Twitter Sentiment Analizi

Muhammet Zahit Aydın
190201039

Kerem Karataş
190201076

Fatih Güner
190201074

***Abstract*— Bu yazıda, Spark kütüphanesi kullanarak Twitter Sentiment Analizi uygulamasının nasıl geliştirileceği anlatılmaktadır. Uygulama, 1.600.000 veri içeren bir veri seti aracılığıyla tweet girdisine bağlı olarak duygu analizi tahmini yapmayı planlar. İşletmeler, siyasi kampanyalar ve toplum bilimleri gibi birçok alanda kullanılabilir, bilgiler sağlayabilir. Uygulamanın geliştirilmesi için öncelikle verilerin toplanması, temizlenmesi ve özellik vektörüne dönüştürülmesi gerekmektedir. Daha sonra veriler eğitim ve test veri kümelerine ayrılır ve logistic regression kullanılarak model eğitilir. Son olarak, modelin performansı test veri kümesi kullanılarak değerlendirilir.**

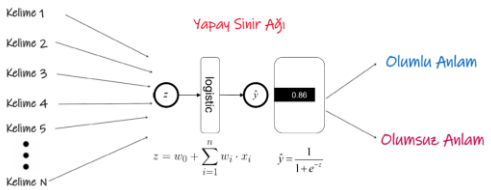
Keywords*— *Spark, Twitter, Logistic Regression, Sentiment Analysis

I. GİRİŞ

Sosyal medya platformları, kullanıcıların düşüncelerini, fikirlerini ve duygularını ifade etmek için kullandıkları bir alan olarak giderek daha önem kazanıyor. Bu nedenle, bu platformlarda paylaşılan içeriklerin anlamını anlamak ve yorumlamak, birçok alanda kullanılabilir bilgiler sağlayabilir. Twitter, dünya genelinde milyonlarca kullanıcısı olan popüler bir sosyal medya platformudur. Bu nedenle, Twitter'da yapılan sentiment analizleri, işletmeler, siyasi kampanyalar ve toplum bilimleri gibi birçok alanda kullanılabilir bilgiler sağlayabilir.

- Sentiment Analizi:

Sentiment analizi, doğal dil işleme alanında kullanılan bir tekniktir. Bu teknik, bir metnin duygu durumunu tanımlamak ve analiz etmek için kullanılır. Bu teknik, metindeki kelimeleri ve ifadeleri analiz ederek pozitif, negatif veya nötr olup olmadıklarını belirler.



- Spark Kütüphanesi:

Spark, büyük veri işleme için açık kaynaklı bir veri işleme motorudur. Spark, çok sayıda veri işleme işlemcisi üzerinde eşzamanlı olarak çalışarak büyük veri kümelerini hızlı bir şekilde işleyebilir. Spark kütüphanesi, veri işleme, makine öğrenimi ve grafik işleme gibi birçok alanda kullanılan bir kütüphanedir.

- Logistic Regression:

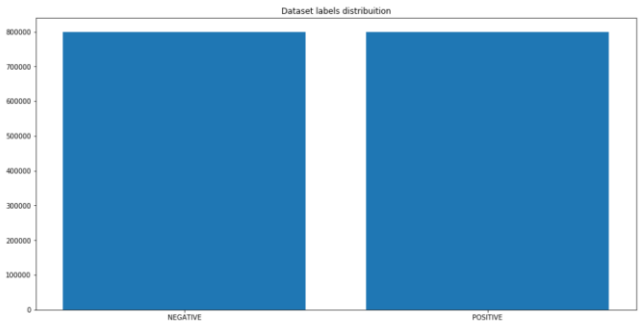
Logistic regression, makine öğrenimi alanında kullanılan bir tekniktir. Bu teknik, bağımsız değişkenlerin veri kümesindeki bir bağımlı değişkene olan etkisini analiz etmek için kullanılır. Bu teknik, sınıflandırma problemleri için kullanılır. Sınıflandırma problemleri, veri kümesindeki örnekleri belirli sınıflara ayırmayı amaçlayan problemlerdir.

- Twitter Sentiment Analizi Uygulaması:

Twitter Sentiment Analizi uygulaması, Spark kütüphanesi ve logistic regression kullanılarak geliştirilebilir. Bu uygulama, veri setindeki etiketli veriler aracılığıyla Logistic Regression modelini eğitir. Eğitilen model daha sonra Twitter'dan toplanan tweetlerin duygu durumunu belirlemek için kullanılır.

II. PROJE MIMARISI

- Kullanılan Veri Seti:
  Kullandığımız veri seti 1.600.000 tweet ve bu tweet'e karşılık gelen olumlu olumsuz etiketlerini içermektedir.




```
In [7]: df.head(5)
```

Out[7]:

	target	ids	date	flag	user	text
0	0	1467810369	Mon Apr 06 22:19:45 PDT 2009	NO_QUERY	_TheSpecialOne_	@switchfoot http://twitpic.com/2y1zl - Awww, t...
1	0	1467810672	Mon Apr 06 22:19:49 PDT 2009	NO_QUERY	scotthamilton	is upset that he can't update his Facebook by ...
2	0	1467810917	Mon Apr 06 22:19:53 PDT 2009	NO_QUERY	mattycus	@Kenichan I dived many times for the ball. Man...
3	0	1467811184	Mon Apr 06 22:19:57 PDT 2009	NO_QUERY	ElleCTF	my whole body feels itchy and like its on fire
4	0	1467811193	Mon Apr 06 22:19:57 PDT 2009	NO_QUERY	Karoli	@nationwideclass no, it's not behaving at all....

- Pozitif ve Negatif Veriler için Veri Bulutu(WordCloud):

Veri Bulutu(Wordcloud), belirli bir metnin veya kelime kümesinin sıklığını görsel olarak yansıtan bir veri

görselleştirme yöntemidir. Bu yöntemde, metindeki kelimelerin sıklıkları dikkate alınarak, daha sık kullanılan kelimeler daha büyük fontlarla ve daha az sık kullanılan kelimeler daha küçük fontlarla yazılmaktadır. Bu sayede, metindeki en önemli kelimeler kolayca fark edilebilir hale gelir.

- Negatif Veriler için Kelime Bulutu:

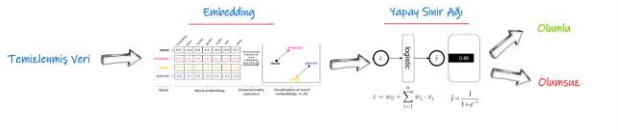
- Pozitif Veriler için Kelime Bulutu:

- Proje Adımları :

1. Verilerin ön işleme adımlarını gerçekleştirin. Bu adımlar, tweetleri tokenize etmek, kelimeleri küçük harfe çevirmek, özel karakterleri kaldırmak ve durak kelimeleri(stopwords) çıkarmak gibi işlemleri içerir.

2. Verileri özellik vektörüne dönüştürün. Bu adım, verileri sayısal verilere dönüştürmek için kullanılır. Özellik vektörü, veri kümesindeki her örnek için sayısal bir vektördür. Bu adımda, verileri özellik vektörüne dönüştürmek için TF-IDF yöntemi kullanılabilir. Projemizde Word2Vec, CountVectorizer ve HashTF yöntemleri kullanılmış ve en iyi sonuç veren model seçilmiştir.

3. Verileri eğitim ve test veri kümelerine ayırın. Eğitim veri kümesi, modelin eğitiminde kullanılırken, test veri kümesi, modelin performansını değerlendirmek için kullanılır.



4. Modeli eğitin. Logistic regression, veri kümesindeki örnekleri belirli sınıflara ayırmak için kullanılır. Bu adımda, model eğitim veri kümesindeki örnekleri kullanarak eğitilir.

5. Modelin performansını değerlendirmek için test veri kümesi kullanılır. Modelin doğruluğu, hassasiyeti, özgüllüğü ve F1 skoru gibi performans ölçüleri kullanılarak hesaplanabilir. Modelimizi eğitirken Grid-Search Hiperparamtetre Optimizasyonu algoritmasını kullandık. Logistic Regression modelinde regParam, elasticNetParam ve Tol gibi parametreleri değiştirerek en iyi sonuç veren modeli elde ettik.

III. SONUÇ

Twitter Sentiment Analizi uygulaması, Spark kütüphanesi ve logistic regression kullanılarak geliştirilebilir. Bu uygulama, Twitter'dan toplanan tweetlerin duygu durumunu belirlemek için kullanılır. Bu uygulama, işletmeler, siyasi kampanyalar ve toplum bilimleri gibi birçok alanda kullanılabilir bilgiler sağlayabilir.

REFERENCES

[1] https://en.wikipedia.org/wiki/Apache_Spark
[2] https://sparkbyexamples.com/pyspark-tutorial/
[3] https://spark.apache.org/docs/3.4.0/
[4] https://en.wikipedia.org/wiki/Logistic_regression
[5] https://www.kaggle.com/datasets/kazanova/sentiment140