.82	0.65	34
4	0.87	0.89	0.88	38
5	1.00	0.94	0.97	34
6	0.89	0.80	0.84	40
7	0.92	0.95	0.93	37
8	0.97	0.90	0.94	41
9	0.95	0.95	0.95	42
accuracy			0.89	400
macro avg	0.90	0.89	0.89	400
weighted avg	0.90	0.89	0.89	400

Confusion Matrix



4. BERT MODELİ

- Hugging Face platformu üzerinden hazır, önceden türkçe cümleler ile eğitilmiş BERT modeli ('dbmdz/bert-base-turkish-128k-cased') kullanılmıştır.
- Modelin mevcut versiyonu, Türkçe OSCAR corpus'unun (Vikipedi'den ve Kemal Oflazer tarafından sağlanan corpus'tan oluşmaktadır) filtrelenmesi ve segmente edilmesi ile eğitilmiştir.
- Hazır modelin eğitiminde 35 GB boyutunda ve 44 milyon token'a sahip corpus yer almıştır ve 128 bin boyutunda sözlük kullanılmıştır.
- Metinlerin tokenize edilmesi için yine hazır, Türkçe için kullanılan Bert Tokenizer'ından yararlanıldı.
- Her bir şiiri temsil eden embedding'in uzunluğu 128'dir.
- Adam optimizer'I kullanılmış, yapılan eğitimlerin sonuçlarına bağlı olarak learning rate 2e-5 olarak seçilmiştir. Mevcut model, batch size'ı 32, epoch sayısı 4 belirlenerek eğitilmiştir.

Accuracy = 0.9381

Loss = 0.27

Modelin Özeti

```
Train inp shape (1600, 128) Val input shape (400, 128)
Train label shape (1600,) Val label shape (400,)
Train attention mask shape (1600, 128) Val attention mask shape (400, 128)
Model: "tf_bert_for_sequence_classification_1"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 bert (TFBertMainLayer)      multiple                  184345344

 dropout_75 (Dropout)        multiple                  0

 classifier (Dense)          multiple                  7690

=================================================================
Total params: 184,353,034
Trainable params: 184,353,034
Non-trainable params: 0
_________________________________________________________________
```

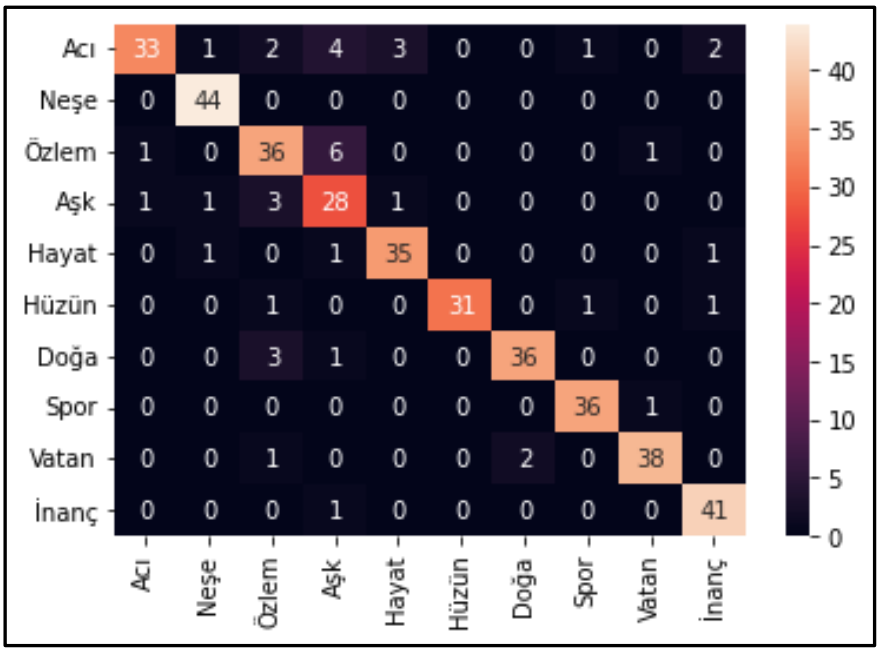
Modelin Eğitimi

```
Epoch 1/4
50/50 [==============================] - 107s 2s/step - loss: 2.1180 - accuracy: 0.2900 - val_loss: 1.7289 - val_accuracy: 0.5275
Epoch 2/4
50/50 [==============================] - 84s 2s/step - loss: 1.1238 - accuracy: 0.7462 - val_loss: 0.6107 - val_accuracy: 0.8500
Epoch 3/4
50/50 [==============================] - 84s 2s/step - loss: 0.4314 - accuracy: 0.9050 - val_loss: 0.4199 - val_accuracy: 0.8825
Epoch 4/4
50/50 [==============================] - 84s 2s/step - loss: 0.2729 - accuracy: 0.9381 - val_loss: 0.3976 - val_accuracy: 0.8950
```

Classification Report

	precision	recall	f1-score	support
Acı	0.94	0.72	0.81	46
Neşe	0.94	1.00	0.97	44
Özlem	0.78	0.82	0.80	44
Aşk	0.68	0.82	0.75	34
Hayat	0.90	0.92	0.91	38
Hüzün	1.00	0.91	0.95	34
Doğa	0.95	0.90	0.92	40
Spor	0.95	0.97	0.96	37
Vatan	0.95	0.93	0.94	41
İnanç	0.91	0.98	0.94	42
accuracy			0.90	400
macro avg	0.90	0.90	0.90	400
weighted avg	0.90	0.90	0.90	400

Confusion Matrix



Özet Tablo

Yöntem	Doğruluk (%)
SVM (TF-IDF)	88.50%
SVM (CountVectorizer)	85.25%
Naive Bayes (TF-IDF)	86.50%
Naive Bayes (CountVectorizer)	83.75%
Random Forest (TF-IDF)	88.00%
Random Forest (CountVectorizer)	88.50%
BERT Model	**93.81%**

Yorumlar

- En yüksek doğruluğu, önceden eğitilmiş BERT modeli üzerinde yapılan eğitim vermiştir. Bununla birlikte modelin hata oranı (0.27) da dikkate alınmalıdır. Bu hata oranı daha çok verinin sağlanması ve epoch sayısının artırılması ile iyileştirilebileceği düşünülmektedir.
- Diğer yöntemler arasında yüksek çoğunlukla TF-IDF yöntemi ile encode edilmiş şiirler daha yüksek başarı oranı göstermiştir. Bunun sebebi, CountVectorizer'ın yalnızca kelimelerin ne kadar sıklıkla geçtiğine yer veriyor olması, TF-IDF yönteminin ise önemli ve daha az önemli kelimeleri de dikkate alması olabilir.
- BERT modeli dışındaki diğer yöntemler birbirlerine oldukça yakın olmakla birlikte, yüksek başarılar elde etmişlerdir. Bu yöntemlerin sıklıkla metin sınıflandırma, spam tespiti, duygu analizi gibi konularda kullanıldığını göz önünde bulundurduğumuzda, sonuçlar anlam kazanmaktadır.
- Seçtiğimiz konulardan bazıları birbirine yakın olmakla birlikte, genellikle ayrık içeriklere sahip olduklarından yöntemlerin başarılarının olumlu etkilendiğini söyleyebiliriz. Örneği vatan, doğa, inanç konulu şiirlerin kelime bulutu çıkarıldığında elde edilen sonuç aşağıda yer almaktadır.

- Sonuç olarak derin öğrenmenin, klasik makine öğrenmesi yöntemlerine göre daha başarılı olduğu ortaya çıkarılmıştır. Ancak bu durum, makine öğrenmesi algoritmalarının kullanışsız olduğu anlamına gelmez; Hatta küçük verisetleri için oldukça iyi başarı değerleri elde edebilir.