



**KISA METİNLERDEN SOSYAL DUYGU
SINIFLANDIRMA İÇİN MAKİNE ÖĞRENMESİ
TABANLI YÖNTEMLERİN GELİŞTİRİLMESİ**

Fatma BAŞKAYA

**Yüksek Lisans Tezi
Bilgisayar Mühendisliği Anabilim Dalı
Danışman: Doç. Dr. İlhan AYDIN
AĞUSTOS-2017**

**T.C.
FIRAT ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ**

**KISA METİNLERDEN SOSYAL DUYGU SINIFLANDIRMA İÇİN
MAKİNE ÖĞRENMESİ TABANLI YÖNTEMLERİN GELİŞTİRİLMESİ**

YÜKSEK LİSANS TEZİ

Fatma BAŞKAYA

(131129115)

Anabilim Dalı: Bilgisayar Mühendisliği

Programı: Yazılım

Danışman: Doç. Dr. İlhan AYDIN

Tezin Enstitüye Verildiği Tarih: 26 Temmuz 2017

AĞUSTOS-2017

T.C.
FIRAT ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

KISA METİNLERDEN SOSYAL DUYGU SINIFLANDIRMA İÇİN
MAKİNE ÖĞRENMESİ TABANLI YÖNTEMLERİN GELİŞTİRİLMESİ

YÜKSEK LİSANS TEZİ
FATMA BAŞKAYA
(131129115)

Tezin Enstitüye Verildiği Tarih : 26 Temmuz 2017
Tezin Savunulduğu Tarih: 09 Ağustos 2017

Tez Danışmanı : Doç. Dr. İlhan AYDIN
Diğer Jüri Üyeleri : Doç. Dr. Murat KARABATAK
Yrd. Doç. Dr. Kazım HANBAY



AĞUSTOS- 2017

ÖNSÖZ

Bu çalışmada, makine öğrenmesi teknikleriyle duygu analizi yaklaşımı, yöntemleri, işlem adımları ve literatür çalışmaları incelenmiştir. Bu yönde yapılan çalışmalarla duygu analizi konusu ve alt çalışma alanlarıyla ilgili önemli iki farklı uygulama yapılmıştır. İlk çalışmada, öncelikle makine öğrenmesi teknikleriyle, duygu analizi konusuna temel oluşturan metin madenciliği ve metin sınıflandırma tabanlı haber metinlerinin farklı öznitelik türü ve yöntemleriyle sınıflandırılması konulu bir uygulama yapılmıştır. Uygulamada, farklı öznitelik türleri olan kelime kökü ve kelime seviye bigramlar kullanılmıştır. Farklı terim ağırlıklandırma yöntemleri ve farklı sınıflandırma algoritmalarıyla sistem test edilmiş ve metin sınıflandırma yaklaşımında en etkili yöntemler bulunmaya çalışılmıştır. İkinci uygulamada ise ana çalışma alanı olan duygu analizi konusu detaylarıyla ele alınmıştır. Bu anlamda Twitter ağının kullanıldığı iki ayrı veri seti üzerinde duygu analizi çalışması yapılarak gönderiler pozitif, negatif ve nötr olarak sınıflandırılmıştır. İki çalışmanın da kullanılan algoritmalarla başarılı sonuçlar verdiği ispatlanmıştır.

Çalışma süresince desteğiyle yanımda olan danışman hocam sayın Doç. Dr. İlhan Aydın'a teşekkür ederim.

Ayrıca, her anımda yanımda olan, bu zorlu yolculukta yalnız olmadığımı hissettiren desteğini, kalbini ve dualarını benden esirgemeyen aileme çok teşekkür ederim.

FATMA BAŞKAYA

AĞUSTOS - 2017

İÇİNDEKİLER

	<u>Sayfa No</u>
ÖNSÖZ	II
İÇİNDEKİLER.....	III
ÖZET	VI
SUMMARY	VII
ŞEKİLLER LİSTESİ	VIII
TABLolar LİSTESİ	X
KISALTMALAR	XI
SEMBOLLER LİSTESİ	XII
1. GİRİŞ	1
1.1. Duygu Analizi	3
1.2. Literatür Özeti	7
1.3. Tezin Amacı ve Kapsamı	10
1.4. Tezin Yapısı	11
2. METİN MADENCİLİĞİ.....	13
2.1. Metin ve Veri Madenciliği	14
2.2. Metin Madenciliği Adımları	14
2.3. Metin Madenciliği Uygulama Alanları	15
3. MAKİNE ÖĞRENMESİ TEKNİKLERİ İLE METİN SINIFLANDIRMA.....	16
3.1. Makine Öğrenmesi ve Metin Sınıflandırma.....	16
3.1.1. Denetimli Öğrenme	17
3.1.2. Denetimsiz Öğrenme.....	18

3.2. Metin Sınıflandırma Aşamaları	19
3.2.1. K En Yakın Komşu Algoritması (k-NN)	20
3.2.2. Naive Bayes (NB).....	22
3.2.3. Destek Vektör Makinesi (DVM)	23
3.2.4. Karar Ağacı Algoritması.....	24
3.2.5. Rastgele Orman Ağacı Algoritması (RO)	25
3.3. Metin Sınıflandırmanın Kullanım Alanları	26
3.4. Duygu Analizi ve Metin Sınıflandırma İlişkisi	26
4. UYGULAMA ÇALIŞMALARI	28
4.1. Uygulama Çalışması 1	29
4.1.1. Veri Seti Oluşturma	31
4.1.2. Metin Ön İşleme	31
4.1.2.a. Veri Temizliği.....	32
4.1.2.b. Dizge Parçalama	32
4.1.2.c. Durak Kelimeleri Çıkarma (Stopwords).....	32
4.1.2.d. Kök Bulma.....	33
4.1.3. Öznitelik Çıkarımı	34
4.1.4. Terim Ağırlıklandırma	36
4.1.5. Boyut İndirgeme	37
4.1.6. Otomatik Haber Sınıflandırma	38

Sayfa No

4.1.7. Başarım Ölçütleri.....	39
4.1.8. Uygulama Çalışması 1 Sonuçları	40
4.2. Uygulama Çalışması 2.....	41
4.2.1. Twitter Veri Setleri.....	42
4.2.2. Ön İşleme.....	43
4.2.3. Özellik Çıkarımı	45
4.2.4. Parçacık Sürü Optimizasyonu Tabanlı Sınıflandırma	46
4.2.5. Uygulama Çalışması 2 Sonuçları	48
5. SONUÇLAR	52
KAYNAKLAR.....	54
ÖZGEÇMİŞ	59

ÖZET

Teknolojinin gelişmesiyle, insan hayatı sanal dünyaya daha çok girmektedir. İnternetin gelişmesiyle birlikte ise Twitter, Facebook, Instagram, Tumblr, Google+, vb. gibi sosyal medya ağları insan hayatının vazgeçilmez parçası olmuştur. Bir günde milyonlarca mesajın dolaştığı bu ağlarda, insanlar hayatlarını arkadaşlarıyla, aileleriyle ve hatta tanımadığı insanlarla paylaşmakta ve bundan memnuniyet duymaktadırlar. İnsanların belli konular ve ürünler hakkında neler düşündüğünü ve nelerden hoşlanıp nelere ihtiyaç duyduğunu sosyal medya ortamlarında paylaşması ticaret, üretim ve hizmet sektörleri için ilgi çekici olmaya başlamıştır. Bu durum, Duygu Analizi konusunun ve bunu destekleyen Metin madenciliği ile Metin Sınıflandırma alanlarının önem kazanmasını sağlamıştır. Bu çalışmada yapılan ilk uygulama, makine öğrenmesi teknikleri ile haber metinlerinin farklı öznitelik ve terim ağırlıklandırma yöntemleriyle sınıflandırılması, yöntemlerin verimliliğinin ve başarısının test edilmesi açısından önem arz etmiştir. İkinci uygulama ile yine makine öğrenmesi teknikleriyle duygu analizi çalışması yapılarak, Twitter gönderilerini içeren farklı iki veri seti pozitif, negatif ve nötr sınıflarla etiketlenmiş ve Parçacık Sürü Optimizasyonu tabanlı K-En Yakın Komşu Algoritması ile sınıflandırılmıştır. Daha önce aynı veri kümeleri üzerinde önerilen guguk kuşu algoritması ile karşılaştırıldığında önerilen yöntemin daha başarılı sonuçlar verdiği gözlemlenmiştir.

Anahtar Kelimeler: Duygu Analizi, Makine Öğrenmesi, Metin Madenciliği, Metin Sınıflandırma, Sosyal Duygu Sınıflandırma, Öznitelik Çıkarımı, Terim Ağırlıklandırma.

SUMMARY

Development of Machine Learning Based Methods for Social Sentiment Classification from Brief Texts

With the development of technology, human life enters the virtual world more and more. With the development of the Internet, social media networks such as Twitter, Facebook, Instagram, Tumblr, Google+, etc. have become indispensable parts of human life. In these networks, where millions of messages are circulating in one day, people share their lives with their friends, family and even people they do not know and are happy with it. The sharing of what people think about specific topics and products and what they like and what they need is starting to become interesting for the trade, manufacturing and service sectors. This has made important Text Mining and Text Analysis as well as the Sentiment Analysis. In the first study, classification of news texts with different feature extraction methods and term weighting methods which based machine learning methods has been important in terms of testing the efficiency and success of the methods. With the second study, two different data sets including Twitter posts are also classified by positive, negative, and neutral classes and the nearest neighbor algorithm based on particle swarm optimization, by conducting sentiment analysis with machine learning techniques. It has been observed that the proposed method yields more successful results when compared to the cuckoo algorithm proposed previously on the same data sets.

Key Words: Sentiment Analysis, Machine Learning, Text Mining, Text Classification, Social Emotion Classification, Feature Extraction, Term Weighting.

ŞEKİLLER LİSTESİ

	<u>Sayfa No</u>
Şekil 1.1. Bir ürün hakkındaki kullanıcı görüşleri	4
Şekil 1.2. Duygu analizi evrimi	5
Şekil 1.3. Duygu analizi yöntemleri	6
Şekil 1.4. Literatürde izlenen işlem adımları.....	8
Şekil 2.1. Metin madenciliği süreci	13
Şekil 2.2. Metin madenciliği adımları	14
Şekil 3.1 Makine öğrenmesi ile metin sınıflandırma.....	16
Şekil 3.2 MÖ teknikleri kullanılarak yapılan MS türleri.....	17
Şekil 3.3. Denetimli öğrenme modeli (a), denetimsiz öğrenme modeli (b)	18
Şekil 3.4. Metin sınıflandırma aşamaları.....	19
Şekil 3.5. MS alanında yaygın olarak kullanılan bazı algoritmalar	20
Şekil 3.6. İki boyutlu örnekler (a), Gelen yeni veri (b)	21
Şekil 3.7. Yeni veriye en yakın 3 komşu üye (a), Yeni üyenin sınıflandırılması (b)	21
Şekil 3.8. DVM ile ikili sınıflandırma.....	23
Şekil 3.9. Karar ağacı örneği	24
Şekil 3.10. Metin sınıflandırma (a), Duygu analizi (b)	27
Şekil 4.1. Uygulama Çalışması 1 modeli	29
Şekil 4.2. Uygulama Çalışması 1 işlem adımları.....	30
Şekil 4.3. Temel ön işleme adımları	31
Şekil 4.4. Türkçe durak kelimeler	32
Şekil 4.5. Zemberek kök aday bulucu	33
Şekil 4.6. Örnek metin verisi (a) , Metnin etkisiz kelime ve noktalama işaretlerinden temizlenmesi (b)	34
Şekil 4.7. Bazı öznitelik çıkarım yöntemleri	34
Şekil 4.8. Kelime kökü öznitelik çıkarımı (a), Kelime seviye n-gram (b)	36
Şekil 4.9. Önerilen yöntemin akış şeması	42
Şekil 4.10. Ön işleme aşaması işlem adımları	43

Şekil 4.11. İngilizce durak kelimeler	44
Şekil 4.12. Örnek tweet üzerinden ön işleme adımları	45
Şekil 4.13. Çoklu PSO tabanlı ağırlıklı k-NN	46
Şekil 4.14. PSO tabanlı ağırlıklı k-NN sınıflandırma performansı	49
Şekil 4.15. Twitter veri kümesi-1 için ROC eğrisi	50
Şekil 4.16. İki veri kümesi üzerinde farklı yöntemlerin karşılaştırılması	51



TABLÖLAR LİSTESİ

Sayfa No

Tablo 1.1. Örnek cümlelerde yer alan anlamsal ifadeler	3
Tablo 1.2. Duygu analizi ile ilgili literatürde yapılan bazı çalışmalar	9
Tablo 4.1. Uygulama Çalışması 1 veri seti detayları.....	31
Tablo 4.2. Kelime seviye N-gram kullanımları	35
Tablo 4.3. Çalışmada kullanılan öznitelik türü ve sayıları	35
Tablo 4.4. Terim frekansı oluşturma	37
Tablo 4.5. Boyutu indirgenen veri.....	37
Tablo 4.6. Çalışmada kullanılan yöntem kombinasyonları	38
Tablo 4.7. Öznitelik yöntemlerine ve TA seçimine göre sınıflandırma başarı oranları	38
Tablo 4.8. Sınıflandırma başarımlar ölçütleri	39
Tablo 4.9. Kullanılan twitter veri kümelerinin özellikleri.....	43
Tablo 4.10. Çıkarılan 9 öznitelik	45
Tablo 4.11. Önerilen PSO'nun parametreleri	49
Tablo 4.12. Veri kümesi-1 için karmaşıklık matrisi	49
Tablo 4.13. Veri kümesi-2 için karmaşıklık matrisi	50

KISALTMALAR

DA	: Duygu Analizi
MS	: Metin Sınıflandırma
MÖ	: Makine Öğrenimi
DDİ	: Doğal Dil İşleme
MM	: Metin Madenciliği
TA	: Terim Ağırlıklandırma
Tf	: Terim Frekansı
Idf	: Ters Doküman Frekansı
$Tf-Idf$	: Terim Frekansı-Ters Doküman Frekansı
k-NN	: K En Yakın Komşu sınıflandırması
DVM	: Destek Vektör Makineleri
NB	: Naive Bayes
RO	: Rastgele Orman Algoritması
PSO	: Parçacık Sürü Optimizasyonu
KA	: Karar Ağacı
GKA	: Guguk Kuşu Algoritması

SEMBOLLER LİSTESİ

N	: Parçacık sayısı
K	: Sınıf sayısı
c_1, c_2 ve c_3	: Sabit ağırlık faktörlerini
χ	: Kısıtlama faktörü
W	: Moment ağırlığını
p_i^t	: i. parçacığın en iyi pozisyonu
ψ_1, ψ_2 ve ψ_3	: [0,1] aralığında rastgele üretilen sayılar
$\{g\}$	: Doğru bir şekilde sınıflandırılan örnek sayısı
$\{b\}$	: Yanlış olarak sınıflandırılan örnek sayısını
d_{ij}	: Bir parçacık ile en yakın örnek arasındaki mesafe

1. GİRİŞ

Facebook, Instagram, Twitter, vb. gibi sosyal ağ platformlarının günlük yaşam içerisinde kabul edilmesi insanların çevrimiçi haberleşme biçimini değiştirmiştir [1]. İnsanlar, görüşlerini ve hissettiği duyguları, internet üzerinden dilediği zaman ve dilediği yerden paylaşabilme imkânına sahip olmakta ve tecrübe, duygu ve görüşlerini anında yüzlerce kişiye aktarabilmektedirler. Bu duruma olanak sağlayan sosyal medya ağları böylece bilginin internette paylaşımı konusunda çok güçlü ve önemli araçlar olmaya başlamışlardır. Politikacılar, hükümet kuruluşları ve tüketiciler gibi sosyal medya kullanıcıları her an çeşitli konulardaki duygularını ifade ederek büyük miktarda veri oluşumuna katkı sağlamaktadırlar [2].

İnsanların belirli bir konu, marka, ürün, film, kitap, vb. gibi herhangi bir alanda neler düşündüğünü sosyal medya ağlarında paylaşması birçok sektör için ilgi çekici olmaya başlamıştır. Artık firmalar kendileri ile ilgili neler düşünüldüğünü bilmek ve bu düşüncelere göre hareket etmek istemektedirler. Örneğin, çekilecek bir sinema filminin başarılı olması için filmin konusu ve oyuncu kadrosu hakkında sinema izleyicilerinin düşünceleri göz önünde bulundurulmak istenmektedir. Bunun için ise sosyal medya ağları veya blog sitelerinde bulunan izlenme reytingleri veya izleyici yorumları kullanılmaktadır. Öte yandan siyasi bir parti sosyal medya ağlarından halkın düşünce ve tepkisini öğrenip yeni politikalar üretebilmekte veya politikalarında değişikliğe gidebilmektedir [3]. Tüm bu olanaklar ve insanların kişisel düşüncelerini, ilgi duydukları alanlar hakkındaki duygularını paylaşmak istemesi sosyal ağları güçlü birer platform haline getirmiştir [4].

Sosyal medya, halkın duygularını temsil etmenin hayati bir platformu olarak ortaya çıkmış ve duygu analizi alanındaki veri madenciliği gereksinimlerini artırmıştır [1]. Yapılan paylaşımlar ile kişilerin bir konu hakkındaki düşünce ve görüşlerini sosyal medya üzerinden paylaşması, sosyal medyayı ekonomi, pazarlama, politika, vb. araştırma alanlarında kullanılabilecek zengin veri kaynağı durumuna getirmekte ve bu yorumların sınıflandırılmasını sağlayan duygu analizi alanının önem kazanmasını sağlamaktadır [5].

Teknolojinin gelişmesiyle birlikte insanların sanal dünyaya bu denli bağımlı hale gelmesi interneti daha büyük bir veri merkezi hâline getirmektedir. Bu durum aslında avantajlı birçok sonuç doğursa da olumsuz durumları da beraberinde getirmektedir.

Çok büyük miktarda işlenmiş ve işlenmemiş verilerin hızlı bir şekilde [6], internet üzerinden akması, bu büyüklüğe bağlı olarak verilere karmaşık bir yapı kazandırmaktadır. Karmaşıklaşan veri ile istenilen bilgiye ulaşmak gittikçe önemli bir problem haline gelmektedir. Bu probleme çözüm bulmanın kullanışlı yollarından biri karmaşık halde bulunan bu verileri metin madenciliği yöntemleri ile kategorilere ayırmaktır. Elde edilen kategorilendirme ile bilgi üzerinde anlamsal benzerlikler oluşturarak bilgiye, etkili ve hızlı bir şekilde ulaşılabilir. Bu tür veriler, çeşitli uygulamalarda Veri Madenciliği ve Metin Madenciliği teknikleri kullanılarak analiz edilebilir. Örneğin, müşteri görüşlerine ilişkin büyük miktarda veriyi analiz etmek oldukça zordur ve genelleştirilmiş bir görüş özeti elde etmek için mevcut yaklaşımlara ihtiyaç duyulur [6].

Kamuoyunun ve tüketicilerin toplumsal olaylar, siyasi hareketler, şirket stratejileri, pazarlama kampanyaları, ürün tercihleri konusundaki görüşlerini anlamak için araştırma toplulukları ve akademisyenler son yıllarda titizlikle çalışmaktadırlar. Duygu analizi (DA), bir varlığa yönelik, metinlerde ifade edilen düşüncelerin, duyguların ve tutumun hesaplamalı bir çalışmasıdır. DA, politika, pazarlama piyasası ve müşteri memnuniyeti ölçümü, film satış tahminleri ve daha birçok konuda kamuoyunu gözlemlemek gibi çeşitli hedeflere ulaşmak için kullanılan bir araştırma ve inceleme alanıdır. Dolayısıyla, tüketicilerin davranışlarını ve toplumun eğilimlerini öğrenmek için halkın görüşünü analiz etmektedir [6].

Sosyal medya ağlarının güçlü platformlarından biri olan Twitter'ın 255 milyon aylık aktif kullanıcısı bulunmaktadır [6]. Bu nedenle, metin altında yatan görüşleri ortaya çıkarmak için iyi bir kaynak olarak hizmet etmektedir. Twitter'dan başka, çeşitli topluluk blogları ve Facebook gibi mecralar da duygu analizi için zengin veri kaynakları sunmaktadırlar.

Bir marka, ürün, siyasi görüş veya kişi hakkında yapılan olumlu ve olumsuz milyonlarca yorumun değerlendirilerek analiz edilmesi ve kamuoyu algısının ölçülmesi duygu analizini hemen hemen her sektör için vazgeçilmez bir alan haline getirmiştir. Hangi firma, politikacı, siyasi parti, marka veya film yapımcısı kamuoyu algısını ölçerek insanların yaptıkları iş hakkında ne düşündüğünü bilerek adımlarını daha doğru bir şekilde atmak ya da adımlarını düzeltmek istemez ki? Bu ve benzeri avantaj sağlayan birçok sebepten dolayı duygu analizi her alanda çağımızın gerekliliği haline gelmiştir.

1.1. Duygu Analizi

Duygu veya görüş analizi, kişilerin, ürünlere, organizasyonlara, bireylere ve diğer konulara yönelik öznel görüş ve duygularının analiz edilmesi ve sınıflandırılmasıyla ilgilenen bir analiz dalıdır [7] ve insanların eğilimlerine ulaşmayı mümkün kılar [8]. Duygu analizi, bir metin içinde geçen kelime ve kelime gruplarını analiz ederek metnin barındırdığı duyguyu ortaya çıkarmaktadır.

Duygu sınıflandırma ise bir konu hakkında yazılmış ifadelerin analiz edilmesiyle yazarın sahip olduğu duyguyu genellikle olumlu, olumsuz veya tarafsız gibi kategorilere sınıflandırmayı amaçlayan ve duygu analizinin alt dalı olarak ele alınan bir çalışma alanıdır [7]. Metinlerin altında yatan duygu ve görüşün olumlu, olumsuz veya tarafsız olup olmadığını belirlemeye yarayan [5] ve doğal dil işleme ile metin madenciliği alanlarının alt dalı olan bir analiz metodudur [7]. Örneğin, sosyal medyada paylaşılan gönderiler ham veri olarak adlandırılmakta ve bu gönderiler üzerinde anlamsal bir etiket bilgisi bulunmamaktadır. Dolayısı ile bir konu ile ilgili düşüncenin ne olduğu bilgisi sosyal medyada bulunan verinin içerisinde gizlidir. Tablo 1.1’de örnek cümlelerde metin içinde yatan duygular verilmiştir.

Tablo 1.1. Örnek cümlelerde yer alan anlamsal ifadeler

Örnek Cümle	Anlamsal İfade
Yemek çok lezzetliydi.	Pozitif Anlam
Yediğim en kötü pastaydı.	Negatif Anlam
Ürünün özellikleri iyi değil.	Negatif Anlam
Çok kaliteli bir filmidir.	Pozitif Anlam

Görüldüğü gibi duygu analizi gizli olan anlamsal bilgileri ortaya çıkararak, konuyla ilgili görüş sınıflandırması yapar ve konu hakkındaki görüşün yüksek oranda ne olduğunu ortaya çıkarır [9]. Şekil 1.1’de görüldüğü gibi herhangi bir konu hakkında farklı kullanıcılar farklı görüş ve duyguya sahip olabilir. Bu görüş farklılığı kullanılarak o konu hakkındaki genel düşüncenin ne olduğu yine duygu analizi ile saptanabilir.

Duygu analizi günümüzde ticari şirketler, politikacılar, basın-yayın kuruluşları, kamu kurumları ve güvenlik teşkilatları gibi birçok sektör tarafından [10] pazarlama analizi, bilimsel, tıbbi ve medikal araştırmalar, suç tespiti, siyasi görüş belirleme ve sosyolojik anket çalışmalarında yaygın bir şekilde kullanılmaktadır.



Şekil 1.1. Bir ürün hakkındaki kullanıcı görüşleri

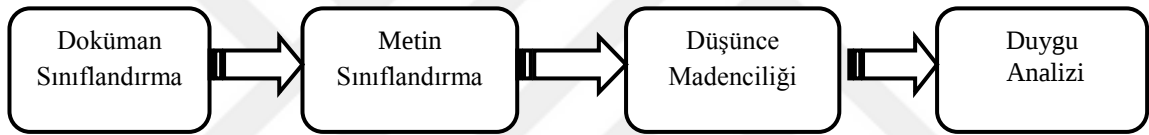
Bu alanlara olan ihtiyacın oldukça fazla olması ve sosyal medya kullanımına olan ilginin her geçen gün artması, duygu analizi çalışmalarının önemini ortaya koymaktadır [11]. Bu duruma, hem kullanıcıların görüşlerini paylaşıp sonuç elde edebileceklerini bildikleri birçok sosyal ağ ortamı ve platformun bulunması; hem de yorumlanan marka veya ürünün, internetin ve sosyal medyanın günümüzdeki gücüyle yayılabilecek olumsuz bir durumu göze almaktan çekinmesi neden olmaktadır.

Sosyal medya ağları kullanılarak yapılan duygu analizinde en zengin veriyi, en popüler sosyal ağlardan biri olan ve günde yaklaşık 500 milyon gönderiyi içeren [8] 255 milyon aylık aktif kullanıcısıyla Twitter [6] sunmaktadır. Twitter, kayıtlı kullanıcıların kısa mesajlar gönderebilmesine olanak tanıyan çok popüler bir sosyal ağ sitesidir ve Twitter veritabanı, sosyal ağ platformları içindeki en büyük veritabanlarından biridir [12]. Kullanıcılar Twitter üzerinde 140 karakterle sınırlı bir paylaşımına sahip olduğundan görüşlerini emoji adı verilen ve duygu ifade eden semboller kullanarak daha yoğun bir şekilde ifade etmektedirler. Bu nedenle, herhangi bir konuyla ilgili duygu ve görüş analizi yapmak için yoğun duygu ve güçlü bakış açısına sahip olan *tweet*'ler en kullanışlı araçlardır.

İnsanlar birebir yaptığı konuşmalarda diyalogda bulunduğu kişinin jest ve mimiklerinden veya ses tonundan ne tür duyguya sahip olduğunu kolay anlayabilmektedir. Ancak, metin türündeki veride böyle bir tanımlama mümkün olmamaktadır. Çünkü bu bilginin ve altında yatan anlamın bilgisayar tarafından kendine has özellikleriyle anlaşılması ve tanınması gerekmektedir [13].

Sosyal ağda paylaşılan veriler, işlenmemiş haliyle yapısal olmayan bir formdadır ve dolayısıyla üzerinde çalışılması oldukça zordur. Ayrıca veriler incelediğinde birçok sözcüğün doğru olmayan kelimeler ve kısaltmalardan oluştuğu görülmüştür.

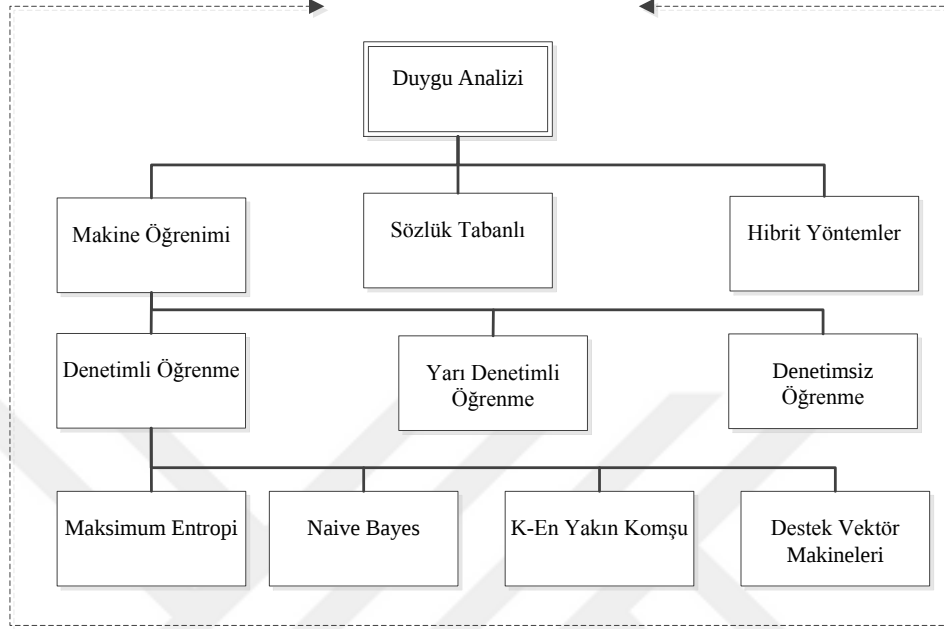
Verinin sınıflandırmaya hazır yapısal forma dönüşmesi için Doğal Dil İşleme (DDİ) ve Metin Madenciliği (MM) yöntemleri ile işlenmesi gerekmektedir. Dolayısıyla, bu durum bir DDİ problemi olarak ele alınarak öncelikle metinsel veri, yapısal olmayan ifadelerden temizlenmelidir. Yapısal hale dönüşen metinsel veri, hatalardan arındırılan ve sınıflandırmaya hazır veri anlamına gelir. Böylece, bu tür metinlerin altında yatan duygunun türü belirlenebilmektedir. Dolayısıyla, bir duygu analizi problemi birçok disiplini bünyesinde barındırdığından dolayı doküman sınıflandırma, metin sınıflandırma ve düşünce madenciliği alanlarından türeyerek ortaya çıkmıştır. Bu durum Şekil 1.2’de şematik olarak verilmiştir [15].



Şekil 1.2. Duygu analizi evrimi

Duygu analiz yöntemleri, Şekil 1.3'te belirtildiği gibi, makine öğrenme tabanlı yöntemler, sözlük tabanlı yöntemler ve hibrit yöntemler olarak sınıflandırılmaktadır. Bu yöntemler içerisinde, literatürde yapılan çalışmalar göz önüne alındığında makine öğrenimi tabanlı yöntemlerin maksimum doğruluk verdiği görülmüştür [6]. Makine öğrenme (MÖ) tabanlı yöntemler ise denetimli, yarı denetimli ve denetimsiz olmak üzere üçe ayrılmaktadır. Denetimli (gözetimli) yöntem ise analiz çalışmalarında sıkça kullanılan bir alandır ve bu yaklaşım için, hangi sınıfa ait oldukları belli, alan ve etiket bilgisine sahip olan eğitim verilerin olması gerekmektedir [8]. Denetimli öğrenme için yaygın olarak uygulanan sınıflandırıcılardan bazıları Maksimum Entropi, Naive Bayes (NB), K-en yakın komşu (k-NN) ve Destek Vektör Makineleri (DVM) dir [8]. Bu algoritmalar, sınıflandırma yapmak için kullanılan zamanın minimize edilmesi ve güvenilir sonuçlar vermeleri açısından oldukça yaygın olarak kullanılmaktadırlar. Ayrıca sınıflandırılmak istenen veriler bu algoritmalar kullanılarak otomatik olarak sınıflandırılabilirler.

Ancak bir sonraki bölümlerde detaylı bir şekilde anlatılacağı gibi sınıflandırma aşamasına geçmeden önce metnin bazı ön işlemlerden geçmesi gerekmektedir.



Şekil 1.3. Duygu analizi yöntemleri

Sosyal medya ağları kullanılarak yapılan duygu analizi:

- Bir ürün veya marka ile ilgili yorumların duygularının saptanması,
- Toplumun ruh durumunun tespit edilmesi,
- Firmaların sosyal medya mesajlarından kamuoyu algılarını ölçmeleri,
- Film değerlendirmeleri,
- Pazar-fiyat dengesi
- Borsa tahminleri,
- Bilimsel ve medikal araştırmalar,
- Suç analizi, güvenlik, istihbarat, vb. gibi hem genel hem de özel anlamda birçok alanda kullanılarak kendini göstermeye başlamıştır. Duygu analizi günümüzde kâr amaçlı şirketler, siyasetçiler, medya kuruluşları, kamu kurumları ve güvenlik teşkilatları gibi birçok sektör tarafından yaygın olarak kullanılmaktadır [18].

1.2. Literatür Özeti

Duygu analizi ile ilgili literatürde çeşitli alanlara ve amaçlara yönelik çalışmalar yapılmıştır. Türkçe dili için yapılan çalışmalarda metin sınıflandırma yöntemleri kullanılarak hem sözlük tabanlı hem de makine öğrenimi yöntemleri ile ilgili farklı çalışmalar yapılmıştır. Ancak, Türkçeye ait bir sözlük son zamanlara kadar olmadığından daha çok makine öğrenimi metotları üzerine yoğunlaşmıştır.

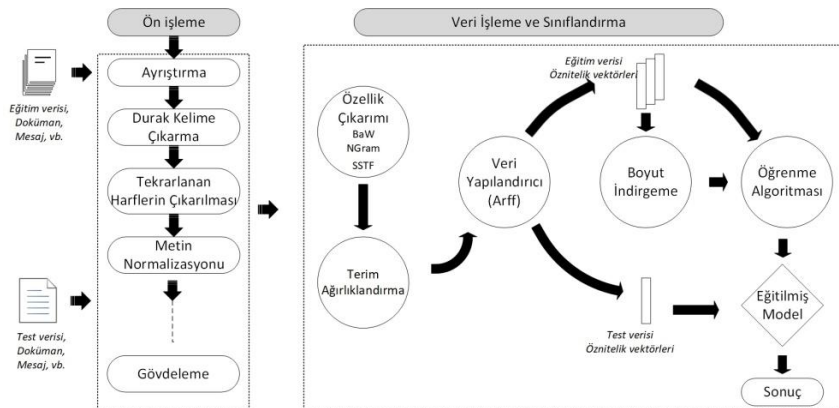
Metin sınıflandırma alanında, Diri ve Amasyalı [19], 18 yazarın gazetede yayımladıkları köşe yazılarından elde ettikleri metinsel veri üzerinde elde ettikleri 22 öznitelik içerisinde en etkili 4 özniteliği kullanarak, *Yapay Sinir Ağları* ile yazar sınıflandırmaya çalışmışlardır. Yapılan yazar tespitinde %84 başarı oranı elde etmişlerdir. Kaşıkçı ve Gökçen [20], e-ticaret sitelerini belirlemek için yaptıkları çalışmada, 273 e-ticaret sitesini inceleyerek 110 kelimelik bir kelime sözlüğü oluşturmuşlardır. Herhangi bir sitenin e-ticaret sitesi olup olmadığını ise listedeki kelimeler ile yaptıkları otomatik karşılaştırma ile sınıflandırmışlardır. Böylece *Naive Bayes* algoritmasıyla %85.30 başarı elde etmişlerdir. Güran ve diğerleri [21], internet ortamından elde ettikleri 5 farklı kategoriye ait metinleri *N-gram* öznitelik yöntemiyle sınıflandırmışlardır. Unigram, bigram ve trigram yöntemleri sonucu yaptıkları sınıflandırmada, en yüksek başarıya unigram modelde ve %95.83 *Multinom Basit Bayes* ile ulaşmışlardır.

İngilizce için yapılan metin sınıflandırma çalışmaları yine Türkçede olduğu gibi farklı amaçlar için kullanılmıştır. Nigam ve ekibi [22], Reuters gibi 3 yaygın haber veri kümelerini kullanarak *Beklenti-Maksimizasyonu* adı verilen algoritma ile eğitim aşamasında etiketi belli olmayan veriler ile çalışmışlardır. Lin ve ekibi [23], metin sınıflandırma için yeni bir yöntem önermiş, *k-NN* ve *k-means* ile test etmişlerdir. RCV1, Reuters-8 ve WebKB veri setleri üzerinde yapılan sınıflandırmanın başarılı sonuçlar verdiği gözlemlenmiştir. Ko [24], farklı *Terim Ağırlıklandırma* yöntemlerini karşılaştırarak yaptığı metin sınıflandırma çalışmasında ve önerdiği yöntemle başarılı sonuçlar elde etmiştir.

Literatürde duygu analizi alanında yapılan çalışmalarda en etkili sonuçlarının alındığı makine öğrenimi yöntemine Pang ve Lee öncülük etmiştir. Pang ve Lee [25], IMDb.com internet sitesinden topladıkları film görüşlerini duygu analizi için kullanmış ve en iyi başarı sonucuna %82.9 ile DVM ile ulaşmışlardır. Pak ve ekibi [26] 300.000 adet Twitter gönderisini etiketsiz olarak sınıflandırmışlardır.

N-gram öznitelik çıkarımı ve *Terim Frekansı* ağırlıklandırma yöntemini kullanarak DVM ile yapılan sınıflandırmada %61 başarı oranı elde etmişlerdir. Saif ve ekibi [27], Twitter ağından alına 60.000 adet gönderiyi olumlu ve olumsuz şekilde etiketleyerek bir eğitim kümesi oluşturmuşlardır. N-gram öznitelik çıkarım yöntemi ve naive bayes sınıflandırıcı ile 1.54 milyon gönderiyi sınıflayarak %86.3 doğruluk oranı tespit etmişlerdir. Habernal ve ekibi [28] Facebook üzerinden aldıkları 10.000 adet gönderiyi denetimli makine öğrenimi sınıflandırması yapmak için olumlu, olumsuz ve nötr olarak etiketlemişlerdir. *N-gram*, *PoS Tags* ve *Tf-Idf* yöntemleriyle olumlu, olumsuz ve nötr sınıflamada en iyi %69 başarı değerini elde etmişlerdir. Basari ve ekibi [29], EMOT [30] veri setinde metinsel veri temizliği yaptıktan sonra *Parçacık Sürü Optimizasyonu* ve DVM ile %76.20'lik bir doğruluk oranı elde etmişlerdir.

Türkçe için yapılan duygu analizi çalışmalarında, Eroğul [31], film yorumları üzerinden yaptığı olumlu ve olumsuz görüşleri farklı makine öğrenmesi yöntemlerini uygulayarak %85 başarı elde etmiştir. Şimşek ve Özdemir [32], borsa iniş-çıkışının Twitter üzerinden atılan tweetler ile ilgisini araştırmak için, kızgınlık, üzüntü, aşk, korku, tikslenme, utanç, sevinç ve şaşkınlık duygularına ait 113 adet özniteliği mutlu ve mutsuz olarak sınıflandırmışlardır. Kaya ve diğerleri [33], politika haberleri üzerinde olumlu-olumsuz yorum tespitinde bulunmuş ver N-gram öznitelikle yapılan sınıflandırmada %65 ve %77 aralığında başarı elde etmişlerdir. Türkmenoğlu ve Tantuğ [34], tweetler ve film yorumları üzerinde yaptıkları duygu analizi çalışmasında makine öğrenimi yönteminin, sözlük tabanlı yöntemden daha etkili olduğunu göstermişlerdir. Çoban ve ekibi [35], Türkçe Twitter gönderilerinden oluşan verileri olumlu ve olumsuz bir şekilde sınıflandırarak en başarılı sonuca %66 ile *Multinom Naive Bayes* ile ulaşmışlardır.



Şekil 1.4. Literatürde izlenen işlem adımları [35]

Çoban ve ekibinin çalışmada izlediği işlem adımları Şekil 1.4’te verilmiştir [35]. Duygu analizi alanında farklı dillerde ve farklı amaçlar doğrultusunda yapılan birçok çalışma bulunmaktadır. Tablo 1.2’de yapılan çalışmalar kategorize edilmiştir.

Tablo 1.2. Duygu analizi ile ilgili literatürde yapılan bazı çalışmalar

Yazar	Konu	Dil	Yöntem	Referans
Diri ve Amasyalı	MS Yazar Tanıma	Türkçe	Yapay Sinir Ağları	[19]
Kaşıkcı ve Gökçen	MS E-Ticaret Sitelerinin Belirlenmesi	Türkçe	BoW Kelime Çantası Naive Bayes	[20]
Güran vd	Metin Sınıflandırma	Türkçe	N-gram Multinom Naive Bayes	[21]
Nigam ve ekibi	Etiketli ve Etiketsiz Dokümanlar ile Metin Sınıflandırma	İngilizce	BoW Beklenti Maksimizasyonu	[22]
Lin ve ekibi	MS ve Kümeleme için Yeni Bir Benzerlik Metriği	İngilizce	BoW k-NN, k-Means	[23]
Ko	Terim Ağırlıklandırma Yöntemlerinin Karşılaştırılması	İngilizce	BoW k-NN, DVM	[24]
Pang ve Lee	Film Görüşleri Üzerine DA	İngilizce	Makine Öğrenimi DVM	[25]
Pak ve ekibi	Twitter Gönderileri Üzerine DA	İngilizce	N-gram Terim Frekansı DVM	[26]
Saif ve ekibi	Twitter DA	İngilizce	N-gram Naive Bayes	[27]
Habernal ve ekibi	Facebook Gönderileri Üzerine DA	İngilizce	Makine Öğrenimi N-gram, PoS Tags ve TFIDF	[28]
Basari ve ekibi	Film Yorumları Üzerine Düşünce Madenciliği	İngilizce	Parçacık Sürü Optimizasyonu DVM	[29]
Eroğul	Film Görüşleri Üzerine DA	Türkçe	Makine Öğrenimi	[31]
Şimşek ve Özdemir	Twitter Ekonomi Tweetlerinin Borsaya Etkisi	Türkçe	-	[32]
Kaya ve diğerleri	Politika Haberleri Üzerine DA	Türkçe	Ngram DVM	[33]
Türkmenoğlu ve Tantuğ	Twitter ve Film Yorumları Üzerine DA	Türkçe	N-gram, BoW, DVM	[34]
Çoban ve ekibi	Twitter Gönderileri Üzerine DA	Türkçe	N-gram Multinom Naive Bayes	[35]

1.3. Tezin Amacı ve Kapsamı

Bu tez çalışmasının amacı, sosyal medyada bulunan ve genelde etiketsiz bir şekilde bir anlam ifade etmeyen metinsel verilerin altında yatan kullanıcı duygusunu, makine öğrenmesi tekniklerine dayanan, metin madenciliği, metin sınıflandırma ve duygu analizi teknikleri ile belirlemektir. Bu çalışmada, en aktif sosyal medya platformlarından biri olan Twitter verileri, metin sınıflandırma ve metin madenciliği tekniklerine dayanarak yapısal forma dönüştürüldükten sonra makine öğrenmesi teknikleriyle sınıflandırılarak duygu tespitinin yapıldığı duygu analizi çalışması hedeflenmiştir. Hedeflenen sosyal duygu analizi ile, var olan yorumların içerisindeki olumlu, olumsuz veya tarafsız (duygu içermeyen, nötr) yorumlar en iyi başarı oranıyla tespit edilip, bu üç kategoride istenen veri çıkarımı yapılmıştır.

Açıklanan hedefler doğrultusunda ilk çalışmada 5 farklı kategoriden alınarak oluşturulan Türkçe haber metinleri, farklı öznitelik, terim ağırlıklandırma yöntemleri ve gözetimli makine öğrenimi ile sınıflandırılmıştır. Yapılan çalışma, yöntemlerin verimliliğinin ve başarısının test edilmesi açısından önem arz etmiştir. İkinci çalışmada ise 2 farklı veri setine ait Twitter gönderileri altında yatan olumlu, olumsuz ve nötr duyguların belirlenmesi amaçlanmıştır. Veriler metin sınıflandırma teknikleri ile yapısal forma dönüştürüldükten sonra *Parçacık Sürü Optimizasyonu* (PSO) tabanlı k-NN ile sınıflandırmak için bir yöntem önerilmiştir. Önerilen çalışma, veri seti üzerinde farklı yöntemlerle karşılaştırılmış ve yöntemin doğruluğu kanıtlanmıştır.

Yapılan çalışmalar metin madenciliği, metin sınıflandırma ve duygu analizi konularına, kullanılan farklı yöntem ve algoritmaların karşılaştırılması ve başarı oranlarının test edilmeleri açısından katkı sağlamıştır.

Bu tez çalışması kapsamında yapılan akademik yayın çalışmaları aşağıda verilmiştir [36-37].

- “Haber Metinlerinin Farklı Metin Madenciliği Yöntemleriyle Sınıflandırılması”, International Artificial Intelligence and Data Processing Symposium (IDAP), Türkiye.
- “PSO Tabanlı Ağırlıklandırılmış K-EYK ile Duygu Sınıflandırma”, Uluslararası Bilgisayar Bilimleri ve Mühendisliği Konferansı (UBMK), Türkiye.

1.4. Tezin Yapısı

Bu tez çalışmasında metin sınıflandırma ve duygu analizi tekniklerinin Türkçe ve İngilizce metinler üzerinde birlikte uygulanabilirliği araştırılmış ve bu problemin üstesinden gelebilmek adına gözetimli makine öğrenimi yöntemi ve farklı özellik çıkarımı yöntemlerinin uygulandığı bir metin sınıflandırma çalışması ve pozitif, negatif ve nötr etiketli Twitter verileri üzerinde Parçacık Sürü Optimizasyonu tabanlı k-NN duygu sınıflandırması yapan duygu analizi çalışması yapılmıştır.

Bu tez çalışması beş bölümden oluşmaktadır. Giriş bölümünde, tez çalışması kapsamında ele alınan metin sınıflandırma ve duygu analizi konusu ile ilgili problemlerden bahsedilerek bu problemlerin çözümü için geliştirilen yöntemler olduklarına yer verilmiştir. Duygu analizine duyulan ihtiyacından önemine, günümüz için vazgeçilmez bir alan olduğuna ve bu önemi tetikleyen başlıca sebeplere değinilmiştir. Literatür çalışması yapılarak, duygu analizini tanımı ve kullanım alanları ile ilgili genel bilgiler verilmiştir. Bu anlamda metin sınıflandırma ve duygu analizi çalışmalarıyla ilgili daha önce literatürde yapılan yayınlar ve çalışmalar hakkında bilgi verilmiştir. Ayrıca tez çalışmasının amacı ve bu tez çalışması kapsamında yapılan yayınlara yer verilmiştir.

İkinci bölümde, metin üzerinde metin madenciliği konusuna yer verilmiştir. Ayrıca, metin madenciliğinin uygulanışı, kullanım alanları ve işlem adımları ile ilgili detaylardan bahsedilerek veri madenciliği ile olan ilişkisine yer verilmiştir. Metin madenciliğinin metin işleme için gerekli bir alan olduğuna ve doğal dil işleme alanının alt disiplini olduğundan bahsedilmiştir. Ayrıca metin madenciliği ve doğal dil işleme adımları olmadan sağlıklı bir metin sınıflandırma yapılamayacağının, tüm bu adımların metnin yapısal forma dönüşmesi için gerekli olduğunun altı çizilmiştir.

Üçüncü bölümde, metin sınıflandırma, yöntemleri, kullanılan algoritmalar ve işlem adımları hakkında bilgiler verilmiştir. Ayrıca, duygu analizi ve metin sınıflandırma alanları arasındaki ilişkiye yer verilmiştir. Makine öğrenimi yöntemleriyle yapılan metin sınıflandırma türlerine yer verilerek tez çalışmasında denetimli öğrenme modeli tabanlı olan etiketli veriler üzerinde uygulamalar geliştirileceğinden bahsedilmiştir. Bazı önemli metin sınıflandırma algoritmaları ile ilgili kapsamlı bilgilere ve eşitliklere yer verilmiştir.

Dördüncü bölümde, tez çalışması kapsamında yapılan uygulama çalışmalarına yer verilmiştir. İlk uygulama çalışmasında haber metnlerinin farklı yöntemlerle sınıflandırılması sağlanmış ve ilgili çalışma ile ilgili işlem adımları adım adım açıklanarak her adımın öneminden bahsedilmiştir. İkinci uygulama çalışmasında ise İngilizce diline ait Twitter veri seti üzerinde duygu analizi çalışmasına yer verilmiştir. Uygulamalarda işlenen işlem adımları, yapılan sınıflandırma türleri ve kullanılan yöntemler detaylı olarak anlatılmıştır. Ayrıca kullanılan veri setleri ile ilgili detaylar sözel olarak ve tablolarla açıklanmıştır. Kullanılan yöntemler ve sınıflandırma algoritmaları karşılaştırılarak test edilmiştir.

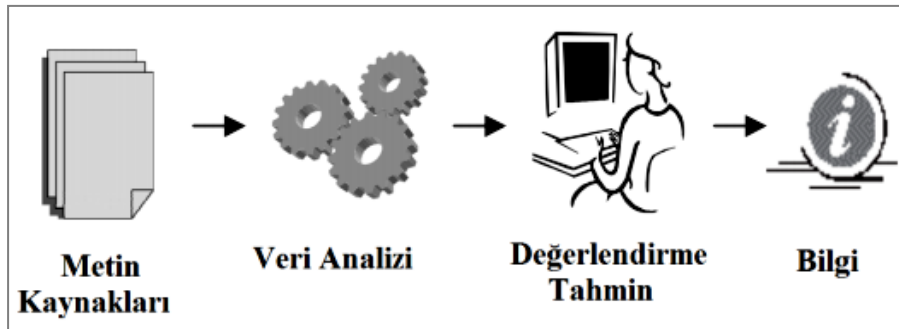
Beşinci bölümde, çalışmalar sonucu elde edilen sonuçlar değerlendirilerek yapılan uygulamaların katkılarından bahsedilmiştir.

2. METİN MADENCİLİĞİ

Metin madenciliği (MM) metinsel verilerden bilgi keşfi, genellikle işlenmemiş metin belgelerinden bilgi ve anlam çıkarma sürecidir. Bilgiyi depolamanın en doğal biçimi metin olduğu için, metin incelemesinin veri madenciliği açısından daha yüksek bir potansiyele sahip olduğuna inanılmaktadır. MM, doğal olarak işlenmemiş ve bulanık olan metin verileriyle uğraşmayı içerdiğinden veri madenciliğinden çok daha karmaşık bir görevdir. Metinsel veri analizi; bilgi çıkarımı, kümeleme, kategorizasyon, veritabanı teknolojisi, makine öğrenimi ve veri madenciliği gibi disiplinler arası bir alandır [38].

Metin madenciliği süreci ilk olarak 1960’larda başlamıştır. Doğal dil işleme alanındaki çalışmaların başlamasıyla metin analizi daha güvenilir yöntemlerle yapılmıştır ve bu süreçte geliştirilen metotlar günümüzde mevcut metin madenciliği araçlarında hâlâ kullanılmaktadır [39].

Metin analizinde önemli bir yere sahip olan metin madenciliğinin genel süreci Şekil 2.1’de verilmektedir. Metin madenciliğinde, metinsel verilerin elde edilmesinden sonra bir veri analizi süreci başlar. Bu analiz sürecinde, ayrıştırma, sözdizimsel analiz, semantik analiz ve kümeleme gibi doğal dil işleme işlemlerinden oluşan birçok alt süreç uygulanır. Veri analizi sürecini takriben elde edilen sonuçlar değerlendirilir ve önceden bilinmeyen bilgi keşfedilir [40]. Elde edilen bilgi, istatistik, karşılaştırma, değerlendirme vb. birçok amaç doğrultusunda etkin bir şekilde kullanılır.



Şekil 2.1. Metin madenciliği süreci

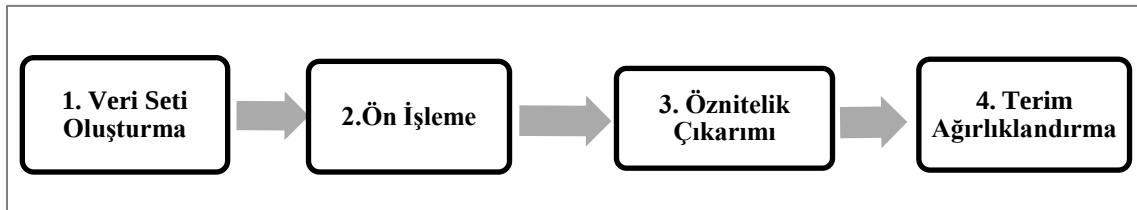
2.1. Metin ve Veri Madenciliği

Veri madenciliği, analiz ve bilgi çıkarımı amacıyla veriden yararlı bilgiler elde etmek için kullanılan algoritmalara dayalı bir işlemdir. Verilerdeki gizli kalıpları ve ilişkileri otomatik olarak keşfetmek ve büyük veri kümeleri üzerinde metinsel analiz yapmak için kullanılmaktadır. Metin madenciliği ise işlenmemiş metin belgelerini veya kaynakları, değerli işlenmiş bilgilere dönüştürmek için gereken süreçler dizisidir ve veri madenciliği disiplini içerisinde yer almaktadır [38]. DDİ teknikleri kullanılarak metinsel dokümanlar üzerinden yapılan işlemlerle veri yapısal forma dönüşür ve bir çıkarım yapmaya hazır hale getirilir.

Metin madenciliği aşamalarından geçen metinsel veri yapısal şekle dönüştükten sonra veri madenciliğinde metin incelemesi için kullanılır [41]. Bu da iki alan arasında önemli bir ilişki olduğunun göstergesidir. Ayrıca, metin madenciliği verileri düzgün veritabanlarından çok, doğal dil metinlerinden çıkarılmaktadır. Bu durum iki alan arasındaki en belirgin farklardan biridir. Metin madenciliği, metinlerin işlenebilir hale getirilmeleri için kullanılır. Dolayısıyla, metnin hazırlanması metin madenciliği ile metnin işlenmesi ise veri madenciliği ile gerçekleştirilir.

2.2. Metin Madenciliği Adımları

Metin madenciliği işlem adımları Şekil 2.2’de gösterildiği gibi verinin toplanması, ön işlemden geçirilerek gürültüden arındırılması, metni temsil edecek kelimelerin seçilmesi ve vektör oluşturulması şeklinde ifade edilir. Metin madenciliği çalışmalarında toplanan metinsel verinin yapısal veriye dönüştürülmesi ön işleme aşaması ile sağlanır. Bu aşamada metin üzerinde noktalama işareti, rakam, url vb. gibi temizleme işlemleri ve küçük harfe dönüştürme işlemleri yapılır. Temizlenen metin dizgilere ayrılır ve kelime kökleri üzerinde çalışılır. Kökleri elde edilen kelimeler terim adını alır.



Şekil 2.2. Metin madenciliği adımları

Sınıflandırma işleminin sağlıklı bir şekilde yapılabilmesi için öncelikle metni temsil eden kelimeleri başarılı bir şekilde tespit eden öznitelik çıkarımı işleminin yapılması gerekmektedir. Böylelikle cümlelerin içermiş olduğu duygu doğru bir şekilde analiz edilebilir duruma gelir. Öznitelikler elde edildikten sonra terim ağırlıklandırma (TA) işlemi uygulanır. TA ile her bir özneliğin ilgili metin ve veri setinde sahip olduğu önemi gösteren bir ağırlık değeri elde edilir. Terim ağırlıklandırma yapılırken terim frekansı (Tf), ters doküman frekansı (Idf), terim frekansı-ters doküman frekansı ($Tf-Idf$), binary, normalizasyon, vb. faktörlerden yararlanılır [12]. Böylece, metin sayısal formata dönüştürülerek vektör uzay modeli şeklinde ifade edilmiş olur.

Bu bileşenlerden Tf , bir t teriminin bir d dokümanındaki ağırlığını, Idf ise veri setindeki ağırlığını temsil eder. Tf ile Idf bileşenlerinin birleşimiyle ortaya çıkan $Tf-Idf$ ve benzeri yöntemler de kullanılmaktadır. Bu yöntemlerle ilgili detaylı bilgiler uygulama aşamasında verilmiştir.

2.3. Metin Madenciliği Uygulama Alanları

MM, metinden nitelikli bilgi çıkarımının yapıldığı her alanda kullanılmaktadır. Bununla birlikte, dijitalleşmenin ve sosyal ağların artışı ile bunlara olan bağımlılığın artması çoğu sektörü müşterilerin sadakatini artırma yoluna gitmeye sevk etmektedir. Sonuç olarak, duygu analizi metin madenciliğinin yeni odak noktasıdır. Artık bilginin metinden elde edilen stratejik bir varlık olduğu ve metin madenciliğinin artık lüks değil, bir zorunluluk olduğu fark edilmiştir. Ayrıca metnin analiz edildiği her ortamda metin madenciliği, uygulanması kaçınılmaz bir alandır. Bu doğrultuda metin madenciliği kullanım alanlarından birkaçı şu şekildedir:

- Müşteri ilişkileri yönetimi,
- Pazar araştırması,
- Sahtekârlık tespiti,
- Bilimsel ve medikal araştırmalar,
- Metinlerden bilgi çıkarımı,
- Doküman özetleme,
- Doküman sınıflandırma,
- Web içerikleri sınıflandırma,
- Yazar tanıma sistemleri.

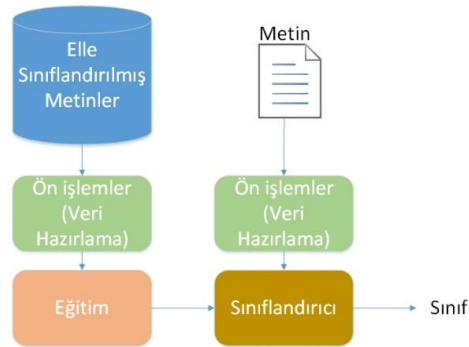
3. MAKİNE ÖĞRENMESİ TEKNİKLERİ İLE METİN SINIFLANDIRMA

Çevrimiçi bilgilerin hızlı bir şekilde büyümesiyle erişilebilir veri miktarı astronomik boyutlara ulaşmış ve bu sebeple metin sınıflandırma çalışmaları da büyük önem ve hız kazanmıştır. Metin sınıflandırma, metin verilerini işleme ve düzenleme için en önemli tekniklerden biri haline gelmiştir. Metinleri elle sınıflandırmak hem zor hem de zaman alıcı bir iş olacağından makine öğrenmesi teknikleriyle eğitim verileri üzerindeki örneklerden yararlanarak sınıflandırma yapmak daha avantajlı ve güvenilirdir.

Metin sınıflandırma, metnin içeriğine dayalı olarak önceden tanımlanmış bir sınıf kümesine atama yapmaktır. Başka bir deyişle, doğal dil metinlerini önceden tanımlanmış bir kümeyle etiketleme faaliyetidir. Metin sınıflandırma, metin madenciliğinin en popüler kısmıdır ve bu durum otomatik olarak metin belgelerinin birbiriyle ilişkili kategorilere sınıflandırılabilmesi anlamına gelmektedir [42].

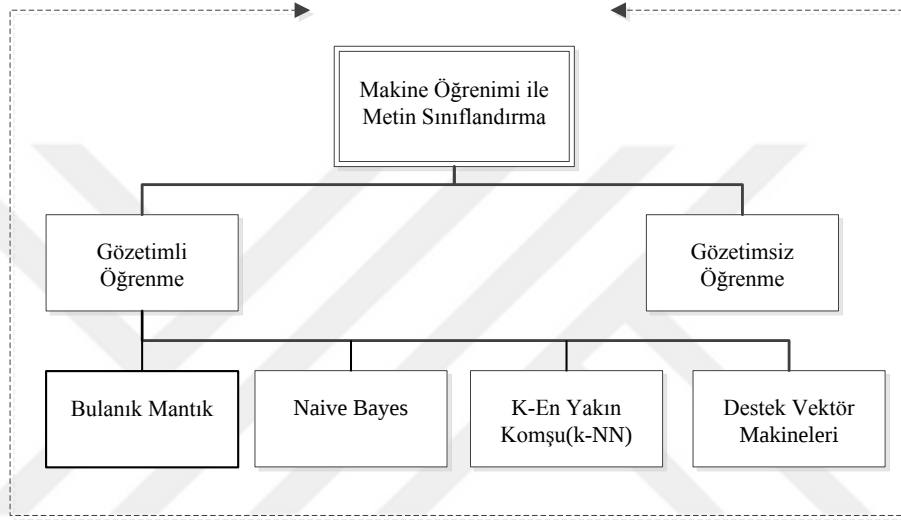
3.1. Makine Öğrenmesi ve Metin Sınıflandırma

Makine öğrenmesi, bir kısım eğitim verisi ile sistemi eğiten ve bu işlemler sonucunda bilgi çıkarımı yapan bir tekniktir. Makine öğrenimi süreci, veri madenciliği süreci ile benzerlik göstermektedir. Her iki çalışma alanında da model çıkarımı için veri üzerinde tarama yapılır. Farklı olarak ise, veri madenciliği insanların bilgi yorumu yapması için sonuç elde eder. Ancak makine öğrenmesi sahip olduğu bilgi ile sistemi eğitir [3]. Şekil 3.1’de makine öğrenmesi ile metin sınıflandırma işleminin çalışma şekli gösterilmiştir [44].



Şekil 3.1. Makine öğrenmesi ile metin sınıflandırma

Makine öğrenmesinin birçok alanda işe yarayan başarılı sonuçlar vermesi, doğal dil işleme alanı için de kullanılmasını yaygınlaştırmıştır. Şekil 3.2’de görüldüğü gibi, bir sınıflandırma konusu, makine öğrenmesinde denetimli veya denetimsiz öğrenme/sınıflandırma olmak üzere iki grupta incelenmektedir [3]. Gözetimsiz (denetimsiz, unsupervised) öğrenme yönteminde, herhangi bir etiket bilgisine sahip olmadan sınıflandırma yapılırken, gözetimli (denetimli, supervised) öğrenme yönteminde etikete sahip veriler üzerinde eğitim ve sınıflandırma yapılmaktadır.



Şekil 3.2 MÖ teknikleri kullanılarak yapılan MS türleri

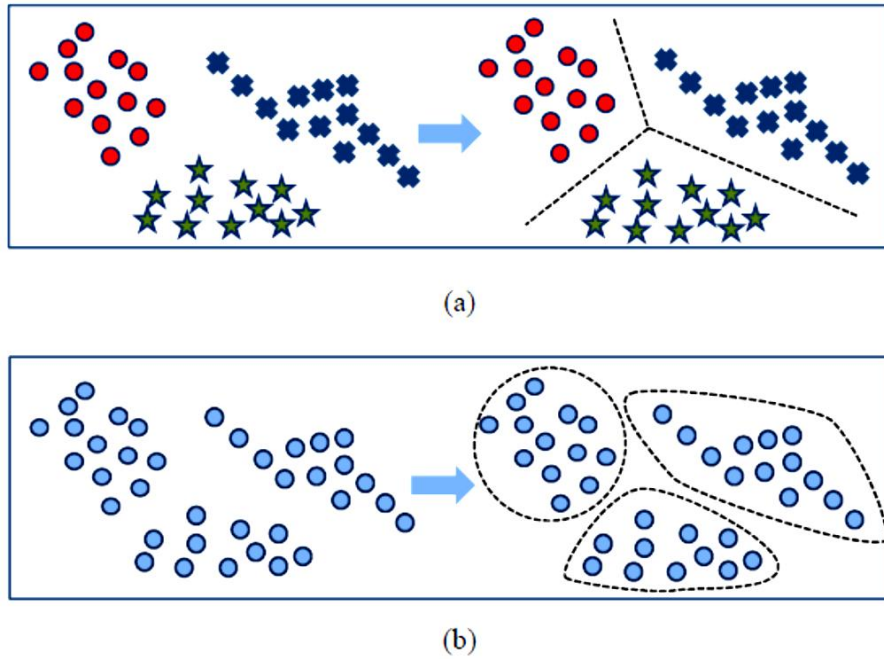
3.1.1. Denetimli Öğrenme

Denetimli öğrenme etiketli eğitim verisi ile yapılan sınıflandırmadır. Denetimli sınıflandırma algoritması, bu eğitim kümesindeki etiket bilgisinden yola çıkarak ilgili sınıf bilgisini öğrenir ve bir model oluşturur. Oluşan bu model sınıf ve etiket verisine sahip olmayan test örneklerini, eğitim kümesinden öğrendiği örüntülere göre sınıflandırır. Reklam e-postaların filtrelenmesi örneğinden yola çıkarak açıklamak gerekirse, e-posta olarak gelen mesajların reklam olup olmadığına karar veren sınıflandırma işlemi bir denetimli sınıflandırma örneğidir. Reklam e-posta ve reklam olmayan e-posta sınıflandırılacak iki etiketi temsil eder. İlgili makine öğrenmesi algoritması, veri setinde bulunan reklam ve reklam olmayan e-postalardan yola çıkarak bu iki sınıfın özelliklerine göre gerekli sınıfları öğrenir ve gelen bir e-postanın reklam olup olmadığına karar verir [12].

Şekil 3.2’de verildiği üzere bulanık mantık, naive bayes, k-en yakın komşu algoritması ve destek vektör makineleri denetimli öğrenme yaparken yaygın olarak kullanılan algoritmalardan bazılarıdır.

3.1.2. Denetimsiz Öğrenme

Denetimsiz öğrenme yönteminde, sistem etiketsiz veriler kullanılarak eğitilir. Denetimsiz öğrenmenin amacı sınıflandırma değildir. Çünkü elde bulunan veriler herhangi bir sınıf ve etiket bilgisine sahip değildir. Bu model genellikle kümeleme, yoğunluk tahmini, özniteliklerin birbirleriyle olan ilişkilerinin belirlenmesi ve boyut indirgeme gibi amaçlar için kullanılmaktadır. Ayrıca denetimsiz öğrenme algoritması sonucu elde edilen sonuçlar denetimli öğrenme için de kullanılabilir. Denetimli ve denetimsiz öğrenme modellerinin kullanım amaçları arasındaki fark Şekil 3.3’de gösterilmiştir [10]:



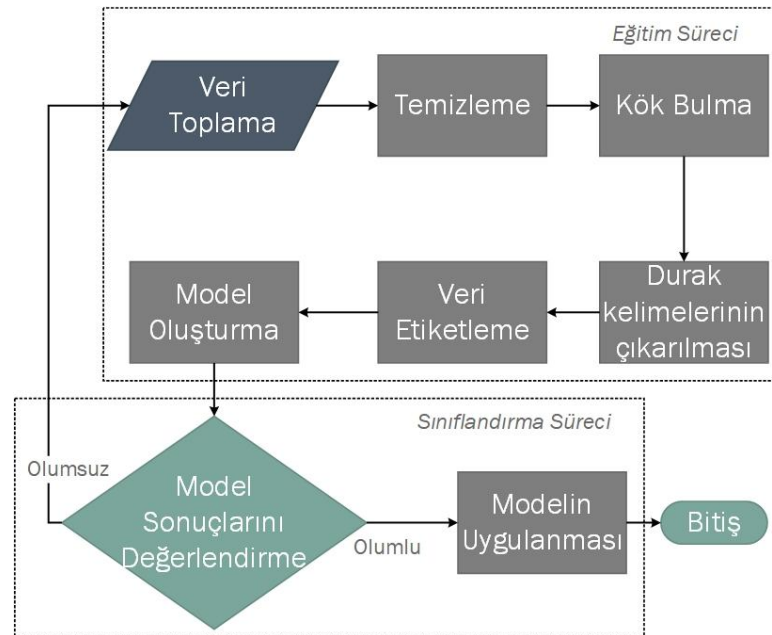
Şekil 3.3. Denetimli öğrenme modeli (a), denetimsiz öğrenme modeli (b)

Şekil 3.3’te görüldüğü gibi denetimli öğrenmede direkt veriler üzerinde bir sınıflandırma söz konusu iken; denetimsiz öğrenmede veri sınıfı belli olmadığından amaç sınıflandırma değil veriyi kümelemektir.

İngilizce, Türkçe ve diğer başka dillerde duygu analizi ile ilgili farklı alanlarda ve farklı amaçlarda çalışmalar mevcuttur. Çalışmalarda kullanılan yöntemler karşılaştırıldığında bazı yöntem ve algoritmaların sahip olduğu güçlü ve zayıf yönlerin bulunduğu görülmüştür. Gözetimli öğrenme tabanlı yöntemler yeni durumlara uyum sağlama gibi bir avantaja sahipken; verilerin etiketlenmesi ve modeli oluşturmak için gerekli süre ve maliyetin fazla olması yönünden dezavantaja sahiptir. Denetimsiz öğrenme tabanlı yöntemlerde ise, etiketsiz veriler kullanıldığından daha düşük performans verebileceği görülmüştür. Sözlük tabanlı yöntemlerin kullanımı basit ve anlaşılması kolay olmasına rağmen kullanılacak dil için bir sözlük tabanı oluşturmak sorun olabilmektedir.

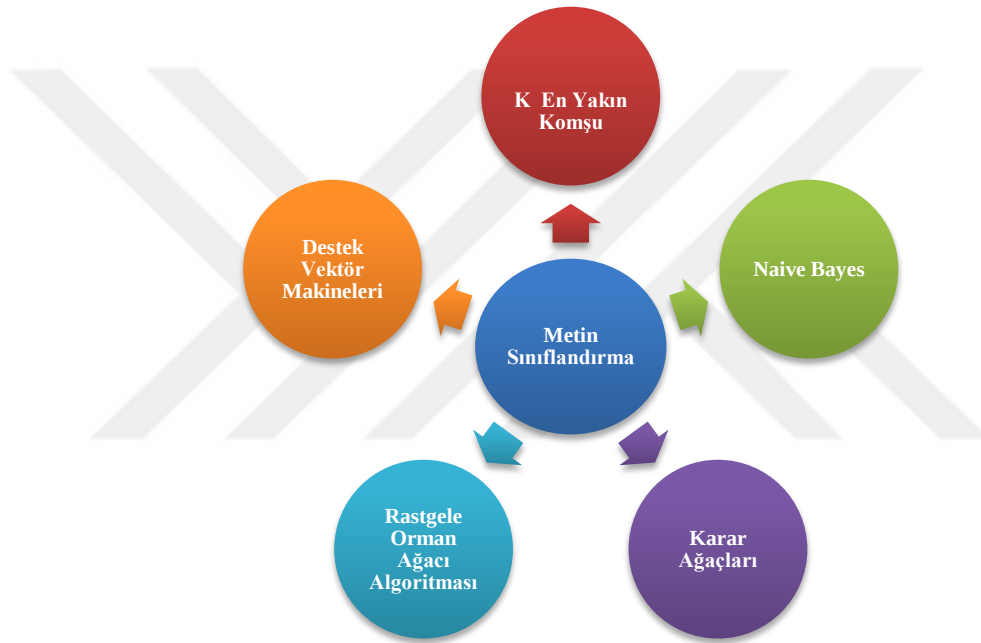
3.2. Metin Sınıflandırma Aşamaları

Metin sınıflandırma Şekil 3.4’de görüldüğü gibi iki aşamadan oluşmaktadır. Eğitim aşamasında hangi sınıf etiketine sahip olduğu bilinen bazı veriler eğitim kümesi olarak belirlenir. Belirlenen bu modelle sistem eğitilir. Sınıflandırma aşamasında ise ilgili veri setinden, ilk aşamada belirlenen eğitim kümesinde bulunmayan başka veriler alınarak test kümesi oluşturulur [20]. Böylece eğitim kümesi ile eğitilen ve öğrenme yetisi kazanan sistem test kümesi ile sınıflandırılır.



Şekil 3.4. Metin sınıflandırma aşamaları

Sınıflandırma problemlerinde daha çok makine öğrenmesi teknikleri kullanılır. Makine öğrenmesinin bu konuda yaygın olarak kullanılmasının sebebi ise daha önce bahsedildiği gibi verileri manuel olarak sınıflandırmanın pahalı ve zaman alan bir iş olmasıdır. Ayrıca, el ile sınıflandırma güvenilir nitelikte sonuçlar vermemektedir. Çünkü sınıflandırmayı yapan kişilere göre değişen sonuçlar güvenilir olmaktan çok uzaktır. Bu sebeplerden ötürü sınıflandırma işlemlerini yapan algoritmalar ve büyük miktarda veriler üzerinde çalışan sistemler önemli bir konuma yükselmiştir [43]. Metin sınıflandırma alanında yaygın olarak kullanılan bazı önemli algoritmalar Şekil 3.5'te verilmiştir [44].



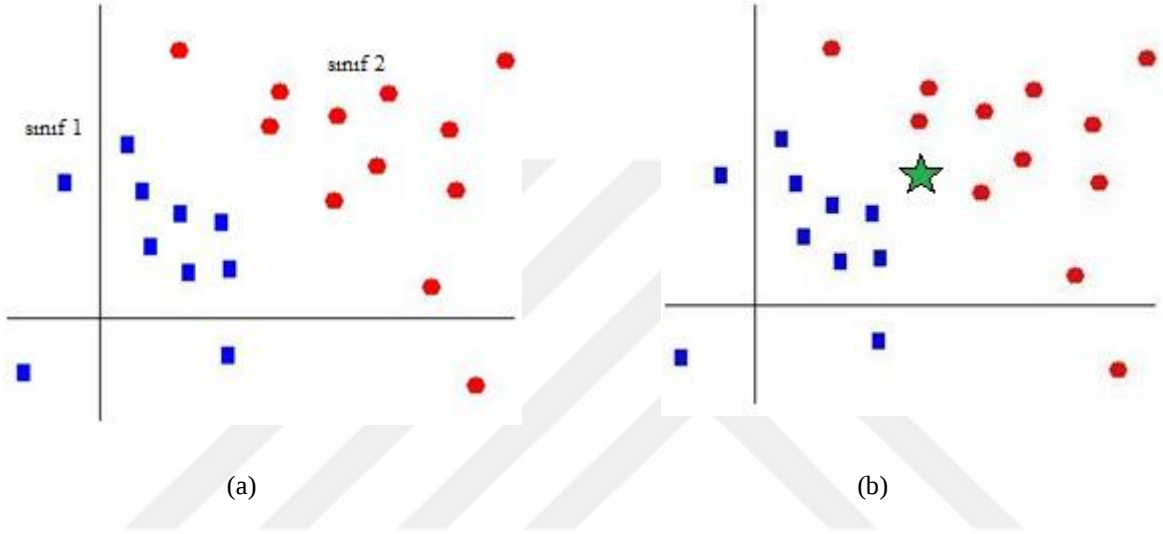
Şekil 3.5. MS alanında yaygın olarak kullanılan bazı algoritmalar

3.2.1. K - En Yakın Komşu Algoritması (k-NN)

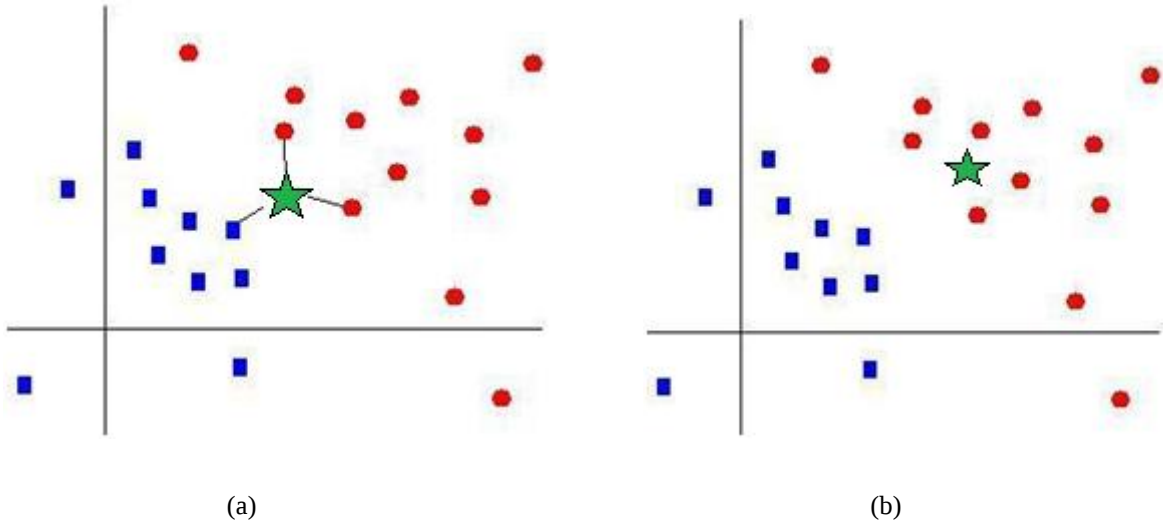
K-NN algoritması, denetimli ve örnekleme alanlı çalışan bir sınıflandırıcıdır. Eğitim setinde test edilecek bir örnek, eğitim setinde bulunan bütün örneklerle karşılaştırılır. Bir test örneğinin ait olduğu sınıfı belirlemek için eğitim setinden en yakın k tanesi seçilir. Sınıflandırılmak istenen örnek hangisine en yakınsa sınıf ataması yapılır. K-NN'nin amacı yeni bir vektörü sınıflandırmaktır [45].

Örneğin Şekil 3.6'da bulunan 2 boyutlu örnekleri ele alalım. k-NN yöntemine dayanarak gelen yeni bir örneğin hangi sınıfa ait olduğunu saptanması için en yakın 3 tane komşuyu alalım. k=3 alınarak komşulara olan uzaklık genelde eşitlik 3.1'de verilen *Öklid Bağıntısı* ile hesaplanır.

$$d(i, j) = \sum_{k=1}^p \sqrt{(x_{ik} - x_{jk})^2} \quad (3.1)$$



Şekil 3.6. İki boyutlu örnekler (a), Gelen yeni veri (b) [46]



Şekil 3.7. Yeni veriye en yakın 3 komşu üye (a), Yeni üyenin sınıflandırılması (b) [46]

Şekil 3.7 ‘te belirtildiği gibi yeni gelen veriye en yakın olan 3 üye öklid mesafesi ile tespit edilir. En yakın 3 üyenin iki tanesi sınıf 2 üyesine ait olduğuna göre yeni gelen veri, sınıf 2 üyesi olarak sınıflandırılır [46].

3.2.2. Naive Bayes (NB)

NB, makine öğrenmesinde gözetimli sınıflandırma metotlarından biridir. Bir çeşit olasılıkçı sınıflandırıcıdır ve Bayes teoremini uygulamaya dayalıdır.

Naive Bayes, makine öğrenmesinde kategorizasyon yaparak etkili sonuçlar almaktadır [47]. Verileri sınıflandırmak için yaygın olarak kullanılan bu algoritma kolay bir kullanıma sahiptir. Naive Bayes ile her özelliğin sonuca olan etkilerinin olasılık değerleri hesaplanır. Bayes teoremi eşitlik 3.2 ile ifade edilir:

$$P(x|y) = \frac{(P(y|x) * P(x))}{P(y)} \quad (3.2)$$

$P(x)$, x olayının olasılık değeri, $P(y)$ ise, y olayının olasılık değeri, $P(y|x)$, x olayının olduğu bilindiğinde y olayının olasılığı; $P(x|y)$ ise, x ’in şartlı olasılığıdır. Bu değerin en büyük olduğu noktalarda sınıf tahmini yapılabilir [48].

Sınıflandırma üzerinden naive bayes kullanımını örneklendirecek olursak bir y örneğinin D_i sınıfına ait olma olasılığı eşitlik 3.3 ile hesaplanır [63].

$$P(D_i|Y) = \frac{(P(Y|D_i) * P(D_i))}{P(Y)} \quad (3.3)$$

Herhangi bir Y dokümanının hangi sınıfta olduğunu bulmak için her D_i sınıfına eşitlik 3.4 uygulanır. Buradan hareketle Y dokümanı olasılık değeri en yüksek olan sınıfa atanır [63].

$$P(Y|D_i) = \prod_{k=1}^m P(Y_k|D_i) \quad (3.4)$$

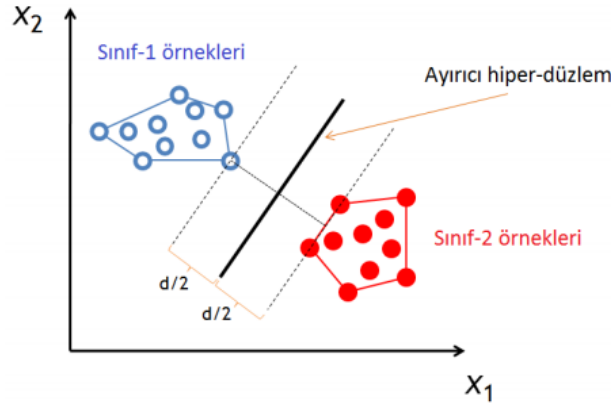
3.2.3. Destek Vektör Makinesi (DVM)

DVM, genellikle metin sınıflandırma alanında kullanılan denetimli öğrenme algoritmalarından biridir. Basit ve etkili bir sınıflandırma algoritmasıdır.

Etiketli eğitim verileri, DVM'yi kullanarak sınıflandırma yapmak için bir vektör modeli oluşturur. Bu gruplar *hiper düzlem* adı verilen bir sınır çizgisi ile ayrılır.

Vapnik tarafından geliştirilen bu algoritma temelde, sınıflar arasındaki en büyük ayırımı oluşturan hiper düzlemi bulmayı amaçlar [49]. Dolayısıyla destek vektör makinesi sistemi, sınıfların birbirine en yakın örneklerini alarak hiper düzlem çizgisini her iki sınıfa da eşit mesafede alır.

Algoritma, doğrusal olarak ayrılan iki sınıflı problemlerin çözümü için düşünülmüş, ancak daha sonra doğrusal olmayan ve ikiden fazla sınıf etiketine sahip olan veriler için de kullanılmaya başlanmıştır. Şekil 3.8'de doğrusal olarak ayrıştırılabilen iki sınıfı en iyi şekilde ayıran hiper düzlem gösterilmiştir.



Şekil 3.8. DVM ile ikili sınıflandırma

Bu düzlemde bulunan her bir noktanın tanımı eşitlik 3.5 ile yapılır. Her x, c değişken ikilisi için x , vektör uzayında bulunan ve $i = 1$ 'den n 'ye kadar giden herhangi bir noktadır. C değeri ise bu noktanın -1 veya $+1$ olduğunu gösteren değerdir [46].

En basit destek vektör makinesi uygulaması doğrusal olanlar için uygulanır. Lineer sistemlerde -1 ve $+1$ olmak üzere, iki sınıf etiketi mevcuttur. Lineer sistemlerde sınıflar, w hiper düzlem çizgisine dik olan normal vektörü, x bahsedilen noktanın değişken parametresi ve b ise kayma oranı olmak üzere, 3.6'da verilen eşitlik ile ayrılır.

$$D=\{(xi, ci)|xi \in R^P, ci \in \{-1,1\}\}_{i=1}^n \quad (3.5)$$

$$f(x) = wx + b \quad (3.6)$$

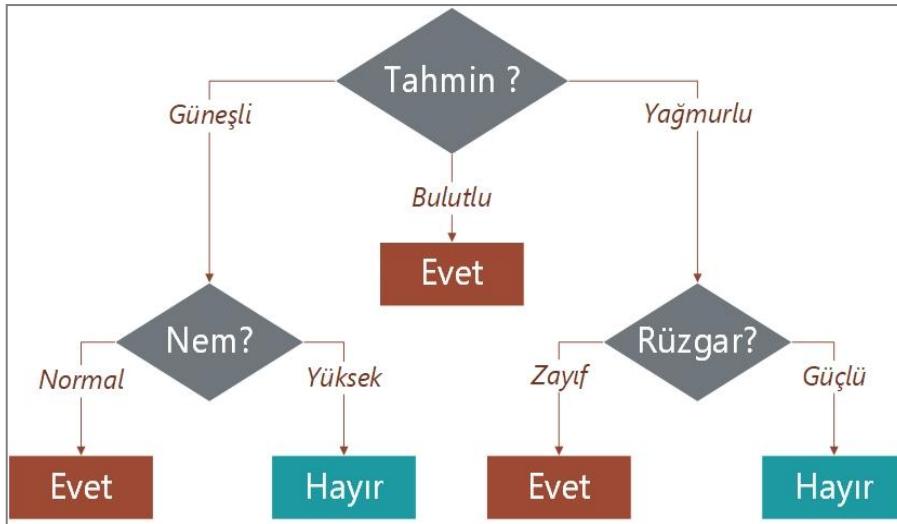
İki sınıf arasında bulunan mesafenin en büyük olması için eşitlik 3.7’de verilen ifadenin minimum olması gerekmektedir. Ayrıca, verilerin doğru bir şekilde sınıflandırılabilmesi için her bir örneğin 3.8’de verilen eşitsizliğe uygun olması gerekmektedir [64].

$$\frac{1}{2} \|w\|^2 \quad (3.7)$$

$$y_i (w^T x_i + b) \geq 1 \quad (3.8)$$

3.2.4. Karar Ağacı Algoritması

Karar ağacı, tahmin etme temeline dayanan bir makine öğrenmesi algoritmasıdır. Örnek veriler, mevcut verilerin niteliklerinden oluşturulmuştur. Bu modelde öğrenilen bilgi bir ağaç üzerinde modelize edilir. Şekil 3.9’da üç düğümlü bir karar ağacı gösterilmiştir. Bu karar ağacında hava durumuna göre oyun oynayıp oynamama kararı verilmiştir. Karar ağacı sistemlerinde yapraklar sınıf etiketlerini tutar (evet/hayır) ve dallar sınıflara götüren özelliklerin birleşim durumlarını temsil eder.



Şekil 3.9. Karar ağacı örneği

Karar ağaçlarında en yaygın olarak kullanılan algoritma ID3 algoritmasıdır. ID3 algoritması entropi ve bilgi çıkarımı tabanlı bir algoritmadır.

Entropi kavramı, rastgele ve belirsiz olan bir durumun meydana gelme olasılığını ifade eder. Eğer örneklerin hepsi düzenli olursa entropi değeri sıfır olur. Örneğin, Şekil 3.9’da verilen karar ağacında oyun oynama kararı için tüm hepsi “Evet” veya “Hayır” olursa entropi sıfır olur [65].

3.2.5. Rastgele Orman Ağacı Algoritması (RO)

Rastgele orman ağacı algoritması birden fazla karar ağacının meydana gelmesiyle oluşur. Kullanılan ağaçlar arasında oylama yaparak sınıflandırma yapar. Bu algoritma öncelikli olarak birden fazla ağaç oluşturur. Her bir ağaca atanan giriş vektörü ile bu ağaçların birer sonuç üretmeleri sağlanır. Bütün ağaçlar arasında en fazla oyu alan ağaç sınıfı, karar verilen sınıf olarak seçilir. Buradaki her bir ağaç, veri setinin eğitim kümesinde bulunan örneklerin rastgele, bir yenisiyle değiştirilmesi sonucu oluşturulur. Sınıflandırma için eğitim seti hesaplanır [63].

Rastgele orman ağacı algoritması başarılı sonuçlar verdiği için günümüzde birçok uygulamada etkin olarak kullanılmaktadır. Literatürde kullanılan algoritmalar arasında eşsiz bir sınıflandırıcı olmuştur. RO algoritmasında, istenildiği kadar ağaç ile çalışılabilir. Algoritma Breiman tarafından önerilmiştir ve temeli karar ağaçlarına dayanır. Eğitim örnekleri ile eğitilen karar ağaçları her düğümde bütün özellikleri incelemek yerine rastsal olarak aldığı alt kümeyi inceler. Böylece, karar ağacının üretimi için geçen süre azalır. RO algoritması aşağıdaki adımlar ile gerçekleştirilir.

1. N kadar orijinal veriden yer değişimli bir şekilde, rastgele olarak N tane eğitim kümesi elde edilir.
2. Düğümlerin her biri için M toplam girdi değerlerinden rastgele bir şekilde $m \leq M$ olacak biçimde değerler oluşturulur. Bu değer orman süresince sabit değerde kalır.
3. Ağaçların her biri olabildiği kadar geniş bir şekilde oluşturulur. Sınıflandırma esnasında ağaç işlemlerinde budama yapılmaması rastgele orman ağacı algoritmasını öteki sınıflandırıcılardan ayıran en önemli özelliklerden biridir [66].

3.3. Metin Sınıflandırmanın Kullanım Alanları

Metin sınıflandırma alanında yapılan birçok farklı uygulama vardır. Örneğin istenmeyen reklam e-postaların filtrelenmesi, bir metin yazarının belirlenmesi anket çalışmaları, cinsiyet belirlenmesi, haber metinlerinin kategorize edilmesi, vb. birçok uygulama MS uygulamasına örnektir. Metin sınıflandırma, metinlerin kullanıldığı her işlemde sınıflandırma problemlerini etkin bir şekilde çözebilmektedir. Doğal dil işlemlerinin bir parçası olarak MS'nin kullanıldığı bazı alanları şunlardır.

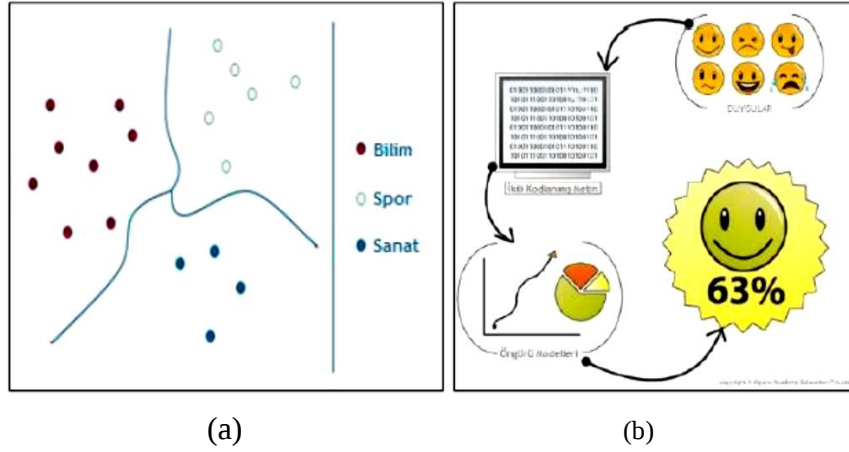
- Metinde istenmeyen terimlerin filtrelenmesi,
- Metin kategorizasyonu,
- Sözcük analizi işlemleri [20],
- İntihal tespiti.

Literatürde metin sınıflandırma alanında, ilgili köşe yazılarından yazar adı ve cinsiyet tahmini, kitap türü ve internet sitelerinin sınıflandırılması, soru algılama sistemleri, belge sınıflandırma ve kategorize etme, haber metinlerinin sınıflandırılması, vb. birçok alanda uygulamalar yapılmıştır.

3.4. Duygu Analizi ve Metin Sınıflandırma İlişkisi

Duygu analizi çalışmalarında veri setinde bulunan bütün örnekler, pozitif, negatif ve nötr (tarafsız) olmak üzere 3 veya daha fazla kategoride sınıflandırılır. Duygu analizi, her bir mesajı içinde barındırdığı baskın olan duygu ile etiketlendirir. Bu yönüyle DA, temelde metin sınıflandırma işlemlerine dayanan bir sınıflandırma problemi olarak ele alınır [38].

Metin sınıflandırma ve duygu analizi çalışmalarının ortak yönü ise kullanılan verinin tamamen metin tabanlı olmasıdır. Metin sınıflandırmada metinler için önceden belirlenen etiket ifadesi, duygu analizinde duygu ile ifade edilir. Literatürde duygu analizi çalışmalarında, duygu analizi teknikleri ile beraber metin sınıflandırma yöntemleri kullanılmaktadır. Şekil 3.10'da gösterildiği gibi metin sınıflandırma ve duygu analizi çalışma alanlarının metin tabanlı veri ve makine öğrenmesi yöntemlerinin kullanılması gibi ortak yönleri bulunmaktadır.



Şekil 3.10. Metin sınıflandırma (a), Duygu analizi (b)

Metin sınıflandırma ve duygu analizi arasında benzerliklere rağmen yapılan uygulamaya göre, kullanılan metin, belge veya mesajın kendine has özellikleri olabileceği için kullanılan teknikler açısından özellikle ön işleme aşamasında farklılıklar gözlenebilmektedir. Çünkü duygu analizi çalışmalarında kullanılan metinler, dil bilgisi kurallarına uyulmadan kısa bir şekilde ifade edilir. Ayrıca çoğu gönderiler sosyal medya jargonuyla yazıldığından bu ifadeleri yapısal formata dönüştürüp hazırlamak, farklı ön işleme işlemleri gerektirebilir. Dolayısıyla, tez çalışmasında kullanılan Twitter mesajlarından anlamlı bilgi çıkarımı yapmak, bu mesajların 140 karakter ile sınırlı olması, kullanıcı adı ve “hashtag” gibi özel terimler ve “emoji” adı verilen duygu ifade eden simgeler içermesi sebebiyle, klasik ön işleme aşamalarından daha kompleks ve zordur.

4. UYGULAMA ÇALIŞMALARI

Teknoloji ve internetin büyük oranda gelişmesiyle bilgiye ulaşmak kolaylaşmıştır. İstatistiksel verilere göre internet son 5 yılda 50 milyonun üzerinde kullanıcıya erişmiş olup, dünyada her kesim tarafından kullanılan bir araç olmuştur. İnsanların aradıkları bilginin sadece birkaç tık uzakta olduğunu bilmesi ve internetteki verinin devasa boyutlara ulaşması beraberinde farklı problemler getirmiştir. Problem, istenilen yapısal bilgiyi elde etmedeki zorluktur. Bunun nedeni ise internetteki bilginin artmasıyla ortaya çıkan bilgi kirliliği ve özensiz bilgi paylaşımıdır [20]. Var olan bu bilgi kirliliğinin içinde istenileni elde etmenin yolu, metin madenciliği ve metin sınıflandırma ile veriyi yapısal hale getirip kategorize etmekten geçmektedir. Bu işlemler sonucu elde edilen yapısal ve temiz veri duygu analizinde metinlerin barındırdığı duyguyu belirlemede kullanılmaktadır.

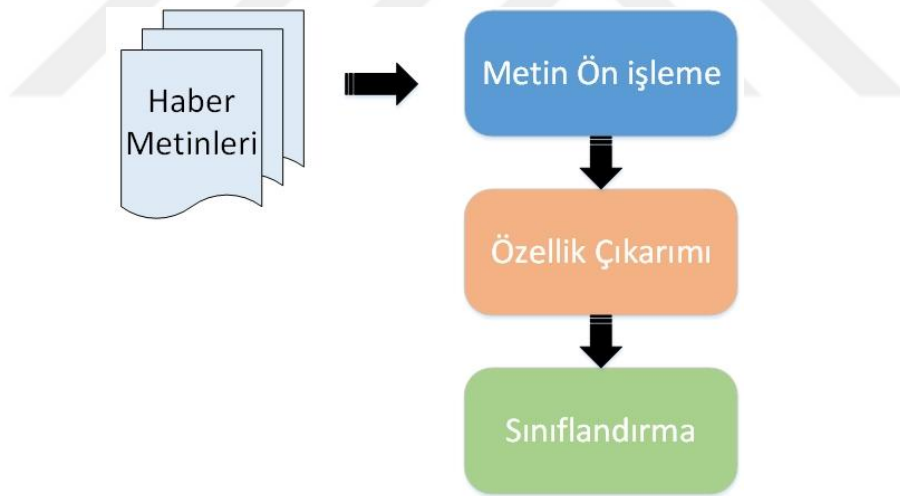
Bahsedilen problemlere etkili çözümler oluşturmak amacıyla bu tez çalışması kapsamında farklı çalışmalar yapılmıştır. İlk çalışmada, doğal dil işleme, metin madenciliği ve metin sınıflandırma alanlarına dayanan haber metinlerinin farklı yöntemlerle sınıflandırılması; ikinci çalışmada ise en büyük sosyal ağ ortamlarından olan Twitter gönderilerinden alınan tweetler üzerinde yapılan duygu analizi çalışması yapılmıştır.

Metin tabanlı veriler üzerinde çalışan her uygulamanın temelde aynı adımlardan geçtiği görülmüştür. Ancak veri setine göre temel adımlar farklı alt dallara ayrılabilir. Örneğin haber metinleri ile yapılan ilk uygulamada haber metinleri Türkçe dilinin kurallarına uygun bir şekilde ve herhangi bir ifade kısaltması yapılmadan yazılırken, Twitter gönderileri kullanıcıların kısalttıkları birçok kelime ve emoji adı verilen duygu ifadeleri ile yazılmaktadır. Dolayısıyla metin ön işleme aşamasında Twitter verileri daha detaylı işlemlerden geçerek yapısal forma ulaşmıştır. Sınıflandırma aşamasında çeşitli sınıflandırma algoritmaları kullanılarak yöntemlerin başarı oranları karşılaştırılmıştır. Yapılan bilgi çıkarımı ile başarılı sonuçlar elde edilmiştir.

Yapılan çalışmalar kapsamında bulunan tüm işlemler, işlem adımları, kullanılan algoritmalar ve başarı oranları Uygulama Çalışması 1 ve Uygulama Çalışması 2 bölümlerinde detaylıca anlatılmıştır.

4.1. Uygulama Çalışması 1

Bu çalışmada metin madenciliği ve metin sınıflandırma teknikleri kullanılarak, farklı haber siteleri ve gazetelerden alınan haber metinlerinin makine öğrenmesi teknikleriyle sınıflandırılması gerçekleştirilmiştir. Bu çalışmadaki amaç internet ortamında dağınık halde bulunan bazı haber metinlerini otomatik olarak kategorize etmektir. Bu amaçla, her birinden 20 tane olmak üzere dört farklı kategoriye ait (ekonomi, spor, politika, sağlık) toplam 80 haber metni toplanmıştır. Bu haber metinlerinden rastsal olarak belirlenmiş 60 metin eğitim; 20 metin ise test amaçlı kullanılmıştır. Eğitim verileri kullanılarak eğitilen sistem, sonrasında test verileriyle sınıflandırılmıştır. Farklı öznitelik ve terim ağırlıklandırma yöntemleri ve sınıflandırma algoritmaları kullanılarak, bu algoritmaların başarı oranları istatistiksel olarak karşılaştırılmıştır. Ayrıca, boyut indirgeme yapılarak öznitelik boyutu indirgenmiş ve deneysel çalışmalar boyutu azalan veri seti üzerinde tekrar edilerek karşılaştırılmıştır. Çalışma modeli Şekil 4.1’de gösterilmiştir.



Şekil 4.1. Uygulama Çalışması 1 modeli

Çalışma kapsamında izlenen adımlar Şekil 4.2’de verildiği gibi, öncelikle uygulama kapsamında veri seti oluşturulmuş ve metin verileri çeşitli ön işleme aşamalarından geçirilerek yapısal formata dönüştürülmüştür.

Devam eden işlemlerde, özellik çıkarımı ve terim ağırlıklandırma yöntemleriyle öznitelikler vektörel uzayda ifade edilmiş ve sayısal veriler elde edilmiştir. En son aşamada ise çeşitli sınıflandırma algoritmaları kullanılarak başarı oranları karşılaştırılmıştır.



Şekil 4.2. Uygulama Çalışması 1 işlem adımları

4.1.1. Veri Seti Oluřturma

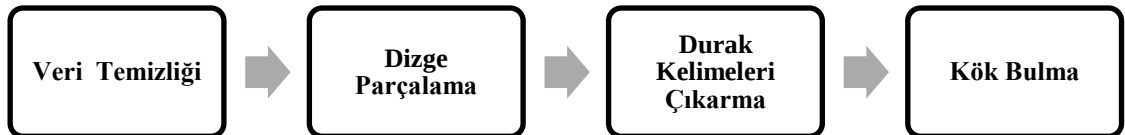
Uygulamada, farklı haber sitesi ve gazetelerden 4 farklı haber kategorisine ait 20’şer haber metni alınmıřtır. Bu metinlerin 60’ı eğitim; 20’si ise test verisi olarak kullanılmıřtır. Veri seti ile ilgili detaylar Tablo 4.1’de verilmiřtir.

Tablo 4.1. Uygulama Çalışması 1 veri seti detayları

Haber Kategorisi	Haber Metni Sayısı	
	Eğitim	Test
Ekonomi	15	5
Politika	15	5
Sağlık	15	5
Spor	15	5
Toplam	60	20
	80	

4.1.2. Metin Ön İşleme

Ön işleme aşamasının temel amacı metinsel verilerden, kategorilerinin birbirinden ayırt edilmesini sağlayabilecek önemli özniteliklerin ortaya çıkarılması ve makine öğrenmesi algoritmalarının işleyip yorum yapabileceği şekilde sayısallaştırılarak uygun formata dönüřtürülmesidir. Ön işleme aşaması, sınıflandırmanın başarısı ve etkili özniteliklerin elde edilebilmesi sebebiyle önemli bir aşamadır. Bu adım, metin verisinin çalışma amacına ve yapısına baėlı olarak deėişebilen alt adımlardan oluřmaktadır [12]. Şekil 4.3’te temel ön işleme adımları verilmiřtir.

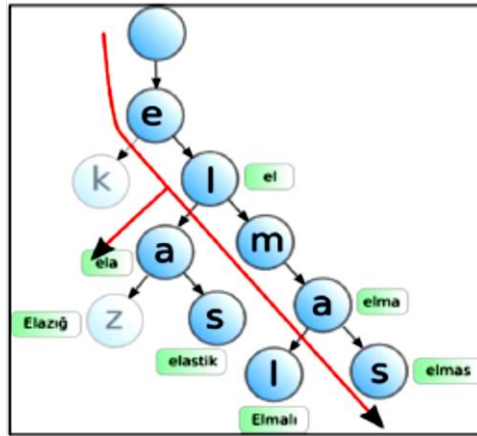


Şekil 4.3. Temel ön işleme adımları

4.1.2.d. Kök Bulma

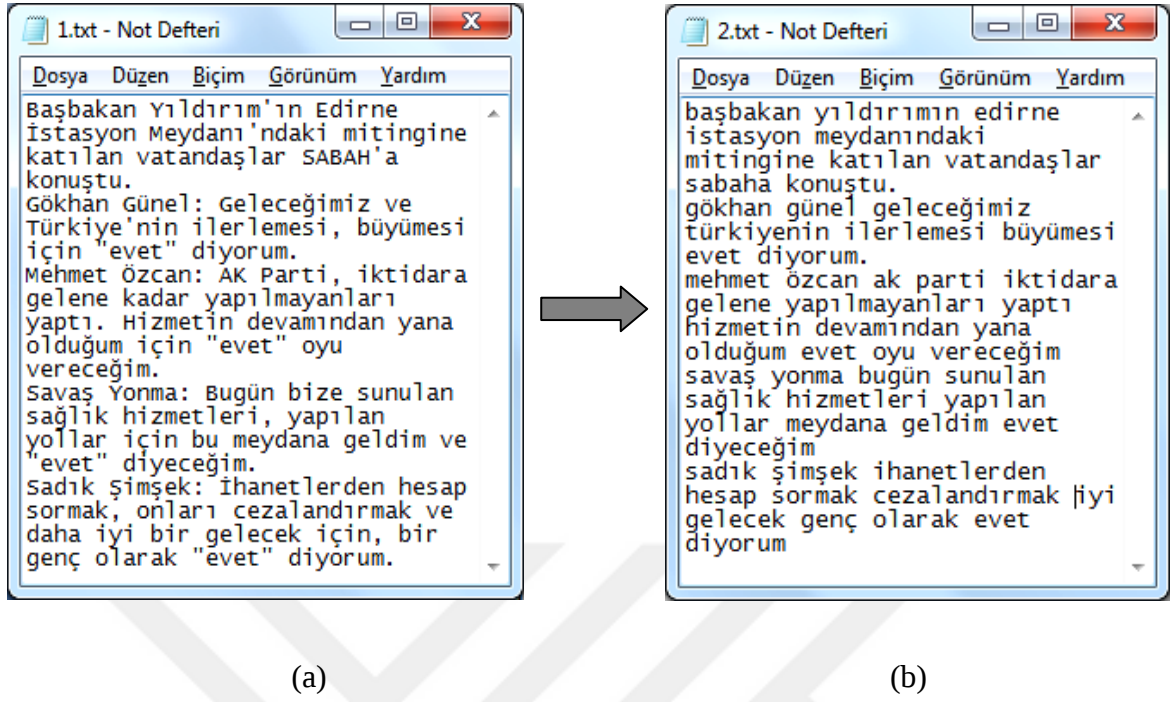
Yapılan çalışmalarda metinsel analiz ön planda olduğundan dolayı, her kelimenin kökü ve eklerini elde etmek çok önemlidir. Bu çalışmada kelimeler köklerine ayrılarak, sınıflandırma yapılabilecek terimler elde edilmiştir. Türkçe sondan eklemeli bir dil olarak iki tür ek ile oluşur. Yapım eki alan kelimeler yeni anlamlara sahip olurken; çekim eki alan kelimenin anlamında herhangi bir değişiklik oluşmaz. Bu sebeple, MS çalışmalarında kelime yerine kelime kökü kullanılır. Çünkü kelime kökleri yerine kelimelerle çalışıldığı takdirde farklı çekim eki almış kelimeler uygulama tarafından farklı kelimeler olarak değerlendirilir. Ancak, kök olarak aynı kelime temsil edildiğinden etkili bir öznelilik olmaz. Böylece hem sözcük hem de vektör boyutu artmış olur ve yanlış sonuçlar elde edilir [51].

Haber metinleri içerisinde yer alan kelimelerin dizgelere ayrıştırılmasında ve köklerinin bulunmasında Türkçe doğal dil işleme kütüphanesi olan Zemberek kütüphanesinden faydalanılmıştır [52,53]. Gövdeleme işleminin kelimeler üzerinde Zemberek kök aday bulucu ile uygulanışı Şekil 4.5’ te gösterilmiştir [16].



Şekil 4.5. Zemberek kök aday bulucu

Zemberek açık kaynak kodlu Türkçe DDİ kütüphanesidir [53] ve kelimeleri analiz ederek sözcüklerle kök-ek birleşmeleri sağlar. Uygulamada bulunan verinin temizliği ve etkisiz kelimelerin çıkarılması işlemleri Visual Studio ortamında C# programlama dili ile gerçekleştirilmiştir. İşlem sonrası metnin durumu Şekil 4.6’da verilmiştir:



Şekil 4.6. Örnek metin verisi (a) , Metnin etkisiz kelime ve noktalama işaretlerinden temizlenmesi (b)

4.1.3. Öznitelik Çıkarımı

Öznitelik çıkarımı metni en iyi şekilde temsil eden özellikler bütünüdür. Literatürde metin sınıflandırma çalışmalarında farklı öznitelik çıkarımları kullanılmıştır. Şekil 4.7’de kullanılan bazı öznitelikler gösterilmiştir.



Şekil 4.7. Bazı öznitelik çıkarım yöntemleri

Kelime çantası öznitelik çıkarım modelinde öznitelikler metni en iyi şekilde temsil eden anlamlı kelimelerden çıkarılır [12]. Bu çalışmada metin içindeki her bir kelimenin kökü ve kelime seviye N-gram olmak üzere iki tür öznitelik yöntemi kullanılmıştır. Zemberek kütüphanesi ile veri setinden çıkarılan kelime kökü ilk öznitelik olarak kullanılmıştır. İkinci öznitelik olarak ise, yine metin içerisinde çıkarılan farklı uzunluklardaki kelime kombinasyonlarından oluşan kelime seviye N-gram kullanılmıştır. Kelime seviye N-gram işlemi, örnek bir cümleyle açıklanacak olursa, “Türkçe sondan eklemeli bir dildir” cümlesi kelime seviye 1-gram (unigram), 2-gram (bigram) ve 3-gram (trigram)’lar ile Tablo 4.2’de olduğu gibi ifade edilir. Bu çalışmada bigram öznitelik çıkarım yöntemi kullanılmıştır.

Tablo 4.2. Kelime seviye N-gram kullanımları

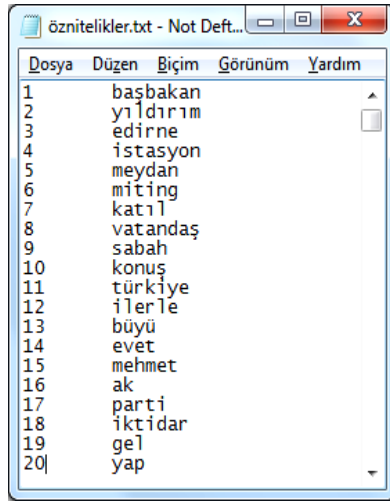
Unigram	“Türkçe”, “sondan”, “eklemeli”, “bir”, “dildir”
Bigram	“Türkçe_sondan”, “sondan_eklemeli”, “eklemeli_bir”, “bir_dildir”
Trigram	“Türkçe_sondan_eklemeli”, “sondan_eklemeli_bir”, “eklemeli_bir_dildir”

Tablo 4.3’te haber metinlerden çıkarılan öznitelik türlerinin ham sayıları ve rakamlardan, noktalama işaretlerinden ve etkisiz kelimelerden arındırılıp ön işleme aşamasından geçirildikten sonraki sayıları verilmiştir.

Tablo 4.3. Çalışmada kullanılan öznitelik türü ve sayıları

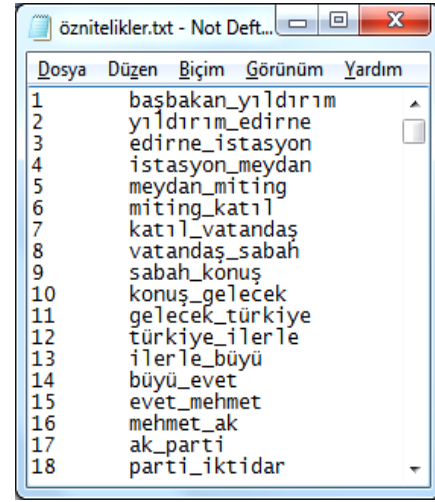
Öznitelik Türü	Ham Veri	Ön İşlemeden Sonra
Kelime Kökü	1785	1693
Kelime Seviye n-Gram (n=2, bigram)	6065	5395

Uygulamada öznitelik çıkarımı olarak kullanılan kelime kökleri ve kelime seviye bigram yöntemleri Netbeans platformunda java programlama dili ile gerçekleştirilmiştir. Şekil 4.8’de çıkarılan öznitelik vektörleri verilmiştir.



Dosya	Düzen	B biçim	Görünüm	Yardım
1	başbakan			
2	yıldırım			
3	edirne			
4	istasyon			
5	meydan			
6	miting			
7	katıl			
8	vatandaş			
9	sabah			
10	konus			
11	türkiye			
12	ilerle			
13	büyü			
14	evet			
15	mehmet			
16	ak			
17	parti			
18	iktidar			
19	gel			
20	yap			

(a)



Dosya	Düzen	B biçim	Görünüm	Yardım
1	başbakan_yıldırım			
2	yıldırım_edirne			
3	edirne_istasyon			
4	istasyon_meydan			
5	meydan_miting			
6	miting_katıl			
7	katıl_vatandaş			
8	vatandaş_sabah			
9	sabah_konus			
10	konus_gelecek			
11	gelecek_türkiye			
12	türkiye_ilerle			
13	ilerle_büyü			
14	büyü_evet			
15	evet_mehmet			
16	mehmet_ak			
17	ak_parti			
18	parti_iktidar			

(b)

Şekil 4.8. Kelime kökü öznitelik çıkarımı (a), Kelime seviye n-gram (b)

4.1.4. Terim Ağırlıklandırma

Terim Ağırlıklandırma, bir özniteliğin metinde görülme sıklığıdır. Bir terimin bir dokümandaki ağırlığının yüksek olması, bu terimin o metin için ayırt edici özelliğe sahip olması anlamına gelmektedir. Dolayısıyla, terimlerin ağırlıklandırılması sınıflandırma başarımı üzerinde etkili olan faktörlerden biridir [54]. Bu çalışmada, özniteliklerin ağırlıklandırılmasında Tf ve $Tf-Idf$ olmak üzere iki farklı yöntem kullanılmıştır.

Terim Frekansı (Tf), eşitlik 4.1’de ifade dildiği gibi veri içinde bulunan özniteliklerden kaç defa geçtiği bilgisini temel alan yöntemdir. *Terim Frekansı-Ters Doküman Frekansı ($Tf-Idf$)* ise, Tf ile tüm doküman içerisinde geçen kelimelerin, aranan kelimelere oranının belirlenmesidir. Ağırlıklandırma içerisinde kullanılan Idf değeri, ilgili verinin, bütün veri kümesinde kaç defa geçtiğini gösteren değeri ifade etmektedir [52].

Df , bir c teriminin veri havuzunda gözlemlendiği doküman sayısı olmak üzere eşitlik 4.2’de Tf ve Idf hesaplanması, eşitlik 4.3 ve eşitlik 4.4’te ise ağırlıklandırma hesaplanması gösterilmiştir [55].

$$W_{Tf}(c, d) = Tf(c, d) \quad (4.1)$$

$$W_{Df} = \sum_{i=1}^n Tf(d_i, c) \quad (4.2)$$

$$W_{Idf}(c, d) = \log \frac{N}{Df(c, d)} \quad (4.3)$$

$$W_{Tf-Idf}(c, d) = Tf(c, d) * Idf(c, d) \quad (4.4)$$

Bu çalışmada, Tf ve $Tf-Idf$ olmak üzere iki farklı ağırlıklandırma yöntemi uygulanmıştır. Tablo 4.4'te Tf ağırlıklandırma yönteminin oluşma şekli ile ilgili örnek verilmiştir.

Tablo 4.4. Terim frekansı oluşturma

Veri	"Akşam güzel bir konser var. Akşam konsere gidelim mi?"
Kelime kökü öznitelikler	{Akşam, Pazar, Konser, Güzel}
Terim Frekansı	{2, 0, 2, 1}

4.1.5. Boyut İndirgeme

Yapılan çalışmada öznitelik seçme işlemiyle, veri kümesini en iyi şekilde ifade eden öznitelikler seçilmiştir. Öznitelik boyutu indirgenerek zaman ve performans artışı amaçlanmıştır. Bu amaçla, öznitelikleri seçmek için $CfsSubset$ algoritması kullanılmıştır [56]. İndirgenen boyutlar Tablo 4.5'te gösterilmiştir.

Tablo 4.5. Boyutu indirgenen veri

Öznitelik Türü	Yapısal Veri	Boyut İndirgeme
Kelime Kökü	1693	32
Kelime Seviye bigram (n=2)	5395	16

4.1.6. Otomatik Haber Sınıflandırma

Farklı haber siteleri ve gazetelerden alınan 4 kategoriye ait haber yazıları veri olarak alınmış ve gerekli ön işleme adımları izlendikten sonra veri seti sınıflandırmaya hazır hale getirilmiştir. Sınıflandırma için Waikato Üniversitesi tarafından geliştirilen Weka arayüzü [57] kullanılmıştır. Farklı 2 öznitelik yöntemi, 2 terim ağırlıklandırma yöntemiyle öznitelikleri çıkarılan veriler, 4 farklı sınıflandırma algoritmasına tabi tutulmuştur. 16 farklı kombinasyona ait durumlar Tablo 4.6’da verilmiştir.

Oluşturulan haber metni veri kümesinin 60 adeti öğrenme, geriye kalan 20 adeti ise test süreci için kullanılmıştır. Öğrenme seti için her bir kategoriden 15’er tane; test kümesi için ise 5’er tane haber metni alınmıştır.

Tablo 4.6. Çalışmada kullanılan yöntem kombinasyonları

	TF	TF-IDF
Kelime Kökü	NB	NB
	DVM	DVM
	J48	J48
	RO	RO
Kelime Seviye N-gram	NB	NB
	DVM	DVM
	J48	J48
	RO	RO

Yapılan çalışmada öznitelik, vektör oluşturma ve sınıflandırma yöntemlerinin 16 farklı kombinasyonunda ortaya çıkan sonuçlar karşılaştırılmış ve başarı oranları Tablo 4.7’de, sınıflandırma başarımları ölçütleri ise Tablo 4.8’de verilmiştir.

Tablo 4.7. Öznitelik yöntemlerine ve TA seçimine göre sınıflandırma başarımları

				Boyutu İndirgenen Veri	
Öznitelik	Sınıflandırma	TA Yöntemi		TA Yöntemi	
		TF	TF-IDF	TF	TF-IDF
Kelime Kökü	NB	90	75	85	90
	DVM	95	90	80	80
	J48	65	65	65	65
	RO	100	100	90	85
Kelime Seviye N-Gram	NB	85	85	50	50
	DVM	65	50	40	40
	J48	50	50	50	50
	RO	75	50	50	65

Tablo 4.8. Sınıflandırma başarımlar ölçütleri

Doğruluk (DY), Duyarlılık (D), Anma Değeri (A), F1 Ölçütü (F),

Sınıflandırma Algoritması	Başarımlar Ölçütleri			
	DY	D	A	F
NB	90	0.9	0.9	0.9
DVM	95	0.95	0.95	0.94
J48	65	0.75	0.65	0.65
RO	1	1	1	1

Tüm deneyler sonucunda ortalama olarak bakıldığında kelime kökü öznelilik türünün kelime seviye bigram'a göre daha başarılı sonuçlar verdiği gözlemlenmiştir. Yapılan dört farklı sınıflandırma türünde ise en başarılı sonuçların rastgele orman ağacı sınıflandırma algoritmasına ait olduğu görülmüştür. Rastgele orman ağacı sınıflandırma algoritması, tek karar ağacı oluşturmak yerine birden fazla karar ağacını her bir düğümde rastgele alınan niteliklerin en iyisini seçerek oluşturduğundan %100 başarı oranıyla verimli bir şekilde çalıştığını ispat etmiştir. Ayrıca çalışma kapsamında öznelilik seçme işlemi için *CfsSubset* algoritması kullanılarak boyut indirgemesi yapılmıştır ve indirgenen boyutların sınıflandırmaya etkisi Tablo 4.7'de gösterilmiştir.

4.1.7. Başarımlar Ölçütleri

Çalışmada kullanılan başarımlar ölçütleri duyarlılık değeri, anma değeri, F1-ölçütü ve doğruluk faktörleridir. Bu ölçütler algoritmaların başarısını ve güvenilirliğini arttıran değerlerdir. Kullanılan modelin başarısı temelde, doğru sınıfa atanan örnek sayısına ve yanlış sınıfla etiketlenen örneklerin sayısına bağlıdır.

Duyarlılık değeri, elde edilen deneysel sonuçların birbirine yakınlık derecesini ölçen başarı değeridir. Bu ölçüt, eşitlik 4.5'te verildiği gibi doğru bir şekilde sınıflandırılmış pozitif örnek sayısının (TP), toplam pozitif örnek sayısına (TP+FP) oranı ile bulunur.

$$\text{Duyarlılık} = \frac{TP}{TP+FP} \quad (4.5)$$

Anma değeri olarak bilinen başarıım ölçütü ise ulaşılması gerekene hangi ölçüde ulaşıldığını gösterir. Anma değeri, eşitlik 4.6'da bulunan doğru bir şekilde sınıflandırılmış pozitif örnek sayısının (TP), toplam dokuman sayısına (TP + FN) bölünmesiyle elde edilir.

$$Anma\ Değeri = \frac{TP}{TP+FN} \quad (4.6)$$

Doğruluk değeri ise yapılan çalışmanın gerçeğe hangi oranda uygun olduğunu göstermek için hesaplanır. Eşitlik 4.7'de verildiği gibi bu değere, doğru bir şekilde belirlenen pozitif ve negatif değerlerin toplam sayısının, tüm verilere oranı ile ulaşılır.

$$Doğruluk = \frac{TP+TN}{TP+FP+TN+FN} \quad (4.7)$$

F1 ölçütü ise yukarıda denklemleri verilen duyarlık ve anma değerlerinin harmonik ortalamaları ile hesaplanır. Makine öğrenmesi için yapılan çalışmalarda en yaygın olarak kullanılan başarıım ölçütü F1-Measure ölçütüdür ve eşitlik 4.8 ile hesaplanır.

$$F1\ Ölçütü = 2 * \frac{Duyarlılık * Anma}{Duyarlılık + Anma} \quad (4.8)$$

4.1.8. Uygulama Çalışması 1 Sonuçları

Teknolojik gelişmelerin büyük boyutlara ulaşması internet kullanımı arttırmıştır. Böylece eldeki verilerin analizine ve kategorizasyonuna olan ihtiyaç kaçınılmaz olmuştur. Metin verilerinin metin madenciliği yöntemleri ile analizi ve metin sınıflandırma yöntemleriyle sınıflandırılması bu ihtiyaçlara çözüm olarak literatüre girmiştir.

Veri Madenciliğinin alt dalları olan bu çalışma alanları yine ihtiyaçlara yönelik olarak gelişmiş ve yelpazesini genişleterek fikir, görüş ve duygu analizi kavramlarının da hayatımıza girmesini sağlamıştır. Böylece metinsel veriler analiz edilerek sahip olduğu etiket, sınıf ve barındırdığı duyguya ulaşılabilir. Bu amaçla yapılan çalışmalardan ilki haber metinlerinin farklı metin madenciliği yöntemleriyle sınıflandırılmasıdır.

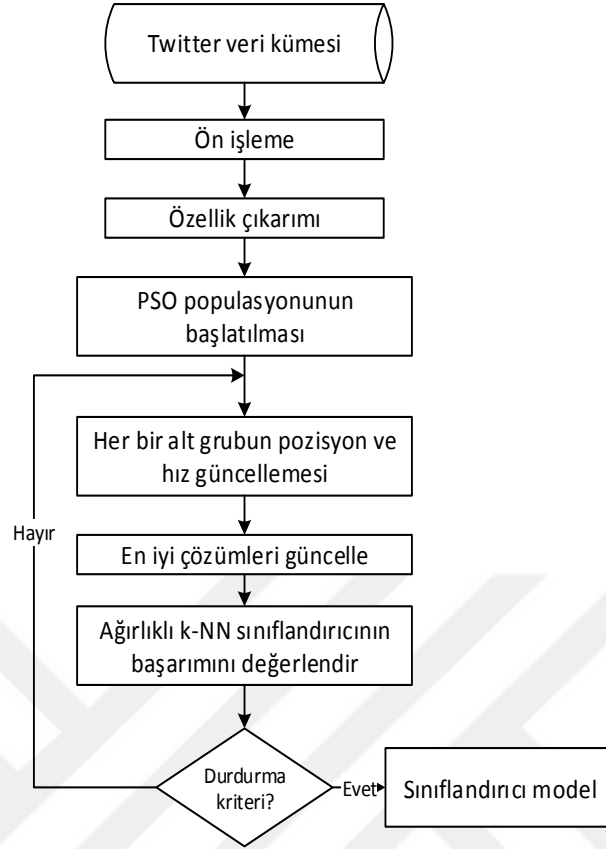
Yapılan çalışmada, metin madenciliği ve metin sınıflandırma kavramlarına, yöntemlerine ve işlem adımlarına yer verilmiştir. Bu anlamda, haber metnlerinin farklı öznitelik türü ve yöntemleriyle sınıflandırılması konulu bir uygulama yapılarak farklı sınıflandırma algoritmaları üzerinde test edilmiştir. Haber sitelerinden alınan, ekonomi, sağlık, politika ve spor başlıklı 4 kategori ve toplam 80 haber metnine ait sınıflandırma testleri sonucunda en iyi sonuçlara, Kelime Kökü öznitelik yöntemi, Terim Frekansı terim ağırlıklandırma yöntemi ve Rastgele Orman Sınıflandırma algoritması kullanılarak, %100'lük başarı oranıyla ulaşıldığı gözlemlenmiştir. Ayrıca, *CfsSubsetEval* boyut indirgeme algoritması kullanılarak boyut indirgeme çalışması yapılmıştır. Boyutu indirgenmiş verilerin zaman ve performans açısından avantaj oluşturduğu gözlemlenmiştir.

4.2. Uygulama Çalışması 2

Tez çalışması kapsamında yapılan bir başka uygulama ile, çoklu popülasyon tabanlı *Parçacık Sürü Optimizasyonu* (PSO) algoritması ile duygu analizi işlemi yapılmıştır [58]. Twitter gönderilerinden oluşan tweetlerin bulunduğu veri setleri pozitif, negatif ve nötr şeklinde üç etiketle sınıflandırılmıştır. Önerilen yöntemin performansı farklı sınıflandırıcılar ile karşılaştırılmış ve daha iyi sonuçlar elde edildiği deneysel olarak ispatlanmıştır.

Önerilen bu yöntem temelde iki adımdan oluşmaktadır. İlk aşamada tweetlerin bulunduğu veri setleri, metin madenciliği ve metin sınıflandırma bölümlerinde bahsedildiği gibi yapısal forma dönüştürülmüştür. Böylece, veri seti gereksiz olan harf, karakter, rakam, noktalama işareti vb. gibi bütün gürültülerden arındırılmıştır. Arındırılan ve yapısal forma dönüştürülen metin verileri üzerinde 9 adet özellik çıkarımı yapılmıştır. Elde edilen özellikler PSO tabanlı ağırlıklı k-NN ile sınıflandırılarak duygu analizi işlemleri tamamlanmıştır.

Çalışma kapsamında önerilen yöntemin blok şeması Şekil 4.9'da verilmiştir. Akış diyagramında gösterildiği gibi, öncelikle veriler ön işleme aşamasına tabii tutulup gürültülü bütün faktörlerden temizlenir. Çıkarılan öznitelikler ile PSO ağırlıklı en yakın komşu algoritması için gerekli olan her bir sınıfa ait küme merkezleri belirlenir ve sınıflandırma işlemi yapılır.



Şekil 4.9. Önerilen yöntemin akış şeması

4.2.1. Twitter Veri Setleri

Önerilen yöntemi test etmek için Twitter ağından alınan İngilizce tweetleri içeren iki veri kümesi kullanılmıştır. Twitter-sanders-apple2 ve Twitter-sanders-apple3 adı verilen bu iki veri seti ile ilgili detaylı bilgi Tablo 4.9’da verilmiştir.

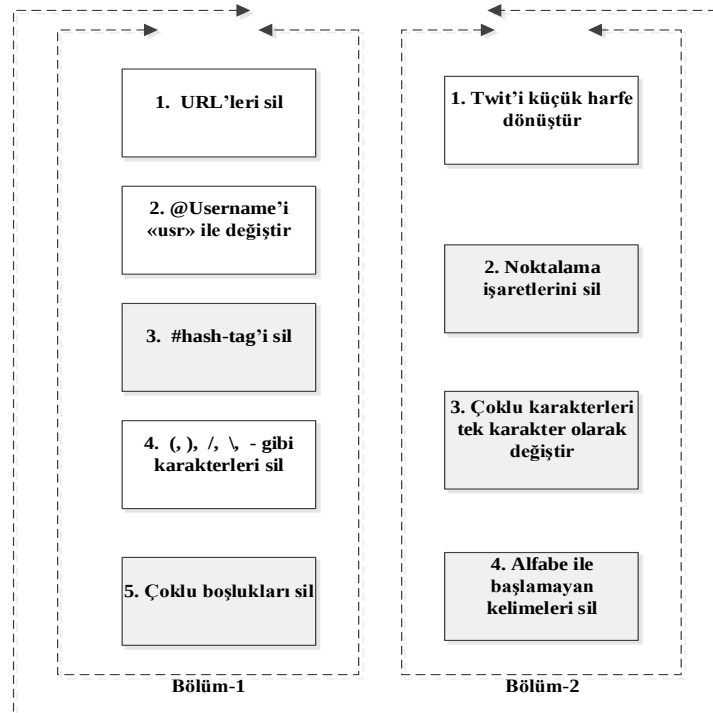
İlk veri kümesi pozitif ve negatif olmak üzere iki sınıf içerirken ikinci veri seti pozitif, negatif ve nötr sınıflarına ait veriler içermektedir. Veri setleri kullanıcıların Twitter üzerinde paylaştıkları tweetlerden oluştuğundan kısaltmaların, sosyal medya jargonunun ve duygu ifade eden emojilerin fazlaca kullanıldığı görülmüştür. Veriler üzerinde bu şekilde sınıflandırma yapmak etkili sonuçlar doğurmayacağından tweetleri öncelikle ön işleme aşamasıyla temizlemek ve yapısal forma dönüştürmek gerekmektedir.

Tablo 4.9 Kullanılan twitter veri kümelerinin özellikleri

Özellikler	Twitter-sanders-apple2 (Veri Kümesi-1)	Twitter-sanders-apple3 (Veri Kümesi-2)
Örnek sayısı	479	988
Sınıf sayısı	2	3
Pozitif	163	163
Negatif	316	316
Nötr	-	509
Dil	İngilizce	İngilizce
Alınma Tarihi	15.10.2011-20.05.2011	
Konu	Apple, Google, Microsoft, Twitter	

4.2.2. Ön İşleme

Gönderiler üzerinde yapılan ön işleme aşamasında Şekil 4.10’da belirtilen işlem adımları izlenerek Twitter veri setinde bulunan istenmeyen gürültü unsurları ortadan kaldırılır. Böylece veri seti sınıflandırmaya hazır bir yapıya dönüşür.

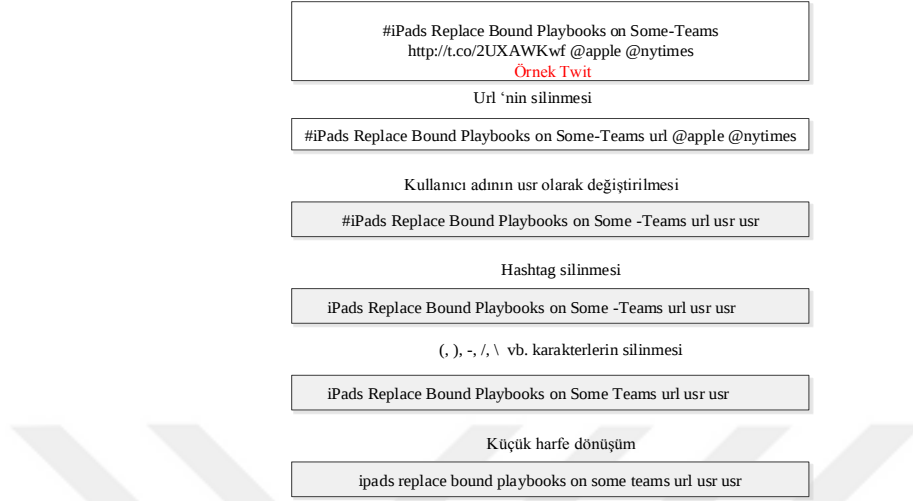


Şekil 4.10. Ön işleme aşaması işlem adımları

Ön işleme aşamasındaki bazı işlemler, düzenli ifade adı verilen ve metinsel veriler üzerinde işlem yapılmasına olanak veren ifadeler yardımıyla yapılmıştır. Çalışmada kullanılan düzenli ifadeler şu şekilde verilmiştir.

44

Veri setinde bulunan örnek bir *tweet* üzerinden Şekil 4.12’de belirtilen ön işlem adımları gösterilmiştir.



Şekil 4.12. Örnek tweet üzerinden ön işleme adımları

4.2.3. Özellik Çıkarımı

Twitter verileri üzerinde ön işleme yapıp gürültüler temizlendikten sonra gönderileri en iyi şekilde temsil eden öznitelikler çıkarılır. Bu çalışmada, her bir tweet için karakter sayısı, pozitif emoji(:,), :D vb.) sayısı, negatif emoji(:(, :'(, >:(vb.) sayısı, nötr emoji sayısı, pozitif ve negatif ünlem ifade eden kelime sayısı, olumsuzluk ekleri, pozitif ve negatif kelimeler olmak üzere toplam dokuz özellik çıkarılmıştır. Çalışma kapsamında çıkarılan 9 öznitelik ve açıklamaları Tablo 4.10’da verilmiştir.

Tablo 4.10. Çıkarılan 9 öznitelik

1	Toplam Karakter Sayısı	Tweetlerdeki toplam karakter sayısını ifade eder.
2	Pozitif Emoji	:),), :D , vb gibi mutlu anları ifade eden sembollerdir [59].
3	Negatif Emoji	:(, :_(, vb. gibi üzgün ve olumsuz bir durumu ifade etmek için kullanılan sembollerdir [59].
4	Nötr Emoji	Herhangi bir duygu ifade etmeyen sembollerdir [59].
5	Pozitif Ünlem	hurrah! wow! ,yeah, waww, vb. gibi güçlü duygu veya görüşleri yansıtmak için kullanılan ifadelerdir [60].
6	Negatif Ünlem	aah!, argh, aw, vb. gibi olumsuz durum tepkilerini ifade ederler [60].
7	Olumsuzlama	no, not, hardly, vb. gibi olumsuzlama eklerini ifade ederler.
8	Pozitif Kelimeler	adaptable, adequate, confidence, vb. gibi pozitif anlam ifade eden tüm kelimeler bu grupta bulunur [61].
9	Negatif Kelimeler	angry, bad, awful, vb. gibi negatif anlam ifade ederler [62].

4.2.4. Parçacık Sürü Optimizasyonu Tabanlı Sınıflandırma

Parçacık sürü optimizasyonu popülasyon tabanlı bir optimizasyon yöntemidir. Her parçacığın pozisyon ve hız bilgisi kendisi ve komşu parçacıkların bilgilerine göre güncellenmektedir. Bu çalışmada kullanılan PSO’da ise her bir parçacık verilen bir örüntüyü tanımak için *k-En Yakın Komşu Algoritmasını (k-NN)* kullanır. Dolayısıyla parçacığın boyutu veri boyutu ile aynı olmalıdır. Amaç, tek bir çözüm bulmak yerine sınıf etiketi kadar optimal çözüm bulmaktır. Her bir sınıf için, N parçacık ve K sınıf sayısı olmak üzere $N \times K$ adet parçacık üretilmektedir. Önerilen yöntemin algoritması Şekil 4.13’te verilmiştir.

```
Algoritma: Çoklu PSO tabanlı ağırlıklı k-NN
VeriKümesi ← Veri kümesini yükle
Populasyon ← rand(N,K)
While(Maksimum adım sayısına erişilmeyene kadar)
  Her sınıf için yarışan/yarışmayan parçacıkları bul
  For her parçacık için do
    Geçerli parçacığın uygunluğunu bul
    Sosyal adaptasyon parametresini güncelle
    Her parçacık sınıfı için yarışan ve yarışmayan en yakın
    parçacıkları
    Yeni en iyi pozisyonu bul
  endfor
  Parçacıkları taşı
  Ağırlıklı k-NN’nin sınıflandırma performansını değerlendir
  Her grubun en iyi pozisyonunu güncelle
endwhile
```

Şekil 4.13. Çoklu PSO tabanlı ağırlıklı k-NN

Önerilen çoklu PSO tabanlı duygu analizi çalışmasında her bir parçacığın hız güncellemesi eşitlik 4.9’a göre yapılmaktadır.

$$\begin{aligned} v_i^{t+1} = & \chi(w \cdot v_i^t + c_1 \cdot \psi_1 \cdot (p_i^t - x_i^t) \\ & + c_2 \cdot \psi_2 \cdot \text{sign}(a_i^t - x_i^t) \cdot S f_i \\ & + c_3 \cdot \psi_3 \cdot \text{sign}(x_i^t - r_i^t) \cdot S f_i) \end{aligned} \quad (4.9)$$

Eşitlik 4.9’ da, c_1 , c_2 ve c_3 sabit ağırlık faktörlerini gösterirken; χ kısıtlama faktörünü ve w ise moment ağırlığını gösterir. Parçacığın t anındaki pozisyonu x_i^t ve i . parçacık tarafından o ana kadar en iyi pozisyon p_i^t ile verilmiştir.

Eşitlikte ψ_1 , ψ_2 ve ψ_3 değerleri ise $[0,1]$ aralığında rastgele üretilen sayıları gösterir. Sosyal uyum parametresi, i . parçacık için çekim ve itme merkezleri sırasıyla Sf_i , a_i ve r_i ile verilmiştir. Ayrıca, c_1 çarpanının olduğu terim bir parçacığın o ana kadarki en iyi pozisyonuna göre güncellemeyi ifade ederken c_2 ve c_3 çarpanlarının olduğu terim ise parçacığın ait olduğu sınıfa göre bir çekim merkezine çekilmesi ve ait olmadığı sınıf tarafından itilmesini sağlar. Hız güncellemesi kullanılarak parçacığın yeni pozisyonu eşitlik 4.10 ile verilebilir.

$$x_i^{t+1} = x_i^t + v_i^{t+1} \quad (4.10)$$

Eşitlik 4.10’da verilen pozisyon güncellemesi standart PSO’da olduğu gibidir ve her bir parçacığın yerel uygulğunun hesaplanması gerekmektedir. Yerel uygunluk parçacığın sınıflandırmadaki başarımını gösterir. Bunun için parçacığın, verileri ne kadar iyi ve kötü sınıflandırdığını gösteren iki parametre kullanılmaktadır. Bu parametreler eşitlik 4.11’de verilmiştir.

$$\begin{aligned} I_f &= \sum_{j \in \{g\}} \frac{1}{d_{ij}+1} \\ K_f &= \sum_{j \in \{b\}} \frac{1}{d_{ij}+1} \end{aligned} \quad (4.11)$$

Eşitlikte $\{g\}$, i . parçacık tarafından doğru bir şekilde sınıflandırılan örnek sayısını gösterirken; $\{b\}$ ise i . parçacık tarafından yanlış olarak sınıflandırılan örnek sayısını göstermektedir. Bir parçacık ile en yakın örnek arasındaki mesafe ise d_{ij} ile gösterilir. Elde edilen parametrelere göre her bir parçacığın uygunluğu eşitlik 4.12’ye göre hesaplanmaktadır.

$$Uygunluk = \begin{cases} \frac{I_f}{toplamlam} + 2, & \text{Eğer } \{g\} \neq 0, \{b\} = 0 \\ \frac{I_f - K_f}{toplamlam} + 1, & \text{Eğer } \{b\} \neq 0 \\ 0, & \text{Eğer } \{b\} = \{g\} = 0 \end{cases} \quad (4.12)$$

Eşitlik 4.12’de *toplamlam* değişkeni, toplam örnek sayısını göstermektedir. Verilen denklemlere göre eğer bir parçacık yanlış örneği sınıflandırmıyorsa, parçacığın uygunluğu en az 2 değerini alır. Eğer parçacık doğru örneklerin yanında yanlış örnek de sınıflandırıyorsa uygunluğu 1 ile 2 arasında olur. Eğer parçacığın sınıflandırdığı bir örnek yok ise uygunluğu 0 olur.

Parçacıkların komşuluk durumu için itme ve çekme parametreleri kullanılmaktadır. Sosyal parametre faktörü parçacıkların sınıf için yarışmayan parçacıklara doğru hareketini ve yarışanlardan uzak olmalarını kontrol eder. Bu parametre uygunluk ile ters orantılıdır ve eşitlik 4.13’te verilmiştir.

$$Sf_i = \frac{1}{1+Uygunluk_i} \quad (4.13)$$

Her bir adımdan sonra parçacıklar sınıf verilerine göre durumlarını güncellemekte ve parçacıkların yeni pozisyonlarına göre ağırlıklı en yakın komşu algoritması ile sınıflandırma yapılmaktadır.

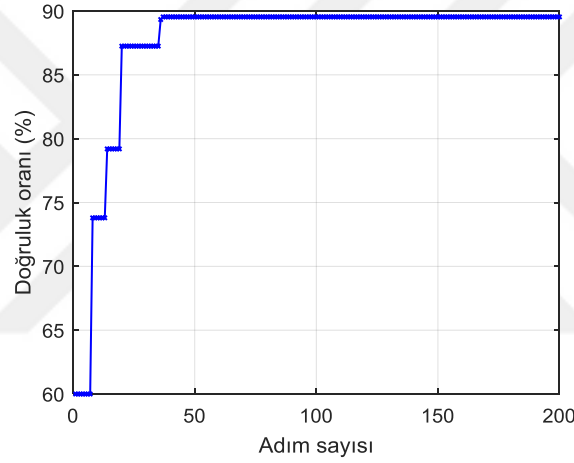
4.2.5. Uygulama Çalışması 2 Sonuçları

Parçacık sürü optimizasyonunda her bir sınıf başına 10 parçacık kullanılmıştır. Dolayısıyla ilk veri kümesinde popülasyon boyutu 20; diğer veri kümesinde ise popülasyon boyutu 30 olarak belirlenmiştir. Parçacık sürü optimizasyonunun diğer özellikleri Tablo 4.11’de verilmiştir.

Tablo 4.11. Önerilen PSO'nun parametreleri

Parametre	Değeri
χ	0.5
w	0.1
c_1	1.0
c_2	1.0
c_3	0.25

Her bir veri kümesi için parçacık sürü optimizasyonu tabanlı yöntem 10 kez çalıştırılmış ve 10 noktalı çapraz doğrulama ile doğrulanmıştır. Her bir veri kümesi için PSO tabanlı ağırlıklı k-NN algoritmasının başarımlı performansı Şekil 4.14'te verilmiştir.



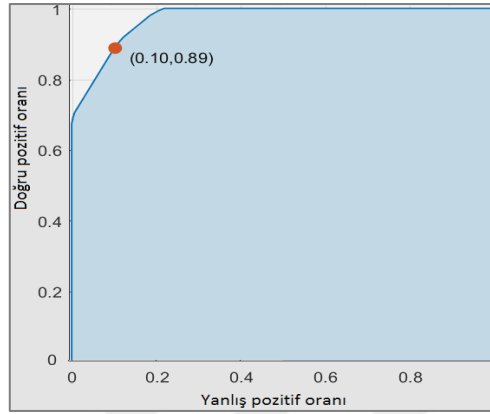
Şekil 4.14. PSO tabanlı ağırlıklı k-NN sınıflandırma performansı

Şekil 4.14'de elde edilen sonuçlara göre önerilen sınıflandırma yöntemi ortalama 50 adım sonrasında maksimum performansa ulaşmaktadır. Tablo 4.12'te bu veri kümesi için elde edilen karmaşıklık matrisi verilmiştir.

Tablo 4.12. Veri kümesi-1 için karmaşıklık matrisi

		Tahmin Edilen Sınıf	
		Pozitif	Negatif
Doğru Sınıf	Pozitif	145	18
	Negatif	32	285

Tablo 4.12’de gösterildiği gibi Twitter veri kümesi-1 üzerinde önerilen yöntem ile yapılan sınıflandırmada %89.6 başarımla elde edilmiştir. Önerilen yöntem için elde edilen ROC eğrisi ise Şekil 4.15’te verilmiştir.



Şekil 4.15. Twitter veri kümesi-1 için ROC eğrisi

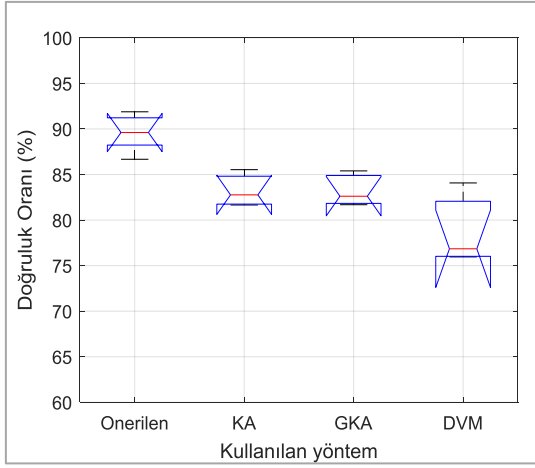
Şekil 4.15’te verilen ROC eğrisine göre önerilen yöntemin verilen veri kümesi üzerinde oldukça iyi sonuçlar elde ettiği görülmektedir.

Twitter veri kümesi-2 ise nötr sınıfının da yer aldığı bir veri setidir ve yine aynı yöntemle test edilmiştir. Önerilen yöntemin twitter veri kümesi-2 üzerindeki başarımlarını gösteren karmaşıklık matrisi Tablo 4.13’te verilmiştir.

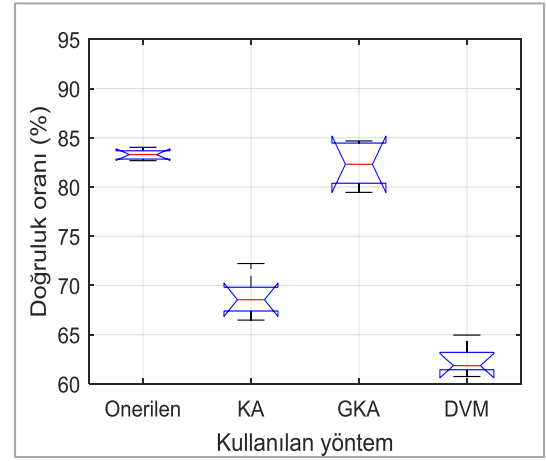
Tablo 4.13. Veri kümesi-2 için karmaşıklık matrisi

		Tahmin Edilen Sınıf		
		Pozitif	Negatif	Nötr
Doğru Sınıf	Pozitif	120	13	31
	Negatif	21	257	38
	Nötr	29	49	445

Tablo 4.13’te gösterildiği gibi twitter veri kümesi-2 üzerinde yöntemin başarımlar performansı ise %83.19 olarak elde edilmiştir. Önerilen yöntem aynı veri seti üzerinde uygulanan farklı yöntemler ile karşılaştırılmıştır. İki veri seti için önerilen yöntemin performans karşılaştırma sonuçları Şekil 4.16’da verilmiştir.



(a) Twitter veri seti-1



(b) Twitter veri seti -2

Şekil 4.16. İki veri kümesi üzerinde farklı yöntemlerin karşılaştırılması

Şekil 4.16’da önerilen yöntem, karar ağacı (KA), gugukkuşu optimizasyon algoritması tabanlı kümeleme (GKA) [1] ve DVM ile iki veri seti üzerinde karşılaştırılmıştır. GKA algoritmasında 11 öznitelik çıkarılmasına rağmen bu çalışmada 9 nitelik çıkarılmış ve elde edilen sonuçların daha iyi olduğu gözlemlenmiştir.

Çalışma kapsamında, farklı iki twitter veri kümesi üzerinde duygu analizi için PSO tabanlı ağırlıklı k-NN algoritması önerilmiştir. Önerilen yöntem ile Twitter veri kümesindeki her bir twetten 9 öznitelik elde edilmiştir. Yöntem, literatürde var olan diğer yöntemler ile karşılaştırılmış ve başarılı sonuçlar elde ettiği görülmüştür. Daha önce aynı veri kümesi üzerinde uygulanan guguk kuşu algoritmasında 11 öznitelik çıkarılmıştır. Önerilen yöntemde ise özellik sayısı azaltılmasına rağmen hem çalışma süresi açısından hem de doğruluk oranı açısından daha iyi bir performans elde edilmiştir. Ayrıca bu yöntemle, parçacıkların bütün veri kümesini sınıflandıracak şekilde dağılımı sağlanmıştır. İkinci veri kümesi ise ilk veri kümesine nötr sınıfı da eklenerek test edilmiştir. Nötr sınıfı performansı düşürmesine rağmen önerilen yöntemin daha iyi olduğu deneysel olarak ispatlanmıştır.

5. SONUÇLAR

İnternetin gün geçtikçe insan hayatına daha çok girdiği yadsınamaz bir gerçektir. Bu durum teknolojik gelişmelerle birleştiğinde internet ortamında boyutu maksimuma ulaşan bir veri olarak karşımıza çıkmaktadır. Veri boyutundaki artış veriyi kalitesizleştirmiş ve veriye olan ulaşımı zorlaştırmıştır. Bunun yanında, sosyal medya kullanımının bağımlılık seviyesine ulaştığı günlerde birbirinden farklı onlarca sosyal medya platformunda her an her dakika binlerce gönderi paylaşılmaktadır. Hem internet ortamında bulunan metinsel verilerin hem de sosyal ağ ortamlarında paylaşılan metinsel gönderilerin bu denli artması, metinleri analiz etmeyi ve bu verilerin istatistiki çalışmalar için kullanılmasını zorunlu kılmıştır. Metin analizi amacı ile literatürde çeşitli çalışmalar başlamış ve metinsel verilerin metin madenciliği yöntemleri ile analizi ve metin sınıflandırma yöntemleriyle sınıflandırılması, belirtilen problem ve ihtiyaçlara çözüm olmuştur.

Metinsel analizin başarılı sonuçlar verdiğini ve başarılı kararlar alınmasına yardım ettiğini gören her sektör, metin analizinin kaçınılmaz bir gerçek olduğunu fark etmiştir. Toplumun görüşü, fikri, yorumu ve en önemlisi duygusu ile çalışan bütün sektörler, duygu analizi konusu ile yakından ilgilenmeye başlamıştır. Temelde doğal dil işleme, metin madenciliği ve metin sınıflandırma esaslarına dayanan duygu analizi, metinsel analiz yaparak metnin barındırdığı duyguyu ön plana çıkarmayı hedeflemektedir. Metinsel veri kaynağı olarak ise güçlü sosyal ağ platformlarından faydalanılmaktadır. Facebook, Twitter, vb. birçok ağ, bu çalışma için zengin veri kaynağı sunmaktadır. Herhangi bir firma duygu analizi ile, çıkarmış olduğu bir ürünle ilgili müşteri görüşlerini değerlendirerek üründe güncelleme yapabilir veya ürün çok kötüyse kaldırabilir. Herhangi bir yapımcı, bir sinema filmini internette bulunan film sitelerindeki izleyici yorumlarına göre detaylandırabilir, oyuncu kadrosunu, bütçesini ve daha birçok detayı kullanıcı yorumları ile oluşturabilir. İşte tüm bu imkânlar duygu analizinin tüm sektörler üstündeki gücünü ortaya koymaktadır. Duygu analizinin kullanıcı açısından öneminden bahsetmek gerekirse, bir okuyucu okumayı düşündüğü bir kitapla ilgili yapılan kullanıcı yorumlarını inceleyip okuyup okumamaya ya da herhangi bir ürün hakkında kullanıcı görüşlerini okuyarak ürünü kullanıp kullanmamaya karar verebilir.

Tüm bu faydaları düşünüldüğünde duygu analizi ve metin sınıflandırmanın hem günümüzün hem de geleceğin vazgeçilmez çalışma alanları olduğu ve olacağı görülmektedir. Bu anlamda tez çalışması için metin madenciliği, metin sınıflandırma ve duygu analizi çalışmaları yapılmıştır.

Yapılan ilk uygulamada internet üzerinden bulunan farklı haber sitelerinden alınan haber metinleri farklı metin madenciliği yöntemleriyle kategorize edilmiştir. İlk aşamada her türlü gürültülü unsurdan temizlenen veri için iki tür öznitelik kullanılmıştır. Kelime kökü ve kelime seviye N-gram olarak kullanılan bigram ile gerekli öznitelikler çıkarılmıştır. Farklı terim ağırlandırma yöntemleriyle sayısal formata dönüştürülen ve boyutu indirgenen öznitelikler, 4 farklı sınıflandırma algoritması ile Weka platformu üzerinde sınıflandırılmıştır. Elde edilen sınıflandırma sonuçları karşılaştırılarak değerlendirilmiştir. Yapılan değerlendirme sonrasında en iyi başarı oranına Rastgele Orman Algoritması ile ulaşılmıştır.

Yapılan ikinci uygulamada ise Twitter üzerinde belirli konular ve ürünler hakkında kullanıcılar tarafından paylaşılan gönderilerin oluşturduğu iki veri seti Parçacık Sürü Optimizasyonu tabanlı, ağırlıklı en yakın komşu algoritması ile sınıflandırılmıştır. Ön işleme aşamasından geçirilerek gereksiz tüm unsurlardan temizlenen veri setlerinde bulunan her bir tweet için 9 ayrı öznitelik çıkarılmıştır. Çıkarılan öznitelikler çalışma için önerilen PSO tabanlı k-NN ile sınıflandırılmıştır. Daha önce aynı veri kümeleri üzerinde önerilen guguk kuşu algoritması ile karşılaştırıldığında önerilen yöntemin daha başarılı sonuçlar verdiği gözlemlenmiştir.

Duygu analizi ve metin sınıflandırma temeline dayalı olarak yapılan çalışmalar metinsel verilerin farklı yöntemler altında sınıflandırılması ve metin altında yatan duyguların duygu analizi yöntemleriyle keşfedilmesi konularına katkı sağlamıştır.

KAYNAKLAR

- [1] **Pandey A.C, Rajpoot D.S., Saraswat D.**, “Twitter Sentiment Analysis Using Hybrid Cuckoo Search Method”, *Information Processing and Management* 53 764–779, 2017.
- [2] **B. Liu**, *Sentiment Analysis and Opinion Mining*, Morgan & Claypool Publishers, 2012 .
- [3] **Türkmenoğlu C.**, “Türkçe Metinlerde Duygu Analizi”, Yüksek Lisans Tezi, İstanbul Teknik Üniversitesi Fen Bilimleri Enstitüsü, 2015.
- [4] **Akgül E.S., Ertano C. ve Diri B.**, “Sentiment Analysis with Twitter”, *Pamukkale Üniversitesi Mühendislik Bilimleri Dergisi* 22(2), 106-110, Denizli, 2016.
- [5] **Güran A., Uysal M., Doğrusöz Ö.**, “Destek Vektör Makineleri Parametre Optimizasyonunun Duygu Analizi Üzerindeki Etkisi”, *Dokuz Eylül Üniversitesi Mühendislik Fakültesi Mühendislik Bilimleri Dergisi* Cilt: 16 Sayı: 48 sh. 86-93, İzmir, 2014.
- [6] **Kumar R., Vadlamani R.**, “A Survey on Opinion Mining and Sentiment Analysis: Tasks, Approaches and Applications”, *Knowledge-Based Systems* 89, 14–46, 2015.
- [7] **Onan A., Korukoğlu S., Bulut H.**, “A Hybrid Ensemble Pruning Approach Based on Consensus Clustering and Multi-Objective Evolutionary Algorithm for Sentiment Classification” , *Information Processing and Management* 53, 814–833, 2017.
- [8] **Keshavarz H., Abadeh M.S.**, “ALGA: Adaptive Lexicon Learning Using Genetic Algorithm for Sentiment Analysis of Microblogs”, *Knowledge-Based Systems* 122 1–16, 2017.
- [9] **F.H. Khan , U. Qamar , S. Bashir** , “SWIMS: Semi-Supervised Subjective Feature Weighting and Intelligent Model Selection for Sentiment Analysis, *Knowledge-Based Systems*, 100 97–111, 2016.
- [10] **Nasukawa T., ve Yi J.**, “Sentiment Analysis: Capturing Favorability Using Natural Language Processing”, *Proceedings of The 2nd International Conference on Knowledge Capture*, 494–500, 2003.
- [11] **Beyhan H.D.**, “Sosyal Medya Üzerinden Metin Madenciliği ve Duygu Analizi ile Pazar Değerlendirme”, Yüksek Lisans Tezi, İstanbul Teknik Üniversitesi Fen Bilimleri Enstitüsü, 2014.

- [12] **Çoban Ö.**, “Metin Sınıflandırma Teknikleri ile Türkçe Twitter Duygu Analizi”, Yüksek Lisans Tezi, Atatürk Üniversitesi Fen Bilimleri Üniversitesi, 2016.
- [13] **Giriş S.F.**, “Sentimedia:Opinion Mining and Sentiment Analysis on Social Media”, Fatih University, Master Thesis, 2014.
- [14] **Çetin M., Amasyalı M.F.**, “Supervised and Traditional Term Weighting Methods for Sentiment Analysis”, YTÜ, 2013.
- [15] **Nalçakan Y., Bayramoğlu Ş.S., Tuna S.**, “Sosyal Medya Verileri Üzerinde Yapay Öğrenme ile Duygu Analizi Çalışması”, Trakya Üniversitesi, 2015.
- [16] **B. Pang, L. Lee, S. Vaithyanathan**, “Sentiment Classification Using Machine Learning Techniques”, Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing, vol. 10, Association for Computational Linguistics, pp. 79–86, 2002.
- [17] **Saif, H. et. al.**, “Alleviating Data Sparsity for Twitter Sentiment Analysis”, The 2nd Workshop on Making Sense of Microposts, 2012.
- [18] **Kennedy H.**, “ Perspectives on Sentiment Analysis” , Journal of Broadcasting & Electronic Media, 2012 , 56(4), 435-450.
- [19] **Diri B., Amasyalı, M. F.**, “Automatic Author Detection for Turkish Texts”, Artificial Neural Networks and Neural Information Processing, (pp. 138-141), 2003.
- [20] **Kasıkçı T., Gökçen H.**, “Metin Madenciliği ile E-Ticaret Sitelerinin Belirlenmesi” Bilişim Teknolojileri Dergisi, 2014.
- [21] **Güran, A., Akyokuş, S., Bayazıt, N. G., Gürbüz, M. Z.**, “Turkish Text Categorization Using N-Gram Words”, Proceedings of the International Symposium on Innovations in Intelligent Systems and Applications (INISTA2009) (pp. 369-373), 2009.
- [22] **Nigam, K., McCallum, A. K., Thrun, S., Mitchell, T.**, “Text Classification from Labeled and Unlabeled Documents Using EM”, Machine learning, 39(2-3), 103-134, 2000.
- [23] **Lin, Y. S., Jiang, J. Y., Lee, S. J.**, “A Similarity Measure For Text Classification And Clustering. Knowledge and Data Engineering”, IEEE Transactions on, 26(7), 1575-1590, 2014.
- [24] **Ko, Y.,”** August. A Study of Term Weighting Schemes Using Class Information for Text Classification”, Proceedings of The 35th International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 1029- 1030), 2012.

- [25] **B. Pang, L. Lee, S. Vaithyanathan**, “Sentiment Classification Using Machine Learning Techniques”, Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing, vol. 10, Association for Computational Linguistics, pp. 79–86, 2002.
- [26] **Pak, A., and Paroubek, P.**, “Twitter as a Derlem for Sentiment Analysis and Opinion Mining”, LREC, 2010.
- [27] **Saif, H. et. al.**, “Alleviating Data Sparsity for Twitter Sentiment Analysis”, The 2nd Workshop on Making Sense of Microposts, 2012.
- [28] **Habernal, I. et. al.**, Sentiment Analysis in Czech Social Media Using Supervised Machine Learning, WASSA: 65, 2013.
- [29] **Abd. S.H. Basari, B. Hussin, I.G.P. Ananta, J. Zeniarja**, “Opinion Mining of Movie Review Using Hybrid Method of Support Vector Machine and Particle Swarm Optimization”, P roc. Eng. 53, 453–462, 2013
- [30] <http://cs.stanford.edu/people/alecmgo/trainingandtestdata.zip>.
- [31] **Erogul U.**, “Sentiment analysis in Turkish”, Middle East Technical University, Ms Thesis, 2009.
- [32] **Simsek M. ve Ozdemir S.**, “Analysis of The Relation Between Turkish Twitter Messages and Stock Market İndex”, Application of Information and Communication Technologies (AICT), 6th International Conference, s.1–4, 2012.
- [33] **Kaya, M., Fidan, G., Toroslu, I. H.**, “December. Sentiment Analysis of Turkish Political News”, Proceedings of the The 2012 IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technology- Volume 01 (pp. 174-180), 2012.
- [34] **Türkmenoğlu, C., Tantuğ, A. C.**, “Sentiment Analysis in Turkish Media”, International Conference on Machine Learning(ICML), Beijing, 2014.
- [35] **Çoban, Ö., Tümüklü G., Özyer B.**, “Sentiment Analysis for Turkish Twitter Feeds”, 978-1-4673-7386-9/15 IEEE, 2015.
- [36] **Başkaya, F., Aydın, İ.**, “Haber Metinlerinin Farklı Metin Madenciliği Yöntemleriyle Sınıflandırılması”, International Artificial Intelligence and Data Processing Symposium (IDAP), Türkiye (Hakem değerlendirmesinde).
- [37] **Aydın, İ., Başkaya, F., Salur, M.U.**, “PSO Tabanlı Ağırlıklandırılmış K-EYK ile Duygu Sınıflandırma”, Uluslararası Bilgisayar Bilimleri ve Mühendisliği Konferansı (UBMK), Türkiye (Hakem değerlendirmesinde).
- [38] **Tan, H.**, Text Mining: The State of The Art and The Challenges, PAKDD'99 workshop on Knowledge Discovery from Advanced Databases, 1999.

- [39] **Melek, C.**, “Metin Madenciliği Teknikleri ile Şirketlerin Vizyon İfadelerinin Analizi”, Yüksek Lisans Tezi, Dokuz Eylül Üniversitesi Sosyal Bilimler Enstitüsü, 2012.
- [40] **Sancar, Y.**, “Metin Madenciliği Kullanılarak Talep Tanıma ve Yönlendirme Sistemi”, Yüksek Lisans Tezi, Atatürk Üniversitesi Fen Bilimleri Enstitüsü, 2016.
- [41] **Dolgun, M.Ö., Özdemir, T., Oğuz, D.**, “Veri Madenciliğinde Yapısal Olmayan Verinin Analizi: Metin ve Web Madenciliği”, İstatistikçiler Dergisi 2, s:48-58, 2009.
- [42] **Ramesh, B., Sathiaselvan, J.**, “An Advanced Multi Class Instance Selection based Support Vector Machine for Text Classification”, 3rd International Conference on Recent Trends in Computing 2015 (ICRTC-2015), 2015.
- [43] **Lagus, K.**, “Text Mining with the WEBSOM”, Helsinki University of Technology, Finland, 2000.
- [44] **Tantuğ A.C.**, “Metin Sınıflandırma”, İTÜ Bilgisayar ve Bilişim Fakültesi, Bilgisayar Mühendisleri Bölüm Başkanlığı Kurulu Dergisi, 2014.
- [45] **P. Hall, B. U. Park, R. J. Samworth** "Choice of Neighbor Order in Nearest-Neighbor Classification", vol.36, pp.2135-2152, **2008**.
- [46] **Şeker, Ş.E.**, <http://bilgisayarkavramlari.sadievrenseker.com/>, erişim:20.07.2017.
- [47] **McCallum A., Nigam K., (2003)**, “A Comparison of Event Models for Naive Bayes Text Classification”, Journal of Machine Learning Research, vol.3, pp.1265-1287.
- [48] **Cooper G. F.**, “A Bayesian Method for the Induction of Probabilistic Networks from Data”, Kluwer Academic Publishers,, Boston Manufactured in the Netherlands, 1992.
- [49] **Haltaş A., Alkan A.**, “Karabulut M., “Metin Sınıflandırmada Sezgisel Arama Algoritmalarının Performans Analizi”, Journal of the Faculty of Engineering and Architecture of Gazi University ,Cilt 30, No 3, 417-427, 2015.
- [50] **Erten F.**, “Metin Madenciliği Tabanlı Bir Web Sitesi Sınıflandırma Aracı Tasarımı”, Maltepe Üniversitesi, YL Tezi, 2015.
- [51] **Karaca, M.F.**, “Metin Madenciliği Yöntemi ile Haber Sitelerindeki Köşe Yazılarının Sınıflandırılması”, Yüksek Lisans Tezi, Karabük Üniversitesi Fen Bilimleri Enstitüsü, 2012.
- [52] **Çalış, K, Gazdağı, O. Ve Yıldız O.**, “Reklam İçerikli E-Postaların Metin Madenciliği Yöntemleri ile Otomatik Tespiti”, Gazi Üniversitesi, Bilişim Teknolojileri Dergisi, Cilt: 6, Sayı: 1, 2013.

- [53] code.google.com/p/zemberek/
- [54] **Bozan., Y.S., Çoban, O., Tümöklü, G. ve Özyer.,B.**, “SMS Spam Filtering based on Text Classification and Expert System”, Atatürk Üniversitesi, IEEE, 2015.
- [55] **Deng, Z. ve Zhang, M.**, “Improving Text Categorization Using the Importance of Words in Different Categories”, Computational Intelligence and Security, 3801: 458-463, 2005.
- [56] **Hall, M.A.**, “Correlation-Based Feature Selection for Machine Learning”, The University of Waikato, 1999.
- [57] <http://www.cs.waikato.ac.nz/ml/weka/>
- [58] **Cervantes,A., Galvan, I.M. ve Isasi, P.**, “AMPSO: A New Particle Swarm Method for Nearest Neighborhood Classification”, IEEE Trans. On Systems, Man, and Cybernetics-Part B, Vol. 39, No.5, pp. 1082-1091, 2009.
- [59] **Pozitif, Negatif ve Nötr Emojiler :** <http://www.netlingo.com/smileys.php>
- [60] **Pozitif ve Negatif Ünlemler :** <http://www.vidarholen.net/contents/interjections/>
- [61] **Pozitif Kelimeler :** twitter-sentiment-analysis-tutorial-201107/data/opinion-lexicon-English/positive-words.txt. 2011, erişim tarihi: 25.06.2017.
- [62] **Negatif Kelimeler :** twitter-sentiment-analysis-tutorial-201107/data/opinion-lexicon-English/Negative-words.txt. , erişim tarihi: 25.06.2017.
- [63] **Varol, M.**, “Metin Madenciliği Yöntemlerini Kullanarak”, Yüksek Lisans Tezi, Süleyman Demirel Üniversitesi Fen Bilimleri Enstitüsü, 2011.
- [64] **Aydın, İ., Karaköse, M. ve Akın E.**, “Zaman Serisi Veri Madenciliği ve Destek Vektör Makineleri Kullanan Yeni Bir Akıllı Arıza Sınıflandırma Yöntemi”, Gazi Üniversitesi Mühendislik Mimarlık Fakültesi Dergisi, Cilt:23, No 2, 431-440, 2008.
- [65] **Uzun, E.**, www.e-adys.com, erişim tarihi: 25.07.2017
- [66] **Breiman, L.**, “Random Forests. Machine Learning”, 45(1), 5-32. doi: 10.1023/A:1010933404324, 2001.

ÖZGEÇMİŞ

Fatma BAŞKAYA Elazığ' da doğdu. İlk, orta ve lise öğrenimini Elazığ'da tamamladı. 2008 yılında Elazığ Mehmet Akif Ersoy Yabancı Dil Ağırlıklı Lisesi'nden mezun oldu. 2013 yılında ise Süleyman Demirel Üniversitesi Mühendislik Fakültesi Bilgisayar Mühendisliği Bölümünden mezun oldu. 2011 yılında ilk stajını Aselsan A.Ş. Aviyonik Sistemler Bölümü'nde yaptı. İkinci stajını ise 2012 yılında Fırat Üniversitesi Bilgi İşlem Daire Başkanlığı'nda yaptı. 2013 yılında Fırat Üniversitesi Fen Bilimleri Enstitüsü Bilgisayar Mühendisliği Ana Bilim dalında yüksek lisans eğitimine başladı ve Türkiye Bilişim Derneği koordinatörlüğünde yürütülen Turkcell Teknoloji Liderleri programına katılmaya hak kazandı. 2015 yılında Fırat Üniversitesi Bilgi İşlem Daire Başkanlığı kapsamında yarı zamanlı çalıştı.