KOCAELİ ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

JEODEZİ VE JEOİNFORMASYON MÜHENDİSLİĞİ
ANABİLİM DALI

YÜKSEK LİSANS TEZİ

SOSYAL MEDYA VERİLERİNDEN DUYGU ANALİZİ YÖNTEMİ İLE SEÇİM SONUÇLARININ MEKANSAL TAHMİNİ: KOCAELİ İLİ ÖRNEĞİ

TUBA BETÜL ÖZKAN

KOCAELİ 2019

KOCAELİ ÜNİVERSİTESİ

FEN BİLİMLERİ ENSTİTÜSÜ

JEODEZİ VE JEOİNFORMASYON MÜHENDİSLİĞİ

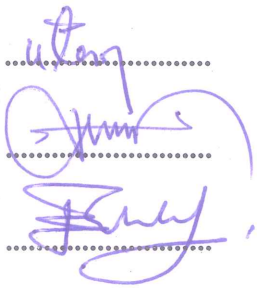
ANABİLİM DALI

YÜKSEK LİSANS TEZİ

SOSYAL MEDYA VERİLERİNDEN DUYGU ANALİZİ YÖNTEMİ İLE SEÇİM SONUÇLARININ MEKANSAL TAHMİNİ: KOCAELİ İLİ ÖRNEĞİ

TUBA BETÜL ÖZKAN

Doç. Dr. Taner ÜSTÜNTAŞ		
Danışman,	Kocaeli Üniversitesi	.....
Doç. Dr. Ozan ARSLAN		
Jüri Üyesi,	Kocaeli Üniversitesi	.....
Doç. Dr. Bahadır ERGÜN		
Jüri Üyesi,	Gebze Teknik Üniversitesi	.....



Tezin Savunulduğu Tarih: 24.12.2019

ÖNSÖZ VE TEŞEKKÜR

Bu tez çalışması, Twitter verileri kullanılarak, Kocaeli ilinin İzmit ilçesindeki 2019 yerel seçimler hakkındaki olumlu ve olumsuz durumların saptanması, halkın belediye başkan adayları ve partiler hakkındaki duygu ve düşüncelerinin belirlenip, konumsal analizinin gerçekleştirilmesi suretiyle, sorunların hızlı ve etkin şekilde teşhis edilmesi amaçlanarak yapılmıştır.

Tez çalışmamda desteğini esirgemeyen, çalışmalarıma yön veren, bana güvenen ve yüreklendiren danışmanım Doç. Dr. Taner Üstüntaş'a teşekkürlerimi sunarım.

Yüksek lisans öğrenimim boyunca, karşılaştığım her zorlukta desteğini esirgemeyen hocam Doç. Dr. Ozan Arslan'a, tez çalışmamda gösterdiği anlayış ve destek için sayın Prof. Dr. Ersoy Arslan'a ve Prof. Dr. Ömer Yıldırım'a, ayrıca teşekkürlerimi sunarım.

Hayatım boyunca bana güç veren, örnek aldığım en büyük destekçilerim, her aşamada sıkıntılarımı ve mutluluklarımı paylaşan sevgili babam Prof. Dr. Ömer Zorba'ya, annem Hatice Zorba'ya, kardeşlerim Ahmet ve Zeynep'e, teşekkür ederim. Ayrıca akademik çalışmalarım sırasında, birçok aşamada beni destekleyen Kocaeli'de biricik ailem olan, İnşaat Mühendisliği bölümünden canım kardeşlerim Hilal Çelik'e ve Nezahat Bilen Demirci'ye teşekkürlerimi sunuyorum.

Ayrıca her aşamada bana desteğini esirgemeyen sevgili eşim Mesut Özkan'a teşekkür ederim.

Aralık – 2019 Tuba Betül ÖZKAN

İÇİNDEKİLER

ÖNSÖZ VE TEŞEKKÜR i
İÇİNDEKİLER ii
ŞEKİLLER DİZİNİ v
TABLOLAR DİZİNİ vi
SİMGELER VE KISALTMALAR DİZİNİ vii
ÖZET viii
ABSTRACT ix
GİRİŞ 1
1. DUYGU ANALİZİNDE YÖNTEM VE YAKLAŞIMLAR 5
1.1. Literatür 5
1.1.1. Genel duygu analizi 12
1.1.2. Duygu analizinin zorlukları 13
1.1.3. Duygu analizinin uygulama alanları 15
1.1.4. Duygu analizi ile ilgili çalışmalar 16
1.1.5. Duygu analiz yöntemleri 16
1.1.5.1. Makine öğrenme yaklaşımı 17
1.1.5.2. Sözlüğe dayalı yaklaşım 17
1.1.5.3. Hibrit (karma) yaklaşımlar 19
1.1.6. Duygu analiz tekniklerinin sınıflandırılması 19
1.1.6.1. Kontrolsüz (denetimsiz) teknikler 19
1.1.6.2. Kontrollü (denetimli) teknikler 20
1.2. Duygu Analizinde Mekansal Analiz ve Coğrafi Bilgi Sistemleri 20
2. DUYGU ANALİZİ VE VERİ MADENCİLİĞİ 23
2.1. Veri Madenciliğine Genel Bakış ve Tarihsel Süreci 23
2.2. Veri Madenciliği Süreci 24
2.2.1. Problemin tanımlanması 25
2.2.2. Verilerin hazırlanması 25
2.2.3. Modelin kurulması ve değerlendirilmesi 26
2.3. Veri Madenciliği Uygulama Alanları ve Veri Madenciliği ile İlgili Yapılan Çalışmalar 27
2.3.1. Veri madenciliği uygulama alanları 27
2.4. Görsel Veri Madenciliği (Visual Data Mining) 28
2.4.1. Verilerin görselleştirilmesi 29
2.4.2. Ara sonuçlarının görselleştirilmesi 29
2.4.3. Veri madenciliği sonuçlarının görselleştirilmesi 29
2.5. Veri Madenciliğinde Kullanılan Temel Yöntemler 29
2.5.1. Tahmini yöntemler 29
2.5.1.1. Makine öğrenmesi 30
2.5.1.2. Doğal dil işleme 39
2.5.1.3. Makine öğrenmesinde doğal dil işleme 40
2.5.2. Tanımlayıcı yöntemler 43
2.5.3. Kümeleme yöntemi 43
2.5.3.1. Hiyerarşik algoritmalar 44

2.5.3.2. Bölümlemeli algoritmalar 45
2.5.4. Veri madenciliğinde kullanılan teknikler 45
2.5.4.1. Sınıflandırma 45
2.5.4.2. Regresyon 45
2.6. Sosyal Medyanın Veri Madenciliğinde Kullanılması 46
2.6.1. Sosyal medya kavramı 47
2.6.2. Sosyal ağ siteleri 48
2.6.3. İçerik topluluğu 49
2.6.4. Bloglar 49
2.6.5. Mikrobloglar 49
2.6.5.1. Twitter 50
2.6.6. Gerçek Zamanlı powertrack API Tarafından Sağlanan Özellikler 52
2.6.7. Standart akış API istek parametreleri 53
2.6.7.1. Delimited 53
2.6.7.2. Stall_Warnings 54
2.6.7.3. Filter_Level 54
2.6.7.4. Language 54
2.6.7.5. Follow 54
2.6.7.6. Track 54
2.6.7.7. Location 55
2.6.7.8. Count 55
2.6.7.9. With (deprecated) 56
2.6.7.10. Replies 56
2.6.7.11. Stringify_Friend_ids 56
2.6.8. Tweetleri bölgelere göre filtreleme 56
2.6.8.1. Tweet konumları ("coğrafi etiketli" tweetler) 56
2.6.8.2. Twitter yerinde JSON 57
2.6.8.3. Tam konum JSON 58
2.6.9. Tweet konum operatörleri 58
2.6.9.1. Place 58
2.6.9.2. Place_contains 58
2.6.9.3. Place_country 58
2.6.9.4. Has:geo 59
2.6.9.5. Point_radius 59
2.6.9.6. Bounding_box 59
3. UYGULAMA 61
3.1. İşlem Adımları 62
3.1.1. Twitter verileri toplama: 63
3.1.2. Verilerin temizlenmesi 67
3.1.2.1. Metnin normalleştirilmesi 68
3.1.3. Duygu analizi (sentiment analysis) 69
3.1.4. Twitter' da duygu analizi için değerlendirme ölçütleri 72
3.1.4.1. Doğruluk (accuracy) 73
3.1.4.2. Kesinlik (precision) 73
3.1.4.3. Duyarlılık (recall) 73
3.1.4.4. F-Skoru 73
3.2. Kocaeli Büyükşehir Belediyesi Seçim Sonuçları 74
4. SONUÇLAR VE ÖNERİLER 75
KAYNAKLAR 84

KİŞİSEL YAYIN VE ESERLER.....91
ÖZGEÇMİŞ.....92

ŞEKİLLER DİZİNİ

Şekil 1.1. Duygu analizi yöntemleri 17
Şekil 1.2. USA obezite haritası (2013) 21
Şekil 1.3. Depresyon bozukluğu hot spot haritası 22
Şekil 2.1. Veri madenciliği sürecinin adımları 24
Şekil 2.2. Doğrusal olarak ayrılabilen veri setleri için hiper-düzlemin belirlenmesi 31
Şekil 2.3. Doğrusal olarak sınıflandırılamayan girdi uzayının bir üst boyuta taşınması 32
Şekil 2.4. Biyolojik sinir hücresi ve yapay sinir ağı karşılaştırması, (a) insan nöronu, (b) yapay nöron veya gizli birlik, (c) biyolojik sinaps, (d) YSA sinaps 37
Şekil 2.5. JSON kod örneği 57
Şekil 2.6. Tam konum JSON kod örneği 58
Şekil 3.1. 2019-03-01 ile 2019-03-31 tarihleri arasında ham tweet dağılımı ve eğilim 61
Şekil 3.2. Duygu analizi adımları 62
Şekil 3.3. Twitter hesabına ulaşımı sağlayan kod parçası 63
Şekil 3.4. Twitter API'lerine ulaşmak için izlenen işlem adımları 64
Şekil 3.5. Anahtar kelimelere göre arama sağlayan kod parçası 66
Şekil 3.6. Anahtar kelimeler vasıtasıyla toplanan tweet örnekleri 66
Şekil 3.7. Twitter'da yayınlanan tweet örnekleri 67
Şekil 3.8. Eğitim yöntemleri WEKA programı ara yüzü 71
Şekil 3.9. Duygu analizinde kullanılan kod parçası 71
Şekil 4.1. AK Partiye ait tweetlerin kişilere göre dağılımı 76
Şekil 4.2. İYİ Partiye ait tweetlerin kişilere göre dağılımı 76
Şekil 4.3. Cross Validation eğitim yöntemine göre Naive Bayes sınıflandırmasında confusion matrisi 78
Şekil 4.4. Cross Validation eğitim yöntemine göre destek vektör makineleri sınıflandırmasında confusion matrisi 79
Şekil 4.5. Percentage Split eğitim yöntemine göre Naive Bayes sınıflandırmasında confusion matrisi 80
Şekil 4.6. Percentage Split eğitim yöntemine göre destek vektör makineleri sınıflandırmasında confusion matrisi 81

TABLOLAR DİZİNİ

Tablo 1.1. Twitter'da yapılan duygu analizi için denetimli makine öğrenmesi 16
Tablo 2.1. Location listesine örnek gösterim 55
Tablo 3.1. Tezde sorgulama amaçlı kullanılan twitter anahtar kelimeleri 66
Tablo 3.2. Düzenli ifade desenleri 68
Tablo 3.3. Gürültülü tweet örnekleri 69
Tablo 3.4. Sınıflandırma ve eğitim yöntemlerinin başarı oranları 71
Tablo 3.5. Confusion matrix (karmaşıklık matrisi) 72
Tablo 3.6. Kocaeli ili 31 Mart 2019 yerel seçim sonuçları 74
Tablo 4.1. Percentage Split eğitim yöntemine göre seçim sonuç tahmini 77
Tablo 4.2. Cross Validation eğitim yöntemine göre seçim sonuç tahmini 77
Tablo 4.3. Cross Validation eğitim yöntemine göre Naive Bayes sınıflandırma başarısı 77
Tablo 4.4. Cross Validation eğitim yöntemine göre Naive Bayes sınıflandırmasında doğruluk oranları 78
Tablo 4.5. Cross Validation eğitim yöntemine göre destek vektör makineleri sınıflandırma başarısı 78
Tablo 4.6. Cross Validation eğitim yöntemine göre destek vektör makineleri sınıflandırmasında doğruluk oranları 79
Tablo 4.7. Percentage Split eğitim yöntemine göre Naive Bayes sınıflandırma başarısı 79
Tablo 4.8. Percentage Split eğitim yöntemine göre Naive Bayes sınıflandırmasında doğruluk oranları 80
Tablo 4.9. Percentage Split eğitim yöntemine göre destek vektör makineleri sınıflandırma başarısı 80
Tablo 4.10. Percentage Split eğitim yöntemine göre destek vektör makineleri sınıflandırmasında doğruluk oranları 81
Tablo 4.11. Manuel yöntemle toplanan tweetler 82

SİMGELER VE KISALTMALAR DİZİNİ

Kısaltmalar

a	: Ak Parti
API	: Application Programing Interfaces-Uygulama Programlama Arayüzleri
API	: Application Programming Interface (Twitter Uygulama Programlama Arayüzü)
BoW	: Bag of Words (Kelime Torbası)
DDI	: Doğal Dil İşleme
DF	: Document Frequency (Belge Sıklığı)
DVM	: Destek Vektör Makineleri
GDA	: Gizli Dirichlet Ataması
HS	: His Simgelerini Kullanarak Etiketleme Yöntemi
i	: İyi Parti
id	: İlgi Derecesi
KNN	: K-En Yakın Komşu
LDA	: Latent Dirichlet Allocation (Gizli Dirichlet Tahsisi)
MaxEnt	: Maksimum Entropi Sınıflandırması
POS	: Part Of Speech (Dilin Parçası)
SMO	: Sosyal Medya Optimizasyonu
SVM	: Support Vector Machine (Karar Destek Makineleri)
TF	: Term Frequency (Terim Sıklığı)
TF-IDF	: Term Frequency–Inverse Document Frequency (Terim Sıklığı - Ters Belge Sıklığı)
TM	: Topik Bilgisine Dayalı Etiketleme Yöntemi
TV	: Televizyon
üd	: Üyelik Derecesi

SOSYAL MEDYA VERİLERİNDEN DUYGU ANALİZİ YÖNTEMİ İLE SEÇİM SONUÇLARININ MEKANSAL TAHMİNİ: KOCAELİ İLİ ÖRNEĞİ

ÖZET

Son dönemlerde oldukça yaygınlaşan sosyal medya ortamları gerek Türkiye'de gerekse Dünya'da sıklıkla kullanılmaktadır. Bu bağlamda Twitter, kullanıcılarla doğrudan iletişime girme imkânı verdiği için bu mecra, seçim dönemlerinde de adayların ve partilerin kurtarıcısı haline gelmiştir. Bunun altında yatan sebep, Twitter'ın kullanıcılara herhangi bir konu üzerinde fikirlerini açıkça beyan etme olanağı sunmasıdır. Ayrıca seçim dönemlerinde halkın da dilediğini yazma imkânı bulduğu bu ortam sayesinde, parti ve adayları halkın isteklerini görebilmekte, bu istek ve şikâyetlere göre önlemler alabilmektedir. Bu tez kapsamında 31 Mart 2019'da yapılan yerel seçimlerde Twitter'da yayınlanan tweetlerden yararlanılarak, Kocaeli ili seçim sonuçlarının tahmini yapılmıştır. Tahminlerin yapılması için; 2019-03-01 ile 2019-03-31 tarihleri arasındaki ham tweetler, Python dilinde yazılan bir program ile json formatında toplanıp, günlük olarak gruplandırılmıştır. Seçim sonucunu belirlemeye yönelik anahtar kelimeler belirlenmiştir. Bu ham tweetler gruplandırıldığında AKP 878, İyi Parti 283, Saadet Partisi 208 adet, tarafsız 600 adet, toplamda 1969 adet tweet toplanmıştır. Seçim sonucunu belirlemeye yönelik anahtar kelimeler belirlenmiştir. Veriler üzerinde Duygu Analizi gerçekleştirilmiş, tweetler pozitif ve negatif olarak ayrılmıştır. Tweetlerin sınıflandırılmasında Naive Bayes ve Destek Vektör Makinesi yöntemleri kullanılarak, başarıları kıyaslanmıştır. Ayrıca tweetler manuel olarak da sınıflandırılarak makine öğrenme yöntemi ile manuel yöntem arasında kıyaslama yapılmıştır. Farklı sınıflandırma eğitim yöntemleri de denenmiş, başarıları kıyaslanmıştır. Çalışma sonucunda, makine öğrenme yöntem sonuçları pek fark etmemiş, fakat eğitim yöntemlerinde, Percentage-Split yöntemi daha başarılı sonuç vermiştir. Ayrıca manuel yöntem ile seçim sonuç tahmini, makine öğrenme yöntemlerine göre daha başarılı sonuç verdiği de saptanmıştır.

Anahtar Kelimeler: Duygu Analizi, Naive Bayes, Seçim Tahmini, Twitter.

THE SPATIAL ESTIMATION OF SELECTION RESULTS BY EMOTION ANALYSIS METHOD FROM SOCIAL MEDIA DATA: A CASE OF KOCAELI PROVINCE

ABSTRACT

Recently,social media is used frequently in turkey and the world. In this context, twitter has become the liberator of candidates and political parties during the election periods and it provides the opportunity to communicate directly with the users. The reason for this is that twitter allows users to express their opinions on any issue. In addition, thanks to this environment in which the people have the opportunity to write whatever they wish during the election periods, political parties and its candidates can see the wishes of the people and can take measures according to these requests and complaints. Within the scope of this thesis, the election results of kocaeli province were estimated by using tweets published on twitter in local elections held on 31 march 2019. To make predictions; the raw tweets between 2019-03-01 and 2019-03-31 were collected in json format with a program written in python and grouped daily. Keywords were determined to determine the outcome of the selection. When these raw tweets were grouped, for akp 878,for ıyi parti 283,for saadet partisi 208, 600 neutral, total 1969 tweets were collected. Then, executed " sentiment analysis " on classified tweets. The success of the tweets was compared by using Naive Bayes and Support Vector Machine methods. In addition, tweets were also classified manually and compared between machine learning method and manual method. Different classification training methods were also tried and their successes were compared. As a result of the study, machine learning method results did not notice much, but in education methods, Percentage-Split method gave more successful results. In addition, it was found out that selection results estimation by manual method was more successful than machine learning methods.

Keywords: Sentiment Analysis, Naive Bayes, Election Predictions, Twitter.

GİRİŞ

Bilgiler, teknolojinin gelişmediği zamanlarda kâğıt ortamlar kullanılarak muhafaza edilmiş ve saklanmıştır. Teknolojinin gelişmesiyle birlikte bilgi aktarımı, korunup saklanması da daha kolay hale gelmiştir. Gelişen internet teknolojisiyle birlikte insan hayatı da büyük ölçüde kolaylaşmıştır. İnternet sayesinde aradaki mesafeleri gözetmeksizin insanlar birbirleriyle rahatça görüşüp, iletişim kurabilmektedir. Sosyal medya aracı olan Facebook, Twitter, İnstagram gibi platformlarda insanlar mesafeye bakmaksızın istediği herkesle görüşebilmekte, fikirlerini rahatça beyan edebilmektedirler. Duygu ve düşüncelerini sadece yazı ile değil, ses veya resimlerle de paylaşabilmektedirler. Duygu ve düşünceler yazı ile ifade edilirken, tam olarak hissettirilmek istenilen duygular bazen karşı tarafa yansıtılamayabilmektedir. Bunun için ise ekstra açıklama ya da emoji adı verilen duygu belirten ifade ve simgelerin kullanımına ihtiyaç duyulmaktadır.

Bloglar, istenilen alanda yorum, paylaşım yapılıp, fikirlerin yazılabildiği çevrim içi günlüklerdir. Mikrobloglar ise, blogların uzunluk kısıtlaması getirilmiş halidir. Günümüzde Twitter, mikroblog sitesinden çok daha fazlasını ifade etmektedir. Bulunduğu konuma bakılmaksızın, dünyanın neresinde olursa olsun, paylaşım yapan insanlardan haberdar olmayı sağlamaktadır. Dünya üzerindeki gündemin takibi, politik, magazinsel veya bilimsel gelişmelerin takibi için sıkça tercih edilen bir platformdur. Twitter'da kullanıcılar herhangi bir konu hakkında fikirlerini belirtebilmekte, başkalarının fikirlerini öğrenebilmektedirler. Kullanıcılar istedikleri konuma göre, gündemi politikacıların ağzından öğrenebilmekte, sevinçlerini, üzüntülerini, destek veya kınama gibi duygu veya hislerini Twitter'da özgürce paylaşabilmektedirler. Twitter, duygu ve hislerin paylaşılmasına izin veren bir platform olması sebebiyle, anlık olarak dünyanın her yerinde farklı kullanıcılar tarafından önemli ölçüde kullanılmakta ve tweetler paylaşılmaktadır. Bu özelliği, çeşitliliğin fazla olmasıyla beraber duygu analizi için çok ideal bir platform olmaktadır. Ayrıca Twitter'ın konum bilgisini de içermesi sebebiyle bu çalışmada

Twitter sosyal sitesi üzerinden paylaşılan mesajların, duygu analizi için kullanılması uygun görülmüştür.

Reklam, artık insan hayatında önemli bir yer kaplamaktadır. Siyasetten, ticarete, sağlığa ya da eğitime hemen hemen her alanda artık reklam kaçınılmaz bir parça olmuştur. Gazete, televizyon, radyo ya da internet üzerinden günümüzde her amaçla reklam yapılmaktadır. Reklamın bu denli yaygın ve etkili olması, siyasi ortamlarda da kullanılmasına uygun bir ortam hazırlamıştır.

Özellikle son zamanlarda kitle iletişim araçlarının hızla gelişip, hayatımızda önemli bir yere sahip olmasıyla birlikte, yüz yüze ilişkiler ve grupların birbirleriyle olan iletişimi zayıflamıştır. İletişim daha çok internet kaynağı üzerinden olmaya başlamıştır. Böylece internet ortamı aracılığıyla insanların birçok farklı konu hakkında fikir alışverişine imkân sağlanmıştır. Bu durum hayatın her alanına yansıdığı gibi siyasete de yansımıştır. Siyasi partiler, internet aracılığı ile insanlara daha rahat ulaşabilmekte ve propagandalarını daha etkili şekilde yapabilmektedirler. Siyasal reklamcılık olarak tabir edilen bu durum, seçmeni bilgilendiren önemli bir araç olarak görülmüş ve uygulanmıştır. Politikacılar bu sayede interneti ve özellikle sosyal medyayı amaçları doğrultusunda kullanarak kendilerine taraftar toplamaktadırlar. Böylece sosyal medya siyasi partiler için önemli bir propaganda aracı haline gelmiştir. Demokrasinin hâkim olduğu toplumlarda, siyasi partiler ya da siyasi parti adayları, seçmenlerden desteklerini ve güvenlerini almak isterler. Bundan dolayıdır ki seçmenlere ikna edici mesaj iletmek zorundadırlar. Buradaki amaç, adayın daha geniş bir seçmen kesimine tanıtılıp, kendi partileri ve adayları ile diğer parti ve adaylar arasındaki farkın gösterilerek, seçimi kazanma arzularıdır. Bundan dolayı ülkedeki yerel ve genel seçimlerde sosyal medyanın kullanılması, büyük kitlelere çok hızlı ve kolay ulaşılabilmesi, etkinliği ve reklam maliyetinin çok ucuz olması gibi nedenlerle artık çok yaygın hale gelmiştir.

Sosyal medya, seçim kampanyalarında siyasi partilerin seçmenlere nasıl hizmet sağlayacağı, seçmenler tarafından nasıl anlaşılacağı gibi konular sayesinde popüler olarak kullanılmaya başlanılmıştır. Sosyal medya insanların çoğuna ulaşma konusunda başarılı olduğu için, bu durum siyasal alana da kaymıştır. Bu mecra sadece seçmen için değil, politikacılara ulaşmak için de kullanılmaktadır. Bundan dolayı politikacılar

sosyal medyayı kendilerinin yararına da kullanmaya başlamışlardır. Seçmeni etkilemek, diğer partilerin adaylarından haberdar olup, ona göre önlem almak için de kullanımı yaygınlaşmıştır. Sosyal medya sayesinde artık çoğu ülkede, önceden hangi partinin ya da parti adayının seçimi kazanacağının tahmini rahatlıkla yapılmaktadır. Halkın sosyal medya aracılığı ile paylaştığı fikir ve görüşleri bu tahmini kolaylaştırmaktadır. Sosyal medya üzerinden yapılan anketlerin sonuçları, seçimlerden önce seçim sonuçları hakkında doğru tahminlerde bulunulmasını sağlamaktadır. Partiler sosyal medyayı; seçmenleri, adayları hakkında değişik stratejik slogan ve vaatler ile bilgilendirerek onları kendi adaylarına oy vermeye yöneltmek için kullanabilmektedirler. Hatta siyasi partiler rekabet açısından sosyal medyayı kullanmak zorundadırlar. Bunun sebebi, seçimi kazanmak için parti ve adaylarını seçmene tanıtmak, kendi partilerini diğer partilerden ayıran özellikleri göstermek veya adaylar arasındaki farkları belirtmek zorunda olmalarıdır. Bunları belirtirken de çok geniş bir kesime hitap eden interneti ve özellikle sosyal medyayı kullanmak zorunda kalırlar. Bu özellikler de internet ve sosyal medyayı olmazsa olmaz yapan bir iletişim aracı kılmaktadır.

Hızla gelişen sosyal medya araçları her şeyin öğrenilmesini mümkün kılmakta, olaylara yön vermekte, düşünceleri şekillendirip, büyük kitlelere sunmaktadır. Bu özelliği ile yaşantımızın olmazsa olmazları arasında yerini almayı başarmıştır. Sosyal medya sadece parti ve adaylar tarafından değil, aynı zamanda halk tarafından da aktif şekilde kullanılıp, faydalanılmaktadır. Halk fikirlerini özgür bir şekilde, sosyal medya aracılığı ile siyasilere ulaştırmaktadır. İnsanların fikir ve düşünceleri doğrultusunda seçime katılan partiler de ona göre propaganda yapmakta, seçim vaatleri düzenlemektedirler. Sosyal medya sayesinde insanlar olumlu, olumsuz tüm durumlardan bahsederek, parti ve adayları da bu durumlardan haberdar etmektedirler. Bu yönleri ile kullanıldığı zaman, sosyal medya gerçekten çok faydalı bir mecra haline gelmektedir. İnsanların yaptıkları yorumlarla, ileri sürdükleri fikirlerle seçim sonuçlarından önce fikir sahibi olunabilmektedir. Sosyal medya araçlarından biri olan Twitter'da konum bilgisinin de olmasıyla birlikte, insanların yapmış oldukları yorumlar daha anlamlı hale gelmektedir. Yapılan yorumların hangi bölgeden olduğuna bakılarak, ülkedeki seçim sonuçlarının bölgelere göre tahmin edilmesi kolaylaşmaktadır. Bölgelerdeki istek ve şikâyetlere göre, partiler seçim çalışmalarını

o yönde geliştirme imkânı bulmaktadırlar. Böylece ülke çapında aynı yönlü kampanya yerine, bölgesel çözümlere gidilerek her bölgeye farklı kampanyalar düzenlenmek suretiyle bölgesel oy oranları arttırılabilmektedir.

Bu avantajlardan yola çıkarak çalışmamızda, 31 Mart 2019 yerel seçiminde yayınlanan tweetler kullanılarak, Kocaeli ilini kapsayan bir çalışma yapılmıştır. Çalışmamızda ilk olarak, Kocaeli ilinin seçime girecek olan parti adayları belirlenmiştir. Her bir aday için, Twitter'da yayınlanmış tweetler sorgulanıp, bir havuzda toplanarak veri bulutu oluşturulmuştur. Veri bulutunda bulunan veriler üzerinde duygu analizi yapılmıştır. Duygu analizi yapılırken, sınıflandırıcı olarak Naive Bayes ve Destek Vektör Makinesi yöntemleri kullanılarak, sınıflandırma başarı oranları karşılaştırılmıştır. Ayrıca sınıflandırma yapılırken, veriler çeşitli yöntemlerle eğitilmektedirler. Bu eğitim yöntemlerinden olan Cross-Validation ve Percentage-Split yöntemleri de ayrı ayrı kullanılarak, sınıflandırma başarısı üzerindeki etkileri karşılaştırılmıştır. Bu yöntemlerle beraber veriler, pozitif ve negatif olarak sınıflara ayrıştırılmıştır. Pozitif tweetlerin tüm tweetlere oranı hesaplanarak seçim sonuç tahmininde bulunulmuştur. Bunun dışında ayrıca toplanan tweetler manuel olarak pozitif, negatif ve tarafsız sınıflara ayrıştırılarak, pozitif tweetlerin tüm tweetlere oranı hesaplanmış ve seçim sonuç tahmininde bulunulmuştur. Böylece makine öğrenmesi yöntemleri ile manuel yöntemin seçim sonuç tahmin sonuçlarının karşılaştırılmasına olanak sağlanmıştır.

Çalışmanın ilk bölümünde, duygu analizinde kullanılan yöntemlerden, uygulama alanlarından genel olarak bahsedilmiştir. İkinci bölümde, duygu analizinin veri madenciliği ile arasındaki ilişki incelenmiş, veri madenciliği süreci hakkında detaylı bilgiler verilmiştir. Ayrıca sosyal medya kavramı ve bu kavramın veri madenciliğinde kullanılması ile alakalı bilgiler de bu bölümde detaylandırılmıştır. Üçüncü bölümde, uygulama yapılmış olup, kullanılan metotlar üzerinde durulmuştur. Uygulama hakkından geniş bilgiler sunulmuştur. Son bölümde çalışmanın sonuçları gösterilmiş ve daha sonraki çalışmalara katkı sağlayacak önerilerde bulunulmuştur.

1. DUYGU ANALİZİNDE YÖNTEM VE YAKLAŞIMLAR

1.1. Literatür

Çoban ve diğ., (2015) yaptıkları çalışmada, Türkçe Twitter mesajlarından oluşturdukları veri seti üzerinde metin sınıflandırma yöntemlerini uygulamış, olumlu veya olumsuz yönlerinin var olup olmadığını incelemişlerdir. Deneysel sonuçlara, SVM, Naive Bayes, Multinom Naive Bayes ve KNN algoritmalarıyla ulaşmışlardır. Öznitelikleri, Kelime Torbası (Bag of Words, BoW) ve N-Gram model olarak iki farklı yöntemle elde etmiş ve bunun sınıflandırma sonuçlarına etkilerini incelemişlerdir. Bu çalışmayı literatürdeki diğer çalışmalardan ayıran özellik ise, N-Gram modelinde özniteliklerin kelime seviyesinde olmayıp, karakter seviyesinde olmasıdır. Tüm sınıflandırıcılar için, N-Gram modelin BoW modelden daha iyi sonuç verdiğini gözlemlemişlerdir. Makine öğrenmesi için ise her iki model üzerinde de en iyi sonucu Multinom Naive Bayes yönteminin verdiğini tespit etmişlerdir.

Türkmenoğlu (2015) yapmış olduğu yüksek lisans tez çalışmasında, Twitter'dan alınan yorum ve Twitter mesajlarına oranla daha kurallı yazılmış olan film yorumlarının oluşturduğu veri seti kullanmıştır. Daha sonra, sözlük tabanlı duygusal analiz ve makine öğrenmesi metotlarına yeni özellikler ekleyip, farklı iki veri kümesi üzerinde değerlendirme yapmıştır. Duygusal analiz yöntemini, hem kısa hem de uzun metinler üzerine uygulayarak başarısını ölçmüştür. Böylece metotlar arası güçlü ve zayıf yönler aranmıştır. Film yorumlarının, Twitter verilerine oranla daha düzenli ve hedefinin belli olmasından ötürü, her iki yöntemde de daha iyi sonuç verdiğini saptamıştır. Makine öğrenmesi KDM sınıflandırıcısıyla birlikte kullanıldığında yüksek başarı elde edilmiştir. Sözlük tabanlı duygu analizi, denetimsiz olmasına rağmen, bu çalışmada iyi bir sonuç vermiştir.

Adak Kaplan (2016) yapmış olduğu yüksek lisans tez çalışmasında, içeriği Türkçe olan tweetlerin duygu analiziyle değerlendirilmesini asıl amaç olarak belirlemiştir. Bu bağlamda, sosyal medya araçlarından biri olan Twitter'ın kullanıcılarının yayınladığı tweetleri analiz etmiştir. Analiz edilen tweetleri; mutluluk, kızgınlık, üzüntü ve

şaşkınlık olarak dört grup etrafında toplamıştır. Sınıflandırmaların doğruluğu açısından tweetleri, Zemberek Kütüphanesi'ne sokarak yazım hatalarını gidermiştir. Sınıflandırılmaya hazır olan verileri, Karar Ağacı ve Fuzzy Ruler Learner yöntemlerine tabii tutarak analiz etmiş ve sonuçları değerlendirmiştir. Değerlendirmeler sonucunda, Karar Ağacı yönteminin doğruluğu %85,372 çıkarken, Fuzzy Ruler Learner yönteminde doğruluk %83,608 çıkmıştır. Araştırıcı ayrıca, çalışma yapılmadan önce verilerin ön hazırlığı sırasında özel isimlerin de temizlenmesi gerektiği sonucuna ulaşmıştır. Araştırıcı, duygu belirtmemesine rağmen, özel isimlerin araştırma yapılan verilere dâhil olması nedeniyle sonuçlara etkisinin olduğunu gözlemlemiştir.

Onan ve Korukoğlu (2015) yaptıkları bir çalışmada, görüş madenciliği alanında, temel makine öğrenmesine yönelik yapılan çalışmaları incelemiş ve incelenen yöntemlerin güçlü ve zayıf yönlerini ele almıştır. Bu yöntemler; öğreticili, yarı-öğreticili ve öğreticisiz yöntemler olarak üç temel grup altında incelenmektedir. Eğitim verisi ile test verisi arasındaki dağılım değiştikçe, öğreticili öğrenme yöntemlerinin sonuçlarına etkisinin kaybolduğu görülmüştür. Bundan dolayı, görüş madenciliğindeki çalışmalarda daha çok yarı-öğreticili ya da öğreticisiz yöntemlerin geliştirilmesine odaklanılmaktadır. Buna paralel olarak görüş madenciliği çalışmalarında kullanılacak olan eğitim seti örnekleri etkin şekilde seçilmelidir. Günümüzde, eğitim seti seçme yöntemlerinin geliştirilmesi başlı başına önemli çalışma konusu haline gelmiştir.

Uçan (2014) yapmış olduğu yüksek lisans tez çalışmasında, sözlük kullanımıyla duygu analizi çalışması yapmıştır. Yazar, son onbeş yıl boyunca yapılmış çalışmaları incelediğinde, duygu analizi çalışmalarında makine öğrenmesi yöntemlerine sıkça rastlarken, duygu sözlüğü kullanılan çalışmaya rastlamamıştır. Bundan ötürü bu eksikliği gidermek adına doğrudan sözlük kullanarak duygu analizi yapmıştır. Türkçe diline ait duygu sözlüğü bulunmadığı için, duygu ifadelerinin evrenselliğini göz önüne alarak, İngilizce olan duygu sözlüğünü Türkçe'ye otomatik çevirisini yapmak suretiyle, Türkçe Duygu Sözlüğü oluşturmuştur. Hem çeviri yaparak hem de farklı çeviri algoritmaları ile oluşturulan sözlüklerin doğruluğunun karşılaştırılmasını deneyler vasıtasıyla gerçekleştirmiştir. Deneyler için iki farklı veri sınıfı oluşturmuş ve bu kümeler ile oluşan kelimeleri deneyerek başarılı olduğunu görmüştür. Elde ettiği sonuçların doğruluk derecesini kontrol etmek adına aynı deneysel adımları makine

öğrenmesi yöntemi ile de tekrar etmiştir. Bu iki sonucu karşılaştırdığı zaman yöntemin başarılı olduğunu görmüştür.

Bilgin ve Çamurcu (2008) yapmış oldukları araştırmada, veri madenciliğinin araştırma alanlarından olan, çok boyutlu veri tabanlarını ve bu veri tabanlarının görselleştirilmesinde kullanılan teknikleri incelemişlerdir. Verilerin görselleştirmesi, veriler arasındaki ilişkinin daha iyi anlaşılmasını sağlamaktadır. Bu yüzden bu çalışmada görselleştirmede kullanılan teknikler açıklanarak, okuyucuya bu konu hakkında bilgi sunmuşlardır.

Jiangl ve diğ., (2011) yapmış oldukları çalışmada, hedefe bağlı Twitter duyarlılık sınıflandırılması yapmışlardır. Bu amaçla bir sorgu verildiği zaman, tweetlerin duygularını olumlu, olumsuz ve tarafsız olarak sınıflandırmışlardır. Buradaki sorgu, duyguların hedefi olarak işlev görmüştür. Çalışmalarında farklı olarak, grafik tabanlı optimizasyon yöntemi kullanmışlar, bu sayede ilgili tweetleri dikkate alarak performanslarını önemli ölçüde arttırmışlardır.

Albayrak (2017) yaptığı çalışmasında; veri tabanları, veri ambarlarında veri işlenmesi için kullanılan yöntemleri, veri tabanlarında bilgi keşfi için kullanılan yöntemleri konu almıştır. Hangi tür veriler üzerinde hangi sınıflandırma ve kümeleme algoritmalarının kullanılacağı sorusuna cevap aramıştır. Kullanım alanlarına değinen yazar, veri madenciliğinde kullanılan programlardan da bahsetmiştir.

Elabiad (2013) yapmış olduğu yüksek lisans tezinde, kayıtlı veriler yerine, web tabanlı anket yöntemiyle cinsiyet, yaş, medeni durum, eğitim durumu, gelir aralığı, GSM operatör türü ve abonelik süresi, GSM numara taşıma oranı, tarife türü gibi anket soru başlıkları belirlemiştir. Anket cevapları kullanılarak, toplanan veriler veri madenciliği yöntemiyle analiz edilip, gerekli bilgiye ulaşılmasını hedeflemiştir. Ulaşılan bilgiler ışığında, GSM şirketlerinin müşteri kaybetme potansiyeli olan müşteri gruplarını ve alışkanlıklarını veri madenciliği yöntemleri ile tespit etmiştir. Web tabanlı anket sistemi ile elde edilen verileri WEKA programıyla işleyip analiz etmiştir. Analiz sırasında karar ağacı algoritması ve birliktelik kuralları çıkarımı için apriori algoritması kullanmıştır. Web tabanlı anket yöntemi sayesinde, birden fazla öznitelik verileri toplanabilmektedir. Bu öznitelikler arasında amaca uygun öznitelik seçimi yapılarak birbirleri arasındaki ilişki, güven bulunabilmektedir ve bu algoritmalar

sayesinde geleceğe yönelik tahminlerde bulunulabilmektedir. Bu çalışmanın sonucunda, ankete katılan müşterilerin özelliklerine göre, GSM şirketleri müşterilere özel kampanya yaparak memnuniyetini sürdürebilir. Yazar bu çalışmayı, sadece GSM şirketleri için değil diğer ticari alanlarda, eğitim sektöründe vs. kullanılabileceğini öne sürmüştür.

Şengür (2013), yapmış olduğu yüksek lisans tezinde; Fırat Üniversitesi, Eğitim Fakültesi, Bilgisayar ve Öğretim Teknolojileri Eğitimi Bölümü (BÖTE) öğrencilerinin mezuniyet notlarının tahmin edilmesini gerçekleştirmiştir. Not tahmininde Yapay Sinir Ağları (YSA) ve Karar ağaçları yöntemlerini kullanmıştır. Yazar tezinde, mezun durumunda olamayan öğrencilerin uyarılması veya belli not ortalaması altında kalan öğrencilerin saptanarak çalışmalarına yoğunlaşmasının sağlanmasını hedeflemiştir. İki farklı senaryo öngörmüştür. İlk senaryosunda, öğrencilerin ilk iki yıldaki derslerinin yılsonu notlarını kullanarak tahmini, ikinci senaryoda ise ilk üç yıldaki sınıf notları ile mezuniyet notlarının tahminini gerçekleştirmiştir. Modellerin doğruluğunu bulmak için geçerlilik analizi yapılmış ve hata oranlarını bulmuştur. Başarım değerlendirmesi için, korelasyon kat sayısı ve ortalama karesel hata fonksiyonlarının yanı sıra ortalama mutlak hata yöntemlerini kullanmıştır. Yapılan çalışmanın sonucunda YSA'nın Karar ağaçlarına oranla daha etkili tahminde bulunduğunu ortaya koymuştur. Ayrıca ikinci senaryonun da birinci senaryoya göre daha etkili tahminde bulunduğunu görmüştür.

Çoban ve Tümüklü Özyer (2016), yaptıkları çalışmada Türkçe Twitter mesajlarında LDA (Latent Dirichlet Allocation) algoritmasını kullanarak duygu sınıflandırması yapmışlardır. Çalışmalarındaki amaç, tweetlerin içeriğini otomatik olarak belirlemek ve içerdikleri duyguları sınıflandırmak olmuştur. Bu tweetlerin duygu analizi için bir sistem önermişlerdir. Ayrıca duygu sınıflandırmasını topik modelleme ile yapmış ve sonuçlara olan etkilerini incelemişlerdir. Topik model kurulumu aşamasında, LDA (Latent Dirichlet Allocation) algoritmasını kullanmışlardır. Topik modelin diğer modellere oranla daha başarılı olduğu sonucuna varmışlardır. Yapmış oldukları çalışmada, tweetleri topik bilgisine dayalı duygu sınıflandırmasından geçirerek etiketlemişlerdir. Etiketleme yöntemi olarak ise his simgelerini kullanarak etiketleme yönteminin yerine (HS), GDA (Gizli Dirichlet Ataması) ile oluşturulan topik bilgisine dayalı etiketleme yöntemini (TM) önermişler ve sonuçlara etkisini incelemişlerdir. HS yönteminde; negatif kategorisinde yer alan mesajları da çoğu zaman pozitif alması,

sonuçların olumsuz etkilenmesine yol açtığını saptamışlardır. TM yönteminde ise tweetlerin yüksek oranda doğru etiketlenmesiyle beraber, duygu analizi başarısının %26 oranında arttırdığını belirlemişlerdir.

Akgül ve diğ., (2015) yaptıkları çalışmada, Twitter'da belli bir konu üzerine yazılmış olan tweetlerin duygusal yönden olumlu, olumsuz ve tarafsız olarak sınıflara otomatik şekilde ayırmayı amaçlamışlardır. Ayrıca kullanıcılara hizmet versin diye, kullanıcıların sektörlerine ait kelime kalıplarını ve sözlüklerini oluşturarak, bu çalışmanın farklı alanlarda kullanımının arttırılmasını da hedeflemişlerdir. Veri seti olarak, Twitter'dan Türkçe olarak belirli kelime gruplarına göre yapılan aramalar sonucunda elde ettikleri, dört aylık süreci kapsayan ve farklı adetlerde üç ayrı veri seti oluşturmuş ve kullanmışlardır. Çalışmalarında hem sözlük hem de n-gram modeli olmak üzere iki farklı yöntem kullanmışlardır. Sonuçların başarısını F- Ölçüm metriğini kullanarak hesaplamışlardır. Sözlük yönteminin daha başarılı olduğu sonucuna varmışlardır. Eğitim verisi olarak küçük veri setleri kullanıldığı zaman, iki yöntem arasındaki farkın az çıktığını görmüşlerdir. Fakat veri seti sayısı arttıkça iki yöntem arasındaki farkın açıldığını saptamışlardır.

Koçak ve diğ., (2016) yaptıkları çalışmada, Twitter kullanıcılarının havayolu ulaşımı ile ilgili attıkları tweetler toplayarak, duygu analizi çalışması yapmışlardır. Tweetleri Twitter'ın sunmuş olduğu API (Application Programing Interfaces-Uygulama Programlama Arayüzleri) hizmeti aracılığıyla, java tabanlı program kullanarak belli tarihler arasında almışlardır. Topladıkları tüm tweetleri olumlu, olumsuz ve tarafsız sınıflarına ayırmışlardır. Bu sınıfları etiket bulutunda toplamış ve sonuçlarını Makine Öğrenmesi Yöntemi ve SMO sınıflandırmasında standart ve normalize Kernel Polinomları kullanarak analizlerini gerçekleştirmişlerdir. SMO algoritmasına göre, standart Kernel Polinomu daha iyi performans gösterdiği sonucuna ulaşmışlardır. Fakat genel olarak sınıflandırma başarısının düşük olduğunu gözlemlemişlerdir. Bunun nedeni olarak ise veri setinin dar olması ve verilerin dengesiz dağılışını göstermişlerdir.

Onan (2017) yaptığı çalışmada, Türkçe Twitter verilerini farklı algoritmalar kullanarak sınıflandırmış ve bu algoritmaları karşılaştırmıştır. Sınıflandırma algoritması olarak; Naive Bayes, Destek Vektör Makineleri ve Lojistik Regresyon

kullanmıştır. Ayrıca metni temsil etmede farklı öznitelikler kullanarak, farklı özniteliklerden çıkan farklı öznitelik setlerini değerlendirmiştir. Twitter API kullanarak bir ay süren mesaj toplama sürecinden sonra mesajları pozitif ve negatif olmak üzere sınıflara ayırmıştır. Sınıflara ayırırken sadece his simgelerini göz önüne alarak yapmıştır. Öznitelik setlerini oluştururken N-gram modelini kullanmıştır. Çalışma sonucunda Naive Bayes algoritmasının en iyi sonucu verdiğini görmüştür. N-gram temsil modellerinde ise en iyi sonucu 1-gram temsil yönteminin verdiğini saptamıştır.

Pak ve Paroubek (2010), duygu analizi için en popüler platform olan Twitter kullanarak verilerin otomatik toplanabilmesi için, duyarlılık sınıflandırıcısının eğitilmesinde de kullanılabilecek bir yöntem geliştirmişlerdir. Bu sınıflandırıcı verilerin olumlu, olumsuz ve tarafsız duyguları belirleyebildiği ileri sürülmüştür. Geliştirdikleri sınıflandırıcı, N-gram ve POS etiketlerini özellik olarak kullanan Multinomlu Naive Bayes sınıflandırıcısına dayanmaktadır. Araştırmalarında İngilizce çalışmalarına rağmen kullandıkları teknikler diğer dillerde de kullanılacağını söylemektedirler. Çalışmalarında ilk olarak, verilerdeki kelime frekans dağılımı yapmışlardır. Daha sonra geliştirdikleri Multinomlu Naive Bayes sınıflandırıcısı kullanılarak verileri sınıflandırmışlardır. N- gramların performanslarını karşılaştırmış, en iyi performansı bigramların verdiği sonucuna varmışlardır. Son olarak veri kümesinin büyüklüğünün performansa olan etkisini de araştırmışlar ve bunun için F-ölçütü kullanmışlardır. Örneklem büyüklüğü ne kadar fazla olursa, performansın da o kadar fazla olacağını alışmalarında kanıtlamışlardır.

Gokulakrishnan ve diğ., (2012) yaptıkları çalışmada, Twitter'da halka açık eğitim veri seti kullanmışlardır. Topladıkları veriler el yordamıyla toplanmış olup, duygusal içeriklerine dayanarak, pozitif-negatif tweet sayısı dengesizlik içermiştir. Bu dengesizliği azaltmak için farklı yöntemler denemişlerdir. Çalışmalarındaki amaç, etkili şekilde kullanılabilecek popüler sınıflandırma tekniklerinden en uygun tekniği belirlemek olmuştur. Duygu analizinin uygulanacağı tweetler sınırlı halka açık veri kümesinden oluştuğu için, Neutral Polar-Alakasız eğitim veri seti Twitter Uygulama Programlama Arayüzü (API) sorgulayarak elle toplama işlemi gerçekleştirmişlerdir. Bu sorguları veri çeşitliliği sağlaması açısından keyfi olarak farklı alanlardan seçmişlerdir. Toplama sürecinde dil, yer gibi kısıtlamalarda bulunmamışlardır.

Dolayısıyla veri setinde her türlü dilden tweet bulunmaktadır. Daha sonra her bir tweeti kutupsal, tarafsız ve alakasız olarak etiketlemişlerdir. İkinci aşama olarak ise, kutup altında sınıflandırılan verileri pozitif ve negatif olmak üzere ikinci bir sınıflandırıcıya tabi tutmak olmuştur. Alakasız ve tarafsız veriler erken aşamada elimine edildiği için, pozitif negatif sınıflandırma, diğer geleneksel yöntemlerdeki sınıflandırmaların doğruluğuna oranla daha yüksek başarı sağladığı sonucuna varmışlardır. Sınıflandırıcı çeşitlerinden en iyi sonucu Bayesian sınıflandırıcı çeşitlerinden olan SMO verirken, en başarısız ağaç tabanlı J48 sınıflandırıcısı olduğu sonucuna varmışlardır.

Go ve diğ., (2009) yaptıkları çalışmada, Twitter mesajlarını pozitif ve negatif sınıflara ayırmışlar ve makine öğrenme algoritmalarının sonuçlarını tartışmışlardır. Ayrıca yüksel doğruluk elde etmek için gereken ön işlem adımlarını da bu çalışmada anlatmaktadırlar. Bu çalışmalarındaki asıl amaçları ise, uzaktan denetimli öğrenme için, duygu belirten tweetleri kullanma fikridir. Sadece 6 Nisan 2009- 25 Haziran 2009 tarihleri arasında, İngilizce tweetleri veri seti olarak kullanmışlardır. Duygu sembolü olarak kullanılan ifadeleri doğruluğu arttırmak için, veri setinden çıkarmışlardır. Daha sonra hem olumlu hem de olumsuz cümle bulunduran tweetleri de veri setinden çıkarmışlardır. Ayrıca retweetleri de veri setinden ayırmışlardır. Ön işlemden geçen tweetleri manuel olarak 177 negatif, 182 de pozitif olmak üzere işaretlemişlerdir. Test verilerini toplarlarken; belirlenen konu başlıklarına göre sorgu yapmışlardır. Daha sonra sorgudan çıkan sonuca göre ifadelerin varlığından bağımsız olarak pozitif ya da negatif olarak işaretlemişlerdir. Ayrıca Unigram, Bigram, Unigram+Bigram, ve POS özelliklerinin sonuçlarını da karşılaştırmışlardır. Unigram'ın yöntem olarak basit olduğunu fakat doğruluğu arttırmadığını saptamışlardır. Bigramlar olumsuz ifadeleri saptamak için kullanmışlar fakat seyrek olma eğiliminde olduğu için MaxEnt ve SVM'de genel doğruluğu düşürdüğünü görmüşlerdir. Daha sonra Unigram ve Bigram'ları birlikte kullanmayı denemişler ve daha başarılı sonuç verdiğini görmüşlerdir. POS etiketinin ise kullanışlı olmadığını görmüşlerdir. Makine öğrenme algoritmaları (Naive Bayes, maksimum entropi sınıflandırması ve destek vektör makineleri), uzaktan denetimli öğrenme yöntemini kullanırken duyarlılığı sınıflandırmak için yüksek doğruluk sağladıkları sonucuna varmışlardır. Bu tarz

çalışmanın 2009 yılından önce bulunmadığını ve bu çalışmanın yeni bir yaklaşım olduğunu söylemektedirler.

Eliaçık ve Erdoğan (2015) yapmış oldukları çalışmada, mikroblog siteleri üzerinden kullanıcı bilgilerini kullanan kendilerine özgün duygu analizi yöntemi önermişlerdir. Bu yöntemi, mikrobloglardaki finans topluluklarının duygu polaritesinin ölçümünde kullanmışlardır. Çalışmalarında kullanmak üzere model oluşturmak amacıyla Twitter servisinden kısa Türkçe tweetler elde etmişlerdir. Daha sonra elde edilen veriler finans konusunda uzman üç kişi tarafından etiketlenmiş ve onaya sunulmuştur. Sadece onaydan geçen tweetleri çalışmalarında kullanmışlardır. Özellik çıkarımı için unigram ve bigram yöntemlerini birleştirerek kullanmışlardır. Etiketli verilerin sınıflandırma performanslarını yükseltmek için 10-katlamalı çapraz doğrulama yöntemini kullanmışlardır. Daha sonra kullanıcıların herhangi bir konu hakkındaki ilgi ve topluma karşı inandırıcılığının duygu polaritesine olan etkisini gözlemlemek için üd (üyelik derecesi) ve id (ilgi derecesi) değerlerini hesaplamışlardır. Hesapladıkları değeri normal duygu analizi sonucunda çıkan polarite değeriyle çarparak ağırlıklandırılmış polarite değerlerini bulmuşlardır. Haftalık duygu polarite değişimini hesaplamışlardır. Yaptıkları analizler sonucunda önerdikleri yöntemin önceki çalışmalara oranla finansal toplulukların duygu polaritesiyle borsa fiyatları arasında bağdaşıklık oranını daha hassas hesaplandığını görmüşlerdir.

1.1.1. Genel duygu analizi

Duygu analizi, belirli bir olay, durum ya da herhangi bir nesne hakkında, kişilerin iç dünyasında uyandırdığı izlenim olarak tanımlanabilmektedir. Kişilerin belli bir konuya yönelik tutumlarının olumlu, olumsuz ya da tarafsız olup olmadığının belirlenmesi için, metin olarak ifade edilen görüşleri hesaplama yoluyla tanımlayıp, sınıflara ayırmaktadır (Sağlam, 2019). Kişinin duygu ve düşünce gibi öznel bilgilerinin saptanması, veri madenciliğinde duygu analizi yapılarak mümkün hale getirilmiştir. Sosyal medya izleme aracı olarak, kullanışlı bir yöntemdir (Pang ve Lee, 2008; Khan ve diğ., 2015; Santos ve Gatti, 2014). Fikir madenciliği olarak da bilinen duygu analizi yöntemi ile bir çevrimiçi metnin öznel mi yoksa nesnel mi olduğunu ve ifade edilen herhangi bir fikrin olumlu mu yoksa olumsuz mu olduğunu belirlemek için ve metindeki duyarlılığın otomatik algılanması için birçok algoritma

geliştirilmiştir (Pang ve Lee, 2008; Çoban ve Tümüklü Özyer, 2017). Bazıları tartışılan nesneleri ve bunlar hakkında ifade edilen duyguların kutupsallıklarını (pozitif, negatif veya tarafsız) tanımlarken, diğer algoritmalar film incelemesi gibi bir metne genel bir kutupluluk atar (Pang ve Lee, 2004; Çoban ve Tümüklü Özyer, 2017). Bireysel düzeyde duygu analizi, bir konuya veya ürüne dair bilgi toplamada ve bu bilgiler ışığında karar verme sürecinde etkin şekilde rol alır.

Birçok araştırmacıya göre, duygu analizi sınıflandırma problemi olarak görülmektedir. Bu sınıflandırmalar birçok kombinasyondan oluşabilir. İkili olumlu ve olumsuz olarak sınıflandırılacağı gibi üçlü olarak olumlu, olumsuz ve tarafsız olarak da sınıflandırılabilmektedir. Çalışmanın amacına yönelik bu sınıf sayıları dörtlü hatta beşli sınıflara bile ayrılabilmektedir. Dörtlü sınıflandırmada, olumlu, olumsuz, tarafsız ve karışık olarak sınıflandırılırken, beşli sınıflandırmada çok negatif, negatif, tarafsız, pozitif, çok pozitif olarak sınıflandırılmaktadır. Genel olarak duygu analizinde asıl hedeflenen, metindeki içeriğin analizi sağlanarak yazarın duygusunun tespitinin gerçekleştirilmesidir (Sağlam, 2019).

1.1.2. Duygu analizinin zorlukları

Duygu analiz denildiği zaman akla gelen ilk şey metin üzerinde yürütülen çalışmalar olsa da, aslında bu analiz sadece metinler üzerinde değil, beraberinde video, ses ve görsel tabanlı öğeleri de içermektedir. Metin tabanlı çalışmalarda, İngilizce dili haricindeki diğer diller henüz tam anlamıyla gelişmemiştir. Dolayısıyla, duygu analizinde karşılaşılan sorunlar derken, aslında içeriği metin tabanlı olan durumlardaki zorlukları ifade etmektedir. Zorluklardan bahsedilecek olursa (Sağlam, 2019);

- Olumsuzluk Durumları (negation): Dillerde olan bazı olumsuzluk ifadeleri, kendilerini takip eden kelimelerin duygu yönünü tersine çevirmektedir. İngilizce' de "never" ve "not", Türkçe' de "değil" ve "me/ma" terimleri örnek olarak gösterilebilir. Metin içinde geçen "kötü değil" gibi bir ifade, aslında negatif skora sahip "kötü" sözcüğünü olumlu hale getirerek pozitif anlam içermesine sebep olmuştur. Dolaysıyla, olumsuzluk durumlarının tespiti, metnin kutuplarını tersine çevirmesine neden olacağı için önemlidir.

- Kuvvetlendirme (intensification): Kendinden sonra gelecek olan ifadenin duygu skorunu kuvvetlendiren ya da zayıflatan ifadelere kuvvetlendiriciler (intensifier) ve zayıflatıcılar (diminisher) adı verilmektedir. Kuvvetlendiriciler, pozitif ve negatif içerikli duyguların şiddetini arttırırken, zayıflatıcılar ise pozitif ve negatif içerikli duyguların şiddetini azaltmaktadır. Bu duruma örnek gösterilecek olursa; Türkçe de "çok", İngilizce de "very" kelimeleri verilebilir. Örneğin; "güzel" pozitif kelimesi, "çok güzel" şeklinde kuvvetlendirici terim ile kullanıldığı zaman daha pozitif bir anlam yansıtmaktadır. Öte yandan, "nispeten güzel" şeklinde kullanıldığı zaman ise zayıflatıcı terim etkisiyle duygu, daha az pozitif hale gelmiştir.
- Koşul Cümleleri (conditional sentences): Günlük konuşma dilinde sıkça kullanılan koşul cümleleri, terimlerin zaman ekleri ve kelime türleri önemli bir hal almaktadır. Örneğin; "Çok güzel olmasaydı sevmezdim." Koşul cümlesine yapılacak olan çözümleme, normal cümlelere göre farklıdır. Cümleler zamanlarına göre gruplandırılıp, duygu belirten terimler cümledeki pozisyonlarına göre öznitelik olarak kullanılmaktadır.
- Onaylayıcı Sorular (rhetorical questions): Onaylayıcı sorular soru formatında görünmelerine rağmen, aslında duygu aktaran, mesaj veren yapılardır. Bu duruma örnek verilirse, "Bu kıyafet giyilmez mi?" cümlesinde soru sormak yerine pozitif duyguyu güçlendirme söz konusudur. Yüksek pozitif duygu cümle içinde gizli kalmıştır.
- Kinayeli İfadeler (sarcastic phrases): genelde bu yapılar pozitif gibi görünse de, işin aslında iğneleyici bir duygu vardır. Örneğin, "Ne demezsin çok güzel görünüyor." ve "Konu seçimin HARİKA!" cümleleri verilebilir. Literatürde ünlem ve büyük harfle yazılmış terimlerin, kinayeli olarak kabul eden çalışmalar mevcuttur.
- Deyim İfadeleri (idiomatic issues): Deyimler doğal dildeki çeşitliliği gösteren en önemli yapılardan biridir. Deyimi oluşturan kelimeler genelde gerçek anlamlarından çok uzak anlamlarda kullanılmaktadır ve bu durum araştırmacılar için oldukça güç bir problem olarak ortaya çıkmaktadır. Örneğin, "Bıyık altından gülmek" cümlesi anlamsal olarak negatif duygu belirtirken, biçimsel olarak "gülmek" kelimesi pozitif yönde algılanır. Dolaysıyla deyimler, Duygusal Analiz yapıldığı zaman ciddi problemlere sebep olmaktadırlar.
- Arka Plan Bilgisi (background knowledge): kişiler birbirleriyle iletişim kurdukları süre zarfında farkında olmadan, oldukça yoğun arka plan bilgisi kullanmaktadırlar.

Örneğin, “3-3 berabere kaldık” ifadesinin aslında bir müsabaka olduğu, hatta bu müsabakanın spor olduğu ve dahası bu spor dalının hentbol olduğu bilgisi, arka planda okuyan ve yazanın bilgisine dayanmaktadır. Bundan dolayı bu durum aslında Duygu Analizinde karşılaşılan en zor problemlerden birisidir.

- Sosyal platformlarda genelde dil bilgisi kurallarına uyulmamasıyla beraber bu durum, analizde sıkıntı çıkarmaktadır.
- Bir dil için geliştirilen bir metodolojinin, başka dillerde uygulanamaması da sorun olarak karşımıza çıkmaktadır.

Bu zorluklarla beraber Türkçe’nin dil yapısından kaynaklı ilave zorluklar da beraberinde gelmektedir. Türkçe sondan eklemeli bir dildir. Biçimsel olarak oldukça zengin olan Türkçe’de kelimelerin arkasına ekler eklenerek türetilmektedir. Bu ekler, yapım eki ve çekim eki olarak iki gruba ayrılmaktadır. Yapım ekleri kelimenin anlamını değiştirirken, çekim ekleri anlamda değişiklik yapmamakta fakat türlerini değiştirebilmektedir. Bu durumlardan dolayı duygu analizi yapmak zahmetli bir süreçtir.

1.1.3. Duygu analizinin uygulama alanları

Duygu analizi, karar verici mecralara ya da politika belirleyicilere öngörü imkânı veren, yol gösteren, bakış açısını geliştirmeye olanak tanıyan büyük bir kapasiteye sahiptir. Gelişim ve değişime açık her birim için objektif bir geri dönüt mekanizmasıdır. Diğer yandan, sosyal medya platformları düşünüldüğü zaman, bu analiz yöntemi bireysel yönden de çok değerli veriler sunmaktadır. Genel olarak duygu analizinin uygulama alanları şu şekilde sıralanabilir (Sağlam, 2019):

- Ekonomi haberlerinin doğru yorumlanarak, tüketici, üretici ve yatırımcılar için sağlıklı öngörü imkânının sunulması,
- Ürün ve hizmet yönünden değerlendirmeler doğrultusunda, müşteri ilişkilerine yönelik en etkin yöntemin bulunabilmesi, ürün geliştirilmesi,
- Kullanıcıların duygularına göre, en uygun iletişim sistemlerinin geliştirilebilmesi,
- Çevrimiçi platformlarda eğitim gören kişiler için, kişilerin duygularına göre sistemin eğitim içeriğini otomatik olarak güncellediği sistemlerin geliştirilmesi,

Duygu analizi için gerekli veri setleri genellikle; ürünler hakkında yapılan yorumlardan, film yorumlarından, forumlardan, bloglardan, Facebook gibi sosyal paylaşım sitelerinden, haber sitelerinde haberlere yapılan yorumlardan ve Twitter gibi mikroblog sitelerinden elde edilmektedir.

1.1.4. Duygu analizi ile ilgili çalışmalar

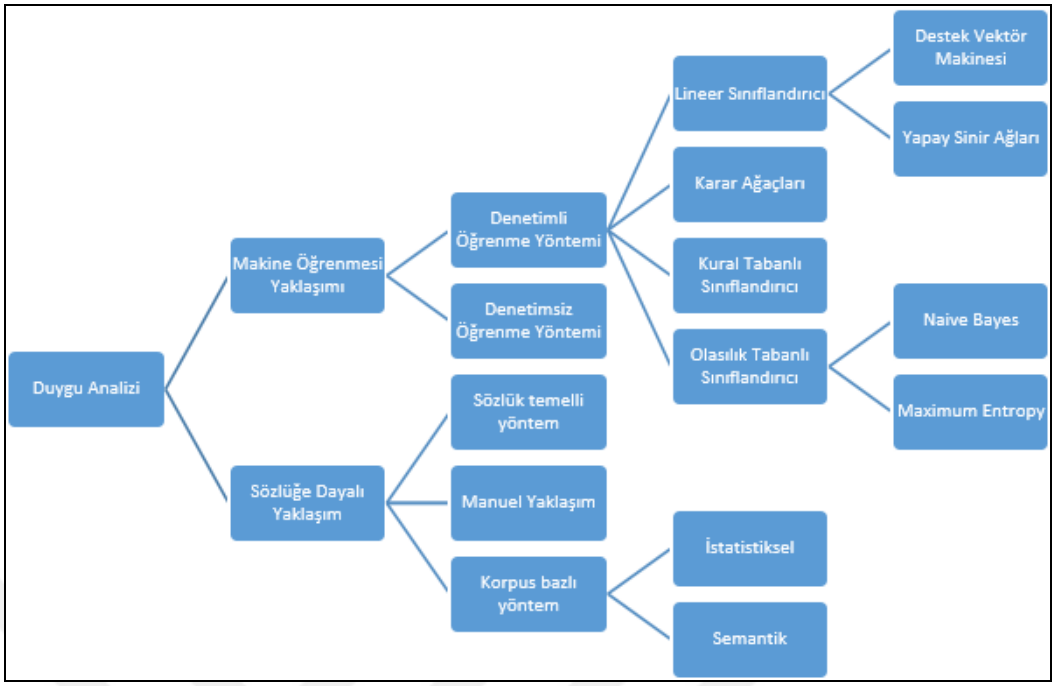
Çalışmanın bu bölümünde, Duygu Analizi ile ilgili yapılan çalışmalar ve uygulanan yöntemler Tablo 1'de özetlenmiştir.

Tablo 1.1. Twitter'da yapılan duygu analizi için denetimli makine öğrenmesi

Yazar	Metotlar	Algoritmalar	Özellikler	Veriler	Sonuçlar
Go ve diğ.	Denetimli Makine Öğrenmesi	Naive Bayes, MaxEnt, SVM	Unigrams, bigrams, POS	Twitter API'sı kullanılarak toplanan tweetler	Hem unigramlarda hem de bigramlarda bulunan MaxEnt %83'lük doğruluk elde edilirken, NB'ye %82.7'lik doğruluk elde etmiştir.
Malhar and Ram	Denetimli Makine Öğrenmesi	Naive Bayes, MaxEnt, SVM ve Yapay Sinir Ağları	Unigrams, bigrams, hybrids (unigrams+ bigrams)	Twitter API'sı kullanılarak toplanan tweetler	Hibrit özellik seçimini kullanan SVM,% 88 doğruluk elde ederken, PCA ile SVM% 92 doğruluk elde etmiştir.
Anton and Andrey	Denetimli Makine Öğrenmesi	Naive Bayes ve SVM	Unigrams, bigrams, hybrids (unigrams+ bigrams)	Online sistem ile toplanan tweetler	Unigramlara sahip SVM,% 81'lik bir doğruluğa ulaşmıştır.
Pak and Paroubek	Denetimli Makine Öğrenmesi	Multinomial Naive Bayes ve SVM	Unigrams, bigrams, trigrams	Twitter API'sı kullanılarak toplanan tweetler	Bigramlı multinom NB, unigram ve trigramlara göre daha iyi bir performans sağlamıştır.
Saif ve diğ.	Denetimli Makine Öğrenmesi	Naive Bayes	Unigrams, POS sentiment-topic features semantic features	STS, HCR ve OMD veri kümeleri	Anlamsal özellikler, unigramları ve POS'ları geride bıraktı. sentiment-topic yaklaşımı, semantic features'den daha iyi sonuç vermiştir.
Hamdan ve diğ.	Denetimli Makine Öğrenmesi	Naive Bayes ve SVM	Unigrams, wordNet, and SentiWordNet	SemEval-2013 veri setleri	Deneyler, DBpedia, WordNet ve SentiWordNet gibi özelliklerin eklenmesinin F ölçüm doğruluğunda SVM ile %2 ve NB ile %4'lük küçük bir artış göstermiştir.

1.1.5. Duygu analiz yöntemleri

Farklı duyarlılık sınıflandırma görevlerini gerçekleştirmek için çeşitli duyarlılık algoritmaları geliştirilmiştir. Duygu analizi yöntemleri Şekil 1.1'de özetlenmiştir.



Şekil 1.1. Duygu analizi yöntemleri (Song ve Xia, 2016)

1.1.5.1. Makine öğrenme yaklaşımı

Genel olarak, makine öğrenmesi yöntemleri, metnin duygu ve görüşlerini öğrenmek amacıyla büyük verilerdeki duyguların kutuplarını, desenlerini otomatik olarak kurallarla keşfetmek için kullanılmaktadır (Sağlam, 2019; Alaei ve diğ., 2019). Bunun için çeşitli algoritmalar geliştirilmiştir. Algoritmaların çoğu denetimli öğrenme yöntemleri başlığı altında toplanmaktadır. Bu başlıklardan, Karar Ağacı, Destek Vektör Makineleri, Yapay Sinir Ağları ve Naive Bayes sınıflandırma yöntemleri bölüm 2.5'de detaylı bir şekilde açıklanmıştır. Kullanılan bu yöntemler sonucunda, metinlerde duygu tahminlerinin yapılması sağlanmaktadır. Genelde veri setinde bulunan tüm kelimeler, birbirlerinden bağımsız şekilde teker teker öznitelik olarak değerlendirilerek sınıflandırma işlemi yapılmaktadır. Sınıflandırma yapılırken; Bu yöntemde ana belirleyici unsur, sınıflandırma sürecinde kullanılacak özniteliklerin doğru bir şekilde belirlenmesinin altında yatmaktadır. Öznitelikler ise duygu belirten terimler ile dilbilimsel özelliklerden oluşmaktadır (Pang ve Lee, 2004).

1.1.5.2. Sözlüğe dayalı yaklaşım

Sözlük temelli yaklaşımda, cümle içerisinde geçen duygu içerikli kelimelerin ele alınıp, cümleyi pozitif veya negatif olarak ayırarak duygu sözlüğü oluşturulmaktadır.

Bu yaklaşımda temel fark, duygu sözlüğünün önceden oluşturularak, analizin oluşturulan duygu sözlüğüne göre gerçekleştirilmesidir. Bu yöntemde kullanılan kelime ve cümlelerin anlamlarına yönelik istatistiksel olarak hesaplamalar yapılır ve metnin görüş kutbu belirlenir. Görüş kutbu belirlenirken, tüm metinlerde ne kadar negatif ne kadar pozitif sözcük varsa incelenip, yüksek sayıda olan duygu sınıfına ataması sağlanmaktadır. Bu şekilde, incelenen metin için, daha fazla negatif terim içeriyorsa metnin duygusu negatif, daha çok pozitif içeriyorsa duygusu pozitif, eşit miktarda ise tarafsız olarak belirlenmektedir (Akın Karaöz ve Şimşek Gürsoy, 2018; Alaei ve diğ., 2019). Sözlüğe dayalı yöntem, yüksek ölçeklenebilirliğinden dolayı sıkça uygulanmaktadır (Pang ve Lee, 2004).

Bu yöntemde, manuel, sözlük temelli ve korpus temelli olmak üzere üç tür yaklaşım bulunmaktadır. Manuel yaklaşım çok zaman alır ve bu nedenle genellikle tek başına kullanılmazlar. Ancak otomatik yöntemler tek başına kullanıldığı zaman hata yaptığı için, bu yöntemleri kontrol etmek amacıyla manuel yaklaşım otomatik yöntemlerle beraber kullanılabilmektedir. Otomatik yöntemlerden biri olan korpus temelli yaklaşım, bilinen kutupluluklara sahip bir dizi duygusal sözcük grubunu kullanmaktadır. Yeni gelen duygu sözcüklerini ve büyük bir korpusta kutupsallıklarını tanımlamak için, modellerin sözdizimsel kalıplarını kullanmaktadır. Herhangi bir kelimenin kutbunun belirlenip, değerlendirmenin ne kadar güçlü olduğunun saptanması, kelimenin "duygusal yönelimi" ile eş anlam taşıması için, sözcüklerin karşılıklı bilgisi hesaplanarak belirlenmiştir. Bu yöntemde, konuşma kalıplarının (sıfat ve zarf) belirli bir kısmına uyan cümleleri taranıp ve daha sonra belgenin yönlendirilmesini hesaplamak adına tüm duyarlılık yönelimleri eklenmektedir. Otomatik yöntemlerden bir diğeri olan sözlüğe temelli yöntemlerde, WordNet veya HowNet gibi sözlük kaynaklarından yararlanılmaktadır. Bu metottaki temel mantık, başlangıçta duyarlılık kelime kümesini elle topladıktan sonra bu kümeyi genişletip, eş anlamlarını ve zıt anlamlarını bulmak için sözlük aramaktır. Yeni sınıflama seti, başka duygu içeren farklı kelimeler üretmek için tekrar tekrar kullanılabilmektedir. Bütün fikir kelimelerini tanımlamak için tek başına korpus yaklaşımını kullanmak, sözlüğe dayalı yaklaşım kadar etkili değildir. Bunun sebebi, bütün kelimeleri kapsayacak şekilde büyük bir korpusun hazırlanmasındaki büyük zorluktur. Buna karşın, bu yaklaşım, keşif sürecinde, yalnızca belirli bir alanda korpus kullanılıyorsa, alana özgü

görüş kelimelerini ve kutuplarını bulmada yardımcı olacağı için büyük bir avantaj sağlamaktadır (Zhang ve diğ., 2014).

1.1.5.3. Hibrit (karma) yaklaşımlar

Karma yaklaşımlarda, sözlük ve makine öğrenmeye dayalı yaklaşımlar, iki duyarlılık kutupluluğunu hesaplamak için paralel olarak çalışabilmektedirler. Sözlükten ve makine öğrenmeye dayalı yöntemlerden elde edilen sonuçlar daha sonra nihai bir duyarlılık polaritesi sağlamak için birleştirilmektedir. Modelin farklı aşamalarında hem sözlük hem de makine öğrenmeye dayalı yöntemleri dahil ederek bir duyarlılık analizi modeli tasarlamak da mümkündür. Bu yaklaşım, başlangıçta sözlüğe dayalı ve öğrenmeye dayalı yaklaşımların bir birleşimidir. Bir metni öznel veya nesnel inceleyerek sınıflandırmaktadır. Metin nesnelse, o zaman duygu analizi görevi sona ererken, bununla birlikte, eğer metin öznel ise, daha sonra pozitif veya negatif olarak sınıflandırılır. Sıfır kutuplamalı bir metin için ise, tarafsız etiket atanmaktadır (Alaei ve diğ., 2019).

1.1.6. Duygu analiz tekniklerinin sınıflandırılması

Kontrolsüz (denetimsiz) ve kontrollü (denetimli) olmak üzere iki farklı yöntem kullanılmaktadır.

1.1.6.1. Kontrolsüz (denetimsiz) teknikler

Denetimsiz öğrenme yöntemi; belirsizlik durumlarında ya da istenilen özel bir sonuç tanımlanmamış ise bu yöntemden söz edebilir (Hastie ve diğ., 2001). Bu yöntemde, ilgili veriler gözlemlenir. Bu gözlemler sonucu bulunan özellikler arasındaki benzerliklerden yola çıkılarak sınıflar tanımlanır. Denetimsiz yöntemde, veriler eğitilirken etiketsiz veriler kullanılmaktadır. Bu yöntem, daha çok keşif amaçlıdır. Yani verilerdeki önceden bilinmeyen yapı ve ilişkilerin bulunmasını sağlamaktadır. Amaç, veri içinde bulunan birbirlerine benzer grupları bulup, verilere ilişkin daha detaylı bilgi edinmek için verilerin temel yapısını ve dağılımlarını modellemektir. Genellikle kümeleme, öznitelikler arasındaki benzer ilişkilerin bulunması, veri boyutlarının indirgenmesi ya da olasılık yoğunluk tahmini gibi amaçlara yönelik kullanılmaktadır. Bu yöntemde kullanıcı sisteme müdahale etmez. Sadece girdi

verilerini sisteme girer, fakat işaretlemelerde bulunmaz. Sistem kendi kendine otomatik olarak keşif yapar ve veriler arasındaki ilişki ağını ortaya çıkarmaya çalışır. Denetimsiz olarak elde edilen veriler, denetimli öğrenme algoritmaları için tekrar kullanılabilmektedir (Çağlayan, 2018; Kızılkaya ve Oğuzlar, 2018). Örneğin, veri tabanı kayıtlarında genç yaşlı bilgisi yoksa çıkarılan kurallar denetimsiz olarak yapılır. Bu yöntem daha çok, sonraki yöntemler için fikir vericidir.

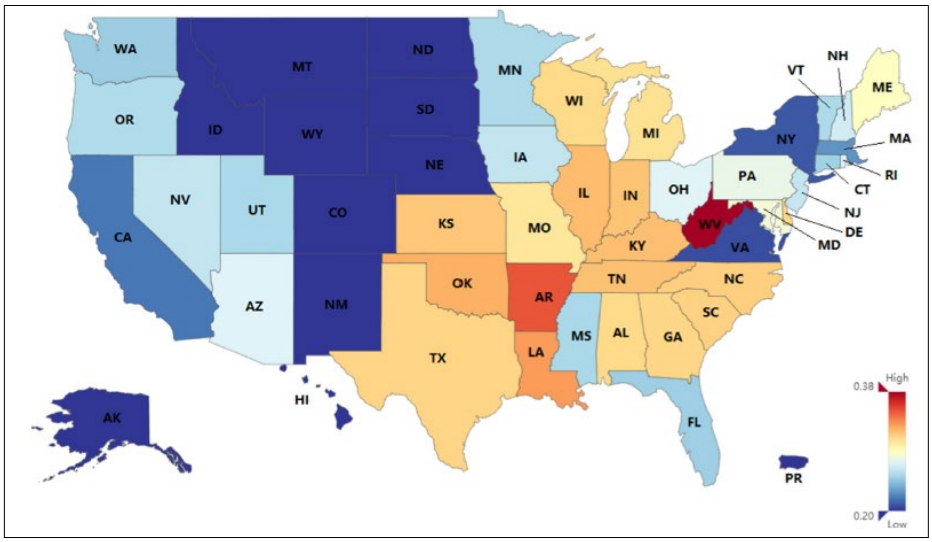
1.1.6.2. Kontrollü (denetimli) teknikler

Denetimli öğrenme yöntemi; kesin veya iyi tanımlanmış bir hedef olduğunda denetimli öğrenmeden bahsedilebilir (Hastie ve diğ. , 2001). Veri örneklerinden hareket ederek, her bir sınıfa ait özellikler bulunur. Bu özellikler kural cümleleriyle ifade edilir. Buradaki asıl amaç; veriden bilgi almak ve anlamlı bir sonuç çıkarmaktır. Denetimli öğrenme, eğitimli veri setinden yani etiketli veri setinden yola çıkarak bilinmeyen veri setinin eğitilip, modelinin çıkarılmasını sağlamaktadır. Eğitim veri seti, kullanılacak olan algoritma için ele alınan gözlemlerden oluşmaktadır. Bu verilerin kullanılmasıyla oluşturulan algoritma ile çıkarımlarda bulunulmaktadır ve böylece bir model oluşturulmaktadır. Eğitim verilerinden oluşturulan model ise test verisi ile denetlenmektedir. Bu denetim elde edilen modelin, gerçek değerlere ne kadar yaklaştığını yani modelin ne kadar başarılı çalıştığını kontrol etmek amacıyla kullanılmaktadır. Denetimli öğrenme kısaca; temelde amacının belli olduğu veri setinin sınıflandırmasından yola çıkarak, sonuçları bilinmeyen veri setlerine yönelik tahminlerin yapılmasına dayanmaktadır. Bu tahminler yapılırken ayrıca test verileri ile de kontrollü gidilmekte ve başarı oranı kontrol edilmektedir. (Çağlayan, 2018; Kızılkaya ve Oğuzlar, 2018). Veri tabanı kayıtlarındaki genç yaşlı kategorisi bu duruma örnek olarak gösterilebilir. Bu veri tabanı üzerinden genç yaşlı olduğuna dair kural çıkarma işleminde denetim kullanılmaktadır (Argüden ve Erşahin, 2013).

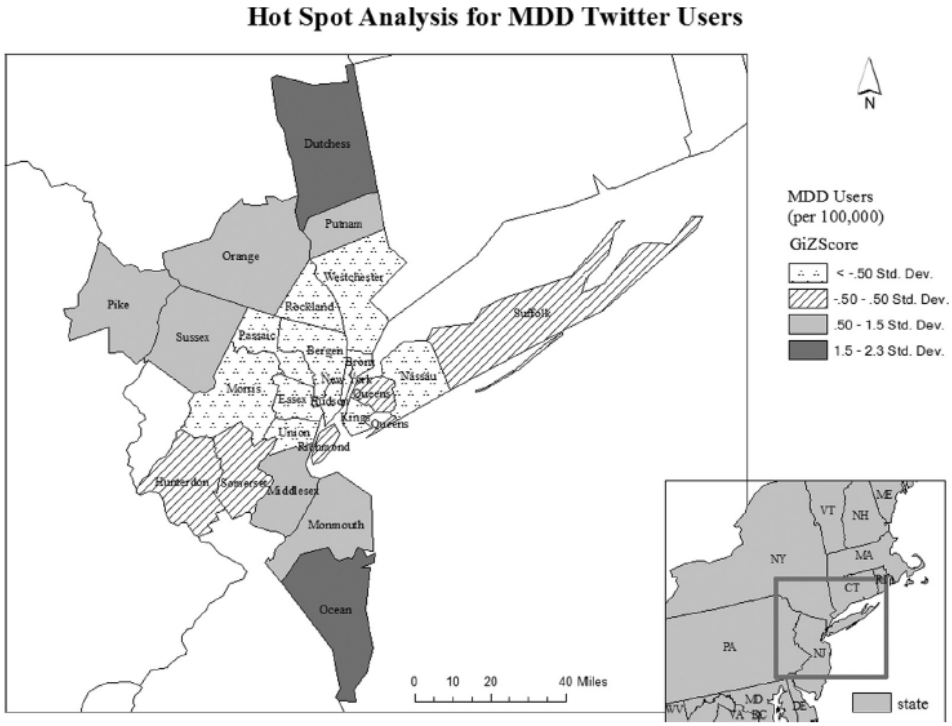
1.2. Duygu Analizinde Mekansal Analiz ve Coğrafi Bilgi Sistemleri

Klasik duygu analizinde güncel bir yaklaşım sosyal medya verilerinin CBS teknolojisi yardımıyla mekânsal analizine yönelik çalışmalardır. Mekânsal analiz ve duygu konusu, insanların mekânı nasıl deneyimlediklerini ölçmek ve politika için potansiyel olarak önemlidir. Tweeter verilerine dayalı olarak oluşturulmuş coğrafik veritabanları, anlık dinamik haritalama için yeni ve güncel bilgi paylaşım platformu olma

yolundadır. Coğrafi konumlu tweeter verilerinden yararlanarak global boyutta demografik verilere dayalı çalışmalar, sosyal araştırmalar, sağlık araştırmaları vb. ile lokal anlamda belli bir temaya ilişkin kişisel görüşlerin değerlendirildiği çalışmalar (geo-located tweet data analysis, geospatial text mining) bu kapsamda ele alınabilir (Şekil 1.2). Ayrıca duygulardaki değişimin zamansal olarak izlenmesi de önemli bir konudur. Diğer önemli bir konu coğrafi etiketli tweetlere eklenmiş mekânsal bilgi kullanımıyla, coğrafik detayların (tweetler için) belli bir komşuluğunda mekânsal bağımlılık (mekânsal otokorelasyon) ilişkisinden kaynaklanan duygu analizinin gerçekleştirilmesidir. Örneğin sosyal medya analizi, hizmet sağlayıcıların müşterileri ve hizmet kullanıcılarını olumsuz yönde etkileyen olaylara nasıl yanıt verdiğini ve olay algısını da etkilediğini göstermiştir. Mekânsal yakınlık/ komşuluk, hizmetlerin olumlu/ olumsuz şekilde etkilendiği yerlere en yakın olanların verdiği tepkilerin ölçülmesinde önemli bir göstergedir. Tepki verme hızı ve sağlayıcıların olaya nasıl tepki vereceği hususu, hizmet sağlayıcıların algılarını da etkilediği gösterilmiştir. Mekânsal regresyon ve otokorelasyon, yöntemleri bilinen konumların veya mekânların kişilerin tercihlerini nasıl etkileyebileceğini araştırmak için kullanılabilir. Entegre bir mekânsal ve duyarlılık analizi yaklaşımı kullanarak coğrafi etiketli Twitter kullanıcılarının görüşleri ile mekân ilişkisi kurulabilmektedir (Şekil 1.3).



Şekil 1.2. USA obezite haritası (2013) (Wanga ve diğ., 2018)



Şekil 1.3. Depresyon bozukluğu hot spot haritası (Yang ve Mu, 2015)

2. DUYGU ANALİZİ VE VERİ MADENCİLİĞİ

2.1. Veri Madenciliğine Genel Bakış ve Tarihsel Süreci

Veri madenciliğinin çıkışı, verinin bilgiye dönüşüm süreciyle yakından alakalıdır. Veri madenciliğinde hedef önemlidir. Hedefler sayesinde, veri madenciliği çok farklı amaçlara hizmet edebilmektedir. Veri madenciliğinin hedefleri arasında, somut olan modelleri mantıksal kurallara oturtmak ya da bu somut modellerin görselleştirilmesini sağlamak vardır. Dolayısıyla bu yönüyle insan merkezlidir. Veri madenciliği; istatistik, makine öğrenimi, yapay zekâ, örüntü tanımlama ve veri görselleştirmesi gibi pek çok teknik alan ile yakından ilgili olmasından ötürü çok disiplinli bir alandır. Veri madenciliği tek başına çözüm ve sonuç vermez, fakat bilgiye ulaşmak için bir adımdır.

Veri madenciliğinde, verilerde bulunan bilgilerin türüne göre farklı modeller uygulanmaktadır. Veri madenciliği modelleri tahmin edici ve tanımlayıcı olmak üzere iki ana başlık altında toplanmıştır. Tahmin edici modeller, varlıkların davranışlarının tahmininde kullanılması için örüntü bulmak üzerine kurulmuştur. Tanımlayıcı modellerde ise; bu örüntülerin insanlar tarafından anlaşılıp, yorumlanmasını sağlayan örüntü oluşturmak üzerine kurulmuştur (Fayyad ve diğ.,1996a, 1996b).

Tarihsel süre boyunca veriler yorumlanmış ve bu yorumlanan anlamlı bilgiler üzerinde bir sistem oluşturulmuştur. Böylece zaman gözetmeksizin, bilgi taşınır hale gelmiştir. Veri yönetimi tarihsel olarak sıralanacak olursa başlangıcı 1950'lerin sonlarına dayanmaktadır. Bilgisayarların sayım için kullanıldığını göz önüne aldığımız zaman, bugüne kıyasla başlangıçta, oldukça ilkel ve maliyeti yüksek olan bir iş olarak görülmekteydi. 1960'larda veri tabanı ve depolama kavramları ortaya çıkmıştır. Bu kavramlarla beraber, basit öğrenmeli bilgisayarlar da gelişmeye başlamıştır. 1970'li yıllara gelindiğinde, ilişkisel veri tabanı kavramı ortaya çıkmış ve bu sayede uzman kişiler basit kurallar üzerine kurulu sistemler geliştirmişlerdir. 1980'lerde artık veri tabanı yönetim sistemi popüler hale gelmiş, birçok alanda kullanılmaya başlanmıştır. 1990'larda, veri yığınları oluşmaya başlamış ve bu yığınlar arasından faydalı, işe yarayan bilginin nasıl çıkarılabileceği üzerine çalışmalar başlamıştır. 2000'li yıllarda

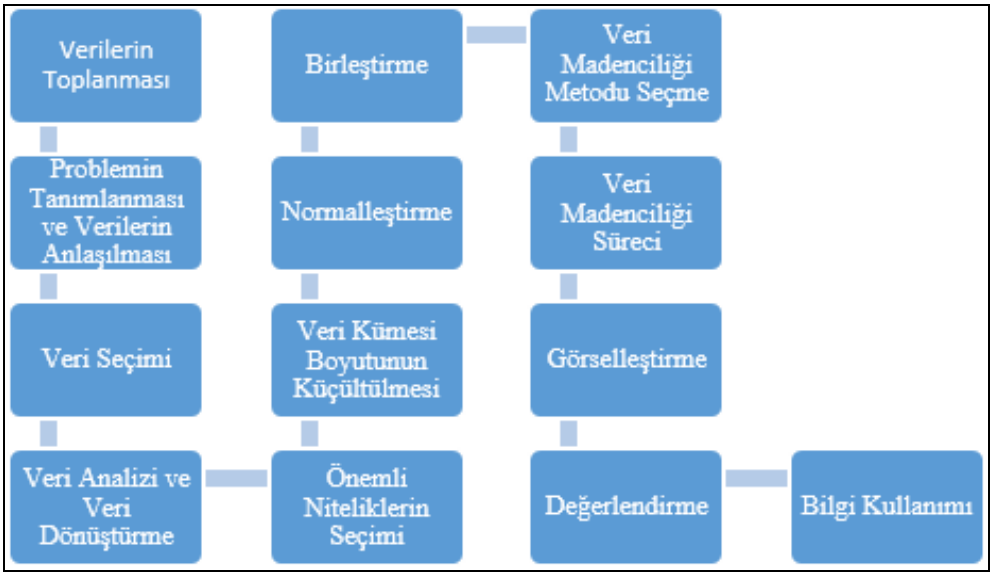
ise sürekli gelişmeler olmuş ve veri yönetimi yaygın hale gelmiştir (Savaş ve diğ., 2012).

2.2. Veri Madenciliği Süreci

Veri madenciliği aslında bir süreçten meydana gelmektedir. Veriler veri ambarlarından alınır, derlenir, düzenlenir ve yorumlanırlar. Veri madenciliği algoritmalarının uygulanması için verilerin bazı ön işlemlerden geçirilmesi gerekir. Bunun nedeni ise, veriler toplanırken bilgiler eksik, yanlış veya tekrarlı olabilirler. Bundan dolayı ham veriler ön işlemden geçirilir. Daha sonra algoritmalar uygulanarak örüntüler ortaya çıkarılır. Bu örüntüler yorumlanıp, değerlendirildikten sonra faydalı bilgi ortaya çıkarılır. Veri madenciliğinde başarılı olmak için; işin ve verilerin özelliklerinin ayrıntılı bir şekilde analizi gerekmektedir. Veri madenciliği sürecinde genellikle şu adımlar izlenmektedir (Savaş ve diğ., 2012).

- Problemin tanımlanması
- Verilerin hazırlanması
- Modelin kurulması ve değerlendirilmesi
- Modelin kullanılması
- Modelin izlenmesi

Veri Madenciliği sürecinin adımları Şekil 2.1.'de gösterilmiştir.



Şekil 2.1. Veri madenciliği sürecinin adımları (Dondurmacı ve Çınar, 2014)

2.2.1. Problemin tanımlanması

Veri madenciliğinde, problemin belirlenip, tanımlanması sonuca giden yolda en önemli kısımdır. Varılmak istenen sonucun bilinmesi ve işin başarılı olması için gerekli olan kriterler bu aşamada planlanır. Analiz için gerekli verilerin araştırılması, problemin tanımlanması ile mümkündür. Bu adımda, kaynak yeterliliği, kaynak ihtiyacının saptanması, sonuçların nasıl raporlanması gerektiği araştırılır. Çalışma sırasında ne tarz bir yol izleneceğine bu aşamada karar verilir. Ayrıca bu çalışmanın ne kadar maliyetli olacağı da problemin tanımlanmasıyla araştırılabilir. Çalışma için risk teşkil eden durumlar önceden saptanıp, alınacak önlemlere ilişkin öngörülere de bu aşamada yer verilmektedir (Ayık ve diğ., 2007).

2.2.2. Verilerin hazırlanması

Bu aşamada, ham veriden başlayıp en son çıkan veriye kadar yapılması gereken düzenlemeler yer alır. Model kurulması sürecinde en önemli kısımdır. Toplama, değer biçme, birleştirme, seçim, dönüşüm başlıkları altında toplanabilir. Model kurulumunda bu aşama %90'lık kısmı oluşturmaktadır. Dolayısıyla herhangi bir problem çıktığında bu aşamaya sıklıkla geri dönülür.

Toplama (collection); tanımlanmış problem için gereken verilerin ve veri kaynaklarının belirlenmesi aşamasıdır. Değer biçme (assessment) ise; toplanan verilerin, birbiriyle uyumunun değerlendirildiği, uyumsuzluklar ve eksik verilerin giderildiği aşamadır. Veri madenciliğinde veriler birçok kaynaktan toplandığı için veri uyumsuzlukları oluşmaktadır. Bu uyumsuzlukların başlıca nedenleri; farklı zaman diliminde elde edilmiş olmaları ve kodlamalardaki veya ölçü birimlerindeki farklılıklardır. Bazı uygulamalarda toplanan verilerin istenilen özelliklere sahip olmadıkları görülür. Uygun olmayan veriler ile eksik veriler tutarsızlık oluşturabilir. Bu tutarsız ve hatalı veriler "gürültü" diye adlandırılmaktadır. Bu nedenlerden ötürü, iyi sonuç alınacak model oluşturmak için, verilerin gürültülerden temizlenmiş ve eksik verilerin tamamlanmış olması şarttır (Ayık ve diğ., 2007). Eksik verilerin tamamlanmasında ise, aşağıdaki yöntemler uygulanabilir (Çilingirtürk ve Altaş, 2010).

- Eksik veriler, veri kümesinden atılabilir.
- Kayıp değerler yerine bir genel sabit kullanılabilir. Tüm kayıp değerler yerine aynı

sabit kullanılabilir. Ancak bu kullanım ilerde sorun oluşturabilir.

- Değişkenin tüm verilerinin ortalaması alınarak eksik değer yerine bu ortalama kullanılabilir.
- Değişkenin tüm verileri yerine, sadece bir sınıftaki verilerin ortalaması alınarak eksik veri yerine kullanılabilir.
- Verilere regresyon analizi veya karar ağacı modeli kurularak, eksik değer tahmin edilebilir ve eksik değer yerine kullanılabilir.

Birleştirme (consolidation) aşamasında ise, değer biçme aşamasında belirlenen sorunlar giderilip, tüm veriler tek bir veri tabanında toplanmaktadır. Sorun giderme işlemleri titizlikle yapılmazsa ilerideki aşamalar için büyük sorunlara zemin hazırlanmış olur.

Seçim (Selection) aşamasında; kurulacak modele bağlı olacak şekilde veri seçiminde bulunulur.

Son olarak dönüştürme (transformation) aşamasında; çözümlemede kullanılacak verilere ilişkin değişkenler uygun şekle dönüştürülmelidirler. Veri dönüştürmede kullanılan bazı yöntemler şu şekilde örneklendirilebilir;

- Veriler gürültü veya aşırı değer içeriyorsa düzleştirme işlemi uygulanır. Düzleştirme işlemi, kümeleme ve regresyon yöntemlerini kullanır.
- Veriler aşırı detay içeriyorsa özetlenerek sade hale getirilir.
- Verilere normalleştirme işlemi uygulanarak, 0-1 veya 1-1 aralıklarına indirgenir (Babaoğlu,2015).

2.2.3. Modelin kurulması ve değerlendirilmesi

Veri madenciliğinde, verilerin çözümlendiği en önemli aşamadır. Tanımlı problemde en uygun modeli bulmak için çok sayıda model kurulup denenmektedir. Bundan dolayı model kurma aşaması en iyisi bulunana kadar yinelenen bir süreçtir. Model kurulduktan sonra o modelin doğruluğunun kontrolü için birçok yöntem bulunmaktadır. Sıkça kullanılan yöntemlerden birisi basit geçerlilik testidir. Basit olarak bilinen bu yöntemde, verilerin %5 ile %33 arasındaki kısmı test verisi olarak ayrılır ve kalan kısım üzerinden modelin öğrenimi gerçekleştirilir. Daha sonra ise bu

veriler üzerinde test işlemi yapılır. Bir sınıflama modelinde hata oranı, yanlış olarak sınıflandırılmış olay sayısının, tüm olay sayısına bölünmesi ile hesaplanmaktadır. Doğruluk oranı ise doğru olarak sınıflandırılan olay sayısının, tüm olay sayısına bölünmesi ile hesaplanmaktadır (Doğruluk oranı = 1- Hata oranı). Oluşturulan model uygulamaya koyulmadan önce son kez her yönüyle değerlendirilmektedir. Modelin amaca yönelik olup olmadığı ve problem için her yönüyle çözüm sağlayıp sağlamadığına karar verilir (Babaoğlu,2015).

2.3. Veri Madenciliği Uygulama Alanları ve Veri Madenciliği ile İlgili Yapılan Çalışmalar

2.3.1. Veri madenciliği uygulama alanları

Veri madenciliği, büyük hacimde veri bulunduran her yerde kullanılabilir. Karar verme sürecinde ihtiyaca cevap vermektedir. Birçok alanda rekabet ortamının artışı, veri tabanı yönetim sistem teknolojilerinin hızla gelişmesi, verilerin sürekli artışı, çok sayıdaki verilerin rahatlıkla toplanıp saklanabilmesini sağlamasıyla veri madenciliği uygulamalarına olan ilgi her geçen gün artmaktadır. Veri madenciliğinin uygulama alanları aşağıdaki gibi sayılabilir: (İnan, 2003; Albayrak, 2008; Akgöbek ve Çakır, 2009).

Pazarlama: Veri madenciliği uygulaması sayesinde, hangi tarzda müşterinin yapılan kampanyalara geri dönüt vereceğini ortaya çıkarmaktadır. Müşterilerin yaş, cinsiyetlerine göre satın alma tarzlarının belirlenmesinde, yapılan kampanyalara cevap verme oranları, mevcut müşteri üzerine yeni müşterilerin kazanılması, satış tahmini gibi alanlarda kullanıldığı zaman oldukça başarılı şekilde sonuç vermektedir.

Banka ve Sigortacılık: Birbirinden farklı finansal göstergeler arasındaki korelasyonun hesaplanması ve yorumlanması, kredi kartlarındaki dolandırıcılıkların tespit edilmesi, talep edilen kredilerin değerlendirilmesi, müşterilerin kredi kartı harcamalarına göre profillerinin saptanması, usulsüzlüklerin saptanması, riskli müşterilerin saptanması gibi konularda kullanıldığında faydalı sonuçlar vermektedir.

Sağlık: Yapılan testlerin sonuçlarının değerlendirilmesinde, gen haritasının çözümlenip yorumlanmasında, yeni hastalıkların ve bunlara neden olan

mikroorganizmaların keşfedilmesi ve sınıflandırılmasında, tedavi sürecinde izlenecek adımların belirlenmesinde, çeşitli hastalıkların ön tanısında, kalp krizi risk haritası çıkarımında, acil servislerde hastaların hastalıklarına göre risk ve öncelik tespitinin yapılmasında vs. çok geniş bir uygulama alanına sahiptir.

Telekomünikasyon: Abonelik tespitlerinde, hat yoğunluğunun tespitinde, kalite ve iyileştirme analizlerinde, telefon dolandırıcılığında etkili şekilde kullanılmaktadır.

Coğrafi Bilgi Sistemleri: Bölgelerin coğrafik özelliklerinin saptanıp sınıflandırılmasında, yeni yerleşim yerlerinin belirlenmesinde, kentlerdeki suç oranlarının tespitinde, kentlerde verilecek olan posta kutusu, bankamatik, park, otobüs durakları gibi hizmetlerin konumlarının belirlenmesinde başarılı şekilde uygulanmaktadır.

Web Madenciliği: İnternet üzerindeki bilgiler gün geçtikçe artmakta, hacimce yer kaplamaktadır. Bu kalabalık bilgilerin ayıklanıp, çözümlenmesi, web sayfalarının tasarlanması gibi alanlarda kullanılmaktadır.

Görüldüğü üzere veri madenciliği birçok alanda, çok farklı amaçlara hizmet etmektedir. Veri madenciliği sayesinde birçok alanda birbirinden farklı konular üzerinde yöntem ve teknikler geliştirilmektedir. Veri madenciliği konularıyla yapay zekâ konularının birbirlerine teknik ve algoritma olarak yakın olmasından dolayı her alanda rahatça kullanılmaktadır (Babaoğlu, 2015).

2.4. Görsel Veri Madenciliği (Visual Data Mining)

Veri madenciliğini daha etkili hale getirmek için görsel veri madenciliği kullanılmaya başlanmıştır. Görsel veri madenciliği, görselleştirme ve veri madenciliğinin birlikte etkili şekilde kullanılmasıyla ortaya çıkmış olup, görselliği bilgisayar ve kullanıcının arasında iletişim aracı olarak kullanmaktadır. Bu sayede veriler daha etkili şekilde yorumlanabilmektedir. Amaç, verilerin daha hızlı ve kolay şekilde keşfedilmesini sağlamaktır. Başka bir ifade ile görsel veri madenciliği, algılanabilirliği arttırmak amacı ile diğer ilişkili verileri de içerecek şekilde verilerin görsel olarak temsil edilmesini sağlamaktadır. Bu süreçte çok boyutlu veriler, aralarındaki ilişkiler korunarak 2 ya da 3 boyuta indirgenip algılanabilir hale getirilmektedir (Vatansever,

2008). Görsel veri madenciliği, veriyi hazırlama, model çıkarma ve onaylama aşamalarının hepsini görsel şekilde göstererek keşfetmeye çalışır. Görselleştirme yaklaşımlarına göre üç sınıfta incelemek mümkündür.

- Verilerin görselleştirilmesi
- Ara sonuçlarının görselleştirilmesi
- Veri madenciliği sonuçlarının görselleştirilmesi

2.4.1. Verilerin görselleştirilmesi

Görsel veri madenciliği, görsel veriler üzerinde somut bir bakış açısı sağlamaktadır. Gürültülü, homojen dağılmamış veriler üzerinde etkili şekilde uygulanabilir. Görsel veri madenciliği karmaşık algoritmalara ihtiyaç duymamaktadır.

2.4.2. Ara sonuçlarının görselleştirilmesi

Veri analizi algoritmalarında isteğe ve amaca göre ara sonuçlar üretilebilir. Ara sonuçlar uygun teknikler kullanılarak görselleştirilebilir. Ara sonuçlar için görselleştirme yapmanın altında yatan sebep, yapılan uygulamadan bağımsız bir algoritmik parça elde etmektir. Yapılan çalışmalarda her zaman uygun algoritmalar bulunamayabilir. Böyle durumlarda kullanılan algoritmaların ara sonuçlarının görselleştirilmesi çok amaçlı sonuçlar üretebilir. Daha sonra üretilen ara sonuçlar için uygun algoritmalar seçilerek veri madenciliği çalışması başarılı şekilde tamamlanabilir.

2.4.3. Veri madenciliği sonuçlarının görselleştirilmesi

Veri madenciliğinde yapılan görselleştirme sayesinde veriler yorumlanabilir hale gelmektedir. Böylece kullanılan algoritmalar parametreleri değiştirilerek sonuçlar gözlemlenir ve en uygun algoritma kolaylıkla seçilebilir. Görselleştirme sayesinde sonuçlar algısal olarak da değerlendirildiği için etkin şekilde kullanılır.

2.5. Veri Madenciliğinde Kullanılan Temel Yöntemler

2.5.1. Tahmini yöntemler

Tahmini yöntemlerde, bağımlı değişkenler, bağımsız değişkenlerin birer fonksiyonu

olarak tahmin edilir. Önceki kayıt işlemlerine bakarak tahmin yoluyla boşlukları doldurur (Özcan, 2014). Tahmini yöntemlerin amacı, sonuçları bilinen veri kümesinden yola çıkarak bir model kurulması ve kurulan bu modelden faydalanılarak, sonuçları bilinmeyen veri kümeleri için sonuçların tahmin edilmesini sağlamaktır. Örneğin, bankalar daha önce verdikleri tüm kredilere yönelik müşterilere ait tüm verilere sahiplerdir. Kredi alan müşteriler bu verilerin bağımsız değişken kısmını oluştururken, alınan kredilerin geri ödenip ödenmemesi bağımlı değişken kısmını oluşturmaktadır. Kurulan bu model banka için sonraki kredi taleplerinde müşterilerin aldığı krediyi ödeyip ödemeyeceğinin tahmininde kullanılmasını sağlamaktadır (Özekes, 2003).

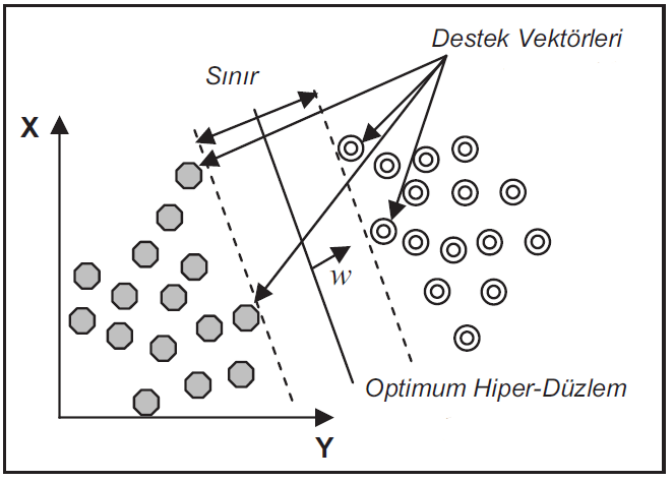
2.5.1.1. Makine öğrenmesi

Makine öğrenmesi, matematiksel ve istatistiksel yöntemlerin, veriler üzerinden çıkarım yaparak, yaptığı çıkarımlarla bilinmeyenlere yönelik tahminlerde bulunan sistemlerin bilgisayarlar vasıtasıyla modellenmesidir (URL-3). Yapay zekâ tekniklerinden olan makine öğrenmesi, bilgisayarlar için programlama yapılmasına ihtiyaç duyulmadan, öğrenme yetkisi sağlamaktadır. Makine öğrenmesinde, önceden yapılmış olan gözlemler göz önüne alınarak, doğru tahminlerde bulunmak için otomatik olarak teknikler geliştirilerek gerçekleştirilmektedir. Makine öğrenmesindeki asıl hedef, karmaşık örüntüleri bilgisayarlara algılatıp, veriye dayalı akılcı kararlar alma becerisi kazandırmaktır. Veri madenciliğinde örüntü tanıma alanı ile yakından ilişkilidir (Türkmenoğlu, 2015).

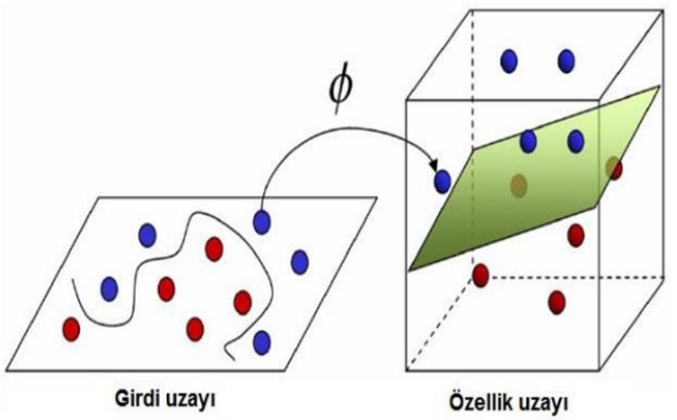
Makine öğrenme süreci, tıpkı veri madenciliği süreci gibidir. Her iki sistemde de desenleri bulmak için veriler üzerinde tarama işlemi yapılmaktadır. Fakat aralarında fark vardır. Veri madenciliğinde, insanlar verileri karşılaştırarak bilgi çıkarımı sağlarlar, fakat makine öğrenmesinde; elde edilen bilgi, programın öğrenme becerisini geliştirmesi için kullanılır. Denetimli ve denetimsiz öğrenme yöntemleri kullanılarak makine öğrenmesi oluşur. Makine öğrenmesinde; Destek Vektör Makinaları, Naive Bayes, Karar Ağaçları, Yapay Sinir Ağları ve Mesafeye Dayalı Algoritmalar sıkça kullanılan yöntemlerdendir. Bu yöntemlerden ilk olarak,

Destek Vektör Makineleri:

Destek vektör makineleri, çok boyutlu ve doğrusal olmayan, sınıflandırma yapmak için geliştirilmiş ve sınıflar arasında ki sınırın belirlenmesi için en uygun algoritmaların kullanılmış olduğu makine öğrenmesi yöntemlerinden birisidir. V. Vapnik tarafından 1990'larda ortaya atılmıştır. Temeli, iki sınıfı birbirinden ayırmak için, hiper düzlem belirlenmesi mantığına dayanmaktadır. Ayırıcı bir model ile tanımlanan destek vektör makineleri ayrıştırıcı bir sınıflandırıcı olarak kullanılır. Girdi olarak farklı iki kategoriye ayrılmış veri verildiği zaman destek vektör makineleri, sınıflandırılmamış yeni verileri sınıflara ayırmak için model üretir. Destek vektör makinelerinin ayırma mekanizmaları, uzaydaki boyutuna göre değişiklik göstermektedir. Ayırma mekanizmaları, iki boyutlu uzayda doğrusal, üç boyutlu uzayda düzlemsel ve çok boyutlu uzayda hiperdüzlem şeklindedir. Bu mekanizmalar sayesinde veriler iki ya da daha fazla sınıfa ayrılabilme yeteneğine sahip olmaktadırlar. Şekil 2.2'de doğrusal olarak ayrıştırılabilen veri setleri için hiper düzlemin belirlenmesi gösterilmektedir. Girdi uzayı doğrusal olarak ayrıştırılamazken, Kernel fonksiyonları sayesinde daha yüksek boyutlu lineer olarak ayrıştırılabilen bir uzaya taşınarak, doğrusal olmayan veriler başarılı olarak sınıflandırılabilmektedir. Şekil 2.3'de doğrusal olmayan veri setinin üst uzaya taşınması gösterilmektedir. Veri setini sonsuz sayıda sınıflara ayıracak çoklu düzlem vardır. Fakat buradaki amaç, bilinmeyen sınıflama hatasını en küçük yapacak hiperdüzlem seçmektir (Kaynar ve diğ., 2017).



Şekil 2.2. Doğrusal olarak ayrılabilen veri setleri için hiper-düzlemin belirlenmesi (Kavzoğlu ve Çölkesen, 2010)



Şekil 2.3. Doğrusal olarak sınıflandırılamayan girdi uzayının bir üst boyuta taşınması (Güldoğan, 2017)

Makine öğrenmesine, girdi veriyi reel sayılarla ifade eden öznitelik vektörü olarak verilir. Verinin, sayılardan oluşan vektör olarak ifadesi zordur. Bag-Of-Words metodu bir metni öznitelik vektörüne dönüştürmek için kullanılabilir. Her kelime bu metotta öznitelik vektörünün birer elemanı olarak görülür. Kelimelerin ifadesi için, öznitelik vektöründe reel sayılar belirlenir. Bu sayılar belirlenirken; o metnin içindeki frekansı (TF), o metinde bulunup bulunmama durumu (binary), ya da tüm eğitim setindeki frekansına göre (DF) değerler hesaplanıp kullanılır. Günümüzde performansından dolayı Destek Vektör Makineleri oldukça popüler bir metottur (Doğan ve Diri, 2010). Destek Vektör Makineleri günümüzde, yüz tanıma sistemleri ya da ses analizleri gibi birçok uygulamada başarılı şekilde uygulanmaktadır. Boyutu yüksek olan uzaylarda etkili şekilde kullanılabilmektedir. Örneklem sayısının az olduğu durumlarda, boyut sayısı fazla olsa bile bu yöntem başarılı şekilde kullanılmaktadır. Karar fonksiyonunda, eğitim noktaları kullanıldığı için bellek verimli şekilde kullanılmaktadır. Bu yönleri ile tercih edilmesi büyük avantaj sağlamaktadır (URL-1).

Naive Bayes:

Herhangi bir belgenin ya da dokümanın sınıflandırılması için kullanılan, “Sınıf Koşullu Bağımsızlık” olarak da adlandırılan basit bir sınıflandırıcıdır. Bu önerme, sınıflandırmada kullanılan özniteliklerin istatistiksel açıdan birbirinden bağımsız olması gerektiğini ifade eder. Temeli, her özniteliğin sonuca olan etkilerinin hesaplanması üzerine kurulmuştur. İstatistikteki Bayes teoremine dayanmaktadır. Bu teorem; belirsizlik halindeki herhangi bir durumun modelini oluşturmaktadır. Modeli

oluşturduktan sonra, bu durumla alakalı evrensel doğrular ve gerçekçi gözlemler doğrultusunda belirli sonuçların elde edilmesine fırsat verir. Belirlenmiş değişkenler ile tanımlanan bağımsız değişkenler arasındaki ilişkilerin analizi sağlanır. Belge veya dokümanlardan elde edilen analizleri, tahmine yönelik ve tanımlayıcı olarak ilişkilendirir (Lewis, 2005; Gülçe, 2010). Naive Bayes sınıflandırıcısı, eğitim kümesi ile eğitilir. İlk olarak, kullanılan her bir öznitelik değerlerinin sınıflarla olan ilişkilerinin olasılık oranını hesaplar. Daha sonra hesaplanan olasılık değerlerini içeren modeli çıktı olarak verir. Ölçülmek istenen verilerin birbirlerinden bağımsız olmaları gerekmektedir. Sisteme belli miktarda sınıfı belli, öğrenilmiş bilgi sunulmaktadır. Yani sınıflandırılması gereken örnek verilerin hangi sınıflara ait olduğu bellidir. Sisteme sunulan yeni test verileri ile öğretilmiş verileri üzerinde değerlendirilen olasılık işlemleri, daha önceki olasılık değerlerine göre işlenir. Daha sonrasında verilerin, hangi test verisi kategorisinde olduğunun tespiti yapılır. Test verisinin gerçek sınıfının bulunmasındaki kesinlik, öğretilmiş verinin sayısının fazlalığıyla doğru orantılıdır. Diğer sınıflandırıcılara göre avantajı; eğitim kümesi çok az miktarda olsa bile gerekli parametreler tahmin edilebilir. Bağımsızlık önermesinden ötürü kullanım alanı sınırlı gözükse de, yüksek boyutlu uzayda, yeterli sayıdaki öznitelik kümesi bileşenlerinin bağımsızlık koşulu esnetilerek başarılı sonuçlar elde edilebilir. Belirsizlik durumlarındaki karar verme aşamasında çok kullanışlı bir yöntemdir. Diğer sınıflandırıcılara göre zayıf yönü ise, değişkenlerin birbirlerinden tamamen bağımsız olması ve değişkenler arasındaki ilişkinin modellenemiyor olmasıdır (Argüden ve Erşahin, 2008).

Bayes yöntemi koşullu olasılık durumları ile ilgilidir. Herhangi bir koşullu olasılık durumu P(X=x | Y=y) = R şeklinde tanımlanır. Bu ifade; "Eğer Y = y doğru ise, X = x olma olasılığı R'dir" anlamına gelmektedir. X ve Y'nin alabileceği değerlerin her kombinasyonu için koşullu olasılıkları belirleyen tabloya verilen isim koşullu olasılık dağılımıdır ve P(X|Y) ile ifade edilir. Bayes Kuralı Formül (2.1)'de şu şekilde tanımlanır.

$$P(XIY) = (P(YIX) \times P(X)) / P(Y) \quad (2.1)$$

Formül (2.1)'de belirtilen durum; Y'nin gerçekleşmesi halinde X'in gerçekleşme ihtimalinin ne olduğudur. Bu değer "X'in gerçekleştiği durumda Y'nin gerçekleşme

ihtimali" ile X'in gerçekleşme ihtimali ile çarpıp, Y'nin gerçekleşme ihtimaline bölerek bulunmaktadır. Örneğin; herhangi cep telefonu operatör şirketlerinden biri, müşterileri arasında yapmış olduğu araştırmada cep telefonu kullanımında arka arkaya 4 ay boyunca sürekli düşüş gösteren müşterilerinin %30'unun hatlarını kapatıp başka operatörlere geçtiğini tespit etmiştir. Ayrıca her 100 müşteriden 8'inin farklı sebeplerden ötürü hatlarını kapattığı ve her 100 müşteriden 17'sinde arka arkaya 4 aylık sürede sürekli düşüş yaşandığını tespit etmişlerdir. Bu bilgilerden hareketle hattını kapatan bir müşterinin, son 4 ayda sürekli azalma yaşanan müşterinin olmuş olma olasılığı nedir?

P(Düşüş I Kapatmış) = P(Kapatmış I Düşüş) x P(Düşüş) / P(Kapatmış)

P(Düşüş I Kapatmış) = (0,3 x 0,17) / 0,08 = % 64

Çıkan bu değerde, müşterilerin yarısından fazlasının son 4 aydaki kullanımının sürekli azalış gösteren müşterilerden geldiği görülmektedir. Çıkan oran oldukça yüksek bir orandır. Şirket bu müşterilerin kullanım eğilimlerinden kim olduklarını tahmin edebilmektedir. Bu müşteriler için cazip kampanyalar yapılarak müşteriler elde tutulabilirse, şirket toplam kaybettiği müşterilerinin yarısından fazlasını elde tutabilecektir (Argüden ve Erşahin, 2008). Diğer bir yöntem olan

Karar Ağaçları:

Uygulanması ve anlaşılması konusunda kolaylık sağladığı için, sınıflandırma algoritmaları içerisinde en yaygın olarak kullanılan yöntemlerden biridir. Bu yöntem, açgözlü öğrenme temeline dayanan, sürekli ve kesik değerlerle çalışabilen sınıflandırıcıdır. Yani az kuralla sonuca gitme yaklaşımıyla hareket eder. Bu öğrenme algoritması, birden fazla sınıflandırıcı üretip, daha sonra onların tahminleriyle yeni veriyi sınıflandırır. Dengesiz veri setinde hata dengeleme yöntemlerine sahip olduğundan büyük veri tabanlarında kusursuz çalışmaktadır. Kaybolan verilerin doğruluğu büyük olasılıkla korunduğundan dolayı, kayıp verilerin tahmininde etkili bir metot olarak kullanılmaktadır (Sun ve Li, 2008; Türe ve diğ. 2008).

Verileri sınıflandırmak için ilk olarak ağaç oluşturulmaktadır. Veriler, bu ağaca uygulanıp, sonuçlar sınıflandırılmaktadır. Karar ağaçları kendi içinde düğüm, dallar

ve yapraktan oluşmaktadır. İlk olarak düğümler oluşturulur. Düğümler soruları oluşturmaktadır. Daha sonra farklı cevaplara göre, dallar ortaya çıkmaktadır. Dalların sonucunda yaprak oluşmuyorsa, başka bir düğüm meydana getirerek sonuca ulaşılmaya çalışılmaktadır. Yapraklar ise hangi sınıfa ait olduğu sonucunu barındırır. İç düğümler birer girdiyi ifade etmektedir. Yaprakların haricindeki tüm düğümler, belli özniteliğe göre bir kural ifade etmektedir. Bu kural, karar aşamasından sonra seçilecek dalı belirler. Son durumdaki sınıflandırma sonucunu taşır. Tüm karar düğümlerindeki kurallar ile sınıflandırma sonuçlarına ulaşılmaktadır. Örneklerin öznitelik değerlerine göre, kural düğümleri oluşturulur. Karar düğümünde oluşturulan özniteliğin değerinin ayırt etme gücü ne kadar fazlaysa, özniteliğin kullanımı o kadar avantajlıdır ve iyi sonuç verir. Özniteliklere göre izlenecek yolu, düğümler arasındaki bağlantılar gösterir. Öznitelik seçimi karar ağaçları için önemlidir. Öznitelik seçimi ise; Bilgi Kazancı Oran değeriyle belirlenmektedir. Daha açık şekilde anlatılacak olursa, ilk etapta elde bulunan veriler üzerinde uygulanacak test tanımlanır. Her düğüm farklı özellikteki testi gösterir. Testin sonucunda ise ağacın dalları meydana gelir. Dallar oluşturulurken veri kaybının en aza indirilmesi için tüm verileri kapsayacak sayıda birbirinden farklı dal oluşturulmalıdır. Oluşan dallar uygulanan testin sonucunu göstermektedir. Ortaya çıkan her dal ile tanımlanacak olan sınıfların belirlenmesi hedeflenmektedir. Elde edilen dalın sonucunda sınıflandırma işlemi tamamlanmıyorsa tekrardan karar düğümü oluşturulmalıdır. Sınıflandırma olana kadar aynı işleme devam edilir. Bu işlem dalların sonucunda sınıflandırmanın tamamlanmasıyla son bulur. Sınıflandırmanın elde edilmesiyle birlikte yapraklar da oluşmuş olur. Elde edilmek istenen sınıflandırmadaki sınıfların tanımlanması için, yapraklar elde var olan verileri kullanmaktadır (Argüden ve Erşahin, 2008).

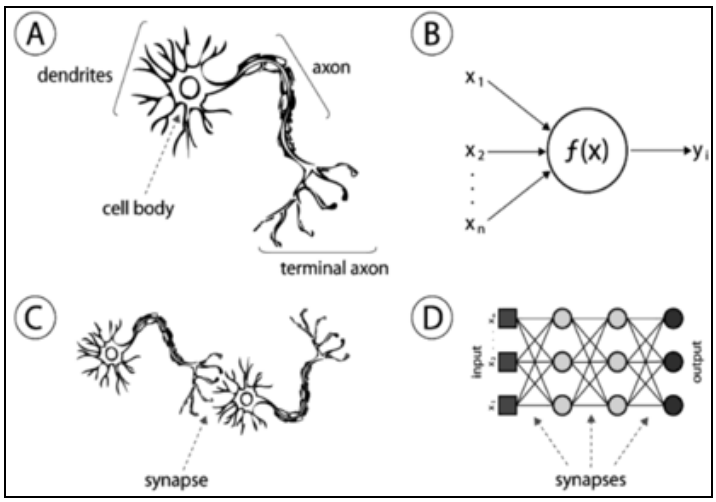
Verilerin karar ağacı teknikleri kullanılarak sınıflandırılması işlemi iki basamaktan oluşmaktadır. Birinci basamak öğrenme basamağıdır. Bu basamakta model oluşturmak için, sınıflama algoritmasının yardımıyla önceden bilinen eğitim verisi analiz edilir. Öğrenilen model, karar ağacı ya da sınıflandırma kuralı olarak gösterilir. İkinci basamak sınıflandırma basamağı olarak bilinir. Test verileri, bu basamakta sınıflandırma kurallarının ya da karar ağacının doğruluğunu saptamak amacıyla kullanılır. Doğruluk seviyesi kabul edilir oranda ise, oluşturulan bu kurallar yeni verilerin sınıflandırılması için kullanılır.

Karar ağaçları birçok durumda kullanılmaktadır. Bunlardan bazıları; belirli bir sınıfın olası üyesinin belirlenmesinde, çeşitli vakaların yüksek, orta, düşük gibi çeşitli kategorilere ayrılmasında, parametrik modellerde kullanılacak çok sayıdaki değişkenlerden en önemlilerinin seçilmesinde, yalnızca belirli alt gruplara özgü olan ilişkilerin tanımlanması gibi durumlarda sıkça kullanılmaktadır (Albayrak ve Koltan Yılmaz, 2009). Veri madenciliğinin bazı durumlarında bu analiz sıkça kullanılır. Örneğin; bir kişinin kredi geçmişi değerlendirilerek yeni kredi durumunun belirlenmesi, satışları etkileyen durumların belirlenmesi ya da ürün hatalarına yol açan etmenlerin belirlenmesinde bu yöntem sıklıkla kullanılmaktadır. Veri madenciliği uygulama kapsamında bu yöntemin kullanılmasının faydalı yönlerine; kuruluşlarının ucuz olması, sonuçların yorumlanmasının daha kolay olması, veri tabanları sistemlerine kolayca entegre edilebilir olması, gelecek tahmini için kurallar tanımlaması, belli alt gruplara özgü ilişkilerin tanımlanmasının bu yöntemde daha kolay olması, sonucun doğrulanması için işlem kalabalığının önüne geçmesi ve son olarak sınıflandırma yapmaya olanak sağlaması örnek verilebilir. Tahmin analizlerinde öngörü yapılacak değişken için değerlerin sürekli oluşu, doğru sonuç verme olasılığını düşürdüğü için bu husus, karar ağacı sınıflandırıcısının zayıf yönü olarak sayılabilir. (Argüden ve Erşahin, 2008).

Yapay Sinir Ağları:

Yapay sinir ağları, insan beyninden yola çıkarak geliştirilmiş olan, ağırlıklandırılmış bağlantılar sayesinde birbirlerine bağlanan ve her birine ait bellek bulunan işlem elemanlarından oluşan bilgisayar yazılımlarıdır (Maltarollo, 2013). İnsan beynindeki biyolojik sinir ağı ile yapay sinir ağı modeli karşılaştırması Şekil 2.4'de gösterilmiştir. Sonuç katmanını elde etmek için ağırlıkların hesaplanmasına dayanmaktadır. Bu yöntemle beraber kurulan modelin kontrolü sağlanmakta olup ayrıca öğrenme faaliyetiyle beraber model geliştirilmektedir. Bu süreçte temel olan davranış biçimlerini anlamak ve hatanın en az olmasını sağlamaktır. Bu yöntem sınıflandırmayı doğru yapmakta ve doğru sonuçlar vermektedir. Fakat bu yöntemin zayıf yönü olarak öğrenme süresinin uzunluğu ve çıkan sonucun ifadesinin zor oluşu gösterilebilir. Öğrenmenin ne kadar gerçekleştirildiği, eğitim veri kümesinde hesaplanan ağırlıkların, test veri kümesi üzerinde kullanılmasıyla saptanır. Elde edilen ağırlıkların etkinliği doğru olana kadar ağırlıklar üzerinde düzeltmeler yapılarak tekrardan

hesaplama yapılır. Ağırlıklar sayesinde verinin hangi sınıfa ait olduğu belirlenebilir. Öğrenme süreci uzun sürmesine rağmen, oldukça duyarlı sınıflandırma yapılmaktadır. Yapay Sinir Ağlarında hesaplama yapılırken belli bir yöntem adımı izlenmez. Sinir ağları dağınık olduğundan, ilişkilendirme yapmak için iç kuralları kendisi üretir. Akabinde bu kurallar daha önce yapılan örneklerle kıyaslanarak yeniden düzenlenir. Deneme yanılma yoluyla işin nasıl yapılması gerektiği öğrenilir. Yapay Sinir Ağları, geçmişteki tecrübelerinden yararlanarak tahminde bulunmayı amaçlar ve belli amaçlar için oluşturulur. Yapay Sinir Ağları, öğrenebilme ve bilgi işleyebilme özelliği sayesinde yapısı ve ağırlığı değişkendir. Bu sayede Yapay Sinir Ağları karmaşık bir problem karşısında, problemi çözebilme yeteneğine sahiptir. Bu yöntem öğrenme mekanizmasıyla geliştirilmiş olan nöronlar arası bağlantı ve bu bağlantıların ağırlıkları üzerine kurulmaktadır. Bu bağlantı ne kadar karmaşık yapıda ise kurulmuş model de o denli karmaşık olmaktadır. Nöronlar katman adı verilen alanlarda bir araya gelir (giriş, çıkış ve gizli katman). Kurulan modele sürekli olarak eğitim verileri girilir ve çıkan sonuçlar gerçek sonuçlarla kıyaslanarak kurulan modelde düzeltmelere gidilerek ağırlıklarda değişimler yapılır. Hata seviyesi minimum kabul edilir bir düzeye ulaştığı zaman model sona ermiş ve kurulmuş olur. Yapay sinir ağları hisse senetleri ile piyasanın tahmin edilmesinde, kredilerin puanlandırılması ve değerlendirilmesinde, semptomlara göre hastalıkların tahmin edilmesinde kullanılmaktadır (Argüden ve Erşahin, 2008).



Şekil 2.4. Biyolojik sinir hücresi ve yapay sinir ağı karşılaştırması, (a) insan nöronu, (b) yapay nöron veya gizli birlik, (c) biyolojik sinaps, (d) YSA sinaps (Maltarollo, 2013)

Mesafeye Dayalı Algoritmalar ile Sınıflandırma:

Bu yöntemdeki algoritmalar, yeni gözlemlerin hangi sınıfa ait olduğunu bulmak için, sınıfı belli olan ve komşu olarak adlandırılan örnek kümedeki gözlem değerlerini kullanır (Demircan,2015). Yani mevcut verilerin birbirlerine olan uzaklıklarını hesap ederek sınıflandırma yapılır. K- en yakın komşu (KNN) algoritması sıkça kullanılan bir yöntemdir. KNN algoritması, öğrenme temelli bir algoritmadır. Eğitim verisinde öğrenilen model, test verileri üzerinde kullanılarak, sınıf etiketi ataması yapılır. Belirlenmiş olan bu model etiketlerine göre, her yeni gelen veriler için etiket ataması yapılır. Algoritmanın temel yapısı, yeni gelen verilerin en yakınında bulunan k komşuya bakılarak, yeni verinin sınıf etiketi için tahminde bulunmaktan ibarettir. Sınıfı belli olmayan bir veri için şu yol takip edilir: İlk olarak bu yöntemde veri kümesi rastgele k adet alt kümeye ayrılır. Küme sayısı "k" değeri olarak isimlendirilir. Her kümedeki nesnelerin özelliklerinin ortalaması olan merkez noktalar hesaplanır. Kümelerdeki nesnelerin küme merkezine olan uzaklıkları bulunur ve hangi küme merkezine yakınsa, yakın olan kümeye dâhil edilir. Yeni gelen nesnelerle artış ya da dışarı nesne vererek azalış gösteren kümelerin ortalaması olan merkez noktalar tekrar tekrar hesaplanır ve nesnelerin kümelerinde değişiklik sabit olana kadar devam edilir. En yakın uzaklığa sahip verilerin olduğu sınıfa göre, yeni gelen örnek için sınıf tahmininde bulunulur. Bu algoritma türü eskiden beri sıkça kullanılan etkili ve basit bir yöntemdir. Bu yöntemin performansını etkileyen etkenler vardır. Mesafe ölçütü, bilinmeyen noktanın en yakınındaki komşu sayıları (k) ve ağırlıklandırma bu etkenlerden bazılarıdır (Argüden ve Erşahin, 2008). Uzaklık ölçütünde genellikle Öklid mesafesi kullanılır. Uzaklık hesaplamasında, i ve j noktaları için eşitlik Formül (2.2)'de verilen Öklid uzaklık formülü kullanılabilir (Demircan, 2015).

$$D(i,j)=\sqrt{\sum_{k=1}^{p}\left(\chi_{ik}-\chi_{jk}\right)^2} \qquad (2.2)$$

KNN algoritmasının tercih edilme nedenleri arasında; eğitimin olmayışı, uygulamasının kolay oluşu, analitik olarak izlenebilmesi, gürültülü eğitim verilerine karşı dirençli olması gibi etmenler vardır. Fakat buna rağmen dezavantajları da bulunmaktadır. Çıkan sonuçlar algoritmada en başta seçilen merkez noktalara bağımlıdır. Bu noktaların seçimiyle sonuçlarda değişiklikler meydana gelebilmektedir.

Yüksek miktarda bellek alanına olan ihtiyaç, veri setinin boyutu arttıkça artan işlem yükü ve maliyet, performansının; k komşu sayısı, uzaklık ölçütü ve öznitelik sayısına bağlı olarak değişmesi gibi etmenler dezavantajları arasında gösterilebilir (Argüden ve Erşahin, 2008).

2.5.1.2. Doğal dil işleme

Bilgisayar ve insan dillerinin etkileşimini inceleyen bir alandır. Doğal dil işleme, dillerin kurallı yapısını çözümleyerek; anlaşılabilmesi, işlenebilmesi ya da tekrar üretilebilmesi için çalışan bir alandır. Bu alanın asıl amacı, bilgisayar ile doğal dildeki iletişimin kurulmasını sağlamaktır. Bunun için ise doğal dil kurallarının bilgisayara öğretilmesi gerekmektedir. Bunun için, genel bir sözlük ve bu sözlüğün kullanılabilmesi için çeşitli algoritmalara ihtiyaç duyulmaktadır. Ayrıca dilin yapısından bağımsız olarak algılanması gereken özel bir bilgi tabanına da ihtiyaç vardır. Otomatik çeviri, soru-cevap makineleri, konuşma tanıma, metin özetleme, duygu analizi gibi birçok konudaki çalışmalarda kullanılmaktadır. Doğal dil işlemenin kapsamları içine, biçimbirimsel çözümleme, biçimbirimsel anlam belirsizliği giderme, POS etiketleme (part-of-speech) gibi metotlar girmektedir(Türkmenoğlu, 2015). Doğal dil işleme; biçimbirimsel çözümleme, biçimbirimsel belirsizlik giderme ve POS etiketleme yöntemlerinden oluşur.

Biçimbirimsel Çözümleme:

Bu metot, cümle içinde bulunan her kelimenin kök ve eklerini ayrıştırıp, görevlerini belirlemek için uygulanan bir süreçtir. Kelimelerin kök ve ekleri çözümlendikten sonra kelime tipleri (isim, sıfat, zarf, fiil, edat vs.) de belirlenir. Biçimbirim yalnızca kelime içine girdiği zaman anlamlı olan, tek başına anlam ifade etmeyen en küçük dilsel birimdir. Bu metotta girdiler ile isim soylu, fiil soylu kelimeler ve rakamlar için tasarlanan Sonlu Durum Makineleri kullanılarak sonuca ulaşılır. Örneğin “arabaya bindim” cümlesindeki gibi kullanıldığında eylem, “bin tane balon” cümlesindeki gibi kullanıldığı zaman ise isim olarak kullanıldığı, biçimbirimsel çözümleme sonucunda çıkacaktır (Türkmenoğlu, 2015).

Biçimbirimsel Belirsizlik Giderme:

Bu metot, biçimbirimsel çözümleyicinin cümle içindeki her kelime için verilen birden fazla sonuçtan, doğru olanı bulmak için kullanılır (Türkmenoğlu, 2015).

POS Etiketleme:

Doğal dil işleme sürecinde kelimelerin dilbilimsel kategorilere ayrıştırılıp, atanması bu sürece büyük katkı sağlamaktadır. Bu kategoriler isim, fiil, sıfat, zamir, edat, bağlaç olarak ayırılabilir. Bu metotla beraber, kelimeler cümle içerisindeki dilbilgisel özelliklerine göre ayrıştırılır. Bu işlem aslında Biçimbirimsel çözümleme içinde yapılan bir işlem adımıdır. Kelimelerin farklı kategorilerde farklı anlamlar taşıması duygu analizi için önemli bir durumdur ve bu durumun saptanıp kullanılması yapılan çalışmanın başarısını olumlu yönde etkilemektedir (Türkmenoğlu, 2015).

2.5.1.3. Makine öğrenmesinde doğal dil işleme

Makine öğrenmesi reel sayılarla çalışırken doğal dil işleme kelimelerle çalıştığı için ortaya sorunlar çıkmaktadır. Bu sorunlar, makine öğrenmesi için belirlenen reel değerli öznitelik vektörleri seçilerek giderilebilir. Öznitelik vektörleri oluşturulurken; metindeki sözcükler, sözcük sayıları, sözcük türleri, N-gram gibi özellikler kullanılır. Vektör oluşturulurken ise; "Vektör Uzay Modeli" kullanılır. Her nesne, vektör uzay modelinde vektör olarak tanımlanmaktadır. Vektör uzayının eksenlerini, nesnelerdeki farklı özellikler oluşturmaktadır. Böylece her nesne vektör uzayında belli bir konumda bulunmaktadır. Vektör uzay modeli uygulanan nesneler yapısal hale gelmiş olur. Ayrıca kelimelerin, metinlerin öznitelik vektörüne dönüşüm sürecinde; seçilen öznitelik veri setinin veri kaybını önlemek amacıyla fazlaca bilgi içermesi gerekmektedir. Fakat makine öğrenmesinin düzgün çalışabilmesi için veri boyutunun mümkün olduğunca küçük olması gerekmektedir. Bunun için ise boyut indirgeme metotları kullanılır.

Öznitelik eleme ve seçme için; ilgili özniteliğin ilgili metinde bulunup, bulunmama durumu (Presence), ilgili metindeki Frekans (TF-Term Frequency), tüm metinlerdeki Frekans (DF-Document Frequency), bu değerlerin kombinasyonları (TF-IDF) gibi yöntemler vardır. Farklı vektör oluşturma yöntemleri mevcuttur.

- Binary Vektör Oluşturma

Bu yöntemde, metinsel veriler 1 ve 0'lar ile ifade edilir. Kelimelerin sözlükteki varlıklarına göre, veri içinde barındırdığı bu değerleri almaktadır. Bu şöyle örneklenebilir:

Veri: "Ayşegül bugün hava çok bulutlu şemsiye almayı unutmamalısın."

Sözlük: {Bugün, Akşam, Şemsiye, Güneş}

Binary Tanımlama: {1, 0, 1, 0} (Çalış ve diğ., 2013).

- Frekans Sıklığına göre Vektör Oluşturma

Binary tanımlamadan farklı olarak, veri içindeki kelime köklerinin kaç defa geçtiği bilgisinin de tutulduğu bir tanımlama biçimidir.

Veri: "Bugün hava çok güneşli, bugün gezmeye çıkalım mı? "

Sözlük: {Bugün, Akşam, Şemsiye, Güneş}

Frekans Tanımlama: {2, 0, 0, 1} (Çalış ve diğ., 2013).

- TF-IDF Ağırlıklandırma ile Vektör Oluşturma

Bu yöntemde ilk adım; ağırlıklandırma çalışması yapılarak tüm dokümanda geçen veriyle, aranan verinin oranının belirlenmesidir. Bu sebeple, veri havuzu içinde ayırt edici özelliği çok olan ya da doküman içinde sıkça geçtiği için ayırt edici özelliği olmayan kelimeler belirlenir. Ağırlıklandırma içerisinde kullanılan TF değeri, belirlenen veri içeriğinde kaç kere kelime köklerinin geçtiğini yani frekans bilgisini tutmaktadır. IDF değeri ise; aranan kelime, sözcüğün tüm veri havuzu içinde kaç kez geçtiği bilgisi ile ilgili değeri vermektedir. Bu değer sayesinde tüm veri havuzu içinde bulunan ve ayırt edici özelliği fazla olan kökler ile ayırt edici özelliği olmayan her yerde var olan kelime kökleri rahatlıkla belirlenmektedir (Çalış ve diğ., 2013). Makine öğrenmesinde doğal dil işleme yönteminde birçok model kullanılır. Bunlardan en sık kullanılanlar N-Gram Model, Bag-Of-Words ve Olumsuzluk Durumlarıdır. İlk yöntem olan;

N-Gram Model:

İstatistiksel temelli olan, komşu sözcüklere (önceki-sonraki veya önceki iki ve daha

fazla) bakarak sözcüğün niteliğini anlamaya ve bulmaya çalışan bir yöntemdir. N-Gram tabanlı sınıflandırma yöntemi, doküman içerisindeki n-gram karakterlerinin kullanım sıklığına dayanmaktadır. N burada komşu sözcük sayısını belirtmektedir. Farklı uzunluk olarak 2-gram, 3-gram, 4-gram şeklinde modeller kurulabilir. N-Gramların yeterli şekilde toplanabilmesi için ise uzun paragraflara ihtiyaç duyulmaktadır. N-Gram frekans yaklaşımı dilden bağımsız olarak çalışır. Yani belirli bir dil hakkında detaylı dilbilgisine veya sözlük yapısına ihtiyacı yoktur. N-Gram yöntemi, dokümanların sınıflandırılması için basit ve güvenilir bir yöntem olarak kullanılmaktadır (Doğan, 2006). Diğer yöntem olan;

Bag-Of-Words metodu (Kelime Torbası):

Verileri reel sayılardan oluşmuş vektör olarak ifade etmek, genelde veri kümelerinde zor bir işlem olarak karşımıza çıkmaktadır. Metin işlemede, kelimeleri öznitelik vektörüne dönüştürmek için bag-of-words metodu kullanılmaktadır. Bu metot, sadece frekans bilgilerini kullanarak, metin içindeki tüm sözcükleri sıralarından bağımsız şekilde ele almaktadır. Bu yöntemde, tüm kelimeler öznitelik vektörünün birer elemanı olarak yer almaktadır. Öznitelik vektöründe kelimelerin ifadesi için reel bir sayı belirlenirken; o metin içerisindeki frekansı (TF), o metinde bulunup bulunmama durumu (Binary) ya da tüm eğitim setindeki frekansına göre (DF) vb. değerler hesaplanıp kullanılabilmektedir (Türkmenoğlu, 2015). Ayrıca bu modelde kelimelerin metin içerisindeki sırasının önemli olmadığı kabul edilmektedir (Lewis, 1992). Yani metinler sırasız ve gramer bilgisinden yoksun şekilde ele alınır. Bundan ötürü yüksek oranda bilgi kaybı kaçınılmaz olur. Olumsuzluk bildiren sözcüklerin ilgili olduğu, sözcük ve sözcük gruplarını yakalamak da mümkün olmamaktadır. Bu durumun üstesinden, kullanılan n-gram veya olumsuzluk durumlarının ön işlenmesiyle işaretleme metotları gelir (Türkmenoğlu, 2015). Son olarak bir diğer yöntem olan;

Olumsuzluk Durumları:

Doğal dilde, bazı olumsuzluk içeren sözcük ya da eklerin katıldığı kelimenin ya da cümlenin anlamlarını tersine çevirmesiyle oluşur. Kelime ya da cümle haricinde olumsuzluk durumu oluşturan bir de ifadeler bulunmaktadır. Türkçe Dilinde olumsuzluk 2 farklı şekilde yapılmaktadır.

- “değil” ve “yok” kelimelerinin cümleye eklenmesi,
- “me/ma” olumsuzluk ekinin kelimelere eklenmesi ile yapılır.
- Ayrıca olumsuzluk durumu içeren ifadelerle de, olumsuzluk durumu pekiştirilebilir. Olumsuzluk durumu içeren ifadelerden bazıları “!, ☹, :-(, :*(, :’-(, :-/” dır.

2.5.2. Tanımlayıcı yöntemler

Karar vermemize yardımcı olacak veri setindeki veriler arasında olan ilişkileri, bağlantıları, davranışları bularak mevcut verilerdeki örüntünün tanımlanmasını sağlayan yöntemdir. Fikirlerin ne olduğu belirlenmemiş olmasına rağmen, veri tabanında gizli örüntüler bulmaya çalışır (Özcan, 2014). Bu yöntemin hedefleri arasında, veriler üzerinde aranan desenlerin çeşitliliği, bu desenlerden çıkarılacak bilgi kalitesi, elde edilen bilgilerin yorumlanarak davranış biçimlerinin tespit edilmesi vardır. Ayrıca tespit ettiği davranış biçimlerinin alt veri setlerinin özelliklerini de tanımlamak vardır. Tanımlayıcı yöntemler tekrarlı faaliyetlerde ya da özelliği bilinen yeni bir verinin yapıya katılmasında nasıl hareket edileceği konusunda karar vermeye yardımcı olmaktadır (Argüden ve Erşahin, 2008).

2.5.3. Kümeleme yöntemi

Kümeleme analizi, gruplarının kesin şekilde belli olmadığı, değişkenleri benzer alt kümelere ayırmaya yardımcı olan çok değişkenli istatistiksel bir analiz yöntemidir. Verilerdeki gruplamayı ortaya çıkarmaktadır. Bu yöntem; küme içindeki benzerliğin maksimum olması, yani küme içindeki mesafelerin minimum olması, kümeler arası benzerliğin ise minimum yani kümeler arasındaki mesafenin maksimum olması mantığına dayanmaktadır (Argüden ve Erşahin, 2008). Bu analizin temel amacı, değişkenlerin sahip oldukları karakteristik özelliklerini baz alarak aralarında farklılık bulunan ya da birbirleri arasında yüksek oranda benzerlik bulunan verileri gruplara ayırmak ve araştırmacıya özet bilgi sunmaktır. Gruplara ayırma işlemi, aynı grupta yer alan gözlemlerin birbirine benzemesini, farklı gruplardaki gözlemlerin ise birbirinden farklı olmasını dikkate almaktadır. Kümelere ayrıldıktan sonra ise, kümelerde bulunan her verinin aynı duyarlılığa sahip olması ve verilerin homojen kümeler oluşturması beklenmektedir. Seçim ölçütleri önceden belirlenerek, birbirlerine benzer verilerin aynı küme içinden toplanmasıyla birlikte, aynı kümede olan birimler birbirleri ile

örtüşürken, diğer küme birimleri birbirleri ile örtüşmemektedir. Dolayısıyla, her küme kendi içinde homojenlik gösterirken, kümeler arasında heterojenlik söz konusu olmaktadır (Berberoğlu, 2011; Terzi, 2012). Çok boyutlu uzayda oluşturulan kümeler gösterildiği zaman, kümeleme yöntemi başarılı gerçekleştirilmişse, aynı küme içinde bulunan birimler birbirlerine çok yakın çıkacaktır (Turanlı ve diğ., 2006). Bu yöntem birçok alanda rahatlıkla uygulanan, kolay yorumlanan etkili bir yöntem özelliği taşımaktadır.

Kümeleme modelindeki veri sınıflarının olmayışı, kümeleme modelini sınıflandırma modelinden ayıran bir özelliktir. Kümeleme modelinde, verilerin hangi kümeye ayrılacağı veya kümelerin hangi değişkenlere göre ayrılacağı bilinmemektedir. Kümeler konunun uzmanları tarafından tahmin edilmektedir. Sınıflandırma modelinde ise, her verinin sınıfı bellidir. Yeni veri geldiğinde, verinin hangi sınıfa dâhil olacağı tahmin edilmektedir (Argüden ve Erşahin, 2008).

Tüm bilim alanlarında kümeleme yöntemi kullanılmaktadır. Tıp, biyoloji, psikoloji, sosyoloji, arkeoloji gibi belirsizlik koşullarının ve karmaşık oluşumların bulunduğu bilim alanlarında ise daha sıkça tercih edilen bir yöntemdir (Turanlı ve diğ., 2006). Marketlere gelen farklı müşteri gruplarının keşfedilmesi ve yapmış oldukları alışveriş örüntüsünün ortaya konulmasında da bu yöntem kullanılır. Dahası biyolojideki bitki ve hayvan sınıflandırılmasının yapılmasında ve bu sınıfların işlevlerine göre benzer genetik yapıdaki canlıların sınıflandırılmasında da sıkça kullanılır. Şehir planlanmasında ev tipleri, konumları ya da değerlerinin gruplandırılması için bu yöntem başarılı şekilde sonuç vermektedir. Kümeleme modelleri ayrıca web üzerinde yapılan bilgi keşiflerinde dokümanların sınıflandırılmasında da kullanılmaktadır (Özekes, 2003).

2.5.3.1. Hiyerarşik algoritmalar

Kümeleme yöntemleri içerisinde en yaygın olarak kullanılan yöntemdir. Bu yöntemde, öncelikli olarak değişken veya birim arasındaki mesafe hesabı yapılır. Daha sonra "dendrogram" adı verilen ağaç grafiği üzerinde oransal uzaklıklar gösterilir. Görsel algı, dendrogramlar yardımıyla birbirlerine yakın değişkenlerin veya birimlerin birbirlerine olan yakınlık oranları ile gruplandırılmasıyla arttırılır. Başlangıçta, her gözlem tek başına küme oluşturmaktadır. Daha sonrasında yakın olan iki küme, tek

küme altında birleştirilir. Bu şekilde küme sayısı gitgide azaltılır. Bu algoritmalar sayesinde birbirine en çok benzeyen iki nesne aynı kümede toplanır. İşlem yükü, değişken sayısı veya gözlem sayısıyla doğru orantılı olarak artmaktadır. Bundan dolayı büyük veri setlerinde çok zaman almaktadır (Demircan, 2015).

2.5.3.2. Bölümlemeli algoritmalar

Bu yöntemde, işe tüm gözlemlerin oluşturduğu küme ile başlanır. Daha sonrasında benzerlik bulunmayan gözlemlerin kümeden atılmasıyla mevcut kümenin küçülmesi sağlanır. Her gözlem tek başına küme oluşturana kadar bu işleme devam edilir. Bu algoritmalarda, kümeler arasındaki maksimum ve minimum uzaklıklar, kümelerin iç benzerlik kriterleri ve küme sayısı önceden kullanıcı tarafından belirlenir. Bu sayede daha hızlı çalışan bölümlemeli algoritmalar büyük veri tabanlarının kümelenmesi için uygun bir yöntem olarak seçilebilir (Demircan, 2015).

2.5.4. Veri madenciliğinde kullanılan teknikler

2.5.4.1. Sınıflandırma

Yeni toplanan verilerin özelliklerini incelemek ve önceden tanımlanıp, belirlenmiş bir sınıfa atamak için kullanılan bir model şeklidir. Önemli olan, sınıf özelliklerinin önceden net bir şekilde belirlenmiş olmasıdır. Sınıflandırma, kategorik özellikteki verileri tahmin eder. Buradaki amaç, önceden sınıfı belli olmayan verileri doğru sınıfa atayan modelin oluşturulmasını sağlamaktır. Sınıflandırma işlemi, kategorik değerleri tahmin etmektedir. Veri üzerinde çalışılıp, öğrenme sürecini tamamladığımız zaman sınıflama kuralları oluşturulmuş olur. Bu kurallar dikkate alınarak, belirlenmiş sınıflara yeni durum ve değerlendirmeler atanır. Sınıflandırma, günlük yaşamda sıklıkla tercih edilen bir işlemdir. Nitelikler kategorize edilerek, karar verme işlemi pratiklik kazanır (Yıldırım ve Yıldız, 2018). Sınıflandırma modellerinde örneğin, bankalardaki kredi uygulamalarının riskli ya da güvenli olmalarına göre kategorilere ayırmak için kullanılmaktadır (Özekes, 2003).

2.5.4.2. Regresyon

Regresyon analizi, bağımlı değişken ile başka değişkenler arasındaki ilişkinin matematik fonksiyon şeklinde yazılıp, bu fonksiyon yardımıyla bağımlı değişkenin

ulaşacağı değerin tahmininin sağlanmasıdır (Arı ve Önder, 2013). Daha açık bir ifade ile regresyon analizini açıklayacak olursak, bir ya da daha fazla değişkenin, başka değişkenler cinsinden tahmin edilmesine yardımcı olan ilişkilerin oluşturulmasını sağlayan analiz yöntemidir. Bu analiz, değişkenler arasındaki niteliğin türünü saptamayı amaçlar. Bir bağımlı ve bir ya da birden fazla bağımsız değişkenlerden oluşmaktadır. Bağımlı değişkenlerin; sayılıp, ölçülebilen cinsten olması tercihen daha iyi bir seçenektir. Bağımlı değişkenler "sonuç" iken, bağımsız değişkenler "girdiler" olarak adlandırılabilir. Sonucun alacağı değer, güven aralığı içinde olmalıdır. Girdiler problemin amacına göre birden fazla olabilir. Model oluşturulurken bir girdi ile oluşturulabilir. Fakat gerçek hayatta bu pek mümkün olmamaktadır. Çünkü gerçek hayatta çözülmesi gereken problemlerin doğru tahminine ulaşmak için birden fazla girdiden yararlanmak gerekmektedir. Bu noktada mühim olan girdilerin, sonucun tahmininin doğruluğuna olan katkılarıdır (Argüden ve Erşahin, 2008). Regresyon modellerinde süreklilik gösteren değerler tahmin edilir. Bu yönüyle, sınıflandırma fonksiyonlarından ayrılmaktadırlar (Özekes, 2003).

Değişkenler arasındaki ilişkinin bulunması ve varsa bu ilişkinin gücünün belirlenmesinde, ilişkinin türünün belirlenmesinde ya da ileriye yönelik değerlerin tahmininde kullanılır. Regresyon analizi ayrıca; finansal tahminlerde, zaman serisi tahminlerinde, konut emlak fiyat değerlendirmelerinde, atmosferdeki zararlı gaz oranlarının tahmininde, biyomedikal ve ilaç sektöründeki tahminlerde kullanılmaktadır (Argüden ve Erşahin, 2008).

2.6. Sosyal Medyanın Veri Madenciliğinde Kullanılması

Günümüzde internet, insanların fikir ve görüşlerini açıkça ifade edebildikleri küresel bir forma dönüşmüştür. Bu özelliği ile beraber internet hızla gelişmekte ve veri madenciliği için önemli bir bilgi kaynağı haline gelmektedir. Facebook, Twitter gibi mikroblog sitelerinde insanlar gerçek zamanlı paylaşımlar yapmaktadır. Bu siteler, kişilerin bir konu hakkında tepki ya da cevap vermesine olanak sağlamaktadır. (Çoban ve diğ., 2015). Bu kişisel bloglar birçok araştırma için geniş ve kullanışlı kaynak oluşturmaktadır. Bir ürünün olumlu ya da olumsuz yönlerine yapılan yorumlar, sosyal medya kullanıcılarının ruh halleri, toplumun politik bir konu hakkında verdiği tepkilerin tespiti gibi konu başlıkları araştırma için örnek teşkil etmektedir. Bundan

dolayı duygu analizi tekniklerinde, özellikle son yıllarda Twitter mesajları gibi sosyal medya ortamlarından elde edilen veriler sıkça kullanılmaktadır.

2.6.1. Sosyal medya kavramı

İnternet, gün geçtikçe büyüyen bilgisayar ve bilgisayar sistemlerinin birbirlerine bağlı olduğu iletişim ağıdır. İletişim teknolojilerindeki gelişmelerle birlikte, toplumsal iletişimde internet önemli bir rol almıştır (Dal ve Dal, 2014). 1990'lı yıllarından itibaren kamusal alanda kullanımı artmış olup, sadece iletişim teknolojisi alanında yenilik sağlamamıştır. Aynı zamanda akademik alanda da ilgi merkezi haline gelmiştir. İlk olarak internete olan akademik ilgi, mühendislik ve iletişim disiplinlerinde yoğun olarak başlamış, fakat daha sonrasında psikolojiden, sosyolojiye, siyaset bilimine kadar birçok disiplinde kullanılmaya başlanmıştır. İnternetin yaygın kullanımıyla birlikte popülerliği artmış, kişisel kullanıcının birçok ihtiyacına cevap vermesiyle beraber önemi daha da artmıştır (Timisi, 2003).

Gündem belirlemede, medyanın toplum üzerindeki etkisi sürekli tartışılan bir konu olmuştur. Fakat yeni medyanın gelişimiyle birlikte, saptayan ve okuyan arasındaki ilişki bir takım değişikliğe uğramıştır. 2000'li yılların başlarında kapalı uçlu seyreden internet sayfaları, sonrasında haber sitesi, blog ve mikroblog siteleri sayesinde kapalı uçlu formu değiştirerek, bireylerin medya üzerinde gündeme dahil olması sonucunu getirmiştir (Çomu ve Halaiqa, 2014). Sosyal medya etkileşimli bir medya türüdür. Geleneksek elektronik medya olan radyo ve televizyonlar bireye tek yönlü iletişim sunarken, sosyal medya çift yönlü iletişim sunmaktadır (Lee ve Cho, 2011). Sosyal medya kavramı, Web 2.0 kavramı yerine sıklıkla kullanılmaktadır. Web 2.0 ile sosyal medya eş anlamlı olarak düşünülmesine rağmen, bu durum doğru bir yaklaşım değildir. Web 2.0, çevrimiçi hizmetleri içererek, sosyal medyanın teknik yönünü göstermektedir. Ayrıca sosyal medya gibi sosyal aktivite içermemektedir (Hazar, 2011). Web 2.0 diğer kullanıcılar hakkında bilgi toplayıp, iletişime geçerek, bunlara ilişkin programları nitelendirmeye yaramaktadır. Sosyal medya ise, daha çok bu teknolojiler üstüne kurulmuş, daha geniş sosyal etkileşimi olan, işbirliği projelerinin gerçekleştirilmesini sağlayan siteler olarak tanımlanmaktadır (Hazar, 2011). Web 2.0 sayesinde geliştirilen sosyal medya, kullanıcılar için içerik yayını ve değişim olanağına imkân sunan yeni internet dalgasını tanımlamak için üretilmiş bir dönem

olarak tanımlanabilmektedir. Sosyal medyanın içine; içerik blogları, mikrobloglar, sosyal ağ siteleri, iş ağ siteleri ve iş birliği ile oluşturulmuş siteler girmektedir (Kwon ve Sung, 2011). Sosyal medya, içindeki teknolojiyi vurgulamak yerine, içerik üretimi ve aktif sosyal rolleri vurgulamaktadır. Sosyal medyada hedef kitlesi olarak TV, radyo, dergi ve gazete de vardır. Fakat sosyal medyanın kitlesindeki en önemli fark; insanların kendilerine oluşturduğu veya başkalarından kopyaladığı içerikleri paylaşmayı sevmeleridir. Bu fark en büyük değişiklik olarak gösterilebilir. İnsanlar kendi başlarına içerik oluşturabilir veya başka bir yerden içerik kopyalayabilirler (Hazar, 2011). Kullanıcıların paylaşım yapmasına, yapılan paylaşımların yorumlanabilmesine olanak sağlayan sosyal medya dört kategoride incelenebilir. Bunlar; sosyal ağ siteleri, içerik topluluğu, bloglar ve mikrobloglardır.

2.6.2. Sosyal ağ siteleri

Sosyal bir çevre oluşturmak için kurulmuş olan sosyal ağ siteleri, büyük kitlelerin birbirleriyle iletişime geçebildikleri, birbirlerinden etkilendikleri elektronik ortamlardır. Sosyal ağlar, kişilere kendilerine ait bir alan oluşturmasına müsaade etmektedir. Bu sayede kullanıcılar bir araya gelerek, dijital ortamda video, fotoğraf, haber vs. materyaller paylaşabilmektedirler. Böylece insanların birbirleriyle etkileşime girmesine imkân sunulmaktadır (Kara ve Coşkun, 2012). Sosyal ağ siteleri, kişilerin sınırlandırılmış bir sistem üzerinde kamuya açık profiller inşa etmelerine olanak tanır (Boyd ve Ellison, 2008). Kullanıcı profil bilgilerinde doğum tarihi, cinsiyet, siyasi görüş, inanç ve doğum yeri bilgilerinden, en sevilen filmlere, kitaplara, sevdiği müzik tarzlarına ve boş zamanlarda neler yapıldığına kadar oldukça geniş bilgiler yer alabilmektedir (Akar, 2010).

Kişilerin başka kişilerle görüşmesine imkân vermesi haricinde aynı zamanda kendi sosyal ağını kurup, fikirlerini rahatlıkla ifade etmesini sağlaması, sosyal ağ sitelerini, diğer medya platformlarından ayıran bir özelliktir. Sosyal ağ siteleri zaman geçtikçe daha fazla hayatımıza girmektedir. Genç nüfus başta olmak üzere gelişmiş ve gelişmekte olan ülkelerde kişilerin sosyal medya karşısında geçirdikleri sürenin ve kullanıcı sayısının gün geçtikçe daha da artmasıyla yeni bir yaşam şekline yönelişin olduğu görülmektedir.

2.6.3. İçerik topluluğu

Videolar oluşturulup video dosyalarının yüklenmesiyle, kullanıcıların paylaşım ve yorum yapması sağlanır. Örneğin; Youtube vs.

2.6.4. Bloglar

İstenilen alanda yorum, paylaşım yapılıp, fikirlerin yazılabildiği çevrim içi günlüklerdir. Bloglar, kendilerine ait tarzları ile diğer internet sitelerinden ayrılmaktadır. Bloglar, kişinin kendi kod becerisiyle oluşturulup kullanılabileceği gibi, hazır tasarım halinde de tercih edilip kullanılabilir. Birçok blog türü mevcuttur. Amacı bilgi vermek olan bloglar, genelde bilgiye dayalı olup, tasarım bakımından ayrıntılı değillerdir. Kişisel bloglar ise genelde sağlık, hayata dair konuların olduğu, giyim, moda ya da spor ile ilgili değişik birçok formatta olabilir. Özel amaçlara yönelik de blog sayfaları oluşturulup kullanılabilmektedir. İnsanlar kendilerini ifade etmek, belli konularda fikirlerini paylaşmak amacıyla blog oluşturup kullanabilirler. Örneğin; kişisel blog sayfaları, blog yazarının kendi ismini taşıdığı, genelde bireysel özelliklerin ön planda tutulduğu sayfalardır. Genelde en çok tercih edilen blog sayfası çeşididir. Diğer bir çeşit olan temasal blog sayfaları, uzmanlık gerektirmektedir. Moda, ekonomi, siyaset, tarih gibi birçok alanda uzman kişilerce belli konular üzerinde hazırlanmaktadır. Kurumsal blog sayfaları ise, kurumların ya da firmaların kendileri hakkında bilgi vermek, tanıtım yapmak amacıyla tasarlanmışlardır. Örneğin; www2.mdanderson.org/, cancerwise/ vs. (URL-1, 2019).

2.6.5. Mikrobloglar

Blogların uzunluk kısıtlaması getirilmiş halidir. Mikrobloglar bugün internet kullanıcıları arasında çok popüler bir iletişim aracı haline gelmiştir. Milyonlarca kullanıcı her gün yaşamın farklı yönleriyle ilgili görüşlerini bu platformlarda paylaşmaktadır. Twitter, Tumblr gibi mikroblog hizmeti veren popüler web sitelerinde her gün milyonlarca mesaj yer almaktadır. İnternet yazarları, hayatları hakkında yazıp, çeşitli konularda görüşlerini paylaşarak, güncel konular hakkında tartışabilmektedirler. Mikroblog platformlarının kolay erişimi ve serbest mesaj biçiminden ötürü, internet kullanıcıları geleneksel iletişim araçlarından (geleneksel bloglar veya postalama gibi) mikroblog hizmetlerine kaymaktadır. Gittikçe daha fazla

kullanıcı, kullandıkları ürünler ve hizmetler hakkında mesaj gönderdikleri için veya politik ya da dini görüşlerini ifade ettikleri için, mikroblog siteleri aracılığıyla, insanların görüş ve duyguları değerli kaynaklar haline dönüşmektedir. Bu veriler pazarlama veya sosyal çalışmalar için verimli bir şekilde kullanılabilir. Bu nedenle, mikroblog web siteleri, fikir madenciliği ve duyarlılık analizi için zengin veri kaynaklarıdır. Bu kaynaklardan elde edilen veriler, fikir madenciliği ve duyarlılık analiz çalışmalarında etkin şekilde kullanılmaktadır. Siyasi partiler, programlarını destekleyip desteklemediklerini bilmek isteyebilirler veya sosyal organizasyonlar, insanlardan mevcut tartışmalar hakkında görüşlerini isteyebilirler. Kullanıcıların her gün sevdiklerini veya beğenmediklerini ve hayatlarının birçok yönüyle ilgili görüşlerini yayınladıkları için, tüm bu bilgiler mikroblog hizmetlerinden elde edilebilir (Pak ve Paroubek, 2010).

2.6.5.1. Twitter

Twitter, tweet olarak adlandırılan ve en fazla 140 karakterden oluşan kullanıcıların herhangi bir konu hakkındaki fikirlerini paylaştığı yaygın olarak kullandıkları sosyal paylaşım ağlarından birisidir. Bu tweetler kullanıcının duygu ve düşüncelerini içermektedir. Twitter bir mikroblog veya sosyal ağ sitesi olarak tanımlanabilir. Merkezi bir faaliyet ile web ya da bir cep telefonu aracılığıyla kısa durum güncelleme mesajları (tweetler) yayınlanması yönüyle mikroblog kategorisinde değerlendirilebilir. Sosyal ağ siteleri kategorisinde değerlendirilebilmesinin nedeni ise, üyelerin kişisel bilgilerini içeren profil sayfalarının bulunmasıdır (Adak Kaplan, 2016). Bu profil sayfalarını diğer kullanıcılar görebilmektedir. Böylece içeriklere rahatça erişim imkânı sunulmaktadır. Twitter ayrıca bilgi paylaşımı, günlük aktiviteleri tanımlama, ya da devlet kurumları tarafından bilgi yayılması amacıyla kullanılabilmektedir.

Twitter'ın özelliklerinden birisi "retweet'tir". Yani yayınlamış tweetin tekrar kendi sayfamızda yayımlanmasıdır. Böylece başkasının tweetini sayfamızda yayımlanmasıyla birlikte kendi takipçilerimiz de rahatça görmüş olacaktır. Retweet sayesinde kendi takipçimiz olmasa dahi başka kişilerin takipçilerine de ulaşmak daha kolay olacaktır. Twitter'ın bir başka iletişim özelliği de "hashtag'dir". Hashtag adı verilen "#" işaret; belli bir konudan bahsedilirken o konunun daha rahat aranmasını

sağlayan, bahsedilen konudaki tweeti etiketlemek anlamına gelmektedir. Yani bir konu hakkındaki konuşmayı kolaylaştırmak için kullanılır. Örneğin tweet yazarken #AnahtarSözcük diye girildiğinde Twitter "#" işareti konulduğundan dolayı bu anahtar sözcüğe link koyar, bu linke tıklandığında dünya üzerindeki bu anahtar sözcüğü içeren bütün tweetler zaman sırasına göre sıralanır. Hashtag oluşturulduğunda bütün kullanıcılar o konu hakkında fikirlerini belirtebilirler. Buradaki amaç; tweetleri kategorilere ayırarak arama sisteminde istenilen kelimelerin daha rahatça bulunmasını sağlamaktır. Sosyal medyada istenilen kelime başına hashtag işareti konulduğu zaman o kelime ile alakalı tüm sonuçlara ulaşılabilir. Bu özellik sayesinde, daha önceden o kelimeyle alakalı kimlerin yazı yazdığı veya fotoğraf paylaştığı görülebilmektedir. Bu özellik sayesinde, sosyal medya üzerindeki iletişim kolaylaştığı için yeni bir iletişim aracı haline dönüşmüştür. Hashtag belli bir konuda fikirleri bulmaya yardımcı olduğu gibi, aynı fikirdeki insanları bir araya toplamak için de kullanılabilmektedir. Buna karşılık, "@" sembolü kullanılarak gönderilen mesajlar "mention" olarak adlandırılır. Bu kullanım; mesajların, yalnızca @ işaretinin yanında yazılan kişi ya da kişilere gönderilmesini sağlar. Kısaca bu kullanım hedefli kişilere tweet atmayı sağlamaktadır (Adak Kaplan, 2016).

Özet olarak, Twitter sosyal amaçlar için kullanılmasına rağmen, kişisel bilgiler de dahil olmak üzere çeşitli türde bilgilerin yayılması için önemli bir araç olarak kullanılabilmektedir. Bu yüzden, kullanıcılar tarafından gönderilen tweetlerin analizlerini yapmak mantıklıdır. Twitter'da tweetler aracılığı ile dünya çapında sohbetler gerçekleşmektedir. Kullanıcılar için daha iyi çözümler oluşturulması, tweetlerden gelen verilerin kilidinin açılması için, uç noktalar bulunmalıdır. Bu uç noktalar, tweetlerin yönetilmesini, yayınlanmasını, tweet konularının filtrelenmesini ve aranmasını sağlamaktadır.

Twitter, tweetlere sorgu terimine göre programlı olarak erişmek için bir Uygulama Programlama Arayüzüne (Application Programming Interfaces - API) sahiptir. Bu arayüz, veri tabanları ya da yazılımlar arasında iletişimi sağlamaya yarayan yapılar olarak karşımıza çıkmaktadır. Bu arayüz kullanılarak, Twitter kullanılmasına gerek olmadan mikroblog sitesinde yayınlanan tweet paylaşımı veya paylaşılmış olan tweetlerin çekilmesi, yazılımlar sayesinde gerçekleştirilmektedir. Twitter API'si tweetlere programlı bir şekilde okuma ve yazma yapılmasını sağlamaktadır. Twitter'ın

Rest API'lerine ve Streaming API'lerine ulaşılabilmesi için kullanılan, Python ile yazılmış bir paket mevcuttur. Bu API'ler yazılımcılar tarafından Twitter'ın çekirdek verilerine erişmek için kullanılabilir. Çekirdek verilerden kasıt kullanıcı bilgileri ve kullanıcıların takipçilerinin bilgileridir (Adak Kaplan, 2016). Ayrıca, Twitter üzerinde kullanıcı işlemlerinden olan çoklu paylaşımda bulunmak, API'ler vasıtasıyla otomatik bir şekilde gerçekleştirilebilir. Twitter API'sinde, tweetlerin alınacağı dili belirten ve sorgulama sıklığının belirtilmesine izin veren bir parametresi vardır (Go ve diğ., 2009). Twitter API platformu, üç aşamalı arama API'si sunmaktadır:

Standart; bu arama API'si, son 7 gün içinde yayınlanan en son Tweet örneklemesine göre arama yapar. 'Public'API kümesinin bir parçasıdır.

Premium; son 30 günlük tweetlere ücretsiz veya 2006'nın başlarından itibaren tweetlere ücretli erişim imkânı sunmaktadır. Kurumsal veri API'lerinin güvenilirliği ve eksiksizliği sayesinde, uygulama veya işletme büyüdükçe erişim yükseltme fırsatı vermektedir.

Kurumsal; ücretli (ve yönetilen) son 30 günlük tweete erişim veya 2006'nın başlarından itibaren tweetlere erişim imkânı sunmaktadır. Tam tutarlılık verileri, doğrudan hesap yönetimi desteği ve entegrasyon stratejisine yardımcı olmak için özel teknik destek sağlamaktadır.

Twitter API'sini kullanarak haberleşmenin sağlanabilmesi ve yetkinin alınabilmesi için bir key alınması gerekmektedir. Yetkilendirmede ise OAuth kullanılmaktadır. OAuth, kullanıcılardan şifrelerini paylaşmasına gerek duymadan masaüstü, web ya da mobil uygulamalarında kullanılabilecek, bir yetkilendirme, yani kimlik doğrulama protokolüdür. Bu protokol, 2010 yılından bu yana bütün Twitter uygulamaları için zorunludur.

2.6.6. Gerçek Zamanlı powertrack API Tarafından Sağlanan Özellikler

Kurumsal lisanslı erişim teklifi olarak gerçek zamanlı PowerTrack API; dinamik filtreleme, tutarlı bağlantı, veri kurtarma ve veri uygunluk yönetimi için araçlar içermektedir. Doğrudan servislere bağlanılarak bu özelliklerden yararlanılmaktadır.

Bu teknoloji sayesinde, işletmeler müşterilerine hizmet etmesi için sağlam bir temel hazırlamaktadır. Bu özelliklerden bazıları aşağıda sıralanmıştır (URL-2, 2018):

- Akışın kural seti güncellenirken, akışın bağlantısının kesilmesine gerek yoktur.
- Herhangi bir aksilik durumunda otomatik kurtarma devreye girmektedir. Geciken zaman içinde kaçırılan veriler kurtarılmaktadır.
- Beş günlük bir geçmişe ait veri penceresinin kurtarılmasını sağlayan Replay akışları mevcuttur.
- Test ve geliştirme için ek akışlar bulunmaktadır.
- Ayrıca, operasyonel konular hakkında müşterilerle iletişim kurmak için durum panosu bulunmaktadır.

2.6.7. Standart akış API istek parametreleri

Streaming API bitiş noktaları tarafından hangi verilerin döndürüleceğini tanımlamak için aşağıdaki istek parametreleri kullanılmaktadır (URL-2, 2018):

1. delimite
2. stall_warnings
3. filter_level
4. language
5. follow
6. track
7. locations
8. count
9. with (deprecated)
10. replies (deprecated)
11. stringify_friend_id (deprecated)

2.6.7.1. Delimited

Durumların akışlarda sınırlandırılması gerektiğini göstermek, dize uzunluğunu ayarlamak için kullanılmaktadır. Böylece istemciler durum iletisinin sonundan önce kaç bayt okuması gerektiğini bilmektedir (URL-2, 2018).

2.6.7.2. Stall_Warnings

Bu parametre true değerine ayarlandığı zaman, müşterinin bağlantısı kesilse dahi periyodik mesajlar iletilmeye devam edecektir. Bu parametre, yüksek bant genişliğine sahip bağlantıları olan müşteriler için en uygun olanıdır (URL-2, 2018).

2.6.7.3. Filter_Level

Bu parametrenin yok, düşük veya orta değerlerinden birine ayarlanması, akışa dahil edilmesi gereken filter_level tweet niteliğinin minimum değerini belirler. Varsayılan değer, mevcut tüm tweetleri içermektedir (URL-2, 2018).

2.6.7.4. Language

Bu parametrenin, Twitter'ın gelişmiş arama sayfasında listelenen dillerden herhangi birine karşılık gelen virgülle ayrılmış BCP 47 dil tanımlayıcıları listesine ayarlanması, yalnızca belirtilen dillerde yazılmış olduğu tespit edilen tweetleri döndürür. Örneğin, language = en ile bağlanmak yalnızca İngilizce dilinde olduğu tespit edilen tweetleri yayınlayacaktır (URL-2, 2018).

2.6.7.5. Follow

Tweetlerin akışta yayınlanması gereken kullanıcıları belirtir. Belirtilen her kullanıcı için, akışın içeriği aşağıdaki gibidir (URL-2, 2018).

- Kullanıcı tarafından oluşturulan tweetler.
- Kullanıcı tarafından retweetlenen tweetler.
- Kullanıcı tarafından oluşturulan herhangi bir tweetin yanıtı.

2.6.7.6. Track

Akışta hangi tweetlerin yayınlanacağını belirlemek için kullanılacak virgülle ayrılmış kelime öbeği listesidir (URL-2, 2018).

2.6.7.7. Location

Tweetleri filtrelemek üzere bir dizi sınırlama kutusu belirten, enlem ve boylam çifti listesidir. Kısacası, yalnızca istenen sınırlama kutularına giren coğrafi konumlu tweetler dâhil edilmektedir. Tablo 2.1'de örnek olarak gösterilmiştir (URL-2, 2018).

Tablo 2.1. Location listesine örnek gösterim

Parametre Değeri	**Tweetlerin Geldiği Yer**
-122.75,36.8,-121.75,37.8	San Francisco
-74.40,-73.41	New York City
-122.75,36.8,-121.75,37.8,-74,40,-73,41	San FranciscoOR New York City

Akış API'si, belirli bir tweetin sınırlayıcı bir kutuya girip girmediğini belirlemek için aşağıdaki buluşsal yöntemlerden birini kullanır:

Koordinatlar alanı doldurulursa; orada bulunan değerler sınırlayıcı kutuya karşı test edilecektir. Bu alan geoJSON sırasını (boylam, enlem) kullanmaktadır.

Koordinatlar boş ancak yer doldurulursa; yerinde tanımlanmış olan bölge, sınırlama kutusundaki konumlarla kesişim açısından kontrol edilir. Herhangi bir çakışma ile eşleşmektedir.

Bu kurallardan hiçbiri uymazsa; coğrafi alan kullanım dışıdır ve akış API'sı tarafından göz ardı edilmiştir. Eğer yalnızca sınırlayıcı kutunun içine giren yerlerin dahil olması istenirse, kod için ek bir filtreleme adımı gerçekleştirilmesi gerekmektedir (URL-2, 2018).

2.6.7.8. Count

Bu parametrenin kullanımı yüksek erişim gerektirmektedir. Akışlı bir uç noktaya bağlanırken; yükseltilmiş erişim istemcileri, bağlantı kesme süresi içinde meydana gelen cevapsız mesajları doldurmaya çalışmak için count parametresini kullanmaktadır (URL-2, 2018).

2.6.7.9. With (deprecated)

With parametresi, Kullanıcı ve Site Akışı istemcilerine gönderilen mesaj türlerini kontrol etmektedir (URL-2, 2018).

2.6.7.10. Replies

Kullanıcı ve site akışlarında kullanılmaktadır. Varsayılan olarak "@"cevaplar sadece mevcut kullanıcı cevabın hem göndereni hem de alıcısını takip ederse gönderilir. Örneğin, Ali'nin Buse'yi takip ettiği, ancak Ali'nin Can'ı takip etmediği varsayılsın. Buse, @Can şeklinde yanıtlarsa, Ali tweeti göremez. Bu Twitter.com ve api.Twitter.com davranışını taklit etmektedir. Bu tür tweetlerin bir akış bağlantısında geri döndürülmesini sağlamak için, bağlanırken yanıtları = tümü belirtilmesi gerekmektedir (URL-2, 2018).

2.6.7.11. Stringify_Friend_ids

Varsayılan olarak, kullanıcı ve site akışları, arkadaş listesi girişini bir tamsayı dizisi olarak göndermektedir (stringify_friend_ids = false değerine eşdeğer).

2.6.8. Tweetleri bölgelere göre filtreleme

Tweet verileriyle çalışırken, iki coğrafi meta veri sınıfı vardır:

- Tweet location (tweet konumu): Kullanıcı, tweet sırasında konumu paylaştığında kullanılabilir.
- Account Location (Hesap Konumu): Kullanıcı tarafından herkese açık profillerinde verilen "ev" konumuna göredir. Bu, serbest biçimli bir karakter alanıdır ve coğrafi olarak yönlendirilebilecek meta veriler içerebilir veya içermeyebilir.

Coğrafi koordinatlar [LONG, LAT] sırasına göre verilmiştir. Bunun tek istisnası, [LAT, LONG] sırasına göre verilen 'coğrafi' özniteliğidir (URL-2, 2018).

2.6.8.1. Tweet konumları ("coğrafi etiketli" tweetler)

Twitter, kullanıcıların bireysel tweetler için konum belirlemesini sağlar. PowerTrack, çeşitli operatörleri aracılığıyla tweete özgü konum verilerine göre tweetleri filtrelemek için birden fazla yol sunar. Tweete özgü konum bilgileri iki genel kategoriye ayrılır:

- Belirli bir enlem / boylam "Nokta" koordinatına sahip tweetler.
- Twitter “Place” adlı tweetler

“Nokta” koordinatlı tweetler, GPS özellikli cihazlardan gelir ve söz konusu tweetin tam GPS konumunu temsil etmektedir. Bu konum türü, tam olarak bir Twitter yeri ile ilişkilendirilemezse, referans alınan GPS konumuyla ilgili (şehir, ülke vb.) bağlamsal bilgi içermemektedir.

Twitter “Place” olan tweetler, kullanıcının tweeti gönderdiği genel alanı (“yer”) tanımlayan yalnızca 4 koordinattan oluşan bir çokgen içermektedir. Ek olarak, yerin diğer alanların yanı sıra yerin bulunduğu ülkeye karşılık gelen bir görünen adı, türü ve ülke kodu da olacaktır (Örneğin; şehir, mahalle) (URL-2, 2018).

2.6.8.2. Twitter yerinde JSON

Aşağıda "Boulder, CO" Twitter Place ile etiketlenmiş bir tweetten JSON örneği Şekil 2.5’de verilmiştir.

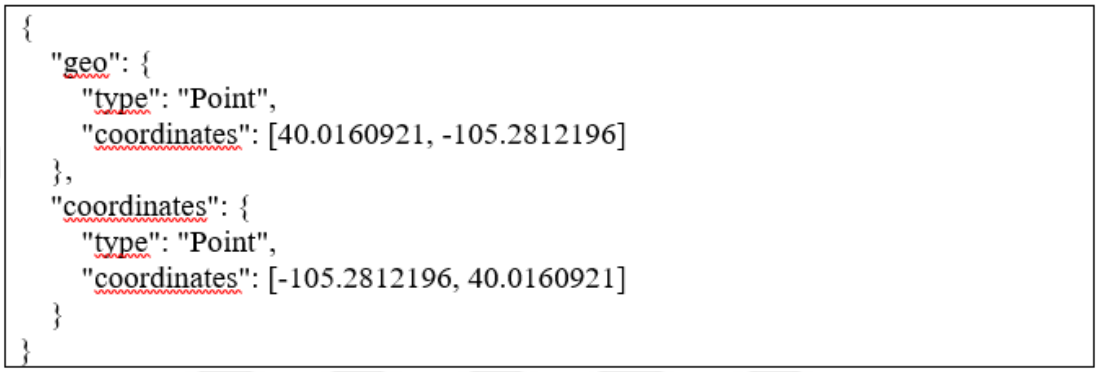

```
{
  "place": {
   "id": "fd70c22040963ac7",
   "url": "https:\/\/api.Twitter.com\/1.1\/geo\/id\/fd70c22040963ac7.json",
   "place_type": "sehir",
   "name": "Ankara",
   "full_name": "Ankara, CO",
   "country_code": "TR",
   "country": "TURKIYE",
   "contained_within": [    ],
   "bounding_box": {
    "type": "Polygon",
    "coordinates": [       [
      [-105.301758, 39.964069],
      [-105.301758, 40.094551],
      [-105.178142, 40.094551],
      [-105.178142, 39.964069]
     ]
    ]
   },
   "attributes": {}
  }
}
```


Şekil 2.5. JSON kod örneği

2.6.8.3. Tam konum JSON

Twitter'ın, kök düzey "geo" ve "koordinat" özellikleri sayesinde, tam konum için ondalık derece koordinatlarını sağlamaktadır.

"Koordinatlar" niteliklerinin [LONGITUDE, enlem], "geo" niteliğinin [enlem, LONGITUDE] olarak biçimlendirildiğine dikkat edilmesi gerekmektedir. Kod örneği Şekil 2.6'da gösterilmiştir.



```
{
  "geo": {
    "type": "Point",
    "coordinates": [40.0160921, -105.2812196]
  },
  "coordinates": {
    "type": "Point",
    "coordinates": [-105.2812196, 40.0160921]
  }
}
```

Şekil 2.6. Tam konum JSON kod örneği

2.6.9. Tweet konum operatörleri

2.6.9.1. Place

Belirli yerlerin, isim ya da kimlikleri kullanılarak filtreleme yapılmaktadır. Belirli bir alanla ilişkili "yerler" i keşfetmek için, REST API'sinde Twitter'ın reverse_geocode bitiş noktası kullanılmaktadır. Daha sonra başvurulan belirli bir yeri içeren tweetleri izlemek için, place: operatör ile bulunan Place ID'leri kullanılmaktadır (URL-2, 2018).

2.6.9.2. Place_contains

Bu operatör sayesinde belirli yerler için eşleşme yerine, yer adlarına karşı bir alt dize eşleşmesi gerçekleştirilir.

2.6.9.3. Place_country

Twitter'daki her tweet "Place", Yerin bulunduğu ülkeyi belirten bir ülke kodu ile birlikte gelir.

2.6.9.4. Has:geo

Has:geo operatörü, Twitter yükü içinde Point veya Place geo bilgilerinin varlığıyla eşleşmektedir. Bu operatör, belli konum ya da coğrafi veri türlerinin belirtilmesine izin vermemektedir.

2.6.9.5. Point_radius

Bu operatör, dairesel bir coğrafi bölge belirlenip, bu alana giren özel konum verilerini içeren tweetlerin eşleştirilmesini sağlamaktadır. Kullanmak için merkezi bir lon-lat koordinatı tanımlanarak, yarıçapı ayarlanması gerekmektedir (en fazla 25 mil). Bu bölgeye giren coğrafi bir nokta içeren herhangi bir tweet eşleştirilecektir. Ek olarak, Twitter Yerleri içeren tweetler, Yer için tanımlanan coğrafi poligonun tanımlanmış nokta yarıçapı alanına tam olarak düştüğü yerde eşleşir. Poligonları tanımlanan nokta yarıçapı alanının dışına düşen yerler uyuşmaz (URL-2, 2018).

2.6.9.6. Bounding_box

Bu operatör, 4 taraflı bir coğrafi bölge belirlenmesini ve bu alana giren tweete özgü konum verilerini içeren tweetlerin eşleştirilmesini sağlamaktadır. Kullanmak için, kutunun her iki tarafının uzunluğu 25 mil olacak şekilde kutunun karşıt köşelerini temsil eden yalnız koordinatlarla tanımlanır. Bu bölgeye giren coğrafi bir nokta içeren herhangi bir tweet eşleştirilecektir. Ek olarak, Twitter yerleri içeren tweetler, yer için tanımlanan coğrafi poligonun tanımlanmış nokta yarıçapı alanına tam olarak düştüğü yerde eşleşmektedir. Poligonları tanımlanan nokta yarıçapı alanının dışına düşen yerler uyuşmaz (URL-2, 2018).

Kullanımı; bounding_box: [west_long south_lat east_long north_lat] şeklindedir.

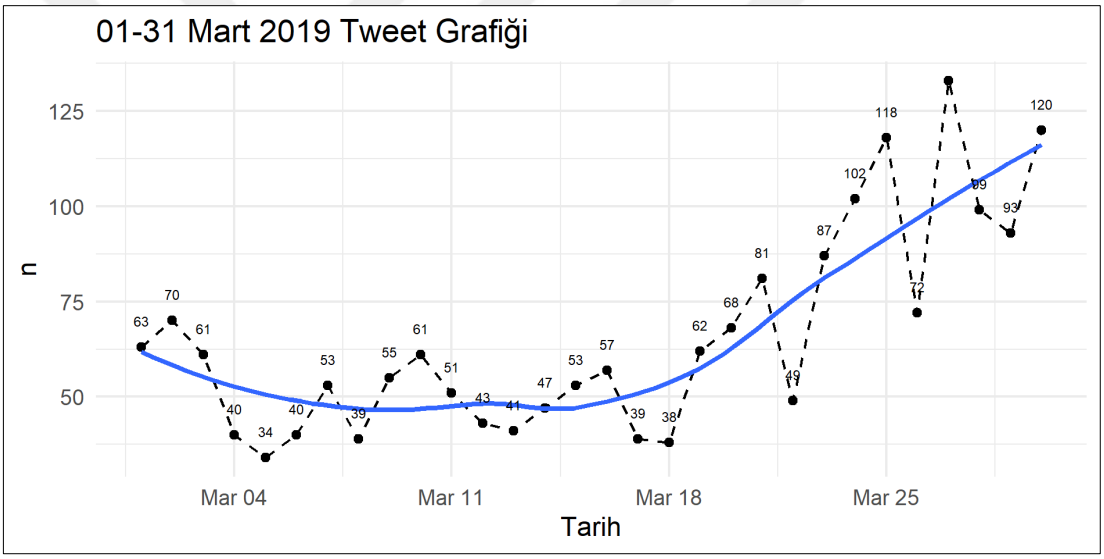
- west_long south_lat sınır noktasının güneybatı köşesini, batı uzunluğunun bu noktanın boylamı, south_lat ise enlem olduğunu göstermektedir.
- east_long ve north_lat sınırlayıcı kutunun kuzeydoğu köşesini temsil etmekte; burada east_long bu noktanın boylamı ve north_lat enlemi olarak ifade edilmektedir.
- Sınırlama kutusunun genişliği ve yüksekliği 25mil'den az olmalıdır.
- Boylam ± 180 Aralığında olmalıdır.
- Enlem ± 90 Aralığında olmalıdır.

- Tüm koordinatlar ondalık derecedir.
- Kural argümanları, boşlukla ayrılmış, parantez içinde bulunmalıdır.

Not: Yerinde eşleşen operatörler (Tweet geo) yalnızca orijinal tweetlerden gelen eşleşmeleri içerecektir. Retweetler herhangi bir yer verisi içermemektedir (URL-2, 2018).

3. UYGULAMA

Çalışmamızda, 31 Mart 2019 da yapılan yerel seçim sonuçları, Twitter verileri üzerinden tahmin edilmeye çalışılmıştır. Bunun için ilk olarak 01.03.2019 ile 31.03.2019 tarihleri arasında, Python dilinde yazılmış bir program ile otuz gün boyunca, günlük olarak toplamda 1968 adet ham tweet toplanmıştır. Ham tweetler ilk olarak Json formatındayken, verilerin kolay işlenebilmesi için veri formatı xls formatına çevrilmiştir. Ham veriler toplanırken ayrıca günlük olarak da gruplandırılmıştır. Otuz günlük tweet dağılımı ve eğilimi Şekil 3.1' de verilmiştir.



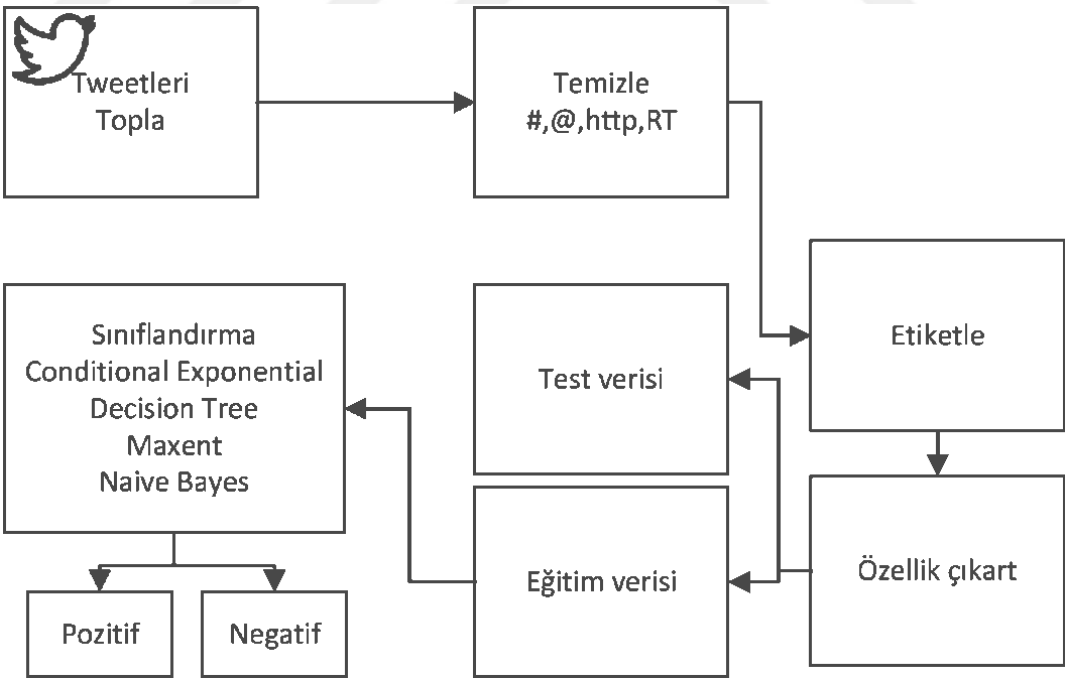
Şekil 3.1. 2019-03-01 ile 2019-03-31 tarihleri arasında ham tweet dağılımı ve eğilim

Seçim sonucu tahmini tweetler üzerinden yapılması amacıyla, arama yapmak için anahtar kelimeler kullanılmıştır. Kullanılan anahtar kelimeler Tablo 3.1'de verilmiştir. Anahtar kelimeler seçilirken, ilgisiz tweetlerin önüne geçmek ve verilere karışmasını önlemek amacıyla kelimeler, seçime giren parti adaylarının Twitter hesaplarından seçilmiştir. Bu şekilde bir miktar faydalı tweetlere ulaşılamamış olmasına rağmen, verileri bozacak ilgisiz çok sayıda tweet de filtrelenmiştir. Anahtar kelime seçiminde, birçok deneme süreci gerçekleştirildikten sonra, bu şekilde daha uygun olacağı kararına varılmıştır. Bu ham tweetler gruplandırıldığı zaman, AK Parti 878, İyi Parti 282, Saadet Partisi 208 adet, tarafsız 600 adet, toplamda ise 1968 adet tweet

toplanmıştır. Toplanan bu veriler temizlenip analize hazır hale getirilmiştir. Hazır olan veriler üzerinde daha sonra "Duygu Analizi" yapılmıştır. Duygu analizi sayesinde, toplanan tweetler olumlu ve olumsuz olarak sınıflandırılmıştır. Sınıflandırma işleminde yöntem olarak Naive Bayes ve Destek Vektör Makinesi yöntemleri kullanılmıştır. Bu sınıflandırma işlemi Ak Parti ve İyi Parti için yapılmıştır. 31 Mart 2019 yerel seçimlerinde CH ile İyi Parti, diğer taraftan ise Ak Parti ile MHP ittifak kurduğu için, CHP ve MHP ile ilgili tweetler ayrıca bir başlık olarak gösterilmemiştir. Çalışma lokal seviyede, Kocaeli ili bazında, yapıldığı ve seçime katılan diğer parti adaylarına yeteri sayıda tweet atılmadığı için, değerlendirmeye alınmamıştır.

3.1. İşlem Adımları

Bu tez kapsamında yapılan işlem adımları grafik olarak Şekil 3.2.'de gösterilmiştir. Bu işlem adımlarının tamamı Python dilinde yazılan kodlarla yapılmıştır. Duygu analizi yapılırken ise, açık kaynak kodlu Natural Language Tool Kit (NLTK) kütüphanesi kullanılmıştır. Sınıflandırma başarı oranları ise WEKA paket programında değerlendirilmiştir.



Şekil 3.2. Duygu analizi adımları

3.1.1. Twitter verileri toplama:

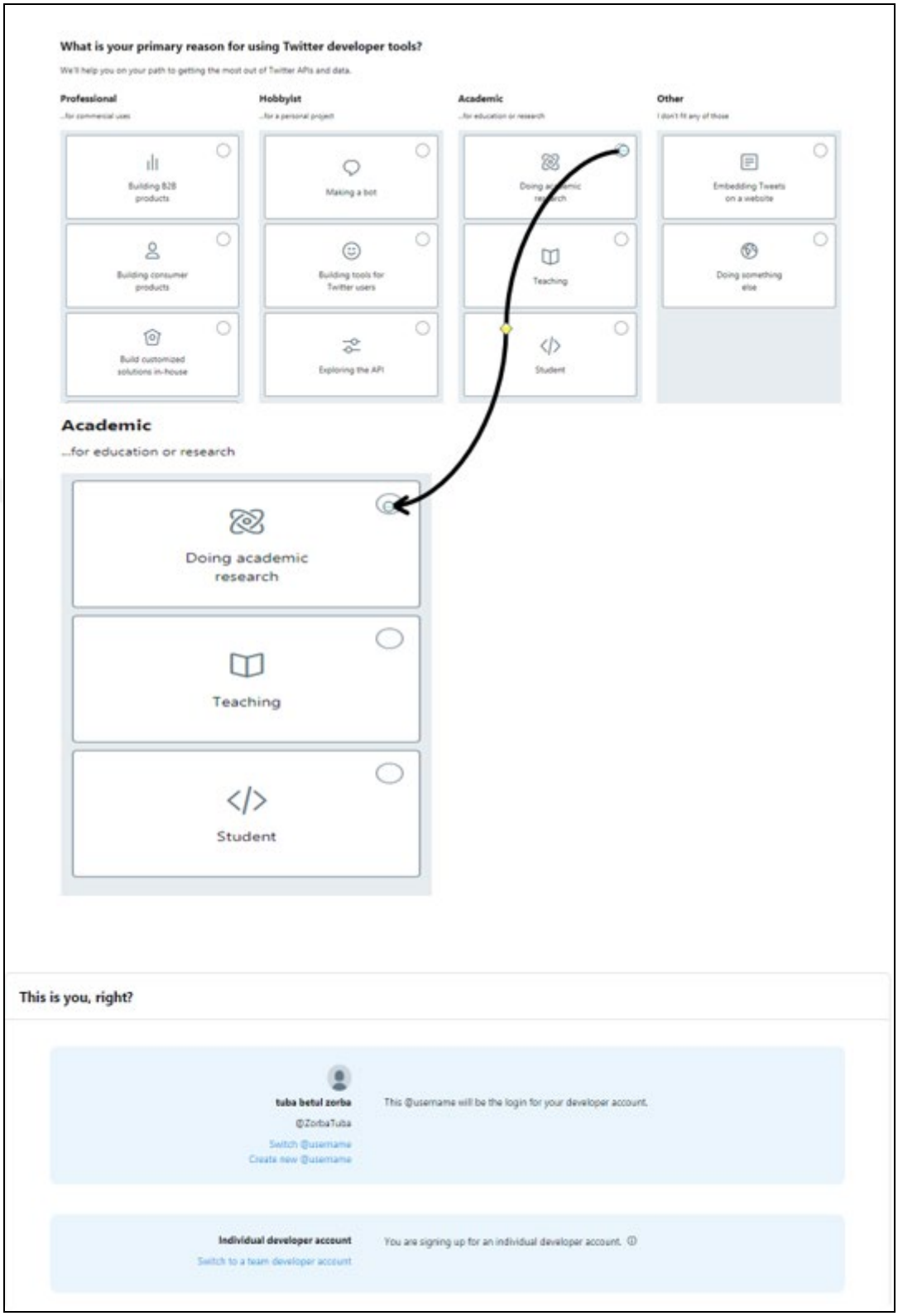
Twitter ile veri toplama üç adımda gerçekleştirilmiştir.

A) Twitter hesabı https://twitter.com/i/flow/signup açılmıştır.
B) https://developer.twitter.com/ sayfasında form doldurulması gerekmektedir. Form doldurulduktan sonra Twitter API'lere ulaşmak için izin alınması gereklidir. Gerekli izin alınması için, Twitter API'lerine ulaşım sağlamamızın nedenlerini sorgulayan bir yazışma yapılmıştır. Bu yazışma adımları, Şekil 3.3'te detaylı şekilde gösterilmiştir. Twitter hesabı kullanılarak, Geliştirici Erişimi için başvuru yapılmıştır. Daha sonra Twitter'a Python programından erişim sağlamak için kullanılacak olan API kimlik bilgileri oluşturulmuştur. Bu kimlik bilgilerinin oluşturulması için ise, Tweepy kütüphanesinin API'leri kullanılarak Python'da program yazılmıştır. Twitter hesabımıza ulaşmamızı sağlayan kod parçası Şekil 3.3'de gösterilmiştir.

```
import tweepy as tw
import pandas as pd
consumer_key        = "XzMEXXXT1cXZHJWk10phsXXXf"
consumer_secret     =
"BmrR2UGhTF7Y90ZwOWw5GealaR9XhCYxq3hCiPpZ2EBSXXXX41"
access_token        = "193206391-
BeuFPyhYtiuc5pggwsehg41pT10XXXMY6L3h25Xm"
access_token_secret=
"EqLK1eqEkU99rwkwhekcm6XXXpzCEGrZFh59XXct44S80"
auth = tw.OAuthHandler(consumer_key, consumer_secret)
auth.set_access_token(access_token, access_token_secret)
api = tw.API(auth, wait_on_rate_limit=True)
```

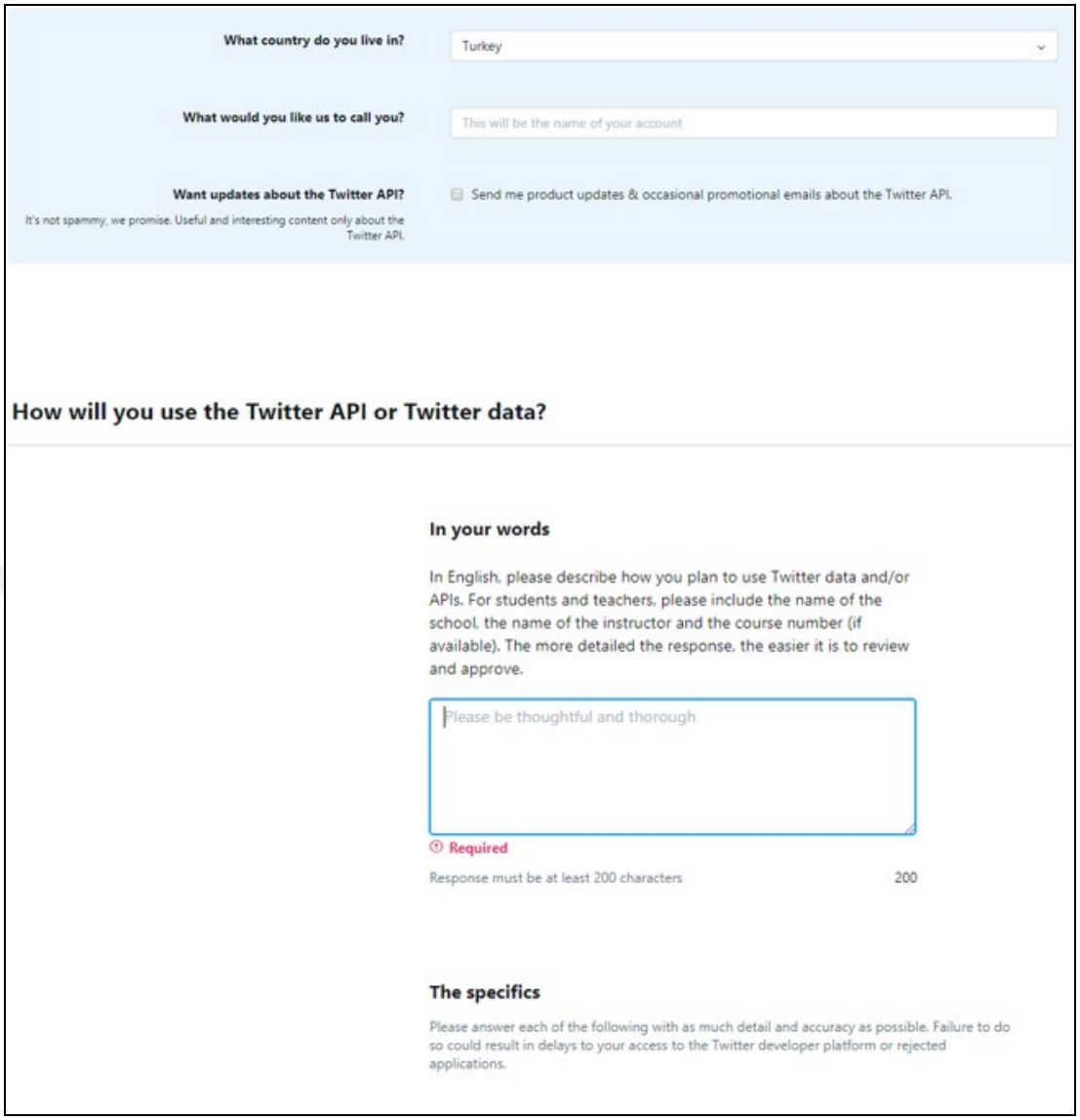
Şekil 3.3. Twitter hesabına ulaşımı sağlayan kod parçası

Şekil 3.4'te ne amaçla API isteğinde bulunulduğu sorusunu yöneltip, seçeneklerden yapılacak işleme yönelik seçim yapılması istenmektedir. Çalışmamızda akademik çalışma seçeneğini kullanılmıştır.



Şekil 3.4. Twitter API'lerine ulaşmak için izlenen işlem adımları

Twitter'a kayıt olunan profil bilgilerinin doğruluğu kontrol edilmektedir.



Şekil 3.4. (devam) Twitter API'lerine ulaşmak için izlenen işlem adımları

Son olarak API isteğinin sebeplerinin detaylı şekilde yazılması gerekli olup, en az 200 karakter yazılması istenmektedir. API isteğinin sebepleri yazıldıktan sonra geri dönüt verilmesi beklenmektedir. Geri dönüt alındıktan sonra API'ler kullanılarak tweetler toplanmıştır.

C) Twitter'da tweet toplamak için anahtar kelimeler kullanılarak arama yapılmıştır. Bu aramayı sağlayan kod parçası şekil 3.5'de verilmiştir.

```
search_words = " AkPartiKocaeli akgenckocaeli
tahirbuyukakin …"
date_since = "2019-03-01"
tweets = tw.Cursor(api.search,
                   q=search_words,
                   lang="tr",
                   since=date_since)
```

Şekil 3.5. Anahtar kelimelere göre arama sağlayan kod parçası

Tablo 3.1. Tezde sorgulama amaçlı kullanılan twitter anahtar kelimeleri

Parti	**Anahtar Kelimler**
Ak Parti	AkPartiKocaeli akgenckocaeli tahirbuyukakin mehmetyasinozlu ciftcibnymn muzafferbiyik aygunzeki hamzasayir041 zinnurbuyukgoz AvYildirimSEZER mimarsibelgonul adnanturankb
İyi Parti	serdarkaman41 iyigenclik_41 iyikocaeli41 IyiKocaeli iyi_parti_41 iyi_kandira basiskele_iyi iyicayirova41 iyikorfez41
Saadet Partisi	birolaydinSP av_zafermutlu YusufAksuTR selimcetinkya halilkayin ercan_dalkilic OpDrBurcu NihatYildiz41 MehmetOzalay41 ibrahimbiyiklii RecepSaridogan saadetkadin41

Anahtar kelimeler kullanılarak Python dilinde yazılan bir program vasıtasıyla yapılan aramalarda 1968 adet ham veri toplanmış, toplanan ham veriler xls ve Json formatında kaydedilmiştir. Toplanan tweetler örnekler Şekil 3.6' da gösterilmiştir. Ayrıca Twitter'da yayınlanan tweetlere örnekler de Şekil 3.7' de gösterilmiştir.

	A
1	Screen Name,Name,Status ID,Text,Local Timestamp,Status Permalink,Status Hidden Link,Conversation ID,Permalink Path,Twitter Time,Image 1,Image
2	,user-info profile lists moments profile,,,,,https://twitter.com,,,,,,,,,
3	SERKANEGE17,SERKAN EGE,1112134631724384256,"DÃ¼nyalÄ±k Ä°ÅŸler ile meÅŸgulken ahiret iÃ§in bu fani tarlada GÃ¶nÃ¼ller alÄ±p ahiretide kazanmayÄ±
4	SERKANEGE17,SERKAN EGE,1112133215739891712,AblacÄ±m ben ÅŸimdiden sizi tebrik ediyorum kazanacaÄŸÄ±ndan hiÃ§ ÅŸÃ¼phem yok gerÃ§ekten Ã§ok
5	GokcenParlakk tahirbuyukakin,GÃ¶kÃ§en Parlak,1112127376668925959 1108046206851256320,YarÄ±n yine hep birlikte kutlayacaÄŸÄ±z Ä°nÅŸAllah...
6	fatmahafsa17,Fatma Ø-ÙØµØ© â˜ª,1112120907584229376,HiÃ§ yakmadÄ±m Ã§Ã¼nkÃ¼ hiÃ§ olmadÄ±. RTE da bizim gibi 15 temmuzda gÃ¶rdÃ¼

Şekil 3.6. Anahtar kelimeler vasıtasıyla toplanan tweet örnekleri

Şekil 3.7. Twitter'da yayınlanan tweet örnekleri

3.1.2. Verilerin temizlenmesi

Twitter'dan elde edilen veriler genelde imla ve dilbilgisi kuralları bakımından oldukça zayıftır. Twitter'daki 140 karakter sınırlamasından dolayı yapılan kısaltmalar, harf eksiklikleri, kuralsız yapılar gibi birçok sorun Duygu Analizi yönteminde zorluk çıkarmaktadır. Bu nedenlerden ötürü, kullanıcılar düşüncelerini, kelimeleri kısaltarak ya da işaretler kullanarak ifade etmektedir. Bu seviyedeki metinlerde Duygu Analizi yöntemi kullanıldığı zaman sonuçlar düşük başarı göstermektedir. Bundan dolayı,

temiz bir Corpus/külliyat oluşturmak için tweetlerin temizlenmesi gerekmektedir. Metin temizleme işlemi Regular Expression (RegEx) açık kaynaklı kütüphanesi yardımı ile yapılmıştır. Temizleme işlemi, metnin normalleştirilmesi olarak da ifade edilebilmektedir.

3.1.2.1. Metnin normalleştirilmesi

Twetlerde standart olmayan birçok kelime bulunmaktadır. Duygu analizi sürecini kolaylaştırmak adına standart olmayan kelimeler, standart hale dönüştürülmelidir. Standart kelimelerden kasıt, farklı karakterlerin kullanılması, kısaltılmış kelimeler ya da ifadeler olabilir. Karakter tekrarının önüne geçmek için, yinelemeli ifadelerin kaldırılması gerekmektedir. Diğer taraftan büyük küçük harf uyumu da önemli bir konudur. Bu hususta tüm veriler küçük harfe dönüştürülmelidir. Ayrıca metinde bulunan sembollerin de kaldırılması gerekmektedir. Bunun nedeni sembollerin metin içerisinde anlam bilgisinin olmamasıdır. Bu aşamada, noktalama işaretleri ve harf olmayan tüm karakterler temizlenmektedir (Pratama ve diğ., 2019). Bir sonraki ön işleme adımı, bağlantıyı kaldırma, tekrarlanan tweetler veya retweetler, hashtagler ve bir tweetin duygularının çıkarılmasında hiçbir etkisi olmayan kelimeler gibi anlamsız ifade veya karakterlerin ortadan kaldırılması aşamasıdır. Temizlenmesi gereken tweetler tablo 3.2 de örnek olarak gösterilmiştir.

Tablo 3.2. Düzenli ifade desenleri

TİP	**ÖRNEK**	**SİLİNEN**
URL	https://twitter.com/koraytonyali/status/1112102273608548354	https://.*
HASHTAG	#MarttınSonuBahar, #31MarttaEnGüçlüDestekGençlerdenGelecek	#.*
MENTION	@fatmakaplan, @mimarsibelgonul, @ekrem_imamoglu @mansuryavas06	@.*
RAKAM	112093596021338118, 1553977319	[0-9]+

Verilerin temizlenmeden önceki hallerine örnekler Tablo 3.3'de gösterilmiştir.

Tablo 3.3. Gürültülü tweet örnekleri

Gürültülü Tweet Örnek Türleri	**Örnekler**
Tarafsız	YaklaÅŸÄ±k 1000 kiÅŸi ile deplasman yapan Kocaelispor taraftarÄ±nÄ± emniyet kontenjan dan kaynaklÄ± iÃ§eri almadÄ±.
Pozitif	Millet ittifakÄ± adayÄ±mÄ±z @fatmakaplan ile buralara Ã§oktan bahar gelmiÅŸ. ParticiliÄŸe deÄŸil, belediyeciliÄŸe oy vereceÄŸimiz 31 Mart gÃ¼nÃ¼, zafer halkÄ±mÄ±zÄ±n olacak.
Negatif	BRturku,BozkÄ±rTÃ¼rkÃ¼ğŸ‡¹ğŸ‡·,1109740239646609410,Kocaelisp ordan size oy Ã§Ä±kmaz sibel hanÄ±m zorlamayÄ±n!

Temizleme işlemleri ise aşağıdaki sırada yapılmıştır.

a) Hashtags/Başlık Etiketi (#) silindi,
b) Mentions/İmalar (@) silindi,
c) Linkler (http://..) silindi,
d) Retweets/ (RT ..) silindi,
e) Tüm sayılar silindi,
f) Noktalama işaretleri ve özel karakterler kaldırıldı,
g) Büyük harfle yazılan tüm tweetler küçük harfe çevrildi,
h) Karakter sınırlaması yapılarak (length(word) > 2), iki ve daha az harften oluşan kelimler silindi,
i) Etkisiz/Dolgu Kelimeler (Stop words); "ama, az, bir, biz, bu, gibi, hala" gibi Türkçe'de çok sık kullanılan ve anlama etkisi olmayan kelimler veri setinden çıkartıldı.

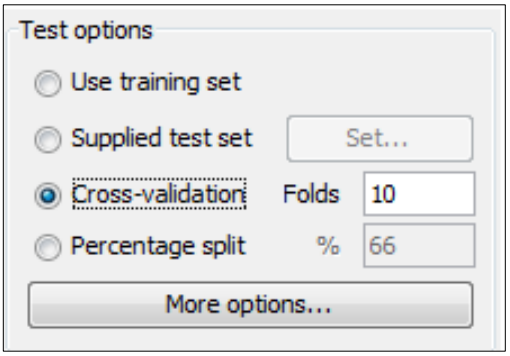
3.1.3. Duygu analizi (sentiment analysis)

Bu çalışmada, Doğal Dil işleme (DDİ)'ye yönelik geliştirilen açık kaynak kodlu kütüphane olan Natural Language Tool Kit (NLTK) kütüphanesi kullanılmıştır. Kullanılan kütüphane sayesinde verilere biçimbirimsel çözümleme işlemi yapılmamıştır. Bu kütüphane, Python programlama dili için yazılmış bir kütüphane olup, kelimeleri köklerine ayırarak, etiketleyip sınıflandırmaktadır. NLTK kütüphanesi, ayrıca yapay sinir ağlarına dayanan yöntemleri de desteklemektedir. Diğer kütüphanelerle kıyaslandığı zaman, birçok avantajı bulunmaktadır. Problemlere daha akademik şekilde yaklaşması, çok sayıda konuşma dilini desteklemesi, kütüphane içinde birçok metin sınıflandırma yöntemlerinin olması avantajları arasında

gösterilebilir. Diğer kütüphanelere göre dezavantajı ise çok yavaş çalışmasıdır. NLTK kütüphanesinin yetenekleri aşağıda sıralanmıştır:

a) Metinler ve kelimeler ile işlem yapabilmektedir,
b) Metin külliyatlarına (Corpus) ve farklı dillerde sözlük kaynaklarına sahiptir,
c) Farklı ortamlardaki metinlere ulaşıp işleme yeteneği vardır,
d) Yapısal programlamaya izin vermektedir,
e) Kelimeleri etiketleyip ve sınıflandırabilmektedir, (isim zarf vb.)
f) Metin sınıflandırma yöntemleri çok çeşitlidir,
g) Metinden bilgi çıkarma yeteneği vardır,
h) Cümle yapısını analiz edebilmektedir,
ı) Dilbilgisi özelliklerini çıkarabilmektedir,
i) Farklı lehçeler içinde cümleyi tanıyabilmektedir.

Temizlenen veriler etiketlendikten sonra, etiketli veriler üzerinden bilgi çıkarımı yapılarak sınıflandırılmıştır. Sınıflandırma yapılırken veriler çeşitli yöntemlerle eğitilmektedir. Şekil 3.8'de kullanılan yöntemler WEKA programı ara yüzünde gösterilmiştir. Bu yöntemlerden ilk olarak, Training set yöntemi, seçilen verilerin yalnızca eğitim amaçlı olarak kullanılmasına müsaade etmektedir. Hangi küme üzerinde çalışılmak istenirse, o küme seçilerek çalıştırılmaktadır. Diğer bir yöntem Cross-Validation yöntemidir. Bu yöntemde, öncelikle küme sayısı olarak tanımlanan k değeri seçilmektedir. Seçilen k değeri ile birlikte veriler k adet parçaya ayrılmaktadır. Alt kümlerden biri eğitim sınıfı olarak kabul edilip, sistem eğitilmektedir. Daha sonra bu eğitim sonucu, diğer alt kümeleri test etmekte ve bu işlem belirtilen küme sayısı kadar tekrarlanarak sistemin iyileştirilmesi sağlanmaktadır. Bir başka yöntem ise Percentage Split yöntemidir. Bu yöntem veri kümesini seçilen oranda ikiye bölüp, ilk kümeyi eğitim ikinci kümeyi ise test amaçlı kullanmaktadır. Genelde bu oran %66'ya %34 olarak belirlenmektedir. Kümenin %66'lık kısmı eğitim, %34'lük kısmı ise test amaçlı kullanılmaktadır. Eğitim verileri, test verileri ile kontrol edildikten sonra veriler sınıflandırma için hazır hale gelmektedir. Çalışmamızda Cross-Validation ve Percentage Split yöntemi kullanılarak başarı oranları karşılaştırılmıştır. Yöntemlerin karşılaştırılması Tablo 3.4' de gösterilmiştir.



Şekil 3.8. Eğitim yöntemleri WEKA programı ara yüzü

Tablo 3.4. Sınıflandırma ve eğitim yöntemlerinin başarı oranları

	Eğitim Yöntemleri	
Sınıflandırma Yöntemleri	**Cross- Validation Başarı Oranı**	**Percentage Split Başarı Oranı**
Destek Vektör Makinesi	%61.79	%65.62
Naive Bayes	%61.79	%65.62

Veriler bu iki eğitim yöntemi ile eğitildikten sonra, Naive Bayes ve Destek Vektör Makineleri ile pozitif ve negatif sınıflara ayrılmış ve bu iki yöntem başarısı karşılaştırılmıştır. Ayrıca pozitif tweetlerin tüm tweetlere oranı hesaplanarak, seçim sonuç tahmininde bulunulmuştur. Duygu analizi yapılırken ise Şekil 3.9'da verilen kod parçası kullanılmıştır.

```
# Veri analizi için gerekli kütüphaneler
import pandas as pd
import matplotlib.pyplot as plt

#Veri, hazırlama ve Özellik çıkarımı için gerekli
from textblob import TextBlob
from nltk.corpus import stopwords
from nltk.stem.wordnet import WordNetLemmatizer
from sklearn.feature_extraction.text import CountVectorizer,
TfidfTransformer
import re

# Model oluşturma ve sonucu doğrulama kütüphaneleri
from sklearn.naive_bayes import MultinomialNB
from sklearn.model_selection import train_test_split
from sklearn.pipeline import Pipeline
from sklearn.metrics import confusion_matrix,
classification_report,accuracy_score
```

Şekil 3.9. Duygu analizinde kullanılan kod parçası

3.1.4. Twitter' da duygu analizi için değerlendirme ölçütleri

Twitter'da duygu analizi, sınıflandırma problemi olarak düşünülebilir. Bunun sebebi; bir tweet de ifade edilen görüşün pozitif veya negatif olarak sınıflandırılmasının hedef olarak belirlenmesidir. En sık olarak kullanılan değerlendirme ölçütleri, geleneksel sınıflandırma problemlerinden esinlenmiştir. Bunlar, doğruluk, kesinlik, duyarlılık ve F-skorudur. Bu değerler aşağıda açıklanmıştır (Gıachanou ve Crestani, 2016).

Bir sınıflandırıcının ya da genel olarak ifade edilecek olursa, bir metnin pozitif ya da negatif duygularını sınıflandırmak için birçok sınıflandırma yöntemi kullanılmaktadır. Bu yöntemler kullanılırken ayrıca yöntemlerinde değerlendirilmesi, kontrol edilmesi gerekmektedir. Tablo 3.5.' de confusion matrisine örnek gösterilmektedir. Bu confusion matrisi, metodun tahminlerini temel gerçeğiyle karşılaştırmak için, duyarlılığı bilinen bir dizi test verisindeki performansı açıklamada kullanılmaktadır. Gerçek değer ve öngörü değerleri arasında 4 farklı kombinasyondan oluşmaktadır. Tahminlerin doğruluğu hakkında bilgi vermektedir. Gerçek Pozitif (TP), Gerçek Negatif (TN), Yanlış Pozitif (FP) ve Yanlış Negatif (FN) örneklerinin sayısını göstermektedir. TP, pozitif olarak tahmin edilen ve gerçekten de pozitif olan örneklerin sayısını temsil ederken, FP pozitif olarak tahmin edilen ancak gerçekte negatif olan örneklerin sayısıdır. TN, negatif olarak tahmin edilen ve gerçekten de negatif olan örneklerin sayısını tahmin ederken, FN negatif olarak tahmin edilen fakat gerçekte pozitif olan örneklerin sayısıdır.

Tablo 3.5. Confusion matrix (karmaşıklık matrisi)

		ÖN GÖRÜLEN (PREDICTED)	
		Pozitif	**Negatif**
GERÇEK (ACTUEL)	**Pozitif**	**T**rue **P**ozitive (**TP**) (Gerçek Pozitif)	**F**alse **N**egative (**FN**) (Yanlış Negatif)
	Negatif	**F**alse **P**ozitif (**FP**) (Yanlış Pozitif)	**T**rue **N**egative (**TN**) (Doğru Negatif)

3.1.4.1. Doğruluk (accuracy)

Doğruluk en sık kullanılan değerlendirme ölçütüdür. Değerlendirilen yöntemin ne kadar sıklıkla doğru tahmin yapıldığını ölçer. Doğru sınıflandırılmış örnek sayısının (TP+TN), toplam örnek sayısına (TP+TN+FP+FN) oranıdır. Doğruluk değeri, analizin gerçek değere ne kadar yakın olduğunu göstermektedir. Formül (3.1)'de doğruluk değerinin formülü verilmektedir.

$$\text{Doğruluk} = \frac{(TP+TN)}{(TP+TN+FP+FN)} \tag{3.1}$$

3.1.4.2. Kesinlik (precision)

Yöntemin kesinliğini temsil eden ölçüttür. Pozitif olarak tahmin edilen ve gerçekten de pozitif olan örneklerin (TP), pozitif olarak tahmin edilen toplam örnek sayısına (TP+FP) bölünmesi ile hesaplanmaktadır. Formül (3.2)'de kesinlik değerinin formülü verilmektedir.

$$\text{Kesinlik} = \frac{TP}{(TP+FP)} \tag{3.2}$$

3.1.4.3. Duyarlılık (recall)

Gerçek pozitif oran olarak da bilinmektedir. Pozitif olarak tahmin edilen pozitif örneklerin oranını belirtmektedir. Doğru sınıflandırılmış pozitif örnek sayısının (TP), toplam pozitif örnek sayısına (TP+FN) bölünmesi ile hesaplanmaktadır. Formül (3.3)'de duyarlık değerinin formülü verilmektedir.

$$\text{Duyarlılık} = \frac{TP}{(TP+FN)} \tag{3.3}$$

3.1.4.4. F-Skoru

Genellikle duyarlılık ve kesinlik hesaplaması tek başına anlamlı karşılaştırma sonucu vermek için yetersizdir. İkisinin bir kombinasyonu, yöntemlerin performansını değerlendirmek için daha uygundur. Bunun için F-Skoru ölçütü tanımlanmıştır. F-skoru, duyarlılık ve kesinliği birleştiren bir ölçüttür. Formül (3.4)'de F-skoru değerinin formülü verilmektedir.

$$\text{F-Skoru}=2*\frac{\text{Kesinlik}*\text{Duyarlılık}}{\text{Kesinlik}+\text{Duyarlılık}} \quad (3.4)$$

Son olarak, duyarlılık sınıflandırması çok sınıflı bir problem olarak formüle edildiği zaman, genelde pozitif, negatif ve tarafsız sınıflar için F-Skoru puanının hesaplanması yaygın olarak kullanılmaktadır. Ancak, tarafsız sınıfı öngöremeyen yaklaşımlar da bulunmaktadır. Bu yaklaşımlar tarafsız tweetleri içeren tüm temel gerçeklere göre değerlendirilirler.

3.2. Kocaeli Büyükşehir Belediyesi Seçim Sonuçları

31 Mart 2019 yerel seçim sonuçlarına göre, seçime katılan seçmen sayısı: 1.350.376, kullanılan oy sayısı: 1.152.218, açılan sandık sayısı: 4.196, katılım oranı: %85,25, geçerli oy sayısı: 1.098.317, geçersiz oy sayısı: 52.901'dur. Yerel seçim sonuçları Tablo 3.6.'da detaylı bir şekilde gösterilmiştir.

Tablo 3.6. Kocaeli İli 31 Mart 2019 yerel seçim sonuçları

Sıra	PARTİLER	ADAY	OY	ORAN
1	AK PARTİ	Tahir Büyükakın	610.350	%55,57
2	İYİ PARTİ	Serdar Kaman	359.010	%32,69
3	SAADET	Birol Aydın	54.640	%4,97
4	HDP	Züleyha Gülüm	41.851	%3,81
5	DSP	Ziya Topal	10.875	%0,99
6	DP	Nail Baki	6.212	%0,57
7	BTP	Metin Yılmazlar	4.692	%0,43
8	TKP	Mustafa Tozkoparan	3.470	%0,32
9	BĞMSZ	Reyhan Başaran Koç	3.135	%0,29
10	VATAN	Ahmet Şahin	2.983	%0,27
11	BĞMSZ	Ersoy Kandemir	1.099	%0.10

Veri toplama işlemi ilk üç parti için yapılmıştır. SP'nin oy oranı yüzde 4.97 olduğu için değerlendirmeye almaya gerek duyulmamıştır. İlk iki parti AK Parti ve İYİ Parti için değerlendirme yapılmıştır.

4. SONUÇLAR VE ÖNERİLER

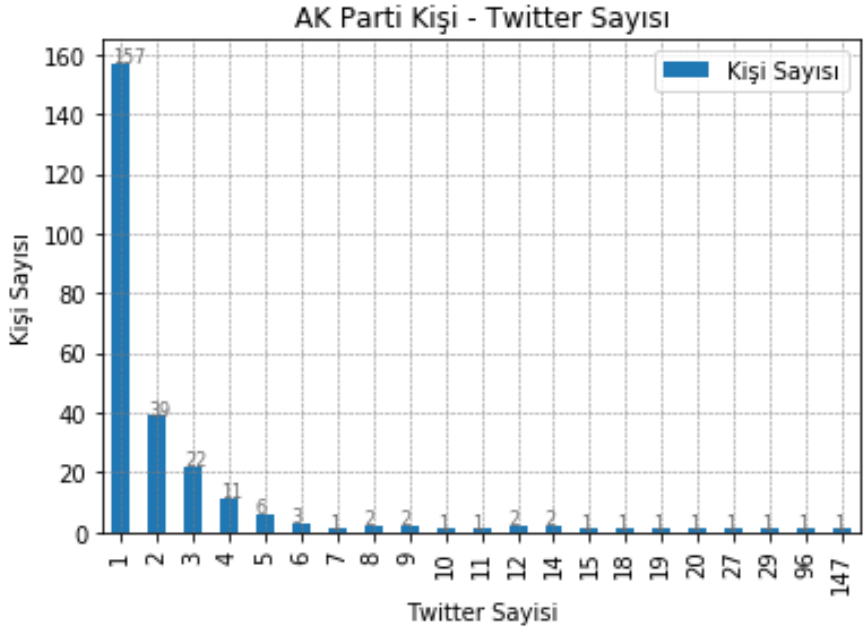
Otuz gün boyunca toplanan tweetler üzerinde, Python programı yardımıyla duygu analizi yapılmıştır. Duygu analizi yapılırken, pozitif ve negatif olmak üzere iki sınıf oluşturulmuştur. Duygu analizi sonucunda Formül (4.1) kullanılarak pozitif tweetlerin tüm tweetlere oranı hesaplanmıştır. Hesaplanan bu oran ile seçim sonuç değeri hesaplanmıştır. Ardından iki satır iki sütundan oluşan Confusion Matrisi oluşturulmuştur. Sınıflandırma yöntemi olarak Destek Vektör Makineleri ve Naive Bayes Yöntemi kullanılarak iki yöntem birbiri ile kıyaslanmıştır. Bu karşılaştırmalar yapılırken diğer bir kıstas olarak da eğitim şekillerine göre sınıflandırma başarıları da kıyaslanmıştır. Percentage Split eğitim yöntemi hem Destek Vektör Makinelerinde hem de Naive Bayes sınıflandırma yönteminde Cross-Validation eğitim yöntemine göre yaklaşık %4 oranında daha başarılı bir sonuç vermiştir. Naive Bayes yöntemi ile Destek Vektör Makinelerinin sınıflandırma başarısı ise birbirlerine çok yakın çıkmıştır. Bu karşılaştırmalar Tablo 3.4'de detaylı bir şekilde görülebilmektedir.

B = Pozitif Tweet Sayısı / Toplam Tweet sayısı (4.1)

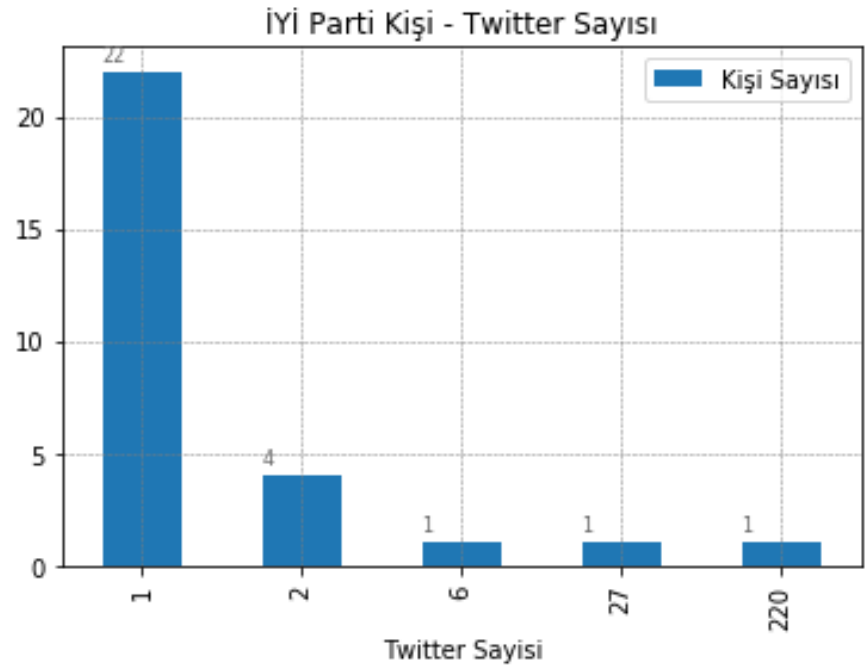
Şekil 4.1'de AK Parti için tweet atanların grafiği görülmektedir. Bu şekilde 157 kişi 1 tane, 39 kişi 2 tane, 22 kişi ise 3 tane tweet attığı görülmektedir. Bu grafikteki veriler, seçtiğimiz anahtar kelimelerin uygun olduğunun kanıtıdır. Ayrıca, toplanan tweetlerin sahte ve şişirilmiş veriler olmadığını da göstermektedir. Grafikte, anormal bir durum söz konusu değildir. Grafikte 96 ve 147 adet tweetin atıldığı görülmektedir. Bu şekilde çok fazla tweet atanlar genellikle haber servisleridir. AK Parti verilerinden 147 adet tweetin çıkarılması, sonucu anlamlı şekilde değiştirmeyeceği için, veri grubundan çıkartılmamıştır.

Şekil 4.1 için belirtilen hususlar Şekil 4.2 için de geçerlidir. Veri grubu tutarlı görünmektedir. Şekil 4.2.'de 22 kişi 1 adet, 4 kişi ise 2 adet tweet atmıştır.

Tablo 4.1 ve Tablo 4.2' de ise eğitim yöntemi ve sınıflandırma yöntemlerine göre seçim tahmin sonuçları gerçek seçim sonuçları ile karşılaştırılmış ve hata oranları gösterilmiştir.



Şekil 4.1. AK Partiye ait tweetlerin kişilere göre dağılımı



Şekil 4.2. İYİ Partiye ait tweetlerin kişilere göre dağılımı

Tablo 4.1. Percentage Split eğitim yöntemine göre seçim sonuç tahmini

	Naive Bayes Tahmin Sonucu	Destek Vektör Makineleri Tahmin Sonucu	Gerçek Seçim Sonuçları	Seçim Sonuç Tahmin Hatası
Ak Parti	%41.26	%41.26	%55.57	%14.31
İyi Parti	%24.36	%24.36	%32.69	%8.33

Tablo 4.2. Cross-Validation eğitim yöntemine göre seçim sonuç tahmini

	Naive Bayes Tahmin Sonucu	Destek Vektör Makineleri Tahmin Sonucu	Gerçek Seçim Sonuçları	Seçim Sonuç Tahmin Hatası
Ak Parti	%40.24	%40.24	%55.57	%15.33
İyi Parti	%21.54	%21.54	%32.69	%11.15

Ayrıca tweet oranlarına göre sonuçlar tahmin edildiği zaman, Ak Partinin tahmin sonucu %46.09 olarak tahmin edilirken bu oran İYİ Parti için %53.90 tahmin edilmiştir. Atılan tweetlere göre sonuçlar karşılaştırıldığında, İYİ Parti seçmeni Twitter'da etkin olduğu kadar seçimlerde etkili olamamıştır. Ayrıca tarafsız sınıfta değerlendirilen tweetlerde İYİ Parti etiketinin fazla olmasından dolayı, İYİ Parti tweetleri sağlıklı şekilde ayrıştırılamamıştır. Tablo 4.3' de Cross Validation eğitim yöntemine göre Naive Bayes sınıflandırma başarısı, Tablo 4.4' de Doğruluk oranları ve Şekil 4.3'de confusion matrisi verilmektedir.

Tablo 4.3. Cross Validation eğitim yöntemine göre Naive Bayes sınıflandırma başarısı

```
===Katmanlı Çapraz Doğrulama ===
===Özet ===

Doğru Sınıflandırılmış Örnek       1216               61.7886  %
Kappa İstatistiği                     0.2619
Ortalama Mutlak Hata                  0.4234
Karesel Ortalama Hata                 0.4884
Göreceli Mutlak Hata                 85.205   %
Bağıl Karesel Hata                   97.9817  %
Kapsamı                             100       %
Ortalama Bölge Büyüklüğü            100       %
Toplam Örnek Sayısı                1968
```

Tablo 4.4. Cross Validation eğitim yöntemine göre Naive Bayes sınıflandırmasında doğruluk oranları

=== Sınıflara Göre Detaylı Doğruluk==

	TP Oranı	FP Oranı	Kesinlik (Precision)	Duyarlılık (Recall)	F-Skoru	MCC	ROC Alanı	PRC Alanı	Sınıf
	0,873	0,600	0,554	0,873	0,678	0,305	0,607	0,520	a
	0,400	0,127	0,787	0,400	0,530	0,305	0,607	0,652	i
Ağırlıklı Ortalama	0,618	0,345	0,680	0,618	0,598	0,305	0,607	0,591	

```
=== Confusion Matrix ===

   a   b   <-- classified as
 792 115 |   a = a
 637 424 |   b = i
```

Şekil 4.3. Cross Validation eğitim yöntemine göre Naive Bayes sınıflandırmasında confusion matrisi

Burada a ile gösterilen sınıf Ak Parti'yi temsil ederken, i ile gösterilen sınıf İYİ Parti'yi temsil etmektedir. Diğer taraftan Tablo 4.5' de Cross Validation eğitim yöntemine göre Destek Vektör Makine yönteminin sınıflandırma başarısı, Tablo 4.6' da Doğruluk oranları ve Şekil 4.4'de confusion matrisi verilmektedir.

Tablo 4.5. Cross Validation eğitim yöntemine göre destek vektör makineleri sınıflandırma başarısı

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        1216               61.7886 %
Kappa statistic                          0.2619
Mean absolute error                      0.3821
Root mean squared error                  0.6182
Relative absolute error                 76.893  %
Root relative squared error            124.0108 %
Coverage of cases (0.95 level)          61.7886 %
Mean rel. region size (0.95 level)      50      %
Total Number of Instances             1968
```

Tablo 4.6. Cross Validation eğitim yöntemine göre destek vektör makineleri sınıflandırmasında doğruluk oranları

=== Sınıflara Göre Detaylı Doğruluk==

	TP Oranı	FP Oranı	Kesinlik (Precision)	Duyarlılık (Recall)	F-Skoru	MCC	ROC Alanı	PRC Alanı	Sınıf
	0,873	0,600	0,554	0,873	0,678	0,305	0,636	0,542	a
	0,400	0,127	0,787	0,400	0,530	0,305	0,636	0,638	i
Ağırlıklı Ortalama	0,618	0,345	0,680	0,618	0,598	0,305	0,636	0,594	

```
=== Confusion Matrix ===

   a   b   <-- classified as
 792 115 |   a = a
 637 424 |   b = i
```

Şekil 4.4. Cross Validation eğitim yöntemine göre destek vektör makineleri sınıflandırmasında confusion Matrisi

Diğer bir analizimiz ise Percentage Split eğitim yöntemine göre sınıflandırmadır. Tablo 4.7'de Percentage Split eğitim yöntemine göre naive bayes sınıflandırma başarısı, Tablo 4.8'de Doğruluk oranları ve Şekil 4.5'de confusion matrisi verilmektedir.

Tablo 4.7. Percentage Split eğitim yöntemine göre Naive Bayes sınıflandırma başarısı

```
Correctly Classified Instances         439               65.6203 %
Kappa statistic                          0.3253
Mean absolute error                      0.409
Root mean squared error                  0.4718
Relative absolute error                 82.1836 %
Root relative squared error             94.3916 %
Coverage of cases (0.95 level)         100      %
Mean rel. region size (0.95 level)     100      %
Total Number of Instances              669
```

Tablo 4.8. Percentage Split eğitim yöntemine göre Naive Bayes sınıflandırmasında doğruluk oranları

=== Sınıflara Göre Detaylı Doğruluk==

	TP Oranı	FP Oranı	Kesinlik (Precision)	Duyarlılık (Recall)	F-Skoru	MCC	ROC Alanı	PRC Alanı	Sınıf
	0,868	0,536	0,595	0,868	0,706	0,360	0,658	0,572	a
	0,464	0,132	0,795	0,464	0,586	0,360	0,658	0,664	i
Ağırlıklı Ortalama	0,656	0,324	0,700	0,656	0,643	0,360	0,658	0,620	

```
=== Confusion Matrix ===

   a   b   <-- classified as
 276  42 |   a = a
 188 163 |   b = i
```

Şekil 4.5. Percentage Split eğitim yöntemine göre Naive Bayes sınıflandırmasında confusion matrisi

Öte yandan Tablo 4.9'da Percentage Split eğitim yöntemine göre Destek Vektör Makine yönteminin sınıflandırma başarısı, Tablo 4.10' de Doğruluk oranları ve Şekil 4.6'da confusion matrisi verilmektedir.

Tablo 4.9. Percentage Split eğitim yöntemine göre destek vektör makineleri sınıflandırma başarısı

Correctly Classified Instances	439	65.6203 %
Kappa statistic	0.3253	
Mean absolute error	0.3438	
Root mean squared error	0.5863	
Relative absolute error	69.0762 %	
Root relative squared error	117.2992 %	
Coverage of cases (0.95 level)	65.6203 %	
Mean rel. region size (0.95 level)	50 %	
Total Number of Instances	669	

Tablo 4.10. Percentage Split eğitim yöntemine göre destek vektör makineleri sınıflandırmasında doğruluk oranları

=== Sınıflara Göre Detaylı Doğruluk==

	TP Oranı	FP Oranı	Kesinlik (Precision)	Duyarlılık (Recall)	F-Skoru	MCC	ROC Alanı	PRC Alanı	Sınıf
	0,868	0,536	0,595	0,868	0,706	0,360	0,666	0,579	a
	0,464	0,132	0,795	0,464	0,586	0,360	0,666	0,650	i
Ağırlıklı Ortalama	0,656	0,324	0,700	0,656	0,643	0,360	0,666	0,616	

```
=== Confusion Matrix ===

   a   b   <-- classified as
 276  42 |   a = a
 188 163 |   b = i
```

Şekil 4.6. Percentage Split eğitim yöntemine göre destek vektör makineleri sınıflandırmasında confusion matrisi

Ayrıca makine öğrenmesi ile sınıflandırılan veriler ile manuel olarak sınıflandırılan verilerin seçim sonucu tahminleri de karşılaştırılmıştır. Tahminler Formül (4.1)'e göre yapılmıştır. Manuel olarak toplanan verilerin seçim sonuç tahminlerine göre Ak Parti oy oranları %47.16 olarak hesaplanırken, İYİ Parti oy oranı %32.02 olarak hesaplanmıştır. Sonuçlar Tablo 4.11' de gösterilmiştir. Bu sonuçlara göre insan eliyle yapılan değerlendirmelerin, makinelere göre daha iyi sonuç verdiği ortaya çıkmıştır. Duygu analizi çalışmalarında verilerin iyi filtrelenip, ayrıştırılması önem arz etmektedir. Ayrıca bu işlemlerin makine öğrenmeleri yöntemleriyle yapılması, analizin hızlandırılması ve kısa sürede yapılması da ayrıca önem arz etmektedir. Manuel olarak işlenen verilerin tahmin sonuçları oranları daha yüksek çıkmış olmasına rağmen, çok fazla zaman almıştır. Dolayısıyla pratik bir yöntem asla değildir. Makine öğrenme yöntemleri daha hızlı sonuç vermelerine karşın, tahmin sonuçları manuel sonuçlar kadar iyi olmamıştır. Dolayısıyla makine öğrenme yöntemlerinin geliştirilmesi bu tarz çalışmaların sonucuna önemli ölçüde katkı sağlayacaktır.

Tablo 4.11. Manuel yöntemle toplanan tweetler

	Pozitif Tweet Sayısı	Negatif Tweet Sayısı	Tarafsız
Ak Parti	623	57	183
İyi Parti	423	35	
Toplam	1321		

Duygu analizinin başarısını artırmak için;

- Çok büyük miktardaki tweete sahip kullanıcıların tweetleri veri gurubundan çıkartılmalıdır. Çalışmamızda bu durumdan etkilenmemek için, sorgulamada kullanılan anahtar kelimeler uygun ve dar kapsamda tutulmuştur. Seçilen tweetler yalnızca parti adaylarının attığı ve kendi isimlerinin geçtiği tweetlerdir.
- Seçime yakın tarihlerdeki veri gurubu ile yapılan analizler daha başarılı olmaktadır.
- Farklı anahtar kelime gruplarında analiz sonucu farklı olacaktır.
- Logistic Regression veya K-Nearest Neighbors (KNN) yöntemleri de denenerek varsa analizin başarısını artıracak yöntem belirlenebilir.
- Twitter'da tweet toplarken, Twitter bir hafta sınırlandırması getirmekte, bu da geçmişteki verilere ulaşamaya engel teşkil etmektedir. Twitter'da veri toplarken, günlük olarak verilerin çekilmesi ile birlikte bu problemin önüne geçilebilir.
- Bu tarz kutuplu çalışmalarda, atılan tweetlerdeki duygular çalışma sonucunu etkilemektedir. Bu çalışmada, anahtar kelime olarak parti adayları kullanılmıştır. Fakat bazı Twitter kullanıcıları, attıkları tweetlerin daha geniş kesime ulaşması açısından konu ile alakasız olmasına rağmen, parti adaylarını etiketlemişlerdir. Bu da çalışma sonucunu etkileyebilmektedir. Bu tarz olayların önüne geçilmesi adına etkili filtreleme yöntemleri geliştirilebilir.
- Kocaeli dışından atılan, yani seçmen olmayan kişilerin tweetleri veri gurubunda olmaması gerekir. Bu durumu sağlamak için ise Twitter'ın konumsal özelliği kullanılabilir.
- Konumsal olarak çalışılmak istenirse, konum bilgisi için Twitter sıkıntı çıkarmakta ve yaklaşık tüm verilerin %7'si kadar sınırlı sayıda konumlu tweet vermektedir. Dolayısıyla, konumsal olarak çalışılmak istenirse, tweet toplama periyodunun oldukça uzun tutulması gereklidir.
- Ayrıca konum bazlı çalışmalarda, kişinin konum erişimine izin vermemesi durumunda, konum bilgisi alınamamaktadır. Diğer bir sıkıntı ise, profildeki konumu

ile tweet attığı konumun uyuşmazlığı söz konusu olursa, konum bazlı çalışmaların doğruluğu, hassasiyeti sıkıntıya girmektedir. Bu tarz konulara engel oluşturulacak filtreleme yöntemleri geliştirilebilir.

KAYNAKLAR

Adak Kaplan B., Twitter Üzerindeki Türkçe Mesajlarda Veri Madenciliğiyle Duygu Analizi, Yüksek Lisans Tezi, Beykent Üniversitesi, Fen Bilimleri Enstitüsü, Ankara, 2016, 450859.

Akar E., Sanal toplulukların Bir Türü Olarak Sosyal Ağ Siteleri-Bir Pazarlama İletişim Kanalı Olarak İşleyişi, *Anadolu Üniversitesi, Sosyal Bilimler Dergisi*, 2010,**10**(1), 107-122.

Akgöbek Ö., Çakır, F., Veri Madenciliğinde Bir Uzman Sistem Tasarımı, *Akademik Bilişim 09*, Harran Üniversitesi, Şanlıurfa, 11-13 Şubat 2009.

Akgül E.S., Ertano C., Diri B., Twitter Verileri İle Duygu Analizi, *Pamukkale Üniversitesi Mühendislik Bilim Dergisi*, 2016, **22**(2), 106-110.

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

Alaei A.R., Becken S., Stantic B., Sentiment Analysis in Tourism: Capitalizing on Big Data, *Journal of Travel Research*, 2019, **58**(2), 175-191.

Albayrak A.S., Koltan Yılmaz Ş., Veri Madenciliği: Karar Ağacı Algoritmaları ve İMKB Verileri Üzerine Bir Uygulama, *Süleyman Demirel Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi,* 2009, **14**(1), 31-52.

Albayrak M., Bilimsel Araştırmalarda Veri Madenciliği Kullanımı, *International Journal of Social Sciences and Education Research,* 2017, **3** (3), 751-760.

Albayrak M., EEG Sinyallerindeki Epileptiform Aktivitenin Veri Madenciliği Süreci ile Tespiti, Doktora Tezi, Sakarya Üniversitesi, Fen Bilimleri Enstitüsü, Sakarya, 2008, 216012.

Argüden Y., Erşahin B., Veri Madenciliği Veriden Bilgiye, Masraftan Değere, ARGE Danışmanlık Yayınları No: 10, http://www.arge.com/wp-content/uploads/2013/03/VeriMadenciligi.pdf, (erişim tarihi: 26.12.2018).

Argüden Y., Erşahin B., Veri Madenciliği Veriden Bilgiye, Masraftan Değere*,* 1rd ed., ARGE Danışmanlık A.Ş., İstanbul, 2008.

Arı A., Önder H., Farklı Veri Yapılarında Kullanılabilecek Regresyon Yöntemleri, *Anadolu Tarım Bilim Dergisi,* 2013, **28**(3), 168-174.

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

Babaoğlu A., Veri Madenciliği Yöntemleri ve Bir Uygulama, Yüksek Lisans Tezi, Selçuk Üniversitesi, Fen Bilimleri Enstitüsü, Konya, 2015, 422159.

Baykal A., Veri Madenciliği Uygulama Alanları Application Fields Of Data Mining, *D.Ü.Ziya Gökalp Eğitim Fakültesi Dergisi,* 2006, **7**, 95-107.

Berberoğlu B., 2008 Global Krizinin Türkiye ve Avrupa Birliği'ndeki Etkilerinin Kümeleme Analizi İle İncelenmesi, *Anadolu Üniversitesi sosyal Bilimler Dergisi,* 2011, **11** (1), 105-130.

Bilgin T.T. , Çamurcu A.Y., Çok Boyutlu Veri Görselleştirme Teknikleri, *Çanakkale Onsekiz Mart Üniversitesi Akademik Bilişim Dergisi,* 2008, 107-112.

Boyd, D.M., Ellison N.B, Social Network Site: Definition, History, and Scholarship, *Journal of Computer- Mediated Communication*, 2008, DOI: 10.1111/j.1083-6101.2007.00393.x.

Çağlayan Akay E., Ekonometride Yeni Bir Ufuk: Büyük Veri ve Makine Öğrenmesi, *Social Sciences Research Journal*, 2018, **7**(2), 41-53.

Çalış K., Gazdağı O., Yıldız O., Reklam içerikli epostaların metin madenciliği yöntemleri ille otomatik tespiti, *Bilişim Teknolojileri Dergisi*, 2013, **6**(1), 3.

Çilingirtürk A.M., Altaş D., Makro İktisat Verilerinde Kayıp Verilerin Regresyona Dayalı En Yakın Komşu "Hot Deck" Yöntemi İle Tamamlanması, *Dokuz Eylül Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi,* 2010, **25**(2), 73-83.

Çoban Ö., Özyer B. ve Özyer G. T., Sentiment analysis for Turkish Twitter feeds, *2015 23nd Signal Processing and Communications Applications Conference (SIU),* DOI: 10.1109/SIU.2015.7130362

Çoban Ö., Tümüklü Özyer G., Türkçe Twitter Mesajları için LDA ile Duygu Sınıflandırması, *Signal Processing and Communication Application Conference (SIU 2016)*, 2016, 126-132.

Çoban Ö., Tümüklü Özyer G., Twitter Duygu Analizinde Terim Ağırlıklandırma Yönteminin Etkisi, *Pamukkale Üniversitesi Mühendislik Bilim Dergisi,* 2017, **24** (2), 283-291.

Çomu T., Halaiqa İ., Web İçeriklerinin Metin Temelli Çözümlemesi**,** Editör: M. Binark, *Yeni Medya Çalışmalarında Araştırma Yöntem ve Teknikleri*, Ayrıntı Yayınları, İstanbul, 26-87, 2014.

Dal N. E., Dal V., Kişilik Özellikleri ve Sosyal Ağ Sitesi Kullanım, Alışkınlıkları: Üniversite Öğrencileri Üzerine Bir Araştırma, *Mehmet Akif Ersoy Üniversitesi Sosyal Bilimler Enstitüsü Dergisi*, 2014, **6**(11), 144-162.

Demircan S., TÜİK Yaşam Memnuniyet Anketleri Üzerine Veri Madenciliği Uygulaması, Yüksek Lisans Tezi, Erciyes Üniversitesi, Fen Bilimleri Enstitüsü, Kayseri, 2015, 394295.

Doğan, S , Diri, B., Türkçe Dokümanlar İçin N-gram Tabanlı Yeni Bir Sınıflandırma(Ng-ind): Yazar, Tür ve Cinsiyet, *Türkiye Bilişim Vakfı Bilgisayar Bilimleri ve Mühendisliği Dergisi*, 2010, **3**, 11-19.

Doğan, S., "Türkçe Dokümanlar için N-gram Tabanlı Sınıflandırma: Yazar, Tür ve Cinsiyet", Yüksek Lisans Tezi, Yıldız Teknik Üniversitesi., Fen Bilimleri Enstitüsü, İstanbul, 2006, 182635.

Dondurmacı G. A., Çınar A., Finans Sektöründe Veri Madenciliği Uygulaması, *Akademik Sosyal Araştırmalar Dergisi*, 2014, **2** (1), 258-271.

Elabiad Z., Web Tabanlı Anket Sistemi İle Elde Edilen Verilerin Veri Madenciliği Yöntemi İle Analizi, Yüksek Lisans Tezi, İstanbul Aydın Üniversitesi, Fen Bilimleri Enstitüsü, İstanbul, 2013, 342511.

Eliaçık A.B., Erdoğan N., Mikrobloglardaki Finans Toplulukları için Kullanıcı Ağırlıklandırılmış Duygu Analizi Yöntemi, 2015, *Ulusal Yazılım Mühendisliği Sempozyumu,* 782-793.

Fayyad U., Piatetsky-Shapiro G. and Smyth P., The KDD Process for Extracting Useful Knowledge from Volumes of Data, *Communications of The ACM* , 1996a, **39** (11), 27-34.

Fayyad U., Shapiro G.P and Smyth P., From Data Mining to Knowledge Discovery in Databases, *Al Magazine*, 1996b, **17** (3), 37-54.

Giachanou A., Crestani F., Like It or Not: A Survey of Twitter Sentiment Analysis Methods, *ACM Computing Surveys*, 2016, **28** (2), 28-41.

Go A., Bhayani R., Huang L., Twitter Sentiment Classification using Distant Supervision, *Technical Report*, Stanford University, California, 2009.

Gokulakrishnan B., Priyanthan P., Ragavan T., Prasath N., Perera A., Opinion Mining and Sentiment Analysis on a Twitter Data Stream, 2012, *The International Conference on Advances in ICT for Emerging Regions*, 182-188.

Gülçe G., Veri Ambarı ve Veri Madenciliği Teknikleri Kullanılarak Öğrenci Karar Destek Sistemi Oluşturma, Yüksek Lisans Tezi, Pamukkale Üniversitesi, Fen Bilimleri Enstitüsü, Denizli, 2010, 275300.

Güldoğan E., Çeşitli Çekirdek Fonksiyonları İle Oluşturulan Destek Vektör Makinesi Modellerinin Performanslarının İncelenmesi: Bir Klinik Uygulama, doktora Tezi, İnönü Üniversitesi ve Mersin Üniversitesi, Sağlık Bilimleri Enstitüsü, Malatya, 2017, 462673

Hastie T., Tibshirani R., Friedman J., The Elements Of Statistical Learning; Data Mining, Inference And Prediction, *Springer Series In Statistics*, 2001.

Hazar M., Sosyal Medya Bağımlılığı-Bir Alan Çalışması, 2011, *İletişim Kuram ve Araştırma Dergisi*, **32**,151-175.

İnan, O., Öğrenci işleri veri tabanı üzerinde veri madenciliği uygulamaları, Yüksek Lisans Tezi, Selçuk Üniversitesi, Fen Bilimleri Enstitüsü, Konya, 2003, 134156.

Jiangl J., Yu M., Zhoul M., Liu X., Zhao T., Target-dependent Twitter Sentiment Classification, *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, 2011, 151-160.

Joachims T., A Probabilistic Analysis of the Rocchio Algorithm with TFIDF for Text Categorization. *Carnegie- Mellon Univ Pittsburgh Pa Dept of Computer Science*, 1996, 96-118.

Kara Y., Coşkun A., Sosyal Ağların Pazarlama Aracı Olarak Kullanımı: Türkiye'deki Hazır Giyim Firmaları Örneği, *Afyon Kocatepe Üniversitesi IIBF Dergisi*, 2012, **15** (2), 73-90.

Kavzaoğlu T., Çölkesen İ., Destek Vektör Makineleri ile Uydu Görüntülerinin Sınıflandırılmasında Kernel Fonksiyonlarının Etkilerinin İncelenmesi, *Harita Dergisi,* 2010, **144**, 73-82.

Kaynar O., Tuna M.F., Görmez Y., Deveci M.A., Makine Öğrenmesi Yöntemleriyle Müşteri Kaybı Analizi, *Cumhuriyet Üniversitesi İktisadi ve İdari Bilimler Dergisi,* 2017, **18**(1), 1-14.

Khan, A. Z. H., Atique, M., Thakare V.M., Combining lexicon-based and learning-based methods for Twitter sentiment analysis, *International Journal of Electronics, Communication and Soft Computing Science & Engineering (IJECSCSE)*, 2015, Special Issue, 89-91.

Kızılkaya Y.M., Oğuzlar A., Bazı Denetimli Öğrenme Algoritmalarının R Programlama Dili İle Kıyaslanması, *Karadeniz Uluslararası Bilimsel Dergi*, 2018, **37**(37), 90-98.

Kocak B.B., Polat İ., Kocak C.B., Twitter Kullanıcılarının Havayolu Pazarına Yönelik Duygu Kutuplarının Belirlenmesi: Bir Fikir Madenciliği Örneği, *Global Business Research Congress(GBRC),* İstanbul, Türkiye, 26-27 Mayıs 2016.

Kwon E., Sung Y., Follow Me! Global Marketers Twitter Use, *Journal of Interactive Advertising*, 2011, **12** (1), 4-16.

Lee S., Cho M., Social Media Use in a Mobile Broadband Environment: Examination of Determinants of Twitter and Facebook Use, *Mobile Marketing Association, IJMM*, 2011, **6**(2), 72-87.

Lewis D. D., Naive (Bayes) at forty: The independence assumption in information retrieval, *European Conference on Machine Learning 1998,* 2005, DOI: 10.1007/BFb0026666.

Lewis, D.D., An Evaluation Of Phrasal And Clustered Representations On A Text Categorization Task., *Proceedings Of The 15th Annual İnternational ACM SIGIR Conference On Research And Development İn İnformation Retrieval. ACM*, Copenhagen, Denmark, June 21-24 1992.

Maltarollo V. G., Honório K. M., Da Silva A. B. F., Bölüm 10, Editör: Suzuki K., *Artificial Neural Networks: Architectures and Applications*, 9, BoD – Books on Demand, Rijeka-Croatia, 203-216, 2013.

Onan A., Korukoğlu S., Makine Öğrenmesi Yöntemlerinin Görüş Madenciliğinde Kullanılması Üzerine Bir Literatür Araştırması, *Pamukkale Üniversitesi Mühendislik Bilimleri Dergisi,* 2016, **22** (2), 111-122.

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

Özcan C., Veri Madenciliğinin Güvenlik Uygulama Alanları ve Veri Madenciliği ile Sahtekârlık Analizi, Yüksek Lisans Tezi, İstanbul Üniversitesi, Sosyal Bilimler Enstitüsü, İstanbul, 2014, 380714.

Özekes S., Veri Madenciliği Modelleri ve Uygulama Alanları*, İstanbul Ticaret Üniversitesi Dergisi*, 2003, **3**, 65-82.

Pak A., Paroubek P., Twitter as a Corpus for Sentiment Analysis and Opinion Mining., *Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC 2010)*, Mediterranean Conference Centre,Valletta, Malta, May 17-23 2010.

Pak A., Paroubek P., Twitter As A Corpus For Sentiment Analysis And Opinion Mining, *In Proceedings of Language Resources and Evaluation Conference*, 2010, 1320-1326.

Pang B., Lee L., Opinion Mining And Sentiment Analysis, *Foundations and Trends in Information Retrieval*, 2008, **2**(1-2), 1–135.

Pang B., Lee L., Sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. *Proceedings of the 42nd Annual Association for Computational Linguistics Conference,* Barcelona, Spain, July 21 – 26 2004.

Pratama Y., Tampubolon A.R., Sianturi L.D., Manalu R.D., Pangaribuan D.F., Implementation of Sentiment Analysis on Twitter Using Naive Bayes Algorithm to Know the People Responses to Debate of DKI Jakarta Governor Election*, IOP Conf. Series: Journal of Physics: Conf. Series,* 2019, 1-7.

Sağlam F., Otomatik Duygu Sözlüğü Geliştirilmesi ve Haberlerin Duygu Analizi, Doktora Tezi, Hacettepe Üniversitesi, Fen Bilimler Enstitüsü, Ankara, 2019, 587951.

Santos, C.N., Gatti, M., Deep convolutional neural networks for sentiment analysis of short texts, *In Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, Dublin, Ireland, August 23-29 2014.

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

Song Z., Xia J., Spatial and Temporal Sentiment Analysis of Twitter data, *European Handbook of Crowdsourced Geographic Information,* 2016, 205-221.

Sun J., Li H., Data Mining Method for Listed Companies, Financial Distress Prediction, *Knowledge-Based Systems*, 2008, **21**(1), 1-5.

Sun Q., Wang N., Li S., Zhou H., Local Spatial Obesity Analysis And Estimation Using Online Social Networksensors, *Journal of Biomedical Informatics,* 2018, **83**, 54-62.

Terzi Ö., Küçüksille E.U., Ergin G., İlker A., Veri Madenciliği Süreci Kullanılarak Güneş Işınımı Tahmini, *SDU International Technologic Science*, 2011, 3(2), 29-37.

Terzi S., Hile ve Usulsüzlüklerin Tespitinde Veri Madenciliğinde Kullanımı, *Muhasebe ve Finansman Dergisi*, 2012, **54**, 51-64.

Timisi N., Yeni İletişim Teknolojileri ve Demokrasi: İnternet Ortamında Kamusal Katılım, Doktora Tezi, Ankara Üniversitesi, Sosyal Bilimler Enstitüsü, Ankara, 2003, 82088.

Turanlı M., Özden Ü.H., Türed S., Avrupa Birliği'ne Aday ve Üye Ülkelerin Ekonomik Benzerliklerinin Kümeleme Analizi ile İncelenmesi, *İstanbul Ticaret Üniversitesi Sosyal Bilimler Dergisi*, 2006, **9**, 95-108.

Türe M., Tokatlı F., Kurt Ü., Using Kaplan-Meirer Analysis Together With Decision Tree Methods (C&Rt, Chaid, Quest, C4.5 and Id3) In Determining Recurrence-Free Survival of Breast Cancer Patients, *Expert Systems With Applications*, 2008, **36** (2), 2017-2026.

Türkmenoğlu C., Türkçe Metinlerde Duygu Analizi, Yüksek Lisans Tezi, İstanbul Teknik Üniversitesi, Fen Bilimleri Enstitüsü, İstanbul, 2015, 389371.

Uçan A., Otomatik Duygu Sözlüğü Çevirimi ve Duygu Analizinde Kullanım, Yüksek Lisans Tezi, Hacettepe Üniversitesi, Fen Bilimleri Enstitüsü, Ankara, 2014, 379634.

URL-1: https://scikit-learn.org/stable/modules/svm.html, (Ziyaret tarihi: 5 Kasım 2019).

URL-2:https://developer.Twitter.com/en/docs/tweets/filter-realtime/guides/basic-stream-parameters, (Ziyaret tarihi: 12 Nisan 2018).

URL-3:https://medium.com/t%C3%BCrkiye/makine-%C3%B6%C4%9Frenmesi-nedir-20dee450b56e, (Ziyaret tarihi: 19 Kasım 2019).

Vatansever M., Görsel Veri Madenciliği Tekniklerinin Kümeleme Analizlerinde Kullanımı ve Uygulaması, Yüksek Lisans Tezi, Yıldız Teknik Üniversitesi, Fen Bilimleri Enstitüsü, İstanbul, 2008, 237170.

Yang W., Mu L., GIS Analysis of Depression Among Twitter Users, *Applied Geography,* 2015, **60**, 217-223.

Yıldırım S., Yıldız T., Türkçe İçin Karşılaştırmalı Metin Sınıflandırma Analizi, *Pamukkale Üniversitesi Mühendislik Bilim Dergisi,* 2018, **24** (5), 879-886.

Zhang H., Gan W., Jiang B., Machine Learning and Lexicon based Methods for Sentiment Classification: A Survey, *11th Web Information System and Application Conference,* 2014, 262-265.

KİŞİSEL YAYIN VE ESERLER

Uzun A., Zeybek H.İ., Bahadır M., Gürgöze S., **Zorba T.B.**, Chapter 133, Editörler: Arapgırlıoğlu H., Atik A., Hızıroğlu S., Elliott R., Atik D., *The Most Recent Studıes In Scıence And Art*, 2nd ed., Gece Kitaplığı, Ankara, 1714-1725, 2018.

ÖZGEÇMİŞ

Tuba Betül ÖZKAN (ZORBA) 1993'de Erzurum'da doğdu. Lise öğrenimini Bolu Anadolu Öğretmen Lisesi'nde tamamladı. 2011 yılında girdiği Kocaeli Üniversitesi Harita Mühendisliği Bölümü'nden 2015 yılında üçüncülük ile mezun oldu. Aynı yıl içerisinde Kocaeli Üniversitesi Jeodezi ve Jeoinformasyon Mühendisliği Anabilim Dalı'nda yüksek Lisans eğitimine başladı. 2019 yılında Tokat Gaziosmanpaşa Üniversitesi Mühendislik ve Doğa Bilimleri Fakültesi'nde başladığı araştırma görevlisi görevini halen sürdürmektedir.