T.C.

TRAKYA ÜNİVERSİTESİ

FEN BİLİMLERİ ENSTİTÜSÜ

TWITTER PLATFORMU ÜZERİNDEN MAKİNE ÖĞRENMESİ ALGORİTMALARI İLE CİNSİYET VE İLGİ ALANI ANALİZİ

ENES GÜNÇE

YÜKSEK LİSANS TEZİ

BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI

TEZ DANIŞMANI: Dr. Öğretim Üyesi Aydın CARUS

EDİRNE-2021

ENES GÜNÇE ’nin hazırladığı “Twitter Platformu Üzerinden Makine Öğrenmesi Algoritmaları ile Cinsiyet ve İlgi Alanı 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 Aydın CARUS ……

Doç. Dr. Erdinç UZUN ……

Dr. Öğretim Üyesi Altan MESUT ……

Tez Savunma Tarihi: 05/02/2021

Bu tezin Yüksek Lisans tezi olarak gerekli şartları sağladığını onaylarım.

İmza

Dr. Öğretim Üyesi Aydın CARUS

Tez Danışmanı

……

Trakya Üniversitesi Fen Bilimleri Enstitüsü onayı

……

Doç. Dr. H. R. Ferhat KARABULUT

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 tamamı ya da herhangi bir bölümünün daha önceden Trakya Üniversitesi ya da farklı bir üniversitede tez çalışması olarak sunulmadığını beyan ederim.

05 / 02 / 2021

Enes GÜNÇE

İmza

Yüksek Lisans Tezi

Twitter Platformu Üzerinde Makine Öğrenmesi Algoritmaları ile Cinsiyet ve İlgi Alanı Analizi

T.Ü. Fen Bilimleri Enstitüsü

Bilgisayar Mühendisliği Anabilim Dalı

ÖZET

Twitter, 2006 yılından itibaren hayatımızda bulunan, 330 milyondan fazla aktif kullanıcısı olan, kısa cümleler ile yaratıcı anlatımlarda bulunulmasını hedefleyen mikro blog sitesidir. İlk kurulduğunda daha çok eğlence ve arkadaşlar arası iletişim amaçlı kullanılan Twitter'ın bu özelliği zamanla evrilmiştir. Günümüzde insanlar, Twitter'ı daha çok; haber takibi, reklam, kurum ve kuruluşların marka bilinirliğini arttırma, kişisel kullanım ve özel hayatın paylaşımının ötesinde, profesyonel hayatın akışı için de kullanmaktadır. Twitter'ın evrim geçiren bu modelinin fırsat olarak görülmesi, artık Twitter'ı eğlenceden ziyade daha fazla insana ulaşma aracı ve sürekli analizleri yapılan bir platform haline getirmiştir.

Twitter, kullanıcıları hakkında birçok bilgi sağlamasına rağmen; kişinin yaşı, eğitim durumu, cinsiyeti ve kullanıcıların ilgi alanları gibi bilgileri sağlamamaktadır. Bu sebeple, bu çalışmada, bu bilgilerden cinsiyet ve tweet ilgi alanı çıkarılması hedeflenmiştir. Twitter üzerinden çıkarılan bu bilgiler, günümüzde kişiye özel ürün tanıtım ve tavsiyesi, yasal soruşturma gibi birçok amaca hizmet edecek şekilde kullanılabilmektedir.

Bu çalışmada; Tweetlerden cinsiyet ve tweetin ilgi alanı hakkında bilgi elde edilmesine yönelik, Türkçe tweetler toplanmış ve bu tweetlerin el ile etiketleme işlemi yapılarak, denetimli makine öğrenmesi algoritmaları için farklı özniteliklere sahip eğitim setleri oluşturulmuştur. Bunlardan, cinsiyet eğitim seti; Erkek, Kadın ve diğer

(kurum/kuruluş) olarak 3 sınıf içerecek şekilde tasarlanmıştır. İlgi alanı eğitim seti ise Bilim ve Teknoloji, Siyaset, Spor, Eğitim, Ekonomi ve Diğer olmak üzere toplamda 6 sınıftan oluşmaktadır. Bu eğitim setleri kullanılarak, makine öğrenmesi algoritmaları ile modeller çıkarılmış ve sınıflandırma başarımları bulunmuştur. Cinsiyet eğitim seti için en iyi sınıflandırma doğruluk başarımının; Naive Bayes algoritması için %91,4, K-En Yakın Komşu Algoritması (KNN- K-Nearest Neighbors) algoritması için %95,7, C4.5 karar ağacı algoritması için %99,4, Destek Vektör Makineleri (SVM- Support Vector Machine) algoritması için %81,9 ve Ardışık Minimal Optimizasyon (SMO- Sequential Minimal Optimization) algoritmaları %99,9 olduğu tespit edilmiştir. Benzer şekilde ilgi alanı eğitim seti için en iyi sınıflandırma başarımının: Naive Bayes algoritması için %84,7, KNN algoritması için %81,9, C4.5 karar ağacı algoritması için %92,1, SVM algoritması için %90,0 ve SMO algoritması için %93,18 doğrulukta olduğu tespit edilmiştir.

Yıl : 2021

Sayfa Sayısı : 92

Anahtar Kelimeler : Twitter, Makine Öğrenmesi, Cinsiyet Analizi, İlgi Alanı Analizi

Master Thesis

Gender and Interest Analysis with Machine Learning Algorithms on Twitter Platform

Trakya University Institute of Naturel Sciences

Computer Engineering Department

ABSTRACT

Twitter is a micro blog social network platform that has been in our lives since 2006, has more than 330 million active users and aims to allow its users to communicate through creative expressions with short sentences. This feature of Twitter, which was used mostly for entertainment and communication between friends when it was first established, has evolved over time. Beyond news follow-up, advertising, increasing brand awareness of institutions and organizations, today Twitter is also used for professional life. Consequently, Twitter has become an essential platform for a variety of use cases other than entertainment whereas a deep data investigation and analysis can be helpful to understand the public.

Although Twitter provides a lot of information about its users; It does not provide information such as the age, education, gender and interests of the users. Therefore, in this study, it is aimed to extract gender and area of interest from these informations. This information obtained through Twitter can be used today to serve many purposes such as personalized recommender systems, also for legal investigation.

In this study; Turkish tweets were collected to obtain information about gender and the area of interest of the tweet from the tweets, and by manually tagging these tweets, training sets with different attributes were created for supervised machine learning algorithms. The first datasete contains the gender training data set is designed to include 3 classes as Male, Female and other (institution / organization). The area of interest training set consists of 6 classes in total: Science and Technology, Politics,

Sports, Education, Economy and Other. By using these training sets, models with machine learning algorithms have been created and classification performance has been measured. The best classification accuracy achievement for the gender training set; 91.4% for the Naive Bayes algorithm, 95.7% for the K-Nearest Neighbor Algorithm algorithm, 99.4% for the C4.5 decision tree algorithm, Support Vector Machines Machine algorithm is 81.9% and Sequential Minimal Optimization algorithms were found to be 99.9%. Similarly, the best classification performance for the interest data set: 84.7% for the Naive Bayes algorithm, 81.9% for the KNN algorithm, 92.1% for the C4.5 decision tree algorithm, 90.0% for the SVM algorithm and SMO It was determined to be 93.18% accurate for the algorithm.

Year : 2021

Number of Pages : 92

Keywords : Twitter, Machine Learning, Gender Analysis, Interest Area 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 değerli katkılarından dolayı Kırklareli Üniversitesi Yazılım Mühendisliği Bölümü Dr. Öğretim Üyesi Edip Serdar GÜNER'e teşekkürü bir borç bilirim.

Çalışmalarım esnasında katkılarından dolayı ablam Şuheda GÜNÇE'ye ve değerli fikirlerini bana aktaran arkadaşım Rıza KORKUSUZ'a teşekkürlerimi sunarım.

Enes GÜNÇE

Edirne, 2021

İÇİNDEKİLER

ÖZET ... iv

ABSTRACT ... vi

TEŞEKKÜR ... viii

İÇİNDEKİLER ... ix

KISALTMALAR DİZİNİ ... xii

ÇİZELGELER DİZİNİ ... xiii

ŞEKİLLER DİZİNİ ... xiv

BÖLÜM 1 ... 1

GİRİŞ ... 1

1.1. Sosyal Medya Hakkında ... 1

1.2. Twitter Nedir? ... 1

1.3. Neden Twitter Analizi Yapılıyor? ... 2

1.4. Twitter Hakkında Yapılan Bazı Çalışmalar ... 3

1.5. Çalışmanın Amacı ... 4

BÖLÜM 2 ... 5

METODOLOJİ ... 5

2.1. Mimari Yapı ... 5

2.2. Tweet Toplama ... 8

2.2.1. Tweet Toplama Modülü 9

2.3. Tweetlerin Temizlenmesi 11

2.3.1. Düzenli İfadeler 11

2.4. Öznitelik Çıkarımı 12

2.5. Tweetlerin Etiketlenmesi 13

2.6. Eğitim Setleri 16

BÖLÜM 3 20

MAKİNE ÖĞRENMESİ 20

3.1. Makine Öğrenmesi Nedir? 20

3.2. Sınıflandırma 21

3.3. Makine Öğrenmesi Algoritmaları 22

3.3.1. NAIVE BAYES Algoritması 23

3.3.2. KNN Algoritması 24

3.3.3. SVM Algoritması 25

3.3.4. SMO Algoritması 26

3.3.5. C4.5 Algoritması 26

BÖLÜM 4 27

CİNSİYET VE İLGİ ANALİZİ İÇİN YAPILAN ÇALIŞMA 27

4.1. Geliştirilen Uygulama 27

4.2. JSON Nedir? 32

4.2.1. JSON Twitter Veri Biçimi 33

4.2.2. JSON Biçimindeki Alanlar ve Anlamları 34

4.3. WEKA Nedir? 35

4.3.1. WEKA Arayüzü ve Modülleri 35

4.3.2. ARFF Formatı Nedir? 37

BÖLÜM 5 39

SONUÇLAR 39

5.1. Deneysel Sonuçlar 39

5.2. Cinsiyet Eğitim Seti Sonuçları 41

5.3. Cinsiyet Eşsiz Kullanıcı Eğitim Seti ve Sonuçları 49

5.4. İlgi Alanı Eğitim Seti Sonuçları 50

5.5. Sonuç ve Değerlendirme 66

KAYNAKLAR 68

EKLER 70

ÖZGEÇMİŞ 76

KISALTMALAR DİZİNİ

API	Uygulama Programlama Arayüzü (Application Programming Interface)
ARFF	Özellik-İlişki Dosyası Biçimi (Attribute-Relation File Format)
CSV	Virgülle Ayrılmış Değerler (Comma Seperated Values)
JSON	Javascript Nesnesi Gösterimi (Javascript Object Notation)
KNN	En Yakın k Komşu (K Nearest Neighborhood)
NB	Naive Bayes
SMO	Ardışık Minimal Optimizasyon (Sequential Minimal Optimization)
SVM	Destek Vektör Makineleri (Support Vector Machine)
WEKA	Waikato Environment for Knowledge Analysis

ÇİZELGELER DİZİNİ

Çizelge 2.1. Cinsiyet eğitim seti için belirlenen öznitelikler 12

Çizelge 2.2. Toplanan cinsiyet eğitim seti örnek adedi 16

Çizelge 2.3. Toplanan ilgi alanı eğitim seti örnek adedi 17

Çizelge 2.4. İlgi alanı anahtar kelimeler 18

Çizelge 5.1. Kappa değeri uyuşma tablosu 40

Çizelge 5.2. Cinsiyet doğruluk ve kappa değerleri 42

Çizelge 5.3. Eşsiz kullanıcı eğitim seti örnek adetleri 49

Çizelge 5.4. Cinsiyet eğitim setlerinin doğruluk sonuçları 50

Çizelge 5.5. 30.000 ilgi alanı eğitim seti doğruluk ve kappa sonuçları 51

Çizelge 5.6. 90.000 ilgi alanı alanı eğitim seti doğruluk ve kapp sonuçları 58

Çizelge 5.7. 30.000 ve 90.000 örneğe sahip eğitim setlerinin doğruluk karşılaştırması 65

ŞEKİLLER DİZİNİ

Şekil 2.1. Tweet toplama ve önişleme süreci mimari yapısı 6

Şekil 2.2. Cinsiyet analizi mimari yapısı 7

Şekil 2.3. İlgi alanı analizi mimari yapısı 8

Şekil 2.4. Tweetinvi kütüphane kullanımı 9

Şekil 2.5. App.config dosyasında twitter api tanımlama 9

Şekil 2.6. Twitter api'ye c# programlama dili içinden erişim..... 10

Şekil 2.7. Toplanan tweet verisinin bazı alanları 13

Şekil 2.8. Profil linkinin aratılmasından sonra görünen resim 14

Şekil 4.1. Tweet toplamak için geliştirilen uygulama arayüzü 28

Şekil 4.2. Geliştirilen uygulamanın oluşturduğu dosyalar..... 29

Şekil 4.3. ArffText dosya formatı içeriği..... 30

Şekil 4.4. Etiketsiz veriler dosya formatı içeriği..... 30

Şekil 4.5. Temiz veriler dosya formatı içeriği..... 31

Şekil 4.6. Ham veriler dosya formatı içeriği..... 31

Şekil 4.7. JSON biçiminde bir tweet verisi..... 33

Şekil 4.8. Belli alanlar içeren json tweet biçimi..... 34

Şekil 4.9. WEKA arayüzü..... 36

Şekil 4.10. ARFF Formatı 37

Şekil 5.1. NB algoritmasının cinsiyet eğitim seti doğruluk ve kappa değerleri 43

Şekil 5.2. SMO algoritmasının cinsiyet eğitim seti doğruluk ve kappa değerleri 44

Şekil 5.3. KNN algoritmasının cinsiyet eğitim seti doğruluk ve kappa değerleri 45

Şekil 5.4. SVM algoritmasının cinsiyet eğitim seti doğruluk ve kappa değerleri 46

Şekil 5.5. C4.5 algoritmasının cinsiyet eğitim seti doğruluk ve kappa değerleri 47

Şekil 5.6. Cinsiyet doğruluk değerlerinin karşılaştırılması 48

Şekil 5.7. 30.000 ilgi alanı eğitim seti NB doğruluk ve kappa değerleri 52

Şekil 5.8. 30.000 ilgi alanı eğitim seti KNN doğruluk ve kappa değerleri 53

Şekil 5.9. 30.000 ilgi alanı eğitim seti SMO doğruluk ve kappa değerleri 54

Şekil 5.10. 30.000 ilgi alanı eğitim seti C4.5 doğruluk ve kappa değerleri 55

Şekil 5.11. 30.000 ilgi alanı eğitim seti SVM doğruluk ve kappa değerleri 56

Şekil 5.12. 30.000 ilgi alanı eğitim seti doğruluk karşılaştırması 57

Şekil 5.13. 90.000 ilgi alanı eğitim seti NB doğruluk ve kappa değerleri. 59

Şekil 5.14. 90.000 ilgi alanı eğitim seti KNN doğruluk ve kappa değerleri 60

Şekil 5.15. 90.000 ilgi alanı eğitim seti SMO doğruluk ve kappa değerleri 61

Şekil 5.16. 90.000 ilgi alanı eğitim seti C4.5 doğruluk ve kappa değerleri 62

Şekil 5.17. 90.000 ilgi alanı eğitim seti SVM doğruluk ve kappa değerleri 63

Şekil 5.18. 90.000 ilgi alanı eğitim seti doğruluk karşılaştırma 64

BÖLÜM 1

GİRİŞ

1.1. Sosyal Medya Hakkında

Günümüzde, sosyal medya platformlarının hızlı yükselişi ile bu platformlardaki kullanıcı sayısının ve yapılan paylaşımların miktarında büyük bir artış meydana gelmiştir. İnsanların, birçok konuda kişisel fikirlerini paylaştıkları bu platformlar, önemli birer veri kaynağı haline gelmiştir. Sosyal ağ; insanları bir araya getirmek, fikir ve ilgi alanlarını paylaşmak veya yeni arkadaşlar edinmek için bir araya getiren bir platform haline gelmiştir. Bu tür bir iş birliği ve paylaşım, sosyal medya olarak bilinmektedir. Aynı zamanda Sosyal ağlar, insanların arkadaşlarıyla ve aileleriyle bağlantılarını sürdürmelerine yardımcı olmaktadır. Ayrıca kullanıcılar sosyal çevrelerinde her gün neler olduğunu kolay bir şekilde izleyebilmektedir. Sosyal ağlar, eğlence ve kullanıcıların ilgilerini çekebilecek birçok konuyu aramak için de kullanılabilmektedir.

1.2. Twitter Nedir?

Mart 2006 tarihinde San Francisco'da prototip olarak ortaya çıkan Twitter'ın resmi başlangıcı, Ağustos 2006 tarihine denk gelmektedir. Bu başlangıcın, şirket haline dönüşmesinin Mayıs 2007 tarihi olduğu değerlendirildiğinde, bu kısa sürede geldiği noktayla başarı kelimesinin ne kadar örtüştüğü görülmektedir. Twitter'ın 2017 yılında yaklaşık 330 milyon üyesi mevcuttur. Birçok kişi ve kurum, açtığı Twitter hesabıyla kendisi hakkındaki gelişmeleri paylaşmaktadır. Twitter, uzun yıllar tüm kullanıcılar için 140 karakterlik kısa mesaj sınırlaması koymuştur. Fakat, 2017 itibari ile bu 140

karakterlik sınır 280 karakter olarak değiştirilmiştir. Twitter'da kullanıcıların birbirlerinin hesaplarını takip etmesine dayanan bir sistem mevcuttur. Site tarafından, kullanıcılara, takip ettikleri kişilerin paylaştıkları tweetler, kullanıcının kendi ana sayfasında gösterilmektedir.

Twitter'da kullanıcılar, tweet paylaşmak dışında, başkaları tarafından paylaşılmış olan tweetleri, retweet edebilmekte veya favorileyebilmektedir. Retweet etme işlemi, tweetin altındaki "retweet" tuşuna basılarak, başkasına ait bir tweetin, retweet eden kullanıcının profilinde de görünmesini sağlayan işlemdir. Favorileme ise, tweetin altındaki favori tuşuna basılarak, tweeti favorileyen kullanıcının favoriler listesinde görünmesini sağlayan işleme denilmektedir. Kullanıcılar, beğendikleri tweetleri retweet ederek ya da favorileyerek takipçilerine paylaşabilmektedir. Ayrıca Twitter, anlık olarak en fazla 10 konuyu listelediği ve "Gündem" adını verdiği listeler barındırmaktadır. Bu listeler, tweetler içinde en çok kullanılan sözcüklerden oluşmaktadır. Bunlarda şehir bazında, ülke bazında ve dünya genelinde en çok konuşulan konuları gösteren sözcük listeleri mevcuttur. Kullanıcılar, aynı sözcükleri barındıran çok sayıda tweet atarak bu listeye girip, dikkat çekmeye çalışmaktadır. İnsanların Twitter'ı tercih etme sebepleri genel olarak; hız, kolay erişim, anında bildirim, doğru kaynak, eğlence gibi geniş yelpazeye yayılmaktadır. İnsanlar, artık haberleri anında alabilme gibi teknolojik çağın sağladığı imkânlardan faydalanınca, zamanlarının büyük bir kısmını bu teknoloji ile geçirmeye başlamıştır (Andrews, 2009).

1.3. Neden Twitter Analizi Yapılıyor?

Eskiden, anketler ile elde edilen kullanıcı görüşlerinin, kullanıcılar tarafından gönüllü olarak sosyal medyada paylaşılması ile firmalar tarafından daha çabuk ulaşılabilir hale gelmesi, firmaların kullanıcı isteklerine göre ürün ve hizmet geliştirmesini kolaylaştırmıştır.

Twitter'daki veri miktarının büyüklüğü ve çeşitliliği, birçok araştırmacının ilgisini çekmektedir. Sosyal medya verileri ile insanların çeşitli konulardaki düşünceleri ölçülebilir fakat bu verilerin miktarı, çoğu zaman gözle kontrol edilip anlamlandırılamayacak kadar büyük olabilmektedir. Sosyal medya verileri kullanılarak, sadece mevcut düşünceler değil, geleceğe yönelik çeşitli çıkarımlar da elde edilebilmektedir. Analiz yapılmasının önemli sebeplerinden birisi ise kişilerin herhangi

bir konu üzerine kendi fikirlerini yazmış olmasından kaynaklanmaktadır. Genelde, anketler sokaklarda ya da kurumların içinde müşterilerine yaptıkları kâğıt üzerinde ya da elektronik ortamda bir analiz üzerine kuruludur. Fakat yapılan bu anketlerin doğruluk değerlerinin düşük olduğu gözlemlenmiştir. Bunun nedeni, kişilerin analiz yapmak için hazırlanan sorulara yanlış ya da kendi fikrinden farklı cevap vermesidir. Kullanıcıların bu şekilde yanıltıcı cevap verme sebeplerinden birisi ise anket sorularının fazlalığına bağlıdır. Youtube videolarında, 5 saniyelik reklamlara bile sabrı olmayan insanların, 100 soruluk bir anketi sabır içerisinde cevaplamaları beklenemez. Bu sebeple bu anketten de üst seviye bir doğruluk oranı beklenmesi inandırıcı olmayacaktır. Diğer sebeplerden birisi ise uzun ve karmaşık soruların sorulmuş olmasıdır (Cox, 2015).

İnsanların Twitter üzerinden analiz yapmalarının ana faktörlerinden biri, şirketlerin marka bilinirliliğini arttırmak istemeleridir. Bu sayede insanlar ile direk iletişim halinde olmaları, arada herhangi bir aracının olmaması, şirketler için bulunmaz bir fırsata dönüşmektedir.

1.4. Twitter Hakkında Yapılan Bazı Çalışmalar

Sosyal medya verileri üzerinde sınıflandırma ve verileri kullanarak çıkarımlar yapmak amacıyla yapılmış çok sayıda çalışma mevcuttur. Twitter kullanıcısının dinamik profili için ilgi tahmini adlı çalışma (Khodra, Siswanto, & Dewi, 2014), ilgi alanı ile ilgili yapılan çalışmalardan bir tanesidir. Bu çalışmada kullanıcıların tweetlerini çok etiketli sınıflandırma metodu ve belirli kullanıcı hesaplarını temel alarak sınıflandırma işlemleri yapılmıştır. Cinsiyet tahmini için yapılan farklı bir çalışmada ise; ekran ismi (screen_name), profil resmi (image) ve tweet kullanılmıştır. Sonuç olarak %89,9 gibi bir doğruluk değerine ulaşılabilmiştir (Sayyadiharikandeh, Ciampaglia, & Flammini, 2016). “Who Likes What” isminde milyonlarca Twitter kullanıcısının paylaştığı tweetleri toplayan bir sistem tasarlayıp, insanların ilgi alanlarını tahmin edebilen çalışma da yapılmıştır (Bhattacharya, Zafar, & Ganguly, 2014). Bir başka çalışmada; yazarların cinsiyetlerini, toplanan tweet örneklerinin el ile etiketlenmesinin ardından %82,48 oranında bir doğruluk yüzdesi ile tahmin edebilmişlerdir (Deitrick, Miller, Valyon, Dickinson, Munson, & Hu, 2012). Cinsiyet alanında yapılan bir diğer çalışmada (Wood-Doughty, Andrews, Marvin, & Dredze, 2018) ise, kişilerin sadece

isim ve ekran isimlerinden yola çıkarak cinsiyet ve etnik köken hakkında tahminde bulunulmuştur.

1.5. Çalışmanın Amacı

Kolay kullanımı ve paylaşımları hızlı bir şekilde birçok kişiye ulaştıran tasarımı sayesinde Twitter, insanların kendini ifade etmek için kullandığı ortamlardan biri haline gelmiştir. Aynı zamanda birçok firma ve bilinen kişilerinde birer kullanıcı olarak yer aldığı bu ortam, kullanıcıların fikirlerini, şikâyetlerini, övgülerini ve sorularını istedikleri kişiye hızlı ulaştırıp, cevap alabildiği bir iletişim aracı olarak da görev yapmaktadır. Twitter'da paylaşılan tüm bu veriler analiz edilerek, kişilerin spesifik özelliklerini tahmin etmek mümkündür. Bu çalışmanın amacı, toplanan türkçe tweetler üzerinden, tweeti paylaşan kullanıcının cinsiyet ve tweetlerin ilgi alanı üzerinden kullanıcının ilgi alanı tahmininin yapılmasıdır. Bunun sağlayacağı en önemli avantaj, kullanıcının cinsiyeti ve ilgi alanına bağlı olarak kişiye özel ürün öneri modeli verilebilir.

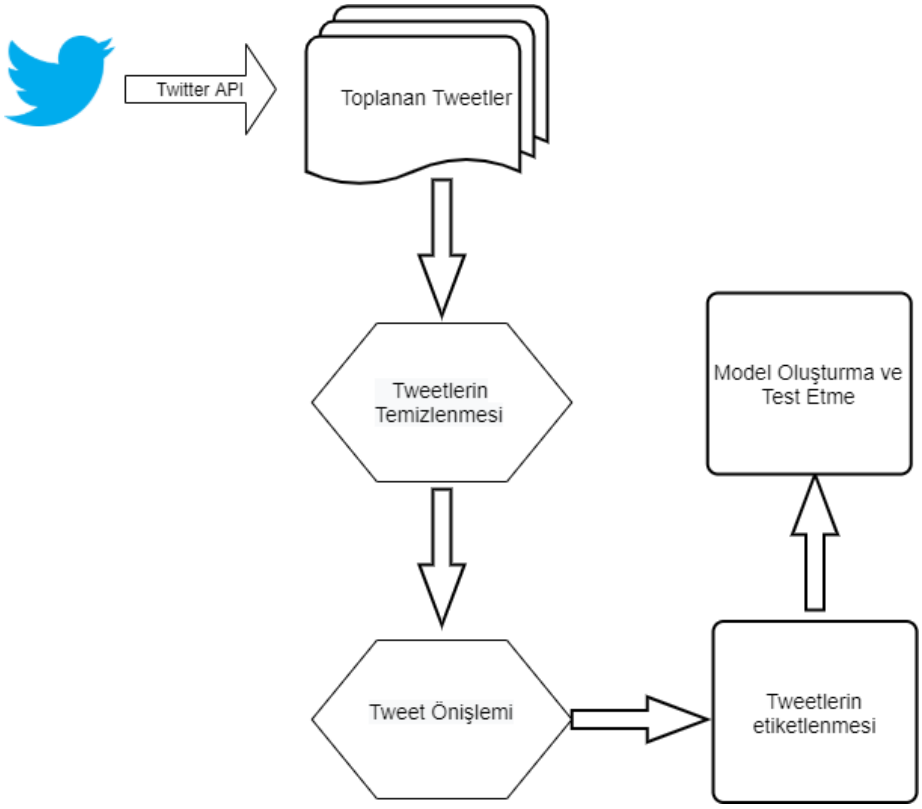
Yapılan çalışmada, makine öğrenmesi algoritmaları ile cinsiyetin tahmin edilmesi ve paylaşılan içeriğin ne ile ilgili olduğunun tahmin edilmesi amaçlanmıştır. Bu amaçla, Twitter yazılım araçlarını kullanarak geliştirilen yazılımlar aracılığı ile, Twitter platformundan tweetler toplanarak ve bu tweetler ile, makine öğrenmesi için gerekli olan eğitim setleri hazırlanmıştır. Bunun için cinsiyet ve ilgi alanı tahmini için kullanıcı verileri de dahil tweetler toplanıp, ön işlem aşamasında temizlenerek, el ile etiketlenip sonrasında öznitelik çıkarımı yapılmıştır. Bu işlem sonrasında oluşturulan eğitim setleri kullanılarak makine öğrenmesi algoritmaları ile modeller oluşturulmuştur. Elde edilen modeller değerlendirilerek en etkin algoritmanın tespit edilmesi amaçlanmıştır. Bu amaçla, farklı makine öğrenmesi algoritmaları ve farklı örneklem sayısına sahip eğitim setleri kullanılarak model çıkarımı yapılmıştır. Elde edilen modelin başarımı farklı ölçütler (Kappa İstatistiği, Doğruluk) göz önünde bulundurularak değerlendirilmiştir. Deneysel Sonuçlar bölümünde, Kappa istatistiği, doğruluk kavramları açıklanmış ve elde edilen sonuçlar verilmiştir.

BÖLÜM 2

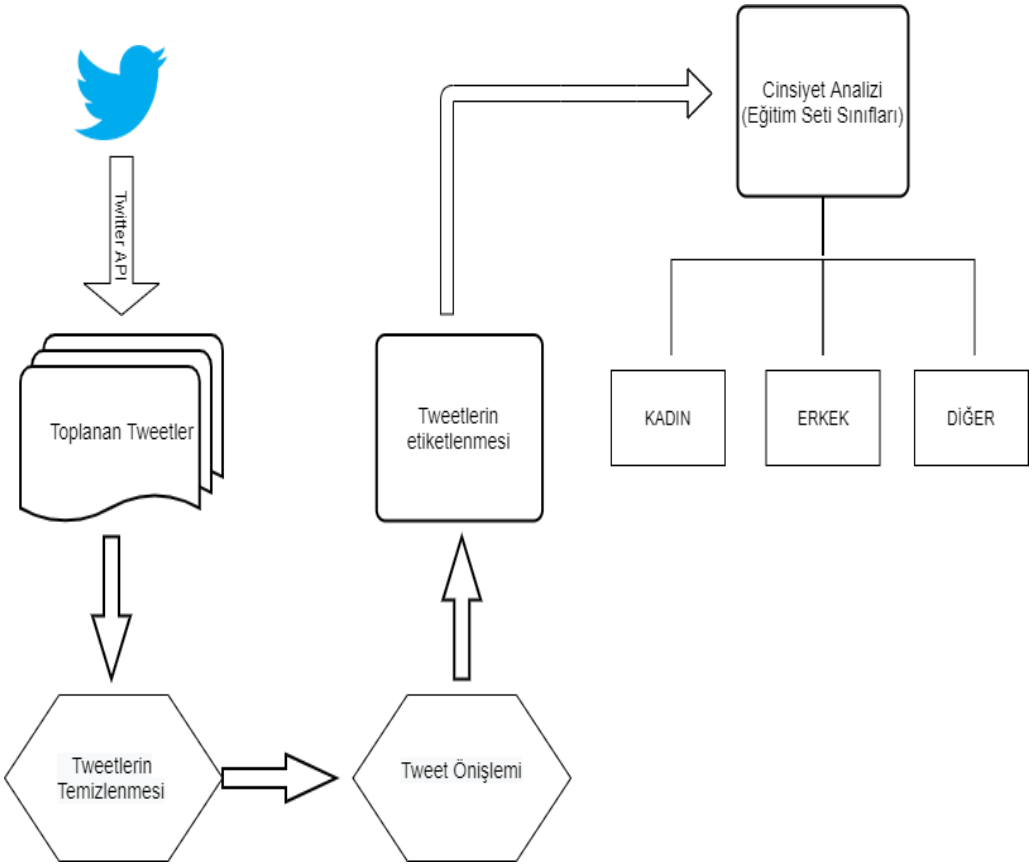
METODOLOJİ

2.1. Mimari Yapı

Bu çalışmada Twitter platformunda paylaşılan tweetler toplanarak cinsiyet ve tweet ilgi alanı tahmini yapılmıştır. Makine öğrenmesi algoritmaları için yapılan işlemlerde Weka yazılımından faydalanılmıştır. Çalışmada ilk aşama olan Twitter platformundan tweet toplama aşamasından başlayan ve son aşama olan model oluşturma arasında birden fazla süreç vardır. Birden fazla modülün birleşiminden oluşan mimari yapının diyagramı, Şekil 2.1'de gösterilmiştir.



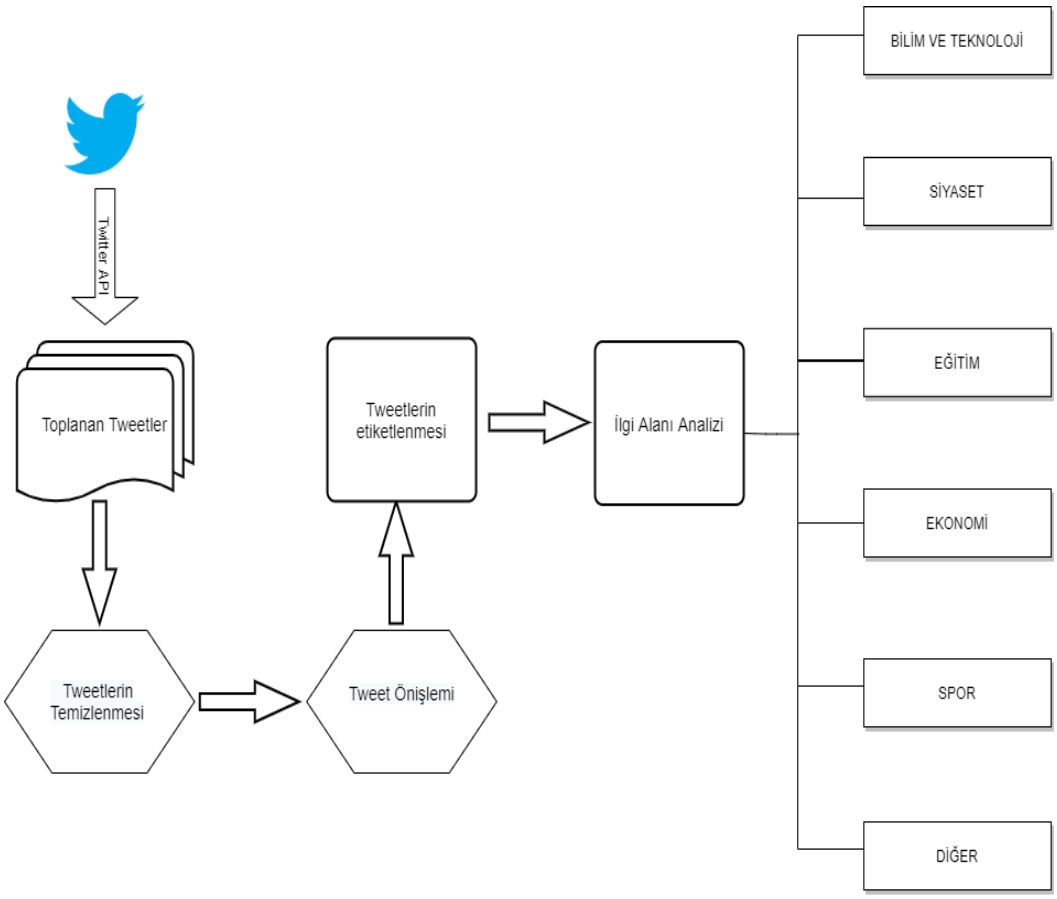
Şekil 2.1. Tweet toplama ve önişleme süreci mimari yapısı



Şekil 2.2. Cinsiyet analizi mimari yapısı

Twitter'ın sunduğu API kullanılarak tweetler toplanmış, bu tweetlerin temizlenmesi ve önişleminin ardından etiketleme işlemi yapılmıştır, sonuç olarak, cinsiyet analiz sonuçlarının kadın, erkek ve diğer olarak sınıflandırılması gerçekleştirilmiştir. Bu mimari yapı Şekil 2.2'de gösterilmiştir.

Şekil 2.3'de gösterilen blok diyagram incelendiğinde ise, Şekil 2.2'de bahsedilen aşamalardan benzer şekilde geçildikten sonra ilgi alanı analizi sonuçlarının; bilim ve teknoloji, siyaset, eğitim, spor ve diğer olduğu görülmektedir.



Şekil 2.3. İlgi alanı analizi mimari yapısı

2.2. Tweet Toplama

Tweet toplama, eğitim setlerinin oluşturulması için gereken tweetlerin toplanma aşamasıdır. Visual Studio ortamında C# programlama dili kullanılarak bir uygulama geliştirilmiş ve veriler, geliştirilen bu uygulama sayesinde toplanmıştır. Tweetlerin toplanması için Twitter'ın arayüzüne ulaşılması için izin alınması gerekmektedir. Bu amaçla Twitter API (Wood-Doughty, Andrews, Marvin, & Dredze, 2018) kullanılmıştır. API (Application Programming Interface), bir uygulamaya ait fonksiyonların, başka bir uygulamada da kullanılabilmesi için, bilgilerin paylaşıldığı, uygulamanın sağladığı bir arayüzdür. Bu, tweetlerin toplanabilmesi için yapılması gereken önemli bir aşamadır.

2.2.1. Tweet Toplama Modülü

Veriler, Visual Studio ortamında, Tweetinvi kütüphanesi kullanılarak toplanmıştır. Tweetinvi, C# programlama dili ile Twitter API'si kullanmak için hazırlanmış güncel bir kütüphane haline gelmiştir. Tweetinvi ile .NET C# programlama dilikullanarak kolayca etkileşime girebilir, tweet atabilir, mevcut tweetler alınabilmektedir.

```
1   System;
2   Tweetinvi;
3   Tweetinvi.Models;
4   Tweetinvi.Parameters;
5   System.Windows.Forms;
6   System.Configuration;
7   System.IO;
```

Şekil 2.4. Tweetinvi kütüphane kullanımı

Tweetinvi kütüphanesini C# dili ile geliştirilen uygulamaya dahil etmek için Şekil 2.4'de görüldüğü gibi, bu kütüphane dahil edebilir ve daha sonra da bu kütüphane içerisindeki oluşturulmuş tüm fonksiyonlar kullanılabilmektedir.

Twitter'dan istenilen veriyi özelleştirerek, kişi bunu kolayca kendi sistemine dahil edebilmektedir. Fakat bu kütüphanenin dahil edilmesi, direk olarak Twitter'ın arayüzüne bağlanılmasına yetmeyecektir. Bunun için kişiye özel olarak oluşturulan, eşsiz anahtarların sisteme dahil edilmesi gerekmektedir.

```
<add key="ConsumerKey" value="sdfsefNHD7884HKKD8jhdj8949"/>
<add key="ConsumerSecret" value="lıblmbdf6546654SDFG00sdf45454dfVDVDSsdfsg"/>
<add key="AccessToken" value="78787566468644545-Ukndnfhe63jklza"/>
<add key="AccessTokenSecret" value="toırmmmsw8785655FYQW645yklkvb"/>
```

Şekil 2.5. App.config dosyasında twitter api tanımlama

Twitter'ın kişiye sağlamış olduğu anahtarların kullanımı Şekil 2.5'te görüldüğü gibi App.config dosyasında kullanılmıştır. App.config, proje içerisinde konfigürasyonların tanımlandığı dosyadır. Twitter API'nın kişiye sağladığı Consumer

Key, Consumer Secret Key, Access Token, Access Token Secret eşsiz anahtarların ve karşılığında verilen değerlerin kullanımı “anahtar (key)” ve “değer (value)” isimli iki özelliğin içine gömüldükten sonra, *tweetinvi* kütüphanesinin kullanım aşaması tamamlandığı gibi, bu aşamanın da tamamlamasını sağlamıştır.

Bu işlem tamamlanmasına rağmen Twitter’ın arayüzüne bağlanılmasına yetmeyecektir. Çünkü anahtar ve değer, App.config’de kullanılmış ve yapılandırılmıştır. Bunun dışında, asıl ana sistemde bu değerleri çağırarak, doğrulama işlemi yapılmamıştır. Twitterdan tweetlerin çekilmesine yardımcı olacak ve tweetlerin sorunsuz bir şekilde çekilmesini sağlayacak fonsiyonların içinde kullanmadıktan sonra, sistem Twitter’dan istenilen şekilde veriler sağlamayacaktır. Şekil 2.6’de, Twitter API’ye C# programlama dili içerisinden erişimin nasıl yapıldığı gösterilmiştir.

```
private void TweetAlBtn_Click(object sender, EventArgs e)
{

    StreamWriter Dosya = File.CreateText("C:\\Users\\DELLULTRA\\Desktop\\tweet.txt");
    Dosya.Close();

    try
    {
        // Authentication
        Auth.SetUserCredentials(ConfigurationManager.AppSettings["ConsumerKey"],
                                ConfigurationManager.AppSettings["ConsumerSecret"],
                                ConfigurationManager.AppSettings["AccessToken"],
                                ConfigurationManager.AppSettings["AccessTokenSecret"]);
```

Şekil 2.6. Twitter api’ye c# programlama dili içinden erişim

Visual Studio aracılığı ile, C# programlama dili kullanarak, App.config içinde yapılandırılan hesap kullanıcısına özel olarak üretilen anahtarları, *Auth.SetUserCredentials* fonksiyonunun içinde kullanılan, *ConfigurationManager.AppSettings* aracılığı ile kullanılmıştır. Bu sayede, Twitter kullanıcısı için üretilen bu anahtarlara erişim, C# programlama dili içinden mümkün kılınmıştır. Bu işlemlerden sonra Twitter’dan istenilen tweetlerin sorunsuz olarak çekilmesi ve tweetinvi kütüphanesi ile aranılan tweetlerin alınmasında gerekli olan konfigürasyon yapılmış olur.

2.3. Tweetlerin Temizlenmesi

Twitter verilerinin temizlenmesi aşamasında, kullanıcıların paylaştığı tweetlerin içerisinde bulunan, sınıflandırma başarımının doğruluğunu olumlu yönde etkilemeyen fakat işleme dahil edildiğinde, başarımı düşüren bu fazladan olan Hashtag (#), Mentions (@), Link (http:\\…), etkisiz kelimeler gibi tweetlerin temizlenmesi işlemi yapılmıştır. Bu durum aşağıda verilen örnek bir tweet ile açıklanmıştır.

> *Galata kulesine gittiniz mi? @cmylmz #SoundCloud? https://soundcloud.com/cmylmz/galata*

Yukarıdaki örnekte görüldüğü üzere, atılan tweetin içerisinde, ana bilgiyi ayrıştırmak için; hashtag (#), mentions (@), link (http:\\....) gibi alanlar çıkarıldığı zaman tweet, aşağıdaki hale dönüşecektir:

> *Galata kulesine gittiniz mi?*

Bu temizleme işlemi, birçok yöntem ile yapılabilmektedir. Fakat bu çalışmada, Twitter tweetlerin temizlenme işlemi düzenli ifadeler (Regular Expression) ile yapılmıştır.

2.3.1. Düzenli İfadeler

Düzenli, kurallı ifadeler anlamına gelen karmaşık bilgi içerisinde, gerekli olan bilgilerin çekilmesi, kullanıcılar tarafından girilen bilgilerin denetimini ve verileri ihtiyaca uygun bir yapıya getiren bir yapıdır. Bu yapının daha iyi anlaşılması için aşağıdaki kod parçası ile bu durum daha ayrıntılı açıklanmıştır.

```
tweet = Regex.Replace(tweet, ",", " ");
```

Yukarıdaki örnekte görüldüğü gibi, bu düzenli ifadeler yapısı bir tweet içerisindeki tüm virgülleri, boşluk ile yer değiştirir. Yer değiştirme *Replace* metodu ile yapılmaktadır. Bu fonksiyon 3 argüman almaktadır. Bu alandan birincisi *tweet* verisidir. Bu parametre içeriğinde değişiklik yapmak istediğimiz veriyi içermektedir. İkinci alan olan “,” ifadesinin anlamı, tweet içerisinde, virgül karakteri ile ilgili bir işlem yapacağımız anlamına gelmektedir. Üçüncü alanda kullanılan “ ” ifadesi ise virgül

karakterini hangi ifade ile yer değiştirmek istediğimiz anlamına gelmektedir. Aşağıda gösterilen 3 madde ile, bir tweet içerisinde link, mention ve hashtag ifadelerinin düzenli ifadeler ile nasıl temizlendiği gösterilmektedir;

1. Link kısmının silinmesi, *Regex.Replace(tweet, @"https?://[-\w]+(\.\w[-]**\", " ");* kod parçası ile yapılmıştır.

2. Mention kısmının silinmesi, *Regex.Replace(tweet, @"\@\w+", " ");* kod parçası ile yapılmıştır.

3. Hashtag kısmının silinmesi, *Regex.Replace(tweet, @"#\w+", " ");* kod parçası ile yapılmıştır.

2.4. Öznitelik Çıkarımı

Cinsiyet eğitim seti için amaç, toplanan tweetler üzerinden, tweeti paylaşan kullanıcının yazım stiline bağlı olarak, cinsiyet analizini mümkün kılmaktadır. Eğitim seti; kadın, erkek ve diğer olarak 3 adet sınıftan oluşmaktadır. "*DİĞER*" olarak belirtilen sınıf; kurum, kuruluş, resmi sayfaların tweetlerinin olduğu sınıftır. Cinsiyet eğitim seti için Çizelge 2.1'de görüldüğü gibi, toplamda 50 adet öznitelik belirlenmiştir. İlgi alanı eğitim setinde; bilim ve teknoloji, siyaset, spor, eğitim, ekonomi ve diğer olmak üzere 6 sınıftan oluşmaktadır ve öznitelik olarak, paylaşılan "*tweet*" alınmıştır.

Çizelge 2.1. Cinsiyet eğitim seti için belirlenen öznitelikler

1- İsim alanı (erkek, kadın, diğer, erkek-kadın, kadın-erkek, erkek-sayı, kadın-sayı, diğer-sayı, erkek-emoji, kadın-emoji, diğer-emoji, erkek-noktalama, kadın-noktalama)	26- Ekran ismindeki sayı adedi
2- İsim	27- Ekran ismindeki sayı adedi aralığı
3- Ekran ismi	28- Ekran ismindeki büyük harf adedi
4- Tweet	29- Ekran ismindeki büyük harf adedi aralığı
5- Tweetteki kelime adedi	30- Ekran ismindeki küçük harf adedi
6- Tweetteki kelime adedi aralığı	31- Ekran ismindeki küçük harf adedi aralığı
7- Tweette tekrarlı kelime var mı?	32- Tweet uzunluğu
8- Tweetteki farklı kelime adedi	33- Tweet uzunluğu aralığı
9- Tweetteki sayı adedi	34- Kaynak
10- Tweetteki sayı adedi aralığı	35- Arkadaş sayısı
11- Tweetteki büyük harf adedi	36- Arkadaş sayısı aralığı
12- Tweetteki büyük harf adedi aralığı	37- İsim'deki emoji sayısı
13- Tweetteki küçük harf adedi	38- İsim'deki emoji sayısı aralığı
14- Tweetteki küçük harf adedi aralığı	39- Tweetteki emoji sayısı

15- Tweetteki hashtag adedi	40- Tweetteki emoji sayısı aralığı
16- Tweetteki hashtag adedi aralığı	41- Tweet hangi karakter ile başlamış? (büyük/küçük/rakam)
17- Tweetteki mention adedi	42- Noktalama adedi
18- Tweetteki mention adedi aralığı	43- Noktalama adedi aralığı
19- Tweet media (resim, video) içeriyor mu (var/yok)?	44- Büyük kelime adedi
20- İsim'deki sayı adedi	45- Büyük kelime adedi aralığı
21- İsim'deki sayı adedi aralığı	46- Tweette büyük harfle başlayan kelime sayısı
22- İsim'deki büyük harf adedi	47- Tweette büyük harfle başlayan kelime sayısı aralığı
23- İsim'deki büyük harf adedi aralığı	48- Tweette tekrarlı harf sayısı
24- İsim'deki küçük harf adedi	49- Tweette etkisiz kelime sayısı
25- İsim'deki küçük harf adedi aralığı	50- Tweette etkisiz kelime sayısı aralığı

2.5. Tweetlerin Etiketlenmesi

Tweetlerin etiketlenmesi aşaması bu çalışmanın önemli bir sürecini oluşturmaktadır. Bu aşamada tüm tweetler insan tarafından okunup etiketleme işlemi yapılmıştır.

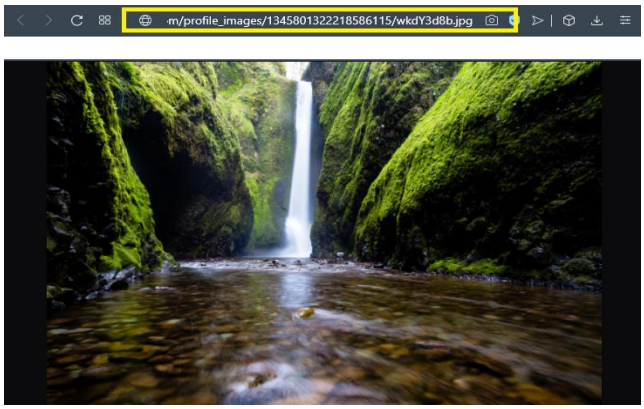
DİĞER	http://pbs.twimg.com/profile_images/1052180246999572480/p3lni64d.jpg	diğer-sayı	habercim19@	habercim19	çorum'da 3 dönen
ERKEK	http://pbs.twimg.com/profile_images/1023030128891314176/H-swste_.jpg	erkek-sayı	adil ciddi 0.19	semboluludivan1	önceki sağ iktidarl
KADIN	http://pbs.twimg.com/profile_images/1040336003905540096/O0y7I-pR.jpg	kadın	Şeyda Aksoy	MaviMorpho	en geç güne sunu
KADIN	http://pbs.twimg.com/profile_images/1052629533566914561/PFf9YgQv.jpg	kadın-noktalama	EMEL <emoji> (◠‿◠)	EmeL_GRT_	iyi parti: chp'nin iş
KADIN	http://pbs.twimg.com/profile_images/1043344836223533061/qH-BdGHI.jpg	kadın	Zeynep Karataş	zeyneprkaratas	haber gün milliyet
DİĞER	http://pbs.twimg.com/profile_images/926719904321101829/h4Pe91k7.jpg	diğer	dielectricspor	dielectricspor	mike dean yönetiy
KADIN	http://pbs.twimg.com/profile_images/898580850752802817/ugrqFpgB.jpg	kadın-noktalama	❁Mehmed (Mihêê)	mehmedak63	bilim aşkın gözü k
KADIN	http://pbs.twimg.com/profile_images/1053982742076145664/cJmoV7Nr.jpg	kadın-emoji	Ela <emoji>	bencebarlas	anatomi al canımı
DİĞER	http://pbs.twimg.com/profile_images/1023612261489934342/qMo_2jIF.jpg	diğer-sayı	haber1001.com	haber1001com	cumhurbaşkanı er
KADIN	http://pbs.twimg.com/profile_images/995038195611066368/CHGEHbcZ.jpg	kadın	Leyla Tutal	tibbi_sekreter0	tibbi sekreter atan
KADIN	http://pbs.twimg.com/profile_images/2735906839/54f26749cfdec2e0a047415cae9448d1.png	kadın-emoji	<emoji> YAŞANABİLİR BİR TÜRKİYE	fatihsunar	sahi 24 haziranda
DİĞER	http://pbs.twimg.com/profile_images/1036153438000828416/Um_DX6I0.jpg	diğer	enBursa	enBursaHaber	cumhurbaşkanı er
KADIN	http://pbs.twimg.com/profile_images/1027806934580256768/zAJ6JsIq.jpg	kadın	Feyza	guneysakini	fakülte kantinin d
DİĞER	http://abs.twimg.com/sticky/default_profile_images/default_profile.png	diğer-noktalama	siyasetcafe.com	siyasetcafecom	mhp'li enginyurt't
ERKEK	http://pbs.twimg.com/profile_images/933812210643632128/rigqWXXE.jpg	erkek	Cem BAKIRCI	CemBAKIRCI	temizlik işçisi çöpt
ERKEK	http://pbs.twimg.com/profile_images/2810389890/4257015b83e4b02788ad2e7e7795e6cf.png	erkek	İzzet Can Ljajic	IzzetCanArvas	1- hoca yönetim g
DİĞER	http://pbs.twimg.com/profile_images/819602153614544896/mm8G90im.jpg	diğer-sayı	Ankaradanhaber1	Ankaradanhaber1	istanbulda 259 üy
ERKEK	http://pbs.twimg.com/profile_images/954814095554957314/l-5C_d8E.jpg	erkek-noktalama	• Muhammet ÖZEN	muhammetozen_	indirimle kasım a
ERKEK	http://pbs.twimg.com/profile_images/1042793148295143424/fqY_ckFC.jpg	erkek	murat topal	murattopal28	sene rahat maç i
ERKEK	http://pbs.twimg.com/profile_images/936169579402944512/jo8LwRmY.jpg	erkek-noktalama	M.Hakan USCA	HakanUSCA	bankaları kapatırın
KADIN	http://pbs.twimg.com/profile_images/888847482532356096/YabNkHp8.jpg	kadın	Suna İzol	SunaIzol	paramedikler atan
KADIN	http://pbs.twimg.com/profile_images/1048000545187876865/jXMgdV6r.jpg	kadın-noktalama	EMEL <emoji> (◠‿◠)	EmeL_GRT_	yunanistan'daki ga
ERKEK	http://pbs.twimg.com/profile_images/988756273003945985/7zleYpM4.jpg	erkek	Turan Dursun	turandursun06	mhp siyasi metres

Şekil 2.7. Toplanan tweet verisinin bazı alanları

Etiketleme işlemi yapılırken Şekil 2.7'de görüldüğü gibi birçok farklı öznitelik ile işlem yapılmıştır. Bu özniteliklerden ilki olan sınıf özniteliği için, geliştirilen bir uygulama ile kullanıcı ve ekran ismine dayalı otomatik olarak kural tabanlı ön bir etiketleme işlemi yapılmıştır. Bu etiket bu ön etiketleme işlemi sonrası, el ile etiketleme

aşamasında insan tarafından kontrol edilerek, gerekli ise düzeltilmiştir. Bu kontrol işleminin yapılması zorunludur. Bunun nedeni etiketlemeye kolaylık sağlaması için kural tabanlı yapılan bu ön etiketlemenin başarısız olması ve bunun eğitim setinin kalitesini düşürmesidir. Kullanıcı ve ekran ismine dayalı, kural tabanlı ön etiketleme işleminin başarısız olması beklenen bir durumdur. Kural tabanlı yapılan bu ön etiketlemenin gözlemlenen başarısız olma nedenlerinden bazıları; bazen Twitter kullanıcıları, kullanıcı isimlerine cinsiyetleri ile bağdaşmayan farklı bir isim verebilmektedir. Benzer şekilde kullanıcı, bir ticari markaya ait bir kullanıcı olabilmektedir. Ayrıca bazen, kullanıcı isminin cinsiyet hakkında bir veri barındırmaması ve cinsiyeti belirlemeye yönelik bir ekran ismine sahip olmaması söylenebilir.

Etiketleme aşamasında kullanıcının cinsiyeti hakkında tam olarak fikir edinebilmek için, bu aşamada yapılan bir diğer işlem, json dosyasından aldığımız profile_image alanında görülen, profil resminin URL adresini Şekil 2.8'de gösterildiği gibi adres çubuğuna yazmaktır.



Şekil 2.8. Profil linkinin aratılmasından sonra görünen resim

Etiketleme işleminin zaman alması, manuel olarak bu işlemin yapılmasından kaynaklanmaktadır. Eğer verilerimiz metin değil, sayısal değerlerden oluşan veriler olsaydı o zaman sayısal bazlı verilerde otomatik etiketleme yapan algoritmalar ile bu veriler etiketlebilirdi. Bu işlem hem zamandan tasarruf hem de daha fazla veri işlenmesine fayda sağlardı. Ancak veriler sayısal olmadığı için URL üzerinden insan kontrolü ile daha yüksek kesinlikte etiketleme yoluna gedilmiştir. Bunun için URL

adresinin arama motorunda bulunan adres çubuğuna yazılmasıyla, kullanıcının ya da markanın fotoğrafının görülmesi ile cinsiyeti etiketlemeye yönelik önemli bir görsel bilgiye ulaşılmıştır. Bu aşamada ekrana gelen fotoğraf eğer erkek fotoğrafı ise sınıf etiketi olarak erkek, kadın fotoğrafı ise kadın, başka bir görsel ise diğer olarak etiketlenmiştir.

Kural tabanlı ön etiketleme sürecinde, elimizde mevcut kadın ve erkek isimlerinin bulunduğu listedeki isimler (İsimler Listesi, 2011) ile kullanıcı ve ekran isimleri düzenli ifadeler kullanılarak karşılaştırılmış ve eşleşme durumuna otomatik etiketlenmiştir. Bu etiketlendikten sonra, bu etiketi doğrulamak için belirtilen URL üzerinden kontolü yapılmıştır. Eğer marka logosu, manzara resmi, araç fotoğrafı ya da herhangi bir nesne fotoğrafı ile karşılaşılmış ise bunlar diğer olarak etikelenmiştir. Eğer isimlerde kadın veya erkek ismi geçiyorsa, bunlar ismin verildiği cinsiyet baz alınarak etiketlenmiştir. Ayrıca bazı hesapların ekran ve kullanıcı isimleri mevcut karşılaştırma listesinde olmasına rağmen URL üzerinden yapılan kontrolde kullanıcının kendi fotoğrafı yerine tanınan bir kişiliğin fotoğrafı olduğuda görülmüştür. Bu durumda isme bağlı kalınarak cinsiyet etiketi verilmiştir. Kural tabanlı otomatik ön etiketleme için kullanılan örnek kaynak C# programlama dili kod parçası aşağıda verilmiştir:

Aşağıda gösterilen kod parçası, düzenli ifade içinde verilen ismi barındıran tweetler, ön etiketlemede erkek olarak etikelenmektedir.

```
if (Regex.IsMatch(name_field, @"^(?i)\b(?:avni|ergün|huso|tarkan|refik|özhan|...)))
{
        file.Write("ERKEK" + ",");
}
```

Erkek sınıfı için, düzenli ifade içinde verilen isim ile uyuşma olmaması durumunda, program aşağıda belirtilen kadın isimlerini kontrol etmek için, düzenli ifade içinde verilen isimleri aşağıda gösterildiği gibi kontrol etmeye başlayacaktır.

```
else if (Regex.IsMatch(name_field, @"^(?i)\b(?:yurdagül|Adelina|Seyhan|Olga|...)))
{
        file.Write("KADIN" + ",");
}
```

Benzer şekilde, kadın sınıfı için uyuşma olmadığı takdirde, düzenli ifade içinde verilen ismi barındıran tweetler, ön etiketlemede aşağıda gösterildiği gibi diğer sınıfı olarak etikelenmektedir.

```
else(Regex.IsMatch(name_field, @"(?i)\bYeni Bildirim|yedi24|Webtekno|Webrazzi|...)))
{
        file.Write("DİĞER" + ",");

}
```

2.6. Eğitim Setleri

Bu çalışmada, Twitter ortamında toplanan tweetlerin, eğitim seti olarak kullanılabilmesi için, tweet temizleme ve öznitelik belirleme ön işlemleri yapılmıştır. Eğitim seti oluşturulması aşamasından sonra, WEKA yazılım aracı ile makine öğrenmesi algoritmalarının denemeleri yapıldıktan sonra, elde edilen modeller test edilmiş ve sonuçlar değerlendirilmiştir. Cinsiyet ile ilgili toplamda 171.200 adet tweet, ilgi alanı ile ilgili toplamda 90.000 adet eğitim seti örneği toplanmış olup, Çizelge 2.2 ve Çizelge 2.3'de ise, toplanan tweetlerin sınıflara göre örnek sayıları gösterilmiştir.

Çizelge 2.2. Toplanan cinsiyet eğitim seti örnek adedi

Sınıf	Tweet Adedi
KADIN	54.370
ERKEK	48.985
DİĞER	67.845

Çizelge 2.3. Toplanan ilgi alanı eğitim seti örnek adedi

Sınıf	Tweet Adedi
SİYASET	15.000
SPOR	15.000
EĞİTİM	15.000
EKONOMİ	15.000
BİLİM VE TEKNOLOJİ	15.000
DİĞER	15.000

Eğitim setleri, kullanılmadan önce tweetlerin temizlenmesi sürecinde; tweet içerisindeki hashtag, etkisiz kelimeler, mention ve linklerin temizlenmesi işlemi gerçekleştirilmiştir. Tweet temizleme ön işleminin yapılması, sınıflandırma başarımının doğruluğunu arttırmak amacıyla yapılmıştır.

Cinsiyet eğitim setinin ön etiketlenmesi aşamasında, mevcut bilinenen isimlerden oluşan bir isim listesi kullanılmış ve buna göre ön etiketleme işlemi yapılmıştır. Toplamda 2.400 adet erkek, kadın ve diğer (kurum ve kuruluş vb.) isimleri bu listede bulunmaktadır. Geliştirilen uygulama ile verilerin toplandığı aşamada, birincil olarak kullanıcı ismi ve ikincil olarak ekran ismi, bu isim listesi içinde aranmıştır eğer bunlardan biri listedeki isimlerden biri ile uyuşuyorsa; erkek, kadın veya diğer olacak şekilde otomatik olarak ön etiketlemesi yapılmıştır. Ardından, kural tabanlı etiketlenen bu veriler tek tek incelenmiş ve doğruluğunun el ile kontrolü yapılarak, bu hesaplar için tek tek nihai sınıf etiketine karar verilmiştir.

İlgi alanı eğitim seti için; herbir alan için önceden belirlenmiş anahtar kelimeler yazılmış ve bu şekilde tweetler toplanmıştır. Eğitim seti içerisinde, verinin daha sade ve anlaşılabilir olması için, cümle yapısını bozan hashtag'ler, mention'lar, media linkleri ve etkisiz kelimeler temizlenmiştir. Bu işlemin yapılmasının amacı, makine öğrenmesi algoritmalarının sınıflandırma başarımını olumsuz etkilememesi içindir. İlgi alanı eğitim seti için; Siyaset, Bilim ve Teknoloji, Spor, Eğitim, Ekonomi ve Diğer olmak

üzere toplamda 6 adet sınıf kullanılmıştır. "*Diğer"* olarak kullanılan sınıfın içerisinde Siyaset, Bilim ve Teknoloji, Spor, Eğitim, Ekonomi alanları ile bir ilgisi olmayan, anahtar kelimeler kullanılıp tweetler toplanmıştır. Benzer şekilde cinsiyet eğitim seti etiketlemesindeki gibi ilgi alanı eğitim seti için kullanılan veriler de tek tek incelenmiş ve doğruluğunun el ile kontrolü yapılarak nihai sınıf etiketine karar verilmiştir. Tweet toplama sürecinde, toplanan bu tweetlerin nispeten istenen alan ile ilgili olmasını sağlamak amacıyla kullanılan anahtar kelimeler, sınıf bazlı olarak Çizelge 2.4'te verilmiştir.

Çizelge 2.4. İlgi alanı anahtar kelimeler

Sınıf	Anahtar Kelimeler
EKONOMİ	hisse senedi, ekonomi, sermaye, mali gelir, vadeli fiyat, vadeli hesap, dolar, vergi, cari kur, cari işlemler hesabı, enflasyon, dalgalı kur, deflasyon, döviz kuru, ekonomik güven endeksi, faiz oranı, finans, finansal, gayri safi milli hasıla, gsmh, hazine bonosu, ihracat, iskonto, ithalat, konut fiyat, konut kira, kredi riski, likidite, merkez bankası, para kurulu, para piyasası, para politikası
EĞİTİM	Yüksek öğretim kurumu, ösym, ösys, dgs, lisans eğitimi, önlisans eğitimi, ortaöğretim, ilköğretim, milli eğitim bakanı, milli eğitim, öğretmen ataması, öğretmen atamaları, atama, yüksek lisans, fakülte, yüksekokul, meslek yüksekokulu, örgün öğretim, ikinci öğretim, açıköğretim, devlet üniversitesi
SİYASET	Anayasa, kararname, milletvekili, tutanak, siyaset, bürokrasi, cumhurbaşkanı, politika, akp, chp, mhp, sayıştay, danıştay, parlamenter, demokrat, sosyalizm, radikalizm, laik, parlamento, emperyalist, referandum, sosyalist, radikalist, tbmm
BİLİM VE TEKNOLOJİ	anakart, işlemci, ram, bilişim, fiber, siber, veri tabanı, HDMI, algoritma, BIOS, IOS, ios, linux, root, yazılım, endüstri 4.0, inovasyon, sanal gerçeklik, yapay zeka, çözelti, elektrot, elektroliz, elektrolit, elektron, atom

SPOR	Asist, penaltı, defans, smaç, faul, şut, hakem, gol, kaleci, korner, rövaşata, UEFA, lig, ofsayt, şut, steps, ribaund, deplasman, turnuva, fair play, antrenör, hazırlık maçı, hat trick, oyun kurucu, galatasaray, fenerbahçe, beşiktaş
DİĞER	Pasaport, akrostiş, alafranga, argo, asalet, betimleme, biyografi, fıkra, gazel, kafiye, kaside, şiir, mahlas, masal, roman, satranç, şarkı, akustik, magazin, acun, altın kelebek, demet akalın, ahmet kural, sıla, Seyahat Acentası, Konaklama, Rezervasyon, resepsiyon, suit oda, tek kişilik oda, film, dizi, cilt bakımı, hava durumu, namaz, tik tok, müzik, şarkı, yılbaşı, necip fazıl, kim milyoner olmak ister, namık kemal, flört, fragman

BÖLÜM 3

MAKİNE ÖĞRENMESİ

3.1. Makine Öğrenmesi Nedir?

Makine öğrenmesi, 1959 yılında bilgisayar biliminin yapay zekâ da sayısal öğrenme ve model tanıma çalışmalarından geliştirilmiş bir alt dalıdır. Makine öğrenmesi, yapısal işlev olarak öğrenebilen ve veriler üzerinden tahmin yapabilen algoritmaların, çalışma ve inşalarını araştıran bir sistemdir (Edgar, & Manz, 2017). Bu tür algoritmalar statik program komutlarını harfiyen takip etmek yerine, örnek girişlerden, veri tabanlı tahminleri ve kararları gerçekleştirebilmek amacıyla bir model inşa ederek çalışmaktadırlar. Arthur Samuel'e göre makine öğrenmesi, bilgisayarları programlamadan onlara öğrenme becerisini öğreten çalışma alanıdır (Simon, 2013). Bishop'a göre ise makine öğrenmesindeki amaç, geçmiş tecrübelerden öğrenerek, gelecek veriler için genelleme yapmaktır (Bishop, 2006). Makine öğrenmesi, günlük hayatta birçok alanda kullanılmaktadır. Kötü amaçlı e-postaların tanınması, konuşmaların tanınması ve işlenmesi, ürün tavsiye etme, yüz tanıma gibi alanlar makine öğrenmesinin kullanıldığı alanlardan bazılarıdır.

Makine öğrenmesinin uygulanabilmesi için eğitim setindeki örneklerin temizlenmesi ve ön işlemden geçmesi gerekmektedir. Ön işlemden geçtikten sonra, eğitim setindeki örnekler makine öğrenmesi algoritmalarına uygun hale gelmektedir. Makine öğrenmesi, denetimli ve denetimsiz öğrenme olmak üzere iki ana öğrenme türünden oluşmaktadır.

Denetimli Öğrenme: Denetimli öğrenme, sınıflandırılmış eğitim setinden oluşmaktadır. Bunun anlamı, girdi olarak verilen verinin hangi amaç için kullanıldığı bellidir. Denetimli öğrenme, tahmin modelleri geliştirmek için iki tür teknik kullanmaktadır. Sınıflandırma teknikleri ve Regresyon teknikleri. Eğitim seti, etiketlenebilir, kategorilere ayrıabiliyorsa denetimli öğrenme tekniklerinden olan sınıflandırma tekniği kullanılabilir. Regresyon tekniklerinde ise, sürekli olarak devam eden bir değerin tahmin edilmesi yapısına dayanmaktadır. Örneğin, arabanın fiyatı, evin fiyatı vb.

Denetimsiz Öğrenme: Denetimsiz öğrenme, denetimli öğrenmenin aksine sınıflandırılmamış eğitim setinden oluşmaktadır. Verilerin, bilgiyi keşfetmesi için kendi kendilerinde öğrenmesine dayalıdır. Denetimsiz öğrenme, denetimli öğrenmeye kıyasla daha karmaşık işlemlerin gerçekleştirilmesini sağlar ve sisteme sınıflandırma etiketleri verilmediği için verilerden bilinmeyen bilgiler çıkarılması beklenir. Denetimsiz öğrenme tekniklerinden olan kümeleme, kategorilere ayrılmamış eğtim setlerinden model oluşturmak için kullanılan bir tekniktir.

3.2. Sınıflandırma

Makine öğrenmesi algoritmaları ile oluşturulan modellerin, yeni gelen veriler için, bu modeldeki sınıflardan hangisine ait olduğunu belirlemeye sınıflandırma denilmektedir (Kotsiantis, Zaharakis, & Pintelas, 2007). Sınıflandırma, denetimli öğrenme tekniklerinden birisidir. Sınıflandırmanın amacı, bir eğitim setinde tanımlı olan sınıflar arasında, yeni gelen veriyi doğru sınıfa yerleştirmektir. Makine öğrenmesinde, sınıflandırma ile ilgili birçok terim yer almaktadır. Bunlardan bazıları aşağıda açıklanmıştır:

Sınıflandırıcı (Classifier): Giriş verilerini belirli bir kategoriye eşlemek için kullanılan bir algoritmadır. Örneğin, e-postaları tarayıp bunların zararlı veya zararsız olarak sınıflandırıcı algoritması ile sınıflandırılabilir.

Sınıflandırma Modeli (Classification Model): Verilerin sınıfını veya kategorisini tahmin etmektedir.

Öznitelik (Feature): Öznitelik, gözlemlenen olgunun, ayrı ayrı ölçülebilir bir özelliğidir. Aynı zamanda, eğitim seti içerisinde bulunan her bir alandır da denilebilmektedir. Bölüm 2.4'te gösterildiği üzere, bu çalışmada cinsiyet eğitim seti için toplamda 50 adet öznitelik çıkarımı yapılmıştır.

İkili Sınıflandırma (Binary Classification): İki sonucu olan bir denetimli makine öğrenmesi algoritmasıdır. Örneğin, Twitter'da duygu analizi yapıldığı zaman sınıflar olumlu tweetler için pozitif veya olumsuz tweetler için negatif olarak ayrılabilmektedir. Benzer şekilde, görüntü işleme eğitim setlerinde, görüntüdekilerin insan olup olmadıkları sınıflandırılmak isteniyorsa, eğer insan ise evet, fakat insan değil ise hayır olarak etiketlenmektedir.

Çok Sınıflı Sınıflandırma (Multi Class Classification): İkiden fazla sınıflı sınıflandırma, çok sınıflı sınıflandırmada her bir veriye bir etiket atanmaktadır. Örnek olarak, yapılan bu çalışmada, cinsiyet ve ilgi alanı analizinde kullanılan eğitim setlerinde, çok sınıflı sınıflandırma kullanılmıştır. Cinsiyet analizi için kadın, erkek ve diğer, ilgi alanı analizinde ise bilim ve teknoloji, spor, siyaset, eğitim, ekonomi ve diğer olarak çoklu sınıflardan oluşmaktadır.

Değerlendirmek (Evaluate): Bir modeli değerlendirmek, etkili bir makine öğrenimi modeli oluşturmanın temel bir parçasıdır. Genel olarak, modelin değerlendirilmesi anlamına gelir. Aynı zamanda, bir modeli eğitildiğinden farklı veriler ile test yapmak anlamına da gelmektedir.

3.3. Makine Öğrenmesi Algoritmaları

Makine öğrenimi algoritmaları, daha fazla veriye maruz kaldıkça, daha iyi performans gösterecek şekilde kendilerini ayarlayan programlardır. Makine öğreniminin "öğrenme" kısmı, bu programların verileri nasıl işlediklerini, tıpkı insanlar öğrenerek verileri işleme şeklini değiştirdikçe, zaman içinde değiştirdiği anlamına gelmektedir. Bundan dolayı, bir makine öğrenimi algoritması, bir eğitim seti hakkında tahminlerde

bulunurken önceki performansı hakkında geri bildirim verilen, kendi parametrelerini ayarlamanın belirli bir yoluna sahip bir programdır.

Literatürde birçok sınflandırma algoritması vardır. Fakat bu çalışmada 5 adet sınıflandırma algoritması kullanılmıştır. Bunlar: KNN (En yakın Komşu- k nearest neigborhood), Naive Bayes, C4.5 karar ağacı, SVM (Destek Vektör Makinesi- Support Vector Machine), SMO (Ardışık minimal optimizasyon- Sequential minimal optimization) algoritmalarıdır.

3.3.1. NAIVE BAYES Algoritması

Naive Bayes, tahmine dayalı sınıflandırma için basit fakat güçlü bir makine öğrenmesi algoritmasıdır. Bu algoritma, ismini Matematikçi Thomas Bayes'den almıştır. Naive Bayes modelinin oluşturulması kolaydır, karmaşık veri tahmini yapmadan, çok büyük veri kümeleri için özellikle yararlı olmaktadır. Naive Bayes makine öğrenmesi algoritması çoğu zaman, karmaşık algoritmalara karşılaştırılınca iyi sonuç verdiği ve kolay uygulanabildiği için yaygın olarak kullanılmaktadır. Naive Bayes'in formüle edilmiş hali olan Bayes Teoremi, denklem 3.0'daki gibidir. Naive Bayes makine öğrenmesi algoritmasının popüler kullanımları arasında spam filtreleri, metin analizi vb. yer almaktadır. Bu sınıflandırıcılar, makine öğrenmesi için yaygın olarak kullanılmaktadır, çünkü uygulanması kolaydır.

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{3.0}$$

P(A|B) = B verisinin, A sınıfında olma olasılığı
P(A) = A sınıfının tüm veride geçme olasılığı
P(B|A) = A sınıfının kaç adet B verisinde geçtiğinin olasılığı
P(B) = B verisinin tüm veride geçme olasılığı

Denklem 3.0'a göre Naive bayes makine öğrenmesi algoritması, netice olarak hedef sınıfın hangi değerinin gerçekleşme olasılığının ne olduğunu bildirir.

3.3.2. KNN Algoritması

KNN makine öğrenmesi, verileri benzerlik ölçütüne göre (örneğin, mesafe) sınıflandıran basit bir makine öğrenmesi algoritmasıdır. KNN makine öğrenmesi algoritması, 1970'lerin başında, istatistiksel tahmin ve örüntü tanıma işlemlerinde kullanılmıştır. Sınıflandırma yapılırken herbir verinin diğer verilere olan uzaklığı hesaplanır. Fakat, bir veri için diğer verilerden sadece k adedi gözönüne alınmaktadır. Bu k adet veri, mesafesi hesaplanan noktaya, diğer verilerden daha yakın olan verilerdir.

KNN Parametreleri:

Öklid Mesafesi: Öklid mesafesi, iki noktanın birbirine olan doğrusal mesafedir. Denklem 3.1'de öklid mesafenin formülü görülmektedir.

$$\sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2} \tag{3.1}$$

Komşu Adedi (k): KNN, k parametrenin değerine bakarak sınıflandırma yapmaktadır.

Algoritmanın Çalışma Adımları:

1. K değeri belirlenir
2. Öklid uzaklıkları tespit edilir
3. En yakın komşular belirlenir
4. En yakın komşu kategorileri toplanır
5. En uygun komşu kategorisi seçilir.

Sonuç olarak KNN makine öğrenmesi algoritması, en basit sınıflandırma algoritmalarından biridir. Basit olmasının yanında, yukarıda belirtilen parametreler sayesinde oldukça rekabetçi sonuçlar verebilmektedir. KNN makine öğrenmesi algoritması, denetimli öğrenme tekniklerinden olan regresyon problemleri için de kullanılabilmektedir.

3.3.3. SVM Algoritması

SVM, sınıflandırma ve regresyon analizi için verileri analiz eden makine öğrenme algoritmasıdır. SVM makine öğrenmesi algoritması, verilere bakan ve bunları iki kategoriden birine ayıran denetimli bir öğrenme yöntemidir. Bir SVM makine öğrenmesi algoritması, sıralanan verilerin bir haritasını mümkün olduğunca birbirinin arasındaki boşluklarla göstermektedir. SVM algoritmalarının genel olarak metin kategorizasyonu, görüntü sınıflandırması, el yazısı tanıma gibi alanlarda kullanılmaktadır. SVM makine öğrenmesi algoritması daha çok derin öğrenme alanında yaygın olarak kullanılmaktadır (Giachanou, 2016).

Avantajları:

- Veriler hakkında hiçbir fikrimiz olmadığında SVM'ler çok iyi sonuç vermektedir.
- Metin, Görüntüler ve karar ağaçları gibi yapılandırılmamış ve yarı yapılandırılmış verilerle iyi çalışmaktadır.
- Yüksek boyutlu verilerde çok iyi sonuçlar verebilmektedir.
- SVM makine öğrenmesi algoritması her zaman yapay sinir ağları ile kıyaslanmaktadır. Yapay sinir ağları modelleriyle karşılaştırıldığında, SVM makine öğrenmesi algoritmaları daha iyi sonuç vermektedir.

Dezavantajları:

- Büyük eğitim setleri için uzun eğitim süresi gerekmektedir.
- Çok sınıflı SVM makine öğrenmesi algoritmaları için en uygun tasarımı bulmak zordur.
- SVM makine öğrenmesi algoritması her ne kadar iyi olsa da test aşamasında çok yavaş olabilmektedir.

3.3.4. SMO Algoritması

SMO makine öğrenmesi algoritması 1998'de John C. Platt tarafından önerildi. Özellikle doğrusal SVM ve seyrek veriler için iyi performans göstermiş ve en hızlı ikinci dereceden programlama optimizasyonu olmuştur.

SMO, karesel programlama (QP) problemlerini çözmek için kullanılmaktadır. QP sorunlarının olduğu bir örnek, SVM eğitim süreci sırasındadır. SMO algoritması bu örnekte bir kısıtlama optimizasyon problemini çözmek için kullanılmaktadır. SVM, eğitim süreci özellikle büyük veriler için yavaştır. Yavaş olmanın sebeplerinden biri, son derece büyük bir QP optimizasyon probleminin çözümünü gerektirmesidir. N veri noktaları göz önüne alındığında ve N çok büyük bir sayı ise, gerçek büyük bir veri sorunu olmaktadır. SMO büyük miktarlarda veri kümesiyle başa çıkmak için bir yaklaşımdır. SMO algoritmasını uygulayan en ünlü kütüphanelerden biri LibSVM kütüphanesidir.

3.3.5. C4.5 Algoritması

Karar ağacı, sınıflandırma tekniklerinden biridir. Bölme kriterleri ile yapılmaktadır. Karar ağacı örnekleri sıralayarak sınıflandıran bir ağaç yapısı gibi öznitelik değerine göre değerlendiren bir akış şemasıdır. Bir karar ağacındaki her düğüm, sınıflandırılacak bir örnekte bir özelliği temsil etmektedir. Tüm yapraklar, testin sonucunu belirtir. Her bir yaprak düğümü, sınıf etiketini tutmaktadır. Örnekler, öznitelik değerlerini baz alarak başlangıçtan itibaren sınıflandırılmaktadır. Karar ağacı, veri kümesinin sınıflandırılması için bir kural oluşturmaktadır. ID3, C4.5 karar ağacı ve CART olan üç temel algoritma yaygın olarak kullanılmaktadır (Horton, & Nakai, 1996).

C4.5, entropi ile değerlendirilen bilgi kazanma oranına dayanmaktadır. Entropi, işlenen bilgilerdeki rastgeleliğin bir ölçüsüdür. Entropi ne kadar yüksek olursa, o bilgiden sonuç çıkarmak o kadar zor olmaktadır. Bilgi kazanma oranı ölçüsü, ağaçtaki her düğümdeki test öznitelik seçmek için kullanılmaktadır. Böyle bir ölçü, öznitelik seçim ölçüsü olarak adlandırılmaktadır. En yüksek bilgi kazanma oranına sahip öznitelik, mevcut düğüm için test özelliği olarak seçilmektedir.

BÖLÜM 4

CİNSİYET VE İLGİ ANALİZİ İÇİN YAPILAN ÇALIŞMA

4.1. Geliştirilen Uygulama

Cinsiyet ve ilgi alanı tahmini için yapılan bu çalışmada Twitter'ın sunduğu API kullanılmıştır. Tweet analizi yapmak amacı ile kullanıcı tweetlerini toplayan sistemler, genel olarak arayüzleri bulunmayan arkaplan uygulamaları olarak tasarlanır. Bu çalışma için geliştirilen uygulamanın amacı, tweetlerin toplanması ve toplanan tweetlere ait verilerden görsel gösterimler elde etmektir. Bu sayede, yapılacak analizlerinin kolay ve hızlı bir şekilde yapılması sağlanmıştır. Tweet toplama ve görselleştirme amacıyla geliştirilen uygulama, Visual Studio ortamında, C# programlama dili kullanılarak geliştirilmiştir. Şekil 4.1'de geliştirilen uygulama arayüzü ile cinsiyet tahmini eğitim seti için tweet toplama ve ön etiketleme süreci uygulama arayüzü görülmektedir.



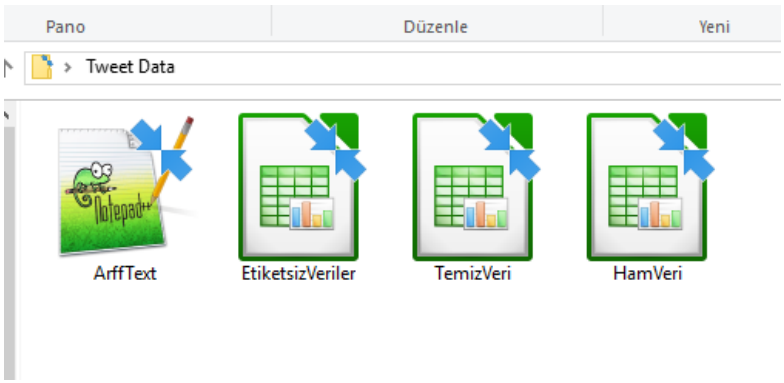
Şekil 4.1. Tweet toplamak için geliştirilen uygulama arayüzü

Gerekli olan anahtar kelimeler, metin kutusuna yazılıp, “Tweet Topla” butonuna basıldığı andan itibaren, yazılan anahtar kelimeleri içeren tweetler toplanmaya başlamaktadır. Uygulamanın başlatıldığından itibaren geçen süre Şekil 4.1’de formun sağ üst kısmında verilmiştir.

Etiketleme aşamasında kolaylık sağlanması amacı ile erkek, kadın, kurum ve kuruluşların listesi tutulup, tweeti paylaşan hesapların, kullanıcı ve ekran isimleri ile karşılaştırma yapılmış ve sonucunda eğer tweeti paylaşan hesabın ismi, listede tutulan isim ile uyuşuyor ise, ilgili alan sayacı bir arttırılmaktadır. Şekil 4.1’de görüldüğü üzere, anlık erkek 2723, kadın 3726, diğer 407, etiketli 6866, etiketsiz 2524, retweet 23092 ve toplamda ise 32482 adet eğitim seti örneği toplanmıştır. Etiketli olan alan; erkek, kadın ve diğer alanlarının toplamından oluşmaktadır. Etiketsiz olarak belirtilen alan ise, tweeti paylaşan hesabın kullanıcı ve ekran ismi, isim listesinde yok ise, etiketsiz alan sayacı bir arttırılmaktadır. Tweet toplama ve ön etiketleme işlemi bittikten sonra, toplanan tüm etiketli ve etiketsiz veriler tekrar gözden geçirilmiştir. Etiketli olarak işaretlenen veriler

tekrar kontrol edilmiş ve etkiketsiz olarak işaretlenen veriler ise etiketlenip, etiketli verilere dahil edilmiştir.

Uygulamanın arayüzünde bulunan ekran görüntüsü tuşu ile, istenirse tweet toplama esnasında tıklandığında, 3 adet uygulama ekran çıktı dosyası oluşturulabilmektedir. İlk çıktı; Erkek, Kadın ve Diğer sınıfının bulunduğu grafik, ikinci çıktı; Etiketli ve Etiketsiz verilerinin bulunduğu grafik, üçüncü çıktı; Retweet ve Toplam (Total) grafiğinin çıktısıdır.



Şekil 4.2. Geliştirilen uygulamanın oluşturduğu dosyalar

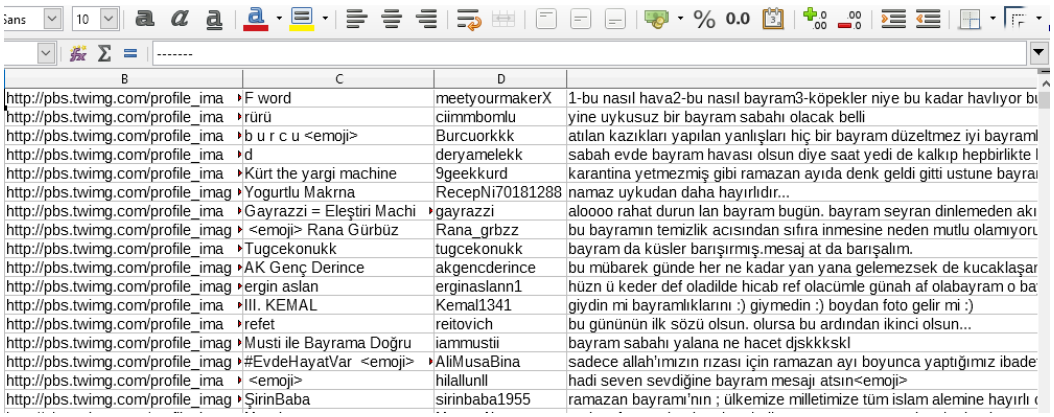
Uygulama çalışmaya başladığında arka planda indirilen tweetlere ait verileri uygun formatlarda dosyalara yazmaktadır. Şekil 4.2'de geliştirilen uygulama tarafından oluşturulan dosyalar gösterilmiştir. Uygulama tarafından oluşturulan dosyalar; 3 adet CSV (Virgülle ayrılmış veriler- Comma seperated value), 1 adet ARFF (Özellik İlişkili Dosya Formatı- Attribute Relationship File Format) dosya olmak üzere 4 adettir. Bölüm 6'da ARFF dosya formatı detaylarından bahsedilmiştir. Şekil 4.3, Şekil 4.4, Şekil 4.5 ve Şekil 4.6'da sırasıyla, "ArffText", "EtiketsizVeriler", "TemizVeri" ve "HamVeri" isimli dosyaların içerikleri gösterilmiştir.

Şekil 4.3'te toplanan tweetlerin, geliştirilen uygulama aracılığı ile arff dosya formatına dönüştürülmüş hali gösterilmektedir.



Şekil 4.3. ArffText dosya formatı içeriği

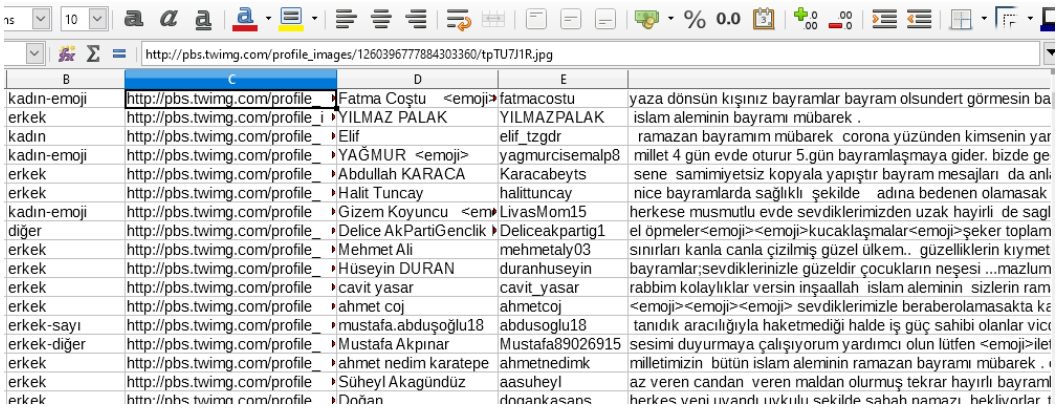
Şekil 4.4'te geliştirilen uygulama aracılığı toplanan, ön işleme sırasında sınıfı belirlenemeyen tweelerin, etiketsiz olarak belirtildiği dosya içeriği gösterilmektedir.



B	C	D	
http://pbs.twimg.com/profile_ima	F word	meetyourmakerX	1-bu nasıl hava2-bu nasıl bayram3-köpekler niye bu kadar havlıyor bu
http://pbs.twimg.com/profile_ima	rürü	ciimmbomlu	yine uykusuz bir bayram sabahı olacak belli
http://pbs.twimg.com/profile_ima	b u r c u <emoji>	Burcuorkkk	atılan kazıkları yapılan yanlışları hiç bir bayram düzeltmez iyi bayraml
http://pbs.twimg.com/profile_ima	d	deryamelekk	sabah evde bayram havası olsun diye saat yedi de kalkıp hepbirlikte k
http://pbs.twimg.com/profile_ima	Kürt the yargi machine	9geekkurd	karantina yetmezmiş gibi ramazan ayıda denk geldi gitti ustune bayra
http://pbs.twimg.com/profile_imag	Yogurtlu Makrna	RecepNi70181288	namaz uykudan daha hayırlıdır...
http://pbs.twimg.com/profile_ima	Gayrazzi = Eleştiri Machi	gayrazzi	aloooo rahat durun lan bayram bugün. bayram seyran dinlemeden akı
http://pbs.twimg.com/profile_imag	<emoji> Rana Gürbüz	Rana_grbzz	bu bayramın temizlik acısından sıfıra inmesine neden mutlu olamıyoru
http://pbs.twimg.com/profile_ima	Tugcekonukk	tugcekonukk	bayram da küsler barışırmış.mesaj at da barışalım.
http://pbs.twimg.com/profile_imag	AK Genç Derince	akgencderince	bu mübarek günde her ne kadar yan yana gelemezsek de kucaklaşar
http://pbs.twimg.com/profile_imag	ergin aslan	erginaslann1	hüzn ü keder def oladilde hicab ref olacümle günah af olabayram o ba
http://pbs.twimg.com/profile_ima	III. KEMAL	Kemal1341	giydin mi bayramlıklarını :) giymedin :) boydan foto gelir mi :)
http://pbs.twimg.com/profile_ima	refet	reitovich	bu gününün ilk sözü olsun. olursa bu ardından ikinci olsun...
http://pbs.twimg.com/profile_imag	Musti ile Bayrama Doğru	iammustii	bayram sabahı yalana ne hacet djskkkskl
http://pbs.twimg.com/profile_imag	#EvdeHayatVar <emoji>	AliMusaBina	sadece allah'ımızın rızası için ramazan ayı boyunca yaptığımız ibade
http://pbs.twimg.com/profile_ima	<emoji>	hilallunll	hadi seven sevdiğine bayram mesajı atsın<emoji>
http://pbs.twimg.com/profile_imag	ŞirinBaba	sirinbaba1955	ramazan bayramı'nın ; ülkemize milletimize tüm islam alemine hayırlı

Şekil 4.4. Etiketsiz veriler dosya formatı içeriği

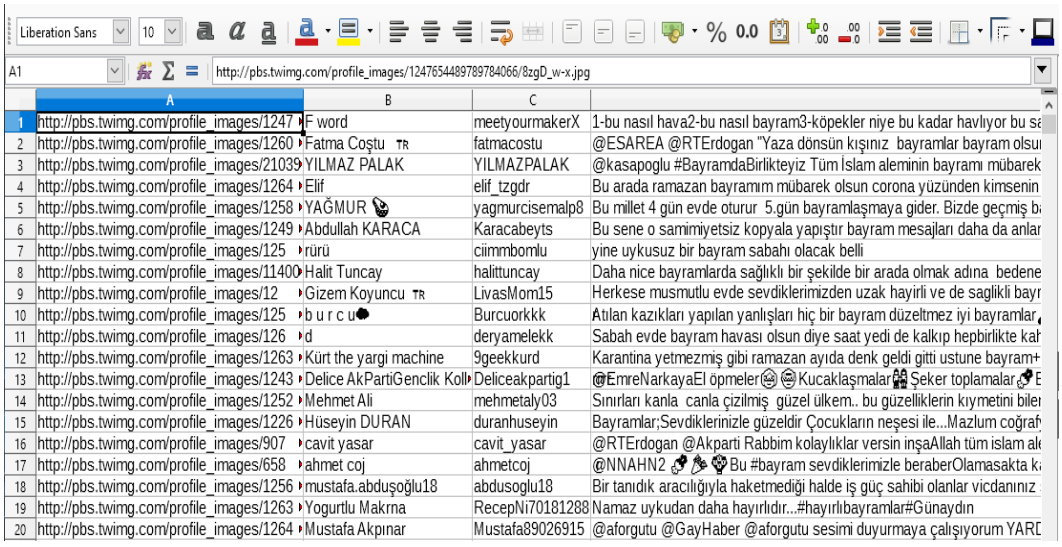
Şekil 4.5'te toplanan tweetlerin temizlenmesi ardından, kaydedildiği temiz verilerin bulunduğu dosya içeriği görülmektedir.



	B	C	D	E	
	kadın-emoji	http://pbs.twimg.com/profile_	Fatma Coştu <emoji>	fatmacostu	yaza dönsün kışınız bayramlar bayram olsundert görmesin ba
	erkek	http://pbs.twimg.com/profile_	YILMAZ PALAK	YILMAZPALAK	islam aleminin bayramı mübarek .
	kadın	http://pbs.twimg.com/profile_	Elif	elif_tzgdr	ramazan bayramım mübarek corona yüzünden kimsenin yar
	kadın-emoji	http://pbs.twimg.com/profile_	YAĞMUR <emoji>	yagmurcisemalp8	millet 4 gün evde oturur 5.gün bayramlaşmaya gider. bizde ge
	erkek	http://pbs.twimg.com/profile_	Abdullah KARACA	Karacabeyts	sene samimiyetsiz kopyala yapıştır bayram mesajları da anl
	erkek	http://pbs.twimg.com/profile_	Halit Tuncay	halittuncay	nice bayramlarda sağlıklı şekilde adına bedenen olamasak
	kadın-emoji	http://pbs.twimg.com/profile	Gizem Koyuncu <em	LivasMom15	herkese musmutlu evde sevdiklerimizden uzak hayirli de sagl
	diğer	http://pbs.twimg.com/profile_	Delice AkPartiGenclik	Deliceakpartig1	el öpmeler<emoji><emoji>kucaklaşmalar<emoji>şeker toplam
	erkek	http://pbs.twimg.com/profile_	Mehmet Ali	mehmetaly03	sınırları kanla canla çizilmiş güzel ülkem.. güzelliklerin kıymet
	erkek	http://pbs.twimg.com/profile_	Hüseyin DURAN	duranhuseyin	bayramlar;sevdiklerinizle güzeldir çocukların neşesi ...mazlum
	erkek	http://pbs.twimg.com/profile	cavit yasar	cavit_yasar	rabbim kolaylıklar versin inşaallah islam aleminin sizlerin ram
	erkek	http://pbs.twimg.com/profile	ahmet coj	ahmetcoj	<emoji><emoji><emoji> sevdiklerimizle beraberolamasakta ka
	erkek-sayı	http://pbs.twimg.com/profile_	mustafa.abduşoğlu18	abdusoglu18	tanıdık aracılığıyla haketmediği halde iş güç sahibi olanlar vic
	erkek-diğer	http://pbs.twimg.com/profile_	Mustafa Akpınar	Mustafa89026915	sesimi duyurmaya çalışıyorum yardımcı olun lütfen <emoji>ile
	erkek	http://pbs.twimg.com/profile_	ahmet nedim karatepe	ahmetnedimk	milletimizin bütün islam aleminin ramazan bayramı mübarek .
	erkek	http://pbs.twimg.com/profile	Süheyl Akagündüz	aasuheyl	az veren candan veren maldan olurmuş tekrar hayırlı bayraml
	erkek	http://pbs.twimg.com/profile	Doğan	dogankasaps	herkes yeni uyandı uykulu şekilde sabah namazı bekliyorlar t

Şekil 4.5. Temiz veriler dosya formatı içeriği

Şekil 4.6'da toplanan tweetlerin içeriği temizlenmemiş, ham verilerin dosya içeriği görülmektedir.



	A	B	C	
1	http://pbs.twimg.com/profile_images/1247	F word	meetyourmakerX	1-bu nasıl hava2-bu nasıl bayram3-köpekler niye bu kadar havlıyor bu sa
2	http://pbs.twimg.com/profile_images/1260	Fatma Coştu TR	fatmacostu	@ESAREA @RTErdogan "Yaza dönsün kışınız bayramlar bayram olsu
3	http://pbs.twimg.com/profile_images/21039	YILMAZ PALAK	YILMAZPALAK	@kasapoglu #BayramdaBirlikteyiz Tüm İslam aleminin bayramı mübarek
4	http://pbs.twimg.com/profile_images/1264	Elif	elif_tzgdr	Bu arada ramazan bayramım mübarek olsun corona yüzünden kimsenin
5	http://pbs.twimg.com/profile_images/1258	YAĞMUR	yagmurcisemalp8	Bu millet 4 gün evde oturur 5.gün bayramlaşmaya gider. Bizde geçmiş b
6	http://pbs.twimg.com/profile_images/1249	Abdullah KARACA	Karacabeyts	Bu sene o samimiyetsiz kopyala yapıştır bayram mesajları daha da anlar
7	http://pbs.twimg.com/profile_images/125	rürü	ciimmbomlu	yine uykusuz bir bayram sabahı olacak belli
8	http://pbs.twimg.com/profile_images/11400	Halit Tuncay	halittuncay	Daha nice bayramlarda sağlıklı bir şekilde bir arada olmak adına bedene
9	http://pbs.twimg.com/profile_images/12	Gizem Koyuncu TR	LivasMom15	Herkese musmutlu evde sevdiklerimizden uzak hayirli ve de saglikli bayr
10	http://pbs.twimg.com/profile_images/125	b u r c u	Burcuorkkk	Atılan kazıkları yapılan yanlışları hiç bir bayram düzeltmez iyi bayramlar
11	http://pbs.twimg.com/profile_images/126	d	deryamelekk	Sabah evde bayram havası olsun diye saat yedi de kalkıp hepbirlikte kah
12	http://pbs.twimg.com/profile_images/1263	Kürt the yargi machine	9geekkurd	Karantina yetmezmiş gibi ramazan ayıda denk geldi gitti ustune bayram+
13	http://pbs.twimg.com/profile_images/1243	Delice AkPartiGenclik Koll	Deliceakpartig1	@EmreNarkayaEl öpmeler Kucaklaşmalar Şeker toplamalar E
14	http://pbs.twimg.com/profile_images/1252	Mehmet Ali	mehmetaly03	Sınırları kanla canla çizilmiş güzel ülkem.. bu güzelliklerin kıymetini bile
15	http://pbs.twimg.com/profile_images/1226	Hüseyin DURAN	duranhuseyin	Bayramlar;Sevdiklerinizle güzeldir Çocukların neşesi ile...Mazlum coğraf
16	http://pbs.twimg.com/profile_images/907	cavit yasar	cavit_yasar	@RTErdogan @Akparti Rabbim kolaylıklar versin inşaAllah tüm islam al
17	http://pbs.twimg.com/profile_images/658	ahmet coj	ahmetcoj	@NNAHN2 Bu #bayram sevdiklerimizle beraberOlamasakta k
18	http://pbs.twimg.com/profile_images/1256	mustafa.abduşoğlu18	abdusoglu18	Bir tanıdık aracılığıyla haketmediği halde iş güç sahibi olanlar vicdanınız
19	http://pbs.twimg.com/profile_images/1263	Yogurtlu Makrna	RecepNi70181288	Namaz uykudan daha hayırlıdır...#hayırlıbayramlar#Günaydın
20	http://pbs.twimg.com/profile_images/1264	Mustafa Akpınar	Mustafa89026915	@aforgutu @GayHaber @aforgutu sesimi duyurmaya çalışıyorum YARD

Şekil 4.6. Ham veriler dosya formatı içeriği

4.2. JSON Nedir?

JSON (Javascript Nesnesi Gösterimi- Javascript Object Notation) biçiminde, veriler düzenli ve erişimi kolay bir şekilde saklanmaktadır. JSON evrensel bir veri saklama biçimidir. Neredeyse tüm modern programlama dilleri, JSON biçimini destekleyen kütüphanelere sahiptir. Programlama dilleri ile değiştirilebilen bir veri yapısı olması, JSON biçimini daha cazip bir hale getirmektedir (Lv, Yan, & He, 2018). JSON biçimi iki yapı üzerine kurulmuştur:

- Ad / değer çiftlerinden oluşan bir koleksiyon, çeşitli dillerde, bu bir nesne, kayıt, yapı, sözlük, karma tablo, anahtarlı liste veya ilişkilendirilebilir dizi olarak gerçekleştirilir.
- Sıralı değerler listesi, çoğu dilde, bir dizi, vektör veya liste olarak gerçekleştirilir.

JSON biçiminin kolay kullanımından dolayı birçok avantajı bulunmaktadır. Bunlar genel olarak:

- Tüm tarayıcılar tarafından desteklenmektedir.
- Basit bir sözdizimine sahiptir.
- Modern programlama dilleri ile basit şekilde kullanılabilir.
- Sunucu ile web uygulaması arasında JSON biçimi kullanılarak veri iletilebilmektedir.

Uygulama tarafından indirilen bir tweet verisi Şekil 4.7'de JSON dosya biçiminde görülmektedir. Burada kullanıcı ve paylaştığı tweet hakkında bilgiler görülmektedir.

```
{
  "id": 464733361420787712,
  "text": "#GoodMorning from #Dante and hulk http:\/\/t.co\/tP2vhUgQ4Q",
  "user": {
        "id": 291386753,
        "name": "Anthony Piper",
        "screen_name": "piper89ad",
        "followers_count": 856,
        "friends_count": 1954,
        "verified": false,
        "statuses_count": 10717,
  },
  "retweet_count": 0,
  "favorite_count": 0,
  "entities": {
        "hashtags": [
                {
                "text": "GoodMorning",
                "indices": [   0, 12   ]
                },
                {
                "text": "Dante",
                "indices": [   18, 24  ]
                }
        ]
  }
}
```


Şekil 4.7. JSON biçiminde bir tweet verisi

4.2.1. JSON Twitter Veri Biçimi

Twitter'da üzerinden veri analizi yapılmak istenildiği zaman, Twitter'ın sunmuş olduğu API üzerinden bağlantı kurulup, tweetler toplanmaya hazır hale gelmektedir. Twitter, toplanan tweetleri, JSON biçiminde sunmaktadır. Şekil 4.7'de görüldüğü gibi Twitter'dan toplanan verilerin JSON biçiminde sağladığı bilgiler görülmektedir. Bir tweet verisinin tüm alanları içerir hali, istenen alanları belirtilmiş şekilde alınan tweet verisine göre boyutu çok daha fazladır. Bunun sebebi ise tüm alanların tweet verisi olarak gelmesi ile ihtiyaç olmayan alanların da alınmasına neden olur bu durum da verinin boyutunu arttırır. Twitter'dan, alan belirterek çekilen tweete ait bir yapı Şekil 4.8'de verilmiştir.



Şekil 4.8. Belli alanları içeren json tweet biçimi

4.2.2. JSON Biçimindeki Alanlar ve Anlamları

Şekil 4.8'de gösterildiği gibi Twitter'ın, JSON biçiminde sunulan alanlarının herbiri tweet ve kullanıcı hakkında bilgi sağlamaktadır. JSON biçiminde sunulan kullanıcı ile ilgili alanların bir kısmı aşağıda açıklanmıştır:

Id: Kullanıcı eşşiz numarası,

Name: Kullanıcının tam adı,

Screen_name: Kullanıcının eşsiz olarak kullandığı ad,

Description: Kullanıcı hesabı için girilen tanımlama metni,

Followers_count: Kullanıcı hesabının takipçi sayısı,

Location: Kullanıcı hesabı için girilmiş olan lokasyon,

Statuses_count: Kullanıcının paylaştığı toplam tweet sayısı,

Friends_count: Kullanıcı tarafından takip edilen hesapların sayısı,

Created_at: Kullanıcı hesabının oluşturulma zamanı.

Bunların dışında olan alanlar, çalışmada kullanılmadığı için açıklanmamıştır.

4.3. WEKA Nedir?

Weka, yapay zekânın önemli alanlarından biri olan makine öğrenmesine ait birçok algoritmayı içerisinde barındıran bir yazılım aracıdır. Weka aracı Waikato üniversitesinde açık kaynak kodlu olarak JAVA dili ile geliştirilmiştir ve genel kamu lisansı ile dağıtılmaktadır (Şeker, 2013).

Yapay zekâ giderek birçok farklı alanda kullanılmaktadır. Yapay zekanın; nispeten farklı alanlara yönelik Derin Öğrenme, Makine Öğrenmesi, Yapay Sinir Ağları, Genetik Algoritmalar gibi birçok alt alanı mevcuttur. Yapay zekâ, bu alt alanlar sayesinde güçlü ve zeki çıkarımlar yapabilme kapasitesine sahip olmuştur. Bu alt alanların biri de Veri Madenciliği ve Makine Öğrenmesidir. Bu alanda çalışma yürüten kişilerin büyük oranda kullandığı uygulama WEKA yazılımıdır. WEKA yazılımı içerisinde makine öğrenmesi ve İstatistik ile ilgili pekçok kütüphane hazır olarak gelmektedir (Weka, 2007).

4.3.1. WEKA Arayüzü ve Modüleri

Şekil 4.9'da WEKA yazılımının arayüzü gösterilmiştir. WEKA yazılımında birçok modül vardır. Veri ön işlemesi (data preprocessing), görselleştirme (visualization), sınıflandırma (classification), gruplandırma (clustering), öznitelik seçimi ve öznitelik çıkarımı bunlardan bazılarıdır. Ayrıca bu işlemler sonucunda çıkan sonuçların görsel olarak ifade edilmesini sağlayan görselleştirme araçları bulunmaktadır (Weka, 2008). Şekil 4.9'daki arayüzde kullanılan bazı modüller aşağıda maddeler halinde açıklanmıştır:

Ön işlem (Preprocess): Önişlem sekmesi, veri dosyasının seçilmesini sağlamaktadır. Bu çalışmada önişlem aşamasında StringToWordVector kullanılmıştır. StringToWordVector, kelimelerin, sunulan bilgi içerisindeki bulunma sıklığını, sayısal özelliğe çevirmeyi sağlayan bir yapıdır.

Sınıflandır (Classify): Sınıflandır sekmesi, önceden işlenmiş veri dosyalarında, farklı makine öğrenmesi algoritmaları uygulanmasına ve bunlarla denemeler yapılmasına olanak tanımaktadır.

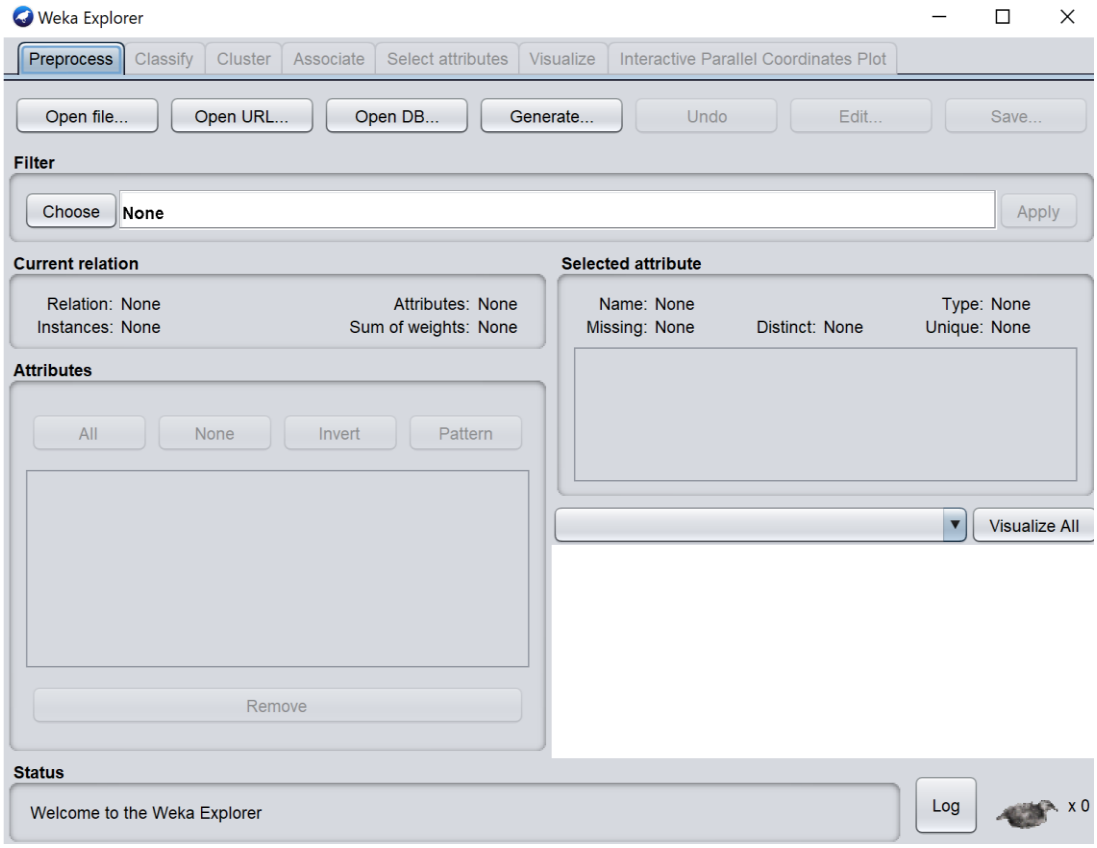
Küme (Cluster): Küme sekmesi, veri dosyasındaki kümeleri tanımlayan, farklı kümeleme algoritmalarının uygulanmasına olanak tanımaktadır.

İlişkilendirme (Association): İlişkilendirme sekmesi, verilerdeki ilişkilendirmeyi tanımlayan, ilişkilendirme kurallarının uygulanmasını sağlamaktadır.

Öznitelikleri seçin (Select Attributes): Öznitelikleri seçin sekmesi, özniteliklerin denemeye dahil edilmesi ve hariç tutulmasındaki değişikliklerin görülmesini sağlamaktadır.

Görselleştir (Visualize): Görselleştir sekmesi ile, eğitim setine ait görselleştirme yapılabilmektedir.

Eğitim setinin ilk ön işleme süreci tamamlanana kadar kullanıcı, farklı sekmeler arasında dolaşamaz veya işlem yapamaz.



Şekil 4.9. WEKA arayüzü

4.3.2. ARFF Formatı Nedir?

Weka, girdi olarak birçok dosya biçimine ilave olarak, arff (Özellik İlişkili Dosya Biçimi- Attribute Relationship File Format) biçiminde de dosyalar ile işlem yapabilmektedir. ARFF dosya biçimine geçmeden önce, veri madenciliği alanında sık kullanılan ve ARFF formatına çevrilen **.***csv* dosya formatından bahsedilecek olursa, CSV dosya formatı verileri virgül ile ayırarak belirli bir düzen içerisinde bulunmalarını sağlayabilmektedir. CSV dosya formatını elektronik tablo programları yardımı ile kolayca elde edilebilmektedir. Şekil 4.10'da arff formatı hakkında bir ekran çıktısı görülmektedir.

```
@relation whatever

@attribute R numeric
@attribute G numeric
@attribute B numeric
@attribute color { blue , orange , violet , yellow , green , red }

@data
105 , 30 , 255 , blue
255 , 183 , 17 , orange
255 , 53 , 221 , violet
244 , 254 , 100 , yellow
251 , 255 , 48 , yellow
158 , 255 , 155 , green
255 , 106 , 108 , red
159 , 150 , 255 , blue
71 , 228 , 109 , green
255 , 45 , 178 , violet
120 , 255 , 30 , green
214 , 50 , 250 , violet
255 , 176 , 40 , orange
255 , 79 , 79 , red
255 , 162 , 22 , orange
255 , 247 , 114 , yellow
61 , 180 , 254 , blue
255 , 97 , 12 , red
```

Şekil 4.10. ARFF formatı

Bir ARFF dosyası @relation, @attribute ve @data yapısından oluşur. Relation (İlişki) veri bütününe bir isim verilmesini sağlayan alandır.

Relation (İlişki): Veri bütününe bir isim verilmesi gereken alandır.

Attribute (Öznitelik): Öznitelik yardımıyla bir değişken oluşturulur. Aynı zamanda o değişken için bir veri tipi belirlenir. Weka yazılımındaki öznitelikler şunlardır;

NUMERIC: Sayısal değerlere verilen değerleri içerir.

REAL: Tüm reel sayısal değerleri içerir.

STRING: Metinsel değerlerdir.

NOMINAL: Küme şeklindeki değerlerdir.

DATE: Tarih tip değerlerdir.

Data (Veri): Mevcutta varolan veriler bütünüdür.

BÖLÜM 5

SONUÇLAR

5.1. Deneysel Sonuçlar

Bu bölümde geliştirilen uygulama ile Twitter üzerinden toplanan verilerden oluşturulan eğitim setleri için JSON biçiminde toplanan tweetlerin CSV formatına dönüştürülmesi ardından, WEKA yazılımı aracılığı ile modeller oluşturulmuştur. Modellerin oluşturulması aşamasında, önceden belirlenen makine öğrenmesi algoritmaları kullanılmıştır. Modellerin oluşturulmasında çapraz doğrulama yöntemi kullanılmıştır. Oluşturulan bu modellerin sınıflandırma başarımı sonuçları üzerinden algoritmaların karşılaştırılması yapılmıştır. Modellere ait sonuçların karşılaştırılması doğruluk ve kappa değerleri üzerinden yapılmıştır.

Kappa İstatistiği: Cohen'in kappa katsayısı, iki değer arasındaki uyuşmanın güven ölçütünü baz alan bir istatistiksel yapıdır (Cohen, 1960). Bu iki değer 0 ile 1 arasında değişmektedir. 0 değeri, hiç uyuşma olmaması, 1 değeri ise mükemmel uyuşma olması anlamına gelmektedir. Landis ve Koch, bu uyuşmayı Çizelge 5.1'deki gibi sunmuşlardır (Landis, &Koch, 1977). Denklem 3.2'da kappa değerinin elde edilmesi için kullanılan ifade verilmiştir.

$$K = \frac{(P_0 - P_c)}{1 - P_c} \tag{3.2}$$

P_0 : Kabul edilen oran

P_c : Kabul edilmesi beklenen oran

Çizelge 5.1. Kappa değeri uyuşma tablosu

K	Uyuşma
<0	Hiç uyuşma olmaması
0,0- 0,20	Önemsiz uyuşma olması
0,21 – 0,40	Orta derecede uyuşama olması
0,41 – 0,60	Ekseriyetle uyuşma olması
0,61 – 0,80	Önemli derecede uyuşma olması
0,81 – 1,00	Neredeyse mükemmel uyuşma olması

Doğruluk: Makine öğrenmesi algoritmasının doğruluğu, algoritmanın bir veri noktasını hangi sıklıkla doğru sınıflandırdığını ölçmenin bir yoludur. Doğruluk, tüm veri noktalarından doğru tahmin edilen veri noktalarının sayısının toplam veri noktası sayısına oranıdır. Aynı zamanda sınıflandırma modellerini değerlendirmek için bir ölçü birimidir. Doğruluk hesaplaması için kullanılan ifade, denklem 3.3 verilmiştir.

$$\text{Doğruluk} = \frac{\text{Doğru Tahmin Adedi}}{\text{Toplam Tahmin Adedi}} \tag{3.3}$$

5.2. Cinsiyet Eğitim Seti Sonuçları

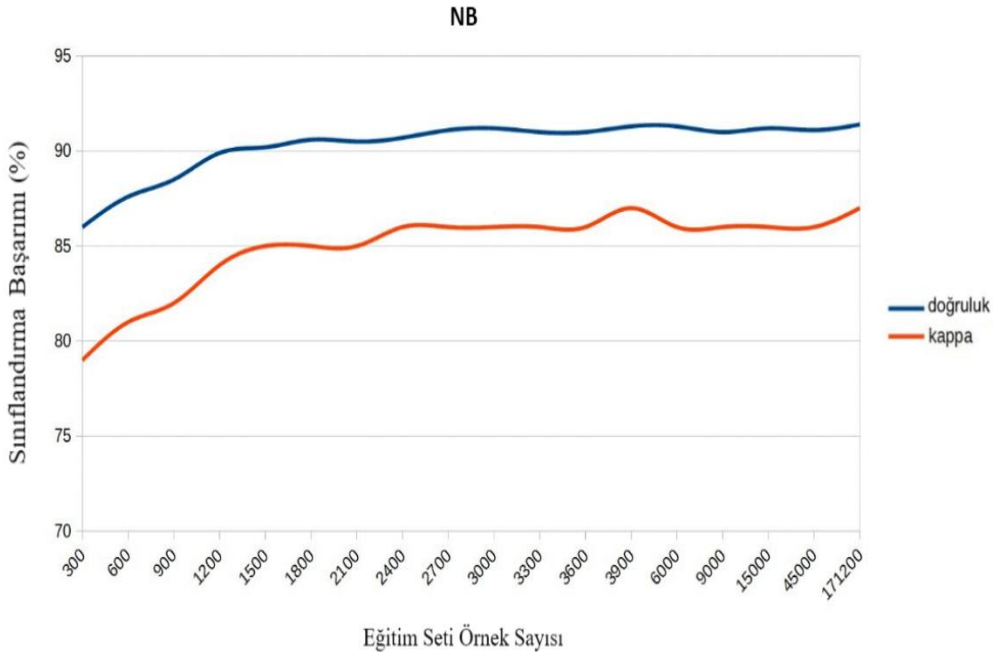
En fazla örnek sayısı barındıran cinsiyet eğitim seti; farklı zaman dilimlerinde Twitter platformunda en çok kullanılan konuları esas alarak rastgele olarak toplanan 172.200 adet tweetten oluşmaktadır. Bu eğitim setinde 54.370 adet tweet kadın sınıfına, 48.985 adet tweet erkek sınıfına ve 67.845 adet tweet ise diğer sınıfına aittir. Eğitim setleri içindeki sınıflar arası örnek sayısı farkının az ya da eşit olmasına dikkat edilmiştir. Eğer herhangi bir sınıf (kadın, erkek, diğer), diğerine göre orantısız bir örnek fazlalığı gösterirse, bu da o sınıf için sonuçlara yanlılık olarak etki edecektir.

Bu çalışmada, cinsiyet eğitim setine ait modellerin algoritmalara göre başarımının nasıl değiştiğini ortaya koyabilmek için eğitim seti örnek sayıları azdan çoğa doğru olacak şekilde ayarlanmıştır. Bunun bu şekilde yapılmasının amacı algoritmalara bağlı olarak örnek sayısına bağlı doygunluk noktasının ortaya koyulabilmesi içindir. Çizelge 5.2'de görüldüğü gibi eğitim seti; 300, 600, 900, 1.200, 1.500, 1.800, 2.100, 2.400, 2.700, 3.000, 3.300, 3.600, 3.900, 6.000, 9.000, 15.000, 45.000, 171.200 şeklinde farklı sayıda örnek içeren 18 farklı eğitim seti oluşturulmuştur. Kendi içinde dengeli bir şekilde ayarlanan eğitim setlerinden makine öğrenmesi algoritmaları ile modeller oluşturulmuştur. Eğitim seti örnek adedi 300 olan set; 100 adet kadın, 100 adet erkek ve 100 adet diğer sınıfından örnek içermektedir. Benzer şekilde eğitim seti örnek adedi 600 olan set ise; 200 adet kadın, 200 adet erkek ve 200 adet diğer sınıfından örnekten oluşmaktadır. Bu şekilde, tüm eğitim setleri kendi içinde dengeli olarak ayarlanmıştır.

Çizelge 5.2. Cinsiyet doğruluk ve kappa değerleri

Eğitim seti	NB		SMO		KNN		C4.5		SVM	
	Doğruluk (%)	Kappa	Doğruluk (%)	Kappa	Doğruluk (%)	Kappa	Doğruluk (%)	Kappa	Doğruluk (%)	Kappa
300	86,0	0,79	98,3	0,97	82,3	0,73	97,6	0,96	44,6	0,17
600	87,6	0,81	98,8	0,98	83,0	0,74	98,1	0,97	44,0	0,16
900	88,5	0,82	99,4	0,99	86,4	0,79	98,4	0,97	47,1	0,20
1.200	89,9	0,84	99,4	0,99	88,0	0,82	98,4	0,97	48,4	0,22
1.500	90,2	0,85	99,8	0,99	87,3	0,81	98,4	0,97	48,6	0,23
1.800	90,6	0,85	99,9	0,99	86,8	0,80	98,7	0,98	49,2	0,23
2.100	90,5	0,85	99,9	0,99	86,7	0,80	98,9	0,98	49,9	0,24
2.400	90,7	0,86	99,9	0,99	87,3	0,81	99,0	0,98	50,9	0,26
2.700	91,1	0,86	99,9	0,99	87,6	0,81	99,0	0,98	51,5	0,27
3.000	91,2	0,86	99,9	0,99	88,0	0,82	99,4	0,99	51,1	0,26
3.300	91,0	0,86	99,9	0,99	88,3	0,82	99,4	0,99	52,1	0,28
3.600	91,0	0,86	99,9	0,99	88,4	0,82	99,3	0,99	52,4	0,28
3.900	91,3	0,87	99,9	0,99	88,5	0,82	99,4	0,99	51,7	0,27
6.000	91,3	0,86	99,9	0,99	89,8	0,84	99,6	0,99	54,2	0,31
9.000	91,0	0,86	99,9	0,99	90,3	0,85	99,4	0,99	56,6	0,34
15.000	91,2	0,86	99,9	0,99	91,1	0,86	99,6	0,99	60,2	0,40
45.000	91,1	0,86	99,9	0,99	93,5	0,90	99,6	0,99	69,6	0,54
171.200	91,4	0,87	99,9	0,99	95,7	0,93	99,4	0,99	81,9	0,72

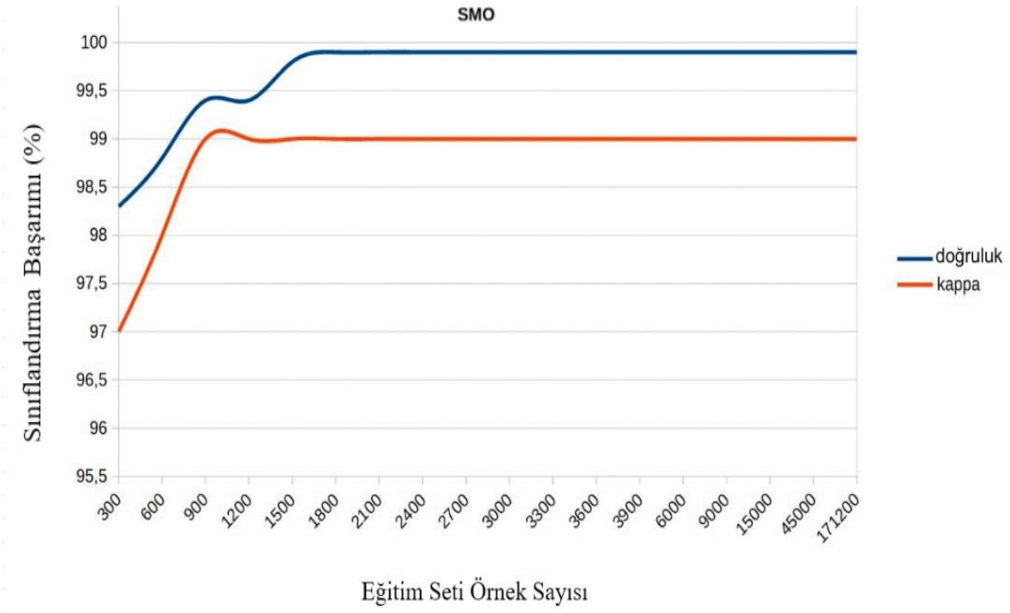
Cinsiyet tahmini için oluşturulan bu 18 farklı eğitim seti için makine öğrenmesi algoritmalarına bağlı olarak elde edilen sınıflandırma başarımları bu bölümde algoritma bazında ele alınmıştır. Bu eğitim setlerine ait modellerin örnek sayılarına bağlı sınıflandırma başarımları, doğruluk ve kappa değeri üzerinden Şekil 5.1'de Naive Bayes algoritması için verilmiştir.



Şekil 5.1. NB algoritmasının cinsiyet eğitim seti doğruluk ve kappa değerleri

Şekil 5.1'de görüldüğü gibi x ekseni; eğitim seti örnek adedini, y ekseni ise doğruluk yüzdelik değerini ve kappa değerini göstermektedir. Makine öğrenmesi algoritmalarından, olasılık tabanlı olan Naive Bayes algoritmasının veri sonuçlarına bakıldığında, eğitim seti örnek sayısı yaklaşık 1.500 olduğunda doygunluğa ulaştığı görülmektedir. Bu noktadan sonra, 169.700 adet eğitim seti daha eklendiğinde, doğruluk değeri yaklaşık %1 gibi, küçük bir oranda artmaktadır. Bunun anlamı ise daha fazla örnek eklenmesi durumunda sonuca, beklenildiği gibi çok az bir oranda etki edeceği söylenebilir.

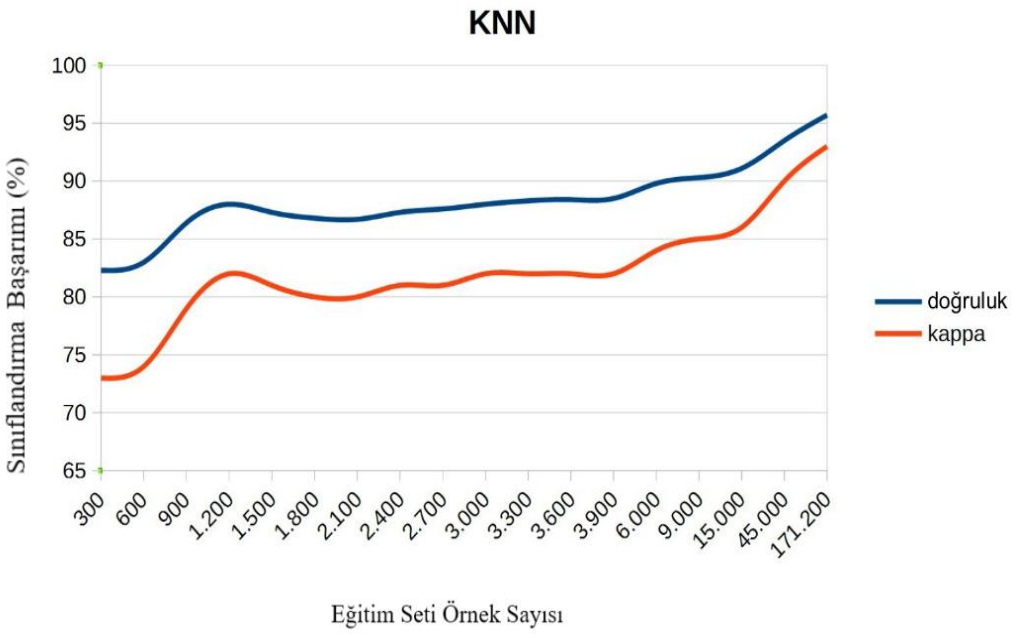
Cinsiyet eğitim setlerine ait modellerin örnek sayılarına bağlı sınıflandırma başarımları, doğruluk ve kappa değeri üzerinden Şekil 5.2'de SMO algoritması için verilmiştir.



Şekil 5.2. SMO algoritmasının cinsiyet eğitim seti doğruluk ve kappa değerleri

Şekil 5.2'de makine öğrenmesi algoritmalarından olan SMO algoritmasının doğruluk ve kappa sonuç değerlerine ait grafik gösterilmektedir. Eğitim seti örnek sayısı yaklaşık 1.800 olduğunda SMO algoritmasının örnek doygunluğuna ulaştığı görülmüştür. Eğitim seti örnek sayısı daha fazla arttırıldığında, doğruluk ve kappa değerlerinin hemen hemen değişmediği grafikten açık olarak görülmektedir. Bundan sonra ne kadar örnek eklenirse eklensin, bu sonucu değiştirmeyecektir.

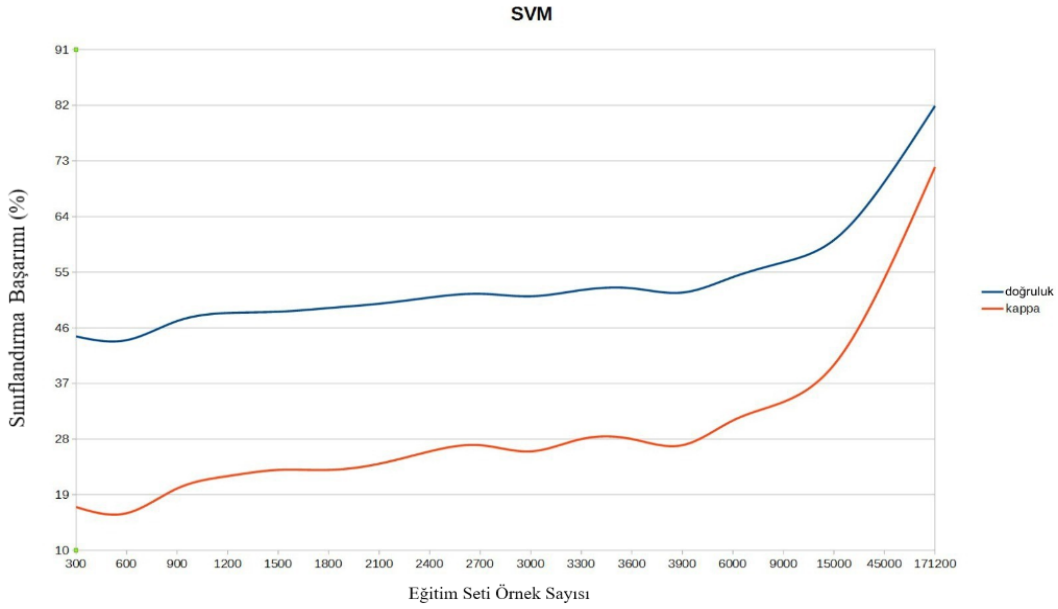
Cinsiyet tahmini için oluşturulan bu 18 farklı eğitim seti için makine öğrenmesi algoritmalarına bağlı olarak elde edilen sınıflandırma başarımları doğruluk ve kappa değerleri Şekil 5.3'de KNN makine öğrenmesi algoritması için verilmiştir.



Şekil 5.3. KNN algoritmasının cinsiyet eğitim seti doğruluk ve kappa değerleri

Şekil 5.3'de, makine öğrenmesi algoritmalarından olan KNN algoritmasının cinsiyet eğitim setlerinden oluşturulan modellere ait doğruluk ve kappa değerlerinden oluşan grafik verilmiştir. 15.000 ve 171.200 eğitim setleri arasında, doğruluk ve kappa değeri artışının olduğu görülmektedir. Bunun anlamı, eğitim setine örnek eklendikçe, doğruluk ve kappa değerinin de artacağı öngörülmektedir.

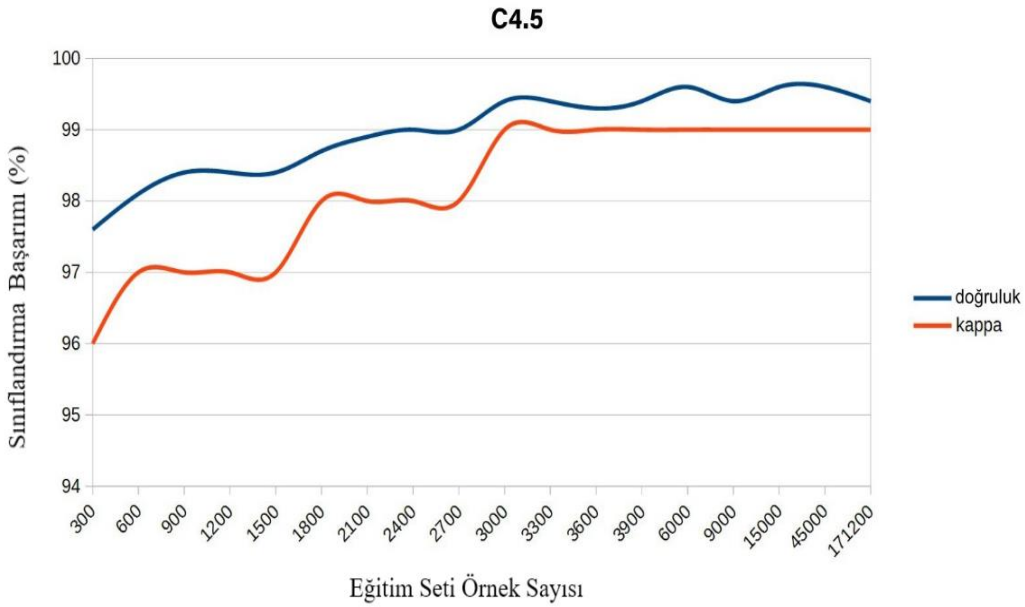
Cinsiyet eğitim setlerine ait modellerin örnek sayılarına bağlı sınıflandırma başarımları, doğruluk ve kappa değeri üzerinden Şekil 5.4'de SVM makine öğrenmesi algoritması için verilmiştir.



Şekil 5.4. SVM algoritmasının cinsiyet eğitim seti doğruluk ve kappa değerleri

SVM makine öğrenmesi algoritması ile, cinsiyet eğitim setleri üzerinden modeller oluşturulduğunda ise modellerin doğruluk ve kappa değerlerinden elde edilen grafik Şekil 5.4'te verilmiştir. SVM makine öğrenmesi algoritmasının, düşük örnek sayısı olan eğitim setlerinde, doğruluk ve kappa değerlerinin nispeten düşük olduğu görülmektedir. Bunun anlamı, SVM makine öğrenmesi algoritmasının, eğitimindeki örnek sayısı doygunluğa ulaşamadığından dolayı, doğruluk ve kappa değerlerinin yükselmesi için daha çok örnek adedine ihtiyaç duyduğundandır. Şekil 5.4'te, eğitim seti örnek adedi arttırıldıkça, doğruluk ve kappa değerlerininde yükseldiği görülmüştür. Daha fazla eğitim seti örneği eklenmesi durumunda, doğruluk ve kappa değerlerinin de oransal artacağı öngörülmektedir.

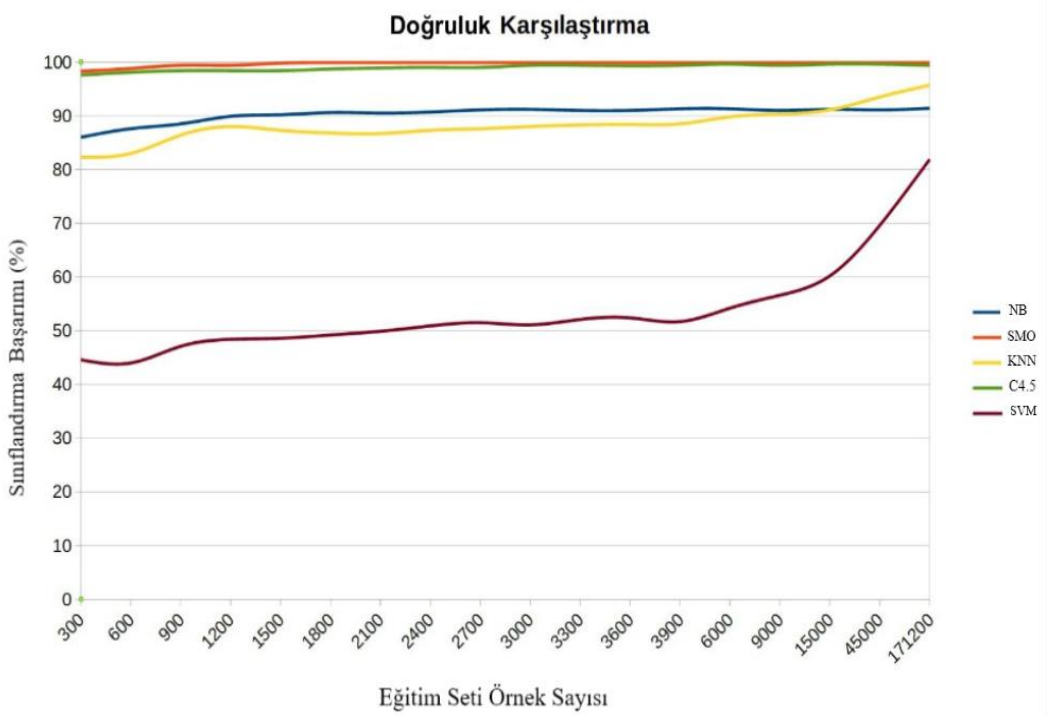
Cinsiyet eğitim setlerine ait modellerin örnek sayılarına bağlı sınıflandırma başarımları, doğruluk ve kappa değeri üzerinden Şekil 5.5'de C4.5 makine öğrenmesi algoritması için verilmiştir.



Şekil 5.5. C4.5 algoritmasının cinsiyet eğitim seti doğruluk ve kappa değerleri

Şekil 5.5'de, makine öğrenmesi algoritmalarından olan C4.5 karar ağacı algoritmasının cinsiyet eğitim setleri modellerine ait doğruluk ve kappa değerlerinin grafiği gösterilmektedir. C4.5 algoritmasında, 3.000 örnek adedinden sonra, örnek eklenmesi durumunda, kappa değeri sabit kaldığı için, öğrenmeyi tamamladığı görülmektedir. Bundan sonra, eğitim setine örnek eklenmesi doğruluk ve kappa değerlerinin başarım oranlarına ciddi etkisi olmayacağı öngörülmektedir.

Cinsiyet tahmini için oluşturulan bu 18 farklı eğitim seti için makine öğrenmesi algoritmalarına bağlı olarak elde edilen sınıflandırma başarımları doğruluk değerlerinin karşılaştırılması Şekil 5.6'da gösterilmiştir.



Şekil 5.6. Cinsiyet doğruluk değerlerinin karşılaştırılması

Şekil 5.6'da, tüm eğitim setleri için Naive Bayes, SMO, SVM, KNN ve C4.5 karar ağacı makine öğrenmesi algoritmalarına ait modellerin sınıflandırma doğruluk değerlerinden elde edilen grafik verilmiştir. Bu grafikte, algoritmaların doğruluk başarımları toplu olarak görülmektedir. En düşük sınıflandırma başarımı olan algoritmanın SVM algoritması olduğu görülmüştür. SVM algoritmasının, düşük örnek sayısına sahip eğitim setlerinde yeterli başarıda sınıflandırma yapamadığı görülmektedir. SVM makine öğrenmesi algoritması için eğitim seti örnek sayısı daha da arttıldığında, diğer makine öğrenmesi algoritmalarının sınıflandırma başarımı seviyesine ulaşabileceği öngörülmektedir. Grafik incelendiğinde, sınıflandırma başarımı en yüksek olan algoritmaların ise SMO ve C4.5 karar ağacı algoritmalarının olduğu görülmektedir.

5.3. Cinsiyet Eşsiz Kullanıcı Eğitim Seti ve Sonuçları

Twitter platformu kullanıcıları, düşüncelerini ifade edebilmek için birden fazla tweet paylaşabilmektedir. Şu ana kadar sonuçları değerlendirilen cinsiyet eğitim setleri olarak belirtilen setlerin içinde bir kullanıcıya ait birden fazla farklı örnek bulunmaktadır. Ancak yukarıda anlatılan 171.200 örnek içeren cinsiyet eğitim setinin bir alt seti olarak cinsiyet eşsiz kullanıcı eğitim seti isimli bir set daha oluşturulmuştur. Bu set içinde bir kullanıcı sadece bir örnek olarak bulunmaktadır. Oluşturan bu cinsiyet eşsiz kullanıcı eğitim seti öznitelikleri Çizelge 2.1'de verilenler ile aynıdır. Cinsiyet eşsiz kullanıcı eğitim seti oluşturulmasının amacı, bu iki eğitim seti arasındaki farkın, makine öğrenmesi algoritmalarının başarımlarını nasıl etkilediğini anlamak içindir.

Cinsiyet tahmini için 171.200 örneğe en büyük cinsiyet eğitim setinden, kullanıcı numaraları kontrol edilerek 59.750 adet eşsiz örnek olduğu tespit edilmiştir. Çizelge 5.3'de bu eşsiz örneklerden oluşturulan cinsiyet eşsiz kullanıcı eğitim setinin örnek sayılarının sınıflara göre dağılımları verilmiştir.

Çizelge 5.3. Eşsiz kullanıcı eğitim seti örnek adetleri

Sınıf	Tweet Adedi
KADIN	24.520
ERKEK	26.740
DİĞER	8.490

Cinsiyet eğitim seti, eşsiz kullanıcı eğitim setine indirgenmeden önce; Kadın sınıfında bulunan eğitim seti örnek adedi 54.370, Erkek sınıfında bulunan eğitim seti örnek adedi 48.985 ve Diğer sınıfında ise 67.845 eğitim seti örnek içermektedir. Fakat, eşsiz eğitim setine indirgendikten sonra, sınıflarda kalan eğitim seti örnek adetleri Çizelge 5.3'de gösterildiği gibi; Kadın sınıfında 24.520 adet, Erkek sınıfında 26.740 adet ve Diğer sınıfında 8.490 adet örnek olduğu görülmektedir. 171.200 ve 59.750 örneğe sahip olan 2 eğitim setinin, Naive Bayes, SMO, KNN ve C4.5 makine öğrenmesi

algoritmaları için modeller oluşturulmuş ve modellerin elde edilen sınıflandırma başarım sonuçları Çizelge 5.4'te gösterilmiştir.

Çizelge 5.4. Cinsiyet Eğitim Setlerinin Doğruluk Sonuçları

Makine Öğrenmesi Algoritması	Cinsiyet Eşsiz Kullanıcı Seti (59.750)		Cinsiyet Eğitim Seti (171.200)	
	Doğruluk (%)	Kappa	Doğruluk (%)	Kappa
NB	90,9	0,85	91,4	0,87
SMO	99,9	0,99	99,9	0,99
KNN	92,6	0,88	95,7	0,93
C4.5	99,8	0,99	99,6	0,99

Çizelge 5.4'te, iki eğitim seti arasında, makine öğrenmesi algoritmalarının doğruluk ve kappa değerleri arasında önemli bir fark görülmemektedir. 171.200'lik eğitim setinde, Naive Bayes algoritmasının sınıflandırma doğruluk başarımı %91,4 iken, eşsiz kullanıcı eğitim setinde ise, bu değerinin %90,9 olduğu görülmüştür. Bu sonuçlara göre eğitim setinde aynı kullanıcıya ait örnek sayısının tek olması ile birden çok olmasının, sınıflandırma başarımına etki etmediği söylenebilir.

En büyük örnek sayısı 171.200 olan cinsiyet eğitim seti içindeki örneklerin bazı özniteliklere bağlı olarak yapılan analizinden elde edilen verilere göre hazırlanmış grafikler, çalışmanın ekler bölümünde verilmiştir.

5.4. İlgi Alanı Eğitim Seti Sonuçları

Twitter'dan toplanan ilgi alanı verileri; bilim ve teknoloji, siyaset, spor, eğitim, ekonomi ve diğer alanlarından oluşmuştur. Diğer sınıfı; bu belirtilen 5 sınıf dışında kalan, film, dizi, müzik, sanat, edebiyat, turizm ve daha birçok farklı alanı içeren örnek verilerin eğitim setine dahil edilmesiyle oluşturulmuştur. İlgi alanı eğitim seti 90.000 örnekten oluşmaktadır. İlgi alanı eğitim seti olarak kulanılan eğitim setleri en büyük 30.000 ve en büyük 90.000 olarak iki farklı eğitim seti halinde ayarlanmıştır. Bu iki

eğitim setinin her biri az örnek sayısından çok örnek sayısına doğru artan şekilde 16 farklı alt eğitim seti haline getirilmiştir. Bunun amacı ise algoritmaların örnek sayısına bağlı doygunluğunun belirlenebilmesi içindir. Bununla ayrıca eğitim setinde örnek miktarı artınca doğruluk ve kappa değerlerinde meydana gelen değişikliğin hassas olarak izlenebilmesi amaçlanmıştır. Çizelge 5.5'te 30.000 örnekten oluşan ilgi alanı eğitim setinin sınıflandırma doğruluk değeri ve kappa değeri sonuçları gösterilmektedir.

Çizelge 5.5. 30.000 ilgi alanı eğitim seti doğruluk ve kappa sonuçları

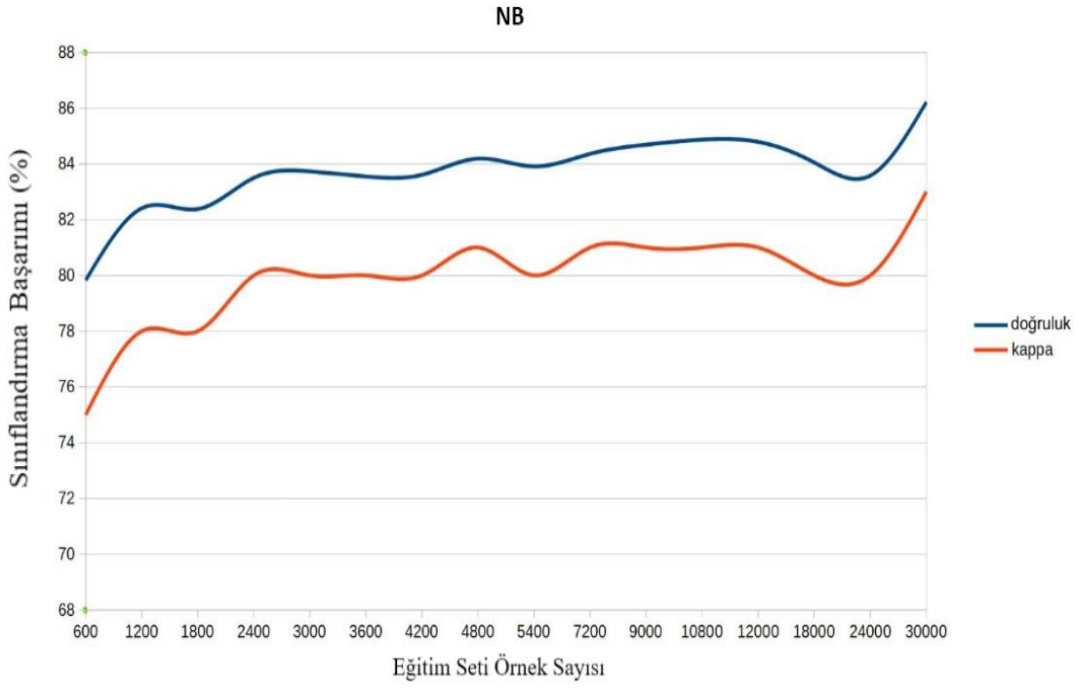
Eğitim Seti	NB		KNN		SMO		C4.5		SVM	
	Doğruluk (%)	Kappa	Doğruluk (%)	Kappa	Doğruluk (%)	Kappa	Doğruluk (%)	Kappa	Doğruluk (%)	Kappa
600	79,83	0,75	50,16	0,40	81,33	0,77	80,16	0,76	62,50	0,55
1.200	82,41	0,78	48,00	0,37	85,66	0,82	81,50	0,77	57,58	0,49
1.800	82,38	0,78	52,22	0,42	86,33	0,82	87,50	0,85	58,44	0,50
2.400	83,50	0,80	53,29	0,43	87,45	0,84	88,95	0,86	59,00	0,50
3.000	83,73	0,80	54,76	0,45	88,33	0,86	90,36	0,88	64,13	0,56
3.600	83,55	0,80	56,08	0,47	88,22	0,85	89,00	0,86	65,58	0,58
4.200	83,61	0,80	56,02	0,47	88,80	0,86	89,90	0,87	67,23	0,60
4.800	84,19	0,81	57,67	0,49	89,81	0,87	90,10	0,88	67,73	0,61
5.400	83,91	0,80	56,60	0,47	89,61	0,87	90,50	0,88	69,30	0,63
7.200	84,36	0,81	58,35	0,50	90,54	0,88	90,15	0,88	73,30	0,67
9.000	84,69	0,81	59,66	0,51	91,04	0,89	90,64	0,88	79,02	0,74
10.800	84,87	0,81	62,15	0,54	91,01	0,89	90,10	0,88	81,52	0,77
12.000	84,79	0,81	60,80	0,52	91,43	0,89	90,70	0,88	82,48	0,78
18.000	84,05	0,80	60,75	0,52	91,48	0,89	90,80	0,89	84,53	0,81
24.000	83,58	0,80	61,29	0,53	91,76	0,90	91,14	0,89	85,19	0,82
30.000	86,22	0,83	73,88	0,68	92,25	0,90	91,39	0,89	84,85	0,81

30.000 örneğe sahip ilgi alanı eğitim setine ait modellerin, başarımının algoritmalara göre nasıl değiştiğini ortaya koyabilmek için eğitim seti örnek sayıları

azdan çoğa doğru olacak şekilde ayarlanmıştır. Bunun bu şekilde yapılmasının amacı, algoritmalara göre örnek sayısına bağlı doygunluk noktasının ortaya koyulabilmesi içindir. Çizelge 5.5'te görüldüğü gibi eğitim seti; 600, 1.200, 1.800, 2.400, 3.000, 3.600, 4.200, 4.800, 5.400, 7.200, 9.000, 10.800, 12.000, 18.000, 24.000 ve 30.000 şeklinde farklı sayıda örnek içeren 16 farklı eğitim seti oluşturulmuştur. Kendi içinde dengeli bir şekilde ayarlanan eğitim setlerinden makine öğrenmesi algoritmaları ile modeller oluşturulmuştur.

İlgi alanı ile ilgili toplanan 30.000 örnekten oluşan eğitim setinde, herbir sınıf için 5.000 örneğin olduğu eğitim seti oluşturulmuştur. Bunun sebebi, eğitim setlerini kendi içinde dengeli hale getirip, yanlılıktan kaçınmaktır. Eğitim seti örnek sayısı 600 olan ilgi alanı seti; 100 adet bilim ve teknoloji, 100 adet spor, 100 adet eğitim, 100 adet siyaset, 100 adet ekonomi ve 100 adet ise diğer sınıfına ait örnek içermektedir.

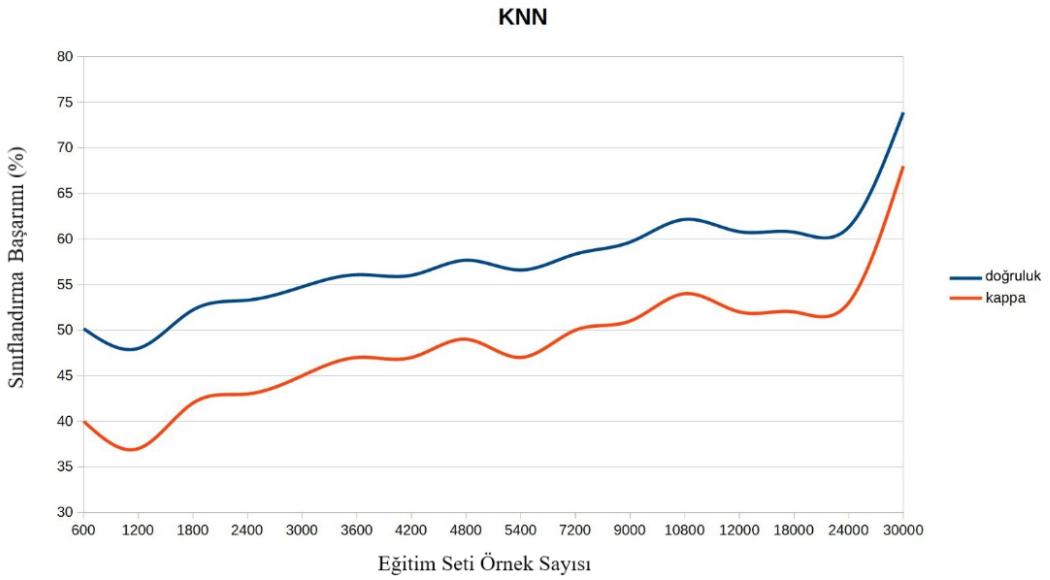
İlgi alanı tahmini için oluşturulan bu 16 farklı eğitim seti için makine öğrenmesi algoritmalarına bağlı olarak elde edilen sınıflandırma başarımları bu bölümde algoritma bazında ele alınmıştır. Bu eğitim setlerine ait modellerin örnek sayılarına bağlı sınıflandırma başarımları, doğruluk ve kappa değeri üzerinden Şekil 5.7'de Naive Bayes makine öğrenmesi algoritması için verilmiştir.



Şekil 5.7. 30.000 ilgi alanı eğitim seti NB doğruluk ve kappa değerleri

Şekil 5.7’de, 30.000 örneğe sahip ilgi alanı eğitim setinin Naive Bayes makine öğrenmesi algoritmasına ait doğruluk ve kappa değerlerine ait grafik gösterilmektedir. Grafik incelendiğinde, Naive Bayes algoritmasının 30.000 örnek ile doygunluğa ulaşmadığı ve bu sebeple, doğruluk ve kappa değerlerinin artmaya devam ettiği görülmektedir. Bundan dolayı, örnek sayısı arttırıldığında doğruluk ve kappa değerlerinde de artış olması beklenmektedir. Doğruluk ve kappa değerlerinin birlikte hareket etmesi, eğitim setinin uyumluluğunu göstermektedir.

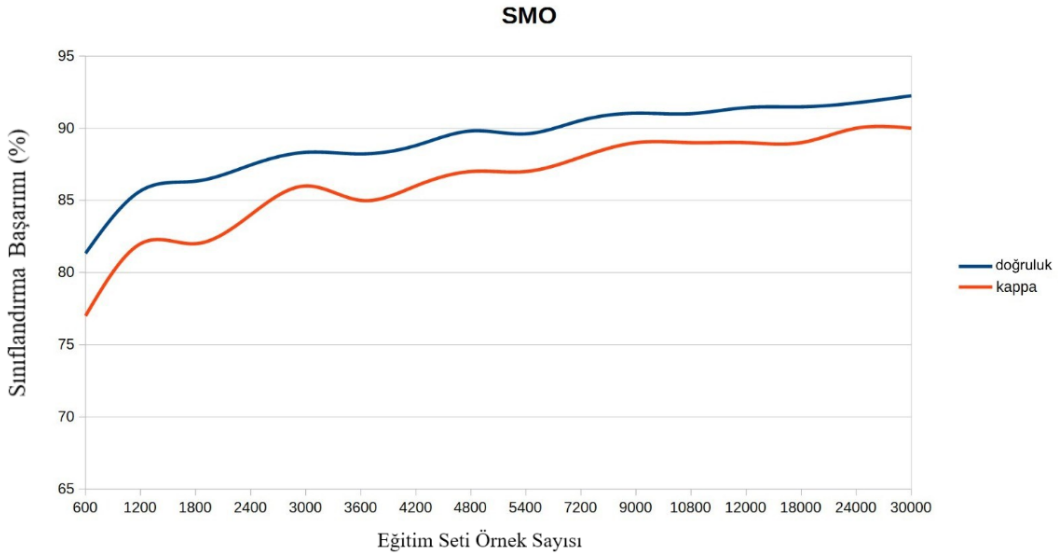
İlgi alanı tahmini için oluşturulan 16 farklı eğitim seti için makine öğrenmesi algoritmalarına bağlı olarak elde edilen sınıflandırma başarımları doğruluk ve kappa değerleri Şekil 5.8’de KNN makine öğrenmesi algoritması için verilmiştir.



Şekil 5.8. 30.000 ilgi alanı eğitim seti KNN doğruluk ve kappa değerleri

Şekil 5.8’de, benzer şekilde 30.000 örnek içeren eğitim setinin KNN makine öğrenmesi algoritması için doğruluk ve kappa değerlerine ait grafik gösterilmektedir. KNN algoritmasının az sayıda örnek içeren setlerde örnek sayısının yetersiz kalması nedeniyle doğruluk ve kappa değerlerinin düşük olduğu görülmektedir. Benzer şekilde bu algoritma için de örnek sayısı arttırılır ise başarımın artacağı söylenebilir.

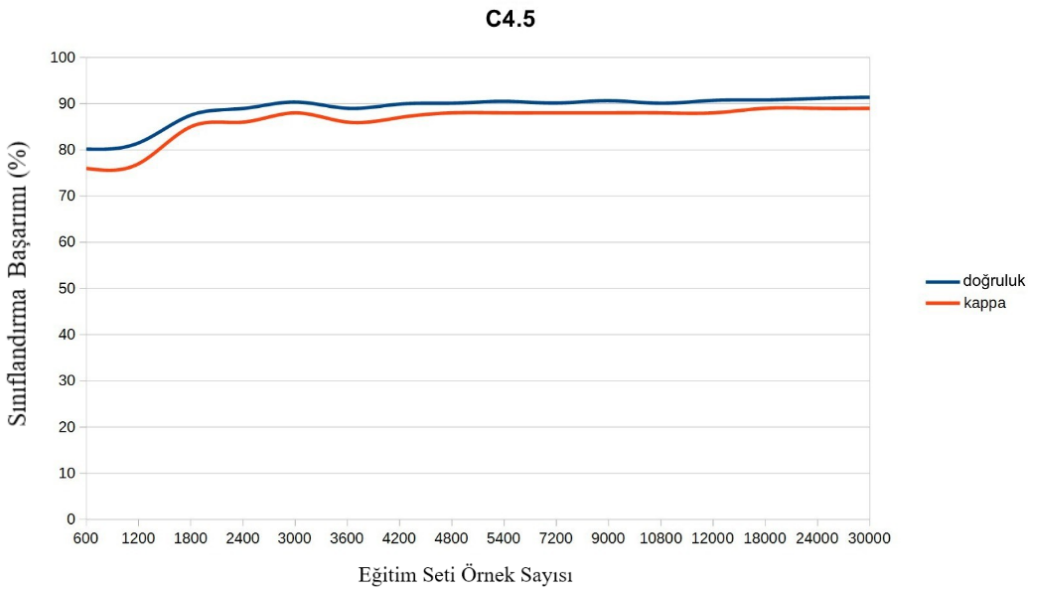
SMO makine öğrenmesi algoritması için belirtilen eğitim setine bağlı olarak elde edilen sınıflandırma başarımları doğruluk ve kappa değerleri Şekil 5.9'daki grafikte verilmiştir.



.Şekil 5.9. 30.000 ilgi alanı eğitim seti SMO doğruluk ve kappa değerleri

SMO makine öğrenmesi algoritmasının doğruluk değeri ve kappa değerinin eğitim seti örnek sayısına paralel şekilde artışta olduğu görülmektedir. Bu algoritmanın önemli özelliklerinden biri, az örnek sayısı ile yüksek sınıflandırma başırımı elde edebilmesidir. Grafiğe bakarak eğitim setinin örnek sayısı daha da arttırıldığı durumda, doğruluk ve kappa değerlerinin de artacağı öngörülmektedir.

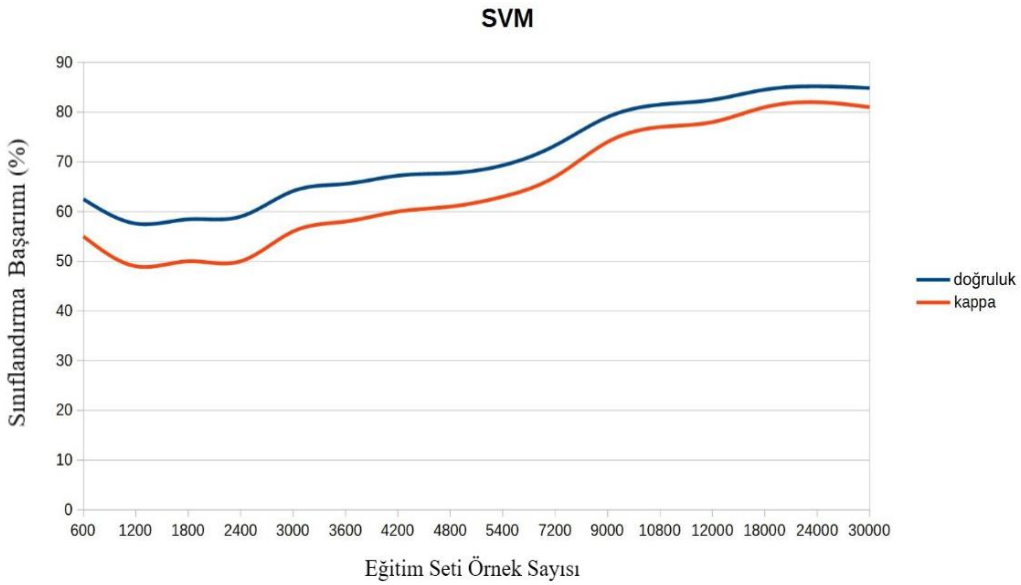
Makine öğrenmesi algoritmalarından C4.5 karar ağacı algoritması için 30.000 örnek içeren eğitim setine bağlı olarak elde edilen sınıflandırma başarımlarına ait doğruluk ve kappa değerleri için oluşturulan grafik Şekil 5.10'de verilmiştir.



Şekil 5.10. 30.000 ilgi alanı eğitim seti C4.5 doğruluk ve kappa değerleri

Grafikte verilen değerler incelendiğinde, eğitim seti örnek sayısı yaklaşık 3.000 olduğunda, doygunluğa ulaştığı ve 30.000 örnek sayısına kadar arttırıldığında ise bu örnek artışının, doğruluk ve kappa ölçütlerine önemli bir etki etmediği görülmektedir.

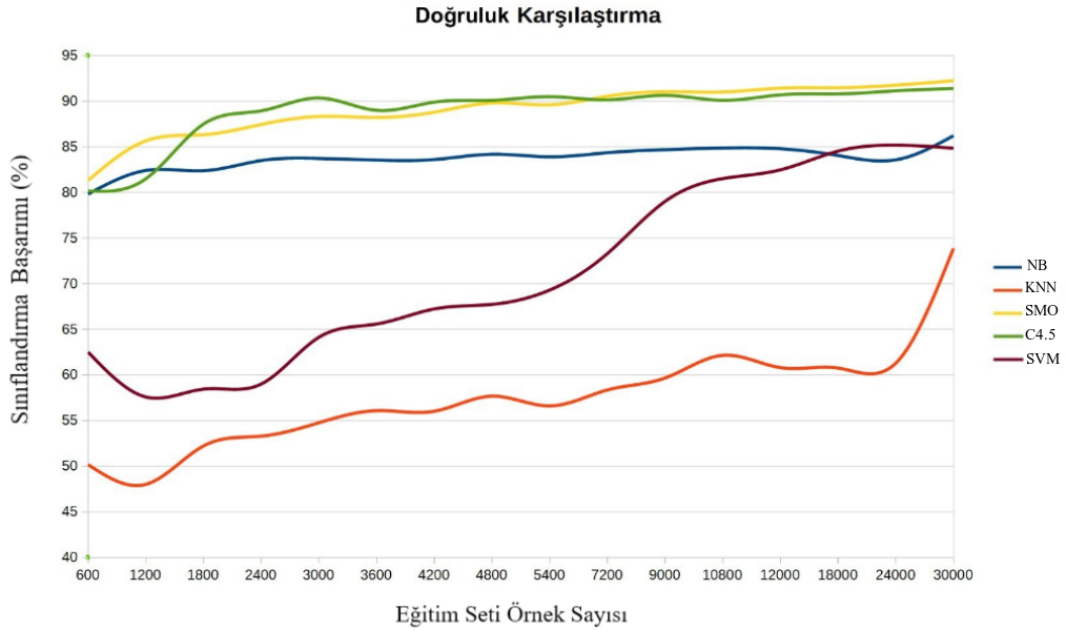
İlgi alanı tahmini için oluşturulan 30.000 örneğe sahip eğitim setinin alt seti olan 16 farklı eğitim seti için, SVM makine öğrenmesi algoritmasına bağlı olarak elde edilen sınıflandırma başarımları için doğruluk ve kappa değerleri Şekil 5.11'de verilmiştir.



Şekil 5.11. 30.000 ilgi alanı eğitim seti SVM doğruluk ve kappa değerleri

Şekil 5.11'de SVM algoritması için eğitim setinin doğruluk ve kappa değerleri 2.400 örnek sayısından sonra örnek sayısı arttıkça bu ölçütlerin de sürekli olarak arttığı görülmüştür. Eğitim setine 30.000 örnekten sonra örnek eklendiği zaman, doğruluk değerinin artıp artmayacağı, eğitim setinin örnek sayısına bağlı olarak değişeceği öngörülmektedir.

30.000 örneğe sahip ilgi alanı tahmini için oluşturulan 16 farklı eğitim seti sınıflandırma başarımı için yapılan denemelerden elde edilen sonuçlara ait grafik Şekil 5.12'de verilmiştir. Makine öğrenmesi algoritmalarına bağlı olarak elde edilen bu sonuçlar, algoritmaların sınıflandırma başarımlarını karşılaştırabilmek için doğruluk ölçütü değerleri üzerinden tek bir grafik üzerinde gösterilmiştir.



Şekil 5.12. 30.000 ilgi alanı eğitim seti doğruluk karşılaştırma

Şekil 5.12'deki grafik incelendiğinde, sınıflandırma başarımı doğruluk değerinin en yüksek olduğu algoritmanın SMO, başarım oranının en düşük olduğu algoritmanın ise KNN makine öğrenmesi algoritması olduğu görülmektedir. Fakat KNN makine öğrenmesi algoritmasının az sayıda örnek içeren eğitim setinde doğruluk başarım oranı düşük olmasına rağmen, örnek sayısı artması durumunda, doğruluk değerinin de arttığı görülmektedir. Eğer örnek sayısı arttırılmaya devam edilirse, doğruluk değerinin artıp artmayacağı 90.000 adet örnek içeren eğitim seti için yapılan denemelerden elde edilen sonuçlarda gösterilecektir.

90.000 örnek sayısına sahip ilgi alanı eğitim setine ait modellerin algoritmalara göre başarımının nasıl değiştiğini ortaya koyabilmek için eğitim seti örnek sayıları azdan çoğa doğru olacak şekilde ayarlanmıştır. Bunun bu şekilde yapılmasının amacı, örnek sayısına bağlı olarak makine öğrenmesi algoritmalarının doygunluk noktasının ortaya koyulabilmesi içindir. Çizelge 5.6'da görüldüğü gibi ilgi alanı için; 3.000, 6.000, 12.000, 18.000, 24.000, 30.000, 36.000, 42.000, 48.000, 54.000, 60.000, 66.000, 72.000, 78.000, 84.000 ve 90.000 örnek içeren 16 farklı eğitim seti oluşturulmuştur. Kendi içinde dengeli bir şekilde ayarlanan eğitim setlerinden makine öğrenmesi

algoritmaları ile modeller oluşturulmuştur. Çizelge 5.6'da oluşturulan 16 farklı eğitim setinin doğruluk ve kappa değerleri gösterilmiştir.

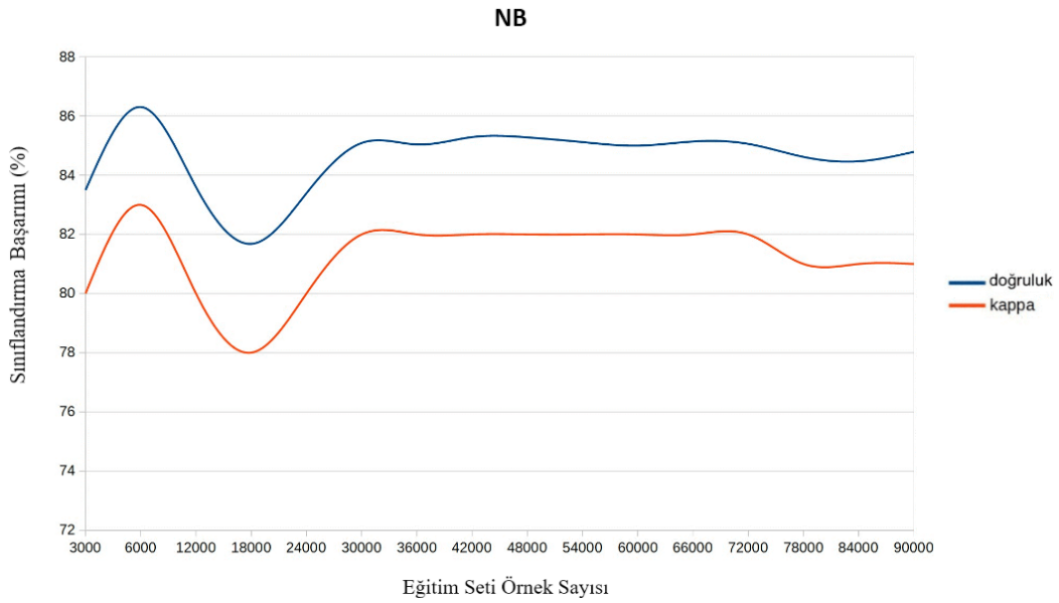
Çizelge 5.6. 90.000 ilgi alanı eğitim seti doğruluk ve kappa sonuçları

Eğitim Seti	NB		KNN		SMO		C4.5		SVM	
	Doğruluk (%)	Kappa	Doğruluk (%)	Kappa	Doğruluk (%)	Kappa	Doğruluk (%)	Kappa	Doğruluk (%)	Kappa
3.000	83,50	0,80	67,00	0,60	88,50	0,86	88,36	0,86	47,90	0,37
6.000	86,31	0,83	65,23	0,58	89,63	0,87	89,18	0,87	65,38	0,58
12.000	83,60	0,80	74,55	0,69	91,21	0,89	89,29	0,87	74,42	0,69
18.000	81,68	0,78	77,80	0,73	91,51	0,89	89,81	0,81	81,03	0,77
24.000	83,39	0,80	78,04	0,73	92,06	0,90	90,32	0,88	83,42	0,80
30.000	85,09	0,82	78,52	0,74	92,16	0,90	90,69	0,88	84,76	0,81
36.000	85,04	0,82	80,10	0,76	92,53	0,91	91,18	0,89	84,84	0,81
42.000	85,29	0,82	80,60	0,76	92,65	0,91	91,43	0,89	86,04	0,83
48.000	85,28	0,82	81,28	0,77	92,69	0,91	91,72	0,90	86,39	0,83
54.000	85,12	0,82	80,75	0,76	92,54	0,91	91,71	0,90	86,93	0,84
60.000	85,00	0,82	80,86	0,77	92,49	0,90	92,06	0,90	87,64	0,85
66.000	85,14	0,82	81,39	0,77	92,69	0,91	92,20	0,90	88,73	0,86
72.000	85,06	0,82	81,56	0,77	92,74	0,91	92,17	0,90	89,08	0,86
78.000	84,62	0,81	81,72	0,78	92,94	0,91	91,92	0,90	89,26	0,87
84.000	84,47	0,81	81,66	0,77	93,03	0,91	92,11	0,90	89,69	0,87
90.000	84,79	0,81	81,90	0,78	93,18	0,91	92,14	0,90	90,07	0,88

30.000 örneğe sahip olan eğitim setinde, her bir sınıf için 5.000 örnek bulunuyorken 90.000 örneğe sahip olan eğitim setinde ise, her bir sınıfa ait 15.000 örnek bulunmaktadır. 30.000 örneğe sahip eğitim setinin, 90.000 örnek sayısına

arttırılmasıyla, doğruluk ve kappa ölçütü değerlerindeki değişim üzerinden algoritmaların sınıflandırma başarımlarının değişimi gözlemlenmiştir.

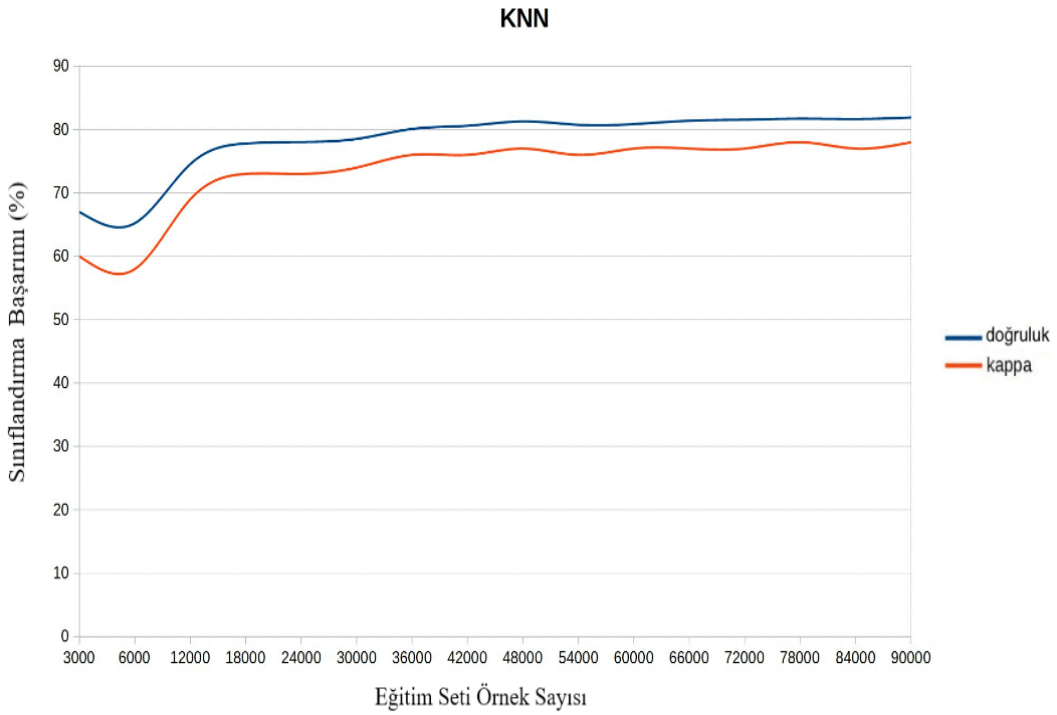
90.000 örnek sayısına sahip ilgi alanı tahmini için oluşturulan 16 farklı eğitim seti, makine öğrenmesi algoritmalarına bağlı olarak elde edilen sınıflandırma başarımları, algoritma bazında ele alınmıştır. Bu eğitim setlerine ait modellerin örnek sayılarına bağlı sınıflandırma başarımları, doğruluk ve kappa değerleri üzerinden değerlendirilmiştir. Şekil 5.13'de Naive Bayes makine öğrenmesi algoritması için grafik verilmiştir.



Şekil 5.13. 90.000 ilgi alanı eğitim seti NB doğruluk ve kappa değerleri

Naive Bayes makine öğrenmesi algoritmasında 6.000 ile 18.000 örneğe sahip eğitim setleri arası doğruluk değerinde, %2 oranında bir düşüş yaşanmıştır. Bazen eğitim setinin örnek sayısı arttırıldığında doğruluk değeri düşebilmektedir. Bunun nedeni, eğitim setine ait örneklerin bazen birbirleri ile uyuşmamasından kaynaklanmasıdır. Fakat, grafik incelendiğinde 30.000 örnekten sonra Naive Bayes makine öğrenmesi algoritmasının, doygunluk seviyesine ulaştığı görülmektedir.

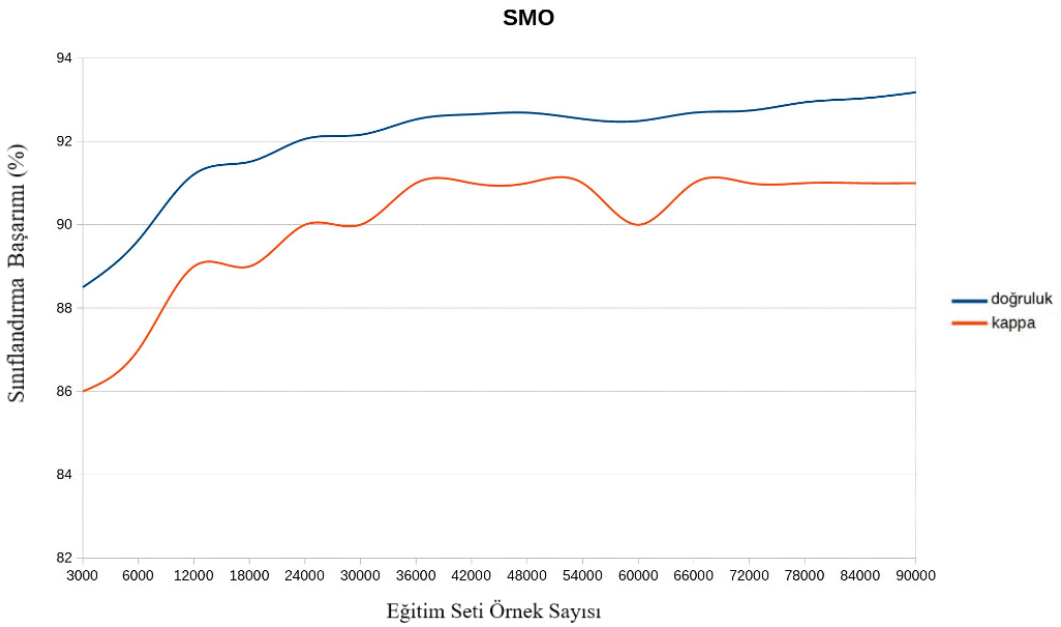
90.000 örneğe kadar eğitim alt setlerine ait modellerin, örnek sayılarına bağlı sınıflandırma başarımları, doğruluk ve kappa değerleri üzerinden Şekil 5.14'de KNN makine öğrenmesi algoritması için verilmiştir.



Şekil 5.14. 90.000 ilgi alanı eğitim seti KNN doğruluk ve kappa değerleri

KNN makine öğrenmesi algoritmasının Şekil 5.14'te gösterilen sonuçları, yaklaşık 18.000 örneğe ait eğitim setinden sonra doygunluğa ulaştığı görülmektedir. 18.000 örnek sayısından sonra örnek sayısının arttırılması, sınıflandırma başarımına yani doğruluk ve kappa değerlerine çok küçük katkı sağladığı söylenebilir.

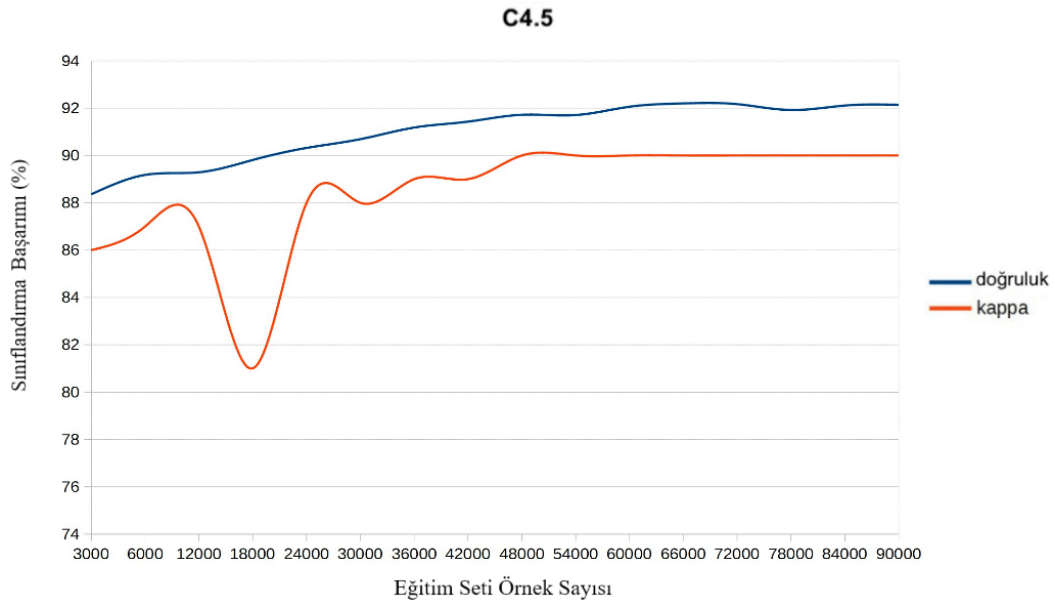
İlgi alanı eğitim setleri için örnek sayılarına bağlı sınıflandırma başarımları, doğruluk ve kappa değerleri üzerinden Şekil 5.15'de SMO makine öğrenmesi algoritması için verilmiştir.



Şekil 5.15. 90.000 ilgi alanı eğitim seti SMO doğruluk ve kappa değerleri

Grafik incelendiğinde, makine öğrenmesi algoritmasının eğitimi, kappa değeri bakımından tam bir uyum göstermemiştir. Sürekli bir artış ve azalış içerisindedir. Fakat 72.000 örnek sayısına ait eğitim setinden sonra örnek sayısının arttırılması durumunda doğruluk ve kappa değerlerinin oransal olarak düzenli hareket ettiği görülmektedir. 90.000 örneğe ait eğitim setinden sonra, örnek sayısı arttırıldığında doğruluk ve kappa değerlerinin başarım sonuçlarının oransal olarak uyumlu artacağı öngörülmektedir.

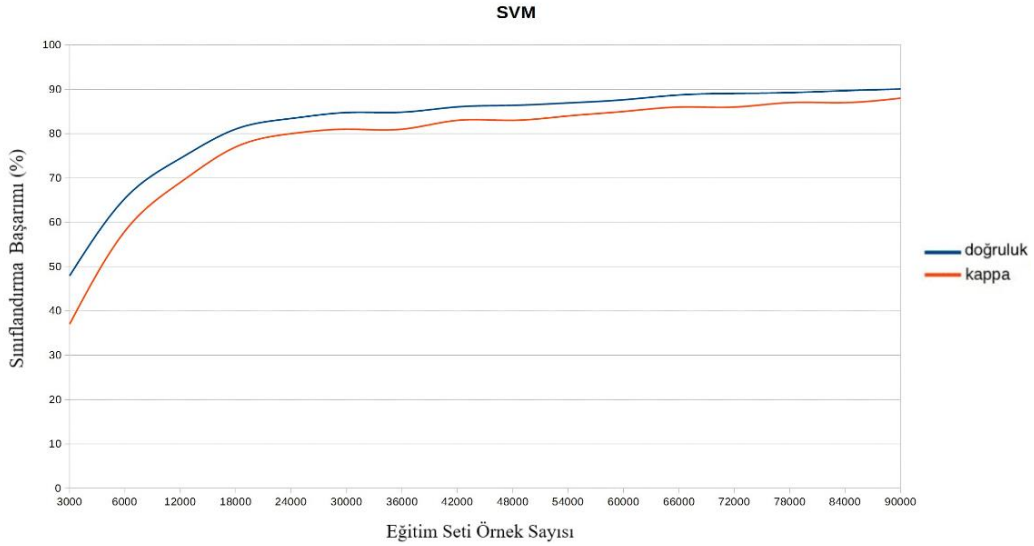
Benzer şekilde ilgi alanı eğitim setleri için C4.5 makine öğrenmesi algoritması ile yapılan denemelerden elde edilen doğruluk ve kappa ölçütlerine ait grafik Şekil 5.16'da verilmiştir.



Şekil 5.16. 90.000 ilgi alanı eğitim seti C4.5 doğruluk ve kappa değerleri

C4.5 karar ağacı makine öğrenmesi algoritması, 12.000 ve 25.000 örneğe ait eğitim setleri arasında doğruluk ve kappa değerlerinde uyuşmazlık olduğu görülmektedir. Bunun nedeni bu eğitim setleri arasında örneklerin birbirleri ile uyuşmadığı düşünülmektedir. Fakat örnek sayısı arttırıldığı zaman, kappa değerinin doğruluk değeri ile oransal olarak beraber hareket ettiği görülmektedir.

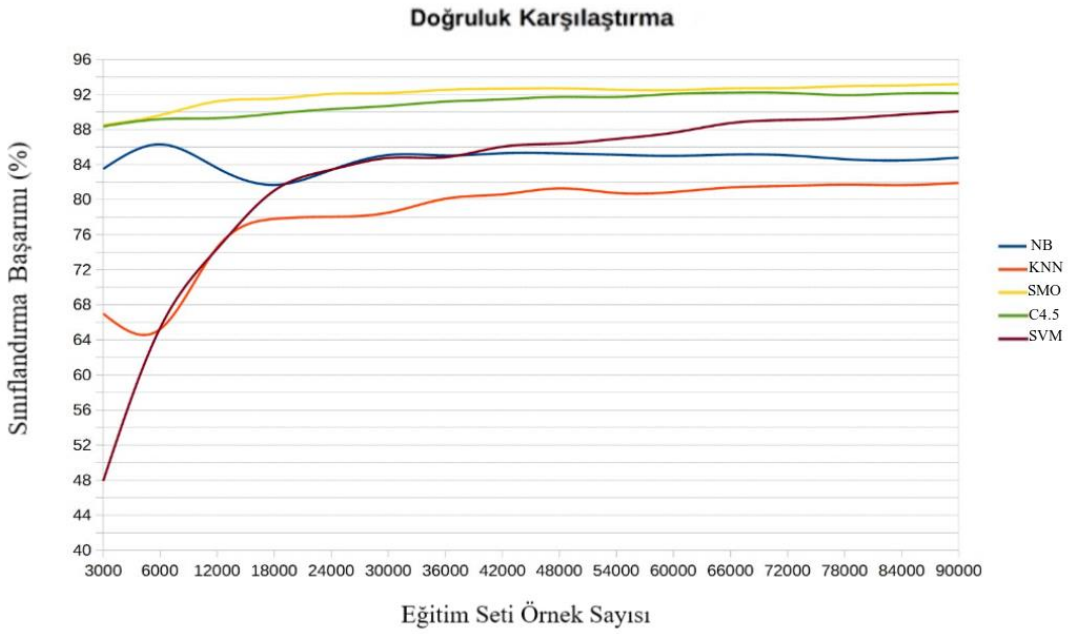
Şekil 5.17'deki grafikte ise SVM makine öğrenmesi algoritmasına ait elde edilen sonuçlar aynı şekilde eğitim seti örnek sayısına bağlı olarak, doğruluk ve kappa değerleri için verilmiştir.



Şekil 5.17. 90.000 ilgi alanı eğitim seti SVM doğruluk ve kappa değerleri

SVM algoritmasında, 3.000 ile 24.000 örneğe sahip eğitim setleri arasında, doğruluk ve kappa değerlerinde artış olduğu farkedildiği gibi, 24.000 örneğe sahip eğitim setinden sonra örnek adedi arttırıldığında ise, sınıflandırma başarımının çok etkilenmediği görülmektedir.

90.000 örneğe sahip ilgi alanı tahmini için oluşturulan 16 farklı eğitim seti için makine öğrenmesi algoritmalarına bağlı olarak elde edilen sınıflandırma başarımlarına ait doğruluk değerleri karşılaştırması Şekil 5.18'de verilmiştir.



Şekil 5.18. 90.000 ilgi alanı eğitim seti doğruluk karşılaştırma

Şekil 5.18'de tüm eğitim setleri için Naive Bayes, SMO, SVM, KNN ve C4.5 karar ağacı makine öğrenmesi algoritmalarına ait modellerin sınıflandırma doğruluk değerlerinden elde edilen grafik verilmiştir. Grafiğe göre, doğruluk başarım oranının en iyi sonucu verdiği Makine Öğrenimi algoritması SMO olduğu görülmektedir. Doğruluk başarım oranının en düşük olduğu algoritma ise KNN olduğu görülmüştür. KNN makine öğrenmesi algoritmasının doğruluk başarımı, düşük örnek sayısına ait eğitim setlerinde, SVM makine öğrenmesi algoritmasından daha yüksek başarım göstermesine rağmen, örnek sayısı arttığında SVM makine öğrenmesi algoritmasının başarım oranının daha yüksek olduğu görülmektedir. Benzer şekilde düşük örnek sayısına ait eğitim setlerinde, Naive Bayes makine öğrenmesi algoritmasının doğruluk başarımının, SVM makine öğrenmesi algoritmasından daha yüksek olduğu görülmektedir. Fakat örnek sayısı arttırıldığında, SVM makine öğrenmesi algoritmasının doğruluk başarımının Naive Bayes makine öğrenmesi algoritmasından daha etkin olduğu görülmektedir. Eğitim setindeki örnek sayısı arttırıldığında, SVM makine öğrenmesi algoritmasının ise, C4.5 karar ağacı algoritmasından daha etkin olabileceği öngörülmektedir.

Çizelge 5.7’de, 30.000 ve 90.000 örnek sayısına sahip ilgi alanı eğitim setlerinin, çalışmada kullanılan makine öğrenmesi algoritmaları için sınıflandırma başarımları doğruluk ölçüt değerleri verilmiştir.

Çizelge 5.7. 30.000 ve 90.000 örneğe sahip eğitim setlerinin doğruluk karşılaştırması

Makine Öğrenmesi Algoritmaları	30.000 Doğruluk (%)	90.000 Doğruluk (%)
NB	86,22	84,79
KNN	73,78	81,90
SMO	92,25	93,18
C4.5	91,39	92,14
SVM	84,85	90,07

Çizelge 5.7’de verilen sonuçlar değerlendirildiğinde:

Naive Bayes makine öğrenmesi algoritması doğruluk değerlerine göre 30.000 örnekten 90.000 örneğe çıkıldığında, yaklaşık %2 doğruluk başarımı düşüşü yaşamıştır. Örnek sayısının arttırılması, Naive Bayes makine öğrenmesi algoritmasının doğruluk başarımını negatif yönde etkilemiştir. Gelecek çalışmalar için eğitim setinin örnek sayısı arttırılıp, doğruluk başarımlarındaki değişimler gözlemlenebilir.

Benzer şekilde KNN makine öğrenmesi algoritmasında, yaklaşık %8 oranında doğruluk başarımı artışı görülmektedir. Bunun anlamı, eğitim setinin örnek sayısının arttırılması, KNN makine öğrenmesi algoritmasını olumlu yönde etkilemiştir. Eğitim setinin örnek sayısının arttırılması durumunda doğruluk değerinde artış olması öngörülmektedir.

Aynı şekilde, SMO makine öğrenmesi algoritmasında sınıflandırma doğruluk başarımı açısından yaklaşık %1 oranında artış olmuştur. Eğitim seti örnek sayısı artması doğruluk başarım oranını negatif yönde etkilemediği görülmektedir.

C4.5 karar ağacı makine öğrenmesi algoritması sınıflandırma başarımı sonucu, SMO makine öğrenmesi algoritması sonucuna benzer şekilde dikkate değer değişim gözlemlenmemiştir.

SVM makine öğrenmesi algoritması, örnek sayısı 30.000'den 90.000'e çıkarıldığında yaklaşık %6 oranında doğruluk başarımı artışı görülmektedir. Eğitim setinin örnek sayısı arttırılması, SVM makine öğrenmesi algoritmasını olumlu yönde etkilemiştir.

5.4. Sonuç ve Değerlendirme

Bu çalışmada, Twitter'dan, kullanıcıların cinsiyetlerini tahmin etmek ve paylaşılan içeriklerin ne ile ilgili olduğunu belirlemek için sırası ile 171.200 ve 90.000 örnek sayısından oluşan eğitim setleri ve bunların alt eğitim setleri oluşturulmuştur. Bu eğitim setleri, temizlenip, önişlemden geçirildikten sonra alt eğitim setlerine bölünmüştür. Cinsiyet eğitim seti kendi içerisinde 18, ilgi alanı eğitim seti ise kendi içerisinde 16 ayrı alt eğitim setlerine bölünüp, makine öğrenmesi algoritmaları ile doğruluk ve kappa değerlerindeki değişimler gözlemlenmiştir. Eğitim setleri; Naive bayes, KNN, SVM, SMO ve C4.5 karar ağacı olmak üzere, toplamda 5 adet makine öğrenmesi algoritması ile test edilmiştir.

Cinsiyet eğitim setinde en iyi sonucu veren makine öğrenmesi algoritması %99,9 doğruluk değeri ile SMO makine öğrenmesi algoritması olmuştur. %81,9 orana sahip doğruluk başarımının en düşük olduğu algoritma ise SVM makine öğrenmesi algoritması olmuştur. Fakat 171.200 örnek sayısına sahip eğitim setinin örnek sayısı arttırılması durumunda, SVM makine öğrenmesi algoritmasının doğruluk ve kappa değerlerinin de artacağı öngörülmektedir. İlgi alanı eğitim seti incelendiğinde, en iyi sonucu veren makine öğrenmesi algoritması, %93,18 doğruluk başarım değeri ile SMO makine öğrenmesi algoritması olmuştur. %81,90 doğruluk başarım değeri ile KNN makine öğrenmesi algoritmasının en düşük sonucu veren algoritma olduğu görülmüştür.

Kullanılan bu makine öğrenmesi algoritmaların hangisinin kullanılan eğitim setlerine göre en iyi olduğunu belirlemek, ancak daha çok eğitim setindeki örnek sayısının arttırılıp, doygunluğa ulaştıkları noktaları belirlemek ile mümkün olmaktadır. Bazı makine öğrenmesi algoritmaların, aynı eğitim setindeki doğruluk ve kappa başarımlarının iyi, bazılarının ise doğruluk ve kappa başarımlarının kötü performans göstermelerinin sebebi, algoritmaların eğitim setine bağlı olmasından kaynaklanmasıdır. Çünkü SVM gibi bazı makine öğrenmesi algoritmaları temelinde metin bazlı, Naive Bayes gibi bazıları ise nümerik bazlı eğitim setlerinde doğruluk başarımları açısından iyi performans göstermektedir. Bu çalışmada, SMO ve C4.5 gibi makine öğrenmesi algoritmalarında örnek doygunluğunun oluştuğu görülmektedir. Bunun nedeni, algoritmanın eğitim setine bağlı olarak bir kural varlığını ifade etmektedir. Gelecek çalışmalar için eğitim setlerinin örnek sayıları arttırılıp herbir makine öğrenmesi algoritmasının doygunluk oranı belirlenebilir. Aynı zamanda İlgi alanı için ise daha çok sınıf eklenip farklı sonuçlar gözlemlenebilir. Benzer şekilde gelecek çalışmalar için, kullanıcıların profil resimlerinin de işlenmesi ve bu sayede farklı bir örüntü ortaya çıkarılıp eğitim setlerinden makine öğrenmesi algoritmaları ile oluşan modellerin başarım oranları arttırılabilir.

Elde edilen cinsiyet ve ilgi alanı eğitim setleri kullanılarak oluşturulan bu modeller, geliştirilecek olan uygulamalarda tweet sahibinin cinsiyeti ve ilgi alanı hakkında tahminlerinin yapılıp, kurum ve kuruluşlar için hedef kitlelerin belirlenmesi için kullanılabilir. Aynı zamanda bu kurum ve kuruluşların, hedef kitlelere özel olarak reklam çıktısı üretip doğru kullanıcıya ulaşılabilirler.

KAYNAKLAR

Andrews, R. (2009, 27 March). Twitter Brings Back UK SMS; Vodafone First, Others To Follow. www.theguardian.com

Bhattacharya, P., Zafar, B., Ganguly, N., Ghosh, S., & Gummadi, P.K. (2014). Inferring User Interest in Twitter Social Network, *RecSys'14: Proceedings of the 8th ACM Conference on Recommender Systems*, 357-360, https://doi.org/10.1145/2645710.2645765

Bishop, C.M. (2006). *Pattern Recognition and Machine Learning*. New York: Springer.

Cox, D. (2015). *Surveys – what makes them fail or succeed?*. 17 Şubat 2019 tarihinde https://www.poppulo.com adresinden erişildi.

Cohen, J. (1960). A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement, 20(1),* 37-46.

Deitrick, W., Miller, Z., Valyou, B., Dickinson, B., Munson, T., & Hu, W. (2012). Gender Identification on Twitter Using the Modified Balanced Winnow, *Communications and Networks, 4(3),* 189-195.

Edgar, T.W., & Manz, D.O. (2017). *Research Methods for Cyber Security*. United States: Cambridge.

Giachanou, A., & Crestani, F. (2016). Like It or Not: A Survey of Twitter Sentiment Analysis Methods, *ACM Computing Surveys, 49(2),* 1-41. doi: 10.1145/2938640

Horton, P., & Nakai, K. (1996). A Probabilistic Classification System for Predicting the Cellular Localization Sites of Proteins. *Proceedings of International Conference on Intelligent Systems for Molecular Biology*, 109-115.

İsimler Listesi. (2011). 16 Ocak 2019 tarihinde https://gist.github.com/ismailbaskin/1325813 adresinden erişildi.

Kotsiantis, S. B., Zaharakis, I. D., & Pintelas, P. E. (2007). Machine Learning: A Review of Classification and Combining Techniques. *Artificial Intelligence Review, 26(3),* 159-190. doi: 10.1007/s10462-007-9052-3

Landis, J.R., & Koch, G.G. (1977). The Measurement of Observer Agreement for Categorical Data. *Biometrics, 33(1),* 159-174. doi: 10.2307/2529310

Lv, T., Yan, P., & He, W. (2018). Survey on JSON Data Modelling, *Journal of Physics Conference Series. 1069(1),* doi: 10.1088/1742-6596/1069/1/012101.

Sayyadiharikandeh, M., Ciampaglia, G.L., & Flammini, A. (2016). Cross-domain Gender Detection in Twitter. *Proceedings of the Workshop on Computational Approaches to Social Modeling*. doi: 10.1007/978-3-319-31232-3_97

Simon, P. (2013). *Too Big to Ignore*: *The Business Case for Big Data*. New Jersey: Wiley.

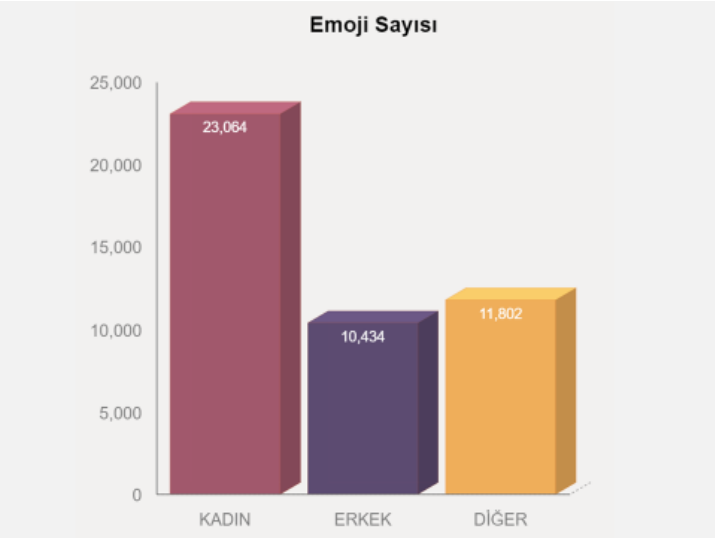
Şeker, Ş.E. (2013). *İş Zekâsı ve Veri Madenciliği*. İstanbul: Cinius.

Weka. (2008). 15 Nisan 2019 tarihinde https://wiki.pentaho.com/display adresinden erişildi.

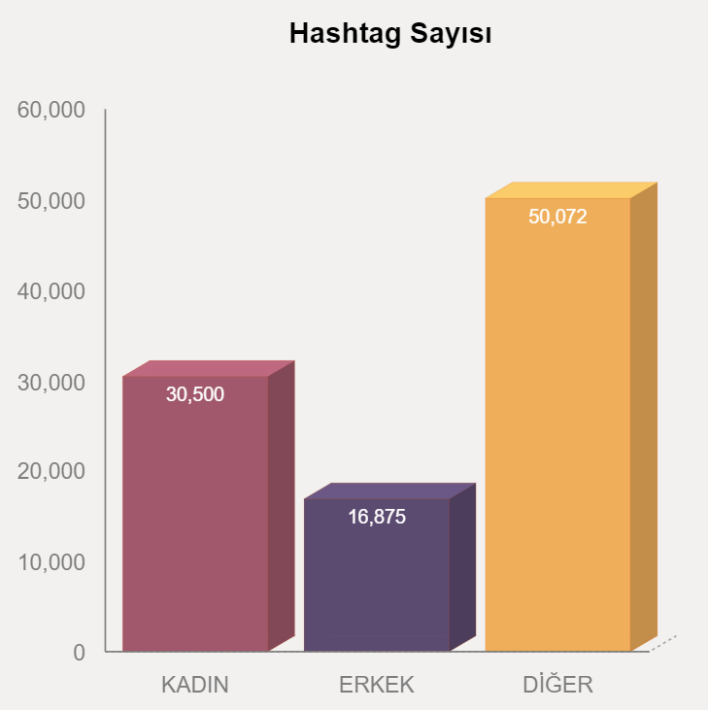
Weka. (2007). 12 Mayıs 2019 tarihinde https://waikato.github.io/weka-wiki/packages/ adresinden erişildi.

Wood-Doughty, Z., Andrews, N., Marvin, R., & Dredze, M. (2018). Predicting Twitter User Demographics from Names Alone. *Proceedings of the Second Workshop on Computational Modeling of People's Opinions, Personality and Emotions in Social Media,* 105-111, doi: 10.18653/v1/W18-1114

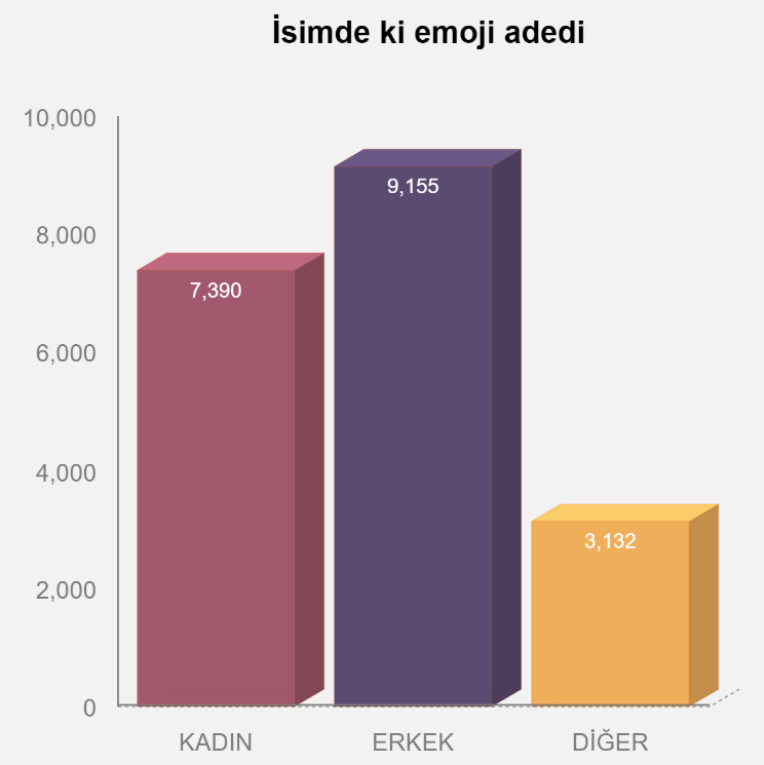
EKLER



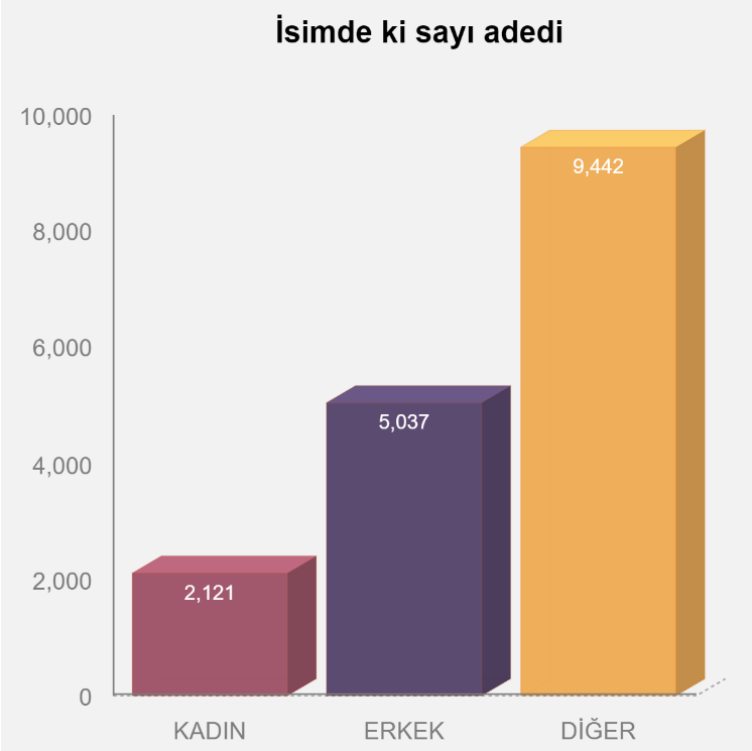
Cinsiyet eğitim setinde kullanılan emoji sayısı



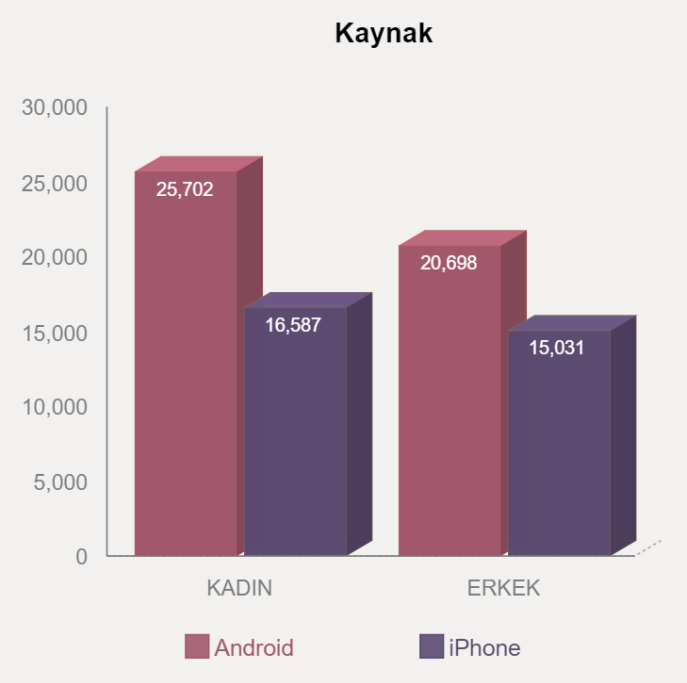
Cinsiyet eğitim setinde kullanılan hashtag sayısı



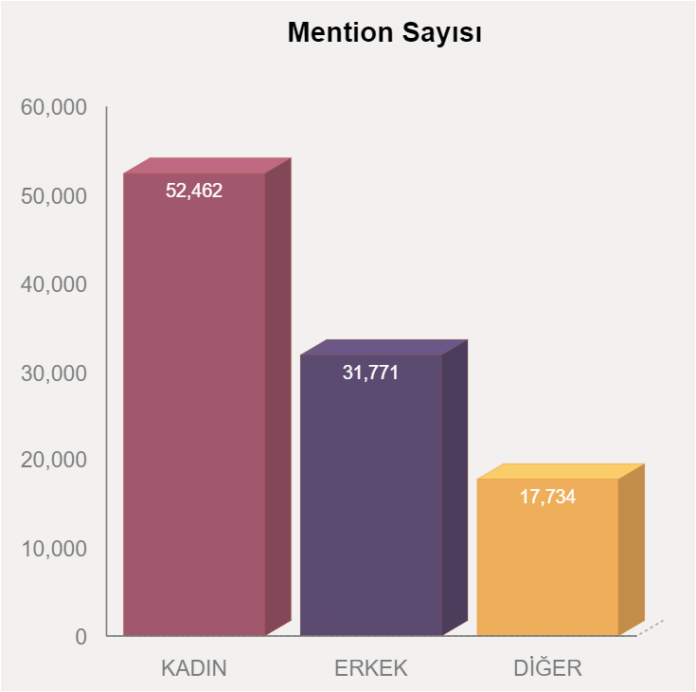
Cinsiyet eğitim setinde kullanıcıların isimlerinde kullandıkları emoji sayısı



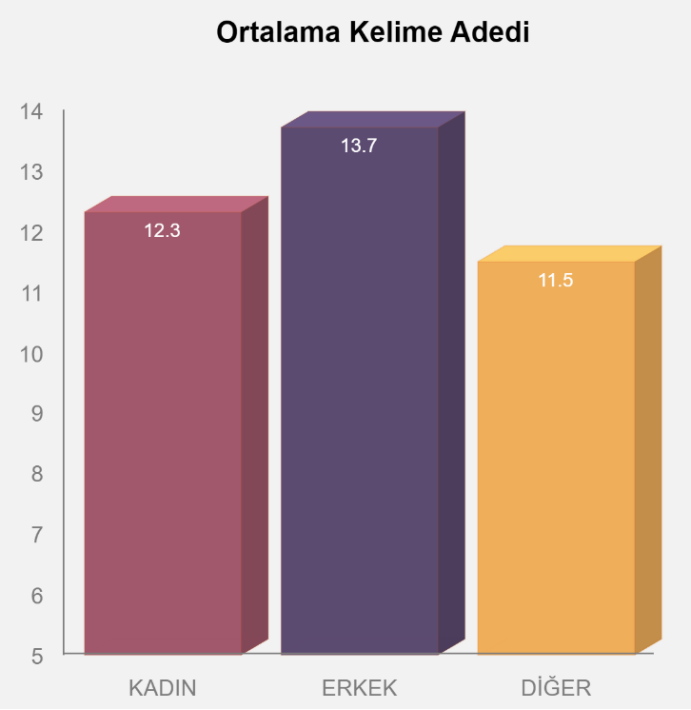
Cinsiyet eğitim setinde kullanıcıların isimlerinde kullandıkları sayı adedi



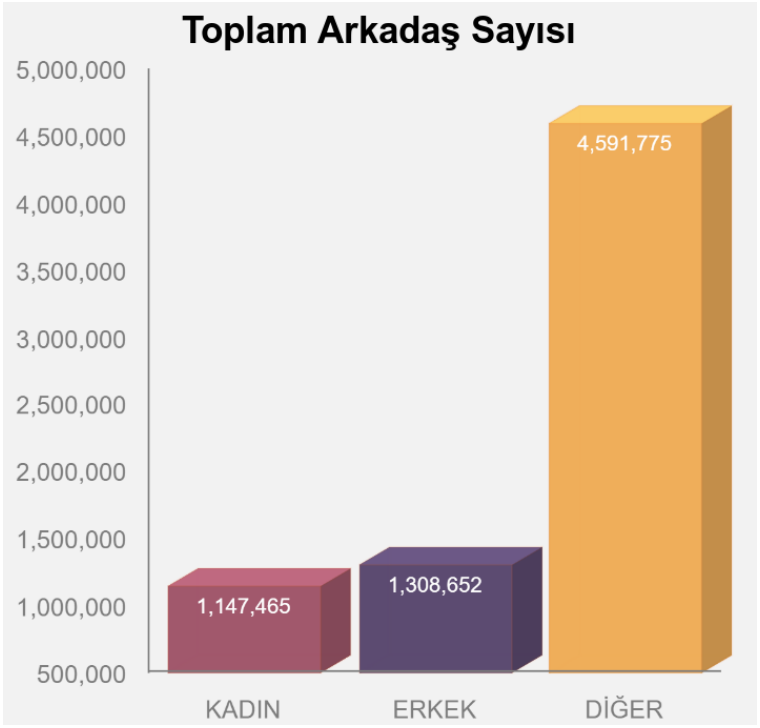
Cinsiyet eğitim setinde kullanıcıların tweetleri paylaştığı kaynak



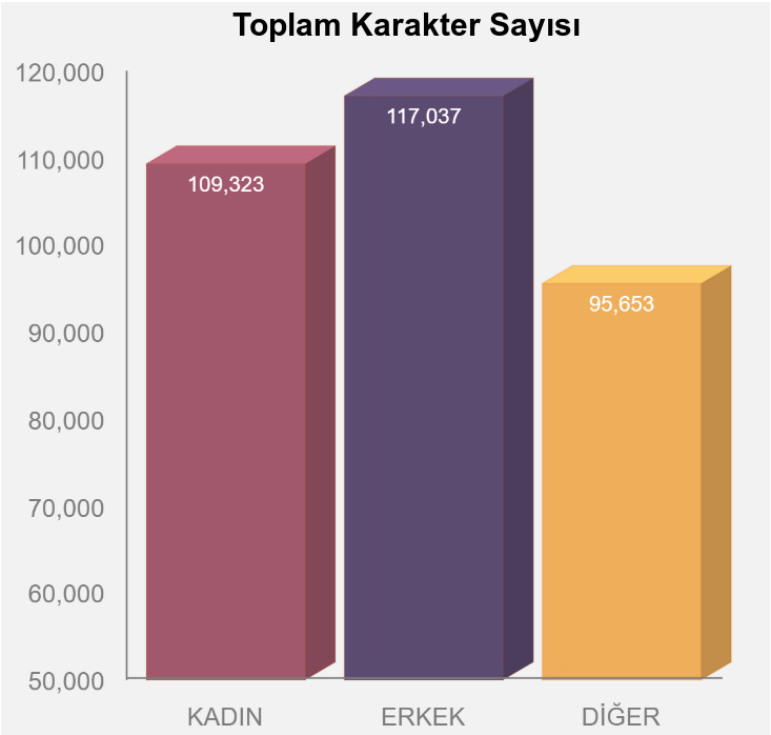
Cinsiyet eğitim setinde kullanıcıların tweetlerde kullandığı mention sayısı



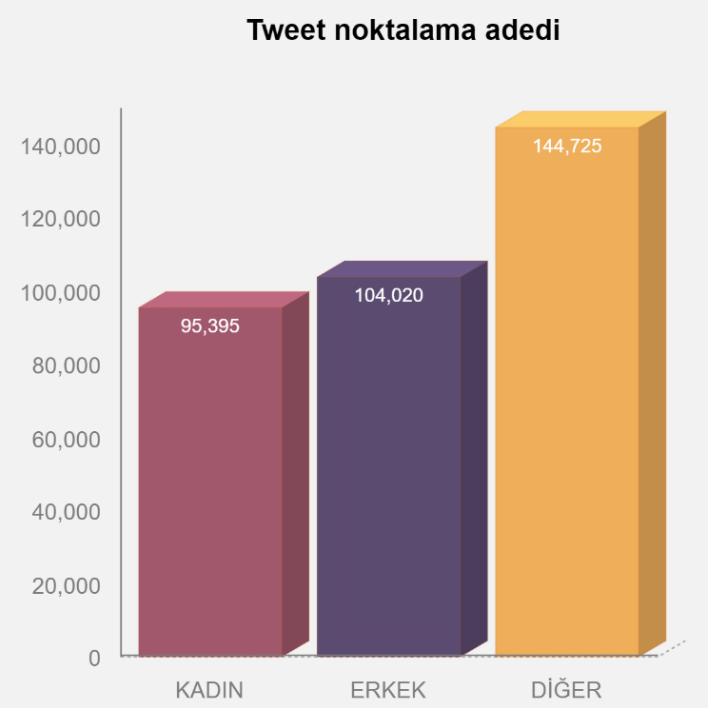
Cinsiyet eğitim setinde kullanıcıların kullandığı ortalama kelime adedi



Cinsiyet eğitim setinde kullanıcıların ortalama arkadaş sayısı



Cinsiyet eğitim setinde kullanıcıların ortalama kullandığı karakter sayısı



Cinsiyet eğitim setinde kullanıcıların tweet içinde kullandığı noktalama adedi

ÖZGEÇMİŞ

Kimlik Bilgileri

Adı Soyadı: Enes GÜNÇE

Doğum Yeri/Yılı: Bingöl, 01/08/1994

E-posta: enesguncee@gmail.com

Eğitim

Lise: 75. Yıl Cumhuriyet Anadolu Lisesi (2007-2011)

Lisans: Trakya Üniversitesi, Bilgisayar Mühendisliği (2012-2016)

Y. Lisans: Trakya Üniversitesi, Bilgisayar Mühendisliği (2017-2021)

Yabancı Dil ve Düzeyi

İngilizce, İyi seviyede

İş Deneyimi

Indisera, Veri Bilimci, 2016-2017

Ronwell Digital Bilgi Teknolojisi ve Hizmetleri, Yazılım Test Otomasyon Mühendisi, 2018-2020

Tezden Üretilmiş Yayınlar

1. Günçe, E, Carus, A. (2020). Twitter Platformunda Makine Öğrenmesi Algoritmalarıyla Cinsiyet ve İlgi Analizi. *Avrupa Bilim ve Teknoloji Dergisi, Ejosat Özel Sayı 2020 (ISMSIT),* 187-194. DOI: 10.31590/ejosat.819722