KARADENİZ TEKNİK ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

İSTATİSTİK VE BİLGİSAYAR BİLİMLERİ ANABİLİM DALI

TÜRKÇE SOSYAL MEDYA METİNLERİNDE DUYGU ANALİZİ

YÜKSEK LİSANS TEZİ

Hasan AMANET

HAZİRAN 2017
TRABZON

KARADENİZ TEKNİK ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

İSTATİSTİK VE BİLGİSAYAR BİLİMLERİ ANABİLİM DALI

TÜRKÇE SOSYAL MEDYA METİNLERİNDE DUYGU ANALİZİ

Hasan AMANET

Karadeniz Teknik Üniversitesi Fen Bilimleri Enstitüsünce
"YÜKSEK LİSANS (İSTATİSTİK)"
Unvanı Verilmesi İçin Kabul Edilen Tezdir.

Tezin Enstitüye Verildiği Tarih : 29 / 05 /2017
Tezin Savunma Tarihi : 16 / 06 /2017

Tez Danışmanı : Yrd. Doç. Dr. Tolga BERBER

Trabzon 2017

KARADENİZ TEKNİK ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

İstatistik ve Bilgisayar Bilimleri Anabilim Dalında
Hasan AMANET Tarafından Hazırlanan

TÜRKÇE SOSYAL MEDYA METİNLERİNDE DUYGU ANALİZİ

başlıklı bu çalışma, Enstitü Yönetim Kurulunun 16 / 05 / 2017 gün ve 1702 sayılı kararıyla oluşturulan jüri tarafından yapılan sınavda
YÜKSEK LİSANS TEZİ
olarak kabul edilmiştir.

Jüri Üyeleri

Başkan : Prof. Dr. Osman MERT .....

Üye : Doç. Dr. Zafer KÜÇÜK .....

Üye : Yrd. Doç. Dr. Tolga BERBER .....

Prof. Dr. Sadettin KORKMAZ
Enstitü Müdürü

ÖN SÖZ

Burada sunulan çalışmayla metinlerin içerdikleri duygular, Metin Madenciliği ve Makine Öğrenmesi teknikleri kullanılarak tespit edilmiştir. Duygu Analizi adı verilen bu analiz işlemi yapılırken Twitter ortamından elde edilen metinler kullanılmıştır.

Sosyal medya ortamlarında paylaşılan metinlerin Türkçe yazı diline bağlı olmadan yazıldığı ve sosyal medya ortamına özgü bir yazı dilinin kullanıldığı tespit edilmiştir. Sözlü anlatım şekilde yazılan metinlerin, Türkçe yazım kurallarına uymaması Metin Madenciliğinin aşamalarından olan metin ön işleme aşamasını zorlaştırmaktadır. Türkçe yazım dilinde olmayan kısaltmaların kullanılması ve harf tekrarlarının olması gibi durumlar ön işleme adımını zorlaştırmaktadır. Karşılaşılan zorluklar, ön işleme teknikleri geliştirilerek giderilmiştir.

Çalışma olarak "Türkçe Sosyal Medya Mesajlarında Duygu Analizi" konusu seçilmiş, bu bağlamda sosyal medya üzerinde duygu analizi ve nitelik seçimi işlemleri irdelenmiştir. Öncelikle elde edilen Twitter metinlerinin duygu durumları olumlu ve olumsuz olmak üzere iki ana gruba ayrılmıştır. Yapılan saha çalışması ile olumlu ve olumsuz duygu durumlarına ait 10 duygu kategorisi belirlenmiştir. Bu duygular "Mutluluk / Neşe Duygusu, Takdir Duygusu, Dilek Duygusu, Güven Duygusu, Gurur Duyma Duygusu, Aşağılama Duygusu, Merak Duygusu, Öfke Duygusu, Hayal Kırıklığı Duygusu, Tavsiye Duygusu" olarak tespit edilmiştir.

Her bir duygu için nitelikler belirlenmiş ve sosyal medyadan elde edilen verilerin sistemde otomatik anlamlandırılması için pratik çözüm önerileri sunulmuştur.

Bu tez çalışmasında bilgisi ve yol göstericiliğiyle bana destek olan değerli görüşlerinden yararlandığım tez danışmanım Yrd. Doç. Dr. Tolga BERBER'e teşekkür ederim.

Tezin oluşmasında desteğini esirgemeyen, yol almamda yapıcı katkıları için Prof. Dr. Asiye Mevhibe ÇOŞAR'a teşekkür ederim.

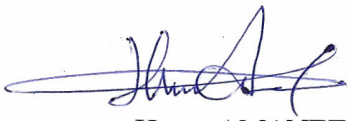
Eğitim hayatım boyunca ihmal ettiğim aileme, eğitimim boyunca bana verdikleri destekten dolayı teşekkür ederim.

Hasan AMANET

Trabzon 2017

TEZ ETİK BEYANNAMESİ

Yüksek Lisans Tezi olarak sunduğum "Türkçe Sosyal Medya Metinlerinde Duygu Analizi" başlıklı bu çalışmayı baştan sona kadar danışmanım Yrd. Doç. Dr. Tolga BERBER'in sorumluluğunda tamamladığımı, verileri kendim topladığımı, analizleri ilgili laboratuvarlarda yaptığımı, başka kaynaklardan aldığım bilgileri metinde ve kaynakçada eksiksiz olarak gösterdiğimi, çalışma sürecinde bilimsel araştırma ve etik kurallara uygun olarak davrandığımı ve aksinin ortaya çıkması durumunda her türlü yasal sonucu kabul ettiğimi beyan ederim. 16/06/2017



Hasan AMANET

İÇİNDEKİLER

Sayfa No

ÖN SÖZ III
TEZ ETİK BEYANNAMESİ IV
İÇİNDEKİLER V
ÖZET VIII
SUMMARY IX
ŞEKİLLER DİZİNİ X
ÇİZELGELER DİZİNİ XII
SEMBOLLER DİZİNİ XIV
1. GENEL BİLGİLER 1
1.1. Giriş 1
1.2. Sosyal Medya Platformu: Twitter 2
1.3. Önceki Çalışmalar 3
1.4. Veri Madenciliği 7
1.4.1. Veri Madenciliğinin Kullanım Alanları 8
1.4.2. Veri Madenciliğinin Adımları 9
1.4.3. Veri Madenciliği Yaklaşımları 12
1.4.4. Veri Madenciliği Yöntemleri ve Kullanılan Algoritmalar 13
1.4.4.1. Sınıflandırma 13
1.4.4.2. Kümeleme Analizi 13
1.4.4.3. Birliktelik Analizi 14
1.4.5. Metin Madenciliği 15
1.4.5.1. Sosyal Medyada Metin Madenciliği 16
1.4.5.2. Sosyal Medya Metinlerinin Farklı Yönleri 16
1.4.6. Duygu Analizi 17
1.4.6.1. Duygu Analizi İle Yapılan İşlemler 21
1.4.7. Duygu Sınıflandırma Teknikleri 21
1.4.7.1. Makine Öğrenmesi Yaklaşımı 22
1.4.7.2. Denetimli Öğrenme 22
1.4.7.3. Karar Ağaçları 23
1.4.7.4. Doğrusal Sınıflandırıcılar 24

1.4.7.5. Destek Vektör Makineleri 24
1.4.7.6. Yapay Sinir Ağları 26
1.4.7.7. Olasılıksal Sınıflandırıcılar 27
1.4.7.8. Yalın Bayes Sınıflandırıcısı 27
1.4.7.9. Maksimum Entropi 30
1.4.7.10. Sözcük Tabanlı Yaklaşım 31
1.4.7.11. Sözlük Tabanlı Yaklaşım 31
1.4.7.12. Korpus Yaklaşımı 32
1.4.8. Sosyal Medyada Duygu Analizi 33
1.4.8.1. Metin Ön İşleme 33
1.4.8.2. Doküman Normalizasyonu 34
1.4.8.3. Belirteçleme 34
1.4.8.4. Soyutlama 35
1.4.8.5. Kelime Köküne Ulaşma 36
1.4.8.6. Ek Analizi 36
1.4.8.7. Durak Kelimelerinin Çıkarılması 37
1.4.8.8. Metnin Vektörel Olarak İfade Edilmesi 37
1.4.8.9. Vektör Uzay Modeli 38
1.4.8.10. N-gram Modeli 40
1.5. Türkçenin Yapısı 41
1.5.1. Türkçenin Ünlüleri / Vokalleri 41
1.5.2. Türkçenin Ünsüzler 43
1.5.3. Türkçede Hece Yapısı 44
1.5.4. Türkçede Ses Uyumları 45
1.5.4.1. Ünlü Uyumları 45
1.5.4.2. Kalınlık-İncelik Uyumu 45
1.5.4.3. Düzlük–Yuvarlaklık Uyumu 46
1.5.4.4. Düzlük Uyumu 46
1.5.4.5. Ünlü–Ünsüz Uyumu 47
1.5.4.6. Ünsüz – Ünlü Uyumları 47
1.5.4.7. Fillerde Olumluluk, Olumsuzluk, Genellik 48
1.5.4.8. Fillerde Olumluluk 48
1.5.4.9. Fiillerde Olumsuzluk 48

1.5.4.10. Olumsuzluğun Morfolojik Usulle İşaretlenmesi 48
1.5.4.11. Olumsuzluğun Semantik Usulle İşaretlenmesi 49
1.6. Modelin Başarım Ölçütleri 50
1.6.1. Doğruluk ve Hata Oranı 50
1.6.2. Kesinlik 51
1.6.3. Duyarlılık 51
1.6.4. F - Ölçeği 51
1.6.5. ROC (Receiver Operating Characteristics) Eğrisi 51
1.7. Öznitelik Seçimi 52
2. YAPILAN ÇALIŞMALAR 53
2.1. Veri Elde Etme 54
2.2. Veri Ön İşleme 55
2.3. Sayısallaştırma 56
2.4. Analiz 56
3. SONUÇLAR 57
3.1. Mutlu / Neşeli Duygusu 57
3.2. Takdir Duygusu 60
3.3. Dilek Duygusu 62
3.4. Güven Duygusu 65
3.5. Gurur Duyma Duygusu 67
3.6. Aşağılama Duygusu 70
3.7. Merak Duygusu 73
3.8. Öfke Duygusu 75
3.9. Hayal Kırıklığı Duygusu 78
3.10. Tavsiye Duygusu 81
4. ÖNERİLER 84
5. KAYNAKÇA 86
6. EKLER 95
ÖZGEÇMİŞ

Yüksek Lisans

ÖZET

TÜRKÇE SOSYAL MEDYA METİNLERİNDE DUYGU ANALİZİ

Hasan AMANET

Karadeniz Teknik Üniversitesi
Fen Bilimleri Enstitüsü
İstatistik ve Bilgisayar Bilimleri Anabilim Dalı
Danışman: Yrd. Doç. Dr. Tolga BERBER
2017, 94 Sayfa, 1 Ek Sayfa

Bu çalışmada, sosyal medyadaki yazılı Türkçe metinlerde duygu analizi yapabilmek için bir yöntem önerilmiştir. Yapılan saha çalışması ile belirlenen duygu kategorileri kullanılarak Twitter metin verileri üzerinde duygu analizi yapılmıştır. Çalışma kapsamında belirlenen "Mutlu", "Güvenmek", "Takdir Etmek", "Gurur Duyma", "Beklenti", "Tavsiye", "Merak", "Hayal Kırıklığı", "Öfke" olmak üzere 10 duygu kategorisinde sosyal medya metinleri sınıflandırılmış ve her duygu için öz nitelikler belirlenmiştir. Duygu durumları R. Plutchik'ın duygu teorisine dayanarak belirlenmiştir.

Çalışmada gerçekleştirilen duygu analizi için metin madenciliği sınıflandırma yöntemlerinden Yalın Bayes, Karar Ağaçları, K- en yakın komşu ve Destek Vektör Makineleri kullanılmıştır. Elde edilen sonuçlara göre *takdir* duygusu %65 oranında doğru sınıflandırılmıştır. Sınıflandırma işlemi tamamlandıktan sonra duyguyu ifade eden en önemli kelimelerin bulunması için ileri doğru seçim yöntemi kullanılmıştır. Elde edilen sonuçlar ışığında 10 duygu için etkin kelimeler belirlenmiştir. Belirlenen kelimeler içerisinde en başarılı olan duygu sınıfı yine *takdir* duygusudur. Bu duygu için bulunan kelimeler "teşekkürler, helal, teşekkür, tebrikler, adamsın, bravo, davranış" şeklindedir.

Bir GSM firmasına ait twitter metinleri ile yapılan duygu analizi sonucunda elde edilen öznitelikler incelenmiştir. Bu öznitelikler değerlendirildiğinde, kullanıcıların paylaştıkları metinleri hangi duygu durumuna ve hangi konuya bağlı olarak yazdığı tespit edilmiştir.

Anahtar Kelimeler: Duygu Analizi, Metin Madenciliği, Vektör Uzayı Modeli, Sınıflandırma

Master Thesis

SUMMARY

SENTIMENT ANALYSIS IN TURKISH SOCIAL MEDIA TEXTS

Hasan AMANET

Karadeniz Technical University
The Graduate School of Natural and Applied Sciences
Statistical and Computer Science Graduate Program
Supervisor: Assist. Prof. Tolga BERBER
2017, 94 Pages, 1 Pages Appendix

In this study, a method for sentiment analysis of Turkish social media texts is proposed. Sentiment analysis was performed on the Twitter text data using the emotion categories determined by the field study. Social media texts were classified into 10 emotion categories as "Happy", "Trust", "Appreciation", "Pride", "Expectation", "Recommendation", "Curiosity", "Disappointment" and the most effective words for each emotion are determined. Emotions used in this study are based on R. Plutchik's emotion theory.

The text classification methods for sentiment analysis used in the study are Naïve Bayes, Decision Trees, K-Nearest Neighbors and Support Vector Machines. According to the results, correct classification performance of Appreciation is 65%. After completing the classification process, the forward selection method was used to find the most important words expressing the emotion. The most effective words for 10 emotions were determined using the results. Appreciation is the most successful emotion class among the other emotions considering the words found in the process. The most effective words for appreciation are “teşekkürler, helal, teşekkür, tebrikler, adamsın, bravo, davranış”.

In this study, Twitter messages posted to formal account of a GSM company is analyzed. According to the analysis results, it has been shown that emotional states and subjects of the tweets could be determined by the words of user tweets.

Key Words: Sentiment Analysis, Text Mining, Vector Space Model, Classification

ŞEKİLLER DİZİNİ

Sayfa No

Şekil 1. Örnek twitter mesajı 2
Şekil 2. Twitter kullanıcı profili 3
Şekil 3. Veri madenciliğinin ilgilendiği disiplinler 7
Şekil 4. 2016 yılına ait veri madenciliği kullanım oranları [29] 8
Şekil 5. Tipik veri madenciliği adımları 10
Şekil 6. Metin analizi ile yapılan işlemler ve ilgili disiplinler 15
Şekil 7. Dr. Robert Plutchik'ın duygu şeması [61] 20
Şekil 8. Duygu analizi ile yapılan işlemler 21
Şekil 9. Duygu sınıflandırma teknikleri 22
Şekil 10. Hasta veri tabanı için bir karar ağacı ve kurallar 23
Şekil 11. Hiper-düzlemler 24
Şekil 12. İki sınıflı bir problemin hiper-düzlemleri [70] 26
Şekil 13. Optimum hiper-düzlem ve destek vektörleri [70] 26
Şekil 14. Geleneksel metin madenciliği 33
Şekil 15. Vektör uzay modeli 39
Şekil 16. Mutluluk duygusu için kullanılan sınıflandırma yöntemleri ve başarımları 58
Şekil 17. Mutluluk duygusu kelime bulutu 59
Şekil 18. Takdir duygusu için kullanılan sınıflandırma yöntemleri ve başarımları 61
Şekil 19. Takdir duygusu kelime bulutu 62
Şekil 20. Dilek duygusu için kullanılan sınıflandırma yöntemleri ve başarımları 63
Şekil 21. Dilek duygusu kelime bulutu 65
Şekil 22. Güven duygusu için kullanılan sınıflandırma yöntemleri ve başarımları 66
Şekil 23. Güven duygusu kelime bulutu 67
Şekil 24. Gurur duyma duygusu için kullanılan sınıflandırma yöntemleri ve başarımları 68
Şekil 25. Gurur duyma duygusu kelime bulutu 70
Şekil 26. Aşağılama duygusu için kullanılan sınıflandırma algoritmaları ve başarımı 71
Şekil 27. Aşağılama duygusu için kelime bulutu 72
Şekil 28. Merak duygusu için kullanılan sınıflandırma yöntemleri ve başarımları 74
Şekil 29. Merak duygusu kelime bulutu 75

Şekil 30. Öfke duygusu için kullanılan sınıflandırma yöntemleri ve başarımları 76
Şekil 31. Öfke duygusu kelime bulutu 77
Şekil 32. Hayal kırıklığı duygusu için kullanılan sınıflandırma yöntemleri ve başarımları..... 79
Şekil 33. Hayal kırıklığı duygusu kelime bulutu..... 80
Şekil 34. Tavsiye duygusu için kullanılan sınıflandırma yöntemleri ve başarımları 81
Şekil 35. Tavsiye duygusu kelime bulutu..... 83

ÇİZELGELER DİZİNİ

Sayfa No

Tablo 1. Farklı teorisyenlerden temel duygular [58] 18
Tablo 2. İMDEA duygu kutupları ve duygular [59] 19
Tablo 3. Metni ifade eden vektör 38
Tablo 4. Örnek metindeki n-gramlar 40
Tablo 5. Türkiye Türkçesindeki ünlüler ve özellikleri [109] 41
Tablo 6. Standart Türkiye Türkçesinde kullanılan ünsüzler [109] 43
Tablo 7. Türkçedeki hece tipleri [109] 44
Tablo 8. Düzlük-yuvarlaklık uyumuna göre bir sözcüğün birinci ve diğer hecelerinde bulunabilecek ünlüler [109] 46
Tablo 9. Düzlük uyumuna göre bir sözcüğün birinci ve diğer hecelerinde bulunabilecek ünlüler [109] 47
Tablo 10. Karışıklık matrisi 50
Tablo 11. Duygu analizinde kullanılan olumlu duygu sınıfları 54
Tablo 12. Duygu analizinde kullanılan olumsuz duygu sınıfları 54
Tablo 13. Mutluluk duygusu için sınıflandırma sonuçları 57
Tablo 14. Mutluluk duygusu için ileri doğru seçim yöntemi ile elde edilen kelimeler ve başarımları 59
Tablo 15. Takdir duygusu sınıflandırma sonuçları 60
Tablo 16. Takdir duygusu için ileri doğru seçim yöntemi ile elde edilen kelimeler ve başarımı 61
Tablo 17. Dilek duygusu sınıflandırma sonuçları 63
Tablo 18. Dilek duygusu için ileri doğru seçim yöntemi ile elde edilen kelimeler ve başarımları 64
Tablo 19. Güvenmek duygusu sınıflandırma sonuçları 65
Tablo 20. Güven duygusu için ileri doğru seçim yöntemi ile elde edilen kelimeler ve başarıları 66
Tablo 21. Gurur duyma duygusu sınıflandırma sonuçları 68
Tablo 22. Gurur duygusu için ileri doğru seçim yöntemi ile elde edilen kelimeler ve başarımları 69
Tablo 23. Aşağılama duygusu sınıflandırma sonuçları 70

Tablo 24. Aşağılama duygusu için ileri doğru seçim yöntemi ile elde edilen kelimeler ve başarımı 71
Tablo 25. Merak duygusu sınıflandırma sonuçları 73
Tablo 26. Merak duygusu için ileri doğru seçim yöntemi ile elde edilen kelimeler ve başarımı 74
Tablo 27. Öfke duygusu sınıflandırma sonuçları 75
Tablo 28. Öfke duygusu için ileri doğru seçim yöntemi ile elde edilen kelimeler ve başarımı 76
Tablo 29. Hayal kırıklığı duygusu sınıflandırma sonuçları 78
Tablo 30. Hayal kırıklığı duygusu için ileri doğru seçim yöntemi ile elde edilen kelimeler ve başarımı 79
Tablo 31. Tavsiye duygusu sınıflandırma sonuçları 81
Tablo 32. Tavsiye duygusu için ileri doğru seçim yöntemi ile elde edilen kelimeler ve başarımı 82

SEMBOLLER DİZİNİ

API	: Uygulama Ara Yüzü
DA	: Duygu Analizi
DDİ	: Doğal Dil İşleme
DVM	: Destek Vektör Makineleri
DATD	: Duygu Açıklaması ve Temsil Dili
EAA	: Eğri Altındaki Alan
HKT	: Hata Kareler Toplamı
HUMAIN	: İnsan – Makine Duygu Etkileşim Ağı
KNB	: Karşılıklı Noktasal Bilgi
ROC	: Receiver Operating Characteristic
POS	: Part of Speech
SÇM	: Sözde Çanta Modeli
VUM	: Vektör Uzayı Modeli
YBS	: Yalın Bayes Sınıflandırıcısı
YSA	: Yapay Sinir Ağları
WordNet	: Kelimelerin Kavramsal İlişkisi Veri Tabanı

1. GENEL BİLGİLER

1.1. Giriş

İletişim; temel prensibi paylaşım, etkileşim ve ortaklık kurma olan, çeşitli semboller ve araçlarla dünyayı daha yaşanır kılan, ileti alışverişine dayanan bir süreçtir. İletişim kurmak için insanoğlu yaratılışından bugüne kadar posta güvercini, mektup, telgraf, internet gibi çeşitli araçları kullanmıştır. 2016 yılı itibariyle kullanılan bu iletişim araçlarından en önemlisi internettir. İnternet, birçok bilgisayar sisteminin birbirine bağlı olduğu, dünya çapında yaygın olan ve sürekli büyüyen bir iletişim ağıdır. Bu teknoloji yardımıyla pek çok alandaki bilgilere insanlar kolay, ucuz, hızlı ve güvenli bir şekilde ulaşabilir [1].

Günümüzde internet, insanların görüşlerini ifade edebildiği küresel bir yapıya dönüşmüştür. Bu sayede insanlar Facebook ve Twitter gibi sosyal medya ortamlarında herhangi bir konu ile ilgili duygu, düşünce ve fikirlerini ifade edebilmektedir [2]. Her geçen gün kullanımı artan bu ortamlar reklam ve kampanyaları yürütme, haber paylaşımı gibi amaçlarla da kullanılmaktadır. Bu açıdan sosyal medya ekonomi, siyaset, ticaret ve duygu analizi gibi alanlarda sıklıkla kullanılmaktadır.

Sosyal medya platformlarını aktif olarak kullananların sayısı sürekli artmakta ve bu platformlar geniş bir iletişim ağı olarak büyümeye devam etmektedir. Dijital pazarlama alanında dönüşüm ajansı olarak faaliyet gösteren We Are Social'ın 2016 yılına ilişkin verileri incelediğinde ülkemizde aktif sosyal medya kullanıcı sayısının 42 milyon olduğu görülmektedir. Sosyal medya platformlarının kullanım oranları incelendiğinde, %32 ile Facebook'un birinci sırada yer aldığı, %17 ile Twitter'ın ikinci, %16 ile Instagram'ın üçüncü ve %16 ile Google+'ın dördüncü sırada yer aldığı görülmektedir. Facebook ve Twitter'ın dünya çapında kullanıcı sayılarına baktığımızda; Facebook, 1.59 milyar aktif kullanıcıya sahipken, Twitter 320 milyon aktif kullanıcıya sahiptir [3]. Bu rakamlardan sosyal medya platformlarında büyük kitlelere ulaşmak, onlarla iletişime geçmek, reklam, kampanya ve kamuoyu araştırmaları gibi faaliyetlerin hızlı ve etkin bir şekilde yapılabileceği anlaşılır.

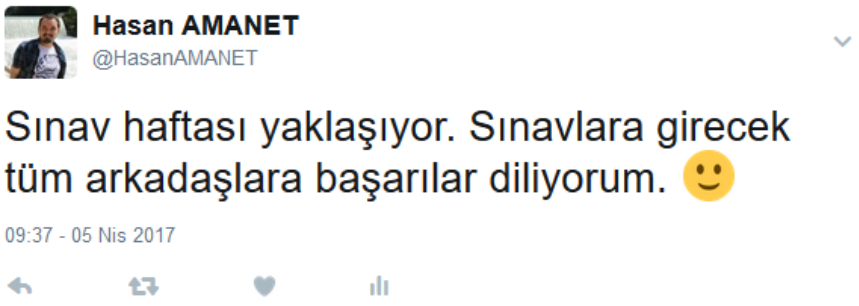
Bu çalışmada kullanılan "Tweet" kavramının Türkçe karşılığı olmadığı, bu nedenle günlük yazılı / sözlü kaynaklarda ortak kabul göreceği için çalışmamızda bu kavrama karşılık öneri sunulmamıştır.

1.2. Sosyal Medya Platformu: Twitter

Twitter, 2006 yılında Jack Dorsey tarafından kurulan, kullanıcılarına 140 karakterle sınırlı metin, video ve fotoğraf paylaşabilme imkânı sağlayan mikro blogdur. Kullanıcılar Twitter'ı akıllı telefon, tablet ve bilgisayar aracılığıyla internet bağlantısının olduğu herhangi bir yerden takip ederek, paylaşımda bulunabilir ve insanlarla iletişime geçebilirler.

Twitter'da kullanılan bazı terimleri açıklamak konunun daha iyi anlaşılmasına yardımcı olacaktır;

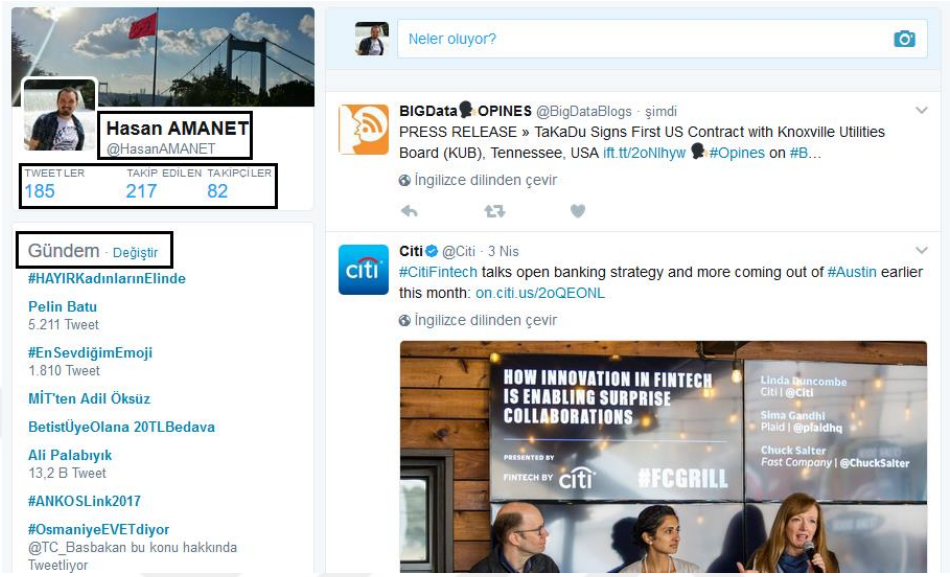
- Tweet, Twitter'da yazılan her mesaja verilen addır. Metin, video ve fotoğraf ayırt etmeksizin paylaşılan her mesaja tweet denir.
- Retweet, başka bir kullanıcı tarafından gönderilen bir tweet'in, kendi hesabınızdaki kişilere aynen iletilmesidir.
- Yanıtlama, Tweet metninde başka bir kullanıcının @kullanıcıadı içeren bir mesajdır. Bir mesajda "@" işareti kullanıldığında o mesaj yalnızca o kişi veya kişilere gönderilmektedir.
- Direkt mesaj, Twitter üzerinden gizli mesaj gönderme işlemine verilen isimdir.
- Etiket, Twitter'da başına "#" sembolü konularak yazılan sözcüklerle ifade edilir. İnsanlar, Tweetleri kategorilere ayırmak ve Twitter aramasında tweetleri daha kolay gösterilmesine yardımcı olmak için mesajlarındaki uygun anahtar sözcüğün veya cümlenin başına etiket sembolü olan "#" koymaktadırlar.
- Trending topic, en çok hangi konuların konuşulduğunu gösteren, popüler konu başlıkları listesidir. Şekil 1'de örnek bir tweet metni gösterilmektedir.



Şekil 1. Örnek twitter mesajı

Duygu analizi konusunda araştırmacılar tarafından en yaygın kullanıma sahip sosyal medya platformu Twitter'dır. Twitter, sistemindeki verileri UAY (Uygulama Ara Yüzleri)

yardımıyla araştırmacılarla paylaşmaktadır. Bir Twitter metniyle ilgili birçok veriye UAY'leri sayesinde ulaşmak mümkündür.



Şekil 2. Twitter kullanıcı profili

UAY'ler sayesinde Şekil 2.'deki kullanıcı profilinde görülen kullanıcı adı, paylaşılan tweet sayısı, takip eden sayısı, takipçi sayısı gibi kullanıcıya ait bilgilere ulaşmak mümkündür. Bunların yanı sıra UAY'ler, paylaşılan tweet ID (Identification), paylaşım zamanı ve paylaşım konumu gibi bilgileri de elde etmeye imkân sağlamaktadır.

Sosyal medya kullanımının yaygınlaşmasıyla beraber kişiler aynı anda birden fazla kişiyle iletişime geçme imkânı bulmuşlar; bu da ortaya başka ihtiyaçların çıkmasına sebep olmuştur. Milyonlarca kişiden oluşan bu iletişim ağında saniyede binlerce mesaj paylaşılmaktadır. Büyük miktarda verinin olduğu bu ortamda, paylaşılan mesajlardaki duygu ve düşünceyi analiz etmek önem kazanmıştır.

1.3. Önceki Çalışmalar

Duygu Analizi (DA), fikir madenciliği olarak bilinir; fikirler, değerlendirmeler, tutum, düşünce tahmini, ürünler ve davranışlar gibi konu ve niteliklere yönelik duyguları analiz etmeye çalışılan bir alandır. Duygu analizi; görüş belirleme, etki analizi, görüş madenciliği gibi farklı isim ve görevlere sahiptir. Tüm bu farklı isim ve görevler duygu analizi ya da görüş incelemesi olarak kullanılmaktadır. Akademik alanda duygu analizi terimi daha

yaygın bir kullanıma sahiptir. DA çoğunlukla olumlu veya olumsuz düşünceleri ifade eden veya ima eden görüşlere odaklanmaktadır.

DA terimi ilk kez Nasukawa ve Yi'nin [4] çalışmasında ortaya çıkmış ve görüş analizi terimi ise ilk kez Dave ve arkadaşları [5] tarafından yapılan çalışmada kullanılmıştır. Ancak duygular üzerine daha önce yapılan araştırmalar bulunmaktadır [5–8]. Doğal dil işleme (DDİ) ve dilbilimi, DA'nın önemli bir kısmını oluşturmaktadır. DDİ ve dilbilimi uzun bir geçmişe sahip olmasına rağmen 2000 yılından önce insanların görüş ve düşünceleri hakkında az sayıda çalışma yapılmıştır. Bu dönemden sonra gelişen internet kullanımı ve sosyal medya platformlarının çoğalmasına paralel olarak DA alanında yapılan çalışmalar da çoğalmıştır. Son yıllarda, sosyal medyadaki görüş ve fikirlerin işletmeleri şekillendirdiği, toplumsal ve siyasi sistemlere etki ettiği görülmektedir. Bu durum insanların duygu durumunun sosyal medya üzerinden kolay bir şekilde değiştirilebildiğini göstermiştir. Örneğin, 2011 yılında Kuzey Afrika ve Körfez bölgesindeki ülkelerde kitlelerin sosyal medya üzerinden harekete geçerek, yaşanan siyasi değişimleri başlattıkları bilinmektedir. Bu durum sosyal medya üzerinde duygu ve fikir toplamanın önemli bir araştırma alanı olduğunu göstermiştir. DA sadece web verileri için değil aynı zamanda kurumlarda bulunan dahili veriler içinde de kullanılır. Örneğin, e-postalar, müşteri geri bildirimleri veya kuruluşlar tarafından yapılan anket sonuçları DA için kullanılan verilerdir.

DA uygulamaları, müşteri-ürün memnuniyeti, sağlık hizmetleri, finansal hizmetler, sosyal etkinlikler, siyasi seçimler gibi birçok alanda yaygın bir şekilde kullanılmaktadır. Microsoft, Google, Hewlett-Packard gibi dünyaca ünlü firmalar kendi bünyelerinde DA için birimler oluşturmuş ve çeşitli yazılımlar geliştirmişlerdir.

DA ile ilgili olarak literatürde çok sayıda çalışma bulunmaktadır. Liu ve arkadaşları [9], satış performansını tahmin etmek için bir DA modeli önermiştir. Hong ve Skiena [10], Amerikan Ulusal Futbol Ligi (NFL)'deki bahis oranları ile sosyal medya blogları ve Twitter mesajları arasındaki ilişkiyi incelemiştir. O'Connor ve arkadaşları [11], 2008-2009 yılları arasında tüketiciler ve siyasi seçimler üzerine yapılan anket sonuçlarıyla, aynı döneme ait Twitter mesajlarındaki duygu sözcük frekanslarının birbiriyle ilişkili olduğunu tespit etmişlerdir. Bu iki bilgi arasındaki ilişkinin %80 gibi yüksek bir oran olduğu tespit edilmiştir. Tumasjan ve arkadaşları [12], seçim sonuçlarını tahmin etmek için Twitter'da DA uygulaması gerçekleştirmiştir. Bu çalışmada amaç Twitter'ın siyasi tartışmalar için bir forum olarak kullanılıp kullanılmadığını tespit etmek ve Twitter'daki siyasi konuşmaların, çevrimdışı siyasi düşüncelere olan etkisini araştırmaktır. Veri olarak Almanya Federal

seçimleri anket verileri kullanılmıştır. Bu çalışma sonucunda Twitter'ın siyasi tartışmalar için yoğun şekilde kullanıldığı belirlenmiş ve bir partiden bahseden mesajların seçim sonuçlarını yansıttığı tespit edilmiştir. Bu sayede araştırmacılar, Twitter'daki siyasi düşüncelerin, çevrimdışı siyasi durumu etkilediği göstermişlerdir. Filmlerin gişe hasılatlarını tahmin etmek için Twitter verileri, film yorumları ve bloglarındaki metinleri kullanan araştırmalar bulunmaktadır [13–15]. Bu çalışmalarda filmler için yapılan eleştiriler ve yorumlar kullanılarak analizler yapılmıştır. Korelasyon, kümeleme ve zaman serisi gibi analizler kullanılmıştır. Mohammad ve Yang [16], e-posta metinlerini kullanılarak duygu durumlarının cinsiyete göre farlılıklarını incelemiştir. Bu çalışmada duygu kelimeleri sözlüğü oluşturulmuş sonrasında sevgiye ve nefrete duyarlı olan kelimeler tespit edilmiştir. Çalışma sonunda iş yerindeki e-postalarda duygu kelimelerinin nasıl kullanıldığı ve cinsiyetler arasında belirgin farkların olduğu tespit edilmiştir. Örneğin, kadınlar yazdıkları e-postalarda neşe - hüzün ekseninde kelimeler kullanırken, erkeklerin korku - güven eksenindeki terimleri tercih ettikleri belirlenmiştir. Mohammad [17], romanlardaki ve masallardaki duygu durumlarını araştırmıştır. Metin koleksiyonu olarak Romanları ve Brothers Grimm masallarını örnek alarak onlar üzerinden işlemler yapmıştır. Bu çalışma ile metinler içinde arama yaparken, duygu ifadeleri birbiriyle bağlantılı olan kelimelerin nasıl belirleneceği gösterilmiştir. Bu sayede metin koleksiyonlarının nasıl daha iyi kategorize edileceği ve aramalarda nasıl daha hızlı sonuçlar vereceği ifade edilmektedir. Masallar ve romanlar arasındaki duygu yoğunluğu karşılaştırılmış ve masalların romanlara oranla daha geniş bir duygu yoğunluğuna sahip olduğu tespit edilmiştir. Paltoglu [18], Twitter'da duygu durumuna dayalı olarak yaşanan olayları tespit etmeye çalışmıştır. Bu çalışma, duygu analizinin dünyada önemli olayların tespiti için kullanılacağını göstermiştir. Ayrıca bu çalışma ile sosyal medyada yönlendirilen olayların tespit edildiği gösterilmiştir. Olumsuz mesajların, olumlu mesajlara göre daha hızlı yayıldığı da bu çalışmada tespit edilmiştir.

Kullanılan teknik açısından yapılan çalışmalar incelendiğinde, DA'da iki tür tekniğin yaygın olarak kullanıldığı görülmektedir. Bu yaklaşımlardan birincisi makine öğrenmesine dayalı olan tekniktir. Bu teknikte, duygu ifade eden kelimeler belirlendikten sonra onlara ait öznitelikler tespit edilir; işaretlenmiş veri seti ve test verileri kullanılır. Bir sonraki adım da makine öğrenmesi algoritmalarıyla duygu sınıflandırması gerçekleştirilir. Geleneksel sınıflandırma algoritmaları Yalın Bayes (YB, Naive Bayes), Destek Vektör Makinesi (DVM, Support Vector Machine) ve Maksimum Entropi (ME)'dir. Lin ve arkadaşları [19], bahsedilen teknikleri kullanarak bir yazının hangi konuda yazıldığını %81 doğruluk ile tespit

etmiştir. Bu çalışmada veri seti olarak Reuters haber ajansına ait veritabanından, İsrail ve Filistin arasında yaşanan olayları içeren veriler kullanılmıştır. Riloff ve arkadaşları [20], yapıkları çalışmada öznel cümleler ile nesnel cümleleri sınıflandıran bir model geliştirmişlerdir. Bu modeli kullanarak Yalın Bayes sınıflandırma algoritmasıyla elde ettikleri başarı %81 düzeyindedir. Pang ve arkadaşları [8], filmlere yapılan yorumları analiz ederek yorumları olumlu ve olumsuz olarak sınıflandırmışlardır. Çalışmada, YB, DVM ve ME yöntemleri kullanılmış başarım sırasıyla, %81, %82 ve %80 olarak elde edilmiştir. DA'da kullanılan diğer bir teknik ise sözlük tabanlı yaklaşımdır. Nasukawa ve Yi [4], İngilizce dili için yaptıkları çalışmalarında, duygu sözlüğü kullanarak web sayfaları ve haber makaleleri üzerinde olumlu ve olumsuz duygu durumunu incelemişlerdir. Kelimelerin tür bilgilerinin sisteme dahil edilmesinden sonra yapılan analizde başarımın arttığı görülmüştür. Kullandıkları yönteme bağlı olarak başarım %91 olarak elde edilmiştir. Sözlük tabanlı yaklaşım, basit, kolay anlaşılabilir ve elde edilen sonuç olarak yüksek doğruluk elde edilen bir yöntemdir. Ancak, duyguları ifade eden yeni bir sözcüğün tanınması açısından performansı iyi değildir.

Literatürde Türkçe DA ile ilgili yapılan çalışmaların sayısı azdır. Kaşıkçı ve Gökçen [21], tarafından yapılan çalışmada e-ticaret sitelerinin otomatik olarak belirlenmesini sağlayacak bir model geliştirilmiş ve başarım k-NN Yakın Komşu için %83.8, Yalın Bayes için %85,3 olarak elde edilmiştir.

Güran ve arkadaşları [22], Türkçe için yaptıkları metin sınıflandırma çalışmasında, N-gram modelini kullanarak 600 doküman ve 6 kategoride %95,8 oranında bir başarı elde etmişlerdir. Doğan ve Diri [23], Türkçe metinlerin türü, yazarı ve yazarın cinsiyetini N-gram modeli kullanılarak belirlemeye çalışmışlardır. N-gram modelinde 2, 3 ve 4 gramları kullanmışlardır. Üç farklı veri seti ve altı kategori ile yaptıkları çalışma sonucunda Ng-ind yöntemini geliştirmişler ve bu yöntemle en yüksek başarım olarak %93,8 değerini elde etmişlerdir. Şimşek ve Özdemir [24], yaptıkları çalışmada borsadaki verilerin değişimi ile Twitter'da yazılan mesajlar arasında bir ilişkinin olup olmadığını araştırmıştır. Sözlüğe dayalı yaklaşım ile yapılan analizde sekiz farklı duyguya ait, 113 öznitelik belirlenmiş ve bu öznitelikler kullanılarak mesajlar mutlu veya mutsuz olarak sınıflandırılmıştır. Çalışma sonucunda borsadaki değişim ile insanların duygu durumları arasında %45 oranında bir ilişki olduğu saptanmıştır.

1.4. Veri Madenciliği

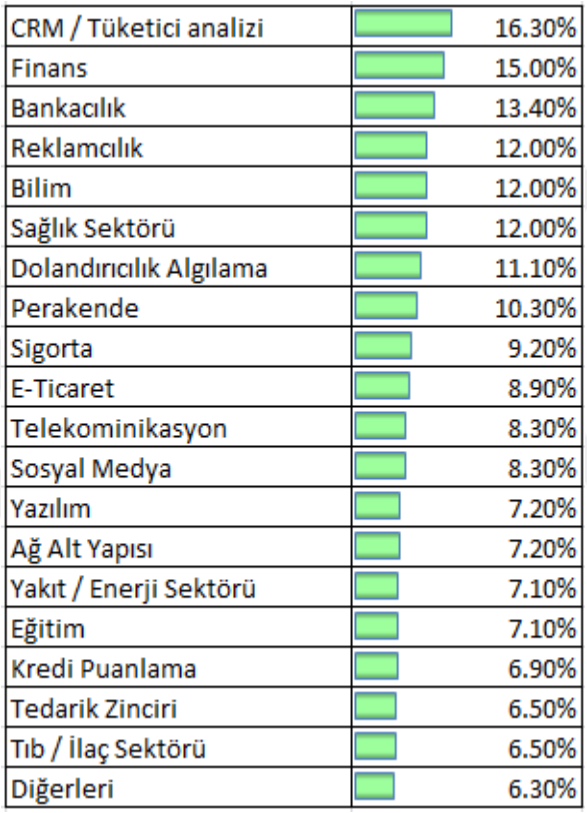
Veri madenciliği, bilgisayarlarda depolanan büyük miktardaki verilerden bilgiye ulaşmak olarak ifade edilmektedir. Veri madenciliği, veri keşfi analizi olarak da adlandırılmaktadır. Veri madenciliği, iş dünyası için fayda sağlamakta ve her geçen gün hızlı bir şekilde kullanımı yaygınlaşmaktadır [25]. Veri madenciliği, donanım ve yazılım teknolojisindeki büyük gelişmelere bağlı olarak son yıllarda hızlı ilerlemeler gösteren ve farklı veri türlerine erişilebilen bir alanı oluşturmaktadır [26]. Veri madenciliği, verileri araştıran, matematiksel modeller geliştiren ve yararlı bilginin özü olan önemli kalıpları (örtülü veya açık) keşfeden, çıkarım algoritmalarını içeren veri tabanından bilgi bulma sürecinin matematiksel özüdür [27]. Veri madenciliği, istatistik, makine öğrenimi, yapay zekâ ve veri tabanı teknolojisini birleştiren çok disiplinli bir alandır [28].

Şekil 3. Veri madenciliğinin ilgilendiği disiplinler

Veri madenciliği pazarlama organizasyonları tarafından müşteri sınıflandırılması ve bankacılık sektöründe, müşterilerin kredi tekliflerine cevap verme olasılığını tahmin etmek amacıyla kullanılmaktadır. Bankalardaki verilerin büyük miktarda ve güvenilir olmasından dolayı veri madenciliği teknikleri tercih edilmektedir. Veri madenciliği, verinin bilgisayar ortamına aktırıldığı tüm sektörlerde uygulanabilir.

1.4.1. Veri Madenciliğinin Kullanım Alanları

Veri madenciliği, verinin bilgisayar ortamına aktarıldığı her alanda etkin bir şekilde kullanılmaktadır. İş analistliği, büyük veri, veri madenciliği ve veri bilimi alanlarında lider olan [29] yayınlanan rapora göre 2016 yılına ait veri madenciliği kullanım alanları ve oranları Şekil 4.'te gösterilmiştir.



Şekil 4. 2016 yılına ait veri madenciliği kullanım oranları [29]

Veri madenciliğini kullanım alanları maddeler hâlinde şöyle sunulabilir:

Finans Sektöründe Veri Madenciliği: Bankacılık ve finans sektöründeki veriler genellikle güvenilirdir ve sistematik veri analizi ve veri madenciliğini kolaylaştıran yüksek kalitededir. Bankacılık ve finans sektöründe veri madenciliği çok boyutlu veri analizi ve veri madenciliği için veri ambarı tasarımı ve inşaatı, kredi ödeme tahmini ve müşteri kredi politikası analizi, hedefli pazarlama için müşterilerin sınıflandırılması ve kümelenmesi, kara para aklama ve diğer mali suçların tespit edilmesi amacıyla kullanılmaktadır [30].

Perakende Sektöründe Veri Madenciliği: Veri madenciliği satış, müşteri satın alma geçmişi, tüketim ve hizmetler olmak üzere çok miktarda veri toplayan perakende sektöründe geniş bir uygulama alanına sahiptir. Perakende sektöründe veri madenciliği yaygın olarak satışların, müşterilerin, zamanın ve bölgenin çok boyutlu analizinde, satış kampanyalarının etkinlik analizinde, müşteri tutmada, ürün tavsiyesi ve ürünlerin çapraz referanslanması amacıyla kullanılmaktadır.

Müşteri İlişkileri Yönetimi ve Tüketici Analizinde Veri Madenciliği: Müşteri İlişkileri Yönetimi, müşteriler kazanma ve bunları koruma, aynı zamanda müşterilerin sadakatini geliştirme ve müşteri odaklı stratejileri uygulama ile ilgilidir. Müşteriyle uygun bir ilişki sürdürmek için veri toplamak ve bilgiyi analiz etmek gerekir. İşte bu noktada veri madenciliği rol oynamaktadır. Veri madenciliği teknolojileri ile toplanan veriler analiz için kullanılmaktadır [31].

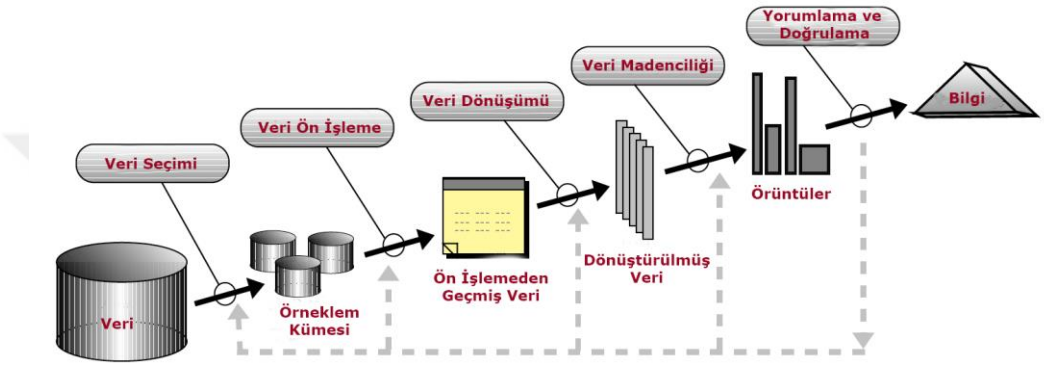
Sağlık Sektöründe Veri Madenciliği: Hastaneler, sağlık kuruluşları, sigorta şirketleri ve hükûmetler, hastalar, sağlık sorunları, kullanılan klinik prosedürleri, maliyetleri ve sonuçları hakkında çok fazla veriye sahiptirler. Bu veriyle ilişkileri anlamak, hangi prosedürlerin ve klinik müdahalelerinin en etkili olduğunun belirlenmesinden, azalan kaynakların bulunduğu dönemde insanlara sağlık hizmetinin en iyi şekilde nasıl ulaştırılacağına kadar farklı ve önemli bilgileri sunmaktadır [32].

Eğitimsel Veri Madenciliği: Eğitimsel veri madenciliği adı verilen ve eğitim ortamlarından kaynaklanan veriden bilgi keşfeden, yöntemler geliştirmekle ilgili yeni ortaya çıkan bir alan vardır. Eğitimsel veri madenciliğinin hedefleri, öğrencilerin gelecekteki öğrenme davranışlarını öngörme, eğitim desteğinin etkilerini inceleme ve öğrenmeyle ilgili bilimsel bilgiyi ilerletme olarak tanımlanmaktadır. Veri madenciliği bir kurum tarafından doğru kararlar almak ve öğrencinin başarı durumlarını tahmin etmekte kullanılabilir. Sonuçlar ile kurum ne öğreteceğine ve nasıl öğretileceğine odaklanabilir. Öğrencilerin öğrenme kalıpları yakalanabilir ve onlara öğretilecek teknikler geliştirmek için kullanılabilir [31].

1.4.2. Veri Madenciliğinin Adımları

Veri madenciliği süreci üç temel adımdan oluşur. İlk işlem adımı sıklıkla verilerin temizlenmesi olarak adlandırılır. Veri madenciliğinin en çok zaman harcanan kısmı veri madenciliği için verilerin hazırlanmasıdır. İkincisi veri madenciliği algoritması ile

hazırlanan veriyi işlemek, sıkıştırmak ve herhangi bir gizli değerli bilgiler kümesini tanımlamayı kolaylaştıracak şekilde dönüştürmek için kullanılır. Veri madenciliğinin ikinci adımında, veriler toplanıp ön işlemeden geçirildikten sonra, veri madenciliği algoritmaları, gerçek elde işlemini gerçekleştirir. Üçüncü adımda, veri madenciliği çıktıları ile gizli bilginin keşfedilip edilmediğini görmek ve veri madenciliği algoritmaları tarafından üretilen olguların önemini belirlemek üzere değerlendirildiği veri analizi aşamasıdır. Veri madenciliği adımları Şekil 5.'te ayrıntılı şekilde görülmektedir.



Şekil 5. Tipik veri madenciliği adımları

Uygulama Alanının Geliştirilmesi ve Anlaşılması: Bu aşama veri madenciliğine başlamadan önce dönüştürme ve algoritmalar ile yapılması gerekenleri anlamak için olayları belirleyen hazırlık aşamasıdır. Veri madenciliği ile bilgi keşfine başlamadan önce son kullanıcıların amaçlarını, bilgi keşfi sürecinin nerede gerçekleşeceği, ön bilgilerin anlaşılması ve tanımlanması gerekmektedir [33].

Veri Seçimi: Veri madenciliği işleminin ilk aşaması, belirli bir görevi doğru bir şekilde tanımlamak için mevcut birçok veri tabanından ilgili verileri seçmektir [25]. Bu aşamada sistemdeki bilgiler iyi analiz edilmeli ve son kullanıcının amacına uygun şekilde ilişkilendirilmelidir [34]. Bu aşama veri keşfi sonucunda elde edilecek bilgiyi önemli ölçüde etkileyeceği için son kullanıcın amacına uygun veri seçilmesine dikkat edilmelidir.

Veri Ön İşleme: Veri madenciliğinin en çok zaman alan kısmı verinin ön işleme sürecidir. Kullanılacak veri kaynakları belirlendikten sonra işlenmemiş bu verilerin temizlenmesi, istenilen formda oluşturulmaları ve biçimlendirilmeleri gerekir. Bu aşama kayıp değerlerin işlenmesi ve aykırı değerlerin kaldırılması gibi veri temizleme işlemlerini

içerir [33]. Veri ön işleme aşaması, veri madenciliği sonucunda elde edilecek bilgiyi doğrudan etkilemektedir.

Ön işleme aşamasının başarılı olması doğru ve kesin sonuçlara ulaşmaya imkân sağlamaktadır.

Veri Dönüşümü: Veri indirgeme adımı olarak da adlandırılan bu adımda veri madenciliği için daha iyi veriler üretilir ve hazırlanır. Verilerin direkt olarak veri madenciliği çalışmalarına katılması yanlış sonuçlar elde edilmesine neden olur. Sayılar üzerinde işlem yapılırken çok büyük sayılar sonucu daha çok etkileyecek, küçük sayıların sonucu etkilemesi çok daha az olacaktır. Bu işlemler için verilerin normalleştirilmesi gerekir. Çeşitli teknikler uygulanarak veriler normalleştirilir. Örneğin; bir tablodaki bir alanın 5-10, diğer alanın 1-10000 arası değer almış olsun. Bu durumda verilerin sonuç üzerinde etkisi farklı olacaktır. Bu verilere min-max normalleştirme uygulanması, bütün alanların sonucu aynı oranda etkilemesi, aynı alandaki verilerin 0-1 arasındaki karşılıklarına dönüştürülmesi şeklinde gerçekleştirilir [35].

Veri Madenciliği: Veri bu aşamada kullanılabilir hâldedir. Çalışmanın sonuncunda ulaşılmak istenen bilgiye göre veri madenciliği teknikleri seçilmelidir. Örneğin, regresyon veya kümeleme seçilebilir. Buradaki seçim buraya gelene kadar izlenen adımlara bağlıdır.

Yorumlama ve Değerlendirme: Bu aşama veri madenciliğinin veri üzerinde uygulandığı, elde edilen sonuçların yorumlandığı ve doğrulandığı son adımdır. Elde edilen sonucun çalışmanın amacına uygunluğu ve doğruluğu araştırılır. Daha önce yapılmış olan çalışmalar mevcut ise o çalışmalarla karşılaştırmalar yapılarak çalışmanın doğruluğu ispat edilir.

Veri Madenciliği Kullanım Amaçları: Veri madenciliğinin kullanım amaçları sırasıyla özetleme, kümeleme, sınıflandırma ve bağımlılık modelleme olarak adlandırılır. Kullanım amaçları maddeler hâlinde şöyle sunulabilir:

Özetleme: Verilerin soyutlanması ve genellemesidir. Amacımızla ilgili veri özetlenerek genel bir yapı olarak sunulabilir. Bu elde edilen özet bilgi tüm verinin genel görünümünü veren daha küçük bir kümesini oluşturur. Belli bir veri kümesi için özlü ve uygun bir açıklama üretmeyi amaçlar [32].

Kümeleme: Kümeleme, veri madenciliğinde, bir veri kümesinde yeni ve gerçekleştirilebilir alt grupları keşfetmenin bir yöntemi olarak her yerde kullanılır [36]. Kümeleme problemi, birtakım özellikleri paylaşan benzer nesneleri veri kümelerinde bulmayı amaçlayan denetimsiz bir öğrenme problemidir.

Sınıflandırma*:* Sınıflandırma veri madenciliğinin en çok kullanıldığı alandır. Var olan bir veri tabanının bir kısmı eğitim olarak kullanılarak sınıflandırma kuralları oluşturulur. Bu kurallar yardımıyla yeni bir durum ortaya çıktığında nasıl karar verileceği belirlenir [37]. Bu yöntemde, eğitim kümesi olarak adlandırılan ve önceden etiket değeri bilinen bir veri kümesine ihtiyaç duyulur [38].

Bağımlılık Modelleme: Bir bağımlılık modelleme problemi, nitelikler arasındaki önemli bağımlılıkları tanımlayan bir model keşfetmeyi içerir. Bu bağımlılıklar genellikle öncesi doğruysa, sonrası da doğrudur şeklinde ifade edilir [25].

1.4.3. Veri Madenciliği Yaklaşımları

Veri madenciliği; istatistiği, makine öğrenimi, veri tabanı sistemleri, sinir ağları gibi birçok araştırma alanından gelen teknikleri benimsemiştir. Veri madenciliği genel olarak aşağıdaki yaklaşımlarla kullanılmaktadır:

İstatistiksel yaklaşımlar: Veri madenciliği için Bayes ağı, regresyon analizi, küme analizi ve korelasyon analizi de dahil olmak üzere birçok istatistiksel araç kullanılmıştır. Genellikle istatistiksel modeller bir dizi eğitim verilerinden oluşturulmuştur [32].

Makine öğrenmesi yaklaşımı: Makine öğrenmesi, verilen bir problemi ortamdan edindiği bilgiye göre modelleyen yapay zekâ disiplinin bir alt dalıdır. Makine öğrenmesi teknikleri denetimli ve denetimsiz öğrenme metotlarından oluşur. Denetimli öğrenme, önceden gözlemlenmiş ve sonuçları bilinen (etiketlenmiş) verileri kullanarak bu verileri ve sonuçlarını kapsayan bir fonksiyon oluşturmayı amaçlayan makine öğrenimi metodudur. Denetimsiz öğrenme, etkilenmemiş verideki gizli yapıyı bulma işlemidir. Yani veriler arasında var olan ama gözle görülemeyen bağlantının açığa çıkarılması işlemidir [39].

Veri tabanı odaklı yaklaşım: Veri tabanı odaklı yöntemler, diğer iki yöntemde olduğu gibi en iyi model için arama yapmaz. Bunun yerine, eldeki verilerin özelliklerini kullanmak için veri modelleme veya veri tabanına özgü sezgisel yöntemler kullanılır.[40]

Diğer yaklaşımlar: Veri madenciliği için sinir ağları da dahil birçok başka teknik benimsenmiştir. Sinir ağı, nöronlar olarak adlandırılan birbirine bağlı düğümler dizisidir. Bir nöron, girişlerinin bir fonksiyonunu hesaplayan basit bir cihazı temsil eder. Girişler, başka nöronların çıktıları veya bir nesnenin özellik değerleridir [40].

1.4.4. Veri Madenciliği Yöntemleri ve Kullanılan Algoritmalar

Veri madenciliği genel olarak üç temel grupta değerlendirilmektedir. Bunlar sınıflandırma, kümeleme, birliktelik kurallarıdır. Sınıflandırma, kümele ve birliktelik kuralı için farklı algoritma ve yöntemler kullanılır.

1.4.4.1. Sınıflandırma

Karmaşık veri kaynaklarının analizinde kullanılan en uygun ayıklama yöntemlerimden biri sınıflandırmadır. Sınıflandırma işlemi daha önce görülmemiş ve kategorisi bilinmeyen her bir örneğin eğitim verisi kategorilerinde en uygun olan kategoriye atanması işlemidir [41].

Sınıflama iki adımda gerçekleşmektedir. Bunlar verilerin eğitimi ve modelin testidir. Eğitim, eğitim kümesinden çıkarımla modelin oluşturulması, test ise test kümesini kullanarak modelin kesinliğinin kontrol edilmesidir. Modelin kesinliğinin belirlenmesi için test örneklerinin iyi bilinen sınıfı, model tarafından tahmin edilen sınıf ile karşılaştırılır [42].

Sınıflama ile ilgili farklı yaklaşımlar söz konudur. Bu yaklaşıma göre sınıflama; tam sınıflama ve kısmi sınıflama olmak üzere ikiye ayrılmaktadır. Buradaki tam sınıflandırma kavramı veri içindeki tüm sınıflar ve örnekleri kapsayan model ile ilgilidir. Bunlara örnek olarak, yapay sinir ağları, C&RT, CHAID, C4.5, C5.0 ve diğer karar ağaçları gibi akıllı teknikler verilebilir. Kısmi sınıflamada, tam sınıflamada olduğu gibi veri sınıflarının özellikleri gösterilmektedir. Ancak kısmi sınıflandırma modellerinde tüm sınıflar veya verilen sınıfın tüm örnekleri kapsanmayabilir. Bu modellerden biri olarak birliktelik kuralları verilebilir [43].

1.4.4.2. Kümeleme Analizi

Verinin ait olduğu sınıfa ile ilgili yeterli bilgi bulunmayan durumlarda, veriye ilişkin tahminlerin yapılmasında kullanılan bir yöntemdir. Sınıflamanın aksine, sınıflanmış veriye dayalı bir yöntem değildir. Kümeleme denetimsiz öğrenme yöntemidir [44]. Kümeleme önceden belirlenen seçme kriterine göre birbirine çok benzeyen verileri aynı küme içinde gruplandırmaktadır. Analiz sonucunda bir kümeye ait elemanlar birbirine benzer, diğer

kümelere ait elemanlar ile çok farklıdır. Kümeleme analizi firmaların müşteri kategorisi ve dolandırıcılık tespiti gibi sorunların çözümünde yaygın biçimde kullanılmaktadır [38].

Kümeleme analizinin geniş kullanım alanına sahip olması farklı kümeleme yaklaşımlarının doğmasına neden olmuştur. Kümeleme analizinde en yaygın kullanılan iki yaklaşım mevcuttur. Hiyerarşik yaklaşım, birbirine en çok benzeyen iki elemanı aynı kümeye atarak başlayıp tüm benzer elemanları aynı gruba atanması ile biten bir yaklaşımdır. Hiyerarşik olmayan yaklaşım, tüm verilerin ortalama değerlerine en yakın değere sahip elemanların aynı kümeye atanması kuralına dayalı bir yaklaşımdır. Hiyerarşik olmayan kümeleme analizi yöntemlerinde en yaygın kullanılan yöntem K-ortalamalar yöntemidir [45].

K-ortalamalar yöntemi, veri içindeki kümeleri bulmayı sağlayan en net ve etkili kümeleme algoritmasıdır. K-ortalama kümeleme yöntemi değerlendirmede yaygın olarak hata kareler toplamını (HKT) kullanmaktadır. En düşük hataya sahip olan kümeleme sonucu en iyi sounucu vermektedir. Nesnelerin bulundukları kümenin merkez noktalarına olan uzaklıklarının kareleri toplamı (1) eşitliğindeki gibi hesaplanmaktadır [46].

$$HTK = \sum_{i=1}^{K} \sum_{x \in C_i} dist^2(m_i, x) \quad (1)$$

Burada dist, iki eleman arasındaki standart Öklid uzaklığı, x değeri C_i kümesinde bulunan bir elaman, m_i değeri C_i kümesinin merkez noktasıdır. K-ortalama algoritması iki boyutlu veriler üzerinde Öklid uzak ölçütüne göre çalışmaktadır [47].

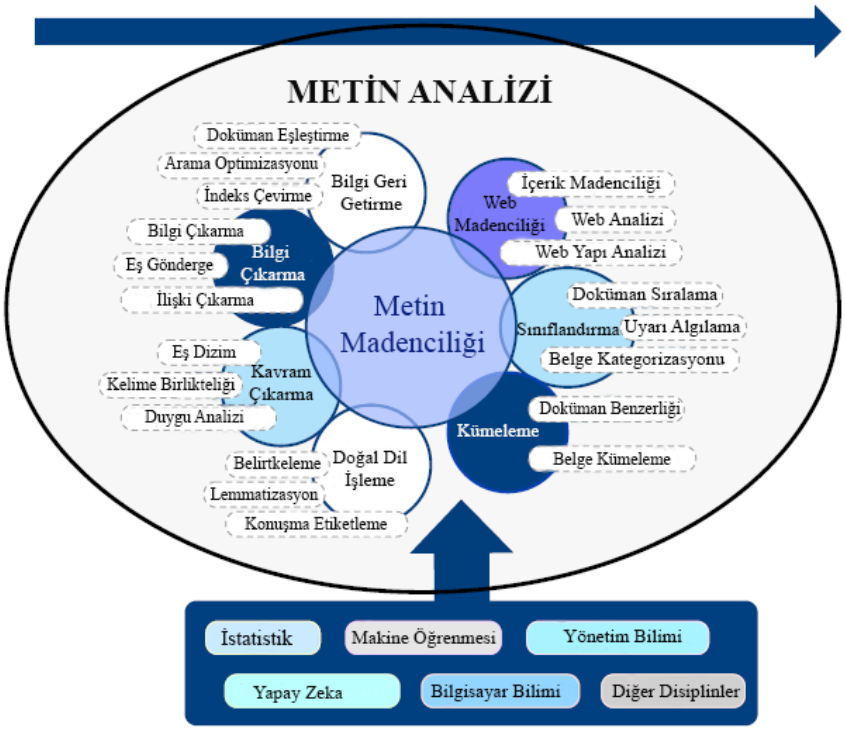
1.4.4.3. Birliktelik Analizi

Birliktelik kuralları veri madenciliğinin önemli örneklerinden biridir. Birliktelik kuralları veriler arasındaki potansiyel ilişkileri tanımlar ve büyük veri tabanlarından müşterilerin alışveriş davranışları keşfetmeyi sağlar. Birliktelik kuralları bir ürün satın alındığında o ürünle beraber başka hangi ürünlerin satın alındığının belirlenmesi amacıyla kullanılır [48]. Örneğin, tatil dolasıyla tüm aile bireylerine uçak bileti alan bir müşteri, %95 olasılıkla tatil beldesinde araba da kiralayacaktır. Perakende sektöründeki büyük mağazalar bu tekniği müşterilerinin satın alma eğilimlerini belirlemek için kullanmaktadır [49].

1.4.5. Metin Madenciliği

Veri madenciliği, donanım ve yazılım teknolojisindeki büyük gelişmelere bağlı olarak son yıllarda hızlı ilerlemeler gösteren ve farklı veri türlerine uygulanılabilen bir alan oluşturmaktadır. Bu durum özellikle web ve sosyal ağlar için donanım ve yazılım platformlarının geliştirilmesiyle büyük miktardaki metin türündeki veriler için de geçerlidir. Farklı uygulamalardan elde edilen metin verilerinin artan miktarı, verilerin dinamik ve ölçülebilir bir şekilde öğrenebilen algoritmik tasarımların geliştirilmesine ihtiyaç duyulmuştur [50].

Metin madenciliği uygulaması insan beynindeki en karmaşık analitik işleme sisteminden anlayışları ve anlayışı, yazı dilinde analiz etmeyi amaçlamaktadır. Metin türündeki verilerde kelime ve cümleler her zaman doğal sıralanışında değildir. Bu nedenle metin madenciliğinde kullanılacak istatistiksel yöntemlerin doğru sonuçlar verebilmesi için doğal dil işleme teknikleriyle kelime ve cümlelerin ön işlemesinin dikkatli bir şekilde yapılması gerekmektedir [51]. Metin madenciliği teknikleri kullanılarak yapılan işlemler ve ilişkili olan disiplinler Şekil 6.'da gösterilmektedir.



Şekil 6. Metin analizi ile yapılan işlemler ve ilgili disiplinler

1.4.5.1. Sosyal Medyada Metin Madenciliği

Sosyal medya, sosyal etkileşim için erişilebilir ve ölçeklenebilir iletişim teknikleri kullanılarak sosyal etkileşim içinde kullanılan ortamdır. Sosyal medya, bilgi ve deneyimleri diğer insanlarla daha verimli şekilde paylaşmak ve tartışmak amacıyla, internet araçlarının kullanılması olarak da tanımlanabilir [52].

Gazete, televizyon ve radyo gibi geleneksel medya, iş dünyasından tüketiciye tek bir dağıtım paradigmasını izlemektedir. Bilgi, medya kaynaklarından veya reklam verenlerden üretilir ve medya tüketicisine iletilir. Bu geleneksel yöntemden farklı olarak internet kullanımının yaygınlaşmasıyla birlikte bu paradigma tüketiciden tüketiciye hizmet şeklini almıştır. Bu nedenle sosyal medya ortamları üreticilere pazar araştırması, kampanya takibi, müşteri memnuniyeti gibi önemli konularda çok büyük kolaylık sağlamaktadır [53].

1.4.5.2. Sosyal Medya Metinlerinin Farklı Yönleri

Metin analizi teknikleri, araştırma veya iş amaçlı sosyal medyadaki metin verilerini verimli bir şekilde ele almamıza yardımcı olmaktadır. Sosyal medyadaki metin verileri, daha önce hem ölçek olarak hem de kapsam olarak ait olduğu sosyal ağlar ve gruplar hakkında bilgi verir. Fakat sosyal medyadaki metin verileri, belirgin özelliklerinden dolayı birçok yeni zorluk getirmektedir [26].

Sosyal medyada kullanılan dil ve üslubun doğru yazım kurallarına sahip olmaması nedeniyle metin analizi için zorlayıcıdır. Genellikle twitler kısaltmalar, argo, alana özgü terimler, yazım ve dilbilgisi hataları bakımından oldukça zengindir. DDİ teknikleri genellikle standart dili ele almak üzere geliştirilir ve bu nedenle bu türden düzensiz metinlerde daha düşük kaliteli sonuçlar üretme eğilimindedir [54].

Metnin ön işlemesinde bileşenleri kendine özgü olan twitter için kabul edilen birtakım teknikler bulunmaktadır. Bunlar sms işlemlerinde kullanılan teknikler, emoji listeleri ekleme, @ bahsetmek, # etiket gibi nesneleri ayrı ayrı tanımlamak, ortak kısaltmaları tam sözleriyle değiştirilmesi (Tşk=teşekkürler) gibi yöntemlerdir. Ön işleme aşamasında kurallara uygun yazılmamış ifadelerin tamamen çıkarılıp yerine doğru yazım şeklinin koyulması büyük bir hatadır. Sistemin hatalı verilere göre geliştirilip otomatik olarak kurallı yazım şeklini tespit etmesi sağlanmalıdır [55].

1.4.6. Duygu Analizi

Duygu analizi duyguların metinlerde hangi yollarla anlatıldığını ve bu anlatımlarda olumlu veya olumsuz durumların tespit etmeyi sağlayan bir analizdir [4]. Duygu analizi ya da fikir araştırması, insanların görüşlerinin, değerlendirmelerinin, tutumlarının ve duygularının hesaplamaya dayalı olarak çalışılmasıdır. Duygu analizi son on beş yılda, özellikle web verisi için popüler hâle gelmiştir. Bir duygu analizi programı kullanılan kelimelerin ve ifadelerin özelliklerine dayalı olarak metinlerin duygu içeriğini tahmin etmeyi amaçlamaktadır [56].

Duygu analizinde görev teknik açıdan zor ama pratik olarak çok faydalıdır. Örneğin, işletmeler her zaman ürün ve hizmetleri hakkında tüketici görüşlerini öğrenmeyi ve potansiyel müşterilerini tespit etmeyi istemektedir. Potansiyel müşteriler, bir hizmeti kullanmadan veya bir ürünü satın almadan önce mevcut kullanıcıların görüşlerini öğrenmek için sosyal medyayı kullanmaktadır. Bundan dolayı yazılı metinlerin duygu durumunun tespit edilmesi işletmeler için büyük öneme sahiptir [53].

İnsanlar her duygu için farklı yoğunluklara sahip olabilir. "Bu dükkâna çok kızıyorum", "Turkcell'den çok mutluyum" ve "mevcut ekonomik durumla birlikte, işimi kaybetmekten korkuyorum", bu cümlelerde görüldüğü gibi görüşlerin gücü sevinç, öfke ve korku duygularının yoğunluklarıyla ilgilidir [57].

Psikolojist teorisyenler, duyguları kategorilere ayırmıştır. Bununla birlikte teorisyenler arasında kabul görmüş temel duygular mevcut değildir. Temel duygular konusunda teorisyenlerin ortaya koyduğu temel duygular Tablo 1'de gösterilmiştir.

Tablo 1. Farklı teorisyenlerden temel duygular [58]

Teorisyenler	Temel Duygular
McDougall (1926)	Öfke, iğrenme, gurur, korku, boyun eğme, öneri duygusu, merak
Watson (1930)	Korku, aşk, öfke
Mowrer (1960)	Acı, zevk
Arnold (1960)	Öfke, nefret, cesaret, keder, arzu, umutsuzluk, korku, nefret, umut, aşk, üzüntü
Izard (1971)	Öfke, küçümseme, iğrenme, sıkıntı, korku, suçluluk, ilgi, sevinç, utanç, sürpriz
Plutchik (1980)	Güven, öfke, beklenti, nefret, sevinç, korku, üzüntü, sürpriz
Ekman vd. (1982)	Öfke, iğrenme, korku, sevinç, hüzün, sürpriz
Gray (1982)	Endişe, sevinç, öfke, korku
Panksepp (1982)	Beklenti, korku, öfke, panik
James (1884)	Korku, keder, aşk, öfke
Tomkins (1984)	Öfke, ilgi, küçümseme, iğrenme, sıkıntı, korku, sevinç, utanç, sürpriz
Weiner and Graham (1984)	Mutluluk, üzüntü
Oatley and Jobnson-Laird (1987)	Öfke, iğrenme, kaygı, mutluluk, üzüntü
Parrott (2001)	Öfke, korku, sevinç, aşk, üzüntü, sürpriz

Duygu analizi konusunda teorisyenlerin görüşlerindeki farklılıklar yapılan çalışmaları etkilememektedir. Yapılacak olan uygulamalar için gerekli gördüğümüz duyguları seçmemizde herhangi bir sakınca yoktur [58].

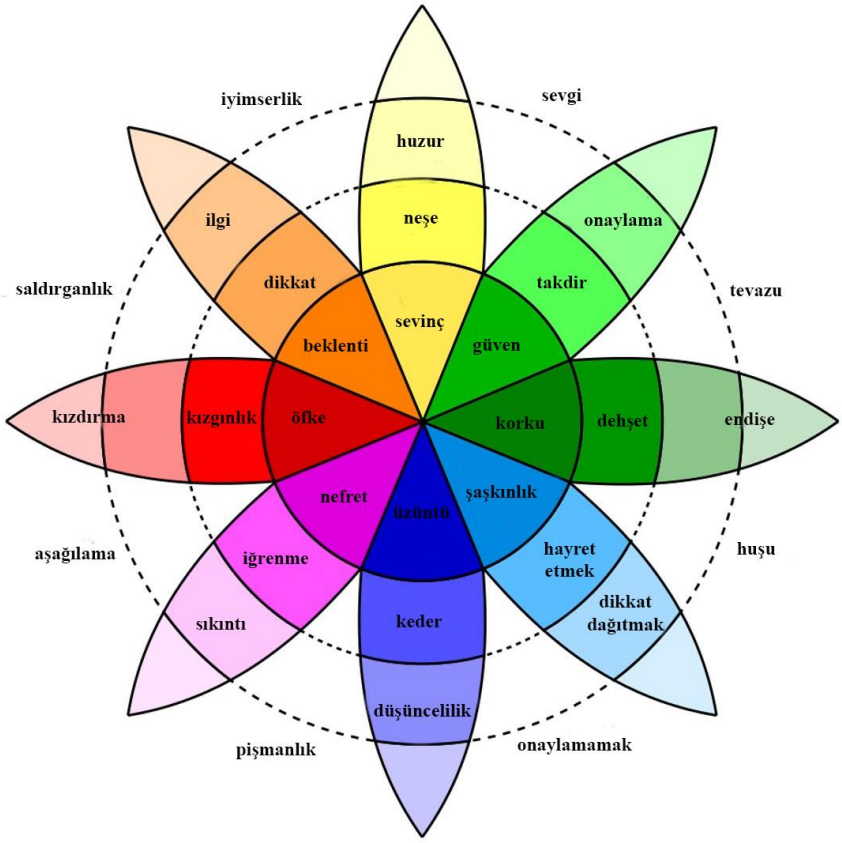
İnsan-Makine Duygu Etkileşim Ağı (İMDEA) 48 duyguyu farklı pozitif ve negatif kutuplara ayırmak için Duygu Açıklaması ve Temsil Dili (DATD) önermiştir (İMDEA, 2006). Tablo 2.'de İMDEA tarafından önerilen duygu kutupları ve duygular listelenmiştir.

Tablo 2. İMDEA duygu kutupları ve duygular [59]

Olumsuz ve Güçlü	**Olumsuz ve edilgen**	**Durgun Olumlu**
Öfke	Can sıkıntısı	Sakin
Sıkıntı	Umutsuzluk	Memnun
Aşağılama	Hayal kırıklığı	Rahatlamış hissetmek
İğrenme	İncitmek	Huzurlu
Nefret	Üzüntü	Sempatik
Kızgınlık		Şefkatli
		Dostça
Olumsuz ve Kontrol Altında Değil	**Olumlu ve Neşeli**	Sevgi
Kaygı	Keyif	
Sıkıntı	Sevinç	Tepkili
Korku	Heyecan	İlgili
Çaresizlik	Mutluluk	İnce
Güçsüzlük	Eğlenceli	Şaşkın
Endişelenmek	İstek	
Olumsuz Düşünceler	**Olumlu Düşünceler**	
Şüphe	Cesaret	
İmrenme	Umut	
Hüsran	Gurur	
Suç	Memnuniyet	
Utanç	Güven	
Endişe		
Stres		
Şok		
Gerginlik		

Tablo 2.'deki duygu kutupları ve duygular kullanılarak metinlerin duygu durumuna göre sınıflandırılması çok kolay ve kullanışlı olmaktadır. Bununla birlikte bazı duyguların olumlu veya olumsuz eğilimleri olmadığından dolayı dikkate almamalıyız. Sürpriz ve ilgi durumları bunlara örnek olarak verilebilir. Bazı psikologlar, bunların olumlu ya da olumsuz değerlere sahip olmadığı için duygu olarak görülmemesi gerektiğini düşünmektedir [60]. Bu nedenle duygu analizinde bu duyguların kullanılması çok yaygın değildir.

Duyguların birbiriyle ilişkisini anlamak, tespit edilen duygu durumlarının olumsuz duruma yoğunlaşmadan engellememizi kolaylaştıracaktır. Bu durum işletmelere büyük bir avantaj sağlayacaktır. Duyguların birbiriyle olan ilişkilerini anlamak için Şekil 7.'den yararlanılmıştır.



Şekil 7. Dr. Robert Plutchik'ın duygu şeması [61]

R. Plutchik’in modeli, duygular arasındaki ilişkileri anlatmakta ve bu duygular arasındaki karmaşık ilişkileri, duyguların zaman içinde nasıl etkileşimde bulunduğunu ve değiştiğini anlamada oldukça yardımcı olan bir modeldir. Merkezde sekiz bölüm bulunmakta ve sekiz temel duygu boyutu olduğunu belirtmektedir. Bu duygular öfke, beklenti, sevinç, güven, korku, şaşkınlık, üzüntü ve nefrettir. Duygular dışarıdan merkeze doğru ilerledikçe yoğunlaşmaktadır. Örneğin, sıkıntı duygusu, göz ardı edilirse nefret duygusuna doğru yoğunlaşmaktadır. Bu ilişkilerde dikkat edilmesi gereken kural, kontrol edilmeyen duygular yoğunlaşabilir.

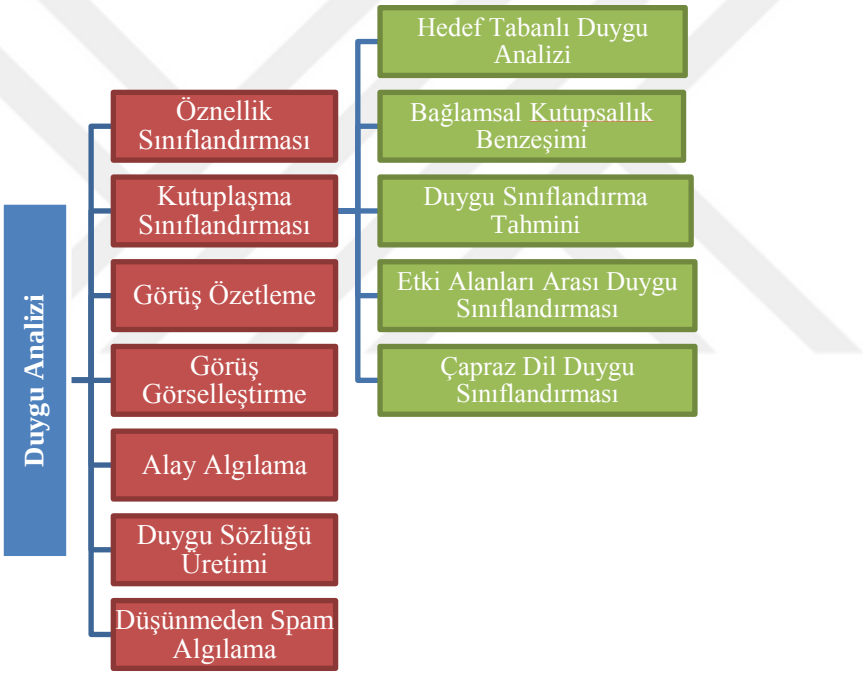
Bu şemada her daire sektörünün karşısında zıt bir duygu bulunmaktadır. Üzüntünün karşısında sevinç ve güven duygusunun karşısında nefret bulunmaktadır. Hiç renk göstermeyen duygular iki temel duygunun karışımı olan bir duyguyu temsil etmektedir. Örneğin, beklenti ve sevinç iyimser duygusu için birleştirilir. Duygular genellikle karmaşıktır ve bir duygu aslında iki veya daha fazla farklı duygunun birleşiminden oluşabilmektedir.

Bir metin çözümleme altyapısıyla ara birim oluşturan web servisi Convey API, Plutchik'in duygu çarkını kullanarak sosyal medya takibi için duygu analizi yapmaktadır [61].

1.4.6.1. Duygu Analizi İle Yapılan İşlemler

Duygu analizi uygulama alanlarına bağlı olarak, görüş inceleme, fikir çıkarma, duygu madenciliği, öznel analiz ve etki analizi gibi birkaç isimle adlandırılmaktadır [62].

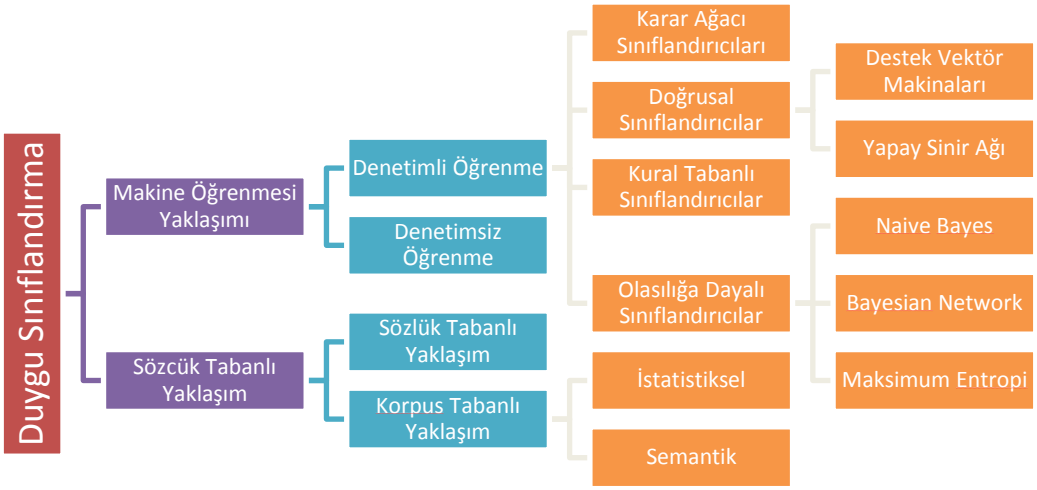
Duygu analizi ile yapılan işlemlerinin sınıflandırması Şekil 8'de sunulmuştur.



Şekil 8. Duygu analizi ile yapılan işlemler

1.4.7. Duygu Sınıflandırma Teknikleri

Genel olarak kabul edilen iki çeşit duygu sınıflandırma yöntemi kullanılmaktadır. Bunlar sözlük temelli yaklaşım ve makine öğrenmesi yaklaşımıdır [63]. Duygu sınıflandırmada en çok tercih edilen teknik makine öğrenmesi yöntemindeki denetimli öğrenme yöntemidir [58]. Duygu sınıflandırmada kullanılan tüm teknikler Şekil 9'de gösterilmiştir.



Şekil 9. Duygu sınıflandırma teknikleri

1.4.7.1. Makine Öğrenmesi Yaklaşımı

Makine öğrenme yaklaşımı, makine öğrenme algoritmalarına dayanarak söz dizimi kuralları ve dil özelliklerini kullanarak duygu analizi gibi bir metin sınıflandırma problemini çözmek için kullanılmaktadır.

Metin sınıflandırma problemi tanımlanırken ait olduğu sınıf etiketleri bilinen X_1, X_2,…,X_n dokümanlarından oluşan $D=\{X_1, X_2,\ldots,X_n\}$ şekilde bir eğitim kümesi belirlenir.

Bu eğitim kümesindeki dokümanların temel özellikleri kullanılarak sınıflandırma modeli oluşturulur. Daha sonra sınıflandırma modeli, sınıfı bilinmeyen bir metnin sınıf etiketini tahmin etmek için kullanılır [64].

1.4.7.2. Denetimli Öğrenme

Denetimli öğrenme yöntemleri, etiketli eğitim kümesini temel almaktadır. Denetimli öğrenme yönteminde, bir eğitim seti kullanılır ve amaç, önceden görülmeyen örnekleri tahmin etmektir. Bu yöntemde, öznitelikleri belirlenen mevcut metin verisi ile başka bir hedef öznitelik arasındaki ilişkiyi keşfetmeye çalışan yöntemlerdir. Bulunan ilişki, model olarak belirtilen yapıda temsil edilmektedir. Genellikle modeller, veri kümesinde gizlenmiş ve belirlenen öznitelikler sayesinde hedef öznitelik arasındaki kullanılabilir olayları ortaya çıkarır [27].

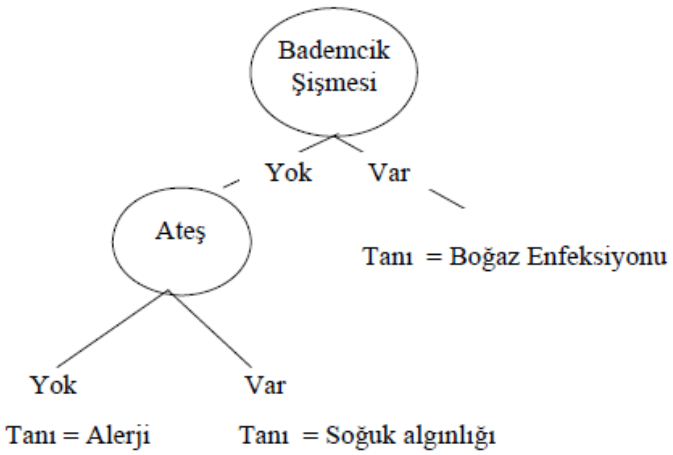
Denetimli öğrenme yöntemleri, pazarlama, finans ve imalat gibi çeşitli alanlarda uygulanabilir. Örneğin, sınıflandırıcılar banka kredisi kullananların iyi ve kötü olarak sınıflandırması için kullanılabilir.

1.4.7.3. Karar Ağaçları

Karar ağaçları sınıflandırma problemlerinde en çok kullanılan algoritmalardan biridir. Diğer yöntemlerle kıyaslandığında karar ağaçlarının yapılandırması ve anlaşılması daha kolay denilebilir [65].

Karar ağaçları bilgi keşfi sırasında pek çok test gerçekleştirilerek, hedefi tahmin etmede en iyi sırayı bulmaya çalışırlar [66]. Karar ağaçları, verileri özellik değerlerine göre hiyerarşik olarak sıralayarak sınıflandırır. Ağacın her düğümü verinin bir özelliğini temsil eder ve her dalı o özellik için bir değeri temsil eder.

En iyi özelliği seçmek için farklı yöntemler kullanılmaktadır. En sık kullanılan yöntemler entropi ve bilgi kazanımı ölçütüdür. Her adımda, en iyi özellik, özelliği kullanarak verileri ayırarak elde edilen bilgi kazanımına göre seçilir. Özellik seçimi, durma durumu elde edilinceye kadar sürdürülür. Yeni veri ağacın kök düğümünden başlayarak sınıflandırılır. Her düğümde veri düğümündeki seçilen özelliğin değeri, o düğümün dalının değeri ile karşılaştırılır [67]. Şekil 10'da örnek bir karar ağacı sunulmuştur.



Şekil 10. Hasta veri tabanı için bir karar ağacı ve kurallar

Aşağıdaki kurallar şeklinde ağaç ağacı tamamlanmaktadır.

Eğer bademcik şişmesi = var

Sonra tanı = Boğaz enfeksiyonu

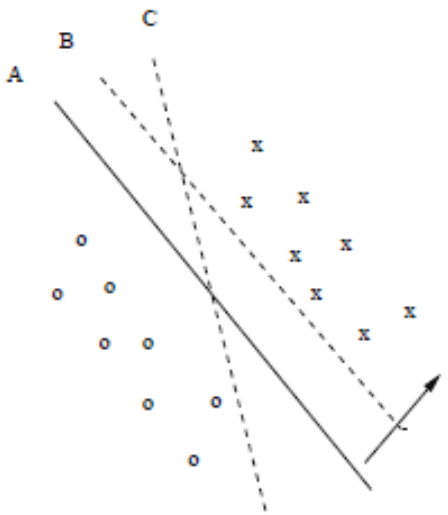
Eğer bademcik şişmesi = yok ve ateş = var

Sonra tanı = soğuk algınlığı

Eğer bademcik şişmesi = yok ve ateş = yok [42].

1.4.7.4. Doğrusal Sınıflandırıcılar

$\overline{X} = \{x_1, \ldots, x_n\}$ normalize edilmiş belge sözcük frekansı, $\overline{A} = \{a_1, \ldots, a_n\}$ nitelik boyutu ile aynı boyuta sahip doğrusal katsayıların bir vektörü, b eğilim değeridir. Doğrusal tahmin edicinin çıktısı $p = \overline{A}\,\overline{X} + b$ formülü ile tahmin edilir. Birçok doğrusal sınıflandırıcı türü vardır [64]. Destek Vektör Makineleri (DVM) farklı sınıflar arasındaki en iyi doğrusal sınıflandırma yöntemlerinden biridir [26]. Şekil 11'de hiper düzlemlerin örnek görseli sunulmuştur.



Şekil 11. Hiper-düzlemler

1.4.7.5. Destek Vektör Makineleri

DVM ilk olarak sayısal veriler için 1995 yılında kullanılmıştır [68]. DVM istatistiksel öğrenme teorisine dayalı bir kontrollü sınıflandırma algoritmasıdır. DVM'nin sahip olduğu matematiksel algoritmalar başlangıçta iki sınıflı doğrusal verilerin sınıflandırılması problemi için tasarlanmıştır. Daha sonra çok sınıflı ve doğrusal olmayan verilerin sınıflandırılması için genelleştirilmiştir. DVM'nin çalışma prensibi iki sınıfı birbirinden ayıran en uygun

karar fonksiyonunu tahmin edilmesi, başka bir ifadeyle en uygun şekilde ayırabilen hiper-düzlemin tanımlanması esasına dayanmaktadır [69].

DVM ile sınıflandırmada genellikle {-1, +1} şeklinde sınıf etiketleri ile gösterilen iki sınıfa ait örneklerin, eğitim verisi ile elde edilen bir karar fonksiyonu yardımıyla birbirinden ayrılması amaçlanır. Belirlene karar fonksiyonu kullanılarak eğitim verisini en uygun şekilde ayırt edebilecek hiper-düzlem bulunur [70].

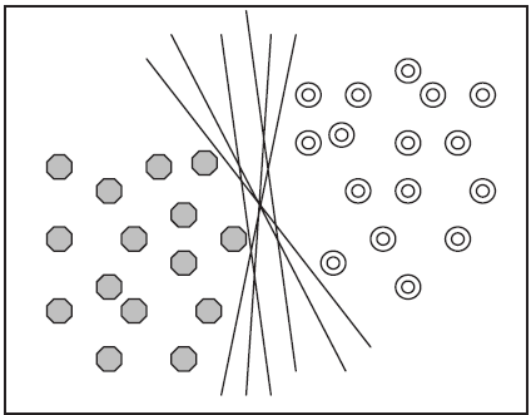
DVM'lerin temel ilkesi, arama alanındaki farklı sınıfları en iyi şekilde ayıran nitelikleri belirlemektedir. Örneğin, Şekil 11'deki 'x' ve 'o' ile gösterilen iki sınıfa sahip olduğumuzu düşünelim. Sırasıyla A, B ve C ile gösterilen üç farklı düzlem belirlenmiştir. A düzleminin farklı sınıflar arasında en iyi ayrımı sağladığı görülmektedir. Çünkü herhangi bir veri noktasına uzaklığı maksimum olan düzlem A düzlemidir. Bu nedenle, A düzlemi maksimum ayrım sınırına sahiptir [26].

Doğrusal olarak ayrılabilen iki sınıflı bir sınıflandırma probleminde DVM'nin eğitimi için k sayıda örnekten oluşan eğitim verisinin $\{x_i, y_i\}$, i=1,…k, olduğu kabul edilirse, optimum hiper-düzleme ait eşitsizlikler aşağıdaki şekilde olur:

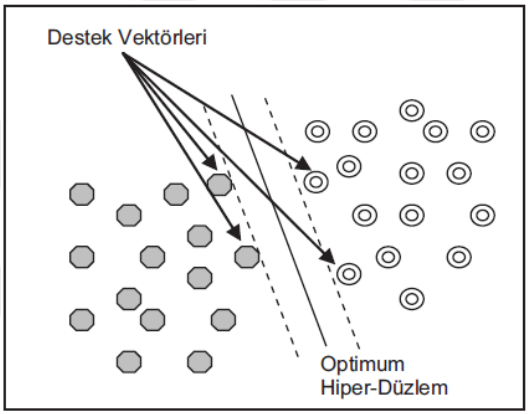
$$w.x_i + b \geq +1 \text{ her } y = +1 \text{ için} \quad (2)$$

$$w.x_i + b \leq +1 \text{ her } y = -1 \text{ için} \quad (3)$$

Burada $x \in R^N$ olup N-boyutlu bir uzayı $y \in \{-1, +1\}$ise sınıf etiketlerini, w ağırlık vektörünü ve b eğilim değerini göstermektedir [71]. En uygun hiper-düzlemin belirlenebilmesi için bu düzeleme paralel ve sınırlarını oluşturacak iki hiper-düzlemi belirlemek gerekmektedir. Bu hiper-düzlemleri oluşturan noktalar destek vektörleri olarak adlandırılır ve bu düzlemler $w.x_i + b = \pm 1$ şeklinde ifade edilir. Şekil 12 ve Şekil 13'de hiper-düzlemlerin çeşitlerine ait görsel sunulmuştur.



Şekil 12. İki sınıflı bir problemin hiper-düzlemleri [70]



Şekil 13. Optimum hiper-düzlem ve destek vektörleri [70]

1.4.7.6. Yapay Sinir Ağları

İnsan beyninin fonksiyonundan esinlenen Yapay Sinir Ağları (YSA), deneme yoluyla öğrenme ve genelleştirme yapabilmektedir. YSA'nın kullanıldığı önemli alanlardan biri de geleceği tahmin etmektir. YSA, veriler arasındaki bilinmeyen ve fark edilmesi zor olan bağlantıları ortaya çıkarabilir [72].

YSA'nın temel yapısına nöron veya birim adı verilir. Metin analizinde kullanılan YSA'da her birim i. dokümana karşılık gelen ve X_i terim frekansları ile gösterilen bir girdi seti alır. Her nörona bir ağırlık değeri A verilir ve bu ağırlıklandırma değeri fonksiyon yardımıyla hesaplanır.

$$p_i = A.\overline{X}_i \quad (4)$$

İkili bir sınıflandırma probleminde X_i sınıfının etiketi y_i ile gösterilmekte ve p_i tahmin fonksiyonu ile etiket tahmin edilmektedir [26].

1.4.7.7. Olasılıksal Sınıflandırıcılar

Olasılıksal Sınıflandırıcılar, sınıflandırma için karışım modelini kullanılır. Karışım modelinde her sınıfın karışımın bir parçası olduğu varsayılır. İstatistiksel teorilere dayanan hesaplamalar yapılarak sınıflandırma gerçekleştirilir [64].

1.4.7.8. Yalın Bayes Sınıflandırıcısı

Yalın Bayes Sınıflandırıcısı (YNS), dokümanların sınıflandırılmasında yaygın olarak kullanılan bir sınıflandırıcıdır. YNS'nin en önemli özelliği sınıfın içerisindeki verinin özniteliklerini birbirinden bağımsız olduğunu varsaymasıdır. Bu durum "Yalın Bayes varsayımı olarak adlandırılmaktadır.". Bağımsızlık varsayımı nedeniyle, her özellik için ayrı ayrı öğrenilebilir ve bu özellikle niteliklerin fazla olduğu durumlarda öğrenmeyi büyük ölçüde basitleştirir [73].

Doküman sınıflandırması, çok sayıda öznitelik içeren bir alandır. Sınıflandırılacak örneklerin nitelikleri kelimelerdir ve farklı kelimelerin sayısı çok büyük olabilir. YNS bu alanlarda oldukça başarılıdır [74].

Bayes teoremi verilen kelimelerden oluşan nitelik kümesinin belirli etikete ait olma olasılığını tahmin etmek için kullanılır. (5)'te Bayes sınıflandırıcısının formülü sunulmuştur.

$$P(etiket \setminus nitelik) = \frac{P(etiket) * P(nitelik \setminus etiket)}{P(nitelik)} \quad (5)$$

P(etiket), bir etiketin oluşma olasılığı veya rasgele bir özelliğin etiket olarak belirlenmesi olasılığıdır. $P(etiket \setminus nitelik)$ belirli bir niteliğin etiket olarak sınıflandırılması olasılığını ifade etmektedir. P(nitelik), ise belirli bir özellik kümesinin oluşması olasılığıdır.

YNS için ikili bağımsızlık modeli ve multinominal model olmak üzere iki farklı modeli bulunmaktadır. YNS'ye göre kelimeler sınıftan bağımsızdır [75].

İkili Yalın Bayes Modeli: Bu yöntemde kelimelerin dokümana ait olup olmaması temel alınmaktadır. Dokümanlardan oluşturulan sözlükteki kelimeler, sınıflandırılacak olan dokümanla karşılaştırılır. Sözlükteki kelime dokümanda varsa 1, yoksa 0 değeri verilir ve dokümanlar 1'ler ve 0'larla ifade edilir. Bu işlemin sonunda sözlük boyutu kadar 1'ler ve 0'lardan oluşan vektörler elde ederiz.

$V = \{0,0,0,1,0,1,1\}$ şeklinde elde edilen vektörün hangi sınıfa ait olduğu (6), (7) ve (8)'de gösterildiği gibi hesaplanmaktadır.

$$P(d\backslash c_j) = \prod_{t=1}^{|V|} P(w_t\backslash c_j)^{x_t} \left(1 - P(w_t \setminus c_j)\right)^{(1-x_t)} \quad (6)$$

$$P(w_t\backslash c_j) = \frac{1 + B_{jt}}{2 + |c_j|} \quad (7)$$

Burada;

|V|: Sözlükteki kelimelerin sayısı

B_{jt}: c_j kategorisinde bulunan ve w_t kelimesini içeren eğitim doküman sayısı

$|C_j|$: c_j sınıfında bulunan eğitim dokümanı sayısı

x_t: kelime ağırlığı $x_t \in \{1,0\}$ ifade etmektedir.

$$M(C) = P(word1\backslash C)^{n1} P(word1\backslash C)^{n2} \dots P(word1\backslash C)^{nv} \quad (8)$$

Bu algoritma iki durumlu problemlerde kullanılmaktadır. Örnek ile açıklayalım:

Sözlük: Akıllı, aptal, evcil, güzel, pis, zeki

Test dokümanı: Sen çok akıllısın.

Bu dokumanı 1'ler ve 0'lardan oluşan vektörle ifade edersek;

DS (1,0,0,0,0,0)

(6) ve (7)'den test dokümanı için hesaplamalar yapılırsa;

P(olumlu)=$\frac{1}{2}$

P(olumsuz)=$\frac{1}{2}$

M(olumlu)= $\frac{1}{2} * \frac{1}{2} * \frac{3}{4} * \frac{1}{2} * \frac{1}{2} * \frac{3}{4} * \frac{1}{2} = 0{,}075$

M(olumsuz)= $\frac{1}{2} * \frac{1}{4} * \frac{1}{4} * \frac{3}{4} * \frac{3}{4} * \frac{1}{2} * \frac{3}{4} = 0{,}0065$

P(d\Olumlu)>P(d\Olumsuz) olduğundan dolayı olumlu kategori olarak etiketlenecektir [76].

Yalın Bayes Frekans Ağırlıklandırma: İkili YBS'den farklı olarak hesaplamaya kelimelerin tekrar sayılarının da dahil edildiği geliştirilmiş bir sınıflandırma algoritmasıdır. Tekrar sayılarının hesaplamaya dahil edilmesi İkili YBS'na göre daha başarılı olduğu görülmüştür [77].

Bu modelde sözlük boyutu kadar kelime tekrarlarından oluşan bir vektör elde edilmektedir. Örneğin; (2, 3, 0,,30,90) şeklinde bir vektör elde edilir. Bayes sınıflandırıcısı için değerler (9), (10) ve (11)'deki şekilde hesaplanmaktadır.

$$P(d\backslash c_j) = P(d)d! \prod_{i=1}^{|V|} \frac{P(w_t\backslash c_j)^{xt}}{X_t!} \tag{9}$$

$$P(w_t\backslash c_j) = \frac{1 + N_{jt}}{|V| + N_J} \tag{10}$$

d: kategori sayısı

N_{jt}: j sınıfındaki dokümanlar içinde t kelimesinin görülme sıklığı

N_j: j sınıfındaki toplam kelime sayısı

P(d): kategori olasılığı

X_t: kelimenin frekansı

V: kelime sayısı

$$M(C) = P(word1\backslash C)^{n1} P(word1\backslash C)^{n2} \dots P(word1\backslash C)^{nv} \tag{11}$$

M(C) değeri en büyük olan kategoriye doküman atanır.

D1=(0,2,0,1,0,0) Sağlık

D2=(0,0,0,1,0,0) Sağlık

D3=(1,0,0,0,0,0) Ekonomi

D4=(0,0,0,0,0,2) Ekonomi

D5=(0,0,2,0,0,0) Spor

D6=(0,0,0,1,2,0) Spor

Ağırlıklandırılmış mevcut eğitim verilerimiz olsun. Test verimiz ise; DQ=(0,0,2,0,1,0) olarak ağırlıklandırılsın.

$$M(Sağlık) = \frac{1}{3} * 3! * 1 * 1 * \left(\frac{1+0}{6+4}\right)^2 \frac{1}{1!} * 1 = 0{,}0010$$

$$M(Ekonomi) = \frac{1}{3} * 3! * 1 * 1 * \left(\frac{1+0}{6+3}\right)^2 \frac{1}{2!} * 1 * \left(\frac{1+0}{6+3}\right) \frac{1}{1!} = 0{,}0013$$

$$M(Spor) = \frac{1}{3} * 3! * 1 * 1 * \left(\frac{1+2}{6+5}\right)^2 \frac{1}{2!} * 1 * \left(\frac{1+2}{6+5}\right)^2 \frac{1}{1!} = 0{,}02$$

M(Spor) değeri en büyük olduğun için dokümanımız 0.02 olasılıkla Spor kategorisine aittir.

Bayes Ağı: Yalın Bayes sınıflandırıcının temel varsayımı özelliklerin bağımsızlığıdır. Bayes Ağı (BA) ise tüm özelliklerin tamamen bağımlı olduğunu varsaymaktadır. Bu modelde oluşturulan her bir düğüm bir rastgele değişkeni temsil etmektedir. BA, değişkenler ve bunların ilişkilerini gösteren tam bir model olarak kabul edilmektedir. Tüm değişkenler üzerinde ortak olasılık dağılımı belirtilir. Metin madenciliğinde, BA'nın hesaplama karmaşıklığı ve hesaplama maliyeti çok yüksektir. Bu nedenle, metin madenciliğinde çok kullanılmamaktadır [64].

1.4.7.9. Maksimum Entropi

Entropi bir sistemdeki belirsizliğin veya düzensizliğin ölçüsüdür. Metin verilerinin C_1, $C_{2,\ldots,}$ C_n şekilden sınıflardan oluştuğunu ve T'nin sınıf değerini gösterdiğini varsayalım. Bu durumda bir sınıfa ait olasılık $P_i = (C_i \,/\, |T|)$ şeklinde hesaplanır. Bir sınıfa ait entropi değeri (12)'deki gibi hesaplanır.

$$Entropi(T) = -\sum_{i=1}^{n} \log_2(p_i) \tag{12}$$

Metin verisindeki B özniteliğine göre T sınıf değerleri T_1,T_2,…,T_n şeklinde alt kümelere ayrıldığı göz önüne alınır. B öznitelik değerleri kullanılarak T sınıf değerlerinin bölünmesi sonucunda elde edilecek kazanç (13)'deki gibi hesaplanır.

$$Kazanç(B,T) = Entropi(T) - \sum_{i=1}^{n} \frac{|T_i|}{|T|} Entropi(T_i) \tag{13}$$

T kümesi için B özniteliğinin değerini belirlenmesinde bölümlenme bilgisi kullanılır. Bölünme bilgisi (14)'deki gibi hesaplanır.

$$Bölünme\ Bilgisi\ (B) = -\sum_{i=1}^{k} \frac{|T_i|}{|T|} \log_2 \left(\frac{|T_i|}{|T|} \right) \tag{14}$$

Bu durumda kazanç oranı (15)'deki gibi hesaplanır.

$$Kazanç\ Oranı = \frac{Kazanç(B,T)}{Bölünme\ Bilgisi\ (B)} \tag{15}$$

Bu eşitlik sınıflandırma işleminde kullanılacak ayırma ile elde edilecek bilgi oranını verir. Bu ölçüt kullanılarak kazanç oranı maksimum olacak şekilde T eğitim kümesi tekrarlı şekilde ayrılır [78].

1.4.7.10. Sözcük Tabanlı Yaklaşım

Düşünce kelimeleri, duygu sınıflandırma alanında yaygın olarak kullanılır. İstenen durumlar olumlu kelimelerle ifade edilirken, istenmeyen durumlar olumsuz kelimeler ile ifade edilmektedir. Aynı şekilde düşünce ifade eden fikir cümleleri ve deyimlerde kullanılmaktadır. Düşünce kelime listesini oluşturmak için üç temel yaklaşım vardır. Manuel yaklaşım çok zaman alıcıdır ve tek başına kullanılmaz. Otomatik yöntemlerden kaynaklanan hataları önlemek için genellikle diğer iki otomatik yaklaşımla birlikte son adımda kontrol amaçlı kullanılmaktadır [64].

1.4.7.11. Sözlük Tabanlı Yaklaşım

Bu yaklaşımda bilinen yöntemlerle manuel olarak küçük fikir sözcüklerinden bir sözlük oluşturulur. Daha sonra bu küme eş anlamlıları ve zıt anlamları geliştirilmelidir. Kelimeler arasındaki ilişkileri sayısallaştırmak için kelime ağları kullanılmaktadır. Örneğin, masa ve tahta kelimelerin arasındaki benzerlik durumunu tespit için kelime ağı kullanmak gerekmektedir. İngilizce için iyi sonuçlar veren kelime ağları mevcuttur. Türkçe'de kabul görmüş bir kelime ağı uygulaması bulunmadığı için kelime ağına dayalı işlemler yapılacağı zaman genellikle kelimelerin İngilizce karşılıklarıyla işlemler yapılmaktadır. Kelime ağı

yardımıyla başlangıçta küçük olan fikir sözlüğümüz zıt anlamlı, eş anlamlı ve bağlantılı kelimeler eklenerek genişletilir [79].

1.4.7.12. Korpus Yaklaşımı

Korpus yaklaşımı metin içeriğine özgü görüş sözcükleri bulma sorununu çözmek amacıyla kullanılmaktadır. Korpus yönteminde büyük bir korpus içindeki fikir sözcüklerini bulmak için fikir sözcüklerinin temel listesinden faydalanılır. Bu liste ile birlikte kelimelerin söz dizimsel kalıp veya kalıpları kullanılarak diğer düşünce ifade eden kelimeler tespit edilir [80].

İstatistiksel Yaklaşım: Düşünce ifade eden temel sözcükleri veya bu sözcüklere eşlik eden kalıplar istatistiksel teknikler kullanılarak tespit edilebilir. Bir kelimenin kutupsallığı kelimelerin geniş bir metinde bulunma sıklığı incelenerek bulunabilir. Sözcük pozitif metinler arasında daha sık ortaya çıkarsa, sözcüğün kutupsallığı pozitiftir. Negatif metinler arasında daha sık görülürse sözcüğün kutupsallığı negatif olarak işaretlenir. Eşit frekansa sahipse, o zaman tarafsız kelime şekilde işaretleme yapılır [81].

Benzer duygu kelimeleri sık sık birlikte aynı kutupsallığa sahip dokümanlar içerisinde yer alır. Bu yönteme Karşılıklı Noktasal Bilgi (NKB) yöntemi adı verilmektedir. Eğer iki kelime aynı bağlamda sık sık birlikte görülüyorsa, muhtemelen aynı kutupsallığa sahip olacaktır. Bu yöntemle birlikte kutupsallığı bilinmeyen bir kelimenin kutupsallığı, başka bir kelimeyle birlikte görülme sıklığı hesaplanarak bulunabilir [82].

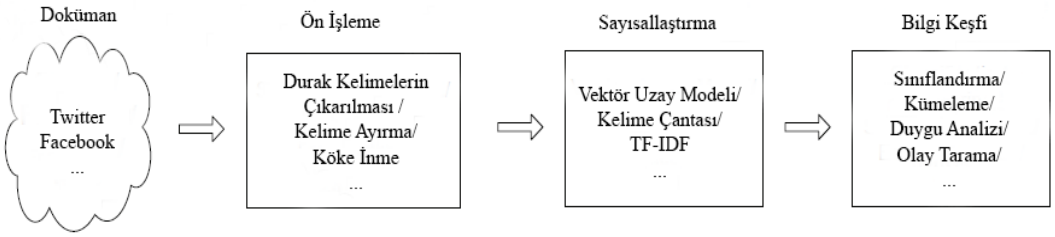
İstatistiksel yöntemler duygu analizi ile ilgili bir çok uygulamada kullanılmaktadır. Bunlardan biri, Run Testi olarak adlandırılan rastgelelik testidir. Run Testi kullanılarak inceleme yorumlarının manipülasyonu tespit edilmektedir. Hu vd. tarafından yapılan çalışmada, inceleme yorumlarının yazım tarzları müşterinin geçmişte yaşadıkları tecrübelerden dolayı rasgele olacağı ve yorumların müşteriler tarafından yazıldığı düşünülüyordu. Amazon.com'daki kitap incelemeleri üzerine yaptıkları çalışma sonucunda ürünlerin yaklaşık %10,3'ünün online inceleme manipülasyonuna maruz kaldıklarını keşfetmişlerdir [83].

Anlamsal (Semantik) Yaklaşım: Anlamsal yaklaşım doğrudan duygu durumlarını verir ve kelimeler arasındaki benzerliği hesaplamak için farklı ilkelere dayanır. Bu ilke anlamsal olarak yakın kelimelere benzer duygu değerleri verilmesidir. Örneğin WordNet (Kelimelerin Kavramsal İlişkisi Veri Tabanı), duygu kutuplarını hesaplamak için kelimeler

arasındaki farklı anlamsal ilişkileri sağlamaktadır. WordNet ile kelime eş anlamlıları ve zıt anlamlıları başlangıç sözcükleri olarak kullanılıp, daha sonra bu kelimenin pozitif ve negatif eş anlamlılarının sayımı ile bilinmeyen bir kelimenin duygu kutupsallığı belirlenebilir [84].

1.4.8. Sosyal Medyada Duygu Analizi

Sosyal medyada değişen çeşitli veri biçimleri arasında metin önemli rol oynamaktadır. Birçok sosyal medya sitesinde bilgi metin biçiminde saklanmaktadır. Geleneksel metin madenciliğinde üç aşama vardır: Metin ön işleme, metin sayısallaştırılması, bilgi keşfi aşamalarıdır.



Şekil 14. Geleneksel metin madenciliği

1.4.8.1. Metin Ön İşleme

Metin ön işleme, metin analizi işlemleri için gerekli olan bir adımdır. Metnin sayısallaştırılmasını kolaylaştırmak için giriş verilerinin tutarlı hâle getirilmesi aşamasıdır. Geleneksel metin ön işleme yöntemleri, kelime ayırma ve dolgu kelimelerinin çıkarılmasıdır. Durak kelimelerinin çıkarma, tek başına anlamsız olan kelimelerin listesini oluşturarak metin içinden bunların çıkarılması işlemidir. Köke inme işlemi; türetilen kelimeyi kök veya gövde formuna indirgeme işlemlerini kapsamaktadır [85]. Örneğin; "izlemek", "izliyor", "izledi" kelimeleri "izlemek" olarak temsil edilebilir.

Ön işleme işlem yöntemlerinin özel uygulamaları vardır. Duygu analizi veya DDİ gibi birçok uygulamada iletiyi söz dizimsel olarak analiz etmek gerekir. Bu bilgi olmadan, "Hangi üniversite başkanını mezun etti?" ve "Hangi başkan Karadeniz Teknik Üniversitesinden mezun oldu?" cümlelerini birbirinden ayırt etmek zordur. Bu cümlelerde

üst üste binen kelimeler vardır. Bu durumda doğru analizi yapmak için söz dizimi önemli olan kelimeleri kaldırmamak gerekir [26].

1.4.8.2. Doküman Normalizasyonu

Dokümanlar toplandıktan sonra dokümanın nasıl üretildiğine bağlı olarak farklı biçimde dokümanlar elde edilebilir. Örneğin; bazı belgeler, kendine özgü karaktere sahip bir kelime işlemci tarafından oluşturulmuş olabilir, diğerleri daha basit bir editörle oluşturulmuş veya taranmış görüntü olarak saklanmış olabilir. Buradan anlaşılacağı gibi tüm belgeleri işleme koyabilmek için verileri standart bir formata dönüştürmek gerekmektedir. Verilerin standardize edilmesinin en önemli avantajı, metin madenciliği araçlarının dokümanın oluş biçimi dikkate alınmadan uygulanabilmesidir [86].

1.4.8.3. Belirteçleme

Metnin işlemesindeki ilk adım, karakter akışını kelimelere veya belirteçlere bölmektedir. Metni daha basit birimler hâline getirme işlemine belirteçleme denir. Daha sonraki analizler için bu adım çok önemlidir. Belirleyicileri, tanımlamaksızın belgeden üst düzey bilgi çıkarmak çok zordur [87].

Karakter akışını belirteçlere bölmek, dilin yapısını bilen biri için çok kolaydır. Bilgisayar programı için bu durum daha karmaşıktır. Bunun nedeni programlama dillerindeki bazı özel karakterlerin belirteç olarak kullanılmasıdır. Örneğin, “(,), < , >, !, ?” gibi karakterler programlar için sınırlayıcı olurken, aynı zamanda belirteçde olabilir.

Sayılar arasındaki virgül veya iki nokta üst üste işareti genellikle bir sınırlayıcı olarak değil, sayının bir parçası olarak düşünülür. Sayılar haricinde kullanılan virgül veya iki nokta üst üste işareti belirteç olabilir. Nokta işareti, her iki tarafta da büyük harf varsa bir kısaltma olarak kullanılabilir. Bunun yanı sıra “Doktor” unvanın kısaltması “Dr.” şeklinde kullanımı da mevcuttur.

İngilizce metinlerde bu işlemleri yapmak çok daha kolaydır, çünkü boşluklar ve noktalama işaretleri kelimeleri birbirinden ayırmaktadır. Çince ve Japonya gibi boşlukların kelimeleri ayırmadığı dillerde bu işlem daha karmaşıktır. Ayrıca Almanca ve Hollandaca gibi bazı dillerde, kelimeleri birleştirmek için kısa çizgiler kullanılmaktadır. Buradan anlaşılacağı gibi yazı sisteminin kendine özgü özellikleri vardır. Belirteçleme işlemine

başlamadan önce doküman dilinin kuralları, yapısı, noktalama işaretleri göz önünde bulundurulmalıdır [27].

1.4.8.4. Soyutlama

Bir karakter akışı belirteç dizisine bölündükten sonraki adım, belirteçlerin her birini standart bir forma dönüştürmektedir. Bu sürece genellikle soyutlama denilir. Bu adımın gerekip gerekmediği uygulamaya bağlıdır. Belge sınıflandırmasında bazı durumlarda küçük bir fayda sağlayabilir.

Çekim Eklerini Soyutlama: Çoğu dilde olduğu gibi Türkçe'de de kelimeler birden fazla formda görülebilir. Örneğin, "kitaplar" ve "kitap" kelimeleri aynı kelimenin farklı formlarıdır. Bu iki kelimeyi "kitap" şekilinde normalleştirmek kelime çeşitliliğini ortadan kaldırmak için avantaj sağlamaktadır. Normalleştirme, tekil, çoğul, zaman ekleri gibi dil bilimsel değişkenlerin düzenlenmesi "köke inme" olarak adlandırılır. Dil bilim terminolojisinde bu sürece "morfolojik analiz" adı verilir. Bu analiz Türkçe için oldukça zordur [86].

Türkçe açık, fakat oldukça karmaşık morfolojik yapıyı sahiptir. Bir köke eklenen morfemler, kelimenin yapısını değiştirmektedir. Morfolojik yapılarda yapılan işlemler birkaç morfofonemik kuralla sınırlandırılmıştır. Türkçede ünlü uyumu kuralları bu kapsamda değerlendirilmektedir. Fakat bazen kelimelerde çeşitli ses olayları olmakta ve bazı harfler değişmekte veya silinmektedir. Buda dile, çeşitli dillerde yabancı kelimelerin girmesine, dilde farklı istisnaların oluşmasına neden olmaktadır [88].

Part of Speech (POS) Etiketleme: POS etiketleme, bir kelimenin dahil olduğu dilbilimsel kategoriyi ifade etmektedir. POS etiketleme, cümle içerisindeki her kelimenin hangi dilbilimsel gruba ait olduğunu tespit etme sürecidir. Bu gruplar, "isim", "fiil", "sıfat", "edat", "zamir", "zarf", "bağlaç" ve "ünlem" olarak sıralayabiliriz. Dilbilimsel bu kategorilerin belirlenmesi ve kullanılması duygu analizinde performansa doğrudan etki eder [53].

1.4.8.5. Kelime Köküne Ulaşma

Bu adımda kelimedeki çekim ve yapım ekleri atılarak kök formuna ulaşmak amaçlanmaktadır. Kök formuna ulaşmak için kök veri tabanı oluşturulup, kelimeleri karşılaştırarak ekleri silip köke ulaşılabilir. Bu aşamada dikkat edilmesi gereken kelime kökünün doğru tespit edilmesidir. Örneğin, "çalışmak" kelimesinin kök analizi yapıldığında "çal" tespit edilebilir. Bu durumda kelimenin kökü yanlış belirlenmiş olur. Bu nedenle kök analizi yapılırken, ek analizi de yapılmalıdır. Türkçe, Fince ve Macarca gibi sondan eklemeli ve çekimli dillerde bu işlemi yapmak zordur [89].

1.4.8.6. Ek Analizi

Türkçede ekler yapım eki ve çekim eki olmak üzere ikiye ayrılır. Çekim ekleri sözcüklerin sonuna eklenerek cümle içerisinde anlam bağlantısı kurmalarını sağlar. Fakat, yapım ekleri isim ya da fiil kök veya gövdelerine eklenerek onların anlamını değiştirmektedir. Morfolojik analiz yapılırken çekim ekleri atılır ama yapım ekleri atılmaz. Türkçe kelimelerin kök ve gövdelerin belirlenmesinde çeşitli yaklaşımlar bulunmaktadır.

Eklemeli dillerle ilgili üç farklı yaklaşım 1980'lerin başında geliştirilmiştir. Keçuva dili için [90], Fince için [91] ve Türkçe için [92]. Türkçe kelimelerin kök ve gövdelerini bulmak için kullanılan yöntemler şöyle sunulabilir:

AF Algoritması: AF algoritması Solak ve Oflazer [93] tarafından geliştirilmiştir. Algoritma, Türkçede sık kullanılan gövdelerden oluşan sözlük ve her bir gövdenin oluşturduğu formları gösteren 64 etiketle çalışmaktadır. Verilen bir kelimeyi, sağdan sola her adımda bir budama yapıp tekrar tekrar sözlüğe bakılmaktadır. Sözcük, kök kelimelerinin herhangi biriyle eşleşirse o kelimenin morfolojik analizi yapılır. Eğer kök formlarından herhangi biri sözlükteki kelime ile uyuşuyorsa, bulunan kök sözcüğünün söz konusu kelime için uygun gövde olduğu varsayılır. Kelime bir harfe düşene kadar süreç tekrarlanır [94].

L-M Algoritması: 1995 yılında geliştirilen ve Longest-Match olarak adlandırılan ilk algoritmadır. Türkçe kelime gövdelerinin ve olası formlarını kapsayan sözlük üzerinden kelime arama mantığına dayanmaktadır [95].

Gövde Arama Algoritması: Gövde Arama algoritması, sözcüğün tüm harflerini küçük parçalar hâline dönüştüren ön işleme aşamasından oluşmaktadır. Bunlar "Kök Bul", "Morfolojik Analiz" ve "Gövde Seçme" olmak üzere üç adımdır.

Algoritmanın ilk adımı, incelenen kelimenin tüm olası köklerini bulmaktır. Daha sonra incelenen kelimeyi bulmak için kökler ve kelime türetme kuralları kullanılmaktadır. Gövdeleme algoritmaları, sözcük anlamını görmezden gelmekte ve bazı hatalara neden olmaktadır [96]. Türkçeye özgü gövdeleme algoritmalarında, sözlük gövdeleme işlmemi için yardımcı bir yapı olarak kullanılmaktadır. Sözlükte her kök, kelimeye ait 10'lu tip bilgisi ve muhtemel kök dekorfolojik analizin kullanımı için kodlanır. Türkçede kök ve son ek kombinasyonu işleminde, kök sözcük yapısında iki değişiklik yapılabilir: kök sözcüğünün son harfi veya harflerin değişmesi, ortadaki harfin düşmesi gibi ses olayları olabilir [97].

Sözlükten muhtemel kök kelimelerinin seçimi, sözlüğün kodlanmış bilgi kısmını kullanan arama algoritması tarafından gerçekleştirilir. Algoritma, incelenen kelimenin ilk karakteriyle başlar ve bu karakter için sözlükte arama yapar. Ardından, bu ilk karaktere sonraki karakter eklenir. Bu işlem öge, incelenen kelimeye eşit olana kadar devam veya kelime sözlüğünde incelenen kelime için ilgili köklerinin olmaması durumuna kadar devam eder [98].

1.4.8.7. Durak Kelimelerinin Çıkarılması

Durak kelimeler dilde yaygın kullanılan fakat tek başına anlam ifade etmeyen kelimelerdir. Duygu analizi gibi çalışmalarda durak kelimelerin çıkarılması işlemi yapılmaktadır. Durak kelimelerini çıkarırken, durak kelimelerin bir listesine ihtiyaç duyulmaktadır. Türkçe durak kelimeleri [99] tarafından belirlenen Ek 1'deki kelimeler kullanılmaktadır. Apache firmasının metin analizi kütüphanesi olan Lucene Kütüphanesinde de bu liste kullanılmaktadır [100].

1.4.8.8. Metnin Vektörel Olarak İfade Edilmesi

Belgeleri modellemenin en yaygın yolu onları sayısal vektörlere dönüştürmek ve onlarla lineer cebirsel işlemler yapmaktır. Bu işlemler "Sözde Çanta Modeli" (SÇM) veya "Vektör Uzayı Modeli" (VUM) denir.

1.4.8.9. Vektör Uzay Modeli

Bu modelde her metin n boyutlu uzayda bir vektör olarak tanımlanmaktadır. Bu modelde metinler arasındaki benzerlik, metinler olarak temsil edilen iki vektör arasındaki açının değeri ile ifade edilir. Vektörle arasındaki açı me kadar büyükse vektörler yani metinler arasındaki benzerlik o kadar az olacaktır. İki vektörün aynı olası durumunda ise aradaki açı sıfır değerini alacak ve iki metnin birbirine benzediği tespit edilecektir [101].

Benzerlik formülü (16)'daki gibi ifade edilmektedir.

$$benzerlik(V_1, V_2) = Cos(\theta) = \frac{V_1 * V_2}{||V_1||||V_2||} \quad (16)$$

Metinleri vektörler olarak ifade edilmek için metinlerdeki terimleri tespit etmek gerekmektedir. Bu aşamada metni oluşturan tüm kelimeler terim olarak kabul edileceğinden vektör boyutu büyüyecek ve bu durum yapılan işlemin karmaşıklığını artıracaktır ve elde edilecek başarıyı olumsuz etkileyecektir. Bunun önüne geçmek için daha önce belirlenmiş olan durak kelimeleri ve metnin yapısına bağlı olarak belirlenen gereksiz kelimeler terim olarak kabul edilmez. Geriye kalan kelimelere terim ağırlıklandırma işlemi yapılarak metin içeriğiyle ilgili olan terimler tespit edilir [102].

Geleneksel olarak terim ağırlıklandırma işlemi için iki yöntem kullanılmaktadır. Bunlar Terim Frekansları (TF) ve Ters Doküman Frekansı (TDF) olarak adlandırılır. Örnek bir metin ile TF ve TDM'nin elde edelim. Örnek metnimiz, "Türk Milletinin dili, Türkçedir Türk Dili dünyanın en güzel, en zengin ve en kolay dilidir." olsun.

Tablo 3. Metni ifade eden vektör

türk	milletinin	dili	türkçedir	dünyanın	en	güzel	zengin	kolay	dilidir
2	1	2	1	1	3	1	1	1	1

Sonuç olarak örnek metin için 10 boyutlu bir vektör oluşturulacaktır ve (2,1,2,1,1,3,1,1,1) şeklinde ifade edilecektir.

TF yönteminde metinler içerdikleri terimlerin frekanslarıyla ifade edilmektedir. Bu yönteme göre satırlarında metinlerin, sütunlarında terimlerin yer aldığı bir matris elde edilir. Matrisin [i, j] gözünde i. metinde j. kelimenin kaç kere geçtiği bilgisi tutulur. Matrisin satır

sayısı metin sayına, sütun sayısı ise tüm metinlerdeki terim sayısına eşittir. Bu yöntemde metinde daha sık geçen ifadelerin daha önemli olduğu düşünülür [103].

$$TF_{ij} = \frac{n_{ij}}{|d_i|} \tag{17}$$

d_i: i. doküman

n_{ij}: d_i dokümanındaki w_j kelimesinin geçme sıklığı

TF_{ij}: terim sıklığı

TDF yönteminde ise metinde az sayıda geçen kelimenin ayırt edici özelliğinin diğer kelimelere göre daha fazla olduğu düşünülmektedir.

$$TDF_j = \log \frac{n}{n_j} \tag{18}$$

n_j: w_j kelimenin geçtiği doküman sayısı

n: tüm doküman sayısı

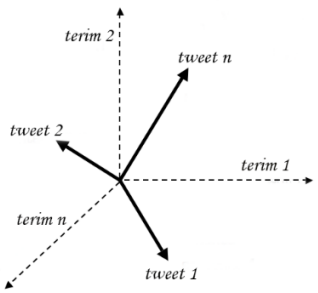
TDF_j: w_j kelimesinin ters doküman frekansı

TF-TDF yönteminde bir dokümanda bulunan ancak diğer dokümanlarda daha az görülen bir terimin ağırlığının daha fazla olacağı düşünülmektedir.

$$x_{ij} = TF_{ij}.TDF_j \tag{19}$$

x_{ij}: w_j kelimesinin d_i dokümanındaki TF-TDF ağırlığı

Bu yöntemlerle yapılan ağırlıklandırma işleminden sonra ön işleme süreci tamamlanır ve ön işlemden sonra öz nitelikler birer vektör olarak temsil edilir [104]. Bu durum Şekil 15'de gözükmektedir.



Şekil 15. Vektör uzay modeli

Öznitelik uzayının büyük boyutlara ulaşması metinden anlamlı bilgi çıkarmayı zorlaştırmıştır. Bu sorunu aşmak için genellikle öznitelik seçme ve öznitelik çıkarma yaklaşımları kullanılır [105].

1.4.8.10. N-gram Modeli

N-gram Modeller olasılık ve istatistiksel doğal dil işleme gibi belirli dizilimlerin olasılıklarını inceleyerek modelleyen alanlarda kullanılmaktadır. N-gram Modeli, önceki n elemanlı sıralamanın olma olasılığı bilindiğinde sıradaki olayın olma olasılığını tahmin etmeye çalışmaktadır. N-gram Modeli doğal dil işleme kullanıldığı zaman n-1. sıradan daha önceki kelimeler ile bağımsızlık varsayımı uygulanır. Kelimenin olma olasılığı sadece kendinden önceki n-1 kelimeye bağlıdır. Tablo 4'te örnek metinler için N-gram Modelleri sunulmuştur.

Tablo 4. Örnek metindeki n-gramlar

Metin	"Okuldan sonra sinemaya gitti. Eve gelmedi."
Unigramlar	"okuladan", "sonra", "sinemaya", "gitti", "eve", "gelmedi"
Bigram	"okuldan sonra", "sonra sinemaya", "sinemaya gitti", "gitti eve", "eve gelmedi"
Trigramlar	"okuldan sonra sinemaya", "sonra sinemaya gitti", "sinemaya gitti eve", "gitti eve gelmedi"
N-gramlar (n=4)	"okuldan sonra sinemaya gitti", "sonra sinemaya gitti eve", "sinemaya gitti eve gelmedi"

Kelime N-gramlarını açıklamak gerekirse, unigram modeli kendisinden önce 0 kelime sırasına bağlıdır. Bigram modeli ise kendinden önceki 1 kelimeye bağlı, trigram modelinde kelime kendinden önceki son 2 kelime sırasına bağlıdır. N-gram Modelleri konuşma tanıma problemlerinde harf sıralamasının tahmini için kullanılmaktadır [106].

Doğal dil işleme çalışmalarında sözde çanta modeli çok sık kullanılmaktadır. Fakat bu model metinleri sırasız ve gramer kuralları olmaksızın incelemektedir. Bu durumda analizde bilgi kaybı yüksek olacaktır. Kelimeler tek tek ele alındıklarında yeterli bilgiyi

taşımayabilirler. N-gramlar kelimeleri bileşik kelime ve deyimler şeklinde incelemeye olanak verdiği için daha fazla bilgi kazancı sağlamaktadır.

Duygu analizinde N-gramlar, duyguyu taşıyan kelimeleri sıralı olarak elde etmemizi sağlayan metot olarak kullanılabilir. Bu aşamada tüm n-gramlar elde edilir ve dokümandaki ağırlıkları TF ve TDF ile hesaplanıp verimiz için uygun öznitelikler belirlenebilir [107].

1.5. Türkçenin Yapısı

Türkçe Ural-Altay dil grubuna ait ve sondan eklemeli bir dildir. Her kelime bir kök ve bu kökün sonuna eklenen eklerden oluşmaktadır. Kelimeyi oluşturan köke gelen ekler yapım ve çekim ekleri olarak iki ayrılır. Yapım ekleri eklendiği kökün hem anlamını hem de türünü değiştirirler. Çekim ekleri ise kelimenin anlamını değiştirmez fakat türünü değiştirebilirler. Türkçe kendiliğinden gelişmiş bir dilden çok akademik topluluk tarafından oluşturulmuş gibi kural tabanlı bir dildir [108].

1.5.1. Türkçenin Ünlüleri / Vokalleri

Ünlüler, telaffuzları sırasında ses tellerinin titreştirildiği, ses yolunun herhangi bir noktasında kapanmanın ya da engelin bulunmadığı sesledir.

Tükçede ünlüler; tek başına telaffuz edilebilirler, hece, (o zamiri gibi) sözcük ya da ek oluşturabilirler, telaffuzları sırasında herhangi bir kapanma söz konusu olmadığı için uzatılabilirler. Standart Türkiye Türkçesinde kapalı é dâhil dokuz ünlü vardır. Ancak é sesi standart konuşma dilinde olmasına rağmen işaretlenmemiştir. Türkiye Türkçesindeki ünlüler ve özellikleri aşağıdaki gibidir: [109].

Tablo 5. Türkiye Türkçesindeki ünlüler ve özellikleri [109]

		Dilin Durumuna Göre			
		İnce (Ön) Ünlüler		Kalın (Art) Ünlüler	
Ağız Açıklığına Göre		Dar	Geniş	Dar	Geniş
Dudakların Durumuna Göre	Düz	i	e	ı	a
	Geniş	ü	ö	u	o

Ünlüler ve Türkiye Türkçesinin ünlülerle ilgili özelliklerinden bazıları şunlardır:

- Türkiye Türkçesinde sözcüklerde kalınlık incelik uyumu vardır.
- Türkiye Türkçesinde sınırlı sayıda örneğin dışında düzlük-yuvarlaklık uyumu vardır.
- Türkiye Türkçesinde ünlü-ünsüz uyumu vardır.
- Eski Türkçe Dönemi'nde; Çuvaşça, Yeni Uygurca, Kırgızca, Türkmence gibi çağdaş lehçelerde ve Türkiye Türkçesinin ağızlarında uzun ünlü vardır.
- Türkiye Türkçesinde uzun ünlüyle ya:d, ya:rın gibi sınırlı sayıda kelimede karşılaşılır. Bu da genellenebilecek bir nitelik arz etmez. Dolayısıyla Türkiye Türkçesinde uzun ünlü yoktur denilebilir.
- Türkiye Türkçesindeki uzun ünlülü sözcükler genellikle Arapça ve Farsça kökenlidir.
- Teşekkülleri sırasında ses telleri titreştiği için bütün ünlüler tonlu; küçük dil burun yolunu kapattığı içinde ağız sesleridir.
- Burun, karın, boyun gibi ikinci hecesi dar ünlülü sözcükler, üzerlerine ek aldıklarında vurgusuz orta hecenin dar ünlüsü düşer: karın > karnım : ı>ø; boyun > boynum: u>ø gibi.
- Türkçedeki en zayıf ünlü ı'dır.
- Standart konuşma dilinde kapalı é olmasına rağmen standart alfabede işaretlenmemiştir.
- Diğer dünya dillerinde olduğu gibi Türkçede de sözcüklerde bilgi yükü büyük oranda ünsüzler üzerindedir.
- Türkçe sözcüklerde uzun ünlü işaretlenmek istendiğinde iki nokta işareti ile (da:hil), karakter üzerine eklenen kısa çizgi ile (dâhil) veya düzeltme işaretiyle (dâhil) şeklinde gösterilir.
- Türkçede birleşik sözcüklerin dışında iki ünlü yan yana gelmez.

Yarı Ünlüler: Ünlüler gibi telaffuz edilen, ancak ünsüzler gibi kullanılan ğ, w, h, y gibi seslerdir [109].

1.5.2. Türkçenin Ünsüzler

Diyafram ve kaburga kaslarının baskısıyla akciğerden dışarı doğru gönderilen havanın ses yolunun herhangi bir noktasının daraltılması ya da kapatılmasıyla şekillendirilen seslerdir. 1 Kasım 1928 tarihindeki 1353 sayılı Kanun'la kabul ettiğimiz Latin kökenli Türk Alfabesinde 21 ünsüz işaretlenmiştir. Ancak bu hareketle Standart Türkiye Türkçesinde 21 standart ünsüz vardır demek mümkün değildir. Zira Latin kökenli Türk Alfabesiyle yazılan yazılarda bütün ünlüler yazıldığında / işaretlendiğinden kalın ķ ve ince k, kalın ţ ve ince t, kalın ş ve ince s, kalın ý ve ince y gibi hem kalın hem de ince şekli olan ünsüzler tek karakter ile gösterilmekte; bu da yazıda ünsüzlerle ilgili karakter sayısını azaltmaktadır.

Türkçede ünsüzler;

- Boğumlanma / teşekkül noktalarına göre,
- Tonlu ya da tonsuz olmalarına (Ses tellerini titreştirip titreştirmemelerine) göre,
- Patlayıcı / süreksiz ya da sızıcı / sürekli olmalarına göre,
- Ağız ya da geniz ünsüzü olmalarına göre sınıflandırılır.

Tablo 6. Standart Türkiye Türkçesinde kullanılan ünsüzler [109]

	Tonlu			Tonsuz	
	Sürekli		Süreksiz	Sürekli	Süreksiz
	Akıcı	Sızıcı	Patlayıcı	Sızıcı	Patlayıcı
Çift-dudak	m (geniz)	-	b	-	p
Diş-dudak	-	v	-	f	-
Diş	n (geniz)	z	d	s	t
Diş-damak	-	j	c	ş	ç
Ön damak	l, r, y		g	-	k
Art damak		ğ	ġ	-	ķ
Gırtlak				h	
Yarı ünlü	y	ğ		h	

Türkçenin ünsüzlerle ilgili bazı özellikleri şöyledir:

- Köktürk Alfabesi'nde f, h, j, v, ğ sesleri işaretlenmemiştir.
- Türkiye Türkçesinde içinde j bulunan sözcükler Türkçe değildir.
- Türkiye Türkçesinde c,ğ, l, m, n, r, z sesleriyle sözcük başlamaz.

- Türkiye Türkçesinde b, c, d, g sesleri genellikle sözcüklerin sonunda bulunmaz.
- Türkçede çift ünsüz ile hece başlamaz.
- Türkçede şedde yoktur.
- Türkiye Türkçesinde ünsüz uyumu vardır.
- Türkçede sözcük kökünde ya da hece sonunda üç ünsüz yan yana bulunmaz [109].

1.5.3. Türkçede Hece Yapısı

Hece, bir nefes hamlesi içinde çıkan, tek bir ses veya ses grubundan oluşan yalnız başına kelime olabilen veya kelime oluşumunda yer alabilen ses birliğidir [110].

Hecelerin yapısı, vocal (V) (ünlü) ve consonant (C) (ünsüz) harfleri kullanılarak gösterilir. Ünlü ile biten hecelere açık hece, ünsüz ile veya uzun ünlü ile biten hecelere ise kapalı hece adı verilir. Açık hece (.) işareti ile, kapalı hece (-) işareti ile gösterilir. Türkçede de diğer dünya dillerinde olduğu gibi en çok karşılaşılan hece tipi (CV= ünsüz + ünlü) yapısındaki hecedir. Her dilde ses birimlerin hece içindeki sıralanışında belirli sınırlamalar vardır. Örneğin CC, CV, CVC, V heceleri her dilde yoktur. Bazı dillerde de heceler ünsüzle bitemez, Arapçada olduğu gibi bazı dillerde hece yalnızca ünsüzlerle başlayabilir [111].

Dilin geriye çekilmesiyle ortaya çıkan seslere kalın sesler denir. a, ı, o, u sesleri kalın seslilerdir. Dilin ileriye hareketiyle ortaya çıkan seslere ise ince sesler denir. e, i, ö, ü sesleri ince seslilerdir.

Tablo 7. Türkçedeki hece tipleri [109]

V (ünlü)	o,
CV (ünsüz + ünlü)	su, şu, bu
VC (ünlü + ünsüz)	us, at, ak
CVC (ünsüz + ünlü + ünsüz)	süt, tip, kök
VCC (ünlü + ünsüz + ünsüz)	ast, üst, alt
CVCC (ünsüz + ünlü + ünsüz + ünsüz)	Türk, kırk

Tabloda da görüldüğü gibi Türkçede çift ünsüzle başlayan bir hece tipi bulunmamaktadır [109].

1.5.4. Türkçede Ses Uyumları

Türkçede ünlülerle ilgili olarak kalınlık-incelik, düzlük-yuvarlaklık ve düzlük; ünsüzlerle ilgili olarak tonluluk – tonsuzluk uyumları vardır [109].

1.5.4.1. Ünlü Uyumları

Türkçede ünlülerin özellikleri üzerine düşen bilgi yükünün oranına göre kalınlık-incelik, düzlük ve düzlük-yuvarlaklık uyumu vardır. Ancak Türkçede ünlülerin genişlik darlık özelliklerine düşen bilgi yükü diğer özellikleri üzerine düşen bilgi yükünden daha az olduğu için genişlik-darlık uyumu (örnekleri olsa da) yağın değildir [109].

1.5.4.2. Kalınlık-İncelik Uyumu

Türkçenin en önemli ses özelliklerinden biri kalınlık-incelik ya da artlık-önlük uyumudur. Bu uyuma göre Türkçe sözcük kökündeki birinci hecenin ünlüsü sonraki hecelerin ünlüsünü belirler. Yani sözcüğün birinci hecesinde kalın bir ünlü varsa o heceyi takip eden hecelerin ünlüsü de kalın; ince bir ünlü varsa takip eden hecelerin ünlüsü de ince olur [109].

- Türkçede /-yor/, /-ken/, /+mtrak/, /+Taş/, /+ki/, /+leyin/, /+gen/ gibi ekler farklı nedenlere bağlı olarak tek şekilli olduklarından kalınlık incelik uyumunu bozarlar. Türkçe sözcüklerde bir önceki hecenin ünlüsü daima bir sonraki hecenin ünlüsünü belirlediği için bu eklerin ünlüleri de üzerlerine gelen hecenin ünlüsünü kendisine benzer:

gel–iyor+**u**m

altı+gen+de

kapının+k**i**+n**i**

meslek+t**a**ş+**ı**m

- Bazı yabancı kökenli sözcüklerin son hecesinde kalın sıradan ünlü bulunmasına rağmen sözcüğün sonunda yer alan ince ünsüzden dolayı sözcük üzerine gelen eklerin ünlüleri ince olur:

h**â**l+den

kontrol+den

iştirak+**i**

- Yabancı dillerden gelen sözcüklerin üzerlerine Türkçe ek getirilirken "Türkçe sözcüklerde bir önceki hecenin ünlüsü daima bir sonraki hecenin ünlüsünü belirler." Kuralından hareketle sözcüğün son hecesindeki ünlü esas alınır. Son hecedeki ünlü kalınsa sözcük üzerine getirilecek ekin ünlüsü kalın; ince ise de ekin ünlüsü ince olur [109]:

beraber+ce

otomob**i**l+**i**

kuaf**ör**+**ü**

telefon+**a**

1.5.4.3. Düzlük–Yuvarlaklık Uyumu

Düzlük-yuvarlaklık uyumu, kalınlık-incelik uyumu çerçevesinde bir sözcüğün ilk hecesinde düz bir ünlü (a, e, ı, i) varsa, takip eden hecelerin ünlülerinin düz (a, e, ı, i) yuvarlak bir ünlü (o, ö, u, ü) varsa takip eden hecelerin ünlülerinin ya düz-geniş (a, e) ya da dar-yuvarlak (u, ü) olmasını gerektirir [109].

Düzlük-yuvarlaklık uyumu, uyumdan ziyade bir özelliktir. Zira Türkçede geniş yuvarlak ünlülerle ilgili tam bir uyum olsaydı Moğolcada (doloo (Mo.): yedi) ve Kırgızcada (kol+dor(Kr.): kollar) olduğu gibi birinci hece dışında da geniş yuvarlak ünlü (o, ö) olabilirdi. Ancak Türkiye Türkçesinde geniş yuvarlak ünlüler sadece ilk hecede bulunabilmektedir [109].

Tablo 8. Düzlük-yuvarlaklık uyumuna göre bir sözcüğün birinci ve diğer hecelerinde bulunabilecek ünlüler [109]

I. hecede bulunabilecek ünlüler	II. ve diğer hecelerde bulunabilecek ünlüler	Örnek
o,u	a, u	soluk, sulu, uçak, solak
ö, ü	e, ü	bölüm, önce, üçüncü

1.5.4.4. Düzlük Uyumu

Büyük ünlü uyumuna uymak koşuluyla bir sözcüğün ilk hecesinde düz bir ünlü varsa takip eden diğer hecelerde de düz bir ünlünün kullanılmasıdır:

Tablo 9. Düzlük uyumuna göre bir sözcüğün birinci ve diğer hecelerinde bulunabilecek ünlüler [109]

I. hecede bulunabilecek ünlüler	II. ve diğer hecelerde bulunabilecek ünlüler	Örnek
a, ı	a, ı	kapı, yapı, ışık, ırak
e, i	e, i	ekin, elek, iye, iyi

b, m, p çift dudak ve v, f diş-dudak ünsüzlerinin yanlarındaki ünlüleri yuvarlaklaştırma özelliğine sahip olmaları nedeniyle günümüzde yağmur, çamur, hamur, kavun, avuç, kabuk, tapu, avun-, kavur-, savur-, gibi sözcükler düzlük uyumuna uymazlar [109].

1.5.4.5. Ünlü–Ünsüz Uyumu

Türkçe sözcüklerde ağzın arka kısmında / yumuşak damakta teşekkül eden g, k gibi kalın ünsüzler ile ince ünlüler; ağzın ön kısmında teşekkül eden ince ünsüzlerle de ağzın arka kısmında teşekkül eden kalın ünlüler aynı hece içinde bulunamazlar. Bu uyuma uymayan sözcükler yabancı kökenlidir [109]:

ķriz: ***ķ*** kalın; ***i*** ince

ġrev: ġ kalın; *e* ince

ķredi: ***ķ*** kalın; *e* ince

inkılap: ***i*** ince; ***ķ*** kalın

1.5.4.6. Ünsüz – Ünlü Uyumları

Standart Türkiye Türkçesinde tonsuz ünsüzlerle (f, s, t, k, ç, ş, h, p) biten sözcüklere tonlu ünsüzle başlayan herhangi bir ek getirildiğinde ekin ünsüzü sözcüğün ötümsüz olan son sesinin ilerleyici etkisiyle (ötümsüz karşılığın varsa) ötümsüzleşir [109]:

fd > ft: sahaf+da > sahaf+ta

sd > st: kafes+de > kafes+te

td > tt: süt+de > süt+te

kd > kt: sokak+da > sokak+ta

çd > çt: ağaç+da > ağaç+ta

şd > şt: kış+da > kış+ta

hc > hç: sabah+cı > sabah+çı

pc > pç: kitap+cı > kitap+çı…

1.5.4.7. Fillerde Olumluluk, Olumsuzluk, Genellik

Alyılmaz [112], yaptığı çalışmada Orhun Yazıtlarının Söz Dizimi'ni incelemiştir ve Standart Türkiye Türkçesi için örnekler vermiştir.

1.5.4.8. Fillerde Olumluluk

Yazıtlarda fillerin olumlu şekillerini gösteren özel bir eke rastlanmamaktadır. Yani fiillerin olumsuz şekilleri karşısında olumlu şekilleri işaretsiz (Ø) dir. Olumlu fiil yapılarını olumsuzluk eki almayan ve olumsuzluk anlamı taşımayan fiiller oluştururlar [112].

süle-Ø- : asker sevk et-, ordu yürüt-, sefer et-

olur-Ø- : tahta otur-, tahta çık-

1.5.4.9. Fiillerde Olumsuzluk

Fillerde Olumsuzluk: Yazıtlarda olumsuzluk bildiren fiillersin sayısı oldukça fazladır. Bu fiiller, iki grupta incelendi:

1- Olumsuzluk ögesi (eki-edatı) alarak olumsuzluk bildiren fiiller
2- Herhangi bir olumsuzluk eki almadan olumsuzluk bildiren (bünyesinde olumsuzluk anlamı taşıyan fiiller [112].

1.5.4.10. Olumsuzluğun Morfolojik Usulle İşaretlenmesi

Olumlu anlama sahip herhangi bir fiilin olumsuzluk eki /-mA-/ ile geçici olarak olumsuz yapılması şeklinde gerçekleşir.

Fiillerin olumsuz şekilleri genelde fiil kök veya gövdelerine /-mA-/ olumsuzluk eki getirilerek yapılmıştır.

yorıt-ma- : yürütme-, ilerletme-

(e)dgü bilge kişig (e)dgü (a)lp kişig yor(ı)tm(a)z (e)rm(i)ş: İyi (ve) akıllı kişileri, iyi (ve) cesur kişileri ilerletmezler imiş.

kal-ma- : kalma-

türk sir bod(u)n y(e)rinte bod k(a)lm(a)tı: Türk-Sir halkının yerinde boy (kabile) kalmadı.

Yazıtlarda olumsuzluğun ifadesinde bazan da /yok/ kelimesinin kullanıldığı görülür.

ol (a)mtı (a)ny(ı)g yok [erür]: Onlar şimdi (hiç de) kötü (durumda) değiller.

Ancak bu kullanımın fazla işlek olmadığını ve yok kelimesinin genelde birleşik bir fiilin isim kısmını oluşturduğunu da hemen kaydetmek gerekir.

türk bod(u)n ölti (a)lk(ı)ntı yok boltı: Türk halkı öldü, bitti, yok oldu [112].

1.5.4.11. Olumsuzluğun Semantik Usulle İşaretlenmesi

Olumsuzluğun bünyesinde olumsuz anlam taşıyan fillerle işaretlenmesidir.

Bu türden fiillerin sayısı oldukça fazladır.

öl- : öl-

(e)kinti ışb(a)ra y(a)mt(a)r boz (a)t(ı)g bin(i)p t(e)gdi ol (a)t (a)nta ölti: ikinci olarak İşbara Yamtar'ın boz atına binip hücum etti. O at (da) orada öldü.

alkın- : azal-, tüken-, mahvol-

arıl- : azal-, tüken-, mahvol-

(a)nt(a)g(ı)ñ(ı)n üçün ig(i)dm(i)ş k(a)g(a)n(ı)ñ(ı)n s(a)bın (a)lm(a)tın yir s(a)yu b(a)rd(ı)g koop (a)nta (a)lk(ı)nt(ı)g (a)r(ı)lt(ı)g: Böyle olduğun için (seni) beslemiş olan kağanının sözünü dinlemeden (beklemeden) her yere gittin (ve) oralarda hep mahvoldun tükendin.

ilsire- : devletsiz kal-, devletsiz ol-

kagansıra- : kağansız kal-, kağansız ol- küñed- : cariye ol-

kulad- : köle ol-

y(i)ti yüz (e)r bol(u)p (i)ls(i)r(e)m(i)ş k(a)g(a)ns(ı)r(a)m(ı)ş bod(u)n(u)g küñ(e)dm(i)ş kuul(a)dm(ı)ş bod(u)n(u)g türük törüsin ıçg(ı)nm(ı)ş bod(u)n(u)g (e)çüm (a)pam törüsinçe y(a)r(a)tm(ı)ş boşgurm(ı)ş: Yedi yüz kişi olup devletsiz kalmış, kağansız kalmış milleti, cariye olmuş, kul olmuş milleti, Türk örf ve âdetlerini bırakmış milleti, atalarımın dedelerimin töresince (yeniden) yaratmış eğitmiş (yetiştirmiş) [112].

1.6. Modelin Başarım Ölçütleri

Model başarımını değerlendirirken kullanılan temel kavramlar hata oranı, kesinlik duyarlılık ve F-ölçütüdür. Modelin başarısı, doğru sınıfa atanan örnek sayısı ve yanlış sınıfa atanan örnek sayısı nicelikleriyle alakalıdır.

Test sonucunda ulaşılan başarım bilgileri karışıklık matrisi ile ifade edilir. Karışıklık matrisinde satırlar test kümesindeki örneklere ait gerçek sayıları, kolonlar ise modelin tahminini ifade etmektedir. Tablo 10'da örnek bir karışıklık matrisi görülmektedir.

Tablo 10. Karışıklık matrisi

		Öngörülen Sınıf	
		Sınıf=1	Sınıf=0
Doğru Sınıf	Sınıf=1	a	b
	Sınıf=0	c	d

a: TP (True Pozitif) c: FP (False Pozitif)

b: FN (False Negatif) d: TN (True Negatif)

1.6.1. Doğruluk ve Hata Oranı

Model başarımı ölçülmesinde kullanılan en popüler ve basit yöntem, modele ait doğruluk oranıdır. Doğruluk oranı, doğru sınıflandırılmış örnek sayısının (TP + TN), toplam örnek sayısına (TP+ TN+ FP+ FN) oranıdır. Hata oranı ise bu değerin 1'tamlayanıdır. Diğer bir ifadeyle yanlış sınıflandırılmış örnek sayısının (FP +FN), toplam örnek sayısına (TP+ TN+ FP+ FN) oranıdır.

$$Doğruluk = \frac{TP + TN}{TP + FP + FN + TN} \tag{20}$$

$$Hata\ Oranı = \frac{FP + FN}{TP + FP + FN + TN} \tag{21}$$

1.6.2. Kesinlik

Kesinlik, sınıfı 1 olarak tahmin edilmiş True Pozitif örnek sayısının, sınıfı 1 olarak tahmin edilmiş tüm örnek sayısına oranıdır.

$$Kesinlik = \frac{TP}{TP + FP} \tag{22}$$

1.6.3. Duyarlılık

Doğru sınıflandırılmış pozitif örnek sayısının toplam pozitif örnek sayısına oranıdır.

$$Duyarlılık = \frac{TP}{TP + FN} \tag{23}$$

1.6.4. F - Ölçeği

F – ölçeği, kesinlik ve duyarlılığın harmonik ortalamasıdır. Kesinlik ve duyarlılık ölçütleri tek başına anlamlı bir karşılaştırma sonucu çıkarmamız için yeterli değildir. Her iki ölçütü beraber değerlendirmek daha doğru sonuçlar verir. Bunun için f – ölçeği tanımlanmıştır [113].

$$F - \text{Ölçeği} = \frac{2 \times Duyarlılık \times Kesinlik}{Duyarlılık + Kesinlik} \tag{24}$$

1.6.5. ROC (Receiver Operating Characteristics) Eğrisi

Her sınıflandırma işleminde metotlar, kesinlik (yanlış pozitif eleme kabiliyeti) ve hassasiyet (doğru pozitifleri tespit etme kabiliyeti) arasındaki dengeyi kurmakla uğraşmaktadır. Veri kümesindeki pozitif ve negatif örnekler, eşit bir şekilde dağılım göstermediğinden dolayı, kesinlik ve hassasiyet ölçütlerinden önce, kesinlik ve hassasiyet arasındaki dengeyi değerlendirmek için kullanılmıştır. ROC eğrisi altında kalan ROC puanı olarak adlandırılmaktadır. ROC eğrisi değişen sınıflandırma eşik değerine göre doğru pozitiflerin sayısının, yanlış pozitiflerin bir fonksiyonu olarak çizilmesiyle oluşmaktadır. ROC puanı 1 olduğunda anlamı, pozitifler mükemmel bir şekilde ve negatiflerden

ayrılmıştır, olmaktadır. ROC puanı 0 olduğunda ise herhangi bir pozitif bulunamadı anlamına gelmektedir [114].

Bu çalışmada sınıflandırma performansları değerlendirilirken ROC puanı yani, Eğri Altındaki Alan (EAA) kullanılmıştır.

1.7. Öznitelik Seçimi

Metin sınıflandırma algoritmasının performansını iyileştirmek, etkin ve ölçeklenebilir bir sınıflandırma modeli oluşturmak için gerekli en temel aşamadır. Metin sınıflandırmada görülen yüksek boyutluluk problemini ortadan kaldırmak için öznitelik yöntemleri sıklıkla kullanılmaktadır.

Öznitelik seçimi ile veri setinden uygun bir öznitelik altkümesi elde edilir. Böylelikle, hem doğru sınıflandırma oranı bakımından hem de ölçeklenebilirlik bakımından daha iyi bir sınıflandırma modeli elde edilmiş olur.

Öznitelik seçim yöntemleri, öznitelik altkümesi değerlendirmede kullanılan stratejiye dayalı olarak filtre tabanlı ve sarmalama tabanlı olmak üzere iki temel sınıf altında incelenmektedir. Filtre – tabanlı öznitelik seçim yöntemlerinde, belirli bir öznitelik altkümesinin yararlılığı, belirli bir sezgisel arama aracılığıyla belirlenmektedir. Sarmalama – tabanlı öznitelik seçim yöntemlerinde ise belirli bir sınıflandırma algoritmasının başarımına dayalı olarak öz nitelikler belirlenmektedir.

Bu çalışmada sarmalama tabanlı öznitelik seçim yöntemi ve sezgisel arama algoritmalarından ileri doğru seçim yöntemi kullanılmıştır [115].

2. YAPILAN ÇALIŞMALAR

Son yıllarda DA ile ilgili yapılan çalışmaların birçoğu İngilizce için yapılmıştır. Bu çalışmadaki amaç sosyal medyadaki Türkçe mesajların duygu durumlarını tespit ederek otomatik değerlendirecek bir sistemin geliştirilmesidir. Mesajlardaki duygu durumlarının tespit edilmesi işletmelerin, siyasi partilerin ve kuruluşların reklam ve kampanya gibi kitleleri etkileyen olaylara insanların verdikleri tepkileri belirlenmesine katkı sağlayacaktır. Örneğin, çalışma [116], duygu çarkında "sıkıntı" durumunun, göz ardı edilmesi durumunda zamanla nefrete dönüşeceğini ortaya koymaktadır.

Bu çalışmayla firmalar, müşterilerinin sosyal medya üzerinde paylaştığı mesajların duygu durumunu kolayca tespit edebilecek ve müşterilerin sorununu çözüp firmadan nefret etmesinin önüne geçebilecektir. Bunun yanı sıra firmalar sosyal medya mesajları üzerinden analizler yaparak firma durumunu ve kamuoyu tepkilerini kolayca belirleyebilecektir.

Klasik duygu analizinde olumlu veya olumsuz sınıflandırma yapılırken, bu çalışmada klasik yaklaşımın yanı sıra mesajlardaki esas duygu durumuna göre de sınıflandırma yapılmıştır. Çalışma klasik duygu analizinden bu yönüyle farklılaşmaktadır. Çalışmada kullanılan duygu kategorileri [59] ve [116] baz alınarak belirlenmiştir

Bu çalışmada duygu kategorileri Tablo 9 ve Tablo 10'daki gibi belirlenmiştir. Bu duygular Karadeniz Teknik Üniversitesi Edebiyat Fakültesi Türk Dili ve Edebiyatı bölümünde eğitim gören 243 öğrenciye yapılan anket uygulaması sonucunda belirlenmiştir. İlk çalışmada 23 olan duygu durumu yapılan anket çalışması sonucunda 10 sınıf olarak belirlenmiştir. Anket uygulamasında yakın anlamlı ifadeler kullanılmış daha sonra alınan geri bildirimler ile Tablo 11 ve Tablo 12'deki olumlu ve olumsuz duygular listesi belirlenmiştir. Metinlerin duygulara göre sınıflandırılmasında bu listedeki duygular kullanılmıştır.

Listede bulunan diğer kategorisi duygu durumunu temsil etmeyip anlamsız ve farklı dillerdeki metinlerin işaretlenmesi için kullanılmaktadır. Bu kategori oluşturulan modelde anlamsız ve farklı dildeki metinlerin otomatik olarak sınıflanmasını ve analiz dışı tutulmasını sağlamaktadır.

Tablo 11. Duygu analizinde kullanılan olumlu duygu sınıfları

Mutluluk ve Neşe	Güvenmek	Takdir Etmek
Gurur Duyma	Beklenti	Tavsiye

Tablo 12. Duygu analizinde kullanılan olumsuz duygu sınıfları

Aşağılama	Merak
Hayal Kırıklığı	Öfke

Bu tez kapsamında Türkçenin yapısına uygun önişleme, öznitelik belirlenme tekniklerinin tespit edilmesi ve sosyal medyadaki duygu durumlarının detaylı bir şekilde belirlenmesi amaçlanmıştır.

2.1. Veri Elde Etme

Bu aşama, hangi tür verilerin kullanılacağı ve elde edilmesi için gereken işlemlerin belirlendiği aşamadır. Bu çalışmada JAVA programlama dili ve Twitter API'si kullanılarak "Turkcell" firmasına ait olan 2100 mesaj, veri tabanına kaydedilmiştir. Veri tabanına kaydedilen mesajlar üzerinde herhangi bir ön işleme veya eleme yapılmamıştır. Türkçe haricindeki dillerde yazılan mesajlar da veri tabanına kaydedilmiş, fakat analiz kısmında Türkçe haricindeki mesajlar analiz dışı tutulmuştur. Bu çalışma için "Turkcell" firmasına ait twitter mesajları 2016 yılının Temmuz ayı boyunca belirli aralıklarla kaydedilmiştir. Turkcell firması tarafından, 15 Temmuz gecesi abonelerine ücretsiz dakika ve internet paketi gönderilmiştir. Ayrıca 15 Temmuz sonrası, şehitlerimiz için "5 lira bağışınıza 5 lira bizden" kampanyası başlatılmıştır. Firmanın yapmış olduğu kapanmayaların, müşterilerin duygu durumlarını etkileyeceği düşünülmektedir.

Bir veri madenciliği modeli, veriye ilişkin ne kadar çok bilgiye sahip olursa performansı o derece artar. Bu ilkeyle Türkçe haricindeki mesajlar, "diğer" kategorisi olarak işaretlenmiş ve sistemin bu işaretlemeyi otomatik olarak yapması için eğitim aşamasında bu mesajlar kullanılmıştır. Sistemin devamlığı açısından bu işaretleme büyük katkı sağlamıştır.

2.2. Veri Ön İşleme

Veri ön işleme aşaması, veri madenciliğinin en önemli ve analiz sonucunda elde edilecek başarıyı doğrudan etkileyen aşamadır. Sosyal medya platformlarının kendine has özelliklerinin olması, bu ortamlarda kullanılan dili de etkilemiştir. Özellikle Twitter gibi sınırlı karakterle mesaj yazma imkânı veren mikro blog sitelerinin kendine özgü yazım dili ortaya çıkmıştır.

Elde edilen verilerden ve yapılan araştırmalar sonucunda sosyal medya ortamında en sık kullanılan kısaltmalar belirlenmiştir. Mevcut mesajlar içerisinde tespit edilen kısaltma ifadeleri, aynı mesaj içerisinde açık bir şekilde yazılarak sentetik mesajlar oluşturulmuştur. Örneğin; "Tşk Turkcell" şeklinde yazılmış bir mesaj, "Teşekkür Turkcell" gibi kısaltmaların açık hâlde yazıldığı bir mesaj olarak veri tabanına kaydedilmiştir.

Bu çalışmada 1990'lı yılların sonunda Japonya'da yaratılan küçük simgeler olan "emoji" karakterleri de sistemin içerisine dahil edilmiştir. İnsanlar küçük bir simgeyle birçok şey ifade edebiliyorlar bu durum DA için göz ardı edilemez. Örneğin, Twitter'da en çok kullanılan emoji "😂" sevinç göz yaşlarıdır. Bu emoji 2015 yılında Oxford English Dictionary tarafından 2015 yılının kelimesi seçilmiştir. Bu çalışmada, mesajlar içerisinde yer alan emojilerde kullanılmış ve nitelik olabilecek olanlar sınıflandırma adımında kullanılmıştır [117]. Ayrıca bu aşama metin mesajlarında kullanılan duygu karakterleri de emojiler gibi sisteme dahil edilmiştir.

Bu aşamada yazılan mesajlardaki harf tekrarları ortadan kaldırılmıştır. Kişinin harf tekrarını kullanmaktaki amacının duyguyu baskın bir şekilde ifade ettiği düşünülmektedir. Bundan dolayı vurguyu artırmak için harf tekrarının yapıldığı kelime, harf tekrar sayısı kadar mesaj içerisine eklenmiştir. Örneğin, "Türkiye'nin Turkceli teşekkürlerrr" şeklindeki mesaj "Türkiyenin Turkceli teşekkürler teşekkürler teşekkürler" şekilde sentetik bir mesaj olarak veri tabanına kaydedilir. Bu sayede "teşekkür" kelimesinin ağırlığı artırılmış ve bir nitelik olarak belirlenmesi sağlanmıştır.

Ön işleme aşamasında kelimenin Türkçe olup olmadığının tespit edilmesi ve bazı gerekli olan hecele adımlarında Zemberek Kütüphanesi kullanılmıştır.

2.3. Sayısallaştırma

Ön işlemeden sonra elde edilen düzenli veri, bu aşamada terimlere ayrılır. Her bir mesaj için terimler elde edilerek genel bir terimler sözlüğü oluşturulur. Bu aşamada veri üzerinde analizler yapabilmek için metin verisinin sayısal ifadelere çevrilmesi gerekmektedir. Her bir terim için terim frekansı ve doküman terim matrisi elde edilerek metin verilerinin analize hazır hâle gelmesi sağlanmıştır. 10 duygu durumu için öz nitelik olan kelimeler ve emojiler belirlenmiştir.

Bu çalışmada toplam 2100 doküman ve 6988 terim elde edilmiştir. Terim frekansı ve terim doküman matrisi elde edilerek analiz için veri hazırlanmıştır.

2.4. Analiz

Bu aşamada veri üzerinde veri madenciliği yaklaşımlarından olan sınıflandırma, kümeleme ve birliktelik analizi işlemleri yapılmaktadır. Sayısallaştırma aşamasında elde edilen terim frekansı ve doküman terim matrisleri kullanılarak metin sınıflandırma işlemi gerçekleştirilmiştir.

Sınıflandırma işlemi için makine öğrenmesi, veri madenciliği ve metin madenciliği amaçlarına yönelik geliştirilmiş olan Rapid Miner 7.5 yazılım platformu kullanılmıştır. Rapid Miner kullanılarak, Karar Ağaçları, Yalın Bayes, 1- En Yakın Komşu ve Destek Vektör Makineleri sınıflandırma algoritmalarının performansları karşılaştırılmıştır [118]. İleri doğru seçim yöntemiyle her duygu için öz nitelikler belirlenerek, algoritmaların performansları iyileştirilmeye çalışılmıştır.

3. SONUÇLAR

Bu bölümde belirlenen 10 duygu için sınıflandırma algoritmalarının, performansları karşılaştırılmış ve her duygu durumu için öznitelik olan kelimler belirlenmiştir. Değerlendirme yapılırken her sınıflandırma algoritmasının F- ölçeği, Doğruluğu ve EAA değerleri göz önünde bulundurulmuştur. Sınıflandırma performansının değerlendirmesinde F-ölçeği ve EAA değerleri dikkate alınmaktadır.

Kurulan modelin iyileştirilmesi için Türkçe'nin kurallı yapısının sosyal medya metinlerine özgün olacak şekilde geliştirilmesi gerekmektedir. Türkçe'nin yapısına uygun olarak sosyal medya metinlerini analiz etmek duygu analizi, fikir madenciliği ve düşünce manipülasyonu tespiti gibi birçok alanda başarılı bir şekilde uygulanabilir.

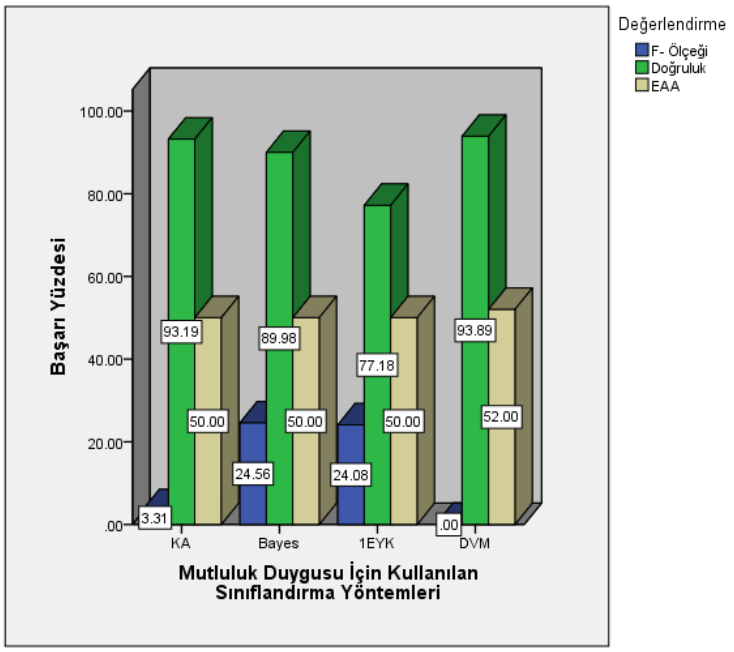
3.1. Mutlu / Neşeli Duygusu

Mutluluk duygusu için sınıflandırma sonuçları Tablo 13'deki gibi elde edilmiş ve ileri doğru seçme yöntemiyle elde edilen sonuçlar Tablo 14'deki gibi elde edilmiştir.

Tablo 13. Mutluluk duygusu için sınıflandırma sonuçları

Yöntem	F- Ölçeği	Doğruluk	EAA
KA	3,31	93,19	50,0
Bayes	**24,56**	89,98	50,0
1 EYK	24,08	77,18	50,0
DVM	-	**93,89**	**52,0**

Mutluluk duygusu için Tablo 13 incelendiğinde, KA için F-ölçeği %3,31, Doğruluğu %93,19 ve EAA'sı 0,50, Bayes için, F-ölçeği %24,56, Doğruluğu %89,98 ve EAA'sı 0,50, 1-En Yakın Komşu için, F-ölçeği %24,08, Doğruluğu %77,18 ve EAA'sı 0,50 ve DVM için F- ölçeği hesaplamamış, doğruluğu %93,89 ve EAA'sı 0,52 olarak elde edilmiştir. Şekil 16'da her sınıflandırma yöntemi için karşılaştırma yapılmış ve sonuçlar sunulmuştur.



Şekil 16. Mutluluk duygusu için kullanılan sınıflandırma yöntemleri ve başarımları

Mutluluk Duygusu için F-ölçeği en yüksek olan sınıflandırma algoritmasının Bayes sınıflandırıcısı, Doğruluk ve EAA değeri en yüksek olan sınıflandırma algoritmasının DVM olduğu görülmektedir.

Mutluluk Duygusu sınıflandırılmasında Bayes ve 1EYK algoritmalarının F-ölçeği değerlerinin birbirine yakın olduğu tespit edilmiştir. Dolayısıyla veri sayısı artırıldığında Mutluluk Duygusuna ait verilerinin yapısı gereği, Bayes ve 1EYK algoritmaları bu duygu için daha iyi sonuç verecektir.

Mutluluk Duygusu için ileri doğru seçim yöntemi ile etkin olan kelimeler tespit edilmiştir. Tablo 14'te her sınıflandırma yöntemi için elde edilen etkin kelimeler ve başarımlar sunulmuştur.

Tablo 14. Mutluluk duygusu için ileri doğru seçim yöntemi ile elde edilen kelimeler ve başarımları

	F Ölçeği	Doğruluk	EAA	Kelimeler
KA	39,26	**94,64**	64,7	canım, 😂 , mutluyum, 😀 , ☺
Bayes	**43,29**	94,18	**83,1**	turkcell, 😀 , 😂 , canım, teşekkürler, seni, gb, i
1EYK	8,49	88,71	50,0	turkcell
DVM	25,52	94,18	72,3	canım, 😀 , internet, iyiki, yarısı

Mutluluk Duygusu için ileri doğru seçme yöntemi ile Tablo 14'teki sonuçlar elde edilmiştir. KA için F-ölçeği %39,26, Doğruluğu %94,64 ve EAA'sı 0,64,7, Bayes için, F-ölçeği %43,29, Doğruluğu %94,18 ve EAA'sı 0,83, 1-En Yakın Komşu için, F-ölçeği %8,49, Doğruluğu %88,71 ve EAA'sı 0,50 ve DVM için F- ölçeği %25,52, Doğruluğu %94,18 ve EAA'sı 0,72 olarak elde edilmiştir. En yüksek F-ölçeğine ve EAA'ya sahip olan sınıflandırma yönteminin Bayes sınıflandırıcısı olduğu tespit edilmiştir. Mutluluk Duygusu için ileri doğru seçme yönteminin sınıflandırma performanslarını artırdığı görülmüştür.

Mutluluk Duygusuna ait etkin kelimeler ileri doğru seçme yöntemiyle tespit edilmiştir. Bu kelimelerin "canım, 😂, mutluyum, 😀, ☺, teşekkürler, i, turkcell, iyiki" olduğu tespit edilmiştir. Belirlenen bu kelimeler Mutluluk Duygusu için ayırt edici niteliklerdir. Bu nitelikler ile mutluluk duygusunu içeren mesajlar tespit edilmektedir.

Şekil 17'de Mutluluk Duygusuna ait kelimelerden oluşturulmuş kelime bulutu sunulmuştur.

Şekil 17. Mutluluk duygusu kelime bulutu

Turkcell firması için mutluluk duygusu, mesajların kayıt edildiği dönem göz önüne alınarak incelendiğinde, yapılan kampanyaların müşteri mutlu ettiği görülmüştür.

3.2. Takdir Duygusu

Takdir Duygusu için sınıflandırma sonuçları

Tablo 15'deki gibi elde edilmiş ve ileri doğru seçme yöntemiyle elde edilen sonuçlar Takdir Duygusu için elde edilen F-ölçeği sonuçları incelendiğinde Takdir Duygusuna ait verilerin Bayes ve 1EYK sınıflandırıcıları ile daha iyi sonuçlar verdiği tespit edilmiştir. Dolayısıyla Takdir Duygusuna ait veriler yapılarından dolayı, Bayes ve 1EYK komşu algoritmalarıyla daha iyi sonuçlar verecektir.

Takdir Duygusu için ileri doğru seçim yöntemi ile etkin olan kelimeler tespit edilmiştir. Tablo 16'da her sınıflandırma yöntemi için elde edilen etkin kelimeler ve başarımlar sunulmuştur.

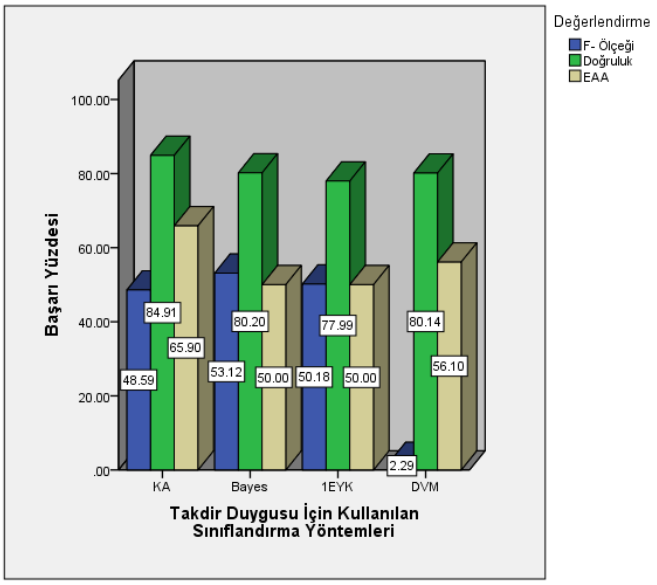
Tablo 16

Tablo 15. Takdir duygusu sınıflandırma sonuçları

Yöntem	F- Ölçeği	Doğruluk	EAA
KA	48,59	**84,91**	**65,9**
Bayes	**53,12**	80,20	50,0
1 EYK	50,18	77,99	50,0
DVM	2,29	80,14	56,1

Takdir Duygusu için

Tablo 15 incelendiğinde, KA için F-ölçeği %48,59, Doğruluğu %84,91 ve EAA'sı 0,65.9, Bayes için, F-ölçeği %53,12, Doğruluğu %80,20 ve EAA'sı 0,50, 1-En Yakın Komşu için, F-ölçeği %50,18, Doğruluğu %77,99 ve EAA'sı 0,50 ve DVM için F- ölçeği %2,29, doğruluğu %80,14 ve EAA'sı 0,56,1 olarak elde edilmiştir. Şekil 18'de her sınıflandırma yöntemi için karşılaştırma yapılmış ve sonuçlar sunulmuştur.



Şekil 18. Takdir duygusu için kullanılan sınıflandırma yöntemleri ve başarımları

Takdir Duygusu için F-ölçeği en yüksek olan sınıflandırma algoritmasının Bayes sınıflandırıcısı, Doğruluk ve EAA değeri en yüksek olan sınıflandırma algoritmasının KA sınıflandırıcısının olduğu görülmektedir.

Takdir Duygusu için elde edilen F-ölçeği sonuçları incelendiğinde Takdir Duygusuna ait verilerin Bayes ve 1EYK sınıflandırıcıları ile daha iyi sonuçlar verdiği tespit edilmiştir. Dolayısıyla Takdir Duygusuna ait veriler yapılarından dolayı, Bayes ve 1EYK komşu algoritmalarıyla daha iyi sonuçlar verecektir.

Takdir Duygusu için ileri doğru seçim yöntemi ile etkin olan kelimeler tespit edilmiştir. Tablo 16'da her sınıflandırma yöntemi için elde edilen etkin kelimeler ve başarımlar sunulmuştur.

Tablo 16. Takdir duygusu için ileri doğru seçim yöntemi ile elde edilen kelimeler ve başarımı

	F Ölçeği	Doğruluk	EAA	Kelimeler
KA	**65,00**	**88,53**	75,90	teşekkürler, helal, teşekkür, tebrikler, adamsın, bravo, 👏, davranış
Bayes	57,74	86,08	**83,10**	teşekkürler, helal, teşekkür, güzel, tebrikler,
1EYK	-	79,91	50,00	turkcell
DVM	57,91	86,95	76,50	teşekkürler, helal, teşekkür, 👏, tebrik, turkcell,

Takdir Duygusu için Tablo 16 incelendiğinde KA için F-ölçeği %65, Doğruluğu %88,53 ve EAA'sı 0,75,9, Bayes için, F-ölçeği %57,74, Doğruluğu %86,08 ve EAA'sı 0,83.1, 1-En Yakın Komşu için, F-ölçeği hesaplanamamış, Doğruluğu %79,91 ve EAA'sı 0,50 ve DVM için F- ölçeği %57,91, Doğruluğu %86,95 ve EAA'sı 0,76,5 olarak elde edilmiştir. Takdir Duygusu için en yüksek F-ölçeğine sahip olan sınıflandırma yönteminin KA sınıflandırıcısı olduğu tespit edilmiştir.

Takdir Duygusu için ileri doğru seçme yönteminin sınıflandırma performanslarını artırdığı görülmüştür. Ayrıca ileri doğru seçme yöntemi ile elde edilen kelimelerin "teşekkürler, helal, teşekkür, tebrikler, adamsın, bravo, 👏, turkcell" olduğu görülmüştür. Belirlenen bu kelimeler Takdir Duygusu için ayırt edici niteliklerdir. Bu nitelikler ile Takdir Duygusunu içeren mesajlar tespit edilmektedir. Şekil 19'da Takdir Duygusu için tespit edilen kelimelerden oluşturulan kelime bulutu sunulmuştur.

Şekil 19. Takdir duygusu kelime bulutu

Turkcell firması için takdir duygusu incelendiğinde, firmanın 15 Temmuz ve sonrasında yapmış olduğu kampanya ve açıklamaların insanlar tarafından takdir edildiği görülmüştür.

3.3. Dilek Duygusu

Dilek Duygusu için sınıflandırma sonuçları

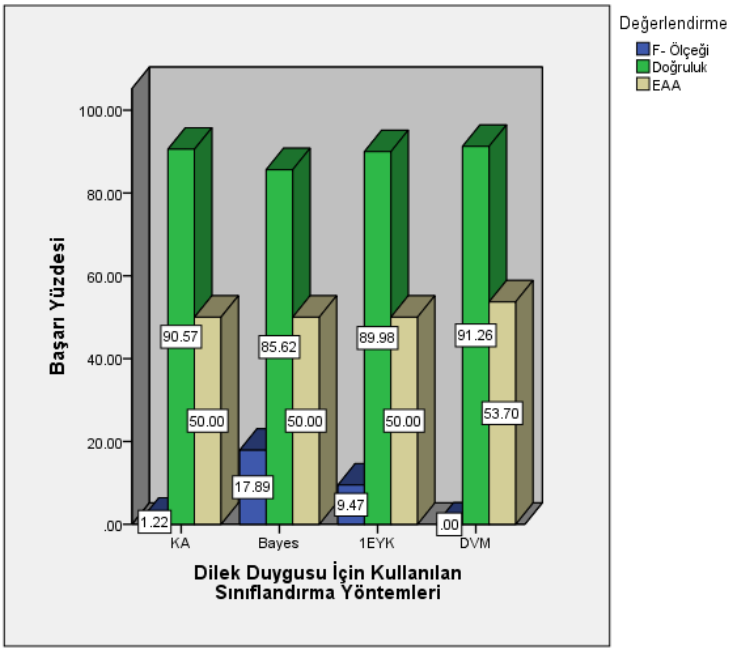
Tablo 17'deki gibi elde edilmiş ve ileri doğru seçme yöntemiyle elde edilen sonuçlar Tablo 18'deki gibi elde edilmiştir.

Tablo 17. Dilek duygusu sınıflandırma sonuçları

Yöntem	F- Ölçeği	Doğruluk	EAA
KA	1,22	90,57	50,0
Bayes	**17,89**	85,62	50,0
1 EYK	9,47	89,98	50,0
DVM	-	**91,26**	**53,7**

Dilek Duygusu için

Tablo 17 incelendiğinde, KA için F-ölçeği %1,22, Doğruluğu %90,57 ve EAA'sı 0,50, Bayes için, F-ölçeği %17,89, Doğruluğu %85,62 ve EAA'sı 0,50, 1-En Yakın Komşu için, F-ölçeği %9,47, Doğruluğu %89,98 ve EAA'sı 0,50 ve DVM için F- ölçeği hesaplanamamış, Doğruluğu %89,98 ve EAA'sı 0,50 olarak elde edilmiştir. Şekil 20'de her sınıflandırma yöntemi için karşılaştırma yapılmış ve sonuçlar sunulmuştur.



Şekil 20. Dilek duygusu için kullanılan sınıflandırma yöntemleri ve başarımları

Dilek Duygusu için F-ölçeği en yüksek olan sınıflandırma algoritmasının Bayes sınıflandırıcısı, Doğruluk ve EAA değeri en yüksek olan sınıflandırma algoritmasının DVM sınıflandırıcısının olduğu görülmektedir. Dilek Duygusuna ait verilerin yapısının Bayes sınıflandırıcısı ile daha iyi sonuç verdiği tespit edilmiştir.

Dilek Duygusu için ileri doğru seçim yöntemi ile etkin olan kelimeler tespit edilmiştir. Tablo 18'de her sınıflandırma yöntemi için elde edilen etkin kelimeler ve başarımlar sunulmuştur.

Tablo 18. Dilek duygusu için ileri doğru seçim yöntemi ile elde edilen kelimeler ve başarımları

	F Ölçeği	Doğruluk	EAA	Kelimeler
KA	24,90	**91,79**	57,70	lütfen, hadi, herşeyi, çözün, istiyorum
Bayes	**37,11**	88,28	**73,50**	turkcell, lütfen, hadi, acil, arkadaşlar, bana, istiyorum
1EYK	-	91,26	50,00	-
DVM	8,48	91,21	63,10	lütfen, turkcell, özel

Dilek Duygusu için ileri doğru seçme yöntemi ile Tablo 18'deki sonuçlar elde edilmiştir. KA için F-ölçeği %24,9, Doğruluğu %91,79 ve EAA'sı 0,57.7 , Bayes için, F-ölçeği %37,11, Doğruluğu %86,28 ve EAA'sı 0,73.5, 1-En Yakın Komşu için, F-ölçeği hesaplanamamış, Doğruluğu %91,26 ve EAA'sı 0,50 ve DVM için F- ölçeği %8,48, Doğruluğu %91,21 ve EAA'sı 0,63.1 olarak elde edilmiştir. Dilek duygusu için en yüksek F-ölçeğine ve EAA değerine sahip olan sınıflandırma yönteminin Bayes sınıflandırıcısı olduğu tespit edilmiştir.

Dilek Duygusu için ileri doğru seçme yönteminin sınıflandırma performanslarını artırdığı görülmüştür. Ayrıca ileri doğru seçme yöntemi ile elde edilen kelimelerin "lütfen, hadi, herşeyi, çözün, istiyorum, acil, turkcell" olduğu görülmüştür. Belirlenen bu kelimeler Dilek Duygusu için ayırt edici niteliklerdir. Bu nitelikler ile Dilek Duygusunu içeren mesajlar tespit edilmektedir. Şekil 21'de Dilek Duygusu için tespit edilmiş baskın kelimelerden oluşturulan kelime bulutu sunulmuştur.

Şekil 21. Dilek duygusu kelime bulutu

Turkcell firması için dilek duygusu incelendiğinde, müşterilerin kullandıkları hizmet ile ilgili yaşadıkları problemleri çözmeyi talep ettiği görülmüştür.

3.4. Güven Duygusu

Güven Duygusu için sınıflandırma sonuçları

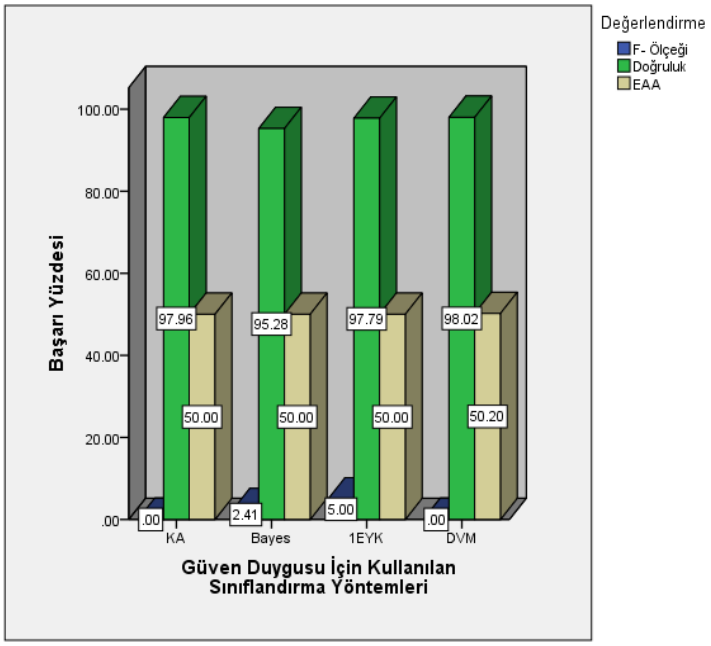
Tablo 19'deki gibi elde edilmiş ve ileri doğru seçme yöntemiyle elde edilen sonuçlar Tablo 20'daki gibi elde edilmiştir.

Tablo 19. Güvenmek duygusu sınıflandırma sonuçları

Yöntem	F- Ölçeği	Doğruluk	EAA
KA	-	97,96	50,0
Bayes	**2,41**	95,28	50,0
1 EYK	5,0	97,79	50,0
DVM	-	**98,02**	**50,2**

Güven Duygusu için

Tablo 19 incelendiğinde, KA için F-ölçeği hesaplanamamış, Doğruluğu %90,96 ve EAA'sı 0,50, Bayes için, F-ölçeği %2,41, Doğruluğu %95,28 ve EAA'sı 0,50, 1-En Yakın Komşu için, F-ölçeği %5, Doğruluğu %97,79 ve EAA'sı 0,50 ve DVM için F- ölçeği hesaplanamamış, Doğruluğu %98,02 ve EAA'sı 0,50.2 olarak elde edilmiştir. Şekil 22'de her sınıflandırma yöntemi için karşılaştırma yapılmış ve sonuçlar sunulmuştur.



Şekil 22. Güven duygusu için kullanılan sınıflandırma yöntemleri ve başarımları

Güven Duygusu için F-ölçeği en yüksek olan sınıflandırma algoritmasının Bayes sınıflandırıcısı, Doğruluk ve EAA değeri en yüksek olan sınıflandırma algoritmasının DVM sınıflandırıcısının olduğu görülmektedir.

Güven Duygusu için ileri doğru seçim yöntemi ile etkin olan kelimeler tespit edilmiştir. Tablo 20'de her sınıflandırma yöntemi için elde edilen etkin kelimeler ve başarımlar sunulmuştur.

Tablo 20. Güven duygusu için ileri doğru seçim yöntemi ile elde edilen kelimeler ve başarımları

	F Ölçeği	Doğruluk	EAA	Kelimeler
KA	-	**98,02**	55,70	geçmek
Bayes	**4,00**	5,13	**83,00**	turkcell, hep, veren, millî, gsm
1EYK	-	**98,02**	50,00	-
DVM	-	**98,02**	71,30	sizden, çeker

Güven Duygusu için ileri doğru seçme yöntemi ile Tablo 20'daki sonuçlar elde edilmiştir. KA için F-ölçeği hesaplanamamış, Doğruluğu %98,02 ve EAA'sı 0,55.7 , Bayes için, F-ölçeği %4, Doğruluğu %5,13 ve EAA'sı 0,83, 1-En Yakın Komşu için, F-ölçeği hesaplanamamış, Doğruluğu %98,02 ve EAA'sı 0,50 ve DVM için F- ölçeği hesaplanamamış, Doğruluğu %98,02 ve EAA'sı 0,71.3 olarak elde edilmiştir. Güven duygusu için en yüksek F-ölçeğine ve EAA değerine sahip olan sınıflandırma yönteminin Bayes sınıflandırıcısı olduğu tespit edilmiştir.

Güven Duygusu için ileri doğru seçme yönteminin sınıflandırma performanslarını artırdığı görülmüştür. Ayrıca ileri doğru seçme yöntemi ile elde edilen kelimelerin "geçmek, turkcell, millî, veren, çeker" olduğu görülmüştür. Belirlenen bu kelimeler Güven Duygusu için ayırt edici niteliklerdir. Bu nitelikler ile güven duygusunu içeren mesajlar tespit edilmiştir. Şekil 23'te Güven Duygusu için tespit edilen etkin kelimelerden oluşturulan kelime bulutu sunulmuştur.

Şekil 23. Güven duygusu kelime bulutu

Turkcell firması için güven duygusu incelendiğinde, sağladığı şebeke hizmeti ve millî bir firma olduğu düşüncesinden dolayı firmaya güven duyulduğu söylenebilir.

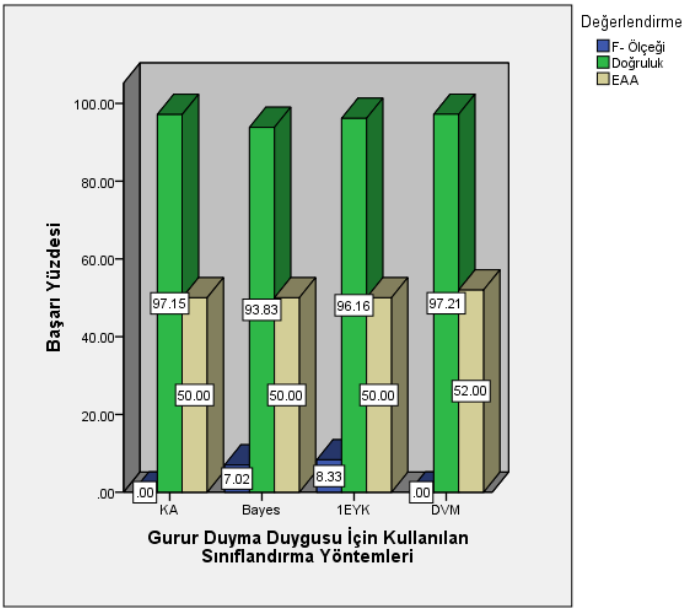
3.5. Gurur Duyma Duygusu

Gurur Duyma Duygusu için sınıflandırma sonuçları Tablo 21'deki gibi elde edilmiş ve ileri doğru seçme yöntemiyle elde edilen sonuçlar Tablo 22'deki gibi elde edilmiştir.

Tablo 21. Gurur duyma duygusu sınıflandırma sonuçları

Yöntem	F- Ölçeği	Doğruluk	EAA
KA	-	97,15	50,0
Bayes	7,02	93,83	50,0
1 EYK	**8,33**	96,16	50,0
DVM	-	**97,21**	**52,0**

Gurur Duyma Duygusu için Tablo 21 incelendiğinde, KA için F-ölçeği hesaplanamamış, Doğruluğu %97,15 ve EAA'sı 0,50, Bayes için, F-ölçeği %7,02, Doğruluğu %93,83 ve EAA'sı 0,50, 1-En Yakın Komşu için, F-ölçeği %8,33, Doğruluğu %96,16 ve EAA'sı 0,50 ve DVM için F- ölçeği hesaplanamamış, Doğruluğu %97,21 ve EAA'sı 0,52 olarak elde edilmiştir. Şekil 24'te her sınıflandırma yöntemi için karşılaştırma yapılmış ve sonuçlar sunulmuştur.



Şekil 24. Gurur duyma duygusu için kullanılan sınıflandırma yöntemleri ve başarımları

Gurur Duyma Duygusu için F-ölçeği en yüksek olan sınıflandırma algoritmasının 1EYK sınıflandırıcısı, Doğruluk ve EAA değeri en yüksek olan sınıflandırma algoritmasının DVM sınıflandırıcısının olduğu tespit edilmiştir.

Gurur Duyma Duygusu için ileri doğru seçim yöntemi ile etkin olan kelimeler tespit edilmiştir. Tablo 22'de her sınıflandırma yöntemi için elde edilen etkin kelimeler ve başarımlar sunulmuştur.

Tablo 22. Gurur duyma duygusu için ileri doğru seçim yöntemi ile elde edilen kelimeler ve başarımı

	F Ölçeği	Doğruluk	EAA	Kelimeler
KA	3,77	97,03	56,0	yüzden, iyiki, hizmet
Bayes	**33,33**	96,04	**77,0**	bu, Turk, iyiki, gurur, yapti, 15, 20, karalama, olsun
1EYK	-	**97,21**	50,0	-
DVM	3,85	97,09	70,0	@kaan_terzioglu, işlem, turkcell

Gurur Duyma Duygusu için ileri doğru seçme yöntemi ile Tablo 22'deki sonuçlar elde edilmiştir. KA için F-ölçeği %3,77, Doğruluğu %97,03 ve EAA'sı 0,56 , Bayes için, F-ölçeği %33,33, Doğruluğu %96,04 ve EAA'sı 0,77, 1-En Yakın Komşu için, F-ölçeği hesaplanamamış, Doğruluğu %97,21 ve EAA'sı 0,50 ve DVM için F- ölçeği %3,85, Doğruluğu %97,09 ve EAA'sı 0,70 olarak elde edilmiştir. Guru duyma duygusu için en yüksek F-ölçeğine ve EAA değerine sahip olan sınıflandırma yönteminin Bayes sınıflandırıcısı olduğu tespit edilmiştir.

Gurur Duyma Duygusu için ileri doğru seçme yönteminin sınıflandırma performanslarını artırdığı belirlenmiştir. Ayrıca ileri doğru seçme yöntemi ile elde edilen kelimelerin "yüzden, Türk, iyi ki, gurur, yaptı, 15, olsun, @kaan_terzioglu, işlem, turkcell" olduğu tespit edilmiştir. Belirlenen bu kelimeler gurur duyma duygusu için ayırt edici niteliklerdir. Bu nitelikler ile güven duygusunu içeren mesajlar tespit edilmiştir.

Şekil 25'te Gurur Duyma Duygusu için tespit edilen baskın kelimelerden oluşturulan kelime bulutu sunulmuştur.

Şekil 25. Gurur duyma duygusu kelime bulutu

Turkcell firması için Gurur Guyma Duygusu incelendiğinde, verilerin elde edildiği tarihle ilgili öz niteliklerin ağırlıkta olduğu görülmektedir. 15 Temmuz ve sonrasında firma olarak yapılan kampanya ve açıklamaların, kişilerde millî bir firma olma duygusu oluşturduğu ve gurur duyulduğu tespit edilmiştir.

3.6. Aşağılama Duygusu

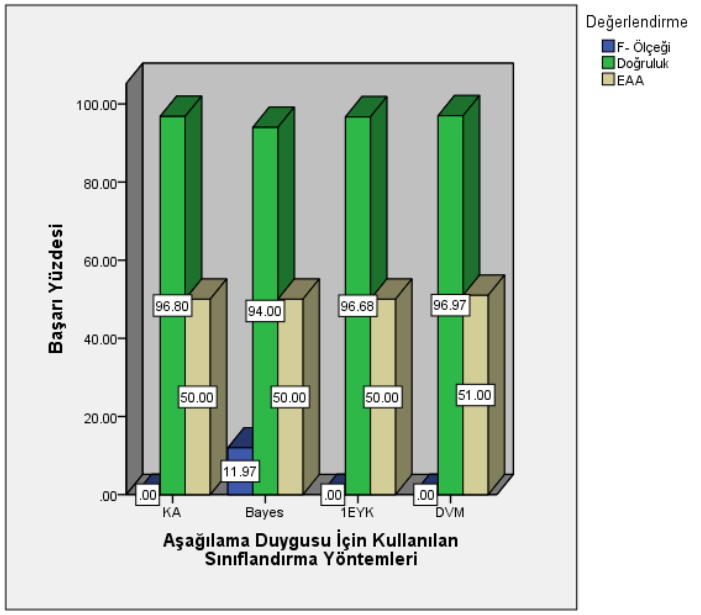
Aşağılama Duygusu için sınıflandırma sonuçları Tablo 23'deki gibi elde edilmiş ve ileri doğru seçme yöntemiyle elde edilen sonuçlar Tablo 24'deki gibidir.

Tablo 23. Aşağılama duygusu sınıflandırma sonuçları

Yöntem	F- Ölçeği	Doğruluk	EAA
KA	-	**96,80**	50,0
Bayes	**11,97**	94,00	50,0
1 EYK	-	96,68	50,0
DVM	-	96,97	**51,0**

Aşağılama Duygusu için Tablo 23 incelendiğinde, KA için F-ölçeği hesaplanamamış, Doğruluğu %96,8 ve EAA'sı 0,50, Bayes için, F-ölçeği %11,97, Doğruluğu %94 ve EAA'sı 0,50, 1-En Yakın Komşu için, F-ölçeği hesaplanamamış, Doğruluğu %96,68 ve EAA'sı 0,50 ve DVM için F- ölçeği hesaplanamamış, Doğruluğu %96,97 ve EAA'sı 0,51 olarak elde

edilmiştir. Şekil 26'da her sınıflandırma yöntemi için karşılaştırma yapılmış ve sonuçlar sunulmuştur.



Şekil 26. Aşağılama duygusu için kullanılan sınıflandırma algoritmaları ve başarımları

Aşağılama Duygusu için F-ölçeği en yüksek olan sınıflandırma algoritmasının Bayes sınıflandırıcısı, Doğruluk değeri için KA ve EAA değeri en yüksek olan sınıflandırma algoritmasının DVM sınıflandırıcısının olduğu tespit edilmiştir.

Aşağılama Duygusu için ileri doğru seçim yöntemi ile etkin olan kelimeler tespit edilmiştir. Tablo 24'te her sınıflandırma yöntemi için elde edilen etkin kelimeler ve başarımlar sunulmuştur.

Tablo 24. Aşağılama duygusu için ileri doğru seçim yöntemi ile elde edilen kelimeler ve başarımı

	F Ölçeği	Doğruluk	EAA	Kelimeler
KA	-	96,80	54,50	ensar, ?, kabul
Bayes	**6,43**	11,88	**77,80**	turkcell, abv, ensar, para
1EYK	-	**96,97**	50,00	-
DVM	-	**96,97**	63,40	turkcell, kanal

Aşağılama Duygusu için ileri doğru seçme yöntemi ile Tablo 24'deki sonuçlar elde edilmiştir. KA için F-ölçeği hesaplanamamış, Doğruluğu %96,8 ve EAA'sı 0,54.5 , Bayes

için, F-ölçeği %6,43, Doğruluğu %11,88 ve EAA'sı 0,77,8, 1-En Yakın Komşu için, F-ölçeği hesaplanamamış, Doğruluğu %96,97 ve EAA'sı 0,50 ve DVM için F- ölçeği hesaplanamamış, Doğruluğu %96,97 ve EAA'sı 0,63.4 olarak elde edilmiştir. Aşağılama duygusu için en yüksek F-ölçeğine ve EAA değerine sahip olan sınıflandırma yönteminin Bayes sınıflandırıcısı olduğu tespit edilmiştir.

Aşağılama Duygusu için ileri doğru seçme yönteminin sınıflandırma performanslarını artırdığı görülmüştür. Ayrıca ileri doğru seçme yöntemi ile elde edilen kelimelerin "ensar, turkcell, avb, para, ?, kanal" olduğu görülmüştür. Belirlenen bu kelimeler aşağılama duygusu için ayırt edici niteliklerdir. Bu nitelikler ile Aşağılama Duygusunu içeren mesajlar tespit edilmektedir. Şekil 27'de Aşağılama Duygusu için tespit edilen baskın kelimelerden oluşturulan kelime bulutu sunulmuştur.

Şekil 27. Aşağılama duygusu için kelime bulutu

Turkcell firması için aşağılama duygusu incelendiğinde, Ensar Vakfı ve Turkcell firması ile ilgili olumsuz haberin çıkması bu duygu durumunun niteliklerini etkilemiştir. Elde edilen nitelikler göz önüne alındığında sosyal medya Turkcell'e bu haberlerden dolayı yoğun bir tepki olduğu tespit edilmiştir.

3.7. Merak Duygusu

Merak Duygusu için sınıflandırma sonuçları

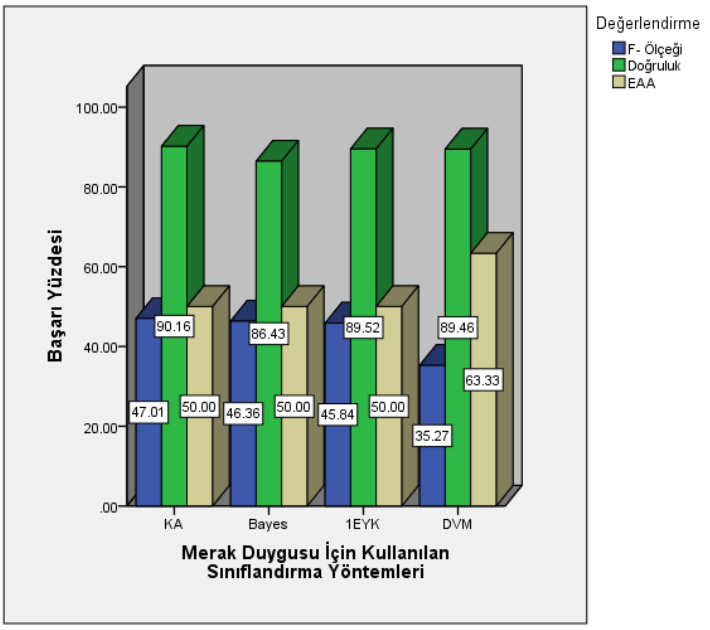
Tablo 25'deki gibi elde edilmiş ve ileri doğru seçme yöntemiyle elde edilen sonuçlar Tablo 26'deki gibidir.

Tablo 25. Merak duygusu sınıflandırma sonuçları

Yöntem	F- Ölçeği	Doğruluk	EAA
KA	**47,01**	**90,16**	50,00
Bayes	46,36	86,43	50,00
1 EYK	45,84	89,52	50,00
DVM	35,27	89,46	**63,33**

Merak Duygusu için

Tablo 25 incelendiğinde, KA için F-ölçeği %47,01, Doğruluğu %90,16 ve EAA'sı 0,50, Bayes için, F-ölçeği %46,36, Doğruluğu %86,43 ve EAA'sı 0,50, 1-En Yakın Komşu için, F-ölçeği%45,84, Doğruluğu %89,52 ve EAA'sı 0,50 ve DVM için F- ölçeği %35,27, Doğruluğu %89,46 ve EAA'sı 0,63.3 olarak elde edilmiştir. Şekil 28'de her sınıflandırma yöntemi için karşılaştırma yapılmış ve sonuçlar sunulmuştur.



Şekil 28. Merak duygusu için kullanılan sınıflandırma yöntemleri ve başarımları

Merak Duygusu için F-ölçeği en yüksek olan sınıflandırma algoritmasının KA sınıflandırıcısı ve EAA değeri en yüksek olan sınıflandırma algoritmasının DVM sınıflandırıcısının olduğu görülmektedir.

Merak Duygusu için ileri doğru seçim yöntemi ile etkin olan kelimeler tespit edilmiştir. Tablo 26'da her sınıflandırma yöntemi için elde edilen etkin kelimeler ve başarımlar sunulmuştur.

Tablo 26. Merak duygusu için ileri doğru seçim yöntemi ile elde edilen kelimeler ve başarımı

	F Ölçeği	Doğruluk	EAA	Kelimeler
KA	**57,10**	**91,55**	70,80	acaba, peki, olayı, dakika, bilginiz, varmı
Bayes	56,53	88,93	**85,60**	neden, ne, mi, nasıl, mu, ?, niye, varmı,
1EYK	-	86,49	50,00	-
DVM	48,43	89,92	76,00	öğren, ne, mi

Merak Duygusu için ileri doğru seçme yöntemi ile Tablo 26 'deki sonuçlar elde edilmiştir. KA için F-ölçeği %57,1, Doğruluğu %91,5 ve EAA'sı 0,70.8 , Bayes için, F-ölçeği %56,43, Doğruluğu %88,93 ve EAA'sı 0,85.6, 1-En Yakın Komşu için, F-ölçeği hesaplanamamış, Doğruluğu %86,49 ve EAA'sı 0,50 ve DVM için F- ölçeği %48,43,

Doğruluğu %89,92 ve EAA'sı 0,76 olarak elde edilmiştir. Merak Duygusu için en yüksek F-ölçeği değerine KA sınıflandırıcısı, en yüksek EAA değerine sahip olan sınıflandırma yönteminin Bayes sınıflandırıcısı olduğu tespit edilmiştir. Şekil 29'da Merak Duygusu için tespit edilen baskın kelimelerden oluşturulan kelime bulutu sunulmuştur.

Şekil 29. Merak duygusu kelime bulutu

Merak Duygusu için ileri doğru seçme yönteminin sınıflandırma performanslarını artırdığı görülmüştür. Ayrıca ileri doğru seçme yöntemi ile elde edilen kelimelerin "acaba, peki, dakika, ne, neden, mi, mı, nasıl, niye, ?, öğren" olduğu görülmüştür. Belirlenen bu kelimeler merak duygusu için ayırt edici niteliklerdir. Bu nitelikler ile merak duygusunu içeren mesajlar tespit edilmektedir.

Turkcell firması için merak duygusu incelendiğinde, kampanya ve ürünleri ile ilgili bilgi talebinde bulunduğu görülmüştür.

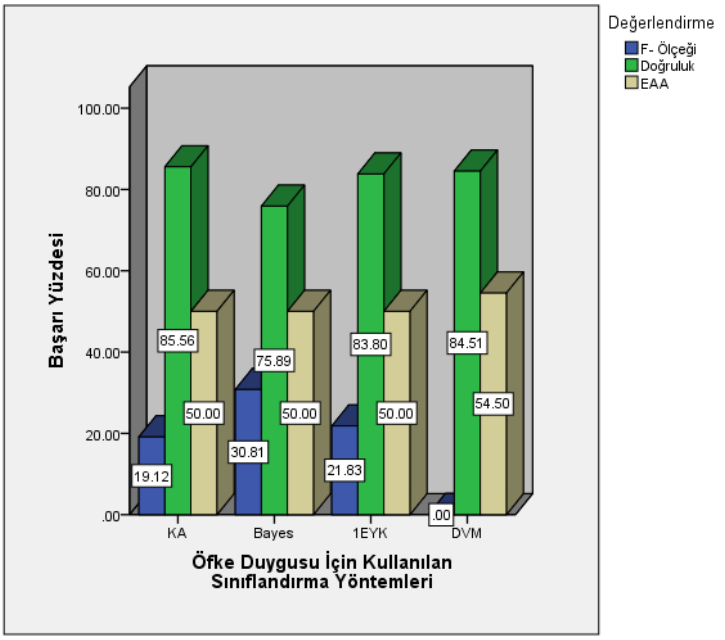
3.8. Öfke Duygusu

Öfke Duygusu için sınıflandırma sonuçları Tablo 27'teki gibi elde edilmiş ve ileri doğru seçme yöntemiyle elde edilen sonuçlar Tablo 24'teki gibi elde edilmiştir.

Tablo 27. Öfke duygusu sınıflandırma sonuçları

Yöntem	F- Ölçeği	Doğruluk	EAA
KA	19,12	**85,56**	50,00
Bayes	**30,81**	75,89	50,00
1 EYK	21,83	83,80	50,00
DVM	-	84,51	**54,50**

Öfke Duygusu için Tablo 27 incelendiğinde, KA için F-ölçeği %19,12, Doğruluğu %85,56 ve EAA'sı 0,50, Bayes için, F-ölçeği %30,81, Doğruluğu %75,89 ve EAA'sı 0,50, 1-En Yakın Komşu için, F-ölçeği%21,83, Doğruluğu %83,8 ve EAA'sı 0,50 ve DVM için F- ölçeği hesaplanamamış, Doğruluğu %84,51 ve EAA'sı 0,54.5 olarak elde edilmiştir. Şekil 30'da her sınıflandırma yöntemi için karşılaştırma yapılmış ve sonuçlar sunulmuştur.



Şekil 30. Öfke duygusu için kullanılan sınıflandırma yöntemleri ve başarımları

Öfke Duygusu için F-ölçeği en yüksek olan sınıflandırma algoritmasının Bayes sınıflandırıcısı ve EAA değeri en yüksek olan sınıflandırma algoritmasının DVM sınıflandırıcısının olduğu görülmektedir.

Öfke Duygusu için ileri doğru seçim yöntemi ile etkin olan kelimeler tespit edilmiştir. Tablo 28'de her sınıflandırma yöntemi için elde edilen etkin kelimeler ve başarımlar sunulmuştur.

Tablo 28. Öfke duygusu için ileri doğru seçim yöntemi ile elde edilen kelimeler ve başarımı

	F Ölçeği	Doğruluk	EAA	Kelimeler
KA	**31,68**	**87,01**	59,50	fatura, nefret, soygun, haber

Bayes	28,99	84,80	**72,60**	turkcell, yok, turkcellhizmet, fatura, hakaret, artık
1EYK	23,78	50,75	50,00	-
DVM	10,49	85,09	63,80	fatura, değil, müdahale

Öfke Duygusu için ileri doğru seçme yöntemi ile Tablo 24'deki sonuçlar elde edilmiştir. KA için F-ölçeği %31,68, Doğruluğu %87,01 ve EAA'sı 0,59.5, Bayes için, F-ölçeği %28,99, Doğruluğu %84,8 ve EAA'sı 0,72.6, 1-En Yakın Komşu için, F-ölçeği %23,78, Doğruluğu %50,75 ve EAA'sı 0,50 ve DVM için F- ölçeği %10,49, Doğruluğu %85,09 ve EAA'sı 0,63.8 olarak elde edilmiştir. Öfke duygusu için en yüksek F-ölçeği değerine KA sınıflandırıcısı, en yüksek EAA değerine sahip olan sınıflandırma yönteminin Bayes sınıflandırıcısı olduğu tespit edilmiştir.

Öfke Duygusu için ileri doğru seçme yönteminin sınıflandırma performanslarını artırdığı görülmüştür. Ayrıca ileri doğru seçme yöntemi ile elde edilen kelimelerin "fatura, nefret, soygun, turkcell, turkcellhizmet, artık, hakaret, değil, müdahale" olduğu görülmüştür. Belirlenen bu kelimeler öfke duygusu için ayırt edici niteliklerdir. Bu nitelikler ile öfke duygusunu içeren mesajlar tespit edilmektedir. Şekil 31'de Öfke Duygusu için tespit edilen baskın kelimelerden oluşturulan kelime bulutu sunulmuştur.

Şekil 31. Öfke duygusu kelime bulutu

Turkcell firması için Öfke Duygusu incelendiğinde, faturadan kaynaklanan sorunların müşterileri öfkelendirdiği görülmektedir. Turkcell firmasının müşteri memnuniyetini

arttırması için firmaya öfke duyan kişilerin sorununu hızlı bir şekilde çözmesi gerekmektedir.

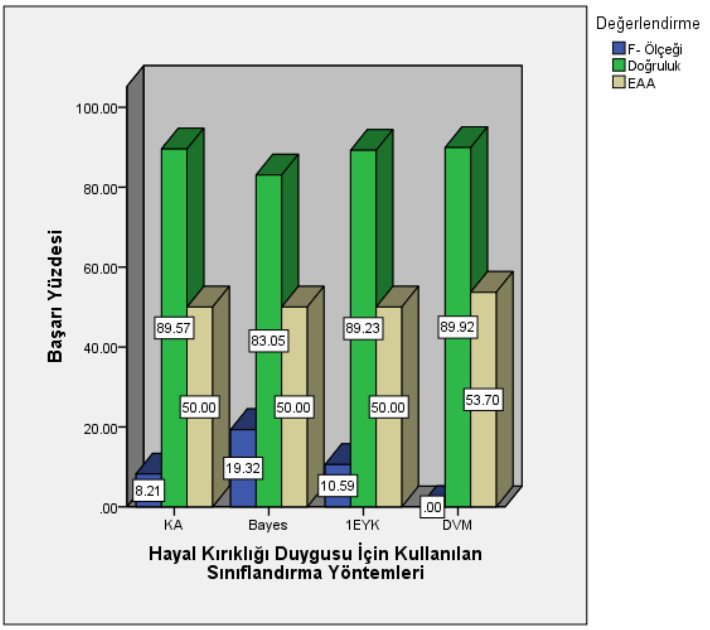
3.9. Hayal Kırıklığı Duygusu

Hayal Kırıklığı Duygusu için sınıflandırma sonuçları Tablo 29'teki gibi elde edilmiş ve ileri doğru seçme yöntemiyle elde edilen sonuçlar Tablo 30'daki gibidir.

Tablo 29. Hayal kırıklığı duygusu sınıflandırma sonuçları

Yöntem	F- Ölçeği	Doğruluk	EAA
KA	8,21	89,57	50,00
Bayes	**19,32**	83,05	50,00
1 EYK	10,59	89,23	50,00
DVM	-	**89,92**	**53,70**

Hayal Kırıklığı Duygusu için Tablo 29 incelendiğinde, KA için F-ölçeği %8,21, Doğruluğu %89,57 ve EAA'sı 0,50, Bayes için, F-ölçeği %19,32, Doğruluğu %83,05 ve EAA'sı 0,50, 1-En Yakın Komşu için, F-ölçeği%10,59, Doğruluğu %89,23 ve EAA'sı 0,50 ve DVM için F- ölçeği hesaplanamamış, Doğruluğu %89,92 ve EAA'sı 0,53.7 olarak elde edilmiştir. Şekil 32'de her sınıflandırma yöntemi için karşılaştırma yapılmış ve sonuçlar sunulmuştur.



Şekil 32. Hayal kırıklığı duygusu için kullanılan sınıflandırma yöntemleri ve başarımları

Hayal Kırıklığı Duygusu için F-ölçeği en yüksek olan sınıflandırma algoritmasının Bayes sınıflandırıcısı ve EAA değeri en yüksek olan sınıflandırma algoritmasının DVM sınıflandırıcısının olduğu görülmektedir.

Hayal Kırıklığı Duygusu için ileri doğru seçim yöntemi ile etkin olan kelimeler tespit edilmiştir. Tablo 30'da her sınıflandırma yöntemi için elde edilen etkin kelimeler ve başarımlar sunulmuştur.

Tablo 30. Hayal kırıklığı duygusu için ileri doğru seçim yöntemi ile elde edilen kelimeler ve başarımı

	F Ölçeği	Doğruluk	EAA	Kelimeler
KA	-	**89,92**	50,00	çekmiyor, anca, bölgede, ☹, yollayın, oo, istanbulun, pü
Bayes	**31,98**	87,30	**75,60**	turkcell, yok, ☹, çekmiyor, superonlinetr, turktelekom
1EYK	-	**89,92**	50,00	-
DVM	-	**89,92**	55,10	aldım, bitireyim

Hayal Kırıklığı Duygusu için ileri doğru seçme yöntemi ile Tablo 30'daki sonuçlar elde edilmiştir. KA için F-ölçeği hesaplanamamış, Doğruluğu %89,92 ve EAA'sı 0,50,

Bayes için, F-ölçeği %31,98, Doğruluğu %87,3 ve EAA'sı 0,75.6, 1-En Yakın Komşu için, F-ölçeği hesaplanamamış, Doğruluğu %89,92 ve EAA'sı 0,50 ve DVM için F- ölçeği hesaplanamamış, Doğruluğu %89,92 ve EAA'sı 0,55.1 olarak elde edilmiştir. Hayal kırıklığı duygusu için en yüksek F-ölçeği değerine ve EAA değerine sahip olan sınıflandırma yönteminin Bayes sınıflandırıcısı olduğu tespit edilmiştir.

Hayal Kırıklığı Duygusu için ileri doğru seçme yönteminin sınıflandırma performanslarını artırdığı görülmüştür. Ayrıca ileri doğru seçme yöntemi ile elde edilen kelimelerin "çekmiyor, anca, yollayın, pü, ☹, superonlinetr, turktelekom" olduğu görülmüştür. Belirlenen bu kelimeler hayal kırıklığı duygusu için ayırt edici niteliklerdir. Bu nitelikler ile hayal kırıklığı duygusunu içeren mesajlar tespit edilmektedir. Şekil 33'te Hayal Kırıklığı Duygusu için tespit edilen baskın kelimelerden oluşturulan kelime bulutu sunulmuştur.

Şekil 33. Hayal kırıklığı duygusu kelime bulutu

Turkcell firması için hayal kırıklığı duygusu incelendiğinde, şebekenin çekmemesi durumu, Turkcell kullanıcıları için hayal kırıklığı oluşturmaktadır. Bu sorunu yaşayan kullanıcıların tespit edilmesi ve sorunun giderilmesi Turkcell'in marka değerini artıracaktır.

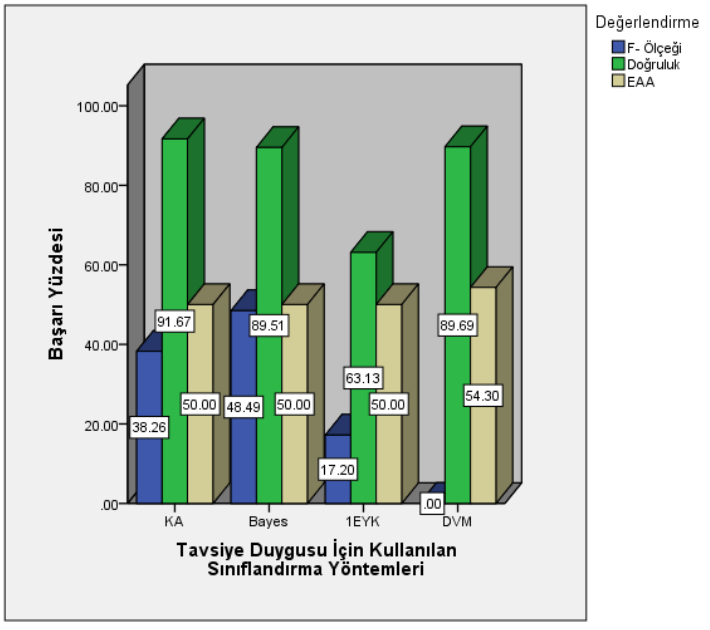
3.10. Tavsiye Duygusu

Tavsiye Duygusu için sınıflandırma sonuçları Tablo 31'deki gibi elde edilmiş ve ileri doğru seçme yöntemiyle elde edilen sonuçlar Tablo 32'deki gibidir.

Tablo 31. Tavsiye duygusu sınıflandırma sonuçları

Yöntem	F- Ölçeği	Doğruluk	EAA
KA	38,26	**91,67**	50,00
Bayes	**48,49**	89,51	50,00
1 EYK	17,20	63,13	50,00
DVM	-	89,69	**54,30**

Tavsiye Duygusu için Tablo 31 incelendiğinde, KA için F-ölçeği %28,26, Doğruluğu %91,67 ve EAA'sı 0,50, Bayes için, F-ölçeği %48,49, Doğruluğu %89,51 ve EAA'sı 0,50, 1-En Yakın Komşu için, F-ölçeği%17,2, Doğruluğu %63,13 ve EAA'sı 0,50 ve DVM için F- ölçeği hesaplanamamış, Doğruluğu %89,69 ve EAA'sı 0,55.3 olarak elde edilmiştir. Şekil 34'te her sınıflandırma yöntemi için karşılaştırma yapılmış ve sonuçlar sunulmuştur.



Şekil 34. Tavsiye duygusu için kullanılan sınıflandırma yöntemleri ve başarımları

Tavsiye Duygusu için F-ölçeği en yüksek olan sınıflandırma algoritmasının Bayes sınıflandırıcısı ve EAA değeri en yüksek olan sınıflandırma algoritmasının DVM sınıflandırıcısının olduğu görülmektedir.

Tavsiye Duygusu için ileri doğru seçim yöntemi ile etkin olan kelimeler tespit edilmiştir. Tablo 32'de her sınıflandırma yöntemi için elde edilen etkin kelimeler ve başarımlar sunulmuştur.

Tablo 32. Tavsiye duygusu için ileri doğru seçim yöntemi ile elde edilen kelimeler ve başarımı

	F Ölçeği	Doğruluk	EAA	Kelimeler
KA	**47,94**	**92,90**	66,60	bağlantısından, inceleyebilirsiniz, ?
Bayes	23,16	32,96	**78,60**	turkcell, ne, öğren, olsun, yok,
1EYK	-	89,81	50,00	-
DVM	24,00	91,15	70,30	bağlantısından, Turkcell, az, mıydı,

Tavsiye Duygusu için ileri doğru seçme yöntemi ile Tablo 30'daki sonuçlar elde edilmiştir. KA için F-ölçeği %47,94, Doğruluğu %92,9 ve EAA'sı 0,66.6, Bayes için, F-ölçeği %23,16, Doğruluğu %32,96 ve EAA'sı 0,78.6, 1-En Yakın Komşu için, F-ölçeği hesaplanamamış, Doğruluğu %89,81 ve EAA'sı 0,50 ve DVM için F- ölçeği %24, Doğruluğu %91,15 ve EAA'sı 0,70.3 olarak elde edilmiştir. Tavsiye duygusu için en yüksek F-ölçeği değerine sahip olan sınıflandırıcının KA, en yüksek EAA değerine sahip olan sınıflandırma yönteminin Bayes sınıflandırıcısı olduğu tespit edilmiştir.

Tavsiye duygusu için ileri doğru seçme yönteminin sınıflandırma performanslarını artırdığı görülmüştür. Ayrıca ileri doğru seçme yöntemi ile elde edilen kelimelerin "bağlantısı, inceleyebilirsiniz, turkcell, olsun, yok" olduğu görülmüştür. Belirlenen bu kelimeler tavsiye duygusu için ayırt edici niteliklerdir. Bu nitelikler ile tavsiye duygusunu içeren mesajlar tespit edilmektedir. Şekil 35'te Tavsiye Duygusu için tespit edilen baskın kelimelerden oluşturulan kelime bulutu sunulmuştur.

Şekil 35. Tavsiye duygusu kelime bulutu

4. ÖNERİLER

Bu bölümde tez kapsamında gerçekleştirilen modelin başarımı ve performansını artıracak çalışmalardan bahsedilmektedir. Tez kapsamında Turkcell firmasına ait 2100 adet twitter mesajı incelenmiş, bu mesajlar 10 farklı duygu durumuna göre analiz edilmiş ve sonuçlar karşılaştırılmıştır.

Sosyal medya kullanımının yaygınlaşması ile birlikte sosyal medyaya özgü yazım dili ortaya çıkmıştır. Özellikle Twitter gibi karakter sınırlamasının olduğu mikro blog siteleri ile birlikte kısaltmalar, bitişik yazım, sesli harflerin yazılmaması şeklinde yazım tarzları ortaya çıkmıştır. Türkçe'nin kurallı yapısının tersi metinlerin olduğu sosyal medyada, bu metinlerin analizleri zor olmaktadır.

Türkçe'nin kurallı yapısı temel alınarak, sosyal medya metinlerine özgü bir doğal dil işleme kütüphanesinin oluşturulması durumunda, sınıflandırma modelinin performansının artacağı düşünülmektedir. Örneğin, Türkçe yazı dilinde tek başına bir harfin anlamı olmayabilir, fakat sosyal medyada "i" harfi ile "iyi" kelimesi kısaltılmakta ve mesajlar bu şekilde yayınlanmaktadır. Analiz sonuçlarında elde edilen başarımlara direkt etkisi olan ön işleme aşamasının, doğal dil işleme kütüphanesinin içerisinde yapılması hem zaman hem de performans açısından fayda sağlayacaktır. Bu çalışmanın devamında sosyal medya metinlerine özgü Türkçe doğal dil işleme kütüphanesinin geliştirilmesi planlanmaktadır. Ayrıca oluşturulacak kütüphanede duygulara ait niteliklerin, diğer kelimelerle olan anlamsal ilişkilerini de gösterecek bir eklentinin oluşturulması amaçlanmaktadır.

Sosyal medyada bir haberin yayılma hızı göz önüne alındığında, yayınlanan mesajın firmanın imajını kötü etkilemesi çok kolay olmakta ve sonucunda firma, borsadaki hisse değerinin düşmesinden, müşteri sayısının azalmasına kadar birçok maddi ve manevi zarara uğramaktadır. Duygu durumları tespit edildikten sonra duyguların birbiriyle olan ilişkileri ayrıca irdelenerek, insanları o mesajı yazmaya iten esas neden tespit edilebilir. Bu sayede firmanın imajı sarsılmasında, müşteri memnuniyeti sağlanmış ve firmanın imajı korunmuş olunacaktır.

Duygu durumlarının ayrı ayrı tespit edilmesi, kişinin yazdığı mesajların tutarlılığını tespit etmeyi sağlayabilir. Bu sayede sadece firmayı kötülemek için açılan hesaplar tespit edilerek firmanın imajı korunabilir. Firmaların imajını kötülemek için açılan bu hesaplar hem firma çalışanlarını hem de firmayı zor duruma sokmaktadır.

Bu çalışma sonucunda elde edilen sonuçların daha fazla veri ve saha çalışmasıyla desteklenmesi durumunda, önerilen yönteminin başarımının artacağı düşünülmektedir. Özellikle mevcut verilerin miktarının az olmasına rağmen bazı duygularda elde edilen sonuçlar bu varsayımı güçlendirmektedir.

5. KAYNAKÇA

1. Https://www.turkcebilgi.com/iletişim İletişim Nedir?, İletişim Nedir?, 08 Şubat 2017.
2. Https://support.twitter.com/categories/ 281 Tweeter Nasıl Kullanılır?, Twitter'ı Kullanma, 20 Ocak 2017.
3. Kemp, S., Digital in 2016: We Are Social's Compendium of Global Digital, Social, and Mobile Data, Trends, and Statistics, Www.Wearesocial.Com, (2016) 537.
4. Nasukawa, T. and Yi, J., Sentiment Analysis: Capturing Favorability Using Natural Language Processing, Proceedings of the 2nd International Conference on Knowledge Capture, October 23-25, 2003, Sanibel Island, FL, USA,70–77.
5. Dave, K., Lawrence, S. and Pennock, D.M., Mining the Teanut Gallery: Opinion Extraction and Semantic Classification of Product Reviews, Proceedings of the 12th international conference on World Wide Web, May 20 - 24, 2003, Budapest, Hungary,519–528.
6. Das, S. and Chen, M., Yahoo! For Amazon: Extracting Market Sentiment From Stock Message Boards, Proceedings of the Asia Pacific Finance Association Annual Conference (APFA), July 22, 2001, Vol. 35, 43-57, New York.
7. Morinaga, S., Yamanishi, K., Tateishi, K. and Fukushima, T., Mining Product Reputations On The Web, Proceedings of The Eighth ACM SIGKDD International Conference On Knowledge Discovery And Data Mining, July 23 - 25, 2002, Edmonton, AB, Canada, 341-349.
8. Pang, B., Lee, L. and Vaithyanathan, S., Thumbs up?: Sentiment Classification Using Machine Learning Techniques, Proceedings of the ACL-02 Conference On Empirical Methods in Natural Language Processing, Volume 10, July 2002, Philadelphia USA, 79 - 86.
9. Liu, Y., Huang, X., An, A. and Yu, X., ARSA: A Sentiment-Aware Model For Predicting Sales Performance Using Blogs, Proceedings of the 30th Annual International ACM SIGIR Conference On Research And Development In Information Retrieval, July 23 - 27, 2007, Amsterdam, 607-614.
10. Hong, Y. and Skiena, S., The Wisdom of Bookies? Sentiment Analysis Versus vs. the NFL Point Spread, International Conference on Weblogs and Social Media (ICWSM), Vol 4, May 23-26, 2010, George Washington University, Washington, DC, 251 - 254.

11. O'Connor, B., Balasubramanyan, R., Routledge, B.R. and Smith, N.A., From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series, International Conference on Weblogs and Social Media (ICWSM), Vol 4, May 23-26, 2010, George Washington University, Washington, DC, 122-129.
12. Tumasjan, A., Sprenger, T.O., Sandner, P.G. and Welpe, I.M., Predicting Elections With Twitter: What 140 Characters Reveal About Political Sentiment, International Conference on Weblogs and Social Media (ICWSM), Vol 4, May 23-26, 2010, George Washington University, Washington, DC, 178-185.
13. Asur, S. and Huberman, B.A., Predicting the Future with Social Media, 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, August 31- September 3, 2010, Toronto Canada, 492-499.
14. Sadikov, E., Parameswaran, A.G. and Venetis, P., Blogs as Predictors of Movie Success, International Conference on Weblogs and Social Media (ICWSM), Vol 3, May 17 - 20, 2009, San Jose, California, 303-307.
15. Joshi, M., Das, D., Gimpel, K. and Smith, N.A., Movie Reviews And Revenues: An Experiment In Text Regression, Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, une 1–6, 2010, Los Angeles, USA, 293-296.
16. Mohammad, S.M. and Yang, T.W., Tracking Sentiment in Mail: How Genders Differ On Emotional Axes, Proceedings of The 2nd Workshop On Computational Approaches To Subjectivity And Sentiment Analysis, 24 June, 2011, Oregon, USA, 70-79.
17. Mohammad, S., From Once Upon A Time To Happily Ever After: Tracking Emotions in Novels And Fairy Tales, Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities, June 24, 2011, Oregon, USA, 105-114.
18. Paltoglou, G., Sentiment- Based Event Detection in Twitter, Journal of the Association for Information Science and Technology, Volume 67,Issue 7 (2016) 1576–1587.
19. Lin, W.-H., Wilson, T., Wiebe, J. and Hauptmann, A., Which Side Are You On?: Identifying Perspectives At The Document And Sentence Levels, Proceedings Of The Tenth Conference On Computational Natural Language Learning, 8-9 June 2006 New York City, USA, 109-116.

20. Riloff, E., Wiebe, J. and Wilson, T., Learning Subjective Nouns Using Extraction Pattern Bootstrapping, Proceedings of The Seaventh Conference On Natural Language Learning At HLT-NAACL,Volume 4, 2003, Edmonton, Canada, 25-32.
21. Kaşıkçı, T. and Gökçen, H., Metin Madenciliği İle E-Ticaret Sitelerinin Belirlenmesi, International Journal of Informatics Technologies, 7,1 (2014) 25–32.
22. Güran, A., Akyokuş, S., Bayazıt, N.G. and Gürbüz, M.Z., Turkish Text Categorization Using N-Gram Words, Proceedings of the International Symposium on Innovations in Intelligent Systems and Applications (INISTA), 29 June - 1 July, 2009, Trabzon, Turkey, 369-373.
23. Doğan, S. and Diri, B., Türkçe Dokümanlar İçin N-Gram Tabanlı Yeni Bir Sınıflandırma (Ng-Ind): Yazar, Tür ve Cinsiyet, Türkiye Bilişim Vakfı Bilgisayar Bilimleri Ve Mühendisliği Dergisi, 3,1 (Basılı 3) (2010).
24. Şimşek, M.U. and Özdemir, S., Analysis Of The Relation Between Turkish Twitter Messages And Stock Market İndex, The 6th International Conference on Application of Information and Communication Technologies AICT2012, October 2012, Tbilisi, Georgia, 1-4.
25. Olson, D.L. and Delen, D., Advanced Data Mining Techniques, Springer Science & Business Media, Verlag Berlin Heidelberg, 2008.
26. Aggarwal, C. and Zhai, C., Mining Text Data, Springer Science & Business Media, New York Dordrecht Heidelberg London, 2012.
27. Maimon, O. and Rokach, L., Data Mining and Knowledge Discovery Handbook, Second Edi, New York Dordrecht Heidelberg London, 2009.
28. Sayad, S., Data Mining, Http://www.saedsayad.com/data_mining.htm, 12 Nisan 2017.
29. KDnuggets.com, Where Analytics, Data Mining, Data Science Were Applied in 2016, http://www.kdnuggets.com/2016/12/poll-Analytics-Data-Mining-Data-Science-Applied-2016.html 12 Ocak 2017.
30. http://www.tutorialspoint.com/data_mining/dm_applications_trends.htm, Data Mining Applications, 24 Nisan 2017, .
31. Rajkumar, P., Rajkumar, P., 14 Useful Applications of Data Mining, (2014).
32. Sumathi, S. and Sivanandam, S.N., Introduction to Data Mining and Its Applications, Springer Science & Business Media, Verlag Berlin Heidelberg, 2006.
33. Marsh, J., Marsh, J., Knowledge Discovery in Databases,

Https://blog.udemy.com/knowledge-Discovery-in-Databases, 1 Nisan 2017.

34. Şimşek, M.U., Sosyal Ağlarda Veri Madenciliği Üzerine Bir Uygulama, Yüksek Lisans Tezi, Gazi Üniversitesi, Fen Bilimleri Enstitüsü, Ankara, 2012.
35. Karaca, M, F., Metin Madenciliği Yöntemi İle Haber Sitelerindeki Köşe Yazılarının Sınıflandırılması, Yüksek Lisans Tezi, Karabük Üniversitesi, Fen Bilimleri Enstitüsü, Karabük, 2012.
36. Srivastava, A.N. and Sahami, M., Text Nining: Classification, Clustering, and Applications, Ed. V. Kumar, CRC Press, Boca Raton, 2009.
37. Çoşlu, E., Veri Madenciliği, 15. Akademik Bilişim Konferansı, Ocak 2013, 15, Bildiriler Kitabı, Akdeniz Üniversitesi, Hukuk Fakültesi, Antalya, 573-578.
38. Çalış, A. and Baynal, K., Veri Madenciliğinde Kümeleme Analizi Ile Bankacılık Sektöründe Bir Uygulama, Beykent Üniversitesi Fen Ve Mühendislik Bilimleri Dergisi, 9,1 (2016) 13–41.
39. Nizam, H. and Akın, S.S., Sosyal Medyada Makine Öğrenmesi Ile Duygu Analizinde Dengeli ve Dengesiz Veri Setlerinin Performanslarının Karşılaştırılması, XIX. Türkiye'de İnternet Konferansı, Kasım 2014 , Yaşar Üniversitesi, İzmir.
40. Fu, Y., Data Mining, IEEE Potentials, 16,4 (1997) 18–20.
41. Çoban, Ö., Özyer, B. and Özyer, G.T., Sentiment Analysis for Turkish Twitter Feeds, Signal Processing and Communications Applications Conference (SIU), 2015 23th, Mayıs 2015, Malatya, Türkiye, 2388-2391.
42. Emel, G.G. and Taşkın, Ç., Veri Madenciliğinde Karar Ağaçları Ve Bir Satış Analizi Uygulaması, Sosyal Bilimler Dergisi, 6,2 (2005).
43. Bloemer, J.M.M., Brijs, T., Vanhoof, K. and Swinnen, G., Comparing Complete and Partial Classification for Identifying Customers at Risk, International Journal of Research in Marketing, 20,2 (2003) 117–131.
44. Albayrak, A.S. and Yilmaz, S.K., Veri Madenciliği: Karar Ağacı Algoritmaları Ve İmkb Verileri Üzerine Bir Uygulama, Suleyman Demirel University Journal of Faculty of Economics & Administrative Sciences, 14,1 (2009).
45. Atbaş, A., Kümeleme Analizinde Küme Sayısının Belirlenmesi Üzerine Bir Çalışma, Yüksek Lisans Tezi, Ankara Üniversitesi, Fen Bilimleri Enstitüsü, Ankara, 2008.
46. Tan, P.-N., Steinbach, M. and Kumar, V., Introduction to Data Mining, Pearson Addison Wesley, New York, 2006.
47. Meltem, I. and Çamurcu, A.Y., K-Means Ve Aşırı Küresel C-Means Algorıtmaları İle

Belge Madenciliği, Marmara Fen Bilimleri Dergisi, 22,1 (2010) 1–18.

48. Yen, S.-J. and Lee, Y.-S., An Efficient Data Mining Approach for Discovering Interesting Knowledge from Customer Transactions, Expert Systems with Applications, 30,4 (2006) 650–657.
49. Akpınar, H., Veri Tabanlarında Bilgi Keşfi ve Veri Madenciliği, İÜ İşletme Fakültesi Dergisi, 29,1 (2000) 1–22.
50. Bender, O., Och, F.J. and Ney, H., Maximum Entropy Models For Named Entity Recognition, Proceedings Of The Seventh Conference On Natural Language Learning at HLT-NAACL,Volume 4, 27 May- June 1, 2003, Edmonton, Canada ,148-151.
51. Berger, A.L., Pietra, V.J. Della and Pietra, S.A. Della, A Maximum Entropy Approach to Natural Language Processing, Computational Linguistics, 22,1 (1996) 39–71.
52. Demner-Fushman, D. and Lin, J., Knowledge Extraction For Clinical Question Answering: Preliminary Results, Proceedings of the AAAI-05 Workshop on Question Answering in Restricted Domains, July 2005, Pennsylvania, 9-13.
53. Aggarwal, C.C. and ChengXiang, Z., Mining Text Data, Ed. C.Z. Charu C. Aggarwal, Springer, 415-463, New York, 2012.
54. Maynard, D., Bontcheva, K. and Rout, D., Challenges in Developing Opinion Mining Tools For Social Media, The International Conference on Language Resources and Evaluation, May 2008, Slovenia, 15-22.
55. Weller, K., Bruns, A., Burgess, J.E., Mahrt, M. and Puschmann, C., Twitter and Society, Ed. S. Jones, 89, Peter Lang, New York, 2014.
56. Kouloumpis, E., Wilson, T. and Moore, J.D., Twitter Sentiment Analysis: The Good The Bad And The Omg, Fifth International AAAI Conference on Weblogs and Social Media, ICWSM, July 2011, Barcelona, 538-541.
57. Parrott, W.G., Emotions in Social Psychology: Essential Readings, Psychology Press, 2001.
58. Liu, B., Sentiment Analysis: Mining Opinions, Sentiments, And Emotions, Cambridge University Press, New York, 2015.
59. Http://emotion-Research.net/projects/humaine, 20 Nisan 2017.
60. Ortony, A. and Turner, T.J., What's Basic about Basic Emotions?, Psychological Review, 97,3 (1990) 315.
61. Rouse, M., Margaret, R., Plutchik's Wheel of Emotions, Http://whatis.techtarget.com/definition/ Plutchiks-Wheel-of-Emotions, 5 Mayıs

2017.

62. Pozzi, F.A., Fersini, E., Messina, E. and Liu, B., Sentiment Analysis in Social Networks, Morgan Kaufmann, Cambridge, USA, 2016.
63. Pang, B. and Lee, L., Opinion Mining and Sentiment Analysis, Foundations and Trends® in Information Retrieval, 2,1–2 (2008) 1–135.
64. Medhat, W., Hassan, A. and Korashy, H., Sentiment Analysis Algorithms and Applications: A Survey, Ain Shams Engineering Journal, 5,4 (2014) 1093–1113.
65. Döven, S., Metin Madenciliği İle Dokümanlar Arasındaki Benzerliklerin Bulunması, Yüksek Lisans Tezi, Bahçeşehir Üniversitesi, Fen Bilimleri Enstitüsü, İstanbul, 2013.
66 Bounsaythip, C. and Rinta-Runsala, E., Overview Of Data Mining For Customer Behavior Modeling, VTT Information Technology Research Report, 1 (2001) 1–53.
67. Akbaş, E., Aspect Based Opinion Mining On Turkish Tweets, Yüksek Lisans Tezi, Bilkent Üniversitesi, Fen Bilimleri Enstitüsü, Ankara, 2012.
68. Cortes, C. and Vapnik, V., Support-Vector Networks, Machine Learning, 20,3 (1995) 273–297.
69. Vapnik, V., The Nature Of Statistical Learning Theory, Springer Science & Business Media, 2013.
70. Kavzoğlu, T. and Çölkesen, İ., Destek Vektör Makineleri Ile Uydu Görüntülerinin Sınıflandırılmasında Kernel Fonksiyonlarının Etkilerinin Incelenmesi, Harita Dergisi, 144,7 (2010) 73–82.
71. Osuna, E., Freund, R. and Girosi, F., Support Vector Machines: Training and Applications, ,Massachusetts Institute of Technology and Artificial Intelligence Laboratory, (1997) 144.
72. Zhang, G., Patuwo, B.E. and Hu, M.Y., Forecasting with Artificial Neural Networks:: The State of the Art, International Journal of Forecasting, 14,1 (1998) 35–62.
73. Friedman, N., Geiger, D. and Goldszmidt, M., Bayesian Network Classifiers, Machine Learning, 29,2–3 (1997) 131–163.
74. McCallum, A. and Nigam, K., A Comparison Of Event Models For Naive Bayes Text Classification, AAAI-98 Workshop On Learning For Text Categorization, July 1998, Madison, Wisconsin, 41-48.
75. Eyheramendy, S., Lewis, D.D. and Madigan, D., On The Naive Bayes Model For Text Categorization, (2003).
76. Pilavcılar, I.F., Metin Madenciliğiyle Metin Sınıflandırma, Yüksek Lisans Tezi,

Yıldız Teknik Üniversitesi, Fen Bilimleri Enstitüsü, 2007.

77. Schneider, K.-M., Advances in Natural Language Processing, Springer, 2004.
78. Kavzoğlu, T. and Çölkesen, İ., Karar Ağaçları İle Uydu Görüntülerinin Sınıflandırılması, Harita Teknolojileri Elektronik Dergisi, 2,1 (2010) 36–45.
79. Hu, M. and Liu, B., Mining And Summarizing Customer Reviews, Proceedings of the Tenth ACM SIGKDD inTernational Conference On Knowledge Discovery And Data Mining, August 2004, Seattle, WA, USA, 168-177.
80. Hatzivassiloglou, V. and McKeown, K.R., Predicting The Semantic Orientation Of Adjectives, Proceedings Of The Eighth Conference On European Chapter Of The Association For Computational Linguistics, 1997, 174-181.,174–181.
81. Read, J. and Carroll, J., Weakly Supervised Techniques For Domain-Independent Sentiment Classification, Proceedings of the 1st International CIKM Workshop On Topic-Sentiment Analysis For Mass Opinion, November 2009, Hong Kong, China, 45-52.
82. Turney, P.D., Thumbs Up Or Thumbs Down?: Semantic Orientation Applied To Unsupervised Classification Of Reviews, Proceedings Of The 40th Annual Meeting On Association For Computational Linguistics, July 2002, Philadelphia, Pennsylvania, 417-424.
83. Hu, N., Bose, I., Koh, N.S. and Liu, L., Manipulation of Online Reviews: An Analysis of Ratings, Readability, and Sentiments, Decision Support Systems, 52,3 (2012) 674–684.
84. Kim, S.-M. and Hovy, E., Determining The Sentiment Of Opinions, Proceedings of the 20th international conference on Computational Linguistics, August 23, 2004, Geneva, Switzerland, 1367-1370.
85. Porter, M.F., An Algorithm For Suffix Stripping, Program, 14,3 (1980) 314–316.
86. Weiss, S.M., Indurkhya, N. and Zhang, T., Fundamentals Of Predictive Text Mining, Ed. S.F.B. David Gries, 41, Springer, London, 2010.
87. Reese, R.M., Natural Language Processing With Java, Packt Publishing Ltd, 2015.
88. Oflazer, K., Two-Level Description of Turkish Morphology, Literary and Linguistic Computing, 9,2 (1994) 137–148.
89. Kolyiğit, Ö., Türkçe Dokümanlar İçin Yazar Tanıma, Yüksek Lisans Tezi, Adnan Menderes Üniversitesi, Fen Bilimleri Enstitüsü, Aydın, 2013.
90. Kasper, R. and Weber, D., Programmer's Reference Manual for the C Quechua

Adaptation Program, Occasional Publication in Academic Computing, 9 (1982).

91. Koskenniemi, K., A General Computational Model For Word-Form Recognition And Production, Proceedings of the 10th international conference on Computational Linguistics, July 1984, Stanford, California, 178-181.
92. Hankamer, J., Turkish Generative Morphology And Morphological Parsing, Second International Conference on Turkish Linguistics, 1984 Istanbul, Turkey.
93. Solak, A. and Oflazer, K., Design and Implementation of a Spelling Checker for Turkish, Literary and Linguistic Computing, 8,3 (1993) 113–130.
94. Birant, Ç.C., Root-Suffix Seperation of Turkish Words, Yüksek Lisans Tezi, Dokuz Eylül Üniversitesi Fen Bilimleri Enstitüsü, Bilgisayar Mühendisliği, İzmir Türkiye, 2008.
95. Alpkoçak, A., Kut, A. and Özkarahan, E., Bilgi Bulma Sistemleri İçin Otomatik Türkçe Dizinleme Yöntemi, Türkiye Bilişim Derneği, 12. Ulusal Bilişim Kurultayı, Bildiriler Kitabı, 1995, İstanbul, 247-253.
96. Hull, D.A., Stemming Algorithms: A Case Study for Detailed Evaluation, JASIS, 47,1 (1996) 70–84.
97. Banguoğlu, T., Türkçenin Grameri, Türk Dil Kurumu Yayınları, Ankara, 1998.
98. Sever, H. and Bitirim, Y., Findstem: Analysis And Evaluation Of A Turkish Stemming Algorithm, String Processing and Information Retrieval (SPIRE) October 2003, Manaus, Brazil, 238-251.
99. Can, F., Kocberber, S., Balcik, E., Kaynak, C., Ocalan, H.C. and Vursavas, O.M., Information Retrieval on Turkish Texts, Journal of the American Society for Information Science and Technology, 59,3 (2008) 407–421.
100. Çoban, Ö., Metin Sınıflandırma Teknikleri İle Türkçe Twitter Duygu Analizi, Yüksek Lisans Tezi, Atatürk Üniversitesi, Fen Bilimleri Enstitüsü, Erzurum, 2016.
101. Jackson, P. and Moulinier, I., Natural Language Processing For Online Applications: Text Retrieval, Extraction And Categorization, 5, John Benjamins Publishing, 2007.
102. Kesgin, F., Türkçe Metinler İçin Konu Belirleme Sistemi, Yüksek Lisans Tezi, İstanbul Teknik Üniversitesi, Fen Bilimleri Enstitüsü, İstanbul, 2007.
103. Çetin, M. and Amasyalı, M.F., Eğiticili ve Geleneksel Terim Ağırlıklandırma Yöntemleriyle Duygu Analizi, 2013, 21. Sinyal İşleme ve İletişim Uygulamaları, Nisan 2013, Kuzey Kıbrıs Türk Cumhuriyeti.
104. Korde, V. and Mahender, C.N., Text Classification and Classifiers: A Survey,

International Journal of Artificial Intelligence & Applications, 3,2 (2012) 85.

105. Yang, Y. and Pedersen, J.O., A Comparative Study On Feature Selection İn Text Categorization, 1997, ICML ’97 Proceedings of the Fourteenth International Conference on Machine Learning, July 1997, San Francisco, CA, USA, 412-420 .,412–420.
106. Türkmenoğlu, C., Türkçe Metinlerde Duygu Analizi, Yüksek Lisans Tezi, Fen Bilimleri Enstitüsü, İstanbul Üniversitesi, İstanbul, 2015.
107. Rosenfeld, R. and Clarkson, P., Statistical Language Modeling Using the CMU-Cambridge Toolkit, (1997).
108. Ergün, K., Metin Madenciliği Yöntemleri İle Ürün Yorumlarının Otomatik Değerlendirilmesi, Doktora Tezi, Fen Bilimleri Enstitüsü, Sakarya Üniversitesi, Sakarya, 2012.
109. Mert, O., Türk Dili I, Atatürk Üniversitesi Açıköğretim Fakültesi Yayınları, Erzurum, 2016.
110. Korkmaz, Z., Türk Dili ve Arap Alfabesi, Dil ve Alfabe Üzerine Görüşler, (1991) 11–20.
111. Akalın, H., Türk, V., Eker, S. and Demir, S.A., Türk, Dili Kitabı I, Anadolu Üniversitesi Yayınları, Eskişehir, 2012.
112. Alyılmaz, C., Orhun Yazıtlarının Söz Dizimi, Doktora Tezi, Sosyal Bilimler Enstitüsü, Atatürk Üniversitesi, Erzurum, 1994.
113. Coşkun, C. and Baykal, A., Veri Madenciliğinde Sınıflandırma Algoritmalarının Bir Örnek Üzerinde Karşılaştırılması, Akademik Bilişim, Malatya, (2011).
114. Anonim, ROC Eğrisi, https://tr.wikipedia.org/wiki/ROC, 20.05.2017.
115. Onan, A. and Korukoğlu, S., Metin Sınıflandırmada Öznitelik Seçim Yöntemlerinin Değerlendirilmesi, Akademik Bilişim, (2016).
116. Plutchik, R., A General Psychoevolutionary Theory of Emotion, Theories of Emotion, 1,3–31 (1980) 4.
117. Http://emojipedia.org/face-with-Tears-of-Joy, Face With Tears of Joy, 10 Mayıs 2017.
118. Mierswa, I., Wurst, M., Klinkenberg, R., Scholz, M. and Euler, T., YALE: Rapid Prototyping for Complex Data Mining Tasks, Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, August 2006, Philadelphia, PA, USA, 935–940.

6. EKLER

Ek 1. 147 kelimeden oluşan durak kelime listesi

ama	böyle	dolasıyla	her	ki	olmak	sadece	yaptığı
ancak	böylece	edecek	herhangi	kim	olması	şey	yaptığını
arada	bu	eden	herkesin	kimse	olmayan	siz	yaptıkları
ayrıca	buna	ederek	hiç	mı	olmaz	şöyle	yerine
bana	bundan	edilecek	hiçbir	mi	olsa	şu	yine
bazı	bunlar	ediliyor	için	mu	olsun	şunları	yoksa
belki	bunları	edilmesi	ile	mü	olup	tarafından	zaten
ben	bunların	ediyor	ilgili	nasıl	olur	üzere	
beni	bunu	eğer	ise	ne	olursa	var	
benim	bunun	etmesi	işte	neden	oluyor	vardı	
beri	burada	etti	itibaren	nedenle	ona	ve	
bile	çok	ettiği	itibariyle	o	onlar	veya	
bir	çünkü	ettiğini	kadar	olan	onları	ya	
birçok	da	gibi	karşın	olarak	onların	yani	
biri	daha	göre	kendi	oldu	onu	yapacak	
birkaç	de	halen	kendilerine	olduğu	onun	yapılan	
biz	değil	hangi	kendini	olduğunu	öyle	yapılması	
bize	diğer	hatta	kendisi	olduklarını	oysa	yapıyor	
bizi	diye	hem	kendisine	olmadı	pek	yapmak	
bizim	dolayı	henüz	kendisini	olmadığı	rağmen	yaptı	

ÖZGEÇMİŞ

Hasan AMANET, 5 Temmuz 1990 tarihinde İstanbul'da doğdu. Orta öğrenimini Barbaros Hayrettin Paşa Lisesi'nde tamamladıktan sonra 2010 yılında Giresun Üniversitesi Tirebolu Meslek Yüksekokulundan okul birinciliği derecesiyle mezun oldu. Sonrasın 2014 yılında Karadeniz Teknik Üniversitesi İstatistik ve Bilgisayar Bilimleri Bölümünde lisans eğitimini tamamlayarak mezun oldu. 2015 yılında Karadeniz Teknik Üniversitesi, Fen Bilimleri Enstitüsü, İstatistik ve Bilgisayar Anabilim Dalı'nda tezli yüksek lisans programına başladı.