T.C.
SÜLEYMAN DEMİREL ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

İNTERNET ORTAMINDAKİ MÜŞTERİ YORUMLARININ FİKİR MADENCİLİĞİ İLE ANALİZ EDİLMESİNE YÖNELİK BİR ÇALIŞMA

MUSTAFA ERDOĞMUŞ

Danışman
Prof. Dr. Gültekin ÖZDEMİR

YÜKSEK LİSANS TEZİ
ENDÜSTRİ MÜHENDİSLİĞİ
ISPARTA- 2019

TEZ ONAYI

MUSTAFA ERDOĞMUŞ tarafından hazırlanan **" İnternet Ortamındaki Müşteri Yorumlarının Fikir Madenciliği ile Analiz Edilmesine Yönelik Bir Çalışma"** adlı tez çalışması aşağıdaki jüri üyeleri önünde Süleyman Demirel Üniversitesi Fen Bilimleri Enstitüsü **Endüstri Mühendisliği Anabilim Dalı**'nda **YÜKSEK LİSANS TEZİ** olarak başarı ile savunulmuştur.

Danışman	**Prof. Dr. Gültekin Özdemir** Süleyman Demirel Üniversitesi	
Jüri Üyesi	**Dr. Öğr. Üyesi Erdal Aydemir** Süleyman Demirel Üniversitesi	
Jüri Üyesi	**Dr. Öğr. Üyesi Mehmet Fatih Demiral** Burdur Mehmet Akif Üniversitesi	

Enstitü Müdürü **Doç. Dr. Şule Sultan UĞUR** .....

TAAHHÜTNAME

Bu tezin akademik ve etik kurallara uygun olarak yazıldığını ve kullanılan tüm literatür bilgilerinin referans gösterilerek tezde yer aldığını beyan ederim.

Mustafa ERDOĞMUŞ

İÇİNDEKİLER

Sayfa

ÖZET

Yüksek Lisans Tezi

İNTERNET ORTAMINDAKİ MÜŞTERİ YORUMLARININ FİKİR MADENCİLİĞİ İLE ANALİZ EDİLMESİNE YÖNELİK BİR ÇALIŞMA

Mustafa ERDOĞMUŞ

Süleyman Demirel Üniversitesi
Fen Bilimleri Enstitüsü
Endüstri Mühendisliği Anabilim Dalı

Danışman: Prof. Dr. Gültekin ÖZDEMİR

Günümüz teknolojisinde duygu analizi konusu şirketler için önem kazandıkça büyük bilişim firmaları da bu alana ciddi yatırımlar yapmaya devam etmektedirler. Akademik dünyanın da bu alana son 5-10 yıldır ciddi yönelimleri olduğu yapılan literatür taramasında görülmektedir. Büyük oranla akademik çalışmalar sosyal medya analizlerine dayanmaktadır.

Bu çalışmada hedeflenen; internet platformlarında hızla biriken ham yorum verilerinin duygu analizini yapmaktır. Bunun sonucunda ise insanların ve şirketlerin ürünler hakkında bilgiye kolay ulaşmasını sağlamaktır. Müşteriler bir ürünü almaya karar vermeden önce fazlaca yorum okuyarak karar verme eğilimindedirler. Şirket yöneticileri ürünleri hakkında piyasadan bilgi toplamak için sadece belli bir kesime uygulanan anket araştırması ile bilgiye ulaşmaya çalışırlar. Bu anketler ya da geri bildirim formlarını işleyerek duygu bilgisini ortaya çıkarmaya çalışırlar.

Hedefe ulaşmak için yapılandırılmamış veri kaynakları tespit edildi. Bu kaynaklara otomatik erişimi sağlayacak bir web kazıma programı geliştirilerek müşteri yorumları veri tabanına kayıt edildi. Ardından veri madenciliği, metin madenciliği ya da fikir madenciliği konularında yapılması gereken aşamalardan olan metin ön işleme adımları uygulandı. Yapılandırılmış veriye dönüşüm aşamalarında dönüştürme, kelime köklerini bulma, durak kelimeleri ayrıştırma yöntemleri uygulanarak yarı yapılandırılmış veri türüne çevrildi. Ardından, terim ağırlıklandırma, Python NLTK Vader ve Stanford NLP, kelime bulutları yöntemleri kullanılmıştır. Elde edilen sonuçlar istatiksel grafiklerle karşılaştırılmıştır.

Anahtar Kelimeler: Metin madenciliği, fikir madenciliği, müşteri memnuniyeti, duygu analizi, duygu skoru, büyük veri

2019, 85 sayfa

ABSTRACT

M.Sc. Thesis

A STUDY TO ANALYZE CUSTOMERS ON THE INTERNET BY USING OPINION MINING

Mustafa ERDOĞMUŞ

Süleyman Demirel University
Graduate School of Natural and Applied Sciences
Department of Industrial Engineering

Supervisor: Prof. Dr. Gültekin ÖZDEMİR

As the issue of sentiment analysis becomes more important for today's technology, large IT companies continue to make serious investments in this field. It is seen in the literature review that the academic world has had serious orientations in this field for the last 5-10 years. Academic studies are mostly based on social media analysis.

Targeted in this study; is to analyze the emotions of raw comment data which is accumulated rapidly on internet platforms. As a result, it is to ensure that people and companies have easy access to information about products. Customers tend to decide to buy by reading a lot of comments before deciding. In order to gather information about their products from the market, company managers try to access information only through survey resarch applied to a certain segment. They try to reveal emotion information by processing questionnaires or feedback forms.

Unstructured data sources were identified to achieve the goal. A web scraping program was developed to provide automatic access to these resources and customer reviews were recorded in the database. Then, data and text mining stages were applied. Four different methods were used to extract emotion information from the interpretation data obtained and the results were evaluated. Term frequency, Python NLTK Vader and Stanford NLP, word cloud techniques were applied. The results obtained were compared with the statistical graphs.

Keywords: Text mining, opinion mining, customer satisfaction, sentiment analysis, semantic score, big data

2019, 85 pages

TEŞEKKÜR

Bu araştırma için beni yönlendiren, karşılaştığım zorlukları bilgi ve tecrübesi ile aşmamda yardımcı olan değerli Danışman Hocam Prof. Dr. Gültekin ÖZDEMİR' e teşekkürlerimi sunarım. Araştırmalarımda fikirleri ve yardımları için Endüstri Mühendisliği Bölümü öğretim üyesi Doktor Öğretim Üyesi Erdal Aydemir ve Arş. Gör. Yusuf KARADEDE' ye teşekkür ederim.

Tezimin her aşamasında beni yalnız bırakmayan aileme sonsuz sevgi ve saygılarımı sunarım.

Mustafa ERDOĞMUŞ
ISPARTA, 2019

ŞEKİLLER DİZİNİ

Sayfa

ÇİZELGELER DİZİNİ

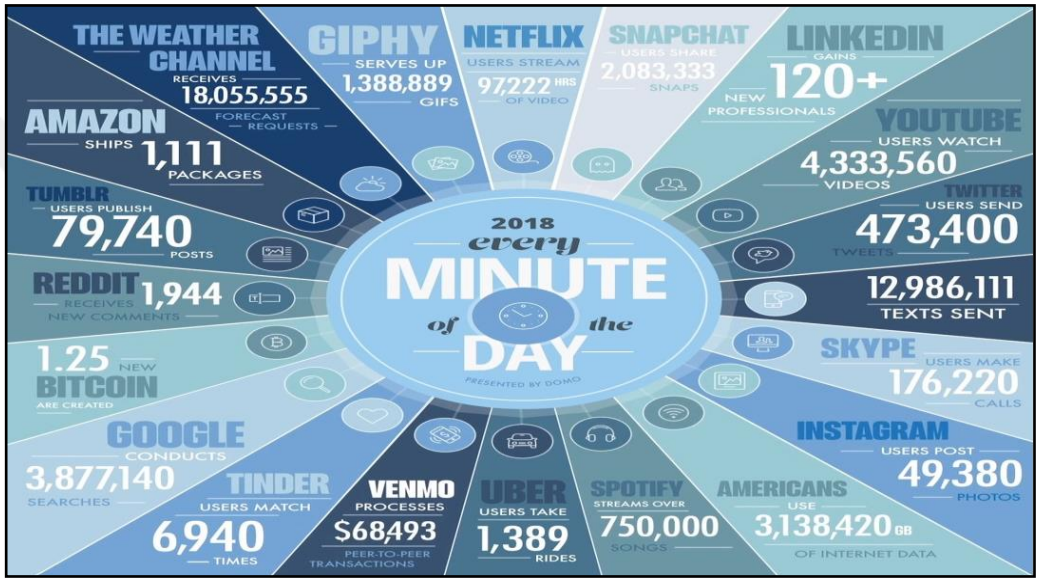
Sayfa

SİMGELER VE KISALTMALAR DİZİNİ

NLP	Doğal Dil İşleme	(Naturel Language Process)
NLTK	Doğal Dil Aracı	(Natural Language Toolkit)
IoT	Nesnelerin İnterneti	(Internet of Things)
GA	Genetik Algoritma	(Genetic Algorithm)
AHP	Analitik Hiyerarşi Süreci	(Analytic Hierarchy Process)

1. GİRİŞ

Teknolojik gelişmeler hızla ilerledikçe bilgisayar ve internet ortamında kayıt altına alınan verilerde gün geçtikte çok büyük veri boyutlarına ulaşmaktadır. Özellikle bilgisayar teknolojisinin ucuzlayıp, internete ve teknolojiye erişim imkânı artan kullanıcılar tarafından oluşturulan verilerin boyutları DOMO tarafından her yıl yayınlanan Data Never Sleeps raporunda inanılmaz boyutlara nasıl ulaştığını göstermektedir (Şekil 1.1).



Şekil 1.1. DOMO Data Never Sleeps raporu (Domo, 2019)

Günümüzde sadece kullanıcıların ürettiği veriler değil cihazların ürettiği veriler de internet ortamında yer alıyorlar. Nesnelerin interneti (Internet of Things: IOT) denilen bu teknoloji kavramı; "Akla gelecek her nesnenin akıllı bir nesne olarak dönüştürülmesi ve o nesnelerin artık akıllı bir nesne olmasına" denilmektedir. Bugün insan eliyle oluşturulan verilerin boyutları bu kadar fazla iken nesnelerin üreteceği verilerin boyutları çok daha fazla olacaktır. Bu nedenle veri işlemek yeterli olmamaktadır. Hızlı ve doğru işlenmesi gerekmektedir.

Yakın geleceğimizin en büyük teknoloji ekonomisinin "Büyük Veri (Big Data)" teknolojisi üzerinde oluşması beklenmektedir. Önümüzdeki beş on yıl içinde de bu teknoloji ekonomisindeki pazar payının elli milyar doları aşması

beklenmektedir. Dünya çapında yıl bazlı veri hacminin büyüme oranı %59 olması ve bunun sürekli artarak devam etmesi tahmin ediliyor. Bu büyümenin merkezinde hem geleneksel tek merkezden üretilen veri [uydu görüntüleri vb.] hem de yeni veri kaynakları (sosyal medya vb.) yatmaktadır. 2018 yılında 33 zettabyte olan global veri boyutu 2025 yılında 175 zettabyte boyutuna ulaşacağı, sonraki on sene içinde de 44 katına çıkacağını tahmin edilmektedir. "Günümüzde, 5 milyardan fazla tüketici her gün verilerle etkileşim kurmaktadır. 2025'e kadar, bu rakam 6 milyar veya dünya nüfusunun %75'i seviyesinde olacağı öngörülmektedir. 2025'te, ağa bağlı her insanın her 18 saniyede en az bir etkileşimi olacaktır. Bu etkileşimlerin çoğu, tüm dünyada ağa milyarlarca IoT cihazının bağlı olmasından kaynaklanacağı araştırmalar sonucunda ortaya koyulmaktadır. Bu cihazların, 2025'te 90 ZB'tan fazla veri oluşturması bekleniyor." (Date Age 2025, 2019). Bu büyümenin asıl kaynağı yapısal olmayan verilerden oluşmaktadır (Wikipedia, 2017). Çizelge 1.1'de bilgisayar depolama boyutlarının büyüklükleri görülmektedir. 1 Zettabyte 1 trilyon Gigabyte büyüklüğündedir.

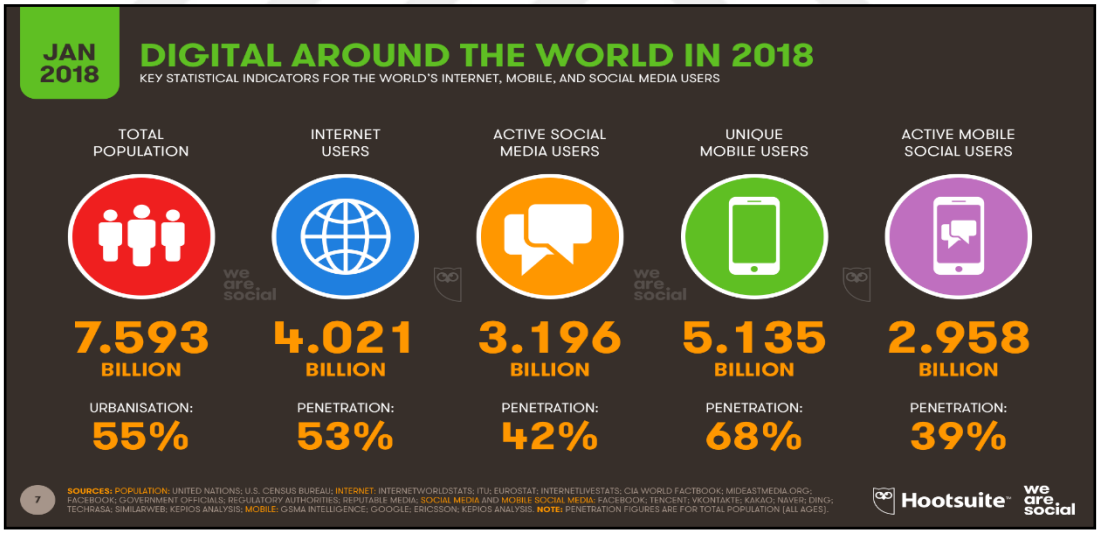
Çizelge 1.1. Bilgisayar depoloma boyutları

Kapasite	**Sembolü**	**Değeri**
1 bit	bit	0 veya bir
1 byte	Byte	8 bit
1 Kilobyte	KB	1024 byte
1 Megabyte	MB	1024 KB
1 Gigabyte	GB	1024 MB
1 Terabyte	TB	1024 GB
1 Petabyte	PB	1024 TB
1 Zettabyte	ZB	1024 PB
1 Yottabyte	YB	1024 ZB
1 Brontobyte	BB	1024 YB

İnsan eliyle artan yapılandırılmamış veri, her geçen gün, bilgisayar ve internet kullanmaya başlayan yeni insanlar, dünya nüfusunun artması bunlara ek olarak cihazların üreteceği veriler hesaba katıldığında büyüme hızı her yıl bir önceki yıldan çok daha fazla olacaktır. Araştırmalara göre 2025 yılına kadar IoT cihazlar 90ZB veri oluşturulacak, verilerin yüzde 49'u genel bulut ortamlarında depolanacak, üretilen verilerin yaklaşık yüzde 30'u 2025 yılına kadar gerçek

zamanlı olarak tüketilecektir. Bu nedenle dünya üzerinde 4 milyar kullanıcının internet kullandığı ve oluşan yapılandırılmamış veriyi yapılandırıp içinden anlamlı ve kullanılabilir bilgiler elde etmek günümüz ve gelecek için büyük önem arz etmektedir (Şekil 1.2).

Veri boyutlarının bu kadar büyümesi veriye olan bilimsel bakış açısını değiştirmektedir. Veriyi işlenebilir hale getirmenin ve ham veriden bilgi elde etmenin incelendiği bilim dallarından olan veri madenciliği, metin madenciliği, doğal dil işleme bu alanda bilimsel çalışmaların çokça yapıldığı bir alan haline gelmiştir. Bilgiyi elinde tutan toplumlar daha refah bir düzeyde yaşamlarını sürdürdükleri bir gerçektir. Ülkemiz adına, internet dünyasında biriken verileri işlemek bilgiye dönüştürmek, elde edilen bilgileri kullanmak gelecek nesillerimiz için önem arz etmektedir. Veriyi işlemek kadar veriye ulaşmakta önemlidir. Bugün akıllı cep telefonu kullanan her bireyin verileri (konum, rehber, internet aramaları vb.) yabancı firmaların elinde bulunmaktadır.

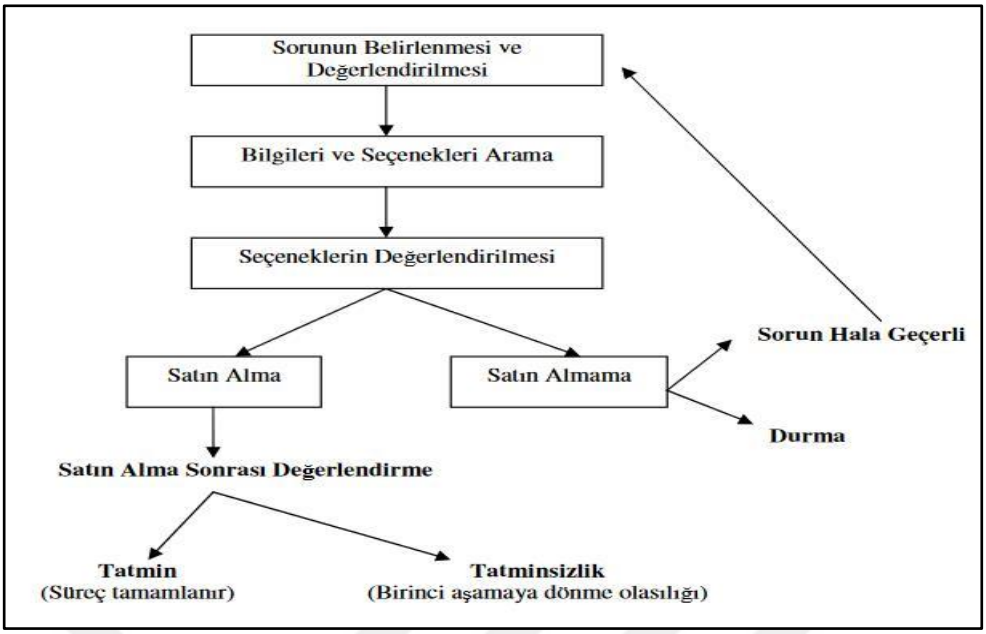


Şekil 1.2. Dünya internet kullanıcıları araştırma raporu (We Are Social, 2018)

İnternet kullanıcıları bilgilerini, düşüncelerini, fikirlerini, deneyimlerini özellikle sosyal medya üzerinde paylaşırlar. Sosyal medya platformlarına paralel şekilde gelişmekte olan diğer sistemlerde de işlenmemiş veri miktarı gün geçtikte artmaktadır.

İnternet ortamında, hizmet ve sanayi alanında faaliyet gösteren firmaların ürettikleri marka ve ürünlerin müşteri yorumları farklı internet platformlarında memnuniyet ve şikâyet bağlamında yerlerini almaktadır. İnsanların ihtiyaç duydukları bilgiyi araştırma alışkanlığı, satın almak istedikleri ürünü araştırmada da ortaya çıkmaktadır. Tüketiciler genellikle alacağı ürün türünü belirledikten sonra bu türe ait markaları araştırmaya başlamaktadırlar. Bu konuyla ilgili yapılan literatür taramasında, **Tüketici Satın Alma Karar Süreci** altında yapılan araştırmalarda tüketicilerin bilgi kaynaklarının; yakın çevre, tv-radyo-internet reklamları, internet haberleri vb. kaynaklardan oluştuğu görülmüştür. Günümüzde fiyat, tüketici için ne kadar önemli ise daha önce bu ürünü ya da hizmeti kullanan tüketicilerin beyan ettikleri fikirler ve yorumlar da önemli hale gelmektedir. Örneğin; bir cep telefonu ya da bilgisayar satın alınması düşünüldüğünde özellikle o markayı satın almış ve kullanan tüketicilerin yorumlarını birden fazla platformdan bulup okuyarak sonrasında o ürünü satın alıp almamaya karar vermekteyiz. Buna karşın, konuyla ilgili çok fazla veri bulunduğundan ve her geçen gün bu veri miktarları daha da arttığından dolayı insanların bu yorumları okuyup karar vermesi özellikle zaman açısından zor hale gelmektedir.

Yapılan araştırmalara göre tüketicilerin satın alma süreci aşamaları Şekil 1.3'te görüldüğü gibidir (Odabaş ve Barış, 2002). Bu aşamalardan "seçeneklerin değerlendirilmesi" ve "satın alma sonrası değerlendirme" aşamalarının günümüzde çok büyük oranda teknoloji kullanılarak yapıldığı aşikârdır.



Şekil 1.3. Tüketici satın alma karar süreci aşamaları

Geçmişten günümüze insanlar her alanda "karar verme" eğilimindedirler. Verilen bu kararlar doğru olabileceği gibi yanlışta olabilmektedir. Doğru karar verebilmek için bilgiye ihtiyaç, bilgi isteniyorsa veriye ihtiyaç doğar. Karar destek sistemleri ve veri madenciliği alanları tam da bu hiyerarşik talepleri karşılamada rol oynar. Veri varsa ya da elde edilebiliyorsa; veri madenciliği ve doğal dil işleme yöntemleriyle programlama araçları kullanılarak işe yarar doğru bilgiler ortaya çıkartılabilir. Doğru bilgi varsa karar verme kolaylaşır ve verilen karar sonucunda da kurumlar ya da bireyler hedeflerine daha doğru ve hızlı bir şekilde ulaşabilirler.

2. KAYNAK ÖZETLERİ

Veri ve metin madenciliği birçok disiplinle çalıştığı ve birçok alana hitap eden çözümler önerdiği için literatürde çokça veri ve metin madenciliği tanımı bulunmaktadır. Bunlardan bazıları derlenerek aşağıda sunulmuştur.

- Jacobs (1999), "Veri madenciliğini, ham verinin tek başına sunamadığı bilgiyi çıkaran, veri analizi süreci" olarak tanımlamıştır.
- Veri madenciliği, "Büyük veri yığınları arasından gelecekle ilgili tahminde bulunabilmesini sağlayabilecek bağlantıların, bilgisayar programı kullanarak aranması işidir" olarak tanımlamıştır (Doğan ve Türkoğlu, 2008).
- Hand (1998), "Veri madenciliğini istatistik, veri tabanı teknolojisi, örüntü tanıma, makine öğrenme ile etkileşimli yeni bir disiplin ve geniş veri tabanlarında önceden tahmin edilemeyen ilişkilerin ikincil analizi" olarak tanımlamıştır."
- Kittler ve Wang (1999), "Veri madenciliğini oldukça tahminci anahtar değişkenlerin binlerce potansiyel değişkenden izole edilmesini sağlama yeteneği" olarak tanımlamışlardır".
- "Veri madenciliği, daha önceden bilinmeyen, geçerli ve uygulanabilir bilgilerin geniş veri tabanlarından elde edilmesi ve bu bilgilerin işletme kararları verirken kullanılmasıdır" (Akpınar, 2014).
- "Veri madenciliği, birimin sahip olduğu veri veya bilgi kaynaklarında yönetici ve analistin sormayı düşünmediği sorular hakkındaki cevapların aranmasıdır" (Gray ve Watson, 1997).

Şekil 2.1'de görüleceği üzere birden fazla disiplin bir araya gelerek veri madenciliği kavramını oluşturmaktadır. Yapılan literatür çalışmasında görülmüştür ki; ticaretten, pazarlamaya, bilişim sektöründen tıp alanına, bankacılık ve finans alanından eğitim faaliyetlerine, genetik, telekomünikasyon, kriminoloji, endüstri, istihbarat vb. daha birçok alanda veri, metin, fikir madenciliği çalışmaları görülmektedir. Bunlardan yapılacak olan çalışmayla paralel olanlar aşağıda incelenmiştir.

Şekil 2.1. Veri madenciliği yardımcı disiplinler

Sinan Aydın ve Ali Ekrem Özkul tarafından 2005 yılında yapılan çalışmada Açık Öğretim sisteminde öğrenci performansını değerlendirmeye yönelik farklı sınıflama modelleri kullanılarak karşılaştırma yapılmıştır. Öğrenci başarısını tahmin etmeye yönelik olarak tahmin performansı en yüksek C5.0 karar ağacı algoritmasının kullanıldığı modelle en yüksek başarıyı elde ettiği bir çalışma ortaya koymuşlardır (Aydın & Özkul, 2015)

Kıyas Kayaalp tarafından 2007 yılında yapılan bir çalışmada, veri madenciliği tekniği ile üç fazlı asenkron motordaki sargı spirleri arasında oluşabilecek kısa devre veya yalıtım bozuklukları ve motor milinde oluşabilecek mekanik dengesizlik hatalarının tespiti gerçekleştirilmiştir. Weka yazılımına akım ve gerilim değerleri verilmiş sınıflandırma algoritmalarından karar ağacı algoritması kullanılmış ve başarılı sonuç elde edilmiştir. Bu çalışma ile asenkron motorlarda oluşan hataların erken teşhisi amaçlanmış ve başarılı olduğu beyan edilmiştir (Kayaalp, 2007).

Eyüp Sırımkaya tarafından 2005 yılında yapılan bir çalışmada, internet üzerinden ulaşılabilen basın-yayın kaynaklarında yer alan görsel ve metinsel verilerden anlamlı ve önemli bilgilerin çıkartılması hedeflenmiştir. Bu çalışmanın sonuçlarına göre "büyük-küçük harf" duyarlı bir karşılaştırma yapan klasik karşılaştırma algoritmasına göre "bulanık mantık" algoritması kullanılarak yapılan karşılaştırma algoritması çok önemli bir üstünlük sağladığı sonucu elde edilmiştir (Sıramkaya, 2005).

Alaettin Uçan 2014 yılında gerçekleştirdiği çalışmada, sözlük kullanarak Duygu Analizi yapılmasını amaçlamıştır. Dilimiz için daha önce hazırlanmamış ve hazırlanması oldukça zahmetli olan "Türkçe Duygu Sözlüğü", İngilizce için hazırlanmış bir duygu sözlüğünü otomatik tercüme yöntemiyle Türkçeleştirerek oluşturulmuştur. Farklı tercüme algoritmalarıyla daha doğru çeviri yapılmaya çalışılmıştır. Sözcüklerin karşılığı birden fazla sözlükte aranmış, bu sayede kontrollü çeviri yapılması hedeflenmiştir. Oluşturulan Türkçe Duygu Sözlüğü ile yapılan duygu analizi sonucunda %80 civarında başarı elde edilmiştir. Hem İngilizce için yapılan deneylerin sonuçları ile hem de tez kapsamında yapılmış olan makine öğrenmesi deneylerinin sonuçları ile karşılaştırıldığında Türkçe Duygu Sözlüğünün başarısının yaklaşık olarak aynı düzeyde olduğu ortaya konulmuştur (Uçan, 2014).

Mehmet Nanğır 2013 yılında gerçekleştirdiği çalışmada, duygu analizini metin madenciliği ile inceleyerek üç tane makine öğrenmesi algoritması kullandığını belirtmiştir. Bu çalışma kapsamında, Türk dili için farklı veri kümeleri üzerinde çoklu sınıflandırıcı makine öğrenmesi algoritmaları uygulanmıştır. Daha önce uygulanan çalışmalardan farklı olarak, performansı yüksek üç tane makine öğrenmesi algoritması birlikte kullanılarak özgün bir çoklu sınıflandırıcı makine öğrenmesi algoritması tasarlanmıştır. Bu özgün sınıflandırıcı yaklaşımının yanı sıra, makine öğrenmesi algoritmalarının parametre optimizasyonu gerçekleştirilerek performans arttırılmıştır. Bu yeni yaklaşım sayesinde, daha önce tek sınıflandırıcı ile elde edilen doğruluk oranı %86,13 lük bir doğruluk oranına yükseltilmiştir. Bu doğruluk oranı, yeni yaklaşımın performansı

iyileştirdiğini ve birçok çalışmada kullanılabileceğini ortaya koymuştur (NANĞIR, 2013).

Burak Çağrı Okur tarafından 2013 yılında yapılan çalışmada İngilizce alfabesi ile yazılan Türkçe metinler her ne kadar insanlar tarafından kolay anlaşılsa da bu işlemin otomatik olarak yapılması günümüzde hala tam çözülmemiş Sözcük Anlamı Belirleme problemi üzerine bir çalışma ortaya koymuştur. İngilizce alfabesi ile yazılmış olan metinlerin Türkçe alfabesi ile yeniden yazılması, Türkçe 'ye özgü bir Doğal Dil İşleme çalışmasıdır. Farklı Türkçe kelime seçenekleri içinden, uygun olanın bulunması için metnin anlamsal açıdan ele alınması gerekmektedir. Bu çalışmada, metnin cümle bazlı veya tüm parça olarak incelenmesinin doğru kelime tercihi üzerindeki etkileri araştırılmıştır. İstatistiğe dayalı yöntemler ile makina öğrenmesi yöntemlerinin doğru kelime tercihi üzerindeki başarısı incelenmiştir. Bir metnin tüm parça olarak incelenmesinin, bize metin hakkında cümle bazlı yönteme göre daha fazla bilgi verdiği; ayrıca makina öğrenmesi yöntemlerinin, istatistiksel bazlı yapılan çalışmalara göre daha iyi sonuçlar sağladığı deneylerle gösterilmiştir (OKUR, 2013).

V.L. Miguéis ve arkadaşları 2012 yılında yayınlanan makalelerinde yapmış oldukları çalışmayla perakendeciliğin müşterinin yaşam tarzına dayalı olarak pazar segmentasyonu için büyük bir işlem veri tabanından alınan bilgilerle desteklenen bir yöntem önermektedir. Değişken bir kümeleme algoritması kullanarak, veri tabanından bir dizi tipik alışveriş sepeti çıkarılır ve bunlar müşterilerin yaşam tarzlarını belirlemek için kullanılır. Müşteriler satın alma geçmişlerine dayanarak bir yaşam biçimi segmentine atanır. Bu çalışma bir Avrupa perakende şirketi ile iş birliği içinde yapıldığı görülmüştür (Miguéis, Camanho, & Falcão e Cunha, 2012).

Hossein Alizadeh ve Behrouz Minaei-Bidgoli 2016 yılında yayınlanan makalelerinde yapmış oldukları çalışmanın temel amacının, farklı kümeleme yöntemleri performansının değerlendirilmesi ve karşılaştırılmasına dayanan, banka müşteri sadakat değerlendirmesinin kapsamlı bir modelini ortaya

çıkarmak olduğunu belirtmişlerdir. Bu çalışma aynı zamanda aşağıdaki spesifik hedefleri de yerine getirmektedir:

a) farklı kümeleme yöntemlerini kullanarak ve bunları müşteri sınıflaması için karşılaştırarak,
b) müşteri sadakati belirlemede etkili değişkenleri bulma ve
Sadık müşterilerden daha fazla kâr elde edildiğinden bu çalışma, müşterilerin sınıflandırılması ve sadakatine yönelik iki aşamalı bir model sunmayı amaçlamaktadır. Bu amaçla, K-Medoid, X-Means ve K-Means gibi çeşitli kümeleme yöntemleri kullanılmış ve sonuncusu Davis-Bouldin indeksiyle karşılaştırarak diğerlerinden daha iyi performans göstermiştir. Müşteriler K-araçları kullanılarak kümelenmiş ve bu dört kümenin üyeleri analiz edilmiştir. Daha sonra, DT (Karar Ağacı), YSA (Yapay Sinir Ağı), NB (Naive Bayes), KNN (K-En Yakın Komşular) ve SVM (Destek Vektörü) gibi çeşitli sınıflandırma yöntemlerini kullanarak müşterilerin demografik değişkenlerine dayalı bir tahmin modeli uygulandı. Sonuçlar, ANN 'nin (Artificial Neural Networks-Yapay Sinir Ağları) sadık müşterileri tahmin etmede en doğru yöntem olduğunu gösterdi. Bu iki aşamalı model, gelecekteki müşteri türünü belirlemek için benzer veriler içeren bankalarda ve finansal kuruluşlarda kullanılabilir (Alizadeh & Minaei-Bidgoli, 2016).

Bankacılık işlemlerinde bankaların vermesi gereken önemli bir karar vardır. Kredi başvurusu yapan müşterilerini nasıl sınıflandırılacakları konusunu otomatik gerçekleştirmeleri gerekir. Abedini Mohammadali 2016 yılında yayınlanan makalelerinde bu karar verme sürecini desteklemek için dört aşamalı bir hibrid veri madenciliği yaklaşımı önermektedir (Abedini, Ahmadzadeh, & Noorossana, 2016).

Mustafa Çetingöz 2011 yılında yaptığı: “Makine öğrenmesi ile Türkçe haber metinlerinde anahtar ifade çıkarımı”. Başlıklı yüksek lisans tez çalışmasında Türkçe haber metinlerinden elde edilen eğitim ve test verileri kullanılarak, KEA algoritması ile ve ilave bir özellik eklenerek oluşturulan KEA-SPR algoritması ile

uygulama geliştirmiş, ilave edilen özellik için performans karşılaştırılması yapmıştır (Çetingöz, 2011).

Betül Güvenç 2016 yılında yaptığı: “Doğal dil işlemede makine öğrenmesi yöntemleri” başlıklı yüksek lisans tez çalışmasında; Doğal Dil İşleme tekniklerinde kullanılan kelime çıkarma ve metin özetleme algoritmalarını incelemiş ve Word2Vec ve PageRank algoritmalarını kullanarak anahtar kelime çıkartmak için yeni bir yöntem önermişlerdir (Güvenç, 2016).

3. KONU VE KAPSAM

Bu tez kapsamında değerlendirilen "yorum verilerinin" toplam sayısı 2.174.713 adettir. Bu verilerin diskte kapladıkları alan ise 5,92 GB tır. Bu yoğunluktaki ham yapılandırılmamış ve yarı yapılandırılmış verinin fikir madenciliği ve doğal dil işleme aşamaları ve süreçleri Python, C# ve R dili kullanılarak yazılmıştır. Veri işleme aşamalarında Python ve R dilinin tercih edilmesinin nedeni daha hızlı ve veri işleme konusunda diğer dillere göre daha yetenekli olmasıdır.

3.1. Metin Madenciliği

Metin madenciliği, metni veri kaynağı olarak kabul eden veri madenciliği çalışmasıdır. Diğer bir tanımla metin üzerinden yapısal hale çevrilmiş veri elde etmeyi amaçlar. Metin madenciliği, metinlerin sınıflandırılması, bölümlenmesi (İng. clustering), metin içinden konu çıkarılması (İng. concept/entity extraction), metinler için sınıf taneciklerinin üretilmesi (İng. production of granular taxonomy), metinlerde görüş analizi yapılması (İng. sentimental analysis), metin özetlerinin çıkarılması (İng. document summarization) ve metinin özü ile ilgili ilişki modellemesi (İng. entity relationship modelling) gibi çalışmaları hedefler (Wikipedia, 2019). Veri madenciliği tanımında da anladığımız üzere veriler yapılandırılmış ya da yapılandırılmamış veri olarak ikiye ayrılır.

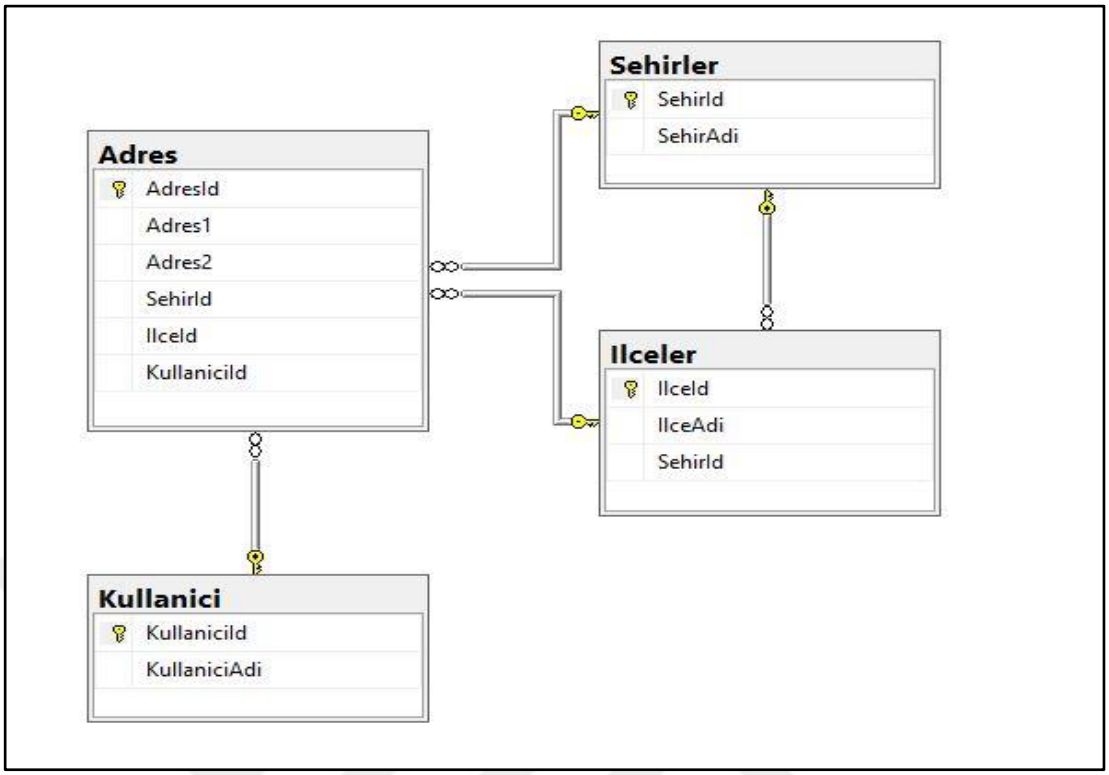
3.1.1. Yapılandırılmış Veri

Belli bir ilişki içerisinde depolanan, istenildiğinde rahatlıkla erişilebilen, işlenmesi gayet kolay olan veriye yapılandırılmış veri denilmektedir. Genellikle ilişkisel veri tabanlarında karşımıza çıkmaktadırlar. Günümüzde kullanılan yazılım sistemlerinin neredeyse tamamı ilişkisel veri tabanları kullanmaktadırlar. Veriler tablolarda ve bu tablolar arasında kurulan veri ilişkileri ile sistemli bir şekilde tutulmaktadırlar. Veri üzerinde sorgulama yaptığınızda net bir yanıt alabiliyorsak elimizde yapılandırılmış bir veri var demektir. Örneğin; öğrencilerin notlarının Çizelge 3.1'de verildiği şekilde olduğu kabul edilsin.

Çizelge 3.1. Yapılandırılmış veri örneği

Adı Soyadı	**Not 1**	**Not 2**	**Ortalama**
Mustafa ERDOĞMUŞ	100	90	95
Zümra ERDOĞMUŞ	100	100	100
Perihan ERDOĞMUŞ	100	90	95
Ahmet YIKILMAZ	40	60	50
Cenk YILMAZ	10	20	15

Çizelge 3.1'deki verilere, "puan ortalaması 95 üzerinde olan öğrenciler kimlerdir?" sorusu sorulduğunda cevabı rahatlıkla alınabilmektedir. Veri üzerinde sorduğumuz sorulara yanıt alabiliyorsak bunlar yapılandırılmış veri adını almaktadır. Genelde bu tarz verilere yönetim sistemlerinde rastlanılmaktadır. Elektronik Belge Yönetim Sistemi, Öğrenci Bilgi Sistemi, Yönetim Bilişim Sistemi, ERP Sistemleri vb. bir sürü sistemde sistematik ve ilişkisel veri bağlantısı kullanarak işlem yapılır. Kullanıcılar bu gibi sistemlerde işe yarar bilgileri tablo örneğinde olduğu gibi birkaç butona tıklayarak istedikleri soruların cevaplarını sistem üzerindeki veriden alabilirler. Çünkü o veriler belli kurallar çerçevesinde birbirleri ile bağlantıları sağlanmıştır. Belli bir düzen içinde saklanmaktadırlar. Bu nedenle Personel Bilgi Sistemine maaşı 5000 TL'den yüksek kimler çalışıyor sorusunu sorduğumuzda rahatlıkla cevap alabiliriz. Çünkü veriler belli bir düzen ve ilişki içinde tutuluyorlar. Şekil 3.1'de görüldüğü gibi tablolar birbirleri ile belli kurallar içerisinde ilişkilendirilmiş durumdalar. Adres tablosu, kullanıcı tablosuna, şehirler tablosuna ve ilçeler tablosuna bağlı. Veriler bu şekilde düzenli tutulmaktadır. Bu nedenle veri bütünlüğü sağlanmaktadır. İlişkisel veri tabanı kullanmanın bir diğer nedeni veri tekrarlarından kaçınmaktır. Yapılandırılmamış veri de ise veri tekrarları çoktur. Aralarında herhangi bir ilişki bulunmadığı için veriye soru sorulduğunda yanıt almak mümkün değildir. Bu nedenle yapılandırılmamış veri işlenmelidir.



Şekil 3.1. İlişkisel veri modeli örneği

3.1.2. Yarı Yapılandırılmış Veri

Yarı yapılandırılmış veri; ilişkisel veri tabanlarında kullanılan tablolar arası ilişkilerin bulunmadığı, veri tabanı gibi bir mekanizmada saklanmayan fakat ham verilerden daha kullanışlı etiketlerle, meta üst bilgilerle birbirinden ayrılan nispeten yapılandırılmamış veriye göre daha düzenli veri yapılarıdır. Yazılımların birbirleri ile haberleşmelerinde kullanılan JSON, XML, RSS veri haberleşmesi örnek olarak verilebilir. Şekil 3.2'de json veri türü örneğinde gördüğünüz gibi veriler key, value yapısında tutulmakta, birbirleri ile hiçbir ilişkileri bulunmamaktadır.

```
id: "adresSokak",
group: [
    new DynamicInputModel({

        id: "sokakAdi",
        label: "Sokak Adı",
        placeholder: "Sokak Adı"
    }),
    new DynamicInputModel({

        id: "sokakNumarasi",
        label: "Sokak Numarası",
        placeholder: "Sokak Numarası"
    })
]
```

Şekil 3.2. Json veri modeli örneği

Şekil 3.3'de görülen XML veri türünde de aynı json veri türü benzer bir yapıda etiketler arasında veri saklama işini yapmaktadır. Yine birbirleri ile bir bağlantıları bulunmamaktadır.

```
<?xml version="1.0" encoding="UTF-8"?>
<kullanicilar>
    <kullanici id="1">
        <ad>Mustafa</ad>
        <soyad>Erdoğmuş</soyad>
    </kullanici>
    <kullanici id="2">
        <ad>Perihan</ad>
        <soyad>Erdoğmuş</soyad>
    </kullanici>
    <kullanici id="5">
        <ad>Zümra</ad>
        <soyad>Erdoğmuş</soyad>
    </kullanici>
    <kullanici id="8">
        <ad>Nuray</ad>
        <soyad>Karakuş</soyad>
    </kullanici>
</kullanicilar>
```

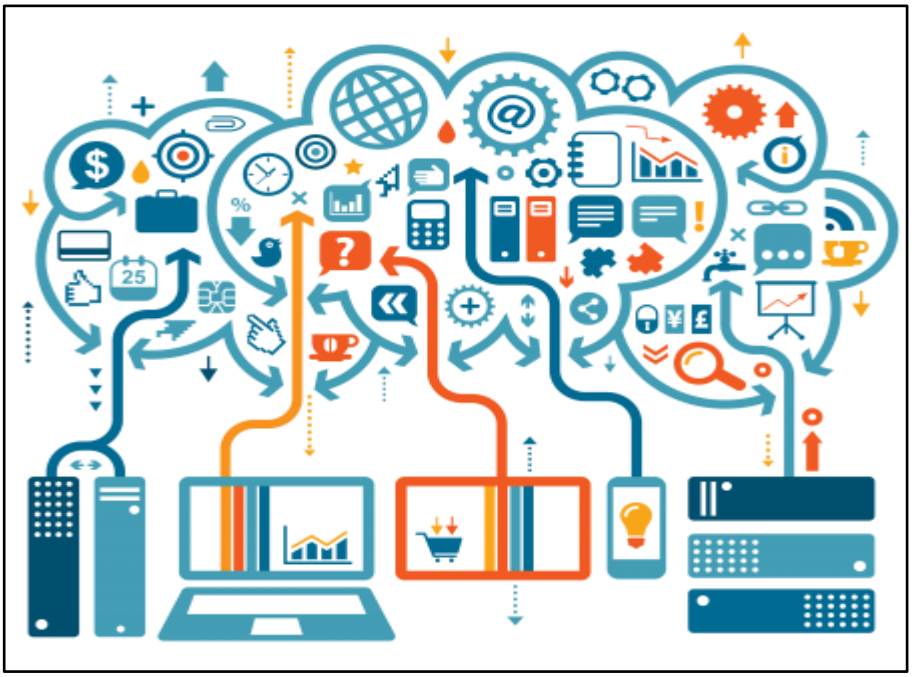
Şekil 3.3. Xml veri modeli örneği

3.1.3. Yapılandırılmamış Veri

Herhangi bir biçimsel kurala bağlı olmayan, büyük ölçüde metinsel içerik taşıyan, işlenip çözümlenebilmesi için gelişmiş metin işleme, doğal dil işleme ve yapay zekâ yöntemleri gereken veri türüdür (Ergüler, 2019). Genellikle sosyal medyada, gelen maillerimizde, internet dünyasında karşılaştığımız veri türü işlenmediği sürece işe yarar bir sonuç elde edemeyiz. Standart bir yapıda olmadıkları için saklandıkları sistemde büyük yer kaplarlar.

Yapılandırılmamış veriler olarak kabul edilen dosya türlerine örnekler: kitaplar, sağlık kayıtları, uydu görüntüleri, Adobe PDF dosyaları, word dosyaları, müşteri yorumları, bloglar, metin mesajları, videolar, fotoğraflar vb. dosya türleridir. Aslında baktığımızda arama motorlarının bünyesinde tuttuğu veriler tamamına yakını yapılandırılmamış veri grubunda yer almaktadır.

Bilgisayar sistemlerinden ve mobil cihazlardan üretilen yapılandırılmamış veriler keşfedilmeyi, yarı ya da tam yapılandırılmış veri formatına dönüştürülüp bilgi çıkarılması için "veri bilimcileri" bekliyor.



Şekil 3.4. Yapılandırılmamış veri kaynakları

İnsan eliyle oluşturulan ve depolanan yapılandırılmamış veriler günümüzün en büyük veri kaynaklarındandır (Şekil 3.4). Artık nesnelerin interneti (Internet of Things, IoT) denilen kavram hayatımıza girmeye başladı. İnsanların ürettiği verilerin çok daha fazlasını makinalar tarafından üretilmesi öngörülüyor. Günümüzde yaygınlaşmaya başlayan akıllı bileklikler, ev otomasyon sistemleri, akıllı arabalar veri üreten ve gelecekte daha fazla üretmesi kesin gözüyle bakılan veri üreticiler olarak karşımıza çıkıyor. Pek bilinmeyen ama büyük teknoloji firmaları tarafından yatırım yapılan bazı projeler hayatımızda yer almaktalar. Bunlardan biri NEST firmasının akıllı ev otomasyon sistemi, Google firması tarafından 2004 yılında satın alındı. HAPIFORK akıllı çatal hızlı yemek yendiğinde ya da günlük kalorinin üzerinde yemek yendiğinde uyaran bir sistem. MICOACH akıllı top Adidas markası bünyesinde geliştirilen top, attığınız kaç penaltının gol olduğunu, kaç km/s hız ile topa vurduğunuzu, hangi ayağınız ile kaç gol attığınızı takip edebildiğiniz bir uygulama. BABOLAT ise akıllı bir kort tenisi raketi. Sporcunun topa vuruş hızını ve açısını, hangi stil ve hangi eli ile topa vurduğunun bilgisini alabildiğiniz bir nesnelerin interneti örneklerindendir.

Görüldüğü üzere artık insan faydasına kullanılan cihazların/makinaların ürettiği verilerde hayatımızda yer almaya başladılar. Sade Group sade.io adında geliştirdiği ürünler, Reengen firmasının enerji alanında geliştirdiği ürünler, Pubinno, Cosa, Evreka, Iven ve Arçelik gibi yerli markalarında bu teknolojide girişimleri bulunmaktadır. İşte bu cihazlardan üretilen veriler network ortamında katlanarak büyüyen hızlarda veri merkezlerine akacaklar. 2000'li yıllarda scale-out NAS teknolojisinin sürekli büyüyen yapılandırılmamış veriyi depolama açısından çözüm olacağı düşünülse de günümüzde bunun yeterli bir çözüm olmadığı ortaya çıkmıştır. Bu nedenle bulut teknolojileri hayatımıza hızlı bir giriş yaptı ve halen güncelliğini koruyarak hayatımızın hayatımızın merkezinde bir konumdalar. Bulut teknolojisi hayatımıza Obje Tabanlı veri depolama teknolojisini beraberinde getirdi. Bu noktada veri depolama mimarileri geçmişten günümüze alt başlıklarda incelenecektir.

3.1.3.1. Scale-Out NAS Mimarisi

Scale-Out NAS mimarisi, sınırsız depolama, kapasite ve performans ile depolama kolaylığı sağlar. Sürekli yeni donanım eklemeleri yapılarak sistemin depolama, işlem kapasitesi ve performansı yükseltilebilir. Günümüzde benzer şekilde cloud depolama olarak karşımıza çıkan sistemler bu şekilde çalışmaktadır. Büyüyen veri miktarlarına karşın sürekli yeni teknolojiler gelişmekte. Tek bir küme içinde ölçeklendirme yaparak, bellek ve ağ kaynakları işlemler arasında optimize edilir. Yapılandırılmamış veriler günümüzde bu şekilde saklanmaktadır. Bu mimarinin yararları şu şekilde sıralanabilir.

- **Sınırsız büyük kapasite.** Ölçeklenebilir ve milyarlarca dosyayı destekleyen tek noktada, depolama sınırı olmayan bir platform oluşturabilir.
- **İş süreçlerinin performans taleplerini karşılar.** Aynı depolama sistemi içinde, öngörülebilir, tutarlı, düşük gecikme süresi ve yüksek verim sunar.
- **Dağıtımı ve yönetimi basitleştirir.** Büyüyen veri gereksinimlerine ayak uydurmak için tek bir arabirim kullanarak yönetim karmaşıklığı ortadan kalkar.
- **Kullanıcı verimliliğini artırın.** Tüm veriler kullanıcılara aynı konumdaymış gibi sunulur.
- **Donanım arızalarından etkilenmeyin.** Tüm veriler birden fazla disk ortamında yazılabildiği için (RAID) veri kaybolma riskini minimize edilebilir.

3.1.3.2. Obje tabanlı veri depolama mimarisi

Bulut tabanlı sistemler veri erişimi için web tabanlı protokollerin kullanımı yaygınlaşmıştır. Web tarayıcısı üzerinden birçok bulut sistemine veri gönderip veri okuyabiliriz. Bu sistemlerin en büyük özelliği dosya erişimindeki karmaşık CIFS, NFS protokolleri yerine PUT, GET, DELETE gibi birkaç basit komut ile veriyi bulut sistemine yükleyen ya da bulut sisteminden okuyan bu yapı bulut mimarilerin depolama alt yapısını oluşturur.

Kullanıcı tarafından buluta gönderilen verinin hangi lokasyondaki hangi sunucuda, sunucunun hangi dizininde saklanacağı bilinmez, bu bulut sisteminin sorumluluğunda olan bir durumdur. Peki, o zaman veri nasıl geri çağırılacaktır? Buluta yüklenen dosyanın güvenliği nasıl sağlanacaktır? Bu soruların cevapları aşağıda özetlenmiştir:

Bir markette her ürün için bir barkod numarası veriliyor ve kasada barkod okutulduğunda o ürün ile ilgili bilgiler okunuyorsa bulut sistemlerinin de benzer mantıkta çalıştığı söylenebilir. Barkodun veri depolama sistemlerindeki karşılığı olan meta veri kullanılır. Obje tabanlı sistemlerde veri belli bir lokasyon üzerinde tanımlı olmayıp çok sayıda lokasyon üzerine dağıtılmış olabilir. Bu hem yük dengelemeyi hem de objenin birden fazla kopyasının farklı lokasyonlarda tutulabilmesini sağlar.

Bu sistemler ile hem limitsiz ölçeklenebilen hem de limitsiz bir kullanım ömrü olan bir yapı kurulmuş olur (Çoşkun, 2019).

3.1.4. Metin Madenciliği Aşamaları

Metin madenciliği yazılı dijital metinleri kaynak olarak alan, basit ve karmaşık bir sürü teknik işlemlerin ardından bilgisayar sistemlerinin asıl anlayacağı veri türü olan sayısal verilere dönüştürülürler. Nihai olarak sayısallaştırılan metinler istatiksel süreçler yardımıyla işe yarar bilgiler arasında seçim yapılabilir hale gelirler. H. P. Luhn 1958 yılında IBM dergisinde yayınlanan makalesinde literatür özetlemenin otomatik gerçekleştirilmesi konulu bir çalışma yayınlamıştır (Luhn, 1958). O yıllarda kısıtlı kalan bu tarz çalışmalar günümüzde teknolojik araçların daha etkin kullanılabiliyor olması sebebiyle yapılandırılmamış veriden daha fazla işe yarar bilgiler elde edilmeye başlanmıştır. Yapılandırılmamış veriden işe yarar bilgiyi elde etmenin yöntemleri hakkında literatürde farklı bir sürü yöntem görülmüştür. Metin madenciliği aşamaları, metin işleme alanında önemli yeri olan Doğal Dil İşleme (DDİ) aşamalarına paralellik gösterir. Doğal Dil İşleme yapay zekâ başlığı altında yer alan daha özel bir çalışma alanıdır. Çünkü her konuşulan dilin kendine özel kuralları vardır. Bu kurallara göre üzerinde özel

çalışmaların yapılması gereken bir çalışma alanıdır. Literatürde ki bir diğer ismi ise Fikir Madenciliğidir. Bu çalışmada kullanılan metin madenciliği ve fikir madenciliği aşamaları aşağıda tartışılmıştır.

3.1.4.1. Yapılandırılmamış Verileri Keşfetme ve Kayıt Altına Alma

Yapılandırılmamış veri tanımını ve açıklamalarını 3.1.3 başlığı altında incelemiştik. Bu verilerin nasıl saklanacağı konusunu ise günümüzde ve gelecekte nasıl olacak alt başlıkları olan 3.1.3.1 ve 3.1.3.2' de incelenmiştir. Bu süreç yapılan tüm çalışmalarda verilerin internet ortamından elde edildiği, özellikle sosyal medya üzerinden elde edilmiş veriler olduğu görülmektedir. Bu tarz çalışmalarda, çalışmayı yapan kişi ya da kişiler öncelikle ne tür veri aradıklarını, ardından bu tür verileri hangi web sitelerinde veya sistemlerden elde edebileceklerini belirlemeliler. Bu iki aşama tamamlandıktan sonra veriye erişim yönteminin belirlenmesi gerekmektedir. Bunun için hazır paket programlar ya da herhangi bir programlama dili ile bu verilere erişim sağlanabilmektedir. Bu tez kapsamında belirlenen veri türü olan müşteri yorumları sosyal medya ve şikâyet portallarından C# programlama diliyle geliştirilen web kazıma programı ile elde edildi. Böylelikle daha veriye erişim sağlandığı anda bir nebzede olsa veri daha yapılandırılmış bir hale getirildi.

3.1.4.2. Metin Ön İşleme

Metin ön işleme veri madenciliği ve doğal dil işleme gibi metin analizi yapılacak alanlarda olmazsa olmaz, eksik ya da yanlış yapıldığında çıkacak sonuca doğrudan negatif etki edecek bir adımdır. Bu nedenle elde edilen veri üzerinde ön işleme adımlarının doğru şekilde uygulanması şarttır. Aksi takdirde elde edilmeye çalışılan yapılandırılmamış veriden bilgi çıkarma amacı başarısızlık ile sonuçlanacaktır. Ön işleme kendi içinde aşağıdaki bölümlere ayrılmaktadır.

3.1.4.2.1. Dönüştürme

İnternet üzerinde veriler belli dosya uzantıları ile erişilebilirler. Herhangi bir web sitesine girdiğimizde en temelde html uzantılı dosyalar sayesinde resimleri, yazıları, videoları vb. tüm formatları internet tarayıcılarının (Chrome, Mozilla, Explorer vb.) yorumlaması sonucu düzgün anlaşılır bir yapıda görüntüleyebilmekteyiz. Veriler geliştirilen web kazıma programı sayesinde o web sitesine gerçek bir kullanıcı giriyormuş gibi davrandırıldı ve sunucu üzerinde bulunan html uzantılı sayfa sunucudan talep edildi. Veri çekilen sitenin sunucusunda bir web kazıma (bot) programı mı bu sayfayı talep ediyor yoksa gerçek bir kullanıcı mı talep ediyor ayırt edemez. Bu nedenle html kodunu üretir ve talep edilen yere gönderir. Gelen html kod örneği Şekil 3.5' de görüleceği üzere hiç anlaşılır biçimde değildir ve veri işlemek için hiçte uygun olmayan bir yapı karşımızdadır. Bize Şekil 3.5' de mavi ile işaretlenmiş yerde bulunan "Turizm Değil Çile Turizm!" bölümü ve 2 satır altında bulunan <p><p/> etiketleri arasındaki "1 Mart 2019" şeklinde başlayan bölüm gereklidir. Aynı şekilde html dışında diğer web veri taşıma teknolojilerinden XML ve JSON türler içinde benzer çalışmalar yapılmalıdır. Tez kapsamındaki veri kaynağımız html sayfalar olduğu için geliştirilen web kazıma programı sadece HTML verileri ayrıştırıp dönüştürme işlemini yapmaktadır. Bunun etiketsiz şekilde ve saf metin halinde siteden çekilmesi gerekmektedir. Web veri kaynaklarından (HTML, XML, JSON) saf metinleri çekebilmek için paket programlar bulunsa da çok amatör kalacaktır. Hemen hemen her web destekli programlama dilinde veri çekme ve ayrıştırma işini yapabilecek kütüphaneler bulunmaktadır. Daha önceden deneyim kazanmış olduğumuz dil olan C# Programlama Dili ile bu veri çekme ve ayrıştırma işlemini gerçekleştirdik.

Şekil 3.5. Örnek html kod yapısı

3.1.4.2.2. Kelime Köklerini Bulma

Dilimiz Türkçe, sondan eklemeli bir dildir. Kelime köklerine yapım ve çekim ekleri gelerek kelimeler yeni anlamlar kazanmaktadır. Bu nedenle dilimizde duygu analizi çalışmalarını daha büyük emek ve çaba gerektirmektedir. Dilimizde kökler kendi başlarına anlam taşıyabilirler. Otur, kalk, yürü, oku, gör vb. Bu kelimelere birden fazla yapım ve çekim ekleri geldiğinde yeni anlam kazanmış birçok kelime ortaya çıkmaktadır. Kök bulma işlemi için Türkçe dili için geliştirilmiş birçok araç bulunmaktadır. Akademik çalışmalara bakıldığında Zemberek açık kaynak kodlu yazılımının bu konuda açık ara önde olduğu görülmektedir. Aynı zamanda Python NLTK kütüphanesi de kelime köklerini bulmak için kullanılmaktadır.

3.1.4.2.3. Yazım Kurallarına Uygunluk ve Türkçe Tespiti

Cümle içerisinde geçen kelimeler Türkçe kelimeler olmayabilir. Türkçe yapısına uymama ihtimali yüksek bu yabancı kelimeler tespit edilmelidir. Aynı zamanda yanlış yazılmış, kısaltma kullanılmış kelimelerinde düzeltilmesi gerekmektedir. Bu işlem adımı için Zemberek doğal dil işleme kütüphanesi kullanılmıştır.

3.1.4.2.4. Kelime ve Kelime Gruplarının Anlamsal Değerlerini Bulma

Cümle içerisinde geçen kelime ve kelime gruplarının isim, sıfat, fiil, zarf, zamir, gibi anlamsal değerlerinin tespit edilmesi gerekir. Bu işlem adımı için Zemberek doğal dil işleme kütüphanesi kullanılmıştır.

3.1.4.2.5. Durak Kelimeleri Ayrıştırma ve Cümleden Çıkartma

Cümle içerisinde tek başına anlamı olmayan ve bir bilgi içermeyen ancak yanına başka bir kelime geldiğinde anlam kazanabilen kelimelerdir. Örnek olarak ve, veya, ancak, lakin, çünkü, ile gibi sözcüklerdir. Bu kelimelerin cümle içinden çıkartılması gerekmektedir. Yine bu aşama için en yaygın Zemberek açık kaynak kodlu yazılım ve Python NLTK kütüphaneleri kullanılmaktadır.

3.1.4.2.6. Terim Ağırlıklandırma

Metin madenciliği ve doğal dil işleme alanlarının temel problemi, bir belgenin ne hakkında, neler içerdiğini, istatiksel olarak hesaplamak ya da duygu skorunu ölçerek belgeden bilgi alma tekniğidir. Terim ağırlıklandırma yöntemi istatiksel olarak belgeyi incelemeye izin verir.

3.1.4.2.6.1. Terim Frekansı

Terim Frekansı, bir terimin belgede ne sıklıkta yer ya da ne kadar geçtiğini ölçen bir yöntemdir.

3.1.4.2.6.2. Ters Doküman Sıklığı

Ters doküman sıklığı (Inverse Document Freqency, IDF), bir terimin derlemdeki (corpus) diğer belgelerde geçme sıklığını ifade eder. IDF hesaplanırken eşitlik-3.1' deki formül kullanılır.

$$idf(w) = \log_{10}\left(\frac{N}{df(w)}\right) \quad (3.1)$$

Burada; N, derlemde bulunan toplam belge sayısını, df(w), w teriminin geçtiği belge sayısını ifade eder. Ölçeği küçültmek için N değerinin df(w) değerine bölümünün logaritması alınır (Şeker, 2019).

Bir terim, bir belge içerisinde ne kadar çok geçiyorsa o kadar değerlidir. Bir terim, derlemdeki diğer belgelerde ne kadar çok geçiyorsa belge için ayırt edici olma gücü o kadar düşüktür (Şeker, 2019).

Bu iki değer Python dili kullanarak hesaplanmış ve bir örneği Çizelge 3.2' de verilmiştir.

3.1.4.2.6.3. Terim Frekansı ve Ters Doküman Sıklığı Ağırlıklandırma Örneği

TF-IDF değeri hesaplanmasında iki sayının tespit edilmesi gerekmektedir. Bu iki önemli sayıdan birincisi incelenecek olan dokümandaki terimin sayısı diğeri ise bu terimi içeren toplam dokümanların sayısıdır (Şeker, 2019).

Çizelge 3.2. Telekominasyon sektörüne ait markanın IDF x DF değerleri

No	Telekom	Internet	Müşteri	Fatura	Iptal
0	0,117622	0	0	0	0
1	0,075651	0	0	0	0
2	0,042564	0,096142	0,052795	0	0,0784507
3	0,071945	0,081253	0	0	0
4	0,139319	0,039336	0,043201	0	0
5	0,034676	0	0,043011	0	0
6	0	0	0,038573	0,051173	0,2865904
7	0,089458	0	0	0	0
8	0,116439	0,131503	0	0	0
9	0	0	0	0,144342	0
10	0,117622	0	0	0	0

3.1.4.2.7. Terim Ayıklama

Bir metin içerisinde bir terim sadece bir kez geçiyor ise. Metin içerisinden çıkartılabilir. Bu terim artık o metin için göz ardı edilebilir. Bu şekilde işlenecek terim sayısında ciddi düşüşler gerçekleştirilebilir.

3.1.4.3. Özellik Çıkarma

Veri ve metin madenciliği işlemlerinde ilgisiz olabilecek veri kümelerini, işlenecek veri kümesinden çıkartmak bunun sonucunda da daha küçük veri boyutları ile çalışmak daha sağlıklı ve sonuca odaklı bir çalışma yapılmasını sağlayacaktır. Veri ve metin madenciliği işlem adımlarının en önemli adımlarından biri bu nedenle özellik çıkarma işlemidir.

Özellik Çıkarma İşleminin Faydaları şunlardır (L.Ladha & T.Deepa, 2011):

- Özellik kümesinin boyutunu azaltır,
- Algoritma hızını arttırır,
- Gereksiz ve gürültülü veriyi ortadan kaldırır,
- Veri analizi görevlerinin sürelerini kısaltır,
- Veri kalitesini artırır,
- Veri kümesini oluşturmak için gerekli olan veri toplama işleminde kaynak tasarrufu sağlar,
- Elde edilen modelin doğruluğunu artırır,
- Veri kümesini daha basit bir şekilde tanımlanabilir, görselleştirilebilir ve anlaşılabilir hale getirir,
- Veri depolamak için gerekli olan hafıza miktarını azaltır.

3.1.4.4. Sınıflandırma

Veri ve metin madenciliğinde sınıflandırma; veri kümesi üzerinde bulunan ortak özelliklere göre veriyi sınıflar arasında olabilecek en iyi tanımlı sınıfa dağıtma işlemi diyebiliriz. Bu işlem adımından sonra bile veri kümesi üzerinde gizli bilgiler görünür hale gelmektedir.

3.1.4.4.1. Sınıflandırma Süreci

Sınıflandırma için tarihsel verinin (historical data) önemi büyüktür. Örneğin hava tahmin raporlarının günümüzde büyük başarı ile doğru tahmin edilmesinin en büyük neden tarihsel verilerin saklanıyor ve işleniyor olmasıdır.

Verinin sınıflandırılma süreci iki adımdan oluşur. Birincisi veri kümesine uygun bir modelin ortaya koyulması ikincisi ise belirlenen test verilerine göre sınıflandırma kurallarının uygulanmasıdır.

3.1.4.4.1.1. Uygun Model Belirlenmesi

Uygun model belirlenmesi şöyledir;

- Model, database üzerindeki verilerin öznitelikleri kullanılarak gerçekleştirilir.
- Sınıflandırma modelinin oluşturulması için verilerin bir bölümü eğitim verisi olarak kullanılır.
- Eğitim verisi veriler arasından random seçilir.
- Veriye bir sınıflandırma algoritması uygulanarak sınıf modeli elde edilir.

Çizelge 3.3'teki veri modeli, banka müşterilerinin sınıflandırılması için kullanılabilir. Müşterinin kredi talebine sistem tarafından olumlu ya da olumsuz yanıt verilmesi sınıf modeli eğitilerek verilebilir.

Çizelge 3.3. Örnek eğitim veri modeli

Müşteri	Borç	Gelir	Risk
A	Düşük	Yüksek	İyi
B	Yüksek	Düşük	Kötü
C	Düşük	Düşük	Kötü
D	Yüksek	Düşük	Kötü
E	Yüksek	Yüksek	Kötü
F	Düşük	Yüksek	İyi

Örneğin sınıf modeli şu şekilde:

- Eğer borç yüksekse; risk kötüdür,
- Eğer borç ve gelir düşükse; risk kötüdür,
- Eğer borç düşük gelir yüksekse; risk iyidir

Bu şekilde tanımlanan modele göre müşteri talebi olumlu ya da olumsuz karşılanabilecektir.

3.1.4.4.2. Vektör Oluşturma

Derlemlerimiz içindeki cümlelerde ve cümle öbeklerinde bulunan kelimelerden vektör oluşturulması işlemidir. Unigram, bigram, trigram yöntemleri ile bu vektörler oluşturulur. Cümlelerin içinde bulunan kelimeler ikili, üçlü halde de anlam kazanmış olabilirler. Bunu bölümleyip anlamanın yöntemi unigram, bigram, trigram olarak gruplamaktır. Örneğin "HP Bilgisayarı aldığım günden beri ne zaman video izlesem sürekli ses geliyor, rahat rahat müzik dinleyemiyorum. Bir şey izleyemiyorum sıkıldım artık" cümlesini Şekil 3.6, Şekil 3.7 ve Şekil 3.8'da inceleyelim. Bu tez kapsamında incelenen metinlerin Vektörleri R Dili kullanılarak oluşturulmuştur.

```
 1  hp
 2  bilgisayarı
 3  aldığım
 4  günden
 5  beri
 6  ne
 7  zaman
 8  video
 9  izlesem
10  sürekli
11  ses
12  geliyor
13  rahat
14  rahat
15  müzik
16  dinleyemiyorum
17  bir
18  şey
19  izleyemiyorum
20  sıkıldım
21  artık
```

Şekil 3.6. Cümlenin Unigram (1-Gram) gösterimi

```
hp bilgisayarı
bilgisayarı aldığım
aldığım günden
günden beri
beri ne
ne zaman
zaman video
video izlesem
izlesem sürekli
sürekli ses
ses geliyor
geliyor rahat
rahat rahat
rahat müzik
müzik dinleyemiyorum
dinleyemiyorum bir
bir şey
şey izleyemiyorum
izleyemiyorum sıkıldım
sıkıldım artık
```

Şekil 3.7. Cümlenin Bigram (2-Gram) gösterimi

```
Bilgisayarı aldığım günden
Bir şey izleyemiyorum
HP Bilgisayarı aldığım
aldığım günden beri
beri ne zaman
dinleyemiyorum. Bir şey
geliyor, rahat rahat
günden beri ne
izlesem sürekli ses
izleyemiyorum sıkıldım artık.
```

Şekil 3.8. Cümlenin Trigram (3-Gram) gösterimi

3.1.4.4.2.1. N-Gram Oluşturma

N-gram tahmin etmeye ve olasılığa dayanan bir sistemdir. Kelimeleri ya da harfleri birli, ikili, üçlü ya da daha fazla bölümleme gruplama işlemidir. Bunun en

önemli örneği Google aramalarımızda arama motorunun bize bunu mu demek istediniz şeklinde sorduğu sorudur. Şekil 3.9'da görülen Isparta aramasında bize 1-Gram, 2-Gram, 3-Gram, 4-Gram olarak arama motorunun tahminini göstermektedir. Bir diğer karşımıza çok çıkan örnek ise akıllı telefon klavyelerinin yanlış yazılan bir kelimeyi düzeltmesi, yine akıllı telefon klavyelerinin çokça kullandığımız "eve gidiyorum" gibi cümleleri eve yazdığımız anda "eve gidiyorum" demek isteyebileceğimizi tahmin edip öneride bulunmaktadır.

Şekil 3.9. Google N-Gram Örneği

3.1.4.4.3. Sınıflandırma Yöntemleri

İşlenecek verinin ya da metinin sahip olduğu ortak özelliklere göre gruplara ayrılıp bu gruplarının etiketlenmesi işlemi sonucunda veri içerisinde gizli bilgilerin ortaya çıkartılmasına yardımcı olan yöntemlerin genel adına sınıflandırma denilmektedir. Elde bulunan verinin ya da metnin bir kısmı eğitim amaçlı kullanılır. Bölüm 3.1.4.4.1.1' de örnek veri üzerinde öznitelikler belirlenmiş ve sınıflandırmak için modele uygun hale getirilmiştir. Sınıflandırma için literatürde birçok algoritma bulunmaktadır. Bu algoritmalar, bu bölümün alt başlıklarında açıklanmıştır.

3.1.4.4.3.1. Naive Bayes

Naive Bayes sınıflandırıcısı temel olarak Bayes teoremini temel alır. Öğrenme algoritmalarından bir tanesidir. Çalışma sistemi; veri içinde ki her bir eleman için olasılık değerini hesaplar. Sonuç olarak olasılık değeri en yüksek olana göre sınıflandırmasını yapar. Bayes formülü şu şekildedir.

$$P\,(\,c\,\mid x\,)=\frac{P\,(\,c\,\mid x\,)\,P(c\,)}{P\,(x)} \tag{3.2}$$

Burada;

c: Tahmin edilmeye çalışılan sınıf,

x: Tahmin eden sınıf,

P(c|x): x olayı gerçekleştiğinde c olayının gerçekleşme olasılığı,

P(x|c): c olayı gerçekleştiğinde x olayının gerçekleşme olasılığı,

P(c): c olayının gerçekleşme olasılığı,

P(x): x olayının gerçekleşme olasılığıdır.

3.1.4.4.3.2. K-En Yakın Komşuluk

Eğitim ve test kavramlarının bir arada olduğu, örnek tabanlı öğrenme algoritmasıdır. Yeni karşılaşılacak bir örnek, eğitim verisinde bulunan örnekler arasında ki benzerlik oranına göre sınıflandırılırlar.

3.1.4.4.3.3. Destek Vektör Makinesi

1963 yıllarında Vladimir Vapnik ve Alexey Chervonenkis tarafından geliştirilen "Destek Vektör Makineleri (DVM)" istatiksel öğrenmeye dayalı gözetimli öğrenme algoritmasıdır. Çoğunlukla sınıflandırma ve regresyon zorluklarının aşılması amacıyla kullanılır.

3.1.4.4.3.4. Karar Ağacı

Karar ağacı, veri kümesinin çok kayıt içerdiği durumlarda, karar kuralları uygulanarak daha küçük kümelere ayrılıp kullanılan bir yapıdır. Basit karar verme adımları uygulanarak, büyük miktarlardaki verileri, çok küçük veri kümelerine bölerek kullanılır.

3.1.4.4.3.5. Yapay Sinir Ağları

İnsan beyninin çalışma ve öğrenme şeklinin modellenmeye çalışıldığı bir yöntemdir. Aynı insanlarda olduğu gibi yapay sinir ağları ile sistemlerin eğitilmesi, öğrenmesi ve karar vermesi amaçlanmaktadır.

3.1.4.4.3.6. Genetik Algoritmalar

Genetik Algoritma (GA), Darwin'in evrim teorisine dayanmaktadır. Mühendislik problemlerinde optimizasyon amacıyla kullanılmaktadır. GA'lar çok karmaşık optimizasyon problemleri için bile çözüm bulabilirler. GA'ların ihtiyaç duyduğu şey problemin karar değişkenlerinin uygun bir yöntemle kodlanması ve neyin iyi olduğunu GA'ya belirtmek üzere tasarlanan bir uygunluk (amaç) fonksiyonudur. GA'lar çözüm uzayını taramaya bir topluluk ile başladıkları için global optimum çözüme yaklaşmak diğer yöntemlere göre daha kolay olmaktadır. Genel olarak global optimum çözümü bulmayı garanti etmezlerse de buna yakın bir sonucu bulduğu birçok araştırmayla ispatlanmıştır. GA'lar bir topluluk (başlangıçta bu topluluk genelde rastgele oluşturulur) ile başlar ve bu topluluk üzerinde çaprazlama, seçme ve mutasyon gibi yöntemlerin uygulanmasıyla problemin her aşamasında en iyiye doğru gidiş sağlanır.

3.2. Doğal Dil

Doğal dil insanların birbiriyle iletişim kurması için gerekli bir araçtır. İnsanların günlük hayatlarında, sosyal çevrelerinde, iş yaşamlarında sessiz kalmaları mümkün değildir. İnsan olarak birbirimizle konuşarak, yazarak anlaşırız.

Geçmişten günümüze yeryüzünde konuşulan ve metin halinde yazılan dil sayısı oldukça fazladır. Teknolojinin gelişmesiyle kullandığımız dilin alfabesindeki tüm harflerin dijitalleştiğini, dilin kurallarında olduğu gibi birleştirilerek karşımıza e-posta, SMS, internet sayfası, reklam ve daha nice vb. örnekle karşımıza çıktığını görmekteyiz. İletişim araçlarının neredeyse tümden teknoloji araçlarıyla yapılmasından dolayı dilin işlenmesi gereğini ve dilin işlenmesi konusunda çalışan bilim insanı sayısını artırmaktadır. Bu çalışma da tamda bunu hedeflemekteyiz. İnsanlar artık bir ürün alırken araştırma yapmadan o ürünü tercih etmemekte. Tercih eden insanlarda ürünle ilgili fikirlerini ve yorumlarını yine iletişim araçlarını ve interneti kullanarak paylaşmaktadırlar.

3.2.1. Duygu Analizi

Duygu analizi üç temel seviyede incelenebilir. Bunlar, Belge (doküman) seviyesi, Cümle seviyesi, Varlık ve özellik seviyesidir.

3.2.1.1. Belge Seviyesi

Tüm belgenin herhangi bir kategoriye dahil edilmeden olduğu gibi bütün şekilde analiz edildiği seviyedir. Örneğin bir üniversiteye ait tüm yorumların toplandığı dokümanın analiz edilip üniversite hakkında pozitif ya da negatif duygu ifade edilip edilmediği araştırılır. Bu seviyede detaya inilmez, doküman bir bütün olarak ele alınır.

3.2.1.2. Cümle Seviyesi

Belge seviyesinde analiz bazı durumlarda yetersiz kalmaktadır. Her cümlenin ya da yorumun ayrı ayrı kendi içinde değerlendirilip pozitif, negatif ya da nötr olduğu değerlendirilir. Her cümle değerlendirildikten sonra doküman genelinde ağırlıklı duygu belirlenmeye çalışılır. Doküman seviyesine göre daha belirgin sonuçlar elde edilen yöntemdir.

3.2.1.3. Varlık ve Özellik Seviyesi

Doküman ya da cümle seviyesi analizde yorumun pozitif çıkıyor olması yorumun barındırdığı varlık ya da özelliğin tümünün pozitif olduğu anlamına gelmeyebilir. Böyle bir durumda varlığın hangi özelliği için pozitif hangi özelliği için negatif duygu içerdiği tespit edilmelidir. Örneğin, yine üniversite örneğini ele alacak olursak; üniversite hakkında yorum yapan bir öğrenci işleri hizmeti için negatif, yemekhane hizmetleri için pozitif, kütüphane hizmetleri için nötr duygu barındıran yorum yapmış olabilir. Bu gibi durumlarda varlık özellik seviyesi incelenmelidir. Seviyeler arasında en zor olan analiz yöntemi bu yöntemdir.

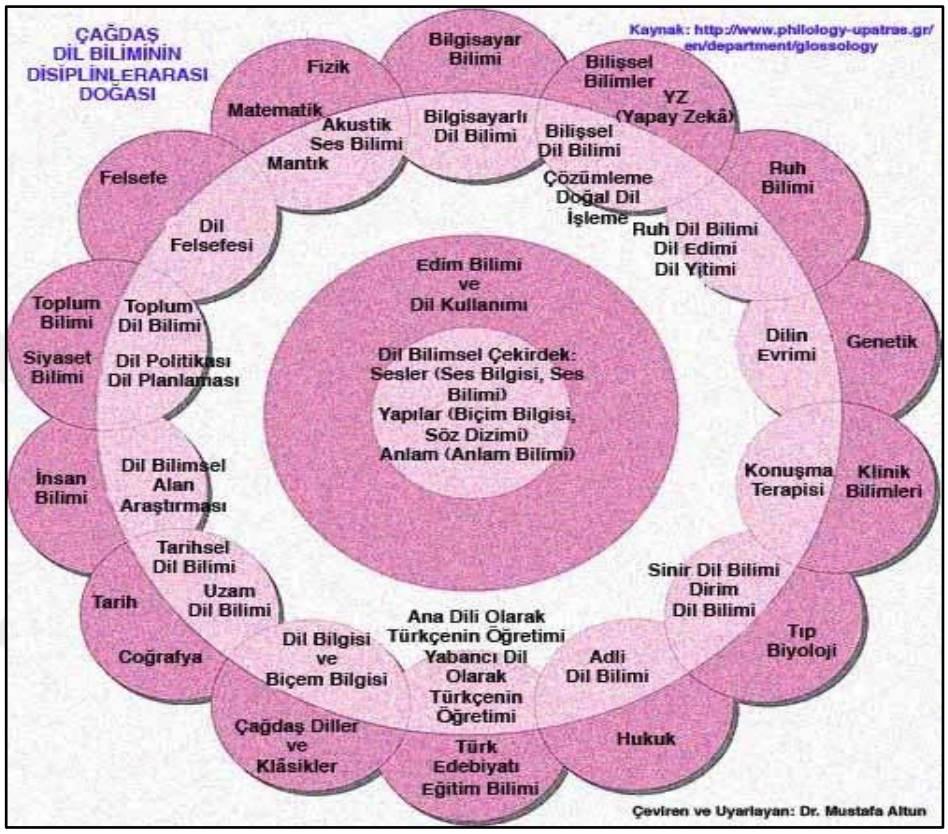
3.2.2. Doğal Dilin Zorluğu

Doğal dil insanın doğumundan itibaren kazanmaya başladığı yetenek olarak tanımlanabilir. Okul çağına gelesiye kadar yalnızca konuşmayı, konuşulanı anlamayı öğreniriz. Bu süreç insan hayatı boyunca sürecek bir öğrenmedir. Her dilin kendine özgü kuralları, sembolleri, araçları bulunmaktadır. Dili edinmek ve doğru kullanmak insanın doğuşu ile başlayan ve eğitim-öğretim hayatında filizlenip, insanın ölümüne kadar devam eden bir süreçtir. Bu süreç zor ve uzun bir zaman diliminde gerçekleşmektedir. Kaldı ki Türkçe sondan eklemeli bir dil olduğu için doğal dili bilgisayar ortamına taşıyabilmek diğer dillere göre daha zor bir hale gelmektedir. Bu nedenle hem bilim adamlarının bu konu üzerinde daha fazla durmasını gerektirme hem de bu alanda çalışan bilim adamı sayısının artması gerekmektedir.

3.2.3. Dil Bilim

Dil bilimi "lengüistik ya da lisaniyat; dilleri dil bilgisi, söz dizimi (sentaks), ses bilgisi (fonetik), ses bilimi (fonoloji), biçim bilimi (morfoloji) ve edim bilimi (pragmatik) gibi çeşitli yönlerden yapısal, anlamsal ve bildirişimin çıkış bağlamını temel alarak sözlerin gönderimlerini ve iletişimde dilin yaptırım gücünü inceleyen bilim dalı" (TDK, 2015) .

Genel (veya kuramsal) dil bilimi dillerin yapılarını (dil bilgisi) ve anlamlarını (anlam bilimi) inceler. Dil bilgisinin incelenmesi, biçim bilimi (sözlerin oluşumu ve değişimi) ve söz dizimini (sözlerin ifade veya cümle oluşturmak için bir araya getirilmesi ile ilgili kurallar) kapsar. Dili sesler aracılığıyla ifade etmek için kullanılan sistem olan ses bilimi de bu alanın bir parçasıdır (Wikipedia, 2019).



Şekil 3.6. Çağdaş dil biliminin disiplinler arası doğası (Dil Bilimi, 2019)

Dil biliminin ve doğal dil işlemenin gelişmesi yapay zekâ biliminin gelişmesine katkı sağladığı bir gerçektir. Şekil 3.10'da dil biliminin disiplinler arası etkileşimini görmektesiniz. Dil biliminin genetik, tıp, tarih, coğrafya, bilgisayar bilimleri gibi birçok disiplinler ile iç içe bir çalışma alanıdır.

Bir robot düşünelim. Basit komutları yerine getirebiliyor olsun; otur, kalk, yürü, dur, zıpla vb. Verdiğimiz komut Türkçe diline ait bir komut eğer bizim verdiğimiz komutu işleme alırken otur komutu verildiğinde "oturma" eylemini yapması gerektiğini anlamaz ise bu robot yapay zekalı bir robot tanımı içerisine girmeyecektir. İşte dilbilimi ve doğal dil işleme arasında ki bağ yapay zekâ ile

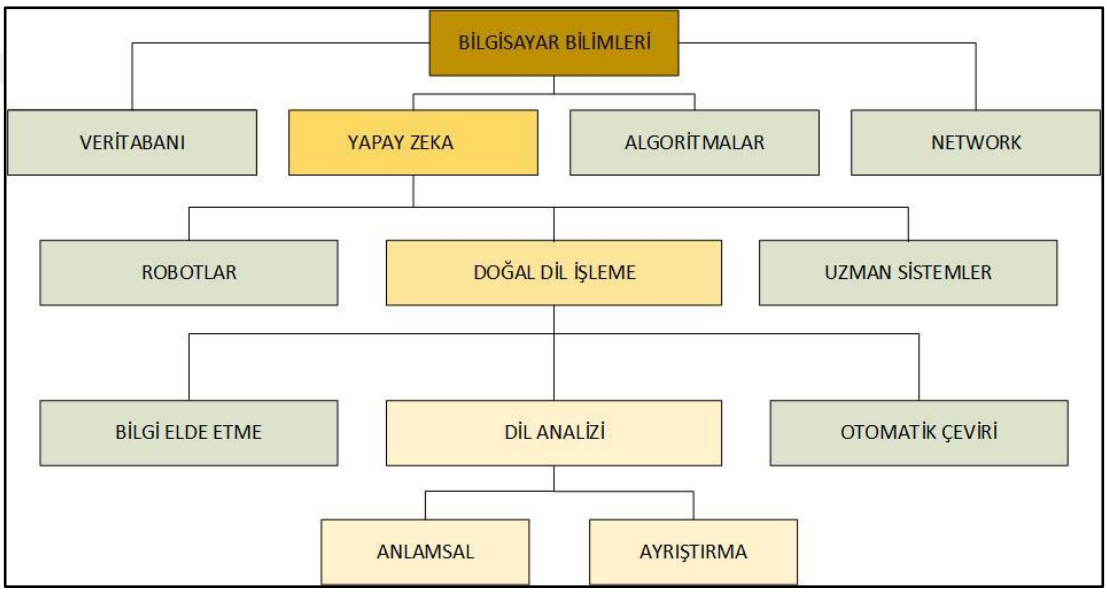
Türkçe doğal dil işleme arasında kurulabilir ve çalıştırılabilirse yapay zekâ konusunda dünya ile yarışıyor hale gelebiliriz. Benzer çalışmalar ülkemizde Türkçe için yapılıyor olsa da Türkçe'nin dil bilimsel zorlukları ve bu alanda yapılan çalışmaların azlığı nedeniyle kısıtlı kalmakta. Yapay zekanın gelişimi açısından çalışılması gereken asıl konu literatürde Sentiment Analyzer Türkçe adıyla Duygu Analizi olarak geçen diğer bir tanımıyla bilgisayarın yazılı olan metini anlayabilmesi diyebiliriz. Öznel bilgileri, duygusal durumları bilgisayar tarafından tespit edilebilmesine olanak sağlayan bir alt disiplindir. Duygu analizi konusuna ayrı bir başlıkta detaylı şekilde inceleyeceğiz.

3.2.4. Doğal Dil İşleme

Doğal Dil İşleme (DDİ), İngilizce adıyla NLP (Natural Language Processing); yapay zekâ ve dilbilimin alt kategorilerinden biridir. Türkçe, İngilizce, Almanca, Fransızca gibi doğal dillerin (insana özgü tüm diller) insan ile bilgisayar arasında ki iletişimi sağlamak için o dilin Dil Bilimini inceleyerek kurallarını çözerek bilgisayar tarafından işlenmesini, anlaşılmasını sağlayan bilim dalıdır. Doğal dillerdeki metinlerin bilgisayarın diline aktarılması süreci her dil için farklılık göstermektedir.

Dil bilim ve bilgisayar bilimlerinin ortak çalışmalarıyla ortaya çıkan doğal dili bilgisayar ortamında analiz edebilme amacını taşır. DDİ Dil Bilim, Bilgisayar Bilimi ve Yapay Zekâ ile iç içe bir çalışma alanıdır (Şekil 3.11). Bu alanın en güzel örneklerinden biri Apple firmasının mobil cihazlarında bulunan Siri komutuyla devreye giren ve Google firmasının Android işletim sistemli mobil cihazlarında bulunan OK Google komutuyla devreye giren insan hayatını kolaylaştırıcı özellikleri bünyesinde bulunan sesli komut ile çalışan dijital asistan yazılımlarıdır. Başka bir deyişle, yalnızca sizden seslerinizle komutları alan, yorumlamaya çalışan ve mümkünse gerekli görevi yerine getiren IOS ya da ANDROID cihazlar için sesle çalışan dijital bir asistandır. Konuşulan dili anlayarak cihazlarda yapmak istediğimiz alarm kurmak, ismini söylediğimiz kişiyi rehberden bularak aramak, söylediklerimizi mesaj olarak yazmak ardından söylediğimiz kişiye o mesajı göndermek, internet üzerinde eller serbest şekilde

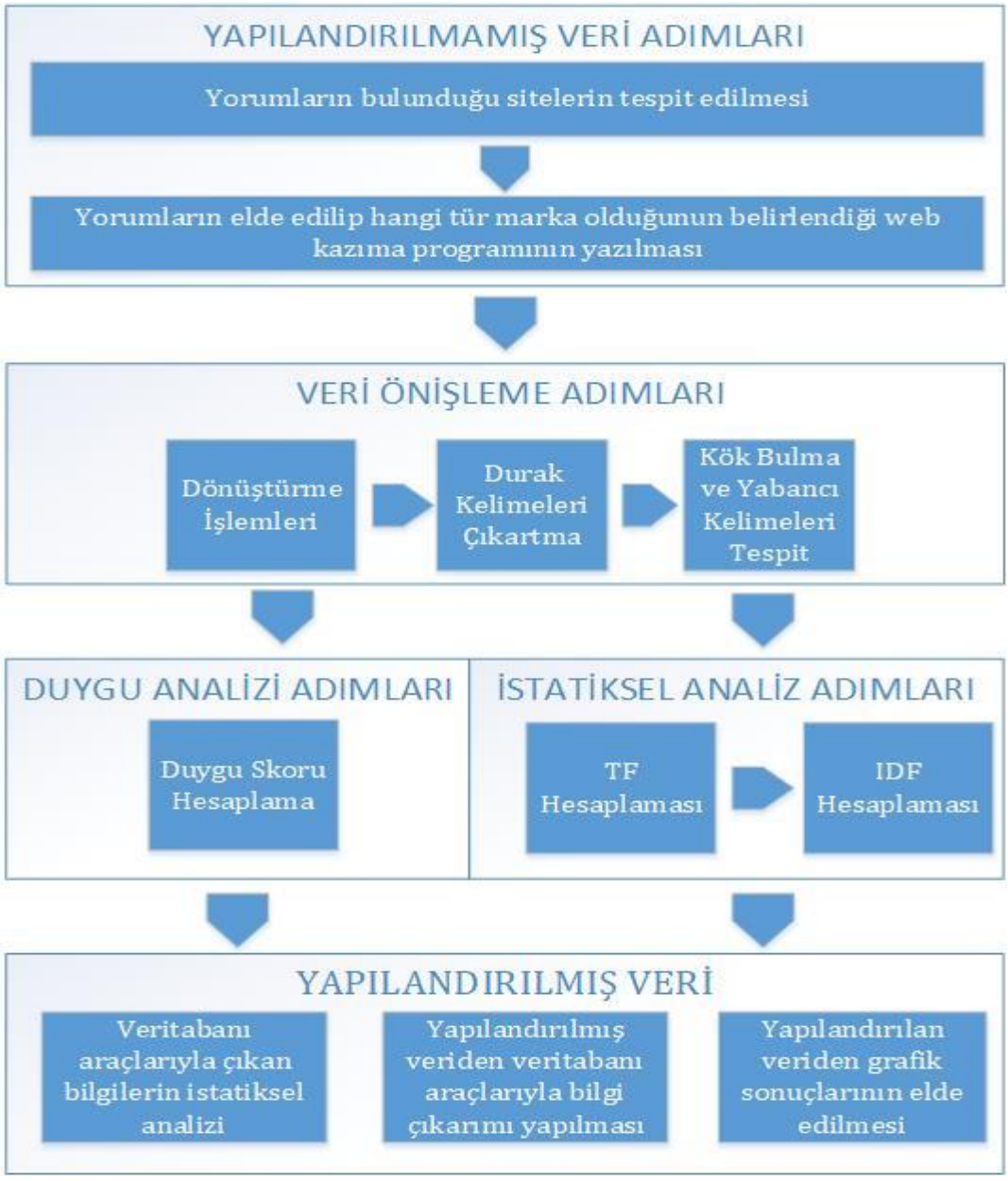
gezinti yapmak gibi bir sürü özelliği bünyesinde barındıran bir sistemdir. Aynı şekilde yine Google ve Yandex vb. firmaların dilden dile çeviri sistemleri de DDİ'nin en büyük öncüleridir. Günümüzde bir sürü uzman sistemin yerini alan bu tarz yazılımlar yavaş yavaş insan eliyle yapılan bir sürü işlemi de ortadan kaldırıyor ve insana olan bağımlılığı azaltıyor. DDİ; duygu analizi, otomatik çeviri, soru cevap, siber güvenlik alanında; domain isimlerinin doğruluğunu tespit etmek, açık kaynak kod güvenliğinin analiz edilmesi, oltalama vb. alanlarda, mail spamları ile mücadele edilmesi, makine diline çeviri yapılması, bilgi çıkarımı, özet çıkarımı vb. alanlarda kullanımı yaygındır.



Şekil 3.7. Doğal dil işleme üst dalları

4. YÖNTEM

Bu tez çalışmasında bahsedilen karar verme sorunu için 13 sektör ve 167 markaya ait 2.089.326 tüketici yorum ve şikâyetleri yazılım geliştirme teknolojileri ile elde edildi. Elde edilen bu yorum ve şikâyetler veri tabanında işlenmek üzere kayıt altına alındı. Herhangi bir ürünü satın almayı hedefleyen muhtemel tüketiciye doğru karar vermesine yardımcı olacak bir sistem haline getirildi. Bu aşamalar şu şekilde gerçekleştirildi.



Şekil 4.1. Gerçekleştirilen çalışmanın adımları

Detaylı Adımlar:

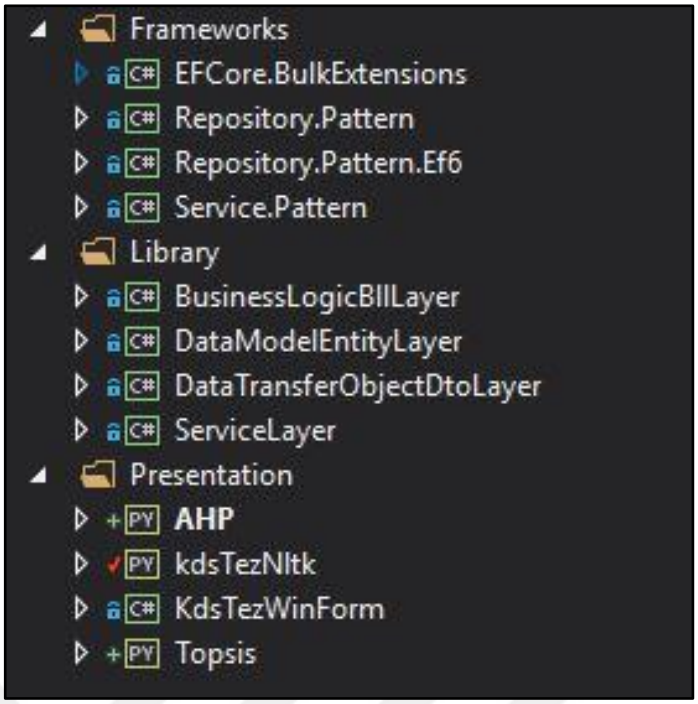
1. **Yapılandırılmamış Veri Adımları**
    - **a.** Yorumların bulunduğu sitelerin tespit edilmesi
    - **b.** Yorumların elde edileceği ortak kodun yazılması
        i. Elde edilen yorumların ham haliyle hangi türe, markaya ait olduğu bilgisiyle beraber veri tabanına kayıt edilmesi. Veri tabanı performansı için indexlerin oluşturulması. Tablo ilişkilerinin oluşturulması
2. **Metin Madenciliği Veri Ön İşleme Adımları**
    - **a.** Gereksiz karakterlerden temizleme araçlarının geliştirilmesi ve uygulanması
        i. Dönüştürme İşlemleri
        ii. Durak Kelimeleri Çıkartma
        iii. Kök Bulma ve Yabancı Kelimeleri Tespit Etme
3. **Duygu Analizi ve İstatiksel Analiz Adımları**
    - **a.** Duygu Analizi & Doğal Dil İşleme
        i. Duygu Skoru Hesaplama
    - **b.** İstatiksel Analiz
        i. TF Hesaplaması
        ii. IDF Hesaplaması
4. **Yapılandırılmış Veri**
    - **a.** Veri madenciliği ve doğal dil işlemleri yapıldıktan sonra elde edilen yapılandırılmış veriden bilgi çıkarımı araçlarının geliştirilmesi ve uygulanması.
    - **b.** Yapılan bilgi çıkarımlarının veri tabanı sorgu araçları ile istatiksel olarak hangi ürün türü içinde hangi marka tercih edilmeli hesaplamalarının yapılması.
    - **c.** Yapılan hesaplamalar sonucunda oluşan verilerin marka ve kategori bazlı grafiklere dönüştürülmesi.

Bu adımların ardından kullanıcının rahatlıkla almak istediği markanın iyi yanlarını kötü yanlarını tercih edip etmemesi sorununa tavsiye niteliğinde yönlendirmeler yapılabilir hale gelmiştir.

Tez çalışmasının tüm yazılım süreçlerini ve verileri tek bir proje çatısı altında toplamak öncelikli olması gereken bir durumdur. Microsoft firmasının geliştirmeye devam ettiği Visual Studio programının 2017 versiyonunda tez çalışmasını gerçekleştirme evresinde ihtiyaç duyulan tüm alt yapı ve birden fazla programlama dili desteğini sağladığı için bu programın kullanılması tercih edilmiştir. Şekil 4.2 'de on iki katmandan oluşan bir programlama yapısı ile tez çalışması gerçekleştirilmiştir.

Frameworks ve Library gruplarının altında bulunan katmanlarda veri tabanı bağlantısı, güvenlik, ortak veri tabanı işlemlerinin tümü tek noktada yönetilebilir olarak kodlandı. Tüm ekleme, silme, güncelleme, listeleme vb. işlemler bu bölümlerde gerçekleştirildi. İstenilen programlama dilinden bu kodlara güvenli ve sağlıklı erişmek için bu şekilde bir yazılım mimarisi tasarlandı.

Presentation grubunun altında bulunan katmanlarda veri ile yapılacak tüm işlemler; veri çekme, veri işleme, veri analiz, doğal dil işleme, istatistiksel analizlerin gerçekleştirilmesi Form ve Console ara yüzleri aracılığıyla görsel olarak gerçekleştirildi. C#, TSQL, PYTHON, R dilleri kullanıldı.



Şekil 4.2. Tez çalışmasında kullanılan yazılım geliştirme katmanları

4.1.Verilerin Elde Edilmesi

Tezin tamamlanması sonucunda ulaşılması gereken hedef hangi ürün ya da marka tercih edilmeli sorusuna yanıt bulmaktır. Bu soruya yanıt bulmak için yazılımsal, istatistiksel süreçler çalışıldı. Bu süreçlerin en başında yer alan veriye ulaşmak gerekmekteydi. Müşterilerin hizmet ve ürün satın alması sonucu şikâyet ve memnuniyetlerini metinsel olarak kayıt ettikleri internet siteleri keşfedildi. Bu internet sitelerinden sektör, marka ve ürün bazlı yorumların hepsi geliştirilen program ile elde edilip veri tabanına kayıt edildi. İlişkisel veri tabanı üzerinde veriler düzenli ve yönetilebilir sorgulanabilir şekilde kayıt altına alındı.

4.1.1. Verilerin Tutulduğu SQL Server Programı

Veri tabanı, verilerin düzenli bir arada tutulduğu bir yönetim sistemidir. İnternet üzerinde dağıtık vaziyette olan verileri bir araya getirip, düzenli ve birbiriyle ilişkili şekilde bir veri tabanı sistemi üzerinde tutmamız gerekiyor. Doğal Dil İşleme ve veri işleme adımlarının ardından, her yorum için 0-1 değerleri arasında ne kadar pozitiflik ne kadar negatiflik ne kadar nötrlük durum içerdiği bilgisini

kayıt altında tutacağız. Düzenli olarak bu verileri tutmaz isek hangi markanın ya da hizmetin toplam pozitif değeri rakip markaya oranla düşük ya da yüksek olduğunu tespit edemeyiz. Bu nedenle yukarda bahsedilen tablolar aracılığıyla veriler düzenli hale getirildi. Tüm bu işlemleri gerçekleştirebilmek geliştirme aşamasında kolaylık sağlayan özellikleri nedeniyle Microsoft SQL Server yazılımı tercih edildi.

SQL Server Microsoft tarafından uzun yıllardır geliştirilen ilişkisel veri tabanı yönetim yazılımıdır. Tüm veri tabanı işlemlerinin sağlıklı şekilde yürütüldüğü, veri saklama ve veri güvenliği yetenekleri açısından ön planda olan genişletilebilir yönetilebilir bir sistemdir.

4.1.2. Kullanılan Programlama Dilleri ve Geliştirilen Ekranlar

Proje kapsamında birden fazla dilin öne çıkan özellikleri ihtiyaca göre kullanıldı. Proje derleyicisi olarak Visual Studio 2017 versiyonu kullanıldı. Visual Studio programı Microsoft firması tarafından geliştirilen çoklu dil desteği ile geliştiricilere kolaylık sağlayan yazılım geliştirme platformudur. Tez uygulamasının yazılım proje mimarisi, Visual Studio çatısı altında 3 ana katmandan oluşmaktadır.

EFCore.BulkExtensions bu bölümde C# Programlama ile internet sitelerinden elde edilen veriler SQL SERVER veri tabanı programına daha hızlı kayıt edilmesi için zorunlu olmayan lakin performansı artırmak için yazılmış bir katmandır. Asıl işlevi SQL SERVER tarafında bulk operation sistemini kullanmaktır. Bulk operation sistemi verileri önce geçici tablolara kayıt ettikleri için diğer yöntemlere göre çok daha fazla hızlıdırlar.

Repository.Pattern bu bölümde tekrarlı kod yazmanın önüne geçerek kodun yönetilebilirliğini kolaylaştırmaktadır. Ayrıca bu sınıf oluşturduğumuz entitylerde yapılacak CRUD (ekleme, silme, güncelleme) işlemlerin tek noktadan yönetilmesini sağlayacaktır. Böylelikle tekrarlı kod yazmanın önüne geçilmiş olmaktadır.

Repository.Pattern.Ef6 Repository.Pattern classı ile CRUD işlemleri tek bir noktada topladık ve veri tabanına kaydedilmek üzere kuyruğa aldık Repository.Pattern.Ef6 UnitofWork ile de bu kuyruğu tek noktadan veri tabanına kaydedeceğiz. Sıralı işlemlerde kayıt sırasında hata oluşması durumunda kuyrukta bulunan tüm işlemleri sırasıyla otomatik geri alınmasını sağlayacak ve veri tabanımızda veri bütünlüğünü korumayı garanti etmiş olacağız.

Service.Pattern repository katmanı ile oluşturduğumuz tüm hizmetleri bir servis mantığı içerisinde kullanmamızı sağlayan bölümdür.

BusinessLogicBllLayer iş katmanında tüm ana işlemler bu bölümde gerçekleştirilir. Veriyi internet üzerinde bulma, çekme, kaydetme gibi ana işlemler bu bölümde gerçekleştirilmiştir.

DataModelEntityLayer Veri tabanı şemamızın bir örneğinin bulunduğu bölümdür. Veri tabanında bulunan her tablo bir class a her tabloda bulunan bir kolon, bir class propertysine karşılık gelir (Şekil 4.3).

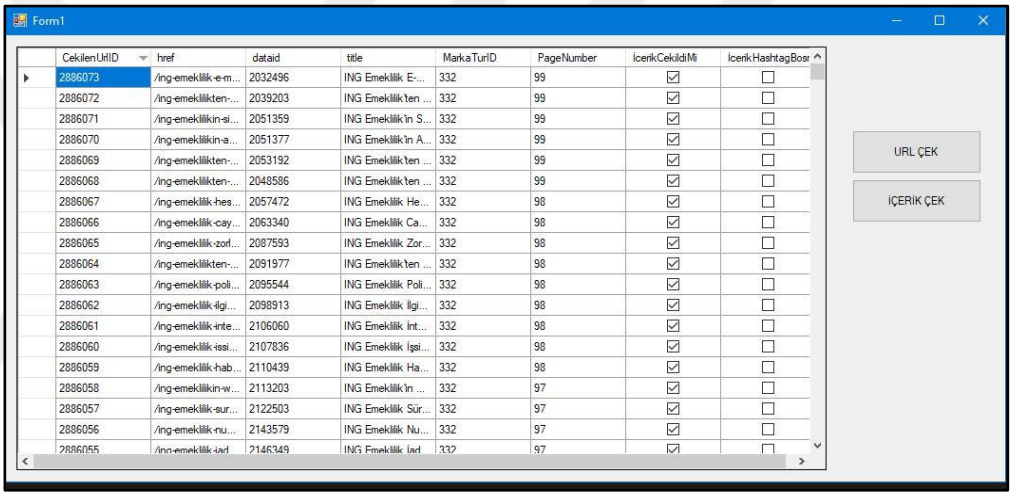
```
    [Table("Test")]
    public partial class Test // Test tablosu
    {
        public long ID { get; set; } //Test tablosunun ID kolonu
        public string Adi { get; set; } //Test tablosunun Adi
kolonu
    }
```

Şekil 4.3 Örnek entity model yapısı

Böyle bir yazılım mimarisi ve ihtiyaç duyulan tüm dilleri destekleyen bir platform kullanmak daha hızlı sonuç alınmasını kolaylaştırdı. Normal şartlarda R dili için R-Studio gibi bir platform, Pyhton için Paycharm gibi bir platform, C# içinse Visual Studio platformunu kullanmamız gerekiyordu. Visual Studio programının güncel sürümlerinde R Dili ve Python dillerine olan desteği ve bu platform üzerinde birden fazla proje geliştirmeye izin vermesi kolaylık sağladı. Her dilin diğer dile oranla başarılı olduğu bölümler bulunmaktadır. Bu nedenle farklı dilleri kullanarak tez çalışmasını tamamladık. Örneğin veri bilimi işlemlerinde

Python ve R dilinin açık ara önde olduğunu belirtmek gerekir. Ama veri tabanı işlemlerini diğer dillere göre daha kolay şekilde C# dili ile gerçekleştirdik. Bu tez kapsamında yapılan tüm çalışmalarda herhangi bir hazır araç (Weka vb.) kullanılmadı. Tüm işlemler hangi dil ile etkili ve başarılı şekilde yapılabiliyorsa o dil ile ilgili kodlar yazıldı.

KdsTezWinForm veri çekme işleminin yapıldığı arayüz bu ekrandadır. Şekil 4.4'de görülen ekranda veri çekilecek URL (linkler) tıklanarak çekiliyor ardından içerik çek butonu ile istenildiği kadar farklı sektörlere ve markalara ait veriler internetten bulunur ve veritabanına kaydedilir.



Şekil 4.4. Veri bulma ve çekme arayüzü

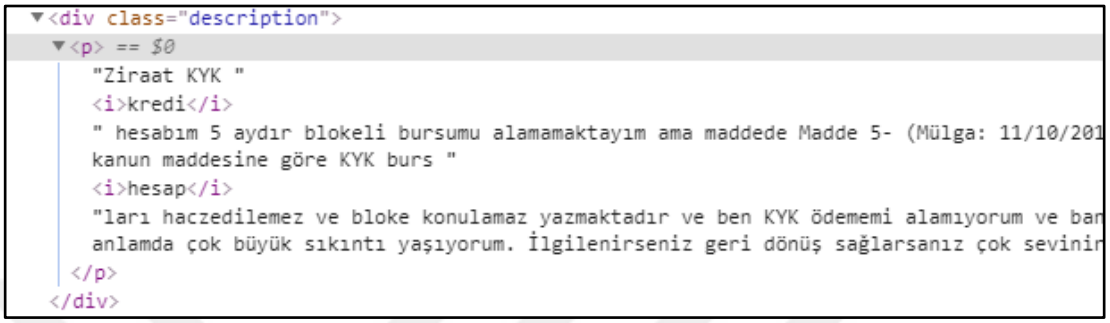
KdsTezNltk doğal dil işleme adımlarının uygulandığı analiz sonuçlarının tekrar veri tabanına kayıt edildiği bölümdür. Bu ekran console uygulaması olarak çalışmaktadır.

4.2. Metin Ön İşleme

Veriler yazılım teknolojileri kullanılarak veri tabanına kayıt edilirken ya da edildikten sonra metin ön işlem adımlarının yapılması gerekmektedir. Daha az efor sarf edilmesi ve gereksiz verilerden kaynaklı veri boyutu sorunu yaşamamak için veri çekme aşamasında gereksiz bölümlerin veri içerisinden atılması gerekir.

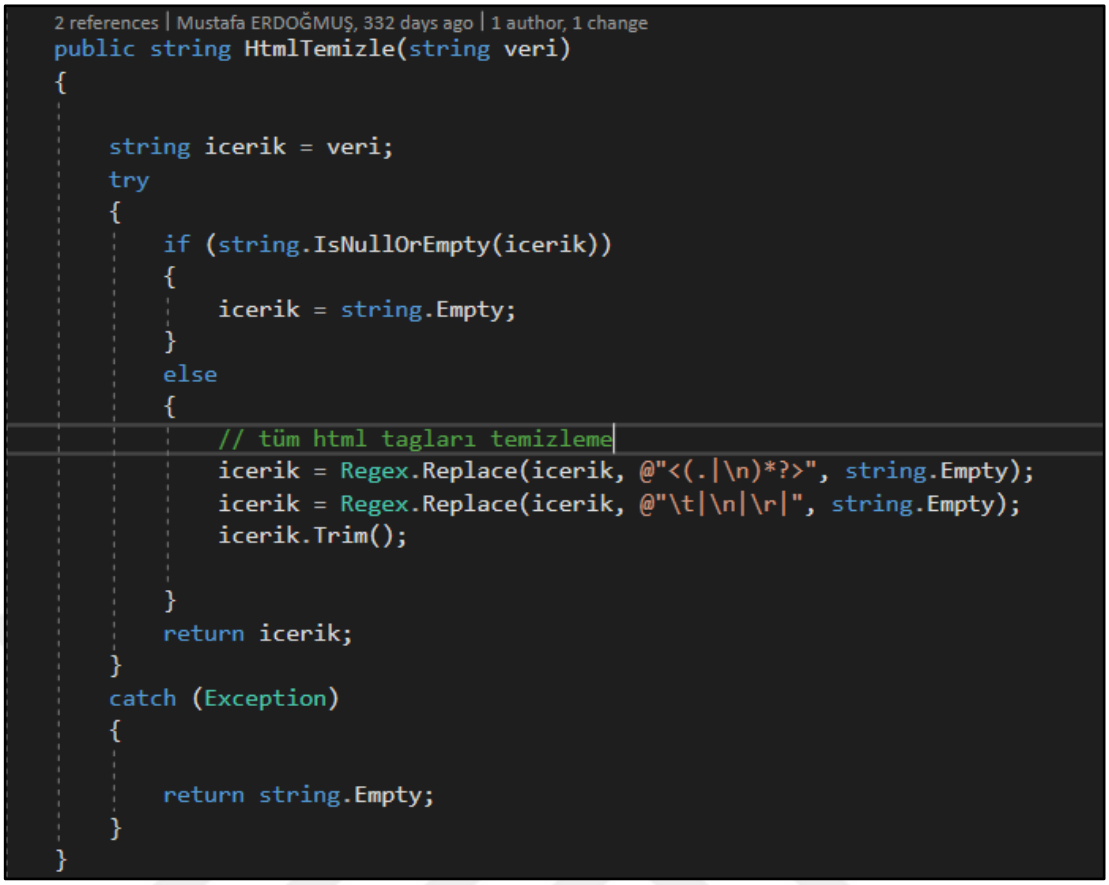
4.2.1. Dönüştürme

Hangi web sitelerinden veri çekilecekse öncelikle o web sitesinin yapısı incelenmelidir. Sonuç olarak geliştirilen bot yazılım gerçek bir kullanıcı gibi davranıp siteden istenilen sayfayı html olarak isteyecektir.



Şekil 4.5. Web kazıma programı tarafından yapılan bir istek sonucunda gelen html kodun bir bölümü

Şekil 4.5'de görülen html kodu için bizin ilgilendiren bölüm … etiketleri arasında kalan metinlerin olduğu kısımdır. Metin haricinde noktalama işaretleri hariç diğer tüm bölümler bizim için gereksizdir. Bu nedenle gereksiz kod parçalarının veri tabanına kayıt edilmeden önce temizlenmesi gerekir. Bu işlem yine bot yazılım içinde bulunan Şekil 4.6'de görülen HtmlTemizle metodu tarafından yapılmaktadır. Bu kod parçası geriye kalan tüm html etiketlerini temizleyip istediğimiz hale getirecektir. Bundan sonra ki aşama da ise veri tabanına istediğimiz "kredi hesabım 5 aydır blokeli bursumu alamamaktayım ama maddede Madde 5- (Mülga: 11/10/2011-khk-666/1 Md.) (1). Yukarıda belirtilen kanun maddesine göre KYK burs hesapları haczedilemez ve bloke konulamaz yazmaktadır ve ben KYK ödememi alamıyorum ve bana üniversite öğrencisi olduğum için maddi anlamda çok büyük sıkıntı yaşıyorum. İlgilenirseniz geri dönüş sağlarsanız çok sevinirim" yorumunu saf metin halinde kayıt edilmektedir. Bu şekilde 2 milyon civarında yorum temizlenerek veri tabanına kayıt edilmiştir.

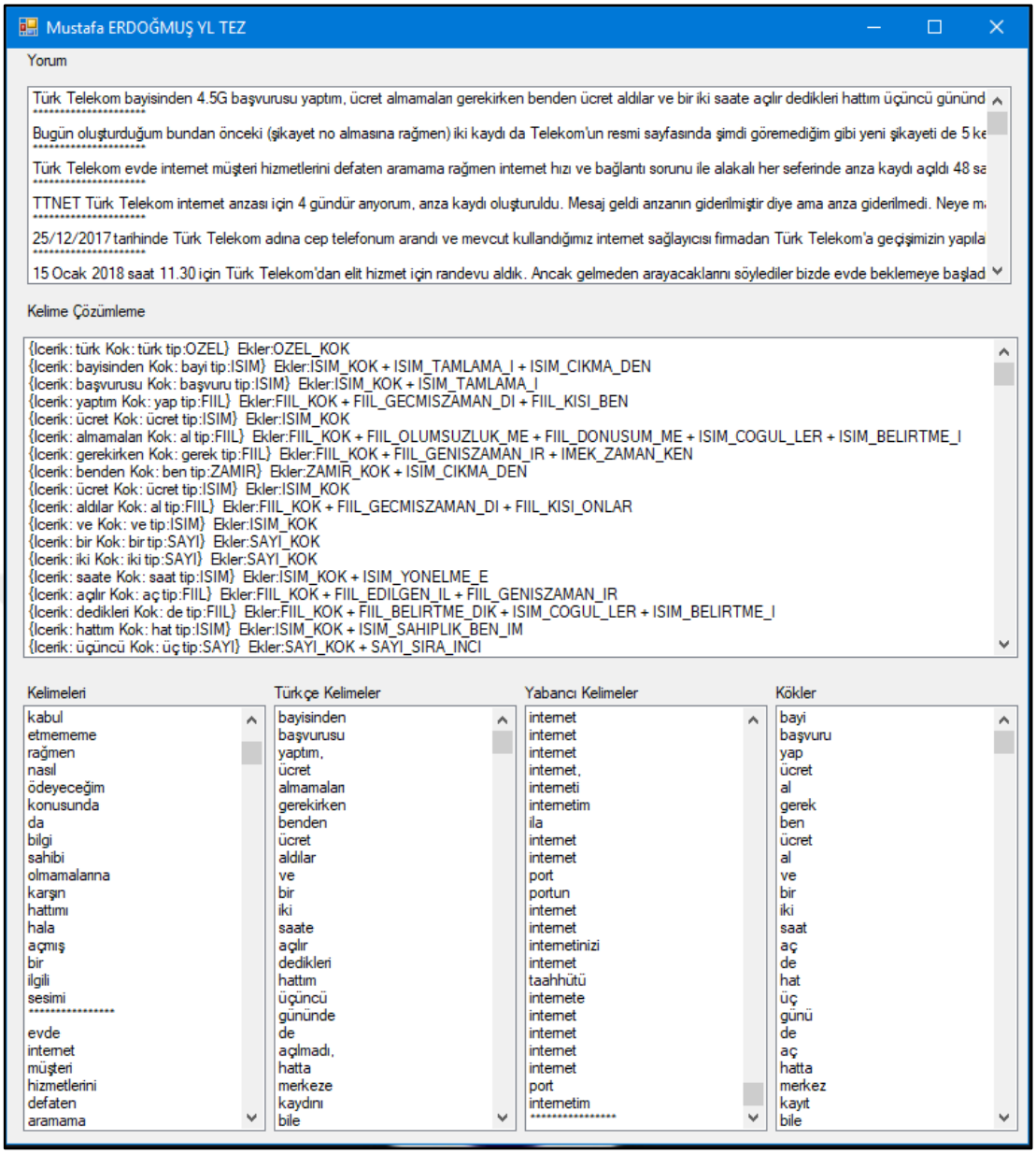


Şekil 4.6. Html tag temizleme metodu

4.2.2. Kelime Köklerini Bulma

Türkçe dili, sondan eklemeli bir dildir. Kelime köklerine yapım ve çekim ekleri gelerek kelimeler yeni anlamlar kazanmaktadır. Bu nedenle dilimizde duygu analizi çalışmalarını daha büyük emek ve çaba gerektirmektedir.

Geliştirilen modül ile metin içinde geçen kelimeler gruplanır ve noktalama işaretlerinden arındırılır. Ardından ilgili metin için Türkçe kelime mi yabancı kelime mi olduğu ayrıştırılır. Nihai olarak kelime kökü bulunarak Şekil 4.7'de görüleceğe üzere kayıt edilir.



Şekil 4.7. Doğal dil işleme adımları modülünün ekran görüntüsü

4.2.3. Durak Kelimeleri Ayrıştırma ve Cümleden Çıkartma

Cümle içerisinde tek başına anlamı olmayan ve bir bilgi içermeyen ancak yanına başka bir kelime geldiğinde anlam kazanabilen kelimeler işlem süresini uzatacağı için cümle içerisinden çıkartılması gerekmektedir. Bunun için Python dilinde bir metot yazılarak sorun çözülmüştür.

```
def stopWords(words,language):
    #dilde anlam içermeyen kelimeler cümleden ayıklanır
    wordsFiltered = []
    stopWords = set(stopwords.words(language))
    for w in words:
        if w not in stopWords:
            wordsFiltered.append(w)
            #anlamsız tüm kelimeler cümleden çıkartılıyor
              wordFiltered içine yazılıyor
        #print(wordsFiltered)
    return wordsFiltered
```

Şekil 4.8. Durak kelimeleri cümleden çıkartma kodu

Şekil 4.8'de görülen kod bloğunda stopWord metodu iki parametre almakta. Bunlardan birincisi words parametresi ile cümlenin kelimeleri dizi halinde metoda gönderiliyor, ikincisinde ise language parametresi ile hangi dilin durak kelimeleri bu cümleden çıkartılacak bilgisi gönderiliyor. Ve en son wordsFiltered değişkeni içerisinde cümlemiz içinde istemediğimiz ve, ile vb. kelimeler ayıklanmış ve cümleden çıkartılmış oluyor.

4.3. Doğal Dil İşlemleri ve Duygu Analizi

Gerçekleştirilecek çalışmanın iki ana temel unsurundan biri Metin Madenciliği diğeri Doğal Dil İşleme konusu altında bulunan duygu analizi işlemleridir. Duygu Analizi üç seviyede incelenebilir. Bunlar Bölüm 3.2 de detaylı şekilde anlatılmıştır.

Bu tez kapsamında cümle seviyesinde doğal dil işleme adımları uygulanarak yorum genelinde pozitif, negatif, nötr duygu içerdiği 2 milyon yorum için ele alınmış markaların her bir yorumu bu şekilde incelenmiş markalara ait yorumların toplam polarite değerleri bir listede toplanmıştır. Bu şekilde o markaya ait toplanan yorum verilerine göre toplam ne kadarlık pozitif, negatif, nötr polarite değerine sahip olduğu belirlenmiştir.

4.3.1. Doğal Dil İşlemleri ve Duygu Analizi Adımları

Duygu analizinde analiz edilecek cümle-yorum havuzunun genişliği sonuca ulaşmak için önemli noktalardan biridir. Biz bu çalışmayı gerçekleştirme aşamasın da 2 milyon yorumu internet ortamından derledik ve veri tabanına kayıt ettik. Bölüm 3.1.4' te bulunan metin madenciliği aşamalarının Metin Ön İşleme aşamasında gerçekleştirilen çalışmaların tamamı, duygu analizi içinde gerçekleştirildi. Dönüştürme işlemi gerçekleştirildi, durak kelimeler ayrıştırıldı ve cümleden çıkartıldı, kelime kökleri bulundu, kelimelerin cümle içerisinde hangi öğe (özne, yüklem, zarf) olarak yer aldığı tespit edildi.

Doğal Dil İşlemi (Natural Language Processing NLP) araçları kullanılarak her yorum için pozitif negatif ve nötr oldukları belirlenip veri tabanına kayıt edilmesi tez çalışmasının en zor adımlarından biridir. Çünkü bu adım bilgisayar mimarisinin de temelinde yatan girdi olarak giren verilerin 0 ve 1 lere çevrilerek işlenmesi ve sonuç olarak bilginin üretilip çıktı olarak kullanıcıya verilmesi mantığında çalışmaktadır. Ham yorumlar Doğal Dil İşleme adımları ile sayısallaştırılıp her yorum için 0 ila 1 arasında pozitif, negatif ve nötr değerler geliştirilen yazılım tarafından oluşturuldu. Örneğin A firmasının X ürünü için 685475 numaralı yorum için DDİ adımlarının uygulanması sonrasında Çizelge 4.1' de görüldüğü gibi geliştirilen yazılım tarafından 0 – 1 arasında değerler oluşturuldu. Elde edilen 2.089.326 yorum için bu işlem tekrarlandı ve ham verilerden 0-1 arasında ne kadar pozitiflik ne kadar negatiflik ne kadar nötrlük içerdiği bilgisi elde edildi. Pozitif, negatif, nötr değerlerin toplamı yapılan tüm yorumlar için toplam değerleri 1 olmak zorundadır. Örneğin Çizelge 4.1. de 685475 id li yorumun değerleri (0,596 + 0+ 0404) şeklindedir. Bu değerlerin toplamı 1 dir.

Çizelge 4.1. Ddi işlemleri sonrasında elde edilen değerler

Yorum Id	Pozitif	Negatif	Nötr
685475	0,596	0	0,404
685476	0,125	0,158	0,717
685477	0,139	0,251	0,61
685478	0,025	0,358	0,617

4.3.1.1. Python NLTK Vader Kütüphanesi Kullanılarak Gerçekleştirilen Çalışma

Python NLTK Vader Kütüphanesi kullanılarak gerçekleştirilen çalışmanın özeti ve sonuçları bu bölümde verilmiştir.

```
def SentimentAnalysis(sentences):
    for sentence in sentences:
        data = [sentence][0][2]
        tokenized_sentence = nltk.word_tokenize(data)
        worldFiltered = ddiEnglish.stopWords(tokenized_sentence,'turkish')
        sid = SentimentIntensityAnalyzer()
        pos_word_list=[]
        neu_word_list=[]
        neg_word_list=[]
        for word in tokenized_sentence:
            #word = word.replace("'","")
            if (sid.polarity_scores(word)['compound']) >= 0.1:
                pos_word_list.append(word)
            elif (sid.polarity_scores(word)['compound']) <= -0.1:
                neg_word_list.append(word)
            else:
                neu_word_list.append(word)
        score = sid.polarity_scores(data)
        for k in sorted(score):
             print('{0}: {1}, '.format(k, score[k]), end='')
        sql = "INSERT INTO [dbo].[SentimentAnalysis]([CekilenIcerikID],[Pos],[Neg],[Neu],"
        "[Compound],[Pos_word_list],[Neu_word_list],[Neg_word_list]) VALUES (?, ?, ?, ?, ?, ?, ?, ?)"
        compound = score['compound']
        neg = score['neg']
        neu = score['neu']
        pos = score['pos']
        pos_word_list_db =','.join(pos_word_list)
        neu_word_list_db =','.join(neu_word_list)
        neg_word_list_db =','.join(neg_word_list)
        val = ([sentence][0][0],pos,neg,neu,compound,pos_word_list_db,neu_word_list_db,neg_word_list_db)


        #sorgu =
        db.dbIslemleri.dbSetDataInsertConnetcion(sql,val)
        print('\nScores:', score)
```

Şekil 4.9. Duygu analizi kodları ve kayıt edilmesi

Şekil 4.9'de görülen kod parçasında SentimenAnalysis isimli metoda gelen dönüştürme işlemleri yapılmış her yorum için pozitif, negatif, nötr olup olmadığı Python NLTK kütüphanesi yardımıyla tespit edildi. Burada kullanılan yöntem cümleyi olumsuz yapan ekler ve kelimeler tespit edilip her bir yorumun 0 ile 1 arasında ne kadar pozitif ne kadar negatiflik içerdiği bulundu. Ve elde edilen her yorum için sonuçlar veri tabanına kayıt edildi. Şekil 4.10'da kayıt edilmiş verilerin örnek bir parçası verilmiştir. Örneğin 2174712 numaralı yorum için pozitif değer 0,107, negatif değer 0,056, nötr değer 0,837 olarak bulunmuştur.

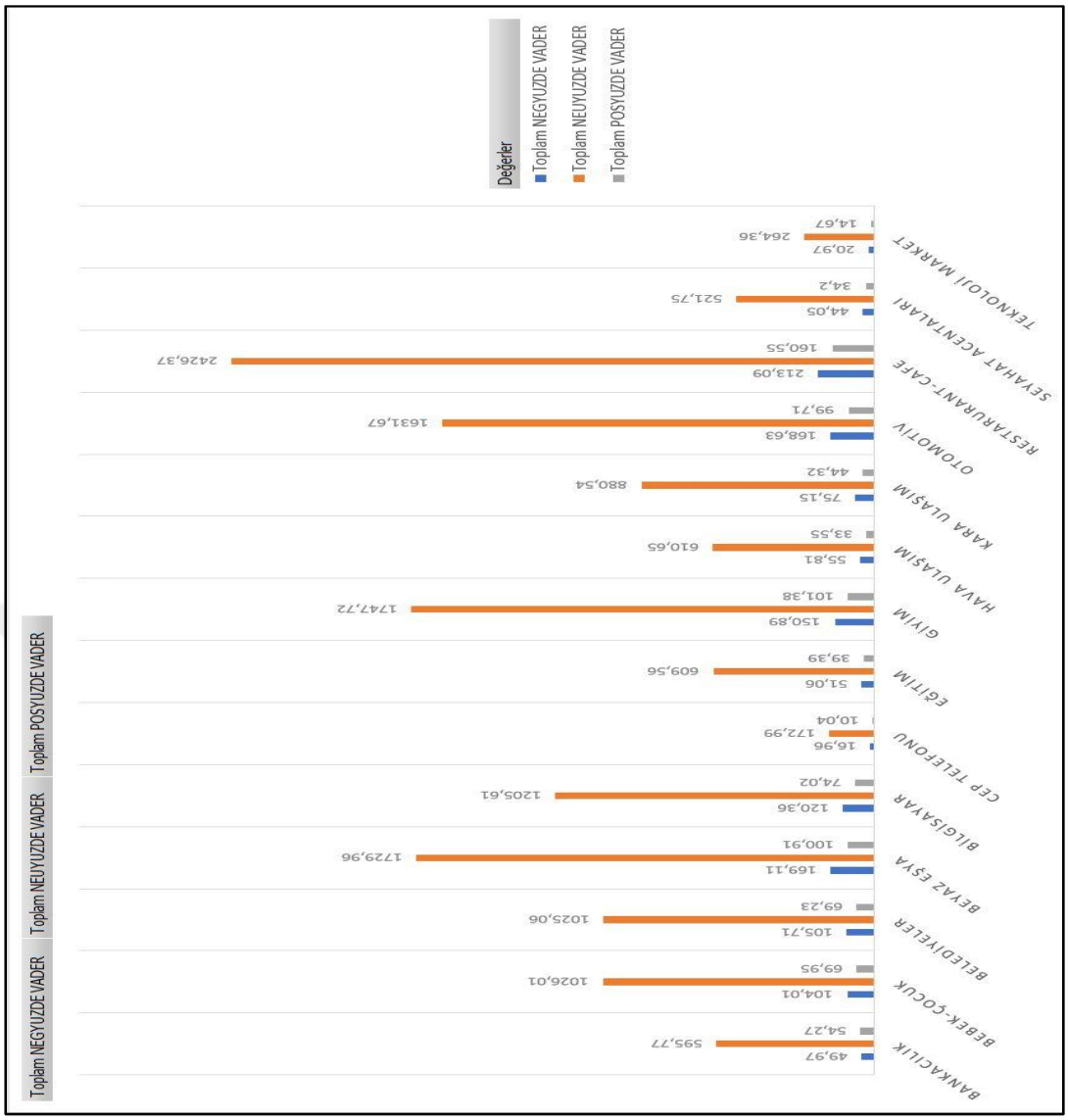
	AnalysisId	CekilenIcerikID	Pos	Neg	Neu	Compound
	1	2174718	0,396	0	0,604	0,9469
	2	2174717	0,746	0	0,254	0,8316
	3	2174716	0	0,529	0,471	-0,5216
	4	2174715	0	0	1	0
	5	2174714	0	0,286	0,714	-0,3412
	6	2174713	0	0,114	0,886	-0,4023
	7	2174712	0,107	0,056	0,837	0,4767
	8	2174711	0,053	0	0,947	0,0772
	9	2174709	0,04	0,237	0,722	-0,93
	10	2174705	0,036	0,173	0,791	-0,8225
	11	2174704	0,224	0,121	0,654	0,0772
	12	2174703	0,024	0,21	0,766	-0,9404
	13	2174701	0,018	0,115	0,867	-0,875
	14	2174697	0,05	0,036	0,914	0,1901
	15	2174696	0,077	0,041	0,882	0,6652
	16	2174694	0,037	0,174	0,789	-0,7458
	17	2174692	0,034	0,11	0,855	-0,836
	18	2174690	0	0,148	0,852	-0,8625
	19	2174687	0	0,108	0,892	-0,8176
	20	2174686	0	0,087	0,913	-0,7032
	21	2174684	0,151	0,133	0,716	0,5671
	22	2174681	0,107	0,221	0,671	-0,9482

Şekil 4.10. Duygu analizi sonucunda her bir yorum için oluşan pozitif, negatif, nötr listesi

Çizelge 4.2. Python nltk vader duygu analizi değerleri

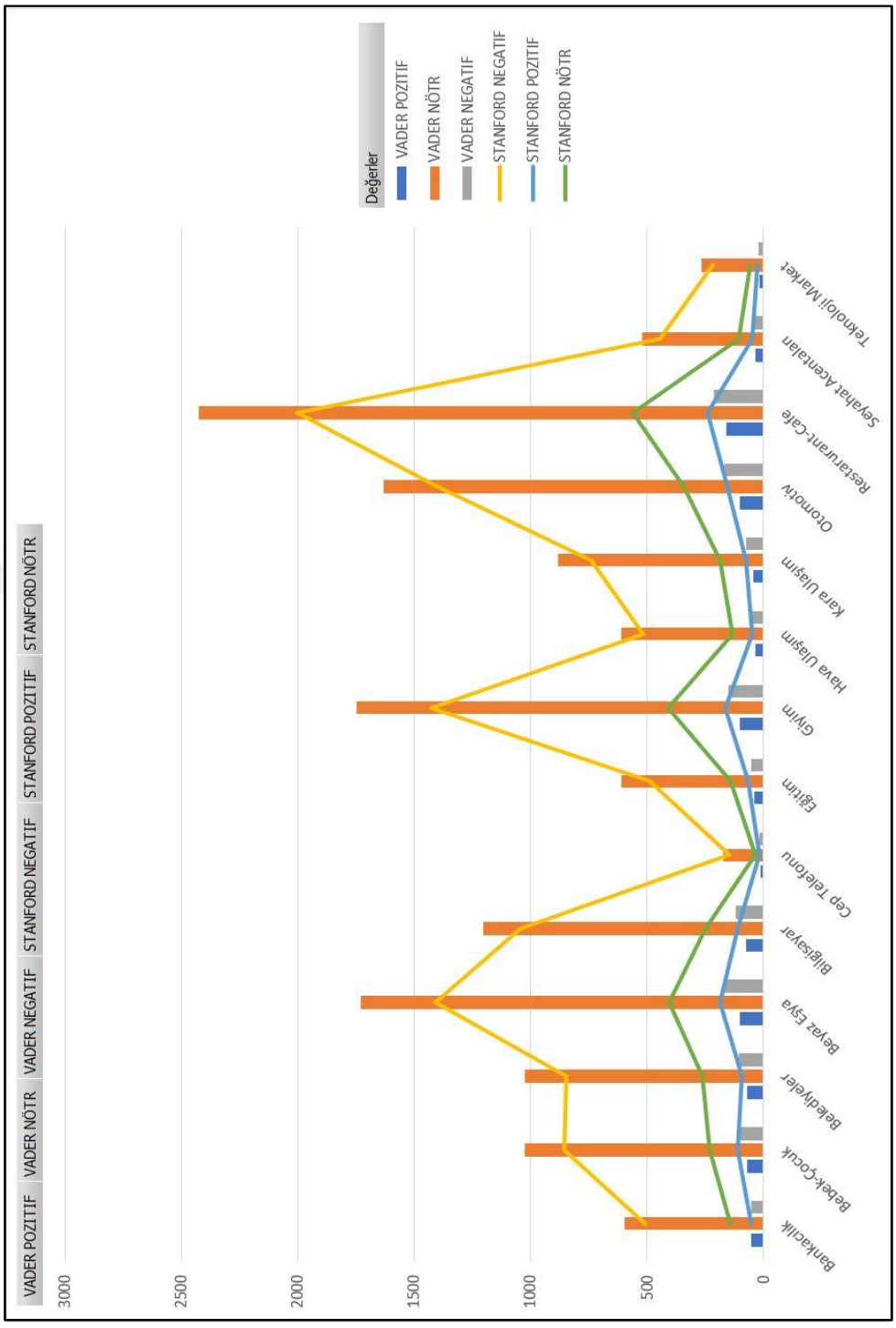
İfade & Metin	**Deneyler Sonucunda Oluşan Ortalama**	**Normal Dağılım Standart Sapması**	**Deneyler sırasında alınan 10 kişi değerlendirmesinin listesidir.**
(-:	1.6	0.8	[2, 2, 1, 3, 1, 1, 1, 3, 1, 1]
)-:	-2.1	0.9434	[-3, -2, -4, -1, -3, -2, -2, -2, -1, -1]
death	-2.9	1.04403	[-3, -4, -4, -3, -3, -1, -1, -4, -3, -3]
happing	1.1	0.83066	[1, 1, 1, 0, 2, 2, 0, 2, 0, 2]
happing	1.1	0.83066	[1, 1, 1, 0, 2, 2, 0, 2, 0, 2]
deny	-1.4	0.4899	[-1, -1, -1, -1, -2, -1, -2, -2, -1, -2]
denying	-1.4	0.4899	[-1, -1, -1, -2, -2, -2, -1, -2, -1, -1]
deadlock	-1.4	0.8	[-2, -2, -1, -3, -1, 0, -2, -1, -1, -1]

Çizelge 4.2'nin birinci sütununda yer alan ifade ve metinleri, ikinci sütun dördüncü sütunda yer alan 10 kişinin o ifade ve metinlere vermiş oldukları duygu değerlerinin ortalamasını, üçüncü sütun ise bunların normal dağılım standart sapmasını ifade etmektedir. Bu değerler kullanılarak Python Nltk Vader kütüphanesi ile birlikte tüm yorumlarda bulunan ifadeler ve kelimeler 7517 adet kelime ve ifadenin bulunduğu duygu sözlüğü ile karşılaştırılıp kelime skorlarının ortalamaları alınarak her yorum için değerler üretildi ve kayıt edildi. Alınan sonuçlar aşağıda gösterilmiştir.



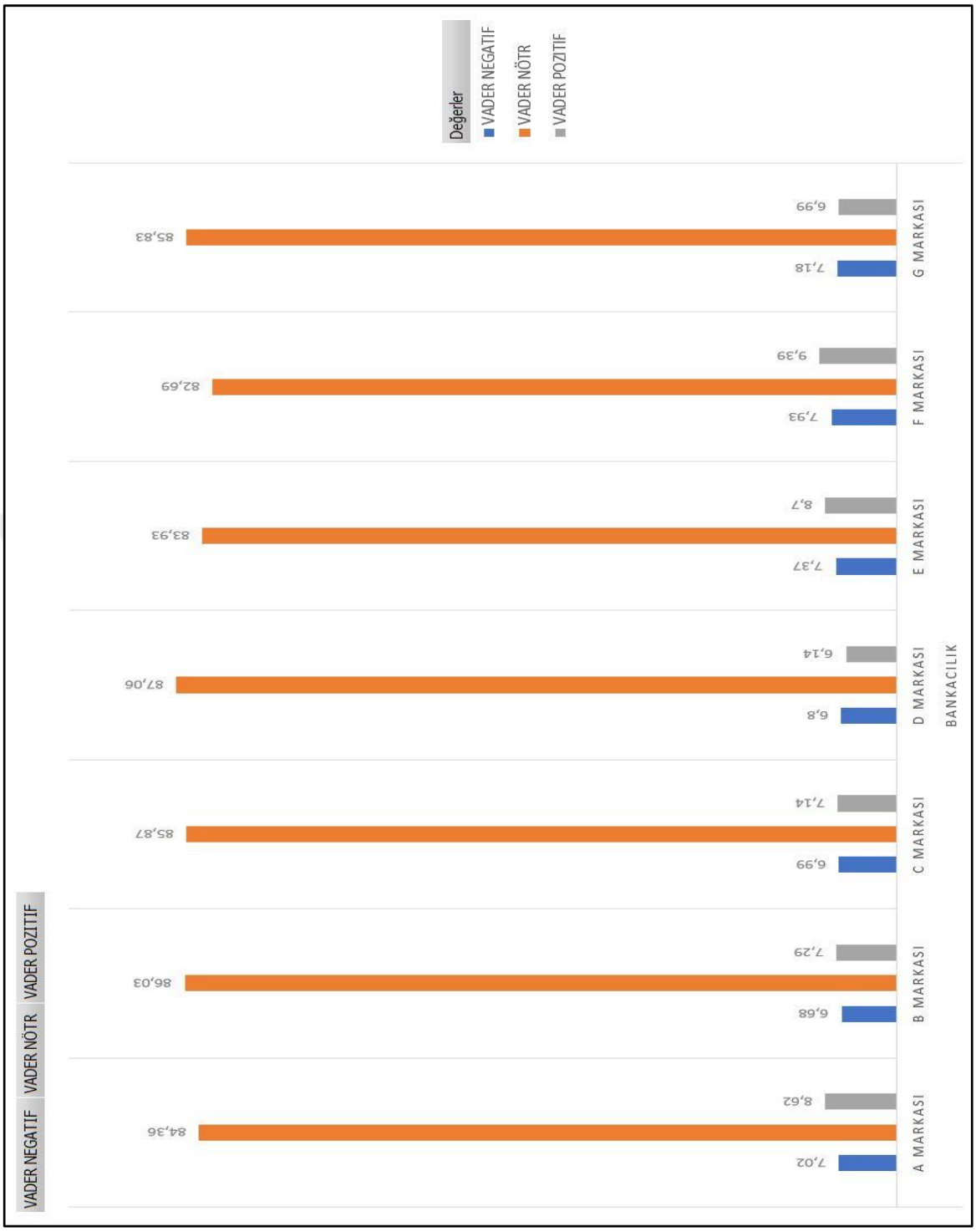
Şekil 4.11. Python nltk vader analizi sonuçları

Şekil 4.11'de görülen grafikte 13 sektöre ait yorumların Python Nltk Vader Kütüphanesi kullanılarak oluşturulan duygu analizi sonuçları verilmiştir. Grafikte en dikkat çekici bölümü nötr yorum yoğunluğunun toplam pozitif ve negatif yoğunluğa göre 2-3 kat daha fazla olmasıdır. Buradan anlaşılacağa üzere bu yöntem çok işe yarar bilgiler sunmamaktadır. Toplanan yorumlar arasında en çok pozitif ve negatif yorum alan sektör Restaruant & Cafe sektörüdür. Ardından giyim ve üçüncü olarak beyaz eşya sektörü gelmektedir. Grafik detaylandırıldığında sektörlere ait hangi markaların daha pozitif ya da negatif yorumlar aldığı Şekil 4.12, 4.13, 4.14 ve 4.15'te görülebilmektedir.



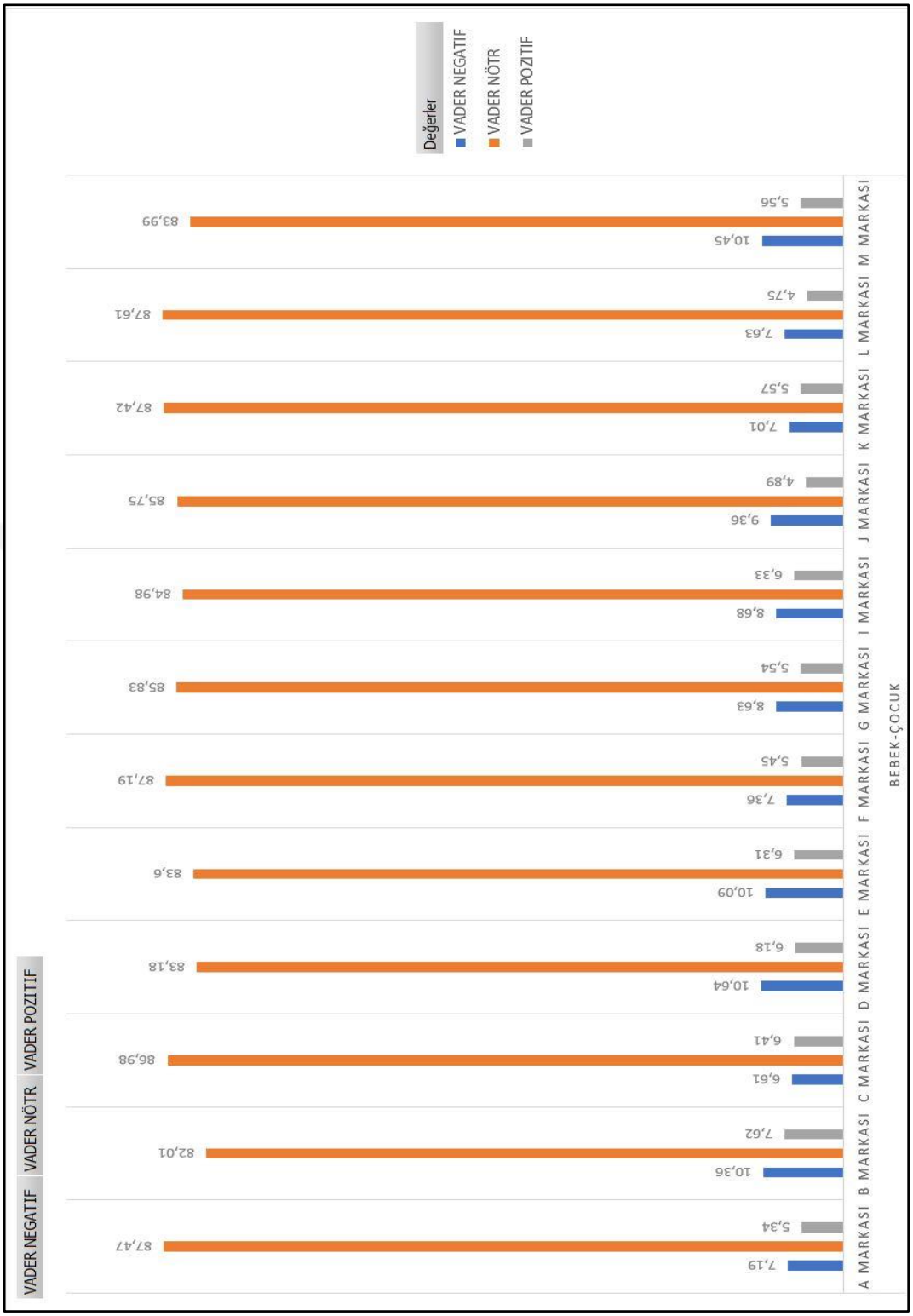
Şekil 4.12. Vader ve stanford nlp analizi sonuçlarının karşılaştırılması

Şekil 4.12'de görülen grafikte çizgi ekseni Stanford NLP analizini temsil ediyor. Sütun ekseni ise Vader NLP yi temsil ediyor. Vader sonuçlarında nötr en fazla, Stanford sonuçlarında ise negatif en fazla görülmektedir. Bu şekil bize Stanford NLP'nin daha başarılı sonuçlar verdiğini ispatlamaktadır.



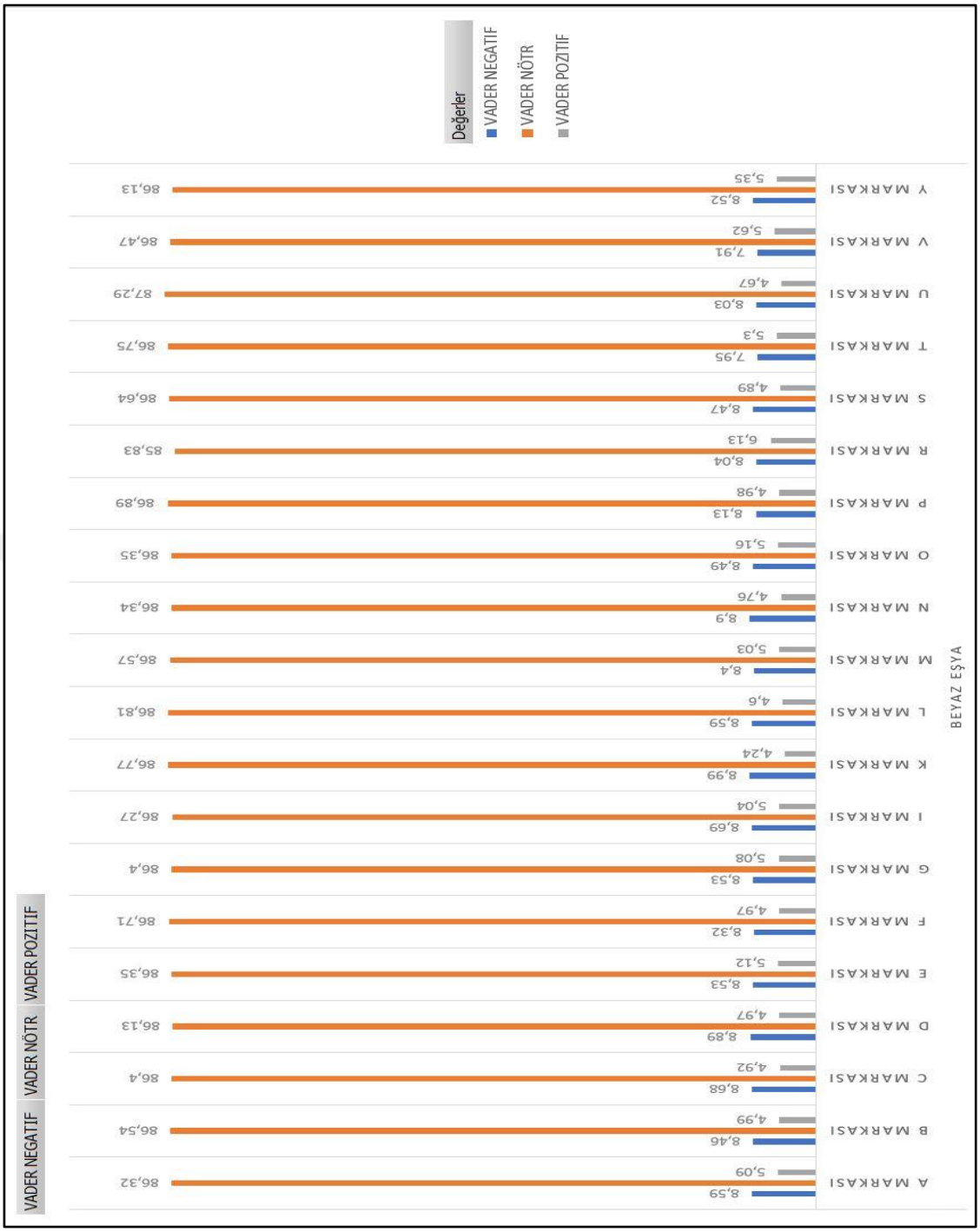
Şekil 4.13. Python nltk vader banka yorumlarının duygu analizi

Şekil 4.13'de grafiğe bakıldığında toplanan yorumlar arasında pozitif yorumları fazla olan banka F Bankası, negatif yorumları fazla olan banka yine F bankası olarak sonuç bulunmuştur.



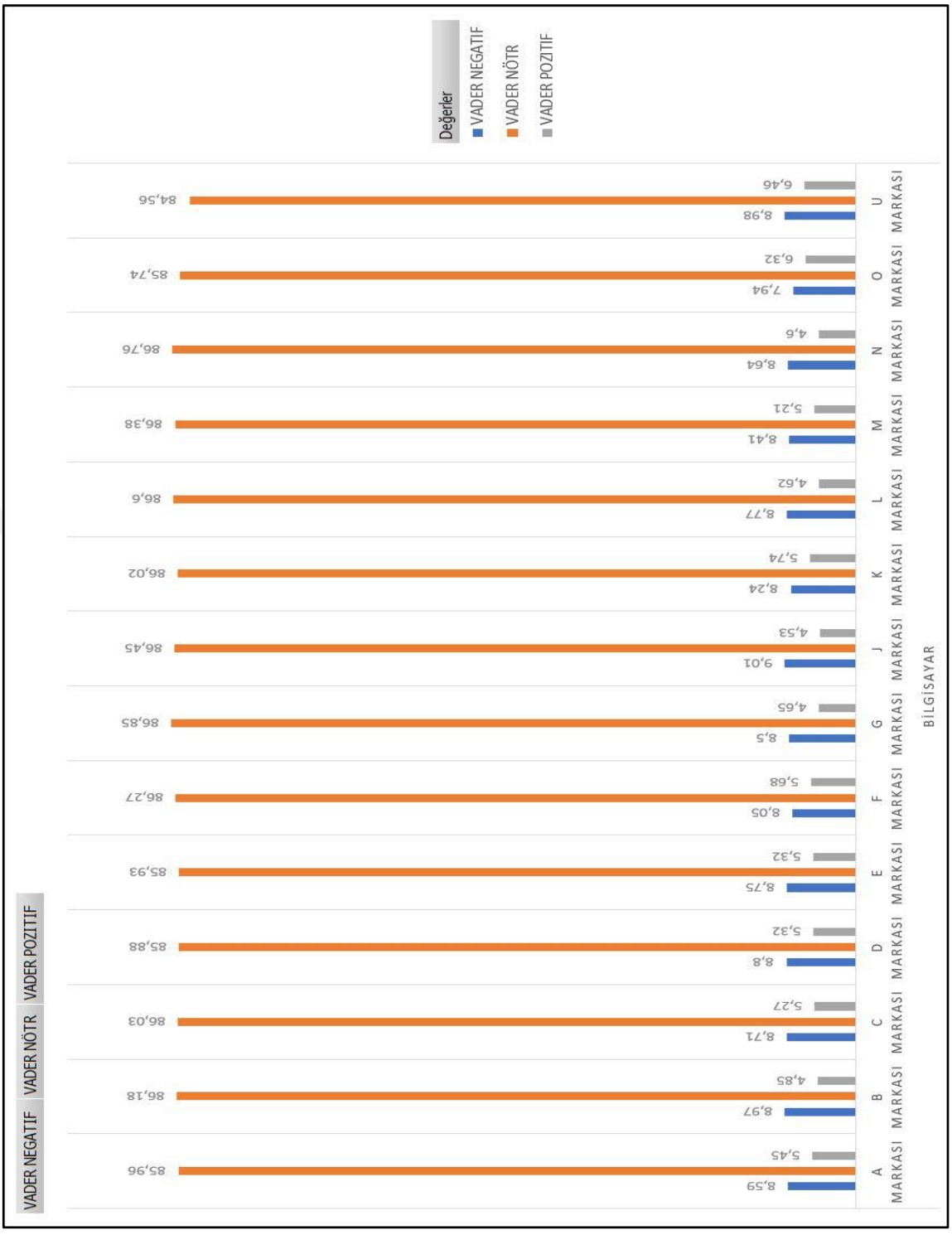
Şekil 4.14 Python nltk vader bebek çocuk marka yorumlarının duygu analizi

Şekil 4.14'de grafiğe bakıldığında toplanan yorumlar arasında pozitif yorumları fazla olan bebek çocuk markası B markası, negatif yorumları fazla olan bebek çocuk markası D markası olarak sonuç bulunmuştur.



Şekil 4.15. Python nltk vader beyaz eşya marka yorumlarının duygu analizi

Şekil 4.15'de grafiğe bakıldığında toplanan yorumlar arasında pozitif yorumları fazla olan beyaz eşya markası R markası, negatif yorumları fazla olan beyaz eşya markası K markası olarak sonuç bulunmuştur.



Şekil 4.16. Python nltk vader bilgisayar marka yorumlarının duygu analizi

Şekil 4.16'de grafiğe bakıldığında toplanan yorumlar arasında pozitif yorumları fazla olan bilgisayar markası U markası, negatif yorumları fazla olan bilgisayar yine J markası olarak sonuç bulunmuştur.

Python Nltk Vader kullanılarak hesaplanan duygu analizlerine göre tüm sektörler bazında tercih edilmesi tavsiye edilen markalar ya da tercih edilmemesi tavsiye edilen markalar Çizelge 4.3'te verilmiştir.

Çizelge 4.3. Python nltk vader kütüphanesinin sonuçlarına göre tavsiye edilen ve edilmeyen markalar

Sektörler	**Tercih Edilmesi Tavsiye Edilen Marka**	**Tercih Edilmesi Tavsiye Edilmeyen Marka**
Bankacılık	F MARKASI	F MARKASI
Bebek-Çocuk	B MARKASI	D MARKASI
Belediyeler	F MARKASI	O MARKASI
Beyaz Eşya	R MARKASI	K MARKASI
Bilgisayar	U MARKASI	J MARKASI
Cep Telefonu	A MARKASI	A MARKASI
Eğitim	F MARKASI	F MARKASI
Giyim	Y MARKASI	C MARKASI
Otomotiv	U MARKASI	S MARKASI
Restarurant-Cafe	D MARKASI	A4 MARKASI
Seyahat Acentaları	C MARKASI	A VE B MARKASI
Teknoloji Market	A MARKASI	C MARKASI
Hava Yolu Ulaşım	D MARKASI	D MARKASI
Kara Yolu Ulaşım	I MARKASI	L MARKASI

4.3.1.2. Stanford Üniversitesi Doğal Dil İşleme Yöntemi

Bir diğer yöntemde ise Stanford Üniversitesinin yapmış olduğu çalışma temel alınarak yapılmıştır. Her iki yöntem kıyas edilecek olursa Stanford NLP kütüphanesi daha başarılı sonuçlar vermiştir. Vader NLP ile yapılan çalışmada yorumların ortalama %80 e yakını nötr olarak tespit edilmiştir. Lakin aynı kategori ve aynı markaya ait yorumlara Standord Nlp ile analiz edildiğinde, yorumların %80 civarında negatif yorumlar olduğu tespit edilmiştir. Analiz sonuçları aşağıda verilmiştir.

```
def SentimentAnalysis(sentences):
    nlp = StanfordCoreNLP('http://localhost:9000')
    say=0
    for sentence in sentences:
        say=say+1
        print("id: " + str([sentence][0][0]))
        veryNegative=0
        negative=0
        neutral=0
        positive=0
        veryPositive=0
           #Dokümandan belirlenen özel karakterleri ve sayıları at
        comment = re.sub("[^a-zA-Z0-9_ğüşöçİĞÜŞÖÇ.,?!]", # Search for all non-letters
                         " ", # Replace all non-letters with spaces
                         str([sentence][0][2]))
        res = nlp.annotate([sentence][0][2],
                   properties={
                       'annotators': 'sentiment',
                       'outputFormat': 'json',
                       'timeout': 50000,
                   })
        for s in res["sentences"]:
             if s["sentimentValue"]=='0':
                 veryNegative=veryNegative+1
             elif s["sentimentValue"]=='1':
                 negative=negative+1
             elif s["sentimentValue"]=='2':
                 neutral=neutral+1
             elif s["sentimentValue"]=='3':
                 positive=positive+1
             elif s["sentimentValue"]=='4':
                 veryPositive=veryPositive+1
        val = ([sentence][0][0],veryNegative,negative,neutral,positive,veryPositive)
        stanfordNlp.DbKayit(val)
        print("say",say)
```

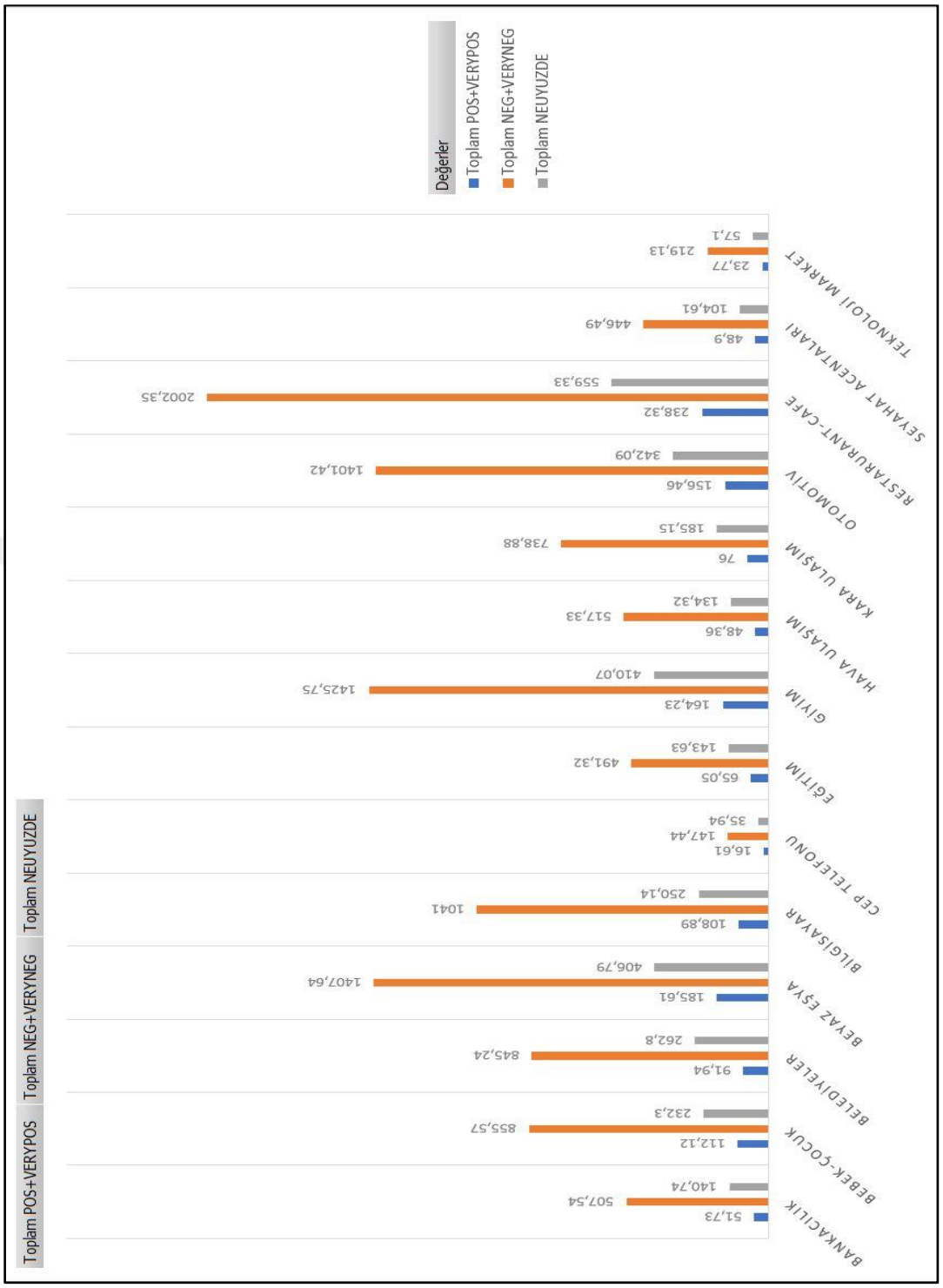
Şekil 4.17. Stanford nlp ile oluşturulan yöntemin kod parçasının bir bölümü

Şekil 4.17'de görülen kod parçasında Stanford Üniversitesi tarafından oluşturulan yöntemden yardım alındı. Bu yöntem bilgisayar üzerinde bir sunucu gibi davrandırıldı. Deneme ve sonuçlar bu sunucu üzerinden alındı. Python NLP yöntemi kullanılarak cümleler kelimelere bölümlendi her kelimenin duygu skoru duygu sözlüğüne göre hesaplandı. Cümledeki tüm kelimelerin duygu skorlarının ortalaması alındı, ardından kelimenin geçtiği yorum için ortalaması alınarak yorumun pozitif negatif skoru belirlendi. Stanford NLP kütüphanesinde ise yorumlar kelime bazlı duygu skoru hesaplama işlemi olmadan cümle bazlı duygu skorunun hesaplandığı yöntemdir.

Çizelge 4.4. Stanfor nlp ile oluşturulmuş duygu analizi örneği

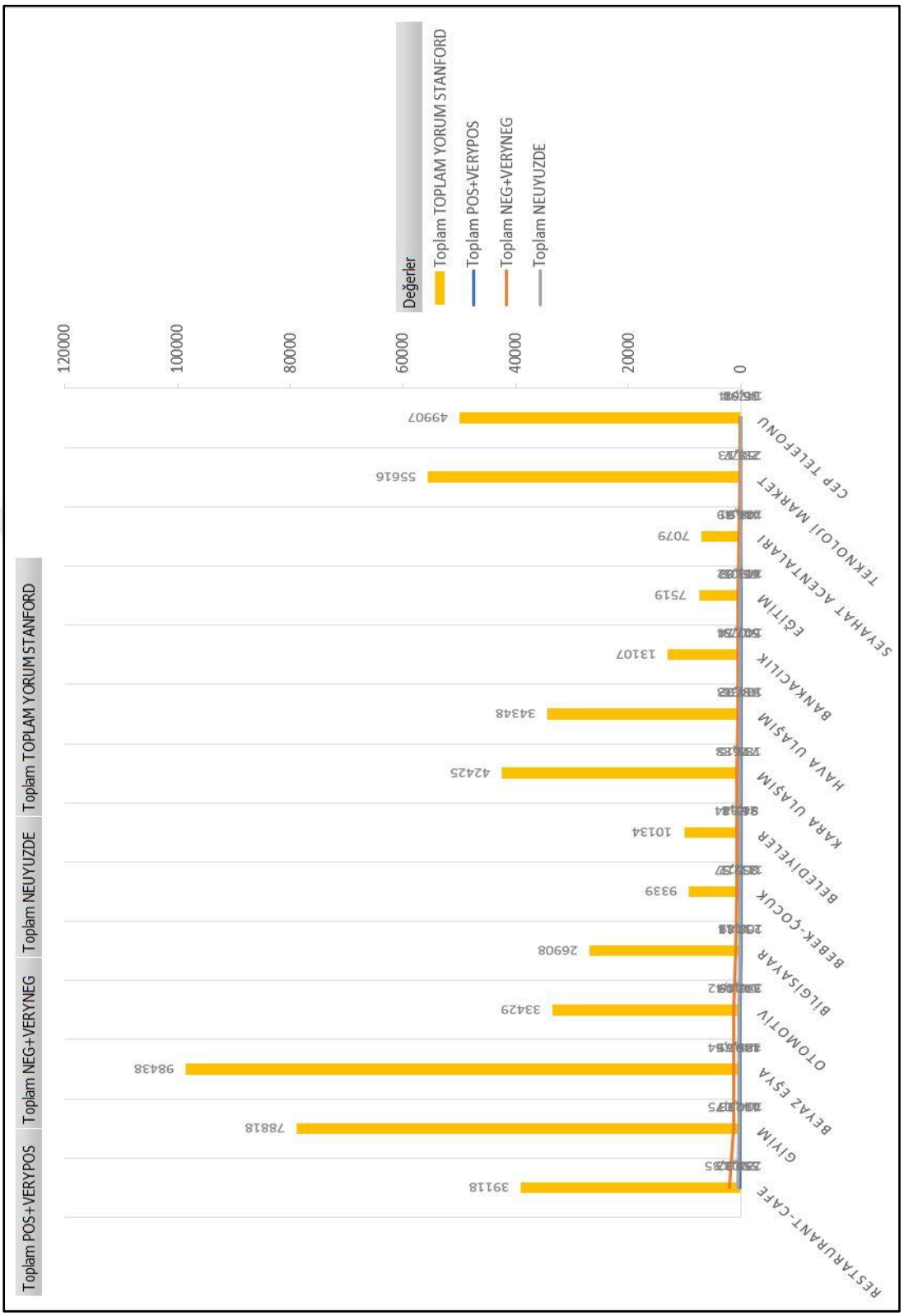
Yorumun Tamamı		
I sent a gift from the D&R for my wife's birthday. The product I received came after a rather painful process. Even though I wanted it to be in gift package, it was not. Also the price label was still on it. Futhermore the note that I wanted to be sticked on package wasnt there.		
Eşimin doğum günü için D&R hediye gönderdim. Aldığım ürünüm bayağı sancılı bir süreçten sonra geldi ama hediye paketi yapılmasını istediğim halde yapılmamış ve üstünde fiyat vardı ve not yazmıştım.		
No	**Cümle**	**Duygu Analizi**
1	I sent a gift from the dnr for my wife 's birthday.	Positive
2	The product I received came after a rather painful process.	Negative
3	Even though I wanted it to be in gift package, it was not.	Negative
4	Also the price label was still on it.	Neutral
5	Futhermore the note that I wanted to be sticked on package wasnt there.	Neutral

Çizelge 4.4 de görülen cümle önce varsa gereksiz karakterlerden temizleniyor Bölüm 4.2 de bulunan metin ön işleme adımları uygulanıyor ardından. Yorum cümlelere bölümleniyor. Her cümle için duygu skoru hesabı yapılıyor. Örnek verilen yorumda 1 pozitif, 2 negatif, 2 nötr cümle bulundu. Müşterinin aldığı hizmet ile ilgili yaşadığı sorunları dile getirdiği yorumdan negatif duygu skoru elde edilmiş. Bu durumda, bu yorumu negatif yorum olarak sınıflandırabileceğiz. Bu şekilde markalara ait tüm yorumlar için duygu analizleri yapıldı. Çıkan sonuçlar aşağıda değerlendirilmiştir.



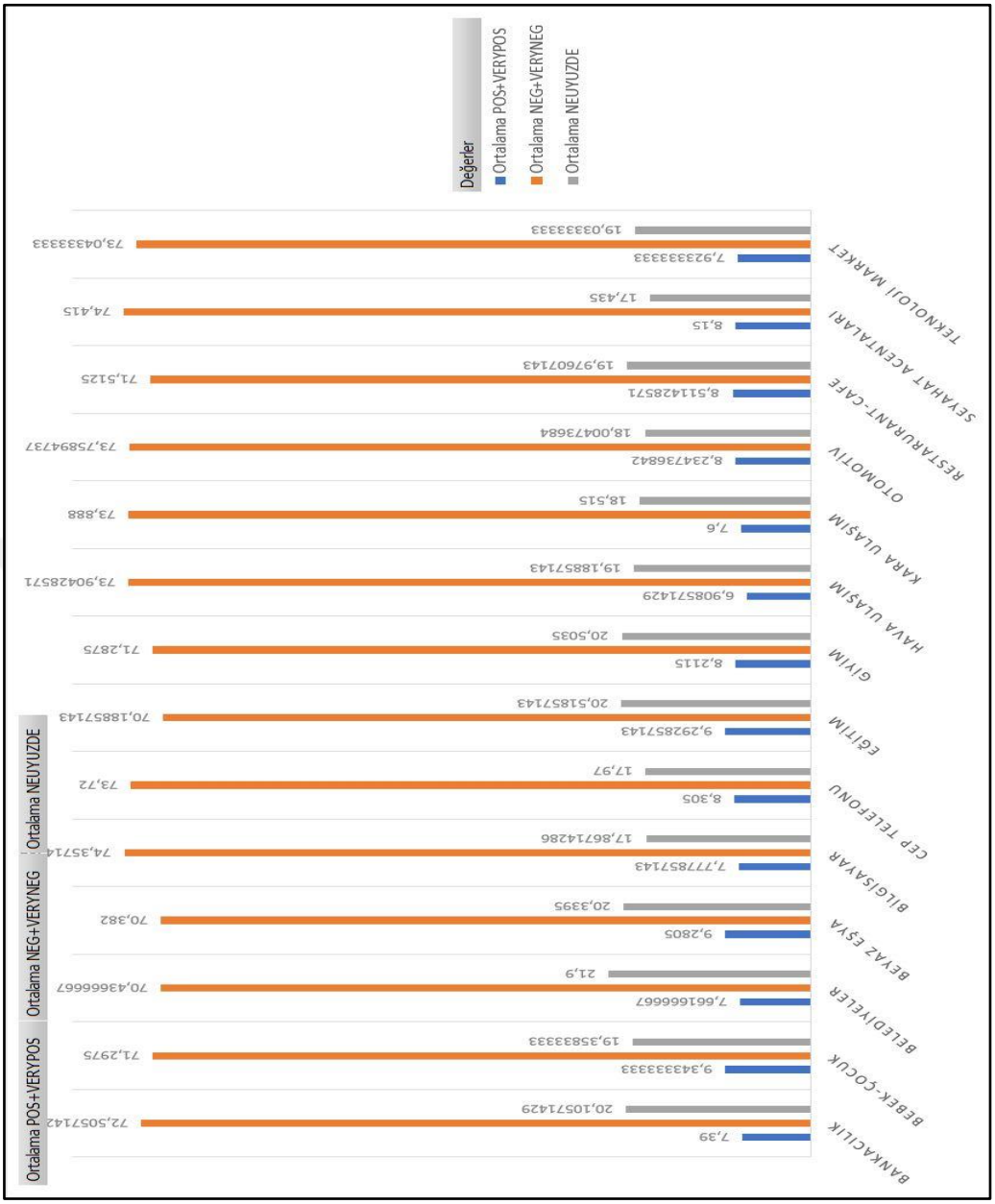
Şekil 4.18. Stanford nltk analizi sonuçlarının toplamı

Şekil 4.18’ de Stanford Nlp yöntemiyle elde edilen sonuçların sektörlere göre ortalamaları verilmiştir.



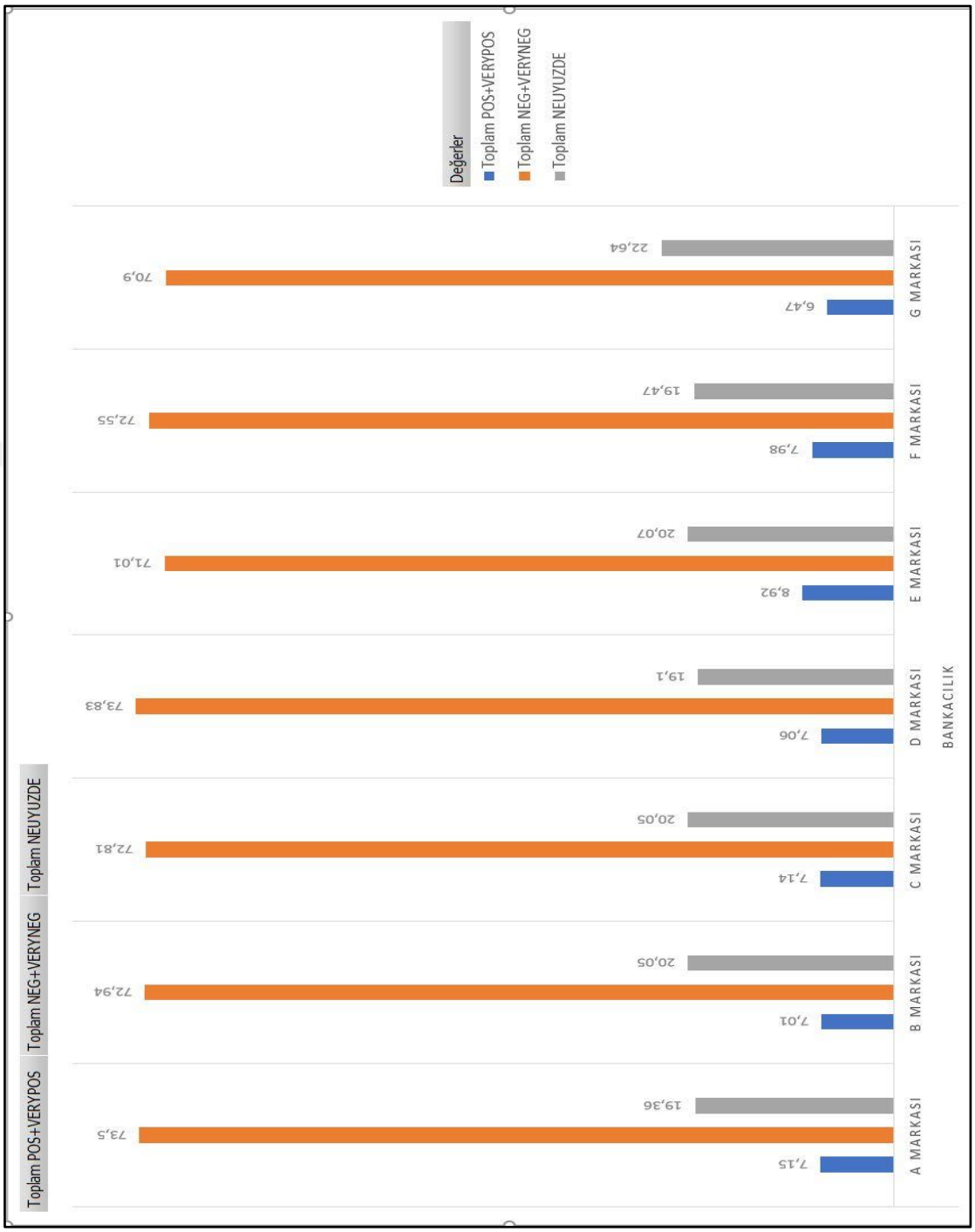
Şekil 4.19. Sektörlere göre toplam yorum sayıları

Şekil 4.19’ da sektörlere göre elde edilen yorumların sayıları verilmiştir.



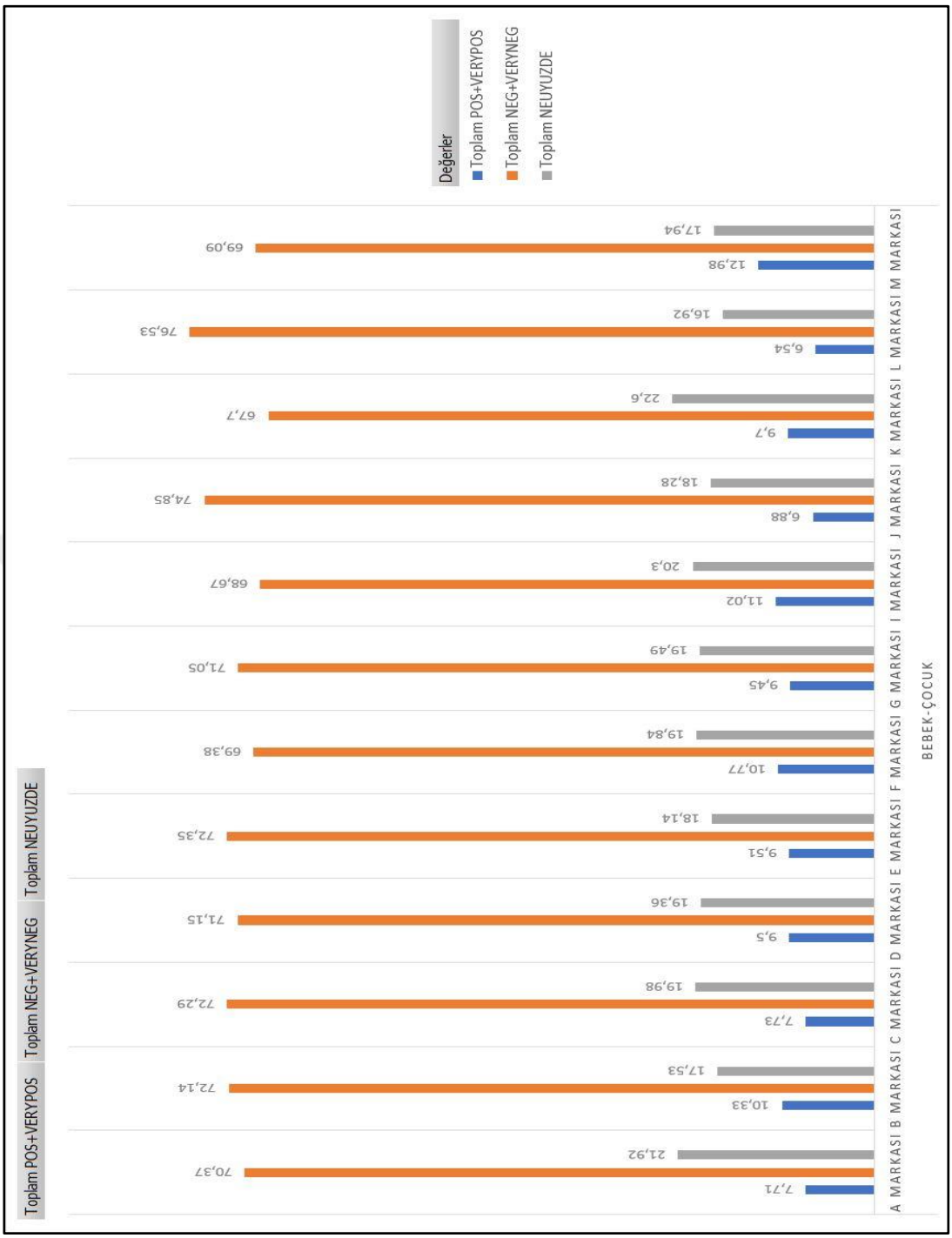
Şekil 4.20. Stanford nltk analizi sonuçlarının ortalamaları

Şekil 4.20'de görülen grafikte tüm kategorilere ait pozitif, negatif ve nötr sonuçlar verilmiştir. Grafikte en çok şikâyet almış sektör beyaz eşya sektörü (Şekil 4.19) ve negatif yorumları en fazla olan sektör seyahat acente sektörü olarak bulunmuştur (Şekil 4.20). Bebek-çocuk sektörü en fazla pozitif yorum alan sektör olmuştur (Şekil 4.20).



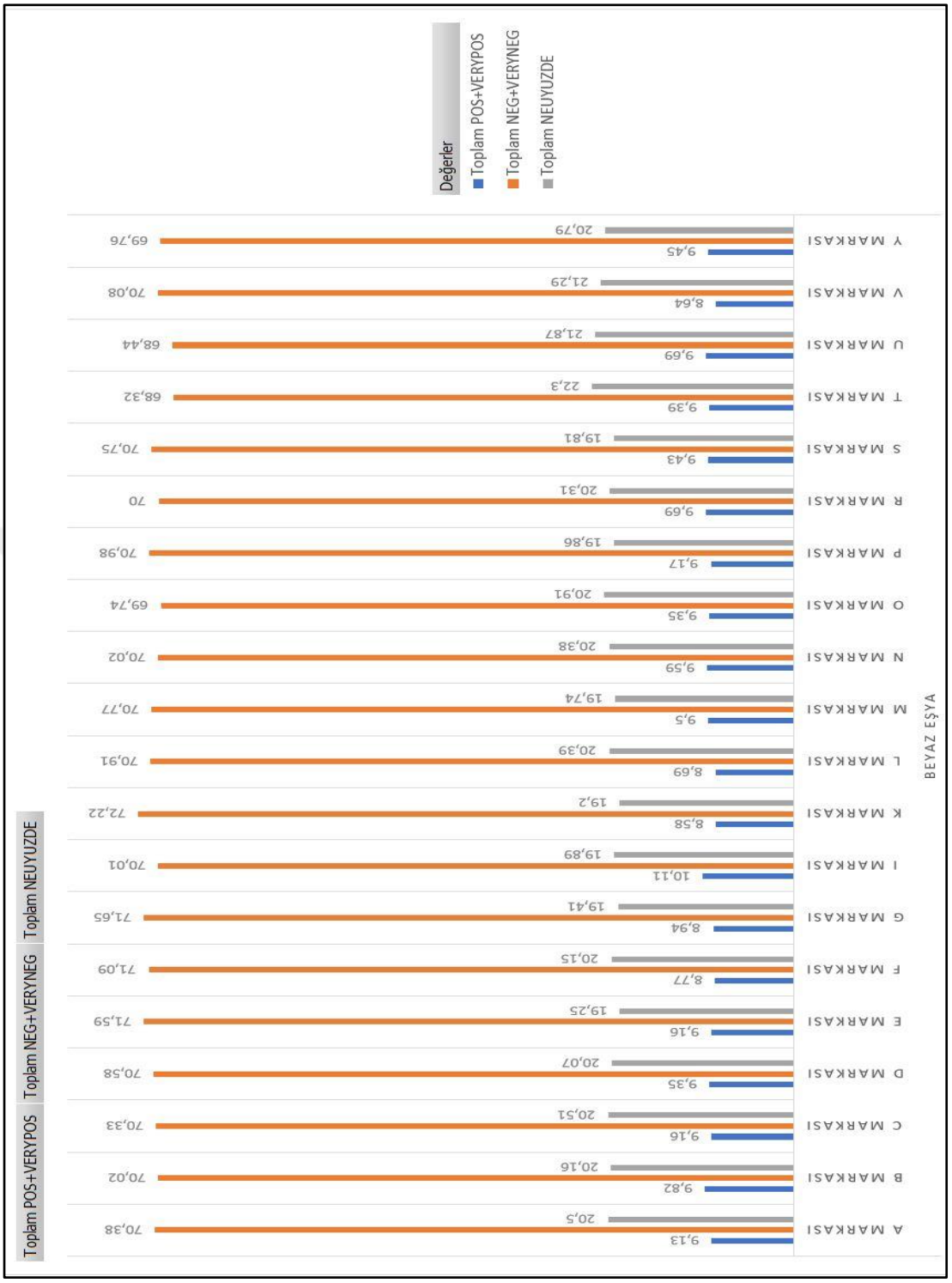
Şekil 4.21. Stanford nltk banka yorumlarının duygu analizi

Şekil 4.21'de D bankası toplam aldığı yorumların %73,83 Ü olumsuz yorum olduğu görülmektedir. E bankası %8,92 ile en yüksek olumlu yorum aldığı görülmektedir.



Şekil 4.22. Stanford nltk bebek çocuk yorumlarının duygu analizi

Şekil 4.22'de L Bebek Çocuk markası toplam aldığı yorumların %76,53 ü olumsuz yorum olduğu görülmektedir. K markası %67,7 ile en az olumsuz yorum alan marka olarak görülmektedir. %12,98 ile M markası en fazla olumlu yoruma sahip olduğu görülmektedir.



Şekil 4.23. Stanford nltk beyaz eşya yorumlarının duygu analizi

Şekil 4.23'de K beyaz eşya markası toplam aldığı yorumların %72,22 si olumsuz yorum olduğu görülmektedir. T markası %68,32 ile en az olumsuz yorum alan marka olarak görülmektedir. %10,11 ile I markası en fazla olumlu yoruma sahip olduğu görülmektedir.

Çizelge 4.5. Stanford nltk kütüphanesinin sonuçlarına göre tavsiye edilen edilmeyen markalar

Sektörler	Tercih Edilmesi Tavsiye Edilen Marka	Tercih Edilmesi Tavsiye Edilmeyen Marka
Bankacılık	E MARKASI	D MARKASI
Bebek-Çocuk	M MARKASI	L MARKASI
Belediyeler	K MARKASI	F MARKASI
Beyaz Eşya	I MARKASI	K MARKASI
Bilgisayar	U MARKASI	B MARKASI
Cep Telefonu	A MARKASI	B MARKASI
Eğitim	F MARKASI	D MARKASI
Giyim	L MARKASI	C MARKASI
Otomotiv	S MARKASI	O MARKASI
Restarurant-Cafe	V MARKASI	M MARKASI
Seyahat Acentaları	E MARKASI	B MARKASI
Teknoloji Market	A MARKASI	C MARKASI
Hava Yolu Ulaşım	C MARKASI	G MARKASI
Kara Yolu Ulaşım	A MARKASI	L MARKASI

Çizelge 4.5'de Stanford NLTK yöntemi ile yapılan analizin sonuçları verilmiştir. Doğruluk payının Vader yöntemine göre daha başarılı olduğu görülmüştür (Çizelge 4.6).

Çizelge 4.6. Her iki sistemin karşılaştırmalı tercih tablosu

Sektörler	Stanford Nlp Tercih Edilen	Stanford Nlp Tercih Edilmeyen	Vader Nltk Tercih Edilen	Vader Nltk Tercih Edilmeyen
Bankacılık	E MARKASI	D MARKASI	F MARKASI	F MARKASI
Bebek-Çocuk	M MARKASI	L MARKASI	B MARKASI	D MARKASI
Belediyeler	K MARKASI	F MARKASI	F MARKASI	O MARKASI
Beyaz Eşya	I MARKASI	K MARKASI	R MARKASI	K MARKASI
Bilgisayar	U MARKASI	B MARKASI	U MARKASI	J MARKASI
Cep Telefonu	A MARKASI	B MARKASI	A MARKASI	A MARKASI
Eğitim	F MARKASI	D MARKASI	F MARKASI	F MARKASI
Giyim	L MARKASI	C MARKASI	Y MARKASI	C MARKASI
Otomotiv	S MARKASI	O MARKASI	U MARKASI	S MARKASI
Restarurant-Cafe	V MARKASI	M MARKASI	D MARKASI	A4 MARKASI
Seyahat Acentaları	E MARKASI	B MARKASI	C MARKASI	A VE B MARKASI
Teknoloji Market	A MARKASI	C MARKASI	A MARKASI	C MARKASI
Hava Yolu Ulaşım	C MARKASI	G MARKASI	D MARKASI	D MARKASI
Kara Yolu Ulaşım	A MARKASI	L MARKASI	I MARKASI	L MARKASI

Grafik verilerinden anlıyoruz ki elimizde bulunan 2 milyon civarındaki yapılandırılmamış veriyi yapılandırarak, anlamlı anlaşılır bilgiler çıkarmayı başardık. Hedefimiz insanların kolay karar vermesine yardımcı olmak olan proje ile sektöründe en başarısız olan markaları, en başarılı olan markaları bu ve buna benzer çok bilgiyi artık yapılandırdığımız veriler içerisinden çıkartabiliriz.

4.4. Kelime Bulutları

Kelime bulutları kelimenin metin içinde geçme sıklığına göre analiz edildikten sonra, farklı renk şekil ve desenlerle oluşturulan görsellerdir. Bu şekilde veri görselleştirilerek öne çıkan kelimeler kullanıcılara daha net fikirler verilmesi sağlanabilir. Amaçları, herhangi bir metin okunmadan hangi konudan sıklıkla bahsettiği, hangi kelimelerin ve kelime öbeklerinin çoğunlukta olduğunu analiz etmemizi kolaylaştırır.

Şekil 4.24. Banka kelime bulutu uygulaması 1

Şekil 4.25. Banka kelime bulutu uygulaması 2

Şekil 4.26. Banka kelime bulutu uygulaması 3

Şekil 4.24, Şekil 4.25 ve Şekil 4.26'da görülen kelime bulut uygulamaları tüm markalar için gerçekleştirilmiştir. Banka sektöründen üç markaya ait sonuçlar eklenmiştir. Bankaların hepsinde "hesap işletim ücreti iadesinin talebi" ile ilgili çok fazla yorum bulunduğu açıkça görülebilmektedir. Kelime bulutları sayesinde tüm yorumları okumadan yorum kümesi ile ilgili bu şekilde fikirler elde edebiliriz.

5. ARAŞTIRMA BULGULARI VE TARTIŞMA

Bu çalışmanın amacı yapılandırılmamış verileri yapılandırılmış veri türüne çevirmek, ardından bu verileri geliştirilen doğal dil işleme, fikir madenciliği, grafik araçları ile 2.174.713 yorum & 20.958.140 adet cümleyi pozitif, negatif ve nötr olarak işaretleyip markaları kendi arasında tercih edilebilirliğini insan gücüne ihtiyaç duymadan ölçmektir. İnsanların ürün satın alırken karar verme süreçlerini basitleştirmek ve daha etkili hale getirmek bu tezin diğer amaçlarından biriydi. Bu amaçları gerçekleştirmek için daha önce yapılan duygu analizi, doğal dil işleme, fikir madenciliği uygulamaları incelenmiş çalışmada ki uygulama ile karşılaştırmalar yapılmıştır. Kullanılan iki adet duygu analizi yönteminin hangisinin daha başarılı olduğu da bu tez kapsamında gerçekleştirilen çalışmadır. Tezde kullanılan iki adet duygu analizi yöntemlerinin başarım oranları aşağıda açıklanmıştır.

Çizelge 5.1. Standord nlp ve vader nltk sistemlerinden elde edilen sonuçlar

Sektörler	**Stanford Nlp Tercih Edilen**	**Stanford Nlp Tercih Edilmeyen**	**Vader Nltk Tercih Edilen**	**Vader Nltk Tercih Edilmeyen**
Bankacılık	E MARKASI	D MARKASI	F MARKASI	F MARKASI
Bebek-Çocuk	M MARKASI	L MARKASI	B MARKASI	D MARKASI
Belediyeler	K MARKASI	F MARKASI	F MARKASI	O MARKASI
Beyaz Eşya	I MARKASI	K MARKASI	R MARKASI	K MARKASI
Bilgisayar	U MARKASI	B MARKASI	U MARKASI	J MARKASI
Cep Telefonu	A MARKASI	B MARKASI	A MARKASI	A MARKASI
Eğitim	F MARKASI	D MARKASI	F MARKASI	F MARKASI
Giyim	L MARKASI	C MARKASI	Y MARKASI	C MARKASI
Otomotiv	S MARKASI	O MARKASI	U MARKASI	S MARKASI
Restarurant-Cafe	V MARKASI	M MARKASI	D MARKASI	A4 MARKASI
Seyahat Acentaları	E MARKASI	B MARKASI	C MARKASI	A VE B MARKASI

Çizelge 5.1. (Devamı) Standord nlp ve vader nltk sistemlerinden elde edilen sonuçlar

Sektörler	**Stanford Nlp Tercih Edilen**	**Stanford Nlp Tercih Edilmeyen**	**Vader Nltk Tercih Edilen**	**Vader Nltk Tercih Edilmeyen**
Teknoloji Market	A MARKASI	C MARKASI	A MARKASI	C MARKASI
Hava Yolu Ulaşım	C MARKASI	G MARKASI	D MARKASI	D MARKASI
Kara Yolu Ulaşım	A MARKASI	L MARKASI	I MARKASI	L MARKASI

Çizelge 5.1 de her iki sistemden alınan sonuçlarında tercih edilen grubunda sadece 4 tanesinde iki sistemde o kategori için aynı markayı önermektedir. Yani bilgisayar, cep telefonu, eğitim, teknoloji market kategorilerinde iki sistemde aynı markayı önermektedir. Aynı çizelgede tercih edilmesi tavsiye edilmeyen markalar arasında beyaz eşya, giyim, teknoloji ve kara ulaşım kategorilerinde aynı markayı önermemektedir.

Çizelge 5.2. Duygu analizi yöntemlerinin örnek yorumlar ile kıyaslanması

YORUM_ID	PYTHON NLTK VADER			STANFORD NLP		
	Negatif	Nötr	Pozitif	Negatif	Nötr	Pozitif
2174371	10,90%	87,00%	2,10%	100,00%	0,00%	0,00%
2174374	17,80%	78,30%	3,90%	66,67%	0,00%	33,33%
2174376	6,70%	81,00%	12,30%	66,67%	33,33%	0,00%
2174378	15,00%	83,80%	1,20%	70,00%	20,00%	10,00%
2174379	9,70%	83,00%	7,20%	100,00%	0,00%	0,00%
2174382	5,90%	92,60%	1,50%	40,00%	50,00%	10,00%

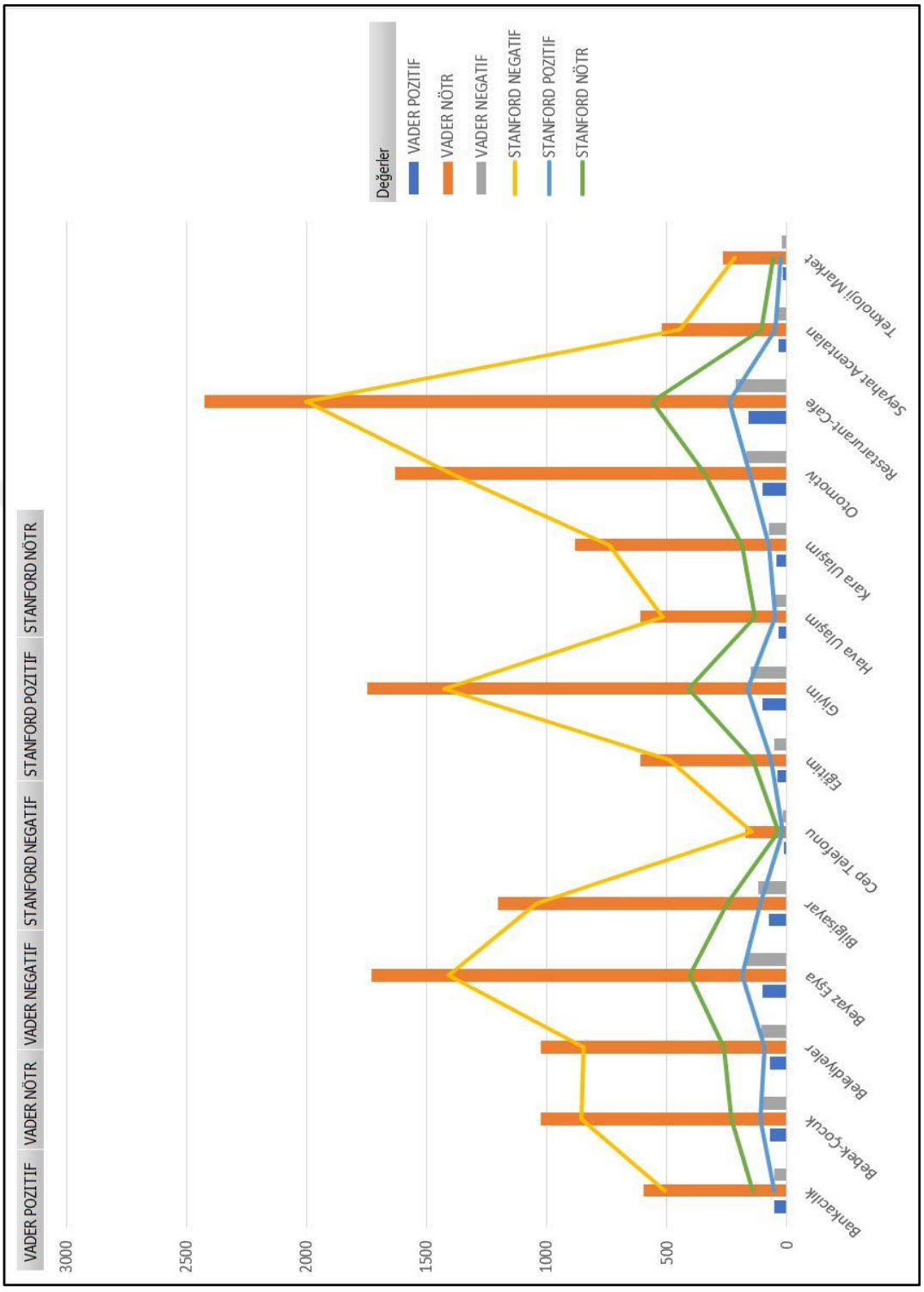
Çizelge 5.2 de 7 adet yoruma ait yöntemlerden alınan sonuçlar verilmiştir. Çizelgeye bakıldığında Python Nltk Vader yönteminden alınan sonuçların büyük çoğunluğu nötr olarak işaretlendiği görüldü. Cümlelere bakıldığında ise negatif

işaretlenmesi beklenen yorum Nltk Vader yöntemiyle nötr işaretlendiği görüldü. Nltk Vader yöntemiyle sağlıklı bilgi çıkarımı yapılamadığı tabloya göre açık şekilde görülmektedir. Önerilen Stanford Nlp yönteminin, 2174376 numaralı yorum için verdiği sonuçlar ve kullanıcıların cümleyi okuduklarında verdikleri yanıtlar Çizelge 5.3' te verilmiştir.

Çizelge 5.3. 2174376 numaralı yorumun Stanford nltk sonuçlarının incelenmesi ve gerçek kullanıcı yorumlarıyla karşılaştırılması

Yorumun Cümleleri	**Yöntemin Bulduğu**	**10 Adet Farklı Kullanıcının Söyledikleri**
1- Yaklaşık 7-8 ay önce bir HP Notebook aldım.	Nötr	Nötr
2- 6c04st model aldığıma bin pişmanım her türlü problem var ısınması performans pil bir dahaki seçimim kesinlikle HP olmayacak.	Negatif	Negatif
3- Yaptığınız ürün iyi değil.	Negatif	Negatif
4- Bari bunu telafi etmek için şu pili değiştirin 1 saat gitmiyor şarj.	Negatif	Negatif
5- Sıkıldım gerçekten bu durumdan ürünümün garantisi hala devam ediyor.	Negatif	Negatif
6- Bu pili göndermek zorundasınız bana ben bu ürünü sürekli prize takılı olarak kullanmak istemiyorum.	Negatif	Negatif
7- 2.200 TL verip aldım bir istek hakkım vardır herhalde.	Nötr	Nötr
8- En kısa zamanda iletişime geçin ve yeni pilimi gönderin bana.	Nötr	Nötr
9- Eğer göndermiyorsanız her türlü yola başvurabilirim.	Negatif	Negatif
10- Tüketici hakları ve kanunlar hakkında yeterince bilgim var	Nötr	Nötr

Çizelge 5.3'de sistemden çıkan sonuçlarla birlikte 10 adet kullanıcıya sorduğumuz sonuçlar 2174376 numaralı yorum için yüzde yüz oranla eşleşme sağlamıştır. Bu yöntem ile analiz edilen bu yorum için yüzde yüz doğruluk sağlamıştır. Genele bakıldığında ise büyük oranla doğruluk oranı yüksektir.



Şekil 5.1. Vader ve Stanford nlp analizi sonuçlarının karşılaştırılması

Şekil 5.1'de görülen grafikte çizgi ekseni Stanford NLP analizini temsil ediyor. Sütun ekseni ise Vader NLP yi temsil ediyor. Vader sonuçlarında nötr en fazla, Stanford sonuçlarında ise negatif en fazla görülmektedir. Şekil 5.1 ve Tablo 5.2'de

yer alan veriler bize Stanford NLP nin daha başarılı sonuçlar verdiğini ispatlamaktadır.

Kullanılan Terim Ağırlıklandırma yönteminde duygu skoru ölçülemeyeceği, sadece istatiksel olarak bir metinde en değerli kelime nedir sorusunun cevabının bulunabileceği. En değerli kelimelerin duygu değerleri el ve göz yardımıyla tekrar değerlendirilip sisteme girilmesi gerektiği gözlemlenmiş ve bu yöntem başarısız bulunmuştur.

Kullanılan diğer yöntem olan AHP' de ise; sektöre ait olumsuz söz öbeklerinin belirlenmesi, ardından bu söz öbeklerinin metin içerisinde geçme oranı tespit edilmesi gerekiyordu. Yine insan eli ve göz yardımıyla değerlendirilecek bir yöntem olduğu görüldü.

6. SONUÇ VE ÖNERİLER

Bu tez çalışmasında başka kullanıcılar tarafından oluşturulan marka & ürün yorumları insan emeğine ihtiyaç duyulmadan analiz edilmesi amaçlanmıştır. İnsanlar satın almak istedikleri ürünlere ait yorumların internet ortamında yorumlarına ve değerlendirmelere bakarak alıp almamaya karar vermektedirler. Bu son derece zahmetli ve uzun sürecek bir aşamadır. Bu süreci tamamen kısaltmanın hedeflendiği çalışmada Stanford Nlp yöntemi diğer yöntemlere göre daha üstün bir başarı sağlamıştır. Çizelge 5.3. de bu başarı tablo ile gösterilmiştir.

Bilgisayar sistemlerinde veri işleyebilmek için verilerin sayısallaştırılması gerekmektedir. Bu çalışmanın 4. Yöntem başlığında verilen tüm aşamaların tamamlanması hazır paket programlar kullanılmadan gerçekleştirilmiştir. Bu aşamaların hepsi için modüller Python, C#, R, Sql programlama dilleri kullanılarak yazılmıştır. Uzun bir araştırma sonrasında kullanılacak teknolojilere ve araçlara karar verilmiştir. Veri işleme konusunda esneklik bu sayede sağlanmış farklı yöntemler kolaylıkla denenebilmiştir.

Python dili veri analizi araçlarında diğer tüm dillere olan üstünlüğü keşfedilmiş ve veri ayrıştırma, işleme, duygu analizi, veri madenciliği işlemlerinin modüller bu dil ile yazılmıştır. Ancak karşılaşılan en büyük sorun bu işlemlerin kişisel bilgisayarlar aracılığıyla yapılmasından dolayı hız ve boyut sorunu ile karşılaşılmıştır. Ham veri boyutu 5GB civarında olmasına rağmen her yeni aşamada yeni veriler kayıt edildiğinden proje tamamlandığında 50GB veri boyutuna ulaşılmıştır. Daha performanslı bilgisayarların bu tarz veri işleme problemlerinde kullanılması gerektiği anlaşılmıştır.

R dilininde azımsanmayacak kadar performans ve başarı sağladığını söylemek gerekir. ~200 milyon kelime için kelime bulutlarını oluşturmada büyük başarım göstermiştir. Bu çalışma içerisinde R dili kelime bulutlarını oluşturmak için kullanılmıştır.

Türkçe diliyle yazılmış metinler için analiz yapabilmek için çalışmalar yapıldığı görülmüştür. Bunun en güzel örneği ve yaygın kullanılanı Zemberek isimli çalışmadır. Türkçe dili ile yazılmış metinlerde duygu bilgisini ortaya çıkartmak için kapsamlı bir araç geliştirilmesi gerektiği görülmüştür. Bunun İngilizce dili için en güzel ve başarılı örneği Stanford Üniversitesinin gerçekleştirdiği çalışma olan Stanford Nlp araçlarıdır.

KAYNAKLAR

(2018). We Are Social: https://wearesocial.com/ https://www.slideshare.net/wearesocial adresinden alındı

Abedini, M., Ahmadzadeh, F., & Noorossana, R. (2016). Customer credit scoring using a hybrid data mining approach. *Emerald Insight, 45*(10), 1576-1588.

Akpınar, H. (2014). *Veri Madenciliği Veri Analiz.* İstanbul: Papatya Yayıncılık.

Alizadeh, H., & Minaei-Bidgoli, B. (2016). Introducing A Hybrid Data Mining Model to Evaluate. *Engineering, Technology & Applied Science Research, 6*(6), 1235-1240.

Aydın, Y. D., & Özkul, P. (2015). Veri Madenciliği Ve Anadolu Üniversitesi Açıköğretim Sisteminde Bir Uygulama. *Eğitim ve Öğretim Araştırmaları Dergisi, 4*(3).

Çetingöz, M. (2011). Makine öğrenmesi ile Türkçe haber metinlerinde anahtar ifade çıkarımı.

Çoşkun, D. (2019, 02 26). *CIO*. http://www.cio.com.tr/blog/yapilandirilmamis-veri-nerede-saklanacak-scale-out-nas-mi-obje-tabanli-veri-depolama-mi/ adresinden alındı

Date Age 2025. (2019, 01 01). Seagate: https://www.seagate.com/tr/tr/our-story/data-age-2025/ adresinden alındı

Dil Bilimi. (2019, 02 16). Dil Bilimi: http://www.dilbilimi.net/anasayfa.htm adresinden alındı

Dogan, S., & Turkoglu, I. (2008). Karar Ağacı Yöntemini Kullanarak Biyokimya Verileri ile Anemi Teşhisi. *International Journal of Science & Technology, 3*(1), 85-92.

Domo. (2019, 01 26). 01 26, 2019 tarihinde Dünya üzerinde ki kullanıcılar tarafından oluşturulan veri raporu: https://www.domo.com/blog/data-never-sleeps-6/ adresinden alındı

Ergüler, C. (2019, 02 23). *biznet*. https://www.biznet.com.tr/veri-yonetisimi-data-governance-nedir/ adresinden alındı

Gray, P., & Watson, H. (1997). *Decision Support in the Data Warehouse.* Usa, United States of America: Prentice Hall, Inc.

Güvenç, B. (2016). Doğal dil işlemede makine öğrenmesi yöntemleri.

Hand, D. J. (1998). Data Mining: Statistics and More. *The American Statistician, 52*(2), 112-118.

Jacobs, P. (1999). Data Mining: What General Managers Need To Know. *Harvard Management Update, 4*(10), 8.

Kayaalp, K. (2007). Asenkron Motorlarda Veri Madenciliği İle Hata Tespiti. Isparta.

Kittler, R., & Wang, W. (1999). “The Emerging Role of Data Mining. *Solid State Technology*, 45.

L.Ladha, & T.Deepa. (2011). Feature Selection Methods And Algorithms. *International Journal on Computer Science and Engineering*, 1787-1788.

Luhn, H. P. (1958). The Automatic Creation of Literature Abstracts. *IBM Journal*.

Miguéis, V., Camanho, A., & Falcão e Cunha, J. (2012). Customer data mining for lifestyle segmentation. *Expert Systems with Applications*, 9359-9366.

Nanğır, M. (2013). Türk Dili İçin Çoklu Sınıflandırıcı Yöntemler ile Duygu Sınıflandırma.

Odabaş, Y., & Barış, G. (2002). *Tüketici Davranışı.* İstanbul: Mediacat Yayınları.

Okur, B. Ç. (2013). İngiliz alfabesi kullanılarak yazılmış Türkçe metinlerin Türk alfabesine göre yeniden oluşturulması. *IEEE*.

Sıramkaya, E. (2005). Veri Madenciliğinde Bulanık Mantık Uygulaması. Konya.

ŞEKER, Ş. E. (2018, 12 12). *Bilgisayar Kavramları*. http://bilgisayarkavramlari.sadievrenseker.com/2012/10/22/tf-idf/ adresinden alındı

Şeker, Ş. E. (2019, 03 30). *TF-IDF*. Bilgisayar Kavramları: http://bilgisayarkavramlari.sadievrenseker.com/2012/10/22/tf-idf/ adresinden alındı

TDK. (2015, 12 19). Türk Dil Kurumu: tdk.gov.tr adresinden alındı

Uçan, A. (2014). Otomatik duygu sözlüğü çevirimi ve duygu analizinde kullanımı.

Wikipedia. (2017, 04 01). 04 2017 tarihinde https://tr.wikipedia.org/wiki/B%C3%BCy%C3%BCk_veri adresinden alındı

Wikipedia. (2019, 02 16). https://tr.wikipedia.org/wiki/Do%C4%9Fal_dil_i%C5%9Fleme adresinden alındı

Wikipedia. (2019, 02 16). Wikipedia: https://tr.wikipedia.org/wiki/Dilbilim adresinden alındı

Wikipedia. (2019, 02 23). Wikipedia: https://tr.wikipedia.org/wiki/Metin_madencili%C4%9Fi adresinden alındı

EKLER

EK A. Python Kodları

EK B. C# Kodları

EK C. R Kodları

EK A. Python Kodları

https://github.com/mustafaerdogmus adresinde Tez Projesi Python kategorisi altında yayınlanmıştır

EK B. C# Kodları

https://github.com/mustafaerdogmus adresinde Tez Projesi C# kategorisi altında yayınlanmıştır

EK C. R Kodları

https://github.com/mustafaerdogmus adresinde Tez Projesi R kategorisi altında yayınlanmıştır

ÖZGEÇMİŞ

Adı Soyadı : Mustafa ERDOĞMUŞ

Doğum Yeri ve Yılı : UŞAK-BANAZ, 1988

Medeni Hali : Evli

Yabancı Dili : İngilizce

E-posta : erdogmusmustafa@gmail.com

Eğitim Durumu

Lise : Uşak Atatürk Lisesi, 2005

Lisans : Ahmet Yesevi Üniversitesi, Mühendislik Fakültesi, Yönetim Bilişim Sistemleri, 2014

Mesleki Deneyim

Akışık Bilgisayar	2006-2008
SDÜ Enformatik Bölüm Başkanlığı	2008-2014
SDÜ Bilgi İşlem Daire Başkanlığı	2014-(halen)

Yayınlar