T.C
KONYA TEKNİK ÜNİVERSİTESİ BİLGİSAYAR MÜHENDİSLİĞİ
BİLGİSAYAR MÜHENDİSLİĞİ UYGULAMASI – 2
(BİTİRME PROJESİ-2) ARA RAPOR FORMU

Öğrencinin Adı- Soyadı	Dilemre Ülkü
Numarası:	171213055
Danışmanı Adı Soyadı:	Sedat Korkmaz
Sınav Tarihi:	
Projenin Konusu: Sayısal Üslup Analizi (Stilometri) Yöntemleri ve Yapay Zekâ Kullanılarak Eser – Yazar Eşleştirme	

DÖNEM İÇİ YAPILAN ÇALIŞMALARIN ÖZETİ

Mevcut projede daha önceden, doğal dil işleme kütüphanesi olarak kullanılan NLTK kütüphanesi yerine Türkçe metinlerde morfolojik analiz yapmamıza olanak sağlayan Zemberek kütüphanesine geçilmeye karar verilmiştir. Zemberek Java tabanlı ve proje Python tabanlı olduğundan dolayı Zemberek, Jpype kütüphanesinin yardımı ile projeye entegre edilmiştir.

Daha önceden veri kazıma ile elde edilen verilerde 204 yazara ait 2415 köşe yazısı bulunmaktadır. Veriler dengesizdir: 15'ten az köşe yazısı bulanan yazar sayısı 95, bu yazarların toplam köşe yazısı sayısı 328'tir. 15'ten az köşe yazısı bulunan yazarların sayısı fazla, fakat toplam köşe yazısı sayısı veri setini küçültmediği için bu veriler veri setinden çıkartılmıştır.

Yazı formatındaki verilerin makine öğrenmesine hazırlanması için uygulanan işlemler şunlardır; metni küçük harfe çevirme, kelime belirteçlerine ayırma, kelimelerin eklerini atarak köklerini çıkarma, stopword kelimelerin atılması, cümle belirteçlerine ayırma, kök uzunluk dağılımlarının çıkartılması, kelime bazında cümle uzunluk dağılımlarının çıkartılması, kelime zenginlik oranlarının çıkartılması, ortalama kök uzunluğu, kelime olarak ortalama cümle uzunluğu, toplam noktalama işareti sayısının çıkartılması, toplam kullanılan stopword sayısının çıkartılması, tamamen büyük harfte yazılmış kelime sayısının çıkartılmasıdır. Yazarlar pozitif tam sayılar olarak kodlanmıştır.

Ön işlemden geçirilerek kökleri çıkartılan her metnin TF-IDF, kelime torbası vektörleri, kelime, cümle dağılımları ve diğer özniteliklerin vektörleri çıkartılmış, tüm vektörler kendi aralarında min-max normalizasyon yöntemi kullanılarak normalleştirme yapılmıştır.

Verilerin %77'si eğitim (1398 köşe yazısı), %33'ü test (689 köşe yazısı) olmak üzere ikiye ayrılmıştır. Scikit-learn kütüphanesindeki SVM modeli varsayılan parametrelerde eğitilerek özniteliklerin farklı kombinasyonları ile alınan doğruluk puanları karşılaştırılmıştır. En yüksek %57,4 doğruluk oranına ulaşılmıştır. SVM ile en iyi doğruluk oranına ulaşılan öznitelikler ile, yine Scikit-learn kütüphanesinde bulunan MLP Classifier (Multi-Layer Perceptron, Çok Katmanlı Algılayıcılar), sinir ağı modeli varsayılan parametrelerinde eğitilerek %77 doğruluk oranına ulaşılmıştır.

Ardından eğitim ve test verilerinde, veri dağılıma bağlı sapmalar ve hatalardan kaçınmak, ayrıca SVM modelinde verilere en uygun parametreleri belirlemek için Scikit-learn kütüphanesinde bulunan RandomizedSearchCV fonksiyonu kullanılmıştır. Sürekli bir olasılıksal dağılım oluşturan SciPy kütüphanesinde bulunan log-uniform fonksiyonu ile C değerleri 1-1000 arasında, gamma değeri 0.0001-0.01 arasında seçilmiş ve toplam 30 iterasyonda, 10 kat çapraz doğrulama ile hiper-parametreler aranmıştır. En yüksek doğruluk puanı %85 ile linear çekirdekte C = 216 değeridir. Linear çekirdekte gamma değeri bulunmamaktadır.

KAYNAK ARAŞTIRMASI

Stilometri, öncelikle yazılı edebiyat alanında olmak üzere, resim ve müzik gibi sanat dallarında, tarih, din ve hukuk alanında ve adli bilimlerde kullanılan bir üslup belirleme çalışmasıdır ve stilometri analizi ise, üslup belirteçlerinin (syle markers) değişken olarak alınıp bu değişkenlerin istatistiksel ve bilişimsel metotlar kullanılarak çeşitli amaçlarla incelenmesine dayanan bir yöntemdir. (İşi, Çemrek, Yıldız, 2013, s: 272)

Referans	Veri	Özellikler	Yöntem	Başarı Düzeyi
Agün, H. V., Yılmazel, S. ve Yılmazel, O. (2017)	En az 1.000 karakterli köşe yazıları, her yazar için 60 adet.	Sözcüksel ve sözdizimsel özellikler	LR, çok değişkenli NB ve çok katmanlı sinir ağı	F-skoru 0,37-0,95 (10 kat çapraz doğrulama)
Amasyalı, M. F., ve Diri, B. (2006)	Dört kadın, 14 erkek yazardan politika, spor ve genel kültür üzerine, 35'er köşe yazısı	Sözcüksel özellikler (2-grams ve 3-grams)	NB, SVM, C4.5 ağacı, rastgele orman	Yazar doğrulaması için %59-83
Aslantürk, O., Sezer, E. A., Sever, H., ve Raghavan, V. (2010)	Dokuz yazardan, politika ve yaşam konularında toplam 513 köşe yazısı	Sözcüksel ve sözdizimsel özellikler	Kaba küme tabanlı sınıflandırma	%70 doğruluk oranı
Aslantürk, O. (2014)	Sekiz yazarın 12.115 adet yaşam ve siyaseti konu alan köşe yazıları	Sözcüksel ve sözdizimsel özellikler	Kaba küme tabanlı sınıflandırma	Toplam 1.134 deneyden 498 tanesi için %70 üzerinde doğrulukla oranı
Bay, Y., ve Çelebi, E. (2016)	17 farklı yazardan toplam 850 köşe yazısı	Sözcüksel özelikler	NB, SVM ve Karar ağacı ve KNN	%96-100 arası doğruluk oranı (10 kat çapraz doğrulama)
Bozkurt, I. N., Bağlıoğlu, O., ve Uyar, E. (2012)	18 farklı yazardan her biri için 500 köşe yazısı	Sözcüksel ve sözdizimsel özellikler, İşlevsel sözcük sıklığı	Histogram metodu, KNN, Bayes sınıflandırma, KM, bu algoritmaların kombinasyonu ve SVM	En yüksek SVM ile %95,7 (10-kat çapraz doğrulama)
Canbay, P., Sever, H., ve Sezer, E. A. (2018)	10 farklı blog yazarından her biri için 50 blog yazısı	Sözcüksel özellikler (Noktalama işaretleri ve kelime çantası)	SVM ve yapay sinir ağı	Ortalama %25-75 doğruluk oranı (10-kat çapraz doğrulama)

Diri, B., ve Amasyalı, M. F. (2003)	18 farklı yazardan her biri için 20 metin	Sözcüksel ve sözdizimsel özellikler	Skor tabanlı metot	En yüksek %84 doğruluk oranı
Küçükyılmaz, T., Cambazoğlu, B. B., Aykanat, C., ve Can, F. (2008)	Sohbet mesajları koleksiyonu	Sözcüksel özellikler, karakter özellikleri, dijital ifadeler	Patient Rule Induction Method, SVM, NB ve KNN	Yazar ataması %100-97
Taş, T., ve Görür, A. K. (2007)	20 yazardan her biri için 25 köşe yazısı	Sözcüksel ve sözdizimsel özellikler, farklı kelime zenginliği özellikleri	Bayes ağı, NB, MNB, NB güncellenebilir lojistik regresyon, Çok katmanlı öğrenme, Radyal temel fonksiyon ağı, Basit lojistik, Regresyon, DECORATE, Çok sınıflı sınıflandırıcı	Özellik seçimi sonrası %80-57 doğruluk oranı (10-kat çapraz doğrulama)
Taşçı, H. ve Ekinci, E. (2012)	10 farklı yazardan 10 ayrı köşe yazısı	Karakter özellikleri ve işlevsel sözcükler	Kosinüs uzaklığı	Karakter özellikleri ile ortalama %86, işlevsel sözcükler ile ortalama %53 doğruluk oranı
Türkoğlu, F. (2006)	18 yazara ait, 35 adet doküman alınarak 630 metin	Sözcüksel ve sözdizimsel özellik, n-gramlar, işlevsel kelimeler	NB, SVM, Rastgele Orman, Çok Katmanlı Algılayıcı ve Öz Düzenleyici Özellik Haritası ve KNN	Farklı veri seti kombinasyonları için en iyi sonuçlar %82,1, %85 ve %89,2 doğruluk oranları
Türkoğlu, F., Diri, B., ve Amasyalı, M. F. (2007)	18 farklı yazardan her biri için 35 köşe yazısı	Çok sayıda sözcüksel ve sözdizimsel özellik, n-gramlar, işlevsel kelimeler	NB, SVM, Rastgele Orman ve Çok Katmanlı Algılayıcı KNN	SVM ile ortalama %88,9 doğruluk oranı
Yavanoğlu, O. (2016)	Dokuz yazardan ekonomi, yaşam ve politika kategorilerinde 20000'i aşkın köşe yazısı	Sözcüksel ve sözdizimsel özellikler	Yapay sinir ağları	Ekonomi için %98, politika için %97, yaşam için %81 ve kategoriler arası %80 doğruluk oranları (10-kat çapraz doğrulama)

Tablo 1: Türkçe metinler üzerinde yazar ataması çalışmaları (Çalışkan, Can, 2018)

PROJEDE KULLANILAN MATERYAL VE METOTLAR

Projenin kodlanması Anaconda dağıtımı 2.14 versiyonunda, Python 3.9.7 versiyon ile Jupyter Notebook 6.4.5 versiyon geliştirme ortamında yapılmıştır. Java tabanlı 0.17.1 versiyon Zemberek kütüphanesi Jar dosyası olarak indirilmiş, Jar dosyasındaki modülleri okumak için zipfile kütüphanesi; Python'a entegre edilmesi için 1.3 versiyon Jpype kütüphanesi kullanılmıştır. Veri manipülasyonu için 1.4.1 versiyon Pandas kütüphanesi; üst düzey matematiksel işlemler için 1.20.3 versiyon Numpy kütüphanesi; makine öğrenmesi modelleri ve bazı ön işleme modülleri için 1.0.2 versiyon Scikit-learn kütüphanesi kullanılmıştır.

Proje dört adımdan oluşmaktadır:

1. Verilerin toplanması
2. Ön işleme ve öznitelik çıkartma
3. Makine öğrenmesi uygulaması
4. Ara yüz oluşturma

Veriler bir gazetenin internet sitesinden veri kazıma yöntemi ile elde edilmiştir. 14 Ocak 2017 ile 19 Mart 2021 tarihleri arasında toplam 204 yazara ait en fazla 20, en az 1 olmak üzere toplam 2415 derlenmiştir. Veriler CSV formatında bir tabloda tutulmuştur. Tablo içeriğinde tarih, yazar, başlık, link ve metin bilgileri bulunmaktadır.

Veri ön işleme sırasında aşağıdaki işlemler yapılmıştır.

1. 15 tane köşe yazısından az köşe yazısına sahip yazarların veri setinden kaldırılması
2. Satır boşluklarını kaldırılması, metin küçük harfe çevrilmesi, noktalama işaretlerinin kaldırılması (toplam 95 yazar 328 köşe yazısı)
3. Kelime belirteçlerinin elde edilmesi
4. Zemberek'ten alınan bir Türkçe stopword listesi ile metindeki stopword kelimelerinin kaldırılması
5. Zemberek kütüphanesi ile kelime köklerinin elde edilmesi

Ön işlemenin ardından aşağıdaki tabloda gösterilen öznitelikler elde edilmiştir.

Öznitelik	Açıklama
Kök sıklığı vektörü	Uzunluğu bütün köşe yazılarında geçen benzersiz kök sayısı kadar olan bir vektör oluşturulur. Vektörün her bir elemanı belirli bir kökün metinde kaç defa geçtiğini belirtir.
TF-IDF vektörü	TF (terim sıklığı) ilgili kelimenin işlem yapılan metindeki frekansıdır. DF (doküman sıklığı) ilgili kelimenin diğer metinlerdeki frekansıdır. IDF (ters doküman sıklığı) DF değerinin logaritmasıdır. TF-IDF bu iki değerin çarpımı ile hesaplanır.ve belirli bir sözcüğün bulunduğu dokümanı ne kadar temsil ettiğinin istatistiksel bir değeridir. TF_IDF matrisi ise bu değerlerin bütün kelimeler ve metinler için hesaplanmış halidir.
Harf bazında kelime uzunluk dağılımı	Vektörün her bir elemanı, belirli bir uzunluktaki her bir kelimenin metinde kaç defa geçtiğini belirtir.
Kelime bazında cümle uzunluk dağılımı	Vektörün her bir elemanı, belirli bir uzunluktaki her bir cümlenin metinde kaç defa geçtiğini belirtir.
Kelime zenginliği ölçüsü	Metindeki benzersiz kelime sayısının toplam kelime sayına oranıdır.
Kök zenginliği ölçüsü	Metindeki benzersiz kök sayısının toplam kök sayına oranıdır
Harf bazında ortalama kelime uzunluğu	Metindeki ortalama kelime uzunluğudur.
Kelime bazında ortalama cümle uzunluğu	Metindeki ortalama cümle uzunluğudur.
Noktalama işareti sayısı	Metinde geçen noktalama işaretlerinin sayısıdır.
Stopword sayısı	Metinde geçen stopword kelimelerinin sayısıdır.
Tamamı büyük yazılmış kelime sayısı	Metinde geçen tamamı büyük harfle yazılmış kelimelerin sayısıdır.

Tablo 2: Elde edilen öznitelikler ve açıklamaları.

Elde edilen öznitelikler Min-Max Normalization yöntemi ile normalleştirildikten sonra Scikit-Learn kütüphanesinde bulunan varsayılan parametrelerdeki SVM modeli ile özniteliklerin performansı değerlendirilmiştir. SVM iki ya da daha fazla sınıfı birbirinden ayıran hiper-düzlemin belirlenmesine dayalı bir sınıflandırma algoritmasıdır. Sınıfları birbirinden ayırmak için sonsuz adet düzlem belirlenebilmektedir (Şenel, 2020). Varsayılan parametreler şöyledir: Kernel = RBF, C = 1.0, Gamma = scale ($\frac{1}{n_{features} \cdot var(X)}$).

Aşağıdaki tabloda özniteliklerin farklı kombinasyonlarının performansı vardır.

Öznitelik	Doğruluk	Öznitelik	Doğruluk
Kök sıklığı vektörü	%42,80	Stopword sayısı	%5,74
TF-IDF vektörü	%44,72	Tamamı büyük yazılmış kelime sayısı	%1,43
Harf bazında kelime uzunluk dağılımı	%7,66	Kök sıklığı vektörü, TF-IDF vektörü	%51,00
Kelime bazında cümle uzunluk dağılımı	%11,00	TF-IDF vektörü, Kelime bazında ortalama cümle uzunluğu	%50,46
Kelime zenginliği ölçüsü	%1,44	TF-IDF vektörü, Harf bazında ortalama kelime uzunluğu	%50,46
Kök zenginliği ölçüsü	%2,87	TF-IDF vektörü, Kök sıklığı vektörü, Harf bazında kelime uzunluk dağılımı, Kelime bazında cümle uzunluk dağılımı, Kelime zenginliği ölçüsü, Kök zenginliği ölçüsü, Harf bazında ortalama kelime uzunluğu, Kelime bazında ortalama cümle uzunluğu, Noktalama işareti sayısı, Stopword sayısı, Tamamı büyük yazılmış kelime sayısı	%52,86
Harf bazında ortalama kelime uzunluğu	%5,72	TF-IDF vektörü, Kök sıklığı vektörü, Harf bazında kelime uzunluk dağılımı, Kelime bazında cümle uzunluk dağılımı, Kök zenginliği ölçüsü, Harf bazında ortalama kelime uzunluğu, Kelime bazında ortalama cümle uzunluğu, Noktalama işareti sayısı, Stopword sayısı, Tamamı büyük yazılmış kelime sayısı	%57,47
Kelime bazında ortalama cümle uzunluğu	%1,44	TF-IDF vektörü, Kök sıklığı vektörü, Harf bazında kelime uzunluk dağılımı, Kelime bazında cümle uzunluk dağılımı, Harf bazında ortalama kelime uzunluğu, Kelime bazında ortalama cümle uzunluğu, Noktalama işareti sayısı, Stopword sayısı, Tamamı büyük yazılmış kelime sayısı	%51,00
Noktalama işareti sayısı	%4,03	TF-IDF vektörü, Kök sıklığı vektörü, Harf bazında kelime uzunluk dağılımı, Harf bazında ortalama kelime uzunluğu, Kelime bazında ortalama cümle uzunluğu, Noktalama işareti sayısı, Stopword sayısı, Tamamı büyük yazılmış kelime sayısı	%44,00

Şekil 3: Scikit-learn SVC varsayılan parametrelerde farklı özniteliklerin doğruluk puanları

Hiper-parametreler, makine öğrenmesi modellerinde doğrudan öğrenilmeyen parametrelerdir. Çapraz doğrulama daha güvenilir bir başarı düzeyi değerlendirmesini yapabilmek için kullanılır. Sonuçlar eğitim ve test veri setlerine bağımlı olabileceği için farklı veri setleri ile model denenerek elde edilen sonuçların ortalaması alınır. En yüksek doğruluk puanını veren öznitelikler ile SVC için hiper-parametre ayarını çapraz doğrulama ile birlikte yapabilmek için Scikit-learn kütüphanesinde bulunan RandomizedSearchCV modülü kullanılmıştır. İşlem sonucu elde edilen en iyi sonuç, linear çekirdek ve C = 216 değeri ile %85 doğruluk puanıdır.

MLP (Multi-Layer Perceptron), tam bağlantılı, ileri beslemeli yapay sinir ağı (YSA) sınıfıdır. Scikit-learn MLPClassifier modülünü varsayılan ayarlarda kullanarak, en yüksek doğruluk oranı veren öznitelikler ile eğitilmiş ve %76,19 doğruluk oranı elde edilmiştir.

DÖNEM SONU HEDEFLERİNİN DEĞERLENDİRİLMESİ

1. Zemberek kütüphanesi kullanılarak morfolojik analiz yapılmıştır.
2. Yeni öznitelikler eklenmiştir.
3. Öznitelikler makine öğrenmesi için hazırlanmıştır.
4. İlk hedef olan SVM modelinde %57'lik doğruluğa ulaşılmıştır.
5. Daha iyi parametreler elde etmek için hiper-parametre ayarlaması yapılmış ve 5 kat çapraz doğrulama ile %82'lik bir doğruluk sonucuna ulaşılmıştır.

Böylece 2. dönem hedeflerinden tamamlanan hedefler şunlardır:

1. Verilerin SVM modelini eğitecek şekilde düzenlenmesi
2. Yapay zekâ modelinde mevcut problem için en iyi parametrelerin belirlenmesi

Tamamlanmayan hedefler ise şunlardır:

1. Ara yüz geliştirme

KAYNAKLAR

Agün,, Yılmazel S., Yılmazel O., 2017; Effects oflanguage processing in Turkish authorship attribution. 2017 IEEE International Conference on Big Data (Big Data), (1), 1876- 1881, https://doi.org/10.1109/BigData.2017.8258132

Amasyalı, Diri., 2006; Automatic Turkish text categorization in terms of author, genre and gender. Natural Language Processing and Information Systems, Proceedings, 3999, 221-226, https://doi.org/10.1007/11765448_22

Aslantürk, Sezer, Sever, Raghavan, 2010; Application of cascading rough set-based classifiers on authorship attribution. Proceedings - 2010 IEEE International Conference on Granular Computing, GrC 2010, 656-660, https://doi.org/10.1109/GrC.2010.110

Aslantürk, 2014; Tamgacı: artırımsal ve geri beslemeli Türkçe yazar çözümleme (Doktora tezi). Hacettepe Üniversitesi, Bilgisayar Mühendisliği Bölümü.

Bozkurt, Bağlıoglu, Uyar, 2012; Authorship attribution. 22nd International International Symposium on Computer and Information Sciences, 1-5, https://doi.org/10.1109/ISCIS.2007.4456854

Canbay, Sezer, Sever, 2018; Authorship modelling approach for authorship verification on the Turkish texts. 2018 26th Signal Processing and Communications Applications Conference (SIU), İzmir, 1-4, DOI: 10.1109/SIU.2018.8404436

Çalışkan, Can, 2018; "Türkçe Metinler Üzerine Yapılan Sayısal Üslup Araştırmalarını İnceleyen ve Benim Adım Kırmızı Çevirilerinin Aslına Olan Sadakatini Ölçen Bir Çalışma", Türk Kütüphaneciliği, 2018, DOI: 10.24146/tkd.2018.41, erişim tarihi: 19.04.2022

Diri, Amasyalı, 2003; Automatic author detection for Turkish texts. Artificial Neural Networks and Neural Information, 1., https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.551.9280

İşi, Çemrek, Yıldız, 2013; "İstatistik'ten Edebiyat'a Bir Köprü: Stilometri Analizi", Uşak Üniversitesi Sosyal Bilimler Dergisi, 2013, https://dergipark.org.tr/tr/download/article-file/202266, erişim tarihi: 19.04.2022

Küçükyılmaz, Cambazoğlu, Aykanat, Can, 2008; Chat mining: Predicting user and message attributes in computer-mediated communication. Information Processing and Management, 44(4), 1448-1466, https://doi.org/10.1016/J.IPM.2007.12.009

Şenel, 2020; Makine Öğrenmesi Algoritmaları Kullanılarak Kayısı İç Çekirdeklerinin Sınıflandırılması, BEÜ Fen Bilimleri Dergisi, 9 (2), 807-815, 2020, https://dergipark.org.tr/en/download/article-file/1149998

Taş, Görür, 2007; Author identification for Turkish Texts. Çankaya Üniversitesi Fen-Edebiyat Fakültesi, 7, 151-161., https://dergipark.org.tr/tr/download/article-file/45267

Taşçı, Ekinci, 2012; Character level authorship attribution for Turkish text documents. The Online Journal of Science and Technology, 2(3), 12-16.

Türkoğlu, 2006; Melez yaklaşımlarla Türkçe dokümanlarda yazar tanıma. (Yüksek lisans tezi). Yıldız Teknik Üniversitesi, Bilgisayar Mühendisliği Bölümü.

Türkoğlu, Diri, Amasyalı, 2007; Author attribution of Turkish texts by feature mining. Advanced Intelligent Computing Theories, 1086-1093., https://doi.org/10.1007/978-3-540-74171-8_110

Yavanoğlu, 2016; Intelligent authorship identification with using Turkish newspapers metadata. İçinde Proceedings - 2016 IEEE International Conference on Big Data, Big Data 2016, 1895- 1900., https://doi.org/10.1109/BigData.2016.7840809