T.C. DOĞUŞ ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ
BİLGİSYAR MÜHENDİSLİĞİ ANABİLİM DALI

KELİME KULLANIM ORANLARI VE KULLANICI İSTATİSTİKLERİ KULLANILARAK TÜRKÇE TWITTER VERİSİ ÜZERİNDE DUYGU ANALİZİ

YÜKSEK LİSANS TEZİ

CEM GÜMÜŞ

201091001

Tez Danışmanı:
PROF. DR. SELİM AKYOKUŞ

İstanbul, Haziran 2017

T.C. DOĞUŞ ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ
BİLGİSYAR MÜHENDİSLİĞİ ANABİLİM DALI

KELİME KULLANIM ORANLARI VE KULLANICI İSTATİSTİKLERİ KULLANILARAK TÜRKÇE TWITTER VERİSİ ÜZERİNDE DUYGU ANALİZİ

YÜKSEK LİSANS TEZİ

CEM GÜMÜŞ

201091001

JÜRİ ÜYELERİ
PROF. DR. SELİM AKYOKUŞ (DANIŞMAN)
PROF. DR. MİTAT UYSAL
YRD. DOÇ. DR. MURAT CAN GANİZ

İstanbul, Haziran 2017

ÖNSÖZ

Desteğini ve yardımlarını esirgemeyen tez danışmanım Prof. Dr. Selim AKYOKUŞ'a ve Yrd. Doç. Dr. Murat Can GANİZ'e kıymetli vakitlerini ayırdıkları ve sabırla yol gösterdikleri için çok teşekkür ederim.

Tez çalışmamın gizli kahramanları Doğuş Üniversitesi öğrencilerine katkılarından dolayı teşekkür ederim.

Eşim Ayfer, çocuklarım Deniz ve Bora'ya destekleri ve hoşgörüleri için çok teşekkür ederim.

İstanbul, Haziran 2017 Cem GÜMÜŞ

ÖZET

İnternetin hızla gelişmesi ve mobil cihazların kullanımındaki artış ile birlikte sosyal ağların kullanımı son yıllarda büyük bir artış göstermiştir. İnsanların kişisel fikir, görüş ve önerilerini başka insanlar ile paylaşmak ve başka insanların bir konu üzerindeki görüş ve önerilerini öğrenmek istemeleri, sosyal medyayı önemli bir bilgi deposu haline getirmiştir. Bu bilgi deposu, araştırmacıların yanı sıra geleneksel yöntemlerle müşteriye ulaşmanın yeterli olmadığını gören firmaların da ilgisini büyük ölçüde çekmektedir. Bu bilgi deposunda yapılan çalışmalar sonucunda firmalar, müşterilerinin ürün ve hizmetleri hakkındaki görüş ve düşüncelerini öğrenebilmekte, elde edilen verileri sınıflandırarak ürün ve hizmetlerini geliştirmede kullanabilmektedirler. Sosyal ağlardan elde edilecek veriler ile yapılacak çalışmalarda en etkili yöntemlerden biri duygu analizidir. Duygu analizi, bu bilgi deposundan elde edilen metinsel verilerin yansıttığı duyguların, bilgisayar yardımıyla otomatik olarak tespit edilmesini amaçlamaktadır.

Günümüzde Facebook, Instagram, Tumblr, Twitter gibi birçok popüler sosyal ağ bulunmaktadır. Mesajların 140 karakter ile sınırlanmış olması, bu sınırlandırma sayesinde paylaşılmak istenen bilginin etkin ve hızlı bir şekilde anlatılması Twitter'ı sosyal ağlar arasında popüler bir hale getirmiştir. Duygu analizi konusunda İngilizce için yapılmış birçok çalışma olmasına karşın Türkçe için yapılan çalışma sayısı sınırlıdır. Türkçe duygu analizi konusunda yeterli çalışma olmamasından dolayı bu tez çalışmasında Türkçe metinler için duygu analizi çalışması yapılmıştır. Bu tez kapsamında yapılacak çalışmada kullanılacak Türkçe mesajlar, popülaritesi, etkin kullanımı ve sağladığı API'den dolayı Twitter sosyal ağından toplanmıştır. Twitter sosyal ağından toplanan tweetler pozitif, negatif ve nötr olmak üzere 3 sınıfa ayrılmıştır. Bu etiketli veriler kullanılarak dengesiz ve dengeli veri kümeleri oluşturulmuştur. Çalışmanın başarısını arttırmak için yeni özellikler veri kümelerine eklenmiştir. Oluşan veri kümeleri makine öğrenmesi (MÖ) yöntemlerinden denetimli öğrenme (supervised) ve yarı-denetimli öğrenme (semi-supervised) yöntemleri ile analiz edilmiştir. Elde edilen sonuçlar karşılaştırılmış ve yeni eklenen özelliklerin deney sonuçlarına etkileri incelenmiştir.

Anahtar Kelimeler: Twitter; Duygu Analizi; Makine Öğrenmesi; Duygu Sınıflandırma; Özellik Çıkartma;

ABSTRACT

With the rapid growth of the Internet and the increase in the use of mobile devices, the use of social networks has increased significantly in recent years. The sharing of people's personal ideas, opinions and suggestions with other people and the desire of other people to learn opinions and suggestions on a topic have made social media an important information repository. This information repository attracts a great deal of interest from companies that see that it is not enough to reach customers with traditional methods as well as researchers. As a result of the studies conducted in this information warehouse, companies can learn opinions and thoughts about customers' products and services, classify the obtained data and use them to improve their products and services. Sentiment analysis is one of the most effective methods to work with data obtained from social networks. Sentiment analysis aims to automatically detect emotions reflected by textual data obtained from this information repository by computer.

Today, there are many popular social networks like Facebook, Instagram, Tumblr, Twitter. By limiting the number of messages to 140 characters, this limitation makes Twitter efficient and fast to share information popular with social networks. Despite the fact that there are many works on English for sentiment analysis, the number of works done for Turkish is limited. Since there is not sufficient study on Turkish sentiment analysis, sentiment analysis study was done for Turkish texts in this thesis study. In this thesis, the Turkish messages to be used in the study are gathered from the Twitter social network because of the popularity, the effective use and the Application Programming Interface (API) that it provides. Tweets collected from Twitter social network are divided into 3 classes as positive, negative and neutral. Using these labeled data, unbalanced and balanced data sets were created. New features have been added to the data sets to enhance the performance of the work. The resulting data sets were analyzed by supervised learning and semi-supervised methods of machine learning methods. The results obtained were compared and the effects of the newly added properties on the test results were examined.

Key Words: Twitter; Sentiment Analysis; Sentiment Classification; Machine Learning; Feature Extraction;

İÇİNDEKİLER

ŞEKİLLER

TABLOLAR

SEMBOLLER

$\|DT\|$	Tüm veri kümesindeki terim sayısı
$\|DP\|$	Pozitif veri kümesindeki terim sayısı
$\|DN\|$	Negatif veri kümesindeki terim sayısı
$\|DR\|$	Nötr veri kümesindeki terim sayısı
d_i	Tüm veri kümesindeki i'nci mesaj
$\|d_i\|$	i'nci mesajdaki terim sayısı
$d_{i,j}$	i'nci mesajdaki j'nci terim
$f_D(t)$	t teriminin tüm veri kümesindeki toplam sayısı
$f_{DP}(t)$	t teriminin pozitif veri kümesindeki toplam sayısı
$f_{DN}(t)$	t teriminin negatif veri kümesindeki toplam sayısı
$f_{DR}(t)$	t teriminin nötr veri kümesindeki toplam sayısı
$\mathrm{U_t(i)}$	Bir kullanıcının toplam tweet sayısı
$\mathrm{U_p(i)}$	Bir kullanıcının toplam pozitif tweet sayısı
$\mathrm{U_n(i)}$	Bir kullanıcının toplam negatif tweet sayısı
$\mathrm{U_r(i)}$	Bir kullanıcının toplam nötr tweet sayısı
W	Ağırlıklandırma fonksiyonu
d	Bir doküman
C	Bir terim
D	Veri kümesindeki tüm dokümanlar
N	Toplam doküman sayısı
WT	Terim Kullanım Oranı
WP	Terim Pozitif Kullanım Oranı
WN	Terim Negatif Kullanım Oranı
WR	Terim Nötr Kullanım Oranı

KISALTMALAR

DDİ	Doğal Dil İşleme
DA	Duygu Analizi (Sentiment Analysis)
MÖ	Makine Öğrenmesi
MS	Metin Sınıflandırma
POS	Cümlenin Öğeleri (Part Of Speech)
NB	Naïve Bayes
SVM	Destek Vektör Makinesi (Support Vector Machine)
RF	Rastgele Orman (Random Forrest)
J48	Karar Ağacı (Desicion Tree)
URL	Standart Kaynak Bulucu(Uniform Resource Locator)
ARFF	Özellik–İlişki Dosya Formatı (Attribute-Relation File Format)
API	Uygulama Programlama Arayüzü (Application Programming Interface)
TF	Terim Frekansı (Term Frequency)
IDF	Ters Doküman Frekansı (Inverse Document Frequency)
IMDB	İnternet Film Veri Tabanı (Internet Movie Database)

1. GİRİŞ

1.1. Tezin Amacı ve Kapsamı

Günümüzde sosyal medya; mobil cihazlar ve internetin hızla gelişmesi ile birlikte insanların hayatında büyük bir önem kazanmaya başlamıştır. Birçok insanın hayatında sosyal ağlar, hava, su gibi vazgeçilemez bir öneme sahiptir. Sosyal ağlar sayesinde insanlar kendi görüş ve önerilerini paylaşırken, başkalarının da bir ürün veya hizmet hakkında ne düşündüğünü bilmek istemektedirler. İnsanların bir konu hakkında çevrelerinden alabilecekleri bilgi ya da danışabilecekleri insan sayısı sınırlıdır. Bu bağlamda; daha fazla bilgi, görüş ve öneri elde edebilmek için sosyal ağlar büyük ölçüde kullanılmaktadır. Bunun yanında, insanların görüş ve önerilerini paylaşmaları, bu görüş ve önerilerin başkaları tarafından değerli kabul edilerek, önem arz etmesi insanlardaki psikolojik tatmini de arttırmaktadır. Bu nedenlerin sonucu olarak yapılan paylaşımlar, sosyal ağları önemli ve büyük bir bilgi deposu haline getirmiştir. Bu bilgi deposu araştırmacıların yanı sıra şirketlerin de ilgisini çekmektedir. Şirketler, bu büyük bilgi deposunu, müşterilerin kendileri ve rakipleri hakkındaki düşüncelerini öğrenebilmek, ürün ve hizmetlerini geliştirebilmek amacıyla kullanmaktadırlar. Ayrıca bu büyük veri deposu üzerinde yapılan çalışmalar yeni ürün ve hizmet fikirlerinin ortaya çıkmasını da sağlayabilmektedir.

Sosyal ağlar önemli bir veri kaynağı olmakla beraber bu verilerin işlenmeden kullanılması zordur. Yazılan mesajların içinde bulunan yazım hataları, kısaltmalar, duygu ifadeleri ve günlük konuşma dilinin kullanılması, bu veriler üzerinde çalışmayı zorlaştırmaktadır. Bu büyük veriler üzerinde çalışma yapabilmek için metin sınıflandırma ve doğal dil işleme teknikleri kullanmak gerekmektedir. Günümüzde yazım yardımcı araçlarının geliştirilmesi, yazım yanlışlarının düzeltilmesi, bir metnin özetini çıkarma, metnin içerdiği bilgiyi çıkarma, bilgiye erişim, metni anlama, bilgisayarla sesli etkileşim, çeviri ve soru yanıtlama gibi birçok konu doğal dil işlemenin (DDİ) kapsamındadır.

Duygu analizi (DA) konusunda İngilizce için birçok çalışma yapılmakla beraber, Türkçe için yapılan duygu analizi çalışması yeterli sayıya ulaşmamaktadır. Bu nedenle Makine Öğrenmesi yöntemleri ile Türkçe Twitter verileri kullanılarak duygu analizi çalışması yapılması ve mesajlardaki semantik bilginin ortaya çıkarılması bu tezde hedeflenmiştir.

1.2. Tezin Yöntemi

Bu tez ile metin sınıflandırma (MS), doğal dil işleme (DDİ) ve duygu analizi (DA) teknikleri kullanılarak Türkçe Twitter verilerinin analizi hedeflenmiştir. Bu amaçla Twitter'dan verileri toplayan bir yazılım geliştirilmiş ve elde edilen verilerden deneylerde kullanılmak üzere veri kümeleri oluşturulmuştur. Bu veri kümelerine, başarı oranını arttırmak için yeni özellikler eklenmiştir. Eklenen yeni özelliklerin sonuçlara etkisini bulabilmek için çalışmalar yapılmış ve sonuçlar değerlendirilmiştir.

2. KAYNAK TARAMASI

Duygu Analizi (DA), metnin ifade ettiği duyguyu otomatik olarak tespit etme ve yorumlama işlemine verilen isimdir. Duygu analizi (DA) çalışmaları doğal dil işleme (DDİ), makine öğrenmesi (MÖ), hesaplamalı bilim, sembolik teknikler…vb gibi yaklaşımları kullanır.

Duygu analizi (DA) problemleri ile ilgili birçok akademik çalışma yapılmakla beraber, Duygu analizi (DA) alanında ilk çalışmalardan biri 2002 yılında İnternet Film Veri Tabanından (IMDB) alınan veriler üzerinde yapılmıştır (Pang vd., 2002). Bu çalışmada unigram, bi-gram, Cümlenin Öğeleri (POS) ve makine öğrenmesi kullanılarak metinler pozitif ve negatif olarak sınıflandırılmıştır. Bu çalışma için makin öğrenmesi yöntemlerinden Naïve Bayes, Maximum Entropi ve Destek Vektör Makinesi (SVM) kullanılmış, çalışma sonucu olarak %82,9 ile en iyi sonuç Destek Vektör Makinesi (SVM)'de elde edilmiştir. Terim olarak Duygu Analizi ise ilk olarak 2003 yılında yapılan çalışmada kullanılmaya başlanmıştır (Nasukawa ve Yi, 2003).

Pang ve Lee (2004) yaptıkları çalışmada, veriyi öznel ve nesnel olarak sınıflandırmışlardır. Bu çalışma sonrası bu öznel verileri olumlu ve olumsuz olarak sınıflandırmışlardır. Destek Vektör Makinesi (SVM) ve Naïve Bayes (NB) sınıflayıcıları ile %85 başarı elde etmişlerdir.

Hu ve Liu (2004) ticari ürünler hakkında yapılan yorumları olumlu ve olumsuz olarak derlemişlerdir. Bu veriler üzerinde sınıflandırma yaparak %70 başarı elde etmişlerdir. Bunun dışında, bir miktar olumlu ve olumsuz kelime alınarak WordNet (Miller vd., 1990) yardımı ile kelime kümesi genişletilmiş ve deneylerde kullanılmıştır. Sonuçta %84 gibi bir başarı elde edilmiştir.

Mishne ve Glance (2006), yaptıkları çalışmada İnternet Film Veri Tabanı (IMDB) verisini kullanarak, vizyona giren filmlerin başarısını tahmin etmeye çalışmışlardır.

Nguyen vd. (2012), yaptıkları çalışmada sosyal medya dinamikleri ile insan algısının zamanla değişimini %85 başarı ile tahmin etmişlerdir.

Günümüzde Twitter üzerinden toplanan veriler ile Doğal Dil İşleme (DDİ) ve Veri Madenciliği alanlarında birçok çalışma yapılmıştır. Szomszor vd. (2010), Mayıs 2009 ve Aralık 2009 tarihleri arasında atılan tweetler üzerinde çalışma yaparak salgın hastalıkları tahmin etmişlerdir. Bian vd. (2012), Mayıs 2009 ve Ekim 2010 tarihleri arasında atılan tweetlerden yaptıkları çalışmada ilaçlar ve onların bilinmeyen yaz etkilerini analiz etmişlerdir.

Claster vd. (2010), yaptıkları çalışmada, bir turistik bölgeyi ziyaret eden turistlerin attığı 70 Milyon tweet üzerinde algı analizi yapmışlardır.

Go vd. (2009), otomatik olarak Twitter mesajlarını sınıflandırmak için bir yaklaşım geliştirmişlerdir. Unigram, bigram ve unigram-bigram birleştirerek deneyler yapmışlar ve %83 ile Maximum Entropi algoritması ile en iyi sonucu elde etmişlerdir.

İngilizce ile yapılan Duygu Analizi (DA) çalışmalarına karşı Türkçe için sınırlı sayıda çalışma yapılmıştır. Türkçe veriler ilk Duygu Analizi (DA) çalışmasını Eroğlu (2009) yapmıştır. N-gram modelini kullanarak film yorumlarını olumlu ve olumsuz olarak sınıflandırmış ve bu işlem soncu %85 başarı elde etmiştir.

Şimşek ve Özdemir (2012), çalışmalarında Twitter mesajları ile borsadaki değişim arasındaki ilişki incelenmiştir. Sonuç olarak bu ilişkinin %45 oranında mevcut olduğu bulunmuştur.

Akbaş (2012), Twitter üzerinden veri toplamış ve makine öğrenmesi (MÖ) yöntemleri ile çalışma yapmıştır. Özellik seçiminde hibrit yapı seçilmiş ve deneyler yapılmıştır. %85 civarı başarı elde edilmiştir.

Çakmak vd. (2012), Türk dili için kelime kökleri ve cümle bazında deneyler yapılmıştır. Kelime kökleri ve cümleler arasında yüksek ilişki tespit edilmiştir.

Çetin ve Amasyalı (2013), Türkçe Twitter verisi üzerinde eğiticili yöntemler ile çalışmalar yapılmıştır. Eğitici yöntemlerin daha başarılı olduğu tespit edilmiştir. Bir sonraki çalışmada Amasyalı (2013), Naïve Bayes (NB) ile eğitim kümesi azaltılarak deneyler yapmıştır. Çalışma sonunda %64 başarı oranı elde edilmiştir.

Vural vd. (2013), Türkçe film yorumları için sözlük tabanlı Duygu Analizi (DA) çalışması yapmışlardır. Çalışmalarında SentiStrength (Thelwall vd., 2010) kütüphanesini Türkçeye çevirerek, olumlu ve olumsuz sınıflandırma senaryosunu denemişler ve %76 başarı elde etmişlerdir.

Meral ve Diri (2014), yaptıkları çalışmada, Tweetleri etiketlemiş ve Makine öğrenmesi yöntemlerinden Destek Vektör Makinesi (SVM), Naïve Bayes (NB) ve Rastgele Orman (RF) sınıflandırıcıları kullanarak deneyler yapmışlardır. Deney sonuçlarında %90 civarı başarı elde etmişlerdir.

3. YAPILAN ÇALIŞMA

Duygu analizi (DA) ve metin sınıflandırma (MS) çalışmaları, makine öğrenmesi (MÖ) teknikleri kullanılarak gerçekleştirilmektedir. Tez'de veri kümesi olarak Twitter sosyal ağından toplanan veriler kullanılmıştır. Bu bölümde, tezde kullanılan makine öğrenmesi (MÖ) teknikleri ile Twitter'dan verilerin toplanması, toplanan verilerin saklanması, çalışma öncesi toplanan veriler üzerinde yapılan ön işlemler, veri kümelerinin oluşturulması ve sınıflandırma için yapılan çalışmalar anlatılmıştır.

3.1. Twitter

Twitter günümüzde kullanılan en popular sosyal ağlarından biridir. Statista[1] verilerine göre günlük 319 Milyon aktif kullanıcı tarafından atılan 500 Milyon üzeri tweet bu popülaritesinin en büyük kanıtıdır. Twitter'ın bu popülaritesine bağlı olarak, akademik çalışmalar için de önemli bir veri kaynağı haline gelmiştir. Twitter, bu veri kaynağına erişim için Twitter Uygulama Programlama Arayüzü'nü (Twitter API[2]) sunmaktadır.

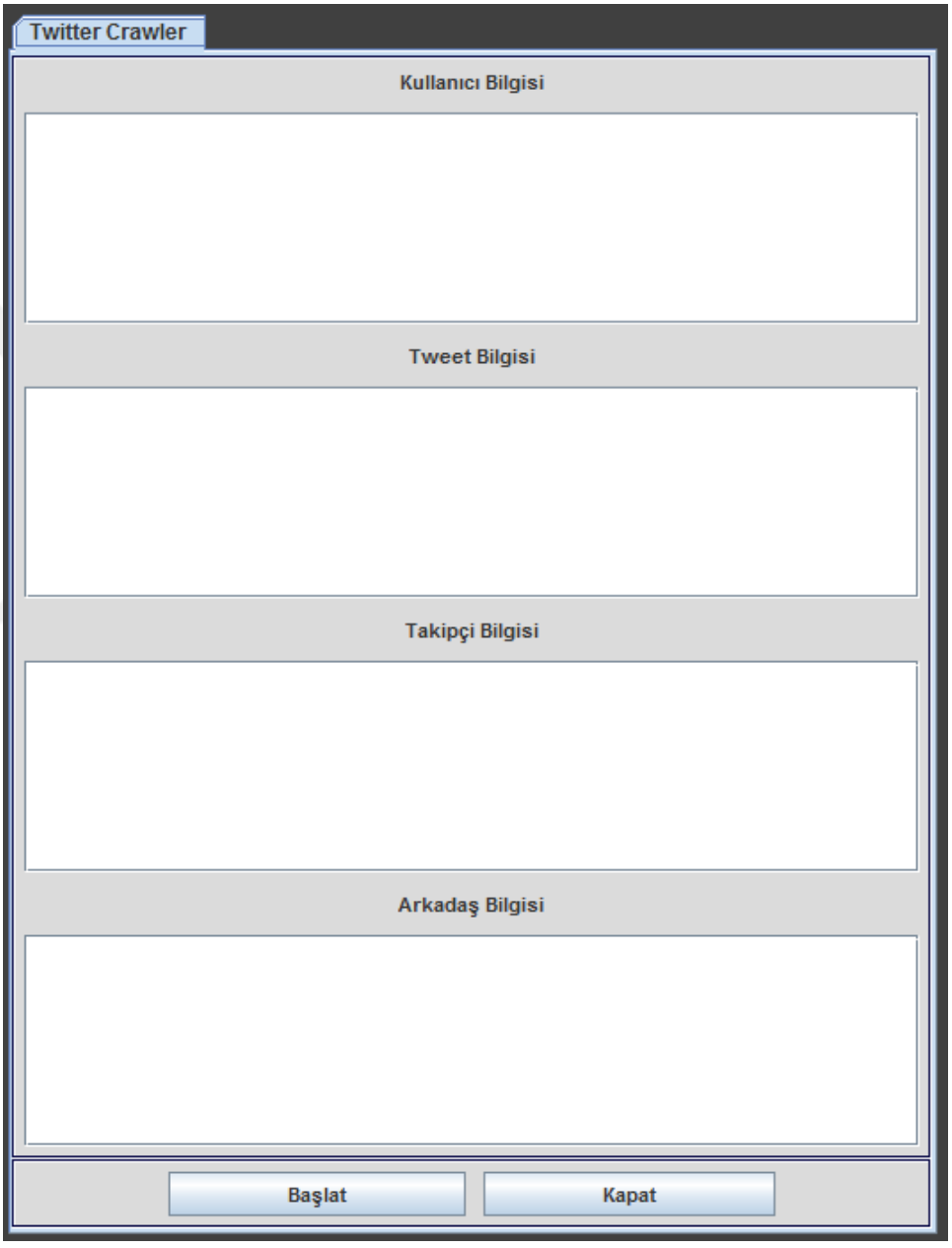
Twitter Uygulama Programlama Arayüzü (Twitter API), sahip olduğu yetenekler ile kullanıcılara sınırlı da olsa Twitter veri kaynağına erişim sağlamaktadır. Twitter Uygulama Programlama Arayüzü (Twitter API) sayesinde, bir kullanıcıya ait mesajlar, takipçileri, takip ettikleri ve kullanıcı detay bilgiler (konum, tarih, vb.) alınabilmektedir. Bu bilgiler dışında sorgu çalıştırma ve mesaj gönderme gibi özellikleri de kullanıcılara sunmaktadır. Twitter, bu yetenekleri sadece genel hesaplar için sağlamaktadır. Özel hesaplar için bu detaylara erişmek mümkün değildir.

Bu tez çalışması kapsamında üzerinde çalışılan veriler, Twitter Uygulama Programlama Arayüzü (Twitter API) kullanılarak Twitter sosyal ağından toplanmıştır.

[1] https://statista.com (15 Şubat 2017)
[2] https://dev.twitter.com

3.2. Twitter Verisi Toplama

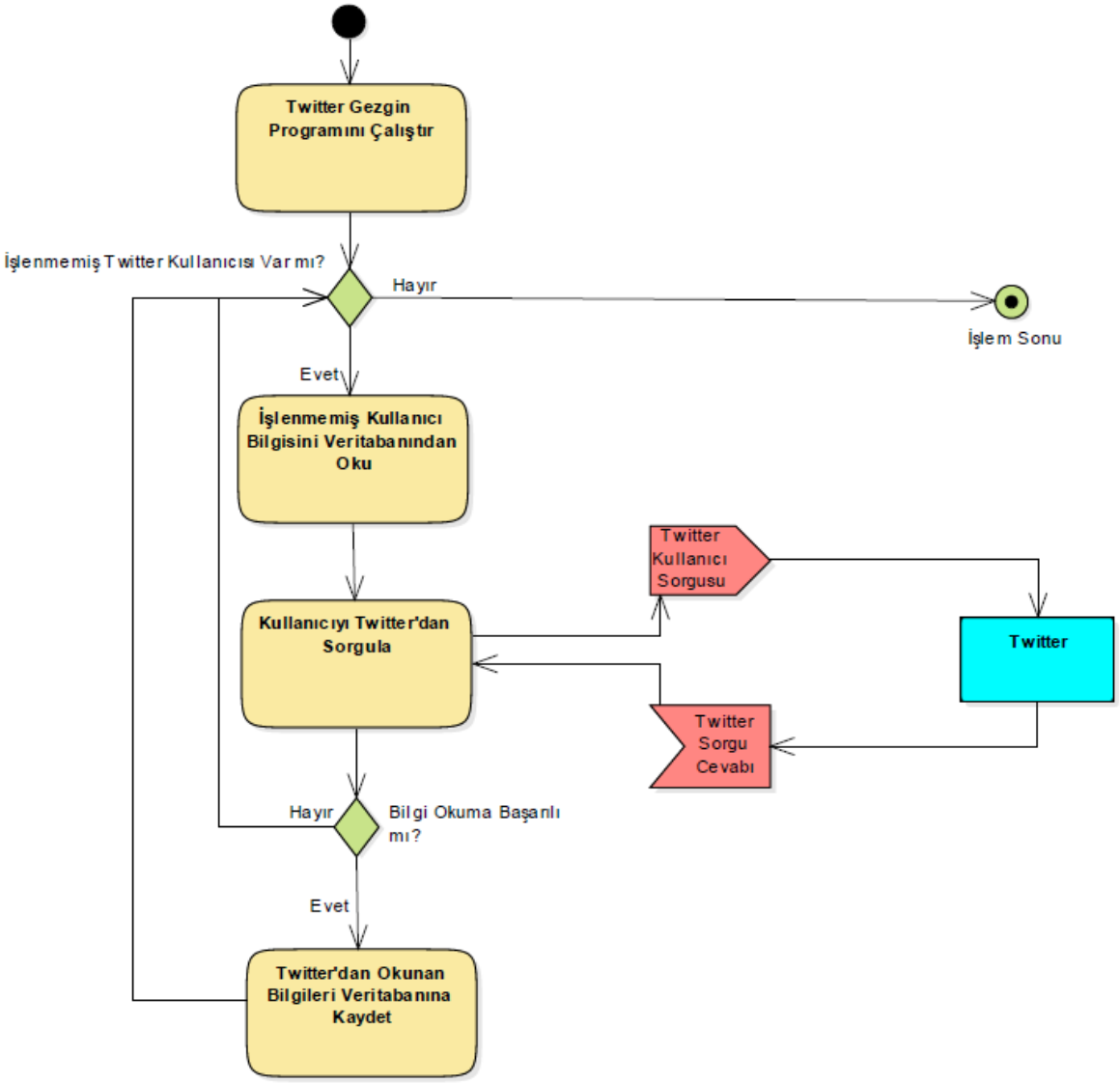
Tez kapsamında Java programlama dili, Twitter Uygulama Programlama Arayüzü (Twitter API) ve PostgreSQL veritabanı kullanılarak Twitter Gezgin (Twitter Crawler) yazılımı geliştirilmiştir (Şekil 3.1).



Şekil 3.1 Twitter gezgin programı

Twitter Gezgin programı ile Twitter soysal ağındaki kayıtlar üzerinde gezinmeye başlanmadan önce bir Twitter hesabı oluşturulmuş ve bu hesap ile dönemin en popüler

hesapları takip edilmeye başlanmıştır. Twitter Gezgin programına bu hesap başlangıç noktası olarak verilmiştir. Twitter Gezgin programı bu hesabın takip ettiği hesaplar üzerinde gezinmeye başlamıştır. Ele aldığı her hesap için, hesabın attığı tweetler, takip ettiği kullanıcılar ve bu hesabı takip eden kullanıcılar bilgisini alarak veritabanına kaydetmiştir. Twitter Gezgin yazılımı, bir hesap için işlemi bitirdikten sonra, veritabanına kaydedilen Twitter kullanıcıları içinden bir sonraki işlenmemiş kayıt alınarak aynı işlem tekrarlamıştır. Bu işleme ait akış Şekil 3.2'de gösterilmektedir.



Şekil 3.2 Twitter gezgin programı akış şeması

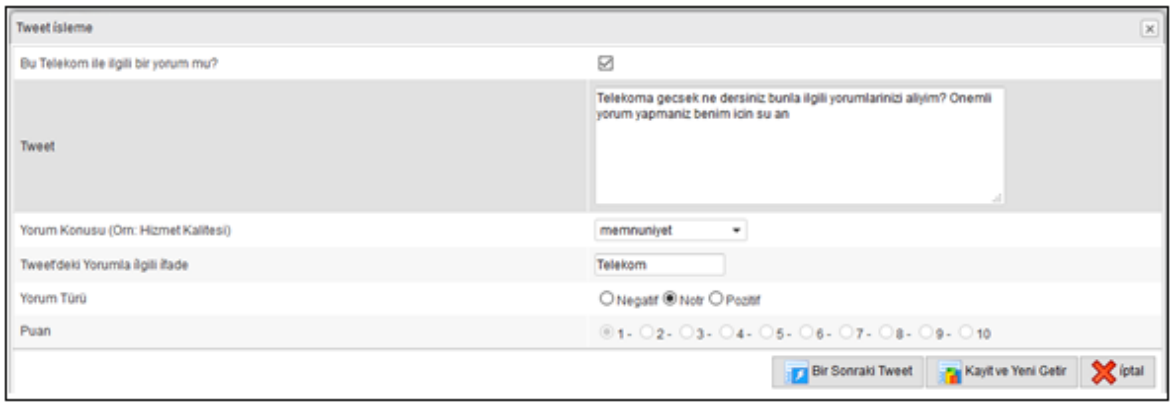
Twitter Gezgin programı Haziran 2012 ve Nisan 2013 tarihleri arasında çalıştırılmış ve bu gezinti sonucunda Twitter'dan alınan veriler PostgreSQL veritabanına kaydedilmiştir. Bu kaydedilen verilere ait detaylar Şekil 3.3'de yer almaktadır.

KULLANICI (USER)	TAKİPÇİ (FOLLOWER)	TAKİP EDİLEN (FOLLOW)	TWEET
• 6050975	• 11630935	• 7725908	• 15445646

Şekil 3.3 Toplanan veri detayları

3.3. Veri Etiketleme

Tez kapsamında, Java programlama dili, ZK Framework ve PostgreSQL veritabanı kullanılarak veri etiketleme yazılımı geliştirilmiştir. Doğuş Üniversitesi öğrencilerinden, bu uygulamayı kullanarak kendilerine belli bir konu hakkında gösterilen tweetleri pozitif, negatif ve nötr olarak sınıflandırmaları istenmiştir. Bu çalışma kapsamında öğrencilere, banka, üniversite, telekom ve mobil cihazlarla ilgili tweetler gösterilmiştir. Bu etiketleme programa ait ekran görünümü Şekil 3.4'de gösterilmektedir.



Şekil 3.4 Veri etiketleme programı

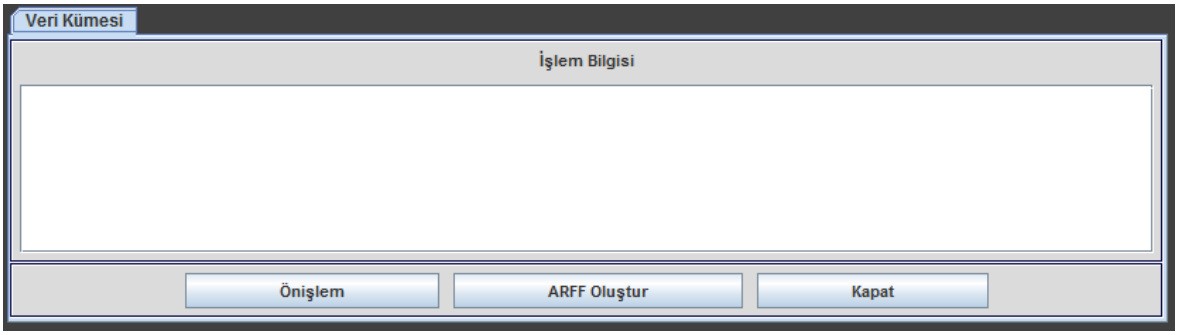
Bu etiketleme çalışması ile toplam 20204 tweet etiketlenmiştir. Etiketleme işleminin sonucunda elde edilen verilere ait detaylar Tablo 3.1'de detaylı olarak gösterilmektedir.

Tablo 3.1 Twitter verisi etiketleme çalışması detayları

Konu	Pozitif	Negatif	Nötr	Toplam
Bankacılık	1451	4603	1997	**8051**
Telekom	2226	2738	884	**5848**
Üniversiteler	1429	2230	1332	**4991**
Mobil Cihazlar	586	322	406	**1314**
Toplam	**5692**	**9893**	**4619**	**20204**

3.4. Veri Kümesi

Tez kapsamında yapılacak çalışmalarda veri kümesi oluşturmak için, Java programlama dili, Zemberek (Akın ve Akın, 2007) kütüphanesi ve PostgreSQL veritabanı kullanılarak, etiketli veriler üzerinde önişleme ve veri kümesi oluşturma işlemlerini yapan bir yazılım geliştirilmiştir. Bu yazılıma ait ekran görüntüsü Şekil 3.5'de gösterilmektedir.



Şekil 3.5 Veri işleme yazılımı

Bu yazılım ile etiketlenmiş telekom ve bankacılık ile ilgili tweetler kullanılarak Özellik–İlişki Dosya Formatı (ARFF) dosyaları oluşturulmuştur.

Denetimli öğrenme yöntemi için telekom verilerinden dengeli ve dengesiz veri kümeleri oluşturulmuştur. Telekom şirketlerine ait tweetler ile oluşturulan Özellik–İlişki Dosya Formatı (ARFF) dosyalarına ait veri detayları Tablo 3.2'de gösterilmektedir.

Tablo 3.2 Telekom dengeli - dengesiz veri kümesi detayları

Telekom	**Tip**			
	Pozitif	**Negatif**	**Nötr**	**Toplam**
Dengeli Veri Kümesi	504	504	504	**1512**
Dengesiz Veri Kümesi	1272	1140	504	**2916**

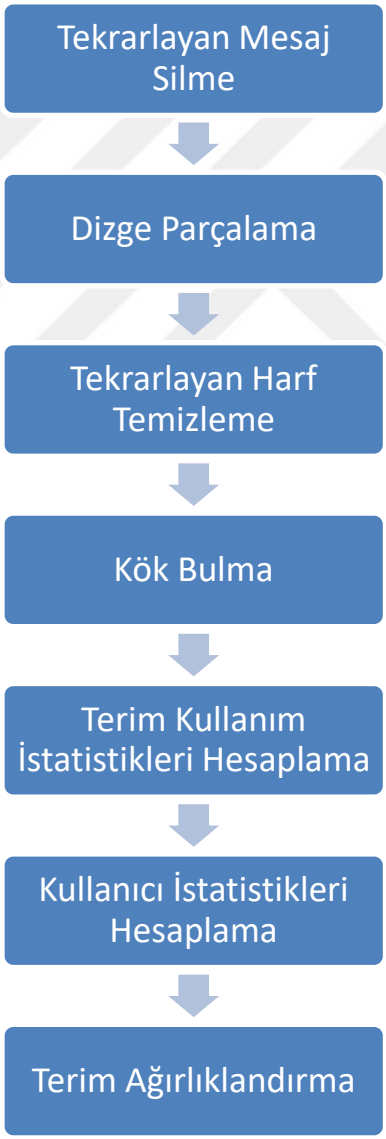
Yarı-denetimli öğrenme yöntemi için bankacılık verileri kullanılarak oluşturulan Özellik–İlişki Dosya Formatı (ARFF) dosyasına ait detay ise Tablo 3.3'de detaylı olarak gösterilmektedir.

Tablo 3.3 Bankacılık veri kümesi detayları

Bankacılık	**Tip**			
	Pozitif	**Negatif**	**Nötr**	**Toplam**
Veri Kümesi	3554	2948	1257	**7759**

3.4.1. Önişleme Adımları

Twitter mesajlarında, sınıflandırma işlemi için değerli veya bilgilendirici olmayan bazı alakasız terimler ve karakter dizeleri bulunabilmektedir. Sınıflandırma işleminin başarısını arttırabilmek için mesajlar üzerinde ön çalışma yapılması gerekmektedir. Bu çalışma yapılmadan önce, ilk olarak bir kişi tarafından paylaşılan aynı mesajlar ve bunlara ek olarak içeriğinde sadece Standart Kaynak Bulucu (URL), hashtag, özel karakter, sayı ve duygu ifadeleri içeren mesajlar silinmiştir. Gereksiz mesajların silinmesinden sonra deneylerde kullanılacak veriler Şekil 3.6'da gösterildiği sıra ile işlemlere tabi tutulmuştur.



Şekil 3.6 Önişlem adımları

3.4.1.1. Dizge Parçalama (Tokinize Strings)

Bu aşamada sınıflandırma işleminin başarısını arttırabilmek için anlamsız içerik temizlenmektedir. Bu bağlamda aşağıda listelenen karakterler mesajların içinden silinmiştir.

- "@" karakteri ile başlayan kullanıcı adları,
- "#" karakteri ile başlayan Hashtag'ler,
- Duygu ifadesi (emoji) simgeleri,
- Standart Kaynak Bulucular (URL),
- Rakamlar,
- Noktalama işaretleri,
- Özel karakterle,

Bu silme işlemi tamamlandıktan sonra tüm mesaj içerikleri küçük harfe dönüştürülmüştür.

3.4.1.2. Tekrarlayan Harfleri Temizleme

Sosyal medya kullanıcıları bazı duygularını daha güçlü ifade edebilmek için kelime içindeki harfleri tekrarlayarak yazmaktadırlar. Bu sosyal medya kullanıcıları arasında sık karşılaşılan bir durumdur. Özellik uzayını azaltmak için bu kelimeler üzerinde tekrarlayan harfler çıkartılarak bire düşürülmüştür. Bu işleme ait örnek Tablo 3.4'de gösterilmiştir.

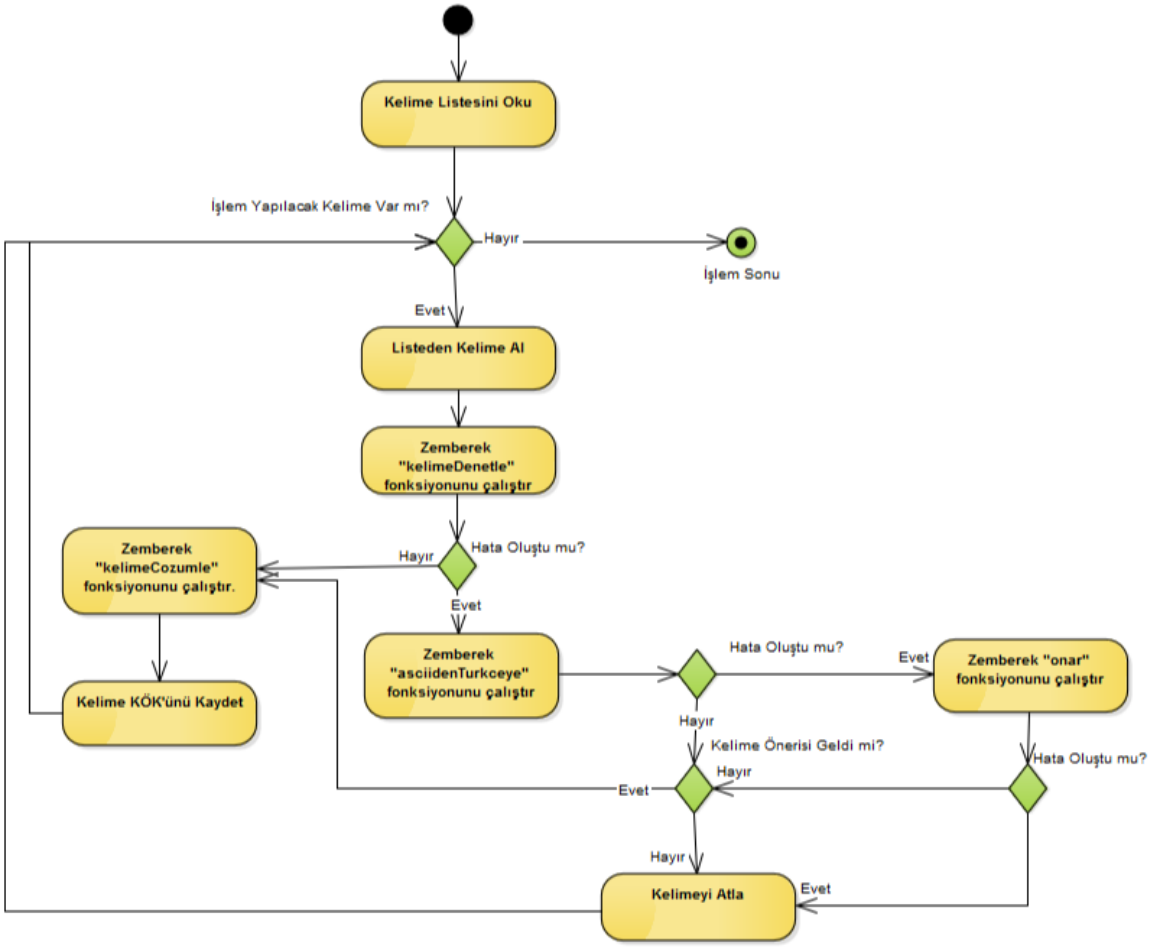
Tablo 3.4 Tekrarlayan harflerin temizlenmesi

Tekrarlayan Harf Bulunan Kelime	Düzeltilmiş Kelime
Seeeellllaaammm	Selam
Güüüünnnnaaayyyydddıınn	Günaydın

3.4.1.3. Kök Bulma (Stemming)

Veri kümesi içerisinden alınan mesajlar öncelikle kelimelere bölünmüş ve her kelime için yazım kontrolü yapılmıştır. Yazım kontrolü sonucunda yanlış veya eksik yazıldığı tespit edilen kelimeler düzeltilmeye çalışılmıştır. Düzeltme işlemi sonrası tüm kelimelerin kökleri bulunmuştur. Tüm bu işlemler için TÜBİTAK Bilgem tarafından desteklenen açık kaynak kodlu Türkçe doğal dil işleme kütüphanesi Zemberek (Akın ve Akın, 2007) kullanılmıştır. Zemberek kütüphanesi kullanılarak yapılan işlemler sırasıyla aşağıda listelenmiş ve Şekil 3.7'de detaylı olarak gösterilmiştir.

- Kelime, Zemberek Kütüphanesi içerisindeki "kelimeDenetle" fonksiyonu ile denetlenmiştir. Bu işlem sonucu hata bulunmaz ise kök bulma adımına geçilmiştir. Hata bulunur ise bir sonraki adım işletilmiştir.
- Yazım yanlışını düzeltebilmek için ilk olarak "asciidenTurkceye" fonksiyonu çalıştırılmıştır. Bu fonksiyon bir kelime önerisi döner ise kök bulma adımına geçilmiştir. Herhangi bir öneri yapamaz ise bir sonraki adım işletilmiştir.
- Yazım yanlışını düzeltebilmek için "onar" fonksiyonu çalıştırılmıştır. Bu fonksiyon bir kelime önerisi döner ise kök bulma adımına geçilmiştir. Herhangi bir öneri yapamaz ise o kelime işlenmemiştir.
- Yazım denetleme sonucu doğru olduğu tespit edilen kelimeler için "kelimeCozumle" fonksiyonu kullanılarak kök bulma işlemi yapılmıştır.



Şekil 3.7 Zemberek kütüphanesi ile kelime işleme adımları

3.4.1.4. Tweet Terim Kullanım İstatistiklerinin Hesaplanması

Bu tez kapsamında yapılacak çalışmanın başarısını arttırabilmek için terim kullanım oranları hesaplanmıştır. Her terim için pozitif, negatif, nötr ve toplam olmak üzere 4 çeşit terim kullanım oranı hesaplanmış ve hesaplanan bu değerler oluşturulan Özellik–İlişki Dosya Formatı (ARFF) dosyalarına özellik olarak eklenmiştir.

3.4.1.4.1. Terim Kullanım Oranı (WT)

Veri kümesindeki her terim için toplam kullanım oranı hesaplanmıştır. Bu değer hesaplanırken sırasıyla;

- Tüm veri kümesindeki toplam terim sayısı bulunmuştur.
- Her terimin tüm veri kümesindeki frekansı bulunmuştur.
- Her terim için toplam kullanım oranı hesaplanmıştır.

Özellik–İlişki Dosya Formatı (ARFF) dosyası oluşturulurken her mesajın içindeki terimlerin toplam kullanım oranı değeri toplanarak o mesaj için özellik olarak eklenmiştir. Bu hesaplama veri kümesindeki her mesaj için yapılmıştır (3.1).

$$WT(d_i) = \sum_{j=0}^{|d_i|} \log_{10} \frac{|DT|}{f_D(d_{i,j})} \quad (3.1)$$

$|DT|$ Tüm veri kümesindeki terim sayısı
d_i Tüm veri kümesindeki i. mesaj
$|d_i|$ i. mesajdaki terim sayısı
$d_{i,j}$ i. mesajdaki j.terim
$f_D(t)$ t teriminin tüm veri kümesindeki toplam sayısı

3.4.1.4.2. Terim Pozitif Kullanım Oranı (WP)

Pozitif olarak etiketlenmiş veri kümesindeki her terim için pozitif kullanım oranı hesaplanmıştır. Bu değer hesaplanırken sırasıyla;

- Pozitif veri kümesindeki toplam terim sayısı bulunmuştur.
- Her terimin pozitif veri kümesindeki frekansı bulunmuştur.
- Her terim için pozitif kullanım oranı hesaplanmıştır.

Özellik–İlişki Dosya Formatı (ARFF) dosyası oluşturulurken her mesajın içindeki terimlerin pozitif kullanım oranı değeri toplanarak o mesaj için özellik olarak eklenmiştir. Bu hesaplama veri kümesindeki her mesaj için yapılmıştır (3.2).

$$WP(d_i) = \sum_{j=0}^{|d_i|} \log_{10} \frac{|DP|}{f_{DP}(d_{i,j})} \quad (3.2)$$

$|DP|$ Pozitif veri kümesindeki terim sayısı

d_i Tüm veri kümesindeki i. mesaj

$|d_i|$ i. mesajdaki terim sayısı

$d_{i,j}$ i. mesajdaki j.terim

$f_{DP}(t)$ t teriminin pozitif veri kümesindeki toplam sayısı

3.4.1.4.3. Terim Negatif Kullanım Oranı (WN)

Negatif olarak etiketlenmiş veri kümesindeki her terim için negatif kullanım oranı hesaplanmıştır. Bu değer hesaplanırken sırasıyla;

- Negatif veri kümesindeki toplam terim sayısı bulunmuştur.
- Her terimin negatif veri kümesindeki frekansı bulunmuştur.
- Her terim için negatif kullanım oranı hesaplanmıştır.

Özellik–İlişki Dosya Formatı (ARFF) dosyası oluşturulurken her mesajın içindeki terimlerin negatif kullanım oranı değerleri toplanarak o mesaj için özellik olarak eklenmiştir. Bu hesaplama veri kümesindeki her mesaj için yapılmıştır (3.3).

$$WN(d_i) = \sum_{j=0}^{|d_i|} \log_{10} \frac{|DN|}{f_{DN}(d_{i,j})} \quad (3.3)$$

$|DN|$ Negatif veri kümesindeki terim sayısı

d_i Tüm veri kümesindeki i. mesaj

$|d_i|$ i. mesajdaki terim sayısı

$d_{i,j}$ i. mesajdaki j.terim

$f_{DN}(t)$ t teriminin negatif veri kümesindeki toplam sayısı

3.4.1.4.4. Terim Nötr Kullanım Oranı (WR)

Nötr olarak etiketlenmiş veri kümesindeki her terim için nötr kullanım oranı hesaplanmıştır. Bu değer hesaplanırken sırasıyla;

- Nötr veri kümesindeki toplam terim sayısı bulunmuştur.

- Her terimin nötr veri kümesindeki frekansı bulunmuştur.
- Her terim için nötr kullanım oranı hesaplanmıştır.

Özellik–İlişki Dosya Formatı (ARFF) dosyası oluşturulurken her mesajın içindeki terimlerin nötr kullanım oranı değerleri toplanarak o mesaj için özellik olarak eklenmiştir. Bu hesaplama veri kümesindeki her mesaj için yapılmıştır (3.4).

$$WR(d_i) = \sum_{j=0}^{|d_i|} \log_{10} \frac{|DR|}{f_{DR}(d_{i,j})} \quad (3.4)$$

$|DR|$ Nötr veri kümesindeki terim sayısı

d_i Tüm veri kümesindeki i. mesaj

$|d_i|$ i. mesajdaki terim sayısı

$d_{i,j}$ i. mesajdaki j.terim

$f_{DR}(t)$ t teriminin nötr veri kümesindeki toplam sayısı

3.4.1.4.5. Kullanıcı İstatistiklerini Hesaplanması (Tweet Sayıları)

Bu tez kapsamında yapılacak çalışmanın başarısını arttırabilmek için terim kullanım oranlarına ek olarak kullanıcı istatistikleri hesaplanarak oluşturulan Özellik–İlişki Dosya Formatı (ARFF) dosyalarına özellik olarak eklenmiştir. Deneylerde kullanılmak üzere her kullanıcı için pozitif, negatif, nötr ve toplam tweet sayılarını içeren kullanıcı istatistikleri hesaplanmıştır. Bu istatistikler sırasıyla aşağıda listelenmiştir.

- $U_t(i)$ Bir kullanıcı tarafından atılan toplam tweet sayısı
- $U_p(i)$ Bir kullanıcı tarafından atılan toplam pozitif tweet sayısı
- $U_n(i)$ Bir kullanıcı tarafından atılan toplam nötr tweet sayısı
- $U_r(i)$ Bir kullanıcı tarafından atılan toplam negatif tweet sayısı

3.4.1.5. Terim Ağırlıklandırma

Burada önemli olan nokta özniteliklerin belirlenmesi sürecidir. Terim Ağırlıklandırma ile her bir terim; ilgili terimin önemini ölçen ve gözlendiği dokümanın sınıflandırılmasına

yaptığı katkıyı belirten bir ağırlık değeri ile ilişkilendirilir. Bu bağlamda ağırlıklandırma yöntemleri olarak İkili (Binary/Boolean) (Liao vd., 2001), Terim Frekansı (TF) ve Terim Frekansı-Ters Doküman Frekansı (TF-IDF) (Liu ve Yang, 2012) kullanılmıştır.

3.4.1.5.1. İkili Ağırlıklandırma (Binary / Boolean)

İkili Ağırlıklandırma yönteminde bir terimin doküman ya da mesajda geçip geçmediğine bakılır. Eğer terim dokümanda ya da mesajda bulunursa ağırlık değeri 1, bulunmazsa ağırlık değeri 0 olur. (3.5)

$$W_{boolean}(c,d) = \begin{cases} 1, & TF \geq 1 \\ 0, & Diğer \end{cases} \tag{3.5}$$

3.4.1.5.2. Terim Frekansı (TF) (Term Frequency)

Terim Frekansı (TF) ağırlıklandırma yönteminde, bir terimin bir doküman içinde toplam kaç kere görüldüğü bulunur. Terimin ağırlığı, bir dokümanda kaç kere kullanıldığı ile orantılı olarak atanmaktadır. Böylece dokümanda daha fazla geçen kelimeler varsa onlar daha değerli olmaktadır. (3.6)

$$W_{TF}(c,d) = TF(c,d) \tag{3.6}$$

3.4.1.5.3. Terim Frekansı-Ters Doküman Frekansı (TF-IDF) (Term Frequency-Inverse Document Frequency)

Ters Doküman Frekansı (IDF) yöntemi, terimlerin ağırlıklarını bulunduğu doküman kapsamında değil tüm veri setinde bulunan dokümanlar kapsamında hesaplar. Bu sayede veri setinde çok kullanılan terimlerin ağırlığı düşerken, nadir kullanılan terimlerin ağırlığı yükselir.

$$W_{DF} = \sum_{i=1}^{n} TF(d_i, c) \tag{3.7}$$

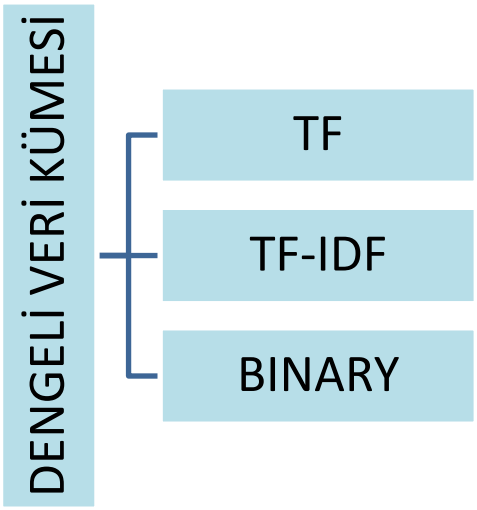
$$W_{IDF}(c,d) = log\frac{N}{DF(c,d)} \tag{3.8}$$

Terim Frekansı-Ters Doküman Frekansı (TF-IDF) ağırlıklandırma sadece belge içindeki terim sıklığını hesaba katmaz bununla birlikte tüm belgelerde terimin sıklığını göz önünde bulundurur. Bu yöntem de, Ters Doküman Frekansında (IDF) olduğu gibi çok kullanılan terimlerin ağırlığını düşürmeyi amaçlar ama bu değeri hesaplarken terimin frekansını da kullanır. (3.9)

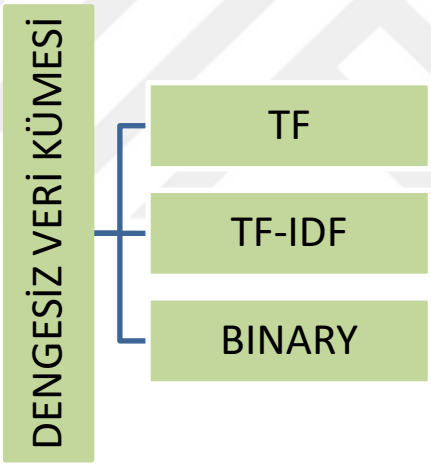
$$W_{TF-IDF}(c,d) = TF(c,d) * IDF(c,d) \quad (3.9)$$

3.4.2. Veri Kümesi Oluşturma

Tüm bu önişleme çalışmasından sonra deneylerde kullanmak üzere telekom ve bankacılık alanında atılan Twitter mesajları kullanılarak Özellik–İlişki Dosya Formatı (ARFF) dosyaları oluşturulmuştur. Bu Özellik–İlişki Dosya Formatı (ARFF) dosyalarının içeriği, veri boyutu ve terim ağırlıklandırma özelliklerine göre değişiklik göstermektedir. Denetimli öğrenme yöntemi ile yapılacak deneylerde kullanılmak üzere telekom ile ilgili 3 dengeli ve 3 dengesiz veri kümesi içeren toplam 6 farklı Özellik–İlişki Dosya Formatı (ARFF) dosyası oluşturulmuştur. Yarı-denetimli öğrenme yöntemi ile yapılacak deneyler için ise bankacılık ile ilgili mesajları içeren 3 Özellik–İlişki Dosya Formatı (ARFF) dosyası oluşturulmuştur. Bu dosyalar İkili (Binary/Boolean), Terim Frekansı (TF) ve Terim Frekansı-Ters Doküman Frekansı (TF-IDF) terim ağırlıklandırma yöntemleri kullanılarak hazırlanmıştır. Şekil 3.8'de Dengeli Veri Kümesi, Şekil 3.9'da Dengesiz Veri Kümesi için oluşturulan Özellik–İlişki Dosya Formatı (ARFF) dosyalarının ağırlıklandırma yöntemleri detaylı olarak gösterilmiştir.



Şekil 3.8 Dengeli veri kümesi



Şekil 3.9 Dengesiz veri kümesi

Bir veri kümesi içinde bulunan pozitif, negatif ve nötr olarak etiketlenmiş kayıt sayıları farklılık gösteriyor ise bu veri kümesine dengesiz veri kümesi denilmektedir. Eğer bir veri kümesi içindeki pozitif, negatif ve nötr olarak etiketlenmiş kayıtların sayısı eşit olursa bu veri kümesine dengeli veri kümesi denilmektedir. Bu tez kapsamında denetimli öğrenme yöntemi ile yapılacak deneyler için telekom şirketleri hakkında atılan mesajlar kullanılarak dengeli ve dengesiz veri kümeleri oluşturulmuştur. Bu veri kümelerine ait detaylar Tablo 3.2'de gösterilmiştir.

Yarı-Denetimli öğrenme yöntemi ile yapılacak deneylerinde kullanılmak üzere oluşturulan bankacılık veri kümesinin detayları Tablo 3.3'de detaylı olarak gösterilmiştir.

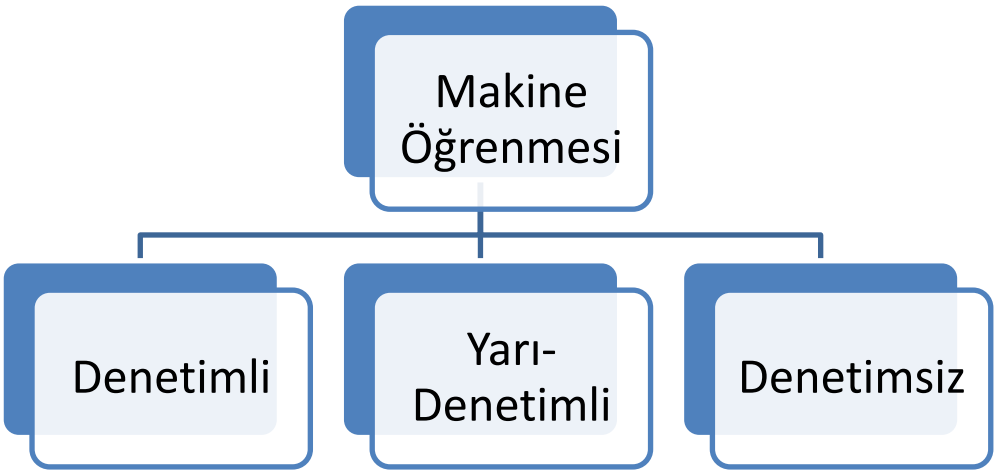
Hesaplanan Terim Kullanım Oranı ve Kullanıcı İstatistiklerinin deneylerde kullanılmak üzere oluşturulan Özellik–İlişki Dosya Formatı (ARFF) dosyalarına eklenmesi sonucu oluşan dosya yapısı Tablo 3.5'de gösterilmiştir.

Tablo 3.5 Özellik–İlişki dosya yapısı

$d_{1,1}$	$d_{1,2}$	$d_{1,3}$	….	….	$d_{1,j}$	WT_1	WP_1	WN_1	WR_1	$\mathrm{U_t}(1)$	$\mathrm{U_p}(1)$	$\mathrm{U_t}(1)$	$\mathrm{U_n}(1)$	$\mathrm{U_r}(1)$
….	….	….	….	….	….	….	….	….	….	….	….	….	….	….
$d_{i,1}$	$d_{i,2}$	$d_{i,3}$	….	….	$d_{i,j}$	WT_j	WP_j	WN_j	WR_j	$\mathrm{U_t}(i)$	$\mathrm{U_p}(i)$	$\mathrm{U_t}(i)$	$\mathrm{U_n}(i)$	$\mathrm{U_r}(i)$

3.5. Makine Öğrenmesi

Makine öğrenmesi (MÖ), önceki gözlemleri kullanarak doğru tahminlerin yapılmasını sağlamak amacıyla geliştirilen tekniklerdir. Makine öğrenmesi (MÖ) yönteminde, etiketlenmiş veri kümesi ile makine öğrenmesi (MÖ) yöntemi kullanılarak bir model oluşturulur ve bu model kullanılarak etiketsiz verilerin sınıflandırılması yapılır. Makine öğrenmesi (MÖ) Şekil 3.10'da gösterildiği üzere 3 bölümde ele alınmaktadır. Bu makine öğrenmesi (MÖ) yöntemleri; denetimli (supervised), denetimsiz (unsupervised) ve yarı-denetimli (semi-supervised) olarak adlandırılmaktadır. Ayrıca denetimli öğrenmeye, öğreticili ya da gözetimli öğrenme, denetimsiz öğrenmeye ise öğreticisiz veya gözetimsiz öğrenme de denilebilmektedir. Yarı-denetimli öğrenme yerine ise, yarı-öğreticili ya da yarı-gözetimli öğrenme kullanılabilmektedir.



Şekil 3.10 Makine öğrenmesi yöntemleri

3.5.1. Denetimli Öğrenme (Supervised)

Denetimli Makine Öğrenmesi yönteminde etiketli veriler kullanılarak sistem eğitilmesi ve öğrenmenin sağlanması amaçlanır. Eğitim yapılan veri kümesinin çıktısının bilindiği durumda kullanılır. Sistem eğitim kümelerinin çıktıları doğrultusunda eğitilir ve tahmin yapılır. Bu öğrenme yönteminde başarılı sonuç elde edebilmek için etiketli veri seti yeterince büyük olmalıdır. Bu öğrenme yönteminde problem, sınıflandırma problem olarak ele alınır ve eğitilmiş sistem test kümesinin tahmin edilmesinde kullanılır. Naïve Bayes (NB), Destek Vektör Makinesi (SVM), k-En Yakın Komşu, Rastgele Orman (RF) ve Karar Ağacı (J48) algoritmaları denetimli öğrenme yönteminde yaygın olarak kullanılır.

3.5.2. Denetimsiz Öğrenme (Unsupervised)

Sistem eğitilirken etiketsiz veri kullanılarak öğrenme çalışması yapılır. Eğitim yapılan veri kümesinin çıktısının bilinmediği durumda kullanılır. Denetimsiz öğrenmede amaç sınıflandırma değildir. Denetimli ve yarı-denetimli öğrenme yöntemlerinde karşılaşılan alan bağımlılığı problemini ortadan kaldırmak amaçlanmaktadır. Denetimsiz öğrenme, kümeleme, olasılık yoğunluk tahmini, öznitelik ilişki bulunması ve boyut indirgeme için kullanılır. Parçalayıcı ve hiyerarşik kümeleme algoritmaları denetimsiz öğrenmede kullanılan algoritmalardır.

3.5.3. Yarı-Denetimli Öğrenme (Semi-Supervised)

Denetimli ve denetimsiz öğrenme yöntemlerinin birleştirilmesi sonucu ortaya çıkmıştır. Bu öğrenme yönteminde etiketli ve etiketsiz veriler beraber kullanılmaktadır. Bu öğrenme yönteminde etiketli verilerin sayısı etiketsizlere göre daha azdır. Oluşturulan model ile etiketsiz verilerin etiketleri tahmin edilmektedir.

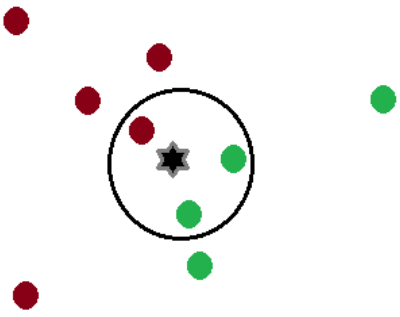
3.6. Sınıflandırma

Sınıflandırma, mevcut örneklerin birbirlerinden yüksek doğruluk derecesi ile ayırt edilmeye çalışılmasıdır. Tez kapsamında, Naïve Bayes (NB), k-En Yakın Komşu (k-NN: k-Nearest Neighbor), Destek Vektör Makineleri (SVM: Support Vector Machines) ve Maksimum Entropi (ME: Maximum Entropy) sınıflandırıcıları kullanılmıştır.

3.6.1. K – En yakın komşu (K-NN) (IB1)

Basitliği ve doğruluğu sayesinde en çok kullanılan sınıflandırma algoritmalarından biridir. Sınıflandırılmak istenen bir test örneği için en yakın örneği, Öklid mesafesi kullanarak bulmaya çalışır (3.10). Sınıflandırma öncesi bir kural veya fonksiyonlar kümesi oluşturmaz. Bu sayede diğer yöntemlere göre daha etkin olmakla beraber sınıflandırma aşamasında her örnek için yeniden hesaplama yapıldığından daha yavaş çalışmaktadır.

$$D(a,b) = \sqrt{\sum_{i=1}^{n}(b_i - a_i)^2} \tag{3.10}$$



Şekil 3.11 K-En yakın komşu yöntemi, k=3

3.6.2. Naïve Bayes (NB) Sınıflandırma Algoritması

Naïve Bayes (NB) sınıflandırıcısı, olasılık tekniklerini kullanır. Sınıflandırılması gereken verilerin sınıfları bellidir. Öznitelik seçme işlemi kolay, verimli ve hassastır. Çok boyutlu verilerde genelde iyi sonuç vermez. Metin sınıflandırma (MS) işlemi için çok başarılı sonuç verdiği kanıtlanmıştır (Dai vd., 2007). $c_1 \ldots . c_n$ olmak üzere; n tane sınıf ve test örneği x=($x_1 \ldots . x_n$) olduğunda sınıflandırma (3.11) ile gerçekleştirilir.

$$P(c_k|x) = \frac{p(x|c_k)p(c_k)}{p(x)} \tag{3.11}$$

Bayes sınıflandırıcı sadece $p(x|c_k)p(c_k)$ ilgilenir. p(x) sınıf ve özniteliklere bağlı değildir. Bu nedenle sabit bir değer olarak kabul edilir ve kullanılmaz. c_k Kategorisinde bulunan doküman sayısının tüm doküman sayısına oranı $p(c_k)$ 'yı temsil eder (3.12). Sonuçta özniteliklerin bağımsız olduğu varsayımı kullanılarak sınıflandırma (3.13) ile ifade edilir.

$$P(c_k) = \frac{c_k}{|c|} \tag{3.12}$$

$$P(c_k|x_1 \ldots x_n) = p(c_k) \prod_{i=1}^{n} p(x_i|c_k) \tag{3.13}$$

3.6.3. Rastgele Orman (RF) (Random Forest) Sınıflandırma Algoritması

Breiman tarafından geliştirilmiştir. Tek bir karar ağacı (J48) üretmek yerine her biri farklı eğitim kümelerinde eğitilmiş olan çok sayıda çok değişkenli ağacın kararlarını birleştirmeyi amaçlamaktadır (Breiman, 2001). Bir sınıflandırıcı yerine birden çok sınıflandırıcı üreten ve sonrasında onların tahminlerinden alınan oylar ile yeni veriyi sınıflandıran öğrenme algoritmasıdır. Büyük veri tabanlarında eşsiz olarak çalışır ve dengesiz veri seti sınıfında hata dengeleme yöntemlerine sahiptir.

3.6.4. SMO (Sequential Minimal Optimization) Sınıflandırma Algoritması

SMO, ekstra matris depolama ihtiyacı yoktur ve büyük ölçekli QP (Quadratic Programming) problemini çözmek için üretilmiştir (Platt, 1998). Öğrenme işleminde elimizde bulunan m

veri için m boyutlu bir uzayda çözüm gerekmektedir. Bu da büyük veri kümelerinde problem oluşturmaktadır. Bu algoritma global olarak bütün kayıp değerleri yenisiyle değiştirir ve nominal öznitelikleri ise ikili olanlara dönüştürür. Bütün özellikleri tanımlanmış değerlerle normalize eder.

3.6.5. Karar Ağacı (J48) (Desicion Tree) Sınıflandırma Algoritması

Temelinde C4.5 algoritmasına dayanan, J. Ross Quinlan tarafından geliştirilmiş bir karar ağacı algoritmasıdır. Karar ağaçları bir makine öğrenmesi algoritmasından bilgi temsil etmede klasik bir yoldur ve veri yapılarını ifade etmekte güçlü ve hızlı bir yol sunar. Öznitelikler bu algoritmada düğüm noktası oluşturacak şekilde yerleştirilir ve eğitim verisine göre yapraklar meydana gelir. Yapraklar aynı zamanda sınıf etiketlerini belirtir (Kargupta vd., 2008).

3.7. Sınıflandırma Algoritmalarının Karşılaştırılmasında Kullanılan Kriterler

Duygu analizi (DA) kapsamında yapılan deney sonuçlarının doğruluğunun ölçülmesi gerekmektedir. Etiketli veriler ile yapılan deneylerde duygu analizi (DA) sonucunda yorumların polaritesi hesaplanır. Eğer pozitif etiketli bir veri pozitif olarak bulunur ise bu Doğru Pozitif (TP – True Positive), pozitif etiketli bir veri negatif bulunur ise bu Yanlış Pozitif (FP – False Positive) olarak gösterilir. Diğer taraftan negatif etiketli bir veri negatif bulunur ise Doğru Negatif (TN – True Negative), negatif etiketli bir veri pozitif olarak bulunur ise Yanlış Negatif (FN – False Negative) olarak gösterilir. Model başarım ölçütleri sırası ile

- Doğruluk-Hata Oranı (Accuracy Error Rate)
- Kesinlik (Precision)
- Duyarlılık (Recall)
- F-Ölçütü (F-Measure)

3.7.1. Doğruluk–Hata Oranı (Accuracy-Error Rate)

Sınıflandırma çalışmasının başarısını ölçmek için kullanılan basit ve popüler bir yöntemdir. Doğruluk ve hata oranı olarak ikiye ayrılır. Doğruluk, başarılı sınıflandırılan verilerin toplam sayısının tüm veri sayısına oranıdır (3.14). Hata oranı ise doğruluk değerinin 1'e tamlayanıdır (3.15).

$$Dogruluk = \frac{TP+TN}{TP+TN+FP+FN} \tag{3.14}$$

$$Hata = 1 - Dogruluk \tag{3.15}$$

3.7.2. Kesinlik (Precision)

Sınıflandırmanın ne ölçüde başarılı tahmin edildiğini bulmak için kullanılır. Doğru sınıflandırılan Doğru Pozitif (TP) mesajların sayısının, Doğru Pozitif (TP) ve Yanlış Pozitif (FP) toplamına oranı bulunarak hesaplanır (3.16).

$$\pi = \frac{TP}{TP+FP} \tag{3.16}$$

3.7.3. Duyarlılık (Recall)

Kategoriye ait elemanların ne kadarının tahmin edildiğini hesaplar. Doğru tahmin edilen mesaj sayısının kategorideki toplam veri sayısına oranı ile bulunur. Doğru Pozitif (TP) mesajların sayısının, Doğru Pozitif (TP) ve Yanlış Negatif (FN) toplamına oranı bulunarak hesaplanır (3.17)

$$\rho = \frac{TP}{TP+FN} \tag{3.17}$$

3.7.4. F-Ölçütü (F-Measure)

Kesinlik (Presicion) ve Duyarlılık (Recall) ölçütleri tek başına anlam ifade etmemektedirler. Bu değerler bir karşılaştırma sonucu çıkarmamız için yeterli değildir. Bu nedenle Kesinlik

ve Duyarlılık değerlerini beraber değerlendirmek daha doğru sonuç verir. F-ölçütü, Kesinlik ve Duyarlılığın harmonik ortalamasıdır (3.18).

$$F = 2\ X\ \frac{\pi\ X\ P}{\pi + P} = \frac{2TP}{2TP+FP+FN} \tag{3.18}$$

4. SONUÇ VE ÖNERİLER

4.1. Deneysel Çalışmalar

Bu tez kapsamında geliştirilen yazılım ile Twitter sosyal ağından Türkçe mesajlar toplanmıştır. Toplanan bu Türkçe Twitter mesajları Doğuş Üniversitesi öğrencileri tarafından pozitif, nötr ve negatif olarak etiketlenmiştir. Etiketleme işlemi sonucunda bu mesajlar metin sınıflandırma (MS) ve doğal dil işleme (DDİ) teknikleri kullanılarak önişlem adımlarından geçirilmiştir. Önişleme adımlarından sonra veri kümesindeki her kelime için kelime kullanım oranları ve her kullanıcı için toplam, pozitif, nötr ve negatif kullanıcı istatistikleri hesaplanmış.

Makine öğrenmesi (MÖ) yöntemlerinden denetimli öğrenme yöntemi için telekom firmaları hakkında atılan tweetler alınarak deneylerde kullanılmak üzere dengeli ve dengesiz Özellik–İlişki Dosya Formatı (ARFF) dosyaları oluşturulmuştur. Bu Özellik–İlişki Dosya Formatı (ARFF) dosyalarına;

- WT Kelime toplam kullanım Oranı,
- WP Kelime pozitif kullanım oranı,
- WN Kelime negatif kullanım oranı,
- WR Kelime nötr kullanım oranı,
- $U_t(i)$ Bir kullanıcı tarafından atılan toplam tweet sayısı,
- $U_p(i)$ Bir kullanıcı tarafından atılan toplam pozitif tweet sayısı,
- $U_n(i)$ Bir kullanıcı tarafından atılan toplam nötr tweet sayısı,
- $U_r(i)$ Bir kullanıcı tarafından atılan toplam negatif tweet sayısı,

değerleri özellik olarak eklenmiştir. Telekom şirketleri hakkındaki veri kümesine ait detaylı bilgi Tablo 3.2'de verilmiştir.

Makine öğrenmesi (MÖ) yöntemlerinden yarı-denetimli öğrenme yöntemi için bankacılık ile ilgili atılan tweetler alınarak deneylerde kullanılmak üzere Özellik–İlişki Dosya Formatı (ARFF) dosyaları oluşturulmuştur. Bu Özellik–İlişki Dosya Formatı (ARFF) dosyalarına

kelime kullanım oranları WT, WP, WN ve WR ile kullanıcı istatistik değerleri U_t, U_p,$U_n$ve U_r özellik olarak eklenmiştir. Bankacılık hakkındaki veri kümesine ait detaylı bilgi Tablo 3.3'de verilmiştir.

Denetimli ve yarı-denetimli öğrenme yöntemleri ile yapılacak deneylerde kullanılmak üzere hazırlanan Özellik–İlişki Dosya Formatı (ARFF) dosyalarında İkili, Terim Frekansı (TF) ve Terim Frekansı-Ters Doküman Frekansı (TF-IDF) ağırlıklandırma yöntemleri kullanılmıştır. Tüm bu işlemlerin sonunda Weka (Witten vd., 2016) yazılımı kullanarak denetimli ve yarı-denetimli öğrenme yöntemleri ile deneyler yapılmıştır.

4.1.1 Denetimli Makine Öğrenmesi

Makine öğrenmesi (MÖ)yöntemlerinden denetimli öğrenme ile yapılacak deneyler için dengeli ve dengesiz Özellik–İlişki Dosya Formatı (ARFF) dosyaları İkili, Terim Frekansı (TF) ve Terim Frekansı-Ters Doküman Frekansı (TF-IDF) ağırlıklandırma yöntemleri kullanılarak hazırlanmıştır. Hazırlanan bu dosyalar ile 10-Kat Çapraz Doğrulama ve %80 bölümlendirme ile deneyler yapılmıştır. Yapılan bu deneyler sonucunda %80,8201 ile en iyi sonuç, dengeli veri kümesi ile 10-Kat Çapraz Doğrulama yöntemi yapılan deneyler sonucunda elde edilmiştir.

4.1.1.1. İkili Veri Kümesi (Binary/Boolean)

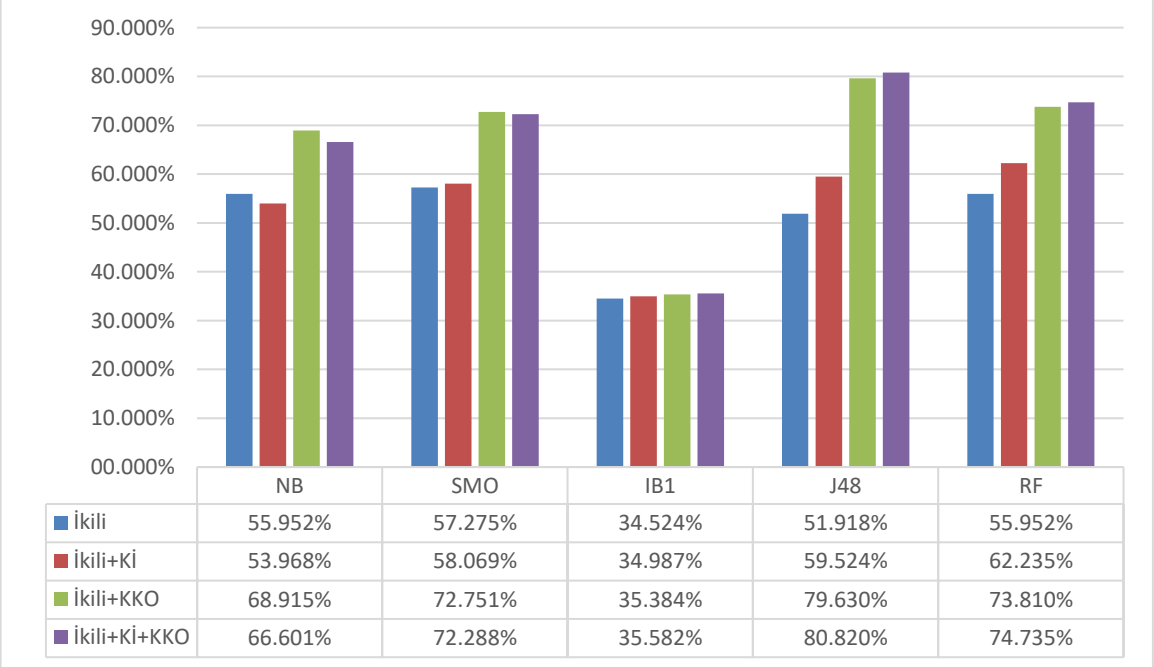
Türkçe Twitter verileri kullanılarak İkili(Binary/Boolean) ağırlıklandırma yöntemi ile oluşturulan Dengeli ve Dengesiz veri kümeleri ile Weka yazılımı kullanılarak 10-kat çapraz doğrulama ve %80 bölümlendirme ile deneyler yapılmıştır. Bu deneyler sırası ile

- İkili,
- İkili+Kİ (İkili + Kullanıcı İstatistiği),
- İkili+KKO (İkili + Kelime Kullanım Oranı),
- İkili+Kİ+KKO (İkili + Kullanıcı İstatistiği + Kelime Kullanım Oranı),

için yapılmıştır.

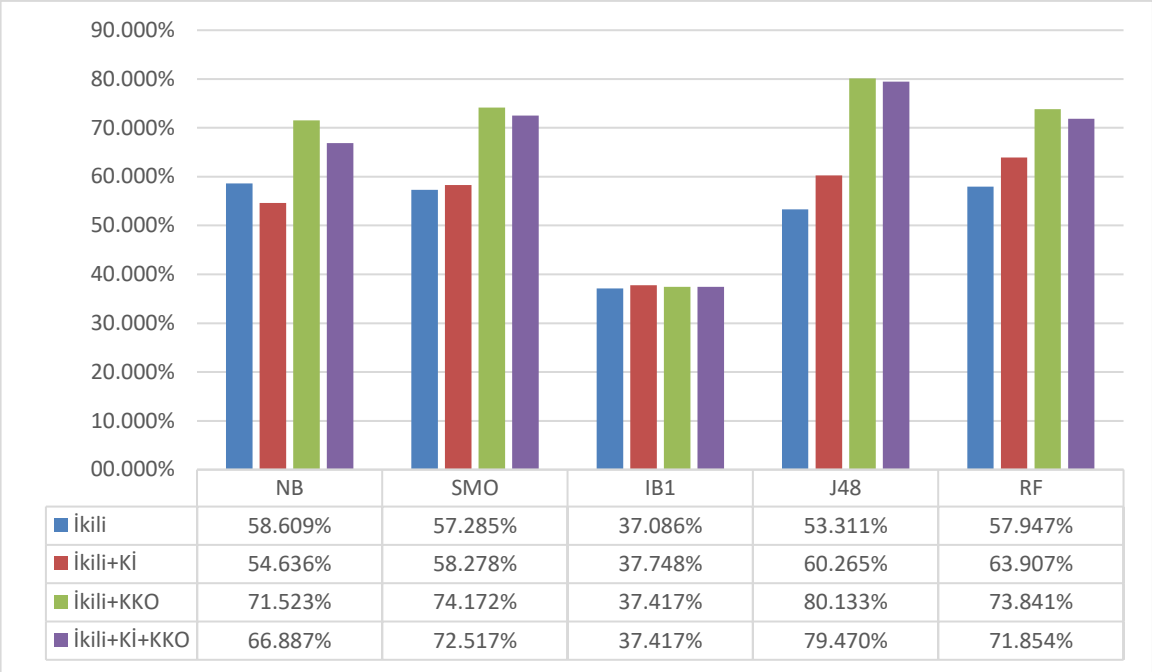
Dengeli veri kümesi, 10-kat çapraz doğrulama yöntemi ve Karar Ağacı (J48) ağırlıklandırma algoritması ile yapılan deneyde, tüm özelliklerin eklendiği İkili+Kİ+KKO (İkili + Kullanıcı İstatistiği + Kelime Kullanım Oranı) durumunda %80,8201 ile en iyi sonucu vermiştir. %80 bölümlendirme yöntemi ile yapılan deneylerde ise Karar Ağacı (J48) algoritması İkili+KKO (İkili + Kelime Kullanım Oranı) eklendiği durumda %80,1325 en iyi sonucu vermiştir. Dengeli veri kümesi ile yapılan deneylerde Kullanıcı İstatistiği (Kİ) özelliğinin eklenmesi sonucu olumsuz etkilemiştir.

Dengeli veri kümesi ile 10-kat çapraz doğrulama yöntemi kullanılarak yapılan deney sonuçları Şekil 4.1'de, detayları ise Tablo 4.1'de gösterilmiştir. Dengeli veri kümesi ile %80 bölümlendirme yöntemi kullanılarak yapılan deney sonuçları Şekil 4.2'de, detayları ise Tablo 4-2'de gösterilmiştir.



	NB	SMO	IB1	J48	RF
İkili	55.952%	57.275%	34.524%	51.918%	55.952%
İkili+Kİ	53.968%	58.069%	34.987%	59.524%	62.235%
İkili+KKO	68.915%	72.751%	35.384%	79.630%	73.810%
İkili+Kİ+KKO	66.601%	72.288%	35.582%	80.820%	74.735%

Şekil 4.1 İkili ağırlıklandırma dengeli veri kümesi 10-kat çapraz doğrulama sonuçları



	NB	SMO	IB1	J48	RF
İkili	58.609%	57.285%	37.086%	53.311%	57.947%
İkili+Kİ	54.636%	58.278%	37.748%	60.265%	63.907%
İkili+KKO	71.523%	74.172%	37.417%	80.133%	73.841%
İkili+Kİ+KKO	66.887%	72.517%	37.417%	79.470%	71.854%

Şekil 4.2 İkili ağırlıklandırma dengeli veri kümesi %80 bölümlendirme doğrulama sonuçları

Tablo 4.1 İkili ağırlıklandırma dengeli veri kümesi 10-kat çapraz doğrulama sonuçları

İKİLİ (BINARY)							
DENGESİZ VERİ KÜMESİ	10-KATMANLI ÇAPRAZ DOĞRULMA						
	TİP		TP	FP	KESİNLİK	DUYARLILIK	F-ÖLÇÜTÜ
	İKİLİ + KKO + Kİ	NB	0,519	0,274	0,547	0,519	0,459
		SMO	0,65	0,214	0,644	0,65	0,646
		IB1	0,454	0,321	0,469	0,454	0,456
		J48	0,456	0,196	0,676	0,681	0,678
		RF	0,688	0,208	0,666	0,688	0,661
	İKİLİ + KKO	NB	0,57	0,262	0,563	0,57	0,562
		SMO	0,65	0,214	0,644	0,65	0,646
		IB1	0,439	0,327	0,449	0,439	0,442
		J48	0,65	0,214	0,647	0,65	0,648
		RF	0,651	0,237	0,632	0,651	0,625
	İKİLİ + Kİ	NB	0,493	0,284	0,52	0,493	0,411
		SMO	0,551	0,271	0,546	0,551	0,548
		IB1	0,466	0,319	0,472	0,466	0,466
		J48	0,647	0,227	0,64	0,647	0,64
		RF	0,657	0,224	0,636	0,657	0,637
	İKİLİ	NB	0,588	0,259	0,568	0,588	0,588
		SMO	0,553	0,27	0,549	0,553	0,551
		IB1	0,438	0,339	0,436	0,438	0,436
		J48	0,524	0,309	0,515	0,524	0,515
		RF	0,569	0,287	0,549	0,569	0,55

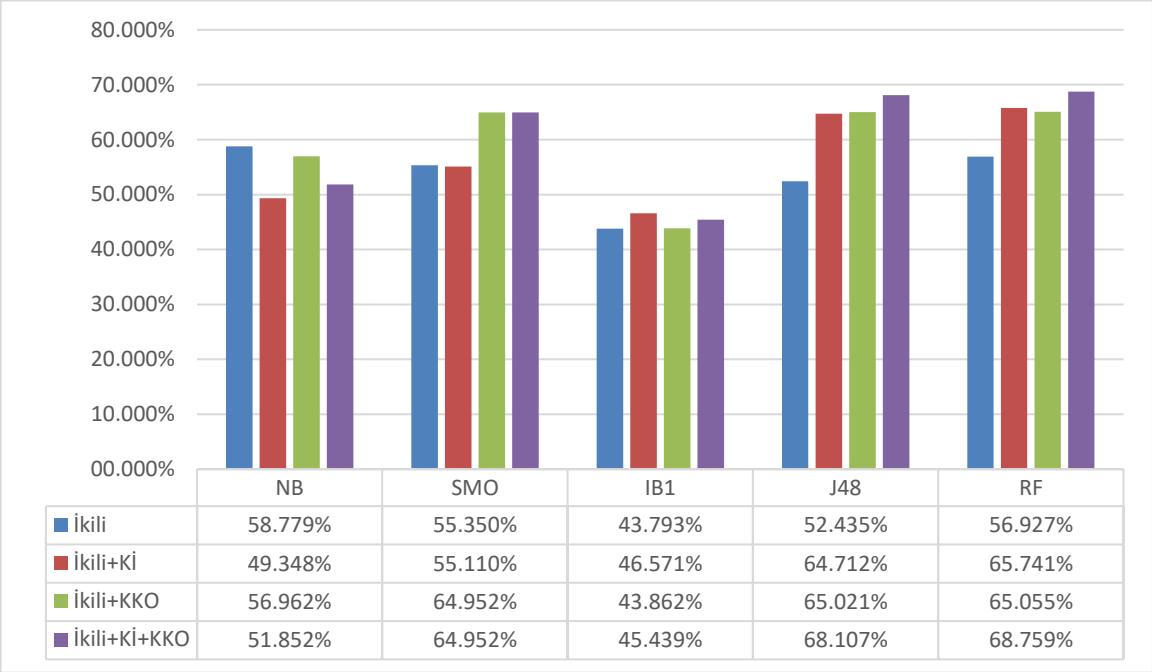
Tablo 4.2 İkili ağırlıklandırma dengeli veri kümesi %80 bölümlendirme doğrulama sonuçları

İKİLİ (BINARY)							
DENGESİZ VERİ KÜMESİ	%80 BÖLÜMLENDİRME						
	TİP		TP	FP	KESİNLİK	DUYARLILIK	F-ÖLÇÜTÜ
	İKİLİ + KKO + Kİ	NB	0,556	0,253	0,613	0,556	0,509
		SMO	0,64	0,223	0,634	0,64	0,635
		IB1	0,419	0,352	0,419	0,419	0,418
		J48	0,717	0,184	0,712	0,717	0,709
		RF	0,707	0,2	0,695	0,707	0,677
	İKİLİ + KKO	NB	0,566	0,277	0,562	0,566	0,556
		SMO	0,64	0,221	0,634	0,64	0,636
		IB1	0,395	0,361	0,395	0,395	0,395
		J48	0,624	0,237	0,622	0,624	0,622
		RF	0,643	0,246	0,635	0,643	0,617
	İKİLİ + Kİ	NB	0,491	0,299	0,515	0,491	0,395
		SMO	0,559	0,266	0,552	0,559	0,555
		IB1	0,461	0,321	0,475	0,461	0,462
		J48	0,623	0,249	0,631	0,623	0,618
		RF	0,654	0,233	0,634	0,654	0,627
	İKİLİ	NB	0,578	0,269	0,561	0,578	0,555
		SMO	0,554	0,274	0,547	0,554	0,549
		IB1	0,401	0,37	0,395	0,401	0,396
		J48	0,525	0,311	0,508	0,525	0,51
		RF	0,552	0,302	0,532	0,552	0,53

Dengesiz veri kümesin ile 10-kat çapraz doğrulama yöntem ile yapılan deneyde Rastgele Orman (RF) ağırlıklandırma algoritması tüm özelliklerin eklendiği İkili+KKO+Kİ (İkili + Kullanıcı İstatistiği + Kelime Kullanım Oranı) durumunda %68,7586 ile en iyi sonucu

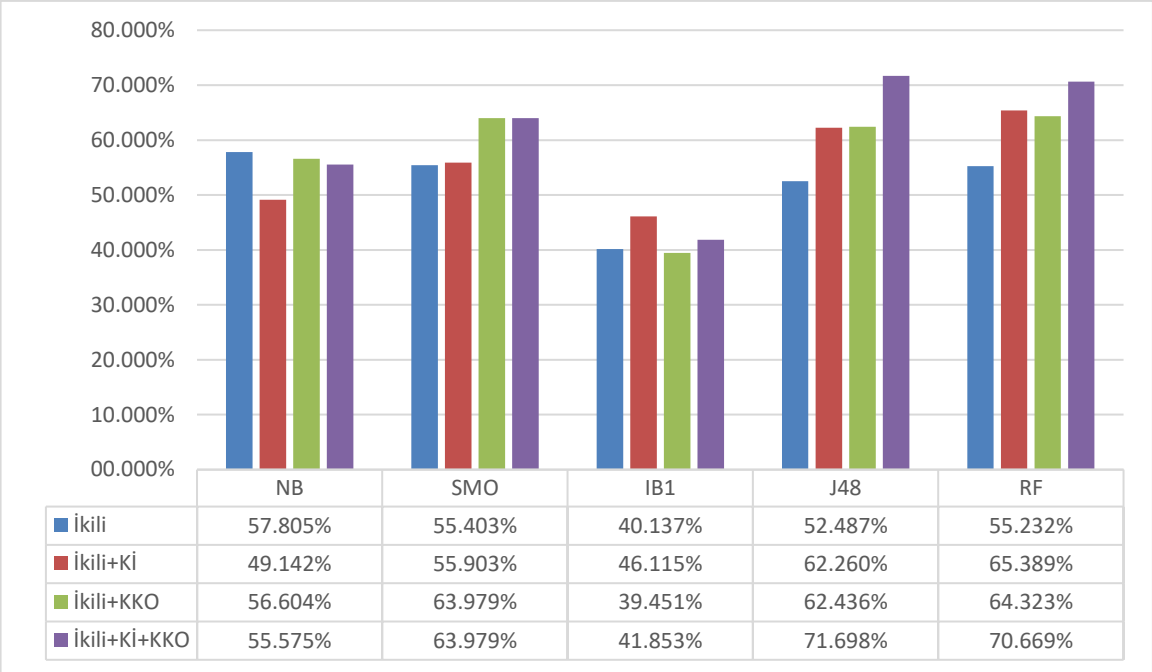
vermiştir. %80 bölümlendirme ile yapılan deneylerde ise Karar Ağacı (J48) algoritması tüm özelliklerin eklendiği İkili + Kullanıcı İstatistiği + Kelime Kullanım Oranı (İkili+KKO+Kİ) durumunda %71,6981 en iyi sonucu vermiştir.

Dengesiz veri kümesi ile 10-kat çapraz doğrulama yöntemi kullanılarak yapılan deney sonuçları Şekil 4.3'de, detayları ise Tablo 4.3'de gösterilmiştir. Dengesiz veri kümesi ile %80 bölümlendirme yöntemi kullanılarak yapılan deney sonuçları Şekil 4.4'de, detayları ise Tablo 4.4'de gösterilmiştir.



	NB	SMO	IB1	J48	RF
İkili	58.779%	55.350%	43.793%	52.435%	56.927%
İkili+Kİ	49.348%	55.110%	46.571%	64.712%	65.741%
İkili+KKO	56.962%	64.952%	43.862%	65.021%	65.055%
İkili+Kİ+KKO	51.852%	64.952%	45.439%	68.107%	68.759%

Şekil 4.3 İkili ağırlıklandırma dengesiz veri kümesi 10-kat çapraz doğrulama sonuçları



	NB	SMO	IB1	J48	RF
İkili	57.805%	55.403%	40.137%	52.487%	55.232%
İkili+Kİ	49.142%	55.903%	46.115%	62.260%	65.389%
İkili+KKO	56.604%	63.979%	39.451%	62.436%	64.323%
İkili+Kİ+KKO	55.575%	63.979%	41.853%	71.698%	70.669%

Şekil 4.4 İkili ağırlıklandırma dengesiz veri kümesi %80 bölümlendirme doğrulama sonuçları

Tablo 4.3 İkili ağırlıklandırma dengesiz veri kümesi 10-kat çapraz doğrulama sonuçları

İKİLİ (BINARY)							
DENGELİ VERİ KÜMESİ	10-KATMANLI ÇAPRAZ DOĞRULMA						
	TİP		TP	FP	KESİNLİK	DUYARLILIK	F-ÖLÇÜTÜ
	İKİLİ + KKO + KI	NB	0,666	0,167	0,676	0,666	0,652
		SMO	0,723	0,139	0,722	0,723	0,722
		IB1	0,356	0,322	0,479	0,356	0,264
		J48	0,808	0,096	0,809	0,808	0,808
		RF	0,747	0,126	0,745	0,747	0,738
	İKİLİ + KKO	NB	0,689	0,155	0,7	0,689	0,681
		SMO	0,728	0,136	0,726	0,728	0,726
		IB1	0,354	0,323	0,465	0,354	0,26
		J48	0,796	0,102	0,796	0,796	0,796
		RF	0,738	0,131	0,734	0,738	0,731
	İKİLİ + KI	NB	0,54	0,23	0,566	0,54	0,509
		SMO	0,581	0,21	0,575	0,581	0,576
		IB1	0,35	0,325	0,461	0,35	0,251
		J48	0,595	0,202	0,593	0,595	0,594
		RF	0,622	0,189	0,618	0,622	0,593
	İKİLİ	NB	0,56	0,22	0,549	0,56	0,546
		SMO	0,573	0,214	0,568	0,573	0,569
		IB1	0,345	0,327	0,445	0,345	0,245
		J48	0,519	0,24	0,51	0,519	0,512
		RF	0,56	0,22	0,569	0,56	0,526

Tablo 4.4 İkili ağırlıklandırma dengesiz veri kümesi %80 bölümlendirme doğrulama sonuçları

İKİLİ (BINARY)							
DENGELİ VERİ KÜMESİ	%80 BÖLÜMLENDİRME						
	TİP		TP	FP	KESİNLİK	DUYARLILIK	F-ÖLÇÜTÜ
	İKİLİ + KKO + KI	NB	0,669	0,163	0,684	0,669	0,66
		SMO	0,725	0,131	0,735	0,725	0,728
		IB1	0,374	0,341	0,485	0,374	0,284
		J48	0,795	0,099	0,798	0,795	0,794
		RF	0,719	0,138	0,72	0,719	0,715
	İKİLİ + KKO	NB	0,715	0,137	0,731	0,715	0,71
		SMO	0,742	0,123	0,751	0,742	0,744
		IB1	0,374	0,341	0,453	0,374	0,285
		J48	0,801	0,098	0,803	0,801	0,801
		RF	0,738	0,128	0,739	0,738	0,736
	İKİLİ + KI	NB	0,546	0,229	0,574	0,546	0,516
		SMO	0,583	0,204	0,588	0,583	0,584
		IB1	0,377	0,34	0,49	0,377	0,283
		J48	0,603	0,193	0,612	0,603	0,606
		RF	0,639	0,187	0,624	0,639	0,617
	İKİLİ	NB	0,586	0,2	0,596	0,586	0,584
		SMO	0,573	0,208	0,584	0,573	0,576
		IB1	0,371	0,344	0,437	0,371	0,272
		J48	0,533	0,234	0,536	0,533	0,532
		RF	0,579	0,223	0,592	0,579	0,554

4.1.1.2. Terim Frekansı (TF) Veri kümesi

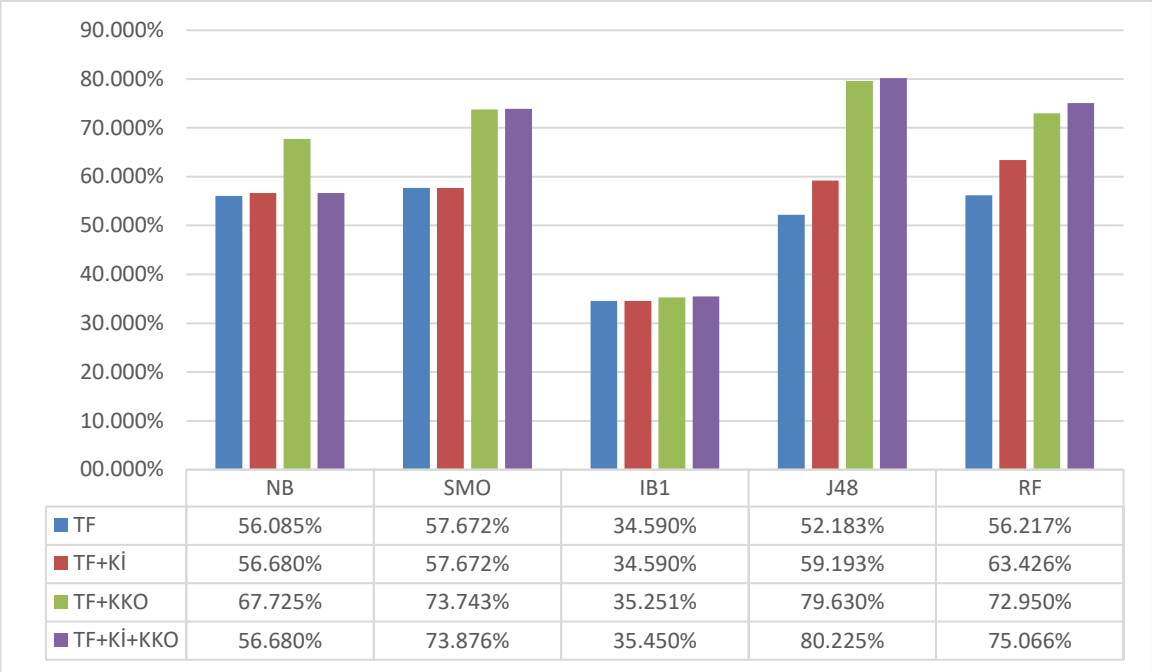
İkili veri kümesi ile yapılan deneylerden sonra Denetimli Makine Öğrenmesi yöntemi için Terim Frekansı (TF) ağırlıklandırma yöntemi kullanılarak deneyler yapılmıştır. Türkçe Twitter verileri için dengeli ve dengesiz veri kümeleri, Terim Frekansı (TF) ağırlıklandırma yöntemi kullanılarak oluşturulmuştur. Oluşturulan bu veri kümeleri üzerinde, Weka yazılımı kullanılarak 10-kat çapraz doğrulama ve %80 bölümlendirme yöntemleri ile deneyler yapılmıştır. Bu deneylerde sırasıyla;

- TF (Terim Frekansı),
- TF+Kİ (Terim Frekansı + Kullanıcı İstatistiği),
- TF+KKO (Terim Frekansı + Kelime Kullanım Oranı),
- TF+Kİ+KKO (Terim Frekansı + Kullanıcı İstatistiği + Kelime Kullanım Oranı),

için yapılmıştır.

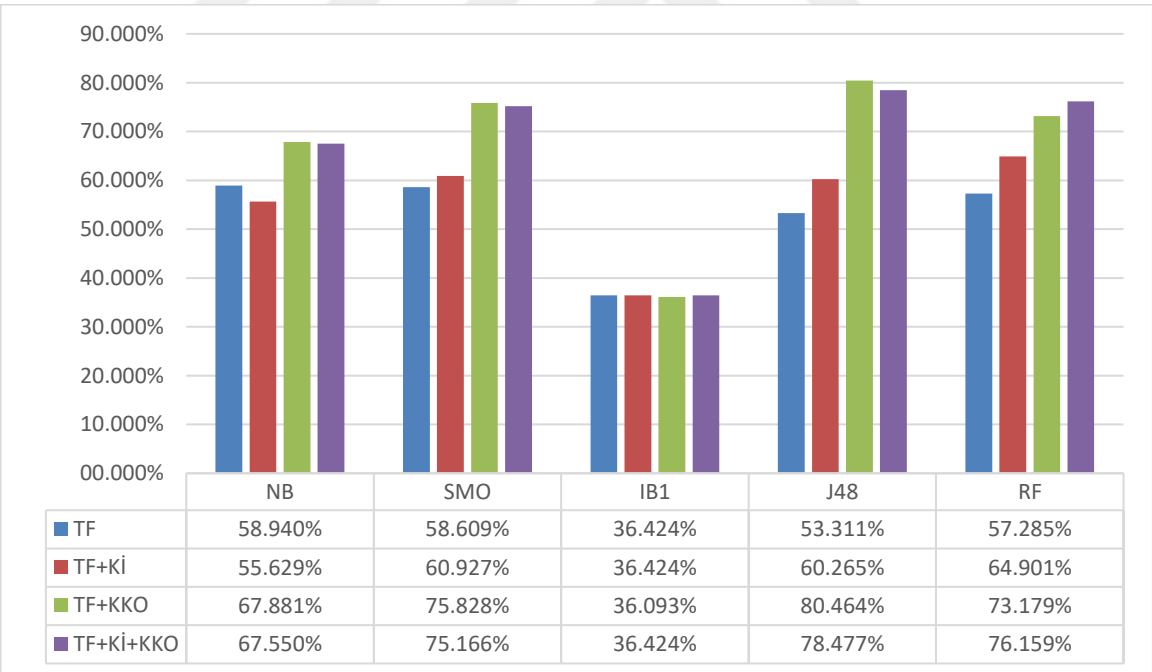
Dengeli veri kümesin ile 10-kat çapraz doğrulama yöntem ile yapılan deneyde Karar Ağacı (J48) ağırlıklandırma algoritması tüm özelliklerin eklendiği TF+Kİ+KKO durumunda %80,2249 ile en iyi sonucu vermiştir. %80 bölümlendirme ile yapılan deneylerde ise yine Karar Ağacı (J48) algoritması TF+KKO durumunda %80,4636 en iyi sonucu vermiştir. Kİ özelliğinin eklenmesi sonucu olumsuz etkilemiştir.

Dengeli veri kümesi ile 10-kat çapraz doğrulama yöntemi kullanılarak yapılan deney sonuçları Şekil 4.5'de, detayları ise Tablo 4.5'de gösterilmiştir. Dengeli veri kümesi ile %80 bölümlendirme yöntemi kullanılarak yapılan deney sonuçları Şekil 4.6'da, detayları ise Tablo 4.6'da gösterilmiştir.



	NB	SMO	IB1	J48	RF
TF	56.085%	57.672%	34.590%	52.183%	56.217%
TF+Kİ	56.680%	57.672%	34.590%	59.193%	63.426%
TF+KKO	67.725%	73.743%	35.251%	79.630%	72.950%
TF+Kİ+KKO	56.680%	73.876%	35.450%	80.225%	75.066%

Şekil 4.5 Terim Frekansı ağırlıklandırma dengeli veri kümesi 10-kat çapraz doğrulama sonuçları



	NB	SMO	IB1	J48	RF
TF	58.940%	58.609%	36.424%	53.311%	57.285%
TF+Kİ	55.629%	60.927%	36.424%	60.265%	64.901%
TF+KKO	67.881%	75.828%	36.093%	80.464%	73.179%
TF+Kİ+KKO	67.550%	75.166%	36.424%	78.477%	76.159%

Şekil 4.6 Terim Frekansı ağırlıklandırma dengeli veri kümesi %80 bölümlendirme doğrulama sonuçları

Tablo 4.5 Terim Frekansı ağırlıklandırma dengeli veri kümesi 10-kat çapraz doğrulama sonuçları

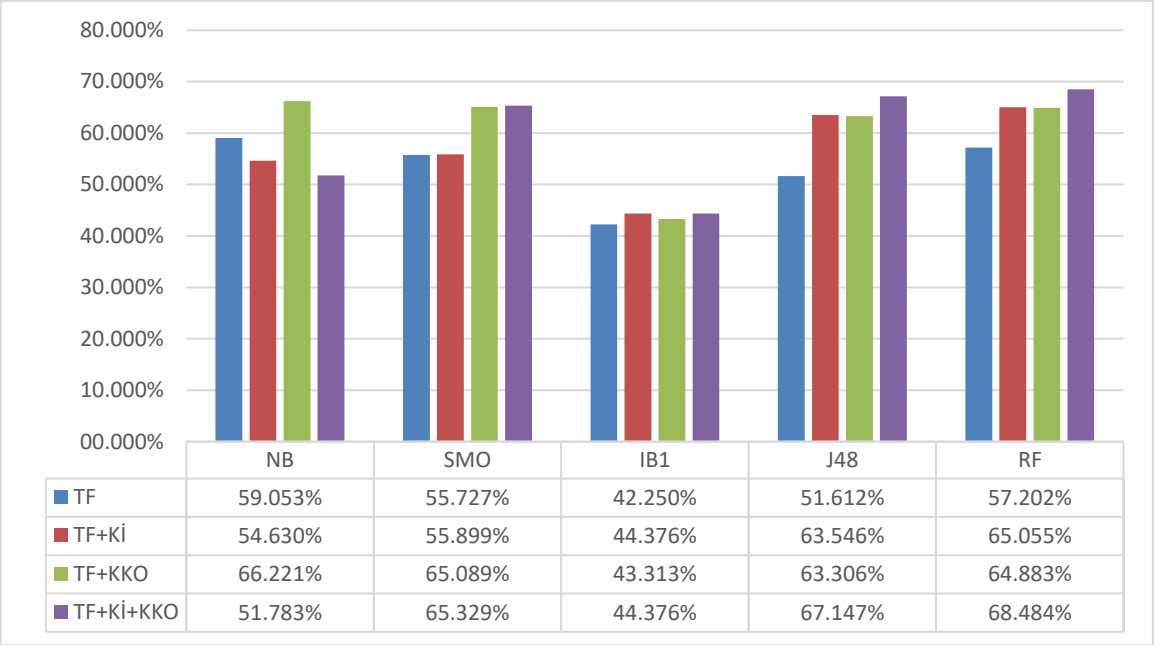
TF (TERİM FREKANSI)							
DENGELİ VERİ KÜMESİ	10-KATMANLI ÇAPRAZ DOĞRULMA						
	TİP		TP	FP	KESİNLİK	DUYARLILIK	F-ÖLÇÜTÜ
	TF + KKO + KI	NB	0,667	0,167	0,672	0,667	0,657
		SMO	0,739	0,131	0,739	0,739	0,738
		IB1	0,354	0,323	0,459	0,354	0,267
		J48	0,802	0,099	0,803	0,802	0,802
		RF	0,751	0,125	0,748	0,751	0,742
	TF + KKO	NB	0,677	0,161	0,685	0,677	0,676
		SMO	0,737	0,131	0,737	0,737	0,737
		IB1	0,353	0,324	0,45	0,353	0,263
		J48	0,796	0,102	0,796	0,796	0,796
		RF	0,729	0,135	0,726	0,729	0,722
	TF + KI	NB	0,567	0,217	0,571	0,567	0,548
		SMO	0,577	0,212	0,571	0,577	0,573
		IB1	0,346	0,327	0,446	0,346	0,247
		J48	0,592	0,204	0,59	0,592	0,591
		RF	0,634	0,183	0,634	0,634	0,606
	TF	NB	0,561	0,22	0,563	0,561	0,556
		SMO	0,577	0,212	0,572	0,577	0,573
		IB1	0,346	0,327	0,442	0,346	0,245
		J48	0,522	0,239	0,513	0,522	0,515
		RF	0,562	0,219	0,57	0,562	0,531

Tablo 4.6 Terim Frekansı ağırlıklandırma dengeli veri kümesi %80 bölümlendirme doğrulama sonuçları

TF (TERİM FREKANSI)							
DENGELİ VERİ KÜMESİ	%80 BÖLÜMLENDİRME						
	TİP		TP	FP	KESİNLİK	DUYARLILIK	F-ÖLÇÜTÜ
	TF + KKO + KI	NB	0,675	0,155	0,694	0,675	0,677
		SMO	0,752	0,12	0,756	0,752	0,753
		IB1	0,364	0,343	0,462	0,364	0,28
		J48	0,785	0,104	0,789	0,785	0,784
		RF	0,762	0,116	0,763	0,762	0,759
	TF + KKO	NB	0,679	0,149	0,71	0,679	0,685
		SMO	0,758	0,115	0,764	0,758	0,76
		IB1	0,361	0,345	0,42	0,361	0,274
		J48	0,805	0,095	0,808	0,805	0,806
		RF	0,732	0,13	0,737	0,732	0,729
	TF + KI	NB	0,556	0,219	0,568	0,556	0,548
		SMO	0,609	0,192	0,613	0,609	0,61
		IB1	0,364	0,345	0,452	0,364	0,268
		J48	0,603	0,191	0,622	0,603	0,608
		RF	0,649	0,183	0,636	0,649	0,626
	TF	NB	0,589	0,198	0,596	0,589	0,585
		SMO	0,586	0,203	0,594	0,586	0,588
		IB1	0,364	0,345	0,415	0,364	0,269
		J48	0,533	0,235	0,533	0,533	0,531
		RF	0,573	0,226	0,582	0,573	0,551

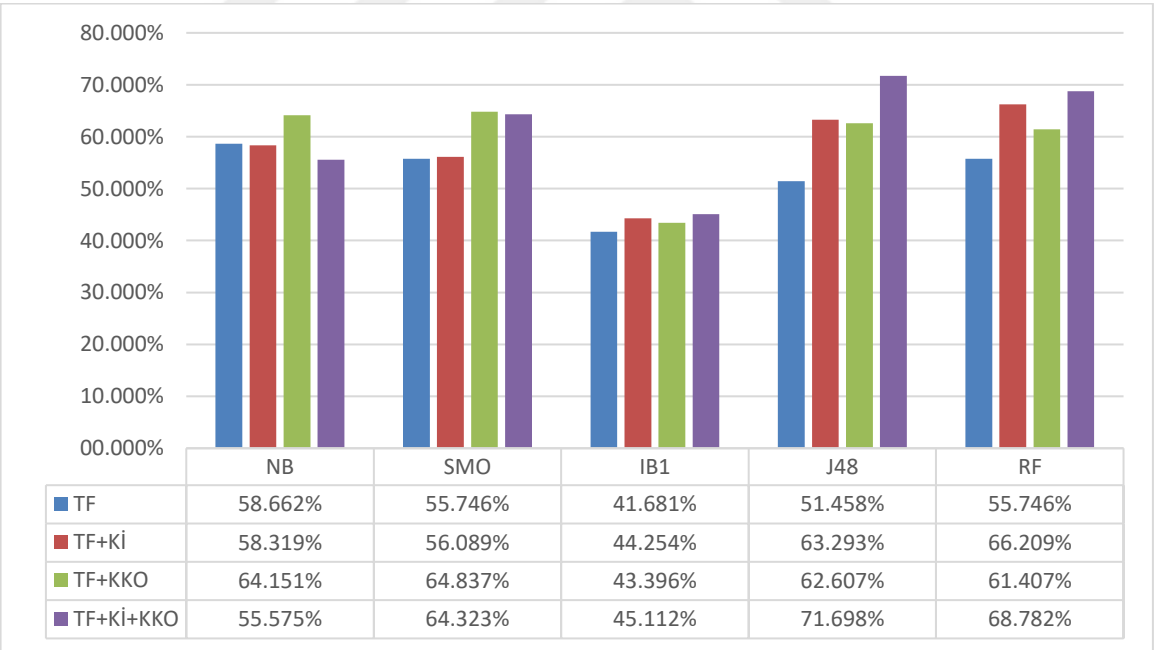
Dengesiz veri kümesi ile 10-kat çapraz doğrulama yöntem ile yapılan deneylerde Rastgele Orman (RF) ağırlıklandırma algoritması tüm özelliklerin eklendiği Terim Frekansı + Kullanıcı İstatistiği + Kelime Kullanım Oranı (TF+KKO+Kİ) durumunda %68,4842 ile en iyi sonucu vermiştir. %80 bölümlendirme ile yapılan deneylerde ise Karar Ağacı (J48) algoritması tüm özelliklerin eklendiği Terim Frekansı + Kullanıcı İstatistiği + Kelime Kullanım Oranı (TF+KKO+Kİ) durumunda %71,6981 en iyi sonucu vermiştir.

Dengesiz veri kümesi ile 10-kat çapraz doğrulama yöntemi kullanılarak yapılan deney sonuçları Şekil 4.7'de, detayları ise Tablo 4.7'de gösterilmiştir. Dengesiz veri kümesi ile %80 bölümlendirme yöntemi kullanılarak yapılan deney sonuçları Şekil 4.8'de, detayları ise Tablo 4.8'de gösterilmiştir.



	NB	SMO	IB1	J48	RF
TF	59.053%	55.727%	42.250%	51.612%	57.202%
TF+Kİ	54.630%	55.899%	44.376%	63.546%	65.055%
TF+KKO	66.221%	65.089%	43.313%	63.306%	64.883%
TF+Kİ+KKO	51.783%	65.329%	44.376%	67.147%	68.484%

Şekil 4.7 Terim Frekansı ağırlıklandırma dengesiz veri kümesi 10-kat çapraz doğrulama sonuçları



	NB	SMO	IB1	J48	RF
TF	58.662%	55.746%	41.681%	51.458%	55.746%
TF+Kİ	58.319%	56.089%	44.254%	63.293%	66.209%
TF+KKO	64.151%	64.837%	43.396%	62.607%	61.407%
TF+Kİ+KKO	55.575%	64.323%	45.112%	71.698%	68.782%

Şekil 4.8 Terim Frekansı ağırlıklandırma dengesiz veri kümesi %80 bölümlendirme doğrulama sonuçları

Tablo 4.7 Terim Frekansı ağırlıklandırma dengesiz veri kümesi 10-kat çapraz doğrulama sonuçları

TF (TERİM FREKANSI)							
DENGESİZ VERİ KÜMESİ	10-KATMANLI ÇAPRAZ DOĞRULMA						
	TİP		TP	FP	KESİNLİK	DUYARLILIK	F-ÖLÇÜTÜ
	TF + KKO + KI	NB	0.518	0.274	0.546	0.518	0.458
		SMO	0,653	0,217	0,648	0,653	0,648
		IB1	0,444	0,321	0,454	0,444	0,448
		J48	0,671	0,202	0,667	0,671	0,668
		RF	0,685	0,211	0,668	0,685	0,659
	TF + KKO	NB	0,662	0,231	0,656	0,662	0,632
		SMO	0,651	0,218	0,645	0,651	0,646
		IB1	0,433	0,325	0,443	0,433	0,437
		J48	0,633	0,223	0,631	0,633	0,632
		RF	0,649	0,237	0,632	0,649	0,625
	TF + KI	NB	0,546	0,254	0,57	0,546	0,498
		SMO	0,559	0,276	0,55	0,559	0,552
		IB1	0,444	0,326	0,449	0,444	0,446
		J48	0,635	0,23	0,627	0,635	0,629
		RF	0,651	0,23	0,629	0,651	0,63
	TF	NB	0,591	0,255	0,571	0,591	0,573
		SMO	0,557	0,278	0,548	0,557	0,551
		IB1	0,422	0,334	0,426	0,422	0,423
		J48	0,516	0,312	0,505	0,516	0,507
		RF	0,572	0,285	0,55	0,572	0,552

Tablo 4.8 Terim Frekansı ağırlıklandırma dengesiz veri kümesi %80 bölümlendirme doğrulama sonuçları

TF (TERİM FREKANSI)							
DENGESİZ VERİ KÜMESİ	%80 BÖLÜMLENDİRME						
	TİP		TP	FP	KESİNLİK	DUYARLILIK	F-ÖLÇÜTÜ
	TF + KKO + KI	NB	0,556	0,254	0,609	0,556	0,511
		SMO	0,643	0,231	0,638	0,643	0,637
		IB1	0,451	0,325	0,453	0,451	0,452
		J48	0,717	0,177	0,713	0,717	0,713
		RF	0,688	0,214	0,675	0,688	0,655
	TF + KKO	NB	0,642	0,247	0,651	0,642	0,602
		SMO	0,648	0,227	0,644	0,648	0,642
		IB1	0,434	0,331	0,436	0,434	0,434
		J48	0,626	0,23	0,623	0,626	0,624
		RF	0,614	0,264	0,603	0,614	0,589
	TF + KI	NB	0,583	0,233	0,623	0,583	0,546
		SMO	0,561	0,28	0,551	0,561	0,553
		IB1	0,443	0,334	0,439	0,443	0,44
		J48	0,633	0,229	0,628	0,633	0,625
		RF	0,662	0,231	0,661	0,662	0,64
	TF	NB	0,587	0,263	0,568	0,587	0,565
		SMO	0,557	0,282	0,547	0,557	0,549
		IB1	0,417	0,343	0,413	0,417	0,412
		J48	0,515	0,317	0,5	0,515	0,501
		RF	0,557	0,296	0,529	0,557	0,532

4.1.1.3. Terim Frekansı-Ters Doküman Frekansı (TF-IDF) Veri Kümesi

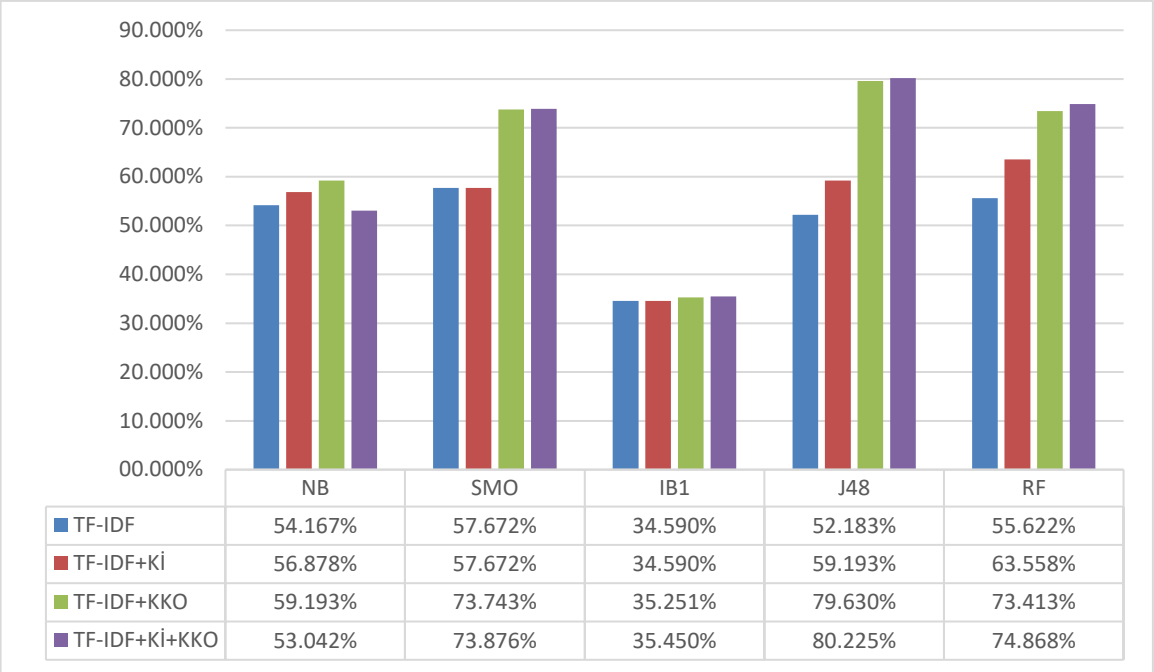
İkili ve Terim Frekansı (TF) ağırlıklandırma yöntemleri için yapılan deneylerden sonra Terim Frekansı-Ters Doküman Frekansı (TF-IDF) ağırlıklandırma yöntemi kullanılarak deneyler yapılmıştır. TF-IDF ağırlıklandırma yöntemi kullanılarak oluşturulan dengeli ve dengesiz veri kümeleri ile Weka yazılımı kullanılarak 10-kat çapraz doğrulama ve %80 bölümlendirme yöntemleri kullanılarak deneyler yapılmıştır. Bu deneylerde;

- TF- IDF (Terim Frekansı-Ters Doküman Frekansı),
- TF-IDF+Kİ (Terim Frekansı- Ters Doküman Frekansı + Kullanıcı İstatistiği),
- TF-IDF+KKO (Terim Frekansı- Ters Doküman Frekansı + Kelime Kullanım Oranı),
- TF-IDF+Kİ+KKO (Terim Frekansı- Ters Doküman Frekansı + Kullanıcı İstatistiği + Kelime Kullanım Oranı),

özellikleri kullanılmıştır.

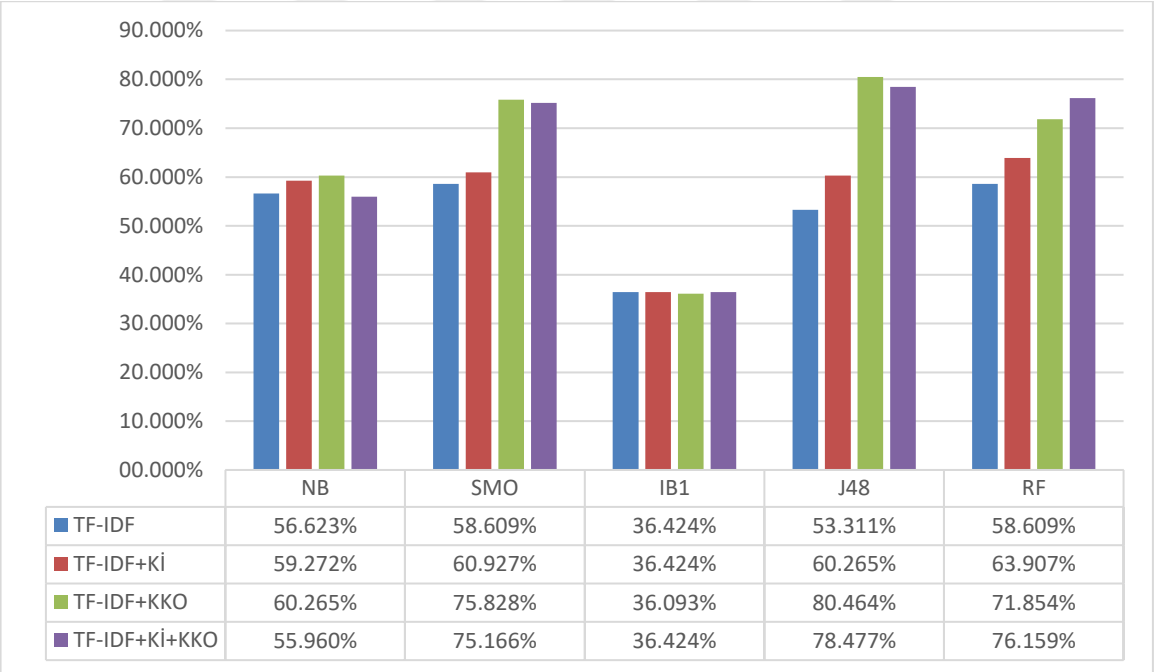
Dengeli veri kümesin ile 10-kat çapraz doğrulama yöntem ile yapılan deneyde Karar Ağacı (J48)ağırlıklandırma algoritması tüm özelliklerin eklendiği Terim Frekansı-Ters Doküman Frekansı + Kullanıcı İstatistiği + Kelime Kullanım Oranı(TF-IDF+Kİ+KKO)durumunda %80,2249 ile en iyi sonucu vermiştir. %80 bölümlendirme ile yapılan deneylerde ise Karar Ağacı (J48) algoritması TF-IDF+KKO eklendiği durumda %80,4636 en iyi sonucu vermiştir. Kİ özelliğinin eklenmesi diğer deneylerde olduğu gibi bu deneyde de sonucu olumsuz etkilemiştir.

Dengeli veri kümesi ile 10-kat çapraz doğrulama yöntemi kullanılarak yapılan deney sonuçları Şekil 4.9'da, detayları ise Tablo 4.9'da gösterilmiştir. Dengeli veri kümesi ile %80 bölümlendirme yöntemi kullanılarak yapılan deney sonuçları Şekil 4.10'da, detayları ise Tablo 4.10'da gösterilmiştir.



	NB	SMO	IB1	J48	RF
TF-IDF	54.167%	57.672%	34.590%	52.183%	55.622%
TF-IDF+Kİ	56.878%	57.672%	34.590%	59.193%	63.558%
TF-IDF+KKO	59.193%	73.743%	35.251%	79.630%	73.413%
TF-IDF+Kİ+KKO	53.042%	73.876%	35.450%	80.225%	74.868%

Şekil 4.9 Terim Frekansı-Ters Doküman Frekansı ağırlıklandırma dengeli veri kümesi 10-kat çapraz doğrulama sonuçları



	NB	SMO	IB1	J48	RF
TF-IDF	56.623%	58.609%	36.424%	53.311%	58.609%
TF-IDF+Kİ	59.272%	60.927%	36.424%	60.265%	63.907%
TF-IDF+KKO	60.265%	75.828%	36.093%	80.464%	71.854%
TF-IDF+Kİ+KKO	55.960%	75.166%	36.424%	78.477%	76.159%

Şekil 4.10 Terim Frekansı-Ters Doküman Frekansı ağırlıklandırma dengeli veri kümesi %80 bölümlendirme doğrulama sonuçları

Tablo 4.9 Terim Frekansı-Ters Doküman Frekansı ağırlıklandırma dengeli veri kümesi 10-kat çapraz doğrulama sonuçları

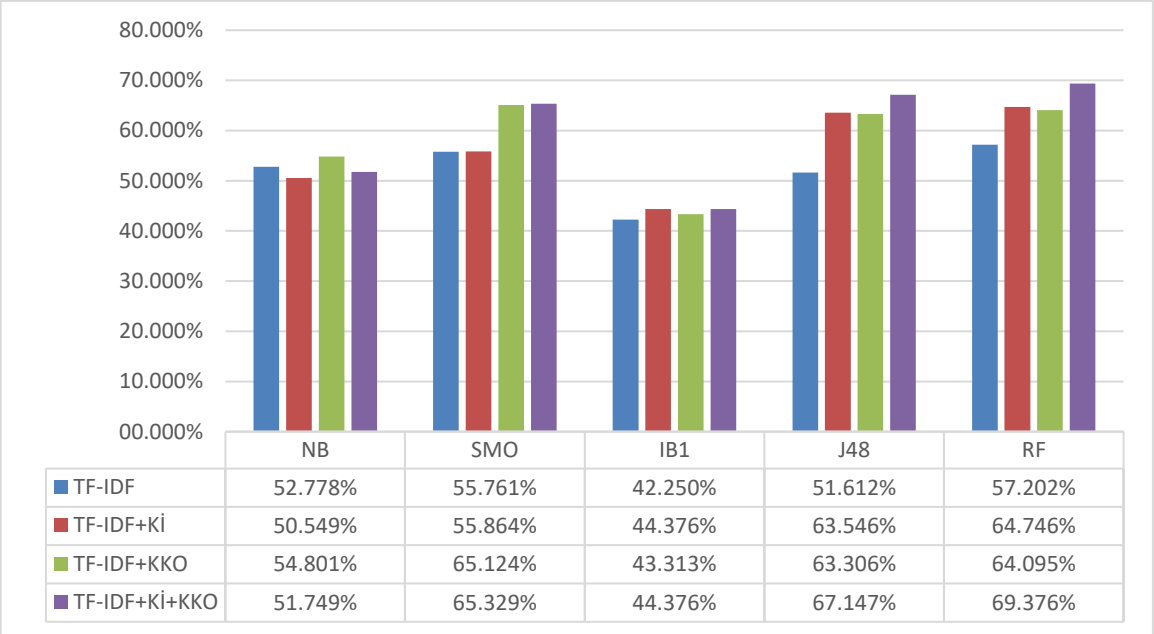
TF-IDF (TERİM FREKANSI-TERS DOKÜMAN FREKANSI)							
DENGELİ VERİ KÜMESİ	10-KATMANLI ÇAPRAZ DOĞRULMA						
	TİP		TP	FP	KESİNLİK	DUYARLILIK	F-ÖLÇÜTÜ
	TF-IDF+KKO+KI	NB	0,53	0,235	0,549	0,53	0,485
		SMO	0,739	0,131	0,739	0,739	0,738
		IB1	0,354	0,323	0,459	0,354	0,267
		J48	0,802	0,099	0,803	0,802	0,802
		RF	0,749	0,126	0,746	0,749	0,74
	TF-IDF + KKO	NB	0,592	0,204	0,586	0,592	0,586
		SMO	0,737	0,131	0,737	0,737	0,737
		IB1	0,353	0,324	0,45	0,353	0,263
		J48	0,796	0,102	0,796	0,796	0,796
		RF	0,734	0,133	0,731	0,734	0,727
	TF-IDF + KI	NB	0,569	0,216	0,576	0,569	0,535
		SMO	0,577	0,212	0,571	0,577	0,573
		IB1	0,346	0,327	0,446	0,346	0,247
		J48	0,592	0,204	0,59	0,592	0,591
		RF	0,636	0,182	0,637	0,636	0,607
	TF-IDF	NB	0,542	0,229	0,532	0,542	0,533
		SMO	0,577	0,212	0,572	0,577	0,573
		IB1	0,346	0,327	0,442	0,346	0,245
		J48	0,522	0,239	0,513	0,522	0,515
		RF	0,556	0,222	0,562	0,556	0,521

Tablo 4.10 Terim Frekansı-Ters Doküman Frekansı ağırlıklandırma dengeli veri kümesi %80 bölümlendirme doğrulama sonuçları

TF-IDF (TERİM FREKANSI-TERS DOKÜMAN FREKANSI)							
DENGELİ VERİ KÜMESİ	%80 BÖLÜMLENDİRME						
	TİP		TP	FP	KESİNLİK	DUYARLILIK	F-ÖLÇÜTÜ
	TF-IDF+KKO+KI	NB	0,56	0,233	0,568	0,56	0,518
		SMO	0,752	0,12	0,756	0,752	0,753
		IB1	0,364	0,343	0,462	0,364	0,28
		J48	0,785	0,104	0,789	0,785	0,784
		RF	0,762	0,116	0,761	0,762	0,758
	TF-IDF + KKO	NB	0,603	0,193	0,606	0,603	0,6
		SMO	0,758	0,115	0,764	0,758	0,76
		IB1	0,361	0,345	0,42	0,361	0,274
		J48	0,805	0,095	0,808	0,805	0,806
		RF	0,719	0,139	0,714	0,719	0,714
	TF-IDF + KI	NB	0,593	0,211	0,58	0,593	0,564
		SMO	0,609	0,192	0,613	0,609	0,61
		IB1	0,364	0,345	0,452	0,364	0,268
		J48	0,603	0,191	0,622	0,603	0,608
		RF	0,639	0,189	0,626	0,639	0,615
	TF-IDF	NB	0,566	0,214	0,561	0,566	0,563
		SMO	0,586	0,203	0,594	0,586	0,588
		IB1	0,364	0,345	0,415	0,364	0,269
		J48	0,533	0,235	0,533	0,533	0,531
		RF	0,586	0,217	0,584	0,586	0,563

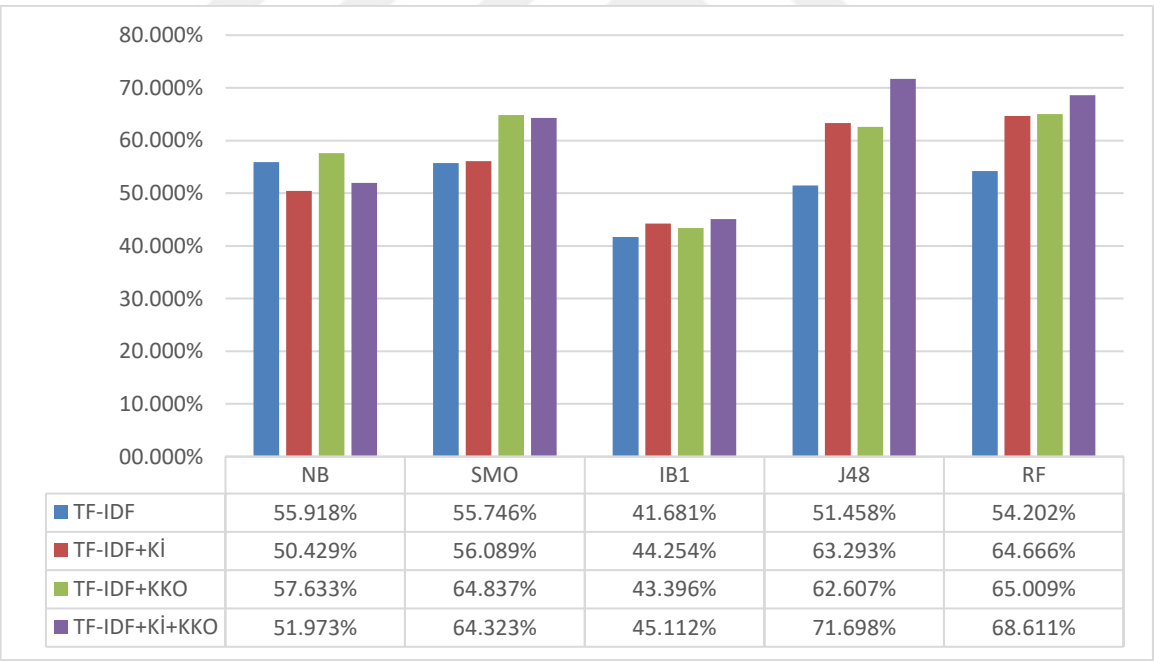
Dengesiz veri kümesi ile 10-kat çapraz doğrulama yöntem kullanılarak yapılan deneylerde Rastgele Orman (RF) ağırlıklandırma algoritması tüm özelliklerin eklendiği Terim Frekansı-Ters Doküman Frekansı + Kullanıcı İstatistiği + Kelime Kullanım Oranı (TF-IDF+KKO+KI) durumunda %69,3759 ile en iyi sonucu vermiştir. %80 bölümlendirme ile yapılan deneylerde ise Karar Ağacı (J48) algoritması tüm özelliklerin eklendiği TF-IDF+KKO+KI durumunda %71,6981 en iyi sonucu vermiştir.

Dengesiz veri kümesi ile 10-kat çapraz doğrulama yöntemi kullanılarak yapılan deney sonuçları Şekil 4.11'de, detayları ise Tablo 4.11'de gösterilmiştir. Dengesiz veri kümesi ile %80 bölümlendirme yöntemi kullanılarak yapılan deney sonuçları Şekil 4.12'de, detayları ise Tablo 4.12'de gösterilmiştir.



	NB	SMO	IB1	J48	RF
TF-IDF	52.778%	55.761%	42.250%	51.612%	57.202%
TF-IDF+Kİ	50.549%	55.864%	44.376%	63.546%	64.746%
TF-IDF+KKO	54.801%	65.124%	43.313%	63.306%	64.095%
TF-IDF+Kİ+KKO	51.749%	65.329%	44.376%	67.147%	69.376%

Şekil 4.11 Terim Frekansı-Ters Doküman Frekansı ağırlıklandırma dengesiz veri kümesi 10-kat çapraz doğrulama sonuçları



	NB	SMO	IB1	J48	RF
TF-IDF	55.918%	55.746%	41.681%	51.458%	54.202%
TF-IDF+Kİ	50.429%	56.089%	44.254%	63.293%	64.666%
TF-IDF+KKO	57.633%	64.837%	43.396%	62.607%	65.009%
TF-IDF+Kİ+KKO	51.973%	64.323%	45.112%	71.698%	68.611%

Şekil 4.12 Terim Frekansı-Ters Doküman Frekansı ağırlıklandırma dengesiz veri kümesi %80 bölümlendirme doğrulama sonuçları

Tablo 4.11 Terim Frekansı-Ters Doküman Frekansı ağırlıklandırma dengesiz veri kümesi 10-kat çapraz doğrulama sonuçları

TF-IDF (TERİM FREKANSI-TERS DOKÜMAN FREKANSI)							
DENGESİZ VERİ KÜMESİ	10-KATMANLI ÇAPRAZ DOĞRULMA						
	TİP		TP	FP	KESİNLİK	DUYARLILIK	F-ÖLÇÜTÜ
	TF-IDF+KKO+KI	NB	0,517	0,266	0,55	0,517	0,456
		SMO	0,653	0,217	0,648	0,653	0,648
		IB1	0,444	0,321	0,454	0,444	0,448
		J48	0,671	0,202	0,667	0,671	0,668
		RF	0,694	0,204	0,674	0,694	0,668
	TF-IDF + KKO	NB	0,548	0,269	0,546	0,548	0,546
		SMO	0,651	0,218	0,646	0,651	0,646
		IB1	0,433	0,325	0,443	0,433	0,437
		J48	0,633	0,223	0,631	0,633	0,632
		RF	0,641	0,243	0,619	0,641	0,613
	TF-IDF + KI	NB	0,505	0,273	0,532	0,505	0,439
		SMO	0,559	0,277	0,549	0,559	0,552
		IB1	0,444	0,326	0,449	0,444	0,446
		J48	0,635	0,23	0,627	0,635	0,629
		RF	0,647	0,231	0,624	0,647	0,626
	TF-IDF	NB	0,528	0,279	0,526	0,528	0,526
		SMO	0,558	0,277	0,548	0,558	0,551
		IB1	0,422	0,334	0,426	0,422	0,423
		J48	0,516	0,312	0,505	0,516	0,507
		RF	0,572	0,283	0,546	0,572	0,551

Tablo 4.12 Terim Frekansı-Ters Doküman Frekansı ağırlıklandırma dengesiz veri kümesi %80 bölümlendirme doğrulama sonuçları

TF-IDF (TERİM FREKANSI-TERS DOKÜMAN FREKANSI)							
DENGESİZ VERİ KÜMESİ	%80 BÖLÜMLENDİRME						
	TİP		TP	FP	KESİNLİK	DUYARLILIK	F-ÖLÇÜTÜ
	TF-IDF+KKO+KI	NB	0,52	0,261	0,565	0,52	0,455
		SMO	0,643	0,231	0,638	0,643	0,637
		IB1	0,451	0,325	0,453	0,451	0,452
		J48	0,717	0,177	0,713	0,717	0,713
		RF	0,686	0,214	0,673	0,686	0,657
	TF-IDF + KKO	NB	0,576	0,257	0,571	0,576	0,572
		SMO	0,648	0,227	0,644	0,648	0,642
		IB1	0,434	0,331	0,436	0,434	0,434
		J48	0,626	0,23	0,623	0,626	0,624
		RF	0,65	0,239	0,64	0,65	0,623
	TF-IDF + KI	NB	0,504	0,267	0,55	0,504	0,433
		SMO	0,561	0,28	0,551	0,561	0,553
		IB1	0,443	0,334	0,439	0,443	0,44
		J48	0,633	0,229	0,628	0,633	0,625
		RF	0,647	0,243	0,636	0,647	0,621
	TF-IDF	NB	0,559	0,269	0,551	0,559	0,554
		SMO	0,557	0,282	0,547	0,557	0,549
		IB1	0,417	0,343	0,413	0,417	0,412
		J48	0,515	0,317	0,5	0,515	0,501
		RF	0,542	0,301	0,519	0,542	0,521

4.1.2 Yarı-Denetimli Makine Öğrenmesi

Makine öğrenmesi (MÖ) yöntemlerinden yarı-denetimli öğrenme ile yapılacak deneyler için Bankacılık hakkında etiketlenmiş veriler kullanılarak Özellik–İlişki Dosya Formatı (ARFF) dosyaları İkili, Terim Frekansı (TF) ve Terim Frekansı-Ters Doküman Frekansı (TF-IDF) ağırlıklandırma yöntemleri kullanılarak hazırlanmıştır. Bankacılık hakkındaki veri kümesine ait detaylı bilgi Tablo 3.3'de verilmiştir. Bu veri kümeleri deneylerde kullanılmak üzere etiketli eğitim veri kümesi ve etiketsiz veri kümesi olarak 2'ye bölünmüştür. Deneylerde, eğitim verisi, toplam veri kümesinin %1, %5, %10, %20, %30, %40 ve %50'si olarak ele alınmıştır. Eğitim verisi dışında kalan verilerin etiket bilgileri silinmiş ve bu veriler için deneyler yapılmıştır. Yapılan bu deneyler sonucunda en iyi sonuç, %50 eğitim kümesi ve ikili ağırlıklandırma yöntemi kullanılarak yapılan deneyde %72,055 olarak bulunmuştur.

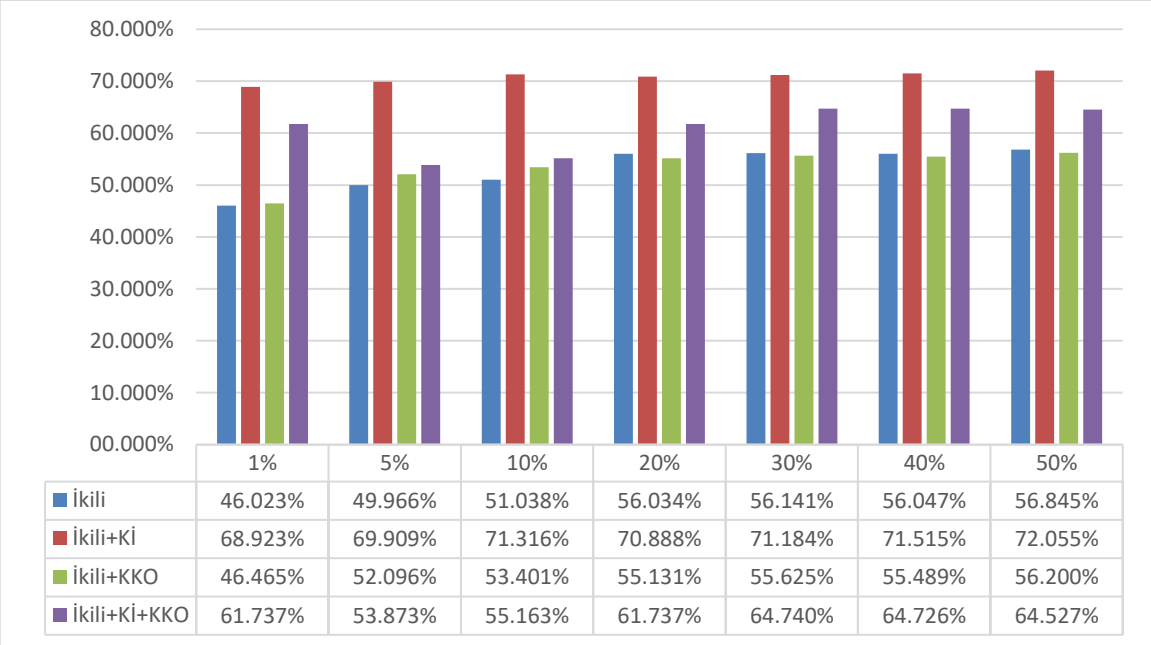
4.1.2.1 İkili Veri Kümesi (Binary/Boolean)

Türkçe Twitter verileri kullanılarak İkili (Binary/Boolean) ağırlıklandırma yöntemi ile oluşturulan veri kümesi ile Weka yazılımı kullanılarak deneyler yapılmıştır. Bu deneylerde veri kümesinin %1, %5, %10, %20, %30, %40 ve %50'si eğitim verisi olarak ayrılmıştır. Bu eğitim verisi dışında kalan verilerin etiketleri silinmiş ve etiketsiz verilerin etiketlerinin bulunması için deneyler yapılmıştır. Bu deneyler sırasıyla;

- İkili,
- İkili+Kİ (İkili + Kullanıcı İstatistiği),
- İkili+KKO (İkili + Kelime Kullanım Oranı),
- İkili+Kİ+KKO (İkili + Kullanıcı İstatistiği + Kelime Kullanım Oranı),

için yapılmıştır.

İkili ağırlıklandırma yöntemi ile yapılan deneylerde en iyi sonuç İkili+Kİ (İkili + Kullanıcı İstatistiği) için %50'lik eğitim kümesi ile yapılan deneylerde %72,055 ile elde edilmiştir. KKO (Kelime Kullanım Oranı) bilgisinin eklenmesi deneylere olumsuz yönde etki etmiştir. İkili ağırlıklandırma yöntemi ile yapılan deney sonuçları Şekil 4.13'de, detayları ise Tablo 4.13'de gösterilmiştir.



	1%	5%	10%	20%	30%	40%	50%
İkili	46.023%	49.966%	51.038%	56.034%	56.141%	56.047%	56.845%
İkili+Kİ	68.923%	69.909%	71.316%	70.888%	71.184%	71.515%	72.055%
İkili+KKO	46.465%	52.096%	53.401%	55.131%	55.625%	55.489%	56.200%
İkili+Kİ+KKO	61.737%	53.873%	55.163%	61.737%	64.740%	64.726%	64.527%

Şekil 4.13 Yarı-Denetimli öğrenme İkili ağırlıklandırma doğrulama sonuçları

Tablo 4.13 Yarı-Denetimli öğrenme İkili ağırlıklandırma doğrulama sonuçları

İKİLİ (BINARY)							
	TİP		TP	FP	KESİNLİK	DUYARLILIK	F-ÖLÇÜTÜ
YARI-DENETİMLİ	İKİLİ + KKO + Kİ	1%	0,617	0,275	0,593	0,617	0,588
		5%	0,539	0,320	0,558	0,539	0,521
		10%	0,552	0,317	0,554	0,552	0,533
		20%	0,617	0,275	0,593	0,617	0,588
		30%	0,647	0,251	0,640	0,647	0,632
		40%	0,647	0,252	0,643	0,647	0,632
		50%	0,645	0,252	0,637	0,645	0,629
	İKİLİ + KKO	1%	0,465	0,440	0,468	0,465	0,344
		5%	0,521	0,361	0,511	0,521	0,485
		10%	0,534	0,353	0,536	0,534	0,494
		20%	0,551	0,352	0,571	0,551	0,509
		30%	0,556	0,344	0,575	0,556	0,518
		40%	0,555	0,349	0,575	0,555	0,513
		50%	0,562	0,332	0,581	0,562	0,531
	İKİLİ + Kİ	1%	0,689	0,226	0,679	0,689	0,662
		5%	0,699	0,205	0,695	0,699	0,689
		10%	0,713	0,193	0,702	0,713	0,702
		20%	0,709	0,201	0,702	0,709	0,698
		30%	0,712	0,198	0,705	0,712	0,702
		40%	0,715	0,198	0,711	0,715	0,705
		50%	0,721	0,193	0,714	0,721	0,710
	İKİLİ	1%	0,460	0,441	0,449	0,460	0,344
		5%	0,500	0,411	0,588	0,500	0,400
		10%	0,510	0,407	0,587	0,510	0,413
		20%	0,560	0,343	0,564	0,560	0,516
		30%	0,561	0,341	0,564	0,561	0,518
		40%	0,560	0,342	0,569	0,560	0,518
		50%	0,568	0,338	0,594	0,568	0,525

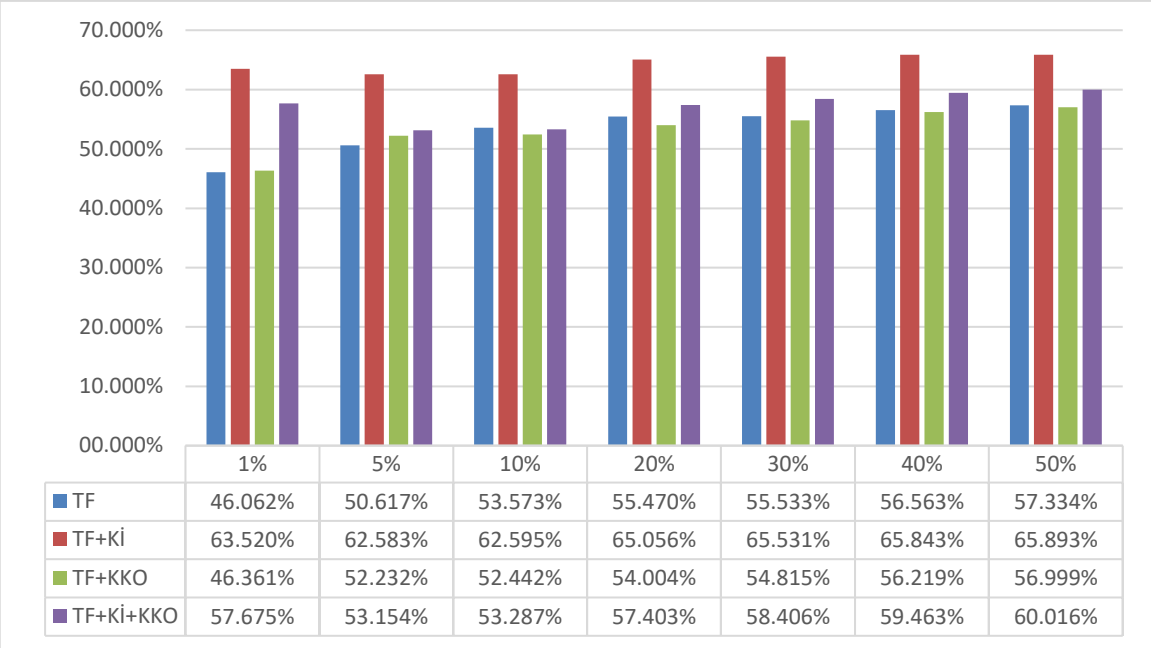
4.1.2.2 Terim Frekansı (TF) Veri Kümesi

Bankacılık hakkındaki Türkçe Twitter verileri kullanılarak Terim Frekansı (TF) ağırlıklandırma yöntemi veri kümesi oluşturulmuştur. Oluşturulan bu veri kümesi üzerinde Weka yazılımı kullanılarak deneyler yapılmıştır. Bu deneylerde veri kümesinin %1, %5, %10, %20, %30, %40 ve %50'si eğitim verisi olarak ayrılmıştır. Bu eğitim verisi dışında kalan verilerin etiketleri silinmiş ve etiketsiz verilerin etiketlerinin bulunması için deneyler yapılmıştır. Bu deneylerde sırasıyla;

- TF (Terim Frekansı),
- TF+Kİ (Terim Frekansı + Kullanıcı İstatistiği),
- TF+KKO (Terim Frekansı + Kelime Kullanım Oranı),
- TF+Kİ+KKO (Terim Frekansı + Kullanıcı İstatistiği + Kelime Kullanım Oranı),

için yapılmıştır.

Terim Frekansı (TF) ağırlıklandırma yöntemi ile yapılan deneylerde en iyi sonuç TF+Kİ (Terim Frekansı + Kullanıcı İstatistiği) için %50'lik eğitim kümesi ile yapılan deneylerde %65,893 ile elde edilmiştir. KKO (Kelime Kullanım Oranı) bilgisinin eklenmesi deneylere olumsuz yönde etki etmiştir. Terim Frekansı (TF) ağırlıklandırma yöntemi ile yapılan deney sonuçları Şekil 4.14'de, detayları ise Tablo 4.14'de gösterilmiştir.



	1%	5%	10%	20%	30%	40%	50%
TF	46.062%	50.617%	53.573%	55.470%	55.533%	56.563%	57.334%
TF+Kİ	63.520%	62.583%	62.595%	65.056%	65.531%	65.843%	65.893%
TF+KKO	46.361%	52.232%	52.442%	54.004%	54.815%	56.219%	56.999%
TF+Kİ+KKO	57.675%	53.154%	53.287%	57.403%	58.406%	59.463%	60.016%

Şekil 4.14 Yarı-Denetimli öğrenme Terim Frekansı ağırlıklandırma doğrulama sonuçları

Tablo 4.14 Yarı-Denetimli öğrenme Terim Frekansı ağırlıklandırma doğrulama sonuçları

TF (TERİM FREKANSI)							
	TİP		**TP**	**FP**	**KESİNLİK**	**DUYARLILIK**	**F-ÖLÇÜTÜ**
YARI-DENETİMLİ	**TF + KKO + Kİ**	**1%**	0,577	0,307	0,553	0,577	0,543
		5%	0,532	0,308	0,528	0,532	0,526
		10%	0,533	0,32	0,519	0,533	0,519
		20%	0,574	0,29	0,559	0,574	0,562
		30%	0,584	0,285	0,571	0,584	0,571
		40%	0,595	0,285	0,585	0,595	0,580
		50%	0,600	0,277	0,589	0,600	0,587
	TF + KKO	**1%**	0,464	0,443	0,449	0,464	0,339
		5%	0,522	0,338	0,496	0,522	0,497
		10%	0,524	0,333	0,502	0,524	0,504
		20%	0,54	0,338	0,53	0,54	0,515
		30%	0,548	0,325	0,532	0,548	0,527
		40%	0,562	0,316	0,550	0,562	0,541
		50%	0,570	0,315	0,562	0,570	0,548
	TF + Kİ	**1%**	0,635	0,267	0,616	0,635	0,604
		5%	0,626	0,244	0,62	0,626	0,619
		10%	0,626	0,254	0,618	0,626	0,615
		20%	0,651	0,241	0,641	0,651	0,639
		30%	0,655	0,239	0,647	0,655	0,643
		40%	0,658	0,242	0,653	0,658	0,645
		50%	0,659	0,241	0,651	0,659	0,644
	TF	**1%**	0,461	0,447	0,439	0,461	0,329
		5%	0,506	0,396	0,531	0,506	0,429
		10%	0,536	0,358	0,514	0,536	0,487
		20%	0,555	0,343	0,557	0,555	0,516
		30%	0,555	0,333	0,537	0,555	0,521
		40%	0,566	0,330	0,569	0,566	0,531
		50%	0,573	0,323	0,575	0,573	0,541

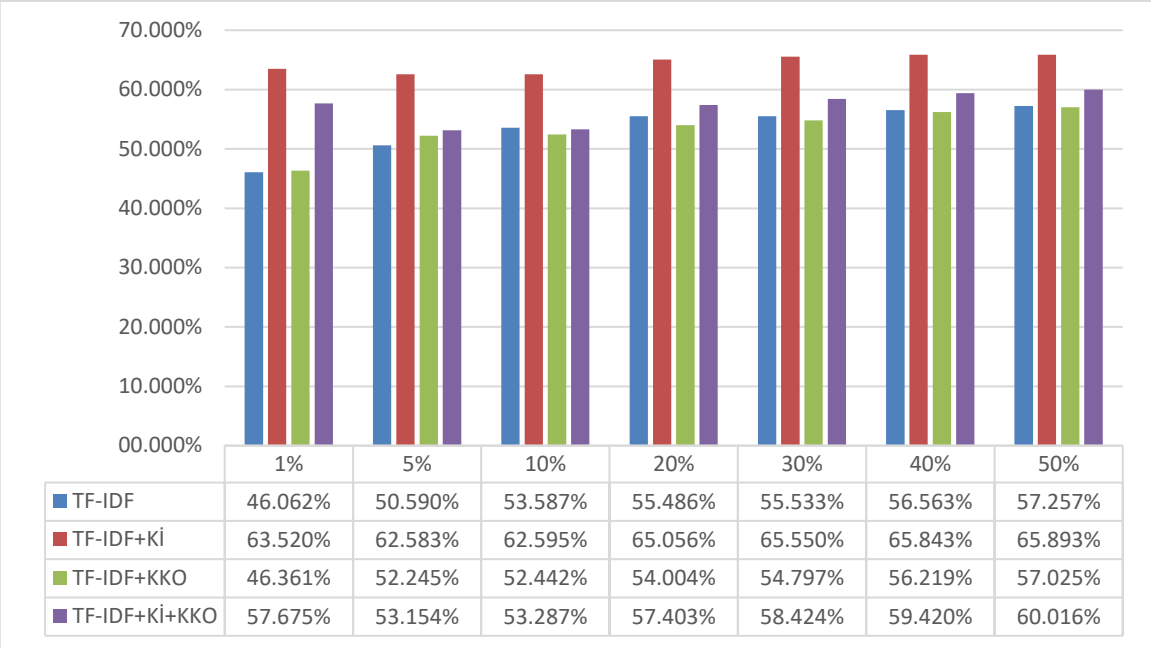
4.1.2.3 Terim Frekansı-Ters Doküman Frekansı (TF-IDF) Veri Kümesi (Terim Frekansı – Ters Doküman Frekansı)

Bankacılık hakkındaki Türkçe Twitter verileri kullanılarak TF-IDF (Terim Frekansı-Ters Doküman Frekansı) ağırlıklandırma yöntemi veri kümesi oluşturulmuştur. Weka yazılımı kullanılarak, oluşturulan bu veri kümesi ile deneyler yapılmıştır. Bu deneylerde veri kümesinin %1, %5, %10, %20, %30, %40 ve %50'si eğitim verisi olarak ayrılmıştır. Bu eğitim verisi dışında kalan verilerin etiketleri silinmiş ve etiketsiz verilerin etiketlerinin bulunması için deneyler yapılmıştır. Bu deneylerde;

- TF- IDF (Terim Frekansı-Ters Doküman Frekansı),
- TF-IDF+Kİ (Terim Frekansı- Ters Doküman Frekansı + Kullanıcı İstatistiği),
- TF-IDF+KKO (Terim Frekansı- Ters Doküman Frekansı + Kelime Kullanım Oranı),
- TF-IDF+Kİ+KKO (Terim Frekansı- Ters Doküman Frekansı + Kullanıcı İstatistiği + Kelime Kullanım Oranı),

özellikleri kullanılmıştır.

TF-IDF ağırlıklandırma yöntemi ile yapılan deneylerde en iyi sonuç TF-IDF+Kİ (Terim Frekansı- Ters Doküman Frekansı + Kullanıcı İstatistiği) için %50'lik eğitim kümesi ile yapılan deneylerde %65,893 ile elde edilmiştir. KKO (Kelime Kullanım Oranı) bilgisinin eklenmesi deneylere olumsuz yönde etki etmiştir. TF-IDF (Terim Frekansı-Ters Doküman Frekansı) ağırlıklandırma yöntemi ile yapılan deney sonuçları Şekil 4.15'de, detayları ise Tablo 4.15'de gösterilmiştir.



	1%	5%	10%	20%	30%	40%	50%
TF-IDF	46.062%	50.590%	53.587%	55.486%	55.533%	56.563%	57.257%
TF-IDF+Kİ	63.520%	62.583%	62.595%	65.056%	65.550%	65.843%	65.893%
TF-IDF+KKO	46.361%	52.245%	52.442%	54.004%	54.797%	56.219%	57.025%
TF-IDF+Kİ+KKO	57.675%	53.154%	53.287%	57.403%	58.424%	59.420%	60.016%

Şekil 4.15 Yarı-Denetimli öğrenme Terim Frekansı-Ters Doküman Frekansı ağırlıklandırma doğrulama sonuçları

Tablo 4.15 Yarı-Denetimli öğrenme Terim Frekansı-Ters Doküman Frekansı ağırlıklandırma doğrulama sonuçları

TF-IDF (TERİM FREKANSI-TERS DOKÜMAN FREKANSI)							
	TİP		TP	FP	KESİNLİK	DUYARLILIK	F-ÖLÇÜTÜ
YARI-DENETİMLİ	TF-IDF + KKO + Kİ	1%	0,577	0,307	0,553	0,577	0,543
		5%	0,532	0,308	0,528	0,532	0,526
		10%	0,533	0,320	0,519	0,533	0,519
		20%	0,574	0,290	0,559	0,574	0,562
		30%	0,584	0,284	0,571	0,584	0,572
		40%	0,594	0,285	0,585	0,594	0,580
		50%	0,600	0,277	0,589	0,600	0,587
	TF-IDF + KKO	1%	0,464	0,443	0,449	0,464	0,339
		5%	0,522	0,338	0,497	0,522	0,497
		10%	0,524	0,333	0,502	0,524	0,504
		20%	0,540	0,338	0,530	0,540	0,515
		30%	0,548	0,326	0,532	0,548	0,527
		40%	0,562	0,316	0,550	0,562	0,541
		50%	0,570	0,314	0,563	0,570	0,548
	TF-IDF + Kİ	1%	0,635	0,267	0,616	0,635	0,604
		5%	0,626	0,244	0,620	0,626	0,619
		10%	0,626	0,254	0,618	0,626	0,615
		20%	0,651	0,241	0,641	0,651	0,639
		30%	0,655	0,239	0,647	0,655	0,643
		40%	0,658	0,242	0,653	0,658	0,645
		50%	0,659	0,241	0,651	0,659	0,644
	TF-IDF	1%	0,461	0,447	0,439	0,461	0,329
		5%	0,506	0,396	0,531	0,506	0,428
		10%	0,536	0,358	0,514	0,536	0,487
		20%	0,555	0,343	0,557	0,555	0,516
		30%	0,555	0,333	0,537	0,555	0,521
		40%	0,566	0,330	0,572	0,566	0,530
		50%	0,573	0,324	0,574	0,573	0,540

4.2. Değerlendirmeler

Bu tez kapsamında metin sınıflandırma (MS), doğal dil işleme (DDİ) ve duygu analizi (DA) teknikleri kullanılarak sosyal medyada otomatik duygu analizi üzerinde çalışılmıştır. Bu amaçla bu alanda yapılan araştırmalar incelenmiş ve duygu analizi (DA) alanında Türkçe ile ilgili yeterli çalışmanın bulunmadığı gözlemlenmiştir. Bu nedenle bu eksiklik göz önüne alınarak sosyal medyadan elde edilen Türkçe Twitter verileri ile duygu analizi (DA) çalışması yapılmıştır.

Duygu analizi (DA) çalışmasına başlamadan önce deneylerde kullanılacak veri kümelerinin oluşturulması çalışması yapılmıştır. Geliştirilen yazılım ile Twitter sosyal ağından veriler toplanmıştır. Toplanan bu verilerin etiketlenmesi Doğuş Üniversitesi öğrencileri tarafından yapılmıştır. Özellik boyutunu azaltmak ve çalışmanın başarısını arttırmak için etiketlenen veriler üzerinde ön işleme çalışması yapılmıştır. Bu ön işleme adımlarından sonra deneylerde kullanılmak üzere kelime kullanım oranları ve kullanıcı istatistik bilgileri hesaplanmıştır. Veri kümeleri oluşturulurken İkili (Binary/Boolean), Terim Frekansı (TF) ve Terim Frekansı-Ters Doküman Frekansı (TF-IDF) ağırlıklandırma yöntemleri kullanılmıştır. Oluşturulan veri kümeleri ile yapılan deneyler, makine öğrenmesi (MÖ) yöntemlerinden denetimli ve yarı-denetimli öğrenme yöntemleri için yapılmış ve yeni eklenen özelliklerin deney sonuçlarına etkisi incelenmiştir.

Tez kapsamında denetimli öğrenme yöntemi için oluşturulan dengeli ve dengesiz veri kümesi kullanılarak 10-Kat Çapraz Doğrulama ve %80 bölümlendirme ile Naïve Bayes (NB), Sequential Minimal Optimization (SMO), k-En Yakın Komşu (KNN) , Rastgele Orman (RF) ve Karar Ağacı (J48) algoritmaları kullanılarak deneyler yapılmıştır.

Dengeli veri kümesi ile yapılan deneylerde en iyi sonuç 10-kat çapraz doğrulama yöntemi ve Karar Ağacı (J48) ağırlıklandırma algoritması ile yapılan deneyde, tüm özelliklerin eklendiği İkili+Kİ+KKO (İkili + Kullanıcı İstatistiği + Kelime Kullanım Oranı) durumunda %80,8201 ile elde edilmiştir.

Dengesiz veri kümesi ile yapılan deneylerde ise en iyi sonuç %80 bölümlendirme yöntemi ve Karar Ağacı (J48) ağırlıklandırma algoritması ile yapılan deneyde, tüm özelliklerin eklendiği İkili+Kİ+KKO (İkili + Kullanıcı İstatistiği + Kelime Kullanım Oranı) durumunda %71,6981 ile elde edilmiştir.

Denetimli öğrenme yöntemi ile yapılan deneylerde, veri kümesinin dengeli olması, Kullanıcı İstatistiği (Kİ) ve Kelime Kullanım Oranı (KKO) özelliklerinin sonuca olumlu yönde etki ettiği gözlemlenmiştir.

Yarı-denetimli öğrenme yöntemi ile yapılan deneylerde veri kümesinin %1, %5, %10, %20, %30, %40 ve %50'si eğitim verisi olarak ayrılmıştır. Eğitim verisi dışında kalan etiketsiz kısımın etiketleri silinmiş ve etiketlerinin bulunması için deneyler yapılmıştır. Yapılan bu deneylerde en iyi sonuç, %50 eğitim kümesi ve ikili ağırlıklandırma yöntemi kullanılarak yapılan deneyde İkili+Kİ (İkili + Kullanıcı İstatistiği) özelliklerinin eklendiği durumda %72,055 olarak bulunmuştur.

Yarı-denetimli öğrenme yöntemi için yapılan deneylerde hesaplanan Kullanıcı İstatistiği (Kİ) özelliğinin deney sonuçlarına olumlu yönde etki ettiği gözlemlenmiştir.

4.3. Sonuç

Bu çalışmalar sonucu, yeni eklenen Kullanıcı İstatistiği (Kİ) ve Kelime Kullanım Oranı (KKO) özelliklerinin deney sonuçlarına olumlu olarak etkilediği gözlemlenmiştir. İleriki çalışmalarda, ilk olarak daha büyük bir veri kümesi oluşturmak hedeflenmektedir. Bu veri kümesi dışında mesaj ve kullanıcı bilgilerine göre yeni özellikler hesaplanarak bu özelliklerin sonuçlara etkisinin incelenmesi düşünülmektedir.

KAYNAKLAR

[1] Akbaş, E., (2012), "Aspect Based Opinion Mining on Turkish Tweets", *Bilkent University Doctoral Dissertation*, July, Ankara, Turkey.

[2] Akın, A. A., and Akın, M. D., (2007), "Zemberek, an Open Source NLP Framework for Turkic Languages", *Structure, 10*, 1-5.

[3] Amasyalı, M.F., (2013), "Active Learning for Turkish Sentiment Analysis", *In Innovations in Intelligent Systems and Applications (INISTA), 2013 IEEE International Symposium on IEEE*, June 19-21, Albena, Bulgaria, p. 1-4.

[4] Bian, J., Topaloglu, U., and Yu, F., (2012), "Predicting Collective Sentiment Twitter Mining for Drug-Related Adverse Events", *Proceedings of the 2012 International Workshop on Smart Health and Wellbeing*, October 29-November 2, Hawaii, USA.

[5] Breiman, L., (2001), "Random Forests", *Machine learning*, *45*(1), 5-32.

[6] Claster, W.B., Dinh, H., and Cooper, M., (2010), "Naïve Bayes and Unsupervised Artificial Neural Nets for Cancun Tourism Social Media Data Analysis", *In Nature and Biologically Inspired Computing (NaBIC), 2010 Second World Congress on IEEE*, December, Kitakyushu, Japan, pp. 158-163.

[7] Çakmak, O., Kazemzadeh, A., Can, D., Yıldırım, S., and Narayanan, S., (2012), "Root-Word Analysis of Turkish Emotional Language", *Corpora for Research on Emotion Sentiment & Social Signals*, May 26, Istanbul, Turkey.

[8] Çetin, M., and Amasyali, M.F., (2013), "Supervised and Traditional Term Weighting Methods for Sentiment Analysis", *In Signal Processing and Communications Applications Conference (SIU), 2013 21st IEEE*, pp. 1-4.

[9] Dai, W., Xue, G.R., Yang, Q., and Yu,Y., (2007), "Transferring Naive Bayes Classifiers for Text Classification", *In AAAI* , July, Vancouver, Canada, Vol. 7, pp. 540-545.

[10] Eroğul U., (2009), "Sentiment Analysis in Turkish"

[11] Go, A., Bhayani, R., and Huang, L., (2009), "Twitter Sentiment Classification Using Distant Supervision", *CS224N Project Report, Stanford, 1*(12).

[12] Hu, M., and Liu, B., (2004), "Mining and Summarizing Customer Reviews", *In Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, August, Wahington, USA, pp. 168-177.

[13] Kargupta, H., Han, J., Philip, S.Y., Motwani, R., and Kumar, V., (2008), *Next Generation of Data Mining*, CRC Press.

[14] Liao, C., Alpha, S., and Dixon, P., (2001), *Feature Preparation in Text Categorization*, Oracle Text Selected Papers and Presentations.

[15] Liu, M., and Yang, J., (2012), *An Improvement of TFIDF Weighting in Text Categorization*, In IPCSIT, 44-47, IACSIT Presss.

[16] Meral, M., and Diri, B., (2014), "Sentiment Analysis on Twitter", *In Signal Processing and Communications Applications Conference (SIU), April 23-25, Trabzon, Turkey*, pp. 690-693.

[17] Miller, G.A., Beckwith, R., Fellbaum, C., Gross, D., and Miller, K.J., (1990), *Introduction to WordNet: An on-line Lexical Database*, International Journal of Lexicography, Oxford University Press, 1990, *3*(4), 235-244.

[18] Mishne, G., and Glance, N.S., (2006), "Predicting Movie Sales from Blogger Sentiment", In *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*, March, California, USA, pp. 155-158.

[19] Nasukawa, T., and Yi, J., (2003), "Sentiment Analysis: Capturing Favorability Using Natural Language Processing", *In Proceedings of the 2nd International Conference on Knowledge Capture*, October, Florida, USA, pp. 70-77.

[20] Nguyen, L. T., Wu, P., Chan, W., Peng, W., and Zhang, Y., (2012), "Predicting Collective Sentiment Dynamics from Time-Series Social Media", *In Proceedings of the First International Workshop on Issues of Sentiment Discovery and Opinion Mining, August*, Beijing, China, p. 6.

[21] Uğuz, H., (2011), "A Two-Stage Feature Selection Method for Text Categorization by Using Information Gain, Principal Component Analysis and Genetic Algorithm", *Knowledge-Based Systems*, *24*(7), 1024-1032.

[22] Pang, B., Lee, L., and Vaithyanathan, S., (2002), "Thumbs up?: Sentiment Classification Using Machine Learning Techniques", *In Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing-Volume 10*, July, USA, pp. 79-86.

[23] Pang, B., and Lee, L., (2004), "A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts", *In Proceedings of the 42nd annual meeting on Association for Computational Linguistics*, July, Barcelona, Spain, p. 271.

[24] Platt, J, (1998), "Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines".

[25] Szomszor, M.N., Kostkova, P., and De Quincey, E., (2010), "# swineflu: Twitter Predicts Swine Flu Outbreak in 2009", *In 3rd International ICST Conference on Electronic Healthcare for the 21st Century,* December, Casablanca, Morocco, pp. 18-26.

[26] Şimşek, M.U., and Özdemir, S., (2012), "Analysis of the Relation Between Turkish Twitter Messages and Stock Market Index", *In Application of Information and Communication Technologies (AICT), 2012 6th International Conference on IEEE*, October 17-19, Tbilisi, Georgia, pp. 1-4.

[27] Thelwall, M., Buckley, K., Paltoglou, G., Cai, D., and Kappas, A., (2010), "Sentiment Strength Detection in Short Informal Text", *Journal of the American Society for Information Science and Technology*, *61*(12), 2544-2558.

[28] Vural, A.G., Cambazoglu, B.B., Senkul, P., and Tokgoz, Z.O., (2013), "A Framework for Sentiment Analysis in Turkish: Application to Polarity Detection of Movie Reviews in Turkish", *In Computer and Information Sciences III*, Springer London, pp. 437-445.

[29] Witten, I.H., Frank, E., Hall, M.A., and Pal, C.J., (2016), *Data Mining: Practical machine learning tools and techniques*, Morgan Kaufmann.

ÖZGEÇMİŞ

Kişisel Bilgiler :

Adı Soyadı	Cem GÜMÜŞ
Doğum Yeri	İstanbul
Doğum Tarihi	08.02.1978
Eposta	cemgumus@gmail.com

Eğitim :

Lise	1993 - 1995	İstanbul Üsküdar Burhan Felek Lisesi
Lisans	1996 - 2000	İstanbul Üniversitesi Bilgisayar Bilimleri Mühendisliği Bölümü
Yüksek Lisans	2010-	Doğuş Üniversitesi Bilgisayar Mühendisliği Bölümü

İş Tecrübesi :

2001 - 2002	SIT Bilişim Teknolojileri Yazılım Geliştirme
2004 -	TÜBİTAK – BİLGEM - UEKAE Başuzman Araştırmacı

Yayınlar :

Akyokuş, S., Ganiz, M.C., and Gümüş, C., (2017), "Improving Twitter Sentiment Classification Using Term Usage And User Based Attributes", IMMM, June 25-29, Venice, Italy