T.C.

KONYA TEKNİK ÜNİVERSİTESİ

MÜHENDİSLİK VE DOĞA BİLİMLERİ FAKÜLTESİ

BİLGİSAYAR MÜHENDİSLİĞİ

3. SINIF BAHAR DÖNEMİ

BİLİŞİM TEKNOLOJİLERİ UYGULAMASI ARASINAV/FİNAL(BÜT) RAPORU

Öğrencinin Adı- Soyadı	Uğur Kaval
Numarası:	201220049
Danışmanı Adı Soyadı:	Ahmet Babalık
Sınav Tarihi:	22.06.2023

Projenin Adı: Doğal Dil İşleme ve Metin Madenciliği ile Duygu Analizi

DÖNEM İÇİ YAPILAN ÇALIŞMALARIN ÖZETİ

Veri Setinin Hazırlanması:
Veri seti, "train.tsv" adlı bir dosyadan yüklendi.
Veri seti üzerinde bazı ön işleme adımları gerçekleştirildi: büyük-küçük harf dönüşümü, noktalama işaretlerinin ve sayıların kaldırılması, stop words'lerin çıkarılması, seyrek kelimelerin kaldırılması gibi.
Veri Setinin Bölünmesi:
Veri seti, eğitim ve test veri kümelerine ayrıldı.
Eğitim veri kümesi "train_x" ve "train_y" olarak, test veri kümesi "test_x" ve "test_y" olarak tanımlandı.
Metin Temsillerinin Oluşturulması:
CountVectorizer ve TfidfVectorizer kullanılarak metin verileri sayısal temsillerine dönüştürüldü.
CountVectorizer ile metinlerdeki kelime frekanslarını hesaplandı ve bir "count matrix" oluşturuldu.
TfidfVectorizer, metinlerdeki terim frekansını ters belge frekansı ile çarparak bir "tf-idf matrix" oluşturuldu.
Makine Öğrenimi Modellerinin Oluşturulması ve Eğitimi:
Çeşitli makine öğrenimi modelleri (Logistic Regression, Naive Bayes) kullanılarak eğitim veri kümesi üzerinde modeller oluşturuldu.
Cross-validation kullanılarak model performansı değerlendirildi.
Duygu Analizi:
Kullanıcıdan bir cümle alınarak bu cümle, önceki ön işleme adımlarına tabi tutulur ve sayısal temsili alındı.
Eğitilmiş model kullanılarak cümlenin duygusu (pozitif veya negatif) tahmin edildi.
Arayüz:
Tkinter kullanılarak basit bir grafik arayüz oluşturuldu.
Kullanıcının bir cümle girmesi için bir metin kutusu ve analizi yapmak için bir buton eklendi.
Kullanıcı düğmeye bastığında, girilen cümlenin duygusu hakkında bir ileti kutusu görüntülendi.
Daha sonra cümlenin analizi yapıldı ve sonuç olarak döndürüldü.

PROJENİN AMACI ve ÖNEMİ

Projenin Amacı:

Proje, gerçek hayatta karşılaşılan bir metin sınıflandırma sorununa çözüm sunmak amacıyla bir makine öğrenmesi modeli geliştirmeyi hedeflemektedir. Temel olarak, bir metnin pozitif veya negatif olarak sınıflandırılmasıyla ilgilenmektedir.

Proje, metin verilerinin işlenmesi ve analizi için çeşitli yöntemleri kullanmaktadır. İlk adım, veri ön işleme tekniklerini kullanarak metinleri temizlemek ve düzenlemektir. Veri setindeki metinlerin sentiment skorları hesaplanır. Bu sentiment skorları, metnin duygusal tonunu temsil eder ve pozitif veya negatif olarak sınıflandırma için bir ölçüt sağlar. Veri seti, değişken mühendisliği teknikleri kullanılarak çeşitli özelliklerle zenginleştirilir. En yüksek doğruluk oranı elde edilen model ile yeni yorumların sınıflandırılması yapılacaktır. Proje, metin sınıflandırma problemlerinde kullanılan temel teknikleri ve makine öğrenmesi modellerini anlamak için örnek bir uygulama olarak da değerlendirilebilir.

Projenin Önemi:

Proje, metin sınıflandırma problemlerine çözüm üretmek amacıyla kullanılan önemli bir uygulama örneğidir. Metin sınıflandırma, günümüzde çok sayıda alanda kullanılan önemli bir teknik olup, örneğin, sosyal medya analizi, müşteri yorumlarının incelenmesi, spam filtreleme, haber makalelerinin sınıflandırılması gibi birçok alanda kullanılmaktadır.

Bu projenin bir diğer önemi ise, veri ön işleme, değişken mühendisliği ve makine öğrenmesi teknikleri gibi temel teknikleri kullanarak, bir gerçek hayat problemini çözmeye yönelik bir örnek sunmasıdır. Bu proje, Apple, Netflix, Amazon gibi büyük şirketlerin yaptığı gibi insan duygularını, yorumlarını ve tepkilerini toplayarak bunları bir veri setine dönüştürüp analiz etme sürecine odaklanmaktadır. Veri ön işleme teknikleri, sentiment analizi, makine öğrenmesi algoritmaları ve doğruluk oranı hesaplamaları kullanılarak, bu veri setleri daha yalın ve anlaşılır bir hale getirilerek pozitif veya negatif duygulara sahip oldukları sınıflandırılmaktadır. Bu sayede, insanların hislerini ve düşüncelerini anlamak için kullanılan veri analizi süreci, bu projede kullanılan tekniklerle daha verimli ve etkili hale getirilmektedir.

KAYNAK ARAŞTIRMASI

Duygu analizi (sentiment analysis) projesi için Python'da kullanılan bazı kütüphaneler ve teknikler kullanılmıştır. Proje, veri kümesindeki metinleri pozitif veya negatif olarak sınıflandırmayı amaçlamaktadır.

Kullanılan Kütüphaneler:
pandas: Veri analizi ve işleme için kullanılan bir kütüphane.
sklearn: Makine öğrenimi modellerini ve metrikleri uygulamak için kullanılan bir kütüphane.
textblob: Metin işleme ve duygu analizi için kullanılan bir kütüphane.
keras: Derin öğrenme modelleri oluşturmak ve eğitmek için kullanılan bir kütüphane.
numpy: Sayısal hesaplamalar için kullanılan bir kütüphane.
string: Metin işleme işlemleri için kullanılan bir kütüphane.
nltk: Doğal dil işleme (NLP) işlemleri için kullanılan bir kütüphane.
tkinter: Grafiksel kullanıcı arayüzü (GUI) oluşturmak için kullanılan bir kütüphane.
Veri Hazırlama:
data = pd.read_csv("train.tsv",sep = "\t"): "train.tsv" adlı bir dosyadan veri okunur.
Veri kümesi üzerinde bazı temizlik işlemleri yapılır. Örneğin, gereksiz sütunlar çıkarılır ve etiketlerin sınıflarına karşılık gelen değerleri değiştirilir.
Metin Ön İşleme:
Metinlerdeki büyük/küçük harf dönüşümü yapılır.
Noktalama işaretleri ve sayılar çıkarılır.
Stopwords (durak kelimeler) temizlenir.
Seyrek (sparse) kelimeler çıkarılır.
Lemmatizasyon işlemi uygulanır.
Veri Bölme:
Veri kümesi, eğitim ve test veri setlerine bölünür.
Özellik Çıkarımı:
Sayma vektörleri (Count Vectors) ve TF-IDF vektörleri (Term Frequency-Inverse Document Frequency) oluşturulur.
Model Eğitimi:
Çeşitli sınıflandırma modelleri (Logistic Regression, Naive Bayes) eğitilir ve doğruluk oranları hesaplanır.
Metin Analizi:
Kullanıcının girdiği bir cümlenin analizi için bir fonksiyon oluşturulur.

Makale: "Deep Learning for Sentiment Analysis: A Survey"
(https://arxiv.org/abs/1801.07883)
Makale: "Sentiment Analysis and Opinion Mining"
(https://www.morganclaypool.com/doi/abs/10.2200/S00616ED1V01Y201409HLT026)
Makale: "Sentiment Analysis: Concept, Analysis and Applications"
(https://www.researchgate.net/publication/334156230_Sentiment_Analysis_Concept_Analysis_and_Applications)
Kaynak: Stanford Sentiment Analysis Dataset
(https://ai.stanford.edu/~amaas/data/sentiment/)

MATERYAL VE METOT

MATERYAL

Veri Seti ve Metin Ön İşleme:
Python ile veri seti işleme: Pandas kütüphanesi belgeleri (https://pandas.pydata.org/docs/)
Metin ön işleme adımları: NLTK (Natural Language Toolkit) belgeleri (https://www.nltk.org/)
Metin temizleme ve normalizasyon yöntemleri: NLTK belgeleri (https://www.nltk.org/book/ch03.html)

Sentiment Analizi:
Sentiment analizi için Python kütüphaneleri:
TextBlob: TextBlob belgeleri (https://textblob.readthedocs.io/)
NLTK Sentiment Analysis: NLTK belgeleri (https://www.nltk.org/howto/sentiment.html)
Makine Öğrenimi ve Sınıflandırma Modelleri:

Scikit-learn kütüphanesi: Scikit-learn belgeleri (https://scikit-learn.org/stable/documentation.html)
Lojistik regresyon modeli: Scikit-learn belgeleri (https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html)
Count Vectors ve TF-IDF: Scikit-learn belgeleri (https://scikit-learn.org/stable/modules/classes.html#module-sklearn.feature_extraction.text)
N-gram TF-IDF: Scikit-learn belgeleri (https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html)
Char-level modeller: Scikit-learn belgeleri (https://scikit-learn.org/stable/modules/classes.html#module-sklearn.feature_extraction.text)

Model Eğitimi ve Değerlendirme:
Makine öğrenimi modelinin eğitimi ve değerlendirilmesi: Scikit-learn belgeleri (https://scikit-learn.org/stable/modules/classes.html#module-sklearn.model_selection)

METOT

Pandas Kütüphanesi: Veri setini okuma, temizleme, filtreleme gibi veri işleme işlemlerini gerçekleştirmek için kullanılır. Özellikle pandas DataFrame ve pandas.Series veri yapılarını kullanarak veriler yönetilir.
NLTK (Natural Language Toolkit): Metin ön işleme adımlarını gerçekleştirmek için kullanılan popüler bir kütüphanedir. Metinleri cümlelere ayırma, kelime köklerini bulma, noktalama işaretlerini kaldırma gibi işlemleri yapabiliriz.
TextBlob: Metinlerin sentiment analizini yapmak için kullanabiliriz. Sentiment skorlarına erişebilir ve metinleri pozitif veya negatif olarak sınıflandırabiliriz.
Scikit-learn: Makine öğrenimi modellerini oluşturmak, eğitmek ve değerlendirmek için kullanabileceğiniz kapsamlı bir kütüphanedir. Lojistik regresyon, Naive Bayes, SVM gibi sınıflandırma algoritmalarını kullanabiliriz. Ayrıca, Scikit-learn ile TF-IDF vektörleştirme ve N-gram özellikleri oluşturma işlemlerini de gerçekleştirebiliriz.
Scikit-learn: Scikit-learn kütüphanesi ile oluşturduğunuz modeli eğitebilir ve veri setinin bir bölümünü kullanarak modelin performansını değerlendirebiliriz. Örneğin, train_test_split fonksiyonunu kullanarak veri setini eğitim ve test kümelerine ayırabilirsiniz.

KAYNAKLAR

https://pandas.pydata.org/docs/

https://www.nltk.org/

https://textblob.readthedocs.io/

https://github.com/you915/Sentiment-Analysis-of-Twitter-Data-for-predicting-Apple-stock-price

https://scikit-learn.org/stable/documentation.html

https://www.veribilimiokulu.com/derin-ogrenme-ile-duygu-analizi/

https://numpy.org/

https://www.nltk.org/

https://medium.com/

https://towardsdatascience.com/

Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit (NLTK) by Steven Bird, Ewan Klein, and Edward Loper

NLP with PyTorch: Build Intelligent Language Applications Using Deep Learning by Delip Rao and Brian McMahan

Python ile Makine Öğrenmesi-Sadi Evren şeker

Veri Bilimi Okulu

Veri Bilimi ve Machine Learning-Vahit Keskin