T.C.
İSTANBUL KÜLTÜR ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

TÜRK DİLİ İÇİN ÇOKLU SINIFLANDIRICI YÖNTEMLER İLE DUYGU SINIFLANDIRMA

YÜKSEK LİSANS TEZİ

Mehmet NANĞIR

1009051003

Anabilim Dalı: Bilgisayar Mühendisliği

Programı: Bilgisayar Mühendisliği

Tez Danışmanı: Yrd. Doç. Dr. Çağatay ÇATAL

EYLÜL 2013

T.C.

İSTANBUL KÜLTÜR ÜNİVERSİTESİ

FEN BİLİMLERİ ENSTİTÜSÜ

TÜRK DİLİ İÇİN ÇOKLU SINIFLANDIRICI YÖNTEMLER İLE DUYGU SINIFLANDIRMA

YÜKSEK LİSANS TEZİ

Mehmet NANĞIR

1009051003

Anabilim Dalı: Bilgisayar Mühendisliği

Programı: Bilgisayar Mühendisliği

Tez Danışmanı: Yrd. Doç. Dr. Çağatay ÇATAL

EYLÜL 2013

ABSTRACT

SENTIMENT CLASSIFICATION WITH MULTIPLE CLASSIFIER SYSTEMS FOR TURKISH LANGUAGE

Mehmet NANĞIR - 2013

Sentiment analysis is to obtain individual information and inferences using natural language processing methods from raw data sources.

User reviews are valuable resource for commercial, social, political analysis and text mining. Consumer reviews, book reviews, social media analysis, political research, news reviews, movie reviews and stock market predictions can be given as examples fort he research and analysis topics in sentiment analysis.

With the explosive growth of social media and internet, the value of personal reviews is increased. The effect of the internet for the commerce changed the brand-consumer relationship significantly. Positive and negative experiences are not only between brand and consumers, but also they spread rapidly to the social environment. Analysis and evaluation of this data began to offer more important benefits for individuals and companies.

There are several studies in this area in literature for English language. This field was not investigated so much for Turkish language and there is not enough number of research studies. According to our literature survey, we only reached two studies for Turkish language. First study used a machine learning algorithm for a specific type and focused on a domain including single dataset [19]. In the second study, several machine learning algorithms are tested on three datasets from different domains. 85% accuracy was obtained with Naive Bayes machine learning algorithm [20].

In this thesis, multiple classifier machine learning algorithms have been applied for Turkish Language on different domains. As distinct from existing studies, a novel multiple classifier system (MCS) was designed by using three high-performance machine learning algorithms all together. In addition to this novel MCS approach, performance was increased by performing parameter optimization of machine learning algorithms. With this new approach, previous accuracy rate was increased to 86.13% accuracy. This accuracy rate revealed that this approach improves the performance and can be used in many studies.

Key Words: Sentiment Classification, Turkish, Naive Bayes, Support Vector Machines, Decision Tree, Multiple Classifier Systems, Parameter Optimization, Machine Learning, Natural Language Processing, Data Mining, Weka

ÖZET

TÜRK DİLİ İÇİN ÇOKLU SINIFLANDIRICI YÖNTEMLER İLE DUYGU SINIFLANDIRMA

Mehmet NANĞIR - 2013

Duygu analizi, doğal dil işleme yöntemlerinin kullanılarak kaynaklarda yer alan ham veriden kişisel bilgi ve çıkarımların elde edilmesidir.

Kullanıcı yorumları; ticari, sosyal, siyasi analizler ve metin madenciliği için çok değerli bir kaynaktır. Duygu analizinin araştırma ve inceleme alanına giren konulara; tüketici yorumları, kitap yorumları, sosyal medya analizi, siyasi araştırmalar, haber yorumları, film değerlendirmeleri ve borsa tahminleri örnek olarak verilebilir.

Son zamanlarda internet ve sosyal medya kullanımının artması, kişisel değerlendirmeleri önemli bir konuma getirdi. Internet kullanımının ticarete etkisi, marka-tüketici ilişkisini de önemli ölçüde değiştirdi. Olumlu ve olumsuz deneyimler artık marka ile tüketici arasında kalmıyor, sosyal çevreye hızla yayılıyor. Bu verinin analizi ve değerlendirilmesi, gerek birey gerekse şirketler için gittikçe daha fazla önemli kazançlar sunmaya başladı.

Bu alanda genel olarak İngilizce için çeşitli çalışmalar literatürde mevcuttur. Bu konu, Türk dili için henüz derinlemesine incelenmemiş ve yeterli sayıda araştırmanın yapılmadığı bir konudur. Yapılan literatür taramasında, Türk dili için gerçekleştirilen sadece iki çalışmaya ulaşabildik. İlk çalışma, sadece bir alan üzerine yoğunlaşıp, tek tipte veri seti üzerinde sadece belirli tipte bir makine öğrenmesi algoritması kullanmıştır [19]. İkinci çalışmada ise üç farklı veri seti üzerinde birden fazla makine öğrenmesi tek tek denenmiş ve Naive Bayes isimli makine öğrenmesi yöntemi ile yaklaşık olarak % 85 doğruluk oranı elde edilmiştir [20].

Bu tez çalışması kapsamında, Türk dili için farklı veri kümeleri üzerinde çoklu sınıflandırıcı makine öğrenmesi algoritmaları uygulanmıştır. Daha önce uygulanan çalışmalardan farklı olarak, performansı yüksek üç tane makine öğrenmesi algoritması birlikte kullanılarak özgün bir çoklu sınıflandırıcı makine öğrenmesi algoritması tasarlanmıştır. Bu özgün sınıflandırıcı yaklaşımının yanı sıra, makine öğrenmesi algoritmalarının parametre optimizasyonu gerçekleştirilerek performans arttırılmıştır. Bu yeni yaklaşım sayesinde, daha önce tek sınıflandırıcı ile elde edilen doğruluk oranı % 86,13'lük bir doğruluk oranına yükseltilmiştir. Bu doğruluk oranı, yeni yaklaşımın performansı iyileştirdiğini ve birçok çalışmada kullanılabileceğini ortaya koymuştur.

Anahtar Kelimeler: Duygu Sınıflandırma, Türkçe, Naive Bayes, Karar Destek Makineleri, Karar Ağacı, Çoklu Sınıflandırıcı Sistemler, Parametre Optimizasyonu, Makine Öğrenmesi, Doğal Dil İşleme, Veri Madenciliği, Weka

İçindekiler

Kısaltmalar

ARFF	: Attribute Relation File Format
ÇO	: Çoğunluk Oylaması
DDİ	: Doğal Dil İşleme
DVM	: Destek Vektör Makineleri
KA	: Karar Ağacı
MÖ	: Makine Öğrenmesi
NB	: Naive Bayes
NBM	: Naive Bayes Multinomial
PO	: Parametre Optimizasyonu

Tablolar Listesi

Şekiller Listesi

1 Giriş

1.1 Duygu Analizi

Duygu analizi (sentiment analysis) ve düşünce madenciliği (opinion mining), yazı dilinden ve metinlerden insanların düşüncelerini, duygularını, değerlendirmelerini ve tutumlarını analiz eden bir doğal dil işleme alanıdır.

Duygu analizi; duygu analizi başta olmak üzere, düşünce madenciliği, düşünce çıkarımı, duygu madenciliği, etki analizi gibi isimlerle de adlandırılır. Endüstriyel uygulamalarda, daha yaygın olarak duygu analizi ifadesi kullanılmaktadır. Akademik çalışmalarda ise duygu analizi ve düşünce madenciliği isimlendirmeleri kullanılmaktadır [1].

Duygu analizi ismi ilk olarak Nasukawa ve arkadaşlarının [2] 2003 yılındaki çalışmasında kullanılmıştır. Düşünce madenciliği ismine ise ilk olarak 2003 yılında Dave ve arkadaşlarının [3] çalışmasında rastlanmaktadır. Bununla beraber, düşünce ve duygu kavramları daha erken çalışmalarda da görülmektedir [4, 5, 6, 7, 8, 9].

Duygu analizinin çalışma alanı; tüketici yorumları, kitap yorumları, sosyal medya analizi, siyasi araştırmalar, haber yorumları, film değerlendirmeleri ve borsa tahminleri olmak üzere insanların fikirlerinin yer aldığı geniş bir problem uzayına sahiptir.

1.2 Duygu Analizinin Düzeyleri

Duygu analizi çalışmaları, şu ana kadar yapılan araştırmalara bakıldığında üç düzeyde ele alınmaktadır. Bu düzeyler; belge düzeyinde duygu analizi, cümle düzeyinde duygu analizi ve varlık düzeyinde duygu analizi olarak sıralanabilir. Bu bölümün alt başlıklarında bu düzeyler açıklanmaktadır.

1.2.1 Belge Düzeyinde Duygu Analizi

Bu düzeydeki duygu analizi çalışmalarında, tüm belge bir bütün olarak ele alınmaktadır. Belgenin işlem sonucunda sadece tek bir sonucu oluşur. Belge içeriği olumlu ya da olumsuzdur, belge bazında sonucu tektir. Pang [6] ve Turney'in [8] çalışmalarında belge bazında duygu analizi çalışmaları yapılmıştır.

1.2.2 Cümle Düzeyinde Duygu Analizi

Bu duygu analizi düzeyinde; belgede ya da veri metninde yazan her bir cümle ayrı ayrı değerlendirilir ve cümle bazında olumlu, olumsuz ya da nötr sonuç alınabilir. Nötr düşünce, konu hakkında fikir elde edilemediği anlamına gelmektedir. Ana cümlelere göre karar oluşturulmaya çalışılmaktadır. Cümleler, her durumda tek bir karar cümlesinden oluşmaz. Bu cümlelere ek olarak, yan cümlecikler de içerebilir. Araştırmacılar yan cümleciklere göre karar oluşturmayı incelemişlerdir ancak şu anda yeterli düzeyde sonuç alınamamıştır [10].

1.2.3 Varlık ve Görünüm Düzeyinde Duygu Analizi

Bu düzey, diğer iki düzeye göre daha ayrıntılı bir duygu analizi işlemidir. Varlık düzeyi (entity level), ilk olarak özellik düzeyi olarak literatürde yerini almıştır [11]. Elde var olan varlığın, özellik ve detaylarına göre analizin yapıldığı bir işlemdir.

Örneğin; "Samsung Galaxy S3 ses kalitesi çok iyi ama batarya süresi çok kısadır" cümlesinde ana varlık Samsung Galaxy S3 model telefondur.

Analiz aşamasında telefonun cümle içindeki özellikleri çıkarılır. Bu cümlede telefona ait özellikler ses kalitesi ve batarya ömrüdür. Ses kalitesi olumlu, batarya ömrü ise olumsuzdur. Cümle görünüm düzeyi içerisinde bu detayda incelenir. Uygulanması belge ve cümle düzeyine göre daha zordur.

1.3 Düşünce Tanımı

Duygu analizi, çalışma alanı olarak düşünce ve görüşleri değerlendirir. Bu nedenle, bu bölümde düşüncenin formel tanımı ve düşünceyi oluşturan alt bileşenler açıklanmaktadır.

Düşünce; insanlar arasında, kişisel değerlendirme olarak adlandırılır. Düşüncenin genel tanımı bu şekilde verilirken, düşünceyi oluşturan bileşenlere de bir örnek üzerinden bakmakta fayda bulunmaktadır.

Örnek:

“ Mehmet NANGIR   26 ağustos 2013

(1) Altı ay önce HTC One cep telefonu satın aldım.
(2) Genel olarak telefonu sevdim.
(3) Resim kalitesi mükemmel.
(4) Batarya süresi oldukça uzun.
(5) Bunun yanında, eşim onun için bu telefonun çok büyük ebatlarda olduğunu düşünüyor”.

Genel olarak bakıldığında iki, üç ve dördüncü cümleler olumlu, beşinci cümle ise olumsuzdur.

a. Bu cümlelere bakıldığında bir düşünce iki kısımdan oluşur: Varlık (Target) ve duygu (sentiment). Varlık g harfi ile gösterilmektedir ve varlık ya da varlığın özelliği olarak adlandırılır. Duygu; olumlu, olumsuz ya da nötr ifadelerin belirtildiği kısımdır. Olumlu, olumsuz ya da nötr ifadeleri duygu kutupları olarak adlandırılmaktadır [1].
b. Kim ve Hovy [12], Wiebe [13] yaptığı çalışmalarda, düşünceleri belirten düşünce sahiplerinin de önemli olduğunu raporlamışlardır. Düşünce sahibi (opinion holder) ya da düşünce kaynağı (opinion source) olarak adlandırılmaktadır. Örnek üzerinden baktığımızda, iki, üç ve dördüncü cümlelerde düşünce sahibi, telefonu satın alan kişidir, beşinci cümledeki görüş sahibi telefonu alan kişinin eşidir [1].

c. Görüşler belirtildikleri zamana göre değerlendirildikleri ve o zaman diliminde değerli oldukları için zaman kavramına sahiptirler. Görüşlerin zamanla nasıl değiştiğinin anlaşılması için görüşün zamanı önemli bir özellik olarak karşımızda durmaktadır [1].

Bu açıklamalara ve çalışmalara göre basit bir düşünce dört kısımdan oluşur. Düşünce, dört bileşenli bir nesnedir. Düşüncenin bileşenleri; varlık (g), düşünce (s), düşünce sahibi (h) ve zamandır (t).

Bu tanım bazı durumlarda yetersiz kalmaktadır. Bazı cümlelerde varlık doğrudan yer almaz, varlığın bir özelliği yer alabilir (Üçüncü cümledeki resim kalitesi gibi). Böyle durumlarda ana varlığın da bilinmesi gerekmektedir.

Varlık, bir ikili nesne olarak tanımlanır: Varlık ya da varlığın parçaları (T) ve varlığın özellikleridir (W). Örnek olarak, HTC One telefon ana varlıktır (T), resim kalitesi ana varlığa ait bir özelliktir (W). Bu tanımlama ile birlikte ana varlığın belirtilmediği cümlelerde de ana varlığa ulaşılabilir.

Varlığın detaylı tanımının yapılması ile birlikte, düşünce beşli bir tanım içerir:

T	->	Varlık ya da varlığın parçaları
W	->	Varlığın özellikleri
s	->	Düşünce
h	->	Düşünce sahibi
t	->	Zaman

1.4 Motivasyon

Düşünceler, davranışlarımızı önemli ölçüde etkilediği için insan eylemlerinin neredeyse tamamının merkezinde farklı kişilerin düşünceleri söz konusu olabilmektedir. Diğer bir ifadeyle, insanlar bir konuda karar vereceği zaman, aynı konu hakkında başkalarının düşüncelerini ve deneyimlerini detaylı olarak bilme ihtiyacı duyar. Uygulamada, farklı şirketler ve organizasyonlar ürünleri hakkında müşterilerinin ve halkın düşüncelerini, ürünleri hakkındaki olumlu veya olumsuz görüşleri öğrenmek istemektedir. Web 2.0'dan önce, bir konuda farklı düşüncelere ve kişisel deneyimlere ulaşmak isteyenler, arkadaşlarına tanıdıklarına ya da çevresindeki tüm insanlara konu hakkındaki görüşlerini sözlü olarak sorarlardı ya da internet üzerinde kısa bir araştırma ile bu tür bilgileri elde etmeye çalışırlardı.

Sosyal medyanın hızla büyümesinden dolayı değerlendirmeler ve görüşlerle ilgili bilgi ve veri içeriği gün geçtikçe çok hızlı artmaktadır. Twitter, 2013 Mart bilgilerine göre 200 milyon aktif kullanıcıya sahiptir. Günde 400 milyon tweet atılmaktadır [14]. Bu veriler internet üzerindeki trafiğin artışını gözler önüne sermektedir. Ayrıca, küresel internet trafiğinin 2015'te 4 katına çıkarak 1 zetabayta ulaşacağı öngörülmektedir [15].

İnternet kullanımının ve verinin artması, bu verinin işlenmesini ve analiz edilmesini daha önemli hale getirmiştir. Bu büyüklükte bir verinin elle analiz edilip sonuçları alınamaz. Bu durum ve şartlar duygu analizi çalışmalarını önemli bir konuma getirmiş ve bu alandaki çalışmaların artmasını sağlamıştır. Şirketler ve bireyler için herkesin ne düşündüğü ve ne konuştuğu önemlidir. Dolayısıyla bu bilgilerin analiz edilmesi gerekir. Veriden en yüksek verimi elde etmek ve yarar sağlamak için otomatik analiz gereklidir.

Yapılan son çalışmalara göre, "Amerikalıların % 58'i bir servis ya da ürün satın almadan önce ürün hakkında araştırma yapıp bilgi toplamaktadırlar" [16]. Diğer bir araştırmaya göre, twitter mesajlarından borsadaki tahminler % 87,6 doğrulukla tahmin edilebilmektedir [17]. Bir başka araştırma raporuna göre [18], sosyal medya kullanıcılarının % 66'sı siyasi konularda görüşlerini internet üzerinden belirtmişlerdir. Bu çalışmalar ve raporlar, doğru bilgi ve doğru bilgi kaynağı oluşturmak için duygu analizinin çok önemli bir çalışma alanı olduğunu göstermektedir.

Duygu analizi alanında İngilizce için birçok çalışma mevcuttur. Yapılan literatür incelemesinde, Türkçe için şu ana kadar sadece iki çalışmaya ulaşılabilmiştir. Bu çalışmalardan ilki [19] tek bir alan üzerinde tek bir sınıflandırıcı kullanılarak gerçekleştirilmiştir. Diğer çalışma [20] ise birde fazla alan üzerinde farklı sınıflandırıcılar tek tek denenerek yapılmıştır ve en yüksek doğruluk değeri elde edilmeye çalışılmıştır. Naive Bayes sınıflandırıcısı ile % 85 oranında doğruluk değeri ilgili çalışmada elde edilmiştir.

Bu tezin amacı ise sınıflandırıcı birlikteliği (ensemble of classifiers) kavramından yararlanarak, çoklu sınıflandırıcılar yardımıyla doğruluk oranını ve dolayısıyla düşünce sınıflandırmadaki performansı arttırmaktır. Tez kapsamında çoklu sınıflandırıcılar Türkçe için ilk defa incelenmiş ve çok sayıda analizin neticesinde % 86,13 doğruluk değerinin elde edildiği yeni bir model ortaya konulabilmiştir. Tez içerisinde; çoklu sınıflandırıcı yaklaşımı, kullanılan sınıflandırıcılar, sınıflandırıcılar üzerinde uygulanan parametre optimizasyon yaklaşımları ve performansın hesaplanma yöntemleri ayrıntılı olarak verilmektedir.

1.5 Taslak

Tezin ilk bölümünde yapılan çalışmayla ilgili genel tanımlar sunulmaktadır. İkinci bölümde literatürde geçen ilişkili çalışmalar verilmektedir. Tezin üçüncü bölümünde tez çalışmasının altyapısını oluşturan bilgiler ortaya konulmaktadır. Bu bilgiler; makine öğrenmesinin genel tanımını, makine öğrenmesi algoritmaları, destek vektör makineleri ve uygulanan yöntemler olarak sıralanabilir. Dördüncü bölümde, WEKA makine öğrenmesi kütüphanesi açıklanmaktadır. Beşinci bölümde deneysel çalışmalarda kullanılan veri kümeleri ve özellikleri verilmektedir. Altıncı bölümde, deneysel çalışmalar ve bu çalışmaların neticesinde tespit edilen gözlemler sunulmaktadır. Son bölüm olan yedinci bölümde ise sonuç ve bu alanda yapılabilecek gelecek çalışmalar ortaya konulmaktadır.

2 LİTERATÜR TARAMASI

Duygu sınıflandırma çalışmalarına çok erken yıllarda başlanmıştır. En erken çalışma Carbonell [21] tarafından doktora tezi olarak yapılmıştır. Carbonell, politik olaylar veri kümesi üzerinde doğal dilin insan anlayışını modellemeye çalışmıştır.

Web 2.0 ile internet kullanımın artmasıyla birlikte analiz edilmesi gereken veri kümeleri artmaya başlamıştır. Analiz edilmesi gereken veri kümelerinin artmasıyla duygu sınıflandırmaya olan ihtiyaç açığa çıkmış ve çalışmalar hızlanmıştır.

Duygu sınıflandırma alanında önemli bir çalışma Pang ve arkadaşlarına [6] aittir. Pang ve arkadaşları bu çalışmalarında IMDB film yorumları veri kümesi üzerinde eğiticili öğrenme algoritmalarını incelemişlerdir. Çalışmalarında Naive Bayes, Destek Vektör Makineleri ve Maksimum Entropi olmak üzere farklı makine öğrenmesi algoritmalarını denemişlerdir. Çalışmalarında Bag of Words doğal dil işleme yaklaşımını kullanmışlar. Çalışma sonunda DVM'ler ile 82.9%'lik doğruluk oranına ulaşmışlardır.

Turney ve arkadaşları [8] eğiticisiz öğrenme algoritmalarını denemişlerdir. Çalışmalarının sonunda %80'lik doğruluk oranı gerçekleştirmişler.

Pang ve Lee [22], sınıf sayısını arttırarak üç ve dört sınıflı sınıflandırmaları incelemişlerdir. Üç sınıflı sınıflandırmada %70, dört sınıflı sınıflandırmada %50 doğruluk oranı elde etmişlerdir.

Duygu analizi konusu insan etkileşiminden ve düşüncelerinden kaynaklı bir çalışma alanı olduğundan dolayı insanın bulunduğu tüm alanlarla doğrudan ilişkilidir. Bu yüzden günlük hayattan gerçek yaşam uygulamaları da yayınlanmıştır. Bu kapsamdaki çalışmalardan biri Liu ve arkadaşları [23] tarafından yapılan satış performansının tahmin edilmesi için duygu modeli geliştirme çalışmasıdır. Tumasjan ve arkadaşları [24] seçim sonuçlarını tahmin etmek için Twitter üzerinde duygu sınıflandırması işlemi gerçekleştirmişlerdir. Bollen ve arkadaşları [25] borsadaki işlemlerin durumunu tahmin edebilmek için Twitter verileri üzerinde duygu sınıflandırması yapmışlardır.

Li ve arkadaşları [26], çoklu sınıflandırıcılar ile çoklu etki alanı üzerinde çalışmışlar ve tek etki alanı üzerinde çalışıldığında alınan hata oranını %27.6 düşürmüşlerdir. Li ve arkadaşları [27] bir diğer çalışmalarında çoklu sınıflandırıcılar ile film veri kümesi üzerinde çalışmışlar ve en iyi tek sınıflandırıcıda 2.56% daha fazla doğruluk oranına ulaşmışlardır.

Kittler ve arkadaşları [28], çoklu sınıflandırıcı altyapısı ve birleştirme kuralı geliştirmişlerdir.

İngilizce dilinde duygu analizi çalışmaları yeterli sayıda olmasına rağmen, Türk dili için duygu analizi çalışmaları hala yetersiz seviyededir. Türk dili için çok fazla sayıda çalışmaya rastlanmamaktadır, kısıtlı sayıda çalışma yapılmıştır.

Yaptığımız literatür taramasına göre Türk dili için şu ana kadar iki çalışma yapılmıştır. Çalışmalardan ilki Umut Eroğul [19] tarafından eğiticili ve eğiticisiz makine öğrenmesi algoritmaları üzerinde gerçekleştirilmiştir. Film yorumları veri kümesi üzerinde tek alandaki veriler alınarak yapılmıştır. Çalışma sonunda %85'lik doğruluk oranına ulaşılmıştır.

İkinci çalışma 2013 yılının Ocak ayında Hakan Çelik [20] tarafından gerçekleştirilmiştir. Bu tez çalışmasında kitap, sinema ve alışveriş veri kümeleri üzerinde üç farklı alanda makine öğrenmesi algoritmaları incelenmiştir. Çalışmada Naive Bayes makine öğrenmesi algoritması ile %85'lik bir başarı sağlanmıştır.

Türk dili için çoklu sınıflandırıcı sistemlerin uygulandığı bir çalışmaya lüteratür taraması sırasında rastlanmamıştır.

3 ALTYAPI

3.1 Makine Öğrenmesi

Çok büyük miktarlardaki verinin elle işlenmesi ve analizinin yapılması mümkün değildir. Amaç geçmişteki verileri kullanarak gelecek için tahminlerde bulunmaktır. Bu problemleri çözmek için makine öğrenme yöntemleri geliştirilmiştir. Makine öğrenmesi yöntemleri, geçmişteki veriyi kullanarak yeni veri için en uygun modeli bulmaya çalışır.

Makine öğrenmesi 1959 yılında Arthur Samuel tarafından "Bilgisayarlara öğrenme yeteneğinin kazandırıldığı çalışma alanı" olarak tanımlanmıştır [29].

Verinin incelenip içerisinden işe yarayan bilginin çıkarılmasına da veri madenciliği (data mining) adı verilmektedir.

Makine öğrenmesi iki ana başlık altında incelenmektedir: eğiticili öğrenme(supervised learning) ve eğiticisiz öğrenme(unsupervised learning)'dir.

Eğiticili öğrenme, sistemin en başta eğitildiği ve bu eğitim sonucuna göre karar vermesinin sağlandığı öğrenme tekniğidir. Sistem; öncelikle bir dizi girdi verileri ve bu girdi verilerine uygun çıktı verileri ile eğitilir. Sistem, bu girdi ve çıktı verilerine uygun öğrenme fonksiyonları üretir. Bu fonksiyonlar yardımı ile daha sonra gelecek verilere karar verir. Naive Bayes, Destek Vektör Makineleri (Support Vector Machines) ve Karar Ağaçları (Decision Tree) yaygın olarak kullanılan eğiticili öğrenme tekniklerindendir.

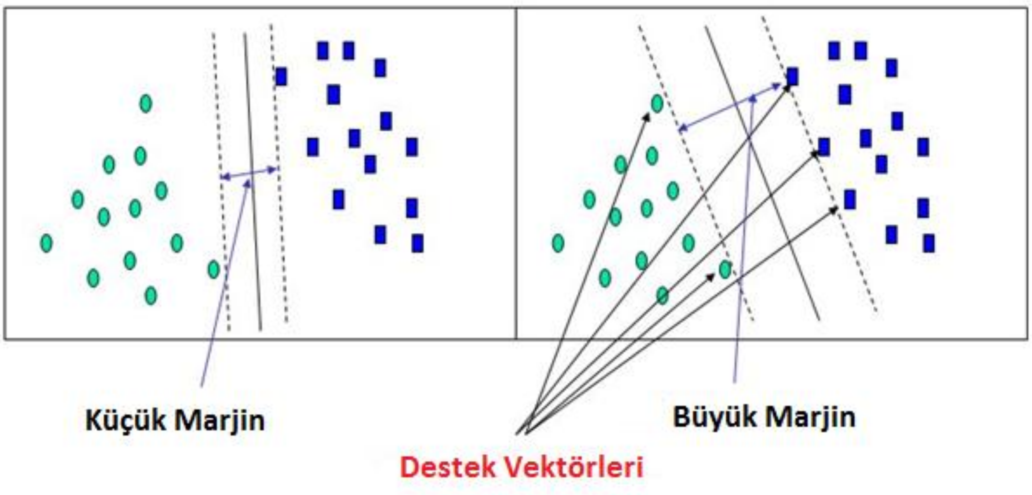
Eğiticisiz öğrenme, sistemin başta eğitilmediği, sistemin kendi kendine öğrenmesinin sağlandığı öğrenme tekniğidir. Sisteme sadece girdi verileri verilir, çıktı verileri sisteme verilmez. Girdi verileri ve örnekler arasındaki ilişkiler ile sistemin kendi kendisine öğrenmesi sağlanır. Kümeleme (Clustering) yaygın olarak kullanılan eğiticisiz öğrenme tekniğidir.

Tez kapsamında test ve analiz işlemleri için eğiticili makine öğrenmesi yöntemleri kullanılmıştır.

3.2 Destek Vektör Makineleri

Destek Vektör Makineleri (DVM), öğrenme, sınıflandırma, yoğunluk tahmini ve kümeleme için kullanılan eğiticili bir öğrenme algoritmasıdır. Destek Vektör Makineleri, sınıflandırılacak kategorileri aşırı düzlem ile ayırarak yapılandırılır. Aşırı düzlem her iki sınıftaki en uç destek vektörlerine eşit uzaklıkta konumlandırılır.

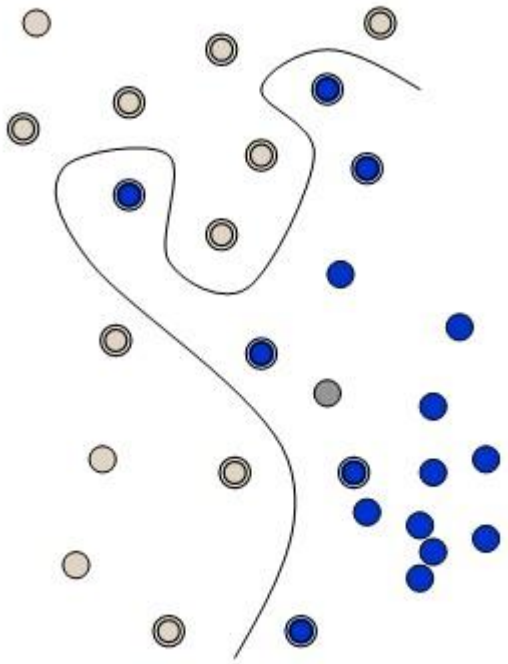
Şekil 3.1'de doğrusal düzlemdeki aşırı düzlemler ve destek vektörleri gösterilmektedir. Aynı lineer düzlem için birden fazla aşırı düzlem çizilebilir. Destek vektörlerine en uzak olan aşırı düzlem seçilir.



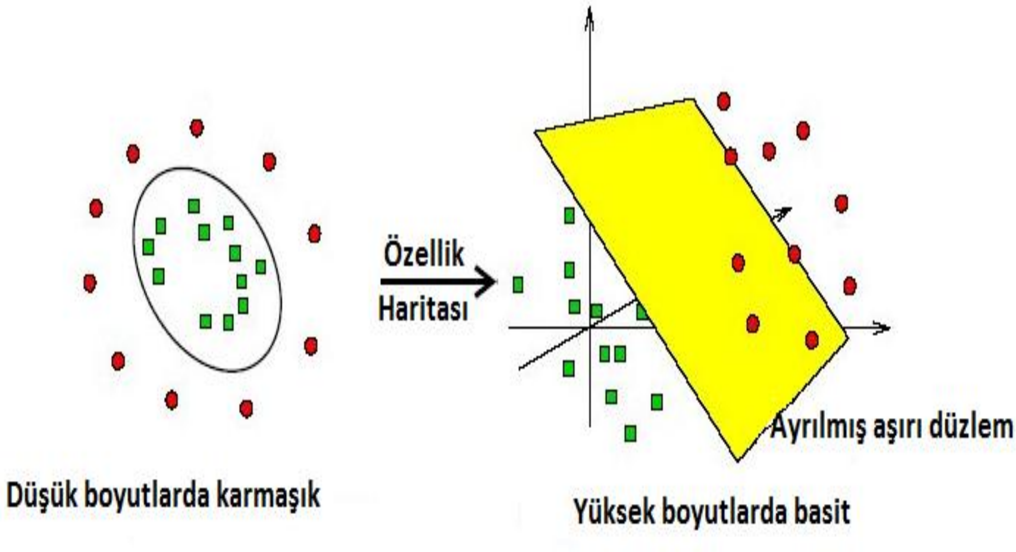
Şekil 3.1 DVM Aşırı Düzlemleri ve Destek Vektörleri [30]

Destek vektör makinelerinin ana prensibi iki sınıfı birbirinden ayırmaya yarayan aşırı düzlemlerin belirlenmesidir [31]. Doğrusal düzlem iki sınıfın sınıflandırılması için tasarlanmıştır fakat daha sonra doğrusal olmayan verilerin sınıflandırılması için de geliştirilmiştir.

Şekil 3.2'de yer alan karmaşık veri, düz bir aşırı düzlemle ayrılamamaktadır. Bu gibi durumlarda çekirdek (kernel) fonksiyonları yardımı ile farklı uzaydaki veriler eşleştirilir. Şekil 3.3'de bu durum gösterilmiştir.



Şekil 3.2 Karmaşık Veri [30]



Şekil 3.3 Doğrusal Olmayan Verilerin Aşırı Düzlem ile Ayrılması [30]

3.3 Karar Ağaçları

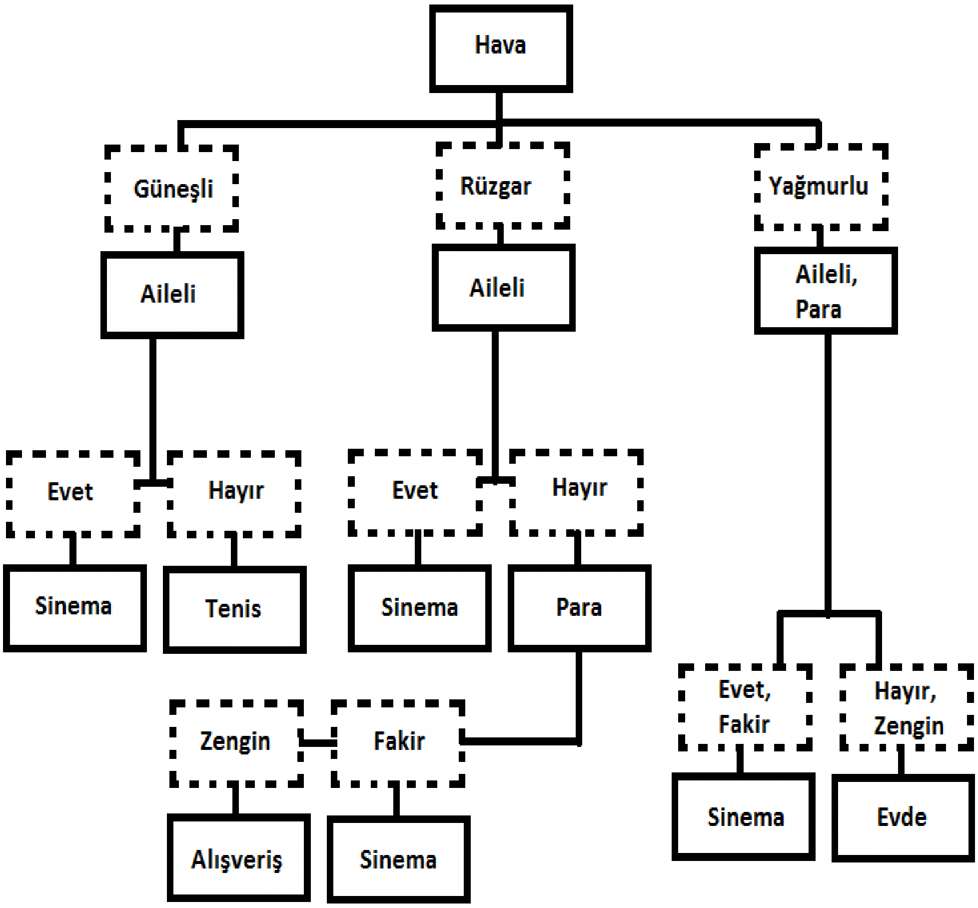
Karar ağaçları makine öğrenmesi tekniklerinden bir tanesidir. Karar ağaçlarında bir ağaç yapısı oluşturularak ağacın yaprakları seviyesinde sınıf etiketleri ve bu yapraklara giden ve başlangıçtan çıkan kollar ile de özellikler üzerindeki işlemeler ifade edilmektedir.

Eldeki veri kümesinden bir sınıf seçilerek kök düğümü oluşturulur. Bu kök düğümünden sorular sorulup cevaplar alınarak, yaprak düğümler oluşturularak bir karar verme yapısı geliştirilir.

Şekil 3.4 ve Şekil 3.5'te karar ağacını oluşturacak veri kümesi ve bu veri kümesinden oluşturulan karar ağacı örnek olarak gösterilmiştir.

Hafta (Örnek)	Hava	Aileli	Para	Karar
H1	Güneşli	Evet	Zengin	Sinema
H2	Güneşli	Hayır	Zengin	Tenis
H3	Rüzgarlı	Evet	Zengin	Sinema
H4	Yağmurlu	Evet	Fakir	Sinema
H5	Yağmurlu	Hayır	Zengin	Evde
H6	Yağmurlu	Evet	Fakir	Sinema
H7	Rüzgarlı	Hayır	Fakir	Sinema
H8	Rüzgarlı	Hayır	Zengin	Alışveriş
H9	Rüzgarlı	Evet	Zengin	Sinema
H10	Güneşli	Hayır	Zengin	Tenis

Şekil 3.4 Karar Ağacı Veri Kümesi [32]



Şekil 3.5 Karar Ağacı Örneği[32]

3.4 Naive Bayes Algoritması

Naive Bayes (NB), niteliklerin birbirinden bağımsız ve nitelikleri hepsinin aynı derecede önemli olduğunu kabul eden eğiticili makine öğrenmesi tekniğidir. İsmini İngiliz matematikçi Thomas Bayes'ten alır [33]. Bayes kuralında, Eşitlik 1'deki denkleme göre koşullu olasılık hesaplanır.

$$\mathrm{P(A \mid B)} = \frac{P(B \mid A)\ P(A)}{P(B)} \qquad (1)$$

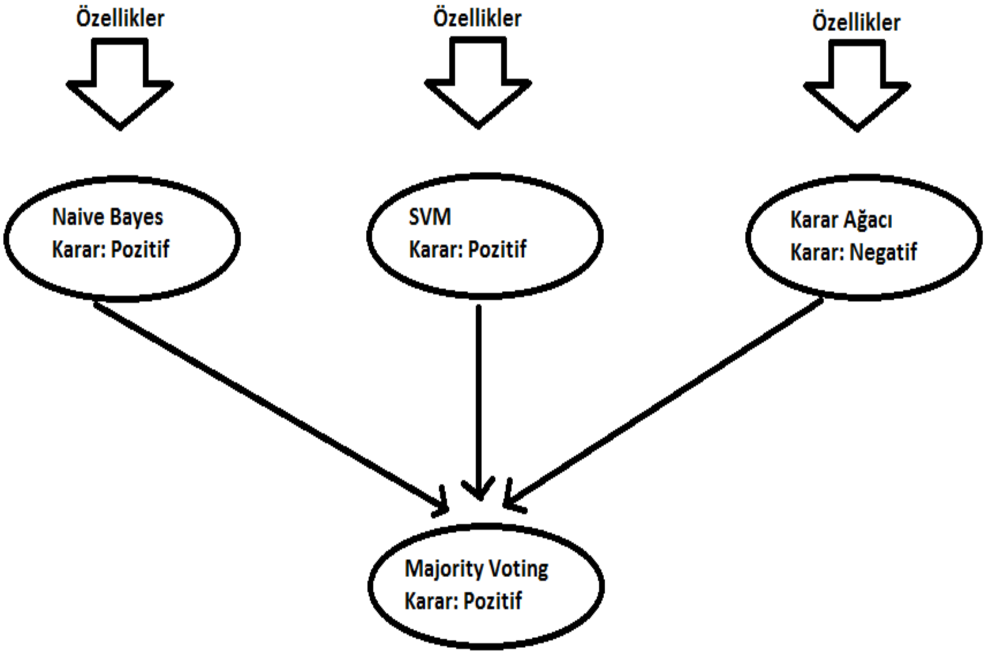
Metin belgelerinin sınıflandırılmasında yaygın olarak kullanılır. Uygulanabilirliği ve performansı ile ön plana çıkan bir algoritmadır. İstatistiksel yöntemler yardımı ile sınıflandırma yapar.

3.5 Çoğunluk Oylaması Kuralı

Sistemin genel performansını arttırmak için tekil sınıflandırıcıların kararlarının birleştirilmesiyle çoklu sınıflandırıcı yöntemler geliştirilmiştir. Oylama (Vote) yöntemi çoklu sınıflandırıcı yöntemlerden bir tanesidir. Oylama yöntemi içerisine birden fazla tekil sınıflandırıcı alarak, bu tekil sınıflandırıcıların kararlarını farklı birleştirme kurallarına göre birleştirir ve genel sistem kararının verilmesini sağlar. Oylama yöntemi, altı farklı birleştirme kuralı içerir.

- Çoğunluk oylaması (Majority Voting)
- Olasılıkların ortalaması (Average of Probabilities)
- Olasılıkların Çarpımı (Product of Probabililities)
- En yüksek olasılık (Maximum Probability)
- En düşük olasılık (Minimum Probability)
- Ortanca (Median)

Bu tez kapsamında çoklu sınıflandırıcı yöntemler çoğunluk oylaması kuralına göre birleştirilmiştir. Çoğunluk oylaması (ÇO) denilen bu yöntem, tekil sınıflandırıcıların kararlarının verilmesi ve en fazla olan kararın sistemin genel kararı olduğunun belirlenmesi şeklinde çalışmaktadır. Her bir sınıflandırıcının kararı eşit öneme sahiptir. Şekil 3.6'da çoğunluk oylaması yönteminin genel çalışma şekli gösterilmektedir.



Şekil 3.6 Çoğunluk Oylaması

3.6 Parametre Optimizasyonu

Sınıflandırıcıların performansını arttırmak için sınıflandırıcıların parametreleri optimize edilebilmektedir. Bu kapsamda sınıflandırıcının daha iyi performansla çalışacağı parametre değerlerinin bulunması gerekmektedir.

Weka makine öğrenmesi kütüphanesi parametre optimizasyonu için iki tane algoritma sunmaktadır [34].

- weka.classifiers.meta.CVParameterSelection
- weka.classifiers.meta.GridSearch

Bu tez kapsamında parametre optimizasyonu (PO) işlemi CVParameterSelection algoritması ile gerçekleştirilmiştir. Destek vektör makinesi yöntemleri ve karar ağacı yönteminin parametre optimizasyonu gerçekleştirilmiş ve test edilmiştir.

3.7 Bagging

Sınıflandırma ve regresyon kullanıldığında makine öğrenmesi algoritmalarının kararlılığını ve doğruluğunu iyileştirmek için tasarlanmış meta algoritmalardan birisi de Bagging algoritmasıdır [33].

Bagging [35] metodu, var olan eğitim verisinin örneklerin yer değiştirilmesiyle eğitim verisinin farklı kombinasyonları oluşturularak elde edilen eğitim verilerinin sınıflandırıcılar tarafından öğrenilmesi sonucu oluşan modellerin sonuçlarının karşılaştırılması yöntemine dayanır.

3.8 Boosting

Boosting algoritmaları, topluluk yöntemi olarak adlandırılan makine öğrenmesi algoritmalarıdır. Topluluk yöntemi algoritmalarının amacı, birçok yetersiz makine öğrenmesi algoritmasını toplayarak güçlü bir öğrenme algoritması oluşturmaktır.

Veri kümesindeki her bir verinin bir ağırlığı bulunmaktadır. Öğrenme işleminden sonra her sınıflandırıcı için yapılan sınıflandırma hatasına bağlı olarak verilerin ağırlığı güncellenmektedir. Bundan dolayı öğrenilen modellerin ağırlıkları bir modelden diğerine değişiklik göstermektedir. Sisteme gelen yeni bir veriyi sınıflandırmak için her sınıflandırıcının doğruluğuna bağlı olarak ağırlıklı ortalaması alınmaktadır.

Tez kapsamında boosting algoritması olarak Adaptive Boosting (AdaBoost) algoritması kullanılmıştır. AdaBoost algoritması Freund ve Schapire [36] tarafından sunulmuştur.

3.9 Bag of Words ve N-Gram Model

Bag of words, doğal dil işleme alanında belgelerin ve cümlelerin basitleştirilmiş temsili için kullanılmaktadır. Bu işlem sırasında kelime sırası, noktalama işaretleri ve dil bilgisi kuralları göz ardı edilir [33].

Örnek:

Birinci cümle: Mehmet kitap okumayı sever.

İkinci cümle: Tuğba fotoğraf çekmekten ve doğa yürüyüşünden zevk alır.

Bag of words:

Birinci cümle: “Mehmet”, ”kitap” , “sever”, “okumayı”

İkinci cümle: “Tuğba”, “doğa”, “zevk”, “fotoğraf”, “çekmekten”, “ve”, “yürüyüşünden”, “alır”

Sözlük:

{ “Mehmet”:1, ”kitap”:2 , “sever”:3, “okumayı”:4, “Tuğba”:5, “doğa”:6, “zevk”:7, “fotoğraf”:8, “çekmekten”:9, “ve”:10, “yürüyüşünden”:11, “alır”:12 }

Bag of words genellikle n-gram modellerle birlikte kullanılırlar. N-gram model, bir cümleden n tane kelimenin seçilmiş şekline verilen isimdir. N-gram modellerin tek kelimeden oluşanlarına unigram, iki kelimeden oluşanlarına bigram, üç kelimeden oluşanlarına trigram ismi verilmektedir.

Var olan metinlerden önce n-gramlar oluşturulur. Bu n-gramlar bag of words yaklaşımında eleman olarak gösterilir.

Örnek: Mehmet kitap okumayı sever.

Unigrams: { “Mehmet”, “kitap”, “okumayı”, “sever” }

Bigrams: { “Mehmet kitap”, “kitap okumayı”, “okumayı sever” }

Trigrams: {“Mehmet kitap okumayı”, “kitap okumayı sever” }

3.10 Eşikleme

Eşikleme, bag of words yaklaşımının özellik sıralamasına verilen isimdir. N-gram modeller veri kümesindeki frekanslarına göre sıralanır. N-gram modellerin eşik değerleri frekanslarıdır. N-gram modellerde eşik değeri arttırıldığında sistemlerin doğruluğunda ve performansında artış olur. Bunun sebebi eşik değerini arttırarak, az kullanılmış kelimelerin veri kümesinden atılmasıdır. Sınıflandırma doğruluğu böylece daha da artacaktır.

3.11 K-Katlamalı Çapraz Doğrulama

Makine öğrenmesinde veri kümeleri; sistemin eğitilmesi ve test edilmesi için eğitim ve test veri kümeleri olarak ayrılmaktadır. Bu ayırma işleminin birden fazla yolu vardır. K-katlamalı çapraz doğrulama bu yöntemlerden bir tanesidir.

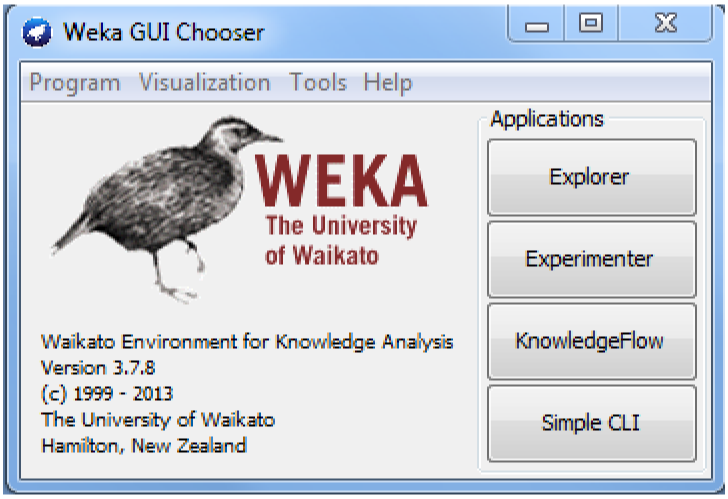
Sistemde öncelikli olarak bir k değeri belirlenir. Veri kümesi k değeri kadar parçaya ayrılır. Bu k değerlerinden bir tanesi test, diğer k-1 tanesi eğitim veri kümesi olarak kullanılır. Sistemdeki sınıflandırma yöntemi k defa çalışır. Sistemin doğruluk oranı bu k defa çalışmanın ortalaması olarak Eşitlik 2'deki denklemle bulunur.

$$t_i \in VK \; olmak \; \text{ü}zere, Sonu\c{c} = \frac{\sum_{i=0}^{k} SF(t_i, VK - t_i)}{k} \qquad (2)$$

Genel olarak en fazla kullanılan k değeri 10'dur. Bu tez kapsamında k değeri 10 olarak kullanılmıştır. Veri kümesi 1 test, 9 eğitim kümesi olarak parçalanıp sınıflandırma yapılacaktır.

4 Weka Makine Öğrenmesi Kütüphanesi

Weka, veri madenciliği işlemleri için kullanılan makine öğrenmesi algoritmalarının bulunduğu bir kütüphanedir [38]. Waikato üniversitesinde Java ile geliştirilmiştir. GPL lisansı ile dağıtılmaktadır. İsmi de Waikato Environment of Knowledge Analysis kelimelerinin baş harflerinden oluşmaktadır. Tez kapsamında Weka programının 3.7.8 versiyonu kullanılmıştır. Şekil 4.1'de Weka 3.7.8 programının ana menüsü yer almaktadır.

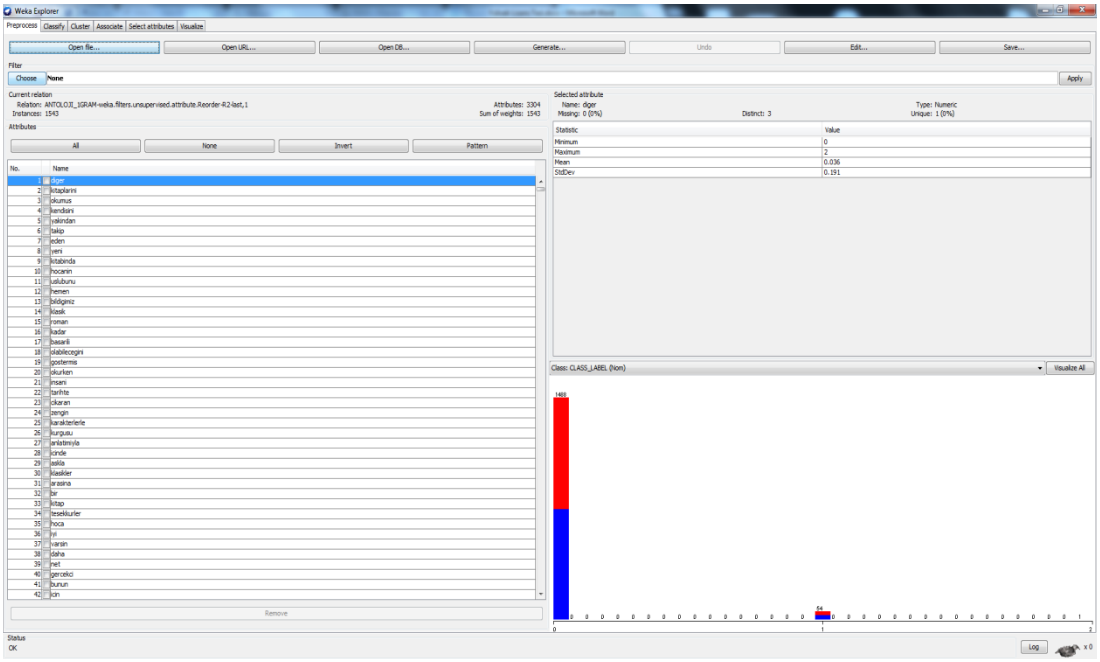


Şekil 4.1 Weka Menüsü

Weka içerisinde şu araçları barındırır:

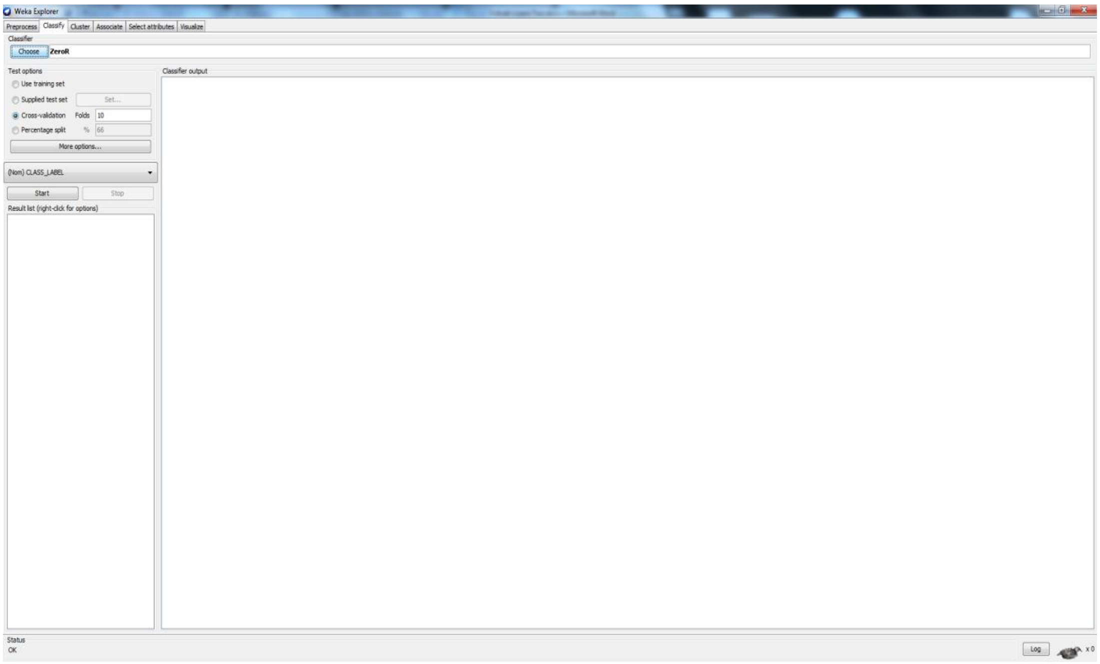
- Veri ön işlemesi
- Sınıflandırma
- Kümeleme
- İlişkisel kurallar
- Özellik seçimi
- Görselleştirme

Şekil 4.2'de Weka programının ana ekranlarından veri ön işleme ekranı yer almaktadır. Veri ön işleme ekranında işlem yapılacak ARFF dosyası programa yüklenir ve yapılmak istenen ön işlemler gerçekleştirilir.



Şekil 4.2 Weka Veri Ön İşleme Ekranı

Şekil 4.3'de Weka programının tez kapsamında kullanılan sınıflandırma penceresi yer almaktadır. Bu ekran üzerinden sınıflandırma algoritması seçilerek Weka programı üzerinde analiz işlemi yapılabilmektedir.



Şekil 4.3 Weka Sınıflandırma Ekranı

5 Veri Kümesi, Eğitim ve Test İşlemleri

5.1 Veri Kümesi

Bu tez çalışmasında Hakan Çelik tarafından Türk dili için yüksek lisans çalışmasında oluşturduğu veri kümesi kullanılmıştır [20]. Veri kümesi içeriği internet üzerinden üç siteden çekilen kitap, film ve alışveriş yorumlarından oluşmaktadır. Aşağıda örnek yorumlar gösterilmiştir.

ÖRNEK VERİ

Kitap Yorum Örneği

- Çok ilginç konuların bir arada toplandığı güzel bir kitap olmuş.
- Hiç beğenmediğim kitaplardan biriydi zaten doğru düzgün okuyamadım kötü içeriklerinden ötürü adeta midemi bulandırdı.

Film Yorum Örneği

- Şüphesiz serinin en güzeli…İzlemeyen varmı hala merak ediyorum…hele o eşsiz müziğiyle finali yokmu tekrar tekrar izleyesi geliyor insanın.
- Zaman kaybı arkadaşlar..kesinlikle tavsiye etmiyorum..

Alışveriş Yorum Örneği

- Ürünün minik adaptörü ile pil sorunu yaşamadan çalışması en sevdiğim tarafı oldu.
- Ürünün elime ulaşma hızı ile kalitesi aynı orantıda değil maalesef.

Web sitelerinden çekilen verilere göre olumlu yorumlar çoğunluktadır. Olumsuz yorumların oranı tüm yorumların sadece %4-5 seviyesinde kalmıştır. Tez çalışmasında olumlu ve olumsuz yorumlar eşit sayılarda alınarak bir eşitlik ve denge kurulmuştur [20]. Tüm yorumların sayısı ve dengeleme işlemi yapıldıktan sonra çalışmada kullanılmak için belirlenen veri kümesi sayısı Tablo 4.2'de gösterilmiştir. Bu dengeleme sayesinde, doğruluk parametresi bu çalışmalarda kullanılabilmiştir.

Tablo 5.1 Veri Sayıları [32]

	Ham Veri	**Dengelemeden Sonra**
Kitap	20623	1548
Film	13156	2248
Alışveriş	51879	5256
Hepsi	85658	9624

Veri yapısı aşağıdaki şekilde oluşturulmuştur. Girdi cümlelerinden bir sözlük oluşturulmuştur. Sözlükte ilk eleman 0 indeksli sınıf etiketidir (CLASS_LABEL). Sonrasında 1, 2, 3 indeksleri ile sırasıyla tüm n-gramlar sözlüğe yerleştirilmiştir. Bir sonraki adımda girdi cümleleri işlenmeye başlanmıştır. Cümlenin n-gramları sözlükte aranır ve arff satırı oluşturulur.

{0 1,1 1,2 1,3 1,4 2,5 1 ...}

Bu arff satırına göre sıfırıncı elemanın değeri birdir. Sıfırıncı eleman sınıf etiketidir. Cümlenin olumlu ya da olumsuz olduğunu bu bilgi göstermektedir. Ardından gelen veriler bu metinde, sözlükteki ID'si bir olan n-gramdan bir tane var, iki ID'li n-gramdan bir tane var, üç ID'li n-gramdan 1 tane var, dört ID'li n-gramdan 2 tane var anlamına gelmektedir [20].

5.2 Eğitim ve Test İşlemleri

Hakan Çelik tarafından yapılan yüksek lisans çalışmasında [20] eşik değerleri olarak 0, 10, 50 ve 100 eşik değerleri denenmiştir. En iyi değerler 10 eşik değerinde elde edildiğinden dolayı bu tez çalışması kapsamında da eşik değeri 10 olan veri kümeleri kullanılmıştır.

Sınıflandırma çalışmaları için Naive Bayes (Multinomial)(NBM), Destek vektör makineleri, Karar ağaçları, Bagging ve Boosting eğiticili makine öğrenmesi algoritmaları incelenmiştir. Çoklu sınıflandırıcı sistem için Vote algoritması, oy çoğunluğu birleştirme metodu ile kullanılmıştır. Algoritmaların performanslarını arttırmak için parametre optimizasyonu işlemi CVParameterSelection meta algoritması ile gerçekleştirilmiştir. Eğitim ve test işlemlerinde 10 katlamalı çapraz doğrulama tekniği uygulanmıştır. 10 veri kümesinin 9 tanesi eğitim veri kümesi, bir tanesi test veri kümesi olarak kullanılmıştır.

6 Testler ve Sonuçlar

6.1 Testler

6.1.1 Test Ortamı

Testler ve analiz işlemleri üç farklı makine üzerinde gerçekleştirilmiştir. İşlemlerin gerçekleştirildiği makinelerin özellikleri Tablo 6.1, Tablo 6.2 ve Tablo 6.3'te gösterilmiştir.

Tablo 6.1 Makine 1'in Özellikleri

İşlemci	Intel Ivy Bridge Q3630 i7 işlemci
Ram	16 gb Corsair Vengence 1600 mhz
Hard disk	256 GB Ocz Vertex 4 SSD

Tablo 6.2 Makine 2'nin Özellikleri

İşlemci	Intel Q720 i7 işlemci
Ram	8 gb Corsair Vengence 1600 mhz
Hard disk	300 GB SATA HDD

Tablo 6.3 Makine 3'ün Özellikleri

İşlemci	Intel Core i5 işlemci
Ram	8 gb Kingston 667 mhz
Hard disk	160 GB SATA HDD

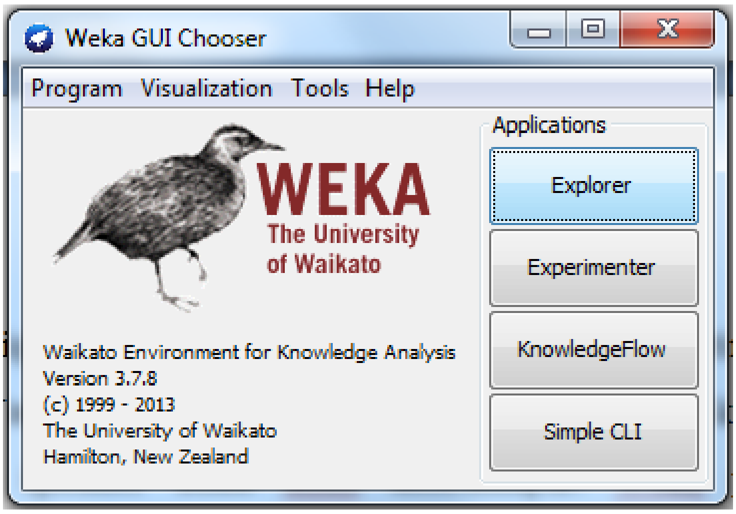
Makineler arasındaki işlem süreleri kıyaslandığında en performanslı olarak Makine 1 çalışmıştır ve işlem süresi diğer cihazlara göre 3-4 kat daha hızlıdır. Bu fark makine üzerinde solid state diskten(SSD) kaynaklanmaktadır. SSD disk, işlem sürelerini çok kısaltmış ve performans sağlamıştır. WEKA programının daha kısa sürede ve daha verimli çalışmasında SSD ve yüksek RAM kapasitesinin etkili olduğu testlerde görülmüştür. Makineler üzerindeki WEKA programı test ve analiz performans karşılaştırması aşağıdaki şekilde sıralanmıştır.

Makine 1 > Makine 3 > Makine 2

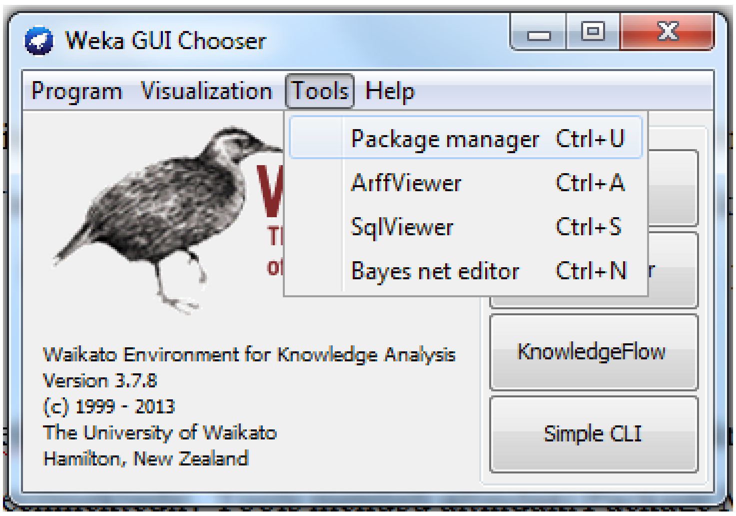
6.1.2 Weka ile Sınıflandırma

Test ve analiz işlemleri Weka programı üzerinde gerçekleştirilmiştir. Weka programı Java dili ile geliştirilmiş açık kaynaklı makine öğrenmesi kütüphanesidir. Weka üzerinde gerçekleştirilen işlemler adım adım bu bölümde açıklanmıştır. Weka programının 3.7.8 versiyonu kullanılmıştır.

Şekil 6.1'de Weka ana ekranı gösterilmektedir. Weka ana ekranı üzerinden işlem yapılacak olan menülere geçilmektedir. Weka programı ilk indirildiğinde içinde tüm algoritmalar ve kütüphaneler gelmemektedir. Şekil 6.2'de Tools menüsü altındaki Package Manager sekmesinden kütüphaneler listelenebilir ve ek paketler indirilebilmektedir. Ek sınıflandırıcı kütüphaneler bu aşamalar izlenerek indirilmiştir.

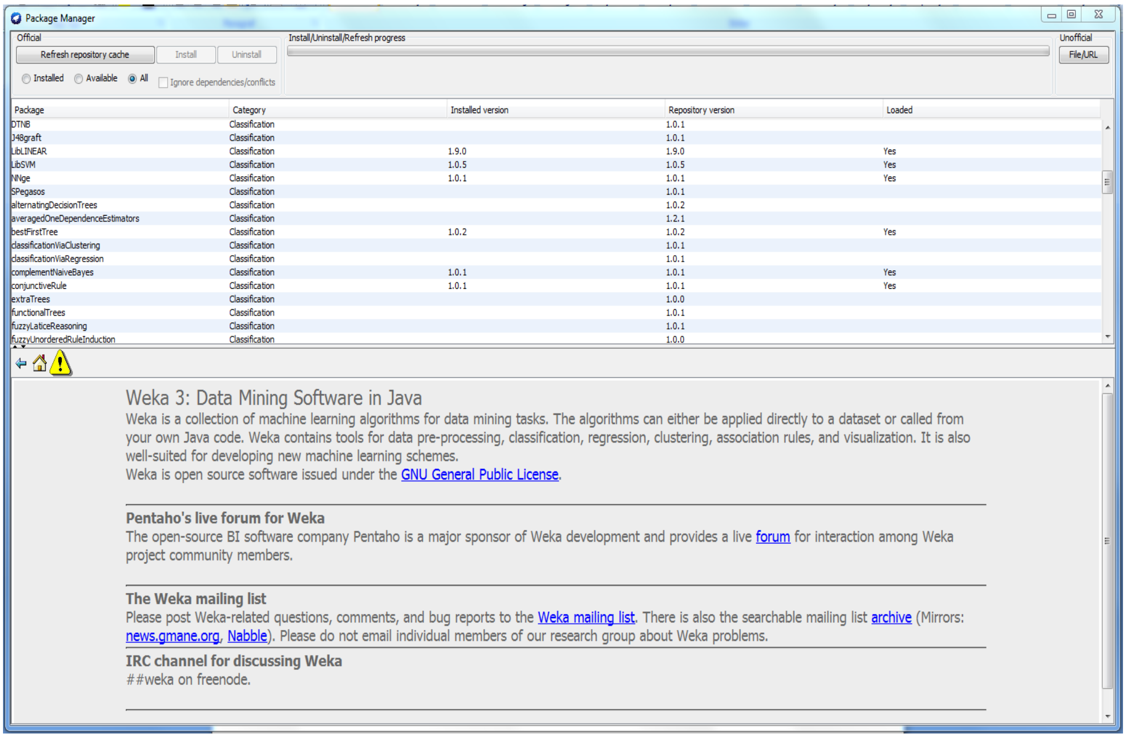


Şekil 6.1 Weka Ana Ekranı



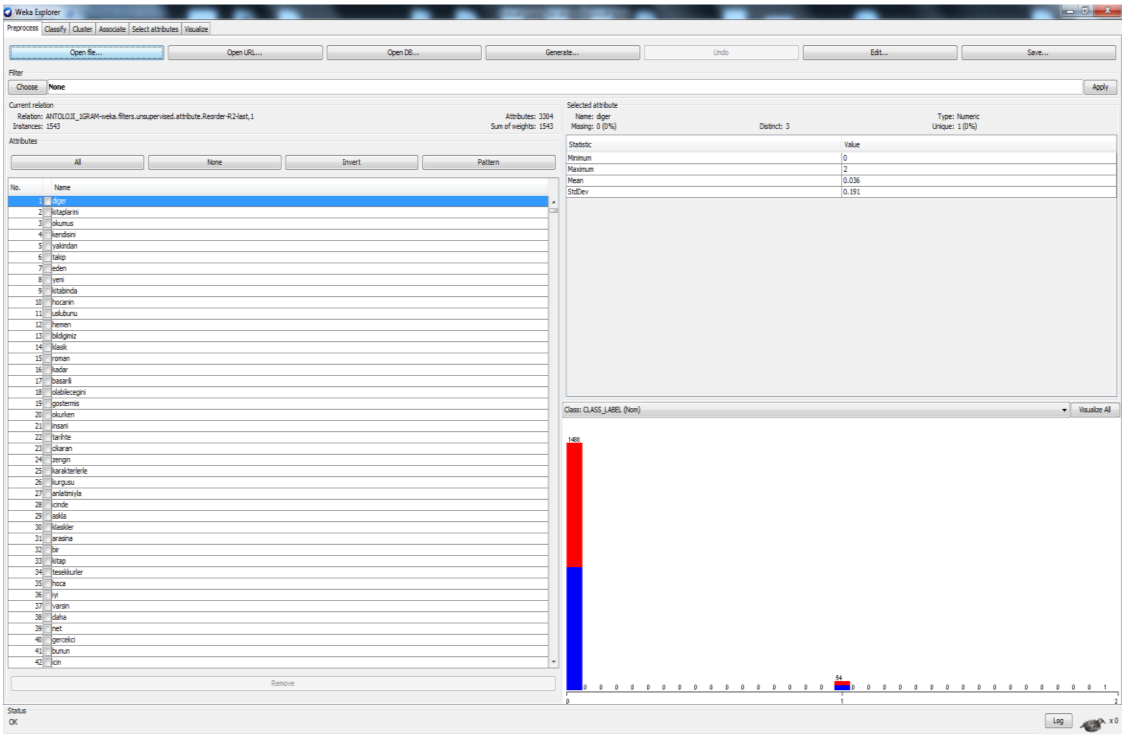
Şekil 6.2 Weka- Paket Yöneticisine Geçiş

Şekil 6.3'de kütüphane indirme penceresi gösterilmektedir.



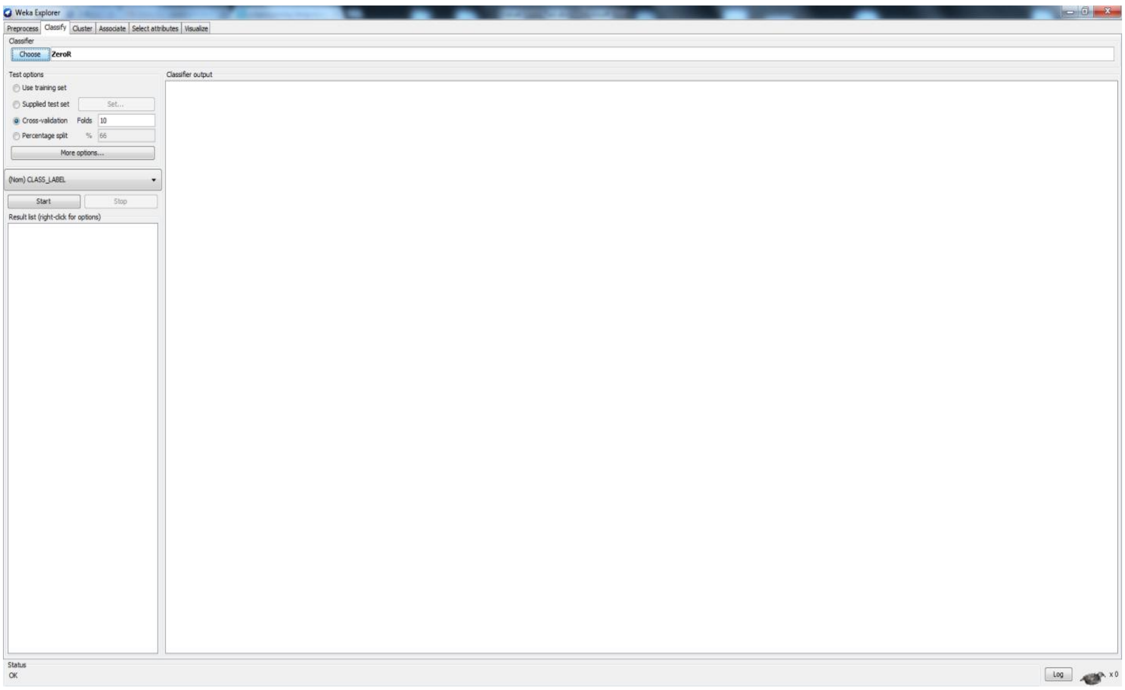
Şekil 6.3 Weka - Paket Yöneticisi

Şekil 6.1 deki Explorer alanına basılarak Weka programı analiz ve test ekranına geçilmektedir. Şekil 6.4'de Weka veri ön işleme ekranı gösterilmiştir. Bu ekran üzerinden testi ve analizi yapılmak istenen veri kümesi yüklenmektedir. Veri kümesi yüklendikten sonra veri kümesinin bilgileri ekran üzerinde gösterilir ve veri kümesinde ön işlemler bu ekran üzerinden gerçekleştirilebilir.



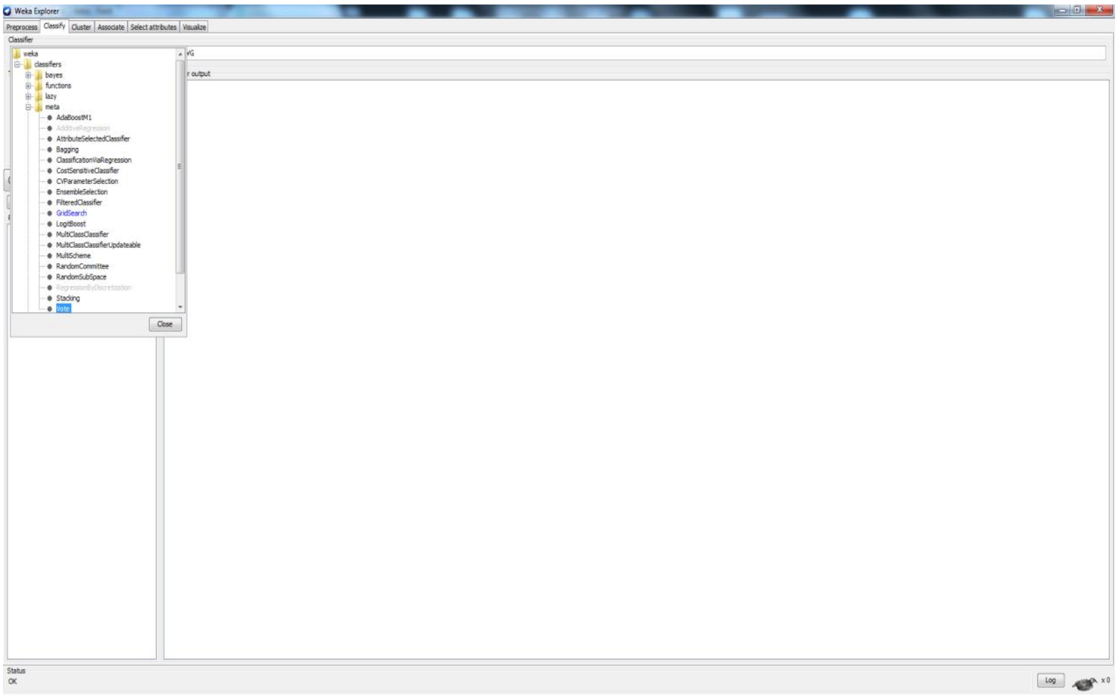
Şekil 6.4 Weka Ön İşleme Ekranı

Veri kümesi yüklendikten sonra sınıflandırma işlemlerini uygulanacağı Şekil 6.5’deki sınıflandırma sekmesine geçilmektedir. Sınıflandırma ekranının üst kısmında sınıflandırıcı seçim alanı bulunmaktadır.



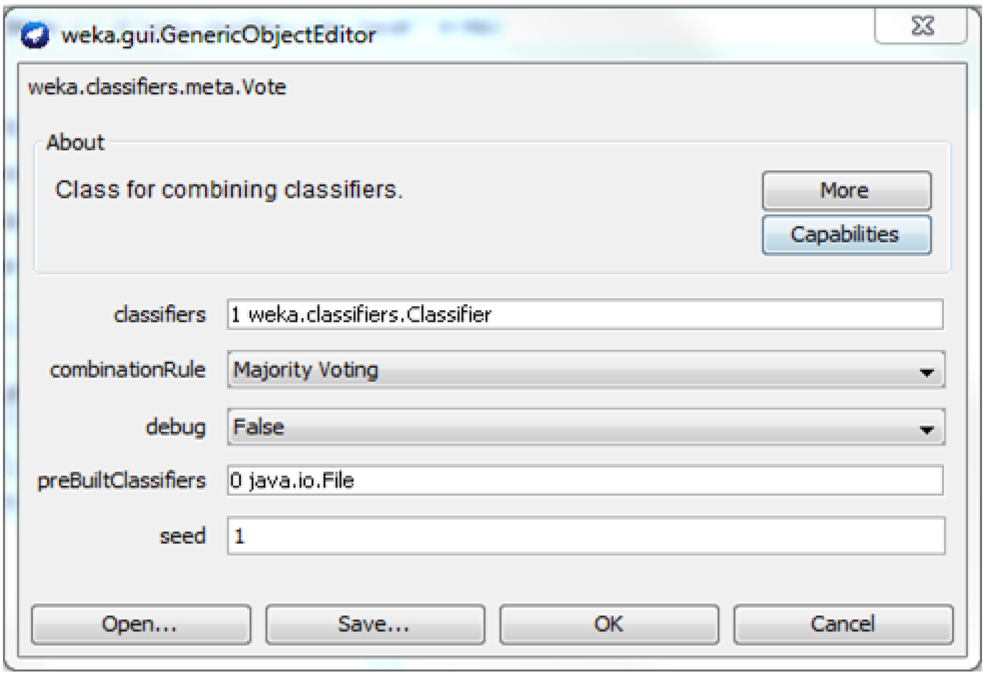
Şekil 6.5 Weka Sınıflandırma Ekranı

Şekil 6.6'de sınıflandırıcı seçme sekmesi açılmıştır ve meta sınıflandırıcılar içindeki Vote sınıflandırıcısı seçilmiştir.



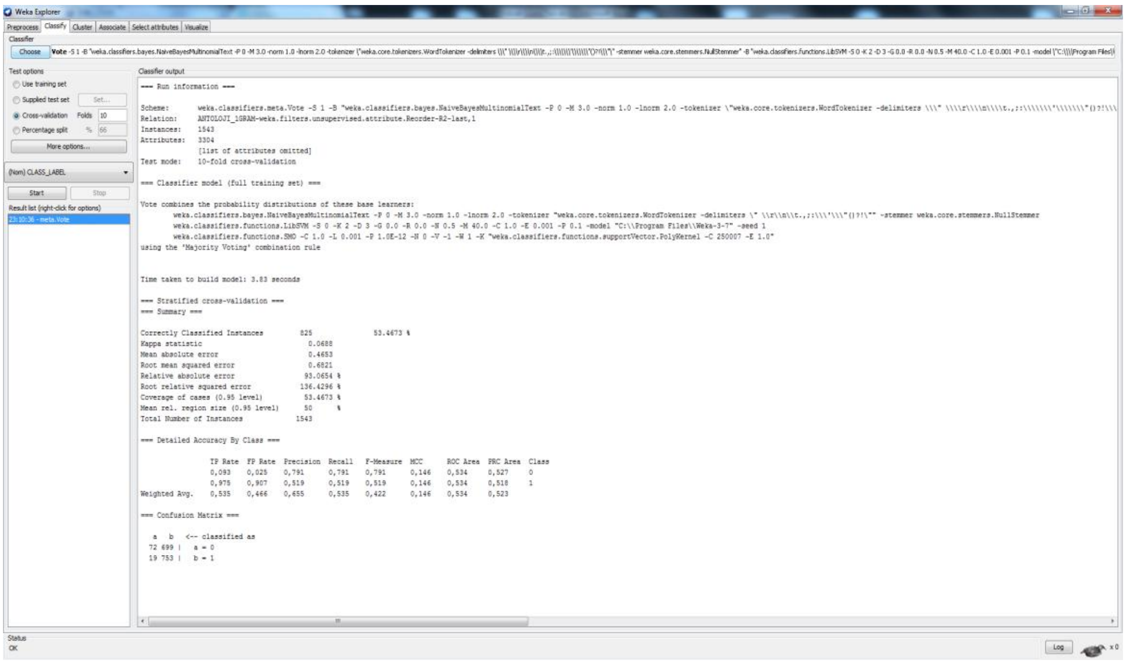
Şekil 6.6 Weka Sınıflandırıcı Seçim Ekranı

Vote sınıflandırıcı seçildikten sonra detaylarının ve içerisindeki çoklu sınıflandırıcıların seçilmesi için üzerine tıklanarak detay ekranı açılmaktadır. Şekil 6.7'de Vote sınıflandırıcısının detay ekranı görüntülenmektedir. Detay ekranındaki birleştirme kuralından (combination rule) Vote sınıflandırıcısının çalışma kuralı seçilmektedir. Tez kapsamında birleştirme kuralı olarak çoğunluk oylaması (majority voting) kullanılmıştır. Sınıflandırıcı (Classifier) alanından Vote çoklu sınıflandırıcısı içerisinde çalışması istenen sınıflandırıcılar seçilmektedir.



Şekil 6.7 Vote Sınıflandırıcısı Detay Ekranı

Vote çoklu sınıflandırıcısı yapılandırıldıktan sonra ana ekrana geçilir ve ekranın sol kısmındaki test seçenekleri kısmından testin ve analizin nasıl uygulanacağı seçilmektedir. Tez çalışmasında test seçeneği olarak 10-katlamalı çapraz doğrulama uygulanmıştır. Sınıflandırıcılar ve test seçeneği seçildikten sonra start düğmesine basılarak ekran üzerinde test ve analiz işlemi başlatılmaktadır. Test ve analiz sonuçları ekranda görüntülenmektedir. Şekil 6.8'de bu detaylar gösterilmektedir.

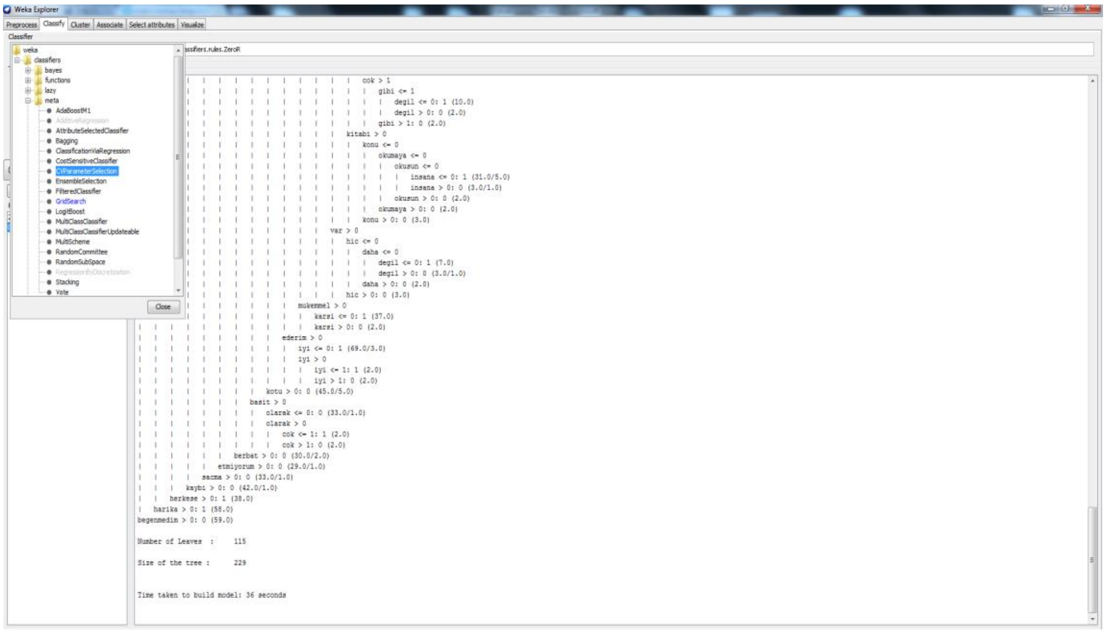


Şekil 6.8 Weka Analiz Ekranı

6.1.3 Weka ile Parametre Optimizasyonu

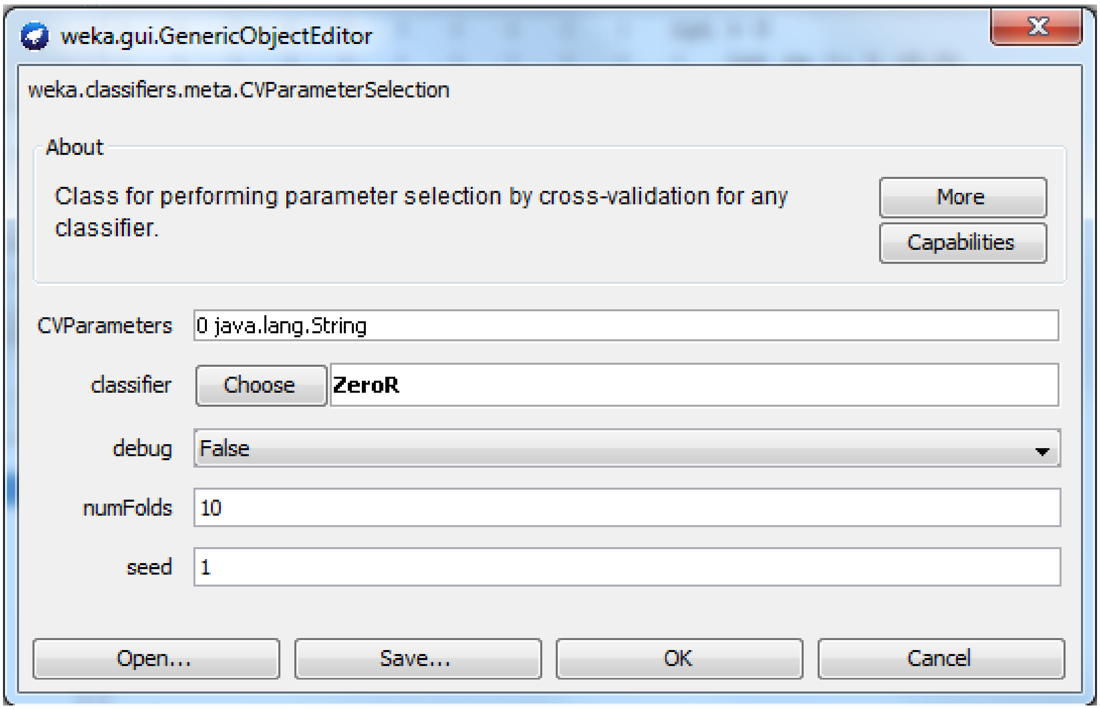
Weka programı içerisinde parametre optimizasyonu CVParameterSelection ve GridSearch algoritmaları ile yapılmaktadır. Tez çalışması kapsamında parametre optimizasyonu işlemi için CVParameterSelection algoritması tercih edilmiştir. Destek vektör makineleri ve karar ağaçları algoritmalarının parametre optimizasyonu yapılmış ve algoritmaların performansları arttırılmıştır.

Weka programı içerisinde parametre optimizasyon işlemi sınıflandırma ekranı üzerinden yapılmaktadır. Sınıflandırma ekranında Şekil 6.9'da görüldüğü gibi meta algoritmalar içerisinden CVParameterSelection algoritması seçilmelidir.



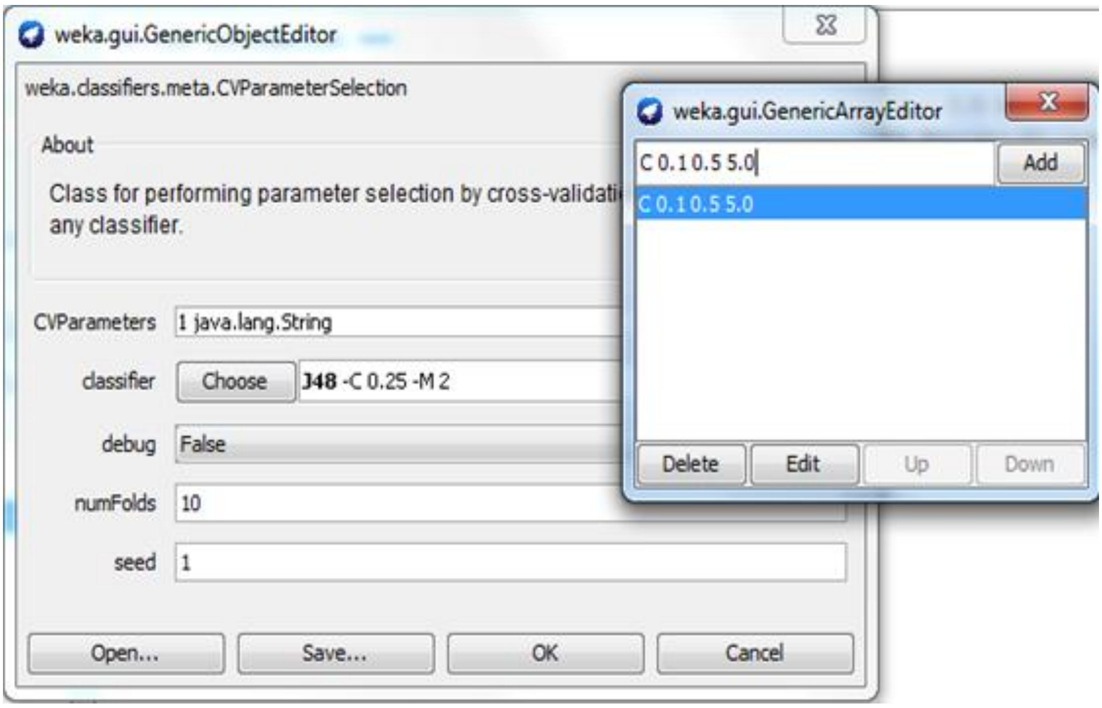
Şekil 6.9 Weka - Parametre Optimizasyonu

CVParameterSelection algoritması seçildikten sonra optimizasyonu yapılacak sınıflandırıcının seçilmesi ve parametre değerlerinin girilmesi için detay alanına girilmelidir. Şekil 6.10'da CVParameterSelection algoritmasının detay ekranı görülmektedir. CVParameters alanına parametre değerleri girilmeli, classifier seçeneğinden de parametre optimizasyonu yapılacak sınıflandırıcı seçilmelidir.



Şekil 6.10 Weka - Parametre Optimizasyonu Detayı

Şekil 6.11'de parametre optimizasyonu için sınıflandırıcı seçilmiş ve parametre girilmiş olarak gösterilmektedir.



Şekil 6.11 Weka - Parametre Optimizasyonu Veri Girişi

6.2 Bulgular

Bu bölümde tez çalışmasında yapılan testlerin ve analizlerin sonuçları sunulmaktadır. Tez çalışması, Türk dili için daha önce uygulanmamış özgün bir çoklu sınıflandırıcı makine öğrenmesi algoritması tasarlanmasına ve tek sınıflandırıcılar ile elde edilen doğruluk oranının arttırılmasına odaklanmıştır. Türk dili için daha önce önerilen tek sınıflandırıcılar ile yapılan çalışma, çoklu sınıflandırıcı yaklaşımı ile uygulanmış ve performansta iyileştirme gerçekleştirilmiştir.

Çoklu sınıflandırıcı yaklaşımı ile %86,13'lük doğruluk oranı elde edilmiştir. Bu oran çalışmanın performansta iyileştirme sağladığını ve geliştirilebileceğini göstermiştir. Çoklu sınıflandırıcı algoritması içerisinde üç sınıflandırıcı birleştirilerek, bu üç sınıflandırıcının doğruluk oranından daha performanslı bir sınıflandırıcı modeli geliştirilmiştir. Sınıflandırıcı modeli üç farklı veri kümesi üzerinde denenmiştir ve tüm veri kümelerinde performans artmıştır.

Tablo 6.4, Tablo 6.5 ve Tablo 6.6'da antoloji, beyazperde ve hepsiburada veri kümelerinin en yüksek doğruluk değerlerinin bulunduğu örnekler tablolarda gösterilmiştir. Çoklu sınıflandırıcı yaklaşımı, tüm veri kümelerinde en iyi sınıflandırıcı olan Naive Bayes algoritmasından daha yüksek doğruluk oranı sunabilmiştir.

Tez kapsamında, bu özgün sınıflandırıcı yaklaşımının yanında parametre optimizasyonu ile makine öğrenmesi algoritmalarının performansı arttırılmıştır.

Tablo 6.4 Antoloji Veri Kümesi Sonuçları 1

Kitap Yorumları	**Doğru Sınıflandırılan Veri Sayısı**	**Yanlış Sınıflandırılan Veri Sayısı**	**Doğruluk Oranı**
Naive Bayes	1319	224	85.48 %
Bagging - Destek Vektör Makineleri	1295	248	83.92 %
CVParameterSelection – Destek Vektör Makineleri	1279	264	82.89 %
Vote – Çoklu Sınıflandırıcı Algoritması	1329	214	**86.13 %**

Tablo 6.5 Antoloji Veri Kümesi Sonuçları 2

Kitap Yorumları	Doğru Sınıflandırılan Veri Sayısı	Yanlış Sınıflandırılan Veri Sayısı	Doğruluk Oranı
Naive Bayes	1319	224	85.48 %
Bagging - Destek Vektör Makineleri	1295	248	83.92 %
Karar Ağacı(J48)	1141	402	73.94 %
Vote – Çoklu Sınıflandırıcı Algoritması	1326	217	**85.93 %**

Tablo 6.6 BeyazPerde Veri Kümesi Sonuçları 1

Sinema Yorumları	Doğru Sınıflandırılan Veri Sayısı	Yanlış Sınıflandırılan Veri Sayısı	Doğruluk Oranı
Naive Bayes	1847	390	82.56 %
Destek Vektör Makineleri	1791	446	80.06 %
CVParameterSelection – Destek Vektör Makineleri	1813	424	81.04 %
Vote – Çoklu Sınıflandırıcı Algoritması	1874	363	**83.77 %**

Tablo 6.7 BeyazPerde Veri Kümesi Sonuçları 2

Sinema Yorumları	Doğru Sınıflandırılan Veri Sayısı	Yanlış Sınıflandırılan Veri Sayısı	Doğruluk Oranı
Naive Bayes	1847	390	82.56 %
Bagging - Destek Vektör Makineleri	1800	437	80.46 %
CVParameterSelection – Destek Vektör Makineleri	1813	424	81.04 %
Vote – Çoklu Sınıflandırıcı Algoritması	1872	365	**83.68 %**

Tablo 6.8 HepsiBurada Veri Kümesi Sonuçları 1

Alışveriş Yorumları	Doğru Sınıflandırılan Veri Sayısı	Yanlış Sınıflandırılan Veri Sayısı	Doğruluk Oranı
Naive Bayes	4188	1069	79.66 %
Destek Vektör Makineleri	3948	1309	75.09 %
CVParameterSelection – Destek Vektör Makineleri	4161	1094	79.15 %
Vote – Çoklu Sınıflandırıcı Algoritması	4210	1047	**80.08 %**

Tablo 6.9 HepsiBurada Veri Kümesi Sonuçları 2

Alışveriş Yorumları	Doğru Sınıflandırılan Veri Sayısı	Yanlış Sınıflandırılan Veri Sayısı	Doğruluk Oranı
Naive Bayes	4188	1069	79.66 %
Bagging - Destek Vektör Makineleri	4021	1236	76.48 %
CVParameterSelection – Destek Vektör Makineleri	4161	1094	79.15 %
Vote – Çoklu Sınıflandırıcı Algoritması	4204	1053	**79.96 %**

Tablo 6.7, Tablo 6.8 ve Tablo 6.9’da parametre optimizasyonu ile performansı arttırılmış olan makine öğrenmesi algoritmaları gösterilmiştir. Parametre optimizasyonu bu makine öğrenmesi algoritmaları için başarılı şekilde performansta artış sağlamıştır.

Tablo 6.10 Antoloji Veri Kümesi - Parametre Optimizasyonu

Kitap Yorumları	Normal Doğruluk Oranı	Parametre Optimizasyonu Yapılmış Doğruluk Oranı
Destek Vektör Makineleri	51.58 %	82.89 %

Tablo 6.11 BeyazPerde Veri Kümesi - Parametre Optimizasyonu

Sinema Yorumları	Normal Doğruluk Oranı	Parametre Optimizasyonu Yapılmış Doğruluk Oranı
Destek Vektör Makineleri	56.68 %	81.04 %

Tablo 6.12 HepsiBurada Veri Kümesi - Parametre Optimizasyonu

Alışveriş Yorumları	Normal Doğruluk Oranı	Parametre Optimizasyonu Yapılmış Doğruluk Oranı
Destek Vektör Makineleri	66.44 %	79.15 %

7 SONUÇ VE GELECEK ÇALIŞMALAR

7.1 SONUÇ

Bu tez çalışması, Türk dili için farklı veriler üzerinde çoklu sınıflandırıcı yöntemler kullanılarak sınıflandırma çalışmasındaki doğruluk oranının arttırılabileceğini deneysel olarak ortaya koymuştur.

Ayrıca parametre optimizasyonu işlemi ile sınıflandırıcıların performansının yükseltilebileceğini ve sınıflandırma çalışmasının doğruluk oranının arttırılabileceğini göstermiştir.

Üç sınıflandırıcı birleştirilerek geliştirilen özgün çoklu sınıflandırıcı yaklaşımının, performans konusunda iyileştirme sağladığı analizlerle tablolar halinde sunulmuştur.

Türk dili için yapılan önceki çalışmalarda [20] Naive Bayes sınıflandırıcısı ile elde edilen doğruluk oranı, %86.13 doğruluk oranına yükseltilmiştir. Bu doğruluk oranı, çoklu sınıflandırıcı yaklaşımının ve parametre optimizasyon işleminin performans artışı sağladığını ve Türk dili için yeni gelişmekte olan duygu analizi konusunda birçok çalışmada kullanılabileceğini ortaya koymuştur.

Bu tez çalışması sırasında rastlanan bir konu da, Weka programının performanslı çalışabilmesi için RAM gereksiniminin önemidir. Weka programı doğrudan RAM üzerinde çalışan bir programdır ve RAM kapasiteniz ne kadar fazla ise o kadar hızlı işlem süresi elde edilmektedir. Weka programı kurulumu yapıldığında 1024 maksimum heap ayarında gelmektedir. Bu ayarıyla çalıştırıldığında bazı büyük veri kümelerinin çalışması sırasında uygulama kapanmaktadır. Bu yüzden çalışma sırasında bu ayar makinedeki en yüksek RAM miktarını kullanacak şekilde güncellenmiş ve program en yüksek performans ile tüm veri kümelerini çalıştırabilmiştir. Weka programının RAM gereksinimi ve maksimum heap ayarı tez çalışması sırasında rastlanan bir konudur ve ileride yapılacak çalışmalara yardımcı olması açısından ek bilgi olarak belirtilmesinde fayda görülmüştür.

7.2 Gelecek Çalışmalar

Bu tez kapsamında, Türk dili içi daha önce incelenmemiş bir yaklaşım olan çoklu sınıflandırıcı yaklaşımı denenmiş ve başarı sağlanmıştır. Bu başarı çoklu sınıflandırıcı yaklaşımının başarılı bir makine öğrenmesi yaklaşımı olduğunu ve birçok çalışmada kullanılabileceğini göstermiştir. Gelecek çalışma olarak, sunulan çoklu sınıflandırıcı yaklaşımı farklı dillerdeki veri kümelerinde denenebilir ve başarısı gözlenebilir. Çoklu sınıflandırıcı yaklaşımı bu tez kapsamında üç tekil sınıflandırıcı ve oy çoğunluğu birleştirme yöntemi ile kullanılmıştır. Önümüzdeki çalışmalarda sınıflandırıcı sayısı arttırılabilir ve oy çoğunluğu yöntemi yerine yeni birleştirme kuralları incelenebilir.

Ayrıca, bu çalışmanın performansına büyük katkı yapan parametre optimizasyon algoritmaları iyileştirilerek daha iyi sonuçlar alınabilir.

Çalışmalara ek olarak önerilen yöntem, İngilizce veri kümeleri için de sınanarak literatür ile kıyaslaması sağlanabilir.

8 Referanslar

[1] Liu, B. (2012). Sentiment Analysis and Opinion Mining, ISBN: 9781608458851

[2] Nasukawa, T. ve Yi, J. (2003). Sentiment analysis: Capturing favorability using natural language processing. In Proceedings of the K-CA P-03, 2nd International Conference on Knowledge Capture.

[3] Dave, K., Lawrence, S., Pennock D.M. (2003). Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In Proceedings of International Conference on World Wide Web

[4] Das, S. ve Chen, M. (2001) Yahoo! for Amazon: Extracting market sentiment from stock message boards. In Proceedings of APFA-2001.

[5] Morinaga, S., Yamanishi, K., Tateishi, K., Fukushima T. (2002) Mining product reputations on the web. In Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.

[6] Pang, B., Lee, L., Vaithyanathan, S. (2002). Thumbs up? Sentiment classification using machine learning techniques. In Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP) , (pp. 79–86).

[7] Tong, R.M. (2001). An operational system for detecting and tracking opinions in on-line discussion. In Proceedings of SIGIR Workshop on Operational Text Classification.

[8] Turney, P. ve Littman, M.L. (2002). Unsupervised Learning of Semantic Orientation from a Hundred-Billion-Word Corpus, NRC/ERB-1094.

[9] Wiebe, J. (2000). Learning subjective adjectives from corpora. In Proceedings of National Conf. on Artificial Intelligence.

[10] Wilson, T., Wiebe, J., Hwa, R. (2004). Just how mad are you? Finding strong and weak opinion clauses. In Proceedings of National Conference on Artificial Intelligence.

[11] Hu, M. ve Liu, B. (2004). Mining and summarizing customer reviews. In Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.

[12] Kim, S.M. ve Hovy, E. (2004). Determining the sentiment of opinions. In Proceedings of Interntional Conference on Computational Linguistics.

[13] Wiebe, J. Ve Riloff, E. (2005). Creating subjective and objective sentence classifiers from unannotate texts. Computational Linguistics and Intelligent Text Processing. (pp. 486–497).
[14] https://blog.twitter.com/2013/celebrating-twitter7
[15] http://www.cisco.com/web/TR/news/press/archive/2011/020611.html
[16] Jansen, J. Online Product Research, Pew Research Center's Internet & American Life Project, http://www.pewinternet.org/Reports/2010/Online-Product-Research.aspx
[17] Bollen, J., Mao, H., Zeng, X.J. (2010). Twitter mood predicts the stock market.
[18] Rainie, L., Smith, A., Schlozman, K.L., Brady, H., Verba, S. Social Media and Political Engagement, http://pewinternet.org/Reports/2012/Political-engagement.aspx
[19] Eroğul, U. (2009). Sentiment Analysis in Turkish, METU Master's Thesis.
[20] Çelik, H. (2013). Sentiment Analysis for Turkish Language, IKU Master's Thesis.
[21] Carbonell, J. (1979). Subjective Understanding: Computer Models of Belief Systems. PhD thesis, Yale.
[22] Pang, B. ve Lee, L. (2005). Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In Proceedings of the ACL, (pp. 115–124).
[23] Liu, Y., Huang, X., An, A., Yu, X. (2007). ARSA: a sentiment-aware model for predicting sales performance using blogs. In Proceedings of ACM SIGIR Conf. on Research and Development in Information Retrieval.
[24] Tumasjan, A., Sprenger, T.O., Sandner P.G., Welpe, I.M. (2010). Predicting elections with twitter: What 140 characters reveal about political sentiment. In Proceedings of the International Conference on Weblogs and Social Media
[25] Bollen, J., Mao, H., Zeng X.J. (2011). Twitter mood predicts the stock market. Journal of Computational Science.
[26] Li, S.S., Huang, C.R., Zong, C.Q. (2010). Multi-Domain Sentiment Classification with Classifier Combination, Journal of Computer Science and Technology.
[27] Li, S., Zong, C., Wang, X. (2007). Sentiment Classification through Combining Classifiers with Multiple Feature Sets, Natural Language Processing and Knowledge Engineering

[28] Kittler, J., Hatef, M., Duin R.P.W., Matas, J. (1998). On Combining Classifiers, IEEE Transactions on Pattern Analysis and Machine Intelligence.

[29] Simon, P. (2013). Too Big to Ignore: The Business Case for Big Data. (pp. 89). ISBN 978-1118638170.

[30] Software For Predictive Modeling and Forecasting , <http://www.dtreg.com>

[31] Vapnik, V. ve Cortes, C. (1995). Support Vector Networks, AT&T Labs-Research.

[32] http://aozsoyler.blogspot.com/2011/04/karar-agaclar_30.html

[33] Wikipedia, the free encyclopedia, <http://en.wikipedia.org>

[34] http://weka.wikispaces.com/Optimizing+parameters

[35] Breiman, L. (1996). Bagging Predictors, Journal Machine Learning, 24, (pp. 123–140).

[36] Freund, Y. ve Schapire, R.E. (1997). A Decision-Theoretic Generalization of on-Line Learning and an Application to Boosting, Journal of Computer and System Sciences 55, (pp. 119-139).

[37]http://www.bilgisayarkavramlari.com/2013/03/31/k-fold-cross-validation-k-katlamali-carpraz-dogrulama/

[38] Data Mining Software in Java, <http://www.cs.waikato.ac.nz/ml/weka/>

9 Özgeçmiş

16.05.1987 tarihinde Denizli'de doğdum. İlkokulu Muğla'nın Ortaca ilçesinde, ortaokulu Denizli'de tamamladım. Liseyi 2004 yılında Denizli'de bitirdikten sonra İstanbul Kültür Üniversitesi Bilgisayar Mühendisliği bölümünden 2010 yılında mezun oldum. Mezuniyetimden bu yana Ozon Giyim A.Ş-Defacto bünyesinde yazılım uzmanı olarak çalışmaktayım. Microsoft teknolojileri ile yazılım projeleri gerçekleştirmekteyim. Temel ilgi alanlarım yazılım teknolojileri, sinema, kitap, futbol ve tarihtir.