

T.C.
MARMARA ÜNİVERSİTESİ
SOSYAL BİLİMLER ENSTİTÜSÜ
İŞLETME ANABİLİM DALI
SAYISAL YÖNTEMLER BİLİM DALI

**METİN MADENCİLİĞİ YÖNTEMLERİ İLE E-TİCARET
MARKALARINA YÖNELİK SOSYAL MEDYA YORUMLARININ
ANALİZİ**

Yüksek Lisans Tezi

NURFER IŞIK

İstanbul, 2019

T.C.
MARMARA ÜNİVERSİTESİ
SOSYAL BİLİMLER ENSTİTÜSÜ
İŞLETME ANABİLİM DALI
SAYISAL YÖNTEMLER BİLİM DALI

**METİN MADENCİLİĞİ YÖNTEMLERİ İLE E-TİCARET
MARKALARINA YÖNELİK SOSYAL MEDYA YORUMLARININ
ANALİZİ**

Yüksek Lisans Tezi

NURFER IŞIK

Danışman: Doç. Dr. Özgür ÇAKIR

İstanbul, 2019



T.C.
MARMARA ÜNİVERSİTESİ
SOSYAL BİLİMLER ENSTİTÜSÜ MÜDÜRLÜĞÜ

TEZ ONAY BELGESİ

İŞLETME Anabilim Dalı SAYISAL YÖNTEMLER Bilim Dalı TEZLİ YÜKSEK LİSANS öğrencisi NURFER IŞIK'ın METİN MADENCİLİĞİ YÖNTEMLERİ İLE E - TİCARET MARKALARINA YÖNELİK SOSYAL MEDYA YORUMLARININ ANALİZİ adlı tez çalışması, Enstitümüz Yönetim Kurulunun 19.07.2019 tarih ve 2019-22/7 sayılı kararıyla oluşturulan jüri tarafından oy birliği ile kabul edilmiştir.

Tez Savunma Tarihi ...05.08.2019...

Öğretim Üyesi Adı Soyadı

			İmzası
1.	Tez Danışmanı	Doç. Dr. ÖZGÜR ÇAKIR	
2.	Jüri Üyesi	Prof. Dr. HAKAN YILDIRIM	
3.	Jüri Üyesi	Prof. Dr. UMMAN TUĞBA GÜRSOY	

ÖZET

METİN MADENCİLİĞİ YÖNTEMLERİ İLE E-TİCARET MARKALARINA YÖNELİK SOSYAL MEDYA YORUMLARININ ANALİZİ

Metin verilerinden anlam çıkarılması adına yapılan analizler teknolojik gelişmeler ile hızla değişebilmektedir. Müşterilerin satın aldıkları ürün/hizmetler hakkında yazdıkları geri bildirimlerin ve sosyal medya platformlarında yazdıkları mesajların/yorumların içerdiği duygunun araştırılabilir ve yorumlanabilir oluşu bu metin verilerine anlam katmaktadır. Bu metin verilerinin analiz edilmesi ile elde edilen bilgileri işletmelerin kullanması ise işletmelere değer katmaktadır. Bu tez çalışmasında, makine öğrenmesi tekniklerinden denetimli öğrenme yaklaşımı kullanılarak sosyal medya yorumlarının duygu analizi yapılmıştır. Denetimli öğrenme sınıflandırma algoritmalarından Naive Bayes, Sıralı Minimal Optimizasyon(SMO), k-en yakın komşu (kNN=IBk) algoritmaları kullanılmıştır. Bazı e-ticaret firmalarına, ürünlerine/hizmetlerine yönelik yapılan yorumlardan oluşturulan veri kümesi Twitter platformu kullanılarak elde edilmiştir. Sosyal medya yorumları olumlu, olumsuz, nötr olarak el yordamı ile etiketlenerek üç sınıfta toplanmıştır. Bu çalışmada ‘sınıflardaki veri dağılımının’ ve ‘öznitelik seçiminin’ sınıflandırma üzerindeki etkileri incelenmiştir. Bu incelemeler Weka 3.8 yazılımında yer alan Naive Bayes (NB), Sıralı Minimal Optimizasyon (SMO) ve 1-en yakın komşu (IB1) sınıflandırma algoritmaları kullanılarak ve 16 farklı model oluşturularak yapılmıştır. Elde edilen deneysel sonuçlarda dengesiz veri kümesinin, dengeli veri kümesine göre daha iyi performans sağladığı gözlemlenmiştir. Ayrıca veri kümelerinde öznitelik seçimi yapıldığı durumlarda da veri kümelerinin daha iyi performans sağladığı gözlemlenmiştir. En iyi performansı gösteren sınıflandırma algoritması ise dengesiz veri kümesi üzerinde öznitelik seçimi yapıldığında ortalama %93,52 sınıflandırma doğruluğu ile kNN olmuştur.

Anahtar Kelimeler: Metin Madenciliği, Duygu analizi, Metin Sınıflandırma, Öznitelik Seçimi, Sosyal Medya Analizi, Naive Bayes, Sıralı Minimal Optimizasyon, k-En Yakın Komşu

ABSTRACT

ANALYSIS OF SOCIAL MEDIA COMMENTS AN E-COMMERCE BRANDS WITH TEXT MINING METHODS

The analysis for extracting meaning from text data can change rapidly with technological developments. The customers' feedback about the products/services that purchase and the messages/comments that write on social media platforms are searchable and interpretable. This situation adds meaning to the text data. The use of the information obtained by the analysis of this text data adds value to the enterprises. In this thesis, a sentiment analysis of social media comments is performed by using supervised learning approach from machine learning techniques. Naive Bayes, Sequential Minimal Optimization (SMO), k-nearest neighbor (kNN = IBk) algorithms are used in the supervised learning classification algorithms. The data set created from the comments made for some e-commerce companies, their products / services is obtained by using Twitter platform. Social media interpretations are gathered into three groups, labeled manually as positive, negative and neutral. In this study, the effect of 'data distribution in groups' and 'attribute selection' on the success results of Naive Bayes (NB), Sequential Minimal Optimization (SMO) and 1-nearest neighbor (IB1) classification algorithms in Weka 3.8 software are examined by creating 16 different models. It is observed that the unbalanced data set provided better performance than the balanced data set. In addition, it is observed that data sets perform better when attribute selection is made in data sets. The best performing classification algorithm is kNN with 93,52% classification accuracy rate when the attribute selection is made on unbalanced data set.

Key Words: Text Mining, Sentiment Analysis, Text Classification, Feature Selection, Social Media Analysis, Naive Bayes, Sequential Minimal Optimization, k-Nearest Neighbor

ÖNSÖZ

Yüksek lisans tez sürecimde metin madenciliği alanında çalışma olanağı sağlayan, bu sürecin tamamında değerli bilgilerini benimle paylaşan, her konuda destek olan, yol gösteren ve beni yönlendiren değerli hocam Doç. Dr. Özgür Çakır'a, tüm katkılarından dolayı teşekkür ederim.

Yüksek lisans dahil olmak üzere tüm eğitim-öğretim hayatımda desteklerini benden esirgemeyen annem Ayfer Işık'a, babam Köksal Işık'a, kardeşlerim Nursal Işık ve Muammer Cemil Işık'a ve sevgili eşim Enis Altıok'a destekleri için teşekkür ederim.

Sevgili arkadaşlarım Mert Aliciklioğlu ve Fuat Kestane'ye çalışmanın hazırlanması sırasındaki yardımlarından dolayı teşekkür eder, bu tez çalışmasının konuyla ilgilenen herkese faydalı olmasını dilerim.

NURFER IŞIK

İÇİNDEKİLER

SAYFA NO

ÖZETiii
ABSTRACT	ii
ÖNSÖZ	iii
İÇİNDEKİLER	iii
TABLolar	vii
ŞEKİLLER	ix
GİRİŞ	1
1. METİN MADENCİLİĞİ, SÜRECİ VE YÖNTEMLERİ	2
1.1. METİN MADENCİLİĞİ	2
1.2. DUYGU ANALİZİ	5
1.2.1. Duygu Analizi Süreci	6
1.2.1.1. İşi Anlama.....	6
1.2.1.2. Veriyi Anlama.....	7
1.2.1.3. Veri Hazırlama	8
1.2.1.4. Duygu Analizi - Modelleme	11
1.2.1.5. Değerlendirme.....	12
1.2.1.6. Uygulama (Konuşlandırma)	12
1.2.2. Duygu Sınıflandırma Yöntemleri.....	12
1.2.2.1. Naive Bayes	14
1.2.2.2. Sıralı Minimal Optimizasyon.....	15
1.2.2.3. K-En Yakın Komşu (kNN).....	17
1.2.3. Model Başarısını Değerlendirme	20
1.2.3.1. Karşılıklı Matrisi	20
1.2.3.2. Değerlendirme Ölçütleri.....	21
1.2.3.3. Model Geçerliliğini Doğrulama	24
1.3. YAPILAN ÇALIŞMALAR	25
1.4. DESTEK ARAÇLARI	28
1.4.1. Açık Kaynak Kodlu Araçlar	28
1.4.2. Doğal Dil İşleme Araçları.....	30

2. E-TİCARET MARKALARINA YÖNELİK SOSYAL MEDYA YORUMLARININ ANALİZİ	32
2.1. AMAÇ VE KAPSAM	32
2.2. VERİ KÜMESİ	33
2.2.1. Başlangıç Verileri	33
2.2.2. Verilerin Hazırlanması	34
2.2.2.1. Veri Etiketleme	34
2.2.2.2. Veri Önışleme	35
2.2.2.3. Veri Kümesinin Seçilmesi	38
2.2.2.4. Veri Kümesini Özniteliklerine Ayırma (Tokenization)	38
2.2.2.5. Öznitelik Seçimi	40
2.3. MODELLEME	41
2.3.1. Dengesiz Veri Kümesi ile Analiz	42
2.3.1.1. Dengesiz Veri Kümesi - Öznitelik Seçiminin Yapılmadığı Modeller	42
2.3.1.2. Dengesiz Veri Kümesi - Öznitelik Seçiminin Yapıldığı Modeller	46
2.3.2. Dengeli Veri Kümesi ile Analiz	50
2.3.2.1. Dengeli Veri Kümesi - Öznitelik Seçiminin Yapılmadığı Modeller	50
2.3.2.2. Dengeli Veri Kümesi - Öznitelik Seçiminin Yapıldığı Modeller	55
2.4. BULGULAR ve TARTIŞMA	60
SONUÇ	63
EKLER	65
EK 1: Dengesiz Veri Kümesi Üzerinde Öznitelik Seçimi Yapılmadan Naive Bayes Sınıflandırma Sonucunun WEKA Ekran Görüntüsü	65
EK 2: Dengesiz Veri Kümesi Üzerinde Öznitelik Seçimi Yapılmadan SMO Sınıflandırma Sonucunun WEKA Ekran Görüntüsü	65
EK 3: Dengesiz Veri Kümesi Üzerinde Öznitelik Seçimi Yapılmadan kNN(k=1) Chebyshev Uzaklık Ölçütü ile Sınıflandırma Sonucunun WEKA Ekran Görüntüsü	66
EK 4: Dengesiz Veri Kümesi Üzerinde Öznitelik Seçimi Yapılmadan kNN(k=1) Öklid Uzaklık Ölçütü ile Sınıflandırma Sonucunun WEKA Ekran Görüntüsü	66
EK 5: Dengesiz Veri Kümesi Üzerinde Öznitelik Seçimi Yapılarak Naive Bayes Sınıflandırma Sonucunun WEKA Ekran Görüntüsü	67
EK 6: Dengesiz Veri Kümesi Üzerinde Öznitelik Seçimi Yapılarak SMO Sınıflandırma Sonucunun WEKA Ekran Görüntüsü	67
EK 7: Dengesiz Veri Kümesi Üzerinde Öznitelik Seçimi Yapılarak kNN(k=1) Chebyshev Uzaklık Ölçütü ile Sınıflandırma Sonucunun WEKA Ekran Görüntüsü	68

EK 8: Dengesiz Veri Kümesi Üzerinde Öznitelik Seçimi Yapılarak kNN(k=1) Öklid Uzaklık Ölçütü ile Sınıflandırma Sonucunun WEKA Ekran Görüntüsü	68
EK 9: Dengeli Veri Kümesi Üzerinde Öznitelik Seçimi Yapılmadan Naive Bayes Sınıflandırma Sonucunun WEKA Ekran Görüntüsü	69
EK 10: Dengeli Veri Kümesi Üzerinde Öznitelik Seçimi Yapılmadan SMO Sınıflandırma Sonucunun WEKA Ekran Görüntüsü	69
EK 11: Dengeli Veri Kümesi Üzerinde Öznitelik Seçimi Yapılmadan kNN(k=1) Chebyshev Uzaklık Ölçütü ile Sınıflandırma Sonucunun WEKA Ekran Görüntüsü	70
EK 12: Dengeli Veri Kümesi Üzerinde Öznitelik Seçimi Yapılmadan kNN(k=1) Öklid Uzaklık Ölçütü ile Sınıflandırma Sonucunun WEKA Ekran Görüntüsü	70
EK 13: Dengeli Veri Kümesi Üzerinde Öznitelik Seçimi Yapılarak Naive Bayes Sınıflandırma Sonucunun WEKA Ekran Görüntüsü	71
EK 14: Dengeli Veri Kümesi Üzerinde Öznitelik Seçimi Yapılarak SMO Sınıflandırma Sonucunun WEKA Ekran Görüntüsü	71
EK 15: Dengeli Veri Kümesi Üzerinde Öznitelik Seçimi Yapılarak kNN(k=1) Chebyshev Uzaklık Ölçütü ile Sınıflandırma Sonucunun WEKA Ekran Görüntüsü	72
EK 16: Dengeli Veri Kümesi Üzerinde Öznitelik Seçimi Yapılarak kNN(k=1) Öklid Uzaklık Ölçütü ile Sınıflandırma Sonucunun WEKA Ekran Görüntüsü	72
KAYNAKÇA	73

TABLolar

Tablo 1: Karşıtlık Matrisi	20
Tablo 2: Kappa İstatistik Deęeri Yorumları	23
Tablo 3: ITU Turkish NLP Web Service kullanılarak düzenlenmiş örnek sosyal medya yorumları.....	35
Tablo 4: Manuel olarak düzenlenen örnek kelimeler	36
Tablo 5: Türkçe Durak Kelimeler	37
Tablo 6: Model 1.1.1. Hata Matrisi	42
Tablo 7: Model 1.1.1. Deęerlendirme Ölçütleri.....	43
Tablo 8: Model 1.1.2. Hata Matrisi	43
Tablo 9: Model 1.1.2. Deęerlendirme Ölçütleri.....	44
Tablo 10:Model 1.1.3.1. Hata Matrisi	44
Tablo 11:Model 1.1.3.1. Deęerlendirme Ölçütleri.....	45
Tablo 12:Model 1.1.3.2. Hata Matrisi	45
Tablo 13:Model 1.1.3.2. Deęerlendirme Ölçütleri.....	46
Tablo 14:Model 1.2.1. Hata Matrisi	47
Tablo 15:Model 1.2.1. Deęerlendirme Ölçütleri.....	47
Tablo 16:Model 1.2.2. Hata Matrisi	48
Tablo 17:Model 1.2.2. Deęerlendirme Ölçütleri.....	48
Tablo 18:Model 1.2.3.1. Hata Matrisi	49
Tablo 19:Model 1.1.3.1. Deęerlendirme Ölçütleri.....	49
Tablo 20: Model 1.2.3.2. Hata Matrisi	50
Tablo 21:Model 1.2.3.2. Deęerlendirme Ölçütleri.....	50
Tablo 22:Model 2.1.1. Hata Matrisi	51
Tablo 23: Model 2.1.1. Deęerlendirme Ölçütleri.....	52
Tablo 24: Model 2.1.2. Hata Matrisi	52
Tablo 25: Model 2.1.2. Deęerlendirme Ölçütleri.....	53
Tablo 26:Model 2.1.3.1. Hata Matrisi	53
Tablo 27:Model 2.1.3.1. Deęerlendirme Ölçütleri.....	54
Tablo 28:Model 2.1.3.2. Hata Matrisi	54

Tablo 29:Model 2.1.3.2. Deęerlendirme Ölçütleri	55
Tablo 30:Model 2.2.1. Hata Matrisi	56
Tablo 31 : Model 2.2.1. Deęerlendirme Ölçütleri.....	56
Tablo 32 : Model 2.2.2. Hata Matrisi	57
Tablo 33 : Model 2.2.2. Deęerlendirme Ölçütleri.....	57
Tablo 34 : Model 2.2.3.1. Hata Matrisi	58
Tablo 35:Model 2.2.3.1. Deęerlendirme Ölçütleri.....	58
Tablo 36:Model 2.2.3.2. Hata Matrisi	59
Tablo 37:Model 2.2.3.2. Deęerlendirme Ölçütleri.....	59
Tablo 38:Model Sonuçlarının Karşılaştırılması	60

ŞEKİLLER

Şekil 1: Metin madenciliğinin ilişkili olduğu disiplinler	4
Şekil 2 : CRISP-DM: Referans Modelinin Adımları	7
Şekil 3: Duygu Sınıflandırma Yöntemleri	13
Şekil 4: SMO Optimizasyonu	16
Şekil 5: kNN Sınıflandırma	18
Şekil 6: ITU Turkish NLP Web Service kullanılarak metin düzenleme	36
Şekil 7: Veri Kümesini Özniteliklerine Ayırma.....	39
Şekil 8: Dengesiz Veri Kümesinin Özniteliklerine Ayrılması.....	40
Şekil 9:Dengeli Veri Kümesinin Özniteliklerine Ayrılması.....	40
Şekil 10:WEKA Ara Yüz Görüntüsü - Öznitelik Seçimi Yöntemi.....	41
Şekil 11:Dengesiz Veri Kümesi Seçilen Öznitelikler	46
Şekil 12: Dengeli Veri Kümesinin Özniteliklerine Ayrılması.....	51

KISALTMALAR

NLP: Doğal Dil İşleme

SMO: Sıralı Minimal Optimizasyon

DVM: Destek Vektör Makineleri

QP: İkinci Dereceden Programlama (Quadratic Programming)

CRISP-DM: Cross Industry Process for Data Mining

GİRİŞ

Duygu Analizi (Sentiment Analysis), metin madenciliğinin önemli bir alanıdır. Duygu analizi terim olarak, “kişinin bir şey hakkında ne hissettiği”, “bir şeye yönelik tutum” veya “bir fikir” anlamına gelmektedir. Literatürde, duygu durum analizi, fikir madenciliği, duygu sınıflandırma, kanaat çıkarımı gibi farklı adlarla da yer almaktadır. Duygu analizinde temel görev, metinde ifade edilen görüşlerin, cümlenin veya işletme unsuru özelliğinin olumlu, olumsuz veya nötr olup olmadığını sınıflandırmaktır.

Bu tez çalışmasında e-ticaret firmalarına yönelik sosyal medya yorumlarının duygu analizi çalışması yapılmıştır. Sosyal medya platformlarından Twitter seçilmiştir. Twitter’ın seçilmesinin nedeni hem erişim kolaylığı, hem popüler olması, hem de yorum çeşitliliğinin fazla olmasıdır. Veri kümesi olarak e-ticaret firmalarına yönelik yazılan yorumlar kullanılmıştır. Veri kümesi olumlu, olumsuz, nötr olarak üç sınıftan oluşturulmuştur.

Bu tez çalışmasında, sınıflardaki veri dağılımları nasıl olmalıdır, öznitelik seçimi yapılmalı mıdır, veri dağılımlarının ve öznitelik seçiminin makine öğrenmesi sınıflandırma algoritmalarının performanslarına nasıl bir etkisi vardır sorularının cevabı aranmıştır.

Çalışmamız iki bölümden oluşmaktadır. İlk bölümde metin madenciliği tanımlanarak, anlam ve önemi özetlenmiştir. Daha sonrasında duygu analizi başlığı altında , duygu analizi süreci, makine öğrenmesi temelli duygu sınıflandırma yöntemlerinden ve model başarısını değerlendirme ölçütlerinden bahsedilmiştir. Son olarak da duygu analizinde daha önceden yapılan çalışmalardan örnekler verilmiştir. İkinci bölümde ise veri kümeleri, veri kümesinin özellikleri, veri kümesinin analize hazırlanması, veri kümesinin özniteliklerine ayrılması, öznitelik seçimi aşamalarından bahsedilmiştir. Bu aşamalardan sonra veri kümeleri üzerinde modeller oluşturulmuş ve oluşturulan 16 model tanıtılarak her bir modelin sonuçları verilmiştir. Model sonuçları Bulgular ve Tartışma başlığı altında değerlendirilerek ikinci bölüm bitirilmiştir. Sonuç bölümünde ise tez çalışmasının sonuçları değerlendirilmiştir.

1. METİN MADENCİLİĞİ, SÜRECİ VE YÖNTEMLERİ

Bu bölümde ilk olarak metin madenciliği tanımlanarak, anlam ve önemi özetlenmiştir. Daha sonrasında duygu analizi başlığı altında , duygu analizi süreci, duygu sınıflandırma yöntemlerinden ve model başarısını değerlendirme ölçütlerinden bahsedilmiştir. Son olarak, duygu analizi konusunda akademik yazın taraması yapılarak bu alanda yapılmış olan çalışmalar aktarılmış ve metin madenciliği sürecini destekleyici araçlar tanıtılmıştır.

1.1. METİN MADENCİLİĞİ

Metin madenciliği, daha önce bilinmeyen, potansiyel olarak yararlı ve değerli bilgileri, büyük miktarlardaki metin verilerinden çıkarma işlemidir.¹ Metin madenciliği, metin belgelerinin (örneğin bir kütüphanedeki kitapların) kataloglanması ihtiyacı ile geliştirilmeye başlatılmıştır.² Metin madenciliği, metinlerdeki makine destekli bilgi keşfi ile ilgilenmektedir. Metin madenciliği, büyük metin verilerindeki bilgileri keşfetme ve metin verilerindeki ilişkileri otomatik olarak belirleme işlemidir.

Bilgiye Erişim (IR-Information retrieval) ve Bilgi Çıkarımı (IE-Information extraction) metin madenciliğinin iki köküdür.³ Metin madenciliğinin ilk adımı olan bilgiye erişim, bilgi ihtiyacını karşılayan yapısal olarak düzensiz dokümanları/metinleri geniş bir kaynak içerisinde bulmaktır. Bilgi çıkarımı ise büyük veri yığınları içerisinde özet bilgiler elde etmektir. Anahtar kelimeler gibi kullanıcı girişleriyle bağlantılı olan bilgi veya metinlerin bulunması bilgi çıkarımı örnekleridir.

Delen ve Crossland'e⁴ göre, kaliteli bir metin madenciliği uygulaması şunları yapabilir:

- Global veri tabanlarından bilgi alımını iyileştirir.
- Teknik bir alanın teknoloji altyapısını (yazarlar, dergiler, organizasyonlar)

tanımlar.

¹ Feldman, R., J., Sanger. (2007). s.1

² Miner, G., D., Delen, J., Elder, A., Fast, T., Hill, R., Nisbet. (2012). s.3

³ Miner, G., D., Delen, J., Elder, A., Fast, T., Hill, R., Nisbet. (2012). s.4

⁴ Delen, D., M., Crossland. (2008). s.1708

- İlişkili veya farklı teknik literatürlerdeki yeni teknik kavramları ve ilişkileri keşfeder..
- Ana teknik konuları ve alt konuları geniş bir teknik literatürde tanımlar ve sınıflandırır.
- Teknik konuları ve altyapı bileşenleri arasındaki ilişkileri tanımlar..
- Teknik konular ve altyapı bileşenleri arasındaki zamana göre değişen farklılıkları (veya benzerlikleri) tanımlar..
- Yeni araştırma yönleri ve potansiyel etkilerin çıktısını sağlar.

Metin madenciliğinin amacı, yapılandırılmamış enformasyonu (bilgiyi) işleyerek metinden anlamlı sayısal endeksler çıkararak, çeşitli veri madenciliği (istatistiksel ve makine öğrenmesi) algoritmaları tarafından metnin içerdiği enformasyonu erişilebilir kılmaktır.⁵

Metin madenciliğinin birçok faydası bulunmaktadır. Metin madenciliğinin faydaları, ticari işlemlerden ve satış sonrası hizmetlerden çok sayıda metin verisinin toplandığı alanlarda açıkça görülmektedir. Müşterilerin serbest formdaki yorumları, zaman içinde şikayet, övgü, garanti talepleri, hata takibi vb. alanları ile ilgili olabilmektedir. Bu alanların tümü açık bir şekilde işletmelerin ürün geliştirme, pazarlama, müşteri hizmetleri, lojistik bölümlerine girdiler sağlayabilmektedir.

Miller⁶'e göre, metin madenciliği (en azından genel işletme uygulamaları için) aşağıdaki genel faaliyet sınıflarını kapsar:

- Çeşitli metinsel veri hacimlerini organize etmek ve anlamak
- Bilgi çıkarımı yapmak
- Bağlamsal (içeriksel) bilgiyi sayısal bilgiye dönüştürmek için kullanılacak metin ölçülerini geliştiren ve tanımlayan ölçme işlemi yapmak

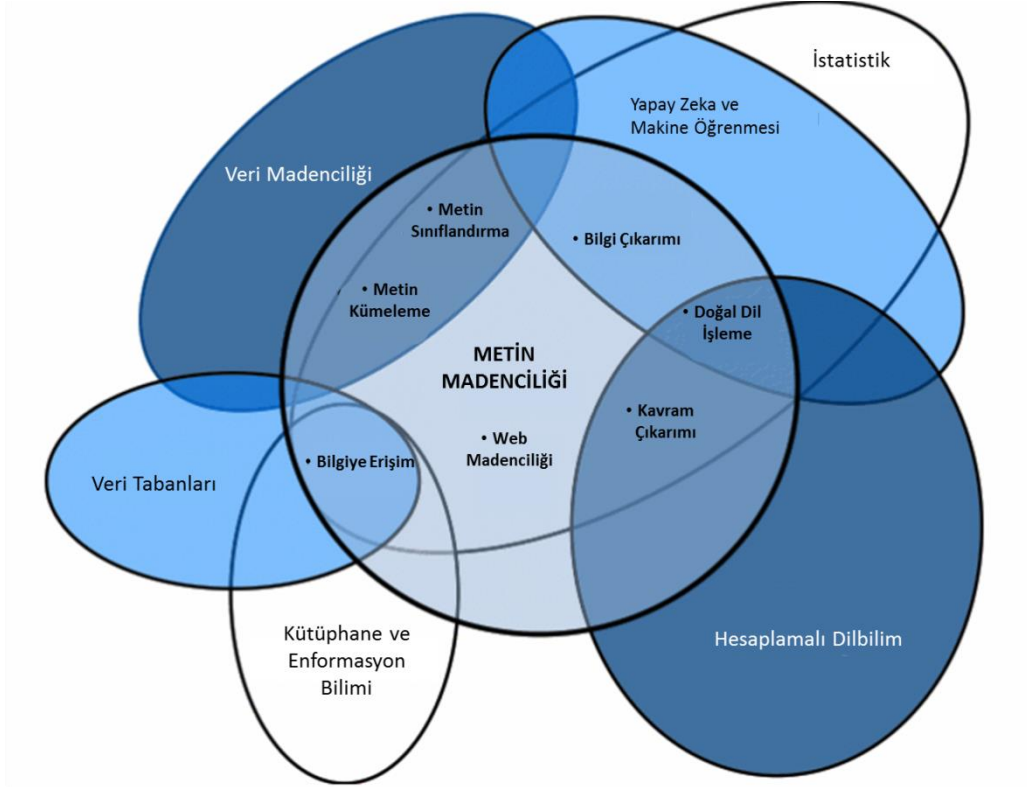
Miner vd. (2012)⁷'e göre metin madenciliği ile veri madenciliği, istatistik, hesaplamalı dilbilim gibi altı adet ilgili alanın kesişiminin Venn şeması Şekil 1'de verilmiştir. Bu Venn

⁵ Gürsakal, N. (2014). s.46

⁶ Miller, T. W. (2005). s.105

⁷ Miner, G., D., Delen, J., Elder, A., Fast, T., Hill, R., Nisbet. (2012). s.31

şemasında yedi adet metin madenciliği uygulama alanı, metin madenciliği ile altı adet ilgili alanın kesişme noktalarında verilmiştir.



Şekil 1: Metin madenciliğinin ilişkili olduğu disiplinler

Kaynak: Miner vd.(2012), s.31

Venn şemasında bulunan metin madenciliğinin yedi uygulama alanı aşağıdaki gibidir:⁸

- Bilgiye Erişim (IR-Information retrieval): Arama motorları ve anahtar kelime arama dâhil, metin belgelerinin saklanması ve temin edilmesi.
- Metin kümeleme: Veri madenciliği kümeleme yöntemleri kullanılarak terimlerin, paragrafların veya belgelerin gruplandırılması.
- Metin sınıflandırma: Etiketli örnekler üzerinde eğitilmiş modellere dayalı olarak veri madenciliği sınıflandırma yöntemleri kullanılarak paragrafların veya belgelerin sınıflandırılması.
- Web madenciliği: Web üzerindeki bilgilerin işlenmesi ve analiz edilmesi.

⁸ Miner, G., D., Delen, J., Elder, A., Fast, T., Hill, R., Nisbet. (2012). s.32

- Bilgi çıkarımı (IE-Information extraction): Yapılandırılmamış metinden ilgili gerçeklerin ve ilişkilerin belirlenmesi ve çıkarılması.
- Doğal dil işleme (NLP-Natural Language Process): Doğal dili çözümleyip yorumlayacak bilgisayar sistemlerinin tasarlanması ve uygulaması.
- Kavram çıkarımı: Kelimelerin ve cümlelerin anlamsal olarak benzerliklerine göre gruplandırılması.

1.2. DUYGU ANALİZİ

Duygu Analizi (Sentiment Analysis), metin madenciliğinin önemli bir alanıdır. Duygu analizi terim olarak, “kişinin bir şey hakkında ne hissettiği”, “kişisel deneyim, kendi hissi”, “bir şeye yönelik tutum” veya “bir fikir” anlamına gelir.⁹ Literatürde, duygu durum analizi, fikir madenciliği, duygu sınıflandırma, kanaat çıkarımı gibi farklı adlarla da yer almaktadır.

Duygu analizi, Hesaplamalı Dilbilim¹⁰'in ve bir konuşmacının fikrini veya tutumunu inceleyen fikir madenciliğinin bir parçasıdır. Agarwal vd.¹¹'e göre, duygu analizi, duygu incelemelerinden duyguların çıkarılmasını veya sınıflandırılmasını otomatikleştirmek için doğal dil işlemeyi, metin analizini ve hesaplama tekniklerini kullanmaktadır.

Duygu analizinde temel görev, belgedeki metnin polaritesini, cümle ya da özellik düzeyi yönünü yani metinde ifade edilen görüşlerin, cümlelerin veya işletme unsurunun özelliğinin olumlu, olumsuz veya nötr olup olmadığını sınıflandırmaktır.

Duygu analizi, doğal dil işleme, veri madenciliği ve metin madenciliği gibi çeşitli araştırma alanlarını bir araya getirmektedir ve işletmeler hesaplamalı zeka yöntemlerini faaliyetlerine entegre etmeye, ürün ve hizmetlerini daha fazla aydınlatmaya ve iyileştirmeye çalıştıkça duygu analizi işletmeler için hızla önem kazanmaktadır.¹²

Son yılların önemli araştırma konularından biri de duygu analizidir. Günümüzde işletmeler/kişiler tarafından üretilen ürünler çok hızlı bir şekilde tüketiciye ulaşmaktadır. Bu

⁹ Farhadloo, M., E., Rolland. (2016). s.2

¹⁰ Hesaplamalı Dilbilim; doğal dili istatistiksel veya kural tabanlı modeller ile hesaplamalı olarak inceleyen bir bilgisayar bilimidir.

¹¹ Agarwal, B., N., Mittal, P., Bansal, S., Garg. (2015).

¹² Farhadloo, M., E., Rolland. (2016). s.2

ürünlere yapılmış olan yorumlar gelişen teknoloji ile beraber internet dünyasına hızlı bir şekilde yansımaktadır. Bu yorumların ne anlam ifade ettiği üreticiler için önem arz etmektedir.

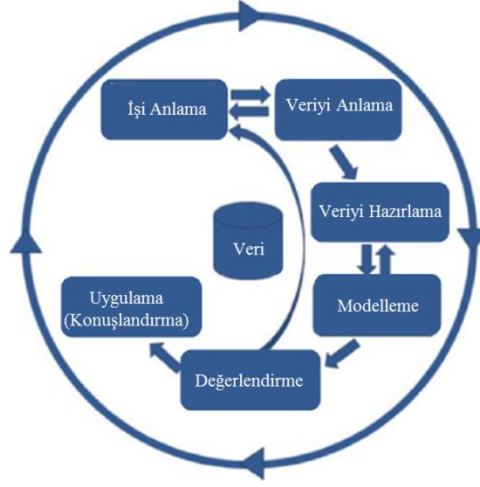
1.2.1. Duygu Analizi Süreci

Bu başlık altında veri madenciliğinin uluslararası düzeyde standardı olarak kabul edilmiş olan CRISP-DM (CRoss Industry Process for Data Mining) referans modeli ile duygu analizinin nasıl yapıldığı anlatılmaktadır. CRISP-DM (CRoss Industry Process for Data Mining) süreç modeli altı adımdan oluşmaktadır. Bu süreç aşağıda bulunan Şekli 2’de görülmektedir.

1.2.1.1. İşi Anlama

Herhangi bir çalışmada/projede yapılması gereken ilk şey, tam olarak neyi başarmaya çalıştığımızı tanımlamak yani problemin tanımıdır. Gerçekleştirilecek duygu analizi çalışmasının amacı ve gereksinimlerinin anlaşılması ve daha sonra da bu bilgilerin hedeflenen amacı net olarak ifade etmesi gerekmektedir.

Bir duygu analizi çalışmasında, başarmaya çalıştığımız şey müşteri şikâyetleri belirleyip, müşterilerin memnuniyeti için şikâyet kategorisine göre hızlıca aksiyon almak olabilir. Bu amaç doğrultusunda işi anlama sürecinde, müşteri yorumları gibi metin mesajlarının incelenerek anlaşılması, bu metin mesajları arasındaki benzerliklerin çıkarılması, metin mesajlarındaki niteliklerin tespit edilmesi, şikâyet nedenlerinin tespiti vb. çalışmalar yapılmalıdır.



Şekil 2 : CRISP-DM: Referans Modelinin Adımları

Kaynak: Şeker, S.E. (2018), s.11

1.2.1.2. Veriyi Anlama

İkinci aşama, başlangıç verilerinin toplanmasıyla beraber sahip olunan verilerin uygunluğunun değerlendirilmesidir. Düşünce sürecinden geçirilerek hedef çalışmada kullanılacak verilere aşinalık kazanmaktır. Şekil 2’de veriyi anlama süreci işi anlama sürecinin bütünleyici bir parçası olduğu için çift yönlü oklar ile gösterilmektedir. İş anladıkça farklı verilere bakmak veya verilerin gösterdiklerini anlamak, verilere baktıkça iş ile ilgili farklı bakış açıları kazanmak mümkündür.¹³

Veriyi anlama süreci dört aşamadan oluşmaktadır¹⁴:

- Başlangıç verilerinin toplanması,
- Toplanan verinin tanımlanması ve bu verilerin ihtiyaçları karşılama yeterliliğinin değerlendirilmesi,
- Çalışmanın gerçekleştirilebilmesi için veri anlamında eksiklerin tespit edilmesi,
- Veri tam mı, doğru mu, hatalar içeriyor mu, hatalar içeriyorsa ne tür hatalar içeriyor, veride eksik bölümler var mı şeklindeki sorular ile verinin kalitesinin tespit edilmesi

¹³ Argüden, Y., B., Erşahin. (2008). s.21

¹⁴ Argüden, Y., B., Erşahin. (2008). s.21

Veriyi anlama süreci içerisinde, başlangıç verilerinin toplanması aşamasında seçilmiş olan konuya uygun bir şekilde veriler toplanmaktadır. Konu seçimi yapıldıktan sonra, konuya uygun anahtar kelimeler belirlenmekte ve bu anahtar kelimeler ile farklı kaynaklardan/platformlardan veri çekilerek başlangıç veri kümesi oluşturulmaktadır. Veriler toplandıktan sonra, veri tam mı, doğru mu, hatalar içeriyor mu, hatalar içeriyorsa ne tür hatalar içeriyor, veride eksik bölümler var mı şeklindeki sorular ile kontrol yapılarak verinin kalitesi belirlenmektedir.

1.2.1.3. Veri Hazırlama

Üçüncü adım verinin analiz edilebilmesi için hazırlanmasıdır. Duygu analizinde veri hazırlama aşaması üç ana adımdan oluşmaktadır: Veri Etiketleme, Veri Önışleme ve Öznitelik seçimi.

Veri Etiketleme:

Veri etiketlemenin temel amacı yeni bir nesnenin özelliklerini açıklamak ve bu nesnenin daha önceden tanımlanmış olan sınıf etiketlerinden birine atamasını yapmaktır. Duygu analizinde hedef değişkenin kategorik olması şartıyla genellikle olumlu, olumsuz ve nötr etiketleri kullanılmaktadır. Medhat vd. (2014)¹⁵ ‘ne göre, duygu analizinde üç tane sınıflandırma (veri etiketleme) seviyesi bulunmaktadır: Doküman Seviyesi, Cümle Seviyesi ve Görüş Seviyesi. Medhat vd. bu seviyeleri aşağıdaki şekilde açıklamışlardır:

- Doküman Seviyesi: Bir fikir (duygu) belgesini, olumlu ya da olumsuz bir görüş ya da duyguları ifade edecek şekilde sınıflandırmayı amaçlamakta ve belgenin tamamını bir bilgi birimi olarak görmektedir.
- Cümle Seviyesi: Her bir cümledeki duyguyu sınıflandırmayı amaçlamakla beraber ilk olarak cümlenin sübjektif ya da objektif olup olmadığı belirlenmektedir. Cümle sübjektif ise, cümlenin olumlu ya da olumsuz bir fikir (duygu) ifade edip etmediğine karar verilmektedir.
- Görüş Seviyesi: Duyguyu (fikri) varlıkların belirli yönlerine (özelliklerine) göre sınıflandırmayı amaçlanmakta ve varlıkların özellikleri belirlenmektedir. Yorumcular aynı

¹⁵ Medhat, W., A., Hassan, H., Korashy,.(2014). s.1093

varlığın farklı özellikleri için farklı yorumlar yapabilmektedir. “Telefonun ses kalitesi hiç iyi değil ama şarj ömrü uzun” cümlesi bu duruma bir örnektir.

Veri Önileme:

Veri önileme, basitçe ham olan başlangıç verisini anlaşılabilir formata dönüştürme aşamasıdır. Başka bir deyişle, başlangıç verilerinin çalışmalara temel oluşturacak final verilere dönüştürülmesi aşamasıdır.¹⁶

Gerçek dünya verileri bazen eksik, tutarsız, gereksiz ve gürültülü olabilmektedir. Veri önileme aşaması, verilerin analiz edilebilmesi için başlangıç verilerinin uygun bir formata dönüştürülmesine yardımcı olan çeşitli adımlar içermektedir. Bu adımlar;

Metin Normalleştirme: Metin normalleştirme, metni daha önce sahip olmadığı tek bir kanonik¹⁷ forma dönüştürme işlemidir. Depolanmadan veya işlenmeden önce metni normalleştirmek metnin tutarlı olması garanti etmektedir.

Durak Kelimelerin Çıkartılması: Türkçede çok sık kullanılan zarflar, edatlar, zamirler yani tek başına anlam ifade etmeyen durak kelimeler metinden çıkartılır.

Dönüştürme: Veri dönüşümü aşaması verinin işlenmesi ve üzerinde analiz yapılabilmesi için uygun hale getirilmesi amacıyla uygulanan yöntemdir. Dönüştürme yöntemi uygulanarak metinlerin standartlaştırılması veya normalleştirilmesi hedeflenmektedir.

Tarama ve İşaretleme: Tarama ve işaretleme aşamasında, metin içindeki terimler temizlenir. Metin içerisinde bulunan tüm noktalama işaretleri, varsa emojiler, semboller, simgeler, sayılar temizlenir.

¹⁶ Argüden, Y., B., Erşahin. (2008). s.21

¹⁷ Kanonik sözcüğü, Yunanca kanon, (kanun, kural) kökünden türemiş bir sıfattır. "genel olarak kabul edilen" veya "otoritelerce doğrulanmış" anlamlarında kullanılır.

Kök Bulma: Kök bulma aşaması, kelimelerin morfolojik analizini göz önünde bulundurarak, yani bir kelimenin çeşitli şekillendirilmiş biçimlerini bir arada toplayarak tek bir madde olarak analiz edilmelerini sağlayan görevdir.¹⁸

Veri Kümesini Özniteliklerine Ayırma (Tokenization): Metnin boşluk, “-”, “?” vb. ayraçlara göre parçalara(kelimelere) ayrılması işlemidir. Her bir parça(kelime) bir özniteliktir. Burada her bir öznitelik birbirinden bağımsız değişkendir. Her bir bağımsız değişkenin bağımlı olduğu değişken ise ait olduğu sosyal medya yorumudur.

Öznitelik/Özellik Seçimi:

Öznitelik seçimi (feature selection), orijinal veri kümesini temsil edebilecek en iyi altkümenin seçimi olarak tanımlanmaktadır.¹⁹ Genel özellik seçim prosedürü, her bir potansiyel özelliği, belirli bir özellik seçim metriğine göre puanlamak ve ardından en iyi k özelliğini almaktır.²⁰ Öznitelik seçimi aşamasında, ilgisiz ve gereksiz niteliklerin silinerek veri kalitesinin artması hedeflenmektedir.

Öznitelik seçme işleminin avantajları şunlardır: ²¹

- Öznitelik kümesinin boyutunu düşürür ve algoritma hızını artırır,
- İlgili olmayan ve gürültülü veriyi ortadan kaldırır,
- Veri kalitesini geliştirir,
- Veri kümesini daha basit bir şekilde tanımlanabilir, görselleştirilebilir ve anlaşılabilir hale getirir,
- Veri kümesini oluşturmak için gerekli olan veri toplama işleminde kaynak tasarrufu sağlar,
- Veri depolamak için gerekli olan hafıza miktarını azaltır,
- Elde edilen modelin başarısını artırır.

s.3

¹⁸ Allahyari, M., S., Pouriyeh, M., Assefi, S., Safaei, E.D., Trippe, J.B., Gutierrez, K., Kochut. (2017).

¹⁹ Budak, H. (2018). s.1

²⁰ Forman, G. 2003, s.1291

²¹ Ladha, L., T., Deepa. (2011). s.1788

1.2.1.4. Duygu Analizi - Modelleme

Bu aşamada duygu analizindeki sınıflandırma problemleri için uygun modeller seçilmekte, bu modeller kullanılmakta ve değerlendirilmektedir. Tanımlanmış bir problem için en uygun modelin bulunabilmesi, olabildiği kadar çok sayıda modelin kurularak denenmesi ile mümkün olmaktadır ve bu nedenle veri hazırlama ve model kurma aşamaları, en iyi ve en doğru olduğu düşünülen modele ulaşıncaya kadar tekrarlanan bir süreçtir.²²

Makine öğrenmesi teknikleri denetimli (supervised) ve denetimsiz (unsupervised) öğrenme yaklaşımlarından oluşmaktadır. Denetimli (supervised) ve denetimsiz (unsupervised) öğrenimin kullanıldığı modellere göre model oluşturma süreci farklılık göstermektedir.

Denetimsiz öğrenimde, kümeleme analizinde de olduğu gibi ilgili örneklerin gözlenmesi ve bu örneklerin öznitelikleri arasındaki benzerliklerden hareket edilerek sınıfların tanımlanması amaçlanmaktadır.

Denetimli öğrenimde ise, veriler seçilen algoritma için hazır hale getirildikten sonra, verinin bir kısmı modelin öğrenimi, bir kısmı ise modelin geçerliliğinin testi için ayrılmaktadır. Modelin öğrenimi öğrenme kümesi kullanılarak yapıldıktan sonra, sınama kümesi ile modelin doğruluk derecesi (accuracy) belirlenmektedir.

Denetimli öğrenimde sınırlı sayıda veriye sahip olunması durumunda, kullanılacak diğer bir yöntem çapraz doğrulama (cross validation) yöntemidir. Bu yöntemde veri kümesi rasgele olarak n adet eşit parçaya ayrılmaktadır. İlk olarak, a parçası üzerinde modelin öğrenimi yapılmaktadır, diğer parçalar üzerinde modelin geçerliliği test edilmektedir. İkinci olarak ise b parçası üzerinde modelin öğrenimi yapılmakta, diğer parçalar üzerinde test yapılmaktadır. Bu şekilde her veri alt kümesi için yapılan işlemler sonucunda elde edilen hata oranlarının ortalaması kullanılmaktadır.

Model başarısının değerlendirilmesi için değerlendirme ölçütleri hesaplanmaktadır. Özellikle sınıflama problemleri için kurulan modellerin doğruluk derecelerinin değerlendirilmesinde basit fakat faydalı bir araç olan risk matrisi kullanılmaktadır.²³ Risk

²² Akpınar, H. (2000). s.10

²³ Akpınar, H. (2000). s.12

matrisi literatürde karşıtlık matrisi, hata matrisi gibi farklı isimlerle de kullanılmaktadır. Karşıtlık matrisinden yararlanılarak hesaplanan birçok değerlendirme ölçütü bulunmaktadır. Bunlardan bazıları; sınıflandırma doğruluğu, hata oranı, duyarlılık, kesinlik, F-ölçütüdür.

1.2.1.5. Değerlendirme

Modelin uygulama aşamasına geçmeden önce, modeli daha ayrıntılı bir şekilde değerlendirmek ve modelin oluşturulması için atılan adımları gözden geçirerek, iş hedeflerinin doğru bir şekilde gerçekleştiğinden emin olmak önemlidir.²⁴ Bir veya daha fazla model oluşturulduğunda bunların yüksek kalite açısından değerlendirilmesi gerekmektedir. Birden fazla model oluşturulduğunda, test sonuçlarına bakılarak en verimli modelin seçilmesi bu aşamada yapılmaktadır.

Bu aşamanın sonunda duygu analizi sonuçlarının kullanımına dair karar verilmesi gerekmektedir. Elde edilen model başarısı ile mevcut karar sürecinden (mevcut yapıdan) ve daha önceki çalışmalardan daha yüksek başarı elde ediliyorsa uygulama aşamasına geçilmelidir.

1.2.1.6. Uygulama (Konuşlandırma)

Uygulama (konuşlandırma), bir önceki aşamada karar sürecinden geçirilen başarılı modelin bu sürecin girdisi olarak kullanılmasıdır. Modelin oluşturulmasıyla kazanılan bilginin, müşterinin/kullanıcının kullanabileceği şekilde organize edilmesi ve sunulması gerekmektedir.

Uygulama adımı gerçekleştirildikten sonra, güncelliğini yitirmiş olan modellerin tespiti için modellerin devamlı izlenmesi gerekmektedir. Zaman içinde tüm sistemlerin özelliklerinde ve paralel olarak ürettikleri verilerde değişiklikler ortaya çıkabilmektedir.

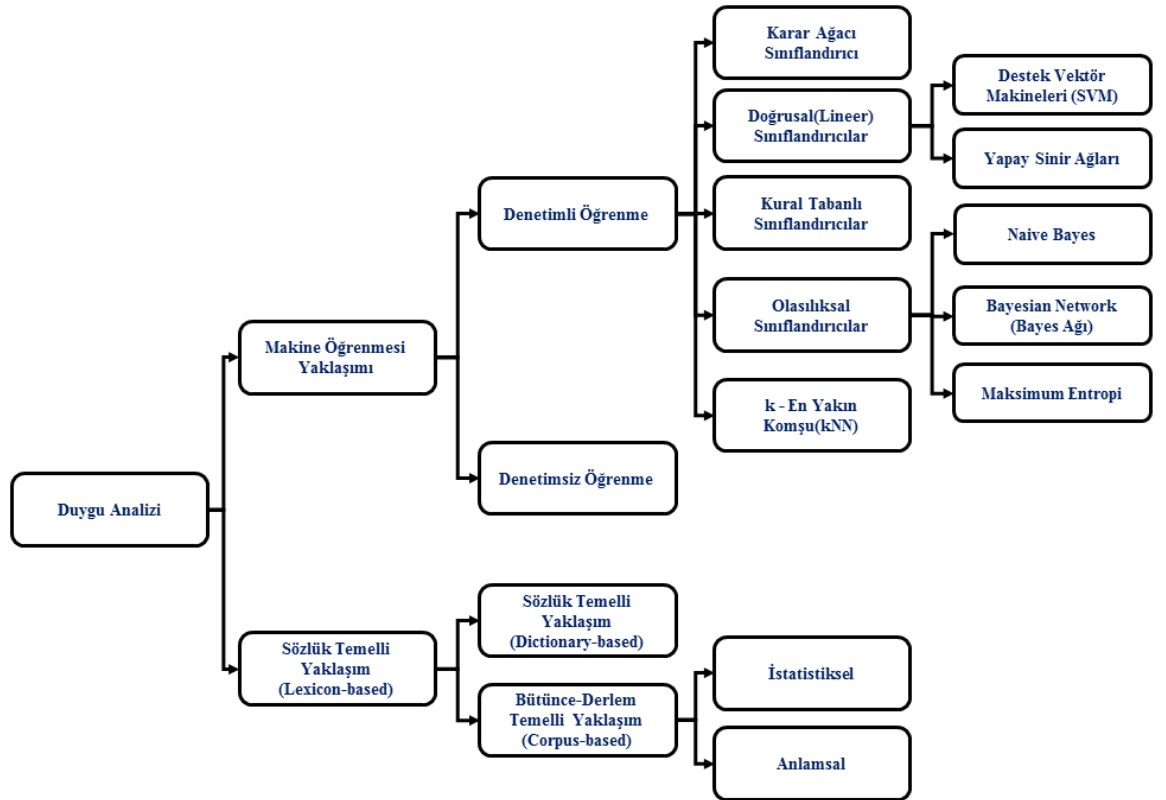
1.2.2. Duygu Sınıflandırma Yöntemleri

Duygu sınıflandırması yöntemi, makine öğrenmesi temelli ve sözlük temelli yaklaşım olmak üzere iki ana dala ayrılmaktadır. Makine öğrenmesi temelli yaklaşım sınıflandırma sırasında makine öğrenme algoritmalarını ve dilbilimsel özelliklerini kullanmaktadır. Sözlük

²⁴ Akdemir, E,B. (2019). s.19

tabanlı yaklaşım ise sınıflandırma sırasında önceden hazırlanmış duygu kavramlarından oluşan sözlüklerden yararlanmaktadır.

Makine öğrenmesi uygulamalarında metin sınıflandırma için çok farklı yöntemler kullanılmaktadır. Naive Bayes, karar ağaçları (decision tree), yapay sinir ağları (artificial neural network), örnek tabanlı sınıflandırıcılar (example based classifier), destek vektör makineleri (support vector machine) ve istatistiksel dil modeli (statistical language model) tabanlı sınıflandırıcılar yaygın olarak tercih edilmektedir.²⁵ Şekil 3 'te Duygu Sınıflandırma Yöntemleri verilmiştir.



Şekil 3: Duygu Sınıflandırma Yöntemleri

Kaynak: Grljević, O. ,Bošnjak, Z. (2018). s.42

²⁵ Tantuğ, A. C. (2012).

1.2.2.1. Naive Bayes

Naive Bayes, etkili bir sınıflandırma algoritmasıdır. Algoritmanın isminde bulunan 'Naive', veri kümesindeki özniteliklerin birbirinden bağımsız olduğu varsayımından gelmektedir. Veri kümesindeki bir özneliğin varlığı diğer özniteliklerden herhangi birine bağlı değildir. Sınıflandırılmış örnek verileri kullanarak yeni bir verinin mevcuttaki sınıflardan herhangi birine ait olma olasılığını hesaplayan bir yaklaşımdır. Bu sınıflandırıcıda öznitelikler birbirinden bağımsız olarak kabul edilmektedir.²⁶

Bayes modeli basit bir model olmakla beraber iteratif değildir, dolayısı ile büyük veri kümeleri için de kullanışlıdır ve bu basitliği ile birlikte, karışık diğer yöntemlere oranla gayet iyi sonuçlar verdiği için yaygın bir kullanım alanı bulmuştur.²⁷

Bayes teoremi bize, $P(c|x)$ olasılığını, $P(c)$, $P(x)$ ve $P(x|c)$ olasılıklarından hesaplama imkanı verir. Naive Bayes sınıflandırıcısı, bir tahmincinin (x) değerinin, belirli bir sınıf (c) üzerindeki etkisinin, diğer tahmin edicilerin değerlerinden bağımsız olduğunu varsayar. Bu varsayımına sınıf şartlı bağımsızlık denir. Bayes teoremi şu şekilde ifade edilebilir:²⁸

$$P(c|x) = \frac{P(x|c) P(c)}{P(x)} \quad (1.1)$$

$P(c|x)$; x olayı gerçekleştiği durumda c olayının meydana gelme olasılığıdır

$P(x|c)$; c olayı gerçekleştiği durumda x olayının meydana gelme olasılığıdır

$P(c)$; c sınıfının önsel²⁹ olasılığıdır.

$P(x)$; x tahmin edicisinin önsel olasılığıdır.

²⁶ Karakoyun, M., M., Hacıbeyoğlu. (2005). s.34

²⁷ Arslan, N. (2018). s.6

²⁸ Muthén, B., T., Asparouhov. (2011). s.8

²⁹Önsel olasılık, Bayesci istatistikte gözlemlere atıf yapmadan önce değerlendirilen özellikle önsel olabilen olasılıktır.

1.2.2.2. Sıralı Minimal Optimizasyon

Sıralı Minimal Optimizasyon (SMO) algoritması ilk kez John Platt tarafından 1998'de yayınlanmıştır. John Platt, bir destek vektör sınıflandırıcısını eğitmek için SMO algoritmasını uygulamıştır. Platt³⁰'a göre, SMO algoritması, Destek Vektör Makineleri ikinci dereceden programlama problemini (DVM QP) herhangi bir ekstra matris depolama alanı olmadan ve tüm sayısal ikinci dereceden programlama optimizasyon adımları kullanmadan hızlı bir şekilde çözebilen basit bir algoritmadır.

SMO algoritması, ayrıştırma metodu fikrinin en uç noktaya çekilmesi ve her bir yinelemede sadece iki noktadan oluşan minimum bir alt kümenin optimize edilmesiyle elde edilir.³¹ Bu algoritma, tüm kayıp değerlerini yenisiyle değiştirirken, nominal olan öznelikleri ikili olanlara dönüştürür ve tüm öznelikleri daha önce tanımlanmış değerlerle normalize eder.

Platt'a göre, SMO her adımda mümkün olan en küçük optimizasyon problemini çözmeyi seçer. Standart DVM QP problemi için, mümkün olan en küçük optimizasyon problemi iki Lagrange çarpanı içerir, çünkü Lagrange çarpanları doğrusal bir eşitlik sınırına uymak zorundadır. SMO her adımda, iki Lagrange çarpanını ortaklaşa optimize etmek için seçer, bu çarpanlar için en uygun değerleri bulur ve yeni optimum değerleri yansıtacak şekilde DVM'yi günceller.

SMO'nun avantajları:³²

- İki Lagrange çarpanının analitik olarak çözülmesini sağlayarak sayısal QP optimizasyonu tamamen önler.
- Algoritmanın iç döngüsü, bütün bir QP kütüphanesi yordamını çağırarak yerine, kısa bir miktarda C koduyla ifade edilebilir.
- Her ne kadar algoritma süresince daha fazla optimizasyon alt problemi çözülsede, her alt problem o kadar hızlıdır ki, genel QP problemi hızlı bir şekilde çözülür.
- SMO hiçbir ekstra matris depolaması gerektirmez. Bu nedenle, çok büyük DVM eğitim sorunları, sıradan bir kişisel bilgisayar veya iş istasyonunun hafızasına sığabilir.

³⁰ Platt, J. (1998). s.6

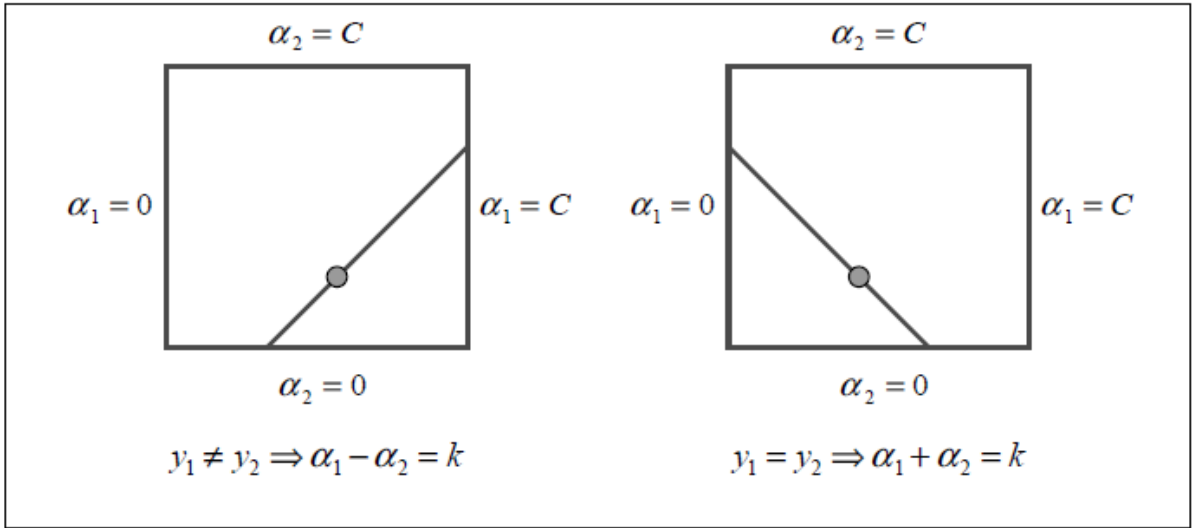
³¹ Lee, Y-J. (2017). s.4

³² Platt, J. (1998). s.6

- SMO'da hiçbir matris algoritması kullanılmadığından, sayısal hassasiyet problemlerine karşı daha az hassastır.

SMO' nun iki bileşeni vardır: İki Lagrange çarpanı için bir analitik yöntem ve optimize edilebilecek çarpanları seçmek için bir buluşsal yöntem.

İki Lagrange çarpanı, problemin tüm kısıtlarını yerine getirmelidir. Şekil 4'te görüldüğü gibi eşitsizlik kısıtlamaları, Lagrange çarpanlarının kutunun içinde yer almasına neden olur. Doğrusal eşitlik kısıtı, çapraz çizgi üzerinde bulunmasına neden olur. Bu nedenle, bir SMO adımı, diyagonal bir çizgi segmentindeki amaç fonksiyonun optimumunu bulmalıdır.



Şekil 4: SMO Optimizasyonu

Kaynak: Platt, J., (1998), s.6

SMO' nun iki bileşeninin detayları şu şekildedir.³³ Şekil 4'te çapraz çizgi bölümünün uçları oldukça basit bir şekilde ifade edilebilir. Genelliği kaybetmeden, SMO algoritması önce ikinci Lagrange çarpanı α_2 'yi hesaplar ve çapraz çizgi bölümünün uçlarını α_2 cinsinden hesaplar.

$$y_1 \neq y_2 \text{ ise, } L = \max\{0, \alpha_2^{old} - \alpha_1^{old}\}, H = \min\{C, C - \alpha_1^{old} + \alpha_2^{old}\} \quad (2.3)$$

³³ Platt, J. (1998). s.7

$$y_1 = y_2 \text{ ise, } L = \max\{0, \alpha_2^{old} + \alpha_1^{old} - C\}, H = \min\{C, \alpha_1^{old} + \alpha_2^{old}\} \quad (2.4)$$

Köşegen çizgi boyunca amaç fonksiyonunun ikinci türevi şöyle ifade edilebilir:

$$\eta = K(\vec{x}_1, \vec{x}_1) + K(\vec{x}_2, \vec{x}_2) - 2K(\vec{x}_1, \vec{x}_2) \quad (2.5)$$

Normal şartlar altında, amaç fonksiyonunun pozitif olacağı kesindir. η , doğrusal eşitlik kısıtlaması yönünde minimum bir değer olmak üzere sıfırdan büyük olacaktır.

$$E_i = f(x_i) - y_i = \left(\sum_{j=1}^l a_j y_j + b\right) - y_i \quad i = 1,2 \quad (2.6)$$

Denklem (2.6), i . eğitim örneğindeki hata olmak üzere α_2^{new} denklem (2.7) deki gibi hesaplanır.

$$\alpha_2^{new} = \alpha_2^{old} + \frac{y_2\{E_1 - E_2\}}{\eta} \quad (2.7)$$

Son adım olarak ise, sınırlandırılmış $\alpha_2^{new,clipped}$, sınırlandırılmamış α_2^{new} 'in çapraz çizgi bölümünün uçlarına kesilmesiyle bulunur:

$$\alpha_2^{new,clipped} = \begin{cases} H & \text{eğer } \alpha_2^{new} \geq H \\ \alpha_2^{new} & \text{eğer } L \leq \alpha_2^{new} \leq H \\ L & \text{eğer } \alpha_2^{new} \leq L \end{cases} \quad (2.8)$$

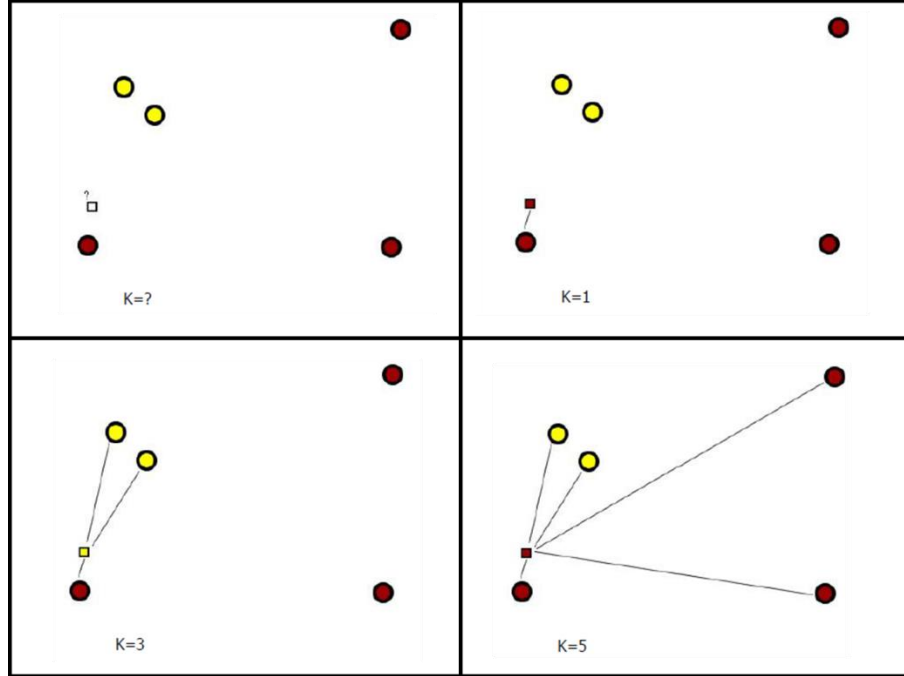
Şimdi de, $s = y_1 y_2$ olmak üzere, α_1^{new} değeri sınırlandırılmış $\alpha_2^{new,clipped}$ 'den hesaplanır:

$$\alpha_1^{new} = \alpha_1^{old} + y_1 y_2 (\alpha_2^{old} - \alpha_2^{new,clipped}) \quad (2.9)$$

1.2.2.3. K-En Yakın Komşu (kNN)

K-En Yakın Komşu (kNN), metin kategorizasyonu için en popüler algoritmalarından biridir. K-En Yakın Komşu algoritmasının arkasındaki fikir oldukça basittir. K değeri dışardan istenir. Bilinmeyen noktadan bilinen veri noktalarına kadar mesafeler hesaplanır. Hesaplanan mesafeler artan sırada düzenlenir, listenin en üstünde K mesafeleri seçilir (en yakın olanlar).

Seçilen veri noktalarının sınıfları arasında, baskın sınıf, bilinmeyen noktanın sınıfı olarak belirlenir.



Şekil 5: kNN Sınıflandırma

Kaynak: <http://bmb.cu.edu.tr/uorhan/DersNotu/Ders02.pdf>

kNN algoritmasının performansında etkili ve önemli parametreler uzaklık ölçütü, k komşu sayısı, ve ağırlıklandırma yöntemidir. Uzaklık ölçütlerinden bazıları Euclid (Öklid), Minkowski, Chebyshev ve Manhattan'dır. K-NN algoritmasında, komşu sayısı (k) parametresinin değerine bağlı olarak sınıflandırma yapılmaktadır. Sınıflandırma sürecinde, k=1 olduğunda sadece en yakın komşunun bulunduğu sınıfa atama yapılırken, k sayısı örnek sayısına (N) yaklaştıkça veri setinde yer alan tüm veriler dikkate alınmakta ve oylamaya göre seçim yapılmaktadır.³⁴ Ağırlıklandırma ile sınıflandırılan niteliğe daha yakın olan komşu niteliklerin, çoğunluk oylamasına daha fazla katkı sağlaması amaçlanmaktadır. En çok kullanılan ağırlık değeri, d:komşular arası uzaklık olmak üzere, $1/d$ dir.³⁵

³⁴ Taşçı, E., A., Onan. (2016). s.4

³⁵ Doad, P.K., M.M., Bartere. (2013). s.3142

En yaygın kullanılan uzaklık ölçüleri şu şekilde özetlenebilir,

Minkowski Uzaklığı; Öklid uzayında tanımlı olan bir dizidir. Genelleştirilmiş metrik mesafesidir. Sıklıkla kullanılan Öklid uzaklığı ve Manhattan uzaklığının bir genelleştirilmesi ile oluşturulmaktadır. n boyutlu uzayda herhangi $A = (a_1, a_2, a_3, \dots, a_n)$ ve $B = (b_1, b_2, b_3, \dots, b_n)$ noktaları arasındaki uzaklık şu şekilde hesaplanır:

$$d(A, B) = \left(\sum_{i=1}^n |a_i - b_i|^p \right)^{1/p} \quad (3.1)$$

Manhattan Uzaklığı; Minkowski uzaklık ölçütünün $p=1$ olduğu özel durumu Manhattan uzaklığını vermektedir. n boyutlu bir uzayda iki nokta arasındaki farkların mutlak değerlerinin toplamıdır. n boyutlu uzayda herhangi $A = (a_1, a_2, a_3, \dots, a_n)$ ve $B = (b_1, b_2, b_3, \dots, b_n)$ noktaları arasındaki uzaklık şu şekilde hesaplanır:

$$d(A, B) = \left(\sum_{i=1}^n |a_i - b_i| \right) \quad (3.2)$$

Öklid Uzaklığı; İki nokta arasındaki doğrusal uzaklıktır. Minkowski uzaklık ölçütünün $p=2$ olduğu özel durumu Öklid uzaklığını vermektedir. Sınıflandırma ve kümeleme algoritmalarında en sık kullanılan uzaklık ölçütüdür. n boyutlu bir Öklid uzayında herhangi $A = (a_1, a_2, a_3, \dots, a_n)$ ve $B = (b_1, b_2, b_3, \dots, b_n)$ noktaları arasındaki uzaklık şu şekilde hesaplanır.

$$d(A, B) = \sqrt{\sum_{i=1}^n (a_i - b_i)^2} \quad (3.3)$$

Chebyshev Uzaklığı; p sonsuza ulaşırken Minkowski uzaklık ölçütünün limitini aldığımızda Chebyshev uzaklığını elde ederiz yani Minkowski uzaklık ölçütünün $p \rightarrow \infty$ olduğu özel durumu chebyshev uzaklığını vermektedir. İki nokta arasındaki farkların mutlak değerlerinin maksimumu şeklinde tanımlanabilmektedir. n boyutlu uzayda herhangi $A = (a_1, a_2, a_3, \dots, a_n)$ ve $B = (b_1, b_2, b_3, \dots, b_n)$ noktaları arasındaki uzaklık şu şekilde hesaplanır:

$$d(A, B) = \lim_{p \rightarrow \infty} \left(\sum_{i=1}^n |a_i - b_i|^p \right)^{1/p} = \max_{i=1}^n |a_i - b_i| \quad (3.4)$$

1.2.3. Model Başarısını Değerlendirme

Model başarısının değerlendirilmesi için farklı metotlar ve ölçütler kullanılabilir. Bu başlık altında karışıklık matrisi, değerlendirme ölçütleri ve k-katlı çapraz doğrulama anlatılmaktadır.

1.2.3.1. Karışıklık Matrisi

Makine öğrenmesi algoritmaları kullanılarak verinin hangi sınıfa ait olduğu tahmini yapılmaktadır. Karışıklık matrisi; bir sınıflama algoritması tarafından oluşturulan gerçek değerler ile tahmin değerleri hakkında bilgiyi gösterir. Karışıklık matrisini Tablo 1’de görüldüğü gibi doğru pozitif (TP), yanlış pozitif (FP), yanlış negatif (FN) ve doğru negatif (TN) olmak üzere dört adet sınıflandırma değeri içerir. Bu dört değer genel olarak iki sınıflı sınıflandırma problemini değerlendirmek için kullanılır.

Tablo 1: Karışıklık Matrisi

		Tahmin değerleri		
		Pozitif	Negatif	Toplam
Gerçek değerler	Pozitif	TP	FN	Toplam Pozitif Gerçekler
	Negatif	FP	TN	Toplam Negatif Gerçekler
	Toplam	Toplam Pozitif Tahminler	Toplam Negatif Tahminler	Tüm Veri Adedi

Önceden belirlenmiş sınıfları pozitif ve negatif şeklinde adlandıracak olursak karşıtlık matrisindeki değerler aşağıdaki şekillerde ifade edilebilir:

TP (True Positive) : Doğru tahmin edilen pozitif değer sayısı

TN (True Negative) : Doğru tahmin edilen negatif değer sayısı

FP (False Positive) : Yanlış tahmin edilen pozitif değer sayısı

FN (False Negative) : Yanlış tahmin edilen negatif değer sayısı

Toplam Pozitif Gerçekler = TP + FN

Toplam Negatif Gerçekler = FP + TN

Toplam Pozitif Tahminler = TP + FP

Toplam Negatif Tahminler = FN + TN

Tüm Veri Adedi (N) = TP + TN + FP + FN

1.2.3.2. Değerlendirme Ölçütleri

Değerlendirme ölçütleri, makine öğrenmesi algoritmalarının değerlendirilmesinde ve bu algoritmaların karşılaştırılmasında ihtiyaç duyulan ölçütlerdir. Bu ölçütlerin çoğu karşıtlık matrisinden yararlanılarak hesaplanmaktadır.

Sınıflandırma doğruluğu; doğru olarak tahmin edilen verilerin tüm veri adedine oranıdır. Tüm veriler doğru olarak sınıflandırılmış olursa bu değer 1 (maksimum) olur. Verilerden hiçbiri doğru olarak sınıflandırılmamış ise bu değer 0 (minimum) olur.

$$\text{sınıflandırma doğruluğu} = \frac{TN+TP}{N} \quad (4.1)$$

Tamamlayıcı hata oranı; yanlış olarak tahmin edilen verilerin tüm veri adedine oranıdır. Tüm veriler yanlış olarak sınıflandırılmış olursa bu değer 1 (maksimum) olur. Verilerden hiçbiri yanlış olarak sınıflandırılmamış ise bu değer 0 (minimum) olur.

$$Hata\ oranu = \frac{FN+FP}{N} \quad (4.2)$$

$$Hata\ oranu = 1 - \frac{TN+TP}{N} \quad (4.3)$$

Duyarlılık (recall); doğru olarak sınıflandırılan pozitiflerin tüm pozitif verilere oranıdır. Pozitiflerin tamamı doğru olarak sınıflandırılmış olursa bu değer 1 (maksimum) olur. Pozitiflerin hiçbiri doğru olarak sınıflandırılmamış olursa bu değer 0 (minimum) olur.

$$Duyarlılık = \frac{TP}{TP+FN} \quad (4.4)$$

Kesinlik (precision); pozitif tahmin değeridir. Doğru olarak sınıflandırılan pozitiflerin, sınıflandırılan tüm pozitif sınıflandırılan verilere oranıdır. Maksimum kesinliğe ulaşmak için tüm pozitiflerin doğru ve hiçbir negatifin de hatalı sınıflandırılmamış olması gerekir.

$$Kesinlik = \frac{TP}{TP+FP} \quad (4.5)$$

Doğru pozitif oranı (TP Rate); modelin pozitif sınıfı doğru olarak sınıflandırma başarısıdır. Aslında, Recall ile aynı değerdir.

$$Doğru\ pozitif\ oranı(TPR) = \frac{TP}{TP+FN} \quad (4.6)$$

Yanlış pozitif oranı (FP Rate); modelin pozitif olarak sınıflandırdığı fakat gerçek değeri negatif olan verilerin tüm negatif değerlere oranıdır. Tüm negatiflerin tamamı hatalı sınıflandırıldığında bu oran 1 (maksimum) olur. Negatiflerin hiçbiri yanlış olarak sınıflandırılmamış olursa bu oran 0 (minimum) olur.

$$Yanlış\ pozitif\ oranı(FPR) = \frac{FP}{FP + TN} \quad (4.7)$$

F-ölçütü, duyarlılık ve kesinlik değerlerinin harmonik ortalaması olup bu iki ölçütü aynı anda ele alma imkanı vermektedir.

$$F - \text{Ölçütü} = 2 \times \frac{\text{Kesinlik} \times \text{Duyarlılık}}{\text{Kesinlik} + \text{Duyarlılık}} \quad (4.8)$$

Kappa istatistiği; iki veya daha fazla gözlemci arasındaki uyumun güvenilirliğini ölçen bir istatistik yöntemidir.

Kappa değeri -1 ile +1 arasında değer alabilir ve bulunan değer şu şekilde yorumlanır:³⁶

$\kappa = 1$ ise iki gözlemcinin sonuçları tümüyle birbiri ile uyumludur.

$\kappa = 0$ ise iki gözlemci arasındaki uyum sadece şansa bağlıdır.

$\kappa = -1$ ise iki gözlemci tümüyle birbirinin tersini değerlendirmektedir.

Kılıç³⁷; kappa katsayısı hesaplanırken iki farklı olasılıktan bahsetmiştir. Bunlar $P_r(a)$ ve $P_r(e)$ 'dir. $P_r(a)$ iki değerlendirici için gözlemlenen uyumların toplam orantısı iken, $P_r(e)$ bu uyumun şansa bağlı ortaya çıkma olasılığıdır ve formül şu şekildedir:

$$\kappa = \frac{P_r(a) - P_r(e)}{1 - P_r(e)} \quad (4.9)$$

Elde edilen κ (kappa) değerlerini yorumlamak için Landis ve Koch³⁸ tarafından şu tablo sunulmuştur:

Tablo 2: Kappa İstatistik Değeri Yorumları

κ değeri	Yorum
< 0	Şansa bağlı olabilecek uyumdan daha kötü uyum olması
0.01 – 0.20	Önemsiz düzeyde uyum olması
0.21 – 0.40	Zayıf düzeyde uyum olması
0.41 – 0.60	Orta düzeyde uyum olması
0.61 – 0.80	İyi düzeyde uyum olması
0.81 – 1	Çok iyi düzeyde uyum olması

³⁶ Sim, J., C.C., Wright. (2005), s.259

³⁷ Kılıç, S. (2015). s.142

³⁸ Landis J.R., G.G., Koch. (1977). s.165

1.2.3.3. Model Geçerliliğini Doğrulama

Modelin doğruluk oranı sınıflandırmadaki performans ile ilgileniliyorsa tahminlenen değerler üzerinden, eğer modelin performansı ile ilgileniyorsa gerçek değerler üzerinden hesaplanmaktadır.

Basit Doğrulama:

Basit doğrulama yöntemi büyük veri setleri için kullanılmaktadır. Bu yöntemde tipik olarak verilerin %5 ile % 33 arasındaki bir kısmı test verileri olarak ayrılır ve kalan kısım üzerinde modelin öğrenimi gerçekleştirildikten sonra, bu veriler üzerinde test işlemi yapılır.³⁹ Sınıflama modelinde yanlış olarak sınıflanan veri sayısının, tüm veri sayısına bölünmesiyle hata oranı, doğru olarak sınıflanan veri sayısının tüm veri sayısına bölünmesi ile doğruluk oranı hesaplanmaktadır.

Çapraz Doğrulama:

Model sınırlı sayıda veri ile kurulmuş olduğunda, kullanılabilir yöntem çapraz doğrulamadır.⁴⁰ Bu yöntemde veri kümesi rasgele iki eşit parçaya (a ve b parçası) ayrılmaktadır. İlk olarak a parçası üzerinde model eğitimi ve b parçası üzerinde test işlemi; ikinci olarak ise b parçası üzerinde model eğitimi ve a parçası üzerinde test işlemi yapılarak elde edilen hata oranlarının ortalaması kullanılır.

K-katlı Çapraz Doğrulama:

Denetimli öğrenmede birkaç bin veya daha az satırdan oluşan küçük veri kümelerinde, verilerin k gruba ayrıldığı k-katlı çapraz doğrulama (k-fold cross validation) yöntemi tercih edilebilmektedir.⁴¹ Literatürde en çok tercih edilen k değeri 10'dur.⁴² K-katlı çapraz doğrulama için k-kere yöntem tekrarlanır. Her adımda veri kümesinin 1/k kadarı, daha önce test için kullanılmamış parçası, test için kullanılırken, geri kalan k-1'lik kısmı eğitim için kullanılır. Hangi parçadan başlandığının bir önemi yoktur.

³⁹ Akpınar, H. (2000). s.11

⁴⁰ Akpınar, H. (2000). s.12

⁴¹ Akpınar, H. (2000). s.12

⁴² Şeker, S.E. (2013).

SF (test, eğitim), sınıflandırma fonksiyonu; VK, veri kümesi; k, kaç parça kat kullanıldığı; t, veri kümesi üzerinden seçilen her bir test kümesi olmak üzere aşağıda formülize edilen sonuç, bütün sınıflandırma fonksiyonlarının performanslarının toplamının, k sayısına bölünerek ortalamasının alınmasıdır.⁴³

$t_i \in VK$ olmak üzere,

$$Sonuç = \frac{\sum_{i=0}^k SF(t_i, VK - t_i)}{k} \quad (5)$$

1.3. YAPILAN ÇALIŞMALAR

Nizam ve Akın, çalışmalarında veri dağılımının sınıflandırma algoritmasındaki başarı oranına etkisi olup olmadığı araştırmıştır. Gıda sektöründeki bazı firmaların farklı ürünlerine ait tweetleri kullanmışlardır. Dengeli ve dengesiz olmak üzere iki farklı veri seti kullanılmıştır ve olumlu, olumsuz ve tarafsız olmak üzere bu iki veri seti 3 ayrı sınıfa tek tek ayrılmıştır. Dengesiz veri seti için 1113 olumlu, 277 olumsuz ve 610 tarafsız veri olmak üzere toplam 2000, dengeli veri seti içinse 257 olumlu, 277 olumsuz ve 299 tarafsız veri olmak üzere toplam 824 tweetten oluşan veri kümesini incelemişlerdir. Weka kütüphanesinde yer alan, Naive Bayes (NB), Rassal Orman (RF), Sıralı Minimal Optimizasyon (SMO), Karar Ağacı (J48) ve IB1 (k=1 için En Yakın Komşu) sınıflandırma algoritmalarını kullanmışlardır. Model başarı ölçütleri olarak ise doğruluk (accuracy), kesinlik (precision), duyarlılık (recall) ve kappa istatistiği kullanmışlar ve bu ölçütlere göre sonuçları karşılaştırmışlardır. Başarı ölçütleri ve kappa istatistiği sonuçlarına göre dengeli veri seti, dengesiz veri setine kıyasla daha iyi performans göstermiştir. Veri dağılımının sınıflandırma algoritmalarının başarısı üzerindeki etkilerden biri olduğunu göstermişlerdir. En iyi performansı %72,33 doğruluk oranı ile SMO sınıflandırma algoritmasından elde etmişlerdir.⁴⁴

Nalçakan vd., çalışmalarında teknoloji sektöründeki farklı üç firma (Samsung, Apple, LG) hakkında paylaşılan tweetlerden oluşturulan ve her firma için ayrı bir veri seti kullanmıştır. Veri setlerinde bulunan tweetler el yordamı ile pozitif, negatif ve nötr olmak üzere üç sınıfa ayrılmıştır. Her firmanın veri seti için 500'er tweetten oluşturulmuş ve anlamlı bir veri seti

⁴³ Şeker, S.E. (2013).

⁴⁴ Nizam, H., S.S., Akın. (2014).

olması için olumlu, olumsuz, nötr tweet sayılarının birbirine yakın olmasına dikkat edilmiştir. WEKA yazılımını kullanan Nalçakan, veri setleri üzerinde NB (Naive Bayes), RF (Random Forest), LibSVM, J48, KStar sınıflandırma yöntemlerini uygulamıştır. Alınan sonuçlar değerlendirildiğinde, en iyi sonucu 3 firma içinde Naive Bayes (NB) algoritmasının verdiği görülmektedir. Apple için düzenlenen veri seti için %52.90, Samsung için düzenlenen veri seti için %56.53 ve LG için düzenlenen veri seti için %65.63 ile Naive Bayes algoritması en iyi sonucu vermiştir.⁴⁵

Meral ve Diri'nin çalışmasında dokuz farklı alan için belirlenen anahtar sözcükler ile elde edilen 8321 adet tweet mesajları her alan için el yordamıyla gönüllü kişilerce olumlu, olumsuz, nötr olarak etiketlenmiştir. Deneysel sonuçların elde edilmesinde Weka veri madenciliği aracı içerisinde yer alan yöntemler varsayılan parametre değerleri ile kullanılmıştır. Doğal dil işleme yöntemlerinden iki farklı yöntem olan kelime tabanlı yöntem ve n-gram tabanlı yöntem kullanmışlardır. Sınıflandırma yöntemlerinden ise Rastgele Orman, Naive Bayes ve Destek Vektör Makinesi'ni kullanmışlardır. Sınıflandırıcının başarısının artmasını hedefleyen Meral ve Diri, kolerasyon tabanlı özellik seçimi (CFS) kullanarak başarının %8'e kadar arttığını belirtmişlerdir. Kelime tabanlı ve farklı n-gram yöntemleri üzerinde denemeler yapan Meral ve Diri, her bir denemede Destek Vektör Makinesinin daha yüksek sonuç verdiğini görmüştür. Deneysel sonuçlara göre ise en yüksek başarı 3-gram yöntemi ile elde edildiği denemede Destek Vektör Makinesi %90 f-ölçüm değeri ile elde edilmiştir.⁴⁶

Duygu analizi alanında yapılan ilk temel çalışmalardan biri 2002 yılında Pang, Lee ve Vaithyanatham tarafından yapılmıştır. Bu çalışmada makine öğrenmesi yöntemlerinden unigram, bigram, Part of Speech (POS) ve ikili birleşimler kullanılarak metinler pozitif veya negatif olarak iki duygu sınıfına atanmıştır. Bu çalışmada İngilizce film yorumları veri seti olarak kullanılmıştır. Naive Bayes, Maximum Entropi ve SVM (Destek vektör makineleri) makine öğrenmesi algoritmaları kullanılarak veri setinin başarımları elde edilmiştir. Yapılan duygu analizi çalışmasında sınıflandırma açısından en iyi sonucu unigram özelliğine sahip

⁴⁵ Nalçakan, Y., Ş.S., Bayramoğlu, S., Tuna. (2015).

⁴⁶ Meral, M., B., Diri. (2014).

makine öğrenmesi yöntemi vermiş olup sınıflandırma algoritmalarından en iyi performansı %82.9 olarak SVM göstermiştir.⁴⁷

Go, Bhayani ve Huang (2006), twitter mesajlarını otomatik olarak sınıflandırmak için bir yaklaşım öne sürmüşler. Twitter mesajlarını pozitif ve negatif olarak iki sınıfta sınıflandırmışlardır. Twitter üzerinde uzaktan denetimli öğrenmeyi yöntemini kullanarak duygu analizi yapmayı amaçlamışlardır. Veri kümesi, 800.000 pozitif tweet, 800.000 negatif tweet olmak üzere 1.600.000 tweetten oluşmaktadır. Çalışmalarını İngilizce metinler üzerinde yapmışlardır. Tweetleri unigram, bigram ve unigram ile bigramı birleştirerek analiz etmişler ve elde ettikleri deneysel sonuçlara göre unigram kullanımı sonucu %82.2 doğruluk oranıyla SVM algoritması, bi-gram kullanımı sonucu %81.6 doğruluk oranıyla Naive Bayes algoritması, unigram ile bigramın birlikte kullanımı sonucu %83 doğruluk oranıyla Maximum Entropi algoritması en iyi performansı göstermiştir.⁴⁸

Sevindi (2013) tez çalışmasında, BeyazPerde.com'dan çeşitli filmler için yazılmış olan toplamda 2305 yorum elde etmiş ve bu yorumlar elle etiketlenerek 1057 pozitif yorum ve 978 negatif yorum elde etmiştir. Geriye kalan 270 yorum, hem pozitif hem negatif görüş içermesi nedeniyle kutbu belirlenemeyen, herhangi bir görüş içermeyen veya anlaşılamayan yorum değerlendirilmeye alınmamıştır. Pozitif ve Negatif türkçe film yorumlarını kullanarak duygu kutuplarını çeşitli makine öğrenmesi yöntemleri ile belirlemeye çalışmış ve bu yöntemleri karşılaştırmıştır. Makine öğrenmesi yöntemlerinden, Karar Ağacı (C4.5), k-En Yakın Komşu (KNN), Naive Bayes ve Destek Vektör Makineleri (SVM) yöntemlerini kullanmış olup, model başarımları ölçütleri olarak ise doğruluk, kesinlik, duyarlılık, F-ölçütü kullanmıştır. Makine öğrenmesi yaklaşımları için yapılan denemelerde, en iyi sonucu SVM sınıflandırıcısından almıştır. SVM sınıflandırıcısı ile kelimelerin köklerinin kullanılarak bilinmeyen kelimelerin elendiği durumda, n-gram boyu 1 alındığında 0,8061'lik bir F-ölçütü değeri elde etmiştir. N-gram boyu artırıldıkça bu skorun düştüğü görülmüştür. N-gram boyu 2 yapıldığında 0,7630 değeri, 3 yapıldığında ise 0,6829 değerini elde etmiştir.⁴⁹

⁴⁷ Pang, B., L., Lee, S., Vaithyanathan. (2002).

⁴⁸ Go, A., R., Bhayani, L., Huang. (2009).

⁴⁹ Sevindi, B.İ. (2013).

1.4. DESTEK ARAÇLARI

Bu başlık altında, duygu analizi çalışmalarında kullanılan açık kaynak kodlu veri madenciliği araçları ve doğal dil işleme araçları kısaca tanıtılmıştır.

1.4.1. Açık Kaynak Kodlu Araçlar

Günümüzde çok sayıda ve çeşitli açık kaynak kodlu yazılım ile makine öğrenmesi yöntemleri uygulanabilmektedir. Bu bölümde en yaygın kullanılan yazılımlardan kısaca söz edilecektir.

WEKA:

Yeni Zelanda Waikato Üniversitesinde geliştirilmiş olan Weka, veri madenciliği görevleri için makine öğrenmesi algoritmaları topluluğudur. Veri hazırlama, sınıflandırma, regresyon, kümeleme, birleşme kuralları madenciliği ve görselleştirme için araçlar içerir.⁵⁰

Weka, Java programlama dili kullanılarak geliştirilmiştir ve diğer java uygulamalarına kolayca entegre edilebilir. Ayrıca yeni algoritmalarla kolayca genişletilebilir. WEKA'nın tek başına grafiksel kullanıcı arayüzü deneyi kolaylaştırır. WEKA, Öznitelik İlişkisi Dosya Biçimi (ARFF) kullanıyor. Basit bir ASCII metin dosyası formatıdır. Hem sayısal hem de nominal özellikleri destekler.

Weka aşağıdaki özelliklere sahiptir:⁵¹

- Veritabanındaki analiz ve ön işleme özelliklerinin ve verinin doğruluğunu değerlendirme.
- Örnek setlerin uygun sınıflara bölünüp sınıf niteliklerinin tanımlanması
- Sınıflandırma için kullanılacak özelliklerin çıkarılması
- Öğrenme işleminde kullanılması için özelliklerin bir alt set olarak seçilmesi
- Seçilen veri seti için mümkün sapmaların araştırılması ve etkisinin nasıl önlenebileceği

⁵⁰ <https://www.cs.waikato.ac.nz/ml/weka/> (2019)

⁵¹ Patil, B. M., D., Toshniwal, R.C., Joshi. (2009). s.1354

- Örnek alt setin seçilmesi, örneğin makine öğrenmesi baz alınarak yapılması ve kayıtlanması
- Öğrenme işlemi için sınıflandırma algoritması programı
- Seçilen algoritmanın performansını tahmin etmek için bir test yöntemine karar verilmesi

RAPID MINER:

Rapid Miner, Dortmund Teknoloji Üniversitesi Yapay Zekâ biriminde geliştirilmiş bir veri madenciliği yazılımıdır ve Java programlama dili ile yazılmıştır. Rapid Miner veri madenciliği, metin madenciliği, makine öğrenmesi, tahmin edici analiz ve iş analizi amaçlarına yönelik olarak geliştirilmiştir.

RapidMiner eklentileri ile Veri Madenciliği'nin bütün yönleri için 400 den fazla operatör sunmaktadır.⁵² Makine öğrenme algoritmaları olarak destek vektör makinelerini içeren büyük sayıdaki öğrenme modelleri için sınıflandırma ve regresyon, Karar Ağaçları, Bayesian, Mantıksal Kümeler, İlişkilendirme Kuralları ve Kümeleme için bir çok algoritma (k-means, k-medoids, dbscan), WEKA'da olan her şey, veri ön işleme için ayırma, normalleştirme, filtreleme gibi özellikler, genetik algoritma, yapay sinir ağları, 3D ile verileri analiz etme gibi birçok özelliği bulunmaktadır.⁵³

KNIME(Konstanz Information Miner):

Knime, Konstanz Üniversitesi görsel veri madenciliği araştırma grubu tarafından Eclipse Rich Client Platform üzerinde Java ile yazılmış açık kaynak kodlu ücretsiz bir yazılımdır.⁵⁴

Knime ile dosyalardan veya veritabanlarından veri alışverişi yapılabilmektedir. Ayrıca veri ön işleme, veri gruplama, pivot, normalleştirme, örnekleme, bölümlenme fonksiyonlarına da sahiptir. Kullanıcıya görsel veri akışı sağlamakla beraber analiz adımlarının tamamını veya bir kısmı üzerinde seçim yapılarak yürütülmesini de sağlamaktadır.

⁵² <http://rapid-i.com/content/view/12/69/> (2019)

⁵³ Tekerek, A. (2011). s.163

⁵⁴ Yıldız, M., S.E., Şeker. (2016). s.16

R Programlama Dili:

R dili ilk olarak Yeni Zelanda'da bulunan Aucland Üniversitesi İstatistik Bölümü'nden Ross Ihaka ve Robert Gentleman tarafından yazılmıştır.⁵⁵ Daha sonra çeşitli araştırmacılar R dilini geliştirmek için bir araya gelerek 1997'de "R Core Team" isimli bir grup kurmuşlardır. R programlama dili günümüzde kullanılan biçimini bu grupta yer alan araştırmacıların katkısı ile almıştır.

R programlama dilinin bazı temel özellikleri şunlardır:⁵⁶

- Etkin veri işleme ve saklama özelliğine sahiptir.
- Dizi ve özellikle matris hesaplamalarında kullanılacak özel operatörler mevcuttur.
- Veri analizi için kullanılacak uyumlu ve bir arada kullanılabilen araçlar içerir.
- Veri çözümlemede kullanılacak grafiksel araçlara sahiptir.

1.4.2. Doğal Dil İşleme Araçları

Doğal dil işleme, yaygın olarak Natural Language Processing (NLP) olarak bilinen yapay zekâ ve dilbilimin bir alt kategorisidir. Doğal dillerin kurallı yapısının çözümlenerek anlaşılması ya da yeniden üretilmesi amacını taşımaktadır. Metin içerisindeki kelimelerin anlamlandırılabilmesi için çeşitli işlemlerden geçmesi gerekir.

- Normalization (metin normalleştirme), metni daha önce sahip olmadığı tek bir kanonik forma dönüştürme işlemidir.
- Lemmization, kelimenin sözlükteki doğru halinin tespit edilmesi işlemidir.
- Stemming (Kök Bulma), kelimelerin kök halinin bulunması işlemidir.

Doğal dil işleme, bilgisayarı kullanarak dil işlemeye odaklanır. Doğal dil işlemenin asıl görevi, doğal dili çözümleyip yorumlayacak bilgisayar sistemleri tasarlamak ve uygulamaktır.

⁵⁵ Özdemir, A.F., E., Yıldıztepe, M. Binar. (2010). s.293

⁵⁶ Özdemir, A.F., E., Yıldıztepe, M. Binar. (2010). s.294

Mevcutta kullanılan halka açık iki popüler Türkçe destekli NLP aracı bulunmaktadır: Zemberek ve ITU Turkish NLP Web Service.⁵⁷

Zemberek:

Zemberek, Türkçe için popüler bir açık kaynak kodlu NLP kütüphanesidir. Ahmet Afşın Akın ve Mehmet Dündar Akın liderliğindeki Zemberek ekibi tarafından geliştirilmiştir. Çerçeve yazım denetimi, morfolojik ayrıştırma, kökten çıkarma, kelime inşası, kelime önerme, yalnızca ASCII karakterleri kullanılarak yazılan sözcükleri dönüştürme ve heceleri çıkarma gibi temel NLP işlemlerini sağlamaktadır.⁵⁸

ITU Turkish NLP Web Service API:

ITU Turkish NLP Web Service API, İstanbul Teknik Üniversitesi'nde Natural Language Processing grubu tarafından geliştirilen Türk dil işleme araçları ve uygulama programlama arayüzlerini (API) içermektedir.⁵⁹ ITU Türk NLP Web Servisi, “hizmet olarak yazılım” esasına dayanan ve birçok katmanda en gelişmiş NLP araçlarını sunan, kamuya açık bir NLP kütüphanesidir: ön işleme, morfoloji, sözdizimi ve varlık tanıma olmak üzere bu servis tarafından sağlanan bileşenler dört katmanda gruplanabilmektedir.

⁵⁷ Kalender, M., E.E., Korkmaz. (2017). s.2391

⁵⁸ Akın, A.A., M.D., Akın. (2007). s.1

⁵⁹ Yüksel, A.S., F.G., Tan. (2018).

2. E-TİCARET MARKALARINA YÖNELİK SOSYAL MEDYA YORUMLARININ ANALİZİ

Bu bölümde, e-ticaret firmalarına yönelik sosyal medya yorumlarının sınıflandırılması ile ulaşılmak istenen amaca değinilip daha sonra analizlerimizde kullandığımız veri kümesi tanıtılmıştır. Gerçekleştirdiğimiz çalışmada sınıflandırma yöntemleri ile analizler yapıldıktan sonra elde edilen bulgular paylaşılmış ve sonuçlar tartışılmıştır.

2.1. AMAÇ VE KAPSAM

İnternet kullanan bireylerin büyük bir kısmı sosyal medya platformlarında zaman geçirmektedir.⁶⁰ Satış sonrası süreçlerde birçok müşteri sosyal medya aracılığı ile aldıkları ürünleri/hizmetleri yorum yazarak değerlendirebilmektedir.

Sosyal medya, geleneksel yöntemler ile derlenmesi güç olan milyonlarca veriye hızlı ve kolay bir şekilde ulaşabilmeyi mümkün hale getirmiştir. Sosyal medya platformları ile kullanıcıların paylaştığı öznel metinler çoğalmış, bu da ilgili metinlerin analizi şeklinde bir araştırma alanı oluşturmuştur.⁶¹ Kullanıcıların/müşterilerin sosyal medya platformlarında istedikleri konuda paylaştıkları yorumların analiz edilmesi ile kullanıcıların görüş ve duyguları hakkında bilgi edinmek mümkündür.

Bu tez çalışmasındaki amacımız, duygu analizi sınıflandırma yöntemleri ile e-ticaret markalarına yönelik sosyal medya yorumlarının otomatik değerlendirilmesi ve analiz edilmesidir. Bu bağlamda kullanıcı veya müşteri yorumlarının otomatik değerlendirilmesi ile işletmelerde müşteri yorumlarının değerlendirilmesi için harcanan zamanın daha efektif kullanılabilmesi, işletmenin müşterilerine daha hızlı geri dönüş sağlayarak müşteri memnuniyetinin artırılması ve çalışan iş yükünün azaltılması amaçlanmıştır. Bu şekilde ulaşılabilecek bilgi ile işletmelerin iş süreçlerini iyileştirmeleri de mümkün olabilecektir.

⁶⁰ http://www.eticaretraporu.org/wp-content/uploads/2017/04/TUSIAD_E-Ticaret_Raporu_2017.pdf

⁶¹ Xia, R., C., Zong, S., Li, (2011).

2.2. VERİ KÜMESİ

Bu çalışmada kullanılan veri kümesi, “Trendyol, GittiGidiyor, Hepsiburada ve n11” markalarına yönelik müşteri yorumlarından oluşturulmuştur. Bu başlık altında başlangıç verilerinin nasıl oluşturulduğu anlatıldıktan sonra, bu verilerin analize hazır hale nasıl getirildiği anlatılmıştır.

2.2.1. Başlangıç Verileri

Bu çalışmada, sosyal medya aracı olarak günümüzün en çok kullanılan sosyal medya platformlarından Twitter seçilmiştir. Twitter’den veri çekme Twitter API (Application Programming Interface) kullanılarak yapılmıştır. Twitter verilerini çekmek (toplamak) üzere Twitter API’sine client üzerinden istekte bulunan npm js (node package modules javascript) modüllerinden twit modülü kullanılmıştır. Bu modülün amacı girilen anahtar kelimelere göre Twitter’den kullanıcıların attığı sosyal medya yorumlarını çekmektir.

Npm.js twit modülü kullanmadan önce Twitter üzerinde bir app oluşturulması gereklidir. Twitter üzerinde yeni bir app⁶² oluşturulduktan sonra bu app’e bağlı olan “consumer_key, consumer_secret, access_token, access_token_secret” değerlerinin twit modülüne bağlanarak yetkilendirme yapılması gereklidir. Twit modülü yüklendikten (npm intall twit) sonra bu modülün kullanılabilmesi için daha önceden oluşturulmuş bir twitter app’inin access key (consumer_key, consumer_secret, access_token, access_token_secret) değerleri twit modülüne bağlanarak yetkilendirme yapılmıştır. Bu yetkilendirmenin ardından npm.js de belirtilen T.get 'search/tweets' metodu⁶³ kullanılarak sosyal medya yorumları (tweet⁶⁴) Twitter API’sinden çekilmiştir.

Oluşturulmuş olan yapıda 5 Aralık 2018 – 19 Aralık 2018 tarihleri arasında yazılmış, “trendyol, n11, gittigidiyor, hepsiburada” anahtar kelimelerini barındıran sosyal medya yorumları, Twitter API’den çekilerek başlangıç verileri elde edilmiştir. Başlangıç verilerinin depolanması için npm js (node package modules javascript) modüllerinden mongodb modülü

⁶² <https://developer.twitter.com/apps>

⁶³ <https://www.npmjs.com/package/twit>

⁶⁴ Tweet, Twitter üzerindeki kayıtlı kullanıcıların yazdıkları mesajlara/gönderilere verilen kısa addır.

kullanılmıştır. Bu modülün amacı MongoDB⁶⁵ veri tabanına bağlantı kurmak ve verilerimizi bu veritabanına kaydetmemizi sağlamaktır. Npm.js de belirtilen metotlar⁶⁶ kullanılarak Twitter API'sinden çekilen sosyal medya yorumları MongoDB veritabanına kaydedilmiştir.

Depolanmış olan veriler içerisinde full_text olarak toplamda 5854 adet sosyal medya yorumu (tweet) yer almıştır. Bu tweetler içerisinde bulunan RT-retweet'ler⁶⁷, satıcıların paylaşmış olduğu ürün reklam paylaşımları, e-ticaret firmalarının kullanıcılarına herhangi bir konuda cevap olarak yazmış oldukları iletiler başlangıç verilerine dâhil edilmemiştir. Kullanıcıların/müşterilerin sadece ürün/hizmet ile ilgili olan kişisel yorumlarının başlangıç verilerine dâhil edilmesi amaçlanarak ve geriye kalan 4168 adet sosyal medya mesajı ile başlangıç verileri oluşturulmuştur.

2.2.2. Verilerin Hazırlanması

Verilerin hazırlanması sürecinde, başlangıç verilerinin nasıl etiketlendiği, nasıl ve hangi yöntemlerle temizlendiği ve verilerin özniteliklerinin nasıl çıkarılarak seçildiği anlatılmıştır.

2.2.2.1. Veri Etiketleme

Veri kümesinde bulunan tüm sosyal medya yorumları (tweetler) Microsoft Excel 2010 programında el yordamı ile olumlu, olumsuz ve nötr olarak tek tek okunarak etiketlenmiştir. Etiketleme yapılırken sosyal medya yorumlarında bulunan açık bir şekilde duygu ifade eden “mutluyum, mutsuzum, sinirliyim, ...” vb. kelimelerden/ifadelerden yardım alınmıştır. Bu gibi ifadelerin yanı sıra duygu ifade eden mutlu, mutsuz, kızgın, ağlayan, şaşkın vb. anlamlara karşılık gelen emoji⁶⁸ de dikkate alınmıştır. Emojiler gibi simgeler ile ifade edilen “:), :(, :/, :D” vb duygu ifadeleri de dikkate alınmıştır. Öncelikli olarak gönderilerin açık bir şekilde ne anlam ifade ettiğine dikkat edilmiştir. Kullanıcıların e-ticaret markalarını veya aldıkları hizmeti/ürünü olumlu mu yoksa olumsuz mu eleştirdiklerine dikkat edilmiştir. Sosyal medya

⁶⁵ <https://www.mongodb.com/>

⁶⁶ <https://www.npmjs.com/package/mongodb>

⁶⁷ RT-retweet, bir kullanıcının yararlı ya da ilginç mesajını kendi hesabından tekrar yayımlamasıdır.

⁶⁸ Emojiler, elektronik mesajlarda ve web sitelerinde yer alan mesaj, diğer iletileri zenginleştirmek adına kullanılabilir ideografi ve smiley içeren uygulamadır.

yorumlarının herhangi bir duygusu olmadığında ya da duygusunun yönüne karar verilmediğinde Nötr olarak etiketleme yapılmıştır.

Sonuç olarak, 4168 adet sosyal medya yorumundan 329 tanesi olumlu, 2865 tanesi olumsuz, 974 tanesi nötr etiketine atanmıştır.

2.2.2.2. Veri Önişleme

Başlangıç verileri üzerinde aşağıda belirtilen önişlemler uygulanarak veri temizliği yapılmıştır:

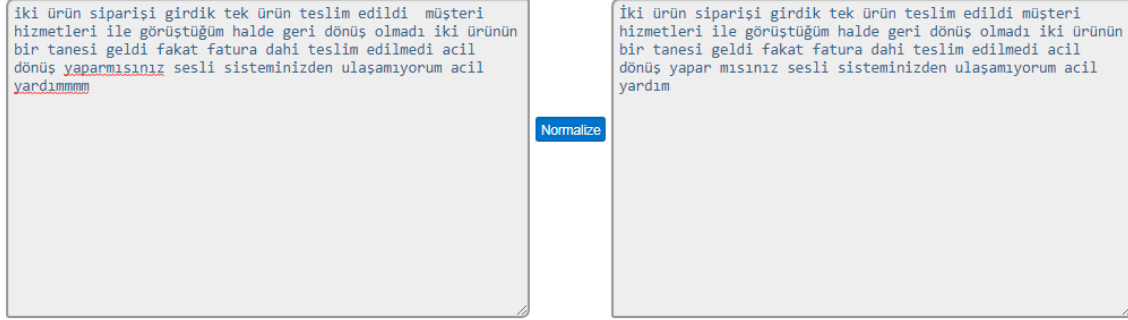
Metin Normalleştirme: ITU Turkish NLP Web Service'nin Normalization metodu kullanılarak başlangıç verileri içerisinde bulunan metinler normalleştirilmiştir. Metin normalleştirme sayesinde sosyal medya yorumlarındaki yanlış yazılan kelimeler bu metot yardımıyla düzenlenmiş ve metin normalleştirme gerçekleştirilmiştir.

Tablo 3: ITU Turkish NLP Web Service kullanılarak düzenlenmiş örnek sosyal medya yorumları

Sosyal Medya Yorumu	Düzenlenmiş Sosyal Medya Yorumu
tamam alicam soz allah kahretsin yaaaaa	Tamam alacağım söz Allah kahretsin ya
eyyyyyy çalışanlar yok mu dönüş allah ın kulu	Ey çalışanlar yok mu dönüş Allah in kulu
tiksindim yemin ediyorum reklamları heryerde bi bırakın yav	Tiksindim yemin ediyorum reklamları heryerde bir bırakın ya

ITU Turkish NLP Web Service arayüzü kullanılarak nasıl metin normalleştirildiği Şekil 6'da gösterilmiştir.

Normalization



Şekil 6: ITU Turkish NLP Web Service kullanılarak metin düzenleme

Normalleştirme aşamasından sonra metinler tek tek yeniden incelenerek ITU Turkish NLP Web Service kullanılarak düzenlenemeyen kelimeler manuel olarak düzenlenmiştir. Tablo 3’teki üçüncü cümlede bulunan yanlış yazılmış ve doğru şekilde normalleştirilemeyen “heryerde” kelimesi buna örnektir. Tablo 4’te manuel olarak düzenlenen örnek kelimeler verilmiştir.

Tablo 4: Manuel olarak düzenlenen örnek kelimeler

Metin içindeki kelime	ITU Turkish Web Service normalleştirme sonucu	Manuel olarak düzenlenmiş hali
heryerde	heryerde	her yerde
bide	bide	bir de
başlıyo	başlama	başlıyor
deniycem	deniysem	deneyeceğim
saolsun	salsın	sağolsun
sarfetmeden	sarfetmeden	sarf etmeden

Durak kelimelerin çıkarılması: Uygulamada kullanılan sosyal medya yorumlarının içerisinde yer alan fakat sınıflandırma işleminde bir anlamı olmayıp türkçede sık kullanılan zarflar, edatlar, zamirler yani tek başına anlam ifade etmeyen durak kelimeler temizlenmiştir.

Bu durak kelimeler arasında, Fazlı Can ve arkadaşları tarafından 2008 yılında yayınlanan “Information Retrieval on Turkish Texts” adlı makalede Türkçe dili için oluşturulmuş olan 147 durak kelime ile “merhaba, iyi geceler, iyi günler, günaydın,...” vb. ifadelerde yer almaktadır.

Tablo 5: Türkçe Durak Kelimeler

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

Kaynak: Can, F.K. vd. (2008). s.407-421.

Dönüştürme: Dönüştürme aşamasında aşağıdaki işlemler yapılmıştır.

- Tweetler HTML ve XML etiketlerinden temizlenmiştir.
- Tüm tweetler küçük harfe dönüştürülmüştür.
- Tüm Türkçe karakterler İngilizce karakterlere dönüştürülmüştür.
- Tweetlerin içinde bulunan tüm kullanıcı adı ve hashtag’ler silinmiştir.
- Sosyal medya genellikle informal bir dile sahip olduğu için “asdfghjkl” gibi kelime anlamı olmayan ifadeler yer almaktadır. “asdfghjkl, sajfhasf, sdfuhjsdjkdsjx...” gibi ifadelerin tümü silinmiştir.
- “ulen, ulan, bok” gibi argo ifadeler kaldırılmıştır.
- Veri seti içerisindeki bir ve iki harf uzunluğunda olan kelimeler silinmiştir.

Tarama ve İşaretleme: Tweetler içerisinde bulunan tüm noktalama işaretleri, tüm emojiler, tüm semboller, simgeler, sayılar silinmiştir.

Kök Bulma: Bu çalışmada kök bulma işlemi yapılmamıştır.

2.2.2.3. Veri Kümesinin Seçilmesi

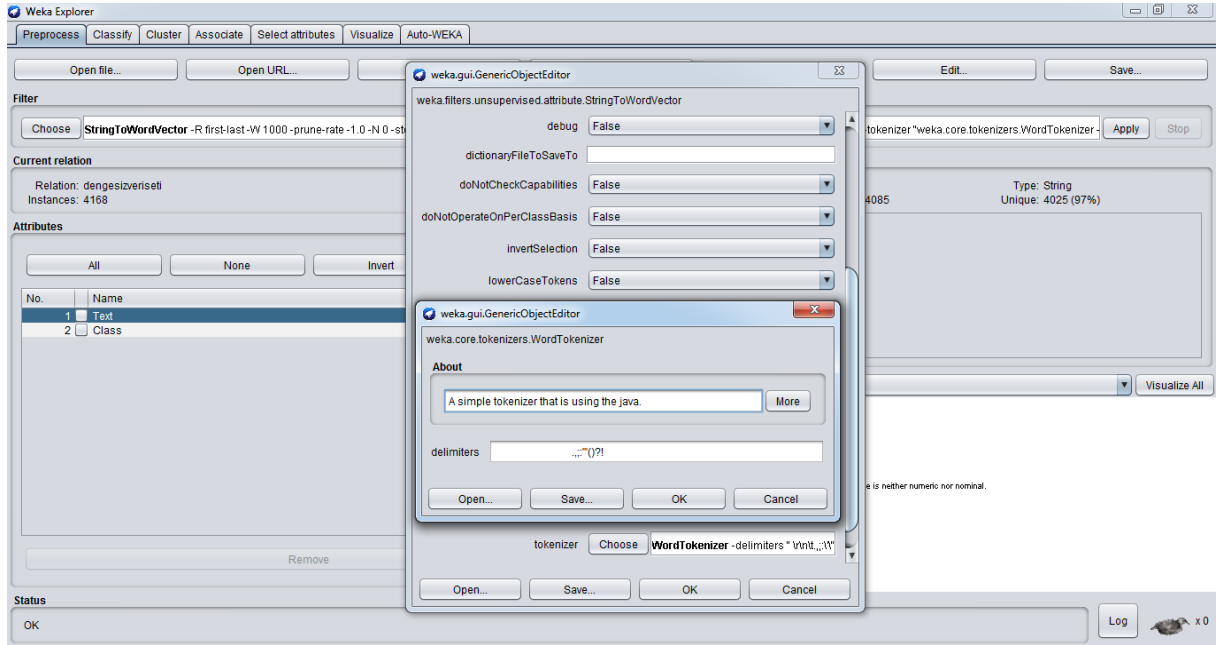
Bu çalışmanın dengeli ve dengesiz olmak üzere iki ayrı veri kümesi üzerinde yapılması ve her iki veri kümesine de sadece olumlu ve olumsuz etiketlerin dahil edilmesi kararlaştırılmıştır.

Dengesiz Veri Kümesi: Dengesiz veri kümesinde herhangi bir seçim olmaksızın tüm olumlu ve olumsuz sosyal medya yorumları veri kümesine dahil edilmiştir. Bu bağlamda dengesiz veri kümesi 329 Olumlu ve 2865 Olumsuz olmak üzere toplam 3194 sosyal medya yorumundan oluşmaktadır. Veri dağılımı dengesizdir.

Dengeli Veri Kümesi: Dengeli veri kümesinde olumlu ve olumsuz sosyal medya yorum sayısının aynı(dengeli) olmasına dikkat edilmiştir. Buradaki amaç veri dağılımını dengeli yapmak olmuştur. 2865 Olumsuz sosyal medya yorumu içerisinde 329 adet Olumsuz sosyal medya yorumu seçilmiştir. Bu seçim Microsoft Excel 2010 kullanılarak her sosyal medya yorumuna rasgele sayı atandıktan sonra bu sayılar sıralanarak ilk 329 Olumsuz sosyal medya yorumu seçilmiştir. Bu bağlamda dengeli veri kümesi 329 Olumlu ve 329 Olumsuz olmak üzere toplam 658 sosyal medya yorumundan oluşmaktadır.

2.2.2.4. Veri Kümesini Özniteliklerine Ayırma (Tokenization)

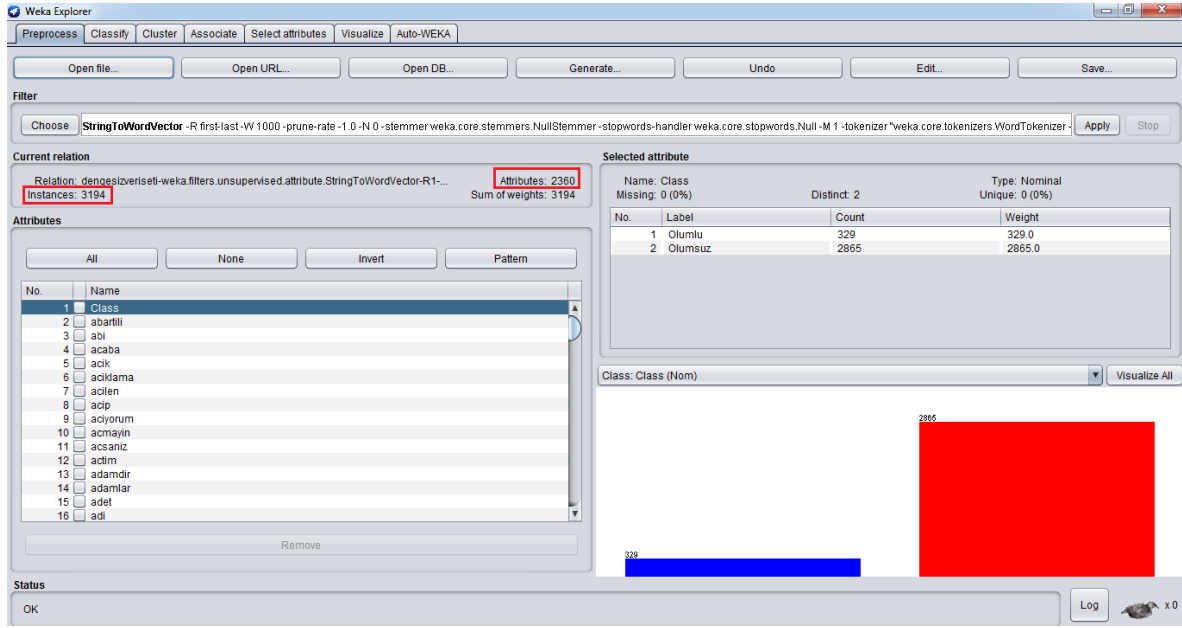
Dengeli ve Dengesiz veri kümelerini özniteliklerine ayırmak için Weka 3.8 programının StringToWordVector filtresi kullanılmıştır. Filtre yapılırken ayraç olarak boşluk kullanılmıştır.



Şekil 7: Veri Kümesini Özniteliklerine Ayırma

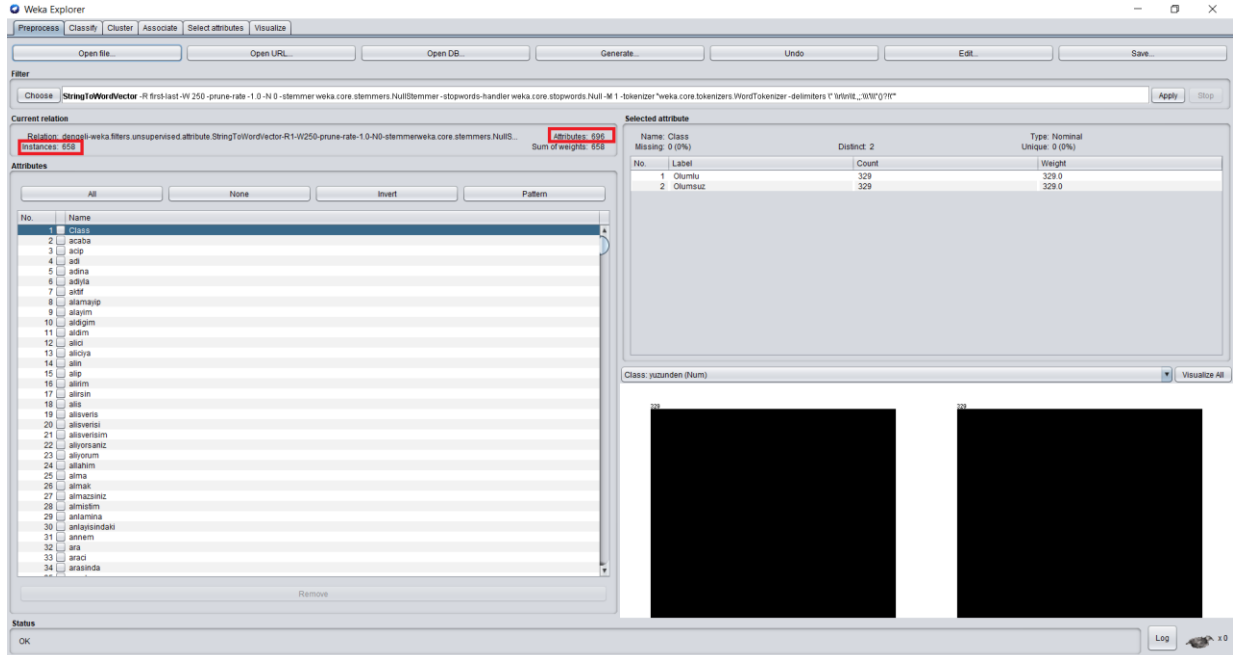
StringToWordVector filtresi içerisindeki wordtokeep sınıf başına kaç tane öznitelik tutulacağını belirlemektedir. Öznitelikler her sınıf için sıklıklarına göre sıralandıktan sonra yalnızca sık kullanılan öznitelikler tutulmaktadır. Her sınıf için bu işlem ayrı ayrı yapıldıktan sonra öngörülen öznitelikler birleştirilerek Weka 3.8 programının önyüzünde bulunan Attributes alanının altında sıralanmaktadır. Weka 3.8 programı wordtokeep değerini default olarak 1000 vermektedir.

Dengesiz veri kümesi wordtokeep değeri 1000(default) kabul edilerek özniteliklerine ayrılmıştır. Dengesiz veri kümesi özniteliklerine ayrıldığında 3194 sosyal medya yorumunu 2360 özniteliğe ayırdığı görülmektedir.



Şekil 8: Dengesiz Veri Kümesinin Özniteliklerine Ayrılması

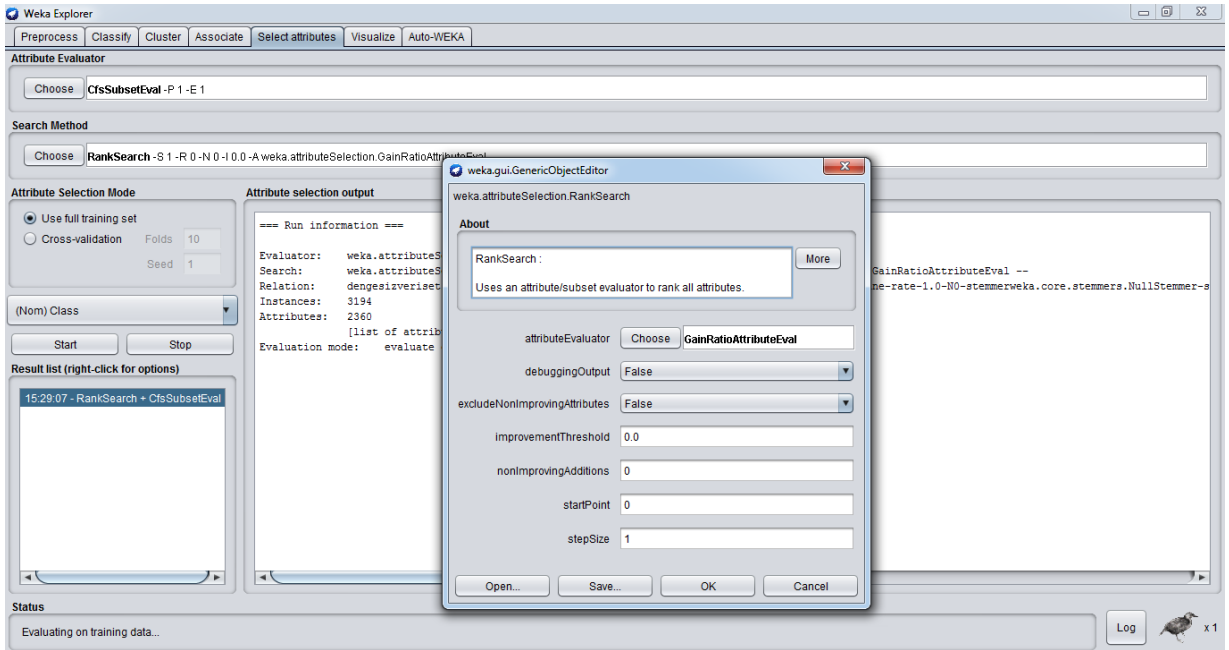
Dengeli veri kümesi için wordtokeep değeri 1000(default) olduğunda öngörülerin olumsuz etkilendiği görüldüğünden wordtokeep değeri 250 kabul edilerek şekilde özniteliklerine ayrılmıştır. Dengeli veri kümesi özniteliklerine ayrıldığında 658 sosyal medya yorumunu 696 öznitelige ayırdığı görülmektedir.



Şekil 9: Dengeli Veri Kümesinin Özniteliklerine Ayrılması

2.2.2.5. Öznitelik Seçimi

Öznitelik seçimi aşaması, açık kaynak kodlu bir uygulama olan WEKA 3.8 programının “Select attributes” metodu kullanılarak yapılmıştır. WEKA yazılımının kurulumu içerisinde doğrudan bulunmayan öznitelik alt kümesi seçme algoritmaları, WEKA yazılımının Package Manager aracı içerisindeki “attributeSelectionSearchMethods” paketi yüklenerek temin edilmiştir. Sezer⁶⁹ tarafından yapılmış olan tez çalışmasında en iyi sonucu veren CfsSubSetEval – GainRatioAttributeEval yöntemi ve arama yöntemi olarak da RankSearch kullanılmıştır. Dengesiz ve Dengeli olmak üzere iki veri seti içinde ayrı ayrı öznitelik seçimi yapılarak sınıflandırma analizi yapılmıştır.



Şekil 10: WEKA Ara Yüz Görüntüsü - Öznitelik Seçimi Yöntemi

2.3. MODELLEME

Modellerde, makine öğrenmesi temelli denetimli öğrenme işlemi yapılmıştır. Sınıflandırma yöntemlerinden Naive Bayes, SMO ve farklı ölçütler ile kNN (k=1) algoritması kullanılmıştır.

⁶⁹ Sezer, E. (2018). s.46

Tüm modelleme işlemleri WEKA 3.8 yazılımı kullanılarak yapılmıştır. Model geçerliliğini doğrulamak için k-katlı çapraz doğrulama yöntemi kullanılmıştır. k değeri literatürde en çok kullanılan 10 olarak kabul edilmiştir. WEKA yazılımının sınıflandırma sonucu olarak sunduğu hata matrisi ve birinci bölümde 1.2.3.2.Değerlendirme Ölçütleri başlığı altında anlatılmış olan tüm değerlendirme ölçütleri kullanılarak modeller değerlendirilmiştir.

2.3.1. Dengesiz Veri Kümesi ile Analiz

329 Olumlu ve 2865 Olumsuz sosyal medya yorumundan oluşan dengesiz veri kümesi öznitelik seçimi yapılmadan ve öznitelik seçimi yapılarak iki ayrı yöntem ile analiz edilmiştir.

2.3.1.1. Dengesiz Veri Kümesi - Öznitelik Seçiminin Yapılmadığı Modeller

Dengesiz veri kümesi, öznitelik seçimi yapılmaksızın 2360 özniteliğin tümü sınıflandırılarak üç ayrı modelde değerlendirilmiştir.

Model 1.1.1: Naive Bayes Sınıflandırma Yöntemi ile Analiz

Dengesiz veri kümesi 2360 öznitelik ile WEKA 3.8. yazılımında Naive Bayes sınıflandırma algoritması kullanılarak sınıflandırıldığında elde edilen hata matrisi Tablo 6’da sunulmuştur.

Tablo 6: Model 1.1.1. Hata Matrisi

Tahmin Değerleri

		Olumlu	Olumsuz	Toplam
Gerçek Değerler	Olumlu	114	215	329
	Olumsuz	63	2802	2865
	Toplam	177	3017	3194

Model 1.1.1. ile elde edilen model değerlendirme ölçütleri ise Tablo 7’de sunulmuştur.

Tablo 7: Model 1.1.1. Değerlendirme Ölçütleri

Sınıflandırma Doğruluğu(%):	91,30
Hata Oranı(%):	8,70
Duyarlılık:	0,91
Kesinlik:	0,90
TPR:	0,91
FPR:	0,59
F-Ölçütü:	0,90
Kappa İstatistiği:	0,41

Model 1.1.2: SMO Sınıflandırma Yöntemi ile Analiz

Dengesiz veri kümesi 2360 öznitelik ile WEKA 3.8. yazılımında SMO sınıflandırma algoritması kullanılarak sınıflandırıldığında elde edilen hata matrisi Tablo 8’de sunulmuştur.

Tablo 8: Model 1.1.2. Hata Matrisi

Tahmin Değerleri

		Olumlu	Olumsuz	Toplam
Gerçek Değerler	Olumlu	128	201	329
	Olumsuz	50	2815	2865
	Toplam	178	3016	3194

Model 1.1.2. ile elde edilen model değerlendirme ölçütleri ise Tablo 9’de sunulmuştur.

Tablo 9: Model 1.1.2. Değerlendirme Ölçütleri

Sınıflandırma Doğruluğu(%):	92,14
Hata Oranı(%):	7,86
Duyarlılık:	0,92
Kesinlik:	0,91
TPR:	0,92
FPR:	0,55
F-Ölçütü:	0,91
Kappa İstatistiği:	0,47

Model 1.1.3: kNN (k=1) Sınıflandırma Yöntemi ile Analiz

kNN algoritması WEKA içerisinde IBk olarak adlandırılmaktadır. k değeri literatürde en çok kullanılan 1 olarak kabul edilmiştir. kNN modeli Chebyshev ve Öklid olmak üzere iki farklı uzaklık ölçütü kullanılarak analiz edilmiştir.

Model 1.1.3.1. Chebyshev Uzaklık Ölçütü ile Sınıflandırma

Dengesiz veri kümesi 2360 öznitelik ile WEKA 3.8. yazılımında IB1 sınıflandırma algoritması kullanılarak Chebyshev uzaklık ölçütü ile sınıflandırıldığında elde edilen hata matrisi Tablo 10’da sunulmuştur.

Tablo 10: Model 1.1.3.1. Hata Matrisi

		Tahmin Değerleri		
		Olumlu	Olumsuz	Toplam
Gerçek Değerler	Olumlu	42	287	329
	Olumsuz	3	2862	2865
	Toplam	151	3043	3194

Model 1.1.3.1. ile elde edilen model değerlendirme ölçütleri ise Tablo 11’de sunulmuştur.

Tablo 11:Model 1.1.3.1. Deęerlendirme Ölçütleri

Sınıflandırma Doğruluęu(%):	90,92
Hata Oranı(%):	9,08
Duyarlılık:	0,91
Kesinlik:	0,91
TPR:	0,91
FPR:	0,78
F-Ölçütü:	0,88
Kappa İstatistięi:	0,20

Model 1.1.3.2. Öklid Uzaklık Ölçütü ile Sınıflandırma

Dengesiz veri kümesi 2360 öznitelik ile WEKA 3.8. yazılımında IB1 sınıflandırma algoritması kullanılarak Öklid uzaklık ölçütü ile sınıflandırıldığında elde edilen hata matrisi Tablo 12’da sunulmuştur.

Tablo 12:Model 1.1.3.2. Hata Matrisi

		Tahmin Deęerleri		
		Olumlu	Olumsuz	Toplam
Gerçek Deęerler	Olumlu	116	213	329
	Olumsuz	35	2830	2865
	Toplam	151	3043	3194

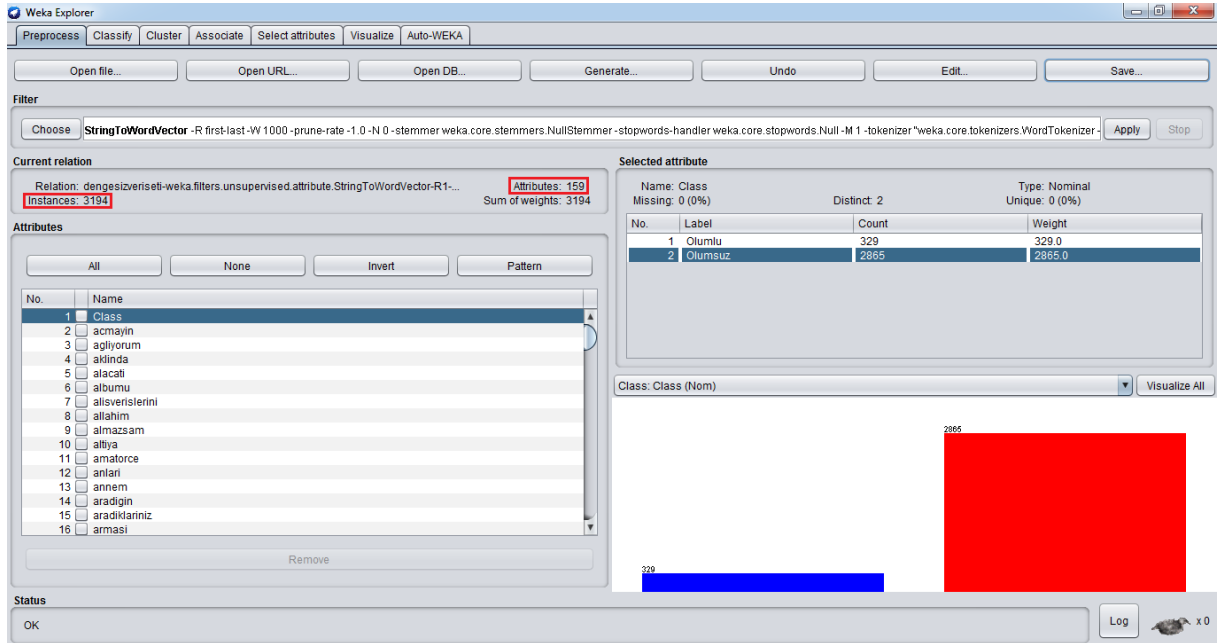
Model 1.1.3.2. ile elde edilen model deęerlendirme ölçütleri ise Tablo 13’de sunulmuştur.

Tablo 13:Model 1.1.3.2. Değerlendirme Ölçütleri

Sınıflandırma Doğruluğu(%):	92,24
Hata Oranı(%):	7,76
Duyarlılık:	0,92
Kesinlik:	0,91
TPR:	0,92
FPR:	0,58
F-Ölçütü:	0,91
Kappa İstatistiği:	0,45

2.3.1.2. Dengesiz Veri Kümesi - Öznitelik Seçiminin Yapıldığı Modeller

2.2.2.5.Öznitelik Seçimi başlığı altında belirtilen CfsSubSetEval-GainRatioAttributeEval yöntemi ile öznitelik seçimi yapılmış ve 2360 öznitelik arasından 158 öznitelik seçilmiştir. Dengesiz veri kümesi seçilen öznitelikler ile üç ayrı sınıflandırma algoritması kullanılarak sınıflandırılmıştır.



Şekil 11:Dengesiz Veri Kümesi Seçilen Öznitelikler

Model 1.2.1: Naive Bayes Sınıflandırma Yöntemi ile Analiz

Dengesiz veri kümesi seçilen 158 öznitelik ile WEKA 3.8. yazılımında Naive Bayes sınıflandırma algoritması kullanılarak sınıflandırıldığında elde edilen hata matrisi Tablo 14’da sunulmuştur.

Tablo 14:Model 1.2.1. Hata Matrisi

		Tahmin Değerleri		
		Olumlu	Olumsuz	Toplam
Gerçek Değerler	Olumlu	96	233	329
	Olumsuz	46	2819	2865
	Toplam	142	3052	3194

Model 1.2.1. ile elde edilen model değerlendirme ölçütleri ise Tablo 15’de sunulmuştur.

Tablo 15:Model 1.2.1. Değerlendirme Ölçütleri

Sınıflandırma Doğruluğu(%):	91,26
Hata Oranı(%):	8,74
Duyarlılık:	0,91
Kesinlik:	0,90
TPR:	0,91
FPR:	0,64
F-Ölçütü:	0,90
Kappa İstatistiği:	0,37

Model 1.2.2: SMO Sınıflandırma Yöntemi ile Analiz

Dengesiz veri kümesi seçilen 158 öznitelik ile WEKA 3.8. yazılımında SMO sınıflandırma algoritması kullanılarak sınıflandırıldığında elde edilen hata matrisi Tablo 16’da sunulmuştur.

Tablo 16:Model 1.2.2. Hata Matrisi

		Tahmin Değerleri		
		Olumlu	Olumsuz	Toplam
Gerçek Değerler	Olumlu	148	181	329
	Olumsuz	42	2823	2865
	Toplam	190	3004	3194

Model 1.2.2. ile elde edilen model değerlendirme ölçütleri ise Tablo 17’de sunulmuştur.

Tablo 17:Model 1.2.2. Değerlendirme Ölçütleri

Sınıflandırma Doğruluğu(%):	93,02
Hata Oranı(%):	6,98
Duyarlılık:	0,93
Kesinlik:	0,92
TPR:	0,93
FPR:	0,50
F-Ölçütü:	0,92
Kappa İstatistiği:	0,54

Model 1.2.3: kNN (k=1) Sınıflandırma Yöntemi ile Analiz

kNN algoritması WEKA içerisinde IBk olarak adlandırılmaktadır. k değeri literatürde en çok kullanılan 1 olarak kabul edilmiştir. kNN modeli Chebyshev ve Öklid olmak üzere iki farklı uzaklık ölçütü kullanılarak analiz edilmiştir.

Model 1.2.3.1. Chebyshev Uzaklık Ölçütü ile Sınıflandırma

Dengesiz veri kümesi seçilen 158 öznitelik ile WEKA 3.8. yazılımında IB1 sınıflandırma algoritması kullanılarak Chebyshev uzaklık ölçütü ile sınıflandırıldığında elde edilen hata matrisi Tablo 18’de sunulmuştur.

Tablo 18:Model 1.2.3.1. Hata Matrisi

		Tahmin Değerleri		
		Olumlu	Olumsuz	Toplam
Gerçek Değerler	Olumlu	148	181	329
	Olumsuz	26	2839	2865
	Toplam	150	3020	3194

Model 1.2.3.1. ile elde edilen model değerlendirme ölçütleri ise Tablo 19’de sunulmuştur.

Tablo 19:Model 1.1.3.1. Değerlendirme Ölçütleri

Sınıflandırma Doğruluğu(%):	93,52
Hata Oranı(%):	6,48
Duyarlılık:	0,94
Kesinlik:	0,93
TPR:	0,94
FPR:	0,49
F-Ölçütü:	0,93
Kappa İstatistiği:	0,56

Model 1.2.3.2. Öklid Uzaklık Ölçütü ile Sınıflandırma

Dengesiz veri kümesi seçilen 158 öznitelik ile WEKA 3.8. yazılımında IB1 sınıflandırma algoritması kullanılarak Öklid uzaklık ölçütü ile sınıflandırıldığında elde edilen hata matrisi Tablo 20’de sunulmuştur.

Tablo 20: Model 1.2.3.2. Hata Matrisi

		Tahmin Değerleri		
		Olumlu	Olumsuz	Toplam
Gerçek Değerler	Olumlu	150	179	329
	Olumsuz	28	2837	2865
	Toplam	178	3016	3194

Model 1.2.3.2. ile elde edilen model değerlendirme ölçütleri ise Tablo 21’de sunulmuştur.

Tablo 21: Model 1.2.3.2. Değerlendirme Ölçütleri

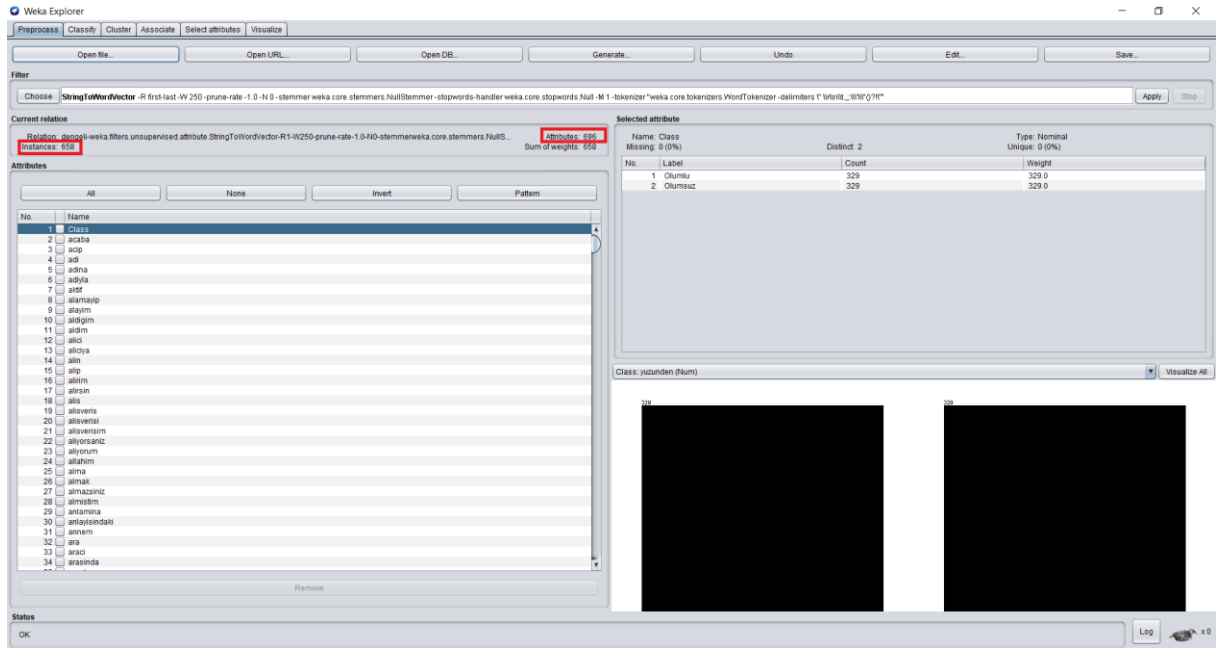
Sınıflandırma Doğruluğu(%):	93,52
Hata Oranı(%):	6,48
Duyarlılık:	0,94
Kesinlik:	0,93
TPR:	0,94
FPR:	0,49
F-Ölçütü:	0,93
Kappa İstatistiği:	0,56

2.3.2. Dengeli Veri Kümesi ile Analiz

329 Olumlu ve 329 Olumsuz yorumdan oluşan dengesiz veri kümesi öznitelik seçimi yapılmadan ve öznitelik seçimi yapılarak iki ayrı yöntem ile analiz edilmiştir.

2.3.2.1. Dengeli Veri Kümesi - Öznitelik Seçiminin Yapılmadığı Modeller

Dengeli veri kümesi, öznitelik seçimi yapılmaksızın 696 özniteliğin tümü sınıflandırılarak üç ayrı modelde değerlendirilmiştir.



Şekil 12: Dengeli Veri Kümesinin Özniteliklerine Ayrılması

Model 2.1.1: Naive Bayes Sınıflandırma Yöntemi ile Analiz

Dengeli veri kümesi 696 öznitelik ile WEKA 3.8. yazılımında Naive Bayes sınıflandırma algoritması kullanılarak sınıflandırıldığında elde edilen hata matrisi Tablo 22’da sunulmuştur.

Tablo 22:Model 2.1.1. Hata Matrisi

		Tahmin Değerleri		
		Olumlu	Olumsuz	Toplam
Gerçek Değerler	Olumlu	250	79	329
	Olumsuz	78	251	329
	Toplam	328	330	658

Model 2.1.1. ile elde edilen model değerlendirme ölçütleri ise Tablo 23’de sunulmuştur.

Tablo 23: Model 2.1.1. Değerlendirme Ölçütleri

Sınıflandırma Doğruluğu(%):	76,14
Hata Oranı(%):	23,86
Duyarlılık:	0,76
Kesinlik:	0,76
TPR:	0,76
FPR:	0,24
F-Ölçütü:	0,76
Kappa İstatistiği:	0,52

Model 2.1.2: SMO Sınıflandırma Yöntemi ile Analiz

Dengeli veri kümesi 696 öznitelik ile WEKA 3.8. yazılımında SMO sınıflandırma algoritması kullanılarak sınıflandırıldığında elde edilen hata matrisi Tablo 24’da sunulmuştur.

Tablo 24: Model 2.1.2. Hata Matrisi

Tahmin Değerleri

		Olumlu	Olumsuz	Toplam
Gerçek Değerler	Olumlu	279	50	329
	Olumsuz	77	252	329
	Toplam	356	302	658

Model 2.1.2. ile elde edilen model değerlendirme ölçütleri ise Tablo 25’de sunulmuştur.

Tablo 25: Model 2.1.2. Değerlendirme Ölçütleri

Sınıflandırma Doğruluğu(%):	80,70
Hata Oranı(%):	19,30
Duyarlılık:	0,81
Kesinlik:	0,81
TPR:	0,81
FPR:	0,19
F-Ölçütü:	0,81
Kappa İstatistiği:	0,61

Model 2.1.3: kNN (k=1) Sınıflandırma Yöntemi ile Analiz

kNN algoritması WEKA içerisinde IBk olarak adlandırılmaktadır. k değeri literatürde en çok kullanılan 1 olarak kabul edilmiştir. kNN modeli Chebyshev ve Öklid olmak üzere iki farklı uzaklık ölçütü kullanılarak analiz edilmiştir.

Model 2.1.3.1. Chebyshev Uzaklık Ölçütü ile Sınıflandırma

Dengeli veri kümesi 696 öznitelik ile WEKA 3.8. yazılımında IB1 sınıflandırma algoritması kullanılarak Chebyshev uzaklık ölçütü ile sınıflandırıldığında elde edilen hata matrisi Tablo 26’da sunulmuştur.

Tablo 26: Model 2.1.3.1. Hata Matrisi

		Tahmin Değerleri		
		Olumlu	Olumsuz	Toplam
Gerçek Değerler	Olumlu	291	38	329
	Olumsuz	289	40	329
	Toplam	580	78	658

Model 2.1.3.1. ile elde edilen model değerlendirme ölçütleri ise Tablo 27’de sunulmuştur.

Tablo 27:Model 2.1.3.1. Deęerlendirme Ölçütleri

Sınıflandırma Doğruluęu(%):	50,30
Hata Oranı(%):	49,70
Duyarlılık:	0,50
Kesinlik:	0,51
TPR:	0,50
FPR:	0,50
F-Ölçütü:	0,42
Kappa İstatistięi:	0,0061

Model 2.1.3.2. Öklid Uzaklık Ölçütü ile Sınıflandırma

Dengeli veri kümesi 696 öznitelik ile WEKA 3.8. yazılımında IB1 sınıflandırma algoritması kullanılarak Öklid uzaklık ölçütü ile sınıflandırıldığında elde edilen hata matrisi Tablo 28’da sunulmuştur.

Tablo 28:Model 2.1.3.2. Hata Matrisi

		Tahmin Deęerleri		
		Olumlu	Olumsuz	Toplam
Gerçek Deęerler	Olumlu	248	81	329
	Olumsuz	126	203	329
	Toplam	374	284	658

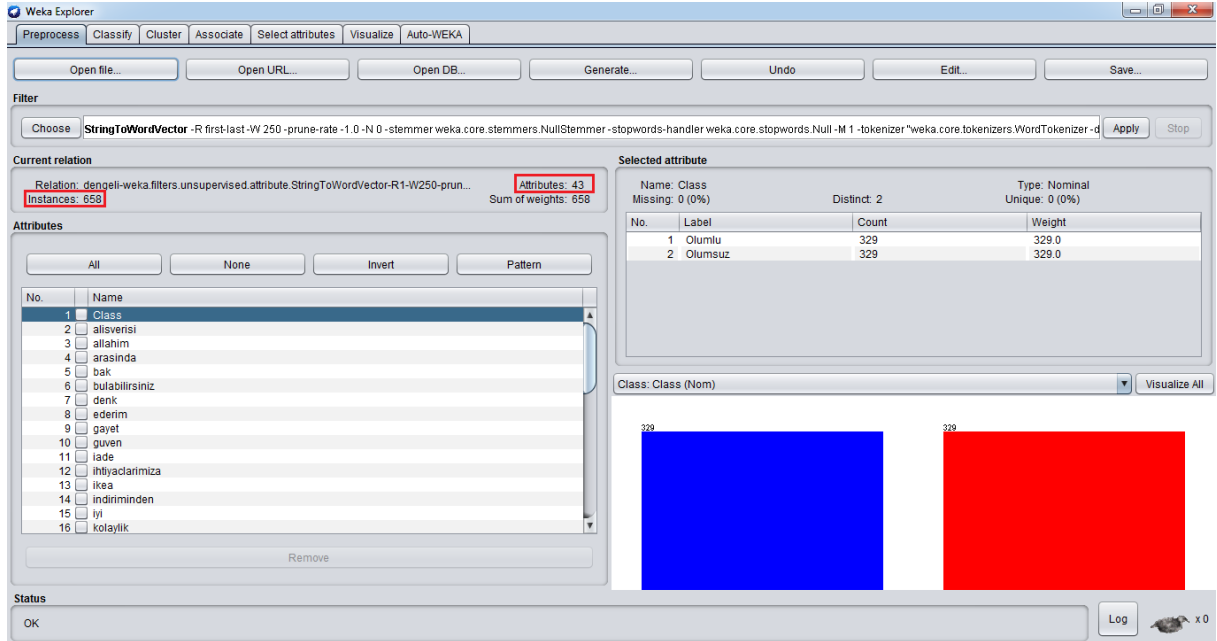
Model 2.1.3.2. ile elde edilen model deęerlendirme ölçütleri ise Tablo 29’de sunulmuştur.

Tablo 29:Model 2.1.3.2. Değerlendirme Ölçütleri

Sınıflandırma Doğruluğu(%):	68,54
Hata Oranı(%):	31,46
Duyarlılık:	0,69
Kesinlik:	0,69
TPR:	0,69
FPR:	0,32
F-Ölçütü:	0,68
Kappa İstatistiği:	0,37

2.3.2.2.Dengeli Veri Kümesi - Öznitelik Seçiminin Yapıldığı Modeller

2.2.2.5.Öznitelik Seçimi başlığı altında belirtilen CfsSubSetEval–GainRatioAttributeEval yöntemi ile öznitelik seçimi yapılmış ve 696 öznitelik arasından 42 öznitelik seçilmiştir. Dengeli veri kümesi seçilen öznitelikler ile üç ayrı sınıflandırma algoritması kullanılarak sınıflandırılmıştır.



Şekil 13: Dengeli Veri Kümesi Seçilen Öznitelikler

Model 2.2.1: Naive Bayes Sınıflandırma Yöntemi ile Analiz

Dengeli veri kümesi 42 öznitelik ile WEKA 3.8. yazılımında Naive Bayes sınıflandırma algoritması kullanılarak sınıflandırıldığında elde edilen hata matrisi Tablo 30'da sunulmuştur.

Tablo 30: Model 2.2.1. Hata Matrisi

Tahmin Değerleri

		Tahmin Değerleri		
		Olumlu	Olumsuz	Toplam
Gerçek Değerler	Olumlu	255	74	329
	Olumsuz	98	231	329
	Toplam	353	305	658

Model 2.2.1. ile elde edilen model değerlendirme ölçütleri ise Tablo 31'de sunulmuştur.

Tablo 31 : Model 2.2.1. Değerlendirme Ölçütleri

Sınıflandırma Doğruluğu(%):	73,86
Hata Oranı(%):	26,14
Duyarlılık:	0,74
Kesinlik:	0,74
TPR:	0,74
FPR:	0,26
F-Ölçütü:	0,74
Kappa İstatistiği:	0,48

Model 2.2.2: SMO Sınıflandırma Yöntemi ile Analiz

Dengeli veri kümesi 42 öznitelik ile WEKA 3.8. yazılımında SMO sınıflandırma algoritması kullanılarak sınıflandırıldığında elde edilen hata matrisi Tablo 32'da sunulmuştur.

Tablo 32 : Model 2.2.2. Hata Matrisi

		Tahmin Değerleri		
		Olumlu	Olumsuz	Toplam
Gerçek Değerler	Olumlu	311	18	329
	Olumsuz	145	184	329
	Toplam	456	202	658

Model 2.2.2. ile elde edilen model değerlendirme ölçütleri ise Tablo 33'de sunulmuştur.

Tablo 33 : Model 2.2.2. Değerlendirme Ölçütleri

Sınıflandırma Doğruluğu(%):	75,23
Hata Oranı(%):	24,77
Duyarlılık:	0,75
Kesinlik:	0,80
TPR:	0,75
FPR:	0,25
F-Ölçütü:	0,74
Kappa İstatistiği:	0,50

Model 2.2.3: kNN(k=1) Sınıflandırma Yöntemi ile Analiz

kNN algoritması WEKA içerisinde IBk olarak adlandırılmaktadır. k değeri literatürde en çok kullanılan 1 olarak kabul edilmiştir. kNN modeli Chebyshev ve Öklid olmak üzere iki farklı uzaklık ölçütü kullanılarak analiz edilmiştir.

Model 2.2.3.1. Chebyshev Uzaklık Ölçütü ile Sınıflandırma

Dengeli veri kümesi 42 öznitelik ile WEKA 3.8. yazılımında IB1 sınıflandırma algoritması kullanılarak Chebyshev uzaklık ölçütü ile sınıflandırıldığında elde edilen hata matrisi Tablo 34'da sunulmuştur.

Tablo 34 : Model 2.2.3.1. Hata Matrisi

		Tahmin Değerleri		
		Olumlu	Olumsuz	Toplam
Gerçek Değerler	Olumlu	315	14	329
	Olumsuz	173	156	329
	Toplam	488	170	658

Model 2.2.3.1. ile elde edilen model değerlendirme ölçütleri ise Tablo 35’de sunulmuştur.

Tablo 35:Model 2.2.3.1. Değerlendirme Ölçütleri

Sınıflandırma Doğruluğu(%):	71,58
Hata Oranı(%):	28,42
Duyarlılık:	0,72
Kesinlik:	0,78
TPR:	0,72
FPR:	0,28
F-Ölçütü:	0,70
Kappa İstatistiği:	0,43

Model 2.2.3.2. Öklid Uzaklık Ölçütü ile Sınıflandırma

Dengeli veri kümesi 42 öznitelik ile WEKA 3.8. yazılımında IB1 sınıflandırma algoritması kullanılarak Öklid uzaklık ölçütü ile sınıflandırıldığında elde edilen hata matrisi Tablo 36’da sunulmuştur.

Tablo 36:Model 2.2.3.2. Hata Matrisi

		Tahmin Değerleri		
		Olumlu	Olumsuz	Toplam
Gerçek Değerler	Olumlu	312	17	329
	Olumsuz	149	180	329
	Toplam	461	197	658

Model 2.2.3.2. ile elde edilen model değerlendirme ölçütleri ise Tablo 37’de sunulmuştur.

Tablo 37:Model 2.2.3.2. Değerlendirme Ölçütleri

Sınıflandırma Doğruluğu(%):	74,77
Hata Oranı(%):	25,23
Duyarlılık:	0,75
Kesinlik:	0,80
TPR:	0,75
FPR:	0,25
F-Ölçütü:	0,74
Kappa İstatistiği:	0,50

2.4. BULGULAR ve TARTIŞMA

2.3. Modelleme başlığı altında yapılan model çalışmalarında elde edilen tüm çıktılar Tablo 38’de verilmiştir.

Tablo 38: Model Sonuçlarının Karşılaştırılması

Sınıflandırma Yöntemi	Veri Kümesi	Öznitelek Seçimi	Sınıflandırma Doğruluğu (%)	Hata Oranı (%)	Duyarlılık	Kesinlik	TP Oranı	FP Oranı	F-Ölçütü	Kappa İstatistiği
Naive Bayes	Dengesiz	Yok	91,30	8,70	0,91	0,90	0,91	0,59	0,90	0,41
		Var	91,26	8,74	0,91	0,90	0,91	0,64	0,90	0,37
	Dengeli	Yok	76,14	23,86	0,76	0,76	0,76	0,24	0,76	0,52
		Var	73,86	26,14	0,74	0,74	0,74	0,26	0,74	0,48
SMO	Dengesiz	Yok	92,14	7,86	0,92	0,91	0,92	0,55	0,91	0,47
		Var	93,02	6,98	0,93	0,92	0,93	0,50	0,92	0,54
	Dengeli	Yok	80,70	19,30	0,81	0,81	0,81	0,19	0,81	0,61
		Var	75,23	24,77	0,75	0,80	0,75	0,25	0,74	0,50
kNN (IBk k=1) Chebyshev	Dengesiz	Yok	90,92	9,08	0,91	0,91	0,91	0,78	0,88	0,20
		Var	93,52	6,48	0,94	0,93	0,94	0,49	0,93	0,56
	Dengeli	Yok	50,30	49,70	0,50	0,51	0,50	0,50	0,42	0,0061
		Var	71,58	28,42	0,72	0,78	0,72	0,28	0,70	0,43
kNN (IBk k=1) Öklid	Dengesiz	Yok	92,24	7,76	0,92	0,91	0,92	0,58	0,91	0,45
		Var	93,52	6,48	0,94	0,93	0,94	0,49	0,93	0,56
	Dengeli	Yok	68,54	31,46	0,69	0,69	0,69	0,32	0,68	0,37
		Var	74,77	25,23	0,75	0,80	0,75	0,25	0,74	0,50

Dengesiz veri kümesinde yapılan tüm çalışmalarda elde edilen sınıflandırma doğruluğunun %90'nın üzerinde olduğu görülmektedir.

Dengesiz veri kümesinde öznitelik seçimi yapılmadığı durumda yapılan çalışmalarda ilk sırada %92,24 sınıflandırma doğruluğu ile kNN (IBk k=1) Öklid uzaklık ölçütüyle yapılan k-en yakın komşu algoritması bulunurken, ikinci sırada ise çok az bir fark ile %92,14 sınıflandırma doğruluğu ile SMO algoritması bulunmaktadır. Üçüncü sırada %91,30 ile Naive Bayes, son sırada ise %90,92 ile kNN (IBk k=1) Chebyshev uzaklık ölçütüyle yapılan k-en yakın komşu algoritması bulunmaktadır.

Dengesiz veri kümesinde öznitelik seçimi yapıldığı durumda yapılan çalışmalarda ise ilk sırada %93,52 sınıflandırma doğruluğu ile kNN(İBk k=1) Öklid ve Chebyshev uzaklık ölçütleriyle yapılan k-en yakın komşu algoritmaları bulunurken, ikinci sırada ise az bir fark ile %93,02 sınıflandırma doğruluğu ile SMO algoritması bulunmaktadır. Son sırada ise %91,26 sınıflandırma doğruluğu ile Naive Bayes bulunmaktadır.

Dengesiz veri kümesi üzerinde yapılan çalışmalarda öznitelik seçiminin başarı oranları üzerinde çok büyük bir değişiklik olmasada oranları pozitif etkilediği görülmektedir.

Dengeli veri kümesinde öznitelik seçimi yapılmadığı durumda yapılan çalışmalarda ilk sırada %80,70 başarı oranı ile SMO algoritması bulunurken, ikinci sırada ise %76,14 başarı oranı ile Naive Bayes algoritması bulunmaktadır. k-en yakın komşu algoritmasının Öklid uzaklık ölçütü ile %68,54, Chebyshev uzaklık ölçütüyle ise %50,30 olduğu dikkat çekmektedir.

Dengeli veri kümesinde öznitelik seçimi yapıldığı durumda yapılan çalışmalarda ilk sırada %75,23 başarı oranı ile SMO algoritması bulunurken, ikinci sırada ise %74,77 başarı oranı ile kNN (IBk k=1) Öklid uzaklık ölçütüyle yapılan k-en yakın komşu algoritması bulunmaktadır. Üçüncü sırada %73,86 ile Naive Bayes, son sırada ise %71,58 ile kNN (IBk k=1) Chebyshev uzaklık ölçütüyle yapılan k-en yakın komşu algoritması bulunmaktadır.

Dengeli veri kümesinde yapılan çalışmalarda öznitelik seçiminin k-en yakın komşu algoritmaları üzerinde Öklid uzaklık ölçütü ile yaklaşık %7'lik, Chebyshev uzaklık ölçütü ile yaklaşık %21'lik bir artış ile pozitif etkilediği görülmektedir.

Dengeli veri kümesinin sınıflandırma doğruluğunun dengesiz veri kümesine göre daha düşük olduğu görülmektedir. Bunun nedeninin dengeli veri kümesinde veri sayısının yetersiz olmasından kaynaklandığı tahmin edilmektedir.

Model başarımlarını ölçütlerinden recall(duyarlılık), precision(kesinlik) ve f-ölçüt değerlerinin 1'e yakın olması istenmektedir. Recall, Precision ve F-Ölçüt değerleri incelendiğinde, dengesiz veri kümesi üzerinde yapılan tüm modellerde algoritmaların recall, precision ve f-ölçüt değerlerinin %88 - %93 arasında değişkenlik gösterdiği görülmektedir. Dengeli veri kümesinde ise recall(duyarlılık), precision(kesinlik) ve f-ölçüt değerleri öznitelik seçimi yapılmadığı durumda %42 - %81 aralığında değişkenlik gösterdiği görülmektedir. öznitelik seçimi yapıldığı durumda ise %70 - %74 aralığında değişkenlik gösterdiği görülmektedir. Dengeli veri kümesinde bu değerlerin dengesiz veri kümesine göre daha düşük olmasının nedeninin veri sayısının yetersiz olmasından kaynaklandığı tahmin edilmektedir.

Kappa istatistik değerleri incelendiğinde ise, bu değerler veri kümeleri ve öznitelik seçimine göre değişkenlik göstermektedir. Veri kümeleri ve öznitelik seçimine göre ortalama kappa istatistik değerleri alındığında SMO algoritması ile en iyi uyumun olduğu görülmektedir.

SONUÇ

Duygu analizi ile müşterilerin satın aldıkları ürün/hizmetler hakkında sosyal medya platformlarında yazdıkları yorumlar incelenerek müşterilerin hangi duygu ile ilgili bu yorumları yazdığı araştırılabilmektedir. Kullanıcı veya müşteri yorumları otomatik değerlendirilerek işletmelerde bu yorumların manuel değerlendirilmesi için harcanan zaman daha efektif kullanılabilen ve çalışan iş yükü azaltılabilmektedir. Doğru sınıflandırılan yorumlar analiz edilerek müşteri memnuniyetinin artırılmasına yönelik çalışmalar yapılabilir.

Bu çalışmada, duygu analizi Weka 3.8 yazılımı kullanılarak yapılmıştır. El yordamı ile sosyal medya yorumlarının her biri “olumlu, olumsuz, nötr” etiketlerinden biri ile etiketlenmiştir. Dengesiz ve dengeli veri seti oluşturulduktan sonra Weka 3.8 yazılımı kullanılarak her iki veri seti ayrı analiz edilmiştir. Her iki veri kümesi Weka 3.8 de bulunan stringtowordvector filtresi kullanılarak özniteliklerine ayrılmıştır. Özniteliklerine ayrılan veri kümeleri, ilk olarak öznitelik seçimi yapılmadan sınıflandırma algoritmaları ile sınıflandırılarak algoritmalarının performansı karşılaştırılmıştır. Sonrasında her iki veri kümesi için de öznitelik seçimi yapılmış ve sınıflandırma algoritmaları ile sınıflandırılarak algoritmalarının performansı karşılaştırılmıştır.

Sınıflandırma algoritmalarının performansı 16 farklı model oluşturularak incelenmiştir. Oluşturulan tüm modellerde model geçerliliğini doğrulamak için 10-katlı çapraz doğrulama yöntemi kullanılmıştır. Oluşturulan modellerde SMO algoritması diğer sınıflandırma algoritmalarından daha yüksek sınıflandırma doğruluğuna sahip olmuştur. Modeller karşılaştırıldığında dengesiz ve dengeli veri kümelerindeki örnek sayısının farklılığından en az etkilenen SMO algoritması olmuştur.

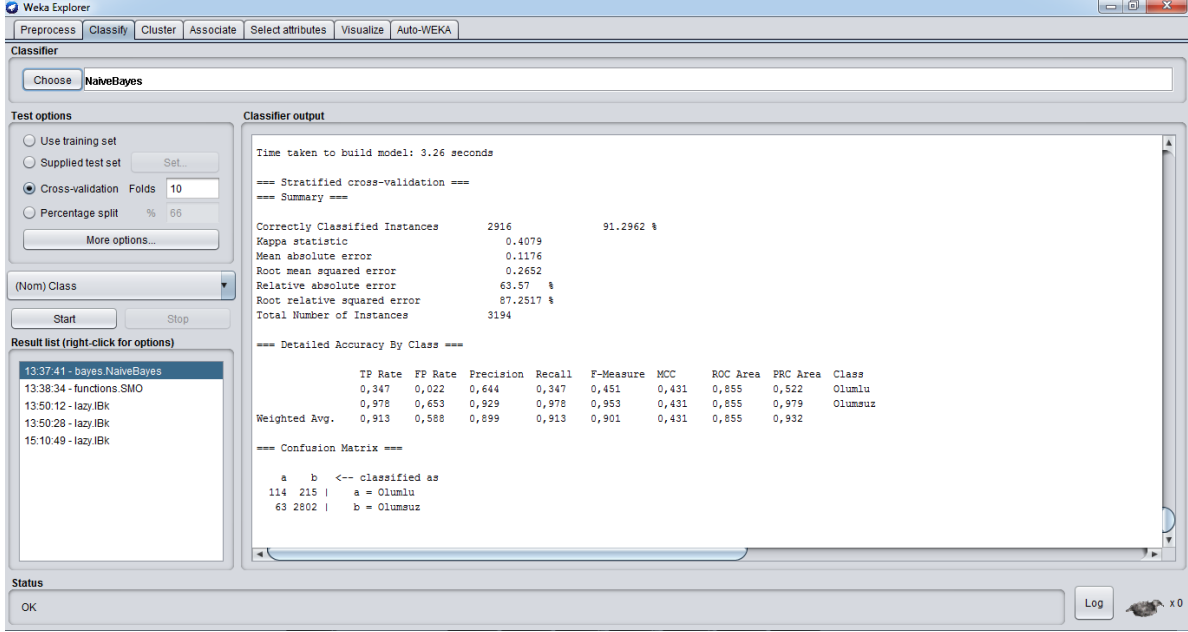
Sonuç olarak, e-ticaret firmalarına yönelik elde edilen sosyal medya yorumları analiz edildiğinde; dengesiz veri kümesinde öznitelik seçimi yapıldığında, kNN algoritması %93,52 sınıflandırma doğruluğu ile en yüksek başarıyı gösterse de tüm kriterler göz önüne alındığında en iyi ortalama performansı SMO algoritmasının gösterdiği söylenebilir.

Gelecek çalışmalarda yeterli veri sayısı ile dengeli veri kümesi oluşturularak sınıflandırma doğruluğu yükseltilebilir. Kelime kökleri bulunarak veri ön işleme teknikleri iyileştirilip, temsil gücü daha yüksek öznitelikler ile daha kaliteli çalışmalar yapılabilir.

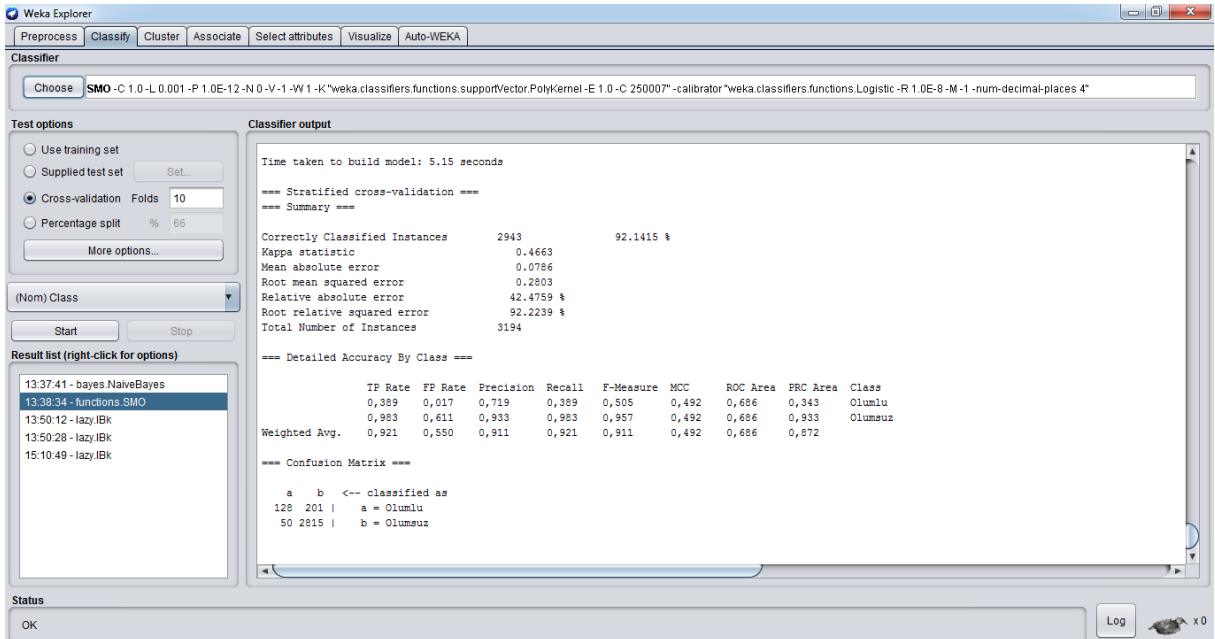
Olumsuz etiketli sosyal medya mesajları iade-tedarik-kargo-müşteri hizmetleri vb. şekilde kategorileştirilerek şikayetlerin hangi kategoride olduğu incelenebilir. Şikayetler kategorileştirilerek ilgili işletme departmanlarına ilgili müşteri şikayetlerinin otomatik dağıtılması hedeflenebilir. Firmalara göre duygu durum yüzdeleri belirlenebilir. Hangi firmada olumlu-olumsuz-nötr etiketli sosyal medya mesajının daha fazla atıldığı incelenebilir. Firmalar sektörlerine göre değerlendirilebilir. En çok içerik paylaşan sektörün/firmanın hangisi olduğu incelenebilir. Firmaların sosyal medya paylaşım gün ve zamanları dikkate alınarak değerlendirme yapılabilir. Etkileşimlerin özgün mesajlardan mı, retweetlerden mi oluştuğu tespit edilebilir. Sektörler/firmalar arasındaki farklılıklar Kikare uygunluk testi ile incelenebilir.

EKLER

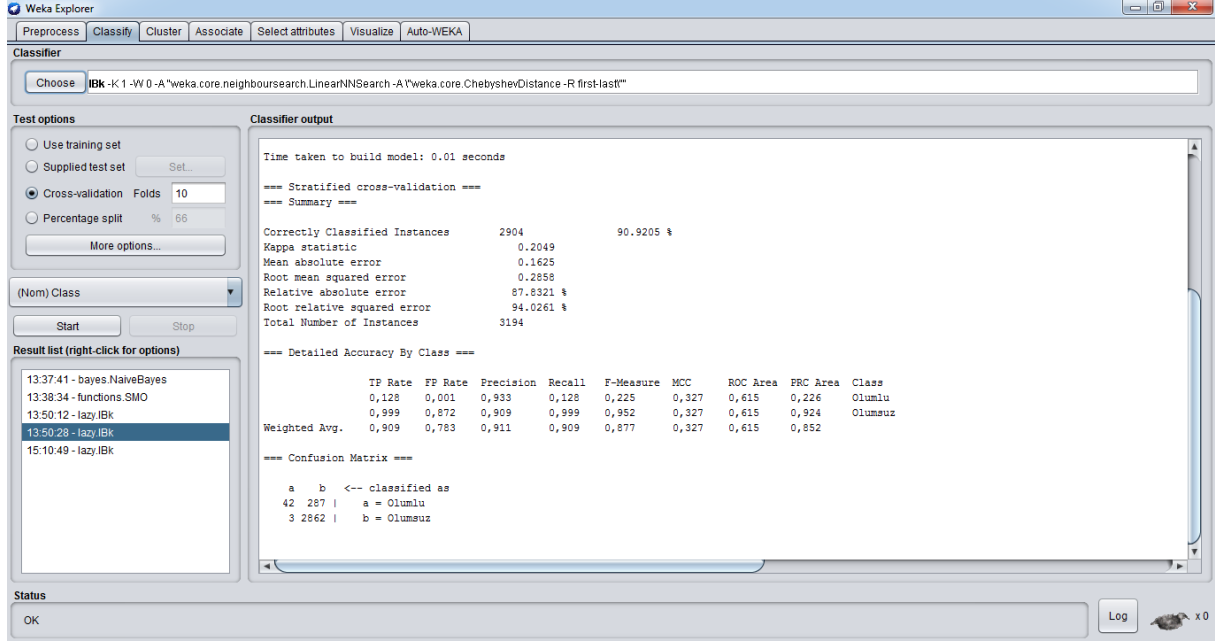
EK 1: Dengesiz Veri Kümesi Üzerinde Öznitelik Seçimi Yapılmadan Naive Bayes Sınıflandırma Sonucunun WEKA Ekran Görüntüsü



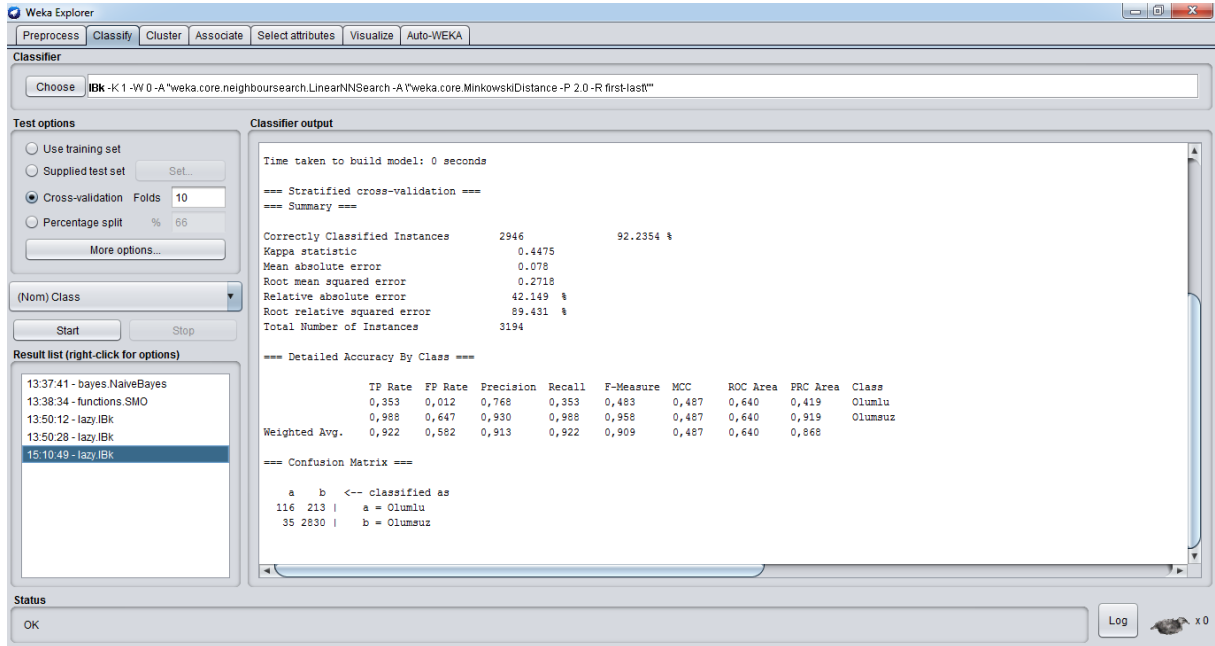
EK 2: Dengesiz Veri Kümesi Üzerinde Öznitelik Seçimi Yapılmadan SMO Sınıflandırma Sonucunun WEKA Ekran Görüntüsü



EK 3: Dengesiz Veri Kümesi Üzerinde Öznitelik Seçimi Yapılmadan kNN(k=1) Chebyshev Uzaklık Ölçütü ile Sınıflandırma Sonucunun WEKA Ekran Görüntüsü



EK 4: Dengesiz Veri Kümesi Üzerinde Öznitelik Seçimi Yapılmadan kNN(k=1) Öklid Uzaklık Ölçütü ile Sınıflandırma Sonucunun WEKA Ekran Görüntüsü



EK 5: Dengesiz Veri Kümesi Üzerinde Öznitelik Seçimi Yapılarak Naive Bayes Sınıflandırma Sonucunun WEKA Ekran Görüntüsü

Classifier
Choose **NaiveBayes**

Test options
 Use training set
 Supplied test set (Set...)
 Cross-validation Folds **10**
 Percentage split % **66**
 More options...
 (Nom) Class
 Start Stop

Result list (right-click for options)
 12:51:37 - bayes.NaiveBayes
 12:51:48 - functions.SMO
 12:52:04 - lazy.IBK
 12:52:17 - lazy.IBK

Classifier output
 Time taken to build model: 0.11 seconds
 === Stratified cross-validation ===
 === Summary ===
 Correctly Classified Instances 2915 91.2649 %
 Kappa statistic 0.3684
 Mean absolute error 0.1362
 Root mean squared error 0.2716
 Relative absolute error 73.6335 %
 Root relative squared error 89.3453 %
 Total Number of Instances 3194
 === Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
Weighted Avg.	0,292	0,016	0,676	0,292	0,408	0,407	0,829	0,515	Olumlu
	0,984	0,708	0,924	0,984	0,953	0,407	0,831	0,973	Olumsuz

 === Confusion Matrix ===

a	b	-- classified as	
96	233	a = Olumlu	
46	2819	b = Olumsuz	

Status
OK Log x0

EK 6: Dengesiz Veri Kümesi Üzerinde Öznitelik Seçimi Yapılarak SMO Sınıflandırma Sonucunun WEKA Ekran Görüntüsü

Classifier
Choose **SMO - C 1.0 - L 0.001 - P 1.0E-12 - N 0 - V -1 - W 1 - K "weka.classifiers.functions.supportVector.PolyKernel -E 1.0 - C 250007" - calibrator "weka.classifiers.functions.Logistic -R 1.0E-8 - M -1 - num-decimal-places 4"**

Test options
 Use training set
 Supplied test set (Set...)
 Cross-validation Folds **10**
 Percentage split % **66**
 More options...
 (Nom) Class
 Start Stop

Result list (right-click for options)
 12:51:37 - bayes.NaiveBayes
 12:51:48 - functions.SMO
 12:52:04 - lazy.IBK
 12:52:17 - lazy.IBK

Classifier output
 Time taken to build model: 0.82 seconds
 === Stratified cross-validation ===
 === Summary ===
 Correctly Classified Instances 2971 93.0182 %
 Kappa statistic 0.5353
 Mean absolute error 0.0698
 Root mean squared error 0.2642
 Relative absolute error 37.7376 %
 Root relative squared error 86.9279 %
 Total Number of Instances 3194
 === Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
Weighted Avg.	0,450	0,015	0,779	0,450	0,570	0,559	0,718	0,407	Olumlu
	0,985	0,550	0,940	0,985	0,962	0,559	0,718	0,939	Olumsuz

 === Confusion Matrix ===

a	b	-- classified as	
148	181	a = Olumlu	
42	2823	b = Olumsuz	

Status
OK Log x0

EK 7: Dengesiz Veri Kümesi Üzerinde Öznitelik Seçimi Yapılarak kNN(k=1) Chebyshev Uzaklık Ölçütü ile Sınıflandırma Sonucunun WEKA Ekran Görüntüsü

Classifier
Choose: `IBK -K 1 -W 0 -A "weka.core.neighboursearch.LinearNNSearch -A "weka.core.ChebyshevDistance -R first-last"`

Test options
 Use training set
 Supplied test set
 Cross-validation Folds: 10
 Percentage split % 66
 More options...
 (Nom) Class
 Start Stop

Result list (right-click for options)
 12:51:37 - bayes.NaiveBayes
 12:51:48 - functions.SMO
 12:52:04 - lazy.IBk
 12:52:17 - lazy.IBk

Classifier output
 Time taken to build model: 0.01 seconds
 === Stratified cross-validation ===
 === Summary ===
 Correctly Classified Instances 2987 93.5191 %
 Kappa statistic 0.5569
 Mean absolute error 0.0789
 Root mean squared error 0.2315
 Relative absolute error 42.6234 %
 Root relative squared error 76.1463 %
 Total Number of Instances 3194
 === Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
Weighted Avg.	0,935	0,494	0,931	0,935	0,926	0,590	0,928	0,988	Olumsuz

 === Confusion Matrix ===
 a b <-- classified as
 148 181 | a = Olumlu
 26 2839 | b = Olumsuz

Status: OK Log x0

EK 8: Dengesiz Veri Kümesi Üzerinde Öznitelik Seçimi Yapılarak kNN(k=1) Öklid Uzaklık Ölçütü ile Sınıflandırma Sonucunun WEKA Ekran Görüntüsü

Classifier
Choose: `IBK -K 1 -W 0 -A "weka.core.neighboursearch.LinearNNSearch -A "weka.core.MinkowskiDistance -P 2.0 -R first-last"`

Test options
 Use training set
 Supplied test set
 Cross-validation Folds: 10
 Percentage split % 66
 More options...
 (Nom) Class
 Start Stop

Result list (right-click for options)
 12:51:37 - bayes.NaiveBayes
 12:51:48 - functions.SMO
 12:52:04 - lazy.IBk
 12:52:17 - lazy.IBk

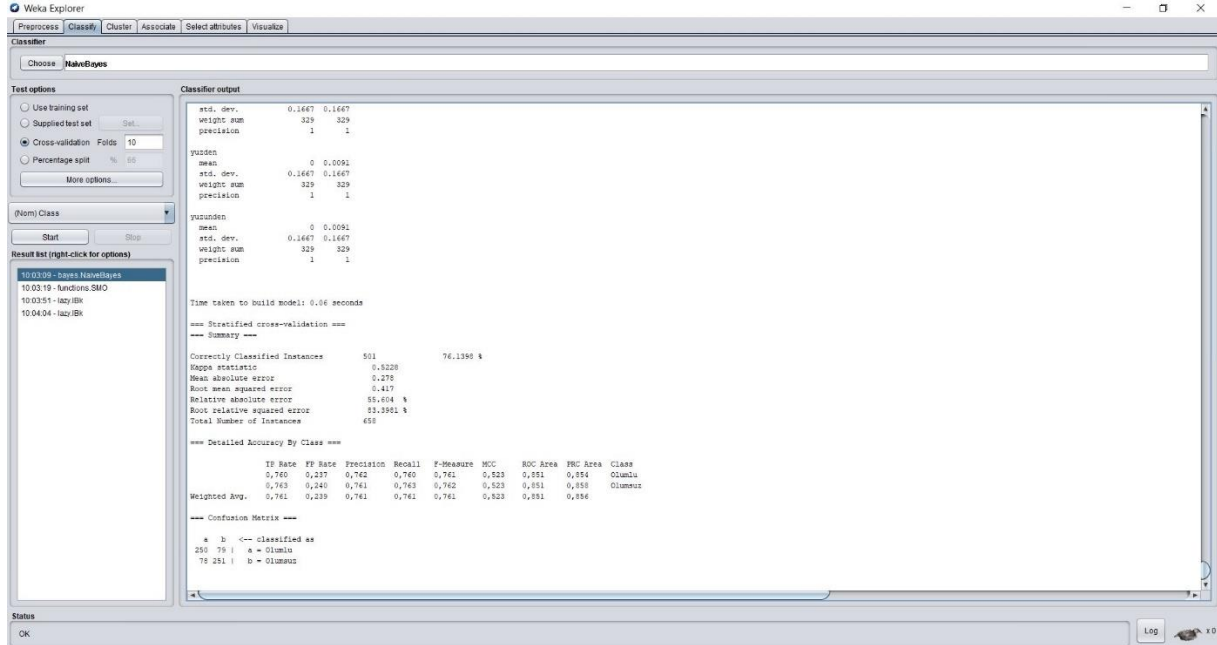
Classifier output
 Time taken to build model: 0 seconds
 === Stratified cross-validation ===
 === Summary ===
 Correctly Classified Instances 2987 93.5191 %
 Kappa statistic 0.5599
 Mean absolute error 0.0789
 Root mean squared error 0.2436
 Relative absolute error 42.6376 %
 Root relative squared error 80.1472 %
 Total Number of Instances 3194
 === Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
Weighted Avg.	0,935	0,489	0,931	0,935	0,926	0,591	0,807	0,969	Olumsuz

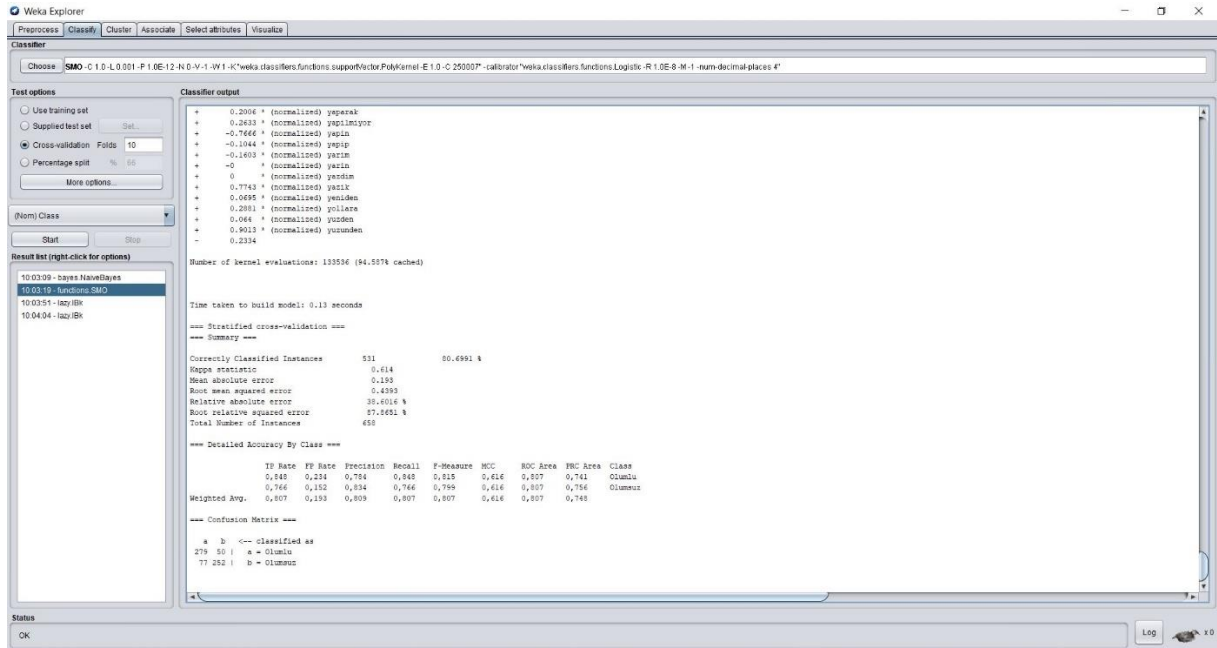
 === Confusion Matrix ===
 a b <-- classified as
 150 179 | a = Olumlu
 28 2837 | b = Olumsuz

Status: OK Log x0

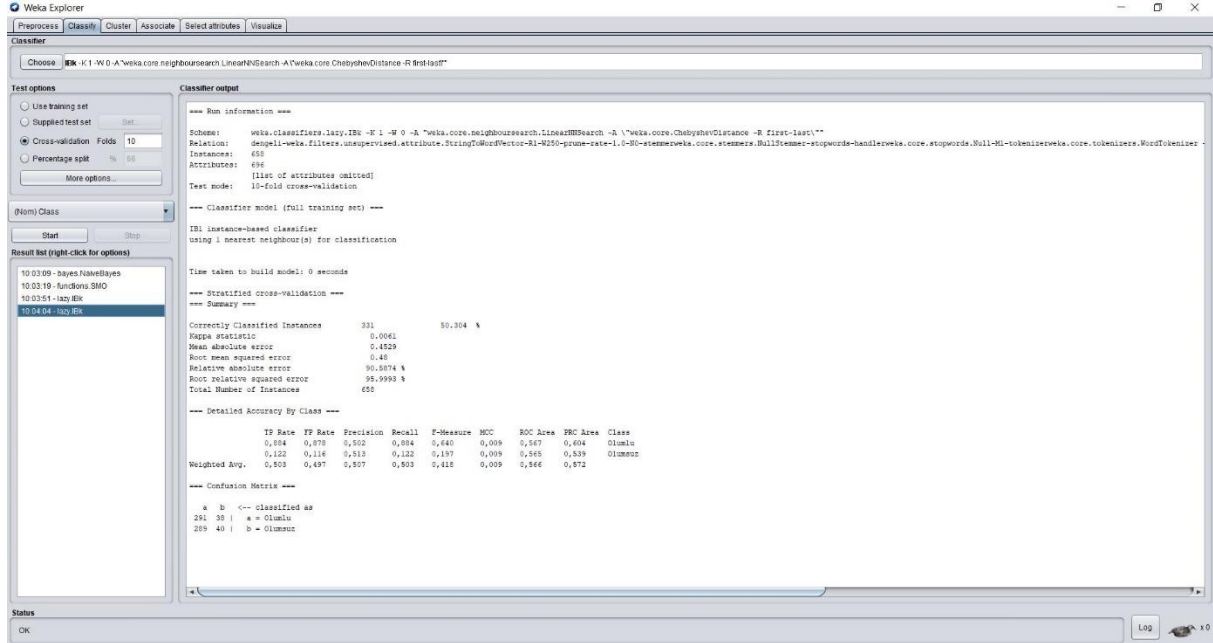
EK 9: Dengeli Veri Kümesi Üzerinde Öznitelik Seçimi Yapılmadan Naive Bayes Sınıflandırma Sonucunun WEKA Ekran Görüntüsü



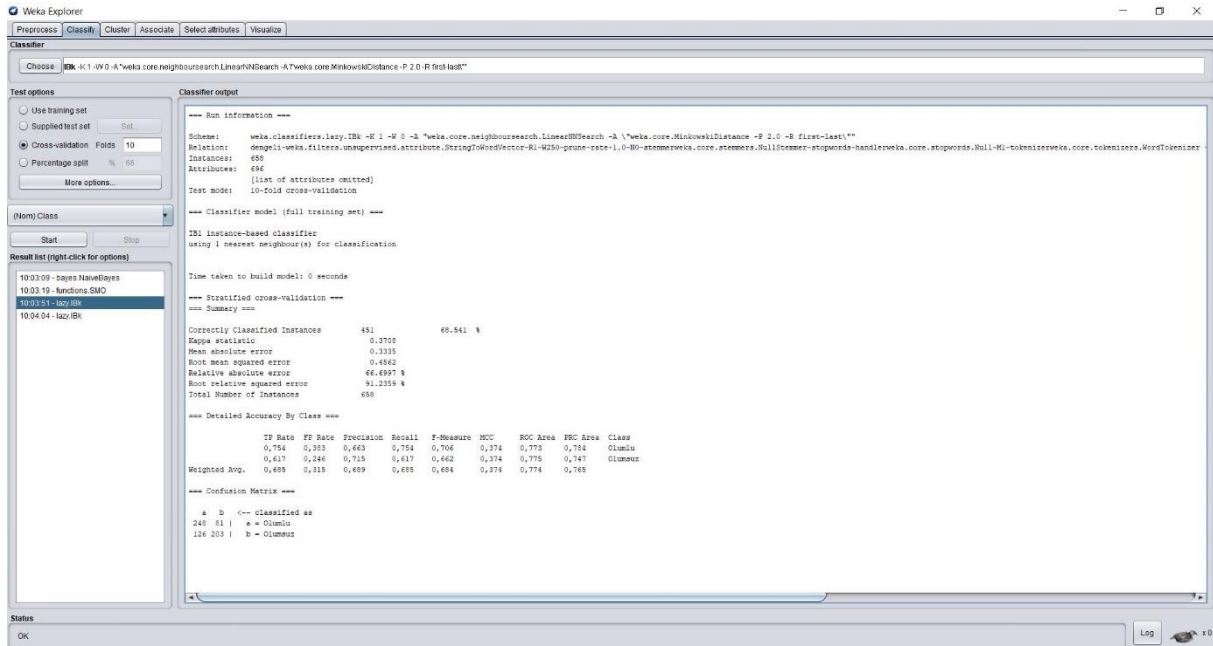
EK 10: Dengeli Veri Kümesi Üzerinde Öznitelik Seçimi Yapılmadan SMO Sınıflandırma Sonucunun WEKA Ekran Görüntüsü



EK 11: Dengeli Veri Kümesi Üzerinde Öznitelik Seçimi Yapılmadan kNN(k=1) Chebyshev Uzaklık Ölçütü ile Sınıflandırma Sonucunun WEKA Ekran Görüntüsü



EK 12: Dengeli Veri Kümesi Üzerinde Öznitelik Seçimi Yapılmadan kNN(k=1) Öklid Uzaklık Ölçütü ile Sınıflandırma Sonucunun WEKA Ekran Görüntüsü



EK 13: Dengeli Veri Kümesi Üzerinde Öznitelik Seçimi Yapılarak Naive Bayes Sınıflandırma Sonucunun WEKA Ekran Görüntüsü

Classifier

Choose **NaiveBayes**

Test options

Use training set
 Supplied test set (Sel...)
 Cross-validation Folds: 10
 Percentage split % 66
 More options...

(Nom) Class

Start Stop

Result list (right-click for options)

13:54:27 - bayes.NaiveBayes

Classifier output

Time taken to build model: 0.01 seconds

=== Stratified cross-validation ===
 === Summary ===

Correctly Classified Instances	486	73.8602 %
Kappa statistic	0.4772	
Mean absolute error	0.3341	
Root mean squared error	0.4188	
Relative absolute error	66.8216 %	
Root relative squared error	83.7529 %	
Total Number of Instances	658	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	FRC Area	Class
Weighted Avg.	0,775	0,298	0,722	0,775	0,748	0,478	0,843	0,846	Olumlu
	0,702	0,225	0,757	0,702	0,729	0,478	0,843	0,834	Olumsuz

=== Confusion Matrix ===

a	b	-- classified as	
255	74	a = Olumlu	
98	231	b = Olumsuz	

Status: OK Log

EK 14: Dengeli Veri Kümesi Üzerinde Öznitelik Seçimi Yapılarak SMO Sınıflandırma Sonucunun WEKA Ekran Görüntüsü

Classifier

Choose **SMO-C 1.0-L 0.001-P 1.0E-12-N 0-V-1-W 1-K weka.classifiers.functions.supportVector.PolyKernel-E 1.0-C 250007-calibrator weka.classifiers.functions.Logistic-R 1.0E-8-M 1-num-decimal-places 4**

Test options

Use training set
 Supplied test set (Sel...)
 Cross-validation Folds: 10
 Percentage split % 66
 More options...

(Nom) Class

Start Stop

Result list (right-click for options)

13:54:27 - bayes.NaiveBayes
 13:55:25 - functions.SMO

Classifier output

Time taken to build model: 0.02 seconds

=== Stratified cross-validation ===
 === Summary ===

Correctly Classified Instances	495	75.228 %
Kappa statistic	0.5046	
Mean absolute error	0.2477	
Root mean squared error	0.4977	
Relative absolute error	49.5438 %	
Root relative squared error	99.5425 %	
Total Number of Instances	658	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	FRC Area	Class
Weighted Avg.	0,945	0,441	0,682	0,945	0,792	0,547	0,752	0,672	Olumlu
	0,559	0,055	0,911	0,559	0,693	0,547	0,752	0,730	Olumsuz

=== Confusion Matrix ===

a	b	-- classified as	
311	18	a = Olumlu	
145	184	b = Olumsuz	

Status: OK Log

EK 15: Dengeli Veri Kümesi Üzerinde Öznitelik Seçimi Yapılarak kNN(k=1) Chebyshev Uzaklık Ölçütü ile Sınıflandırma Sonucunun WEKA Ekran Görüntüsü

The screenshot shows the WEKA Explorer interface with the Classifier tab selected. The classifier chosen is 'IBK -K 1 -W 0 -A "weka.core.neighboursearch.LinearNNSearch -A "weka.core.ChebyshevDistance -R first-last"'. The test options are set to 'Cross-validation' with 10 folds. The classifier output is as follows:

```
Time taken to build model: 0 seconds
=== Stratified cross-validation ===
=== Summary ===
Correctly Classified Instances      471          71.5805 %
Kappa statistic                    0.4316
Mean absolute error                 0.3002
Root mean squared error            0.3911
Relative absolute error            60.0407 %
Root relative squared error       78.2204 %
Total Number of Instances         658

=== Detailed Accuracy By Class ===
               TP Rate  FP Rate  Precision  Recall  F-Measure  MCC      ROC Area  PRC Area  Class
Weighted Avg.   0,474  0,043  0,918  0,474  0,625  0,493  0,839  0,828  Olumsuz
               0,716  0,284  0,782  0,716  0,698  0,493  0,839  0,836

=== Confusion Matrix ===
      a  b  <-- classified as
315 14 | a = Olumlu
173 156 | b = Olumsuz
```

EK 16: Dengeli Veri Kümesi Üzerinde Öznitelik Seçimi Yapılarak kNN(k=1) Öklid Uzaklık Ölçütü ile Sınıflandırma Sonucunun WEKA Ekran Görüntüsü

The screenshot shows the WEKA Explorer interface with the Classifier tab selected. The classifier chosen is 'IBK -K 1 -W 0 -A "weka.core.neighboursearch.LinearNNSearch -A "weka.core.MinkowskiDistance -P 2.0 -R first-last"'. The test options are set to 'Cross-validation' with 10 folds. The classifier output is as follows:

```
Time taken to build model: 0 seconds
=== Stratified cross-validation ===
=== Summary ===
Correctly Classified Instances      492          74.772 %
Kappa statistic                    0.4954
Mean absolute error                 0.2811
Root mean squared error            0.3799
Relative absolute error            56.2102 %
Root relative squared error       75.9788 %
Total Number of Instances         658

=== Detailed Accuracy By Class ===
               TP Rate  FP Rate  Precision  Recall  F-Measure  MCC      ROC Area  PRC Area  Class
Weighted Avg.   0,547  0,052  0,914  0,547  0,684  0,541  0,847  0,840  Olumsuz
               0,748  0,252  0,795  0,748  0,737  0,541  0,847  0,846

=== Confusion Matrix ===
      a  b  <-- classified as
312 17 | a = Olumlu
149 180 | b = Olumsuz
```

KAYNAKÇA

- Agarwal, B., N., Mittal, P., Bansal, S., Garg. (2015). *Sentiment Analysis Using Common-Sense and Context Information*. Hindawi Publishing Corporation Computational Intelligence and Neuroscience Volume 2015, Article ID 715730, 9 pages, <https://www.hindawi.com/journals/cin/2015/715730/> (15 Nisan 2019).
- Akdemir, E.B. (2019). *Veri Madenciliği Yaklaşımı ile Sosyal Ağ Analizi*. İstanbul: İstanbul Şehir Üniversitesi Fen Bilimleri Enstitüsü.
- Akın, A.A., M.D., Akın. (2007). Zemberek, an open source nlp framework for Turkic languages. Structure, 2007. Ulaşılabilir kaynak: https://s3.amazonaws.com/academia.edu.documents/34521696/zemberek_makale.pdf?AWSAccessKeyId=AKIAIWOWYYGZ2Y53UL3A&Expires=1556743707&Signature=oJr8zmV3tAKGuC2zFaz2nL2%2F8i0%3D&response-content-disposition=inline%3B%20filename%3DZemberek_makale.pdf (01.05.2019).
- Akpınar, H. (2000). *Veri Tabanlarında Bilgi Keşfi ve Veri Madenciliği*. İstanbul Üniversitesi İşletme Fakültesi Dergisi, C:29, S. 1 Nisan 2000, s. 1-22.
- Allahyari, M., S., Pouriye, M., Assefi, S., Safaei, E.D., Trippe, J.B., Gutierrez, K., Kochut. (2017). *A Brief Survey of Text Mining: Classification, Clustering and Extraction Techniques*. Bigdas at KDD 2017, Halifax, Kanada.
- Alpaydin, E. (2010). *Introduction to Machine Learning*. MIT Press. 2nd Ed. ISBN 978-0-262-01243-0.
- Argüden, Y., B., Erşahin. (2008). *Veri Madenciliği: Veriden Bilgiye, Masraftan Değere*. ARGE Danışmanlık Yayınları, İstanbul.
- Arslan, N. (2018). *Özelleştirilmiş Naive Bayes Algoritması*, <https://www.researchgate.net/publication/326635209> (20 Nisan 2019).
- Budak, H. (2018). *Özellik Seçimi Yöntemleri ve Yeni Bir Yaklaşım*. Journal of Natural and Applied Sciences Volume 22, Special Issue, s.21-31.
- Can, F., S., Kocberber, E., Balcık, C., Kaynak, H.C., Ocalan, O.M., Vursavas. (2008). *Information Retrieval on Turkish Texts*. Journal of the American Society for Information Science and Technology. Vol.59, No.3, s.407-421.
- Çetin, M., M. F., Amasyalı. (2013). *Active Learning for Turkish Sentiment Analysis*. IEEE International Symposium on Innovations in Intelligent Systems and Applications (INISTA), Albena-Bulgaria, s.1-4.
- Delen, D., M., Crosslan. (2008). *Seeding the survey and analysis of research literature with text mining*. Expert Systems with Applications, s.1707-1720.

- Doad, P.K., M.M., Bartere. (2013). *A Review : Study of Various Clustering Techniques*. International Journal of Engineering Research & Technology, 2(11). s.3141-3145.
- Farhadloo, M., E., Rolland. (2016). *Fundamentals of Sentiment Analysis and Its Applications*. ReserarchGate:Chapter-March 2016, DOI: 10.1007/978-3-319-30319-2_1. https://www.researchgate.net/publication/300965436_Fundamentals_of_Sentiment_Analysis_and_Its_Applications (10 Nisan 2019).
- Feldman, R., I., Dagan. (1995). *Knowledge Discovery in Textual Databases (KDT)*. In Proceeding of the 1st International Conference on Knowledge Discovery in Databases and Data Mining.
- Feldman, R., J., Sanger. (2007). *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data*. Cambridge University Press.
- Forman, G. (2003). *An Extensive Empirical Study of Feature Selection Metrics for Text Classification*. Journal of Machine Learning Research, 3, s.1289–1305.
- Go, A., R., Bhayani, L., Huang. (2009). *Twitter sentiment classification using distant supervision*. CS224N Project Report, Stanford(2009): s.1-12.
- Göker, H., H., Tekedere. (2017). *FATİH Projesine Yönelik Görüşlerin Metin Madenciliği Yöntemleri İle Otomatik Değerlendirilmesi*. Bilişim Teknolojileri Dergisi, Cilt: 10, Sayı: 3, s.291-299.
- Gürsakal, N. (2014). *Büyük Veri*. Dora Yayıncılık. 2.Baskı. ISBN 978-605-4798-80-3, Bursa.
- Gürsoy Şimsek, U.T., B., Karagöz Akın. (2018). *Adaptif Öğrenme Sözlüğü Temelli Duygu Analiz Algoritması Önerisi*. Bilişim Teknolojileri Dergisi, Cilt: 11, Sayı: 3, s.245-253.
- Güven, Z.B., T.T., Bilgin. (2014). *Zaman Serileri Madenciliği Kullanılarak Nüfus Artışı Tahmin Uygulaması*. Akademik Bilişim'14 - XVI. Akademik Bilişim Konferansı Bildirileri.
- Hall, M., E., Frank, G., Holmes, B., Pfahringer, P., Reutemann, I. H., Witten. (2009). *The weka data mining software: an update*. ACM SIGKDD explorations newsletter, 11(1), s. -10-18.
- Hearst, M. (2013). *What is text mining?*. <https://www.jaist.ac.jp/~bao/MOT-Ishikawa/FurtherReadingNo1.pdf> (03.05.2019).
- Hotha, A., A., Nurnberger, G., Paaß. (2005). *A Brief Survey of Text Mining*. LDV Forum – GLDV Journal for Computational Linguistics Language Technology 20(1) , s.19-62.
- İnternet: <http://www.cs.waikato.ac.nz/ml/weka/> (20 Nisan 2019).
- İnternet: *Rapid - I - Operator Overview*. <https://rapidminer.com/> (29.04.2019).

- Joachims T. (1998). Text Categorization with Support Vector Machines. Learning with Many Relevant Features [A]. In:Proceedings of the European Conference on Machine Learning [C].
- Kalender, M., E.E., Korkmaz. (2017). *Turkish entity discovery with word embeddings*. Turkish entity discovery with word embeddings (2017) 25: s.2388-2398.
- Karakoyun, M., M., Hacıbeyoğlu. (2005). Biyomedikal Veri Kümeleri ile Makine Öğrenmesi Sınıflandırma Algoritmalarının İstatistiksel olarak Karşılaştırılması, Dokuz Eylül Üniversitesi Mühendislik Fakültesi Dergisi, 16(48), s.30-41.
- Kaynar, O., M., Yıldız, Y., Görmez, A., Albayrak. (2016).*Makine Öğrenmesi Yöntemleri ile Duygu Analizi*. International Artificial Intelligence and Data Processing Symposium (IDAP'16), Sivas.
- Kılıç, S. (2015). *Kappa Test*. Journal of Mood Disorders Volume: 5, Number: 3, 2015 - www.jmood.org, s.142-144.
- Kılınç, D., E., Borandağ , F., Yücalar, V., Tunalı, M., Şimşek, A., Özçift. (2016). *KNN Algoritması ve R Dili ile Metin Madenciliği Kullanılarak Bilimsel Makale Tasnifi*. Marmara Fen Bilimleri Dergisi, 28(3),s.89-94.
- Ladha, L., T. Deepa. (2011). *Feature Selection Methods And Algorithms*. International Journal on Computer Science and Engineering, 3(5), 1787-1797.
- Landis J.R., G.G., Koch. (1977). *The measurement of observer agreement for categorical data*. Biometrics.1977;33, s.159-74.
- Lee, Y-J. (2017). *Sequential Minimal Optimization (SMO)*. http://jupiter.math.nctu.edu.tw/~yuhjye/assets/file/teaching/2017_machine_learning/SMO_algorithm.pdf (20 Nisan 2019).
- Manning C. D., H., Schütze. (1999). *Foundations of Statistical Natural Language Processing*. Cambridge: MIT Press.
- Medhat, W., A., Hassan, H., Korashy. (2014). *Sentiment analysis algorithms and applications. A survey*. Ain Shams Engineering Journal (2014) 5, s.1093–1113.
- Meral, M., B., Diri. (2014). *Twitter Üzerinde Duygu Analizi Sentiment Analysis on Twitter*. IEEE 22nd Signal Processing and Communications Applications Conference (SIU 2014).
- Miller, T. W. (2005). *Data and text mining: a business applications approach*. New Jersey: Pearson/Prentice Hall.
- Miner, G., D., Delen, J., Elder, A., Fast, T., Hill, R., Nisbet. (2012). *Practical Text Mining and Statistical analysis for Non-Structured Text Data Applications*. Waltham, USA: Elsevier Science & Technology, ProQuest Ebook Central,

<http://ebookcentral.proquest.com/lib/bogazici-ebooks/detail.action?docID=842198>
(20 Nisan 2019).

Naive Bayesian. (2019). http://www.saedsayad.com/naive_bayesian.htm (19 Nisan 2019).

Nalçakan, Y., Ş.S., Bayramoğlu, S., Tuna. (2015). *Sosyal Medya Verileri Üzerinde Yapay Öğrenme ile Duygu Analizi Çalışması.* Erişim: <https://www.researchgate.net/publication/280938376> (25.04.2019).

Nizam, H., S.S., Akın. (2014). *Sosyal Medyada Makine Öğrenmesi ile Duygu Analizinde Dengeli ve Dengesiz Veri Setlerinin Performanslarının Karşılaştırılması.* XIX. Türkiye'de İnternet Konferansı, İzmir.

Özdemir, A.F., E., Yıldıztepe, M. Binar. (2010). *İstatistiksel Yazılım Geliştirme Ortamı: R.* Akademik Bilişim'10 - XII. Akademik Bilişim Konferansı Bildirileri, 10 - 12 Şubat 2010 Muğla Üniversitesi. s.293-297.

Pang, B., L., Lee, S., Vaithyanathan. (2002). *Thumbs up?: sentiment classification using machine learning techniques.* Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10. Association for Computational Linguistics.

Patil, B. M., D., Toshniwal, R.C., Joshi. (2009). *Predicting Burn Patient Survivability Using Decision Tree In WEKA Environment.* 2009 IEEE International Advance Computing Conference (IACC 2009), Patiala, India, s.1353-1356.

PhD. Umut Orhan Çukurova Üniversitesi Ders Notları: <http://bmb.cu.edu.tr/uorhan/DersNotu/Ders02.pdf> (08.05.2019).

Platt, J.C., (1998), *Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines,* Technical Report MSR-TR-98-14.

Rich,E., K., Knight. (1990). *Artificial Intelligence Second Edition.* Columbus. McGraw-Hill Higher Education, Columbus USA.

Romero, C., S. Ventura. (2007). *Educational data mining: a survey from 1995 to 2005.* Expert Systems with Applications, 33(1), s.135–146.

Sahami, M. (1996). *Learning Limited Dependence Bayesian Classifiers.* KDD-96 Proceedings. Copyright © 1996, AAAI (www.aaai.org), Stanford, CA.

Sevindi, B.İ. (2013). *Türkçe Metinlerde Denetimli ve Sözlük Tabanlı Duygu Analizi Yaklaşımlarının Karşılaştırılması,* Ankara: Gazi Üniversitesi Fen Bilimleri Enstitüsü.

Sezer, E. (2018). *Sınıflandırma Sorunu İçin En Uygun Değişken Alt Kümesi Seçimi Üzerine Bir Uygulama.* Marmara Üniversitesi Sosyal Bilimler Enstitüsü.

Sim, J.,C.C., Wright. (2005). *The Kappa Statistic in Reliability Studies. Use, Interpretation, and Sample Size Requirements.* Physical Therapy, 85, 257-268.

- Şeker, S.E. (2008). *Çok sınıflı DVM (Multiclass SVM)*. <http://bilgisayarkavramlari.sadievrenseker.com/2008/12/01/cok-sinifli-dvm-multiclass-svm/> (23 Nisan 2019).
- Şeker, S.E. (2013). *K Fold Cross Validation (K Katlamalı Çarpaz Doğrulama)*. <http://bilgisayarkavramlari.sadievrenseker.com/2013/03/31/k-fold-cross-validation-k-katlamali-carpraz-dogrulama/> (23 Nisan 2019).
- Şeker, S.E. (2018). *CRISP-DM: Endüstriler Arası Standart İşleme – Veri Madenciliği için (Cross Industry Standard Processing – Data Mining)*. YBS Ansiklopedi www.YBSAnsiklopedi.com, Cilt 5, Sayı 2, Temmuz 2018, s.10-16.
- Tan, A.H., P.S., Yu,. (2004). *Guest Editorial: Text and Web Mining*. Applied Intelligence 18, Kluwer Academic Publisher, s.239-241.
- Tantuğ, A. C. (2012). *Metin Sınıflandırma Text Classification*. Türkiye Bilişim Vakfı Bilgisayar Bilimleri ve Mühendisliği Dergisi, 5(2).
- Taşçı,E., A., Onan. (2016). K-en yakın komşu algoritması parametrelerinin sınıflandırma performansı üzerine etkisinin incelenmesi. Akademik Bilişim 2016, Türkiye.
- Tekerek, A. (2011). *Veri Madenciliği Süreçleri ve Açık Kaynak Kodlu Veri Madenciliği Araçları*. Akademik Bilişim’11 - XIII. Akademik Bilişim Konferansı Bildirileri 2 - 4 Şubat 2011, İnönü Üniversitesi, Malatya, s.161-169.
- Thomas C.W., L., Paclik, P., Paclik, P.W., Robert. (2006). *Precision-recall operating characteristic (P-ROC) curves in imprecise environments*.School of Info. Tech. and Elec. Eng., The University of Queensland, 4072, Australia.
- Tripathy, A., A., Agrawal, S.K., Rath. (2016). *Classification of Sentiment Reviews using N-gram Machine Learning Approach*. Expert Systems with Applications, DOI:10.1016/j.eswa.2016.03.028.
- Tunç, A., İ., Ülger, (2016). *Veri Madenciliği Uygulamalarında Özellik Seçimi İçin Finansal Değerlere Binning ve Five Number Summary Metotları ile Normalizasyon İşleminin Uygulanması*. Kuveyttürk Katılım Bankası Konya AR-GE Merkezi, Konya.
- Uzer, M.S. (2014). *Örüntü Tanıma Uygulamalarında Yapay Zeka ve Öznitelik Dönüşüm Metotları Kullanılarak Geliştirilen Öznitelik Seçme Algoritmaları*, Konya: Selçuk Üniversitesi Fen Bilimleri Enstitüsü.
- Xia, R., C., Zong, S., Li. (2011). *Ensemble of feature sets and classification algorithms for sentiment classification*. Information Sciences, 181(6), s.1138-1152.
- Yang, Y., X., Liu. (1999). *A Re-examination of Text Categorization Methods*. In: Proceedings of 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval [C]. s.42-49.

- Yıldız, M., S.E., Şeker. (2016). *Veri Madenciliği Araçları (Data Mining Tools)*. YBS Ansiklopedisi, www.YBSAnsiklopedi .com, Cilt:3, Sayı:4, s.10-19.
- Yuzhong, C., L., Baoli, Y., Shiwen. (2003). *An Improved k-Nearest Neighbor Algorithm for Text Categorization*. To appear in the Proceedings of the 20th International Conference on Computer Processing of Oriental Languages, Shenyang, China.
- Yuzhong, C., L., Baoli, Y., Shiwen. (2002). *A Comparative Study on Automatic Categorization Methods for Chinese Search Engine*. In: Proceedings of the Eighth Joint International Computer Conference [C]. Hangzhou: Zhejiang University Press, 117-120.
- Yüksel, A.S., F.G., Tan. (2018). *Metin Madenciliği Teknikleri ile Sosyal Ağlarda Bilgi Keşfi*. Journal of Engineering Sciences and Design, 6(2), 324-333.