T.C.
İSTANBUL ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

DOKTORA TEZİ

MİKROBLOG HİZMETLERİNDEKİ ÖRTÜK BİLGİNİN VERİ MADENCİLİĞİ TEKNİKLERİ İLE KEŞFİ

Feridun Cemal ÖZÇAKIR

Enformatik Anabilim Dalı

Enformatik Programı

DANIŞMAN
Prof. Dr. Sevinç GÜLSEÇEN

Ekim, 2016

İSTANBUL

Bu çalışma 17/10/2016 tarihinde aşağıdaki jüri tarafından Enformatik Anabilim Dalı Enformatik Programında Doktora Tezi olarak kabul edilmiştir.

Tez Jürisi:

İmza
Prof. Dr. Sevinç GÜLSEÇEN (Danışman)
İstanbul Üniversitesi
Enformatik Bölümü

İmza
Prof. Dr. Ş. Alp BARAY
İstanbul Üniversitesi
Mühendislik Fakültesi

İmza
Prof. Dr. İ. Müfit GİRESUNLU
İstanbul Üniversitesi
Fen Fakültesi

İmza
Prof. Dr. Oya KALIPSIZ
Yıldız Teknik Üniversitesi
Elektrik Elektronik Fakültesi

İmza
Doç. Dr. Ahmet Tuğrul SAVAŞ
Okan Üniversitesi
Uygulamalı Bilimler Yüksekokulu

20.04.2016 tarihli resmi gazetede yayımlanan Lisansüstü Eğitim ve Öğretim Yönetmeliğinin 9/2 ve 22/2 maddeleri gereğince; Bu Lisansüstü teze, İstanbul Üniversitesi'nin abonesi olduğu intihal yazılım programı kullanılarak Fen Bilimleri Enstitüsü'nün belirlemiş olduğu ölçütlere uygun rapor alınmıştır.

ÖNSÖZ

Doktora tez çalışmam boyunca gösterdiği her türlü destek ve yardımdan dolayı çok değerli hocam Prof. Dr. Sevinç GÜLSEÇEN'e, tez izleme komitemde yer alan ve çalışmaya değerli fikirleri ile katkıda bulunan hocalarım Prof. Dr. Alp BARAY'a, Doç. Dr. Ahmet Tuğrul SAVAŞ'a ve Prof. Dr. Gonca TELLİ YAMAMOTO'ya en içten dileklerimle teşekkür ederim.

Tezin hazırlanmasında bütün sıkıntılarıma katlanan ve çalışma boyunca her türlü desteğini veren sevgili eşim Şahine ÖZÇAKIR'a ve çocuklarım Dilara ve Kerem'e ayrıca teşekkür ederim.

Ekim, 2016 Feridun Cemal ÖZÇAKIR

İÇİNDEKİLER

Sayfa No

ŞEKİL LİSTESİ

Sayfa No

TABLO LİSTESİ

Sayfa No

SİMGE VE KISALTMA LİSTESİ

Simgeler	Açıklama
$\boldsymbol{A}$	: Doğruluk Ölçüsü
$\boldsymbol{c}$	: Sınıf Etiketi
$\boldsymbol{C}$	: Sınıf Uzayı
$\boldsymbol{d}$	: Doküman – Mesaj
$\boldsymbol{D}$	: Mesaj uzayı
F	: F-Ölçüsü
$\boldsymbol{H}$	: Hipotez veya Entropi
$\boldsymbol{K}, \boldsymbol{k}$	: Sabit, Katsayı
$\boldsymbol{N}, \boldsymbol{n}$	: Öğe Sayısı
$\boldsymbol{P}, \boldsymbol{p}$	: Olasılık
$\boldsymbol{t}$	: Terim (Kelime veya Kelime Grubu)
$\boldsymbol{x}$	: Gözlenen Değer
$\boldsymbol{w}$	: Ağırlık
$\boldsymbol{Z}$	: Normalleştirme Faktörü
$\boldsymbol{\lambda}$	: Ağırlık Parametresi

Kısaltmalar	Açıklama
ASP	: Aktif Sunucu Sayfaları (Active Server Pages)
API	: Uygulama Programlama Arayüzleri (Application Programming Interface)
$\boldsymbol{CVA}$	: Çapraz-Geçerlilik Doğruluğu (Cross-Validation Accuracy)
IDF	: Ters Doküman Frekansı (Inverse Document Frekans)
DDA	: Doğal Dil Ayrıştırma
DDI	: Doğal Dil İşleme (Natural Processing Language – NLP)
JSON	: Java Script Nesne Gösterimi (Java Script Object Notation)
LI	: Olabilirlik
LS	: Laplace Düzgünleştirme (Smoothing)
MAP	: En Yüksek Sonsal Olasılık
MaxEnt	: Maksimum Entropi
ML	: En Yüksek Olabilirlik (Maximum Likelihood)
NB	: Naïve Bayes
OAuth	: Twitter Kimlik Doğrulama Yapısı
neg	: Olumsuz – Negatif
poz	: Olumlu - Pozitif
SQL	: Yapısal Sorgulama Dili
TF	: Terim Frekansı
TF-IDF	: Terim Bileşik Ağırlık Değeri
VBTK	: Veritabanlarında Bilgi Keşfi
VM	: Veri Madenciliği
XML	: Genişletilebilir İşaretleme Dili (Extensible Markup Language)

ÖZET

DOKTORA TEZİ

MİKROBLOG HİZMETLERİNDEKİ ÖRTÜK BİLGİNİN VERİ MADENCİLİĞİ TEKNİKLERİ İLE KEŞFİ

Feridun Cemal ÖZÇAKIR

İstanbul Üniversitesi

Fen Bilimleri Enstitüsü

Enformatik Anabilim Dalı

Danışman : Prof. Dr. Sevinç GÜLSEÇEN

Günümüzde internet, mobil teknolojiler ve sosyal medya günlük yaşamın ayrılmaz birer parçalarıdır. Bu ortamlarda üretilen veriler üzerinden kişi veya kurumlar için değerli olacak bilgiye veri madenciliği ve makine öğrenmesi teknikleri ile ulaşılabilir. Bu tez çalışmasında sosyal medya ortamı Twitter mikroblog hizmetini kullanan insanların ürettiği görüşlerin analizini yapan ve analiz sonuçlarına göre mesajlar içindeki örtük bilgiyi keşfedebilen bir veri analiz aracı geliştirilmiştir. Geliştirilen yazılım aracı ile insanların ürünler, hizmetler, kuruluşlar, bireyler, sorunlar ve toplumsal olaylar ile ilgili duygularının, fikirlerinin, görüşlerinin, tepkilerinin ve değerlendirmelerinin analizi gerçekleştirilmektedir. Analiz işlemi esnasında metin madenciliğine yönelik veri madenciliği ve doğal dil işleme teknikleri kullanılmaktadır.

Web tabanlı bir yazılım uygulaması olarak geliştirilen analiz aracı ile belirtilen anahtar kelime (kilit terim) veya kelime gruplarını içeren görüşler Twitter mikroblog uygulamasından alınmakta, Türkçeye özel olarak geliştirilen Anahtar Kelime ve Fiil Tabanlı model ile alınan görüşler üzerinden kişilerin/fikir gruplarının/toplumun olumlu/olumsuz yorum veya istek gibi duyguları belirlenmektedir. Analiz aracına ayrıca N-gram tekniği ile oluşturulmuş eğitim verisi ile çalışan Naïve Bayes ve Maksimum Entropi makine öğrenmesi teknikleri yazılım olarak entegre edilmiştir. Entegre edilen bu modeller ile analiz aracı Twitter mikroblog uygulamasından alınan görüşler üzerinde duygu analizi yapmaktadır.

Modellerin geçerliliği için performans ölçümleri hesaplanmış, modellerin başarı seviyesi değerlendirilmiştir. Anahtar Kelime ve Fiil Tabanlı model ile N-gram yapılı Naïve Bayes ve Maksimum Entropi modellerinin karşılaştırılması yapılmış, elde edilen bulgulara göre Naïve Bayes ve Maksimum Entropi modellerinin Anahtar Kelime ve Fiil Tabanlı modele göre daha başarılı sonuçlar ürettikleri görülmüştür.

Ekim 2016, 161 sayfa.

Anahtar kelimeler: Veri Madenciliği, Duygu Analizi, Mikroblog, Makine Öğrenmesi, N-gram, Naïve Bayes, Maksimum Entropi

SUMMARY

Ph.D. THESIS

DISCOVERY OF TACIT KNOWLEDGE IN THE MICROBLOGGING SERVICES BY DATA MINING TECHNIQUES

Feridun Cemal ÖZÇAKIR

İstanbul University

Institute of Graduate Studies in Science and Engineering

Department of Informatics

Supervisor : Prof. Dr. Sevinç GÜLSEÇEN

Today, internet, mobile technology and social media are integral parts of daily life. The valuable knowledge for individuals and organizations can be reached by data mining and machine learning techniques from generated data via social media. In this thesis, a data analysis tool has been developed, makes the analysis of opinions produced by the people who use the Twitter microblogging service, is a social media environment, and discovers the tacit knowledge in the messages according to analysis results. The developed software tool is realized analysis of people's sentiments (ideas, opinions, reactions, evaluations) about products, services, organizations, individuals, problems, and social events. Data mining and natural language processing techniques towards text mining are used during the analysis process.

The data analysis tool was developed as a web-based software application. The opinions (messages), contain specified keyword (key term) or phrase, are taken from the Twitter microblogging application with this tool. On opinions, are taken with Keyword and Verb Based model is developed for Turkish language, to determine sentiments of people / opinion groups / society. Naïve Bayes and Maximum Entropy models, are machine learning techniques and work with training dataset was established with N-gram structures, have been integrated a software in the data analysis tool. By these integrated

models, the data analysis tool makes sentiment analysis over opinions are taken from the Twitter microblogging application.

The performance metrics of models have been calculated for the validity of the models, the success levels of the models have been evaluated. Keyword and Verb Based model was compared with N-gram models (Naïve Bayes and Maximum Entropy models). According to the findings, Naïve Bayes and Maximum Entropy models have produced more successful results by Keyword and Verb Based model.

October 2016, 161 pages.

Keywords: Data Mining, Sentiment Analysis, Microblog, Machine Learning, N-gram, Naïve Bayes, Maximum Entropy

1. GİRİŞ

"*İyi bir fikre (düşünceye) sahip olmanın en iyi yolu, birçok fikre sahip olmaktır.*"

Albert Einstein

Türk Dil Kurumu'nun güncel Türkçe sözlüğünde düşünce (fikir); "uzay ve zamanın ötesinde, öznenin dışında, kendiliğinden var olan, duyularla değil, yalnızca ruhen algılanabilen asıl gerçekliktir. Dış dünyanın insan zihnine yansımasıdır." şeklinde tanımlanmaktadır (Türk Dil Kurumu, 2015). Duyularımız ile topladığımız veriler zihin içinde analiz edilir, gerekli ve gereksiz veriler ayıklanır, kalan verilerin aralarındaki ilişkiler değerlendirilir. Bu sürecin sonucunda insan zihni gelişimine bağlı olarak ilgili olay (durum), sorun veya nesne için bir veya birden fazla düşünce üretir. Bu düşünceler mevcut olayı veya nesneyi irdeleyerek yorum niteliğinde çıkarımlar olarak ortaya sunulabildiği gibi, bir sorunu çözmeye yönelik çözümler olarak da karşımıza çıkmaktadır. Önceden gerçekleşmiş bir olay veya sorunun irdelenmesi sonucunda gelecekte benzer özelliklere sahip veya geçmişteki olay veya sorunun devamı niteliğindeki olay veya sorunların tahminine yönelik varsayımlar olarak da karşılaşabilmekteyiz. Orhan Hançerlioğlu *Düşünce Tarihi* kitabında "*Dil ve düşünce, insanı insan eden insanca özelliklerin başında gelir*" ifadesi ile düşüncenin insanın en önemli varlığı olduğuna işaret etmektedir (Hançerlioğlu, 2000). Bir olay, sorun veya nesne karşısında kişilerin zihinlerinde ürettiği düşünce; yetiştiği sosyal çevreye, aldığı eğitime, sahip olduğu inanca ve zamanla geliştirdiği dünya görüşüne göre kişiden kişiye göre değişmektedir. Bu doğrultuda birçok etkene göre değişken olan ve insan olmanın en önemli özelliklerinden biri olan düşüncelerin her biri değerlidir. Bir ortak düşünceye (veya çözüm yoluna) ihtiyaç duyulduğunda tek bir düşünce ile süreci planlamak veya sorunu çözmek yerine, konu ile ilgili tüm bireylerin düşüncelerini almak, iyi bir fikre ulaşmanın en iyi yoludur. Belli bir düzen içinde düşüncelerin toplanması, toplanan düşüncelerden sonuca yönelik olmayanların elenmesi ve sonuca yönelik olanlar içinde en fazla aynı fikri savunanların tespit edilip, çoğunluğun görüşü olabilecek, ortak düşünce olarak kabul edilecek düşünceye karar verilmesi en değerli düşünceye ulaşmamızı sağlayacaktır.

Knowledge Forum (2012)'a göre, günümüzde birçok başarılı araştırma ekibi, ticari, bilimsel ve eğitim kurumunun ortak noktası bireysel fikirleri topluluk için geçerli fikirler (örgüt fikri) haline dönüştürebilmektir. Bu kurumların her biri çalışanları ile birlikte bir örgüttür. Bu örgütler, her bireyin fikrinin örgüte katkı sağladığı, yeni bir fikrin veya bilginin oluşturulmasında herkesin katkısının olmasıyla çıkan bilgi veya fikrin herkesin ortak eseri olacağını, ayrıca paylaşılan bilginin yenilik ve büyümeye yol açacağını savunmaktadırlar.

Pazarın ihtiyaçları doğrultusunda pazara sunulacak yeni bir ürün tasarlamak, bir kuruma ait pazarlama stratejisini olumlu yönde geliştirmek, herhangi bir konunun öğrenimi esnasında bu konunun eğitimini alanların daha etkin ve kalıcı öğrenmeleri sağlamak gibi birçok çözüm bekleyen sorunu çözmek için kurumlar ve sosyal topluluklar içinde bireylerden oluşan fikir grupları oluşturulmaktadır. Bu gruplar içinde yer alan bireylerden çözüme yönelik görüşler toplamak günümüzde sıklıkla tercih edilen paylaşımlı ortamlarda ortak görüşü ortaya çıkarma yöntemidir. Bu ortamların oluşturulmasındaki amaç, örgüt bireyleri veya ortam katılımcıları tarafından belirtilen fikirlerin içinden değerli olabilecek gerçek anlamdaki örtük ve açık bilgiyi ortaya çıkarmaktır.

Belli bir amacı hedefleyen, bir soruna ait çözümü bulmayı amaçlayan veya bir durum karşısında eleştiri (veya tepki) niteliğindeki düşünceler ile olması gereken süreci biçimlendirmeye çalışan fikir gruplarından (bu gruplar bir organizasyon aracılığı ile veya kendiliğinden internet ortamında oluşmuş olabilir) görüş toplamak ve toplanan bu görüşleri analiz etmek için bilgi teknolojilerinden yararlanılması, birçok fikir içinden en iyi fikre (veya toplumun çoğunluğunun benimsediği ortak fikre) ulaşmamızı sağlayacaktır. İlgili alanda çalışma yapan analistlerin de zaman ve iş gücü problemlerini etkili bir şekilde çözümleyecektir. Fikirlerin (düşüncenin) analiz edilmesi sonucunda ulaşılmak istenen gerçekte değerli olan bilgidir.

Son yıllarda yaşamın her alanında bilişim teknolojilerinin vazgeçilmez bir öğe olduğu görülmektedir. Günümüzde hızla gelişen elektronik teknoloji ile birlikte özellikle mobil cihazların hayatımızın bir parçası olmaya başladığı görülmektedir. 7-8 yıl öncesinde mobil cihazlar sadece sesli görüşme ve kısa mesajlaşmalar için kullanılırken, günümüzde artık görüntülü görüşme ve özellikle internet ağ yapısı içinde sürekli bulunma amacıyla kullanılmaktadır. İnternet kullanımının yaygınlaşması ile kişilerin sosyal medya ve sosyal

ağları kullanma oranları hızla artmaktadır, özellikle Facebook ve Twitter'ın günümüzde dünya üzerinde en çok kullanılan sosyal ağlar olduğu görülmektedir (Narayanan, 2010). Bu sosyal ağların popülerliğinin hızla artması ile bilginin de bu ortamlar üzerinden hızla yayıldığı bir durum ortaya çıkmıştır.

Sosyal ağlar üzerinde bir konu, olay veya sorun ile ilgili fikir grupları oluşturulabilir, Fikir gruplarının sosyal ağlar üzerinde kurulması katılımcıların kendilerini özgür bir ortamda hissetmeleri sağlayacak, zaman kısıdı olmayacağından katılımcılardan daha verimli görüşler toplanacaktır. Analiz açısından uzun metin halindeki görüşleri analiz etmek zor ve çok zaman almaktadır, daha kısa ve net mesajlar üzerinden görüşleri analiz etmek sürecin işleyişini kolaylaştıracaktır. Fikir gruplarının görüşlerini veya genel olarak gündemi yakalayan bir konu ile ilgili toplumsal görüşleri toplamak için mikroblog ortamının araç olarak kullanılması etkin bilgi toplama yöntemi olacaktır. Bu doğrultuda en popüler mikroblog uygulaması olan Twitter sosyal ağ sitesinin en uygun veri toplama ortamı olduğu görülmektedir.

Mobil teknolojileri ve interneti sıklıkla kullanan bireyler herhangi bir konu hakkında görüşlerini belirtmek için mikroblogları özellikle de günümüzde Twitter mikroblog hizmetini tercih etmektedir. Bu hizmette yer alan Türkçe mesajlar üzerinde bilgi teknolojileri ve makine öğrenmesi teknikleri kullanılarak, duygu analizi çalışması yapılması toplumsal, politik veya herhangi özel bir konu hakkında genel görüş hakkında tahminde bulunma imkânı oluşturacaktır.

Bu tez çalışmasının amacı, mikroblog hizmetleri üzerinde insanların üreteceği görüşlerin analizi ile örtük bilgiyi keşfedecek bir veri analiz aracı geliştirmektir. Analiz aracı ile metin madenciliğine yönelik veri madenciliği teknikleri ve doğal dil işleme kullanılarak insanların ürünler, hizmetler, kuruluşlar, bireyler, sorunlar, olaylar, konular ve bu nitelikteki varlıklarla ilgili fikirlerinin (düşüncelerinin, görüşlerinin), duygularının ve değerlendirmelerinin analizini gerçekleştirmek hedeflenmektedir.

Tez çalışması kapsamında analizi yapan kişi veya ekip tarafından belirtilen anahtar kelime (kilit terim) veya kelime gruplarını içeren görüşlerin mikroblog (Twitter) uygulamasından getirilmesi, getirilen görüşler üzerinde doğal dil işleme yöntemleri yardımı ile veri madenciliği sınıflama tekniklerinin uygulanması, bu teknikler aracılığı

ile kişilerin/fikir gruplarının/toplumun olumlu/olumsuz yorum, soru veya istek gibi duygularının sınıflandırılması amaçlanmaktadır.

Veri madenciliği kavramı ile bütünleşik olan "veritabanlarında bilgi keşfi" Usame Fayyad'a göre; "veri seçimi", "veri önişleme", "veri indirgeme", "veri madenciliği" ve "değerlendirme (yorumlama, doğrulama)" aşamalarından oluşur (Fayyad ve diğ. , 1996). Sürecin başındaki "veri" keşif süreci sonucunda daha önceden bilinmeyen, değerli olan "bilgi" şekline dönüşmektedir.

Tez çalışması kapsamında süreç, tez konusu ile ilgili alanyazın (literatür) taraması sonrası veritabanlarında bilgi keşfi yapısında yer alan aşamalar dikkate alınarak oluşturulmuştur. Tez çalışması kapsamında oluşturulan etkileşimli ortam ve analiz aracı ile örtük bilginin ortaya çıkarılması, anlamlı ve yararlı olacak bilginin üretiminin gerçekleştirilmesi amaçlanmaktadır.

Amaç doğrultusunda tezin kavramsal temellerini oluşturmak için GENEL KISIMLAR bölümünde, tez çalışmasının temel amacı olan problemin tanımlanması için "Bilgi Toplumu", "Veritabanlarında Bilgi Keşfi Süreçleri", "Veri Madenciliği Algoritmaları", "Doğal Dil İşleme", "Duygu Analizi", "Makine Öğrenmesi" konularına değinilmiştir. "Naïve Bayes", 'Destek Vektör Makineleri' ve "Maksimum Entropi" makine öğrenmesi metotları hakkında temel bilgilere yer verilmiştir.

MALZEME VE YÖNTEM bölümünde, tez çalışması kapsamında geliştirilen "Anahtar Kelime ve Fiil Kelime Tabanlı Duygu Analizi" modeli ve modelin uygulaması, veritabanlarında bilgi keşfi süreçleri takip edilerek açıklanmıştır. Ayrıca geliştirilen modelin "Naïve Bayes" ve "Maksimum Entropi" makine öğrenmesi metotları ile karşılaştırılmasına karar verilmiş ve bu modeller hakkında detaylı bilgilere yer verilmiştir. Naïve Bayes ve Maksimum Entropi makine öğrenmesi modellerinin geliştirilen modelle bütünleşik çalışması için uygulama içine geliştirme yapılarak (kodlama) aktarılması bu bölümde açıklanmıştır.

BULGULAR bölümünde geliştirilen model ve uygulama içine entegre edilen diğer modeller üzerinden elde edilen verilerin analizi, bu analizlere göre modellerin uygulanabilirliğine ilişkin sonuçlara yer verilmiştir. Ayrıca bu bölümde geliştirilen model ile makine öğrenmesi tekniklerinin birbirleriyle karşılaştırması yapılarak en iyi

tahminleyici model hakkında karar vermemizi sağlayacak veri analiz sonuçları gösterilmiştir.

Son olarak TARTIŞMA VE SONUÇ' bölümünde tez çalışmasının değerlendirilmesi yapılarak, mevcut çalışmanın üzerinden geleceğe yönelik olabilecek yeni geliştirmeler hakkında tahminlerde bulunulmuş ve bu çalışmanın gelecekteki akademik çalışmalara katkı sağlaması için önerilerde bulunulmuştur.

2. GENEL BİLGİLER

2.1. BİLGİ TOPLUMU, SOSYAL AĞLAR VE SOSYAL MEDYA

Geçmişte insanlar düşüncelerini diğer insanlara aktarabilmek için seslerini (sözel) veya yazılı belgeleri kullanırdı. Günümüzde ise bu metotlara ve araçlara ek olarak elektronik ve iletişim teknolojisindeki gelişmeler ile oluşan dijital ortamlarda düşüncenin aktarımına hizmet etmektedir. 1800'lü yıllarda kitap, dergi gibi kâğıda basılı belgeler aracılığı ile topluma sunulan düşünce ve bilgi, 1900'lü yılların ortasında radyo, sinema ve televizyon ile, 1980 ve 1990 yıllardan itibaren dijital teknoloji ürünleri olan disket, CD (compact disk), DVD (digital versatile-video disc) aracılığı ile, 1990 ortasından sonra da internet aracılığı ile topluma sunuldu. Elektrik sinyallerinin taşındığı kablo ve radyo sinyallerinden oluşan ağların birbirleri ile iletişim içinde olduğu ağlar arası büyük ağ olarak nitelendireceğimiz internetin yaygın olarak dünya üzerinde alt yapısının oluşturulup, toplumların hizmetine sunulması, düşüncenin ve bilginin yayılmasına büyük katkıda bulunmaktadır.

İnternet, 2004 yılından itibaren kullanılan Web 2.0 teknolojisi ile kişilerin ve kurumların oluşturduğu verilerin yayınlandığı bir ortam olmaktan çıkmış, artık süreklilik içeren birden fazla kullanıcının katılımıyla güncel ve anlık değişen verilerin olduğu bir ortama dönüşmüştür. Kısaca durağan verilerin yayınlandığı bir ortamdan daha fazla güncel veri paylaşımının olduğu bir ortam olmuştur. Video paylaşım siteleri, wiki uygulamaları, forumlar, bloglar bu ortamlara örnek olarak verilebilir. Kişisel paylaşımların bir araya gelerek oluşturduğu bu ortamlarda sosyal medya kavramını ortaya çıkarmıştır. Bu ortamlar ve yeni paylaşım alanı uygulamaları ile kişiler bu sanal ortamları toplumsal buluşma alanları haline dönüştürmüşlerdir. Bu doğrultuda teknolojik alt yapı sayesinde kişilerin birbirinden haberdar olduğu, görüşünü belirttiği, resimlerinin ve videolarını o ortamı kullananlar ile paylaştığı sosyal ağ yapıları oluşmuştur.

2.1.1. Mikrobloglar

Günümüzde kişiler kendilerine ait fikirlerini, yorumlarını ve paylaşmak istedikleri yazı, görüntü ve diğer çalışmalarını web üzerinde yer alan **blog** adı verilen kişisel çevrimiçi (online) ortamlarda yayınlamaktadırlar. Blog alanlarında yayınlanan yazı ve diğer

materyaller bir moderatör veya editör onayı olmadan yayınlanabilmektedir (EduCause Learning Initiative, 2005).

Reporter Without Borders topluluğuna göre, Blog veya Weblog, çoğunlukla haber (mesaj) içeren, düzenli olarak güncellenen, bir günlük şeklinde son yazılanın üstte olduğu, kategori tabanlı düzenlenebilen, genellikle anonim özellikli, tek kişi tarafından işletilen kişisel web sitesidir (Reporter Without Borders, 2005).

Mikrobloglar da geleneksel bir blog çeşididir. Mikroblog ve geleneksel blog arasındaki temel fark mesajların uzunluğudur. Bloglarda yayınlanacak yazının uzunluğunda kısıtlama yoktur fakat mikroblog ortamlarında daha hızlı ve etkili iletişim sağlanabilmesi için yazı boyutunda karakter sayısı kısıtlamaları uygulanır. (Java ve diğ., 2007)

Mikrobloglar kullanıcılara kısa cümleler, anlık fotoğraflar veya video linkleri gibi küçük içerik parçalarını paylaşmak için ortam sağlar (Kaplan Andreas M., 2011). Mikroblog hizmetlerinde mesajların ücretsiz olması, biçim ile ilgili kısıt olmaması ve mikroblog platformlarına kolay erişilebilirlik internet kullanıcılarının geleneksel iletişim araçları (geleneksel bloglar veya e-posta listeleri gibi) yerine mikroblog hizmetlerini tercih etmelerine neden olmuştur (Pak ve Paroubek, 2010). Dünya üzerinde kayda değer mikroblog hizmetleri olarak Tumblr, Cif2.net, Plurk, Jaiku ve identi.ca bulunmaktadır (Stieglitz ve Dang-Xuan, 2014).

Efron (2011)'a göre, mikrobloglar kişisel veya bir kuruma ait olabilir. Mikroblog yazarı periyodik olarak paylaşımlar yapar, okuyucuları da (takipçi olarak isimlendirilirler) toplu olarak bu paylaşımları görürler. Çoğu mikroblog oldukça kişiseldir ve okuyucu sayısı azdır. Toplum içinde tanınmışlığı olan kişilerin mikroblogları ise geniş bir kesime hitap etmektedir. Bu kişilerin mikroblogları hayranları tarafından yoğun bir şekilde takip edilmekte, hayranları kişinin birebir görüş, bilgilendirme ve medya içeren paylaşımlarını anında görebilmektedir. Genel olarak birçok kişi az sayıda seçkin mikroblog yazarlarını takip etmektedir.

Niu ve diğ. (2012)'ne göre mikrobloglar beş özelliğe sahiptir, bunlar;

1. **Büyük ölçekte veri içerirler.** Mikroblog kullananların sayısı, bilgisayar ve akıllı mobil iletişim cihazlarının kullanımının artmasıyla doğru orantılı olarak hızla

artmakta, bu durumda da mikrobloglar üzerinde yer alan görüşlerin veri boyutu büyümektedir.

2. **İnternet üzerindeki dile ait özellikler kullanılır.** Kişiler günlük yaşamdaki konuşma veya yazım dilinden farklı olarak siber dili kullanılır. Geleneksel derlem üzerinde yapılan makine öğrenmesi çalışmalarından farklı olarak siber dili dikkate alan makine öğrenmesi çalışmaları yapılması gerekir.
3. **İçerikte çeşitlilik içerirler.** Bilindik sosyal yaşama ait sabit konu başlıkları hakkında yazılan yazılar olduğu gibi çoğunlukla kişiye göre değişen (kişilerin kendi yaşamlarının bir yönü veya ilgi alanları gibi) farklı konular içerik olarak paylaşılmaktadır.
4. **Gerçek zamanlıdırlar.** Bir mikroblog mesajı yayınlandığında, güncellendiğinde veya silindiğinde diğer insanlar tarafından aynı anda bu işlemler görülmektedir.
5. **Oldukça kısa mesajlardan oluşurlar.** Her türlü ayrıntıyı ve kişisel duyguları ayrıntılı bir şekilde açıklayan mesajlar yerine, öz ama etkili kısa mesajlar halinde yer alırlar.

2.1.2. Twitter

2006 yılında kurulan ABD merkezli mikroblog sitesi Twitter da, kullanıcılar tweet (tivit) adı verilen 140 karakterlik mesaj niteliğinde içerik yayınlayabilirler, kullanıcıların attığı tweetler herkes tarafından görülebilir, kullanıcılar istedikleri kullanıcıların attığı tweetlerden haberdar olmak için takipçi olabilirler (Twitter, 2016). Tüm kullanıcılar Twitter'ın web sitesi aracılığıyla tweet atabilir (mesaj yayınlayabilir) veya başkalarının mesajlarını izleyebilir. Kullanıcılar Twitter uygulamasına harici uygulamalar kullanarak da erişim sağlayabilirler (akıllı telefonlar gibi).

Twitter mikroblog hizmetini kullanan kullanıcıların kısa mesajlar ile kendi düşünce ve durumlarını milyonlar seviyesine kadar çıkabilen "takipçi" lerine aktarabilmeleri, takipçi konumundaki kullanıcıların da takip ettikleri kişilerin yazmış oldukları mikroblog mesajlarına kendi hesapları üzerinden anlık ve aracısız direkt ulaşabilmeleri, bu servisi kullananlara son derece çekici gelmektedir (Efron M. , 2011). Twitter mikroblog yapısı üzerinde paylaşılan içerik farklı birçok alan (kişisel yaşam, güncel eğilimler, önemli dünya olayları, vb.) ile bağlantılı olmaktadır (Korenek ve Simko, 2014).

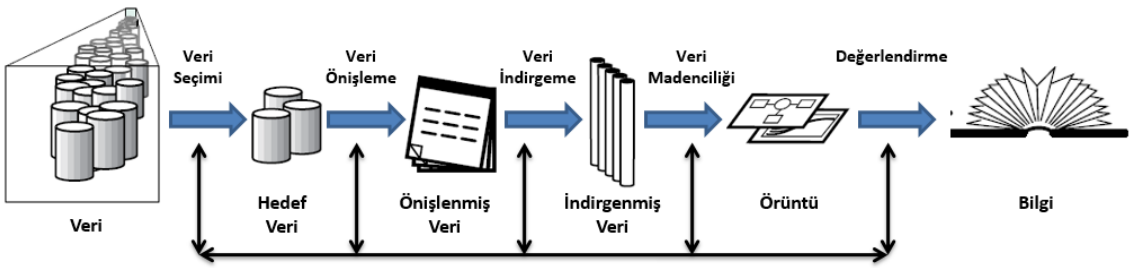
Twitter'ın Eylül 2016 sonunda açıkladığı kendi resmi rakamlarına göre aylık aktif kullanıcı sayısı 313 milyondur. Bu sosyal ağ sitesinin aktif kullanıcılarının %82'si Twitter uygulamasını mobil cihazlar üzerinde de kullanmaktadır (Twitter, 2016).

2.2. VERİTABANLARINDA BİLGİ KEŞFİ

Günlük yaşamımız içinde yer alan birçok süreçte sürekli olarak veri üretilmektedir. Bir kelime işlemcide, bir hesap tablolama programında açıp incelediğimiz veya oluşturduğumuz her doküman, internet ortamındaki kısa mesajlaşma (chat), görüntülü görüşme yazılımları üzerinden yapılan her yazışma ile konuşma metinleri ve görüntüler bizim için veridir (Gülseçen, 2012). Internet tarayıcıları ile açtığımız her sayfanın içindeki dokümanlar ve her sayfa üzerinde yapmış olduğumuz tıklamalarda veridir.

Fikir grupları ve fikirleri ortaya çıkarmak için oluşturulan ortamlardaki görüşler de veri olarak değerlendirilir. Fikirlerin (düşüncenin) analiz edilmesi sonucunda ulaşılmak istenen, değeri olan bilgidir. Büyük veri setlerinden yararlanmak için şablonlar ve kurallar uygulanarak değerli bilginin keşfedilmesine **veritabanlarında bilgi keşfi - VTBK** (*Knowledge Discovery in Databases – KDD*) adı verilmektedir (Fayyad ve diğ., 1996). Bazı araştırmacılar veritabanlarında bilgi keşfi ile veri madenciliğini eş anlamlı olarak kabul etmelerine rağmen genel görüş, veri madenciliğinin veritabanlarında bilgi keşfi sürecinin bir aşaması olduğu şeklindedir (Özçakır, 2006).

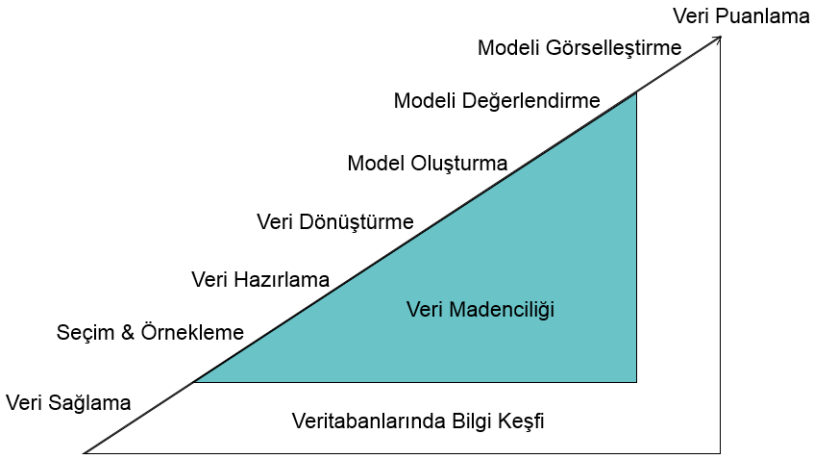
Fayyad ve diğ.,(1996)'ne göre veritabanlarında bilgi keşfi sürecinde yer alan adımlar; veri seçimi, veri önişleme, veri indirgeme, veri madenciliği ve değerlendirme (yorumlama, doğrulama) şeklinde sıralanmaktadır (Şekil 2.1)



Şekil 2.1: Veritabanlarında bilgi keşfi (Fayyad ve diğ., 1996; Gülseçen, 2012).

Frawley ve diğ. (1992)'ne göre bilgi keşfi, önceden bilinmeyen, önemi yüksek, örtük ve potansiyel olarak kullanışlı enformasyonu veriden çıkarma işlemidir. Nispet ve diğ.

(2009)'ne göre ise bilgi keşfinin tüm süreçleri, veriye erişim, veri arama, veri hazırlama, modelleme, model dağıtım ve model izleme olarak sıralanabilir. Şekil 2.2'de gösterildiği gibi bu geniş işlem içinde veri madenciliği veritabanlarında bilgi keşfinin bir süreci değil, bazı süreçleri içinde barındıran bütünleşik bir katmandır.



Şekil 2.2: Veritabanlarında bilgi keşfi ile veri madenciliğinin ilişkisi (Nisbet ve diğ., 2009).

Maimon ve Rokach (2005)'a göre veritabanlarında bilgi keşfi hazırlık aşaması ile başlar. Bu aşama, çalışma kapsamında yapılacak işlemleri anlama, veri dönüşümü, algoritmalar ve sonucun gösterimi gibi konulara karar verme sürecidir. Çalışmadan sorumlu olan kişiden son kullanıcıya kadar herkesin çalışmanın hedeflerini ve bilgi bulma işleminin gerçekleşeceği çevreyi anlaması ve tanımlaması bu süreçte gerçekleşir.

Zaki ve Meira (2014)'ya göre ise veritabanlarında bilgi keşfi veri çıkarma, veri temizleme, veri bütünleştirme, veri azaltma, büyük bir oranda veri madenciliği ve elde edilen örüntü (desen) ve modelin yorumlanması aşamalarından oluşmaktadır. Bilgi keşfi ve veri madenciliği süreci arasında etkileşim yüksektir.

2.2.1. Veri Seçimi

Veri olmadan hiçbir süreç başlamayacaktır, ilk olarak kullanılabilir veri kaynaklarının belirlenmesi veya veri kaynakları biliniyorsa verinin bulunduğu kaynaklardan üzerinde çalışma yapılacak verinin seçilmesi gerekir (Nisbet ve diğ., 2009). Amaç hedef veri kümesini oluşturmak, veri seti seçilirken hangi değişken, özellik ve veri noktalarından (örneklerinden) yararlanacağına karar vermektir (Fayyad ve diğ., 1996a). Bu adımda genellikle alt veri kümelerini seçmek için mevcut büyük veri kümeleri sorgulanır (Cios

ve diğ., 2007). Veri düzenli bir kurumsal veri ambarında yer alabileceği gibi, ideal yapıya sahip olmayan farklı veri kaynaklarında da olabilir.

Hedef verilere erişmek için ek çalışmalar ve yapılar kurulması gerekebilir. Farklı kaynaklardan sağlanacak verilerin birbirleri ile entegre olması gerekecektir, bu neden ile kaynaklardan tam ve eksiksiz veri alma işlemi için çok iyi planlama yapılmalıdır. Eğer hatalı veya eksik veri alımı gerçekleşirse sürecin tekrarlanması gerekecek, bu durum da zaman kaybına ve farklı kaynaklarda kayıplar oluşmasına sebep olacaktır. Doğru bir değerlendirme yapılmaz ise tüm çalışma başarısız olabilir.

Veritabanlarında bilgi keşfi ve veri madenciliği çalışmalarında veri seçimi (veri toplama) genellikle tüm süreçte fazla çaba harcanan aşamalardan biridir. Çalışmanın bu aşamasına başlarken bir dizi halinde tüm verileri bir araya getirmek gerekir. Gerçek bir iş uygulamasında, farklı bölümlerden (pazarlama, satış, müşteri fatura, müşteri hizmetleri, vb.) gelen verileri bir araya getirmek veri seçimine örnek olarak verilebilir. Farklı kaynaklardan gelen verileri entegre etmede genellikle karşılaşılan sorunlar, kayıt tutma ortamı farklılıklarından, farklı veritabanı yapısından, farklı zaman dilimlerinde veri toplanmasından veya farklı biçimdeki verilerden kaynaklanmaktadır. Kurum çapında verilerin entegre, düzenli ve temiz biçimde depolanması veri ambarı sayesiyle gerçekleşir. Veri ambarları ilgili bölümlerin kurumsal ya da örgütsel verilere tek bir noktadan tutarlı olarak erişim yapılmasına imkân sağlar. Veri ambarlarında depolanan veriler iş kararları oluşturmak ve yöneticileri bilgilendirmek amacıyla kullanıldığı gibi, ayrıca depolanan stratejik değere sahip veriler günlük operasyonel süreci desteklemek içinde kullanılır. Özellikle veri madenciliği çalışmaları için veri ambarları sürecin başarılı bir şekilde işleyebilmesi için her adımda yer alması gereken bir yapıdır (Witten ve diğ., 2011).

Veri toplama işlemine mevcut en iyi veri seti ile başlanmalı ve devamında işlemi genişleterek diğer veri setlerinden veri alımı yapılmalıdır (Maimon ve Rokach, 2005).

2.2.2. Veri Hazırlama (Önişleme)

Bugünün dünyasındaki veritabanlarında genellikle büyük boyutta, çok farklı kaynaklardan gelen, muhtemelen gürültülü, eksik ve tutarsız veriler tutulmaktadır. Bu veriler üzerinde bütünleşik bir çalışma yapılabilmesi için birden fazla işlem yapılması

gerekmektedir. Veri üzerinde modelleme yapılarak veri madenciliği çalışması öncesi yapılan bu işlemlere veri önişleme adı verilmektedir. Farklı veri önişleme teknikleri vardır, bunlar (Han ve Kamber, 2006);

1. **Veri Entegrasyonu (Bütünleştirme):** Veri ambarları veya tutarlı veri depoları gibi birden çok kaynaktan gelen verileri birleştirmek için kullanılır.
2. **Veri Temizleme:** Veriler üzerindeki gürültü ve tutarsızlıkları ortadan kaldırmak için uygulanır.

Veri setlerini birleştirmek kolay bir işlem değildir, önişleme süreci oldukça zaman almaktadır. Genellikle veri farklı biçimlerde, farklı seviyelerde ya da farklı birim cinsinden ifade edilmiştir. Bütünleme faaliyetlerinin büyük bir bölümü veri haritası inşa etmek üzerinedir. Bu harita her veri setindeki birbiri ile ilişkili her verinin ortak bir biçimde ve ortak bir kayıt yapısında olmasını nasıl sağlanacağını gösteren yapıdır. İlişkisel veritabanlarında veriye ya da kayıta (bir tablodaki tek satıra) veri haritası doğrultusunda erişmek ve ortak yapıya dönüştürmek için doğrudan veritabanlarına erişim gerekir. Bu işlem bazen veri analiz araçları ile çoğunlukla da veritabanına doğrudan erişen araçlar ile gerçekleşir. Önişleme faaliyetlerini doğru ve uyumlu bir şekilde gerçekleşmesi doğrultusunda birçok veri madenciliği aracı veriler üzerinde oldukça iyi ve doğru şekilde çalışabilir ama düzgün bir veri önişleme yapılmamış ise veri madenciliği çalışmasında dezavantajlar oluşabilir. Ayrıca büyük boyuttaki veri üzerinde veri önişleme yaparken kullanılan donanım ve bellek alanı işlemin önişleme süresinin hızlı veya çok zaman almasında etkendir (Nisbet ve diğ.,2009).

Veri entegrasyonu çok sayıda bilgi kaynağından gelen verileri en uygun şartlarda bir araya getirme aşamasıdır (Maimon ve Rokach, 2005). Veri madenciliğinde zaman ve iş açısından maliyet, modelleme aşamasında daha fazla veri hazırlama, temizleme ve bütünleştirme aşamasındadır (Jackson, 2002). Birçok farklı veri kaynağından gelen veriyi tek bir veri seti üzerinde birleştirme işleminde, veri kaynaklarının teknolojisi (veritabanı teknolojileri), verinin biçim farklılığı (veri tipleri: tamsayılar, ondalık sayılar, karakterler, tarih/saat, boolean, vb.) ve verinin büyüklüğü gibi faktörlere bağlı olarak zorluk ve problemler çıkmaktadır. Çözüm için arabirim-arayüz yazılımları yazılarak bu sıkıntılı süreç rahatlatılabilir, bu işlem için veritabanı yönetim sisteminin alt yapısı ve programlanabilir yapısı kullanılmaktadır. Farklı yazılım firmalarına ait veritabanı yönetim sistemleri birbirleri ile farklı teknolojiler aracılığı ile entegre olabilme yeteneğine

sahiptir, veri madencisi bu teknolojileri doğru bir şekilde kullanırsa bu süreç daha az sıkıntılı olarak çözümlenebilir.

Zafarani ve diğ. (2014)'ne göre, veri üzerinde yapılacak aşağıdaki temizlik işlemleri, veri setinde yer alan verinin kalitesini artıracaktır.

1. **Gürültü**, verilerin bozuk ve çarpık olmasıdır, verinin bulunduğu özellikle ilgisinin sağlanamadığı hallerdir. Örnek olarak kişisel veri tutulan bir veri setinde telefon numarası verisi içinde harflerin yer alması verilebilir. Bu istenmeyen durumlara müdahale edilmemesi durumunda veri madenciliği algoritmalarının performansı olumsuz etkilenebilir. Filtreleme algoritmaları veya veri madencisi tarafında oluşturulacak filtreleme prosedürleri ile gürültü veriler veri setinden çıkarılabilir.
2. **Aykırı Veriler**, veri seti içinde diğer verilerden aşırı seviyede farklı olan örneklerdir. Twitter'da kullanıcıların takipçi sayıları üzerine yapılan bir çalışmada, ortalama takipçi sayısı tespit edilmek istendiğinde çok ünlü bir kişinin takipçi sayısı ortalama değeri olması gerekenden yükseğe çekecektir, bu gibi aykırı verileri veri setinden çıkarmak normal sonuçların elde edilmesini sağlayacaktır.
3. **Kayıp Değerler**, veri seti içinde özellik olarak belirttiğimiz veri alanlarında olması gereken verilerin olmaması durumudur. Özellikler içinde eksik verilerin bulunmasıdır. Kişisel veri tutulan bir veri setinde bazı kişilerin yaşadığı il bilgisinin bulunmaması örnek olarak verilebilir. Bu sorunu çözmek için, (1) eksik olan veriler veri setinden çıkarılabilir, (2) eksik veri tahmin edilebilir (en yaygın değer veya veri sayısal ise ortalama değer tahmini veri olarak atanır) veya (3) eksik veriler görmezden gelinir.
4. **Yinelenen (Tekrarlayan) Veri**, bir özelliğe ait aynı değerler birden çok defa tekrarlandığı durumlardır. Özellikle Twitter ortamında yinelenen tweet mesajları veya tekrar tweetleri (retweet) örnek olarak verilebilir. Özel olarak veri madencisinin yazacağı sorgular aracılığı ile yinelenen veriler veritabanında çıkarılması verinin kalitesini ve tutarlılığını artıracaktır.

Maimon ve Rokach (2005)'a göre de veri önişleme veritabanlarında bilgi keşfi çalışmalarında en fazla zamanın harcandığı süreçtir. Bu aşamada, eksik değerlerin tamamlanması ve temizlenmesi, gürültülü ve aykırı verilerin programlama yapıları ile veya tek tek incelenerek temizlenmesi veri güvenilirliğini artırır. Karmaşık istatistiksel yöntemler ve farklı veri madenciliği algoritmaları ile birçok eksik ve şüpheli veri ya da

güvenirliliği yetersiz özellikler veri setinden çıkarılır. Veri madenciliği algoritmaları bu işlem esnasında eksik veri tahmini yapma noktasında tahminleme modeli olarak kullanılabilir.

2.2.3. Veri Dönüşümü

Veri madenciliği esnasında sıklıkla farklı yapıdaki veri kaynakları içindeki büyük boyuttaki veri setlerinden, özel ve hassas verilerden daha önceden bilinmeyen enformasyonu çıkarmak esastır (Saranya ve Manikanda, 2013). Bu veri kaynaklarındaki veriler birden fazla veritabanı içinde farklı tasarlanmış veri küpleri halinde veya standart düz metin dosyalar şeklinde olabilirler (Han ve Kamber, 2006).

Olson ve Delen (2008)'e göre, veri dönüşümü veri analizi amacıyla veriler üzerinde basit matematiksel formüllerin ve farklı sayısal ölçeklerin kullanılmasıdır. Veri dönüşümü kapsamında sayısal veriler farklı ölçekteki başka bir sayısal veri biçimine dönüştürülebildiği gibi, kategorik veriler sayısal ölçeklere, sayısal veriler kategorik verilere dönüştürülebilir. Veriyi daraltmak ve küçültmek için de veri dönüşümü kullanılabilir.

Veri önişleme teknikleri ile veri dönüşümü ve veri indirgeme teknikleri birbirlerinden tamamen ayrık değildir, kullanım açısından bu aşamaların tamamı geçerlidir, veri seti ve çalışılan veri madenciliği algoritmasına göre bu teknikler bu üç aşamada farklı amaç için uygulanabilir. Veri dönüşümü verileri madencilik için uygun biçimlere dönüştürme sürecidir. Bu süreçte, düzgünleştirme (smoothing), bütünleme (aggregation), genelleme (generalization), normalleştirme (normalization) ve özellik oluşturma (attribute/feature construction) gibi teknikler kullanılır. Bu tekniklerden düzgünleştirme ayrıca veri temizleme esnasında kullanılan gürültülü veri temizleme işlemidir, bütünleme, genelleme ve özellik çalışması ise veri indirgeme esnasında uygulanan veri setini küçültme işlemleridir (Han ve Kamber, 2006).

Veri temizleme ve veri indirgeme esnasında mevcut veri değiştirilerek yeni bir biçim kazandırıldığından yapılan işlemler esasen bir veri dönüşüm işlemidir. Bu neden ile veri dönüşümü, veri temizleme ve veri indirgeme aşamaları içinde kullanılan ara bir aşama olarak göze çarpmaktadır. Bu aşamada kullanılan teknikler duruma göre veri temizleme veya veri indirgeme teknikleri olarak ifade edilmektedirler.

Veri dönüştürme ile veri madenciliği algoritmalarının doğruluğunu ve etkinliğini artırabilmek için veri üzerinde normalleştirme işlemleri uygulanır (Han & Kamber, 2006). Veri normalleştirmesi ile veri setindeki bir özelliğe ait verilerin orijinal değerleri yerine özel bir aralık içinde gösterilmesi ile daha geniş bir alana dağılmış veri yerine belli sınırları olan anlamlı veri kümeleri oluşmaktadır. Normalleştirme ile karmaşık veritabanlarına yerine basit, anlaşılır ve üzerinde kolay analiz yapılabilir veritabanları oluşturulur (Goyal ve diğ., 2014).

Veri normalleştirme için birçok yöntem vardır, bunlardan sıklıkla kullanılanlar aşağıdadır (Han ve Kamber, 2006);

Minumum Maksimum (Min-Max) Normalleştirme: Özelliğe ait verileri daha küçük ve dar bir aralık içinde temsil etme, orijinal veriler üzerinde doğrusal dönüşüm gerçekleştirme işlemidir (-1, +1 aralığı veya 0,+1 aralığı gibi).

Z-Skor (Zero Score - Mean) Normalleştirme: Bir özelliğin değerlerini (verilerini) ortalama ve standart sapma işlemleri uygulayarak belli bir aralığa çekme işlemidir.

Ondalık Ölçekleme (Decimal Scaling) Normalleştirme: Özelliğe ait tamsayı niteliğindeki değerleri yüzdesel olarak ifade edilebilecek ondalık sayıya bölme işlemi yaparak dönüştürme işlemidir. Bu işlem sonucunda değer büyüklüğüne bakılmaksızın tüm değerler 0 ile ±1 arasında bir değere dönüşecektir.

2.2.4. Veri İndirgeme

Veri indirgeme, veri üzerindeki gereksiz özellikleri ya da kümeleri ortadan kaldırarak veri boyutunu azaltmaktır (Han ve Kamber, 2006). Veri madenciliği uygulamasına yarar sağlayacak özellikleri ortaya çıkarma, veri seti üzerinde boyut azaltma ve dönüşüm teknikleri ile ihtiyaç dışı özellikleri veri setinden çıkararak gerçek veri setini temsil edecek veri büyüklüğü açısından normal veri setinden daha küçük temsili veri seti oluşturma aşamasıdır (Fayyad ve diğ.,1996a).

Veri madenciliği için sağlanan veriler çoğu zaman çalışmak istenen boyuttan oldukça fazladır. Veriyi ideal çalışma yapılacak boyuta getirmek için uygulanabilecek işlemlerden birkaçı şunlardır: (Zafarani ve diğ., 2014)

1. **Birleştirme:** Birden fazla özellik tek tek ölçekleri değiştirilerek tek bir özellik altında birleştirilebilir
2. **Ayrıklaştırma:** Çok fazla detay veri içeren bir özellik farklı bir aralık değerine eşlenerek daha anlaşılır olması sağlanabilir (detay veri "Yüksek", "Normal" ve "Düşük" şeklinde ayrıklaştırılabilir).
3. **Özellik Seçimi ve Çıkarma:** Veri setindeki tüm özellikler çoğu zaman yararlıdır, fakat bazen alakasız özellikler olabilir. Genellikle bütün veri setini işleme almak veri madenciliği algoritmalarının performansını etkiler, bu neden ile sadece yarar sağlayacak özellikler ile devam edilip, alakasız özellikler veri setinden çıkarılabilir.
4. **Örnekleme:** Büyük boyuttaki verinin işlenmesi bazen imkânsız olabilir, tüm veri seti üzerinde verinin belli bir bölümü ile çalışma yapmak için örnekleme işlemi gerçekleştirilebilir. Örneklem seçiminde amaç seçilen örneklem üzerinden elde edilecek sonuçların tüm veri seti üzerinden elde edileceklere yakın olmasını sağlamaktır.

Veri indirgeme teknikleri ile veri setinin orijinal veri bütünlüğünü koruyarak hacmi daha küçük temsili veri kümesi oluşturmak amaçlanır, oluşan bu temsili veri seti üzerinde uygulanacak veri madenciliği teknikleri ile elde edilecek analitik sonuçların tüm veri seti üzerinde uygulandığında çıkacak sonuçlar kadar verimli olması gerekir (Han ve Kamber, 2006).

Zafarani ve arkadaşlarından farklı olarak Han ve Kamber (2006)'e göre ayrıca aşağıdaki veri indirgeme stratejileri de kullanılabilir:

1. Veri küpü bütünleştirme
2. Boyut indirgeme
3. Çokluk indirgeme

Veri küpleri çok boyutlu bilgileri depolamaktadır. Her bir hücre çok boyutlu bütünleştirilmiş bir veri değerini içerir. Veri küpleri özellikle **Çevrimiçi Analitik İşleme** (OLAP) süreçlerinde kullanıldığı gibi veri madenciliği çalışmalarında da verilere hızlı erişim imkânı sağlar.

Tipik veri madenciliğinde kullanılan veritabanlarında milyonlarca kayıt (satır) ve binlerce değişken olabilir. Genel de değişkenler arasında herhangi bir korelasyon yapısı olmadığı, bağımsız bir yapıya sahip oldukları görülür. Boyut indirgeme yöntemleri belirleyici

değişkenler arasındaki korelasyon yapısını kullanarak belirleyici değişkenlerin sayısını azaltmak, bu bileşenlerin bağımsız olmasını sağlamak, sonuçların yorumlanabileceği bir çerçeve sağlamak için kullanılır. Bu amaçları gerçekleştirmek için temel bileşenler analizi, faktör analizi ve kullanıcı tanımlı bileşikler yöntemleri kullanılır (Larose, 2006).

Çokluk indirgeme, verilerin tamamı yerine kümeleme, örnekleme veya histogram kullanarak yapılan veri daraltma işlemleridir (Han & Kamber, 2006). Ayrıca ayrıklaştırma ve kavram hiyerarşisi de güçlü araçlardır ve veri madenciliği ile elde edilecek sonuçları güçlendirmektedirler.

2.2.5. Değerlendirme

Veri Madenciliği yöntemleri uygulandıktan sonra gelen adım olan değerlendirme, veri madenciliği teknikleri ile elde edilen örüntülerin değerlendirildiği ve yorumlandığı, veri tabanlarında bilgi keşfi anlayışı ile oluşturulan modelin sonuçlarının görselleştirildiği aşamadır (Fayyad ve diğ. 1996a). Oluşturulan modelin sonuçlarının beklenen hedefler doğrultusunda değerlendirildiği, sıklıkla elde edilen değerlendirme sonuçları doğrultusunda tekrardan önceki adımlarda düzeltme ve değişiklik yapılması ihtiyacını belirleyen ve önceki adımların tekrar edilmesini sağlayan, çeşitli görselleştirme teknikleri, istatistik ve yapay zekâ araçları ile sonuçları son kullanıcılara aktaran işlemler bütünlüğüdür (Olson ve Delen, 2008).

2.3. VERİ MADENCİLİĞİ

Veri madenciliği, büyük veri yığınları içerisinden gelecek ile ilgili tahminde bulunabilmemizi sağlayabilecek bağıntıların bilgisayar programı kullanarak aranmasıdır, veri madenciliği veritabanlarında bilgi keşfinin bir adımını simgelemektedir (Han & Kamber, 2006).

Veri madenciliği, makine öğrenmesi algoritmalarının kullanarak büyük, gürültülü ve dağınık veri setleri içindeki veriler arasındaki daha önceden görülmeyen ilişki örüntülerini bulma, bulunan örüntüler ile geleceğe yönelik doğru kararlar verilmesini sağlayacak tahmin modelleri geliştirmektir. Veri madenciliği 1990 yıllarda gelişen yeni bir alandı, 2000 yılların başından itibaren farklı alanlar ile birlikte kullanılan çok işlevsel bir alan şekline dönüştü, bu alanlar (Nisbet ve diğ., 2009);

- Geleneksel istatistiksel analiz (veri setleri arasındaki ilişki tespiti için tümdengelim yöntemi izlenir),
- Yapay zekâ (uzman sistemler, vb.),
- Makine öğrenmesi teknikleri (yapay sinir ağları, karar ağaçları, vb. tekniklerde tümevarım yöntemi izlenir),
- Büyük veri tabanlarının geliştirilmesi.

Hand ve diğ. (2001)'ne göre veri madenciliği faaliyetleri aşağıdaki genel işlemleri kapsamaktadır.

1. **Veri Analizi Araştırmaları:** Bu veri arama faaliyetleri interaktif ve görsel teknikler içerir. Veri setlerinden fikir almak için istatistiksel parametre ve grafik gösterimler ile örüntü veya eğilimler alınır.
2. **Tanımlayıcı Modelleme:** Bu faaliyet ile üst düzey veri seti görünümleri biçimlendirilir. Bunlar;
    - Verinin genel olasılık dağılımlarının belirlenmesi (yoğunluk tahminleri),
    - Değişkenler arasındaki ilişkiyi açıklayan modeller (bağımlı-ilişkisel modelleme),
    - Gruplar içindeki verinin bölümlendirilmesi ya da küme analizinin yapılması,
3. **Tahmin Edici Modelleme (Sınıflandırma ve Regresyon):** Burada amaç bir değişkenin değeri ile diğer değişkenlerin değerlerini tahmin edecek bir model oluşturmaktır. Sınıflandırma *kategorik* değişkenler için kullanılır (örneğin "Evet/Hayır" gibi iki seçimli değişkenler veya "1 ile 5 arası", "Çok Beğenilen – Hiç Beğenilmeyen" gibi çok seçimli cevaplar içeren değişkenler). Regresyon ise "sürekli" değişkenler için kullanılır (örneğin, bir sayısal değer ile farklı bir sayısal değer arasında ondalık sayı olabilecek herhangi bir sayısal değer, bir kişinin yaşı veya kan basıncı gibi değerler).
4. **Desenleri ve Kuralları Keşfetme:** Bu faaliyet işlemsel veritabanlarında sıklıkla birlikte bulunan öğe kombinasyonlarını (örneğin bir markette müşterilerce aynı zaman diliminde genellikle birlikte satın alınan ürünleri) bulma gibi işlemleri kapsar
5. **İçerik Alma:** Bu faaliyet tipi bilinen örüntüler doğrultusunda yeni veri setlerindeki benzer örüntüleri bulmayı hedefler. Bu örüntü tanımlama yaklaşımı sıklıkla metin

çalışmalarında kullanılır (örneğin Word, PDF veya web sayfası niteliğindeki yazılı belgelerdeki metin içeriği analiz edilir).

Veri madenciliği büyük veri içindeki bilgi ve temel anlayışları çekirdek algoritma yapıları kullanarak ortaya çıkarmaktır. Bu işlemi gerçekleştirirken veritabanı sistemleri, istatistik, makine öğrenmesi ve örüntü tanımlama disiplinleri arasında birleştirici bir alan olarak yer alır (Zaki ve Meira, 2014).

Veri madenciliği otomatik veya yarı otomatik sistemler aracılığı ile veri setleri içindeki veri örüntülerini keşfetmektir (Witten ve diğ., 2011). Veri madenciliğinde genel de regresyon (tahminleme için normal regresyon, sınıflama için lojistik regresyon), yapay ağlar, kümeleme ve sınıflama (karar ağaçları ile sınıflama, istatistiki yöntemler ile sınıflama) gibi çok bilindik analitik bilgisayar modelleri kullanılmıştır, ayrıca birliktelik-ilişki kuralları (association rules), sıralı örüntüler (sequential patterns), zaman serileri (time series-sequence), bulanık (fuzzy) veri madenciliği yaklaşımları, destek vektör makineleri ve genetik algoritmaları gibi farklı yöntemlerde kullanılmaktadır (Olson ve Delen, 2008).

2.3.1. Sınıflama

Veri madenciliği için en yaygın kullanılan uygulamalardan biri olan (Bremer, 2007) ve tahmin edici bir yöntem olan **sınıflama** (*classification*), ortak özelliklere sahip kayıtların farklı sınıflar içine aktarılmasını belirleyen algoritmalardır, sınıf olmak için her kaydın sınıf içinde yer alan diğer kayıtlarla belirlenmiş bir ortak özelliği olması gerekir (Özçakır, 2006). Sınıflama tekniklerinde sınıflar daha önceden belirlenmiştir, veri setine gelen yeni verinin hangi sınıf içinde yer alacağı tahmin edilir (Ramkumar ve Swami, 1998).

Önceden tanımlanmış sınıflardaki her bir veri öğrenme işlevine yöneliktir ve bu sınıflar ile içindeki veriler **eğitim (veya öğrenme) seti** olarak tanımlanır, herhangi bir sınıfa ait olmayan verilerin otomatik olarak hangi sınıfın üyesi olması gerektiğini tahmin edilmesini sağlarlar (Olson ve Delen, 2008). Eğitim seti olarak yorumlanan veriler **etiketlenmiş** veriler olarak da isimlendirilir, veri setine yeni gelen veya henüz eğitim verisi şekline dönüşmemiş verilere ise **etiketlenmemiş** veri adı verilir, ayrıca eğitim setini test etmek amacı ile kullanılan veriler ise **test seti** olarak tanımlanır.

Sınıflandırma yöntemlerinde amaç eğitim seti verilerinden sınıflandırma kuralları geliştirmektir, bu kurallar genellikle bir karar ağacı içinde örtülü şekilde yer alır ve bu yapı içinden bu kurallar elde edilir (Bremer, 2007). Kurallar içinde eğitim veri setinde yer alan özellikler parametre şeklinde kullanılır ve parametrelerin aldığı değerler doğrultusunda kurallar genellikle şart cümleciği (if/else yapıları) şeklinde oluşturulur (Witten ve diğ., 2011). Geliştirilen sınıflama modelleri ile test seti olarak belirtilen veri veya verilerin hangi sınıflar içinde yer alacakları sınıflama kurallarını sağlamaları doğrultusunda otomatik olarak tahmin edilerek sınıflama işlemi test edilmektedir (Bremer, 2007).

Sınıflama modelinin inşa edilmesinde karar ağaçları, yapay sinir ağları, doğrusal programlama ve istatistik gibi matematik teknikleri kullanılır (Olson ve Delen, 2008). Sınıflandırma için ayrıca olasılığa dayalı sınıflandırıcılar, destek vektör makineleri ve benzeri bir birçok farklı yöntem önerilir (Zaki ve Meira, 2014). Bilindik karar ağacı algoritmaları ID3, CART, C4.5, C5, vb. algoritmalardır. İstatistiğe dayalı algoritmalar ise Regresyon, Lojistik Regresyon, Zaman Seri Analizi, Bayes, Naïve Bayes, vb. algoritmalardır.

2.3.2. Kümeleme

Aydoğan (2003)'a göre **kümeleme** (*clustering*) tanımlayıcı bir yöntemdir, küme veya sınıflar içindeki verinin belirlenmiş benzerlik kriterleri doğrultusunda gruplanmasıdır. Amaç, verileri alt kümelere ayırmaktır. Sınıflama algoritmasında olduğu gibi ortak özellikleri olan veriler bir kümeye girer. Alt kümelere ayırmak için keşfedilen kurallar ile bir verinin hangi alt kümeye girdiği kümeleme algoritmaları ile bulunur.

Kümeleme modelinde, sınıfları bulunmayan veriler belirlenen benzerlik-yakınlık kriterlerine göre gruplar halinde kümelere ayrılırlar (Ramkumar ve Swami, 1998). Kümeleme veri madenciliği bölümlemeli (ağırlık merkezli) yöntemler (K-means, K-medoids, PAM, CLARA, CLARANS vb. algoritmalar), hiyerarşik yöntemler (AGNES, DIANA vb. algoritmalar), yoğunluk tabanlı yöntemler (DBSCAN, OPTICS, DENCLUE vb. algoritmalar), ızgara tabanlı (grid temelli) yöntemler, model tabanlı yöntemler şeklinde sınıflandırılmaktadır.

2.3.3. Birliktelik-ilişki kuralları

Birliktelik-ilişki kuralları (Association Rules), geçmiş verilerin analiz edilerek bu veriler içindeki birliktelik davranışlarının tespiti ile geleceğe yönelik çalışmalar yapılmasını destekleyen bir yaklaşımdır (Sever ve Oğuz, 2002). Birliktelik-ilişki kuralının uygulandığı algoritmaların bazıları şunlardır; AIS, Apriori, DHP, Partition, Eclat, FP-Growth (Hipp ve diğ., 2000).

2.3.4. Metin Madenciliği

Veri madenciliğinin alt dalı olan metin madenciliği metinden yüksek kaliteli, kullanışlı bilgi çıkarma işlemidir (Narayanan, 2010). Metin madenciliğini veri madenciliğinden ayıran en büyük fark metin madenciliğinde kalıpların düzgün veritabanlarından çok, doğal dil metinlerinden çıkarılmasıdır (İlhan ve diğ., 2008). Veri madenciliği çözümleri ve algoritmaları metin veya web verisindeki kalıpları bulmadan veya model oluşturmadan önce metin veya web verisinin yapısal olması gerekmektedir, metin ve web madenciliği işlemleri, veri madenciliğinde kullanılacak yapısal veriye ulaşmak için kullanılan araçlar olarak tanımlanabilir (Dolgun ve diğ., 2009).

Öğüdücü (2012)'ye göre metin (text) madenciliği, veri madenciliği teknikleri ile yazılı belgeler arasındaki (içindeki) ilişkileri, örüntüleri bulma tekniğidir. Metin madenciliğinde doğal dilde yazılmış metinler üzerinde aynı konudaki belgeleri bulma, birbiriyle ilişkili belgeleri bulma, bulunan belgeleri sıralama işlemleri yapılır. Metin madenciliğinin belirgin sınırlılıkları ve sorunları vardır. Metinler yapısal değildir, sorgulama sonucuyla ilintili metinleri bulmak zordur, önemsiz veri (kelime) çok fazladır, metin içinde hatalar vardır, kavram oluşturmak zordur (eş anlamlı, eş sesli, üst grup, vb.), anlam çıkarmak zordur.

Metin madenciliğinde, veriden bilgi çıkarma yöntemlerinden biri olan doğal dil işleme disiplini ile bilgi çıkarımında daha anlamlı sonuçlar elde edilmeye başlanmış, insan-bilgisayar etkileşiminin arttırılması başarılmıştır (İlhan ve diğ., 2008).

2.4. DOĞAL DİL İŞLEME

Doğal Dil İşleme (DDI, *Natural Language Processing - NLP*), ana işlevi bir doğal dili çözümleme, anlama, yorumlama ve üretme olan bilgisayar sistemlerinin tasarımını konu alan bir mühendislik alanıdır (Rich ve diğ., 1991).

Dolgun ve diğ. (2009)'ne göre metin yazımında standart kurallar olmadığından dolayı bilgisayar bunları anlayamamaktadır. Her bir metnin dili ve içerdiği anlam amaca bağlı olarak çeşitlilik göstermektedir. Yapısal olmayan bilgiden içerik çıkarmak için kullanılan geleneksel yöntemler; anahtar kelimeler veya mantıksal aramalar, istatistiksel veya olasılıksal algoritmalar, sinir ağları ve kalıp keşfedici sistemler gibi dilbilimsel olmayan yöntemlerdir. Bu yöntemler, hem sorgudaki hem de metindeki kelimelerin karakterlerini karşılaştıran bir temele dayanır. Bundan dolayı içeriği açıklayıcı sonuçlar elde edemez. Dili anlamanın temeli dilbilimsel yollara dayanır ve bu çoğunlukla **Doğal Dil İşleme** olarak ifade edilir. DDI'yı içeren bir sistemde, karmaşık yapıların bulunduğu ifadeler (örneğin; duştan akan soğuk su ile içilen soğuk su arasındaki fark gibi) akıllı olarak çıkarabilmekte ve terimleri sınıflayarak; ürünler, organizasyonlar veya kişiler gibi sınıflara atamaktadır.

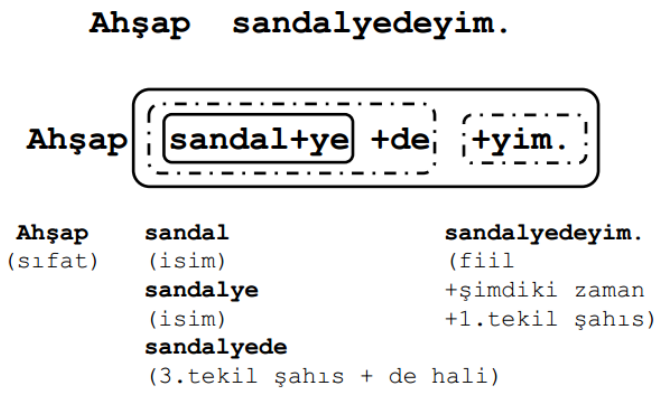
2.4.1. Türkçe Dil Yapısı

Türk Dil Kurumu güncel Türkçe sözlüğüne göre *morfoloji*, dil yapı bilgisidir (Türk Dil Kurumu, 2015). Yunan etimolojisinde *morfem*; şekil, biçim anlamına gelmekte, biyoloji de *morfoloji*; biçim üzerinde çalışma ve organizma yapısı olarak geçmekte, dilbilim de ise *morfoloji*; kelimenin oluşumunu, kelimenin iç yapısını tanımlar ve kelimenin nasıl oluştuğunu inceler (Aranoff ve Fudeman, 2011).

Koç (2007) yeryüzü dilleri ile ilgili çalışmasında, yeryüzündeki dillerin ses sistemi, biçim yapısı ve söz dizimi bakımından bazı yakınlıklar ve benzerlikler gösterdiğini belirtmiştir. Yeryüzündeki diller (dil aileleri) bazı yakınlık ve benzerliklerine göre yapı ve köken olmak üzere iki bakımdan incelenir. Yapı bakımından diller tek heceli, eklemeli (bitişken) ve çekimli (bükümlü) diller şeklinde kendi içinde de ayrılmaktadır. Eklemeli (bitişken) dillerde bir veya daha çok heceli köklere yapım ve çekim ekleri eklenir. Getirilen ekler kökle kaynaşmışlardır. Köke getirilen yapım ekleri ile yeni sözcükler, yeni kavramlar türetilir. Yeni ekler ulandığında kökte bir değişiklik olmaz. Türkçe eklemeli (bitişken) dil

ailesi içinde yer almaktadır ve bu dil yapısının en uyumlu ve güzel örneğidir. Türkçeye yabancı dillerden giren bazı sözcük köklerine de ekler getirilerek yeni sözcükler türetilir.

Eryiğit ve diğ. (2006)'ne göre, Türkçenin eklemeli bir dil olmasına bağlı olarak kelimelerin sonlarına eklenen ekler ile yeni kelimeler oluşabildiği gibi eklere bağlı olarak kelimelerin türleride (isim, fiil, sıfat vb.) değişebilmektedir (Şekil 2.3). Türkçede "özne-nesne-yüklem (fiil)" veya "nesne-özne-yüklem" yapısında olan cümlelerde içeriğe ve vurguya göre cümle öğeleri yer değiştirebilmektedir. Bu neden ile Türkçenin düzenli bir kalıp dili olduğundan bahsetmek oldukça zordur.



Şekil 2.3: Ekler ile kelime çekimi (Eryiğit ve diğ., 2006).

Türkçe gibi eklemeli dillerde kelimeler bir kök kelimenin sonuna eklenen morfemler morfemler kök kelimenin anlamını, türünü, olumlu – olumsuz – soru olma durumunu ve zamanını belirler (Oflazer ve Bozşahin, 1994). Bu şekilde oluşan bir kelimedeki anlam farklı dillerde bir cümle ile anlatıldığı görülmektedir. Örneğin; "**gözlemlenmeyeceklerdendir**" kelimesi içinde isim, yüklem, olumsuzluk, çoğul kişi, vb. birçok yapı barındırır.

Türkçe cümle yapısını elektronik ortamda çözümlemeye yönelik 1990 yılından bu yana morfolojik çalışmalar yapılmaktadır (Zafer, 2012). Türkçe üzerine yapılan morfolojik çalışmalarda; bütün bir cümleyi kelime kelime, hece hece parçalara ayırarak anlamsal ve yapısal olarak çözümlemek istenmektedir. Kemal Oflazer'in "kural tabanlı gerekirci ayrıştırıcı" sı Türkçede bağlılık analizi konusunda ilk yapılan çalışmadır (Oflazer, Dependency Parsing with An Extended Finite-State Approach, 2003). Günümüzde de bu çalışmalar devam etmektedir, tam çözüme yönelik bir ürün henüz tam olarak

geliştirilmemiştir, birçok akademik çalışma ekibi benzer konu üzerine çalışmalarını sürdürmektedir.

2.5. DUYGU ANALİZİ

Alanyazında **görüş madenciliği** (*opinion mining*) adı da verilen **duygu analizi** (*sentiment analysis*) insanların ürünler, hizmetler, kuruluşlar, bireyler, sorunlar, olaylar, konular ve bu nitelikteki varlıklarla ile ilgili görüşlerini, duygularını ve değerlendirmelerini analiz etmeye yönelik çalışmadır (Liu, 2012). Aggarwal'a göre görüş madenciliği, kısa metin mesajları halindeki müşteri yorum ve görüşleri üzerinden yararlı enformasyon çıkarma işlemidir (Aggarwal, 2012). Duygu analizi farklı isimler halinde de ifade edilir (görüş çıkarma-*opinion extraction*, duygu madenciliği-*sentiment mining*, öznellik analizi-*subjectivity analysis*, etki analizi-*affect analysis*, hissiyat analizi-*emotion analysis*, yorum madenciliği-*review mining*, vb.) (Liu, 2012).

Duygu analizi bir doğal dil işleme ve enformasyon çıkarma işidir. Duygu analizi ile çok fazla sayıdaki doküman üzerinde dokümanı oluşturan kişinin, olumlu/olumsuz yorum, soru veya istek gibi duygularını doğal dil işleme yöntemleri kullanarak yapılan analizler ile elde etmek amaçlanmaktadır (Mukherjee, 2012).

Duygu sınıflandırma, farklı alanlarda çeşitli biçimlerde sunulmaktadır (pozitif /negatif, iyi/köyü, hoş/antipatik, alınır/alınmaz, önerilir/önerilmez, mükemmel/sıkıcı, destekçi/karşı, iyimser/kötümser, olumlu/olumsuz, uygun/uygunsuz vb.). Ayrıca duygu sınıflandırması, tüm dokümanın olumlu veya olumsuz sınıflandırılması (doküman seviyesi), cümlelerin olumlu, olumsuz veya nötr olarak sınıflandırılması (cümle seviyesi) ve ürünlerin özelliklerini belirleme, ayıklama şeklinde sınıflandırılması (görünüş/özellik seviyesi) gibi farklı seviyelendirmeler içinde kullanılır. Duygu sınıflandırma teknikleri üç kategoriye ayrılır: makine öğrenmesi (machine learning) algoritmaları, bağlantı analiz (link analysis) metotları ve skor-tabanlı (skor-based) yaklaşımlar (Singh ve diğ., 2014).

Vinodhini ve Chandrasekaran (2012)'a göre, duygu analizi ile ilgili literatür taramalarında makine öğrenmesi ve anlamsal oryantasyon (semantic orientaion) teknikleri kullanıldığı görülmektedir. Doküman içinde duygu bulmak için genellikle doğal dil işleme teknikleri kullanılır. Makine öğrenmesi yaklaşımlarında duygu analizi

aşamasında genel olarak denetimli sınıflama ve metin sınıflama teknikleri uygulanır. Duygu analizinde doğal dil işleme diğer bilinen bir makine öğrenmesi metodudur. Doğal dil işleme alanında K-En Yakın Komşuluk (K-Nearest Neighborhood) ve Merkezi Sınıflandırıcı (Centroid Classifier) gibi kümeleme, Naïve Bayes (Saf-Yalın Bayes), ID3 ve C5 gibi sınıflama veri madenciliği teknikleri kullanıldığı gibi bu işlem esnasında metin yapısı N-gram modeline göre düzenlenmektedir.

Columbia Üniversitesi'nde Yoon (2011) tarafından yapılan doktora çalışmasında, ABD genelinde fiziksel aktiviteler (aerobik, jogging, yüzme, koşma, basketbol oynama vb.) ile ilgili Twitter soysal ağı ile oluşturulan tweetler üzerinde metin madenciliği kullanılarak analizler yapılmıştır. Analiz öncesinde Visual Basic programlama dilinde yazılan bir uygulama ve WEKA[1] yazılımı ile mesajlar üzerinde temizleme ve düzenleme çalışması yapılmış sonra da doğal bayes ve maksimum entropi ile sınıflama çalışmaları yapılmıştır. Mesajlar üzerinde fiziksel aktivitelerin olumlu ve olumsuz yanları ile ilgili duygu analiz çalışması gerçekleştirilmiştir. (Yoon, 2011)

IOWA Üniversitesi'nde Mayıs 2012'de Yelena Aleksandrovna Mejova tarafından yapılan doktora çalışmasında Doğal Dil İşleme yardımı ile Twitter ortamında yer alan tweetler üzerinde kutuplaşma (polarity classification), sosyal akış madenciliği (mining social stream) işlemleri ile siyasi (politik) konular ile ilgili duygu analizi yapmıştır. Bu çalışma esnasında Etki Kontrol Teorisi (Affect Control Theory) ile çalışılmıştır. Bu teoriye göre ortak dil payı içindeki insanların belirli kültürel normlar doğrultusunda dikte edilen sözcüklerin duygusal anlamlarını kabul ettiği varsayılır. (Mejova, 2012)

Mikrobloglar da yer alan görüşler üzerinde varlık felsefesi (ontoloji) tabanlı duygu analizi ve görüş madenciliği çalışması yapan Uluslararası Hellenic Üniversitesi'ndeki bir grup araştırmacı biçimsel kavram analizi (formal concept analysis) ve varlık felsefesi öğrenimi (ontology learning) yaklaşımlarını kullanmışlardır. Çalışmada mesajlar üzerinde tartışılan konuya karar vermek için ontoloji tabanlı tekniklerin kullanılması savunulmaktadır. Mesajların birbirleriyle ilişkilendirme oranlarına puan verilmesi

[1] WEKA Waikato Üniversitesinde makine öğrenmesi üzerine çalışan ekip tarafından geliştirmiş veri madenciliği görev ve algoritmalarını çalıştırılmasını sağlayan bir veri madenciliği platformudur (The Waikato University, 2013).

öngörülmekte, verilen puanlar doğrultusunda mesajlar arasındaki konu tabanlı bağları ortaya çıkarmak hedeflenmektedir. (Kontopoulos ve diğ., 2013)

2010 yılında Albert Bifet ve Eibe Frank tarafından yapılan bir çalışmada Twitter Akış API'si ile büyük miktarda gerçek zamanlı tweet sağlanmış, Twitter veri akışı üzerinde duygu analizi için Kappa istatistiği kullanılmıştır. Çalışma içinde mesajlar üzerinde Multinomial Naïve Bayes (Çok Terimli Saf-Yalın Bayes), Stochastic Gradient Descent (SGD, Rastlantısal Değişim Eğimi) ve Hoeffding Tree (Hoeffding Ağacı) modelleri ile çalışılmıştır. Çalışma kapsamında yapılan testler yorumlandığında uygun bir öğrenme oranı ile kullanılan SGD tabanlı modelin, Twitter verileri üzerinde duygu analizi için tavsiye edilebilir olduğu görüşü paylaşılmaktadır. (Bifet ve Frank, 2010)

Stanford Üniversitesi kapsamında bir araştırma grubu 2009 yılının bahar aylarında Twitter üzerinde yayınlanan mesajlar üzerinde duygu analizi (sentiment analysis) çalışması başladı (Stanford University, 2013). Bu çalışma kapsamında belirlediğiniz belli bir konu üzerine atılmış son 100 tweet mesajı üzerinde mesajların olumlu veya olumsuz olduğuna yönelik analiz işlemi yapılmaktadır. Sadece İngilizce ve İspanyolca tweetler üzerinde mesajın "olumlu mu?" veya "olumsuz mu?" olduğu tespit edilmektedir. Duygu analizi esnasında sınıflamaya yönelik Anahtar Kelime-Tabanlı (Keyword-Based), Naïve Bayes, Maksimum Entropi ve Destek Vektör (Support Vector Machines) makine öğrenmesi metotları kullanılmıştır. (Go ve diğ., 2009)

İngiltere'de bir üniversiteye ait Bilgi Medya Enstitüsü (The Open University, Knowledge Media Institute) kapsamındaki bir araştırma ekibi 2012 yılında Stanford üniversitesinin Twitter Duygu Analizi (Stanford Twitter Sentiment Corpus – STS) çalışması ile olumlu ve olumsuz ifadeleri (emoticons - :), :-), :D vb.) içeren Twitter dan rastgele çekilmiş 60.000 tweet mesajının oluşturduğu veri seti, Amerika'daki Sağlık Reformu (Health Care Reform – HCR) ile ilgili atılan tweet mesajlarının oluşturduğu veri seti ve Obama-McCain televizyon tartışması (Obama-McCain Debate –OMD) esnasında atılan tweet mesajlarının oluşturduğu veri seti üzerinde duygu analizi çalışması yapmıştır. Çalışmada anlamsal özellikler ve kavramlar (semantic features and concepts) üzerinden çalışma yapılmış, tweet mesajlarındaki pozitif ve negatif duygular ile anlamsal kavramlar (kişi, şirket, şehir, ülke, organizasyon, teknoloji, spor, vb. kavramlar) arasında korelasyon ölçümleri gerçekleştirilmiştir. Anlamsal yaklaşım ile gerçekleştirilen duygu analizi

çalışması büyük veri setlerinde daha uyumlu sonuçlar ürettiği belirtilmektedir. (Saif ve diğ., 2012)

2011 yılında Columbia Üniversitesi'ndeki araştırma ekibi Twitter mesajları üzerinde kutuplaşma (olumluluk, olumsuzluk) yaklaşımı ile duygu analizi gerçekleştirmiştir. Çalışma Twitter ortamında gerçek yazım şekli ile yazılmayan kelimeler (bff-best friend forever, gr8- great) ve şekiller (:), :-), :D) ile ilgili özel sözlükler oluşturulmuştur. Ayrıca Whissel'ın 1989 yılında oluşturduğu Dictionary of Affect in Language isimli bir sözlük çalışma içinde kullanılmıştır. Bu sözlükte yer alan 8000 kelimeye 1 (negatif – olumsuz) ile 3 (pozitif – olumlu) arasında bir puan verilmektedir, mesajlar içindeki kelime bu doğrultuda puanlandırılmaktadır, sonra her kelimenin puanı 3 değerine bölünerek normalleştirilmiştir. Kelime normalleşme sonucunda 0,5 değerinin altında kalırsa olumsuz, 0,8 den yüksek olursa olumlu olarak yorumlanmış. Kelimelerin olumlu, olumsuz ağırlıkları üzerinden de mesajların duygu analizi yapılmıştır. (Agarwal ve diğ., 2011)

2.5.1. Anahtar Kelime-Tabanlı (Keyword-Based) Duygu Analizi

Anahtar kelime-tabanlı duygu analizi literatürde **sözlük-tabanlı** (*lexicon-based*) olarak ta isimlendirilir. Cümle veya metne olumlu veya olumsuz anlam katan kelimelerin metin içinde kullanım oranlarından yola çıkarak, metnin duygu analiz sonucu belirlenir. Genelde kelime-tabanlı çalışmaların çoğu metnin anlamsal yönelimini bulmak için sıfat niteliğindeki kelimeleri gösterge olarak kullanmaktadır (Hatzivassiloglou ve McKeown, 1997).

Türk Dil Kurumuna göre sıfat; "nitelik, nicelik, yer, sıra vb. niteleyen, belirten kelimelerdir" (Türk Dil Kurumu, 2015). Örneğin "Mutlu", "Hüzünlü", "Pişman", "Zalim" gibi sıfatlar bir kişiyi nitelendirmekte ve ayrıca o kişi hakkında bize olumlu veya olumsuz duygu aktarmaktadır.

2.5.2. Kelime Torbası (Bag of Words)

Çok büyük sayıda (milyon veya milyarlarca) metin içerikli belge üzerinde işlem yaparken farklı teknikler uygulanarak etkili sonuçlar elde edilebilir. Metin belgeleri içinde yer alan kelimelerin dağılımları ve kullanım sayıları ile yapılan değerlendirmeler de kullanılan teknik **Kelime Torbası** (*Bag of Words*) olarak isimlendirilir. Bu teknik ile farklı belgeler

içindeki belirli kelimelerin seyrek veya sık bulunmaları, benzer kelimelerin tekrar sayılarının eşleşmesi gibi çıktılar bu farklı belgelerin arasındaki içeriksel ilişkilere ait sonuçlar üretmektedir (Grauman ve Leibe, 2011).

Kelime torbası kavramı bir sözlük içindeki kelimelerin tekrar sayısıdır (Salton ve McGill, 1986). Sınıflandırma çalışmalarında özellikler set olarak yer alır ve metne dayalı sınıflandırma çalışmalarında bu özellikler kelime torbaları şeklinde kullanılır, belge veya cümle içinde sıralı veya sırasız haldeki tüm kelimeler çoklu setler halinde bu torbalar içinde gösterilir (Yoshikawa ve diğ., 2014).

2.5.3. N-gram

Metin içindeki dilin tanınması harf, harf dizileri, kelimeler ve **N-gram** adı verilen kelime veya harf dizileri ile gerçekleştirilir (Bayrak ve diğ., 2012).

Doğan ve Diri (2010)'ye göre; **N-gram**, bir karakter katarının n adet karakter dilimidir ve bu yöntem ile doküman içerisindeki karakterler ile oluşturulmuş N-gram değerlerinin kullanım sıklığına dayalı bir sınıflandırma işlemi gerçekleştirilir.

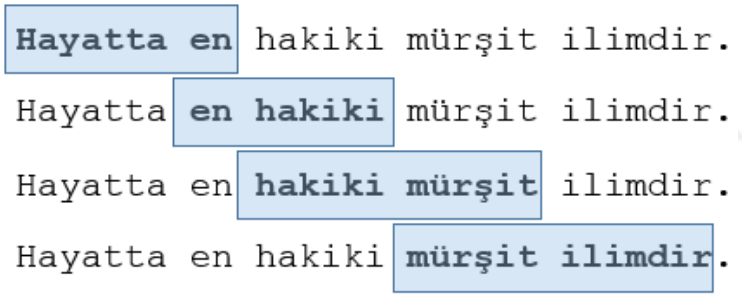
Bayrak ve diğ. (2012)'ne göre; genelde metinsel veri boşluk, noktalama (bazı durumlarda noktalama işaretlerine de anlam yüklenebilir), sayısal karakterlerden arındırma işleminden sonra N-gram değerleri oluşturulur. N-gram değerleri peş peşe gelen n adet harf veya kelimeden meydana gelir. Metinler üzerinde genişliği n (harf veya kelime) olan kayan pencereler yardımıyla o metinde yer alan bütün N-gram değerleri elde edilir. Elde edilen bu N-gram değerlerinden tekrar edenler sayılarak her bir N-gram değerinden kaç adet yer aldığı bulunur ve bu sayısal değer **N-gram profili** adı verilen bir vektörde tutulur. Eğitim ve test metinleri için elde edilen N-gram profilleri doğrultusunda bir sonraki süreç; bu verilere makine öğrenmesi teknikleri ile sınıflandırılmasıdır.

N-gram modelleri cümle ya da bir metin içine aynı anda sadece ***n*** adet kelimenin görünür olduğu küçük bir pencere yerleştirmek olarak hayal edilebilir. En basit N-gram modeli, aynı zaman diliminde sadece bir sözcüğe bakılan **unigram** modelidir. Örneğin, Mustafa Kemal Atatürk'ün "*Hayatta en hakiki mürşit ilimdir*" cümlesi beş unigram içerir: "Hayatta", "en", "hakiki", "mürşit" ve "ilimdir". Cümlenin sadece kelimeler halinde ayrılması çok bilgilendirici değildir ama N-gram özelliği iki veya daha büyük olduğunda

ilginç olmaya başlar. Unigram modeline benzer olarak aynı zaman diliminde pencerede iki kelimenin görünmesi **bigram** olarak ifade edilir. Cümlemiz bu durumda dört bigram içerir:

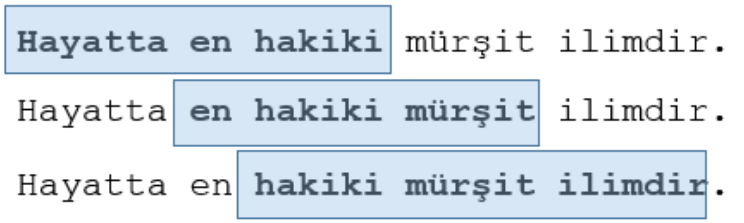
- Hayatta, en
- en, hakiki
- hakiki, mürşit
- mürşit, ilimdir

Cümle üzerinde yer alan tüm bigramlar iki kelime boyutunda bir pencerenin ilk iki kelime üzerinde konumlandığı varsayılırsa, son iki kelimeye ulaşılana kadar her adımda pencere bir kelime sağ doğru kaydırılır. Her adımda pencere içinde kalan kelimeler bigram setini oluşturacaktır (Şekil 2.3) (Kok ve Brouwer, 2010). Aslında unigram modelinde de benzer işlem yapılmaktadır, aynı işlemler **trigram** (üçlü) ve üstü içinde benzer şekilde gerçekleştirilir (Şekil 2.4).



Şekil 2.4: N-gram – 2 gram (bigram) modeli.

N-gram dil modellemesi konuşma tanıma, dil tanımlama, makine çevirisi, karakter tanıma ve konu sınıflandırma gibi birçok alanda kullanılmaktadır ve çoğu dil modelleme tekniği N-gram yaklaşımını benimser (Steffen, 2004).



Şekil 2.5: N-gram – 3 gram (trigram) modeli.

2.6. MAKİNE ÖĞRENMESİ

Teknolojideki hızlı gelişmelere bağlı olarak veritabanlarında depolanan verilerin boyutları büyümekte, bu doğrultu da veritabanı teknolojileri endüstrisi gelişmekte, depolanan bu veriler üzerinde pazarın ihtiyaçları doğrultusunda değerli olan bilgiyi elde etmek için bilgisayar bilimleri alanındaki araştırmalarda veri madenciliği ve makine öğrenmesi gibi yöntemler kullanılmaktadır. **Makine öğrenmesi** (*machine learning*) ağırlıklı olarak veri modellerinin ve desenlerinin keşfedilmesi ile ilgili bilgisayar biliminin oturmuş ve bilindik bir araştırma alanıdır (Fürnkranz ve diğ., 2012).

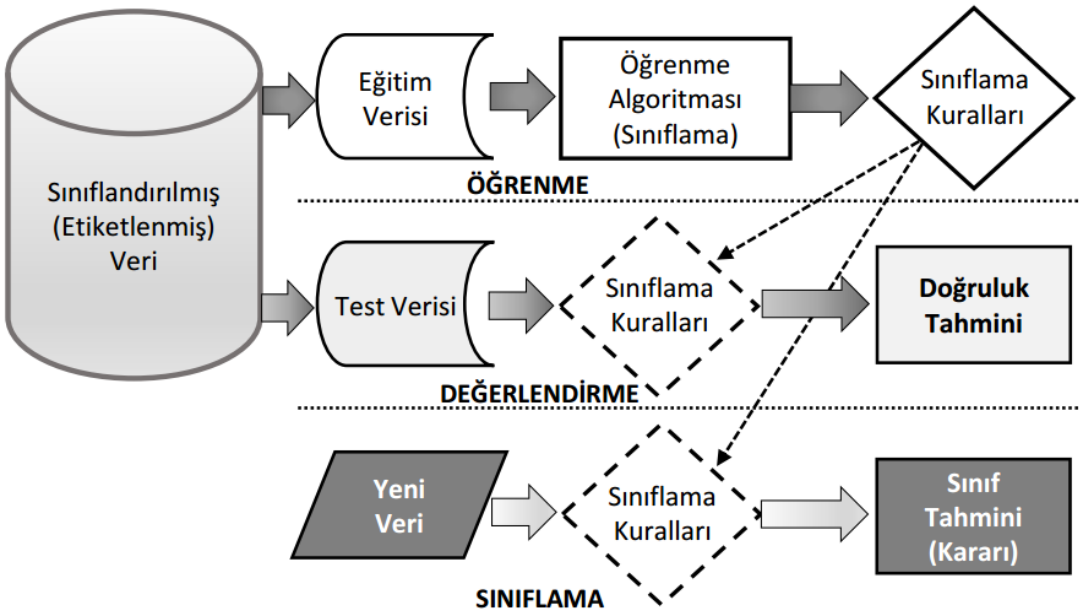
Smola ve Vishwanathan (2008)'a göre, makine öğrenmesi teknikleri bu süreçte kalıcı veri analizi yöntemleri açısından zengin bir yapı ortaya koymaktadır. İnternet üzerinde kullandığımız arama motorları ile arama yaptığımızda kelime veya kelimeler ile ilişkili web sayfalarının çok hızlı olarak bulunup belli bir kural veya öncelikle listelenmesinde, web üzerinden satış yapılan web sitelerinde bir ürünü inceleme esnasında veya gelecekte satın alma kararı verebilmemizi etkileyecek önerilerde bulunulmasında, cümle şeklindeki bir metni bir dilden başka bir dile çevirme sürecinde, bir finans kuruluşuna gelen kredi taleplerinin kredi geri ödeme risklerinin hesaplanmasında veya görsel bir veriden (fotoğraf veya video), el yazısından, belgeden, konumdan ve benzeri verilerden yola çıkarak bir kişiyi ya da bir sorunu ya da gelecekte oluşabilecek bir durumu ortaya çıkarma süreçlerinde veri destekli yazılım ve donanımlarla oluşturulmuş modeller karar verme işlevini yerine getirmektedir. Bu modeller geliştirilirken makine öğrenmesi teknikleri ve algoritmalarından yararlanılmaktadır.

Makine öğrenmesi yaklaşımlarının temeli veri madenciliği sınıflama yöntemi ve algoritmalarına dayanmaktadır. Sınıflama işlemi herhangi bir yapı veya daha önceden oluşturulmuş bütünlüklere (sınıflara) dâhil olmayan bir veri veya öğenin yapı veya sınıfını tahmin etmektir (Zaki ve Meira, 2014).

Makine öğrenmesi teknikleri eğitim ve test seti olmak üzere iki veri seti ile gerçekleştirilir (Pang ve Lee, 2008). **Eğitim (öğrenme) seti** veri veya belgenin farklılaşan özelliklerinin otomatik **sınıflandırıcı** doğrultusunda sınıflandırılmasını öğrenmek için kullanılır. **Test seti** ise sınıflandırıcının performansının ne kadar iyi olduğu kontrol etmek için kullanılır.

Biçimsel olarak, sınıflandırıcı M olarak ifade edilen bir model veya bir fonksiyon olursa, bir giriş değeri olan x'in y olarak ifade edilen sınıflar içinden hangi sınıf içinde olacağını belirler. $y = M(x)$, burada $y \in (c_1, c_2, \ldots, c_k)$ ve her bir c_i ifadesi bir sınıf etiketini (kategorik özellik değeri) temsil eder (Zaki ve Meira, 2014). Model oluşturmak için doğru sınıflandırılmış bir veri veya öğe setine ihtiyaç vardır bu öğe seti **eğitim seti** olarak isimlendirilir. Öğrenme modeli M sonrası, yeni bir veri veya öğenin sınıfını otomatik olarak tahmin edilebilir. Bu süreç esasen veri madenciliği yöntemlerinden sınıflama işlemidir.

Han & Kamber (2006)'e göre, makine öğrenmesi süreci ve sınıflandırma algoritmaları **öğrenme modeli oluşturma** ve **tahmin etme** (*sınıfa yerleştirme*) olmak üzere iki aşamadan oluşur. Mevcut veri setlerinin değişken veya özellik niteliğindeki veri alanları doğrultusunda sınıflara ayrılması, oluşan sınıfların hangi kurallar bütünlüğü doğrultusunda oluştuğunu bir karar modeli olarak tanımlanması **öğrenme modeli** oluşturmaktır. Yeni elde edilen veriler üzerinde oluşturulan öğrenme modelinin çalıştırılması sonrası yeni verinin hangi sınıf içinde olması gerektiğini hesaplamak ise tahmin etme aşamasıdır.



Şekil 2.6: Öğrenme modeli (denetimli).

Denetimli (Supervised) Öğrenme: Bir denetçi tarafından veri seti önceden belirlenen kriterlere göre sınıflara ayrılır, her sınıf bu kriterlere uyan verilerden (örneklerden) oluşur.

Model ile sınıflarda yer alan veriler üzerinden ilgili sınıfa ait özelliklerin bulunmasını, bu özellikler doğrultusunda sınıflama kurallarının oluşturulması amaçlanır (Hastie ve diğ., 2008). Oluşturulan kurallar (şart cümlecikleri) aracılığı ile sisteme alınan/girilen yeni veriler üzerinde model uygulanır ve yeni verilerin hangi sınıfa ait olduğu kurulan model tarafından belirlenir (Şekil 2.6). Bu modelin kurulması için farklı sınıflandırma algoritmalarından biri seçilir, genellikle de bu algoritmalar istatistik tabanlıdır.

Denetimsiz (Unsupervised) Öğrenme: Bu öğrenme modelinde eğitim verisi içindeki sınıflar bilinmez, verilerin gözlenmesi ve verilerin özellikleri arasındaki benzerliklerden dikkate alınarak veri seti içindeki sınıfların belirlenmesidir, bu özelliği ile sınıflandırma işleminden daha çok kümeleme sürecine hizmet eden bir yapıdır (Manning ve Schütze, 1999).

Görüşlerin (fikirlerin -duyguların) sınıflandırılmasında Destek Vektör Makineleri (SVM - Support Vector Machines), Naïve Bayes (Saf-Yalın Bayes) ve Maksimum Entropi (Maximum Entropy) gibi bir dizi makine öğrenmesi yaklaşımı kullanılır (Pang ve Lee, 2008). Makine öğrenmesi yaklaşımları eğitim setinde veri toplanması ile başlar, sonra eğitim seti üzerinde sınıflandırıcı çalıştırılır. Denetimli bir sınıflandırma tekniği seçildikten sonraki önemli adım karar verebilmek için özellik seçimidir. Denetimli sınıflandırıcı sadece dokümanın nasıl temsil edildiğini gösterir.

Sınıflama teknikleri genel olarak nicel veya kategorik veriler üzerinde uygulanır, metinler üzerinde sınıflama gerçekleştirirken metni oluşturan kelimelerin metin içinde tekrar sayıları üzerinden frekans değerleri oluşturulmuşsa nicel bir modelleme yapılabilir. Metin boyutunun yüksek olduğu çok fazla farklı kelime kullanıldığı metinlerde kelime frekansları düşük kalmaktadır, bu neden ile etkin bir sınıflandırma yöntemi belirlenmesi gerekir (Aggarwal, 2012).

2.6.1. Destek Vektör Makineleri

Destek vektör makineleri sınıflandırıcıları farklı sınıflar arasındaki doğrusal veya doğrusal olmayan yapıları kullanıp, ideal sınırları belirleyerek sınıflanmamış veriyi sınıflandırmayı amaçlar (Aggarwal, 2012).

Destek vektör makineleri ile sınıflandırmada genellikle {-1,+1} şeklinde sınıf etiketleri ile gösterilen iki sınıfa ait örneklerin, eğitim verisi ile elde edilen bir karar fonksiyonu yardımıyla birbirinden ayrılması amaçlanır. Söz konusu karar fonksiyonu kullanılarak eğitim verisini en uygun şekilde ayırabilecek hiper-düzlem bulunur. İki sınıflı verileri birbirinden ayırabilen birçok hiper-düzlem çizilebilir. Ancak destek vektör makinesinin amacı kendisine en yakın noktalar arasındaki uzaklığı maksimuma çıkaran hiper-düzlemi bulabilmektir. Sınırı maksimuma çıkararak en uygun ayrımı yapan hiper-düzleme optimum hiper-düzlem ve sınır genişliğini sınırlandıran noktalar ise destek vektörleri olarak adlandırılır (Kavzoğlu ve Çölkesen, 2010).

Destek Vektör Makinesi (SVM) tekniği ile n adet sınıf için $n(n-1)/2$ sınıflandırıcı oluşturulur. Bu yöntem sayesiyle çoklu sınıf yapısı ikili sınıf yapısına çevrilmiş olur. Bu yöntemde n adet sınıf için n adet destek vektör makinesi kurulur ve $i.$ destek vektör makinesi, i sınıfındaki verileri kendi sınıf verisi olarak kullanırken, diğer sınıflardan gelen verilerin hepsini sanki 2.sınıfa ait veriymiş gibi kabul eder, kendi verilerine +1 etiketi verirken, diğer sınıflara ait olan tüm verilere -1 etiketini verir ve bu işlemi bu şekilde n adet destek vektör makinesi için yapar (Bayrak ve diğ., 2012).

Destek Vektör Makinesi sınıflandırıcı fonksiyonu: (Niu ve diğ., 2012)

$$dx = \sum_{i=1}^{m} a_i y_i K(x_i, x) + b \tag{2.1}$$

Yukarıdaki eşitlikte a_i ve b destek vektör makine algoritması tarafından elde edilebilir, $K(x_i, x)$ çekirdek fonksiyonudur. a_i değeri 0 değilse, örneklemler "destek vektör"e dönüşür.

2.6.2. Bayes Teoremi

A ve B gibi iki olay olsun, A'nın gerçekleşmesi durumunda B'nin gerçekleşme olasılığı **koşullu (conditional) olasılık** olarak tanımlanır ve B olayının, A olayının gerçekleşmesi halindeki koşullu olasılığı $P(A|B)$ şeklinde gösterilir (Vuran, 1983).

Bayes teoremi, bir A olayının ortaya çıkması halinde belirli bir B olayının gerçekleşme olasılığının hesaplanmasında kullanılır. Olasılık teorisi kapsamında A olayına koşullu bir B olayı (A olayı önceden biliniyor bu durumda B olayı) için olasılık değeri ile B olayına

koşullu olarak $\boldsymbol{A}$ olayı ($\boldsymbol{B}$ olayı önceden biliniyor bu durumda $\boldsymbol{A}$ olayı) için olasılık değerleri birbirinden farklıdır. Fakat bu iki ters koşulluluk arasında çok belirli bir ilişki vardır ve bu ilişkiye **Bayes Teoremi** denilmektedir. (Triola, 2008)

Apaydın (2010)'ın çalışmasında verdiği örneğe benzer şekilde örneğin, kurumların kredi notu değerlendirilmesi yapılırken gelir ve giderleri olmak üzere iki özelliği dikkate alalım. Kurumları düşük riskli veya yüksek riskli olmak üzere iki sınıfa ayırmak istendiğinde kurumlara göre farklı değerler olarak gelecek gelir (X_1) ve gider (X_2) parametreleri rastgele değişkenler olarak yorumlanır.

Bir olayın gerçekleşmesi halinde 1, gerçekleşmemesi halinde 0 değeri alan değişkene Bernoulli değişkeni adı verilir (Vuran, 1983). Bu doğrultuda gözlenen değerlerin ($\boldsymbol{X} = \{X_1, X_2\}$) sağladığı koşullara bağlı olarak kurumların kredi notu yüksek riskli ise *C* olarak ifade ettiğimiz Bernoulli rastgele değişkeni olarak *C*=1, düşük riskli ise *C*=0 olarak gösterilir, *C* burada sınıflarımızı temsil etmektedir. Farklı bir kuruma ait yeni gözlenen değerler olarak gelir ve gider özellikleri ($\boldsymbol{x} = \{x_1, x_2\}$) geldiğinde Bernoulli değişkeni olarak *C* 'nin alacağı değer $P(C|\boldsymbol{x})$ hesaplaması ile elde edilir, bu durumda **bayes kural**ının gösterimi aşağıdadır (Apaydın, 2010).

$$P(C|\boldsymbol{x}) = \frac{P(C)p(\boldsymbol{x}|C)}{p(\boldsymbol{x})} \tag{2.2}$$

Önsel (Önceki) Olasılık (*Prior Probability*): Herhangi bir ek enformasyon elde edilmeden önceki başlangıç aşamasındaki olasılık değeridir (Triola, 2008).

Örneğimizdeki *C* değerinin yeni gözlenecek değerler ($\boldsymbol{x}$) ne olursa olsun ($\boldsymbol{x}$ değerleri elde edilmeden önce) 1 olduğu durumda $P(C = 1)$ ifadesi **önsel olasılık** olarak adlandırılır, çünkü gözlemlenecek değerden önce *C* değeri bilgi olarak bir önceki adımda elde edilmiştir, bu doğrultu da;

$$P(C = 0) + P(C = 1) = 1 \tag{2.3}$$

Bayes kuralında $p(\boldsymbol{x}|C)$ ifadesi yeni gözlemlenecek $\boldsymbol{x}$ değerlerinin *C* ile ifade edilen sınıflarla ilişkilendirme olasılığıdır. Farklı ifade ile **sınıf ihtimali-olasılığı** (*class likelihood*) dır. $p(\boldsymbol{x})$ ifadesi ise **ispat** (*evidence*) olarak adlandırılır (Apaydın, 2010).

$$p(\boldsymbol{x}) = \sum_c p(\boldsymbol{x}, C) = p(\boldsymbol{x}|C = 1)P(C = 1) + p(\boldsymbol{x}|C = 0)P(C = 0) \quad (2.4)$$

Sonsal (Sonraki) Olasılık (*Posterior Probability*): Ek enformasyon kullanılarak revize edilmiş olasılık değeridir (Triola, 2008). Rastgele bir ***A*** olayının, farklı bir rastgele ***B*** olayına bağlı gerçekleşmesi ihtimalini ifade etmek için önsel olasılıklar yeterli olmaz, **koşullu** veya **sonsal** olasılık olarak ifade edilen $P(A|B)$ gösterimi kullanılır (Vuran, 1983).

Apaydın'ın da açıkladığı gibi örneğimizden devam edersek, gözlemlenen yeni gelen $\boldsymbol{x}$ değerleri sonrasında bu değerlerin etkisiyle hesaplanan sonsal olasılık bayes kuralında $P(C|\boldsymbol{x})$ ile ifade edilir.

$$sonsal = \frac{önsel \times ihtimal}{ispat} \quad (2.5)$$

sınıfların sonsal değerlerinin toplamı 1 değerini verir.

$$P(C = 0|\boldsymbol{x}) + P(C = 1|\boldsymbol{x}) = 1 \quad (2.6)$$

Genel yapıda *K* adet birbirinden farklı sınıf olduğu düşünülürse, $C_i, i = 1, \dots, K$ sınıflar şeklinde ifade edilebilir. Bu doğrultuda her sınıfın önsel (prior) olasılığı;

$$P(C_i) \geq 0 \; ve \; \sum_{i=1}^{K} P(C_i) = 1 \quad (2.7)$$

$\boldsymbol{x}$ değerleri C_i sınıfına ait girdi olarak bilindiğinde, $\boldsymbol{x}$ değerlerini görme olasılığı $P(\boldsymbol{x}|C_i)$ ile ifade edilir. C_i sınıfının sonsal (posterior) olasılığı ise aşağıdaki şekilde hesaplanır.

$$P(C_i|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|C_i)P(C_i)}{p(x)} = \frac{p(\boldsymbol{x}|C_i)P(C_i)}{\sum_{k=1}^{K} p(\boldsymbol{x}|C_k)P(C_k)} \quad (2.8)$$

Bayes sınıflandırıcı en az hata için en yüksek sonsal (posterior) olasılığı seçer. Eğer $P(C_i|\boldsymbol{x}) = \max P(C_k|\boldsymbol{x})$ ise C_i sınıfı seçilir (Apaydın, 2010).

Makine öğrenmesi kapsamında en iyi hipotezimizi ***H***, önceden gözlenen eğitim verimizi de ***D*** ile ifade edelim (Mitchell, 1997). Mevcut eğitim verisi, verilerin dağılımı doğrultusunda sınıflandırılır, her sınıf farklı bir hipotezimizi temsil edecektir. Her sınıfın eğitim veri setinde bulunma olasılıkları doğrultusunda farklı hipotezler içinden önceki-önsel olasılıkları dikkate alarak en olası hipotez bulunabilir. Bayes teoremi bu tür

olasılıkları hesaplamak için doğrudan bir yöntem sağlar. Her biri bir hipotezi temsil eden $\boldsymbol{D}$ eğitim veri seti içindeki sınıfları $\boldsymbol{c}$ ile ifade edelim, her bir $\boldsymbol{c}$ sınıfı hipotezimizin olma olasılığını da $\boldsymbol{P(c)}$ olarak ifade edelim. $\boldsymbol{P(c)}$ ifadesi $\boldsymbol{c}$ sınıfının önsel olasılığı olarak isimlendirilir ve her bir $\boldsymbol{P(c)}$ değeri ulaşmak istediğimiz hipotezimiz için aday hipotez olacaktır. Benzer şekilde eğitim verisinin gözlemlenme olasılığını da $\boldsymbol{P(d)}$ şeklinde gösterebiliriz. Bu doğrultuda her $\boldsymbol{d}$ belgesinin hangi sınıfta ($\boldsymbol{c}$) olma olasılığını bulmak için (2.8)'deki bayes kuralı aşağıdaki şekilde gösterilebilir.

$$P(c|d) = \frac{P(c)\ P(d|c)}{P(d)} \tag{2.9}$$

Aggarwal (2012)'a göre, bayes sınıflandırıcılar (üretken sınıflandırıcılar olarak da adlandırılır) ile farklı sınıflar içinde yer alan terimlerden yola çıkarak bir **olasılıksal model** inşa edilir. Olasılıksal sınıflandırıcılar temel dokümanların oluşumu için örtülü karışım modeli kullanmak amacı ile tasarlanırlar. Bu model kapsamında genellikle her sınıfın karışımının bir bileşeni olduğu varsayılır. Her karışım bileşeni esasen bir üretken modeldir ve bu üretken model de bileşen ya da sınıf için özel bir terim örnekleme olasılığı sağlar. Bu tür sınıflandırıcılara genellikle **üretken sınıflandırıcılar** denir.

Bayes sınıflandırıcı sınıf seçimi esnasında bayes karar verme kuralını uygulayarak hata olasılığını en aza indirir (Duda ve Hart, 1973). Metin çalışmalarında bayes sınıflayıcı geniş bir yapı içinde yer alan belirsiz terimleri uygun konumlara yerleştirmek için kullanılır ve her sınıflandırılmış terim belirsiz terimlerin sınıflandırılmasında potansiyel olarak yararlı veri olarak nitelendirilir (Manning ve Schütze, 1999).

2.6.3. Naïve Bayes Sınıflandırıcı

Naïve Bayes sınıflandırıcı farklı terimlerin dağılımları hakkındaki bağımsız varsayımlar ile her sınıf içindeki dokümanların dağılımını modelleyen basit ve yaygın olarak kullanılan üretken bir sınıflandırıcıdır (Aggarwal, 2012). Naïve Bayes metin sınıflandırma için basit bir yöntem sağlar, Naïve Bayes ile her metinin ait olduğu sınıfın belirlenmesi öncelikli olarak olasılık ve özellik dağılımı oluşturma ile sağlanır (Mullen ve Collier, 2004).

Naïve Bayes daha önceden sınıflara ayrılmış verileri kullanarak yeni elde edilen veya analiz için üzerinde çalışılan verinin hangi sınıf içinde yer alacağının olasılığını

hesaplayan algoritmadır. Bu algoritma ile yapılan hesaplamanın kapsamı; bir sonucun çıkma olasılığı o sonucu etkileyen tüm faktörlerin o sonucu sağlama olasılıklarının çarpımı şeklindedir. Farklı bir ifade ile analiz için üzerinde çalışılan verinin daha önceden oluşturulmuş olan sınıflardan hangi sınıfta olacağı, verinin tüm değişken niteliğindeki değerlerinin her bir sınıfta olma olasılığının çarpımları ile hesaplanır. Her bir sınıf için elde edilen bu değerlerden büyük olan hangi sınıfa ait ise üzerinde analiz yapılan veri o sınıfa yerleşir.

Aggarwal'a göre Naïve Bayes kapsamında iki sınıf modeli (*Çok Değişkenli Bernoulli Modeli* ve *Çok Terimli Model*) yaygın olarak kullanılır. Her iki modelde de esasen belgedeki kelimelerin dağılımına dayalı bir sınıflama olasılığı hesaplanır. Belgedeki kelimelerin gerçek konumu görmezden gelinir ve **kelime torbası** varsayımı ile çalışılır. Bu iki model arasındaki en önemli fark ise kelime frekanslarını dikkate alma (ya da almama) varsayımlarıdır.

> **Çok Değişkenli (Multivariate) Bernoulli Modeli:** Bir metin doküman içindeki kelimelerin varlığı ve yokluğu dokümanı temsil eden özellikler olarak kullanılır. Kelimelerin doküman içinde bulunma sıklığı belgeyi modellemek amacı ile kullanılmaz ve bir kelimenin varlığını veya yokluğunu simgeleyen iki değer ile gösterilir. Özellik olarak da ifade edilen kelime veya terimler ikili (binary) olarak modellenmiş olacaktır, bu şekilde oluşturulmuş her sınıf **Çok Değişkenli Bernoulli** modelidir.
>
> **Çok Terimli (Multinomial) Model:** Bu modelde, **kelime torbası** ile temsil edilen bir belgedeki terimlerim frekansları ile işlem yapılmaktadır. Her sınıf içindeki belgeler çok terimli kelime dağılımdan çekilmiş örnekler olarak modellenebilir. Bu durumda, bir sınıf içinde verilen bir belgenin koşullu olasılığı eşleşen sınıftaki her gözlenen kelimenin olasılığının basitçe bir sonucudur. (Aggarwal, 2012)

Naïve Bayes algoritması Bayes teoremine dayalıdır ve aşağıdaki temel eşitlikle ifade edilebilir.

Vektör $\boldsymbol{d}(\boldsymbol{w_1}, \boldsymbol{w_2} \ldots \boldsymbol{w_n})$ metin $\boldsymbol{A}$ ve $\boldsymbol{w_i}$ öğe özelliğidir. Eğer $\boldsymbol{A}$ grup $\boldsymbol{C_k}$'ya ait ise;

$$C_k = \underset{c \in \mathbb{C}}{\operatorname{argmax}} P(c|d) \ , P(c|d) = \frac{P(c)\ P(d|c)}{P(d)} \tag{2.10}$$

Aynı metin için, $\boldsymbol{P(d)}$ farklı olmaz. Tüm özelliklerin her biri birbirinden bağımsızdır, $\boldsymbol{P(d|c)}$ değerini $\boldsymbol{d}$'ye bölerek hesaplarız:

$$P(d|C_k) = P(w_1, w_2, .., w_n|C_k) = \prod_{i=1,k=1}^{n} P(w_i|C_k) \tag{2.11}$$

Bayes sınıflandırıcı kolayca hesaplanabilir ve farklı alanlarda yaygın olarak kullanılmaktadır, böylece yüksek bir hassasiyete sahiptir (Niu ve diğ., 2012).

Bu algoritmanın en önemli özelliği kullanılan veri seti içeresindeki az miktardaki gürültünün sonuç üzerindeki etkilerinin az olmasıdır (Kavzoğlu ve diğ., 2013).

2.6.4. Maksimum Entropi Yaklaşımı

Maksimum Entropi (MaxEnt) dil modellemesi ve doğal dil işleme süreçlerinde kullanılan çok güçlü bir araçtır (Berger ve diğ., 1996). Naïve Bayes doğrusal bir sınıflayıcı iken, Maksimum Entropi belirlenen koşullar doğrultusunda değer alan özelliklerin eğitim verisinden elde edilen ağırlık değerlerinin birleştirilerek oluşturulduğu, özelliklerin birbirine bağlı olduğu bir modeldir (Jurafsky ve Martin, 2015). Maksimum Entropi prensibinde önceki verilerin (eğitim verisi) olasılık dağılımlarındaki enformasyonu gösteren kısıtlara bağlı entropi değerlerinin maksimum edilmesi ile oluşan karar verme sürecidir (Malouf, 2010).

2.6.4.1. Entropi

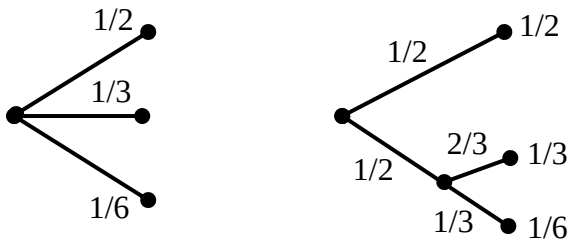
Entropi bilgi, bilgi eksikliği veya belirsizlik ile ilişkilidir, bazen çelişki ve birden fazla yön göstermektir (Brissaud, 2005). Boltzmann'a göre düzensizliğin özümsenmesi (Boltzmann, 1964), Shannon'a göre olumlu bilgi (Shannon, 1948), Brillouin'e göre ise bilgi eksikliği veya bilgisizlik (Brillouin, 1956) olarak ifade edilmektedir. Entropi belirsizliğin (tahmin edilemezliğin) en yüksek (maksimum) olması olarak anlaşılabilir (Fiske, 1990). Entropi kavramı, eldeki bilginin sayısallaştırılmasıdır (Dunham, 2003).

Entropi, bir sistemdeki rastgelelik ve düzensizliktir. Sınıflandırmada en ayırıcı özelliğe sahip değişkeni bulmak için veri kümesi içindeki belirsizliği ve rastgeleliği ölçmek gerekir, bu işlem içinde entropi kavramından yararlanır (Danacı ve diğ., 2010).

Her bilim alanında entropi farklı yorumlanır ve tanımlanır. Farklı bilim adamları entropiyi zaman içinde farklı şekilde tanımlamışlardır, karar verme sürecindeki entropi Shannon'un tanımındaki entropidir (Baray, 2003). Shannon'ın entropi tanımı ile olasılık arasındaki ilişkiyi Ross (1998) "Olasılığı düşük olan olayın sürpriz olarak değerlendirilmesi, bu olayın olasılığının ne kadar düşük olduğuna bağlıdır. Gerçekleşme olasılığı az olan bir olayın meydana gelmesi halinde oluşacak sürpriz de bu olasılıkla ters orantılı olarak büyük olacaktır" şeklinde belirtmiştir (Baray, 2003).

Meydana gelme olasılıkları $p_1, p_2, .., p_n$ olan bir dizi olay olsun. Bu olayların olasılıkları bilinirken, hangisinin gerçekleşeceği tahmin edilebilir. Seçilmiş olayın gerçekleşmesinde veya belirsiz olarak kalmasında kaç farklı seçeneğimiz olduğunun ölçüsünü bulabilmek için, aşağıdaki özellikler $H(p_1, p_2, .., p_n)$ olarak ifade edilen olasılıklar değerinın makul gerekliliklerdir (Shannon, 1948). Buradaki *H* belirsizliği verecek olan ölçüdür (Baray, 2003),

1. p_i olasılıklarında *H* değeri sürekli olmalıdır.
2. Eğer tüm p_i olasılıkları birbirine eşitse, $p_i = \frac{1}{n}$ olur. Bu durumda *H* monoton artan bir *n* fonksiyonudur.
3. Eğer bir seçenek iki başarılı seçenek olarak ayrılabiliyorsa, orijinal H değeri bireysel H değerlerinin ağırlıklı toplamı olacaktır. Şekildeki gösterim doğrultusunda üç olasılığın değerleri $p_1 = \frac{1}{2}$, $p_2 = \frac{1}{3}$, $p_3 = \frac{1}{6}$ tır.



Şekil 2.7: Üç olasılıklı bir tercih ayrışması (Shannon, 1948).

İlk tercih her biri $\frac{1}{2}$ olasılık değerine sahip iki olasılık arasında kalıyor, eğer ikinci tercih meydana gelirse diğer tercih $\frac{2}{3}$ ve $\frac{1}{3}$ olan iki olasılık ihtimali olarak ortaya çıkıyor. Nihai sonuçlar daha önce olduğu gibi aynı olasılıkları veriyor. Bu özel durumda,

$$H\left(\frac{1}{2},\frac{1}{3},\frac{1}{6}\right) = H\left(\frac{1}{2},\frac{1}{2}\right) + \frac{1}{2}H\left(\frac{2}{3},\frac{1}{3}\right). \quad (2.12)$$

Katsayı değeri $\frac{1}{2}$, çünkü ikinci tercih sadece yarı zaman diliminde meydana geliyor.

Bu durumda yukarıdaki varsayımlar doğrultusunda *H* aşağıdaki şekilde ifade edilebilir.

$$H = -K\sum_{i=1}^{n} p_i \log p_i \quad (2.13)$$

Eşitlikte *K* pozitif bir sabittir ve sadece bir ölçü birimi kadar değere sabittir (Shannon, 1948).

Bu doğrultu da *H(x)* aşağıdaki eşitlikle ifade edilir,

$$H(x) = -\sum_{x\in R} p_i \log(p_i) \quad (2.14)$$

bu eşitlik Shannon belirsizliği, Shannon Entropisi veya Bilgi Entropisi olarak ta isimlendirilir (Baray, 2003).

2.6.4.2. Maksimum Entropi Prensibi

Maksimum Entropi prensibi, bütün kısıtlamaları karşılayan bir istatistiksel model seçmek için kullanılır (Khundanpur ve Wu, 1999). Ortadan kalkması istenen belirsizliğin en üst seviyede bulunması bir çelişkidir, Jaynes (1995)'e göre ise "entropinin maksimum olması, güvenilir bilgi aracılığı ile karar vericinin doğru sonuca ulaşması için her olasılığa en uygun mesafede durması anlamına gelmektedir" (Baray, 2003).

Maksimum entropinin ana görevi mevcut durumun tüm ihtiyaçlarına cevap verebilecek en uygun sonucu bulmaktır. Maksimum entropi modelleri özellik-tabanlı modellerdir, lojistik regresyon ve sınıflar üzerinde dağılım olmak üzere iki sınıf senaryosu vardır (Go ve diğ., 2009).

Maksimum entropinin üssel modeli: (Pietra ve diğ., 1997)

$$P_{ME} = P_{(c|d)} = \frac{1}{Z(d)}\exp[\sum_i \lambda_i f_i(d,c)]] \quad (2.15)$$

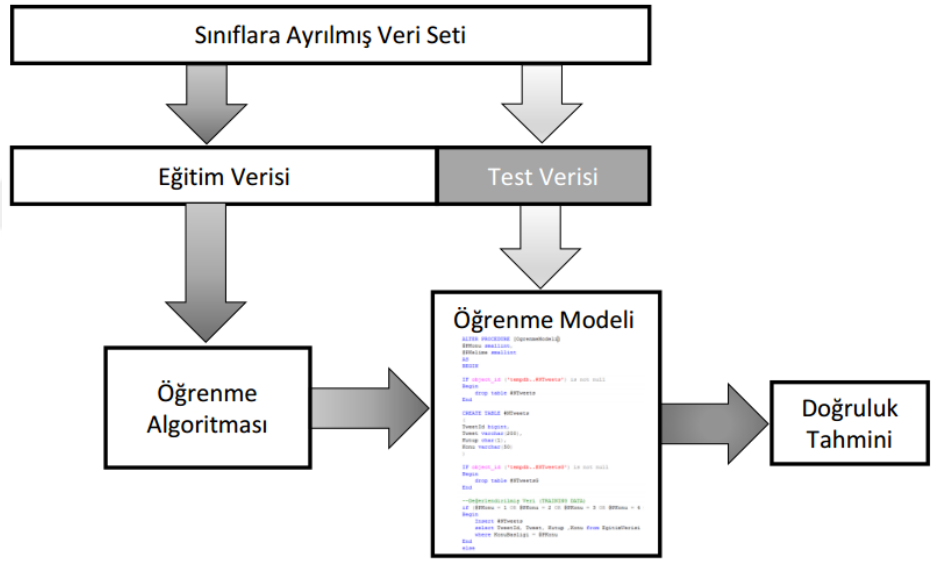
$P_{(c|d)}$ d dokümanının c sınıfında bulunma olasılığı, $Z(d)$ normalleştirme faktörü, $f_i(d,c)$ ise i teriminin c sınıfı ile ilgili koşullu durumunun karakteristik fonksiyonudur.

$$f_i(d,\mathrm{c}) = \begin{cases} 1,| \; eğer\ ki\ d\ dokümanı\ ile\ \boldsymbol{c}\ sınıfı\ arasındaki\ ilişki\ varsa \\ 0,| \; diğer\ durumlarda \end{cases} \quad (2.16)$$

Maksimum entropinin genel kullanımında boolean olarak eğer d dokümanının sınıfı c ise fonksiyonun değeri 1, diğer durumlarda ise 0 dır. λ_i ise karakteristik fonksiyonunun ağırlığıdır.

2.6.5. Modeli Değerlendirme

Makine öğrenmesi yöntemleri ile sınıflandırma işlemi gerçekleştiren modellerin başarısını ölçmek için performans değerlendirme ölçüleri ve yöntemleri bulunmaktadır.



Şekil 2.8: Model değerlendirme.

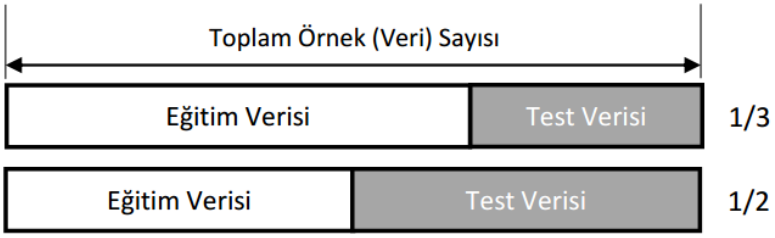
2.6.5.1. Model Değerlendirme Yöntemleri

Sınıflandırma çalışmaları genellikle eğitim ve test verisi temeline dayanır. Denetimli öğrenmede sınıflandırma amacı ile belirlenen algoritmaya uygun olarak hazırlanan veri iki bölüme ayrılır. İlk bölüm (eğitim verisi) modelin öğrenilmesi, kalan bölüm (test verisi) ise modelin geçerliliğinin test edilmesi için kullanılır. Eğitim verisi ile model öğrenimi gerçekleştirildikten sonra, test verisi ile modelin doğruluk derecesi belirlenir (Şekil 2.8). Model değerlendirme yöntemleri model çalışmasında belirlenen örneklemi farklı yapıda

bir örneklem şekline dönüştürmek, büyütmek veya parçalı birden fazla örneklem olarak tekrar oluşturmak amacı ile kullanılır.

Bootstrap: Veri setinde yer alan verilerin rastgele yer değişimi ile birden fazla yeni veri seti oluşturulmasıdır, bu yöntem ile çok fazla sayıda çok yeni örneklem niteliğinde veri seti oluşturulabilir (Temel ve diğ., 2012). ***N*** adet örnekten (veriden) oluşan bir veri kümesinden ***N*** adet yeni veri kümesi oluşturulması önerilir, fakat bu işlem örnek sayısı az olduğunda tercih edilir, büyük örnek sayısında problemdir (Efron ve Tibshirani, 1993).

Holdout (Ayırma): Orijinal sınıflandırılmış verinin (eğitim verisinin) iki ayrı sete ayrılıp, birinin eğitim diğerinin test seti olarak tanımlandığı yöntemdir, bu yöntemde ayırma Şekil 2.9'da yer aldığı gibi yarı yarıya olabileceği gibi 1/3 oranında da olabilir. Bu yöntemin belli kısıtları vardır, eğer ki ilk veri seti az kayıta sahipse, bölünme ile doğru orantılı olarak yeni oluşan eğitim seti de az veriden oluşacaktır, eğer ki ilk veri çok sayıda veriye sahipse 1/2 veya 1/3 oranına bağlı olarak yeni oluşan test verisi de oldukça büyük olacaktır (Tan ve diğ., 2003).



Şekil 2.9: Holdout yöntemi.

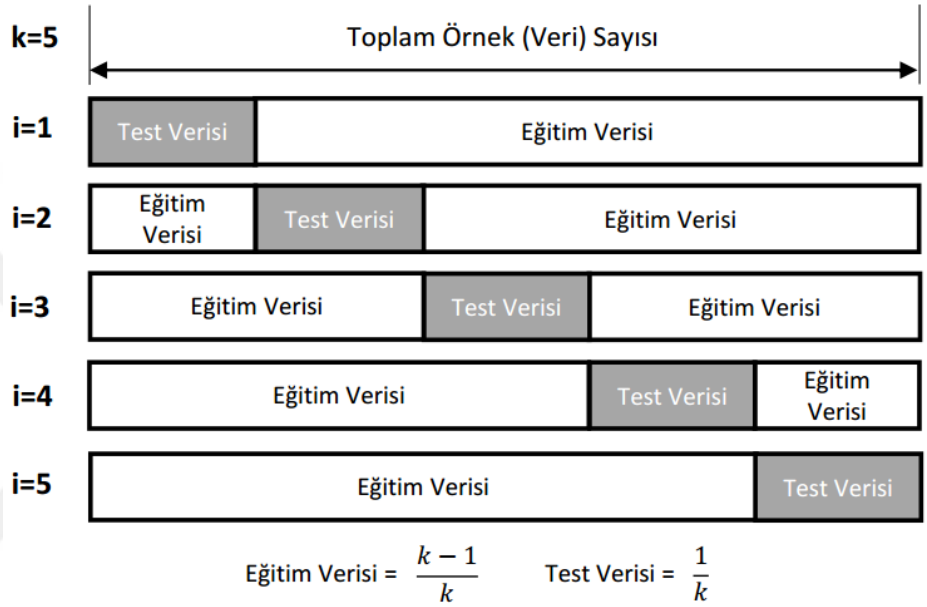
***k*-Katlı Çapraz Geçerlilik (*k*-Fold Cross Validation):** Bu yöntemde orijinal eğitim setinin rastgele bir şekilde k adet birbirinden ayrık eşit büyüklükte veri setleri şeklinde ayırımı yapılır, oluşan her veri seti kabaca aynı sınıf dağılımına sahiptir (Shakev-Shwartz ve Ben-David, 2014).

Sınıflandırıcı her defasında farklı bir test verisi ile k defa eğitilir. Şekil 2.10'da yer aldığı gibi sınıflandırılmış verinin $\frac{1}{k}$ oranı test verisi, $\frac{k-1}{k}$ oranı ise eğitim verisi olarak kullanılır (Kohavi, 1995). k değeri genelde 5 veya 10 seçilir, yöntemin bu şekilde kullanımı bir bölümün eğitim seti dışında bırakılarak kalan bölümlerin eğitim seti yapıldığı (**leave-one-out**) çapraz geçerlilik olarak isimlendirilir (Zaki ve Meira, 2014). Tahmin hata değeri k adet hata değerinin ortalamasının alınması ile hesaplanır. Aynı şekilde modelin

tamamının doğruluk değeri tahmini tüm ayrık veri setlerinin doğruluk ölçü değerinin ortalamasına eşittir,

$$CVA = \frac{1}{k}\sum_{i=1}^{k} A_i \tag{2.17}$$

Eşitlikte CVA , **çapraz geçerlilik doğruluğu** (*cross validation accuracy*), k, kullanılan kat sayısı, A, doğruluk ölçüsüdür (Olson ve Delen, 2008).



Şekil 2.10: 5-katlı çapraz geçerlilik (5-fold cross validation) yöntemi

2.6.5.2. Modelin Geçerliliğini Sınama Yöntemleri

Makine öğrenmesi algoritmaları ile oluşturulan modelin geçerli sayılabilmesi için başarı seviyesinin ve performansının ölçülmesi gerekir. Algoritmaların performans değerlendirilmesinde genelde hata matrisinin (error matrix) (Stehman, 1997) geliştirilmesi ile oluşan **karışıklık matris** (*confusion matrix*) yapısı kullanılır (Tablo 2.1). Bu yapı içinde kullanılan ölçüler; hata oranı (error-rate), doğruluk-kesinlik (accuracy), kesinlik-hassaslık (precision), duyarlılık (sensitivity–recall) ve F-ölçüsüdür (Chawlave diğ., 2002).

Karışıklık matrisinin farklı bir gösterimi olan **olabilirlik tablosu** (*contingency table*) değişkenlerin frekans dağılımın gösteren matris bir yapıdır (Jurafsky ve Martin, 2015).

Tablo 2.1: İki kategorili karışıklık matrisi (confusion matrix).

		Tahmin	
		Negatif (Olumsuz)	Pozitif (Olumlu)
Gerçek	Negatif (Olumsuz)	Doğru (True) Negatif **TN**	Yanlış (False) Pozitif **FP**
	Pozitif (Olumlu)	Yanlış (False) Negatif **FN**	Doğru (True) Pozitif **TP**

Tablo 2.1 ve 2.2'de yer alan parametreler doğrultusunda modelin geçerliliğini sınama ölçüleri diğer sayfadaki şekilde formüle edilmektedir.

Tablo 2.2: İki kategorili olabilirlik tablosu (contingency table).

		Gerçek Sonuç			
		Pozitif (Olumlu)	**Negatif** (Olumsuz)	**Toplam**	
Model Sonucu	**Pozitif** (Olumlu)	Doğru (True) Pozitif **TP**	Yanlış (False) Pozitif **FP**	**MP**=TP+FP	**kesinlik(precision)** $\frac{TP}{TP+FP}$
	Negatif (Olumsuz)	Yanlış (False) Negatif **FN**	Doğru (True) Negatif **TN**	**MN** =TN+FN	$\frac{TN}{TN+FN}$
Toplam		**GP**=TP+FN	**GN**=FP+TN	TP+FP+TN+FN	
		duyarlılık $\frac{TP}{TP+FN}$			**doğruluk(accuracy)** $\frac{TP+TN}{TP+FP+TN+FN}$

Doğruluk (*accuracy*)**:** Model tarafından tüm doğru tahmin edilenlerin, tüm sonuçlara olan oranıdır.

$$doğruluk(accuracy) = \frac{TP+TN}{TP+FP+TN+FN} \tag{2.18}$$

Hata oranı (*error rate*)**:** Doğruluk değeri üstteki eşitlik sonucunda 0-1 aralığında bir değer olarak elde edilecektir, bu değer yüzdesel olarak bir doğruluk değerini temsil edecektir, bu doğrultuda modelin hata oranı aşağıdaki şekilde formüle edilir.

$$hata\ oranı\ (error\ rate) = 1 - doğruluk(accuracy) \tag{2.19}$$

Kesinlik (*precision*)**:** Model tarafından doğru tahmin edilenlerin kendi sınıf sonuçlarına olan oranıdır.

$$kesinlik(precision)_{pozitif-olumlu} = \frac{TP}{TP+FP} \quad (2.20)$$

$$kesinlik(precision)_{negatif-olumsuz} = \frac{TN}{TN+FN} \quad (2.21)$$

Duyarlılık (*sensitivity-recall*) – **doğru pozitif oranı** (*true positive rate*)**:** Model tarafından pozitif (olumlu) sınıf için doğru tahmin edilenlerin negatif (olumsuz) sınıfı için yanlış tahmin edilenlerle olan oranıdır.

$$duyarlılık(sensitivity - recall) = duyarlılık_{pozitif} = \frac{TP}{TP+FN} \quad (2.22)$$

Belirleyicilik (*specificity*) – **doğru negatif oranı** (*true negative rate*)**:** Model tarafından negatif (olumsuz) sınıf için doğru tahmin edilenlerin pozitif (olumlu) sınıfı için yanlış tahmin edilenlerle olan oranıdır.

$$belirleyicilik(specificity) = duyarlılık_{negatif} = \frac{TN}{TN+FP} \quad (2.23)$$

Yanlış pozitif/negatif oranları (*false positive/negative rate*)**:** Bu oranlar duyarlılık ve belirleyicilik oranlarının tam zıttı olan oranlardır. Gerçek veri olarak pozitif (olumlu) olan ama model tarafında negatif (olumsuz) olarak sınıflandırılan veriler ve gerçekte negatif (olumsuz) olup model tarafında pozitif (olumlu) olarak sınıflandırılan veriler için bu oranlar kullanılmaktadır.

$$yanlış\ pozitif\ oranı\ (false\ positive\ rate) = \frac{FP}{FP+TN} = 1 - belirleyicilik \quad (2.24)$$

$$yanlış\ negatif\ oranı\ (false\ negative\ rate) = \frac{FN}{FN+TP} = 1 - duyarlılık \quad (2.25)$$

F-Ölçüsü (*F-Measure*)**:** kesinlik ve duyarlılık ölçülerinin harmonik ortalamasıdır. Harmonik ortalama; n adet değerin birbirleri ile çarpımını n ile çarpımı sonucunda elde edilen değerin n adet sayının toplamına bölümüdür.

$$F\ Ölçüsü\ (F\ Measure) = \frac{2 \times kesinlik \times duyarlılık}{kesinlik+duyarlılık} \quad (2.26)$$

Yukarıdaki eşitlikler iki sınıflı (kategorili) modellerde kullanılmaktadır. Eğer ikiden fazla sınıf ile oluşturulan bir modelin geçerliliğinin sınanması yapılacak ise Tablo 2.3'deki gibi bir karışıklık matris ve bu matrise göre hesaplanan geçerlilik sınama ölçüleri kullanılır.

Tablo 2.3'de yer alan parametreler doğrultusunda üç sınıflı (kategorili) modelin geçerliliğini sınama ölçüleri aşağıdaki şekilde formüle edilmektedir (Jurafsky ve Martin, 2015).

$$doğruluk(accuracy) = \frac{TP+TN+T0}{TP+FP+TN+FN+T0+F0} \quad (2.27)$$

Tablo 2.3: Üç kategorili karışıklık matrisi (confusion matrix).

		Gerçek Sonuç			
		Pozitif (Olumlu)	**Negatif** (Olumsuz)	**Nötr**	**Toplam**
Model Sonucu	**Pozitif** (Olumlu)	Doğru (True) Pozitif **TP**	Yanlış (False) Pozitif **FP**	Yanlış (False) Pozitif **FP**	**Model Pozitiflerinin Toplamı**
	Negatif (Olumsuz)	Yanlış (False) Negatif **FN**	Doğru (True) Negatif **TN**	Yanlış (False) Negatif **FN**	**Model Negatiflerinin Toplamı**
	Nötr	Yanlış (False) Nötr **F0**	Yanlış (False) Nötr **F0**	Doğru (False) Nötr **T0**	**Model Nötrlerinin Toplamı**
Toplam		**Gerçek Pozitifleri Toplamı**	**Gerçek Negatiflerin Toplamı**	**Gerçek Nötrlerin Toplamı**	

Her bir sınıf için kesinlik ve duyarlılık eşitlikleri aşağıdadır.

$$kesinlik(precision)_{pozitif-olumlu} = \frac{TP}{TP+FP} \quad (2.28)$$

$$kesinlik(precision)_{negatif-olumsuz} = \frac{TN}{TN+FN} \quad (2.29)$$

$$kesinlik(precision)_{nötr} = \frac{T0}{T0+F0} \quad (2.30)$$

$$duyarlılık(sensitivity - recall) = duyarlılık_{pozitif} = \frac{TP}{TP+FN+F0} \quad (2.31)$$

$$duyarlılık(sensitivity - recall) = duyarlılık_{negatif} = \frac{TN}{TN+FP+F0} \quad (2.32)$$

$$duyarlılık(sensitivity - recall) = duyarlılık_{nötr} = \frac{T0}{T0+FP+FN} \quad (2.33)$$

2.7. TÜRKÇE VE DUYGU ANALİZİ

Son yıllarda Türkçe üzerine de duygu analizi çalışmaları yapılmıştır. Özellikle Yıldız Teknik Üniversitesi Bilgisayar Mühendisliği bölümü kapsamında birçok çalışmanın yapıldığı görülmektedir.

2006 yılında yapılan bir yüksek lisans tez çalışması sonrasında 2010 yılında Yıldız Üniversitesi'nde Türkçe bir dokümanın türü, yazarı ve doküman yazarının cinsiyetini Türkçe üzerinde N-gram modeli kullanılarak belirlenmeye çalışılmıştır. Çalışma kapsamında sınıflandırma yapmak için Weka sınıflandırma aracı içerisinde yer alan Naïve Bayes, Destek Vektör Makineleri, K-En Yakın Komşuluk ve Rastgele Orman algoritmaları kullanılmıştır (Doğan ve Diri, 2010).

Kaya vd. (2012) açık kaynak kodlu doğal dil işleme yazılımı Zemberek aracılığı ile Türk politikası için Twitter ortamında yazılmış mesajlar üzerinde Naïve Bayes, Karar Destek Makinesi ve Maksimum Entropi algoritmaları ve N-gram yapısı kullanılarak duygu analizi çalışması gerçekleştirmişlerdir.

Yine Yıldız Üniversitesi'nde 2013 yılında yapılan bir duygu analizi çalışmasında geleneksel terim ağırlıklandırma yöntemi ile birlikte makine öğrenmesi ve sınıflandırma algoritmaları olarak Naïve Bayes, Karar Destek Makinesi, Karar Ağacı (J48) ve Rastgele Orman kullanılmıştır. Çalışma kapsamında ilgili algoritmalar Weka yazılımı üzerinde çalıştırılmıştır (Çetin ve Amasyalı, 2013).

Çoğunlukla yukarıdaki çalışmalarda da görüldüğü gibi duygu analizi daha önceden yazılmış ve araştırmacıların kullanımına sunulmuş hazır yazılımlar aracılığı ile yapılmakta, bu hazır yazılımların kısıt ve şartları ile çalışmalar gerçekleştirilmektedir.

2.7.1. Kelime Tabanlı Yaklaşım

"Bölüm 2.5 Duygu Analizi" başlığı altında belirtilen Stanford Üniversitesi çalışmasındaki duygu analizi, cümle genelinin olumlu veya olumsuz anlamına bakılarak gerçekleştirilmektedir. Çalışma kapsamında mesajların olumlu veya olumsuz olduğuna karar verilirken mesaj içindeki olumsuzluk anlamı veren kelimelerin (olumluluk için laugh, cheap, luck, vb. kelimelerin, olumsuzluk içinse; hate, no, poor, lost, RIP, vb. kelimelerin) arandığı görülmektedir. (Stanford University, 2013)

Türkçede de birçok kelime anlamsal olarak incelendiğinde cümleye olumlu veya olumsuz duygu kattığı görülür. Örneğin “iyi”, “sakin”, “tatlı”, “coşku”, gülümse” gibi kelimeler anlamsal olarak “Olumlu” dur, “tuzak”, “kötü”, “talihsiz”, “ucube” gibi birçok kelimede anlamsal olarak “Olumsuz” dur. Bu kelimeler bir cümle içinde yer alırsa o cümleye belli bir oranda olumluluk veya olumsuzluk duygusu katmaktadır. Dikkat edilirse bu kelimeler **sıfat** niteliğindeki kelimelerdir.

2.7.2. Fiil Tabanlı Yaklaşım

Türk Dil Kurumuna göre fiil; “olumlu veya olumsuz olarak çekimli durumda zaman kavramı taşıyan veya zaman kavramı ile birlikte kişi kavramı veren kelime, eylem” dir (Türk Dil Kurumu, 2015). Türkçede bir cümledeki duygu, şahıs, zaman ve belli bir oranda da anlam cümlenin yapısı içindeki fiil veya fiilin çekilmiş (eklemeler almış) halinden tespit edilir. Görüşler üzerinden bir analiz yapılacaksa bu analiz o görüşün içinde yer alan fiil niteliğindeki kelimelerin olumlu, olumsuz, soru, zaman ve şahıs (kişi) ekleri doğrultusunda gerçekleşecektir.

“Bölüm 2.4 Doğal Dil İşleme” başlığı altında belirtildiği gibi Türkçe sondan eklemeli bir dildir. Türkçede fiillerin yalın hallerine olumsuzluk eki eklenerek cümle olumsuzlaştırılır veya fiilin yalın haline zaman (dili veya mişli geçmiş zaman, şimdiki zaman, geniş zaman, gelecek zaman ve diğer zaman) ekleri eklenerek cümlenin olumlu olması ve zamanı belirlenir.

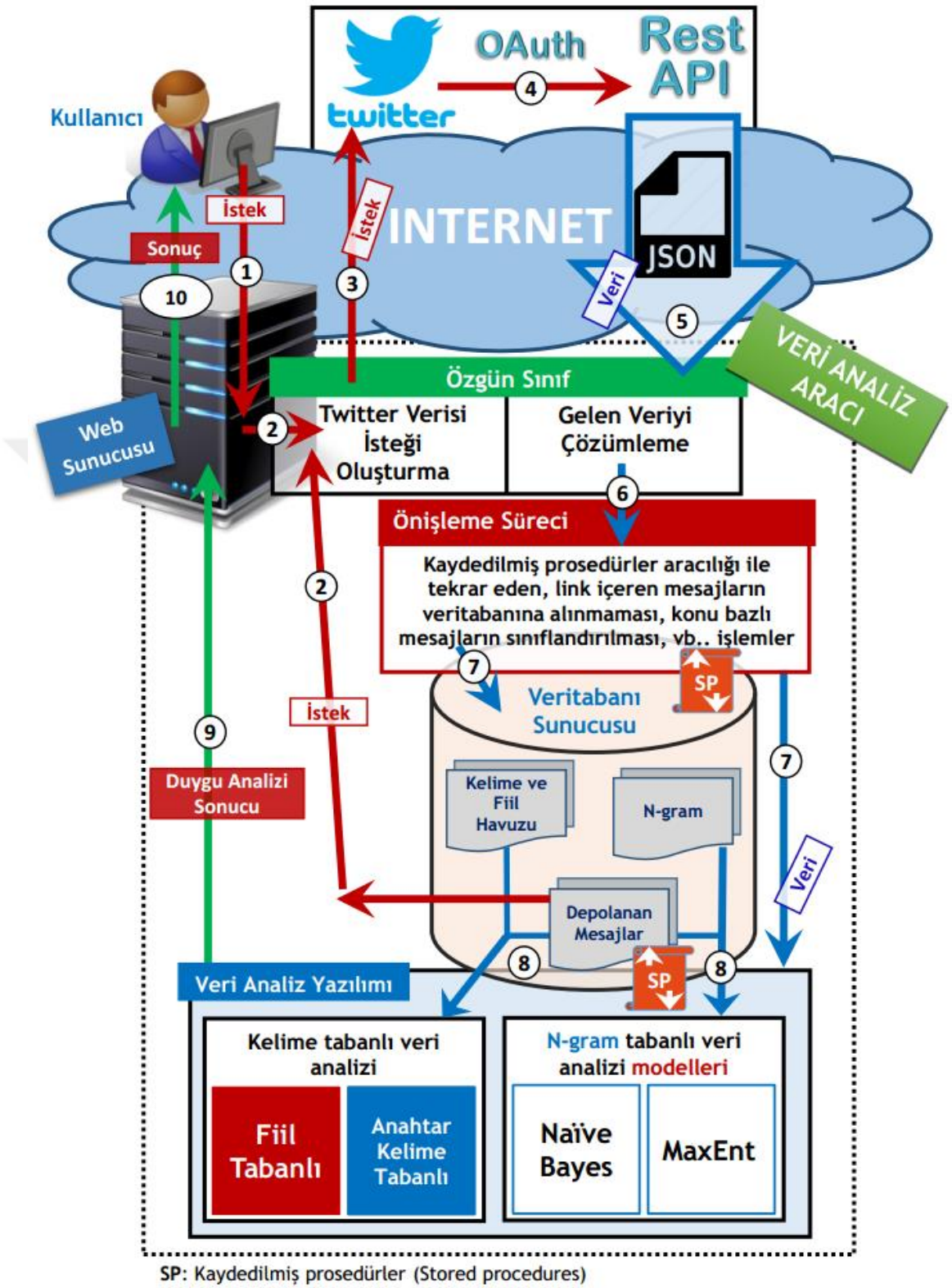
Stanford Üniversitesi’nin çalışmasındaki (Stanford University, 2013) İngilizce tweetler üzerindeki sonuçlar incelendiğinde, mesajların olumlu veya olumsuz olduğuna karar verilirken mesaj içindeki olumsuzluk ekinin (İngilizce -not ekinin) de arandığı görülmektedir. –not ekinin ve Bölüm 2.7.1 de belirtilen olumlu veya olumsuz kelimelerin kullanım sıklığına bakılarak cümlenin olumlu veya olumsuz olduğuna karar verilmektedir. Fakat Türkçede İngilizcede olduğu gibi cümle yapısı içinde olumsuzluk anlamı katan ayrık veya birleşik olduğunda belirgin bir şekilde olan sabit bir olumsuzluk eki yoktur. Türkçede yer alan olumsuzluk ekleri fiilin ana kök kısmından hemen sonra gelen eklerdir (-ma, -me, -mı, -mi, -mu, -mü, -maz, -mez ve bu eklere gelen kaynaşmaya yönelik harf veya heceler). Eğer ki bu ekler fiilin hemen kökünden sonra gelmiyor ve kök sonrasında zaman ve şahıs ekleri geliyorsa cümlenin olumlu bir cümle olduğu düşünülür.

3. MALZEME VE YÖNTEM

Son yıllarda hızla artan ve büyüyen veri miktarları doğrultusunda bu veriler üzerinde hızlı analizler yapabilmek için, akıllı veri analizi araçlarına ihtiyaç duyulmaktadır. Özellikle mikroblog hizmeti Twitter'ın kullanım sıklığı doğrultusunda herhangi bir konu da milyonlarca hatta milyarca mesaj kısa sürede kullanıcılar tarafından üretilebilmektedir, bu veriler üzerinde analiz yapabilmek için yukarıda bahsedildiği gibi hızlı sonuç üreten veri analiz araçlarına ihtiyaç vardır. Bu tez çalışması kapsamında veritabanlarında bilgi keşfi süreçleri doğrultusunda Türkçe Twitter mesajları üzerinde hızlı sonuç üreten bir duygu analizi aracı geliştirilmiş (Şekil 3.1) ve bu aracın ürettiği sonuçlar ile metin veri üzerinde daha önceden literatürde duygu analizi için kullanılmış makine öğrenmesi tekniklerinden Naïve Bayes ve Maksimum Entropi (MaxEnt) yöntemleri de analiz aracına entegre edilerek geliştirilen yöntemin karşılaştırılması yapılmaktadır.

Tezin malzeme ve yöntem bölümünde, tez çalışmasında kapsamında geliştirilen "Anahtar Kelime ve Fiil Tabanlı Duygu Analizi" model ve modelin uygulaması, veritabanlarında bilgi keşfi süreçleri takip edilerek açıklanmıştır. Ayrıca geliştirilen modelimizi karşılaştıracağımız makine öğrenmesi metotları Naïve Bayes ve Maksimum Entropi makine öğrenmesi metotları hakkında Bölüm 2.6.3 ve 2.6.4'de verilen bilgilere ek olarak uygulamaya yönelik daha detaylı bilgilere yer verilmiştir. Bu makine öğrenmesi modellerinin geliştirilen modelle entegre çalışması için uygulama içine geliştirme yapılarak (kodlama) aktarılması süreci de bu bölümde açıklanmıştır.

Bu bölüm kapsamında Şekil 3.1'de yapısı ve çalışma aşamaları yer alan "Duygu Analizi Aracı" nın katmanları detaylı olarak açıklanmakta, katmanların geliştirilmesi esnasında kullanılan bilişim altyapısı, programlama teknikleri, veri yapıları, veri değişim modelleri, veritabanı teknolojileri ve ileri veri sorgulama teknikleri anlatılmaktadır.



Şekil 3.1: Geliştirilen duygu analizi aracı.

3.1. PROBLEMİN TANIMLANMASI

30 Haziran 2016 verilerine göre 7 milyar 340 milyon nüfusa sahip dünyamızda yaklaşık 3,61 milyar (%49,2) insan internet kullanıcısıdır (Haziran 2012 verilerine göre ise dünyada yaşayan yaklaşık 7 milyar insanın yaklaşık 2,5 milyarı (%34,3) internet kullanıcısıydı) (Internet World Stats, 2016). Türkiye Bilgi Teknolojileri ve İletişim Kurulu'nun 2016 yılı ikinci çeyrek sonu pazar verileri raporuna göre Türkiye'de aktif internet kullanıcı sayısı 55,3 milyona ulaşmıştır (2013 yılı üçüncü çeyreğinde bu değer 33,7 milyon kişiydi) (Bilgi Teknolojileri ve İletişim Kurumu, 2016). Ayrıca aynı raporda mobil operatörlerdeki 3G abone sayısının 28,6 milyona, 4.5G abone sayısının 38,6 milyona ulaştığı, 3G ve 4.5G hizmetiyle birlikte mobil bilgisayardan ve cepten internet hizmeti alan abone sayısının 45.322.149'a yükseldiği, 2016 ikinci çeyrekte toplam mobil internet kullanım miktarının ise 255.376 TByte olarak gerçekleştiği belirtilmektedir (2013 üçüncü çeyrek raporunda mobil operatörlerdeki 3G abone sayısı 47,5 milyon, 3G hizmetiyle birlikte mobil bilgisayardan ve cepten internet hizmeti alan abone sayısının 25.492.162, toplam mobil internet kullanım miktarı ise 38.944 TByte olarak gerçekleşmişti). TÜİK verilerine göre 2015 yılı sonu itibari ile Türkiye nüfusu 78.741.053'dür (2013 sonu itibari ile TÜİK verilerine göre Türkiye nüfusu 75.627.384'dü) (Türkiye İstatistik Kurumu, 2016), bu doğrultuda Türkiye nüfusun %70,23'ü aktif internet ağ kullanıcısıdır, nüfusun %85,34'ü de mobil operatörler üzerinden internet kullanımı gerçekleştirme imkânına sahiptir.

Yukarıdaki rakamlarında gösterdiği gibi bir teknoloji çağının içinde yaşamaktayız. Bu teknoloji çağı doğrultusunda fikir toplama yöntemleri fiziksel ortamdan elektronik ortama aktarılmaya başlamıştır. Toplum içinde kullanılan birçok araç-gereç (otomobil, televizyon, cep telefonu, vb.), yazılım (oyun, programlama dili vb.), donanım, siyasi fikir, para yönetimi, finans, tatil, turizm ve benzeri birçok konu hakkında web ortamında forum siteleri ile karşılaşılmaktadır. Teknolojik altyapıya sahip kurumlar içinde de çalışanlarının fikirlerini toplamaya yönelik elektronik forum ortamları oluşturulmaktadırlar. Bu ortamlar sayesiyle fikirler analiz amacı ile toplanmaktadır.

T.C. Gençlik ve Spor Bakanlığı (2013)'nın yayınladığı "Gençlik ve Sosyal Medya Araştırma Raporu"nda, sosyal ağ sitelerini, kullanıcıların kendilerini tanıtacakları bir profil sayfası oluşturarak, bu sayfayı arkadaşları veya herkesle paylaştığı, kullanıcı

iletişimlerinin e-posta veya anlık mesajlaşma teknolojileri sağlandığı ortamlar olarak tanımlanmaktadır. comScore Inc. firmasının Mart 2013 verilerine göre dünya genelinde sosyal ağ lideri 818,2 milyon aktif kullanıcı sayısı ile insanların başka insanlarla iletişim kurmasını ve bilgi alışverişi yapmasını amaçlayan bir sosyal paylaşım web sitesi olan Facebook'tu. 189,8 milyon kullanıcı ile bir mikroblog sitesi olan Twitter ikinci, 164,2 milyon kullanıcı ile profesyonellerin meslek profillerini paylaştığı Linkedln üçüncü sırada yer almaktaydı.

Facebook sosyal ağ uygulamasının kurumsal web sitesindeki Haziran 2016 sonu verilerine göre aylık aktif kullanıcı sayısı 1,71 milyar kişidir (Facebook, 2016).

Statista Inc. firmasının Eylül 2016'da yayınladığı verilerine göre ise Facebook 1712, Twitter 313 ve Instagram 500 milyon aktif kullanıcıya sahip oldukları görülmektedir. Çevrimiçi mesajlaşma uygulaması olarak kullanılmaya başlayan Whatsapp uygulamasını kullanan aktif kullanıcı sayısının ise 1 milyar olduğu, son iki yıl içinde kullanım oranının çok hızlı arttığı göze çarpmaktadır. (Statista Inc., 2016)

Sosyal ağ sitelerinden mikroblog hizmetleri, kullanıcılara kısa cümleler, anlık fotoğraflar veya video linkleri gibi küçük içerik parçalarını paylaşmak için ortam sağlar. Ipsos firmasının 2013 yılında yaptığı araştırmaya göre Endonezya, Hindistan, Türkiye, Güney Afrika ve Polonya'daki internet kullanıcıları sosyal ağlar üzerinde kamu politikaları, toplumsal ve politik konularda aktif veya pasif olarak yer aldığı raporlanmıştır (T.C. Gençlik ve Spor Bakanlığı, 2013).

Mikroblog hizmetlerinden en bilinenlerden biri olan Twitter uygulaması için sosyal medya ölçümleme firması Somera'nın Digital Age dergisi için 2015 yılı başında yaptığı 2014 yılı sonu araştırma verilerine göre, Türkiye'de 11,5 milyon Twitter kullanıcısı bulunuyor, aktif kullanıcı sayısı 5.6 milyon olarak belirtilmiştir. Türkiye'de saniye de atılan tweet sayısı 152, günlük atılan tweet sayısı ise 13,1 milyona ulaşmıştır (Digital Age Dergisi, 2015). Londra merkezli We Are Social firmasının Ocak 2016'da yayınladığı küresel dijital ortam verilerine göre Türkiye'deki Twitter kullanıcı sayısı 7,14 milyon olarak belirtilmektedir (We Are Social, 2016).

Dünyada Twitter kullanımı ile ilgili 2013 yılının son aylarında eMarketer firması tarafından yapılan bir araştırmaya göre, ülkelerin internet kullanıcı sayısına göre Twitter

kullanım oranlarında Türkiye ilk sırada yer almaktadır (T.C. Gençlik ve Spor Bakanlığı, 2013). 2016 yılı başı verilerine göre Türkiye'deki Twitter kullanıcısı sayısını internet kullanan nüfusa oranlandığımızda %14,7 değeri elde edilmektedir, bu doğrultu da Türkiye'deki her yedi internet kullanıcısından birinin aktif Twitter hesabı vardır.

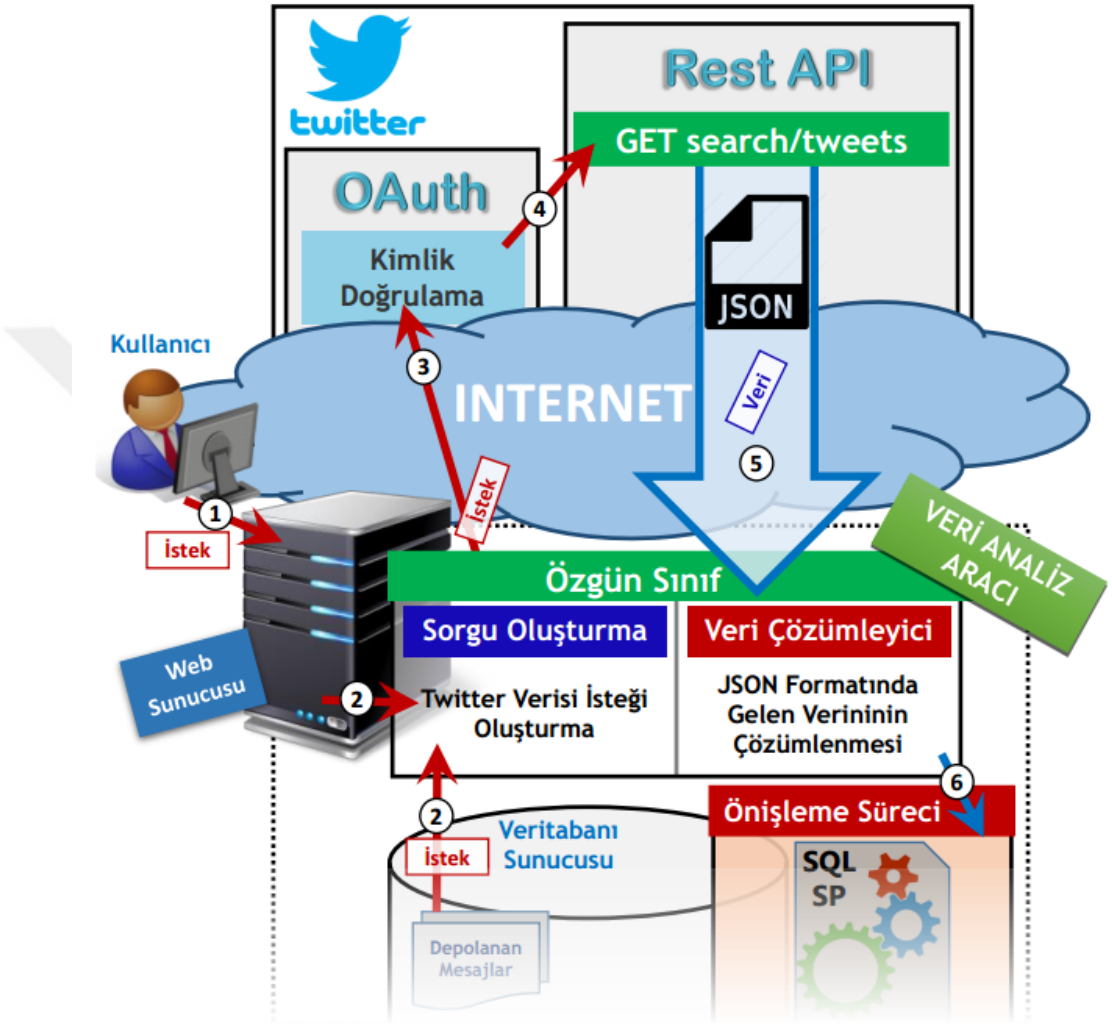
Etiya firmasının Somemto isimli birimi tarafından Haziran 2013 yapılan istatistik çalışmada, 2013 yılının mayıs ayında günlük ortalama Türkçe tabanlı 7 milyon adet Tweet gönderilmiştir. İstanbul'daki toplumsal olayların başladığı 31 Mayıs 2013 Cuma günü ise bu rakamın 15 milyonlara yükseldiği görülmektedir, Türkiye genelinde gerginliğin arttığı 1 Haziran 2013 Cumartesi günü tweet mesajı sayısı 18.8 milyona ulaşmıştır, bu tarih tüm zamanların Türkçe tabanlı tweet mesajı gönderilen en yüksek günü olmuştur. İlgili hafta Türkçe tabanlı 91.3 milyon tweet mesajı atılmıştır. (Somemto, 2013)

Sosyal ağların çok popüler olduğu günümüzde fiziksel ortamlarda veya sabit web sayfaları üzerinden fikir toplamak yerine bu sosyal ağları kullanmak verimlilik açısından çok daha başarılı sonuçlar elde edilmesini sağlayacaktır. Sosyal ağlar içinde kişiler kendilerini daha özgür hissetmekte, görüşlerini daha net ve açık bir şekilde ifade edebilmektedirler. Sosyal ağlara insanlar günün 24 saati erişim sağlayabilmektedirler, bu durum zaman kısıdını ortadan kaldırmaktadır. Çoğunlukla katılımcılar bu sosyal ağların aktif kullanıcısı oldukları için belirtilen yeni bir görüşe anında olumlu veya olumsuz cevaplar dönebiliyorlar. Bir olay, sorun veya nesne hakkında toplanmak istenen veri niteliğindeki görüşlerin bu ortamlar üzerinden yapılması alınmak istenen sonucun normal tartışma ortamlarına göre çok daha kısa sürede elde edilmesini sağlamaktadır.

İnsanlar kişisel, siyasi ve dini görüşlerini mikroblog ortamlarında paylaştıkça mikroblog sitelerinin insanların düşünce ve duygularını barındırma nitelikleri daha da değerlenmekte, mikroblog siteleri içindeki bu veriler verimli pazarlama ya da sosyal çalışmalar için de kullanılabilir konuma gelmektedir (Pak ve Paroubek, 2010). Twitter mikroblog uygulamasının aktif kullanıcı sayısı ve paylaşılan mesaj sayıları dikkate alındığında herhangi bir konu başlığı için çok kısa sürelerde çok fazla veri toplanabileceği görülmektedir, bu verilerin doğru bir şekilde analiz edilmesi yararlı olan bilgiye ulaşılmasını sağlayacaktır.

3.2. VERİ SEÇİMİ

Usame Fayyad'ın tanımladığı **veritabanlarında bilgi keşfi** süreçleri doğrultusunda tez kapsamında **veri seçimi** aşaması için gerekli ortam hazırlama çalışmaları yapılmıştır.



Şekil 3.2: Duygu analizi aracının veri seçimi katmanları (Twitter verisi seçim arayüzü).

Twitter sosyal ağ sitesi geliştiricilerle API'ler (Application Programming Interface - Uygulama Programlama Arayüzleri) aracılığı ile belirlenen belli bir taslak yapı içinde Twitter ortamında atılan tweet mesajlarını, dünya geneli, ülkeye göre veya bölgesel anlık güncel konu başlıklarını (Trend Topics) paylaşmaktadır. Bu doğrultuda tez çalışması kapsamında Twitter'ın API yapısını kullanarak Twitter ortamından belirlenen bir veya birden fazla konuyu içeren tweet mesajları üzerinde duygu analizi gerçekleştirebilmek için mesajları veri olarak bir veritabanı içine alan arayüz yazılımı geliştirilmiştir. Şekil 3.2'de duygu analizi aracının veri seçimi ile ilgili katmanları yer almaktadır.

3.2.1. Twitter REST API Arayüzü

Twitter üzerindeki bazı işlevlerin Twitter web sitesi ve uygulaması haricinden de kullanılabilmesi için Twitter'ın sunduğu REST API arayüz yapısı kullanılmaktadır (Twitter, 2013). REST API arayüzü programsal olarak Twitter verilerine okuma ve yazma şeklinde erişim yapılmasını sağlar. Bu arayüzün aktif olarak kullanılan v1.1 sürümü ile geliştirici (developer) olarak kendilerini Twitter uygulaması içinde tanımlayan Twitter kullanıcıları ile veri paylaşımı işlemi sağlanmaktadır. API içinde yer alan farklı sorgu linkleri aracılığı ile Twitter kullanıcıların (hesapların) durumları, ayarları okunabilmekte veya güncellenebilmekte, hatta bağlı hesaba ait Twitter ortamı kullanılmadan hesaba ait mesajlar okunabilmekte ve mesaj gönderilebilmektedir (get/post statuses/users/account, vb.), ilgili hesabı takip eden veya hesabın takip ettiği kullanıcı hesaplarının listeleri alınabilmekte (get friends / followers, vb.), Twitter ortamında yer alan bir mesajın kimin tarafında kime veya hangi hashtag ile gönderildiği, tekrar edilerek gönderilen mesajların (retweet) kaçıncı defa tekrar edildiğine kadar daha birçok Twitter ortamında herkese açık olarak yayınlanan verilere ulaşılabilmekte, Twitter ortamında da güncellenebilme imkânı olan tanım ve opsiyonların güncellenmesi yapılabilmektedir.

Twitter REST API v1.0 sürümü 12 Haziran 2013 tarihinden itibaren kullanımdan kaldırılmıştır (Twitter, 2013). REST API v1.1 sürümüne geçilmesinin nedenlerinin bazıları;

- Twitter ortamından veri alan uygulamaların Twitter geliştiricileri için tanımlanan kimlik doğrulama (authentication) tanımlarına ihtiyaç duymaması,
- Tek bir sorgulama ile 1000 adet tweet mesajının Twitter veritabanından alınabilmesidir.

v1.0 ile yapılan aramalarda aramayı yapan Twitter hesabının kimlik doğrulama bilgileri Twitter firmasına aktarımı gerçekleşmez iken, kullanıma alınan v1.1 ile artık bu bilgiler Twitter'a bildirilmeden sorgulama yapma imkânı verilmemektedir. Ayrıca tek bir sorgu ile 1000 tweet mesajı yerine son 100 tweet mesajı veya 100 adedi geçmemek kaydı ile son 7 günlük mesajlar Twitter veritabanından alınabilmektedir.

Twitter geliştiricileri v1.1 ile Twitter veritabanından veri almak için birçok sorguda (veri isteğinde) kimlik doğrulaması yapma süreci Twitter yapısında yer alan OAuth kimlik doğrulama yapısı ile güvenli bir şekilde gerçekleştirilir (Twitter, 2016). OAuth yapısı ile

Twitter uygulama arayüzlerine yetkilendirilmiş erişim imkânı sağlanmaktadır. v1.1 öncesinde de Twitter kurumsal uygulamaları haricinde kullanıcı niteliğinde firma dışı kişisel geliştiricilerin yazılımları ile Twitter işlemlerini gerçekleştirirken (tweet gönderme-alma, profil bilgisi değiştirme, vb.) OAuth kullanılmaktaydı.

Twitter ana veritabanında her kullanıcıya bir tekil kullanıcı numarası (User ID) atanmaktadır (Twitter, 2013). Veritabanlarında verilerin tutulacağı alanlar tanımlanırken tutulacak verinin özelliğine göre bir veri tipi tanımı içerir ve sayısal veri tiplerinin üst ve alt sınırı vardır. Kullanıcı sayısının hızla artması sonucunda sayısal bir veri olan bu numaranın üst sınırını artırmak için Twitter 21 Ekim 2013 tarihinde veri tipi değişikliğine gitmiştir. Maksimum 32-bit olan User ID değeri, 64-bit boyutuna çıkarılmıştır (Oozeman-Kurrik, 2013). Bu değişim Twitter ortamından veri alan sistemleri ilk aşamada olumsuz yönde etkilemiştir.

3.2.2. JSON - Java Script Nesne Gösterimi

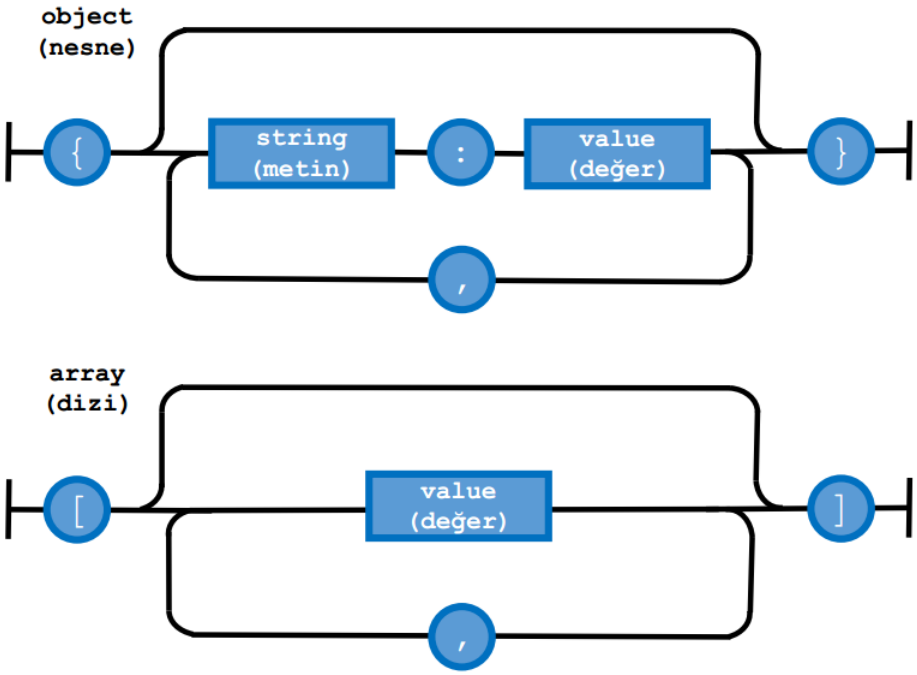
Twitter veriyi JSON (Java Script Object Notation – Java Script Nesne Gösterimi) isimli bir veri değişim biçimi ile paylaşmaktadır. Bu formatı insanların okuyup, yazabilmesi ve bu formatı makinelerin tarayıp, oluşturabilmesi oldukça kolaydır. JSON, programlama dillerinden bağımsızdır ve C türevi dillere (C, C++, C#, Java, JavaScript, Perl, Python ve daha pek çoğu) yazılış bakımından çok benzeyen bir veri tanımlama formatıdır (JSON Organization, 2013). Bu özellikler, JSON veri değişim biçimini veri alış verişi için ideal hale getirmektedir.

```
{
   "isim":"Feridun",
   "yas":45,
   "telefon":[  "0(216) 6771630",
                "0(216) 6771647"]
}
```

Şekil 3.3: Basit JSON veri değişim biçimi (Oracle, 2013).

Veri değişim biçimleri bilginin yapısal özelliğini tanımlayan meta-verinin kodlanmasını desteklerler ve geliştirirler, özellikle JSON ve XML en çok kullanılan veri değişim biçimidir (Nurseitovve diğ., 2009). Şekil 3.3'de görüldüğü gibi JSON insanlar ve makineler tarafından kolay okunup yazılması sağlayacak yapıda oluşturulan bir veri değişim biçimidir, JSON ile veri transferi yapılırken Şekil 3.4'de yer aldığı gibi

değişkenler (nesneler) ile bu değişkenlerin değerlerinin gösterimi ile dizi nesnelerinin gösterim biçimleri birbirinden farklıdır, bu tanım bloklarında yer alan her noktalama işaretinin yapı içinde farklı anlamları vardır (Oracle, 2013).



Şekil 3.4: JSON yapısında nesne (değişken) ve dizi nesnesine değer atama (Oracle, 2013).

Programcı dostu JSON yapısında tekrarlayan değişken isimleri olmaz, veri tipleri direkt olarak değişkenlerle ilişkilendirilmiştir, programcı dostu olmayan JSON yapılarında veri yapısı basit değildir ve çözümlenmesi için çaba gerektirirler (Boyer, ve diğ., 2011). Şekil 3.5'de çözümlenmesi kolay olan bir JSON örnek veri değişim yapısı yer almaktadır.

3.2.3. Twitter REST API – GET search/tweets Sorgusu

Twitter mikroblog sitesi kullanıcıların attığı mesajlar ile oluşturduğu veritabanı içindeki verileri geliştirici olarak sisteme kendini tanımlayan kullanıcılar ile bu verileri paylaşmakta veya kullanıcıların Twitter web sitesini ya da akıllı cihazlar üzerinde yüklü olan Twitter uygulamasını kullanmadan Twitter veritabanına normal kullanıcı veri aktarabilmesine imkân tanımaktadır. Bu işlemler gerçekleştirilir iken REST API içinde yer alan farklı sorgular kullanılmaktadır.

```
{ "Kitaplar":
  [
     { "ISBN":"ISBN-6-05-422038-0",
       "Fiyat":26,
       "Basım":2,
       "Adı":"Bilgi ve Bilginin Yönetimi",
       "Yazarlar": [ {"Ad":"Sevinç", "Soyad":"Gülseçen"}]}
  ,
     { "ISBN":"ISBN-0-12-374856-0",
       "Fiyat":132,
       "Açıklama":"Makine Öğrenmesi Teknikleri",
       "Basım":3,
       "Adı":"Data Mining",
       "Yazarlar": [ {"Ad":"Ian H.", "Soyad":"Witten"},
                     {"Ad":"Frank", "Soyad":"Eibe"},
                     {"Ad":"Mark A.", "Soyad":"Hall"}]}
  ]
}
```

Şekil 3.5: Programcı dostu JSON örneği

Bu sorgular get (veri alma) ve post (veri gönderme) olmak üzere iki türlüdür (Twitter, 2013). Örneğin sorguya parametre olarak girilen ifadeyi içeren son tweet mesajını Twitter veritabanından sorguyu çalıştıran kullanıcıya aktaran sorgu GET search/tweets sorgusudur. Twitter ortamında her mesajın tekil bir numarası (ID) vardır, bu doğrultuda GET search/tweets sorgusuna belli bir Tweet ID aralığında arama yapmak için parametre girilebilmektedir, ayrıca veritabanında arama yapmak istenilen ifade veya ifadeler de parametre olarak girilmektedir. Tercihe bağlı olarak aramanın yapılacağı dil, coğrafi bölge ve son kaç adet ifade ile bağlantılı tweet mesajının getirileceği parametre olarak tanımlanabilmektedir. Bu sorgunun veritabanı üzerinde çalıştırılabilmesi için sorgu öncesinde kullanıcının Twitter ortamında kimlik doğrulamasından geçmesi gerekmektedir, doğrulama sonrasında ilgili parametreler ile sorgu çalıştığında JSON formatında verilerin kullanıcıya aktarımı sağlanmaktadır. Aktarılan veriler ilgili ifade veya ifadeleri mesajın gövdesinde barındıran son 100 tweet mesajıdır, ayrıca bu mesajlar sorgunun çalıştırıldığı günden en fazla geriye dönük olarak 7 gün öncesine ait olabilirler.

GET search/tweets sorgusuna ait örnek soru linki aşağıdadır. Bu soru linki ile “Veritabanı Yönetimi” ifadelerine ait son 2 tweet mesajına ait verilerin veritabanından alınması gerçekleştirilmiştir.

```
https://api.twitter.com/1.1/search/tweets.json?q=veritabanı
yönetimi&since_id=0&max_id=9999999999999999&lang=tr&locale=tr&count=2
```

Yukarıdaki link içinde yer alan Türkçe karakter gerçekte link çalıştırılırken aşağıdaki şekilde oluşmakta ve linki çalıştıracak uygulamanın kodlama yapısı içinde de aşağıdaki şekilde yer almaktadır.

```
https://api.twitter.com/1.1/search/tweets.json?q=veritaban%C4%B1%20y%C3%B6
netimi&since_id=0&max_id=9999999999999999&lang=tr&locale=tr&count=2
```

Yukarıdaki sorgu çalıştırıldığında Ekler bölümünde yer alan JSON yapısı Twitter mikroblog sitesi tarafında oluşturularak kullanıcılarla paylaşılmaktadır.

3.2.4. Twitter Verisi Seçme Arayüz Uygulaması

Twitter ortamında yer alan tweet mesajlarına ulaşmak ve üzerinde duygu analizi gerçekleştirebilmek için verinin Twitter ortamından çekilmesi gerekmektedir, bu işlem için bir veri seçme arayüz uygulaması geliştirilmiştir. "Bölüm 3.1 Problemin Tanımlanması" başlığı altında belirtilen istatistiksel verilere göre Türkiye'de farklı zamanlarda gündemi yakalayan tweet mesajlarında çok büyük artışlar yaşanmaktadır, bu neden ile gündemi yakalayan tweet mesajlarına ulaşmak için ise bu uygulamanın sürekli olarak Twitter veritabanında sorgu yapması ve veri çekmesi gerekmektedir. Bu doğrultuda uygulama yazılımı belli zaman aralıklarında Twitter veritabanından belirlenen kelime veya kelime öbeklerine göre arama yapıp veri çekmesini sağlayacak bir motor uygulaması (engine application) şekline dönüştürülmüştür. Bu motor uygulaması aktif olarak çalışmakta ve her 30 dakikada bir Twitter veritabanından veri çekme işlemi gerçekleştirmektedir.

Veri seçme arayüz uygulamasının Twitter dan veri alma aşamaları;

- Twitter'ın OAuth yapısı aracılığı ile Twitter üzerinde kullanıcı kimliğini doğrulama, (veri seçimi)
- Twitter REST API arayüzü içinde yer alan GET search/tweets sorgusunu belirlenen konu başlıkları için çalıştırma, (veri seçimi)
- Sorgular sonucunda Twitter API'si tarafında üretilen JSON biçimindeki (formatındaki) veriyi uygulamanın çalıştırıldığı sunucuya getirme, (veri seçimi)
- Gelen mesajlar (Tweets) verisini (JSON biçimindeki veriyi) veri çözümleme (serialize) teknikleri ile çözümleme, (veri seçimi – veri önişleme)

- Veritabanı sunucusu bağlantısı ve veritabanı sunucu platformu içindeki kaydedilmiş (stored) prosedürler, tetikleyiciler (triggers) ve imleçler (cursors) gibi veri teknolojileri aracılığı ile tekrar eden, link içeren mesajların gelen veri içinden temizleme, (veri önişleme)
- Temizleme sonrası mesaj verilerinin veritabanı tablolarına konuya göre sınıflandırılma yapılarak aktarma, (veri önişleme)

şeklinde sıralanmaktadır.

Veri seçme arayüz uygulaması genel olarak tez sürecinin veri seçimi aşamasına hizmet etmekte fakat yukarıda görüldüğü gibi veri önişleme aşamasına da hizmet etmektedir.

Veri seçme arayüz motor uygulaması ile uygulama içinde tanımlanan konu (kelime) veya konu başlıkları (birden fazla kelimeden oluşan bütünlük) her 30 dakikada bir Twitter mikroblog sitesinde sorgulanmakta, sorgulama sonucunda ilgili konu veya konu başlığını mesaj gövdesinde veya kullanıcının adında barındıran son 7 gün içinde yayınlanan 100 adet tweet mesajı veritabanına aktarılmaktadır.

3.2.5. Veri Çözümleme (Serialize) - Özgün Sınıf

JSON formatında gelmiş veriyi çözümleme (serialize) işlemi için programlama dillerinde özel hazır sınıflar (class) kullanılmaktadır. Örneğin Microsoft.net ortamı için geliştirilen NuGet Galary (Outercurve Foundation , 2013) yazılım paketi içinde yer alan TweetSharp (Crenna, 2013) isimli sınıf da bu amaç için kullanılan en popüler sınıflardan biridir. Twitter REST API arayüzünde Bölüm 3.2.1 de belirtildiği gibi farklı zamanlarda Twitter tarafından güncellemeler yapılmakta ve bazen radikal bir kararla üst sürüme yükseltilebilmektedir. Bu tip köklü gelişmeler sonrası TweetSharp gibi hazır veri çözümleme sınıflarının REST API ile uyum sağlayacak güncellemelerini zamanında gerçekleştiremedikleri görülmektedir. Geliştiricilerin çalışmalarını etkileyen bu gibi durumlar ile karşılaşmamak ve hazır yazılım paketlerine bağımlı kalmamak için C# programlama dili kullanılarak Twitter'ın JSON formatındaki verisini çözümleyen özgün bir **çözümleyici sınıf** (*serialize class*) geliştirilmiştir. Bu çözümleyici sınıfın veri seçme arayüz uygulamasına entegre edilmesi ile arayüz yazılımı tamamen bağımsız doğrudan Twitter ortamının paylaştığı veriyi transfer edebilen özel bir yazılım özelliği kazanmıştır.

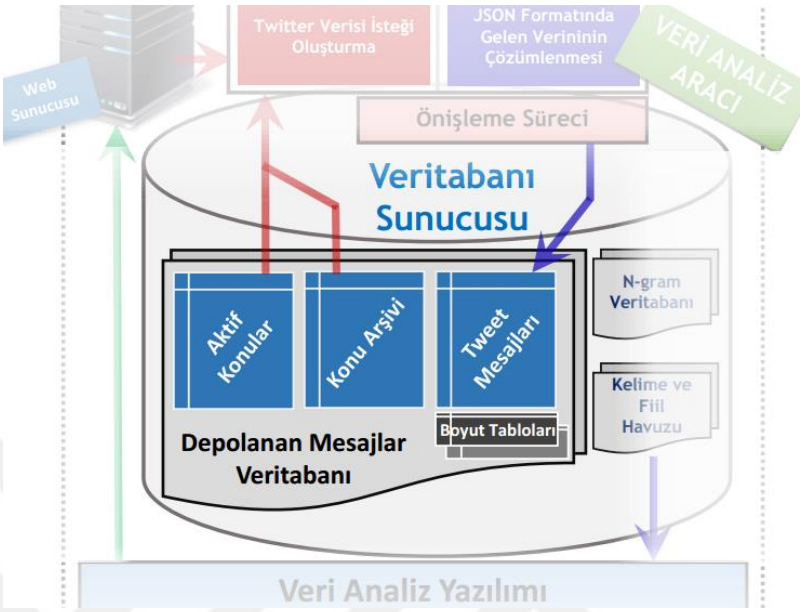
Twitter sosyal ağ sitesinin JSON formatında geliştiriciler ile paylaştığı veriyi bir grup geliştirici tarafından hazırlanmış hazır yazılım paketlerinde yer alan çözümleyici sınıflar kullanılarak çözümlemesi yapıldığında, tweet mesajlarına ait her bir tweet mesajına özel olan tekil bir tamsayı olarak Twitter veritabanında tutulan ID, oluşturulma tarihi (CreatedDate), tweet mesajını gönderen kullanıcının tekil bir tamsayı olan kullanıcı numarası (UserId), kullanıcının ekranda görünen adı-takma adı (ScreenName), kullanıcın sistemde tanımlı adı (UserName), mesajın gönderildiği kişiye ait kullanıcı no (InReplyToUserId) ve adı (InReplyToScreenName), mesajın kendisi (Text), mesajın gönderildiği lokasyonun enlem ve boylam bilgisi (Location), mesajın gönderildiği ortamın-mobil cihaz veya web uygulaması (Source) bilgileri alınabiliyor.

Özgün olarak kodlama yapılarak oluşturulan çözümleyici sınıf ile Twitter sosyal ağ sitesinin JSON formatında verisi çözümlendiğinde yukarıdaki bilgilere ek olarak; mesajı gönderen veya alan kişinin hesabına ait tüm veriler (lokasyonu, hesabının korumalı olup olmadığı, takipçilerinin sayısı, takipçilerinin adı ve diğer bilgileri, takip ettiklerinin sayısı ve diğer bilgileri, hesabın oluşma tarihi, favorileri gibi 39 farklı veriye ve bu verilerin alt kırılımlarına) ulaşılmaktadır. Ayrıca retweet edilmiş mesajın kaçıncı defa retweet edildiği, kim tarafında retweet edildiği, orijinal mesajı oluşturanın tüm açık verilerine ve daha birçok detay veriye ve alt kırılımlara geliştirilen özgün sınıf ile erişim yapılabilmektedir. Çözümleyici sınıf ile bir tweet mesajı ile ilgili olarak değerlendirilebilecek her türlü detay ve alt detay veriye hiçbir kısıtlama olmaksızın erişebilmektedir.

3.2.6. Mesajlar Veritabanı

REST API aracılığı ile alınan bilgilerin depolanması için bir veritabanı tasarlanmış ve bu veritabanı Microsoft SQL veritabanı sunucusu üzerinde oluşturulmuştur. Oluşturulan veritabanı, Twitter üzerinde arama yapılacak kelimelerin tutulduğu "**Aktif Konular**", Twitter üzerinde daha önceden araması yapılan ama aktif zamanda araması yapılmayan arama işlemi açısından güncel olmayan kelimelerin (eski aramalarda kullanılan kelimelerin) arşiv amacı ile yedeklendiği "**Konu Arşivi**" ve uygulama yazılımı aracılığı ile belirlenen kelime veya kelime öbekleri (bu kelimeler "Aktif Konular" tablosundan besleniyor) doğrultusunda Twitter veritabanında yapılan arama sonucuna göre belirlenen kelime/kelime öbeklerini mesajın içinde veya hashtagin içinde veya gönderen/gönderilen

tweet hesabı ifadesi içinde barındıran mesajların depolandığı "**Tweet Mesajları**" isimli üç ana tablodan oluşmaktadır (Şekil 3.6).



Şekil 3.6: Twitter ortamından veri alımı için hazırlanan veritabanı ve tablolar.

Ayrıca veritabanı içinde veri toplama aşamasının ve sorgulama işlemlerinin daha etkin gerçekleşebilmesi için boyut (dimension) tabloları kullanılmıştır. Bu veri tabloları aracılığı ile veri toplama işleminin ve daha sonradan analiz esnasında kullanılacak ileri düzey sorgulamaların verimli çalışması hedeflenmiştir. Veritabanındaki ana (fact-olgu tablosu niteliğindeki "**Tweet Mesajları**") tablosundaki veri sayısı arttıkça bu tablo içindeki verilere yönelik gerçekleştirilen sorgulama işlemlerinin sonuçlarının dönme süreleri uzamaktadır, bu zaman kaybını önlemeye yönelik ana tablonun sorgulama esnasında kullanılan kritik veri alanlarından oluşan özet tablo niteliğinde boyut tablolar oluşturulmuştur.

3.3. VERİ ÖNİŞLEME

Veri seçme (toplama) esnasında oluşturulan veritabanı ile ilgili tüm yordamsal (programlama gerektiren) işlemler için veritabanı üzerinde oluşturulan kaydedilmiş prosedürler, tetikleyiciler ve imleçler kullanılmıştır. Veritabanı sunucusunun bu teknik alt yapısı ile büyük veri üzerindeki işlemler çok daha hızlı gerçekleşmekte ve zaman açısından yüksek performanslı sonuçlar elde edilmektedir.

REST API ve bu API ile entegre çalışan uygulama yazılımı aracılığı ile yapılan sorgulamalar ile seçilen (toplanan) ve tez kapsamında hazırlanan veritabanına transfer edilen mesajlar Twitter mikroblog uygulamasında yayınlandığı şekilde alınmaktadır. Twitter'ın veri paylaşım için sunduğu API'sine, mesajların yapısına, içeriğine ve yazılan motor uygulamasına bağlı olarak veri transferi esnasında temiz veri toplanmasını engelleyen aşağıdaki durumlar oluşmaktadır:

- Konu farklı olmasına rağmen aynı mesaj içinde birden fazla konuya ait terim geçtiğinde her veri alma işleminde aynı mesajlar birden fazla aktarılabilmektedir,
- İlgili konu başlığında çok fazla yeni mesaj atılmadığında her veri alma işleminde daha önceki aktarımlarda alınan mesajlar tekrar veritabanına aktarılabilmektedir,
- Aynı mesaj olmasına rağmen tekrar tekrar retweet edilen mesajlar farklı mesajlar gibi aktarılabilmektedir,
- Mesajlar içinde yer alan link veya paylaşılan resimler "http://t.co" ve "https://t.co" ile başlayan, direkt bakıldığında çözümlenemeyen, bir internet tarayıcıya kopyalanıp çalıştırıldığında gerçek link adresine veya resme ulaşılabilen ifadeler olarak gelmekte, hatta tüm mesaj link şeklinde olabilmektedir,
- Mesajlar yazılırken dilbilgisi kurallarına göre yazılmadığından mesajlar içinde gereksiz sayıda fazladan kelimeler arasında boşluklar, noktalama işaretleri olabilmektedir,
- Mesajların tamamı noktalama işaretleri ile oluşabilmektedir.

Analiz yapabilmek için bu şekilde gelen mesajların analiz yapılabilir düzeye getirilmesi gerekir. Veri niteliğindeki bu mesajlardan tekrar edenlerin bir kopyasının bırakılarak geri kalanlarının veri setinden çıkarılması (veri temizleme – data cleaning), analize uygun olmayanların analize uygun hale dönüştürülebiliyorsa dönüşümünün yapılması (veri dönüşümü – data transformation) işlemleri veritabanlarında bilgi keşfi süreçlerinde veri önişleme (data preprocessing) aşamasına karşılık gelmektedir.

Analiz için gerekli veri seti oluşturulurken geliştirilen uygulama yazılımı içindeki ileri veritabanı teknikleri ile aşağıdaki veri önişleme işlemleri gerçekleştirilmiştir.

- Veri seçimi esnasında ara tampon geçici veritabanı tabloları ve imleç (cursor) uygulamaları ile mükerrer gelen mesajları daha mesajlar alınırken filtrelemesini yapan kodlama geliştirilmiştir.

- Daha önceden alınmış mesajlar ile yeni alınan mesajların hızlı bir şekilde karşılaştırmasını yapan, gelen mesajın mevcut veritabanında varlığını tespit edip, bu mesajı filtreleyen kodlama geliştirilmiştir. Bu kodlama gün içinde veri alımı haricinde de veritabanı tarayarak mükerrer mesaj varlığını kontrol etmektedir.
- Veri seçimi esnasında retweet mesajlar gibi benzer olan mesajların filtrelemesi için gerekli kodlama geliştirilmiştir.
- Mesajlar içinde yer alan link veya paylaşılan resimlere ait "http://t.co" ve "https://t.co" ile başlayan linkler her veri seçiminde filtrelenmektedir.
- Mesajlar içinde yer alan gereksiz boşluk ve noktalama işaretleri geliştirilen filtreleme yazılımı ile analiz edilebilir mesajların veritabanına aktarılmasını sağlamaktadır.

3.4. MODEL OLUŞTURMA

Tez veritabanında toplanan tweet mesajları ya da belirlenen kriterlere göre güncel Twitter mikroblog sitesinin veritabanından çekilen anlık tweet mesajları üzerinde **duygu analizi** yapılmasını sağlayacak web tabanlı **Veri Analiz** yazılımı geliştirilmiştir. Bu yazılım iki katmandan oluşmaktadır.

- Kelime - Fiil tabanlı veri analizi
- N-gram tabanlı veri analizi

3.4.1. Anahtar Kelime ve Fiil Tabanlı Veri Analizi

Mikroblog ortamlarında yer alan Türkçe görüşler (mesajlar) üzerinde fikir analizi (duygu analizi) yapmak temel amaç olduğundan tez kapsamında Türkçe mesajlar üzerinde doğal dil işleme ve metin madenciliği çalışmaları yapılmıştır.

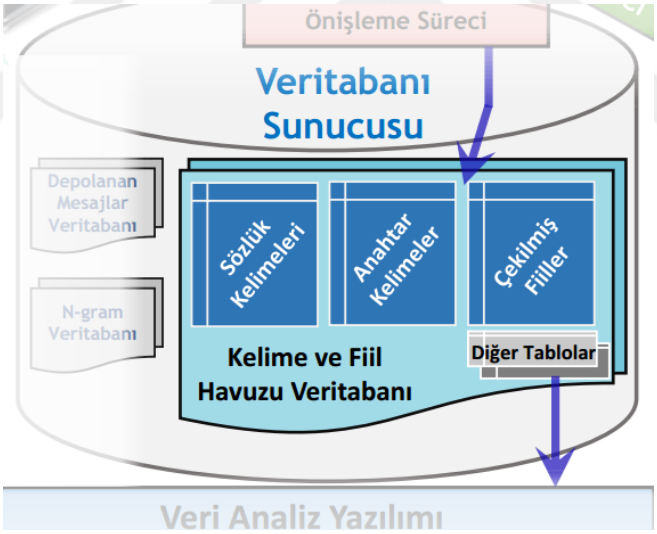
"Zihin, cümleyi oluşturan sözcükler ve komşuları arasında ilişkiler bulur ve bu ilişkilerin bütünü cümlenin iskeletini oluşturur. Her bir ilişki bir alt terimi bir üst terime bağlamaktadır." Günümüzde doğal dil ayrıştırma (DDA) alanında kullanılan bağlılık gramerlerinde bu ilişki bağımlı (alt terim) - sahip (üst terim) ilişkisi olarak tanımlanmaktadır (Eryiğit ve diğ., 2006).

Bölüm 2.7'de belirtildiği gibi Türkçede bir cümledeki duygu, şahıs, zaman ve anlam, cümlenin yapısı içindeki fiil veya fiilin çekilmiş halinden tespit edilir. Türkçe metinler

üzerinde duygu analizi çalışması yapılırken, cümleye olumlu veya olumsuz duygu katan kelimeler ve cümledeki eylemi ifade eden fiiller (yüklemler) aracılığı ile çalışma gerçekleştirilmelidir.

"Bölüm 2.4 Doğal Dil İşleme" başlığında belirtildiği gibi Türkçe üzerindeki doğal dil işleme çalışmalarında, bütün bir cümle kelime seviyesinde, hece seviyesinde parçalara ayırarak, anlamsal ve yapısal olarak çözümlenmek istenmektedir. Bu tez çalışmasında ise cümleleri (mesajları) anlamsal olarak çözümleme sürecinde genel dilbilimcilerin kullandığı bütünden en küçük parçaya (hecelere ve kaynaştırma harflerine kadar) ayrıştırma yöntemi yerine cümleyi sadece kelimelere kadar ayrıştırması yapılmış, Türkçede yer alan bir cümleye anlamsal olarak olumluluk veya olumsuzluk duygusu veren kelimelerden ve cümlede yer alan fiillerin olumsuzluk eki alıp almamasına bakılarak cümle genelinin duygu analizi yapılmıştır.

3.4.1.1. Kelime ve Fiil Havuzu Veritabanı



Şekil 3.7: Kelime ve fiil havuzu veritabanı.

Elektronik ortamlarda yayınlanan "Türkçe Sözlük" verilerinden derlenerek Türkçede yer alan yaklaşık 65000 kelime ile "**kelime havuzu**" niteliğinde bir veritabanı oluşturulmuştur (Türk Dili ve Edebiyatı Dersleri Kaynak Eğitim Sitesi, 2007; Dil Bilgisi.net, 2014; Vikisözlük, 2015). Bu veritabanında asıl amaç, metine (görüşe-cümleye) olumlu veya olumsuzluk duygusu katan tüm kelimeleri ve çekimli/çekimsiz fiilleri depolamaktır.

Geliştirilen yazılım aracılığı ile mesaj içinde ayrıştırılan her kelime veritabanı içinde niteliklerine göre sınıflandırılmış (fiiller, olumlu veya olumsuzluk anlamı veren kelimeler gibi) kelimelerden oluşan tablolardaki kayıtlar ile karşılaştırılmaktadır. Karşılaştırılan kelime tablolarda yer alan kelime ile eşleşirse, tabloda yer alan kelimelerin özelliklerine bakılarak mesaj içindeki kelimenin anlamsal olarak yorumu yapılmaya çalışılmaktadır.

Şekil 3.8: Anahtar kelimelerin değerlendirildiği anket.

Anahtar Kelimeler

Kelime Havuzu veritabanına yaklaşık 1000'e yakın olumlu veya olumsuz duygu katan kelime eklenmiştir. Bu kelimeler tez kapsamında oluşturulan bilişim alt yapısı içindeki web sunucusu üzerinde geliştirilen anket yazılımı aracılığı (Şekil 3.8) ile 36 adet lisans ve yüksek lisans öğrencisi tarafından kişi üzerinde oluşturmuş oldukları "Olumlu" , "Olumsuz" ve "Nötr" duygular açısından değerlendirilmiştir. Değerlendirme seçenekleri "Çok olumsuz duygu (-2)", "Olumsuz duygu (-1)", "Nötr duygu (0)", "Olumlu duygu (+1)" ve "Çok olumlu duygu (+2)" şeklinde belirlenmiştir.

Tablo 3.1: Anahtar kelimeler için duygu sınıf aralıkları.

Duygu Sınıfı	Alt Değer	Üst Değer
Olumsuz	-2,00	-0,41
Nötr	-0,40	+0,40
Olumlu	+0,41	+2,00

Anket sonucunda her kelimeye ait elde edilen sayısal değerlendirmeler üzerinden yapılan hesaplama doğrultusunda kelimeler için çıkan sayısal sonuçlar Tablo 3.1'de belirtilen aralıklarda bulunma durumlarına göre duygu analizi açısından hangi duygu sınıfını (sentiment polarity) temsil ettiği belirlenmiştir (Tablo 3.2).

Tablo 3.2: Anahtar kelimelerden bazılarının hesaplanan duygu sınıfları.

Kelime	Duygu	Değer
tekin	Olumlu	+0,57
çatlak	Nötr	-0,31
beddua	Olumsuz	-1,79
sarhoş	Olumsuz	-1,64
erdem	Olumlu	+1,50
pusu	Olumsuz	-1,29
çekingen	Olumsuz	-0,54
köpek	Nötr	+0,15

Depolanan mesajlara ve çevrimiçi ile alınan mesajlar incelendiğinde Türkçe metin yazılırken, bazı kullanıcıların latin alfabesinden yer almayan Türkçe karakterleri kullanmak yerine latin alfabesinde o karaktere karşılık gelen karakterleri kullandıkları gözlemlenmektedir. Anahtar kelimeler yapısı içinde yer alan her bir kelimenin latin alfabesi ile yazılabilecek formları da veri yapısı içine alınmıştır. Bu formdaki kelimeler anket çalışması içine dahil edilmemiş, olumluluk (duygu) değeri olarak ta asıl olan kelimelerin hesaplanan değerleri bu kelimelere de atanmıştır.

Fiiller

Bölüm 2.7'de belirtildiği gibi Türkçede fiillerin yalın hallerine olumsuzluk eki eklenerek cümleler olumsuzlaştırılır veya fiilin yalın haline zaman ekleri eklenerek cümlenin olumlu olması ve zamanı belirlenir. Aynı şekilde olumsuzluk ekinden sonra zaman eki koyularak cümlenin olumsuz olması ve eklenen zaman ekine göre de zamanı belirlenir. Fiillere zaman eki eklendikten sonra cümlenin genelde öznesini belirten şahıs eki (ben,

sen, o, biz, siz, onlar) eklenmektedir. Örneğin konuşmak fiilinin 1.tekil şahısa ve gelecek zamana göre olumlu çekilmiş hali "konuş-a-cağ-ım" şeklindedir, burada "konuş" kök, "-a" kaynaşma eki, "-cağ" gelecek zaman eki olan "-cak" tır, fakat buradaki "-k" şahıs ekine bağlı olarak yumuşamış "-ğ" olmuştur. "-ım" ise 1.tekil şahıs ekidir. Aynı fiilin geniş zamanda 2.çoğul şahısa göre çekilmiş hali "konuş-maya-cak-sınız" şeklindedir, burada "konuş" kök, "-maya" olumsuzluk eki, "-cak" gelecek zaman eki, "-sınız" ise 2.çoğul şahıs ekidir.

Tablo 3.3: Bazı fiillerin 3.tekil şahısa göre bazı zamanlara göre çekimi.

Fiil	Geçmiş Zaman (-di li)	Geçmiş Zaman (-miş li)	Şimdiki Zaman	Geniş Zaman	Gelecek Zaman
git	**git**-ti	**git** -miş	**gi-d**-iyor	**gi-d**-er	**gi-d**-ecek
öt	**öt**-tü	**öt**-müş	**öt**-üyor	**öt**-er	**öt**-ecek
it	**it**-ti	**it**-miş	**it**-iyor	**it**-er	**it**-ecek
acele **et**	**et**-ti	**et**-miş	**e-d**-iyor	**e-d**-er	**e-d**-ecek

Türkçe sondan eklemeli bir dil olduğundan fiillerin sonuna gelecek ekler sona eklenirken kök kelimenin sonunda veya ekin kendisinde ses etkileşimine bağlı olarak özgün hallerine göre değişimler (yumuşama, kaynaşma için harf düşümü veya harf ekleme gibi değişimler) yaşanmaktadır (Tablo 3.3). Bu değişimler kelimeden kelimeye göre farklı şekilde gerçekleşmektedir. Örneğin, "git" fiilini şimdiki zamanın 3.şahısına göre çektiğimizde "gid-iyor" şeklini alır, buradan "köklerinin sonu -t harfi ile biten fiillerin şimdiki zaman 3.şahısa göre çekimi yapıldığında kökün son harfi –d ye dönüşür" genelleme kuralını oluşturursak, "öt" fiilinin şimdiki zaman 3.şahısa göre çekimini yaptığımızda "öt-üyor" şeklinde olduğunu görürüz ve oluşturduğumuz genellemenin geçersiz olduğu ortaya çıkar. Kökün son iki harfine (-it) göre bir genel kural tayin edilirse ("-it şeklinde biten fiil kökleri çekim sonunda -id olur" şeklinde bir kural oluşturulursa) bu kuralında "it" fiilinin çekiminde (it-iyor) işlevselliğini kaybeder.

Fiillerin çekimini zorlaştıran yukarıdaki açıklanan durumlara benzer sebepler, elektronik ortam (yazılım) aracılığı ile sistematik bir algoritma yapısı doğrultusunda yalın haldeki

fiillerin zaman ve şahıslara göre çekiminin yapılıp, veritabanında kategorize edilmiş olarak saklanmasını oldukça fazla zorlaştırmaktadır.

Türkçede fiillerin mastar durumu almış oldukları "-mek" veya "-mak" ekleri ile ifade edilir. Elektronik ortamlarda yayınlanan "Türkçe Sözlük" verilerinden derlenerek oluşturulan "kelime havuzu" veritabanı içinde mastar eki almış kelimelerden yola çıkılarak fiil özellikli kelimeler belirlenmiş ve bu fiillerin çekimleri işlevsel veritabanı sorguları ile gerçekleştirilmiştir. Bu işlem gerçekleşirken uygulanan aşamalar;

- Yaklaşık 65000 kelime içinde mastar eklerine sahip kelimeler SQL sorguları ile tespit edilmiş ve bir veritabanı tablosunun ("Fiiller" tablosu) içine yalın halleri ile (-mek ve –mak ekleri yer alamadan, kök halleri ile) aktarılmıştır.
- Aynı tablo içine her bir fiilin yalın hali için Türkçenin çekim kuralları doğrultusunda olumlu ve olumsuz (Türkçede cümleyi olumsuz yapan ekler dikkate alınarak) halleri ileri veritabanı sorgulama teknikleri ile oluşturulup aktarılmıştır. Bu aktarım esnasında fiillerin dil içinde kullanımları tek tek dikkate alınarak fiil çekim kuralları oluşturulmuş, bu kurallar doğrultusunda veritabanı sorguları yazılmış, bu sorgular ile çekimler yapılarak aktarım gerçekleştirilmiştir.
- "Fiiller" tablosundaki fiillerin olumlu ve olumsuz halleri temel alınarak fiillerin zamana (dili geçmiş zaman, mişli geçmiş zaman, şimdiki zaman, geçmiş zaman, gelecek zaman ve diğer zamanlara) ve şahıslara (ben, sen, o, biz, siz, onlar şahıs yapılarına) göre çekilmiş halleri ileri veritabanı sorgulama teknikleri kullanılarak kaydedilmiş (stored) prosedürler ile sınıflandırılmış olarak oluşturulup veritabanına aktarılmıştır. Bu işlem esnasında da tüm zamanlarda ve şahıs çekimlerinde fiillerin dil içinde kullanımları her bir fiil dikkate alınarak fiil çekim kuralları ve kurallara uygun yazılan veritabanı sorguları ile çekimler gerçekleştirilmiş ve "Fiiller" tablosuna yeni formlar aktarılmıştır.

Kelime havuzu veritabanında yalın (eksiz) halde 6000'e yakın fiil bulunmaktadır, ama bu fiil niteliğindeki kelimelerin tüm zaman, tüm şahıs (kişi) ve olumsuzluk ekleri eklenerek fiil çekimi yapıldığında ise yaklaşık 585.000'e yakın fiil verisi oluşturulmuştur. Veritabanı içinde sınıflandırılmış olarak yer alan kelimenin (bu bir çekilmiş veya çekilmemiş bir fiil olabilir ya da yalın halde bile olumlu veya olumsuz anlam ifade eden bir kelime olabilir) olumlu/olumsuz bilgisi, zaman eki almışsa hangi zamana ait olduğu

bilgisi, şahıs eki almış ise hangi şahısa ait olduğu bilgisi veritabanında kelime ile birlikte aynı satır içinde tutulmuştur.

Yukarıdaki işlemler sonucunda elde edilen çekilmiş fiillerin olumlu/olumsuzluk bilgisi olumsuzluk eki alıp almamasına bağlı olarak ek alanlar olumsuz anlamında sayısal olarak -1,20 değeri, ek almamış çekilmiş fiiller ise olumlu anlamında sayısal +1,20 değeri ile temsil edilmiştir (Tablo 3.4).

Tablo 3.4: Fiillerin olumlu/olumsuzluk durumlarının sayısal karşılıkları.

Fiilin Ek Durumu	Sayısal Karşılık
Fiil olumsuzluk eki almamış ise, (Örnek: gel-i-yor-um, koş-muş-tu)	**+1,20** (Olumlu)
Fiil olumsuzluk eki almış ise, (Örnek: gel-**mi**-yor-um, koş-**ma**-mış-tı)	**-1,20** (Olumsuz)

Türkçede bazı fiiller olumsuzluk eki almış olsa dahi anlamsal olarak olumlu olabilmektedir. Veritabanı oluşturulurken ters duygu ifade eden fiiller dikkate alınarak duygu sınıflandırma işlemleri yapılmıştır. Tablo 3.5'de bu fiiller ile ilgili birkaç örneğe yer verilmiştir.

Tablo 3.5: Ters duygu ifade eden fiil örnekleri.

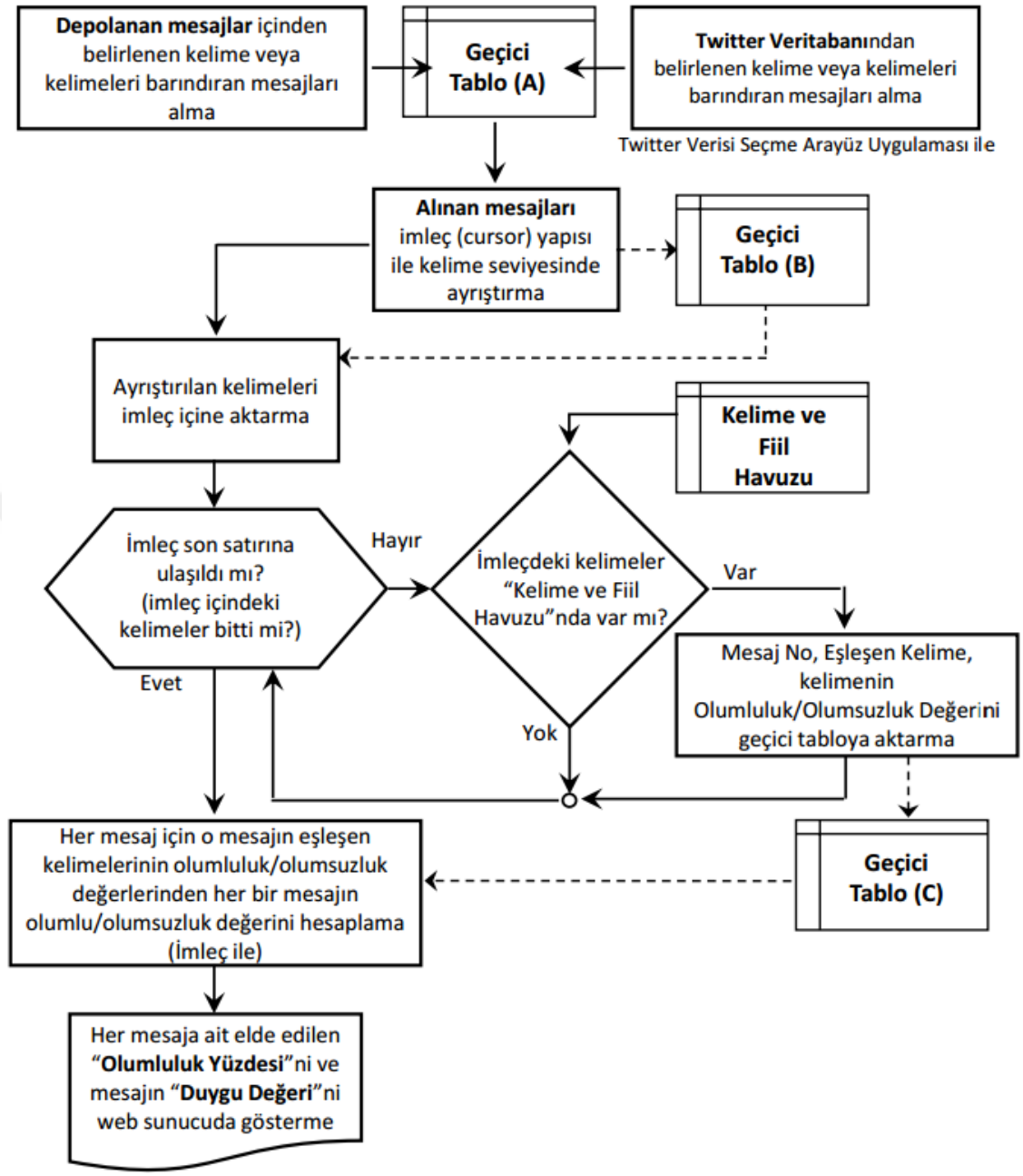
Fiil (Mastar Hali)	Geçmiş Zaman	Geniş Zaman	Duygu Durumu
üz-mek (1.tekil şahıs için)	üz-dü-m üz-**me**-di-m	üz-er-im üz-**me**-m	Olumsuz (Negatif) Olumlu (Pozitif)
kız-mak (3.çoğul şahış için)	kız-dı-lar kız-**ma**-dı-lar	kız-ar-lar kız-**maz**-lar	Olumsuz (Negatif) Olumlu (Pozitif)
kandır-mak (2.tekil şahış için)	kandır-dı-n kandır-**ma**-dı-n	kandır-ır-sın kandır-**maz**-sın	Olumsuz (Negatif) Olumlu (Pozitif)

3.4.1.2. Anahtar Kelime ve Fiil Tabanlı Veri Analizi Yazılımının Çalışması

Bölüm 3.2.6'da belirtilen belli bir konu veya terime ait Twitter sosyal ağ veritabanından yer alan tweet mesajlarını tez kapsamında geliştirilen "**Twitter Verisi Seçme Arayüz**

Uygulaması" ile Microsoft SQL veritabanı sunucusu üzerinde oluşturulan "mesajlar" veritabanında belirlenen herhangi bir konu veya terime ait depolanmış tweet mesajlarının veya mikroblog (Twitter sosyal ağı) ortamında yer alan anlık (çevrimiçi-online) Türkçe tweet mesajlarının (belirlediğimiz herhangi bir konu veya terime ait tweet mesajlarının) olumlu veya olumsuz olduğuna karar verecek yordamsal bir veritabanı uygulaması tez çalışması içinde geliştirilmiştir (**duygu analizi –sentiment analysis- uygulama yazılımı**) (Şekil 3.9). Bu uygulama oluşturulurken ASP.NET teknolojisi ve kodlama işleminde C# programlama dili kullanılmıştır. Mesajları sorgulama ve karşılaştırma işlemleri için "mesajlar" ve "kelime ve fiil havuzu" veritabanlarındaki tablolardan yararlanılmıştır. Geliştirilen uygulama web tabanlı olarak çalışmakta, mesaj içinde yer alan olumlu veya olumsuz kelimeler ve ayrıca fiillerin zamana, şahısa göre çekilmiş hallerinden olumlu veya olumsuz halleri aranmaktadır.

Duygu analizi uygulama yazılımı "mesajlar" veritabanında yer alan ana tweet mesajlarının tutulduğu veritabanı tablosuyla ve "kelime ve fiil havuzu" veritabanında yer alan çekilmiş "fiiller" ve "anahtar kelimeler" verilerini tutan veritabanı tablolarını kullanmaktadır. Uygulama esnasında "kelime ve fiil havuzu" veritabanındaki anahtar kelimelerin ve fiillerin olumlu ve olumsuz olarak sınıflandırılmış kayıtlarından yararlanılmaktadır. Her bir mesaj için fiilleri içeren tabloda yer alan her bir fiil mesaj içinde aranmakta, eğer ki mesaj içinde fiiller tablosu ile eşleşen bir fiil varsa mesaj, fiil ve fiilin tablodaki olumlu/olumsuzluk bilgisi veritabanı geçici tablo yapıları içine aktarılmaktadır. Bir mesajda birden fazla fiil olabilir, her yakalanan fiil için mesaj, fiil, fiilin olumluluk/olumsuzluk durumu tabloya kayıt olarak aktarılır. Bu işlem gerçekleştirilirken "fiiller" tablosunda yer alan yaklaşık 585.000 adet fiilin (ve ek olarak "anahtar kelimeler" tablosundaki 1000'e yakın kelimenin) her birinin mesajın gövdesinde varlığı kontrol edilir. Bu işlem standart veritabanı sorgulama yapısı (select, join komut yapıları ve SQL döngü yapıları) ile yapıldığında veya uygulama yazılımı içinde uygulanacak döngüsel yapılar ile ayrıştırma (parsing) ile gerçekleştirildiğinde oldukça çok zaman almaktadır.



Şekil 3.9: Kelime ve Fiil tabanlı veri analiz modeli.

Duygu analizi işleminin çok daha hızlı gerçekleşmesi için Şekil 3.9'da görüldüğü gibi tweet mesajları SQL yordamsal sorgulama dili içinde yer alan imleç (cursor) yapıları (oluşturulan imleç yapıları kaydedilmiş prosedürler içinde barındırılıyor - stored procedure) ile kelimeler halinde ayrıştırılmakta, her bir mesaj barındırdığı kelime sayısı kadar kayıt adedi olarak işlem esnasında veritabanı sunucusunun belleğinde oluşturulan bir geçici tabloya (B) "mesaj no" ve "kelime" verileri ile birlikte aktarılmaktadır. Bu yordamsal programlama ile geçici tablo (B) içinde yer alan mesajdan ayrıştırılmış

kelimeler “kelime ve fiil havuzu” veritabanında yer alan “fiiller” ve “anahtar kelimeler” tablolarında yer alan kayıtlar ile karşılaştırmak için imleç yapısı içine alınmaktadır. İmleç üzerinde oluşturulan bir veritabanı döngü yapısı ile imleç içindeki her bir kelime “kelime ve fiil havuzu” veritabanında yer alan “fiiller” ve “anahtar kelimeler” tablolarında varlıkları kontrol edilir. Eşleşen kelimeler, ait olduğu mesajın nosu, veritabanı tarafında eşleşen kelime ve olumluluk/olumsuzluk özelliği (sayısal değerleri) geçici tablo (C) içine tek satırlık bir kayıt olarak aktarılmaktadır. Son olarak yine imleç yapılarında yararlanarak, elde edilen son geçici tablo (C) içindeki kayıtlar ait oldukları mesajlar bazlı gruplanarak eşleşen kelimelerin olumluluk/olumsuzluk değerleri ve eşleşen kelime sayısı üzerinden yapılan hesaplama ile her bir mesajın olumlu ve olumsuzluk değeri sayısal olarak hesaplanmakta, çıkan sonuçlar üzerinden yüzde olarak olumluluk oranı bulunmakta, Tablo 3.6’da yer alan aralık değerlerine göre her mesajın olumlu, olumsuz veya nötr olduğuna karar verilmektedir.

Tablo 3.6: Tweet mesajları duygu sınıf aralıkları.

Duygu Sınıfı	Alt Değer	Üst Değer	Olumluluk %
Olumsuz	-2,00	-0,41	< %40
Nötr	-0,40	+0,40	%40 - %60
Olumlu	+0,41	+2,00	> %60

Tablo 3.7’de de yer alan mesajlar Türkiye’de hizmet veren iki farklı mobil operatöre ait abonelerin operatörler hakkında Twitter mikroblog sitesinde yayınladıkları mesajlardır.

Tablo 3.7: Örnek tweet mesajları.

Mesaj No	Tweet Mesajı
m_1	1 ayı X_Operatör'ün bedava internetleri ile devirdim seviyorum seni X_Operatör; bir de çok hızlı gitmese paket daha iyi olacak.
m_2	Y_Operatör online da berbatsın..Bu kadar kilit, geç açılan dandik bi site daha görmedim..Kapat o siteyi bi halta yaramıyor zaten..Dönence gibi

Bu mesajlar üzerinde "Kelime – Fiil Veri Analizi" sürecini işlettiğimizde mesajların her biri kelime bazlı ayrıştırılmaktadır. Her bir kelime Tablo 3.8'de küçük bölümleri görünen "anahtar kelimeler" ve "fiiller" tabloları ile karşılaştırılmakta, bu tablolar ile eşleşen kelime ve fiillerin olumluluk değerleri üzerinden her mesajın duygu (olumluluk) değeri *D(m)* aşağıdaki eşitlikle hesaplanmaktadır.

$$D(m) = \frac{\sum_{i=1}^{n} d_i}{n} \tag{3.1}$$

Bu eşitlikte $\boldsymbol{m}$ değerlendiren her bir mesaj, $\boldsymbol{n}$ her bir mesajda eşleşen toplam kelime ve fiil sayısı, $\boldsymbol{d}$ ise her kelime veya fiilin tablolarda yer alan duygu (olumluluk) değeridir. Hesaplanan bu değer -2,00 ile +2,00 arasında olur. Bu değer ile ilgili mesajın Tablo 3.6'da yer alan duygu sınıflarından hangisinde yer aldığına karar verilmektedir.

Tablo 3.7'de yer alan birinci mesajın duygu değerini hesaplamak için eşitlik 3.1 kullanıldığında; Table 3.8'de yer alan anahtar kelimelerden "bedava", "hızlı", "iyi", fiillerden de "devirdim", "olacak", "seviyorum" ifadelerinin duygu değerlerinden yararlanılır.

> $\boldsymbol{m_1}$: "*1 ayı X_Operatör'ün bedava internetleri ile devirdim seviyorum seni X_Operatör; bir de çok hızlı gitmese paket daha iyi olacak*"

$$D_1 = \frac{d_{(bedava)} + d_{(\boldsymbol{devirdim})} + d_{(hızlı)} + d_{(iyi)} + d_{(\boldsymbol{olacak})} + d_{(\boldsymbol{seviyorum})}}{6}$$

$$= \frac{(1{,}25) + (\boldsymbol{-1{,}20}) + (1{,}20) + (1{,}92) + (\boldsymbol{1{,}20}) + (\boldsymbol{1{,}20})}{6} = \frac{5{,}57}{6} = 0{,}9283$$

Elde edilen D_1 değeri $0{,}41 > D_1 > 2{,}00$ aralığında olduğundan 1 nolu mesajın duygu sınıfı "**Olumlu**" dur.

Aynı işlemleri ikinci mesaj içine gerçekleştirelim;

> $\boldsymbol{m_2}$: "*Y_Operatör online da berbatsın..Bu kadar kilit, geç açılan dandik bi site daha görmedim..Kapat o siteyi bi halta yaramıyor zaten..Dönence gibi*"

Anahtar kelimeler; "berbat", "dandik", "halt", fiiller; "geç", "görmedim", "kapat", "yaramıyor":

$$D_2 = \frac{d_{(berbat)} + d_{(dandik)} + d_{(\boldsymbol{geç})} + d_{(\boldsymbol{görmedim})} + d_{(halt)} + d_{(\boldsymbol{kapat})} + d_{(\boldsymbol{yaramıyor})}}{7}$$

$$= \frac{(-1{,}08) + (-1{,}00) + (\mathbf{1,20}) + (\mathbf{-1,20}) + (-1{,}29) + (\mathbf{1,20}) + (\mathbf{-1,20})}{7}$$

$$= \frac{(-3{,}37)}{7} = (-0{,}48)$$

Elde edilen D_2 değeri -2,00 > D_2 > -0,41 aralığında olduğundan 2 nolu mesajın duygu sınıfı "**Olumsuz**" dur.

Tablo 3.8: Anahtar kelime ve fiiller veri tablo parçaları.

Anahtar Kelimeler

Kelime	Olumluluk Değeri
bedava	1,25
berbat	-1,08
canım	1,31
dandik	-1
halt	-1,29
hızlı	1,20
hizli	1,20
iyi	1,92
mutlu	1,75
pişman	-0,80
pisman	-0,80
sevin	1,25
sömürü	-1,50
somuru	-1,50
şikayet	-1,25
sikayet	-1,25
yok	-1
zalim	-1,67
zalım	-1,67
zavallı	-0,92
zavalli	-0,92

Fiiller

Fiil	Mastar Hali	Zaman	Şahıs	Olumluluk Değeri
aldım	almak	dili geçmiş	1.Tekil	1,20
aradım	aramak	dili geçmiş	1.Tekil	1,20
bağladın	bağlamak	dili geçmiş	3.Tekil	1,20
çekmiyor	çekmek	şimdiki	3.Tekil	-1,20
devirdim	devirmek	dili geçmiş	1.Tekil	-1,20
duy	duymak	kök, emir		1,20
ettim	etmek	dili geçmiş	1.Tekil	1,20
geç	geçmek	kök, emir		1,20
gerekiyor	gerekmek	şimdiki	3.Tekil	1,20
girmiyor	girmek	şimdiki	3.Tekil	-1,20
kapat	kapatmak	kök, emir		1,20
konuşmak	konuşmak	kök, emir		1,20
kullanma	kullanmak	kök, emir		-1,20
olacak	olmak	gelecek	3.Tekil	1,20
oldum	olmak	dili geçmiş	1.Tekil	1,20
seviyorum	sevmek	şimdiki	1.Tekil	1,20
yapışmam	yapışmak	geniş	1.Tekil	-1,20
yaramıyor	yaramak	şimdiki	3.Tekil	-1,20
yıkıldım	yıkılmak	dili geçmiş	1.Tekil	-1,20

Sayısal değerleri 0 – 1 aralığında ifade edilecek şekilde dönüştürmek, değerin yüzde olarak ifade edilmesini kolaylaştıracaktır. Yüzde olarak ifade edilen değerler daha anlaşılır olacaktır.

Şekil 3.10'da olduğu gibi -2.00 değeri 0, +2.00 değeri 1 değeri olarak ifade edildiğinde örnek mesajların D_1 ve D_2 değerleri için;

Duygu Değeri	-2,0	-1,6	-1,2	-0,8	-0,4	0,0	0,4	0,8	1,2	1,6	2,0
	Olumsuz					Nötr				Olumlu	
0 - 1 Aralığı	0,0	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9	1,0
Yüzde (%)	0	10	20	30	40	50	60	70	80	90	100

Şekil 3.10: Duygu (olumluluk) değerinin 0 – 1 aralığı ve yüzde olarak gösterimi.

- D_1 = 0,928 , 0 – 1 aralığında gösterim değeri 0,7321, yüzde olarak ise %73,21 olumlu bir mesajdır.
- D_2 = - 0,48 , 0 – 1 aralığında gösterim değeri 0,38, yüzde olarak ise %38 olumlu bir mesajdır, Tablo 3.6'ya göre %40'ın altında kalan mesajlar olumsuz olarak kabul edildiğinden, zayıf bir nitelikle bu mesaj olumsuzdur.

Web tarafındaki uygulama yazılımının web ara yüzü sadece SQL sunucudaki kaydedilmiş prosedürlerin çalışmaları için gerekli parametrelerin SQL sunucuya gönderilmesini ve sunucudaki işlemler sonucunda oluşan rapor verisinin görsel olarak son kullanıcıya sunulmasını sağlamaktadır. Veri üzerinde yapılan tüm işlemler veritabanı programlaması ile gerçekleştirilmektedir.

Geliştirilen bu yordamsal veritabanı yapısı aracılığıyla duygu analizi uygulaması ile Twitter sosyal ağ veritabanından anlık (online) çekilen 100 adet tweet mesajı üzerinde olumlu/olumsuz fiil ve olumlu/olumsuz/nötr kelimeleri tespit edilmesi ve sonuçların web ortamında gösterilmesi yaklaşık 10-15 saniyelere indirgenmiştir. Sürenin kısalmasındaki en büyük etken özellikle imleç yapılarının çok etkin olarak kullanılmasıdır.

Duygu analizi, tez kapsamında geliştirilen "Twitter Verisi Seçme Arayüz Uygulaması" aracılığı ile depolanan tweet mesajları üzerinde uygulanabildiği gibi, çevrimiçi (anlık-online) olarak ta Twitter sosyal ağ veritabanında belli bir konu için yazılmış mesajlar üzerinde de gerçekleştirilebilmektedir.

3.4.2. N-gram Tabanlı Veri Analizi

Bölüm 2.5 ve 2.6'da genel bilgileri verilen N-gram modelleme ve makine öğrenmesi tekniklerinden veri madenciliği sınıflama algoritmalarından Naïve Bayes ve Maksimum Entropi algoritmaları "Duygu Analizi Aracı" uygulamamıza entegre edilmiştir. Bölüm 3.4.1'de yapısı ve çalışma şekli belirtilen "Kelime – Fiil tabanlı veri analizi" yapısı ile elde ettiğimiz mesaj analiz sonuçlarını makine öğrenmesi süreci ile elde edeceğimiz analiz sonuçları ile karşılaştırarak bizim için değerli olan mesajlardaki fikirleri en doğru bir şekilde analiz etmek amaçlanmaktadır.

3.4.2.1. Eğitim ve Test Veri Setleri

Makine öğrenimi eğitim veri setinin toplanması ile başlar ve sonra eğitim verileri doğrultusunda bir sınıflandırıcı oluşturmaktır (Vohra ve Teraiya, 2013). Bölüm 2.6'da belirtildiği gibi makine öğrenmesi tekniklerinde sınıflama algoritmaları kullanılır, bu algoritmaların kullanımda eğitim ve test olmak üzere iki veri seti yer alır. Eğitim veri setindeki verilerin hangi sınıfta olduğu bellidir, test veri seti de eğitim veri setinin bir bölümünden oluşturulmuş verilerdir, sınıf olma özelliklerinden elde edilecek sınıflama kuralları ile bu verilerin hangi sınıf içinde olmaları gerektiğine karar verilir. Bu işlem ile kurulan modelin doğruluğu test edilir.

Eğitim veri setinin içindeki verilerin sınıflandırılması belli bir süreç ile gerçekleşir. Bu süreç çoğunlukla otomatikleşmemiş aşamalar içerir;

- Nitel veriler ise, verilerin tek tek izlenmesi veya incelenmesi ile
- Nicel veriler ise, veriler üzerinde yapılan hesaplamaların sonuçları üzerinden karar verilmesi ile

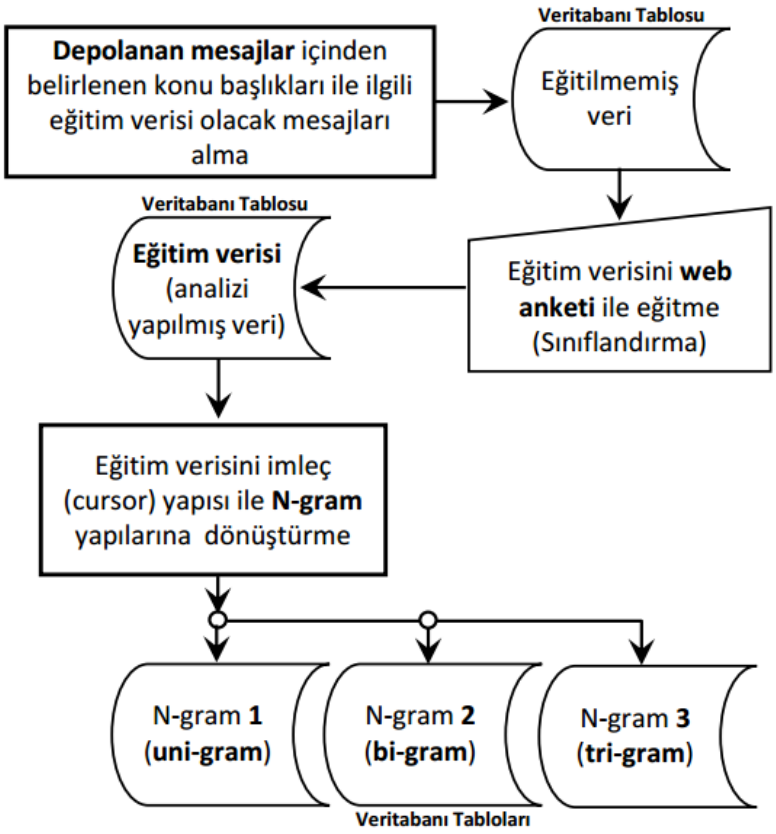
elde edilir. Bu elde edilen sonuçlar izlemeyi veya incelemeyi yapan veya kararı veren kişinin kişisel görüşü niteliğindedir. Bu durumda kişilerin veri ile ilgili olarak uzmanlaşmış olması doğru sınıfların oluşmasını sağlayacaktır.

Tez kapsamında da N-gram tabanlı veri analizi için Türkiye Büyük Millet Meclis'inde temsil edilen 4 siyasi partinin liderleri, Türkiye banka sektörünün önde gelen iki bankası, iki yerli teknoloji firması, Türkiye'de hizmet veren mobil operatörler ve halk tarafından yoğun olarak takip edilen üç yerli dizi ile ilgili 2015 ve 2016 yılında yazılan Twitter mesajlarından oluşan bir veri seti (eğitilmemiş veri) oluşturulmuştur (Tablo 3.9).

Tablo 3.9: N-gram veri analizi eğitim veri setinin konu başlıkları ve tweet sayıları.

Konu Başlığı	Tweet Sayısı
Siyasi Liderler	300
Banka-Finans	690
Yerli Teknoloji	208
Mobil Operatörler	312
Yerli Diziler	484

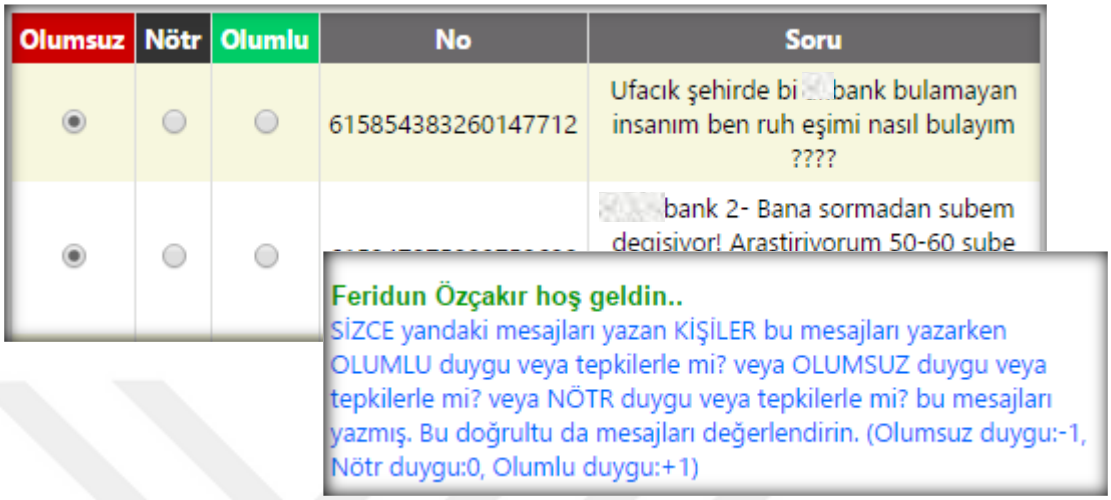
Şekil 3.11'de yer alan süreç doğrultusunda eğitilmemiş veri setini N-gram modeli için eğitim veri setine dönüştürmek amacı ile her bir tweet mesajı okunarak olumlu, olumsuz



Şekil 3.11: Eğitilmemiş veriden N-gram veri setinin elde edilme süreci.

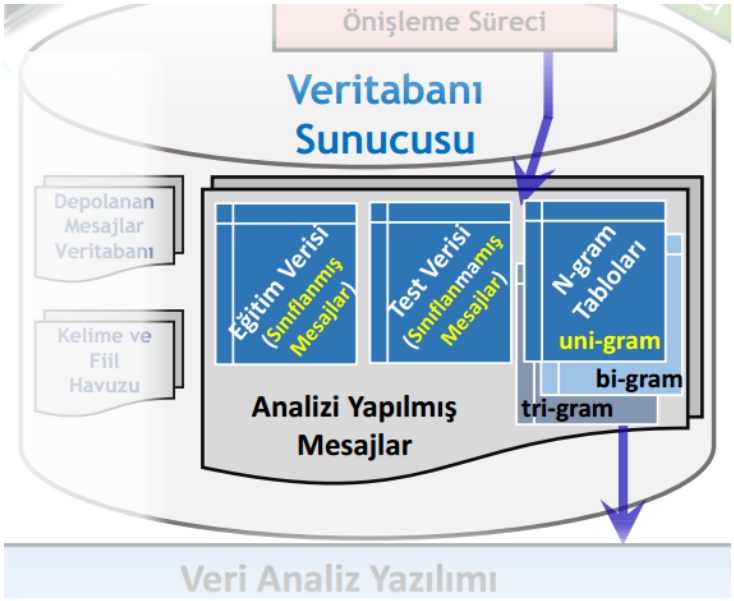
veya nötr olduğuna karar verilmiştir. Bu işlem gerçekleştirilirken belirlenen kelimeler üzerindeki duygu kutbu tespiti çalışmasına katılan lisans ve yüksek lisans öğrencilerinin görüşleri alınmıştır. Değerlendirmeye yapanlardan ilgili tweet mesajlarını yazan kişilerin bu mesajları yazarken hangi duygu (olumlu, olumsuz, nötr) veya tepki ile yazdıklarını

belirlemeleri istenmiştir. Değerlendiriciler bu işlemleri web sunucusu üzerinde geliştirilen interaktif (etkileşimli) bilgi toplama uygulaması ile gerçekleştirmişlerdir (Şekil 3.12).



Şekil 3.12: Eğitim verisinin değerlendirildiği anket.

Değerlendirilen bu mesajlar "N-gram" veritabanı içine aktarılmıştır (Şekil 3.13). Bu mesajlar N-gram modeli doğrultusunda 1 gram (unigram) , 2 gram (bigram) ve 3 gram (trigram) yapıları şeklinde geliştirilen yazılım aracılığı ile kelime seviyesinde ayrıştırılması ve gram seviyelerinde gruplamaları yapılmıştır.



Şekil 3.13: Eğitim - test verisi ve N-gram veritabanı.

Her gram boyutunu temsil eden veriler veritabanında farklı tablolarda tutulmuştur (Şekil 3.11), her gram satırında ilgili gramın eğitim verisi içinde toplam kaç adet olumlu,

olumsuz ve nötr mesaj içinde yer aldığı ayrı birer kolon (olumlu/olumsuz/nötr) olarak oluşturulmuştur. Bu işlem ile normal Twitter mikroblog sitesinden çekilecek anlık veya daha önce depolanmış mesajlar üzerinde analiz yapılmasını sağlayacak *gram seviyesindeki eğitim verileri* oluşturulmuştur.

Tablo 3.10: Farklı iki bankaya ait eğitilmiş örnek mesajlar (eğitim verisi).

Mesaj No	Tweet Mesajı	Sınıf
1	**XBank** Personel Alım İlanları	Nötr (O)
2	**XBank** Dünyada İlk 100e Girdi	Olumlu (P)
3	**XBank** 1000 Kişi Alacak	Olumlu (P)
4	Etkinlik Önerisi – **YBank** Sanat Atölyeleri	Nötr (O)
5	@**YBank**Destek #**YBank**DirektMobil haraç kesen banka **YBank**	Olumsuz (N)
6	@**YBank**Destek #**YBank**DirektMobil müşterisini dolandıran banka **YBank**	Olumsuz (N)
7	**XBank**'dan kaleleri fetheden kart	Olumlu (P)
8	Birisi **YBank**'a söyleyebilir mi, Babalar Günü geçti gitti. YouTube'dan çeksin reklamlarını.	Olumsuz (N)
9	Otomatik talimata hesap işletim ücretini iade ederiz yalanı ile 60 TL mi gasp ettin ya @**YBank** @**YBank**Destek Bi daha zor bulursun beni.!	Olumsuz (N)
10	Cazın özgün sesi Karima **YBank** Caz Günleri'nde	Olumlu (P)
11	@**YBank**Destek Maaş müsterisiyim.Sitedeki bilgilendirmede hesap işletim ücreti alınmayacağı sonucuna varılmakta. İade talep ediyorum. #**YBank**	Olumsuz (N)
12	@safaktas **YBank**Direkt şifre talebinizi 444YYYY Telefon Şubemizden ya da linkimizden iletebilirsiniz.	Nötr (O)
13	BİR BİREY OLARAK **YBANK** VE DİĞER BÜTÜN BÜYÜK SERMAYENİN ESİRİ OLMAYACAĞIM TÜM YASAL HAKLARIMI KULLANACAĞIM	Olumsuz (N)
14	@**YBank**Destek benim limiti niye yükseltmiyorsunuz sayın **YBank** :)	Olumsuz (N)
15	Paralarımı sevmeye geldim:) (@**YBank**)	Olumlu (P)
16	Brand Finance araştırmasına göre Türkiyenin En Değerli Markası **YBank**	Olumlu (P)
17	Bıktım yıldım !!! (@**YBank** in İstanbul)	Olumsuz (N)
18	3500 TL kredi için kefil isteyen bir banka varmı ya varmış o banka **XBank**.	Olumsuz (N)
19	#**YBank** reklam filminin maliyetini milletten kestigi 60 liralarla odemistir net!	Olumsuz (N)
20	@tgctbn **XBank** şubesinde göreve başlayacağım bi sorun çıkmazsa :)	Olumlu (P)

3.4.2.2. N-gram Yapısı ile Öğrenme Modeli Oluşturma

Tablo 3.10'da yer alan mesajlar Türkiye'de hizmet veren iki farklı bankaya ait müşterilerinin bankalar hakkında Twitter mikroblog sitesinde yayınladıkları mesajlardır.

Bu mesajlar ön işlemeden geçirilerek hangi duygu sınıfında olduğu belirlenerek eğitim verisi şekline dönüştürülmüştür. Bir sonraki aşama eğitim verisinin gramlara ayrıştırılmasıdır, bu doğrultuda Tablo 3.10'da yer alan mesajlar verisi Şekil 3.11'de yer alan yapı ile 1-gram (unigram) seviyesinde ayrıştırılır. Bu ayrıştırma sonucunda elde edilen kelimeler bir kelime torbasını oluştururlar. Bu kelimelerin tümü işlem öncesi küçük harfe dönüştürülebilir veya dönüştürmeden de devam edilebilir. Dönüşüm işlemlerin daha rahat yapılmasını sağlayacaktır.

Tablo 3.11: Eğitim verisine göre 1-gram (unigram) kelime torbası.

1-gram (uni-gram)	**Tekrar Sayısı (frekans)**	**1-gram (uni-gram)**	**Tekrar Sayısı (frekans)**
ybank	12	bi	2
xbank	6	bir	2
ybankdestek	5	hesap	2
banka	4	iade	2
ya	3	işletim	2
:)	3	mi	2
ybankdirektmobil	2	TL	2

Kelime torbası içindeki terimler (kelimeler veya eğer 2-gram ve üstü çalışma yapıyorsak kelime grupları) eğitim veri seti içinde yer alma sıklıkları (terim frekansları -$tf_{t,d}$) dikkate alınarak sıralanır. Kelime torbasının büyüklüğü eğitim verisinin büyüklüğü ile doğru orantılıdır, her mesajın birden fazla kelimeden oluştuğu düşünülürse çok fazla öğe kelime torbasında yer alacaktır. Tüm öğeleri kullanarak analiz yapılmak istenirse süreç uzun olacaktır, bu neden ile veri analizini gerçekleştiren uzman tecrübesi ile kelime torbasından en sık kullanılan belli bir sayıdaki kelime ile süreci devam edilmesini sağlayacak kısıtlamaya gidebilir. Tablo 3.10'da yer alan eğitim verisi setinde 136 farklı kelime (1-gram) bulunmaktadır, Tablo 3.11'de ise bu eğitim veri setinde en sık kullanılan ilk 14 kelimenin (1-gram) oluşturduğu kelime torbası görülmektedir.

Kelime torbası içinde yer alan kelimelerin (kelime gruplarının) hangi mesajlarda içinde yer aldığını gösteren vektör gösterim Tablo 3.12'de yer almaktadır.

Tablo 3.12: Eğitim verisine göre 1-gramların vektör gösterimi.

Mesaj No	Sınıf	ybank	xbank	ybankdestek	banka	ya	:)	yabankdirektmobil	bi	bir	hesap	iade	işletim	mi	TL
1	Nötr	0	1	0	0	0	0	0	0	0	0	0	0	0	0
2	Olumlu	0	1	0	0	0	0	0	0	0	0	0	0	0	0
3	Olumlu	0	1	0	0	0	0	0	0	0	0	0	0	0	0
4	Nötr	1	0	0	0	0	0	0	0	0	0	0	0	0	0
5	Olumsuz	1	0	1	1	0	0	1	0	0	0	0	0	0	0
6	Olumsuz	0	0	1	1	0	0	1	0	0	0	0	0	0	0
7	Olumlu	0	1	0	0	0	0	0	0	0	0	0	0	0	0
8	Olumsuz	1	0	0	0	0	0	0	0	0	0	0	0	1	0
9	Olumsuz	1	0	1	0	1	0	0	1	0	1	1	1	1	1
10	Olumlu	1	0	0	0	0	0	0	0	0	0	0	0	0	0
11	Olumsuz	0	0	1	0	0	0	0	0	0	1	1	1	0	0
12	Nötr	1	0	0	0	1	0	0	0	0	0	0	0	0	0
13	Olumsuz	1	0	0	0	0	0	0	0	1	0	0	0	0	0
14	Olumsuz	1	0	1	0	0	1	0	0	0	0	0	0	0	0
15	Olumlu	1	0	0	0	0	1	0	0	0	0	0	0	0	0
16	Olumlu	1	0	0	0	0	0	0	0	0	0	0	0	0	0
17	Olumsuz	1	0	0	0	0	0	0	0	0	0	0	0	0	0
18	Olumsuz	0	1	0	2	1	0	0	0	1	0	0	0	0	1
19	Olumsuz	1	0	0	0	0	0	0	0	0	0	0	0	0	0
20	Olumlu	0	1	0	0	0	1	0	1	0	0	0	0	0	0

Tablo 3.12'de her kelimenin (1-gramların) mesajlarda bulunma değerleri bulunmakta, eğer bir kelime bir mesajda birden fazla tekrarlanıyorsa (18 nolu mesajda "banka" kelimesi gibi) tekrar sayısı vektöre aktarılır.

Her bir gram (kelime – kelime grupları) eğitim verisindeki mesajların sınıfları doğrultusunda her sınıfta (olumlu, olumsuz, nötr) bulunma sayısı hesaplanır (Tablo 3.13).

Tablo 3.13: 1-gramların (unigram) sınıflarda (olumlu, olumsuz, nötr) bulunma sayıları.

1-gram (uni-gram)	Olumlu (P)	Olumsuz (N)	Nötr (O)	Toplam
ybank	3	9	1	13
xbank	4	1	1	6
ybankdestek	0	5	0	5
banka	0	4	0	4
ya	0	2	1	3
:)	2	1	0	3
ybankdirektmobil	0	2	0	2
bi	1	1	0	2
bir	0	2	0	2
hesap	0	2	0	2
iade	0	2	0	2
işletim	0	2	0	2
mi	0	2	0	2
TL	0	2	0	2

Tüm eğitim verisindeki mesajların duygu sınıflarına göre dağılımı, örnek verimiz doğrultusunda hesaplamış değerler Tablo 3.14 görülmektedir.

Tablo 3.14: Eğitim verisindeki mesajların duygu sınıf dağılımları.

	Olumlu Sınıfı	Olumsuz Sınıfı	Nötr Sınıfı	Toplam Mesaj Sayısı
Sınıftaki Mesaj Sayısı	7	10	3	20

1-gram niteliğindeki kelime torbasındaki kelimelerin sınıflarda bulunma değerlerinden ilgili kelimelerin her duygu sınıfını temsil etme olasılığı o sınıfta yer alan toplam kelime sayısına oranı ile hesaplanır (Tablo 3.15). Kelimelerin bu olasılık değerleri ve genel sınıf dağılım değerleri aracılığı ile yeni bir tweet mesajında bu kelimelerin yer alması durumunda yeni mesajın hangi duygu sınıfında olduğu farklı sınıflandırma teknikleri ile tahmin edilmektedir.

Tablo 3.15: 1-gramların (unigram) sınıflarda (olumlu, olumsuz, nötr) bulunma olasılıkları.

	Olumlu Sınıfı		**Olumsuz Sınıfı**		**Nötr Sınıfı**	
Sınıftaki Toplam Kelime Sayısı	**44**		**109**		**20**	
1-gram (uni-gram)	**Kelimenin Frekansı**	**P(Olumlu)**	**Kelimenin Frekansı**	**P(Olumsuz)**	**Kelimenin Frekansı**	**P(Nötr)**
ybank	3	3/44	9	9/109	1	2/20
xbank	4	4/44	1	1/109	1	1/20
ybankdestek	0	0/44	5	5/109	0	0/20
banka	0	0/44	4	4/109	0	0/20
ya	0	0/44	2	2/109	1	1/20
:)	2	2/44	1	1/109	0	0/20
ybankdirektmobil	0	0/44	2	2/109	0	0/20
bi	1	1/44	1	1/109	0	0/20
bir	0	0/44	2	2/109	0	0/20
hesap	0	0/44	2	2/109	0	0/20
iade	0	0/44	2	2/109	0	0/20
işletim	0	0/44	2	2/109	0	0/20
mi	0	0/44	2	2/109	0	0/20
TL	0	0/44	2	2/109	0	0/20

3.4.3. Naïve Bayes Sınıflandırma Modeli

Naïve Bayes metin sınıflamayı sağlayan basit bir modeldir (Niu ve diğ., 2012). Bir metin içindeki duyguyu geliştirilen bilgisayar yazılımları ile analiz etmek için o metin ile anlamsal benzerlik içeren metinler üzerinde yapılan analizlerden yola çıkarak olasılık hesaplamaları ile sonuca ulaşılmaktadır. Geçmiş ve düzenlenmiş veriden yola çıkıp, mevcut verinin durumu tahmin edilmektedir, bu süreç makine öğrenmesidir.

Analiz uygulamamız ile Twitter mikroblog sitesinden yeni alınan veya veritabanına daha önceden depolanmış tweet mesajlarının hangi duygu sınıfı (olumlu, olumsuz, nötr) içinde yer alacakları Naïve Bayes sınıflandırıcı ile tahmin edilmektedir. Bölüm 3.4.2 başlığındaki süreçle elde edilen eğitim verisi ve bu eğitim verisi üzerinden elde edilen N-gram yapıları sınıflandırma algoritmalarındaki öğrenme modelinin en önemli parçalarıdır. Sınıflandırma esnasında, duygu analizi yapılacak veri niteliğindeki mikroblog sitesinden alınan mesajlarda 1-gram, 2-gram veya 3-gram parçalarının bulunması doğrultusunda bu parçaların eğitim verisi üzerinde bulunma olasılık değerlerini kullanarak mesajlarının olumlu/olumsuz/nötr duygu sınıflarında bulunma olasılıkları hesaplanmaktadır.

1-gram yapısında her bir kelime tek olacağından kelimelerin sıralanışı,

$$t_{1-gram} = \{w_1, w_2, \dots, w_{n-1}, w_n\} \tag{3.2}$$

şeklinde ifade edilir. w gram seviyesindeki kelimeleri simgeler. Öğrenme modelimizde yer alan eğitim veri seti doğrultusunda 1-gram seviyesindeki terimlerimiz aşağıdaki yapıdadır.

$t_{1-gram} = \{ybank,\ xbank,\ ybankdestek,\ banka,\ ya,\ :),\ ybankdirektmobil,\ bi,\ bir,\ hesap,\ iade,\ işletim,\ mi,\ TL\}$

2-gram olduğunda ise terimlerin sıralanışı,

$$t_{2-gram} = \{(w_1w_2), (w_2w_3) \dots, (w_{n-1}w_n)\} \tag{3.3}$$

şeklinde ifade edilir. Öğrenme modelimizde yer alan eğitim seti doğrultusunda 2-gram seviyesindeki terimlerimiz ise aşağıdaki yapıda bulunurlar.

$t_{2-gram} = \{(ybank - xbank), (xbank - ybankdestek),\ (ybankdestek - banka),$
$(banka - ya), (ya - :)), (:), ybankdirektmobil),\ (ybankdirektmobil, bi),$
$(bi, bir), (bir, hesap), (hesap, iade), (iade, işletim), (işletim, mi), (mi, TL)\}$

Bölüm 2.6.3 de yer alan bayes eşitliğini (2.2) Twitter mesajları üzerinde duygu analizi çalışmamıza göre düzenlediğimizde,

$$P(c|\boldsymbol{d}) = \frac{P(c)p(\boldsymbol{d}|c)}{p(\boldsymbol{d})} \tag{3.4}$$

Burada c her bir duygu sınıfını-kutbunu (olumlu, olumsuz, nötr) temsil etmekte, $\boldsymbol{d} = \{d_1, d_2, .., d_n\}$ şeklinde twitter mesajlarını ifade etmektedir.

$P(c|\boldsymbol{d})$, duygu kutuplarının sonsal (posterior) olasılığı (mesaja bağlı olarak),

$P(c)$, duygu kutuplarının önsel (prior) olasılığı,

$P(\boldsymbol{d})$, mesajın önsel olasılığı,

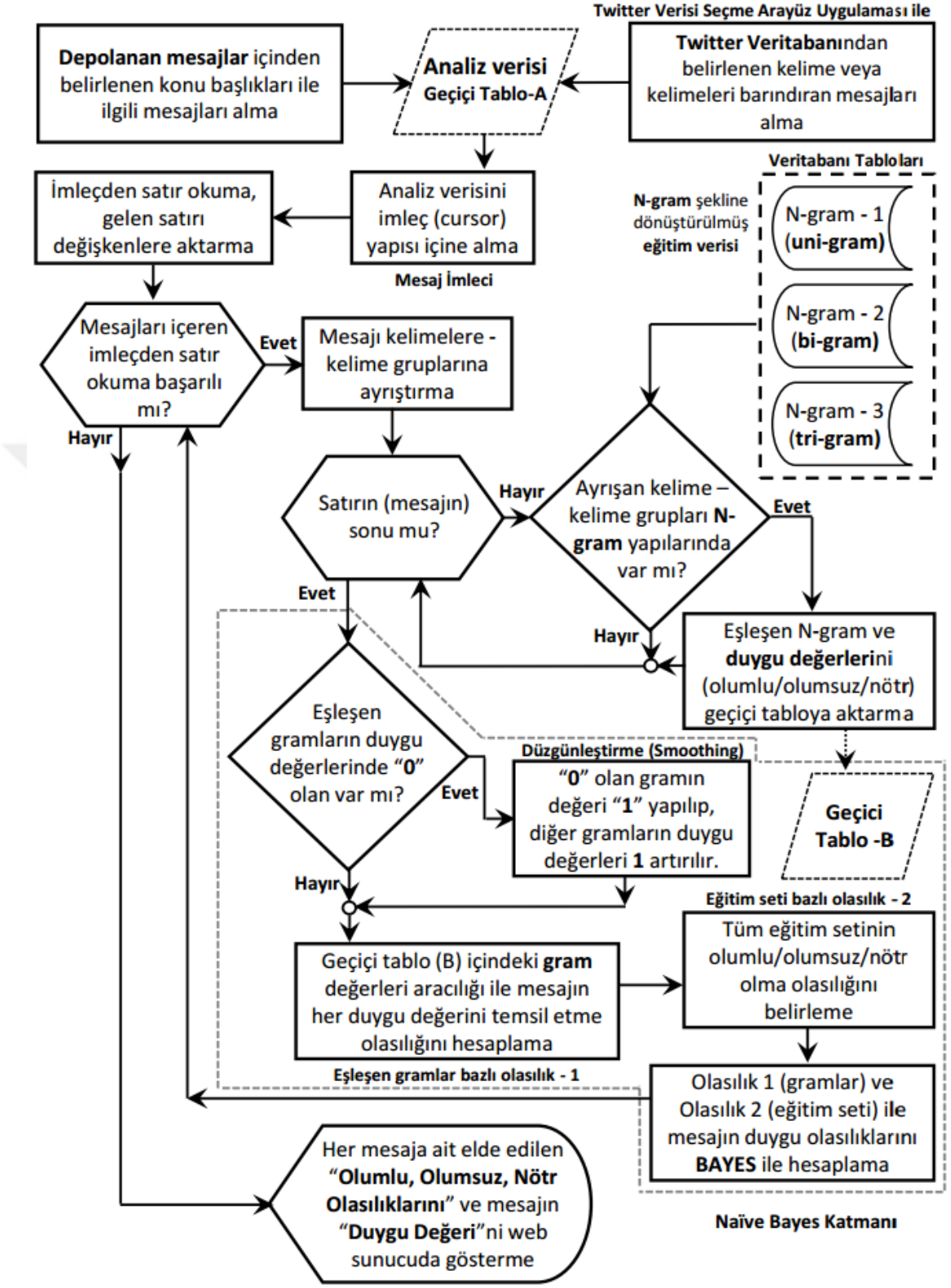
$P(\boldsymbol{d}|c)$, mesajın verilen duygu kutbuna göre olabilirlik (likelihood) olasılığıdır. (Neapolitan, 1990)

Tablo 3.16'da yer alan mesajların Türkiye'de hizmet veren iki farklı banka hakkında Twitter mikroblog sitesinde yayınlanan mesajlardır.

Tablo 3.16: Naïve Bayes ile sınıflandırılacak örnek tweet mesajları.

Mesaj No	Tweet Mesajı
$\boldsymbol{d_1}$	Kredimi eksik alacaksınız çünkü üstünü ödemeyeceğim yasal olmayan hesap işletim ücretini iade edip tamamlamanız gerekiyor. Bilginize. @YBank.
$\boldsymbol{d_2}$	Sıcaktan bunalınca XBank'a girer, sıra alır serinlerim:) Sıra gelmiyor nasıl olsa. Gerçi gelse ne yapacağım o da ayrı bi ironi tabi, serinliyoruz :)

Şekil 3.14'de tez kapsamında oluşturulan N-gram tabanlı duygu analizinde Naïve Bayes sınıflandırıcı modelinin kullanımı yer almaktadır.



Şekil 3.14: Naïve Bayes sınıflandırıcı modeli.

Tablo 3.16'da yer alan mesajlar duygu analizi açısından sınıflandırıldığında üç farklı sınıfta bulunma ihtimalleri vardır. Naïve Bayes ile sınıflandırma yaparken ilk olarak Şekil

3.15'de belirtildiği gibi bu mesajların çalışılacak gram seviyesinde ayrıştırılması ve daha sonra da veritabanında yer alan gram verileri karşılaştırılmaları gerekir. Bölüm 3.4.2.2'de yer alan örnek verimiz üzerinden bu mesajlar için sınıflama işlemi gerçekleştirilirse, 1-gram seviyesine indirgenmiş eğitim verisi ile bu mesajların kelimeleri ile karşılaştırıldığında,

- $\boldsymbol{d_1}$ için "ybank", "hesap", "iade", "işletim" kelimeleri,
- $\boldsymbol{d_2}$ için "xbank", "bi", ":)" kelimeleri 1-gram verileri ile eşleşmektedir.

Naïve Bayes yöntemi metin sınıflandırılmasında genellikle hız ve basitlik nedeniyle metin sınıflandırılmasında kullanılır. Bu yöntemde özellik (örneğin her bir gram) olarak kullanılan yapılar her biri birbirinden bağımsızdır, Twitter mesajlarında da her kelimenin (ya da N-gramın) durumu bağımsız olarak dikkate alınır. Bayes teoreminden yola çıkarak belirlenen her sınıfta verinin olma olasılığı aşağıdaki şekilde formüle edilir (Manning ve diğ., 2009);

$$P(c|\boldsymbol{d}) = P(c)(\prod_{1\leq k\leq n_d} P(t_k|c)) \tag{3.5}$$

Burada $P(c)$ duygu sınıflarında olma olasılığı, $\boldsymbol{d}$ mesajlar, $\boldsymbol{t}$ ise her bir terimdir (kelime veya kelime grupları, 1-gram veya n-gram), $\boldsymbol{k}$ terim indeksi, $\boldsymbol{n_d}$ ise mesajdaki toplam terim sayısıdır. Metin sınıflandırma da amacımız metin için en iyi sınıfı bulmaktır. Naïve Bayes sınıflandırıcısı da mesaj için en yüksek olasılıklı (maximum a posterior – MAP) sınıf değerini üretir.

$$c_{MAP} = \underset{c\in \mathbb{C}}{\operatorname{argmax}}\, \hat{P}(c|d) = \underset{c\in \mathbb{C}}{\operatorname{argmax}}\, \hat{P}(c) \prod_{1\leq k\leq n_d} \hat{P}(t_k|c) \tag{3.6}$$

Yukarıdaki eşitlikte $\boldsymbol{c_{MAP}}$ en yüksek olasılıklı sınıfı, $\mathbb{C}$, ifadesi ise tüm sınıfları simgelemektedir, $P(c)$ ve $P(t_k|c)$ bilinmediğinden olasılık ifadesi P, $\hat{P}$ olarak belirtilmiştir. Eşitlikten anlaşılacağı gibi her terim ($1 \leq k \leq n_d$) için bir çok koşullu olasılığın çarpımı ile sonuç elde edilmektedir. Çok fazla terim olduğunda bu çarpma işlemleri ondalıklı değer taşmasına (floating point underflow) neden olabilir, olasılıkların logaritmalarını hesaplamaya ekleyerek hesaplamayı gerçekleştirmek bu problemi ortadan kaldırır.

$$c_{MAP} = \underset{c \in C}{\text{argmax}}[\log \hat{P}(c) + \sum_{1 \le k \le n_d} \log \hat{P}(t_k|c)] \quad (3.7)$$

Her koşullu parametre $\log \hat{P}(t_k|c)$, t_k'nın c sınıfı için iyi bir gösterge olması için bir ağırlıktır. Benzer şekilde, $\log \hat{P}(c)$'de c sınıfının frekansını gösteren bir ağırlıktır. Önsel olasılıkların logoritmalarının toplamı ve terimlerin (kelime veya kelime gruplarının) ağırlıkları mesajın hangi sınıf içinde yer alacağına dair birer ölçüdür.

Eğitim verisi içinde ayrıştırılmış sınıfların toplam veri setine oranları aşağıdaki eşitlikle ifade edilir.

$$\hat{P}(c) = \frac{N_c}{N} \quad (3.8)$$

N_c, c ile ifade edilen duygu sınıfında yer alan toplam mesaj sayısını, N ise eğitim seti içindeki toplam mesaj sayısını simgelemektedir.

Bölüm 3.4.2.2'deki öğrenme modelinde yer alan eğitim verisi üzerinde eşitlik (3.8) uygulanırsa, Tablo 3.17'deki olasılık değerleri elde edilir.

Tablo 3.16'da yer alan her mesaj için eşleşen kelimeler doğrultusunda her duygu sınıfı için aşağıdaki eşitlikler yazılır.

Tablo 3.17: Eğitim verisi doğrultusunda duygu sınıf olasılıkları.

	Olumlu Sınıfı	**Olumsuz Sınıfı**	**Nötr Sınıfı**	**Toplam Mesaj Sayısı**
Sınıftaki Mesaj Sayıları	7	10	3	20
	P(Olumlu)	**P(Olumsuz)**	**P(Nötr)**	
(Sınıftaki Mesaj Sayısı) / (Toplam Mesaj Sayısı)	7 / 20	10 / 20	3 / 20	$\hat{P}(c) = \frac{N_c}{N}$
Yüzde Gösterimi	%35	%50	%15	

İlk olarak $\boldsymbol{d_1}$ mesajının olumlu (**poz**itif) duygu sınıfı için;

$$P(poz|\boldsymbol{d_1}) = Ppoz\ (\prod_{1 \le k \le 4} P(t_k|poz)) \quad (3.9)$$

$$P(poz|\boldsymbol{d_1}) = Ppoz * (Ppoz_{ybank} * Ppoz_{hesap} * Ppoz_{iade} * Ppoz_{işletim}) \quad (3.10)$$

"ybank", "hesap", "iade" ve "işletim" kelimelerinin olumlu sınıfında olma olasılık değerleri Tablo 3.15'den ve eğitim verisinin tamamında olumlu olma olasılığı ($Ppoz$) Tablo 3.17'den alınır.

$$P(poz|\boldsymbol{d_1}) = Ppoz * (Ppoz_{ybank} * Ppoz_{hesap} * Ppoz_{iade} * Ppoz_{işletim})$$

$$P(poz|\boldsymbol{d_1}) = \frac{7}{20} * \left(\frac{3}{44} * \frac{0}{44} * \frac{0}{44} * \frac{0}{44}\right) = 0$$

Olumsuz (**neg**atif) duygu sınıfı için;

$$P(neg|\boldsymbol{d_1}) = Pneg * \left(Pneg_{ybank} * Pneg_{hesap} * Pneg_{iade} * Pneg_{işletim}\right) \quad (3.11)$$

"ybank", "hesap", "iade" ve "işletim" kelimelerinin olumsuz sınıfında olma olasılık değerleri Tablo 3.15'den ve eğitim verisinin tamamında olumsuz olma olasılığı ($Pneg$) Tablo 3.17'den alınır.

$$P(neg|\boldsymbol{d_1}) = \frac{10}{20} * \left(\frac{9}{109} * \frac{2}{109} * \frac{2}{109} * \frac{2}{109}\right)$$

$$P(neg|\boldsymbol{d_1}) = \frac{720}{2823163220} = 0{,}000000255$$

Nötr (**nötr**) duygu sınıfı için;

$$P(nötr|\boldsymbol{d_1}) = Pnötr * (Pnötr_{ybank} * Pnötr_{hesap} * Pnötr_{iade} * Pnötr_{işletim}) \quad (3.12)$$

"ybank", "hesap", "iade" ve "işletim" kelimelerinin nötr sınıfında olma olasılık değerleri Tablo 3.15'den ve eğitim verisinin tamamında nötr olma olasılığı ($Pnötr$) Tablo 3.17'den alınır.

$$P(nötr|\boldsymbol{d_1}) = \frac{3}{20} * \left(\frac{1}{20} * \frac{0}{20} * \frac{0}{20} * \frac{0}{20}\right) = 0$$

Mesajın her duygu sınıfı için hesaplanan olasılık değerleri doğrultusunda mesajın hangi duyguyu içerdiğine karar verilir. $P(poz|\boldsymbol{d_1}) = 0$, $P(\text{neg}|\boldsymbol{d_1}) = 0{,}000000255$ ve $P(nötr|\boldsymbol{d_1}) = 0$ olduğuna göre $\boldsymbol{d_1}$.mesajı "**olumsuz**" duygu içermektedir.

$\boldsymbol{d_2}$ mesaj için aynı işlemleri gerçekleştirirsek (eşleşen kelimeler: "xbank", "bi", ":)");

$$P(poz|\boldsymbol{d_2}) = Ppoz * (Ppoz_{xbank} * Ppoz_{bi} * Ppoz_{:)})$$

$$P(poz|\boldsymbol{d_2}) = \frac{7}{20} * \left(\frac{4}{44} * \frac{1}{44} * \frac{2}{44}\right) = \frac{56}{1703680} = 0{,}00003287$$

$$P(neg|\boldsymbol{d_2}) = Pneg * (Pneg_{xbank} * Pneg_{bi} * Pneg_{:)})$$

$$P(neg|\boldsymbol{d_2}) = \frac{10}{20} * \left(\frac{1}{109} * \frac{1}{109} * \frac{1}{109}\right) = \frac{1}{2590058} = 0{,}00000039$$

$$P(nötr|\boldsymbol{d_2}) = Pnötr * (Pnötr_{xbank} * Pnötr_{bi} * Pnötr_{:)})$$

$$P(nötr|\boldsymbol{d_2}) = \frac{3}{20} * \left(\frac{1}{20} * \frac{0}{20} * \frac{0}{20}\right) = 0$$

Yukarıdaki sonuçlara göre $\boldsymbol{d_2}$ mesajı "olumlu" duygu içermektedir.

3.4.3.1. Çok Terimli (Multinomial) Model

Bölüm 3.4.3'de başlığı altında yapılan işlemlerde $\boldsymbol{d_2}$ mesajında iki adet " :) " özel karakteri yer almaktadır, fakat Naïve Bayes yöntemi ile hesaplama esnasında bu dikkate alınmamıştır. Naïve Bayes yönteminin bu şekilde kullanımında terimlerin (kelimelerin , n-gramların) analizi yapılan mesaj içinde bulunma sıklığının (tekrarlanma sayısının – frekansının) bir önemi yoktur. Bir terimin duygu sınıflarına ayrılmış eğitim setinde ilgili sınıfa ait mesajlar içindeki varlığı veya yokluğu iki değer ile gösterilir, bu değerler 0 (yok) ve 1 (var)'dir. Bu modele bölüm 2.6.4'e belirtildiği gibi **Çok Değişkenli Bernoulli Modeli** adı verilir.

Çok terimli (multinomial) modelinde ise Bernoulli modelinde farklı olarak kelime frekansları dikkate alınarak hesaplama yapılır. $\boldsymbol{d_2}$ mesajının duygu sınıflarını bu doğrultuda tekrar hesaplayalım.

$$P(poz|\boldsymbol{d_2}) = Ppoz * (Ppoz_{xbank} * Ppoz_{bi} * Ppoz_{:)}^2)$$

$$P(poz|\boldsymbol{d_2}) = \frac{7}{20} * \left(\frac{4}{44} * \frac{1}{44} * \left(\frac{2}{44}\right)^2\right) = \frac{112}{74961920} = 0{,}0000014941$$

$$P(neg|\boldsymbol{d_2}) = Pneg * (Pneg_{xbank} * Pneg_{bi} * Pneg_{:)}^2)$$

$$P(neg|\boldsymbol{d_2}) = \frac{10}{20} * \left(\frac{1}{109} * \frac{1}{109} * \left(\frac{1}{109}\right)^2\right) = \frac{1}{282316322} = 0{,}0000000035$$

$$P(nötr|\boldsymbol{d_2}) = Pnötr * (Pnötr_{xbank} * Pnötr_{bi} * Pnötr_{:)}^2)$$

$$P(nötr|\boldsymbol{d_2}) = \frac{3}{20} * \left(\frac{1}{20} * \frac{0}{20} * \left(\frac{0}{20}\right)^2\right) = 0$$

Yukarıdaki sonuçlara göre $\boldsymbol{d_2}$ mesajı Bernoulli modeline göre daha güçlü bir değer ile "olumlu" duygu içerdiği tahmin edilmektedir. Tez kapsamında geliştirilen "N-Gram Veri Analizi" yapısında "Çok Terimli (Multinomial) Model" kullanılmıştır.

3.4.3.2. Düzgünleştirme (Smoothing)

Zhai ve Lafferty (2004)'ye göre, metin ve dil modelleme de en yüksek olabilirliğin (maximum likelihood) tahmin edilmesinde düzgünleştirme son derece önemlidir. Kelimeler metin belgeleri üzerinden tahminleme yapılırken yöntemin çalışmasındaki kritik öğelerdir, bir metin içinde hiç görünmeyen veya az sıklıkla yer alan bir kelime tahmin işlemi esnasında parametre olarak yer aldığı durumlarda olasılık değeri sıfır olabilir, sıfır olasılık değeri de tahminleme işleminde tutarlı sonuçlar elde edilmesini etkiler. Olasılık düzgünleştirmesi eğitim verilerinde görünmeyen özelliklerin olasılık değerlerinin sıfır olmasına izin vermeyen modelleme tekniğidir. Düzgünleştirme seyrek veri sorunun üstesinden gelen bir yöntemdir. Düzgünleştirme etkisi özellikle dil modelleme çalışmalarında performansı ve tutarlılığı olumlu yönde etkilemektedir. **Doğrusal ara değer bulma** (*linear interpolation smoothing*) ve **Dirichlet düzgünleştirmesi** gibi farklı düzgünleştirme yöntemleri kullanılmaktadır.

En yüksek olabilirlik (maximum likelihood) tahmini aşağıdaki şekilde formüle edilmektedir (düzgünleştirme yapılmamış yalın hali ile);

$$P_{ML}(t|c) = T_{ct} \;/\; \sum_{t'} T_{ct'} \tag{3.13}$$

T_{ct} , t teriminin eğitim veri seti içindeki c ile ifade edilen duygu sınıfları içinde bulunma sayısıdır.

En yüksek olabilirlik tahmininin doğrusal ara değer bulma yöntemi ile düzgünleştirmesi;

$$P_{LI}(t|c) = \lambda P_{ML}(t|c) + (1-\lambda)\, P_{ML}(t|C) \tag{3.14}$$

Yukarıdaki eşitlikte C ifadesi tüm eğitim veri setini simgeler. λ (lambda) ise ayırt edici niteliğinde bir parametredir ve λ değeri $0 < \lambda < 1$ aralığında olmalıdır.

En yüksek olabilirlik tahmininin Dirichlet yöntemi ile düzgünleştirmesi;

$$P_{Di}(t|c) = \left(T_{ct} + \mu P_{ML}(t|C)\right) / \left((\sum_{t'} T_{ct'}) + \mu\right) \quad (3.15)$$

μ bir parametredir ve $\mu > 0$ olmalıdır.

3.4.3.3. Laplace Düzgünleştirmesi - (Laplace Smoothing)

Bölüm 3.4.3 ve 3.4.3.1 başlıkları altında yapılan işlemlerde mesajların duygu sınıflarından hangisinde olduğunu hesaplanırken, mesajların bazı sınıflarda yer alma olasılık değerleri 0 (sıfır) çıkmıştır. Naïve Bayes yönteminde mesaj içinde yer alan terimlerden N-gram tabloları ile eşleşen her terimin duygu sınıflarına ayrıştırılmış eğitim verisi içinde ilgili sınıfa ait mesajlarda olma olasılık değerlerinin çarpımları elde edilen sonuç mesajın ilgili sınıfta bulunma olasılığını hesaplar. Bu hesaplama da bir terimin bile ilgili sınıfa ait mesajlarda yer almaması işlemin bir çarpma işlemi olmasında dolayı toplam sonucun 0 çıkmasına neden olur.

Manning ve diğ. (2009)'ne göre, Laplace düzgünleştirmesi, metinler üzerindeki tahminlemede çoğunlukla tercih edilen yöntemdir. Bu yöntem ile sıfır değer üreten olasılık sonuçları her işleme giren kelimenin metinde görünme sıklık değerine bir (1) eklenerek bertaraf edilir. En yüksek olabilirlik tahmininin Laplace yöntemi ile düzgünleştirmesi aşağıdaki şekilde formüle edilir;

$$P_{LS}(t|c) = (T_{ct} + 1) / \sum_{t'}(T_{ct'} + 1) = (T_{ct} + 1)/((\sum_{t'} T_{ct'}) + V) \quad (3.16)$$

V ifadesi eğitim veri setindeki mesajlar içinde yer alan tüm terimlerin sayısını ifade etmektedir.

Çoklu terimli Naïve Bayes yöntemi ile Bölüm 3.4.3'deki hesaplamalarda $\boldsymbol{d_1}$ mesajının olumlu (pozitif) ve nötr sınıflarında olma ihtimali 0 değerini vermişti.

$$P(poz|\boldsymbol{d_1}) = Ppoz * (Ppoz_{ybank} * Ppoz_{hesap} * Ppoz_{iade} * Ppoz_{işletim})$$

$$P(poz|\boldsymbol{d_1}) = \frac{7}{20} * \left(\frac{3}{44} * \frac{0}{44} * \frac{0}{44} * \frac{0}{44}\right) = 0$$

$$P(nötr|\boldsymbol{d_1}) = Pnötr * (Pnötr_{ybank} * Pnötr_{hesap} * Pnötr_{iade} * Pnötr_{işletim})$$

$$P(nötr|\boldsymbol{d_1}) = \frac{3}{20} * \left(\frac{1}{20} * \frac{0}{20} * \frac{0}{20} * \frac{0}{20}\right) = 0$$

Yukarıdaki sonuçlar $\boldsymbol{d_1}$ mesajının hangi duygu sınıfında yer aldığını tahminleme işleminde tutarlı sonuçlar elde edilmesini etkilemiştir. $\boldsymbol{d_1}$ mesajının hangi duygu sınıfında yer aldığını Laplace düzgünleştirme ile tekrar gerçekleştirelim. 0 olasılık değerine neden olan terimler "hesap", "iade" ve "işletim" dir. Bu terimlerin 0 olasılık değeri üretmemesi için olasılık hesabındaki tüm pay değerlerine "1" değeri eklenir. Paydada ise hesaplamanın yapıldığı sınıfda yer alan mesajlar içinde yer alan toplam terim sayısı yer almaktadır, düzgünleştirme için bu değere tüm eğitim veri setinde yer alan terim sayısı (tekil olarak hesaplanmış terim sayısı) eklenir. Bölüm 3.4.2.2 Tablo 3.10'da yer alan eğitim veri setindeki tekilleştirilmiş toplam terim sayısı 136 dır. Bu işlem herhangi bir mesajın sınıf olasılık değeri 0 çıktığında kullanılır, fakat ilgili mesajın sadece 0 olasılık üreten sınıf olasılık hesaplamalarında değil, 0 dan farklı değer üreten hesaplamalarında da kullanılması gerekir. Bu doğrultuda $\boldsymbol{d_1}$ mesajının duygu sınıfı olasılıklarının hesaplanmasında Laplace düzgünleştirmesi aşağıda yer almaktadır.

$$P(poz|\boldsymbol{d_1}) = \frac{7}{20} * \left(\frac{3+\mathbf{1}}{44+\mathbf{136}} * \frac{0+\mathbf{1}}{44+\mathbf{136}} * \frac{0+\mathbf{1}}{44+\mathbf{136}} * \frac{0+\mathbf{1}}{44+\mathbf{136}}\right) = 0{,}0000000013$$

$$P(neg|\boldsymbol{d_1}) = \frac{10}{20} * \left(\frac{9+\mathbf{1}}{109+\mathbf{136}} * \frac{2+\mathbf{1}}{109+\mathbf{136}} * \frac{2+\mathbf{1}}{109+\mathbf{136}} * \frac{2+\mathbf{1}}{109+\mathbf{136}}\right) = 0{,}0000000375$$

$$P(nötr|\boldsymbol{d_1}) = \frac{3}{20} * \left(\frac{1+\mathbf{1}}{20+\mathbf{136}} * \frac{0+\mathbf{1}}{20+\mathbf{136}} * \frac{0+\mathbf{1}}{20+\mathbf{136}} * \frac{0+\mathbf{1}}{20+\mathbf{136}}\right) = 0{,}0000000005$$

Yukarıdaki sonuçlara göre $\boldsymbol{d_1}$ mesajı "olumsuz" duygu içermektedir. Bazen bu çıkan sonuç Laplace düzgünleştirme uygulanmamış hesaplamada çıkan sınıftan farklı olur, çünkü Laplace düzgünleştirmesi olmadan yapılan hesaplamada tüm terimlerin güçlü olasılık değerine sahip olmasına rağmen sadece bir terim bile sonucun 0 çıkmasına neden olmaktadır. Eğitim veri seti büyük olduğunda terim sayısı da büyüyecektir, yukarıdaki hesaplamalarda da görüldüğü gibi sınıf içindeki ve tüm veri setindeki terimlerin sayıları payda da yer aldığında eğitim seti büyüdüğünde payda değeri de büyüyecek, sonuç değeri çok küçük olacaktır.

Hesaplamalarda da görüldüğü gibi kısıtlı bir eğitim setinde bile çıkan sonuç değerleri sayısal olarak çok küçüktür, çok fazla veride bu sonuç üzerinde ondalıklı değer taşması (floating point underflow) oluşacaktır. Bu durumun önüne geçmek ve mesajların hangi sınıfta olduğu olasılığını hesaplamak için eşitlik (3.7) kullanılır. Örneğin $\boldsymbol{d_1}$ mesajının

laplace düzgünleştirmeli olarak olumlu sınıfında olma olasılığı eşitlik (3.7)'ye göre aşağıdaki eşitlikler ile formüle edilip, hesaplanır.

$$P(poz|\boldsymbol{d_1}) = \log Ppoz + (\log Ppoz_{ybank} + \log Ppoz_{hesap} + \log Ppoz_{iade} + \log Ppoz_{işletim})$$

$$P(poz|\boldsymbol{d_1}) = \log\frac{7}{20} + \left(\log\frac{3+\mathbf{1}}{44+\mathbf{136}} + \log\frac{0+\mathbf{1}}{44+\mathbf{136}} + \log\frac{0+\mathbf{1}}{44+\mathbf{136}} + \log\frac{0+\mathbf{1}}{44+\mathbf{136}}\right)$$

$$= -0{,}4559 + (-1{,}6532 - 2{,}2553 - 2{,}2553 - 2{,}2553)$$

$$= -8{,}8745$$

$\boldsymbol{d_1}$mesajının olumsuz sınıfında olma olasılığı;

$$P(neg|\boldsymbol{d_1}) = log\ Pneg + (log\ Pneg_{ybank} + log\ Pneg_{hesap} + log\ Pneg_{iade} + log\ Pneg_{işletim})$$

$$P(neg|\boldsymbol{d_1}) = \log\frac{10}{20} + \left(\log\frac{9+\mathbf{1}}{109+\mathbf{136}} + \log\frac{2+\mathbf{1}}{109+\mathbf{136}} + \log\frac{2+\mathbf{1}}{109+\mathbf{136}} + \log\frac{2+\mathbf{1}}{109+\mathbf{136}}\right)$$

$$= -0{,}3010 + (-1{,}3892 - 1{,}9120 - 1{,}9120 - 1{,}9120)$$

$$= -7{,}4263$$

$\boldsymbol{d_1}$mesajının nötr sınıfında olma olasılığı;

$$P(nötr|\boldsymbol{d_1}) = \log Pnötr + (\log Pnötr_{ybank} + \log Pnötr_{hesap} + \log Pnötr_{iade} + \log Pnötr_{işletim})$$

$$P(nötr|\boldsymbol{d_1}) = \log\frac{3}{20} + \left(\log\frac{1+\mathbf{1}}{20+\mathbf{136}} + \log\frac{0+\mathbf{1}}{20+\mathbf{136}} + \log\frac{0+\mathbf{1}}{20+\mathbf{136}} + \log\frac{0+\mathbf{1}}{20+\mathbf{136}}\right)$$

$$= -0{,}8239 + (-1{,}8921 - 2{,}1931 - 2{,}1931 - 2{,}1931)$$

$$= -9{,}2954$$

Logaritma kullanılarak yapılan hesaplamalar sonucunda da $\boldsymbol{d_1}$ mesajı için aynı şekilde "olumsuz" duygu içermektedir.

Tez kapsamında geliştirilen "N-Gram Veri Analizi" yapısında Naïve Bayes sınıflandırıcı modelinde "Çok Terimli (Multinomial) Model" ile birlikte "Laplace Düzgünleştirme" ve ondalıklı değer taşmasını önlemek için en yüksek olasılıklı sınıfı bulma işleminde logaritma formu kullanılmıştır.

3.4.4. Maksimum Entropi (MaxEnt) Sınıflandırma Modeli

Maksimum Entropi üssel ve logaritmik doğrusal (log-linear) sınıflandırıcılar ailesine aittir ve yaygın olarak **Çok Terimli Mantıksal** (*Multinomial Lojistic*) Regresyon olarak ta bilinir, bu niteliği konuşma ve dil işleme gibi ardışık olmayan sınıflandırma süreçlerindeki etkisini de simgelemektedir (Jurafsky ve Martin, 2006).

Eğitim verisinden elde edilecek kısıtlar ile oluşturulacak bir model ile tahminde bulunmak asıl amaçtır. Duygu analizi çalışmasında mikroblog hizmetinden alınan bir mesajın hangi duyguyu içerdiğini belirlemek için MaxEnt yönetimi uygulamak için ilk olarak sınıfları belirlemek gerekir. Duygu analizinde olumlu, olumsuz ve nötr olmak üzere üç duygu kutbu vardır, bu duygu sınıflarını c_1, c_2, c_3 olarak ifade edersek ve eğitim verisi olmadığı düşünülürse, yeni mesajın her hangi bir duygu sınıfında olma olasılığı 1/3 dür. Ama bu olasılık gerçekçi olmayacaktır, sınıflandırılması yapılmış benzer metin içeren eğitim verisi yardımı ile mesajın duygusu tespit edilmektedir.

Eğitim verisini aşağıdaki gibi ifade edelim,

$$D = \{(d_1, c_1), (d_2, c_2), \dots, (d_{\mathrm{N}}, c_{\mathrm{N}})\} \tag{3.17}$$

d_{i} eğitim verisindeki mesajları, c_{i} duygu sınıflarını, mesajların içindeki terimlerin oluşturduğu özelliklerin karakteristik fonksiyonunu ise $f_{\mathrm{i}}(d, c)$ temsil etmektedir. c sınıfında verilen d mesajı için istatistiksel model $p(c|d)$ ile ifade edilir. Eğitim verisinden elde edilecek olasılıksal dağılımın parametrik üssel gösterimi;

$$p(c|d) = \frac{1}{Z(d)} \exp\left(\sum_i \lambda_i f_i(d, c)\right) \tag{3.18}$$

şeklindedir. Bu eşitlikteki normalleştirme faktörü $Z(d)$ aşağıdaki şekilde formüle edilir.

$$Z(d) = \sum_c \exp\left(\sum_i \lambda_i f_i(d, c)\right) \tag{3.19}$$

Bu doğrultuda eşitlik (3.18) tekrar düzenlenirse,

$$p(c|d) = \frac{1}{\sum_c \exp\left(\sum_i \lambda_i f_i(d,c)\right)} \times \exp\left(\sum_i \lambda_i f_i(d, c)\right)$$

$$p(c|d) = \frac{\exp\left(\sum_i \lambda_i f_i(d,c)\right)}{\sum_c \exp\left(\sum_i \lambda_i f_i(d,c)\right)} \tag{3.20}$$

λ_i her özellik için ifade edilen karakteristik fonksiyonu $f_{\mathrm{i}}(d, c)$ için ağırlık değerini temsil eden bir parametredir, bu parametre değerinin büyüklüğü özelliğin fonksiyonu f_{i}'ye etkisi de büyüktür (Berger ve diğ., 1996). λ_i ağırlık vektörü olarak ta ifade edilir, özelliklerin sınıf içindeki yüksek ağırlık değerleri karar vermede güçlü göstergeler olarak karşımıza çıkmaktadır (Go ve diğ., 2009).

Karmaşık kısıtlar problemlerin çözümünü zorlaştırır, eğer ki karmaşık kısıtlar arındırılırsa problemin çözümü daha kolay olacaktır, bu tip problemlerin çözmenin en basit yollarından biri de Lagrange çarpan yöntemidir (Timor, 1994). Sınıflandırılmış eğitim veri seti üzerinden yeni mesajın duygu sınıfının belirlenmesi eğitim veri setinden kısıtlar (koşullar) oluşturulmasıdır, mesaj hangi sınıfın koşulunu sağlıyorsa o sınıf içinde yer alacaktır, kısacası her sınıf olma kuralı bir kısıttır. Bu şekilde oluşturulan kısıtlar ile maksimum entropi aracılığı ile tahminde bulunma aşamasında Lagrange çarpanları λ_i parametresi olarak kullanılabilir.

Maksimum Entropi prensibinde koşullu (kısıtlı) olasılık esastır, bu yapı içinde koşullar $f_{\mathrm{i}}(d, c)$ karakteristik fonksiyonunun durumunu belirlemek amacı ile kullanılır.

$$f_{t_i}(d, \mathrm{c}) = \begin{cases} 1, \mid eğer\ t_i = \text{":)" veya "-)"} \quad ise \\ 0, \mid diğer\ durumlarda \end{cases} \tag{3.21}$$

Yukarıdaki koşul ile terim t_i tanımda belirtilen karakter grupları ise fonksiyonun değeri 1, diğer durumlarda ise 0 olacaktır, bu durum bir koşuldur, bu şekilde her terim ve sınıf için birçok koşul ile fonksiyonun durumu belirlenebilir.

Berger vd. (1996) göre koşullu dağılımın matematiksel ölçüsü koşullu entropi ile sağlanır.

$$H(p) \equiv -\sum_{d,c} \tilde{p}(d) \times p(c|d) \times \log p(c|d) \tag{3.22}$$

Burada $\tilde{p}$ ampirik (gözleme dayalı) olasılık dağılımıdır. C ile ifade edilen veri setinden gelen olasılık dağılımları maksimum entropi $H(p)$ ile modellendiğinde,

$$p^* = \underset{p \in C}{\operatorname{argmax}}\, H(p) \tag{3.23}$$

Maksimum entropi ile bir kısıtlı set C 'nin içinde tekil bir model olarak p^* her zaman iyi tanımlanmış olabilir (Berger ve diğ., 1996).

Koşullu olabilirlik tahmini, eğitim setinde yüksek (maksimum) sayıda gözlenen terimlerin ağırlık değerlerini belirlemektir. Maksimum olabilirlik (likelihood) aşağıdaki şekilde ifade edilir;

$$L(p) \equiv \log \prod_{d,c} p(c|d)^{\tilde{p}(d,c)} \equiv \sum_{d,c} \tilde{p}(d,c) \times \log p(c|d) \quad (3.24)$$

C, veri setindeki, D mesajları için maksimum olabilirlik modelinin özellik ağırlık parametresi λ ile birlikte üssel gösterimi; (Manning ve Schütze, 1999)

$$L(p) \equiv \log P(C|D,\lambda) = \sum_{(c,d)\in(C,D)} \log \frac{\exp \sum_i \lambda_i f_i(c,d)}{\sum_{c'} \exp \sum_i \lambda_i f_i(c',d)} \quad (3.25$$

3.4.4.1. Terim Frekans Ağırlıkları Yaklaşımı

λ_i parametresi özellik fonksiyonunun ağırlık parametresidir, bu parametre Lagrange çarpanları haricinde terim frekansları (TF-IDF) aracılığı ile de oluşturulabilir.

Her terimin veri seti içinde yer alan dokümanlarda bulunma sayılarının (TF-IDF) bir vektör yapısı içinde gösterilmesi ve bu doğrultuda terimlerin veri seti içindeki ağırlıklarının bulunması süreci **vektör uzayı** (*vector space*) modeli olarak tanımlanmaktadır. Basit anlamda bu yaklaşımda doküman d içinde bir t teriminin görünme sayısı **ağırlık** (*weight*) değerini vermektedir. Metin halindeki bir veri seti (derlem) içindeki her doküman d_i ile tüm set içinde yer alan terimler $(t_1, t_2, \dots, t_\mathrm{n})$ şeklinde ifade edilirse, d_i dökümanı içinde görünen t_i terimlerinin bir yapı içinde gösterilmesi vektör uzayını temsil eder. t_i terimi için TF-IDF terim frekans yöntemi ile hesaplanan ağırlık değeri w_i ile gösterilirse, n adet terimden oluşan n–boyutlu vektör d_i dökümanı için aşağıdaki gibi gösterilir (Zhang ve diğ., 2005).

$$d_\mathrm{i} = (w_1, w_2, \dots, w_\mathrm{n}) \quad (3.26)$$

Manning vd. (2009) göre, bir belge içinde, diğer terimlere göre daha sık yer alan bir terimin önemi diğer terimlere göre daha fazladır. TF ifadesi **terim frekansı** (*term frequency*) ifadesini simgeler, TF terim t'nin d dökümanında görülme sayısıdır ($\mathrm{tf}_{t,d}$). Terim frekansı farklı yöntemler ile hesaplanabilir, terimin ilgili doküman içerisinde bulunma (geçme) sayısı doğrudan kullanılabildiği gibi, terimin ilgili doküman içerisinde bulunma sayısının ilgili dokümandaki toplam kelime sayısına bölümü ile veya ilgili

dokümanda en çok tekrar eden terimin tekrar sayısına bölümü ile hesaplanabilir, bu bölüm işlemleri ile terimin dokümanda bulunma sayısı değeri normalleştirilir.

$$\mathrm{tf}_{t,d} = d\ \text{dökümanında}\ t\ \text{teriminin bulunma sayısı},\quad \text{veya} \quad (3.27)$$

$$\mathrm{tf}_{t,d} = \frac{d\ \text{dökümanında}\ t\ \text{teriminin bulunma sayısı}}{d\ \text{dökümanındaki toplam terim sayısı}} = \frac{\mathrm{f}_{t,d}}{\mathrm{N}_d} \quad \text{veya} \quad (3.28)$$

$$\mathrm{tf}_{t,d} = \frac{d\ \text{dökümanında}\ t\ \text{teriminin bulunma sayısı}}{d\ \text{dökümanında en çok tekrar eden terimin tekrar sayısı}} \quad \text{şeklindedir.} \quad (3.29)$$

$\mathrm{f}_{t,d}$, t teriminin d dokümanında bulunma (görülme) sayısıdır, N_d , d dokümanının tamamında yer alan terim sayısıdır.

t teriminin içinde geçtiği (bulunduğu) doküman sayısına doküman frekansı (df_t) adı verilir.

$$\mathrm{df}_t \rightarrow \{\mathrm{d} \in \mathrm{D}: \mathrm{t} \in \mathrm{d}\} \quad (3.30)$$

d doküman, D veri setinin tamamıdır, t ise d dokümanında yer alan terimdir. Veri setinde daha az sayıda belgede görülen bir terimin ayırt edici özelliği daha fazladır, bir terimin veri setinde görünüm özelliği (sık veya seyrek olma durumu) **ters doküman frekansı** idf_t (*inverse document frequency*) ile tespit edilir. Veri setinde yer alan doküman sayısı değerinin doküman frekansına bölümünün logaritması ters doküman frekansı (idf_t) değerini verir.

$$\mathrm{idf}_t = \log \frac{N}{\mathrm{df}_t} \quad (3.31)$$

N veri setindeki doküman sayısını temsil etmektedir. Seyrek terim de bu değer yüksek çıkar. Terim frekansı ile ters doküman frekansı birleştirilerek oluşturulacak hesaplamada her terimin her doküman içinde bulunmasına yönelik bileşik ağırlık değeri elde edilir, ifade TF-IDF olarak gösterilir.

$$\text{TF-IDF} = \text{tf-idf}_t = \mathrm{tf}_{t,d} \times \mathrm{idf}_t \quad (3.32)$$

Az sayıda dokümanda t terimi çok sayıda yer alırsa TF-IDF değeri yüksek olur, bir veya birden fazla dokümanda t terimi daha az sayıda yer alırsa TF-IDF düşük olur, t terimi her dokümanda bulunursa TF-IDF en düşük değerde olur (Manning vr diğ., 2009).

Tez kapsamında Twitter mikroblog sitesinden alınan aşağıdaki iki mesaj üzerinden TF-IDF ağırlık hesaplama süreci aşağıdadır.

Tablo 3.18: Örnek tweet mesajlarında TF-IDF.

Mesaj No	Tweet Mesajı	Terim Sayısı
$\boldsymbol{d_1}$	Y_Marka buzdolabı iyi değil, çok kötü çok.	7
$\boldsymbol{d_2}$	Çok çok iyi klima istiyorsan klima kralı X_Marka klima	9
$\boldsymbol{d_3}$	Yerli malı, yurdun malı	4

Doküman olarak nitelendireceğimiz Tablo 3.18'deki $\boldsymbol{d_1}$ ve $\boldsymbol{d_2}$ mesajlarına ait terimlerin (kelimelerin) terim frekansları Tablo 3.19'dadır,

Tablo 3.19: Örnek $\boldsymbol{d_1}$, $\boldsymbol{d_2}$ tweet mesajlarında TF.

Terimler ($\boldsymbol{d_1}$)	**Terim Sayısı** $\mathrm{tf}_{t,d}$ (eşitlik 3.27)	Normalleştirme $\mathrm{tf}_{t,d}$ (eşitlik 3.28)
Y_Marka	1	1/7 = 0,143
buzdolabı	1	1/7 = 0,143
iyi	1	1/7 = 0,143
değil	1	1/7 = 0,143
çok	2	2/7 = 0,286
kötü	1	1/7 = 0,143

Terimler ($\boldsymbol{d_2}$)	**Terim Sayısı** $\mathrm{tf}_{t,d}$ (eşitlik 3.27)	Normalleştirme $\mathrm{tf}_{t,d}$ (eşitlik 3.28)
çok	2	2/9 = 0,222
iyi	1	1/9 = 0,111
Klima	3	3/9 = 0,333
istiyorsan	1	1/9 = 0,111
Kralı	1	1/9 = 0,111
X_Marka	1	1/9 = 0,111

Yukarıdaki ayrım doğrultusunda "iyi" teriminin $\boldsymbol{d_1}$ mesajı için ters doküman frekansını ve TF-IDF ağırlık değerini hesaplayalım.

$$\mathrm{idf}_{(iyi)} = \log \frac{N}{\mathrm{df}_{(iyi)}} = \log \frac{mesaj\ sayısı}{\text{doküman frekansı}_{(iyi)}}$$

$$\mathrm{idf}_{(iyi)} = \log \frac{3}{2} = \log_{10} \frac{3}{2} = \log_{10}\ 1{,}5 = 0{,}176$$

$$\text{TF-IDF} = \text{tf-idf}_{(iyi)} = \ \mathrm{tf}_{(iyi,m1)} \times \mathrm{idf}_{(iyi)}$$

$$\text{TF-IDF} = \mathbf{1} \times 0{,}176 = 0{,}176 \qquad \text{(eşitlik 3.27'ye göre)}$$

$$\text{TF-IDF} = \mathbf{0{,}143} \times 0{,}176 = 0{,}025 \qquad \text{(eşitlik 3.28'ye göre)}$$

"klima" teriminin $\boldsymbol{d_2}$ mesajı için ters doküman frekansını ve TF-IDF ağırlık değerini hesaplayalım.

$$\text{idf}_{(klima)} = \log\frac{3}{1} = \log_{10}\frac{3}{1} = \log_{10} 3 = 0{,}477$$

$$\text{TF-IDF} = \text{tf-idf}_{(klima)} = \text{tf}_{(klima,m2)} \times \text{idf}_{(klima)}$$

$$\text{TF-IDF} = \mathbf{3} \times 0{,}477 = 1{,}431 \qquad \text{(eşitlik 3.27'ye göre)}$$

$$\text{TF-IDF} = \mathbf{0{,}333} \times 0{,}477 = 0{,}159 \qquad \text{(eşitlik 3.28'ye göre)}$$

Ters terim frekansı (idf_t) hesaplanırken logaritma tabanı 10 olarak kabul edilmiştir. Her terimin her doküman içindeki ağırlıkları ağırlık vektörünü oluşturacaktır. Örnek mesajlarımız ve yalın TF (eşitlik 3.22) doğrultusunda oluşan terim ağırlık vektörü Tablo 3.20'de görülmektedir.

Tablo 3.20: Örnek tweet mesajların terim ağırlık vektörü (TF-IDF) - 1.

	Terimler												
Mesajlar	Y_Marka	Buzdolabı	iyi	değil	çok	kötü	klima	istiyorsan	kralı	X_Marka	Yerli	malı	yurdun
$\boldsymbol{d_1}$	0,477	0,477	0,176	0,477	0,352	0,477							
$\boldsymbol{d_2}$			0,176		0,352		1,431	0,477	0,477	0,477			
$\boldsymbol{d_3}$											0,477	0,954	0,477

Örnek mesajlarımız ve normalleştirilmiş TF (eşitlik 3.28) doğrultusunda oluşan terim ağırlık vektörü Tablo 3.21'de görülmektedir.

Eşitlik 3.27, 3.28 ve 3.29 'da görüldüğü gibi terim frekansı (TF) değeri birçok farklı formda hesaplanabilmektedir, logaritmik ve boolean formları da aşağıdadır (Manning ve diğ., 2009).

$$\text{Logaritmik form:} \quad \text{TF} = \text{tf} = \text{tf}_{t,d} = 1 + \log(\text{f}_{t,d}) \qquad (3.33)$$

Eşitlikte; $\text{f}_{t,d}$, t terimin d dökümanında bulunma sayısıdır, diğer ifade ile ham terim frekansıdır.

Tablo 3.21: Örnek tweet mesajların terim ağırlık vektörü (TF-IDF) - 2.

	Terimler												
Mesajlar	Y_Marka	Buzdolabı	iyi	değil	çok	kötü	klima	istiyorsan	kralı	X_Marka	yerli	malı	yurdun
$\boldsymbol{d_1}$	0,068	0,068	0,025	0,068	0,050	0,068							
$\boldsymbol{d_2}$			0,020		0,039		0,159	0,053	0,053	0,053			
$\boldsymbol{d_3}$											0,119	0,239	0,119

Boolean Form: $$\text{TF} = \text{tf} = \text{tf}_{t,d} = \begin{cases} 1, & \text{f}_{t,d} > 0 \\ 0, & bunun\ dışında \end{cases} \quad (3.34)$$

Terim frekansı değerinin logaritmik formu doğrultusunda terim ağırlığı (TF-IDF) eşitliği tekrar düzenlenirse,

$$\text{TF-IDF} = \text{tf}_{t,d} \times \text{idf}_t = \left(1 + \log\left(\text{tf}_{t,d}\right)\right) \times \log \frac{N}{\text{df}_t} \quad (3.35)$$

şekline dönüşür.

Witten ve diğ. (2011)'ne göre enformasyon elde etme çalışmalarında terim frekansı ve terim ağırlığı aşağıdaki eşitlikler doğrultusunda hesaplanır.

Terim Frekansı: TF $$\text{tf}_{t,d} = \log(1 + \text{f}_{t,d}) \quad (3.36)$$

Terim Ağırlığı: $\text{tf}_{t,d} \times \text{idf}_t = \text{f}_{t,d} \times \log \frac{N}{\text{df}_t}$ eşitlik (3.36)'ya göre düzenlenirse aşağıdaki şekilde oluşur.

$$\text{tf}_{t,d} \times \text{idf}_t = \log(1 + \text{f}_{t,d}) \times \log \frac{N}{\text{df}_t} \quad (3.37)$$

Eşitlikte N, toplam doküman sayısı, df_t, t teriminin içinde bulunduğu doküman sayısını temsil etmektedir.

3.4.4.2. N-gram Yapılı Eğitim Setinde TF-IDF Terim Ağırlığı

Bölüm 3.4.4.1'de doküman olarak nitelendirdiğimiz her mesaj içindeki terimlerin (gramlar -kelime veya kelime grupları) ağırlıklarının TF-IDF özelliği ile hesaplanması açıklanmıştır. Eğitim setimizde üç farklı duygu sınıfına ayrılmış mesajlarımız yer almaktadır, bu mesajlar N-gram yapılarına ayrılmıştır, maksimum entropi yönteminde her terimin sınıflarda bulunma değerleri doğrultusunda sınıflardaki bulunma ağırlıklarının hesaplanması gerekir. Martineau ve Finin (2009) yaptıkları çalışma da TF-IDF ağırlık hesaplamasını terimlerin bulundukları dokümanlarındaki ağırlıklarını hesaplamak yerine, olumlu ve olumsuz olmak üzere iki sınıfa ayırdıkları bir eğitim setindeki terimlerin duygu sınıfları içindeki ağırlıklarını hesaplamak için kullanmışlardır. Eşitlik 3.28 doğrultusunda TF-IDF ağırlık eşitliği aşağıdaki gibi düzenlenebilir.

$$\text{TF-IDF} = \text{tf}_{t,c} \times \text{idf}_t = \frac{\text{f}_{t,c}}{n_c} \times \log \frac{N_c}{\text{df}_{t,c}}, \tag{3.38}$$

veya eşitlik (3.36)'ya göre aşağıdaki eşitlik oluşur.

$$\text{TF-IDF} = \text{tf}_{t,c} \times \text{idf}_t = \log(1 + \text{f}_{t,c}) \times \log \frac{N_c}{\text{df}_{t,c}} \tag{3.39}$$

$\text{f}_{t,c}$, c sınıfında t teriminin frekans değeridir (c sınıfında yer alan mesajlarda t terimin bulunma – görünme sayısı) , n_c, c sınıfında yer alan tekil haldeki toplam terim sayısını, $\text{df}_{t,c}$, c sınıfı içinde t terimi bulunan mesaj sayısını, N_c, c sınıfı içinde yer toplam mesaj sayısını temsil etmektedir.

Tablo 3.22: Örnek eğitim verisi özeti.

Sınıflardaki	Olumlu Sınıfı	Olumsuz Sınıfı	Nötr Sınıfı	Toplam
Mesaj Sayısı	7	10	3	20
Terim Sayısı	44	109	20	173
Frekans "ybank"	3	9	1	13
"ybank" içeren mesaj sayısı	3	9	1	13
Frekans "banka"	0	4	0	4
"banka" içeren mesaj sayısı	0	3	0	3

Tablo 3.10'da duygu sınıflarına ayrılmış eğitim veri seti ve bu set içinden elde edilmiş 1-gram verilerinin duygu sınıflarında bulunma frekansları (Tablo 3.13) doğrultusunda "ybank" teriminin TD-IDF ağırlık değerleri aşağıda hesaplanmıştır (hesaplama da logaritma 10 tabanına göredir), hesaplama esnasında ihtiyaç duyulan değerlerin özet gösterimi Tablo 3.22 yer almaktadır.

$$\mathrm{tf}_{ybank,\boldsymbol{olumlu}} \times \mathrm{idf}_{ybank,\boldsymbol{olumlu}} = \frac{\mathrm{tf}_{ybank,olumlu}}{n_{olumlu}} \times \log \frac{N}{\mathrm{df}_{ybank,olumlu}}$$

$$\mathrm{tf}_{ybank,olumlu} \times \mathrm{idf}_{ybank,olumlu} = \frac{3}{44} \times \log \frac{7}{3}$$

$$\mathrm{tf}_{ybank,olumlu} \times \mathrm{idf}_{ybank,olumlu} = 0{,}0682 \times 0{,}368 = 0{,}0251$$

$$\mathrm{tf}_{ybank,\boldsymbol{olumsuz}} \times \mathrm{idf}_{ybank,\boldsymbol{olumsuz}} = \frac{\mathrm{tf}_{ybank,olumsuz}}{n_{olumsuz}} \times \log \frac{N}{\mathrm{df}_{ybank,olumsuz}}$$

$$\mathrm{tf}_{ybank,olumsuz} \times \mathrm{idf}_{ybank,olumsuz} = \frac{9}{109} \times \log \frac{10}{9}$$

$$\mathrm{tf}_{ybank,olumsuz} \times \mathrm{idf}_{ybank,olumsuz} = 0{,}0826 \times 0{,}00378 = 0{,}0038$$

$$\mathrm{tf}_{ybank,\boldsymbol{nötr}} \times \mathrm{idf}_{ybank,\boldsymbol{nötr}} = \frac{\mathrm{tf}_{ybank,nötr}}{n_{nötr}} \times \log \frac{N}{\mathrm{df}_{ybank,nötr}}$$

$$\mathrm{tf}_{ybank,nötr} \times \mathrm{idf}_{ybank,nötr} = \frac{1}{20} \times \log \frac{3}{1}$$

$$\mathrm{tf}_{ybank,nötr} \times \mathrm{idf}_{ybank,nötr} = 0{,}05 \times 04771 = 0{,}0239$$

Eğer terimler ilgili sınıflar içinde yer almazlar ise df_tdeğeri 0 olacak, $\mathrm{idf}_t = \log(\frac{N}{\mathrm{df}_t})$ hesaplaması da sıfıra bölüm hatası verecektir. Bu durumun oluşmaması için **düzgünleştirme** (*smoothing*) yapılmalıdır. Robertson ve arkadaşlarının **olasılıksal enformasyon alma** (*probabilistic information retrieval*) hesaplama yöntemi **BM25** (*BM-Best Matching – En İyi Eşleşme*) kullanılarak düzgünleştirme aşağıdaki eşitlikle sağlanır (Robertson ve diğ., 1994; Robertson ve diğ., 1998; Robertson ve Zaragoza, 2009).

$$\mathrm{idf}_t = \log \frac{N - \mathrm{df}_t + 0{,}5}{\mathrm{df}_t + 0{,}5} \tag{3.40}$$

Tablo 3.23: Eğitim setindeki terimlerin (unigram) TF-IDF ($\mathrm{tf}_{t,d} \times \mathrm{idf}_t$) ağırlık vektörü

Mesaj Sayısı (20)	**Olumlu Sınıfı (7)**			**Olumsuz Sınıfı (10)**			**Nötr Sınıfı (3)**		
Sınıftaki Toplam Terim (Kelime)	**44**			**109**			**20**		
1-gram (uni-gram)	**Kelimenin Frekansı**	**Mesaj Sayısı**	**TF-IDF**	**Kelimenin Frekansı**	**Mesaj Sayısı**	**TF-IDF**	**Kelimenin Frekansı**	**Mesaj Sayısı**	**TF-IDF**
ybank	3	3	0,0251	9	9	0,0038	1	1	0,0239
xbank	4	4	0,0221	1	1	0,0092	1	1	0,0239
ybankdestek	0	0	0,0000	5	5	0,0138	0	0	0,0000
banka	0	0	0,0000	**4**	**3**	0,0192	0	0	0,0000
ya	0	0	0,0000	2	2	0,0128	1	1	0,0239
:)	2	2	0,0247	1	1	0,0092	0	0	0,0000
ybankdirektmobil	0	0	0,0000	2	2	0,0128	0	0	0,0000
bi	1	1	0,0192	1	1	0,0092	0	0	0,0000
bir	0	0	0,0000	2	2	0,0128	0	0	0,0000
hesap	0	0	0,0000	2	2	0,0128	0	0	0,0000
iade	0	0	0,0000	2	2	0,0128	0	0	0,0000
işletim	0	0	0,0000	2	2	0,0128	0	0	0,0000
mi	0	0	0,0000	2	2	0,0128	0	0	0,0000
TL	0	0	0,0000	2	2	0,0128	0	0	0,0000

idf_t eşitliğine eklenen düzgünleştirme faktörünün hesaplama sonucunda elde edilen değere etkisi azdır, fakat faktör kullanılmadığında df_tdeğeri 0 olarak geldiğinde sıfıra bölüm hatasının etkisi çok büyük olacaktır (Paltoglou ve Thelwall, 2010). Eşitlik (3.38) doğrultusunda TF-IDF ağırlık eşitliği aşağıdaki gibi düzenlenebilir.

$$\text{TF-IDF} = \mathrm{tf}_{t,c} \times \mathrm{idf}_t = \frac{\mathrm{f}_{t,c}}{n_c} \times \log \frac{N - \mathrm{df}_{t,c} + 0{,}5}{\mathrm{df}_{t,c} + 0{,}5} \tag{3.41}$$

"banka" teriminin eğitim setindeki olumlu ve olumsuz sınıflarındaki TF-IDF terim ağırlıkları eşitlik (3.33)'e göre aşağıdadır.

$$\mathrm{tf}_{banka,\boldsymbol{olumlu}} \times \mathrm{idf}_{banka,\boldsymbol{olumlu}} = \frac{\mathrm{tf}_{banka,olumlu}}{n_{olumlu}} \times \log \frac{N - \mathrm{df}_{bank,olumlu} + 0{,}5}{\mathrm{df}_{bank,olumlu} + 0{,}5}$$

$$\mathrm{tf}_{banka,olumlu} \times \mathrm{idf}_{banka,olumlu} = \frac{0}{44} \times \log \frac{7 - 0 + 0{,}5}{0 + 0{,}5} = 0$$

$$\mathrm{tf}_{banka,\boldsymbol{olumsuz}} \times \mathrm{idf}_{banka,\boldsymbol{olumsuz}} = \frac{\mathrm{tf}_{banka,olumsuz}}{n_{olumsuz}} \times \log \frac{N - \mathrm{df}_{bank,olumsuz} + 0{,}5}{\mathrm{df}_{bank,olumsuz} + 0{,}5}$$

$$\mathrm{tf}_{banka,olumsuz} \times \mathrm{idf}_{banka,olumsuz} = \frac{4}{109} \times \log \frac{10 - 3 + 0{,}5}{3 + 0{,}5} = 0{,}0192$$

Eğitim seti içinde sık geçen ilk 14 kelimenin (1-gramın) Tablo 3.13'de yer alan duygu sınıflarında bulunma frekansları doğrultusunda oluşturulan eşitlik 3.38'e göre TF-IDF ($\mathrm{tf}_{t,d} \times \mathrm{idf}_t$) ağırlık vektörü Tablo 3.23 görülmektedir.

3.4.4.3. TF-IDF Ağırlık Vektörü ile Maksimum Entropi Sınıflandırma

Şekil 3.15'de tez kapsamında oluşturulan N-gram tabanlı duygu analizinde Maksimum Entropi sınıflandırıcı modelinin kullanımı yer almaktadır.

Tablo 3.16'da yer alan tweet mesajlarından $\boldsymbol{d_2}$'nin Maksimum Entropi sınıflandırıcı ve TD-IDF terim ağırlık yöntemi kullanılarak hangi duygu sınıfında olduğunu hesaplamak için Tablo 3.23'de yer alan 1-gram şeklindeki terimlerin ağırlık değerleri kullanılır. Mesajlar içindeki terimlerden 1-gram verileri ile Bölüm 3.4.2'de eşleşmesi tespit edilen terimlerin ağırlık değerleri hesaplamada en etkili parametredir. $\boldsymbol{d_2}$ mesajını ve bu mesaj ile eşleşen terimleri tekrar hatırlayalım.

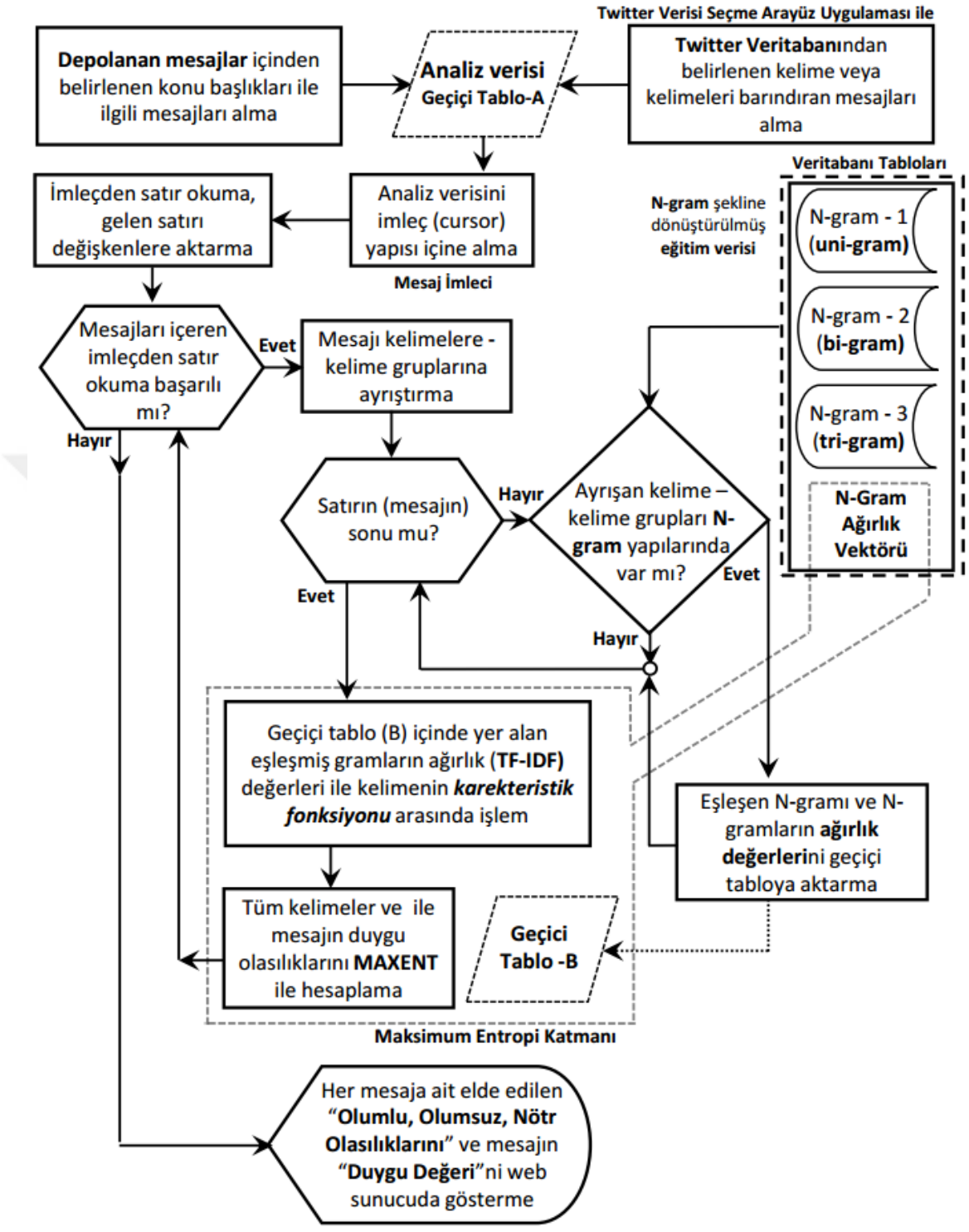
Mesaj ($\boldsymbol{d}$): "*Sıcaktan bunalınca XBank'a girer, sıra alır serinlerim:) Sıra gelmiyor nasıl olsa. Gerçi gelse ne yapacağım o da ayrı bi ironi tabi, serinliyoruz :)*"

Mesajdaki "xbank", "bi" ve ":)"kelimeleri 1-gram verileri ile eşleşmektedir.

Bu doğrultuda $\boldsymbol{d}$ mesajının olumlu olma ihtimalini eşitlik (3.20) aracılığı ile hesaplamak için eşitlik aşağıdaki şekilde düzenlenir.

$$p(olumlu|\boldsymbol{d}) = \frac{\exp(\sum_i \lambda_i f_i(\boldsymbol{d}, olumlu))}{\sum_c \exp(\sum_i \lambda_i f_i(\boldsymbol{d}, c))} \tag{3.42}$$

i, eşleşen kelimelerin indeks sayısı, c ise veri setinde yer alan sınıflara ait indeks sayısıdır (olumlu, olumsuz ve nötr olmak üzere üç sınıfımız olduğundan c=3'dür), λ_i mesaja ait eşleşen terimlerin eğitim verisi aracılığı ile hesaplanmış ağırlık değeridir.



Şekil 3.15: Maksimum entropi sınıflandırıcı modeli.

Eşleşen terimler doğrultusunda eşitlik tekrar yazıldığında;

$$p(olumlu|\boldsymbol{d}) = \frac{\exp(\lambda_{xbank} f_{xbank}(\boldsymbol{d}, olumlu) + \lambda_{bi} f_{bi}(\boldsymbol{d}, olumlu) + \lambda_{:)} f_{:)}(\boldsymbol{d}, olumlu))}{\begin{matrix}\exp(\lambda_{xbank} f_{xbank}(\boldsymbol{d}, olumlu) + \lambda_{bi} f_{bi}(\boldsymbol{d}, olumlu) + \lambda_{:)} f_{:)}(\boldsymbol{d}, olumlu)) \\ + \exp(\lambda_{xbank} f_{xbank}(\boldsymbol{d}, olumsuz) + \lambda_{bi} f_{bi}(\boldsymbol{d}, olumsuz) + \lambda_{:)} f_{:)}(\boldsymbol{d}, olumsuz)) \\ + \exp(\lambda_{xbank} f_{xbank}(\boldsymbol{d}, nötr) + \lambda_{bi} f_{bi}(\boldsymbol{d}, nötr) + \lambda_{:)} f_{:)}(\boldsymbol{d}, nötr))\end{matrix}}$$

$f_i(\boldsymbol{d}, c)$ karakteristik fonksiyonu aşağıdaki şart doğrultusunda oluşturulursa,

$$f_{t_i}(\boldsymbol{d}, \text{olumlu}) = \begin{cases} 1 \mid eğer\ ki\ \boldsymbol{d}\ mesajındaki\ t\ terimi\ \boldsymbol{olumlu}\ sınıfında\ ise \\ 0 \mid diğer\ durumlarda \end{cases}$$

$f_i(\boldsymbol{d}, c)$ fonksiyonu eşitlik 2.16 ve 3.19'da olduğu gibi belirlenen koşul sağlandığında 1, sağlanmadığında 0 değeri (binary değerler) almaktadır, ayrıca mesaj içinde eşleşen terimlerin her sınıfa ait ağırlık değerleri (λ_i) terim frekans yöntemi ile hesaplanmış olarak Tablo 3.23'de yer almaktadır. Bu doğrultuda yukarıdaki eşitliğe ilgili değerlerin aktarılmış hali aşağıdadır.

$$p(olumlu|\boldsymbol{d}) = \frac{\exp((0{,}0221 \times 1) + (0{,}0192 \times 1) + (0{,}0247 \times 1))}{\begin{matrix}\exp((0{,}0221 \times 1) + (0{,}0192 \times 1) + (0{,}0247 \times 1)) \\ + \exp((0{,}0092 \times 1) + (0{,}0092 \times 1) + (0{,}0092 \times 1)) \\ + \exp((0{,}0239 \times 1) + (0 \times 0) + (0 \times 0))\end{matrix}}$$

$$p(olumlu|\boldsymbol{d}) = \frac{exp\ 0{,}0660}{exp\ 0{,}0660 + exp\ 0{,}0275 + exp\ 0{,}0239}$$

$$p(olumlu|\boldsymbol{d}) = \frac{1{,}0682}{1{,}0682 + 1{,}0279 + 1{,}0242} = 0{,}3423$$

$\boldsymbol{d}$ mesajının olumsuz olma ihtimali;

$$p(olumsuz|\boldsymbol{d}) = \frac{exp(\lambda_{xbank} f_{xbank}(\boldsymbol{d}, olumsuz) + \lambda_{bi} f_{bi}(\boldsymbol{d}, olumsuz) + \lambda_{:)} f_{:)}(\boldsymbol{d}, olumsuz))}{\begin{matrix}exp(\lambda_{xbank} f_{xbank}(\boldsymbol{d}, olumlu) + \lambda_{bi} f_{bi}(\boldsymbol{d}, olumlu) + \lambda_{:)} f_{:)}(\boldsymbol{d}, olumlu)) \\ + exp(\lambda_{xbank} f_{xbank}(\boldsymbol{d}, olumsuz) + \lambda_{bi} f_{bi}(\boldsymbol{d}, olumsuz) + \lambda_{:)} f_{:)}(\boldsymbol{d}, olumsuz)) \\ + exp(\lambda_{xbank} f_{xbank}(\boldsymbol{d}, nötr) + \lambda_{bi} f_{bi}(\boldsymbol{d}, nötr) + \lambda_{:)} f_{:)}(\boldsymbol{d}, nötr))\end{matrix}}$$

$$p(olumsuz|\boldsymbol{d}) = \frac{exp((0{,}0092 \times 1) + (0{,}0092 \times 1) + (0{,}0092 \times 1)}{\begin{matrix}exp((0{,}0221 \times 1) + (0{,}0192 \times 1) + (0{,}0247 \times 1)) \\ + exp((0{,}0092 \times 1) + (0{,}0092 \times 1) + (0{,}0092 \times 1)) \\ + exp((0{,}0239 \times 1) + (0 \times 0) + (0 \times 0))\end{matrix}} = 0{,}3294$$

$\boldsymbol{d}$ mesajının nötr olma ihtimali;

$$p(nötr|\boldsymbol{d}) = \frac{\exp(\lambda_{xbank} f_{xbank}(\boldsymbol{d}, nötr) + \lambda_{bi} f_{bi}(\boldsymbol{d}, nötr) + \lambda_{:)} f_{:)}(\boldsymbol{d}, nötr))}{\begin{array}{c}\exp(\lambda_{xbank} f_{xbank}(\boldsymbol{d}, olumlu) + \lambda_{bi} f_{bi}(\boldsymbol{d}, olumlu) + \lambda_{:)} f_{:)}(\boldsymbol{d}, olumlu)) \\ + \exp(\lambda_{xbank} f_{xbank}(\boldsymbol{d}, olumsuz) + \lambda_{bi} f_{bi}(\boldsymbol{d}, olumsuz) + \lambda_{:)} f_{:)}(\boldsymbol{d}, olumsuz)) \\ + \exp(\lambda_{xbank} f_{xbank}(\boldsymbol{d}, nötr) + \lambda_{bi} f_{bi}(\boldsymbol{d}, nötr) + \lambda_{:)} f_{:)}(\boldsymbol{d}, nötr))\end{array}}$$

$$p(nötr|\boldsymbol{d}) = \frac{\exp((0{,}0239 \times 1) + (0 \times 0) + (0 \times 0))}{\begin{array}{c}\exp((0{,}0221 \times 1) + (0{,}0192 \times 1) + (0{,}0247 \times 1)) \\ + \exp((0{,}0092 \times 1) + (0{,}0092 \times 1) + (0{,}0092 \times 1)) \\ + \exp((0{,}0239 \times 1) + (0 \times 0) + (0 \times 0))\end{array}} = 0{,}3282$$

Yukarıdaki hesaplamalar doğrultusunda $\boldsymbol{d}$ mesajı için $p(olumlu|\boldsymbol{d}) > p(olumsuz|\boldsymbol{d}) > p(nötr|\boldsymbol{d})$ olduğu görülmektedir, bu hesaplamalar sonrasında $\boldsymbol{d}$ sınıfının duygu sınıfının "olumlu" olduğu söylenebilir.

Doğal dil işleme aşamalarında maksimum entropi özelliği fonksiyon değeri olarak genellikle binary özelliği kullanılmaktadır fakat hesaplamanın daha tutarlı olması için $f_i(d, c)$ fonksiyonundaki koşulun sağlanması durumunda aşağıda belirtilen hesaplama kullanılır (Nigam ve diğ., 1999).

$$f_{t_i}(d, \mathrm{c}) = \begin{cases} \frac{N(d,t)}{N(d)} & | \; eğer\ ki\ d\ mesajındaki\ t\ terimi\ \ \boldsymbol{c}\ sınıfında\ ise \\ 0 & | \; diğer\ durumlarda \end{cases}$$

$N(d, t)$, d dokümanı içinde t teriminin bulunma sayısı, $N(d)$ ise d dokümanı içindeki toplam terim sayısıdır. Bu doğrultuda $\boldsymbol{d}$ mesajının olumlu sınıfında olma olasılığı aşağıdaki şekilde hesaplanır. (Mesajdaki toplam terim sayısı=22)

$$p(olumlu|\boldsymbol{d}) = \frac{\exp(\lambda_{xbank} f_{xbank}(\boldsymbol{d}, olumlu) + \lambda_{bi} f_{bi}(\boldsymbol{d}, olumlu) + \lambda_{:)} f_{:)}(\boldsymbol{d}, olumlu))}{\begin{array}{c}\exp(\lambda_{xbank} f_{xbank}(\boldsymbol{d}, olumlu) + \lambda_{bi} f_{bi}(\boldsymbol{d}, olumlu) + \lambda_{:)} f_{:)}(\boldsymbol{d}, olumlu)) \\ + \exp(\lambda_{xbank} f_{xbank}(\boldsymbol{d}, olumsuz) + \lambda_{bi} f_{bi}(\boldsymbol{d}, olumsuz) + \lambda_{:)} f_{:)}(\boldsymbol{d}, olumsuz)) \\ + \exp(\lambda_{xbank} f_{xbank}(\boldsymbol{d}, nötr) + \lambda_{bi} f_{bi}(\boldsymbol{d}, nötr) + \lambda_{:)} f_{:)}(\boldsymbol{d}, nötr))\end{array}}$$

$$p(olumlu|\boldsymbol{d}) = \frac{\exp((0{,}0221 \times (\frac{1}{22})) + (0{,}0192 \times (\frac{1}{22})) + (0{,}0247 \times (\frac{2}{22})))}{\begin{array}{c}\exp((0{,}0221 \times (\frac{1}{22})) + (0{,}0192 \times (\frac{1}{22})) + (0{,}0247 \times (\frac{2}{22}))) \\ + \exp((0{,}0092 \times (\frac{1}{22})) + (0{,}0092 \times (\frac{1}{22})) + (0{,}0092 \times (\frac{2}{22}))) \\ + \exp((0{,}0239 \times (\frac{1}{22})) + (0 \times (\frac{1}{22})) + (0 \times (\frac{2}{22})))\end{array}} = 0{,}5998$$

Yukarıdaki eşitlik doğrultusunda $p(olumsuz|\boldsymbol{d}) = 0{,}2425$ ve $p(nötr|\boldsymbol{d}) = 0{,}1577$ olacaktır.

Tez kapsamında geliştirilen "N-Gram Veri Analizi" yapısında Maksimum Entropi sınıflandırıcı modelinde ağırlık parametresini hesaplamak için "TF-IDF – Terim Frekansı ile Ağırlık" yöntemi kullanılmış, yöntem içinde terim frekansı (TF) hesaplama aşamasında logatitmik form tercih edilmiş, ters doküman frekansı (IDF) hesaplama aşamasında ise normalleştirme kullanılmıştır.

4. BULGULAR

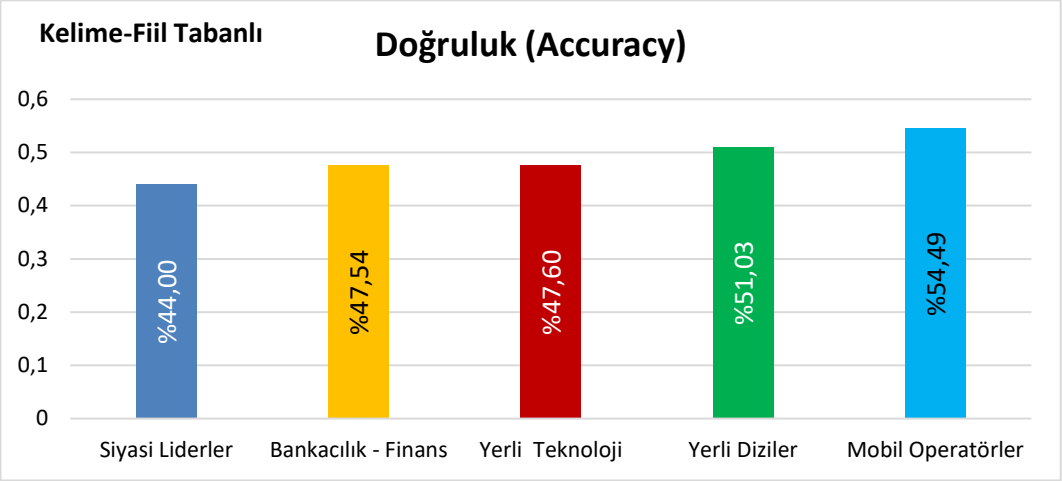
Tez kapsamında bir duygu analizi aracı geliştirilmiştir. Veri analiz aracında Türkçe dil yapısına uygun Anahtar Kelime ve Fiil Tabanlı duygu analizi yapısı ve metin sınıflandırma sürecinde kullanılan makine öğrenmesi yöntemlerinden Naïve Bayes ve Maksimum Entropi algoritmaları doğrultusunda yazılım olarak geliştirilen duygu analizi yapıları yer almaktadır. Ayrıca tez kapsamında geliştirilen veri çözümleme özgün sınıfı ile Twitter mikroblog sitesinden Türkiye ile bağlantılı konularla (siyasi liderler, banka-finans, yerli teknoloji, mobil operatörler, yerli diziler) ilgili alınan tweet mesajları aracılığı ile ilgili konu başlıklarına ait veri setleri oluşturulmuştur. Bu bölümde geliştirilen veri analiz aracının ilgili veri setleri üzerinde uygulanması ve bu süreç doğrultusunda yapılan analizlerin sonucunda elde edilen bulgulara yer verilmiştir.

Tüm analizlerde sınıflandırılması yapılmış veri setleri **5-katlı çapraz geçerlilik** model değerlendirme yöntemi ile değerlendirilmiştir. Sınıflama modelleri test verisi durumuna dönüştürülen $k = 1,2,..,5$ katları için ayrı ayrı uygulanmıştır.

Geliştirilen veri analizi aracında üç farklı sınıf ile çalışıldığından model performansını ve geçerliliğini belirlenmesini sağlayan ölçüler üç kategorili (sınıflı) karışıklık matris (confusion matrix) aracılığı ile hesaplanmaktadır.

4.1. ANAHTAR KELİME VE FİİL TABANLI DUYGU ANALİZİ BULGULARI

Anahtar Kelime ve Fiil tabanlı duygu analizi modeli kapsamında eğitim verisinden yararlanmadan tahmin yapıldığından, modelin değerlendirilmesinde Naïve Bayes ve Maksimum Entropi sınıflandırıcıları için kullanılan eğitim verisi test verisi olarak kullanılmıştır (5 farklı konu başlığına ait 5 eğitim seti analiz amacı ile test veri seti olarak kullanılmıştır). Bu doğrultuda bu modelin değerlendirilmesinde herhangi bir eğitim–test verisi ile gerçekleştirilen bir model değerlendirme yöntemi kullanılmamıştır. Tamamı eğitim verisi olan test verisi üzerinde modelin uygulanmasıyla elde edilen sonuçlar ile gerçek sonuçların karşılaştırılmasından elde edilen her konu başlığına ait doğruluk (accuracy) ölçüsü değerlerinin grafiği Şekil 4.1'de yer almaktadır.



Şekil 4.1: Anahtar Kelime ve Fiil tabanlı model ile her konu başlığına ait doğruluk.

Veri setleri üzerinde modelin uygulanması doğrultusunda elde edilen doğruluk değerleri düşük çıkmıştır. Türkçe dilinin yapısı ve konu başlıkları ile ilgili Twitter mesajlarının çoğunun tepki mesajı olması çıkan sonucu etkilemiştir.

Üç sınıflı (kategorili) Anahtar Kelime ve Fiil tabanlı modelin her bir konu başlığı için geçerliliğini sınama (performans) ölçüleri Tablo 4.1'de yer almaktadır.

Tablo 4.1: Anahtar Kelime ve Fiil tabanlı modelin geçerliliğini sınama (performans) ölçüleri.

Konular	Doğruluk (Accuracy)	Kesinlik (Precision)			Duyarlılık (Sensitivity)		
		Olumlu	Olumsuz	Nötr	Olumlu	Olumsuz	Nötr
Siyasi Liderler	0,4400	0,2011	0,8700	0,6364	0,6909	0,3595	0,0409
Bankacılık - Finans	0,4754	0,2987	0,8673	0,5476	0,7500	0,3360	0,0628
Yerli Teknoloji	0,4760	0,3481	0,8980	0,0000	0,9016	0,2973	0,0000
Yerli Diziler	0,5103	0,5292	0,4577	0,5882	0,6719	0,2889	0,0417
Mobil Operatörler	0,5449	0,4030	0,8866	0,2143	0,7864	0,3963	0,0224

Analiz sonuçlarına genel olarak bakıldığında düşük doğruluk, olumsuz sınıfında da yüksek kesinlik tespit edilmektedir, bu durum sistematik bir önyargıyı temsil etmektedir.

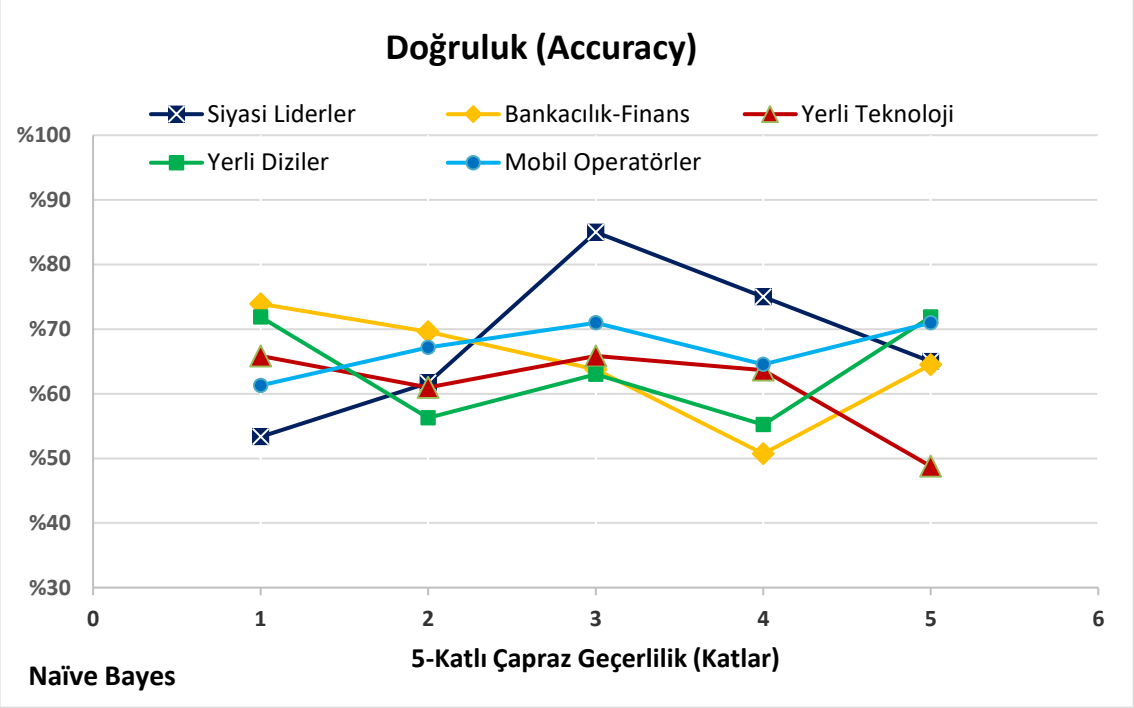
Kesinlik; model ile elde edilen sonuçların birbirlerine yakınlığını, tekrarlanabilirliğini ifade eder, Duyarlılık ise modelin küçük farkları ayırt edebilme kabiliyetinin bir

ölçüsüdür, F-Ölçüsü geçerlilik sınama değeri ise, kesinlik ve duyarlılık ölçü değerleri ile elde edilmektedir, bu doğrultu da bu değer bu iki ölçünün etkisindedir (Tan ve diğ., 2014). Anahtar Kelime ve Fiil tabanlı model için her konu başlığı ve her sınıf için F-Ölçüsü değerleri Tablo 4.2'de yer almaktadır.

Tablo 4.2: Anahtar Kelime ve Fiil tabanlı modele ait F-Ölçüsü değerleri.

	Konular	**Siyasi Liderler**	**Bankacılık - Finans**	**Yerli Teknoloji**	**Yerli Diziler**	**Mobil Operatörler**
F-Ölçüsü	Olumlu	0,3115	0,4272	0,5023	0,5921	0,5329
	Olumsuz	0,5088	0,4844	0,4467	0,3542	0,5478
	Nötr	0,0769	0,1127	0,0000	0,0779	0,0406

F-Ölçüsü değeri sınıflandırma modelinin ilgili sınıf için hangi seviyede doğru sınıflandırma tahmini yaptığını belirten bir ölçüdür. Bu değer maksimum 1 olabilir, değerin 1'e yakın çıkması sınıflandırma modelinin başarılı olduğu anlamına gelir. Her bir sınıf açısında bakıldığında analiz sonuçlarına göre modelin başarısı düşük çıkmaktadır.

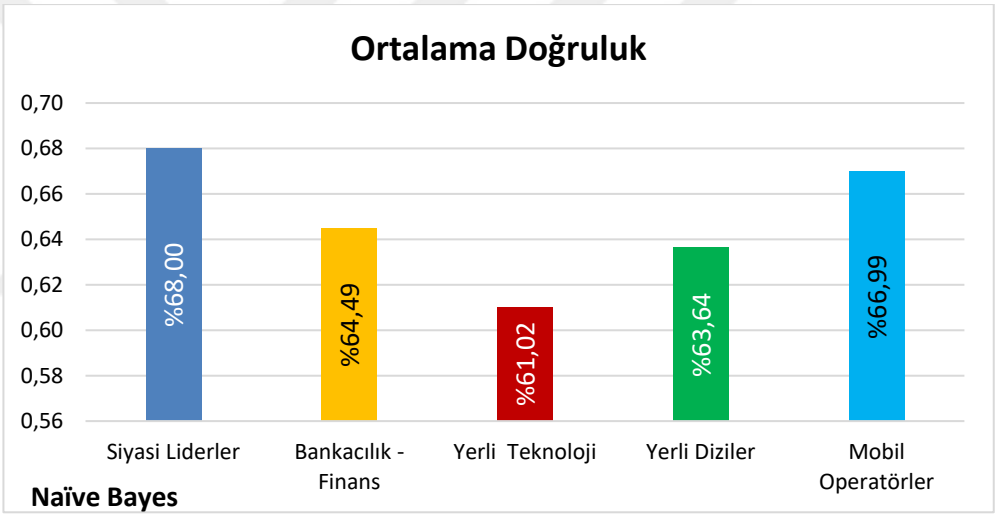


Şekil 4.2: Naïve Bayes sınıflandırıcı ile her konu başlığına ait doğruluk.

4.2. NAÏVE BAYES SINIFLANDIRICI İLE DUYGU ANALİZİ BULGULARI

Makine öğrenmesi algoritmalarından Naïve Bayes sınıflandırıcısı ile 5 farklı konu başlığına ait 5 farklı eğitim veri seti üzerinden gerçekleştirilen duygu analizinde 5-kat çapraz geçerlilik yöntemi uygulanarak her kat aşamasında oluşturulan eğitim ve test veri setleri ile elde edilen doğruluk (accuracy) değerlerine ait grafik Şekil 4.2'de yer almaktadır.

Modelin tamamının doğruluk değeri tahmini tüm ayrık veri setlerinin doğruluk ölçü değerinin ortalamasına eşittir, bu doğrultuda 5 katlı çapraz geçerleme yönteminin her katında elde edilen doğruluk değerinin ortalaması elde edilen her konu başlığı için modelin doğruluk değeri grafiği Şekil 4.3'de yer almaktadır.



Şekil 4.3: Naïve Bayes sınıflandırıcı ile her konu başlığına ortalama doğruluk.

Naïve Bayes modelinin ortalama doğruluk sonuçlarına göre bu model tez kapsamında geliştirilen Anahtar Kelime ve Fiil Tabanlı modele göre daha tutarlı sonuçlar vermiştir. İngilizce üzerine yapılan çalışmalara göre ise bu model düşük sayılabilecek değerler üretmiştir.

Üç sınıflı (kategorili) Naïve Bayes modelin her bir kat için geçerliliğini sınama (performans) ölçüleri Tablo 4.3'de yer almaktadır. Bu esnada olumlu sınıfı için hesaplanan duyarlılık ölçüsü *doğru pozitif (olumlu) oranı* olarak, olumsuz sınıfı için hesaplanan duyarlılık ölçüsü *doğru negatif (olumsuz) oranı* olarak, nötr sınıfı için hesaplanan duyarlılık ölçüsü *doğru nötr oranı* olarak ta ifade edilir. Tüm bu

hesaplamalarda her sınıf için tespit edilen doğru tahminlerin diğer sınıflarda yapılan yanlış tahminlere oranlaması yapılır.

Aynı şekilde yanlış sınıf tahmin değerleri içinde *yanlış pozitif (olumlu)*, *yanlış negatif (olumsuz)* ve *yanlış nötr* oranları hesaplanmaktadır, bu değerler duyarlılık ölçüsünün tersidir ve modelin geçerliliği için düşük değer olması beklenen ölçülerdir. Eğer ki bu değerler yüksek ise modelin hata oranı da yüksektir.

Tablo 4.3: Naïve Bayes modelinin geçerliliğini sınama (performans) ölçüleri.

Konular	Katlar	Doğruluk (Accuracy)	Kesinlik (Precision)			Duyarlılık (Sensitivity)		
			Olumlu	Olumsuz	Nötr	Olumlu	Olumsuz	Nötr
Siyasi Liderler	Kat 1	0,5333	0	0,5614	0	0	0,9143	0
	Kat 2	0,6167	0	0,6316	0	0,0417	0,9474	0
	Kat 3	0,8500	0	0,8772	0	0,1111	0,9615	0
	Kat 4	0,7500	0	0,7500	0	0	1	0
	Kat 5	0,6500	0	0,6610	0	0	0,9750	0
Bankacılık - Finans	Kat 1	0,7391	0,2857	0,7692	0	0,0606	0,9434	0
	Kat 2	0,6957	0,4375	0,7395	0,3333	0,1750	0,8889	0,0244
	Kat 3	0,6377	0,8261	0,6000	0	0,2923	0,9452	0
	Kat 4	0,5072	0,5000	0,5128	0	0,1471	0,8451	0
	Kat 5	0,6449	0,3600	0,7080	0	0,2143	0,8333	0
Yerli Teknoloji	Kat 1	0,6585	0,4118	0,8261	1	0,6364	0,6552	0,0667
	Kat 2	0,6098	0,7500	0,5714	0	0,4091	0,8000	0
	Kat 3	0,6585	0,7143	0,6471	0	0,2941	0,9167	0
	Kat 4	0,6364	0,5556	0,6970	0	0,2941	0,7931	0
	Kat 5	0,4878	0,7778	0,4063	0	0,2692	0,8667	0
Yerli Diziler	Kat 1	0,7188	0,7821	0,4667	0,3333	0,8592	0,2692	0,0385
	Kat 2	0,5625	0,6000	0,4286	0	0,7895	0,2308	0
	Kat 3	0,6300	0,6667	0,5238	0	0,8254	0,2895	0
	Kat 4	0,5521	0,5543	0,5000	0	0,9623	0,0465	0
	Kat 5	0,7188	0,7442	0,5000	0	0,9275	0,1111	0,0741
Mobil Operatörler	Kat 1	0,6129	0,2222	0,6792	0	0,1053	0,8372	0
	Kat 2	0,6719	0,6000	0,6939	0	0,3750	0,8500	0
	Kat 3	0,7097	0,7778	0,6981	0	0,3043	0,9487	0
	Kat 4	0,6452	0,7500	0,6200	0	0,3214	0,9118	0
	Kat 5	0,7097	0,6429	0,7292	0	0,4091	0,8750	0

Naïve Bayes model için her konu başlığı ve her sınıf için F-Ölçüsü değerleri Tablo 4.2'de yer almaktadır.

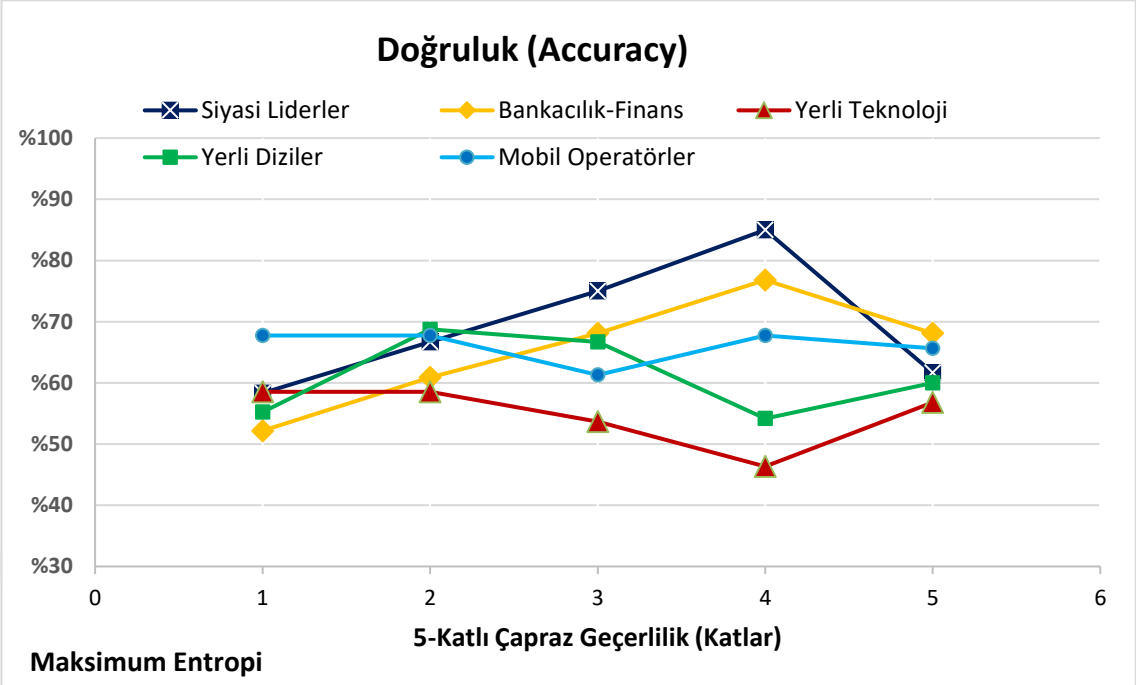
Tablo 4.4: Naïve Bayes modele ait F-Ölçüsü değerleri.

		Konular	Siyasi Liderler	Bankacılık - Finans	Yerli Teknoloji	Yerli Diziler	Mobil Operatörler
Kat 1	**F-Ölçüsü**	Olumlu	0	0	0	0	0
		Olumsuz	0,6957	0,7579	0,9174	0,8571	0,7879
		Nötr	0	0	0	0	0
Kat 2	**F-Ölçüsü**	Olumlu	0,1000	0,2500	0,4318	0,2273	0,2687
		Olumsuz	0,8474	0,8073	0,7340	0,6383	0,7656
		Nötr	0	0	0	0	0
Kat 3	**F-Ölçüsü**	Olumlu	0,5000	0,5294	0,4167	0	0,4000
		Olumsuz	0,7308	0,6666	0,7587	0,7420	0,5532
		Nötr	0,1251	0	0	0	0
Kat 4	**F-Ölçüsü**	Olumlu	0,8188	0,6818	0,7376	0,7034	0,8258
		Olumsuz	0,3414	0,3000	0,3729	0,0851	0
		Nötr	0,0690	0	0	0	0
Kat 5	**F-Ölçüsü**	Olumlu	0	0,4615	0,4375	0,4500	0,5000
		Olumsuz	0,7500	0,7641	0,8043	0,7381	0,7955
		Nötr	0	0	0	0	0

Naïve Bayes modeline ait F-Ölçüsü değerlerine bakıldığında katların çoğunda özellikle olumsuz sınıfı tahmininde F-Ölçüsü değerinin 1'e yakın olduğu görülmektedir, bu sonuç modelin olumsuz sınıfının tahmin edilmesinde başarılı olduğu anlamına gelir. Bu durumun sadece olumsuz sınıfının belirlenmesinde yüksek çıkmasının nedeni ise Türkiye'de toplumsal konular, firmalar ve markalar ile ilgili olarak Twitter üzerinden yazılan mesajların genelde tepki, şikayet içermesinden kaynaklanmaktadır.

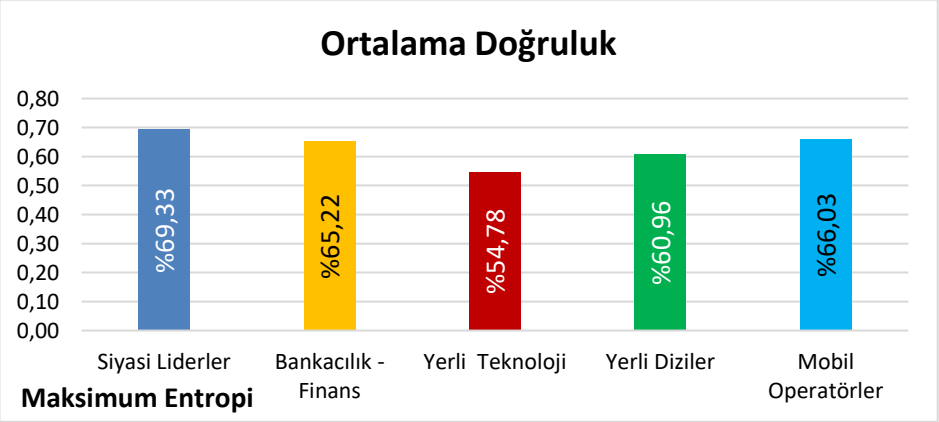
4.3. MAKSİMUM ENTROPİ SINIFLANDIRICI İLE DUYGU ANALİZİ BULGULARI

Makine öğrenmesi algoritmalarından Maksimum Entropi sınıflandırıcısı ile 5 farklı konu başlığına ait 5 farklı eğitim veri seti üzerinden gerçekleştirilen duygu analizinde 5-kat çapraz geçerlilik yöntemi uygulanarak her kat aşamasında oluşturulan eğitim ve test veri setleri ile elde edilen doğruluk (accuracy) değerlerine ait grafik Şekil 4.4'de yer almaktadır.



Şekil 4.4: Maksimum Entropi sınıflandırıcı ile her konu başlığına ait doğruluk.

Modelin tamamının doğruluk değeri tahmini tüm ayrık veri setlerinin doğruluk ölçü değerinin ortalamasına eşittir, bu doğrultuda 5 katlı çapraz geçerleme yönteminin her katında elde edilen doğruluk değerinin ortalaması elde edilen her konu başlığı için modelin doğruluk değeri grafiği Şekil 4.5'de yer almaktadır.



Şekil 4.5: Maksimum Entropi sınıflandırıcı ile her konu başlığına ortalama doğruluk.

Maksimum Entropi modelinin doğruluk değerleri Naïve Bayes modeline yakın çıkmıştır. Teoride Maksimum Entropi modeli Naïve Bayes'e göre daha doğru sonuçlar üretmektedir (Nigam ve diğ., 1999), fakat literatürde farklı çalışmalarda Naïve Bayes modelinin daha başarılı sonuçlar ürettiği de görülmektedir (Go ve diğ., 2009).

Üç sınıflı (kategorili) Maksimum Entropi modelin her bir kat için geçerliliğini sınama (performans) ölçüleri Tablo 4.5'de yer almaktadır.

Tablo 4.5: Maksimum Entropi modelinin geçerliliğini sınama (performans) ölçüleri.

Konular	Katlar	Doğruluk (Accuracy)	Kesinlik (Precision)			Duyarlılık (Sensitivity)		
			Olumlu	Olumsuz	Nötr	Olumlu	Olumsuz	Nötr
Siyasi Liderler	Kat 1	0,5833	0	0,5833	0	0	1	0
	Kat 2	0,6667	0	0,6667	0	0	1	0
	Kat 3	0,7500	0	0,7500	0	0	1	0
	Kat 4	0,8500	0	0,8500	0	0	1	0
	Kat 5	0,6167	0	0,6167	0	0	1	0
Bankacılık - Finans	Kat 1	0,5217	0,8000	0,5113	0	0,0580	0,9855	0
	Kat 2	0,6087	0,8667	0,5772	0	0,2000	0,9726	0
	Kat 3	0,6812	0,4286	0,7054	0	0,0698	0,9381	0
	Kat 4	0,7681	0,5000	0,7863	0	0,0938	0,9626	0
	Kat 5	0,6812	1	0,6788	0	0,0222	1	0
Yerli Teknoloji	Kat 1	0,5854	0,3077	0,7037	1	0,3333	0,6786	0,0556
	Kat 2	0,5854	0,5000	0,5946	0	0,1176	0,9167	0
	Kat 3	0,5366	0,7143	0,5000	0	0,2273	0,8947	0
	Kat 4	0,4634	0,7500	0,3939	0	0,2308	0,8667	0
	Kat 5	0,5682	0	0,6098	0	0	0,8929	0
Yerli Diziler	Kat 1	0,5521	0,5647	0,4545	0	0,8889	0,1190	0
	Kat 2	0,6875	0,7241	0,2000	0,5000	0,9130	0,0370	0,0667
	Kat 3	0,6667	0,7033	0	0	0,9275	0	0
	Kat 4	0,5417	0,5556	0,3333	0	0,9259	0,0476	0
	Kat 5	0,6000	0,6000	0,6000	0	0,9661	0,0732	0
Mobil Operatörler	Kat 1	0,6774	0	0,6885	0	0	0,9767	0
	Kat 2	0,6774	0,8000	0,6667	0	0,1739	0,9744	0
	Kat 3	0,6129	1	0,5789	0	0,1724	1	0
	Kat 4	0,6774	0,8000	0,6667	0	0,1739	0,9744	0
	Kat 5	0,6563	1	0,6557	0	0,0833	0,9756	0

Modelin performans ölçü değerleri incelendiğinde yüksek doğruluk değerlerinin genel de tepki mesajı içeren konu başlıklarına ait veri setlerinden elde edildiği görülmektedir, kesinlik değerlerinin genelde bir sınıfta yüksek diğer sınıflarda düşük olduğu sınıflamanın rahat ayrıştırılabildiği veri setlerinde model daha başarılıdır. Duyarlılık değerleri bazı veri setlerinde %95'lerin üstüne çıktığı görülmüştür, bu oranın yüksekliği ilgili sınıf için modelin başarılı çalıştığını gösterir.

Maksimum Entropi modeli için her konu başlığı ve her sınıf için F-Ölçüsü değerleri Tablo 4.6'da yer almaktadır.

Tablo 4.6: Maksimum Entropi modeline ait F-Ölçüsü değerleri.

		Konular	**Siyasi Liderler**	**Bankacılık - Finans**	**Yerli Teknoloji**	**Yerli Diziler**	**Mobil Operatörler**
Kat 1	**F-Ölçüsü**	Olumlu	0	0	0	0	0
		Olumsuz	0,7368	0,8000	0,8571	0,9189	0,7629
		Nötr	0	0	0	0	0
Kat 2	**F-Ölçüsü**	Olumlu	0,1082	0,3250	0,1200	0,1580	0,0434
		Olumsuz	0,6733	0,7245	0,8053	0,8656	0,8087
		Nötr	0	0	0	0	0
Kat 3	**F-Ölçüsü**	Olumlu	0,3200	0,1904	0,3449	0,3530	0
		Olumsuz	0,6909	0,7213	0,6415	0,5416	0,7247
		Nötr	0,1053	0	0	0	0
Kat 4	**F-Ölçüsü**	Olumlu	0,6906	0,8077	0,8000	0,6945	0,7403
		Olumsuz	0,1886	0,0624	0	0,0833	0,1305
		Nötr	0	0,1177	0	0	0
Kat 5	**F-Ölçüsü**	Olumlu	0	0,2857	0,2941	0,2857	0,1538
		Olumsuz	0,8077	0,7917	0,7333	0,7917	0,7843
		Nötr	0	0	0	0	0

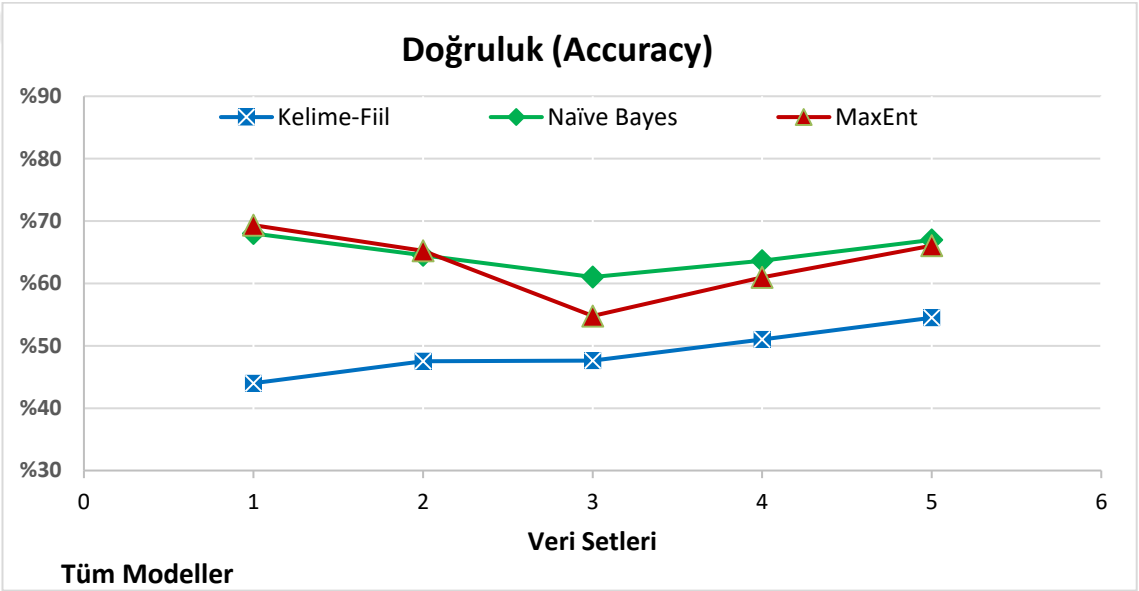
Naïve Bayes modelinde olduğu gibi F-Ölçüsü değerleri özellikle olumsuz sınıfının tahmin edilmesinde modelin başarılı olduğunu göstermektedir.

4.4. MODELLERİ KARŞILAŞTIRMA

5 farklı veri seti üzerinde yapılan analizler doğrultusunda Anahtar Kelime ve Fiil Tabanlı model ile iki adet makine öğrenmesi modelinin geçerliliğini sınama (performans) değerleri üzerinden yapılan karşılaştırmalı veriler Tablo 4.7'de, grafik gösterimde Şekil 4.6 yer almaktadır. Tablo incelendiğinde makine öğrenmesi (Naïve Bayes ve Maksimum Entropi) yöntemlerinin Anahtar Kelime ve Fiil Tabanlı modele göre daha başarılı sonuçlar ürettiği görülmektedir.

Tablo 4.7: Modellerin doğruluk değerleri (ortalama).

Veri Setleri	Doğruluk		
	Kelime Fiil	Naïve Bayes	MaxEnt
Siyasi Liderler	0,4400	0,6800	0,6933
Bankacılık - Finans	0,4754	0,6449	0,6522
Yerli Teknoloji	0,4760	0,6102	0,5478
Yerli Diziler	0,5103	0,6364	0,6096
Mobil Operatörler	0,5449	0,6699	0,6603



Şekil 4.6: Modellerin doğruluk değerlerinin (ortalama) grafik gösterimi.

Makine öğrenmesi teknikleri eğitim veri setlerindeki duygu sınıfı yapılarından oluşan kurallar ile çalışır. Performans ölçümü esnasında ise eğitim veri setinden üretilmiş test verisi üzerinde sınıflama kuralları çalıştırılarak ölçüler hesaplanır. Fakat Anahtar Kelime Tabanlı modellerin karar yapısı oluşturulurken eğitim verisinden yararlanılmaz, model duygu içeren kelime ve fiil havuzunu kullanılır. Eğitim veri setindeki mesajların dilbilgisi açısından doğru yazılmaması, mesajların üç sınıftan özellikle olumsuz sınıfında yoğun olarak yer alması ve kelime havuzunun oluşumunda eğitim seti kelimelerin etkisinin olmaması Anahtar Kelime ve Fiil Tabanlı modelin başarısını etkilemiştir.

Makine öğrenmesi modelleri kendi içlerinde karşılaştırıldıklarında ise her iki model eğitim veri setleri ile test edildiğinde birbirine çok yakın ölçü değerleri elde edildiği görülmüştür. Her iki modelin 5 farklı eğitim setindeki ortalama doğruluk değerleri %60'ın üzerinde çıkmıştır.

Türkçe tweet mesajları üzerinden iki makine öğrenmesi modeli karşılaştırıldığında teorinin aksine Naïve Bayes modelinin Maksimum Entropi modeline göre çok az da olsa daha başarılı sonuçlar ürettiği de görülmektedir.

5. TARTIŞMA VE SONUÇ

Gelişen ve yaygın olarak toplum tarafından kullanılan mobil teknolojiler ve yüksek hızlı internet aracılığı ile bireyler sanal ortamlarda geçmişe göre çok daha fazla zaman geçirmektedirler. İnternet ve sosyal medya yaşamın her katmanında yer almaya başlamış, günlük yaşama ait birçok süreç bu ortamlar aracılığı ile bireylerin yaşamlarının bir parçası şekline dönüşmüştür. Sosyal paylaşım ortamları ve sosyal medyayı kullanan insan sayısı hızla artmaktadır (Statista Inc., 2016).

Bölüm 3.1'de belirtilen veriler doğrultusunda da tüm dünya genelinde olduğu gibi Türkiye'de de sosyal ağ sitelerinden Twitter'ın çok popüler olduğunu görülmektedir. Mobil teknolojiyi ve interneti kullanan bireyler herhangi bir konu hakkında görüşlerini belirtmek için günümüzde Twitter mikroblog hizmetini kullanmaktadırlar. Mesajlar sadece belli bir konuya ait değildir, mesajların içeriklerinde siyasete, teknolojiye, sanata, spora, eğitime, vb. tüm toplumsal yaşama ait görüşler yer almaktadır.

Twitter da yer alan Türkçe mesajlar üzerinde bilgi teknolojileri ve makine öğrenmesi teknikleri kullanılarak duygu analizi çalışması yapılması toplumsal, politik veya herhangi özel bir konu hakkında genel görüş hakkında tahminde bulunma imkânı oluşturacaktır.

Bu tez çalışmasında mikroblog hizmeti üzerinde üretilen mesajların metin madenciliği, doğal dil işleme ve makine öğrenmesi teknikleri kullanılarak analiz edilmesi ve analiz sonucunda mesajlarda yer alan örtük bilginin bulunması amaçlanmıştır. Bu doğrultuda tez kapsamında mesajları Twitter ortamında alan, alınan mesajları belirlenen modeller doğrultusunda analiz eden, analiz sonucunda mesajları duygu sınıflarına ayıran bir analiz aracı geliştirilmiştir. Verinin mikroblog ortamından alınması ve alınan verinin analiz edilmesi şeklinde analiz aracının iki farklı işlevsel yapısı bulunmaktadır. Verinin alınması esnasında JSON veri değişim biçimi kullanılmış, alınan veriler ise bir Microsoft SQL veritabanı sunucusu içine aktarılmıştır. Alınan verinin analizi esnasında ise üç model oluşturulmuş, bu modellerden ikisi makine öğrenmesi tekniğidir. SQL sunucuya ait ileri veritabanı teknikleri kullanılarak "Anahtar Kelime ve Fiil Tabanlı" modeli oluşturulmuş, model ile alınan mesajların üzerinde duygu sınıflarını belirleme çalışması yapılmıştır. Ayrıca geliştirilen bu modele ek olarak bir grup veri eğitim verisi olacak şekilde eğitilerek N-gram yapıları şekline dönüştürülmüş, ortamdan yeni alınan verilerin hangi duygu

sınıfında yer alacağını belirlemek için de literatürde yer alan makine öğrenmesi teknikleri "Naïve Bayes" ve "Maksimum Entropi" modelleri oluşturulup analiz aracına yazılım olarak eklenmiştir. Sonra da bu teknikler uygulanarak mesajların duygu sınıflarına ayrıştırılması yapılmıştır. Son olarak ise modellerin performans ölçü değerleri doğrultusunda modeller arasında karşılaştırma yapılmıştır.

Tez çalışması veritabanlarında bilgi keşfi aşamaları dikkate alınarak oluşturulmuştur. Bu sürecin en çok zaman alan ve performans harcanan aşamaları "veri seçimi" ve "veri önişleme" aşamalarıdır, tez çalışmasında da bu süreçler benzer şekilde gerçekleşmiştir.

Analizi yapılacak verinin seçimi için Twitter ortamı ile konuşacak, ilgili ortamda sorgu çekecek ve sorgu ile elde edilecek mesajları veritabanı alanına depolayacak bir ara yazılım ihtiyacı oluşmuştur. Benzer çalışmalar da bu arayüz ihtiyacı için hazır arayüz sınıf yazılımları kullanılmaktadır, fakat bu yazılımlar mikroblog sitesi içinde yer alan belli başlıklara ait verileri transfer etmektedir, ayrıca belli zamanlarda Twitter'ın paylaştığı verinin yapısında yaptığı düzenleme ve eklemelerin hazır yazılımlar tarafından zamanında belirlenmemesi ve arayüz yazılımların bu değişiklikler doğrultusunda zamanında güncellenmemesi nedeni ile sürekli Twitter ortamından veri çeken kurum ve araştırmacılar sıkıntılar yaşamaktadır. Hazır yazılım yerine analiz aracı içine entegre edilecek özel bir veri alma (toplama) sınıfı geliştirmek uygulamanın daha özgür ve etkin (performanslı) çalışmasını sağlayacaktır.

Twitter ortamında alınan veri temiz bir veri değildir, veri içinde özel karakterler, farklı site veya resim objelerine erişimi sağlayan linkler, hashtag olarak nitelendirilen etiketler, kullanıcıları simgeleyen mentionlar gibi mesaj metninde belirtilen görüş haricinde birçok farklı yapı yer almaktadır. Mesaj içindeki metin ile analiz çalışması yapılacak ise mesajın bu yapılardan temizlenmesi gerekir, bu süreç bir veri önişleme sürecidir. Çok sayıda mesaj üzerinde yapılacak bu işlem için yazılım ve veritabanı teknolojilerini kullanmak süreci hızlandıracak ve oluşturulacak temizleme algoritması ile her mesaj üzerinde benzer işlemleri gerçekleştiren bir temizleme süreci gerçekleşecektir.

Bu kapsamda tez çalışmasında etkin ve tutarlı veri toplama ve veri önişleme yapılması için aşağıdaki yapılar oluşturulmuştur.

- Hazır veri transfer yazılımı yerine veri analiz aracı ile entegre çalışan Twitter ortamından sorgulanan veriyi transfer eden bir özgün arayüz yazılımı geliştirilmiştir. Bu yazılım ile Twitter mikroblog sitesinin paylaştığı tüm veriler sürekli olarak alınabilmektedir. Bu arayüz Twitter ortamından belli konu başlıkları ile ilgili yazılan mesajların belli aralıklar ile bir depolama alanında toplanması için zamanlanmış bir süreç içinde kullanılmaktadır. Ayrıca bu arayüz anlık olarak herhangi bir konu, kelime veya kelime grubu ile ilgili Twitter mikroblog sitesine sorgu gönderilip, en son Twitter da ilgili konu, kelime veya kelime grubu ile ilgili yayınlanan mesajların alınmasını sağlamaktadır.
- Mesajları depolamak (veri seçimi) veya mesajlar üzerinde anlık analiz yapmak amacı ile kullanılan veri transfer arayüzü içinde alınan mesajların temizlenmesini sağlayan bir modül geliştirilmiştir. Modüldeki temizleme işlemi esnasında ise her mesaj harf bazında ayrıştırılmakta mesajın analiz edilmesini zorlaştıracak tüm karakter ve yapılar mesaj gövdesinden çıkarılmaktadır. Bu işlemleri gerçekleştirmek için bir temizleme algoritması tasarlanmıştır. Bu algoritma anlık veri çekme işlemi esnasında bile oldukça verimli ve hızlı çalışmaktadır. Temizleme işlemi esnasında her mesaj kelime seviyesinde ayrıştırılıp bu mesajla ilişkili olacak şekilde farklı tablo yapıları içinde depolanmaktadır.

Analiz aracı çalışmasının ikinci bölümü olarak nitelendirilen sınıflama için model geliştirme veya daha önceden kullanılan modelleri uygulama aşamasında aşağıda yer alan modeller tasarlanmış ve uygulanmıştır.

- Anahtar Kelime ve Fiil Tabanlı Model; ilk olarak bu model ile duygu sınıfı oluşmasına etken olacak kelime, kelime gruplarının belirlenmesi yapılmıştır, bu kelimeler anlam yüklü kelimeler olarak ta nitelendirilebilir. Modelin bu şekilde uygulanmasına Bölüm 2.5'de belirtildiği gibi literatürde sıklıkla karşılaşılmaktadır, fakat tez kapsamında modele entegre edilen fiil-yüklem tabanlı modele literatürde rastlanılmamaktadır.
- Anahtar Kelime ve Fiil tabanlı modelin başarısını ve performansını ölçmek için özellikle İngilizce metinler üzerinde kullanılan makine öğrenmesi tekniklerinden Naïve Bayes ve Maksimum Entropi modelleri veri analiz aracına yazılım olarak entegre edilmiştir.

Anahtar Kelime ve Fiil tabanlı model iki veri bölümünden oluşmaktadır, ilk bölüm anlam yüklü kelimelerdir, bu kelimeler olumlu, olumsuz veya nötr duygu veren kelimelerdir, Türkçe içinde bu özelliğe sahip olan bu kelimeler tez kapsamında oluşturulmuş ve bu kelimelerin hangi oranda, hangi duyguyu temsil ettiği bir grup denek aracılığı belirlenmiştir. Bir mesaj içinde bu kelimelerle sıklıkla karşılaşıldığında bu kelimelerin duygu ağırlıklarının mesajın duygusunu temsil ettiği varsayımı yapılmıştır. Modelin ikinci bölümü ise Türkçe diline özel olarak tasarlanmış modeldir, Bölüm 3.4.1'de belirtildiği gibi Türkçede fiil (yüklem) cümlenin en önemli yapı taşıdır, cümle içindeki fiil üzerinde zaman, şahıs bilgilerini barındırdığı gibi olumlu veya olumsuz duyguyu da temsil eder. Türkçe içinde genelde her cümlenin içinde fiil bulunur, mesajlarda cümlelerden oluştuğu için fiil özellikli kelime barındırırlar. Fiilin durumu cümlenin anlamını da etkiler, tez çalışmasında fiillerin cümle içindeki durumu doğrultusunda cümlenin olumlu, olumsuz, nötr olduğuna karar verilebileceği varsayılmıştır. Cümle içinde bir fiilin tek bir durumu yoktur, zamana, şahısa, tekil/çoğul durumuna ve olumlu/olumsuz olmasına göre farklı yazımları vardır. Her bir fiilin her türlü cümle içinde bulunma yapıları veritabanı içinde barındırılacak şekilde veritabanı sorgulama ve programlama yapıları ile oluşturulmuş ve depolanmıştır. Bu süreç günümüz Türkçe dili üzerinden gerçekleştirilirken belli bir algoritma ile gerçekleştirilmeye çalışılmış, ama bazı fiillerin istisnai durumlar sergilediği görülmüştür. Bu şekilde sorunlu kelimelerin çekimleri özel kodlama ile desteklenerek yapılmıştır. Literatürde Türkçe üzerine yapılan doğal dil işleme çalışmalarında Türkçenin sondan eklemeli bir dil olması nedeni ile cümleler, kelime ve harf seviyesinde ayrıştırılarak anlamları çözümlenmeye çalışılmaktadır. Fakat günümüzde veritabanı sunucu teknolojilerindeki yenilikler, performans artırıcı kodlama ve yapısal değişiklikler cümle içindeki kelimeleri ayrıştırmak yerine, her bir kelimeyi ve duygusunu veritabanı tabloları içinde tutulmasının daha başarılı sonuç vereceği varsayılmıştır. Cümle içinde yer alan her kelime bu fiiller ile karşılaştırılmakta, eşleşmelerde fiilin duygu durumu cümlenin duygusunu temsil etmekte etkisi olduğu varsayılmıştır. Türkçe içinde 6000'e yakın fiil vardır, bu fiiller tüm zaman, şahıs ve olumlu/olumsuz ekleri aldığında sayıları yaklaşık 585.000 adede ulaşmaktadır, fakat iyi bir veritabanı kodlaması ile bu kelimelerin her biri çok kısa sürelerde cümlede yer alan tüm kelimeler ile karşılaştırılabilmektedir.

Bu modeldeki en büyük problem kullanıcıların Türkçeye uygun şekilde kelime ve fiilleri doğru şekilde mesajlar içine yazmamalarıdır. Yazılan kelime ile sistemde depolanmış Anahtar Kelime ve Fiil tabanlı kelimeler uyuşmadığından eşleştirme olmamaktadır. Veritabanı sunucusu ve veritabanı sorgulama dil yapıları efektif olarak kullanıldığından veri boyutunda oluşacak büyüme, karşılaştırma sorgularının çalışmasını olumsuz olarak etkilemeyecektir. Bu doğrultuda her kelime ve fiilin kişiler tarafından yanlış yazılma ihtimallerinin de veritabanına aktarılması ve gerçek kelime ile ilişkilendirilmesinin yapılması ile doğru değerlendirme sonuçları elde edilecektir. Fakat bu eklemeler için zamana ve emeğe ihtiyaç vardır.

Anahtar Kelime ve Fiil Tabanlı modeli karşılaştırmak amacı ile analiz aracına eklenen makine öğrenmesi teknikleri Bölüm 2.6, 3.4.3 ve 3.4.4'de yapısal ve matematiksel formları ile örnekler üzerinde uygulanarak açıklanmıştır. Bölüm 2.5 ve 2.7'de belirtildiği gibi literatürde bu makine öğrenmesi teknikleri Weka, Matlab, R, vb. hazır istatistik ve veri madenciliği yazılımları üzerinde çalıştırılmaktadır. Bu öğrenme teknikleri ile yapılan bilimsel çalışmalarda ilgili algoritmaların yapısal ve matematiksel formlarına değinilmekte, hazırlanmış veri hazır yazılımlar üzerinde çalıştırılıp sonuç alınmaktadır. Tez kapsamında ise her bir makine öğrenmesi algoritmasının veri analiz aracı içinde kodlaması (programlaması) yapılmış, ileri veritabanı sorgulama yapıları, kaydedilmiş prosedürler ve imleç yapıları ile aktif çalışan modüler bir yapıya dönüştürülmüştür. Algoritmalar tamamen veritabanı sunucusu içinde kod şekline dönüştürüldüğünden hızlı ve tutarlı çalışmaktadır. Veritabanı sunucularının görevi sadece veri depolamak değildir, asıl görevleri depoladıkları veriyi en etkin ve hızlı şekilde işlemek ve veri üzerinde talep edilen her türlü seçme, ekleme, silme ve güncelleme işlemini en yüksek verimlilikte gerçekleştirmektir. Veritabanının bu özelliği kullanılarak veri analiz işlemleri çok hızlı şekilde gerçekleşmektedir. Veritabanı programlama ile geliştirilen bu yazılım modülü farklı veri setleri ile çalışan yazılımların içine modüler olarak eklendiğinde de çalışacak yapıdadır.

Makine öğrenmesi tekniklerinden Naïve Bayes veri analizinde uygulanırken eğitim verisi olarak eğitilmiş mesajlar N-gram yapıları olarak kelime veya kelime grupları seviyesinde ayrıştırılmaktadır. Eğitim verisi doğrultusunda her gramın duygu sınıflarında bulunma olasılıkları hesaplanmakta ve bu gram yapıları veritabanında tutulmaktadır. Mikroblog

ortamından analiz edilmek amacı ile alınan yeni mesajlar çalışılan gram seviyesinde ayrıştırılmakta gram tabloları ile karşılaştırılarak eşleşen kelimelerin duygu sınıflarındaki olasılık değerleri doğrultusunda mesajın duygu sınıfı tayin edilmektedir. Naïve Bayes tekniğinin matematiksel formunda her eşleşen terimin sınıfa ait olasılık değerleri çarpma işlemine alınmaktadır. Olasılık değeri 0 ile 1 arasındaki bir ondalıklı bir sayısal değerdir. Eğer ki bir gram yapısı sınıfların hepsinde yer almıyorsa, yer almadığı sınıflardaki olasılık değerleri 0 olacaktır, 0 olan bu değer terimlerin çarpma sonucunu da 0 yapacaktır. Bu durumun önüne geçmek için Laplace düzgünleştirmesi yapılarak 0 olan değerler varsa tüm değerler +1 artırılır. Ayrıca eğer ki eşleşen terim sayısı fazla ise çarpma işleminin sonucunda ondalıklı bölümdeki basamak sayısı sürekli büyüyeceğinden sonuç 0'a doğru gidecektir ve algoritmanın çalıştığı bilgisayar ve yazılımın sayısal sınırlarını aşacağından 0 olarak kabul edilecektir. Bu durumda, sonuç yanlış olacağından Naïve Bayes algoritması ile hesaplama yapılırken her terime ait olasılık değerinin çarpılması yerine, olasılık değerlerinin logaritmalarının alınıp bu değerlerin toplanması sağlanarak çıkacak sonucun normalleşmesi sağlanmalıdır. Tez kapsamındaki tüm analiz ve işlemlerde Naïve Bayes algoritmasında logaritma form yapısı ile işlemler gerçekleştirilmiştir.

Makine öğrenmesi tekniklerinden Maksimum Entropi veri analizinde ise gram seviyesinde ayrıştırılmış eğitim verisinde gramların duygu sınıflarında bulunma değerlerinden diğer bir ifade ile ağırlık değerlerinden yararlanarak işlemler gerçekleştirilir. Ağırlık değeri Lagrange çarpanları kullanılarak hesaplanabildiği gibi terim frekans yöntemi (TF-IDF) de kullanılmaktadır. Eğitim verisinden üretilen gramların sınıflarda bulunma durumları bir fonksiyon olarak tanımlanır, bu fonksiyon ikilik (binary) seviyede tasarlanabilir. Diğer bir ifade ile gram sınıfta var ise 1 yoksa 0 değeri üreten bir fonksiyon olarak tanımlanır ya da gramın sınıfta bulunma olasılık değeri fonksiyonun değeri olarak üretir, gram sınıfta yoksa ikilik yapıda olduğu gibi fonksiyon 0 değeri üretecektir. Maksimum Entropi tekniğinde analizi yapılacak mesaj karşılaştırılması yapılacak gram seviyesinde ayrıştırıldıktan sonra eşleşen terimlerin ağırlık ve fonksiyon değerlerinin çarpımından elde edilen değerlerin her terim için toplanması ile elde edilen değerin üssel sonucu mesajın duygu sınıfını belirleyecektir. TF-IDF yönteminde TF değeri de farklı şekillerde hesaplanabilmektedir, her farklı yöntemde çıkan tahmin değeride değişmektedir, tez kapsamımda en fazla doğru sonuç üreten yöntem

belirlenmeye çalışılmış, elde edilen analiz sonuçlarından TF hesabında logaritma formunun en verimli sonuçları verdiği görülmüştür.

Tez çalışmasında kullanılmak amacı ile 5 farklı konu başlığına (Tablo 3.9) ait 5 farklı eğitim veri seti oluşturulmuştur. Makine öğrenmesi tekniklerinde 5 katlı çapraz geçerleme yöntemi ile her eğitim seti içinden 5 farklı test veri seti oluşturulmuştur. Geliştirilen modellerin test verileri üzerinde çalıştırılması ve **Üç Kategorili Karışıklık Matris** (*Confusion Matrix*) yapısı ile performans ölçümleri yapılmıştır.

Literatür araştırmasında (Bölüm 2.5) genelde duygu analizi çalışmalarının iki kutuplu (olumlu, olumsuz) olduğu görülmektedir. Tez çalışmasında ise duygu analizi üç kutuplu (olumlu, olumsuz ve nötr) olarak tasarlanmış ve çalışılmıştır. Bu doğrultu da genelde iki kategorili (kutuplu) olarak kullanılan olabilirlik tablosu Tablo 2.3'de görüldüğü gibi üç kategorili olarak yapılandırılmıştır. Ayrıca matematiksel olarak iki kategorili çalışmalarda sınıfların olasılık ihtimali üç kategorili çalışmalara göre daha yüksektir. Aynı şekilde iki kategorili çalışmalarda modelin doğruluk oranları üç kategorili çalışmalara göre daha yüksektir.

Tez çalışmasında modeller oluşturulup, veri analiz aracına yazılımsal olarak entegre edildikten sonra yapılan geçerlilik analizlerinde Anahtar Kelime ve Fiil Tabanlı modeli makine öğrenmesi teknikleri için hazırlanan eğitim verisi üzerinde çalıştırılmıştır, fakat bu çalıştırma esnasında makine öğrenmesi tekniklerinde olduğu gibi analize yön veren bir eğitim verisi olmamasından, mesajlar içinde kullanılan kelime ve fiillerin imla açısından genelde yanlış kullanılmasından, özellikle Türkiye'de bir marka, kurum veya siyasi kişilik ile ilgili mesajların tepki ve şikayet içermesinden ve mesajların kinaye (üstü kapalı, sitemli, dokunaklı ifadeler) içeren şekilde yazılmasından dolayı tutarlı sonuçlar oluşmamıştır. Ayrıca özellikle sosyal medya ortamında Türkçe yazılan mesajlar bölgelere göre değişen şive ve ağız farklılıkları kullanılarak yazılmaktadır. Bu durumda yazım dili ve dil bilgisi doğrultusunda hazırlanan kelime havuzu ile yazılan mesajlar arasında uyuşmazlıklar oluşmaktadır. Bu etkenler doğrultusunda, analiz sonucunda doğruluk (accuracy) değerleri düşük çıkmıştır.

Makine öğrenmesi tekniklerinden Naïve Bayes modelinin performans ölçüleri incelendiğinde Anahtar Kelime ve Fiil Tabanlı modele göre yüksek çıkmıştır. İngilizce

dili üzerinde yapılan duygu analizi çalışmalarında bu değerlerin daha yüksek çıktığı görülmektedir, fakat Türkçe'nin dil yapısı, anlam ve duygunun cümlenin farklı öğeleri üzerinde oluşması nedeni ile daha düşük değerler elde edilmiştir.

Makine öğrenmesi tekniklerinden Maksimum Entropi modelinin performans ölçüleri incelendiğinde ise elde edilen sonuçların Naïve Bayes modeline yakın olduğu görülmektedir. Özellikle TF değeri hesaplamasında kullanılan yönteme göre bu değerler değişmekte fakat değişim genel sonucu çok fazla etkilememektedir.

Bölüm 4'de gerçekleştirilen modeller arasındaki karşılaştırma analizinde makine öğrenmesi modelleri Naïve Bayes ve Maksimum Entropi modellerinin Anahtar Kelime ve Fiil Tabanlı modele göre daha başarılı sonuçlar ürettikleri görülmektedir. Fakat bu sonucu değerlendirirken makine öğrenmesi tekniklerinin eğitim veri setlerindeki duygu sınıfı yapılarından oluşan kurallar ile çalıştığı ve performans ölçümü esnasında bu kurallar aracılığı ile yine aynı veri setinden üretilmiş test verisi üzerinde bu kuralların çalıştırıldığının dikkate alınması gerekir. Çünkü anahtar kelime tabanlı modelin eğitim verisi ile hiçbir ilişkisel bağı yoktur, model tamamen daha önceden oluşturulmuş duygu açısından sınıflandırılmış kelime ve çekimi yapılmış fiiller ile veri setinin karşılaştırılması prensibi ile çalışmaktadır.

Anahtar kelime tabanlı duygu analizi çalışmalarında çoğunlukla duygu ifade eden kelimelerden oluşan bir havuz oluşturulur ve yeni elde edilen doküman (mesaj) içindeki kelimeler bu havuz ile karşılaştırılarak dokümanın duygu sınıfı belirlenir. Tez çalışmasında sadece duygu ifade eden kelimeler ile çalışılmamış, ayrıca Türkçe içinde birçok durum, eylem ve duyguyu ifade eden fiillerde (yüklemlerde) diğer akademik çalışmalardan farklı olarak çalışmaya dahil edilmiştir.

Literatür incelendiğinde Twitter üzerinde farklı akademik çalışmalar kapsamında bu tez çalışmasına benzer duygu analizi çalışmaları yapıldığı görülmekte fakat bu çalışmalarda kullanılan derlemin genelde daha önce üzerinde çalışılmış derlemler olduğu görülmektedir. Anlık Twitter verisi üzerinden analiz yapan akademik çalışma sayısının azlığı dikkat çekmektedir. Çalışma kapsamında anlık istenen herhangi bir konuya ait mesajların Twitter ortamında alınması ve çok kısa süre içinde üç farklı model ile analiz edilerek mesajların duygu sınıflarının tahmin edilmesi çalışmanın diğer çalışmalara göre

dikkat çeken farklılığıdır. Ayrıca veri alımı için üçüncü kişilerce yazılmış hazır uygulama kullanılmamış olması çalışmaya özgünlük kazandırmıştır.

Makine öğrenmesi algoritmalarını çalıştırmak için profesyonel yazılımlar (Weka, Matlab, R, vb.) kullanılmamış, üç modelde geliştirilen yazılım içine entegre edilmiştir. Bu entegrasyon ile modellerin tüm matematiksel formları ve veri etkileşimleri yakından incelenmiş, parametrik sayılacak farklı yöntemler test edilerek modelin çalışması en ince detayına kadar incelenmiştir. Böylece Twitter ortamından alınan mesajların analizi esnasında kullanılan modellerin geliştirilen yazılımın içine entegre edilmesi farklı yazılımlara olan bağımlılığı ortadan kaldırmıştır.

Uygulama bu yönü ile gelecekte bir ticari uygulama ürünü olarak piyasaya sunulabilir. Bu ürün farklı uygulama ve platformlardan bağımsız olarak herhangi bir sektörün, markanın, siyasi kurumun, yeni bir ürünün veya kişilerin sosyal medya üzerindeki popülerliği, olumlu veya olumsuz görünümü hakkındaki tüm değerlendirmeleri hızlı ve sürekli olarak toplayabilmekte ve toplanan bu görüşler üzerinden geleceğe yönelik kullanılabilecek değerli bilgileri sağlamaktadır. Sosyal ağlar ve sosyal medya ile bağlantılı her alanda bu veri analiz aracı rahatlıkla kullanılabilir.

Gelecekte analiz aracına farklı makine öğrenmesi yöntem ve algoritmaları eklenebilir, ana yapının veritabanı üzerinde kurulu olması her türlü yöntem ve tekniğin veritabanı programlaması yapılarak analiz aracına eklenmesine imkân tanımaktadır. Yeni eklenecek yöntem ve algoritmalar ile analizler ve tahminler daha tutarlı hale getirilebilir.

İçinde bulunulan dijital teknoloji çağında kişilerin internet ve internet alt yapısı üzerine kurulu paylaşım ortamlarını çok yoğun bir şekilde kullandığı görülmektedir. Bu ortamlarda üretilen veri sürekli büyümektedir, bu büyük veri içinde gizli kalmış, ama değerli olan bilgi yer almaktadır. Bu değerli bilginin elde edilmesi için kullanılması gereken model veritabanlarında bilgi keşfi ve veri madenciliğidir. Tez çalışmasında sürekli büyüyen sosyal medya verileri üzerinde bilgi keşfi yapmak amaçlanmış, bu doğrultuda Twitter mikroblog sitesindeki görüşler ile çalışılmıştır. Twitter mikroblog ortamında kısa ama etkili görüşler zaman zaman yer almaktadır. Sonuç olarak bu hizmet üzerinde yer alan Türkçe mesajlar üzerinde bilgi teknolojileri ve makine öğrenmesi teknikleri kullanılarak duygu analizi çalışması yapılması toplumsal, politik veya herhangi

özel bir konu hakkında genel görüş hakkında tahminde bulunma imkânı oluşturacaktır. Elde edilen bu tahminler ile geleceğe yönelik planlama, düzeltme, iyileştirme gibi aksiyonlar alınabilir. Geçmiş veriden geleceğe yönelik farklı yollar belirlenebilir.

Makine öğrenmesi süreci elektronik ortamda yer alan geçmiş veriden yararlanarak geleceğe yönelik tahmin yapılmasını sağlar. Sonuç olarak bu çalışma ile geleceğe yönelik süreci belirleyecek bilginin geçmiş veriden elde edilmesi farklı makine öğrenmesi modelleri ile yazılımsal olarak gerçekleştirilmiştir.

KAYNAKLAR

Agarwal, A., Xie, B., Vovsha, I., Rambow, O., & Passonneau, R. (2011). Sentiment Analysis of Twitter Data. *Association for Computational Linguistics*, 30–38.

Aggarwal, C. C. (2012). *Mining Text Data.* Boston: Kluwer Academic Publisher.

Apaydın, E. (2010). *Introduction to Machine Learning.* The MIT Press.

Aranoff, M., & Fudeman, K. (2011). *What is Morfology?* Wiley-Blackwell.

Aydoğan, F. (2003). *E-Ticarette Veri Madenciliği Yaklaşımlarıyla Müşteriye Hizmet Sunan Akıllı Modülleri Tasarımı ve Gerçekleştirmi.* Ankara: Hacettepe Üniversitesi Fen Bilimleri Enstitüsü.

Baray, A. (2003). Entropi ve Karar Verme. *İstanbul Üniversitesi Yönetim Dergisi*, 7-21.

Bayrak, Ş., Takçı, H., & Eminli, M. (2012). Makine Öğrenme Yöntemleriyle N-Gram Tabanlı Dil Tanıma. *Elektrik - Elektronik ve Bilgisayar Mühendisliği Sempozyumu - ELECO 2012* (s. 534-538). Bursa: Elektrik Mühendisleri Odası.

Berger, A. L., Pietra, S., & Pietra, V. (1996). A Maximum Entropy Approach to Natural Language Processing. *Association for Computational Linguistics*.

Bifet, A., & Frank, E. (2010). Sentiment Knowledge Discovery in Twitter.

Bilgi Teknolojileri ve İletişim Kurumu. (2016). *Türkiye Elektronik Haberleşme Sektörü Üç Aylık Pazar Verileri Raporu.* Ankara: Bilgi Teknolojileri ve İletişim Kurumu.

Boltzmann, L. (1964). *Lectures on Gas Theory.* Berkeley: University of California Press.

Boyer, J., Gao, S., Malaika, S., Maximilien, M., Salz, R., & Simeon, J. (2011). Experiences with JSON and XML Transformations. *IBM Submission to W3C Workshop on Data and services Integration.* Bedford, ABD.

Bremer, M. (2007). *Undergraduate Topics in Computer Science Principles od Data Mining.* Springer-Verlag London Limited.

Brillouin, L. (1956). *Science and Information Theory.* New York: Academic Press.

Brissaud, J.-B. (2005, Şubat). *The Meaning of Entropy*. Open Access Publishing. adresinden alındı

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. (2002). SMOTE: Syntetic Minority Over-Sampling Technique). *Journal of Artificial Intelligence Research 16*, 321-357.

Cios, K. J., Pedrycz, W., Kurgan, L. A., & Swiniarski, R. W. (2007). *Data Mining: A Knowledge Discovery Approach.* Springer.

Crenna, D. (2013). *Github TweetSharp*. GitHub: https://github.com/danielcrenna/tweetsharp adresinden alındı

Çetin, M., & Amasyalı, M. (2013). Eğiticili ve Geleneksel Terim Ağırlıklandırma Yöntemleriyle Duygu Analizi. *Proceedings of Signal Processing and Communications Applications Conference (SIU)*.

Danacı, M., Çelik, M., & Akkaya, A. E. (2010). Veri Madenciliği Yöntemleri Kullanılarak Meme Kanseri Hücrelerinin Tahmin ve Teşhisi. *Akıllı Sistemlerde Yenilikler ve Uygulamaları Sempozyumu*, (s. 21-24). Kayseri.

Digital Age Dergisi. (2015, Ocak 20). *Twitter'ın 2014 Türkiye Karnesi | Digital Age*. Haziran 06, 2015 tarihinde Digital Age: http://www.digitalage.com.tr/twitterin-2014-turkiye-karnesi/ adresinden alındı

Dil Bilgisi.net. (2014). *Yazım / İmla Kılavuzu.* Dilbilgisi dilin net adresi: http://www.dilbilgisi.net/sozlukler/yazim-kilavuzu/ adresinden alındı

Doğan, S., & Diri, B. (2010). Türkçe Dokümanlar İçin N-gram Tabanlı Yeni Bir Sınıflandırma(Ng-ind): Yazar, Tür ve Cinsiyet. *Türkiye Bilişim Vakfı Bilgisayar Bilimleri ve Mühendislik Dergisi*, 11-19.

Dolgun, M. Ö., Özdemir, T. G., & Oğuz, D. (2009). Veri Madenciliğinde Yapısal Olmayan Verinin Analizi: Metin ve Web Madenciliği. *İstatistikçiler Dergisi*, 48-58.

Duda, R. O., & Hart, P. E. (1973). *Pattern Classification and Scene Analysis.* Menjo Park, California: A Wiley-Interscience Publications.

Dunham, M. H. (2003). *Data Mining Introductory and Advanced Topics.* Prentice Hall / Pearson Education.

EduCause Learning Initiative. (2005, Ağustos). *7 things you should know about Blogs.* EduCause Learning: https://net.educause.edu/ir/library/pdf/ELI7006.pdf adresinden alındı

Efron, B., & Tibshirani, R. J. (1993). *An Introduction to the Bootstrap.* London: Chapman & Hall.

Efron, M. (2011, Haziran). Information Search and Retrieval in Microblogs. *Journal of The American Society for Information Science and Technology*, 997-1008. Haziran 26, 2015 tarihinde alındı

Eryiğit, G., Adalı, E., & Oflazer, K. (2006). Türkçe Cümlelerin Kural Tabanlı Bağlılık Analizi. *Proceedings of the 15th Turkish Symposium on Artificial Intelligence and Neural Networks*, (s. 17-24). Muğla.

Facebook. (2016, Haziran 30). *Company Info - Facebook*. Facebook: http://newsroom.fb.com/company-info/ adresinden alındı

Fayyad, U., Piatetsky-Shapiro, G., & Padhraic, S. (1996, Kasım). The KDD Process for Extracting Useful Knowledge from Volumes Data. *Communications of The ACM, 78*, 27-34.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996a). From Data Mining to Knowledge Discovery in Databases. *Al Magazine, 17*.

Fiske, J. (1990). *Introduction to Communication Studies.* Taylor & Francis Group.

Frawley, W. J., Piatetsky-Shapiro, G., & Matheus, C. J. (1992). Knowledge Discovery in Databases: Overview. *AI Magazine, 13*, 57-70.

Fürnkranz, J., Gamberger, D., & Lavrac, N. (2012). *Foundations of Rule Learning.* Berlin Heidelberg: Springer-Verlag.

Go, A., Bhayani, R., & Huang, L. (2009). *Twitter Sentiment Classification using Distant Supervision.* Stanford: Stanford University.

Goyal, H., Venu, S. D., Pokuri, R., & Kathula, S. (2014). Normalization of Data in Data Mining. *International Journal of Software and Web Sciences (IJSWS)*.

Grauman, K., & Leibe, B. (2011). Visual Object Recognition. R. Brachman, W. W. Cohen, & P. Stone içinde, *Synthesis Lectures on Artificial Intelligence and Machine Learning* (Cilt 5, s. 1-11). Morgan&ClayPool Publishers.

Gülseçen, S. (2012). *Bilgi ve Bilginin Yönetimi.* (S. Gülseçen, Dü.) İstanbul: Papatya Yayıncılık.

Han, J., & Kamber, M. (2006). *Data Mining Concepts and Techniques.* San Francisco: Morgan Kaufmann Publishers.

Hançerlioğlu, O. (2000). *Düşünce Tarihi.* İstanbul: Remzi Kitabevi.

Hand, D., Mannila, H., & Smyth, P. (2001). *Principle of Data Mining.* London: A Bradford Book The MIT Press.

Hastie, T., Tibshirani, R., & Friedman, J. (2008). *The Elements of Statistical Learning.* California: Springer.

Hatzivassiloglou, V., & McKeown, K. (1997). Predicting The Semantic Orientation of Adjectives. *Proceedings of 35th Meeting of the Association for Computational Linguistics*, (s. 174–181). Madrid.

Hipp, J., Güntzer, U., & Gholamreza, N. (2000). Algorithms for Association Rule Mining - A General Survey and Comparison. *SIGKDD Explorations - ACM SIGKDD*, 58-64.

İlhan, S., Duru, N., Karagöz, Ş., & Sağır, M. (2008). Metin Madenciliği ile Soru Cevaplama Sistemi. *ELECO Elektrik - Elektronik ve Bilgisayar Mühendisliği Sempozyumu* (s. 356-359). Bursa: Elektrik Mühendisleri Odası.

Internet World Stats. (2016, Haziran 30). *World Internet Users Statistics and 2016 World Population Stats*. Internet World Stat Usage and Population Statistics: http://www.internetworldstats.com/stats.htm adresinden alındı

Jackson, J. (2002). Data Mining: A Conceptual Overview. *Communications of the Association for Information Systems, 8*, 267-296.

Java, A., Song, X., Finin, T., & Tseng, B. (2007). Why We Twitter: Understanding Microblogging Usage and Communities. *9th WEBKDD and 1st SNA-KDD Workshop '07.* San Jose, California.

Jaynes, E. T. (1995). *Probabilty Theory: The Logic of Science.* St. Louis: Edwin T. Jaynes.

JSON Organization. (2013). *Introducing JSON*. JSON: http://www.json.org/index.html adresinden alındı

Jurafsky, D., & Martin, J. (2006). *Speech and Language Processing: An Introduction to Natural Language Processing.*

Jurafsky, D., & Martin, J. (2015). *Speech and Language Processing - 3rd Edition.*

Kaplan Andreas M., H. M. (2011). The early bird catches the news: Nine things you should know about micro-blogging. *Business Horizons.*

Kavzoğlu, T., & Çölkesen, İ. (2010). Destek Vektör Makineleri ile Uydu Görüntülerinin Sınıflandırılmasında Kernel Fonksiyonlarının Etkilerinin İncelenmesi. *Harita Dergisi*, 73-82.

Kavzoğlu, T., Çölkesen, İ., & Şahin, E. K. (2013). Multispektral Uydu Görüntüleri için En Uygun Bant Seçiminin Sınıflandırma Doğruluğuna Etkilerinin İncelenmesi. *Türkiye Ulusal Fotogrametri ve Uzaktan Algılama Birliği VII. Teknik Sempozyumu (TUFUAB'2013).* Trabzon.

Kaya, M., Fidan, G., & Toroslu, İ. H. (2012). Sentiment Analysis of Turkish Political News. *International Conferences on Web Intelligence and Intelligent Agent Technology*, 174-180.

Khundanpur, S., & Wu, J. (1999). A Maximum Entropy Language Model Integrated N-Grams and Topic Depencies for Conversational Speech Recognition. *IEEE*, 553-556.

Knowledge Forum. (2012). *Welcome to Knowledge Forum*. Knowledge Forum: http://www.knowledgeforum.com/ adresinden alındı

Koç, H. (2007). Dillerin Sınıflandırılması ve Türkçe'nin Dünya Dilleri Arasındaki Yeri (Ünite II). H. Koç içinde, *Dil ve Anlatım I.* Ankara: Milli Eğitim Bakanlığı.

Kohavi, R. (1995). A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. *International Joint Conference on Artificial Intelligence (IJCAI).*

Kok, D., & Brouwer, H. (2010). *Natural Language Processing for the Working Programmer.* Online: http://nlpwp.org/.

Kontopoulos, E., Berberidis, C., Dergiades, T., & Bassilliades, N. (2013). Ontology-based sentiment analysis of twitter posts. *Expert Systems with Applications*, 4065-4074.

Korenek, P., & Simko, M. (2014). Sentiment Analysis on Microblog Utilizing Appraisal Theory. *World Wide Web*, 847-867.

Largeron, C., Moulin, C., & Gery, M. (2011). Entropy Based Feature Selection for Text Categorization. *ACM Symposium on Applied Computing*, (s. 924-928). TaiChung.

Larose, D. T. (2006). *Data Mining Methods and Models.* New Jersey: A John Wiley & Sons. Inc Publication.

Liu, B. (2012). *Sentiment Analysis and Opinion Mining.* Chicago : Morgan & Claypool Publishers.

Maimon, O., & Rokach, L. (2005). *The Data Mining and Knowledge Discovery Handbook A Complete Guide for Practitioners and Researchers.* Springer.

Malouf, R. (2010). Maximum Entropy Model. A. Clark, C. Fox, & S. Lappin içinde, *The Handbook of Computational Linguistics and Natural Language Processing* (s. 133-153). Singapore: A John Wiley & Sons, Ltd, Publications .

Manning, C. D., & Schütze, H. (1999). *Foundations of Statistical Natural Language Processing.* The MIT Press.

Manning, C. D., Raghavan, P., & Schütze, H. (2009). *An Introduction to Information Retrieval.* Cambridge: Cambridge University Press.

Martineau, J., & Finin, T. (2009). Delta TFIDF: An Improved Feature Space for Sentiment Analysis. *Third AAAI International Conference on Weblogs and Social Media.* San Jose.

Mejova, Y. A. (2012). *Sentiment Analysis Within and Across Social Media Streams.* The University of Iowa.

Mitchell, T. M. (1997). *Machine Learning.* McGraw-Hill.

Mukherjee, S. (2012). *Sentiment Analysis.* Bombay: Indian Institute of Technology.

Mullen, T., & Collier, N. (2004). Sentiment Analysis Using Support Vector Machines with Diverse Information. *Proceedings of The Conference on Emprical Methods in Natural Language Processing*, (s. 412-418). Barcelona.

Narayanan, R. (2010). *Mining Text for Relationship Extraction and Sentiment Analysis.* United States-Illinois: ProQuest, UMI Dissertations Publishing.

Neapolitan, R. (1990). *Probabilistic Reasoning in Expert Systems.* John Wiley.

Nigam, K., Lafferty, J., & McCallum, A. (1999). Using Maximum Entropy for Text Classification. *Proceedings of IJCAI-99 Workshop on Machine Learning for Information Filtering*, (s. 61-67).

Nisbet, R., Elder, J., & Miner, G. (2009). *Handbook of Statistical Analysis and Data Mining Applications.* San Diego, California: Elsevier Inc.

Niu, Z., Yin, Z., & Kong, X. (2012). Sentiment Classification for Microblog by Machine Learning. *Fourth International Conference on Computational and Information Sciences* (s. 286-289). IEEE.

Nurseitov, N., Paulson, M., Reynolds, R., & Izurieta, C. (2009). Comparison of JSON and XML Data Interchange Formats: A Case Study. *ISCA 22nd International Conference on Computer Applications in Industry and Engineering, CAINE 2009*, (s. 157-162). San Francisco, California.

Oflazer, K. (2003). Dependency Parsing with An Extended Finite-State Approach. *Association for Computational Linguistics - Volume 29, Number 4*, 515-544.

Oflazer, K., & Bozşahin, H. (1994). Türkçe Doğal Dil İşleme. *Proc. of Turkish Informatics Society TBD'94*.

Olson, D. L., & Delen, D. (2008). *Advanced Data Mining Techniques.* Springer.

Oozeman-Kurrik, A. (2013). *Moving to 64-bit Twitter User IDs - Twitter Developers.* Twitter: https://dev.twitter.com/blog/64-bit-twitter-user-idpocalypse adresinden alındı

Oracle. (2013). *JSR-353 : Java API for Processing JSON* . JSON Processing: https://jsonp.java.net/ adresinden alındı

Outercurve Foundation . (2013). *NuGet Galary*. NuGet Galary Home: http://www.nuget.org/ adresinden alındı

Öğüdücü, Ş. G. (2012). *Veri Madenciliği Metin Madenciliği*. (Ş. G. Öğüdücü, Düzenleyen) Veri Madenciliği Metin Madenciliği. adresinden alındı

Özçakır, F. C. (2006). *Müşteri İşlemlerindeki Birlikteliklerin Belirlenmesinde Veri Madenciliği Uygulaması.* İstanbul: Marmara Üniversitesi Fen Bilimleri Enstitüsü.

Öztürk, A. (2009). *Kalite Yönetimi ve Planlaması.* Bursa: Ekin Yayınevi.

Pak, A., & Paroubek, P. (2010). Twitter as a Corpus for Sentiment Analysis and Opinion Mining. *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10).* Valletta, Malta: European Language Resources Association (ELRA).

Paltoglou, G., & Thelwall, M. (2010). A study of Information Retrieval Weighting Schemes for Sentiment Analysis. *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics* (s. 1386–1395). Uppsala: Association for Computational Linguistics.

Pang, B., & Lee, L. (2008). Opinion Mining and Sentiment Analysis. *Foundations and Trends in Information Retrieval* (s. 1-135). içinde

Pietra, S., Pietra, V., & Lafferty, J. (1997). Inducing Features of Random Fields. *IEEE Transactions Pattern Analysis and Machine Intelligence*.

Prasad, S. (2010). *Micro-blogging Sentiment Analysis Using Bayesian Classification Methods.* San Francisco: Stanford University. Stanford University. adresinden alındı

Ramkumar, G. D., & Swami, A. (1998). Clustering Data Without Distance Function. *IEEE Bulletin of the Technical Committee on Data Engineering* (s. 9-14). IEEE Bulletin of the Technical Committee on Data Engineering, Vol.21 No.l.

Reporter Without Borders. (2005). *Handbook for Bloggerd and Cyber-Dissidents.* Reporters Without Borders (www.rsf.org).

Rich, E., Knight, K., & Nair, S. (1991). *Articial Intelligence.* New Delhi: McGraw-Hill Publishing.

Robertson, S., & Zaragoza, H. (2009). The Probabilistic Relevance Framework: BM25 and Beyond. D. Oard, M. Rijke, & M. Sanderson içinde, *Foundations and Trends in Information Retrieval* (s. 333-389). James Finlay - Now Publisher.

Robertson, S., Walker, S., & Beaulieu, M. (1998). Okapi at TREC-7: Automatic Ad Hoc, Filtering, VLC and Interactive Track. *Proceedings of the Seventh Text Retrieval Conference (TREC 1998)*, (s. 253). Gaithersburg.

Robertson, S., Walker, S., Jones, S., Hancook-Beaulieu, M., & Gatford , M. (1994). Okapi at TREC-3. *Proceedings of The Third Text Retrieval Conference (TREC 1994)*, (s. 109). Gaithersburg.

Ross, S. (1998). *A First Course in Probabilty.* Prentice Hall.

Saif, H., He, Y., & Alani, H. (2012). Semantic Sentiment Analysis of Twitter. *ISWC 2012 - The 11th International Semantic Web Conference.* Boston: International Semantic Web Conference.

Salton, G., & McGill, M. J. (1986). *Introduction to Modern Information Retrieval.* New York: McGraw-Hill.

Saranya, C., & Manikanda, G. (2013). A Study on Normalization Techniques for Privacy Preserving Data Maining. *International Journal of Engineering and Technology (IJET)*, 2071-2074.

Scholz, T., & Conrad, S. (2013). Opinion Mining in Newspaper Articles by Entropy-Based Word Connections. *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, (s. 1828-1839). Seattle.

Sever, H., & Oğuz, B. (2002, Ekim). Veritabanlarında Bilgi Keşfine Formal Bir Yaklaşım, Kısım 1: Eşleşme Sorguları ve Algoritmalar. *Bilgi Dünyası*, 173-204.

Shakev-Shwartz, S., & Ben-David, S. (2014). *Understanding Machine Learning From Theory to Algorithms.* Cambridge University Press.

Shalabi, L. A., Shaaban, Z., & Kasasbeh, B. (2005). Data Mining: A Preprocessing Engine. *Journal of Computer Science 2*, 735-739.

Shannon, C. E. (1948). A Mahematical Theory of Communication. *The Bell System Technical Journal*, 379-423, 623-656.

Singh, S. K., Paul, S., Kumar, D., & Arfi, H. (2014). Sentiment Analysis of Twitter Data Set: Survey. *International Journal of Applied Engineering Research*, 13925-13936.

Smola, A., & Vishwanathan, S. (2008). *Introduction to Machine Learning.* Cambridge: Cambridge University Press.

Somemto. (2013, Ekim 10). *Türkiye Tweet Oldu Yağdı - Somemto ile Gezi Parkı Twitter Analizi*. Somemto: http://somemto.com/turkiye-tweet-oldu-yagdi-somemto-ile-gezi-parki-twitter-analizi/ adresinden alındı

Stanford University. (2013). *Sentiment140 General Information*. Sentiment140: http://help.sentiment140.com/home adresinden alındı

Statista Inc. (2016, Eylül). *Leading social networks worldwide as of September 2016, ranked by number of active users (in millions)*. Mart 06, 2016 tarihinde Statista The Statistics Portal: https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/ adresinden alındı

Steffen, J. (2004). N-Gram Language Modeling Robust Multi-Lingual Document Classification. *4th International Conference on Language Resources and Evalıation (LREC 04).* Lizbon.

Stehman, S. V. (1997). Selecting and Interpreting Measures of Thematic Classification. *Remote Sensing of Environment*, 77-89.

Stieglitz, S., & Dang-Xuan, L. (2014). Emotions and Information Diffusion in Social Media—Sentiment of Microblogs and Sharing Behavior. *Journal of Management Information Systems*, 217-248.

T.C. Gençlik ve Spor Bakanlığı. (2013). *Gençlik ve Sosyal Medya Araştırma Raporu.* Ankara: T.C. Gençlik ve Spor Bakanlığı.

Tan, P.-N., Steinbach, M., & Kumar, V. (2003). *Introduction to Data Mining.* Addison-Wesley Companion.

Tan, P.-N., Steinbach, M., & Kumar, V. (2014). *Introduction to Data Mining.* Pearson Education Limited.

Temel, G. Ö., Erdoğan, S., & Ankaralı, H. (2012). Sınıflama Modelinin Performansını Değerlendirmede Yeniden Örnekleme Yöntemlerinin Kullanımı. *Bilişim Teknolojileri Dergisi*.

Timor, M. (1994). Tamsayılı Doğrusal Programlama Problemlerinin Çözümünde Lagrange Yöntemi (Lagrange Relaxation). *İstanbul Üniversitesi İşletme Fakültesi Dergisi*, 185-193.

Triola, M. F. (2008, Ocak 26). Bayes' Theorem. Washington, Amerika. http://faculty.washington.edu/tamre/BayesTheorem.pdf adresinden alındı

Türk Dil Kurumu. (2015). *Güncel Türkçe Sözlük*. Türk Dil Kurumu: http://www.tdk.gov.tr adresinden alındı

Türk Dili ve Edebiyatı Dersleri Kaynak Eğitim Sitesi. (2007). *Türkçe İmla (Yazım) Kılavuzu.* Türk Dili ve Edebiyatı Dersleri Kaynak Eğitim Sitesi: http://www.turkedebiyati.org/imla-kilavuzu.html adresinden alındı

Türkiye İstatistik Kurumu. (2016, Ocak 10). *Türkiye İstatistik Kurumu*. Türkiye İstatistik Kurumu - Anasayfa: http://www.tuik.gov.tr/Start.do adresinden alındı

Twitter. (2013, Ekim 10). *Documentation | Twitter Developers*. Twitter Developers: https://dev.twitter.com/apps adresinden alındı

Twitter. (2013). *GET search/tweets - Twitter Devolopers.* Twitter: https://dev.twitter.com/rest/reference/get/search/tweets adresinden alındı

Twitter. (2016). *Company | About*. Şubat 15, 2016 tarihinde Twitter: https://about.twitter.com/company adresinden alındı

Twitter. (2016). *OAuth | Twitter Developers.* Twitter: https://dev.twitter.com/oauth adresinden alındı

Vikisözlük. (2015, 07 11). Vikisözlük Özgür Sözlük: https://tr.wiktionary.org/wiki/Vikis%C3%B6zl%C3%BCk adresinden alındı

Vinodhini, G., & Chandrasekaran, R. (2012). Sentiment Analysis and Opinion Mining: A Survey. *International Journal of Advanced Research in Computer Science and Software Engineering*.

Vohra, S., & Teraiya, J. (2013). A Comparative Study of Sentiment Analysis Techniques. *Journal of Information, Konwledge and Research in Computer Engineering*, (s. 313 -317).

Vuran, A. (1983). *İstatistik II.* İstanbul: Nihad Sayar Yayın ve Yardım Vakfı.

We Are Social. (2016). *Digital In 2016.* Londra: We Are Social. http://wearesocial.com/uk/special-reports/digital-in-2016 adresinden alındı

Witten, I. H., Frank, E., & Hall, M. A. (2011). *Data Mining Practical Machine Learning Tools and Techniques.* Morgan Kaufmann Publishers.

Yoon, S. (2011). *Application of Social Network Analysis and Text Mining to Characterize Network Structures and Contents of Microblogging Messages: An Observational Study of Physical Activity-Related Tweets.* Columbia University.

Yoshikawa, Y., Iwata, T., & Sawada, H. (2014). Latent Support Measure Machines for Bag-of-Words Data Classification. *Advances in Neural Information Processing Systems 27 (NIPS 2014).*

Zafarani, R., Abbasi, M. A., & Liu, H. (2014). *Social Media Mining.* Cambridge University Press.

Zafer, H. R. (2012, 03). *Türkçe ve Diğer Türkî Diller için Cümle Çözümleyici*. Harun Reşit Zafer: http://www.hrzafer.com/turkce-ve-diger-turki-diller-icin-cumle-cozumleyici adresinden alındı

Zaki, M. J., & Meira, W. (2014). *Data Mining and Analysis: Fundamental Concepts and Algorithms.* USA: Cambridge University Press.

Zhai, C., & Lafferty, J. (2004). A Study of Smoothing Methods for Language Models Applied to Information Retrieval. *ACM Transactions on Information Systems*, (s. 179-214).

Zhang, Y.-T., Gong, L., & Wang, Y.-C. (2005). An Improved TF-IDF Approach for Text Classification. *Journal of Zhejiang University Science*, 49-55.

EKLER

EK-1. GET search/tweets Sorgusuna ait Örnek JSON Verisi.


```
{
 "statuses": [
  {
   "created_at":"Fri Apr 29 13:09:20 +0000 2016",
   "id":726035641377677312,
   "id_str":"726035641377677312",
   "text":"Veritaban\u0131 Y\u00f6netimi dersi final s\u0131nav\u0131nda dok\u00fcman
kullan\u0131m\u0131 serbest DE\u011e\u0130LD\u0130R. S\u0131nav takvimde belirtilen
g\u00fcn saat 09.30'da ba\u015flayacakt\u0131r. -2-",
   "entities": {
    "hashtags": [

    ],
    "symbols": [

    ],
    "user_mentions": [

    ],
    "urls": [

    ]
   },
   "truncated":false,
   "metadata": {
    "result_type":"recent",
    "iso_language_code":"tr"
   },
   "source":"\u003ca href=\"http:\/\/twitter.com\" rel=\"nofollow\"\u003eTwitter Web
Client\u003c\/a\u003e",
   "in_reply_to_status_id":null,
   "in_reply_to_status_id_str":null,
   "in_reply_to_user_id":null,
   "in_reply_to_user_id_str":null,
   "in_reply_to_screen_name":null,
   "user": {
    "id":1100949518,
    "id_str":"1100949518",
    "name":"Feridun \u00d6z\u00e7ak\u0131r",
    "screen_name":"OzcakirFeridun",
    "location":"",
    "description":"Bili\u015fim d\u00fcnyas\u0131n\u0131n bir yerlerinde..",
    "url":"http:\/\/t.co\/7pZm7ftJga",
    "entities": {
     "url": {
      "urls": [
       {
        "url":"http:\/\/t.co\/7pZm7ftJga",
        "expanded_url":"http:\/\/www.feridunozcakir.com",
        "display_url":"feridunozcakir.com",
        "indices": [
         0,
         22
        ]
       }
      ]
     },
     "description": {
      "urls": [
```

```
      ]
     }
    },
    "protected":false,
    "followers_count":108,
    "friends_count":1,
    "listed_count":0,
    "created_at":"Fri Jan 18 13:14:53 +0000 2013",
    "favourites_count":0,
    "utc_offset":null,
    "time_zone":null,
    "geo_enabled":false,
    "verified":false,
    "statuses_count":163,
    "lang":"tr",
    "contributors_enabled":false,
    "is_translator":false,
    "is_translation_enabled":false,
    "profile_background_color":"C0DEED",
    "profile_background_image_url":"http:\/\/abs.twimg.com\/images\/themes\/theme1\/
bg.png",
    "profile_background_image_url_https":"https:\/\/abs.twimg.com\/images\/themes\/t
heme1\/bg.png",
    "profile_background_tile":false,
    "profile_image_url":"http:\/\/pbs.twimg.com\/profile_images\/3538957171\/8f92573
d1476521ffebdc6724d56cea8_normal.png",
    "profile_image_url_https":"https:\/\/pbs.twimg.com\/profile_images\/3538957171\/
8f92573d1476521ffebdc6724d56cea8_normal.png",
    "profile_banner_url":"https:\/\/pbs.twimg.com\/profile_banners\/1100949518\/1366
666463",
    "profile_link_color":"0084B4",
    "profile_sidebar_border_color":"C0DEED",
    "profile_sidebar_fill_color":"DDEEF6",
    "profile_text_color":"333333",
    "profile_use_background_image":true,
    "has_extended_profile":false,
    "default_profile":true,
    "default_profile_image":false,
    "following":false,
    "follow_request_sent":false,
    "notifications":false},
    "geo":null,
    "coordinates":null,
    "place":null,
    "contributors":null,
    "is_quote_status":false,
    "retweet_count":0,
    "favorite_count":0,
    "favorited":false,
    "retweeted":false,
    "lang":"tr"
   },
  {
   "created_at":"Fri Apr 29 13:06:11 +0000 2016",
   "id":726034849161695232,"id_str":"726034849161695232",
   "text":"Veritaban\u0131 Y\u00f6netimi dersi final s\u0131nav\u0131
yaz\u0131l\u0131 ve uygulama \u015feklinde iki oturum halinde
yap\u0131lacakt\u0131r. -1-",
   "entities": {
    "hashtags": [

   ],
    "symbols": [

    ],
    "user_mentions": [

    ],
    "urls": [

    ]
   },
   "truncated":false,
   "metadata": {
```

```
    "result_type":"recent",
    "iso_language_code":"tr"
   },
   "source":"\u003ca href=\"http:\/\/twitter.com\" rel=\"nofollow\"\u003eTwitter Web
Client\u003c\/a\u003e",
   "in_reply_to_status_id":null,
   "in_reply_to_status_id_str":null,
   "in_reply_to_user_id":null,
   "in_reply_to_user_id_str":null,
   "in_reply_to_screen_name":null,
   "user": {
    "id":1100949518,
    "id_str":"1100949518",
    "name":"Feridun \u00d6z\u00e7ak\u0131r",
    "screen_name":"OzcakirFeridun",
    "location":"",
    "description":"Bili\u015fim d\u00fcnyas\u0131n\u0131n bir yerlerinde..",
    "url":"http:\/\/t.co\/7pZm7ftJga",
    "entities": {
     "url": {
      "urls": [
       {
        "url":"http:\/\/t.co\/7pZm7ftJga",
        "expanded_url":"http:\/\/www.feridunozcakir.com",
        "display_url":"feridunozcakir.com",
        "indices": [
         0,
         22
        ]
       }
      ]
     },
     "description": {
      "urls": [

      ]
     }
    },
    "protected":false,
    "followers_count":108,
    "friends_count":1,
    "listed_count":0,
    "created_at":"Fri Jan 18 13:14:53 +0000 2013",
    "favourites_count":0,
    "utc_offset":null,
    "time_zone":null,
    "geo_enabled":false,
    "verified":false,
    "statuses_count":163,
    "lang":"tr",
    "contributors_enabled":false,
    "is_translator":false,
    "is_translation_enabled":false,
    "profile_background_color":"C0DEED",
    "profile_background_image_url":"http:\/\/abs.twimg.com\/images\/themes\/theme1\/b
g.png",
    "profile_background_image_url_https":"https:\/\/abs.twimg.com\/images\/themes\/th
eme1\/bg.png",
    "profile_background_tile":false,
    "profile_image_url":"http:\/\/pbs.twimg.com\/profile_images\/3538957171\/8f92573d
1476521ffebdc6724d56cea8_normal.png",
    "profile_image_url_https":"https:\/\/pbs.twimg.com\/profile_images\/3538957171\/8
f92573d1476521ffebdc6724d56cea8_normal.png",
    "profile_banner_url":"https:\/\/pbs.twimg.com\/profile_banners\/1100949518\/13666
66463","profile_link_color":"0084B4",
    "profile_sidebar_border_color":"C0DEED",
    "profile_sidebar_fill_color":"DDEEF6",
    "profile_text_color":"333333",
    "profile_use_background_image":true,
    "has_extended_profile":false,
    "default_profile":true,
    "default_profile_image":false,
    "following":false,
    "follow_request_sent":false,
    "notifications":false
   },
```

```
    "geo":null,
    "coordinates":null,
    "place":null,
    "contributors":null,
    "is_quote_status":false,
    "retweet_count":0,
    "favorite_count":0,
    "favorited":false,
    "retweeted":false,
    "lang":"tr"
   }
  ],
"search_metadata": {
  "completed_in":0.017,
  "max_id":999999999999999999,
  "max_id_str":"999999999999999999",
  "next_results":"?max_id=726034849161695231&q=veritaban%C4%B1%20y%C3%B6netimi&lang=tr
&count=2&include_entities=1",
  "query":"veritaban%C4%B1+y%C3%B6netimi",
  "refresh_url":"?since_id=999999999999999999&q=veritaban%C4%B1%20y%C3%B6netimi&lang=t
r&include_entities=1",
  "count":2,
  "since_id":0,
  "since_id_str":"0"
  }
}
```

ÖZGEÇMİŞ

Kişisel Bilgiler	
Adı Soyadı	Feridun Cemal ÖZÇAKIR
Doğum Yeri	Sakarya
Doğum Tarihi	02.07.1971
Uyruğu	☑ T.C. ❒ Diğer:
Telefon	0216 6771630 - 2730
E-Posta Adresi	feridun.ozcakir@okan.edu.tr
Web Adresi	http://www.feridunozcakir.com

Eğitim Bilgileri	
Lisans	
Üniversite	Marmara Üniversitesi
Fakülte	Teknik Eğitim Fakültesi
Bölümü	Elektronik-Bilgisayar Bölümü
Mezuniyet Yılı	13.06.1994

Yüksek Lisans	
Üniversite	Marmara Üniversitesi
Enstitü Adı	Fen Bilimleri Enstitüsü
Anabilim Dalı	Elektronik-Bilgisayar Eğitimi
Programı	Bilgisayar-Kontrol Eğitimi
Mezuniyet Tarihi	19.07.2006

Doktora	
Üniversite	İstanbul Üniversitesi
Enstitü Adı	Fen Bilimleri Enstitüsü
Anabilim Dalı	Enformatik Anabilim Dalı
Programı	Enformatik Programı
Mezuniyet Tarihi	17.10.2016

Makale ve Bildiriler

Makaleler:

Ayvaz Reis, Z., Baktır, H.Ö. , Çelik, B., Erkoç, M.F., Özçakır, F., Özdemir, Ş., Şahin, K., Açık Kaynak Kodlu Öğrenme Yönetim Sistemleri Üzerine Bir Karşılaştırma Çalışması, *JRET - Journal of Research in Education and Teaching*, Sayfa: 42-58, 2012.

Özçakır, F., Çamurcu, Y., Birliktelik Kuralı Yöntemi İçin Bir Veri Madenciliği Yazılımı Tasarımı ve Uygulaması. *İstanbul Ticaret Üniversitesi Fen Bilimleri Dergisi*, 2007.

Çamurcu, Y., Bozkurt, A., Özçakır, F., Şahin, Ş., Karadağ, A. Kişisel Bilgisayarlarda (PC), Paralel Port Deneyleri İçin Bir Deney Sistemi Tasarımı, *Marmara Üniversitesi Fen Bilimleri Dergisi*, 1998.

Bildiriler:

Özçakır, F., Erkoç, M.F., Özçakır, Ş., The Use of Social Media In Education: A Review Of Recent Research, *5th International Conference the Future of Education (FOE 2015)*, 2015, Florence (ITALY).

Erkoç, M.F., Özçakır, F., Erkoç, Ç., The Relationship Between Loneliness and Game Preferences of Secondary School Students, *5th International Conference the Future of Education (FOE 2015)*, 2015, Florence (ITALY).

Özçakır, F., Özçakır, Ş., Bilişim Teknolojileri Eğitiminde Yeterliğe Dayalı Modüler Sistemin Değerlendirilmesi ve İyileştirmeye Yönelik Öneriler, *6th International Computer & Instructional Technologies Symposium (ICITS 12)*, 2012, Gaziantep (TÜRKİYE)

Ozcakir, F., Ozcakir, S., Evalution of Compotency-based Modular System for The Information Technology Area of Vocational and Technical Secondary Education In Turkey, *4th International Conference on Education and New Learning Technologies (EDULEARN 12)*, 2012, Barcelona (SPAIN).

Ayvaz Reis, Z., Baktır, H.Ö. , Çelik, B., Erkoç, M.F., Özçakır, F., Özdemir, Ş., Şahin, K., Açık Kaynak Kodlu Öğrenme Yönetim Sistemleri Üzerine Bir Karşılaştırma Çalışması, *3rd International Conference on New Trends in Education and Their Implications (ICONTE 12)*, 2012, Antalya (TÜRKİYE)

Özçakır, F., Erkoç, M.F., Çelik, B., Baktır, H.Ö., Şahin, K., Ayvaz Reis, Z., Güvenli Bilgisayar ve Internet Kullanımına Yönelik Bir Web Tabanlı Öğretim Ortamının Geliştirilmesi, *5th International Computer & Instructional Technologies Symposium (ICITS 11)*, 2011, Elazığ (TÜRKİYE)

Baktır, H.Ö. , Çelik, B., Erkoç, M.F., Özçakır, F., Docebo Öğrenme Web Tabanlı Öğretim Gereksinimlerini Karşılama Açısından İncelenmesi, *5th International Computer & Instructional Technologies Symposium (ICITS 11)*, 2011, Elazığ (TÜRKİYE)