



**TÜRKÇE TWEETLERİN ANALİZ EDİLEBİLMESİ
İÇİN HADOOP/HIVE KULLANAN MELEZ
BİR SİSTEMİN GELİŞTİRİLMESİ**

Sinan KUL

Doktora Tezi

Yönetim Bilişim Anabilim Dalı

Prof. Dr. Üstün ÖZEN

2018

Her Hakkı Saklıdır

**T.C.
ATATÜRK ÜNİVERSİTESİ
SOSYAL BİLİMLER ENSTİTÜSÜ
YÖNETİM BİLİŞİM SİSTEMLERİ ANABİLİM DALI**

Sinan KUL

**TÜRKÇE TWEETLERİN ANALİZ EDİLEBİLMESİ İÇİN
HADOOP/HIVE KULLANAN MELEZ BİR SİSTEMİN
GELİŞTİRİLMESİ**

DOKTORA TEZİ

**TEZ YÖNETİCİSİ
Prof. Dr. Üstün ÖZEN**

ERZURUM 2018



T.C.
ATATÜRK ÜNİVERSİTESİ
SOSYAL BİLİMLER ENSTİTÜSÜ
TEZ BEYAN FORMU



14/12/2018

SOSYAL BİLİMLER ENSTİTÜSÜ MÜDÜRLÜĞÜNE

BİLDİRİM

Atatürk Üniversitesi Lisansüstü Eğitim ve Öğretim Uygulama Esaslarının ilgili maddelerine göre hazırlamış olduğum **“TÜRKÇE TWEETLERİN ANALİZ EDİLEBİLMESİ İÇİN HADOOP/HIVE KULLANAN MELEZ BİR SİSTEMİN GELİŞTİRİLMESİ”** adlı tezin/raporun tamamen kendi çalışmam olduğunu ve her alıntıya kaynak gösterdiğimi taahhüt eder, tezimin/raporumun kâğıt ve elektronik kopyalarının Atatürk Üniversitesi Sosyal Bilimler Enstitüsü arşivlerinde aşağıda belirttiğim koşullarda saklanmasına izin verdiğimi onaylarım:

Lisansüstü Eğitim ve Öğretim Uygulama Esaslarının ilgili maddeleri uyarınca gereğinin yapılmasını arz ederim *.

☒ Tezimin/Raporumun tamamı her yerden erişime açılabilir.

☐ Tezimin/Raporumun makale için **altı ay**, patent için **iki yıl** süreyle erişiminin ertelenmesini istiyorum.

14.12.2018

Sinan KUL

* LİSANSÜSTÜ TEZLERİN ELEKTRONİK ORTAMDA TOPLANMASI, DÜZENLENMESİ VE ERİŞİME AÇILMASINA İLİŞKİN YÖNERGE

.....

ÜÇÜNCÜ BÖLÜM

Çeşitli ve Son Hükümler

Lisansüstü tezlerin erişime açılmasının ertelenmesi MADDE 6– (1) Lisansüstü teze ilgili patent başvurusu yapılması veya patent alma sürecinin devam etmesi durumunda, tez danışmanının önerisi ve enstitü anabilim dalının uygun görüşü üzerine enstitü veya fakülte yönetim kurulu iki yıl süre ile tezin erişime açılmasının ertelenmesine karar verebilir.

(2) Yeni teknik, materyal ve metotların kullanıldığı, henüz makaleye dönüşmemiş veya patent gibi yöntemlerle korunmamış ve internetten paylaşılması durumunda 3. şahıslara veya kurumlara haksız kazanç imkanı oluşturabilecek bilgi ve bulguları içeren tezler hakkında tez danışmanının önerisi ve enstitü anabilim dalının uygun görüşü üzerine enstitü veya fakülte yönetim kurulunun gerekçeli kararı ile altı ayı aşmamak üzere tezin erişime açılması engellenebilir.

Gizlilik dereceli tezler MADDE 7– (1) Ulusal çıkarları veya güvenliği ilgilendiren, emniyet, istihbarat, savunma ve güvenlik, sağlık vb. konulara ilişkin lisansüstü tezlerle ilgili gizlilik kararı, tezin yapıldığı kurum tarafından verilir. Kurum ve kuruluşlarla yapılan işbirliği protokolü çerçevesinde hazırlanan lisansüstü tezlere ilişkin gizlilik kararı ise, ilgili kurum ve kuruluşun önerisi ile enstitü veya fakültenin uygun görüşü üzerine üniversite yönetim kurulu tarafından verilir. Gizlilik kararı verilen tezler Yükseköğretim Kuruluna bildirilir.

(2) Gizlilik kararı verilen tezler gizlilik süresince enstitü veya fakülte tarafından gizlilik kuralları çerçevesinde muhafaza edilir, gizlilik kararının kaldırılması halinde Tez Otomasyon Sistemine yüklenir.



**T.C.
ATATÜRK ÜNİVERSİTESİ
SOSYAL BİLİMLER ENSTİTÜSÜ**

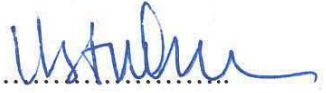


TEZ KABUL TUTANAĞI

SOSYAL BİLİMLER ENSTİTÜSÜ MÜDÜRLÜĞÜNE

Prof. Dr. Üstün ÖZEN danışmanlığında, Sinan KUL tarafından hazırlanan bu çalışma 14 / 12 / 2018 tarihinde aşağıda isimleri yazılı jüri tarafından Yönetim Bilişim Sistemleri Anabilim Dalı'nda Doktora Tezi olarak kabul edilmiştir.

Başkan : Prof. Dr. Üstün ÖZEN

İmza: 

Jüri Üyesi : Prof. Dr. Uğur Yavuz

İmza: 

Jüri Üyesi : Dr. Öğr. Üyesi Mustafa KESKİNKILIÇ

İmza: 

Jüri Üyesi : Prof. Dr. Yılmaz GÖKŞEN

İmza: 

Jüri Üyesi : Prof. Dr. Ahmet BARAN

İmza: 


Prof. Dr. Sait UYLAŞ
Sosyal Bilimler Enstitüsü
Müdürü

Enstitü Müdürü

İÇİNDEKİLER

ÖZET.....	V
ABSTRACT	VI
SİMGELER ve KISALTMALAR DİZİNİ	VII
ŞEKİLLER DİZİNİ	IX
TABLolar DİZİNİ	XIV
ÖNSÖZ.....	XV
GİRİŞ	1

BİRİNCİ BÖLÜM KURAMSAL TEMELLER

1.1. SOSYAL MEDYA.....	4
1.1.1. Sosyal Medya Kullanımının Yaygınlaşma sebepleri	5
1.1.2. Sosyal Medyanın Temel Özellikleri	5
1.1.3. Sosyal Medya Kullanımının Avantaj ve Dezavantajları	6
1.1.4. Sosyal Medyanın İşletmelerde Kullanımı	6
1.1.5. Sosyal Medyanın Tüketici Davranışlarına Etkisi	8
1.1.6. Bir Sosyal Paylaşım Sitesi Olarak Twitter	8
1.2. BÜYÜK VERİ	10
1.2.1. Bilgi Yönetiminin Önemi	10
1.2.2. Aşırı Bilgi Artışı	12
1.2.3. Seri - Paralel Bilgisayarlar.....	14
1.2.4. Paralel Bilgisayar Kümeleri	15
1.2.5. Büyük Veri Tanımı.....	16
1.2.6. Büyük Verinin Boyutları	17
1.2.7. Büyük Verinin İşletmelere Etkileri	19
1.2.8. Sosyal Medya Analizinde Büyük Veri	22
1.2.9. İlişkisel Veri Tabanı Yönetim Sisteminden NoSQL'e	23
1.2.9.1. İlişkisel Veri Tabanı Yönetim Sistemi	23
1.2.9.2. NoSQL	25
1.2.9.3. Döküman Veri Tabanı	27
1.2.9.4. Anahtar Değer Veri Tabanı	27

1.2.9.5. Sütun Veri Tabanı.....	28
1.2.9.6. Çizge (Graph) Veri Tabanı	28
1.2.10. Veri Madenciliğinden Büyük Veri Analizine.....	29

İKİNCİ BÖLÜM

MATERYAL ve YÖNTEM

2.1. HADOOP	31
2.1.1. HDFS.....	33
2.1.2. Hadoop Eşle İndirge	35
2.1.3. Hadoop Eşle İndirge ve Apache Spark.....	37
2.1.4. Hadoop'un Avantajları	38
2.1.5. Hadoop ile İlişkili Projeler	38
2.2. HIVE	39
2.2.1. Veri Tipleri	40
2.2.2. Aritmetik ve Mantıksal İşlemler.....	41
2.2.3. HiveQL Dili.....	42
2.3. DUYGU ANALİZİ.....	44
2.4. MAKİNE ÖĞRENMESİ	50
2.4.1. Destek Vektör Makinesi	50
2.4.2. Naive Bayes Sınıflandırıcısı	51
2.5. DOĞAL DİL İŞLEME	52
2.5.1. Türkçe Dil Yapısı Özellikleri	55
2.5.1.1. Hece Yapısı.....	55
2.5.1.2. Sesli Uyumu.....	56
2.5.1.3. Sessiz Uyumu	57
2.5.1.4. Kelimenin Hecelerine Ayrılması	57
2.5.1.5. Sondan Eklemeli Dil Yapısı	59
2.5.1.6. Cümle Yapısı ve Anlam Vurgusu.....	61
2.5.2. Yazım Yanlışlarının Düzeltilmesi	62
2.5.2.1. Zemberek Kütüphanesi	65

ÜÇÜNCÜ BÖLÜM

UYGULAMA GELİŞTİRME

3.1. ÇALIŞMANIN ÖNEMİ	69
3.2. KAPSAM	69
3.3. MODEL.....	70
3.4. MODÜLLER	71
3.4.1. Tweetlerin Çekilmesi	73
3.4.1.1. Twitter Api Şifresi Alma	73
3.4.1.2. Tweet Çekme Uygulaması.....	74
3.4.1.3. Test Verileri	75
3.4.2. Tweetlerin Düzenlenmesi	76
3.4.2.1. Zemberek Kullanımı ve Tweetlerin Genel Temizliği.....	76
3.4.2.2. Sözcüklerin Hatalı Olma Durumları için Öneriler.....	77
3.4.2.3. Metnin Kelimelerine Ayrıştırılması (Tokenizer).....	80
3.4.2.4. Kelimelerin Hecelerine Ayrıştırılması.....	82
3.4.2.5. Kelimelerin Sözlük ile Eşleştirilmesi	84
3.4.2.6. Kelime Veri Tabanı Tasarımı	87
3.4.2.7. Q ve F Klavye Düzeni ile Hata Tespiti.....	87
3.4.2.8. Python Betiklerinin Yazılması.....	89
3.4.3. Tweetlerin Sınıflandırılması.....	92
3.4.3.1. Eğitim Veri Setinin Oluşturulması	92
3.4.3.2. T-SQL ile Naive Bayes Sınıflandırma.....	93
3.4.3.3. Linux Küme Üzerinde Hadoop Ortamının Oluşturulması.....	97
3.4.3.4. Hadoop Ortamında Hive Kurulumu	102
3.4.3.5. Windows Ortamında Hadoop ve Hive Kurulumu	105
3.4.3.6. HiveQL ile Naive Bayes Sınıflandırma	110
3.4.4. Sonuçların Raporlanması.....	113
3.4.4.1. Marka Karşılaştırmalı Tweetler	114
3.4.4.2. En Çok Okunan Tweetler	115
3.4.4.3. Sık Kullanılan Kelimeler	115
3.4.4.4. Tweet Analiz Sonuçlarının İllere Göre Raporlanması.....	116
3.4.4.5. Tweet Analiz Sonuçlarının İllere ve Tarihe Göre Raporlanması	117

3.4.5. Konfigürasyon Ekranları	117
3.4.5.1. Duygu Belirleyici İfadeler	117
3.4.5.2. Kelime Veri Tabanı	118

DÖRDÜNCÜ BÖLÜM

BULGULAR

4.1. GENEL İSTATİSTİKLER	119
4.1.1. Ek ve Kökler ile İlgili İstatistikler	119
4.1.2. Takipçi Sayılarına Göre Tweet İstatistikleri	123
4.1.3. Okunma Sayılarına Göre Tweet İstatistikleri	124
4.1.4. Favori Yapılma Sayılarına Göre Tweet İstatistikleri.....	125
4.1.5. Konum Bilgilerine Göre Tweet İstatistikleri.....	126
4.1.6. Hashtag (Etiket) İstatistikleri.....	128
4.1.7. Mention (Bahis) İstatistikleri.....	129
4.1.8. Hece İstatistikleri.....	130
4.1.9. Harf İstatistikleri.....	131
4.2. ANALİZ SONUÇLARI	133
4.2.1. Sınıflandırma Başarısı	133
4.2.2. Karşılaştırmalı Analiz.....	133
4.2.3. Tweet Örnekleri.....	134
SONUÇ ve TARTIŞMA	136
KAYNAKÇA	143
ÖZGEÇMİŞ.....	168

ÖZET**DOKTORA TEZİ****TÜRKÇE TWEETLERİN ANALİZ EDİLEBİLMESİ İÇİN
HADOOP/HIVE KULLANAN MELEZ BİR SİSTEMİN GELİŞTİRİLMESİ****Sinan KUL****Tez Danışmanı: Prof. Dr. Üstün ÖZEN****2018, 168 sayfa****Jüri: Prof. Dr. Üstün ÖZEN****Prof. Dr. Uğur YAVUZ****Prof. Dr. Ahmet BARAN****Prof. Dr. Yılmaz GÖKŞEN****Dr. Öğr. Üyesi Mustafa KESKİNKILIÇ**

Günümüzde, tüketiciler, hizmet aldıkları firmalar hakkında beğenilerini ve/veya eleştirilerini e-ticaret sitelerinde veya Facebook ve Twitter gibi sosyal medya ortamlarında paylaşmaktadırlar. Piyasa araştırması yapmak, pazarlama stratejilerini geliştirmek, müşteri memnuniyetini ve firma itibarını ölçmek isteyen firmalar için ise bu paylaşımlar oldukça değerlidir. Bu paylaşımların ayrıca, diğer müşterilerin satın alma davranışını etkilediği ve marka bağlılığını artırdığı da yapılan çalışmalarla ortaya konmuştur.

Bilişim ve iletişim teknolojilerinin ucuzlaması ve yaygınlaşmasıyla sayısı giderek artan sosyal yazışmaların analiz edilmesi ise geleneksel veri analizi tekniklerinin sınırlarını aşmıştır. Bunun için gerek verilerin düşük maliyetli bir şekilde ve anlık olarak toplandığı gerekse işlendiği (doğal dil işleme) ve analiz edildiği (duygu analizi) bilgisayar tabanlı sistemlere ihtiyaç duyulmaktadır.

Bu çalışma ile işletmelerin kendileriyle ve rakipleriyle ilgili paylaşılan Twitter yorumlarını takip etmelerine ve pazardaki konumlarını periyodik olarak yorumlayabilmelerine imkân sağlayan düşük maliyetli bir uygulama geliştirilmesi amaçlanmıştır. Geniş kullanıcı kitlesi ve yaygın kullanımı dolayısıyla da veri kaynağı olarak Twitter seçilmiştir.

Veri kümesi içindeki tüm kelimeler özellik olarak kullanılacağı için veri boyutu çok büyük olacağından, Hadoop kümesi üzerinde Hive betiklerinin koşturulmasına karar verilmiştir. Çalışma kapsamında gerçekleştirilen sistem, özetle dört modülden oluşmaktadır: birinci modül gerekli verilerin periyodik ve otomatik olarak toplanabilmesine imkân vermekte, ikinci modül ile veriler ön işleme tabi tutulmakta; üçüncü ve dördüncü modüllerde ise analiz ve raporlama yapılabilmektedir. Sonuçlar, ilgili sektördeki tüm işletmelerin karşılaştırmalı analizleri ile görselleştirilerek raporlanabilmektedir.

Anahtar Kelimeler: metin madenciliği, doğal dil işleme, Hive, HiveQL, Hadoop, Naive Bayes, makine öğrenmesi, duygu analizi, Twitter, Türkçe, sosyal medya

ABSTRACT**Ph. D. DISSERTATION****DEVELOPMENT OF A HYBRID SYSTEM USING HADOOP / HIVE FOR
ANALYZING TURKISH TWEETS****Sinan KUL****Advisor: Prof. Dr. Üstün ÖZEN****2018, Page: 168****Jury: Prof. Dr. Üstün ÖZEN****Prof. Dr. Uğur YAVUZ****Prof. Dr. Ahmet BARAN****Prof. Dr. Yılmaz GÖKŞEN****Asst. Prof. Dr. Mustafa KESKİNKILIÇ**

Today, consumers share their appreciation or their criticism about the companies they serve at e-commerce sites or social network such as Twitter or Facebook. These comments are quite valuable for companies want to do market research, to develop marketing strategies, to measure customer satisfaction and firm reputation. It has also revealed through studies that these comments affect other customers' buying behavior and increase brand loyalty.

Analysis of increasing number of social correspondence with the cheaper and widespread use of information and communication technologies has already exceeded the limits of traditional data analysis techniques. Therefore, it is a need for computer-based system that the data both at low cost and instantly collected, processed (natural language processing) and analyzed (sentiment analysis).

With this study, it has intended to develop a low-cost application which allows businesses to keep track of shared Twitter comments about themselves and their competitors and allows them to periodically interpret their position in the market. Due to its large user base and widespread using, Twitter has been selected as a data source.

It was decided to run Hive commands on Hadoop cluster since the data size will be very large for all words in the dataset to be used as properties. The developed system in the scope of the study consists of four modules in brief: the first module allows the necessary data to be collected periodically and automatically, the second module allows pre-processing of the data; analysis and reporting can be done in the third and fourth modules. The results can be report by visualizing with comparative analysis of all the firms in the related sector.

Keywords: text mining, natural language processing, Hive, HiveQL, Hadoop, Naive Bayes, machine learning, sentiment analysis, Twitter, Turkish, social media

SİMGELER ve KISALTMALAR DİZİNİ

ACID	: Atomicity Consistency Isolation Durability - Atomiklik Tutarlılık İzolasyon Sağlamlık
ALU	: Arithmetic Logic Unit – Aritmetik Mantık Birimi
API	: Application Programming Interface - Uygulama Programlama Ara yüzü
ASCII	: American Standard Code for Information Interchange - Bilgi Değişimi için Amerikan Standart Kodlama Sistemi
BOW	: Bag of Words – Kelimeler Torbası
CAP	: Consistency Availability Partition tolerance – Tutarlılık Ulaşılabilirlik Bölünebilirlik toleransı
CPU	: Central Processing Unit – Merkezi İşlem Birimi
CRM	: Customer Relationship Management - Müşteri İlişkileri Yönetimi
CSV	: Comma Separated Values – Virgül ile Ayrılan Değerler
DBMS	: Data Base Management System – Veri Tabanı Yönetim Sistemi
DDL	: Data Definition Language - Veri Tanımlama Dili
DML	: Data Manipulation Language - Veri İşleme Dili
DN	: Data Node – Veri Düğümü
EB	: Exa Byte (1 EB= 1024 Peta Byte)
ERP	: Enterprise Resource Planning - Kurumsal Kaynak Planlama
GPS	: Global Positioning System - Küresel Konumlandırma Sistemi
HDFS	: Hadoop Distributed File System – Hadoop Dağıtık Dosya Sistemi
HiveQL	: Hive Query Language – Hive Sorgulama Dili
ID	: Instruction Decode – Komut Çözücü
IE	: Instruction Execute – Komut Yürütücü
IF	: Instruction Fetch – Komut Getirme
JDBC	: Java Database Connectivity – Java Veritabanı Bağlanırlığı
JT	: Job Tracker – Görev Takipçisi
MIMD	: Multiple Instruction Multiple Data – Çoklu Komut Çoklu Veri
MISD	: Multiple Instruction Single Data – Çoklu Komut Tek Veri
NC	: Negative Count – Negatif Sayısı

VIII

NLC	: Negative Listed Count – Nagatif Okunma Sayısı
NLP	: Natural Language Processing – Doğal Dil İşleme
NN	: Name Node – İsim Düğümü
NoSQL	: Not Only SQL – Yalnızca SQL değil
NTRC	: Neutral Count – Nötr Sayısı
NTRLC	: Neutral Listed Count – Nötr Okunma Sayısı
ODBC	: Open Database Connectivity - Açık Veritabanı Bağlanırlığı
OLAP	: Online Analytic Processing - Çevrimiçi Analitik İşleme
OLTP	: Online Transaction Processing – Çevrimiçi İş yığını İşleme
PB	: PetaByte (1 PB= 1024 Tera Byte)
PC	: Positive Count – Pozitif Sayısı
PLC	: Positive Listed Count – Pozitif Okunma Sayısı
SIMD	: Single Instruction Multiple Data - Tek Komut Çoklu Veri
SISD	: Single Instruction Single Data - Tek Komut Tek Veri
SS	: Sert Sessizler
SSH	: Secure Shell – Güvenlik Kabuğu
SQL	: Structured Query Language – Yapısal Sorgulama Dili
SVM	: Support Vector Machine – Destek Vektör Makinesi
T-SQL	: Transact Structured Query Language – İş Gören Yapısal Sorgulama Dili
TT	: Task Tracker – İş Takipçisi
WB	: Write Back – Geri yaz
VTYS	: Veri Tabanı Yönetim Sistemi
XML	: eXtensible Markup Language - Genişletilebilir İşaretleme Dili
VM	: Virtual Machine – Sanal Makine
YSB	: Sert karşılığı bulunan yumuşak sessizler
YSM	: Sert karşılığı bulunmayan yumuşak sessizler
ZB	: Zetta Byte (1 ZB= 1024 Exa Byte)

ŞEKİLLER DİZİNİ

Şekil 1.1. Tweet Görünümü	9
Şekil 1.2. Verinin Büyüme Hızı	12
Şekil 1.3. Toplam Yatırımlar ve Saklama Maliyetleri	13
Şekil 1.4. 2018 Yılı İtibariyle İnternette 1 Dakikada Ne Oluyor?	14
Şekil 1.5. Seri Yürütme ve Boru Hattı (Pipeline) Yürütme	15
Şekil 1.6. Akademik Çalışmalarda Büyük Veri – Dünya Çapında.....	17
Şekil 1.7. Akademik Çalışmalarda Büyük Veri – Türkiye Özelinde.....	17
Şekil 1.8. Veri Çeşitliliği	18
Şekil 1.9. Json Format.....	26
Şekil 1.10. Anahtar-Değer Veritabanı	28
Şekil 1.11. Çizge Veritabanı	28
Şekil 1.12. Veri Madenciliği Süreçleri	29
Şekil 2.1. Namenode ve Datanode Yapıları.....	32
Şekil 2.2. Rack Yapısı.....	33
Şekil 2.3. HDFS Kümesi.....	33
Şekil 2.4. HDFS Mimarisi	34
Şekil 2.5. Eşle İndirge Kelime Sayma Şeması	36
Şekil 2.6. Eşle İndirge Algoritmasının Çalışma Şekli	36
Şekil 2.7. Hive Mimarisi.....	40
Şekil 2.8. Kelime Sayma Örneği	44
Şekil 2.9. BOW Modeli Örneği	46
Şekil 2.10. Destek Vektör Makinesi	51
Şekil 2.11. Büyük (üstte) ve Küçük Ses Uyum (altta) Durum Makineleri	56
Şekil 2.12. Sessiz Uyum Durum Makinesi	57
Şekil 2.13. Heceleme Algoritması Akış Diyagramı	58
Şekil 2.14. “Ruhsatlandırılmamasındaki” Kelimesindeki Ekler	59
Şekil 2.15. Soldan İsim Çekim Eki Sonlu Durum Makinesi	61
Şekil 2.16. Bağlılık İlişkileri.....	62
Şekil 2.17. Her Bir Dil için Zemberek Dosya ve Sınıfları	65
Şekil 2.18. Morfolojik Ayırıştırıcı	66
Şekil 2.19. Zemberek Kök Ağacı Yapısı	67

Şekil 2.20. Kök Ağacı Yapısı, Kitap Kelimesi Örneği	67
Şekil 3.1. Önerilen Model.....	70
Şekil 3.2. Twitter Uygulama Yönetim Sayfası	73
Şekil 3.3. Erişim Şifreleri	73
Şekil 3.4. Tweet Çekme Uygulaması.....	74
Şekil 3.5. Tweet Çekme Kod Satırı	75
Şekil 3.6. Alfa Nümerik Olmayan Karakter Örnekleri	77
Şekil 3.7. Ön İşlem ve Kaydetme Ekranı.....	77
Şekil 3.8. Kelimenin Hatalı Olma Durumunun Durum Makinesi ile Gösterimi	80
Şekil 3.9. Tweetin Kelimelerine Ayrılması (C# Kodu)	80
Şekil 3.10. Tweetin Kelimelerine Ayrılması (Python Kodu)	80
Şekil 3.11. Tweetin Kelimelerine Ayrılması (T-SQL Kodu)	81
Şekil 3.12. Bitişik Kelimelerin Ayrılması (T-SQL Kodu).....	81
Şekil 3.13. Kelimelerin Köklerine Ayrılması (C# Kodu).....	81
Şekil 3.14. Kelimenin Hecelerine Ayrılması (C# Kodu).....	82
Şekil 3.15. Kelimenin Hecelerine Ayrılması (T-SQL Kodu)	82
Şekil 3.16. Kelimenin Hecelerine Ayrılması Algoritması.....	83
Şekil 3.17. Kelimelerin Sözlük ile Eşleştirilmesi (T-SQL Kodu)	84
Şekil 3.18. Kelimelerin Sözlük ile Eşleştirilmesi (Python Kodu)	84
Şekil 3.19. Kelimelerin ASCII Karşılıklarıyla Eşleştirilmesi (Python Kodu).....	84
Şekil 3.20. Kelimelerin ASCII Karşılıklarıyla Eşleştirilmesi (T-SQL Kodu).....	84
Şekil 3.21. Son Harfi Düzeltme Kodu (T-SQL)	85
Şekil 3.22. Kelimelerin Sözlükte Var Diye İşlenmesi (T-SQL)	85
Şekil 3.23. Kelime Düzeltme Algoritması Akış Diyagramı	86
Şekil 3.24. Kelime Veri Tabanının Oluşturulması.....	87
Şekil 3.25. Tweet Dosyasına Erişim	89
Şekil 3.26. Tweet Metnini Elde Etme	89
Şekil 3.27. Tweet İçindeki Emojileri Ayırma.....	89
Şekil 3.28. Tweet İçindeki Nümerik İfadeleri Silme	90
Şekil 3.29. Tweet İçindeki Fazla Boşlukları Teke İndirme	90
Şekil 3.30. Tweeti Kelimelere Ayırma	90
Şekil 3.31. Kelime Hashtag, Mention, Link, vs. Değilse.....	90

Şekil 3.32. Kelimenin Sözlükte Aranması.....	90
Şekil 3.33. Harflerin ASCII Karşılığının Alınması	91
Şekil 3.34. Kelimelerin Türkçe Harf Kullanılmadan Sözlükte Aratılması.....	91
Şekil 3.35. Harflerin Sert Yumuşak Karşılıklarının Alınması.....	91
Şekil 3.36. Kelimenin Son Harfinin Sert Yumuşak Karşılıkları ile Aratılması.....	91
Şekil 3.37. Çoğullayan Harflerin Teke Düşürülmesi.....	92
Şekil 3.38. Kelimeler Tablosunun Oluşturulması.....	94
Şekil 3.39. “TweetKelimeler” Tablosu ile “Kelimeler” Tablosunun Eşleştirilmesi.....	94
Şekil 3.40. “TweetFrekans” Tablosunun Oluşturulması	94
Şekil 3.41. Kelimelerin Tweetlerde Geçme Sayılarının Alınması.....	95
Şekil 3.42. Tweet Sayılarının Elde Edilmesi	95
Şekil 3.43. Kelime Sayılarının Elde Edilmesi	95
Şekil 3.44. Kelime Olasılıklarının Hesaplanması	95
Şekil 3.45. Tweet Olasılıklarının Hesaplanması.....	96
Şekil 3.46. Sınıflara Ait Olma Olasılıklarının Hesaplanması	96
Şekil 3.47. Olasılık Değerlerinin Hesaplanması	96
Şekil 3.48. Tweetlere Duygu Atanması	97
Şekil 3.49. Linux Paketlerinin Güncellenmesi	97
Şekil 3.50. Java kurulumu.....	97
Şekil 3.51. Mevcut Paketlerin Güncellenmesi	97
Şekil 3.52. SSH Kurulumu	98
Şekil 3.53. Rsync Paketinin Kurulumu.....	98
Şekil 3.54. Hadoop Kullanıcısının Tanımlanması	98
Şekil 3.55. Root Kullanıcısının Yetkilerinin Bulunduğu Satır	98
Şekil 3.56. Hadoop Kullanıcı Yetkileri	98
Şekil 3.57. “opt” ve “srv” Klasör Yetkileri.....	98
Şekil 3.58. .bashrc Dosyasını Düzenlemek.....	99
Şekil 3.59. .bashrc Dosyasına Eklenecek Düzenlemeler	99
Şekil 3.60. .bashrc Dosyasının Hafızaya Alınması.....	99
Şekil 3.61. Hadoop Kurulum Dosyası İndirilmesi.....	100
Şekil 3.62. Hadoop Kurulum Dosyasının Açılması.....	100
Şekil 3.63. Hadoop Kurulum Klasörünün Kopyalanması	100

Şekil 3.64. Host Dosyasının Düzenlenmesi	100
Şekil 3.65. Şifre Oluşturulması	100
Şekil 3.66. Şifrenin Diğer Makinelere Gönderilmesi	100
Şekil 3.67. Diğer Makinelere Girişin Test Edilmesi.....	101
Şekil 3.68. Ana Makinenin Başlatılması.....	101
Şekil 3.69. Dosya Sisteminin Başlatılması	101
Şekil 3.70. Dosya Sisteminin Sonlandırılması.....	101
Şekil 3.71. Yarn'ın Başlatılması	101
Şekil 3.72. Jps Komutu ile Çalışan Hizmetlerin Listelenmesi.....	101
Şekil 3.73. Hadoop Web Ara Yüzü	102
Şekil 3.74. Datanode Bilgileri.....	102
Şekil 3.75. Hive Kurulum Dosyasının İndirilmesi	103
Şekil 3.76. Hive Kurulum Dosyasının Açılması.....	103
Şekil 3.77. Hive Klasörünün Oluşturulması ve Yetkilendirilmesi	103
Şekil 3.78. Hive Klasörünün Taşınması	103
Şekil 3.79. .bashrc Dosyasının Açılması	103
Şekil 3.80. Değişken Tanımlamaları.....	104
Şekil 3.81. Değişkenlerin Hafızaya Alınması.....	104
Şekil 3.82. HDFS için Klasör Oluşturulması.....	104
Şekil 3.83. hive-config Dosyasının Açılması	104
Şekil 3.84. Hadoop_Home Değişkeninin Tanımlanması.....	104
Şekil 3.85. Hive Komutunun Çalıştırılması.....	104
Şekil 3.86. Cihazı İçeri Aktar Bağlantısı	105
Şekil 3.87. Oracle VM Virtualbox Yöneticisi Ekranı.....	106
Şekil 3.88. Sandbox 2.0 Sanal Makinesi Ekranı	106
Şekil 3.89. Sorgu Geçmişi Ekranı.....	107
Şekil 3.90. Hive ile Oluşturulan Tabloların HDFS Klasörleri.....	108
Şekil 3.91. “Tweetkelimeler” Tablosu için Dosya Konumu	109
Şekil 3.92. “Tweetkelimeler” Dosyası İçeriği	109
Şekil 3.93. “tweetkelimeler” ve “sozluk” Tablolarının Oluşturulması (HiveQL kodları)	110
Şekil 3.94. Sözlük Tablosu için Veri Dosyası İçeriği.....	110

Şekil 3.95. TweetView Görünümü Oluşturulması.....	111
Şekil 3.96. Tweet Sayılarıyla İlgili Tabloların Oluşturulması	111
Şekil 3.97. Olasılık Tablolarının Oluşturulması	111
Şekil 3.98. Kelme Sayılarıyla İlgili Tabloların Oluşturulması	112
Şekil 3.99. “TweetOlasilik” tablosunun oluşturulması.....	112
Şekil 3.100. “TweetOlasilikSon” Tablosunun Oluşturulması	113
Şekil 3.101. “TweetDuygu” Tablosunun Oluşturulması	113
Şekil 3.102. Tweet Analiz Modülü Ana Ekranı.....	114
Şekil 3.103. Karşılaştırmalı Tweetler Ekranı.....	114
Şekil 3.104. Ençok Okunan Tweetler Ekranı.....	115
Şekil 3.105. Sık Kullanılan Kelimeler Ekranı	116
Şekil 3.106. Tweet – Konum Ekranı.....	116
Şekil 3.107. Tweet – Konum Ekranı (Belirli Bir Tarih için)	117
Şekil 3.108. Duygu Belirleyici İfadeler Ekranı	118
Şekil 3.109. Kelime Veri Tabanı Ekranı.....	118

TABLOLAR DİZİNİ

Tablo 1.1. Sosyal Medya Türleri	4
Tablo 1.2. Veri Biçimleri ve İşletme Birimlerine Etkisi	22
Tablo 1.3. Temel DDL ve DML Komutları	24
Tablo 1.4. Büyük Veriye Kadar Olan Gelişim Süreci	30
Tablo 2.1. Yaygın Kullanılan HDFS Komutları	35
Tablo 2.2. Eşle İndirge ve Spark Karşılaştırması	37
Tablo 2.3. Hadoop ile İlişkili Projeler	38
Tablo 2.4. Veri Tipleri	41
Tablo 2.5. Aritmetik, Mantık ve İlişkisel Operatörler	41
Tablo 2.5. Hece Yapıları	55
Tablo 2.6. Cümle Öğelerinin Dizilişinin Anlam Üzerinde Etkisi	61
Tablo 3.1. Tweet Bilgileri	75
Tablo 3.2. Sözcüklerin Hatalı Olma Durumları ve Örnekler	77
Tablo 3.3. Q Klavye Düzeni Tablosu	87
Tablo 3.4. Duygu Belirleyen Emojiler	93
Tablo 4.1. Kökler ile İlgili İstatistikler	119
Tablo 4.2. En Çok Kullanılan Kelime Kökleri	120
Tablo 4.3. Kelime Kullanım İstatistikleri	120
Tablo 4.4. Ekler ile İlgili İstatistikler	121
Tablo 4.5. Takipçi Sayılarına Göre Kullanıcı İstatistikleri	124
Tablo 4.6: Okunma Sayılarına Göre Tweet Sayıları	124
Tablo 4.7. En Çok Okunan Tweetler	125
Tablo 4.8. Favori Yapılma Sayılarına Göre Tweetler	126
Tablo 4.9. Konum Bilgisine Göre Tweet Sayıları	126
Tablo 4.10: Hashtag (Etiket) Kullanım İstatistikleri	129
Tablo 4.11. Mention (Bahis) Kullanım İstatistikleri	129
Tablo 4.12. Hece Yapılarına Göre Hece Sayıları	130
Tablo 4.13. En Sık Kullanılan Heceler	131
Tablo 4.14. Harf İstatistikleri	132
Tablo 4.15. Samsung ve Apple Aylık Tweet Sayıları	133
Tablo 4.16. Samsung ve Apple Haftalık Tweet Sayıları	134

ÖNSÖZ

Doktora eğitimim sürecinde ve tezimin her aşamasında motive edici yaklaşımları ve desteklerinden dolayı tez danışmanım Prof. Dr. Üstün ÖZEN'e ve değerli katkıları için tez izleme komitesi üyelerine teşekkür ederim.

Ayrıca bu süreçte desteği ve sevgisi ile her zaman yanımda olan değerli anneme, babama, eşime ve kardeşlerime çok teşekkür ederim.

Oyun zamanlarından çaldığım Emir Hattab'a ve Mehlika Sultan'a sevgilerimle...

Erzurum-2018

Sinan KUL



GİRİŞ

İşletmeler üzerine yapılan bir araştırma, sosyal medyayı aktif olarak kullanan işletmelerin net satış ve net kâr değerlerinin ve piyasa değeri ortalamalarının sosyal medyayı kullanmayanlara göre daha yüksek olduğunu ortaya koymuştur (Özmen ve Villi, 2014). Başka bir çalışma ise işletmelerin, sosyal medyayı aktif olarak kullanmasının ve sosyal medyada müşteriler tarafından yapılan paylaşımları analiz etmesinin, işletme başarısını artırdığını belirlemiştir (Atalık, Sak, Bayrak ve Karataş, 2014).

İşletmeler büyük veri analizi için teknolojik yatırım yapmanın ve veri analisti istihdam etmenin maliyet hesabını yaparak işi ağırdan alırken, sektördeki diğer işletmelerin müşteri odaklı olması ve büyük veri teknolojilerine yatırım yapması da işletmeleri baskı altına almaktadır. İşletme üst yöneticileri de bu yatırımın kaçınılmaz olduğunu bilseler de ne zaman ve ne ölçüde yatırım yapmaları gerektiğini kestirememektedirler (Davenport, 2014). Yani öncelikle hangi teknolojiye ihtiyaç duyduklarına ve yatırımın büyüklüğüne karar vermeleri gerekmektedir.

İşletmeler, Twitter gibi sosyal medya sitelerindeki paylaşımları analiz ederek, müşteri memnuniyetini ölçebilecekleri, pazar araştırması yapabilecekleri bir sisteme ihtiyaç duymaktadırlar. Sosyal medya analizi ile özelde ürün ve hizmetlerin takip edilebilmesi; genelde marka bağlılığı hakkındaki müşteri hissiyatlarının değişen trendinin takip edilebilmesi mümkün olabilecektir. Ayrıca işletmeler, gerek fiyat optimizasyonu gerekse risk yönetimi ve yeni tesis için yer belirlenmesi gibi konularda stratejik kararlar alırken analiz sonuçlarından faydalanabilecektir.

Bu çalışma kapsamında, Twitter verilerinin toplanarak, ön işleme tabi tutulduğu ve duygu durumlarının analiz edildiği bir sistem tasarlanması amaçlanmıştır. Türkiye özelinde ve belirli bir sektörde faaliyet gösteren firmalar hakkında paylaşılan tweetler toplanıp analiz edilerek, firmaların zamansal ve mekânsal olarak müşteri memnuniyet düzeylerini takip edebilmesi hedeflenmiştir.

Önerilen sosyal medya duygu analizi sisteminin kullanımının, işletmenin kendisi ve rakipleri hakkında paylaşılan tweetler üzerinden pazardaki konumunun dönemsel ve bölgesel olarak tespit edilmesine, piyasa trendinin takip edilmesine katkı sağlayacağı ve bir tür kıyaslama (benchmarking) olanağı sunacağı öngörülmektedir.

Özetle, bu çalışma ile müşteri yorumlarının analizinin önemli olduğu sektörler için veri analizi sürecinde, işletme yönetimi ve veri analisti dışında kullanıcı müdahalesine ihtiyaç duymayan ve düşük maliyetli teknolojik gereksinimleri olan bir sistem gerçekleştirilmiştir. Sistemin test edilebilmesi için de Türkiye’de faaliyet gösteren telefon markaları hakkında paylaşılan Türkçe tweetler kullanılmıştır.

Çalışmanın ilk bölümünde, kavramsal temeller dâhilinde, internetin yaygınlaşmasıyla kullanımı her geçen gün artan sosyal medya kullanımının işletmelere ve tüketicilere etkileri incelendi. Sosyal medya analizlerinin bugünü ve geleceği olan büyük veri kavramına giriş yapıldı. Yüksek hesaplama gerektiren işler için kullanılan çözümlere (paralel bilgisayarlar), veri depolama yöntemlerindeki değişikliklere (NoSQL yapılar) ve veri analizi tekniklerindeki değişikliklere (büyük veri analizi) ve tarihsel arka planlarına değinildi.

Materyal ve yöntem bölümünde büyük veri için yaygın olarak kullanılan depolama (HDFS), hesaplama (Hadoop Eşle İndirge / Apache Spark) ve programlama (HiveQL) teknolojilerinin kavramsal temelleri anlatıldı. Sosyal medya analizi kapsamında duygu analizi ve makine öğrenmesi yöntemleri incelendi. Yazım hatalarının giderilebilmesi için doğal dil işleme kapsamında yapılan çalışmalar ve Türkçe dilinin gramer yapısı incelendi; morfolojik (biçimbilimsel) ayrıştırıcısıyla doğal dil işlemede sıklıkla kullanılan Zemberek kütüphanesi tanıtıldı.

Çalışmanın üçüncü bölümünde sosyal medya analizi için 4 adet uygulama modülünden oluşan bir model önerildi. Twitter Api’leri kullanılarak geliştirilen tweet çekme uygulaması ve çekilen tweetlerin ön işlenebilmesi için geliştirilen modüller anlatıldı. Ön işlem modülü geleneksel veri tabanı sorgulama dili olan TSQL ile ve python dili ile nasıl kodlanabileceği tecrübe edildi.

Tweet metninin kelimelerine ve hecelerine ayrılması için geliştirilen algoritmalar ve kodlar; kelimelerin doğru yazılıp yazılmadıklarının test edilmesi için arattırılacak sözlük tablosunun oluşturulması için eski çalışmaların derlenmesi, tweetlerin düzenlenmesi bölümünde anlatıldı. Tweetlerin sınıflandırılması bölümü altında ise eğitim veri setinin oluşturulması için hazırlanan duygu belirleyiciler ve durak kelimeleri tablolarından ve duygu emojilerinin kullanılmasından bahsedildikten sonra Naive Bayes sınıflandırma için yazılan TSQL ve HiveQL kodları ve geliştirilen ekranlar tanıtıldı. Bu

bölümde ayrıca, Hadoop ortamının Linux ve Windows ortamlarına sanal ve fiziksel olarak kurulumları etüd edildi. Dördüncü ve son bölümde ise çalışmaya ilişkin bazı istatistiksel sonuçlar tartışıldı.



BİRİNCİ BÖLÜM

KURAMSAL TEMELLER

1.1. SOSYAL MEDYA

İnsanların kendi aralarında, uluslararası ölçekte, mesajlaşabilmeleri, Usenet'in 1979 yılında geliştirilmesiyle birlikte gerçekleşmiştir. Sosyal medya ise ilk olarak Bruce ve Susan Abelson tarafından geliştirilen ve Open Diary ismini verdikleri sosyal paylaşım sitesi ile doğmuştur (Atalık, Sak, Bayrak ve Karataş, 2014). Günümüzde ise sosyal medyanın çok çeşitli uygulamaları bulunmaktadır (Tablo 1.1).

Tablo 1.1. Sosyal Medya Türleri (Dawley, 2009)

<i>Tür Adı</i>	<i>Örnekler</i>
<i>Sosyal Ağ</i>	Twitter, Facebook, MySpace, Friendster, Hi5, Friendfeed, Formspringa, Xing
<i>Fotoğraf Paylaşım</i>	Flicker, PhotoBucket, 1x, Deviantart, photosig, Fotocommunity, Photo, Photodom, Instagram
<i>Video Paylaşım</i>	YouTube, Dailymotion, Google Videos, Yahoo Video, It's On.
<i>Blog</i>	LinkedIn, Ning, Blogger.com, Wordpress, Cnet, TheHuffington Post, BoingBoing, Techcrunch, Kottke
<i>Bilgi Paylaşımı</i>	Wikipedia, Wetpaint, PBWiki, Intelipedia
<i>Sanal Âlem</i>	SL, Active Worlds, There, Whyville, Club Penguin, HiPiHi
<i>İçerik Etiketleme</i>	MERLOT, SLoog

Dülger (2015), büyük veriyi oluşturan kaynakları sıralarken, veriyi bir anlamda, kamusal veri, özel veri, çevresel veri, topluluk verisi ve kişisel veri olarak 5 türe ayırmaktadır. Bu sınıflandırmasıyla Dülger, sosyal medya verilerini, topluluk verisi kategorisi altında zikretmektedir.

İnternet web 1.0 (sabit web sayfaları şeklinde) hayata girdiğinde, kitlelere bilgi paylaşımında kullanılmaktaydı. Web 2.0 ile birlikte ise kullanıcıların da aktif bir şekilde geri bildirim yapmalarına ve kendi aralarında bilgi üreterek paylaşmalarına imkân sağlanmış oldu. Yani internet kullanıcıları tüketicilikten üreticiliğe geçtiler (Aydoğan ve Başaran, 2012). Dahası fikirleriyle üretim süreçlerine katkıda bulunabildiler. Bunun bir örneği, İstanbul Büyükşehir Belediyesi'nin “sizin renginiz hangisi” kampanyasıyla vatandaşlardan aldığı dönüşler ile otobüslerin rengini erguvan olarak belirlemesidir (Karakulakoğlu ve Uğurlu, 2015).

İnterneti aktif olarak kullanan müşteriler, satın aldıkları ürün ve hizmetlerle ilgili sosyal medyada yorum yaparak, varsa sorunları dile getirerek potansiyel müşterileri yönlendirebilmektedirler. Böylece müşteriler, işletme satışlarını belirleyen, işletme itibarını takdir eden ve hatta işletmenin hayatta kalmasını belirleyen bir konuma yükselmiştir (Atalık, Sak, Bayrak ve Karataş, 2014).

İşletmelerin rekabetteki başarısı için daha fazla ürün ve hizmet satışı anlayışı, yerini ulaşılan müşteri sayısına bırakmışken, müşteri odaklı pazarlama stratejileri önem kazanmıştır. Müşterilerle yakınlaşma sağlayarak onların istek ve beklentilerini daha iyi öğrenmek isteyen ve böylece kendilerini hedef kitleye daha iyi anlatmak isteyen işletmeler için sosyal medya önemli bir pazarlama kanalıdır (Çağlıyan, Hassan ve Işıklar, 2016). Geleneksel pazarlama kanalları ile kıyaslandığında ise çok geniş müşteri kitlesine erişim avantajının yanı sıra, sosyal medya oldukça düşük maliyetlidir (Çağlıyan vd., 2016).

1.1.1. Sosyal Medya Kullanımının Yaygınlaşma sebepleri

Çoğunlukla sohbet ve eğlence için kullanılan sosyal medya, internetin hızlanması ve akıllı cep telefonlarının yaygınlaşmasıyla birlikte daha geniş kitlelere ulaşmıştır. Sosyal medya kullanımını yaygınlaştıran diğer bir etmen ise kullanıcıların iletişime geçmekte veya iletişimi sona erdirmekte özgür olmaları ve sosyal medya üzerinden ortak ilgi alanlarına göre gruplara dâhil olarak sosyalleşebilmeleridir (Barutçu ve Tomaş, 2013; Michelson and Macskassy 2010). Dahası, sosyal medya sayesinde kullanıcılar, varolma, görünür olma, tanınma ve beğenilme gibi ihtiyaçlarını karşılayabilmektedir.

2014 yılında yapılan ilk çalışmaya göre; Türkiye'de internet kullanıcılarının %96'sı, ortalama 10 saat sosyal medya kullanmaktadır. Sosyal medyanın bu denli aktif kullanılması ile iştahı kabaran işletmeler, politikalarına yön vermeden önce sosyal medya okuması yapmaktadırlar (Gökkurt ve Demirtel, 2014).

1.1.2. Sosyal Medyanın Temel Özellikleri

Sosyal medyanın beş temel özelliğinden bahseden Kocabaş (2016), bu özellikleri katılımcılık, açıklık, karşılıklı iletişim, topluluk oluşturma ve bağlantılılık olarak listelemektedir. Katılımcılık özelliği, interneti olan herkesin kullanıcı hesabı açarak

sosyal medyaya dâhil olabilmesini ifade ederken, açıklık özelliği katılımın ve geri bildirimin (yorum ve oylama gibi) şeffaflığını anlatır. Karşılıklı iletişim özelliği ise sosyal medyanın, çift yönlü iletişimi desteklemesidir. Topluluk oluşturma özelliği sayesinde ise etkili iletişim için grup oluşturulabilmektedir. Son olarak bağlantılılık özelliği, diğer sitelerle ve kaynaklarla bağlantı kurulabilmesini ifade eder (Kocabaş, 2016).

1.1.3.Sosyal Medya Kullanımının Avantaj ve Dezavantajları

Ticari kurumlar, sosyal medya aracılığıyla reklamlarını, promosyon ve indirim kampanyalarını, geniş kitlelere duyurabilmek için önemli bir imkân elde etmişlerdir. Kamu kurumları ise daha çok bilgi/haber verme; katılıma ve işbirliğine teşvik etme gibi amaçlarla sosyal medyayı kullanmaktadır (Köseoğlu ve Demirci, 2017). Sosyal medyanın, yukarıdan aşağıya doğru mesaj iletim ortamı olmasının yanısıra; bireylerin kendi arasında da iletişim kurabildikleri bir ortam olması, viral reklamlığın da önünü açmaktadır. Ancak, sosyal medya sayesinde olumlu reklamlar kadar olumsuzlar da hızlı bir şekilde geniş kitlelere ulaşabilmektedir. İşletme kontrolü dışında gerçekleşen bu durum, fırsatlar kadar tehditleri de beraberinde getirmektedir. İşletmeler, tüketici eleştirilerini dikkate alarak ürün ve reklam kampanyalarında değişikliğe gidebilmektedir. Ayrıca sosyal medya yazışmalarının kayıtlı (kalıcı) olması yönü dolayısıyla, sosyal medya, pazar hakkında ve rakip işletmeler hakkında araştırma yapılabilmesi için bir veri kaynağı teşkil etmektedir (Çağlıyan vd., 2016).

Sosyal medya manipülatif mesajların yayılabilmesi için uygun bir ortam olması yönüyle de ciddi riskler içermektedir. Geniş kitleler önünde gerçekleştirilen bazı haksız tüketici eleştirileri ve sabote edici mesajlar mesela, en kısa sürede savuşturulamazsa, kurum imajı yara alabilecektir. Sosyal medya kullanan kurum çalışanları ise bilerek veya bilmeyerek, kurumun mahrem bilgilerini tehlikeye düşürebilecek tutum ve davranış içine girebilmektedirler. Sosyal medyada gittikçe artan kirlilik ise, pazarlama kampanyalarının hedef kitleye ulaşp ulaşmadığı konusunu bulanıklaştırmaktadır (Kocabaş, 2016).

1.1.4.Sosyal Medyanın İşletmelerde Kullanımı

1980' lerde işletmelerin belirli bir ürünü üretmek ve müşteriye ulaşımını sağlama ihtiyacına cevap olarak, müşterileri, dağıtım merkezlerini, tedarikçileri ve üretim

süreçlerini bir platformda toplayabilen ERP(Enterprise Resource Planning) sistemleri doğdu. İşletmeler, kendileri için doğru müşteriye aramaya başladıklarında ise, CRM (Customer Relationship Management-Müşteri ilişkileri Yönetimi) sistemleri devreye girdi ve artık ürüne göre müşteri değil, müşteriye göre ürün devri başlamış oldu (Gattiker, Chen ve Goodhue, 2005; Nwankpa, 2018).

Sosyal medyanın doğuşuyla birlikte resmi web siteleri, misyonunu, sosyal medyaya bırakmaya başlamış olsa da, Salman ve Eroğlu (2017) tarafından, Türkiye'nin en büyük 500 şirketi üzerine yapılan bir araştırmaya göre kurum web sitesi, müşteriyle iletişimde hâlâ en yaygın kullanılan iletişim kanalıdır. Kurumların resmi web sitelerinin kullanılabilirliğinin, görsel ve metinsel olarak etkili içeriklere sahip olmasının, ziyaretçiler üzerinde olumlu etki yaratması, güven telkin etmesi ve marka bağlılığı oluşturmada önemli bir faktör olmasına rağmen (Hacıfendioğlu, 2014), kurumların, resmi web siteleri üzerinden müşterilerle iletişim kurma düzeyi oldukça düşük seviyelerde seyretmektedir (Salman ve Eroğlu, 2017).

Müşterileriyle yakın ilişki kurmak isteyen ve böylece müşterilerinin memnuniyet, şikâyet, istek ve önerilerini işletme faaliyetlerine geri bildirim olarak almak isteyen işletmeler ise çift yönlü iletişime izin veren sosyal medya kanalını kullanmaktadırlar. İşletme, yine bu kanal sayesinde takipçilerini fazla bilgiye boğmadan, hikâyesini paylaşabilmekte ve gerektiğinde ise kullanıcılarla diyaloga geçebilmektedir (Twitter, 2017a). İşletme ayrıca, sosyal medyada takipçileri arasında paylaşılacak reklam ve kampanyaların tetikleyicisi ve teşvikçisi olabilmektedir (Barutçu ve Tomaş, 2013).

Kurum içi inovasyonu tetikleyecek bilginin elde edilmesinde aktif rol oynayan “kitle kaynak kullanımı” kavramı da sosyal medya kullanımını teşvik etmektedir. İnsanların farklı konulardaki bilgi ve deneyimlerini elde etmeyi amaçlayan bu yaklaşım, bir tür dış kaynak kullanımıdır ve katılımcılık esasına dayanmaktadır (Banger ve Çalışır, 2014). Marketing Week yazarlarından Steve Hemsley'e göre ise “pazarlamacılar, kapılarının önündeki bu büyük fili görmezden gelmek yerine onunla kucaklaşmalıdırlar” (Demirtaş ve Argan, 2015). Büyük fil olarak tasvir edilense sosyal medya gibi gittikçe büyüyen ve okunması zorlaşan, işletme dışı veri kaynaklarıdır.

İşletmeler, sosyal medya performans takibi için “BoomSonar” gibi bir sistemi kullanarak mesela, yatırımlarının geri dönüşünü alabilmekte; kendileri ve rakipleri (en

fazla 7 işletme) hakkında, bazı kullanım verilerini anlık ve grafiksel olarak takip edebilmektedirler (Barutçu ve Tomaş, 2013).

1.1.5. Sosyal Medyanın Tüketici Davranışlarına Etkisi

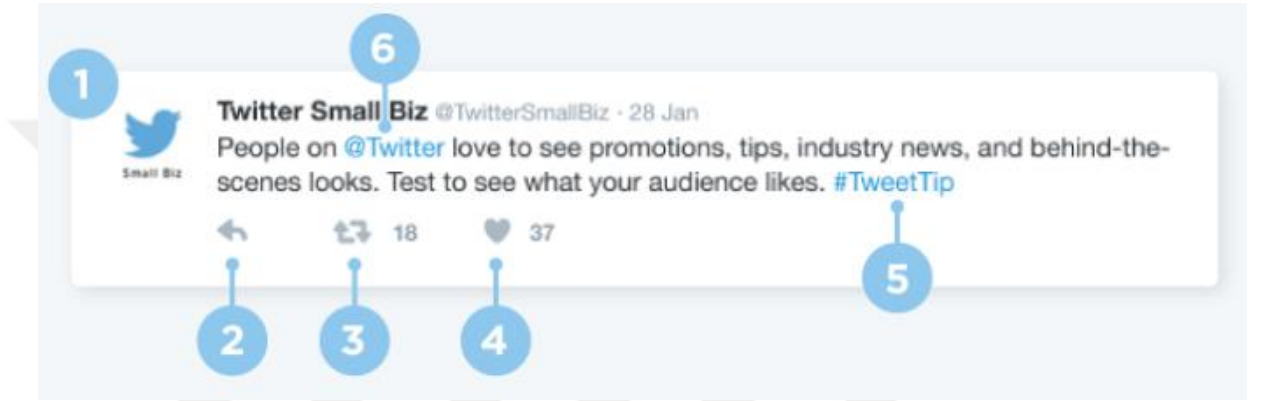
Öğrenme, güdülenme, algılama, tutum ve kişilik gibi psikolojik faktörlerin; kültür, aile, yaşam tarzı ve dâhil olunan sosyal sınıf gibi sosyo kültürel faktörlerin ve cinsiyet, yaş, eğitim, gelir ve statü gibi diğer bazı faktörlerin satın alma davranışını etkilediği bilinmektedir (Keskin ve Baş, 2016). Sosyal medyadaki güçlü iletişim ve etkileşim ağı ise herhangi bir ürünü satın almak isteyen tüketicilerin de ilgi alanındadır. Teknik uzmanlık bilgisine sahip kişilerin önerileri kadar, daha önce ilgili ürünü satın almış olan tüketicilerin deneyimleri de kıymetlidir. Yapılan bir çalışma, tüketicilerin %45' inin aldıkları ürün ve hizmetler hakkında sosyal medya üzerinden yorum yaptıklarını ortaya koymaktadır (Kuş-Khalilov ve Gündebahar, 2014). Paylaşılan olumlu yorumları okuyan kullanıcıların ise ilgili ürün ve hizmete güvenleri artmakta (Kauffman vd., 2010), güven duyduklarında ise satın alma davranışları olumlu etkilenmektedir (Gümüşsoy ve Yeterel, 2016). Sosyal medya kullanıcılarının bilişsellik ve duygusallıkları üzerine yapılan bir çalışma ise tüketicilerin satın alma kararlarını alırken rasyonellikten ziyade duygusallıklarının ön planda olduğunu ortaya koymuştur (Avcıoğlu, 2016).

Yapılan başka bir çalışma ise, müşterilerin %75'inin Twitter üzerinden takip ettikleri işletme hakkında olumlu düşünme eğiliminde olduğunu, %94'ünün ise o işletmenin ürün ve hizmetlerini satın almayı planladığını bulgulamıştır. Twitter' de yürütülen reklam kampanyalarında gördükleri ürünleri satın alma davranışı gösteren müşteri oranı ise %69'dur (Twitter, 2017b; Borthakur, 2008).

1.1.6. Bir Sosyal Paylaşım Sitesi Olarak Twitter

2006 yılında Jack Dorsey, Evan Williams, Biz Stone ve Noah Glass tarafından bir mikro blog sitesi olarak kurulan Twitter, günümüzde insanlar tarafından çok yaygın bir şekilde kullanılmaktadır. Firmalar tarafından da en yaygın kullanılan sosyal medya hesabı Twitter üzerindedir (Salman ve Eroğlu, 2017). Twitter'in, paylaşılacak metinde kısıtlamaya (280 karakter) gitmesine rağmen, bu denli yaygın olmasının başlıca sebepleri, sisteme katılım için ön koşulun bulunmaması, katılım ve kullanım kolaylığı, bir mesajın

aynı anda yüzlerce kişiye ulaştırılabilmesi, sistemdeki kullanıcıların takip edilebilmesinin ve takipçi bulmanın kolaylığıdır. Ortak bir fikri, duyguyu ve ideolojiyi paylaşan kişilerin bir araya getirilebilmesi de Twitter'ı kullanmayı cazip hale getirmiştir (Ünür, 2016). Ayrıca metin yazımında kısıtlamaya gidilmesi, uzun okuma alışkanlığı bulunmayan internet kullanıcıları için bir avantaj da olabilmektedir. Twitter'ın "hiper metin" özelliği sayesinde uzun metinlere (harici web sitelerine) bağlantı kurulabilmesi veya "TwitLonger" benzeri uygulamaların kullanılabilmesi daha detaylı paylaşımlar yapmak isteyenler tarafından aktif kullanılmaktadır (TwitLonger, 2017).



Şekil 1.1. Tweet Görünümü (Twitter, 2016a)

Twitterde, kullanıcıların ana sayfalarında görüntülenen tweet'lerin "zaman akışı" güncelden geçmişe doğru sıralanmaktadır. Yukarıdaki şekilde ise herhangi bir tweet'in olası içeriği çözümlenen ve numaralandırılan her bir bağlantının açıklaması şöyle:

1. Twitter içeriği (yazı, fotoğraf, link ve video içerebilir)
2. Yanıtla: Tweete cevap verilebilir.
3. Retweetle: Tweet bir yorum eşliğinde kullanıcının kendi duvarında paylaşılabilir.
4. Beğen: Tweet beğenilebilir.
5. Başlık etiketi (Hashtag) : “#” sembolü ile başlayıp boşluk karakteri içermeyen ifadedir. Konuşmaları organize eder ve belirli bir konu hakkındaki tüm içeriğe erişmeyi kolaylaştırır. Hashtagli kelimeler popüler olduğunda çok konuşulan konulardan (*trending topics*) olur. Hashtag linki tıklandığında, bu Hashtag ile birlikte paylaşılan bütün tweetler listelenebilmektedir.

6. Söz etmek (*Mention*): Tweet içinde başka birinin ilgisini çekmek için o kişinin kullanıcı adı, “@” işaretiyle birlikte paylaşılır. Kullanıcı adı, Twitter hesabını temsil etmektedir.

Haziran 2016 itibariyle Twitter’in kendi sitesinde paylaştığı verilere göre 313 milyon aktif kullanıcı için 40’ın üzerinde dil desteği verilmektedir. Kullanıcıların %82’si mobil telefon üzerinden erişmektedir (Twitter, 2016b) ve kullanıcılar, günde 500 milyonun üzerinde tweet paylaşmaktadır (Twitter, 2017b). Twitter’in ayrıca kullanıcıların hizmetine sunduğu arşivleme hizmeti sayesinde veriler, CSV formatında yerel bilgisayara indirilebilmektedir.

1.2. BÜYÜK VERİ

1.2.1. Bilgi Yönetiminin Önemi

İnsanlık tarihi üç evreye ayrıldığında, birinci evre tarım toplumunu, ikinci evre sanayi toplumunu ve üçüncü evre de bilgi toplumunu temsil etmektedir. Bilgi toplumu, standartlaşmayı öne çıkaran sanayi toplumunun aksine farklılıklardan beslenmektedir. Bir alandaki uzmanlaşma yerini farklı disiplinlerin ortak çalışmalarına bırakmıştır. Senkronize çalışma zorunluluğu ise yerini zaman ve mekân bağımsız ofis çalışanlarına bırakmıştır. Merkezden kontrol edilen hiyerarşik yapı ise dağıtık yetki ve karar alma mekanizmalarına evrimleşmiştir. Müşteri odaklılık da kişiye özel üretim dönemini getirmiştir (Toffler, 2008). Hatta, bilgi toplumunda, müşteriler, kendi istedikleri ürünlerin üretim sürecine katkı sağlamaya başlamıştır (Gülseçen, 2015).

Bilgi toplumunun en önemli özelliği, bilgiyi yalnızca tüketen değil aynı zamanda üretilmesinde gayret sarf eden bireylerden oluşmasıdır (İ.H. Yücel, 1997). Bu bireyler ise günümüzde büyük oranda dijital yerlilerden oluşmaktadır. Prensky (2001) tarafından kullanılan “dijital yerliler” tabiri 1980 sonrasında bilişim teknolojileri içine doğan, uzaktan ve istedikleri yerden erişebilmeyi seven ve aynı anda birden fazla işe odaklanabilen bireyleri tanımlamaktadır (Prensky, 2001).

Peki, işletmeler bu bilgi denizinin hangi balıklarına oltalarını atmalıdırlar? İşletmeler için ürün ve hizmet kalitesini artıran, maliyeti düşüren ve inovatif değişiklikler getiren bilgi önemlidir. Çünkü günümüz işletmeleri, taşınmaz malları ve sermayeleriyle

değil, ürettikleri ve kullandıkları bilgiler ile rakiplerinden daha üst konuma yükselmektedirler.

İşletmeler, bilgiyi toplamak, işlemek ve yaymak için bilgi sistemlerine ihtiyaç duyarlar ve bilgi sistemlerine 6 stratejik hedefi gerçekleştirmek için yatırım yaparlar: verimlilik sağlamak, yeni ürün hizmet ve iş modeli geliştirmek, tüketici ve tedarikçi yakınlaşması sağlamak, karar vermeyi geliştirmek, rekabetçi avantaj sağlamak ve hayatta kalmak. Bilgi sistemi kaçınılmaz bir zorunluluk halini aldığı anda ise hayatta kalabilmek için işletmeler, bilgi sistemlerine yatırım yapmak zorundadırlar. Finans sektörü buna iyi bir örnektir (K.C. Laudon ve J.P. Laudon, 2004).

Rekabetçi bir piyasa içinde yer alan işletmelerin pazar bilgisine sahip olmaları ve müşterilerin nabzını tutmaları ise bir tür zorunluluktur. İşletmeler, bazı işleri rakiplerinden daha iyi yaparak, müşterilerin isteklerine hızlı ve etkili cevap vererek, daha iyi kalite ve fiyat politikası belirleyerek rekabetçi avantaj sağlayabilmektedirler (Şimşek ve Özdemir, 2012). İşletmeler, herhangi bir ürünü satın alan müşterilerin neden o ürünü veya markayı tercih ettiklerini ve satış sonrası memnuniyet düzeyini belirleyebilmek için, bilgi teknolojilerine yatırım yaparak da rekabetçi avantajı sürdürülebilir kılabilirler (Tekin, Güleş ve Öğüt, 2006).

Bilgi sistemlerine yatırım yapan işletmeler, pazardaki değişimleri, rakiplerin hareketlerini ve pazara dâhil olacak rakipleri (potansiyel rakipleri) tahmin edebilmek, yeni ürün, süreç ve teknolojileri ön görebilmek, yeni pazarlar keşfedebilmek ve rakipler ile kendilerini kıyaslayabilmek imkânını elde edebilmektedirler (Öztürk, 2002). Pazarı yeni sunulan bir ürünün pazar tarafından kabul edilme olasılığının % 15 civarlarında olduğu gerçeği dikkate alındığında, pazara hükmetmenin zorluğu ile beraber zorunluluğu da ortaya çıkmaktadır (Uzkurt, 2010).

Geleneksel bilgi yönetimi daha çok iç kararları desteklemekle ilgilenirken, “büyük veri” yönetimi, stratejik kararların alınmasında daha çok dış veriye odaklanmaktadır (Davenport, 2014). Bir sonraki bölümde dijital ortamda tutulan verinin nasıl büyüdüğü ve büyük veri kavramının nasıl ortaya çıktığı incelenecektir.

1.2.2. Aşırı Bilgi Artışı

İnsanın karar vermesinde ve davranış belirlemede bilgi yol göstericidir. Ancak algılama ve işleme kapasitesinin üzerinde bilgilenmeye maruz kalındığında zihin aktivitelerinde ve karar verme kabiliyetinde düşüş görülmektedir (Taslitz, 2012). Zaten “büyük veri makineler, küçük veri insanlar içindir” (Rouse, 2014) sözü, büyük verinin oluşmasındaki esas etkenin makineler olmasının yanı sıra büyük veriyi analiz edebilecek olanın da yine makineler olduğunu anlatan güzel bir sözdür.

Verinin önemli bir kısmı, makineler tarafından üretilirken, belki lazım olur düşüncesiyle kayıt altına alınabilmesinde depolama maliyetlerinin düşmesinin katkısı büyük. Bilgisayarın icadıyla başlayan ve internetin ortaya çıkışıyla ivmelenen veri artışı, arama motorlarının bilgiye erişimi hızlandırmasının, sosyal ağlar, bloglar ve son olarak mobil teknolojileriyle bulunduğu noktada veri miktarı bugünkü seviyelere ulaştı (Gülseçen, 2015). Verinin küresel ölçekte, bu denli artmasının sebepleri sayıldığında görüldüğü üzere teknolojik gelişmeler başı çekmektedir.

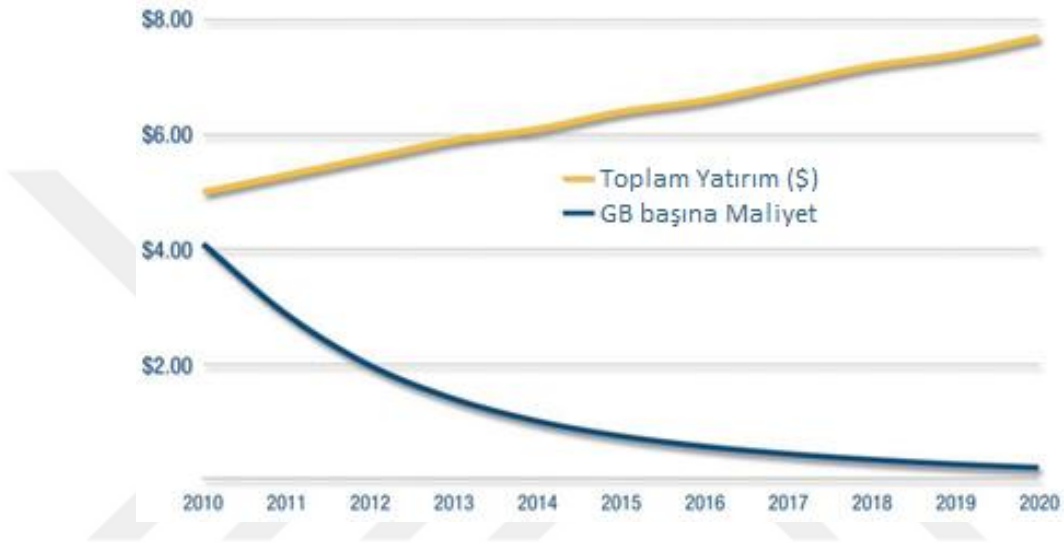


Şekil 1.2. Verinin Büyüme Hızı (Al-Khoury, 2013)

IBM şirketi tarafından yapılan bir çalışmaya göre 2011 yılı itibariyle üretilen verilerin %90'ı son iki yıla ait. Şimdiye kadar oluşan verilerin kapasitesi ölçüsünde yapılacak bir filmin izlenebilmesi için 47 milyon seneye ihtiyaç olması, verinin büyüklüğünü anlatmaktadır (Barske, 2013).

2006 yılı itibariyle dünyadaki bütün bellek birimlerinde toplam 160 Exabyte (1 EB = 2^{20} TB) veri kaydedildiği biliniyor (Al-Khoury, 2013). Google yazılım

mühendislerinden birinin yanlışlıkla 10 Petabyte veri silebildiği (Schutt ve O'Neil, 2013) günümüzde IBM tarafından yapılan bir çalışmaya göre 2020 yılında 40 Zettabayt (1 ZB = 2^{20} EB) veri üretileceği tahmin ediliyor (Rometty, 2014). Seagate firması ise 2016 yılı itibariyle üretilen veri miktarını 16 ZB olarak açıklarken 2025 yılı tahminlerini 163 ZB olarak açıklamaktadır (Seagate, 2018). Başka bir çalışmada ise 2020 yılına kadar bilişime daha fazla harcama yapılacağı ve veri saklama maliyetlerinin düşeceği öngörülmektedir (Şekil 1.3) (Gantz ve Reinsel, 2012). Bu da şüphesiz, veri büyüme ivmesini artıracaktır.



Şekil 1.3. Toplam Yatırımlar ve Saklama Maliyetleri (Gantz ve Reinsel, 2012)

Dijital dünyada temsil edilen her türlü veri bu kapsam içinde değerlendirilmektedir. Peki, verinin bu kadar büyümesinde internetin katkısı nedir? Şekil 1.4'te 2016 yılı itibariyle internette her bir dakikada ne olduğuna ilişkin istatistikler sunulmaktadır.

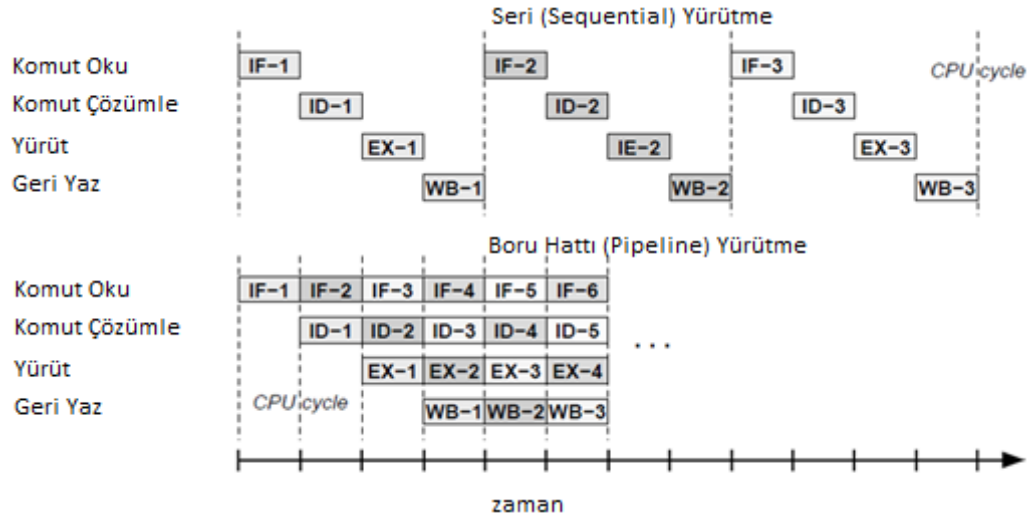


Şekil 1.4. 2018 Yılı İtibariyle İnternette 1 Dakikada Ne Oluyor? (Lewis ve Chadd, 2018)

Bu büyük veri havuzunda boğulmamak içinse sadece gerekli olan verilere odaklanılması şarttır. Bunun için de öncesinde araştırma problemi belirlenmelidir, hipotez kurulmalıdır. Hipotezin test edilmesi için de veriler seçilerek istatistiksel analizler yapılmalıdır (Bayrakçı, 2015).

1.2.3. Seri - Paralel Bilgisayarlar

Von Neumann mimarisine sahip işlemciler, Şekil 1.5' te görüldüğü üzere, yazılım komutlarını “*Fetch*” (getir), “*Decode*” (çözümle), “*Execute*” (yürüt) ve “*WriteBack*” (geri yaz) adımlarını ardışık olarak işleyerek icra ederler. Her bir çevrimde (CPU Cycle) bir adım işletilir. Komut okuma adımında işlem bittikten sonra ise geri yazma adımı son bulana kadar herhangi bir işlem yapılmadan beklenmektedir.



Şekil 1.5. Seri Yürütme ve Boru Hattı (Pipeline) Yürütme (Kıraç, 2015)

“Pipeline” çalıştırma, montaj hattına benzetilebilir. “...CPU'lar her aşama için işlem yollarını birden fazla parçaya (pipeline) ayırarak aynı anda birden fazla komut (instruction) çalıştırabiliyorlar. Her pipeline CPU'nun farklı kısımlarını (ALU, FPU, LSU gibi) farklı zamanlarda meşgul ederek, CPU birimlerinin verimli şekilde kullanılabilmesini sağlıyor. Yeni CPU'ların aynı anda yapabileceği işlem sayısının daha da artırılması için her çekirdeğe birden fazla işlem veya hafıza kontrol birimleri eklenerek pipeline'ların sayısı daha da artırılabilir.” (Kıraç, 2015 s.1)

1.2.4. Paralel Bilgisayar Kümeleri

Paralel bilgisayar kümeleri, yüksek işlem gücü gerektiren işler için hesaplama maliyetini düşürmede bilgisayarların, paralel olarak, yani eşzamanlı çalıştırılarak kullanılmasıdır. Flynn sınıflandırmasına göre bilgisayarlar 4 gruba ayrıldığında son üç grup paralel bilgisayar kümelerini ifade etmektedir (Güneş, 2011; Limet, Merlo ve Spalazzi, 2018).

1. SISD (Single Instruction Single Data - Tek Komut Tek Veri seti): Von Neumann mimarisine sahip paralel olmayan (seri) makineler bu sınıftandır.
2. SIMD (Single Instruction Multiple Data - Tek Komut Çoklu Veri seti): Bir komutun birden fazla makine ve farklı veri setleri üzerinde çalıştırılması esasına dayanır.

3. MISD (Multiple Instruction Single Data - Çoklu Komut Tek Veri seti): Bir veri seti üzerinde birden fazla makinenin farklı komutlar çalıştırması esasına dayanır.
4. MIMD (Multiple Instruction Multiple Data - Çoklu Komut Çoklu Veri seti): Bir veri setinin farklı parçaları üzerinde birden fazla makinenin farklı komutlar çalıştırması esasına dayanır.

1.2.5. Büyük Veri Tanımı

“Veri köken olarak Latince “vermek” anlamındaki “dare” sözcüğünden gelmektedir. İngilizcede “data” anlamına gelen veri kavramı ilk kez on yedinci yüzyılda matematikte kullanılmış olup, daha sonra on sekizinci yüzyılda diğer disiplinlerde de kullanılmaya başlanmıştır.” (Bayrakcı, 2015 s.4)

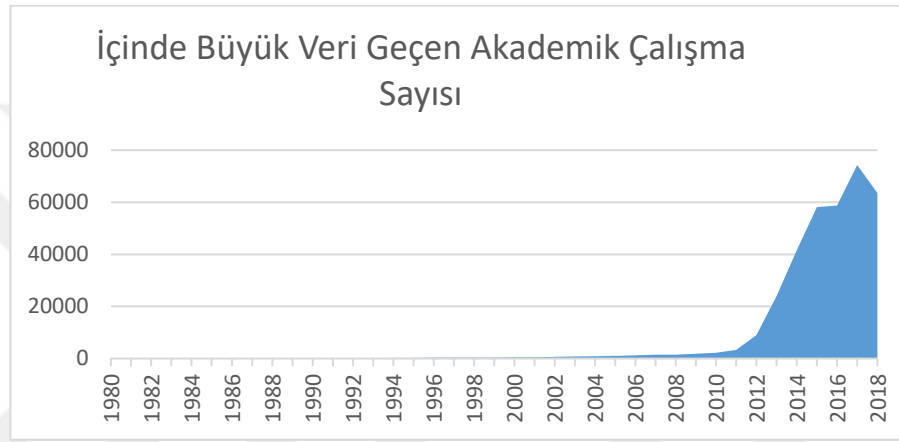
Yılmaz (2009) veriyi, *“tek başına anlam ifade etmeyen veya kullanılamayan, bununla birlikte enformasyona ve bilgiye temel oluşturan ilişkilendirilmeye, gruplandırılmaya, yorumlanmaya, anlamlandırılmaya ve analiz edilmeye gereksinim duyulan ham bilgi”* (s.98) şeklinde tanımlamaktadır.

Önceleri, istenen zaman içinde kolayca işlenemeyen veri setleri için kullanılan büyük veri deyimi ise bugünkü anlamıyla ilk defa 1997 yılında NASA araştırmacıları tarafından (Yıldız ve Topal, 2015) kullanılmıştır. Devenport (2014), büyük veri için tek sunucuya sığmayacak olmasını bir ölçü olarak kabul etmiş ve verinin, büyük veri kabul edilebilmesi için 100 terabyte’tan büyük olması gerektiğini savunmuş. Vinod (2013)’a ve Anuradha (2015) göre ise büyük veri, kapladığı alan itibariyle PB (PetaByte) ve üzeri kavramlarla ifade edilen büyüklükteki verileri anlatmaktadır. Ancak büyük veri kapsamında verinin boyutunu belirli sayılar ile kısıtlamak geçerliliğini kaybetmektedir. Çünkü bugün büyük diye bilinen veri miktarı gelecekte pek büyük görülmeyebilecektir. Zira yakın geçmişte büyüklüğü göz dolduran veriler artık büyük değil. Ayrıca verinin hacmi farklı sektörlerde farklı kapsamlarda değerlendirilebilmektedir.

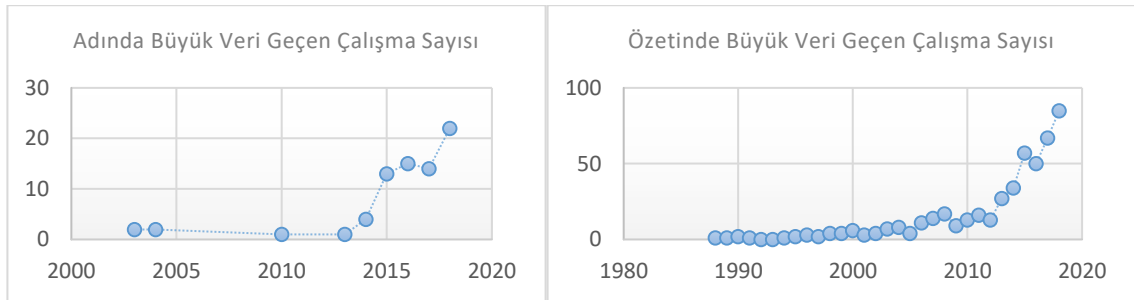
Büyük veri kavramı, hem işlenecek verinin büyüklüğünü hem de analiz için kullanılan teknolojiyi ifade etmek için kullanılmaktadır. Büyük veri analizi, büyük diye nitelendirilen veri kümesinin analitik teknikler ile işlenmesini ifade etmektedir (Casado ve Younas, 2015). Büyük verinin hacmi, yapısı ve hızı, veri saklama, işleme ve analiz

etme yöntem ve araçlarını doğrudan etkileyerek büyük veri teknolojisi kavramının ortaya çıkmasını sağlamıştır. Büyük veri teknolojisiyle, ilişkisiz ve çoğunlukla yapısal olmayan ve hızla akan verinin birleştirilmesi ve gerçek zamanlı analiz edilmesi mümkün olmaktadır (Erl, Khattak ve Buhler 2016).

Akademik çevrelerde de büyük veri konusu 2012 yılı ve sonrasında giderek artan bir ilgi ile çalışılmaya başlanmıştır; 2015-2016 yıllarında ise ilgide azalma olmuştur. Aşağıdaki şekillerde Türkiye ve Dünyada büyük veri çalışmalarının eğilim grafikleri verilmiştir. Bu grafiklerde 2018 yılı için ilk 11 aylık veriler bulunmaktadır.



Şekil 1.6. Akademik Çalışmalarda Büyük Veri – Dünya Çapında



Şekil 1.7. Akademik Çalışmalarda Büyük Veri – Türkiye Özelinde

1.2.6. Büyük Verinin Boyutları

Büyük veri kavramı, yalnızca verinin büyüklüğünü (hacmini) tanımlamak için kullanılmamaktadır. Bir verinin büyük veri kabul edilebilmesi için Laney, 3V ile temsil edilen boyutların bulunması gerektiğini söylemektedir: Volume (hacim, büyüklük), Velocity (hız) ve Variety(çeşitlilik) (Laney, 2001; Beyer ve Laney, 2012). Verinin

büyüklüğü göreceli bir kavram olmakla birlikte, daha önce de bahsedildiği gibi, tek bir makine tarafından işlenemeyecek kadar çok olmasıyla formülize edilmektedir.

Çeşitlilik boyutu ise verinin, farklı kaynaklardan alınmasını ve web içeriği, log dosyası, sosyal medya paylaşımı, e-posta, telefon kayıtları, radyo ve televizyon arşivi vs. formatlarda olmasını ifade eder (Yüzügöldü, 2015). Verinin çeşitliliği, verinin, teknoloji farklılığı dolayısıyla çeşitlenmesini ve yapısal (structured), yarı yapısal (semi-structured) ve yapısal olmayan (unstructured) formatlarda bulunması durumunu ifade eder. Yapısal veriler belirli kurallar ve sistematik yapı içinde kaydedildikleri için veriye erişim ve düzenleme kolaydır. Yapısal veriler, bu yönüyle ilişkisel veri tabanı sistemlerinde depolanabilmektedir; sıradan programlama usulleri ile kolayca okunup, değiştirilebilmektedir.

Yapılandırılmamış (yapısal olmayan) veriler ise geleneksel depolama ve veri analiz yöntemleri ile işlenemeyen verilerdir. Eposta, web sayfası, kitap, fotoğraf ve film gibi dijital veriler, yapılandırılmamış verilere örnek verilebilir (Oğuzlar 2011). Son olarak bir kısmı yapısal formatta ve bir kısmı yapısal olmayan formatta veriler bulunmaktadır ki buna yarı yapısal veriler denir ve xml, csv, xls uzantılı dosyalarda tutulabilmektedir (Doğan, 2014). Aşağıdaki şekilde farklı teknolojilerin ve ortamların ürettiği farklı formatlar ile çeşitlenen veri resmedilmektedir.



Şekil 1.8. Veri Çeşitliliği (Dorenkott, 2013)

Narayanan'a (2014) göre gelenekselden farklı olarak büyük veri, yarı yapılandırılmış ve yapılandırılmamış (büyük verinin %80'i yapısal değil) veri tiplerini de kapsayarak, işletmelerin paradigmalarını değiştirecek boyutta farklılıklar içermektedir. Yapısal olmayan veriler için kullanılan bilgi çöplüğü tabiri ise yakın geçmişimizde yerini sermaye ve petrol kavramlarına bırakmıştır (Rotella, 2016; Palmer, 2016).

Yapısal olmayan verinin, karar almada kullanılarak, değer oluşturmalarının işletmeler tarafından fark edilmesiyle büyük veri tanımına dördüncü bir boyut daha eklendi: Value (değer) (Dumbill, 2012; Doğan, 2014), şirket için artı değer sağlıyor olması durumunu ifade eder. Peki veriler, her hal ve durumda değerlerini koruyabilirler mi? Tekrar kullanılmak üzere saklanan yüksek hacimli verinin, amortisman değerinin dikkate alınması depolama maliyetlerinin düşürülmesi için önemlidir. Bu bağlamda, faydası giderek düşen verilerin ayıklanabilmesi için komplike modeller geliştirilmiştir (Mayer-Schönberger ve Cukier, 2013).

Bazı kaynaklarda büyük verinin 5. ve 6. boyutlardan da bahsedilmektedir. Verinin doğru olması (Verification), yani güvenli bir kaynaktan gelmesi ve ilgili kişilerce okunması; Verinin değişkenliği (Variability) yani veri akışındaki tutarsızlık dolayısıyla veri bütünlüğünün sağlanmasının güçlüğü (Katal, Wazid ve Goudar, 2013; Katal, Wazid ve Goudar, 2013).

Büyük verinin anlaşılmasını kolaylaştıran bu boyutlar kesin sınırlar ile tanımlı değildir. Sosyal medya verileri ise bu bağlamda bütün boyutlarıyla gerçek bir büyük veri olma özelliği göstermektedir.

1.2.7. Büyük Verinin İşletmelere Etkileri

Farklı kaynaklardan beslenen veriler arasındaki ilişkileri anlayabilmek, güncel ve önemli verileri tespit ederek gerçek zamanlı analiz gerçekleştirmek işletmeler için önemlidir. Zira büyük veri analizi, tedarik, üretim, pazarlama, insan kaynakları ve müşteri ilişkileri gibi işletme fonksiyonlarını beslemektedir (Davenport, Barth ve Bean, 2012). Ancak işletmeler, büyük veriyi depolayabilmek ve işleyebilmek için teknolojik yatırım maliyetine katlanmak zorundalar.

Teknolojilerin ayrıca işletmeler üzerinde çok ciddi değişiklik yaptığı öteden beri bilinmektedir. İşletmenin bu bağlamda, bazı kültürel ve organizasyonel değişiklik için

kendini hazırlaması gerekmektedir. Örneğin işletmelerde, veri analisti yerine veri bilimcisi istihdam edilmelidir. Çünkü büyük veri analizlerinde, işletmenin belli bir sorununa dönük olarak verileri analiz eden veri analisti yerine farklı kaynaklardan verileri toplayarak yapılandırılmamış olanları yapılandırılmış formata çeviren, veriyi analiz eden, sonuçları yöneticilere (özellikle üst düzey) görsel olarak raporlayan (Ming vd., 2011) ve bütün bu işleri çok kısa bir süre içerisinde yapan veri bilimcisine ihtiyaç duyulmaktadır. Veri bilimcisi, bilgisayar ve istatistik bilgisine, bilimsel arka plana, iletişim becerisine ve iş uzmanlığına sahip çok donanımlı bir personeldir (De Mauro vd., 2018).

Veri bilimcisinin istatistik bilgisine ihtiyaç duyuluyor çünkü büyük veri analizinin gelişimi, temel istatistiksel yöntemlerini değiştirmemiştir. Ancak veri işleme hızını artırarak maliyetleri düşürmüştür.

Yapılandırılmamış veriyle uğraşmak ise veri bilimciler için verinin büyüklüğü ile uğraşmaktan daha ciddi bir sorundur. Bu bakımdan veri bilimcinin, Pig, Hive, Python ve Java gibi programlama dillerine hâkim olmalarının yanında, büyük veri teknolojilerine aşinalık kazanmış olmaları gerekmektedir. Nitekim, veri bilimcinin gelecekteki önemini anlayan Kaliforniya Üniversitesi Berkeley Bilişim Okulu lisans düzeyinde ve New York Üniversitesi yüksek lisans düzeyinde veri bilimci yetiştirmeye başlamıştır (Davenport, 2014). Ülkemizde ise 5 üniversitede yüksek lisans düzeyinde eğitim verilmektedir.

İşletmeler veri analizi için teknolojik yatırım ve veri bilimcilerini istihdam etmenin maliyetine de katlanmalılar. Büyük veri analizinden faydalanabilirlerse işletme yöneticileri, karar destek sistemlerinden çok daha etkin ve kolay karar verme imkânına kavuşmuş olmaktadır (Demirtaş ve Arğan, 2015). Stratejik kararlar alma konumundaki yöneticiler de büyük verinin kıymeti hakkında olumlu kanaat taşımalarına rağmen, kurumları için ne kadar etkili olabileceğini kestirememektedir (Gillis ve Stephanny, 2015).

İşletmeler karar alma mekanizmalarında ağırlıklı olarak kendi iç veri tabanlarından faydalanmaktadır. Ancak kurum içinde her yıl yaklaşık %40 artan ve bir kısmı otomatik olarak kayıt altına alınmış verilerin kıymetinin de farkında olmadıkları bir vakıadır. Karar alma mekanizmasında bulunanlar, herhangi bir bilgi sistemine yatırım yapabilmeleri için yatırımlarının, yatırım miktarından daha fazla dönüş sağlamasını ve

iřletmeye gerek bir ekonomik deęer katmasını isterler. Onlar iin daha da nemlisi, iřletme deęerinin artırılabilmesidir (K.C. Laudon ve J.P. Laudon, 2004).

Byk veri analizi ile iřletmeler, iř ortamları hakkında daha fazla bilgi edinerek, pazarlama ve tedarik zinciri ynetim srelerini de iine alan tm iřletme srelerinde, kanıta dayalı, isabetli ve gerek zamanlı kararlar alabilmektedir (McAfee ve Brynjolfsson, 2012; Esen ve Trkay, 2017). Bunu, daha karmařık analiz teknikleri ve daha detaylı veriler kullanarak saęlayabilmektedir. Ayrıca varsayımlarını ve aldıkları kararların sonularını test ederek operasyonel sreleri ve yatırımlarını daha etkin ynetebilmektedir (Brown vd., 2011). Ancak btn bunlar iin yeni ynetim yaklaşımlarına ihtiya duyulmaktadır (Grsakal, 2014).

“Endstri devi olsun veya olmasın her boyuttan ve trden firma iin aynı durum geerlidir. Konu artık firmaların byk bir veriyi kullanıp kullanmamaları deęil, onu ne zaman ve nasıl kullanacaklarıdır” (Marr, 2017 s.40). Teknolojik geliřmelerle birlikte yarı yapısal ve yapısal olmayan verilerin iřletme dıřından toplanarak analiz edilmesi hayati nem arz etmeye bařladı. Marr (2017), dięer taraftan zellikle daha az kaynaęa sahip kk iřletmelerin, analitik kapasitelerinin yetersizlięi dolayısıyla, byk veriye karřı temkinli yaklařıklarını sylemektedir. Ancak tketicilerle ilgili zengin bir i gr kaynaęı olması hasebiyle, pazarlamacılar iin altın madeni nitelięinde bulunan verilerin analiz edilmeden pe gitmesi kaınılmazdır.

Davenport’a (2014) gre ise iřletmeler byk veriye geiřte  farklı tutum ve davranıřı benimseyebilirler. Rakiplerin henz byk veriye gemedięi, sektrde teknolojik dnřmn yavaş olduęu, mřteriler ve dięer nemli iř etkenleri ile ilgili fazla verinin olmadığı ve sektrdeki inovasyonlara nclk edilemedięi durumda byk veriye geiř iin temkinli olunmalıdır. Sektrn byk veri ve analitięi alanında aktif olduęu, rekabeti avantajın nemli olduęu, teknolojik yeterlilięin bulunduęu ve byk veri alıřanlarının yeterli olduęu durumda ise byk veriye geilmesi yararlıdır. Ancak iřletme, byk veri kullanımında ok aktif bir sektr iindeyse, gemiřte teknolojiyi kullanarak kendini dnřtrlebildiyse ve gereken tm ekipmana sahipse byk veriye geiřte acele etmelidir (Davenport, 2014).

Büyük veriden bütün sektörler eşit oranda etkilenmemektedir. Büyük verinin farklı formatları için hangi sektörün ve işletmenin hangi biriminin etkilendiği Tablo 1.2’de verilmektedir.

Tablo 1.2. Veri Biçimleri ve İşletme Birimlerine Etkisi (Davenport, 2014)

<i>Veri Biçimi</i>	<i>Veri Kaynağı</i>	<i>Etkilenen Sektör</i>	<i>Etkilenen Birim</i>
<i>Yüksek Hacim</i>	Çevrimiçi	Finansal Hizmetler	Pazarlama
<i>Yapılandırılmamış</i>	Video	Sağlık	Tedarik Zinciri
<i>Sürekli Akan</i>	Sensör	İmalat	İnsan Kaynakları
<i>Çoklu Biçim</i>	Genom	Seyahat/Ulaşım	Finans

Şirketler, büyük veri analizi ile maliyetlerinin %72 oranda düşürülmesini, %54 oranında kaynakların etkili kullanılmasını ve %35 oranında yeni fırsatların oluşturulmasını amaçlamaktadır (Gillis ve Stephanny, 2015). Şirketler, müşteri ihtiyaçlarının belirlenmesinden, müşteriye özel ürün ve hizmet sunumunun gerçekleştirilmesine, kurumsal kaynak planlamasından, yatırım yönetimine kadar birçok konuda büyük veriden faydalanmaktadır (Esen ve Türkay, 2017).

Büyük veriyi kazanca dönüştüren örneklerden biri olan Turkcell, sorun yaşayan müşterilerini tespit edip, onlarla iletişime geçerek çeşitli promosyonlara dâhil ederek, müşteri memnuniyetini artırmış, müşteri kaybını büyük ölçüde önlemiştir. Gartner tarafından ödüllendirildiği bu başarısıyla ayrıca Turkcell müşteri hizmetleri (call center) maliyetlerini de azaltabilmiştir (Barske, 2013). Büyük veriden gerçek değer elde eden şirketlerin başında gelen Wal-Mart ve Amazon ise çoğu ülkenin milli hasılasından fazla ciro yapmaktadır.

1.2.8. Sosyal Medya Analizinde Büyük Veri

Sosyal medya, milyarlarca insanın, yorum, fotoğraf, video ve konum bilgisi gibi anlık bilgi paylaşımlarında bulunabildiği bir platformdur. Yapılan bir çalışmaya göre tüketicilerin % 43’ü, bu platformu aktif olarak kullanan şirketlerin, sosyal ağlar üzerinden kendi problemlerini çözüme kavuşturmalarını beklerken, şirketlerin ancak %7’si müşteri ilişkilerinin öneminin farkındadır (N. Yücel, 2013). Sosyal medyanın verimsiz kullanımının ise hisse senedi fiyatlarını ve marka itibarını olumsuz etkilediği yapılan başka bir çalışmayla ortaya konmuştur (Özmen ve Villi, 2014).

Sosyal medya analizi, veri boyutu ve oluşum hızı faktörleri dolayısıyla, büyük veri teknikleri kullanmayı gerekli kılmaktadır. Sosyal medya verilerinin analizi ile geleneksel yöntemlerle anlaşılamayan tüketici davranışlarının gerçek örüntülerinin belirlenmesi mümkün olmaktadır (Çevik, Öztürkcan ve Kasap, 2015). Büyük veri analizi, denize sarkıtılan bir ağa benzetilmektedir. Ağa hangi balıkların takılacağı ise başlangıçta bilinmemektedir. Yine de veriyle ne yapılmak istendiği ve hangi sorulara cevap arandığı önemlidir (Özdoğan, 2013).

Kurumlar, mevcut müşterileri ile iletişime geçmek, güven ve sadakatlerini kazanmak ve yeni müşterilere ulaşmak için sosyal medya analiz yöntemlerini, geleneksel yöntemlere göre oldukça düşük maliyet ile kullanabilmektedirler. Sosyal medya analizlerinin, mevcut durumun tespit edilmesinden ziyade geleceğin tahmin edilmesinde de aktif rol alması beklenmektedir. Yani örneğin müşterilerin yakın gelecekte hangi ürün ve hizmetleri satın alacağını tahmin edebilmek, yatırımcılar için önem arz etmektedir. Şirketler bu bağlamda büyük veriye sahip olmakla övünmemelidir. Zira sosyal medya özelinde büyük verinin analiz edilmesi ve anlamlandırılarak gerçek değer üretilmesi, veriye yalnızca sahip olunmasından çok daha önemlidir (Davenport, 2014).

1.2.9. İlişkisel Veri Tabanı Yönetim Sisteminden NoSQL'e

1.2.9.1. İlişkisel Veri Tabanı Yönetim Sistemi

İlişkisel veri tabanında, veriler, satır (record - kayıt) ve sütunlardan (field - alan) oluşan iki boyutlu tablolarda tutulmaktadır. Kayıtlar, ortak anahtarlar ile ilişkilendirilirken tablolar da kendi arasında ilişkilendirilmektedir. Böylece veri bütünlüğü sağlanabilmektedir (Oracle, 2015).

Birden fazla kullanıcının aynı anda verilere erişimini sağlayabilmek için her uygulama için farklı dosya kullanmak yerine tek veri tabanı tercih edilmektedir. Böylece veriler, veri tabanına eş zamanlı olarak kaydedilebilmektedir. Veri tabanları ise, veri tabanı yönetim sistemi (VTYS) olarak bilinen yazılımlar tarafından yönetilmektedir. VTYS, veri tabanının ve veri tabanı nesnelerinin oluşturulması, yönetilmesi, verilerin depolanması ve sorgulanması için geliştirilmiştir (Prytherch 2016). VTYS'de ayrıca,

yedekleme mekanizması ve kullanıcı haklarının tanımlandığı güvenlik kontrolleri de bulunmaktadır (Hoche vd., 2006).

İlişkisel veri tabanlarında veriye erişimi hızlandırılabilmesi için indeks yapıları da geliştirilmiştir. Bu yapılar sayesinde, kullanıcı erişim yetkileri ile güvenlik sağlanabilmekte ve veriler sıkıştırılmış olarak saklanabilmektedir. VTYS’de ayrıca ACID olarak bilinen dört temel özellik bulunmaktadır: Atomiklik (Atomicity), Tutarlılık (Consistency), İzolasyon (Isolation) ve Sağlamlık (Durability). Atomiklik, icra edilen komut kümesinin (transaction) çalışmasında, donanımsal veya yazılımsal bir hata olması durumunda, işlemlerin iptal edilmesini ifade etmektedir. Tutarlılık, birbiriyle ilişkili veriler arasında tanımlı kuralların işletilmesinin yazılımsal olarak da ihlal edilememesidir. İzolasyon ise transaction’ların birbirinden izole edilmesi durumudur. Son olarak sağlamlık ise transaction sonuçlarının veri tabanına kaydedildiğini garantilemektedir (Poyraz, 2015).

İlişkisel veritabanları üzerinde sorgulama yapmak için SQL (Structured Query Language – Yapısal Sorgulama Dili) olarak bilinen ve tüm ilişkisel veri tabanları için standart olan bir dil geliştirilmiştir (Oracle, 2015). SQL dili kullanılarak yazılan komutlar ise iki ana kategori altında değerlendirilmektedir: DML (Data Manipulation Language - Veri İşleme Dili) komutları ve DDL (Data Definition Language - Veri Tanımlama Dili) komutları. DML komutları, tablolardaki verilerin çekilmesi ve güncellenmesi gibi işlemler için kullanılırken, DDL komutları tablolar ve diğer veri tabanı nesnelerinin oluşturulması ve değiştirilmesiyle ilgilenir (Atcharyachanvanich, Nalintippayawong ve Permpool, 2017). SQL’de kullanılan temel DDL ve DML komutları aşağıdaki tabloda verilmiştir.

Tablo 1.3. Temel DDL ve DML Komutları

<i>Kategori</i>	<i>Komut</i>	<i>Açıklama</i>
<i>DML</i>	Select	Verilerin sorgulanması
	Insert	Kayıt(satır) eklenmesi
	Delete	Kayıt(satır) silinmesi
	Update	Verilerin güncellenmesi
<i>DDL</i>	Create	Nesne oluşturulması
	Drop	Nesnenin kaldırılması
	Alter	Nesne özelliklerinin değiştirilmesi

SQL sorguları temel olarak üç kelimedenden oluşmaktadır: “Select”, “From” ve “Where”. Listelenmek istenen veriler yani sütun adları “Select” sözcüğü ile birlikte, verinin hangi veri kaynaklarından çekileceği “From” sözcüğü ile ve son olarak sorguda filtreleme yapılacaksa da “Where” sözcüğü ile birlikte belirtilir.

1.2.9.2. NoSQL

Büyük verinin üretim hızı ve akışı, bağlantı hızını ve bant genişliğini de etkilemektedir. İlişkisel veri tabanı teknolojisi ise bu veri hızına ve bant genişliğine çok pahalı çözümler önermektedir (Özdoğan, 2016). NoSQL sistemler, ilişkisel veri tabanı için ölçeklenemeyecek miktarda ve çeşitlilikte verinin uygun bir şekilde kaydedilerek işlenebilmesi ihtiyacına cevap olarak doğmuştur.

NoSQL (Not Only SQL – Sadece SQL Değil) 1998 yılında Carlo Strozzi tarafından açık kaynaklı olarak geliştirilmiş (Uzunbayır, 2015) olan NoSQL sayesinde hızlı akan veriler üzerinde işlem yapabilmek kolaylaşmıştır. NoSQL sistemlerin özelliği yatay olarak ölçeklenebilmesi, dağıtık olarak çalışabilmesi ve düşük ölçekli olmasıdır. Dağıtık çalışabilme özelliği, bu sistemlerin ağ üzerinde birbirine bağımlı olmadan iletişimde bulunan çok sayıda bilgisayar üzerinde çalışabilmesini ifade etmektedir (Öztürk ve Atmaca, 2017). Ölçeklenebilirlik özelliği sayesinde ise yapılan iş, farklı sunuculara bölünebilmektedir (partition).

NoSQL yapılar, ilişkisel veri tabanı yönetim sistemlerinin büyük veri söz konusu olduğunda, veri yazmada yavaşlığına ve yapısal olmayan verilerin işlenememesi problemlerine önemli bir çözüm olarak ortaya çıkmıştır (Yüzügüldü, 2015). NoSQL yapılar, yapısal olmayan verilerin hızlı kaydedilmesini sağlarken, depolama ve işlem maliyetlerini ise düşürmektedir (Mayer-Schönberger ve Cukier, 2013).

NoSQL, veriler üzerinde değişiklik yapılmak için tasarlanmadığından, NoSQL’de SQL’e kıyasla çok daha basit sorgular yazılabilmektedir. Ancak, geleneksel veri tabanlarındaki kısıtlar olmadığından, veriler hızlı bir şekilde sorgulanabilmektedir. Verileri daha hızlı okuyabilmek için de RAM den faydalanılmaktadır (Selçuk, 2015). NoSQL, ilişkisel veritabanlarına kıyasla daha zayıf eş zamanlılık modeli kullanmaktadır. İlişkisel veritabanlarının ACID özelliklerine benzer şekilde, NoSQL’de de BASE

(Basically Available, Soft state, Eventually consistent) özellikleri bulunmaktadır (Gökşen ve Aşan, 2015):

- Kolay Ulaşılabilirlik (Basically Available): Veriye daha kolay erişebilmek için farklı sunuculara dağılmış kopyalardan faydalanır.
- Esnek Durum (Soft state): ACID özelliklerinden farklı olarak tutarsız verilerin bulunmasına da imkân sağlar.
- Eninde sonunda Tutarlı (Eventually consistent): Veriler, anlık olarak tutarlı olmasa da verilerin, bir zaman sonra tutarlı olacağını varsayar.

NoSQL'in dağıtık sunucular üzerinde çalıştığını söylemiştik. Dağıtık veritabanı sistemlerinde ise Professor Eric Brewer tarafından 2000 yılında ortaya konan CAP (Consistency Availability Partition tolerance – Tutarlılık Ulaşılabilirlik Bölünebilirlik toleransı) teoremi geçerlidir (Casado ve Younas, 2015). Bu teoreme göre, dağıtık veritabanı sistemlerinde tutarlılık, ulaşılabilirlik ve bölünebilirlik toleransı özelliklerinden 3'ü aynı anda bulunamaz. İlişkisel veritabanlarında mesela bu özelliklerden ikisi yani tutarlılık ve ulaşılabilirlik (CA: Consistency Availability) özelliği bulunmaktadır. Diğer iki durum (CP ve AP) ise NoSQL veritabanlarına özgüdür (Uzun vd., 2018).

NoSQL veri tabanları çoğunlukla, anahtar-değer veri tabanları üzerine temellendirilmiştir. İlişkisel veri tabanlarında birincil anahtar (primary key) kullanımı zorunlu olmamasına rağmen NoSQL'de verilere başka türlü erişilememektedir. Alışılmışın aksine veriler bu yapıda, sabit tablo yapılarından bağımsız olarak JSON, XML veya BSON formatlarında dokümanlar olarak tutulmaktadır (Şekil 1.10). Dokümanlar belirli bir standarda uygun şekilde tasarlanmadığı için dokümanlar arasında bir ilişki kullanılması yani join komutları uygun değildir. Ayrıca transaction kavramı NoSQL sistemlerde karşılık bulmadığından veri güvenliği sorun olabilmektedir. Özetle VTYS ile çalışan bir sistemin bütünüyle NoSQL'e taşınması, kolay değildir.

```
{“Numara”:452, “1.Vize”:30, “Final”:50}
{“Numara”:453, “1.Vize”:80, “Final”:50}
{“Numara”:454, “1.Vize”:45, “2.Vize”:19, “Final”:50}
```

Şekil 1.9. Json Formatı (Karataş, 2015)

NoSQL sistemler, ilişkisel veritabanlarının yerini dolduramadığından melez bir yazılım ortamında ilişkisel veritabanı sistemlerine destek olacak şekilde geliştirilmesi öngörülebilir (Özdoğan, 2016). Müşteri tabloları gibi daha çok okuma ve yazmanın yoğun olduğu tablolar mesela, NoSQL veritabanlarında tutulurken birden fazla tablo üzerinde sorgulama yapılmasını gerektiren hesaplama tablolarının ise VTYS’de tutulması sağlanabilir (Gökşen ve Aşan, 2015).

NoSQL veri tabanları, doküman, anahtar-değer, sütun ve çizge olmak üzere dört sınıfa ayrılmaktadır (Ward, 2013).

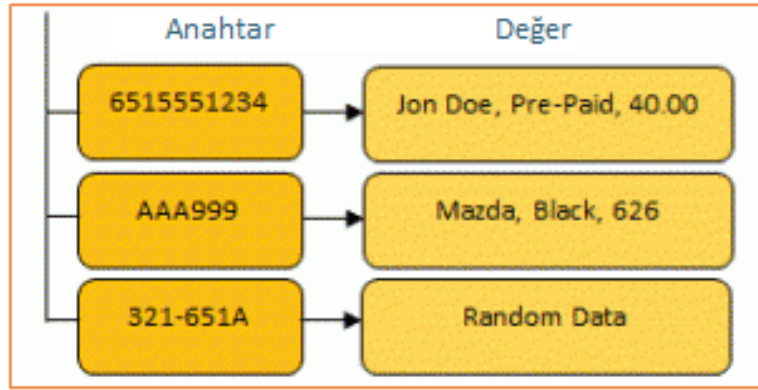
1.2.9.3. Döküman Veri Tabanı

Döküman veri tabanı, çeşitliliği yüksek olan ve özellikle yarı yapısal verilerin kaydedilmesinde ve yönetilmesinde kullanılmaktadır. Veri türüne göre esneklik arz eden bu veri tabanı tipinde her bir satır bir dokümana karşılık gelmektedir ve dokümanlar arasında standart bir yapının ve benzerliğin bulunması zorunlu değildir. MongoDB, eXist, BaseX ve CouchDB bu veri tabanı tipine örnek olarak verilebilir (Parmar ve Roy, 2018).

Birden fazla kullanıcının bir tabloya aynı anda erişmek istemesi durumunda öncelik sırasına göre erişim sağlanabilmektedir. Eğer kullanıcı sayısı çok ise veya yoğun işlemler icra edilmesi gerekiyorsa, sistemde uzun beklemler ve dolayısıyla yavaşlama olacaktır.

1.2.9.4. Anahtar Değer Veri Tabanı

Anahtar değer veri tabanı, bir anahtar alanın ve ona karşılık gelen değer alanının tutulduğu veri tabanı tipidir. Anahtarın tekil olma şartı bulunmaktadır. REDIS, Flare, Memcached örnek olarak verilebilir.



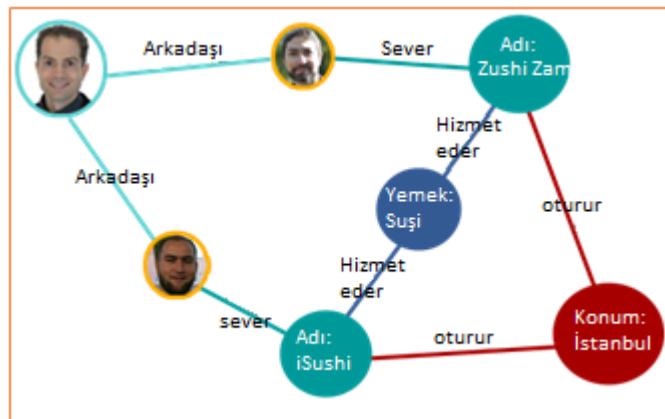
Şekil 1.10. Anahtar-Değer Veritabanı (Valz, 2013)

1.2.9.5. Sütun Veri Tabanı

Sütun veri tabanı, bir satırda çok sayıda sütun bulunması prensibine göre işlemektedir (Sarıbıyık, 2015; Ünalır vd., 2015). HBase, Cassandra, Riak, Hypertable ve Voldemort projesi, anahtar değer veri tabanlarına örnek olarak verilebilir (Ward 2013).

1.2.9.6. Çizge (Graph) Veri Tabanı

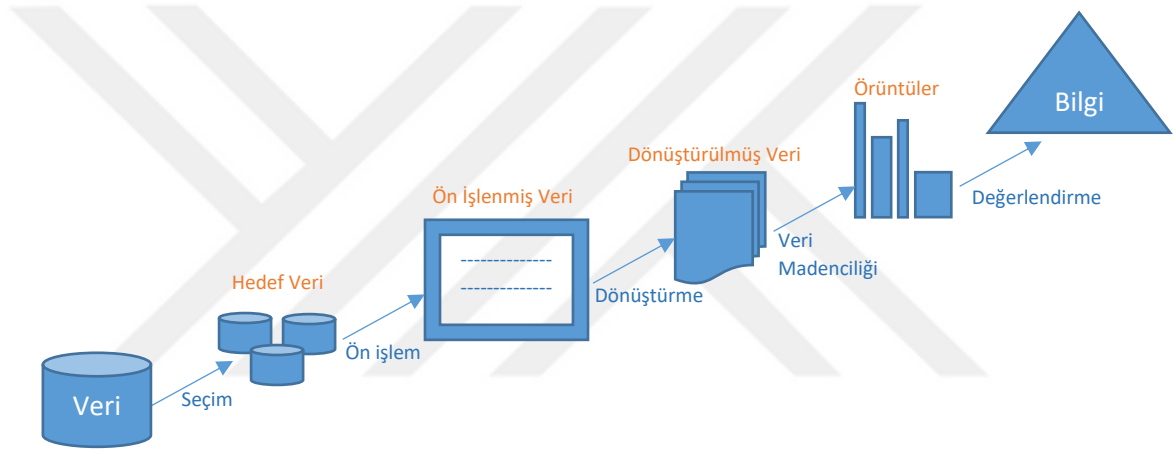
Çizge (Graph) veri tabanında veriler, düğümler ve bağlantılar (çizgiler) ile temsil edilmektedir. Çizge veri tabanında, gösterge (pointer) kullanılarak doğrudan adresleme yapılabildiğinden indexlere gerek yoktur (Sarıbıyık, 2015). Neo4j, DEX, InfiniteGraph, Sesame ve Jena veri tabanları, çizge veri tabanlarına örnek olarak verilebilir (Karataş, 2015).



Şekil 1.11. Çizge Veritabanı (Neo4j, 2017)

1.2.10. Veri Madenciliğinden Büyük Veri Analizine

Veri madenciliği (data mining), veri yığınları içinden, önceden keşfedilmemiş bilgileri ve ilişkileri bulmaya çalışan bir araştırma alanıdır. Diğer bir deyişle veri madenciliği, tanımlama veya tahmin amacıyla, istatistiksel analiz ve yapay zekâ tekniklerinin kullanıldığı bir alandır. Tanımlama, belirli bir veri seti içindeki geçerli ilişkilerin ortaya konması anlamına gelirken; tahmin etme, mevcut verilerden yola çıkarak gelecekte neler olabileceğinin ortaya konmasıdır. İlk çıktığı zamanlarda statik verilerin analizinde kullanılan veri madenciliği, günümüzde akan veriler üzerinde de kullanılabilmektedir (Güven, Bozkurt ve Kalıpsız, 2007).



Şekil 1.12. Veri Madenciliği Süreçleri (Fayyad, Piatetsky-Shapiro ve Smyth, 1996)

Veri madenciliği, eksik veya yanlış verilerin düzeltilmesi, farklı kaynaklardan verilerin bütünleştirilmesi, analiz yapılacak verilerin seçilmesi, verilerin analize uygun hale dönüştürülmesi, veri madenciliğinin uygulanması, ortaya çıkan örüntünün değerlendirilmesi ve son olarak bilginin sunumu işlemleri yapılmaktadır. Veri madenciliği süreç adımları Şekil 1.12’de gösterilmiştir. Veri madenciliği için yaygın kullanılan araçlar ise, SPSS Clementine, SPSS SAS, KXEN, Orange, KNIME, SQL Server ve Excel’dir (Dener, Dörterler ve Orman, 2009).

İstatistik ve veri madenciliği yöntemleri, verilerin az olduğu zamanlarda teorilere olan ihtiyaçtan doğmuştur denebilir. Ancak bugün veriler o kadar büyüdü ki, örneklem alma yöntemi, yerini bütün verinin analiz edildiği tekniklere bıraktı; bu dönemde artık teorilerden çok verinin sesi duyulur oldu (Anderson, 2008). Aşağıdaki tabloda büyük veriye kadar olan tarihsel süreç sunulmaktadır:

Tablo 1.4. Büyük Veriye Kadar Olan Gelişim Süreci (Davenport, 2014)

<i>Terim</i>	<i>Zaman Aralığı</i>	<i>Özel Anlam</i>
<i>Karar destek</i>	1970-1985	Karar vermeyi desteklemek için veri analizi kullanımı
<i>Yönetici Desteği</i>	1980-1990	Üst düzey yönetici kararları için veri analizi kullanımı
<i>Çevrimiçi Analitik İşlem İş zekâsı</i>	1990-2000	Çok boyutlu veri tablolarının analizi için kullanılan yazılımlar
<i>Analitik</i>	1989-2005	Veri odaklı karar vermeyi desteklemek için kullanılan raporlama ağırlıklı araçlar
	2005-2010	İstatistiki ve matematiksel analiz odaklı karar verme
<i>Büyük Veri</i>	2010-	Çok büyük, yapılandırılmamış ve hızlı akan veri kullanımı

İKİNCİ BÖLÜM

MATERYAL ve YÖNTEM

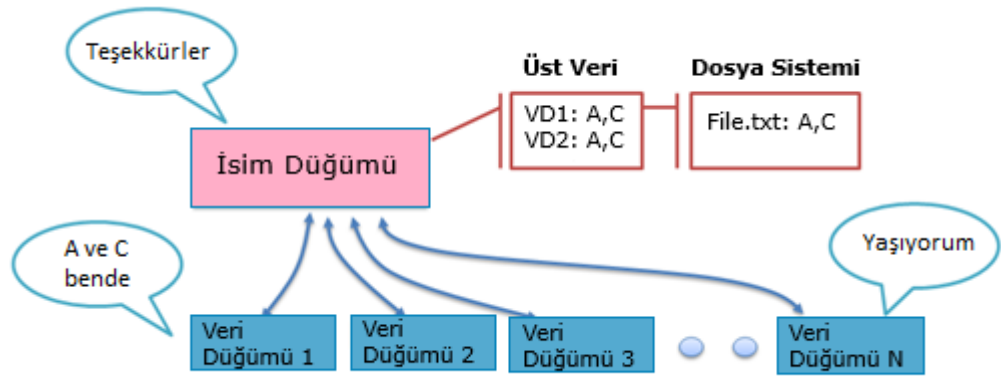
2.1. HADOOP

Apache Hadoop, Google tarafından geliştirilen dosya sistemi (GFS: Google File System – Google Dosya Sistemi) ve Eşle İndirge (MapReduce) modelleri ilham alınarak Java programlama dilinde açık kaynak kodlu olarak geliştirilmiş bir kütüphanedir. Düşük hesaplama gücüne sahip çok sayıda bilgisayarın bir küme içerisinde organize çalışması prensibine dayanmaktadır (Doğan, 2014). Böylece veriler düşük maliyetli ve hızlı bir şekilde işlenebilmektedir. Hadoop, özellikle geleneksel yöntemlerin yetersiz kaldığı veya pahalı olduğu durumlarda tercih edilmektedir. Geleneksel yöntemlerde veri, analiz edilmeden önce veri ambarına taşınarak ETL sürecine tabi tutulurken; Hadoop'ta tutulan veri, taşınamayacak kadar büyüktür. Yani, veri için büyüklük, bir anlamda yerçekimi kanununa bağlı herhangi bir nesne için ağırlık gibidir.

Hadoop isminin bulunması ise biraz ilginçtir. Rivayet edildiğine göre, proje yöneticisinin (Doug Cutting) oğlunun, oyuncağına verdiği anlamsız isim, söylenmesi ve hatırlanması kolay olduğu için projeye isim olmuş (Kapçı, 2015).

Hadoop, büyük verinin standart bilgisayarlardan oluşan bir küme üzerinde depolanmasında ve dağıtık programlama kullanarak işlenmesinde altyapı desteği vermektedir (Kayım, 2015). Hadoop, üç farklı modda yani tekil modda (singlenode mode), sözde dağıtık modda (pseudo-distributed mode) ve tam dağıtık modda (fully-distributed mode) çalışabilmektedir (Yıldırım vd., 2014).

Yahoo, Facebook, Twitter, eBay ve LinkedIn Hadoop'u aktif kullanarak; Oracle, Microsoft ve HP gibi önemli firmalar çeşitli çözümler geliştirerek Hadoop'a destek vermektedir (Yüzüğüldü, 2015).



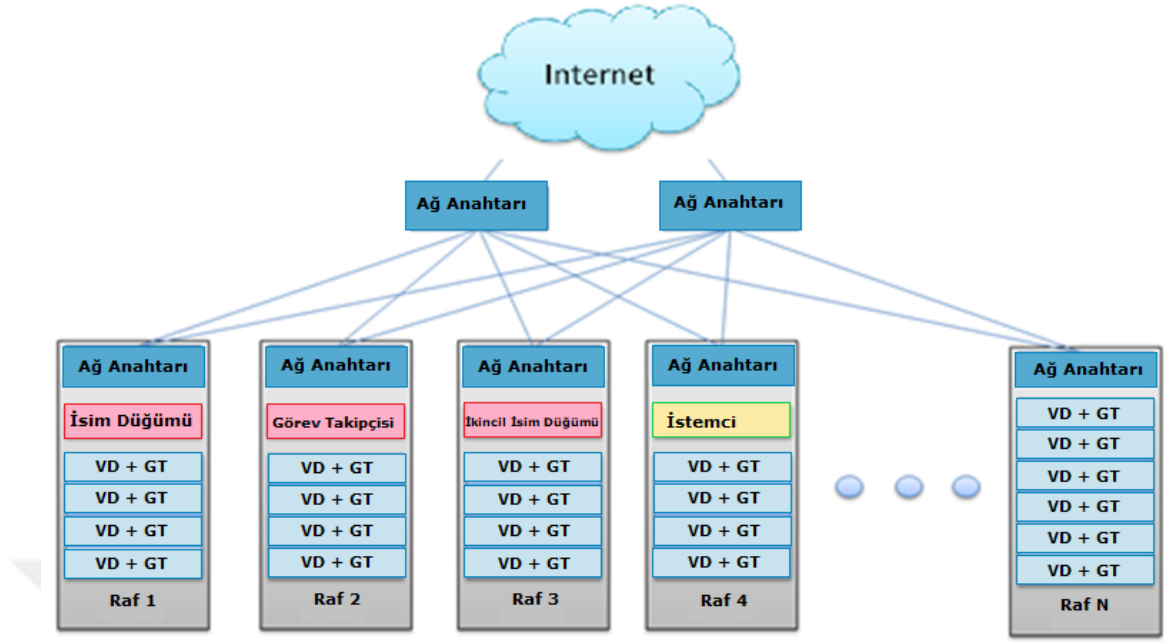
Şekil 2.1. İsim Düğümü ve Veri Düğümü Yapıları (Hedlund, 2011)

Hadoop'un çalışma prensibi ise şöyledir: Hadoop kümesi içerisinde bir makine, isim düğümü (NN: Name Node) diğer adıyla efendi düğüm (Master Node) görevini üstlenir; diğerleri ise veri düğümü (DN: Data Node) yani köle düğüm (Slave Node) görevini üstlenmektedirler. İsim düğümü, veri düğümlerini kontrol eder ve dosya transferinden sorumludur (Uzunbayır, 2015). Veri düğümleri, isim düğümlerine sürekli ileti (heartbeat: kalp atışı) göndererek (ön tanımlı olarak 3 saniyede bir) erişilebilir olduklarını ve hangi veri bloklarını barındırdıklarını bildirmektedirler (Şekil 2.1). Bu ileti sinyalleri blokları ve bloklar da meta verileri (metadata) oluşturmaktadır. On dakika süreyle herhangi bir sinyal alınmayan veri düğümü içinse artık devre dışı kaldığına hükmedilir (Doğan, 2014).

İkincil isim düğümü (Secondary Name Node) ise sistem dosyalarının yedeğini tutmaktadır. İsim düğümü arızalandığında sistemi geri yükleyebilmek için isim düğümünün periyodik olarak (varsayılan olarak saatte bir) anlık görüntülerini (snapshot) alır (Koca, 2014).

İş takipçisi (JT: Job Tracker), Hadoop işlerini (jobs) ve verilerini veri düğümleri arasında paylaştırır. Görev takipçisi (TT: Task Tracker) ise veri düğümü üzerinde çalışarak iş takipçisinden alınan işi (task) yürütür (Gunarathne, 2015).

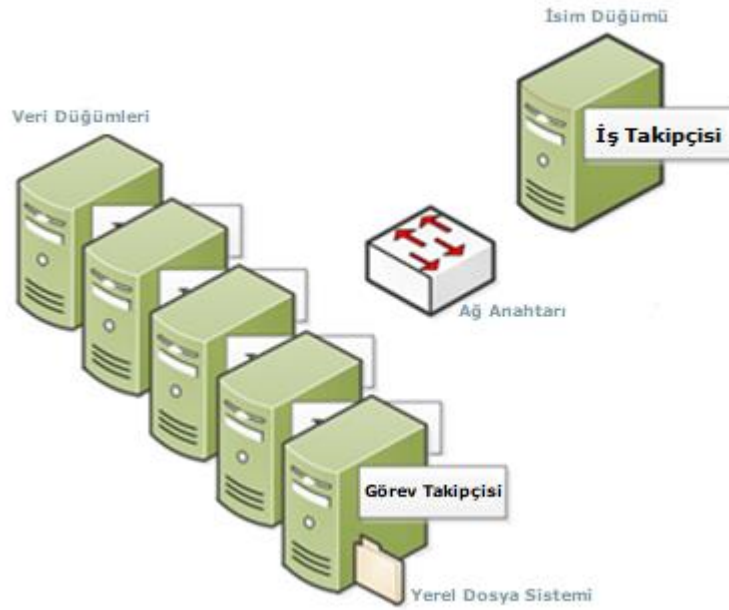
Veri düğümleri ise iş parçalarının çalıştırılması ve verilerin depolanması işini üstlenmektedir. Dolayısıyla Hadoop kümesinin disk kapasitesi ve işlem gücü, veri düğümlerinin kapasitesi ile ölçülmektedir. Veri düğümlerinin birlikte anahtarlandığı yapıya ise “rack” (raf) denilmektedir (Hedlund, 2011). Şekil 2.2’de rack yapısı gösterilmektedir.



Şekil 2.2. Rack Yapısı (Hedlund, 2011)

Hadoop, en temelde HDFS'den (Hadoop Distributed File System – Hadoop Dağıtık Dosya Sistemi) ve Eşle İndirge'den oluşmaktadır.

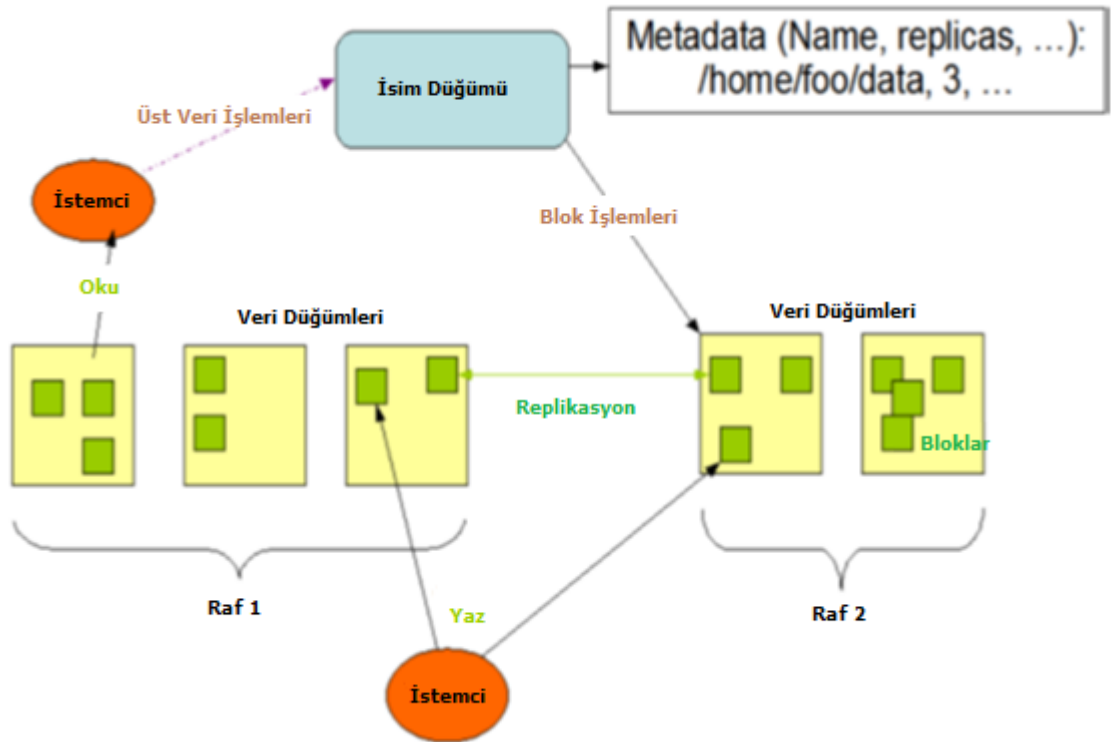
2.1.1.HDFS



Şekil 2.3. HDFS Kümesi (Sur vd., 2010)

Açılımı, Hadoop Dağıtık Dosya Sistemi olan HDFS, büyük verilerin dağıtık olarak saklanması için kullanılan bir sistemdir. Hadoop verileri, HDFS üzerinde tutulmaktadır. HDFS’ de ise yüksek depolama kapasitesine sahip bilgisayarlar yerine standart bilgisayarlardan oluşan bir ağ kümesi kullanılmaktadır. Veriler HDFS' ye kaydedilirken sıkıştırılabildiği için büyük boyutlu dosyalar halinde kaydedilebilmektedir (Kayım, 2015; Shvachko vd., 2010). Veriler bloklar şeklinde dağıtık sunucular üzerinde replikasyon yapılarak (3 kez) saklandığı için de HDFS’nin hızlı veri okuma yeteneği bulunmaktadır. Replikasyon aynı zamanda veri güvenliğini de sağlamaktadır. Replikasyon özelliği sayesinde veri kayıpları en aza indirilirken hata toleransı da artırılabilir (Jayalakshmi ve Srinivasan, 2017).

HDFS’de meta verileri, isim düğümü sunucusunda, uygulama verileri ise veri düğümü sunucularında tutulmaktadır (Şekil 2.4). Veri düğümündeki veriler için dosya, blok ve konum bilgileri de meta verilerinde tutulmaktadır. Sunucuların iletişimi ise TCP/IP üzerinden sağlanmaktadır (Doğan, 2014).



Şekil 2.4. HDFS Mimarisi (Borthakur, 2008)

Aşağıdaki tabloda ise yaygın olarak kullanılan HDFS komutları bulunmaktadır.

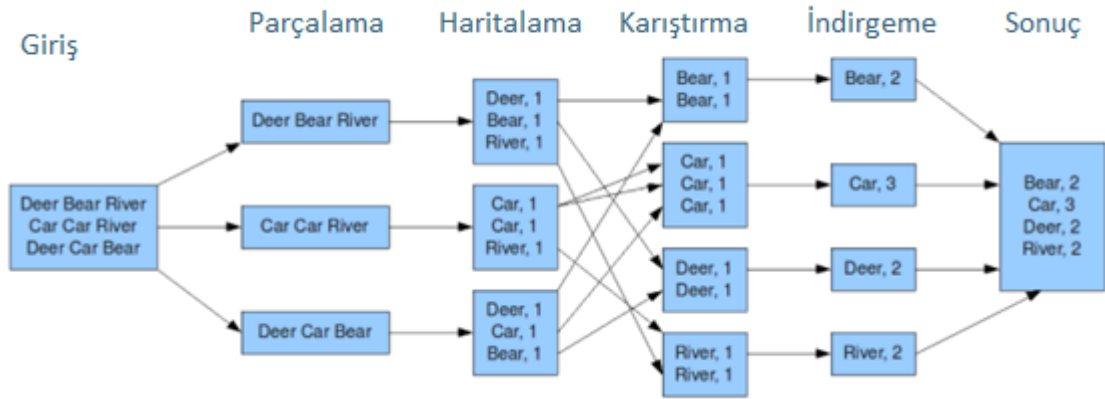
Tablo 2.1. Yaygın Kullanılan HDFS Komutları (Sakınmaz, 2016a)

Komut Adı	Kullanımı	Açıklama
<i>mkdir</i>	hdfs dfs -mkdir /KlasörAdı	klasör oluşturur
<i>copyFromLocal</i>	hdfs dfs -copyFromLocal KaynakDizin/Dosya HedefDizin	yerel makinedeki dosyalar HDFS ye kopyalanır
<i>copyToLocal</i>	hdfs dfs -copyToLocal KaynakDizin/Dosya HedefDizin	HDFS sistemindeki dosyalar yerel bilgisayara kopyalanır
<i>ls</i>	hdfs dfs -ls KlasörAdı	dosyalar listelenir
<i>du</i>	hdfs dfs -du Dizin/*	dosyalar boyutlarıyla beraber listelenir
<i>count</i>	hdfs dfs -count Dizin/*	dizindeki dosya sayısı alınır
<i>chown</i>	hdfs dfs -chown -R owner:group /Dizin/*	dosyaların kullanıcı yetkileri düzenlenir
<i>chmod</i>	hdfs dfs -chmod +x /Dizin/*	dosyaların okuma, yazma, çalıştırma yetkileri düzenlenir
<i>setrep</i>	hdfs dfs -setrep 3 -R /Dizin/*	dosyaların replication factor değeri düzenlenir
<i>cat</i>	hdfs dfs -cat /Dizin/dosya	dosyanın içeriğine erişilir
<i>tail</i>	hdfs dfs -tail /Dizin/dosya	dosyanın içeriğinin son satırlarına erişilir
<i>cp</i>	hdfs dfs -cp KaynakDizin/Dosya HedefDizin	dosya HDFS içindeki başka bir konuma kopyalanır
<i>mv</i>	hdfs dfs -mv KaynakDizin/Dosya HedefDizin	dosya HDFS içindeki başka bir konuma taşınır
<i>touchz</i>	hdfs dfs -touchz /Dizin/dosya	dosya oluşturulur
<i>rmr</i>	hdfs dfs -rmr /Dizin/*	dosya silinir

2.1.2. Hadoop Eşle İndirge

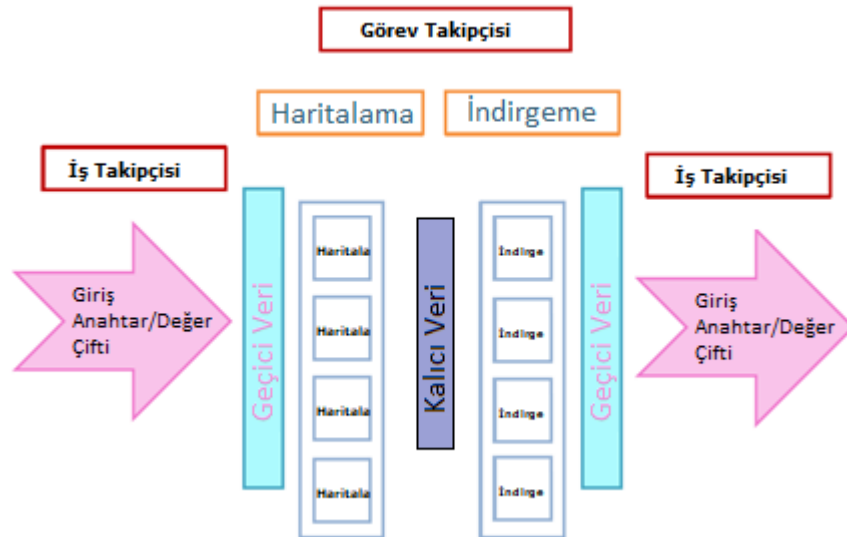
Paralel programlama modellerinden biri olan Eşle İndirge (MapReduce), Google tarafından geliştirilmiş olup, Eşle (Map) ve İndirge (Reduce) olmak üzere iki fonksiyonun birleşiminden oluşmaktadır (Hallaç, 2014; Dean ve Ghemawat, 2008; Lammel, 2008). HDFS’de kayıtlı devasa büyüklükteki verileri işlemek için kullanılan hesaplama modelidir (Gunarathne, 2015). Hadoop işlem framework’u iken, Eşle indirge işlem motorudur (Gurusamy, Kannan ve Nandhini, 2017).

Eşle fonksiyonu ile dağıtılan iş parçacıkları (thread) ve veriler, İndirge fonksiyonu marifetiyle toplanmaktadır. C, C++, Java, Python, Perl ve Ruby gibi programlama dillerine destek vermektedir. Yani bu diller ile Eşle İndirge programları yazılabilmektedir (Yüzügüldü, 2015).



Şekil 2.5. Eşle İndirge Kelime Sayma Şeması (Sakınmaz, 2016b)

Kelime sayma örneğinin çizelgelendirildiği şekil 2.5'te görüldüğü gibi HDFS'de tutulan veriler (persistent data) öncelikle bloklara ayrılır (splitting: parçalama), sonrasında geçici anahtar - değer yapısına (transient data) dönüştürülmektedir (mapping: haritalama). Aynı anahtara sahip olan veriler gruplandıktan (shuffling: karıştırma) sonra da fonksiyona (toplam) tabi tutulur (reducing: indirgeme). Özetle, Eşle (Map) fonksiyonunun çıktısı İndirge (Reduce) fonksiyonu için girdidir yani Eşle fonksiyonu ile gruplanan veriler, İndirge fonksiyonu ile sayılarak raporlanmaktadır. Sonuçlar ise yine HDFS'de tutulmaktadır (Selçuk, 2015).



Şekil 2.6. Eşle İndirge Algoritmasının Çalışma Şekli (Selçuk, 2015)

2.1.3. Hadoop Eşle İndirge ve Apache Spark

HDFS ve HBase gibi dosya sistemlerini kullanan Spark, Scala dili ile yazılmış, bellek içi, gerçek zamanlı ve akan veri işleme özelliğine sahip, açık kaynak kodlu veri işleme aracıdır. Spark'ta akan veri madenciliği (Spark Streaming, SQL, GraphX) ve makine öğrenimi (Mlib) için geliştirilmiş çeşitli kütüphaneler (Spark MLlib, Flink ML) de bulunmaktadır (Özdeş, 2017; Gu ve Li, 2013).

Eşle İndirge ile olan farkı, Eşle İndirgenin, dağıtık sistemler üzerinde paralel olarak, yığın veri işleme yapılabilmesi için tasarlanmış olmasıdır. Eşle İndirge, çoğunlukla raporlama, öngörücü modelleme ve makine öğrenimi gibi alanlarda kullanılmaktadır (Dean ve Ghemawat, 2008; Kayabay vd., 2016). Spark ise akan veri işleme için geliştirilmiştir (Cugola ve Margara, 2012); RAM (Random Access Memory – Rastgele Erişimli Bellek) üzerinde çalıştığı için, disk üzerinde çalışan Hadoop Eşle İndirge'den yaklaşık 100 kat daha hızlı veri işleme yapabildiği bulgulanmıştır (Karabay ve Ulaş, 2018; Pol, 2016; Zaharia vd., 2012; Farhan vd., 2018). Benzerlikleri ve farklılıklarıyla, Eşle İndirge ve Spark özellikleri aşağıdaki tabloda listelenmiştir.

Tablo 2.2. Eşle İndirge ve Spark Karşılaştırması (Karabay ve Ulaş, 2018)

	<i>Hadoop Eşle İndirge</i>	<i>Apache Spark</i>
<i>Performans</i>	Düşük	Yüksek
<i>Ölçeklenebilirlik</i>	Var	Var
<i>Hata Toleransı</i>	Var	Var
<i>Veri Tabanı Teknolojisi</i>	Disk üzerinde	Bellek içi
<i>Açık Kaynak Kodlu</i>	Evet	Evet
<i>Dağıtık Dosya Sistemi</i>	Evet	Evet
<i>Desteklediği Veri Formatı</i>	Yapılandırılmış, Yarı Yapılandırılmış, Yapılandırılmamış	Yapılandırılmış, Yarı Yapılandırılmış, Yapılandırılmamış
<i>Gerçek Zamanlı Analiz</i>	Yok	Var
<i>Lisans</i>	Ücretsiz	Ücretsiz
<i>Paralel İşleme</i>	Var	Var
<i>Desteklediği İşletim Sistemi</i>	Windows, Linux, Mac OS	Windows, Linux, Mac OS
<i>Çevrim İçi Çalışabilme</i>	Bulut üzerinde	Bulut üzerinde
<i>Kullanım Amacı</i>	Veri analizi, veri işleme	Veri işleme, veri yönetimi
<i>ETL Desteği</i>	Var	Var

Tablo 2.2. (Devamı)

	<i>Hadoop Eşle İndirge</i>	<i>Apache Spark</i>
<i>Veri Saklama Teknolojisi</i>	Cassandra, Couchbase, DynamoDB, MongoDB, Neo4j, HBase	HBase, MongoDB
<i>Dosya Sistemi</i>	HDFS, Amazon S3	HDFS, MapR File System, Amazon S3
<i>Akan Veri Desteği</i>	Yok	Var
<i>Programlama Dili Desteği</i>	C++, Java, Python, PHP, Ruby, R	Java, Python, R, Scala

2.1.4. Hadoop'un Avantajları

Hadoop'un avantajları, düşük maliyetli çözüm, hatadan kurtarma, esneklik ve ölçeklenebilirlik olarak özetlenebilir. Hadoop büyük hacimdeki verileri işlerken, yüksek hesaplama gücüne ve büyük depolama ünitelerine sahip makinalara ihtiyaç duymayarak düşük maliyetli bir çözüm sunmaktadır. Hadoop, herhangi bir makine çöktüğünde işi diğer makinalara dağıtarak iş sürekliliği sağlamaktadır. Yapısal olarak farklı veri türlerini birlikte işleyebilme esnekliğine sahiptir. Son olarak ölçeklenebilirlik avantajı, yürürlükteki işlerin aksamasına meydan vermeden ve mevcut sistemde herhangi bir değişiklik yapmayı gerektirmeden yeni düğüm noktalarının kümeye dâhil edilebilmesidir (Doğan, 2014).

Hadoop mimarisi, çoğunlukla veri yoğun projelerde kullanılmasına rağmen hata tespit ve ayıklama kolaylığı sebebiyle yüksek hesaplama gerektiren projelerde de kullanılmaktadır (Kayım, 2015).

2.1.5. Hadoop ile İlişkili Projeler

Tablo 2.3. Hadoop ile İlişkili Projeler (Schneider, 2012)

<i>Proje Adı</i>	<i>Açıklama</i>
<i>Ambari</i>	Hadoop kümesinin kurulum, görüntüleme ve bakımını sağlayan yönetim araçlarının bir setidir.
<i>Avro</i>	Veriyi ikili formata dönüştüren bir serileştirme(serialization) framework'udur.
<i>Flume</i>	Log verilerinin Hadoop'a hareketini sağlayan veri akış servsidir.
<i>HBase</i>	HDFS'yi kullanan dağıtık sütun bazlı veritabanıdır.
<i>HCatalog</i>	Hadoop'da kayıtlı verilerin ilişkisel görünümünü sağlayan bir servis
<i>Hive</i>	HDFS'de kayıtlı veriler için dağıtık bir veri ambarıdır. SQL temelli sorgulama yeteneği bulunmaktadır.

Tablo 2.3. (Devamı)

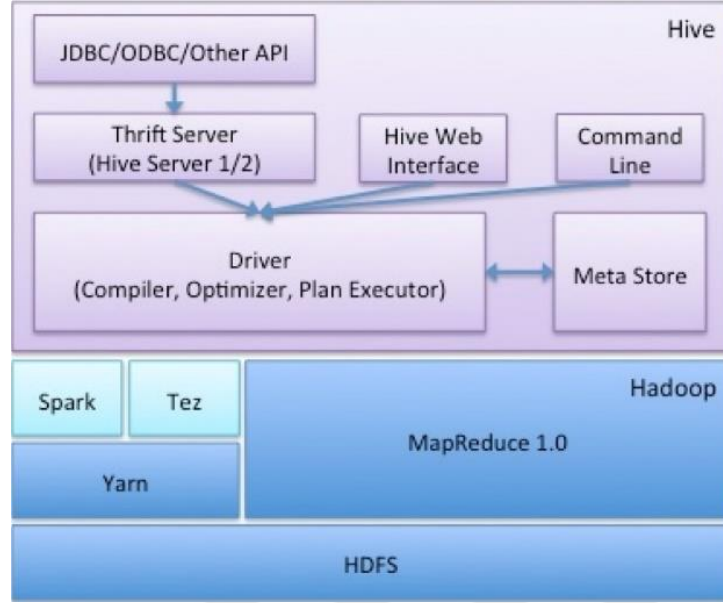
<i>Proje Adı</i>	<i>Açıklama</i>
<i>Hue</i>	Hive ve pig sorgularını yayınlayan, Oozie iş akışlarını geliştiren ve dosyaları görüntüleyen Hadoop yönetim ara yüzüdür.
<i>Mahoot</i>	Hadoop üzerinde doğal olarak çalışabilen, makine öğrenmesi algoritmalarını barındıran bir kütüphanedir.
<i>Oozie</i>	Hadoop uygulamalarını zamanlayan ve birbirine bağlayan bir iş akış yönetimi aracıdır.
<i>Pig</i>	Pig Latin sorgulama dili, Mapreduce komut dizileri ve derleyici altyapısı ile HDFS'deki büyük miktardaki verileri analiz etmeye yarayan bir platformdur.
<i>Sqoop</i>	İlişkisel veri tabanları ile HDFS arasında büyük miktardaki verilerin etkili bir şekilde taşınmasını sağlayan bir araçtır.
<i>ZooKeeper</i>	Dağıtık uygulamalar tarafından kullanılan senkronizasyon ve konfigürasyon benzeri servislerin koordinasyonunu sağlayan basit bir ara yüzüdür.

2.2. HIVE

Facebook tarafından 2008 yılında geliştirilen Hive ile, SQL benzeri kodlar yazılarak HDFS'de tutulan veriler üzerinde sorgulama yapılabilir (Kayım, 2015). Sorgulama, gruplama, filtreleme ve Çevrimiçi Analitik İşleme (OLAP: Online Analytic Processing) fonksiyonlarındaki benzerlikleriyle yaygın kullanıma sahip olan SQL' e çok benzemesi, Hive' nin tercih edilme sebepleri arasında sayılabilir. Çünkü çoğu veri ambarı uygulaması, SQL tabanlı ilişkisel veri tabanı altyapısı kullanmaktadır (Capriolo, Wampler ve Rutherglen, 2012).

Hive, verilerin farklı boyutları kullanılarak, raporlanması esasına dayanan OLAP için uygun ise de Çevrimiçi Hareket İşleme (OLTP: Online Transaction Processing) fonksiyonları için uygun değildir. Hive' de, SQL' deki update (güncelle) ve delete (sil) komutlarından ziyade select (seç) ve insert (ekle) komutlarına odaklanılmıştır. Hive, dosya formatı olarak Json; veri tipi olarak da map, struct ve array tiplerini desteklemektedir (Du, 2015).

Hive, optimize edilmiş index desteği, sorgu vektörizasyonu, maliyet temelli optimizasyon ve TEZ motoru sayesinde daha iyi performans sunmaktadır. ACID özelliklerini kullanarak veri bütünlüğünü garantilemesi dolayısıyla da işletmeler tarafından tercih edilmektedir. Ayrıca nesne bazlı güvenlik özelliği sayesinde işletme bütününde güvenliği temin etmektedir (Binny Mathews, 2017).



Şekil 2.7. Hive Mimarisi (Du, 2015)

Yukarıdaki diyagramda Hadoop ekosisteminde Hive mimarisi görülmektedir. “Metastore”, veri tabanı bilgilerini (tabloların şema bilgilerini) tutmaktadır. Gömülü, yerel ve uzak olmak üzere üç tür “metastore” bulunmaktadır (Bansal, Chauhan ve Mehrotra, 2016). Çalıştırma birimi (Execution) ise HiveQL sorgularını Eşle İndirge kodlarına dönüştürmektedir (Araç, 2016).

2.2.1. Veri Tipleri

Hive tarafından kullanılan bazı veri tiplerinin SQL dilinde karşılığı bulunurken (basit veri tipleri), Hive, SQL’de olmayan kompleks veri tipleri için de destek vermektedir. Bu kompleks veri tipleri ile birden fazla veri tipi birleştirilebilmektedir. Hive’da kullanılan basit ve kompleks veri tipleri Tablo 2.4’ de gösterilmektedir.

Tablo 2.4. Veri Tipleri (Aodba, 2016)

	Kategori	Tip Adı	Açıklama	Boyut
Basit Veri Tipi	Sayısal	Tinyint	İşaretli tam sayı	1 byte
		Smallint	İşaretli tam sayı	2 byte
		Int	İşaretli tam sayı	4 byte
		Bigint	İşaretli tam sayı	8 byte
		Float	İşaretli ondalıklı sayı	4 byte
		Double	İşaretli ondalıklı sayı	8 byte
		Decimal	İşaretli ondalıklı sayı	38 bit
	Karakter	String	Karakter katarı	
		Char	Sabit uzunluklu	255 karakter
		Varchar	Değişken uzunluklu	65535 karakter
Tarih/saat	Timestamp	Tarih ve saat		
	Date	Tarih	0000-01-01 ile 9999-12-31 arasında	
Kompleks Veri Tipi	Diğer	Boolean	Mantıksal veri tipi	
		Binary	İkili sayma	
		Array	Dizi	
		Map	Anahtar-değer çifti	
		Struct	C'deki struct benzeri	
		Uniontype	Birden fazla veri tipinin birleşimi	

2.2.2. Aritmetik ve Mantıksal İşlemler

Tablo 2.5. Aritmetik, Mantık ve İlişkisel Operatörler (Markovitz, 2017)

	<i>Kategori</i>	<i>Tip Adı</i>	<i>Açıklama</i>
<i>İlişkisel Operatör</i>		A=B	Eşitse
		A!=B	Eşit değilse
		A<B	Küçükse
		A<=B	Küçük veya eşitse
		A>B	Büyükse
		A>=B	Büyük veya eşitse
		A is null	A null ise
		A is not null	A null değilse
		A Like B	A, B'ye benzerse
		A rlike B	A'nın alt stringi B'nin regular expressionu ile eşleşirse
		A Regexp B	Rlike benzeri ise
<i>Aritmetik Operatör</i>		A + B	Toplam
		A - B	Fark
		A * B	Çarpım
		A / B	Bölüm
		A % B	Mod
<i>Bit İşlem</i>		A & B	Ve
		A B	Veya
		A ^ B	Xor
		~ A	Değil

Tablo 2.5. (Devamı)

<i>Kategori</i>	<i>Tip Adı</i>	<i>Açıklama</i>
<i>Mantıksal Operatör</i>	A AND B	A ve B doğruysa
	A && B	A ve B doğruysa
	A OR B	A veya B doğruysa
	A B	A veya B doğruysa
	NOT A	A doğru değilse
	! A	A doğru değilse

2.2.3.HiveQL Dili

HQL (Hive Query Language) de denilen ve SQL'e çok benzeyen bu dil sayesinde Hadoop ortamında Eşle İndirge programlara çevrilen komutlar yazılabilmektedir. HiveQL dilinde yazılan komutların bitiminde noktalı virgül kullanılması ise zorunludur.

Tablo yapısı SQL'den farklılık arz etmektedir. SQL' de "schema on write" yapısı kullanılırken, Hive'da "schema on read" yapısı kullanılmaktadır. Yani SQL'de tablo şeması (satır ve sütun yapısı) bellidir ve veriler bu şemaya göre kaydedilir. Hive'da ise gelişmiş güzel kaydedilen veriler okunurken şemaya uydurularak okunur.

Bu bölümde, sıklıkla kullanılan kod örnekleri anlatılacaktır. İlk olarak bir tabloyu şema olarak oluşturmak için:

```
Create Table TabloAdı(
  Kolon1 Veritipi,
  Kolon2 Veritipi
)
Partitioned by (kolon1 veritipi, kolon2 veritipi);
```

Tablo şemasını ortadan kaldırmak için:

```
Drop Table TabloAdı;
```

Tablo oluştururken, bölünmüş tablo oluşturmak için "partitioned" ifadesinden faydalanılmaktadır. Bu sayede tabloda yer alan veriler birden fazla diske dağıtılabilir. Mevcut tablo adlarını görüntülemek için de:

```
Show tables;
```

Herhangi bir tablonun kolon adlarını ve tiplerini görüntülemek için:

```
describe TabloAdı;
```

Tablo adını değiştirmek için:

```
ALTER TABLE TabloAdı1 RENAME TO
```

Tabloya kolon eklemek için:

```
ALTER TABLE TabloAdı ADD COLUMNS (KolonAdı  
VeriTipi);
```

Tabloya veri yüklemek için:

```
Load data local inpath 'dosya yolu'  
Overwrite into table TabloAdı  
Partition (kolon1=değer1, kolon2=değer2 ...)
```

Sorgu sonucunda dönen değerleri tabloya yüklemek için:

```
Insert overwrite table TabloAdı  
Partition (kolon1=değer1, kolon2=değer2 ...)  
If not exists Selectİfadesi;
```

Tablodaki veriler fiziksel dosyaya da çıkarılabilmektedir:

```
Insert overwrite local directory 'KlasörAdı'  
Selectİfadesi;
```

Bişimlendirilmemiş veriler üzerinde işlem yapmak için:

```
Create external table if not exists TabloAdı(  
Kolon1 Veritipi,  
Kolon2 Veritipi  
)  
Row format serde 'serde dosyası'  
Location 'dosyaKonumu';
```

HDFS'deki kayıt dosyalarının (log) silinmemesi için "external" ifadesi kullanılır.
Birden fazla tablodan sorgulama yapmak (join) için:

```

Select tablo1.sütun1, tablo1.sütun2,
tablo2.sütun2
From tablo 1
Join tablo2 (tablo1.sütun1= tablo2.sütun1);

```

Son olarak basit bir kelime sayma örneği HiveQL dili ile aşağıdaki gibi kodlanabilir:

```

/* docs tablosunu yarat */
CREATE TABLE docs (line STRING);

/* docs dosyasini docs tablosu icerisine aktar */
LOAD DATA INPATH 'docs' OVERWRITE INTO TABLE docs;

/* kelimeleri sayarak yeni bir tablo olustur */
CREATE TABLE word_counts AS
SELECT word, count(1) AS count FROM
(SELECT explode(split(line, '\s')) AS word FROM docs) w
GROUP BY word
ORDER BY word;

```

Şekil 2.8. Kelime Sayma Örneği (İlter, 2013)

2.3. DUYGU ANALİZİ

İnsanlar duygularını yüz ifadeleriyle, baş ve vücut hareketleri ile veya statik olarak göstermektedirler. Konuşurken de seslerini değiştirerek veya bazı kelimeleri vurgulayarak duygu durumlarını (sakin, öfkeli, heyecanlı vs.) yansıtabilmektedirler. Düz metinde ise bunlardan hiçbirine yer verilememektedir (Çavdaroğlu, 2014). Dolayısıyla yazarken, ünlem ve soru işareti gibi noktalama işaretleri; gülen yüz, kızgın yüz gibi semboller kullanılarak yazının monotonluğundan kurtulmak ve yüz yüze iletişimin samimiyeti yakalanmak istenmektedir (Karakulakoğlu ve Uğurlu, 2015).

İnsanlar, satın aldıkları ürün ve hizmetler hakkında olumlu veya olumsuz görüşlerini sosyal medya aracılığıyla paylaştıklarında ise hangi duyguyu taşıdıkları açıkça bilinmemektedir. Hatta bazen insan gözüyle dahi duygunun tespit edilmesi mümkün değildir. Yine de tespit edilmeleri önemlidir çünkü bu paylaşımlar, işletmelerin başarıları hakkında somut bir değerlendirmeyi ortaya koymaktadır. Twitter kullanımları üzerine yapılan çalışmalarda insanların çoğunlukla pozitif tweet attığı ve çoğunlukla pozitif tweetleri retweet yaptığı ancak, negatif tweet atanların pozitif tweet atanlara nispetle daha çok paylaşım yaptığı bulgulanmıştır (Quercia vd., 2011; Gruzd vd., 2011).

İşletmeler ise müşterilerinin düşüncelerini öğrenebilmek için önceleri, geri bildirim formları ve anketler için fon ayırmaktaydılar. Ancak geniş bir müşteri kitlesi söz konusu olduğunda geleneksel istatistiksel yöntemler yetersiz kalmaktadır. İşte tam da bu noktada sosyal medyada duygu analizi, müşteri geri bildirimlerini alabilmek için bulunmaz bir fırsat sunmaktadır. Gerçi bireylerin demografik özellikleri ve tek tek ne düşündüğü dikkate alınamamaktadır fakat duygu durum analiziyle kitlesel eğilimin ne olduğu ortaya konulabilmektedir.

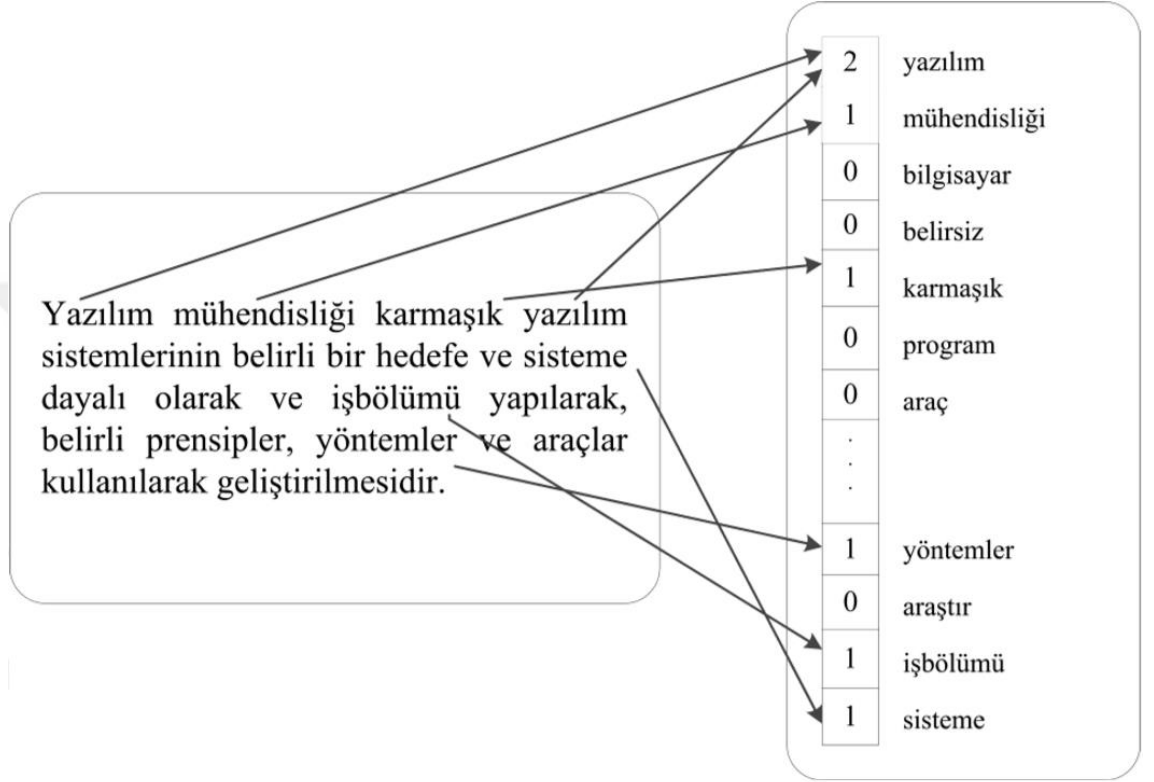
Duygu analizi, sosyal medya verilerinin analiz edilmesinde en yaygın kullanılan fikir madenciliği çalışması olması yönüyle sosyal medyada herhangi bir konu, kişi veya olay hakkında paylaşılan görüş ve değerlendirmelerin, tutum ve duyguların, analiz edilmesinde kullanılmaktadır (Liu, 2012). Duygu analizi ile sosyal medya paylaşımları yani metinler, bilgisayar yazılımları ile işlenerek olumlu, olumsuz ve yansız (nötr) gibi etiketlerle temsil edilen sınıflara ayrılmaktadır (Uçan, 2014).

Duygu analizinde, bağımsız değişkenlerin değerlerinden hareketle bağımlı değişkenin değerinin belirlenmesi için istatistiksel yöntemler ve makine öğrenmesi yöntemleri kullanılmaktadır (Bayrakçı, 2015). Vural ve arkadaşları (2013) tarafından yürütülen bir duygu analizi çalışmasında mesela, bu iki yöntem birlikte kullanılmıştır. Makine öğrenmesi teknikleri içerisinde ise en başarılı yöntemin destek vektör makineleri olduğu bazı çalışmalarda ortaya konulmuştur (Oğul ve Ercan, 2016; Akba vd., 2014). Onan (2017) ise Twitter’lerin duygu analizinde gizli drichlet tahsisine dayalı Naive Bayes yöntemini, destek vektör makinelerinden daha başarılı (%78.3) bulmuş. Türkiye’deki üç GSM operatörü hakkında paylaşılan tweetlerin analiz edildiği bir çalışmada ise destek vektör makinesi, genetik algoritma ile birlikte uygulanmıştır (Yelmen 2016).

Sözlük tabanlı yöntem olarak da bilinen istatistiksel yöntem uygulanırken, pozitif veya negatif bir puana sahip olan kelimeleri içeren bir sözlükten faydalanılmaktadır. Burada ayrıca bir eğitim kümesine ihtiyaç duyulmamaktadır. Ancak sözlük tabanlı “duygu analizi yöntemlerinde bir ifade, bir bağlamda olumlu duygu barındırırken başka bir bağlamda olumsuz duygulara karşılık gelebilmektedir.” (Uçan, 2014 s.4). Bu da sözlük tabanlı duygu analizi için bir handikaptır.

Makine öğrenmesi tekniği ile gerçekleştirilen duygu analizinde ise karakter ve kelime düzeyinde n-gram teknikleri ve BOW (Bag Of Words – Kelimeler Torbası) tekniği

ile öznitelik çıkarılmaktadır. Kelimeler torbası modelinde metin içinde geçen kelimelerin her birinin kaçar defa geçtiği bilgisi elde edildiği ve metni oluşturan kelimelerin sırası dikkate alınmadığı için anlam kaymasına sebep olabilmektedir. Bunun önüne geçmek için de kelime düzeyinde n-gram (word-gram) yöntemi tercih edilmektedir (Türkmenoğlu, 2015).



Şekil 2.9. BOW Modeli Örneği (Çoban 2016)

Word-gram modelinde her bir öznitelikte birkaç kelime birlikte kullanılmaktadır. Model isimlendirilirken de kullanılan kelime sayısı temel alınır. N, 2 ise mesela bigram, 3 ise trigram diye isimlendirilmektedir. Kelime düzeyinde ngram modeli yani word-gram, “yazılım mühendisi olmak zordur” örnek ifadesi için aşağıdaki gibi oluşturulabilir:

Unigram: yazılım – mühendisi – olmak – zordur

Bigram: yazılım mühendisi – mühendisi olmak – olmak zordur

Trigram: yazılım mühendisi olmak – mühendisi olmak zordur

Fourgram: yazılım mühendisi olmak zordur

Karakter düzeyde ngram ise, önışlem gerektirmemesi ve dilden bağımsız olmasıyla; yazım hatalarına ve noktalama işareti kullanımlarına karşı bağımsız olmasıyla öne çıkmaktadır. Karakter düzeyde unigram yani tek karakter modeli çıkarılamaz (Cavnar ve Trenkle 1994; Kanaris vd., 2007). “Yazılım mühendisliği” örnek ifadesinin karakter düzeyde bigram ve trigram gösterimleri (gösterimde boşluk karakteri yerine “_” karakteri kullanılmıştır) ise şöyledir:

Bigram: ya-az-zı-ıl-lı-lım-m_-m-mü-üh-he-en-nd-di-is-sl-li-iğ-ği

Trigram: yaz-azı-zıl-ılı-lım-lım_-m_m_-mü-müh-ühe-hen-end-ndi-dis-isl-sli-liğ-iği

BOW yönteminin uygulandığı metin sınıflandırma çalışmaları içinde, beklenti maksimizasyonu algoritmasının (Nigam vd., 2000), kümeleme tabanlı k-nn algoritmasının (S. Jiang vd., 2012), benzerlik ölçüm yönteminin (Lin vd., 2014) önerildiği; SVM tekniğinin (Argamon vd., 2007; Ko, 2012; Amasyalı, 2012) ve NB tekniğinin (Kaşıkçı ve Gökçen 2014) uygulandığı çalışmalar bulunmaktadır.

Duygu analizi çalışmaları ise her dil için ayrıca yapılmaktadır. Literatür tarandığında görülmektedir ki Türkçe ile ilgili sınırlı sayıda çalışma bulunmaktadır (Çoban, Özyer ve Özyer, 2015) ve çoğu duygu analizi çalışması İngilizce için yapılmıştır. Türkçe için yapılan çalışmalarda ise Wordnet’in Türkçeleştirilmesine yönelik çalışmalar (Amasyalı, 2005; Bilgin, Çetinoğlu ve Oflazer, 2004), duygunun uluslararası olduğu kabulüyle otomatik sözlük çevirme çalışmaları (Uçan, 2014), yabancı dilde hazırlanmış duygu analizi kütüphanesinin (sentistrength) Türkçeye çevrildiği (Vural, Cambazoglu, Senkul ve Tokgoz, 2013) ve SentiTurkNet’in geliştirildiği (Dehkharghani vd., 2016) çalışmalar bulunmaktadır.

Türkçe tweetlerin analiz edildiği bazı duygu analizi çalışmalarında (Çetin ve Amasyalı, 2013; Türkmenoğlu, 2015; Sevindi, 2013) makine öğrenmesi yaklaşımının, sözlük tabanlı yaklaşımdan daha başarılı olduğu, diğer bazı çalışmalarda (Akgül, Ertano ve Banu, 2016; Uçan, 2014) ise aynı başarı düzeyinde olduğu bulgulanmıştır. Mulki ve arkadaşları (2018) ise ister sözlük tabanlı ve ister makine öğrenmesi tabanlı olsun ön işlemenin (yazım hatalarının düzeltilerek normalleştirilenin) duygu analizine olumlu etkisi olduğunu tespit etmiştir. Zira duygu analizi için metinlerin, ilgili dilin gramer kurallarına uygun yazılmış olması gerekmektedir. Bu bakımdan özellikle sosyal medya

yazışmaları üzerine Türkçe için yapılan duygu analizi çalışmalarında ön işleme bağlamında doğal dil işleme de yapılmaktadır.

Çelikyılmaz ve arkadaşları (2010) makine öğrenmesi teknikleriyle yaptıkları analiz çalışmalarının başarısını, elle etiketledikleri tweetler üzerinden test etmişlerdir. Destek vektör makinelerini kullanan Yıldırım ve arkadaşları (2015) ise, 12790 tweeti elle etiketledikleri çalışmalarında en iyi başarıyı (%79), en az öz nitelik sayısı ile elde etmişlerdir. Destek Vektör Makinesi kullanılarak TV programlarının reyting tahminlerini %83'lük başarı ile tahmin edebildiği çalışmada Akarsu (2016), etki puanını hesaplarken tweetlerin retweet sayıları ve favori yapıma sayılarını da dikkate almıştır.

Çoban ve Özyer (2015), denetimli makine öğrenmesi tekniklerini uyguladıkları çalışmalarında, literatürdeki bazı çalışmalarda (Go vd., 2009; Pak ve Paroubek, 2010; Davidov vd., 2010) olduğu gibi, eğitim veri setini oluştururken duygu simgelerinden faydalanmışlar. Duygu analizi için n-gram yönteminin BOW yönteminden daha başarılı sonuç verdiğini rapor etmişler ve ardışık kullanılan kelimelerin duyguyu anlatmada daha çok anlam yüklü olduğu gerçeğine vurgu yapmışlardır (Çoban vd., 2015). Akgül ve arkadaşları (2015) tarafından yapılan çalışmada ise sözlük tabanlı yöntemde %70 ve karakter düzeyde n-gram yönteminde %69'luk başarı elde edilmiştir. Pang ve arkadaşları ise (2002) literatürdeki ilk duygu analizi çalışmasını 2002 yılında gerçekleştirirken, çeşitli makine öğrenmesi tekniklerini denedikleri çalışmalarında unigram yöntemini diğer tekniklerden başarılı bulmuşlardır. Parlak ve Diri (2013) ise insan sesinin yazıdan daha fazla duygu içerdiği kabulü ile yaptıkları çalışmalarında duygu analizi için insan sesinden faydalanmışlardır.

Boynukalın ve Karagöz (2013) ise Türkçe metinleri, öfke, üzüntü, korku ve sevinç olarak dört sınıfa ayırmışlardır. Literatürde çoğunlukla ikili (olumlu ve olumsuz) (Piryanı vd., 2017; Schumaker vd., 2016) ve üçlü sınıflandırma (olumlu, olumsuz ve yansız) (Nizam ve Akın, 2014) çalışmaları bulunmaktadır. Yüksel ve Tan (2018) ise 3'lü sınıflandırmanın (%84,5) ikili sınıflandırmadan (%82) daha başarılı olduğu sonucuna varmışlardır. Nizam ve Akın (2014) ise denetimsiz makine öğrenmesi uyguladıkları 3 sınıflı duygu analizi çalışmalarında sınıfların veri kümesi içinde eşit dağılmasının analizi olumlu etkilediğini tespit etmişler.

Tweetlerin duygu durumunun analiz edilmesine odaklanan diğer çalışmalara bakıldığında, turistler tarafından paylaşılan 80 milyon tweet üzerinde duygu durum analizi yapan bir çalışmada (Claster, Dinh ve Cooper, 2010) duygu belirleyici kelime kütüphanesinden faydalanılmış ve tweetler Naive Bayes ile sınıflandırılmış. Simşek ve Özdemir (2012) ise borsa ile Twitter duygu analizi arasında %40 korelasyonun olduğunu tespit ettikleri çalışmalarında, tweetleri sınıflandırmak için sekiz farklı duygu etiketine sahip 113 duygu belirleyici ifadeden faydalanmıştır.

70 milyon tweet üzerinde yapılan başka bir çalışmada ise Naive Bayes ve Yapay Sinir Ağları kullanılmış (Claster, Hung ve Cooper, 2010). Twitter üzerine yapılan diğer çalışmalara bakıldığında, salgın hastalıkları önceden tahmin etmeye çalışan (Szomszor, Kostkova ve De Quincey, 2010), ilaçların yan etkilerini araştıran (Bian, Topaloglu ve Yu, 2012), zaman serilerini kullanarak duygu dinamiklerini tahmin eden (Nguyen, Wu, Chan, Peng ve Zhang, 2012), borsa hareketlerini tahmin etmeye çalışan (Bollen, Mao ve Zeng, 2011) ve tv program reyting değerleriyle Twitter analizi sonuçlarını karşılaştıran (Akarsu ve Diri, 2016; Kayahan, Sergin ve Diri, 2013) çalışmalar bulunduğu görülmektedir.

Duygu analizi belge, cümle ve varlık seviyelerinde uygulanabilmektedir (Medhat, Hassan ve Korashy, 2014). Belge düzeyinde analiz yapıldığında daha fazla hesaplama maliyeti gerektirirken, başarımlı artmaktadır. Twitter ise karakter kısıtlamasıyla bir bakıma cümle düzeyinde analizi zorunlu kılmaktadır. Abdalnabi (2016) ise bu bağlamda Twitter gibi kısa cümlelerden ibaret metinlerin sınıflandırılması için Konvolüsyonel Sinir Ağı (convolutional neural network) mimarisi önermiştir.

Tweetlerin etki ağırlıklarının tespit edildiği Kwak ve arkadaşları (2010) tarafından gerçekleştirilen çalışmada takipçi sayısı ve retweet sayısından faydalanılmış. Cha ve arkadaşları (2010) ise etki ağırlığına, bahsedilme (mention) sayısını da dâhil etmişlerdir. Sosyal ağı etkileme gücü yüksek olan kullanıcıların tespit edildiği çalışmalarda ise retweet ve mention sayısının takipçi sayısından önemli olduğu tespit edilmiş (Subbian ve Melville, 2011; Weng vd., 2010).

Psikoloji bilimine göre insanlar konuşurken ve yazarken içinde bulundukları duygu durumuna bağlı olarak kelimeleri seçmekte ve farklı şekillerde kullanabilmektedir (Thelwall, 2010). Twitter kullanıcıların duygu durumlarıyla, kullandıkları kelimeler

arasındaki bağlantıyı Türkçe metinler üzerinden araştırdığı çalışmasında ise Albayrak (2011), duygu durumlarıyla kelimeler arasında anlamlı bir ilişki tespit etmiştir.

Akbaş (2012) ise Türkçe tweetler için alan bağımlı (tweetleri konularına göre gruplandırarak) duygu analizi yapabilen bir sistem geliştirmiştir. Meriç ve Diri (2014) ise denetimli makine öğrenmesinin, karakter tabanlı n-gram yöntemiyle alan bağımlı veri setini, sözcük tabanlı yöntemiyle de alan bağımsız veri setini daha başarılı sınıflandırdığını tespit etmişlerdir. L. Jiang ve arkadaşları (2011) ise alan bağımlı çalışmanın daha doğru sınıflandırma başarısı sağlayacağını söyleyerek, odaklanılan konu dışında içeriklerin de tweet içinde bulunabilmesi dolayısıyla çalışmanın zorluğuna dikkat çekmişlerdir.

2.4. MAKİNE ÖĞRENMESİ

Makine öğrenmesi, problemi ortamdan edindiği bilgiye göre modelleyebilen Yapay Zekâ disiplininin bir alt dalıdır (Nizam ve Akın, 2014). Bu teknikte mevcut veriler ile eğitilen bilgisayar yeni verileri sınıflara ayırmaktadır. Yani bir nevi mevcut durumun anlaşılması ve öğrenilmesi ile geleceğin tahmin edilmesine çalışılmaktadır.

Makine öğrenmesi, danışmanlı ve danışmansız öğrenme olarak ikiye ayrılmaktadır. Danışmanlı öğrenmede (supervised learning) mevcut veriler sınıflara ayrılmış olarak sisteme verilirken; danışmansız öğrenmede ise sınıfların tahmin edilmesi tamamıyla sisteme bırakılmıştır (Hallaç, 2014).

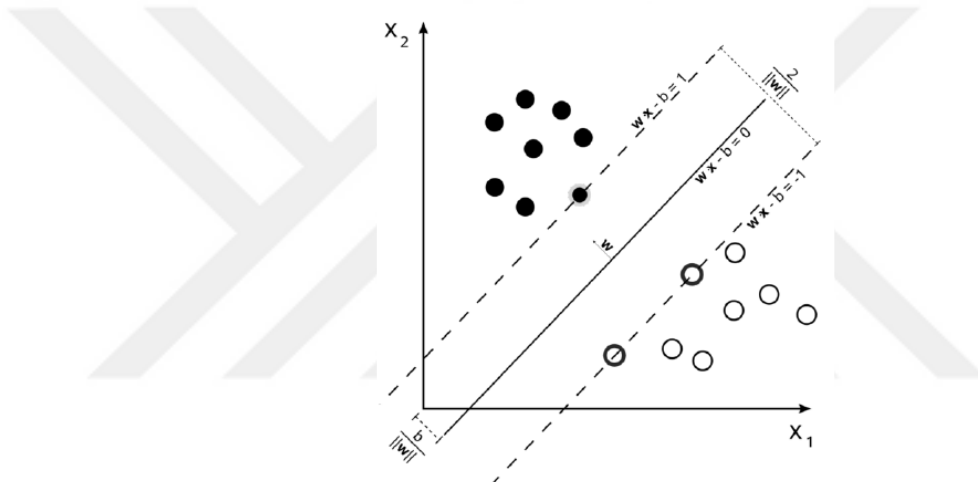
Danışmanlı veya diğer adıyla denetimli makine öğrenmesinde öncelikle eğitim için kullanılacak büyük bir veri setinin oluşturulması gerekmektedir. Bu veri setinde her bir örnek için ait olduğu sınıf etiketi tanımlı olmalıdır. Eldeki veriler kullanılarak sınıfların özellikleri ve örüntüleri ortaya çıkarıldıktan sonra yeni veriler sınıflandırılabilir (Sarkar, Bali ve Sharma, 2018).

Makine öğrenmesi için Destek Vektör Makineleri yöntemi ve Naive Bayes yöntemi sıklıkla kullanılan yöntemlerdir.

2.4.1. Destek Vektör Makinesi

Herhangi iki sınıfı birbirinden ayırabilecek sonsuz sayıda çizgi çizilebilir ancak destek vektör makinesi (SVM: Support Vector Machine), sınıflar arasındaki mesafenin

maksimum olmasını sağlayacak şekilde sınır çizgileri çeker (Şekil 2.9) (Elmas 2012; Wang 2005). “Destek vektör makinesi iki boyutlu uzayda doğrusal, üç boyutlu uzayda düzlemsel ve çok boyutlu uzayda hiperdüzlem şeklindeki ayırma mekanizmaları ile veriyi iki ya da daha çok sınıfa ayırma yeteneğine sahiptir.” (Güran, Uysal ve Doğrusöz, 2014 s. 87). Sınıfların doğrusal olarak birbirinden ayıramadığı problemler içinse doğrusal olmayan (nonlinear) sınıflandırma kullanılmaktadır. Makine öğrenmesi algoritmalarından ve veri madenciliğinde kullanılan sınıflandırma (classification) yöntemlerinden biri olan destek vektör makinesi, doğrusal (lineer) olmayan örnek uzayını, bir üst boyuta aktararak sınıfların doğrusal olarak ayrılabilmesini sağlar (Türkmenoğlu, 2015).



Şekil 2.10. Destek Vektör Makinesi (Türkmenoğlu, 2015)

2.4.2. Naive Bayes Sınıflandırıcısı

Danışmanlı makine öğrenmesi yöntemlerinden biri olan Naive Bayes sınıflandırma yöntemi Bayes teoremine dayanmaktadır. i ve j , 1 ile n arasında birer sayı olmak üzere; x_i nicelik vektörü, ω_j sınıfı temsil ediyor olsun. Dolayısıyla $P(x_i|\omega_j)$, ω_j sınıfına ait olduğu için x_i örnekleminin gerçekleşme olasılığı yani bağıl olasılıktır. Buna göre,

$$P(\omega_j|x_i) = \frac{P(x_i|\omega_j)P(\omega_j)}{P(x_i)} \quad (1)$$

Bayes teorem denkleminde öncül olasılıklar (prior probability) yani ω_j olayının gerçekleşme olasılığı $P(\omega_j)$ ile x_i olayının gerçekleşme olasılığı ise $P(x_i)$ ile temsil edilmektedir (Pawlak, 2003). Buna göre olasılık değerlerinin toplamı 1 olacaktır:

$$1 = \sum_{j=1}^n P(\omega_j) \quad (2)$$

Bayes karar teoremine göre ise x ' in ω_j 'ye ait olma olasılığı aşağıdaki denklemde verilmiştir (Leung, 2007):

$$P(\omega_i|x) > P(\omega_j|x), \forall j \neq i \quad (3)$$

Aynı şekilde 4 numaralı denklem de sağlanacaktır:

$$P(\omega_i) P(x|\omega_i) > P(\omega_j) P(x|\omega_j), \forall j \neq i \quad (4)$$

$P(x|\omega_j)$ terimi yerine yaklaşık değeri yazılırsa;

$$P(x|\omega_i) \approx \prod_{k=1}^n P(x_k|\omega_i) \quad (5)$$

X ' in ω_j 'ye ait olma olasılığı, 6 numaralı eşitsizlik ile hesaplanabilecektir.

$$P(\omega_j) \prod_{i=1}^n P(x_i|\omega_j) > P(\omega_j) \prod_{i=1}^n P(x_i|\omega_j) \quad (6)$$

2.5. DOĞAL DİL İŞLEME

İnsanlar tarafından konuşulan dil, doğal dil olarak adlandırılmaktadır. Bir dilin, bilgisayar algoritmalarıyla analiz edilmesine, anlaşılmasına ve üretilmesine yarayan hesaplamalı dil bilimine de doğal dil işleme (NLP: Natural Language Processing) denilmektedir (Allen 2003).

Doğal dil işleme, özellikle metin madenciliği çalışmalarında yaygın olarak kullanılmaktadır. Metin madenciliği (text mining) ise yapılandırılmamış formatta bulunan metinlerin yapılandırılmış formata dönüştürülerek analiz edilmesi ve veri içindeki gizli örüntülerin elde edilmesi esasına dayanmaktadır (Bayrakçı, 2015; Tan, 1999). Metin madenciliğini veri madenciliğinden ayıran en temel özellik, metin madenciliğinde yapısal olmayan verilerin kullanılmasıdır.

Doğal dil işleme bağlamında 4 farklı seviyede çalışılmaktadır: sözlüksel (lexical), söz dizimsel (syntactic), anlamsal (semantic) ve pragmatik (pragmatic). Sözlüksel seviyede, sondan eklemeli (agglutinative) diller için, kelime kök ve eklerine ayrıştırılarak (biçimbilimsel çözümleyici ile), sözcük türü (POS – Part of Speech) ve anlamı anlaşılmasına çalışılır (Sadredini vd., 2018).

Söz dizimsel seviyede ise cümlelerin dizilişinin anlama yaptığı katkı çözümlenir. Çünkü farklı söz dizimleriyle cümlelerin vurgusu ve anlamı değişebilmektedir. Anlamsal çözümleme ile de metindeki cümlelerde ne anlatılmak istendiği bilgisayar tarafından anlaşılmaaya çalışılır.

Her gün bir insan ölecek $\rightarrow \forall x, \text{gün}(x) \{ \exists (y), \text{insan}(y), \text{ölecek}(y, x) \}$

Bir insan her gün ölecek $\rightarrow \exists y, \text{insan}(y) \{ \forall (x), \text{gün}(x), \text{ölecek}(y, x) \}$

Bir üst seviyede (pragmatik) ise kelimelerin hangi biçimde ve anlamda kullanıldığıyla da ilgilenilir. Örneğin “aslan, ceylanı kovaladı” ve “Aslan, işe geç kaldı” cümlelerindeki “aslan” kelimeleri farklı varlıkları temsil etmektedir.

Türkçe, dil yapısı dolayısıyla (sondan eklemeli olması, olumsuzluk ekleri ve varlık/yokluk ekleri alması) doğal dil işleme çalışmaları açısından zorluk içermektedir (Türkmenoglu ve Tantug, 2014). Bu zorluğu gidermek için (kelimelerin ek ve köklerine ayrılmasında ve hatalı yazılan kelimelerin düzeltilmesinde) birçok Türkçe çalışmada zemberek kütüphanesinden faydalandığı görülmektedir (Meral ve Diri, 2014; Hallaç, 2014).

Türkçe için yapılan çalışmalar incelendiğinde bazı kapsamlı gramer çalışmaları (Güngördü ve Oflazer 1995; Bozşahin 2002) görülmektedir. Özenç (2017) mesela, biçim bilimsel çözümleyici geliştirmiş. M.B. Öztürk (2016) ise geliştirdiği biçim bilimsel çözümleyici modeli ile %82’lik başarıyla sözcük türetebilmektedir. Butt ve King (2005) ise Urdu dili için biçim bilimsel çözümleyici uygulamışlar.

Söz dizimsel düzeyde yapılan bazı çalışmalar bağıllık ayrıştırması alanına odaklanmıştır. İngilizce için Yamada ve Matsumoto’nun (2003), İsveççe için Nivre ve arkadaşlarının (2004), Japonca için Kudo ve Matsumoto’nun (2000, 2002) ve Sekine ve arkadaşlarının (2000), Korece için Chung ve Rim’in (2004), Çekçe için Nivre ve Nilson’ın (2005) çalışmaları bağıllık ayrıştırıcısı çalışmalarına örnek verilebilir. Eryiğit ve arkadaşları ise (2006) Türkçe cümleler için kural tabanlı bağıllık analizi yöntemi geliştirmiştir (Eryiğit vd. 2006). Dönmez ise (2016) Türkçe cümlelerin dil bilgisel ve anlamsal olarak çözümlenmesi için vektör kıyaslanması kullanan bir yöntem geliştirmiştir.

Cümle öğelerinin ağaç yapısıyla analiz edilmesine odaklanan çalışmalar içinde Arapça için (Hajic vd. 2004), Çekçe için (Hajic vd. 2001), Danimarkaca için (Kromann 2003), İtalyanca için (Bosco 2004), Slovakça için (Dzeroski vd. 2006), İsveççe için (Nivre 2006) ve Türkçe için (Eryiğit ve Oflazer 2006) yapılan çalışmalar örnek verilebilir. Yıldız ve arkadaşları (2014) ise Türkçe ve İngilizce için paralel ağaç yapısı geliştirmiştir.

Doğal dil işleme kapsamında ayrıca diller arası çeviri çalışmaları yapılmaktadır. Bilgisayar tarafından otomatik olarak yapılması istenen çeviri çalışmaları ise istatistiksel ve kural tabanlı olmak üzere iki koldan yürütülmektedir. İstatistiksel yöntemde diller için her bir cümle için diğer dilde karşılığının bulunduğu derlem (paralel metinler) kullanılmaktadır. İstatistiksel yöntemde ayrıca dil bilgisine ihtiyaç duyulmaması bir avantaj ise de çeviri başarısı düşüktür. Kural tabanlı yöntemde ise her iki dilin de dil bilgisi kuralları dikkate alınarak çeviri yapılmaktadır (Çalış 2017).

Türk dilleri arasında yapılan ilk bilgisayarlı çeviri çalışması Hamzaoğlu tarafından Türkçe ve Azerice arasında 1993 yılında gerçekleştirilmiştir (Hamzaoğlu 1993). Başka bir çalışma ise Tatarca ve Türkçe arasında yapılmıştır (Altıntaş ve Çiçekli, 2001).

Çekçe ve Slovakça arasında sözcük bazında yapılan bir çalışmada %90'lık başarı sağlanmıştır (Hajic vd., 2000). Sözlük ve biçim bilimsel üreticinin kullanıldığı başka bir bilgisayarlı çeviri çalışması ise İspanyolca ve Katalanca arasında yapılmıştır (Canals-Marote vd., 2000). Çakıcı (2005) ise Türkçe ağaç yapılı derlem çalışması ile otomatik olarak sözlük üretebilmiştir.

El-Kahlout (2006), yaptığı istatistiksel tabanlı çalışmasıyla Türkçeden İngilizceye çeviri sistemi geliştirmiştir. Bilgi, kural ve istatistik temelli yürüttüğü çalışmasında ise Tantı ve arkadaşları (2010) Türk dilleri arasında çeviri sistemi geliştirmiştir.

Doğal dil işleme çalışmalarının diğer bir ilgi odağı yazım ve söz dizim hatalarının düzeltilmesidir. Twitter gibi sosyal medya yazışmalarındaki hataların düzeltilmesi ise oldukça zor bir iştir. Tweetlerin imla ve dilbilgisi bakımından zayıf oluşu ve kendine özgü jargonunun oluşu, işlenmelerini ve analiz edilmelerini zorlaştırmaktadır. Sosyal medya yazışmalarının doğasındaki bu farklılığı Türksever (2013) şu şekilde ifade etmektedir: “Bir haberi veya bir durumu iletmek için saniyeler içinde yazmamız ve göndermemiz gereklidir. Bu hızlı mesaj iletimi sırasında da dildeki kelimelerde kısaltmalar yapıldığını veya yapılan yanlışların önemsenmediğini, kendini ifade edebilmek için sözlükte

olmayan yeni kavramların iletişim dilimize girdiğini görmekteyiz.” (s. 7). Twitter’in kısıtlı yazıma imkân vermesi dolayısıyla da kullanıcılar kısaltmalar kullanmaktadır.

2.5.1. Türkçe Dil Yapısı Özellikleri

Türkçede 8 sesli harf (a, e, ı, i, o, ö, u, ü) ve 21 sessiz (b, c, ç, d, f, g, ğ, h, j, k, l, m, n, p, r, s, ş, t, v, y, z) harf bulunmaktadır. Sesli harfler ses yolunda hiçbir engele uğramadan oluşurken, sessiz harfler ise ses yolunda bazı engellere uğrayarak oluşur. Bu harflerden bazıları (j, f, p gibi) Türkçe kelimelerde nadiren bulunurken bazıları (a, i, e gibi) da çok yaygın geçmektedir. Yine “ğ” harfi, kelime başında geçmezken; bazı harfler (b, c, d ve g) de kelime sonunda bulunmaz (diyalog, radyolog gibi alıntı kelimeler hariç) (Delibaş, 2008).

Yanlış yazılan kelimelerin düzeltilmesinde harf kullanımı ve hece yapısı dikkate alınması gerekirken; duygu analizi söz konusu olduğunda ise eklerin kullanımı ve cümle kuruluşu da dikkate alınmalıdır.

2.5.1.1. Hece Yapısı

Bir solukta çıkan sese hece denir ve her hecede bir adet sesli harf bulunur. Türkçede altı tür hece yapısı bulunmaktadır (Çınar, 2003).

Tablo 2.5. Hece Yapıları

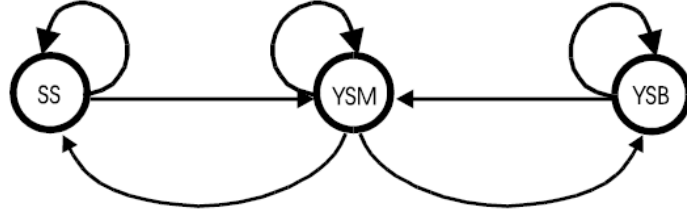
<i>Hece Yapısı</i>	<i>Örnek</i>
<i>1 Sesli harf</i>	A -rıtım
<i>1 Sesli 1 sessiz harf</i>	Ar -tırım
<i>1 Sesli 2 sessiz harf</i>	Art -mak
<i>1 Sessiz 1 sesli harf</i>	Ba -yılmak
<i>1 Sessiz 1 sesli 1 sessiz harf</i>	Bay -rak
<i>1 Sessiz 1 sesli 2 sessiz harf</i>	Kork -mak

Kelime ve hecelerin yapısıyla ilgili temel kurallar ise aşağıdaki gibidir (Çınar, 2003; Büyükkuşçu ve Adalı, 2006):

- Hece sesli harf ile başlıyorsa kelimenin başında gelmek zorundadır.
- İki sessiz harfin yan yana geldiği hecelerde ise iki harf farklı olmak zorundadır.
- İki sessiz harf kelime başında bulunamaz.

2.5.1.3. Sessiz Uyumu

Türkçede sert sessizlerden (SS: ç, f, h, k, p, s, ş, t) hemen sonra sert karşılığı bulunan yumuşak sessizler (YSB: b, c, d, g, ğ, j, v, z) gelmez. Aynı şekilde YSB' den sonra da SS gelmez.



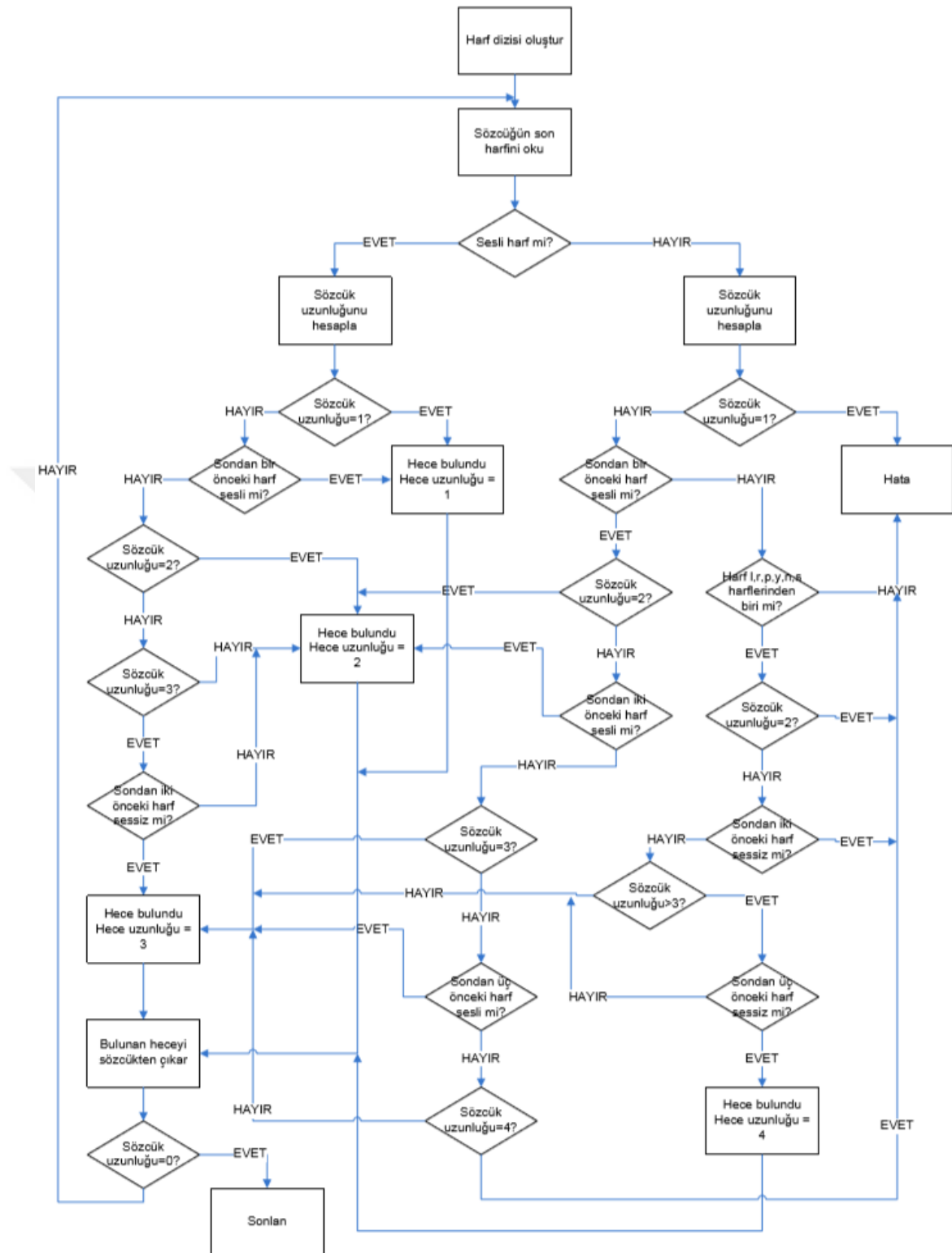
Şekil 2.12. Sessiz Uyumu Durum Makinesi (Çınar, 2003)

Sert karşılığı bulunmayan yumuşak sessizler (YSM) ise l, m, n, r ve y harfleridir. Ayrıca Türkçe kelimeler b, c, d, g harfleriyle bitmez.

2.5.1.4. Kelimenin Hecelerine Ayrılması

Türkçe heceler 6 farklı hece yapısında ve kurallı olduğu için heceleme işlemi için bir algoritma tasarlanabilmektedir. Hece içinde bir sesli harf bulunmaktadır ve hece en fazla 4 harften oluşmaktadır. Bir kelimenin hecelerine ayrılabilmesi şu açıdan önemlidir: hecelenemiyorsa bir kelime, Türkçe değildir.

Çınar (2003) heceleme algoritmasını şöyle formülize etmiş: “Sözcük ilk sesli bulunana kadar sondan başa doğru taranır. Bu sesliden önceki sessiz de alınarak ilk hece bulunur. Ardından yeni bir sesli bulunana kadar başa doğru ilerlenir ve yeni hece bir önceki adımdaki gibi bulunur. Bu işlem sözcük sonlanana kadar sürdürülür.” (Çınar, 2003 s.16). Kelime uzunluğu, hece uzunluğu ve hece içinde kullanılan harfler de dikkate alınarak Delibaş (2008) tarafından hazırlanan akış diyagramı ise şekil 2.12’ de gösterilmektedir.



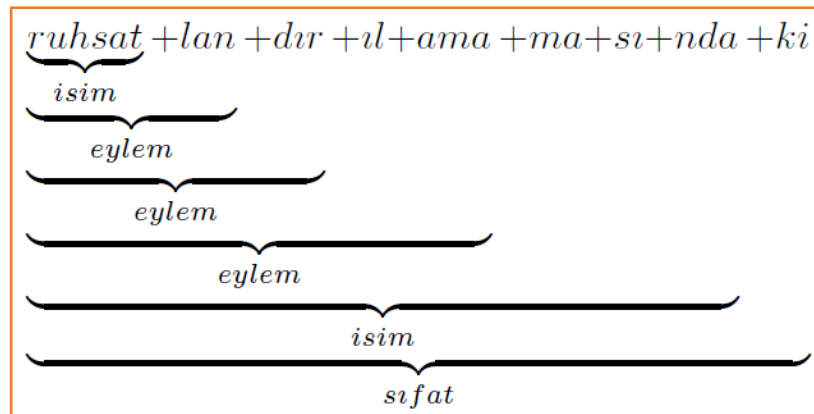
Şekil 2.13. Heceleme Algoritması Akış Diyagramı (Delibaş, 2008)

2.5.1.5. Sondan Eklemeli Dil Yapısı

Türkçe sondan eklemeli dil ailesinde bulunmaktadır. Türkçe kelime köklerine gelen ekler ile yüzlerce farklı kelime türetilabilmektedir. Türemiş kelimeler geriye doğru, kök ve eklerine ayrıştırılırken de farklı ekler alarak aynı şekle giren kelimelerde zorluk yaşanmaktadır. “Okuma” kelimesi örneğin iki farklı anlamda kullanılmaktadır. Fiil kipinde iken “oku-ma” şeklinde eklerine ayrılan kelime, isim halinde “ok-um-a” şeklinde eklerine ayrılabilir. Bu bağlamda kelime türünün tespit edilmesi önemlidir.

“Yapım ekleri, Türkçenin vazgeçilmez unsurlarından biridir. Yapım ekleri, kelime kök/gövdelerine gelerek dilin yeteneklerini genişleten, geliştiren ve zenginleştiren, yeni kelimeler üretilmesine olanak sağlayan eklerdir. Yapım eki almış kelimenin mutlaka bir önceki kök/gövdesi ile anlamsal bağı bulunması ve yeni bir kelime türetilmiş olması gerekir. Örneğin <sandık> kelimesi <san(mak)+dık> şeklinde düşünülemez. Çünkü kök olan <san(mak)> ile gövde olan <sandık> arasında hiçbir anlamsal yakınlık yoktur. Fakat <sanık> kelimesinde <sanmak> ile <sanık> arasında bağ kurulabildiği için türemiş kelime olarak kabul edilir.” (Karaca ve Görgünoğlu, 2016 s.20)

Yapım ekleri, isimden isim, isimden fiil, fiilden isim ve fiilden fiil yapım ekleri olmak üzere 4 kategoriye ayrılmaktadır. Bir kelimeye birden fazla yapım eki gelebilmektedir. “ruhsatlandırılammamasındaki” kelimesinde örneğin, 5 adet yapım eki bulunmaktadır.



Şekil 2.14. “Ruhsatlandırılammamasındaki” Kelimesindeki Ekler (Ofлаzer ve Bozşahin, 1994)

Kelimenin anlamında herhangi bir değişikliğe sebep olmayıp, cümle içinde kelimeleri birbirine bağlayarak anlamı tamamlayan eklere ise çekim ekleri denir. İsim ve

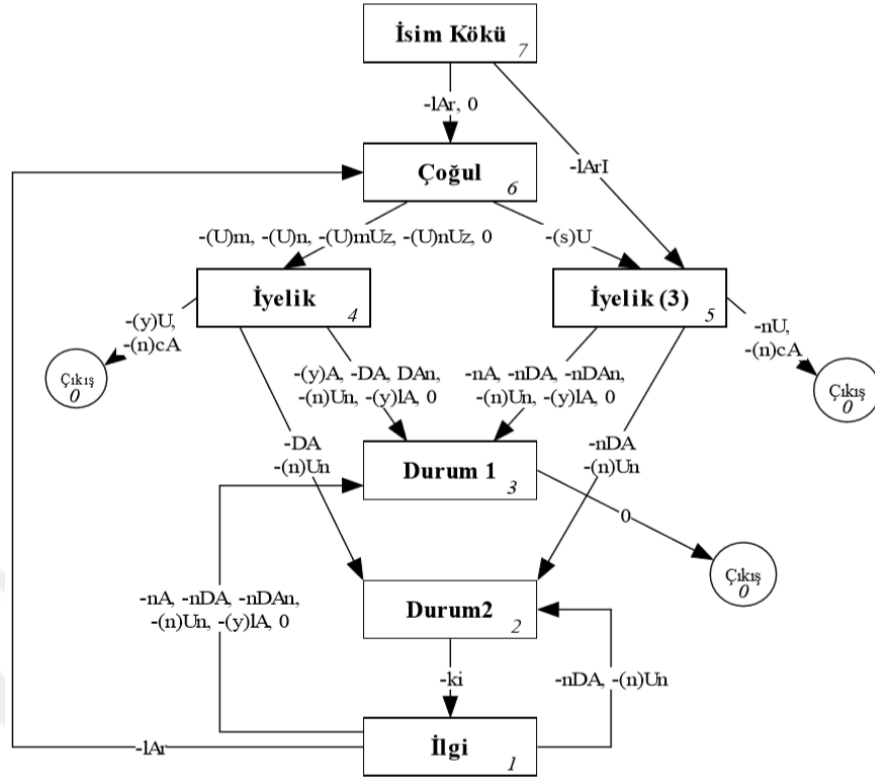
fiil çekim ekleri olmak üzere iki çeşidi bulunmaktadır. Yapım ekleri ve çekim ekleri ile bir kelimedenden yüzlerce farklı kelime türetilmektedir. Örneğin “Oku” kelimesine eklenen sadece 3 yapım ekiyle ve çekim ekleriyle 1461211 adet kelime türetilmektedir. Bu kelimelerin hepsi kullanılsa da Türkçenin biçimbilim üretim gücünü temsil etmesi bakımından önemlidir (Oflazer ve Bozşahin, 1994).

Yapım ekinden sonra yapım eki veya çekim eki gelebilirken; çekim eklerinden sonra sadece çekim eki gelebilmektedir. Eklerde sesli ve sessiz uyumu kuralları ve ses olayları ise çoğunlukla geçerlidir.

Kelime köküne gelen eklerde kullanılan harflerle ilgili de bazı kurallar geçerlidir (Dilsiz, 2005; Cebiroğlu, 2002):

- İki sesli harf yan yana gelmez
- “b”, “v”, “f”, “h” ve “j” harfleri eklerde bulunmaz
- “p” ve “z” harfleri ek başında bulunmaz
- Ek başında “d”, “t”, “c”, “ç”, “g”, “k”, “s”, “m”, “l” ve seyrek olarak r harfi (geniş zaman eki) gelir.

Yapım eklerinin ardışık kullanımıyla ilgili kurallar kesinlik göstermediği için sonlu durum makinesi çizilememektedir. Çekim eklerinin ise her bir türü için ayrı bir durum makinesi tasarlanması gerekmektedir. Şekil 2.13’te isim çekim eki için soldan sağa durum makinesi çizim örneği bulunmaktadır.



Şekil 2.15. Soldan İsim Çekim Eki Sonlu Durum Makinesi (Cebiroğlu, 2002)

2.5.1.6. Cümle Yapısı ve Anlam Vurgusu

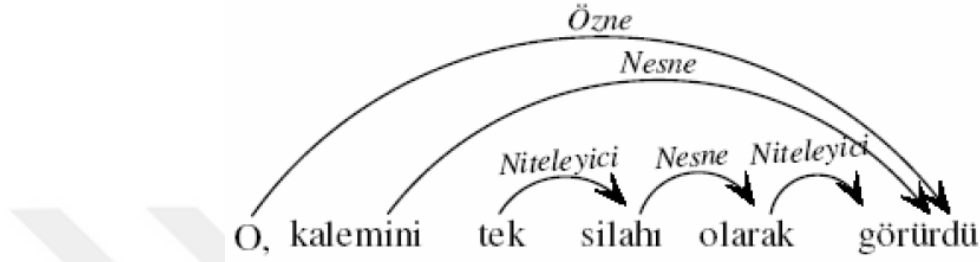
Türkçede cümle yapısı Özne + Nesne + Yüklem şeklinde olmasına rağmen, devrik yapıda cümle kurulumları da sıklıkla karşımıza çıkmaktadır. Üç kelimedenden müteşekkil bir cümlede bile 6 farklı dizilim bulunmaktadır ve altı farklı anlam türetilabilmektedir.

Tablo 2.6. Cümle Öğelerinin Dizilişinin Anlam Üzerinde Etkisi (Of lazer ve Bozşahin, 1994)

Sıra	Cümle	Anlam
1	Ekin, Ayşe'yi gördü	
2	Ayşe'yi Ekin gördü	Gören Ekin'di başkası değil.
3	Gördü Ekin Ayşe'yi	Ama görmemesi gerekiyordu.
4	Gördü Ayşe'yi Ekin	Zaten görmesini bekliyordum.
5	Ekin gördü Ayşe'yi	Başkası da görebilirdi.
6	Ayşe'yi gördü Ekin	Başkasını da görebilirdi.

Cümle öğelerinin dizilişi vurguyu değiştirdiği gibi anlamı da ciddi ölçüde zenginleştirmektedir. Tablo 2.6’da 3. sıradaki cümle dizilişinde anlam olumsuz iken, 4. sıradaki cümlede anlam olumlu olabilmektedir.

Peki, kelimelerin sırası önemli midir? Evet, çünkü cümle içindeki kelimelerin arasında da bağıllık ilişkileri bulunmaktadır. Dolayısıyla kelimelerin diziliş sırası da anlama katkı sağlamaktadır. Şekil 2.13’te bağıllık ilişkilerine bir örnek verilmektedir.



Şekil 2.16. Bağıllık İlişkileri (Eryiğit vd., 2006)

2.5.2. Yazım Yanlışlarının Düzeltilmesi

Metnin kelimelere ayrıştırılması (*tokenize*), durdurma kelimelerinin (*stop words*) çıkarılması ve kelimelerin köklerine inilmesi bu aşamada yapılmaktadır (Hoche vd., 2006).

İnternet sayfaları veya resim linkleri, @ işareti ile başlayan kullanıcı adları ve tarih ifadeleri analize katkı sağlamadıkları için mesaj içeriğinden çıkarılabilmektedir. Bu parazitler, gereksiz şişkinliğe sebep olmaktadır ve analiz sürecini yavaşlatıcı etkiye sahiptir. Ancak duygu sembolleri (☺ ☹ vs.) yazarın duygusunu daha iyi betimledikleri için analize dâhil edilmelidirler (Akarsu ve Diri, 2016).

Türkçe kelimelerin tutulduğu sözlükte yer almayan bir kelimenin hatalı kabul edilip edilmemesi önemli bir sorundur. Zira bir kelime sözlükte bulunmasa da sosyal medya jargonunda kullanılabilmektedir. Kelimenin hatalı olup olmadığını anlamının bir yolu da kelimenin diğer tweetlerdeki kullanım sıklığına bakılmasıdır.

Bir kelimenin birden fazla tweet içinde aynı yanlış kullanımının bulunması durumunda ise bu tweetler birbirini maskeleyebileceklerdir. Dolayısıyla bu kelime jargon zannedilecektir.

“...Küreselleşmenin en etkili araçlarından biri internetle birlikte gelen sosyal medya ağlarıdır. Bu ağlar sayesinde kişiler farklı dillerde konuşan, farklı ülkelerde yaşayan insanlarla etkileşim içine girmiş; bu etkileşimden dil de nasibini almıştır. Çeşitli şekillerde değişime uğrayan dil (kısaltmalar, yeni kelimeler, yeni harfler) sadece Türkiye’de değil her ülkede küreselleşmeden etkilenmiştir... Ancak küresel piyasadaki ürünlerin kullanımını kısıtlamak ne kadar zorsa dile gelen bu yeniliklerin engellenmesi de bir o kadar zordur. Çünkü dilin yoğun olarak kullanıldığı internet ortamı 24 saat boyunca sürekli bir akış halindedir, küreselleşme gibi.” (Türksever, 2013 s.8)

Kelimenin, kendisine en yakın kelime ile değiştirilebilmesi için mesafe veya benzerlik ölçüleri kullanılabilir. Örneğin uzaklık değerleri ölçülecek olursa fonksiyon değerinin 0’a yakın olması iki kelimenin benzerliğini artırmaktadır.

$$d(K_1, K_2)$$

Bazı kelimeler ise eksik veya fazla harf kullanımı dolayısıyla bütünüyle başka bir forma bürünmektedir. Yine Türkçe karakter kullanılmadan yazılan kelimeler de diğer tweetlerde geçen kelimeler ile önemli düzeyde farklı olduğu tespit edilebilecektir. Bu hataların tesadüfen oluştuğu ve düzeltilmesinin zaruri olduğu aşikârdır. Öncelikle yanlış yazılan verilerin tesadüfîlik düzeyi tespit edilerek sıklıkla yapılan yanlışların hangi durumlarda yoğunlaştığı tespit edilmelidir.

Sözlük baz alınarak düzeltilemeyen, kelimelerin, birlikte kullanıldığı kelimelerle birlikte arattırılması doğruluğu artıracaktır. Bunun için de kelime çiftlerinin tweet havuzu içinde arattırılması bir yöntem olarak tercih edilebilir. Örneğin “rakit” şeklinde yazılan bir kelime “rakip” ve “raket” kelimelerine eşit uzaklıktadır. Bu kelimenin düzeltilebilmesi için kendinden sonra gelen “olmak” kelimesi, “rakip” kelimesiyle ikili oluşturacağı için önerinin isabet oranı artırılmış olacaktır.

İlgili dönüşümlerin başarısız olduğu yani yüksek hata düzeyindeki kelimeler için “liste temelli silme” yöntemi uygulanabilir (Akpınar, 2014). Yani sözlükteki herhangi bir kelimeyle eşleşmeyen kelimeler silinebilir. Ancak silinen bu kelime, kaydın toplam duygu değerini değiştirebileceği için sakıncalıdır.

Türkçe sondan eklemeli dil grubunda olduğu için Türkçe kelimeler sonlarına ek olarak anlam değişikliğine gidebilmektedir ve bu durum duygu analizi çalışması için de zorluk teşkil etmektedir. Olumsuzluk ekleri (me/ma) ve varlık (lı/li) yokluk (sız/siz) ekleri

buna örnek verilebilir. Benzer şekilde olumsuzluk bildiren “değil” ve “yok” kelimeleri de dikkate alınmalıdır. Ünlem işareti ise sonuna geldiği cümlelerin veya kelimenin anlamında değişikliğe sebep olabilmektedir (iğneleyici ifadelerde olduğu gibi). "Ama", "fakat" ve "lakin" ifadeleri de cümlelerin gidişatında anlam değişikliğine yol açması hasebiyle ayrıca üzerinde durulmalıdır (Kavasoğlu, 2013). Duygu simgeleri ise metnin duygu durumunun anlaşılmasında belirleyici rol oynamaktadır. Olumsuzluk ekleri veya kelimeleri bulunduğu ise ilgili kelimenin duygu puanı eksi ile çarpılabilir.(Türkmenoğlu, 2015)

Birden fazla kelimenin birlikte kullanıldığı kelime ikililerinin de ortaya çıkarılması gerekmektedir. Kelime ikililerinde kelimeler kendi başlarına bir anlam ve duygu içermedikleri halde birlikte kullanıldıklarında anlamlı olmaktadırlar ("Hayal Kırıklığı" örneğindeki gibi).

Bir harfin gereksiz kullanımı kelimenin yanlış yazılmasına sebep olmaktadır. “Meerhabaaa” kelimesinde olduğu gibi. Bu kelimedeki kullanılan her bir harfi teke indirmek sorunu çözecektir fakat “faaliyet”, “saat”, “kıraat”, “cemaat”, “maaş”, “inşaat” gibi Türkçemize diğer dillerden giren kelimelerde peşpeşe aynı harf bulunabilmektedir. Sözlükte karşılığı bulunmayan kelimelere işlem yapmak bir çözüm olabilirse de tweetlerde sıklıkla karşımıza çıkan “notebook”, “facebook” ve “google” gibi yazım diline sonradan giren yabancı kelimeler için sözlüğün düzenli olarak güncellenmesi gerekmektedir.

Konuşma dilindeki uzatmaların yazıya geçirilmesi de anlamı dönüştürmektedir. “eee hani nerde?” ve “eee napalım?” gibi ifadelerin de kendi başına anlamı bulunmadığı halde cümlelerin anlamını tamamlayıcı bir etkisi bulunmaktadır.

Özetle, yazım hatalarının giderilmesi bilgisayarlar tarafından otomatik olarak yapılamamaktadır. Bu durum insan müdahalesini zorunlu kılmıştır (Çınar, 2003).

Kelimelerin doğruluğu veri tabanında kayıtlı sözlükteki kelimelerle eşleştirilerek tespit edilebileceği belirtildi. Yanlış eşleşmelerin önüne geçebilmek için de kelimeler, veri tabanında en uzun kelimedenden kısaya doğru gidilerek eşleştirilmelidir.

Kelimelerde harf fazlalığı, eksikliği veya ardışık gelen iki harfin yanlış sırada kullanımı şeklinde beliren yazım yanlışlığı iki kelime arasına boşluk bırakmama şeklinde de ortaya çıkabilmektedir. Klavye kullanımından kaynaklanan hataların tespitinde de

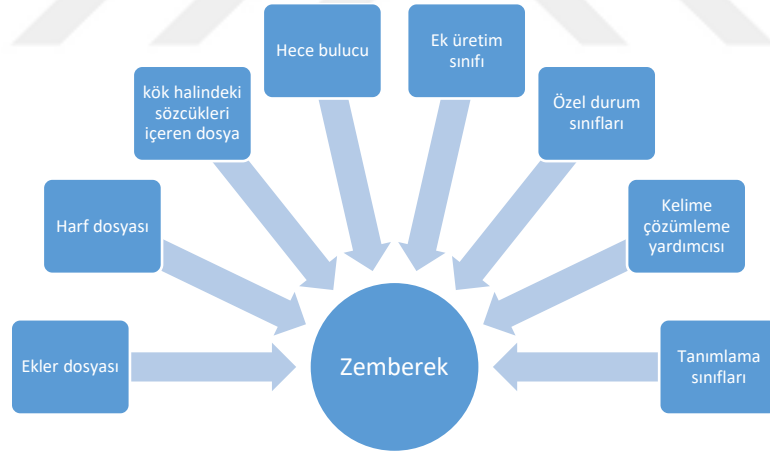
Türkçe karakterden kaçış ASCII benzerleriyle değiştirilmesi ile düzeltilebilecektir. Türkçe dil kurallarına uymayan hataların tespit edilmesi de nispeten daha kolaydır.

Herhangi bir kelimenin veya bütün cümlenin büyük harf ile yazılması çoğunlukla yüksek sesle konuşmak manasında olumsuzluk içermektedir. Böyle yazılan tweetleri olumsuz olarak etiketlemek düşünülebilirse de kısaltmalar ve caps lock (büyük harf kilidi)’un açık unutulması ihtimali dikkatten uzak tutulmamalıdır.

2.5.2.1. Zemberek Kütüphanesi

Türkçe metinler için doğal dil işleme üzerine çok sayıda teorik çalışma bulunmasına rağmen uygulamaya dönük pek çalışma bulunmamaktadır. Zemberek kütüphanesi bu bağlamda 2004 yılında açık kaynak kodlu doğal dil işleme kütüphanesi olarak Java dilinde geliştirilmiştir. OpenOffice, libreOffice ve Firefox eklentisi geliştirilmiştir. Pardus için yazım denetim desteği de bulunan zemberek için NET projesinin (NZemberek) geliştirilmesine ise 2006 yılında başlanmıştır (Zemberek, 2006).

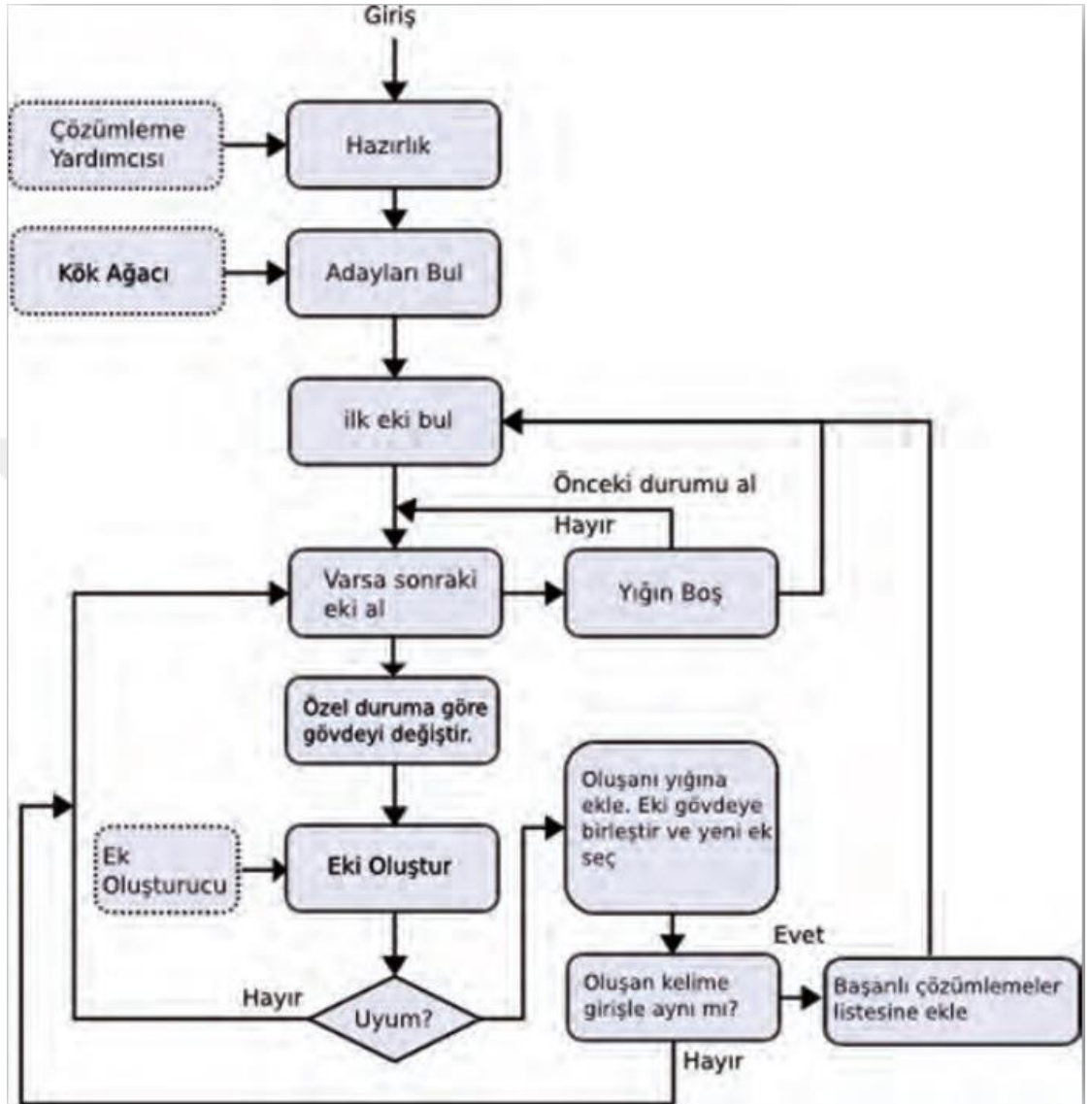
Zemberekte her bir dil için sözlük dosyaları ve sınıflar oluşturulmaktadır.



Şekil 2.17. Her Bir Dil için Zemberek Dosya ve Sınıfları (Akın ve Akın, 2007)

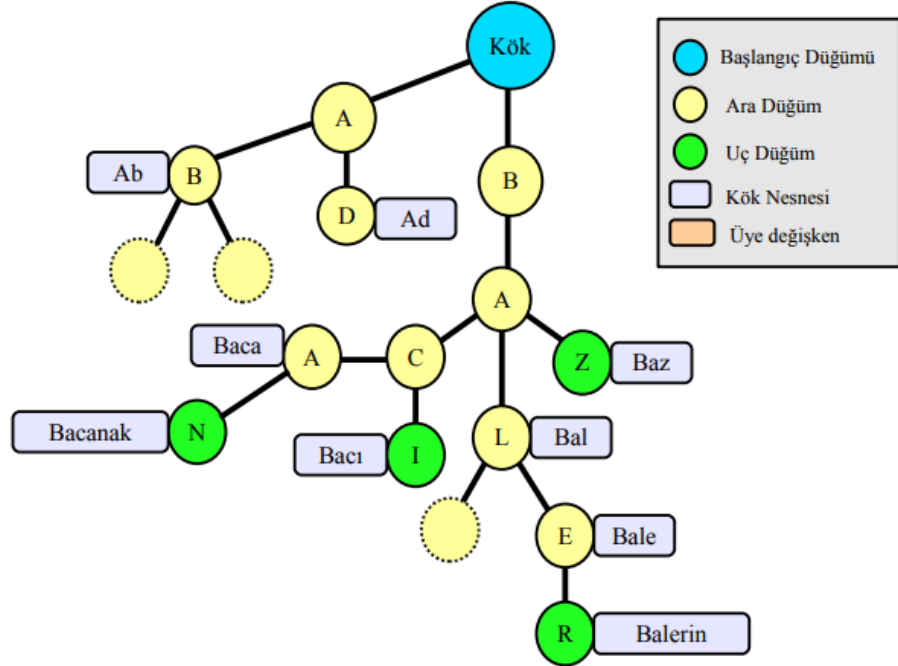
Zemberek, her bir kelime için morfolojik ayrıştırıcı algoritmasını devreye sokarak kelime kökünü ve eklerini bulmaktadır. Kelime öncelikle ön işleme (preprocessing) alınır. Noktalama işaretleri ve diğer özel işaretler kaldırıldıktan sonra kelime, küçük harfe çevrilir. Sonrasında kök ağacı içinden olası aday bulunur. Kelime olası eklerle yeniden oluşturulmaya çalışılır. Bu işlem uyum sağlanıncaya kadar olası her bir ek için sürdürülür.

Böylece kelime kök ve gövdesi tespit edilmiş olur. Bu işlemin algoritması, aşağıdaki şekildeki akış diyagramı ile gösterilmiştir.



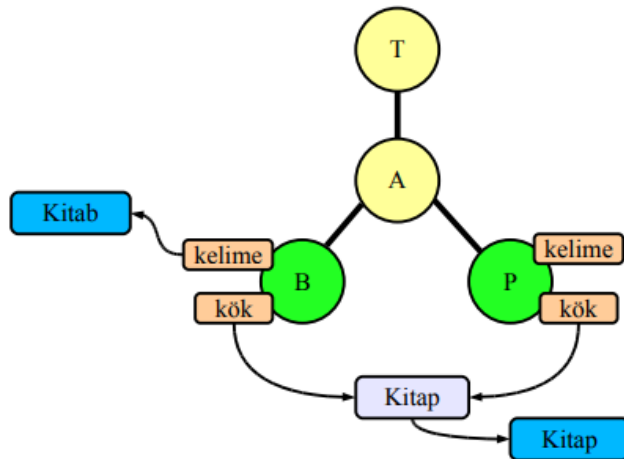
Şekil 2.18. Morfolojik Ayrıştırıcı (Akın ve Akın, 2007)

Zemberek'te her bir kelime kökü, kök ağacına eklenir. Kök ağacı yapısı ise aşağıdaki gibidir:



Şekil 2.19. Zemberek Kök Ağacı Yapısı (Akın, 2007)

Bu ağaç işletildiğinde “bacan” ifadesi “bacanak” kelime kökünü buldurmaktadır. Aynı şekilde “baler” ifadesi de “balerin” kelimesine ulaştırmaktadır. Çünkü bu ifadelerin devamında iki farklı kelime bulunmamaktadır. “Kitap” kelimesinin ek almış hali “kitabı”, “kitaba” şekillerinde tezahür ettiğinden, kelime köküne inildiğinde, “kitab” ifadesinin kitap kelimesine ulaştırması için iki farklı kelimenin uç düğümü aynı kelime kökünü, yani “kitab” kelimesini göstermelidir.



Şekil 2.20. Kök Ağacı Yapısı, Kitap Kelimesi Örneği (Akın, 2007)

Zemberek, Türkçede yaygın kullanımı bulunmayan ön ekler için ayrıştırma yapmamaktadır. Bazı yapım ekleri için de ayrıştırma yapmaz ve sözlükte yapım ekli halini bulundurur. Zemberek, hata toleranslı kök seçicisi için “Damerau-Levenshtein” hata mesafe algoritmasını kullanır. Zemberek geliştiricileri, yanlışlıkla bitişik yazılan kelimeleri ayrıştırabildiklerini ve Türkçe karakter kullanılmadan yazılan kelimeleri de ASCII toleranslı olarak düzeltebildiklerini söylemektedirler (Akin ve Akin, 2007).



ÜÇÜNCÜ BÖLÜM

UYGULAMA GELİŞTİRME

3.1. ÇALIŞMANIN ÖNEMİ

Sosyal medya analizleri alternatif araştırma yöntemlerine kıyasla, daha geniş bir örnekleme sahiptir ve sosyal medya analizi ile çok katmanlı davranış bilgisine, düşük maliyetle ve kısa zamanda ulaşmak mümkün olmaktadır. Ayrıca Twitter, gerçek davranışsal veri içermesi yönüyle, anket çalışmalarından daha güvenilir sebep sonuç ilişkileri ortaya konulmasına olanak vermektedir (Çevik, Öztürkcan ve Kasap, 2015).

Büyük veri teknolojisi kullanılarak Twitter verilerinin analiz edilmesinde Türkçe tweetlerin kullanılması açısından bu sahada yapılacak çalışmalara da katkı sağlayacağı ön görülmektedir. Bu çalışma, ilgili sektördeki firmaların karşılaştırmalı analizinin Türkçe Twitter verilerinin duygu analizi sonuçlarının karşılaştırılabildiği bir çalışması olması bakımından da işletme yönetiminin karar süreçlerine dâhil edilebilecektir.

Sosyal medya verileri akış halinde olduğu için analiz sonuçlarının raporlanması da süreklilik arz etmektedir. İnternet ve sosyal medya kullanımının artmasıyla zorlaşan veri analizi ise büyük veri tekniklerinin kullanımını zorunlu kılmaktadır. Bu çalışma, büyük veri teknolojisine yatırım yapmaktan çekinen küçük ve orta çaplı işletmeler için bir geçiş denemesi niteliği taşımaktadır. Bu çalışma ayrıca duygu analizi sonuçlarının zamansal ve mekânsal sonuçlarının karşılaştırmalı olarak paylaşabilmesi açısından da bir yenilik katmaktadır.

İşletme kendisi ve rakipleri hakkında paylaşılan tweetlerin dönemlik olarak aldığı sonuçları şehirler bazında da görüntüleyebilmektedir. Geliştirilen sistem sayesinde işletme üst yönetiminin izlediği rekabet stratejilerinde yeni adımlar (maliyet liderliği, farklılaştırma ve odaklama gibi) atabilmesine katkı sunmaktadır.

3.2. KAPSAM

Bu çalışma, duygu analizi ve metin madenciliği çalışmasıdır. Bu çalışma ile sosyal medya verileri içinde Twitter verilerinin çekilmesine odaklanıldı. Twitter Api'lerinin kısıtları ve yeterlilikleri ölçüsünde veri çekilebildi. Benzer bir biçimde Facebook

verilerinin çekilmesi için de genişletilebilir. Tweet metni içinde geçen adreslerin temsil ettiği resimler gibi diğer içerikler de bu çalışmanın kapsamı içine alınmadı. Tweet dili olarak da Türkçe seçildi.

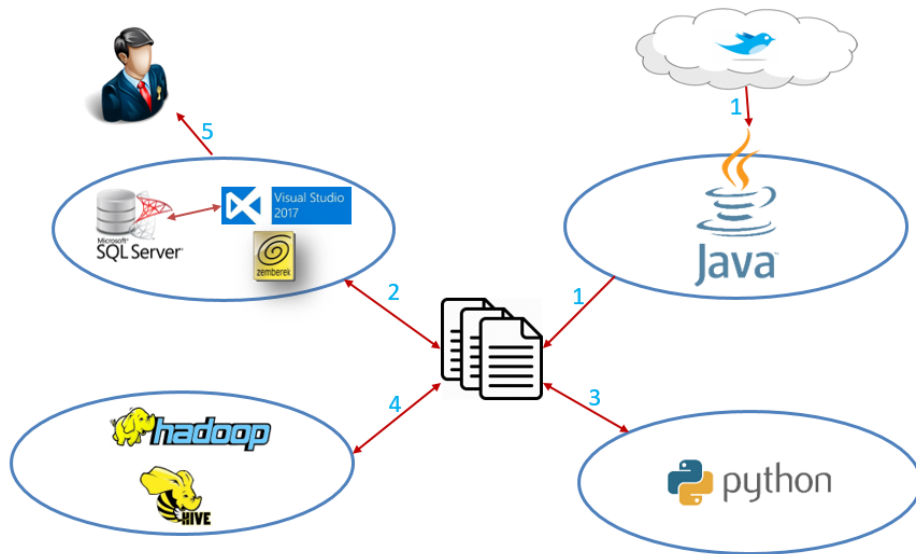
Doğal dil işleme modülleri de İstanbul Türkçesi üzerine tasarlandı. Diğer lehçe ve ağızlar için destek vermemektedir. Yanlış yazılan kelimelerin düzeltilebilmesi için tasarlanan sözlük tablosuna, kelime köklerine gelebilecek eklerle oluşturulabilen kelimelerin tümü dâhil edilmedi. Literatürdeki diğer çalışmalarda da kullanılan sözlük tabloları derlenerek bir sözlük tablosu oluşturuldu.

Tweet atan kullanıcılar ile ilgili, Twitter tarafından paylaşılan kullanıcı Id, konum ve takipçi sayısı gibi bilgiler dışındaki bilgilere erişilemedi. Dolayısıyla yaş, kültür ve eğitim düzeyleri gibi demografik bilgiler üzerinden istatistiksel analizler yapılamadı.

Test verisi için Türkiye’de faaliyet gösteren telefon markaları hakkında public (halka açık) paylaşılan tweetler çekildi. Çalışma, resmi hesaplar için atılan tweetler ile sınırlı tutulmadı. Duygu analizi için sözlük tabanlı yöntem tercih edildi. Ancak söz dizimsel analiz yapılmadığı için birden fazla marka adının geçtiği tweetlerde karşılaştırmalı analiz yapılmadı.

3.3. MODEL

Önerilen Model



Şekil 3.1. Önerilen Model

Önerilen modele göre ilk olarak tweetler, geliştirilen Java uygulaması ile çekilmektedir. Visual Studio.NET uygulaması ile ön işleme tabi tutulan tweetler, Zemberek kütüphanesi API'leri ile köklerine ayrıldıktan sonra geliştirilen bir modül ile SQL Server veri tabanına kaydedilmektedir. Köklerine ayırma işleminden maksat, öznitelik olarak kelimelerin seçildiği analizde, sonuna ekler almış kelime türevleriyle fazla büyümesinin ve analiz başarısını düşürmesinin önüne geçmektir.

Yanlış yazılan kelimelerin düzeltilebilmesi için Python uygulaması ile doğal dil işleme sürecine tabi tutulan veriler, Hadoop ortamında Hive betikleri ile analiz edilmektedir.

Analiz işlemi de bittikten sonra sonuçlar geliştirilen Visual Studio.NET uygulaması ile kullanıcıya raporlanmaktadır. Bu modül, modelin işletme yönünü yansıtmaktadır. Sistemin son kullanıcılarının şirketin üst düzey yöneticileri olması öngörülmektedir.

3.4. MODÜLLER

Çalışma kapsamında tweetlerin çekilmesi modülü, ön işleme modülü, sınıflandırma modülü ve raporlama modülü olmak üzere toplamda dört adet modül geliştirilmiştir.

Tweetlerin çekilmesi modülü için Twitter API'leri (Twitter4J kütüphanesi) kullanan bir java uygulaması geliştirilerek kullanıcı tarafından girilen kelimelerin geçtiği tweetler çekilerek ilgili klasörlere “txt” formatında kaydedilmektedir. Bu modülün periyodik olarak çalıştırılması ile son atılan tweetlerin çekilmesi sağlanabilmektedir.

Elde edilen tweetlerin ön işleme tabi tutulması aşamasında, bazı çalışmalarda Zemberek kütüphanesinden faydalanılırken (Boynukalın, 2012; Çakmak, 2012); Kayahan ve arkadaşları (2013) ise, Zemberek kütüphanesinin eksik harf düzeltme doğruluğu düşük ve sunucu bağlantı hızı ve işlem hızı düşüklüğü gerekçesiyle yanlış yazılan kelimelerin düzeltilmesi için “edit distance” yöntemini tercih etmiştir. Nalçakan ve arkadaşları (2014) ise, ön işlem aşamasında kelime düzeltici programı, Python dilinde hazırlamıştır.

Meral ve Diri (2014) yaptıkları çalışmada Twitter erişim modülü (Twitter 4j kütüphanesi ile), doğal dil işleme modülü, veritabanı (Solr) yönetim modülü ve

sınıflandırma modüllerinden (makine öğrenmesi) oluşan bir sistem geliştirmişlerdir. Kelime tabanlı yöntem ile %89,5 başarı elde etmişlerdir (Meral ve Diri, 2014).

Köse (2015), havayolu şirketleri hakkında paylaşılan İngilizce tweetler (55000 adet) üzerine yaptığı duygu analizi çalışmasında Hadoop mimarisi üzerinde Hive betikleri çalıştırmıştır (Köse, 2015). Bunun için Oracle VirtualBox üzerinde Hortonworks Sandbox 2.0 ortamından faydalanmıştır.

Twitter'in daha ne kadar süre kullanılacağı bilinmemekle birlikte, işletmelerin müşterileriyle iletişim kanalı olarak sosyal medyayı kullanmaya devam edeceği kuvvetle muhtemeldir. Sosyal medya kullanımı arttığı sürece, işletmeler hakkında paylaşılan yorum sayısının zamanla artacağı ve büyüyeceği kabulü ile büyük veri teknolojisinin kullanımının gelecek çalışmalara da ışık tutacağı ön görülmektedir. Ayrıca tüm kelimelerin özellik olarak kullanılması ve ngram yöntemi dolayısıyla işleme dâhil edilen veri boyutunun artması sınıflandırma için paralel makinelerin kullanılmasını yakın bir gelecekte zorunlu kılacaktır. Bu bağlamda, çalışma kapsamında, Linux Ubuntu işletim sistemi kurulu 3 makinelik bir Hadoop kümesi oluşturulması ve bu küme üzerinde HiveQL betiklerinin koşturulması planlandı. Eşle İndirge fonksiyonlarının icra edilmesi için üst seviye betik (script) dillerinden biri olan Hive dilinin tercih edilmesinin sebebi ise SQL diline yakınlığı ve kullanım kolaylığıdır.

Bu çalışmanın ön işlem aşamasında Python dili kullanarak hazırlanan betikler ile analizde kullanılmayan tweetler çıkarılıp, tweetler gereksiz ifadelerden (linkler, harf olmayan karakterler vs.) temizlenerek eksik/hatalı yazımlar düzeltilebilmektedir. Analiz aşamasında ise sınıflandırma algoritması için kelime tabanlı denetimli makine öğrenmesi tekniği kullanıldı.

SQL Server veritabanının kurulu olduğu ve TSQL kodlarının çalıştırıldığı makinenin özellikleri aşağıdaki gibidir:

- CPU (İşlemci): Intel Core i7 2,20 GHz
- Ram (Bellek): 16 GB
- Sistem: 64 bit Windows 10 işletim sistemi

Sanal makinenin konfigürasyon özellikleri ise:

- CPU (İşlemci): 2 çekirdekli

- Ram (Bellek): 2 GB
- Sistem: 64 bit Red Hat
- Görüntü Belleği: 12 MB

3.4.1. Tweetlerin Çekilmesi

3.4.1.1. Twitter Api Şifresi Alma

Twitter tarafından uygulama geliştiricilere açılan servislere erişim için kullanıcı kaydı oluşturulması gerekmektedir. İnternet tarayıcısının adres kısmına “<https://apps.twitter.com/>” değeri girildiğinde açılan sayfada,



Şekil 3.2. Twitter Uygulama Yönetim Sayfası

“Create New App” düğmesine tıklandıktan sonra gelen ekranda ilgili bilgiler doldurulduktan sonra en altta bulunan “Create your Twitter application” düğmesine tıklandığında Twitter Api’si için gerekli erişim şifreleri oluşturulmuştur. Oluşturulan erişim şifreleri aşağıdaki gibidir:

```
consumerKeyStr      = "Nnmrk39Cpr0uH4JSO1tNI41";
accessTokenStr      = "3331570701-pWnh4IddhJeRDABxrOnL8MmqVF3QzbjHPE9v4Hd";
accessTokenSecretStr = "fRxsxjvqbdXQolfE3168WqUpKPBPQhro8VwEFNLYgGjSr";
consumerSecretStr   = "m4thJXg5FvazTkKXiFyOdaw5ZiLwfDoENk7QtVWvj7rgpBNqX";
```

Şekil 3.3. Erişim Şifreleri

Bu şifreler aracılığıyla, Twitter’in kullanıcılara açtığı API’ler (Application Programming Interface – Uygulama Programlama Arayüzü) kullanılarak periyodik olarak (haftanın ilk günü) Twitter verileri toplanabilmektedir.

Bu aşamada hedef ile doğrudan bağlantılı olmayan tweetlerin çekilmesi önlenerek belleğin şişmesine mani olunabilmektedir.

3.4.1.2. Tweet Çekme Uygulaması

Bu uygulama sayesinde, işletme son 10 gün içinde herkese açık olarak paylaşılan tweetler üzerinde tarama yaparak ilgili kelimelerinin geçtiği tweetlerin çekilmesini sağlamaktadır. Şekil 3.4’da gösterilen ekran üzerinde arama sözcükleri eklenip, çıkarılabilmekte ve son düzenleme diske kaydedilebilmektedir. Böylece kurum, sektörde kendisini kıyaslayabileceği markaları belirleyebilmektedir.

İşletme bu uygulama sayesinde kendisi ve rakipleri hakkında paylaşılan tweetlere erişim sağlayabilecektir.

Tweet Çekme Uygulaması

Tweet Arama Parametreleri

Aranacak Metin:

Şu ID'den İtibaren:

Tweet Aramıyor...

ID:894908610886733824/FollowersCount:236/ListedCount:33/FavoriteCount:0/RetweetCount:17/CreatedAt:08.08.2017 16:10/Location:/User/ID:1725834739/Tweet:RT @BayBayNakit: #temassızyaşaya
 ID:89490740932166528/FollowersCount:106/ListedCount:0/FavoriteCount:0/RetweetCount:0/CreatedAt:08.08.2017 16:05/Location:Mersin, Türkiye/User/ID:739526897953067009/Tweet:#samsung gü
 ID:894907321737990144/FollowersCount:353/ListedCount:0/FavoriteCount:0/RetweetCount:0/CreatedAt:08.08.2017 16:05/Location:Ümraniye/İstanbul/Türkiye/User/ID:2336236598/Tweet:İlk de Anta
 ID:89490608790366208/FollowersCount:84/ListedCount:14/FavoriteCount:0/RetweetCount:0/CreatedAt:08.08.2017 16:00/Location:/User/ID:78723742877397193/Tweet:Samsung Galaxy S8 Active Ö
 ID:894904912433344512/FollowersCount:1209/ListedCount:10/FavoriteCount:0/RetweetCount:0/CreatedAt:08.08.2017 15:55/Location:Gaziantep/User/ID:1047673532/Tweet:Samsung telefon pilinin b
 ID:894904758066192384/FollowersCount:291/ListedCount:8/FavoriteCount:0/RetweetCount:17/CreatedAt:08.08.2017 15:54/Location:/User/ID:1269939001/Tweet:RT @BayBayNakit: #temassızyaşayan
 ID:894904755230781440/FollowersCount:856/ListedCount:1/FavoriteCount:0/RetweetCount:0/CreatedAt:08.08.2017 15:54/Location:samsun değil/User/ID:882549758/Tweet:@tugcecelikbass Samsung r
 ID:894904113154187265/FollowersCount:1771/ListedCount:260/FavoriteCount:0/RetweetCount:17/CreatedAt:08.08.2017 15:52/Location:/User/ID:1445327477/Tweet:RT @BayBayNakit: #temassızyaşay
 ID:894902804631896064/FollowersCount:1057/ListedCount:60/FavoriteCount:0/RetweetCount:0/CreatedAt:08.08.2017 15:47/Location:İstanbul, Türkiye/User/ID:37891246/Tweet:Samsun'un bapı dert
 ID:894902685891035136/FollowersCount:2/ListedCount:0/FavoriteCount:0/RetweetCount:0/CreatedAt:08.08.2017 15:46/Location:/User/ID:845947183295803394/Tweet:@TansuYegen Adamlar neler ya
 ID:894902573160742913/FollowersCount:16/ListedCount:0/FavoriteCount:0/RetweetCount:0/CreatedAt:08.08.2017 15:46/Location:/User/ID:73472216403828736/Tweet:@gzsdsam @gobinsevo Ya fot
 ID:894902361872793600/FollowersCount:969/ListedCount:39/FavoriteCount:0/RetweetCount:17/CreatedAt:08.08.2017 15:45/Location:İstanbul/User/ID:42674382/Tweet:RT @BayBayNakit: #temassız
 ID:89490215634431170/FollowersCount:19260/ListedCount:11/FavoriteCount:0/RetweetCount:17/CreatedAt:08.08.2017 15:44/Location:İstanbul, Türkiye/User/ID:49107263/Tweet:RT @BayBayNakit:
 ID:894901263678275584/FollowersCount:432/ListedCount:8/FavoriteCount:0/RetweetCount:17/CreatedAt:08.08.2017 15:41/Location:/User/ID:921774127/Tweet:RT @BayBayNakit: #temassızyaşayan
 ID:894900608079200256/FollowersCount:100/ListedCount:0/FavoriteCount:0/RetweetCount:0/CreatedAt:08.08.2017 15:38/Location:/User/ID:732734737/Tweet:@MunirArkan @TCSanayi @Dr_Faruk_C
 ID:894900407419502592/FollowersCount:26/ListedCount:0/FavoriteCount:0/RetweetCount:0/CreatedAt:08.08.2017 15:37/Location:Türkiye İzmir Karşıyaka/User/ID:2503807630/Tweet:Bir @YouTube v
 ID:894899572580864004/FollowersCount:817/ListedCount:4/RetweetCount:4/CreatedAt:08.08.2017 15:34/Location:Ankara la /User/ID:810641781419286528/Tweet:Samsung tost maki
 ID:894899592949153796/FollowersCount:6577/ListedCount:1/FavoriteCount:5/RetweetCount:1/CreatedAt:08.08.2017 15:32/Location:Türkiye/User/ID:3217937344/Tweet:Samsung U Flex kulaklık hosts
 ID:89489848548893249/FollowersCount:271/ListedCount:16/FavoriteCount:0/RetweetCount:0/CreatedAt:08.08.2017 15:30/Location:İstanbul, Türkiye/User/ID:52474499/Tweet:RT @BayBayNakit: #
 ID:89489830164512768/FollowersCount:84/ListedCount:0/FavoriteCount:0/RetweetCount:0/CreatedAt:08.08.2017 15:23/Location:İskenderiye/User/ID:601054448/Tweet:@SamsungDestekTR @SamsunT
 ID:894896812137558017/FollowersCount:75/ListedCount:1/FavoriteCount:0/RetweetCount:0/CreatedAt:08.08.2017 15:23/Location:İstanbul/Tweet:Samsung Galaxy S8 Active Tanıt
 ID:894896464990304257/FollowersCount:17/ListedCount:0/FavoriteCount:0/RetweetCount:0/CreatedAt:08.08.2017 15:22/Location:/User/ID:3126470031/Tweet:@SamsungTürkiye #samsungpismanlıktr
 ID:894896588111437824/FollowersCount:55/ListedCount:0/FavoriteCount:0/RetweetCount:0/CreatedAt:08.08.2017 15:22/Location:/User/ID:890504570199318528/Tweet:Samsung 9082 MAVI Silikon Kılıf
 ID:8948960944582913/FollowersCount:1621/ListedCount:7/FavoriteCount:0/RetweetCount:0/CreatedAt:08.08.2017 15:20/Location:/User/ID:144454892/Tweet:@HASANYK2763071 @asiruh266 Aha
 ID:894894980455649293/FollowersCount:1641/ListedCount:5/FavoriteCount:0/RetweetCount:0/CreatedAt:08.08.2017 15:16/Location:renklibhaber/User/ID:170981969/Tweet:4000mAh pilli Galaxy S8 Act
 ID:894894634496802816/FollowersCount:9/ListedCount:0/FavoriteCount:0/RetweetCount:0/CreatedAt:08.08.2017 15:14/Location:Sakarya, Türkiye/User/ID:796807462498267136/Tweet:Samsung aske
 ID:89489397702298099/FollowersCount:245/ListedCount:0/FavoriteCount:0/RetweetCount:0/CreatedAt:08.08.2017 15:12/Location:İzmir, Türkiye/User/ID:78827323111356416/Tweet:Samsung telef
 ID:894893898149101568/FollowersCount:1019/ListedCount:3/FavoriteCount:7/RetweetCount:0/CreatedAt:08.08.2017 15:11/Location:Dalaman, Muğla/User/ID:3079046086/Tweet:Allah'tan belasını istedi
 ID:894893208358014977/FollowersCount:2/ListedCount:0/FavoriteCount:0/RetweetCount:0/CreatedAt:08.08.2017 15:09/Location:/User/ID:8886872886826636/Tweet:DEVİ ÇEKİLİŞİ! SAMSUNG GALA
 ID:894893036232159232/FollowersCount:200/ListedCount:3/FavoriteCount:0/RetweetCount:17/CreatedAt:08.08.2017 15:08/Location:/User/ID:50662941/Tweet:RT @BayBayNakit: #temassızyaşayanla
 ID:894892962680786944/FollowersCount:362/ListedCount:16/FavoriteCount:0/RetweetCount:17/CreatedAt:08.08.2017 15:06/Location:/User/ID:593703715/Tweet:RT @BayBayNakit: #temassızyaşayanla
 ID:894892509419061249/FollowersCount:916/ListedCount:43/FavoriteCount:0/RetweetCount:0/CreatedAt:08.08.2017 15:00/Location:İstanbul, Türkiye/User/ID:940431734/Tweet:Galaxy S8 satışı Sa
 ID:89489172417169408/FollowersCount:473/ListedCount:0/FavoriteCount:1/RetweetCount:0/CreatedAt:08.08.2017 14:54/Location:/User/ID:761539687139999744/Tweet:Apple anamı sikim androido tı
 ID:894891571134205953/FollowersCount:531/ListedCount:25/FavoriteCount:0/RetweetCount:0/CreatedAt:08.08.2017 14:54/Location:/User/ID:80768946669910016/Tweet:Samsung, Galaxy Note 8'in Y
 ID:894891452207240864/FollowersCount:310/ListedCount:2/FavoriteCount:0/RetweetCount:0/CreatedAt:08.08.2017 14:54/Location:Norvege /User/ID:75622594941184640/Tweet:Samsung TV'de flash
 ID:8948914519619584/FollowersCount:165/ListedCount:0/FavoriteCount:0/RetweetCount:0/CreatedAt:08.08.2017 14:52/Location:/User/ID:744867338898518016/Tweet:İşte Samsung'un en dayanıklı
 ID:8948892132705852/FollowersCount:1010/ListedCount:12/FavoriteCount:0/RetweetCount:24/CreatedAt:08.08.2017 14:44/Location:İstanbul, Türkiye/User/ID:423038930/Tweet:RT @Basmarcatr_Ri
 ID:894888634266829096/FollowersCount:215/ListedCount:0/FavoriteCount:0/RetweetCount:0/CreatedAt:08.08.2017 14:41/Location:/User/ID:538407946/Tweet:samsung tizen tv ler için ben connect u
 ID:894885986211155970/FollowersCount:4/ListedCount:0/FavoriteCount:0/RetweetCount:0/CreatedAt:08.08.2017 14:40/Location:/User/ID:83268749888724993/Tweet:@YouTubept:/youtu.be:/YT
 ID:8948857786759319552/FollowersCount:1801/ListedCount:26/FavoriteCount:1/RetweetCount:0/CreatedAt:08.08.2017 14:39/Location:Ankara, Türkiye/User/ID:49724677/Tweet:@gkdemircem @ProFD
 ID:894885771962509444/FollowersCount:17/ListedCount:0/FavoriteCount:0/RetweetCount:3/CreatedAt:08.08.2017 14:39/Location:/User/ID:808218533171761152/Tweet:RT @Echomahmut: Apple ve #
 ID:894885341810704384/FollowersCount:1063/ListedCount:4/FavoriteCount:0/RetweetCount:0/CreatedAt:08.08.2017 14:37/Location:/User/ID:576358672/Tweet:Geçici de olsa samsung kullanmak kork
 ID:894884668163497984/FollowersCount:2050/ListedCount:153/FavoriteCount:0/RetweetCount:0/CreatedAt:08.08.2017 14:35/Location:Türkiye/User/ID:104634641/Tweet:Samsun'un vellahtı için 12 y
 ID:894883762122629120/FollowersCount:553/ListedCount:1/FavoriteCount:0/RetweetCount:24/CreatedAt:08.08.2017 14:31/Location:İstanbul /User/ID:177492122/Tweet:RT @asmarcatr: Raki şişesi, k
 ID:894883546191601664/FollowersCount:53987/ListedCount:86/FavoriteCount:0/RetweetCount:0/CreatedAt:08.08.2017 14:30/Location:İzmir/User/ID:67255743/Tweet:Samsung Galaxy S8 Active Tanıtı
 ID:894882302228131840/FollowersCount:75/ListedCount:0/FavoriteCount:0/RetweetCount:5/CreatedAt:08.08.2017 14:25/Location:sehir İstanbul/User/ID:870053901423758272/Tweet:RT @doganburak
 ID:894882163153305601/FollowersCount:18723/ListedCount:26/FavoriteCount:0/RetweetCount:1/CreatedAt:08.08.2017 14:25/Location:/User/ID:581141362/Tweet:Darbeye, suya, toza dayanıklı Samsu
 ID:894881243929686016/FollowersCount:1400/ListedCount:23/FavoriteCount:0/RetweetCount:0/CreatedAt:08.08.2017 14:21/Location:/User/ID:563666329/Tweet:Samsung telefon bataryalarını bir bi
 ID:894881070176501760/FollowersCount:4234/ListedCount:16/FavoriteCount:0/RetweetCount:0/CreatedAt:08.08.2017 14:20/Location:İstanbul/User/ID:2332362883/Tweet:Samsung Galaxy S8 Active d
 ID:894880944696719478/FollowersCount:916/ListedCount:43/FavoriteCount:0/RetweetCount:0/CreatedAt:08.08.2017 14:20/Location:İstanbul, Türkiye/User/ID:940431734/Tweet:Samsun Galaxy S8 Act

Bağlantı sağlandı, tweet aramıyor...

Şekil 3.4. Tweet Çekme Uygulaması

Twitter API' lerinin, saatte 180 sorgu ve sorgu başına 100 mesaj kısıtı bulunması dolayısıyla sorgular, periyodik olarak çalıştırılmaktadır. Bu sorgularda çekilen bütün bilgiler aşağıdaki tablodaki gibidir:

Tablo 3.1. Tweet Bilgileri

KOLON ADI	AÇIKLAMA
ID	Tweet Id'si
FOLLOWERSCOUNT	Takipçi sayısı
LISTEDCOUNT	Listelenme sayısı
FAVORITECOUNT	Favori yapılma sayısı
RETWEETCOUNT	Yeniden Tweet yapılma sayısı
CREATEDATE	Oluşturulma tarihi
LOCATION	Bölge
USERID	Kullanıcı numarası
TWEET	Tweet içeriği

Bu uygulama sayesinde aynı Tweetlerin çekilmesi önlenmiştir. Bunun için çekilen son tweet Id'si hafızaya alınarak bu Id'den büyük değere sahip olan, yani daha sonra paylaşılmış olan, tweetlerin çekilmesi sağlanmıştır.

```
Query Sorgu = new Query("lang:tr AND " + lst.getItem(i).toString() );
Sorgu.setSinceId(SonId);
```

Şekil 3.5. Tweet Çekme Kod Satırı

Geliştirilen modül marifetiyle retweetlerin çekilmesine de izin verilmiştir. Böylece retweetlerin de detaylı bilgisine de erişilebilecektir. Tweetlerin paylaşıldığı bölge ve zaman bilgisi de analizde kullanılmak üzere veri uzayına dâhil edilmiştir.

3.4.1.3. Test Verileri

Çalışma kapsamında test için, 10.02.2017 ile 31.07.2017 tarihleri arasında Türkçe olarak paylaşılan ve içinde 50 adet marka (www.gittigidiyor.com sitesinden alınan telefon markaları) isminin geçtiği 312632 adet tweet çekilmiştir. Bu markalar: Samsung, Lg, Htc,

Apple, Huawei, Asus, Lenovo, Sony, Nokia, Casper, Vestel, General Mobile, Xiaomi, Alcatel, Blackberry, Hiking, Anka, Microsoft, Zte, Piranha, Tinmo, Bb Mobile, Excon, Quatro, Torch, Turkcell, Alcatech, Meizu, Sunny, Tp-Link, Ultron Mobile, Vmaxx, Concord, Concox, Dark, Fly, Gigabyte, Haier, Inew, Jcb, Kaan, Niu, Nubia, Oxiron, Praktica, Pyramid, Ulefone, Umi, Unipad ve Yota'dır.

3.4.2. Tweetlerin Düzenlenmesi

Gerçek kullanıcılar tarafından paylaşılmayan, gündem oluşturma veya başka ürünlerle ilgili reklam amacı taşıyan tweetlerin filtrelenmesi gerekmektedir. Bu aşamada ayrıca kendini tekrar eden tweetler de temizlenmektedir.

Çalışma kapsamında telefon markalarının geçtiği bütün tweetler toplanmıştır. Ancak içinde örneğin, “Anka” kelimesi geçen tweetlerin tamamına yakını Anka kuşundan ve Anka uçağından bahsetmektedir. Samsung ve Apple gibi firmaların ise telefon dışında geliştirdiği teknolojiler de bulunmaktadır. Dolayısıyla konu dışı (irrelevant) tweetlerin bir kısmı başlangıçta el yordamıyla veri uzayından çıkarılmıştır.

Tweet atan kullanıcıların karakteristik olarak özensizlikleri, az kelimeyle çok şeyi ifade etme alışkanlığı (Uçan, 2014), Tweet yazımında kural tabanlı bir veri girişi bulunmaması ve editör kontrolüne tabi tutulmaması hatalı, imla ve dilbilgisi kuralları bakımından zayıf mesajların kullanılmasına yol açmaktadır. Kullanıcı kaynaklı hataların tesadüfilik düzeyi yüksek olduğundan tespit edilmesi ve düzeltilmesi de zordur. Ancak “çöp giren çöp çıkar” prensibinden hareketle, kaliteli veri analizi sonuçları için verinin kalitesine ve dolayısıyla ön işlem aşamasına önem verilmiştir.

3.4.2.1. Zemberek Kullanımı ve Tweetlerin Genel Temizliği

İçinde birden fazla marka adı bulunan ve bu yüzden tekrarlı olarak çekilen tweetlerin teke indirilmesi ve Twitter Api'leri marifetiyle yanlışlıkla veri setine dâhil edilen tweetlerin ayıklanması bu aşamada yapılmaktadır. Tweetler içinde geçen ve analize katkı sağlamayan alfa nümerik olmayan karakterlerin ayıklanması da yine bu aşamada yapılmaktadır. Bu karakterlere örnek:

""^|. +&\$€° •{,|! »><¿^~x©-

Şekil 3.6. Alfa Nümerik Olmayan Karakter Örnekleri

Çalışma kapsamında geliştirilen ve Zemberek kütüphanesinin de kullanıldığı modül sayesinde tweetler kelimelere ayrıştırılarak veri tabanına kaydedilmektedir.

Dosyaların Veri Tabanına Kaydedilmesi Formu

Twitlerin Kayıtlı Olduğu Dizin:

Dosyalar

- C:\\Users\\User\\Desktop\\TwitterVerileri\\alcatel\\03_01_2017_17_46.txt
- C:\\Users\\User\\Desktop\\TwitterVerileri\\alcatel\\12_12_2016_12_24.txt
- C:\\Users\\User\\Desktop\\TwitterVerileri\\alcatel\\12_12_2016_14_32.txt
- C:\\Users\\User\\Desktop\\TwitterVerileri\\alcatel\\24_01_2017_19_59.txt
- C:\\Users\\User\\Desktop\\TwitterVerileri\\alcatel\\27_12_2016_12_15.txt
- C:\\Users\\User\\Desktop\\TwitterVerileri\\alcatel\\27_12_2016_21_18.txt
- C:\\Users\\User\\Desktop\\TwitterVerileri\\alcatel\\28_12_2016_16_41.txt
- C:\\Users\\User\\Desktop\\TwitterVerileri\\alcatel\\30_11_2016_13_41.txt
- C:\\Users\\User\\Desktop\\TwitterVerileri\\anka\\03_01_2017_18_01.txt

Şekil 3.7. Ön İşlem ve Kaydetme Ekranı

3.4.2.2. Sözcüklerin Hatalı Olma Durumları için Öneriler

Sözcüklerin tespit edilen hatalı olma durumları aşağıdaki tabloda listelenmiştir. Bu tabloda, hata durumuna ilişkin tweetler içinden tespit edilen bazı örnekler de sunulmaktadır.

Tablo 3.2. Sözcüklerin Hatalı Olma Durumları ve Örnekler

<i>Hatalı Olma Durumu</i>	<i>Örnek</i>
<i>Eksik Harf Kullanımı</i>	Ingstram Tlfn Aglyrsn Tşkkrr Hoşgeldi Dolaşıyosunuz
<i>Bitişik Kelimeler</i>	Sayenizdetümbilgilerimikaybettğim Vermedensahipolabilmeninkeyfiniyaşasın Birmalınngarantikapsamındateslim Güzelülkemizinherköşesidünyanınengelişmişteknolojisinedünyaları Dolandımbütüngecebütnbiaybütnsenenefarkeder

Tablo 3.2. (Devamı)

<i>Hatalı Olma Durumu</i>	<i>Örnek</i>
<i>Fazla harf kullanımı</i>	Meerhabaaa Saağğğooolluuuuğğğğnnnnn Adamsııınnnnııııızzzz Cocuuuuuuuugggggguuuuuuuu Cıldırıyorumdıdockodkckc Sanırımxlsöfkssşk
<i>Yanlış harf kullanımı</i>	öerhaba
<i>Kısaltma kullanımı</i>	Slm Tlfn Aglyrsn Tşkkrr
<i>Yanlış sırada harf kullanımı</i>	Migernnniiim
<i>Ağız farklılığı</i>	Oğurlayır Alamıcam Degistircem Naptılar Olucak Merheba Eyvallah Kitliyonuz Silmiycem Napiym Çıxarılacaq
<i>Rasgele harf kullanımı</i>	Sañjdhjklashdajskdhlkjasdhasjkdahsld dfndjsnkd kdkdkdkdkdkd dkksllskdjdnndnfnndnc dndkssksksksjdjd

Eksik harf kullanımı: uzun bir kelime içinde bir harf eksikliğinin tespit edilmesi nispeten kolay ise de harf eklenmesiyle birden fazla kelimeye dönüşebilen kelimeler için durum biraz daha zorlaşmaktadır. Örneğin, yanlışlıkla “Yarıldı” diye yazılan kelimenin aslı, “yakıldı” ve “yanıldı” olabilmektedir. Eksik harfin geçtiği yerin tespitinde hece uyumsuzluğunun konumunun tespiti kullanılabilmektedir.

Bitişik kelimeler: bitişik kelimelerin tespitinde döngüsel bir algoritmanın işletilmesi gereği görülmektedir. Tespit edilen her bir kelime, bitişik kelime grubu içinden

çıkarılarak işleme devam edilebilir. Kelime sonuna gelen eklerin ise ayrıca dikkate alınması gerekmektedir.

Fazla harf kullanımı: bir hecede bir harfin birden fazla kullanımı hataya sebep olduğu gibi kelime içinde bir harfin üç defa kullanımı da hatadır. Harf tekrarlarını teke düşürme yöntemi literatürde tercih ediliyorsa da Türkçeye sonradan giren “faaliyet” ve “saat” gibi kelimeler dolayısıyla kontrollü olarak azaltılmasında fayda görülmüştür.

Yanlış harf kullanımı: özellikle klavyede birbirine yakın harfler dolayısıyla oluşan bu durum, Türkiye’de “F” klavyenin de dikkate alınmasını zorunlu kılmaktadır. Harflerin klavyedeki sıralarına göre puanlandırılması ve kelimedeki harf puanlarının toplanması ile bu hata durumu tespit edilebilir.

Kısaltma kullanımı: Eksik harf kullanımına benzese de bazen kelimenin aslından çok uzaklaşmaktadır. Bunun için kelimenin uzun haliyle eşleşme yapılabilen tablo kullanılması zorunludur. Bu tablo diğer bütün hatalı kullanımlar için de otomatik düzeltme yardımı sağlayabilecek bir şekilde doldurulabilir. Bu tablonun oluşturulabilmesi için kullanıcı hatalarından öğrenen bir motora ihtiyaç duyulmaktadır.

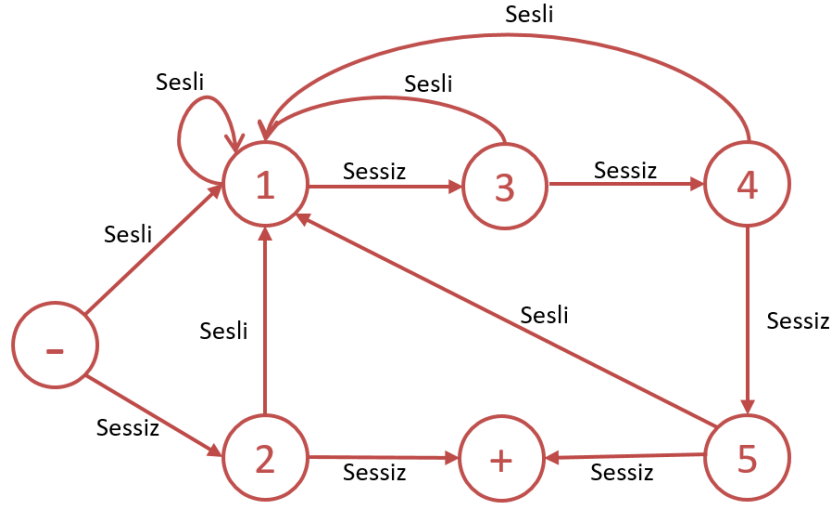
Yanlış sırada harf kullanımı: bu hata durumu, bazen hece uyumsuzluğu ile tespit edilebilirse de çoğu zaman bu yöntem uygun değildir. Harflerin sayı değerleri toplamalarının, sözlükteki kelimelerle kıyaslanması bir çözüm olabilirse de birden fazla hata türünü barındıran “migerennniim” örneğinde bu yöntem, başarısız olacaktır.

Ağız farklılığı: çoğunlukla kelime sonuna gelen eklerde değişikliğe yol açsa da kelime kökünü de değiştirebilmektedir. Tespit edilmesi ve düzeltilmesi için esaslı bir çalışma gerekmektedir.

Rasgele harf kullanımı: rastgele harflere basılarak oluşturulan ifadeler bu sınıftadır. Hece uyumsuzluğu ve tekrarlı harf sayısı üzerinden tespit edilebilmektedir.

Türkçe kelimeler “b”, “c”, “d” ve “g” harfleriyle bitmedikleri için kelime sonunda bu harflerin bulunması durumunda sırasıyla “p”, “ç”, “t” ve “k” harfleriyle değiştirilip sözlükte arattırılabilir.

Kelime içinde dört sessiz harfin ardışık olarak kullanılması, kelime başında iki sessiz harfin kullanılması ve kelime ortasındaki bir hecenin sesli harfle başlaması durumları için çizilen bir durum makinesi Şekil 3.8’de gösterilmektedir.



Şekil 3.8. Kelimenin Hatalı Olma Durumunun Durum Makinesi ile Gösterimi

Bir hece içinde bir harfin peş peşe kullanımı ve bir hecenin beş harfli olması durumları da Türkçe dil yapısına uymamaktadır.

3.4.2.3. Metnin Kelimelerine Ayrıştırılması (Tokenizer)

Tweet içeriklerinin kelimelerine ayrıştırılması için gerekli kod C# programlama dili ile aşağıdaki gibi yazılabilmektedir:

```
foreach (string kelime in twit.Split(' '))
```

Şekil 3.9. Tweetin Kelimelerine Ayrılması (C# Kodu)

Python ile

```
for kelime in twit.split():
```

Şekil 3.10. Tweetin Kelimelerine Ayrılması (Python Kodu)

Aynı kod, T-SQL dili ile aşağıdaki gibi yazılabilecektir:

```

ALTER Function [dbo].[fn_BosluklardanAyir] (
    @Parametre varchar(max)
)
RETURNS
@Veriler TABLE (
    P1 varchar(max)
)
AS
Begin
    Declare
        @Pos int,
        @P1 nvarchar(50)
    Select @Pos = charindex(' ', @Parametre, 0)
    while @Pos!=0
    Begin
        set @P1 = SubString(@Parametre,0, @Pos)
        if rtrim(ltrim( @P1 ))!=' '
            insert into @Veriler
            select @P1
        Set @Parametre = SubString(@Parametre, @pos+1, (LEN(@Parametre)-@pos)+1)
        Select @Pos = charindex(' ', @Parametre, 0)
    End
    RETURN
END

```

Şekil 3.11. Tweetin Kelimelerine Ayrılması (T-SQL Kodu)

Bitişik kelimelerin ayrıştırılması içinse aşağıdaki T-SQL kodu yazılmıştır:

```

select
    Duzeltilmis
    ,left (Duzeltilmis, (select top 1 len(kelime)
        from
            sozluk
        where
            Duzeltilmis like kelime + '%'
        order by len (kelime) desc
    ))
from
    TweetKelimeler
where
    sozlukte is null

```

Şekil 3.12. Bitişik Kelimelerin Ayrılması (T-SQL Kodu)

Kelimelerin köklerine ayrılması (stemmer) için Zemberek kütüphanesi kullanılarak yazılabilecek C# kodu aşağıdaki gibidir:

```

Zemberek zemberek= new Zemberek(new TurkiyeTurkcesi());
kok = zemberek.kelimeCozumle(kelime)[0].kok().icerik();

```

Şekil 3.13. Kelimelerin Köklerine Ayrılması (C# Kodu)

3.4.2.4. Kelimelerin Hecelerine Ayrıştırılması

Kelimelerin hecelerine ayrılması, hatalı yazılıp yazılmadıklarının tespitinde kullanılmaktadır. Türkçe hece yapısına uymayan kelimeler, hecelerine ayıramayacağı için hatalı olarak işaretlenecektir. Kelimelerin hecelerine ayrılması için Zemberek kütüphanesi kullanılarak yazılabilecek C# kodu aşağıdaki gibi tek komut satırından oluşmaktadır:

```
string[] heceler = zemberek.hecele(Kelime);
```

Şekil 3.14. Kelimenin Hecelerine Ayrılması (C# Kodu)

Aynı işlevdeki betik, T-SQL dil ile yazılmak istenirse Şekil 3.15'teki gibi uzun bir kod yazılmalıdır. Ancak her bir kod bloğunun, bir seferde yüzlerce satırı aynı anda etkilediği düşünüldüğünde, eşleniği olan diğer kısa kodlara kıyasla daha etkili çözüm sunacağı görülecektir. Algoritma akış diyagramı ise Şekil 3.16'daki gibidir.

```
update tweetkelimeler set hece = left(hece,len(hece)-2) ,heceler =right(hece,2) +',' +heceler
where hece like '%[bcçdfgğhjklnprsstvyzwx][aeioöüü]'
```

```
update tweetkelimeler set hece = left(hece,len(hece)-3) ,heceler =right(hece,3) +',' +heceler
where
    hece like '%[bcçdfgğhjklnprsstvyzwx][aeioöüü][bcçdfgğhjklnprsstvyzwx]'
```

```
update tweetkelimeler set hece = left(hece,len(hece)-4) ,heceler =right(hece,4) +',' +heceler
where
    hece like '%[bcçdfgğhjklnprsstvyzwx][aeioöüü][bcçdfgğhjklnprsstvyzwx][bcçdfgğhjklnprsstvyzwx]'
```

```
update tweetkelimeler set hece = left(hece,len(hece)-1) ,heceler =right(hece,1) +',' +heceler
where
    hece like '%[aeioöüü][aeioöüü]'
```

```
update tweetkelimeler set hece = left(hece,len(hece)-2) ,heceler =right(hece,2) +',' +heceler
where
    hece like '%[aeioöüü][aeioöüü][bcçdfgğhjklnprsstvyzwx]'
```

```
update tweetkelimeler set hece = left(hece,len(hece)-3) ,heceler =right(hece,3) +',' +heceler
where
    hece like '%[aeioöüü][aeioöüü][bcçdfgğhjklnprsstvyzwx][bcçdfgğhjklnprsstvyzwx]'
```

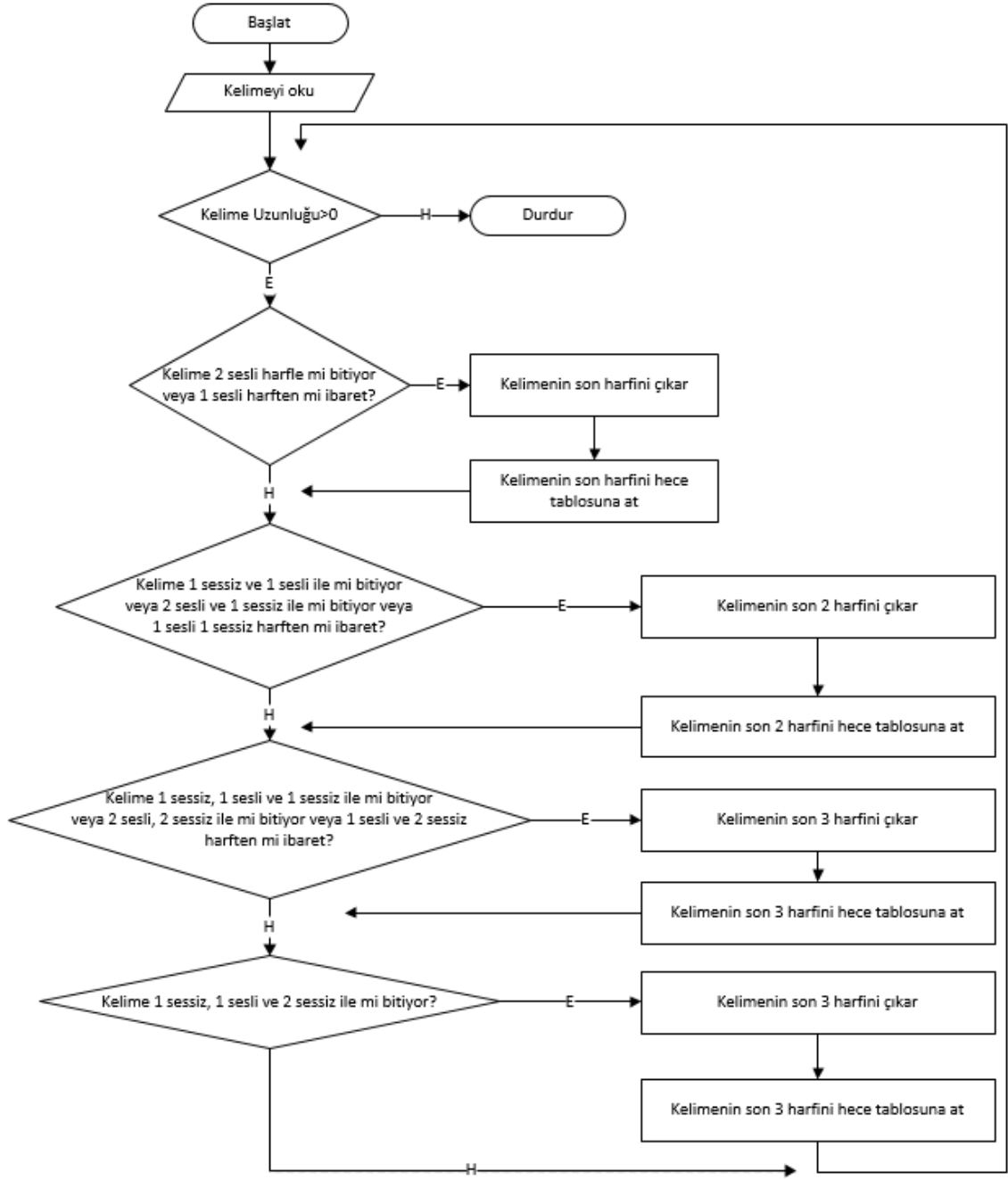
```
update tweetkelimeler set hece = left(hece,len(hece)-1) ,heceler =right(hece,1) +',' +heceler
where
    hece like '%[aeioöüü][aeioöüü]'
```

```
update tweetkelimeler set hece = left(hece,len(hece)-1) ,heceler =right(hece,1) +',' +heceler
where
    hece like '[aeioöüü]'
```

```
update tweetkelimeler set hece = left(hece,len(hece)-2) ,heceler =right(hece,2) +',' +heceler
where
    hece like '[aeioöüü][bcçdfgğhjklnprsstvyzwx]'
```

```
update tweetkelimeler set hece = left(hece,len(hece)-3) ,heceler =right(hece,3) +',' +heceler
where
    hece like '[aeioöüü][bcçdfgğhjklnprsstvyzwx][bcçdfgğhjklnprsstvyzwx]'
```

Şekil 3.15. Kelimenin Hecelerine Ayrılması (T-SQL Kodu)



Şekil 3.16. Kelimenin Hecelerine Ayrılması Algoritması

Çalışma kapsamında toplanan tweetlerde geçen 1591598 adet kelime hecelerine ayrılabilirken, 49561 adet kelime hecelerine ayırlanamamıştır. Yanlış yazılan kelimelere ek olarak internet adresleri ve hashtag gibi ifadeler de hecelerine ayırlanamamıştır.

3.4.2.5. Kelimelerin Sözlük ile Eşleştirilmesi

Sözlükte karşılığı bulunmayan kelimeler, çeşitli ön işlemlerden geçirildikten sonra yine de düzeltilmediyse öznitelik sayısının azaltılabilmesi için analize dâhil edilmemektedir. Kelimelerin sözlükte taranması için T-SQL dili kullanılarak iki tablo arasında “inner join” yöntemi kullanılmaktadır.

```
update tk set Duzeltilmis=s.Kelime, Sozlukte=1
from
    TweetKelimeler tk
    inner join sozluk s on s.kelime=tk.kelime
where
    sozlukte is null
```

Şekil 3.17. Kelimelerin Sözlük ile Eşleştirilmesi (T-SQL Kodu)

Python dili kullanılarak yapılacak eşleştirmede aşağıdaki gibi sözlük dosyasının okutulması gerekmektedir. Bu yöntemde her bir kelime için dosyanın tekrar okunması gerektiği için bu kod parçasığı oldukça düşük performanslı olarak icra edilmektedir.

```
sozluk = open("C:/Users/User/Desktop/TwitterVerileri/Sozluk.txt")
if sozluk.readlines().count(kelime+ "\n") >0:
    print("Bulunan Kelime var: " + kelime)
```

Şekil 3.18. Kelimelerin Sözlük ile Eşleştirilmesi (Python Kodu)

Kelimelerin ASCII karşılıklarıyla eşleştirilmesi için yazılacak python kodu aşağıdaki gibi olabilir:

```
kaynak = "şçöğüŞÇÖĞÜİ"
hedef = "scoguiSCOGUI"
ceviri = str.maketrans(kaynak, hedef)
AsciiKarsiligi=kelime.translate(ceviri)
```

Şekil 3.19. Kelimelerin ASCII Karşılıklarıyla Eşleştirilmesi (Python Kodu)

Aynı kodun T-SQL karşılığı ise:

```
replace (replace (replace (replace (replace (upper(marka),'Ş','S'),'Ç','C'),'İ','I'),'Ü','U'),'Ö','O')
```

Şekil 3.20. Kelimelerin ASCII Karşılıklarıyla Eşleştirilmesi (T-SQL Kodu)

B, c, d ve g harfleriyle biten kelimeler eşleniklerine (p, ç, t ve k) çevrilerek yeniden sözlükte taranmaktadır (Şekil 3.21). Bu sayede, “görüşüb”, “ölüb”, “arab”, “yurd”, “mürid”, “ısıd”, “imzalayıb” gibi kelimeler düzeltilebilmiştir.

```
update tk set Duzeltilmis=s.Kelime, Sozlukte=1
from
    TweetKelimeler tk
    inner join sozluk s on s.kelime = left (tk.kelime,len(tk.kelime)-1) +'p'
where
    sozlukte is null
    and tk.kelime like '%[b]'
```

Şekil 3.21. Son Harfi Düzeltme Kodu (T-SQL)

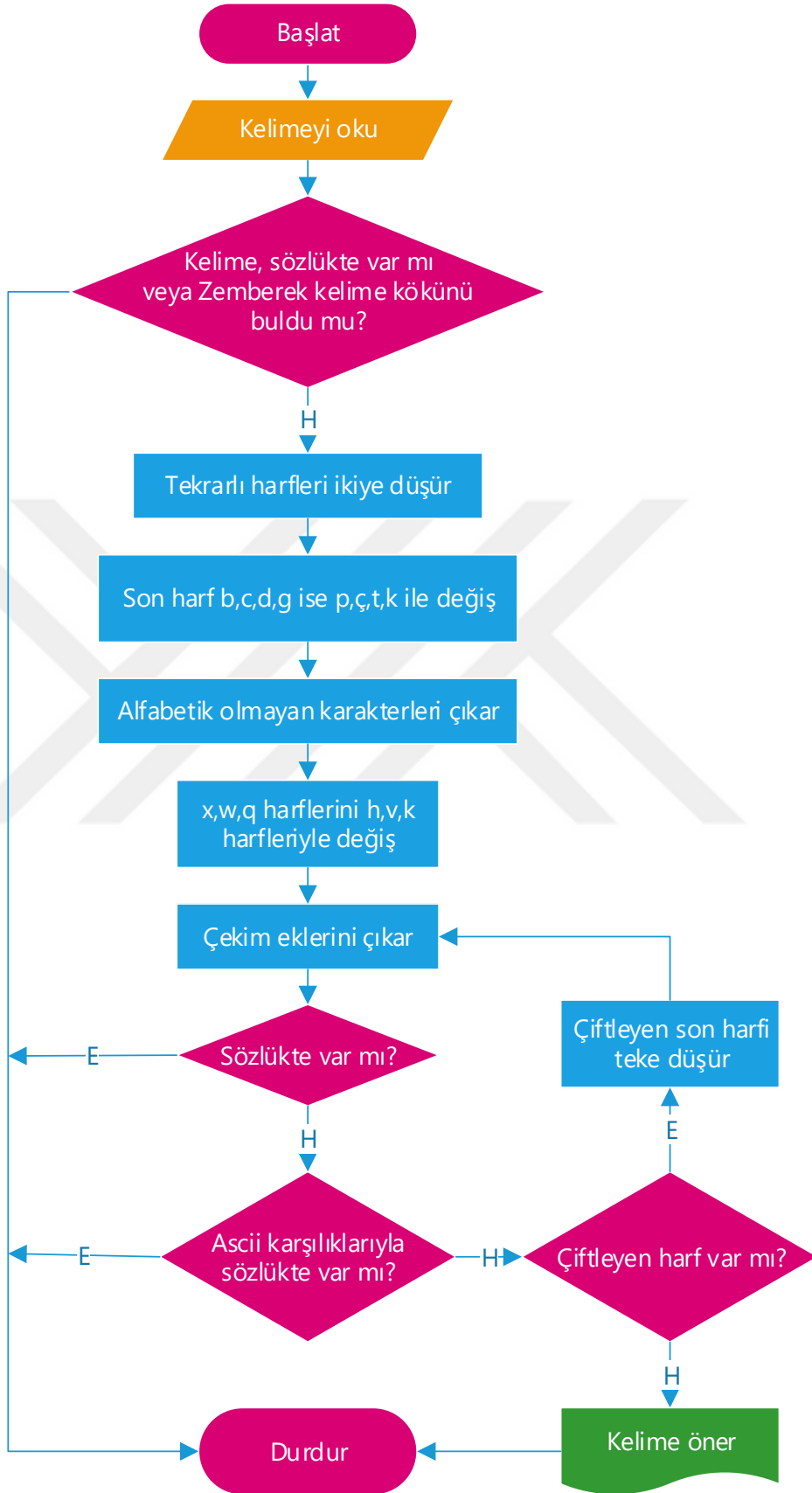
Yukarıda görüldüğü gibi, T-SQL kod karşılığında her bir harfin ayrıca dönüştürülmesi gerekmektedir.

Son olarak, Zemberek tarafından kökü bulunabilen kelimelerin de sözlüğe dâhil edilerek eşleşmesi sağlanmıştır (Şekil 3.22).

```
update tk set duzeltilmis= tk.kelime, sozlukte =1,hatalı= null
from TweetKelimeler tk left join sozluk s on tk.kok =s.kelime where s.kelime is null and len(kok)>2
```

Şekil 3.22. Kelimelerin Sözlükte Var Diye İşlenmesi (T-SQL)

Kelime düzeltme algoritmasının akış diyagramı aşağıdaki gibidir:



Şekil 3.23. Kelime Düzeltme Algoritması Akış Diyagramı

3.4.2.6. Kelime Veri Tabanı Tasarımı



Şekil 3.24. Kelime Veri Tabanının Oluşturulması

Sözlük tablosuna, Zemberek veri tabanındaki 1140253 adet kelimeye, bir çalışmadan (Serdar, 2013) 62784 adet kelime, Türk Dil Kurumundan (TDK, 2017) 72300 adet kelime, Wikisözlükten (Wikisözlük, 2017) 63813 kelime eklendi

Sektörde faaliyet gösteren 50 adet marka adı da listeye dâhil edildi. Türkçede sıklıkla kullanılan ve duygu ifade etmeyen kelimeler, durak kelimeleri (stop words) de veri tabanına eklendi (Sarıgül, 2009). Son olarak NVİ (Nüfus ve Vergi İşleri)’den il ve ilçe adları (Nvi, 2017) ve 7119 adet özel isim de sözlüğe eklendi (Dilipak vd., 2016).

3.4.2.7. Q ve F Klavye Düzeni ile Hata Tespiti

Q ve F klavye düzenine göre harflerin konumları aşağıdaki tabloda işaretlenmiştir. Buna göre x değerleri satır bilgisini, y değerleri ise sütun bilgisini tutmaktadır. Örneğin, solda en üstte bulunan Q harfi, Q klavye düzeninde birinci satır ve birinci sütunda bulunmaktadır. Hemen yanında bulunan W harfi ise birinci satır ikinci sütun harfidir.

Tablo 3.3. Q Klavye Düzeni Tablosu

Satır (x)	Sütun (y)	Q düzeni	F düzeni
1	1	Q	F
	2	W	G
	3	E	Ğ
	4	R	I
	5	T	O

Tablo 3.3. (Devamı)

	6	Y	D
	7	U	R
	8	I	N
	9	O	H
	10	P	P
	11	Ğ	Q
	12	Ü	W
2	1	A	U
	2	S	İ
	3	D	E
	4	F	A
	5	G	Ü
	6	H	T
	7	J	K
	8	K	M
	9	L	L
	10	Ş	Y
	11	İ	Ş
3	1	Z	J
	2	X	Ö
	3	C	V
	4	V	C
	5	B	Ç
	6	N	Z
	7	M	B
	8	Ö	X
	9	Ç	Q

Q klavye düzeninde Q harfi basılması gerekirken W basılması durumu, aynı satırdaki komşu sütunda bulunan harfin basılması durumudur. Q harfi yerine A harfinin basılması durumu ise aynı sütunda ve komşu satırda bulunan harfin basılması manasına gelmektedir. Bu durum satır veya sütun farklarından birinin 1 olmasına karşılık gelmektedir. Aşağıdaki Öklit (Euclid) uzaklığı denkleminde mesafenin 1 olması durumu, klavye düzeninden kaynaklanan hataların tespitinde kullanılmıştır.

$$d(A, B) = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$$

3.4.2.8. Python Betiklerinin Yazılması

Tweetlerin bulunduğu “txt” dosyası açıldıktan sonra, tweet dosyasına satır düzeyinde erişilerek (Şekil 3.25) tweet metni ve marka bilgisi alınmaktadır (Şekil 3.26).

```
tweetler = open("C:/Users/User/Desktop/TwitterVerileri/TumTweet.txt")
```

Şekil 3.25. Tweet Dosyasına Erişim

```
for satir in tweetler:
    twit= satir.partition("/Tweet:") [2]
    marka= satir.partition("/Marka:") [2]
    marka= marka.partition("/") [0]
```

Şekil 3.26. Tweet Metnini Elde Etme

“Retweet” olmayan ve tweet metni içinde marka adı geçen tweet metni içindeki emojiler, tweet metninde ayrıştırılır (Şekil 3.27).

```
if not(twit.startswith("RT ") or twit.count(marka)<1 ):
    if (twit.isalnum()==False):
        twit = twit.replace(":", " : ")
        twit = twit.replace(":D", " :D ")
        twit = twit.replace("(", " ( ")
        twit = twit.replace("/", " / ")
        twit = twit.replace("\\", " \\ ")
        twit = twit.replace("|", " | ")
        twit = twit.replace("-", " - ")
        twit = twit.replace("- (", " - ( ")
```

Şekil 3.27. Tweet İçindeki Emojileri Ayırma

Analize dâhil edilmediği için tweet metni içindeki nümerik karakterler silinir (Şekil 3.28).

```

if (twit.isalpha()==False):
    twit = twit.replace("0", " ")
    twit = twit.replace("1", " ")
    twit = twit.replace("2", " ")
    twit = twit.replace("3", " ")
    twit = twit.replace("4", " ")
    twit = twit.replace("5", " ")
    twit = twit.replace("6", " ")
    twit = twit.replace("7", " ")
    twit = twit.replace("8", " ")
    twit = twit.replace("9", " ")

```

Şekil 3.28. Tweet İçindeki Nümerik İfadeleri Silme

Tweet metni içinde birden fazla boşluk karakteri hatası düzeltilir (Şekil 3.29).

```

for i in range(0,twit.count(" ")):
    twit = twit.replace(" ", " ")

```

Şekil 3.29. Tweet İçindeki Fazla Boşlukları Teke İndirme

Tweet metni, boşluk karakterlerinden ayrıştırılarak kelime dizisi oluşturulur (Şekil 3.30).

```

for kelime in twit.split():

```

Şekil 3.30. Tweeti Kelimelere Ayırma

Kelime içeriğinin, hashtag, mention, link içermesi veya alfabetik olmayan karakterler içermesi durumu kontrol edilir (Şekil 3.31).

```

if not (kelime.startswith("#") or kelime.startswith("@")
or kelime.startswith("http") or kelime.isalnum()==False or len(kelime)<3 ):

```

Şekil 3.31. Kelime Hashtag, Mention, Link, vs. Değilse

Sözlük dosyası belleğe yüklenerek, her bir kelime, sözlük içinde arattırılır (Şekil 3.32).

```

sozluk = open("C:/Users/User/Desktop/TwitterVerileri/Sozluk.txt")
if sozluk.readlines().count(kelime+ "\n") >0:
    print("Bulunan Kelime var: " + kelime)

```

Şekil 3.32. Kelimenin Sözlükte Aranması

Sözlükte karşılığı bulunmayan kelimelerin Türkçeye özgü karakter kullanılmadan yazılmış olması durumu değerlendirilerek karakter dönüşümü için tablo oluşturulur (Şekil 3.33).

```
trchar= "şçöğüıışçöğüı"
enchar = "scoquiSCOGUI"
AsciiTablosu = str.maketrans(enchar, trchar)
```

Şekil 3.33. Harflerin ASCII Karşılığının Alınması

Türkçeye özgü harfler ASCII eşlenikleriyle değiştirilerek sözlükte yeniden taranır (Şekil 3.34).

```
sozluk = open("C:/Users/User/Desktop/TwitterVerileri/Sozluk.txt")
if sozluk.readlines().count(kelime.translate(AsciiTablosu)+ "\n") >0:
    print("Bulunan Kelime var: " + kelime)
```

Şekil 3.34. Kelimelerin Türkçe Harf Kullanılmadan Sözlükte Aratılması

Kelimenin son karakterinin “b”, “c”, “d” veya “g” harfleriyle bitmesi durumunda “p”, “ç”, “t” ve “k” harfleriyle değiştirilerek sözlükte taranır (Şekil 3.35 ve Şekil 3.36).

```
bcdg= "bcdg"
pçtk = "pçtk"
SertYumuşakTablosu = str.maketrans(bcdg, pçtk)
```

Şekil 3.35. Harflerin Sert Yumuşak Karşılıklarının Alınması

```
sozluk = open("C:/Users/User/Desktop/TwitterVerileri/Sozluk.txt")
if sozluk.readlines().count(kelime[1:len(kelime)-1] + kelime[-1].translate(SertYumuşakTablosu)+ "\n") >0:
    print("Bulunan Kelime var: " + kelime)
```

Şekil 3.36. Kelimenin Son Harfinin Sert Yumuşak Karşılıkları ile Aratılması

Son olarak kelime içinde bir harfin yanlışlıkla çoğullaması durumu düzeltilir ve sözlükte taranarak (Şekil 3.37) bir sonraki kelimeye geçilir.

```

yeni=""
for i in range(len(kelime)-1):
    if (kelime[i:i+1]!= kelime[i+1:i+2]):
        yeni =yeni+kelime[i:i+1]
yeni=yeni+ kelime[-1]
sozluk = open("C:/Users/User/Desktop/TwitterVerileri/Sozluk.txt")
if sozluk.readlines().count(yeni+ "\n") >0:
    print("Bulunan Kelime var: " + yeni)

```

Şekil 3.37. Çoğullayan Harflerin Teke Düşürülmesi

3.4.3. Tweetlerin Sınıflandırılması

Sosyal medyada paylaşılan az sayıdaki yorum, gerçeği daha fazla temsil gücüne sahip olması hasebiyle diğer bütün yorumlara baskın gelebilmektedir. Sayıca az olmasına rağmen ikna kabiliyeti yüksek olan bu yorumların bilgisayar yöntemleriyle etkinliğinin ortaya konması ise oldukça zordur. Aynı şekilde takipçi sayısı yüksek kullanıcılar tarafından paylaşılan tweetlerin de diğerlerine kıyasla daha fazla kitleyi etkisi altına aldığı yadsınamaz.

Paylaşılan yorumlar bazen de gerçeği yansıtmaktan ziyade kullanıcıların psikolojisinin bir dışavurumu olabilmektedir.

3.4.3.1. Eğitim Veri Setinin Oluşturulması

Yüzlerce Tweet incelendikten sonra herhangi bir tweetin dâhil olduğu duygu sınıfını doğrudan belirleme becerisi olan ifadelerin bir listesi hazırlandı. Eğitim veri setinin oluşturulmasında da öncelikle bu duygu belirleyici ifadelerden (1766 adet) faydalanıldı. 169875 tweet, duygu belirleyici ifadeler kullanılarak duygulandırıldı (26.4 dakika sürdü). Duygu belirleyici ifadelere örnek:

- Utanç verici
- Sabrımı zorla
- Tahammül edilmez
- Bu ne arkadaş
- Allah belanı
- Rezalet
- Sövmemek için

- Yetmezmiş gibi
- Tüketici hakları
- Utanmadan

Duygu belirleyici ifadelerin geçmediği tweetler içinse duygu simgeleri (emojiler) kullanılarak belirli bir sınıfa (olumlu, olumsuz) atanmaları sağlandı. Sık kullanılan ve duyguyu belirlemede etkili olduğu düşünülen emojiiler, Tablo 3.4'te sunulmuştur.

Tablo 3.4. Duygu Belirleyen Emojiiler

Olumsuz Emoji	Olumlu Emoji
:(	:)
:-)	:)
=)	=)
;(	:D
x(	;))
>:)	
:^o	
l-)	
8o)	
(u)	
:s	

Çekilen tweetler içinden 2700 adet tweet okunarak, 2039 adet tweet manuel olarak duygu sınıflarına atanmıştır. 35514 adet tweet ise duygu belirleyen emojiiler (16 adet) kullanılarak sınıflandırılabilmiştir (13 sn sürdü). Weka programı kullanılarak, Naive Bayes yöntemine göre “cross-validation” değeri 10 seçildiğinde %75,28 başarı ile sınıflandırılabilmiştir. Destek Vektör Makinesi (SVM) yöntemine göre ise %83,86 başarı ile sınıflandırılabilmiştir.

3.4.3.2. T-SQL ile Naive Bayes Sınıflandırma

Sınıflandırma yapılabilmesi için öncelikle tweetler içinde geçen kelimeler bir tabloya atılır (Şekil 3.38). Bu tablodaki kelimeler eşsizdir (unique: bir kelime bir kere geçmektedir). “TweetKelimeler” tablosu ile “Kelimeler” tablosu eşleştirilir (Şekil 3.39).

```
truncate table kelimeler
insert into kelimeler (Kelime)
select
    distinct Kok
from
    TweetKelimeler tk, tweet t
where
    t.TweetId =tk.TweetId and t.Tweet not like 'rt %' and tk.Sozlukte =1
```

Şekil 3.38. Kelimeler Tablosunun Oluşturulması

```
update tk
set tk.kelimeId =k.Id
from
    TweetKelimeler tk, Kelimeler k
where
    tk.kok =k.Kelime
```

Şekil 3.39. “TweetKelimeler” Tablosu ile “Kelimeler” Tablosunun Eşleştirilmesi

Her bir kelimenin tweet içinde kaç defa geçtiği bilgisini tutan “TweetFrekans” tablosu oluşturulur (Şekil 3.40).

```
truncate table [KelimeFrekans]
insert into KelimeFrekans ([KelimeId] ,[TweetId] ,[frekans],duygu)
select
    kelimeId,t.tweetId, count(*), t.uygu_egitim
from
    TweetKelimeler tk, tweet t
where
    t.TweetId =tk.TweetId and t.Tweet not like 'rt %'
    and tk.Sozlukte =1 and kelimeId is not null
group by
    kelimeId, t.tweetId, t.uygu_egitim
```

Şekil 3.40. “TweetFrekans” Tablosunun Oluşturulması

Her bir kelime için tweetlerde toplam kaç defa geçtiği bilgisi hesaplanır. Aynı hesap olumlu, olumsuz ve nötr sınıfları için de tekrarlanır (Şekil 3.41).

```

update Kelimeler
set TweetdeGecmeSayisi = (select sum(frekans) from KelimeFrekans
                           where Kelimeler.Id = KelimeId)

update Kelimeler
set OlumludaGecmeSayisi = (select sum(frekans) from KelimeFrekans
                           where Kelimeler.Id = KelimeId and duygu=1)

update Kelimeler
set OlumsuzdaGecmeSayisi = (select sum(frekans) from KelimeFrekans
                             where Kelimeler.Id = KelimeId and duygu=-1)

update Kelimeler
set NotrdeGecmeSayisi = (select sum(frekans) from KelimeFrekans
                         where Kelimeler.Id = KelimeId and duygu=0)

```

Şekil 3.41. Kelimelerin Tweetlerde Geçme Sayılarının Alınması

Hesaplamalarda kullanılmak üzere tweet sayısına ve her bir sınıf için toplam tweet sayılarına ulaşılır (Şekil 3.42).

```

select @TweetSayi = count(distinct tweetId) from tweet
select @PozitifTweetSayi= count(distinct tweetId) from tweet where duygu_egitim =1
select @NegatifTweetSayi= count(distinct tweetId) from tweet where duygu_egitim =-1
select @NotrTweetSayi = count(distinct tweetId) from tweet where duygu_egitim=0

```

Şekil 3.42. Tweet Sayılarının Elde Edilmesi

Tweetlerde geçen farklı kelime sayısı bulunur. Aynı işlem pozitif, negatif ve nötr sınıflar için de uygulanır (Şekil 3.43).

```

select @KelimeSayi = count(distinct kelimeId) from [KelimeFrekans]
select @PozitifKelimeSayi = count(distinct kelimeId) from [KelimeFrekans] where duygu =1
select @NegatifKelimeSayi = count(distinct kelimeId) from [KelimeFrekans] where duygu =-1
select @NotrKelimeSayi = count(distinct kelimeId) from [KelimeFrekans] where duygu=0

```

Şekil 3.43. Kelime Sayılarının Elde Edilmesi

Her bir kelime için olumlu, olumsuz ve nötr tweetlerde geçme olasılıkları hesaplanır (Şekil 3.44).

```

update kelimeler
set
  olumludaGecmeOlasiligi = (isnull(olumludagecmesayisi,0) +1) / (@PozitifKelimeSayi+@KelimeSayi)
  ,olumsuzdaGecmeOlasiligi = (isnull(olumsuzdagecmesayisi,0) +1)/ (@NegatifKelimeSayi+@KelimeSayi)
  ,NotrdeGecmeOlasiligi = (isnull(NotrdeGecmeSayisi,0) +1) / (@NotrKelimeSayi +@KelimeSayi)

```

Şekil 3.44. Kelime Olasılıklarının Hesaplanması

Tweetlerin, pozitif, negatif veya nötr olma olasılıkları hesaplanır (Şekil 3.45).

```

update tweet
set
  PozitifOlasilik = @PozitifTweetSayi/@TweetSayi
  ,NegatifOlasilik= @NegatifTweetSayi/@TweetSayi
  ,NotrOlasilik   = @NotrTweetSayi/@TweetSayi

```

Şekil 3.45. Tweet Olasılıklarının Hesaplanması

Her bir tweet için pozitif, negatif veya nötr sınıflara ait olma olasılıkları hesaplanır (Şekil 3.46).

```

select
  tweetId
  ,EXP(SUM(LOG(OlumsuzdaGecmeOlasiligi))) OlumsuzdaGecmeOlasiligi
  ,EXP(SUM(LOG(OlumludaGecmeOlasiligi))) OlumludaGecmeOlasiligi
  ,EXP(SUM(LOG(NotrdeGecmeOlasiligi)))   NotrdeGecmeOlasiligi
into #olasilik
from
  KelimeFrekans kf, kelimeler k
where
  kf.KelimeId =k.Id
group by
  tweetId

```

Şekil 3.46. Sınıflara Ait Olma Olasılıklarının Hesaplanması

Bulunan değerler ilgili sınıfın olasılık değerleri ile çarpılarak nihai olasılık değerleri bulunur (Şekil 3.47).

```

update t
set
  PozitifOlasilik = PozitifOlasilik * OlumludaGecmeOlasiligi
  ,NegatifOlasilik = NegatifOlasilik * OlumsuzdaGecmeOlasiligi
  ,NotrOlasilik   = NotrOlasilik   * NotrdeGecmeOlasiligi
from
  tweet t, #olasilik o
where
  t.TweetId =o.TweetId

```

Şekil 3.47. Olasılık Değerlerinin Hesaplanması

Tweetin hangi sınıfta bulunma olasılığı daha yüksekse, tweet o sınıfa dâhil edilir (Şekil 3.48).

```

update tweet set duygu_tahmin = 1
where
    isnull(PozitifOlasilik,0) > isnull(NegatifOlasilik,0)
    and isnull(PozitifOlasilik,0) > isnull(NotrOlasilik,0)

update tweet set duygu_tahmin = -1
where
    isnull(NegatifOlasilik,0) > isnull(PozitifOlasilik,0)
    and isnull(NegatifOlasilik,0) > isnull(NotrOlasilik,0)

```

Şekil 3.48. Tweetlere Duygu Atanması

3.4.3.3. Linux Küme Üzerinde Hadoop Ortamının Oluşturulması

Çalışma kapsamında, Linux Ubuntu işletim sistemine sahip makinelere, Hadoop 1.0.3 sürümü kurulumu yapılmıştır (İlter, 2012; Burgaz, 2016).

Linux'un mevcut paketlerinin güncellenerek, sistemin güncel kernel sürümüne yükseltilmesi ve paket kurulumları esnasında sorun yaşamamak için öncelikle aşağıdaki kod çalıştırılır:

```
apt dist-upgrade
```

Şekil 3.49. Linux Paketlerinin Güncellenmesi

Kurulumu başlamadan önce Java, openssh-server ve rsync paketlerinin kurulması gerekmektedir. Java kurulumu için sırayla aşağıdaki kodlar çalıştırılır:

```

sudo apt-get install software-properties-common
sudo apt-add-repository ppa:webupd8team/java
sudo apt-get update

```

Şekil 3.50. Java kurulumu

Paket depoları ile mevcut paketler karşılaştırılır ve mevcut liste güncellenir. Mevcut paketlerin güncellenmesi için:

```

apt-get upgrade
sudo apt-get install oracle-java8-installer

```

Şekil 3.51. Mevcut Paketlerin Güncellenmesi

Hadoop kümesi içinde ana makineden diğer düğümlere erişmek için SSH kurulumu yapılır:

```
sudo apt-get install ssh
sudo apt-get install openssh-server
sudo apt-get install openssh-client
```

Şekil 3.52. SSH Kurulumu

Rsync paketinin kurulumu yapılır:

```
sudo apt-get install rsync
```

Şekil 3.53. Rsync Paketinin Kurulumu

Paket kurulumları tamamlandıktan sonra ilk olarak Hadoop kullanıcısı tanımlanır:

```
Adduser Hadoop
```

Şekil 3.54. Hadoop Kullanıcısının Tanımlanması

Sonra “visudo” ile

```
root ALL=(ALL:ALL) ALL
```

Şekil 3.55. Root Kullanıcısının Yetkilerinin Bulunduğu Satır

Satırın altına aşağıdaki kod satırı eklenerek hadoop kullancısına yetkiler tanımlanır.

```
hadoop ALL=(ALL:ALL) ALL
```

Şekil 3.56. Hadoop Kullanıcı Yetkileri

Hadoop kullanıcısı için “opt” ve “srv” klasörlerine yetki verilir.

```
sudo chown -R Hadoop:Hadoop /opt
sudo chown -R hadoop:hadoop /srv
```

Şekil 3.57. “opt” ve “srv” Klasör Yetkileri

Kullanıcı profilini güncellemek için .bashrc dosyası açılarak:

```
sudo gedit ~/.bashrc
```

Şekil 3.58. .bashrc Dosyasını Düzenlemek

Aşağıdaki kod satırları eklenir.

```
#HADOOP VAR
export JAVA_HOME=/usr/lib/jvm/java-8-oracle
#HADOOP VARIABLES START
export JAVA_HOME=/usr/lib/jvm/java-8-oracle
export HADOOP_INSTALL=/opt/hadoop-2.7.2
export PATH=$PATH:$HADOOP_INSTALL/bin
export PATH=$PATH:$HADOOP_INSTALL/sbin
export HADOOP_MAPRED_HOME=$HADOOP_INSTALL
export HADOOP_COMMON_HOME=$HADOOP_INSTALL
export HADOOP_HDFS_HOME=$HADOOP_INSTALL
export YARN_HOME=$HADOOP_INSTALL
export
HADOOP_COMMON_LIB_NATIVE_DIR=$HADOOP_INSTALL/lib/native
export HADOOP_OPTS="-Djava.library.path=$HADOOP_INSTALL/lib"
#HADOOP VARIABLES END
```

Şekil 3.59. .bashrc Dosyasına Eklenecek Düzenlemeler

Oluşturulan değişkenler hafızaya alınır:

```
source ~/.bashrc
```

Şekil 3.60. .bashrc Dosyasının Hafızaya Alınması

Yukarıda anlatılan hazırlık işlemleri bitirildikten sonra aşağıdaki gibi bir internet adresinden Hadoop kurulum dosyası indirilir:

```
wget http://ftp.itu.edu.tr/Mirror/Apache/hadoop/common/hadoop-2.7.2/hadoop-2.7.2.tar.gz
```

Şekil 3.61. Hadoop Kurulum Dosyası İndirilmesi

“tar” komutu ile dosya açılır:

```
tar -zxvf hadoop-2.7.2.tar.gz
```

Şekil 3.62. Hadoop Kurulum Dosyasının Açılması

Kurulum Klasörü “opt” klasörü altına kopyalanır:

```
mv hadoop-2.7.2 /opt
```

Şekil 3.63. Hadoop Kurulum Klasörünün Kopyalanması

Hadoop ayarlarını yapmak için “/opt/hadoop-2.7.2/etc/Hadoop” dizininde bulunan core-site.xml, hadoop-env.sh, hdfs-site.xml, mapred-site.xml, yarn-site.xml, master ve slave dosyalarında gerekli düzenlemeler yapılır.

Host dosyası açılır ve gerekli ipler yazılır.

```
sudo gedit /etc/hosts
```

Şekil 3.64. Host Dosyasının Düzenlenmesi

SSH ile ilgili ayarlar yapılır (key oluşturulur):

```
ssh-keygen -t rsa
```

Şekil 3.65. Şifre Oluşturulması

Oluşturulan şifre her bir makineye gönderilir:

```
ssh-copy-id hadoop@diğerMakineAdı
```

Şekil 3.66. Şifrenin Diğer Makinelere Gönderilmesi

Diğer makinelere erişim test edilerek, ssh kurulumunun tamamlandığı teyid edilir:

```
ssh hadoop@ diğerMakineAdı
```

Şekil 3.67. Diğer Makinelere Girişin Test Edilmesi

Hadoop kurulumu tamamlandıktan sonra Hadoop’u başlatmak için ana makinede iken aşağıdaki kod çalıştırılır:

```
hdfs namenode -format
```

Şekil 3.68. Ana Makinenin Başlatılması

Hdfs başlatılır:

```
start-dfs.sh
```

Şekil 3.69. Dosya Sisteminin Başlatılması

Başlatma işlemi başarılı olmadığında önce sonlandırılır:

```
stop-dfs.sh
```

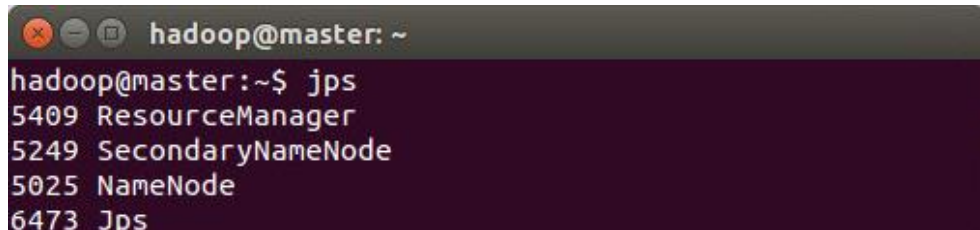
Şekil 3.70. Dosya Sisteminin Sonlandırılması

Son olarak yarn başlatılır:

```
start-yarn.sh
```

Şekil 3.71. Yarn’ın Başlatılması

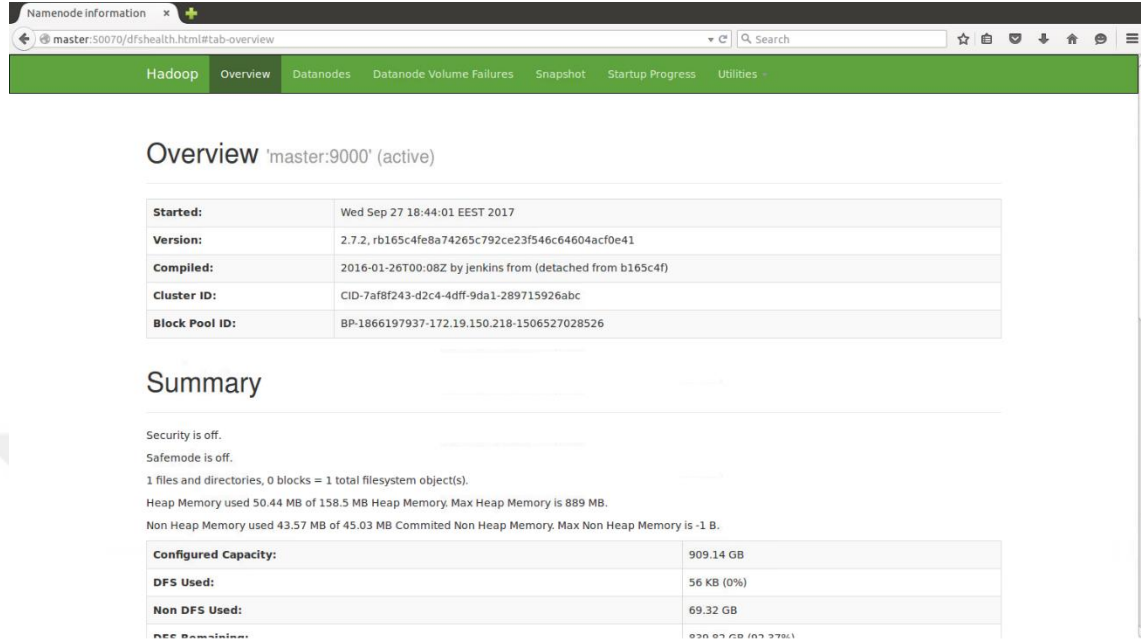
Hadoop ortamının aktif olduğunu anlamak için “jps” komutu çalıştırılabilir.



```
hadoop@master: ~
hadoop@master:~$ jps
5409 ResourceManager
5249 SecondaryNameNode
5025 NameNode
6473 Jps
```

Şekil 3.72. Jps Komutu ile Çalışan Hizmetlerin Listelenmesi

Tarayıcıya “<http://master:50070>” yazıldığında aşağıdaki gibi bir ekran görüntüsü alınabilmektedir.



Şekil 3.73. Hadoop Web Ara Yüzü

2 adet veri düğümü (data node) için bilgiler, web ara yüzünden görüntülenebilmektedir.

Datanode Information										
In operation										
Node	Last contact	Admin State	Capacity	Used	Non DFS Used	Remaining	Blocks	Block pool used	Failed Volumes	Version
slave2:50010 (172.19.150.219:50010)	1	In Service	454.57 GB	28 KB	34.7 GB	419.87 GB	0	28 KB (0%)	0	2.7.2
slave1:50010 (172.19.150.223:50010)	0	In Service	454.57 GB	28 KB	34.62 GB	419.95 GB	0	28 KB (0%)	0	2.7.2

Şekil 3.74. Datanode Bilgileri

3.4.3.4. Hadoop Ortamında Hive Kurulumu

Bu bölümde Hive 2.0 kurulumunun nasıl yapılacağına ilişkin detaylı bilgiler bulunmaktadır (Sarıkaya, 2017). Hive kurulum dosyasını Linux ortamında dosya sunucusundan indirmek için aşağıdaki komut çalıştırılır.

```
wget http://ftp.itu.edu.tr/Mirror/Apache/hive/hive-2.3.0/apache-hive-2.3.0-bin.tar.gz
```

Şekil 3.75. Hive Kurulum Dosyasının İndirilmesi

İndirilen dosyayı açmak için,

```
tar -xvf apache-hive-2.3.0-bin.tar.gz
```

Şekil 3.76. Hive Kurulum Dosyasının Açılması

Hive kurulum dosyaları için bir dizin oluşturulur ve Hadoop kullanıcısına dosyaya erişim yetkisi tanımlanır. Bunun için root kullanıcı yetkisi gerekmektedir.

```
su
mkdir /opt/hive
chown -R hadoop:master /opt/hive
```

Şekil 3.77. Hive Klasörünün Oluşturulması ve Yetkilendirilmesi

Hadoop kullanıcısına dönülerek, oluşturulan klasör içine kurulum dosyaları taşınır:

```
su hadoop
cd apache-hive-2.3.0-bin
mv * /opt/hive
cd ..
```

Şekil 3.78. Hive Klasörünün Taşınması

.bashrc dosyası açılarak:

```
nano .bashrc
```

Şekil 3.79. .bashrc Dosyasının Açılması

Hive ve Hadoop dosya yollarını tanımlamak için değişken tanımlaması yapılır:

```
export HIVE_HOME=/opt/hive
export PATH=$PATH:$HIVE_HOME/bin
export PATH=$PATH:$HADOOP_HOME/bin
```

Şekil 3.80. Değişken Tanımlamaları

Oluşturulan değişkenlerin hafızaya alınabilmesi için:

```
source ~/.bashrc
```

Şekil 3.81. Değişkenlerin Hafızaya Alınması

HDFS sisteminde verilerin tutulacağı klasör oluşturulur:

```
hadoop fs -mkdir /opt/hive/warehouse
```

Şekil 3.82. HDFS için Klasör Oluşturulması

Hive klasöründe bulunan hive-config dosyası açılarak:

```
cd /opt/hive
nano bin/hive-config.sh
```

Şekil 3.83. hive-config Dosyasının Açılması

“HIVE_AUX_JARS_PATH” yazısının bulunduğu satırdan sonra Hadoop’un kurulum dizinini gösteren değişken tanımlaması yapılır:

```
export HADOOP_HOME=/opt/hadoop-2.7.2
```

Şekil 3.84. Hadoop_Home Değişkeninin Tanımlanması

Hive çalıştırılır.

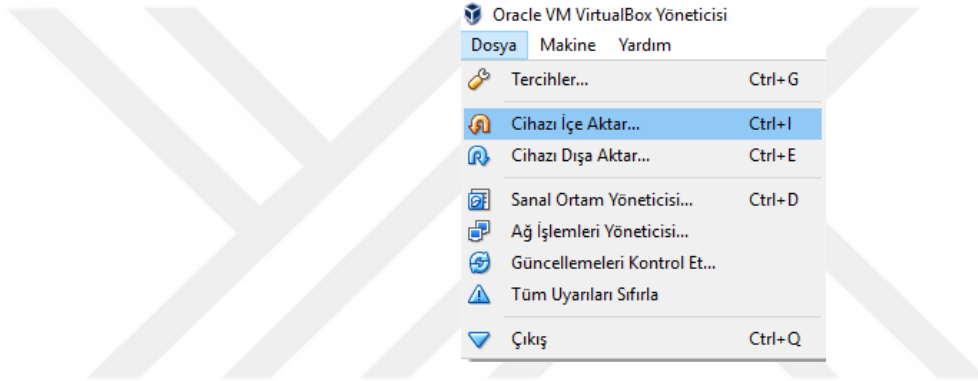
```
bin/hive
```

Şekil 3.85. Hive Komutunun Çalıştırılması

3.4.3.5. Windows Ortamında Hadoop ve Hive Kurulumu

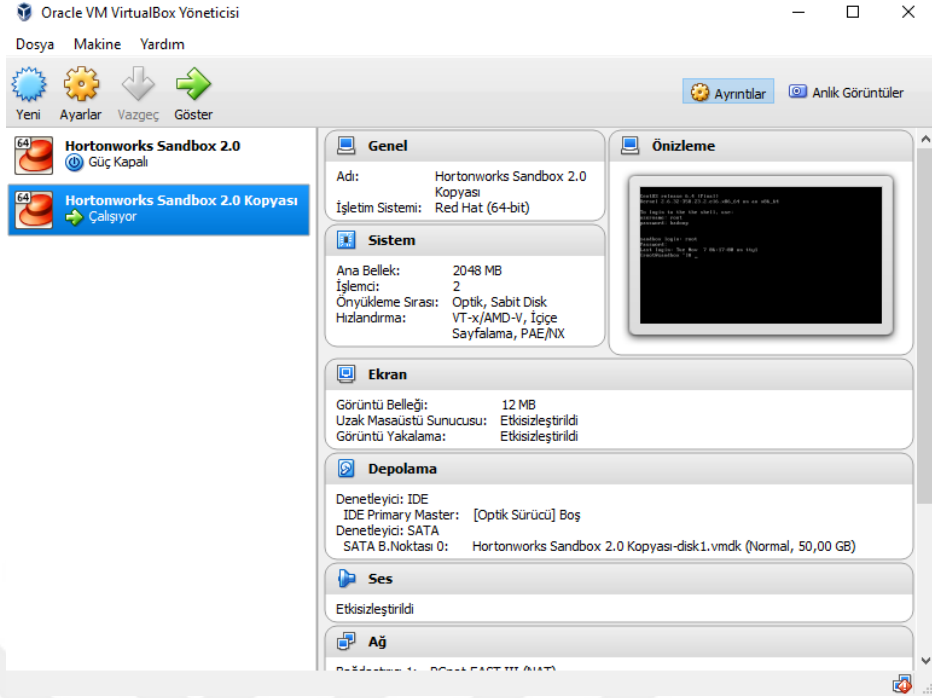
Hadoop Linux işletim sistemleri üzerinde çalışmasına rağmen istenirse Linux sanal makinesi oluşturularak veya Sandbox Hadoop ortamı sanal sürücü olarak indirilerek çalıştırılabilmektedir. Sandbox ortamı <https://hortonworks.com/downloads/#sandbox> adresinden Virtualbox programı için sandbox sürücü dosyası indirilebilmektedir. Sunucuda ayrıca VMware ve Docker programları için de sürücüler bulunmaktadır.

<https://www.virtualbox.org/> sitesinden Oracle VM Virtualbox indirilerek kurulum yapıldı. Program çalıştırıldıktan sonra “Dosya” menüsü altından “Cihazı İçer Aktar” bağlantısı tıklanarak sandbox dosyası programa dâhil edildi.



Şekil 3.86. Cihazı İçeri Aktar Bağlantısı

Sanal makine için “Ayarlar” düğmesine tıklanarak gerekli düzenlemeler yapıldıktan sonra makine başlatıldığında, programın ekran görüntüsü aşağıdaki gibidir:



Şekil 3.87. Oracle VM Virtualbox Yöneticisi Ekranı

Çalışan makinenin ekran görüntüleri ise Şekil 3.88’deki gibidir:

```
Starting the VirtualBox Guest Additions [ OK ]
Starting VirtualBox Guest Addition service [ OK ]
Starting snmpd: [ OK ]
Starting sshd: [ OK ]
Starting postfix: [ OK ]
Starting httpd: [ OK ]
Starting startup_script...
Updating IP...
Starting HDP ...
Starting mysql [ OK ]
Starting Postgre SQL [ OK ]
Starting data node [ OK ]
Starting secondary name node [ OK ]
Starting name node [ OK ]
Starting Oozie [ OK ]
DEPRECATED: Use of this script to execute hdfs command is deprecated.
Instead use the hdfs command for it.

Starting zookeeper nodes [ WARNINGS ]
JMX enabled by default
Using config: /etc/zookeeper/conf/zoo.cfg
Safe mode is OFF
Starting hive server [ OK ]
Starting Hiveserver2 [ OK ]
```

Şekil 3.88. Sandbox 2.0 Sanal Makinesi Ekranı

Makine başlatılması tamamlandığında “Alt + F5” tuş kombinasyonu ile içine girilerek kullanıcı adı (“root”) ve şifre (“hadoop”) bilgileri ile makineye erişilmiş oldu.

Time	Name	Query	User	State	Result
10/26/17 01:43:23	[Unsaved]	select tweetId, olumluolasilik* po.i olumluolasilik, olumsuzolasilik * nfo.i olumsu...	hue	expired	~
10/26/17 01:33:03	[Unsaved]	create table IF NOT EXISTS TweetDuygu as select tweetId, case when olumluolasili...	hue	available	Results
10/26/17 01:32:40	[Unsaved]	select tweetId, case when 2 >1 then 1 else 0 end duygu from TweetOla...	hue	available	Results
10/26/17 01:25:50	[Unsaved]	create table IF NOT EXISTS TweetOlasilikSon as select tweetId, olumluolasilik* po.i ol...	hue	available	Results
10/26/17 01:10:50	[Unsaved]	DROP TABLE 'default.sehir'	hue	available	Results
10/25/17 14:34:06	[Unsaved]	create table IF NOT EXISTS TweetOlasilik as select tweetId, EXP(SUM(LOG(olumluolasilik))) oluml...	hue	available	Results
10/25/17 14:27:20	[Unsaved]	select * from NotrOlasilik	hue	available	Results
10/25/17 14:26:06	[Unsaved]	create table IF NOT EXISTS NotrOlasilik as select mts.i *1.0 / ts.i i from notrTweetSayi mts inner j...	hue	available	Results
10/25/17 14:25:20	[Unsaved]	create table IF NOT EXISTS NegatifOlasilik as select mts.i *1.0 / ts.i i from negatifTweetSayi mts...	hue	expired	~
10/25/17 14:23:43	[Unsaved]	drop view NotrTweetSayi; create table IF NOT EXISTS NotrTweetSayi as select count(distinct tweetId)...	hue	available	Results
10/25/17 14:23:18	[Unsaved]	drop view NegatifTweetSayi; create table IF NOT EXISTS NegatifTweetSayi as select count(distinct tw...	hue	available	Results
10/25/17 14:22:33	[Unsaved]	select * from PozitifOlasilik	hue	available	Results
10/25/17 14:19:04	[Unsaved]	drop view PozitifOlasilik; create table IF NOT EXISTS PozitifOlasilik as select count(distinct tw...	hue	available	Results

Showing 1 to 20 of 89 items, page 1 of 5

← Beginning of List Previous Page Next Page End of List →

Şekil 3.89. Sorgu Geçmiş Ekranı

“8000” portu kullanılarak tarayıcı üzerinden (localhost:8000) sanal makineye bağlanılarak Hive editörü vasıtasıyla, ilgili tablolar oluşturuldu. Şekil 3.90’da kod geçmişi gösterilmektedir. Tablo oluşturma betikleri bir sonraki bölümde anlatılacaktır.

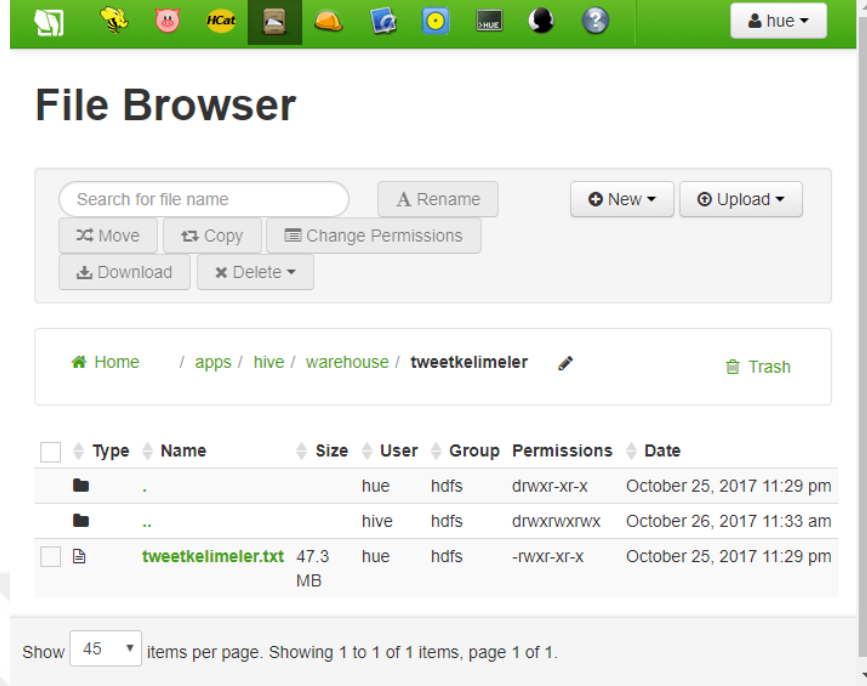
The screenshot shows the Hue File Browser interface. At the top, there is a green navigation bar with various icons and a user profile dropdown labeled 'hue'. Below this, the title 'File Browser' is displayed. The interface includes a search bar and several action buttons: 'Move', 'Copy', 'Change Permissions', 'Download', 'Delete', 'Rename', 'New', and 'Upload'. The breadcrumb navigation shows the path: Home / apps / hive / warehouse. A 'Trash' link is also visible. The main content area displays a table of files and directories in the HDFS system.

Type	Name	Size	User	Group	Permissions	Date
Folder	.		hive	hdfs	drwxrwxrwx	November 07, 2017 06:12 pm
Folder	..		hdfs	hdfs	drwxr-xr-x	October 21, 2013 01:08 am
Folder	negatifolasilik		hue	hdfs	drwxr-xr-x	October 26, 2017 12:25 am
Folder	negatiftweetsayi		hue	hdfs	drwxr-xr-x	October 26, 2017 12:23 am
Folder	notrolasilik		hue	hdfs	drwxr-xr-x	October 26, 2017 12:26 am
Folder	notrtweetsayi		hue	hdfs	drwxr-xr-x	October 26, 2017 12:24 am
Folder	pozitifolasilik		hue	hdfs	drwxr-xr-x	October 26, 2017 12:22 am
Folder	pozitiftweetsayi		hue	hdfs	drwxr-xr-x	October 26, 2017 12:20 am
Folder	sozluk		hue	hdfs	drwxr-xr-x	October 25, 2017 11:22 pm
Folder	tweetduygu		hue	hdfs	drwxr-xr-x	October 26, 2017 11:33 am
Folder	tweetkelimeler		hue	hdfs	drwxr-xr-x	October 25, 2017 11:29 pm
Folder	tweetolasilik		hue	hdfs	drwxr-xr-x	October 26, 2017 12:34 am
Folder	tweetolasilikson		hue	hdfs	drwxr-xr-x	October 26, 2017 11:26 am
Folder	tweets_sentiment		hue	hdfs	drwxr-xr-x	October 12, 2017 06:08 pm
Folder	tweets_text		hue	hdfs	drwxr-xr-x	October 12, 2017 06:02 pm
Folder	tweetsayi		hue	hdfs	drwxr-xr-x	October 26, 2017 12:21 am

At the bottom, there is a pagination bar showing 'Show 45 items per page. Showing 1 to 14 of 14 items, page 1 of 1.'

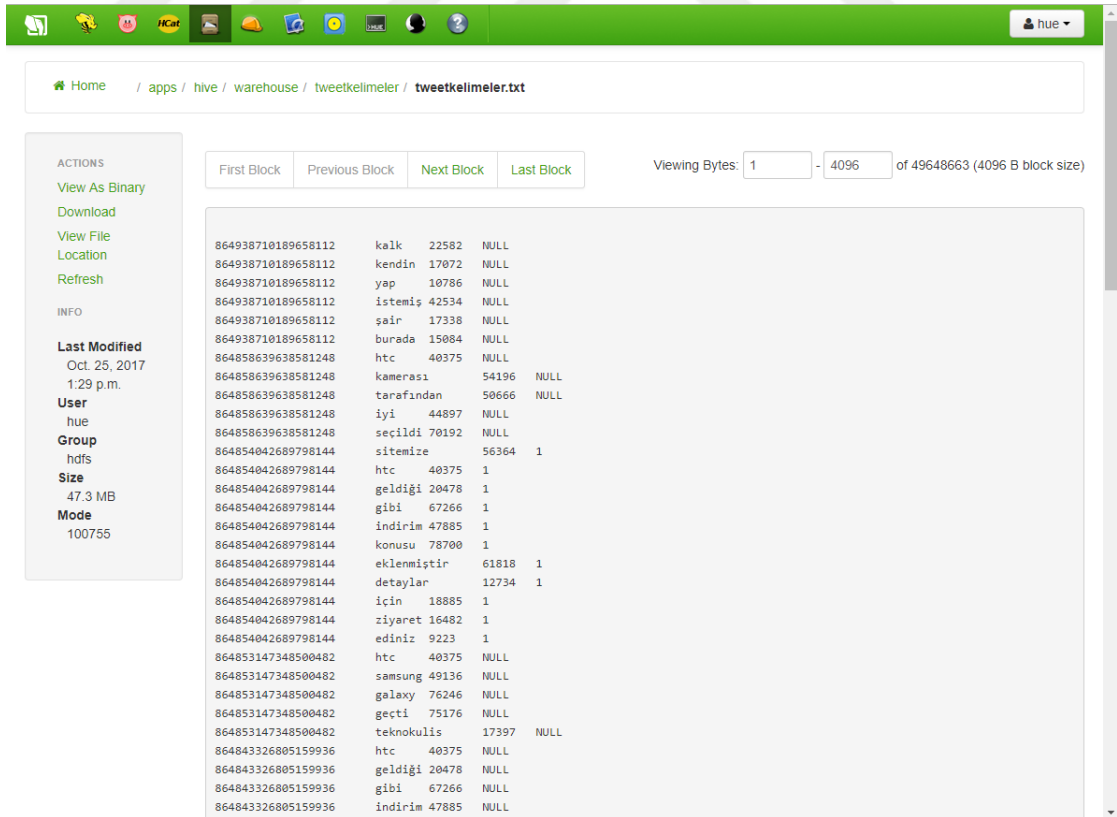
Şekil 3.90. Hive ile Oluşturulan Tabloların HDFS Klasörleri

Dışarıdan veri yüklemesi yapılacak “tweetkelimeler” ve “sozluk” tabloları için tablonun kayıtlı olduğu konuma erişilerek “upload” bağlantısı üzerinden yerel dosya HDFS sistemine yüklendi.



Şekil 3.91. “Tweetkelimeler” Tablosu için Dosya Konumu

“Tweetkelimeler.txt” dosyasının içeriği aşağıdaki gibidir:



Şekil 3.92. “Tweetkelimeler” Dosyası İçeriği

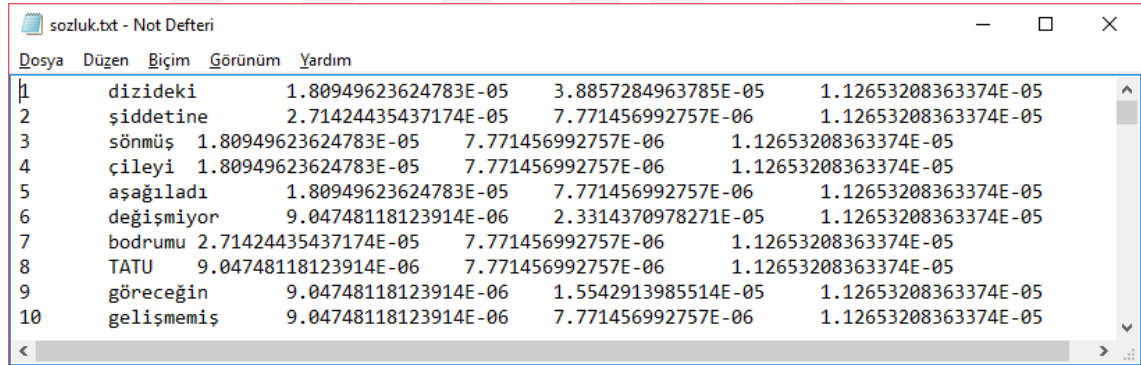
3.4.3.6. HiveQL ile Naive Bayes Sınıflandırma

Veritabanına verileri yüklemek için “tweetkelimeler” ve “sozluk” tabloları oluşturuldu (Şekil 3.93).

```
CREATE TABLE IF NOT EXISTS tweetkelimeler ( tweetid int, kelime String, kelimeId int)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY '\t'
STORED AS TEXTFILE;

CREATE TABLE IF NOT EXISTS sozluk ( id int, kelime String, olumluOlasilik float,
olumsuzOlasilik float, notrOlasilik float)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY '\t'
STORED AS TEXTFILE
```

Şekil 3.93. “tweetkelimeler” ve “sozluk” Tablolarının Oluşturulması (HiveQL kodları)



	olumlu	olumsuz	notr	id
1 dizideki	1.80949623624783E-05	3.8857284963785E-05	1.12653208363374E-05	1
2 şiddetine	2.71424435437174E-05	7.771456992757E-06	1.12653208363374E-05	2
3 sönmüş	1.80949623624783E-05	7.771456992757E-06	1.12653208363374E-05	3
4 çileyi	1.80949623624783E-05	7.771456992757E-06	1.12653208363374E-05	4
5 aşağıladı	1.80949623624783E-05	7.771456992757E-06	1.12653208363374E-05	5
6 değişmiyor	9.04748118123914E-06	2.3314370978271E-05	1.12653208363374E-05	6
7 bodrumu	2.71424435437174E-05	7.771456992757E-06	1.12653208363374E-05	7
8 TATU	9.04748118123914E-06	7.771456992757E-06	1.12653208363374E-05	8
9 göreceğin	9.04748118123914E-06	1.5542913985514E-05	1.12653208363374E-05	9
10 gelişmemiş	9.04748118123914E-06	7.771456992757E-06	1.12653208363374E-05	10

Şekil 3.94. Sözlük Tablosu için Veri Dosyası İçeriği

Aşağıdaki tablo tanımlamaları ve komutlar ise duygu analizi için oluşturulmuştur. Öncelikle “TweetView” görünümü oluşturuldu.

```

create view IF NOT EXISTS tweetView as
select
    TweetId,
    KelimeId,
    OlumluOlasilik,
    OlumsuzOlasilik,
    NotrOlasilik
from tweetKelimeler tk inner join sozluk s on s.Id =tk.kelimeId;

```

Şekil 3.95. TweetView Görünümü Oluşturulması

Tweet sayılarıyla ilgili değişken tanımlanamadığı için tablolar oluşturuldu. Tweetlerin sayıları (TweetSayı), pozitif tweetlerin sayısı (PozitifTweetSayı), negatif tweetlerin sayısı (NegatifTweetSayı) ve nötr tweetlerin sayısı (NotrTweetSayı) değerleri için birer tablo oluşturuldu (Şekil 3.96).

```

create view IF NOT EXISTS TweetSayı as select count(distinct tweetId) i from tweetView
create view IF NOT EXISTS PozitifTweetSayı as select count(distinct tweetId) i from
tweetKelimeler where tweetDuygu=1
create view IF NOT EXISTS NegatifTweetSayı as select count(distinct tweetId) i from
tweetKelimeler where tweetDuygu=-1
create view IF NOT EXISTS NotrTweetSayı as select count(distinct tweetId) i from
tweetKelimeler where tweetDuygu=0

```

Şekil 3.96. Tweet Sayılarıyla İlgili Tabloların Oluşturulması

Hesaplama kullanılmak üzere pozitif olasılık, negatif olasılık ve nötr olasılık değerleri hesaplanarak ilgili tablolar oluşturuldu (Şekil 3.97).

```

create table IF NOT EXISTS PozitifOlasilik as select pts.i *1.0 / ts.i i from pozitifTweetSayı pts
inner join tweetsayı ts on true
create table IF NOT EXISTS NegatifOlasilik as select nts.i *1.0 / ts.i i from negatifTweetSayı
nts inner join tweetsayı ts on true
create table IF NOT EXISTS NotrOlasilik as select nts.i *1.0 / ts.i i from notrTweetSayı nts
inner join tweetsayı ts on true

```

Şekil 3.97. Olasılık Tablolarının Oluşturulması

Tweetlerde kullanılan farklı kelime sayısı (KelimeSayı), pozitif tweetlerde geçen (PozitifteKelimeSayı), negatif tweetlerde geçen (NegatifteKelimeSayı) ve nötr tweetlerde geçen (NötrdeKelimeSayı) kelime sayıları için birer tablo oluşturuldu (Şekil 3.98).

```
create view IF NOT EXISTS KelimeSayı as select count(distinct KelimeId) i from
tweetKelimeler

create view IF NOT EXISTS PozitifteKelimeSayı as select count(distinct KelimeId) i from
tweetKelimeler where tweetDuygu=1

create view IF NOT EXISTS NegatifteKelimeSayı as select count(distinct KelimeId) i from
tweetKelimeler where tweetDuygu=-1

create view IF NOT EXISTS NotrdeKelimeSayı as select count(distinct KelimeId) from
tweetKelimeler where tweetDuygu=0
```

Şekil 3.98. Kelme Sayılarıyla İlgili Tabloların Oluşturulması

Tweetlerin sınıflarda bulunma olasılıkları hesaplanarak bir tabloya atılır (TweetOlasilik). Bunun için tweette geçen kelimelerin sınıf olasılık değerleri çarpılır (Şekil 3.99).

```
create table IF NOT EXISTS TweetOlasilik as
select tweetId,
      EXP(SUM(LOG(olumluolasilik))) olumluolasilik ,
      EXP(SUM(LOG(olumsuzolasilik))) olumsuzolasilik,
      EXP(SUM(LOG(NotrOlasilik))) NotrOlasilik
      from tweetview
      group by tweetId
```

Şekil 3.99. “TweetOlasilik” tablosunun oluşturulması

Tweetlerin sınıflarda bulunma olasılıkları, herbir sınıfın olasılık değeri ile çarpılarak nihai değer hesaplanır ve “TweetOlasilikSon” tablosu oluşturulur (Şekil 3.100).

```

create table IF NOT EXISTS TweetOlasilikSon as
select
    tweetId,
    olumluolasilik* po.i olumluolasilik,
    olumsuzolasilik * nfo.i olumsuzolasilik,
    NotrOlasilik* nto.i NotrOlasilik
from tweetview
    inner join pozitifOlasilik po on true
    inner join Negatifolasilik nfo on true
    inner join NotrOlasilik nto on true

```

Şekil 3.100. “TweetOlasilikSon” Tablosunun Oluşturulması

Son olarak tweetlerin sınıf olasılık değerleri mukayese edilerek tweetin hangi sınıfa ait olduğu belirlenir. “TweetDuygu” tablosunda tweetin ID’si ve tahmin edilen duygusu oluşturulmaktadır (Şekil 3.101).

```

create table IF NOT EXISTS TweetDuygu as
select
    tweetId,
    case
        when olumluOlasilik>OlumsuzOlasilik and olumluOlasilik>NotrOlasilik then 1
        when OlumsuzOlasilik>olumluOlasilik and OlumsuzOlasilik>NotrOlasilik then -1
        else 0 end duygu
from TweetOlasilikSon

```

Şekil 3.101. “TweetDuygu” Tablosunun Oluşturulması

3.4.4. Sonuçların Raporlanması

Rapor ekranları marifetiyle belirli bir marka hakkında paylaşılan tweetler için genel bilgi edinilebilmektedir. Ayrıca belirli özelliklere sahip tweetler listelenebilmektedir.

Visual Studio.Net ortamında C# dili kullanılarak geliştirilen masaüstü (desktop) uygulaması sayesinde raporlar, günlük, haftalık, aylık ve 6 aylık periyotlar şeklinde alınabilmektedir.



Şekil 3.102. Tweet Analiz Modülü Ana Ekranı

Bu konu başlığı altında yukarıdaki ekranda bulunan istatistik ve rapor ekranları incelenecektir. Ayarlar sekmesinde bulunan ekranlara ise konfigürasyon ekranları bölümünde değinilecektir.

3.4.4.1. Marka Karşılaştırmalı Tweetler

Birden fazla marka adının birlikte geçtiği tweetlerin analiz edilmesi zor olduğundan ve bu tweetlerin kıyaslama içerebilmesi dolayısıyla görüntülenmesi önem arz etmektedir.

Bu bağlamda birden fazla marka adının birlikte geçtiği tweetlerin listelenebildiği bir ekran tasarlandı. Tweetlerin dönemlik olarak listelenebildiği bu ekran marifetiyle içinde geçen marka adları seçilerek hangi konumdan atıldığı (il), kaç defa okunduğu (listed count) ve kullanıcı takipçi sayılarıyla (followers count) birlikte ilgili tweetlerin görüntülenmesi sağlanabilmiş oldu.

The screenshot shows the 'Karşılaştırmalı' (Comparative) screen. It has a sidebar on the left with filters for 'Süre Aralığı' (Time Range) set to 'Son 6 Ay' and 'Marka Adı' (Brand Name) set to 'APPLE'. Below the sidebar is a list of tweets. The main area displays a table with the following columns: id, tweet, il, ListedCount, and FollowersCount.

id	tweet	il	ListedCount	FollowersCount
34583	Samsung fiyatla skip atı milletçi götürü satma alanam sorry tasamı özürlü lg desen sorunlu apple telefonlar yatırım aracı haline geldi		5	679
34647	RT @TurkGamerNet: Apple ve Samsung'un arkasından dünyanın 3. büyük akıllı telefon üreticisi olan Huawei, net kârının 5.3 milyar \$'a yüks...		25	678
34982	RT @cemalkahveci7: Samsung klavye gören zengin Apple sahibi https://t.co/NUNQWjZNR	RİZE	1	780
35004	Samsung klavye gören zengin Apple sahibi https://t.co/NUNQWjZNR	SAKARYA	0	142
35054	RT @fikret_karakaya: Bizler fiyat konusunda Apple'a mana bulurken, Samsung bu konuda maalesef bizi başkına geviridi diyebilirim ! @Sams...	İZMİR	1	2009
35090	.@SamsungTürkiye Neyse seviyorsanız samsung, apple'ın Abv		3	344
35093	Apple olma kendin ol @SamsungMobile	ANKARA	0	862
35248	RT @fikret_karakaya: Bizler fiyat konusunda Apple'a mana bulurken, Samsung bu konuda maalesef bizi başkına geviridi diyebilirim ! @Sams...		0	341
35309	RT @fikret_karakaya: Bizler fiyat konusunda Apple'a mana bulurken, Samsung bu konuda maalesef bizi başkına geviridi diyebilirim ! @Sams...		0	5236
35311	Turkcell'lerin 4.5G ile data kullanım 1 yılda 574 Petabayt'a ulaştı https://t.co/qA5e7wLQgR #teknoloji #samsung #apple #ILG	İSTANBUL	11	8885

Şekil 3.103. Karşılaştırmalı Tweetler Ekranı

3.4.4.2. En Çok Okunan Tweetler

id	tarih	tweet	il	ListedCount	FollowersCount
342941	14.03.2017 12:01	RT @NTVBilimveTek: Samsung Galaxy S8 Plus'in bataryası sızdı https://t.co/WJu9elQwwe https://t.co/EXhZULY9Zq		5116	6663472
345075	16.03.2017 17:20	RT @cnrturktekno: Samsung Galaxy S8'in S7'den dokuz farkı https://t.co/KEXQHRt3 https://t.co/vtCLp2NPOS		4242	4480389
341362	16.03.2017 01:58	RT @cnrturktekno: Samsung artık pamak izi okuyucusu kullanmayacak https://t.co/f91yT6QAD1 https://t.co/ZPmX9mq8Au		4236	4477297
341293	16.03.2017 08:15	#Son #Dakika #Samsung #Haberleri 16 #Mart 2017 00:15 #Galaxy s8'e #Iphone islevi geliyor ! samsung 'un yeni amiral https://t.co/...		3846	1297
344834	17.03.2017 08:05	RT @HumiyetTekno: #Samsung'un simsiyah model Galaxy S8'i ilk kez görüntüledi! Bakın tasarımı nasıl görünüyor? https://t.co/Z1M...	ISTANBUL	3842	4071192
341297	16.03.2017 08:07	RT @HumiyetTekno: #Samsung'un beklenen son oyuncu! #GalaxyS8'e iPhone'larda kullanılan çok önemli bir özellik geliyor! https://t.co/...	ISTANBUL	3838	4070054
345157	16.03.2017 15:15	RT @ekonomiHTcom: Yeni iPhone ve Samsung GalaxyS8'e dair son sızıntılar! https://t.co/HdyORgXTc	ISTANBUL	3344	3558076
342879	14.03.2017 14:11	Samsung Galaxy S8'in arayüzü internete sızdı https://t.co/5RmbW44GyG https://t.co/vwt6qMnpV3		2625	2356269
343341	13.03.2017 21:30	Samsung'a ödül yağdı https://t.co/h2q1rHuUXh https://t.co/4gNCL0h12E		2225	1431727
344027	12.03.2017 12:05	RT @hadjordanbecom: "Samsung Geleceğin Mucitlerini Arıyor!" Haberin Tamamı için #samsung, diğer https://t.co/R1NPSeW9ERht...		2167	2560

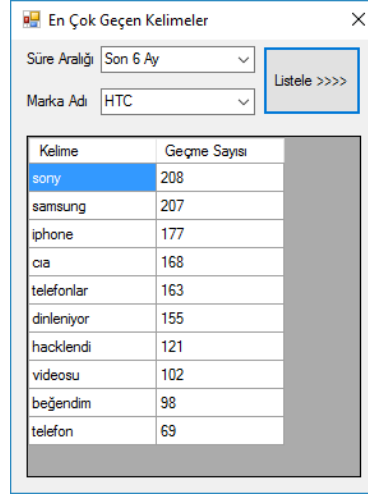
Şekil 3.104. En Çok Okunan Tweetler Ekranı

Çok sayıda kullanıcı tarafından okunan tweetlerin kamuoyu oluşturması bakımından etkili olduğu bilinmektedir. Kullanıcı takipçi sayısı (followers count) ile tweet okunma sayısı (listed count) arasında doğrusal bir ilişki bulunabilmişse de hata yüzdelere bakıldığında, takipçi sayısı düşük olanların da okunma sayılarının yüksek olabildiği görülmektedir.

Bu bakımdan okunma sayılarına göre sıralanarak en üstte kalan 100 tweetin sistem kullanıcısına görüntülenmesinin faydalı olacağı düşüncesinden hareketle “En Çok Okunan Tweetler” ekranı tasarlandı.

3.4.4.3. Sık Kullanılan Kelimeler

Herhangi bir marka hakkında paylaşılan tweetler içinde sıklıkla kullanılan kelimeler, genel kanaatin belirlenebilmesi adına önemlidir. Bu düşünceden hareketle periyodik olarak sık kullanılan kelimelerin görüntülenebildiği bir ekran tasarlanması faydalı görüldü.

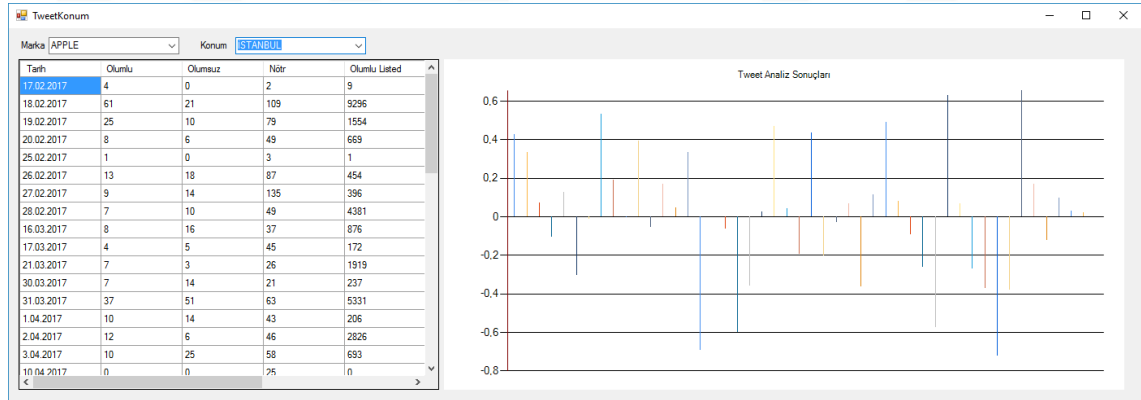


Kelime	Geçme Sayısı
sony	208
samsung	207
iphone	177
cia	168
telefonlar	163
dinleniyor	155
hacklendi	121
videosu	102
beğendim	98
telefon	69

Şekil 3.105. Sık Kullanılan Kelimeler Ekranı

3.4.4.4. Tweet Analiz Sonuçlarının İllere Göre Raporlanması

Bu ekranda tweet verileri çekilen herhangi bir marka için Türkiye’deki herhangi bir ilden paylaşılan tweetlerin analiz sonuçları tablo halinde ve grafiksel olarak görüntülenmektedir. Örnekte “Apple” firması için “İstanbul” vilayetinden paylaşılan tweetler ile ilgili sonuçlar görüntülenmektedir.



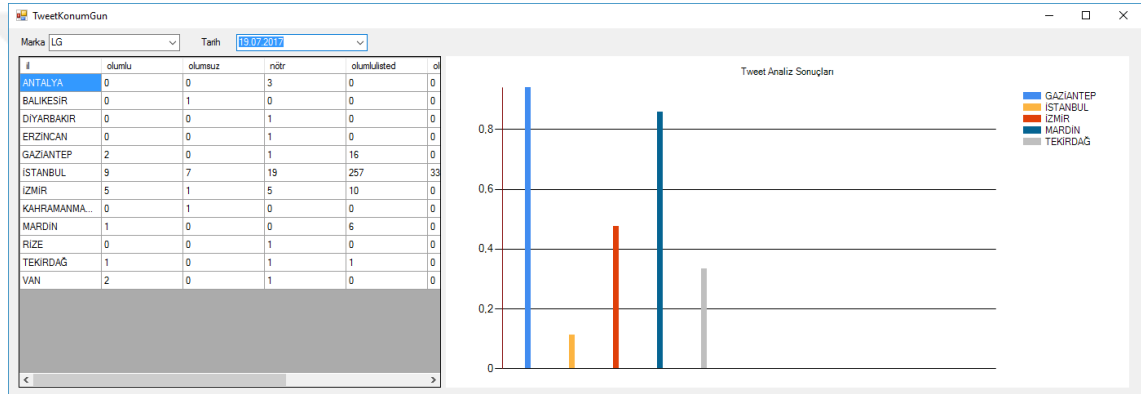
Şekil 3.106. Tweet – Konum Ekranı

Tweetler ile ilgili analiz yapılırken olumlu tweet sayısı (pc) ile olumlu tweetlerin görüntülenme sayıları (plc) toplanarak olumsuz tweet sayısı (nc) ile görüntülenme sayıları (nlc) çıkarılmaktadır. Çıkan sonuç tüm tweet sayısı ile görüntülenme sayılarına bölünmektedir. Toplam tweet sayısı içinde nötr tweet sayıları (ntrc) ve nötr tweetlerin görüntülenme sayıları (ntrlc) da bulunmaktadır.

$$f = \frac{pc + plc - (nc + nlc)}{pc + plc + nc + nlc + ntrc + ntrlc}$$

3.4.4.5. Tweet Analiz Sonuçlarının İllere ve Tarihe Göre Raporlanması

Bu ekranda tweet verileri çekilen herhangi bir marka için ilgili tarihte paylaşılan tweetlerin analiz sonuçları tablo halinde ve grafiksel olarak görüntülenmektedir. Örnekte “LG” firması için “19.07.2017” tarihi itibarıyla paylaşılan tweetler ile ilgili sonuçlar görüntülenmektedir. Gaziantep, İstanbul, izmir, Mardin ve Tekirdağ ile ilgili analiz sonuçları grafikte görüntülenmektedir. Diğer iller için bu değer 0 olduğundan grafikte yer almamıştır.



Şekil 3.107. Tweet – Konum Ekranı (Belirli Bir Tarih için)

3.4.5. Konfigürasyon Ekranları

3.4.5.1. Duygu Belirleyici İfadeler

Duygu belirleyiciler tablosuna (1766 adet duygu belirleyici ifade) yeni ifadelerin de eklenebildiği bu ekran marifetiyle tweetlerin analiz edilmesi sürecine kullanıcı da katkı sağlayabilmektedir. Herhangi bir ifadenin geçtiği tweetler olumlu olarsa bu ifade olumlu sınıfına eklenebilmektedir. Duygu puanı olumlu olmasına rağmen tweeti olumlu sınıfına geçirme gücü olmayan ifadeler bu tabloda bulunmamalıdır.

Duygu Belirleyiciler

İfade: Ekle

Duygu: Olumlu

id	İfade	Duygu
285	1 nisan şakası	Olumsuz
169	10 numara	Olumlu
158	acı kayımız	Olumlu
147	acil şifalar	Olumlu
46	aç köpek	Olumsuz
196	açık sağık	Olumsuz
20	adam değil	Olumsuz
159	adamsın	Olumlu
37	ağızma sıç	Olumsuz
28	ağızma sıçayım	Olumsuz
8	akıldan uzak	Olumsuz
60	alet etme	Olumsuz

Şekil 3.108. Duygu Belirleyici İfadeler Ekranı

3.4.5.2. Kelime Veri Tabanı

Tweetlerde geçen ancak sözlükte karşılığı bulunmayan ifadeler Türkçe hece uyumu bakımından incelendikten sonra bu ekrana getirilmektedir.

Bu ekran vasıtasıyla kullanıcı ön işlem aşamasının geliştirilmesine katkıda bulunmuş olacaktır. Kullanıcı herhangi bir kelimenin olması gerektiği halini düzenledikten sonra kelime veri tabanına ekleyebilmektedir. Böylece aynı yanlış ifade başka tweetlerde geçtiğinde sistem otomatik olarak o kelimeyi düzelterek analize tabi tutacaktır.

Kelime Veri Tabanı

Tweetlerde geçen ifadeler

- abey
- abhey
- abineligiö
- abıyeva
- abibi
- abicio
- abideliği
- abidik
- abiğ
- abiiçm
- abiiğ
- abiiim
- abiiim
- abiiim
- abij
- abilene
- abimme
- abimmin
- abimml

Düzenlenmiş Hali:

abimin

Ekle >>>

<<< Çıkar

Kelime Veri Tabanı

- ab
- aba
- aba güreşi
- abaca
- abacısı
- abacı
- abacıdan
- abacılar
- abacıların
- abacılarla
- abacılık
- abacının
- abacıya
- abacıyı
- abacıyla
- ABAD
- abada
- abadan
- abadın

Şekil 3.109. Kelime Veri Tabanı Ekranı

DÖRDÜNCÜ BÖLÜM

BULGULAR

4.1. GENEL İSTATİSTİKLER

Veri kümesi içinde 312632 tweet bulunmaktadır. Tweetler içinde ise toplam:

- 149515 adet link
- 71677 adet mention
- 140990 adet hashtag
- 72046 adet noktalama işareti

1161 adet noktalama işareti bulunan hatalı yazılmış kelimenin 413'ü başındaki ve sonundaki noktalama işaretlerinden arındırılarak düzeltilebildi.

4.1.1. Ek ve Kökler ile İlgili İstatistikler

Tweetler içinde kullanılan kelimeler zemberek kütüphanesi marifetiyle kök ve eklerine ayrıştırıldığında isim köklerinin fiil köklerinden daha fazla kullanıldığı gözlemlenmiştir. Diğer kök türleri ve kullanım adetleri ise aşağıdaki tabloda sunulmuştur.

Tablo 4.1. Kökler ile İlgili İstatistikler

Kök Türü	Adedi	Yüzde
ISIM_KOK	47234	55
FIIL_KOK	35760	42
OZEL_KOK	728	1
SAYI_KOK	651	1
ZAMIR_KOK	617	1
ZAMAN_KOK	596	1
SORU_KOK	67	0
YANKI_KOK	53	0
EDAT_KOK	29	0
IMEK_KOK	17	0
BAGLAC_KOK	5	0
UNLEM_KOK	2	0

Yukarıda istatistiği verilen kök türleri içinde en çok kullanılan 10 kelime kökü ise aşağıdaki tabloda listelenmiştir. Sektörde en sık kullanılan kelimeler arasında “almak”, “yapmak”, “olmak”, “etmek” ve “kullanmak” kelimeleri geçmektedir. İlk on kelime kökü ise fiil kökü türündedir.

En çok kullanılan kelime kökleri ise aşağıdaki tabloda görüntülenmektedir.

Tablo 4.2. En Çok Kullanılan Kelime Kökleri

<i>Kelime Kökü</i>	<i>Adedi</i>
<i>al</i>	3311
<i>yap</i>	3122
<i>ol</i>	2841
<i>et</i>	2042
<i>kullan</i>	1987
<i>de</i>	1891
<i>anla</i>	1849
<i>gör</i>	1828
<i>gel</i>	1792
<i>çek</i>	1592

En sık kullanılan kelimeler ve tweetler içinde kullanım sıklığı ise aşağıdaki tabloda gösterilmiştir. Telefon markalarıyla ilgili paylaşılan yorumlar içinde en sık kullanılan ilk 20 kelime içinde “telefon” kelimesinin geçmesi ise anlamlıdır.

Tablo 4.3. Kelime Kullanım İstatistikleri

<i>Kelime</i>	<i>Adedi</i>
<i>ve</i>	20699
<i>bir</i>	18160
<i>yeni</i>	12529
<i>için</i>	10793
<i>bu</i>	10764
<i>de</i>	8605
<i>ile</i>	8314
<i>ne</i>	8039
<i>da</i>	7266
<i>mi</i>	7006
<i>çok</i>	5981
<i>var</i>	5782
<i>en</i>	5723
<i>ama</i>	5389
<i>telefon</i>	5289

Kelimelere gelen yapım ve çekim ekleri ve kullanım sıklıkları ise aşağıdaki tabloda listelenmiştir. Kullanım sıklığına göre sıralandığında en üstte isim çoğul eki gelmektedir. Onu tamlama ve belirtme ekleri takip etmektedir. Fiil olumsuzluk ekinin de sıralamada 4. olması dikkate şayandır.

Tablo 4.4. Ekler ile İlgili İstatistikler

<i>Ek Türü</i>	<i>Adedi</i>
ISIM_COGUL_LER	10660
ISIM_TAMLAMA_IN	9185
ISIM_BELIRTME_I	7546
FIIL_OLUMSUZLUK_ME	7337
ISIM_TAMLAMA_I	6498
ISIM_YONELME_E	6404
ISIM_KALMA_DE	5769
ISIM_DONUSUM_LE	4840
FIIL_SIMDIKIZAMAN_IYOR	4737
ISIM_SAHİPLİK_BEN_IM	4361
ISIM_SAHİPLİK_O_I	4351
FIIL_GENİSZAMAN_IR	4312
FIIL_DONUSUM_ME	4283
FIIL_GECMİSZAMAN_DI	4253
ISIM_CIKMA_DEN	3540
FIIL_EDILGEN_IL	3160
FIIL_KISI_BEN	2977
FIIL_ETTIRGEN_TIR	2925
IMEK_HIKAYE_DI	2708
ISIM_BULUNMA_LI	2673
FIIL_EDILGENSESİLI_N	2589
FIIL_BELIRTME_DIK	2519
FIIL_DONUSUM_EN	2340
ISIM_SAHİPLİK_SEN_IN	2243
FIIL_BERABERLIK_IS	2231
ISIM_SAHİPLİK_BİZ_IMİZ	2033
ISIM_BULUNMA_LIK	1998
FIIL_KISI_BİZ	1919
FIIL_YETERSİZLİK_E	1908
ISIM_BIRLIKTELIK_LE	1885
ISIM_KISI_BİZ_İZ	1660
FIIL_DONUSUM_ECEK	1650
FIIL_GECMİSZAMAN_MIS	1617
FIIL_OLDURGAN_T	1583
FIIL_YETENEK_EBİL	1553

Tablo 4.4. (Devamı)

<i>Ek Türü</i>	<i>Adedi</i>
<i>FIIL_DONUSUM_IM</i>	1490
<i>ISIM_TANIMLAMA_DIR</i>	1398
<i>ISIM_KISI_BEN_IM</i>	1350
<i>IMEK_SART_SE</i>	1332
<i>FIIL_MASTAR_MEK</i>	1291
<i>ISIM_SAHİPLİK_SİZ_INİZ</i>	1286
<i>ISIM_ILGI_CI</i>	1277
<i>FIIL_KISI_ONLAR</i>	1251
<i>FIIL_KISI_SEN</i>	1228
<i>FIIL_KISI_SİZ</i>	1094
<i>FIIL_İSTEK_E</i>	1065
<i>FIIL_SART_SE</i>	1044
<i>ISIM_KISI_SEN_SİN</i>	1003
<i>FIIL_GELECEKZAMAN_ECEK</i>	891
<i>FIIL_DONUSUM_MIS</i>	868
<i>ISIM_BULUNMA_Kİ</i>	826
<i>FIIL_EMİR_SİZ_IN</i>	811
<i>ISIM_YOKLUK_SİZ</i>	774
<i>IMEK_RIVAYET_MIS</i>	732
<i>FIIL_IMSİ_IP</i>	701
<i>FIIL_DONUSUM_İS</i>	650
<i>FIIL_EMİR_O_SİN</i>	629
<i>ISIM_KISI_SİZ_SİNİZ</i>	607
<i>ISIM_DURUM_LİK</i>	561
<i>FIIL_TANIMLAMA_İCİ</i>	533
<i>IMEK_ZAMAN_KEN</i>	491
<i>FIIL_DONUSUM_İK</i>	477
<i>FIIL_SUREKLİLİK_EREK</i>	316
<i>ISIM_KISI_ONLAR_LER</i>	294
<i>FIIL_ZAMAN_INCE</i>	263
<i>ISIM_GİBİ_CE</i>	257
<i>FIIL_OLUMSUZLUK_DEN</i>	181
<i>FIIL_DONUSUM_ESİ</i>	153
<i>FIIL_DONUSUM_MEZ</i>	141
<i>ISIM_ILISKILI_SEL</i>	137
<i>FIIL_DONUSUM_INTİ</i>	128
<i>ISIM_DONUSUM_LES</i>	126
<i>FIIL_EMİR_ONLAR_SİNLER</i>	123
<i>ISIM_ANDIRMA_Sİ</i>	123
<i>ISIM_SAHİPLİK_ONLAR_LERİ</i>	105
<i>ISIM_TARAFINDAN_CE</i>	100
<i>FIIL_EMİR_SİZRESMİ_INİZ</i>	99

Tablo 4.4. (Devamı)

<i>Ek Türü</i>	<i>Adedi</i>
ISIM_KUCULTME_CIK	95
FIIL_DEVAMLILIK_DIKCE	92
FIIL_MASTAR_CE	87
FIIL_BERI_ELI	79
FIIL_ETTIRGEN_TEKRAR_T	52
SAYI_SIRA_INCI	51
FIIL_ISTEK_SENE	49
ZAMAN_BELIRTME_KI	43
FIIL_TEZLIK_IVER	29
FIIL_ISTEK_SENIZE	27
ISIM_ANDIRMA_IMSI	25
ZAMIR_SAHİPLİK_IM	24
SAYI_TOPLULUK_IZ	19
SAYI_KESIR_DE	16
YANKI_DONUSUM_DA	15
SAYI_ULESTIRME_ER	12
ISIM_KUCULTME_CEGIZ	11
YANKI_DONUSUM_TI	10
FIIL_SURERLIK_EKAL	8
FIIL_OLUMSUZLUK_SIZIN	8
SAYI_KOSE_GEN	7
FIIL_GIBI_CESINE	4
FIIL_DONUSUM_ESICE	1
FIIL_SURERLIK_EDUR	1
FIIL_SURERLIK_EGEL	1

4.1.2. Takipçi Sayılarına Göre Tweet İstatistikleri

Toplanan tweetler incelendiğinde 140119 kullanıcı tarafından paylaşılan 312632 tweet içinden takipçi sayısı 5 milyonun üzerinde olan 5 kullanıcı olduğu görülmüştür. 1000’den fazla takipçisi olan kullanıcı sayısı ise 30 bine yaklaşmaktadır. 400 ve üzeri takipçisi olan kullanıcı sayısı 150 binin üzerindeyken hiç takipçisi olmayan kullanıcı sayısı ise 2 bine yakındır.

Aşağıdaki tabloda mevcut tweetler incelenerek takipçi sayılarına göre kullanıcı sayıları ve bu kullanıcılar tarafından paylaşılan tweet sayıları yer almaktadır.

Tablo 4.5. Takipçi Sayılarına Göre Kullanıcı İstatistikleri

<i>Takipçi Sayısı</i>	<i>Kullanıcı Sayısı</i>	<i>Atılan Tweet Sayısı (Toplam)</i>	<i>Atılan Tweet Sayısı (Kişi Başı)</i>
<i>5000000 üzeri</i>	5	79	15,8
<i>1000000 üzeri</i>	53	717	13,5
<i>100000 üzeri</i>	615	5294	8,6
<i>10000 üzeri</i>	5685	30382	5,3
<i>1000 üzeri</i>	29381	105362	3,6
<i>100 üzeri</i>	96491	233235	2,4
<i>Takipçisi olmayan</i>	1234	1934	1,6

Tablo incelendiğinde takipçi sayısı ile twitter kullanımının paralel olduğu görülecektir. Hiç takipçisi olmayan kullanıcılar en az tweet atarken, takipçi sayısı yüksek olan kullanıcılar daha aktif paylaşımda bulunmaktadır. Dolayısıyla Twitter analizi için de takipçi sayısı yüksek kullanıcılar tarafından atılan tweetlerin toplam analizine önem verilebilir.

4.1.3. Okunma Sayılarına Göre Tweet İstatistikleri

Toplanan tweetler incelendiğinde 140119 kullanıcı tarafından paylaşılan 312632 tweet içinden hiç okunmayan tweet sayısı 120 bine yakın olduğu (toplam veri setinin yarısından fazlasını teşkil ettiği) bilgisi elde edilmiştir. 10 bin ve üzeri okunma sayısına sahip sadece bir tweet bulunmaktadır. Yine elde edilen istatistiklere göre en fazla okunan tweet 11058 defa okunmuştur.

Tablo 4.6: Okunma Sayılarına Göre Tweet Sayıları

<i>Okunma Sayısı</i>	<i>Tweet Sayısı</i>	<i>Yüzdelik Dilim</i>
<i>0</i>	119605	% 55
<i>10 ve üzeri</i>	78514	% 37
<i>100 ve üzeri</i>	15607	% 0,7
<i>1000 ve üzeri</i>	1261	% 0,1
<i>10000 ve üzeri</i>	1	% 0,0

Tablo 4.6 incelendiğinde 1000-10000 arası okunma sayısına sahip tweetlerin, 10-100 arası okunma sayısına sahip tweetlere göre 1/5 oranında daha az sayıda olmasına rağmen okunma sayılarıyla çarpıldığında Twitter kullanıcıları üzerinde 2 kat etki yaptığı görülmektedir. Hiç okunmayan tweetlerle kıyaslandığında ise 1000 ve üzeri okunan

tweetler yaklaşık olarak 120 kat daha etkilidir. Bu istatistikler de ortaya koymaktadır ki her tweet eşit etki ağırlığına sahip değildir. Bu istatistiklere dayanarak, sadece belirli takipçi sayısı üzerindeki kullanıcılar tarafından paylaşılan ve en az bir kere okunan tweetlerin analize sokulmasının daha anlamlı sonuçlar vereceği söylenebilir.

En çok görüntülenen veya okunan 10 tweet aşağıdaki tabloda gösterilmiştir. En çok okunan tweetler incelendiğinde çoğunun Retweet edildiği ve internet adresi içerdiği görülmektedir.

Tablo 4.7. En Çok Okunan Tweetler

<i>Okunma Sayısı</i>	<i>Tweet içeriği</i>
11058	RT @sadarbeh: Microsoft Designer Bluetooth Desktop ile iki büyük dert son buluyor :) #HayataTarzBirDokunuş https://t.co/LPNlhpY0v3 #Microso...
9592	Umi Kalsum https://t.co/eUqdbGt0cR
8598	@SonyLegacyRecs Bah dum tsssssss.
8136	RT @alphanmanas: Çinli ZTE firması dünyanın ilk 5G telefonunu çıkardı. Ürünler baz istasyonlarından önce çıkacağı benziyor. https://t.co/IC...
6665	RT @AppGaraj: Meizu'nun yeni akıllı telefon modeli: A5 https://t.co/E62xm2rAw7 https://t.co/SKsE7aJi4Z #mobile #tech #startups #news #haber...
6665	RT @AppGaraj: Asus Zenwatch için Wear 2.0 karmaşası! https://t.co/ENd1oKtGXX https://t.co/jVZiTMrkrt #mobile #tech #startups #news #haber #...
6665	RT @AppGaraj: Nokia 3310 parasına ne alınır? https://t.co/trTqQJXNTD https://t.co/yTrarJTZN4 #mobile #tech #startups #news #haber #mobil
6665	RT @AppGaraj: LG, Apple için OLED ekran üretecek! https://t.co/X8OXl5N2 https://t.co/eNWj2hAneA #mobile #tech #startups #news #haber #mob...
6665	RT @AppGaraj: Yeni Çift Kameralı Samsung Telefon! https://t.co/EdledBiHy2 https://t.co/wiwMC9WiEg #mobile #tech #startups #news #haber #mob...
6665	RT @AppGaraj: Xiaomi hızla yükselmeye devam ediyor! https://t.co/iGQlrXD43i https://t.co/bleBYeNKss #mobile #tech #startups #news #haber #m...

4.1.4. Favori Yapılma Sayılarına Göre Tweet İstatistikleri

Toplanan tweetler incelendiğinde 140119 kullanıcı tarafından paylaşılan 312632 tweet içinde en yüksek favori yapılma sayısına sahip tweet, 7476 kişi tarafından favori yapılmıştır. Hiç favori yapılmayan tweet sayısı ise 247316 (%77)'dir.

Tablo 4.8. Favori Yapılma Sayılarına Göre Tweetler

<i>Favori Yapılma Sayısı</i>	<i>Tweet Sayısı</i>	<i>Yüzde</i>
0	247316	77
1	65316	20
10	7941	2
100	712	0
1000	22	0

4.1.5. Konum Bilgilerine Göre Tweet İstatistikleri

200 Bine yakın tweetin, yani kayıtlı toplam tweetlerin yaklaşık %66'sının, lokasyon bilgisi bulunmamaktadır. Lokasyon bilgisi olan tweetler için, illere göre atılan tweet sayıları, illerin 2017 nüfus bilgileriyle (<http://tuik.gov.tr>) Tablo 4.9'da sunulmuştur.

Tablo 4.9. Konum Bilgisine Göre Tweet Sayıları

<i>İl</i>	<i>Nüfus Sayısı</i>	<i>Toplam Nüfusa Oran</i>	<i>Atılan Tweet Sayısı</i>	<i>Toplam Tweete Oran</i>
Adana	2.216.475	2,74%	1582	1%
Adıyaman	615.076	0,76%	136	0%
Afyonkarahisar	715.693	0,89%	375	0%
Ağrı	536.285	0,66%	126	0%
Aksaray	402.404	0,50%	7	0%
Amasya	329.888	0,41%	133	0%
Ankara	5.445.026	6,74%	14586	13%
Antalya	2.364.396	2,93%	2725	2%
Ardahan	97.096	0,12%	2	0%
Artvin	166.143	0,21%	104	0%
Aydın	1.080.839	1,34%	698	1%
Balıkesir	1.204.824	1,49%	891	1%
Bartın	193.577	0,24%	154	0%
Batman	585.252	0,72%	65	0%
Bayburt	80.417	0,10%	63	0%
Bilecik	221.693	0,27%	57	0%
Bingöl	273.354	0,34%	86	0%
Bitlis	341.474	0,42%	71	0%
Bolu	303.184	0,38%	159	0%
Burdur	264.779	0,33%	84	0%
Bursa	2.936.803	3,63%	2892	2%
Çanakkale	530.417	0,66%	528	0%
Çankırı	186.074	0,23%	135	0%

Tablo 4.9. (Devamı)

<i>İl</i>	<i>Nüfus Sayısı</i>	<i>Toplam Nüfusa Oran</i>	<i>Atılan Tweet Sayısı</i>	<i>Toplam Tweete Oran</i>
Çorum	528.422	0,65%	660	1%
Denizli	1.018.735	1,26%	1030	1%
Diyarbakır	1.699.901	2,10%	479	0%
Düzce	377.610	0,47%	199	0%
Edirne	406.855	0,50%	509	0%
Elazığ	583.671	0,72%	271	0%
Erzincan	231.511	0,29%	247	0%
Erzurum	760.476	0,94%	482	0%
Eskişehir	860.620	1,06%	1596	1%
Gaziantep	2.005.515	2,48%	780	1%
Giresun	437.393	0,54%	348	0%
Gümüşhane	170.173	0,21%	9	0%
Hakkâri	275.761	0,34%	36	0%
Hatay	1.575.226	1,95%	498	0%
Iğdır	194.775	0,24%	25	0%
Isparta	433.830	0,54%	273	0%
İstanbul	15.029.231	18,60%	56485	49%
İzmir	4.279.677	5,30%	9352	8%
Kahramanmaraş	1.127.623	1,40%	292	0%
Karabük	244.453	0,30%	104	0%
Karaman	246.672	0,31%	58	0%
Kars	287.654	0,36%	113	0%
Kastamonu	372.373	0,46%	202	0%
Kayseri	1.376.722	1,70%	613	1%
Kırıkkale	278.749	0,34%	1	0%
Kırklareli	356.050	0,44%	165	0%
Kırşehir	234.529	0,29%	64	0%
Kilis	136.319	0,17%	39	0%
Kocaeli	1.883.270	2,33%	2485	2%
Konya	2.180.149	2,70%	911	1%
Kütahya	572.256	0,71%	169	0%
Malatya	786.676	0,97%	281	0%
Manisa	1.413.041	1,75%	761	1%
Mardin	809.719	1,00%	128	0%
Mersin	1.793.931	2,22%	1540	1%
Muğla	938.751	1,16%	1016	1%
Muş	404.544	0,50%	28	0%

Tablo 4.9. (Devamı)

<i>İl</i>	<i>Nüfus Sayısı</i>	<i>Toplam Nüfusa Oran</i>	<i>Atılan Tweet Sayısı</i>	<i>Toplam Tweete Oran</i>
Nevşehir	292.365	0,36%	75	0%
Niğde	352.727	0,44%	110	0%
Ordu	742.341	0,92%	341	0%
Osmaniye	527.724	0,65%	413	0%
Rize	331.041	0,41%	439	0%
Sakarya	990.214	1,23%	761	1%
Samsun	1.312.990	1,62%	221	0%
Siirt	324.394	0,40%	81	0%
Sinop	207.427	0,26%	100	0%
Sivas	621.301	0,77%	241	0%
Şanlıurfa	1.985.753	2,46%	273	0%
Şırnak	503.236	0,62%	36	0%
Tekirdağ	1.005.463	1,24%	1009	1%
Tokat	602.086	0,75%	147	0%
Trabzon	786.326	0,97%	2635	2%
Tunceli	82.498	0,10%	26	0%
Uşak	364.971	0,45%	79	0%
Van	1.106.891	1,37%	285	0%
Yalova	251.203	0,31%	266	0%
Yozgat	418.650	0,52%	82	0%
Zonguldak	596.892	0,74%	181	0%

Nüfus sayısı ile atılan tweet sayısı arasında doğrusal bir ilişki olduğu (%96) görülmektedir. Nüfus yoğunluğu ve internete erişim imkânları dolayısıyla diğerlerine göre önde olan illerden daha fazla tweet atıldığı gözlemlenmiştir. İstanbul, Ankara ve İzmir illerinden atılan tweet sayısı toplam tweet sayısının %70'ini teşkil ediyor. Yaygın olarak Kürtçe konuşulan iller (Ağrı, Batman, Bingöl, Diyarbakır, Hakkâri, Mardin, Muş, Şanlıurfa, Şırnak, Tunceli ve Van) özelinde tablo incelendiğinde Twitter kullanımının ortalamadan yaklaşık olarak 7 kat daha düşük olduğu gözlenmektedir.

4.1.6. Hashtag (Etiket) İstatistikleri

Telefon markalarıyla ilgili toplanan tweetler içinde en çok kullanılan etiketlerin başında “#teknoloji” etiketi bulunmaktadır.

Tablo 4.10: Hashtag (Etiket) Kullanım İstatistikleri

<i>Etiket</i>	<i>Kullanım Sayısı</i>
#teknoloji	4877
#Samsung	1557
#Sunny	1100
#Apple	1082
#haber	1069
#turkcell	930
#Mobil	766
#lg	688
#Sony	666
#microsoft	604
#mobile	565
#tech	515
#shiftdeletenet	494
#sdncevaplıyor	465
#haberdizin	456
#istanbul	426
#xiaomi	412
#Haberler	406
#BlackBerry	403
#News	390

4.1.7. Mention (Bahis) İstatistikleri

İçinde marka adı geçen kullanıcı adı (mention) istatistiği incelendiğinde @Turkcell kullanıcı adının açık ara farkla önde olduğu görülmektedir. Diğer markalara ait diğer kullanıcı adları ve tweetler içinde geçme sayıları ise aşağıdaki tabloda görüntülenmektedir.

Tablo 4.11. Mention (Bahis) Kullanım İstatistikleri

<i>Mention</i>	<i>Kullanım Sayısı</i>
@Turkcell	7512
@igde_agaci_htc	3573
@Vestel	2369
@TurkcellHizmet	1988
@Apple	807
@SamsungTurkiye	695
@sunny_iyp	650
@lgturkiye	421
@HTC_TR	405

Tablo 4.11. (Devamı)

<i>Mention</i>	<i>Kullanım Sayısı</i>
@Sony_Turkiye	378
@Microsoft	305
@Sony	283
@SonyXperiaTR	277
@samsung	264
@LenovoTurkiye	247
@VestelDestek	236
@MicrosoftTR	171
@ASUSTR	163
@RADYOD_SUNNY	156
@SamsungDestekTR	147
@casper_grgn	146
@lgdestek	133
@lenovo	123
@htc	114
@ASUS	112
@casperturkiye	104

4.1.8. Hece İstatistikleri

Hece istatistikleri incelendiğinde en çok kullanılan hece yapısının, bir sesli harften oluşan hece yapısı olduğu; bir sesli ve iki sessiz harften oluşan hece yapısının ise diğer hece yapılarına kıyasla daha az kullanıldığı görülebilmektedir.

Tablo 4.12. Hece Yapılarına Göre Hece Sayıları

<i>Harf Sayısı</i>	<i>Hece Yapısı</i>	<i>Bulunma Sayısı</i>	<i>Yüzde</i>
1	1 Sesli harf	3405274	7%
2	1 Sesli 1 sessiz harf	2773867	6%
3	1 Sesli 2 sessiz harf	542979	1%
2	1 Sessiz 1 sesli harf	19560131	39%
3	1 Sessiz 1 sesli 1 sessiz harf	20936559	42%
4	1 Sessiz 1 sesli 2 sessiz harf	2936036	6%

Kullanım sıklığına göre ilk 20 hece aşağıdaki tabloda listelenmiştir.

Tablo 4.13. En Sık Kullanılan Heceler

<i>Hece</i>	<i>Bulunma Sayısı</i>
<i>a</i>	1209205
<i>le</i>	992991
<i>i</i>	888599
<i>ne</i>	782910
<i>de</i>	779855
<i>da</i>	737867
<i>o</i>	672888
<i>ni</i>	664921
<i>yor</i>	530531
<i>ve</i>	528808
<i>ri</i>	506791
<i>ya</i>	481352
<i>la</i>	475279
<i>mi</i>	467760
<i>li</i>	462902
<i>lar</i>	449743
<i>ma</i>	446827
<i>ye</i>	442533
<i>sung</i>	434504
<i>bir</i>	402957

Tabloda ilk sekiz hece tek sesliden oluşan ve bir sessiz bir sesli harften oluşan heceleri vermektedir. “O”, “bir” ve “ve” kelimelerinin, “de”, “da”, “yor” ve “lar” gibi çekim eklerinin “mi” soru ekinin ve “ma” olumsuzluk ekinin istatistikleri etkilediği görülmektedir. “Samsung” kelimesinde geçen “sung” hecesi ise sektörde kullanım sıklığı dolayısıyla tabloda yerini almıştır.

4.1.9. Harf İstatistikleri

Aşağıdaki tabloda harflerin çekilen tweetler içinde kullanım sıklıkları bulunmaktadır. Bu istatistik büyük ölçüde sosyal medya içinde Türkçe yazışma dilinde kullanılan harflerin kullanım istatistiğini ortaya koymaktadır. Bu harf istatistiklerinin klavye düzeni tercihinde ve sıkıştırma algoritmalarında kullanılabilecektir.

Tablo 4.14. Harf İstatistikleri

<i>Harf</i>	<i>Kullanım sayısı</i>	<i>Yüzde</i>
<i>a</i>	1117988	11%
<i>e</i>	910141	9%
<i>i</i>	820984	8%
<i>n</i>	702278	7%
<i>l</i>	640393	7%
<i>r</i>	589232	6%
<i>s</i>	418550	4%
<i>k</i>	410887	4%
<i>t</i>	401130	4%
<i>m</i>	397164	4%
<i>ı</i>	375526	4%
<i>o</i>	360914	4%
<i>d</i>	354010	4%
<i>y</i>	335723	3%
<i>u</i>	319807	3%
<i>b</i>	209849	2%
<i>g</i>	169586	2%
<i>c</i>	146549	1%
<i>z</i>	144176	1%
<i>p</i>	128382	1%
<i>h</i>	127330	1%
<i>ü</i>	126595	1%
<i>ş</i>	111626	1%
<i>v</i>	106782	1%
<i>f</i>	92105	1%
<i>ç</i>	80542	1%
<i>ğ</i>	57386	1%
<i>ö</i>	50915	1%
<i>w</i>	23094	0%
<i>x</i>	22557	0%
<i>j</i>	20484	0%
<i>q</i>	2665	0%

Tablo incelendiğinde a, e ve i harflerinin diğer sesli harflerden daha çok kullanıldığı n, l ve r sessizlerinin ise diğer sessiz harflerden daha sık kullanıldığı görülmektedir. En seçkin 21 Türk roman ve öykülerinin ve 7 süreli yayının incelendiği bir çalışmadaki (Büyükkuşçu ve Adalı, 2006) harf sıklıkları ile benzerlik göstermektedir. Aynı çalışmada sesli harflerin seçkin eserlerde geçme sıklığı %43,36 iken bizim çalışmamızda bu oran %42'dir.

Türkçe yazışmalarda kullanılabilen yabancı kökenli kelimeler dolayısıyla alfabemizde bulunmayan w, x ve q harfleri de istatistiğe dâhil olmuştur. Bu harflerin Türkçe yazışmalarda az kullanılıyor olması, Türkiye’de kullanılan klavye düzeninin İngilizce konuşulan ülkelerdekinden farklı olması gerektiğini bir kez daha ortaya koymaktadır. Alfabemizde bulunan j harfinin en az kullanılıyor olması da ayrıca dikkat çekicidir.

4.2. ANALİZ SONUÇLARI

4.2.1. Sınıflandırma Başarısı

Naive bayes makine öğrenme algoritması uygulanırken kelimeler kök halleri dikkate alınarak test verileriyle ölçüldüğünde %69’luk başarı elde edilmiştir. Olumsuzluk ekleri ayrı bir kelime olarak alınıp test edildiğinde başarı oranı %40’ düşmüştür. En iyi performans, %76’lık oran ile kelimelerin orijinal halleri ile eğitime tabi tutulduklarında elde edilmiştir.

“FIIL_OLUMSUZLUK_ME” ve “ISIM_YOKLUK_SIZ” eklerini alan kelimeler olumsuz kabul edilerek farklı kelime olarak tabloya eklendi. Bu bağlamda 8110 kelimenin (hecelerine ayrılabilen kelimelerin % 0,5’ i) olumsuzluk eki aldığı tespit edildi.

4.2.2. Karşılaştırmalı Analiz

Samsung ve Apple markaları özelinde aylara göre paylaşılan tweetlerin duygu analizi sonuçları aşağıdaki tablodaki gibidir.

Tablo 4.15. Samsung ve Apple Aylık Tweet Sayıları

Ay	Samsung				Apple			
	Olumlu	Olumsuz	Nötr	Tüm	Olumlu	Olumsuz	Nötr	Tüm
2	737	789	3237	4763	673	377	2245	3295
3	372	658	2148	3178	383	506	944	1833
4	410	300	1711	2421	494	450	2459	3403
5	402	428	1837	2667	416	549	1962	2927
6	714	897	4357	5968	1020	1053	2779	4852
7	2159	1687	6663	10509	372	616	2559	3547

Tablo 4.16. Samsung ve Apple Haftalık Tweet Sayıları

Hafta	Samsung				Apple			
	Olumlu	Olumsuz	Nötr	Tüm	Olumlu	Olumsuz	Nötr	Tüm
6	0	0	0	0	0	0	0	0
7	0	0	0	0	366	94	462	922
8	373	636	2167	3176	177	83	549	809
9	364	153	1070	1587	130	200	1234	1564
10	0	0	0	0	0	0	0	0
11	361	648	2105	3114	60	151	363	574
12	11	10	43	64	24	17	241	282
13	120	46	408	574	344	421	614	1379
14	166	98	552	816	86	146	396	628
15	22	30	178	230	10	35	261	306
16	81	63	393	537	26	93	443	562
17	21	63	180	264	327	93	1085	1505
18	10	27	96	133	78	92	326	496
19	0	0	0	0	0	0	0	0
20	56	62	365	483	171	141	623	935
21	47	57	353	457	167	316	1013	1496
22	434	461	1850	2745	0	0	0	0
23	333	447	2443	3223	837	436	1739	3012
24	78	89	422	589	133	559	668	1360
25	0	0	0	0	0	0	0	0
26	158	182	665	1005	50	58	372	480
27	789	738	2089	3616	59	121	293	473
28	141	83	585	809	51	76	413	540
29	229	308	1809	2346	78	154	880	1112
30	868	450	1739	3057	178	259	941	1378
31	132	108	441	681	6	6	32	44

4.2.3. Tweet Örnekleri

Çekilen tweetlerden 10 tanesi seçilerek aşağıda listelenmiştir:

Apple'ı 1 Nisan'da kurmuşlar. IBM'e yapılmış en uzun süren şaka

<https://t.co/sYKpHDgZvb>

Yaradan zaten sınıyor sana ne oluyor Apple.

#apple #iphone bana Siri'yi çiz kumpanyası başlamalı bence ama yarışması değil özgürce herkes hissettiğini çizmeli

Arkadaşım şarz aleti sorduğunda benimki sana uymaz(apple)haznı??

Samsung S8'in fiyatının 4.999 TL olarak açıklanmasına Apple'dan cevap; "Daha son sözümüzü söylemedik..." ?????? @zaytung

Unutma! Unutulanlar, unutanları asla unutmaz! imza: Apple Şifren.

Siri çok güzel bişey, siri sayesinde uzun zaman sonra kızla konuşan arkadaşlarım var. Tsk @Apple

yeni telefonlar vantilatör ya da klimalı olmalı apple bunu oku

galiba çöktüm,bagcilar cocugu gibi,65yas ustü dedeler gibi ve hatta apple pc gibi

Teknolojiden anlamak ile apple ürünlerini sevmek arasında ters korelasyon mevcuttur ve bunun istisnası yoktur.



SONUÇ ve TARTIŞMA

Stratejik kararlar alan yöneticiler için şirket dışındaki verilerin (müşteri geri dönüşlerinin ve rakip firma verilerinin) değeri ve dolayısıyla dış verilerin analiz edilmesinin önemi açıktır. İşletmeler müşterileri memnuniyet düzeyini ölçmek için uyguladıkları geri bildirim formu ve anket yöntemleri, küreselleşen dünyada yetersiz kaldı. Büyük veri olarak kabul edilen sosyal medya paylaşımları ise işletmelerin iştahını kabartmaya başladı. İşletmeler, bu verilerin temin edilerek analiz edilmesini ve karar verme süreçlerine dâhil edilmesini istemektedirler. Ancak, verinin oluşturulma hızına paralel olarak analizin de anlık olarak yapılması ve rapor edilmesi geleneksel yöntemlerin sınırlarını aşmaktadır.

İşletmeler büyük veriye geçmede tereddüt içindeler. Oysa düzensizliği, boyutu ve büyüme hızı giderek artan sosyal medya verilerinin analizi için geleneksel yöntem ve teknolojilerinin kabiliyetlerinin yetersiz oluşu işletmeleri büyük veriye yatırım yapmaya zorlamaktadır.

Müşteri yorumları analiz edildiğinde mevcut durum tespiti yapılmış olunmuyor, aynı zamanda gelecekle ilgili de bilgi elde edilmiş olunmaktadır. Müşteriler, viral reklam prensibiyle yatay düzlemde birbirlerini etkileyerek bir tür kelebek etkisi oluşturabilmektedir. Sosyal medyanın analiz edilmesinin, yeni risk ve fırsatların tespitinde ve geleceğin tahmin edilmesinde önemli bir yeri bulunmaktadır.

Bu çalışma ile işletmelerin, müşterini daha yakından tanımalarına ve rakiplerini takip etmelerine destek sağlaması için sosyal medya verilerini analiz ederek üst yönetimin stratejik kararlar almasını yönlendirici bir yazılım sisteminin geliştirilmesi amaçlanmıştır. Çalışma kapsamında geleneksel yöntemler ile büyük veri tekniklerinin birleştirildiği melez bir sistem tasarlanmıştır. Dolayısıyla bu çalışma, küçük işletmelerin de büyük veri analizlerine dâhil olabilmesi için düşük maliyetli küçük bir adım mesabesinde. Bu çalışma sayesinde işletmeler, müşterilerinin dilek, şikâyet ve önerilerini dikkate almak için sosyal medya analizlerinden faydalanabilmektedir. İşletmelerin, duygu analizine tabi tutacağı tweetlerin içinde geçen kelimeleri tercih edebilmesiyle, marka, ürün adı veya firmalar ile ilgili filtreleme imkânı elde edebilmektedir. Böylece işletme, ilgisiz tweetleri işleme almayarak, sadece karar problemine uygun olarak ilgilendiği tweetleri

filtreleyebilmektedir. Bu bağlamda Twitter'deki paylaşımların anlık olarak ve filtrelenerek çekilebildiği esnek bir API uygulaması geliştirilmiştir.

Twitter'in katılım ve kullanım kolaylığının bulunması; fırsat eşitliği sunan şeffaf bir ekosistem olması; Twitter'de kolayca takipçi bulunabilmesi ve grup oluşturulabilmesi gibi sebepler dolayısıyla Twitter yaygın bir kullanıma sahiptir. Twitter'de erişime açık yorumlar, pazar ve piyasa araştırması yapmak isteyen işletmelerin iştahını kabartsa da Twitter'de kontrolsüz veri girişi yapıyor olması, tweetlerin analiz edilebilmesini zorlaştırmaktadır. Bu çalışmanın Twitter verilerine odaklanılmasının gerekçesi de yaygın kullanımı yanısıra tweetlerin işlenmesinin ve analiz edilmesinin zorluğudur.

Çalışma kapsamında tweet çekme modülü, ön işlem modülü, sınıflandırma modülü ve raporlama modülü olmak üzere toplamda dört adet modül geliştirilmiştir. Tweetlerin çekilmesi modülü için Twitter API'leri (Twitter4J kütüphanesi) kullanan bir java uygulaması geliştirilerek anahtar kelimelerin (örneğin marka adları) geçtiği tweetler çekildikten sonra ilgili klasörlere "txt" formatında kaydedilmektedir. Tweet çekme modülünün periyodik olarak çalıştırılması ile de son atılan tweetlerin çekilmesi sağlanabilmektedir.

Visual Studio.NET uygulaması ile ön işleme tabi tutulan tweetler gereksiz ifadelerden (linkler, harf olmayan karakterler vs.) arındırıldıktan ve Zemberek kütüphanesi API'leri ile kök ve eklerine ayrıldıktan sonra SQL Server veri tabanına kaydedilmektedir. Köklerine ayırma işleminden maksat, kök ve ekleriyle ayrı ayrı analize sokarak başarıyı test etmek ve ayrıca kelimenin hatalı yazılıp yazılmadığını test etmektir. Bundan sonraki aşamada hatalı yazılan her bir kelime ayrıca ön işleme tabi tutulmaktadır.

Sosyal yazı dilinin doğasındaki çarpıklık, duygu analizi öncesinde yürütülen doğal dil işleme çalışmalarını zorunlu kılmaktadır. Ancak literatür incelendiğinde, Türkçe doğal dil işleme alanında çalışmaların yetersiz olduğu ve özellikle tweetlerin ön işlenmesiyle ilgili sistematik çalışmaların az olduğu tespit edildi. Sosyal medya yazışmalarının çarpıklığına Twitter'in yazım kısıtlaması da eklenince tweetlerin normalleştirilmesi ciddi zorluk teşkil etmektedir. Ayrıca Türkçe'nin sondan eklemeli dil yapısı kelime veritabanı oluşturulabilmesini ve dolayısıyla yazım yanlışlarının tespit edilmesini de zorlaştırmaktadır.

Bu çalışma kapsamında, biçim bilimsel olarak ayrışamayan Türkçe kelimeler için Zemberek'in öneri fonksiyonu yetersiz kaldığı için ayrıca bir ön işlem modülü tasarlanmıştır. Ön işlem aşamasında işlem adımları için algoritma kurulmuş; TSQL ve Python dillerinde kodlanmış ve test edilmiştir. Beklenenin aksine TSQL dili ilişkisel veri tabanı yönetim sisteminin sorgulama kabiliyetiyle daha başarılı sonuçlar vermiştir.

Ön işlem aşaması işletilirken öncelikle her bir tweet metni kelimelerine ayrılmaktadır. Kelimelerin doğru yazılıp yazılmadıklarının tespit edilebilmesi için Türkçe sözlükte arattırılması gerekmektedir. Bu çalışma kapsamında, ön işlem aşamasına hazırlık olması için, yaygın kullanılan Türkçe sözlük tabloları birleştirilerek yeni bir tablo (sözlük) oluşturulmuştur. Oluşturulan bu tabloya ayrıca özel isimler, il ve ilçe adları, marka adları ve durak kelimeleri eklenmiştir.

Kelime düzeltme algoritması işletilmeden önce hecelerine ayrılıp ayrılmadığına bakılmaktadır. Çınar (2003) tarafından sözel ifade edilen Hecelerine ayırma algoritması şematize edilmiş ve algoritmanın, Delibaş (2008) tarafından kurulan algoritmadan daha sade olduğu ortaya çıkmıştır. Çalışma kapsamında tweetler içinde geçen kelimelerin yaklaşık olarak %97' si hecelerine ayrılabilmiştir.

Sözlükte geçmeyen ve Zemberek tarafından kökü tespit edilemeyen kelimeler için kelime düzeltme aşamasına geçilmektedir. Kelime düzeltme algoritmasının işleyişi özetle şöyle: öncelikle tekrarlı harfler ikiye düşürülür. Son harf b, c, d veya g harflerinden biriye sırasıyla p, ç, t ve k harfleriyle değiştirilir. Kelime içinde geçen x, w ve q harfleriye sırasıyla h, v ve k harfleriyle değiştirilir. Kelime sonuna gelen çekim ekleri çıkarıldıktan sonra sözlükte bulunamadığında Türkçe karakterler ASCII eşlenikleriyle değiştirilerek sözlükte yeniden taranır. Son olarak çiftleyen harfler teke düşürülerek sözlükte yeniden taranır. Yine bulunamadıysa da tombul parmak hatası (klavyede yanlışlıkla komşu tuşa basma) durumu dikkate alınarak kelime önerilmeye çalışılır.

Twitterin karakter kısıtlaması ve sosyal medya yazışmalarının doğası, bireyleri kısaltma kullanmaya yönlendirmiştir. Kısaltmalar sözlükte bulunmadığı için de hatalı yazım kabul edilmektedir ve otomatik düzeltme becerisi yetersiz kalmaktadır. Kısaltılmış ifadelerin normalleştirilebilmesi için bu çalışma kapsamında eşleştirme tablosunun oluşturulması ve kullanıcı hatalarından öğrenen bir modül önerilmiş ve geliştirilmiştir.

Çalışmanın duygu analizine hazırlık aşamasında, sınıflandırmada kullanılacak eğitim veri setinin oluşturulmasında ve otomatik sınıflandırmada kullanılmak üzere, duygu belirleyici ifadeler, duygu emojileri ve durak kelimeleri tabloları oluşturulmuştur. Duygu belirleyici ifade tablosu, Şimşek ve Özdemir (2012) tarafından oluşturulan 113 duygu belirleyici ifadeye eklemeler yapılarak oluşturulan 1766 adet duygu belirleyici ifadeden oluşmaktadır. Bu ifadelerden faydalanılarak tweetlerin yaklaşık %54' ü sınıflandırılabilmiştir. Duygu sembolleri (16 adet) marifetiyle de tweetlerin %11' i sınıflandırılabilmiştir. Otomatik sınıflandırma başarısı ise genel olarak yaklaşık %84' tür.

Duygu analizi için kelime torbası yöntemi (bag of words) ile makine öğrenmesi (Naive Bayes sınıflandırma) ve sözlük tabanlı yöntem birlikte uygulanmıştır. Kelimelerin çoğunlukla kök halleriyle tutulduğu sözlükle eşleştirilebilmesi için kelimelerin köküne inilmesi gereği doğmuştur. Kelimelerin kök ve eklerine ayrıştırılmasında ise Zemberek kütüphanesinden faydalanılmıştır.

Naive bayes sınıflandırması önce TSQL dilinde kodlanarak HiveQL diline çevrilmiştir. Her iki dilde ve platformda çalıştırılarak test edilmiştir. Hive ortamında tabloların oluşturulması dahi ayrıca zaman aldığından ve yapısının, birden fazla tablonun birlikte sorgulanmasına uygun olmamasından; HiveQL sorguları daha yavaş neticelenmiştir. Analiz edilecek veri büyüdükçe, Hadoop kümesi ölçeklenerek daha anlamlı performans sonuçlarının elde edilmesi beklenmektedir.

Büyük veri analizine geçiş planlamalarına yardımcı olunması için Hadoop ortamının Windows ve Linux işletim sistemine sahip makineler üzerine nasıl kurulacağı ve Hive kurulumları etüd edilmiştir. Hadoop, fiziksel olarak Ubuntu işletim sistemli 3 makine üzerine kurulmuş; Windows ortamında ise Oracle Virtualbox üzerinde Hortonworks Sandbox 2.0 kullanılmıştır. Sandbox sanal makinesinde Hive, kurulu olarak geldiği için sadece Linux makinelere Hive kurulumu yapılmıştır. Hive'in aynı işlevleri toplu işlem (batch) yönelimiyle yapıyor olması ve Sql'e uygun ilişkisel forma dönüştürebiliyor olması bu çalışmada tercih edilme sebebidir.

Çalışmanın test edilebilmesi ve öğrenme motorunun eğitilerek duygu sözlüğünün oluşturulabilmesi için Türkiye'de faaliyet gösteren markalar hakkında paylaşılan tweetler (son altı ay için) çekilerek sisteme dâhil edilmiştir. 2700 tweet manuel olarak duygu

sınıfına dâhil edilmiştir. Analiz başarısı ölçülürken de manuel olarak duygulandırılan bu veri kümesinden faydalanılmıştır.

Sonuçların zamansal ve mekânsal olarak raporlanabilmesi için tweetlerin paylaşılan konum bilgileri de çekilmiştir (çekilen tüm tweetlerin %33'ünün konum bilgisi elde edilebilmiştir). Konum bilgisi bulunan tweetler incelendiğinde, illerin nüfus sayısı ile atılan tweet sayısının pozitif ilişkili (%96) olduğu görülmektedir. Ayrıca, nüfusu yoğun olan ilk üç ilden (İstanbul, Ankara ve İzmir) atılan tweetlerin, toplam tweet sayısının %70'ini oluşturduğu görülmüştür. Yaygın olarak Kürtçe konuşulan (Ağrı, Batman, Bingöl, Hakkâri, Mardin, Muş, Şanlıurfa, Şırnak, Tunceli ve Van) illerdeki tweeter kullanımının ise ortalamanın çok altında (ortalamanın 1/7'si) olduğu görülmüştür. Bu veriler, tweetlerin konum bilgisinin önemini ayrıca ortaya koymaktadır.

Tweetlerde kullanılan hece yapıları incelendiğinde ise hecelerin önemli bir bölümünün (%86) sessiz harfle başladığı bulgulanmıştır. Harf kullanım istatistikleri incelendiğinde, sesli harf kullanım oranının (%42), Büyükkuşçu ve Adalı (2006) tarafından yapılan en seçkin 21 Türk roman ve öykülerinin ve 7 süreli yayının incelendiği çalışmadaki istatistiksel sonuçlar (%43) ile benzerlik gösterdiği görülmüştür. Türkiye Türkçesinde bulunmayan w, x ve q harflerinin kullanım sıklığının çok düşük çıkması ise Türkiye'de kullanılan klavye düzenine karar verirken ayrıca dikkate alınmalıdır.

Telefon kelimesinin Tweetler içinde geçen en yaygın 20 kelimedenden biri olması, “#teknoloji” etiketinin ve marka adlarının en yaygın kullanılan etiketler içinde olması ve “sung” hecesinin (samsung kelimesinden dolayı) en yaygın kullanılan heceler içinde yer alması duygu analizi çalışmalarının alan bağımlı yürütülmesinin daha doğru olacağı sonucuna ulaştırmaktadır.

Çekilen Tweetler üzerinde yapılan diğer istatistiksel çalışmalar göstermiştir ki hiç takipçisi olmayan kullanıcılar, en az tweet atarken; takipçi sayısı yüksek olan kullanıcılar, daha aktif paylaşımda bulunmaktadır. Buna rağmen 312632 tweet içinden hiç okunmayan tweetlerin toplam veri setinin yarıdan fazlasını teşkil ettiği bilgisi elde edilmiştir. Hiç favori yapılmayan tweet yüzdesi ise %77'dur. Retweet edilme sayıları da incelendiğinde retweet edilenlerin daha çok sayıda kullanıcıya ulaştığı görülmüştür. Dolayısıyla takipçi sayısı yüksek kullanıcılar, okunma, retweet edilme ve favori yapılma sayısı yüksek tweetler, diğer kullanıcılara ve tweetlere nispetle analizler için daha zengin içerikler

sunmaktadır. Bu istatistiklere dayanarak, sadece belirli takipçi sayısı üzerindeki kullanıcılar tarafından paylaşılan ve en az bir kere okunan tweetlerin analize sokulmasının daha anlamlı sonuçlar vereceği söylenebilir.

Yukarıda verilen istatistiksel sonuçlardan hareketle bu çalışmanın analiz sürecinde retweetler ve tweetlerin okunma sayıları da dikkate alınarak tweetlerin gerçek etkisi hesaplanmaya çalışılmıştır. Yani, duygu puanı belirlerken tweetlerin okunma sayıları da dikkate alınmıştır. Tweetlerin etki ağırlığının hesaplandığı diğer çalışmalar da ise takipçi sayısı, retweet sayısı (Kwak vd., 2010) ve bahsedilme sayısı (Cha vd., 2010) da hesaplama katılmıştır.

Duygu sınıflandırmasında kelimeler kök halleriyle ve ekli halleriyle ayrı ayrı test edilmiştir. En iyi performans ise %76'lık oran ile kelimelerin orijinal halleri ile eğitime tabi tutulduklarında elde edilmiştir. Bu oran, alan bağımlı yaklaşık 8 bin tweet üzerinde Naive Bayes sınıflandırma yapan Meral ve Diri (2014)'nin başarısına (%78) ve Onan (2016)'ın başarısına (%78.3) oldukça yakın çıkmıştır. Çalışma başarısı alan bağımsız çalışmalarla kıyaslandığında ise Akarsu ve Diri (2016)'nin sınıflandırma başarısından (%81) düşük, Oğul ve Ercan (2016)'nin sınıflandırma başarısından (%74) ise yüksek çıkmıştır. “FIIL_OLUMSUZLUK_ME” ve “ISIM_YOKLUK_SIZ” eklerini alan kelimelerin ayrı bir kelime olarak kabul edildiği veri setinde ise başarı beklenenin aksine düşük (%40) çıkmıştır.

Analiz sonuçları, geliştirilen Visual Studio.NET uygulaması ile kullanıcıya grafiksel olarak sunulmaktadır. Rapor ekranları marifetiyle sonuçlar, günlük, haftalık, aylık ve 6 aylık periyotlar şeklinde alınabilmektedir. Sistemin son kullanıcılarının da şirketin üst düzey yöneticileri olması öngörülmektedir.

Çalışma kapsamında tweet içerikleri içinden yalnızca metinler analiz edilebilmiştir. Tweet içinde geçen bağlantılar analize dâhil edilememiştir. Ayrıca Twitter tarafından kullanıcıların yaşları, kültür ve eğitim düzeyleri gibi demografik özelliklerini paylaşmadığı için kullanıcıların yorum paylaştıkları konuya ilgi düzeyleri de tespit edilememiştir.

Çalışma kapsamında yalnızca Türkçe tweetler analiz edilmiştir. Gelecek çalışmalarda İngilizcenin de analize dâhil edilmesi önerilmektedir. Ayrıca Türkçenin

yerel kullanım (ağız) farklılıklarından kaynaklanan sapmaların, İstanbul Türkçesine normalleştirilmesi ile ilgili de detaylı bir çalışma da yürütülebilir.

Birden fazla marka adının birlikte geçtiği tweetlerde hangi marka için olumlu söylemde bulunulduğu tam olarak tespit edilememiştir. İronik ifadelerin anlaşılmasında ise noktalama işaretleri ve emojilerden faydalanıldıysa da insan algısını dahi aşan bazı ironik söylemler analizi zorlaştırmıştır.

Uygulamanın yaygın kullanılan ingilizce dili de dâhil edilerek ticari bir paket yazılıma dönüştürülebilmesi sağlanabilir. Analiz sonuçlarının ise web ve mobil uygulamalar ile üst yönetime sunulması sağlanabilir.

Mevcut çalışma pazar ve piyasa hakkında durum tespiti yapmaktadır. Gelecek çalışmalar için de pazar trendinin tahmin edilmesi ve Borsada faaliyet gösteren firmalar için hisse senedi fiyatlarının tahmin edilmesi çalışılabilir. Ayrıca duygu sınıflandırması için büyük veri analizi teknikleri, Yapay Sinir Ağları yöntemi için uygulanıp test edilebilir.

KAYNAKÇA

- Abdulnabi, N.Z.T. (2016). *Semantic Analysis Using Natural Language Processing Methods*. (Yayımlanmış Yüksek Lisans Tezi), İstanbul: Yıldız Teknik Üniversitesi.
- Akarsu, C., & Diri, B. (2016). "Twitter ile Türk Televizyonları Rating Tahmini" [Bildiri]. *The Signal Processing and Communication Application Conference (SIU)*, 2016 24th, Zonguldak.
- Akarsu, C. (2016). *Twitter Verileri İle Türk Televizyonları İzlenme Oranı Sıralamaları Tahmini*. (Yayımlanmış Yüksek Lisans Tezi), İstanbul: Yıldız Teknik Üniversitesi, Fen Bilimleri Enstitüsü, Bilgisayar Mühendisliği.
- Akba F., Uçan, A., Akçapınar Sezer, E. ve Sever, H. (2014). "Assessment of Feature Selection Metrics for Sentiment Analyses: Turkish Movie Reviews" [Bildiri]. *The 8th European Conference on Data Mining*, Lisbon, Portugal.
- Akbaş, E. (2012). *Aspect Based Opinion Mining on Turkish Tweets*. (Yayımlanmış Yüksek Lisans Tezi), Ankara: Bilkent Üniversitesi, Fen Bilimleri Enstitüsü, Bilgisayar Mühendisliği.
- Akgül, E. S., Ertano, C., & Banu, D. (2016). "Twitter Verileri ile Duygu Analizi" *Pamukkale Üniversitesi Mühendislik Bilimleri Dergisi*, 22(2), 106-110.
- Akın, M. D. (2007). *Zemberek Nasıl Çalışır? 1.Sözlük ve Kök ağacı*. <http://zembereknlp.blogspot.com.tr/?view=classic>
- Akın, M. D., & Akın, A. A. (2007). "Türk Dilleri için Açık Kaynaklı Doğal Dil İşleme Kütüphanesi". *Elektrik Mühendisliği*, 431, 38-44.
- Akpınar, H. (2014). *Data: Veri Madenciliği Veri Analizi*. İstanbul: Papatya Yayıncılık Eğitim.
- Al-Khoury, A. M. (2013). "Privacy in The Age of Big Data: Exploring The Role of Modern Identity Management Systems" [Büyük Veri Çağında Gizlilik: Modern Kimlik Yönetim Sistemlerinin Rolünü Keşfetmek]. *World Journal of Social Science*, 1(1), 37-47.

- Allen, J.F. (2003). "Natural Language Processing" [Doğal Dil İşleme]. In Encyclopedia of Computer Science, 1218–1222. John Wiley and Sons Ltd.
- Atan, S. (2016). "Veri, Büyük Veri ve İşletmecilik." *Balıkesir Üniversitesi Sosyal Bilimler Enstitüsü Dergisi*, 19(35), 137-153.
- Altıntaş, K., ve Çiçekli, İ., (2001). "A Morphological Analyser for Crimean Tatar" [Bildiri]. *The 10th Turkish Symposium on Artificial Intelligence and Neural Networks*, Tainn, North Cyprus.
- Amasyalı, M. (2005). "Türkçe Wordnet'in Otomatik Olarak Oluşturulması" [Bildiri]. *The Signal Processing and Communications Applications Conference*, Kayseri.
- Amasyalı, M. F., Balcı, S., Mete, E., Varlı, E. N. (2012). "A Comparison of Text Representation Methods for Turkish Text Classification" [Türkçe Metinlerin Sınıflandırılmasında Metin Temsil Yöntemlerinin Performans Karşılaştırılması]. *EMO Bilimsel Dergi*, 2(4), 95-104.
- Anderson, C. (2008). "The End Of Theory: The Data Deluge Makes The Scientific Method Obsolete" [Teorinin Sonu: Veri Tufanı, Bilimsel Metodu Eski Yapar]. *Wired magazine*, 6.
- Anuradha, J. (2015). "A brief introduction on Big Data 5Vs characteristics and Hadoop technology" [Big Data 5Vs özellikleri ve Hadoop teknolojisi hakkında kısa bir giriş]. *Procedia Computer Science*, 48, 319-324.
- Aodba. (2016). *What are the Hadoop HIVE Data Types ?* <http://www.aodba.com/what-are-the-hadoop-hive-data-types/>
- Araç, B. (2016). *Apache Hive*. <http://www.kodumundunyasi.net/big-data/apache-hive.html>
- Argamon, S., Whitelaw, C., Chase, P., Hota, S. R., Garg, N., Levitan, S. (2007). "Stylistic Text Classification Using Functional Lexical Features." [İşlevsel Sözcük Özellikleri Kullanan Biçimsel Metin Sınıflandırması]. *Journal of the American Society for Information Science and Technology*, 58(6), 802-822.

- Atalık, Ö., Sak, F. S., Bayrak, Ü., & Karataş, E. (2014). "Havayolu İşletmelerinin Sosyal Medya Kullanımları: Twitter Örneği" [Bildiri]. *The II. International Aviation Management Conference*, Ankara.
- Atchariyachanvanich, K., Nalintippayawong, S., & Permpool, T. (2017). "Development of a MySQL Sandbox for Processing SQL Statements: Case of DML and DDL Statements" [Bildiri]. *14th International Joint Conference*, Thailand.
- Avcıoğlu, C. (2016). *Tüketicilerin Markalar ile Sosyal Medya Kanalları Aracılığıyla Kurduğu Bağ ve Pazarlama Stratejilerine Etkileri*. (Yayımlanmış Yüksek Lisans Tezi). İstanbul: Doğu Üniversitesi, Sosyal Bilimler Enstitüsü, İşletme.
- Aydoğan, A., & Başaran, F. (2012). "Yeni Medyayı Alternatif Medya Bağlamında Anlamak" [Bildiri]. *Alternatif Medya Alternatif Gazetecilik: Türkiye'de Alternatif Gazetecilik Üzerine Değerlendirmeler*, İstanbul.
- Banger, G., & Çalışır, G. (2014). "Sosyal Medyanın Kurumsal İnovasyon için Kitle Kaynak Olarak Kullanımı" [Bildiri]. *I. Uluslararası Yeni Medya Yeni Yaklaşımlar Konferansı*, Çanakkale.
- Bansal, H., Chauhan, S., & Mehrotra, S. (2016). *Apache Hive Cookbook*. United Kingdom: Packt Publishing Ltd.
- Barske, A. (2013). *Big Data Whitepaper*. http://www.academia.edu/4556547/Big_Data_Turkce_Whitepaper
- Barutçu, S., & Tomaş, M. (2013). "Sürdürülebilir Sosyal Medya Pazarlaması ve Sosyal Medya Pazarlaması Etkinliğinin Ölçümü". *İnternet Uygulamaları ve Yönetimi Dergisi*, 4(1), 5-24.
- Bayrakcı, S. (2015). *Sosyal Bilimlerdeki Akademik Çalışmalarda Büyük Veri Kullanımı*. (Yüksek Lisans Tezi), İstanbul: Marmara Üniversitesi Sosyal Bilimler Enstitüsü, İletişim Bilimleri.
- Beyer, M. A., & Laney, D. (2012). *The Importance of 'Big Data': a Definition*. <https://www.gartner.com/doc/2057415/importance-big-data-definition>

- Bian, J., Topaloglu, U., & Yu, F. (2012). "Towards large-scale twitter mining for drug-related adverse events" [Bildiri]. *The Proceedings of the 2012 international workshop on Smart health and wellbeing*.
- Bilgin, O., Çetinoğlu, Ö., & Oflazer, K. (2004). "Building a Wordnet for Turkish" [Türkçe için bir Wordnet Oluşturulması]. *Romanian Journal of Information Science and Technology*, 7(1-2), 163-172.
- Binny Mathews, O. A. (2017). *Hive Advantages*. <https://www.dezyre.com/hadoop-course/hive#advantage>
- Bollen, J., Mao, H., & Zeng, X. (2011). "Twitter Mood Predicts the Stock Market" [Twitter Ruh Hali Borsayı Tahmin Eder]. *Journal of Computational Science*, 2(1), 1-8.
- Bosco, C. (2004). *A Grammatical Relation System for Treebank Annotation*. (Yayımlanmış Doktora Tezi) Torino: University of Torino.
- Boynukalın, Z. ve Karagöz, P. (2013). "Emotion Analysis on Turkish Texts" [Türkçe Yazılarda Duygu Analizi]. *Information Sciences and Systems*, 264, 159-168
- Bozşahin, C. (2002). "The combinatory Morphemic Lexicon" [Kombinatoriyel Biçimbirimsel Sözlük]. *Computational Linguistics* 28(2) 145–186
- Borthakur, D. (2008). *HDFS Architecture Guide*. https://hadoop.apache.org/docs/r1.2.1/hdfs_design.html
- Brown, B., Chui, M., Manyika, J. (2011). "Are You Ready for the Era of Big Data?" [Büyük Veri Çağına Hazır mısınız?], *McKinsey Quarterly*, 4(1), 24-35.
- Burgaz, F. (2016). *Linux Paket Güncelleme Komutları*. <http://fatihburgaz.com/linux-paket-guncelleme-komutlari>
- Büyükkuşçu, İ., E. Adalı. "Heceleme Yöntemiyle Kök Sözcük Üretme." *Türkiye Bilişim Vakfı Bilgisayar Bilimleri ve Mühendisliği Dergisi* 2.1 (2006).
- Brown B., Chui M., & Manyika J. (2011). *Are You Ready for the Era of Big Data*, McKinsey Quarterly.
- Butt, M., King, T.H. (2005). "Restriction For Morphological Valency Alternations: The Urdu Causative" [Bildiri]. *10th International LFG Conference*, Bergen, Norway.

- Canals-Marote, R., Esteve-Guillén, A., GarridoAlenda, A., Guardiola--Savall, M. I., IturraspeBellver, A., Montserrat-Buendia, S., PérezAntón-Rojas, P., Ortiz-Pina, S., Pastor-Antón, H., ve Forcada, M. L., (2000). interNOSTRUM: a Spanish-Catalan Machine Translation System, *Machine Translation Review*, 11, 21-25.
- Capriolo, E., Wampler, D., & Rutherglen, J. (2012). *Programming Hive: Data Warehouse and Query Language for Hadoop* (Vol. 1). Amerika: O'Reilly Media, Inc.
- Casado, R., & Younas, M. (2015). Emerging Trends And Technologies In Big Data Processing. *Concurrency and Computation: Practice and Experience*, 27(8), 2078-2091.
- Cavnar, W. B., Trenkle, J. M., 1994. "N-gram-Based Text Categorization" [N-Gram Tabanlı Yazı Sınıflandırması]. *Ann Arbor MI*, 48113(2), 161-175.
- Cebiroğlu, G. (2002). *Sözlüksüz Köke Ulaşma Yöntemi*. (Yayımlanmış Doktora Tezi), İstanbul: İstanbul Teknik Üniversitesi Fen Bilimleri Enstitüsü, Bilgisayar Mühendisliği Bilimleri.
- Cha, M., Haddadi, H., Benevenuto, F., & Gummadi, K. P. (2010). "Measuring User Influence In Twitter: The Million Follower Fallacy" [Bildiri]. *ICWSM 10: Proceedings of international AAAI Conference on Weblogs and Social, Washington*.
- Chung, H. ve Rim, H. (2004). "Unlexicalized Dependency Parser For Variable Word Order Languages Based On Local Contextual Pattern" [Bildiri]. *Computational Linguistics and Intelligent Text Processing*, Seoul.
- Claster, W.B., Cooper, M., Sallis, P. (28-30 Sept. 2010). "Tourism and Conflict: Modeling Sentiment from Twitter Tweets Using Naive Bayes and Unsupervised Artificial Neural Nets" [Bildiri]. *2010 Second International Conference, Bali*.
- Claster, W. B., Dinh, H., & Cooper, M. (2010). "Naive Bayes and Unsupervised Artificial Neural Nets for Cancun Tourism Social Media Data Analysis" [Bildiri]. *The Nature and Biologically Inspired Computing (NaBIC), 2010 Second World Congress on, Kitakyushu, Japan*.

- Claster, WB., Hung, D., & Cooper, M. (2010). "Bayes and unsupervised artificial neural nets for Cancun tourism social media data analysis" [Bildiri]. *Nature and Biologically Inspired Computing (NaBIC), 2010 Second World Congress*, Fukuoka.
- Cugola, G., & Margara, A.(2012). Processing Flows of Information: From Data Stream to Complex Event Processing. *ACM Computing Surveys*, 44(3), 15.
- Çağlıyan, V., Hassan, S. A., & Işıklar, Z. E. (2016). "Üniversite Öğrencilerinin Satın Alma Davranışlarında Sosyal Medya Reklamlarının Etkisi: Selçuk Üniversitesi'nde Bir Araştırma". *Selçuk Üniversitesi Sosyal ve Teknik Araştırmalar Dergisi*, 11(11), 43-56.
- Çakıcı, R. (2005). Automatic Induction of a CCG Grammar for Turkish. [Bildiri]. *Proceedings of the ACL Student Research Workshop*, Michigan.
- Çalış, T. (2017). *Sözdizimsel Aktarıma Dayalı Makine Çevirisi*. (Yayımlanmış Yüksek Lisans Tezi), Edirne: Trakya Üniversitesi, Fen Bilimleri Enstitüsü, Bilgisayar Mühendisliği.
- Çavdaroglu, G. Ç. (2014). "Semantik Web Altyapısıyla Duygu ve Eylem Modelleme" [Bildiri], *Proceedings of the 8th Turkish National Software Engineering Symposium*, KKTC .
- Çelikyılmaz, A., Hakkani-Tur, D. and Junlan, F. (2010). "Probabilistic model-based sentiment analysis of twitter messages" [Bildiri]. *Spoken Language Technology Workshop (SLT), 2010 IEEE*, Berkeley, CA.
- Çetin, M., & Amasyalı, M. F. (2013). "Supervised and Traditional Term Weighting Methods for Sentiment Analysis" [Bildiri]. *The Signal Processing and Communications Applications Conference (SIU)*, Haspolat.
- Çevik, M., Öztürkcan, S., & Kasap, N. (2015). "Sosyal Medya Analitiği: Twitter için Büyük Veri Yaklaşımı" [Bildiri]. *20.Ulusal Pazarlama Kongresi*, Eskişehir.
- Çınar, M. (2003). *Veri Ambarlarında Verilerin Temizlenmesi*. (Yayımlanmış Yüksek Lisans Tezi), İstanbul Teknik Üniversitesi Fen Bilimleri Enstitüsü, Bilgisayar Mühendisliği Bilimleri.

- Çoban, Ö. (2016). *Metin Sınıflandırma Teknikleri ile Türkçe Twitter Duygu Analizi* (Yayımlanmış Yüksek Lisans Tezi), Erzurum: Atatürk Üniversitesi, Fen Bilimleri Enstitüsü, Bilgisayar Mühendisliği Bilimleri.
- Çoban, Ö., Özyer, B., & Özyer, G. T. (2015). "Sentiment Analysis for Turkish Twitter feeds" [Bildiri]. *The Signal Processing and Communications Applications Conference (SIU)*, Malatya.
- Davenport, T. H. (2014). *Big Data @ Work*. Amerika: Türk Hava Yolları Yayınları.
- Davenport, T. H., Barth, P., & Bean, R. (2012). "How Big Data is Different" [Büyük Veri Ne Kadar Farklı]. *MIT Sloan Management Review*, 54(1), 43-47.
- Davidov, D., Tsur, O., Rappoport, A. (2010). "Enhanced Sentiment Learning Using Twitter Hashtags And Smileys" [Bildiri]. *The 23rd International Conference on Computational Linguistics*, Beijing, China.
- Dawley, L. (2009), "Social Network Knowledge Construction: Emerging Virtual World Pedagogy" [Sosyal Ağ Bilgisi İnşası: Gelişen Sanal Dünya Pedagojisi], *the Horizon*, 17 (2), 109-121.
- De Mauro, A., Greco, M., Grimaldi, M., & Ritala, P. (2018). "Human Resources For Big Data Professions: A Systematic Classification Of Job Roles And Required Skill Sets" [Büyük Veri Meslekleri İçin İnsan Kaynakları: İş Rollerinin ve Gerekli Beceri Setlerinin Sistematik Bir Sınıflandırması]. *Information Processing & Management*, 54(5), 807-817.
- Dean, J., & Ghemawat, S. (2008). "MapReduce: Simplified Data Processing On Large Clusters". *Communications of the ACM*, 51(1), 107-113.
- Dehkharghani, R., Saygin, Y., Yanikoglu, B., & Oflazer, K. (2016). "SentiTurkNet: a Turkish Polarity Lexicon For Sentiment Analysis" [SentiTurkNet: Duygu Analizi için Bir Türkçe Polarite Sözlüğü]. *Language Resources and Evaluation*, 50(3), 667-685.
- Delibaş, A. (2008). *Doğal Dil İşleme ile Türkçe Yazım Hatalarının Denetlenmesi*. (Yayımlanmış Doktora Tezi), İstanbul: İstanbul Teknik Üniversitesi Fen Bilimleri Enstitüsü, Bilgisayar Mühendisliği.

- Demirtaş, B., & Arğan, M. (2015). "Büyük Veri ve Pazarlamadaki Dönüşüm: Kuramsal Bir Yaklaşım". *Pazarlama ve Pazarlama Araştırmaları Dergisi*, 15, 1-21.
- Dean, J., & Ghemawat, S.(2008). "MapReduce: Simplified Data Processing On Large Clusters" [Eşle İndirge: Büyük Kümelerde Basitleştirilmiş Veri İşleme]. *Communications of the ACM*, 51(1), 107-113.
- Dener, M., Dörterler, M., & Orman, A. (2009). "Açık Kaynak Kodlu Veri Madenciliği Programları: Weka’da Örnek Uygulama". *Akademik Bilişim*, 9, 11-13.
- Dilipak, A., Dilipak, A., & Meriç, N. (2016). *Türk Evim*. www.turkevim.com/ld/0/90_ANSIKLOPEDIK_IS.doc
- Dilsiz, S. (2005). *Bulanık Mantık ve Yapay Sinir Ağları ile Türkçe Yazım Denetleyicisi*. (Yayımlanmış Doktora Tezi), İstanbul: İstanbul Teknik Üniversitesi, Fen Bilimleri Enstitüsü, Bilgisayar Mühendisliği.
- Doğan, M. (2014). *Büyük Veri'nin Kişiler ve Kurumlar Üzerindeki Etkileri*. (Yayımlanmış Yüksek Lisans Tezi), İstanbul: İstanbul Bilgi Üniversitesi Sosyal Bilimler Enstitüsü, Bilişim ve Teknoloji Hukuku.
- Dorenkott, J. (2013). *Big Data Part 4*. <http://www.relationalsolutions.com/blog/bid/334338/Big-Data-Part-4>
- Dönmez, İ. (2016). *Türkçe Tümcelerin Yükleme Odaklı Anlam ve Dil Bilgisi*. (Yayımlanmış Doktora Tezi), İstanbul: İstanbul Teknik Üniversitesi, Fen Bilimleri Enstitüsü, Bilgisayar Mühendisliği.
- Du, D. (2015). *Apache Hive Essentials*: United Kingdom: Packt Publishing Ltd.
- Dumbill, E. (2012). *Planning for Big Data*: California: O'Reilly Media, Inc.
- Dülger, Ü. (2015). *Stratejik Büyük Veri Yönetiminin Yatırımlar Üzerindeki Etkileri*. (Yayımlanmış Yüksek Lisans Tezi), İstanbul: İstanbul Üniversitesi Fen Bilimleri Enstitüsü, Mühendislik Bilimleri.
- Dzeroski, S., Erjavec, T., Ledinek, N., Pajas, P., Zabokrtsky, Z., Zele, A. (2006). "Towards a Slovene Dependency Treebank" [Bildiri]. *Fifth International Conference on Language Resources and Evaluation*, Genoa, Italy.

- El-Kahlout, D., Dönmez, İ. ve Oflazer, K. (2006). "Türkçe-İngilizce için İstatistiksel Bilgisayarlı Çeviri Sistemi" [Bildiri]. *AINN 2006 Fifteenth Turkish Symposium on Artificial Intelligence and Neural Networks*, Akyaka, Muğla, Türkiye.
- Elmas, M. (2012). *Destek Vektör Makineleri ile Fiyat Tahminleri ve Kuyumculuk Sektöründe Bir Uygulama*, (Yayımlanmış Yüksek Lisans Tezi), İstanbul: İstanbul Üniversitesi, Fen Bilimleri Enstitüsü, Endüstri Mühendisliği.
- Eryiğit, G. ve Oflazer, K. (2006). "Statistical dependency parsing of Turkish" [Bildiri]. *The 11th Conference of the European Chapter of the Association for Computational Linguistics*, Trento, Italy
- Eryiğit, G., Oflazer, K., & Adalı, E. (2006). "Türkçe Cümlelerin Kural Tabanlı Bağlılık Analizi" [Bildiri]. *The Fifteenth Turkish Symposium on Artificial Intelligence and Neural Networks*, Muğla.
- Esen, M. F., & Türkay, B. (2017). "Turizm Endüstrilerinde Büyük Veri Kullanımı", *Journal of Tourism and Gastronomy Studies*, 5(4), 92 - 115.
- Farhan, M. N., Habib, M. A., & Ali, M. A. (2018). "A study and Performance Comparison of MapReduce and Apache Spark on Twitter Data on Hadoop Cluster" [Hadoop Kümesindeki Twitter Verileri Üzerinde MapReduce ve Apache Spark'ın Performans Karşılaştırması ve Bir Çalışma]. *Information Technology and Computer Science*, 7, 61-70
- Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). "From Data Mining to Knowledge Discovery in Databases" [Veri Madenciliğinden Veritabanlarındaki Bilgi Keşfine]. *AI magazine*, 17(3), 37-54.
- Gantz, J., & Reinsel, D. (2012). *Bigger Digital Shadows and Biggest Growth in the Far East*. <https://www.emc.com/collateral/analyst-reports/idc-the-digital-universe-in-2020.pdf>
- Gattiker, T. F., Chen, D., & Goodhue, D. L. (2005). *Agility Through Standardization: A CRM/ERP Application, Strategic ERP Extension and Use*. Stanford: Stanford University Press,.

- Gillis, T., & Stephanny, P. (2015). *Going Beyond the Data*. <https://home.kpmg.com/xx/en/home/insights/2015/03/going-beyond-the-data-indirect-tax.html>
- Gruzd, A., Doiron, S., Mai, P. (2011). "Is Happiness Contagious Online? A Case of Twitter and the 2010 Winter Olympics." [Bildiri]. *44th Hawaii International Conference*, Kauai, HI.
- Go, A., Bhayani, R., Huang, L. (2009). *Twitter Sentiment Classification Using Distant Supervision (Rapor No: CS224N)*. Stanford.
- Gökkurt, Ö., & Demirtel, H. (2014). "Sosyal Medyada Kurumsal Arşivleme Sorunu" [Bildiri]. *Uluslararası Kültürel Mirasın Ve Kültürel Bellek Kurumlarının Yönetimi Kongresi*, Ankara.
- Gökşen, Y., Aşan, H. (2015). "Veri Büyüklüklerinin Veritabanı Yönetim Sistemlerinde Meydana Getirdiği Değişim: NoSQL". *Bilişim Teknolojileri Dergisi*, 8(3), 125-131
- Gu, L., & Li, H., (2013). "Memory or time: Performance evaluation for iterative operation on hadoop and spark" [Bildiri]. *2013 IEEE International Conference on Embedded and Ubiquitous Computing (HPCC_EUC)*, Çin.
- Gunarathne, T. (2015). *Hadoop MapReduce v2 Cookbook*: United Kingdom: Packt Publishing Ltd.
- Gurusamy, V., Kannan, S., & Nandhini, K. (2017). "The Real Time Big Data Processing Framework: Advantages and Limitations" [Gerçek Zamanlı Büyük Veri İşleme Framework'u: Avantajları ve Sınırları]. *International Journal of Computer Science and Engineering*, 5(12), 2347-2693.
- Gülseçen, S. (2015). *Bilgi Yönetimi: Bilgi Türeticileri, Büyük Veri, İnovasyon, Kurumsal Zekâ* (Vol. 1): İstanbul: Papatya Yayıncılık Eğitim.
- Gümüşsoy, Ç. A., & Yeterel, A. C. (2016). "Fırsat Sitelerinden Tekrar Satın Alma Kararını Etkileyen Faktörlerin Araştırılması". *Bilişim Teknolojileri Dergisi*, 9(3), 275-284.

- Güneş, A. (2011). *Paralel Programlama ile El Yazısı Rakamlarının Tanınması*. (Yayımlanmış Yüksek Lisans Tezi), Isparta: Süleyman Demirel Üniversitesi Fen Bilimleri Enstitüsü, Bilgisayar Mühendisliği.
- Güngördü, Z., Oflazer, K. (1995). "Parsing Turkish using the Lexical Functional Grammar Formalism" [Türkçeyi Lexical Fonksiyonel Gramer Biçimciliğini Kullanarak Ayırıştırma]. *Machine Translation* 10(4) 515–544.
- Güran, A., Uysal, M., & Doğrusöz, Ö. (2014). "Destek Vektör Makineleri Parametre Optimizasyonunun Duygu Analizi Üzerindeki Etkisi". *Dokuz Eylül Üniversitesi Mühendislik Fakültesi Fen ve Mühendislik Dergisi*, 16(48), 86-93.
- Gürsakal, N. (2014). *Büyük Veri*, Bursa: Dora Basım Yayın.
- Güven, A., Bozkurt, Ö. Ö., & Kalıpsız, O. (2007). "*Veri Madenciliğinin Geleceği*" [Bildiri]. The IX. Akademik Bilişim Konferansı, Kütahya.
- Hacıfendioğlu, Ş. (2014). "Sosyal Medyanın Marka Bağlılığına Etkisi Üzerine bir Araştırma". *Kocaeli Üniversitesi Sosyal Bilimler Dergisi*, 28, 59-70.
- Hajič, J., Hajičová, E., Mikulová, M., & Mirovský, J. (2017). "Prague Dependency Treebank" [Prag Bağımlı Treebank]. In *Handbook of Linguistic Annotation* (pp. 555-594). Springer, Dordrecht.
- Hajic, J., Homola, P., ve Kubon, V., (2003). "A Simple Multilingual Machine Translation System" [Bildiri]. MT Summit IX, New Orleans, USA.
- Hajic, J., Smrz, O., Zemanek, P., Snajdauf, J., Beska, E. (2004). "Prague Arabic Dependency Treebank: Development in Data and Tools" [Bildiri]. *International Conference on Arabic Language Resources and Tools*, Cairo, Egypt.
- Hamzaoglu, İ., (1993). *Machine Translation From Turkish To Other Turkic Languages And An Implementation For The Azeri Languages*. (Yayımlanmış Yüksek Lisans Tezi), İstanbul: Boğaziçi Üniversitesi.
- Hallaç, İ. R. (2014). *Büyük Veri Analizinde Dağıtık Makine Öğrenmesi Algoritmalarının Kullanılması*. (Yayımlanmış Yüksek Lisans Tezi), Elazığ: Fırat Üniversitesi, Fen Bilimleri Enstitüsü, Bilgisayar Mühendisliği.

- Hedlund, B. (2011). *Understanding Hadoop Clusters and the Network*. <http://bradhedlund.com/2011/09/10/understanding-hadoop-clusters-and-the-network/>
- Hoche, S., Ingolf Geist, Lourdes Peña Castillo, & Schulz, N. (2006). *Databases, Knowledge Discovery, Information Retrieval, and Web Mining*. In *CIGR Handbook of Agricultural Engineering* (Vol. 6). Amerika: American Society of Agricultural and Biological Engineers.
- İlter, H. (2012). *Hadoop 1.0.3 Kurulumu Adım Adım*. <http://devveri.com/hadoop/hadoop-kurulumu-adim-adim>
- İlter, H. (2013). *Pig ve Hive ile Hadoop Üzerinde Veri Analizi*. <http://devveri.com/hadoop/pig-ve-hive-ile-hadoop-uzerinde-veri-analizi>
- Jayalakshmi, D. S., & Srinivasan, R. (2017). "An Analytical Model for HDFS Replication Pipeline based on Queuing Theory" [Kuyruklama Teorisine Dayalı HDFS Çoğaltma Boru Hattı için Analitik Bir Model]. *International Journal of Applied Engineering Research*, 12(24), 15158-15168.
- Jiang, L., Yu, M., Zhou, M., Liu, X. ve Zhao, T. (2011). "Target-Dependent Twitter Sentiment Classification" [Bildiri]. *The 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 1, 151–160.
- Jiang, S., Pang, G., Wu, M., Kuang, L., 2012. "An Improved K-Nearest-Neighbor Algorithm For Text Categorization" [Metin Sınıflandırmasında Geliştirilmiş Bir K-En Yakın Komşu Algoritması]. *Expert Systems with Applications*, 39(1), 1503-1509.
- Kanaris, I., Kanaris, K., Houvardas, I., Stamatatos, E., (2007). "Words Versus Character NGgrams for Anti-Spam Filtering" [Anti-Spam Filtreleme için Karakter NGgramlarına Karşı Kelimeler]. *International Journal on Artificial Intelligence Tools*, 16(06), 1047-1067.
- Kapçı, Ş. S. (2015). *Hadoop Nedir?* <http://sarifil.blogspot.com.tr/2015/04/hadoop-nedir.html>

- Karabay, B. & Ulaş, M. (2018), “Büyük Veri İşlemede Yaygın Kullanılan Araçların Karşılaştırılması” [Bildiri] 8. *International Advanced Technologies Symposium*, Elazığ.
- Karaca, M. F., & Görgünoğlu, S. (2016). "Türkçeden Türk İşaret Dili'ne". *İleri Teknoloji Bilimleri Dergisi*, 5(1), 15-27.
- Karakulakoğlu, S. E., & Uğurlu, Ö. (2015). *İletişim Çalışmalarında Dijital Yaklaşımlar* (Vol. 1): Ankara: Tercan Matbaacılık.
- Karataş, G. (2015). *Dağıtık Veritabanları için Test Uygulaması*. (Yayımlanmış Yüksek Lisans Tezi), İstanbul: İstanbul Kültür Üniversitesi Fen Bilimleri Enstitüsü, Bilgisayar Mühendisliği.
- Kaşıkcı, T., Gökçen, H. (2014). "Metin Madenciliği ile E-Ticaret Sitelerinin Belirlenmesi". *International Journal of Informatics Technologies*, 7(1), 25-32.
- Katal, A., Wazid, M., & Goudar, R. H. (2013). "Big data: issues, challenges, tools and good practices" [Bildiri]. *2013 Sixth International Conference*, Noida, India.
- Kauffman, R.J., Lai, H. & Ho, C. (2010) “Incentive Mechanisms, Fairness And Participation In Online Group-Buying Auctions” [Çevrimiçi Grup Satın Alma Müzayedelerinde Teşvik Mekanizmaları, Dürüstlük ve Katılım], *Electronic Commerce Research and Applications*, 9, 249-262.
- Kavasoğlu, F. Z. (2013). *Kullanıcı Tarayıcı Geçmişine Dayanarak Müşteri Yorumlarının Özetlenmesi*. (Yayımlanmış Yüksek Lisans Tezi), İstanbul: İstanbul Teknik Üniversitesi Fen Bilimleri Enstitüsü, Bilgisayar Mühendisliği.
- Kayabay, K., Gökalp, M. O., Akyol, M. A., Eren, P. E., & Koçyiğit, A. (2016). “Geleceğin Kuruluşları için Büyük Veri: Mevcut Durum ve Eğilimler” [Bildiri], *3rd International Management Information Systems Conference*, İzmir.
- Kayahan, D., Sergin, A., & Diri, B. (2013). "Determination of TV Programme Ratings by Twitter" [Bildiri]. *The Signal Processing and Communications Applications Conference (SIU)*, Haspolat.
- Kayım, F. (2015). *K-Means İle DbSCAN Algoritmasının Parallellleştirilmesi ve Hadoop Üzerinde Büyük Veri Analizinde Kullanılması, Performans ve Yeterlilik*

- Karşılaştırması*. (Yayımlanmış Yüksek Lisans Tezi), İstanbul: Beykent Üniversitesi Fen Bilimleri Enstitüsü, Bilgisayar Mühendisliği.
- Keskin, S., & Baş, M. (2016). "Sosyal Medyanın Tüketici Davranışları Üzerine Etkisinin Belirlenmesi". *İktisadi ve İdari Bilimler Fakültesi Dergisi*, 17(3), 51-69.
- Kıraç, M. (2015). *Büyük Veri Yazılımlarının Donanım İmtihanı*. <https://tr.linkedin.com/pulse/büyük-veri-yazılımlarının-donanım-ımtihanı-mustafa-kıraç>, Erişim Tarihi: 25.09.2018
- Ko, Y. (2012). "A Study Of Term Weighting Schemes Using Class Information For Text Classification". *The 35th International ACM SIGIR Conference On Research And Development In Information Retrieval*, USA.
- Koca, M. (2014). *The Costs and Benefits of Turning Data Into Information Using Big Data Systems*. (Yayımlanmış Yüksek Lisans Tezi), İstanbul: Özyeğin Üniversitesi Fen Bilimleri Enstitüsü, Bilgisayar Mühendisliği.
- Kocabaş, İ. (2016). "Halkla İlişkiler Aracı Olarak Sosyal Medya: Arçelik ve Bosch Markalarının Kurumsal Facebook ve Twitter Hesapları Üzerine Bir İnceleme". *Selçuk İletişim*, 9(2), 69-91.
- Köseoğlu, Ö., & Demirci, Y. (2017). "Türkiye'de Büyük Veri ve Veri Madenciliğine İlişkin Politika ve Stratejiler: Ulusal Politika Belgelerinin İçerik Analizi." *Suleyman Demirel University Journal of Faculty of Economics & Administrative Sciences*, 22(15), 2223-2239.
- Kromann, M.T. (2003). "The Danish Dependency Treebank and the DTAG treebank tool" [Bildiri]. *Second Workshop on Treebanks and Linguistic Theories*, Vaxjo, Sweden.
- Kudo, T., Matsumoto, Y. (2000). "Japanese Dependency Analysis Based On Support Vector Machines" [Bildiri]. *Joint Sigdat Conference On Empirical Methods In Natural Language Processing and Very Large Corpora*, Hong Kong.
- Kudo, T., Matsumoto, Y. (2002). "Japanese Dependency Analysis Using Cascaded Chunking" [Bildiri]. *Sixth Conference on Natural Language Learning*, Taipei, Taiwan.

- Kuş-Khalilov, M. C., & Gündebahar, M. (2014). "Bankacılıkta Büyük Veri Uygulamaları: Bir İnceleme" [Bildiri]. The XVI. Akademik Bilişim, Mersin.
- Kwak, H., Lee, C., Park, H., & Moon, S. (2010). "What is Twitter, a Social Network or a News Media?" [Bildiri]. *19th International Conference on World Wide Web*, New York, USA
- Lammel, R. (2008). "Google's MapReduce Programming Model" [Google'un Eşle İndirge Programlama Modeli]. *Science Of Computer Programming*, 70(1), 1-30.
- Laney, D. (2001). *Application Delivery Strategies*. <http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf>
- Laudon, K. C., & Laudon, J. P. (2004). *Management Information Systems: Managing the Digital Firm* (Vol. 8). Amerika: Global Edition.
- Leung, K. M. (2007). *Naive Bayesian Classifier*. <https://pdfs.semanticscholar.org/4632/2a2c113db0e28dd1b37aeabd5668adf77de6.pdf>
- Lewis, L. & Chadd, O. (2018). What is happening in one minute of Internet in 2018? <https://www.devdiscourse.com/Article/agency-wire/8337-what-is-happening-in-one-minute-of-internet-in-2018>
- Limet, S., Merlo, A., & Spalazzi, L. (2018). "HPC & Co strike back: Where are Distributed Paradigms Heading Toward?" [HPC & Co Karşılık Verir: Dağıtılmış Paradigmalar Nereye Doğru Gidiyor?]. *Concurrency and Computation: Practice and Experience*, 30(9), 1-13.
- Lin, Y. S., Jiang, J. Y., Lee, S. J. (2014). "A Similarity Measure For Text Classification and Clustering" [Metin Sınıflandırması ve Kümeleme için Benzerlik Ölçümü]. *Knowledge and Data Engineering*, 26(7), 1575-1590.
- Liu, B. (2012). *Sentiment Analysis and Opinion Mining*: Chicago: Morgan and Claypool Publishers.
- Markovitz, D. (2017). *Hive Tutorial*. <https://cwiki.apache.org/confluence/display/Hive/Tutorial>

- Marr, B. (2017). *Büyük Veri İş Başında* (Vol. 1): İstanbul: Media Cat Kitapları.
- McAfee, A., & Brynjolfsson, E. (2012). *Big Data: The Management Revolution*, Harvard Business.
- Medhat, W., Hassan, A., Korashy, H. (2014). "Sentiment Analysis Algorithms And Applications: A Survey" [Duygu Analizi Algoritmaları ve Uygulamaları: Bir Anket], *Ain Shams Engineering Journal*, 5(4),1093-1113.
- Meral, M., & Diri, B. (2014). "Twitter Üzerinde Duygu Analizi" [Bildiri]. *The IEEE 22. Sinyal İşleme ve İletişim Uygulamaları Kongresi*, Trabzon.
- Michelson, M., & Macskassy, S. A. (2010). "Discovering Users' Topics of Interest on Twitter: a First Look" [Bildiri]. *The Fourth Workshop on Analytics for Noisy Unstructured Text Data*, Toronto, Canada.
- Ming, H., Rohrdantz, C., Janetzko, H. and Dayal, U., Keim, DA., Haug, L., ve diğerleri. (2011). "Visual Sentiment Analysis On Twitter Data Streams" [Bildiri]. *2011 IEEE Conference*, Providence, RI.
- Mulki, H., Haddad, H., Ali, C. B., & Babaoğlu, İ. (2018). "Preprocessing Impact On Turkish Sentiment Analysis" [Bildiri]. *26th Signal Processing and Communications Applications Conference (SIU)*, İzmir, Türkiye.
- Nalçakan, Y., Bayramoğlu, Ş. S., & Tuna, S. (2014). *Sosyal Medya Verileri Üzerinde Yapay Öğrenme ile Duygu Analizi Çalışması*. https://www.researchgate.net/publication/280938376_Sosyal_Medya_Verileri_Uzerinde_Yapay_Ogrenme_ile_Duygu_Analizi_Calismasi
- Narayanan, V. (2014). "Using Big-Data Analytics to Manage Data Deluge and Unlock Real-Time Business Insights" [Veri Tufanını Yönetmek ve Gerçek Zamanlı İşletme Bilgilerini Kilitlemek için Big-Data Analytics'i Kullanmak]. *The Journal of Equipment Lease Financing*, 32(2), 1-6.
- Neo4j. (2017). *Building a Recommendation Engine with Cypher in Two Minutes*. <https://neo4j.com/developer/guide-build-a-recommendation-engine/>
- Nguyen, L. T., Wu, P., Chan, W., Peng, W., & Zhang, Y. (2012). "Predicting Collective Sentiment Dynamics From Time-Series Social Media" [Bildiri]. *The Proceedings*

of the first international workshop on issues of sentiment discovery and opinion mining, Çin.

- Nigam, K., McCallum, A. K., Thrun, S., Mitchell, T., 2000. Text Classification from Labeled and Unlabeled Documents Using EM" [EM Kullanarak Etiketli ve Etiketsiz Belgelerde Metin Sınıflandırması]. *Machine Learning*, 39(2-3)
- Nivre, J. (2006). *Inductive Dependency Parsing*. (Doktora Tezi) İsveç: Vaxjo University.
- Nivre, J. ve Nilsson, J. (2005). "Pseudoprojective Dependency Parsing" [Bildiri]. *The 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, Ann Arbor, Michigan.
- Nivre, J., Hall, J. ve Nilsson, J. (2004). "Memorybased dependency parsing" [Bildiri]. *8th Conference on Computational Natural Language Learning*, Boston MA.
- Nizam, H., & Akın, S. S. (2014). "Sosyal Medyada Makine Öğrenmesi ile Duygu Analizinde Dengeli ve Dengesiz Veri Setlerinin Performanslarının Karşılaştırılması" [Bildiri]. *The XIX. Türkiye'de İnternet Konferansı*, İzmir.
- Nvi. (2017). *İl ve İlçe Listesi*. www.nvi.gov.tr/PublishingImages/Pages/il-ilce-kod-tablosu/IL_ILCE_LISTESI.xls&usg=AFQjCNGhNZO1se7w2SJ6VRrthACW67NJSw
- Nwankpa, J. K. (2018). "ERP Systems Benefit Realization and the Role of ERP-Enabled Application Integration" [ERP Sistemleri Fayda Gerçekleştirilmesi ve ERP-Etkin Uygulama Entegrasyonu'nun Rolü]. *Encyclopedia of Information Science and Technology*, 4, 2953-2964.
- Oflazer, K., & Bozşahin, H. C. (1994). "Türkçe ve Doğal Dil İşleme" [Bildiri]. *The Proc. of Turkish Informatics Society TBD'94*, Ankara.
- Oğul, B. B., & Ercan, G. (2016). "Sentiment Classification on Turkish Hotel Reviews" [Bildiri]. *The 24th Signal Processing and Communication Application Conference (SIU) Zonguldak*.
- Oğuzlar, A. (2011). *Temel Metin Madenciliği*. Bursa: Dora.
- Onan, A. (2017). "Türkçe Twitter Mesajlarında Gizli Dirichlet Tahsisine Dayalı Duygu Analizi" [Bildiri]. 19. Akademik Bilişim Konferansı, Aksaray.

- Oracle. (2015). *A Relational Database Overview*.
<http://docs.oracle.com/javase/tutorial/jdbc/overview/database.html>
- Özdeş, M. (2017). *Büyük Veri Araçlarını Kullanarak Duygu Analizi Gerçekleştirimi* (Yayımlanmış Yüksek Lisans Tezi), Denizli: Pamukkale Üniversitesi, Fen Bilimleri Enstitüsü, Bilgisayar Mühendisliği.
- Özdoğan, O. (2013). *Büyük Veri (Big Data) Nedir?*
<http://www.oganozdogan.com/2013/07/buyuk-veri-big-data-nedir.html>
- Özdoğan, O. (2016). *Büyük Veri Denizi* (Vol. 1): Ankara: Elma Yayınevi.
- Özenç, B. (2017). *Morphological Analyser For Turkish*. (Yayımlanmış Yüksek Lisans Tezi), İstanbul: Işık Üniversitesi, Fen Bilimleri Enstitüsü, Bilgisayar Mühendisliği.
- Özmen, H. İ., & Villi, B. (2014). "Sosyal Medya ve Finansal Performans: Borsa İstanbul'da İşlem Gören İşletmeler Üzerinde Bir Araştırma". *Abant İzzet Baysal Üniversitesi Sosyal Bilimler Enstitüsü Dergisi*, 14, 269-293.
- Öztürk, M. (2002). *Bilgi Yönetimi: Rekabet Üzerindeki Etkileri ve Bir Uygulama*. (Yayımlanmış Yüksek Lisans Tezi), İstanbul: İstanbul Üniversitesi Sosyal Bilimler Enstitüsü, İşletme.
- Öztürk, M.B. (2016). *Türkçede Morfolojik Analiz Yapan Bir Sistemin Morfolojik Türetme İçin Kullanılması*. (Yayımlanmış Yüksek Lisans Tezi), İstanbul: Hacettepe Üniversitesi, Bilgisayar Mühendisliği.
- Öztürk, S., & Atmaca, H. E. (2017). "İlişkisel ve İlişkisel Olmayan (Nosql) Veri Tabanı Sistemleri Mimari Performansının Yönetim Bilişim Sistemleri Kapsamında İncelenmesi." *Bilişim Teknolojileri Dergisi*, 10(2), 199-209.
- Pak, A., Paroubek, P. (2010). "Twitter as a Corpus for Sentiment Analysis and Opinion Mining" [Duygu Analizi ve Görüş Madenciliği için bir Yapı Olarak Twitter]. *LREC*, 10, 1320-1326.
- Pang, B., Lee, L., & Vaithyanathan, S. (2002). "Thumbs up?: Sentiment Classification Using Machine Learning Techniques" [Bildiri]. *The Proceedings of the ACL-02 Conference on Empirical Methods In Natural Language Processing*, USA.

- Parlak, C., & Diri, B. (2013). "İnsan Sesinden Duygu Çıkarma" [Bildiri]. *The 21st Signal Processing and Communications Applications Conference (SIU)*, Haspolat.
- Parmar, R. R., & Roy, S. (2018). "MongoDB as an Efficient Graph Database: An Application of Document Oriented NOSQL Database" [Verimli bir Grafik Veri Tabanı Olarak MongoDB: Belge Yönelimli bir NOSQL Veri Tabanı Uygulaması]. *Data Intensive Computing Applications for Big Data*, 29, 331.
- Pawlak, Z. (2003). "A Rough Set View on Bayes Theorem" [Bayes Teoremine kabaca bir bakış]. *International Journal of Intelligent Systems*, 18(5), 487-498.
- Pirayani, R., Madhavi, D., Singh, V.K. (2017). "Analytical Mapping of Opinion Mining and Sentiment Analysis Research During 2000-2015" [2000-2015 Döneminde Görüş Madenciliğinin ve Duygu Analizi Araştırmalarının Analitik Haritalaması], *Information Processing and Management*, 53 (1), 122-150.
- Pol, U. R. (2016). "Big Data Analysis: Comparison of Hadoop Mapreduce and Apache Spark" [Büyük Veri Analizi: Hadoop Mapreduce ve Apache Spark Karşılaştırması]. *International Journal of Engineering Science*, 6(6), 6389-6391.
- Poyraz, D. (2015). *Bigdata Analytics Architectures For Hvac Energy Optimization Systems*. (Yayımlanmış Yüksek Lisans Tezi), Ankara: Orta Doğu Teknik Üniversitesi.
- Prabowo, R., Thelwall, M., 2009. "Sentiment Analysis: a Combined Approach" [Duygu Analizi: Kombine Bir Yaklaşım]. *Journal of Informetrics*, 3(2), 143-157.
- Prensky, M. (2001). "Digital Natives, Digital Immigrants Part 1". *On the horizon*, 9(5), 1-6.
- Prytherch, R. (2016). *Harrod's librarians' Glossary And Reference Book: a Directory of Over 10,200 Terms, Organizations, Projects And Acronyms in The Areas of Information Management, Library Science, Publishing And Archive Management*. Routledge.
- Quercia, D., Ellis, J., Capra, L. & Crowcroft, J. (2011). "In the Mood for Being Influential on Twitter" [Bildiri]. *2011 IEEE Third International Conference*, Boston, MA.

- Rometty, G. (2014). *Future of Data Hortonworks Data Platform and Hortonworks Data Flow*. [https://www-01.ibm.com/events/wwe/grp/grp310.nsf/vLookupPDFs/The%20Power%20of%20Open%20Montreal%20Hortonworks/\\$file/The%20Power%20of%20Open%20Montreal%20Hortonworks.pdf](https://www-01.ibm.com/events/wwe/grp/grp310.nsf/vLookupPDFs/The%20Power%20of%20Open%20Montreal%20Hortonworks/$file/The%20Power%20of%20Open%20Montreal%20Hortonworks.pdf)
- Rouse, M. (2014). *Big Data*. <http://searchcloudcomputing.techtarget.com/definition/big-data-Big-Data>
- Sadredini, E., Guo, D., Bo, C., Rahimi, R., Skadron, K., & Wang, H. (2018). "A Scalable Solution for Rule-Based Part-of-Speech Tagging on Novel Hardware Accelerators" [Bildiri]. In Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, London.tera
- Sakınmaz, S. (2016a). *Hadoop HDFS Komutları*. <http://www.buyukveri.co/hadoop-hdfs-komutlari/>
- Sakınmaz, s. (2016b). *Hadoop Nedir?* <http://www.buyukveri.co/hadoop-nedir/>
- Salman G., & Eroğlu, E. (2017). "Sosyal Medyada Kurum Kimliği Yönetimi: Türkiye'nin En Büyük 500 Şirketi Üzerine Bir Araştırma", *Gümüşhane Üniversitesi İletişim Fakültesi Elektronik Dergisi*, 1(5), 525-547.
- Sarıbıyık, H. (2015). *NoSQL'in Kısa Tarihi*. <http://devveri.com/nosql/nosqlin-kisa-tarihi>
- Sarıgül, K. (2009). *Türkçe Etkisiz Kelimeler (Stop Words) Listesi 1.1*. <http://www.turkceogretimi.com/Genel-Konular/article/541-turkce-etkisiz-kelimeler-stop-words-listesi-11/35#>
- Sarıkaya, F. (2017). *Apache Hive 2.0.0 Kurulumu*. <http://ferhatsarikaya.com/apache-hive-2-0-0-kurulumu/>
- Sarkar, D., Bali, R., & Sharma, T. (2018). *Machine learning basics. In Practical Machine Learning with Python*. Apress, Berkeley, CA.
- Schneider, R. (2012). *Custom Hadoop for Dummies* (Vol. 2). Amerika: John Wiley & Sons Incorporated.
- Schumaker, R.P., Jarmoszko, A.T., Labedz, C.S. (2016). "Predicting Wins and Spread in the Premier League Using a Sentiment Analysis of Twitter" [Twitter'ın Bir Duygu

- Analizi Kullanılarak Premier Ligi Dağılımının ve Kazananların Tahmini] *Decision Support Systems*, 88, 76-84
- Mayer-Schönberger, V., & Cukier, K. (2013). *Büyük Veri - Yaşama, Çalışma ve Düşünme Şeklimizi Dönüştürecek Bir Devrim*. Çev. Banu Erol. İstanbul: Paloma.
- Schutt, R., & O'Neil, C. (2013). *Doing Data Science: Straight Talk From the Frontline*: Amerika: O'Reilly Media, Inc.
- Seagate, (2018). *Data Age 2025*. <https://www.seagate.com/tr/tr/our-story/data-age-2025/>
- Sekine, S., Uchimoto, K. ve Isahara, H. (2000). "Backward Beam Search Algorithm for Dependency Analysis of Japanese" [Bildiri]. *17th International Conference on Computational Linguistics*, Saarbrücken.
- Selçuk, A. (2015). *Private Search Over Big Data Leveraging Distributed File System And Parallel Processing*. (Yayımlanmış Yüksek Lisans Tezi), İstanbul: Sabancı Üniversitesi Fen Bilimleri Enstitüsü, Bilgisayar Mühendisliği.
- Serdar, İ. (2013). *TDK İmla Kılavuzu*. <http://www.excelgurusu.com/tdk-imla-kilavuzu/>
- Shvachko, K., Kuang, H., Radia, S., & Chansler, R. (2010). "The Hadoop Distributed File System" [Bildiri]. *Mass Storage Systems and Technologies (MSST), 2010 IEEE 26th Symposium*, Washington DC, USA.
- Sommer, S., Schieber, A., Hilbert, A., Heinrich, K. (2011). "Analyzing Customer Sentiments In Microblogs—A Topic-Model-Based Approach For Twitter Datasets" [Bildiri]. *Proceedings of the Seventeenth Americas Conference on Information Systems*, Detroit, Michigan.
- Subbian, K., Melville, P. (2011). "Supervised Rank Aggregation for Predicting Influencers in Twitter"[Bildiri]. *2011 IEEE Third International Conference*, Boston, MA.
- Sur, S., Wang, H., Huang, J., Ouyang, X., & Panda, D. K. (2010). "Can High-Performance Interconnects Benefit Hadoop Distributed File System" [Bildiri]. *The Workshop on Micro Architectural Support for Virtualization, Data Center Computing and Clouds*, Tayvan.

- Szomszor, M., Kostkova, P., & De Quincey, E. (2010). "Swineflu: Twitter Predicts Swine Flu Outbreak in 2009". *International Conference on Electronic Healthcare*, 69, 18-26.
- Şimşek, M. U., & Özdemir, S. (2012). "Analysis of the Relation Between Turkish Twitter Messages and Stock Market Index" [Bildiri]. *The 6th International Conference: Application of Information and Communication Technologies (AICT)*, Tbilisi, Georgia.
- Tan, A.H. (1999). "Text Mining: The State of the Art and the Challenges" [Bildiri]. *The Proceedings of the PAKDD 1999 Workshop on Knowledge Discovery from Advanced Databases*, Singapore.
- Tantuğ, A. C., Adalı, E., & Oflazer, K. (2010). "Türkmenceden Türkçeye Bilgisayarlı Metin Çevirisi". *İtühergisi*, 7(4), 83-94.
- Taslitz, A. E. (2012). "Information Overload, Multi-Tasking, and the Socially Networked Jury: Why Prosecutors Should Approach the Media Gingerly" [Bilgi Aşırı Yüklenmesi, Çoklu Görevlendirme ve Sosyal Ağ Üzerindeki Jüri: Neden Savcılar Medyaya Zekice Yaklaşmalılar?]. *Journal of the Legal Profession*, 37, 89-138.
- TDK. (2017). *Türk Dil Kurumu*. <http://www.tdk.gov.tr/>
- Tekin, M., Güleş, H. K., & Ögüt, A. (2006). *Değişim Çağında Teknoloji Yönetimi*. Ankara: Nobel Yayın Dağıtım.
- Thelwall, M., Buckley, K., Paltoglou, G., Cai, D. ve Kappas, A. (2010). "Sentiment Strength Detection In Short Informal Text" [Kısa Bilgi Metninde Duygu Gücü Algılama], *Journal of the American Society for Information Science and Technology*, 61(12), 2544–2558.
- Toffler, A. (2008). *3. Dalga: Bir Futürist Ekonomi Analizi Klasığı*: İstanbul: Koridor Yayınları.
- Türkmenoglu, C., & Tantug, A. C. (2014). "Sentiment Analysis in Turkish Media" [Bildiri]. *The International Conference on Machine Learning*, Çin.

- Türksever, B. (2013). *Küreselleşme Etkisi Altında Ötekileşen Ana Dil*.
http://www.academia.edu/download/35028045/KURESELLESME_ETKISI_AL_TINDA_OTEKILESEN_ANADIL.docx
- TwitLonger. (2017). *Sign in to Start Writing Longer Posts with TwitLonger*.
<http://www.twitlonger.com/>
- Twitter. (2016a). *How Your Business Can Use Twitter*.
<https://business.twitter.com/en/basics/intro-twitter-for-business.html>
- Twitter. (2016b). *İşte Olup Bitenler*. <https://about.twitter.com/tr/company>
- Twitter. (2017a). *Create Your Twitter Content Strategy*.
<https://business.twitter.com/en/basics/what-to-tweet.html>
- Twitter. (2017b). *How Twitter Can Help Your Business*.
<https://business.twitter.com/en/basics.html>
- Uçan, A. (2014). *Otomatik Duygu Sözlüğü Çevirimi ve Duygu Analizinde Kullanımı*. (Yayımlanmış Yüksek Lisans Tezi), Ankara: Hacettepe Üniversitesi Fen Bilimleri Enstitüsü, Bilgisayar Mühendisliği.
- Uzkurt, C. (2010). "İnovasyon Yönetimi: İnovasyon Nedir, Nasıl Yapılır ve Nasıl Pazarlanır". *Ankara Sanayi Odası Yayın Organı*, 39, 37-51.
- Uzun, E., Buluş, H.N., & Erdoğan, C. (2018). "Veritabanı Tasarımının Yazılım Performansına Etkisi: Normalizasyona Karşı Denormalizasyon" *Süleyman Demirel Üniversitesi Fen Bilimleri Enstitüsü Dergisi*, 22(2), 887 - 895.
- Uzunbayır, S. (2015). *A Comparison Between Relational Database Models and Nosql Trends on Big Data Design Challenges Using A Social Shopping Application*. (Yayımlanmış Yüksek Lisans Tezi), İzmir: İzmir Ekonomi Üniversitesi Fen Bilimleri Enstitüsü, Bilgisayar Mühendisliği.
- Ünalır, M. O., Inan, E., Yönyül, B., Olca, E., Sentürk, F., Mostafapour, ve diğerleri. (2015). "Veri Yogun Bilgi Sistemleri İçin Melez Bir Veri Mimarisi Önerisi" [Bildiri]. *The Ulusal Yazılım Mühendisliği Sempozyumu*, Bornova, İzmir.

- Ünür, E. (2016). "Geleneksel Medya'nın Sosyal Medyayı Nasıl Kullandığına İlişkin Bir İnceleme: Dizilerin Twitter Kullanımı". *Turkish Journal of Social Research*, 20(1), 1-11.
- Valz, D. (2013). *Rules of Engagement – Key Value Stores*. <http://www.ingenioussql.com/tag/key-value-store/>
- Vikisözlük. (2017). *Viki Sözlük*. <https://tr.wiktionary.org/wiki/Anasayfa>
- Vinod, B. (2013). "Leveraging Big Data for Competitive Advantage in Travel" [Seyahatte Rekabet Avantajı için Büyük Verileri Kaldırabilme]. *Journal of revenue and pricing management*, 12(1), 96-100.
- Vural, A. G., Cambazoglu, B. B., Senkul, P., & Tokgoz, Z. O. (2013). *A framework for Sentiment Analysis in Turkish: Application to Polarity Detection of Movie Reviews in Turkish*. In *Computer and Information Sciences III* (pp. 437-445): Springer.
- Ward, M. (2013). "NoSQL Database in the Cloud: MongoDB on AWS" [Bulutta NoSQL Veritabanı: AWS'de MongoDB], *Amazon Web Services E-Journal*, 2-22
- WANG, L. (2005). "Support Vector Machines: Theory and Applications" [Destek Vektör Makineleri: Teori ve Uygulamalar], *Springer*, New York, 1434-9922.
- Weng, j., Lim, E.P., Jiang, J. & He, Q. (2010). "Twitterrank: Finding topic-sensitive influential twitterers" [Bildiri]. *The Third ACM International Conference on Web Search and Data Mining*, New York, USA.
- Yamada, H. ve Matsumoto, Y. (2003). *"Statistical Dependency Analysis With Support Vector Machines"* [Bildiri]. *8th International Workshop of Parsing Technologies*, Nancy.
- Yelmen, İ. (2016). *Doğal Dil İşleme Yöntemleriyle Türkçe Sosyal Medya Verileri Üzerinde Duygu Analizi*. (Yayımlanmış Yüksek Lisans Tezi), İstanbul: İstanbul Aydın Üniversitesi, Fen Bilimleri Enstitüsü, Bilgisayar Mühendisliği.
- Yıldırım, G., Aydın, G., Alli, H., & Tatar, Y. (2014). "Hadoop ile Kaos Temelli FCW Optimizasyon Algoritmasının Analizi An Analysis of Chaos-Based the FCW

Optimization Algorithm by Hadoop" *Elektrik – Elektronik – Bilgisayar ve Biyomedikal Mühendisliği Sempozyumu Bursa*, 708-712.

Yıldız, U. A., & Topal, S. (2015). "Büyük Veri Kahramanı Veri Bilimci". *Bilim ve Teknik* 76-79.

Yıldız, O. T., Solak, E., Görgün, O., & Ehsani, R. (2014). "Constructing a Turkish-English Parallel Treebank." [Türkçe-İngilizce Paralel Treebank Yapımı]. *52nd Annual Meeting of the Association for Computational Linguistics*, 2(2), 112-117.

Yılmaz, M. (2009) "Enformasyon ve Bilgi Kavramları Bağlamında Enformasyon Yönetimi ve Bilgi Yönetimi." *Ankara Üniversitesi Dil ve Tarih-Coğrafya Fakültesi Dergisi*, 49(1), 95-118.

Yücel, İ. H. (1997). *Bilim-Teknoloji Politikaları ve 21. Yüzyılın Toplumu*: Ankara: Devlet Planlama Teşkilatı. Sosyal Sektörler ve Koordinasyon Genel Müdürlüğü, Araştırma Dairesi Başkanlığı.

Yücel, N. (2013). "Müşteri İlişkileri Yönetimi'nde Yeni Bir Anlayış: Sosyal Müşteri İlişkileri Yönetimi". *The Journal of Academic Social Science Studies*, 6(1), 1641-1656.

Yüksel, A. S. ve Tan, F. G. (2018). "Metin Madenciliği Teknikleri ile Sosyal Ağlarda Bilgi Keşfi". *Journal of Engineering Sciences*, 6(2), 324-333.

Yüzügüldü, M. Z. (2015). *İçerik Tabanlı Resim Arama Motoru*. (Yayımlanmış Yüksek Lisans Tezi), İstanbul: Yıldız Teknik Üniversitesi Fen Bilimleri Enstitüsü, Elektrik ve Elektronik Mühendisliği.

Zaharia M, Chowdhury M, Das T, Dave A, Ma J, McCauley M, Franklin MJ, Shenker S, Stoica I. (2012). "Resilient distributed datasets: a fault-tolerant abstraction for in-memory cluster computing" [Bildiri], *9th USENIX Conference on Networked Systems Design and Implementation*, USA.

Zemberek. (2006). *NZemberek Sürüm 0.1.*
<http://zembereknlp.blogspot.com.tr/2006/12/nzemberek-srm-01.html>

ÖZGEÇMİŞ

Kişisel Bilgiler	
Adı Soyadı	Sinan KUL
Doğum Yeri ve Tarihi	Erzurum/14.01.1984
Eğitim Durumu	
ÖnLisans Öğrenimi	Atatürk Üniversitesi Erzurum Meslek Yüksek Okulu Bilgisayar Programcılığı
Lisans Öğrenimi	Erciyes Üniversitesi Mühendislik Fakültesi Bilgisayar Mühendisliği
Y.Lisans Öğrenimi	Atatürk Üniversitesi Mühendislik Fakültesi Bilgisayar Mühendisliği
Bildiği Yabancı Diller	İngilizce
Bilimsel Faaliyetler	Eserler: A. Ulusal bilimsel toplantılarda sunulan ve bildiri kitaplarında basılan bildiriler: 1. KUL Sinan, ÖZEN Üstün (2015). Öğrenci Bilgi Sisteminde Ders Alma Süreçlerinin İyileştirilmesi. Yönetim Bilişim Sistemleri Kongresi, 547-554. (Tam Metin Bildiri/)(Yayın No:2292175) 2. KUL Sinan, KARAMAN Ersin, ÖZEN Üstün (2015). Genetik Algoritma ile Sınav Takvimi Otomasyonu. Yönetim Bilişim Sistemleri, 579-591. (Tam Metin Bildiri/)(Yayın No:2291999) B. Uluslararası bilimsel toplantılarda sunulan ve bildiri kitaplarında (proceedings) basılan bildiriler: 1. Kul Sinan, Özen Üstün (2015). The Investigation Of Factors Affecting Computer Programming Students Achievement. Building Bridge on Sciences (Özet Bildiri/)(Yayın No:2291630) 2. Kul Sinan, Dal Deniz, Aydın Tolga (2015). Various aspects of computer engineering undergraduate education in Turkey. 2015 10th International Conference on Computer Science & Education (ICCSE), 249-252., Doi: 10.1109/ICCSE.2015.7250251 (Tam Metin Bildiri/)(Yayın No:2291000) 3. Demirdöğen Osman, Erdal Hamit, Kul Sinan (2016). Dağıtım Merkezi Yer Seçimi Problemine Stokastik Bir Model Önerisi Tra Bölgesinde Bir uygulama. II. Uluslararası Kafkasya-Orta Asya Dış Ticaret ve Lojistik Kongresi (Tam Metin Bildiri/)(Yayın No:2963977). D. Ulusal hakemli dergilerde yayımlanan makaleler 1. KESKINKILIÇ Mustafa, YILDIZ İbrahim, KUL Sinan (2015). Bilişim Odaklı Bir Bölümde Okumanın İnternet Bağımlılığına Etkisi Üzerine Karşılaştırmalı Bir Araştırma. Atatürk üniversitesi iktisadi ve idari Bilimler Dergisi, 29(3), 543-556. (Kontrol No: 2292403)

	F. Kitap/Bölüm, madde ve ünite yazarlığı: 1. Veri Tabanı Yönetim Sistemleri, Bölüm adı:(Sorgu Oluşturmak ve Çeşitlerini Kullanmak-2) (2016). KUL Sinan Atatürk Üniversitesi - Açık Öğretim Fakültesi, Editör: Yrd. Doç. Dr. Abdulkadir ÖZDEMİR, Basım sayısı:1, Türkçe(Ders Kitabı) 2. Veri Tabanı Yönetim Sistemleri, Bölüm adı:(Sorgu Oluşturmak ve Çeşitlerini Kullanmak-3) (2016). KUL Sinan Atatürk Üniversitesi - Açık Öğretim Fakültesi, Editör: Yrd. Doç. Dr. Abdulkadir ÖZDEMİR, Basım sayısı:1, Türkçe(Ders Kitabı) 3. Veri Tabanı Yönetim Sistemleri, Bölüm adı:(Sorgu Oluşturmak ve Çeşitlerini Kullanmak-4) (2016). KUL Sinan Atatürk Üniversitesi - Açık Öğretim Fakültesi, Editör: Yrd. Doç. Dr. Abdulkadir ÖZDEMİR, Basım sayısı:1, Türkçe(Ders Kitabı) 4. Kurumsal Bilgi Yönetimi, Bölüm adı:(Bilgi Yönetiminde Temel Kavramlar) (2016). KUL Sinan Atatürk Üniversitesi - Açık Öğretim Fakültesi, Editör: Doç. Dr. Abdullah NARALAN, Basım sayısı:1, Türkçe(Ders Kitabı), (Yayın No: 2964011) 5. Kurumsal Bilgi Yönetimi, Bölüm adı:(Bilgi Yönetiminde Sorunlar ve Geleceği) (2016). KUL Sinan Atatürk Üniversitesi - Açık Öğretim Fakültesi, Editör: Doç. Dr. Abdullah NARALAN, Basım sayısı:1, Türkçe(Ders Kitabı), (Yayın No: 2964027) 6. Programlama Temelleri, Bölüm adı:(Hata Ayıklama) (2018). KUL Sinan Atatürk Üniversitesi - Açık Öğretim Fakültesi, Editör: Dr. Öğr. Üyesi Deniz DAL, Basım sayısı:1, Türkçe(Ders Kitabı) Projeler 1. KUL Sinan, AĞIRMAN Ensar, BAKIRCI Fehim, “Ders Program Çizelgeleme Otomasyonu”, Yükseköğretim Kurumları tarafından destekli bilimsel araştırma projesi(BAP), 2016 Görevler 1. Bilgisayar Araştırma ve Uygulama Merkez Müdürlüğü, Otomasyon yazımı, 2015 G. Tezler 1. Kul, S. (2014). Paralel Hata Ayıklama”, (Yayımlanmamış Yüksek Lisans Tezi), Fen Bilimleri Enstitüsü, Atatürk Üniversitesi, Erzurum. I. Ödüller-Burslar: 1. TEV Bursu. İ. Verdiği Dersler: 1. Veri tabanı Yönetim Sistemleri 2. Algoritma ve Programlamaya Giriş 3. Temel Bilgi Teknolojileri
İş Deneyimi	
Stajlar	İski (2007)
Projeler	
Çalıştığı Kurumlar	Çağtay Bilgisayar Kara Kuvvetleri Komutanlığı Ardahan Üniversitesi (Öğretim Görevliliği)

İletişim	
E Posta Adresi	Sinan.kul@atauni.edu.tr
Tarih	2018

