T.C.
SÜLEYMAN DEMİREL ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

KURAKLIKLA İLGİLİ SOSYAL MEDYA MESAJLARININ DUYGU ANALİZİ

Sevdanur DURAN

Danışman
Dr. Öğr. Üyesi Turgay AYDOĞAN

YÜKSEK LİSANS TEZİ
BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI
ISPARTA - 2023

İÇİNDEKİLER

ÖZET

Yüksek Lisans Tezi

KURAKLIKLA İLGİLİ SOSYAL MEDYA MESAJLARININ DUYGU ANALİZİ

Sevdanur DURAN

Süleyman Demirel Üniversitesi
Fen Bilimleri Enstitüsü
Bilgisayar Mühendisliği Anabilim Dalı

Danışman: Dr. Öğr. Üyesi Turgay AYDOĞAN

Sosyal medya, dünya üzerinde sürekli ve hızla gelişen, insanlarda adeta bağımlılık haline gelen sanal bir sistemdir. İnsanlar herhangi bir konudaki duygularını, düşüncelerini, kişisel ilgi alanlarını, olaylara bakış açılarını, beğenilerini bu sistem içerisinde yer alan sosyal medya platformlarında paylaşırlar. Bu sebeple, sosyal medya paylaşımları duygu analizi alanı için büyük bir veri kaynağı haline gelmiştir. Ülkemizde su kaynakları ve yağış miktarı giderek azalmaktadır. Bu da kuraklık sorununu ortaya çıkarmaktadır. Kuraklık, Türkiye gündemindeki en güncel sorun ve geleceğimizi tehdit eden bir doğa olayıdır. Bu çalışmada küresel bir sorun olan kuraklık hakkında, sosyal medya platformlarından Twitter verileri kullanılarak duygu analizi yapılmıştır.

Bu tez çalışmasında kullanılan veriler 01.01.2019'dan 01.01.2022'ye kadar ki tarih aralığında yer alan tweetlerden oluşmaktadır. Bu tweetlerden içerisinde "kuraklık" etiketi içeren toplam 96.401 adetlik bir veri seti oluşturulmuştur. Python programlama dili kullanılarak bütün tweet toplama, ön işleme ve duygu analizi işlemleri gerçekleştirilmiştir. Oluşturulan veri setinin ön işleminden ve Zemberek kütüphanesi ile yapılan normalleştirme sonrası 82.221 adet tweet üzerinde duygu analizi yapılmıştır. Sınıflandırma aşamasında SWNetTR++ sözlüğünü kullanarak etiketlenen veri seti kullanılmıştır. Özellik çıkarımı için BoW ve TF-IDF yöntemi kullanılmıştır. Ayrıca N-gram yöntemi kullanılarak veriler 1-gram, 2-gram, 3-gram olarak ayrılmıştır. Sınıflandırma işlemlerinde her N-gram değeri için ayrı ayrı hesaplama yapılmıştır.

Naive Bayes, Lojistik Regresyon, Karar Ağacı, Rastgele Orman, Destek Vektör Makinesi, K-En Yakın Komşu makine öğrenmesi algoritmaları ile sınıflandırma yapılmıştır. Yapılan duygu analizi işlemleri sonucunda %56 Negatif, %30 Pozitif ve %14 Nötr duygu sonuçlarına ulaşılmıştır. Elde edilen sonuçlar doğrultusunda insanların kuraklık hakkında bilinçsiz ve negatif düşüncelerinin daha fazla olduğu baskın olan duygu çıktısından görülmektedir. Sınıflandırma performans sonucunda ise sözlük tabanlı etikete sahip olunan yapıda en iyi BoW – Destek Vektör Makinesi ikilisi ile 0,85'lik bir sınıflandırma başarısı elde edilmiştir. Diğer modellerin performans sonuçlarına bakıldığında en iyi sonuçtan azalana doğru sıralama yapılır ise ikinci en iyi değerin 0,84 ile BoW - Lojistik Regresyon'a ait olduğu görülmektedir. Diğer oranların ise 0,70 ile BoW-Naive

Bayes, 0,69 ile TF-IDF- Rastgele Orman, 0,61 ile BoW- Karar Ağacı ve son olarak ise 0,61 ile TF-IDF - K-En Yakın Komşu algoritmalarına ait olduğu hesaplanmıştır.

Anahtar Kelimeler: Kuraklık, duygu analizi, sosyal medya, makine öğrenmesi, twitter.

2023, 86 sayfa

ABSTRACT

M.Sc. Thesis

SENTIMENT ANALYSIS OF SOCIAL MEDIA DATA ABOUT DROUGHT

Sevdanur DURAN

Süleyman Demirel University
Graduate School of Natural and Applied Sciences
Department of Computer Engineering

Supervisor: Asst. Prof. Dr. Turgay AYDOĞAN

Social media is a virtual system that is constantly and rapidly developing in the world and has become an addiction. People share their feelings, thoughts, personal interests, perspectives and likes on any subject on social media platforms. For this reason, social media shares have become a great source of data for the field of sentiment analysis. In our country, water resources and the amount of precipitation are gradually decreasing. This also raises the problem of drought. Drought is the most current problem on the agenda of Turkey and a natural phenomenon that threatens our future. In this study, sentiment analysis was conducted about drought, which is a global problem, using Twitter data from social media platforms.

The data used in this thesis consists of tweets in the date range from 01.01.2019 to 01.01.2022. A total of 96,401 datasets were created from these tweets, including the "drought" tag. All tweet collection, preprocessing and sentiment analysis operations were performed using the Python programming language. Sentiment analysis was performed on 82,221 tweets after the pre-processing of the created data set and normalization with Zemberek library. In the classification phase, the data set tagged using the SWNetTR++ dictionary was used. BoW and TF-IDF method were used for feature extraction. In addition, using the N-gram method, the data were divided into unigram, bigram, trigram. In classification processes, calculations were made separately for each N-gram value.

The classification was carried out with Naive Bayes, Logistic Regression, Decision Tree, Random Forest, Support Vector Machine, K-Nearest Neighbor machine learning algorithms. As a result of emotion analysis, 56% Negative, 30% Positive and 14% Neutral emotion results were obtained. According to the results obtained, it is seen from the dominant emotion output that people have more unconscious and negative thoughts about drought. As a result of the classification performance, a classification success of 0.85 was achieved with the best BoW – SVM duo in the structure with a dictionary-based label. When the performance results of the other models are examined, it is seen that the second best value belongs to BoW - Logistic Regression with 0.84. The other ratios were

calculated to belong to BoW-Naive Bayes with 0.70, TF-IDF- Random Forest with 0.69, BoW-Decision Tree with 0.61 and finally TF-IDF - K-Nearest Neighbor algorithms with 0.61.

Keywords: Drought, sentiment analysis, social media, machine learning, twitter.

2023, 86 pages

TEŞEKKÜR

Tez çalışmam boyunca engin tecrübesi ve bilgisiyle bana yol gösteren, desteğini ve zamanını asla esirgemeyen değerli tez danışmanım Sayın Dr. Öğr. Üyesi Turgay AYDOĞAN'a teşekkürlerimi sunarım.

Eğitimim boyunca her zaman yanımda olan, desteklerini hiçbir zaman esirgemeyen Sayın Prof. Dr. Ecir Uğur KÜÇÜKSİLLE hocama, aileme ve arkadaşlarıma sonsuz sevgi ve saygılarımı sunarım.

Sevdanur DURAN
ISPARTA, 2023

ŞEKİLLER DİZİNİ

ÇİZELGELER DİZİNİ

Sayfa

SİMGELER VE KISALTMALAR DİZİNİ

API	Application Programming Interface (Uygulama Programlama Arabirimi)
BERT	Bidirectional Encoder Representations for Transformers (Çift Yönlü Transformatör Kodlayıcı Temsilleri)
BoW	Bag of Words (Kelime Çantası)
CBoW	Continuous Bag of Words (Sürekli Kelime Çantası)
DDİ	Doğal Dil İşleme
DT	Desicion Tree (Karar Ağacı)
FN	False Negatif (Yanlış Negatif)
FP	False Positive (Yanlış Pozitif)
IDE	Integrated Development Enviroment (Tümleşik Geliştirme Ortamı)
IDF	Inverse Document Frequency (Ters Belge Frekansı)
KNN	K-Nearest Neighbours (K-En Yakın Komşu)
LR	Lojistik Regresyon
LSTM	Long Short-Term Memory (Uzun Kısa Süreli Bellek)
MSSQL	Microsoft SQL Server
NB	Naive Bayes
NLP	Natural Language Processing (Doğal Dil İşleme)
NLTK	Natural Language Toolkit (Doğal Dil Araç Takımı)
RF	Random Forest (Rastgele Orman)
SVM	Support Vector Machine (Destek Vektör Makinesi)
TF	Term Frequency (Terim Frekansı)
TN	True Negatif (Doğru Negatif)
TP	True Positive (Doğru Pozitif)
Vader	Valence Aware Dictionary and Sentiment Reasoner

1. GİRİŞ

Sosyal medya veya farklı bir ifade ile sosyal medya platformu, dünya üzerinde sürekli gelişen sanal bir sistemdir. Bu sistem içerisinde kullanıcılar içerik üretebilmekte, diğer kullanıcılar ile eş zamanlı olarak karşılıklı bilgi akışı sağlayıp interaktif iletişim ve etkileşim kurabilmektedirler. Gün geçtikçe yeni bir sosyal medya platformu ortaya çıkmaktadır. Sosyal medya platformlarının bu artışına bağlı olarak da sanal sistem içerisinde kullanıcıların oluşturduğu veriler de her geçen gün artış göstermektedir.

Twitter, sosyal medya platformları içerisinde en çok kullanılan sosyal ağ platformu olarak ön plana çıkmaktadır. Twitterda kullanıcılarının oluşturduğu milyonlarca veri vardır. Bu platformdaki veriler uzun zamandır analiz edilerek elde edilen sonuçlar paylaşılmaktadır. Dünya genelinde Twitter'ın 330 milyondan fazla kullanıcısı, Türkiye'de ise yaklaşık 12 milyon kullanıcısı bulunmaktadır. Kullanıcılar "tweet" olarak ifade edilen Twitter'daki paylaşılan mesajlarında mesaj içeriğine göre özet kelime veya kelime grupları kullanmaktadırlar. Bu kelime ve kelime grupları etiket olarak tanımlanmaktadır. Neredeyse hemen hemen her tweette bu etiketler kullanılarak paylaşılan mesajların aynı duyguya sahip kullanıcılara veya hedef kitlelerine ulaşması amaçlanmaktadır. Aktif ve sürekli olarak kullanılan bu etiketler ile insanlar birlik olabilmekte, markalar müşteri kitlelerine ulaşabilmekte, sivil toplum örgütleri de hedeflenen kitlelerle etkileşim içine girebilmektedir (Karabulut, 2018).

Sosyal medya platformları, bilgiyi elde etme, doğru kullanma ve paylaşma yönüyle insan yaşamının parçası hâline gelmişlerdir. Bu platformlarda kullanıcılar tarafından her gün kendilerini ifade eden veya toplumsal olaylar ile ilgili duygu ve düşüncelerini içeren oldukça fazla paylaşımlar ortaya çıkmaktadır. Bununla birlikte verileri doğru analiz etme ve verilerden anlamlı sonuçlar çıkarma ihtiyacı da hem ticari hem de bilimsel bir konu olarak ön planda yerini almıştır. Anlamlı sonuçlar çıkarmak için sosyal medya kullanıcı paylaşımlarından yola çıkarak konular ile ilgili değerlendirmelerini,

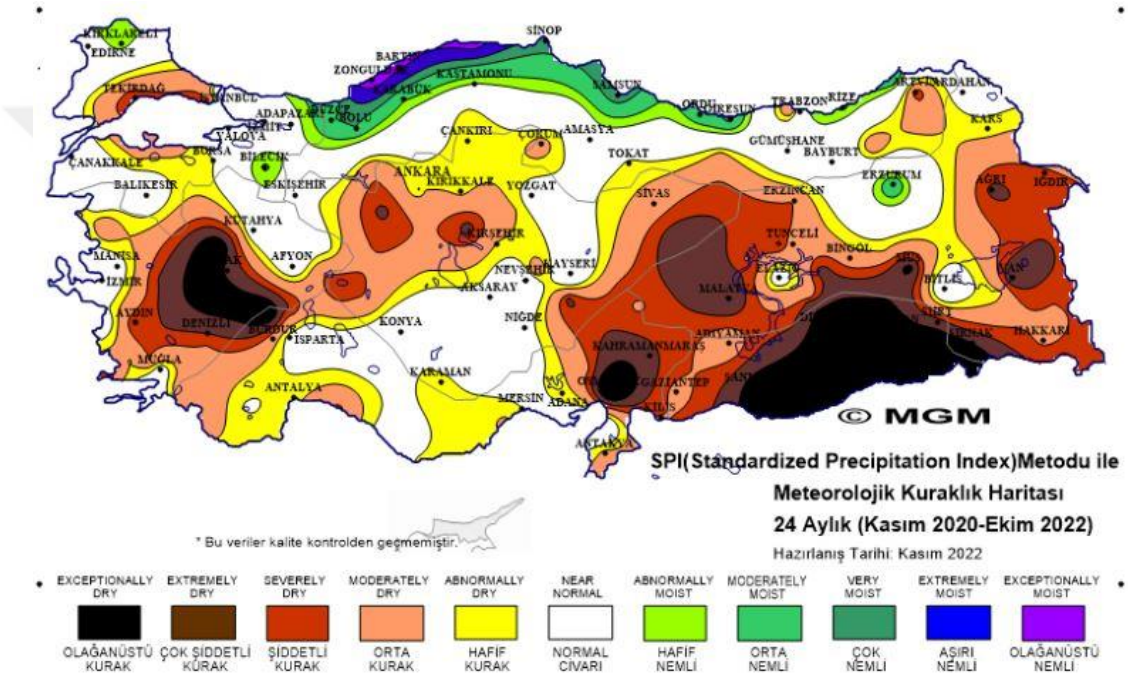
düşüncelerini, duygularını ve tutumlarını analiz eden işlemler bütünü duygu analizi olarak ifade edilmektedir (Kırcı ve Gülbak, 2020).

Çeşitli konulardaki düşünce ve duygular ile ilgili özgün ve güncel konular üzerine birçok çalışma yapılmış ve son zamanlarda duygu analizi çalışmaları hız kazanmıştır. Literatür incelendiğinde duygu analizi çalışmalarında genellikle film, otel, marka vb. konularının ele alındığı görülmektedir. Doğal afetler üzerindeki etkileşimlerine yönelik yapılan çalışmalarda sadece deprem konusunun ele alındığı görülmüştür.

Dünyada meteorolojik ve jeolojik olarak meydana gelen, insanları, insanların çevrelerini ve yaşamlarını ciddi anlamda olumsuz bir şekilde etkileyen, doğal olarak oluşan olaylar "doğal afet" olarak tanımlanabilir. Dünya üzerindeki tüm doğal afetler araştırıldığında 31 çeşit doğal afetin olduğu görülmektedir (Büyükbaş ve Ormanoğlu, 2013). Bunlardan bazıları, kuraklık, deprem, sel, heyelan, çığ vb. dir. Meteoroloji Mühendisleri Odası (MMO) tarafınca oluşturulan rapora göre**,** doğal afetlerin oluşum özelliklerinin ve doğa üzerinde oluşturduğu etkilerin puanlanması sonucundaki risk gruplamasına göre kuraklığın en önemli afet olarak birinci sırada yer aldığı gözlemlenmiştir (MMO, 1999).

Kuraklık, dünyada oluşturabileceği zararlar ve toplumun bu konuda yeterli bilgiye sahip olmaması gibi sebeplerden dolayı insanlığı ciddi boyutta tehdit eden en önemli doğal afetlerden kabul edilmektedir ve 'tabiatın görünmez tehlikesi' olarak adlandırılabilir (Partigöç ve Soğancı, 2019). Kuraklık iklimsel değişimlerden dolayı belirli alanlarda atmosferik, yer altı - yer üstü tüm su kaynaklarının bir bölgedeki insanların ve bitkilerin ihtiyacını karşılayamaması durumudur (Wilhite ve Buchanan-Smith, 2005). Kuraklık günümüzde iklim değişikliğiyle doğal nem dengesinin korunamadığı ve yağış rejiminin tamamen bozulduğu yerlerde görülmeye başlamıştır. Fazla su tüketimi, ormanların yok edilmesi vb. gibi etkenler kuraklığı önemli ölçüde etkilemektedir. Su kaynakları ve yağış miktarı giderek azalmaktadır. Küresel ısınmanın artması ile birlikte iklimde meydana gelen değişimler nedeniyle kuraklık, ülkemiz için en büyük

tehditlerden olmuştur. Şekil 1.1'de Kasım 2020 – Ekim 2022 aralığına ait Türkiye'nin 24 aylık kuralık haritası verilmiştir. Kuraklığın meydana getireceği zararlar bakımından ve farkındalığın yeterli seviyede olmaması, durumun daha da tehlikeli hale gelmesine sebep olmaktadır. Kuraklığın süresi, etkisi ve zamanının tahmini yeterince zor olup kuraklığın etkileri, insan faaliyetleri ile de yakın ilişkilidir (Turan, 2018). Kuraklık sadece fiziksel bir olay veya bir doğa olayı olarak görülmemelidir buna sebep olabilecek insan davranış ve duygularının da olabileceği unutulmamalıdır.



Şekil 1.1. Meteorolojik kuraklık haritası (MGM, 2022)

Yapılan bu tez çalışmasında ise daha önce değinilmemiş Türkiye'deki en güncel sorun olan ve geleceği tehdit eden doğa olaylarından kuraklık ile ilgili duygu analizi yapılmıştır.

Sosyal medya kullanıcıları, Twitter aracılığıyla her konuda kolayca mesaj paylaşımı yapıp yeri geldiğinde toplumu bilinçlendirebilmektedirler. Bu çalışmada kuraklık etiketini içeren sosyal medya mesajları toplanarak bir sosyal medya platformu kitlesinin duygu analizlerinin yapılması amaçlanmıştır. Yapılan analizleri sonucu kuraklık ile ilgili baskın olan duyguları ortaya

çıkarılmıştır. Böylece sosyal medyanın günlük hayatımızda oldukça fazla yer alması ile kuraklığın sosyal medyada ne kadar ön planda olup olmadığının ortaya çıkarılmıştır. Duygu analizi sonuçlarını toplum ile paylaşımı yapılarak ülkemizde, kuraklığa hazırlıklı olmak, önlemek ve zararlarını azaltmayı sağlamak adına katkıda bulunabilmesi imkânı sunmaktadır. Bundan dolayı tez çalışması konusu ve sonuçları itibariyle özgünlüğe sahip bir çalışma olmuştur.

Yapılan tez çalışmasının ikinci bölümünde daha önce literatürde yer alan duygu analizine yönelik akademik çalışmalar hakkında bilgi verilmiştir. Üçüncü bölümde çalışma kapsamında kullanılan materyal ve metot hakkında bilgi verilmiştir. Duygu analizi ve Twitter hakkında bilgilere yer verilip analiz sürecinin gerçekleştirilebilmesi için gerekli yöntemlerden de bahsedilmiştir.

Dördüncü bölümde ise deneysel çalışmalar ve analiz için uygulanan adımlar tek tek ele alınmıştır. Twitter'dan elde edilen verilerin toplanmasından ön işleme adımlarına kadar her detaydan bu kısımda bahsedilmiştir. Gerekli kod yazılımları, sayısal analiz sonuçları ve ekran görüntüsü paylaşımları bu bölümün sonunda verilmiştir. Son bölümde ise literatür kıyaslamaları ile elde edilen sonuçlar tartışılmıştır.

2. KAYNAK ÖZETLERİ

Literatüre bakıldığında, duygu analizi konusunda çeşitli çalışmaların olduğu görülmektedir. Bilimsel çalışmaların birçok farklı yöntemlerle yapıldığı gözlemlenmiştir. Çalışmanın bu kısmında, Twitter verileri ile yapılmış bilimsel çalışmalara yer verilmiştir.

Onan (2017), yaptığı çalışmada, Türkçe tweetler üzerinde duygu analizini makine öğrenmesi yöntemlerini kullanarak gerçekleştirmiştir. Analiz için, Twitter API (Application Programming Interface, Uygulama Programlama Arabirimi) kullanılarak, Python ile bir aylık zaman sürecindeki Twitter mesajları toplanmıştır. 5300 pozitif ve 5300 negatif olmak üzere toplam 10600 Twitter mesajı içeren veri seti oluşturmuşlardır. Çalışmada Zemberek kütüphanesi kullanılmıştır. Çalışma kapsamında tweetlerin sınıflandırılmasında, üç ana makine öğrenmesi sınıflandırıcısı (naive bayes, destek vektör makineleri ve lojistik regresyon) ve üç ana kelime temsil modeli (1-gram, 2-gram ve 3-gram) ile değerlendirilmiştir. Sonuçlara göre, en yüksek başarının NB (Naive Bayes) algoritması %77,78 ile ve veri seti 1-gram ve 2- gram kelime öznitelik setlerinin birleştirilmesiyle oluşturulan veri seti ile gerçekleştiği gözlemlenmiştir. Türkçede sözcük türü etiketleme araçlarının olmaması sebebiyle, çalışma kapsamında yalnızca N-gram modelleri kullanılmıştır.

Alfarrarjeh vd. (2017), duygu analizi üzerine yaptıkları bu çalışmada afetlerle ilgili atılan tweetlerin coğrafi mekânsal özelliklerini ön planda tutmuşlardır. Bunun amacının verileri daha bölgesel iç görülere göre kıyaslayabilmek olduğunu belirtmişleridir. Sandy Kasırgası ve Napa Depremi etiketleri verileri toplamışlardır. Verileri NLTK (Natural Language Toolkit, Doğal Dil Araç Takımı) yöntemi ile sınıflandırmışlardır. Deprem anında aynı bölgede olan kullanıcıların tweetlerini analiz derken birden fazla karmaşık duygular olduğunu fark etmişlerdir. Bu karmaşıklığın çalışmalarındaki en büyük zorluk olduğunu belirtmişlerdir. Sonuçta ise depremin olduğu bölgeye yakın bir coğrafi konumdaki atılan tweetler ile bölgedeki duyguların ileri derecede tutarsız ve zıt duygularda olduğu görülmüştür.

Karaöz ve Şimşek (2018), yaptıkları çalışmada Twitter'dan bir televizyon kanalı seçmişlerdir ve o televizyon kanalında 8 ay boyunca yayınlanan programlar hakkında atılan tweet verilerini toplamışlardır. Duygu analizi ile seyircilerin verilerinin olumlu, olumsuz ya da nötr durumlarını incelemişlerdir. 1000' er adet Negatif ve Pozitif tweet etiketlemişlerdir. Programlama dili olarak R dilini seçmişlerdir. Veri seti metin ön işleme ve etiketleme yapmışlardır. İstatistiksel doğruluk oranı %68 olarak belirtmişlerdir. Sonuç olarak bu bilgiler ile televizyon kanal yöneticileri ve yayınlanan programın sorumluluklarını üstlenen kişilere öngörülerde bulunmuşlardır ve bu doğrultuda gerekli stratejileri geliştirebileceklerini belirtmişlerdir.

Karabulut (2018), geliştirdiği uygulama ile Twitter platformundaki kullanıcıların tweet mesajlarında kullandıkları etiketler (#hashtag) üzerinden analiz yapmıştır. Bu analiz sonucu ile gündemdeki herhangi bir olayın, etkinliğin detaylarına ulaşabilmekte ve atılan tweet sayısı, etkileşim sayısı, duygu analizi ve tweet mesajı yayınlayan kullanıcılar üzerinden cinsiyet analizini yapabilmektedir. Geliştirdiği uygulamada; veri tabanı yönetim sistemi olarak PostgreSql, programlama dili olarak Python ve web altyapısı için ise Odoo Framework kullanmıştır. Örnek etiket analizi olarak #teknoloji seçmiştir. Toplamda 1520 tweet çekmiştir ve cinsiyete göre analiz sonucu tweet atan Twitter kullanıcıların %37,93 'ünün erkek ve %12,70'inin kadın olduğu sonucuna varılmıştır. Atılan tweetlerin duygu analizi sonucunda %25'inin pozitif duygular içerdiği ve %65'inin ise negatif duygularının daha baskın olduğu gözlemlenmiştir. Yaptığı çalışma sonucunda geliştirdiği analiz aracı günümüzdeki analiz araçlarından farklı olarak birden fazla çeşitli birçok analiz türünü aynı panel içinde gerçekleştirebilmektedir.

Jamali vd. (2018), bu çalışmada Twitter kullanıcılarının davranışlarını analiz etmiştir. Sandy Kasırgası ile ilgili tweetleri Twitter API ile toplamıştır. Kullanıcıların konum, cinsiyet vb. bilgilerini de toplamıştır. Yaşanılan felaket için Dirichlet regresyon modelini kullanmayı uygun görmüştür. Kullanıcıları afet deneyimi yaşayanlar ve yaşamayanlar olarak ikiye ayırmıştır. Sonuç olarak afet

hakkında, tecrübesi olan kullanıcılar günlük yaşamlarını etkileyecekleri, ancak afet tecrübesi olmayan kullanıcılar felaketi umursamadıklarını ve günlük yaşamlarını düşündüklerinin sonucuna varmışlardır.

Şavlukbaş (2019), bu çalışmada Rock'n Coke Festivali ile ilgili atılan tweet mesajlarının duygu analizini gerçekleştirmiştir. Amaç olarak ise Twitter sayesinde elde ettiği veriler ile tüketici ve festival arasındaki sorunların analizini sağlayarak etkinlik yönetim sürecinin nasıl olması gerektiğini yöneticilere açıklamayı hedeflemiştir. Veri seti 1895 adet negatif tweet, 4374 adet pozitif tweet, 5918 adet nötr tweet yani toplam 12187 adet tweetten oluşmaktadır. Bu tweetlerden yola çıkarak yönetimin, personelin, tüketicinin, festival ücretinin, mekânın vb. detaylı analizlerini yapmıştır. Yaptığı çalışma sonucunda festival etkinlik organizasyon yöneticisine, konumunun önemi, festival hijyen kuralları, personellerin eğitime alınması gibi konular üzerinden tavsiyelerde bulunmuştur.

Aziz vd. (2019), çalışmalarında 10 adet afet hashtag'i için 32117 adet Twitter verisi toplamışlardır. Bunlar: Deprem, kuraklık, çığ, tsunami, sel, kasırga, volkan, heyelan, orman yangını ve hortumdur. En olumlu ve olumsuz tweetleri görmek için makine öğrenimi algoritması kullanmışlardır. Etkileyici tweetlerden kelime bulutunu çizmişlerdir. Sonuç olarak, afetlerin en çok hâkim olduğu bölgeyi analiz etmişler ve o bölgede yaşayan insanların duygu analizlerini gerçekleştirmişlerdir. Duygu analizi sonucunda halkın %41' lik bir oranla negatif düşüncelere sahip olduğunu vurgulamışlardır.

Khaleq ve Ra (2019) Tweet'lerden afet yönetimi etiketlerini kullanarak bir çalışma yapmışlardır. Kasırga konulu verilerden yola çıkarak tweetleri hazırlık, tepki ve kurtarma olarak sınıflandırmışlardır. Tweet'in konuyla ilgili olup olmadığına karar vermek için bir anahtar kelime listesi oluşturmuşlardır. Analizlerin doğruluğunu ölçmek için ise makine öğrenme yöntemini kullanmayı tercih etmişlerdir. Çalışma sonucunda yaklaşık olarak %85 doğruluk payı bulmuşlardır.

Korkusuz (2019), Twitter üzerinden paylaşılmış olan futbol müsabakaları hakkındaki Türkçe mesajların duygu analizini yapmıştır. Veriler 4 duygu sınıfı kullanarak manuel olarak etiketlenmiştir. Tarafsız, alakasız, olumlu ve olumsuz olarak duygu durumları belirtilmiştir. Farklı sınıflandırma algoritmaları ile modeller çıkarılmış ve bu modellerin başarım oranları bulunmuştur. En iyi sınıflandırma başarım oranının; NB algoritması için %84,30, KNN (K-Nearest Neighbours, K-En Yakın Komşu) algoritması için %87,73 ve SVM (Support Vector Machine, Destek Vektör Makinesi) algoritması için %92,30 olduğu tespit edilmiştir.

Kumar (2020), bu çalışmada 2015 yılında olan Nepal depremi sonrası kültürel mirasa yönelik atılan tweetlerin analizini yapmıştır. 201457 tweet topladığı görülmektedir. Analiz tweetlerinin yaklaşık %4'ünün kültürel mirasla ilgili olduğunu göstermiştir. Kişilerin %89'u zarar gören yapıların tespiti için Twitter'ı kullandığı görülmüştür. Yaptığı çalışma sonucunda kültürel mirasların halktan eşit ilgi görmediğini hasarlı yapıların, depremden etkilenmeyen yapılara göre daha fazla ilgi gördüğünü tespit etmiştir.

Demirci (2020), Twitter'da kullanıcıların cevap etkileşimlerindeki sosyal ağ analizini uygulamıştır. Hedeflenenin, faydalı tweet mesajlarını paylaşarak afet yönetiminde fayda sağlayabilecek belli bir kullanıcı kitlesi ve kategoriler bulmak olduğu görülmektedir. Afet anından önce, afet sırasında ve sonrasında sosyal medya kullanıcılarının davranışını analiz etmiştir. Tweetlerden kelime bulutu oluşturulmuş ve tepkiler analiz edilmiştir. Sonuçlardan, devlet idarecilerinin Twitter hesaplarından afet zamanlarında mesaj gönderilmesi konusunda hızlı ve geniş bir hedefe ulaşabildiği gözlemlenmiştir. Elazığ depremi gibi yaşanan deprem olayları için aynı çalışmaları yaptıklarında, kategorilerin ve aynı kullanıcı hesaplarının afet yönetiminde etkili olduğunu kanıtlamışlardır.

Noor (2020), duygu analiz çalışmasında konu olarak, 2019 Kenya para birimi değişimini seçmiştir. Vatandaşın değişim üzerindeki tepkilerinin analizi hedeflenmiştir. 5 aylık tweet toplama sürecinde toplamda 1087 tweet elde edilmiştir. Negatif, pozitif ve nötr olarak etiketleme yapılmıştır. Sınıflandırma

algoritmaları arasından sadece Multinomial Naive Bayes algoritmasını kullanarak Twitter tweetlerinin duygu analizini yapmıştır. 967 eğitim ve 120 test verilerine göre modelleme yapılmış, unigram modeli ile %70,8, bigram modelinde ise %64,1 doğruluk oranı elde etmiştir.

Ayan (2020), çalışmasında islamofobik tweetlerin sınıflandırılması için gözetimli makine öğrenmesi yöntemini kullanmıştır. İngilizce tweetler arasından Islam, Muslim gibi kelimeleri barındıran 290000 tweet, Twitter API kullanılarak toplanmıştır. Temizlenen veri setinden geriye kalan 162000 tweet, beş kişiden oluşmuş ekip tarafından İslamofobik ve İslamofobik değil şeklinde el ile işaretlenmiştir. Bayes Regresyonu, Ridge Regresyonu ve derin öğrenme modeli olmak üzere 3 farklı model kullanılmıştır. Her üç model için de %95'in üzerinde doğruluğa ulaşılmış. Eğitilen modelin aşırı öğrenme sorununa karşılık, veri seti haricinde yeni 100 adet tweet ile tekrar veri seti oluşturulmuş. Bu tweet veri seti ile yapılan testlerde Bayes Regresyonu'nda %89, Ridge Regresyonu'nda %82, derin öğrenme modelinde ise %68 doğruluğa ulaşılmıştır. Aradaki farkın bu kadar düşüşte olmasının sebebinin ise yeni veri setinin ön işleme adımlarından geçirilmemiş olmasından kaynaklandığı vurgulanmıştır.

Ballı (2021), Türkçe tweetler ile duygu analizi çalışmalarını iki adet veri seti üzerinden yapmış ve çeşitli makine öğrenmesi algoritmalarını kullanmıştır. Veri setlerinden bir tanesi, hazırda bulanan veri setidir. Diğeri ise pandemi, korona gibi seçilmiş kelimelerden oluşan tweetlerin el ile etiketlenmesi gerçekleştirerek oluşturmuş olduğu SentimentSet isimli veri setidir. Veri setlerinin pozitif ve negatif ağırlıkları belirlenmiştir. İki veri seti için de algoritmalar ve LSTM (Long Short Term Memory, Uzun Kısa Süreli Bellek) derin öğrenme ağı kullanılarak eğitimleri yapılmış ve modelleri üretilmiştir. Eğitilmiş olan hazır veri seti ile oluşturulmuş modellerde pozitif tahmin doğruluk oranının daha yüksek olduğu görülmüştür. SentimentSet veri setinde ise negatif tahmin doğruluk oranının daha yüksek olduğu görülmüştür. Bunun sebebinin ise özel seçilen kelimeler negatif olduğu için tweetlerin de genel olarak negatif anlam taşımasından kaynaklandığı belirtilmiştir.

Akdeniz (2021), yaptığı çalışmada Sakarya Büyükşehir Belediyesi'nin ve 14 ilçenin kurumsal Twitter hesapları ve Twitter hesabı içerik analizi ile incelenmesini gerçekleştirmiştir. Belediye ve ilçe hesaplarından 20.10.2020 tarihine kadar atılmış, Twitonomy tweet analiz programından faydalanarak 27058 tweet çekilmiştir. Makine ve derin öğrenmesi yaklaşımlarından biri olan BERT (Bidirectional Encoder Representations for Transformers, Çift Yönlü Transformatör Kodlayıcı Temsilleri) yöntemi ile duygu analizi gerçekleştirilmiştir. Fen İşleri alanında, Sakarya genelinde olumlu, spor ve sosyal hizmetleri hakkında atılan tweetlerde ise olumlu ve olumsuz algının birbirine yakın olduğu, çevre hizmetleri ve kültür işleri hakkında atılan tweetlerde ise olumsuz bir analizin ortaya çıktığı görülmektedir.

Barzenji (2021), tarafından yapılan tez çalışmasında duygu analizi incelemesi, doğal dil işleme ve makine öğrenmesi sınıflandırıcılarını kullanarak yapılmıştır. Analizi 45. Amerika Birleşik Devletleri Eski Başkanı Donald Trump'ın tweetleri üzerinde gerçekleştirmiştir. Veri ön işleme adımlarını uygulayarak az kapasite alma hedefi ile içerik boyutunu küçültmüştür. Sosyal medya platformlarında nefret dolu söylemler fazla olmasından dolayı tweetler için kutuplama yapılmıştır. Her cümle, olumlu, olumsuz veya nötr etiketlerine göre ele alınmıştır. RF (Random Forest, Rastgele Orman) sınıflandırıcı, NB ve SVM gibi makine öğrenme algoritmaları kullanılarak ön işlemden geçmiş veriler eğitilmiştir. Elde edilen eğitim sonuçlarına göre, her sınıflandırıcı için sırasıyla %88, %72 ve %89 oranında başarı elde etmiştir.

Kaşgarlı (2021) , çalışmasında 45. Amerika Birleşik Devletleri Eski Başkanı Donald Trump'ın tweetlerinin duygu analizini yapmıştır. Donald Trump'ın başkanlığını yaptığı 2017 yılından 2020 yılına kadar olan tweetler, yıl bazında veri setleri halinde parçalanmıştır. Duygu analizini tweetlerin negatif ve pozitif olarak ikili sınıflandırma yapılmasıyla gerçekleştirmiştir. Bu sınıflandırmayı yapmak için TextBlob kütüphanesini kullanılmış ve her sözcüğe polatire skoru atanılmıştır. Tweet metin verileri vektörlere dönüştürülerek BoW (Bag of Words, Kelime Çantası) ve TF-IDF (Term Frequency-Inverse Document Frequency, Terim Frekansı-Ters Belge Frekansı) gibi özellik çıkarımı

yöntemlerini kullanmıştır. Makine öğrenmesi algoritmalarından ise SVM, LR (Lojistik Regresyon), NB, RF ve Extreme Gradient Boost algoritmalarını kullanmış ve başarım oranlarını kıyaslamıştır. Doğruluk oranı en yüksek algoritma SVM, en düşük algoritma ise NB olarak sonuçlanmıştır.

Kemaloğlu vd. (2021), yaptıkları çalışmada veri seti için beş farklı sosyal medya platformundan veriler toplanmışlardır. Topladıkları 28189 adet verinin etiketleri uzman tarafından yapılmıştır. Veri setinde 5712 adet pozitif, 11567 adet negatif ve 11247 adet nötr tweet vardır. Çalışmalarında makine öğrenmesi algoritması olarak Lojistik Regresyon ve Rastgele Orman algoritmaları kullanılmıştır. Derin öğrenme algoritması olarak da LSTM kullanılmıştır. Veri seti BoW, TF-IDF ve Kelime İndeksleme kullanılarak ayrı ayrı vektörize edilmiştir. Deneysel sonuçlara göre, derin öğrenme modelinin en iyi performansı gösterdiği gözlemlenmiştir. Doğruluk başarı oranları LR ile %83, RF ile %84 ve LSTM kullanılarak uygulanan modelde ise %84,46'dır.

Kündüm (2021), yaptığı çalışmada kelime gömme modelleri ile geliştirilmiş, geleneksel makine öğrenme algoritmaları ve geleneksel derin öğrenme algoritmaları kullanarak Türkiye'deki büyük üç operatörlerden olan Vodafone, Türk Telekom ve Turkcell şirketleri üzerinden müşteri memnuniyeti tahminlemesi yapmıştır. Twitter platformundan ilgili hashtag'ler ile 3 aylık Türkçe metinleri toplayıp bir veri seti oluşturmuştur. Bu veri setinden %80 ve %97 arasında doğruluk oranı sonucuna varılmıştır. Oluşturulan modellerin kıyaslamasında ise kelime gömme yöntemlerinin, geleneksel derin öğrenme algoritmalarına kıyasla daha iyi sonuçlar ortaya koyduğu gözlemlenmiştir. Gerek doğa olayları gerek teknik arızalardan kaynaklanan olumsuz yorumlara bakıldığında ise, Vodafone operatörünün diğer operatörlere göre daha az bahsinin geçmesi nedeniyle kaliteli hizmet sergilediği öngörülmüştür.

Arı (2022), tüketici yorumlarının fayda düzeyinin tahminlenmesine yönelik bir araştırma gerçekleştirmiştir. Ürüne yapılan yeni bir değerlendirmenin tüketici onayına sunulmadan faydalı bir değerlendirme olup olmadığının tahminlenmesi hedeflenmiştir. Hepsiburada, çevrimiçi alışveriş platformundan satılan Xiaomi

Redmi Airdots Basics 2 TWS Bluetooth 5.0 Kulaklık ürününe ait sadece 1 ve 5 yıldız derecelendirmesine sahip tüketici değerlendirmeleri alınarak veri seti oluşturmuştur. 6 farklı denetimli makine öğrenmesi algoritmaları kullanılmıştır. Test sonuçlarında, en iyi performans %94 doğruluk oranıyla DT (Desicion Tree, Karar Ağacı) algoritmasına ait olduğu ve 39 değerlendirmenin 36'sını doğru bir şekilde faydalı olarak tahmin ettiği görülmüştür. En düşük performans gösteren algoritmanın ise NB olmuştur. Tahmin hata matrisine göre faydalı tahminlerin %86'sının, faydalı değil tahminlerinin ise %97'sinin doğru olduğu görülmektedir.

Ertoy (2022), çalışmasında Twitter kullanıcıları tarafından İngilizce olarak seçilen Kovid19 aşılarına ilişkin tweetlere göre kamu duygu durum algısı analiz edilmiştir. Kullandığı modeli Colab ortamında geliştirmiştir. Vader (Valence Aware Dictionary and Sentiment Reasoner) duygu analiz aracını ve BERT makine öğrenmesi yöntemini kullanmıştır. BERT algoritması, Colab platformunda 40 bin tweetin eğitim aşamasının ortalama 8 saat sürdüğü görülmüştür. Eğitim aşamasını 2000 tweet ile 2 saat 10 dakikalık bir çalışma ile gerçekleştirmiştir. Tahminleme aşamasında ise 212000 tweeti içeren veri seti ile sonuç alınamadığı için veri setinde azaltmaya gidilerek 3109 tweet ile 3 saat süren çalışma sonunda sınıflandırma gerçekleşmiştir. Tüm bunlar göz önüne alındığında bu çalışma için en verimli başarım gösteren analiz yönteminin Vader olduğu ispatlanmıştır.

Sham ve Mohamed (2022), yaptıkları çalışmada iklim değişikliği konusunu ele almışlardır. Twitter platformundan topladıkları verilerin analizini sözlük tabanlı, makine öğrenimi tabanlı yaklaşımlar ve aralarındaki hibritleştirmeyle gerçekleştirmişlerdir. 4 farklı veri seti kullanmışlardır. Kullandıkları sözlükler; TextBlob, SentiWordNet, SentiStrength, Vader, MPQA, Hu ve Liu ve WKWSCI' dır. Kullandıkları makine öğrenmesi algoritmaları ise LR, SVM ve NB algoritmasıdır. Analiz sonucunda en etkili başarımın, TextBlob ve LR hibrit yaklaşımıyla %75,3'lük orana sahip olduğu görülmektedir.

Akdeniz (2022), tez çalışmasında korona virüs salgını ile Mart 2020'den itibaren başlayan uzaktan eğitim sürecine yönelik duygu analizi gerçekleştirmiştir. Sosyal medya platformu olan Twitter üzerinden uzaktan eğitim konulu Türkçe tweetler toplanmıştır. Ön işleme gerçekleştirilmiş ve zemberek kütüphanesi ile normalleştirilme yapılmıştır. Veri seti el ile etiketlenmiştir. İngilizce tweetler çevrilerek duygu analizi için TextBlob, Vader ve BERT ile modeller oluşturulmuştur. Türkçe metinlerin el ile etiketleme sonucunda TF-IDF – LR ikili yöntemiyle %79'luk bir sınıflandırma başarısı elde edilmiştir. El ile etiketlenmiş nötr tweetler analize dahil edilmediğinde ise başarı oranının arttığı ve BoW – LR ikili yöntemin %84'lük oranla en iyi sonucu verdiği görülmüştür.

Dankhara (2022), tarafından yapılan çalışmada veri seti olarak içerisinde 400000 tweet bulunan Sentiment.140 kullanılmıştır. Sözlük tabanlı yaklaşım ve KNN, NB, RF ve SVM algoritmaları kullanılmıştır. Sözlük tabanlı yaklaşım ile %59,5 doğruluk başarım oranı, makine öğrenmesi yaklaşımı ile en yüksek doğruluk başarım oranı; eğitim verisi ile RF %98,8 ve test verisi ile SVM %76,7 olarak elde edilmiştir.

3. MATERYAL VE YÖNTEM

Kuraklıkla ilgili sosyal medya mesajlarının duygu analizine yönelik yapılan bu tez çalışmasında, duygu analizinin yapılabilmesi için sosyal medya platformu olan Twitter aracılığıyla Türkçe tweetler elde edilerek veri seti oluşturulmuştur. Analizin tüm aşamalarında Python programlama dili kullanılmıştır. Programın kod geliştirilme aşamaları ise Anaconda dağıtımı içerisinde yer alan JupyterLab, Spyder ve VSCode geliştirme ortamları ile gerçekleştirilmiştir. Duygu analizi için kullanılan sözlük tabanlı yaklaşım, kelime gömme metotları ve makine öğrenmesi algoritmaları aşağıdaki bölümlerde açıklanmaktadır.

3.1. Doğal Dil İşleme

Modern teknolojinin durmaksızın gelişmesiyle birlikte çeşitli problemler ve bu problemlerin çözümüne yönelik çeşitli ihtiyaçlar doğmuştur. Bu doğrultuda yeni araştırma alanları ve bilgisayar bilimleri ortaya çıkmaktadır. Doğal Dil İşleme'nin (DDİ, Natural Language Processing) ana işlevi, doğal bir dili çözümleme, anlama, yorumlama ve üretme olan bilgisayar sistemlerinin tasarımını ve gerçekleştirilmesini konu alan bir bilim ve mühendislik alanıdır. Bir başka ifade ile DDİ, dillerin ses, biçim, dizilim ve anlam gibi ana başlıkları çözümleyerek dil yapı ve kuralları bilgisayar yardımıyla denetleyen inceleyen yazılımlar olduğu öne sürülmektedir (Delibaş, 2008). Bir bilgisayar bilimi olan bu konu, 1950'li yılların başlarında yapay zekânın bir alt dalı olarak meydana gelmiştir. Araştırmacıların, geliştiricilerin ve yapılmış olan uygulamaların ürettiği başarıları sonucunda artık bilgisayar bilimlerinin ana disiplini olarak kabul görmektedir (Oflazer ve Bozşahin, 2006).

DDİ kısaca; web siteleri, sosyal medya uygulamaları, epostalar, anketler, dergi makaleleri, gazeteler, edebi metinler, kitaplar gibi kaynaklardan elde edilen doğal dil ile oluşturulmuş metinlerle ilgilenen bir alan olarak tanımlanabilir (Korkusuz, 2019).

DDİ alanındaki temel amaçlar şu şekildedir (Delibaş, 2008):

• Doğal dillerin işlevini ve yapısını daha iyi kavramak,
• Bilgisayarlar ile insanlar arasındaki köprü işlevi olarak doğal dil kullanmak ve bilgisayar ile insan arasındaki iletişimi kolaylaştırmak,
• Bilgisayar kullanarak diller arası çeviri yapabilmek amaçlanmaktadır.

DDİ beş ana seviyede incelenebilir:
• Sesbilim,
• Biçimbilim,
• Sözdizimbilim,
• Anlambilim,
• Kullanımbilim.

Kısaca bu seviyeleri özetlemek gerekirse: Sesbilim, harflerin seslerini, seslerin dil içerisindeki kullanımını inceler. Biçimbilim, sözcük kurumudur. Türetme ve ses değişmesi olarak iki çeşit sözcük oluşturma yöntemi mevcuttur. Sözdizimbilim, sözcüklerin cümle içerisinde diziliminin nasıl olması gerektiğini inceler. Anlambilim, dilin iletişime geçmesini sağlar. Kullanımbilim dilin metindeki kullanımını ele alır ve değişimini inceler. Bir sözcük tek başına kullanıldığında veya bir cümle içindeyken farklı anlamlar ifade edebilir (Oğuz, 2009).

DDİ'nin genel kullanım alanları şu şekilde ifade edilebilir:

• **Soru Cevaplama (Question Answering, OA):** Bilgisayara sorulan soru ya da sorulara cevap vermeyi simgeler. Örneğin, herhangi bir web sitesinde bulunan sohbet, yardım veya destek botlarında kullanılır (Kızılırmak, 2020).

• **Bilginin Çıkartılması (Information Extraction, IE):** Bir metinin analizini sağlayarak içerisinden bilgi çıkarımını sağlar. Örneğin, düğün davetiye metini içerisinden düğün davetinin ne zaman, nerede ve nasıl gerçekleşeceği gibi bilgilerin analizini sağlar (Hamde, 2018).

• **Duygu Analizi (Sentiment Analysis, SA):** Metindeki bir cümlenin olumlu ya da olumsuz olup olmadığının ifadesidir. Örneğin, çok iyi bilinen bir markanın ürün fiyatının gayet uygun olduğunu (yani olumlu) ancak ürün kalitesinin kötü olduğunun (yani olumsuz) analiz edilmesidir (Seker, 2016).

• **Makine Diline Çevrim (Machine Translation, MT):** Başka bir dilden diğer başka bir dile çevrilme işleminin DDİ işlemiyle yapılmasıdır. Örneğin, Google Translate, Yandex Translate' tir (Kızılırmak, 2020).

• **Sözcük Anlamı Açıklaştırma (Word Sense Disambiguation, WSD):** Bir kelimenin sözlük anlamı olduğunu eğer yok ise sözlükte bulunmayan bir kelime olmasının analizinin yapılmasıdır. Örneğin, bugün hava almak için çay kenarına gittik. Burada ki "çay" kelimesi, coğrafi terim olan su birikintisi mi yoksa Türkiye'de yetişen "çay" bitkisinin mi olduğunun analizidir (Akçakaya, 2019).

• **Özetleme (Summarization):** Metinsel herhangi bir yazı içerisinden, konu ile ilgili özet çıkarma işlemidir. Örneğin, bir akademik makale içerisinden, çalışma özetini çıkartmak (Seker, 2016).

Özetle, DDİ'nin tüm yöntemleri ve kullanım amaçlarından yola çıkılarak bazı metinsel işleme durumlarının el ile değil, otomatik olarak yapılabilmesi için DDİ büyük kolaylıklar sağlamaktadır. Bu kolaylıklar da DDİ'ye olan ilginin son 10-15 yıl içinde büyük bir hızla artmasını sağlamıştır.

3.2. Veri Madenciliği

Teknoloji günden güne gelişmekte ve buna bağlı olarak da sosyal medya kullanımı çok yaygın hale gelmektedir. Bu da büyük miktarlarda verileri oluşturmaktadır. Bu büyük miktardaki verilerden faydalı bilgiler veya farklı bir ifade ile anlamlı bilgiler çıkarılması gerekmektedir. Bunu sağlayan süreç ise Veri Madenciliğidir. Veri madenciliği, faydalı her bir bilginin büyük veri tabanlarından otomatik olarak keşfedilme sürecidir (Tan vd., 2006). Veri tabanlarında bilgi keşfi (Knowledge Discovery in Databases, KDD) kavramı

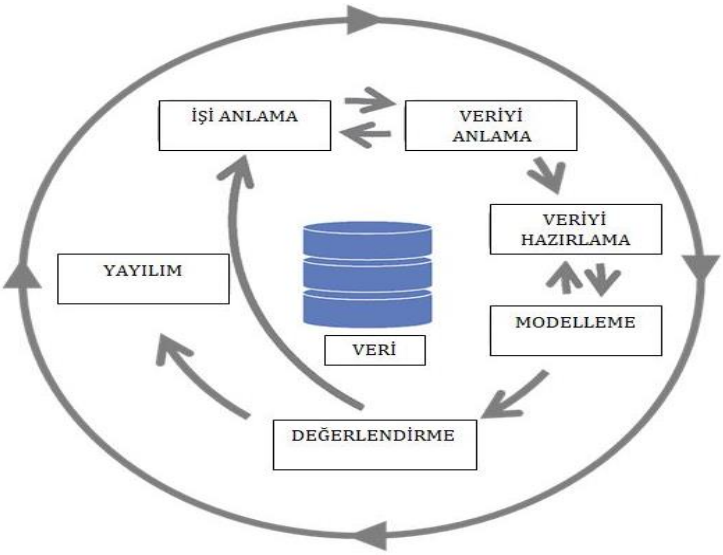
büyük veri tabanlarında kademeli bir şekilde bilginin keşfedilme sürecidir. Veri madenciliği, bu keşif sürecinin en önemli aşamasıdır (Fayyad vd., 1996). Başka bir ifadeyle, veri madenciliği tek başına bir anlam ifade etmeyen veri ya da veriler içindeki gizli örüntüleri ve ilişkileri ortaya koymak için istatistik, makine öğrenmesi ve yapay zekâ gibi yöntemlerin ileri veri çözümleme araçlarıyla kullanılmasını içine alan süreçler topluluğudur (Bozkır vd., 2009).

Veri Madenciliği gün geçtikçe birçok alanda kullanılmaya başlanmıştır (Baykal, 2006). Bunlardan bazıları aşağıdaki gibi maddelendirilmiştir;

- **Pazarlama:** Müşteri hakkında bilgiler, müşteri değerlendirme, satış tahmini vb.
- **Bankacılık:** Kredi/Banka kartı dolandırıcılığı, kredi talep değerlendirilmesi vb.
- **E-Ticaret:** Sunucu saldırı tespiti, kullanıcı üyelik dolandırıcılığı vb.
- **Sigortacılık:** Riskli müşteri tespiti, sigorta dolandırıcılığı vb.

Örnek verilen her alanda oluşan bu karışık ve büyük veriler veri madenciliği sayesinde çok daha hızlı bir süreç içerisinde anlamlı ve faydalı bir bilgiye dönüşebilmektedir.

Veri madenciliğinin uygulanabilmesi için bir süreçten geçmesi gerekmektedir. Aynı zamanda sürecin de kendisidir. Yararlı veriyi ortaya çıkarmanın yanı sıra, veri tabanlarında bilgi keşfi sürecinin gerçekleşmesini sağlamaktadır. Bu süreç Şekil 3.1'de gösterilmiştir.



Şekil 3.1. Veri madenciliği süreci (Hotz, 2022)

Veri madenciliğinde başarıya uzanan yol, işin ve verilerin özelliklerinin çok iyi bilinmesi ve detaylı bir şekilde analiz edilmesinden geçmektedir. Bu süreçte izlenen adımlar şu şekildedir;

• **Problemin Tanımlanması:** Veri madenciliğinde ilk ve en önemli şart, projenin hangi amaç için yapılacağının ve elde edilecek sonuçların başarı düzeylerinin ölçütlerinin tanımlanmasıdır (Ayık vd., 2007).

• **Verilerin Hazırlanması:** Model kurulması sürecinin en önemli kısmıdır**.** Modelin kurulması aşamasında ortaya çıkan problemler, bu aşamaya sürekli geri dönülmesine ve verilerin yeniden düzenlenmesine sebep olacaktır. Bu durum ise analistin enerji ve zamanının %50 ile %85'ini harcamasına neden olmaktadır. Verilerin hazırlanması, "toplama", "değer biçme", "örneklem seçimi", "birleştirme ve temizleme", ve "dönüştürme" gibi 5 aşamadan oluşmaktadır (Ayık vd., 2007).

• **Modelin Kurulması ve Değerlendirilmesi:** Tanımlanmış problem için en doğru modelin seçimi için birden çok modelin kurularak denenmesi gerekmektedir. Bu aşama en iyi olduğunu düşünülen model bulunana kadar devam eden bir döngüdür (Çetin, 2009).

• **Modelin Kullanılması**: Kurulmuş olan model uygulama olabileceği gibi, uygulamalara ait alt parçalar da olabilir (Çetin, 2009).

• **Modelin İzlenmesi:** Bütün sistem özelliklerinde ve üretilen verilerde ortaya çıkan değişiklikler, seçilerek kurulan modellerin daimi olarak izlenmesini ve gerekiyorsa yeniden düzenlenmesini gerektirecektir (Savaş vd., 2012).

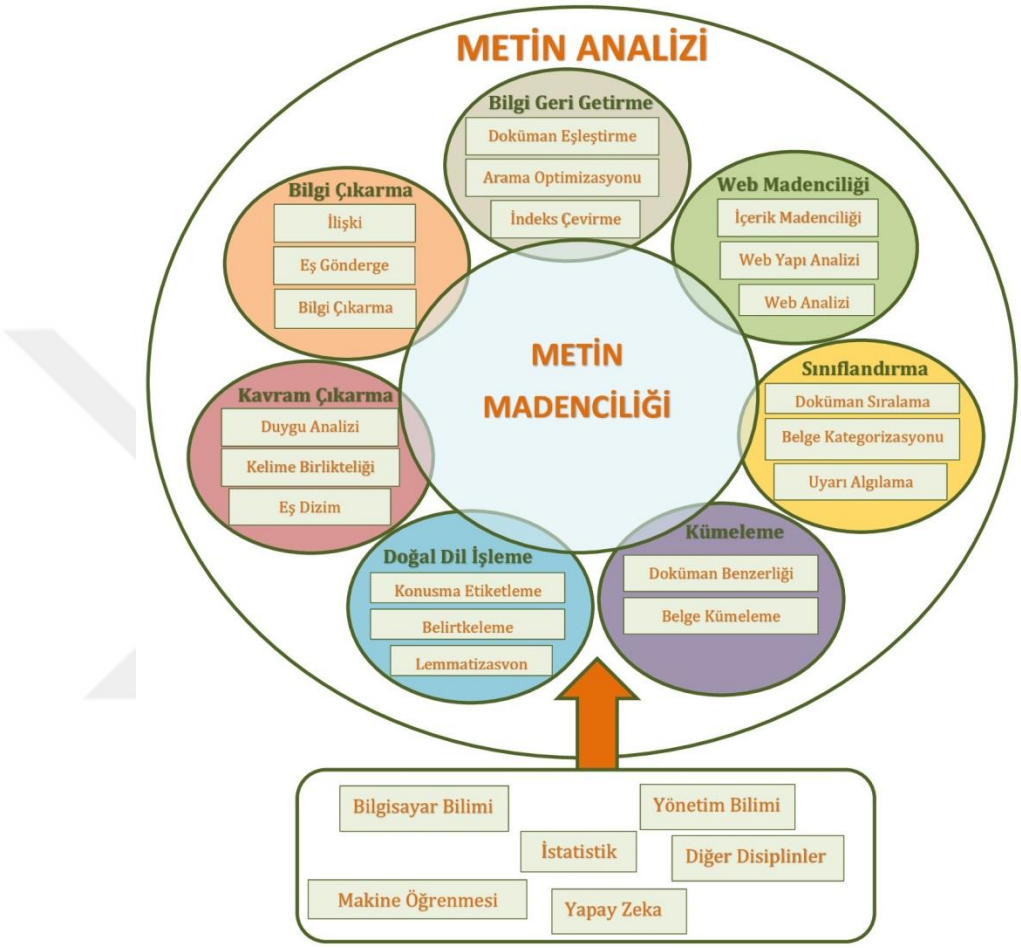
3.3. Metin Madenciliği

Son zamanlarda teknolojideki gelişmeler sonucu hızlı ilerleyen veri madenciliği alanında, elde edilen metin verilerinin miktarının da artmasıyla birlikte bilgilerden veya verilerden anlamlı sonuç çıkarma ihtiyacı da günden güne artmaktadır. Bu ihtiyacı ise metin madenciliği aracılığıyla çözülebilmektedir. Metin madenciliği kavramı yapılanmamış ya da bir kısmı yapılanmış metinlerden anlamlı sonuçlar çıkarabilmek gayesiyle metinlerin belirli bazı süreçlerden geçirildikten sonra yapılandırılmış hale getirilerek çıktılarının analiz edilmesi şeklinde tanımlanır (Karamanlı, 2019).

Başka bir deyişle metin madenciliği, işlenmemiş bir halde bulunan yapısal olmayan DDİ metinlerini, işleyerek karar destek amacıyla değerlendirilmesinde metinlerden anlamlı bilgiler çıkarmaya yarayan yarı otomatik bir analitik süreç olarak ifade edilebilir (Ergün, 2012).

Metin madenciliği, veri madenciliğinin bir alt dalı olarak bilinir fakat bu yöntemlerin birbirinden farklı özellikleri vardır. Aralarındaki temel fark, veri madenciliğinde yapılandırılmış veri tabanları işleme alınır, metin madenciliğinde ise yapılandırılmamış veya yarı yapılandırılmış DDİ metinleri kullanılmaktadır (Hearst, 1999). Diğer farklar ise, metin madenciliğinde birden fazla özellikte farklı biçimlere sahip, belirli bir çerçevede bulunmayan veri seti kullanılır, veri madenciliğinde ise veri tabanında hazır bulunan biçimli veri seti bulunur.

Metin madenciliği teknikleri kullanılarak yapılan yöntemler ve ilişkili olan disiplinler Şekil 3.2'de verilmiştir. Bu şekle göre metin madenciliğinin istatistik, bilgisayar bilimi, yapay zekâ, makine öğrenmesi, yönetim bilimi ve diğer bilimlerin ortak noktası olduğu söylenebilir.



Şekil 3.2. Metin analizi ile metin madenciliği evreni (Anonim, 2015)

Veri madenciliğinin uygulama alanı oldukça geniştir. Bu alanlar içerisinde, Veri Tabanı Sistemleri, Veri Görselliği, Yapay Sinir Ağları, İstatistik, Yapay Öğrenme, vb. gibi disiplinler bulunmaktadır.

3.3.1. Metin Madenciliği Adımları

Şekil 3.3.'de görüldüğü üzere metin madenciliği işlemleri genel olarak beş adımdan oluşmaktadır. Bu adımlar;

• **Metin Koleksiyonu Oluşturma:** Metin madenciliği işleme adımlarını gerçekleştirebilmek için ilk kısım olarak amaca yönelik bir veri kümesinin elde edilmesi gerekmektedir. Yani bu adım, istenilen konularda bilgiye erişilebilecek sistemler kullanılarak metin koleksiyonu elde etme sürecidir. Bu süreç, günümüzde sıkça kullandığımız internet üzerinden, özellikle Google arama motoru, sosyal medya platformları vb. kullanılarak gerçekleştirilmektedir. Koleksiyon için hazır veri kümeleri kullanılabilir ya da özel veri kümeleri oluşturulabilir (Oğuz, 2009).

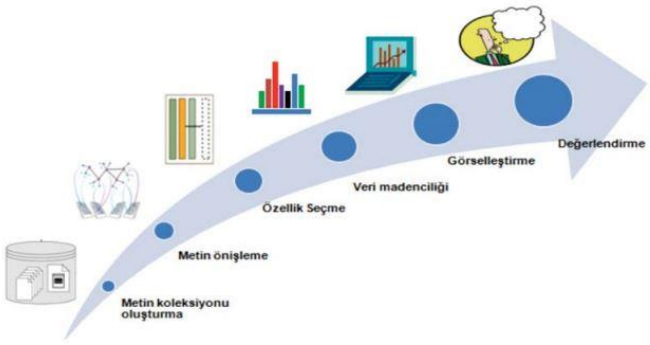
• **Metin Ön İşleme:** Metin madenciliğinin en önemli alt görevidir. Ön işlemenin temel amacı, düzensiz metin verilerini veri madenciliğinde, makine öğrenmesi algoritmalarında kullanılmak üzere uygun bir yapıya dönüştürmektir. Ön işleme, verinin işlenme amacına bağlı olarak çeşitlilik gösterebilen birden fazla alt aşamalara sahiptir (Srividhya ve Anitha, 2010).

• **Özellik Seçme:** Özellik seçme adımında, ön işleme adımından geçen metinlerdeki önemli kelimeleri belirleme (isimler, sayılar, tarihler, kısaltmalar vb.) ve ilişkisi bulunmayan özelliklerin çıkarılması (yalnızca bazı dokümanlarda gözlemlenen özelliklerin çıkarılması, birden fazla dokümanda gözlemlenen özellikleri azaltma vb.) işlemleri yapılmaktadır (Ergün, 2017).

• **Veri Madenciliği:** Düzensiz biçimdeki metinlerden yapılandırılmış bir biçime dönüştürülen metinlerin geleneksel veri madenciliği yöntemleriyle analizi sürecidir (Aydın, 2019).

• **Görselleştirme:** Elde edilen analiz sonuçların kullanıcıya sunumunda en etkileyici ve anlaşılır görselleştirmenin yapılması aşamasıdır (Aydın, 2019).

• **Değerlendirme ve Yorumlama:** Veri madenciliği uygulama adımları ile verilerin analizinden elde edilen sonuçların değerlendirilip kullanıcıya göre uygun ve anlaşılır bir şekilde sunulması, aktarılması işlemidir (Ergün, 2017).



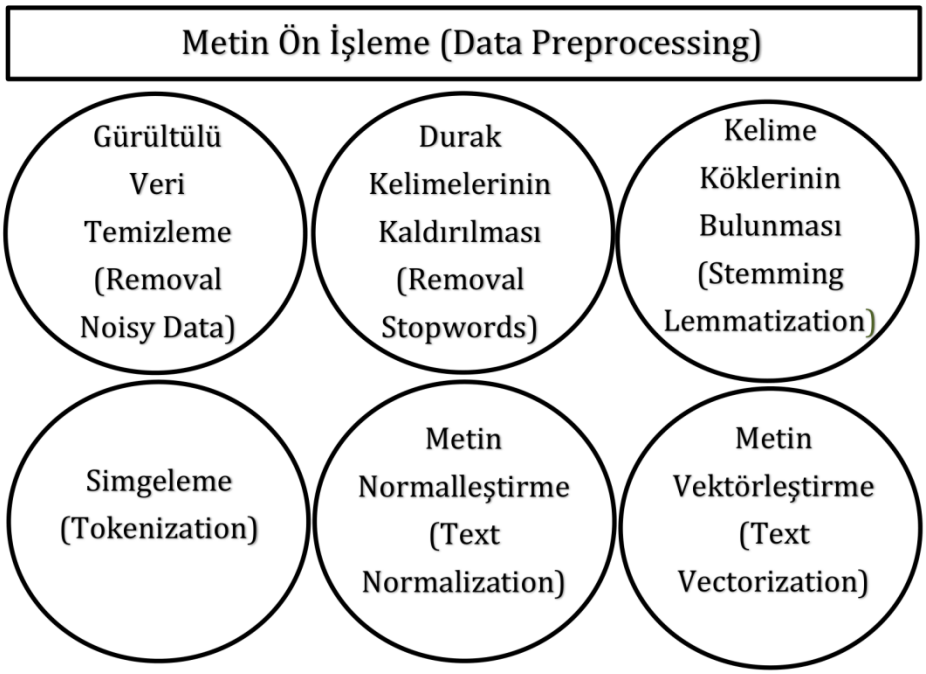
Şekil 3.3. Metin madenciliği süreci (Oğuz vd., 2007)

3.3.2. Veri Ön İşleme

Veri kümesi oluşturma aşamasında elde edilen metinleri kullanabilmek için ilk olarak metinlerin düzenlenmesi gerekmektedir. Oluşturulan modelin başarısı genellikle öğrendiği verinin kalitesine ve ne kadar doğru ve verimli bir ön işlemden geçtiğine göre değişkenlik gösterir (Coşkun ve Baykal, 2011).

Yapılandırılmamış, düzensiz metin veri setlerinin, metin madenciliği aşamalarında işlenerek yapılandırılmış hale getirilebilmesi işlemine ön işleme denilmektedir. Bu aşamada, metinlerdeki gereksiz kelimeleri çıkarma, kısaltma olan kelimeleri doğru kelimeler ile eşleştirme, kelimeleri sözcük öbekleri halinde bölme, kelimenin köklerini bulma, imla ve yazım hatalarını düzeltme vb. işlemler uygulanarak veri seti analiz edilebilir hale getirilir (Çınar, 2020).

Metinler üzerinde yapılacak ön işleme adımları, çalışılacak amaca göre farklılıklar gösterebilir. Genel metin ön işleme adımları Şekil 3.4'de gösterilmektedir. Bu adımları kısaca açıklayalım.



Şekil 3.4. Metin ön işleme

• **Gürültülü Veri Temizleme:** Metnin içerisinden anlamlı olmayan ve istenmeyen verilerin çıkarılması işlemidir. Bu veriler, web sitesi linkleri, emojiler, hashtag'ler, konu başlıkları, çeşitli semboller, noktalama işaretleri bu verilere örnek olarak verilebilir**.** Büyük küçük harf dönüşümü de bu kısımda gerçekleştirilir (Contractor vd., 2010).

• **Durak Kelimelerinin Kaldırılması:** Duraklama kelimeleri, metin içinde metinin anlamını değiştirmeyen fakat Türkçe dil yapısından dolayı kullanılan bazı kelimelerdir. Türkçede bağlaçlar bunlara örnektir. Durak kelimelere hepsi, bir, bazı, sadece, biz, bunlar vb. örnek verilebilir. Bu kelimelerin kaldırılması DDİ'nin hangi amaca yönelik olduğuna göre değişkenlik gösterir. Duygu analizi konusunda, cümlelerden kaldırılacak duygu anlam ifadesini bozan ve yanlış sınıflandırmaya sebep olan kelimeler de duraklama kelimeleri olarak sayılabilir (Ballı, 2021).

• **Kelime Köklerinin Bulunması:** Lemmazation, metindeki kelimeleri morfolojik analizini dikkate alarak köklerine indirgeme işlemidir. Kök bulmak için her zaman sözlüğe ihtiyaç duyulur. Kelimeler bu sözlükteki karşılıklarına

göre sadeleştirilirler. Kelimelerin en sade hallerinin sözlükteki karşılığı bulunarak alınır. Örneğin, alıngan kelimesinin kökü almak kelimesidir, gidiyorlar kelimesinin ise gitmektir (Noyan, 2019).

• **Simgeleme:** Veri setindeki cümlenin istenilen özelliklere, kurallara göre daha küçük anlamlı parçalara bölünmesi işlemine denir. Cümle kelimelere, deyimlere ve söz öbeklerine ayrılır. Örneğin, "Sosyal medya platformlarının kullanıcıları herhangi bir konuda düşünce ve fikirlerini özgürce ifade edebilmektedirler." Kelimelere ayrılan bu cümle şu hale gelir; ['Sosyal', 'medya', 'platformlarının', 'kullanıcıları', 'herhangi', 'bir', 'konuda', 'düşünce', 've', 'fikirlerini', 'özgürce', 'ifade', 'edebilmektedirler', '.']. Bu cümle sonrasında, duraklama kelimelerinin kaldırılması ve kök bulma işlemlerinin yapılması gerekmektedir (Eliöz, 2021).

• **Metin Normalleştirme:** Veri ön işleme aşamalarından olan fakat çoğunlukla uygulanması es geçilen adımlardan biri metin normalleştirmesidir. Metin normalleştirme, basit bir şekilde ifade edilirse metnin standart bir forma dönüştürülmesidir. Örneğin, Türkçe atılmış bir tweette "gzel", "güüzeeel" kelimelerinin kurallı formu olan "güzel" kelimesine dönüştürülmesidir (Anonim, 2022).

• **Metin Vektörleştirme:** Metin Vektörleştirme, veri madenciliği alanında çalışma yapılırken metin halindeki doğal dili bilgisayarın anlayıp yorumlayabileceği sayısallaştırma yani veri setlerini vektörler haline getirme işlemine denir (Anonim, 2022).

3.4. Duygu Analizi

Duygu analizi, DDİ, bilgisayar bilimleri, istatistik vb. alanlardan bazı yöntem ve belirli tekniklerin kullanılması ile duygu ifade sahibinin, metin içerisinde belirttiği duygu, fikir, ifade gibi kişiye ait öznel bilgilerin belirlenmesini hedefleyen en yeni araştırma alanlarındandır (Onan, 2017). Günümüzde ürün ve hizmet kıyaslamalarında, AR-GE dönüş bildirimlerinde, siyasi konularda, film

yorumlamalarında, marka kullanım deneyimlerinde, eğitim içeriklerinde vb. durumlarda duygu analizi sık sık tercih edilmektedir.

Duygu analizi, ilk olarak 2000'li yılların başlarında bir araştırma alanı olarak hayatımıza girmiştir. Ancak 90'lı yıllarda duygu türleri, bakış açıları ve etkileri gibi konular üzerine çalışmalara rastlanmıştır. Nasukawa ve Yi'nin çalışmalarında ilk kez terim olarak karşımıza çıkmıştır (Nasukawa ve Yi, 2003). Duygu analizi terimi 2003 yılında ortaya çıkmış olmasına rağmen bu alandaki çalışmalar daha önceki yıllara dayanmaktadır. 2000 yılında Vasileios ve Janyce (Vasileios ve Janyce, 2000), 2001 yılında Tong ve arkadaşlarının (Tong vd., 2001), 2002 yılında Turney'in (Turney, 2002), 2002 yılında Pang ve arkadaşlarının (Pang vd., 2002) yaptıkları çalışmalar ilk çalışmalardan sayılmaktadır (Özyurt ve Akçayol, 2018).

Duygu analizi metni olumlu, olumsuz ve nötr olarak sınıflandırmayı amaçlamaktadır. Bu sınıflandırma işlemi sadece negatif ve pozitif olarak da yapılabilmektedir. Duygu analizi yapılırken terimlerin sözlük içerisinde karşılık geldiği puanlar bulunmaktadır. Her sözcüğün -1 ile +1 arasında bir değeri vardır. Cümle içerisindeki kelimelerin sözlükte karşılık gelen değerlerinin toplamından oluşan sonucuna polarite skoru denilmektedir, bu skor ile de duygu tanımlaması yapılabilmektedir (Bhadane vd., 2015).

Duygu analizi, genellikle metni nötr, olumlu ve olumsuz olarak sınıflandırmayı amaçlamaktadır. Sınıflandırma işlemi her zaman 3 duygu durumu üzerinden yapılmayabilir. Sadece negatif ve pozitif olarak da yapılabilmektedir. Duygu analizi yapılırken terimlerin sözlükte karşılık geldiği puanlar bulunmaktadır. Sözlükte her sözcüğün -1 ile +1 arasında bir değeri vardır. Cümle içerisindeki kelimelerin sözlükte karşılık gelen değerlerinin toplamından oluşan sonucuna polarite skoru denilmektedir, bu skor duygu tanımlaması yapılırken kullanılmaktadır (Bhadane vd., 2015). Metinlerin altında atan duyguları, duygu analizi kullanarak analiz edilebilir. Örneğin, sosyal medyada platformlarında paylaşılan ifadeler duygusal olarak anlamdırılmamış verilerdir ve bu ifadeleri anlamsal bir etiketle duygusal analizi yapılabilir. Duygular verilerin içinde

gizlidir. Çizelge 3.1'de örnek cümlelerde metin içinde keşfedilen duygular verilmiştir.

Çizelge 3.1. Örnek metinlerde gizli duygu keşfi

Örnek Cümle	Duygu Keşfi
İzlediğim film çok sürükleyiciydi.	Pozitif Duygu
Bu markanın ürünleri hep bozuk çıktı.	Negatif Duygu
Bugün ders yoktu.	Nötr Duygu

Duygu analizi yöntemiyle, herhangi bir dokümanın bütün olarak ele alınabildiği gibi, dokümanların cümlelere ayrılıp, her biri cümle düzeyinde de incelenebilir. Bunlar göz önüne alınırsa duygu analizi yönteminin bütüne ve parçaya odaklanabilme yetisi vardır denilebilir. Diğer bir analiz durumu ise cümlelerin kendi içlerindeki duyguların tanımlanarak varlık veya özellik olarak da duygu analizinin yapılmasıdır (Akdeniz, 2021). Duygu analizi uygulama aşamalarında kullanılan 3 farklı sınıflandırma seviyesi vardır. Bunlar: doküman seviyesi (document-level), cümle seviyesi (sentence-level) ve hedef tabanlı(aspectbased) duygu analizidir.

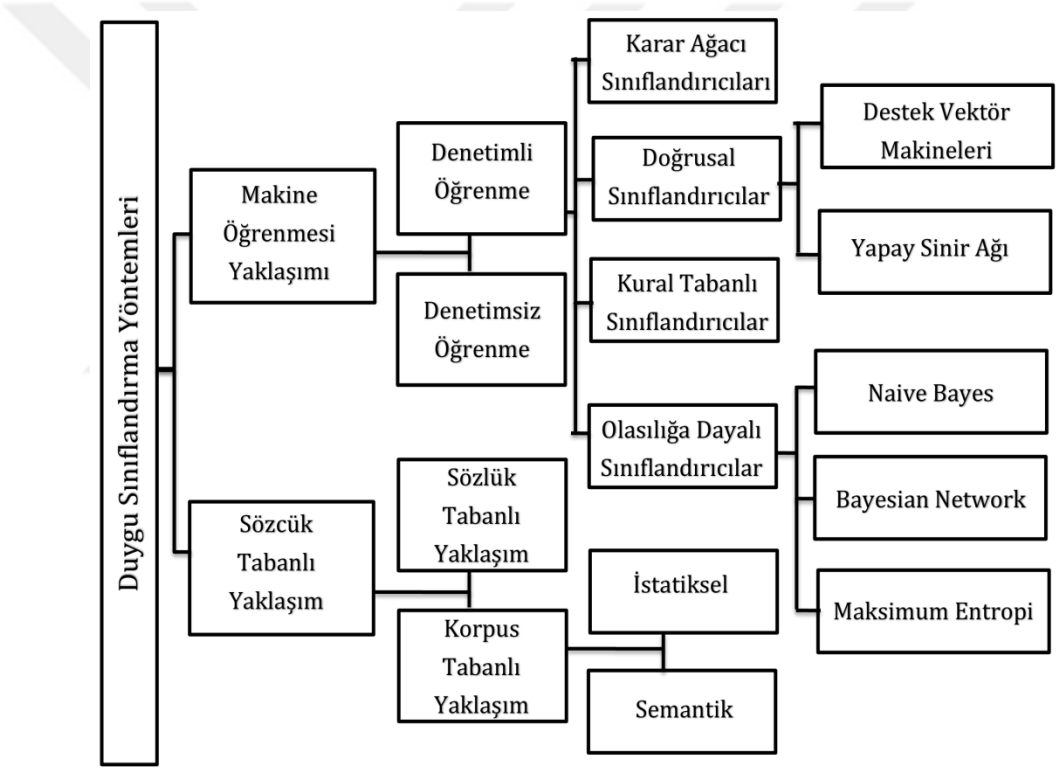
• **Doküman Seviyesi Sınıflandırma**: Dokümanın tamamının işleme alındığı bir yöntemdir. Doküman içeriğinde farklı duygular bulunsa dahi, doküman analiz sonucunda çok daha geniş kapsamlı tek bir duygu (pozitif/negatif/nötr) ile etiketlenmektedir.

• **Cümle Seviyesi Sınıflandırma**: Dokümanı cümleler halinde inceler ve cümleye yönelik duygu analizini gerçekleştirmeyi hedefler. İlk olarak cümlenin objektif ya da sübjektif olup olmadığının belirlenmesi gerekir. Eğer cümle sübjektif ise yani öznel bir cümleyse, cümlenin olumlu veya olumsuz olduğu belirlenir (Sarıman ve Mutaf, 2020).

• **Hedef Tabanlı Sınıflandırma**: Doküman ve cümle tabanlı sınıflandırma seviyelerine göre yaklaşımı biraz farklıdır. Varlığın tüm yönleriyle ele alınmasını

amaçlar. Bir nesnenin belli özelliklerine göre sınıflandırma yapılır. Bir nesnenin hangi bakış açısına göre olumlu ya da olumsuz olduğu önemlidir (Özdeş, 2017).

Duygu analizi sınıflandırma teknikleri Şekil 3.5'de görüldüğü üzere makine öğrenmesine dayalı yöntemler ve sözlük tabanlı yöntemler olarak iki temel sınıfa ayrılmaktadır. Makine öğrenmesine dayalı yöntemlerde algoritma teknikleri kullanılmaktadır. Sonucunda ki istatiksel değerler karar vermeye destek için kullanılır. Sözlük tabanlı yöntemler ise bir duygu sözlüğüne dayanır, bilinen ve önceden derlenmiş duygu terimlerinin toplanması ile duygu analizi gerçekleştirilir (Medhat vd., 2014).



Şekil 3.5. Duygu sınıflandırma yöntemleri

3.4.1. Sözlük Tabanlı Yaklaşım

Sözlük tabanlı yaklaşım, metnin duygu analizini sözlük kullanılarak gerçekleştirdiği bir yöntemdir. DDİ teknikleri kullanılmaktadır. Duygu analizi yapılacak olan metin için daha önceden oluşturulmuş anlamsal sözcükler içeren ve bu sözcüklerin polarite değerleri olan sözlük tercih edilir. Bu analizde

yönteminde, metnin içerisindeki cümle terimlere ayrılır ve her bir terim duygu sözlüğünde aranır. Sözlükte bulunursa, duygu polarite skoru toplam skora eklenir. Sonuçta elde edilen toplam duygu skoru, sıfıra eşit olursa nötr, sıfırdan büyükse pozitif, küçükse de negatif olarak sınıflandırılır (Sağlam vd., 2019).

Duygu sözlükleri incelendiğinde dünyada kullanımı yaygın olan ve Türkçe metinlerin analizi için düzenlenmiş çeşitli açık kaynaklı duygu sözlükleri olduğu görülmektedir. Bunların bazıları Vader, Textblob, WordNet, SentiWordNet, SentiTurkNet ve SWNetTR++' dır.

3.4.1.1. Vader

Açılımı Valence Aware Dictionary and Sentiment Reasoner (Duygu Akıl Yürütme için Değer Duyarlı Sözlük) olan Vader, sözlük ve kural tabanlı bir sözlük modelidir. MIT lisansına sahip açık kaynaklıdır. Metinler üzerinde doğrudan işlem yapabilen ve polarite duygu sınıflarına ilave olarak polarite skoru yoğunluğunu da hesaba katarak daha hassas bir analiz imkânı sağlar. Vader nötr, negatif, pozitif ve bileşik (compound) olarak adlandırılan özellik değeriyle birlikte 4 çıktı vermektedir. Polarite skor yoğunluğu -4 ile +4 arasında değişmektedir. Aşırı olumsuzdan aşırı olumluya göre duygu ifade edilir (Kaşgarlı, 2021).

3.4.1.2. TextBlob

TextBlob, Natural Language Toolkit (NLTK) üzerindeki bir Python kütüphanesidir. TextBlob, her kelime için ne kadar olumlu ya da ne kadar olumsuz olduğunu veren duygu polarite skoru ve öznellik puanı olmak üzere iki çıktıya sahiptir. Polarite skoru -1 ile +1 arasında değişmektedir. -1 çok olumsuz, +1 çok olumlu olarak ifade edilmektedir. Bir kelimenin ne kadar nesnel ya da öznel olduğunu öznellik skoru vermektedir. Bu skor 0 ile +1 aralığındadır. Bu değer 0'a yakınsa nesnel, +1'e yakınsa özneldir (Akdeniz, 2022).

3.4.1.3. WordNet

Princeton Üniversitesi tarafından oluşturulmuş büyük bir İngilizce sözcüksel veri tabanıdır. WordNet, kelimeyi fiiller, isimler, sıfatlar ve zarflar gibi kategorilere ayırır. Veri seti 155287 İngilizce kelime, 175979 eş anlamlı kelimeler ve 207016 kelime-anlam çiftlerinden oluşmaktadır (Kaşgarlı, 2021).

3.4.1.4. SentiWordNet

SentiWordNet, WordNet üzerinden geliştirilmiş bir duygu sözlüğüdür. SentiWordNet'in içerdiği her bir eş anlamlılık kümesi, pozitif, negatif ve nötr değerlere sahiptir. Değerler 0 ve 1 arasında değişir ve bu değerlerin toplamı da 1'e eşittir. SentiWordNet'te toplam 117659 adet anlamdaş kelimeler kümesi ve bu kümeler açıldığında da toplamda 206941 kapasiteli farklı anlama sahip terim mevcuttur. Bir terimin WordNet'te farklı eş anlam kümelerinde bulunabilmesi, bu terimin SentiWordNet'te de farklı duygu skorlarına ve yönüne sahip olabilmesi SentiWordNet'i güçlü kılan bir diğer özelliktir (Ucan vd., 2016).

3.4.1.5. SentiTurkNet

Dehkharghani vd. tarafından oluşturulmuş SentiTurkNet sözlüğü, 3 farklı İngilizce dil kütüphanesinden faydalanarak oluşturulmuştur. Türk dili duygu analizi çalışmaları için geliştirilmiş bir sözlüktür. Bu sözlüğün içerisinde, 14795 tane eş anlamlı kelime yer almaktadır. Duygu kelime skorlarının bulunduğu SentiTurkNet adlı bu sözlük ilk duygu değeri içeren sözlük kaynağı olarak literatürde yer almaktadır (Dehkharghani vd., 2016).

3.4.1.6. SWNetTR++

Sağlam vd. tarafından geliştirilmiş bu sözlük, yaklaşık olarak 49000 Türkçe kelime ve kelime gruplarına ait eğilimleri gösteren polarite ve yoğunluk değerlerini içermektedir. SWNetTR++ sözlüğüne göre bir kelimenin polaritesi, pozitif duygularda 1, negatif duygularda ise -1 olmaktadır. Kelimelerin yoğunluk

değerleri ise [-1, +1] aralığında sürekli değerler almakta ve duygunun gücü yani yoğunluğu hakkında daha fazla bilgi vermektedir**.** Türkçe dili için geliştirilmiş en geniş kelime haznesine sahip sözlüktür (Sağlam vd., 2019).

3.4.2. Makine Öğrenmesi Yaklaşımı

Makine öğrenmesi, Nvidia'ya göre veriyi ayrıştırmak, veriden öğrenmek ve ardından bir tahminleme yapmak için algoritmalar kullanma uygulamasıdır. İlk olarak 1950'nin başlarında yapay zekâ olarak ortaya çıkmış ve ardından makine öğrenmesi sonrasında ise derin öğrenme ile devam eden bir süreç bulunmaktadır (Nvidia, 2019).

Makine öğrenmesi yaklaşımı, makine öğrenmesi algoritmalarını ve dilbilimsel özellikleri uygulamaktadır. Makine öğrenmesinin doğru tahminler yapabilmeyi öğrenmek gibi bir amacı vardır. Bunu da önceki gözlemlere dayanarak otomatik teknikler geliştirip uygulayarak yapar.

Metinler belirli boyutlara ulaştıktan sonra el ile sınıflandırmak zor ve zaman alıcı bir iş haline geldiğinden makine öğrenmesi teknikleri kullanmak daha hızlı, avantajlı ve güvenilir durumdadır. Makine öğrenmesinin birçok alanda işe yarayan başarılı sonuçlar vermesi, DDİ alanı için de kullanılmasını yaygınlaştırmıştır. Veri setlerini makine öğrenmesinde kullanabilmek için etiketli ve etiketsiz olmak üzere iki farklı türden oluşması gerekmektedir. Algoritmayı eğitmek için etiketli veri seti, eğitilmiş algoritmayı test etmek için de etiketsiz veri kullanılmaktadır (Elmas, 2019). Makine öğrenmesi tabanlı yaklaşım genel olarak denetimli veya denetimsiz öğrenme/sınıflandırma olmak üzere iki gruba ayrılır.

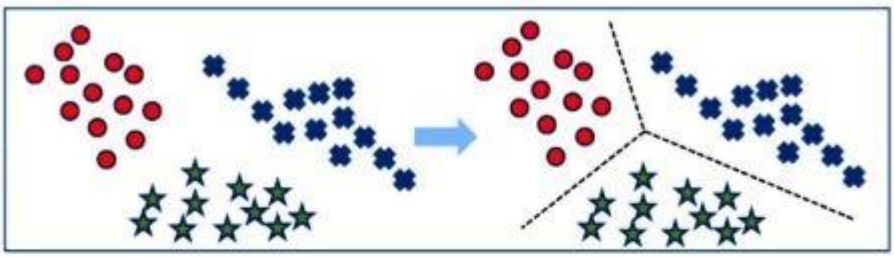
3.4.2.1. Denetimli Makine Öğrenmesi

Denetimli makine öğrenmesi etiketlenmiş eğitim verisi ile yapılan sınıflandırmadır. Denetimli sınıflandırma algoritması, eğitim kümesindeki etiketlerden yola çıkarak alakalı olduğu sınıf bilgisini öğrenir ve bir model

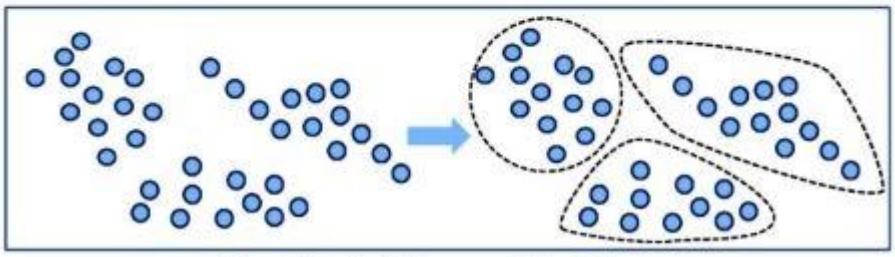
oluşturur. Oluşan bu model sınıf ve etiket verisine sahip olmayan test örneklerini, eğitim kümesinden öğrendiği modele göre sınıflandırır (Çoban, 2016). Tümör örneğinden yola çıkılırsa iyi huylu mu yoksa kanser tümörü mü olduğuna karar veren sınıflandırma işlemi bir denetimli sınıflandırma örneğidir. İlgili makine öğrenmesi algoritması, veri setinde bulunan kanser olan ve iyi huylu olan tümör bilgilerinden yola çıkarak bu iki sınıfın özelliklerine göre gerekli sınıfları öğrenir ve gelen bir hasta tümörünün hangi tür olduğuna karar verir.

3.4.2.2. Denetimsiz Makine Öğrenmesi

Denetimsiz makine öğrenmesi tekniğinde, model etikete sahip olmayan veriler kullanılarak eğitilir. Denetimsiz öğrenmenin hedefi sınıflandırma olamaz, çünkü eldeki verilerin herhangi bir sınıf ve etiket bilgisi yoktur. Bu model genellikle kümeleme, özniteliklerin birbirleriyle olan ilişkilerinin belirlenmesi, yoğunluk tahmini gibi amaçlar için kullanılmaktadır. Denetimsiz öğrenme algoritmasından elde edilen sonuçlar denetimli öğrenme için de kullanılabilmektedir. Denetimli ve denetimsiz makine öğrenmesi modellerinin kullanım amaçları ve kümeleme işlemi arasındaki fark Şekil 3.6'da gösterilmiştir (Başkaya, 2017). Görüldüğü üzere denetimli öğrenmede farklı etiketlere sahip direkt veriler üzerinde bir sınıflandırma söz konusu iken; denetimsiz öğrenmede veri sınıfı belli olmadığından amaç sınıflandırma değil veriyi kümelemektir.



Denetimli Öğrenme Modeli



Denetimsiz Öğrenme Modeli

Şekil 3.6. Denetimli öğrenme ve denetimsiz öğrenme farkı (Başkaya, 2017)

3.5. Twitter

Twitter 2006 yılında kurulmuş anlık paylaşım yapmayı sağlayan ve insanların ortak ilgi alanları doğrultusunda diğer insanlarla etkileşime geçtiği bir sosyal medya platformudur. Twitter' daki ilk tweet kurucuları kabul edilen Jack Dorsey ve Biz Stone tarafından atılmıştır. Twitter'ın asıl amacı kullanıcılarına kısa ve anlamlı mesajlar kullanarak kendilerini ifade etmeleridir (Ayan, 2020).

Dünya üzerinde en çok kullanılan sosyal medya platformları araştırıldığında Twitter ön plana çıkmaktadır. Kullanıcıların oluşturduğu milyonlarca veri vardır. Elde edilen verileri doğru şekilde analiz etme ve kullanma ihtiyacı her zaman için süregelmektedir. Twitter'ın tüm dünya üzerinde 330 milyondan fazla, Türkiye'de ise yaklaşık 12 milyon kullanıcısı bulunmaktadır. Kullanıcılar ilgi alanlarına göre etiketler üzerinden paylaşımlar yapmakta ve topluluklar bir araya gelebilmekte, markalar potansiyel müşterilerine ulaşabilmekte, sivil toplum kuruluşları da bu sayede hedef kitleleri ile etkileşime geçebilmektedir. Gelinen noktada etiketler, mesajlar, ifadeler tüm dünyada aktif ve sürekli bir şekilde kullanılmaktadır.

3.5.1. Twitter Yapısı

Twitter ne kadar bir blog özelliklerini içerse de kendi yapısına ait birçok özelliği vardır. Bunlar; Tweet, retweet, etiket (hashtag), bahsetme (mention), beğenme, trend topic, profil, yanıt, direkt mesaj gibi özelliklerdir.

• **Tweet:** Twitter platformu üzerinde yapılan paylaşımların her birine verilen addır. 140 karakter kullanımına imkân sağlayan Twitter 2017'den sonra 280 karaktere olanak sağlamıştır. 2011 yılının 25 Nisan tarihinden itibaren de Türkçe dil seçeneğini eklemiştir (Ayan, 2020).

• **Retweet (RT):** Kullanıcı tarafından paylaşılan tweetin başka bir kullanıcıya ait olduğu durumlarda başına RT ifadesi getirilerek kendi Twitter sayfasında paylaşmasıdır.

• **Etiket-Hashtag (#):** Hashtag olarak tanımlanan "#" işareti kullanarak; belli bir konudan, başlıktan, durumdan vb. bahsedilirken o konunun daha rahat aranmasını sağlayan, bahsedilen konudaki tweeti etiketlemek anlamına gelmektedir. Hashtag ile etiketleme yapıldığında tüm kullanıcılar o konu hakkında fikirlerini belirtmiş olurlar. Hedef tweetleri kategorize ederek, kullanıcıların arama kısmına istenilen kelimelerin başına # işareti koyarak o kelime ile ilgili bütün sonuçlara rahatça ulaşılabilmesini sağlamaktır. Bu özellik yardımıyla daha önceden o kelimeyi içeren tweet paylaşımlarını kimlerin yazdığı veya o konuyla ilgili ne tür fotoğraf paylaştığı görülebilmektedir. Bu özellikle birlikte sosyal medya iletişimi kolaylaştığı için Twitter bir iletişim aracı haline dönüşmüştür (Özkan, 2019).

• **Bahsetme-Mention (@):** Bahsetmek anlamında olan "@" işareti, bir twitter kullanıcısının adının önüne koyularak o kullanıcıyı tweete dahil etmek anlamına gelir.

• **Beğenme:** Paylaşılmış olan bir tweetin kullanıcılar tarafından beğenilmesidir. Twitter platformundaki kalp sembolü beğeni anlamına gelmektedir.

• **Trend Topic (TT):** Twitter gündemindeki popüler konuyu ifade eder. Bir hashtag defalarca kullanılıyorsa ve kullanıcılar bu konuda tweet atıyorsa o konu TT olabilir. Genellikle bu kısımda ilk 10 başlık gösterilmektedir.

• **Profil:** Twitter kullanıcılarının, en fazla 20 karakter kullanarak oluşturdukları kullanıcı adıyla oluşturdukları kişisel hesaptır. Diğer kullanıcıları takip etmek ve diğer kullanıcılar tarafından takip edilmek gibi özelliğe sahiptir.

• **Yanıt:** Paylaşılan tweete verilen cevaptır. Tweetlerin altında bulunan yanıtla butonuna basarak yapılır.

• **Direkt Mesaj (DM):** DM, twitter platformu kullanıcılarının birbirleri arasındaki mesajlaşmasına verilen addır.

3.5.2. Twitter API

Twitter, tweetlere programlı olarak otomatik bir şekilde erişmek için bir Uygulama Programlama Arayüzüne (Application Programming Interfaces - API) sahiptir. Bu arayüz-API, yazılımlar ya da veri tabanları arasında iletişim kurmaya yarayan yapılar olarak da tanımlanabilir. Bu arayüz kullanılarak, Twitter uygulamasından bağımsız bir şekilde web sitesinde yayınlanan tweet paylaşımı ya da paylaşılmış olan tweetlerin çekilmesi, oluşturulmuş yazılımlar sayesinde gerçekleştirilmektedir. Twitter API'si tweetlere programlı bir şekilde okuma ve yazma yapılmasını sağlamaktadır (Özkan, 2019).

Twitter API, Search, Rest ve Stream kütüphanelerini içeren 3 farklı API özelliklerinin birleşmesiyle oluşmuştur. Search API, belirlenmiş bir arama sorgusuna yönelik olarak geliştirilmiştir. Belirli anahtar kelimeler, belirli kullanıcının Tweetleri veya kullanıcılardan bahsedilen Tweetler gibi sorgular olabilir. Bu API türünde sonuçlar kısıtlanacak ve bundan dolayı sonuçlarda eksikler olabilecektir. Bütün sonuçların eksiksiz dönüşü alınmak istenilirse Streaming API kullanılmalıdır. Streaming API, Twitter platformundaki verilere anlık olarak erişim imkânı sunan API' dir. Arama kriterleri sonucu çok fazla veri

elde edilecekse bu API tercih edilmelidir. Rest API, Twitter veri setinde mevcut olan özel kişisel bilgilere ulaşmak için kullanılmaktadır. Bu API, kullanıcı Tweet oluşturma, gönderme, favori ekleme ve cevap yazma gibi etkileşimlerde kullanılabilmektedir (Olgun, 2018).

Twitter API, üç farklı aşamalı arama API'si sunmaktadır: Standart arama API'si, son bir hafta içinde yayınlanan en son paylaşım yapılmış Tweetler üzerinden arama yapar. Premium arama API'si, son bir aylık tweetlere ücretsiz erişim imkânı sunmaktadır. Kurumsal; ücretli son bir aylık tweete erişim veya 2006'nın başlarından itibaren tweetlere erişim imkânı sunmaktadır (Twitter, 2022).

Twitter API'den yararlanabilmek için ilk olarak Twitter'dan yetki-izin alınması gerekmektedir. 2018 yılına kadar API'ye erişip tweetler çekilebilmekteydi fakat daha sonra erişime yetki kısıtlaması getirildi. Bu yetkinin alınması için Twitter Developer Account denilen, yani Twitter'da bir geliştirici hesabının açılması gerekmektedir. Twitter iki seçenek sunmaktadır. Bunlar standart ve akademik araştırma alanlarıdır (Twitter, 2022). Standart, klasik olarak herkesin öğrenmek veya öğretmek için yazılımcılar da dahil olmak üzere genel olarak başvuruların olduğu alandır. Geliştiriciler twitter tarafından onay geldikten sonra standart proje oluşturabilir ve kullanabilir. Akademik araştırma tarafı ise, tüm arşivde arama yetkisi, daha yüksek aylık tweet sınırı ve gelişmiş filtreleme yetenekleri, yükseltilmiş erişim ve gelişmiş işlevlere erişimin sağlandığı alandır. Twitter, geliştirici hesabı başvurularını hemen kabul etmez, kullanım politikasına uygun olmayan başvurular reddedilmektedir. Akademik Araştırma başvurusu yapılırken, araştırmacının, kimlik bilgileri, oluşturmak istediği proje, Twitter verilerini nasıl kullanmayı planladığı ve çalışmayı nasıl paylaşacağı sorulur. Başvuruya onay verildikten sonra, standart veya akademik araştırma projesi ve API'ye yönelik tüm erişim isteklerinin doğrulanması için kullanılacak bir kimlik bilgisi sağlayacak olan ilişkili bir geliştirici uygulaması oluşturulmalıdır (Ballı, 2021).

3.5.3. Tweepy

Tweetleri toplayabilmek, veri seti oluşturabilmek ve işleyebilmek için Twitter çeşitli kütüphanelere sahiptir. Tweepy bu işlemleri sağlayan Python'da yazılmış kütüphanelerden bir tanesidir. Twitter'dan veri çekebilmek için kimlik doğrulama yapılması gerekmektedir. Bu doğrulama işlemi için Consumer Key olarak ifade edilen kullanıcı anahtarı, Twitter tarafından kullanıcılarına kaynaklara erişebilmesi amacıyla verdiği anahtardır. Consumer Secret olarak ifade edilen kısım ise kaynak erişim talebindeki anahtarla birlikte kullanılacak kullanıcı şifresidir, Access Token ve Acces Token Secret olarak adlandırılan ifadeler ise kullanıcı yetkilendirme işlemleri tamamlandıktan sonra Twitter tarafından kullanıcıya verilen API kimlik bağlantılarıdır. Bu yapıların tümüne Twitter API erişim isteğinde ihtiyaç duyulur. Twitter başvuru onayını verdikten sonra oluşturduğumuz uygulama için kimlik bilgisi ataması yapmaktadır. Bu kimlikler ile Twitter bağlantısı sağlanarak veriler çekilebilmektedir (Roesslein, 2009). Şekil 3.7'de kimlik doğrulaması örneği gösterilmiştir.

```
consumer_key = "******"
consumer_secret = "******"
access_token = "******"
access_token_secret = "******"

auth = tweepy.OAuthHandler(consumer_key, consumer_secret)
auth.set_access_token(access_token, access_token_secret)
api = tweepy.API(auth,wait_on_rate_limit=True)
```

Şekil 3.7. Tweepy API kimlik doğrulaması

3.6. Python Programlama Dili

Python, yaygın olarak kullanılan genel amaçlı, nesne yönelimli yüksek seviyeli bir programlama dilidir. 1990 yılların başında Guido Van Rossum tarafından, Hollanda' da ortaya çıkartılmış bir programlama dilidir. Guido Van Rossum,

Python'u geliştirmeye başladığında, kısa, eşsiz ve gizemli bir isim düşünmüş ve Monty Python adlı komedi grubunun gösterisi olan bir BBC komedi dizisinden esinlenmiş ve bu yüzden dili Python olarak adlandırmaya karar vermiştir (Python, 2022). Python, yorumlanmaya açık ve nesne yönelimli bir programlama dilidir. Python programlama dili birçok alanda kullanılmaktadır. En çok web uygulamaları geliştirirken tercih edilmektedir. En büyük örnekleri Youtube, Spotify'dır. Mobil uygulamaları ilk zamanlar Python ile geliştirmek mümkün değildi fakat zamanla gelişen Python kütüphaneleri sayesinde mümkün hale getirildi. Başka örnekler verecek olursak masaüstü arayüz uygulamaları, oyun geliştirme, akademik alanlarda veri analizi ve veri işleme, veri tabanı, makine öğrenmesi ve yapay zekâ en sık tercih edilen kullanım alanlarındandır (Nagpal ve Gabrani, 2019).

Python programlama dili avantajları (Hari, 2022):

- Açık kaynak kodlu ve ücretsizdir.
- Basit bir söz dizimine sahiptir. Bu sayede kolay öğrenilir ve okunabilir.
- Herhangi bir derleyiciye ihtiyaç duymaz.
- Taşınabilirlik özelliği sayesinde herhangi bir ortamda yazılan bir programı, başka ortamlarda da çalıştırılabilir hale getirebilmektedir.
- Geniş kütüphane alt yapısı sayesinde geliştirme süreci çok hızlıdır.

Python programlama dili dezavantajları(Hari, 2022):

- Yorumlanmış bir dil olmasından dolayı satır işleme de yavaşlığa sebep olur.
- Python'un veri tabanı erişim katmanı, bazı yaygın teknolojilere kıyasla ilkel ve az gelişmiştir.
- Mobil uygulamalarda zayıftır.

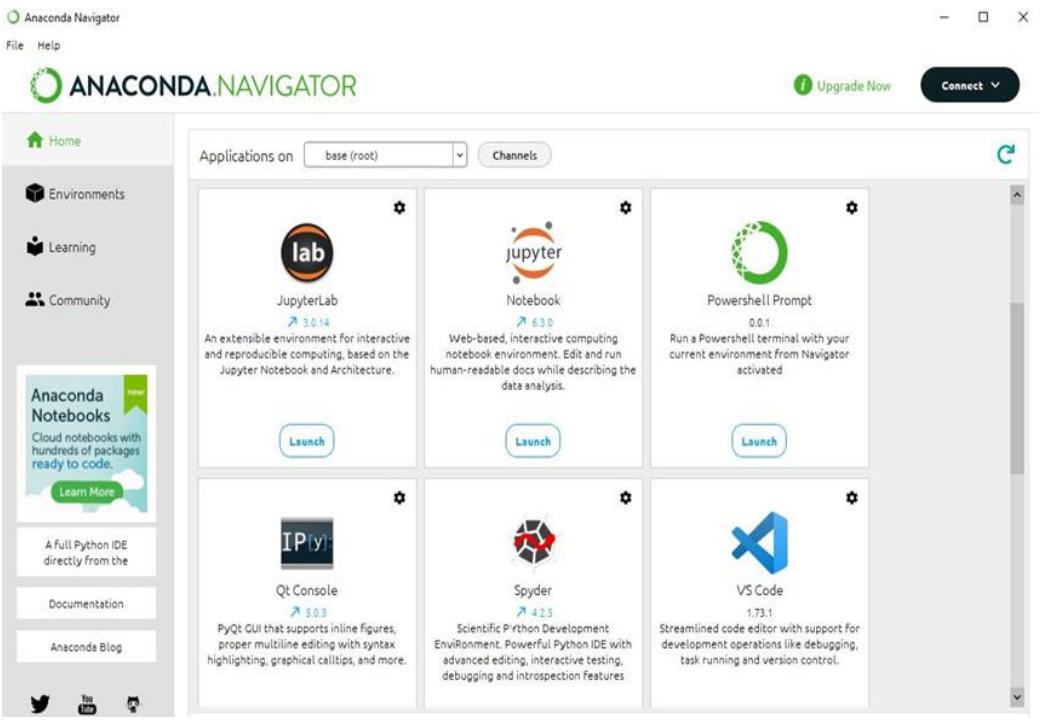
3.6.1. Yazılım Geliştirme Araçları

Programlama dillerinde yazılım geliştirebilmek için bir arayüze sahip olması gerekir. Bu arayüzlere IDE (Integrated Development Enviroment) yani Tümleşik Geliştirme Ortamı adı verilir. Python programlama dili için yerel bilgisayarlarda (local) veya bulut sistemlerde kullanabilecek birçok yazılım geliştirme ortamı

mevcuttur. Bu IDE'lerin çoğu kendi içlerinde bir bilgisayar programını makine koduna çevirip çalıştırmaya yarayan derleyici ve yorumlayıcı barındırmaktadır (Mert, 2021). Bunlardan bazıları; Anaconda, JupyterLab, Spyder, VsCode ortamlarıdır.

3.6.1.1. Anaconda

Anaconda, büyük ölçekli, açık kaynaklı ve ücretsiz olan; Python ve R programlama dillerinin paket yönetimini ve dağıtımını basitleştirmeyi hedefleyen bir dağıtımdır. 2012 yılında Peter Wang ve Travis Oliphant tarafından kurulmuş, geliştirilmiş ve sürdürülmüştür. Anaconda dağıtımında bulunan Anaconda Navigator, kullanıcıların komut satırına ihtiyacı olmadan uygulamaları başlatmasını ve Anaconda paketlerini ve ortamlarını yönetmesini sağlayan masaüstü arayüz uygulamasıdır. Anaconda Navigator'da hazır olarak bulunan birden fazla uygulama vardır. JupyterLab, JupyterNotebook, Spyder ve VsCode vb. varsayılan uygulamalardandır (Anaconda, 2022). Şekil 3.8'de Anaconda Navigator görünümü verilmiştir.



Şekil 3.8. Anaconda navigator görünümü

3.6.1.2. Jupyter Notebook

Jupyterlab, kodu ve metni düzenlemek için birçok uygulamada birlikte çalışabilen temel dosya biçimi olarak da tanımlanabilir. Markdown, JSON, CSV, Vega gibi popüler dosya biçimlerini düzenleyebilir. Jupyter Notebooklar grafikler, çizimler, resimler, videolar ve bağlantı adreslerini barındırır (Project Jupyter, 2022).

3.6.1.3. Spyder

Python programlama dili ile yazılmış ve Python geliştirme için kullanılabilen, ücretsiz ve açık kaynak kodlu bir IDE'dir. Spyder, gelişmiş düzenleme, etkileşimli test, hata yapılan kısmı ve hatayı gösteren hata ayıklama ve iç gözlem özellikleriyle Python programlama dili için güçlü bir etkileşimli geliştirme ortamıdır (Mert, 2021).

3.6.1.4. VsCode

Visual Studio Code, Microsoft tarafından Windows, Linux ve MacOS için geliştirilen tamamen ücretsiz bir kaynak kodu düzenleyicisidir. Gelişmiş web ve bulut uygulamaları ile kodları düzenlemeye, yeniden tanımlamaya ve optimize etmeye yarar. Yaklaşık 30 programlama dili desteği sunar. Bunlardan bazıları: C++, Python, Java, SQL, XML, C#, CSS vb. programlama dilleridir. Zengin bir intellisense özelliği yani kod tanıma içeriği vardır (Atamedya, 2019).

3.6.2. Python Kütüphaneleri

Kütüphaneler çok sık kullanılan fonksiyon, veri tipleri gibi unsurların tekrar tekrar kodlanması yerine, bir yerde tutularak istenilen zamanda çağırılan kodlar topluluğu olarak tanımlanabilir. Kütüphaneler yazılımı daha işlevsel hale getirilerek yazılımcıların iş yükünü hafifletir ve süre konusunda da avantaj sağlar. Bütün programlama dilleri kendilerine ait kütüphanelere sahiptir (Kargın, 2022). Alt başlıklarda Python içerisinde kolaylıkla kullanılabilen ve

yazılımcıların veri madenciliği, makine öğrenimi algoritmaları, duygu analizi gibi çalışmalarda kullandığı bazı Python kütüphaneleri açıklanmıştır.

3.6.2.1. Zemberek-Python

Metin ve veri madenciliği uygulamalarında yapılan durak kelimelerin kaldırılması, cümleleri daha küçük parçalara bölme yani simgeleme, parçaların morfolojik özelliklerini bulma, yazım denetimi, yanlış kelime önerme gibi işlemleri barındıran Java tabanlı, açık kaynak kodlu DDİ kütüphanesidir. Mehmet Dündar Akın ve Ahmet Afşin Akın tarafından geliştirilmiştir. Mehmet Dündar Akın kulağa komik/ilginç geldiği için zemberek kelimesini seçtiğini ifade etmiştir. Zemberek, Türkçede bir saatin ana yayı demektir. Etimolojik olarak da "küçük arı" anlamına gelen Farsça "zanbūrak" kelimesinden gelmektedir (GitHup, 2021).

3.6.2.2. NumPy

NumPy, Python'da temel bilgi işlemleri için yaygın olarak kullanılan pakettir. Çok boyutlu diziler ve matrisler üzerindeki matematiksel, mantıksal, şekil işleme, sıralama ve seçme gibi işlemler için, temel matematik ve istatistik işlemlerini hızlı ve esnek bir şekilde yapabilen kütüphanedir (NumPy, 2022).

3.6.2.3. Pandas

Pandas, veri yapıları ile çalışan ve veri analiz araçlarını barındıran önemli Python kütüphanelerindendir. Pandas, ilişkisel veya etiketli verilerle kolay ve sezgisel olarak çalışmayı sağlamak için tasarlanmış performansı yüksek, hızlı, esnek ve anlamlı bir Python paketidir. Veri yapılarındaki düzensiz biçimdeki verileri DataFrame nesnelerine dönüştürebilmektedir. Dosyadaki eksik verilerin kolay işlenmesi sağlar. Düz metin dosyalarından, excel dosyalarından, veri tabanlarından veri yükleme/kaydetme gibi işlemleri çok hızlı yapabilmektedir (Pandas, 2022).

3.6.2.4. NLTK

DDİ için kullanılan bir Python kütüphanesidir. NLTK (Natural Language Toolkit) içerisinde 50 farklı derlem ve sözlük bulundurmaktadır. Kütüphane içerisinde barındırdığı paketler sayesinde sınıflandırma, parçalama-tokenizasyon, kaynak oluşturma, etiketleme, ayrıştırma, anlamsal akıl yürütme gibi birçok farklı işlem bu kütüphane tarafından gerçekleştirilir (NLTK, 2022).

3.6.2.5. SciKit-Learn

Scikit-learn denetimli ve denetimsiz öğrenmeyi destekleyen açık kaynaklı bir makine öğrenimi kütüphanesidir. Modelleme ve tahmine dayalı analiz için kullanılabilecek çeşitli araçlar sağlar. Sınıflandırma, regresyon ve kümeleme algoritmalarına sahiptir (Scikit-Learn, 2022).

3.6.2.6. Snscrape

2020'de piyasaya sürülen Snscrape, sosyal ağ hizmetleri için bir kazıma aracıdır. Python 3.8 ve üstü sürüm gerektiren, Tweepy'nin kısıtlamaları olmadan tweet'leri kazımaya izin veren bir kütüphanedir (Desai, 2021).

3.6.2.7. Scipy

Scipy, matematik, bilim, mühendislik ve teknik problemleri çözmek için kullanılan açık kaynaklı bir kütüphanedir. Optimizasyon, özdeğer problemleri, diferansiyel denklemler, entegrasyon, enterpolasyon, cebirsel denklemler, istatistikler vb. için yerleşik algoritmalara sahiptir. Veri işleme ve görselleştirme için kolay bir yol sağlar (Gholizadeh, 2022).

3.6.2.8. Matplotlib

Matplotlib, statik, animasyonlu ve etkileşimli görselleştirmeler oluşturmak için oluşturulmuş bir kütüphanedir. John D. Hunter, tarafından oluşturulmuştur.

Dağılım, çizgi, pasta, histogram, güç spektrumları, çubuk grafikler vb. grafikleri oluşturan grafik çizim kütüphanesi olarak da tanımlanabilir (Matplotlib, 2022).

3.6.2.9. Seaborn

Seaborn, Python'da istatistiksel grafikler oluşturmak için bir kütüphanedir. Matplotlib kütüphanesine yüksek seviye arayüz sağlar. Seaborn ayrıca, kullanıcıların grafiklerinin görsel görünümünü değiştirmek için seçebilecekleri birçok yerleşik tema sunar (Waskom, 2021).

3.7. Veri Tabanı

Veri tabanı, bir bilgisayar sisteminde elektronik olarak depolanan organize bir şekilde yapılandırılmış bilgi veya veri koleksiyonudur. Başka bir tanımlamaya göre veri tabanı, birbiriyle ilişkili veri öğeleri topluluğudur. Veri tabanları yaygın olarak; Bankacılık: Müşteri bilgileri, hesaplar, krediler ve bankacılık işlemleri, Havayolları: Rezervasyon ve tarife bilgileri, Üniversiteler: Öğrenci bilgileri, ders kayıtları ve notları vb. alanlarda çok sık kullanılmaktadır (Sharma, 2022).

Veri tabanı genellikle bir veri tabanı yönetim sistemi tarafından kontrol edilir. Veri tabanı yönetim sistemi, veri tabanı ile kullanıcılar arasında bir arayüz görevi görerek kullanıcıların veri tabanı kurması, yönetmesi, güncellemesi ve bakımını yapması gibi işlemlere olanak tanır. Microsoft SQL Server (MSSQL), MySQL ve Oracle bazı popüler veri tabanı yönetim sistemlerindendir (Oracle, 2022). MSSQL veri tabanı yönetim sistemi, Microsoft firması tarafından 1989 yılında ilk olarak SQL Server 1.0 adı altında oluşturulmuştur. Birçok yazılım dilli ile uyumlu bir şekilde çalışabilmektedir. MSSQL, verilerin bütünlüğünün, güvenliğinin korunmasını sağlar ve çok kullanıcılı erişime izin verir. SQL yapısal sorgulama dili olarak tanımlanır. Bu sorgulama dili ile veri tabanında kayıt ekleme, silme, düzeltme gibi işlemler yapılabilir. Veri tabanı yönetim sistemlerinde veriler üzerinde sorgulama işlemleri yapılabilir, tablolar oluşturulabilir ve ilişki diyagramları kurulabilir (Sarpkaya, 2008).

3.8. Kelime Gömme Metotları

Metinlerde geçen ifadeleri sayısal olarak temsil etmek gerekir çünkü bilgisayar sistemleri metinsel ifadeleri algılayamaz. Bu durumu değiştirmek için de bazı yöntemler kullanılır. Bu yöntemlerden biri ve en gelişmişi de kelime gömme metotlarıdır. Kelime ve kelime parçacıklarını sayısal veriler ile ifade etme, kelime vektörleştirme işlemlerine kelime gömme (Word Embedding) denilmektedir (Sar, 2021).

3.8.1. Kelime Çantası

Kelime çantası (Bag of words, BoW), DDİ alanında kullanılan temel metin vektörleştirme modellerinden biridir. Kelime çantası ile metnin tamamı eşsiz terimlere ayrılmaktadır. Oluşturulan her bir terim bir öznitelik olarak kabul edilir ve sonra her bir özniteliğin yani terimin tüm metinde geçme sıklığı bulunur. Böylece kategorik bir veri sayısal hale dönüştürülmektedir (Aksu ve Karaman, 2020). Bu yönteme göre, belgedeki her metinsel ifade eşsiz kelimelere ayrılır ve eşsiz kelime boyutunda bir matrise dönüştürülür. Matrisin sütunlarını belge içindeki benzersiz kelimeler (N), satırlarını ise belge sayısı oluşturur (D). Sonuç olarak tüm veri kümesi DxN boyutunda bir matris olarak tanımlanmış olur (Aydoğan ve Karcı, 2019). Çizelge 3.2'de matris örneği verilmiştir.

Çizelge 3.2. Bag of words matris örneği

Bag Of Words	**Bu**	**Bir**	**Çocuk**	**Ve**	**Daha**	**Küçük**
Belge 1- "Bu Bir Kadın"	1	1	0	0	0	0
Belge 2- "Bu Kadın Ve Çocuk Ne Yapıyor"	1	0	1	1	0	0
Belge 3- "Adam Ve Bir Çocuk Ellerinde Küçük Bir Çanta ve Poşetle Gidiyor"	0	2	1	2	0	1

3.8.2. Terim Frekansı-Ters Doküman Frekansı

TF-IDF (Term Frequency Inverse Document Frequency- Terim Frekansı-Ters Doküman Frekansı) olarak adlandırılan bu yöntem bir dokümandaki her bir kelimenin değerlerini, belirli bir dokümandaki kelimenin geçme sıklığı ile kelimenin bulunduğu belgelerin yüzdesinin tersiyle hesaplama yapar. Çalışma prensibi, belirli bir belgede kelimelerin göreceli sıklığını, bu kelimenin tüm veri seti üzerindeki tersine oranına göre belirlemektir. Bu yöntemin sonucunda kelimenin belirli bir belge ile ne kadar alakalı olduğu bulunabilir diyebiliriz (Çelik ve Koç, 2021). TF-IDF yöntemi hesaplama fonksiyonu denklemi Denklem 3.1'deki gibidir.

$$\boldsymbol{w_{i,j} = tf_{i,j} \times \log \frac{N}{df_i}} \tag{3.1}$$

$\boldsymbol{tf_{i,j}}$ = i kelimesinin j dokümanında bulunma sıklığı

$\boldsymbol{df_i}$ = i kelimesinin geçtiği doküman sayısı

$\boldsymbol{N}$ = toplam doküman sayısı

3.8.3. Kelime Vektörü

Kelimeleri vektör olarak göstermeyi sağlayan yöntemlere kelime vektörü ya da diğer adıyla Word2Vec denir. Word2Vec, kelime gömme işleminde yapay sinir ağları kullanır ve denetimsiz öğrenme tekniğiyle vektöre dönüşüm gerçekleştirir. Kelime vektörünün kullandığı iki model vardır. Bunlar CBoW (Continuous Bag of Words- Sürekli Kelime Çantası) ve Skip-Gram Model'dir. Her iki modelde de pencere boyutu adı verilen giriş parametresi bulunmaktadır. CBoW modelinde, tahmin edilmek istenen kelimenin sağından ve solundan pencere boyutu parametresi kadar kelime girdi olarak verilir ve tahmin edilmek istenen kelimenin vektörü çıktı olarak elde edilmektedir. Skip-Gram modelinde ise CBoW modelinin tam tersi bir işlem gerçekleştirir. CBoW, daha çok küçük veri setlerinde tercih edilirken Skip-Gram büyük veri setlerinde daha başarılıdır (Aksu ve Karaman, 2020).

3.8.4. N-Gram

N-gram, bir metindeki N kelime dizisini ifade eder. N-gramlar, duygu analizi için makine öğrenmesi alanında oldukça sık kullanılmaktadır. N-gramlar kelime veya kelime grupları ve deyimler biçiminde birlikte kullanıldıklarında anlam bakımından daha bilgili ve manalı olabiliyorken, kelimelerin birçoğu da tek başına kullanıldığında bir anlam ifade etmezler. Bu doğrultuda N-gramlar duygu analizi çalışmalarında duygu barındıran kelime sıraları elde etmemizi sağlarlar (Akdeniz, 2022). N-gram modelinde N=1 ise unigram, N=2 ise bigram N=3 ise trigram olarak adlandırılmaktadır. Çizelge 3.3'de N-gram örneği verilmiştir.

Çizelge 3.3. N-gram yöntemi örneği

Metin: "Bu Kitap Çok Sürükleyiciydi"	
N=1(Unigram)	"Bu", "Kitap", "Çok", "Sürükleyiciydi"
N=2 (Bigram)	"Bu Kitap", "Kitap Çok", "Çok Sürükleyiciydi"
N=3 (Trigram)	"Bu Kitap Çok", "Kitap Çok Sürükleyiciydi"

3.9. Makine Öğrenmesi Algoritmaları

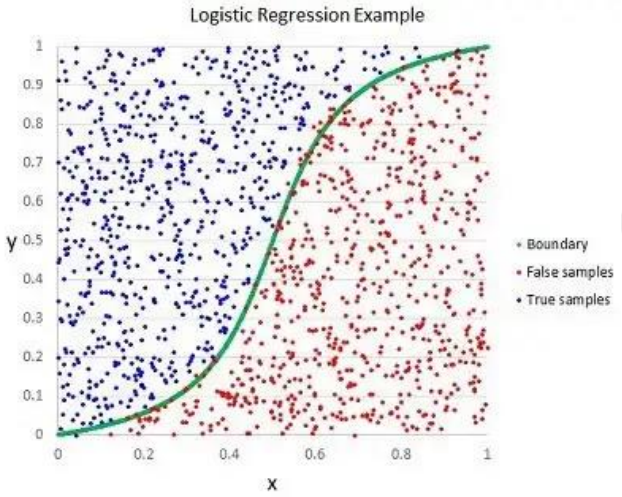
Makine öğrenmesi algoritmaları, karmaşık, anlaşılması zor veri kümelerinin keşfedilmesini, analiz edilmesini ve bunlarda anlam bulunmasına yardımcı olan kod bloglarıdır. Her bir farklı algoritma, bir makinenin belirlenmiş olan hedefi gerçekleştirmek için izleyebileceği kısıtlı ve belirli adım adım ilerleyen yönerge kümesidir. Makine öğrenmesinin hedefi, insanların tahmin yapmak veya bilgileri kategorilere ayırmak için kullanabileceği desenler oluşturmak veya keşfetmektir (GitHup, 2022).

Literatürde birçok makine öğrenmesi algoritması bulunmaktadır. Bu tezde kullanılan makine öğrenmesi algoritmaları alt başlıklarda açıklanmıştır.

3.9.1. Lojistik Resgresyon

Lojistik Regresyon (LR), araştırmalarda yaygın kullanılan bir yöntemdir. Sosyal bilimlerde, sağlık bilimlerinde, ekonomide, pazarlama ve bankacılık vb.

alanlarda verileri kategorik olarak analiz etmeye yarayan yöntemdir. LR, sınıflandırma işlemlerinde tercih edilen bir regresyon yöntemidir. Kategorik veya sayısal verilerin sınıflandırılmasında kullanılır. Bağımlı değişkenin yani sonucun, Evet / Hayır, Erkek / Kadın, Negatif / Pozitif gibi sadece 2 farklı değer alabilmesi gerekir. Şekil 3.9'da LR sınıflandırma örneği verilmiştir (Hatipoğlu, 2018).



Şekil 3.9. Lojistik regresyon sınıflandırma (Hatipoğlu, 2018)

Logistik Regresyon'un bağımlı değişkenin türüne göre 3 farklı analiz yöntemi vardır.

İkili Lojistik Regresyon; kategorik sınıflandırma çıktısı olarak sadece iki olası sonucun olduğu regresyon yöntemidir. Örnek olarak kadın veya erkek çıktılarını verebiliriz. Duygu analizi çalışmasında bu tip LR yöntemi tercih edilecekse sadece pozitif ve negatif tweetlerin verilmesi gerekir.

Çok Terimli Lojistik Regresyon; diğerlerinden daha fazla yani, üç veya daha fazla kategoride sonuç verebilir. Örneğin, yaş durumunu tahmin etmek (çocuk, genç, yaşlı) olabilir. Bu veri tipinde tweet duygu analizi verileri nötr, pozitif ve negatif olmalıdır.

Sıralı Lojistik Regresyon; sıralama ile üç veya daha fazla kategoride sonuç verir. Örnek: Bir markanın ürünün 1'den 5'e kadar puanlanması verilebilir (Swaminathan, 2018).

3.9.2. Naive Bayes Sınıflandırması

Naive Bayes (NB) sınıflandırma algoritması, Bayes teoremine dayanan olasılık tabanlı bir sınıflandırma yöntemidir. Sınıflandırmanın amacı her bir eleman için tüm durumların olasılığını hesapladıktan sonra olasılık değeri en yüksek olana göre sınıflandırmaktır. 3 tür NB yöntemi vardır (DevHunteryz, 2019) ve bunlar aşağıda açıklanmıştır:

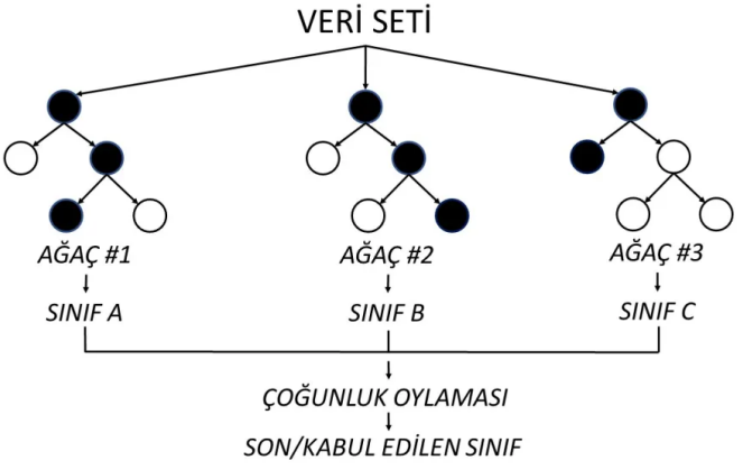
Multinomial Naive Bayes: Doküman sınıflandırma problemi için kullanılır, yani bir dokümanın spor, politika, teknoloji vb. kategorisine ait olup olmadığına bakarak sınıf tahmini yapar. Sınıflandırıcının kullandığı nitelikler dokümanda bulunan kelimelerin sıklığıdır.

Bernoulli Naive Bayes: Tahmin çıktı sonuçları boolean değişkenlerdir. Kullanılan parametreler yalnızca evet/hayır, iyi/kötü gibi net sonuç değerleri almaktadır.

Gauss Naive Bayes: Tahmin çıktıları sürekli bir değer aldıklarında ve ayrık olmadıklarında, bu değerlerin bir gauss dağılımından örneklendiği varsayılır.

3.9.3. Rastgele Orman Algoritması

Rastgele orman (Random Forest, RF) torbalama (bagging) yöntemiyle eğitilmiş karar ağaçlarından orman oluşturan bir denetimli makine öğrenmesi algoritmasıdır. Rastgele ormanda, her karar ağacı kendi kararını verir. Karar ormanında en çok oyu alan sınıf karar sınıfı olarak kabul edilir. Ormandaki her ağaç o sınıfa dahil edilmektedir. Kural olarak her ağaç bir sınıflandırma sonucu, son tahmin için en çok oylanan değeri seçerek sonuç oluşturur (Kemaloğlu vd., 2021).

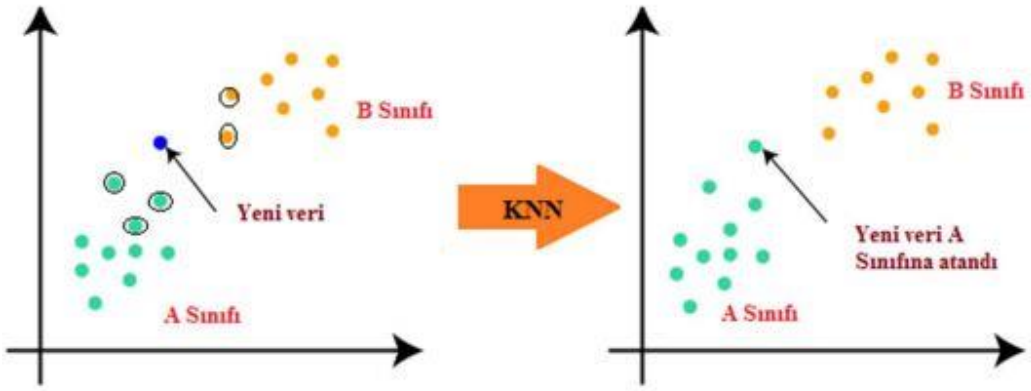


Şekil 3.10. Rastgele orman algoritması (Öztürk, 2022)

3.9.4. K-En Yakın Komşu Algoritması

K-En Yakın Komşu Algoritması (K-Nearest Neighbours, KNN), gözlem benzerliğine göre tahminler yapar. Sınıflandırma problemleri için oluşturulmuş ve daha sonra regresyon problemlerine uygulanmıştır. Adım adım algoritma tanımlanırsa; ilk önce baz alınacak komşuların sayısı yani "k" değeri belirlenir. Daha sonra bilinmeyen nokta ile diğer tüm noktalar arasındaki mesafe hesaplanır. Artan sıralama şeklinde uzaklıklar listelenir ve en yakın k gözlem seçilir. Son olarak ise sınıflandırma en sık sınıf olarak, regresyon ise ortalama değer tahmin değeri olarak verilir (Nacar ve Erdebilli, 2021).

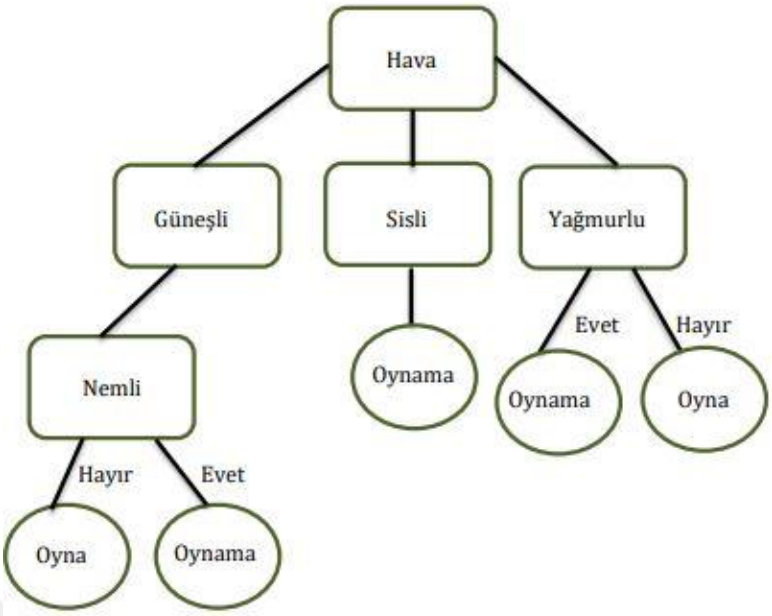
Örneğin, Şekil 3.11'de A ve B olarak adlandırılan iki sınıf ve sınıflandırmaya dahil edilecek bir veri var. Bu yeni veri KNN algoritmasına göre sınıflandırmaya dahil edilmek istenirse, öncelikli olarak komşu sayısı değerinin belirlenmesi gerekmektedir. k = 4 alınacak olursa, en yakın 4 komşu arasındaki uzaklıkların hesaplanması gerekir. Bu hesap sonucu en yakında bulunan komşuların çoğunluk sınıflarına göre yeni verinin ataması yapılır (Ayaz, 2021).



Şekil 3.11. K-En yakın komşu algoritması örneği (Ayaz, 2021)

3.9.5. Karar Ağacı Algoritması

Karar Ağacı algoritması (Desicion Tree, DT), etiketlenmiş verinin özelliklerine göre oluşturulan sınıflandırma algoritmalarıdır. Veri türü ayrımı yapmadığı için sınıflandırma problemlerinde yaygın kullanılan algoritmalardandır. Veri kümesi özelliklerine göre alt ağaçlara bölünerek karar ağaçları oluşturulur. Karar ağaçları kök düğüm, karar düğümleri ve yaprak düğümlerden oluşur. Sınıf etiketleri yaprak düğümleri oluşturur. Ağacın her düğümü verinin bir özelliğini ve her dal o özelliğe ait bir değeri ifade etmektedir. Sadece tek sınıfa ait örnekler kalana kadar veri kümesi ağaçlara bölünmeye devam eder (Elmas, 2019). Şekil 3.12'de hava durumu göre top oyna ya da oynama üzerinden karar ağacına örnek verilmiştir.



Şekil 3.12. Karar ağacı algoritması örneği

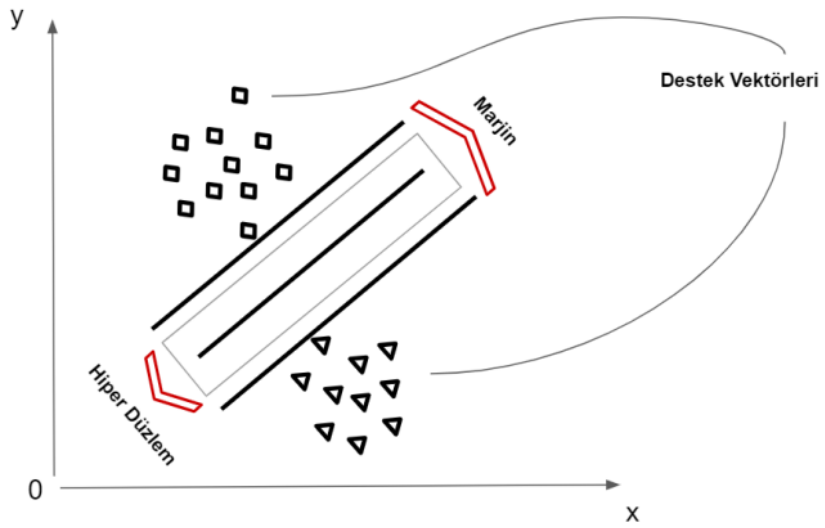
3.9.6. Destek Vektör Makineleri

Destek Vektör Makinesi (Support Vector Machine, SVM), Vladimir Vapnik ve Alexey Chervonenkis tarafından 1963 yılında tanımlanan, istatistiksel tabanlı bir algoritmadır. İlk olarak iki sınıflı doğrusal veriler sınıflandırılmıştır. Fakat daha sonra birçok sınıfı barındıran ve doğrusal olmayan verilerin sınıflandırılması amacıyla geliştirilmiştir. SVM algoritmasında temel amaç iki sınıfı en uygun şekilde ayıran bir hiper düzlem veya bir karar yüzeyi tanımlamaktır (Ayaz, 2021). Genel olarak SVM, Doğrusal Destek Vektör Makineleri ve Doğrusal Olmayan Destek Vektör Makineleri şeklinde ikiye ayrılmaktadır (Öztürk, 2022).

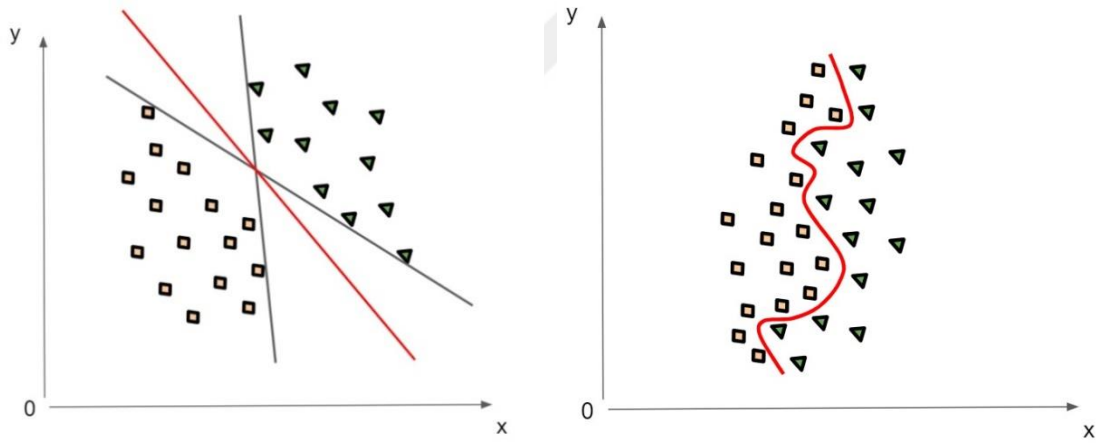
Doğrusal Destek Vektör: Veri grupları çizgiler-doğrular ile kolay bir şekilde ayrılıp, veriler düzlem-hiper düzlem ile sınıflandırılır.

Doğrusal Olmayan Destek Vektör: Veri grupları çizgiler-doğrular ile ayrılamayabilir (eğri olabilir), veriler düzlem-hiper düzlem ile sınıflandırılamaz.

Şekil 3.13'de iki boyutlu SVM algoritması, Şekil 3.14' de de doğrusal olan ve doğrusal olmayan sınıflandırma gösterilmiştir.



Şekil 3.13. İki boyutta SVM örneği (Öztürk, 2022)



(a) (b)

Şekil 3.14. SVM örneği a) Doğrusal olan, b) Doğrusal olmayan (Öztürk, 2022)

3.10. Model Başarım Ölçütleri

Model başarımı değerlendirilirken birkaç terim kullanılır. Bu terimler hata oranı, kesinlik, duyarlılık ve F-skorudur. Modelin başarısı, doğru sınıfa atanan örnek sayısı ve yanlış sınıfa atanan örnek sayısı değerleriyle alakalıdır. Test

sonucunda ulaşılan başarı oranı bilgileri karışıklık (hata) matrisi ile ifade edilir. Çizelge 3.4'de örnek bir karışıklık matrisi görülmektedir.

Çizelge 3.4. Karışıklık matrisi-confusion matrix

		Tahmin	
		Negatif	**Pozitif**
Gerçek	**Negatif**	**TN**	**FP**
	Pozitif	**FN**	**TP**

TN (True Negative): Gerçekte negatif olan ve tahminin de negatif olduğu sayı.
FN (False Negative): Gerçekte pozitif olan ama tahminin negatif olduğu sayı.
FP (False Positive): Gerçekte negatif olan ama tahminin pozitif olduğu sayı
TP (True Positive): Gerçekte pozitif olan ve tahminin de pozitif olduğu sayı.

Sınıflandırma başarım performansını karşılaştırmak için kullanılan metrikler doğruluk, kesinlik, duyarlılık ve F-Skoru'dur (Nalçakan vd., 2015), bunlara ait formüller Denklem 3.2, 3.3, 3.4 ve 3.5'de aşağıda açıklanmıştır.

• **Doğruluk (Accuracy):** Model başarım ölçütleri arasındaki en basit ve yaygın olan yöntemdir. Doğru sınıflandırılma yapılmış örnek sayısının, toplam örnek sayısına oranı doğruluk oranıdır.

$$Doğruluk = \frac{TP+TN}{TP+FP+FN+TN} \quad (3.2)$$

• **Kesinlik (Precision):** Kesinlik, sınıfı pozitif olarak tahmin edilmiş TP örnek sayısının, sınıfı pozitif olarak tahmin edilmiş tüm örnek sayısına (TP+FP) oranıdır.

$$Kesinlik = \frac{TP}{TP+FP} \quad (3.3)$$

• **Duyarlılık (Recall):** Duyarlılık sınıfı pozitif olarak tahmin edilmiş TP örnek sayısının, pozitif örnek sayısının toplamına oranıdır.

$$\boldsymbol{Duyarlılık} = \frac{\boldsymbol{TP}}{\boldsymbol{TP+FN}} \tag{3.4}$$

• **F-Skoru (F-Measure):** Duyarlılık ve kesinlik başarım ölçütleri, performans değerlendirme aşamasında tek başına kullanıldıklarında yeterli değildir. F-Skoru bu iki başarım metriğini kullanarak hesaplanır. Kesinlik ve duyarlılığın harmonik ortalaması olarak da ifade edilir.

$$\boldsymbol{F} = \frac{\boldsymbol{2\times Duyarlılık \times Kesinlik}}{\boldsymbol{Duyarlılık+Kesinlik}} \tag{3.5}$$

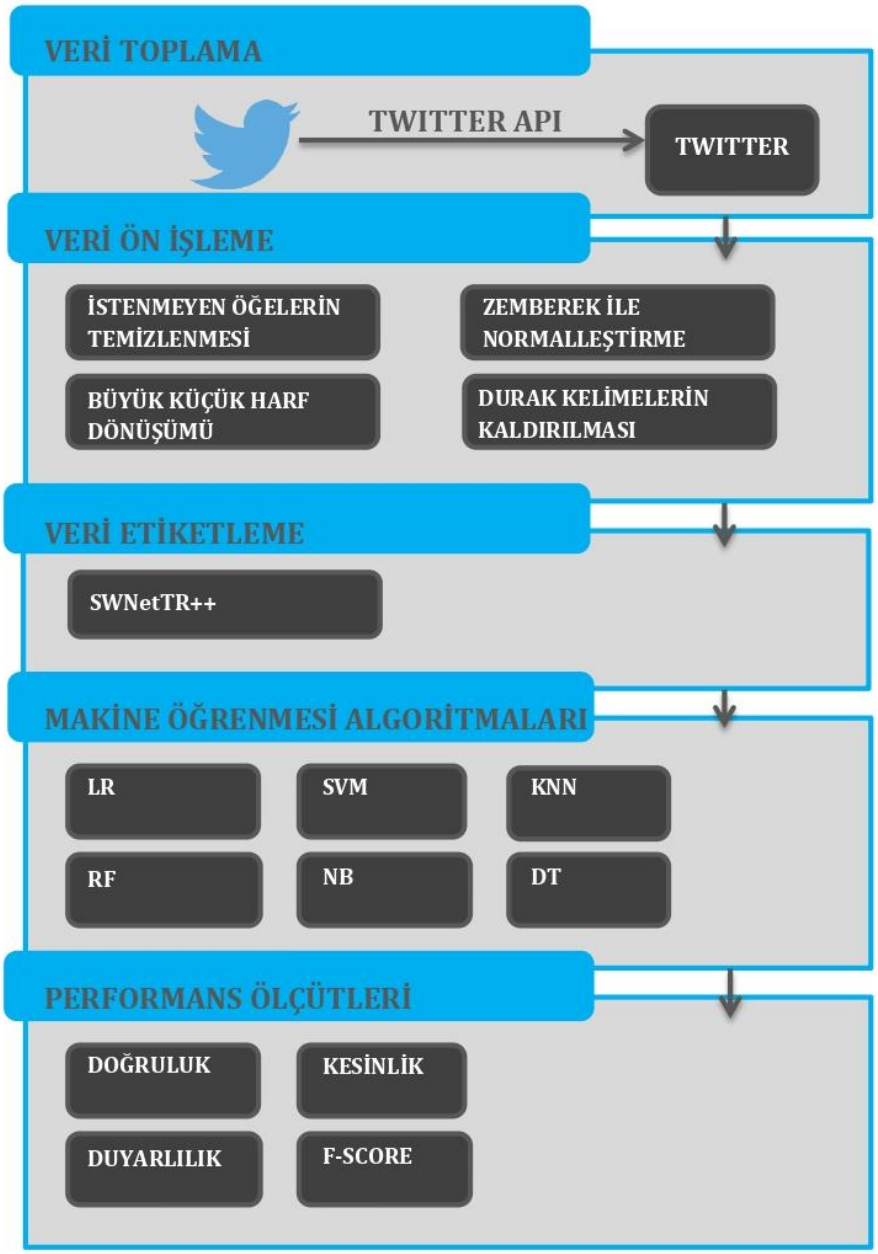
4. ARAŞTIRMA BULGULARI

Bu tez çalışmasında kuraklık etiketi ile toplanan tweet verilerden yola çıkarak, gerekli analizler sonucunda Türkiye-Kuraklık ilişkisini büyük ölçüde ortaya koyan tweet paylaşımlarının duygu analizi yapılmıştır.

Tez çalışmasının en önemli bölümünü kapsayan bu bölümde ise duygu analizi sonuçlarını elde etmek için bir önceki bölümde anlatılan yöntemlerin çalışmada nasıl kullanıldığı ve sonuç çıkarma aşamalarında nasıl uygulandığı detaylı anlatılmaktadır.

4.1. Çalışmanın Mimari Yapısı

Kuraklık üzerine yapılan duygu analizi çalışması, Anaconda ortamında Python programlama dili kullanarak geliştirilmiştir. Sözlük tabanlı duygu analiz yöntemi seçilmesinden dolayı SWNetTR++ sözlüğü kullanılmıştır. Veri seti çeşitli makine öğrenmesi algoritmalarıyla eğitilerek modeller oluşturulmuş ve oluşturulan modellerle sosyal medya kullanıcılarının duygu durumları pozitif, negatif veya nötr olarak sınıflandırılmıştır. Elde edilen sonuçlarla kullanılan algoritmaların başarı oranları karşılaştırılmış ve doğruluk oranları ortaya konulmuştur. Yapılan çalışmaya ait sistem mimarisi Şekil 4.1'deki gibidir.



Şekil 4.1. Çalışmaya ait sistem mimarisi

Geliştirilen sisteminin ilk aşamasında Twitter API üzerinden veri seti oluşturulmak amacıyla yazılım geliştirilmiştir. Toplanan verilerin daha anlamlı ve yapısal olarak düzgün bir biçim haline gelmesi için işleme ve temizleme aşamalarına tabi tutulmuştur. Doğrudan alınan işlenmemiş veri ile elde edilmiş tweetler üzerinde ilk olarak büyük-küçük harf dönüşümü yapılmış ardından da istenmeyen tüm öğeler temizlenmiştir.

Veri ön işleme kısmında tekrar eden aynı satırlar veri setinden arındırılmıştır. Metindeki yazım hataları ve kelime önerilerinde bulunarak normalleştirme yapılmasını sağlayan yöntem için ise zemberek kullanılmıştır.

Veri etiketleme bölümünde makine öğrenmesi algoritmalarına girdi olarak kullanılacak veri seti için üç duygu sınıfına göre, girdiye uygun olacak şekilde duygu etiketi oluşturulmaktadır. Bu çalışmada sözlük tabanlı analiz yöntemin tercih edilmiştir. SWNetTR++ sözlüğü ile etiketleme gerçekleştirilmiştir.

Sistemin diğer bir adımında ise sınıflandırma çıktılarını alabilmek için gerekli modeller oluşturulmuştur ve farklı makine öğrenmesi algoritmaları (LR, RF, SVM, NB, KNN, DT) ile birlikle karşılaştırılması ve bu karşılaştırma sonucunda doğruluk/performans başarım ölçütlerine bakılarak en iyi sınıflandırma modelinin tespit edilmesi amaçlanmıştır.

En son aşamada ise duygu analizi gerçekleştirildikten sonra, Word Cloud ile kelime bulutu oluşturulmuştur.

4.2. Veri Oluşturma

Twitter üzerinde paylaşılan 01.01.2019 ile 01.01.2022 tarihleri arasında, içerisinde “kuraklık” sözcüğü geçen toplam 96401 tweet, Snscrape kütüphanesi kullanan bir Python yazılımı ile toplanmıştır. Anaconda dağıtım sistemi içinde bulunan VsCode kaynak kod düzenleyicisi, Python için iyi bir destek olduğundan tercih edilmiştir. Twitter platformundaki mesajların toplanabilmesi için Twitter tarafından oluşturulmuş olan “Twitter Search API” uygulama programlama arayüzü kullanılmıştır. Veri çekebilmek için Developer Twitter/Geliştirici Twitter üyeliği aktifleştirilmiş ve gerekli izinler alınmıştır. Kullanılan kütüphane ve gerekli kod yazılımları Şekil 4.2’ de gösterilmiştir. Tweet çekme/toplama/kazıma işlemi için öncelikle şekilde de görüldüğü üzere bir sorgu kelimesi (kuraklık) seçilmiş, sonrasında bu sorgu kelimesinin de içinde bulunduğu kazınmış tweetler veri seti içinde detay isteyen bir sorgu yazılır. Tüm bu işlemler bir döngü içerisine alınır ve o şekilde gerçekleştirilir.

```
import snscrape.modules.twitter as sntwitter
for i,tkurak in enumerate(sntwitter.TwitterSearchScraper
(' kurakl k lang:tr since:2019-01-01 until:2022-01-01').get_items()):
    if i>5000000:
        break
    print(tkurak.content)
    print(tkurak.username)
    print(tkurak.date)
    print(i)
    print("\n")

    command = ('insert into [twitterduygu].[dbo].[tkurak](content, username, date) values (?,?,?) ')
    values = [tkurak.content, tkurak.username, tkurak.date]

    cursor.execute(command, values)
    cursor.commit()
```

Şekil 4.2. Tweet çekme yazılımı

Toplanan tweet mesajlarını saklamak için ise MSSQL (Microsoft SQL Server) veri tabanı yönetim sistemi kullanılmıştır. Analiz için "twitterduygu" isimli veri tabanı oluşturulmuştur. Bu veri tabanında "tkurak" isimli tablo oluşturulmuş ve Snscrape kütüphanesi ile kazınan tweetler bu tabloya kaydedilmiştir. Bu tablo içerisine her bir tweet mesajına ait eşsiz kimlik numarası (Tweet id), paylaşılan mesaj (tweet/content), mesajı paylaşan kullanıcı adı (username), ve mesajın paylaşıldığı tarih (date) bilgileri kaydedilmiştir. Şekil 4.3'te MSSQL tablo yapısı verilmiştir. Şekil 4.4'de veri seti yapısı özet bilgileri ve Şekil 4.5'te Twitter veri seti örneği gösterilmiştir.

tkurak

	Column Name	Data Type	Allow Nulls
🔑	id	int	☐
	[content]	nvarchar(MAX)	☑
	username	nchar(100)	☑
	date	date	☑
			☐

Şekil 4.3. MSSQL tablo yapısı

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 96401 entries, 0 to 96400
Data columns (total 4 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   id        96401 non-null  int64
 1   content   96401 non-null  object
 2   username  96401 non-null  object
 3   date      96401 non-null  object
dtypes: int64(1), object(3)
memory usage: 2.9+ MB
```

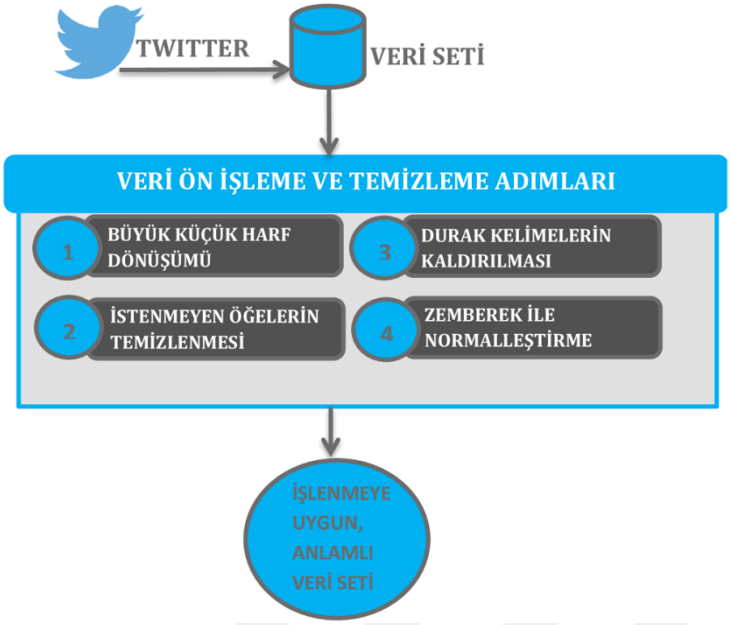
Şekil 4.4. Veri seti özet bilgileri

	id	content	username	date
0	1	dib ali erbas agabey kuraklik icin yagmur duas...	bababanacharal ...	2021-12-31
1	2	@elonue Anlaşılmadı zam mecburi doğalgaz petro...	Nuhokyanus1 ...	2021-12-31
2	3	2021 dileklerim tutmayınca 2022 için yazmaya k...	eneskulc ...	2021-12-31
3	4	#Türkiye'nin doğusundaki en büyük gölü olan #V...	cevretvcomtr ...	2021-12-31
4	5	Enerji krizi..!\n\nDünyada her yıl tüketilen e...	teskit ...	2021-12-31
5	6	kimse aşkı bulamayacak, elimizde kalan son şey...	cayikimdemledi ...	2021-12-31
6	7	@MediaMuhtari Enerji kitligi ve kuraklik	smlkdr ...	2021-12-31
7	8	En dibi gören kuraklık,\n2022 itibari ile yeri...	GizliOperasy0n ...	2021-12-31
8	9	@namyunkiy21 Evet öyle bir şey olucakmış ama b...	jeonsnowy ...	2021-12-31
9	10	◆2022 yılında ormanlar yanmasın\n\n◆Hayvanlara...	KEMZEY_64 ...	2021-12-31

Şekil 4.5. Twitter veri seti örneği

4.3. Veri Ön İşleme ve Metin Temizleme

Tweetler üzerinden çekilen veri setinin, düzensiz ve işlenmeye uygun bir yapısı olmadığı için verilerin analize gönderilmeden önce ön işleme tabii tutulması ve temizlenmesi gerekir. Bu adımın amacı kısaca veriyi çalışmaya/analize uygun hale getirmektir. Tez çalışmasında uygulanan veri ön işleme ve temizleme adımları Şekil 4.6'daki gibidir.



Şekil 4.6. Veri ön işleme ve temizleme adımları

Her tweet mesajında çok sayıda hashtag'ler, bahsetmeler, emojiler, semboller, noktalama işaretleri vb. ifadeler duygu analizi çalışması için anlamı olmadığından dolayı bunları temizlemek gerekir. Bu temizleme işlemlerinin hepsi için Python programlama dili ve Jupyter Notebook ara yüzü kullanılmıştır. Bu adımın önemli bir avantajı ise depolama alanı için gereksiz ifadelerden dolayı oluşan büyük veri boyutunu en aza indirgeyebilmesidir. Veri setinin boyutunun küçültülmesi, duygu analizi üzerindeki çalışmanın daha etkili olması ve doğru sonuçlar alınması açısından önemlidir. Bu aşama sonrası elde etmiş olduğumuz tweetlere ön işleme adımlarına ek olarak, verilerin daha yalın ve anlamsal açıdan daha verimli hale gelmesi için Zemberek ile normalleştirme işlemi gerçekleştirilmiştir. Veri seti içinde tekrar eden satırlar tespit edilerek veri seti eşsiz tweetlere sahip olacak şekilde diğer tweetlerden arındırılmıştır.

• **Büyük Küçük Harf Dönüşümü:** Metin içerisindeki büyük harflerin küçük harflere dönüşümü lower() metodu kullanarak yapılmıştır. Bu dönüşüm kelime anlamında değişikliğe sebep olmayacaktır. Yapılan büyük küçük harf dönüşümü Şekil 4.7'de gösterilmiştir.

id	content	kücük_harf_dönüsümü
1	dib ali erbas agabey kuraklik icin yagmur duas...	dib ali erbas agabey kuraklik icin yagmur duas...
2	@elonue Anlaşılmadı zam mecburi doğalgaz petro...	@elonue anlaşılmadı zam mecburi doğalgaz petro...
3	2021 dileklerim tutmayınca 2022 için yazmaya k...	2021 dileklerim tutmayınca 2022 için yazmaya k...
4	#Türkiye'nin doğusundaki en büyük gölü olan #V...	#türkiye'nin doğusundaki en büyük gölü olan #v...
5	Enerji krizi..!\n\nDünyada her yıl tüketilen e...	enerji krizi..! dünyada her yıl tüketilen ener...
6	kimse aşkı bulamayacak, elimizde kalan son şey...	kimse aşkı bulamayacak, elimizde kalan son şey...
7	@MediaMuhtari Enerji kitligi ve kuraklik	@mediamuhtari enerji kitligi ve kuraklik
8	En dibi gören kuraklık,\n2022 itibari ile yeri...	en dibi gören kuraklık, 2022 itibari ile yerin...
9	@namyunkiy21 Evet öyle bir şey olucakmış ama b...	@namyunkiy21 evet öyle bir şey olucakmış ama b...
10	◆2022 yılında ormanlar yanmasın\n\n◆Hayvanlara...	◆2022 yılında ormanlar yanmasın ◆hayvanlara ez...

Şekil 4.7. Büyük-küçük harf dönüşümü örneği

• **İstenmeyen Öğelerin Temizlenmesi:** Bu işlem adımında etiket, hashtag, RT, url, sayı, noktalama işaretleri ve Türkçe olmayan karakterler kaldırılır. Gereksiz öğeleri temizleme kod bloğu Şekil 4.8'de gösterilmiştir. Kod bloğunda "replace" fonksiyonu kullanılmıştır. Bu fonksiyon istediğimiz herhangi bir alan içerisinde istediğimiz herhangi bir karakteri istediğimiz başka bir karakter ile yer değişimini sağlar. En son kısımda tek tırnak içerisindeki boşluk ifadesi replace ile bulunan ifadelerin yerine boşluk koy yani sil anlamına gelmektedir

```
df['content']=df['content'].str.replace('@[\w:]+', '')
df['content']=df['content'].str.replace('[^\w\s]','')
df['content']=df['content'].str.replace('\d','')
df['content']=df['content'].str.replace('rt','')
df['content']=df['content'].str.replace('((25[0-5]|2[0-4][0-9]|[01]?[0-9][0-9]?)(\.|$)){4}', '')
df['content']=df['content'].str.replace('[a-zA-Z0-9-_.]+@[a-zA-Z0-9-_.]+', '')
```

Şekil 4.8. İstenilmeyen öğeleri temizleyen kod bloğu

Gereksiz ifadelerden sonra oluşan yeni veri setimiz Şekil 4.9'da verilmiştir.

id	content	id	content
1	dib ali erbas agabey kuraklik icin yagmur duas...	1	dib ali erbas agabey kuraklik icin yagmur duas...
2	@elonue anlaşılmadı zam mecburi doğalgaz petro...	2	anlaşılmadı zam mecburi doğalgaz petrol üretm...
3	2021 dileklerim tutmayınca 2022 için yazmaya k...	3	dileklerim tutmayınca için yazmaya karar ver...
4	#türkiye'nin doğusundaki en büyük gölü olan #v...	4	türkiyenin doğusundaki en büyük gölü olan van ...
5	enerji krizi..! dünyada her yıl tüketilen ener...	5	enerji krizi dünyada her yıl tüketilen enerji ...
6	kimse aşkı bulamayacak, elimizde kalan son şey...	6	kimse aşkı bulamayacak elimizde kalan son şeyl...
7	@mediamuhtari enerji kitligi ve kuraklik	7	enerji kitligi ve kuraklik
8	en dibi gören kuraklık, 2022 itibari ile yerin...	8	en dibi gören kuraklık itibari ile yerini rah...
9	@namyunkiy21 evet öyle bir şey olucakmış ama b...	9	evet öyle bir şey olucakmış ama bu sene başla...
10	◆2022 yılında ormanlar yanmasın ◆hayvanlara ez...	10	yılında ormanlar yanmasın hayvanlara eziyet e...

Şekil 4.9. Veri setinden gereksiz öğelerin temizlenmesi

• **Durak Kelimelerinin Kaldırılması:** Durak kelimeleri olarak adlandırılan Stopwords sözcükleri, bir cümleye, metine anlam katmayan bir terimlerdir. Bundan dolayı cümlenin anlamını değiştirmezler ve görmezden gelinebilir veya metin dosyasından kaldırılır. NLTK kütüphanesi içerisinde farklı dillere ait durak kelimeler listesi yer almaktadır. Türkçe durak kelimelere ulaşmak için stopwords.words('Turkish') metodu kullanılır. Şekil 4.10'da Türkçe için durak kelimelerin listesi ve Şekil 4.11'de de durak kelimelerden arındırılmış veri seti gösterilmiştir.

```
stop_words= set (stopwords.words('Turkish'))
```

```
{'acaba', 'ama', 'aslında', 'az', 'bazı', 'belki', 'biri', 'birkaç', 'bir şey', 'biz', 'bu', 'da', 'daha', 'de',
'defa', 'diye', 'en', 'eğer', 'gibi', 'hem', 'hep', 'hepsi', 'her', 'hiç', 'ile', 'ise', 'için', 'kez', 'ki', 'kim',
'mu', 'mü', 'mı', 'nasıl', 'ne', 'neden', 'nerde', 'nerede', 'nereye', 'niye', 'niçin', 'o', 'sanki', 'siz',
'tüm', 've', 'veya', 'ya', 'yani', 'çok', 'çünkü', 'şey', 'şu'}
```

Şekil 4.10. Türkçe durak kelimeler

id	content
1	dib ali erbas agabey kuraklik icin yagmur duas...
2	anlaşılmadı zam mecburi doğalgaz petrol üretmi...
3	dileklerim tutmayınca yazmaya karar verdim cov...
4	türkiyenin doğusundaki büyük gölü olan van göl...
5	enerji krizi dünyada yıl tüketilen enerji üret...
6	kimse aşkı bulamayacak elimizde kalan son şeyl...
7	enerji kitligi kuraklik
8	dibi gören kuraklık itibari yerini rahmet boll...
9	evet öyle bir olucakmış sene başlayacak kurakl...
10	yılında ormanlar yanmasın hayvanlara eziyet ed...

Şekil 4.11. Durak kelimelerin olmadığı veri seti

• **Zemberek İle Normalleştirme**: Veri setinin daha sade ve anlamlı bir mesaj içeriğine sahip olabilmesi için Zemberek-Python kütüphanesi kullanılmıştır. Bu kütüphane ile yazım hataları giderilmiş ve kelime tahmini ile tweet mesajlarında normalleştirme işlemi gerçekleştirilmiştir. Normalleştirme için kullanılmış olan kütüphaneler ve kod yazılımı Şekil 4.12'de verilmiştir. Veri setinin bu adımdan sonraki son hali ise Şekil 4.13'teki gibidir.

```
import time
import logging
import pandas as pd
import csv
from zemberek import (
    TurkishSpellChecker,
    TurkishSentenceNormalizer,
    TurkishSentenceExtractor,
    TurkishMorphology,
    TurkishTokenizer
)

logger = logging.getLogger(__name__)
data=[]
df = pd.read_csv('C:\\Users\\Asus\\Desktop\\yükseklisans\\twitter\\ensonveri\\t--kurak.csv')

morphology = TurkishMorphology.create_with_defaults()

# SENTENCE NORMALIZATION
start = time.time()
normalizer = TurkishSentenceNormalizer(morphology)
logger.info(f"Normalization instance created in: {time.time() - start} s")

start = time.time()
```

Şekil 4.12. Zemberek normalleştirme kod örneği

id	content	content
1	dib ali erbas agabey kuraklik icin yagmur duas...	dibi ali erbaş ağabey kuraklık için yağmur du...
2	anlaşılmadı zam mecburi doğalgaz petrol üretmi...	anlaşılmadı zam mecburi doğalgaz petrol üretm...
3	dileklerim tutmayınca yazmaya karar verdim cov...	dileklerim tutmayınca yazmaya karar verdim ço...
4	türkiyenin doğusundaki büyük gölü olan van göl...	türkiyenin doğusundaki büyük golü olan van gö...
5	enerji krizi dünyada yıl tüketilen enerji üret...	enerji krizi dünyada yıl tüketilen enerji üre...
6	kimse aşkı bulamayacak elimizde kalan son şeyl...	kimse aşkı bulamayacak elimizde kalan son şey...
7	enerji kitligi kuraklik	enerji kıtlığı kuraklık
8	dibi gören kuraklık itibari yerini rahmet boll...	dibi gören kuraklık itibarı yerini rahmet bol...
9	evet öyle bir olucakmış sene başlayacak kurakl...	evet öyle bir olucakmış sene başlayacak kurak...
10	yılında ormanlar yanmasın hayvanlara eziyet ed...	yılında ormanlar yanmasın hayvanlara eziyet e...

Şekil 4.13. Zemberek ile normalleştirilmiş veri seti

Tüm bu adımlardan sonra veri seti içinde aynı mesaj içeriğine sahip tweetler sadece bir adet kalacak şekilde veri setinden çıkartılmıştır. Satır silme işlemi sonucunda veri setindeki 96401 adet tweet mesajından geriye 82221 adet tweet kalmıştır.

4.4. Veri Etiketleme

Bu tez çalışmasında sözlük tabanlı yöntem kullanarak duygu etiketleme yöntemi tercih edilmiştir. Duygu analizinde kelimelerin pozitif, negatif puanlarını içeren bir sözlüğe ihtiyaç duyulmaktadır. Bu çalışmada SWNetTR++ duygu kelime sözlüğü kullanılmıştır. Duygu sözlüğündeki pozitif, negatif ve nötr kelime değerleri, cümlede geçen kelimeler için ayrı ayrı toplanarak hesaplanır. İşlemler sonucunda elde edilen baskın değere göre cümlenin duygu durumu etiketlenir. Sözlük yaklaşık olarak 49227 kelimeden oluşmaktadır. Her kelime/kelime grubu kullanıma hazır bir şekilde -1 ve +1 olarak duygu durumuna göre etiketlenmiştir. Şekil 4.14'de sözlük içeriği örneği verilmiştir.

acele acele	-1	fırlak çeneli	-1
acele aile	-1	fırlama	-1
acele etmek	1	fırlamak	1
aceleci	-1	fırlamış	-1
aceleci kişi	-1	fırlatış	1
aceleciliği	-1	fırlatmak	1
acelecilik	-1	fırlatmalara	-1
aceleyle	1	fırsat	1
acemi	-1	fırsat vermemek	-1
acemi çaylak	1	fırsatçı	1

Şekil 4.14. SWNetTR++ sözlük içerik örneği

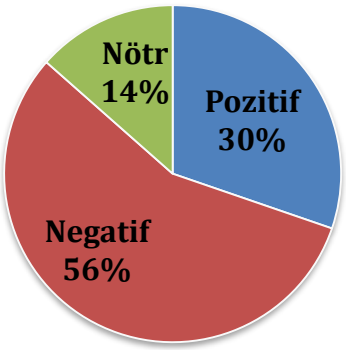
Duygu sınıflandırma yöntemlerinden biri olan sözlük tabanlı yöntemde SWNetTR++ sözlüğü kullanılarak tweetler analize alınmıştır. Oluşturulan bu yöntem modelinde 82221 adet tweetin etiketleme işlemi Anaconda ortamında Python programlama dili kullanılarak Spyder ara yüzünde gerçekleştirilmiştir. Analiz sonucuna göre 24907 tweet pozitif, 46255 tweet negatif, 11059 tweet nötr olarak işaretlenmiştir. Etiketlenen tweetlerin duygu çıktıları Şekil 4.15'de etiketlenmiş tweetlerin duygu örnekleri Çizelge 4.1'de ve duygu dağılımları daire grafiği de Şekil 4.16'daki gibidir.

id	tweet	duygu
256	batı karadeniz bölgesi korkuta…	negatif
259	bir göçün çeşitli nedenleri ol…	pozitif
260	kuraklık sadece benimi korkutu…	negatif
261	ülkemde beton ağaçtan seviliyo…	pozitif
262	kuraklık olacağına sel baskınl…	negatif
263	trakyanın başındaki bir köyde …	pozitif
265	bunlar icat değil planlananmış…	negatif
266	arkadaşlar afrika büyük yer al…	negatif
267	türkiye halleder geriye kalanı…	negatif
268	durağı psikoloji katılıyorum k…	nötr

Şekil 4.15. Sözlük tabanlı duygu çıktıları

Çizelge 4.1. Etiketlenmiş tweetlerin duygu örnekleri

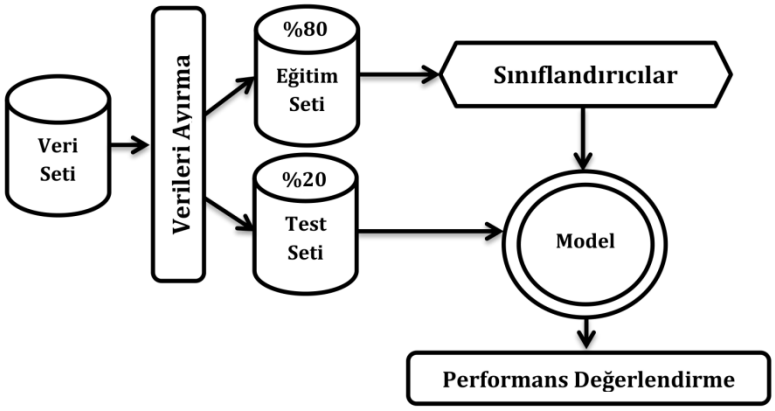
TWEET	DUYGU ETİKETİ
gidişat çok kötü tuz golü kuraklık nedeniyle küçülüyor	Negatif
afetsiz yağmurlar diliyorum kuraklık olmasındansa gri gökyüzü candır	Pozitif
kuraklık ve sellerin ekonomik yansıması	Nötr



Şekil 4.16. Duygu dağılımları daire grafiği

4.5. Verilerin Ayrılması ve Modelleme

Tez çalışmasının veri işleme bölümlerinde veriler yapısal biçime uygun olacak şekle getirilmiştir. Artık veri setimiz modelleme işlemleri için hazır hale gelmiştir. Bu bölümde de en iyi sınıflandırma modelini bulabilmek için etiketlendirilen veri setleri BoW ve TF-IDF ile ayrı ayrı sayısal olarak temsil edilerek Python programlama dili içerisinde bulunan Sklearn kütüphanesinde yer alan train_test_split metodu ile iki parçaya ayrılmıştır. Birinci veri setinde veriler eğitilir, ayrılan diğer veri setinde ise doğruluğu test edilir. Çalışmada test_size=0.2, random_state=0 alınarak veri setinin %80'i eğitim, %20'si de test için ayarlanmıştır. Eğitim ve test verilerinin sınıflandırma için ayrılması Şekil 4.17'de gösterilmiştir.



Şekil 4.17. Eğitim ve test verilerinin ayrılması

Tez kapsamında sınıflandırma işlemi için etiketlendirilen metinlerin pozitif, negatif veya nötr duygu dağılımları makine öğrenmesi algoritmalarının performansları karşılaştırılmıştır. BoW ve TF-IDF veri sayısallaştırma yöntemleri ile LR, SVM, RF, KNN, NB ve DT makine öğrenmesi algoritmaları kullanılarak sınıflandırma modelleri oluşturulmuştur. Sözlük tabanlı yöntem ile etiketlendirilen veri setleri sınıflandırma için girdi olarak kullanılmış ve her birinin sınıflandırma işlemlerindeki performansları değerlendirilmiştir. Sınıflandırma modellerinden en yüksek başarım oranına sahip modeli seçmek için doğruluk oranları dikkate alınmıştır.

4.6. Doğruluk Oranları ve Performans Başarım Ölçütleri

Bu aşamada veri seti Türkçe metinlerden oluşan sözlük tabanlı yöntem kullanarak etiketleme yaptığımız veri seti modeli ele alınmıştır. Bu veri setlerinde BoW ve TF-IDF veri sayısallaştırma yöntemleri kullanıp öznitelikleri çıkartıldıktan sonra LR, SVM, RF, KNN, NB ve DT makine öğrenmesi algoritmaları ile sınıflandırma işlemlerine tabii tutularak en iyi modelin seçilmesi hedeflenmiştir.

Sınıflandırma için SWNetTR++ sözlüğü kullanılarak etiketlemesi yapılan modelin ilk olarak BoW, sonrasında ise TF-IDF veri sayısallaştırma yöntemleri kullanarak öznitelikleri çıkartıldıktan sonra LR, SVM, RF, KNN, NB ve DT makine öğrenmesi algoritmaları ile sınıflandırılması yapılmıştır. Model için yazılan

kodun bir kısmı Şekil 4.18'de gösterilmiştir. Sınıflandırma modeli kod örneğinde Sklearn kütüphanesine bağlı "CountVectorizer" ile veri seti sayısallaştırması yapılmıştır. "CountVectorizer" dosyadaki tüm kelimelerin sıklığını saymaktadır. Bu işlemler sonrasında kelime çantası modeli bir sınıflandırıcı için girdi olarak kullanılabilir hale gelmektedir. Oluşturulan modele ait başarım ölçütleri Çizelge 4.2'de verilmiştir.

```
bow_vectorizer = CountVectorizer()
bow = bow_vectorizer.fit_transform(df['clean'])
bow.shape

df=df.fillna(0)

from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(bow, df['etiket'],
                                                    test_size=0.2, random_state=0)
from sklearn.naive_bayes import MultinomialNB

model_naive = MultinomialNB().fit(X_train, y_train)
predicted_naive = model_naive.predict(X_test)

from sklearn.metrics import accuracy_score

score_naive = accuracy_score(predicted_naive, y_test)
print("Dogruluk Naive-bayes: ",score_naive)

from sklearn.metrics import confusion_matrix

print(mt.classification_report(y_test, predicted_naive))
print('F1 Skoru: %.3f' % f1_score(y_test, predicted_naive,average='macro'))
print('Duyarlılık: %.3f' % recall_score(y_test, predicted_naive,average='macro'))
print('Kesinlik: %.3f' % precision_score(y_test, predicted_naive,average='macro'))
```

Şekil 4.18. Sınıflandırma modeli kod örneği

Çizelge 4.2. Oluşturulan modele ait başarım ölçütleri

Model Başarım Ölçütleri	**Kesinlik**	**Duyarlılık**	**F1-Skoru**	**Toplam**
0	0,71	0,91	0,80	9171
1	0,72	0,62	0,66	5186
2	0,23	0,03	0,06	2088
Doğruluk			0,70	16445
Makro ortalama	0,55	0,52	0,51	16445
Ağırlıklı ortalama	0,65	0,70	0,66	16445

BoW sayısallaştırma yöntemi 4 farklı şekilde ele alınmıştır. İlk olarak sadece BoW sayısallaştırma yöntemi ile sınıflandırma modeli oluşturulmuştur. Diğer iki model ise BoW ile N-gram yöntemlerinin birlikte kullanılmasıyla oluşturulan modellerdir. BoW-N-gram(1,1) ifadesindeki n=1 yani unigramların kullanıldığı modeldir. BoW-N-gram(2,2) ifadesindeki n=2 yani bigramların kullanıldığı modeldir. Son model ise BoW-N-gram(3,3) ifadesindeki n=3 yani trigramların kullanıldığı modeldir. Tüm bu oluşturulan modeller tek tek sınıflandırma algoritmalarına uygulanmıştır. BoW ve N-gram sayısallaştırma yönteminin algoritmalar üzerindeki birlikte kullanımı Şekil 4.19'da verilmiştir.

```
cnb = Pipeline([
    ('3gram', CountVectorizer(ngram_range = (3,3))),
     ('Multi NB', MultinomialNB())])
clr = Pipeline([
    ('3gram', CountVectorizer(ngram_range = (3,3))),
     ('LogisticRegression', LogisticRegression())])
cdt = Pipeline([
    ('3gram', CountVectorizer(ngram_range = (3,3))),
     ('DecisionTree', (DecisionTreeClassifier(criterion = 'entropy'))])
crf = Pipeline([
    ('3gram', CountVectorizer(ngram_range = (3,3))),
    ('random_forest', (RandomForestClassifier()))])
cknn = Pipeline([
    ('3gram', CountVectorizer(ngram_range = (3,3))),
     ('KNeighborsClassifier', (KNeighborsClassifier(n_neighbors=10,
                                                    metric = 'minkowski'))])
cs = Pipeline([
    ('3gram', CountVectorizer(ngram_range = (3,3))),
     ('SVM', SVM(kernel='linear'))])
```

Şekil 4.19. BoW ve N-gram modeli kod örneği

Şekil 4.19'a bakıldığında KNN sınıflandırıcı algoritmasında "n_neighbors=10, metric = 'minkowski'" ifadesi olduğu görülmektedir. Bu ifadeler KNN modelinin özelliklerini belirlenmesine yardımcı olan parametrelerdir. N_neighbors=10 ifadesi en yakın 10 adet komşu seç anlamına gelmektedir. Eğer değer ataması yapılmaz ise varsayılan değer olarak 5 olarak kabul edilir. Metric=minkowski ifadesi "minkowski" uzaklığına göre uzaklık hesapla anlamına gelir. Eğer herhangi bir uzaklık metriği verilmez ise algoritma varsayılan değerini minkowski olarak kabul edilir.

TF-IDF yönteminde LabelEncoder metotu kullanılmıştır. Bu metot veriyi birebir sayısallaştırmaya yarar. Yani kategorik her veriye sayısal bir değer atar. Örneğin Negatif, Nötr ve Pozitif verisi için Negatif: 0, Pozitif: 1, Nötr: 2 değerlerini atayacak ve bu şekilde devam edecektir. TF-IDF sayısallaştırma yönteminin algoritmalar üzerindeki kullanımı Şekil 4.20'de verilmiştir.

```
data['sentiment'] = data.sentiment.map({'negative':0, 'positive':1,
                                        'neutral':2})
le = LabelEncoder()
data['sentiment'] = le.fit_transform(data['sentiment'])

cnb = Pipeline([
     ('vectorizer_tfidf',TfidfVectorizer()),
     ('Multi NB', MultinomialNB())])
clr = Pipeline([
     ('vectorizer_tfidf',TfidfVectorizer()),
     ('Random Forest', RandomForestClassifier())])
cknn = Pipeline([
     ('vectorizer_tfidf',TfidfVectorizer()),
     ('KNN', KNeighborsClassifier())])
cdt = Pipeline([
     ('vectorizer_tfidf',TfidfVectorizer()),
     ('DecisionTree', (DecisionTreeClassifier(criterion = 'entropy')))])
clr = Pipeline([
    ('vectorizer_tfidf',TfidfVectorizer()),
     ('LogisticRegression', LogisticRegression())])
cs = Pipeline([
    ('vectorizer_tfidf',TfidfVectorizer()),
     ('SVM', SVM(kernel='linear'))])
```

Şekil 4.20. TF-IDF kod örneği

Şekil 4.20'deki SVM algoritmasına bakıldığında kernel='linear' ifadesi görülüyor. Bu ifade, SVM modelinde sınıflandırma yaparken veri gruplarının doğrusal olarak ayrılmasını sağlar.

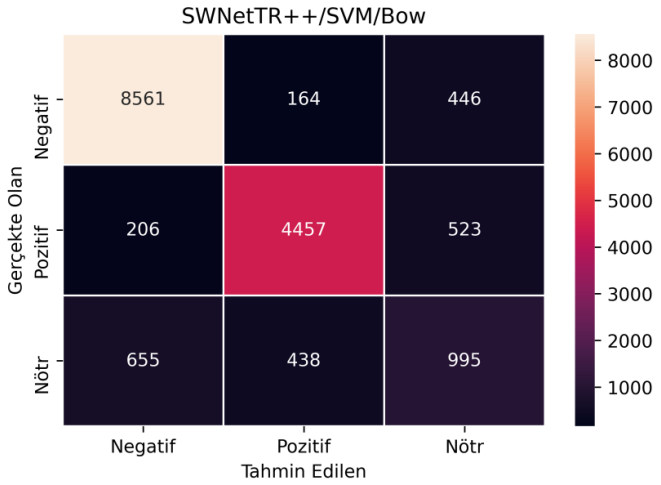
BoW ve TF-IDF ile sayısallaştırılma işlemlerinden sonra NB, LR, DT, RF, KNN ve SVM makine öğrenmesi algoritmaları ile oluşturulan her model sınıflandırma işlemine dahil edilmiştir. Sözlük tabanlı etiketlemede, BoW, BoW N-gram(1,1), BoW N-gram(2,2), BoW N-gram(3,3) ve TF-IDF kelime gömme metotları ile başarım ölçütleri ve doğruluk oranı Çizelge 4.3'de verilmiştir. Çizelge üzerinde ondalık sayılar sıfırdan sonra üç basamağa yuvarlanmış ve yüksek doğruluk oranı koyu renk ile gösterilmiştir.

Çizelge 4.3. SWNetTR++ başarım ölçütleri

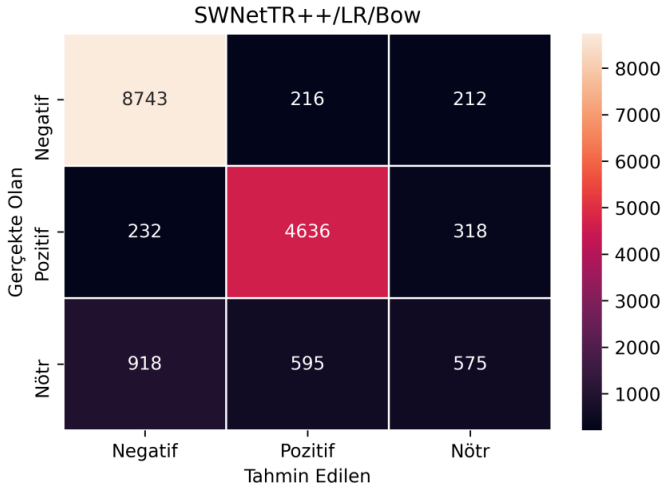
SWNetTR++		Doğruluk	Kesinlik	Duyarlılık	F1-Skoru
NB	**BoW**	**0,704**	0,554	0,519	0,507
	BoW N-gram (1,1)	0,689	0,569	0,494	0,478
	BoW N-gram (2,2)	0,664	0,551	0,474	0,463
	BoW N-gram (3,3)	0,624	0,510	0,442	0,435
	TF-IDF	0,611	0,672	0,386	0,341
LR	**BoW**	**0,848**	0,752	0,708	0,716
	BoW N-gram (1,1)	0,775	0,644	0,616	0,619
	BoW N-gram (2,2)	0,677	0,572	0,495	0,497
	BoW N-gram (3,3)	0,625	0,565	0,442	0,409
	TF-IDF	0,777	0,689	0,592	0,579
DT	**BoW**	**0,616**	0,510	0,503	0,506
	BoW N-gram (1,1)	0,605	0,490	0,485	0,487
	BoW N-gram (2,2)	0,573	0,480	0,473	0,474
	BoW N-gram (3,3)	0,613	0,505	0,434	0,433
	TF-IDF	0,583	0,467	0,462	0,464
RF	**BoW**	0,669	0,579	0,507	0,517
	BoW N-gram (1,1)	0,690	0,593	0,520	0,532
	BoW N-gram (2,2)	0,602	0,507	0,471	0,475
	BoW N-gram (3,3)	0,617	0,529	0,424	0,416
	TF-IDF	**0,697**	0,642	0,513	0,521
KNN	**BoW**	0,590	0,511	0,465	0,464
	BoW N-gram (1,1)	0,590	0,480	0,406	0,393
	BoW N-gram (2,2)	0,373	0,517	0,374	0,298
	BoW N-gram (3,3)	0,575	0,561	0,350	0,281
	TF-IDF	**0,613**	0,488	0,437	0,436
SVM	**BoW**	**0,852**	0,765	0,756	0,761
	BoW N-gram (1,1)	0,753	0,632	0,619	0,624
	BoW N-gram (2,2)	0,659	0,539	0,495	0,502
	BoW N-gram (3,3)	0,553	0,449	0,355	0,450
	TF-IDF	0,773	0,656	0,593	0,584

Çizelge 4.3‘de yer alan doğruluk oranları incelendiğinde SWNetTR++ ile etiketlenen metinler ile kullanılan modelde BoW– SVM ile 0,852‘lik en yüksek doğruluk oranına ulaşılmıştır. Bu değere en yakın diğer değer ise BoW-LR ikilisine ait 0,848’lik başarı oranıdır. Veri sayısallaştırma sonuçlarının doğruluk oranları karşılaştırıldığında BoW ile sayısallaştırma yönteminde makine

öğrenmesi algoritmalarının başarı oranının diğer yöntemlere göre kısmen daha başarılı sonuçlar elde edildiği gözlemlenmiştir. En iyi performansı gösteren BoW– SVM sınıflandırma işlemine ait hata matrisi Şekil 4.21'deki gibidir. Hemen sonrasında en yakın başarı oranına sahip BoW-LR ikilisinin hata matrisi de Şekil 4.22'de verilmiştir.



Şekil 4.21. BoW-SVM hata matrisi

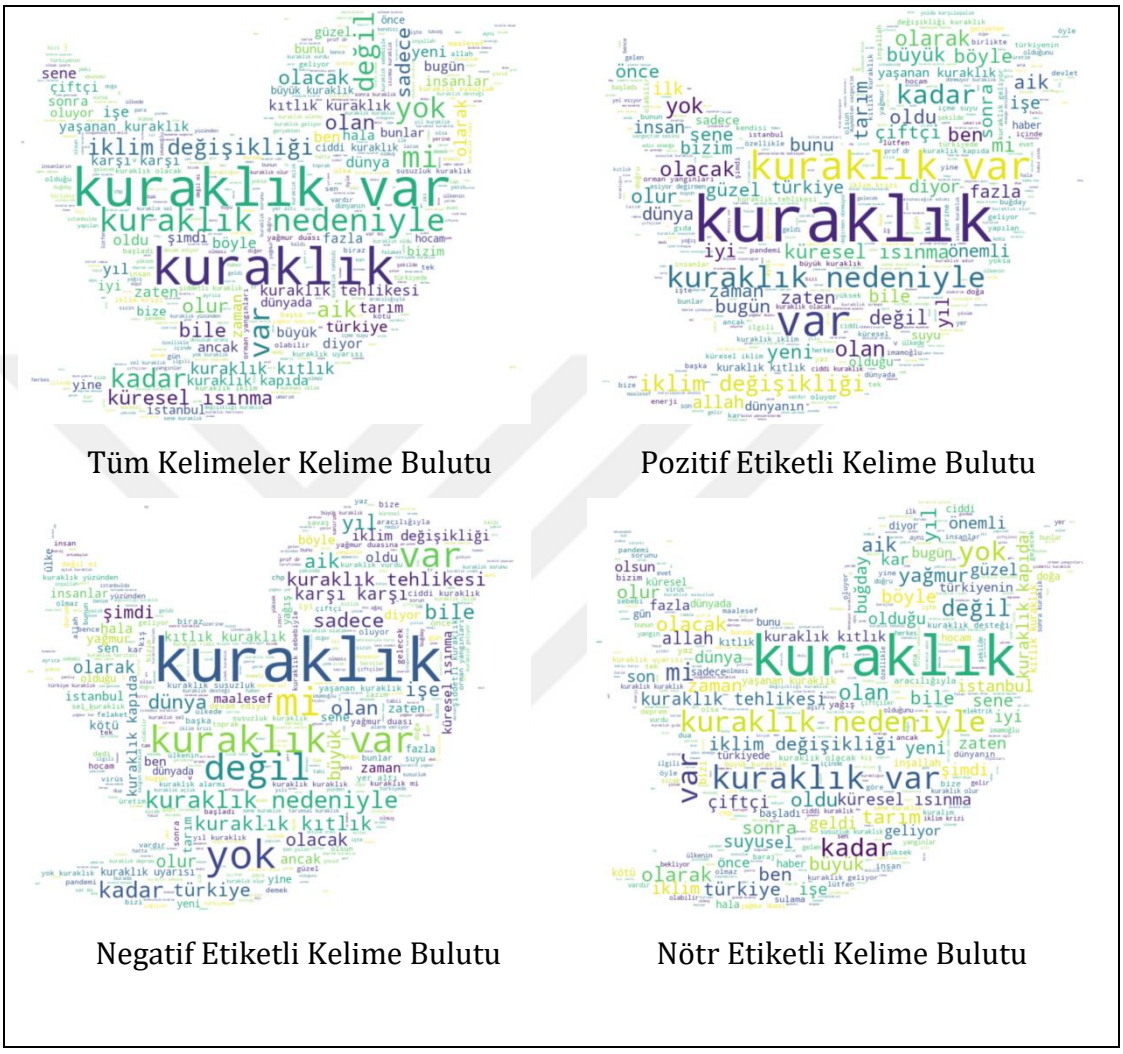


Şekil 4.22. BoW-LR hata matrisi

4.7. Kelime Bulutu

Kelime bulutu, görselleştirme yöntemlerinden uygulaması en kolay olan ve en sık tercih edilen bir yöntemdir. Veri seti içerisinde en sık tekrar eden kelimelerin görseleştirilmesini ve bu kelimelere bakarak veri seti ile ilgili

yorumlar yapılabilmesini sağlar. Duygu analizi gerçekleştirildikten sonra pozitif, negatif ve nötr etiketlere ait kelime bulutu ile görselleştirilme yöntemi uygulanmıştır. Görseleştirme işleminde kelime sınırı 400 olarak belirlenmiştir. Tüm kelimelere ve duygu sınıflarına ait kelime bulutları Şekil 4.23'deki gibidir.



Şekil 4.23. Duygu analizi kelime bulutları

Kelime bulutunda sık geçen pozitif anahtar kelimelerden bazılarının tweet paylaşımlarındaki örneklerine bakıldığında "umarım oraya da yağar gerçekten bu kar çok iyi geldi kuraklık var epey zamandır" tweetini görüyoruz. Negatif anahtar kelimelerden bazılarının tweet paylaşımlarındaki örneklerine bakıldığında ise "kuraklık falan da yok umrumda değil" olduğunu görmekteyiz. Nötr anahtar kelimelerden bazılarının tweet paylaşımlarındaki örneklerinde ise "bir karar ver kuraklık var mı yok mu" tweetini görebiliriz.

5. TARTIŞMA VE SONUÇLAR

Sosyal medya platformları özellikle hedef kitleden geri dönüşüm alabilmek için önemli ve yeterince geniş kaynaklardır. Bu platformlardaki verileri manuel olarak analiz etmek imkânsızdır. Bu noktada sosyal platformları gözlemlemek için duygu analizi araçları devreye girerler. Bir metin hakkındaki tutum, ancak duygu analizi sonucu ile anlaşılabilir. Duygu analizi, DDİ ve metin madenciliği için zor bir çalışma alanıdır.

Özgün ve güncel konular üzerine yapılan çalışmalar son zamanlarda hız kazanmıştır. Duygu analizi çalışmalarında genellikle film, otel, marka vb. konular ön plandadır. Doğal afetler üzerindeki etkileşimlerine yönelik yapılan çalışmalarda sadece deprem konusu ele alındığı görülmüştür. Kuraklık üzerine bir çalışmaya rastlanılmamıştır. Kuraklık insanlığın başladığı zamandan bu zamana kadar meydana gelen tüm afet türleri incelendiğinde, bu türler arasında en önemlisi ve meteorolojik afetler sıralamasında en tehlikelisidir.

Bu tez çalışmasında makine öğrenmesi algoritmalarında kullanılacak veri seti için SWNetTR++ sözlüğü ile etiketlenmiş 82221 adet tweet içeren veri kümesi girdi olarak kullanılmıştır. Duygu analizi sonucunda kuraklık etiketi ile paylaşılmış tweet mesajlarının duygu durumu %56'lık oranla negatif çıktığı sonucuna ulaşılmıştır. Çalışma kapsamında BoW ve TF-IDF yöntemleri kullanılarak veri sayısallaştırma işlemleri gerçekleştirilmiştir. Sonrasında veri seti, eğitim ve test için ayrılıp eğitilerek bir model oluşturuldu ve her oluşturulan model NB, LR, DT, RF, SVM ve KNN makine öğrenmesi algoritmaları ile sınıflandırma işlemine alındıktan sonra performansları karşılaştırılmıştır. Kullanılan makine öğrenmesi algoritmalarının sahip olduğu metriklerin doğru değerlerde seçimi sınıflandırma performanslarını doğrudan etkilediği sonucuna varılmıştır.

Her makine öğrenimi algoritmasının performans başarısının en iyi olduğu modelleri kıyaslanmştır. NB 0,704'lük oranla BoW, LR 0,848'lik oranla BoW, DT 0,616'lık oranla BoW, RF 0,697'lik oranla TF-IDF, KNN 0,613'lük oranla TF-IDF

ve son olarak da SVM 0,852'lik oranla BoW yönteminde en yüksek doğruluk değerlerini elde edilmiştir. Genel olarak sonuçlara bakıldığında en verimli sayısallaştırma yönteminin BoW olduğu görülmektedir.

Tez çalışması literatürdeki bazı çalışmalar ile bazı benzerlikler ve farklılıklar barındırmaktadır. İlhan ve Sağaltıcı (2020), Twitter'da kullanıcılara ait sınıflandırılmış tweet verileri üzerinde çeşitli makine öğrenme teknikleri kullanılarak bir duygu analizi çalışması yapmışlardır. Etiketleme olarak sadece pozitif ve negatif duygu durumlarını ele almışlardır. Bu çalışmada olduğu gibi, onlarda sözlük tabanlı yöntem tercih etmişler fakat sözlük olarak WordNet kullanmışlardır. Sınıflandırmada ise Naive Bayes ve SVM sınıflandırıcıları ile başarı ölçülmüştür. Bu tez çalışmasında altı adet sınıflandırıcı ile modelleme yapılırken, çalışmalarında sadece iki algoritma kullanmışlardır. Naive Bayes ile %42 ve SVM ile de %64 başarı oranları almışlardır.

Kumaş (2021), yaptığı çalışmada 32000 adet Türkçe tweet kullanmıştır. Bu tweetler yarı yarıya negatif ve pozitif olarak etiketli bir şekilde ayrılmıştır. Veri setine metin madenciliği yöntemleri uygulayarak duygu analizi yapmıştır. NB, KNN, SVM, LR ve DT sınıflandırma algoritmalarını kullanmıştır. Sınıflandırma sonuçlarını f1 skoru üzerinden kıyaslamış ve en iyi sonucu %73 ile SVM sınıflandırıcısıyla elde etmiştir. Kullandığı yöntem ve algoritmalar bu tez çalışmasına benzese de f1 skoru üzerinden yaptığı başarım oranıyla farklılık oluşturmaktadır. Bu tez çalışmasında ki en iyi modelleme f1skoru %76'dır. Eryılmaz vd., (2020) yapmış oldukları çalışmada alakası olmayan içerikleri barındıran ve kişiye zarar veren yani spam epostaların makine öğrenmesi yöntemleri ile tespitini amaçlamışlardır. İki adet Türkçe veri seti kullanılmıştır. İlk olarak veri setlerinde öznitelik çıkarımı ve özellik seçimi yapılmıştır. Veri setlerinde çalışılan makine öğrenme algoritmaları RF, DT, SVM, KNN, LR, NB'dir. Veri kümeleri DDİ kütüphanesi olan "Zemberek" yazılımından geçirilip normalleştirilmiştir. Bu tez çalışmasındaki benzer kısımlar TF-IDF ile öznitelik çıkarımı ve Zemberek ile normalleştirmedir. Elde edilen sonuçlara göre SVM algoritması ile %98'in başarım oranı elde etmişlerdir. Yaptıkları çalışmada, bu tezde kullanılandan farklı birçok öznitelik çıkarım yöntemi kullanmışlardır.

Yüksek başarım oranının elde edilmesi de bundan dolayıdır. Çizelge 5.1'de yapılan bazı çalışmalar ve sonuçları gösterilmiştir.

Çizelge 5.1. Çalışmalar ve sonuçları

	Veri Seti	**Ön İşleme**	**Sözlükler Algoritmalar**	**Başarı Sonuçları**
Alfarrarjeh vd. (2017)	Sandy Kasırgası ve Napa Depremi	Detay verilmemiş	SentiBank SentiStrength, CoreNLP NLTK	Sözel yorum yapılmış
Aziz vd. (2019)	10 adet afet etiketi için 32117 adet tweet	Detay verilmemiş	NB	%47 Negatif tahmin
Khaleq ve Ra (2019)	Afet yönetimi ile ilgili etiketlere sahip tweet ler	Manuel	LR, SVM, NB, Stanford	%85 doğruluk
Kumar (2020)	Nepal depremi ile ilgili 201457 tweet	Manuel	Word Tree	Sözel yorum yapılmış
Noor (2020)	Kenya Para Birimi	N-gram	NB	Unigram - %70 Bigram - %64
Sarıman ve Mutaf (2020)	COVID-19	BoW, TF-IDF	LR	LR - %82
Kumaş (2021)	32000 tweet	TF-IDF	NB, KNN, SVM, LR, DT	SVM - %73
Sham ve Mohamed (2022)	Weather Sentiment DecarboNet Kaggle	BoW, TF-IDF	SentiWordNt, TextBlob, VADER, SentiStrength SVM, NB, LR	VADER - %57 LR - %70 TextBlob-LR- %75
Akdeniz (2022)	Uzaktan Eğitim	BoW, TF-IDF, Word2Vec	Manuel, TextBlob, VADER LR, SGD, SVM, RF, NB	Manuel Etiketleme Bow-LR - %84
Dankhara (2022)	Sentiment 140	Count Vectorizer	Subjectivity Lexicon KNN, NB, RF, SVM	Subjectivity Lexicon - %59 SVM - %76
Bu Çalışma (2023)	**Kuraklık ile ilgili 96401 tweet**	**BoW, TF-IDF**	**SWNetTR++ RF, DT, SVM, KNN, LR, NB**	**SVM-BoW - %85 LR-BoW - %84**

Bu tez çalışmasından elde edilen analizin sonuçları göz önüne alınırsa ülkemizde, kuraklığa hazırlıklı olmak, önlemek ve zararlarını azaltmayı sağlamak adına katkıda bulunmamız gerekmektedir. Pozitif etiketli bir tweet örneği verecek olursak: "susuz kalmamak için şimdiden tedbirini al karantina ve kuraklık sebebiyle evlerimizde kaldığımız şu süreçte su tüketimi artmıştır sularımızı israf etmeden özenle kullanalım israfı önleyecek öneri ve tavsiyeleriniz varsa yorum kısmına yazabilirsiniz" tweeti ile binlerce yorumda kullanıcılar önerilerde bulunabilir. Böylelikle kullanıcılar farkındalık oluşturabilir. Yaklaşık olarak Türkiye'de 12 milyon Twitter kullanıcısı vardır. Bu kullanıcılar hep birlikte ele ele verip gündem etiketleri oluşturabilirler. Kuraklık konusunda analizler yaparak farkındalık sağlanabilir ve böylece literatür boşlukları doldurulabilir. Bu tez çalışması benzeri çalışmalar sayesinde kuraklığın olumsuz etkilerini azaltmak adına yapılacak doğru planlamaları kolayca tespit edip kullanıcılarla bu sonuçlar paylaşılabilinir. Kuraklık yalnızca fiziksel bir olay veya bir doğa olayı olarak görülmemelidir buna sebep olabilecek insan davranış ve duygularının da olabileceği unutulmamalıdır.

Tez çalışması ile elde edilen sonuçlar Türkiye-Kuraklık ilişkisin ilişkisini büyük ölçüde yansıtmıştır. Toplumun bu tehditten ne kadar haberdar ve ne derece bilinçli olduklarının analizinin çıkarımlarına ulaşma imkânı sunulmuştur.

Bu çalışmadaki sonuçlar üç adet duygu (pozitif, nötr, negatif) ifadesinin kullanılarak elde edilmiştir. Gelecek çalışmalarda üzgün, mutlu, sinirli, şaşkın gibi ya da emojiler üzerinden daha ayrıntılı duygu ifadeleri belirten çalışmalar yapılabilir. İngilizce metinler için geliştirilmiş TextBlob gibi birçok hazır model bulunmaktadır. Türkçe metinler için de hazır modeller geliştirilebilir. Sözlük tabanlı yaklaşımlarda ise kullanılan genel ve duygu sözlüklerinin Dünyada yaygın olarak kullanılan WordNet ve SentiWordnet kadar geniş kapsamlı olmadığı gözlemlenmiştir. Bu bağlamda, Türkçe duygu analizinde kullanılan duygu sözlükleri daha fazla geliştirilebilir.

KAYNAKLAR

Akçakaya, S., 2019. All-Words Word Sense Disambiguation In Turkish. Işık Üniversitesi, Fen Bilimleri Enstitüsü, Yüksek Lisans Tezi, 43s, İstanbul.

Akdeniz, A., 2022. Makine Öğrenmesi Yöntemleri Kullanılarak Uzaktan Eğitim Konulu Türkçe Tweetlerin Duygu Analizi. Konya Teknik Üniversitesi, Lisansüstü Eğitim Enstitüsü, Yüksek Lisans Tezi, 90s, Konya.

Akdeniz, F., 2021. Belediye Hizmetlerinde Memnuniyetin Sosyal Medya Aracılığı İle Değerlendirilmesi. Sakarya Üniversitesi, İşletme Enstitüsü, Yüksek Lisans Tezi, 106s, Sakarya.

Aksu, M. Ç. & Karaman, E., 2020. FastText ve Kelime Çantası Kelime Temsil Yöntemlerinin Turistik Mekânlar İçin Yapılan Türkçe İncelemeler Kullanılarak Karşılaştırılması. Avrupa Bilim ve Teknoloji Dergisi, (20), 311-320.

Alfarrarjeh, A., Agrawal, S.,Ho Kim, S. And Shahabi. C., 2017. Integrated Media Systems Center, University of Southern California, Los Angeles, CA 90089, USA.

Amanet H., 2017. Türkçe Sosyal Medya Metinlerinde Duygu Analizi, Karadeniz Teknik Üniversitesi, Fen Bilimleri Enstitüsü, Yüksek Lisans Tezi, 94s, Trabzon.

Anaconda, 2022. Erişim tarihi: 04.12.2022. Wikipedia: https://en.wikipedia.org/wiki/Anaconda_(Python_distribution).

Anonim, 2015. Text Mining et Webmarketing, Erişim Tarihi: 23.11.2022. https://www.mauricelargeron.com/analyse-de-texte-et-seo/.

Anonim, 2022. Text Vectorization and Transformation Pipelines. Erişim Tarihi: 22.12.2022. https://www.oreilly.com/library/view/applied-text-analysis/9781491963036/ch04.html.

Arı, O., 2022. Tüketici Yorumlarının Fayda Düzeyinin Tahminlenmesine Yönelik Bir Araştırma: Makine Öğrenmesi Algoritmalarının Karşılaştırılması. Sakarya Üniversitesi, İşletme Enstitüsü, Yüksek Lisans Tezi, 101s, Sakarya.

Atamedya, 2019. Erişim tarihi: 04.12.2022. https://ata.com.tr/blog-detay/visual-studio-code-nedir-144.

Ayan, B., 2020. Twitter Üzerindeki İslamofobik Tweetlerin Duygu Analizi İle Tespiti. Gazi Üniversitesi, Bilişim Enstitüsü, Yüksek Lisans Tezi, 81s, Ankara.

Ayaz, M., 2021. Makine Öğrenmesi Algoritmaları İle Covıd-19 Hastalarının Belirlenmesi. Pamukkale Üniversitesi, Fen Bilimleri Enstitüsü, Yüksek Lisans Tezi, 56s, Denizli.

Aydın, K. A., 2019. Veri Madenciliği ve Uygulamaları. https://www.slideshare.net/KazmAnlAYDIN/metin-madencilii-nedir-sunum.

Ayık, Y. Z., Özdemir, A., & Yavuz, U., 2007. Lise Türü ve Lise Mezuniyet Başarısının, Kazanılan Fakülte ile İlişkisinin Veri Madenciliği Tekniği İle Analizi. Atatürk Üniversitesi Sosyal Bilimler Enstitüsü Dergisi, 10(2), 441-454.

Aydoğan, M. & Karcı, A., 2019. Kelime Temsil Yöntemleri İle Kelime Benzerliklerinin İncelenmesi. Çukurova Üniversitesi, Mühendislik-Mimarlık Fakültesi Dergisi, 34 (2), 181-196.

Aziz, K., Zaidouni, D. and Bellafkih, M., 2019. Social Network Analytics: Natural Disaster Analysis Through Twitter, 2019 Third International Conference on Intelligent Computing in Data Sciences (ICDS), Marrakech, Morocco.

Ballı, Ç., 2021. Doğal Dil İşleme İle Türkçe İçerikli Paylaşımlardan Sosyal Medya Kullanıcılarının Duygu Analizi. Ankara Üniversitesi, Fen Bilimleri Enstitüsü, Yüksek Lisans Tezi, 65s, Ankara.

Barzenji, H., 2021. Sentiment Analysis Of Twitter Texts Using Machine Learning Algorithms. Sakarya University, Institute Of Science And Technology, M.Sc. Thesis, 64s, Sakarya.

Başkaya, F., 2017. Kısa Metinlerden Sosyal Duygu Sınıflandırma İçin Makine Öğrenmesi Tabanlı Yöntemlerin Geliştirilmesi. Fırat Üniversitesi, Fen Bilimleri Enstitüsü, Yüksek Lisans Tezi, 59s.

Baykal, A., 2006. Veri Madenciliği Uygulama Alanları . Dicle Üniversitesi, Ziya Gökalp Eğitim Fakültesi Dergisi, (7), 95-107. Retrieved from https://dergipark.org.tr/tr/pub/zgefd/issue/47963/606848

Bhadane, C., Dalal, H., & Doshi, H., 2015. Sentiment Analysis: Measuring Opinions. Procedia Computer Science, 45, 808-814.

Bozkır, A., Gök, B., & Sezer, E., 2009. Öğrenci Seçme Sınavında (ÖSS) Öğrenci Başarımını Etkileyen Faktörlerin Veri Madenciliği Yöntemleriyle Tespiti, 5. Uluslararası İleri Teknolojiler Sempozyumu (IATS'09), 13-15 Mayıs, Karabük Üniversitesi, Karabük, 37-43

Büyükbaş, E., & Ormanoğlu, B., 2013 . Afetler ve Afet Yönetiminde Meteorolojinin Yeri. Türk İdare Dergisi, 476, 13-46.

Contractor, D., Faruquie, T. A., & Subramaniam, L. V., 2010. Unsupervised Cleansing of Noisy Text. In Coling 2010: Posters (pp. 189-196).

Coşkun, C., & Baykal, A., 2011. Veri Madenciliğinde Sınıflandırma Algoritmalarının Bir Örnek Üzerinde Karşılaştırılması. Akademik Bilişim, 11, 51-58.

Çelik, Ö. & Koç, B. C., 2021. TF-IDF, Word2vec ve Fasttext Vektör Model Yöntemleri İle Türkçe Haber Metinlerinin Sınıflandırılması, Dokuz Eylül Üniversitesi Mühendislik Fakültesi Fen ve Mühendislik Dergisi, 23 (67), 121-127.

Çetin, M., 2009. Bir Üretim İşletmesinde Veri Madenciliği Uygulaması. Sakarya Üniversitesi, Fen Bilimleri Enstitüsü, Yüksek Lisans Tezi, 110s, Sakarya.

Çınar, A., 2020, Sınıflandırma Algoritmaları İle Bir Metin Madenciliği Uygulaması: Veri Madenciliği Ve Makine Öğrenmesi, Temel Kavramlar, Algoritmalar, Uygulamalar, Çağlayan Kitapevi ve Eğitim Çözümleri Tic. A.Ş., İstanbul, 105-140.

Çoban Ö., 2016. Metin Sınıflandırma Teknikleri ile Türkçe Twitter Duygu Analizi, Atatürk Üniversitesi Fen Bilimleri Enstitüsü, Yüksek Lisans Tezi, 105s, Erzurum.

Dankhara, D., 2022. A Review of Sentiment Analysis of Tweets. https://www.researchgate.net/publication/360076165_A_Review_of_Sentiment_Analysis_of_Tweets#fullTextFileContent

Dehkharghani, R., Saygin, Y., Yanikoglu, B., & Oflazer, K., 2016. SentiTurkNet: a Turkish Polarity Lexicon for Sentiment Analysis. Language Resources and Evaluation, 50(3), 667-685.

Delibaş, A. 2008. Doğal Dil İşleme İle Türkçe Yazım Hatalarının Denetlenmesi. İstanbul Teknik Üniversitesi, Fen Bilimleri Enstitüsü, Yüksek Lisans Tezi, 78s, İstanbul.

Demirci, G., 2020. Understanding Twitter Users' Behaviour by Social Network Analysis During Disasters.(M.Sc. Thesis, Department of Industrial Engineering Industrial Engineering Programme)

Desai, R., 2021. How To Scrape Millions of Tweets Using Snscrape. https://medium.com/dataseries/how-to-scrape-millions-of-tweets-using-snscrape-195ee3594721.

DevHunteryz, 2019. Naive-Bayes Sınıflandırıcı. Erişim tarihi: 05.12.2022. https://devhunteryz.wordpress.com/2019/12/02/naive-bayes-siniflandirici/.

Eliöz. R., 2021. NLP ile Metin Verisi Ön İşleme. Erişim Tarihi: 22.12.2022. https://eliozrumeysa.medium.com/nlp-ile-metin-verisi-%C3%B6n-i%CC%87%C5%9Fleme-9c852b6de3b3.

Elmas, Ş., 2019. Sosyal Medya Mesajlarının Veri Madenciliği Yöntemi İle Duygu Analizi (Sivas İli Örneği). Sivas Cumhuriyet Üniversitesi, Sosyal Bilimler Enstitüsü, Yüksek Lisans Tezi, 99s, Sivas.

Ergün, K., 2012. Metin Madenciliği Yöntemleri İle Ürün Yorumlarının Otomatik Değerlendirilmesi. Sakarya Üniversitesi, Fen Bilimleri Enstitüsü, Doktora Tezi, 91s, Sakarya.

Ergün, K., 2017. Pusula Yayıncılık. Erişim tarihi: 22.12.2022. http://kergun.baun.edu.tr/veri_madenciligi_hafta11.pdf.

Ertoy, U., 2022. Twitter Verileri Kullanılarak Koronavirüs Aşıları Hakkındaki Kamu Algısının Zaman İçindeki Değişiminin Yapay Zekâ Destekli Duygu Analizi İle İncelenmesi. Kütahya Dumlupınar Üniversitesi, Lisansüstü Eğitim Enstitüsü, Yüksek Lisans Tezi, 83s, Kütahya.

Eryılmaz, E. E., Şahin, D. Ö., Kılıç, E., 2020. Türkçe İstenmeyen E-postaların Farklı Öznitelik Seçim Yöntemleri Kullanılarak Makine Öğrenmesi Algoritmaları ile Tespit Edilmesi. Türkiye Bilişim Vakfı Bilgisayar Bilimleri ve Mühendisliği Dergisi, 13(2), 57-77.

F. Sağlam, B. Genç, & H. Sever, 2019. Extending a Sentiment Lexicon with Synonym-Antonym Datasets: SWNetTR++, Turkish Journal of Electrical Engineering and Computer Sciences, 27, 1806-1820.

Fayyad, U., Piatetsky-Shapiro, & Smyth, P., 1996. The KDD Process for Extracting Useful Knowledge from Volumes of Data, Communıcatıons Of The ACM, 39(11), 27–34.

Gholizadeh, S., 2022. Top Popular Python Libraries in Research. J Robot Auto Res 3(2), 142-145.

GitHup, 2021. Erişim tarihi: 04.12.2022. https://github.com/ahmetaa/zemberek-nlp.

GitHup, 2022. Erişim tarihi: 05.12.2022. https://github.com/merve-karagoz/makineogrenmesi-TRVersion.

Hamde, M. A., 2018. Kurumsal Belgelere (Metin Verilerine) Metin Madenciliği Tekniği İle Erişimin Değerlendirilmesi: Türk Özel Sektörüne Yönelik Bir İnceleme. İstanbul Üniversitesi, Sosyal Bilimler Enstitüsü, Yüksel Lisans Tezi, 342s, İstanbul.

Hari, K. C., 2022. Python For Data Analysis: Python Programming. Blue Rose Publishers.

Hatipoğlu, E., 2018. Machine Learning — Classification — Logistic Regression — Part 8. Erişim tarihi: 04.12.2022. https://medium.com/@ekrem.hatipoglu/machine-learning-classification-logistic-regression-part-8-b77d2a61aae1.

Hearst, M. A., 1999. Untangling Text Data Mining. In Proceedings Of The 37th Annual Meeting Of The Association For Computational Linguistics, 3-10.

Hotz, N., 2022. What is CRISP DM? Erişim tarihi: 31.12.2022. https://www.datascience-pm.com/crisp-dm-2/.

İlhan N., & Sağaltıcı D., 2020. Twitter'da Duygu Analizi. Harran Üniversitesi Mühendislik Dergisi, 5(2): 146-156.

Jamali, M., Nejat, A., Ghosh, S. and Cao, G., 2018. Social Media Data and Post-Disaster Recovery, International Journal of Information Management, 44, s. 25–37.

Karabulut Y., 2018. Twitter Üzerine Yapılan Türkçe Paylaşımlar İçin Etiket Analiz Aracı. Süleyman Demirel Üniversitesi, Fen Bilimleri Enstitüsü, Yüksek Lisans Tezi, 71s, Isparta.

Karamanlı, E., 2019. Makine Öğrenmesi Algoritmaları Kullanarak, Metin Madenciliği Ve Duygu Analizi İle Müşteri Deneyiminin Geliştirilmesi. İstanbul Üniversitesi, Sosyal Bilimleri Enstitüsü, Yüksek Lisans Tezi, 55s, İstanbul.

Karaöz Akın, B., & G. Şimşek, U., 2018. Adaptif Öğrenme Sözlüğü Temelli Duygu Analiz Algoritması Önerisi. Bilişim Teknolojileri Dergisi, 11(3), s. 245-253.

Kargın, T., 2022. Algoritma ve Programlama Dünyası — Kütüphane Çağırma. Erişim tarihi: 04.12.2022. https://medium.com/kodcular.

Kaşgarlı, K., 2021. Twitter Sentiment Analysis Via Machine Learning. Kadir Has University School Of Graduate Studies Department Of Engineering And Natural Sciences, Master's Degree Thesis, 85s, Istanbul.

Kemaloğlu, N., Küçüksille, E., & Özgünsür, M. 2021. Turkish Sentiment Analysis On Social Media. Sakarya University Journal of Science, 25(3), 629-638.

Khaleq, A. & Ra, I., 2019. Twitter Analytics for Disaster Relevance and Disaster Phase Discovery: Volume 1, Proceedings of the Future Technologies Conference.

Kırcı, P., & Gülbak, E., 2020. Instagram Verileri ile Duygu Analizi. Avrupa Bilim ve Teknoloji Dergisi, 360-364.

Kızılırmak, E., 2020. İngilizce–Türkçe Çeviri Metinlerde Levenshtein Uzaklığı İle Desteklenmiş Çapa Tabanlı Cümle Eşleme. T.C. Maltepe Üniversitesi, Lisansüstü Eğitim Enstitüsü, Yüksek Lisans Tezi, 24s, İstanbul.

Korkusuz, R., 2019. Futbola İlişkin Twitter Paylaşımlarının Duygu Analizi. Trakya Üniversitesi, Fen Bilimleri Enstitüsü, Yüksek Lisans Tezi, 89s, Edirne.

Kumar, P., 2020. Twitter, Disasters And Cultural Heritage: A Case Study Of The 2015 Nepal Earthquake. DOI: 10.1111/1468-5973.12333.

Kumaş, E., 2021. Türkçe Twitter Verilerinden Duygu Analizi Yapılırken Sınıflandırıcıların Karşılaştırılması, ESTUDAM Bilişim Dergisi, 2 (2), 1-5.

Kündüm, D., 2021. Twitter Üzerinden Müşteri Duygularının Analiz Edilerek Türkiye'deki Telekom Operatörleri İle İlgili Müşteri Memnuniyetinin Değerlendirilmesi. T.C. Doğuş Üniversitesi, Lisansüstü Eğitim Enstitüsü, Yüksek Lisans Tezi, 33s, İstanbul.

Matplotlib, 2022. Erişim tarihi: 04.12.2022. https://matplotlib.org/.

Medhat, W., Hassan, A., & Korashy, H., 2014. Sentiment Analysis Algorithms And Applications: A Survey. Ain Shams Engineering Journal, 5(4), 1093-1113.

Mert, 2021. Erişim tarihi: 03.12.2022. https://mertmekatronik.com/programlama-dilleri-icin-gelistirme-araclari.

MGM, 2022. Erişim Tarihi: 21.11.2022. https://mgm.gov.tr/veridegerlendirme/kuraklik-analizi.aspx

MMO, 1999. Meteorolojik Karakterli Doğal Afetler ve Meteorolojik Önlemler Raporu, Ankara

Nacar, E., & Erdebilli, B., 2021. Makine Öğrenmesi Algoritmaları İle Satış Tahmini. Journal of Industrial Engineering 32(2), 307-320.

Nagpal, A., & Gabrani, G., 2019. Python For Data Analytics, Scientific And Technical Applications. In 2019 Amity International Conference On Artificial Intelligence (AICAI), pp. 140-145. IEEE.

Nalçakan, Y., Bayramoğlu, S., & Tuna, S., 2015. Sosyal Medya Verileri Üzerinde Yapay Öğrenme İle Duygu Analizi Çalışması. Trakya Üniversitesi.

Nasukawa, T. & Yi, J., 2003. Sentiment Analysis: Capturing Favorability Using Natural Language Processing. In Proceedings Of The KCAP-03, 2nd Intl. Conf. On Knowledge Capture.

NLTK, 2022. Documentation-Natural Language Toolkit. Erişim tarihi: 04.12.2022. http://www.nltk.org/.

Noor, M., 2020. Sentiment Analysis On New Currency In Kenya Using Twitter Dataset. Istanbul Commerce University, Graduate School Of Natural And Applied Sciences, Master Thesis Computer Engineering Department, 42s, Istanbul.

Noyan, M. 2019. Erişim tarihi: 22.12.2022. https://merveenoyan.medium.com/do%C4%9Fal-dil-i%CC%87%C5%9Fleme-natural-language-processing-2d7c72daf245.

NumPy, 2022. Erişim tarihi: 04.12.2022. https://numpy.org/.

Nvidia. 2019. Erişim tarihi: 30.11.2022. https://blogs.nvidia.com/blog/2016/07/29/whats-difference-artificial-intelligence-machine-learning-deep-learning-ai/.

Oflazer, K., & Bozşahin H.C., 2006. Türkçe Doğal Dil İşleme. Ç.Ü. Türkoloji Makale Bilgi Sistemi. http://turkoloji.cu.edu.tr/DILBILIM/turkce_dogal_dil_isleme.pdf

Oğuz, B., 2009. Metin Madenciliği Teknikleri Kullanılarak Kulak Burun Boğaz Hasta Bilgi Formlarının Analizi. Akdeniz Üniversitesi, Sağlık Bilimleri Enstitüsü, Yüksek Lisans Tezi, 58s, Antalya.

Oğuz, B., Bilge, U., & Saka O., 2007. Tıpta Metin Madenciliği. Tıp Bilişimi, 7, 15-18.

Olgun, O., 2018. Veri Kaynağı Olarak Sosyal Ağlar. Erişim tarihi: 02.12.2022. https://tr.linkedin.com/pulse/veri-kayna%C4%9Fi-olarak-sosyal-a%C4%9Flar-oktay-olgun.

Onan, A., 2017. Twitter Mesajları Üzerinde Makine Öğrenmesi Yöntemlerine Dayalı Duygu Analizi. Yönetim Bilişim Sistemleri Dergisi, 3(2), s. 1-14.

Oracle, 2022. Erişim tarihi: 23.12.2022. https://www.oracle.com/database/what-is-database/.

Özdeş, M., 2017. Büyük Veri Araçlarını Kullanarak Duygu Analizi Gerçekleştirimi, Pamukkale Üniversitesi, Fen Bilimleri Enstitüsü, Yüksek Lisans Tezi,75s, Denizli.

Özkan, T., 2019. Sosyal Medya Verilerinden Duygu Analizi Yöntemi İle Seçim Sonuçlarının Mekânsal Tahmini: Kocaeli İli Örneği. Kocaeli Üniversitesi, Fen Bilimleri Enstitüsü, Yüksek Lisans Tezi, 92s, Kocaeli.

Öztürk, M., 2022. Python ile Sınıflandırma Analizleri – Rastgele Orman (Random Forest) Algoritması. Erişim tarihi: 06.12.2022. https://miracozturk.com/.

Özyurt, B., & Akçayol, M. A., 2018. Fikir Madenciliği ve Duygu Analizi, Yaklaşımlar, Yöntemler Üzerine Bir Araştırma. Selçuk Üniversitesi Mühendislik, Bilim ve Teknoloji Dergisi, 6(4), 668-693.

Pandas, 2022. Erişim tarihi: 04.12.2022. https://pandas.pydata.org/pandasdocs/stable/getting_started/overview.html.

Partigöç, N. S., & Soğancı, S., 2019. Küresel İklim Değişikliğinin Kaçınılmaz Sonucu: Kuraklık. Resilience, 3(2), 287-299.

Project Jupyter, 2022. Erişim tarihi: 04.12.2022. Wikipedia: https://en.wikipedia.org/wiki/Project_Jupyter.

Python, 2022. Erişim tarihi: 03.12.2022. https://docs.python.org/3/.

Roesslein, J., 2009. Tweepy Documentation. http://tweepy. readthedocs. io/en/v3, 5.

Sar, K. T., 2021. Yapay Sinir Ağları Ve BERT Dil Modeli Kullanılarak Zaman Bazlı Duygu Analizi: Whatsapp Yeni Gizlilik Sözleşmesine Yönelik Yorumların Araştırılması, Dokuz Eylül Üniversitesi, Sosyal Bilimler Enstitüsü, Yüksek Lisans Tezi, 70s, İzmir.

Sarıman, G., & Mutaf, E., 2020. COVID-19 Sürecinde Twitter Mesajlarının Duygu Analizi. Euroasia Journal of Mathematics, Engineering, Natural & Medical Sciences, 7(10), 137-148.

Sarpkaya, Y., 2008. Uzaktan Eğitimde Veritabanı Tasarımı ve Örnek Model. Afyon Kocatepe Üniversitesi, Fen Bilimleri Enstitüsü, Yüksek Lisans Tezi, 102s, Afyonkarahisar.

Savaş S., Topaloğlu N., & Yılmaz M., 2012. Veri Madenciliği ve Türkiye'deki Uygulama Örnekleri, İstanbul Ticaret Üniversitesi, Fen Bilimleri Dergisi, 21, 1-23.

Scikit-Learn, 2022. Erişim tarihi: 04.12.2022. https://scikit-learn.org/.

Sharma, D. A., 2022. Database Management Systems/Managıng Database. Lovely Professional University. India.

Seker, S. E., 2016. Duygu Analizi (Sentimental Analysis). YBS Ansiklopedi, 3(3), 21-36.

Sham, N., & Mohamed, A., 2022. Climate Change Sentiment Analysis Using Lexicon, Machine Learning and Hybrid Approaches. Sustainability, 14, 4723.

Srividhya, V., & Anitha, R., 2010. Evaluating Preprocessing Techniques in Text Categorization. International Journal of Computer Science and Application, 47(11), 49-51.

Swaminathan, S. 2018. Erişim tarihi: 05.12.2022 https://towardsdatascience.com/logistic-regressiondetailed-overview 46c4da4303bc.

Şavlukbaş, G., 2019. Rock'n Coke Festivali'nin Twitter Verileri ile Duygu Analizi Örneği. Eskişehir Anadolu Üniversitesi, Sosyal Bilimler Enstitüsü, Yüksek Lisans Tezi

Tan P.T., Steinbach M. & Kumar V., 2006. Introduction to Data Mining. Addison Wesley, Boston.

Tong, R.M., 2001. An Operational System for Detecting and Tracking Opinions in On-Line Discussion, In Proceedings of SIGIR 2001 Workshop on Operational Text Classification, New Orleans, Louisiana, ABD.

Turan, E. S., 2018. Türkiye'nin iklim değişikliğine bağlı kuraklık durumu. Doğal Afetler ve Çevre Dergisi, 4(1), 63-69.

Turney P.D., 2002. Thumbs up or Thumbs down?: Semantic Orientation Applied to Unsupervised Classification of Reviews, In Proceedings of ACL'02, 40th Annual Meeting of the Association for Computational Linguistics, Pennsylvania, ABD, ss. 417-424.

Twitter, 2022. Erişim tarihi: 02.12.2022. https://web.archive.org/web/20210405155212/https://developer.twitter.com/en/docs/twitter-api/getting-started/about-twitter-api.

Ucan, A., Naderalvojoud, B., Sezer, E. A., & Sever, H., 2016. SentiWordNet For New Language: Automatic Translation Approach. In 2016 12th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS) (pp. 308-315). IEEE.

Vasileios, H., Janyce, M.W., 2000. Effects of Adjective Orientation and Gradability on Sentence Subjectivity, In Proceedings of COLING-2000, 18th Interntional Conference on Computational Linguistics, Saarbrücken, Almanya, ss. 299-305.

Waskom, M., 2021. Seaborn: Statistical Data Visualization. Journal of Open Source Software, 6(60), 3021.

Wilhite DA., & Buchanan-Smith M., 2005. Drought as hazard: understanding the natural and social context. In: Wilhite DA (ed) Drought and water crises: science, technology, and management issues. CRC Press, Taylor & Francis Group, Florida, 3–29.