T.C.
KONYA TEKNİK ÜNİVERSİTESİ
LİSANSÜSTÜ EĞİTİM ENSTİTÜSÜ

ŞİİR KATEGORİSİNİN DOĞAL DİL İŞLEME YÖNTEMLERİ KULLANILARAK TAHMİN EDİLMESİ

Emre YÖNET

YÜKSEK LİSANS TEZİ

Bilgisayar Mühendisliği Anabilim Dalı

Temmuz-2023
KONYA

TEZ KABUL VE ONAYI

Emre YÖNET tarafından hazırlanan "Şiir Kategorisinin Doğal Dil İşleme Yöntemleri Kullanılarak Tahmin Edilmesi" adlı tez çalışması 14/07/2023 tarihinde aşağıdaki jüri tarafından oy birliği ile Konya Teknik Üniversitesi Lisansüstü Eğitim Enstitüsü Bilgisayar Mühendisliği Anabilim Dalı'nda YÜKSEK LİSANS TEZİ olarak kabul edilmiştir.

Jüri Üyeleri	**İmza**
Başkan Doç. Dr. Mehmet Akif ŞAHMAN	……..
Üye Doç. Dr. Ersin KAYA	……..
Danışman Dr. Öğr. Üyesi Sedat KORKMAZ	……..

Yukarıdaki sonucu onaylarım.

Prof. Dr. Mevlüt UYAN
Enstitü Müdürü

TEZ BİLDİRİMİ

Bu tezdeki bütün bilgilerin etik davranış ve akademik kurallar çerçevesinde elde edildiğini ve tez yazım kurallarına uygun olarak hazırlanan bu çalışmada bana ait olmayan her türlü ifade ve bilginin kaynağına eksiksiz atıf yapıldığını bildiririm.

DECLARATION PAGE

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Emre YÖNET

Tarih: 14.07.2023

ÖZET

YÜKSEK LİSANS TEZİ

ŞİİR KATEGORİSİNİN DOĞAL DİL İŞLEME YÖNTEMLERİ KULLANILARAK TAHMİN EDİLMESİ

Emre YÖNET

Konya Teknik Üniversitesi
Lisansüstü Eğitim Enstitüsü
Bilgisayar Mühendisliği Anabilim Dalı

Danışman: Dr. Öğr. Üyesi Sedat KORKMAZ

2023, 87 Sayfa

Jüri
Doç. Dr. Mehmet Akif ŞAHMAN
Doç. Dr. Ersin KAYA
Dr. Öğr. Üyesi Sedat KORKMAZ

Doğal dil işleme; insanlar tarafından kullanılan dillerin (doğal dil) bilgisayarlar tarafından anlaşılmasını, yorum ve cevap üretilmesini sağlayan bilgisayar bilimleri ve dilbilim yöntemlerinin birlikte kullanıldığı bir bilim alanıdır. Doğal dil işleme, insanlarla bilgisayarların etkileşimini artırmak ve özellikle büyük veri setleri üzerinde çalışan kurumlar ve araştırmacılar için doğal dil verilerini analiz etmeyi ve anlamayı kolaylaştırmak amacıyla kendine birçok uygulama alanı bulmuştur. Bu uygulama alanlarından bir tanesi de metinlerin sınıflandırılmasıdır. Bu tezde, metin formatında olan ve farklı kategoride yazılmış olan şiirlerin önceden etiketlenmiş olan kategorilere göre sınıflandırılması üzerine çalışma yapılmıştır. Çalışmada web kazıma yöntemi ile elde edilen ve 4198 adet şiirden oluşan veri seti kullanılmıştır. Veri seti üzerinde 13 farklı doğal dil işleme adımları uygulanmıştır. Söz konusu işlemlerin yapılabilmesi için Zemberek Kütüphanesi kullanılmıştır. Sınıflandırma işlemi için altı farklı makine öğrenmesi algoritması kullanılmış, elde edilen sonuçlar değerlendirilmiş ve model performansını artırmaya yönelik hiperparametre analizi yapılmıştır. Hiperparametre analizi için GridSearchCV ve RandomizedSearchCV yöntemleri kullanılmıştır. Sınıflandırma algoritmalarının sonuçları kıyaslandığında en yüksek doğruluk oranını Random Forest ve SVM algoritmalarının verdiği görülmüştür.

Anahtar Kelimeler: Doğal Dil İşleme, Makine Öğrenmesi, Metin Sınıflandırma, Yapay Zekâ

ABSTRACT

MS THESIS

ESTIMATING THE POETRY CATEGORY USING NATURAL LANGUAGE PROCESSING METHODS

Emre YÖNET

Konya Technical University
Institute of Graduate Studies
Department of Computer Engineering

Advisor: Asst. Prof. Dr. Sedat KORKMAZ

2023, 87 Pages

Jury
Assoc. Prof. Dr. Mehmet Akif ŞAHMAN
Assoc. Prof. Dr. Ersin KAYA
Asst. Prof. Dr. Sedat KORKMAZ

Natural language processing is a field of science that combines methods from computer science and linguistics to enable computers to understand, interpret and respond to human language (natural language). Natural language processing has found many applications to improve human-computer interaction and to make it easier to analyse and understand natural language data, especially for organisations and researchers working with large data sets. One such application is text classification. In this thesis, a study was conducted on the classification of poems in text format and written in different categories according to pre-labelled categories. The study used a dataset of 4198 poems obtained by web scraping. Thirteen different natural language processing steps were applied to the dataset. The Zemberek library was used to perform these operations. Six different machine learning algorithms were used for classification, the results obtained were evaluated and hyperparameter analysis was performed to improve model performance. The methods GridSearchCV and RandomizedSearchCV were used for the hyperparameter analysis. When the results of the classification algorithms were compared, it was found that the Random Forest and SVM algorithms gave the highest accuracy rate.

Keywords: Machine Learning, Natural Language Processing, Text Categorization, Artificial Intelligence

ÖNSÖZ

Yüksek lisans öğretim süresi boyunca göstermiş olduğu anlayış ile motive eden, her zaman yol gösteren ve deneyimlerini paylaşan danışman hocam Dr. Öğr. Üyesi Sedat KORKMAZ'a, yapmış olduğum çalışmalarda desteklerini hiçbir zaman esirgemeyen başta annem ve babam olmak üzere aileme ve arkadaşlarıma teşekkürlerimi sunuyorum.

Emre YÖNET
KONYA-2023

İÇİNDEKİLER

SİMGELER VE KISALTMALAR

Kısaltmalar

AI: Artificial Intelligence
BERT: Bidirectional Encoder Representations from Transformers
BOW: Bag-Of-Words
CNN: Convolutional Neural Network
CRF: Conditional Random Field
DDİ: Doğal Dil İşleme
DT: Decision Tree
GRUs: Gated Recurrent Units
K-NN: K-Nearest Neighbours
LR: Logistic Regression
LSTM: Long Short-Term Memory
ML: Machine Learning
NBM: Naive Bayes Multinomial
NLP: Natural Language Processing
RF: Random Forest
SVM: Support Vector Machine
TF-IDF: Term Frequency–Inverse Document Frequency
YSA: Yapay Sinir Ağları

1. GİRİŞ

İnsanlar yaşamları boyunca diğer insanlarla anlaşmaya ve haberleşmeye ihtiyaç duymaktadır. İnsanlar bu ihtiyacını karşılayabilmek için dil olarak adlandırılan iletişim aracını kullanmaktadır. Bu bağlamda dil; insanın bireysel yaşamında ve farklı toplumlar arasında kurulan etkileşimde çok önemli bir yere sahiptir. Dil olmadan birbirinden farklı işaretler ve hareketlerle de iletişim kurulabilmesi mümkün olsa da en etkili ve en basit yöntem dil ile kurulan iletişimdir. Dil olmadan kurulan iletişimin hızlı ve etkili olması beklenemez.

Dil belirli bir kural yapısı olan ve sözcüklerin bu kurallara göre bir araya gelmesinden oluşan temelde seslere dayanan bir iletişim aracıdır. Diğer bir ifadeyle bir toplumun varoluşunun en önemli kavramıdır. Mustafa Kemal Atatürk, dilin toplum için ne anlama geldiğini Türk diline verdiği büyük önem kapsamında yaptığı çalışmalar ve söylediği sözler ile ortaya koymuştur. Türk alfabesinin 1 Kasım 1928 tarihinde kanunlaşarak resmen yürürlüğe girmesi, bizzat Atatürk'ün talimatıyla 12 Temmuz 1932 tarihinde Türk Dili Tetkik Cemiyeti'nin kurulması akla gelen ilk çalışmalarıdır. Türk Dili Tetkik Cemiyeti'nin adı ilerleyen yıllarda Türk Dil Kurumu (TDK) olarak değiştirilmiştir. Türk Dil Kurumu'nun amacı "Türk dilinin öz güzelliğini ve zenginliğini meydana çıkarmak, onu yeryüzü dilleri arasında değerine yaraşır yüksekliğe eriştirmek" olarak belirlenmiştir. Mustafa Kemal Atatürk "Milli duygu ile dil arasındaki bağ çok kuvvetlidir. Dilin milli ve zengin olması, milli duygusunun gelişmesinde başlıca etkendir. Türk dili, dillerin en zenginlerindendir, yeter ki bu dil bilinçle işlensin. Ülkesini yüksek bağımsızlığını korumasını bilen Türk milleti, dilini de yabancı diller boyunduruğundan kurtarmalıdır." ve "Milli bilincin ayakta kalabilmesi ve uyanık bulunması için dil ve tarih uğrunda çalışmaya mecburuz. Türk milletinin milli dili ve milli benliği bütün hayatında egemen ve esas kalacaktır." sözleriyle Türk diline verdiği önemi vurgulamıştır (Tarim, 2013; TDK, 2023c).

Dil kavramı denince insanların kendi aralarında anlaşmak için kullandıkları diller akla ilk gelmektedir. Bu dillere yaygın olarak kullanılmakta olan İngilizce Arapça, Fransızca, Çince, İspanyolca, Farsça, Türkçe başta olmak üzere diğer birçok dil örnek olarak verilebilir. Türkçe ana dili Türkiye olmak üzere toplam sekiz ülkede konuşulan bir dildir. Dünya üzerinde konuşan sayısına bakıldığında başta İngilizce, Arapça, Çince, İspanyolca dilleri gelmektedir. Fransızca, konuşan sayısı bakımına kıyasla bu dillerin

arkasından gelse de konuşulduğu ülkelerin coğrafi konumları dikkate alındığında uluslararası alanda yaygın bir bölgede kullanılmaktadır (Uzun, 2012).

Çizelge 1.1. Diller ve konuşulduğu ülkeler

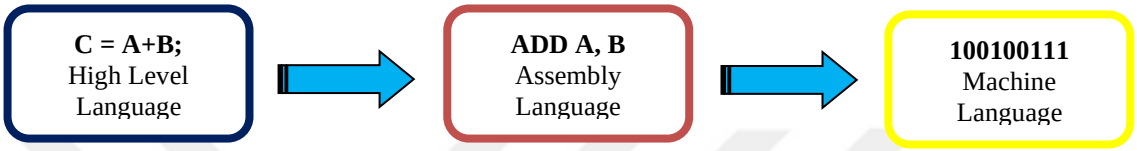
S/N	Dil	Ana Ülke	Toplam Ülke Sayısı
1.	İngilizce	İngiltere	101
2.	Arapça	Suudi Arabistan	59
3.	Fransızca	Fransa	51
4.	Çince	Çin	33
5.	İspanyolca	İspanya	31
6.	Farsça	İran	29
7.	Almanca	Almanya	18
8.	Rusça	Rusya Federasyonu	16
9.	Malayca	Malezya	13
10.	Portekizce	Portekiz	11
11.	İtalyanca	İtalya	10
12.	Türkçe	Türkiye	8
⋮	⋮	⋮	⋮

Bir ülkede birden fazla dil konuşulabilmektedir. Örneğin Papua Yeni Gine Bağımsız Devleti'nde birden fazla resmi dil kullanılmaktadır. Bu diller Tok Pisin, İngilizce, Hiri Motu, Papua New Guinean Sign Language olarak sıralanabilir. Bu ülkenin diğer bir özelliği ise dünya üzerinde en çok dilin kullanıldığı ülkelerin başında gelmesidir. Kültürel çeşitlilik açısından çok zengin olan Papua Yeni Gine'de bilinen 851 dil bulunmaktadır. (Vikipedi, 2023) Ülkemizde ise 46 farklı dil bulunmakta olup en çok dilin bulunduğu ülkeler arasında ilk 50 içerisinde yer almaktadır. Bu diller Türkçe, Kürtçe, Arapça, Çerkezce, Rumca, Gürcüce, Ermenice, Lazca ve diğer diller olarak sıralanabilir. Türkçe, ülke nüfusunun yaklaşık %90'ı tarafından kullanılan bir dildir.

Bilgisayar bilimlerinde dil kavramı programlama dilleri (JavaScript, HTML, CSS, SQL, Java vb.) ve insanlar tarafından kullanılan diller (doğal dil) olarak iki ana başlıkta karşımıza çıkmaktadır. Bilgisayarlar kendilerine verilen komutları anlayabilmek ve bu komutlara karşılık olarak geri dönüş sağlayabilmek için kullanıcılar ile iletişim kurmaktadırlar. Bilgisayarların anlayabildiği makine dili 0 ve 1'lerden oluşmaktadır. Makine dilinin detayları tasarım aşamasında üretici tarafından belirlenmektedir (Eldeniz, 1994). Kullanıcı tarafından verilen komutlar ve herhangi bir programlama dilinde yazılmış programlar bilgisayarlar tarafından anlaşılıp işlem yapılabilmesi için makine diline çevrilirler. Programlama dilini üst seviye, makine dilini ise alt seviye olarak kabul ettiğimizde bu iki seviye arasında dönüşüm derleyici tarafından gerçekleştirilir.

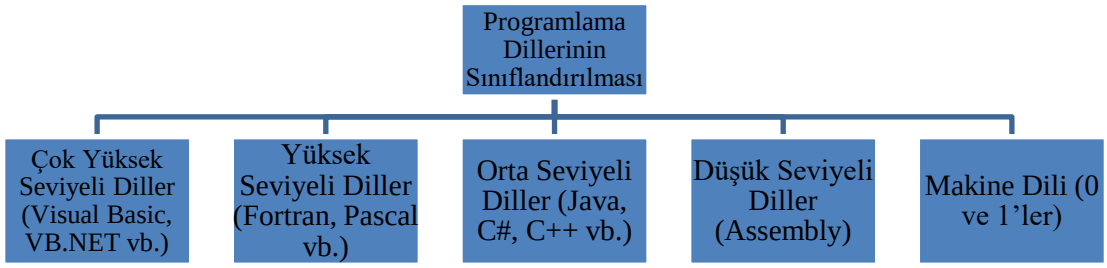
Şekil 1.1. Programlama dilleri ve özellikleri

Her programlama dili kendine özgü yazım kurallarına sahiptir. Tıpkı insanlar tarafından kullanılan dillerde olduğu gibi. Şekil 1.2'de bulunan kod parçaları aynı anlama gelse de farklı kurallar çerçevesinde yazılmıştır.



Şekil 1.2. Programlama dilleri ve kuralları

Programlama dillerini seviyelerine göre sınıflandırmak mümkündür. İnsanların anlaması ile bilgisayarların anlaması aynı olmamaktadır. İnsanlar tarafından anlaşılması daha kolay olan diller yüksek seviyeli programlama dilleri, insanlar tarafından anlaşılması daha zor olan diller ise düşük seviyeli programlama dili olarak sınıflandırılmaktadır. Şekil 1.3'de programlama dillerinin sınıflandırılması şeması yer almaktadır (Eldeniz, 1994).



Şekil 1.3. Programlama dillerinin sınıflandırılması

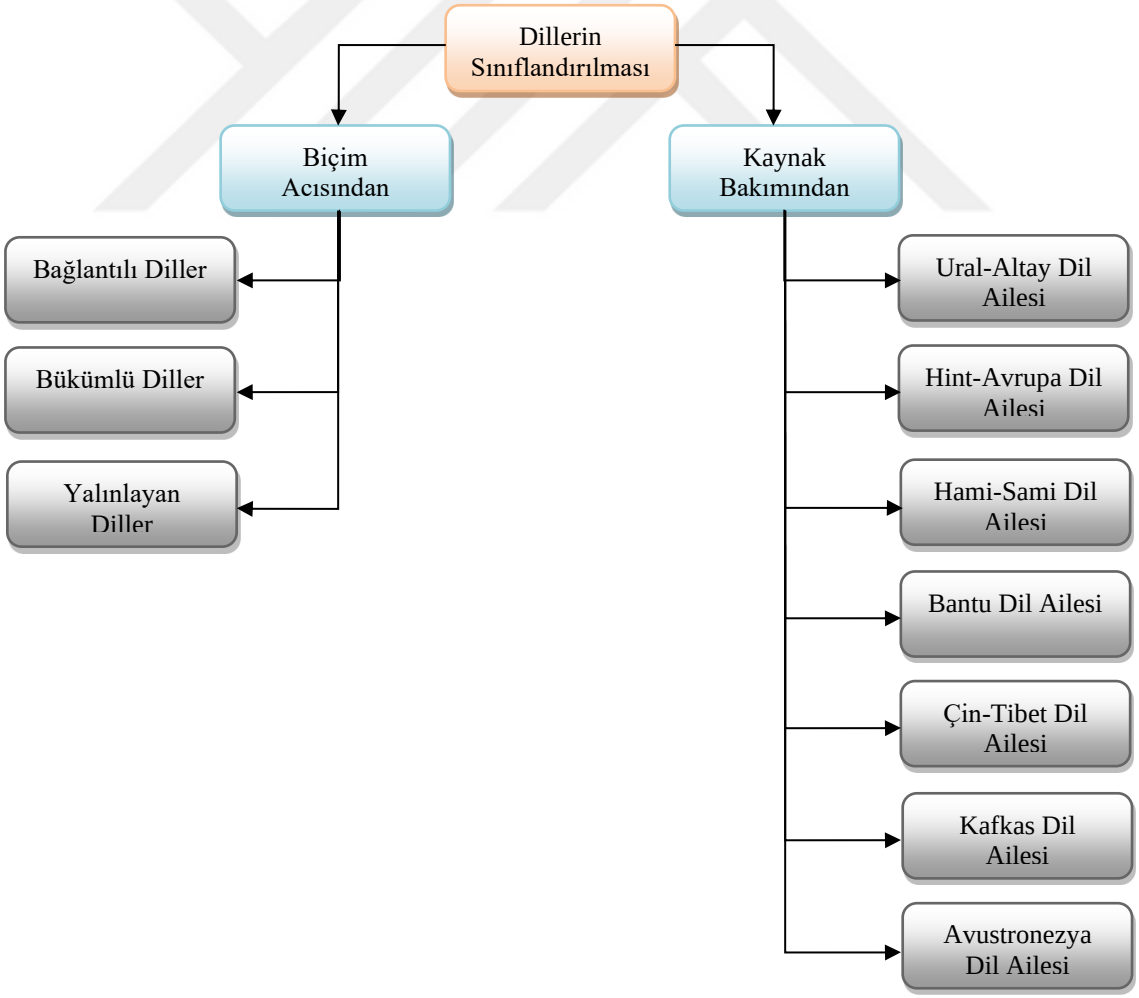
Dilin yaygın olarak kabul edilen tanımı; insanlar tarafından kendi aralarında iletişim kurmak için kullandıkları bir araçtır. Bu iletişim konuşma dilleri ile sağlanabileceği gibi işitme engelli bireyler tarafından kullanılan işaret dilleri aracılığıyla da sağlanabilmektedir. Hem konuşma dilleri hem de işaret dilleri ülkeden ülkeye farklılıklar gösterebilmektedir. İnsanlar tarafından kullanılan bu diller bilgisayar bilimleri dünyasında doğal dil olarak tanımlanmaktadır. Teknolojinin hızla gelişmesi ve dolayısıyla bilgisayarların insan hayatında çok önemli bir yere sahip olması beraberinde insan-bilgisayar iletişimini zorunlu hale getirmiştir. İnsanların kendi aralarında kullandıkları dil birbirlerine kolay ve anlaşılır gelmektedir ancak makineler tarafından doğal dillere herhangi bir işlem yapılmadığında anlaşılması mümkün olmamaktadır. Bunun için doğal dilin belirli kurallar çerçevesinde anlaşılır hale getirilmesi gerekmektedir. İki farklı dili konuşan tarafların iletişim kurmasının zor olduğu düşünülse de yapay zekânın alt dalı olan doğal dil işleme sayesinde bu iletişim mümkün hale gelmektedir. Doğal dil işleme çalışmaları kapsamında konuşma dili, yazılı metinler (kitaplar, makaleler, gazeteler, köşe yazıları) ve işaret dili gibi doğal dilde oluşturulmuş kaynaklar bilgisayarlara girdi olarak verilmekte ve bilgisayar tarafından doğal diller anlaşılmakta, yorumlanmakta ve bunun sonucu olarak çıktı üretilebilmektedir. Ses verileri üzerinde çalışma yapılacağı zaman bu verilerin metin verilerine dönüştürülerek yazılı hale getirilen metinler üzerinden işlem yapılması tercih edilmektedir. Doğal dil İşleme ile bu işlemler yapılırken veri bilimi, dilbilim vb. çok sayıda bilim dalından faydalanılmaktadır. Doğal dil işleme pek çok alanda kullanılmakta ve günlük yaşantımızda kolaylık sağlamaktadır. Metin sınıflandırma, makine çevirisi, soru cevaplama, metin özetleme, yazım denetimi vb. uygulama alanlarından bazılarıdır. Onlarca sayfa bir dokümandan sadece birkaç sayfadan oluşan bir özet çıkarmak istediğimizde doğal dil işleme devreye girmekte ve bu işlemi çok kısa bir süre içerisinde yerine getirebilmektedir.

Bu tez çalışmasında metin sınıflandırma üzerinde çalışılmıştır. İçinde yaşadığımız dijital çağda bilgisayar ortamında üretilmiş çok sayıda metin bulunmaktadır. Bu metinlerin çoğu yapılandırılmamış olduğundan, bu verileri sınıflandırmak son derece yararlı olmaktadır. Metin sınıflandırma, çeşitli konularda yazılmış metinlerin önceden belirlenen etiketler ile etiketlenmesidir. Diğer bir ifadeyle metinlerin önceden belirlenmiş olan kategorilere göre sınıflandırılması işlemidir. Metin sınıflandırma işlemi doğal dil işleme yöntemleri kullanılarak otomatik olarak analiz edilir ve belirli bir sınıfa göre etiketlenir. Bu çalışmada metin olarak farklı kategoride

yazılmış olan şiirler kullanılmıştır. Metinden çıkarılan özellikler metin temsil yöntemleri kullanılarak işlemden geçirilmiş ve birden fazla makine öğrenmesi yöntemi kullanılarak sınıflandırma yapılmıştır. Elde edilen sınıflandırma sonuçları mukayese edilmiştir.

1.1. Türkçe ve Doğal Dil İşleme

Diller, Şekil 1.4'de görüldüğü üzere genel olarak biçim açısından ve kaynak bakımından (köken) iki gruba ayrılırlar (Ercilasun, 2013). Biçim açısından; bitişken diller ve kaynaştıran diller, yalınlayan diller ve bükümlü diller olarak üç grupta sınıflandırılmaktadır. Kaynak bakımından ise Ural-Altay Dil Ailesi, Hint-Avrupa dil ailesi, Hami-Sami dil ailesi, Bantu dil ailesi, Çin-Tibet dil ailesi, Kafkas dil ailesi ve Avustronezya dil ailesi olmak üzere yedi grupta sınıflandırılmaktadır.



Şekil 1.4. Dünya dillerinin sınıflandırılması (Ercilasun, 2013)

Türkçe, kaynak bakımından Ural-Altay dil ailesinin Altay koluna mensup, biçim açısından ise bağlantılı diller grubunda yer alan ve dünya üzerinde yaklaşık 220 milyon kişi tarafından konuşulan dildir (Akalın, 2009). Altay dil ailesi içerisinde Türk dilleri (Türkçe, Çuvaşça, Özbekçe, Uygurca vb.), Moğol dilleri (Darkhat, Buryat, Khamnigan Mongol vb.) ve Tunguz dilleri (Solon, Manegir, Nanai, Akani, Birar, Kile) yer almaktadır. Geçmişte akraba ya da kardeş olan bu diller "Ana Türk Dili"nden türemiştir (Tekin, 1978).

Türkçe dilini diğer dillerden ayıran bazı tipolojik özellikleri bulunmaktadır. Sondan eklemeli bir dil olup kelimeler aldıkları eklerle yeni anlamlar kazanabilmektedir. Türkçe bu özelliği dolayısıyla bitişken bir dil olarak tanımlanabilmektedir. Bunun yanı sıra eklerin art arda sıralanabilmesi, eklerin kademeli olarak niteleme yapabilmesi, eklerin esnek bir yapıya sahip olması, isimlerin bitişik yapıda olabilmesi (ekmek dolabı gibi), belirtili isim tamlaması, gramatikal cinsiyet bulunmaması Türkçenin başlıca genetik özellikleridir. Türkçenin genetik özellikleri doğal dil işleme açısından çeşitli zorlukları beraberinde getirmektedir (Oflazer, 2016). Bu zorluklar Türkçe üzerinde yapılan doğal dil işleme çalışmalarının sayısını olumsuz etkilemektedir. Türkçe diline ait bazı fonetik özelliklere de yer vermek faydalı olacaktır. Türkçede büyük ünlü uyumu-küçük ünlü uyumu bulunmaktadır. Sekiz adet ünlü harf bulunmakta olup bu bakımdan zengin bir dildir. Bazı ünlü harfler sadece ilk hecede bulunabilirken diğerleri tüm hecelerde bulunabilmektedir. Birden fazla ünsüz harf bir arada bulunmamaktadır. Türkçe hem açık hem de kapalı hecelere sahiptir. Türkçede başka fonetik özellikler de bulunmakta olup bunlar ilk akla gelenlerdir. Türkçe dilinde yapılan doğal dil işleme çalışmalarına ikinci bölümde (Kaynak araştırması) detaylı olarak yer verilmiştir.

1.1.1. Türkçe biçimbilim yapısı

Dilbilgisi bilimde sözcüklerin yapısını inceleyen bileşen biçimbilim (Morphology) alt dalıdır. Biçimbilim, sözcüklerde meydana gelen biçimsel ve anlamsal değişiklikleri incelemektedir. Sözcüklerin sözlük anlamı ve dil bilgisi yapısında değişim gözlemlenebilmektedir. Sözcük yapısında değişim meydana geldiğinde bu durumun bilgisayarlar tarafından çözümlenmesi gerekmektedir. Örneğin "kalemler" sözcüğünün kökünün "kalem" olduğu bilgisayarlar tarafından yapılacak olan çözümleme sonucunda

belirlenebilmektedir. Bu durum morfolojik çözümleme olarak tanımlanmaktadır. Birçok dilde morfolojik çözümleme yapabilmek için sözcüğün bulunabileceği formları listelemek yeterli olmaktadır. Türkçe dilinde bu yöntem etkili bir çözüm olmamaktadır. Türkçede bir sözcük birden fazla ek alabildiğinden farklı anlamlar kazanabilmektedir. Bu durum diğer dillerle kıyaslandığında, Türkçe bir sözcüğün başka bir dilde kurulmuş cümlenin anlamını tek başına karşılayabildiği görülmektedir. Türkçe dilinin bu özelliği göz önünde bulundurulduğunda sözcüklerin alabileceği ekleri listelemek çok zordur.

Çizelge 1.2. Ev sözcüğünün çekimli halleri

ev	evler	evim	evin
evi	evimiz	eviniz	evlerde
evlerim	evlerin	evleri	evdeler
evleriniz	evdeyim	evde	evdesiniz

Çizelge 1.2'de ev sözcüğünün bazı çekimli halleri bulunmaktadır. Çizelgeden de görüldüğü üzere, Türkçede bir sözcük onlarca çekim alabilmektedir. Ev sözcüğünün çekimli hallerini artırmak mümkündür.

Sözcükler biçimbilim açısından incelendiğinde birden fazla çözümleme yapılma durumları ile karşılaşılabilmektedir. Örneğin "anı" sözcüğü iki farklı biçimde çözümlenebilir. Birincisi "Zamanın bölünemeyecek kadar kısa olan parçası, lahza, dakika" anlamına gelen "an" kelimesi, ikincisi ise "Geçmişte yaşanmış çeşitli olaylardan belleğin sakladığı her türlü iz, hatıra" anlamına gelen "anı" kelimesi. Hangisinin doğru çözümleme olduğuna, cümle içerisinde kullanımına bakılarak karar verilebilir. "Anı yaşamak bana keyif veriyor." cümlesindeki "anı" sözcüğü çözümlendiğinde birinci anlamda kullanıldığı görülmektedir. "Vakit geçirdiğim kısa süre içerisinde çok sayıda anı biriktirdim bu şehirde." cümlesindeki "anı" sözcüğü çözümlendiğinde ikinci anlamda kullanıldığı görülmektedir. Buna göre çözümleme yapılması doğru olacaktır. Olası bütün çözümlemeler bulunduktan sonra doğru olan çözümlemeye karar verilmesi gerekmektedir. DDİ uygulamalarında sözcüklerin çözümlenmesine ihtiyaç duyulmaktadır. Örneğin bir arama motoru üzerinden internette arama yapıldığında sözcüğün çekimli hallerinin de taranması doğru sonuca ulaşmak için önem arz etmektedir. "Doğayı sevenler derneği" şeklinde bir arama yapılmak istendiğinde, "Doğayı seven derneği" ve "Doğa sevenler derneği" aramalarının da benzer sonuçlar vermesi beklenmektedir. Bu durum için detay seviyesinde biçimbilim çözümlemesi yapılmasına gerek olmasa da sözcüklerin köklerinin doğru belirlenmesi

gerekmektedir. Sözcüklerin çözümlenmesine ihtiyaç duyulan bir diğer uygulama alanı da yazım hatalarının denetlenmesidir. Dijital ortamda belge hazırlarken yazım yanlışları veya imla hataları sıkça yapılan bir davranıştır. Yanlış yazılan sözcüklerin denetlenmesi ve doğru yazılışının bilgisayar tarafından önerilebilmesi için morfolojik çözümleme yapılmalıdır. Bu problem için farklı çözümler de bulunmaktadır. Örneğin ilgili dile ait sözlükte yer alan sözcükler taranarak yazım yanlışı ve doğru yazılışı tespit edilebilir ancak Türkçe gibi dillerde böyle bir sözlük hazırlamak mümkün değildir. Sözcüklerin morfolojik olarak analiz edilmesi (çözümlenmesi), doğal dil işleme alanında çalışma yapabilmek için en temelde yapılan bir işlemdir. Morfolojik çözümleme yapılabilmesi noktasında çeşitli yaklaşımlardan ve bu alanda yapılan çalışmalardan bahsetmek mümkündür.

Gülşen Eryiğit ve Eşref Adalı tarafından 2004 yılında yapılan çalışmada Türkçe için morfolojik çözümleyici tasarımı ve uygulaması üzerine çalışılmıştır. Yapılan bu çalışmada dilin kurallı yapısı temel alınarak bir çözümleyici geliştirilmiştir. Geliştirilen çözümleyici kelimenin köküne ulaşmayı hedeflemektedir Bu hedefe ulaşmak için, kelimenin alabileceği ekler (yapım ekleri, çekim ekleri vs.) veri tabanında tutulmaktadır. Türkçe sözcüklerin çözümlemesini, eklerin sözcükten çıkarılması yaklaşımıyla ve herhangi bir sözlük kullanmadan yapmak için yeni bir yöntem önermişlerdir. Sözlük kullanmak uygulama performansını etkileyebilecek bir kıstastır (Eryiğit ve Adalı, 2004).

Prof. Dr. Aydın Köksal tarafından yapılan *"Türkçenin Özdevimli Biçimbilgisi Çözümlemesine ilk Yaklaşım"* adlı çalışma Türkçenin biçimbilim açısından bilgisayar ortamında incelenmesi üzerinedir (Köksal, 1975).

Kemal Oflazer tarafından 1994 yılında iki seviyeli biçimbilim yaklaşımı ile Türkçenin çözümlenebilmesi üzerine bir çalışma yapılmıştır (Oflazer, 1994).

Ahmet Afşin Akın tarafından geliştirilen Zemberek Kütüphanesi Türkçe metinler üzerinde morfolojik analiz yapılabilmesine olanak sağlamaktadır. Java ortamında geliştirilen ve açık kaynak olan Zemberek ile yazım denetimi, hatalı kelimeler için öneri gibi işlevler yerine getirilebilmektedir (Akın ve Akın, 2007).

Sezgi Yılmaz tarafından 2009 yılında yapılan yüksek lisans tez çalışması kapsamında, var olan çözümleyiciler detaylı olarak incelenmiş ve bu çözümleyicilerin eksik olan yönleri iyileştirilerek yeni bir çözümleyici tasarlanmıştır (Yılmaz, 2009).

Diğer diller için, Daniel Jurafsky ve James H. Martin (2009), Brodda ve Fred (1980), Kaalep (1997) ve Packard (1973) tarafından yapılan çalışmalar örnek olarak gösterilebilir (Eryiğit, 2012).

1.1.2. Türkçe dilbilgisi kuralları

Her dilin kendine özgü kuralları bulunmaktadır. Bir dilde metin oluşturulmak istendiğinde o dilin kurallarına uymak gerekmektedir. Dilbilgisi kurallarına uyulmadan veya dikkat edilmeden oluşturulan metinler ve kurulan cümleler anlamlı olmamaktadır. Cümle içerisinde kullanılan kelimelerin dizilişi, noktalama işaretleri, kelimelerin yazılışı vb. durumlarda bu kurallar dikkate alınmaktadır. İmla kuralları olarak da bilinen bu kurallar kişiden kişiye göre değişmemektedir. Bu kurallar olası yazım yanlışlarını ve yanlış telaffuzu önlemekte dolayısıyla okumayı ve anlamayı kolaylaştırıcı bir rol üstlenmektedir.

1.1.2.1. Büyük ünlü uyumu

Bir kelimenin ilk hecesinde *a, ı, o, u* kalın ünlülerden bir tanesi varsa diğer hecelerdeki ünlülerin de kalın, ilk hecesinde *e, i, ö, ü* ince ünlülerden bir tanesi varsa diğer ünlüler de ince olur. Silgi, vergi, nine, oyuncak, belge gibi kelimeler büyük ünlü uyumuna uymaktadır. Türkçede büyük ünlü uyumuna uymayan kelimeler de bulunmaktadır. Elma, kardeş gibi kelimeler büyük ünlü uyumuna uymamaktadır. Başka dillerden Türkçeye geçmiş olan kelimelerde ve bitişik yazılan kelimelerde büyük ünlü uyumu aranmamaktadır. Çizelge 1.3'de bulunan ekler büyük ünlü uyumuna uymayan eklerdir.

Çizelge 1.3. Büyük ünlü uyumuna uymayan ekler

-gil	-ken	-leyin
-daş (-taş)*	-mtırak	-yor

(*) Bazı kelimelerde (ülkü-daş, gönül-daş)

Türkçeye başka dillerden giren (gazete, kalem, telefon vb.) ve son hecesinde kalın ünlü bulunduran ve son sesleri ince telaffuz edilen sözcükler ince ünlü ek alır (helal/helalimiz, idrak/ idrakimiz vb.) (TDK, 2023b).

1.1.2.2. Küçük ünlü uyumu

Türkçede ünlü harfler düz ünlü (a, e, ı, i) ve yuvarlak ünlü (o, ö, u, ü) olarak iki sınıfa ayrılabilir. Küçük ünlü uyumunun olduğu durumlar şu şekildedir; bir sözcüğün ilk hecesinde düz ünlü bulunuyorsa sonraki hecelerde de düz ünlü bulunduğunda, bir sözcüğün ilk hecesinde yuvarlak ünlü ve sonraki ilk hecede geniş düz ünlü (a, e) veya dar yuvarlak ünlü (u, ü) bulunuyorsa küçük ünlü uyumu bulunmaktadır. Boyunduruk, yuvarlak gibi kelimeler örnek olarak gösterilebilir. Türkçede küçük ünlü uyumuna uymayan sözcükler de bulunmaktadır. Yağmur, kavun, avuç, savurmak gibi bazı kelimeler küçük ünlü uyumuna aykırı kelimelerdir. Başka dillerden Türkçeye geçmiş olan kelimelerde küçük ünlü uyumu aranmamaktadır. Çizelge 1.4'de kurallar bir bütün olarak verilmiştir (TDK, 2023a).

Çizelge 1.4. Ünlü uyumu

a → a, ı	o → u, a
e → e,i	ö → ü, e
ı → ı, a	u → u, a
i → i, e	ü → ü, e

1.1.2.3. Ünlü daralması

Türkçe bir fiilin *"a"* veya *"e"* ünlüsü ile bitmesi durumunda söz konusu fiilin şimdiki zamanda çekiminde *"a"* ünlüsü *"ı, u"*, *"e"* ünlüsü ise *"i, ü"* ünlüsüne dönüşür. Örneğin "izle-" fiili şimdiki zamanda çekimlendiğinde "izleyor" yerine ünlü daralması kuralı dolayısıyla "izliyor" şeklinde çekimlenir. Diğer bir örnek "ilerle-" fiili şimdiki zamanda çekimlendiğinde "ilerleyor" yerine ünlü daralması kuralı dolayısıyla "ilerliyor" şeklinde çekimlenir. Bu kurala ünlü daralması adı verilmektedir. Bir fiilin birden çok heceli olması ve "a, e" ünlüsü ile bitmesi durumunda ünlü bir harfle başlayan ek aldığında söz konusu fiilin telaffuzunda "a, e" ünlülerinde daralma görülse de yazımında bu daralmaya yer verilmez. Örneğin "başla-" fiili "başlayan" olarak yazılsa da genellikle "başlıyan" olarak telaffuz edilir. Buna karşın tek heceli olan "de-" ve "ye" fiillerinde telaffuzda yer alan "i" ünlüsü fiillerin yazımında da yer alır. Örneğin; "yiyor" ve "diyor" sözcüklerinde olduğu gibi (TDK, 2023k).

1.1.2.4. Ünlü düşmesi

İki heceli olan ve ikinci hecesinde "ı, i, u, ü" dar ünlülerinden bir tanesi olan sözcükler ünlü bir harfle başlayan bir ek aldığında ikinci hecesindeki dar ünlünün düşmesi ünlü düşmesi olarak adlandırılır. Örneğin; "karın" kelimesi ek aldığında "karnı" ve "fikir" kelimesi ek aldığında "fikri" haline dönüşür. Buna ilave olarak aşağıdaki çizelgede yer alan sözcükler ek aldıklarında son hecelerinde bulunan ünlü harflerde düşme görülmez (TDK, 2023j). Çizelge 1.5'de ünlü düşmesi görülmeyen bazı kelimeler verilmiştir.

Çizelge 1.5. Ünlü düşmesi görülmeyen bazı kelimeler

aşağı	yukarı	ora
şura	bura	ileri
içeri	dışarı	beri

1.1.2.5. Ünsüz harfler

Türk alfabesinde sekiz adet sesli (ünlü) harf, yirmi bir adet sessiz (ünsüz) harf olmak üzere yirmi dokuz harf bulunmaktadır. Yirmi bir adet ünsüz harfin on üç tanesi yumuşak ünsüz (b, c, d, g, j, l, m, n, r, v, y, z) geriye kalan sekiz adeti ise sert (p, ç, k, h, t, ş, f, s) ünsüzden oluşmaktadır. Türkçe kökenli sözcüklerin sonunda "b, c, d, g" ünsüz harfleri yer almaz. Bazı sözcükler (p, ç, t, k ünsüzleri ile biten sözcükler) ünlü ile başlayan ek aldıklarında yumuşama (b, c, d, g ünsüzlerinden birine dönüşür) olmaktadır. Örneğin; "kitap" kelimesi "ı" iyelik eki aldığında "kitabı" sözcüğüne dönüşür. Ancak Türkçeye başka dillerden geçen sözcüklerde yumuşama olmamaktadır. Örneğin; "sepet" (Farsça seped) kelimesi "i" iyelik eki aldığında "sepeti" sözcüğüne dönüşür. Tek heceli sözcüklerin bazılarında "yumuşama" olurken bazı sözcüklerde sonda bulunan ünsüz harfler olduğu gibi korunur. Örneğin; "kurt" sözcüğü ek aldığında "kurdu" sözcüğüne dönüşürken "saç" kelimesi ek aldığında "saçı" sözcüğü olarak orijinal hali korunarak ek alır. Türkçede; sert ünsüz harflerin bir tanesiyle biten sözcüklerin ek aldığında sert ünsüz bir harfle başlaması, yumuşak ünsüz harflerin bir tanesiyle biten sözcüklerin ek aldığında yumuşak ünsüz bir harfle başlaması ünsüz uyumu olarak tanımlanır. Örneğin; "düş-kün" ve "bil-gi" sözcüklerinde ünsüz uyumu görülmektedir (TDK, 2023i). Türkçeye başka dillerden geçen sözcükler tek ünsüzle kullanılır. Bu sözcükler ünlü bir

harfle başlayan ek aldıklarında ünsüz türemesi görülür. Örneğin; “his/hissetmek” ve “af/affetmek” sözcüklerinde ünsüz türemesi görülmektedir (TDK, 2023h).

1.1.2.6. Büyük harflerin yazımı

Türkçede cümleler, dize, özel adlar, iki noktadan (:) sonra gelen cümle, kişi adlarından sonra gelen saygı sözleri, yapılar, yapıtlar ve tüm yer adları, kitap/dergi adları, devlet adları, millet adları, kurum ve kuruluş adları, dil ve lehçe adları ve tırnak içerisinde yapılan alıntı büyük harfle başlamaktadır. Buna ilave olarak din ve mezhep adları, kanun, tüzük, yönetmelik, yönerge, genelge sözcükleri, milli ve dini bayram isimleri ve bayram niteliği kazanmış günlerin isimleri ve belirli bir tarih bildiren ay ve gün adları büyük harfle başlamaktadır. Ancak cümleye sayıyla başlandığında sayıdan sonra gelen sözcük ve akrabalık bildiren sözcükler küçük harfle başlar. Cümleye mümkün olduğunca sayıyla başlanmaması esastır (Anonim, 2023b).

1.1.2.7. Sayıların yazımı

Sayılar harfle veya rakamla yazılabilir. Ancak para miktarı, istatistiksel veriler gibi sayılar yazılırken rakam kullanılmaktadır. Saatler, dakikalar ve uzun basamaklı sayı bildiren sözcükler (milyon, milyar gibi) tercihen yazıyla yazılabilir. İki veya daha fazla sözcükten oluşan sayılar ayrı yazılırken çek, dekont veya senet gibi belgelerde yer alan sayılar bitişik (yüzyirmibir) yazılır. Özel simgelerle (1÷2) beraber kullanılan sayılarla simge arasında boşluk bırakılmaz. Tarihî olaylarda, kitapların, tezlerin ve dergilerin belirli bölümlerinde ve diğer bazı durumlarda Romen rakamı kullanılır. Kesirli sayılar virgülle (,) ayrılırken dört veya daha fazla basamaklı sayılar sondan üçlü gruplar oluşturularak aralarına nokta (.) konur. Sıra bildiren sayılar tercihen yazıyla veya rakamla yazılabilir. Rakamla belirtilen sıra sayılarından sonra nokta konur. Yazıyla bildirilen sıra sayılarından sonra gelen ek kesme işareti ile ayrılır. Bir rakam üleştirme eki aldığında sayıyla yazılır (TDK, 2023g).

1.1.2.8. Ek veya bağlaç olabilen “ki” yazımı

Ek olan “ki” bitişik yazılırken bağlaç olan “ki” ayrı yazılır. Ancak bazı durumlarda “ki” bağlacının bitişik yazıldığı durumlar bulunmaktadır. Örneğin;

“mademki” sözcüğünde “ki” bağlaç olmasına rağmen kalıplaşmış sözcük olduğu için bitişik yazılmaktadır (TDK, 2023f).

1.1.2.9. Soru olan “mi” yazımı

“mi” edatı soru eki olarak kullanıldığında ve başka amaçla kullanıldığı durumlarda ayrı olarak yazılır. “mi” edatı ek aldığında aldığı ek ile bitişik yazılır (misin, mısın, musun) ve kendinden önceki sözcüğe bağlı olarak ünlü uyumuna uyar (TDK, 2023e).

1.1.2.10. Ek veya bağlaç olabilen “de” yazımı

Ek olan “de” bitişik yazılırken bağlaç olan “de” ile “ya” sözcüğü ile kullanılan “de” ayrı yazılır ve kendinden önceki sözcüğe bağlı olarak ünlü uyumuna uyar. Bağlaç olarak kullanılan “de” cümleden çıkarıldığında anlamda bozulma olmazken ek olarak kullanılan “de” cümleden çıkarıldığında anlamda bozulma olur.

1.1.2.11. Mastar eklerinin yazımı

“-mak” veya “-mek” mastar eklerinden bir tanesi ile biten sözcükler “a, e, ı, i” ünlülerinden bir tanesini ek olarak aldığında araya “y” ünsüzü eklenir. Örneğin; “koşmak” sözcüğü “a” ünlüsünü ek olarak aldığında “koşmaka” değil “koşmaya” olur.

1.1.2.12. Kısaltmaların yazımı

Bir sözcük veya özel ad anlaşılır olacak şekilde kısaltma olarak yazılabilmektedir. Kitap, dergi, kuruluş ve yön adlarının kısaltmaları her sözcüğün birinci harfi büyük olacak şekilde yazılması şeklinde yapılabilir. Örneğin; Türkiye Futbol Federasyonu kuruluşunun kısaltması TFF olarak yazılır. Kısaltma yapılırken harflerin arasına nokta konmaz. Ancak harflerin arasına nokta koyularak yapılan kısaltmalar da bulunmaktadır. Bu durum en çok bilinen örneği Türkiye Cumhuriyeti’nin T.C. şeklinde yapılan kısaltmasıdır. Bazı kısaltmalarda sözcüklerin birkaç harfi alınarak kısaltma yapılabilir. Örneğin; İleri Teknolojiler Araştırma Enstitüsü’nün kısaltması İLTAREN olarak yazılmaktadır. Bazı sözcüklerin kısaltması ilk harf ile birlikte sözcüğü

oluşturan temel harfler alınarak yapılmaktadır. Örneğin, üsteğmen sözcüğünün kısaltması ütğm. olarak yapılır. Bazı sözcüklerin kısaltması da sözcük gibi olabilmektedir. Örneğin; HAVELSAN kısaltmasının açılımı Hava Elektronik Sanayii olarak karşımıza çıkmaktadır. Elementler ve ölçü birimlerinin kısaltması standart olarak kabul edilmiştir. Örneğin; kobalt elementinin kısaltması co, milimetre biriminin kısaltması mm olarak uluslararası alanda kabul edilmektedir. Son olarak özel isimlerin kısaltmaları büyük harfle başlarken özel olmayan isimlerin kısaltmaları küçük harfle başlamaktadır.

1.1.2.13. Birleşik kelimelerin yazımı

Birleşik kelimeler yeni bir kavramı karşılamak üzere birden fazla kelimenin bir araya gelmesiyle oluşmaktadır. Birleşik kelimeler bitişik yazılabildiği gibi ayrı olarak da yazılabildiği durumlar bulunmaktadır. Birleşik kelimelerin bitişik yazıldığı durumlar aşağıda detaylı olarak açıklanmıştır (TDK, 2023d).

a. Birleşik kelimelerin oluşması esnasında ses düşmesi (ünlü düşmesi veya ünsüz düşmesi) varsa söz konusu birleşik kelimeler bitişik yazılır. Örneğin; “cumartesi” kelimesi “cuma” ve “ertesi” kelimelerinin birleşiminden oluşmakta ve ses düşmesi görüldüğünden bitişik yazılmaktadır.

b. Kökeni Arapça olan bazı tek heceli sözcükler ile birleşik kelime oluşturulurken ses düşmesi, ses değişmesi veya ses türemesi olayları görülmektedir. Genel olarak *etmek, edilmek, eylemek, olmak, olunmak* yardımcı fiilleriyle birleşirken bu ses olayları görülmekte olup bu durumlarda oluşturulan birleşik kelimeler bitişik yazılır. Örneğin; “reddetmek” kelimesi bu kurala göre oluşturulmuş olan birleşik kelimedir.

c. Birleşik kelimeyi oluşturan kelimelerden her ikisi veya sadece ikinci kelime benzetme yoluyla anlam değişmesine uğradığında oluşturulan yeni kelime bitişik yazılır. Örneğin; civanperçemi, kamçıkuyruk, itdirseği, kargaburnu, kırlangıçkuyruğu, tavukgöğsü, dokuztaş, samanyolu, vişneçürüğü, hinoğluhin vb. kelimelerde anlam değişmesi olduğu için bitişik yazılmaktadır.

d. Bazı zarf-fiil ekleri ve fiillerle oluşturulan birleşik kelimeler bitişik olarak yazılır. Örneğin; “uyuyakalmak” ve “düşünebilmek” kelimeleri bu kurala göre oluşturulmuş olan birleşik kelimedir.

e. Zamanla kalıplaşmış bazı birleşik kelimelerin bir veya iki ögesi emir kipiyle kurulmuş durumdadır. Bu biçimde oluşturulmuş olan kelimeler bitişik olarak yazılır. Örneğin; “unutmabeni” ve “incitmebeni” kelimeleri bu kurala göre oluşturulmuş olan birleşik kelimedir.

f. Bazı sıfat-fiil ekleri ile oluşturulan birleşik kelimeler bitişik olarak yazılır. Örneğin; “değerbilmez” ve “cankurtaran” kelimeleri bu kurala göre oluşturulmuş olan birleşik kelimedir.

g. Oluşturulan birleşik kelimeler, ikinci kelimesinde geçmiş zaman eki (-dı, -di vb.) alarak kurulmuşsa söz konusu birleşik kelimeler bitişik yazılır. Örneğin; “serdengeçti” ve “hünkârbeğendi” kelimeleri bu kurala göre oluşturulmuş olan birleşik kelimedir.

h. Oluşturulan birleşik kelimelerin, birinci ve ikinci kelimesinin her ikisi de geçmiş zaman eki (-tı, -ti vb.) veya geniş zaman eki (-r vb.) almışsa bahse konu birleşik kelimeler bitişik yazılır. Örneğin; “uyurgezer” ve “konargöçer” kelimeleri bu kurala göre oluşturulmuş olan birleşik kelimedir.

i. Oluşturulan birleşik kelimeler, ikinci kelimesinde alt, üst ve üzeri sözcüklerinden bir tanesi kullanılarak oluşturulmuşsa bu kurala uyan birleşik kelimeler bitişik olarak yazılır. Örneğin; “akşamüzeri” ve “olağanüstü” kelimeleri bu kurala göre oluşturulmuş olan birleşik kelimedir.

j. Birleşik kelime formatında oluşturulmuş kişi adları, soyadları ve lakapları bitişik olarak yazılır. Örneğin; “Atatürk” ve “Karaosmanoğlu” kelimeleri bu kurala göre oluşturulmuş olan birleşik kelimedir.

k. Birleşik kelime formatında oluşturulmuş yer adları (il, ilçe vb.) bitişik yazılır. Örneğin; “Gökçeada” ve “Yenişehir” kelimeleri bu kurala göre oluşturulmuş olan birleşik kelimedir.

l. Kişi adı ve unvanından (unvanın ikinci kelime olması durumunda) oluşan birleşik kelime formatında oluşturulmuş yer adları bitişik yazılır. Örneğin; “Bayrampaşa” ve “Gazi Osmanpaşa” kelimeleri bu kurala göre oluşturulmuş olan birleşik kelimedir.

m. Oluşturulan birleşik kelimenin ara yön belirten bir kelime olması durumunda bu kurala göre oluşturulmuş olan birleşik kelimeler bitişik yazılır. Örneğin; “güneybatı” ve “kuzeydoğu” kelimeleri bu kurala göre oluşturulmuş olan birleşik kelimedir.

n. Oluşturulan birleşik kelimenin her iki kelimesi de esas anlamını korumasına rağmen bitişik olarak yazılan ve kalıplaşmış kelimeler bulunmaktadır. Örneğin; "başöğretmen" ve "hanımefendi" kelimeleri bu kurala göre oluşturulmuş olan birleşik kelimedir.

o. Birleşik kelimenin oluşturulmasında "ev" kelimesi kullanıldığında oluşturulan kelime genelde bitişik yazılır. Örneğin; "öğretmenevi" ve "orduevi" kelimeleri bu kurala göre oluşturulmuş olan birleşik kelimedir.

p. Birleşik kelimenin oluşturulmasında "hane, name ve zade" kelimeleri kullanıldığında oluşturulan kelime genelde bitişik yazılır. Örneğin; "beyanname" ve "seyahatname" kelimeleri bu kurala göre oluşturulmuş olan birleşik kelimedir.

q. Birleşik kelimenin oluşturulmasında "-zede" kelimesi kullanıldığında oluşturulan kelime genelde bitişik yazılır. Örneğin; "kazazede" ve "depremzede" kelimeleri bu kurala göre oluşturulmuş olan birleşik kelimedir.

r. Birleşik kelime Farsça ve Arapça kurala göre oluşturulmuşsa bitişik yazılır. Örneğin; "gayrimeşru" ve "fevkalade" kelimeleri bu kurala göre oluşturulmuş olan birleşik kelimedir.

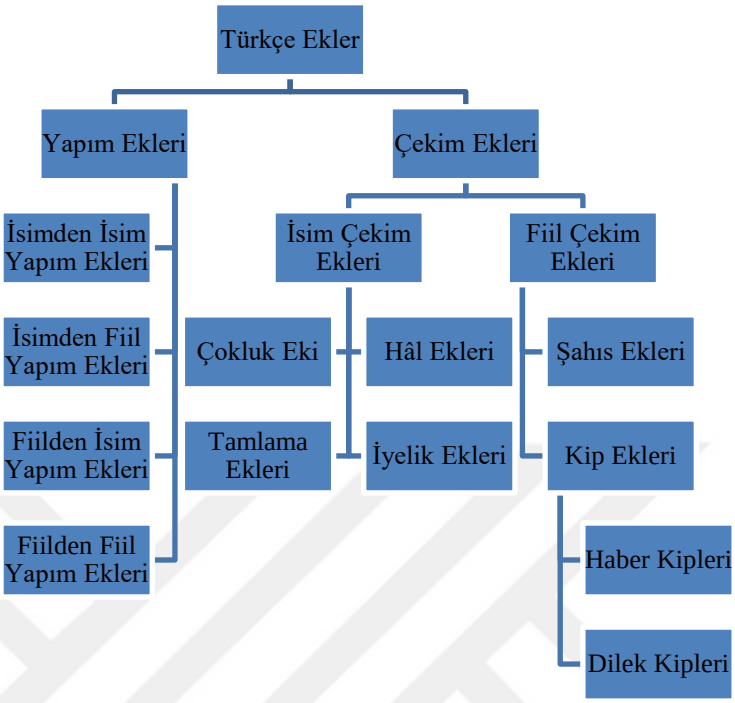
s. Oluşturulan birleşik kelime kanunda bitişik olarak yer alıyorsa ve bitişik tescil edilen kelimeler bitişik yazılır. Örneğin; "Genelkurmay" ve "Yükseköğretim" kelimeleri bu kurala göre oluşturulmuş olan birleşik kelimedir.

t. "Alacamenekşe" ve "sarıçiçek" gibi renk adlarıyla oluşturulan birleşik kelimeler bitişik yazılır.

1.1.2.14. Yapım ve çekim ekleri

Sözcükler ek alabilen yapıya sahiptir. Türkçede ekleri, Şekil 1.5'de görüleceği üzere yapım ekleri ve çekim ekleri olarak iki sınıfa ayırmak mümkündür. Yapım ekleri, eklendiği sözcüğün türünün, anlamının ve görevinin değişmesini sağlar. Örneğin; "sararmış" kelimesinin "sarı" köküne getirilen yapım ekinden türetildiği görülmektedir. Kelimenin kökünün, aldığı yapım eki ile yeni anlam kazandığı söylenebilir. Çekim ekleri, eklendiği sözcüğün türünü ve anlamını değiştirmezler ancak sözcüklerin cümle içindeki kullanımlarını sağlayan eklerdir. Örneğin; "öğrenciyi" sözcüğü aldığı çekim eki ile anlam değişikliğine uğramamış ancak cümle içerisinde kullanımı belirtilmiştir. Yapım ekleri kendi arasında isimden isim yapım eki, isimden fiil yapım eki, fiilden fiil yapım eki ve fiilden isim yapım eki olmak üzere dörde ayrılır. Çekim ekleri ise kendi

arasında isme gelen çekim ekleri (hâl ekleri, çoğul eki, tamlama ekleri, iyelik ekleri) ve fiile gelen çekim ekleri (kip ekleri, şahıs ekleri) olmak üzere ikiye ayrılır.



Şekil 1.5. Yapım ve çekim ekleri

Sözcüğün görevini, türünü ve anlamını değiştiren ekler yapım ekleri olarak tanımlanır. Yapım eki alan sözcükler türetilmiştir demek mümkündür. Bir sözcük birden fazla yapım eki alabilir. Örneğin; "bil-gi-len-dir" sözcüğü üç adet yapım eki almıştır.

a. İsimden isim yapım eki; isim köküne gelerek yeni bir isim oluşturulmasını sağlayan eklerdir. Örneğin; "su" kelimesi isimden isim yapım eki alarak yeni bir isim olan "suluk" kelimesinin türemesine yol açar.

b. İsimden fiil yapım eki; isim köküne gelerek yeni bir fiil oluşturulmasını sağlayan eklerdir. Örneğin; "sarı" kelimesi isimden fiil yapım eki alarak yeni bir fiil olan "sararmak" kelimesinin türemesine yol açar.

c. Fiilden fiil yapım eki; fiil köküne gelerek yeni bir fiil oluşturulmasını sağlayan eklerdir. Örneğin; "çalış-" kelimesi fiilden fiil yapım eki alarak yeni bir fiil olan "çalış-tır" kelimesinin türemesine yol açar.

d. Fiilden isim yapım eki; fiil köküne gelerek yeni bir isim oluşturulmasını sağlayan eklerdir. Örneğin; "sil-" kelimesi fiilden isim yapım eki alarak yeni bir fiil olan "sil-gi" kelimesinin türemesine yol açar.

Çekim ekleri tür ve anlam değişikliğine yol açmazlar. Yani sözcüğün görevini, türünü ve anlamını değiştirmeyen ekler çekim ekleri olarak tanımlanabilir. Bir sözcük birden fazla çekim eki alabilir. Örneğin; "hayvan-lar-ımız-da" sözcüğü çoğul, iyelik ve hal eki olmak üzere üç adet çekim eki almıştır. Çekim ekleri iki gruba ayrılmaktadır. İsim köküne gelen çekim ekleri isim çekim ekleri olarak tanımlanmaktadır ve kendi içerisinde dört gruba ayrılmaktadır.

a. Çokluk eki (-ler, -lar), eklendiği ismin sayıca birden fazla olduğunu belirten eklerdir. Örneğin; "Çıkardığı özet **sayfalarca** tutmuştu." Cümlesinde çıkarılan özetin birden fazla sayfadan oluştuğu isme getirilen çokluk eki ile belirtilmiştir. Net bir sayı belirtilmemektedir.

b. Hâl ekleri (-i, -e, -de, -den), eklendiği isme bulunma, belirtme, ayrılma ve yönelme anlamı kazandıran eklerdir. Örneğin; "Arabay**ı** dün akşam boyadım.", "Hafta içi tatil**e** çıktım.", "Çantamı okul**da** bıraktım." ve "Annesin**den** para istedi." cümlelerinde hâl eki kullanımı görülmektedir.

c. İlgi ekleri (-ın, -in, -un, -ün), eklendikleri isimlerin diğer isimlerle bağ kurmasını sağlayan eklerdir. Örneğin; "Ablam ben**im** elbisemi giymiş." ve "Ceket**in** düğmesi kopmuş." cümlelerinde ilgi eki kullanımı görülmektedir.

d. İyelik ekleri (-m, -n, -i, -miz, -niz, -leri), şahıslara göre çekimlenen bu ekler eklendikleri isme, kime ait olduğunu gösteren anlam kazandırmaktadır. Örneğin; "Okul**umuz**un konferans salonu oldukça büyüktür." ve "Telefon**um**un kılıfı siyah renklidir." Cümlelerinde iyelik eklerinin kullanımı görülmektedir. İyelik ekleri zaman zaman hâl ekleri ile karıştırılabilmektedir. Bu duruma çok dikkat edilmesi gerekmektedir.

Fiil köküne gelen çekim ekleri fil çekim ekleri olarak tanımlanmaktadır ve kendi arasında kip ekleri ve şahıs ekleri olmak üzere iki gruba ayrılmaktadır.

Fiilde belirtilen işin ne zaman ve hangi duyguyla yapıldığını belirten şekiller kip olarak adlandırılır.

a. Haber kipleri zaman anlamı taşımaktadır. Eylemin şimdiki zamanda, geçmiş zamanda, gelecek zamanda veya geniş zamanda yapıldığını gösterir. Örneğin; "Okula gidi**yor**um." cümlesinde eylemin içinde bulunduğumuz şimdiki zamanda yapıldığı görülmektedir.

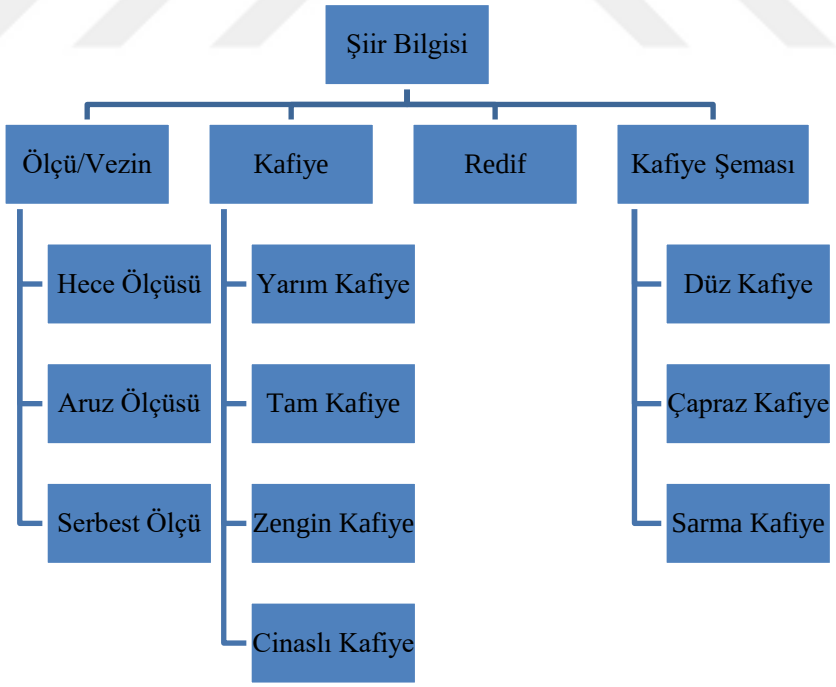
b. Dilek kipleri zaman anlamı ifade etmez. Eylemin emir duygusu, istek duygusu, gereklilik duygusu veya şart duygusu içerisinde yapılıp yapılmadığını gösterir.

Örneğin; “Derlerine çalışır**sa**n başarılı olursun.” cümlesinde eylemin şart içerisinde gerçekleşeceği görülmektedir.

Fiilde belirtilen işin kim tarafından yapıldığını belirten ekler şahıs ekleri olarak adlandırılır. Örneğin; “Bugün maraton koşusuna katıldım.” cümlesinde eylemin birinci tekil şahıs tarafından gerçekleştirildiği görülmektedir.

1.1.3. Türkçe şiirlerin incelenmesi

Şiir, Arapça şiʿr sözcüğünden gelmektedir. İnsanların duygu ve düşüncelerini ifade edebildiği, kendi özgü kuralları olan, sözcüklere şiir içerisinde kullanıldığında farklı anlamlar kazandırabilen, sözcüklerin uyum içerisinde olduğu, çeşitli söz sanatlarından faydalanılan ve insanlarda estetik duygular oluşturan bir edebiyat türüdür. Bir şair olan Birhan Keskin Kargo isimli şiirinde şiiri şöyle tanımlamıştır: “Hem zaten şiir niye var? Dünyanın acısını başkaları da duysun! Acı mıhlanıp bir kalpte durmasın. Ortada dursun. Olur ya biri eline alır okşar, biri alnından öper. Az unutursun.” Şiir yazan kişiler ise şair olarak tanımlanmaktadır.



Şekil 1.6. Şiir bilgisi (Anonim, 2023a)

Şiirin kendine ait yapı unsurları bulunmaktadır. Nazım birimi, nazım şekli/biçimi, kafiye, redif vb. unsurlar yapı unsurlarını oluşturmaktadır. Bir şiirin nazım

birimi dize/mısra, beyit veya bent olabilir. Tek bir satırdan oluşan şiir bölümüne dize veya mısra adı verilmektedir. Bir şiir iki adet dizeden oluşuyorsa yani iki satırdan meydana geliyorsa nazım birimi beyit olarak tanımlanır. Bir şiir üç ila on mısradan oluşuyorsa bu tür şiirlerin nazım birimi bent olarak tanımlanır. Burada ilave bilgi olarak dört dizeden oluşan şiirlerin nazım birimi dörtlük veya kıta olarak da karşımıza çıkabilmektedir. Şiirler nazım şekli/biçimi açısından ele alındığında şiirin dış yapısı (kafiye, kafiye düzeni, vezin vs.) incelenmelidir. İslamiyet öncesi Türk edebiyatında (koşuk, sagu, destan), halk edebiyatında (mani, ninni, tekerleme vb.), divan edebiyatında (gazel, mesnevi, kaside vb.) çeşitli nazım şekli/biçimi bulunmaktadır. Nazım türü ile nazım şekli/biçimi arasındaki fark kısaca özetlenecek olursa, nazım biçimi/şekli şiirin dış yapısıyla ilgili iken nazım türü şiirin içyapısıyla yani konusuyla ilgilidir.

Şiirde ahenk ve dizelerin uyumuna ölçü denir. Bu noktada karşımıza üç adet ölçü çıkmaktadır.

a. Hece ölçüsü; şiirin her bir satırının aynı hece sayısına sahip olmasıdır. Hece ölçüsü Türkçeye özgü bir ölçüdür. Türkçede yaygın olarak yedili, sekizli ve on birli hece ölçüsü kullanılmaktadır. Hece ölçüsü bulunan şiirde dizeler bölümlere ayrılabilir. Bu durum durak olarak tanımlanmaktadır. Yedili hece ölçüsüne sahip şiirlerde “4+3”, sekizli hece ölçüsüne sahip şiirlerde “4+4”, on birli hece ölçüsüne sahip şiirlerde “6+5” veya “4+4+3” durak olabilir. Şiirde durak olması zorunlu değildir. Yani durak olmayabilir.

b. Aruz ölçüsü; Arapçadan Farsçaya, Farsçadan da Türkçeye geçmiş bir şiir ölçüsüdür. Türk edebiyatına, divan edebiyatı döneminde girmiştir. Yani aruz ölçüsü divan edebiyatı döneminde sık kullanılmış bir şiir ölçüsüdür. Şiirde açık (kısa) veya kapalı (uzun) hece kullanımına göre ölçü sağlanır. Bir hecenin sonu kısa ünlü ile bitiyorsa kısa hece, bir hecenin sonu uzun ünlü (â, î, û) veya ünsüzle bitiyorsa kapalı hece olarak tanımlanır. Açık heceler noktayla gösterilirken kapalı heceler kısa çizgi ile gösterilir. Aruz ölçüsünde ulama varsa vasl, vezin gereği açık hecenin kapalı olarak okunması imâle, kapalı bir heceyi açık bir heceye çevirme zihaf ve uzun hecenin biraz daha uzun okunmasına ise medd denir.

c. Serbest ölçü; şiirde hece ölçüsünün aranmadığı ölçüdür. Genellikle cumhuriyet dönemi eserlerinde görülmektedir.

Anlam ve görev bakımından aynı seslerin benzeşmesine redif denir. Anlam ve görev bakımından farklı seslerin benzeşmesine ise kafiye (uyak) denir. Şiirde, önce var

ise redif bulunur ardından kafiye uyumuna bakılır. Şiirde kafiye dört başlıkta incelenmektedir.

a. Yarım kafiye; eğer bir ses benzerliği bulunuyorsa bu benzerliğe yarım kafiye denir.

b. Tam kafiye; eğer iki ses benzerliği bulunuyorsa bu benzerliğe tam kafiye denir.

c. Zengin kafiye; eğer üç ve daha fazla ses benzerliği bulunuyorsa bu benzerliğe zengin kafiye denir.

d. Cinaslı kafiye; dize sonunda yazılışları aynı fakat anlamları farklı sözcükler bulunuyorsa bu benzerliğe cinaslı kafiye denir.

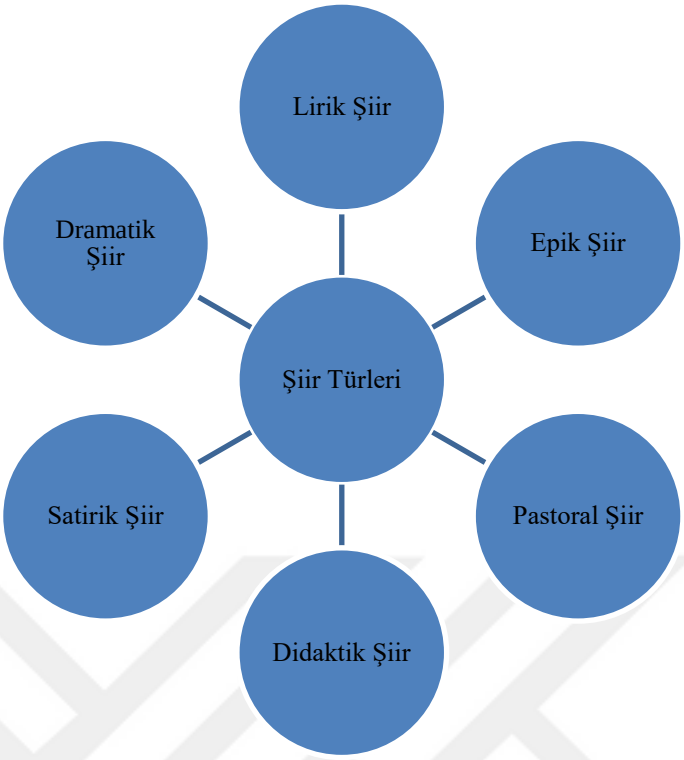
Dizeler anlam ve görev bakımından incelenip kafiye tespiti yapıldıktan sonra bu dizeler harfler kullanılarak gruplandırılmaktadır. Yaygın olarak "a" harfinden başlamak suretiyle gruplandırma yapılmaktadır. Bu işleme kafiye düzeni ya da uyak örgüsü denir. Şiirde uyumun sağlanması açısından önemli unsurlardan bir tanesidir. Düz uyak örgüsü, çapraz uyak örgüsü ve sarma (sarmal) uyak örgüsü olmak üzere üç çeşit kafiye örgüsü bulunmaktadır.

a. Düz uyak; bir kıtayı oluşturan mısraların "a, a, a, a" veya "a, a, a, b" şeklinde uyak örgüsüne sahip olması durumudur. Mısraların tamamı uyumlu olduğunda veya üç mısra uyumlu olduğunda düz uyak vardır denir.

b. Çapraz uyak; bir kıtayı oluşturan mısraların "a, b, a, b" şeklinde uyak örgüsüne sahip olması durumudur. Yani birinci mısra ile üçüncü mısra arasında, ikinci mısra ile dördüncü mısra arasında çapraz bir şekilde ses uyumu olmasına çapraz uyak denir.

c. Sarmal uyak; bir kıtayı oluşturan mısraların "a, b, b, a" şeklinde uyak örgüsüne sahip olması durumudur. Yani birinci mısra ile dördüncü mısra arasında, ikinci mısra ile üçüncü mısra arasında sarmal bir şekilde ses uyumu olmasına çapraz uyak denir.

Her şiirin kendine ait bir konusu vardır. İçerdikleri konu çerçevesinde bazı şiirler doğa ile ilgili temaya sahipken bazı şiirler bayrak temasına sahip olabilmektedir. Bu açıdan şiirler düz metin gibi görünse de duygu yüklü edebi eserlerdir. Duyguların ifade edildiği şiirler birden fazla duyguyu içerebildiğinden okuyucuda birden fazla duygunun oluşmasına yol açarlar. Şiirleri Şekil 1.7'de gösterildiği gibi, sahip oldukları temaya göre lirik şiir, epik şiir, pastoral şiir, didaktik şiir, dramatik şiir ve satirik şiir olmak üzere altı şiir türüne göre sınıflandırmak mümkündür.



Şekil 1.7. Şiir türleri

Lirik şiir; şairin aşk, tabiat, özlem, gurbet, vatan, din, ölüm vb. içten gelen duygularını dokunaklı bir biçimde anlattığı şiir türüdür. Lirik türde şiirler dünyada ilk kez Yunanlar tarafından yazılmıştır. Divan edebiyatında gazel ve şarkı, halk edebiyatında ise güzelleme şiirleri bu türe örnek olarak gösterilebilir. Fuzuli, Âşık Veysel, Karacaoğlan, Yahya Kemal Beyatlı, Cahit Sıtkı Tarancı bu türde şiir yazan şairlerdir. Karacaoğlan tarafından kaleme alınan Şekil 1.8'de bulunan kıta lirik şiir türüne örnek olarak gösterilebilir.

Kömür gözlüm ne salının karşımda,
Gündüz hayalimde, gece düşümde.
Bir güzelin sevdası var başımda,
Yar sevdası çetin olur yaradan.

Şekil 1.8. Lirik şiir örneği (Karacaoğlan)

Epik şiir; şairin savaş, kahramanlık, vatan sevgisi, yiğitlik gibi konuları coşkulu bir biçimde anlattığı şiir türüdür. Halk edebiyatında koçaklama, destan ve varsağı bu

türe örnek olarak gösterilebilir. Şair Mehmet Akif Ersoy tarafından Çanakkale şehitlerine ithafen yazılan ve Şekil 1.9'da bir bölümü yer alan şiir epik şiir türüne örnek olarak gösterilebilir.

Şu Boğaz Harbi nedir? Var mı dünyâda eşi?
En kesîf orduların yükleniyor dördü beşi,
-Tepeden yol bularak geçmek için Marmara'ya-
Kaç donanmayla sarılmış ufacık bir karaya.
Ne hayâsızca tehaşşüd ki ufuklar kapalı!
Nerde -gösterdiği vahşetle "Bu: Bir Avrupalı!"
Dedirir- yırtıcı, his yoksulu, sırtlan kümesi,
Varsa gelmiş, açılıp mahbesi, yâhud kafesi!

Şekil 1.9. Epik şiir örneği (Mehmet Akif Ersoy)

Didaktik şiir; şairin bilgi vermeyi amaçladığı öğretici şiir türüdür. Bu açıdan duygunun ikinci planda olduğu düşüncenin ön planda olduğu şiirlerdir. Okuyucuya belirli bir konuda nasihat vermek veya ders çıkarmak amacıyla yazılır. Düşünceler ve nasihatler daha kalıcı olması açısından şiir olarak aktarılmaktadır. Didaktik şiir türüne örnek olarak divan edebiyatı şairi Nâbi'nin oğlu Ebulhayr Mehmed Çelebî'ye öğüt vermek için kaleme aldığı Hayriyye ve Servet-i Fünûn topluluğundan şair Tevfik Fikret'in oğlu Haluk'a öğüt vermek için kaleme aldığı Halûk'un Defteri didaktik şiir türüne örnek olarak gösterilebilir. Bir diğer örnek, Ziya Gökalp tarafından kaleme alınan Lisan şiirinden bir bölüm Şekil 1.10'da verilmiştir.

Güzel dil Türkçe bize,
Başka dil gece bize.
İstanbul konuşması
En saf, en ince bize.

Şekil 1.10. Didaktik şiir örneği (Ziya Gökalp)

Satirik şiir; bireylerin veya toplumun hoş bulunmayan taraflarının iğneleyici bir dil kullanılarak açıktan veya kapalı bir biçimde eleştirildiği şiir türüdür. Şair tarafından yergi amaçlanmaktadır. Beğenilmeyen özelliklerin zaman zaman gülünç bir tavırla ele alındığı görülmektedir. Bu tür şiirlere halk edebiyatında taşlama ve divan edebiyatında hiciv denilmektedir. Bu şiirlerde şairler ya bireysel ya da toplumsal eleştiride bulunmuştur. Özellikle konum itibariyle önemli görevde bulunan kişiler bu şiirlere konu olmuştur. Şairler Nef'î, Bağdatlı Ruhi ve Şeyhi satirik şiir türünde eserler kaleme

almıştır. Şekil 1.11'de, Tevfik Fikret tarafından kaleme alınan ve bir bölümü yer alan Han-ı Yağma şiiri satirik şiir türüne örnek olarak gösterilebilir.

Bu sofracık, efendiler - ki iltikaama muntazır
Huzurunuzda titriyor - bu milletin hayatıdır;
Bu milletin ki mustarip, bu milletin ki muhtazır!
Fakat sakın çekinmeyin, yiyin, yutun hapır hapır...

Şekil 1.11. Satirik şiir örneği (Tevfik Fikret)

Pastoral şiir; şair tarafından tabiat ve doğal güzelliklerin, köy yaşamının ve çobanların konu alınarak yazıldığı şiir türüdür. Pastoral şiirlerde genel olarak sade bir dil tercih edildiği görülmektedir. Şairler şehir hayatının sıkıcı tarafından uzaklaşmak için bu şiire yönelmiştir. Abdülhak Hamit Tarhan tarafından kaleme alınan Sahra şiiri Türk edebiyatında ilk pastoral şiir olarak kabul edilmektedir. Şair eserinde kır yaşamını dile getirmeyi amaçlamıştır. Pastoral şiirin idil ve eglog olmak üzere iki türü bulunmaktadır. İdil, şairin doğa güzelliklerini kır yaşamını, çoban yaşamını anlatmasıdır. Eglog, iki çobanın şiir aracılıyla konuşmasıdır. Şekil 1.12'de Kemalettin Kamu tarafından kaleme alınan ve bir bölümü yer alan Bingöl Çobanları şiiri pastoral şiir türüne örnek olarak gösterilebilir.

Daha deniz görmemiş bir çoban çocuğuyum.
Bu dağların en eski âşinasıdır soyum,
Bekçileri gibiyiz ebenced buraların.
Bu tenha derelerin, bu vahşi kayaların
Görmediği gün yoktur sürü peşinde bizi,
Her gün aynı pınardan doldurur destimizi
Kırlara açılırız çıngıraklarımızla...

Şekil 1.12. Pastoral şiir örneği (Kemalettin Kamu)

Dramatik şiir; insanların kolay ifade edemediği duygularını ifade edebildiği şiirlere dramatik şiir denir. Acıklı ve korkunç olaylara yer verildiği görülmekle birlikte tiyatroda kullanılmaktadır. Bu şiirler, dramatik bir konuya odaklanarak karakterler, diyaloglar ve olay örgüsü gibi tiyatroya ait unsurları içerir. Dramatik şiir, şiirin duygusal yoğunluğunu, ritmik dilini ve imgesel gücünü kullanarak okuyucuya bir olayı veya hikâyeyi canlı bir şekilde anlatmayı amaçlar. Bu tür şiirler genellikle trajik, acıklı veya duygusal konuları ele alır ve okuyucunun duygusal tepkilerini harekete geçirmeyi amaçlar. Dramatik şiirler lirik şiirler ile benzer özellikler göstermektedir. Cahit Sıtkı

Tarancı ve Nazım Hikmet bu türde eserler vermiş şairlere örnek olarak verilebilir. Şekil 1.13'de Melih Cevdet Anday tarafından kaleme alınan ve bir bölümü yer alan Gelinlik Kızın Ölümü şiiri dramatik şiir türüne örnek olarak gösterilebilir.

Salâ verilirken kalktık kahveden,
Kızın babası yanımızda, boyu uzun,
Zayıf, ağzı mırıltılar.
On köylü, iki subay, bir tezkereci er,
Sıralandık ahşap mescidin avlusunda,
Aldık cenazeyi sarsmandan, iğreti
Ve hafif, gözlerimiz yerde,
Kayıp bir tayın izini süreriz sanki...

Şekil 1.13. Dramatik şiir örneği (Melih Cevdet Anday)

1.2. Doğal Dil İşleme

Doğal dil işleme, insanlar tarafından kullanılan dillerin (doğal dil) bilgisayarlar tarafından işlenebilir bir forma dönüştürmek için bilgisayar bilimleri ve dilbilim yöntemlerinin birlikte kullanıldığı bir bilim alanıdır. Doğal dil işlemenin amacı, doğal dilde üretilmiş olan metinlerin bilgisayarlar tarafından anlaşılmasını, yorum ve cevap üretilmesini sağlamaktır.

Doğal dil işlemenin amacına ulaşabilmesi için bir dizi işlemin gerçekleştirilmesi gerekmektedir. İlk adım, verilerin toplanması ve doğal dilde üretilmiş olan metinlerin elektronik formatta bilgisayara aktarılmasıdır. Daha sonra, bu metinlerin ne anlam ifade ettiğini anlamak üzere dilbilim yöntemleri, matematiksel ve istatistiksel modeller kullanılır. Bu teknikler metindeki kelimelerin anlamlarını, yapılarını, bağlamlarını ve anlamsal ilişkilerini belirlemeye yardımcı olur.

Doğal dil işleme, insanlarla bilgisayarların etkileşimini artırmak ve özellikle büyük veri setleri üzerinde çalışan kurumlar ve araştırmacılar için doğal dil verilerini analiz etmeyi ve anlamayı kolaylaştırmak amacıyla kendine birçok uygulama alanı bulmuştur. Örnek olarak; aramak istediğimiz bilgileri anlayarak buna karşılık en iyi sonuçları getirmeyi sağlayan arama motorları, gerçekleştirmek istediğimiz işlemi konuştuğumuz ve yazdığımız doğal dili anlayarak yerine getiren akıllı sanal asistanlar günlük hayatta en sık kullandığımız doğal dil işleme uygulamalarındandır. İnsanlar tarafından kullanılan farklı diller arasında çeviri yapan otomatik çeviri sistemleri, uzun bir formatta olan metinlerin daha kısa bir biçimde özetlenmesini sağlayan otomatik

metin özetleme, kullanıcılara uygun reklamların gösterilebilmesini sağlayan sosyal medya takibi ve daha birçok uygulama sayılabilir.

1.2.1. Doğal dil işlemenin tarihçesi

Makine çevirisi dönemi olarak adlandırılabilecek yıllarda (1940-1960) bilim insanları çalışmalarında ağırlıklı olarak makine çevirisine odaklanmıştır. Dr. Booth'un, Dr. R. H. Richens'le birlikte yapmış oldukları çalışma ve Warren Weaver tarafından 1949 yılında ortaya atılan istatistiksel makine çevirisi yaklaşımına ilişkin ilk düşünceler doğal dil işleme üzerine yapılan araştırmaların temelini oluşturmuştur. Ardından 1954'te IBM 701 kullanılarak yapılan Georgetown-IBM deneyinde Rusçadan İngilizceye çeviri yapılabilmesi sağlanmıştır. Takip eden yıllarda uluslararası konferanslar yapılmaya başlanmıştır.

Verinin veya bilgi tabanının ele alındığı yıllarda (1960-1970) yapılan çalışmalar yapay zekâdan etkilenmiştir. Bu yıllarda BASEBALL soru cevap sistemi geliştirilmiştir. Genel manada; veri tabanında saklanan verilerle sıradan İngilizcede ifade edilen soruları yanıtlayan bir bilgisayar programıdır. Daha sonra yapılan çalışmalarda daha gelişmiş sistemler ortaya koyulmuş ve bu sistemlerde dil girdisini (ses veya metin) yorumlama ve yanıtlama yöntemiyle veri tabanından bilgi çıkarımı yapma ihtiyacı üzerinde durulmuştur.

Daha sonra yapılan çalışmalarda (1970-1980), pratik sistem çalışmalarında istenen başarı elde edilemediğinden yapay zekâ ile bilgi temsili ve akıl yürütme yapabilmek için mantık kullanımına yönelim olmuştur. Bu yıllarda yapılan sözlük çalışmaları dilbilimsel-mantıksal yaklaşıma işaret olarak gösterilebilir.

Bu aşamaya kadar yapılan çalışmaları takip eden yıllarda doğal dil işleme için makine öğrenmesi algoritmalarının kullanılmaya başlanmasıyla doğal dil işleme çalışmaları ivme kazanmıştır.

Doğal dil, insanlar tarafından sürekli kullanıldığından ve gelişime açık olduğundan dinamik bir doğaya sahiptir. Zamanla değişime uğradığı ve farklılaşmalar olduğu gözlemlenmektedir. Kurallı bir yapıya sahip olan doğal dilin incelenmesi çeşitli disiplinler yardımıyla gerçekleştirilir. Her disiplinin kendine özgü soruları/sorunları ve bunları çözmek için ortaya koydukları çeşitli yöntem ve metotları vardır. Aşağıdaki çizelge 1.6.'da bahse konu disiplinler, ilgilendikleri problemler ve bunlara yönelik çözüm yolları anlatılmıştır.

Çizelge 1.6. Disiplinler ve ilgilendikleri alanlar (Tutorialspoint, 2023)

Disiplin	Problem	Yöntem
Dilbilimciler	Sözcükler kullanılarak cümle nasıl kurulabilir? Bir cümlenin anlamını bozabilecek şeyler nelerdir?	Biçimlilik ve anlamla ilgili sezgiler. Matematiksel yapı modeli. Örneğin, model teorik anlambilim, biçimsel dil teorisi.
Psikodilbilimciler	İnsanlar tarafından cümlelerin yapısı nasıl tanımlanabilir? Kelimelerin ifade ettiği anlam nasıl tespit edilebilir? Anlama eylemi ne zaman gerçekleşir?	Deneysel teknikler esas alınarak insan performansının ölçülmesi. Elde edilen gözlemlerin istatistiksel analizi.
Filozoflar	Sözcükler ve cümleler nasıl anlam kazanır? Nesneler kelimelerle nasıl tanımlanır? Anlam nedir?	Sezginin kullanılması ile doğal dil tartışması. Mantık ve model teorisi gibi matematiksel modeller.
Hesaplamalı Dilbilimciler	Bir cümlenin yapısı nasıl tanımlanabilir? Bilgi ve akıl yürütme nasıl modellenebilir? Belirli görevleri gerçekleştirmek için dili nasıl kullanabiliriz?	Algoritmalar Veri yapıları Biçimsel temsil ve akıl yürütme modelleri. Arama ve temsil yöntemleri gibi yapay zekâ teknikleri.

Dipnot: Psikodilbilim veya dilin psikolojisi psikolojik süreçlerle dilsel etkenler arasındaki iletişimi çalışan disiplindir.

1.2.2. Doğal dil işlemenin uygulama alanları

Teknolojinin gelişimine paralel olarak doğal dil işleme ile geliştirilmiş birçok sistem günlük yaşantımızda sıkça kullandığımız birçok teknoloji sayesinde hayatımızın bir parçası haline gelmiştir. İlerleyen başlıklarda en çok bilinen uygulamamlar açıklanmıştır.

1.2.2.1. Otomatik arama düzeltme ve tamamlama

Milyonlarca internet kullanıcısı istedikleri bilgilere anında erişebilmek için arama motorlarını kullanmaktadır. Arama motorlarını geliştiren yazılımcılar en iyi sonucu sunulabilmek için birçok yöntem kullanmaktadır. Bu yöntemlerden biri de doğal dil işleme uygulamalarından biri olan otomatik arama düzeltme ve tamamlama sistemidir. Arama motoruna bir şeyler yazılmaya başlandığı anda en alakalı başlıklar kullanıcılara sunulmaktadır. Arama işlemi yapıldığında herhangi bir yazım hatası varsa, bunlar düzeltilir ve yine kullanıcılara en alakalı sonuçlar arama motorları tarafından sunulur.

1.2.2.2. Otomatik çeviri

İnsanlar tarafından kullanılan doğal dillerin sayısı oldukça fazladır. Bu durum, farklı dillerde üretilmiş birçok metin ortaya çıkarmıştır. Bir insanın bildiği dil sayısı oldukça kısıtlıdır. Tüm bu durumlar göz önüne alındığında diller arasında tercüme ihtiyacı ortaya çıkmıştır. Bu duruma karşılık en yaygın ve en etkin kullanılan yöntem bir makine tarafından bir metnin bir dilden başka bir dile tercüme edilmesidir. Makine çevirisi alanında yapılan çalışmalar 1950'li yıllara dayanmaktadır. Yapılan ilk çalışmalarda sözlük ve kural tabanlı sistemler kullanılmıştır. Ancak bu sistemin başarısının oldukça düşük olduğu söylenebilir. Yapay sinir ağlarının gelişimi ve büyük verilerin işlenebilir hale gelmesiyle birlikte makine çevirisi oldukça başarılı bir hale gelmiştir. Şekil 1.14.'de makine çevirisi adımları sırasıyla verilmiştir.

Şekil 1.14. Makine çevirisi işlem adımları

1.2.2.3. Duygu analizi

İnsanlar birçok konuda duygu ve düşüncelerini internet tabanlı platformlarda paylaşmaktadır. Bu paylaşımlar herhangi bir konuda duyulan memnuniyet veya memnuniyetsizlik, siyasi görüş gibi konular içerebilmektedir. Artan rekabet ortamında, firma/kurum/kuruluşlar müşteri memnuniyetini artırmak, kendi hakkında yapılan paylaşımları analiz etmek veya ürünleri hakkında istatiksel bilgiler ortaya koyabilmek için bu verilerin analizi üzerinde çalışmaktadır. Bu şekilde oluşturulan verilerin boyutu oldukça büyüktür. Tam da bu noktada doğal dil işleme, bu konuda çalışma yapan araştırmacılara çeşitli yöntemler sunmaktadır.

1.2.2.4. Sohbet robotu

Sohbet robotları, kullanıcıları dijital ortamda ulaşmak istedikleri alana daha hızlı ve kolay bir şekilde yönlendiren bilgisayar yazılımlarıdır. Bir konuda işlem yapma ve

bilgi alma gibi çeşitli amaçlarla kullanılan sohbet robotları basit sorular sorarak bu işlemleri yerine getirir. Bir sohbet robotu kullanarak bankacılık işlemi yapmak, bilet satın almak, uçuş için kontrol işlemlerini yapmak, müşteri hizmetleri işlemlerinin yapılabilmesi mümkün hale gelmektedir. Sohbet robotları işlemlerin daha hızlı ve sorunsuz yapılabilmesine yardımcı olmaktadır.

1.2.2.5. Anket analizi

Anket, araştırmacının belirli bir konuda bilgi toplamak üzere hazırlamış olduğu yazılı soruları kişilere çeşitli yöntemlerle (telefon, internet, yüz yüze vb.) yöneltmek suretiyle yapmış olduğu veri toplama yöntemidir. Yapılan anketler ne kadar çok kişi ile yapılırsa elde edilen verinin boyutu aynı oranda artar. Bu verilerin tek tek analiz edilmesi oldukça zor bir hal almaktadır. Bu işlemin bir makine tarafından yapılması ihtiyacı ortaya çıkmıştır. Tam da bu noktada, verilerin analiz edilmesi ve bu verilerden anlamlı sonuçlar çıkarılabilmesi için doğal dil işleme kullanılmaktadır. Bu yöntem sağladığı yararlar (zaman, iş gücü vb.) nedeniyle sıkça kullanılmaktadır.

1.2.2.6. Metin özetleme

Bir metin içinden ana düşünceyi ve önemli bilgileri içerecek şekilde metni kısaltma işlemi metin özetleme olarak tanımlanmaktadır. Özetlenecek metnin onlarca hatta yüzlerce sayfadan oluşması durumunda özetleme işlemi oldukça zor bir işlem haline gelmektedir. Bu işlemin daha kolay bir şekilde makineler tarafından yapılabilmesi için çeşitli doğal dil işleme teknikleri kullanılmaktadır. Bu sayede blog siteleri, haberler vb. metinler birkaç satır kod kullanılarak özetlenebilmektedir.

1.2.2.7. İşe alım süreci

Şirketler yürütmüş oldukları faaliyetlerde organizasyon yapısına uygun bir şekilde insan kaynaklarını oluşturmaktadır. İşe alım sürecinde ilgili pozisyona en uygun personelin alınması firmalar için önem arz etmektedir. Alım yapılacak bir pozisyon için yüzlerce bazen binlerce başvuru olabilmektedir. Başvuran adaylar arasından en uygun kişiyi seçmek zor bir süreçtir. Doğal dil işleme yardımı ile insan kaynakları doğru adayı çok daha kısa sürede bulabilmektedir.

1.2.2.8. Sesli asistanlar

Doğal dil işlemenin en yaygın kullanılan örneği olarak görülebilir. Kullanıcılar tarafından yöneltilen soruları cevaplama (ör. Bugün hava kaç derece?), verilen komutları yerine getirme (ör. Wi-Fi aç) gibi işlemleri yerine getiren sesli asistan yazılımı doğal dil işleme teknikleri kullanılarak geliştirilmiştir. Bu yazılımların en bilinen örneği Siri yazılımıdır.

1.2.3. Doğal dil işleme sürecinde karşılaşılan zorluklar

Bu bölümde doğal dil işleme sürecinde karşılaşılan çeşitli zorluklara değinilmiştir.

1.2.3.1. Dilde belirsizlik

Doğal dil işlemede belirsizlik, bir sözcüğün veya cümlenin birden fazla anlama gelmesi durumudur.

a. Bir kelime tek başına bir anlam ifade edebilirken aldığı ek sayesinde bambaşka bir anlam kazanabilmektedir. Türkçede birçok kelime bu şekilde yeni anlam kazanabilir. Bu durum kelime kökünün bulunmasında araştırmacı açısından bir zorluk içermektedir. Bu belirsizlik doğal dil işlemede sözcüksel (Lexical) belirsizlik olarak tanımlanmaktadır.

b. Sözdizimsel belirsizlik (Syntactic), bir cümlenin farklı anlamlara gelebilmesi durumudur. Belirsizliğin sebebi cümleyi oluşturan kelimelerin yer değiştirmesi olabileceği gibi kelimeye yapılan vurgu da olabilir. Bu belirsizlik bilgisayarın anlaması açısından zorluk içermektedir.

c. Anlamsal belirsizlik (Semantic), cümlede asıl anlatılmak istenene alternatif olarak başka bir anlamın daha cümleden çıkarılabiliyor olmasıdır.

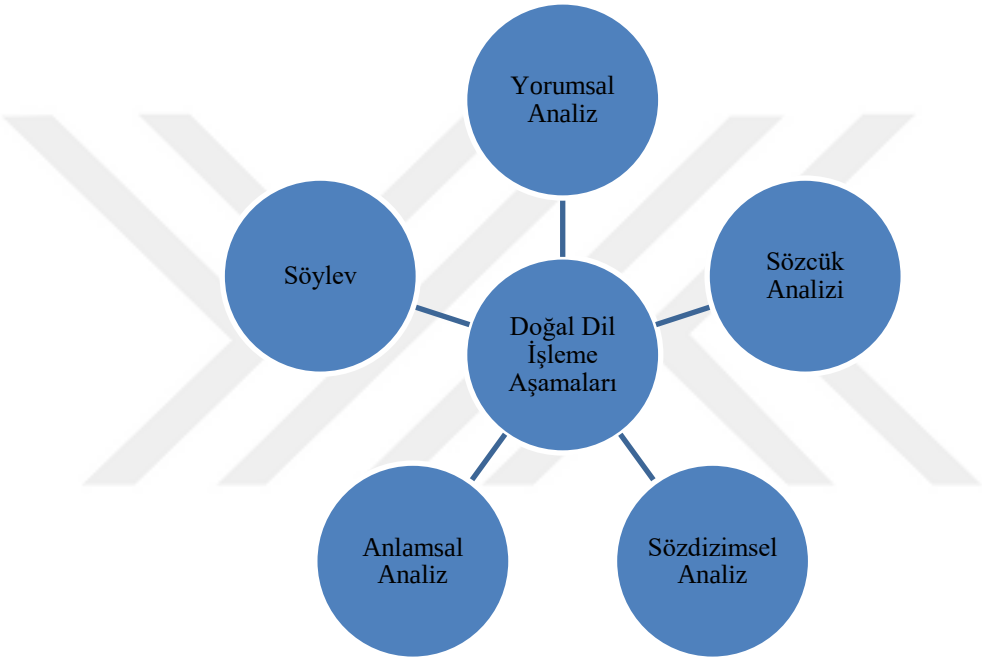
d. Bir dilden başka bir dile tercüme veya çeviri yapılabilmektedir. Böyle durumlarda metinde bulunan örneğin zamir görevindeki kelimenin hangi kelimenin yerine kullanıldığının bulunması kısmi bilgi yorumlanması olarak tanımlanmaktadır. İngilizce “he, she, it” zamirleri örnek olarak düşünülebilir (Şeker, 2015).

e. Pragmatik belirsizlik; kullanılan bir ifadenin, kullanan kişi tarafından taşıdığı doğruluk veya kesinlik taşıması kullandığı kişi tarafından aynı doğruluk veya

kesinlik değerini taşımaması durumudur. Bu tür bir belirsizliğin giderilmesi için bazı yöntemleri bulunmaktadır. Örneğin, dünyaca kabul görmüş birimlerin (yoğunluk, sıcaklık, uzunluk) kullanılarak pragmatik belirsizliğin önüne geçilebilmektedir.

1.2.4. Doğal dil işlemenin aşamaları

Doğal dil işlemenin aşamaları Şekil 1.15'de gösterildiği gibi olup bu aşamalar ilerleyen bölümlerde açıklanmıştır.



Şekil 1.15. Doğal dil işleme aşamaları

1.2.4.1. Sözcük analizi (Lexical Analysis)

Bir metinde geçen sözcüklerin yapısını tanımlamayı ve analiz etmeyi amaçlar. Bir dile ait sözlük, o dilde bulunan kelimelerin ve deyimlerin bir araya gelmesiyle oluşmaktadır. Kısaca sözcük analizi, bir metnin paragraflara ardından cümlelere ve son olarak kelimelere ayrılması işlemidir. Sözcük analizinde yaygın olarak kullanılan iki yöntem (Stemming ve Lemmatization) bulunmaktadır. Stemming; ek almış bir kelimenin kök haline indirgenmesi olarak açıklanabilir. Lemmatization ise morfolojik analiz yardımıyla sözcüklerin temel biçimlerine indirgenmesi adımıdır.

1.2.4.2. Sözdizimsel analiz (Syntactic Analysis)

Sözcüklerin dizilişinin ve sözcükler arasındaki yapısal ilişkinin incelendiği aşamadır. Bu ilişki her dile özgü dilbilgisi (gramer) kuralları dikkate alınarak yapılır. Sözcüklerin belirli bir düzende dizilimi önemlidir çünkü sözcüklerin yer değiştirmesiyle bambaşka bir anlam ile karşılaşılabilir. Cümleyi oluşturan yüklem, özne vb. cümlenin ögeleri dilbilgisi kurallarına uygun dizilime sahip olmalıdır.

1.2.4.3. Anlamsal analiz (Semantic Analysis)

Bir cümlenin tam olarak ne alam ifade ettiği üzerinde durulur. Cümlenin tam olarak ne alam ifade ettiğinin anlaşılması sözcük veya deyim seviyesinde analiz yapmaktan ziyade ifade ettiği anlamın yorumlanması şeklinde gerçekleştirilir.

1.2.4.4. Söylev (Discourse)

Kelimeler ve cümleler arasında bağlantı kurarak, metnin yapısının ve anlamının analizi ile ilgilenir (Özmutlu, 2021). Zamirlerin cümle içinde kullanımı örnek olarak verilebilir. Zamirlerin kimin yerine kullanıldığı insanlar açısından kolay bir şekilde anlaşılıyor olmasının yanında bilgisayarlar tarafından anlaşılması zaman zaman karmaşık bir hal alabilmektedir.

1.2.4.5. Yorumsal (Pragmatic Analysis)

Söylenenden ziyade söylenmek istenen üzerinde durulur. Bilgisayarlar, söylenen cümleden/kelimeden ortaya çıkan anlamlar arasında seçim yapmak durumunda kalırlar. Söylev ile pragmatik analiz birbirlerine yakın alanlarda çalıştığından zaman zaman karıştırılsa da özünde farklı konular olduğu söylenebilir.

1.2.5. Doğal dil işlemenin avantajları ve dezavantajları

Doğal dil işlemenin avantajları aşağıdaki gibi sıralanabilir:

a. Doğal dil işleme, insanların bilgisayarlarla doğal dilde iletişim kurmasına yardımcı olur.

b. Sağladığı birçok çözüm ve kolaylık sayesinde zaman ve iş gücü kazandırmaktadır.

c. Büyük ölçekli verilerin analizinde makine öğrenme algoritmalarıyla birlikte kullanılabilmektedir.

Doğal dil işlemenin dezavantajları aşağıdaki gibi sıralanabilir:

a. Türkçe gibi sondan eklemeli dillerle çalışmak zordur.

b. Dilin kurallara göre kullanılmaması ve anlamsız cümleler kullanılması doğal dil işleme çalışmayı zorlaştırmaktadır. (Nabıon eyi misin?)

c. Dilbilgisi kurallarına ve noktalama işaretlerine uyulmaması doğal dil işleme çalışmayı zorlaştırmaktadır.

2. KAYNAK ARAŞTIRMASI

Türkçe doğal dil işleme alanında yapılan çalışmalara bakıldığında diğer dillere (latin kökenli diller) kıyasla az sayıda olduğu görülmektedir. Bu başlık altında Türkçe ve diğer dillerde doğal dil işleme alanında yapılan çalışmalara değinilmiştir.

M. Fatih Amasyalı ve Banu Diri tarafından Türkçe metinler üzerinde metin sınıflandırma çalışması yapılmıştır. Metin sınıflandırma yapabilmek için n-gram yöntemi kullanılmıştır. Bu yöntem elde edilen veride tekrar oranını hesaplamayı sağlamaktadır. Buradaki n, tekrarın kontrol edildiği değere denk gelirken gram ise bu tekrarın ağırlığına denk gelmektedir. Yöntemin adı n değerine bağlı olarak özel bir isimlendirme ile anılabilmektedir. Değerin bir olması durumunda "unigram", iki olması durumunda "bigram" ve üç olması durumunda "trigram" olarak literatürde geçmektedir (Şeker, 2011). Diğer durumlarda yani n değerinin üçten büyük olduğu durumlarda genellikle n-gram olarak geçmektedir. N-gram doğal dil işlemede metin sınıflandırmadan başka uygulamalarda da kullanılmaktadır. Örneğin telefonunuzda mesaj yazarken son yazmış olduğunuz kelimeden sonra gelmesi muhtemel kelimeyi telefonunuzun size önermesinde istatistiksel verilerle birlikte n-gram kullanılmaktadır. İlk olarak Türkçe metinlerden oluşan veri setinde bigram ve trigram değerleri hesaplanmıştır. Elde edilen bigram ve trigram değerleri kullanılarak üç farklı sınıflandırma (metnin yazarını, metnin türünü ve yazarın cinsiyetini) yapılmış ve iki farklı n-gram modeli için toplamda altı adet veri seti oluşturulmuştur. Yapılan çalışmada Türkçe bir metnin yazarının belirlenmesi, metnin türüne göre sınıflandırılması ve yazarın cinsiyetinin tespit edilmesi üzerinde çalışılmıştır. Naive Bayes, Destek Vektör Makinesi, C 4.5 ve Random Forest sınıflandırma yöntemleri kullanılarak yapılan çalışmada metnin yazarını, metnin türünü ve yazarın cinsiyetini belirlemede başarı oranları sırasıyla %83, %93 ve %96 olarak elde edilmiştir (Amasyalı ve Diri, 2006).

Filiz Türkoğlu, Banu Diri ve M. Fatih Amasyalı tarafından yapılan çalışmanın amacı metnin yazarının belirlenmesidir. Yazarı belirlenecek metinden çeşitli yazarlık öznitelikleri ve n-gram değerleri elde edilerek Naive Bayes, SVM, K-NN, RF ve MLP sınıflandırma metotları karşılaştırmalı olarak analiz edilmiştir. En başarılı sınıflandırıcının MLP ve SVM olduğu, n-gram'ların yazarlık özniteliklerine göre daha yüksek doğruluk oranı verdiği gözlemlenmiştir. Bununla birlikte her ikisinin birlikte kullanılması tek tek kullanılmalarına göre daha iyi sonuçlar vermektedir. Son olarak

kullanılan sınıflandırıcıların etkinliği on kat çapraz doğrulama kullanılarak değerlendirilmiştir (Türkoğlu ve ark., 2007).

Metinlerin otomatik olarak sınıflandırılması mümkündür. Bir metnin yazarı, hangi türde yazıldığı veya yazarının cinsiyeti gibi bilgiler bilgisayarlar tarafından tahmin edilebilmektedir. Bu işlemin yapılabilmesi için sınıfı belirli olan metinler kullanılarak bir veri seti oluşturulur. Bu metinlerin karakteristik özellikleri incelenerek sınıfı bilinmeyen metinler bu özellikler kullanılarak sınıflandırılır. Bu sınıflandırmanın yapılabilmesi için metnin en iyi şekilde nasıl temsil edileceğine karar verilmesi gerekmektedir. Bununla birlikte metin temsil yöntemlerinin sınıflandırma problemleri ile olan etkileşimi ayrı bir araştırma konusudur. Daha önce yapılan metin sınıflandırma çalışmaları incelendiğinde en sık kullanılan metin temsil yöntemlerinin kelime grubu frekansları, n-gram frekansları vb. yöntemler olduğu görülmektedir. M. Fatih Amasyalı ve arkadaşları tarafından yapılan çalışmada Türkçe metinlerin sınıflandırılmasında kullanılan metin temsil yöntemlerinin performans karşılaştırılması yapılmıştır. Sık kullanılan metin temsil yöntemlerine ilave olarak kendi önerdikleri metin temsil yöntemlerinin farklı veri kümeleri üzerinden kıyaslaması yapılmıştır. Çalışmada altı farklı sınıflandırma problemi üzerinde çalışılmış ve deney sonuçları paylaşılmıştır. Türkçe sınıflandırma veri kümesi üzerinde on yedi adet özellik grubunun (cümle, kelime, ek sayıları, n-gramlar, kelimeler, kelime grupları ve saklı anlam indeksi özellik gruplarından bir kaçıdır) etkisi incelenmiştir. Metin temsil yöntemleri arasında n-gramların diğer yöntemlere kıyasla daha başarılı sonuçlar verdiği gözlemlenmiştir (Amasyalı ve ark., 2012).

Torunoğlu ve arkadaşları tarafından yapılan çalışmanın amacı, metin sınıflandırmasında önişleme yöntemlerinin Türkçe metinler üzerindeki etkisini incelemektir. Yapılan çalışmada Türkçe metin sınıflandırması için kök ayırma, durak kelime (stopword) filtreleme ve kelime ağırlıklandırma gibi önişleme yöntemlerinin detaylı analizi yapılmıştır (Torunoğlu ve ark., 2011).

Savaş Yıldırım ve Tuğba Yıldız tarafından yapılan çalışmada, geleneksel kelime torbası yaklaşımı ile sinir ağı tabanlı yaklaşım metin sınıflandırması açısından karşılaştırılmıştır. Çalışmada beş farklı sınıflandırma algoritması kullanılmıştır. N-gram yönteminin diğer özelliklerle birlikte kullanıldığında başarılı sonuçlar verdiği gözlemlenmiştir (Yıldırım ve Yıldız, 2018).

Özcan Kolyiğit tarafından yapılan yüksek lisans tez çalışmasında Türkçe bir metnin yazarının tespit edilebilmesi amacıyla bir sistem geliştirilmiştir. Veri setinin

oluşturulmasında farklı yazarlara ait köşe yazıları bir araya getirilmiştir. Toplamda altı farklı yazara ait yetmişer adet köşe yazısı alınarak 420 adet köşe yazısından oluşan veri seti elde edilmiştir. Veri seti test ve eğitim için bölünerek sistem eğitilmiştir. Geliştirilen sistem SVM ve K-NN sınıflandırma algoritmaları kullanılarak test edilmiştir. SVM algoritmasının K-NN algoritmasına kıyasla daha başarılı sonuçlar verdiği sonucuna ulaşılmıştır (Kolyiğit, 2013).

Salimkan Fatma Taşkıran tarafından yapılan yüksek lisans tez çalışmasında doğal dil işleme ile akademik metinlerin kümelenmesi yapılmıştır. Çalışmada büyük boyutlara ulaşan veriler arasından istenen veriye en hızlı ve kolay bir şekilde ulaşılabilmesi için kümeleme yapılması amaçlanmıştır. Akademik alanda çalışma yapan araştırmacılar açısından bahse konu metinlerin kümelenmesi son derece önemlidir. Metinlerin Türkçe yazılmış olan ön söz bölümlerinden bir veri seti oluşturulmuştur. Veri seti çeşitli ön işlemlerden geçirilmiş ve metin temsil yöntemleri kullanılarak kümeleme yapılmıştır. Kümeleme metodu olarak K-Means, OPTICS, Affinity Propagation ve K-Medoids algoritmaları kullanılmıştır. TF-IDF ve Word2Vec temsil yöntemleri kullanılarak kümeleme yapıldığında OPTICS ve Affinity Propagation algoritmalarının K-Means ve K-Medoids algoritmalarına kıyasla daha iyi sonuçlar verdiği görülmüştür (Taşkıran, 2021).

Ali Öztürk ve arkadaşları tarafından yapılan çalışmada, kullanıcıların Twitter üzerinden paylaşmış oldukları mesajlar içerisinde hastalık konulu olan mesajlar ve hangi hastalık olduğu tespit edilmiştir. Twitter üzerinden paylaşılan mesajlardan (1032 adet mesaj) oluşan veri seti çeşitli ön işlemlerden geçirilmiştir. Bu mesajlar on farklı şehirde bulunan kullanıcılar tarafından atılan mesajlardan oluşmaktadır. Ön işlemeden geçirilen metinler üzerinde öznitelik çıkarımı yapılmış ve makine öğrenmesi algoritmaları kullanılarak sınıflandırma yapılmıştır. Sınıflandırmanın yapılabilmesi için elde edilen veriler araştırmacılar tarafından hastalık belirtisi olmayan mesajlar "sağlıklı", alerji belirtileri içeren mesajlar "alerji" ve nezle belirtilerini içeren mesajlar "nezle" olarak etiketlenmiştir. Sınıflandırma algoritması olarak toplamda 8 adet algoritma (K-Means, K-NN, Karar Ağacı, Topluluk Öğrenmesi, SVM, Rasgele Orman, Lojistik Regresyon, Naive Bayes) kullanılmıştır. Metin temsil yöntemi olarak TF-IDF ve BOW yöntemleri kullanılmıştır. Kullanılan algoritmalar kıyaslandığında gözetimli öğrenme algoritmalarının gözetimsiz öğrenme algoritmalarına göre daha başarılı sonuçlar verdiği görülmüştür. Metin temsil yöntemleri kıyaslandığında ise gözetimsiz öğrenme algoritmalarında TF-IDF yönteminin, gözetimli öğrenme algoritmalarında ise

BOW yönteminin daha iyi sonuçlar verdiği görülmüştür. Yapılan çalışmanın hastalıkların yayılma hızında meydana gelen artışların ve azalmaların tespitinde kullanılabileceği savunulmuştur (Öztürk ve ark., 2020).

Gülşen Eryiğit ve Eşref Adalı tarafından yapılan çalışmada Türkçe için morfolojik çözümleyici ve tasarımı sunulmaktadır. Türkçe kelimelerin analizini ek çıkarma yaklaşımıyla ve sözlük kullanmadan yapmak için yeni bir metodoloji önerilmektedir (Eryiğit ve Adalı, 2004).

Gülşen Eryiğit ve Kemal Oflazer tarafından yapılan çalışma, Türkçe için ilk istatistiksel bağımlılık ayrıştırıcısının sonuçlarını sunmaktadır. Ayrıştırma için farklı temsil birimleri kullanan istatistiksel modeller araştırılmıştır (Eryiğit ve Oflazer, 2006).

Şeker ve Eryiğit tarafından yapılan çalışma, haber metinlerinde kişi, yer ve kuruluş varlıklarını tespit etme amacı taşımaktadır (Şeker ve Eryiğit, 2012).

Yaşar Anıl Sansak tarafından yapılan yüksek lisans tez çalışmasında kekemeliğin doğal dil işleme yöntemleri kullanılarak tespit edilmesi ve makine öğrenmesi yöntemleri kullanılarak sınıflandırılması yapılmıştır. Çalışmada kekemeliğin hangi sınıf kekemelik olduğunun tespit edilerek insanların günlük yaşamında karşılaştıkları bu problemin çözümü noktasında bir potansiyel taşıdığı belirtilmiştir (Sansak, 2023).

Rezan Bakır tarafından yapılan doktora tez çalışmasında doğal dil işleme teknikleri ve derin öğrenme algoritmaları kullanılarak sosyal ağlarda spam tespiti için farklı modeller önerilmiştir. Önerilen modellerin başarısının kıyaslanabilmesi amacıyla farklı sınıflandırma algoritması kullanılmış ve elde edilen sonuçlar mukayese edilmiş ve sonuçlar paylaşılmıştır (Bakır, 2022).

Aybüke Güzel Altıntaş tarafından yapılan yüksek lisans tez çalışmasında doğal dil işleme ile edebi eserlerden otomatik olarak söz tespiti uygulaması yapılmıştır. Web kazıma yöntemi ile çeşitli web sitelerinden veri seti elde edilmiştir. Çevrimiçi platform olan Kaggle üzerinde minimum kullanıcı tarafından oylanan alıntılar elde edilmiştir. Nihai olarak 4554 satırdan oluşan bir veri seti elde edilmiştir. Alıntı çıkarımı için KRA ve MatchSum kullanılmıştır. Bunun sonucunda sırasıyla %27,24 ve %40,54 Rouge-1 skorlarına ulaşılmıştır (Altıntaş, 2022).

Mustafa Özkan tarafından yapılan yüksek lisans tez çalışmasında Türkçe bilimsel metinlerden oluşan veri seti üzerinde derin öğrenme ve doğal dil işleme yöntemleri kullanılarak sınıflandırma uygulaması yapılmıştır. Veri seti son 10 içerisinde Türkçe dilinde yazılmış olan akademik ve bilimsel araştırmalardan oluşmaktadır. Derin

öğrenme tekniği olan BERT algoritması kullanılarak %96 başarı oranı elde edilmiştir (Özkan, 2022).

Hakan Kekül tarafından yapılan doktora çalışmasında doğal dil işleme teknikleri kullanılarak güvenlik açığı vektörlerinin farklı çok sınıflı sınıflandırma algoritmaları ile tahmini gerçekleştirilmiştir. Veri seti, eğitim ve test veri setlerine bölünmüştür. Metin temsil yöntemi olarak BOW, TF-IDF, N-gram, Word2Vec, Doc2Vec ve FastText özellik çıkarımı yöntemleri kullanılmıştır. Sınıflandırma aşamasında MLP, Rasgele Orman, Karar Ağacı, Naive Bayes ve K-NN algoritmaları kullanılmıştır. Sınıflandırma sonuçlarını değerlendirebilmek için F1 değerlendirme (F1 Score), duyarlılık (Recall), kesinlik (Precision) ve doğruluk (Accuracy) hesaplamaları yapılmıştır. Modelin doğruluğu çapraz doğrulama kullanılarak denetlenmiştir. Çalışma kapsamında CVSS skorlama sisteminin 2.0 ve 3.1 versiyonları kullanılmıştır. 2.0 versiyonu ile yapılan deneyde; BOW tekniği kullanıldığında Karar Ağacı algoritmasının, TF-IDF tekniği kullanıldığında K-NN algoritmasının, bigram tekniği kullanıldığında Karar Ağacı, MLP ve K-NN algoritmasının daha başarılı sınıflandırma yaptığı sonucuna ulaşılmıştır. Yapılan çalışmada word embedding yöntemlerinin büyük veri setlerinde kullanılması durumunda daha başarılı sonuçlar elde edildiği sonucuna ulaşılmıştır. 3.1 versiyonu ile yapılan deneyde; BOW, N-gram ve TF-IDF teknikleri kullanıldığında K-NN algoritması ile daha başarılı sınıflandırma yapıldığı sonucuna ulaşılmıştır. Ayrıca önerilen modelin yazılım güvenlik açıklıklarını değerlendirmede etkili olabileceği belirtilmiştir (Kekül, 2022).

Hawar Sameen Ali Barzenji tarafından yapılan çalışmada Twitter üzerinden paylaşılan mesajlar kullanılarak duygu analizi yapılmıştır. Veri seti web kazıma yöntemi kullanılarak Twitter üzerinden paylaşılan mesajların toplanmasıyla elde edilmiştir. Veri seti veri ön işlemeye tabii tutulmuştur. Doğal dil işleme tekniklerinden yararlanılmış ve (lemmatization, tokenization vb.) ve makine öğrenmesi yöntemleri kullanılarak sınıflandırma işlemi yapılmıştır. Sınıflandırma algoritması olarak Rasgele Orman, Naive Bayes ve Destek Vektör Makineleri kullanılmıştır. Bu algoritmalar kullanılarak yapılan sınıflandırmada sırasıyla %88, %72 ve %89 doğruluk oranları elde edilmiştir (Barzenji, 2021).

Serkan Korkmaz tarafından yüksek lisans tezi kapsamında makine öğrenmesi yöntemleri ve doğal dil işleme teknikleri kullanılarak edebi eserlerin şairinin tanınması çalışması yapılmıştır. Çalışmada, yazarı bilinmeyen veya belirlenmesinde tereddütte kalınan eserlerin yazarının bilgisayar tarafından otomatik olarak tespit edilmesi

amaçlanmıştır. Veri seti 75 şaire ait toplam 2805 adet şiirden oluşmaktadır. Veri setinde bulunan kelimelerin köklerinin ve aldığı eklerin tespit edilebilmesi için 64607 adet kelimeden oluşan Türkçe sözlük oluşturulmuştur. Ek almış sözcüklerin aldıkları ekleri tespit etmek amacıyla ek tablosu kullanılmıştır. Edebi metinlerde kullanılan kelimelerin yanlış kullanılmış olanları tespit edilmiş ve düzletilmiştir. Kelimelerin eserlerde kullanılan hali ve kök halini içeren tablo oluşturulmuştur. Beş farklı yazara ait onar adet eserin bulunduğu sekiz farklı veri seti, on farklı yazara ait onar adet eserin bulunduğu on farklı veri seti ve on beş farklı yazara ait onar adet eserin bulunduğu on farklı veri seti olmak üzere toplamda yirmi sekiz adet veri seti oluşturulmuştur. Oluşturulan veri setleri farklı makine öğrenmesi yöntemleri kullanılarak sınıflandırılmış ve doğruluk, duyarlılık ve kesinlik değerleri hesaplanarak karşılaştırılmıştır. Sınıflandırıcılara ait parametreler revize edilerek tekrar tekrar sınıflandırma yapılmış ve elde edilen sonuçlar değerlendirilmiştir. Çapraz doğrulama tekniği kullanılarak modelin performansı değerlendirilmiştir. Beş yazara ait elli eserin bulunduğu veri seti ile 857 öznitelik çıkarılarak yapılan sınıflandırmada Nu-Support Vector Machine algoritması ile %96, Extremely Randomized Trees algoritması ile %83 doğruluk oranı elde edilmiştir. Beş yazara ait 50 eserin bulunduğu veri seti ile 857 öznitelik çıkarılarak eğitim (%80) ve test (%20) modeli ile yapılan sınıflandırmada C Support Vector Machine algoritması kullanıldığında %100 doğruluk elde edilmiştir. Beş yazara ait 50 eserin bulunduğu veri seti ile iki yüz on dokuz öznitelik çıkarılarak yapılan sınıflandırmada Nu Support Vector Machine algoritması ile %96, Extremely Randomized Trees algoritması ile %83 doğruluk elde edilmiştir. Beş yazara ait 50 eserin bulunduğu veri seti ile iki yüz on dokuz öznitelik çıkarılarak eğitim (%80) ve test (%20) modeli ile yapılan sınıflandırmada C Support Vector Machine algoritması %93 doğruluk elde edilmiştir. On yazara ait yüz eserin bulunduğu veri seti ile 468 öznitelik çıkarılarak eğitim (%80) ve test (%20) modeli ile yapılan sınıflandırmada C Support Vector Machine algoritması %95 doğruluk elde edilmiştir. On beş yazara ait yüz eli eserin bulunduğu veri seti ile 2560 öznitelik çıkarılarak eğitim (%80) ve test (%20) modeli ile yapılan sınıflandırmada C Support Vector Machine algoritması %100 doğruluk elde edilmiştir. Sonuç olarak Nu Support Vector Machine ve C Support Vector Machine algoritmaları kullanıldığında daha yüksek doğruluk oranı elde edilmiştir (Korkmaz, 2021).

Atilla Suncak tarafından doktora tezi kapsamında Türkçe cümlelerdeki anlam bozukluklarının derin öğrenme yaklaşımları, makine öğrenmesi yöntemleri ve doğal dil işleme teknikleri kullanılarak tespit edilebilmesi için bir model önerilmiştir. Bu alanda

daha önce yapılan doğal işleme çalışmalarının olmaması beraberinde bir zorluk getirmektedir. Yapılan çalışma öncesinde, eğitim bilimci araştırmacılar tarafından çeşitli analizler yapılmış olsa da yapılan analizlerin hiçbiri makine öğrenimi yöntemleri kullanılarak gerçekleştirilmemiştir. Veri setinin oluşturulmasında kurs, okul veya sınav merkezi gibi eğitim kuruluşlarının çevrimiçi kaynaklarından veri toplama işlemi gerçekleştirilmiştir. Toplanan verinin miktarının modeli eğitmede yetersiz olacağı değerlendirildiğinden veri miktarını artırmaya yönelik işlem yapılmıştır. Word2vec kelime temsil yöntemi ile veri seti kullanılarak özellik çıkarımı için kullanılmak üzere bir kelime gömme derlemi oluşturulmuştur. Makine öğrenmesi yöntemlerinden K-NN, Random Forest ve SVM sınıflandırıcıları ile derin öğrenme yöntemlerinden LSTM ve CNN mimarileri kullanılarak deneysel çalışmalar yapılmıştır. Çalışmalardan elde edilen sonuçlar değerlendirildiğinde derin öğrenme modellerinin makine öğrenmesi sınıflandırıcılarına göre daha başarılı sonuçlar verdiği gözlemlenmiştir. Derin öğrenme modelleri kendi aralarında kıyaslandığında LSTM mimarisinin CNN mimarisine göre ön plana çıktığı görülmektedir. Bunun sebebinin, Türkçe gibi anlamsal bağlam ve morfolojik gramer açısından karmaşık bir yapıya sahip dilde bazı doğal dil işleme işlemleri için geleneksel yöntemlerin iyi bir seçenek olamayabileceği değerlendirilmektedir. Sonuç olarak; Türkçede kusurlu cümlenin otomatik olarak tespit edilebilmesinde derin öğrenme mimarilerinin kullanılması daha kabul edilebilir sonuçlar vermiştir (Suncak, 2022).

Emre Mumcuoğlu tarafından yüksek lisans tezi kapsamında Türk yüksek mahkemelerinde karar tahmini yapılabilmesi için doğal dil işleme yöntemleri kullanılmıştır. Yüksek mahkemelerin ve alt bağlılarının vermiş oldukları kararlar incelenerek bir veri seti oluşturulmuştur. Çalışmada yüksek mahkemelere yapılan 93,340 başvuru kullanılmıştır. Veriler yüksek mahkemelerin web sitesinden alınmıştır. Metin temsil yöntemi olarak unigram kullanılmıştır. Doğal dil işleme işlemlerinin yapılabilmesi için Zemberek aracı kullanılmıştır. Metinlerde geçen rakamsal değerler (tarih vb.) veri setinden çıkarılmıştır. Karar tahmininin yapılabilmesi için Karar Ağaçları, Rastgele Orman ve Destek Vektör Makineleri sınıflandırma algoritmaları ile birlikte LSTM ve GRU derin öğrenme mimarileri kullanılmıştır. Değerlendirme ölçütü olarak accuracy, balanced accuracy ve F1 score hesaplamaları yapılmıştır. Destek Vektör Makinelerinin Karar Ağaçları ve Rastgele orman algoritmalarından daha iyi sonuçlar verdiği gözlemlenmiştir. Kullanılan yöntemler arasında derin öğrenme mimarileri genel olarak yüksek doğruluk oranına sahip olmuştur (Mumcuoğlu, 2022).

Çağla Ballı tarafından yüksek lisans tezi kapsamında sosyal medya kullanıcıları tarafından paylaşılan Türkçe içerikli paylaşımlardan duygu analizi yapılması amaçlanmıştır. Kullanıcıların sosyal medya paylaşımları doğal dil işleme teknikleri ve makine öğrenmesi yöntemleri kullanılarak analiz edilmiş ve elde edilen sonuçlar paylaşılmıştır. Çalışmada iki farklı veri seti kullanılmıştır. Veri setinin birinde on bir bin gönderi bulunurken diğer veri setinde iki bin sekiz yüz gönderi bulunmaktadır. Veri seti Twitter üzerinden paylaşılan Türkçe gönderilerden oluşmaktadır. Elde edilen gönderiler araştırmacı tarafından etiketlenmiştir ve veri ön işlemeden geçirilmiştir. Veri setinde bulunan paylaşımlar makine öğrenmesi yöntemleri kullanılarak pozitif veya negatif olarak etiketlenmiştir. Etiketleme sonucunda paylaşımların %55'i pozitif %45'i ise negatif paylaşımlardan oluşmaktadır. Modelin eğitilebilmesi için veri seti eğitim (%80) ve test (%20) verilerine bölünmüştür. Doğal dil işleme işlemlerinin yapılabilmesi için Zemberek ve SnowBall araçları kullanılmıştır. Doğal dil işleme araçlarının başarısının deney yapılan veri setine göre farklılık gösterdiği tespit edilmiştir. Örneğin; Zemberek Kütüphanesi kullanıldığında Lojistik Regresyon, Rastgele Orman ve Destek Vektör Makineleri algoritmalarının sınıflandırma başarıları sırasıyla %85.51, %86.30 ve %87.47 olarak gerçekleşmiştir. SnowBall Kütüphanesi kullanıldığında ise %85.32, %86.49 ve %86.10 olarak gerçekleşmiştir. Metin temsil yöntemi olarak TF-IDF tekniği kullanılmıştır (Ballı, 2021)

Maher Alrefaai tarafından yüksek lisans tezi kapsamında, manuel olarak yapılan haberlerin sınıflandırılması işleminin makine öğrenmesi ve doğal dil işleme yöntemleri kullanılarak Türkçe haberlerin otomatik olarak sınıflandırılabilmesi için bir çalışma yapılmıştır. Veri setinde olumlu duygu içeren haberler "1" olumsuz duygu içeren haberler "-1" olarak işaretlenmiştir. Veri setinde sınıflandırma başarısına etki etmeyen karakterler çıkarılmıştır. Sınıflandırma başarısı artırabilecek bir işlem olan "stopword" kelimelerin çıkarılması gerçekleştirilmiştir. Makine öğrenimi sürecinin bir parçası olan veri setinin eğitim ve test veri setlerine bölünmesi işlemi gerçekleştirilmiştir. Veri setini oluşturan haberler metin formatında olduğu için bu metinlerin temsil yöntemi kullanılarak vektör haline getirilmesi gerekmektedir. Bunun için TF-IDF metin temsil yöntemi kullanılmıştır. Makinenin eğitilmesi için RF, BERT ve FastText modelleri kullanılmıştır. Modelin performansının ölçülebilmesi çeşitli hesaplamalar yapılmıştır. Kullanılan modeller kıyaslandığında BERT ve FastText'in, RF'e göre daha iyi sonuçlar verdiği görülmüştür. BERT ve FastText modelleri kıyaslandığında ise BERT'in FastText'e göre daha iyi sonuçlar verdiği ancak FastText ile modeli eğitmenin daha

hızlı olduğu tespit edilmiştir. Yüksek lisans tezi sonucunda ulaşılan doğruluk oranları BERT, FastText ve RF algoritmalar için sırasıyla %95, %92 ve %53 olarak verilmiştir (Alrefaai, 2021).

Alican Bozyiğit tarafından doktora tezi kapsamında yapılan çalışmada, sosyal medya üzerinden uygulanan zorbalığın makine öğrenmesi yöntemi kullanılarak otomatik olarak tespit edilmesi amaçlanmıştır. Çalışmada 5000 Türkçe içerikten oluşan ve çeşitli ön işlemlerden geçirilen bir veri seti kullanılmıştır. Makine öğrenimi sonuçlarını iyileştirmek için veri setine bazı ön işleme ve normalleştirme adımları uygulanmıştır. Veri setini oluşturan paylaşımlar metin formatında olduğu için bu metinlerin temsil yöntemi kullanılarak vektör haline getirilmesi gerekmektedir. Bunun için TF-IDF metin temsil yöntemi kullanılmıştır. Veri seti üzerinde yapılan ön işleme ve öznitelik çıkarımı işlemlerinden sonra makine öğrenmesi yöntemleri uygulanmıştır. Makine öğrenmesi yöntemi olarak SVM, LR, K-NN, RF, NBM ve AdaBoost kullanılmıştır. Yapılan deneyler sonucunda AdaBoost algoritmasının en iyi sonucu verdiği görülmüştür. Diğer makine öğrenimi algoritmalarının ise %90 civarında doğruluk oranına sahip olduğu görülmüştür. NBM ve K-NN algoritması diğerlerine kıyasla daha düşük doğruluk oranına sahiptir. Deneysel çalışma sonuçlarının dikkat çeken noktası, deneyde kullanılan her bir makine öğrenmesi algoritmasının, metinsel özelliklerin yanı sıra sosyal medya özelliklerini (takipçi sayısı vb.) de içeren veri kümesi üzerinde daha başarılı bir tahmin performansı vermesidir. Örneğin sosyal medyada takipçi sayısı fazla olan kullanıcıların zorbalık içeren paylaşım yapmaya meyilli olmaması bunun bir sonucudur (Bozyiğit, 2021).

Hatice Elif Ekim tarafından yüksek lisans tezi kapsamında yapılan çalışmada, sosyal medya üzerinden havayolu firmalarını ilgilendiren paylaşımların makine öğrenmesi yöntemleri, derin öğrenme yöntemleri ve doğal dil işleme teknikleri kullanılarak otomatik olarak sınıflandırılması amaçlanmıştır. Çalışmanın aşamaları sırasıyla; veri seti oluşturma ve etiketleme işlemini gerçekleştirme, çeşitli veri ön işleme adımlarının gerçekleştirilmesi, eğitim ve test veri setlerinin oluşturulması, özellik çıkarımı yapılması, terim ağırlıklandırma işlemi, özellik seçimi, sınıflandırma ve sonuçların karşılaştırmalı olarak değerlendirilmesi şeklinde özetlenebilir. Veri seti, Twitter API ve Twitter Archive Google Sheets yöntemleri kullanılarak elde edilen on dört bin dört yüz altı adet Türkçe paylaşımdan oluşmaktadır. Paylaşımlar sefer iptalleri, kampanyalar gibi başlıklar belirlenerek etiketlenmiştir. Veri seti doğal dil işleme teknikleri kullanılarak sınıflandırma işlemine hazır hale getirilmiştir. Sınıflandırma

aşamasında Naive Bayes, DT, SVM, RF ve LR makine öğrenmesi yöntemleri ile birlikte CNN ve LSTM derin öğrenme mimarileri kullanılmıştır. Bagging, Adaboost ve XGBoost topluluk öğrenmesi yöntemleri ile de deneyler yapılmıştır. Sınıflandırma performansının ölçülmesi için çeşitli metrik hesapları yapılmıştır. Doğal dil işleme teknikleri için İstanbul Teknik Üniversitesi Doğal Dil İşleme Araştırma Grubu tarafından geliştirilen İTÜ NLP Kütüphanesi ve Ahmet Akın tarafından geliştirilen Zemberek Kütüphanesi kullanılmıştır. Metin temsil yöntemi olarak N-gram yöntemi, terim ağırlıklandırma yöntemi olarak TF-IDF yöntemi kullanılmıştır. Sınıflandırma işleminde ön plana çıkan SVM ve RF algoritmaları ile %77 doğruluk oranı elde edilmiştir. Derin öğrenme yöntemleri olan LSTM ve CNN mimarilerinde ise %76 doğruluk oranı elde edilmiştir (Ekim, 2021).

Fatih Gürcan tarafından yapılan çalışmada, Türkçe haber metinlerinin beş kategoride (ekonomi, siyaset, teknoloji, spor, sağlık) sınıflandırılması işlemi gerçekleştirilmiştir. Bu bağlamda, Multinomial Naïve Bayes, Bernoulli Naïve Bayes, SVM, K-NN ve DT algoritmalarının Türkçe haber metinleri üzerindeki sınıflandırma performansları değerlendirilmiştir. Multinomial Naïve Bayes algoritması diğer algoritmalara kıyasla en yüksek doğruluk oranına sahip olmuştur. Bu oran yaklaşık %90 olarak elde edilmiştir. Bu durum Multinomial Naïve Bayes algoritmasını Türkçe metinler ile sınıflandırma probleminde ön plana çıkarmaktadır (Gürcan, 2018).

Ahmet Bardız tarafından yapılan yüksek lisans tez çalışmasında hastaların başvuracağı bölümün makine öğrenmesi yöntemleri ile tespit edilmesi hedeflenmiştir. NBM ve SVM algoritmaları tıbbi metinlerin sınıflandırılmasında yaygın kullanılan iki algoritmadır. Veri seti üniversite hastanelerinin gerçek kayıtlarından oluşmaktadır. Diğer bir ifadeyle alanında uzman doktorların hastalarını muayene etmesi sonucu ortaya çıkan verilerdir. Çalışmada; fizik tedavi ve rehabilitasyon, cildiye, gastroenteroloji, jinekoloji, kardiyoloji, kulak burun boğaz, ortopedi, plastik cerrahi, psikiyatri ve üroloji branşlarından toplanan veriler kullanılmıştır. Her bölümden onar bin olmak üzere toplamda yüz bin kayıt toplanmıştır. Her kayıt, hastaların şikâyetlerini ve doktorların hastalarını yönlendirdikleri bölüm bilgisini içermektedir. Doğal dil işleme aşamasında morfolojik analiz için Zemberek Kütüphanesi kullanılmıştır. Veri seti %80 eğitim ve %20 test olmak üzere iki veri setine ayrılmıştır. Metin temsil yöntemi olarak N-gram yöntemi kullanılmıştır. Sınıflandırma işlemi sonucunda SVM algoritmasının %98.16 doğruluk oranı verdiği sonucuna ulaşılmıştır. Yapılan deneylerde SVM algoritmasının fizik tedavi ve rehabilitasyon ile ortopedi bölümlerini birbiri yerine tahmin ettiği

görülmüştür. NBM algoritmasının sınıflandırma doğruluk oranı ise %96.7 olarak elde edilmiştir. Son olarak yapay sinir ağları mimarisi olan CNN algoritmasında ise bu oran %94,6 olarak elde edilmiştir (Bardız, 2019).

Spam filtreleme istenmeyen e-postaların tespit edilerek sınıflandırılması işlemidir. Sevinj Shirzadova tarafından yüksek lisans tezi kapsamında bir video paylaşım platformunda yapılan istenmeyen yorumların sınıflandırılması üzerine çalışma yapılmıştır. Çalışmada kullanılan veri seti, video paylaşım platformunda (YouTube) 2017 yılında izlenme sayısı en yüksek beş videoya yapılan Türkçe yorumlardan oluşmaktadır. Veri seti çeşitli ön işleme (büyük-küçük harf dönüşümü vb.) adımlarından geçirilmiştir. Doğal dil işleme aşamasında Zemberek Kütüphanesi kullanılmıştır. Makine öğrenimi yöntemlerinin uygulanması amacıyla Weka yazılımı kullanılmıştır. Veri seti eğitim (%90) ve test (%10) veri setlerine bölünmüştür. Sınıflandırma işlemi için birden fazla sınıflandırma algoritması kullanılmış ve sonuçlar karşılaştırılmıştır. Kullanılan başlıca algoritmalar Naive Bayes, NBM ve RF algoritmalarıdır. Deneyden elde edilen sonuçlar kıyaslandığında Naive Bayes, Jrip ve SMO algoritmalarının daha yüksek doğruluk oranı verdiği gözlemlenmiştir (Shirzadova, 2020).

Rabia Kontuk tarafından yüksek lisans tezi kapsamında, haberlerin yaş gruplarına göre sınıflandırılması üzerine bir model önerilmiştir. Doğal dil işleme işlemlerinin gerçekleştirilebilmesi için Zemberek Kütüphanesi kullanılmıştır. Veri seti iki farklı haber sitesinin web sayfasında elde edilen haberlerden oluşmaktadır. Veri seti çeşitli ön işleme (eksik verinin tamamlanması vb.) adımlarından geçirilmiştir. Veri seti eğitim (%70) ve test (%30) olarak ikiye ayrılmış ve toplamda üç bin haberden oluşmaktadır. Sınıflandırma; ergenlik, çocukluk ve yetişkinlik yaş grubu olmak üzere üç başlıkta yapılmıştır. Yapılan deney sonucunda ilk olarak %71 doğruluk oranı elde edilmiştir. Yalnızca isim içeren sözlük kullanılarak deney yapıldığında ise %73 doğruluk oranı elde edilmiştir. Ergenlik sınıfı ile çocuk sınıfı birleştirilip sınıflandırma yetişkin ve yetişkin olmayan olmak üzere iki sınıfa indirgendiğinde %83 oranında doğruluk oranı elde edilmiştir (Kontuk, 2020).

Semuel Franko tarafından yapılan yüksek lisans tezi kapsamında, İspanyolca metinlerden oluşan bir derlem üzerinde, makine öğrenmesi yöntemleri kullanılarak sınıflandırma modeli oluşturulması ve modelin parametreleri optimize edilerek kıyaslama yapılması amaçlanmıştır. Açık kaynak bir veri seti bulunmadığından farklı kaynaklardan elde edilen İspanyolca metinler kullanılmıştır. Veri seti oluşturulduktan sonra veri ön işleme ve doğal dil işleme teknikleri uygulanmıştır. Metin temsil yöntemi

olarak birkaç teknik kullanılmıştır. Sınıflandırma aşamasında DT, SVM, Naive Bayes ve Maximum Entropy algoritmaları kullanılmıştır. Bu algoritmaların başarılarının değerlendirilmesi için çapraz doğrulama yöntemi kullanılmıştır. Yapılan analizler sonucunda doğruluk oranlarında %2 ile %16 arasında artış meydana gelmiştir. Sınıflandırma sonucunda en yüksek doğruluk oranı %89 olarak gerçekleşmiştir (Franko, 2018).

3. MATERYAL VE YÖNTEM

Tez kapsamında gerçekleştirilen uygulamanın kodlanması Anaconda dağıtımı versiyon 2.2.0, Python 3.10.4 versiyon ile Jupyter Notebook 6.4.8 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 Pandas kütüphanesi; üst düzey matematiksel işlemler için versiyon Numpy kütüphanesi; makine öğrenmesi modelleri ve bazı önişleme modülleri için versiyon Scikit-learn kütüphanesi kullanılmıştır.

Veri seti olarak internet sitelerinden web kazıma yöntemi ile elde edilen, 21 kategoriden ve toplam 4198 şiirden oluşan bir veri seti kullanılmıştır. Çizelge 3.1'de veri setinde yer alan kategoriler yer almaktadır.

Çizelge 3.1. Veri setinde bulunan kategoriler

1. Anne	**8.** Asker	**15.** Atatürk
2. Aşk	**9.** Baba	**16.** Bayrak
3. Bayram	**10.** Deniz	**17.** Hasta
4. Kadın	**11.** Para	**18.** Savaş
5. Spor	**12.** Tanrı	**19.** Türkiye
6. Vatan	**13.** Yemek	**20.** Yolculuk
7. Öfke	**14.** Öğretmen	**21.** Şehit

Veriler internet sitesinden veri kazıma yöntemi ile elde edilmiştir. Veriler CSV formatında bir tabloda tutulmuştur. Tablo içeriğinde şiir adı, şair adı, bağlantı, şiir kategorisi ve şiir içeriği bulunmaktadır.

Başlık	Şair	Kategori	Şiir
Bayram	Mehmet Akif Ersoy	Bayram	Âfâk bütün hande, cihan başka cihandır;
Bayram	Necip Fazıl Kısakürek	Bayram	Ölüm ölene bayram, bayrama sevinmek var;
Süleymaniye'de Bayram Sabahı	Yahya Kemal Beyatlı	Bayram	Artarak gönlümün aydınlığı her saniyede
Bayram Duası	Ozan Arif	Bayram	Ya Rabbi tadına bütün milletin
Bayramlar Bayram Ola-1	Abdurrahim Karakoç	Bayram	Güneş yükselmeden kuşluk yerine

Şekil 3.1. Veri seti

Doğal dil işleme işlemlerinin yapılabilmesi için Zemberek Kütüphanesi kullanılmıştır. Zemberek, Ahmet Afşın Akın tarafından Java programlama dili

kullanılarak geliştirilmiş açık kaynak Türkçe doğal dil işleme kütüphanesidir. Zemberek Kütüphanesi kullanılarak varlık isim tanıma, metin sınıflandırma, dil tanımlama, cümlenin ögelerine ayırma, kelimelerin kökünü bulma, normalizasyon gibi işlevler yerine getirilebilmektedir. Zemberek sadece Türkçe için değil diğer Türk dilleri için de doğal dil işleme imkânı sunmayı amaçlamaktadır. Açık kaynaklı olmasından kaynaklı, kütüphaneye yeni bir Türk dili eklemek mümkündür. Zemberek kütüphanesinin Python programlama dili ile kullanılabilmesi için Jype modülünün projeye dâhil edilmesi gerekmektedir.

3.1. Veri Setinin Hazırlanması

Büyük boyutlu verilerin kullanılmadan önce çeşitli ön işlemlerden geçirilmesine ihtiyaç duyulmaktadır. Bu işlem daha iyi sonuçlar elde edilebilmesi amacıyla yapılmaktadır. Bu işlemin yapılmaması gerçekten uzak sonuçlar alınmasına yol açabilmektedir. Bu çalışmanın veri önişleme safhasında aşağıda belirtilen işlemler gerçekleştirilmiştir.

a. Elde edilen şiirlerden iki yazara ait birer şiir veri setinden çıkarılmıştır. 21 adet kategoride 200'er adet şiir ve toplamda 4200 adet şiirden oluşan veri seti kullanılması planlanmakta iken veri kazıma esnasında hatalı olarak çekilen iki şiirin çıkarılmasıyla 4198 adet şiirden oluşan veri seti kullanılmıştır. Şekil 3.2'de, veri setinde bulunan kategoriler ve her bir kategoride bulunan şiir sayısı verilmiştir.

S/N	Kategori	Şiir Sayısı	S/N	Kategori	Şiir Sayısı
1	Anne	200	11	Tanrı	200
2	Asker	200	12	Türkiye	200
3	Atatürk	200	13	Vatan	200
4	Aşk	200	14	Yemek	200
5	Baba	200	15	Öfke	200
6	Bayram	200	16	Öğretmen	200
7	Bayrak	200	17	Şehit	200
8	Deniz	200	18	Spor	200
9	Hasta	199	19	Yolculuk	199
10	Kadın	200	20	Savaş	200
			21	Para	200

Şekil 3.2. Kategorilerde bulunan şiir sayısı

b. Satır boşluklarının kaldırılması, metnin küçük harfe çevrilmesi ve noktalama işaretlerinin kaldırılması yapılmıştır. Bu işlem kelimelerin birbirleri ile karşılaştırılması veya vektör değerlerini hesaplarken sağlıklı sonuçlar alınabilmesi için önemlidir.

c. Kelime belirteçlerinin elde edilmesi yani metnin cümlelere, ardından kelimelere ayırma işlemi gerçekleştirilmiştir. Bu işlem ham metnin daha küçük parçalara ayrılması işlemi olarak da ifade edilebilir.

d. Zemberek Kütüphanesinde bulunan bir Türkçe stopword listesi kullanılarak metindeki stopword kelimelerinin kaldırılması gerçekleştirilmiştir. Etkisiz kelimelerin metinlerden çıkarılması işlemi stopword kelimelerin atılması işlemi olarak bilinmektedir. Bu işlem stop-word.txt dosyasında bulunan 1797 adet kelime ile karşılaştırılarak eşleşmesi durumunda veri setinden atılması suretiyle gerçekleştirilmiştir. Türkçede anlama etki etmeyen kelimeler genellikle bağlaçlar, zamirler, edatlar veya soru ekleri gibi metnin konusu ile ilgili herhangi bir anlamı bulunmayan genel kelimelerdir. Metne etki etmeyen kelimelerin çıkarılması, çalışılacak verinin boyutunu düşürmekte ve aynı zamanda anlamda değişikliğe neden olmamaktadır (Taşkıran, 2021).

Çizelge 3.2. Durak kelimeler dosyasında bulunan bazı kelimeler

acaba	sırasıyla	töskürü
abes	sonradan	tükenik
hatta	süratle	tümüyle
acaba	sırasıyla	töskürü
abes	sonradan	tükenik
hatta	süratle	tümüyle

e. Bu aşamada Zemberek Kütüphanesi ile kelime köklerinin elde edilmesi işlemi gerçekleştirilmiştir. Bu aşamanın önemi, ek almış kelimenin makine tarafından farklı bir kelime gibi algılanmasından kaynaklanmaktadır. Ek alan kelime anlam olarak değişikliğe uğrayabileceği gibi uğramadığı durumlar da olabilir. Ancak bu durumun makine tarafından anlaşılması beklenemez. Dolayısıyla kelime frekanslarında farklı dağılımlara sebep olabilmektedir.

Veri önişleme adımlarının tamamlanmasının ardından veri setinde bulunan verilerin temsil edilmesi gerekmektedir. Bu çalışmada doğal dil işlemede sıklıkla

kullanılan TF-IDF (Term Frequency — Inverse Document Frequency) yöntemi kullanılmıştır.

Metinler üzerinde istatistiksel incelemeler konusunda kullanılan TF-IDF kavramı İngilizce Term Frequency — Inverse Document Frequency kelimelerinin baş harflerinden meydana gelmektedir. Temel anlamda, bir metinde geçen kelimelerin bulunması ve kelimelerin tekrarlama sayılarına göre çeşitli hesaplamaların yapılması esasına dayanmaktadır. Bu hesaplamalar; TF ve IDF değerinin hesaplanması ve bu değerlerin çarpımıyla TF-IDF değerinin bulunmasıdır. Böylece; bir sözcüğün, bulunduğu dokümanı ne kadar temsil ettiğinin istatistiksel bir değeri elde edilmiş olur. TF_IDF matrisi ise bu değerlerin bütün kelimeler ve metinler için hesaplanmış halidir. Örnek vererek açıklamak gerekirse, elimizde 200 adet dokümandan oluşan bir veri kümesi olduğunu kabul edelim. TF-IDF değerinin de ilk olarak okul kelimesi için hesaplanacağını kabul edelim. İlk adımda okul kelimesinin birinci dokümanda kaç kez geçtiği bulunur. Bulunan değerin 8 olduğunu varsayalım. İkinci adımda okul kelimesinin elimizde bulunan 200 dokümanın kaçında geçtiği bulunur. Burada dikkat edilmesi gereken nokta kelimenin geçip geçmediğine bakılıyor olmasıdır. Yani ikinci adımda okul kelimesinin diğer dokümanlarda kaç defa geçtiğine değil 200 dokümanın kaçında geçtiği bulunur. Bulunan değerin 20 olduğu kabul edilir ve TF ve IDF değerleri ayrı ayrı hesaplanarak bulunan değerlerin çarpılması suretiyle TF-IDF değeri hesaplanmış olur.

TF değerinin hesaplanabilmesi için ihtiyaç duyulan bir diğer değer, üzerinde çalışılan dokümanda (birinci dokümanda çalışıldığı kabul edilmiş olsun) en sık geçen kelimenin kaç defa geçtiğidir. Bu değer TF-IDF değeri hesaplanacak kelimenin (okul kelimesinin 8 defa geçtiği kabul edilmişti) tekrar sayısının bu değere oranlanabilmesi için gereklidir. Bu değerin 160 olduğunu varsayarak hesaplama işlemine geçelim. TF değeri 8/160 işlem sonucu olan 0,05 olarak hesaplanır.

Ardından IDF değerinin hesaplanmasına geçilir. IDF değeri toplam doküman sayısının ilgili kelimeyi içeren doküman sayısına oranlanarak logaritmasının alınmasıyla hesaplanabilir. Bu formüle göre toplam doküman sayısı 200, terimin geçtiği doküman sayısı 20 olduğu bilindiğinden oran 200/20'den 10 olarak hesaplanır. Logaritma alma işlemi de yapıldığında IDF değeri 1 olarak bulunmuş olacaktır. Burada dikkat edilmesi gereken husus logaritma hesabı yapılırken tabanın ne olduğunun öneminin olmadığıdır. Yani TF-IDF değeri kelimeler arasında kıyas için kullanıldığından tabanın tüm kelimelerde aynı seçilmesi sonucu etkilemeyecektir. Diğer

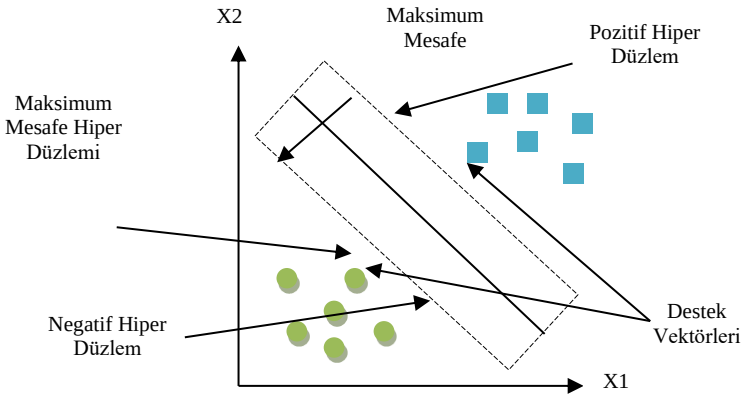
bir dikkat edilmesi gereken nokta ise terimi içeren doküman sayısının sıfır olma ihtimaldir. Bu değer paydada olduğundan belirsizlik durumunun ortaya çıkma ihtimali vardır. Bu sorunun çözümü için bu değere 1 eklenmesi sıkça kullanılan bir yöntemdir.

3.2. Sınıflandırma Algoritmaları

Bir verinin hangi gruba ait olduğunun bilgisayarlar tarafından tespit edilerek otomatik olarak sınıflandırılması için sınıflandırma algoritmaları kullanılır. Bu sınıflar önceden belirlenmiş sınıflardır. Sınıflandırma algoritmaları çeşitli hesaplar sonucunda bir verinin hangi sınıfa ait olduğuna karar verir. Bu tez çalışmasında altı farklı sınıflandırma algoritması kullanılmıştır.

3.2.1. Destek vektör makineleri

Destek vektör makineleri İngilizce literatürde Support Vector Machine, SVM olarak kısaltılmaktadır. SVM, doğrusal ve doğrusal olmayan sınıflandırma problemlerinde kullanılan gözetimli öğrenme yöntemlerinden biridir. Veriyi birbirinden ayırmak için en uygun fonksiyonun tahmin edilmesi esasına dayanmaktadır. Bir düzlemde bulunan iki grup arasında çizilen çizgi grupları ayırmaya imkân tanımaktadır. Burada çizilen çizgi iki grubun üyelerine de maksimum uzaklıkta olmalıdır. Bu sınırın nasıl çizileceği SVM tarafından belirlenmektedir. İki gruba teğet olacak şekilde birer çizgi çizildikten sonra (maksimum uzaklık dikkate alınarak) bu iki çizginin tam ortasına çizilen çizgi sınıflandırmada kullanılan çizgi olacaktır. SVM, günümüzde yüz tanıma sistemlerinden, ses analizine kadar birçok sınıflandırma probleminde kullanılmaktadırlar.

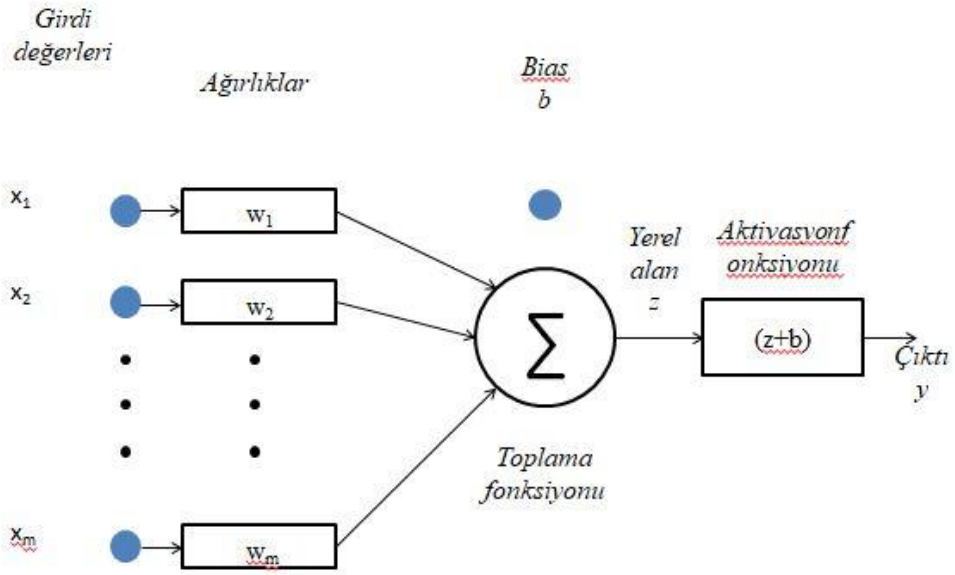


Şekil 3.3 Destek vektör makineleri

3.2.2. Çok katmanlı algılayıcılar

Çok katmanlı algılayıcılar İngilizce literatürde Multi Layer Perceptron olarak geçmekte ve MLP şeklinde kısaltılmaktadır. Makine öğrenmesi algoritması olan çok katmanlı algılayıcılar sınıflandırma problemlerinde etkin çalışan bir algoritmadır. Temelde, öğrenme ve karar verme olmak üzere iki işlevi yerine getirmeyi amaçlamaktadır. MLP, ağırlıklandırma aktivasyon fonksiyonu ve bias sayesinde öğrenme ve karar verme işlevlerini gerçekleştirir. Bir sonraki katmana aktarılmak üzere giriş katmanına gelen veriler bir ağırlık değeri ile çarpılır. Tüm girdiler, kendilerine özgü ağırlık değerleriyle çarpılarak toplanır. Aktivasyon fonksiyonu kendine gelen değere göre karar verir. Bias değeri, kullanıcıdan kullanıcıya, sistemin çalışma biçimine veya amacına göre değişebilen, kullanıcı tarafından eklenen bir parametredir (Bulut, 2016).

Çok katmanlı algılayıcılar; Şekil 3.4'de (Bulut ve Bozyılan) görüldüğü üzere girdi değerleri, ağırlıklar, bias, toplama fonksiyonu ve aktivasyon fonksiyonundan meydana gelmektedir.



Şekil 3.4. Çok katmanlı algılayıcılar

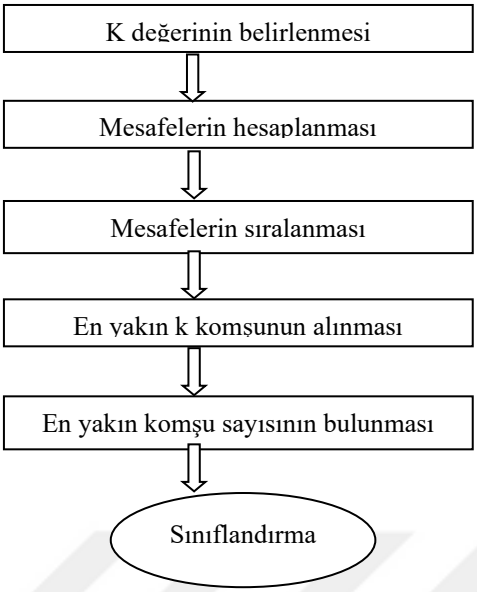
Çok katmanlı algılayıcılar ile yapılan sınıflandırmanın başarını etkileyebilecek çeşitli faktörler bulunmaktadır. Veri setinin kalitesi bu faktörlerden bir tanesidir. Sınıflandırma başarısının artırılması için bu durumu dikkat edilmesi gerekmektedir. Diğer bir faktör ise katman sayısı, bias değeri gibi parametrelerin seçimidir. Bu değerlerde çok katmanlı algılayıcılar arasında farklılıklar gözlemlenebilmektedir.

3.2.3. Naive bayes algoritması

Temelde Bayes teoremine dayanan Naive Bayes algoritması İngiliz matematikçi Thomas Bayes'den adını almaktadır. En çok kullanılan sınıflandırma algoritmalarından bir tanesidir. Olasılıksal olarak işlem yapılmakta ve veriler kategorilere göre sınıflandırılmaktadır. Sistemdeki veriler değiştikçe Naive Bayes algoritması kendini bu değişikliklere adapte edebilmektedir. Yani yeni gelen verilere duyarlılık gösterebilmektedir. Naive Bayes algoritması ile sınıflandırma yapabilmek için öncelikle sisteme öğretilmiş veri sunulmalıdır. Sunulan bu verilerin sınıfının önceden biliniyor olması gerekmektedir. Bu veriler kullanılarak olasılık temelli hesaplamalar yapılır ve sistemin, sınıfını tahmin etmesi istenen veriler sağlandığında Naive Bayes algoritması yapmış olduğu hesaplamalara göre bu verilerin sınıfını tahmin eder. Bu aşamada, tahmin edilen sınıfın doğru olma ihtimali eğitim verisinin ne kadar çok olduğu ile doğru orantılıdır. Diğer bir ifadeyle eğitim verisi ne kadar fazla ise tahmin edilen sınıfın doğru olma ihtimali o kadar yüksektir. Naive Bayes algoritması birçok sınıflandırma probleminde kullanılabilir (Kodedu, 2014).

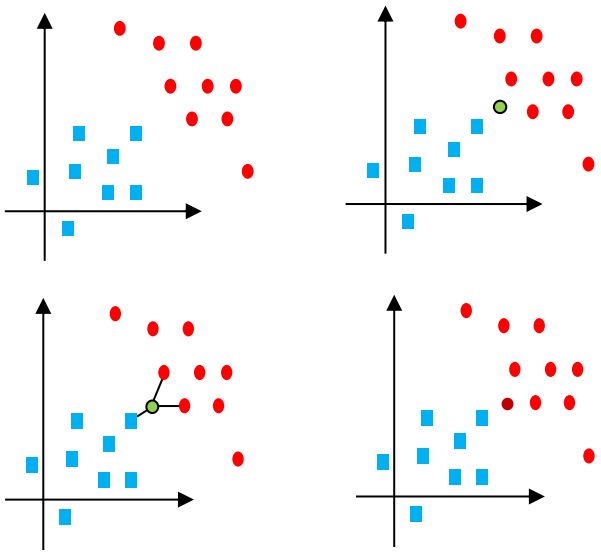
3.2.4. K-en yakın komşu algoritması

En yakın k komşu algoritması İngilizce literatürde K-Nearest Neighbor/K-NN olarak geçmektedir. Gözetimli öğrenme yöntemi olup benzerlik fonksiyonlarının kullanıldığı sınıflandırma algoritmasıdır. Buradaki k değeri sayısal bir değerdir. Sınıflandırılması istenen verinin önceki verilerden k tanesi ile benzerliğine bakılması için kullanılır. Bahse konu k değeri genelde tek sayı olarak tercih edilir. Bunun sebebi eşitlik durumunun önüne geçilmek istenmesidir (Seker, 2008). K-NN algoritması K-Means algoritması ile benzerlikler göstermektedir. Ancak temel farklılıklar bulunmaktadır. K-NN algoritması gözetimli öğrenme algoritması iken K-Means algoritması gözetimsiz öğrenme algoritmasıdır.



Şekil 3.5. K-NN algoritması adımları

K-NN algoritmasının beş adımdan oluştuğunu söylemek mümkündür. Bu adımlar yukarıdaki Şekil 3.5'de verilmiştir. İlk olarak k değeri belirlenir ve ikinci adımda öklid mesafeleri hesaplanır. Üçüncü adımda, elde edilen uzaklıklar sıralanır ve en yakın k komşu bulunur. Dördüncü adımda en yakın komşu kategorileri toplanır. Beşinci adımda en uygun komşu bulunur ve son olarak sınıflandırma işlemi gerçekleştirilir (Uzun, 2023a). Açıklaması yapılan adımları k değerini 3 olarak tanımladıktan sonra aşağıdaki örnek üzerinde inceleyelim (Seker, 2008).



Şekil 3.6. K-NN algoritması örneği

K-NN algoritmasında k değerinin aldığı değer ve eğitim veri setinin boyutu önemli bir husustur. Bu veriler artırılarak algoritmanın sınıflandırma başarısı gözlemlenir. Başarı oranı sabit seyretmeye başladığında doğru değerlere ulaşıldığı sonucu çıkarılır. K-NN algoritmasının dezavantajı yeni veri eklendiğinde uzaklık hesaplamalarının tekrar tekrar yapılmasıdır (Uzun, 2023a).

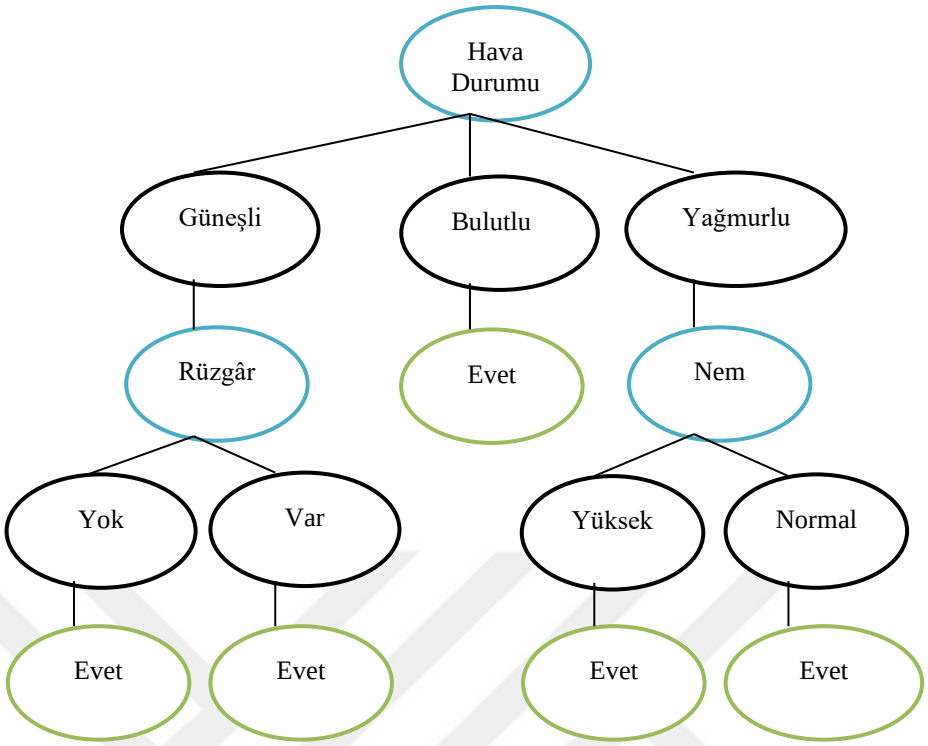
3.2.5. Karar ağacı algoritması

Karar ağacı algoritması İngilizce literatürde Decision Trees olarak geçmektedir. Sınıflandırma ve regresyon problemleri için yaygın olarak kullanılan algoritmalardan bir tanesidir. Bir ağaç yapısı formuna sahip olan model, sisteme sorular sorar ve karşılığında aldığı cevaba göre sonuca ulaşır. Bu ağaç yapısı yaprak düğümlerden ve karar düğümlerinden oluşmaktadır. Karar düğümünün birden fazla dallanmaya sahip olduğu durumlar olabilmektedir. Kök düğüm ağaçtaki ilk düğümdür. Karar ağaçları koşul yapısı (if-else) kullanılarak kodlanabilmektedir (Uzun, 2023b). Aşağıda karar ağacı örneği bulunmaktadır.

Çizelge 3.3. Karar ağacı için örnek veri tablosu

Özellikler				**Hedef**
Hava Durumu	**Sıcaklık**	**Nem**	**Rüzgâr**	**Koşu Yap**
Yağmurlu	Sıcak	Yüksek	Yok	Hayır
Yağmurlu	Sıcak	Yüksek	Var	Hayır
Bulutlu	Sıcak	Yüksek	Yok	Evet
Güneşli	Ilık	Yüksek	Yok	Evet
Güneşli	Soğuk	Normal	Yok	Evet
Güneşli	Soğuk	Normal	Var	Hayır
Bulutlu	Soğuk	Normal	Var	Evet
Yağmurlu	Ilık	Yüksek	Yok	Hayır
Yağmurlu	Soğuk	Normal	Yok	Evet
Güneşli	Ilık	Normal	Yok	Evet
Yağmurlu	Ilık	Normal	Yok	Evet
Bulutlu	Ilık	Yüksek	Var	Evet
Bulutlu	Sıcak	Normal	Yok	Evet
Güneşli	Ilık	Yüksek	Var	Hayır

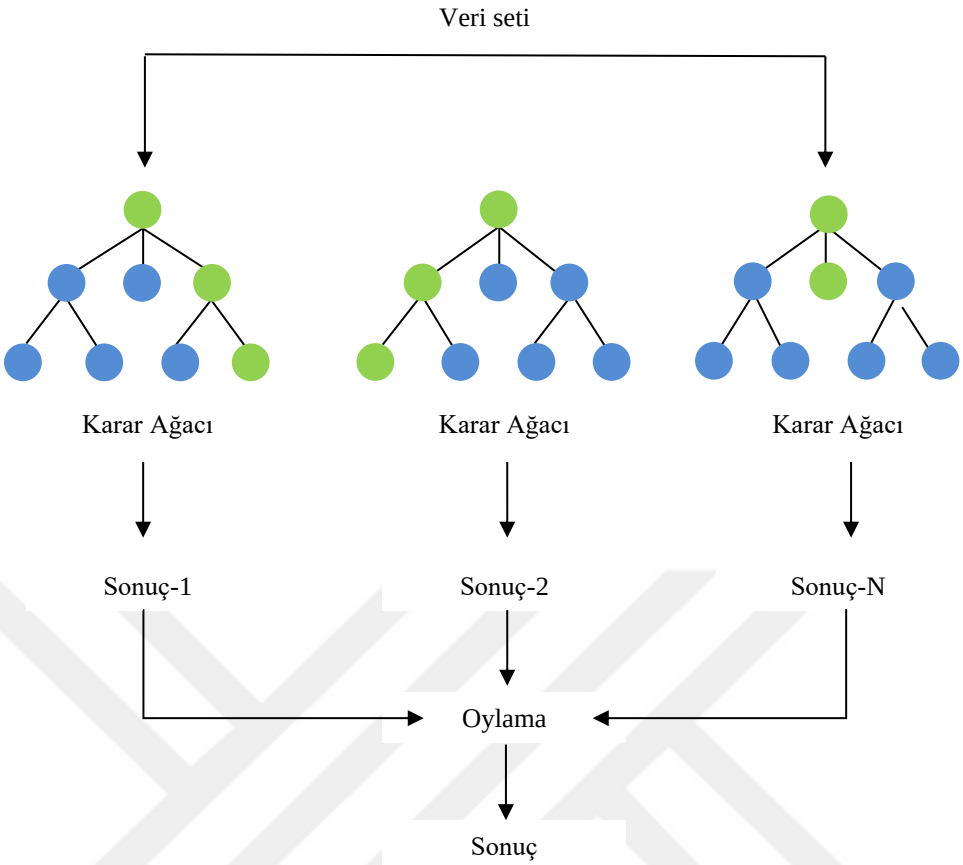
Şekil 3.7’de, Çizelge 3.3’de verilen tabloda bulunan verilerden elde edilmiş karar ağacı bulunmaktadır.



Şekil 3.7. Karar ağacı örneği

3.2.6. Rastgele orman algoritması

Rastgele orman algoritması denetimli bir sınıflandırma algoritması olup İngilizce literatürde Random Forest olarak geçmektedir. Karar ağacı algoritması gibi sınıflandırma ve regresyon problemleri için yaygın olarak kullanılan algoritmalardan bir tanesidir. Algoritmada birden fazla karar ağacı bulunmaktadır. Ağaç sayısının fazla olması elde edilen sonucun başarısıyla doğru orantılıdır. Ağaç sayısının fazla olması aşırı uyum (overfitting) probleminin de önüne geçebilmektedir. Karar ağacı algoritması ile benzerlik göstermektedir. Aralarındaki en temel fark, kök düğümün rastgele orman algoritmasında rastgele çalışıyor olmasıdır (Çebi, 2020). Algoritma temel olarak dört adımdan oluşmaktadır. Birinci adımda veri seti oluşturulur. İkinci adımda karar ağaçları oluşturulur ve bu ağaçların tahmin ettiği değerler elde edilir. Üçüncü adımda elde edilen tahmin sonuçları oylanır. Dördüncü ve son adımda ise en çok oylanan sonuç algoritmanın tahmin sonucudur (Öztürk, 2022). Şekil 3.8'de rastgele orman algoritmasının yapısı gösterilmektedir.



Şekil 3.8. Rastgele orman

3.3. Sınıflandırma Algoritmalarının Performans Ölçümü

Kullanılan algoritmalardan elde edilen sonuçların değerlendirilmesi ve birbiri arasında kıyaslanması için bazı hesaplamalar yapılmaktadır. Bu metrikler aşağıda açıklanmıştır.

3.3.1. Karmaşıklık matrisi

Karmaşıklık matrisi İngilizce literatürde Confusion Matrix olarak geçmektedir. Karmaşıklık matrisi sınıflandırma algoritmasının performansını ölçmek için kullanılan tablodur. Sınıflandırma sonucunda elde edilen sonuçlar ile gerçek sonuçlar tabloda birlikte gösterilmektedir.

		Gerçek Sınıf	
		Pozitif	Negatif
Tahmin Edilen Sınıf	Pozitif	TP	FP
	Negatif	FN	TN

Şekil 3.9. Karmaşıklık matrisi

Gerçek pozitif (True Positive – TP), tahminin pozitif olarak yapılması ve gerçekte de pozitif olması durumudur. Gerçek negatif (True Negative – TN), tahminin negatif olarak yapılması ve gerçekte de negatif olması durumudur. Yanlış pozitif (False Positive – FP), tahminin pozitif olarak yapılması ve gerçekte negatif olması durumudur. Yanlış negatif (False Negative – FN), tahminin negatif olarak yapılması ve gerçekte pozitif olması durumudur. TP ve FN değerlerinin toplanmasıyla gerçek pozitiflerin sayısına ulaşılır. TN ve FP değerlerinin toplanmasıyla da gerçek negatiflerin sayısına ulaşılır. Denklem 3.1'e göre TP, FP, FN ve TN değerlerinin toplanmasıyla da Toplam değişkeninin değerine ulaşılır. (Küçük, 2023).

$$Toplam = TP + FP + FN + TN \tag{3.1}$$

3.3.2. Doğruluk

Doğruluk İngilizce literatürde Accuracy olarak geçmektedir. Doğru yapılan sınıflandırma sayısının (TN+TP) Toplam değerine oranı doğruluğu verir. Örneğin pozitifin pozitif (TP) olarak tahmin edildiği değer 68, negatifin negatif (TN) olarak tahmin edildiği değer ise 24 olsun. FN ve FP değerlerinin ise sırasıyla 15 ve 29 olduğunu kabul edelim. Bu durumda Denklem 3.2'e (Patro ve Patra, 2014) göre Accuracy 67.5 olarak hesaplanmış olur.

$$accuracy = (TP + TN)/(TP + TN + FN + FP) \tag{3.2}$$

3.3.3. Kesinlik

Kesinlik İngilizce literatürde Precision olarak geçmektedir. Model tarafından pozitif olarak tahmin edilen değerlerin başarısını değerlendirmede kullanılır. TP değerinin TP ve FP değerlerinin toplamına oranı Precision değerini verir. Kısaca, tahmin edilen tüm pozitiflerin, yüzde kaçının gerçekten pozitif olduğunun göstergesidir.

3.3.4. Duyarlılık

Duyarlılık İngilizce literatürde Recall olarak geçmektedir. Model tarafından pozitif olarak tahmin edilmesi gerekenlerin ne kadarının pozitif olarak tahmin edildiğinin göstergesidir. Recall, TP değerinin, TP ve FN değerlerinin toplamına oranlanmasıyla bulunur.

3.3.5. F1-skor

F1-skor İngilizce literatürde Recall olarak geçmektedir. F1-skor Recall ve Precision değerlerinin harmonik ortalaması olarak tanımlanmaktadır. F1-skor değerinin harmonik ortalama olması, tüm durumların göz önünde bulundurduğunu göstermektedir. Hem FP hem de FN değerlerini hesaba katar. Bu nedenle, dengesiz veri kümelerinin sınıflandırma performansı ölçümünde yaygın olarak kullanılan bir metriktir.

3.3.6. Min-Max normalizasyonu

Birbirinden farklı skalada bulunan kavramların belirli bir formda bir araya getirilmesi işlemine normalizasyon adı verilmektedir. Diğer bir ifadeyle verilerin birbirleriyle aynı aralıkta ifade edilmesidir. Normalizasyon işlemi için kullanılmakta olan çeşitli yöntemler mevcuttur. Normalizasyon yöntemlerinin çoğu [0-1] aralığına indirgemeye çalışmaktadır. Bu yöntemlerden bir tanesi İngilizce literatürde feature scaling olarak da bilinen min-max normalizasyon yöntemidir. Minimum ve maksimum değeri bilinen bir aralıkta bulunan sayılar Denklem 3.3 (Henderi ve ark., 2021) kullanılarak [0-1] aralığında normalize edilebilmektedir.

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)} \tag{3.3}$$

Denklem 3.3'de yer alan x' normalize edilmiş değeri, x normalize edilecek değeri, min(x) en küçük x değerini ve max(x) ise en büyük x değerini ifade etmektedir.

4. ARAŞTIRMA SONUÇLARI VE TARTIŞMA

İnternet ortamına erişen kişi sayısı artıkça üretilen verinin boyutu da aynı oranda artmaktadır. Bu durum verinin işlenmesi ve bilgisayar tarafından otomatik olarak sınıflandırılması ihtiyacını beraberinde getirmiştir. Doğal dil işlemenin çalışma alanlarından biri de metin sınıflandırma uygulamalarıdır. Metin sınıflandırmanın bilgisayar tarafından otomatik olarak yapılması internet kullanıcıları açısından veriye erişimde kolaylık sağlamaktadır. Bu alanda daha önce yapılan çalışmalara ikinci bölümde yer verilmiştir. Bu çalışmada ise farklı kategorilerde yazılan şiirlerin konusuna göre (Bayrak, şehit vb.) sınıflandırılması üzerine bir çalışma yapılmıştır. Bu sınıflandırmanın yapılabilmesi için Support Vector Machines, Multi Layer Perceptrons, Naive Bayes, K-NN, Decision Trees ve Random Forest olmak üzere altı farklı sınıflandırma algoritması kullanılmıştır. Öznitelik çıkarımı için TF-IDF yöntemi kullanılmıştır. Sınıflandırma algoritmalarından elde edilen sonuçlar karşılaştırılmıştır.

Çalışma kapsamında 4200 adet kayıttan oluşan ve web kazıma yöntemi ile elde edilmiş veri seti kullanılmıştır. Veri setinde şiir başlığı, şair adı, kategorisi, link ve şiir metninden oluşan 5 adet sütun bulunmaktadır. Toplamda 4198 adet kayıttan oluşan veri seti kullanılmıştır. Veri setinde 21 adet kategori bulunmaktadır.

Veri seti toplam 594.130 adet kelimeden oluşmaktadır. Benzersiz kelime sayısı ise 100.812 olarak tespit edilmiştir.

Veri seti elde edildikten sonra çeşitli veri önişleme işlemlerine tabii tutulmuştur. Bu aşamada yapılan işlemler şunlardır:

a. Satır boşlukları kaldırılmış, metin küçük harfe çevrilmiş ve noktalama işaretleri kaldırılmıştır. Satır boşlukları metinin düzeni açısından önem taşısa da doğal dil işlemede aynı önemi taşımamaktadır. Satır boşluklarının kaldırılması, modelin yüksek doğruluk oranına sahip sonuçlar verebilmesi için yapılmıştır. Büyük yazılmış harflerin küçük harflere dönüştürülmesi doğal dil işleme aşamasında yapılan işlemlerden bir tanesidir. Veri setinde kullanılan büyük harfler küçük harflere dönüştürülmüştür.

b. Veri setinde bulunan etkisiz kelimeler (Stopwords) kaldırılmıştır. Sınıflandırma başarısını olumsuz etkileyebilen bu kelimeler veri setinden çıkarılmıştır. Etkisiz kelimeler sık kullanılan “ama”, “bazı”, “belki” gibi kelimelerdir. Bu kelimeler metinden çıkarıldığında daha etkili analizler yapılabilmekte ve doğal dil işleme sürecine olumlu katkı sağlamaktadır. Bu kelimelerin veri setinden çıkarılması için Zemberek

Kütüphanesinin sunmuş olduğu etkisiz kelimeler listesi kullanılmıştır. Listede daha önceden tanımlanan etkisiz kelimeler yer almakta ve veri setinde bulunan kelimeler ile kıyaslanmaktadır. Eşleşen kelimler olması durumunda bu kelimeler veri setinden çıkarılmaktadır.

c. Veri setinde bulunan cümleler kelimelere ayrılmıştır. Kelimeler cümlelerin en anlamlı parçalarıdır.

d. Veri setinde bulunan şiirler cümlelere ayrılmıştır.

e. Veri setinde kullanılan yabancı kelimeler tespit edilmiştir.

f. Kelimelerin dağılımları tespit edilmiştir.

g. Kelimelerin harf olarak uzunluk dağılımları hesaplanmıştır.

h. Cümlelerin kelime olarak uzunluk dağılımları hesaplanmıştır.

i. Kelimelerin toplam kelime sayısına oranı hesaplanmıştır.

j. Ortalama kelime uzunluğu ile birlikte ortalama cümle uzunluğu tespit hesaplanmıştır.

k. Noktalama işareti sayısı, kullanılan stopwords sayısı ve tamamı büyük harf ile yazılan kelime sayısı çıkarılmıştır.

Veri önişleme adımları gerçekleştirildikten sonra metin temsili için önemli olan kök çözümleme aşamasına geçilmiştir. Bu işlem için farklı algoritmalar (ITU Türkçe Doğal Dil İşleme Yazılım Zinciri, NLTK, Spacy vb.) bulunmakla birlikte yapılan çalışmada açık kaynak olan Zemberek Kütüphanesi kullanılmıştır. Ek almış kelimelerin eklerinin tespit edilerek köklerinin bulunması sağlanmıştır. Bu işlemin amacı, ek almış kelimelerin bilgisayar tarafından farklı kelimeler olarak algılanmasının önüne geçmektir. Örneğin “ilişkilendiremediklerimiz” kelimesinin kökünün “ilişki” olduğu bu kütüphane sayesinde bulunabilmektedir. Bilgisayar tarafından farklı kelimeler olarak algılanmasının sebebi harf bazında karşılaştırma yapılmasıdır. Kelimelerin metin içerisinde ne kadar geçtiklerine bakılarak frekansları hesaplanmaktadır. Aynı köke sahip kelimeler tespit edilmediği zaman kelimelerin frekansının çıkarılması adımında temsil kabiliyetini düşürmektedir.

	Başlık	Şair	Kelimeler	Cümleler	Kökler	Kök Dağılımı	Kelime Uzunluk Dağılımı	Cümle Uzunluk Dağılımı
0	Bayram	Mehmet Akif Ersoy	[âfâk, bütün, hande, cihan, cihandır, bayram, ...	[âfâk bütün hande cihan cihandır, bayram kadar...	[afak, bütün, hande, cihan, cihan, bayram, şet...	{'mürüvvet':1, 'salla':1, 'al':1, 'suret':...	{0: 0, 1: 3, 2: 14, 3: 29, 4: 45, 5: 119, 6: 6...	{0: 1, 1: 7, 2: 17, 3: 20, 4: 27, 5: 22, 6: 12...
1	Bayram	Necip Fazıl Kısakürek	[ölene, bayram, bayrama, sevinmek, var, ne, ba...	[ölene bayram bayrama sevinmek var, ne bayramd...	[öl, bayram, bayram, sevin, var, ne, bayram, t...	{'ata':1, 'bin':1, 'tahta':1, 'bayram':3, ...	{0: 0, 1: 0, 2: 1, 3: 3, 4: 0, 5: 2, 6: 2, 7:...	{0: 0, 1: 0, 2: 0, 3: 0, 4: 0, 5: 2, 6: 0, 7:...
2	Süleymaniye'de Bayram Sabahı	Yahya Kemal Beyatlı	[artarak, gönlümün, aydınlığı, saniyede, mehâb...	[artarak gönlümün aydınlığı saniyede, mehâbetl...	[art, gönül, aydın, saniye, mehabet, ol, süley...	{'ses':1, 'bit':1, 'halk':1, 'beri':1, 'za...	{0: 0, 1: 0, 2: 4, 3: 7, 4: 9, 5: 27, 6: 18, 7...	{0: 0, 1: 0, 2: 4, 3: 6, 4: 9, 5: 6, 6: 3, 7:...
3	Bayram Duası	Ozan Arif	[rabbi, tadına, bütün, milletin, varacağı, bay...	[rabbi tadına bütün milletin, varacağı bayraml...	[rabbi, tat, bütün, millet, var, bayram, eriş...]	{'zekat':1, 'al':1, 'rab':1, 'enflasyon':1...	{0: 0, 1: 0, 2: 4, 3: 4, 4: 7, 5: 26, 6: 12, 7...	{0: 0, 1: 0, 2: 2, 3: 21, 4: 11, 5: 2, 6: 0, 7...

Şekil 4.1. Metin üzerinde ön işleme yapılması sonucunda oluşan veri seti (a)

	Başlık	Şair	Kelime Zenginliği	Kök Zenginliği	Ort. Kelime Uzunluğu	Ort. Cümle Uzunluğu	Noktalama İşareti Sayısı	Stopwords Sayısı	Büyük Yazılmış Kelime Sayısı
0	Bayram	Mehmet Akif Ersoy	0.908046	0.652874	6.050575	3.750000	223	142	1
1	Bayram	Necip Fazıl Kısakürek	0.909091	0.727273	5.090909	5.000000	3	4	0
2	Süleymaniye'de Bayram Sabahı	Yahya Kemal Beyatlı	0.958333	0.808333	6.508333	3.928571	62	49	1
3	Bayram Duası	Ozan Arif	0.826772	0.748031	6.692913	3.361111	25	29	0

Şekil 4.1. Metin üzerinde ön işleme yapılması sonucunda oluşan veri seti (b)

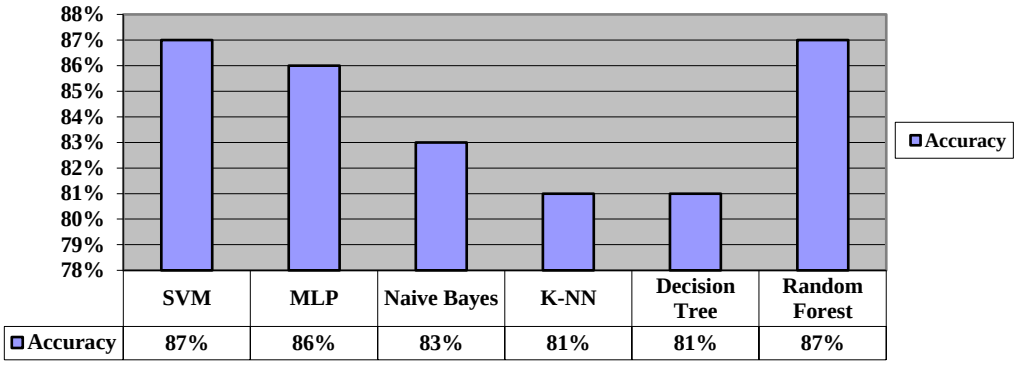
Şekil 4.1'de, veri setinde bulunan metinlerin, Zemberek Kütüphanesi kullanılarak çeşitli işlemlere tabii tutulmuş hali bulunmaktadır. Şekil 4.1 (a ve b)'de birinci sütunda şiirin başlığı ikinci sütunda ise şairin adı bulunmaktadır. Kelimeler sütununda; metnin kelimelere ayrılmış hali bulunmaktadır. Cümleler sütununda; metnin

cümlelerine ayrılmış hali bulunmaktadır. Kökler sütununda; kelime köklerinin bulunmuş hali yer almaktadır. Kök Dağılımı sütununda; Kökler sütununda bulunan kelimelerin harf olarak uzunluk dağılımları yer almaktadır. Kelime Uzunluk Dağılımı sütununda; Kelimeler sütununda bulunan kelimelerin harf olarak uzunluk dağılımları yer almaktadır. Cümle Uzunluk Dağılımı sütununda; cümlelerin kelime olarak uzunluk dağılımları yer almaktadır. Kelime Zenginliği sütununda; Kelimeler sütununda bulunan kelimelerin toplam kelime sayısına oranı yer almaktadır. Kök Zenginliği sütununda; Kökler sütununda bulunan kelimelerin toplam kelime sayısına oranı yer almaktadır. Ort. Kelime Uzunluğu sütununda; Kelimeler sütununda yer alan kelimelerin ortalama kelime uzunluğu yer almaktadır. Ort. Cümle Uzunluğu sütununda; Cümleler sütununda yer alan cümlelerin kelime olarak ortalama cümle uzunluğu yer almaktadır. Noktalama İşareti Sayısı sütununda; metinde bulunan noktalama işareti sayısı yer almaktadır. Stopwords Sayısı sütununda; metinde bulunan etkisiz kelime sayısı yer almaktadır. Büyük Yazılmış Kelime Sayısı sütununda ise metinde bulunan tamamı büyük harf olan kelime sayısı yer almaktadır.

Metnin istenen formata dönüştürülmesinin ardından kelimelerin metni ne kadar temsil ettiğinin bulunması gerekmektedir. Bu temsilin bulunabilmesi için literatürde TF-IDF olarak geçen “Terim Sıklığı-Ters Doküman Sıklığı” kullanılmıştır. TF-IDF Denklem 4.1’e (Trstenjak ve ark., 2014) göre hesaplanmaktadır. Denklemde yer alan $w_{x,y}$ değeri hesaplanan TF-IDF’i, $tf_{x,y}$ x’in y’de kaç defa geçtiğini, df_x kaç adet dokümanda x’in geçtiğini ve N ise toplam doküman sayısını ifade etmektedir. Bu değerlerin hesaplanmasında Python kütüphanesi olan Scikit Learn kullanılmıştır. Yüksek TF-IDF değerine sahip olan kelimenin ilgili metni daha çok temsil ettiği, düşük TF-IDF değerine sahip kelimenin ise ilgili metni temsil açısından az temsil ettiği kabul edilmektedir.

$$w_{x,y} = tf_{x,y} * \log\left(\frac{N}{df_x}\right) \tag{4.1}$$

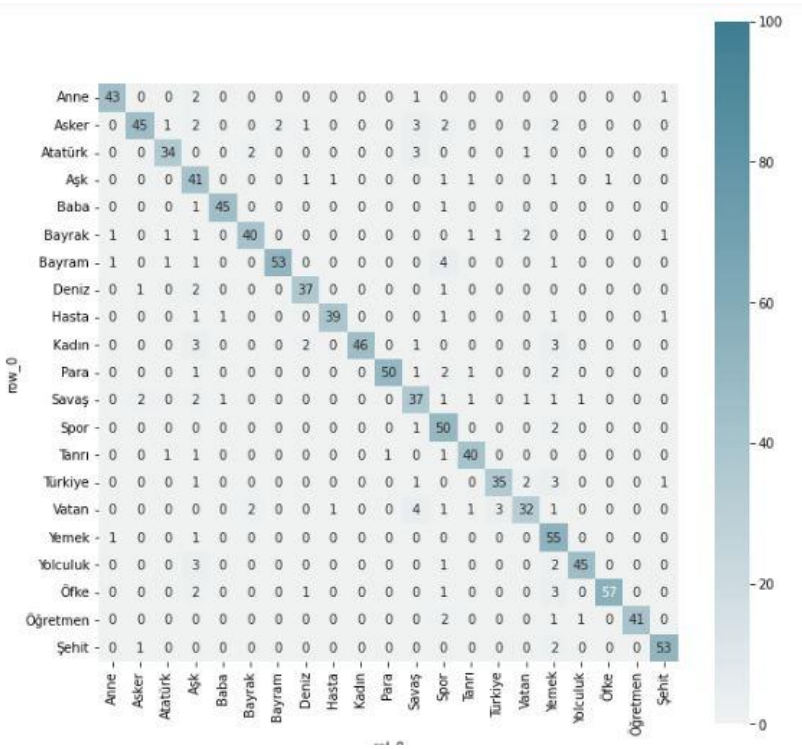
Sınıflandırma algoritmalarının doğruluk oranlarının gösterildiği Şekil 4.2 aşağıda bulunmaktadır.



Şekil 4.2. Sınıflandırma algoritmaların doğruluk (accuracy) sonuçları

Sınıflandırma algoritmaları kıyaslandığında en yüksek doğruluk oranını Random Forest ve SVM algoritmalarının verdiği görülmektedir. Bunu, %86 doğruluk oranı ile MLP ve %83 doğruluk oranı ile Naive Bayes algoritmaları takip etmektedir. En düşük doğruluk oranı %81 ile K-NN ve Decision Tree algoritmalarından elde edilmiştir.

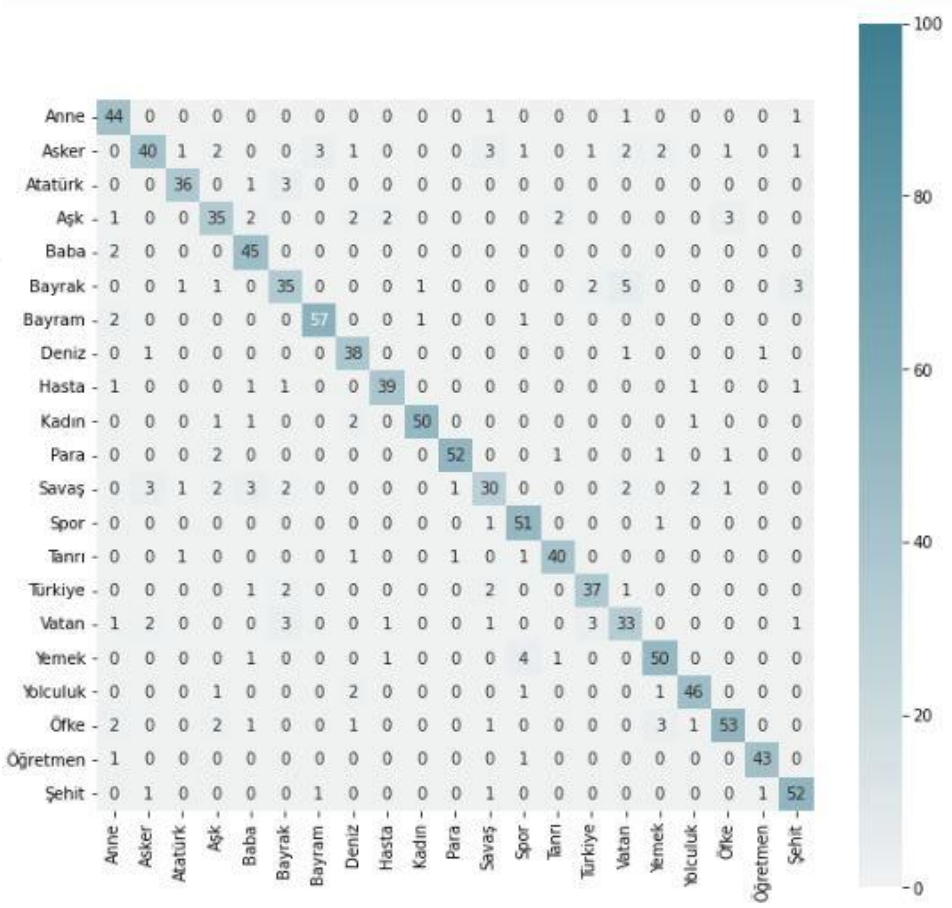
Verilerin %80'i eğitim, %20'i test olmak üzere ikiye ayrılmıştır. Scikit-learn kütüphanesindeki SVM test edilerek %87 doğruluk oranına ulaşılmıştır. Elde edilen öznitelikler Min-Max Normalization yöntemi ile normalize edildikten sonra Scikit-Learn kütüphanesinde bulunan varsayılan parametrelerdeki (C=1.0, kernel='rbf', degree=3, gamma='scale') SVM modelinin sonuçları incelenmiştir.



Şekil 4.3. SVM sınıflandırma algoritması kullanılarak elde edilen karmaşıklık matrisi

Şekil 4.3'de verilen karmaşıklık matrisi incelendiğinde, "Anne" kategorisindeki 46 adet şiirden yanlışlıkla birer tane "Yemek", "Bayram" ve "Bayrak" kategorisinde tahmin edilmiştir. 43 tane şiirin ise doğru tahmin edildiği görülmektedir. SVM algoritması kullanıldığında en doğru tahmin "Kadın" kategorisine ait şiirlerde gerçekleşmiştir. Bu sınıfa ait şiirlerin tamamı doğru tahmin edilmiştir. Bu kategorideki şiirler daha belirgin kelimeler içerdiğinden daha doğru tahminler elde edilmiştir. "Aşk" ve "Savaş" kategorilerine ait şiirlerin doğru sınıflandırma oranı düşük olduğundan başarıyı olumsuz etkilediği görülmektedir.

MLP sınıflandırma algoritması kullanılarak %86 doğruluk oranına ulaşılmıştır. SVM sınıflandırıcının MLP sınıflandırıcıya göre daha yüksek doğruluk oranına sahip olduğu gözlemlenmiştir.

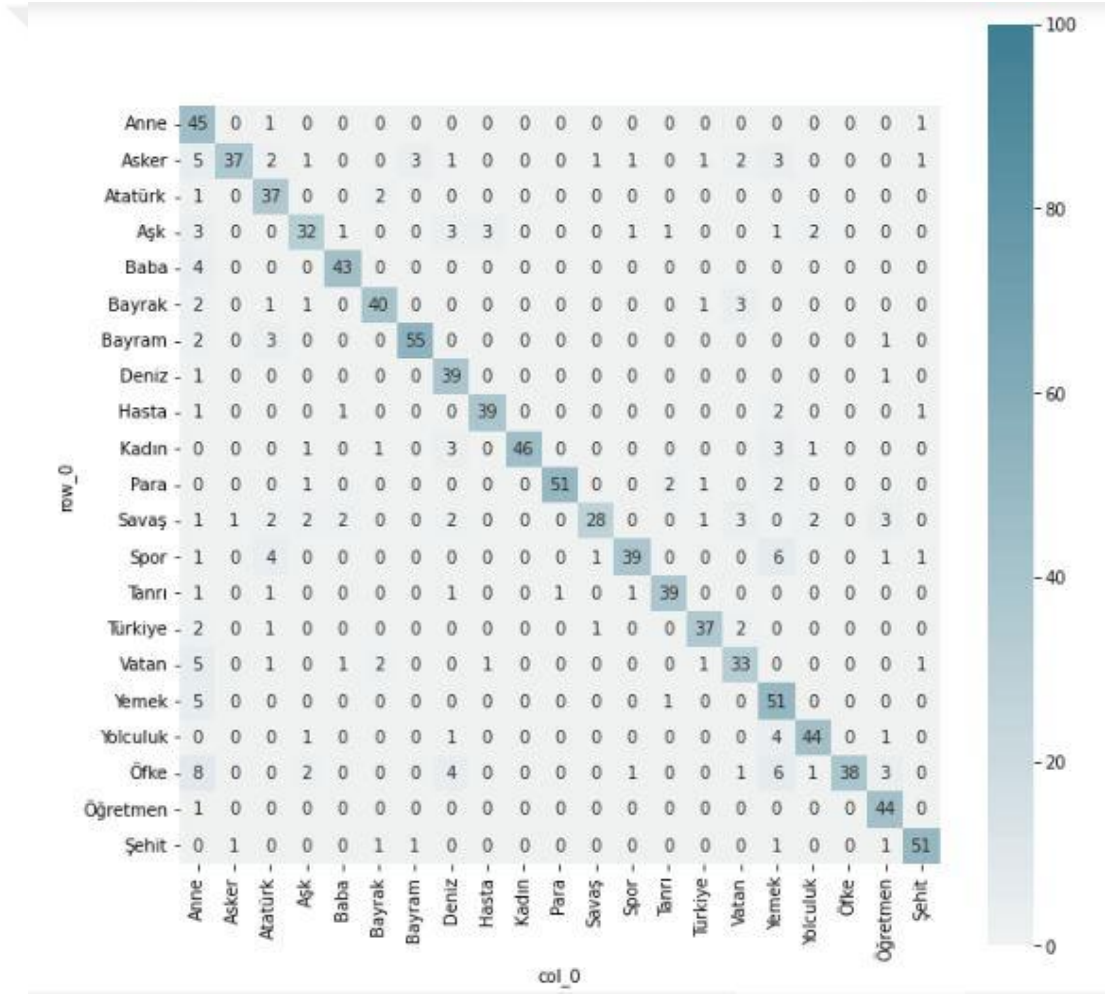


Şekil 4.4. MLP sınıflandırma algoritması kullanılarak elde edilen karmaşıklık matrisi

Şekil 4.4'de verilen karmaşıklık matrisi incelendiğinde, "Anne" kategorisindeki 54 adet şiirden yanlışlıkla birer tane "Öğretmen", "Vatan", "Aşk" ve "Hasta"

kategorisinde ikişer tane şiir ise “Öfke”, “Baba” ve “Bayram” kategorisinde tahmin edilmiştir. 44 tane şiirin ise doğru tahmin edildiği görülmektedir. MLP algoritması kullanıldığında en doğru tahmin “Bayram” kategorisine ait şiirlerde gerçekleşmiştir. Bu sınıfa ait 61 adet şiirin 57’si doğru tahmin edilmiştir. Bu kategorideki şiirler daha belirgin kelimeler içerdiğinden daha doğru tahminler elde edilmiştir. “Vatan” kategorisine ait şiirlerin doğru sınıflandırma oranı düşük olduğundan başarıyı olumsuz etkilediği görülmektedir.

Naive Bayes algoritması ile sınıflandırma yapıldığında ise %83 doğruluk oranına ulaşılmıştır. Naive Bayes algoritması kullanıldığında SVM ve MLP algoritmalarına kıyasla daha düşük doğruluk oranına sahip sonuç elde edilmiştir.

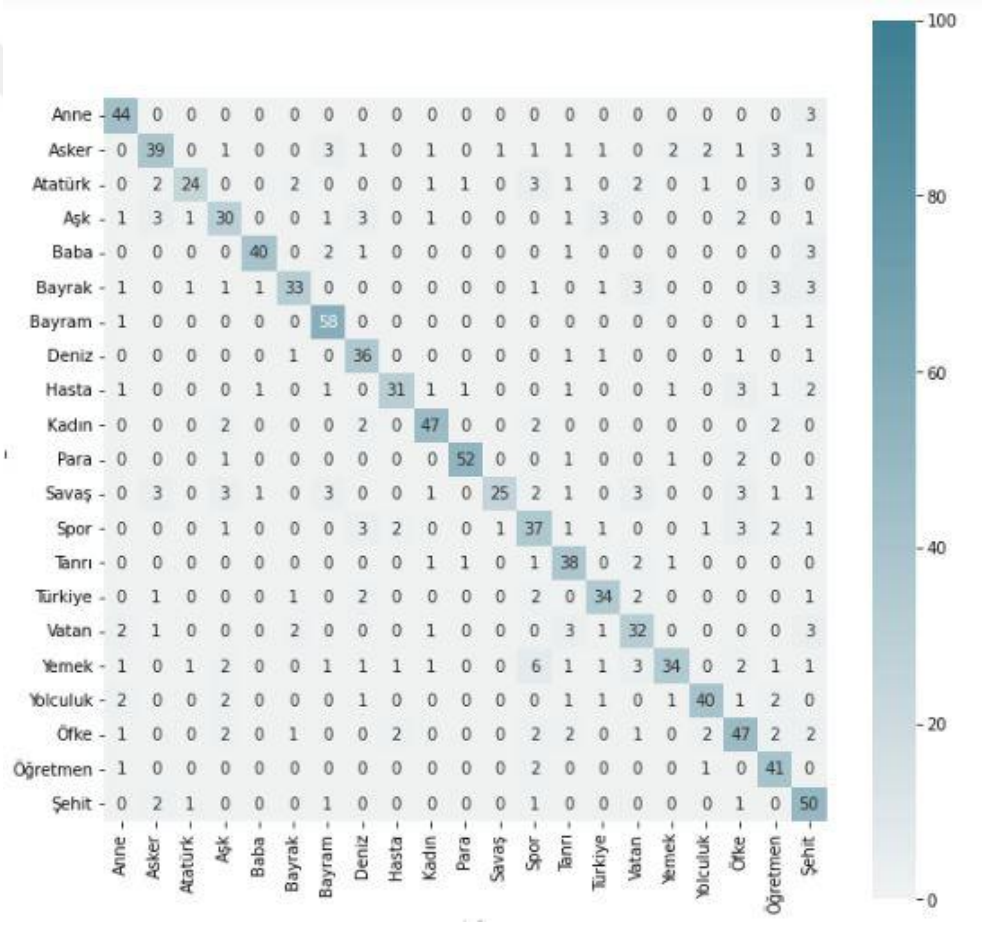


Şekil 4.5. Naive Bayes sınıflandırma algoritması kullanılarak elde edilen karmaşıklık matrisi

Şekil 4.5’de verilen karmaşıklık matrisi incelendiğinde, “Asker” kategorisindeki 39 adet şiirden yanlışlıkla birer tane şiir “Şehit”, ve “Savaş” kategorisinde tahmin edilmiştir. 37 tane şiirin ise doğru tahmin edildiği görülmektedir. Naive Bayes

algoritması kullanıldığında en doğru tahmin "Kadın" kategorisine ait şiirlerde gerçekleşmiştir. Bu sınıfa şiirlerin tamamı doğru tahmin edilmiştir. Bu kategorideki şiirler daha belirgin kelimeler içerdiğinden daha doğru tahminler elde edilmiştir. "Anne" kategorisine ait şiirlerin doğru sınıflandırma oranı düşük olduğundan başarıyı olumsuz etkilediği görülmektedir.

K-NN ile sınıflandırma yapıldığında ise K değeri ilk aşamada 3 olarak seçildiğinde yapılan sınıflandırmada %77 doğruluk oranına ulaşılmıştır. K değerinin 5 olarak seçilmesi sonucunda yapılan sınıflandırmada %80 doğruluk oranı elde edilmiştir. K değerinin 7 olarak seçilmesi sonucunda yapılan sınıflandırmada %81 doğruluk oranı elde edilmiştir.

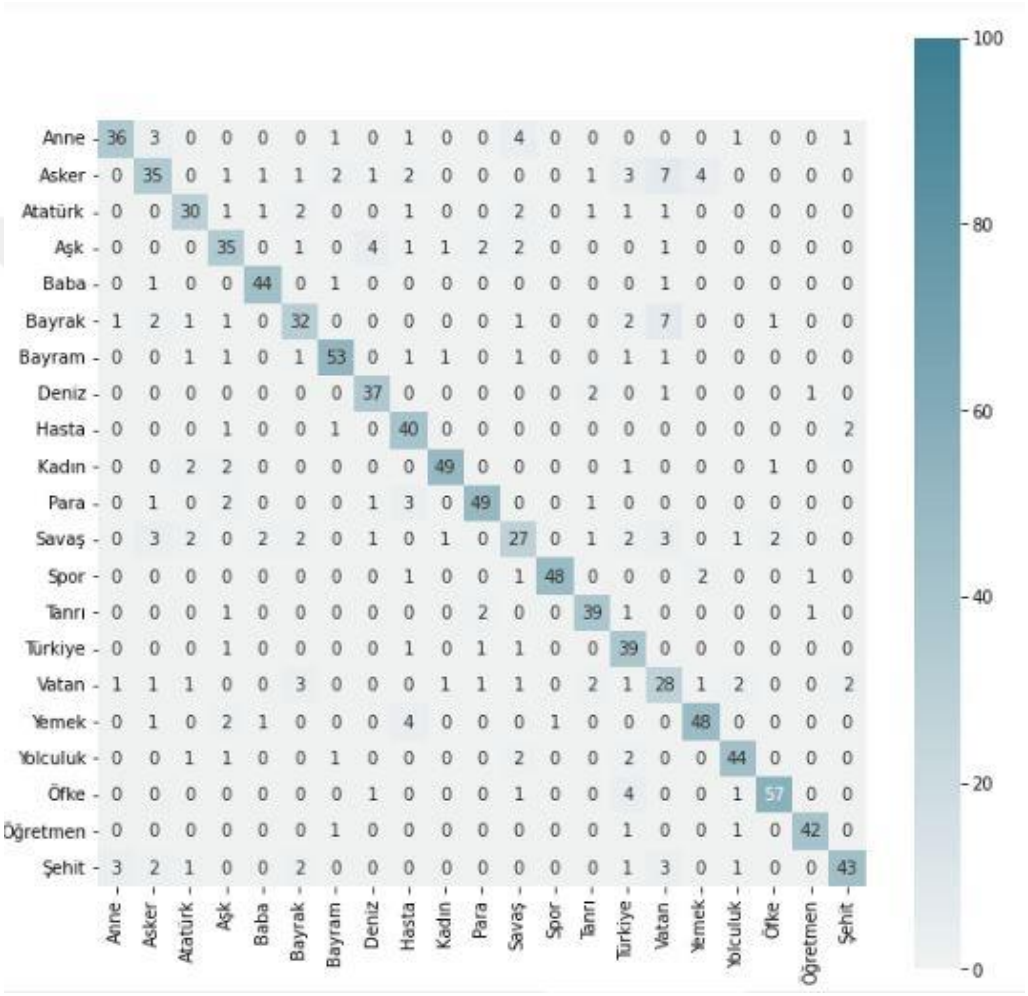


Şekil 4.6. K-NN sınıflandırma algoritması kullanılarak k=7 için elde edilen karmaşıklık matrisi

Şekil 4.6'de verilen karmaşıklık matrisi incelendiğinde, "Savaş" kategorisindeki 27 adet şiirden yanlışlıkla birer tane şiir "Spor", ve "Asker" kategorisinde tahmin edilmiştir. 25 tane şiirin ise doğru tahmin edildiği görülmektedir. K-NN algoritması kullanıldığında en doğru tahmin "Baba" kategorisine ait şiirlerde gerçekleşmiştir. Bu

sınıfa şiirlerin tamamı doğru tahmin edilmiştir. Bu kategorideki şiirler daha belirgin kelimeler içerdiğinden daha doğru tahminler elde edilmiştir. “Şehit” kategorisine ait şiirlerin doğru sınıflandırma oranı düşük olduğundan başarıyı olumsuz etkilediği görülmektedir.

Decision Tree algoritması ile sınıflandırma yapıldığında ise %81 doğruluk oranına ulaşılmıştır.

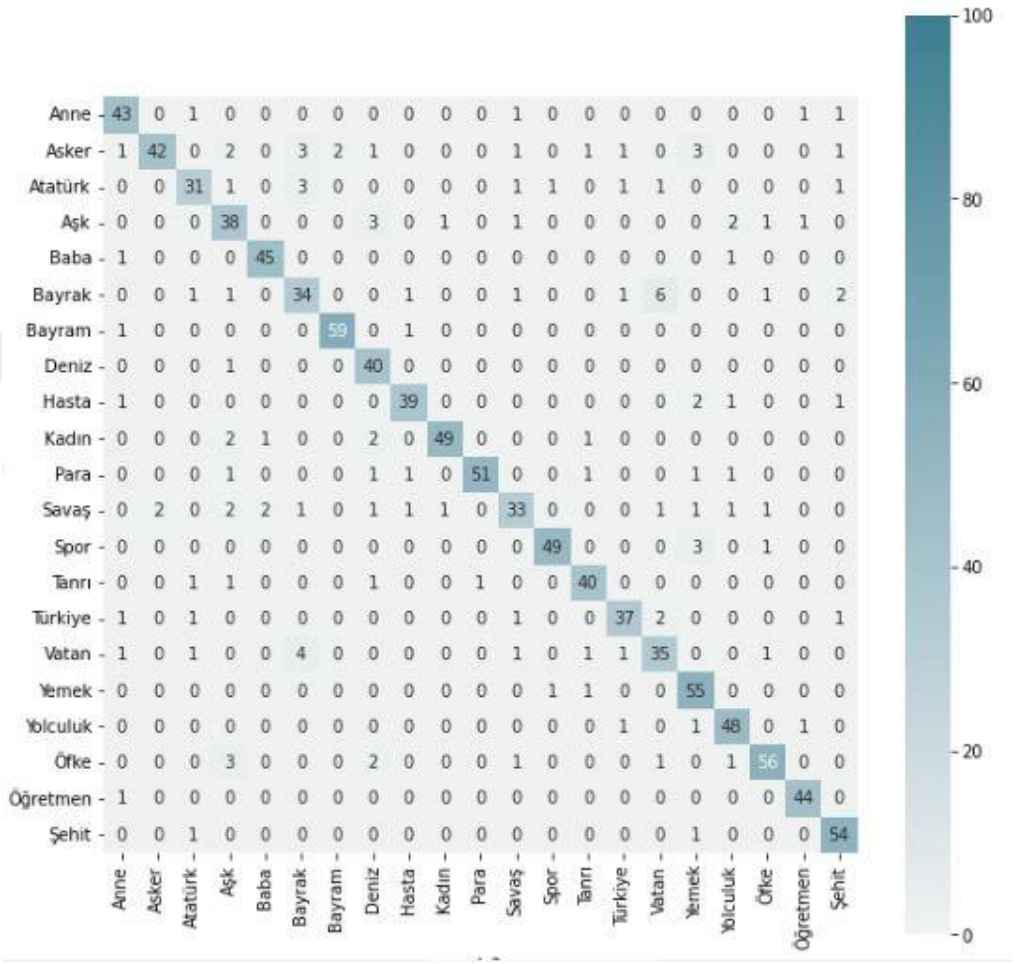


Şekil 4.7. Decision Tree sınıflandırma algoritması kullanılarak elde edilen karmaşıklık matrisi

Şekil 4.7’de verilen karmaşıklık matrisi incelendiğinde, “Anne” kategorisindeki 41 adet şiirden yanlışlıkla birer tane şiir “Vatan”, ve “Bayrak” 3 adet şiir ise ”Şehit” kategorisinde tahmin edilmiştir. 36 tane şiirin ise doğru tahmin edildiği görülmektedir. Decision Tree algoritması kullanıldığında en doğru tahmin “Spor” kategorisine ait şiirlerde gerçekleşmiştir. Bu sınıfa şiirlerin tamamı doğru tahmin edilmiştir. Bu kategorideki şiirler daha belirgin kelimeler içerdiğinden daha doğru tahminler elde

edilmiştir. “Vatan” kategorisine ait şiirlerin doğru sınıflandırma oranı düşük olduğundan başarıyı olumsuz etkilediği görülmektedir.

Random Forest algoritması ile sınıflandırma yapıldığında ise %87 doğruluk oranına ulaşılmıştır.



Şekil 4.8. Random Forest sınıflandırma algoritması kullanılarak elde edilen karmaşıklık matrisi

Şekil 4.8’de verilen karmaşıklık matrisi incelendiğinde, “Anne” kategorisindeki 50 adet şiirden yanlışlıkla birer tane şiir “Öğretmen”, “Vatan”, “Türkiye”, “Hasta”, “Bayram”, “Baba” ve “Asker” kategorisinde tahmin edilmiştir. 43 tane şiirin ise doğru tahmin edildiği görülmektedir. Random Forest algoritması kullanıldığında en doğru tahmin “Para” kategorisine ait şiirlerde gerçekleşmiştir. Bu sınıfa 52 adet şiirin 51 tanesi doğru tahmin edilmiştir. Bu kategorideki şiirler daha belirgin kelimeler içerdiğinden daha doğru tahminler elde edilmiştir. “Vatan” ve “Bayrak” kategorilerine ait şiirlerin doğru sınıflandırma oranı düşük olduğundan başarıyı olumsuz etkilediği görülmektedir.

Sınıflandırma algoritmaları kullanılarak varsayılan parametreler ile sınıflandırma yapıldığında elde edilen sonuçların kıyaslanabilmesi ve daha iyi sonuçlar elde edilebilmesi için kullanılan yöntemlerden bir tanesi hiperparametre analizidir. Bu tez çalışmasında sınıflandırma algoritmalarının doğruluk oranlarının artırılması için hiperparametre analizi çalışması yapılmıştır. Hiperparametre analizinin amacı sınıflandırma algoritmasının performansını artırmaktır. Hiperparametre analizi ile birlikte varsayılan parametreler en iyi sonucu verecek şekilde değiştirilebilmektedir. Hiperparametre analizi için kullanılan yöntemler GridSearchCV ve RandomizedSearchCV'dir. GridSearchCV yöntemi, belirlenen parametreler ile modelleri ayrı ayrı test eder. Bu işlem sonucunda en iyi sonucu veren parametreler belirlenmiş olur. Tüm olasılıklar denendiği için bulunan parametrelerin en iyi sonucu verdiğinden emin olunabilir. Bu yöntem küçük veri setleri ile çok iyi sonuçlar verebilir. Büyük veri setleri ile çalışılmak istendiğinde olasılık sayısına bağlı olarak çok uzun işlem süresi ile karşılaşılabilmektedir. RandomizedSearchCV yönteminde, hiperparametreler rastgele seçilir. Seçilen bu parametreler ile kurulan modeller iterasyon sayısı boyunca test edilir. Büyük veri setlerinde daha az maliyetle iyi sonuçlar verebilir. Parametreler rastgele denendiği için en iyi sonuç garanti edilmez. Hiperparametre ayarlamasının yapılabilmesi için mevcut modelin belirli bir başarıyı sunması gerekmektedir. Başarının çok düşük olduğu durumlarda, bu işlemin model başarısını çok yüksek seviyeye çıkarması beklenemez.

Bu tez çalışmasında kullanılan yöntemlerden Rastgele Orman algoritması için hiperparametre analizi yapılmıştır. Bunun için RandomizedSearchCV yöntemi kullanılmıştır. Random Forest algoritmasında kullanılan "n_estimators" ve "max_depth" değişkenleri değiştirilerek en iyi sonuç bulunmaya çalışılmıştır. "n_estimators" değişkeni algoritmada kullanılan ağaç sayısını ifade eder. Bu hiperparametrenin artırılması genellikle modelin performansını artırır ancak aynı zamanda eğitim ve tahminin hesaplama maliyetini de artırır. "max_depth" değişkeni ise algoritmadaki maksimum derinliği ifade eder. Bu değerin yüksek ayarlanması overfitting (aşırı öğrenme) problemine yol açarken düşük ayarlanması ise underfitting (eksik öğrenme) problemine yol açar. Hiperparametre analizi sonucunda en iyi "max_depth" değeri 17 ve en iyi "n_estimators" değeri ise 254 olarak bulunmuştur. Parametrelerin bu şekilde ayarlanması sonucunda doğruluk oranı %87'den %89'a çıkmıştır.

Çalışmada kullanılan diğer bir algoritma SVM için de hiperparametre analizi yapılmıştır. Bunun için GridSearchCV yöntemi kullanılmıştır. SVM algoritmasında kullanılan “C”, “gamma” ve “kernel” değişkenleri değiştirilerek en iyi sonuç bulunmaya çalışılmıştır. Bu değişkenler sırasıyla varsayılan “1.0”, “scale” ve “rbf” değerlerine sahiptir. GridSearchCV yöntemi ile bu değerlerin farklı kombinasyonları ile model 125 defa eğitilmiştir. “C” ve “gamma” değişkenleri için beşer adet parametre tanımlanmış, “kernel” değişkeni için ise bir adet parametre tanımlanmıştır. Buradan ortaya çıkan 25 adet kombinasyon için 5 kat çapraz doğrulama yapıldığında model toplamda 125 kez test edilerek en iyi parametreler bulunmaya çalışılmıştır. Hiperparametre analizinde kullanılan kombinasyonların bir kısmı Çizelge 4.1’de verilmiştir.

C	Gamma	Kernel	Score	Eğitim Süresi
0.1	1	rbf	0.679	8.1min
0.1	0.01	rbf	0.051	8.2min
1	0.1	rbf	0.812	5.1min
10	0.001	rbf	0.402	8.2min
100	1	rbf	0.858	6.3min
1000	0.001	rbf	0.865	3.9min
1000	0.0001	rbf	0.839	5.0min
1000	0.001	rbf	0.875	4.0min

Çizelge 4.1. SVM algoritmasının hiperparametre analizinde kullanılan parametre örnekleri

Hiperparametre analizi sonucunda “C” parametresi için 1000, “gamma” parametresi için 0.001 ve “kernel” parametresi için “rbf” değerleri ile en iyi kombinasyon sağlanmıştır. Bu kombinasyon ile %87.5 doğruluk oranı elde edilmiştir. Hiperparametre analizinden önce de aynı doğruluk oranı elde edildiğinden herhangi bir artış gözlemlenmemiştir.

5. SONUÇLAR VE ÖNERİLER

5.1 Sonuçlar

Doğal dil işleme günümüzün dikkat çeken alanlarından bir tanesidir. Bunun sebebi uygulama alanının çok geniş olması ve günlük hayatımızı kolaylaştıran birçok örneğinin bulunmasıdır. Bunlardan bir tanesi de doğal dil işleme yöntemleri kullanılarak metinlerin sınıflandırılmasıdır. Günümüzde veri boyutunun büyüklüğü ve artış hızı dikkate alındığında metin sınıflandırma yapmanın önemi ortaya çıkmaktadır. Makine öğrenmesi ve doğal dil işleme yöntemleri kullanılarak metinler hızlı ve maliyet etkin bir şekilde otomatik olarak sınıflandırılabilir. Sınıflandırılmamış metinlerle çalışmak yorucu ve zaman alan bir süreçtir. Metin sınıflandırma, istenen veriye daha hızlı ulaşmayı ve kolay işlem yapabilmeyi daha maliyet etkin bir biçimde sağlar. Metinin sınıflandırılabilmesi için genellikle denetimli makine öğrenmesi algoritmaları kullanılır. Denetimli makine öğrenmesi kullanmadan da metin sınıflandırması yapılabilir. Yani kural tabanlı bir sistem tasarlanarak sınıflandırma işlemi yapılabilir. Ancak bu tez çalışmasında kural tabanlı model yerine makine öğrenimi temelli model ile çalışılmıştır.

Çalışmada farklı kategorilerde yazılmış olan şiirlerin kategorisi, otomatik olarak önceden belirlenmiş kategorilere göre tahmin edilmektedir. Bu sınıflandırmanın yapılabilmesi için altı farklı makine öğrenmesi algoritması ve doğal dil işleme teknikleri birlikte kullanılmıştır. Yapılan çalışma yedi aşamadan oluşmaktadır. Birinci aşamada gerekli kütüphaneler projeye dâhil edilmiştir. İkinci aşamada veri seti projeye dâhil edilmiştir. Veri seti web kazıma yöntemi ile elde edilmiştir. Üçüncü aşamada metin ön işleme yapılmıştır. Bu aşamada doğal dil işleme yapılabilmesini sağlayan Zemberek Kütüphanesi kullanılmıştır. Dördüncü aşamada metnin sayısal bir formatta temsil edilebilmesi için vektörleştirme yapılmıştır. Bilgisayarların sayısal veriler üzerinde işlem yapabilmesi nedeniyle metinlerin sayılardan oluşan vektörlere dönüştürülmesi işlemidir. Metni sayısal forma dönüştürmek için farklı yaklaşımlar bulunmaktadır. Bu çalışmada metinin sayısal forma dönüştürülmesi için TF-IDF yöntemi kullanılmıştır. Beşinci aşamada eğitim ve test veri setleri oluşturulmuştur. Çalışmada kullanılan veri seti %80 eğitim ve %20 test olacak şekilde kullanılmıştır. Altıncı aşamada sınıflandırma algoritmaları kullanılarak oluşturulan model eğitilmiş ve tahmin işlemi gerçekleştirilmiştir. Sınıflandırma yapılabilmesi için SVM, MLP, Naive Bayes, K-NN, Karar Ağacı ve Rastgele Orman algoritmaları kullanılmıştır. En iyi sonucu %87

doğruluk oranı ile SVM ve Rastgele Orman algoritmaları vermiştir. Yedinci ve son aşamada sınıflandırma algoritmalarının performansı değerlendirilmiştir. Değerlendirmenin yapılabilmesi için karmaşıklık matrisi kullanılmıştır.

Sınıflandırma başarısının artırılması için hiperparametre analizi yapılmıştır. Bunun için RandomizedSearchCV ve GridSearchCV yöntemleri kullanılmıştır. Bu analiz sonucunda ise Rastgele orman algoritmasının doğruluk sonucu %89 olarak elde edilmiştir.

5.2 Öneriler

Bu çalışmada makine öğrenmesi ve doğal dil işleme yöntemleri kullanılarak metinlerin sınıflandırılması çalışması yapılmıştır. İnternet sitelerinden elde edilen ve sınıf etiketleri önceden belli olan veri seti eğitim ve test veri setlerine bölünerek sınıflandırma işlemi gerçekleştirilmiştir.

Yapılacak olan sınıflandırma çalışmalarında, veri seti önemli bir yere sahiptir. Veri setinin boyutu sonucu etkileyen bir faktör olup veri setinin boyutu arttıkça zaman maliyeti ve başarı oranı doğru orantılı olarak artmaktadır. Tam tersi durumda ise zaman maliyeti ve başarı oranı doğru orantılı olarak azalmaktadır. Sınıflandırma çalışmalarında veri setinin boyutu kademeli olarak artırılarak elde edilen sonuçlar neticesinde veri seti boyutuna karar verilmelidir.

Çalışmada kullanılacak olan veri seti, veri ön işleme adımından geçirilmelidir. Bu işlem doğru ve güvenilir sonuçlar elde edilmesi bakımından önemli bir aşamadır. Veri ön işleme adımında eksik verilerin tamamlanması, tekrarlayan verilerin kaldırılması gibi düzeltici işlemler yapılabildiğinden veri seti sınıflandırma adımına hazırlanmış olur. Doğal dil işleme adımlarının gerçekleştirilebilmesi için farklı kütüphaneler bulunmaktadır. Çalışmada farklı kütüphaneler kullanılarak elde edilen sonuçlarda oluşan değişim gözlemlenebilir.

Veri ön işleme adımı yapıldıktan sonra çalışmada kullanılacak olan metin vektörleştirme yöntemleri belirlenmelidir. Farklı yöntemler kullanılarak elde edilen sonuçlar kıyaslanmalı ve çalışma için en uygun yönteme karar verilmelidir.

Sınıflandırma işlemi için kullanılabilecek birçok algoritma bulunmaktadır. Algoritmalar farklı problemeler için farklı sonuçlar vermektedir. Bu çalışmada metin sınıflandırma problemi için yaygın olarak kullanılan gözetimli öğrenme algoritmaları (SVM, MLP, K-NN, Decision Tree, Random Forest, Naive Bayes) kullanılmıştır. Farklı

algoritmalar kullanılarak elde edilen sonuçların gözlemlenmesi daha verimli sonuçlar alınmasını sağlamaktadır.

KAYNAKLAR

Akalın, Ş. H., 2009, Türk dili: Dünya dili, *Türk Dili*, 687, 195-204.

Akın, M. D. ve Akın, A. A., 2007, Türk dilleri için açık kaynaklı doğal dil işleme kütüphanesi: ZEMBEREK, *Elektrik mühendisliği*, 431, 38-44.

Alrefaai, M., 2021, Haber Analizi ve Sınıflandırması İçin NLP Makine Öğrenmesinin Uygulanması, *Bahçeşehir Üniversitesi*.

Altıntaş, A. G., 2022, AUTOMATIC QUOTE DETECTION FROM LITERARY WORK, *İzmir Yüksek Teknoloji Enstitüsü*.

Amasyalı, M. F. ve Diri, B., 2006, Automatic Turkish text categorization in terms of author, genre and gender, *International Conference on Application of Natural Language to Information Systems*, 221-226.

Amasyalı, M. F., BALCI, S., Emrah, M. ve VARLI, E., 2012, Türkçe metinlerin sınıflandırılmasında metin temsil yöntemlerinin performans karşılaştırılması/a comparison of text representation methods for turkish text classification, *EMO Bilimsel Dergi*, 2 (4).

Anonim, 2023a, Şiir Bilgisi, Web adresi: https://www.dilbilgisi.net/siir-bilgisi-konu-anlatimi/ [Ziyaret Tarihi: 20.01.2023]:

Anonim, 2023b, Dil Bilgisi ve İmla Kuralları, Web adresi: https://acikders.ankara.edu.tr/pluginfile.php/134072/mod_resource/content/1/TSU156%20YAZI%C5%9EMA%20TEKN%C4%B0KLER%C4%B0-sayfalar-50-87.pdf [Ziyaret Tarihi: 18.01.2023]:

Bakır, R., 2022, DOĞAL DİL İŞLEME TEKNİKLERİNİ VE DERİN ÖĞRENME ALGORİTMALARINI KULLANARAK SOSYAL AĞLARDA SPAM TESPİTİ, *KIRIKKALE ÜNİVERSİTESİ*.

Ballı, Ç., 2021, Doğal Dil İşleme ile Türkçe İçerikli Paylaşımlardan Sosyal Medya Kullanıcılarının Duygu Analizi, *Ankara Üniversitesi*.

Bardız, A., 2019, Türkçe Medikal Metin Ayrıştırma ve Sınıflandırma, *Galatasaray Üniversitesi*.

Barzenji, H. S. A., 2021, Makine öğrenme algoritmaları kullanılan twitter metinlerinin duygu analizi, *Sakarya Üniversitesi*.

Bozyiğit, A., 2021, Sosyal Ağlarda Sanal Zorbalığın Otomatik Olarak Tespit Edilmesi, *Dokuz Eylül Üniversitesi*.

Bulut, F. ve Bozyılan, D., ÇOK KATMANLI ALGILAYICILAR İLE DOĞRU MESLEK SEÇİMİ.

Bulut, F., 2016, ÇOK KATMANLI ALGILAYICILAR İLE DOĞRU MESLEK TERCİHİ, *Anadolu University Journal of Science and Technology A-Applied Sciences and Engineering*, 17 (1), 97-109.

Çebi, C. B., 2020, Rastgele Orman Algoritması, Web adresi: https://medium.com/@cemthecebi/rastgele-orman-algoritmas%C4%B1-1600ca4f4784 [Ziyaret Tarihi: 23.01.2023]:

Ekim, H. E., 2021, Türkiye'deki Havayolu Firmalarıyla İlgili Sosyal Medya Yorumlarının Makine Öğrenmesi Yöntemleriyle Sınıflandırılması, *Kocaeli Üniversitesi*.

Eldeniz, L., 1994, Programlama dilleri, *Marmara İletişim Dergisi*, 7 (7), 139-144.

Ercilasun, A. B., 2013, Türkçenin dünya dilleri arasındaki yeri, *Dil Araştırmaları*, 12 (12), 17-22.

Eryiğit, G. ve Adalı, E., 2004, An affix stripping morphological analyzer for Turkish, *Proceedings of the iasted international conference on artificial intelligence and applications*, 16-18.

Eryiğit, G. ve Oflazer, K., 2006, Statistical dependency parsing of Turkish.

Eryiğit, G., 2012, Biçimbilimsel çözümleme, *Türkiye Bilişim Vakfı Bilgisayar Bilimleri ve Mühendisliği Dergisi*, 5 (2).

Franko, S., 2018, Multiclass Analysis of Automatic Text Classification Techniques, *Galatasaray Üniversitesi*.

Gürcan, F., 2018, Multi-class classification of turkish texts with machine learning algorithms, *2018 2nd international symposium on multidisciplinary studies and innovative technologies (ismsit)*, 1-5.

Henderi, H., Wahyuningsih, T. ve Rahwanto, E., 2021, Comparison of Min-Max normalization and Z-Score Normalization in the K-nearest neighbor (kNN) Algorithm to Test the Accuracy of Types of Breast Cancer, *International Journal of Informatics and Information Systems*, 4 (1), 13-20.

Kekül, H., 2022, Yazılım Güvenlik Açıklarının Skorlanması ve Kategorilerinin Belirlenmesinde Yeni Bir Yöntem, *Fırat Üniversitesi*.

Kodedu, 2014, Naïve Bayes Sınıflandırma Algoritması, Web adresi: https://kodedu.com/2014/05/naive-bayes-siniflandirma-algoritmasi/ [Ziyaret Tarihi: 19.02.2023]:

Köksal, A., 1975, A first approach to a computerized model for the automatic morphological analysis of Turkish, *Unpublished Doctoral Thesis*.

Kolyiğit, Ö., 2013, Türkçe dokümanlar için yazar tanıma, *Adnan Menderes Üniversitesi, Fen Bilimleri Enstitüsü*.

Kontuk, R., 2020, NLP Kullanılarak Haberlerin Yaş Gruplarına Göre Sınıflandırılması, *İstanbul Ticaret Üniversitesi*.

Korkmaz, S., 2021, Makine Öğrenme Yöntemleri Kullanılarak Doğal Dil İşleme Tabanlı Şair Tanıma, *Erciyes Üniversitesi*.

Küçük, Z., 2023, Matris Hatası Nedir? -Confusion Matrix, Web adresi: https://womaneng.com/matris-hatasi-nedir-confusion-matrix/ [Ziyaret Tarihi: 20.02.2023]

Mumcuoğlu, E., 2022, Doğal Dil İşleme Yöntemleri Kullanılarak Türk Yüksek Mahkemelerinde Karar Tahmini, *İhsan Doğramacı Bilkent Üniversitesi*.

Oflazer, K., 1994, Two-level description of Turkish morphology, *Literary and linguistic computing*, 9 (2), 137-148.

Oflazer, K., 2016, Türkçe ve Doğal Dil İşleme, *Türkiye Bilişim Vakfı Bilgisayar Bilimleri ve Mühendisliği Dergisi*, 5 (2).

Özkan, M., 2022, Multiclass classification of scientific texts written in Turkish by applying deep learning technique, *Bahçeşehir Üniversitesi*.

Özmutlu, A. E., 2021, DOĞAL DİL İŞLEME, *Bilgisayar Bilimlerinde Teorik Ve Uygulamalı Araştırmalar*, 129.

Öztürk, A., Durak, Ü. ve Badıllı, F., 2020, Twitter Verilerinden Doğal Dil İşleme ve Makine Öğrenmesi ile Hastalik Tespiti, *Konya Mühendislik Bilimleri Dergisi*, 8 (4), 839-852.

Öztürk, M., 2022, Python ile Sınıflandırma Analizleri – Rastgele Orman (Random Forest) Algoritması, Web adresi: https://miracozturk.com/python-ile-siniflandirma-analizleri-rastgele-orman-random-forest-algoritmasi [Ziyaret Tarihi: 26.02.2023]

Patro, V. M. ve Patra, M. R., 2014, Augmenting weighted average with confusion matrix to enhance classification accuracy, *Transactions on Machine Learning and Artificial Intelligence*, 2 (4), 77-91.

Sansak, Y. A., 2023, AUTOMATIC STUTTERING DETECTION AND CLASSIFICATION, *TED Üniversitesi*.

Şeker, G. A. ve Eryiğit, G., 2012, Initial explorations on using CRFs for Turkish named entity recognition, *Proceedings of COLING 2012*, 2459-2474.

Seker, S. E., 2008, KNN (K nearest neighborhood, en yakın k komşu), Web adresi: https://bilgisayarkavramlari.com/2008/11/17/knn-k-nearest-neighborhood-en-yakin-k-komsu/ [Ziyaret Tarihi: 11.03.2023]

Şeker, S. E., 2015, Doğal Dil İşleme (Natural Language Processing), *YBS Ansiklopedi*, 2 (4), 14-31.

Şeker, Ş. E., 2011, n-gram, Web adresi: https://bilgisayarkavramlari.com/2011/04/23/n-gram/ [Ziyaret Tarihi: 23.03.2023]

Shirzadova, S., 2020, Türkçe YouTube Yorumları Üzerinde Spam Filtreleme, *Eskişehir Teknik Üniversitesi*.

Suncak, A., 2022, Türkçe İçin Doğal Dil Anlamada Anlatım Bozukluklarının Tespiti İçin Yeni Bir Yaklaşım Geliştirilmesi, *Dokuz Eylül Üniversitesİ*.

Tarim, R., 2013, Atatürk ve Türk Dili, *Turkish Studies-International Periodical For The Languages, Literature and History of Turkish or Turkic*, 8 (9), 95-103.

Taşkıran, S. F., 2021, Doğal dil işleme ile akademik metinlerin kümelenmesi, *Konya Teknik Üniversitesi*.

TDK, 2023a, Küçük Ünlü Uyumu, Web adresi: https://www.tdk.gov.tr/icerik/yazim-kurallari/kucuk-unlu-uyumu/ [Ziyaret Tarihi: 27.03.2023]

TDK, 2023b, Büyük Ünlü Uyumu, Web adresi: https://www.tdk.gov.tr/icerik/yazim-kurallari/buyuk-unlu-uyumu/ [Ziyaret Tarihi: 26.02.2023]

TDK, 2023c, Türk Dil Kurumu Tarihçe, Web adresi: https://www.tdk.gov.tr/tdk/kurumsal/tarihce-2 [Ziyaret Tarihi: 25.03.2023]

TDK, 2023d, Bitişik Yazılan Birleşik Kelimeler, Web adresi: https://www.tdk.gov.tr/icerik/yazim-kurallari/bitisik-yazilan-birlesik-kelimeler/ [Ziyaret Tarihi: 12.05.2023]

TDK, 2023e, Soru Eki mı / mi / mu / mü'nün Yazılışı, Web adresi: https://www.tdk.gov.tr/icerik/yazim-kurallari/soru-eki-mi-mi-mu-munun-yazilisi [Ziyaret Tarihi: 23.04.2023]

TDK, 2023f, Bağlaç Olan ki'nin Yazılışı, Web adresi: https://www.tdk.gov.tr/icerik/yazim-kurallari/baglac-olan-kinin-yazilisi [Ziyaret Tarihi: 20.04.2023]

TDK, 2023g, Sayıların Yazılışı, Web adresi: https://www.tdk.gov.tr/icerik/yazim-kurallari/sayilarin-yazilisi/ [Ziyaret Tarihi: 11.04.2023]

TDK, 2023h, Ünsüz Türemesi, Web adresi: https://www.tdk.gov.tr/icerik/yazim-kurallari/unsuz-turemesi/ [Ziyaret Tarihi: 07.04.2023]

TDK, 2023i, Ünsüz Uyumu, Web adresi: https://www.tdk.gov.tr/icerik/yazim-kurallari/unsuz-uyumu [Ziyaret Tarihi: 28.03.2023]

TDK, 2023j, Ünlü Düşmesi, Web adresi: https://www.tdk.gov.tr/icerik/yazim-kurallari/unlu-dusmesi [Ziyaret Tarihi: 27.03.2023]

TDK, 2023k, Ünlü Daralması, Web adresi: https://www.tdk.gov.tr/icerik/yazim-kurallari/unlu-daralmasi [Ziyaret Tarihi: 26.03.2023]

Tekin, T., 1978, Türk Dilleri Ailesi I, *Türk Dili*, 37 (318), 173-183.

Torunoğlu, D., Çakirman, E., Ganiz, M. C., Akyokuş, S. ve Gürbüz, M. Z., 2011, Analysis of preprocessing methods on classification of Turkish texts, *2011 International Symposium on Innovations in Intelligent Systems and Applications*, 112-117.

Trstenjak, B., Mikac, S. ve Donko, D., 2014, KNN with TF-IDF based framework for text categorization, *Procedia Engineering*, 69, 1356-1364.

Türkoğlu, F., Diri, B. ve Amasyalı, M. F., 2007, Author attribution of Turkish texts by feature mining, *International Conference on Intelligent Computing*, 1086-1093.

Tutorialspoint, 2023, Natural Language Processing, Web adresi: https://www.tutorialspoint.com/natural_language_processing/natural_language_processing_introduction.htm [Ziyaret Tarihi: 28.02.2023]:

Uzun, E., 2023a, K-NN Algoritması, Web adresi: https://erdincuzun.com/makine_ogrenmesi/k-nn-algoritmasi/ [Ziyaret Tarihi: 16.05.2023]

Uzun, E., 2023b, Decision Tree (Karar Ağacı): ID3 Algoritması – Classification (Sınıflama), Web adresi: https://erdincuzun.com/makine_ogrenmesi/decision-tree-karar-agaci-id3-algoritmasi-classification-siniflama [Ziyaret Tarihi: 18.05.2023]

Uzun, N. E., 2012, Türkçenin dünya dilleri arasındaki yeri üzerine, *Türkoloji dergisi*, 19 (2), 115-134.

Vikipedi, 2023, Papua Yeni Gine, Web adresi: https://tr.wikipedia.org/wiki/Papua_Yeni_Gine#cite_note-7 [Ziyaret Tarihi: 10.02.2023]

Yıldırım, S. ve Yıldız, T., 2018, Türkçe için karşılaştırmalı metin sınıflandırma analizi, *Pamukkale Üniversitesi Mühendislik Bilimleri Dergisi*, 24 (5), 879-886.

Yılmaz, S., 2009, Türkçe İçin İyileştirilmiş Biçimbirimsel Çözümleyici, *Fen Bilimleri Enstitüsü*.