T.C.

SİVAS CUMHURİYET ÜNİVERSİTESİ

TEKNİK BİLİMLER MESLEK YÜKSEKOKULU

BÜYÜK VERİ ANALİSTLİĞİ BÖLÜMÜ

DOĞAL DİL İŞLEME UYGULAMA RAPORU

GÖKAY İLYAS KOÇ

Danışman

Öğretim Görevlisi İstatistik Bölüm Başkanı Fatih GÖKÇE

Sivas

25. 2025

Twitter Metin Sınıflandırması

Hazırlayan: Gökay İlyas Koç
Öğrenci Numarası: 2024295008
Tarih: 25 Aralık 2025

Özet

Bu projede Twitter tweet verileri kullanarak tweetlerin **Nefret**, **Saldırgan** ve **Nötr** olmak üzere üç farklı sınıfa ayırmayı amaçlıyoruz. Projede Python ve makine öğrenmesi yöntemlerini kullanıyoruz. Tweetler üzerinde çeşitli ön işleme adımlarını deneyip ardından TF-IDF yaparak metinleri sayısal değerlere çeviriyoruz. Farklı sınıflandırma algoritmalarını kullanarak modelimizin performanslarını karşılaştırıp elde ettiğimiz sonuçlara göre en yüksek doğruluk oranı veren modeli belirleyip, performans analizini gerçekleştiriyoruz.

1. Amaç

Projemizin amacı, Tweetler'den elde edilen metinleri kullanarak nefret söylemi mi yoksa saldırgan içerik mi otomatik olarak tespit etmeyi amaçlıyoruz. Bu süreç sonunda makine öğrenmesi yöntemlerini deneyerek, oranları karşılaştırıyoruz ve en başarılı modeli belirliyoruz.

2. Kullanılan Veri ve Ön İşlemler

Projede kaggle üzerinden aldığımız **veri.csv** adlı içinde tweetler bulunan veri setini indirip kullanacağız. İşlemler öncesinde metinleri bir takım ön işleme adımları uygulayıp deneyeceğiz.

Uygulanan Ön İşleme Adımları

- tweet metinlerini küçük harfe dönüştürdüm.
- Linkleri, kullanıcı hashtag , sayıları ve noktalama işaretlerini temizleyip tweetleri analize uygun hale getiriyorum.
- Temizlenen tweetleri **clean_text** adında oluşturduğumuz yeni bir sütuna aktarıyoruz.

Temizlik Kontrolü

Orijinal Tweetler	Temizlenmiş Tweetler
!!! RT @mayasolovely: As a woman you shouldn't...	rt as a woman you shouldnt complain about...
!!!!! RT @mleew17: boy dats cold...	rt boy dats coldtyga dwn bad...
!!!!!!! RT @UrKindOfBrand Dawg!!!!	rt dawg rt you ever fuck...
!!!!!!!! RT @C_G_Anderson:	rt she look like a tranny
!!!!!!!!!!! RT @ShenikaRoberts:	rt the shit you hear about me...

Bu tweetler, temizleme işlemini doğru uyguladığımızı gösteriyor.

3. Özellik Çıkarımı

Tweetleri sayısal hale dönüştürmek için **TF-IDF (Term Frequency – Inverse Document Frequency)** kullanıyoruz. Bu işlemde **5000 kelimelik** bir özellik oluşturuyoruz.

- **TF-IDF matris boyutu:** (19.826, 5000)

Bu boyut, verinin yeterli kelime bilgisiyle eğitildiğini kanıtlar niteliktedir.

4. Kullanılan Modeller ve Performansları

Projemizde beş tane makine öğrenmesi yöntemini kullanarak eğitiyoruz ve test verisi ile doğruluk (accuracy) değerini hesaplıyoruz.

Model	Doğruluk (%)
Lojistik Regresyon	89.75
Naive Bayes	83.50
Destek Vektör Makineleri (SVM)	89.39
Karar Ağacı	87.37
Rastgele Orman	89.51

Sonuçlara göre en yüksek oranı **Lojistik Regresyon** modelinde yakalıyoruz.

5. Detaylı Performans Analizi

En yüksek oran veren modelimiz **Lojistik Regresyon** için sınıf bazlı precision, recall ve F1-score değerlerini hesaplıyoruz ekte:

```
              precision    recall  f1-score   support

Nefret             0.62      0.20      0.31       295
Saldırgan          0.92      0.96      0.94      3814
Nötr               0.84      0.84      0.84       848

accuracy                             0.90      4957
```

Değerlendirme

- Modelimizin yüksek bir doğruluk oranına sahip olduğunu görüyoruz. (%89.75).
- Özellikle saldırgan sınıfında başarılı sonuçlar elde ettiğimizi gördük.
- Nefret sınıfında recall değerinin düşük olması ise bu sınıftaki istisna verilerin doğru tespit edilmediğini görüyoruz.
- Nötr sınıfında ise dengeli bir performans dağılımı gözlemliyoruz.

6. Karışıklık Matrisi

Modelimizin hata analizinde ise karışıklık matrisini inceliyoruz:

Gerçek \ Tahmin	Nefret	Saldırgan	Nötr
Nefret	60	204	31
Saldırgan	37	3673	104
Nötr	0	132	716

Tabloda, nefret içerikli tweetlerin genellikle saldırgan sınıfı ile karıştırıldığını görüyoruz.

7. Görselleştirme

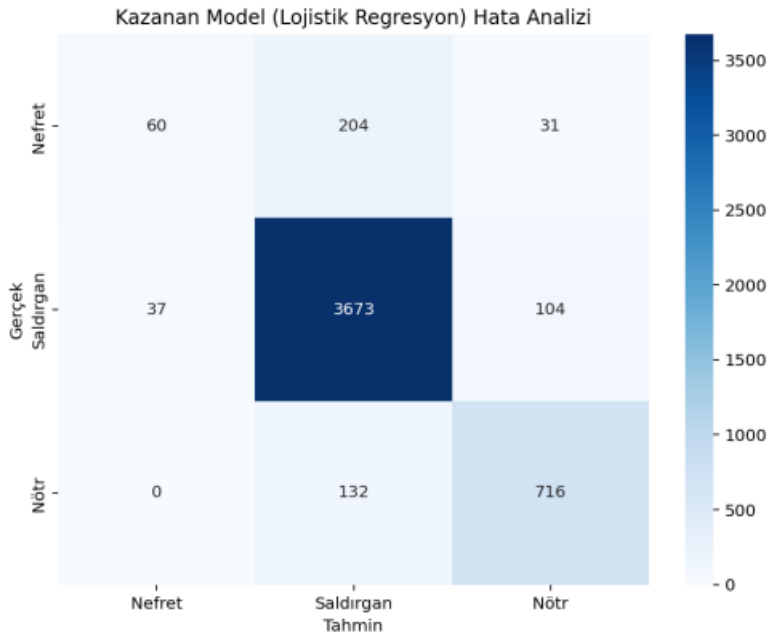
Karışıklık matrisi, **seaborn** kütüphanesiyle ısı haritası (heatmap) şeklinde görselleştiriyoruz. Bu görsel grafik, modelin hangi sınıflarda hata yaptığını daha net bir şekilde gösteriyor.

8. Sonuç

Bu analiz projemizde Twitter tweetleri üzerinde başarılı bir sınıflandırma süreci gerçekleştirdik ve Lojistik Regresyon modelinin en iyi performansını yakaladık.

EKLER

Ek-1: Karışıklık Matrisi Isı Haritası



Ek-2: Variable Explorer Çıktıları

Name	Type	Size	Value
cm	Array of int64	[3, 3]	[[60 204 31] [37 3673 104]
df	DataFrame	[24783, 8]	Column names: Unnamed: 0, count, hate_speech, offensive_language, neit ...
en_yuksek_puan	float	1	0.8975186604801291
isim	str	14	Rastgele Orman
kazanan_model_ismi	str	18	Lojistik Regresyon
kazanan_tahminler	Array of int64	[4957]	[1 1 1 ... 0 1 2]
model	ensemble._forest.RandomForestClassifier	50	RandomForestClassifier object of sklearn.ensemble._forest module
modeller	dict	5	{'Lojistik Regresyon':LogisticRegression, 'Naive Bayes':MultinomialNB, ...
puan	float	1	0.8957030461972968
tahmin	Array of int64	[4957]	[1 1 1 ... 1 1 2]
tfidf	feature_extraction.text.TfidfVectorizer	1	TfidfVectorizer object of sklearn.feature_extraction.text module
X_test	Series	[4957]	Series object of pandas.core.series module
X_test_vec	sparse._csr.csr_matrix	[4957, 5000]	csr_matrix object of scipy.sparse._csr module
X_train	Series	[19826]	Series object of pandas.core.series module
X_train_vec	sparse._csr.csr_matrix	[19826, 5000]	csr_matrix object of scipy.sparse._csr module
y_test	Series	[4957]	Series object of pandas.core.series module
y_train	Series	[19826]	Series object of pandas.core.series module

Confusion Matrix (cm)

`cm` değişkeni, 3x3 boyutunda Confusion Matrix'i temsil etmektedir.

- Satırlar: Gerçek sınıflar
- Sütunlar: Model tarafından tahmin edilen sınıflar

Veri Seti (df)

- Satır sayısı: 24.783
- Sütun sayısı: 8

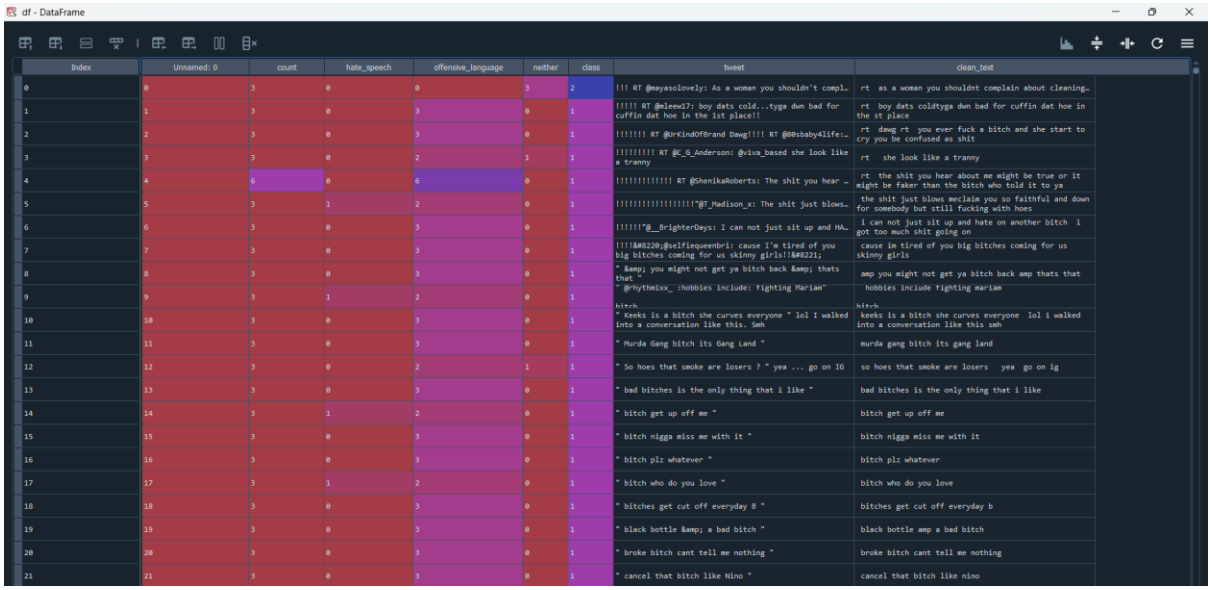
En Yüksek Başarı Skoru (en_yuksek_puan)

- Değer: %89.75

Kazanan Model Bilgileri

- Kazanan Model: Lojistik Regresyon

Ek-3: DataFrame Sütun Açıklamaları



Index	Unnamed: 0	count	hate_speech	offensive_language	neither	class	tweet	clean_text
0	0	3	0	0	3	2	!!! RT @mayasolovely: As a woman you shouldn't compl…	rt as a woman you shouldnt complain about cleaning…
1	1	3	0	3	0	1	!!!!! RT @mleew17: boy dats cold...tyga dwn bad for cuffin dat hoe in the 1st place!!	rt boy dats coldtyga dwn bad for cuffin dat hoe in the st place
2	2	3	0	3	0	1	!!!!!!! RT @UrKindOfBrand Dawg!!!! RT @80sbaby4life:…	rt dawg rt you ever fuck a bitch and she start to cry you be confused as shit
3	3	3	0	2	1	1	!!!!!!!!! RT @C_G_Anderson: @viva_based she look like a tranny	rt she look like a tranny
4	4	6	0	6	0	1	!!!!!!!!!!!!! RT @ShenikaRoberts: The shit you hear …	rt the shit you hear about me might be true or it might be faker than the bitch who told it to ya
5	5	3	1	2	0	1	!!!!!!!!!!!!!!!!!"@T_Madison_x: The shit just blows…	the shit just blows meclaim you so faithful and down for somebody but still fucking with hoes
6	6	3	0	3	0	1	!!!!!!"@__BrighterDays: I can not just sit up and HA…	i can not just sit up and hate on another bitch i got too much shit going on
7	7	3	0	3	0	1	!!!!“@selfiequeenbri: cause I'm tired of you big bitches coming for us skinny girls!!”	cause im tired of you big bitches coming for us skinny girls
8	8	3	0	3	0	1	" & you might not get ya bitch back & thats that "	amp you might not get ya bitch back amp thats that
9	9	3	1	2	0	1	" @rhythmixx_ :hobbies include: fighting Mariam"	hobbies include fighting mariam
10	10	3	0	3	0	1	" Keeks is a bitch she curves everyone " lol I walked into a conversation like this. Smh	keeks is a bitch she curves everyone lol i walked into a conversation like this smh
11	11	3	0	3	0	1	" Murda Gang bitch its Gang Land "	murda gang bitch its gang land
12	12	3	0	2	1	1	" So hoes that smoke are losers ? " yea ... go on IG	so hoes that smoke are losers yea go on ig
13	13	3	0	3	0	1	" bad bitches is the only thing that i like "	bad bitches is the only thing that i like
14	14	3	1	2	0	1	" bitch get up off me "	bitch get up off me
15	15	3	0	3	0	1	" bitch nigga miss me with it "	bitch nigga miss me with it
16	16	3	0	3	0	1	" bitch plz whatever "	bitch plz whatever
17	17	3	1	2	0	1	" bitch who do you love "	bitch who do you love
18	18	3	0	3	0	1	" bitches get cut off everyday B "	bitches get cut off everyday b
19	19	3	0	3	0	1	" black bottle & a bad bitch "	black bottle amp a bad bitch
20	20	3	0	3	0	1	" broke bitch cant tell me nothing "	broke bitch cant tell me nothing
21	21	3	0	3	0	1	" cancel that bitch like Nino "	cancel that bitch like nino

1. **Unnamed: 0** – Otomatik indeks sütunu
2. **count** – Metin içerisindeki kelime sayısı
3. **hate_speech** – Nefret söylemi etiketi
4. **offensive_language** – Saldırgan dil etiketi
5. **neither** – Nötr içerik etiketi
6. **class** – Nihai sınıf etiketi
7. **Metin sütunu** – Ham tweet metni
8. **Ön işlenmiş metin sütunu** – Temizlenmiş metin

Ek-4: Kullanılan Kütüphaneler

- Pandas
- Scikit-learn
- Matplotlib
- Seaborn