T.C.

TRAKYA ÜNİVERSİTESİ

FEN BİLİMLERİ ENSTİTÜSÜ

FUTBOLA İLİŞKİN TWITTER PAYLAŞIMLARININ DUYGU ANALİZİ

RIZA KORKUSUZ

YÜKSEK LİSANS TEZİ

BİLGİSAYAR MÜHENDİSLİĞİ ANA BİLİM DALI

TEZ DANIŞMANI: Dr. Öğretim Üyesi Aydın CARUS

EDİRNE-2019

Rıza KORKUSUZ 'un hazırladığı "Futbola İlişkin Twitter Paylaşımlarının Duygu Analizi" başlıklı bu tez, tarafımızca okunmuş, kapsam ve niteliği açısından Bilgisayar Mühendisliği Anabilim Dalında bir Yüksek Lisans tezi olarak kabul edilmiştir.

Jüri Üyeleri (Ünvan, Ad, Soyad):	İmza
Dr. Öğretim Üyesi Özlem AYDIN	.....
Dr. Öğretim Üyesi Edip Serdar GÜNER	.....
Dr. Öğretim Üyesi Aydın CARUS	.....

Tez Savunma Tarihi: 17/07/2019

Bu tezin Yüksek Lisans tezi olarak gerekli şartları sağladığını onaylarım.

İmza

Dr. Öğretim Üyesi Aydın CARUS
Tez Danışmanı

Trakya Üniversitesi Fen Bilimleri Enstitüsü onayı

Prof. Dr. Murat YURTCAN
Fen Bilimleri Enstitüsü Müdürü

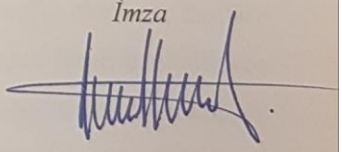
T.Ü. FEN BİLİMLERİ ENSTİTÜSÜ
BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI
YÜKSEK LİSANS PROGRAMI
DOĞRULUK BEYANI

Trakya Üniversitesi Fen Bilimleri Enstitüsü, tez yazım kurallarına uygun olarak hazırladığım bu tez çalışmasında, tüm verilerin bilimsel ve akademik kurallar çerçevesinde elde edildiğini, kullanılan verilerde tahrifat yapılmadığını, tezin akademik ve etik kurallara uygun olarak yazıldığını, kullanılan tüm literatür bilgilerinin bilimsel normlara uygun bir şekilde kaynak gösterilerek ilgili tezde yer aldığını ve bu tezin tamamının ya da herhangi bir bölümünün daha önceden Trakya Üniversitesi ya da farklı bir üniversitede tez çalışması olarak sunulmadığını beyan ederim.

17/07/2019

Rıza KORKUSUZ

İmza



Yüksek Lisans Tezi
Futbola İlişkin Twitter Paylaşımlarının Duygu Analizi
T.Ü. Fen Bilimleri Enstitüsü
Bilgisayar Mühendisliği Anabilim Dalı

ÖZET

İnternetin ve sosyal medyanın hızlı bir şekilde yaygınlaşıp kabul görmesi, önceden bir konu hakkında pek mümkün ve kolay olmayan görüş bildirimini son kullanıcılar için de mümkün kılmıştır. Facebook, Instagram ve Twitter gibi sosyal medya platformları son yıllarda giderek daha popüler hâle gelmiştir. Sosyal medya paylaşımları işlenip duygu bakımından analiz edilerek ilgili konu hakkında kişilerin duyguları ortaya çıkarılabilmektedir.

Bu çalışmada bir sosyal medya platformu olan Twitter üzerinde paylaşılan futbol müsabakaları hakkındaki Türkçe görüşler üzerinden duygu analizi yapılmıştır. Twitter üzerinde futbol müsabakası hakkında paylaşılmış olan Türkçe tweetler alınarak el ile etiketlemesi yapılmıştır. Bu etiketlemede duygu sınıfı olarak olumlu, olumsuz, tarafsız ve alakasız olarak 4 sınıf kullanılmıştır. Bu etiketlenmiş tweetlerden farklı özniteliklere sahip 12 farklı eğitim seti oluşturulmuştur. Oluşturulan bu eğitim setleri kullanılarak farklı sınıflama algoritmaları ile modeller çıkarılmış ve bu modellerin çapraz doğrulama ile sınıflama başarımları bulunmuştur. Eğitim setlerindeki öznitelikseL farklılıklara bağlı olarak bu modeller için yapılan testlerde en iyi sınıflama başarımının; Naive Bayes algoritması için %84.30, KNN algoritması için %87.73, C4.5 algoritması için %89.60, SVM algoritması için %92.30 olduğu tespit edilmiştir. Çalışma sonucunda Türkçe tweetlerde imla ve anlama dayalı yazım denetimi yapılmasının sınıflama başarımında önemli oranda artış sağladığı görülmüştür.

Ayrıca sınıflama için elde edilmiş olan bu modeller içinden en başarılı modelleri kullanan bir uygulama hazırlanmıştır. Bu uygulama futbol müsabakaları hakkında paylaşılan tweetleri anlık olarak toplayıp, oluşturulan modeller ile bu tweetlerin sınıfını tespit ederek sınıflara ait sonuçları gerçek zamanlı görselleştirmektedir.

Yıl : 2019

Sayfa Sayısı : 102

Anahtar Kelimeler : Büyük Veri, Veri Madenciliği, Makine Öğrenmesi, Twitter, Anlık Duygu Analizi

Master Thesis
Sentiment Analysis of Twitter Shares on Football
Trakya University Institute of Naturel Sciences
Computer Engineering Department

ABSTRACT

Facebook, Instagram, and Twitter social media platforms have become increasingly popular in recent years. The widespread use and acceptance of social media have made it possible and easier for end-users to share opinions on a topic. Social media posts can be processed and analyzed to reveal the emotions of people on a certain topic.

In this study, emotion analysis was conducted based on Turkish posts about the football competitions shared on Twitter. Turkish tweets about the football competitions were collected and tagged manually. For this labeling, 4 classes were utilized as emotion classes: positive, negative, neutral and irrelevant.

Twelve different training sets with different features were built based on these tagged tweets. Using these training sets, models were created with different classification algorithms and the classification performances of these models were evaluated by cross-validation. Depending on the qualitative differences in the training sets, the classification accuracies in the tests for these models were found as: 84.30% for the Naive Bayes algorithm, 87.73% for the KNN algorithm, 89.60% for the C4.5 algorithm, and 92.30% for the SVM algorithm.

As a further result of this study, it is demonstrated that checking and correcting the spelling of the tweets provides a significant increase in classification performance. In addition, an application which employs the most successful models of these

classification models was created. This application collects tweets about football competitions instantly, determines the class of these tweets based on the models and visualizes the results of the classes in real time.

Year : 2019

Number of Pages : 102

Keywords : Big Data, Data Mining, Machine Learning, Twitter, Real Time Sentiment Analysis

TEŞEKKÜR

Çalışmalarımı yönlendiren ve araştırmalarımın her aşamasında bilgi, öneri ve deneyimleri ile yol gösterip yardımlarını esirgemeyen danışmanım Dr. Öğretim Üyesi Aydın CARUS'a ve (Kırklareli Üniversitesi Yazılım Mühendisliği Bölümü) Dr. Öğretim Üyesi Edip Serdar GÜNER'e teşekkürlerimi sunarım.

Kodlama konusunda deneyim ve katkılarından dolayı arkadaşım Enes GÜNCE'ye teşekkürü bir borç bilirim.

Çalışmalarım esnasında gösterdiği büyük sabırdan dolayı eşim Meryem TOPÇU KORKUSUZ ve Oğlum Doruk Ata KORKUSUZ'a içtenlikle teşekkür ederim.

Rıza KORKUSUZ

Edirne, 2019

İÇİNDEKİLER

ÖZET iv
ABSTRACT vi
TEŞEKKÜR viii
İÇİNDEKİLER ix
SİMGELER DİZİNİ xi
ÇİZELGELER DİZİNİ xii
ŞEKİLLER DİZİNİ xiii
BÖLÜM 1 1
GİRİŞ 1
1.1. Duygu Analizi ve Twitter 1
1.2. Arka Plan 3
1.3. Sorun bildirimi 4
1.4. Amaç ve Yaklaşım 4
1.5. Varsayımlar 5
1.6. Tez Organizasyonu 6
BÖLÜM 2 8
LİTERATÜR İNCELEMESİ 8
2.1. Genel Bakış 8
2.1.1. Doğrulama 9
2.2. Sosyal Medya 10
2.2.1. Kullanılabilirlik 10
2.2.2. Sosyal Medyanın Faydaları 10
2.2.3. Sosyal Medyanın Sorunları 11
2.2.3.1. Veri ve Veri İşleme Ortamının Önemi 11
2.2.3.2. Veri Kaynağı 12
2.3. Duygu Analizi 12
2.4. Doğal Dil İşleme 13
2.5. Uygulama 13
2.6. Twitter 14
2.6.1. Twitter Duygu Analizi Uygulamaları 14
2.6.1.1. Zorluklar 16
2.6.1.2. Sınıflandırma 19
2.6.1.3. Özellik Seçimi 19
BÖLÜM 3 21

METODOLOJİ 21
3.1. Önerilen Mimari 21
3.2. Twitter API 25
3.2.1. Stream API Kullanarak Twitter'dan Veri Akışı 25
3.3. Tweet Toplama 26
3.4. Veri Ön İşleme 30
3.5. Doğal Dil İşleme Kütüphanesi (Zemberek) 32
3.5.1. Yazım Denetimi 32
3.6. El ile Etiketleme 33
3.7. Öznitelik Seçimi 37
3.8. Oluşturulan Veri Setleri 39
BÖLÜM 4 48
MAKİNE ÖĞRENMESİ 48
4.1. Makine Öğrenmesi 48
4.1.1. Denetimli Makine Öğrenmesi Algoritmaları 49
4.2. Weka 57
4.2.1. StringToWordVector Filtresi 59
4.2.2. Bilgi Kazancına Dayalı Özellik Seçimi 60
4.2.3. Çapraz Doğrulama 61
4.2.4. Duygu Sınıflandırılması Başarımının Değerlendirilmesi 64
4.2.4.1. Doğruluk, Duyarlılık, Geriçağırım ve F-ölçütü 64
4.2.4.2. Kappa İstatistiği 66
4.2.4.3. Karışıklık Matrisi 66
BÖLÜM 5 68
BAŞARIM TESTLERİ 68
5.1. Model Eğitimi 68
5.2. Makine Öğrenmesi Algoritmalarının Başarımı 68
5.3. Anlık Twitter Duygu Analizi 77
BÖLÜM 6 80
SONUÇLAR VE TARTIŞMA 80
6.1. Katkıları 81
6.2. Gelecek Çalışma Önerileri 81
KAYNAKLAR 83
EKLER 86
EK-1 Arff Dosyası İçerik Örneği 86
EK-2 Gereksiz Kelimeler 88
ÖZGEÇMİŞ 89

SİMGELER DİZİNİ

Kısaltmalar

ARFF	Özellik-İlişki Dosyası Biçimi (Attribute-Relation File Format)
CSV	Virgülle Ayrılmış Değer (Comma Seperated Value)
SVM	Destek Vektör Makinesi (Support Vector Machine)
KNN	En Yakın k Komşu (K Nearest Neighbors)
Twitter API	Twitter Uygulama Programlama Arayüzü (Application Programming Interface)
Streaming API	Akış Uygulama Programlama Arayüzü (Streaming Application Programming Interface)
Rest API	Temsili Durum Transferi Uygulama Programlama Arayüzü (Representational State Transfer Application Programming Interface)
F1	F-ölçütü (F-Measure)
FN	Yanlış Negatif (False Negative)
FP	Yanlış Pozitif (False Positive)
TP	Doğru Pozitif (True Positive)
TN	Doğru Negatif (True Negative)
NLP	Doğal Dil İşleme (Natural Language Processing)

ÇİZELGELER DİZİNİ

Çizelge 3.1. Tweet içinden seçilen alanların açıklamaları 29
Çizelge 3.2. Etiketleme işleminden önce yapılan veri ön işleme 31
Çizelge 3.3. Etiketleme işleminden sonra yapılan veri ön işleme 31
Çizelge 3.4. Etiketleme işlemi öncesi ham tweet ve temizlenmiş tweet örneği 32
Çizelge 3.5. Zemberek Kütüphanesi kullanılarak yazım denetimi 33
Çizelge 3.6. Kelimelerin köklerine indirgenmesi 37
Çizelge 3.7. Ham tweet ve etiketleme işleminden önce veri ön işleme yapılan tweet 40
Çizelge 3.8. EğitimSeti-1 arff dosya formatı bir örnek satırı 40
Çizelge 3.9. EğitimSeti-2 arff dosya formatı bir örnek satırı 41
Çizelge 3.10. EğitimSeti-3 arff dosya formatı bir örnek satırı 41
Çizelge 3.11. EğitimSeti-4 arff dosya formatı bir örnek satırı 42
Çizelge 3.12. EğitimSeti-5 arff dosya formatı bir örnek satırı 42
Çizelge 3.13. EğitimSeti-6 arff dosya formatı bir örnek satırı 43
Çizelge 3.14. EğitimSeti-7 arff dosya formatı bir örnek satırı 43
Çizelge 3.15. EğitimSeti-8 arff dosya formatı bir örnek satırı 44
Çizelge 3.16. EğitimSeti-9 arff dosya formatı bir örnek satırı 44
Çizelge 3.17. EğitimSeti-10 arff dosya formatı bir örnek satırı 45
Çizelge 3.18. EğitimSeti-11 arff dosya formatı bir örnek satırı 45
Çizelge 3.19. EğitimSeti-12 arff dosya formatı bir örnek satırı 46
Çizelge 3.20. Etiketleme işleminden sonra dosya isimleri ve işlenmiş tweet içerikleri 47
Çizelge 5.1. Naive Bayes (kernelEstimator = false) sınıflandırma algoritması sonuçları 69
Çizelge 5.2. Naive Bayes (kernelEstimator = true) sınıflandırma algoritması sonuçları 70
Çizelge 5.3. SVM sınıflandırma algoritması sonuçları 71
Çizelge 5.4. Ağaç C4.5 sınıflandırma algoritması sonuçları 72
Çizelge 5.5. IBk (k=1) sınıflandırma algoritması sonuçları 73
Çizelge 5.6. IBk (K=2) sınıflandırma algoritması Sonuçları 74
Çizelge 5.7. IBk (k=3) sınıflandırma algoritması sonuçları 75
Çizelge 5.8. Sınıflandırma algoritmaları doğruluk başarımları 76
Çizelge 5.9. Sınıflandırma algoritmalarına göre en iyi performans gösteren veri setleri 77

ŞEKİLLER DİZİNİ

Şekil 3.1. Twitter'dan tweet toplama ve model oluşturma sistem mimarisi 22
Şekil 3.2. Tweet toplama ve model oluşturma ayrıntıları 23
Şekil 3.3. Anlık duygu analizi sistem mimarisi 24
Şekil 3.4. Tweet toplamak için geliştirilen uygulama arayüzü 28
Şekil 3.5. Veri etiketleme arayüzü 35
Şekil 3.6. Etiketleme işleminden sonra dengeli veri sınıfları 36
Şekil 4.1. SVM algoritması kesik çizgiler ve düz çizgi 51
Şekil 4.2. SVM algoritması doğrusal ayırma çizgi 52
Şekil 4.3. SVM algoritması eğri çizgi 52
Şekil 4.4. SVM algoritması eğriyi düzenleme işlemi 52
Şekil 4.5. Kırmızı ve yeşil sınıfların dağılımı 53
Şekil 4.6. X verisinin sınıflar içindeki konumu 54
Şekil 4.7. K = 5 ile KNN = bilinmeyen veri mavi sınıfa atanmıştır. 56
Şekil 4.8. KNN (k=1, k=2, k=3) algoritması 57
Şekil 4.9. Weka'da yapılan işlem basamakları 59
Şekil 4.10. Sınıflandırma algoritmaları ve çapraz doğrulama işlemi 62
Şekil 4.11. 10 Katlı Çapraz Doğrulama 63
Şekil 4.12. İki yönlü sınıflandırma için bir karışıklık matrisi 67
Şekil 5.1. Anlık duygu analizi döngüsü 78
Şekil 5.2. Anlık duygu analizi sonuçları 79

BÖLÜM 1

GİRİŞ

1.1. Duygu Analizi ve Twitter

Günümüzde sosyal medya platformları, bilgiyi elde etme ve paylaşma yönüyle insan yaşamının bir parçası hâline gelmiştir. Bu platformlar her geçen gün daha fazla kullanıcı sayısına erişmektedir. Kullanıcılar çeşitli şekillerde (metin, resim, ses, vb.) bu ortamlarda veri paylaşmaktadır. Sosyal medya platformları aracılığı ile bireyler kendi sanal profillerini oluşturmaktadır. Twitter, Facebook, Instagram, LinkedIn gibi sosyal medya platformlarında, bireyler yaşam tarzlarını, fikirlerini ve diğer bazı kişisel bilgilerini paylaşarak birbirleriyle etkileşime girmektedir. Bunların en popülerlerinden biri, kullanıcılarına en fazla 280 karakterden oluşan metin girişine olanak tanıyan Twitter'dır.

Bir sosyal medya kullanıcısı Twitter'da ilgi duyduğu alanlarda ne paylaşıldığını bilmek ister ve bunun için bir takip listesi oluşturur. Ayrıca herhangi bir konu hakkındaki fikirlerini ve yaşamlarından kesitleri de bu platformda paylaşabilmektedir. Bu şekilde, kullanıcılar hesapları üzerinden Twitter ortamını takip ederek ilgili olduğu konular hakkında paylaşılanları görebilir. Kullanıcılar, bu konular hakkında kendi paylaşımları ile duygusal durumunu ifade edebilir. Bu paylaşımlardan çeşitli amaçlar için duygu analizi yapılabilmektedir. Bu analizlerin sonuçlarından; kişiler, olaylar, ürünler, markalar, müsabakalar vb. gibi daha birçok varlık için paylaşım yapanların fikirleri hakkında bilgi sahibi olunabilir. Bu amaçla bazı firmalar kendilerini ilgilendiren Twitter paylaşımlarını dinleyerek piyasa eğilimlerine karşı cevap ve firmalarına yön verir (Ficamos, Liu, 2016).

Pew Araştırma Merkezi'ne göre, internet kullanıcılarının yaklaşık dörtte biri Twitter kullanmaktadır (Duggan, Page, Manager, 2019). Kullanıcı sayısı göz önünde bulundurulursa Twitter'ın 280 karakter ile kısıtlı olan mesaj uzunluğunun kullanıcılarını rahatsız etmediği söylenebilir. Firmalar ürünleri hakkında müşterilerini sosyal medya platformlarından bilgilendirme yolunu kullanmaktadırlar. Örneğin, Dell firması müşterilerini ürün indirimleri konusunda bilgilendirmek için Twitter ortamını başarıyla kullanabilmiştir. Ayrıca, şirketler Twitter paylaşımları üzerinden duygu analizi yaparak tüketicilerinin ürünleri hakkında ne söylediklerini kolayca belirleyebilmektedir (Liu, Cheung, Lee, 2010).

Sosyal medya platformları; ürünler hakkında yorum, herhangi bir konu hakkında kullanıcı görüşleri vb. gibi bilgiye dönüştürülebilecek birçok ham veriye sahiptir. Sosyal medya platformlarındaki büyüme nedeniyle özellikle büyük firmalar duygu analizi üzerinde durmakta ve bu alanda çalışmaktadır. Duygu analizinin temel amacı, kişinin bir konu hakkındaki paylaşımından olumlu, olumsuz veya tarafsız gibi bir sonuca varmaktır. Tüketiciler ürün ve ürün fiyatı, ürünü satan güvenilir satıcı vb. konular hakkında araştırma yapmak için sosyal medya platformunu kullanırlar. Ayrıca firmalar, tüketicilerinin, ürünleri hakkındaki görüşlerini bilmek ister ve ürünleri hakkında görüşleri öğrenmek için duygu analizleri yapmaktadır. Pazarlamacılar, duygu analizini etkin kullanarak firma ve ürünleri hakkında pazar araştırması yaparlar (Kethavath, 2015).

Twitter, internette bilgi paylaşmanın en güçlü platformlarından biridir. 2017 yılının dördüncü çeyreğinde aylık 330 milyon aktif kullanıcıya sahip ve Alexa sıralamasına göre dünyanın en çok ziyaret edilen 12. sıradaki web sitesi olmuştur. Twitter; tüm kullanıcılarına her türlü kültürel, politik, ekonomik, sportif ve diğer birçok alandaki olaya ilişkin görüşlerini, fikirlerini, yargılarını, kaygılarını ve tutumlarını ifade etme fırsatı sunmaktadır (Khritantsev, Shehadeh, 2018).

Bazı duygu analizi çalışmaları; hisse senedi fiyatları, sinema gişesi gelirleri ve siyasi seçimler gibi sosyal olayların sosyal medya verilerinden etkilendiğini iddia etmektedir (Bollen, Mao ve Zeng, 2011; Asur ve Huberman, 2010; Gayo-Avello, 2013). Bu platformlarda ifade edilen görüşlerin kamuoyunun fikrini dolaylı olarak anlamak için kullanılabileceği söylenmektedir (O'Connor, Balasubramanyan, Routledge ve

Smith, 2010).

Duygu analizi aynı zamanda; tüketicilerin ürünler hakkındaki görüşleri, seçmenlerin siyasi partilere karşı tutumları veya yatırımcıların hisse senetleri hakkındaki beklentileri ile de ilgilidir. Duygu analizi farklı platformlardan haberler, bloglar, Instagram yorumları, Facebook yorumları, Twitter gönderileri, vb. elde edilen veriler üzerinde de yapılabilir. Duygu analizi için genel olarak iki yaklaşım vardır: (i) sözlük tabanlı duygu analizi ve (ii) makine öğrenmesi tabanlı duygu analizi. Sözlük tabanlı duygu analizi, metindeki duygu içeren kelimelerden hesaplanır ve metinde tanımlanan duyguları ifade eden kelimelerden duygu analizi yapılır. Makine öğrenmesi tabanlı duygu analizinde ise ilk olarak etiketli verilerin oluşturduğu geniş bir kümeden bir duygu modeli oluşturulur ve daha sonra sınıfı bilinmeyen verinin model aracılığı ile sınıfı tahmin edilir. Model, veriden çıkarılan öznitelikleri kullanarak duygu etiketlerini olumlu, olumsuz veya tarafsız olarak tahmin eder. Bu duygu sınıfları, paylaşanın mesaj içeriğinde ele alınan konuya karşı duygusunu ifade eder.

Her iki yaklaşımda, duygu analizi için tek tek kelimelerle veya kısa metinlerle ifade edilen duygular etiketlenmelidir. Bu duygu etiketlemesi; dil, etki alanı ve çoğu zaman konuya özeldir (Mozetič, Grčar, Smailović, 2016). Duygu analizi, metinlerin anlamsal analiz biçimidir. Amacı, konulara yönelik olarak görüş, duygu veya tutumları ortaya çıkarmaktır (Liu, 2015)(Liu, 2012).

Bu çalışma, Twitter'da futbol müsabakaları hakkında paylaşılan tweetler üzerinden, paylaşanların duygularını analiz yoluyla anlık olarak saptamak için gerçekleştirilmiştir. Çalışmada ayrıca paylaşılan tweetlerden anlık olarak tahmin edilen duyguların görselleştirilmiş olarak gösterimi de yapılmaktadır.

1.2. Arka Plan

Son yıllarda, kullanıcıların bireysel görüş ve düşüncelerini paylaşabildiği sosyal medya platformları hızlı bir şekilde büyümüştür. Sosyal medya platformları her kullanıcının dünya çapında görüşlerini dile getirmesi için etkileşimli bir yapıya sahiptir. Paylaşımların niteliği, konuları ve tanımlama biçimleri bölgeden bölgeye değişebilir (Boyd, Ellison 2007). Kullanıcılar tarafından oluşturulan içeriğin bu yeni boyutu; bireyler, gruplar, firmalar, hükümetler, vb. için yeni ve heyecan verici bir ilgi alanı

yaratmaktadır (Giachanou, Crestani 2016).

Sosyal medya platformları, bireylere aşağıda belirtilen web tabanlı hizmetleri sağlar:

- Farklı nitelikte profil oluşturma (açık, kapalı),
- Bir paylaşımın kısıtlanması,
- Paylaşılanları görüntüleyebilme.

1.3. Sorun bildirimi

Sosyal medya platformlarında kullanıcı sayısındaki artış gün geçtikçe; metin, ses, video ve görüntüler şeklinde büyük miktarda veri üretmektedir. Duyguların olumsuz, olumlu veya tarafsız olup olmadığını bulmak için duygu analizi yapılması gerekir.

Duygu analizi, bir insan topluluğunun bir konudaki görüşünü ortaya koyabilmek için bir bakış açısı sağlayabilir. Bu analiz, şirketlere hedef tüketicilerini daha iyi anlamalarını sağlayarak yardımcı olabilir. Benzer şekilde elde edilen bilgi, hükümetlerin bir toplumu daha iyi anlamasını ve böylece o toplum için daha doğru kararlar vermelerini sağlayabilir.

Sosyal medya platformlarında paylaşılan çok sayıda veri olmasından dolayı meşru ve meşru olmayan verileri ayırt etmek çok zordur. Sahte görüş, spam yayın, viral mesajlar, vb. kullanıcıların itibarını tanımlamaya ilişkin bu tür durumları tespit edebilecek duygu analizinde herhangi bir standart yoktur. Ayrıca sürekli aynı paylaşımların retweet yapılması, yeni paylaşımların olmaması, farklı içerik zenginliğini ortadan kaldırmaktadır.

Bu çalışmada da futbol takımların ne kadar beğenildiğini veya beğenilmediğini ortaya koymak için yaygın olarak kabul görmüş olması nedeniyle Twitter paylaşımlarından futbol müsabakaları hakkında kişilerin paylaşımlarından duygu analizi ile görüşlerinin ortaya çıkarılması hedeflenmiştir.

1.4. Amaç ve Yaklaşım

Konu odaklı olarak futbol müsabakaları ile ilgili duyguları ortaya çıkarabilmek amacıyla Twitter API (Application Programming Interface) aracılığıyla anahtar

kelimelere göre filtreleme ile anlık olarak tweetleri toplamak için C# .NET programlama dili kullanılarak bir uygulama geliştirilmiştir. Bu sayede olumlu, olumsuz, tarafsız ve alakasız duygu sınıflarını içeren eğitim setlerini oluşturmak için futbol maçları oynandığı sırada gerçek zamanlı olarak taraftarların attığı tweetler toplanmıştır. Toplanan bu tweetlerin geliştirilen bir başka arayüz kullanılarak el ile sınıfları belirlenerek etiketlemesi yapılmıştır. Makine öğrenmesi algoritmaları ile modeller oluşturmuş ve modeller aracılığı ile anlık olarak paylaşılan yeni tweetlerin sınıflaması yapılarak sonuçlar görsel olarak ifade edilmiştir.

Twitter paylaşımlarının duygu analizi için 6 adımlı bir süreç izlenmiştir. Bu adımlar:

- Tweet toplama,
- Tweet ön işleme (ayrıştırma, temizleme, denetim işlemleri, öznitelik çıkarımları),
- Eğitim seti oluşturma (etiketleme, eğitim kümesi olarak dosya oluşturma),
- Modellerin çıkarılması,
- Oluşturulan modeller ile anlık tweet duygu tahmini,
- Sonuçların görselleştirilmesi.

Öncelikle tweet ön işlemi sürecinde tweetler; yazım hataları, kısaltmalar, ifadeler, bağlantılar, resimler içerebileceği için geliştirilen uygulama ile bu verilerin kullanılabilir hâle getirilmesi sağlanmıştır.

1.5. Varsayımlar

Twitter'da, Twitter API aracılığıyla tweet toplandığından normal olarak gürültü içeren veriler de toplanmaktadır. Tweetler aşağıda belirtilen şekilde görünebilecek kullanıcı adlarını içermektedir: @mr_hasan. Tweetler herhangi bir sayıda #hashtag ve diğer öğeleri de içerebilir. Hashtag, başkalarının belirli bir temayı kolayca referans alması veya devam ettirmesi için içeriğe kolay erişime izin veren, dinamik olarak

kullanılan etikettir. Analiz sırasında "@" ve "#" sembolleri bilgiyi korumak için kaldırılmıştır. Bu yapılan temizliğin duygu analizinin sonucunu değiştirmeyeceği varsayılmıştır.

Twitter ortamında, "RT" harfleri, bir tweetin retweet olduğunu belirtir. Retweet, başka bir kullanıcıdan gelen bir tweetin tekrar paylaşıldığı anlamına gelir. Bu çalışmada retweet edilen tweetler duygu analizine dahil edilmiştir. Çünkü diğer kişilerin orijinal mesajı paylaşmayı kabul ettiklerini, istediklerini, onayladıklarını veya aynı duyguyu paylaştıklarını belirtir ve bu nedenle ilgili duygular bir defa daha tekrar etmiş olur. Bununla birlikte, tweette yer alan medya içerik linkleri temizlenmiştir.

1.6. Tez Organizasyonu

Sosyal medya platformu olan Twitter; politikacılar, spor kulüpleri, şirketler, aktivistler, ünlüler, kısaca tüm bireyler ve kurumlar için kendilerini ifade etme ortamı hâline gelmiştir. Neredeyse tüm futbol takımlarının taraftarları için haberlerini paylaştıkları Twitter hesapları vardır. Twitter hesabı olan kullanıcılar haber paylaşımlarına cevap verebilir veya haber paylaşımlarını retweet edebilir. Hesabı olan Twitter kullanıcıları, başka herhangi bir futbol takımı hakkında mesaj gönderebilirler. Kullanıcılar Twitter'da futbol müsabakaları hakkında mesaj paylaşabilir. Birbirlerinin futbol maçları hakkındaki paylaşımlarını retweet ettiklerinde, aynı duyguları birlikte ifade edebilmeleri de mümkündür. Bu çalışmanın konusu, Twitter ortamında futbol takımları ve futbol müsabakaları hakkında Twitter kullanıcılarının paylaşımlarının duygu analizi ile ilgilidir.

Bu çalışma da futbol müsabakaları hakkında Twitter paylaşımlarından anlık duygu analizi için model bir uygulama sunmuştur. Geliştirilen uygulama, filtrelenerek seçilen tweetlerin analiz edilmesine olanak sağlamıştır. Duygu analizi, toplanan Twitter verilerinde #hashtag ve belirlenen konu başlıklarına da uygulanabilir. Bu uygulama, Twitter kullanıcılarının görüş ve düşüncelerini belirlemek amacıyla duygunun anlık belirlenmesini ve görselleştirilmesini sağlamıştır.

Genel olarak bu çalışmada, anlık olarak futbol maçları oynanırken taraftarların paylaştığı tweetler toplanmıştır. Toplanan tweetler geliştirilen uygulama yardımıyla el ile etiketlenmiştir. Etiketlenen tweetlerden 12 farklı eğitim seti arff (Attribute

Relationship File Format) dosya şeklinde oluşturulmuştur. Oluşturulan 12 farklı arff dosyasına Weka makine öğrenmesi aracı ile ön işlem uygulanarak, makine öğrenmesi algoritmalarından Naive Bayes sınıflandırıcısı, SVM sınıflandırıcısı, C4.5 sınıflandırıcısı ve KNN sınıflandırıcısı kullanılarak Twitter duygu analizi yapılmıştır. Veri setleri için algoritmalar ile elde edilen modellerin sınıflama başarımları çapraz doğrulama ile test edilerek değerlendirilmiştir.

Mayıs 2019 tarihine kadar Türkiye'deki lisansüstü tez çalışmalarında, futbol konusuna özel, anlık Twitter duygu analizi gerçekleştiren çalışmaya rastlanmamıştır. Ayrıca Türkçe yapılan tez çalışmaları incelendiğinde duygu analizinin olumlu, olumsuz ve tarafsız olarak 3 sınıflı yapıldığı görülmüştür. Paylaşılan tweetler üzerinden duygu analizi yapan bu çalışmada olumlu, olumsuz, tarafsız ve alakasız olarak dört sınıf mevcuttur. Bu tez aşağıdaki gibi yapılandırılmıştır:

Giriş bölümünde duygu analizi, sosyal medya ve Twitter platformuna değinilmiş ve Twitter paylaşımlarından makine öğrenmesi algoritmaları kullanılarak yapılan çalışmanın amacı anlatılmıştır.

İkinci bölümde, duygu analizi ve sosyal medya hakkında literatür incelemesi yapılmıştır. Sosyal medya platformlarında duygu analizi alanında şu ana kadar neler yapıldığı incelenmiştir.

Üçüncü bölümde; Twitter platformundan anlık veri toplayan ve bu verileri kaydedip temizlemek, etiketlemek gibi çeşitli ön işlem ve sonrası işlemleri yerine getirmek için geliştirilen uygulama anlatılmıştır.

Dördüncü bölümde; toplanan tweetlerin etiketlendikten sonra öznitelik çıkarımı, oluşturulan arff dosyalarının yapıları ve birbirlerinden farklarından bahsedilmiştir.

Beşinci bölümde, makine öğrenmesi için geliştirilmiş olan Weka yazılım aracı kullanılarak oluşturulan eğitim setlerinin makine öğrenmesi algoritmalarına bağlı sınıflama başarımları analiz edilmiş ve elde edilen sonuçlar birbiriyle kıyaslanmıştır.

Çalışmanın sonuçlar ve tartışma bölümünde sonuçlar değerlendirilerek gelecekte yapılacak çalışmalar hakkında önerilerde bulunulmuştur.

BÖLÜM 2

LİTERATÜR İNCELEMESİ

2.1. Genel Bakış

Bu bölümde, duygu analizi çalışmaları incelenmiştir. Çeşitli alanlarda duygu analizi kullanılarak yapılmış birçok çalışma mevcuttur. Bu alanlardan birisi de sosyal medya platformları üzerinden duygu analizi yapılmasıdır. Sosyal medya platformları üzerinden duygu analizi yapılabilmesi için genellikle sözlük tabanlı, kural tabanlı ve makine öğrenmesi tabanlı yöntemler kullanılmaktadır.

Duygu analizi, bir bireyin ilgi duyduğu bir konu hakkında genel hissiyatını ortaya çıkartmak için bir yöntemdir. Duygu analizi sadece insanların ne söylediklerini değil, insanların söylediklerinin ne anlama geldiğini anlamaya çalışır (Newman, 2018). Sosyal medya platformlarında duygu analizi, anlaşılması zor olan verilerden anlamlı bilgilerin çıkarımını sağlar (Bertot, Jaeger, Hansen, 2012). Sosyal medya platformları, bireylerin duygularını gözlemlemeyi mümkün kılmıştır. Kullanıcılar görüşlerini bu platformlardan paylaşmaktadırlar. Paylaşımlarını tüm kullanıcıların erişimine açık hâle getirebilirler.

Sosyal medya platformları, özgürce yapılan paylaşımlardan duyguları ortaya çıkarmak ve analiz etmek için bilgi ve uzmanlığa sahip olan kullanıcılar için bir fırsattır (Pang, Lee, 2008). Geleneksel yöntemler ile duygu analizi çalışmalarının zorlukları; kısıtlı erişiminin olması, olaylar geliştikçe fark edilmesinin güç olması ve olaylara müdahale edilmesinin zor olmasıdır. Sosyal medya platformlarını kullanarak duygu analizi yapmak, bir bireyin veya toplumun geleneksel yöntemlerle –örneğin; telefon ve posta anketleri ile- toplanamayan görüş ve fikirleri hakkında daha çok bilgi edinmemizi sağlayabilir (Kavanaugh, vd, 2012). Duygu analizi, gerçek zamanlı yaşanan olaylarla

ilgili üretilen veriyi anlamak için uygulanabilir. Örneğin; bunu etkili bir şekilde gerçekleştirerek hükümetler, toplumu ilgilendiren kritik konular ile ilgili kamu güvenliğini sağlamak için önlem alabilirler (Kavanaugh, 2012).

2.1.1. Doğrulama

Costa ve Boiney bir çalışmasında devletlerin sosyal medya platformlarını güvenilir bir şekilde izlemeleri, olaylar meydana gelmeden önce değişimi öngörmeleri ve olaylara uygun eylemler için plan yapmaları gerektiğini savunmuştur (Costa, Boiney, 2012).

Sosyal medya platformlarında bireyleri ve toplumu anlamak için duygu analizi çalışmaları her geçen gün önem kazanmaktadır (Costa, Boiney, 2012). Duygu analizi çalışmaları kültürlerin, toplumların, grupların, dinin ve ideolojinin bir toplumu nasıl etkilediği ile ilgilenmektedir. Toplum anlayışını yakalamak için duygu analizi çok önemlidir (Costa, Boiney, 2012). Bu anlamda duygu analizi bu amaçlar için kullanılır:

- Önemli toplumsal olayları fark etmek ve hızlı bir şekilde duyguyu tespit etmek,
- Bireylerin ve toplumun bir konu hakkında ilgili paylaşımlarını ve duygunun zamansal olarak etkisini tespit etmek ve ölçmek.

Bir toplum hakkında sosyal medya platformlarını kullanarak yapılan paylaşımları toplayarak duygu analizi yapılabilmektedir. Twitter duygu analizi; bir protestonun duygu durumunu, duygunun etkisini ve hangi tür duyguların ifade edildiğini belirleyecek şekilde ayrıntılı bir analiz yapmak için kullanılabilir. Twitter duygu analizi, Arap Baharı örneğinde olduğu gibi bir protestonun büyüyüp büyümeyeceğini veya ülke geneline yayılıp yayılmayacağını belirlemek için kullanılabilir (Aday, Farrell, Lynch, Sides, Freelon, 2012).

Duygu Analizi ile Twitter #hashtag konulara yönelik olarak duygu durumlarını göstermek için kullanılabilir. Ayrıca, kullanıcının kimliği ve konum bilgisi duygu analizi için önemli olabilir. Örneğin; radikal bireylerin paylaşımlarından Bağdat'ta veya Arizona'daki Twitter kullanıcılarından kaynaklı olup olmadığını anlamak önemli olabilir (Lynch, Freelon, Aday, 2014). Newman'a göre Twitter duygu analizi, duygu analizi olarak bir sosyal barometre olarak düşünülmektedir (Newman, 2018). Duygu

analizi sosyal barometre ile filtrelenmemiş sonuçları elde edilebilir (Newman, 2018). Bir işletmenin yalnızca çalışanlarını tanıması önemli değildir; önemli olan sosyal medya platformlarında tüm tüketicilerinin paylaşımlarını takip etmesi ve duygu analizi yapabilmesidir. Bu da şirket olarak tüketicilerini ve kendilerini daha iyi tanımasını sağlar. Sosyal medya platformlarında tüketiciler ürünler hakkında her zaman paylaşım yaparlar. Böyle olunca şirketler sürekli olarak sosyal medya platformlarını takip etmekte ve kendilerini tüketicilerinin isteklerine, şikayetlerine göre ayarlayabilmektedir (Newman, 2018).

2.2. Sosyal Medya

Sosyal medya platformları; insanların özgürce fikirlerini, yorumlarını, görüş ve duygularını paylaşmalarını sağlar. Twitter, Facebook, Instagram, LinkedIn genel olarak en çok kullanılan sosyal medya platformlarıdır (Giachanou, Crestani, 2016).

Sosyal medya platformları, kullanıcıların istedikleri kadar içerik paylaşabilecekleri geleneksel bloglarla karşılaştırıldığında, çok daha küçük içerik paylaşımına da izin verir.

2.2.1. Kullanılabilirlik

Twitter; Mac OS , Windows, Android gibi farklı işletim sistemi kullanan telefon, tablet ve masaüstü cihazları gibi birden fazla farklı cihaz tipi kullanılarak bağlanabilen uygulamaya sahiptir. Twitter'a birçok platformdan erişebilmenin yanı sıra doğrudan web sitesinden de erişilebilir. Twitter, birkaç ülke dışında, tüm dünyada özgürce kullanılan bir sosyal medya platformudur. Şu anda Çin, İran ve Kuzey Kore devletleri vatandaşlarına Twitter'a erişimi engellemiştir. Bazı ülkelerin hükümetleri, bilgi paylaşımını önlemek amacıyla Twitter'a erişimi veya Twitter'daki belirli tweetleri sansürlemiş veya engellemiştir (Twitter sansürü, 2019).

2.2.2. Sosyal Medyanın Faydaları

Twitter, duygu analizi için ham veri kümesidir. Twitter verilerinin niteliği analiz için çeşitli faydalı alanlar içerir. Örneğin; Twitter API kullanılarak elde edilen tweetlerin kullanıcıdan gelen tweet bilgisi olarak bir gönderinin ne zaman oluşturulduğuna, kaç kez yeniden retweetleştirildiğine ve coğrafi konuma (bir kullanıcı bu seçeneği devre dışı bırakmadığında) ilişkin verileri içerir. Coğrafi konumlu tweetler

bir fikrin, yorumun, olayın vb. nasıl yayıldığını tespit etmek için kullanılabilir. Ayrıca, duygu analizi, fikirlerin yayılmasını engellemek veya bir fikrin sosyal medyada nasıl yayıldığını anlamak için kullanılabilir. Elde edilen bilgiler, kötü amaçlı bireyleri tespit etmek veya istenmeyen bir fikrin yayılmasını önlemek için de kullanılabilir (Marcellino, Smith, Paul, Skrabala, 2017).

2.2.3. Sosyal Medyanın Sorunları

Sosyal medya platformları duygu analizi için bir veri kaynağıdır. Sosyal medya platformlarında herhangi bir konuda yapılan paylaşımlar ile genelleme yapılamayacağını ve paylaşımların sosyal medyayı kullananların yaptıkları paylaşımlara doğru duygusal ivme kazandığını bilmek gerekir. Ayrıca, sosyal medya dünyanın tüm bölgelerinde aynı miktarda kullanım ve popülerliğe sahip değildir (Marcellino, 2017).

İnternete erişimi olan birçok ülkeden kullanıcılar sosyal medya platformlarını kullanmaya başlamıştır. Ama sosyal medya duygu analizi, etkin bir şekilde kullanılmamakta ve izlenmemektedir (Marcellino, 2017).

Duygu Analizi, herhangi bir sosyal medya platformundaki paylaşımların olumlu, olumsuz ve tarafsız olarak sınıflandırılmasıdır. Sosyal medya platformlarında veriler hızlı üretilir, veri miktarı çok büyüktür ve çeşitlidir. Bu veriler yapılandırılmamıştır ve analiz edilene kadar anlamlı bir bilgi vermemektedir. Bu nedenle bu büyük miktarda veriden faydalı bilgi elde edebilmek için duygu analizi yapılmalıdır; yani bu veriden, bilgiye ulaşmayı mümkün kılan öznitelikler çıkarılıp bu veriler sınıflandırılmalıdır. İnsanlar çevresindeki düşünce ve düşüncelerden sürekli etkilenirler, bunun için duygu analizi çok önemlidir. Bugün herhangi biri bir ürün satın almak veya oy vermek veya bir film izlemek, vs. istiyorsa ilk önce ürün veya aday ya da film hakkındaki Twitter, Facebook, Instagram gibi sosyal medya platformlarında neler olduğunu, tepkileri ve düşünceleri bilmek istemektedir. Bu nedenle bu büyük miktardaki verilerden otomatik olarak anlık duygu analizi oluşturabilen bir sisteme ihtiyaç vardır.

2.2.3.1. Veri ve Veri İşleme Ortamının Önemi

Sosyal medya platformlarında duygu analizi için verilerin toplanmasında ve sonuçların elde edilmesinde kullanılan veri işleme ortamı önemlidir.

• Veri ve Veri İşleme Ortamı Edinme Maliyeti: Duygu analizi yapmak maliyet

gerektirir. Sosyal medya verilerinin büyük veri kapsamında olması, verilerin toplanmasını ve verileri işlemeyi zorlaştırmaktadır.

• Ölçeklendirme Analizi: Sosyal medya verilerinin artması nedeniyle duygu analizi yapabilecek sistemler geliştirmek zor olabilir. Anlık olarak veri akışlarıyla verileri çözümleyecek sistemlere ihtiyaç vardır.

• Standartlar ve Paylaşma: Sosyal medya platformlarındaki paylaşımlardan duygu analizi yapmak için ortak bir çalışma geliştirilmelidir. Ortak bir veri mimarisiyle standartlar geliştirmek için kurumsal düzeyde bir çözüm getirilmelidir. Bu ortak mimari; kurumların ham verileri, analiz sonuçlarını ve kurum genelinde veri görselleştirmelerini paylaşmalarını sağlayacaktır. Standartlar olmadan yeni veri işleme ortamı veya yöntemleri test etmenin sistematik bir yolu olmayacaktır. Mevcut olan sosyal medya verilerinin miktarını ve gelecekte kullanılabilecek miktarları tam olarak kullanabilmek için ve yönetmek için bir altyapı geliştirmesi gerekmektedir (Marcellino, 2017).

2.2.3.2. Veri Kaynağı

Duygu analizi için sosyal medya platformlarından paylaşılan verileri kullanarak benzersiz ve çok miktarda veriye erişilebilmektedir. Sosyal medya platformları erişilebilir bir bilgi kaynağı olsa da kullanımıyla ilgili olarak büyük dezavantajları ve zorlukları vardır. Dünyada, ülkelerde, şehirlerde sosyal medya platformlarını kullanan kullanıcı sayıları değişkendir. Bu değişkenlik, sosyal medya platformlarının veri kanyağı olarak kullanımında zorluklara neden olmaktadır. Sosyal medya kullanıcılarının paylaştığı veriler doğal olarak içinde bulunduğu toplumun görüşlerini yansıtmaktadır (Marcellino, 2017).

Ayrıca, Twitter API'sine erişim kısıtlıdır ve toplanacak tweet sayısı sınırlıdır. Twitter API'si 15 dakikalık sorgulara ayrılmıştır ve limiti sorgu başına 15 istek için izin vermektedir (Rate Limiting-Twitter Developers, 2019). Ayrıca Twitter'ın kullanımında ülkeden ülkeye, şehirden şehire çok büyük farklılıklar vardır. Bu nedenle Twitter'ın kullanımının değişkenlik göstermesi, veri kaynağı olarak paylaşılan tweetlerin az olmasına neden olabilir.

2.3. Duygu Analizi

Duygu analizi, metinde belirli bir konuya yönelik olarak ifade edilen görüş,

duygu ve tutumların çıkarımıdır. Duygu analizi; farklı konularla ilgili görüş, duygular ve tutumları tespit etmek, çıkarmak ve sınıflandırmaktan ibarettir (Ravi, Ravi, 2015).

Genellikle iki, belki de üç değerin olumlu, olumsuz ve tarafsız olduğu bir metnin duygusunu sınıflandırmak için yeterli olduğu varsayılır (Ravi, Ravi, 2015). Duyguları sınıflandırma yöntemleri, daha güçlü bilgisayar kapasitesiyle gittikçe daha gelişmiş hâle gelmiştir; ancak mevcut sınıflandırma yöntemleri, önceden belirlenmiş bir kelime listesi ile sabit duygu değerleriyle eşleştirme yapmaktadır (Cambria, Das, Bandyopadhyay, Feraco, 2017).

2.4. Doğal Dil İşleme

Doğal dil işleme; sosyal medya platformları, web sayfaları, e-postalar, farklı dillerde yayımlanan gazete makaleleri ve başka birçok kaynaktan elde edilen doğal dilde yazılmış metinler ile ilgilenen bir alandır.

Otomatik soru cevaplamadan e-posta spam tespitine, makine çevirisinden yazım ve dilbilgisi düzeltmeye kadar çok çeşitli doğal dil işleme uygulamaları vardır. Belirli bir metinden görüş almayı amaçlayan duygu analizi de doğal dil işlemenin bir çalışma alanıdır.

Makine öğrenmesi yöntemleriyle doğal dil işleme kullanılması metinlerin analizini yapmayı kolaylaştırır. Makine öğrenme tekniklerini doğal dil işleme yöntemleri ile birleştirerek verilen bir metnin kelime seviyesinden cümle seviyesine ve tüm metnin anlamsal seviyesine kadar olan duygusunu çıkarmak ve değerlendirmek mümkündür.

2.5. Uygulama

Duygu analizi, kullanıcıların duygu durumunu yansıtan ifadeleri tanımlamaya yarayan bir doğal dil işleme uygulamasıdır (Cambria, Das, Bandyopadhyay, Feraco, 2017). Siyasi hareket, pazar araştırması, müşteri memnuniyetinin ölçülmesi, film gişe tahminleri, kamuoyunun eğilimini bilmek ve daha birçok konuda çeşitli hedeflere ulaşmada duygu analizi yardımcı olur (Ravi, Ravi, 2015). Duygu analizi, dil bilimi kadar bilgisayar bilimlerini de ilgilendirmektedir. Duygunun ancak bir duyguya bağlı olarak düşüncenin ima ettiği olumlu, olumsuz veya tarafsız olduğu duygu analizi ile ortaya çıkarılır (Liu, 2015).

2.6. Twitter

Twitter, 2006 yılında kurulan ve kullanıcıların tweet olarak isimlendirilmiş mesaj gönderdiği bir sosyal medya platformudur (Twitter, 2019). Başlangıçta 140 karakterle sınırlı iken şimdi en fazla 280 karakter içeren bir tweet mesajı paylaşılabilmektedir. Kullanıcı adı en fazla 20 karakter olabilmektedir. Bu durum, Twitter'ın popülaritesini azaltmamıştır; çünkü Twitter'ı günde yaklaşık 500 milyon tweet gönderen yaklaşık 328 milyon aktif kullanıcı kullanmaktadır. Günlük oluşturulan bu veri miktarı nedeniyle, Twitter'ın kullanıcılar tarafından oluşturulan en büyük veri kümelerinden birini oluşturduğu düşünülmektedir (Giachanou, Crestani, 2016).

Twitter kullanıcıları, istedikleri kişinin paylaşımlarını daha yakından takip edebilmektedir. Twitter'a mobil cihazlardan, tabletlerden ve web sitesi üzerinden erişilebilir. Twitter'da, kullanıcılar tweetlere cevap verebilir ve tweetleri paylaşabilir. @Kullanıcıadı olarak paylaşılan tweetler, tweetlerin bir cevap olduğunu belirtir. Retweet özelliği ile RT @kullanıcıadı biçimi kullanılarak paylaşımlar yapılabilir. #İstanbul, #Fenerbahçe, #Galatasaray vb. gibi konu etiketleri olan hashtag kelimeleri bir hash sembolüne eklemek mesajın içeriğini vurgular.

Twitter'ın karakter kısıtlaması ve herkese açık paylaşım ortamı sağlaması benzersiz bir lehçe yaratmıştır (Hu, Talamadupula ve Kambhampati, 2013). Twitter, birçok kısaltmayı, yanlış yazılmış kelimeyi ve ifadeyi içerir. Örneğin: Kahv6 (kahvaltı), 1 Şey (bir şey), Gg (güle güle), R yapmak (bir şey hakkında geri adım atmak) veya :), :S, :(vb. Belirli bir zaman dilimi içinde sıkça kullanılan kelimeler, konu etiketleri gündem olur. Bu gündemleri, kullanıcılar her ülkede en çok konuşulan gündemi listeleyip görebilir. Twitter'ın herkese açık olması ve herkesin Twitter'da gündem oluşturabilmesi kullanıcı sayısını her geçen gün artırmaktadır.

2.6.1. Twitter Duygu Analizi Uygulamaları

Twitter'daki paylaşımlar, duygu analizi için çok çeşitli büyük bir kaynak olarak kabul edilir. Aşağıda duygu analizinin başlıca uygulamaları bulunmaktadır.

- Ürün ve Hizmet incelemeleri : Duygu analizi, tüketici ürünlerinin ve hizmetlerinin değerlendirilmesinde kullanılır. Twitter kullanıcıları ürün veya hizmetler hakkında görüşlerini paylaşırlar (Pak ve Paroubek, 2010). Ürünler

hakkında yapılan paylaşımlar ile ilgili otomatik olarak yapılan özetleri sunan birçok web sitesi bulunmaktadır. Bunun kayda değer bir örneği "Google Ürün Arama"dır. Bir çalışmasında Liu ve arkadaşları, ürün satışlarını tahmin etmek için tüketicilerin bloglardan aldığı görüşlerin değerlendirdiğini göstermektedir (Liu, Huang, An ve Yu, 2007) . Ürün incelemelerini içeren bloglardan başka filmlerin incelemelerini içeren bloglar da vardır. Filmler için temel olarak duygu durumu ile ilgili bir filmi iyi ya da kötü olarak sınıflandırmaktan daha çok farklı duygu sınıflandırmalarına yer verilmektedir.

- Marka İtibar İzleme: Twitter ve Facebook, duygu analizi için çok fazla ve önemli veri sağlar. En çok kullanım alanı belirli bir markanın Twitter veya Facebook'taki itibarını izlemektir.

- Sonuç tahmini: Sosyal medya platformlarından yapılan paylaşımların duygularını analiz ederek belirli bir olayın olası sonucunu tahmin edilebilmektedir. Örneğin; duygu analizi, siyasi adaylara önemli bir değerlendirme sağlayabilir. Adayların, seçmenlerin farklı konular hakkında genel olarak ne düşündüklerini bilmesini ve konuşma ve eylemleriyle seçmenleri nasıl etkileyebileceğini sağlayabilir. ABD'deki 2012 başkanlık seçimlerine yönelik kamuoyu tepkilerinin anlık duygu analizi için Twitter'dan yapılan paylaşımlar toplanmıştır. Twitter'da ABD'deki seçim adayları ile ilgili yapılan paylaşımlar, paylaşılan tweetlerden her aday hakkındaki yorumlar ve ayrıca kimin kazanacağının bir tahminini oluşturan büyük miktarda veri toplanmış ve duygu analizi yapılmıştır. (H. Wang, D. Can, F. Bar, S. Narayana, 2012)

- Karar verme: Duygu analizi, karar verme sistemlerine yardımcı olması için kullanılmaktadır. Örneğin; finansal piyasalarda yatırım yapmak için kullanılır. Her şirketle ilgili çok sayıda haberler, makaleler, bloglar ve tweetler vardır. Çeşitli yatırımcıların seçimlerinde karar vermelerine yardımcı olması için duygu analizi kullanılmaktadır (Dergiades, 2012).

Duygu analizi için Twitter'ı kullanmak zordur. Duygu analizi yapabilmek için Twitter aşağıdaki durumlar göz önünde bulundurularak seçilmiştir:

- Twitter herkese açık sosyal medya platformudur.

- Twitter'da duygu paylaşımı 280 karakterle sınırlıdır. Bu karakter kısıtı kullanıcıları az kelime ile duygularını ifade etmeye zorlamaktadır.
- Twitter, gerçek zamanlı olarak paylaşımları kısıtlı da olsa toplamayı kolaylaştıran Streaming API sağlar.

Duygu analizinde kullanılan dil duyguların belirlenmesinde hatalara neden olmaktadır. Dilin geniş ve karmaşık yapıya sahip olmasından dolayı yüksek duygu tespiti yapabilen kesin bir model elde etmek mümkün olmaz. Ayrıca; metinlerin anlamsal bağını, duygu tonunu, şüpheciliğini, alaycı yaklaşımını vb. duyguları ayırt etmek çok zordur (Newman, 2018). Duygu analizi için şirketler ve medya kuruluşları, sosyal medya platformlarını ve Twitter verilerini kullanarak duygu analizini geliştirmek için çalışmaktadır. Spredfast, Microsoft Azure ve Lexalytics gibi şirketler, duygu analiz hizmetleri sunmaktadırlar (Kouloumpis, Wilson, Moore, 2011). Sosyal medya büyümeye devam etmektedir ve bununla birlikte duygu analizi başarımları da artmaktadır. Microsoft gibi büyük şirketler duygu analizini geliştirmekte ve kullanıma sunmaktadır. Duygu analizinin değerli olduğu ve gelecekte önemli bir büyüme göstereceği tahmin edilmektedir.

2.6.1.1. Zorluklar

Duygu analizi; belirli bir konuyla ilgili bir kişinin tutumunu, görüşünü, duygularını veya yargısını belirleme çalışması olduğundan bazı zorluklar içermektedir. Verilerin tweet şeklinde olmasından dolayı ayrıca bir zorluğu da beraberinde getirmektedir. Twitter verisinin zorlukları; yapılandırılmamış veriler olan doğal dilde yazılmış olması, kısıtlı kelime ile yazılması, semboller içermesi ve kelimelerin farklı kullanımlarını ayırt etmenin zor olmasıdır. Örneğin:

"Hayır, kesinlikle! Yeni iPhone X'de gerçekten kullanışlı bir menü var."

"Hayır, kesinlikle yeni iPhone satın almıyorum. X!").

Bu iki tweetteki duyguyu sözlük tabanlı gibi makine öğrenmesi dışındaki yöntemler ile duygu analizi yaparak ifade edilen 2 ayrı duyguyu saptamak zordur.

Duygu analizinde, konu temelli anahtar kelimeler söz konusu olduğunda, çeşitli özelliklerin dikkate alınması gerekir. Makine öğrenmesi, duygu analizin konu temelli metin sınıflandırmadan daha zor olmasının temel nedeni, konu temelli sınıflandırma

anahtar kelimelerle yapılabilir; ancak bu makine öğrenmesi duygu analizinde iyi başarım göstermez (Turney, 2002).

Duygu analizindeki sınıf sayısının konu temelli sınıflandırmadaki sınıf sayısından daha az olmasına rağmen duygu analizinin geleneksel konu temelli metin sınıflandırmadan daha zor olduğunu ortaya koymaktadır (Pang, Lee, 2008). Makine öğrenmesi ile duygu analizi, konu temelli sınıflandırma ile karşılaştırıldığında daha zordur. Çünkü konu temelli duygu analizi anahtar kelimelere dayanmaktadır.

Bir tweette ifade edilebilecek olumlu, olumsuz ve tarafsız vb. farklı duygu sınıfları vardır. Duygu öznel bir yargı olduğundan, bir tweetin duygu sınıfının bir veri analisti tarafından belirlenmesi gerekir ve bu nedenle tweetlerin duygu sınıflarının el ile eklenmesi zaman alıcı ve yoğun emek isteyen bir iştir. Etiketli tweetlerin elde edilmesinin zor olması ve etiketli tweetlerin az olmasından dolayı makine öğrenmesi duygu analizi modelleri elde etmek için bu tweetlerin duygularını sınıflandırıp kullanmak zordur.

Bununla birlikte, el ile etiketlenmiş tweetlerden oluşan bir veri kümesinden eğitilmiş bir sınıflandırıcı modeli, eğitim verilerine dahil edilmemiş konular hakkında veya farklı bir zaman diliminde paylaşılmış tweetler ile iyi bir başarım göstermeyecektir. Bunun nedeni, tweetlerde karşılık gelen duygu etiketleri arasındaki ilişkinin bir konu alanından diğerine veya zaman içinde değişebilmesidir. Sosyal medya platformlarında paylaşımlar, politikacılar, ürünler, film incelemeleri, spor müsabakaları, futbol takımları, diğerleri gibi farklı alanlarla ilgili olabilir. Bir duygu, belirli bir konu hakkında olur ve o konuya bağlı olur. Bu, belirli bir konu alanı içinde olumlu veya olumsuz olarak kabul edilen duygunun farklı bir konuda aynı duyguya sahip olmayacağı anlamına gelir. Bunun sonucu, belirli bir konu alanı verileri üzerinde eğitilmiş bir duygu sınıflandırıcısının diğer konular veya alanlar için aynı sınıflandırma başarımına sahip olmayabilir. Bir kelimenin duygusu, bir insan tarafından belirlenen öznel bir yargıdır.

Duygu analizinde, etiketleme yapan veri analisti kendi duygularını etiketlemeye katabilir. Örneğin; şöyle bir tweet olabilir: “Fenerbahçe dün oyunu nasılll kaybetti”. Tweet net bir duygu durumu belirtir, ancak farklı veri analistleri etiketleme yaparken bu konuda tamamen farklı hissedebilir.

Duygu, bir olumsuz kelime kullanmadan da ifade edilebilir. Belirli bir metnin

nesnel mi öznel mi olduğunu belirlemek çok zordur. Nesnel ve öznel metinler arasında her zaman ince bir çizgi vardır. Paylaşım yapan bireyin duygu durumunu belirlemek çok zordur. Örneğin; bireyin görüşü ile veri analistinin görüşü farklı olabilir. Konu alanına ve kullanılan kelimelere göre başka faktörler de duygu durumunu etkileyebilir (Pang ve Lee, 2008). Duygu analizinin diğer zorlukları; alaycılık, ironi ile duyguların ifade edilebiliyor olmasıdır.

Twitter duygu analizi, Twitter'da paylaşılan tweetlerin ifade ettikleri duyguları analiz etmektir. Twitter, duygu analizi için zorluklar içeren alandır. Twitter duygu analizinin zorluklarından bazıları (Giachanou, Crestani, 2016):

- Metin Uzunluğu: Her tweet en fazla 280 karakterle sınırlıdır. Az kelime kullanılarak yapılan paylaşımlardan duygu sınıflandırması yapmak zordur.
- Konu İlgi Düzeyi: Bir konuyu belirleme yöntemlerinden biri #hashtag kullanmaktır, ancak bunlar bir kullanıcı tarafından belirlenir ve o konuyla ilgili doğru olur ya da olamaz.
- Yanlış Dil: Twitter'da birçok kullanıcı tweetlerinde duygu analizi sınıflandırmasını zorlaştıran çok kötü ve yanlış yazım içeren bir dil kullanır. Bunun nedeni, bir kelimenin doğru bir şekilde yazılmaması durumunda, tweet doğru bir şekilde hesaba katılmayacağı için veri analisti tweetin doğru hissiyatını bilemez. Ayrıca, tweetler; doğru duygu sınıflandırılması zor olan argo kelimeler, kısaltmalar, kelimelerin uzatılması gibi kelimeler de içerir. Birçok tweet; yazım yanlışlarına, yanlış kullanılan kelimelere ve cümlelere sahip olduğu için duygu sınıflandırılmasının yanlış olmasına sebep olmaktadır.
- Olumsuzluk: Tweetlerde alaycılık ve çifte olumsuzlukları tespit etmek çok zordur. Eğer duygu sınıflandırılması doğru bir şekilde tespit edilmezse tweetin gerçek duygu durumunun tam tersi olur (olumlu, olumsuz olur veya olumsuz, olumlu olur).
- Gereksiz Kelimeler: Gereksiz kelimelerin genellikle kaldırılması gerekir, çünkü bunlar bir veri analisti ve sınıflandırıcı algoritma için faydalı bilgiler sağlamaz. Genelde; yani, gibi, bazen vb. gibi duygu belirtmeyen gereksiz kelimelerin kaldırılmaları gerekir.

- Çok Dilli İçerik: Twitter'da tweetler içinde yabancı dillerde kelimeler de olabilir. Bu nedenle bir analiz dili seçilmesi ve aynı dilin tweetlerinin toplanması gerekir. Kullanıcının birden çok dili tek bir tweet hâlinde birleştirdiği bir paylaşım da olabilir.
- Çok Farklı İçerik Türleri: Tweetler birçok kez farklı içerik türleri ile yüklenir. Tweetler; metin, resim, film klipleri, gif formatı, web bağlantıları, #hashtag, vb. içerebilir. Birden fazla farklı içerik türüne sahip olan tweetler temizlenir.

Ayrıca, üzerinde durulması gereken önemli bir nokta, tamamen farklı özellikli konular için özel olarak o konuyla ilgili modeller oluşturulmalı ve kullanılmalıdır. Bu nedenle öncelikle belirli bir konuda ait olduğu alana göre duygu sınıflandırması yapılmalıdır. Örneğin; yapılan analiz teknoloji isimli bir konu içeriyorsa ve iphone marka cep telefonu ürünü ile ilgili araştırma yapılacaksa onu içeren tweet konu odaklı oluşturulmuş "Teknoloji" isimli model kullanılarak sınıflandırılmalıdır. Diğer konular; "Ekonomi", "Eğitim" "Sağlık", "Spor" ve "Müzik" olabilir. Bu amaçla, istenen konulara göre arama yapılabilir. Farklı konuları içeren eğitim veri setleri oluşturmak için çok sayıda tweet toplandıktan sonra sınıflandırıcı eğitimi başlar. Bu veri setlerinin her biri için ayrı bir model daha sonra eğitilerek kullanılabilir. Bu tezde, "Spor" konusu belirlenmiştir ve anlık olarak futbol maçlarında taraftarların attığı tweetler toplanmıştır.

2.6.1.2. Sınıflandırma

Duygu analizi bir anlamda denetimli makine öğrenmesi alanı ile ilgili metin sınıflandırma çalışmasıdır. Bununla birlikte, sınıflandırma, ele aldığımız konuya göre farklılık gösterebilir. Belirlediğimiz konuya göre ve duygu analizi yapan kişinin bakış açısına bağlı olarak sınıflandırma yapılmaktadır.

2.6.1.3. Özellik Seçimi

Twitter duygu analizinde, bir öznitelik kümesine dayanarak duygu tahmini yapılmaktadır. Seçilen öznitelikler ve bunların birleşimi, metnin duygusunu belirlemede önemli bir rol oynamaktadır. Twitter kullanılarak 3 farklı duygu analizi yapılmaktadır (Giachanou, Crestani, 2016):

- Anlamsal özellikler: Analiz için en önemli kelimeler, olumlu veya olumsuz bir duygu içeren duyarlılık kelimeleridir.

- Sözdizimsel özellikler: Sözdizimsel özellikler tipik olarak unigramları, bigramları, n-gramları, terim sıklıklarını ve konuşmanın bir kısmını içerir.
- Stilistik özellikler: Bu elemanlar genellikle standart olmayan yazma ifadeleri olarak görülebilir. Örnek olarak da emoji olarak bilinen ifadelerdir.

BÖLÜM 3

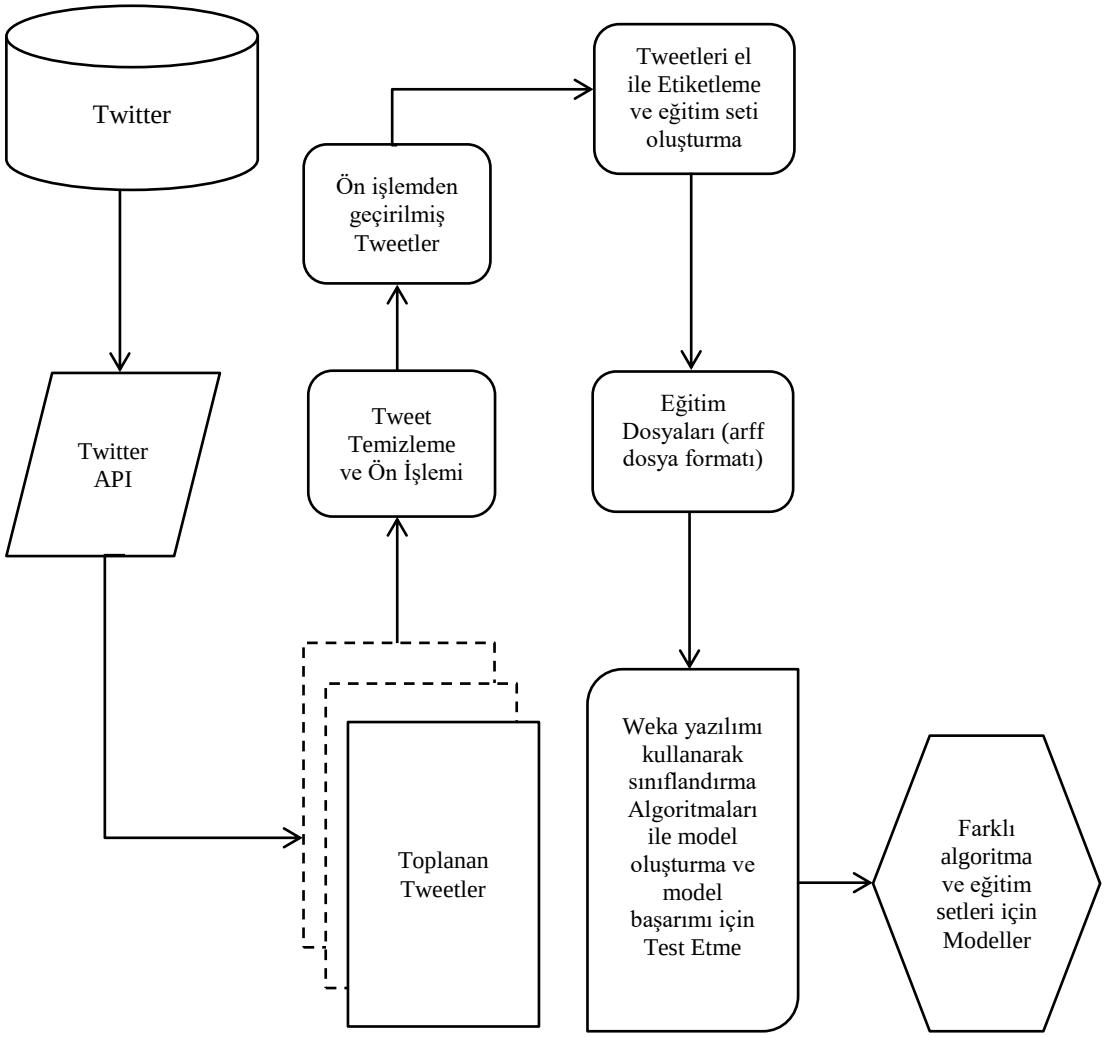
METODOLOJİ

3.1. Önerilen Mimari

Bu çalışmada Twitter platformunda paylaşılan tweetler toplanarak duygu analizi yapılmıştır. Makine Öğrenmesi algoritmaları için yapılan işlemlerde Weka yazılımından faydalanılmıştır. Şekil 3.1'de verilen şekil ile Twitter platformundan tweet toplama ve model oluşturma arasındaki süreç ifade edilmiştir. Bu aşamada farklı özniteliklere sahip eğitim setleri kullanılarak modellerin oluşturulması için:

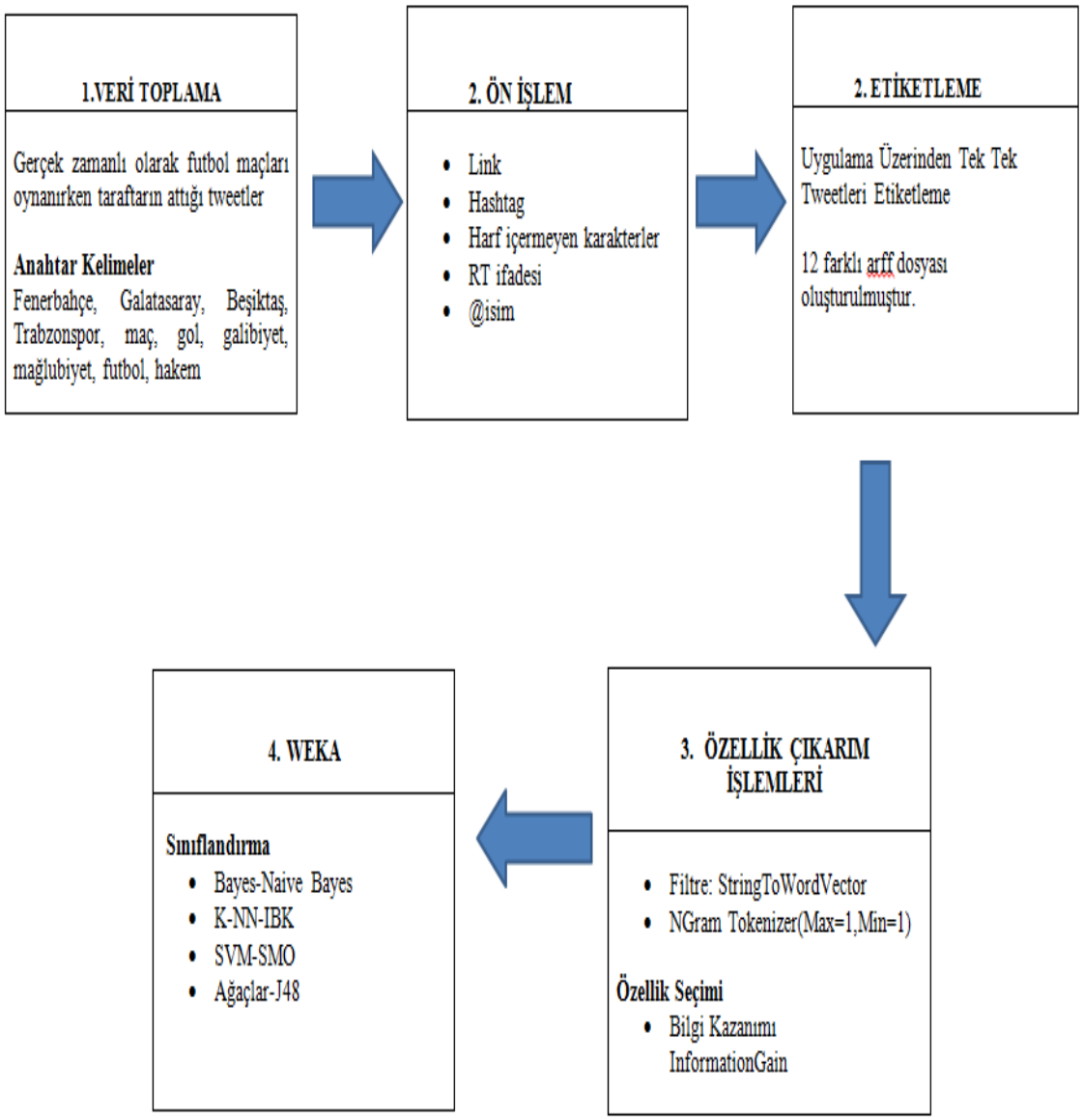
- Tweetlerin toplanması ve ön işlemden geçirilmesi,
- Tweet sınıflarının el ile etiketlenmesi,
- Belirlenmiş algoritmalar kullanılarak sınıflayıcı modellerin oluşturulup modellerin çapraz doğrulama ile testlerinin yapılması işlemleri yerine getirilmektedir.

Şekil 3.1'de farklı eğitim setlerinden farklı Makine Öğrenmesi sınıflandırma algoritmaları ile oluşturulmuş modelleri kullanarak tweetleri anlık sınıflandıran ve sonuçları anlık görselleştiren bir uygulamanın çalışma mimarisi gösterilmektedir.
Şekil 3.2'de ise bu mimarinin ayrıntılarını gösteren blok şema verilmiştir.

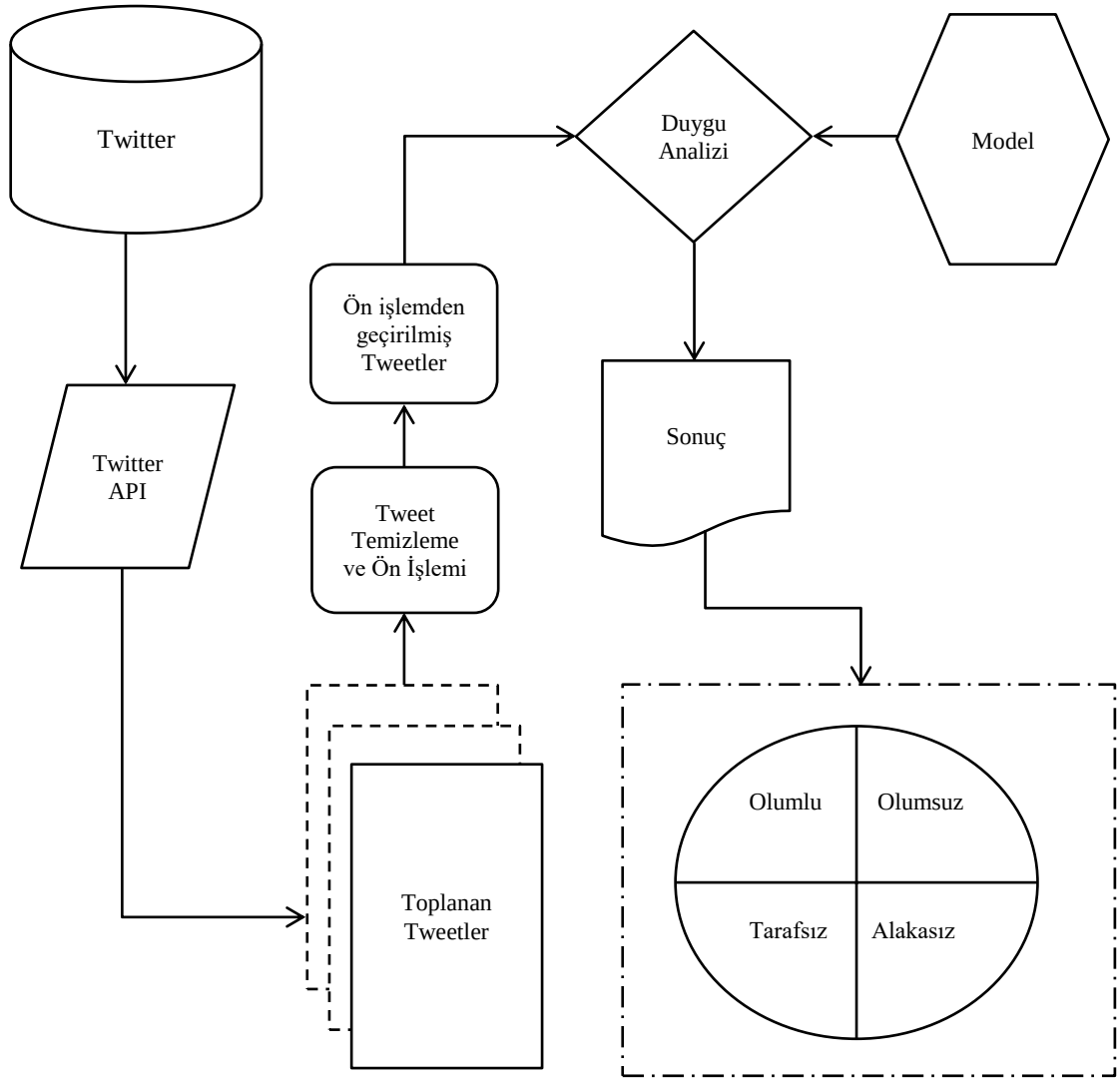


Şekil 3.1. Twitter'dan tweet toplama ve model oluşturma sistem mimarisi

Tweetler twitterdan toplanmıştır. Bunun için C# .NET programlama dili kullanılarak bir uygulama geliştirilmiştir. Twitter Streaming API, bu uygulama ile Twitter API arasındaki arayüzü oluşturur ve bu bağlantıdan sorumludur (kimlik doğrulama, kalıcı bağlantılar vb.). İlk olarak C#. Net'deki Tweetinvi kütüphanesinin yardımıyla Twitter Streaming API kullanarak tweetler toplanmıştır. Sonra veri madenciliği yöntemleri kullanılmış ve metinden özellik çıkarma işlemleri için bu tweetler işlenmiştir. Ön işlemden sonra bu veriler uygulamamız üzerinden tek tek okunarak "Olumlu", "Olumsuz", "Tarafsız", "Alakasız" olarak etiketlenmiştir. Etiketlenen tweetlerden öznitelik seçimi ve özellik çıkarımı yapılmıştır ve 12 farklı arff dosyası oluşturulmuştur. 12 farklı arff dosyası Weka programı kullanılarak makine öğrenmesi algoritmaları ile test edilmiştir ve modeller oluşturulmuştur.



Şekil 3.2. Tweet toplama ve model oluşturma ayrıntıları



Şekil 3.3. Anlık duygu analizi sistem mimarisi

Uygulamamızda anlık olarak tweet toplarken aynı zamanda oluşturduğumuz modelleri kullanarak tweetlerin anlık duygu analizleri elde edilmiştir. Şekil 3.3'te anlık duygu analizi sistem mimarisinde görüldüğü gibi elde edilen sonuçlar pasta grafiği şeklinde olumlu, olumsuz, tarafsız ve alakasız olarak görselleştirilmiştir.

3.2. Twitter API

Twitter, veri analistlerinin Twitter API'si yardımıyla tweet toplamalarına izin verir. Twitter'ın, Twitter Rest API ve Streaming API olmak üzere iki API türü kullanımı mümkündür. Farklı özelliklere sahip olan bu API'lerin özellikleri şunlardır:

- Rest API: Get() ve Post() isimli iki ana işlevi vardır. Twitter verilerini okuma ve yazma işlevlerini yerine getirirler.
- Streaming API: Twitter'ın herkese açık tweet veri akışına erişmek için kısıtlı erişim sağlar.

Bu ikisi arasındaki fark Rest API kısa zaman aralıkları için bağlantıları desteklerken Streaming API ise gerçek zamanlı olarak uzun süre bağlantı sağlar. Rest API kısa süreli bağlantıları desteklemektedir ve istek sayısı sınırlıdır, aynı zamanda belirli miktarda tweet indirilebilir. Rest API, zamandan bağımsız olarak durum güncellemeleri ve kullanıcı bilgileri gibi Twitter verilerinin erişimine izin verir. Ancak bir haftadan daha eski olan Twitter verilerine erişilemez. REST erişimi paylaşıldıktan sonra bir haftayı geçmeyen Twitter verileriyle sınırlıdır. Bu nedenle, Rest API bu bir haftalık toplanmış verilere erişime izin verirken Streaming API sadece gerçek zamanlı olarak verilere erişim sağlar. Bu çalışmada Streaming API kullanılmıştır. Bunun nedeni ise model olarak geliştirilen uygulama ile anlık duygu analizi yapılmaktadır ve büyük miktarda tweet toplamak için veri sınırı olmayan uzun ömürlü bir bağlantıya ihtiyaç duyulmaktadır.

Streaming API kendisine gönderilen istek sayısını sınırlar. Streaming API, uzun süreli bağlantılara uygun limit ve erişim seviyelerine sahiptir. Streaming API'de 15 dakikada 15 sorgu ve 15 dakikada 180 sorgu olmak üzere iki sorgu yapısı mevcuttur (Twitter, 2018). Streaming API kullanarak model olarak geliştirilen uygulamada 15 dakikada 180 sorgu tipi kullanılmış ve geriye kalan sorgu adedinin anlık izlenmesi sağlanmıştır. Bu sayede kalan sorgu adedi izlenebilmektedir.

3.2.1. Stream API Kullanarak Twitter'dan Veri Akışı

Tweetler paylaşıldığı anda Streaming API, neredeyse gerçek zamanlı olarak tweelerin toplanmasını sağlar. Streaming API; tweetlerin duygu analizi için belirli bir

konuya yönelik kelimelerle, kullanıcı adlarıyla, coğrafi konumlarıyla veya zaman dilimleriyle vb. gibi belirlediğimiz özelliklere göre tweetlerin anlık filtrelenmiş olarak toplanmasını sağlar. Böylece, anlık olarak bazı konulara, dile, konuma bağlı olarak tweetlere erişilebilmektedir.

Anlık duygu analizi için Streaming API kullanılarak tweet akışı üzerinden veriler toplamak ve yorumlamak için bir model uygulama geliştirilmiştir. Bu uygulamada, Twitter Streaming API ile verileri toplamak için Tweetinvi isimli bir MS .NET C # kütüphanesi kullanılmıştır.

Twitter Streaming API, Twitter hesabından verilere erişilip belirlenen amaç için kullanabilmeye izin verir. Bunun için Twitter kimlik bilgileri ile dev.twitter.com web sitesinde oturum açılarak bağlantı için gerekli verilerin alınması gerekir. Bu web sitesinde, önce gerekli bilgiler girilerek paylaşılan tweetlere anlık olarak erişim için kullanılacak bir uygulama oluşturulmaktadır. Twitter Streaming API ile Twitter platformuna erişip hesaba ait 4 anahtar bilgi alınması gerekmektedir. Streaming API oluşturulduktan sonra müşteri anahtarı, müşteri gizli anahtarı, erişim belirteci anahtarı ve erişim gizli anahtarı geliştirilen modele uygulama içinde tanımlanıp kullanılmıştır. Bu anahtarlar, twitter verilerine erişilmek istendiği zaman uygulama kimliğini doğrulamak için kullanılır.

3.3. Tweet Toplama

Bu çalışmanın amacı, futbol müsabakaları oynanırken Twitter kullanıcılarının attıkları tweetlerden duygularını analiz etmektir ve bu amaçla bununla ilgili tweetler toplanmıştır. Veri toplama işlemi için Twitter Streaming API kullanılarak Twitter'dan tweetler toplanmıştır.

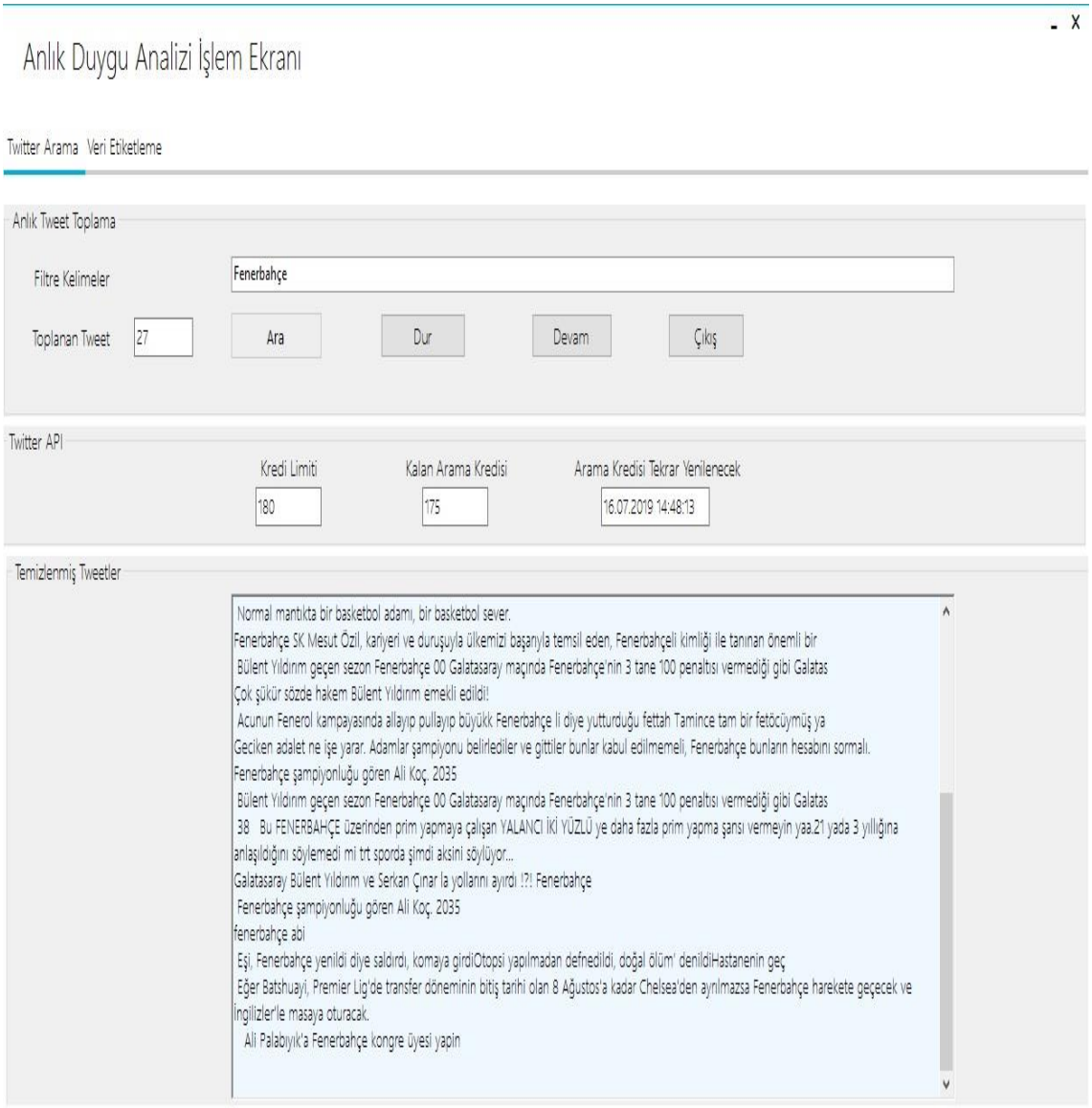
Eğitim veri setlerini oluşturmak için Twitter Streaming API kullanılarak, hashtag anahtar kelime araması yapılarak, farklı öznitelikli eğitim veri setleri oluşturulmuştur. Tweetler anlık olarak futbol müsabakaları oynanırken toplanmış verilerdir. Farklı zamanlarda farklı futbol müsabakaları oynanırken atılan tweetleri depolamak için yerel sunucu bilgisayar kullanılmıştır. Eğitim setleri oluşturmak için futbol müsabakalarında atılan tweetlerden oluşan bir HamTweet isimli dosya oluşturulmuştur. Bu amaçla 30000 tweet toplanmıştır. Tweetler depolandıktan sonra, bu tweetler ön işlemden geçirilmiştir ve olumlu, olumsuz, tarafsız, alakasız olarak el ile sınıfları etiketlenmiştir. Toplanan bu

30000 tweet kullanılarak 12 farklı eğitim seti oluşturulmuştur. Tüm eğitim setlerinde bu 30000 tweet baz alınarak sadece öznitelikleri farklı olan eğitim setleri yaratılmıştır. Toplanan toplam 30000 tweet ile dengeli bir veri seti oluşturulmuştur. Her bir sınıftan etiketli 7500 tweet eğitim setlerine dahil edilmiştir.

Her bir duygu sınıfı için eşit sayıda tweetler ile dengeli bir veri seti oluşturulmuştur. Tüm veri setleri içinde her sınıftan 7500 tweet ve toplam 30000 tweet vardır. Sınıfların herhangi bir sınıf lehine önyargının engellenmesi için her bir sınıftan eşit sayıda (olumlu, olumsuz, tarafsız, alakasız) dengeli bir veri seti oluşturulmuştur.

Twitter API için bir MS C # .NET kütüphanesi olan Tweetinvi, MS Windows, MS Windows Phone, Xamarin Android ve Xamarin iOS üzerinde geliştirme için kullanılabilen bir kütüphanedir. Tweetinvi ayrıca başka platformlar için uygulama geliştirme amacıyla kullanılabilen taşınabilir bir kütüphanedir.

Tweet toplama amacıyla kullanılacak uygulamanın geliştirilmesinde kullanılan bir tweet topla sınıfı oluşturulmuştur. Bu amaçla önce yetkilendirme protokolü olan "Oauth" protokolü kurulmuştur. OAuth, yetkilendirme için kullanılan standart bir protokoldür. Kullanıcının, herhangi bir sosyal medya platformu hesabını kullanarak şifreleri göstermeden üçüncü taraf web sitelerine giriş yapabilmesini sağlar. OAuth, kullanıcıya güvenlik ve yetkilendirme sağlar. Twitter Streaming API kullanarak anlık olarak paylaşılan tweetlere erişmek için kullandığımız model uygulama arayüzü Şekil 3.4'te gösterilmektedir.



Şekil 3.4. Tweet toplamak için geliştirilen uygulama arayüzü

Önceden yaratılmış olan Twitter kullanıcı hesap bilgileri, uygulama içinden Twitter API erişimi için kullanılmıştır. Belirli bir konuyla ilgili tweet toplamak için tweetinvi kütüphanesi kullanılarak kullanıcı hesap bilgileri uygulamamıza eklenmiştir. Ayrıca, anahtar kelimeler belirlenerek anlık olarak paylaşılan tweetlere bu anahtar kelimeler filtre olarak uygulanmıştır.

Bu işlem yapılırken futbol müsabakası ile ilgili olmayan reklam içeren veya farklı konuları içinde barındıran birçok tweet de toplanmıştır. Anlık olarak bir duygu analizi için bu alakasız tweetlerin her zaman olabileceği düşüncesiyle bu tweetler de değerlendirilmeye alınmış ve alakasız olarak etiketlenmiştir.

Toplanan bir tweetin, Makine Öğrenmesi algoritmaları için öznitelik olarak kullanılabileceği öngürülen 16 alanı seçilmiştir. Bu 16 veri alanı adları ve açıklamaları Çizelge 3.1'de görülebilir.

Çizelge 3.1. Tweet içinden seçilen alanların açıklamaları

Değişkenler	**Tanımlama**
Metin(stream.AddTrack(KonuTxt.Text.ToString())	Tweet metni
İsletimson (args.Tweet.Source)	İşletim sistemi
TweetUz(args.Tweet.PublishedTweetLength)	Tweet uzunluğu
Tweetkesildi(args.Tweet.Truncated)	Tweetin devam edip etmeme durumu, kesilmiş metin, tweet uzunluğuna sığmayan retweetin durumunu belirtir.
Tweethastagsay(args.Tweet.Hashtags.Count.ToString())	Tweetdeki hashtag sayısı.
Tweethassasicerik(args.Tweet.PossiblySensitive)	Tweetin hassas içerik içermeme durumu, tweetdeki içeriğin hassas içerik içerip içermediğini belirler.
Tweetgun(args.Tweet.CreatedAt.DayOfWeek.ToString())	Tweetin atıldığı gün, tweet haftanın hangi günü atıldı.
Tweetzamanaralik(01-12)	Tweetin atıldığı zaman aralığı
Tweetsadecesaat(args.Tweet.CreatedAt.Hour)	Tweet atılma saati
Tweetsadecedakika (args.Tweet.CreatedAt.Minute)	Tweet atılma dakikası
Hesapyili(args.Tweet.CreatedBy.CreatedAt.Year)	Tweet atanın hesap açılış yılı
Hesapayi(args.Tweet.CreatedBy.CreatedAt.Month)	Tweet atanın hesap açılış ayı
Hesaptweetsay(args.Tweet.CreatedBy.StatusesCount)	Tweet atanın toplam attığı tweet sayısı
Hesaptakipcisay(args.Tweet.CreatedBy.FollowersCount)	Tweet atanın takipçilerinin sayısı
Hesaparkadassay(args.Tweet.CreatedBy.FriendsCount)	Tweet atanın takip ettiklerinin sayısı
Hesapfavorisay(args.Tweet.CreatedBy.FavouritesCount)	Tweet atanın beğendiği tweet sayısı
Hesaprozetli(args.Tweet.CreatedBy.Verified)	Tweet atan hesabın onaylı rozetli olup olmama durumu
Hesapkorumali(args.Tweet.CreatedBy.Protected)	Tweet atılan hesabın korumalı olup olmama durumu
Hesapkonumacik(args.Tweet.CreatedBy.GeoEnabled)	Tweet atılan hesabın konumunun açık kapalı olma durumu
Yer(args.Tweet.CreatedBy.Location)	Konum yer bilgisi, tweet nereden atıldı.

3.4. Veri Ön İşleme

Tweet verileri içerisinde gürültü olarak nitelendirilebilecek çok fazla anlam ifade etmeyen gereksiz veri vardır. Kısmen veri ön işleme olarak ifade edebileceğimiz verilerin temizlenmesi süreci makine öğrenmesi algoritmaları için önem arz etmektedir. Veri ön işleme işlemleri sınıflandırma algoritmalarının başarımını artırmaktadır. Tweetlerde, gereksiz veriler temizlenip makine öğrenmesi algoritmaları tarafından işlenmeye hazır olduğunda temizlenmiş verilerden eğitim setleri oluşturulmuştur.

Bir duygu analizi yapmak için verilerin temizlenmesi gerekir. Tweetler analiz için gereksiz birçok veri içerir. Ön işleme, etiketleme işlemini daha kolay yapmamızı sağlar. Ham tweet; doğal olarak yazım hatası olan metin, emoji, görüntü, video, web bağlantısı, kullanıcı ismi, boş alan, özel karakter, gereksiz kelimeler, sembol, kısaltma, etiket, tarih, saat, vb. verilerinden oluşur. Verileri uygun hâle getirmek için ham tweetler çeşitli aşamalardan geçirilerek işlenmiştir. Ön işleme sırasında tweet ana metni içinden boş alanlar, gereksiz kelimeler, hashtagler, tekrarlanan kelimeler gibi yukarıda sözü edilen kısımlar temizlenmiştir.

Bir retweet birden fazla farklı kullanıcı tarafından paylaşılan aynı tweettir. Bu çalışmada; tweeti paylaşan kullanıcı ile retweet yapan kullanıcıların aynı fikirde olduğunu gösteren retweetler, veri kümesi içinde tutulmuştur.

Ön işlem sürecinde retweetlerde sadece retweet içinde olan “RT” sözcüğü kaldırılmıştır. Ancak, daha önce de belirtildiği gibi retweetlerin toplam etkisini belirtmesi için tweet dahil edilmiştir. Noktalama işaretlerinin duygu belirtimine katkı sağladığı düşüncesiyle eğitim seti için el ile etiketleme işlemi yapıldıktan sonra noktalama işaretleri ve gereksiz kelimeler kaldırılmıştır.

Eğitim seti oluşturma için yapılan veri ön işlemleri, etiketleme işleminden önce ve sonra olarak 2 aşamada uygulanmıştır. Çizelge 3.2’de yapılan veri işleme işlemleri etiketleme işleminden önce yapılmıştır. Eğitim setlerinin özniteliklerine bağlı olarak Çizelge 3.3’te yapılan işlemler etiketleme işleminden sonra yapılmıştır.

Çizelge 3.2. Etiketleme işleminden önce yapılan veri ön işleme

İçerik	**Ön İşleme**
#kelime (#Fenerbahçe, #TSvFB gibi)	Temizlendi.
@kullanıcıadı kullanıcıların tweetlerine verilen yanıtlar (‘@’ ile başlayanlar)	Temizlendi.
URL’ler, küçük URL’ler ve Link’ler	Temizlendi.
Alfabede olmayan karakterler simgeler	Temizlendi.
Emojiler	Temizlendi.
Birden fazla boşluklar	Temizlendi.

Çizelge 3.3. Etiketleme işleminden sonra yapılan veri ön işleme

İçerik	**Ön İşleme**
Noktalama işaretleri (! ? , . ” : ;)	Temizlendi.
Büyük harfler	Tüm içerik küçük harfe çevrildi.
Sayılar	Temizlendi.
Kelime	Kelimelerin kökleri çıkartıldı.
Tekrar eden kelimeler	Tekillendi.
Gereksiz Kelimeler	Temizlendi.

Tweetlerde bulunan çeşitli içerik türleri ve bu içerikler üzerinde gerçekleştirdiğimiz eylemler Çizelge 3.2 ve Çizelge 3.3’te gösterilmiştir. Temizlenmiş bir tweet örneği Çizelge 3.4’te gösterilmektedir. Bu şekilde, eğitim veri setinin gürültüsünün bir kısmı azaltılmıştır ve işlenmiş tweetler hazırlanmıştır.

Çizelge 3.4. Etiketleme işlemi öncesi ham tweet ve temizlenmiş tweet örneği

Ham Tweet	**Temizlenmiş Tweet**
RT @futbolarena: Trabzonspor, Fenerbahçe kaleisne gonderdıgı 26. şutta 2. golu buldu. Fenerbahçe çok köüt oynyor… #TSLive #TSvFB ☺☺☺ https://t.co/t7fOLB1LYd	Trabzonspor, Fenerbahçe kaleisne gonderdıgı 26.şutta 2.golu buldu. Fenerbahçe çok köüt oynyor...

3.5. Doğal Dil İşleme Kütüphanesi (Zemberek)

Zemberek (Akın, Akın, 2017) kütüphanesi, Türkçe için açık kaynaklı bir doğal dil işleme kütüphanesidir. Bu çalışmada bazı eğitim setlerinin oluşturulması sürecinde Zemberek kütüphanesi kullanılarak yazım denetimi, kelime köklerini bulma, yanlış yazılmış kelimeleri düzeltme ve kelime önerme gibi işlemler yapılmıştır.

Zemberek Kütüphanesi kullanılarak en fazla bir harf ekleme, silme, değiştirme yapılarak ve yazım denetimi ile kelimelerdeki basit yazım hataları düzeltilebilmekte ve yakın kelime önerisinde bulunulabilmektedir.

3.5.1. Yazım Denetimi

Zemberek Kütüphanesi yazım denetimi ile bir metindeki kelimeleri denetlemektir. Yazım denetimi amacıyla Zemberek Kütüphanesi kullanılmıştır. Zemberek Kütüphanesi kullanılarak yapılan denetime örnek Çizelge 3.5'te verilmiştir. Çizelge 3.5'te görüldüğü gibi hem imlaya göre kelime düzeltmesi yapılmış hem de anlama göre kelime düzeltmesi yapılmıştır.

Çizelge 3.5. Zemberek Kütüphanesi kullanılarak yazım denetimi

Ham Tweet	**Zemberek İmlaya Göre Kelime Düzeltmesi Yapılmış ve Temizlenmiş**
RT @futbolarena: Trabzonspor, Fenerbahçe kaleisne gonderdıgı 26. şutta 2. golu buldu. Fenerbahçe çok köüt oynyor… #TSLive #TSvFB ☺☺☺ https://t.co/t7fOLB1LYd	Trabzonspor fenerbahçe kalesine gönderdiği şutta golü buldu fenerbahçe çok kötü oynuyor
Ham Tweet	**Zemberek Anlama Göre Kelime Önerisi Yapılmış ve Temizlenmiş Tweet**
RT @futbolarena: Trabzonspor, Fenerbahçe kaleisne gonderdıgı 26. şutta 2. golu buldu. Fenerbahçe çok köüt oynyor… #TSLive #TSvFB ☺☺☺ https://t.co/t7fOLB1LYd	trabzonspor fenerbahçe kalesine gönderdiği şutta golü buldu fenerbahçe çok kötü oynuyor
Ham Tweet	**Zemberek Yazım Denetimi Yapılmış Tweet**
RT @futbolarena: Trabzonspor, Fenerbahçe kaleisne gonderdıgı 26. şutta 2. golu buldu. Fenerbahçe çok köüt oynyor… #TSLive #TSvFB ☺☺☺ https://t.co/t7fOLB1LYd	trabzonspor fenerbahçe kalesine gönderdiği şutta golü buldu fenerbahçe çok kötü oynuyor

3.6. El ile Etiketleme

Eğitim setleri için el ile etiketleme aşamasında Şekil 3.5'te verilen uygulama arayüzü üzerinden tweetler 4 sınıf olarak etiketlenmiştir. Eğer tweet içinde birden fazla duygu ifade ediliyorsa tweette ifade edilen güçlü olan duygu etiketlenmiştir. Tweetlerde ifade edilen olumlu, olumsuz, tarafsız ve alakasız duygu durumları için el ile etiketleme sürecinde, aşağıdaki belirtilen hususlara dikkat edilmiştir. Bunlar;

Olumlu: Tüm tweette olumlu, mutlu, heyecanlı, neşeli bir tavır varsa veya olumlu çağrışımlarla bir şeyden bahsedilirse o tweet olumlu olarak etikenlenmiştir. Ayrıca, tweette birden fazla duygu durumu ifade edilirse, ancak olumlu duygu daha

baskınsa örneğin; “her şey kötü giderken bana iyi gelen tek şeyin beşiktaş oluşu” şeklinde ise olumlu olarak etiketlenmiştir.

Olumsuz: Eğer tüm tweetin olumsuz, üzgün, memnun olmayan bir tutumu varsa veya olumsuz çağrışımlardan söz edilirse, ayrıca tweette birden fazla duygu ifade edilirse, ancak olumsuz duygusu daha baskınsa o tweet olumsuz olarak etiketlenmiştir.

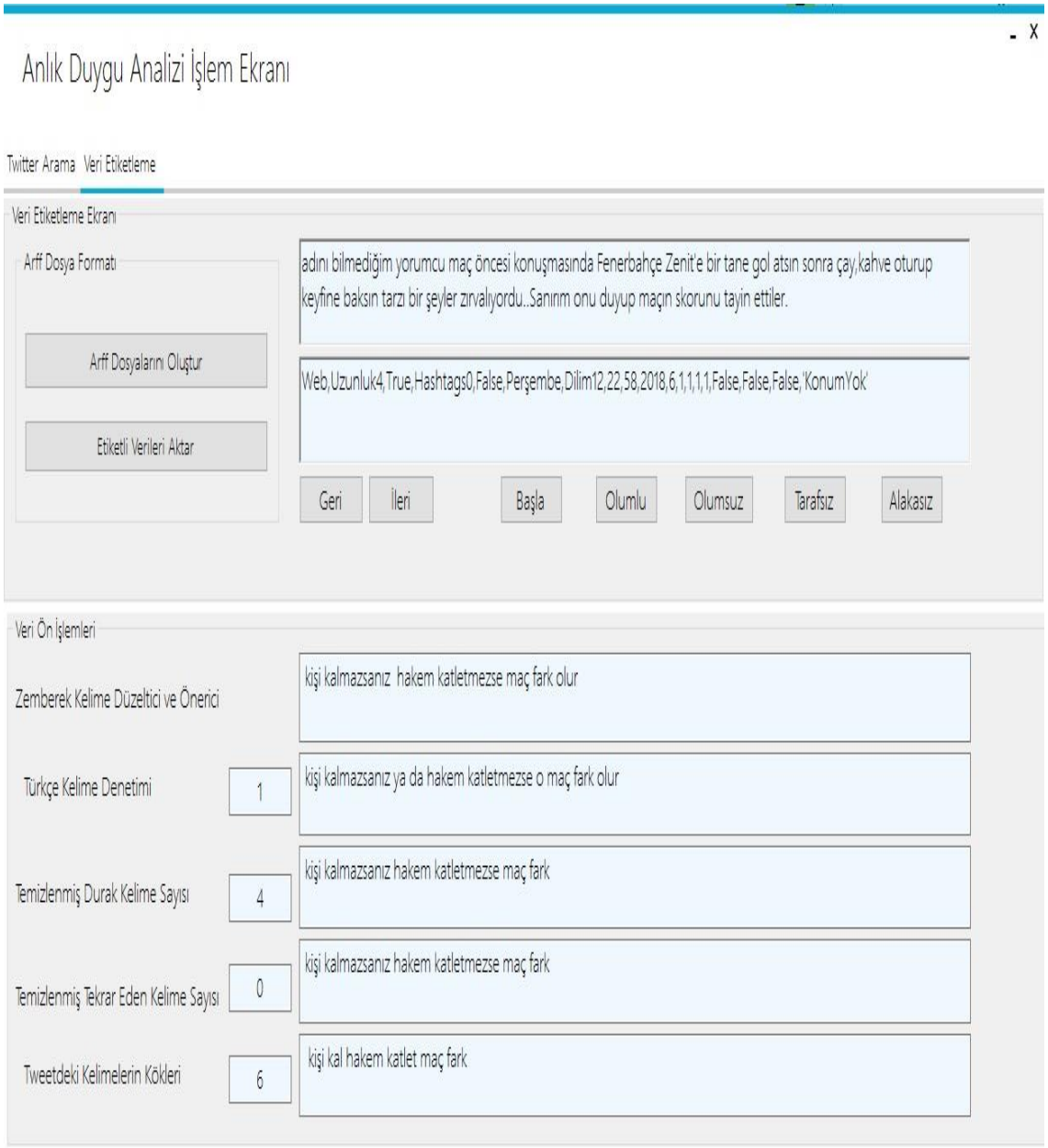
Örneğin; “takımda en hırslı en iyi oynayan adama sallamak ne seviye” şeklinde bir tweet olumsuz olarak etiketlenmiştir.

Tarafsız: Eğer tweeti paylaşan kişi, tweette kişisel bir düşünce, görüş ifade etmez ve sadece bilgi iletirse, ayrıca maçlar ile ilgili tahminler de tarafsız olarak etiketlenmiştir. Örneğin; “hakem ilk yarının sonuna en az 3 dakika uzatma ilave etti” şeklinde bir tweet tarafsız olarak etiketlenmiştir.

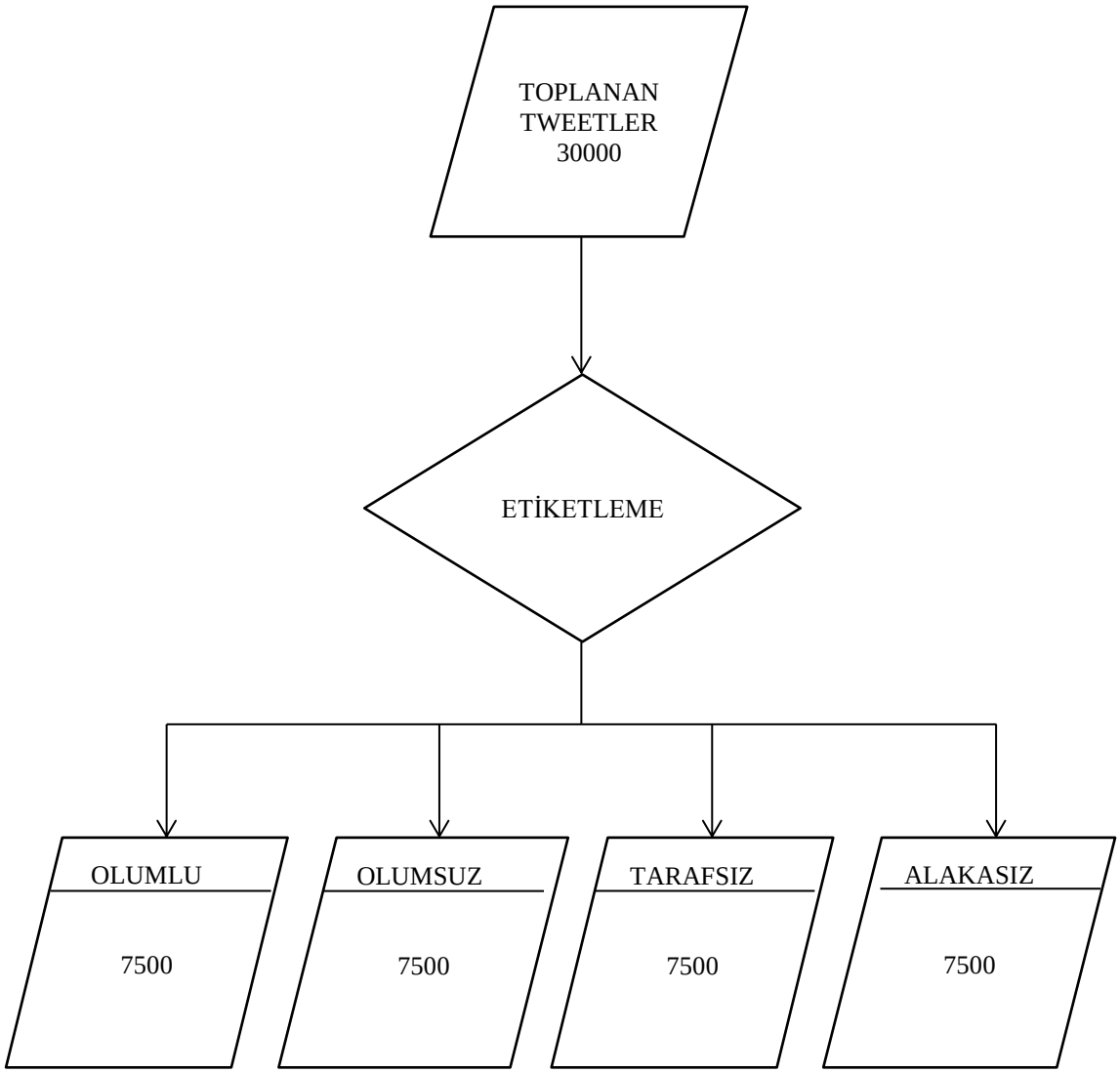
Alakasız: Futbol müsabakalarında taraftarların maçlar ile ilgili attığı tweetler toplandığından, farklı konularda atılan istenmeyen tweetler de alakasız olarak etiketlenmiştir. Ürünlerin reklamları bu kategori altında etiketlenmiştir. Örneğin; “dünyanın en iyi dizi sahnesi aksini söyleyen kiraz mevsimi izliyordur” şeklindeki bir tweet alakasız olarak etiketlenmiştir.

Bunun yanı sıra, etiketleme işleminde kişisel önyargılar etiketlemenin dışında tutulmuştur ve hiçbir varsayım yapılmamıştır. Tweet, geçmiş herhangi bir kişisel bilgiden değil; sadece tweette verilen bilgiden etiketlenmiştir.

Yapılan etiketleme işlemi sırasında farklı özniteliklere sahip eğitim dosyaları arff dosya formatında oluşturulmuştur. Oluşturulan EğitimSeti-1 isimli örnek bir eğitim seti arff dosyası Ek-1’de verilmiştir. Oluşturulan 12 farklı arff dosyasında sınıflar Şekil 3.6’da görüldüğü gibi dengeli dağıtılmıştır.



Şekil 3.5. Veri etiketleme arayüzü



Şekil 3.6. Etiketleme işleminden sonra dengeli veri sınıfları

Tüm veri setleri genel olarak; tweet, işletim sistemi, tweet uzunluğu, tweetin devam edip etmediği durumu, tweetde yer alan hashtag sayısı, tweetin hassas içerik içerip içermediği durumu, tweetin atıldığı gün, tweetin atıldığı zaman aralığı, tweetin atıldığı saat, tweetin atıldığı dakika, tweet atanın hesap açılış yılı, tweet atanın hesap açılış ayı, tweet atanın attığı tweet sayısı, tweet atanın takipçi sayısı, tweet sahibinin takip ettiği kişi sayısı, tweet atanın favori tweet sayısı, tweet atanın hesabının rozetli olup olmadığı durumu, tweet atılan hesabın korumalı olup olmadığı durumu, tweet atılan hesabın konumunun açık olup olmadığı, tweet atılan yer konum bilgisi ve duygu durumu (olumlu, olumsuz, tarafsız, alakasız) özniteliklerine sahiptir.

3.7. Öznitelik Seçimi

Öznitelik seçimi, tüm sahip olunan özniteliklerden bir alt öznitelik seçme işlemidir. Öznitelik seçimi, makine öğrenmesi algoritmalarının sınıflama işleminde bilgi kazancı sağlamayan değersiz ve gereksiz verileri çıkartmayı amaçlamaktadır.

Bu çalışmada bilgi kazançlarına bakılarak bir öznitelik seçme işlemi yapılmıştır ve buna bağlı olarak eğitim setleri oluşturulmuştur. Bu çalışmada kullanılan öznitelik çıkarımında yardımcı olan bazı metin formatlama teknikleri:

Bölme: Bir tweeti sözcüklere, sembollere ve diğer anlamlı ögelere bölme işlemidir. Boşluk karakterleriyle veya noktalama karakterleriyle ayrılabilir. Bu bölme işlemi tweeti oluşturan bileşenlere ayrı ayrı bakabilmek için yapılmıştır (Aday, 2012). Url'ler ve kullanıcı adları sırasıyla "http" ve "@" ile tanımlanır. Bu çalışmada yalnızca tweet metnini analiz etmekle ilgilendiği için "http" ve "@" temizlenmiştir.

- Küçük Harf Dönüştürme: Tweetler küçük harfe dönüştürülmüştür.
- Kelime Köklerini Bulma: Türetilmiş bir kelimeyi köküne indirgeme, metni normalleştirme işlemidir (Hearst, 2009). Bazı eğitim setlerimizde kelimeler Çizelge 3.6'daki gibi kök kelimesine indirgenmiştir.

Çizelge 3.6. Kelimelerin köklerine indirgenmesi

Tweet	**Kök**
Trabzonspor Fenerbahçe Kalesine gönderdiği Şutta golü Buldu	kale gönder şut gol bul
Trabzonspor Fenerbahçe Karşısında Deniyor	karşı dene

Bazı eğitim veri setlerimiz için türemiş kelimeler köklerine indirgenerek kelimenin kök olarak eğitim setinde bir öznitelik olarak bulunmasının sınıflama başarımını nasıl etkileyeceği belirlenmesi araştırılmıştır. Kelimeleri çözümleyip köklerine ayrıştırmak için Zemberek kütüphanesi kullanılmıştır.

- Gereksiz Kelimelerin Kaldırılması: Gereksiz kelimeler, bir metinde kullanıldığında hiçbir değerli bilgi içermeyen ve bu nedenle işe yaramaz olduğu

iddia edilen çok sık kullanılan bazı kelimelerden oluşur (Aday, 2012). Örneğin; "için", "birisi", "ama", "öyle", "yine", "zaten", "mı", vb. gereksiz kelimelerdir. Bu kelimeleri kaldırmak uygundur, çünkü değerli bilgi taşımazlar. Bu nedenle gereksiz kelimeler kaldırılmıştır.

Bir duygu belirtmediği için duygu analizi çalışmasında sınıflama başarımına katkıda bulunmayacağı düşünüldüğünden bazı Türkçe kelimeler çıkarılmıştır. Bu kelimeler gereksiz kelimeler listesine eklenmiştir. Bu çalışmada zamirler ve sayılar da gereksiz kelimeler olarak kabul edilmiştir. Ön işleme sürecinden sonra bunlar da tweetlerden temizlenmiştir. Tüm bu işlemlerde kaldırılan kelimelerin listesi Ek-2'de yer verilmiştir. Öznitelik çıkarımı:

- Boyut azaltmak ve gürültüyü gidermek,
- Sınıflama başarımını artırmak,
- Basitlik ve anlaşılabilirlik sağlamak için yapılmıştır.

Çok sayıda öznitelikten en iyi öznitelikleri seçmek gerekmektedir. Bunun için özniteliklerin bilgi kazancı hesaplanmış ve bilgi kazancı olmayan öznitelikler çıkartılarak değerli öznitelikler seçilmiştir.

Duygu sınıflarına ayırmak için tweetler sayısal vektörler olarak gösterilmektedir. Öznitelik vektörü oluşturma aşaması, metinsel tweetleri uygun vektör formlarına dönüştürmek için uygulanır.

Sınıflandırma için uygun öznitelik vektörleri oluşturmak üzere verilerden çeşitli farklı öznitelikler çıkarılmıştır. Öznitelik vektörlerini oluşturmak için n-gram ayrıştırma kullanılmıştır.

Eğitim veri setleri için belirlenen öznitelikler, açıklamaları aşağıdadır:

Tweet: Paylaşılan tweetin metin kısmı,

İsletimsistemi: Paylaşım yapılan cihazın işletim sistemi,

Tweetuzunluğu: Tweet uzunluğu, tweetin uzunluğunu kategorize ettik,

Tweetkesik: Tweetin devam edip etmeme durumu,

Tweethastagsay: Tweetdeki hashtag sayısı,

Tweethassasicerik: Tweetin hassas içerik içermeme durumu,

Tweetgun: Tweetin atıldığı gün, tweet haftanın hangi günü atıldı,

Tweetzamanaralik: Tweetin atıldığı zaman aralığı, haftanın hangi günü atıldı,

Tweetsadecesaat: Tweet atılma saati,

Tweetsadecedakika: Tweet atılma dakikası 01-59,

Hesapyili: Tweet atanın hesap açılış yılı,

Hesapayi: Tweet atanın hesap açılış ayı,

Hesaptweetsay: Tweet atanın toplam attığı tweet sayısı,

Hesaptakipcisay: Tweet atanın takipçilerinin sayısı,

Hesaparkadassay: Tweet atanın takip ettiklerinin sayısı,

Hesapfavorisay: Tweet atanın beğendiği tweet sayısı,

Hesaprozetli: Tweet atan hesabın onaylı rozetli olup olmama durumu,

Hesapkorumali: Tweet atılan hesabın korumalı olup olmama durumu,

Hesapkonumacik: Tweet atılan hesabın konumunun açık kapalı olma durumu,

Yer: Tweetin atıldığı konum bilgisi.

3.8. Oluşturulan Veri Setleri

Veri setlerinde yer alan verilerimiz birçok ön işlem aşamasından geçerek oluşturulmuştur. Öncelikle Çizelge 3.7'de etiketleme işleminden önce veri ön işleminden geçen tweet durumu görülmektedir.

Çizelge 3.7. Ham tweet ve etiketleme işleminden önce veri ön işleme yapılan tweet

Ham Tweet	Maç özetleri ve Maç yayinlari ligin 20.haftasınnı Tüm Maçları birazdan izlyin bc.vc/THHohfi #maçyayını #maçlinki #yayın #link #bein #beinsports #ligtv #iptv #sopcast #galatasaray #Beşiktaş #sopcast #atacast
Etiketlemeden önce işlenmiş Tweet	Maç özetleri ve Maç yayinlari ligin 20.haftasınnı Tüm Maçları birzadan izlyin

Etiketleme işleminden sonra oluşturulan arff dosya formatındaki 12 farklı eğitim seti arff formatı dosya içerikleri aşağıdaki özniteliklerden oluşturulmuştur.

EğitimSeti-1: Tweet metnine belirlediğimiz öznitelikler eklenmiştir.

Öznitelikler:

“tweet, isletimsistemi, tweetuzunluk, tweetkesik, tweethastagsay, tweethassasicerik, tweetgun, tweetzamanaralik, tweetsadecesaat, tweetsadecedakika, hesapyili, hesapyayi, hesaptweetsay, hesaptakipcisay, hesaparkadassay, hesapfavorisay, hesaprozetli, hesapkorumali, hesapkonumacik, yeradi, sınıf ”

Belirlenen öznitelikler yapılan işlemlerden sonra Çizelge 3.8’de görüldüğü gibi arff dosya formatında ifade edilmiştir.

Çizelge 3.8. EğitimSeti-1 arff dosya formatı bir örnek satırı

'maç özetleri ve maç yayinlari ligin haftasınnı tüm maçları birzadan izlyin' ,Web,Uzunluk4,False,Hashtags1,False,Pazar,Dilim9,17,24,2019,1,1,1,1,1,False,False ,False,'KonumYok',Olumlu

EğitimSeti-2: Tweet metninde Türkçe kelime denetimi yapılmıştır ve belirlediğimiz özniteliklere ayrıca Türkçe kelime denetimi sonunda düzeltilen kelime sayıları da eklenmiştir.

Öznitelikler:

“tweet, isletimsistemi, tweetuzunluk, tweetkesik, tweethastagsay, tweethassasicerik, tweetgun, tweetzamanaralik, tweetsadecesaat, tweetsadecedakika, hesapyili, hesapyayi, hesaptweetsay, hesaptakipcisay, hesaparkadassay, hesapfavorisay, hesaprozetli, hesapkorumali, hesapkonumacik, yeradi, denetimsay, sınıf ”

Belirlenen öznitelikler ve yapılan işlemlerden sonra Çizelge 3.9’da görüldüğü gibi arff dosya formatında ifade edilmiştir.

Çizelge 3.9. EğitimSeti-2 arff dosya formatı bir örnek satırı

'maç özetleri ve maç yayınları ligin haftasınnı tüm maçları birzadan izlyin' ,Web,Uzunluk4,False,Hashtags1,False,Pazar,Dilim9,17,24,2019,1,1,1,1,1,False,False, False,'KonumYok' ,1,Olumlu

EğitimSeti-3: Tweet metninde Türkçe kelime denetimi yapılmıştır.

Öznitelikler:

“tweet, isletimsistemi, tweetuzunluk, tweetkesik, tweethastagsay, tweethassasicerik, tweetgun, tweetzamanaralik, tweetsadecesaat, tweetsadecedakika, hesapyili, hesapyayi, hesaptweetsay, hesaptakipcisay, hesaparkadassay, hesapfavorisay, hesaprozetli, hesapkorumali, hesapkonumacik, yeradi, sınıf ”

Belirlenen öznitelikler ve yapılan işlemlerden sonra Çizelge 3.10’da görüldüğü gibi arff dosya formatında ifade edilmiştir.

Çizelge 3.10. EğitimSeti-3 arff dosya formatı bir örnek satırı

'maç özetleri ve maç yayınları ligin haftasınnı tüm maçları birzadan izlyin' ,Web,Uzunluk4,False,Hashtags1,False,Pazar,Dilim9,17,24,2019,1,1,1,1,1,False,False, False,'KonumYok' ,1,Olumlu

EğitimSeti-4: Tweet metninde gereksiz kelimeler temizlenmiş ve belirlediğimiz özniteliklere temizlenen gereksiz kelime sayısı eklenmiştir.

Öznitelikler:

“tweet, isletimsistemi, tweetuzunluk, tweetkesik, tweethastagsay, tweethassasicerik, tweetgun, tweetzamanaralik, tweetsadecesaat, tweetsadecedakika, hesapyili, hesapyayi, hesaptweetsay, hesaptakipcisay, hesaparkadassay, hesapfavorisay, hesaprozetli, hesapkorumali, hesapkonumacik, yeradi, duraksay, sınıf ”

Belirlenen öznitelikler ve yapılan işlemlerden sonra Çizelge 3.11’de görüldüğü gibi arff dosya formatında ifade edilmiştir.

Çizelge 3.11. EğitimSeti-4 arff dosya formatı bir örnek satırı

'maç özetleri maç yayınları ligin haftasının maçları birzadan izlyin' ,Web,Uzunluk4,False,Hashtags1,False,Pazar,Dilim9,17,24,2019,1,1,1,1,1,False,False, False,'KonumYok',2,Olumlu

EğitimSeti-5: Tweet metninde gereksiz kelimeler temizlenmiş ve belirlediğimiz öznitelikler eklenmiştir.

Öznitelikler:

“tweet, isletimsistemi, tweetuzunluk, tweetkesik, tweethastagsay, tweethassasicerik, tweetgun, tweetzamanaralik, tweetsadecesaat, tweetsadecedakika, hesapyili, hesapyayi, hesaptweetsay, hesaptakipcisay, hesaparkadassay, hesapfavorisay, hesaprozetli, hesapkorumali, hesapkonumacik, yeradi, sınıf ”

Belirlenen öznitelikler ve yapılan işlemlerden sonra Çizelge 3.12’de görüldüğü gibi arff dosya formatında ifade edilmiştir.

Çizelge 3.12. EğitimSeti-5 arff dosya formatı bir örnek satırı

'maç özetleri maç yayınları ligin haftasınnı maçları birzadan izlyin' ,Web,Uzunluk4,False,Hashtags1,False,Pazar,Dilim9,17,24,2019,1,1,1,1,1,False,False, False,'KonumYok',Olumlu

EğitimSeti-6: Tweet metninde yer alan kelimelerinin kelime kökleri elde edilip belirlediğimiz özniteliklere; kelimelerinin kökleri, ayrıca kaç tane kelime kökü olduğunun sayısı eklenmiştir.

Öznitelikler:

"tweet, isletimsistemi, tweetuzunluk, tweetkesik, tweethastagsay, tweethassasicerik, tweetgun, tweetzamanaralik, tweetsadecesaat, tweetsadecedakika, hesapyili, hesapyayi, hesaptweetsay, hesaptakipcisay, hesaparkadassay, hesapfavorisay, hesaprozetli, hesapkorumali, hesapkonumacik, yeradi, koksay, sınıf "

Belirlenen öznitelikler ve yapılan işlemlerden sonra Çizelge 3.13'te görüldüğü gibi arff dosya formatında ifade edilmiştir.

Çizelge 3.13. EğitimSeti-6 arff dosya formatı bir örnek satırı

'maç özet yayın lig maç' ,Web,Uzunluk4,False,Hashtags1,False,Pazar,Dilim9,17,24,2019,1,1,1,1,1,False,False ,False,'KonumYok',5,Olumlu

EğitimSeti-7: Tweet metninde yer alan kelimelerin kelime kökleri elde edilmiş ve belirlediğimiz özniteliklere kelime kökleri eklenmiştir.

Öznitelikler:

"tweet, isletimsistemi, tweetuzunluk, tweetkesik, tweethastagsay, tweethassasicerik, tweetgun, tweetzamanaralik, tweetsadecesaat, tweetsadecedakika, hesapyili, hesapyayi, hesaptweetsay, hesaptakipcisay, hesaparkadassay, hesapfavorisay, hesaprozetli, hesapkorumali, hesapkonumacik, yeradi, sınıf "

Belirlenen öznitelikler ve yapılan işlemlerden sonra Çizelge 3.14'te görüldüğü gibi arff dosya formatında ifade edilmiştir.

Çizelge 3.14. EğitimSeti-7 arff dosya formatı bir örnek satırı

'maç özet yayın lig maç' ,Web,Uzunluk4,False,Hashtags1,False,Pazar,Dilim9,17,24,2019,1,1,1,1,1,False,False, False,'KonumYok',Olumlu

EğitimSeti-8: Tweet metninde tekrar eden kelimeler varsa tekrar eden kelimeler çıkartılmış ve aşağıdaki öznitelikler eklenmiştir.

Öznitelikler:

“tweet, isletimsistemi, tweetuzunluk, tweetkesik, tweethastagsay, tweethassasicerik, tweetgun, tweetzamanaralik, tweetsadecesaat, tweetsadecedakika, hesapyili, hesapyayi, hesaptweetsay, hesaptakipcisay, hesaparkadassay, hesapfavorisay, hesaprozetli, hesapkorumali, hesapkonumacik, yeradi, sınıf ”

Belirlenen öznitelikler ve yapılan işlemlerden sonra Çizelge 3.15’te görüldüğü gibi arff dosya formatında ifade edilmiştir.

Çizelge 3.15. EğitimSeti-8 arff dosya formatı bir örnek satırı

'maç özetleri yayınları ligin haftasınnı maçları birzadan izlyin' ,Web,Uzunluk4,False,Hashtags1,False,Pazar,Dilim9,17,24,2019,1,1,1,1,1,False,False ,False,'KonumYok',Olumlu

EğitimSeti-9: Tweet metninde tekrar eden kelimeler varsa tekrar eden kelimeler çıkartılmış ve belirlediğimiz özniteliklere ayrıca tekrar eden kelime sayısı eklenmiştir.

Öznitelikler:

“tweet, isletimsistemi, tweetuzunluk, tweetkesik, tweethastagsay, tweethassasicerik, tweetgun, tweetzamanaralik, tweetsadecesaat, tweetsadecedakika, hesapyili, hesapyayi, hesaptweetsay, hesaptakipcisay, hesaparkadassay, hesapfavorisay, hesaprozetli, hesapkorumali, hesapkonumacik, yeradi, tekrarsay, sınıf ”

Belirlenen öznitelikler ve yapılan işlemlerden sonra Çizelge 3.16’da görüldüğü gibi arff dosya formatında ifade edilmiştir.

Çizelge 3.16. EğitimSeti-9 arff dosya formatı bir örnek satırı

'maç özetleri yayınları ligin haftasınnı maçları birzadan izlyin' ,Web,Uzunluk4,False,Hashtags1,False,Pazar,Dilim9,17,24,2019,1,1,1,1,1,False,False ,False,'KonumYok',2,Olumlu

EğitimSeti-10: Tweet metninde Zemberek Kütüphanesi kullanılarak Türkçe kelime denetimi yapılmış ve kelimeler hem imlaya hem de anlama göre düzeltilmiş ve belirlediğimiz öznitelikler eklenmiştir.

Öznitelikler:

“tweet, isletimsistemi, tweetuzunluk, tweetkesik, tweethastagsay, tweethassasicerik, tweetgun, tweetzamanaralik, tweetsadecesaat, tweetsadecedakika, hesapyili, hesapyayi, hesaptweetsay, hesaptakipcisay, hesaparkadassay, hesapfavorisay, hesaprozetli, hesapkorumali, hesapkonumacik, yeradi, sınıf ”

Belirlenen öznitelikler ve yapılan işlemlerden sonra Çizelge 3.17’de görüldüğü gibi arff dosya formatında ifade edilmiştir.

Çizelge 3.17. EğitimSeti-10 arff dosya formatı bir örnek satırı

'maç özetleri ve maç yayınları ligin haftasının tüm maçları birazdan izleyin' ,Web,Uzunluk4,False,Hashtags1,False,Pazar,Dilim9,17,24,2019,1,1,1,1,1,False,False ,False,'KonumYok',Olumlu

EğitimSeti-11: Tweet metni ve sadece duygu durumu eklenmiştir.

Öznitelikler:

“tweet, sınıf ”

Belirlenen öznitelikler ve yapılan işlemlerden sonra Çizelge 3.18’de görüldüğü gibi arff dosya formatında ifade edilmiştir.

Çizelge 3.18. EğitimSeti-11 arff dosya formatı bir örnek satırı

'maç özetleri ve maç yayinlari ligin haftasınnı tüm maçları birzadan izlyin',Olumlu

EğitimSeti-12: Tweet metninde sadece url, hashtag temizliği yapılmıştır. Sayılar, sayısal değerler temizlenmemiş ve belirlediğimiz öznitelikler eklenmiştir.

Öznitelikler:

“tweet, isletimsistemi, tweetuzunluk, tweetkesik, tweethastagsay, tweethassasicerik, tweetgun, tweetzamanaralik, tweetsadecesaat, tweetsadecedakika, hesapyili, hesapyayi, hesaptweetsay, hesaptakipcisay, hesaparkadassay, hesapfavorisay, hesaprozetli, hesapkorumali, hesapkonumacik, yeradi, sınıf ”

Belirlenen öznitelikler ve yapılan işlemlerden sonra Çizelge 3.19’da görüldüğü gibi arff dosya formatında ifade edilmiştir.

Çizelge 3.19. EğitimSeti-12 arff dosya formatı bir örnek satırı

'maç özetleri ve maç yayinlari ligin 20 haftasınnı tüm maçları birzadan izlyin' ,Web,Uzunluk4,False,Hashtags1,False,Pazar,Dilim9,17,24,2019,1,1,1,1,1,False,False ,False,'KonumYok',Olumlu

Etiketlendikten sonra on iki birbirinden farklı arff dosyası oluşturulmuştur. Oluşturulan dosya isimleri ve tweete yapılan işlem sonunda oluşan veri setlerinin içerikleri Çizelge 3.20’de gösterilmiştir.

Çizelge 3.20. Etiketleme işleminden sonra dosya isimleri ve işlenmiş tweet içerikleri

Dosya Adı	**Tweet**
EğitimSeti-1	'maç özetleri ve maç yayinlari ligin haftasınnı tüm maçları birzadan izlyin'
EğitimSeti-2	'maç özetleri ve maç yayınları ligin haftasınnı tüm maçları birzadan izlyin'
EğitimSeti-3	'maç özetleri ve maç yayınları ligin haftasınnı tüm maçları birzadan izlyin'
EğitimSeti-4	'maç özetleri maç yayınları ligin haftasınnı maçları birzadan izlyin'
EğitimSeti-5	'maç özetleri maç yayınları ligin haftasınnı maçları birzadan izlyin'
EğitimSeti-6	'maç özet yayın lig maç'
EğitimSeti-7	'maç özet yayın lig maç'
EğitimSeti-8	'maç özetleri yayınları ligin haftasınnı maçları birzadan izlyin'
EğitimSeti-9	'maç özetleri yayınları ligin haftasınnı maçları birzadan izlyin'
EğitimSeti-10	'maç özetleri ve maç yayınları ligin haftasının tüm maçları birazdan izleyin'
EğitimSeti-11	'maç özetleri ve maç yayinlari ligin haftasınnı tüm maçları birzadan izlyin'
EğitimSeti-12	'maç özetleri ve maç yayinlari ligin 20 haftasınnı tüm maçları birzadan izlyin'

BÖLÜM 4

MAKİNE ÖĞRENMESİ

4.1. Makine Öğrenmesi

Makine Öğrenmesi; Nvidia'ya göre veriyi ayrıştırmak, veriden öğrenmek ve ardından dünyadaki bir şey hakkında bir tahmin yapmak için algoritmalar kullanma uygulamasıdır (Nvidia, 2019). İşleyeceğimiz veriyi nasıl işleyip bilgiye dönüştüreceğimiz hakkında elimizde algoritma olmadığı durumlarda, verileri makine öğrenmesi algoritmaları ile işleyerek veriler hakkında karar verebilmemiz mümkün olabilmektedir. Bazı makine öğrenmesi algoritmaları çok sayıda mevcut örnek veriyi işleyerek verinin hakkında bilgi sahibi olup sonraki veriler hakkında karar verirken önceden elde ettiği bu bilgiyi kullanır. Bu tür yapılan makine öğrenmesine denetimli öğrenme, örnek veri setine eğitim seti ve bu amaçla kullanılan algoritmalara sınıflandırıcı algoritmalar denir. Sınıflandırıcı algoritma denilmesi ise mevcut örnek verilerin her birinin bilinen bir sınıfının önceden belirlenmiş olmasındandır. Eğitim setindeki sınıfları belirlenmiş örnek veri için sınıflandırıcı algoritmaları örnek veri ile sınıflar arasındaki ilişkilere bağlı olarak veri setini temsil eden bir model ortaya çıkarmaktadır. Bu modelin başarısı ise veri seti içindeki örnek veri sayısına göre değişiklik göstermektedir. Bu set içindeki örnek veri sayısı az ise öğrenme yetersiz olur ve dolayısıyla başarısız bir model ortaya çıkar ve buna yetersiz öğrenme durumu adı verilir. Aksi durum, yani veri setindeki veri sayısı olması gerekenden daha fazla ise bu durumda aşırı öğrenme olarak ifade edilen, neredeyse her sınıfı temsil edebilecek tüm veriler veri setinde mevcut olur. Bu da makine öğrenmesi yaklaşımından ziyade kural tabanlı bir sonuç doğurur. Yetersiz ve aşırı öğrenme durumlarının önüne geçmek için ise örnek veri seti eleman sayısının en uygun olması gerekir.

Bazı makine öğrenmesi algoritmaları ise verilen veri seti içindeki düzenlilikleri ortaya çıkarmayı hedeflemektedir. Bu amaçla kullanılan algoritmalar ise denetimsiz öğrenme algoritmaları olarak isimlendirilir ve daha çok kümeleme ismi verilen veri gruplama ya da veriler arası ilişkiyi belirtme amacıyla kullanılır. Duygu analizinde makine öğrenmesi; sınıflandırıcı algoritmalar kullanılarak yani denetimli öğrenme yapılarak veriden olumlu, olumsuz, tarafsız vb. gibi duygu sınıflarını belirlemek için kullanılır. Çalışmamızda eğitim veri setleri kullanılarak veriyi sınıflandırmaya çalışan modeller oluşturulmuştur. Daha sonra modeller, yeni gelen verilerin sınıflarını tahmin etmek için kullanılmıştır.

4.1.1. Denetimli Makine Öğrenmesi Algoritmaları

En çok kullanılan denetimli makine öğrenmesi algoritmaları (Wu vd., 2007), veri setine uygun olması göz önünde bulundurularak çalışmada kullanılmıştır. Kullanılan denetimli makine öğrenmesi algoritmaları;

- C4.5 Algoritması,
- Destek Vektör Makineleri (Support Vector Machine - SVM),
- Naive Bayes Algoritması,
- En Yakın k Komşu Algoritması (k-Nearest Neighbors - KNN),

Bu çalışmada sınıflandırma için veriye uygun olan ve farklı yaklaşımlara dayanan sınıflandırma algoritmaları kullanılmıştır. Olasılık tabanlı olması yönüyle Naive Bayes algoritması, karar çizgileri ile bölme yaparak sınıflama yapan ve yüksek boyutlu verilere daha uygun olan SVM, verilerin benzerliklerini kullanarak sınıflandırma yapan KNN algoritması, kullanılan her özniteliğe ait bilgi kazanç oranı ölçütüne dayanarak karar ağacı oluşturan C4.5 algoritmaları seçilmiştir.

C4.5 algoritması: C4.5 sınıflandırıcı bir karar ağacıdır ve bu ağaç kökünden yapraklara ayrılır. Bir karar ağacı; düğüm, kök, dal ve yapraktan oluşan bir veri yapısıdır. Karar ağacı, eğitim veri setinden, sınıflandırma için ayırt edilebilecek en iyi öznitelikleri kullanarak oluşturulur. Burada amaç, ağaç seviyelerinin ve ağaç düğümlerinin sayısını en aza indirgemektir. Seçilen benzersiz bir değeri temsil eden her öznitelikten ağaç düğümü oluşturulur ve seçilen öznitelik alt düğümlere bölünür. Bölünecek öznitelik sayısı belirli bir eşiğin altında olduğunda bölme sona erer. Eğitim

verilerinin eksik öznitelik değerleri ile kullanılması için C4.5 öznitelik değerlerinin eksik olanlarını "?" olarak işaretlenmesine izin verir. Eksik öznitelik değerleri kazanç ve entropi hesaplamalarında kullanılmaz. Hata bazlı budama, büyüme aşamasından sonra gerçekleştirilir. C4.5 sayısal özellikleri işleyebilir. Sürekli ve kategorik verilerle ve eksik değerlerle çalışabilir.

C4.5, eğitim veri setini analiz eder ve bilgi kazanç oranlarını baz alarak sınıflandırma yapacak tek bir ağaç oluşturur. Bu şekilde, daha genelleştirilmiş ağaçlar yaratır. Ayrıca, algoritma kazanım maksimizasyonuna dayanan öznitelikleri nominal özniteliklere dönüştürür ve bu şekilde sürekli verileri işleyebilir. Sınıfı bilinmeyen bir verinin C4.5 ağacı ile sınıf tahminine, karar ağacının kökünden başlanır; karar düğümlerinde ağaçta hangi düğüme yönlenileceği o düğümde yazan kurallara göre belirlenir ve bu sınıfların tutulduğu yaprak düğüme ulaşıncaya kadar devam eder. Ayrıca, düğümleri silmek veya iç yapıyı değiştirmek için ağacın üzerinden geriye gider. C4.5 tek bir ağaçtır ve aşırı büyümeye karşı korunmak için hata budama tekniğini kullanır. Budama sonrası C4.5 ilk karar ağacını yaratır ve yarattıktan sonra ağaçtan geriye doğru gider ve yardım etmeyen dalları yaprak düğümleriyle değiştirerek çıkarmaya çalışır.

Destek vektör makinaları (SVM): SVM, metin içindeki duygu sınıflarına göre kelimeleri saymak ve metnin duygu durumunu belirlemek için duygu durumu belirten kelimelerin sayısını karşılaştırmaktır. SVM, sınıflar arasında en iyi ayrımı yapan doğrusal ayırıcıyı bulur.

SVM, genellikle iki sınıfa ait olan sınıfların sınıf tahminlerini yapar. SVM, verileri ikiye bölen bir hiper düzlemi bulmaya çalışır. Sonra sınıfları karar sınırı oluşturan bir hiper düzlemle ayırır. Hiper düzlem iki sınıfı en iyi ayıran bir sınırdır. Böylece nominal özellikleri ayırır, ikili olarak dönüştürür ve normalleştirir.

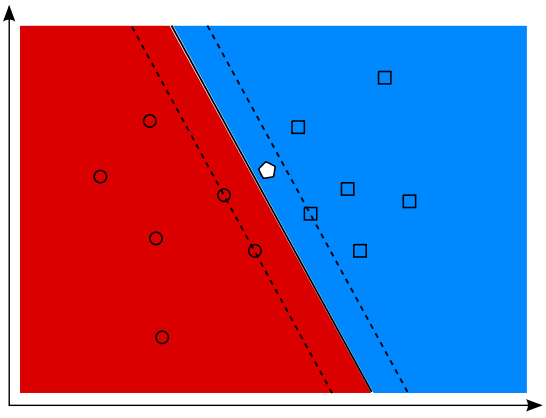
SVM'nin amacı, sınıflandırmadaki hatayı en aza indiren hiper düzlemde sınıfları birbirinden olabildiğince maksimum uzaklıkta sınıflara ayırmaktır. Hiper düzlemi konumlandırmak için SVM her bir sınıftan en kritik örnekleri bulur ve ardından hiper düzlemle ayırır, böylece her bir sınıfın hiper düzlem uzaklığı mümkün olduğu kadar büyük olur. Bu sayede, uzaklığın yeni verilerin doğru şekilde sınıflandırılmasına katkıda bulunmasını sağlar. Şekil 4.1'de görüldüğü gibi bu hiper düzlemin sınırını

belirleyen her sınıftan kritik sınıflar destek vektörleri olarak adlandırılır.

SVM, ayırıcı bir hiper düzlem tarafından ayırt edici bir sınıflandırıcıdır. etiketli eğitim verileri verildiğinde, algoritma yeni metni sınıflandıran en uygun hiper düzlemi çıkarır.

SVM olasılıksal olmayan doğrusal bir sınıflandırıcıdır. Şekil 4.2'de görüldüğü gibi verileri her bir sınıftan ayırmak için düz bir çizgi çizilebilir, ancak verilerin dağılımı her zaman böyle değildir. Veriler çoğu zaman Şekil 4.3'te görüldüğü gibidir, iki sınıfı doğrusal olarak ayırmak mümkün değildir. SVM, sınıfları daha yüksek bir boyuta yansıtan çekirdek işlevi olarak adlandırılan bir işlemi uygulayarak bunu çözer. Bu çekirdek işlevini kullanarak veri kümesini daha yüksek veya sonsuz boyutlu bir uzaya yerleştirir. Bazıları doğrusal, polinom, sigmoid fonksiyonlar ve radyal fonksiyon olan birkaç çekirdek işlevi fonksiyonu vardır. Böylece, sınıflar arasındaki doğrusal ayrımı sağlar (Giachanou, 2016).

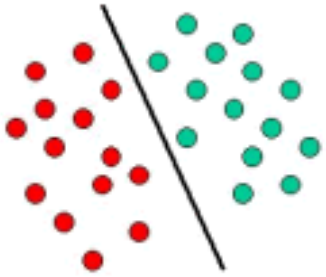
SVM, tüm verileri doğrusal olarak aynı sınıfta olmaya zorlarsa yeni veriler yanlış sınıflandırılabilir. Öte yandan, çok fazla hassasiyet, eğitim hatalarına yol açmaktadır. Şekil 4.1'de, ayırıcı bir hiper düzlem ve destek vektörleri gösterilmiştir. Kesik çizgiler iki sınıf arasındaki en büyük sınırı göstermektedir. Düz çizgi ile ayırma hiper düzlemini gösterir.



Şekil 4.1. SVM algoritması kesikli çizgiler ve düz çizgi

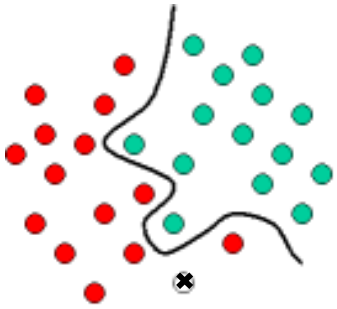
Şekil 4.2'de görüldüğü gibi ayırma çizgisinin sağındaki tüm nesnelerin "yeşil" olduğu ve ayırma çizgisinin solundaki tüm nesnelerin "kırmızı" olduğu bir sınırı belirtir. Sağa gelen her yeni nesne yeşil olarak sınıflandırılmıştır. Ayırma çizgisinin soluna

gelen her yeni nesne kırmızı olarak sınıflandırılmıştır.



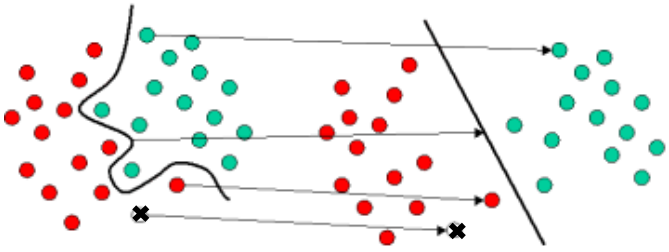
Şekil 4.2. SVM algoritması doğrusal ayırma çizgi

Şekil 4.2'de görüldüğü gibi SVM, sınıfları bir çizgiyle ayıran bir sınıflandırıcıdır. Bununla birlikte çoğu sınıflandırmada doğrusal bir ayrım yapmak, doğru şekilde sınıflandırmak zordur. Şekil 4.2'de, yeşil ve kırmızı nesnelerin ayrılabilmesi için bir eğri gerekmektedir. Farklı sınıfları ayırt etmek için ayrı çizgiler çizmeye dayanan sınıflandırma görevleri, hiper düzlem sınıflandırıcıları olarak bilinir. Destek Vektör Makineleri, bu tür görevleri yerine getirmek için özellikle uygundur.



Şekil 4.3. SVM algoritması eğri çizgi

Şekil 4.3'te SVM, kırmızı ve yeşil nesneleri ayırmak için kullanılan eğriyi yeniden düzenleme işlemiyle en uygun çizgiyi bularak Şekil 4.4'te doğrusal olarak ayırmıştır.

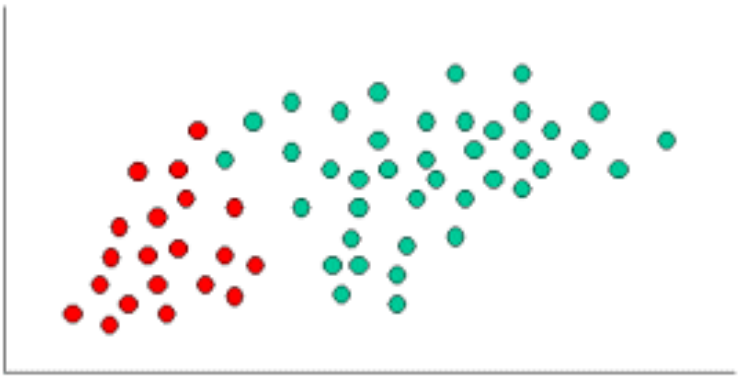


Şekil 4.4. SVM algoritması eğriyi düzenleme işlemi

Naive Bayes: Naive Bayes algoritması, sıralanmamış bir kelime kümesi olan bir kelime torbasıyla çalışır. Her kelimenin frekansı korunur, fakat pozisyonu korunmaz. Bir metin verildiğinde, Naive Bayes, sınıfı tüm sınıfların dışına çıkma ihtimalinin en yüksek olduğu sınıfa geri döndürür. Bu sınıflandırıcı, verilen bir öznitelik setinin ait olduğu etiketi belirlemek için Bayes Teoremi'ni kullanır. Naive Bayes algoritması, Bayes Teoremi ile olasılıkları hesaplar. Her bir sınıf için olasılık hesaplanır. Sınıflandırıcı daha sonra en yüksek olasılığı veren sınıfı tahmin etmektedir (Kamber, Micheline Pei, Jian Han, Jiawei, 2011).

Bayes, teoremi şöyle tanımlamaktadır: Her özniteliğin sınıflandırma problemine eşit ve bağımsız olarak katkıda bulunduğunu varsayan istatistiksel bir yöntemdir. Bu, öznitelikler arasındaki korelasyonun dikkate alınmadığı, bir özniteliğin değerini bilen, diğer öznitelikler hakkında hiçbir şey söylemediğiniz anlamına gelir. Bununla birlikte, Naive Bayes etkili olduğunu kanıtlamıştır ve pratik uygulamalar için kapsamı geniştir (Irina Rish, 2001).

Naive Bayes, çok çeşitli sınıflandırma görevlerinde kullanılabilen olasılıksal bir makine öğrenme algoritmasıdır. Tipik uygulamaları arasında spam filtreleme, dokümanları sınıflandırma, duyarlılık tahmini vb. vardır.



Şekil 4.5. Kırmızı ve yeşil sınıfların dağılımı

Şekil 4.5'te gösterilen örnekte belirtildiği gibi sınıflar "yeşil" veya "kırmızı" olarak sınıflandırılmıştır. Mevcut olan sınıflara dayanarak yeni gelen bilginin hangi sınıfa ait olduklarına karar vermektedir.

Kırmızıdan iki kat daha yeşil sınıf bulunduğundan, gelmemiş yeni bir verinin

"kırmızı" yerine "yeşil" sınıfta olma ihtimalinin iki katı olduğuna inanmak mantıklıdır. Bayes analizinde, bu öncelik olasılık olarak bilinir. Öncelikli olasılıklar önceki deneyime dayanır; bu durumda Naive Bayes algoritması, "yeşil" ve "kırmızı" sınıfların yüzdesi ile sonuçları tahmin etmek için kullanılır. Denklem 4.1'de yeşil sınıf ve Denklem 4.2'de kırmızı sınıf için öncelikli olasılık formülü verilmiştir.

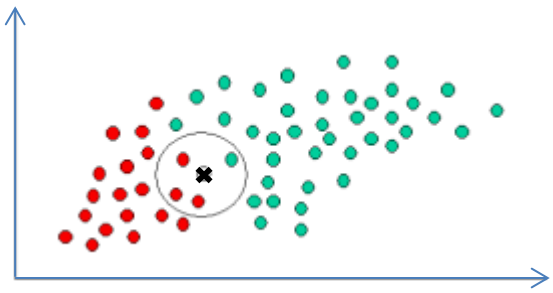
$$Yeşil\ Sınıf\ \text{İçin}\ Öncelikli\ Olasılık\ = \frac{\text{Yeşil Sınıf Sayısı}}{\text{Toplam Sınıf Sayısı}} \tag{4.1}$$

$$Kırmızı\ Sınıf\ \text{İçin}\ Öncelikli\ Olasılık\ = \frac{\text{Kırmızı Sınıf Sayısı}}{\text{Toplam Sınıf Sayısı}} \tag{4.2}$$

40'ı yeşil ve 20'si kırmızı olmak üzere toplam 60 nesne bulunduğundan, sınıf tahmini için öncelikli olasılıklar Denklem 4.3 ve 4.4 görüldüğü gibi hesaplanmıştır:

$$Yeşil\ Sınıf\ \text{İçin}\ Öncelikli\ Olasılık\ = \frac{40}{60} \tag{4.3}$$

$$Kırmızı\ Sınıf\ \text{İçin}\ Öncelikli\ Olasılık\ = \frac{20}{60} \tag{4.4}$$



Şekil 4.6. X verisinin sınıflar içindeki konumu

Öncelikli olasılık belirledikten sonra Şekil 4.6'da görüldüğü gibi yeni bir sınıflandırma için yeni gelen "x" verisinin sınıf tahmini yapılır. Sınıfların iyi bir şekilde kümelenmesi nedeniyle verinin çevresinde ne kadar yeşil veya kırmızı sınıfların daha fazla olması, o belirli renge ait olma ihtimalinin daha yüksek olduğu varsayılır. Bu x verisinin olasılığını ölçmek için sınıf etiketlerine bakılmaksızın bir daire çizilerek her bir sınıf etiketine ait olan dairedeki puan sayısı hesaplanır. Denklem 4.5'te ve Denklem 4.6'da x'in olasılık formülleri verilmiştir:

$$X'\text{in Yeşil Sınıf Olasılığı} = \frac{X\text{'in Yakınındaki Yeşil Sınıf Sayısı}}{\text{Toplam Yeşil Sınıf Sayısı}} \quad (4.5)$$

$$X'\text{in Kırmızı Sınıf Olasılığı} = \frac{X\text{'in Yakınındaki Kırmızı Sınıf Sayısı}}{\text{Toplam Kırmızı Sınıf Sayısı}} \quad (4.6)$$

Yukarıdaki şekilden, x'in yeşil sınıf olasılığı, x'in kırmızı sınıf olasılığından daha küçüktür, çünkü dairede 1 yeşil sınıf ve 3 kırmızı sınıf vardır. Denklem 4.7 ve Denklem 4.8'de x'in olasılıkları hesaplanmıştır:

$$X'\text{inYeşil Sınıf Olasılığı} = \frac{1}{40} \quad (4.7)$$

$$X'\text{in Kırmızı Sınıf Olasılığı} = \frac{3}{20} \quad (4.8)$$

Önceki olasılıklar; x'in, yeşil sınıfın kırmızı sınıftan iki kat fazla olmasından dolayı yeşil sınıf'a ait olabileceğini gösterir. X sınıf üyeliğinin kırmızı olduğu x civarında yeşil'den daha fazla kırmızı nesne olduğu düşünülür. Bayes analizinde son sınıflandırma; her olasılık birleştirilerek, Bayes kullanılarak bir arka olasılık oluşturmak üzere üretilir. Denklem 4.9 ve Denklem 4.10'da x'in arka olasılıkları hesaplanmıştır:

$X'in\ Yeşil\ Sınıf\ Olma\ arka\ olasılığı =$ Yeşil Sınıfın Önceki Olasılığı X "X"'in Yeşil Sınıf Olasılığı

$$X'in\ Yeşil\ Sınıf\ Olma\ arka\ olasılığı = \frac{4}{6} X \frac{1}{40} = \frac{1}{60} \quad (4.9)$$

$X'in\ Kırmızı\ Sınıf\ Olma\ arka\ olasılığı = Kırmızı\ Sınıfın$ Önceki $Olasılığı\ X\ "X"in\ Kırmızı\ Sınıf\ Olasılığı$

$$X'in\ Kırmızı\ Sınıf\ Olma\ arka\ olasılığı = \frac{2}{6} X \frac{3}{20} = \frac{1}{20} \quad (4.10)$$

Sınıf üyeliği x'in kırmızı sınıf olma arka olasılığı daha büyük olduğundan, x kırmızı olarak sınıflandırılır.

Olasılıklı bir model olduğundan, tüm özniteliklerin birbirinden bağımsız olduğunu varsayar; yani belirli bir özniteliğin varlığını veya yokluğunu diğer herhangi bir öznitelik ile tamamen ilgisiz olarak kabul eder.

En yakın k komşu (KNN): KNN uzaklık hesabına dayalı, sınıflandırma yapabileceğimiz makine öğrenme algoritmasıdır.

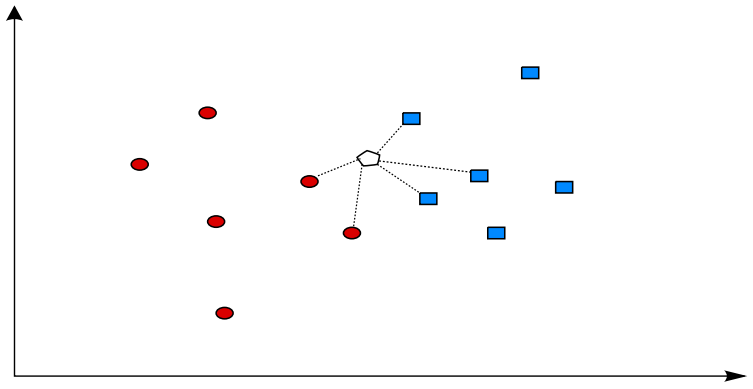
KNN, verideki tüm sınıflar arasındaki mesafeleri bularak, belirlenen k sayısı kadar en yakın olan sınıfları seçip ardından en çok kullanılan sınıflandırma durumuna göre tahmin yapar. Bir sınıf, komşularının çoğunluk oyuyla sınıflandırılır ve yeni gelen bilinmeyen sınıf, en yakın komşuları arasında en yaygın sınıfa atanır. Bunun anlamı, genelleme yapmasıdır.

Sınıfları karşılaştırmak için bir benzerlik ölçütü tanımlanmıştır. Yeni bir veri ortaya çıktığında, benzerlik ölçütüne göre en yakın komşusu seçilmektedir. Sınıflandırılacak yeni veri, en yakın komşusu arasındaki çoğunluğun sınıfına atanır.

KNN'de K en yakın komşu sayısıdır. Komşu sayısı temel karar verme faktörüdür. En yakın benzer noktaları bulmak için öklid mesafesi, Hamming mesafesi, Manhattan mesafesi ve Minkowski mesafesi gibi mesafe ölçütlerini kullanarak noktalar arasındaki mesafe bulunabilir.

Bu algoritmada K seçimi çok önemlidir. K çok küçük seçilirse çok farklı, doğru olmayan sınıflandırma yapılabilir. Öte yandan, K çok büyük seçilirse farklılıklar göz ardı edilir ve sınıfı çok olana doğru eğilim olur. En verimli K değerini seçmek için çapraz doğrulama kullanılmıştır.

Şekil 4.7'de, iki veri sınıfına sahip bir eğitim verisi görülebilir. En benzer 5 örnek, bilinmeyen örnek için seçilir. Örneklerin çoğu mavi sınıftan olduğu için yeni örneğin mavi sınıfta olduğu belirlenir.

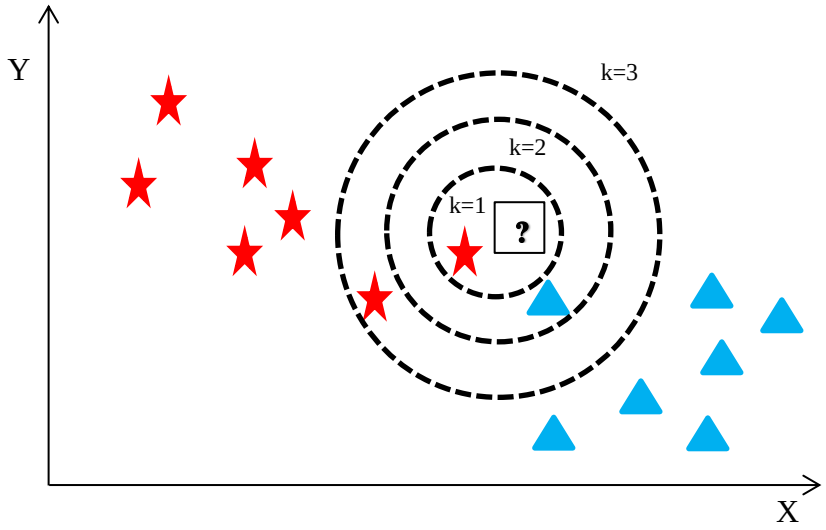


Şekil 4.7. K = 5 ile KNN = bilinmeyen veri mavi sınıfa atanmıştır.

KNN; finans, sağlık, siyaset bilimi, el yazısı tanıma, görüntü tanıma ve video tanıma gibi çeşitli uygulamalarda kullanılmaktadır. Finansal kuruluşlarda müşterilerin kredi derecelendirmeleri KNN ile öngörülmektedir. Ayrıca, bankacılık kuruluşları

kredinin güvenli mi yoksa riskli mi olacağını tahmin etmektedir.

Çalışmamızda, başarımı artırmak için algoritmamız Şekil 4.8'de görüldüğü gibi farklı k=1, 2, 3 değerlerinde kullanılarak model eğitilmiştir.



Şekil 4.8. KNN (k=1, k=2, k=3) algoritması

Bu çalışmamızda denetimli makine öğrenmesi yöntemi kullanılmıştır. Makine öğrenmesi, eğitim veri setinin toplanmasıyla başlamıştır. Daha sonra eğitim verileri üzerinde kullandığımız dört algoritma, eğitim veri seti ile denenmiştir. Denetimli Makine öğrenmesi algoritmaları Naive Bayes, C4.5, SVM , KNN algoritmaları test edilmiştir.

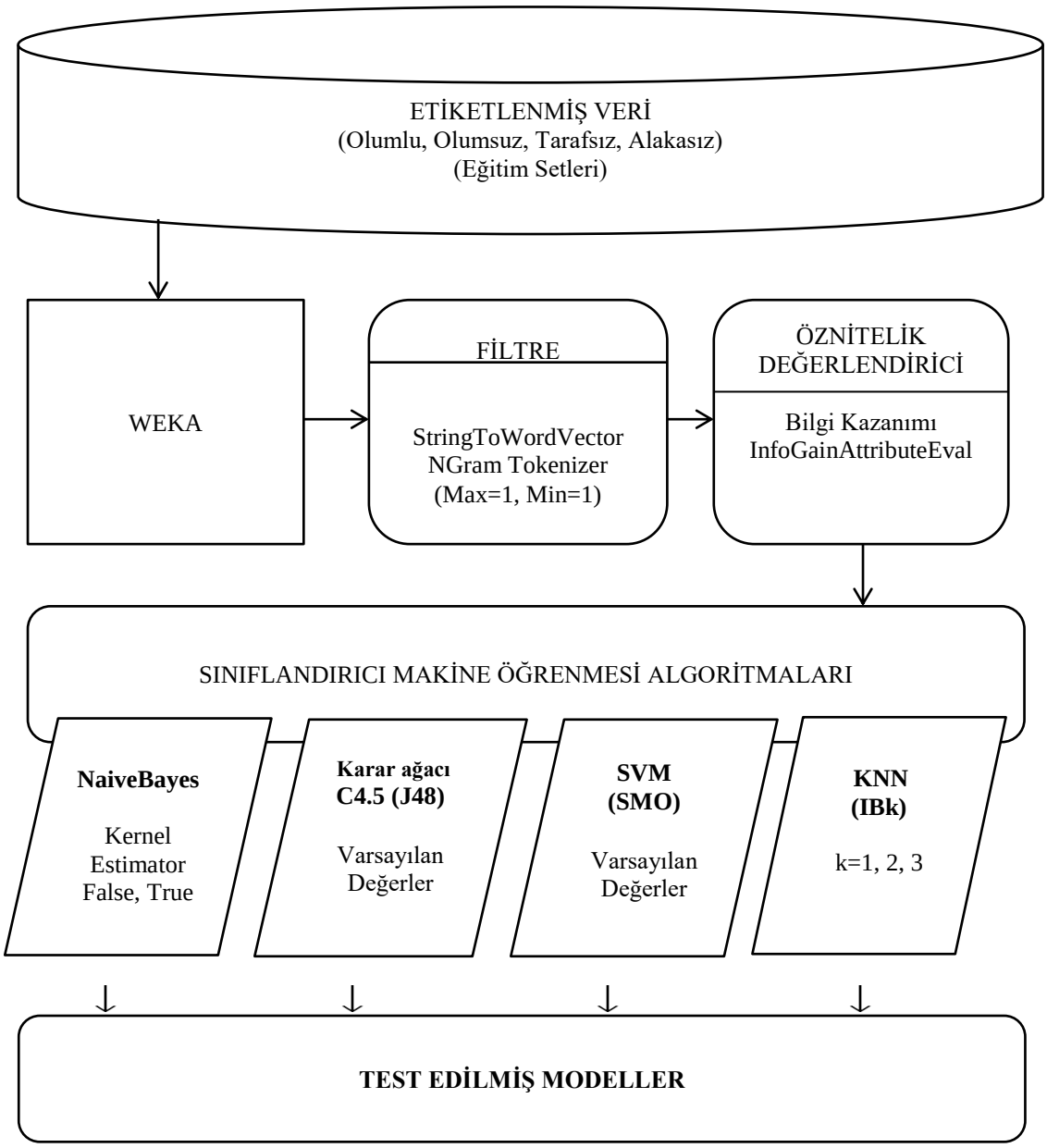
4.2. Weka

Weka (Witten ve Frank, 2005), Waikato Üniversitesi tarafından geliştirilen açık kaynaklı bir makine öğrenme aracıdır. Makine öğrenmesi algoritmalarını barındıran ve kolay kullanımını sağlayan arayüze sahip bir veri madenciliği aracıdır. Bu araç Windows, Linux ve Mac OS dahil olmak üzere birçok platformda çalışır. Veri ön işleme, öznitelik seçimi, kümeleme, sınıflandırma ve veri görselleştirme için çeşitli algoritmaları kullanmayı sağlar. Weka sınıflandırma amacıyla kullanıldığında genellikle; veri setini anlamak ve özniteliklerin bilgi kazançları hakkında bilgi edinmek, veri setlerinden model oluşturmak ve oluşturulmuş modele göre yeni bir verinin sınıfını tahmin etmek için kullanılır. Weka, öznitelik ilişkisi için basit bir metin dosyası formatı olan arff dosya formatını kullanmaktadır. Hem sayısal hem de nominal öznitelikleri destekler.

Yapılan çalışmada Weka; veri ön işleme, sınıflandırma, öznitelik seçimi ve oluşturulan modelleri test etmek için kullanılmıştır. Tüm testlerde, sınıflandırma doğruluğunu değerlendirmek için 10 katlı çapraz doğrulama kullanılmıştır. Öznitelik seçimi ve model değerlendirme için tüm işlemler Weka ile yapılmıştır. Weka makine öğrenmesi algoritmaları için farklı ayarlanabilir parametreler sağlar, böylece parametrelerin değiştirilmesi ile başarımının iyileştirilip iyileştirilmediğinin görülmesini sağlar.

Bu çalışmada geliştirilen model uygulamanın mimarisi Şekil 4.9'da gösterilmiştir. Şekil 4.9 ihtiyaç duyulan çeşitli modelleri oluşturmak ve değerlendirmek için kullanılan adımları kısaca göstermektedir. İlk başta eğitim veri seti bir .arff metin dosyası biçiminde verilmiştir. Bu tür dosyalarda, bir dizi özniteliği paylaşan örneklerin bir listesi açıklanmaktadır (Waikato, 2019).

Daha sonra, seçilen sınıflandırıcılara uygun olduğu için eğitim setini bireysel kelimelere ya da n-gramlara ayıracak olan filtre uygulanmıştır. Bunun için FilteredClassifier->StringToWordVector filtresi seçilmiş ve çıktı olarak elde edilen veriler, belirlenmiş sınıflandırma algoritmasına verilerek modeller elde edilmiştir.



Şekil 4.9. Weka'da yapılan işlem basamakları

4.2.1. StringToWordVector Filtresi

Weka ile sınıflandırma; veri ön işleme, öznitelik belirleme, sınıflandırma, model test etme şeklinde gerçekleştirilir. Ön işleme aşaması, Weka'daki "StringToWordVector" filtre işlevi kullanılarak yapılmıştır. Bu filtre, eğitim setini kelime frekanslarını temsil eden bir vektöre dönüştürür. Bunun sonunda elde edilen dönüştürülmüş veri seti sınıflandırma aşamasında kullanılır. Bu aşamadan sonra sınıflandırma algoritmaları ile modeller oluşturulmuş ve testleri yapılmıştır.

Tweetleri sınıflandırmak için, eğitim setindeki metinler öznitelik olarak kullanılacak kelimelere ayrılmıştır. Weka, bu görevi gerçekleştirmek için kullanılacak StringToWordVector adlı bir filtre sunar. StringToWordVector filtresi ardışık kelime sayısı olarak ifade edebileceğimiz n-gram parametresi ile verilir ve buna bağlı olarak öznitelikler çıkartılır. Basitçe terim varlığının duygu analizi ile ilgili problemlerde kelime sıklığından daha faydalı olduğu gösterilmiş (Pang, Lee ve Vaithyatan, 2002) olduğundan çalışmada bu yaklaşım tercih edilmiştir. Bu çalışmada kullanılan metin veriler tweetlerden oluştuğu için metin uzunlukları kısadır bu sebeple n-gram parametresi max :1, min:1 olarak alınmıştır.

Bu filtre uygulanmadan sınıflandırma başarısı düşmektedir. Bu sebeple duygu analizi için çeşitli parametreler ile denemeler yapılmış ve n-gramlara bölmenin başarımı arttırdığı tespit edilmiştir. Yapılan denemelerde n-gram olarak; max=2, 3, 3 ve min=1, 2, 3 parametreleri verilmiş ve algoritmaların sınıflandırma başarımının düştüğü tespit edilmiştir. Bu sebeple bu sonuçlara çalışmada yer verilmemiştir.

Ayrıca tüm metinler küçük harfe dönüştürülmüştür. Mevcut çalışmaların hemen hemen hepsi bu tekniği kullanmaktadır, ancak bir duygu durumunun metin içerisinde büyük harflerle ifade edilmesi ile küçük harflerle ifade edilmesi duygu ifadesinin farklı anlamlara gelmesine neden olabilir. Ancak bu duygu durumunu ortaya çıkarmak kolay değildir.

4.2.2. Bilgi Kazancına Dayalı Özellik Seçimi

Bilgi kazancı (S. Kullback ve R. A. Leibler, 1951) bir özniteliğin önemini ölçmek için kullanılmakta ve böylelikle öznitelik seçiminde yarar sağlamaktadır. Bilgi kazancı olmayan ve çok düşük olan öznitelikler eğitim veri setinden çıkarıldığında, bu özniteliklerin çıkarılmasından kaynaklanan sınıflandırma başarısı kaybı oluşmamaktadır. En yüksek sınıflandırma başarımını sağlamak için bilgi kazancı sağlayan özniteliklerin seçimi yapılmıştır.

Bu çalışmada hangi özniteliklerin bilgi kazancı olduğunu belirlemek için özniteliklerin bilgi kazancı analizi yapılmıştır. Bilgi kazancına göre öznitelikler sıralanmıştır. Kazanç sağlayan öznitelikler ve hiçbir bilgi kazanımı olmayan özellikler tespit edilmiştir ve bilgi kazancı "0" olan öznitelikler atılmıştır. Bilgi kazancı, olasılıksal bir değişkenin belirsizlik ölçüsü olan entropi ile ifade edilir ve H simgesi ile

gösterilir. Burada H bilgi kazancının dağılımıdır.

$p(x_i)$ herhangi bir olayın gerçekleşme olasılığını göstermek üzere entropi 4.11'de verilen denklem ile hesaplanır:

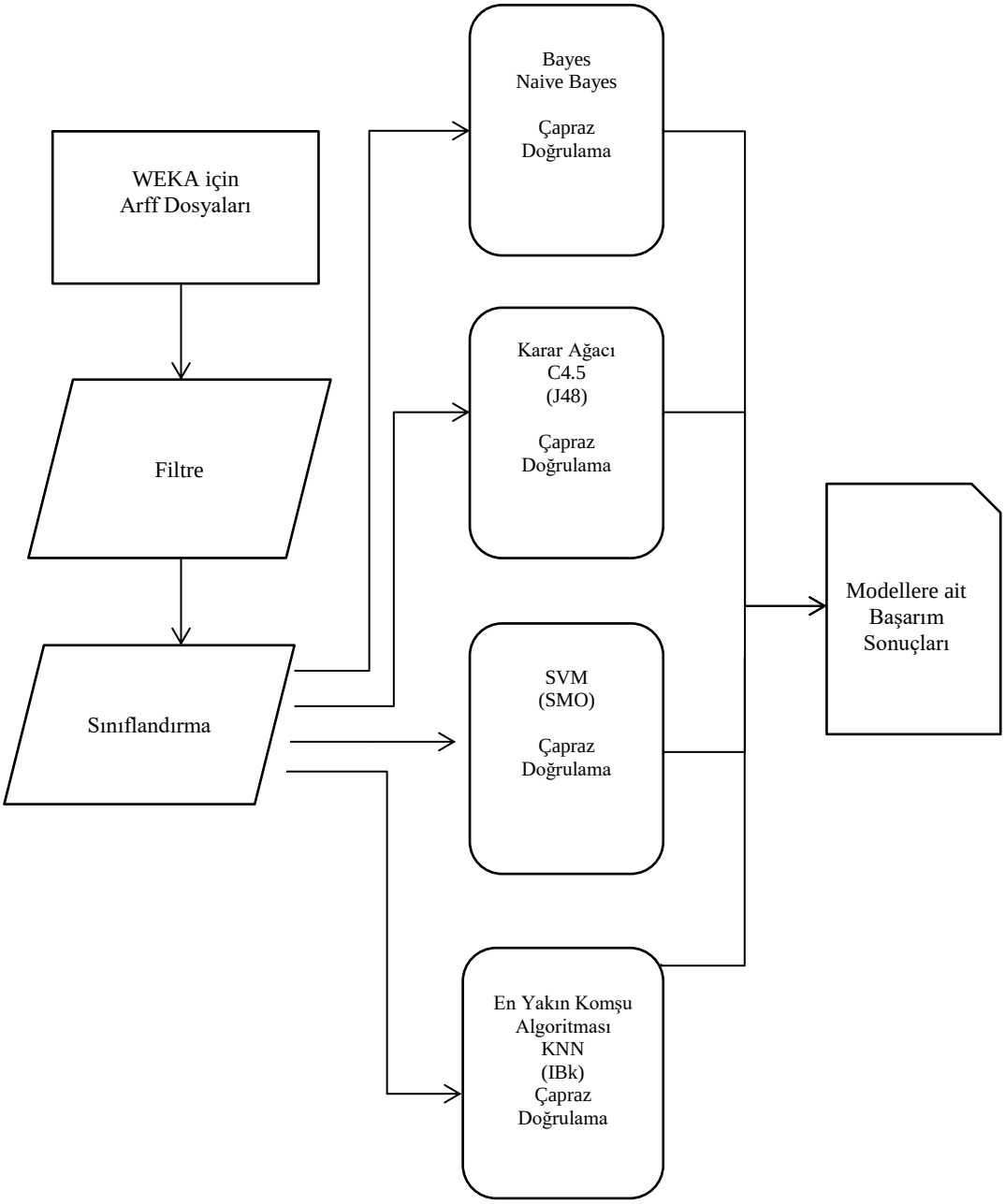
$$H(x) = -\sum_{x=i}^{n} p(x_i) \log_b p(x_i) \quad (4.11)$$

Bu durumda bilgi kazancı aşağıdaki 4.12 denklemi ile hesaplanır:

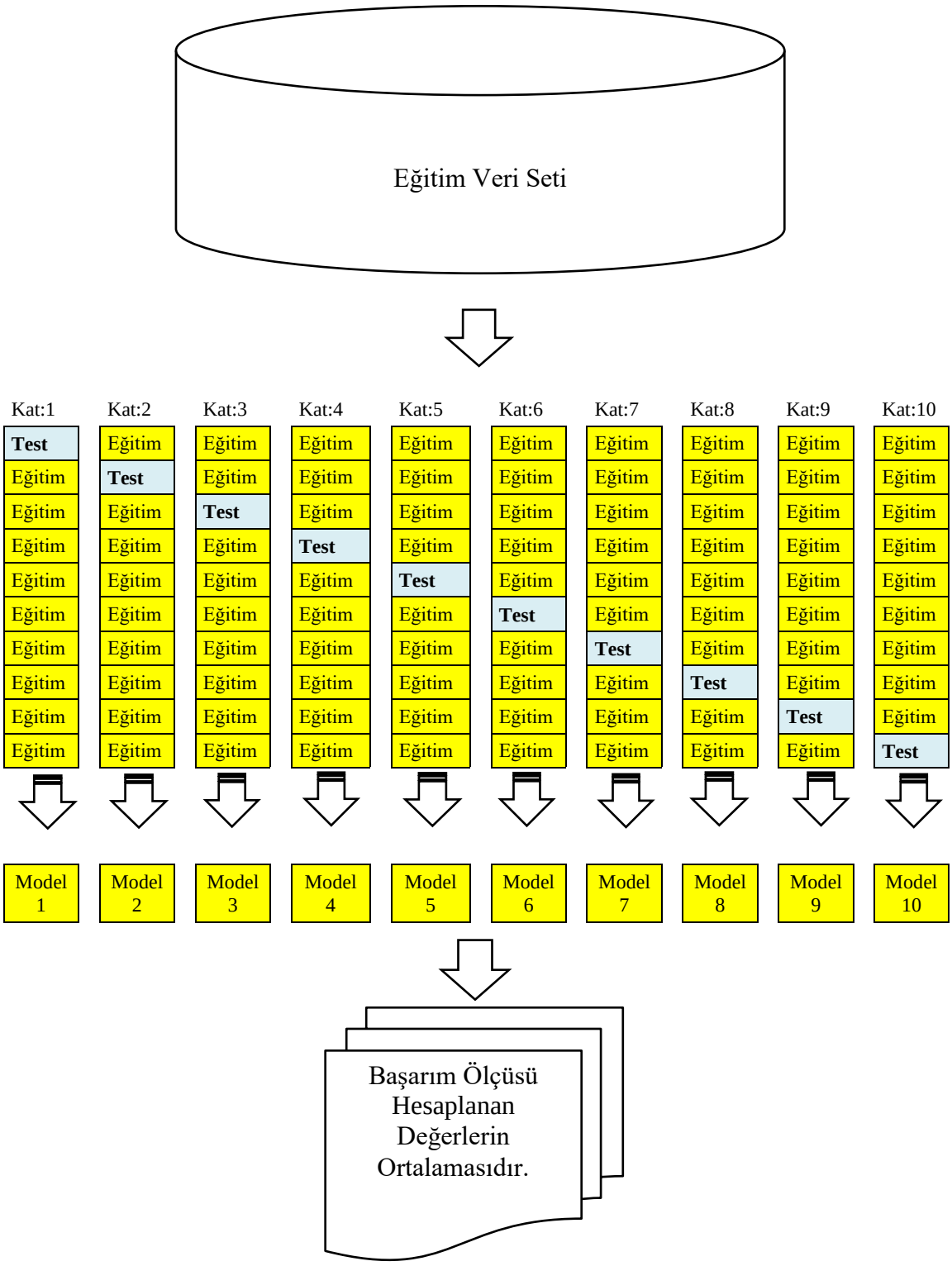
Bilgi Kazancı (Sınıf, Öznitelik) = H (Sınıf) - H (Sınıf | Öznitelik) (4.12)

4.2.3. Çapraz Doğrulama

Çapraz doğrulama, tahmin hatalarını belirlemek için kullanılan bir yöntemdir. Şekil 4.10'da blok şema olarak algoritmalar ve çapraz doğrulama gösterilmektedir. Şekil 4.11'de örnek bir eğitim setinin temsili olarak görselleştirilmesi yapılmıştır. Şekilde çapraz doğrulama için eğitim seti, eşit sayıda örnek içeren bloklara ayrılmaktadır. Blokların bir kısmı modelin eğitimi için ve bir kısmı da testi için kullanılır. Şekildeki her blok bir kez test için kullanılır. İşlem k defa tekrarlanır ve k defalık tahmin hatalarının ortalaması, hata tahmini olarak kullanılır. Örneğin; eğitim setinin örnek sayısı, n = 1000 ve k = 10 ise veriler rastgele olarak her birinde 100 örnekle 10 eşit bloklara bölünür. Daha sonra 9 bloktaki 900 örnek, eğitim için ve geri kalan 100 örnek test için kullanılır. Bu işlem 10 kez tekrarlanır, böylece tüm bloklar bir defa test için kullanılmış olur. Elde edilen modelin başarısı, farklı veriler ile 10 defa oluşturulan ve test edilen modellerin başarı ortalamasıdır (Kamber, Micheline Pei, Jian Han, Jiawei, 2011).



Şekil 4.10. Sınıflandırma algoritmaları ve çapraz doğrulama işlemi



Şekil 4.11. 10 Katlı Çapraz Doğrulama

4.2.4. Duygu Sınıflandırılması Başarımının Değerlendirilmesi

Makine Öğrenmesi algoritmalarının başarımını değerlendirmek için ölçütler mevcuttur. Bu ölçütlerin her biri her sınıf için ayrı ayrı hesaplanır ve ardından algoritmanın başarımı için kullanılır.

Sınıflandırma algoritmasının başarımı, karışıklık matrisi kullanılarak değerlendirilebilir. Karışıklık matrisi gerçek ve öngörülen sınıfların sayılarını gösterir. Ayrıca, testlerimizin sonuçlarını hesaplamak ve karşılaştırmak için doğruluk, duyarlılık, geriçağırım, F-ölçütü ve kappa değeri kullanılmıştır.

Çalışmada kullanılan makine öğrenmesi algoritmalarının karşılaştırılması bu ölçütler üzerinden yapılmıştır. Her bir algoritmanın başarımı, algoritmanın özelliklerine ve veri setinin özniteliklerine göre değişebilir. Bu nedenle, farklı eğitim setleri için en iyi başarımı sağlayan makine öğrenmesi algoritmaları, eğitim setine bağlı olarak farklı olabilmektedir.

4.2.4.1. Doğruluk, Duyarlılık, Geriçağırım ve F-ölçütü

Makine öğrenmesi algoritmalarının başarımını ölçmek için; doğruluk, duyarlılık, geriçağırım, F-ölçütü ve kappa değerleri gibi ölçütler kullanılır.

Bu ölçütlerin belirlenmesinde ise aşağıda belirtilen değerler kullanılmaktadır. Bu değerlerin açıklamaları aşağıda kısaca verilmiştir.

- Doğru Pozitif (TP), gerçekte pozitif olan ve ayrıca sınıflandırıcı tarafından pozitif olarak sınıflandırılan doğru tahmin edilen durumları ifade eder.
- Yanlış Pozitif (FP), sınıflandırıcı tarafından pozitif olarak tahmin edilen, ancak gerçekte negatif olan yanlış tahmin edilen durumları ifade eder.
- Doğru Negatif (TN), gerçekte negatif olan ve ayrıca sınıflandırıcı tarafından negatif olarak sınıflandırılan doğru tahmin edilen durumları temsil eder.
- Yanlış Negatif (FN), sınıflandırıcı tarafından negatif olarak tahmin edilen ancak gerçekte pozitif olan yanlış tahmin edilen durumları ifade eder.

Doğruluk, tahmin sonucunun gerçek değerine yakınlık derecesine denir. Yapılan doğru tahminlerin, yapılan toplam tahmin sayısına oranıdır. Doğruluk, sınıflandırma başarımı için ortak bir ölçüttür. Doğruluk, Denklem 4.13'te görüldüğü gibi

hesaplanabilir:

$$Doğruluk = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}} = \frac{doğru\ sınıflandırılmış\ tweetler}{tüm\ tweet\ sınıflarının\ toplamı} \tag{4.13}$$

Duyarlılık, doğru sınıflandırılmış pozitif tweetlerin, pozitif olarak sınıflandırılmış tüm tweetlere oranıdır. Sınıflandırıcının, veri setinden bağımsız olarak belirli bir sınıfta ne kadar etkin başarım gösterdiğini gösterir. Bu nedenle, bir sınıfın diğerlerinden daha fazla olması yarar sağlar. Ancak, sınıflandırıcı tarafından tahmin edilen yanlış negatifleri dikkate almaz. Bu olumsuzluğu gidermek için genellikle geriçağırım ile birlikte kullanılır. Düşük duyarlılık çok sayıda yanlış pozitif olduğunu da gösterir. Duyarlılık, Denklem 4.14'te görüldüğü gibi hesaplanır:

$$Duyarlılık = \frac{\text{TP}}{\text{TP+FP}} = \frac{doğru\ sınıflandırılmış\ pozitif\ tweetler}{pozitif\ olarak\ sınıflandırılmış\ tüm\ tweetler} \tag{4.14}$$

Geriçağırım, doğru sınıflandırılmış pozitif tahminlerin ne kadar başarılı sınıflandırıldığını gösterir. Geriçağırım, doğru olarak pozitif etiketlenen sınıf sayıları arasındaki oranın, gerçekten doğru olan toplam sınıf sayısına bölünmesiyle tanımlanır. Geriçağırım, Denklem 4.15'te görüldüğü gibi hesaplanır:

$$Geriçağırım = \frac{\text{TP}}{\text{TP+FN}} = \frac{doğru\ sınıflandırılmış\ pozitif\ tweetler}{doğru\ olarak\ sınıflandırılmış\ tüm\ \ tweetler} \tag{4.15}$$

F-ölçütü, sınıflandırıcının doğruluğunu ölçmek için kullanılır. F-ölçütü, hem duyarlılık hem de geriçağırım kullanarak bir testin başarımının ölçümünü sağlayabilir.

F-ölçütü, duyarlılık ile geriçağırım arasındaki dengedir ve her ikisinin de ağırlıklı harmonik ortalamasını hesaplayarak duyarlılık ve geriçağırımı birleştirir. F-ölçütü, 0 ile 1 arasında değerler alır; burada 0 en kötü ve 1 en iyi ölçüt değerleridir. Hatasız duyarlılık ve geriçağırma değerleri olursa F-ölçütü değeri 1 olur. En düşük duyarlılık ve en düşük geriçağırım değerlerinde F-ölçütü en düşük değer alır. F-ölçütü, Denklem 4.16'da görüldüğü gibi hesaplanır:

$$F - \text{Ölçütü} = \frac{2*\text{Geriçağırım}*\text{Duyarlılık}}{(Geriçağırım+Duyarlılık)} \quad (4.16)$$

4.2.4.2. Kappa İstatistiği

Kappa istatistiği (κ) hem çok sınıflı hem de dengesiz sınıflı sınıflandırma problemlerinde her bir sınıfın frekansına göre rastgele tahmin eden sınıflandırıcının, başarımını gösterir. Kappa istatistiği, sınıflandırma doğruluğunu verilerdeki sınıfların dengesine göre normalleştirerek bu sorunu düzeltir. Her zaman doğru olan bir sınıflandırıcı, κ değerine sahip olacaktır. Tersine, rastgele bir sınıflandırıcı ile aynı olasılıkla doğru tahminleri yaparsa κ değeri sıfır olacaktır. Kappa istatistiğinin denklemine gelmeden önce, bir değer daha gereklidir. Bu değer, beklenen doğruluk değeridir. Bu değer, karışıklık matrisine dayanarak herhangi bir rasgele sınıflandırıcının elde etmesi beklenen doğruluk olarak tanımlanır. Gözlemlenen doğruluk, sadece tüm karışıklık matrisi boyunca doğru şekilde sınıflandırılmış verilerin sayısıdır. Denklem 4.17'de verilen kappa, gözlenen doğruluk ve beklenen doğruluk kullanılarak hesaplanabilir (Cohen, 1960). Kappa değeri, Denklem 4.17'de görüldüğü gibi hesaplanır:

$$\text{Kappa} = \frac{(\text{gözlenen doğruluk} - \text{beklenen doğruluk})}{(1 - \text{beklenen doğruluk})} \quad (4.17)$$

Kappa istatistiği, gözlenen doğruluk ile beklenen doğruluğu karşılaştıran bir ölçüttür. Kappa istatistiği, yalnızca tek bir sınıflandırıcıyı değerlendirmek için değil; aynı zamanda kendi aralarında sınıflandırıcıları değerlendirmek için de kullanılır.

Temel olarak kappa istatistiği, makine öğrenmesi algoritmaları tarafından sınıflandırılan metinlerin, tahmini olarak etiketlenen verileri ne kadar yakından eşleştirdiğini ve beklenen doğruluk ile ölçülen rastgele bir sınıflandırıcının doğruluğunu kontrol eden bir ölçüttür.

4.2.4.3. Karışıklık Matrisi

Karışıklık matrisi, bir sınıflandırıcı algoritmanın başarımını gösterir ve bir sınıflandırma modeli tarafından yapılan gerçek ve öngörülen sınıflandırmalar hakkında

bilgi içerir (Bernhard E. Boser, Isabelle M. Guyon, and Vladimir N. Vapnik. 1992). Makine öğrenmesi algoritmalarını değerlendirmek için sonuçlar matriste gösterilmiştir. Her bir değer, Şekil 4.12'de gösterildiği gibi karışıklık matrisi, bir sınıflandırıcının tam başarımını gösterir. Matrisler, sınıflandırıcıların her bir veriyi, verilen sınıfına karşılık tahmin ettiği sınıfı gösterir. Verilen sınıf, y ekseninde ve tahmin edilen sınıf x ekseninde gösterilir.

		Makine Öğrenmesi Tahmin Edilen Sınıf	
		Doğru Tahmin (P)	Yanlış Tahmin (N)
Gerçek Sınıf	Doğru Etiketli (P)	Doğru Pozitif (TP)	Yanlış Negatif (FN)
	Yanlış Etiketli (F)	Yanlış Pozitif (FP)	Doğru Negatif (TN)

Şekil 4.12. İki yönlü sınıflandırma için bir karışıklık matrisi

BÖLÜM 5

BAŞARIM TESTLERİ

Bu çalışmamızda farklı özniteliklere sahip 12 farklı eğitim veri seti kullanılarak sınıflandırıcı modeller oluşturulmuştur. Öznitelik çıkarımları yapıldıktan sonra sınıflandırma için Naive Bayes, C4.5, SVM, KNN algoritmalarının farklı parametreler ile denemeleri yapılmıştır. Bu denemelerden elde edilen sonuçlar çeşitli ölçütler göz önünde bulundurularak aşağıda verilmiştir.

5.1. Model Eğitimi

Öznitelikler seçildikten ve eğitim verileri öznitelikler bakımından temsil edildikten sonra, eğitim başlar. Eğitim veri setleri kullanılarak modeller eğitilmiştir. Eğitim veri setlerinde sınıflar; olumlu, olumsuz, tarafsız ve alakasız olarak etiketlenmiştir. Her bir sınıftaki olumlu, olumsuz, tarafsız ve alakasız sınıfların olasılıkları hesaplanmaktadır. Bu olasılıklar, veri setinde kaçının olumlu, olumsuz, tarafsız ve alakasız sınıf olduğunun tahminidir.

5.2. Makine Öğrenmesi Algoritmalarının Başarımı

Makine öğrenmesi algoritmaları içinden Naive Bayes, SVM, C4.5 ve KNN olmak üzere dört farklı sınıflandırma algoritması kullanılmıştır. Weka'da yer alan kullanımları ile Naive Bayes algoritması için "NaiveBayes", SVM algoritması için "SMO", C4.5 algoritması için "J48", KNN en yakın k komşu algoritması için "IBk" kullanılmıştır. Çapraz Doğrulama işlemi için 10 kat çapraz doğrulama tercih edilmiştir. Sınıflandırıcının başarım değerlendirmesi; doğruluk, duyarlılık, geriçağırım, F-ölçütü gibi ölçütler kullanılarak yapılmış, belirtilen parametreler ile dört sınıflandırma algoritmasının başarım sonuçları aşağıdaki gibi bulunmuştur:

- Naive Bayes (kernelEstimator = false) sınıflandırma algoritması

Naive Bayes sınıflandırma algoritmasında EğitimSeti-1 veri seti en yüksek duyarlılık ve geriçağırım değerlerine sahip olmuştur. En düşük doğruluk %75,21 olmuştur. En yüksek kappa, F-ölçütü ve doğruluk yüzdesine sahip olan EğitimSeti-4 veri seti %76,82 doğruluk değerine sahip olmuş ve en iyi performansı göstermiştir. Sonuçlar Çizelge 5.1'de gösterilmiştir.

Çizelge 5.1. Naive Bayes (kernelEstimator = false) sınıflandırma algoritması sonuçları

Dosya Adı	**Sınıflandırma Algoritması**	**Duyarlılık (Precision)**	**Geriçağırım (Recall)**	**F-ölçütü (F-measure)**	**Kappa**	**Doğruluk (Accuracy) %**
EğitimSeti-1	Naive Bayes (kernelEstimator = false)	0,08	0,78	0,76	0.68	75,75
EğitimSeti-2	Naive Bayes (kernelEstimator = false)	0,79	0,76	0,76	0.68	75,81
EğitimSeti-3	Naive Bayes (kernelEstimator = false)	0,78	0,75	0,76	0.67	75,31
EğitimSeti-4	Naive Bayes (kernelEstimator = false)	0,79	0,77	0,77	0.69	**76,82**
EğitimSeti-5	Naive Bayes (kernelEstimator = false)	0,78	0,75	0,76	0.67	75,38
EğitimSeti-6	Naive Bayes (kernelEstimator = false)	0,78	0,76	0,76	0.68	75,98
EğitimSeti-7	Naive Bayes (kernelEstimator = false)	0,78	0,75	0,76	0.67	75,41
EğitimSeti-8	Naive Bayes (kernelEstimator = false)	0,78	0,76	0,76	0.67	75,50
EğitimSeti-9	Naive Bayes (kernelEstimator = false)	0,78	0,75	0,76	0.67	75,21
EğitimSeti-10	Naive Bayes (kernelEstimator = false)	0,79	0,76	0,77	0.68	76,32
EğitimSeti-11	Naive Bayes (kernelEstimator = false)	0,78	0,74	0,75	0.65	74,09
EğitimSeti-12	Naive Bayes (kernelEstimator = false)	0,79	0,76	0,77	0.68	76,03

- Naive Bayes (kernelEstimator = true) sınıflandırma algoritması

Veri kümesindeki özniteliklerin gerçek dağılımına daha iyi eşleşebilecek KernelEstimator argümanına sahip bir çekirdek tahmincisi kullanılmıştır ve Naive Bayes sınıflandırma algoritmasında öncekine göre daha yüksek doğruluk değeri elde edilmiştir. EğitimSeti-11 veri seti en yüksek duyarlılık ve geriçağrışım değerlerine sahip olmuştur. En düşük doğruluk %82,92 olmuştur. En yüksek kappa, F-ölçütü ve doğruluk yüzdesine sahip olan EğitimSeti-10 veri seti %84,3 doğruluk değerine sahip olmuştur ve en iyi performansı göstermiştir. Sonuçlar Çizelge 5.2'de gösterilmiştir.

Çizelge 5.2. Naive Bayes (kernelEstimator = true) sınıflandırma algoritması sonuçları

Dosya Adı	**Sınıflandırma Algoritması**	**Duyarlılık (Precision)**	**Geriçağırım (Recall)**	**F-ölçütü (F-measure)**	**Kappa**	**Doğruluk (Accuracy) %**
EğitimSeti-1	Naive Bayes (kernelEstimator = true)	0,86	0,84	0,84	0.79	83,91
EğitimSeti-2	Naive Bayes (kernelEstimator = true)	0,86	0,84	0,84	0.78	83,87
EğitimSeti-3	Naive Bayes (kernelEstimator = true)	0,86	0,83	0,84	0.78	83,31
EğitimSeti-4	Naive Bayes (kernelEstimator = true)	0,86	0,84	0,84	0.78	83,60
EğitimSeti-5	Naive Bayes (kernelEstimator = true)	0,86	0,83	0,84	0.78	82,90
EğitimSeti-6	Naive Bayes (kernelEstimator = true)	0,86	0,83	0,84	0.78	83,28
EğitimSeti-7	Naive Bayes (kernelEstimator = true)	0,85	0,83	0,84	0.77	83,02
EğitimSeti-8	Naive Bayes (kernelEstimator = true)	0,86	0,83	0,84	0.77	82,95
EğitimSeti-9	Naive Bayes (kernelEstimator = true)	0,86	0,83	0,83	0.77	82,92
EğitimSeti-10	Naive Bayes (kernelEstimator = true)	0,86	0,84	0,85	0.80	**84,30**
EğitimSeti-11	Naive Bayes (kernelEstimator = true)	0,87	0,84	0,85	0.79	83,99
EğitimSeti-12	Naive Bayes (kernelEstimator = true)	0,86	0,84	0,84	0.79	83,87

- SVM sınıflandırma algoritması

SVM sınıflandırma algoritması ile diğer kullandığımız sınıflandırma algoritmalarından daha yüksek doğruluk değeri elde edilmiştir.

EğitimSeti-2 veri seti en yüksek duyarlılık ve geriçağırım değerlerine sahip olmuştur. En düşük doğruluk %90,37 olmuştur. En yüksek kappa, F-ölçütü ve doğruluk yüzdesine sahip olan EğitimSeti-2 veri seti %92,3 doğruluk değerine sahip olmuştur ve en iyi performansı göstermiştir. Sonuçlar Çizelge 5.3'te gösterilmiştir.

Çizelge 5.3. SVM sınıflandırma algoritması sonuçları

Dosya Adı	**Sınıflandırma Algoritması**	**Duyarlılık (Precision)**	**Geriçağırım (Recall)**	**F-ölçütü (F-measure)**	**Kappa**	**Doğruluk (Accuracy) %**
EğitimSeti-1	SVM	0,92	0,92	0,92	0.90	92,23
EğitimSeti-2	SVM	0,92	0,92	0,92	0.90	**92,30**
EğitimSeti-3	SVM	0,92	0,92	0,92	0.89	91,87
EğitimSeti-4	SVM	0,92	0,92	0,92	0.89	92,06
EğitimSeti-5	SVM	0,92	0,92	0,92	0.89	91,58
EğitimSeti-6	SVM	0,91	0,91	0,91	0.88	90,76
EğitimSeti-7	SVM	0,91	0,90	0,90	0.87	90,37
EğitimSeti-8	SVM	0,92	0,92	0,92	0.89	91,57
EğitimSeti-9	SVM	0,92	0,92	0,92	0.89	91,64
EğitimSeti-10	SVM	0,92	0,92	0,92	0.90	92,17
EğitimSeti-11	SVM	0,92	0,92	0,92	0.89	91,63
EğitimSeti-12	SVM	0,92	0,92	0,92	0.90	92,19

- C4.5 sınıflandırma algoritması Sonuçları

C4.5 karar ağacı için Weka programında J48 olarak açık kaynak kodlu sınıflandırma algoritması kullanılmıştır.

EğitimSeti-1 veri seti en yüksek duyarlılık, geriçağırım, kappa değerine ve en yüksek doğruluk yüzdesine sahip olmuştur. En düşük doğruluk %87,56 olmuştur. EğitimSeti-1 veri seti %89,60 doğruluk değeriyle en iyi performansı göstermiştir. Sonuçlar Çizelge 5.4'te gösterilmiştir.

Çizelge 5.4. Ağaç C4.5 sınıflandırma algoritması sonuçları

Dosya Adı	**Sınıflandırma Algoritması**	**Duyarlılık (Precision)**	**Geriçağırım (Recall)**	**F-ölçütü (F-measure)**	**Kappa**	**Doğruluk (Accuracy) %**
EğitimSeti-1	C4.5	0,90	0,90	0,90	0.86	**89,60**
EğitimSeti-2	C4.5	0,89	0,89	0,89	0.86	89,33
EğitimSeti-3	C4.5	0,89	0,89	0,89	0.86	89,28
EğitimSeti-4	C4.5	0,89	0,89	0,89	0.85	88,64
EğitimSeti-5	C4.5	0,89	0,89	0,89	0.85	88,98
EğitimSeti-6	C4.5	0,88	0,88	0,88	0.84	88,08
EğitimSeti-7	C4.5	0,88	0,88	0,88	0.83	87,56
EğitimSeti-8	C4.5	0,89	0,89	0,89	0.85	88,76
EğitimSeti-9	C4.5	0,89	0,89	0,89	0.85	88,62
EğitimSeti-10	C4.5	0,89	0,89	0,89	0.86	89,34
EğitimSeti-11	C4.5	0,88	0,88	0,88	0.85	88,40
EğitimSeti-12	C4.5	0,90	0,90	0,90	0.86	89,48

- IBk (k=1) sınıflandırma algoritması

KNN en yakın k komşu algoritması için Weka programında IBk olarak isimlendirilen açık kaynak kodlu sınıflandırma algoritması K=1 olarak kullanılmıştır.

EğitimSeti-11 veri seti; en yüksek duyarlılık, geriçağırım, kappa değerlerine ve en yüksek doğruluk yüzdesine sahip olmuş ve en iyi performansı göstermiştir. En düşük doğruluk %82,14 olmuştur. En yüksek F-ölçütü değerine EğitimSeti-10 veri seti olmuştur. EğitimSeti-11 veri seti %87,73 doğruluk değeriyle en iyi performansı göstermiştir. Sonuçlar Çizelge 5.5'te gösterilmiştir.

Çizelge 5.5. IBk (k=1) sınıflandırma algoritması sonuçları

Dosya Adı	**Sınıflandırma Algoritması**	**Duyarlılık (Precision)**	**Geriçağırım (Recall)**	**F-ölçütü (F-measure)**	**Kappa**	**Doğruluk (Accuracy) %**
EğitimSeti-1	IBk (k=1)	0,86	0,86	0,86	0.81	85,68
EğitimSeti-2	IBk (k=1)	0,86	0,86	0,85	0.81	85,59
EğitimSeti-3	IBk (k=1)	0,86	0,85	0,85	0.81	85,40
EğitimSeti-4	IBk (k=1)	0,85	0,85	0,85	0.81	85,39
EğitimSeti-5	IBk (k=1)	0,85	0,85	0,85	0.80	85,14
EğitimSeti-6	IBk (k=1)	0,83	0,82	0,82	0.77	82,42
EğitimSeti-7	IBk (k=1)	0,83	0,82	0,82	0.76	82,14
EğitimSeti-8	IBk (k=1)	0,85	0,85	0,85	0.80	85,11
EğitimSeti-9	IBk (k=1)	0,85	0,85	0,85	0.80	84,89
EğitimSeti-10	IBk (k=1)	0,87	0,86	0,86	0.82	86,22
EğitimSeti-11	IBk (k=1)	0,88	0,88	0,85	0.84	**87,73**
EğitimSeti-12	IBk (k=1)	0,86	0,85	0,85	0.81	85,45

- IBk (k=2) sınıflandırma algoritması

Weka programında, KNN algoritması için açık kaynak kodlu sınıflandırma algoritması olan IBk k=2 olarak uygulanmıştır.

EğitimSeti-11 veri seti; en yüksek duyarlılık, geriçağırım, F-ölçütü, kappa değerlerine ve en yüksek doğruluk yüzdesine sahip olmuştur. En düşük doğruluk %80,84 olmuştur. EğitimSeti-11 veri seti %85,57 doğruluk değeriyle en iyi performansı göstermiştir. Sonuçlar Çizelge 5.6'da gösterilmiştir.

Çizelge 5.6. IBk (k=2) sınıflandırma algoritması Sonuçları

Dosya Adı	**Sınıflandırma Algoritması**	**Duyarlılık (Precision)**	**Geriçağırım (Recall)**	**F-ölçütü (F-measure)**	**Kappa**	**Doğruluk (Accuracy) %**
EğitimSeti-1	IBk (k=2)	0,86	0,84	0,84	0.79	84,43
EğitimSeti-2	IBk (k=2)	0,86	0,84	0,84	0.79	84,42
EğitimSeti-3	IBk (k=2)	0,85	0,84	0,84	0.79	83,91
EğitimSeti-4	IBk (k=2)	0,85	0,84	0,84	0.79	84,07
EğitimSeti-5	IBk (k=2)	0,85	0,84	0,84	0.79	84,10
EğitimSeti-6	IBk (k=2)	0,83	0,81	0,81	0.75	81,12
EğitimSeti-7	IBk (k=2)	0,83	0,81	0,81	0.74	80,84
EğitimSeti-8	IBk (k=2)	0,85	0,84	0,84	0.79	84,05
EğitimSeti-9	IBk (k=2)	0,85	0,84	0,84	0.79	83,98
EğitimSeti-10	IBk (k=2)	0,86	0,85	0,85	0.80	85,00
EğitimSeti-11	IBk (k=2)	0,87	0,86	0,85	0.81	**85,57**
EğitimSeti-12	IBk (k=2)	0,86	0,84	0,84	0.79	84,27

- IBk (k=3) sınıflandırma algoritması

KNN algoritması için k=3 olarak alındığında EğitimSeti-10 veri seti; en yüksek duyarlılık, geriçağırım, kappa değerlerine ve en yüksek doğruluk yüzdesine sahip olmuştur. EğitimSeti-4 veri seti en yüksek F-ölçütü değerine sahip olmuştur. En düşük doğruluk %80,38 olmuştur. EğitimSeti-10 veri seti %85,04 doğruluk ile en iyi performansı göstermiştir. Sonuçlar Çizelge 5.7'de gösterilmiştir.

Çizelge 5.7. IBk (k=3) sınıflandırma algoritması sonuçları

Dosya Adı	Sınıflandırma Algoritması	Duyarlılık (Precision)	Geriçağırım (Recall)	F-ölçütü (F-measure)	Kappa	Doğruluk (Accuracy) %
EğitimSeti-1	IBk (k=3)	0,86	0,85	0,84	0.80	84,66
EğitimSeti-2	IBk (k=3)	0,86	0,85	0,84	0.80	84,71
EğitimSeti-3	IBk (k=3)	0,85	0,85	0,85	0.80	84,74
EğitimSeti-4	IBk (k=3)	0,86	0,85	0,85	0.80	85,03
EğitimSeti-5	IBk (k=3)	0,85	0,85	0,85	0.80	84,80
EğitimSeti-6	IBk (k=3)	0,83	0,81	0,81	0.75	81,01
EğitimSeti-7	IBk (k=3)	0,82	0,80	0,80	0.74	80,38
EğitimSeti-8	IBk (k=3)	0,85	0,85	0,85	0.80	84,69
EğitimSeti-9	IBk (k=3)	0,85	0,85	0,85	0.80	84,73
EğitimSeti-10	IBk (k=3)	0,86	0,85	0,85	0.80	**85,04**
EğitimSeti-11	IBk (k=3)	0,86	0,85	0,84	0.79	84,57
EğitimSeti-12	IBk (k=3)	0,86	0,85	0,84	0.79	84,49

Oluşturduğumuz 12 farklı arff dosyasını kullandığımız tüm sınıflandırma algoritmalarının doğruluk yüzdeleri Çizelge 5.8'de görüldüğü gibi hemen hemen tüm sınıflandırıcılar için ortalama %75 ve üzeri doğruluğa sahiptir. Başka bir ifadeyle, kullandığımız sınıflandırma algoritmaları her bir sınıf için olumlu, olumsuz, tarafsız ve alakasız tweetleri en az %75 doğruluk ile tahmin etmiştir. Dolayısıyla, bu da derleme sınıflandırıcılarımızın hepsinin tamamen eğitilmiş ve twitter verilerinin duyarlılık analizi için hazır olduğunu göstermiştir. Çalışmamızda eğitim setlerini kullanarak tüm algoritmalarının doğruluk başarımları Çizelge 5.8'de gösterilmiştir. Kullandığımız sınıflandırma algoritmaları içinde en iyi başarım gösteren veri setleri aşağıda Çizelge 5.9'da yer almaktadır.

Çizelge 5.8. Sınıflandırma algoritmaları doğruluk başarımları

Dosya Adı	Naive Bayes (kernel Estimator = false) %	Naive Bayes (kernel Estimator = true) %	SVM %	C4.5 %	KNN (k=1) %	KNN (k=2) %	KNN (k=3) %
EğitimSeti-1	75,75	83,91	92,23	**89,60**	85,68	84,43	84,66
EğitimSeti-2	75,81	83,87	**92,30**	89,33	85,59	84,42	84,71
EğitimSeti-3	75,31	83,31	91,87	89,28	85,40	83,91	84,74
EğitimSeti-4	**76,82**	83,60	92,06	88,64	85,39	84,07	85,03
EğitimSeti-5	75,38	82,90	91,58	88,98	85,14	84,10	84,80
EğitimSeti-6	75,98	83,28	90,76	88,08	82,42	81,12	81,01
EğitimSeti-7	75,41	83,02	90,37	87,56	82,14	80,84	80,38
EğitimSeti-8	75,50	82,95	91,57	88,76	85,11	84,05	84,69
EğitimSeti-9	75,21	82,92	91,64	88,62	84,89	83,98	84,73
EğitimSeti-10	76,32	**84,3**	92,17	89,34	86,22	85	**85,04**
EğitimSeti-11	74,09	83,99	91,63	88,40	**87,73**	**85,57**	84,57
EğitimSeti-12	76,03	83,87	92,19	89,48	85,45	84,27	84,49

Yapılan çalışmadan elde edilen sonuçlar değerlendirildiğinde Zemberek kütüphanesini kullanarak hem Türkçe imla denetimi hem de Türkçe anlam denetimi yapmak tüm algoritmaların başarımını artırmıştır. SVM makine öğrenmesi algoritması ile en yüksek doğruluk başarımı elde edilmiştir. Ayrıca KNN algoritmasına öznitelik olarak sadece tweet verildiğinde sınıflama performansının diğer öznitelikli olan eğitim setlerine göre sınıflama başarımının arttığı görülmüştür. C4.5 algoritmasında tweet içindeki sayısal değerleri temizlemenin algoritmanın başarımına katkı sağladığı görülmektedir. Tweet içinden gereksiz kelimeleri temizlemenin Naive Bayes algoritmasının performansını artırdığı tespit edilmiştir.

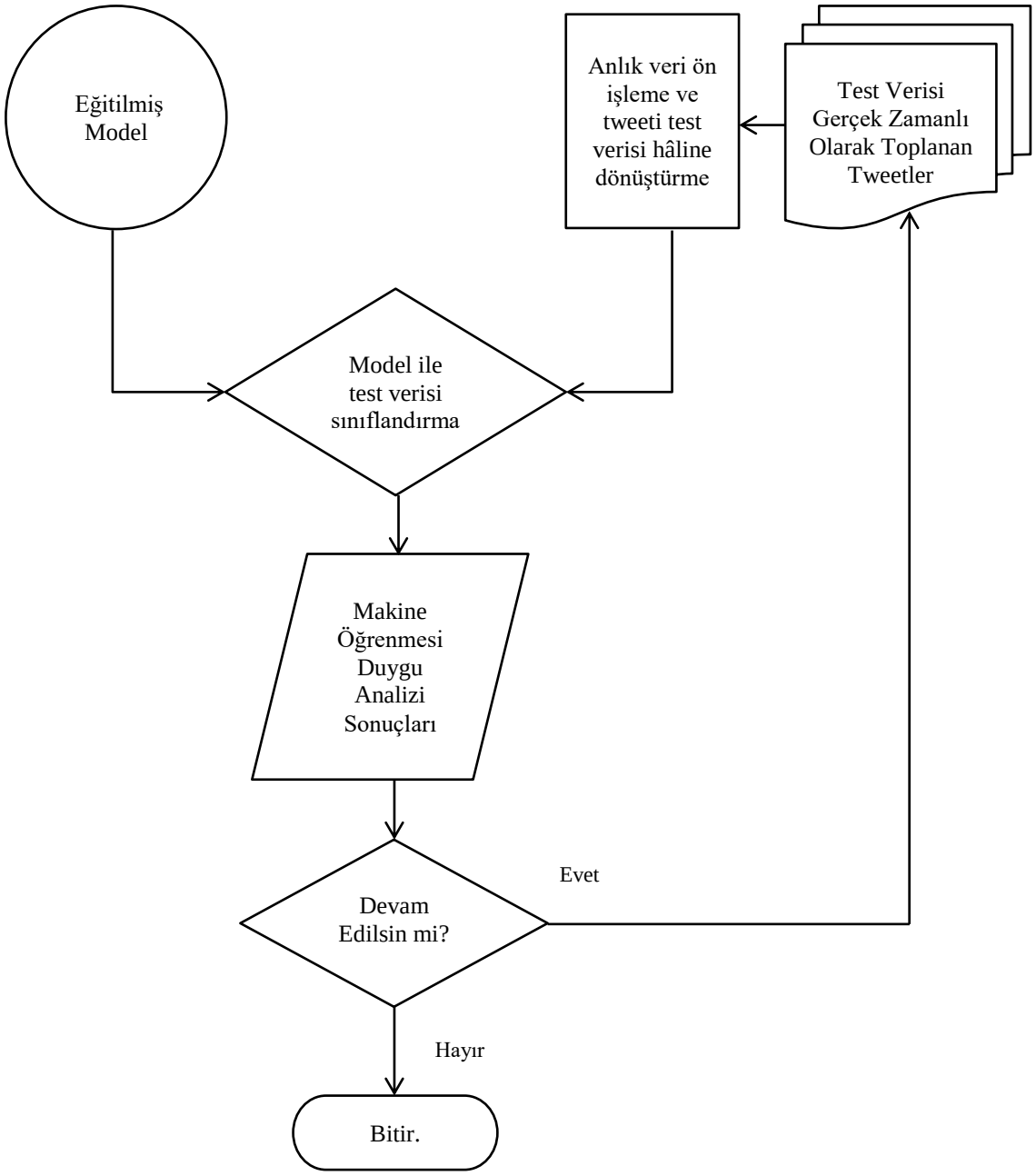
Çizelge 5.9. Sınıflandırma algoritmalarına göre en iyi performans gösteren veri setleri

Dosya Adı	Sınıflandırma Algoritması	Doğruluk (Accuracy) %
EğitimSet-2	SVM	92,30
EğitimSet-1	C4.5	89,60
EğitimSet-11	IBk (K=1)	87,73
EğitimSet-11	IBk (K=2)	85,57
EğitimSet-10	IBk (K=3)	85,04
EğitimSet-10	Naive Bayes (kernelEstimator = true)	84,30
EğitimSet-4	Naive Bayes (kernelEstimator = false)	76,82

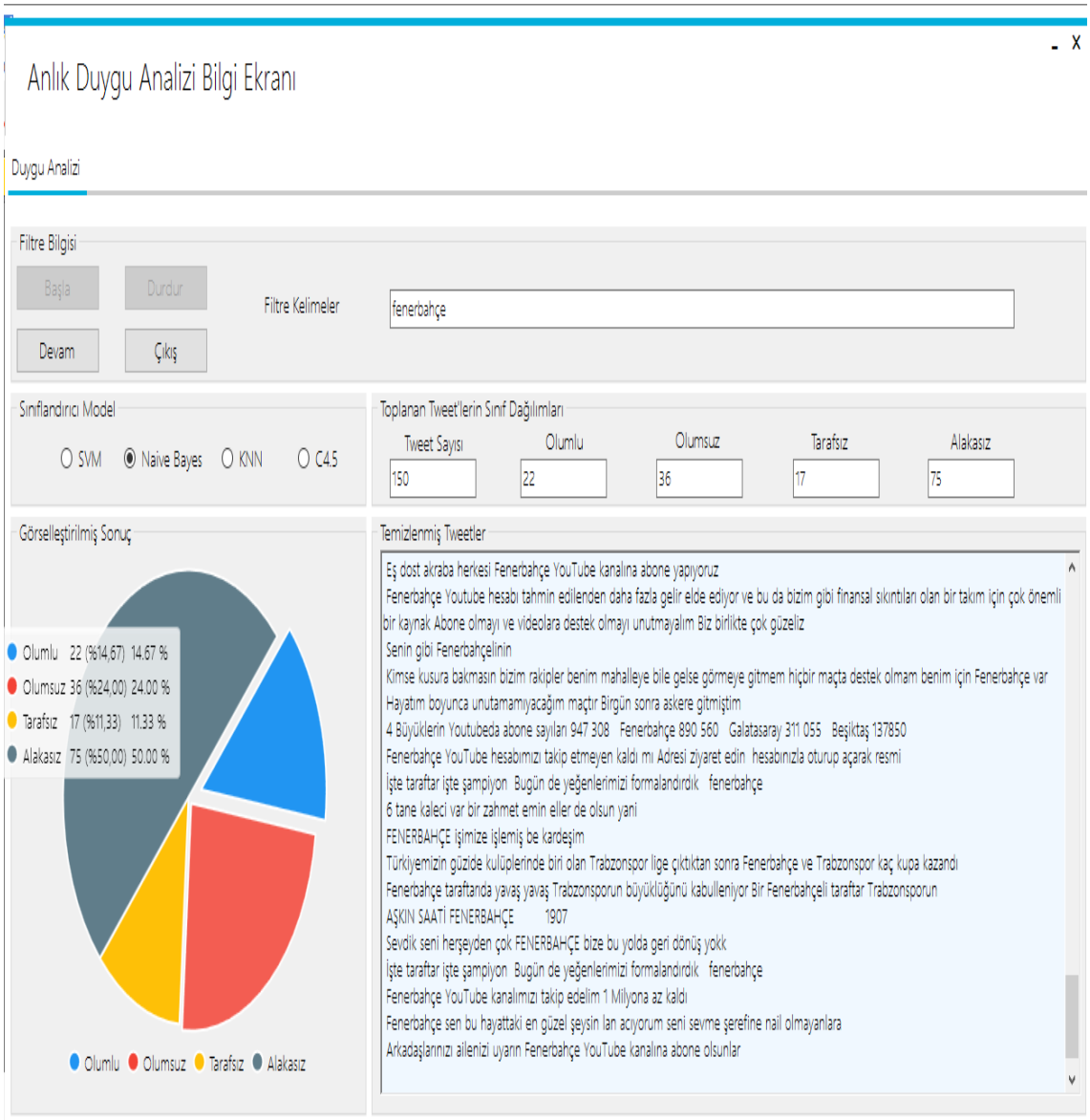
5.3. Anlık Twitter Duygu Analizi

Anlık olarak Twitter'dan paylaşılan tweetlerin duygularını belirlemek için oluşturulan hazır modellerden tercih ettiğimiz modelleri kullanan bir model uygulama geliştirilmiş ve kullanmıştır. Model uygulama, anlık olarak gelen yeni tweet paylaşımının olumlu, olumsuz, tarafsız ve alakasız olarak sınıfının ne olduğunu tahmin etmiştir. Şekil 5.1'de görüldüğü gibi anlık duygu analizi yapabilen bir yapı kurulmuştur. Anlık olarak aranan anahtar kelime ile ilgili olarak toplanan tweetlerin duygu durumu anlık olarak model uygulama arayüzünde her sınıf için grafik olarak ilgili kullanıcıya gösterilmiştir. Şekil 5.2'de model uygulamadan elde edilen sonuçları anlık gösteren örnek bir ekran verilmiştir. Geliştirilen uygulama bir model uygulama olarak düşünüldüğü için düşük donanım özelliklere sahip, yerel sunucu üzerinde çalıştırılmış ve böyle olmasına rağmen yine de çok yoğun olmayan tweet paylaşım dönemlerinde

başarılı olmuştur. Çok yoğun paylaşım olan dönemlerin üstesinden gelebilmek için uygulamanın daha yüksek işlemci gücü ve bellek kapasitesine sahip ortamlara taşınması gerekmektedir.



Şekil 5.1. Anlık duygu analizi döngüsü



Şekil 5.2. Anlık duygu analizi sonuçları

BÖLÜM 6

SONUÇLAR VE TARTIŞMA

Türkçe paylaşılmış tweetler üzerinden duygu analizi yapılan çalışmada tweetleri duygusal içeriklerine göre sınıflandırmak için 4 farklı makine öğrenmesi algoritması kullanılmıştır. Elde edilen sonuçlar, sınıflandırma algoritmalarının 4 sınıflı olarak oluşturulan eğitim setlerinde kabul edilebilir başarı oranlarına sahip olduğunu göstermiştir.

Duygu analizi uzun zamandır üzerinde çalışılan bir çalışma konusu olmasına rağmen özellikle bu alanda Türkçe veriler için yapılmış çok fazla çalışma mevcut olmadığından eğitim seti olarak kullanılabilecek örnek veriler de yeterli değildir. Bu nedenle, Twitter'da spor ile ilgili anahtar kelime tabanlı bir arama kullanarak Twitter'dan toplanan tweetler ile eğitim veri setleri oluşturulmuştur ve duygu sınıfları el ile etiketlenmiştir. Sınıf olarak olumlu, olumsuz, tarafsız ve alakasız gibi 4 sınıf kullanılmış olması önemlidir.

Çalışma, Zemberek Kütüphanesi ile hem Türkçe imla denetimi hem de Türkçe anlam denetimi yapmanın sınıflandırma başarımını artırabileceğini göstermiştir. Ayrıca KNN algoritmasının başarımının, öznitelik sayısının azalması ile arttığı görülmüştür. Çünkü öznitelik sayısının artış göstermesiyle boyut artmakta ve ilişkisiz özniteliklerin artması da ayırt ediciliği yüksek olan öznitelikleri etkisiz hâle getirmektedir. Bu durum KNN algoritmasının başarımını düşürmektedir. C4.5 algoritmasında sayısal değerleri çıkartmanın algoritma başarımına pozitif katkı sağladığı belirlenmiştir. Gereksiz kelimeler temizlenmiş, veri seti sadece Naive Bayes (kernelEstimator = false) algoritması ile en yüksek başarımı elde etmiştir. Çünkü tweetlerde çok fazla gereksiz kelime kullanılmamaktadır.

6.1. Katkıları

Sosyal medya üzerinden duyguları bir zaman diliminde anlık olarak analiz etmek duygular ile ilgili eğilimleri gösterebilir. Elde edilen sonuçlar, sosyal medyadan bir zaman diliminde elde edilen kullanıcı paylaşımlarındaki duyguların gerçek dünyadaki olaylarla paralellik gösterdiğini ifade etmektedir. Bu sebeple, bu çalışma; Twitter kullanıcılarının zaman içindeki duygu analizlerini yapmak ve duygu durumlarının tweetlerin içeriğine göre nasıl değiştiğini görmek için kullanılabilir. Bazı özel olaylara veya konulara yönelik halkın tepkisi gözlemlenebilir ve eğilimler belirlenebilir.

Model uygulama, sadece spor ve özellikle futbol konusu ile ilgilidir; ancak sonraki çalışmalarda eğitim, ekonomi, siyaset, sağlık, müzik vb. gibi farklı alanlarda eğitim veri setleri ve modeller oluşturularak uygulama birçok alana yönelik hâle getirilebilir.

6.2. Gelecek Çalışma Önerileri

Gelecek çalışmalarda; siyaset, müzik, ekonomi, sağlık vb. alanlarda veri setleri ve modeller oluşturulacaktır. Ayrıca çalışmalar için, duygu ifade eden tweetleri toplamak için kullanılan anahtar kelime listesi çok çeşitli tweetlere ulaşmak için geliştirilebilir. Anahtar kelimelerin bir listesinin oluşturulmasından sonra Zemberek kütüphanesi, belirli bir çözümleme işlemleriyle yeni anahtar kelimeler üretmek için kullanılabilir. Daha çeşitli ve daha fazla sayıda tweet içeren bir veri seti, tweetleri Türkçe olarak daha iyi genelleyebilir.

Sosyal medya platformları, insanların iletişim kurma biçimini değiştirmiştir. Bu yüzden kullanıcı görüşlerini analiz etmekte fayda vardır. Twitter, kullanıcıların bir konu hakkındaki görüşlerini paylaşabilecekleri, güncel konular hakkında tartışabilecekleri bir platformdur. Twitter verilerinden anlık olarak duygu analizi, kullanıcı görüşünü otomatik olarak olarak çıkarmakla ilgilenen yeni bir dönemdir.

Gelecekte yapılması planlanan çalışmalar:

1. Yeni özellikler eklemek: Duyguyu daha doğru algılamak ve mevcut sonuçtan daha iyi sonuç almaya yardımcı olacak bazı yeni özellikler ekleyerek çoklu anlam cümleleriyle başa çıkabilecek şekilde bir sistem geliştirilebilir.

2. Bulut Tabanlı Sistem: Bulut tabanlı bir uygulama yapılabilir. Her bir konu

hakkında duygu analizinin sürekliliğini sağlamak için bulut tabanlı bir platform eklenebilir.

3. Emojili tweetler ile çalışmak: Bu çalışmada sadece metin içeren tweetler ile çalışılmıştır. Gelecekte metin tweetleriyle ve emojili tweetler ile çalışılabilir.

4. Kullanıcıların diğer duygu etiketlerini tespit etmek: Bu çalışmada sadece olumlu, olumsuz, tarafsız ve alakasız duygu etiketleriyle çalışılmıştır. Çalışmada diğer duygu etiketleri yoktur. Duygu analizinde duygu sınıflarını artırarak daha iyi sonuçlar alınabilir.

5. Büyük miktarda tweet ile çalışmak: Gelecekte büyük miktarda tweet içeren veri setleriyle çalışılabilir. Daha fazla farklı konu alanları hakkında ve Twitter, Facebook ve Instagram, LinkedIn vb. sosyal medya platformlarından aynı konu ile ilgili anlık olarak duygu analizi sağlamak için çalışmalar yapılabilir.

6. Doğruluk ve başarım değerlendirmesi: Çalışmada doğruluk ve başarımı ölçmek için 4 makine öğrenmesi algoritması kullanılmıştır. Gelecekte diğer makine öğrenmesi algoritmaları da kullanılabilir.

7. Gerçek dünya problemleriyle çalışma: Etkili bir duygu analizi yaparak gerçek dünya problemlerinin çözümüne nasıl faydalı katkı sağlanabileceği ve bu çözümün nasıl uygulanabileceği üzerine çalışılabilir. Örneğin; bir hastalığın salgın olup olmayacağını tespit edebilmek, seçim sonuçlarını öngörebilmek, ürün itibarını tahmin edebilmek için çalışılabilir.

KAYNAKLAR

Akın, A. A. ve Akın, M. D. (2007). Zemberek, an open source nlp framework for turkish languages. Structure, 10:1–5.

A. Giachanou ve F. Crestani, "Like It or Not: A survey of Twitter Sentiment Analysis Methods," ACM Computing Surveys, vol. 49, no. 2, pp. 1–41, 2016.

A. L. Kavanaugh, E. A. Fox, S. D. Sheetz, S. Yang, L. T. Li, D. J. Shoemaker, A. Natsev, ve L. Xie, "Social Media Use by Government: From The Toutine to The Critical," Government Information Quarterly, vol. 29, no. 4, pp. 480–491, 2012.

B. Costa ve J. Boiney, "Social Radar," tech. rep., The MITRE Corporation,2012.

Bernhard E. Boser, Isabelle M. Guyon, ve Vladimir N. Vapnik. 1992. A training algorithm for optimal margin classifiers. In Proceedings of the fifth annual work shop on Computational learning theory (COLT '92). ACM, New York, NY, USA, 144-152. DOI=http://dx.doi.org.proxy.ub.umu.se/10.1145/130385.130401

Bollen, J., Mao, H., Zeng, X.-J. (2011). Twitter mood predicts the stock market. Journal of computational science, 2(1), 1–8.

B. Liu, Sentiment analysis: Mining opinions, sentiments, and emotions. Cambridge University Press, 2015.

B. Pang ve L. Lee., "Opinion mining and sentiment analysis", Foundations and Trends in Information Retrieval, vol. 2, no. 1-2, pp. 1-135, 2008.

Cohen J. A coefficient of agreement for nominal scales. Educ Psychol Meas. 1960;20:37–46.

Dergiades, T. (2012). Do investors' sentiment dynamics affect stock returns? Evidence from the US economy. Economics Letters, 116(3), pp. 404-407.

D. Newman, "Big Data and the Power of Prediction," Forbes, p. 2. Erişildi: 2019-04-18.

D. M. Boyd ve N. B. Ellison, "Social Network Sites: Denition, History, and Scholarship," Journal of Computer-Mediated Communication, vol. 13, no. 1, pp. 210–230, 2007.

Duggan, M., Page, D., Manager, S.C., "Social Media Update 2016." http://www.pewinternet.org/2016/11/11/social-media-update-2016/. Erişim: 2019-04-27.

E. Cambria, D. Das, S. Bandyopadhyay, ve A. Feraco, A practical guide to sentiment analysis, vol. 5. Springer, 2017.

E. Kouloumpis, T. Wilson, ve J. Moore, "Twitter Sentiment Analysis: The Good the Bad and the OMG!," in Artificial Intelligence, pp. 538–541, 2011.

Twitter.(2019). 20 Nisan 2019 tarihinde https://about.twitter.com/company.html/ adresinden erişildi.

Ficamos, P., Liu, Y., A Topic based Approach for Sentiment Analysis on Twitter Data, (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 7, No. 12, pp.201-205, 2016.

Hu, Y., Talamadupula, K., Kambhampati, S. (2013). Dude, srsly?: The surprisingly formal nature of twitter's language. In Proceedings of the Seventh International Conference on Weblogs and Social Media, pp. 244–253. AAAI Press.

H. Wang, D. Can, F. Bar ve S. Narayana, "A system for real-time Twitter sentiment analysis of 2012 U.S.presidential election cycle", Proc. ACL 2012 System Demostration, pp. 115-120, 2012.

Irina Rish. An empirical study of the naive Bayes classifier. IBM Research Report 2001. 22 Nisan 2019 tarihinde http://www.research.ibm.com/people/r/rish/papers/RC22230.pdf/ adresinden erişildi.

J. C. Bertot, P. T. Jaeger, ve D. Hansen, "The Impact of Polices on Government Social Media Usage: Issues, Challenges, and Recommendations," Government Information Quarterly, vol. 29, no. 1, pp. 30–40, 2012.

Kamber, Micheline Pei, Jian Han, Jiawei. Data Mining: Concepts and Techniques. :ElsevierScience; 2011.

Kethavath, S., Classification of Sentiment Analysis on Tweets using Machine Learning Techniques, Department of Computer Science and Engineering National Institute of Technology Rourkela Rourkela, India, 2015.

Khritantsev, M., Shehadeh, M., A Thesis in the Field of Finance For the Degree of Master of Science in Finance Lund University, 2018.

K. Ravi ve V. Ravi, "A survey on opinion mining and sentiment analysis: tasks, approaches and applications," Knowledge-Based Systems, vol. 89, pp. 14–46, 2015.

Liu, B., Sentiment analysis and opinion mining. Synthesis Lectures on Human Language Technologies., Vol. 5, No. 1, Pages 1-167, 2012 (https://doi.org/10.2200/S00416ED1V01Y201204HLT016).

Liu, Y., Huang, X., An, A., & Yu, X. (2007). ARSA: A Sentiment-Aware Model for Predicting Sales (Vol. SIGIR'07). Amsterdam.

Liu, I. L. B. , Cheung, C. M. K., Lee, M. K. O., "Understanding Twitter Usage: What drive people Continue to Tweet," in Proceedings of the Pacific Asia Conference on Information Systems, pp. 928–939, 2010.

M. Lynch, D. Freelon, ve S. Aday, "Syria's Socially Mediated Civil War," United States Institute of Peace, no. 91, p. 38, 2014.

Mozetič, I., Grčar, H., Smailović, J., Multilingual Twitter Sentiment Classification: The Role of Human Annotators, 2016. (https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0155036)

Nvidia. (2019). 17 Mayıs 2019 tarihinde https://blogs.nvidia.com/blog/2016/07/29/whats-difference-artificial-intelligence machine-learning-deep-learning-ai/ adresinden erişildi.

O'Connor, B., Balasubramanyan, R., Routledge, B. R., Smith, N. A. (2010). From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series. In Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media, pp. 122–129. AAAI Press.

Pak, A., Paroubek, P. (2010). Twitter as a corpus for sentiment analysis and opinion mining. In Proceedings of the Seventh International Conference on Language Resources and Evaluation, pp. 1320–1326. European Language Resources Association.

Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up? Sentiment Classification using Machine Learning. Proceedings of EMNLP, (pp. pp. 79–86).

P. Turney., "Thumbs up or thumbs down?" Semantic orientation applied to unsupervised classification of reviews. In Proceedings of the 40th annual meeting of the Association for Computational Linguistics,pp. 417424, 2002.

Rate Limiting - Twitter Developers.(2019). 2 Şubat 2019 Tarihinde https://developer.twitter.com/en/docs/basics/rate-limiting.html/ adresinden erişildi.

S. Aday, H. Farrell, M. Lynch, J. Sides, ve D. Freelon, "New Media and Conflict After the Arab Spring," United States Institute of Peace, no. 80, pp. 1–24, 2012.

S. Kullback ve R. A. Leibler. On information and sufficiency. The Annals of Mathematical Statistics, pages 79–86, 1951.

Twitter sansürü.(2019). 22 Nisan 2019 tarihinde https://en.wikipedia.org/wiki/Censorship_of_Twitter/ adresinden erişildi.

Twitter.(2018). 22 Ekim 2019 tarihinde https://developer.twitter.com/en/docs/basics/rate-limits.html/ adresinden erişildi.

Waikato, D. o. (2019). Weka 3: Data Mining Software in Java. Retrieved from http://www.cs.waikato.ac.nz/~ml/weka/. Erişim: 2019-03-27.

Witten, I.H. ve Frank, E. (2005) Data mining: Practical machine learning tools and techniques. 2nd Edition, Morgan Kaufmann Publisher, Burlington.

W. Marcellino, M. Smith, C. Paul, ve L. Skrabala, "Monitoring Social Media Lessons for Future Department of Defense Social Media Analysis in Support of Information Operations," tech. rep., 2017.

Wu, X., Kumar, V., Quinlan, J. R., Ghosh, J., Yang, Q., Motoda, H., Steinberg, D. (2007). Top 10 algorithms in data mining. Springer-Verlag.

EKLER

EK-1 Arff Dosyası İçerik Örneği

@Relation Twitter
@attribute tweet string
@attribute isletimsistemi {'iPhone','Android','Web','Diğer'}
@attribute tweetuzunluk{Uzunluk1, Uzunluk2, Uzunluk3, Uzunluk4, Uzunluk5, Uzunluk6, Uzunluk7}
@attribute tweetkesik {True,False}
@attribute tweethastagsay {Hashtags0, Hashtags1, Hashtags2, Hashtags3, Hashtags4, Hashtags5, Hashtags5+}
@attribute tweethassasicerik {True,False}
@attribute tweetgun{Pazartesi,Salı,Çarşamba,Perşembe,Cuma,Cumartesi,Pazar}
@attribute tweetzamanaralik{Dilim1, Dilim2, Dilim3,Dilim4, Dilim5, Dilim6, Dilim7, Dilim8, Dilim9, Dilim10, Dilim11, Dilim12}
@attribute tweetsadecesaat numeric
@attribute tweetsadecedakika numeric
@attribute hesapyili numeric
@attribute hesapyayi {1,2,3,4,5,6,7,8,9,10,11,12}
@attribute hesaptweetsay {1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,22,23,24}
@attribute hesaptakipcisay {1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,22,23,24}
@attribute hesaparkadassay {1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,22,23,24}
@attribute hesapfavorisay {1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,22,23,24}
@attribute hesaprozetli {True,False}
@attribute hesapkorumali {True,False}
@attribute hesapkonumacik {True,False}
@attribute yeradi{'İstanbul','Ankara','İzmir','Adana','Adıyaman','Afyonkarahisar','Ağrı','Aksaray','Amasya','Antalya','Ardahan','Artvin','Aydın','Balıkesir','Bartın','Batman','Bayburt','Bilecik','Bingöl','Bitlis','Bolu','Burdur','Bursa','Çanakkale','Çankırı','Çorum','Denizli','Diyarbakır','Düzce','Edirne','Elazığ','Erzincan','Erzurum','Eskişehir','Gaziantep','Giresun','Gümüşhane','Hakkari','Hatay','Iğdır','Isparta','Kahramanmaraş','Karabük','Karaman','Kars','Kastamonu','Kayseri','Kırıkkale','Kırklareli','Kırşehir','Kilis','Kocaeli','Konya','Kütahya','Malatya','Manisa','Mardin','Mersin','Muğla','Muş','Nevşehir','Niğde','Ordu','Osmaniye','Rize','Sakarya','Samsun','Siirt','Sinop','Sivas','Şırnak','Tekirdağ','Tokat','Trabzon','Tunceli','Şanlıurfa','Uşak','Van','Yalova','Yozgat','Zo

nguldak','KonumYok'}
@attribute class{Olumlu,Olumsuz,Tarafsız,Alakasız}

@DATA
' beşiktaş böyle oyna canımı yee'
,Android,Uzunluk1,False,Hashtags0,False,Cumartesi,Dilim10,19,31,2015,2,4,3,3,13,False,False,True,'KonumYok',Olumlu
'madem hakemlerden şikayetçiyiz hadi çıksın hem galatasaray hem beşiktaş yöneticiler yurtdışından hakem istesin'
,Android,Uzunluk4,False,Hashtags0,False,Pazar,Dilim1,0,58,2018,8,3,3,1,10,False,False,False,'KonumYok',Olumsuz
' ankaragücü beşiktaş maçı canlı yayın '
,Diğer,Uzunluk4,False,Hashtags2,False,Cumartesi,Dilim10,18,54,2010,4,24,6,1,1,False,False,False,'KonumYok',Tarafsız
' beşiktaş çok sessiz ne kadar az insan var etrafta ne güzel '
,iPhone,Uzunluk4,True,Hashtags0,False,Pazar,Dilim1,0,27,2016,12,6,1,1,6,False,False,True,'KonumYok',Alakasız

EK-2 Gereksiz Kelimeler

a	birçok	dan	falan	için	mı	olması	sana	şunları
acaba	biri	de	filan	içinde	milyar	olmayan	sanki	şunu
altı	birisi	defa	gene	iki	milyon	olmaz	sanki	ta
altmış	birkaç	değil	gereği	ile	mu	olsa	şayet	tabii
ama	birşey	diğer	gerek	ilgili	mü	olsun	şekilde	tam
ancak	biz	diğeri	gibi	ise	nasıl	olup	sekiz	tamam
arada	bizden	diğerleri	göre	işte	ne	olur	seksen	tamamen
artık	bize	diye	hala	itibaren	neden	olur	sen	tarafından
asla	bizi	doksan	halde	itibariyle	nedenle	olursa	senden	trilyon
aslında	bizim	dokuz	halen	kaç	nerde	oluyor	seni	tüm
aslında	böyle	dolayı	hangi	kadar	nerede	on	senin	tümü
ayrıca	böylece	dolayısıyla	hangisi	karşın	nereye	ön	şey	u
az	bu	dört	hani	kendi	neyse	ona	şeyden	ü
bana	buna	e	hatta	kendilerine	niçin	önce	şeye	üç
bazen	bunda	edecek	hem	kendine	nin	ondan	şeyi	un
bazı	bundan	eden	henüz	kendini	nın	onlar	şeyler	ün
bazıları	bunlar	ederek	hep	kendisi	niye	onlara	şimdi	üzere
belki	bunları	edilecek	hepsi	kendisine	nun	onlardan	siz	var
ben	bunların	ediliyor	her	kendisini	nün	onları	siz	vardı
benden	bunu	edilmesi	herhangi	kez	o	onların	sizden	ve
beni	bunun	ediyor	herkes	ki	öbür	onu	sizden	veya
benim	burada	eğer	herkese	kim	olan	onun	size	ya
beri	bütün	elbette	herkesi	kime	olarak	orada	sizi	yani
beş	çoğu	elli	herkesin	kimi	oldu	öte	sizi	yapacak
bile	çoğunu	en	hiç	kimin	olduğu	ötürü	sizin	yapılan
bilhassa	çok	etmesi	hiçbir	kimisi	olduğunu	otuz	sizin	yapılması
bin	çünkü	etti	hiçbiri	kimse	olduklarını	öyle	sonra	yapıyor
bir	da	ettiği	i	kırk	olmadı	oysa	şöyle	yapmak
biraz	daha	ettiğini	ı	madem	olmadığı	pek	şu	yaptı
birçoğu	dahi	fakat	ın	mi	olmak	rağmen	şuna	yaptığı
zira	yirmi	yerine	zaten	yine	nde	yu	yi	yaptığını
zxtest	yoksa	yetmiş	ye	yedi	yüz	yı	yaptıkları	

ÖZGEÇMİŞ

Adı Soyadı: Rıza KORKUSUZ
Telefon: 0506 351 00 10
e-posta: rizakorkusuz@gmail.com

Aslen Kayserili olup 07.02.1985 yılında Adana'da doğdu. İlk – orta ve lise öğrenimini Hatay ili İskenderun ilçesinde tamamlamıştır. İskenderun Mesleki ve Teknik Anadolu Lisesi Bilişim Teknolojileri Alanından 2003 yılında birincilikle mezun olmuştur. Sonra 2009 yılında Marmara Üniversitesi Teknik Eğitim Fakültesinden mezun oldu. 2014 yılında Trakya Üniversitesi Mühendislik Fakültesi Bilgisayar Mühendisliği Bölümünden mezun oldu. Hâlen Trakya Üniversitesi Fen Bilimleri Enstitüsü Bilgisayar Mühendisliği Anabilim Dalında tez çalışmasına devam etmektedir.

2010 yılında İstanbul'da Fikriye Nüzhet Bilgincan Mesleki ve Teknik Anadolu Lisesinde öğretmen olarak göreve başlamış, 2012 yılında Edirne Selimiye Mesleki ve Teknik Anadolu Lisesinde görevine devam etmiş, 2014-2016 yılları arasında Edirne Selimiye Mesleki ve Teknik Anadolu Lisesinde Bilişim Teknolojileri Alan Şefi olarak görevde bulunduktan sonra 2016 yılında müdür yardımcılığı sınavını kazanarak Edirne Emel Özgür Subaşıay Mesleki ve Teknik Anadolu Lisesine müdür yardımcısı olarak atanmıştır. 20 Eylül 2016 tarihinden itibaren bu görevini sürdürmektedir.

Evli ve bir çocuk babasıdır.