T.C.
SÜLEYMAN DEMİREL ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

FARKLI DİL VE PLATFORMDA SEMANTİK ANALİZ

Volkan ALTINTAŞ

Danışman
Dr. Öğr. Üyesi Mehmet ALBAYRAK

II. Danışman
Dr. Öğr. Üyesi Kamil TOPAL

DOKTORA TEZİ
BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI
ISPARTA- 2020

TEZ ONAYI

Volkan ALTINTAŞ tarafından hazırlanan **"Farklı Dil ve Platformda Semantik Analiz"** adlı tez çalışması aşağıdaki jüri üyeleri önünde Süleyman Demirel Üniversitesi Fen Bilimleri Enstitüsü **Bilgisayar Mühendisliği Anabilim Dalı**'nda **DOKTORA TEZİ** olarak başarı ile savunulmuştur.

Danışman	**Dr.Öğr.Üyesi Mehmet ALBAYRAK** Isparta Uygulamalı Bilimler Üniversitesi	.....
Jüri Üyesi	**Prof.Dr. Ecir Uğur KÜÇÜKSİLLE** Süleyman Demirel Üniversitesi	.....
Jüri Üyesi	**Dr.Öğr.Üyesi Onur SEVLİ** Burdur Mehmet Akif Ersoy Üniversitesi	.....
Jüri Üyesi	**Dr.Öğr.Üyesi Ali KAVURUR** Burdur Mehmet Akif Ersoy Üniversitesi	.....
Jüri Üyesi	**Dr.Öğr.Üyesi Fırat YÜCEL** Akdeniz 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.

Volkan ALTINTAŞ

İÇİNDEKİLER

Sayfa

ÖZET

Doktora Tezi

FARKLI DİL VE PLATFORMDA SEMANTİK ANALİZ

Volkan ALTINTAŞ

Süleyman Demirel Üniversitesi
Fen Bilimleri Enstitüsü
Bilgisayar Mühendisliği Anabilim Dalı

Danışman: Dr. Öğr. Üyesi Mehmet ALBAYRAK

II. Danışman: Dr. Öğr. Üyesi Kamil TOPAL

Teknolojinin hızlı gelişimi ile beraber, internet, yaygın bir şekilde günlük hayatımızda kullanılmaktadır. İnternet ve internet teknolojilerinin yaygınlaşması ve her alanda kullanılması ile birlikte, üretilen veri miktarı her gün artmaya devam etmektdir. Boyut olarak artan verinin biçimlendirilerek analiz edilebilir hale getirilmiş şekli, "Büyük Veri" olarak adlandırılmaktadır. Büyük verinin, bireyler tarafından analiz edilmesi, yorumlanması ve anlamlı sonuçlara varılabilmesi ihtimali, verinin boyutundan dolayı kalmamıştır. Büyük verinin işlenebilmesi, işlenen verilerden anlamlı sonuçlar çıkarılabilmesi ve verilerin içerisinde varolan tematik bilginin ortaya çıkarılması son yıllarda önem kazanmıştır. Devletler, şirketler ve kurumlar, izleyecekleri politikaları depoladıkları verileri analiz ederek belirlemektedir. Bu konuda gelişen teknoloji ile verinin işleneceği donanım özelliklerinin de gelişmesi, araştırmalara katkı sunmaktadır. Algılayıcılardan toplanan veriler, sosyal medya paylaşımları, firmaların ve devlet kurumların barındırdığı veriler, büyük veri için örnek olarak gösterilebilir. Bu verilerin büyük bir çoğunluğu kullanıcılar tarafından oluşturulmaktadır. Kullanıcı tarafından veri paylaşımının en fazla yapıldığı ortamlar olarak sosyal medya platformları ön plana çıkmaktadır. Sosyal medya platformlarında kullanıcılar karşılaştıkları bir problem, güncel bir sorun veya herhangi bir konu ile ilgili yorumlarını ve deneyimlerini paylaşmaktadır.

Bu tez çalışmasında, iki farklı platform ve iki farklı dil için semantik analizi yapılmıştır. Türkçe ve İngilizce dillerinde kullanım oranları dikkate alınarak Reddit ve Ekşi Sözlük sosyal medya platformları seçilmiştir. Çalışmada, bu platformlarda teknoloji kanalında paylaşılan kullanıcı yorumları veri ön işleme adımlarının ardından, Gizli Anlam Analizi (GAA) ve Gizli Dirichlet Ayrımı (GDA) algoritmaları ile konu modellemesi işlemi gerçekleştirilmiştir. İki algoritmanın sonuçlarında oluşan benzerlikler ve farklılıklar hem aynı dilde hem de Türkçe ve İngilizce dilleri için ayrı ayrı incelenmiştir. Konu modellemede öne çıkan yorumlar üzerinden, Varlık İsmi Tanıma (VİT) metotları kullanılarak yorumlar içerisinde geçen varlık isimleri bulunmuştur. Çevrimiçi ansiklopedi olan WikiPedia' daki metinsel bilgilerin semantik algoritmalar yardımıyla formatlı bilgi haline getirildiği DBPedia üzerinde VİT metotları ile tespit edilen varlık

isimleri açıklamaları ile eşleştirilmiştir. Analiz edilen büyük veri üzerinde belirlenen sosyal medya platformlarında konuşulan tematik konular tespit edildiği gibi, ayrıca konuların belirlenmesinde etkin olarak geçen yorumlardaki varlık isimleri ve açıklamaları da belirlenmiştir.

Anahtar Kelimeler: Doğal dil işleme, metin madenciliği, konu modelleme, varlık ismi tanıma, ontoloji.

2020, 84 sayfa

ABSTRACT

PhD Thesis

SEMANTIC ANALYSIS IN DIFFERENT LANGUAGE AND PLATFORM

Volkan ALTINTAŞ

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

Supervisor: Asst.Prof.Dr. Mehmet ALBAYRAK

Co-Supervisor: Asst.Prof.Dr. Kamil TOPAL

With the rapid development of technology, the internet is widely used in our daily life. With the spread of internet and internet technologies and their use in every field, the amount of data produced continues to increase every day. The format of the increasing data in size, which has been formatted and analyzed, is called "Big Data." The possibility of big data being analyzed, interpreted, and meaningful conclusions by individuals are not due to the size of the data. It has gained importance in recent years to be able to process big data, to draw meaningful conclusions from the processed data, to reveal the thematic information existing in the data. States, companies, institutions determine the policies they will follow by analyzing the data they store. In this regard, the development of the technology and the hardware features of the data will contribute to the research. Sensor data, social media shares, data hosted by companies, and government agencies can be shown as examples for big data. Users create the vast majority of this data. Social media platforms come to the fore as environments where data sharing is made most by the user. On social media platforms, users share their comments and experiences about a problem they face, a current situation, or any topic.

In this thesis, the semantic analysis was done for two different platforms and two other languages. Reddit and Ekşi Sözlük social media platforms were selected by taking into consideration the usage rates in Turkish and English languages. In this study, the topic modeling process was carried out with Latent Semantic Analyzer (LSA) and Latent Dirichlet Allocation (LDA) algorithms after user comments data preprocessing steps shared on technology channel in these platforms. The similarities and differences in the results of the two algorithms are examined separately for both the same language and Turkish and English languages. Entity names in the comments were found by using Name Entity Recognition (NER) methods. The text names in Wikipedia, the çevrimiçi encyclopedia, are matched with the descriptions of the asset names determined by NER methods on DBPedia, where semantic algorithms are converted into formatted information. The thematic topics spoken on the social media platforms defined on the big data obtained were identified, as well as the asset names and

their explanations in the comments that were actively involved in the determination of the topic.

Keywords: Natural language processing, text mining, topic modelling, name entity recognition, ontology.

2020, 84 pages

TEŞEKKÜR

Bu araştırma için beni yönlendiren, karşılaştığım zorlukları bilgi ve tecrübesi ile aşmamda yardımcı olan değerli Danışman Hocam Dr. Öğr. Üyesi Mehmet ALBAYRAK ve II. Danışman Hocam Dr. Öğr. Üyesi Kamil TOPAL'a teşekkürlerimi sunarım.

Tez çalışmam süresince desteklerini esirgemeyen anneme ve babama teşekkürü bir borç bilirim. Hayatımın her alanında olduğu gibi çalışmalarım süresince bana destek olan, tezimin her aşamasında beni yalnız bırakmayan sevgili eşim Aslı ALTINTAŞ'a ve çalışmalarım boyunca sabreden, destek olan çocuklarım Batu ve Efe ALTINTAŞ'a sonsuz teşekkür ederim.

Volkan ALTINTAŞ
ISPARTA, 2020

ŞEKİLLER DİZİNİ

ÇİZELGELER DİZİNİ

SİMGELER VE KISALTMALAR DİZİNİ

AA	Anlamsal Ağ
BİT	Bilgi İletişim Teknolojileri
DDİ	Doğal Dil İşleme
GAA	Gizli Anlam Analizi
GDA	Gizli Dirichlet Ayrımı
KDM	Karar Destek Makinesi
KTÇ	Kaynak Tanımlama Çerçevesi
NBMN	Naive Bayes Multinomial
NMF	Non-Matrix Factorization
TDA	Tekil Değer Ayrıştırma
TF/TDF	Terim Frekansı/Ters Doküman Frekansı
VİT	Varlık İsmi Tanımlama

1. GİRİŞ

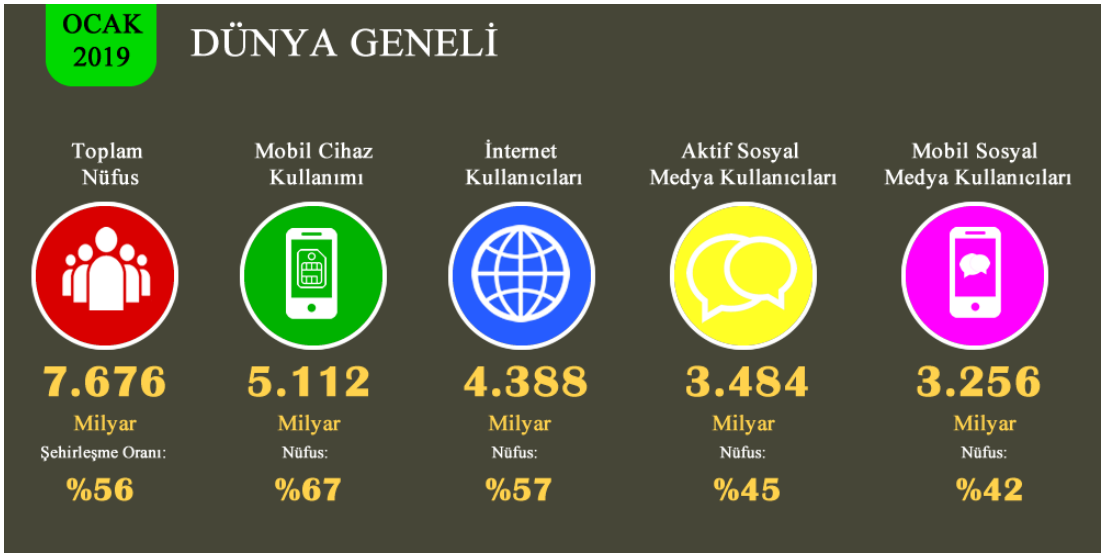
Teknoloji, canlıların ihtiyacı olan mal ve hizmetlerin üretilmesi ve geliştirilmesi için gerekli bilgi, beceri, yetenek ve bilimsel araştırma olarak tanımlanmaktadır (WikiPedia, 2001). Teknoloji, toplumların ekonomik, kültürel, siyasal ve sosyal hayatını 19. yüzyıl' dan itibaren önemli şekilde değiştirmiştir. Teknoloji kavramı insanların hayatlarını kolaylaştırmak amacı ile bulunmuştur. Teknolojideki ilerlemelerin de insan ihtiyaçları ile yakından ilgisi bulunmaktadır. Teknolojinin temelinde, teknolojiyi ortaya çıkaracak kadar bilgi birikimi ve deneyim bulunmaktadır. Bilgi, sürekli kendisini yenilediği ve güncellediği için teknoloji de bilgi ile beraber kendini yenilemekte ve güncellemektedir. Sanayi devriminin ortaya çıkışı ile başlayan teknolojik gelişim süreci, 20. yüzyılın sonlarından itibaren ise önemli ölçüde, Bilgi ve İletişim Teknolojileri (BİT) çerçevesinde şekillenmeye başlamıştır.

BİT, bilgisayar donanımı, yazılımı ve diğer pek çok bileşenin oluşturduğu, bilgiyi depolamak, işlemek ve dağıtmak için gerekli altyapı ile bütünleşik bir sistem olarak tanımlanmaktadır (Wangwe, 2007). 20. yüzyılın ortalarında bilgisayar, akıllı telefon ve internet gibi teknolojiler olmamasına rağmen hızla gelişen teknoloji ile radyodan televizyona geçiş yaşanmıştır. 80' li yıllardan itibaren bilgisayar teknolojisindeki büyüme ve devamındaki yıllarda bilgisayar parçalarının fiyatlarının azalması nedeniyle bilgisayar her kurumun, her evin vazgeçilmez bir parçası haline gelmiştir. 90' lı yıllarda internetin hayatımıza girmesi ile beraber uzak konumlar birbirine dijital olarak yakınlaşmış ve insanların birbirleriyle iletişim kurmalarına olanak sağlamıştır. İnternet sayesinde dünya küçülmüş ve iletişim olanakları artmıştır. İnternet, günümüzde dev bir kütüphane halini almış ve büyümesini her gün devam ettirmektedir.

İnternet, bireylerin üretilen veriyi saklama, paylaşma, kolay ve hızlı bir şekilde ulaşma isteği sonucu ortaya çıkan bir teknolojidir. İnternetle birlikte; zaman ve mekân yeniden yapılandırılmıştır. Teknoloji, bilginin sistematik olarak işlenmesini amaçlamaktadır. Dijital teknolojiyle birlikte karmaşıklıktan çok, az bilgiyle çok işlem yapabilecek arayüzler üretilmektedir. İnternet, web ortamında bulunan bilgi ve belgelerin çoğalmasının arkasındaki en büyük itici güç olmuştur..

Dünya var olduğu andan itibaren sürekli değişim ve dönüşüm içerisinde olmuştur. Ancak son yıllarda iletişim teknolojilerinin gelişmesiyle, dünya adeta küçük bir köy haline gelmiştir. Özellikle 21. yüzyılda geliştirilen teknolojik aletler ve en önemlisi akıllı cep telefonlarıyla kullanılan internet, dünyayı avucunun içine sığdırabilmiştir. İnternet ve sosyal medya araçlarıyla sosyalleşen insanlar artık birey olmaktan ziyade sanal âlemde toplumsal bir güç olmaktadırlar. İş, eğitim, sağlık, ticaret, ziyaret, ulaşım gibi hayatın neredeyse her alanına etki eden sosyal medya bileşenleri internet aracılığıyla insanların hayatlarında çok önemli bir yere sahip olmuştur. Günümüzde dünya nüfusunun yaklaşık olarak %40' ı internete bağlanmaktadır. Bu oran, 1995 yılında %1' ler seviyesinde bulunmaktaydı (Çalışkan ve Mencik, 2015).

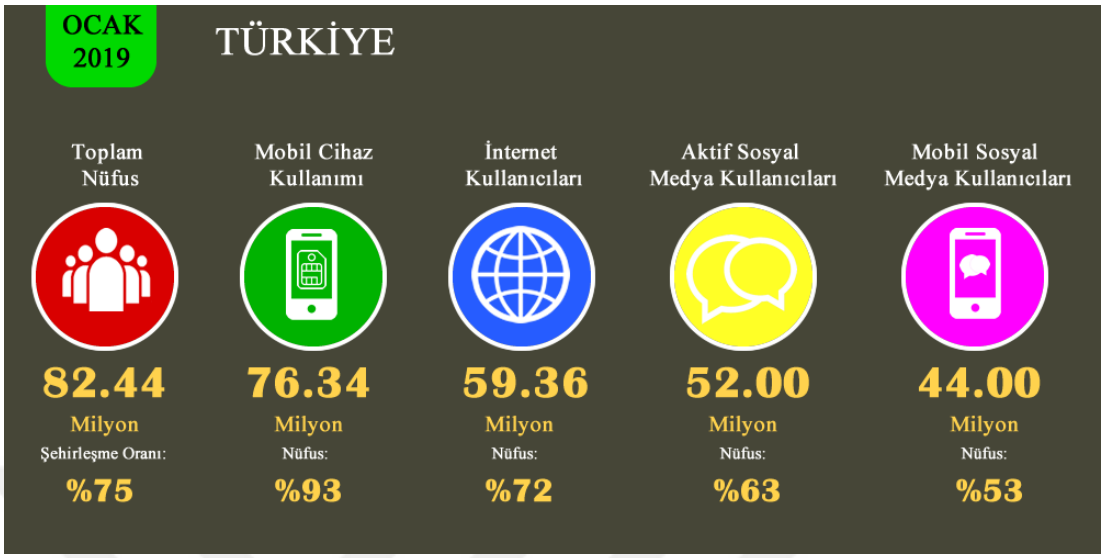
Nüfusunun yaklaşık %56' sı şehirlerde yaşayan dünyada, 5.11 milyar benzersiz mobil kullanıcı vardır. 2019 yılında 4.39 milyar internet, 3.48 milyar sosyal medya kullanıcısı bulunmaktadır. Ocak 2019' da 3.26 milyar kişi mobil cihazlarda sosyal medya kullanmaktadır. 2019 yılında, dünya genelinde mobil cihaz, internet ve sosyal medya kullanımı Şekil1.1' de gösterilmektedir (Wearesocial, 2008).



Şekil 1.1. 2019 yılı dünya geneli dijital kullanımı (Wearesocial, 2008)

Türkiye' de toplam nüfusun %93' ü mobil cihaz kullanmaktadır. 2019 yılında 59,36 milyon internet, 52 milyon aktif sosyal medya kullanıcısı bulunmaktadır. Ocak 2019' da 44 milyon kişi mobil cihazlarda sosyal medya kullanmaktadır.

Şekil 1.2' de ise Türkiye' deki mobil cihaz, internet ve sosyal medya kullanımı gösterilmektedir (Wearesocial, 2008).



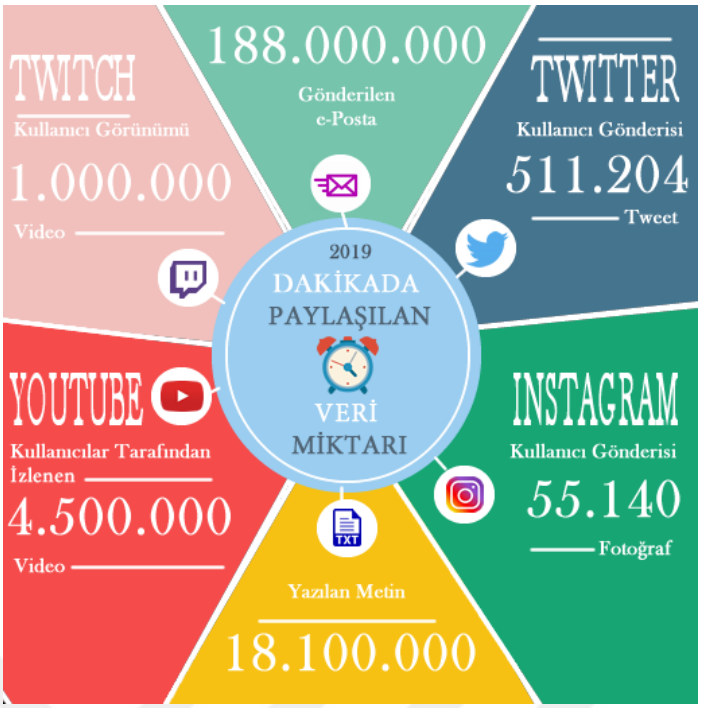
Şekil 1.2. 2019 yılı Türkiye geneli dijital kullanımı (Wearesocial, 2008)

İnternet teknolojileri de bilişim teknolojilerine bağlı olarak büyük bir gelişim göstermiştir. Bu gelişmeler ile birlikte, donanım maliyetlerinin gittikçe ucuzlaması, sürekli olarak ürettiğimiz ve depolanan veri miktarında büyük bir artış meydana getirmiştir. Günümüz bilişim ortamında bilgiye ulaşmak ve bilgiyi kontrol etmek önem kazanmaktadır. Kullanıcılar için bilgiye erişmek basittir, ancak bilgiyi güncellemek ve paylaşmak kullanıcılar için çok kolay olmamaktadır. Web 1.0' ın kişisel eğlence mantığı, Web 2.0' la birlikte kişisel yayıncılığa dönüşmüştür. Web 3.0' ın özellikleriyle birlikte kişisel yayıncılıkta karşılaşılan kodlama ile ilgili karmaşıklık yerini görsel tabanlı sistemlere bırakmaktadır. "Tut-sürükle-bırak" mantığıyla çalışan bu sistemler kullanıcının içeriğe müdahalesini mümkün hale getirmektedir. Web 3.0, toplayıp birleştirme devri olarak ifade edilmektedir (Creamer, 2008). Anlamsal web teknolojisi ile toplanan veriler hem kullanıcılar tarafından kullanılmakta hem de bilgisayarların anlamlandırabileceği bir yapıya ulaşmaktadır. Veri dünyamızı hızlı bir şekilde değiştirmektedir. Zamanın başlangıcından 2003' e kadar ürettiğimizden daha fazla veri, her iki günde bir üretilmektedir. Ürettiğimiz bu veriler devletler, şirketler vb. organizasyonlar için kısacası herkes için değer taşımaktadır. Bütün bu yapı büyük veri kavramını ortaya çıkarmıştır. Web teknolojilerinde yaşanan gelişmeler sonucunda çevrimiçi blog, forum, yorum vb. sitelerin yaygınlaşması ve

bu siteler üzerinden yapılan ürün, hizmet, kalite yorumlarının kullanıcılar üzerinde etkili olduğu görülmektedir.

Hızlı gelişen teknoloji, geleneksel kitle iletişim araçlarının daha özellikli olmasını sağlarken, yeni iletişim araçlarını da kitlelere sunmaktadır. Bunlardan en yenisi, son yıllarda popüler hale gelen, bireyler ve işletmeler tarafından eğitim, eğlence, bilgi edinme, reklam vb. amaçlarla sıklıkla kullanılan sosyal medyadır. Rekabet üstünlüğü sağlayabilmek için teknolojik çevreyi yakından takip etmek isteyen işletmeler, sosyal medyayı da iletişim kanalı olarak en etkin biçimde kullanmak istemektedirler. İnsanlar gerek çevresi gerekse de dünyada olup bitenden haberdar olmasının yanında ürün ya da hizmetlere ilişkin bilgi edinme, eğlenme, arkadaş edinme ve bazen de yaşamın baskısından kurtularak rahatlama gibi birtakım gereksinimleri sosyal medya aracılığıyla sağlamaktadır.

Şekil 1.3' de 2019 yılında, internet ortamında bir dakikada kullanıcılar tarafından üretilen veri miktarı gösterilmektedir. Sosyal medya platformu Twitter' da 511.200 adet gönderi paylaşılmıştır. Instagram üzerinden 55.140 fotoğraf paylaşılmıştır. Gönderilen metinlerin sayısı 18 milyonun üzerindedir. Bir dakikada gönderilen e-posta sayısı yaklaşık 188 milyondur. Teknoloji ile bağımlı olarak paylaşılan veri miktarı her geçen gün daha da artmaktadır. Paylaşılan verilerin önemi ve miktarının büyüklüğü burada rakamlar ile daha anlaşılabilir olmaktadır. Tamamen kullanıcılar tarafından paylaşılan bu verilerin işlenerek anlamlı sonuçlar ortaya konulması daha da önemli hale gelmektedir. Büyük veri analizleri, insanların davranışılarını anlama ve düşüncelerini çözme konusunda önceki zamanlara göre daha fazla katkı sunabilir. Gelecek hakkında daha tutarlı kararlar alma konusunda tahminler geliştirebilir. Diğer bir yönden bakıldığında yapılan analizlerin başarı başarısı ne kadar çok verinin saklandığı değil, bilginin ortaya çıkarılması, davranış ve düşüncelerin tahmin edilebilme kapasitesinde yatmaktadır.



Şekil 1.3. 2019 yılı dakikada üretilen veri miktarı

Sosyal medyanın ortaya çıkışı ile birlikte insanların sosyalleşmek için iletişim kurma şekillerini değiştirip değiştirmediklerine yönelik tartışmalar, sosyal medyanın tüm dünyada ortaya çıkan kitle hareketlerindeki rolü sebebi ile artan bir ilgiye ulaşmıştır. Son dönemlerde meydana gelen Arap Baharı, İspanyol Öfkeliler Hareketi (Indignados Protests), İşgal Et Eylemleri (Occupy Wall Street) gibi kitlesel eylemler sosyal medya ortamlarında yoğunlaşan aşırı oranda iletişim teknolojilerinin kullanılması ve bilgi akışı sebebiyle tüm dünyanın dikkatini çekmektedir. Büyük kitlelerin takip ettiği sosyal medya platformlarında yapılan paylaşımları doğru ölçütlerle anlamak, filtrelemek, analiz etmek, metinler aracılığıyla oluşan anlam ürünleri ile ilgilenen geniş kapsamlı sosyal ve kültürel araştırmalar için kullanılan bir araştırma yöntemi olan semantik analizi yapılarak yorumlamak önemli hale gelmiştir (Conover vd., 2013).

Veri analizi bilgisayar bilimi, biyoloji, tıp, finans ve ülke güvenliği gibi disiplinlerde son derece önemli ve zorlu bir problem olmuştur. Büyük miktarda veriyi analiz etmek oldukça zor bir iş haline gelmiştir. Dünyadaki bilgi miktarı büyüktür ve katlanarak artmaktadır. Örneğin, Facebook gibi çeşitli sosyal ağlar günde terabaytlarca veri, video, duvar yazısı vb. şeklinde veri üretir ve yakın gelecekte önemli ölçüde daha fazla veri üretecektir.

Bugün verilerin boyutunun büyük olmasından dolayı daha önce görülmemiş ve geleneksel veri yönetimi teknikleri ile analiz edilemez. Bununla birlikte, büyük verilerden etkin bir şekilde "anlam çıkarabilmek", farklı alanlarda her zamankinden daha fazla önem kazanmaktadır. Bilgisayar biliminde, küresel eğilimleri ve kullanıcı davranışını anlamak için internet ölçeğinde verilerin analiz edilmesi gerekmektedir. Biyolojide, karmaşık biyolojik sistemleri anlamak için çok miktarda DNA ve RNA dizilim verisinin yorumlanması esastır. Sıralama verilerinin üssel olarak büyümesi, depolama kapasitesinin büyüme oranını aşmıştır. Sağlık alanında sağlık cihazları, uyku, kalp atış hızı ve diğer sağlık koşullarını izleyerek hastaların durumunu yansıtan çok miktarda veri üretir. Finans alanında, borsa, şirketlerin karlarını maksimize etmelerine yardımcı olabilecek büyük miktarlarda işlem verisi üretir. Ülke güvenliğinde, hükümetler her gün kütüphanelerde bulunan metin miktarından daha fazla miktarda terabaytlarca veri toplamaktadır. Bu veriler, daha sonra ülkeye yönelik potansiyel tehditleri belirlemek için analiz edilebilir. Örneklerde görüldüğü gibi, büyük miktarda bilgiden yararlanmaya başlayan birçok alan bulunmaktadır.

Tweetler, sosyal medya mesajları, blog yazıları, forum yazıları ile gittikçe artan miktarda metin verisi üretilmeye devam etmektedir. Verinin içerisinden anlamlı ve istenen bilgilerinin bilgisayarın anlayabileceği anlamların çıkarılması için Doğal Dil İşleme (DDİ) teknikleri kullanılmaktadır. DDİ; sohbet botları (chatbots), makale veya yazıların özeti, dil çeviri ve veriden görüş tanımlama gibi birçok akıllı uygulamada kritik bir rol oynamaktadır. Makine Öğrenmesi modellerini ve algoritmalarını uygulayabilmemiz için öncelikle metinlerin işlenmesi gerekmektedir.

Hesaplamalı dil bilim hem bilimsel hem de mühendislik olarak görülebilir. Genellikle DDİ olarak adlandırılan hesaplamalı dil bilimin mühendislik ve yapay zekâ tarafı, büyük ölçüde, dil ile yararlı şeyler yapan hesaplama araçları oluşturma ile ilgilidir. İstatistiksel DDİ olasılıkları, bir ifadeyi veya metni analiz ederken karşılaşılan alternatiflerle ilişkilendirir ve en olası sonucu doğru olanı kabul eder. Dünyayla yakından ilişkili olan olguları olan sözcükler, sık sık birbirine yaklaşır ve bu türdeki metinleri hızlıca çözmek daha kolay ve güvenilir olmaktadır.

Bu tez çalışmasında, büyük veri niteliğinde iki farklı platformdan (Reddit ve EkşiSözlük) ve iki farklı dil (Türkçe ve İngilizce) elde edilen verilerin DDİ teknikleri ve konu modelleme algoritmaları kullanılarak bilgisayarlar tarafından da anlaşılıp işlenebilmesi sağlanacaktır. Elde edilen analiz sonuçları yorumlanarak konu hakkında yapılan paylaşımların eğilimleri belirlenecektir. Konu başlıklarında geçen varlık isimleri belirlenerek DBPedia projesi üzerinde varlık isimlerine ait tanımlamalar bulunacaktır. Çalışma teknoloji başlığı altında dünya ve Türkiye' de paylaşılan verilerin işlenmesini üzerine olsa da kullanılacak olan yöntem ve tekniklerin ileri de farklı yaklaşımlar için de kullanılması açısından önem taşımaktadır. Bir başka açıdan bakıldığında; teknolojinin gelişmesi ile beraber günlük yaşamımızda sürekli olarak kullandığımız sosyal medya ortamlarında paylaşılan verilerin analizi, yorumlanması ve sonuç çıkarımı konusunda önemli katkı sunması beklenmektedir.

2. KAYNAK ÖZETLERİ

Bu bölümde, doğal dil işleme yöntemleri ve konu modelleme temelli yaklaşımlar ile ilgili literatürde yapılan çalışmalar kapsamlı bir şekilde araştırılmıştır.

Blei vd. (2003), yaptıkları çalışmada, metin topluluğu (corpus) gibi ayrık verilerin toplanması için olasılık modeli olan Gizli Dirichlet Ayrımı (GDA) modelini açıkladılar. Temel GDA modeline göre, her belgedeki her kelime, konu dağılımlarının bir karışımı arasından seçilmektedir. Bir belge, konuların bir dağılımını içerir. GDA, Markov zinciri Monte Carlo simülasyonlarını kullanarak kelime ve belge konu dağılımlarındaki parametreleri tahmin etmektedir.

Li vd. (2010), dokümandaki belirgin özellikleri çıkarmak amacıyla "Sentiment-LDA" ve "Dependent Sentiment-LDA" olmak üzere iki yöntem önermişlerdir. Önerdikleri yöntemler ile ürün özelliklerini çıkartırken eşzamanlı olarak duygu ifadelerini de çıkartmışlardır. Bu yöntemleri; dokümandaki duygu ifadelerinin temel varlık ile ilişkili olması fikrinden yola çıkarak geliştirmişler ve duygu ifadeleri ve belirgin özellikleri bütün olarak ele alınmıştır. "Dependent Sentiment-LDA" ile ayrıca sentiment polaritelerinin belirlenmesini amaçlamışlardır.

Balkan ve Takcı (2010), kavram çıkarmanın bir aşaması olan terim benzerliklerine dayalı kümelemede kullanılan üç yöntemi incelemiş ve en başarılı yöntemi bulmaya çalışmıştır. Çalışma, Türkçe dili üzerinde yapılmıştır. 11 dokümandan oluşan küçük bir veri seti kullanılmıştır. Dokümanların hepsi eğitim konusundadır ve ortalama 1500 kelime içermektedir. Denemelere başlamadan önce her bir dokümanı sunacak terim özellik seti çıkarılmıştır. GAA ve k-Ortalama yöntemi, terimlerin dokümanlara dağıtılması konusunda karşılaştırılmıştır. Daha sonra bir uzman tarafından yapılan dağıtımın doğruluğu kontrol edilmiştir. Sonuç olarak k-ortalama yönteminin bu veri seti ve çalışma için en doğru sonucu verdiği görülmüştür.

Zhao vd. (2011), yaptıkları çalışmada geleneksel medya (New York Times) ile Twitter konu içeriklerini karşılaştırmıştır. Twitter' deki konu başlıklarını ortaya çıkarmak için GDA kullanmışlardır. Kısa tweet' ler için tasarlanmış ve mevcut

modellere kıyasla etkinliğini gösteren yeni bir Twitter-GDA modeli geliştirmişlerdir. Twitter ve geleneksel haber medyaları arasındaki güncel farklılıkların analizini kolaylaştırmak için konu kategorileri ve konu türleri kavramlarını tanıtmışlardır. Twitter üzerinden elde edilen konuların, geleneksel haber medyasında az yer bulan haberlere iyi bir kaynak olabileceğini, Twitter kullanıcıların önemli haberlerin yayılmasında önemli rol oynadığı sonucunu elde etmişlerdir.

Jo ve Oh (2011), aynı başlık altındaki ürün özelliklerinin yorum içerisinde birbirine yakın oldukları fikrinden yola çıkarak bir cümledeki tüm kelimelerin tek bir ürün özelliği ile ilişkili olduğu yaklaşımını varsayan "Sentence-GDA" yöntemini önermişlerdir. Sonra ise bu yöntemin gelişmiş bir hali olan "Aspect Sentiment Unification" geliştirmişlerdir. Bu yöntemde ürün özellikleri ve duygu ifadeleri birlikte modellenerek özellik, duygu ifadesi çiftlerinin çıkartılması sağlanmıştır. Bu amaçla duygu ifadelerinin küçük bir kümesinden de yararlanılmıştır. Yöntemler elektronik ve restoran veri kümelerine uygulanarak ürün özellikleri ve özellik, duygu ifadesi çiftleri başarılı bir şekilde elde edilmiştir.

Ünaldı ve Kırkgöz (2011), yaptıkları çalışmada, anadili İngilizce olan üniversite öğrencileri ile anadili Türkçe olan üniversite öğrencileri tarafından oluşturulan metinleri GAA algoritması kullanarak karşılaştırmışladır. Karşılaştırmanın yapılabilmesi için anadili Türkçe olan öğrencilerin İngilizce olarak yazdığı metinlerden bir derlem oluşturulmuş ve bu derlem anadili İngilizce olan üniversite öğrencileri tarafından yazılmış metinleri içeren başka bir derlemle karşılaştırılmıştır. Bu karşılaştırma sürecinde üç çeşit anlam indeksi kullanılmıştır: tümce, paragraf ve metin geneli. Elde edilen sonuçlar, bu yöntemin literatürde geçen diğer çalışmalarda ortaya çıkarılmış olan yeterliğiyle paralellik göstermektedir.

Kuzu vd. (2012), çağrı merkezi kayıtlarının otomatik ses tanıma sistemiyle yazıya dökülmesinden sonra, metinsel içeriklerin öznitelik vektörleri ile temsili ve farklı örüntü tanıma yöntemleri ile sınıflandırılması üzerine çalışmıştır. Öznitelik çıkarımında vektör uzayı modeli, gizli anlamsal analiz ve tekrarlı artık ölçekleme yöntemlerinden yararlanılırken, k-Ortalama sınıflandırıcı, Karar Destek

Makineleri (KDM) ve YSA kullanılarak da metinsel içerikler, konularına göre sınıflandırılma yoluna gidilmiştir. 6 kategori için, KDM, k-Ortalama ve YSA kullanılarak sınıflandırma eğitimi yapılmış, konuşma tanıma çıktılarından gelen metinlerle bu sınıflandırıcılarda metinlerin sınıflandırma testleri yapılmıştır. KDM ve YSA' nın benzer metin sınıflandırma eğrileri çizdiği gözlemlenmiştir.

Yazdani ve Belis (2013), yaptıkları çalışmada, Wikipedia' yı kullanarak kelimelerin veya metinlerin anlamsal ilişkisini hesaplama yöntemi önermişlerdir. Ansiklopedi makalelerini filtrelemek suretiyle bir kavramlar ağı oluşturulmuştur ve her kavram bir makaleye karşılık gelmektedir. Kavramlar arasında iki tür ağırlıklı bağlantı göz önüne alınmaktadır. Biri yazıların metinleri arasındaki bağlantıları, diğeri metinler içerisinde bulunan sözcüklerin benzerlikleridir. Kavramların birbirine olan bağlantıları üzerinden rastgele hareket eden ve hareket mesafesini hesaplayan bir algoritma geliştirilmiştir. Geliştirilen algoritmanın içerisinde k en yakın komşuluk grafiğinin Wikipedia gibi büyük bir kaynakta kullanılabilirliği de gösterilmiştir.

He vd. (2013), çalışmalarında 3 büyük pizza şirketi arasında sosyal medya verilerini kullanılarak rekabet analizi çalışması yapılmıştır. Ay bazında müşterilerin beğenileri, paylaşımları ve yorumları ele alınarak metin madenciliği ile analiz sonuçlarını ortaya koymuşlardır. Analizler sonucunda firmaların sosyal medyadaki rekabeti geliştirmesine yönelik tavsiyelerde bulunmuşlardır.

Conover vd. (2013), çalışmasında yazarlar 2011 yılında Amerika Birleşik Devletleri' nin New York şehrinde gerçekleşen ¨Occupy Wall Street¨ protestolarını incelemişlerdir. Bu olaylar, finansal çevrelerce ünlü Wall Street caddesinde, tüm dünyadaki ekonomik eşitsizliği protesto etmek için başlamıştı. Protestolar esnasındaki etkin kullanıcıların geçmiş paylaşımları incelenmiş ve bu kullanıcıların birbirleriyle sosyal medyada iletişimlerinin olduğu, ayrıca yerli ve yabancı sosyal hareketlere karşı duyarlı olduklarını gözlemlenmiştir.

Erdur ve Alatlı (2013), yaptıkları çalışmada, birbiri ile ilişkili veri bulutlarında bulunan veri setlerini sorgulamak ve özel veriler üzerine yoğunlaşarak veriler üzerinde oluşan değişimleri izleyebilen mobil uygulamaların geliştirilmesini

basit hale getirmek için bir sistem tasarlamışlardır. Önerilen sistem ile "bağlı veri ortamı" olarak adlandırılan ve ilişkili veri bulutları ile mobil uygulamaları soyut olarak incelenmektedir. Sistemin kullanılabilirliği göstermek için filmler hakkında bilgileri tutan en geniş veri setlerinden biri olan LinkedMDB kullanılmıştır. Düşünülen senaryoda bir Android telefon kullanıcısı, en sevdiği yönetmen olan Robert Redford' un yeni bir filmi vizyona girdiğinde haberdar olmak istemektedir. Mobil uygulama için SPARQL sorgusu oluşturulmuş ve bu sorgunun takip edilmesi sağlanmış ve ilgili yönetmene ait filmler listelenmiştir. Filmlerde herhangi bir değişim olduğunda listenin sonuna eklenmiş ve kullanıcının haberdar edilmesi sağlanmıştır.

Zhou ve Chen (2014), Twitter üzerindeki veri akışına göre olay algılama üzerine bir çalışma yapmışlardır. Gerçek zamanlı uygulamalar için tweet akışları üzerinden, çevrimiçi sosyal olay izlemesi, kriz yönetimi ve karar verme gibi sorunları incelemişlerdir. İlk olarak, atılan tweetlerin temsili için sosyal içerik, yer ve zaman bilgilerini birleştiren yeni bir yer-zaman kısıtlı konu modeli önermişlerdir. İletilerin benzerliğini bulmak için içeriği, yeri, zamanı ve bağlantıyı içeren dört öznitelik üzerindeki fark gözetilmiştir. Son yıllarda Avustralya' da meydana gelen iki kriz sırasında uzun tweet akışları üzerinden kapsamlı deneyler yapılmıştır. GDA tabanlı çevrimiçi olay tespit etmek için oluşturulan model, GDA ile karşılaştırıldığında, çevrimiçi olay tespit etme performansının daha iyi olduğu gözlemlenmiştir.

Sarker vd. (2015), yaptıkları çalışmada sosyal medya verilerini kullanarak "Farmakovijilans" için bir inceleme yapmışlardır. İlaçların neden olduğu zararlı hasta sonuçları olarak tanımlanan ilaçların zararlı reaksiyonlarının otomatik olarak izlenmesi, halen tıp bilişimi topluluğundan ilgi gören zorlu bir araştırma problemi olduğu için son yıllarda sosyal medyada kullanıcı tarafından yayınlanan veriler, ilaçların zararlı reaksiyonlarını izleme için yararlı bir kaynak haline gelmiştir. Sosyal medyadan ilaçların zararlı reaksiyonlarının tespitinde farklı yaklaşımları karakterize etmek için metodolojik bir inceleme gerçekleştirmişlerdir. Medline, Embase, Scopus Google Scholar arama motoru, Web of Science veritabanlarından ve sosyal medyadan ilaçların zararlı reaksiyonlarının tespiti ile ilgili yaklaşımlar ve çalışmalar tespit edilmiştir.

Çalışma, öncelikle ilaçların zararlı reaksiyonlarını bulma yaklaşımı, verinin boyutu, veri kaynağı, kullanılabilirlik ve değerlendirme ölçütleri gibi farklı özelliklere göre sınıflandırılmıştır. İnceleme sonunda, ilaçların zararlı reaksiyonlarını izleme için geniş miktarda sosyal medya verisinin kullanımına olan ilginin arttığı gözlemlenmiştir.

Yalçınkaya ve Singh (2015), tarafından yapılan çalışmada, "Bina Bilgi Modellemesi" literatürünün kapsamlı, sayısal ve sistematik bir sınıflandırması bulunmamasından dolayı 2004' ten 2014' e kadar yayımlanan "Bina Bilgi Modellemesi" araştırma çalışmalarının büyük bir bölümü sentezlenmiş ve etiketlenmiştir. 975 akademik makalenin özetlerine doğal dil işleme tekniği olan Gizli Anlam Analizi (GAA) uygulanmıştır. Yapılan analiz sonucunda, on iki ana araştırma alanı ortaya konmuştur. Her ana alanla ilişkili çeşitli spesifik araştırma temaları belirlenmiştir. Bu temel araştırma alanları ve araştırma temaların, Bina Bilgi Modellemesi araştırmalarındaki model ve eğilimlerini gösterdiği belirtilmiştir.

Hatipoğlu ve Omurca (2015), tarafından yapılan çalışmada Türkçe' nin yapısal özelliklerine göre istatistiksel olarak puanlandırılması ve gizli anlam çıkarım yöntemlerini sezgisel olarak birleştirilerek cümle seçimi yapan melez bir model sunulmuştur. Çalışmada, metin özetleme sorunu üzerine çalışılmıştır. Özet cümlelerin seçimi, özetlenecek metinlerin Türkçe' nin dil özelliklerine dayalı istatistiksel puanlandırılması ve anlamsal puanlandırılması yöntemlerinin melez şekilde değerlendirilmesi ile gerçekleştirilmiştir. Özetlenecek metinlerde yer alan cümlelerin özet cümle adaylığı için aldıkları puanlar, yapısal ve anlamsal özelliklerin sezgisel bir ağırlıklandırma yöntemi ile birleştirilmesi ile belirlenmiştir. Çalışma kapsamında veri ön işleme, yapısal olarak istatistiksel puanlandırma, GAA analiz ve melez cümle seçimi aşamaları Türkçe yazılmış metinler üzerinde başarıyla gerçekleştirilmiştir. Elde edilen sonuçların değerlendirilmesi için özetleme sistemi geliştirilmiştir. Geliştirilen bu sisteme ve farklı kullanıcılara aynı metinler verilmiş ve kullanıcıların önerileri karşılaştırılmıştır. Bu karşılaştırma sonucunda "Güneş Sistemi" metninin özeti ile kullanıcıların bu metinden seçtiği cümleler %77.5, "Charles Bukowski" metninin özeti ile kullanıcıların bu metinden seçtiği cümleler % 82 oranında eşleşmiştir.

Kakisim vd. (2016), twitter sosyal ağında aktif olan bilgisayar korsanlarının ve takipçilerinin paylaşımlarını analiz ederek şüpheli gruba ait yaygın etiketleri tespit eden bir model geliştirmiştir. Şüpheli gruba ait yaygın etiketlerin çıkarılması için Terim Frekansı-Ters Doküman Frekansı (TF-TDF), Twitter' ın yapısında olan favori ve re-tweet sayısı ile yeniden yorumlanmıştır. Kelimenin grup içerisindeki önemini öğrenmek için favori ve re-tweet sayısı, kelimenin TF-TDF değeri ile çarpılmıştır. Elde edilen öznitelik uzayı hedef bilgisayar korsanlarına benzeyen en kuvvetli şüpheli adayların bulunması için kullanılmıştır. Sonuçlar, geliştirilen model ile tespit edilen şüpheli profillerin Twitter tarafından mahkeme kararları ile kapatıldığını göstermektedir.

Ekinci ve Omurca (2016), kullanıcı yorumlarından ürün özelliklerini çıkarmada en popüler konu modelleme yöntemlerinden biri olan GDA kullanılmıştır. Türkçe otel yorumları üzerinden elde edilen deneysel sonuçlar, LDA' nın özellik çıkarmada başarılı olduğunu göstermiştir.

Pavlinek ve Podgorelec (2017), metin sınıflandırma için GDA ile kendisini eğitebilen bir model tasarlamışlardır. Metin sınıflandırması için "seLDA" metodunu, konu modellerine dayalı temsillerle yarı-denetimli bir şekilde göstermişlerdir. Kendi kendine eğitim, etiketsiz verilerin bilgilerini kullanarak küçük ilk etiketli kümeyi büyütmek için kullanılmıştır. Genişletilmiş bir etiketli kümede "Naïve Bayes Multinomial" (NBMN) ve KDM sınıflandırma algoritmaları uygulayarak konu bazlı sunumun tahmin doğruluğunu nasıl etkilediğini araştırmışlar. Çıkan sonuçlar, TF-TDF yöntemiyle oluşan sonuçlarla karşılaştırılmıştır. Sonuçlar, seLDA yönteminin, NBMN ile birlikte kullanıldığında, diğer benzer yöntem ve varyasyonlara göre sınıflandırma doğruluğu açısından önemli ölçüde daha iyi performans gösterdiği sonucu elde edilmiştir.

Tsumoto vd. (2017), hastalıkların teşhisi için metin madenciliği yöntemini kullanmışlardır. Çalışma hastaların elektronik hasta kayıtlarında bulunan şikayetler, fiziksel bulgular, laboratuvar sonuçları vb. yaklaşık 200 GB boyutunda veri üzerinden yapılmıştır. Toplanan veriler üzerinde metin madenciliği işlemi başlıca 4 aşamada gerçekleştirilmiştir. İlk olarak morfolojik analizi yapılarak terimler matrisi elde edilmiştir. Daha sonra benzerlik analizi ile etiketler

sınıflandırılmış ve başlıca anahtar kelimeler oluşturulmuştur. Sıralamalarına göre anahtar kelimeler seçilmiş, eğitim için örnekler oluşturulmuştur. Son aşamada eğitim örneklerine, "Darch Derin Öğrenme", KDM, Yapay Sinir Ağları (YSA) ve Karar Ağaçları gibi öğrenme metotları uygulanmıştır. Uygulama sonunda Darch Derin Öğrenme metotunun diğer metotlara göre hastalıkların teşhisinde daha doğru sonuçlar verdiği görülmüştür.

Tonon vd. (2017), İsviçre Silahlı Kuvvetleri' nin Ar-Ge kurumu olan "Armasuisse Science and Technology" olarak, Twitter verilerini analiz ederek doğal afetler ve terörist faaliyetler gibi olayları tespit etmek için bir Sosyal Medya Analizi sistemi geliştirmişlerdir. Geliştirilen sistemde toplanan twitter paylaşımlarından DDİ yöntemleri kullanılarak kelimeler öğelerine ayrılmaktadır. Öğelerden yüklem "WORDNET", kişi, nesne ve yer verileri DBPEDİA ile eşleştirilerek Kaynak Tanımlama Çerçevesi (KTÇ) modeline dönüştürülmektedir. Güvenlik analistlerine verileri grafik olarak yorumlayabilme olanağı tanımaktadır. Yaklaşım özel olayları tespit etmede başarılı olmuştur. DBPedia için oluşturulan SPARQL sorgularının manuel olarak oluşturulması sistemin zorluklarından biri olarak gösterilmektedir.

Guo vd. (2017), GDA tekniği kullanarak müşterilerin çevrimiçi olarak yaptığı değerlendirme ve derecelendirme bilgilerinin memnuniyet analizini yapmıştır. Yapılan çalışmada kullanılan veri seti 16 ülkede bulunan 25.670 otel için 266,544 çevrimiçi yorum içermektedir. Toplanan tüm otel incelemelerinde müşteri memnuniyetinin boyutlarını çıkartmak ve etiketlemek için GDA kullanılmıştır. GDA tarafından 30 konu tanımlanmıştır ve her bir konu içinde en iyi 20 kelime ve bu kelimelerin ilgili konudaki ağırlıkları gösterilmektedir. Bu konular içerisinde oda durumu ve hizmet kalitesi en önemli boyut olarak bulunmuştur.

Kherwa ve Bansal (2017), GAA üzerine bir çalışma yapmışlardır. Çeşitli DDİ uygulamalarının bilimsel yayınlarından elde edilen bir veri setinde terimlerin birbirleri ile ilişkilerini bulmak için GAA yöntemini kullanmışlardır. Çalışma sonunda GAA, Tekil Değer Ayrıştırma (TDA) ile aynı anlamlı ile birden fazla terimi azalttığını ve birden çok anlamı olan terimleri tanımlayabileceğini ve düşük boyutlu kavramsal alandaki belgeleri temsil ettiğini göstermektedir.

Xie vd. (2018), e-sigara kullanımın zararları üzerine bir çalışma yapmışlardır. Çalışmada e-sigara forumu olan "e-cigarette-forum.com" adresinden 2008 ve 2015 yılları arasındaki, farklı 64 alt forumdan 197106 kullanıcının 6.054.832 gönderisini toplamışlardır. Sosyal medyadan e-sigara hakkında güvenilir bilgi elde etmek için derin yapay sinir ağı modeli tasarlanması amaçlanmıştır. Oluşturulan modele göre 1591 benzersiz zarar ve 9930 e-sigara bileşeni tespit edilmiştir.

Wang vd. (2018), iki rakip ürün için internet üzerinden yapılan yorumları GDA tekniği ile inceleyerek analizini yapmıştır. İki ürünün avantajlı ve dezavantajlı yanları analiz edilmiştir. Veriler amazon.com adresinden toplanmıştır. İki farklı markanın kablosuz bilgisayar faresi ürününe yapılan yorumlar üzerinde çalışma yapılmıştır. Toplanan veriler öncelikle pozitif ve negatif olmak üzere ürünlere verilen puanlara göre 1-2-3 yıldız alan yorumlar negatif olarak, 4-5 yıldız alan yorumlar pozitif olarak sınıflandırılmıştır. Elde edilen veri üzerinde metin madenciliği işlemleri uygulanmış daha sonra GDA tekniği kullanılarak analiz işlemi yapılmıştır. Yorumlar üzerinde uygulanan GDA tekniğinden elde edilen başlıklar birbirine benzeyen, birbirinden farklılık gösteren olmak üzere sınıflandırılmış ve farklı firmalara ait iki ürün hakkında yapılan yorumlar listelenmiştir.

Tasar vd. (2018), yaptıkları çalışmada doğal dil sorgusunun hem dilbilimsel hem de anlamsal teknolojilerden yararlanılarak ontoloji sorgu diline çevrilmesi için bir yöntem ile mimari önerilmiştir. Ontolojilerin bilgi kaynağı olarak kullanıldığı önerilen yöntem ve mimaride, yapısal olmayan doğal dil sorgusunun otomatik olarak yapısal bir sorgu dili olan SPARQL' a dönüştürülmesi ve üretilen sorgu ile bağlı veri üzerinden cevap üretilmesi süreçleri açıklanmıştır. Önerilen mimaride açıklanan "Sorgu Anlamsallaştırma" katmanı, bu katmanda tanıtılan işlemler ve doğal dil farkında bir ontoloji geliştirilerek tasarlanan yöntem ile ilgili çalışmalardan farklı bir yaklaşım izlenmiştir. DDİ teknikleri ve anlamsal web teknolojilerinin birleştirilerek tasarlanan bu farklı yaklaşım ile literatüre katkı hedeflenmektedir.

Yıldıztepe ve Uzun (2018), yaptıkları çalışmada, olasılıksal GAA ve GDA yöntemleri üzerine çalışmışlardır. Farklı haber ajanslarında bulunan Türkçe haber metinlerinin anlamsal benzerliklerine göre kümeleme uygulaması oluşturulmuş ve uygulamadan elde edilen sonuçlar incelenmiştir. Elde edilen sonuçlara göre iki yöntemle de aynı konudan bahseden haber metinleri başarılı bir şekilde sınıflandırılmış ve anlamsal olarak yakın haberler belirlenmiştir.

Merchant ve Pande (2018), DDİ tekniklerini ve GAA algoritmasını kullanarak uzun metinlerden kısa ve faydalı özetler çıkarmak için dava dosyaları üzerinde çalışma yapmışlardır. Yapılan çalışmada ceza ve hukuk mahkemeleri türünde dosyaları incelenmiştir. Oluşturulan modelin ROGUE-1 skoru 0,58 oranına ulaşmıştır. Yapılan çalışma İngilizce metinler üzerinde uygulanmıştır.

Şenel vd. (2019), çalışmalarında ile diyalog tabanlı Türkçe metinler içerisinde konu degişimini otomatik olarak algılayabilen sınıflandırıcılar geliştirilmiştir. Bu sınıflandırıcıların geliştirilebilmesi için öncelikle Türkçe forumlardan konu tabanlı karşılıklı konuşma verileri tasnif edilerek ham bir veri kümesi elde edilmiştir. Oluşturulan veri kümesi üzerinde klasik bir yöntem TD-TDF ile bir derin öğrenme modeli, otomatik konu değişimi tespiti problemi için karşılaştırılmıştır. Klasik yöntem ile test kümesinde %80' lere varan başarı elde edilirken, derin öğrenme yönteminin performansının %76 seviyesinde kaldığı gözlenmiştir.

Yıldız ve Fındık (2019), yaptıkları çalışmada Türkçe soru benzerliği tespiti için çeşitli anlamsal metin benzerliği yöntemlerinin analizini yapmıştır. Bir veri setinde verilen soruya benzer anlama sahip soruları tespit etmek için elle çıkarılmış öznitelikler ve yinelemeli YSA incelemişlerdir. Çıkarımı yapılmış özniteliklerin ve sinirsel yöntemlerin performansını karşılaştırmak için farklı denemeler yapmışlardır. Deneme sonuçlarına göre yinelemeli ağlar, sözcük ve sözcük kökü eşleşme sayıları, TF-TDF vektörleri ve sözcük vektörlerinin benzerlikleri gibi özniteliklere dayanan geleneksel yöntemlerden kayda değer oranda daha başarılı sonuçlara ulaşmaktadır. Yinelemeli ağların başarımının, özniteliklerin sürece dâhil edilmesiyle daha da geliştirilebileceği deney sonuçlarıyla ortaya çıkmıştır.

Ekinci vd. (2020), Türkiye' deki araştırmacılar tarafından yayınlanmış tıp makalelerinin otomatik ve anlamsal analizini gerçekleştiren bir konu modelleme yöntemi olan GDA uygulamıştır. Deneysel çalışma, yıllara göre bir tıp veritabanı olan PubMed' den elde edilen son 11 yıldaki tıp literatüründeki makaleler üzerinde gerçekleşmiştir. Deneysel sonuçlar incelendiğinde, son 11 yılda trend olan çalışma başlıklarının başarılı bir şekilde keşfedildiği gözlenmiştir.

Ulaş ve Karabay (2020), 1970-2017 tarihleri arasındaki farklı haber kaynaklarından oluşan terör verilerinin bir araya gelerek oluşturduğu Global Terör Veritabanı isimli veri kümesini incelemişlerdir. Terör olaylarının büyük veri çerçevesinde makine öğrenmesi teknikleri ile analizi ve sınıflandırması işlemi yapılmıştır. Bir terör olayında saldırının tipi, saldırı yapılan ülke, bölge, saldırının hedef kitlesi ve kullanılan silah türü gibi özellikler ele alınarak tahmin edilmede kullanılmıştır. Çalışmada, Apache Spark ve Python programlama dili kullanılmıştır. Veri kümesi içeriğinde bulunan en çok saldırı gerçekleştiren ilk 10 terör örgütü ele alınmış, altı farklı sınıflandırma algoritması uygulanmıştır. Uygulanan algoritmalar arasından en yüksek ağırlıklı doğruluk oranı olarak k-En Yakın Komşu algoritması %98,2 ile en yüksek değer bulunmuştur.

3. SİSTEMİN GENEL YAPISI VE VERİ TOPLAMA

Bu bölümde, çalışmada genel olarak üzerinde çalışılan Büyük Veri, DDİ konuları ile sistem mimarisine genel bir bakış yapıldıktan sonra veri toplama konusu ve veri ön işlemenin detayları anlatılacaktır.

3.1. Büyük Veri

Günümüzde bilginin gücü, teknolojinin ilerlemesi ve internetin gelişmesi ile beraber daha önemli hale gelmiştir. Bununla birlikte, internet, web ortamında paylaşılan her bilginin kullanılabilir olmamasından dolayı "Bilgi Çöplüğü" olarak anılmaya başlanmıştır. Bu derece fazla bilginin olduğu çöplükten anlamlı bilgiler elde edilebilecek olması "Büyük Veri" kavramını ortaya çıkarmaktadır. Bu verilere, web sunucu ve istemci günlükleri, sosyal medya paylaşımları, algılayıcalardan elde edilen veriler, ses ve görüntü aktarımları, istatistiki bilgiler, haberler, log dosyaları, arşiv sistemleri gibi veriler örnek olarak gösterilebilir. Büyük Veri, bu farklı kaynaklardan elde ettiğimiz tüm bu verilerin anlamlı ve işlenebilir hale dönüştürülmüş biçimidir. Başlangıçta bir bilgisayar ile işlenemeyen veri anlamına gelen Büyük Veri, günümüzde veri analizi veya görselleştirme ile ilgili her şeyi öne çıkarmak için kullanılan bir terim olmuştur (NIST, 2015).

İlişkisel veri tabanlarında veriler yapısal bir şekilde tutularak veri tabanı uzmanları tarafından analizler yapılabilmektedir. Fakat ilişkisel veri tabanlarından tutulamayan birçok veri kümesi bulunmaktadır. Bilgi çöplüğü olarak isimlendirilen bu verilerin bir veri tabanında tutulması ve analiz edilerek raporlanması işlemi çok zordur. Günümüze kadar bir firmanın kendi müşterisine ait bir veriyi, veri tabanında tutarak gerekli analizler yapılabilmekteydi. Fakat günümüz şartlarında ilgili firmanın müşterilerinin sosyal medya ortamlarında paylaştığı veriler, kullandığı akıllı cihazlarda bulanan algılayıcılardan elde edilen veriler önem arz etmeye başladı. Bu yüzden bu verilerden işimize yarayacak olan yararlı verileri çıkarabilmek önemli hale gelmiştir. Sürekli büyüyen fotoğraf algılayıcı bilgisi, metin, video, ses bilgileri kullanarak devletler, firmalar vb. kendilerine belirli bir strateji belirlemektedir. Son zamanlarda ifade edildiği gibi

veriyi, petrolden daha değerli hale getirmektedir. Analitik ve depolama alanındaki gelişmeler, farklı türden birçok veriye ulaşabildiğimiz, saklayabildiğimiz ve üzerinde çalışabileceğimiz şekle dönüşmüştür.

3.1.1 Büyük verinin bileşenleri

Büyük Verinin 5 temel karakteristik özelliği vardır. Bu bileşenler sırasıyla; çeşitlilik (variety), hız (velocity), hacim (volume), doğrulama (verification) ve değer (value) olarak bilinmektedir. Genel olarak 5V şeklinde adlandırılmaktadır. Bir V daha eklenerek değişkenlik (variability) ile 6V olarak adlandırılmaktadır (Demchenko, 2014).

Çeşitlilik (Variety): Üretilen veriler genel olarak yapısal olmadığı ve birçok farklı ortamdan elde edilen veri formatlarından oluştukları için bütünleşik ve birbirlerine dönüştürülebiliyor olmaları gerekmektedir.

Hız (Velocity): Hızlı büyüyen veri, o veriye gerekli işlem sayısının ve çeşitliliğinin de aynı hızda artması sonucunu ortaya çıkartmaktadır.

Veri Büyüklüğü (Volume): Büyük veri her geçen gün hızına hız katarak artmaktadır.

Doğrulama (Verification): Hızla büyüyen verilerin akışı sırasında gelen verilerin güvenli olup olmadığını kontrol etmemiz gerektiği durumlarda da bir diğer veri bileşeni olarak görülmektedir.

Değer (Value): Veriler yukarıdaki veri bileşenlerinden filtrelendikten sonra büyük verinin üretimi ve işlenmesi katmanlarında elde edilen verilerin kuruluşlar için anlamlı sonuçlar vermesi gerekmektedir.

3.1.2. Büyük verinin kullanım alanları

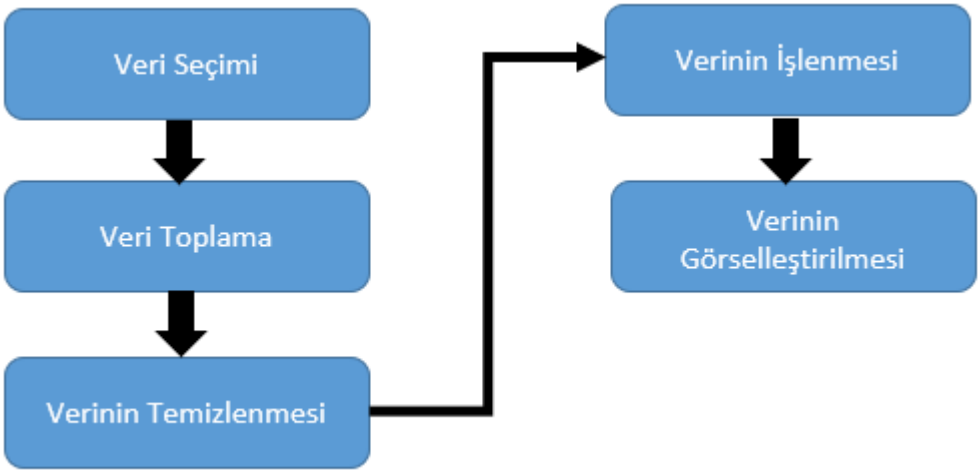
Doğru metot ve teknikler ile uygulandığında firmalar ve devletlerin doğru karar vermelerine, risk analizinin daha düzenli yapılabilmesine büyük veri olanak sağlamaktadır. Büyük veri analizi için birçok yöntem, araç ve teknik geliştirilmiştir. Geliştirilen yöntem ve teknikler kullanılarak firmalar ve

işletmeler kar oranlarının artmasına, hizmet sundukları müşteri kitlesinin hareket ve paylaşımlarına göre tahmin yaparak strateji belirleyebilmektedir. İnternet kullanımının son dönemlerde artması ile beraber kişiye özel hizmet sunabilmek için kişilere ait bireysel veriler saklanarak işlenebilmektedir.

Sosyal medya paylaşımları sayesinde her gün milyarlarca kilobayt veri elde edilmektedir. Twitter' da günlük işlenen veri boyutu: 7 Terabyte' a ulamış durumdadır. Sağlık Kuruluşları; sağlık hizmetlerini bireye özel hale getirebilmek için bireysel verileri saklamaktadır. Bankalar, müşterileri ile ilgili sakladıkları bilgiler aracılığı ile kullanıcılarını tanıyan ve internet şubesinde o gün için hangi hizmeti aldığını bilen aynı zamanda ana sayfayı, menüleri en etkin hale getiren (kişiselleştirme uygulamaları), müşterilerine hatırlatmalar yapan, kişiselleştirilmiş ara yüz deneyimi, zengin içerik ve sürekli hizmet sağlayan şube haline gelmektedir. Arama motorları, arka planda büyük veriyi kullanarak en doğru ve en hızlı sonucu milyonlarca sayfa içerisinden getirebilmektedir. Eğitim sistemi verilerinin toplanması ile birlikte büyük veri sayesinde öğrenciler üzerinde olumlu gelişmeler oluşturabilmektedir. Eğitim verilerinin toplanmasıyla elde edilen büyük veri analiz edilerek sistemin gerisinde olan öğrenciler bulunabilmektedir. Bu öğrencilere kişiye özgü eğitim imkânı sunulabilmektedir. Enerji firmaları, abonelerinin tüketim verilerini depolayarak elde ettikleri büyük veriden anlamlı sonuçlar çıkararak abonelerine kişiye özel tarifeler önerebilmektedir. Otobüs, havayolu, tren ve gemi firmaları yolcuların bilgilerini toplayarak müşteri ile ilişkilerini geliştirebilmektedir. Havayolu şirketleri müşterisi ile olan ilişkisini ilerletmek ve onları daha özel hissettirmek için büyük veriden yararlanmaktadır.

3.1.3. Büyük veri analiz aşamaları

Farklı kaynaklardan elde edilen yapılandırılmamış verilerin yani oluşan büyük verinin analizinin yapılması için belirli işlem basamakları vardır. Büyük veriden anlamlı sonuçlar çıkarmak için analiz süreci büyük önem taşımaktadır. Büyük veri analiz adımları Şekil 3.1' de gösterilmektedir.



Şekil 3.1. Büyük veri analiz aşamaları

3.1.3.1. Veri seçimi

Çalışılmak istenen konu belirlendikten sonra verilerin hangi kaynaktan çekileceğinin belirlendiği aşamadır. Hangi veri türü ile çalışılmak isteniyorsa bu aşamada belirlenmeli ve ilgili veri toplama aşamasına geçilmelidir.

3.1.3.2. Veri toplama

Verilerin toplanması ve kaydedilmesi oldukça önemli bir süreçtir. Günümüzde veri sayısının çok fazla olması, her verinin alınarak kullanılacağı anlamına gelmemektedir. Veri toplama işlemi bilgisayarlar, çevrimiçi kaynaklar, kameralar, çevre kaynakları veya personel gibi çeşitli kaynaklar aracılığıyla yapılabilir. Elde edilen veriler hassas ve birbirleriyle tutarlı olmalıdır.

3.1.3.3. Veri temizleme

Verileri temizleme, düzenleme ve gereksiz bilgileri, mevcut verilerden çıkarma süreci oldukça önemli ve dikkat edilmesi gereken bir süreçtir. Toplanan veya kaydedilen veriler eksik veya yanlış kaydedilmiş olabilir. Toplanan verilerde çok miktarda tekrarlanan veriler olabilir ve bu veriler analizde yanlış sonuçlar ortaya çıkartabilir.

3.1.3.4. Verinin işlenmesi

Veri analizinin önceki süreçleri, veriyi analiz için hazır hale getirmek amacıyla gerçekleştirilen basamaklardır. Çalışmada öncelikle veriler toplanarak temizlenmiş ve analiz için hazır hale gelmiştir. Veri analizi işlemleri sınıflandırma metotları, örüntü tanıma, makine öğrenmesi, yapay sinir ağları gibi çeşitli yöntemler kullanılarak gerçekleştirilmektedir. Veriler belirli yöntem ve tekniklerle filtrelendikten sonra anlamlı hale gelebilmektedir.

3.2. Doğal Dil İşleme

Büyük veri analizi üzerine çalışan firma ve kuruluşlar için toplanan metinlerin analiz edilmesinde DDİ önemli bir aşamadır. Veri miktarı, web teknolojilerinin gelişmesi değişik alanlarda aktif olarak kullanılması ile giderek artmaktadır. Web teknolojilerinde yaşanan bu gelişmeler ile birlikte DDİ öne çıkan çalışma alanları arasında gösterilmektedir. Elektronik ortamdaki dokümanlar, kullanıcı geri bildirimleri ve Twitter, Facebook gibi sosyal medya platformlarının sağladığı veriler/yorumlar ile DDİ' ye yeni uygulama alanları katılmıştır. Verinin içerisinden bilgisayarların anlayabileceği anlamların çıkarılması için DDİ teknikleri kullanılmaktadır. DDİ, uzun yıllardır üzerinde çalışılan Türkçe, İngilizce vb. insanlar tarafından kullanılan doğal dillerde bulunan seslerin ve metinlerin bilgisayar ortamına aktarılması işlemidir. 1950' li yıllarda bu alan yapay zekânın bir alt başlığı olarak görülmekte iken, günümüzde yapılan çalışmalar sonucunda bilgisayar bilimlerinin temel bir çalışma alanı olarak kabul edilmektedir (Adalı, 2012).

Metin madenciliğinin bir parçası olan DDİ, bilgisayarlar ve insan (doğal) dilleri arasındaki etkileşimlerle ilgili bilgisayar bilimi ve yapay zekâ alanıdır. Metinlere ve konuşmaya makine öğrenmesi algoritmalarını uygulamak için kullanılmaktadır. DDİ, dilbilimsel analiz için bilgisayarın insan dilini anlamasına yardımcı olmaktadır. DDİ, insan dillerinin niteliğinden dolayı zor bir problemdir. Doğal dil ile aktarılan bilgileri, bilgisayarın anlaması kolay değildir. İnsan dilini kapsamlı bir şekilde anlamak hem kelimelerin hem de kavramların amaçlanan mesajı iletmek için nasıl birbiri ile kullanıldığını anlamayı gerektirmektedir. DDİ, yapılandırılmamış doğal dil verilerinin bilgisayarların anlayabileceği bir formata

dönüştürülmesi için doğal dil kurallarını tanımlamak ve çıkarmak için algoritmaları kullanmaktadır. Bilgisayarlar, cümlenin anlamını tam olarak anlayamaz ise istenilen doğrulukta sonuçlara ulaşmakta sorun olabilir. DDİ, Google Translate gibi çeviri uygulamalarında, kelime işlemci programlarında metin hatalarının düzeltilmesi, çağrı merkezlerinde müşteri taleplerinin alınmasında, OK Google, Siri, Cortana ve Alexa gibi kişisel asistan uygulamalarında yaygın olarak kullanılmaktadır.

Doğal dil insanların birbirleri ile iletişim kurmasını sağlayan temel özelliklerinden biridir. İletişim için doğal dil kaçınılmaz bir ögedir. Doğal dilin yanında metin belgeleri, eposta, reklamlar vb. farklı türde dil karşımıza çıkmaktadır. Doğal dil öğrenilmesi zor bir süreçtir. Gündelik olarak kullandığımız dilde dahi birçok terim ve kelime günden güne değişmektedir veya yeni bir ifade tarzı ile karşımıza çıkmaktadır. Doğal dilde anlama ve konuşabilme beyin içerisinde karmaşık bir yapıdadır. DDİ, istatistiksel olarak daha baskın olması sebebiyle klasik dil biliminden ayrılmaktadır. DDİ bazı kaynaklarda hesaplamalı dil bilim olarak da geçmektedir. Genel olarak dil ile ilgili yararlı işlemler ile ilgilidir. DDİ' nin amacı istatiksel olarak çıkarım yapmaktır. DDİ' nin başlıca çalışma konuları aşağıdaki gibidir.

- Yazım hatalarının düzeltilmesi
- Spam tespiti
- Metin özetleme
- Bilgi çıkarımı
- Metinin anlamını çıkarma
- Bilgisayar ile sesli iletişim
- Konuşmayı metne dönüştürme
- Soru cevaplama
- Çeviri işlemleri
- Varlık tanıma
- Konu modelleme

Uygulamada değişiklikler olsa da DDİ ile ilgili uygulamalar genel olarak aynı adımlardan oluşur. Bu adımları dört başlık altında toplayabiliriz:

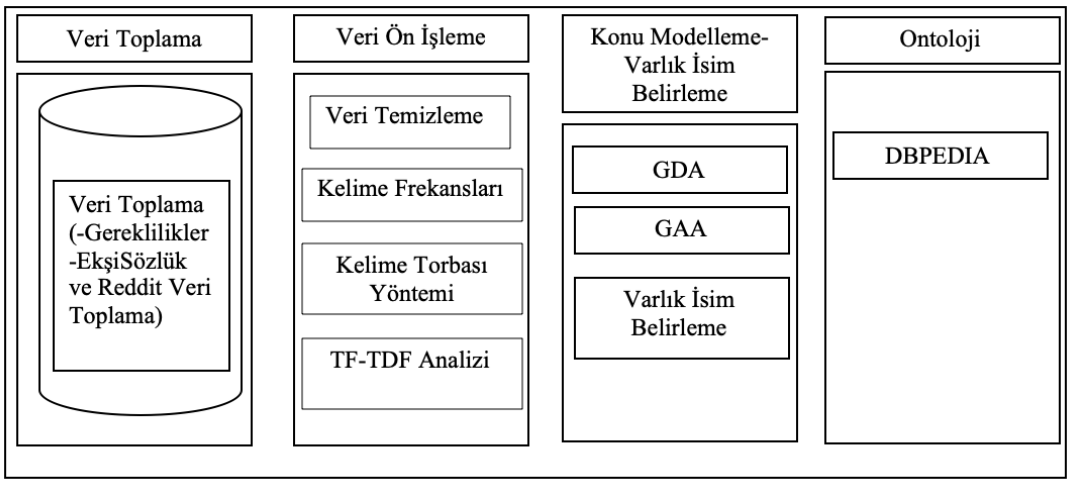
SesBilim (Fonetik): Dil içinde harflerin seslerini ve bu seslerin nasıl kullanıldığını araştırır. Her dilin kendine özgü bir alfabesi bulunmaktadır ve alfabede bulunan tüm harflerin sesleri birbirinden farklıdır. Bu aşamada ulaşılmak istenen amaç konuşulan ifadeleri yazıya dönüştürmektir.

BiçimBirim: Biçimbirim kapsamında her kelime tek başına dilin kurallarına uygun olarak incelenir ve her parçasının çözümlenme işlemi yapılmaktadır. Ekler ve köklerle ilgili kuralların kategorilere ayrılması bu başlıkta incelenir.

SözDizimi: Cümleyi oluşturan kelimelerin sıralanış şekilleri incelenmektedir.

Anlam Bilim (Semantik): Cümlenin genel yapısının bilinmesi sonucu harekete geçme işlemi bu kısımda olmaktadır (Delibaş, 2008).

3.3. Sistem Mimarisi



Şekil 3.2. Sistem mimarisi

Şekil 3.2' de sistem mimarisi genel olarak gösterilmektedir. Sistem genel olarak 4 ana bölümden oluşmaktadır. Bu bölümde veri toplama ve veri ön işleme kısımlarında gerçekleştirilen işlemler anlatılacaktır.

3.3.1. Araçlar ve bağımlılıkları

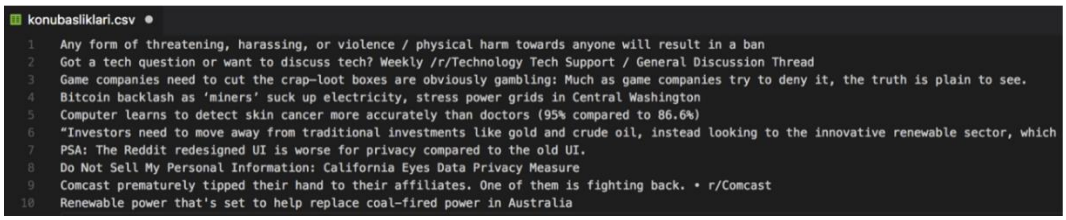
Çalışma Python ile kodlanmışıtr. Python 2.7 ve Python 3.6 sürümleri, Python kurumsal web sayfasından yüklenmiştir (Pyhton, 2001). Python programlama dili not defteri, herhangi bir metin editörü veya farklı platformlar üzerinden kodlanabilmektedir. Bu çalışmada, birden fazla programlama dilinin kodlanmasına, Python modüllerinin eklenmesine olanak sağlayan ücretsiz olarak dağıtımı sağlanan Visual Studio Code platformu kullanılmıştır (Code, 2020).

Visual Studio Code platformu içerisinde pip komutu kullanılarak çalışmayı gerçekleştirmek için gerekli olan Numpy, Pandas, Matplotlib, NLTK, Seaborn, Gensim, Collections, Csv, Praw, Datetime, Json, SPARQLWrapper, Spacy modülleri kurulmuştur.

3.3.2. Veri toplama

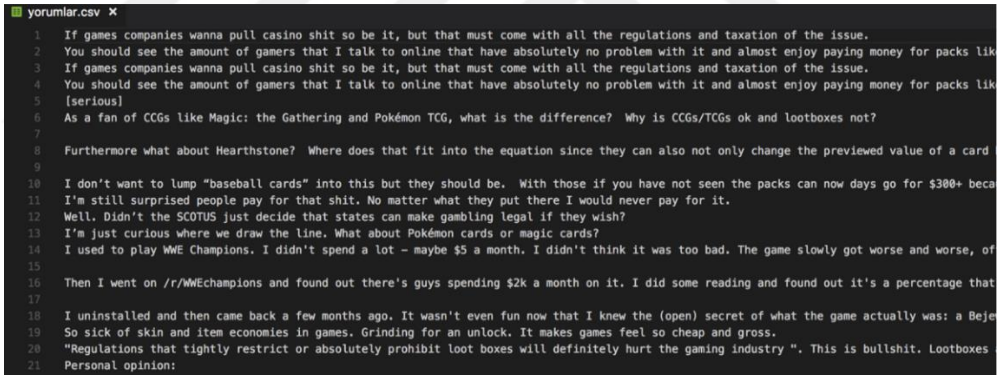
Bu bölümde veri setinin hazırlanması ve veri seti üzerinde yapılan işlemler anlatılmıştır. Tez çalışmasında, farklı dil ve farklı platformlardan veriler toplanarak analizi ve karşılaştırılması yapılmıştır. Dünya çapında yaygın olarak kullanılan, kullanıcıların kendi kimlik bilgilerini gizleyerek veya gizlemeyerek veri girdikleri forum siteleri bulunmaktadır. Çalışmada dünyada popüler olan sosyal medya platformları (Reddit, EkşiSözlük) kullanılacaktır. Farklı platform olarak dünyada yaygın olarak kullanılan Reddit sosyal platformu ve Türkiye' de yaygın olarak kullanılan EkşiSözlük sosyal platformu belirlenmiştir. Bu platformların seçilmesinde kullanıcı sayıları, güncel olarak kullanılmaları, aktif konularda kişilerin herhangi bir karakter sayısı sınırlaması olmadan paylaşım yapmaları göz önünde bulundurulmuştur. Bu iki platformun seçilmesindeki başlıca neden ise EkşiSözlük' te yapılan paylaşımların büyük bir çoğunluğunun Türkçe dilinde, Reddit platformunda ise İngilizce dilinde paylaşımların yapılmasıdır. İki dil ve platform seçildikten sonra, bu platformlarda yapılan paylaşımlar incelenmiştir. Her iki platformda da çok geniş spektrumda paylaşımlar yapılmaktadır. Bu nedenle incelenecek olan konunun sınırlandırılması gerekmektedir. Konu başlığı olarak "Teknoloji" seçilmiştir. Her iki platformda teknoloji alt konu başlığında yapılan yorumlar toplanmıştır.

“Reddit.com” internet sitesi, Alexa istatistiklerine göre dünya genelinde ziyaret edilme sıklığına göre, Mart 2020 itibariyle 18. sırada bulunmaktadır (Alexa, 1996). Reddit forum sitesi “Teknoloji” alt başlığında bulunan açılmış konu başlıklarını çekebilmek için Python dilinde PRAW kütüphanesi kullanılarak script yazılmıştır. Açılan konu başlıklarının verisinin saklandığı *.csv dosyasının ekran görüntüsü Şekil 3.3’ de gösterilmektedir.



Şekil 3.3. Konu başlıklarının saklandığı csv dosyasının ekran görüntüsü

Şekil 3.4’ te Reddit forum sitesi “Teknoloji” alt başlığında bulunan konu baslıklarına ait yorumları gösteren ekran görüntüsü gösterilmektedir.

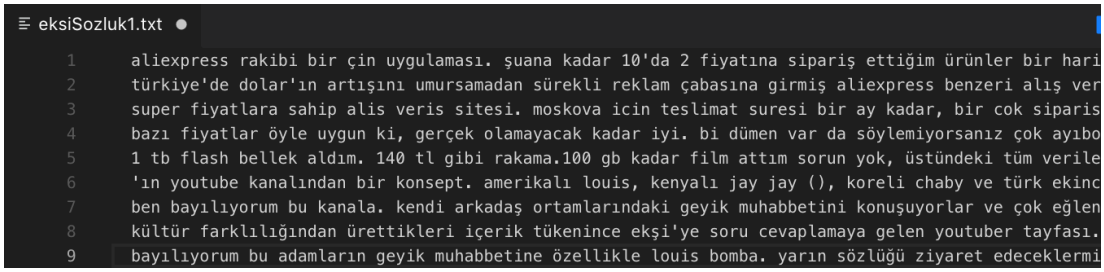


Şekil 3.4. Konu başlıklarına ait yorumları gösteren ekran görüntüsü

EkşiSözlük, her türlü konu ve kavram hakkında, kayıtlı yazarların yorumlarını içeren katılımcı sözlük tarzında bir platform olup, web sitesi Türkiye' deki katılımcı sözlükler arasında en fazla tanımlama (girdi/entry) yapılan sitedir. Kayıtlı yazarlar tarafından yapılan girdiler, paylaşılan bilgiler, yöneticiler ve “gammaz” adı verilen gönüllü kullanıcılar tarafından denetlenmekte uygun olmayanlar silinmektedir. Platformda kayıtlı olan tüm yazarlar gammaz özelliğine sahiptir. Yazar alımı sürekli yapılmamaktadır. Kısa süreli başvurular ile yazar alınmaktadır. Her yazar alınma dönemine “nesil” denilmektedir (Wikipedia, 2001). EkşiSözlük, Alexa istatistiklerine göre dünya genelinde ziyaret

edilme sıklığına göre, Mart 2020 itibariyle 454. sırada, Türkiye genelinde ise 11. sırada bulunmaktadır (Alexa, 1996). Verileri platformdan almak için Python programlama dilinde request (Python Request, 2001) kütüphanesi ile "web crawler" hazırlanmıştır. Hazırlanan web crawler ile yapılan yorumlar toplanarak *.csv uzantılı dosyalarda saklanmıştır. Yapılan yorumlar herkes tarafından okunabilmesi sebebiyle EkşiSözlük verilerinin kullanılması açısından herhangi bir problem bulunmamaktadır. Ayrıca alınan yorumlarla ilgili herhangi bir kişisel veri üzerinde çalışma yapılmamaktadır.

Platformda yapılan yorumlar genellikle Türkçe olarak paylaşılmıştır. Çalışmada seçilen konu kapsamı belirli bir alan olduğu için, konu ile alakasız paylaşım sayısı oldukça azdır. Literatürde Twitter vb. sosyal platformlar üzerine yapılan çalışmalarda karşılaşılan kelimelerin kısaltılması, değiştirilmesi, sadece emoji kullanımı gibi metin analizini zor hale getiren bir durum veri setinde gözlemlenmemiştir. Ekşisözlük platformunda yapılan paylaşımlarda platform kuralları gereği büyük harf olmadığından bütün girdiler küçük harftir. Veri setinde harfleri küçültme ile ilgili bu yüzden herhangi bir işlem yapılmamıştır. Şekil 3.5' te EkşiSözlük' te yapılan yorumları gösteren ekran görüntüsü gösterilmektedir.

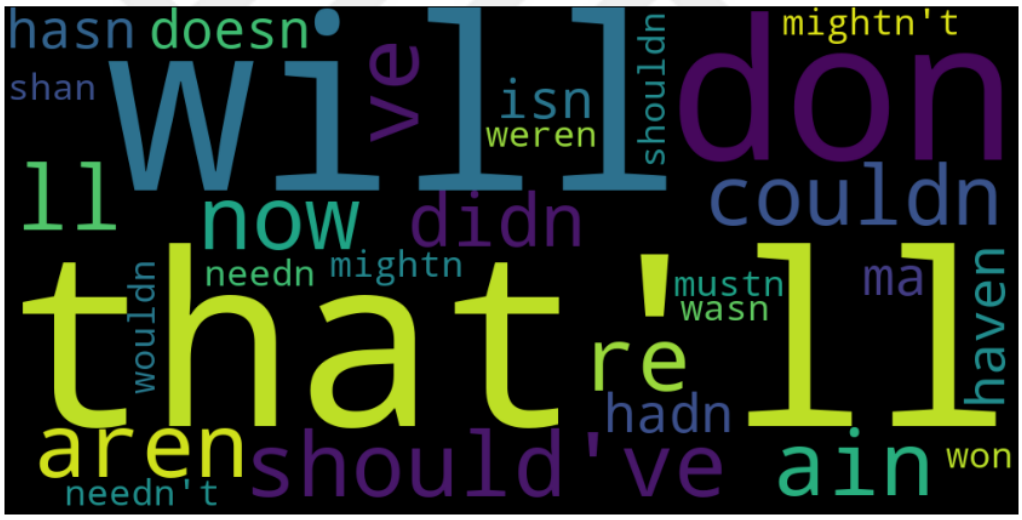


Şekil 3.5. EkşiSözlük yorumlarını gösteren ekran görüntüsü

EkşiSözlük ve Reddit platformlarının teknoloji kanallarından elde edilen Türkçe ve İngilizce içerikli veri öncelikle ön işleme ile parçalara ayrılmıştır (tokenization). Elde edilen kelimeler her iki dilde küçük harfe dönüştürülmüştür. Elde edilen tokenler içerisinden Türkçe ve İngilizce için anlama etki etmeyen durak ifadeler çıkarılmıştır. Türkçe en sık kullanılan kelimeler listesi manuel olarak bir kelime bulutu olarak oluşturulmuştur. Türkçe durak kelimeler listesi Şekil 3.6' da gösterilmektedir. İngilizce durak kelimeler listesi için "StopWords"

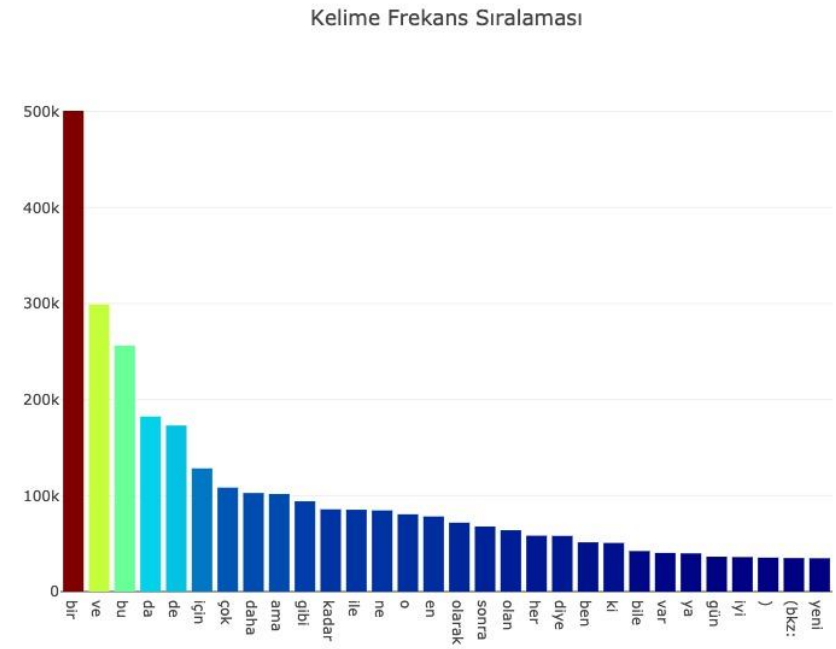
kütüphanesi kullanılmıştır. Bu kütüphanede bulunan kelimelerin listesi Şekil-3.7' de kelime bulutu olarak gösterilmektedir.

Şekil 3.6. Türkçe durak kelimeler listesi

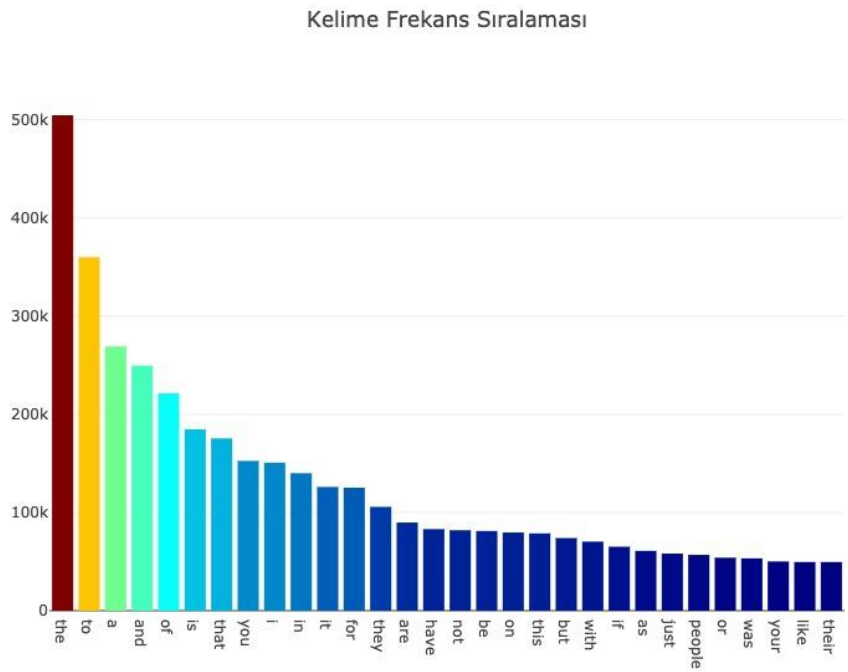


Şekil 3.7. İngilizce durak kelimeler listesi

Verilerin, tokenlara ayrılmasından sonra durak kelimelerinde dahil olduğu Türkçe ve İngilizce kelime sayıları, sırasıyla Şekil 3.8 ve Şekil 3.9' da gösterilmektedir.



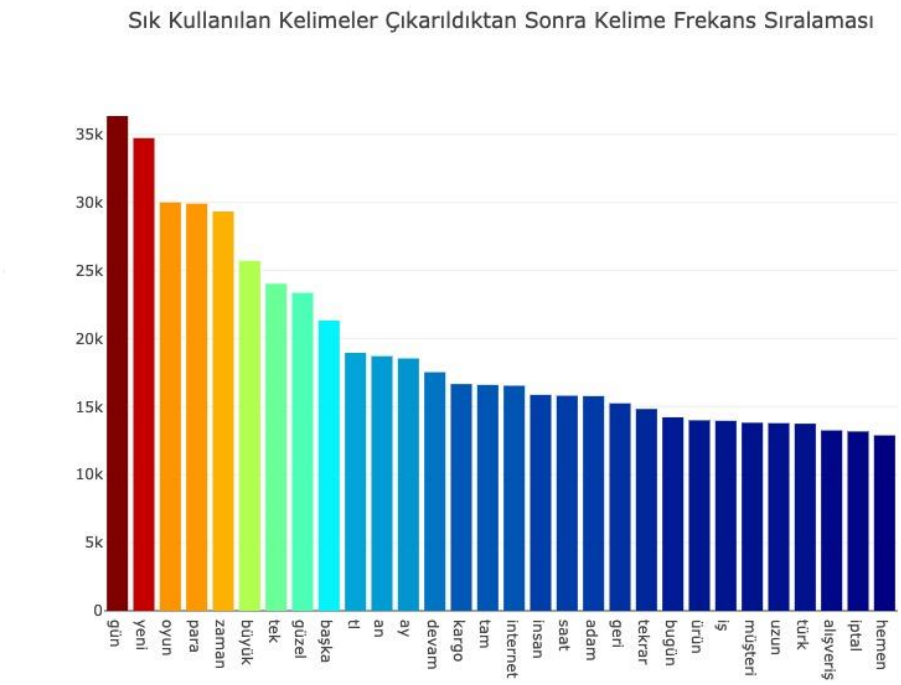
Şekil 3.8. Türkçe kelime frekansları



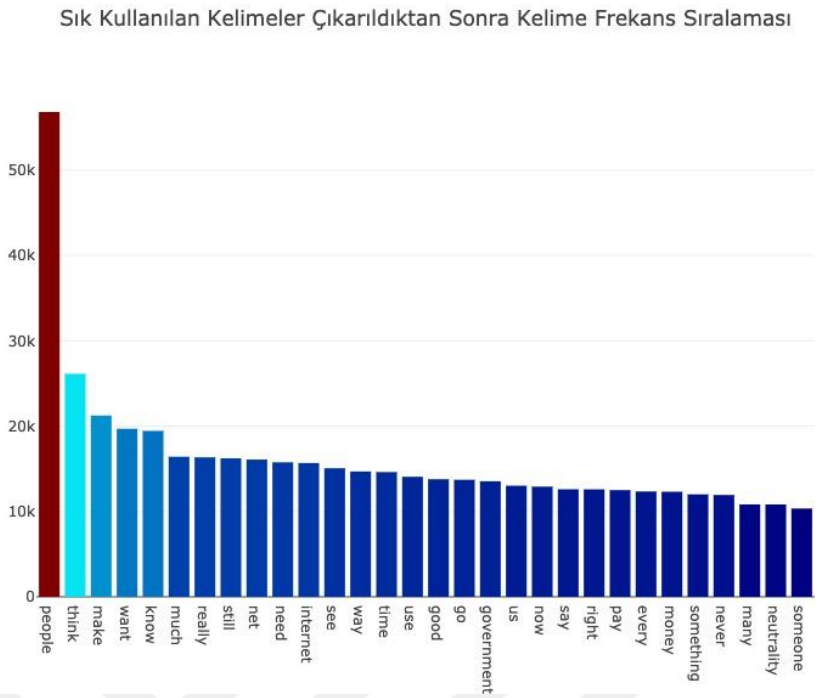
Şekil 3.9. İngilizce kelime frekansları

Şekil 3.8’ de bulunan grafikte görüldüğü gibi veri seti içerisinde en çok kullanılan kelimeler, Türkçe’ de en fazla kullanılan kelimeler seti ile örtüşmektedir. Veri setinde en fazla kullanılan Türkçe kelimeler “bir”, “bu”, “ve” kelimelerinden

oluşmaktadır. Şekil 3.9' da bulunan İngilizce kelime frekansları incelendiğinde "the", "to", "a" ve "and" gibi kelimelerin fazla kullanıldığı görülmektedir. Sık kullanılan kelimeler, Türkçe ve İngilizce veri seti içerisinden çıkarıldığında oluşan kelime sayılarının dağılımları Şekil 3.10 ve Şekil 3.11' da gösterilmektedir. Sık kullanılan kelimeler Türkçe ve İngilizce veri setlerinden temizlendikten sonra kalan kelimelerin tek başlarına kelimeler çıkarılmadan önce elde edilen kelimelere göre daha fazla anlam ifade ettiği görülmektedir. Türkçe durak kelimeler olarak adlandırılan en fazla kullnılan kelimeler çıkarıldığında en çok "gün", "yeni", ve "oyun" kelimelerinin yaklaşık 30.000 civarında kullanıldığı görülmektedir. İngilizce veri setinde ise durak kelimeler çıkarıldığında ise "people" sözcüğün en fazla kullanıldığını bu sözcükten sonra "think" ve "make" sözcüklerinin geldiği görülmektedir.



Şekil 3.10. Durak kelimeler çıktıktan sonra Türkçe kelime frekansları



Şekil 3.11. Durak kelimeler çıktıktan sonra İngilizce kelime frekansları

4. GİZLİ ANLAM ANALİZİ İLE KONU MODELLEME

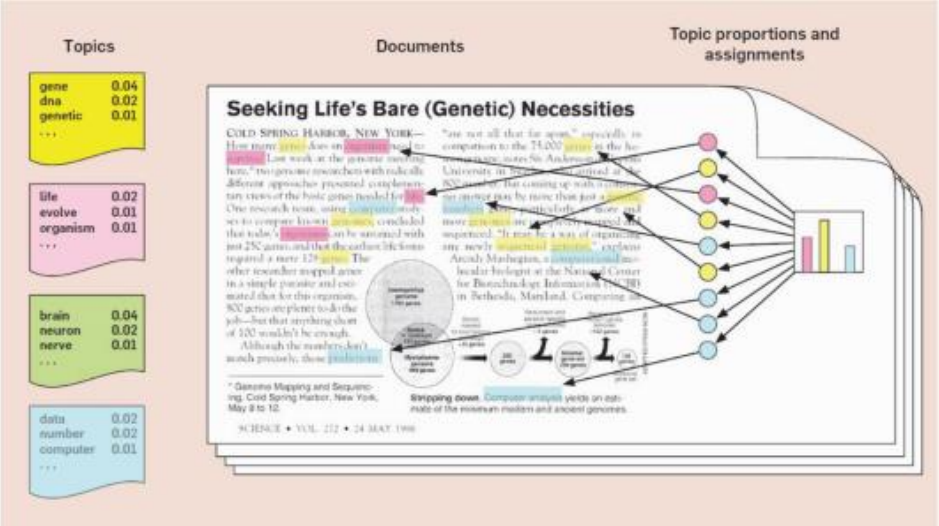
Bu bölümde konu modelleme kavramı, GAA algoritması anlatılmakta ve GAA algoritmasının Türkçe ve İngilizce veri setine uygulanması ile elde edilen sonuçlar analiz edilmektedir.

4.1. Konu Modelleme

Her gün büyük miktarlarda veri toplanmaktadır. Daha fazla veri toplandıkça, aranılan içeriğe erişmek zorlaşmaktadır. Bu nedenle, çok miktarda bilgiyi düzenlemek, aramak ve anlamak için araçlara ve tekniklere ihtiyacımız vardır. Konu modelleme, büyük metinsel bilgi koleksiyonlarını organize etme, anlama ve özetleme olanağı sağlamaktadır.

Makine öğrenmesi ve metin madenciliği uygulamalarında tercih edilen önemli çalışma alanlarından biri doküman koleksiyonlarındaki gizli tematik bilgiyi küçük boyutlu uzaya dönüştürerek ortaya çıkaran algoritmalar olarak bilinen konu modelleme yöntemleridir (Lu vd., 2011). Konu modelleme, etiketsiz halde bulunan dokümanların işlenerek gizli konuların ortaya çıkarılması çalışmasıdır. Konu oluşumunun dokümanda olan kelimelerin kullanımı ile ilişkili olduğu varsayımına göre hareket eder.

Konu modelleme çalışma mantığı Şekil 4.1' de verilmiştir. Makale okunurken karşılaşılan farklı temalara işaret eden anahtar kelimeler, vurgulamak için farklı renkler ile işaretlenir. Makalenin her bir anahtar kelimesini okuduktan ve her bir anahtar kelimeyi vurgulayıcıyla renklendirdikten sonra, ortak bir renkle renklendirilmiş anahtar kelimeler toplanırsa, ortak bir renkle renklendirilmiş her bir anahtar kelime grubunun bir temayı temsil ettiği kabul edilebilir. Farklı renk sayılarının toplamı makalede var olan konu sayısını vermektedir (Bhat vd., 2019). Konu modelleme için GDA, GAA (Landauer ve Dutnais, 1997) ve "Non-negative Matrix Factorization" (NMF) (Lee ve Seung, 2001) yöntemleri ön plana çıkmaktadır (Stevens vd., 2012).



Şekil 4.1. Olasılıksal konu modelleme gösterimi (Blei, 2012)

4.2. Gizli Anlam Analizi

Konu modelleme; verilen dokümanlardan alt konuları otomatik olarak bulmak için kullanılan bir istatiksel makine öğrenmesidir. Bu yöntemle alt konuların önemli özellikleri ve her bir dokümanın hangi alt konuya ait olduğunu bulunabilir. Verilen belgedeki anahtar sözcük grubunu bulmak için kullanılan, denetimsiz öğrenen bir metin analizidir. İşlem sonucu ortaya çıkan kelime grubu (özellikler), alt konuyu temsil etmektedir. Denetimsiz bir öğrenme şekli olduğu için, bulunan konuların bir uzman tarafından değerlendirilmesi gerekebilir. Ayrıca çoğu zaman kaç farklı alt konunun bulunacağı, önceden bilinmesi gerekmektedir.

Konu modellemede kullanılan başlıca modellerden birisi de GAA' dır. GAA; anlaşılması ve uygulanması kolay olan bir yöntemdir. Diğer metotlara göre daha hızlıdır. Çünkü sadece doküman terim matrislerine göre işlem yapmaktadır. GAA için bir doküman-terim matrisine ihtiyaç vardır. Bu matrisin değerleri genel olarak TF-TDF ağırlıkları ile oluşturulur.

TF-TDF, her bir dokümanın içinde yer alan kelimelere birer ağırlık oluşturur. Bu ağırlıklar, kelimelerin o doküman için ne kadar sık geçtiğine ve o kelimenin diğer dokümanlarda ne kadar geçip geçmediğine bakılarak hesaplanır. Bunun için önce Terim Sıklığı (TF) hesaplanır. Bu işlem her bir kelimenin bir doküman içinde kaç

kere geçtiğini hesaplar. Daha sonra Ters Doküman Frekansı (TDF) denklem 4.1' de verilen formülle hesaplanır:

$$IDF(t) = \log(\frac{N}{|\{d \in D: t \in d\}|}) \quad (4.1)$$

t terim (kelime), N doküman sayısı, D tüm dokuman seti, d tek bir dokumanı temsil eder. $|\{d \in D: t \in d\}|$ ifadesi t teriminin, tüm dokümanlarda, kaçının içinde yer aldığını bulur. Eğer bir terim, çok sayıda dokümanda geçiyorsa, payda büyüyecek ve logaritmik ölçekte TDF değeri küçülecektir. Ya da bir kelime az sayıda dokümanda geçiyorsa, o kelime ilgili doküman için ayırt edici ve önemli bir kelime olur. TDF tüm kelimelerin doküman zıtlığıdır. Son olarak, TF ve TDF ağırlıkları çarpılarak, TF-TDF ağırlıkları bulunup TF-TDF matrisi oluşturulur. Doküman sıklığı ve ters doküman sıklığı özelliklerinin çarpımıyla elde edilen TF-TDF matrisinin her bir satırı bir dokümanı, her bir sütunu ise kelimeleri temsil eder.

TF-TDF matrisinden yararlanılarak, Tekil Değer Ayrıştırma (TDA) ile işlemi yapılabilir. Bu ayrıştırma işlemi ile satırlardaki dokümanlar ve sütunlardaki kelimelerin gruplandırılması hedeflenir. TDA matrisi çarpanlarına ayırma çeşitlerinden biridir. TDA, Google PageRank algoritması, insan yüzü modelleme, bilgi çıkarımı, veri sıkıştırma işlemlerinde kullanılan temel bir işlemdir. Bu gruplandırma işlemi yapılırken TDA' nın denklem 4.2' de verilen formül ile elde edilen matrisler yorumlanır:

$$A = U\Sigma V^T \quad (4.2)$$

A, TF-TDF ağırlıklarının olduğu $m \times n$ boyutundaki orijinal matristir. m döküman sayısı, n ise tüm dokümanlardan elde edilen sözcük sayısıdır. U, $m \times m$ boyutunda dik açılı (ortogonal) sol tekil değer matrisidir. Bu matriste dokümanlar ile ilgili ağırlıklar yer almaktadır. Σ matrisi $m \times n$ boyutunda köşegen bir matristir. Köşegende A matrisinin özdeğerleri (eigen values) büyükten küçüğe doğru yer alır. V^T ise $n \times n$ boyutunda dik açılı sağ tekil değer matrisidir. Bu matriste de terimler ile ilgili ağırlıklar yer almaktadır.

Σ matrisinin köşegeninde $\sigma 11 > \sigma 22 > \cdots > \sigma mm$ değerleri yer almaktadır. Belirlenecek bir k sayısı ile bu köşegenin ilk k değeri alınır. Bu değer, kaç farklı konu gösterilmek istendiğidir. Eğer k değerinin ne olacağı bilinmiyorsa, sıralı özdeğerler arasındaki en büyük boşluğa sahip yer k olarak seçilir. Σ matrisinin yeni boyutu $k \times k$ olacaktır. Dolayısıyla U matrisinin ilk k sütununu $m \times k$ ve V^T matrisinin ilk k satırını $k \times n$ seçilmesi gerekir. Bu üç matrisin çarpımı orijinal A matrisine yakınsayacaktır.

U matrisinin ilk sütunu, ilk konunun doküman ağırlıklarını vermektedir. Yani, ilk sütundaki en yüksek değerler, ilk konunun ağırlığı en yüksek dokümanı olacaktır. Aynı şekilde V matrisinin ilk sütunundaki değerler, ilk konunun terim ağırlıklarını gösterecektir. Ağırlıkları yüksek olan kelimeler, o konunun açıklayıcı kelimeleri olacaktır. Bu işleme k. konuya kadar devam edilir. TDA yapıldıktan sonra U ve V matrislerinde negatif değerler yer alacaktır. Ancak A matrisi tamamen pozitif değerlerden oluşmaktadır. U, Σ ve V^Tmatrislerinin çarpımı A'yı vereceği için ve Σ'da negatif değer yer almadığı için U'da negatif bir değer varsa, V^T'de negatif olmak zorundadır. Dolayısıyla negatif değerlerde olsa bile, mutlak değerlerinin alınması sonucu pozitife dönüşmesi U ve V^T analizi için göz önünde bulundurulmalıdır.

4.2. GAA Tekniğinin Veri Setine Uygulanması

Python' da gensim kütüphanesi kullanarak GAA modelinin oluşturulması mümkündür. Gensim kütüphanesi (Gensim, 2009) kullanılarak oluşturulan modelden üretilen EkşiSözlük sosyal platformundan çekilen yorumlar için Türkçe konu başlıkları Çizelge 4.1' de gösterilmektedir. Çizelge 4.1' de görüldüğü gibi EkşiSözlük platformu teknoloji kanalından yapılan kullanıcı paylaşımlarından 3 konu başlığı belirlenmiştir. Belirlenen konu başlıklarına ait kelimeler, ilgili konu başlıklarının sütunlarında görülmektedir. Konu 1 olarak belirtilen kelimelere bakıldığında öncelikle Çin' de başlayan daha sonra bütün dünyayı etkisine altına alan koronavirüs salgını hakkında konuşulduğu görülmektedir. Konu 1 ile ile ilgili yorumlara bakıldığında salgınla 5G teknolojisinin ilgisi olduğuna dair yapılan haberlerle ilgili paylaşımlar yapıldığı gözlemlenmektedir. Konu 2 ile ilgili kelimelere bakıldığında bir oyundan ve

zamandan bahsedildiği görülmektedir. Bu konu ile ilgili paylaşımlar incelendiğinde piyasada var olan çevrimiçi olarak oynanabilen League of Legend oyunu hakkında paylaşımlar başta olmak üzere, farklı oyunlar hakkında konular üzerine yoğunlukla konuşulduğu görülmektedir. Konu başlıkları belirlenirken 3 ana konu olarak belirlenmiş olmasına rağmen bazı konu başlıklarından birden fazla konu hakkında bulgular olabilmektedir. Konu 3 içerisinde bu tarzda bir durum vardır. Farklı iki konu üzerinde bulgular vardır. İlk konu olarak ekonomik konu olarak görebileceğimiz bir dijital para birimi olan Bitcoin üzerine paylaşımlar varken daha sonra sinema filmleri üzerine paylaşımlar görülmektedir.

Çizelge 4.1. GAA algoritması ile üretilen Türkçe konu başlıkları

No	Başlık	Etiket
1	Ölüm, sayısı, vaka, bildirdi, ülkedeki, sayı, fazla, coronavirüs, para, hayatı	Koronavirüs
2	Zaman, fazla, tek, oyun, güzel, yıl, insan, para, doğru, hemen	Oyunlar
3	Bitcoin, para, film, sağ, puan, tarafından, düşman, yapımı, time, seri	Bitcoin

GAA tekniği Reddit platformundan toplanan yorumlar üzerine de uygulanmıştır. Toplanan yorumlar üzerinden elde edilen 3 konu başlığı ve bu konu başlıklarına ait kelimeler Çizelge 4.2' de gösterilmektedir.

Çizelge 4.2. GAA algoritması ile üretilen İngilizce konu başlıkları

No	Başlık	Etiket
1	Lol, people, think, know, make, internet, money, want, really, much	Oyunlar
2	People, think, know, want, make, internet, time, now, really, much	Yasa Teklifi
3	Ads, netflix, people, pay, watch, hulu, content, service, shows, cable	Filmler

Konu 1 olarak konuşulan konu başlığına bakıldığında kişilerin internet üzerinden oynadığı "LOL (League of Legends)" oyunu üzerine başta olmak üzere oyunlar hakkında konuşulduğu görülmektedir. Konu 2 incelendiğinde ise "teknoloji" ile alakalı kelimelerden bazı konu başlıkları çıkmaktadır. Fakat anlamlı konu başlıklarının elde edilebilmesi için kelimelerin yoğun geçtiği paylaşımlar manuel

olarak incelenmiştir. İnceleme sonucunda burada Amerika Hükümeti' nin internet ile ilgili olarak çıkarmak istediği bir yasa teklifine yapılan eleştiriler üzerine paylaşımlar olduğu görülmektedir. Konu 3' ün içerdiği kelimeler incelendiğinde çevrimiçi yayın platformları üzerine konuşulduğu görülmektedir. Çevrimiçi yayın platformu olan Netflix ve Hulu hakkında karşılaştırmaların yapıldığı görülmektedir.

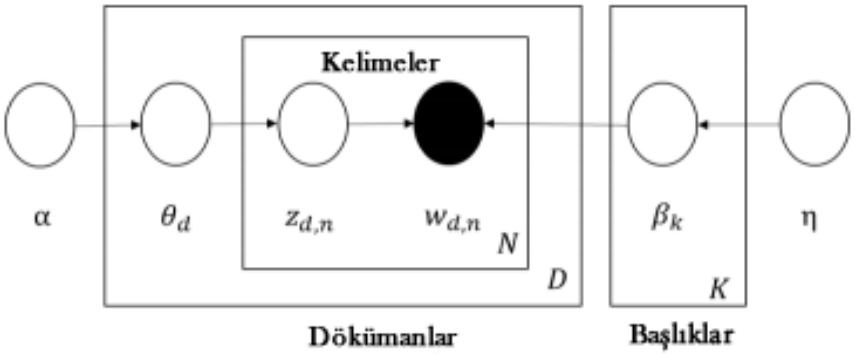
GAA algoritması ile yapılan İngilizce ve Türkçe dillerinin konu modellemeleri sonucu ortaya çıkan konular incelenmiştir. Sonuçlarda Reddit ve EkşiSözlük Sosyal platformlarında teknoloji başlığı üzerine yapılan yorumlardan başlıca 3 konu başlığı elde edilmiştir. Elde edilen konu başlıkları karşılaştırıldığında 3 konu başlığından 2' si birbirine yakın konular olarak görülmektedir. Konulara bakıldığında her iki platfromda "Teknoloji" kanalında benzer konu başlıklarının kullanıcılar tarafından tartışıldığı/konuşulduğu görülmektedir. Farklı konularda ise EkşiSözlük platformunda 5G ve Koronavirüs üzerine paylaşımlar yapılırken, Reddit platformunda ise Amerika' da çıkarılacak olan Yasa Teklifi üzerine paylaşımlar yapıldığı görülmektedir.

5. GİZLİ DIRICHLET AYRIMI

Bu bölümde GDA algoritması hakkında teorik bilgi ve GDA algoritmasının Türkçe ve İngilizce veri setine uygulanması ile elde edilen sonuçlar analiz edilecektir.

5.1. Gizli Dirichlet Ayrımı

Konu modelleme yöntemleri arasında GDA yaygın bir şekilde kullanılmaktadır. GDA, metinsel belge koleksiyonunda bulunan gizli anlamsal yapıları ortaya çıkarmak için olasılıksal bir modelleme yaklaşımı olarak geliştirilmiştir (Blei vd., 2003). Her bir dokuman, her bir konunun doküman koleksiyonundaki benzersiz kelimeler üzerinde dağıtılmasıyla karakterize edilen gizli konuların bir karışımını gösterir. GDA' nın doğrusal gösterimi (plate notasyonu) Şekil 5.1' de gösterilmektedir.



Şekil 5.1. Gizli Dirichlet Ayrımı doğrusal gösterimi

Dokümanda bulunan kelimeler Wd,n ile gösterilmektedir. Gösterimdeki n kelimeyi ($\forall n$ = 1, … , N), d ($\forall$d = 1, … , N) ise dokümanı temsil etmektedir. βk konu başlıklarının içeriğini, θd ise her dokümanın bulunan konu başlıklarına dağılımını göstermektedir. Zd,n ise her kelimenin bağlı olduğu konu başlığını temsil etmektedir. βk, θd, Zd,n önceden bilinmeyen , Wd,n bilinen değişkeninin işlenmesi sonucu elde edilen değerleri temsil etmektedir. η ve α parametreleri ise βk ve θd parametrelerinin önceki değerlerini temsil etmektedir. Plate notasyonunda K kutusu konu başlıklarının sayısını, D kutusu doküman sayısını, N kutusu ise dokümanlarda yer alan benzersiz toplam kelime sayısını belirtmektedir.

Plate notasyonu tekrarlayan yapıları yani aynı tipte birden fazla nesnenin olduğu durumları ifade etmek için kullanılmaktadır. GDA için plate notasyonu ise gözlemlenen verinin rastgele değişkenler ve bu değişkenlerin yönlü kenarlar boyunca yayılımı üzerinden nasıl üretildiğini açıklamaktadır. Konu modellemedeki asıl amaç, doküman koleksiyonundan konuların çıkartılmasıdır. Bunu yaparken elimizde sadece dokümanlar gözlenebilir durumda olup; kelimelerin konulara atanması, konuların dokümandaki ve kelimelerin konulardaki dağılımları gizlidir. Bu nedenle Şekil 5.1' de gözlemlenen değişkenler siyah renkle, gözlenemeyenler beyaz renk ile temsil edilmiştir. Tamamen denetimsiz bir yöntem olan GDA herhangi bir önbilgiye gerek kalmadan, kelime torbası yaklaşımına dayalı olarak çalışmaktadır. Kelimelerin doküman içerisindeki yerleşimi dikkate alınmazken, kelimelerin birlikte bulunması bu yöntemde kullanılmaktadır.

5.2. GDA Algoritmasının Veri Setine Uygulanması

5.2.1. GDA algoritmasının Türkçe veri setine uygulanması

GDA analizi ve veri ön işleme işlemleri Python dilinde yapılmıştır. Veri ön işleme işlemleri için NLTK kütüphanesi, GDA analizi için ise Gensim (Gensim, 2009) kütüphanesi kullanılmıştır. GDA algoritması girdi olarak konu sayısını istemektedir. Literatürde konu sayısını belirleme için çoğunlukla deneme yanılma yoluna başvurulmuştur. Farklı konu sayıları denenerek en anlamlı sonuç veren değer K değeri olarak seçilmiştir. Çalışmamızda, deneme yanılma sonucunda en anlamlı sonuçların görüldüğü 3 değer seçilmiştir.

GDA algoritmasının çıktı olarak verdiği bilgiler dokümanın içerdiği konu başlıkları, her dokümanın konu başlıklarına katkısı ve başlıkların dağılım oranlarıdır. Türkçe yorumlar için konu başlıkları Çizelge 5.1' de gösterilmektedir. Konuların içerdiği kelimelerden anlamlı etiketler çıkarılmaktadır. GDA algoritmasının denetimsiz olarak çalışmasından dolayı konuların otomatik olarak bulunması mümkün değildir. GDA algoritmasından elde edilen başlıkların içerdiği kelimeleri yorumlamak için dışarıdan bir etki gerekmektedir. Çizelge 5.1' de çıkan başlıklara ve ilgili kelimelere bakıldığında 1 nolu konunun ilgili

kelimeler ve katkı yapan yorumlar incelendiğinde çevrimiçi olarak oynanan oyunlar hakkında konuşulduğu görülmektedir. 2 nolu konu başlığında ise son zamanlarda bütün dünyanın uğraştığı korona virüs olduğu görülmektedir. 3 nolu konu başlığında müşteri hizmetleri ile ilgili şikâyetlerden bahsedilmektedir.

Çizelge 5.1. GDA algoritması ile üretilen Türkçe kelimeler ve konu başlıkları

No	Başlık	Etiket
1	oyun, güzel, zaman, tek, adam, para, başka, film, fazla, Türk	Oyunlar
2	virüs, dünya, tarafından, sayısı, fazla, Türk, insan, vaka, yüksek, kişi	Korona Virüs
3	tl, internet, iptal, müşteri, ay, lira, kargo, saat, telefon, sorun	Müşteri Hizmetleri Şikayetleri

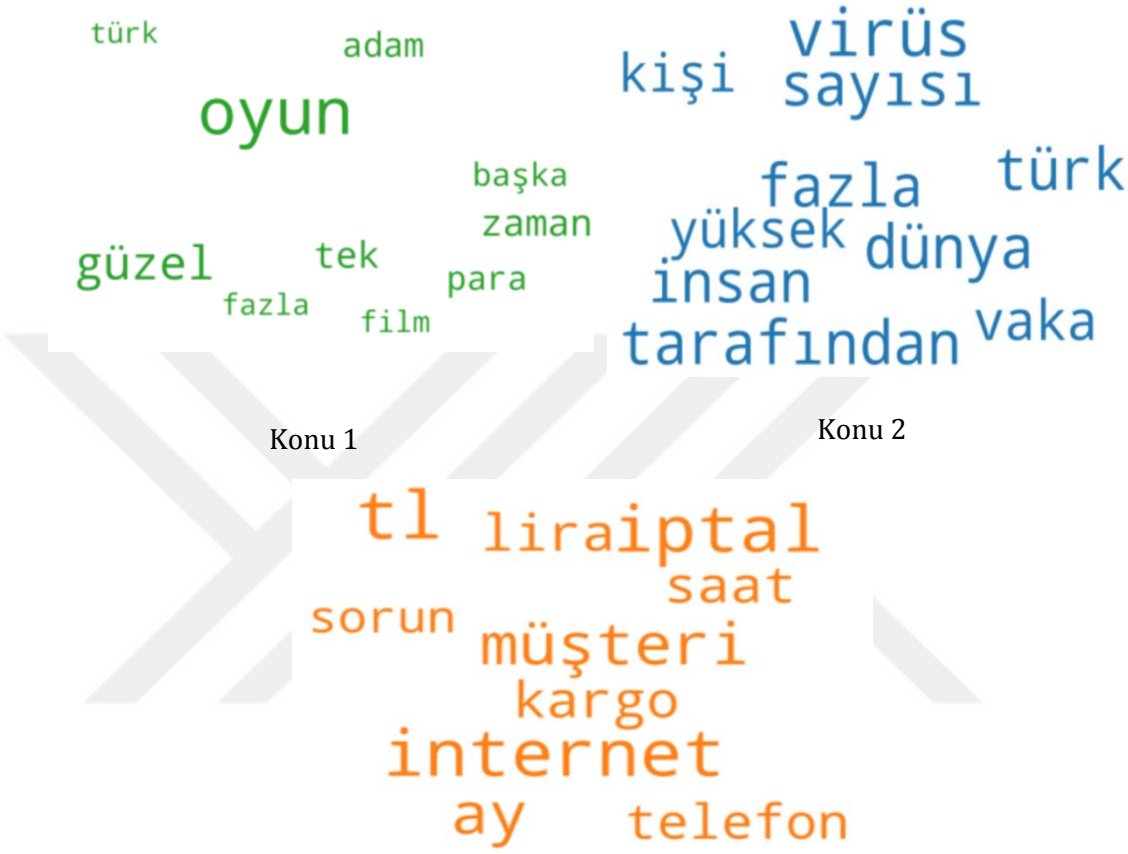
Bu çalışmada 282351 kullanıcı yorumu toplanarak analiz edilmiştir. Bundan dolayı bütün yorumların konu dağılımlarını göstermek imkânsızdır. Her bir konu alt başlığın oluşması esnasında birçok yorum etkin rol oynamaktadır. Bazı yorumlar birden çok başlığa uymaktadır. Her bir başlığın belirlenmesinde en fazla orana sahip olan yorumlar ve başlıklar Çizelge 5.2' de gösterilmektedir.

Çizelge 5.2. Başlıklarda en fazla ağırlığa sahip yorumlar

No	Konuya Uyumluluk	Etiket	Yorum
1	0.9988	Oyunlar	önceki iki oyunu oynamış ve ikinci oyunu ikinci kez bitirdiği gün bu oyuna başlamış birisi olarak özellikle ilk seviyelerde dövüş sisteminde zorlandım. yetenekleriniz gelişene kadar canavar gördünüz mü kaçabilirsiniz yargılamam çünkü assassins of kings ten çıkan ""ooo geralt"","aman geralt"", ""yaman geralt"" ileri geri taklalar atıp vurup kaçmak zorunda kalıyor. drownerlar bile sizi yere seriyor zorunuza gidiyor. bunun dışında oyun zaten harika ilk oyun gibi ikinci oyun gibi. hikayede, questlerde, seçimlerin grinin binbir tonuna ayrılmasında derken bildiğimiz witcher dünyası bildiğimiz geralt. bildiğimiz derken kesinlikle bir aşağı görme ya da beğenmeme durumu yok elbette ben önceden de bayılıyordum şimdi de aynı şekilde bayılıyorum. oyunun başında beni biraz rahatsız eden birkaç noktaya değineyim;--- --- öncelikle zaten ilk oyunda kim kimdir biz neciyiz bilmiyorduk malum hafıza kaybı, ikincide parça parça geldi sonunda gaza geldik ""ooo yennefer'ı bulacağız, wild hunt kork benden"" modunda çıktık oyundan witcher wild hunt'a başladık.
2	0.9945	Korona Virüsü	corona virüsü üst solunum yollarında enfeksiyona yol açabilen viral etkenlerden sadece bir tanesidir. enfeksiyon pek çok bakımdan mevsimsel gripten ayırt edilemez. hastalık daha çok hayvanlarda ortaya çıkar. nadiren hayvanlardan insanlara, insanlardan da diğer insanlara geçiş gösterir. çin'deki bir deniz ürünleri pazarındaki yılanlardan insanlara geçiş yaptığı düşünülmektedir ama insanlara hangi kaynaktan geçiş yaptığına dair varsayımlar iddia olmanın ötesine geçebilmiş değildir.insanlarda ve hayvanlarda hastalığa yol açabilen ve birçok türü bulunan virüslere ""coronavirus"" denilmektedir. insanlarda genellikle soğuk algınlığına yol açan bu virüsler hayat kayıplarına varan tablolara da neden olabilmektedir. 2002 yılında çin'in guangdong eyaleti'nde başlayan sars-cov virüsü salgını dünya genelinde 17 ülkeye yayılmıştır. bu salgında 8098 kişi hastalığa yakalanmış ve 774 kişi hayatını kaybetmiştir. 2012 yılında suudi arabistan'da başlayan mers-cov virüsü salgını ise dünya genelinde 27 ülkeyi etkilemiş, salgına yakalanan 2499 kişiden 861'i hayatını kaybetmiştir. insanlarda ve hayvanlarda hastalığa yol açabilen ve birçok türü bulunan virüslere ""coronavirus"" denilmektedir....
3	0.9991	Müşteri Hizmetleri	hizmet kalitesi ve musteri memnuniyeti konusunda dibe vurduklarini gayet iyi bilen sirket. adima kayitli biri evde adsl biri ofiste fiber olmak uzere kurumsal 2 abonelik var. adsl'i en azindan 8 yildir kullaniyorum, modem falan da bana ait. arada kesilir arar ariza kaydi yaptirirdik cozulurdu. olmadi gorevliler gelir modemden kaynakli ya da telefon hattindan kaynakli falan derlerdi, kullanmaya basladigimdan beri 2 ya da 3 sefer modem degistirdim bu yuzden. uzun yillar cok da memnundum. fiyat politikalarina, akk kalktiktan sonra geldikleri hale deginmeyecegim bile. nisan ortasinda kesinti daha dogrusu kopma problemi yasamaya basladim adsl baglantisinda. yapilacaklari denedik degismedi aradim musteri hizmetlerini. telefondaki gorevli dedi ki 3 gunde 600'den fazla kopma gorunuyor sayi cok fazla ekip yonlendirecegim ariza tespiti icin bakacaklar...

Kelime bulutları, birçok sosyal ağ analitik aracında güçlü bir görselleştirme aracı haline gelmiştir. Genellikle blog yazılarında, belgelerde, sosyal medya konuşmalarında tartışmanın "başlıkları" olarak adlandırılan kelimeleri göstermek için kullanılır. Kelime bulutları, altta yatan tartışma konularını gizleme eğilimindedir ve yalnızca en sık kullanılan sözcüklerin yüzeye çıkmasına izin verir. Şekil 5.2' de her bir konu başlığını oluşturan kelimeler için hazırlanmış olan kelime bulutu gösterilmektedir. Hangi başlıkta hangi kelimenin daha

ağırlıklı olduğunu kelimenin bulut içerisindeki boyutu belirlemektedir. Her konu başlığı farklı bir renkte gösterilmektedir. Konu başlıklarında baskın bulunan kelimeler daha büyük boyutta gösterilmektedir. Kelimelerin ilgili konu başlığı üzerinde ağırlıkları düştükçe kelime bulutu üzerinde gösterilen boyutları azalmaktadır.



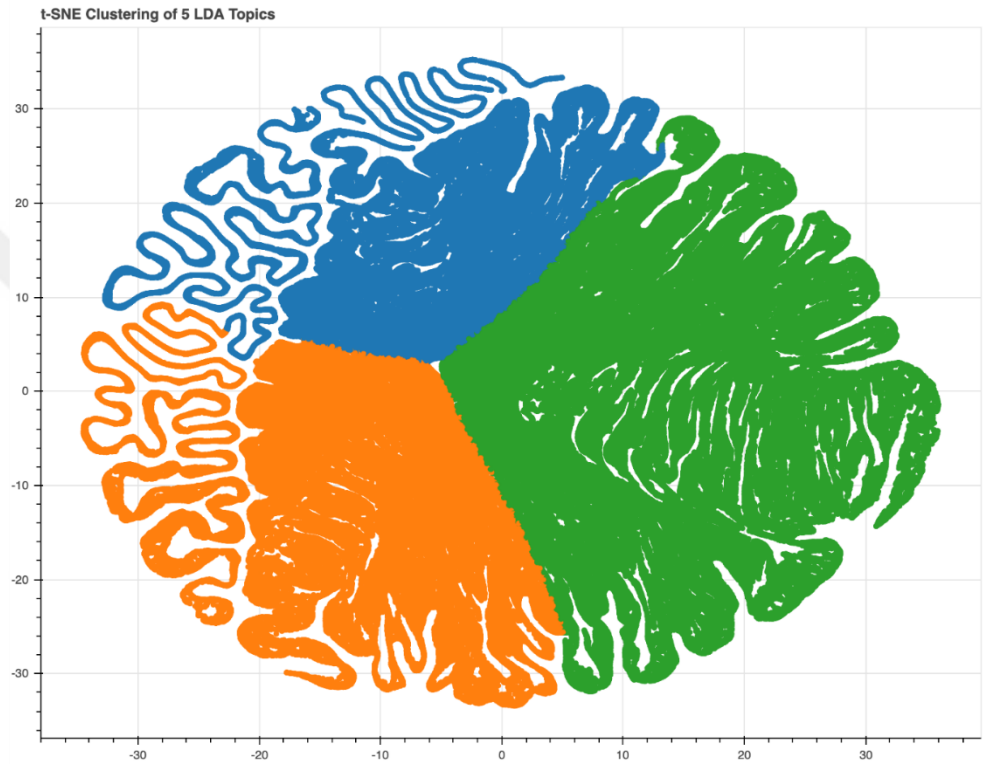
Konu 1 Konu 2

Konu 3

Şekil 5.2. Konu başlıklarında bulunan kelimeler için kelime bulutu

t-SNE boyut sayısının azaltılması için doğrusal olmayan bir teknik olup özellikle yüksek boyutlu veri kümelerinin görselleştirilmesi için çok uygundur (Maaten ve Hinton, 2008). t-SNE algoritmasının ana fikri, noktalar arasındaki uzaklıkları olabildiğince koruyacak bir şekilde düşük boyutlu bir temsil bulmaktır. t-SNE, her bir veri noktası için rastgele bir düşük boyutlu temsil ile başlar ve orjinal uzayda yakın olan noktaları birbirine yakın, uzak olanları ise birbirinden uzak tutmaya çalışır. t-SNE, birbirine uzak noktaların arasındaki uzaklığı korumaktansa birbirine yakın noktalara önem vermektedir. Görüntü işleme, DDİ, konuşma işlemede yaygın olarak uygulanır. t-SNE algoritması kelimelerin yakınlıklarını belirlemek için olasılık dağılımları üzerinden hareket etmektedir. 282351 adet

yorumu 3 konu başlığına t-SNE algoritması yardımıyla ayırdığımız zaman elde edilen grafik Şekil 5.3' te gösterilmektedir. Her bir renk farklı bir alt konu başlığını temsil etmektedir. Grafikte görüldüğü gibi 3 farklı konu birbirinden ayrılmaktadır. t-SNE algoritmasının denetimsiz bir öğrenme metodu olmasına ve veriler hakkında herhangi bir bilgiye sahip olmamasına rağmen yorumları birbirinden başarı ile 3 farklı konu başlığına sınıflandırdığı görülmektedir.



Şekil 5.3. t-SNE algoritması ile konu başlıklarının yakınlık gösterimi

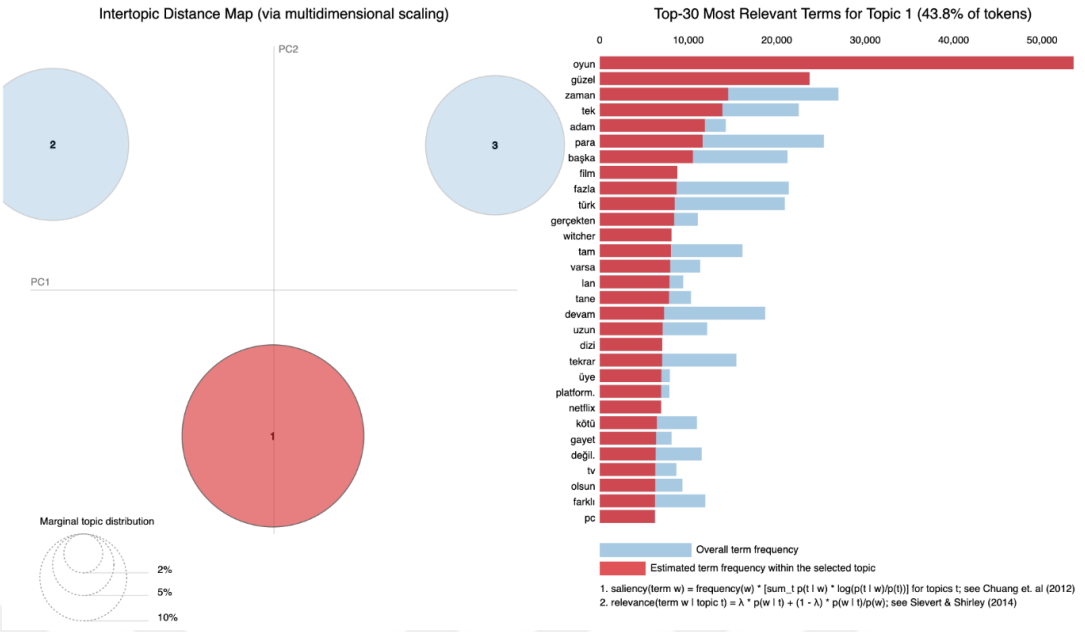
"LDAvis" (Sievert ve Shirley, 2014), kullanıcıların bir metin modeline uygun bir konu modelindeki konuları yorumlamasına yardımcı olmak için tasarlanmıştır. "PyLDAvis", GDA algoritması kullanılarak tahmin edilen konuların web tabanlı etkileşimli bir görselleştirmesi olan LDAvis' i temel alarak oluşturulmuştur. PyLDAvis, kullanıcıların bir konu modelindeki konuları yorumlamasına yardımcı olan etkileşimli konu modeli görselleştirme için bir Python kütüphanesidir (Hidayatullah ve Ma'arif, 2017). PyLDAvis iki farklı sütün içermektedir. İlk sütunda konuların birbirine olan uzaklıkları yer almaktadır. Sağ tarafta bulunan ikinci sütunda ise ilgili konu içerisinde yer alan kelimelerle ilgili veriler bulunmaktadır. PyLDAvis grafiklerinden elde edilen terimlerin konu başlıklarına

dağılımı Çizelge 5.3' te gösterilmektedir. GDA algoritmasından elde edilen konulara, veri önişleme sonucu kalan terimlerin en yüksek %43.8 oranda 1 nolu başlık ile eşleştiği, en düşük ise %25.6 oranında 3 nolu başlık ile eşleştiği görülmektedir.

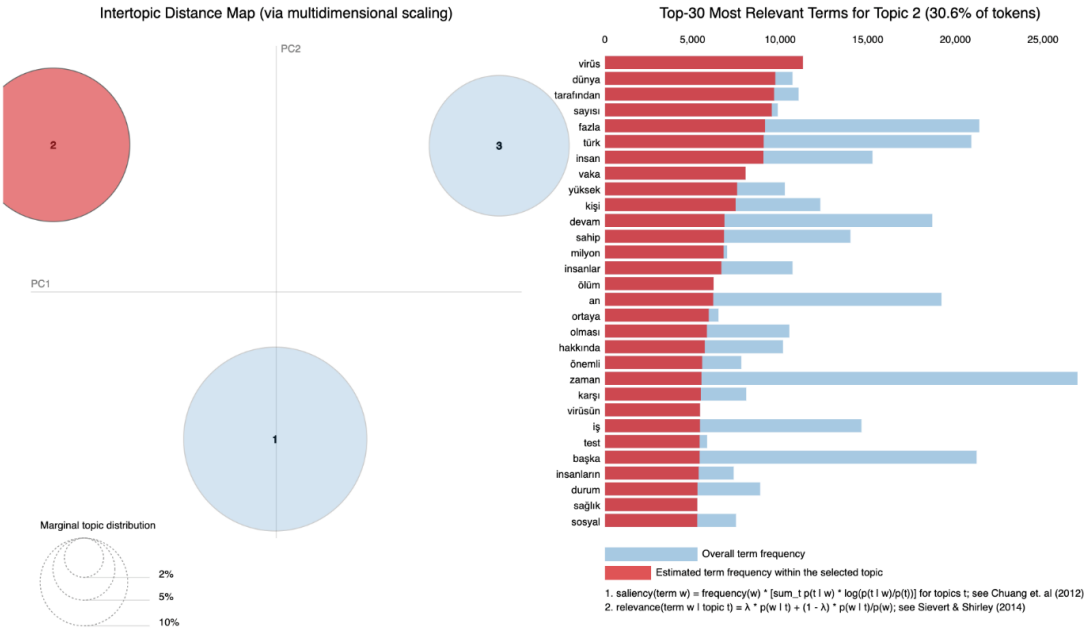
Çizelge 5.3. Terimlerin konulara dağılımı

No	Konu	Yüzde
1	Oyunlar	%43.8
2	Korona Virüsü	%30.6
3	Müşteri Hizmetleri Şikayeti	%25.6

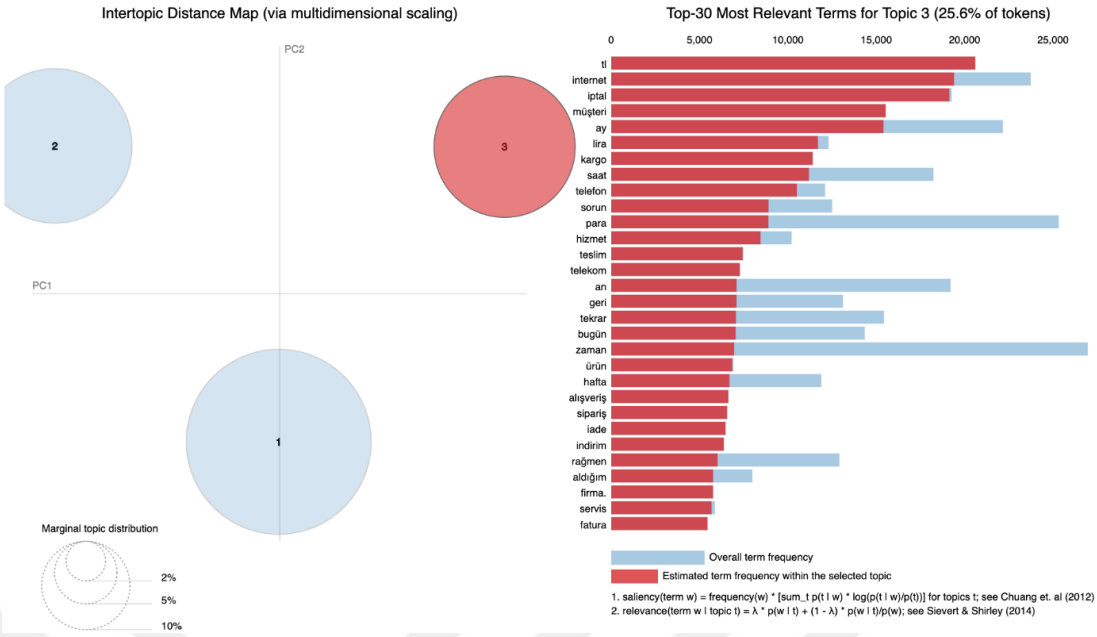
Şekil 5.4, 5.5' te Ekşi Sözlük sosyal platformu "teknoloji" kanalından toplanan yorumlardan oluşturulan 3 konu başlığı için ayrı ayrı PyLDAvis grafikleri gösterilmektedir. Sağ tarafta bulunan sütunda her başlık için en çok ilgili olan 30 terim gösterilmektedir. Yine sağ üst kısımda başlık ile ilgili terimlerin tüm terimlere oranı gösterilmektedir. Örneğin 1 nolu konu başlığına uyan terimler, tüm terimlerin %43.8' ini oluşturmaktadır. 3 Nolu konu başlığında gösterilen terimler, tüm terimlerin %25.6' sına karşılık gelmektedir. Grafik üzerinde sol sütunda konuların birbirine uzaklığı görülmektedir. Grafiğe göre 1,2 ve 3 nolu konuların birbirine yakın konular olmadığı görülmektedir. Şekil 5.4' te 1 nolu konu başlığı için 30 adet kelime listelenmektedir. Bu başlık için en etkili kelimeler "oyun", "güzel" ve "zaman" kelimeleri olarak görülmektedir. Bu konu başlığında bulunan terimler bütün terimlerin %43.8' ini oluşturmaktadır. Çalışmada kullanılan veri seti için konu başlıkları içerisinde öne çıkan başlık olarak görülmektedir. Şekil 5.5' da 2 nolu konu başlığı kanser hastalarının hastalık zamanında moral-motivasyon ve çevrelerinden bekledikleri ilgi üzerine 30 adet kelime listelenmektedir. En etkili kelimeler "virüs", "dünya", "tarafından" kelimeleridir. Analiz sonuçların çıkan "tarafından" kelimesi Türkçe durak kelimeler içerisinde bulunmadığı için görülmektedir. İlerleyen çalışmalarda Türkçe durak kelime listesi, İngilizce durak kelime listesi ile karşılaştırılarak güncellenecektir. Bütün terimlerin %30.6' sı bu konu başlığını oluşturmaktadır. Şekil-5.6' da 3 nolu konu başlığı üzerine 30 adet kelime gösterilmektedir. En etkili kelimeler "tl", "internet" ve "iptal" olarak listelenmiştir.



Şekil 5.4. GDA konu modelleme konu 1 pyldavis grafiği



Şekil 5.5. GDA konu modelleme konu 2 pyldavis grafiği



Şekil 5.6. GDA konu modelleme konu 3 pyldavis grafiği

5.2.2. GDA algoritmasının İngilizce veri setine uygulanması

GDA analizi ve veri ön işleme işlemleri Python dilinde yapılmıştır. Veri ön işleme işlemleri için Nltk kütüphanesi, GDA analizi için ise Gensim (Gensim, 2009) kütüphanesi kullanılmıştır. İngilizce veri seti üzerinde, Türkçe veri setinden farklı olarak kök bulma işlemi gerçekleştirilmiştir. Kök bulma işlemi için "snowball" kütüphanesi kullanılmıştır. İngilizce yorumlar için konu başlıkları Çizelge 5.4' te gösterilmektedir. Bu çalışmada, 245.872 kullanıcı yorumu toplanarak analiz edilmiştir. GDA algoritmasından elde edilen çıktılar ve konuların ağırlıklı yorumları incelendiğinde 1 nolu konunun 2016 yılında yapılan Amerika seçimleri hakkında, Rusya' nın teknolojik olarak yaptığı düşünülen girişimlerin seçim sonuçlarına etkisi olup olmadığı hakkında konuşulduğu, 2 nolu konunun kullanılan mobil cihazların çalışması, insanların yaşamına ve çalışmalarına etkileri/katkıları hakkında konuşulduğu, 3 nolu konu başlığında ise Amerika' da çıkması düşünülen internet servis sağlayıcıları ile ilgili bir yasa teklifi hakkında konulardan oluştuğu gözlemlenmiştir.

Çizelge 5.4. GDA algoritmasından çıkarılan İngilizce kelimeler ve çıkarılan başlıklar

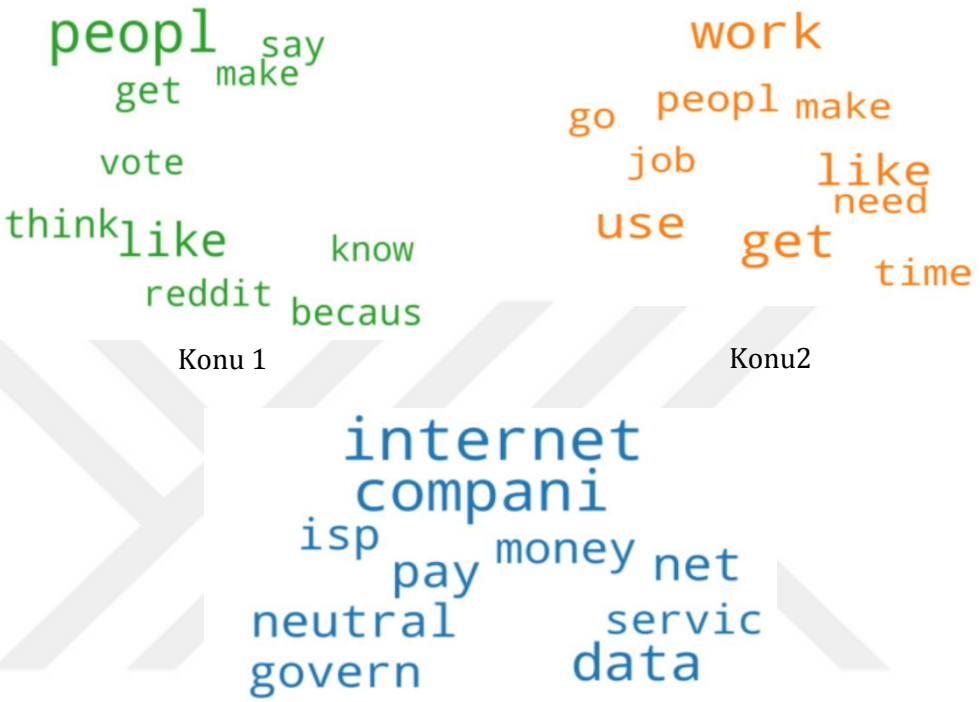
No	Başlık	Etiket
1	peopl, like, think, get, say, becaus, reddit, vote, know, make	Amerikan Seçimleri
2	work, get, use, like, time, peopl, go, make, need, job	Mobil Cihazlar
3	internet, compani, data, pay, net, neutral, govern, money, isp, servic	Yasa Teklifi

Bu çalışmada, Reddit sosyal platformundan 245.872 kullanıcı yorumun analiz edilmiştir. Her başlığın belirlenmesinde en fazla orana sahip olan yorumlar ve başlıklar Çizelge 5.4' te gösterilmektedir. En fazla etkisi olan yorumlar incelendiğinde konu başlıkları ile uyumlu olduğu görülmektedir.

Çizelge 5.5. Başlıklarda en fazla ağırlığa sahip yorumlar

No	Konuya Uyumluluk	Etiket	Yorum
1	0.9957	Amerikan Seçimleri	well, that proves everything. can you give me a reason to believe these reports if i can't see the evidence then? there have been multiple accounts of the government lying to the people.it's the side that's too far up it's own ass to ever consider that their intellectual minds could fall for the same trap that their disgusting, inbred enemies did.like this article? lol.you don't really think all anti trump people only see evidence based information... right? that only pro trump people twist and turn until they see what they want? i won't argue for or against anything except that r/politics is not information based. they are simply anti trump and it's all fake and/or opinion pushing. i actually resubbed because i get a couple of daily laughs there.where's that video of trump pissing on hookers that we heard all about?this is true...i went there yesterday to see what they were saying about the russians being indicted. the linked an article that said the russians were accused of planning pro clinton, sanders, and both pro and anti trump rallies. they immediately started saying the russians were all “bernie bros” and refusing to even acknowledge that the influence was directed to all fronts. that's one of the things that bothers me most about reddit. people will scroll through your entire history, see one post, and then spend 20 minutes insulting you and calling you a racist nazi. it doesn't matter what you posted or why you posted in it. you post in td and you are forever called an inbred racist. it's dumb and i can't believe the people don't see the problem with this attitude. i didn't vote trump and i'm not even kind of a fan, but the way the entire left acts on this site is an embarrassment. when you do a search and it says users are those active one then?i got banned from r/twoxchromosomes or whatever it's called just for having posted on t_d once. ....
2	0.9983	Mobil Cihazlar	https://www.apple.com/support/iphone6s-unexpectedshutdown/how about letting me put in a new batt then?that's not the issuethe problem is apple never told us of this feature until after they were exposed. if this feature was included in the upgrade list people were not be so upset my dad still had the ip4 with ios 7, most o the apps cant use anymore, even the youtube app can't run properly, only call and text msg can hold almost 6-7 hours, so used it as a secondary phone just for works.don't worry, even though i'm not an apple guy, but i can say the 4s (maybe 5s too, but it's bigger) is the best phone design ever, small, thick, and great screen resolution is all you need in a phone. big ups 4s squadi loved the 4s i noticed an immediate downgrade when i got the 6still my favorite iphone. i wish they made a new one with the same design and even a little heft. though i do have a metal case that makes my 7 similar. if you're running a newer version of ios on the 4s, you can actually downgrade back to ios 6, just to let you know!my fiance kept her 4s up until a year ago. she only got rid of it because the screen peeled away from the rest of the phone and i was getting a new phone at the same time. had a 4s for over 5 years now and i've been dragging it through its retirement with a portable charger, it finally ran out of juice and has been on charge for 2 days now without coming back oni miss the smaller phone too my dad's first smart phone was a 4s what seems like a decade ago but he refuses to give it up. it's funny though because i don't really see his phone that often so when ever i do, i'm just amazed at how tiny it is. phones have just grown in size so much after spending a decade shrinking.yup that's still what i use. 4s isn't so bad still. i do like how small it is sometimes.my daughter has a 4s and... wait for it... she's getting upgraded to a 5c for christmas! she's going to be elated!
3	0.9967	Yasa Teklifi	the federal communications commission’s proposed net neutrality rules would, among other things, prohibit broadband access providers from prioritizing traffic, charging differential prices based on the priority status, imposing congestion-related charges, and adopting business models that offer exclusive content or that establish exclusive relationships with particular content providers. the proposed regulations are motivated in part by the concern that the broadband access providers will adopt economically inefficient business models and network management practices due to a lack of sufficient competition in the provision of broadband access services. this paper addresses the competitive concerns motivating net neutrality rules and addresses the potential impact of the proposed rules on consumer welfare. we show that **there is significant and growing competition among broadband access providers and that few significant competitive problems have been observed to date**. we also evaluate claims by net neutrality proponents that regulation is justified by the existence of externalities between the demand for internet access and content services. we show that **such interrelationships are more complex than claimed by net neutrality proponents and do not provide a compelling rationale for regulation.** we conclude that **antitrust enforcement and/or more limited regulatory mechanisms provide a better framework for addressing competitive concerns raised by proponents of net neutrality.

Şekil 5.7’ de her bir konu başlığını oluşturan kelimeler için hazırlanmış olan kelime bulutu gösterilmektedir. Her bir konu başlığı farklı renkle temsil edilmektedir. Her konu başlığı altında bulunan kelimelerin ağırlıklarına göre, kelime bulutu içerisindeki boyutları belirlenmektedir.

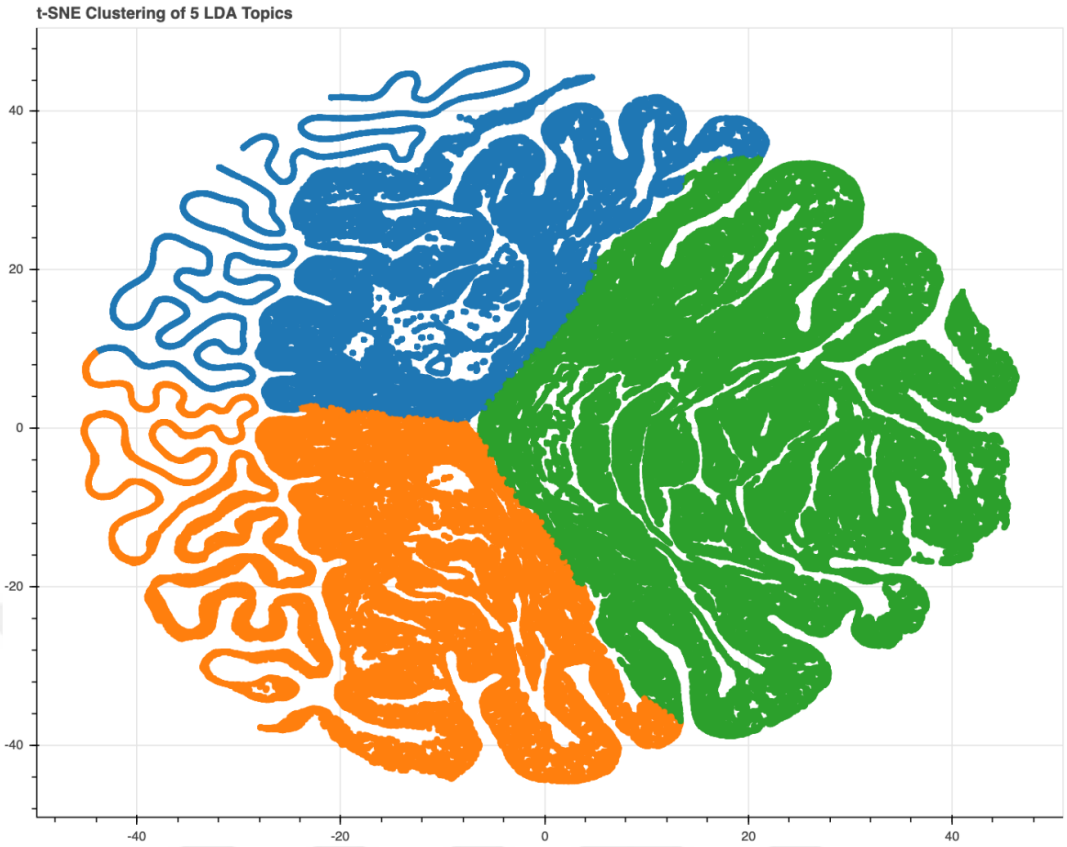


Konu 1 Konu2

Konu 3

Şekil 5.7. Konu başlıkları için kelime bulutu

245.872 adet yorumu 3 konu başlığına t-SNE algoritması yardımıyla ayırdığımız zaman elde edilen grafik Şekil 5.8’ de gösterilmektedir. Her bir renk farklı bir alt konu başlığını temsil etmektedir. Reddit sosyal platformundan t-SNE algoritması ile elde edilen şekil incelendiğinde konuların birbiri ile ayrıldığı gözlemlenmektedir.



Şekil 5.8. t-SNE algoritması ile konu başlıklarının yakınlık gösterimi

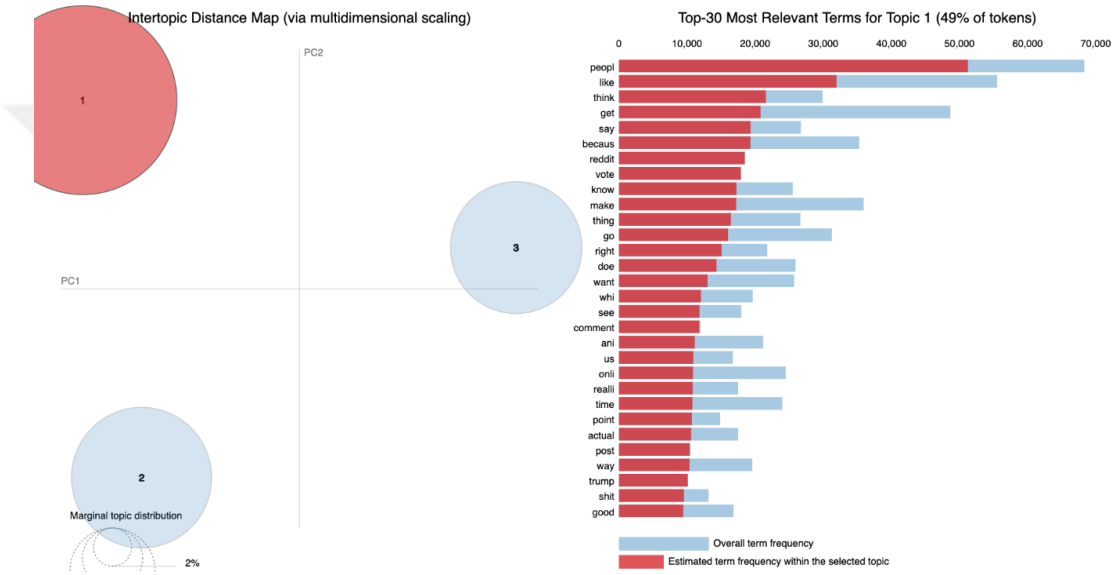
PyLDAvis grafiklerinden elde edilen terimlerin konu başlıklarına dağılımı Çizelge 5.6' te gösterilmektedir. GDA algoritmasından elde edilen konulara, veri önişleme sonucu kalan terimlerin en yüksek %49 oranda 1 nolu başlık ile eşleştiği, en düşük ise %23.9 oranında 3 nolu başlık ile eşleştiği görülmektedir.

Çizelge 5.6. Terimlerin konulara dağılımı

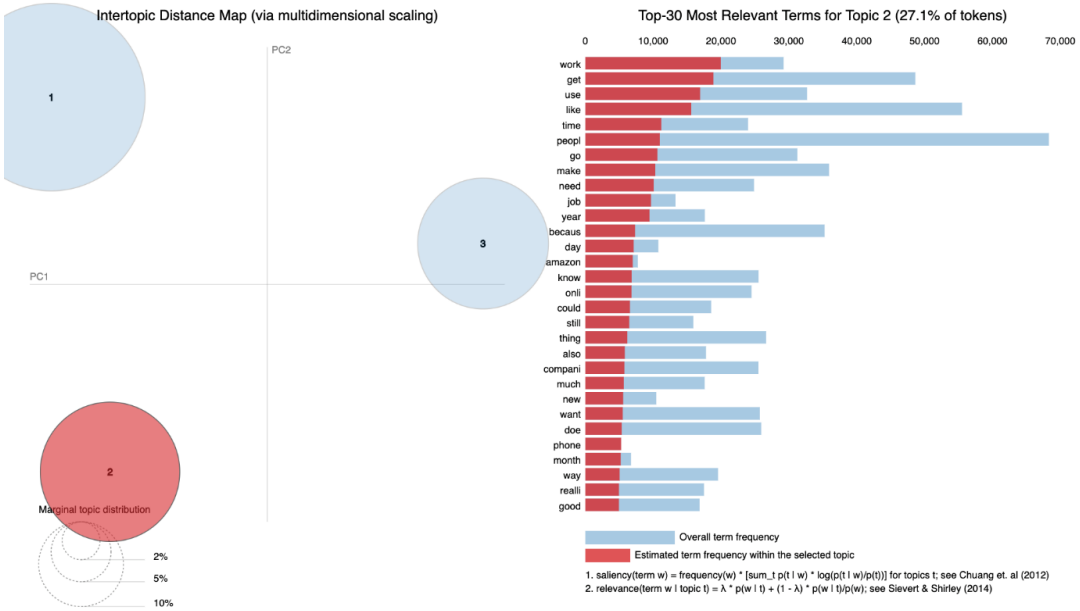
No	Konu	Yüzde
1	Amerikan Seçimleri	%49
2	Mobil Cihazlar	%27.1
3	Yasa Teklifi	%23.9

Aşağıdaki şekillerde Reddit sosyal platformu "technolgy" kanalından toplanan yorumlardan oluşturulan 3 konu başlığı için ayrı ayrı PyLDAvis grafikleri gösterilmektedir. Grafiğe göre 1,2 ve 3 nolu konuların birbirine yakın konular olmadığı görülmektedir. Şekil 5.9' da 1 nolu konu başlığı için 30 adet kelime listelenmektedir. Bu başlık için en etkili kelimeler "peopl", "like" ve "think"

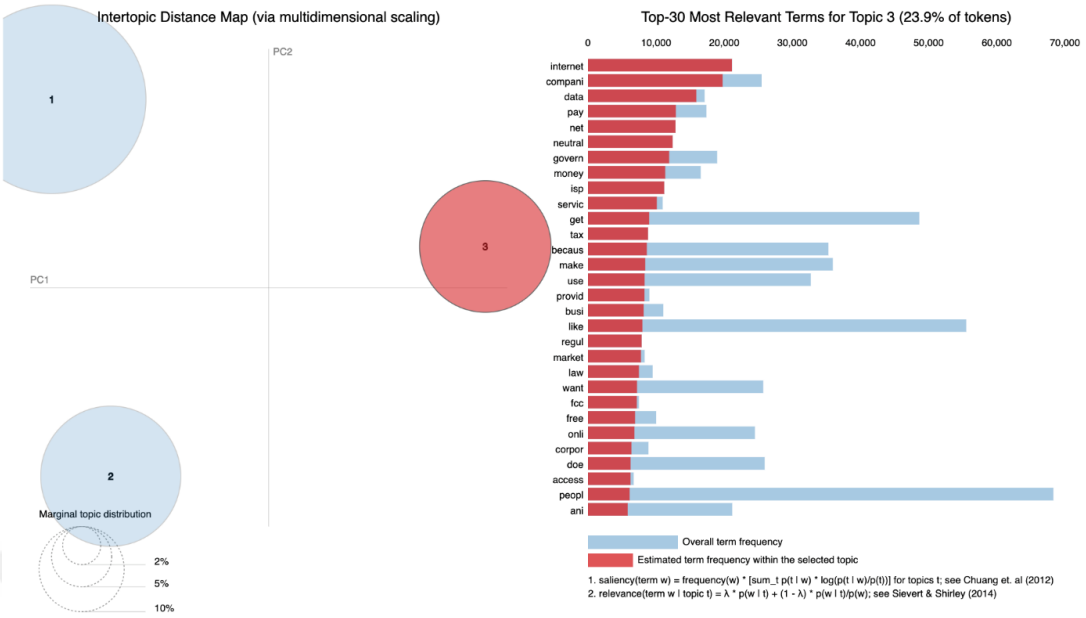
kelimeleri olarak görülmektedir. Bu konu başlığında bulunan terimler bütün terimlerin %49' unu oluşturmaktadır. Çalışmada kullanılan veri seti için konu başlıkları içerisinde öne çıkan başlık olarak görülmektedir. Şekil 5.10' da 2 nolu konu başlığı ile ilgili 30 adet kelime listelenmektedir. En etkili kelimeler "work", "get", "use" kelimeleridir. Bütün terimlerin %27.1' sı bu konu başlığını oluşturmaktadır. Şekil 5.11' de 3 nolu konu başlığı üzerine 30 adet kelime gösterilmektedir. En etkili kelimeler "internet", "compani" ve "data" olarak listelenmiştir. Bütün terimlerin %23.9' u bu konu başlığını oluşturmaktadır.



Şekil 5.9. GDA konu modelleme konu 1 pyldavis grafiği



Şekil 5.10. GDA konu modelleme konu 2 pyldavis grafiği



Şekil 5.11. GDA konu modelleme konu 3 pyldavis grafiği

GDA algoritması ile oluşturulan İngilizce ve Türkçe dillerinin konu modellemeleri sonucu ortaya çıkan konular incelenmiştir. Sonuçlarda Reddit ve EkşiSözlük Sosyal platformlarında teknoloji başlığı üzerine yapılan yorumlardan başlıca 3 konu başlığı elde edilmiştir. Elde edilen konu başlıkları karşılaştırıldığında her platformun kendi içerisinde farklı konuların ön plana çıktığı görülmektedir. Amerika' da yoğun olarak kullanılan Reddit platformunda, elde edilen 3 konu başlığından 2' sinin Amerika' nın kendi içindeki gündemi yansıttığı görülmektedir. Bu gündemlere bakıldığında Amerika' da yapılan seçimlerin öncelikli olarak konuşulduğu görülmektedir. Bir diğer konu başlığı olarak ise çıkarılan bir yasa teklifi üzerine konular olduğu görülmektedir. Bu konu başlığıda Amerika' yı ilgilendiren bir başlık olarak görülmektedir. Diğer konu başlığında ise mobil cihazlar hakkında konuşulmuştur. Türkiye' de yaygın bir şekilde kullanılan EkşiSözlük platformunda ise kullanıcıların "Oyunlar" üzerine konuştuğu görülmektedir. Bir diğer konu olarak ise bugünlerde bütün dünyanın gündemini oluşturan "Koronavirüs" olmuştur. EkşiSözlük platformunda konuşulan diğer konu başlığı ise Türkiye' deki firmaların müşteri hizmetleri üzerine konuşulduğu gözlemlenmiştir.

6. VARLIK İSMİ TANIMA VE DBPEDIA

Bu bölümde bilgi tabanlı veri çıkarımı, ontoloji yapısı, ontoloji sorgulama dili olan SPARQL, tez çalışmasında kullanılan DBPedia projesi ve varlık ismi tanıma hakkında bilgi verildikten sonra veri seti üzerinde VİT ve DBPedia projesi uygulaması anlatılmıştır.

6.1. Bilgi Tabanlı Veri Çıkarımı

Web teknolojisinin gelişim seyrini temsil etmek üzere Web 1.0, Web 2.0 ve Web 3.0 tanımlamaları yapılmaktadır fakat bu durumun herhangi resmi bir temeli bulunmamaktadır. Ancak Web 1.0 "Durağan (Static)"; Web 2.0 "Etkileşimsel (Interactional)"; Web 3.0 ise "Anlamsal (Semantik)" olarak tanımlanmaktadır. Anlamsal Ağ (AA), internet ortamındaki kaynakların daha kolay erişilebilir, makineler tarafından anlaşılabilir ve yazılım ajanları tarafından kullanılabilir hale getirilmesi amacıyla yeniden tanımlanması fikridir. AA, web teknolojisinin güncel bir sürümü; yazılımlar için yeni bir üstveri (metadata) teknolojisi; açık kaynak teknolojiler için farklılık veya yeni nesil bir yapay zekâ teknolojisi olarak değerlendirilebilir (Pollock, 2009).

AA fikrinin ortaya konduğu 1990’ lı yılardan beri çeşitli çalışmalar yapılmış anlamsal ağ ile ilgili teknoloji ve standartlar oluşturulmuştur. Bu teknoloji ve standartlar ontolojileri işleyip sonuç çıkarma yapabilecek olan KTÇ, KTÇ Şeması (KTÇŞ), internet ontoloji dillleri, SPARQL sorgulama dili gibi araçlardır. Buna rağmen bağlı veri kavramının ortaya çıkması ile anlamsal web daha etkin bir şekilde kullanılmaya başlamıştır (Berners, 2019).

AA’ nın günümüzdeki gerçek hali bağlı verilerdir (Berners, 2001). Bağlı verinin temellerini hem yapısal olarak hemde bilgisayarın anlayabileceği KTÇ (W3C,1994) standardına uygun olarak yayınlanması bağlı verinin temeli olarak görülmektedir (Bizer vd., 2009). AA’ daki temel amaç iyi tanımlanmış ve bağlantılandırılmış olan bilgilerin ve servislerin internet ortamında kolay bir şekilde bilgisayarca-okunabilir ve bilgisayarca-anlaşılabilir olmasını sağlayacak standartların ve teknolojilerin geliştirilmesidir.

W3C, 1994 yılında internetin ortaya çıkmasında büyük rol oynayan Tim Berners-Lee tarafından kurulmuş bir topluluktur. Bu topluluğun asli görevi; internet dünyası tarafından ihtiyaç duyulan standartları organize etmek ve bu standartlar için gerekli teknolojilerin ortak projeler içinde oluşmasına liderlik yapmaktır. W3C, internet standartlarının ana kaynağıdır ve HTML dili W3C tarafından bir standart haline getirilmiştir. KTÇ, SPARQL ve ontoloji dili teknolojileri W3C tarafından geliştirilmeye devam etmektedir. W3C, AA projeleriyle etiketleme işaretlerini de sorgulayan yeni standartların geliştirilmesini öngörmektedir. AA ile beraber ortaya çıkan KTÇ, şu anda var olan veya bundan oluşacak verilerin birbirleriyle bağlantılı hale getirilmesini sağlayacaktır. SQL veri tabanı ve HTML işaretlemeleri arasında uyumsuzluklar yeniden düzenlenip, bilgisayar tarafından kullanıcı için otomasyon kolaylığı getirilecektir. Web semantiğin oluşturduğu küresel bazlı veri tabanında tek dile dönüştürülmüş veriler, SPARQL sorgu dili ile sorgu sonuçları raporlanabilecektir. Ontoloji dilleri, ise AA dünyasının yeni temel dil yapısı olarak KTÇ ile oluşturulmuş ve SPARQL ile sorgulanmış verilerin, bilgisayar ve ağlar üzerinde değerlendirmesini yapacaktır. Ontoloji dillerinin kullanımıyla birlikte anlamsal sorgulama yapan makine, insan zekâsını örneklem alacaktır. W3C, AA için bağlı veri kavramını kullanmaktadır. Bağlı veri, web ağı içerisinde her bir bilgiyi belli bir anlama sahip olacak şekilde modelleyerek, bu bilgilerin birbirleriyle ilişkilendirilmesi ve akıllı veri tabanlarının oluşumu hedeflemektedir (W3C,1994; DBPedia,2004; Akçavlı, 2012).

6.2. Ontoloji Nedir?

Varlıkları bağlantıları ile birlikte belirten felsefi bir terimdir ve AA en belirgin özelliğidir (Çoban, 2010). İnternet Ontolojisi, internet üzerindeki bir alanda (domain, özel bir konuya ait bilgi alanı), paylaşılabilinir bilgiye ulaşmak isteyen ihtiyaç sahiplerine, nesnelerin kurallı tanımı yaparak ortak kelimeler ve anlamlar sunmaktadır (Karademirci, 2008). Aşağıda yaygın olarak kullanılan çeşitli ontoloji geliştirme araçları gösterilmektedir (Yüksek ve Karasulu, 2010).

- Protege
- Apollo
- LinkFactory
- OntoEdit

- OpenKnoME
- SESAME
- RDFDB

Anlamsal teknolojilerin arkasındaki temel bileşen, belirli bir alanda tanımlanan kavramlar arasındaki ilişkiler ve taksonomi ile kavramsallaştırmayı temsil eden ontolojilerdir. Farklı platformlar arasında bilgi paylaşımı ve değişimi yapılabilmesinde önemli rol oynarlar. Bu sebepten dolayı ontoloji kavramı, sözdizimi, ontoloji dilleri ya da KTÇ ile temsil edilmesi ve ontoloji sorgu dilleri olan RDQL, SPARQL konularındaki çalışmalarda artış görülmektedir (Berners, 2001). Anlamsal teknolojilerin yaygınlaşması ile DDİ alanındaki gelişmeler yeni bir yön kazanmıştır. Araştırmacılar arasında kullanıcıyı anlamak ve sorularına cevap üretmek için DDİ tekniklerini anlamsal teknolojiler ile birleştiren melez yöntemler tartışılmaya başlanmıştır. Bu aşamada ilk odaklanılan konu anlamsal teknolojilerin DDİ yöntemlerini güçlendirmeye nasıl yardımcı olacağı olmuştur. İkinci önemli konu ise, AA geliştirmede kullanılan ontoloji öğrenmesi, ontoloji sorgulama, çok dilli ontoloji eşleştirme yöntemlerinde DDİ' nin hangi noktalarda katkı sağlayabileceği olmuştur (Guo ve Ren, 2009).

Cümle anlama, içerisinde temel olarak iki aşamadan oluşur. Bunlardan ilki morfolojik olarak incelenmesi, ikincisi ise anlamsal analizdir (Guo vd., 2011). Morfolojik analiz kelimelerin yapılarıyla ilgilenirken anlamsallığa ulaşmak için kelimeler arası ilişkiler, kelimelerin türü etiketleri (isim, fiil, sıfat, vb.) ve cümle içindeki görevlerini (özne, yüklem, nesne, vb.) çıktı olarak üretir. Anlamsallık seviyesine katkıda bulunmak için, varlık ismi tanıma yönteminden faydalanılır. Bu yöntem doğal dil sorgusunda bulunan varlıkların önceden tanımlanmış kategorilerle (kişi, organizasyon, tarih, lokasyon, vb.) işaretlenmesini sağlar (Collobert vd. ,2011). Üretilen bu çıktılar, doğal dil sorgusunun zenginleştirilmesi için ontolojilerle birleştirilerek kullanılır. Bu yaklaşım özellikle doğal dil ile sorgulama imkânı sağlayan soru cevaplama sistemleri için ontolojilerin bilgi kaynağı olarak kullanılmasını sağlar. Doğal dil sorgusu ile temsil edilen bilgi ihtiyacı ontolojinin sorgu diline dönüştürülür. Kullanıcılar ontoloji sorgu dilini, yapıyı ve ontoloji sözlüğünü öğrenme ihtiyacı olmaksızın ontolojilerdeki bilgiye

ulaşabilirler. Kullanıcılar ile sistemler arası köprü kuran bu yaklaşım sayesinde ontolojilerin pratik kullanımı da sağlanmış olmaktadır (Bernstein,2005).

6.2.2. Kaynak tanımlama çerçevesi

KTÇ, 1997 yılında W3C tarafından duyurulmuştur. KTÇ' nin ortaya çıkışında anlamlı bir internete duyulan ihtiyacın artması önemli bir rol oynamıştır. KTÇ' nin ortaya çıkmasında kullanıcı toplulukları, "The Internet Engineering Task Force", W3C gibi birçok kurum ve kullanıcı rol almıştır. 1999 yılında KTÇ web' in işlevi ve iş birliğini arttırmak amacıyla W3C' nin standartları arasına girmiştir. İnternet içerisinde yer alan kaynakların tanımlanabilmesi için geliştirilmiş bir dildir. İnternet kaynakları içerisinde yer alan bir metadatanın saklanabilmesine olanak sağlamış ve bu verilerin farklı uygulamalar arasında değiştirilmesi sırasında yaşanabilecek anlamsal kaybı engelleyen bir yapıya sahiptir (Altay ve Ulaş, 2018).

XML dili verilerin kodlanması ve taşınması için sözdizimi yapısını belirler. KTÇ, bir veri modelidir. Bu model internet ortamındaki nesnelerin (kaynakların), kaynak özelliklerinin ve özellik değerlerinin tanımlanması fikrine dayanır. KTÇ ifadelerinde yer alan nesne, özellik, değer üçlüleri KTÇ' nin temelini oluşturur.

6.2.3. SPARQL

SPARQL, KTÇ Sorgulama Dili' dir (Sparql, 2008). W3C tarafından KTÇ veri modeli için tanımlanmıştır. KTÇ sorgulama dilleri üzerindeki çalışmalar son birkaç yıldır devam etmektedir. Bu süreç içerisinde RDQL, Squish, Versa gibi farklı yaklaşımların bu süreç içerisinde geliştirilmiştir. SQL ile benzer söz dizimine sahip olan SPARQL yaygın bir şekilde kullanılmaktadır (Özpala ve Köker, 2008). KTÇ modeli "subject, predicate, object" (özne, yüklem, nesne) olarak ifade edilen üçlülerden oluşan dizgelere dayanmaktadır. Benzer şekilde SPARQL sorgusu da üçlüler şeklinde ifade edilmekte ancak bu üçlülerden herhangi birisi veya birileri değişken olabilmektedir. SPARQL sorgusu ile KTÇ çizgesindeki üçlüler eşlenerek sorgu cevaplanmaya ve KTÇ çizgesinden ilgili üçlüler çekilmeye çalışılmaktadır (Erdur ve Alatlı, 2013).

“SELECT”, “ASK”, “DESCRIBE” ve “CONSTRUCT” olmak üzere dört çeşit sorgu tipi desteklenmektedir. SELECT sorgusu verilen sorgulama örüntüsüne uyacak şekilde üzerinde çalışılan veri kümesinden istenilen değişkenlerin tamamını ya da bir kısmını döndürmeye yarar. ASK sorgusu veri kümesi içinde sorgulama örüntüsünü karşılayan veri bulunup bulunmadığının cevabını döndürür. DESCRIBE sorgusu, sorgulama örüntüsü ile veya doğrudan URI ile tanımlanan kaynağın KTÇ veri kümesi içindeki tanımlamasını döndürür. CONSTRUCT sorgulaması verilen sorgulama örüntüsünü veri kümesi içinde arar ve yine sorguda verilen şablona uyan bir şema (graph) üretir (Sparql, 2008).

KTÇ sorgulama araçlarının büyük çoğunluğunda SPARQL desteği bulunmaktadır (Battal, 2009). Aşağıda verilen örnek kod bloğunda SPARQL sorgusunun kullanımı verilmiştir.

```
PREFIX a:<http://ornek.com/resources/> ...

SELECT ?x

WHERE { ... }
```

Günümüzde DBpedia, Freebase, Geonames, DBLP gibi çeşitli siteler verilerini KTÇ olarak yayınlanmakta ve SPARQL ile sorgulanabilmektedir.

6.2.4. DBPedia

DBpedia, Wikipedia üzerinde sunulan yapısal verilerin web üzerinden sorgulanabilmesi için semantik veritabanlarında tutulması amacıyla “Free University of Berlin, University of Leipzig” ve “OpenLink Software” işbirliği ile 2007 yılında başlatılan bir projedir. DBpedia, Wikipedia üzerindeki bilgi kutularında bulunan yapısal verilerin semantik veritabanlarında depolanmasıyla oluşturulmuş, Linked Open Data bünyesinde bulunan en önemli merkezi bilgi tabanlarındandır (Bizer vd., 2009). DBpedia üzerinde 2.35 milyonu tanımlı ontoloji içerisinde sınıflandırılmış olmak üzere toplamda 3.77 milyon üçlü tanımlanmıştır. Bu üçlüler içerisinde 764.000 insan, 573.000 yerleşim yeri, 333.000 sanat ürünü ve 192.000 organizasyon tanımlıdır. DBPedia içerdiği 50

milyon KTÇ linki ile diğer bağlı veri setlerine bağlantı sağlamaktadır. DBpedia bilgi tabanının mevcut bilgi tabanlarına göre birkaç avantajı vardır: birçok alanı kapsar; toplulukların katılımını temsil eder. Wikipedia' nın içeriği değiştikçe otomatik olarak gelişir ve çoklu dil desteği bulunmaktadır. DBPedia bilgi setine sorgular oluşturarak sorgular sorulabilmektedir. Bu sorguların cevabı DBPedia aracılığı ile Wikipedia' dan gelmektedir (DBPedia, 2004).

DBPedia, WikiPedia üzerinde yapılandırılmış olan verilerin çıkarılması ile geniş kapsamlı ve çok dilli bilgi tabanı projesidir. Wikipedia' dan yapılandırılmış bilgileri çıkarmak ve bu bilgileri Web' de kullanıma sunmak için bir topluluk çalışmasıdır. DBpedia, Wikipedia' dan türetilmiş veri kümelerine karşı karmaşık sorgular sormanızı ve Web' deki diğer veri kümelerini Wikipedia verilerine bağlamanızı sağlar (Auer vd., 2007).

WikiPedia üzerinde bulunan bir metinden ilgili alanları, özellikleri çıkarmak 4 aşamadan oluşmaktadır. WikiPedia ile DBPedia arasında bulunan sistemin genel çalışma şekli Şekil 6.1' de gösterilmektedir.

Şekil 6.1. DBPedia veri çıkarım sistemi

Bu 4 aşamanın ilk aşaması olan girdi bölümünde Wikipedia, harici kaynaktan veriyi okur. WikiPedia' da bulunan sayfalar bölümler direk okunabilir veya MediaWiki API yardımıyla alınabilir. Bölümleme katmanında her WikiPedia sayfası ayrıştırıcı yardımıyla soyut olarak bir söz dizimi ağacına dönüştürülür. Soyut söz dizimi ayrıştırılan sayfayı çıkarım yapılmak üzere çıkarım kısmına iletir. Bu kısımda söz dizim ağaçından KTÇ ifadeler oluşturulur. Çıktı kısmında ise oluşturulan KTÇ ifadeler bir havuzda toplanır.

DBPedia üzerinde, WikiPedia metinler çıkarım yapıldıktan sonra elde edilen bazı özellikler Çizelge 6.1' de gösterilmektedir. Örnek olarak verilen bilgilerin başında 3 harften oluşan önekler bulunmaktadır. Bu öneklere kaynak tanımlayıcı adı verilmektedir ve diğer kaynaklardan ayıran tanımlayıcı ismini belirtmektedir. DBPedia' da başlıca 3 adet kaynak tanımlayıcı bulunmaktadır.

http://dbpedia.org/resource: Önek olarak "dbr" kullanılır. Makale verilerinin sunumu yapılır. WikiPedia' da bulunan her bir makale için bir önek DBPedia tarafında oluşmaktadır.

http://dbpedia.org/property: Önek olarak "dbp" kullanılır. Satırlarda bulunan özellikler bu önek ile çıkarılır.

http://dbpedia.org/ontology: Önek olarak "dbo" kullanılır. Ontoloji türünde bilgilerin sunumu yapılır.

Çizelge 6.1. DBPedia çıkarıcıları genel görünüm

Ad	**Tanım**	**Örnek**
abstract	WikiPedia yazılarının ilk satırları	dbr:Berlin dbo:abstract "Berlin is the capital city of (...)" .
article categories	Yazıların kategorilerinin çıkarılması.	dbr:Berlin dbo:abstract "Berlin is the capital city of (...)" .
category label	Kategorilerde bulunan etiketlerin çıkarılması.	dbr:Category:English novels rdfs:label "English novels" .
category hierarchy	Kavramların kategorileriler ile ilişkisi çıkarılır.	dbr:Category:World War II skos:broader dbr:Category:Modern history .
disambiguation	Ayırt etme linkleri çıkarılır.	dbr:Alien dbo:wikiPageDisambiguates dbr:Alien (film) .
external links	Dış sayfa linkleri ayırt edilir.	dbr:Animal Farm dbo:wikiPageExternalLink
geo coordinates	Coğrafik koordinat bilgilerini çıkarır	dbr:Berlin georss:point "52.5006 13.3989" .
grammatical gender	Kişilerin cinsiyet bilgileri	dbr:Abraham Lincoln foaf:gender "male" .
homepage	Kurumsal web adresleri	dbr:Alabama foaf:homepage .
image	WikiPedia Sayfasında yer alan ilk resim	dbr:Berlin foaf:depiction http://...../overview-berlin.jpg
infobox	Bilgi kutusunda yer alan veriler	dbr:Animal Farm dbo:date "March 2010" .
interlanguage	İnterWiki linkleri	dbr:Albedo dbo:wikiPageInterLanguageLink dbr-de:Albedo
label	Makale başlığı	dbr:Berlin rdfs:label "Berlin"
mappings	WikiPedia bilgi kutularındaki verileri haritalama yöntemi ile DBPedia Ontolojisi çıkarımı	dbr:Berlin dbo:country dbr:Germany
page ID	Yazının id numarası	dbr:Autism dbo:wikiPageID "25"
page links	WikiPedia sayfaları arasındaki sayfa bağlantıları	dbr:Autism dbo:wikiPageWikiLink dbr:Human brain .
persondata	Person-Data şablonundan kişisel bilgilerin çıkarımı	dbr:Andre Agassi foaf:birthDate "1970-04-29"
PND	Kişi ile ilgili PND bilgileri	dbr:William Shakespeare dbo:individualisedPnd "118613723" .
redirects	WikiPedia 'daki yazılar arası yönlendirme	dbr:ArtificialLanguages dbo:wikiPageRedirects dbr:Constructed language .

6.3. Varlık İsimlerini Tanıma

VİT çalışmaları; DDİ, Veri Madenciliği ve Veri Çıkarımı gibi alanlarda önemli bir yere sahiptir. Bir metinde geçen varlıkların bulunarak önceden tanımlı kişi, yer ve organizasyon gibi sınıflardan bir tanesine atanmasına VİT denilmektedir. Konu üzerine yapılan ilk çalışmalardan biri Lisa F. Rau (Rau, 1991) tarafından 1991 yılında gerçekleştirilmiş olup yazar çalışmasında problemi metin içerisinde geçen şirket isimlerini işaretleme olarak ele almıştır.

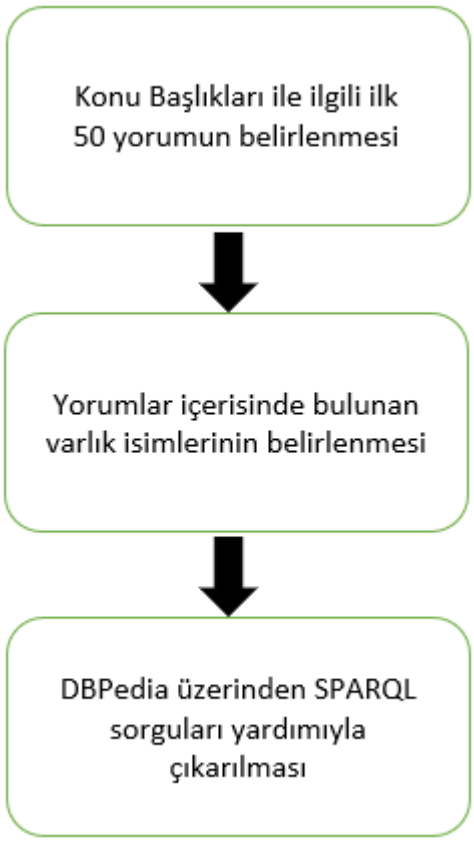
DDİ araştırma alanının önemli bir konusu olan VİT sistemleri; bağımsız sistemler olarak kullanılabilecekleri gibi; otomatik bilgi erişim, medya izleme, soru-cevaplama ve özet çıkarma gibi sistemlerin bünyesinde de önemli birer modül olarak hizmet verebilmektedirler.
VİT konusunda özellikle İngilizce metinler üzerinde; kural-tabanlı, istatistiksel ve makine öğrenmesi yöntemleriyle çok çeşitli çalışmalar yapılmış, temel varlık ismi türleri (sınıfları) olan kişi, yer ve kurum isimleri dışında da birçok tür çalışmaların kapsamına alınmış, ve büyük boyutlarda işaretlenmiş ortak veri kümeleri hem değerlendirmelerde hem de öğrenen sistemlerin öğrenme aşamalarında kullanılmıştır (Marrero vd., 2013).

Türkçe metinlerde varlık ismi tanıma konusunda; istatistiksel yöntemler (Tür vd., 2003), kural-tabanlı (Küçük ve Yazıcı, 2009) ve hibrit yaklaşımlar (Küçük ve Yazıcı, 2012), ve makine öğrenmesi yöntemlerini (Özkaya ve Biri, 2011; Şeker ve Eryiğit, 2012) kullanan çeşitli çalışmalar yayınlanmıştır. Bu çalışmalar çoğunlukla haber metinleri üzerinde değerlendirilmiştir. Sonraki dönemde daha çok sosyal medya metinleri gibi yazım ve dilbilgisi hataları da içeren metinler üzerinde yapılan çalışmalar sunulmaktadır (Küçük vd., 2014). Bununla beraber, Türkçe metinlerde yapılan çalışmalar; İngilizce ve İspanyolca gibi dillerdeki metinler üzerinde yapılan çalışmalara kıyasla hem kapsam hem de sayı olarak oldukça sınırlı kalmaktadır.

6.2.1. VİT yöntemlerinin veri seti üzerine uygulaması

Türkçe ve İngilizce dilinde sosyal medya platformlardan elde edilen veriler üzerinde GAA, GDA algoritmaları kullanılarak konuşulan başlıca 3 konu

belirlenmişti. Bu algoritmalardan GDA algoritması tarafından belirlenen, konuların oluşmasında en etkin olan ilk 50 yorum incelenmiştir. Toplamda 150 yorumda en fazla geçen varlık isimleri belirlenmiştir. Belirlenen varlık isimleri SPARQL sorguları aracılığıyla DBPedia projesi içerisinde sorgulanmıştır. Sorgu sonucunda elde edilen varlık isimlerine karşılık gelen tanımlamalar çıkarılmıştır. VİT yöntemlerinin veri setleri üzerine uygulanması ve DBPedia üzerinden açıklamaların çekilmesine ait işlem basamakları Şekil 6.2' de gösterilmektedir.



Şekil 6.2. İşlem basamakları

6.2.2. İngilizce yorumların varlık isimlerinin belirlenmesi

Konu başlıklarının oluşmasında, kullanıcılar tarafından yapılan yorumların etkisi bulunmaktadır. Belirlenen 3 konu başlığına en fazla etki yapan 50 yorum belirlenmiştir. Toplam 150 yorum üzerinden VİT işlemi uygulanmıştır. İngilizce dili için "Spacy" (Spacy, 2019) kütüphanesi kullanılarak VİT işlemi yapılmıştır. 150 yorumda en fazla bahsedilen 5 varlık ve geçme sayıları Çizelge 6.2' de

gösterilmektedir. En fazla 5 varlık ismine bakıldığında Apple gibi bir teknoloji firması ve yine Apple tarafından piyasaya sürülen Ios işletim sistemi gibi teknoloji ile alakalı varlık isimleri belirlenmiştir. İlk sırada çıkan ”one” kelimesi yorumlar içerisinde manuel olarak tarandığında tek başına kullanıldığı gibi, ayrıca “Samsung” firmasının bir ürününü temsil etmek için kullanıldığı yorumlara bakıldığında görülmüştür. ”Russian” ve “The Federal Communication Commision” varlıkları ise GDA algoritmasının ortaya koyduğu başlıklarla ilgili olarak yoğun bir şekilde kullanıldığı görülmektedir. GDA algoritması tarafından Amerikan Seçimleri ve Yasa Teklifinin konuşulduğu başlıkları tespit edilmişti. “Russian” varlığı Rusya’ nın Amerikan Seçimlerine müdahalesi ile ilgili yorumlarda sık olarak kullanılmaktadır. “The Federal Communication Commision” varlığı da Yasa Teklifini içeren yorumlarda sık geçtiği tespit edilmiştir.

Çizelge 6.2. İngilizce veri seti için varlık isimleri ve sayıları

Varlık İsmi	**Geçme Sayısı**
One	53
Russian	26
Fcc	21
Ios	19
Apple	17

Çizelge 6.3’ de 1 nolu başlık için yani Amerika Seçimleri olarak etiketlenen konu ile ilgili varlık isimleri ve sayıları gösterilmektedir. Başlığın belirlenmesinde en etkili olan 50 yorum üzerinden bulunan varlık isimlerine bakıldığında “Russian” varlığının en çok geçtiği görülmektedir. Amerika seçimleri sonrasında çok konuşulan konulardan biri olan, seçim sonuçlarında Rusya’ nın etkisinin olup olmadığı ile ilgili yorumlardan dolayı 1 nolu başlığın ilk 50 yorumunda en fazla bulunan varlık olduğu düşünülmektedir. İkinci olarak en fazla bulunan varlık adının da yine seçimler ile ilgili olduğu görülmektedir. “Republican” kelimesi, Amerika’ da Cumhuriyetçi Parti için ve bu partiyi destekleyenleri belirtmek için kullanılan bir tanımlama olarak bilinmektedir. Bir diğer bulunan varlık ismi ise “Trump” olarak görülmektedir. Konu ile direkt bağlantılı olduğu ve seçimlerde adaylardan biri olduğu bilinmektedir. Seçimi kazanınca Rusya’ nın seçimlere müdahalesi konuşulmaya başlanmıştır. “One” varlığı da teknoloji ile ilgili

başlıklarda ağırlıklı olarak görünmesine burada da bulunmaktadır. Bu başlık altında bulunan ilk 50 yorum incelendiğinde "one" teriminin aslında teknoloji ile bağlantılı yorumlarda değil, sayı bildiren bir terim olarak kullanıldığı görülmektedir. Son bulunan varlık ise "Government" yine konu başlığı ile alakalı ve seçimler sonucu oluşacak olan varlığı temsil etmektedir. Konu1 için bakıldığında GDA algoritması tarafından bulunan konu başlığı ile, yorumlardan çıkarılan varlık isimlerinin alakalı olduğu görülmektedir.

Çizelge 6.3. Konu 1 ile ilgili varlık isimleri ve sayıları

Varlık İsmi	**Geçme Sayısı**
Russian	22
Republican	11
Trump	7
One	5
Government	4

Çizelge 6.4' de 2 nolu konu başlığı olan "Mobil Cihazlar" olarak etiketlenen konu ile ilgili varlık isimleri ve sayıları gösterilmektedir. En etkili 50 yorum üzerinden yapılan varlık isimi ve sayı belirleme işleminde elde edilen bütün varlık isimlerinin mobil cihazlar konu başlığı ile uyumlu olduğu görülmektedir. En fazla geçen varlık ismi olarak "One" terimi bulunmuştur. Buradaki yorumlar incelendiğinde "Samsung" teknoloji firmasına ait mobil cihazların serisi olan "One" olarak kullanıldığı görülmüştür. Daha sonraki varlıklar incelendiğinde "ios", "apple" ve "4s" varlıkları olduğu görülmektedir. Bu terimler Apple firmasını bu firmanın mobil cihazlarda kullanılan işletim sistemi olan ios ve Apple firmasına ait olan mobil telefonlardan bir model olduğu görülmektedir. Son bulunan varlık ise ilk 50 yorumda 6 defa bulunan "Amazon" varlığı mobil cihazlar ile bir ilgisi olmamasına rağmen teknoloji firmalarının başında gelen farklı alanlarda ürünü olan bir firmadır. Aslında konu başlığı ile uzak bir varlık değildir. Aynı alanda hizmet etmektedir. Konu2 ile ilgili olarak bulunan varlık isimlerinin konu ile ilgili oldukları görülmektedir.

Çizelge 6.4. Konu 2 ile ilgili varlık isimleri ve sayıları

Varlık İsmi	Geçme Sayısı
One	44
Ios	15
Apple	13
4s	7
Amazon	6

Çizelge 6.5' de 3 nolu başlık için yani Amerika çıkan yasa teklifi ile ilgili konu ile alakalı varlık isimleri ve sayıları gösterilmektedir. Bulunan varlıkların hepsinin konu ile alakalı olduğu görülmektedir. "Fcc" bir komisyon olarak yasa çıktıktan sonra belirli bir süre içinde belli işleri yapmakla görevlendirilmiştir. "Broadband Access Services" ise çıkan yasanın içeriği ile ilgili terimlerden oluşmaktadır. "Comcast" ise kablolu tv, internet ve mobil telefon hattı hizmeti sağlayan bir şirkettir. Çıkarılan yasa ile çalışma alanları nedeniyle doğrudan bağlantısı bulunmaktadır. "Tim Wu" ise Columbia Üniversitesi Hukuk Fakültesinde profesörlük görevi yapan akademisyen ve fikir insanı olarak bilinmektedir. Bu konu ile ilgili geçmişte yayınlanmış makalelerinden yorumlar sıkça bahsedildiği görülmektedir. "Congress" varlığı ise yasa ile ilgili kararın alındığı yer olmasına bakımından yorumlarda geçmektedir. Konu3 ile ilgili ilk 50 yorumda tespit edilen varlık isimlerinin de konu başlığı ile ilgii olduğu gözlemlenmiştir.

Çizelge 6.5. Konu 3 ile ilgili varlık isimleri ve sayıları

Varlık İsmi	Geçme Sayısı
Fcc	17
Broadband Access Services	12
Comcast	11
Tim Wu	8
Congress	6

6.2.3. Türkçe yorumların varlık isimlerinin belirlenmesi

İngilizce dilinde uygulandığı gibi, 3 konu başlığının oluşmasında etkili ilk 50 yorum olmak üzere toplam 150 yorum üzerinden varlık ismi tanıma işlemi uygulanmıştır. Türkçe dili için PolyGlot (PolyGlot, 2019) kütüphanesi kullanılarak varlık ismi tanıma işlemi yapılmıştır. 150 yorumda en fazla bahsedilen 5 varlık ve geçme sayıları Çizelge 6.6' da gösterilmektedir. Türkçe' de

belirlenen varlık isimleri incelendiğinde en fazla geçen varlık isminin son zamanlarda bütün dünyanın sorunu olan "Corona" olduğu görülmektedir. Pandemi hastalığının, teknolojisi ile ilişkisi ilk bakışta görülemese de, son zamanlarda hastalığın "5G" teknolojisi nedeniyle yayıldığı bilgisi ortaya atılmıştır. Hatta bazı ülkelerde 5G vericilerine zarar verilmiştir. EkşiSözlük platformunda kullanıcıların 5G' nin pandemi üzerine etkilerini yoğun şekilde "teknoloji" alanında konuştuğunu gözlemlenebilmektedir. Varlık isimleri sıralamasında ikinci olarak "Apple" firması bulunmaktadır. Apple firması teknolojik ürünleri olan bir firma olması bakımından "teknoloji" başlığı ile uyumludur. Diğer varlıklara bakıldığında "ABD" ve "Hong Kong" gibi varlık isimleri ile karşılaşılmaktadır. Bu ülke isimlerinin de pandemi hastalığı ile ilgili yorumlarda sıklıkla geçtiği görülmektedir. Son bulunan varlık ismi olarak ise Türkiye' de yayın yapan özel bir platform olan "Digitürk" olduğu görülmektedir.

Çizelge 6.6. Türkçe veri seti için varlık isimleri ve sayıları

Varlık İsmi	**Geçme Sayısı**
Corona	41
Apple	25
ABD	19
Hong Kong	14
Digitürk	12

Varlık isimleri her konu başlığı için incelenmiştir. Her konu başlığının oluşmasında etkili olan 50 yorum incelenerek en fazla geçen varlık isimleri bulunmuştur. İlk konu başlığımızı "oyunlar" olarak etiketlenmişti. "Oyunlar" konu başlığına ait varlık isimleri Çizelge 6.7' de gösterilmektedir. Bulunan varlık isimleri incelendiğinde teknoloji firmalarının isimleri görülmektedir. "Apple" ve "Microsoft" dünyanın önde gelen teknoloji firmalarındandır. "Ios", Apple firmasının ürettiği mobil cihazlarında kullanılan işletim sistemidir. Bulunan varlık isimlerinden konu başlığına en yakın olan varlık ismi "Fifa" terimidir. "Fifa" dünya üzerinde yaygın olarak oynanmakta olan bir futbol oyunudur. EkşiSözlük kullanıcıları tarafından belirlenen ilk 50 yorum içerisinde en fazla kullanılan oyun ismi olarak görülmüştür. Son bulunan varlık ismi ise "Linux" olarak görülmektedir. "Linux" açık kaynak olarak kullanılan bir işletim sistemidir. Analiz sonuçlarına göre bu başlık altında az sayıda da olsa kullanıldığı görülmektedir.

Çizelge 6.7. Konu1 ile ilgili varlık isimleri ve sayıları

Varlık İsmi	Geçme Sayısı
Apple	19
Ios	9
Fifa	8
Microsoft	4
Linux	2

GDA algoritması ile bulunan ikinci konu başlığı "Koronavirüs" olarak belirlenmişti. 2 nolu başlığına ait etkili 50 yorum incelenip bulunan varlık isimleri Çizelge 6.8' te gösterilmektedir. Bulunan varlık isimlerinin hepsinin konu başlığı ile uyumlu olduğu görülmektedir. En fazla geçen sayı olarak net bir şekilde konu başlığını da yansıtan varlık ismi "Corona" olarak görülmektedir. "ABD", "Hong Kong" ve "Kore" ise hastalık ile ilgili yorumlarda geçen ülke isimleri olarak tespit edilmiştir. "Wuhan" ise bilindiği üzere hastalığın ilk görüldüğü kent olarak bilinmektedir. Bu nedenle incelenen yorumlar içerisinde kullanıldığı görülmektedir.

Çizelge 6.8. Konu2 ile ilgili varlık isimleri ve sayıları

Varlık İsmi	Geçme Sayısı
Corona	39
ABD	16
Hong Kong	13
Wuhan	11
Kore	8

Türkçe dilinde 3. konu başlığımız "Müşteri Hizmetleri Şikayetleri" olarak önceki bölümde etiketlenmişti. 3 nolu konu başlığı için bulunan varlık isimleri Çizelge 6.9' da gösterilmektedir. Bulunan varlık isimleri incelendiğinde en fazla geçen varlık isimleri olarak "Sms" ve "Çözüm" gibi müşteri hizmetlerine ait şikayet yorumları içerisinde geçen varlıklar görülmektedir. Bu konuda en fazla geçen varlık ismi olarak özel bir yayın platformu olan "Digitürk" görülmektedir. "İstanbul" ismi en fazla geçen varlıklar arasındadır. Yorumlarda müşteri şikâyetlerinin İstanbul şehri üzerine yoğunlaştığı görülmektedir. Teknoloji firması olan "Apple" firmasının da bu başlık altında bulunduğu görülmektedir.

Çizelge 6.9. Konu3 ile ilgili varlık isimleri ve sayıları

Varlık İsmi	Geçme Sayısı
Digiturk	11
İstanbul	6
Sms	4
Apple	3
Çözüm	2

3 konu başlığı için yorumlardan elde edilen varlık isimlerinin genel itibariyle ilgili konu başlıkları ile uyumlu olduğu gözlenmiştir.

7. ARAŞTIRMA BULGULARI

Varlık ismi tanıma ile bulunan Türkçe ve İngilizce varlık isimleri hakkında DBPedia projesi üzerinden SPARQL sorgulama dili ile sorgulama işlemi yapılarak, konuşulan yorumlar içerisinden elde edilen varlık isimlerinin açıklamalarının gösterilmesi amaçlanmaktadır.

Çizelge 6.10' da elde edilen Türkçe dili için varlık isimlerinin DBPedia ile eşleştirilmesi işlemi gerçekleştirilmiştir. Bu işlemler için SPARQL sorgu dili yardımıyla, yorumlardan elde edilen 5 varlık için sorgu yapılmıştır. Elde edilen 5 varlık isminin 4 tanesinin karşılığına ulaşılmıştır. "ABD" terimi kullanım şeklinden dolayı bulunamamıştır. Bulunan tanımlama incelendiğinde "Apple" terimi hariç diğerlerinde beklenen tanımalamalara erişildiği görülmektedir.

"Apple" terimi meyve olarak eşleştirilmiştir. Sorgu sonucu beklediğimiz sonuç olan "Apple" teknoloji firması ile açıklamanın geri dönmemesinin sebebi "Apple" firmasının DBPedia üzerinde "Apple Inc" olarak kayıtlı olmasıdır. Bakıldığında yanlış bir sonuç dönmemiştir. "Apple" bir meyve ve meyve ağacıdır. Fakat "teknoloji" başlığı ile yorumlarda geçen varlık ismi ile aynı bilgi değildir.

"Hong Kong" terimi bu haliyle DBPedia üzerinde bulunmamaktadır. DBPedia projesi üzerinde genelde birden fazla kelime barındıran bilgiler "_" işareti ile ayrılmış olduğu görülmüştür. Bu nedenle "Hong_Kong" olarak arama işlemi yapılmış ve sonuca ulaşılmıştır.

Çizelge 7.1. Türkçe veri setinde bulunan varlık isimleri için DBPedia bulgusu

Varlık İsmi	DBPedia Bulgusu
Corona	A corona (Latin, 'crown') is an aura of plasma that surrounds the sun and other stars. The Sun's corona extends millions of kilometres into space and is most easily seen during a total solar eclipse, but it is also observable with a coronagraph. The word "corona" is a Latin word meaning "crown", from the Ancient Greek κορώνη (korōnè, "garland, wreath"). The high temperature of the Sun's corona gives it unusual spectral features, which led some in the 19th century to suggest that it contained a previously unknown element, "coronium". Instead, these spectral features have since been explained by highly ionized iron (Fe-XIV). Bengt Edlén, following the work of Grotrian (1939), first identified the coronal spectral lines in 1940 (observed since 1869) as transitions from low-lying metastable levels of the ground configuration of highly ionised metals (the green Fe-XIV line at 5303 Å, but also the red line Fe-X at 6374 Å). These high stages of ionisation indicate a plasma temperature in excess of 1,000,000 kelvin, much hotter than the surface of the sun. Light from the corona comes from three primary sources, from the same volume of space. The K-corona (K for kontinuierlich, "continuous" in German) is created by sunlight scattering off free electrons; Doppler broadening of the reflected photospheric absorption lines spreads them so greatly as to completely obscure them, giving the spectral appearance of a continuum with no absorption lines. The F-corona (F for Fraunhofer) is created by sunlight bouncing off dust particles, and is observable because its light contains the Fraunhofer absorption lines that are seen in raw sunlight; the F-corona extends to very high elongation angles from the Sun, where it is called the zodiacal light. The E-corona (E for emission) is due to spectral emission lines produced by ions that are present in the coronal plasma;

	it may be observed in broad or forbidden or hot spectral emission lines and is the main source of information about the corona's composition.
Apple	The apple tree (Malus pumila, commonly and erroneously called Malus domestica) is a deciduous tree in the rose family best known for its sweet, pomaceous fruit, the apple. It is cultivated worldwide as a fruit tree, and is the most widely grown species in the genus Malus. The tree originated in Central Asia, where its wild ancestor, Malus sieversii, is still found today. Apples have been grown for thousands of years in Asia and Europe, and were brought to North America by European colonists. Apples have religious and mythological significance in many cultures, including Norse, Greek and European Christian traditions. Apple trees are large if grown from seed. Generally apple varieties are propagated by grafting onto rootstocks, which control the size of the resulting tree. There are more than 7,500 known cultivars of apples, resulting in a range of desired characteristics. Different cultivars are bred for various tastes and uses, including cooking, eating raw and cider production. Trees and fruit are prone to a number of fungal, bacterial and pest problems, which can be controlled by a number of organic and non-organic means. In 2010, the fruit's genome was sequenced as part of research on disease control and selective breeding in apple production. Worldwide production of apples in 2013 was 80.8 million tonnes, with China accounting for 49% of the total.
ABD	Null
Hong Kong	Hong Kong (Chinese: 香港; literally: "Fragrant Harbour" or "Incense Harbour"), officially the Hong Kong Special Administrative Region of the People's Republic of China, is an autonomous territory on the Pearl River Delta in East Asia. The mainland Chinese province of Guangdong borders the territory to the north. With a total land area of 1,106 square kilometres (427 sq mi) and a population of over 7.3 million of various nationalities, it ranks as the world's fourth most densely populated sovereign state or territory. After the First Opium War (1839–42), Hong Kong became a British colony with the perpetual cession of Hong Kong Island, followed by the Kowloon Peninsula in 1860 and a 99-year lease of the New Territories from 1898. Hong Kong was later occupied by Japan during the Second World War until British control resumed in 1945. In the early 1980s, negotiations between the United Kingdom and China resulted in the 1984 Sino-British Joint Declaration, which paved way for the transfer of sovereignty of Hong Kong in 1997, when it became a special administrative region (SAR) with a high degree of autonomy. Under the principle of "one country, two systems", Hong Kong maintains a separate political and economic system from China. Except in military defence and foreign affairs, Hong Kong maintains its independent executive, legislative and judiciary powers. In addition, Hong Kong develops relations directly with foreign states and international organisations in a broad range of "appropriate fields". Hong Kong is one of the world's most significant financial centres, with the highest Financial Development Index score and consistently ranks as the world's most competitive and most laissez-faire economic entity in the World Competitiveness Yearbook. Its legal tender, the Hong Kong dollar, is the world's 13th most traded currency. Hong Kong's tertiary sector dominated economy is characterised by simple taxation with a competitive level of corporate tax and supported by international confidence in its independent judiciary system where the rule of law, not rule by law, applies to legal, contractual proceedings. However, while Hong Kong has one of the highest per capita incomes in the world, it suffers from the most severe income inequality among developed economies. Hong Kong is renowned for its deep natural harbour, which enables ready access by international cargo ships, and its skyline, with a very high density of skyscrapers; the territory boasts the second largest number of high rises of any city in the world. It has a very high Human Development Index ranking and the world's longest life expectancy. Over 90% of the population makes use of well-developed public transportation. Seasonal air pollution with origins from neighbouring industrial areas of Mainland China, which adopts loose emissions standards, has resulted in a high level of atmospheric particulates.
Digitürk	Digiturk is a Turkish satellite television provider founded in 1999, with services starting in mid-2000. They offer both national cable television channels and their own channels, national radio, and music streams of different genres. Digiturk is also the current owner of the broadcasting rights of Turkish Super League. In addition to Turkey, they offer service throughout Europe, mainly for members of the Turkish diaspora. Reportedly, they have over 3.5 million subscribers worldwide. Their service is provided from Eutelsat W3A, positioned some 35 degrees west of the more traditionally used Türksat; and is encrypted via Cryptoworks and Irdeto conditional access systems. Digiturk relays 38 national channels, 22 news channels, 23 film and series channels, 20 sport channels, 14 children channels, 12 music channels, 15 documentary channels, 9 entertainment and life style channels and 65 other international channels. beIN Media Group, another Qatar origined media group acquired Digiturk on July 13, 2015. In October 2015, Digiturk decided to exclude some channels which include Samanyolu TV, Samanyolu Haber TV, Bugun TV, Yumurcak TV(a TV channel for children), and Irmak TV.

Çizelge 6.11' de elde edilen İngilizce dili için bulunan varlık isimlerinin DBPedia projesi ile eşleştirilmesi işlemi gerçekleştirilmiştir. Bu işlemler için SPARQL sorgu dili yardımıyla elde edilen 5 varlık için sorgu yapılmıştır. 5 varlığın 4 tanesinin karşılığına ulaşılmıştır. Diğerleri ile ilgili elde edilen sonuçlara göre

herhangi bir sonuç bulunamamıştır. “Apple” terimi meyve olarak eşleştirilmiştir. “One” terimi bulunamamıştır.

“fcc” terimi arama işlemi yapılırken terimin açılımı kullanılarak tanımlamalara erişilebildiği için “Federal Communications Commission” olarak arama işlemi yapılmıştır.

Çizelge 7.2. İngilizce veri setinde bulunan varlık isimleri için DBPedia bulgusu

Varlık İsmi	**DBPedia Bulgusu**
One	Null
Russian	Russia (/ˈrʌʃə/; Russian: Росси́я, tr. Rossija; IPA: [rɐˈsʲijə]; from the Greek: Ρωσία — Rus'), also officially known as the Russian Federation (Russian: Росси́йская Федера́ция, tr. Rossijskaja Federacija; IPA: [rɐˈsʲijskəjə fʲɪdʲɪˈratsʲɪjə]), is a transcontinental country in Eurasia. At 17,075,200 square kilometres (6,592,800 sq mi), Russia is the largest country in the world, covering more than one eighth of Earth's inhabited land area, and the ninth most populous, with over 146.6 million people at the end of March 2016. Extending across the entirety of northern Asia and much of Eastern Europe, Russia spans eleven time zones and incorporates a wide range of environments and landforms. From northwest to southeast, Russia shares land borders with Norway, Finland, Estonia, Latvia, Lithuania and Poland (both with Kaliningrad Oblast), Belarus, Ukraine, Georgia, Azerbaijan, Kazakhstan, China, Mongolia, and North Korea. It shares maritime borders with Japan by the Sea of Okhotsk and the U.S. state of Alaska across the Bering Strait. The nation's history began with that of the East Slavs, who emerged as a recognizable group in Europe between the 3rd and 8th centuries AD. Founded and ruled by a Varangian warrior elite and their descendants, the medieval state of Rus arose in the 9th century. In 988 it adopted Orthodox Christianity from the Byzantine Empire, beginning the synthesis of Byzantine and Slavic cultures that defined Russian culture for the next millennium. Rus' ultimately disintegrated into a number of smaller states; most of the Rus' lands were overrun by the Mongol invasion and became tributaries of the nomadic Golden Horde in the 13th century. The Grand Duchy of Moscow gradually reunified the surrounding Russian principalities, achieved independence from the Golden Horde, and came to dominate the cultural and political legacy of Kievan Rus'. By the 18th century, the nation had greatly expanded through conquest, annexation, and exploration to become the Russian Empire, which was the third largest empire in history, stretching from Poland on the west to Alaska on the east. Following the Russian Revolution, the Russian Soviet Federative Socialist Republic became the largest and leading constituent of the Union of Soviet Socialist Republics, the world's first constitutionally socialist state. The Soviet Union played a decisive role in the Allied victory in World War II, and emerged as a recognized superpower and rival to the United States during the Cold War. The Soviet era saw some of the most significant technological achievements of the 20th century, including the world's first human-made satellite and the launching of the first humans in space. By the end of 1990, the Soviet Union had the world's second largest economy, largest standing military in the world and the largest stockpile of weapons of mass destruction. Following the partition of the Soviet Union in 1991, fourteen independent republics emerged from the USSR; as the largest, most populous, and most economically developed republic, the Russian SFSR reconstituted itself as the Russian Federation and is recognized as the continuing legal personality and sole successor state of the Soviet Union. It is governed as a federal semi-presidential republic. The Russian economy ranks as the twelfth largest by nominal GDP and sixth largest by purchasing power parity in 2015. Russia's extensive mineral and energy resources are the largest such reserves in the world, making it one of the leading producers of oil and natural gas globally. The country is one of the five recognized nuclear weapons states and possesses the largest stockpile of weapons of mass destruction. Russia is a great power and a permanent member of the United Nations Security Council, as well as a member of the G20, the Council of Europe, the Asia-Pacific Economic Cooperation

	(APEC), the Shanghai Cooperation Organisation (SCO), the Organization for Security and Co-operation in Europe (OSCE), and the World Trade Organization (WTO), as well as being the leading member of the Commonwealth of Independent States (CIS), the Collective Security Treaty Organization (CSTO) and one of the five members of the Eurasian Economic Union (EEU), along with Armenia, Belarus, Kazakhstan, and Kyrgyzstan.
Fcc	The Federal Communications Commission (FCC) is an independent agency of the United States government, created by Congressional statute (see 47 U.S.C. § 151 and 47 U.S.C. § 154) to regulate interstate communications by radio, television, wire, satellite, and cable in all 50 states, the District of Columbia and U.S. territories. The FCC works towards six goals in the areas of broadband, competition, the spectrum, the media, public safety and homeland security, and modernizing itself. The FCC was formed by the Communications Act of 1934 to replace the radio regulation functions of the Federal Radio Commission. The FCC took over wire communication regulation from the Interstate Commerce Commission. The FCC's mandated jurisdiction covers the 50 states, the District of Columbia, and Political divisions of the United States. The FCC also provides varied degrees of cooperation, oversight, and leadership for similar communications bodies in other countries of North America. The FCC is funded entirely by regulatory fees. It has an estimated fiscal-2016 budget of US$388 million. It has 1,720 federal employees.
Ios	iOS (formerly iPhone OS) is a mobile operating system created and developed by Apple Inc. exclusively for its hardware. It is the operating system that presently powers many of the company's mobile devices, including the iPhone, iPad, and iPod touch. It is the second most popular mobile operating system globally after Android by sales. iPad tablets are also the second most popular, by sales, against Android since 2013, when Android tablet sales increased by 127%. Originally unveiled in 2007 for the iPhone, it has been extended to support other Apple devices such as the iPod Touch (September 2007) and the iPad (January 2010). As of June 2016, Apple's App Store contained more than 2 million iOS applications, 725,000 of which are native for iPads. These mobile apps have collectively been downloaded more than 130 billion times. The iOS user interface is based upon direct manipulation, using multi-touch gestures. Interface control elements consist of sliders, switches, and buttons. Interaction with the OS includes gestures such as swipe, tap, pinch, and reverse pinch, all of which have specific definitions within the context of the iOS operating system and its multi-touch interface. Internal accelerometers are used by some applications to respond to shaking the device (one common result is the undo command) or rotating it in three dimensions (one common result is switching between portrait and landscape mode). Major versions of iOS are released annually. The current version, iOS 10, was released on September 13, 2016. It runs on the iPhone 5 and later, iPad (4th generation) and later, iPad Pro, iPad Mini 2 and later, and the 6th-generation iPod Touch. In iOS, there are four abstraction layers: the Core OS, Core Services, Media, and Cocoa Touch layers. iOS 10 dedicates around 1.8GB of the device's flash memory for itself.
Apple	The apple tree (Malus pumila, commonly and erroneously called Malus domestica) is a deciduous tree in the rose family best known for its sweet, pomaceous fruit, the apple. It is cultivated worldwide as a fruit tree, and is the most widely grown species in the genus Malus. The tree originated in Central Asia, where its wild ancestor, Malus sieversii, is still found today. Apples have been grown for thousands of years in Asia and Europe, and were brought to North America by European colonists. Apples have religious and mythological significance in many cultures, including Norse, Greek and European Christian traditions. Apple trees are large if grown from seed. Generally apple varieties are propagated by grafting onto rootstocks, which control the size of the resulting tree. There are more than 7,500 known cultivars of apples, resulting in a range of desired characteristics. Different cultivars are bred for various tastes and uses, including cooking, eating raw and cider production. Trees and fruit are prone to a number of fungal, bacterial and pest problems, which can be controlled by a number of organic and non-organic means. In 2010, the fruit's genome was sequenced as part of research on disease control and selective breeding in apple production. Worldwide production of apples in 2013 was 80.8 million tonnes, with China accounting for 49% of the total.

8. TARTIŞMA VE SONUÇLAR

Hayatın her alanında kullanılan teknolojik cihazlar, internetin yaygınlaşması gibi sebeplerden dolayı kişiler tarafından üretilen veri miktarı her geçen gün artmaktadır. Kullanılan teknolojik cihazların sensörlerinden gelen veriler, kişilerin kendi paylaştığı metin, resim verileri büyük veri adı verilen kavramı ortaya çıkarmıştır. Son zamanlarda büyük verinin algoritmalar yardımıyla işlenerek, bu büyük veriden anlamlı sonuçlar çıkarılması üzerine çalışmalar yoğunlaşmıştır.

Bu tez çalışmasında; kullanıcılar tarafından üretilen metin verileri üzerine farklı iki platform ve farklı iki dilde semantik analiz çalışması yapılmıştır. Farklı platformlar olarak EkşiSözlük ve Reddit sosyal medya platformları seçilmiştir. Dil olarak ise İngilizce ve Türkçe dili seçilmiştir. Reddit ve EkşisSözlük platformları çok farklı ve geniş alanlarda görüş paylaşımı yapılan platformlar olduğu için çalışma alanı olarak "teknoloji" alt başlığı seçilmiştir. Reddit platformu "technology" alt başlığından 245872 adet ingilizce kullanıcı yorumu toplanmıştır. EkşiSözlük platformundan "teknoloji" alt başlığından 282351 adet kullanıcı yorumu toplanmıştır. Her iki dilde toplanan yorumlar önişlemeden geçirilmiştir. Türkçe ve İngilizce dilinde yaygın olarak kullanılan durak kelimeler analiz sonuçlarını etkilememesi için veri setlerinden çıkarılmıştır.

Ön işlemden sonraki adım olarak GAA algoritması ile İngilizce ve Türkçe veri seti üzerinde konuşulan başlıca 3 konu belirlenmiştir. GAA ile yapılan analiz sonuçlarına göre Türkçe ve İngilizce dilinde elde edilen konu başlıkları karşılaştırılmıştır. GAA algoritması ile yapılan konu modelleme işlemi sonucu belirlenen konu başlıklarından ikisi birbirine yakın konular olduğu görülmektedir. Bu iki konuda da oyunlar filmler hakkında görüş paylaşımı yapılmaktadır. Diğer konu başlıkları ise Reddit üzerinde Amerika' daki yasa teklifi konuşulurken, EkşiSözlükte ise 5G teknolojisinin Koronavirüs üzerine etkilerinin konuşulduğu gözlemlenmektedir.

GDA algoritması kullanılarak aynı şekilde iki farklı platform ve iki dil için konu modelleme analizi 3 konu başlığı için yapılmıştır. Analiz sonuçlarına göre bu

algoritmaya göre farklılıklar görülmektedir. Çoğunlukla Amerika' da kullanılan Reddit platformunda kullanıcılar Amerikan Seçimleri başta olmak üzere mobil cihazlar ve Amerika' da Trump tarafından Mart ayında onaylanan yasa teklifini konuşurken, Türkiye' de yaygın şekilde kullanılan EkşiSözlük platformunda ise Oyunlar, Koronavirüs ve Müşteri Hizmetleri Şikâyetleri konusunda konuşmalar ağırlıklı olarak ön plana çıkmaktadır.

GAA ve GDA algoritmaları karşılaştırıldığında platformlar, diller kendi aralarında ve birbirleriyle karşılaştırıldıklarında ortak noktalar ve farklılıklar gözlemlenebilmektedir. Bu farklılıklar algoritmaların çalışma mantıkları arasındaki farktan kaynaklanmaktadır. GAA ve GDA algoritmasının EkşiSözlük platformunda bulduğu konu başlıkları birbirine yakındır. Bir konu başlığında farklılık görülmektedir. GAA algoritması Filmler ve Bitcoin üzerine konu bulurken, GDA algoritması ise Müşteri Hizmetleri Şikâyetleri üzerine bir konu başlığı bulmuştur. Reddit platformu için ise konu başlıklarında farklılık biraz daha fazla göze çarpmaktadır. GAA algoritmasına göre Oyunlar, Yasa Teklifi, Filmler üzerine konu başlıkları elde edilmiştir. GDA algoritması ise Amerika Seçimleri, Mobil Cihazlar ve Yasa Teklifi üzerine konu başlıklarına ulaşmıştır.

VİT, konu modelleme algoritmalarından GDA algoritmasında bulunan konu başlıklarında göre konuların oluşmasında etkili olan ilk 50 yorum bulunarak yapılmıştır. Dil desteğinden dolayı İngilizce için Spacy, Türkçe dili için PolyGlot kütüphaneleri kullanılarak VİT işlemi gerçekleştirilmiştir. VİT işlemi sonucunda her konu başlığı için ve 3 konu başlığının toplam 150 yorumu için en fazla geçen 5 varlık ismi bulunmuştur.

Ontoloji olarak bir WikiPedia projesi olan DBPedia projesi kullanılmıştır. Elde edilen varlık isimleri SPARQL sorgu dili yardımıyla DBPedia projesi üzerinden sorgulanmıştır. Otomatik olarak varlık isimleri DBPedia üzerinde arandığı için bazı varlık isimlerinin karşılıkları bulunamamıştır. Bulunan varlık isimlerinden "Apple" terimi bir meyve olan elma olarak bulunmuş ve bununla ilgili veriler geri döndürülmüştür. Fakat araştırıldığında "Apple Inc" olarak DBPedia sisteminde kayıtlı olduğu görülmektedir. Fakat kullanıcılar yorumlarında sadece terim ismi

kullandığı için beklenenden farklı bir sonuç elde edilmiştir. Diğer varlık isimleri sorunsuz bir şekilde elde edilmiştir.

Veri işlemenin öneminin giderek arttığı günümüzde şirketler, devletler, kurumlar politikalarını, işleyişlerini elde ettikleri verileri işleyerek ve bu veriden anlamlı sonuçlar çıkararak yönlendirmektedir. Bu tez çalışmasında farklı iki platform ve iki dil üzerinde yaklaşık 530.000 kullanıcı yorumu işlenerek farklı dillerde ve farklı platformlarda teknoloji başlığı altında hangi konuların konuşulduğu tespit edilmiştir. Konular arasındaki farklılıklar ve benzerlikler gösterilmiştir. Analiz sonuçlarına bakıldığında son zamanlardaki konuların ön plana çıktığı görülmektedir. Bunun asıl sebebi de güncel konular üzerine paylaşım miktarının fazla olmasıdır. Çalışmanın devamı niteliği taşıyacak çalışmalarda gerçek zamanlı analizler yapılarak, günlük/anlık olarak istenilen alan/alt alan konu başlıklarında konuşulan konular hakkında bilgi sahibi olunarak bu yönde kararlar alınabilir. Böylelikle anlık/günlük vb. veri toplama, veri analizi ve istihbarat gibi alanlarda çalışma derinleştirilebilir. Verinin elde edilmesi içinde farklı sosyal ağlar temel teşkil edebilir. Böylelikle farklı sosyal ağlar ve platformlar arası konu analizleri gerçekleştirilebilir.

KAYNAKLAR

Adalı, E., 2012. Doğal Dil İşleme.Türkiye Bilişim Vakfı Bilgisayar Bilimleri ve Mühendisliği Dergisi, 5(2).

Akçavlı, İ., 2012 W3C Linked Data ve DBPedia Nedir? Erişim Tarihi: 19.10.2019. http://ibrahimakcavli.com.tr/blog/w3c-linked-data-ve-dbpedia-nedir/

Alexa, 1996. Erişim Tarihi:20/03/2020. https://www.alexa.com/siteinfo/ reddit.com

Alexa, 1996. Erişim Tarihi:20.03.2020. https://www.alexa.com/siteinfo/ eksisozluk.com

Altay, O., Ulaş, M., 2018. Anlamsal Web Kullanılarak İlaç Ontolojisi Çıkarılması. Fırat Üniversitesi Mühendislik Bilimleri Dergisi, 30(1), 169-174.

Auer S., Bizer C., Kobilarov G., Lehmann J., Cyganiak R., Ives Z., 2007. DBpedia: A Nucleus for a Web of Open Data. The Semantic Web. ISWC 2007, 11-15 November, Busan, Korea, 722-735.

Balkan, K., Takcı, H., 2010. Obtaining Term Similarities on Concept Extraction Study. National Conference on Electrical, Electronics and Computer Engineering, 2-5 Aralık, Bursa, 78-582.

Battal, A., 2009. Semantik Web ile Geliştirilen Bir Televizyon Program Öneri Sistemi. TOBB Ekonomi ve Teknoloji Üniversitesi, Fen Bilimleri Enstitüsü, Yüksek Lisans Tezi, 78s, Ankara.

Berners-Lee, T., Hendler, J., Lassila, O., 2001. The Semantic Web. Scientific American, 284(5): 34- 43.

Berners-Lee, T., 2006. Linked Data - Design Issues. Erişim Tarihi: 19.04.2019 http://www.w3.org/DesignIssues/ LinkedData.html.

Bernstein, A., Kaufmann, E., Göhring, A., Kiefer, C., 2005. Querying Ontologies: A Controlled English Interface for End-Users. In: International Semantic Web Conference, 6-10 November, Galway, Ireland, 112-126.

Bhat, M.R., Kundroo, M.A., Tarray, T.A., Agarwal, B., 2019. Deep LDA: A New Way to Topic Model. Journal of Information and Optimization Sciences, 1–12.

Bizer, C., Heath, T., Berners-Lee, T., 2009. Linked Data-The Story So Far. International Journal on Semantic Web and Information Systems, 5(3), 1-22.

Bizer, C., Lehmann, J., Kobilarov, G., Auer, S., Becker, C., Cyganiak, R., Hellmann, S., 2009. Dbpedia-A Crystallization Point for the Web of Data. Journal of Web Semantics: Science, Services and Agents on the World Wide Web, 7(3), 154-165.

Blei, D.M., Ng, A. Y., Jordan, M.I. 2003. Latent Dirichlet Allocation. Journal of Machine Learning Research, 3, 993-1022.

Blei, D.M., 2012. Probabilistic Topic Models. Communications of the ACM, 55(4), 77.

Code, 2020. Erişim Tarihi: 17.03.2020. https://code.visualstudio.com/

Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., Kuksa, P., 2011. Natural Language Processing (Almost) from Scratch. Journal of Machine Learning Research, 12, 2493- 2537.

Creamer, M., 2008. It's Web 3.0 and Someone Else's Content is King. Erişim Tarihi: 17.01.2018. http://adage.com/article/digital/web-3-0-s-content-king/126364/

Conover, M., Ferrara, E., Menczer, F., & Flammini, A., 2013. The Digital Evolution of Occupy Wall Street. PLoS ONE8, 8(5), 1-6.

Çalışkan, M., Mencik, Y., 2015. The New Face of a Changing World: Social Media. Akademik Bakış Uluslararası Hakemli Sosyal Bilimler Dergisi, 50, 254-277.

Çoban, H., D., 2010. Bir Anlamsal Web Uygulaması Olarak Türkiye Organik Tarım Bilgi Portalı Tasarımı. Çukurova Üniversitesi, Fen Bilimleri Enstitüsü, Yüksek Lisans Tezi, 87s, Adana.

Demchenko, Y., Laat, C., Membrey, Peter., 2014. Defining Architecture Components of the Big Data Ecosystem. 2014 International Conference on Collaboration Technologies and Systems, 19-23 May, Minneapolis, USA, 104-112.

Delibaş, A., 2008. Doğal Dil İşleme ile Türkçe Yazım Hatalarının Denetlenmesi. İstanbul Teknik Üniversitesi, Fen Bilimleri Enstitüsü, Yüksek Lisans Tezi, 94s, İstanbul.

DBPedia, 2004. Erişim Tarihi: 08.09.2019 www.dbpedia.org

Domo, (2011). Erişim Tarihi: 14.12.2019 https://www.domo.com/learn/data-never-sleeps-7

Ekinci, E., Omurca, İ.S., 2016. Ürün Özelliklerinin Konu Modelleme Yöntemi ile Çıkarılması, Türkiye Bilişim Vakfı Bilgisayar Bilimleri ve Mühendisliği Dergisi, 9(1), 51.

Ekinci, E., Omurca, S.O., Kırık, E., Taşçı, Ş., 2020. Tıp Veri Kümesi için Gizli Dirichlet Ayrımı. Dokuz Eylül Üniversitesi Fen ve Mühendislik Dergisi, 22(64), 67-80.

Erdur, R., Alatlı, O., 2013. Bağlı Veri Bulutu Üzerinde Mobil Uygulamalar Geliştirme için Bir Altyapı. Gazi Üniversitesi Fen Bilimleri Dergisi Part C: Tasarım ve Teknoloji, 1, 129-138.

Gensim, 2009. Erişim Tarihi: 15.01.2019. https://radimrehurek.com/gensim

Guo, Y., Barnes, S. J., Jia, Q., 2017. Mining Meaning from Çevrimiçi Ratings and Reviews: Tourist Satisfaction Analysis Using Latent Dirichlet Allocation. Tourism Management, 59, 467–483.

Guo, Y., Li, Y., Shao, Z., 2011. A Semantic Processing Model for Sentence Understanding Based on Cognitive Learning. In: 2011 IEEE 3rd International Conference on Communication Software and Networks, 27-29 May 2011, Xi'an, China, 106-110.

Guo, R., Ren, F., 2009. Towards The Relationship Between Semantic Web and NLP. In: 2009 IEEE International Conference on Natural Language Processing and Knowledge Engineering. 21-23 August, Beijing, 1-8.

Hatipoğlu, A., Omurca, S., 2015. Türkçe Metin Özetlemede Melez Modelleme. Dokuz Eylül Üniversitesi Mühendislik Fakültesi Fen ve Mühendislik Dergisi, 17, 95-108.

He, W., Zha, S., Li, L., 2013. Social Media Competitive Analysis and Text Mining: A Case Study in The Pizza Industry. Int J Inf Manage, 33(3), 464-472.

Hidayatullah, A.F., Ma'arif, M.R., 2017.Road Traffic Topic Modeling on Twitter Using Latent Dirichlet Allocation. In 2017 International Conference on Sustainable Information Engineering and Technology (SIET), 24-25 December, Batu, Indonesia,47–52.

Jo, Y., Oh, A., 2011. Aspect and Sentiment Unification Model for Çevrimiçi Review Analysis, In Proceedings of 4th ACM International Conference on Web Search and Data Mining, 9-12 February, Hong Kong, 815-824.

Kakisim, Y. O. Ipek and I. Sogukpinar, 2016. Suspect and Popular Tag Detection Model for Social Media. 24th Signal Processing and Communication Application Conference (SIU), 16-19 May, Zonguldak, Turkey, 893-896.

Karademirci, O., 2008. Anlamsal Web Teknikleri Kullanılarak GPS Tabanlı Bağlam Bilinçli Mobil Uygulama. Gazi Üniversitesi, Fen Bilimleri Enstitüsü, Yüksek Lisans Tezi, 97s, Ankara.

Kherwa, P., Bansal, P., 2017. Latent Semantic Analysis: An Approach to Understand Semantic of Text. International Conference on Current Trends in Computer, Electrical, Electronics and Communication, 8-9 September, Karnataka, India, 870-874.

Kuzu, R.S., Haznedaroğlu, A., Arslan, M.L.,2012. Topic Identification for Turkish Call Center Records. 2012 20th Signal Processing and Communications Applications Conference (SIU), 18- 20 April 2012, Fethiye, Muğla, Turkey,1-4.

Küçük, D., Jacquet, G., Steinberger, R., 2014. Named Entity Recognition on Turkish Tweets. Language Resources and Evaluation Conference, 26-31 May, Reykjavik, Iceland 450–454.

Küçük, D., Yazıcı, A., 2012. A Hybrid Named Entity Recognizer for Turkish. Expert Systems with Applications, 39(3): 2733-2742.

Küçük, D., Yazıcı, A., 2009. Named Entity Recognition Experiments on Turkish Texts. International Conference on Flexible Query Answering Systems, 26-28 October, Roskilde, Denmark, 524–535.

Landauer, T.K., Dutnais, S.T., 1997. A Solution to Platos Problem: The Latent Semantic Analysis Theory of Acquisition, Induction, and Representation of Knowledge. Psychol Revision, 211–240.

Lee, D.D., Seung, H.S., 2001. Algorithms for Non-Negative Matrix Factorization. Neural Information Processing Systems, 556-562.

Li, F., Huang, M., Zhu, X., 2010. Sentiment Analysis with Global Topics and Local Dependency. In Proceedings of the 24th AAAI Conference on Artificial Intelligence, March 17–20, Marina del Ray, CA, USA, 1371-1376.

Lu, Y., Mei, Q., Zhai, C., 2011. Investigating Task Performance of Probabilistic Topic Models: An Empirical Study of PLSA and LDA. Inf Retr Boston, 14(2), 178–203.

Marrero, M., Urbano, J., Sánchez-Cuadrado, S., Morato, J., Gómez-Berbís, J.M., 2013. Named Entity Recognition: Fallacies, Challenges and Opportunities. Computer Standards & Interfaces, 35(5): 482-489.

Merchant, K., Pande, Y., 2018. NLP Based Latent Semantic Analysis for Legal Text Summarization. 2018 International Conference on Advances in Computing, Communications and Informatics, 16-18 April, Karnataka, India, 1-8.

National Institute of Standards and Technology (NIST), 2015.NIST Big Data Interoperability Framework: Definitions. Erişim Tarihi: 01.11.2019. https://bigdatawg.nist.gov/_uploadfiles/NIST.SP.1500-1.pdf

Nltk, 2009. Erişim Tarihi: 15.08.2019 www.nltk.org.

Özkaya, S., Diri, B., 2011.Türkçe Metinlerde Şartlı Rasgele Alanlarla Varlık İsmi Tanıma. IEEE Sinyal İşleme ve İletişim Uygulamaları Kurultayı, 20-22 Nisan, Antalya, 662–665.

Özpala, A., Köker, R., 2008. İş Başvuru Sisteminde Anlamsal Ağ Servislerinin Kullanımı. 1. Mühendislik ve Teknoloji Sempozyumu, 24-25 Nisan, Ankara, 382-384.

Pavlinek, M., Podgorelec, V., 2017. Text Classification Method Based on Self-Training and LDA Topic Models. Expert Systems with Applications, 80, 83-93.

Pollock, J.T., 2009. Semantic Web for Dummies. Wiley Publishing, 482s, New Jersey, USA.

PolyGlot, 2014. Erişim Tarihi: 29.11.2019 https://polyglot.readthedocs.io/en/latest/index.html

Python, 2001. Erişim Tarihi: 17.03.2020. https://www.python.org/

Python Request, 2001. Erişim Tarihi: 15.01.2018. https://pypi.org/project/requests/

Rau, L. F., 1991. Extracting Company Names from Text. In Artificial Intelligence Applications, 1991. Proceedings., Seventh IEEE Conference. Miami, Floarida, 29-32.

Sarker, A., Ginn, R., Nikfarjam, A., 2015. Utilizing Social Media Data for Pharmacovigilance: A Review. J Biomed Inform, 54, 202–212.

Sievert, C., Shirley, K., 2014. LDAvis: A Method for Visualizing and Interpreting Topics. In Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces, 63–70.

SpaCy, 2016. Erişim Tarihi: 29.11.2019 https://www.spacy.io

Sparql, 2008. Erişim Tarihi:09.05.2019. http://www.w3.org/TR/rdfsparql-protocol/

Stevens, K., Kegelmeyer, P., Andrzejewski, D., Buttler, D., 2012. Exploring Topic Coherence over Many Models and Many Topics. Association for Computational Linguistics, Proceeding, 952–961.

Şeker, G.A. ve Eryiğit, G., 2012.Initial Explorations on Using Crfs for Turkish Named Entity Recognition. International Conference on Computational Linguistics (COLING), 11-17 March, New Delhi, 2459-2474.

Şenel, L.K., Yücesoy, V., Koç, A., Çukur, T., 2019.Topic Change Detection on Dialog Based Text, 27th Signal Processing and Communications Applications Conference (SIU), 24-26 Nisan, Sivas, 1-4.

Tasar, O. C., Komesli, M., Ünalır, M.O., 2018. Development of Semantic Web Application Architecture for Natural Language Based Querying. Turkish National Software Architecture Conference. 29-30 Kasım, İstanbul, 162-178.

Tonon A., Cudré-Mauroux P., Blarer A., Lenders V., Motik B., 2017. ArmaTweet: Detecting Events by Semantic Tweet Analysis. In: Blomqvist E., Maynard D., Gangemi A., Hoekstra R., Hitzler P., Hartig O. (eds) The Semantic Web. Lecture Notes in Computer Science,10250, 138-153.

Tsumoto, S., Kimura, T., Iwata, H., Hirano, S., 2017. Mining Text for Disease Diagnosis. Procedia Comput Science, 122, 1133-1140.

Tür, G., Hakkani-Tür, D. Oflazer, K., 2003. A Statistical Information Extraction System for Turkish. Natural Language Engineering, 9(2), 181–210.

Uguz, H., 2011. A Two-Stage Feature Selection Method for Text Categorization by Using Information Gain, Principal Component Analysis and Genetic Algorithm. Knowledge-Based Systems, 24, 1024–1032.

Ulaş, M., Karabay, B., 2020. Terör Saldırılarını İçeren Büyük Verinin Makine Öğrenmesi Teknikleri ile Analizi. Fırat Üniversitesi Mühendislik Bilimleri Dergisi, 32(1), 267-277.

Ünaldı, İ., Kırıkgöz, Y., 2011. Latent Semantıc Analysıs: An Analytıcal Tool For Second Language Wrıtıng Assessment. Mustafa Kemal University Journal of Social Sciences Institute, 8(16), 487-498.

Xie, J., Liu, X., Zeng, D.D., 2018. Mining E-Cigarette Adverse Events in Social Media Using Bi-LSTM Recurrent Neural Network with Word Embedding Representation. Journal of the American Medical Informatics Association, 25(1), 72–80.

Wang, W., Feng, Y., Dai, W. 2018. Topic Analysis of Çevrimiçi Reviews for Two Competitive Products Using Latent Dirichlet Allocation. Electronic Commerce Research and Applications, 29, 142–156.

Wang, T., Cai, Y., Leung, H., Lau, R.Y.K., Li, Q., Min, H., 2014. Product Aspect Extraction Supervised with Çevrimiçi Domain Knowledge. Knowledge-Based Systems, 86-100.

Wangwe, S., 2007, A Review of Methodology for Assessing ICT Impact on Development and Economic Transformatio. African Economic Research Consortium Working Papers, Paper No: ICTWP-02, 1-31.

Wearesocial, 2008. Erişim Tarihi: 15.02.2020. https://wearesocial.com/global-digital-report-2019

WikiPedia, 2001. Erişim Tarihi: 15.01.2020. https://tr.wikipedia.org/wiki/ Ekşi_ Sozlük

WikiPedia, 2001. Erişim Tarihi: 15.01.2020.https://tr.wikipedia.org/wiki/ wiki.teknoloji

W3C, 1994. Erişim Tarihi: 11.12.2019.http:// www.w3c.org

Van der Maaten, L., Hinton, G., 2008. Visualizing Data Using t-SEN. Journal of Machine Learning Research, 9, 2579–2605.

Yalcinkaya, M., Singh, V., 2015. Patterns and Trends in Building Information Modeling (BIM) Research: A Latent Semantic Analysis. Automation in Construction, 59, 68–80.

Yazdani, M., Belis, A., 2013. Computing Text Semantic Relatedness using the Contents and Links of A Hypertext Encyclopedia. Artificial Intelligence, 194, 176-202.

Yıldız, E., Fındık, Y., 2019. Question Similarity Detection in Turkish using Semantic Textual Similarity Methods. 27th Signal Processing and Communications Applications Conference (SIU), 24-26 Nisan, Sivas, 1-4.

Yıldıztepe, E., Uzun, V., 2018. Olasılıksal Yöntemler ile Türkçe Metinlerin Anlamsal Benzerliğinin Belirlenmesi. Sinop Üniversitesi Fen Bilimleri Dergisi, 3 (2), 66-78.

Yüksek, Y., Karasulu, B., 2010. Çoklu Ortam Ontolojilerini Kullanan Anlamsal Video Analizi Üzerine Bir İnceleme. Mühendislik Mimarlık Fakültesi Dergisi, 25, 719-739.

Zhao, W.X., Jiang, J., Weng, J., He, J., Lim, E., Yan, H., Li, X., 2011. Comparing Twitter and Traditional Media Using Topic Models. Advances in Information Retrieval, 18-21 April, Dublin, Ireland, 338-349.

Zhou, X., Chen, L., 2014. Event Detection over Twitter Social Media Streams. The International Journal on Very Large Data Bases, 23: 381.

ÖZGEÇMİŞ

Adı Soyadı : Volkan ALTINTAŞ

Doğum Yeri ve Yılı : Aydın, 1983

Medeni Hali : Evli

Yabancı Dili : İngilizce

E-posta : volkan.altintas@cbu.edu.tr

Eğitim Durumu

Lise : Aydın Anadolu Teknik ve Meslek Lisesi, 2001

Lisans : SDÜ, Teknik Eğitim Fakültesi, Elektronik-Bilgisayar Eğitimi, Bilgisayar Sistemleri Öğretmenliği

Yüksek Lisans : SDÜ, Fen Bilimleri Enstitüsü, Elektronik- Bilgisayar Eğitimi Anabilim Dalı

Mesleki Deneyim

Millî Eğitim Bakanlığı	2005-2009
MCBU Akhisar MYO	2009-…….. (halen)

Yayınlar

Sesli, M., Yeğenoğlu, E.D., Altıntaş, V., 2020. Determination of Olive Cultivars by Deep Learning and ISSR Markers. Journal of Envıronmental Bıology, 41(2), 426-431.

Sesli, M., Yeğenoğlu, E.D., Altıntaş, V., 2019. Artificial Neural Networks and Fuzzy Logic Applications Through ISSR Markers on CultİvatedType Olives. Fresenius Environmental Bulletin, 28, 1374-1380.

Gevrekçi, Y., Altıntaş, V., Yeğenoğlu, E.D., Takma, Ç., Atıl, H., Sesli, M., 2019. Yumurtacı Tavuklarda Yumurta Veriminin Tahminlenmesinde Bulanık Mantık Uygulaması. Kafkas Universitesi Veteriner Fakultesi Dergisi, 25(1), 111-118.

Sesli, M., Yeğenoğlu E.D., Altıntaş, V., Gevrekçi, Y., 2017. UPGMA and Artificial Neural Networks Applications on Wild Type Olives. Journal of Environmental Biology, 38(5), 1079-1084.

Altıntaş, V., Topal, K., Albayrak, M., 2019. Sosyal Medya Platformu Üzerinde Gizli Anlam Analizi. European Journal of Science and Technology, 16, 863-869.

Çakır, A., Küçüksille, E.U., Altıntaş, V., 2018. Baskı Devre Yerleşim Optimizasyonu için Genetik Algoritma. Journal of Technical Sciences, 8(2), 5-10.

Abuşka, M., Şevik, S., Altıntaş, V., 2018. The Effect of Blowing Direction on Heat Sink Performance by Thermal Imaging. Journal of Thermal Engineering, 4, 2471-2480.

Albayrak, M., Topal, K., Altıntaş, V., 2017. Sosyal Medya Üzerinde Veri Analizi: Twitter. Süleyman Demirel Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi, 22, 1991-1998.

Albayrak, M., Altıntaş, V., 2017. Artırılmış Gerçeklik Teknolojisinin Veritabanı Dersinde Kullanımı. Istanbul Journal of Innovatıon in Educatıon, 3(1), 13-23.

Abuşka, M., Akgül, M.B., Altıntaş, V., 2017. Yutucu Plaka Üzerine Konik Yayların Yerleştirildiği Güneş Enerjili Hava Kollektörünün Bulanık Mantık ile Modellenmesi. Politeknik Dergisi, 20(4), 907-914.

Okcu, M., Albayrak, M., Topal, K., Turhan, G., Altıntaş, V., 2019. Yapay Zekayla Sosyal Medyadan Şehir Verisini "Dinlemek. Kartepe Zirvesi Şehircilik ve Mutlu Şehir, 326-345

Abuşka, M., Altıntaş, V., Şevik, S., 2017. Prediction of the Outlet Air Temperature of Solar Air Collector with Artificial Neural Network. International Conference on Energy and Thermal Engineering, 204-208.

Abuşka, M., Şevik, S., Altıntaş, V., 2017. The Effect of Blowing Direction on Heat Sink Performance by Thermal Imaging. International Conference on Energy and Thermal Engineering, 563-568.

Albayrak, M., Altıntaş, V., 2016. Augmented Reality Application in Education Sample Preparation Lesson. 6th International Conference on "Innovations in Learning for the Future" 2016: Next Generation, 115-124.

Albayrak, M., Altıntaş, V., Sümen, A. M., Şener, G. 2016. Robotics Education Based on Augmented Reality in Primary Schools. International Conference on Advanced Technology&Sciences (ICAT'16), 55-59.

Abuşka, M., Altıntaş, V., 2016. Outlet Temperature Prediction of Trapeze Solar Air Heater with Fuzzy Logic Model. 8th International Ege Energy Syposium, 11-13 May 2016, Afyon, (CD-ROM)

Abuşka, M., Akgül, M.B., Altıntaş, V., 2015. Artificial Neural Network Modeling of the Thermal Performanceof a Novel Solar Air Absorber Plate. Mühendislik ve Bilim Alanında Yenilikçi Teknolojiler Sempozyumu, 3-5 Haziran 2015, Valencia, İspanya, (CD-ROM).

Dündar, S., Altıntaş, V., 2014. Cep Telefonu Değeri Belirlemek için Mobil Uygulama. Akademik Bilişim 2014, 589-595.

Çakır, A., Akbulut, F.T., Altıntaş, V., 2013. Dokunmatik Ekranda Menü Tasarımı. Akademik Bilişim 2013, 330-335.

Çakır, A., Altıntaş, V., Akbulut, F.T., 2013. İris Tanıma Sistemleri ve Uygulama Alanları. Akademik Bilişim 2013, 423-427.

Çakır, A., Akbulut, F.T., Altıntaş, A., 2012. Dokunmatik Ekran. Akademik Bilişim 2012, 261-267.

Çakır, A., Altıntaş, V., Akbulut, F.T., 2012. Taşınabilir Bilgisayarların Uzaktan Takip ve Kontrol Sistemi. Akademik Bilişim 2012, 319-323.

Altıntaş, V., Yeğenoğlu, E.D., 2011. Performance of Serial and Parallel Programming in Image Processing. 6.Uluslararası İleri Teknolojiler Sempozyumu, 3, 131-134.

Altıntaş, V., Uzunkavak, H., 2009. İnternet Programcılığı. Moss Yayınevi, 105s, İstanbul.