STACKOVERFLOW'DA "BIG DATA" İLE İLGİLİ GÖNDERİLERİN KONU MODELLEME VE BİRLİKTELİK ANALİZİ İLE ÖZELLİKLERİNİN ÇIKARILMASI

Adile GENÇ[1], Ayça YURTSEVEN[2], Hacer ÖZYURT[3], Özcan ÖZYURT[4*]

[1,2,3,4] Karadeniz Teknik Üniversitesi, Teknoloji Fakültesi, Yazılım Mühendisliği Bölümü, Trabzon
[1] ORCID No : https://orcid.org/0009-0001-6520-8596
[2] ORCID No : https://orcid.org/0009-0002-6361-6796
[3] ORCID No : https://orcid.org/0000-0001-8621-2335
[4] ORCID No : https://orcid.org/0000-0002-0047-6813

Anahtar Kelimeler
Konu Modelleme
LDA
Birliktelik Analizi
Büyük Veri
Stackoverflow Gönderileri

Öz

Günümüz teknolojisinde internet kullanımının artması ile birlikte "Büyük Veri" kavramının ortaya çıkması kaçınılmaz olmuştur. 23 milyondan fazla soru ve 35 milyona yakın cevap barındırarak büyük veriye katkı sağlayan StackOverflow'da paylaşılan bilgilerin analiz edilmesi güncel konuların ve eğilimlerin belirlenmesi konusunda ciddi çıkarımlar sunabilmektedir. StackOverflow'daki bu büyük ve dağınık veri kümesi üzerinde tartışmaların elle analiz edilmesi mümkün olmadığı için otomatik analiz yapabilecek yöntemlere ihtiyaç duyulmaktadır. Bu ihtiyacı gidermek için konu modelleme yaklaşımlarına başvurulmuştur. Konu modelleme alanında yapılan çalışmalarda Gizli Dirichlet Ataması (Latent Dirichlet Allocation - LDA) yöntemi oldukça tercih edilmiş ve başarısı ispatlanmıştır. Yürütülen çalışmada LDA yöntemi kullanılarak StackOverflow platformu üzerinde "Big Data" etiketli soruların ve bu soruların cevaplarının anlamsal analizi yapılmış olup büyük veri hakkında en çok konuşulan konuların %16'lık bir oran ile makine öğrenmesi/veri bilimi ve bellek yönetimi olduğu sonucuna varılmıştır. StackOverflow'da gönderilerin etiketleri ile veri seti oluşturulmuş ve birliktelik analizi yapılmıştır. Bu aşamanın asıl amacı Apriori algoritması kullanarak görülemeyen ilişkileri ortaya çıkarmaktır. Elde edilen veriler sonucunda en yüksek oran ile 100 sorunun 25'inde bigdata etiketi ile hadoop etiketinin beraber kullanıldığı görülmüştür. Ek olarak hive etiketini kullanan biri %60 gibi bir ihtimalle hadoop ve bigdata etiketini de kullanmaktadır ve bu etiketlerin kullanım oranını 2.39 artırmaktadır.

EXTRACTING FEATURES OF "BIG DATA" RELATED POSTS ON STACKOVERFLOW WITH TOPIC MODELING AND ASSOCIATION ANALYSIS

Keywords
Topic Modeling
LDA
Association Analysis
Big Data
Stackoverflow Posts

Abstract

With the increasing use of the internet in today's technology, the emergence of the concept of "Big Data" has become inevitable. With more than 23 million questions and nearly 35 million answers, analyzing the information shared on StackOverflow, which contributes to big data, can provide serious inferences about current issues and trends. Since it is not possible to manually analyze discussions on this large and dispersed dataset on StackOverflow, there is a need for methods that can perform automatic analysis. To address this need, topic modeling approaches have been used. Latent Dirichlet Allocation (LDA) method has been highly preferred and proven successful in topic modeling studies. In the current study, the LDA method was used to semantically analyze the questions labeled "Big Data" and the answers to these questions on the StackOverflow platform, and it was concluded that the most talked about topics about big data are machine learning/data science and memory management with a rate of 16%. In StackOverflow, a dataset was created with the tags of the posts and association analysis was performed. The main purpose of this stage is to reveal invisible relationships using the Apriori algorithm. As a result of the data obtained, it was seen that the bigdata tag and hadoop tag were used together in 25 out of 100 questions with the highest rate. In addition, someone who uses the hive tag is 60% likely to use both hadoop and bigdata tags, increasing the usage rate of these tags by 2.39.

Araştırma Makalesi
Başvuru Tarihi : 16.10.2023
Kabul Tarihi : 19.12.2023

Research Article
Submission Date : 16.10.2023
Accepted Date : 19.12.2023

* Sorumlu yazar: oozyurt@ktu.edu.tr
https://doi.org/10.31796/ogummf.1375611

1. Giriş

Son yıllarda büyük hızla gelişen internet teknolojileri; web sayfaları, sosyal medya uygulamaları arama motorları, forumlar, gözlemler ve araştırmalar sayesinde her an bilimsel veya bilimsel olmayan verileri toplar hale gelmiştir (Atalı, 2018). Toplanan veriler sağlık hizmetleri, yapay zeka, eğitim, devlet hizmetleri veya bankacılık gibi alanların yanında araştırmacıların yapmış oldukları araştırmalarda da kullanılabilir nitelik taşıyabilmektedir. Yediden yetmişe herkesin dijital ortama veri kaydedebilmesi ve bu verilerin günümüzde hız, çeşitlilik, kapasite (hacim) açısından büyük artış göstermesi, bu artışa teknolojinin de destek vererek veriyi daha değerli hale getirmek için yeni çözümler üretmesi ile birlikte "Büyük Veri" kavramının ortaya çıkması kaçınılmaz olmuştur (Atalı, 2018).

Büyük veri terimi ilk ortaya çıktığından itibaren üç ana bileşeni (3V) olan çeşitlilik (variety), hız (velocity) ve hacme (volume) ek olarak 5V, 7V, 10V gibi farklı sayıdaki özelliklerle de ifade edilmiştir (Favaretto ve diğ. 2020). Büyük veriler, anlamlı sonuçlara dönüştürülemediği sürece sadece bir maliyet unsuru olarak kalır. Bu yüzden her geçen gün boyutu artan büyük verilerin anlamlı verilere dönüştürülmesi için doğru analiz algoritmaları kullanılmalıdır (Eravcı, 2010). Büyük verinin özelliklerini ele alan araştırmacılar, çok büyük miktarlardaki büyük veriyi yönetmek, veriden değerli bilgi elde etmek ve bilinmeyen ilişkileri keşfetmek için büyük veri analiz yöntemlerini kullanmaktadır (Altunışık, 2015).

Büyük veri; veri madenciliği, makine öğrenmesi, doğal dil işleme gibi disiplinlerde kullanılmaktadır. Bu bağlamda yapılan analizler daha akıllı iş hamlelerine, daha hızlı karar vermeye, daha verimli operasyonlara, mevcut süreçlerin etkinliğinin artmasına, beklentilere uygun ürün geliştirme gibi farklı alanlarda doğru kararlar alınmasına imkan sağlar (Hoş, 2020). Günümüzde bilgiye erişim ve bilgi paylaşımına ihtiyaç göz önünde bulundurulduğunda soru cevap platformlarının popülaritesi gittikçe artmıştır (Bakir, Hakkoymaz, Diri ve Güçlü, 2020).

Yazılım sektöründe dünyanın en popüler soru cevap platformlarından biri olan StackOverflow insanların ihtiyaç duydukları yanıtları, ihtiyaç duydukları anda bulmalarına yardımcı olur. Her ay 100 milyondan fazla kişinin soru sormak, öğrenmek ve teknik bilgileri paylaşmak için ziyaret ettiği Stackoverflow'da kullanıcılar çok çeşitli konuları tartışabilir. Bu kişilerin üzerinde tartıştıkları konuların anlaşılması, trendlerin ve eğilimlerin ortaya çıkarılması amacıyla bu platformda paylaşılan verilerin analiz edilmesi önemli çıkarımlar ortaya koyabilir (Gürcan ve Özyurt, 2021).

Stackoverflow'da soru, cevap ve etiketler üzerinden yapılan analizler araştırmacıların verilerden doğru şekilde yararlanmasında ve bunları yönetmesinde önemli rol oynar. Platform üzerinde her geçen gün kullanıcı sayısı ve sorulan soruların artmasıyla veri miktarı da sürekli artış göstermektedir. Bu veri büyüklüğü göz önünde bulundurulduğunda elle analiz yapmak mümkün olmadığı için otomatik araçlara olan ihtiyaç ortaya çıkmıştır. Bu bağlamda veri işleme, veri analizi, konu modelleme, etiketlerin birliktelik analizi tekniklerinden faydalanılabilir.

Bu çalışma ile StackOverflow üzerinde "Big Data" konusunda yayınlanan soru ve cevaplardan konu modelleme tekniği olan LDA ile analiz yapılarak kullanıcıların konuştukları konuların belirlenmesi, etiketler üzerinde yapılacak birliktelik analizi ile ilişkilerin ortaya çıkarılması, en sık kullanılan etiketler ve bu etiketlerin yıllara göre dağılımlarının bulunması amaçlanmıştır. Elde edilen sonuçların bu konu ile ilgilenen araştırmacılara ve kullanıcılara yararlı bilgi sağlayacağı düşünülmektedir.

2. Bilimsel Yazın Taraması

Literatürde StackOverflow üzerindeki gönderilerin analizine yönelik çok sayıda çalışmadan söz edilebilir. Bunlardan bazıları şu şekilde sıralanabilir: Gürcan ve Özyurt (2021), StackOverflow gönderilerinde en çok kullanılan 50 etiket üzerinde ivmesi artan ve azalan etiketleri analiz etmişlerdir. Bu analiz, kelime frekanslarına dayalı olarak, özellikle yenilikçi web ve mobil teknolojiler ile ilişkisel olmayan veri tabanı teknolojilerinin giderek artan bir ivmeyle kullanıldığı ve güncel teknolojilerin genel olarak benimsediği sonucunu ortaya koymuştur.

Rosen ve Shihab (2016), StackOverflow üzerinde mobil geliştiricilerin tartışmalarını LDA ile sorulan soruların türlerini araştırarak mobil geliştiricilerin karşılaştığı zorlukları vurgulamayı amaçlamışlardır. Yang, Lo, Xia, Wan ve Sun (2016), StackOverflow üzerinde güvenlikle ilgili konuları ve eğilimleri araştıran bu çalışmasında, eğitimcilere ve uygulayıcılara çeşitli çıkarımlarda

bulunmak, konuların popülerliğini ve zorluğunu araştırmak için LDA kullanmıştır.

Bagherzadeh ve Khatchadourian (2019), geliştiricilerin ilgilendiği büyük veri konularını, bu konuların hiyerarşisini, popülerliklerini, zorluklarını ve korelasyonlarını ile bu tür bir anlayışın büyük veri yazılım geliştirme pratiği, araştırması ve eğitimi için etkilerini anlamak için StackOverflow'daki 157.525 büyük veri sorusunu ve cevabını LDA ile analiz edip konu modeli oluşturmuşlardır. Büyük veri geliştiricilerinin sorduğu soruları dokuz alt kategoride gruplandırmışlardır. Buna göre en fazla sorunun "Programlama" kategorisinde, en az soru ise "Günlüğe Kaydetme" kategorisinde sorulduğu görülmüştür. Çalışmanın sonucunda büyük veri konularının popülerliği ve zorluğu arasında istatistiksel olarak anlamlı bir korelasyon bulunmamıştır.

Zhang (2019) çalışmasında StackOverflow üzerinde Java ile ilgili gönderilerin konularını tespit etmek için LDA yöntemini kullanmıştır. Çalışmanın sonuçlarına göre, en popüler dört konu arasında "Sistem Platformu", "Java Arşivi", "Kaynakları İçe Aktarma" ve "Maven Yapısı" yer almıştır.

Syam, Lal ve Chen (2023) StackOverflow'daki Python ile ilgili gönderiler üzerinde yaptıkları çalışmada sorularının temasını anlamak ve geliştiricilerin karşılaştığı zorlukları belirlemeyi amaçlamışlardır. Çalışmanın sonuçlarına göre, bilimsel hesaplama kütüphaneleri "Pandas" ve "TensorFlow" hakkındaki sorularda son zamanlarda bir artış olmuştur.

Ouni, Saidani, Alomar ve Mkaouer (2023), geliştiricilerin sürekli entegrasyon bağlamında karşılaştığı zorlukları anlamak için StackOverflow'da 27.728 gönderinin LDA ile analizini gerçekleştirmişlerdir. En popüler ve zor konuları, yanıtlanmamış sorulara dayanarak araştırıp ilgili iç görüleri ortaya çıkarmışlardır. Çalışma sonucunda "Build", "Testing", "Version Control", "Configuration", "Deployment" ve "CI Culture" olmak üzere altı ana konuda zorluklarla karşılaşıldığı ortaya çıkmıştır.

Ma, Zhou, Tag, Sarsenbayeva, Knibbe ve Goncalves (2023), kullanıcıların konumlarının Weibo'da görüntülenmesine ilişkin tutum ve görüşlerini 20.162 ilgili gönderi ve yorumda LDA aracılığıyla analiz etmişlerdir. Çalışmanın sonucu çoğu insanın olumsuz bir tutum sergilediğini göstermiştir.

Alan ve Yeşilyurt (2019), Apriori algoritması ile hastane veri tabanındaki 28.738 hastaya ait verileri kullanarak birliktelik kuralları analizini yapmışlardır. Çalışma, analiz sonucunda %60 ve üzeri güven seviyesine sahip olarak 64 birliktelik kuralının varlığını ortaya koymuştur. Bu kuralların hasta memnuniyetine katkı sağlayacağı ve sağlık hizmetlerinin iyileştirilmesinde katkısı olacağı düşünülmektedir.

Doğan, Erol ve Buldu (2014) Apriori algoritmasını Türkiye'de faaliyet gösteren bir sigorta şirketinin müşterilerine ait verileri analiz etmek için kullanmışlardır. Analiz sonucunda müşterilerin bir arada almayı tercih ettikleri ürün gruplarını ortaya çıkarmışlardır. Bu sonuçlar, daha verimli satış kampanyaları ve pazarlama stratejileri geliştirmek için kullanılabilecek bilgiler sağlamaktadır.

Konu modelleme, veri kümesi içindeki saklı konuları bulmak için kullanılan metin madenciliğinin istatistiksel bir yöntemi olarak tanımlanmaktadır (Altıntaş, Albayrak ve Topal, 2021). İstatistik tabanlı konu modelleme yöntemleri, metni analiz ederken içeriğini veya anlamını anlayamazlar. Bunun yerine her konunun bir kelime havuzu olduğunu varsayarak, bir dokümanın da o havuzdan kelimelerin seçilerek oluşturulduğunu düşünür.

Konu modelleme, hissiyat analizi, soru cevaplama, bilgi çıkarımı gibi doğal dil işleme uygulamalarında oldukça sık kullanılmaktadır. Konu modelleme yöntemlerinden biri olan LDA'nın en çok tercih edilme sebeplerinden birisi de denetimsiz çeşitlerinin olmasıdır. Literatürde LDA ile yapılan pek çok araştırmaya ulaşmak mümkündür.

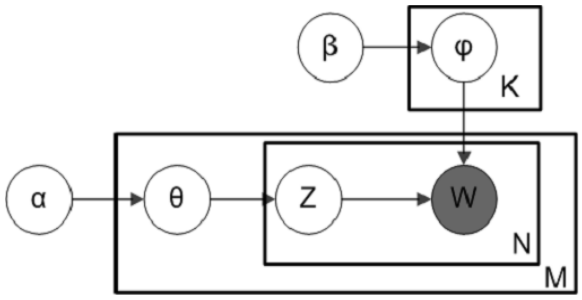
3. Gizli Dirichlet Ataması - Latent Dirichlet Allocation (LDA)

LDA, doküman içerisinden ilişkili konuların ortaya çıkartılmasını sağlayan olasılıksal üretici bir modeldir (Kaya ve Gülbandılar, 2022). Konuların kelimeler üzerinde bir olasılık dağılımına, dokümanların da konular üzerinde bir olasılık dağılımına sahip olduğu fikrine dayanmaktadır. Konu modelleme, kelimelerin olasılık dağılımları üzerinden rastgele birleşerek dokümanların oluşumunu sağlar (Steyvers ve Griffiths, 2007).

LDA denetimsiz öğrenme algoritmasıdır, bu sebeple önceden tanımlanmış kelimelere ihtiyacı yoktur. Kelimeleri bir araya getirerek işleme alır. Bu işlem sırasında kelimelerin cümle içindeki konumları önemsizdir, ancak kelimelerin birlikte bulunma durumları dikkate alınır. (Kaya ve Gülbandılar, 2022). LDA'nın grafiksel temsilinde plate notasyonundan yararlanılmaktadır. Gözlemlenen verinin rastgele değişkenlerinin yönlü kenarlar üzerinden yayılması, Plate notasyonu ile açıklanır (Ekinci ve Omurca, 2017).

LDA'ya ait plate notasyonu Şekil 1'de verilmiştir. Rastgele değişkenler düğümlerle belirtilir. Düğümler arasındaki olası bağlantılar, kenarlar kullanılarak gösterilir (Güven, Diri ve Çakaloğlu, 2020). Konu modellemedeki temel amaç; dokümanın içerdiği kelimelere dayanarak ait olduğu konuları ortaya çıkarmaktır.

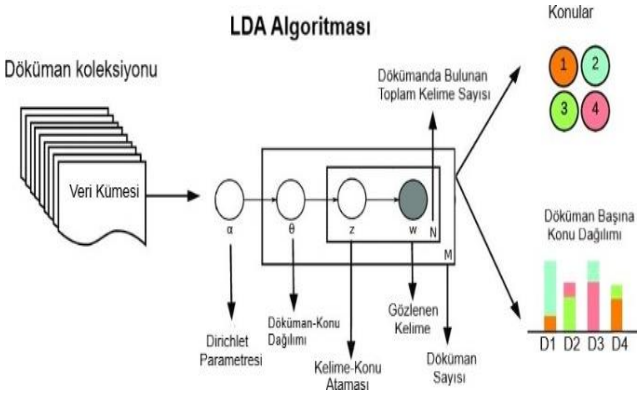
Bu aşamada yalnızca dokümanlar gözlemlenebilir durumdadır. Kelimelerin konuya atanmasıyla ilgili olan konular ve bu konuların dokümandaki ve kelimelerin konulardaki dağılımları gizli olarak kalmaktadır. Şekil 1'de, gözlemlenen değişkenler gri renkle temsil edilirken gözlenemeyenler beyaz renk ile temsil edilmiştir. (Altıntaş ve diğ., 2021).



Şekil 1. LDA İçin Grafiksel Model

Burada M Toplam doküman sayısı, K Gizli konuların sayısı, α ve β Dirichlet parametreleridir. Θ konuların dokümanda bulunma olasılığını, φ ise kelimelerin konulardaki dağılımını göstermektedir. Z her bir kelime için atanan konulardır (Ekinci ve Omurca, 2017). LDA algoritması bir kelimenin bir konuya ait olduğunu ve bir dokümanın en az bir konuya ait olduğunu varsayar. Doküman birden çok konuya da ait olabilir. Bu çokluk nedeni ile Dirichlet dağıtımına ihtiyaç duyarız, bu noktada α Dirichlet parametresi kullanılır.

LDA algoritmasının verimli çalışabilmesi için α değeri özenle seçilmelidir. Yüksek α değeri fazla sayıda başlık bulabilir ve konuların dağılımının homojenliğini sağlar. Düşük α değeri ise daha az alt başlık bulur ve çıkarım sürecinin bazı konularda olasılık dağılımını yapmasını engeller. Şekil 2'de LDA algoritması verilmiştir.



Şekil 2. LDA Algoritması

Kelimeler konulara atanırken ilk olarak mevcut dokümanın konularla ilişkisi hesaplanır. Ardından her kelimenin konularla olan ilişkisi hesaplanır. Bu hesaplama sonucunda kelimenin belirli bir konuyla ilişkili ağırlığı hakkında bilgi elde edilir (Güven ve diğ., 2018).

LDA'da K konu sayısı tutarlılık değeri ile belirlenir. Tutarlılık değeri, kelimeler arasındaki benzerliği ölçer. Sistemi uygun konu sayısı ile modellemek oldukça önemlidir. Bu nedenle, belirli konu sayıları için hesaplanan tutarlılık değerleri arasından en yüksek değere sahip olan K değeri, modelin konu sayısı olarak seçilmektedir (Güven ve diğ., 2018).

4. Araştırma Soruları

Bu çalışmanın amacı, StackOverflow da "Big Data" konusunda yayınlanan soru ve cevaplardan konu modelleme tekniği olan LDA ile analiz yapılarak kullanıcıların konuştukları konuların belirlenmesi, etiketler üzerinde yapılacak birliktelik analizi ile de ilişkilerin ortaya çıkarılmasıdır. Bu amaç doğrultusunda cevap aranan sorular ise şöyledir;

1. StackOverflow'da Big Data hakkında konuşulan konular nelerdir?

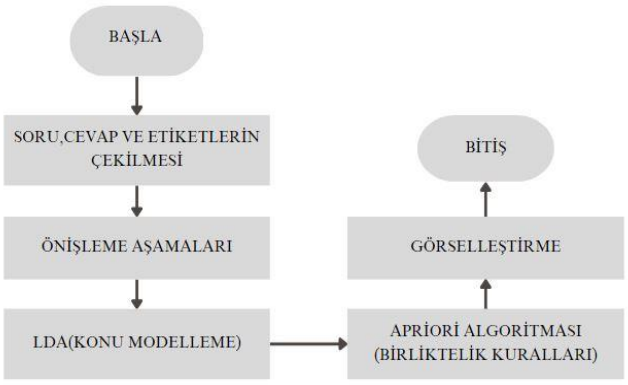
2. Birlikte kullanılan etiketlerin ilişkileri ve dağılımları ne yöndedir?

3. En sık kullanılan etiketler nelerdir?

4. Big Data ile ilgili gönderilerin betimsel karakteristikleri nedir?

5. Yöntem

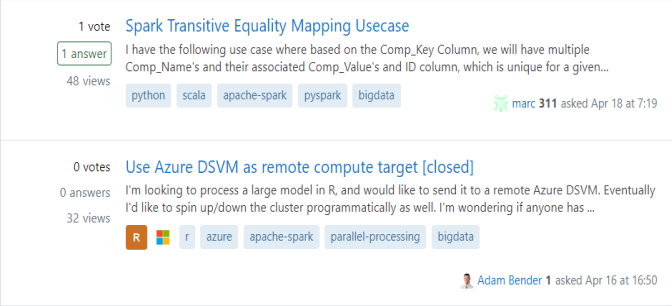
Bu çalışmada StackOverflow soru cevap platformu üzerinde 2023 yılına kadar paylaşılmış olan ve "Big Data" ile ilgili toplamda 7598 veri çekilmiştir. Soru ve cevapların analizi LDA algoritması, etiketlerin birliktelik analizi ise Apriori algoritması kullanılarak yapılmıştır. Çalışma kapsamında izlenilen yöntem Şekil 3'te gösterilmiştir.



Şekil 3. Çalışma Kapsamında İzlenilen Yol

5.1. Soru, Cevap ve Etiketlerin Çekilmesi

StackOverflow üzerinde 2023 yılına kadar paylaşılmış olan tartışmaları Python'ın BeautifulSoup ve Request kütüphaneleri kullanılarak çekilmiştir. Çekilen soru, cevap ve etiketlerin tamamı İngilizce'dir. Tartışmalar çekilirken yalnızca tartışmaların içerikleri değil tartışmaların görüntülenme sayısı, oylanma sayısı ve tarih de çekilmiştir. Toplamda 7598 veri çekilmiştir ve bu veriler CSV formatında saklanmıştır. Şekil 4'te Stackoverflow'dan bir tartışma görseli verilmiştir.



Şekil 4. Örnek Tartışma Görseli

5.2. Ön İşleme Aşamaları

Ön işleme aşamaları 4 adımdan oluşmaktadır. İlk aşama istenmeyen içeriklerin silinmesi aşamasıdır.

Bu aşamada noktalama işaretleri, sayılar, web bağlantıları ve anlamsız karakterler çıkarılmıştır. Büyük harfler, küçük harflere dönüştürülmüştür. İkinci aşama ise etkisiz kelimelerin çıkartılmasıdır. Yorum içerisinde tekrar eden kelimeler ve anlamsal ifadeyi etkilemeyen kelimeler çıkarılmıştır. Üçüncü aşama dizge parçalama aşamasıdır. Dokümandaki metinsel veriler kelime, cümle, sembol veya belirteç adı verilen anlamlı ögelere ayrılır. Dördüncü aşama kök bulma aşamasıdır. Bu aşama kelime köklerinin ayrıştırılmasının ve kelime indirgenmesinin yapıldığı aşamadır. Bu işlem yapılırken Lemmatization yöntemi kullanılmıştır. Sebebi ise bunu Stemming'den farklı olarak kelime anlamının korunmasının sağlamasıdır.

5.2.1. İstenmeyen İçeriklerin Silinmesi

Bu aşamada, konu modelleme algoritmaları büyük-küçük harfe duyarlı olduğu için dokümandaki tüm kelimelerin küçük harfe dönüşümü sağlanmıştır. Tartışmalar içerisinde anlam ifade etmeyen noktalama işaretleri, anlamsız karakterler, web bağlantıları ve sayıların çıkarılmıştır.

5.2.2. Etkisiz Kelimelerin Çıkarılması

Tartışmaların anlamsal ifadesini etkilemeyen etkisiz kelimeler dokümandan çıkartılmıştır. Bu sayede modelin gereksiz yorulması, konuların birbirine benzemesi engellenmiş ve daha iyi bir performansın elde edilmesi sağlanmıştır. Bu işlem için NLTK Kütüphanesi kullanılmıştır.

5.2.3. Dizge Parçalama

Bir LDA modelinin oluşturulup çalıştırılması için dokümandaki verilerin vektörlere dönüştürülmesi gerekir. Doküman parçalama aşamasında daha öncesinde büyük harflerin küçük harflere dönüştürülüp, noktalama işaretleri ve sayıların çıkarılıp, durgun kelimelerin atıldığı tartışmalar kendi içerisindeki kelimelerden oluşan bir vektöre dönüştürülmüştür.

5.2.4. Kök Bulma

Kök Bulma aşamasında kelimelere gelen eklerin, kelimenin anlamını değiştirmediği durumlarda oluşturulacak olan LDA modeli üzerinde olumsuz bir etki yaratmaması için kelimelerin köklerinden ayrılarak düzeltilmesi işlemi yapılmıştır.

Alfa parametresi belge-konu yoğunluğunu, eta ise konu-kelime yoğunluğunu temsil eder. Alfa değeri ile konu sayısı orantılıdır. Alfa değeri ne kadar yüksek olursa, belgeler daha fazla konudan oluşur. Öte yandan, beta değeri yüksek olan konular, derlemde çok sayıda kelimeden oluşur ve beta değeri düşük olduğunda, birkaç kelimeden oluşurlar.

Bu çalışmada Alpha ve eta parametreleri sırasıyla 0.1, 0.01, 0.5, 0.05, symmetric, symmetric değerlerinin kombinasyonları olarak seçilerek çalıştırılmıştır. Bu parametrelerden en uygun olanın Alpha parametresi 'symmetric' ve eta parametresi 'symmetric' olduğuna karar verilmiştir.

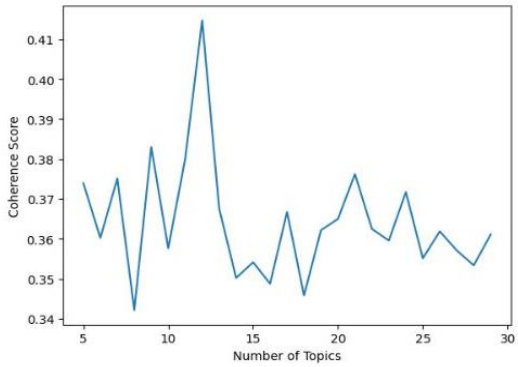
5.3. LDA ile Konu Modelleme

LDA modelleri oluşturulurken gensim kütüphanesi altında "models" modülü kullanılmıştır. LDA modeli oluşturulurken bazı önemli parametreler bulunmaktadır. Corpus parametresi ön işleme aşamalarından sonra oluşan, vektör kullanılarak oluşturulan bir sözlüktür. num_topics parametresi modelin kaç tane konuya sahip olacağını belirlemek için kullanılır. id2word parametresi sözlükteki eşsiz tüm kelimelerin tutulduğu bir yapıdır. Corpus ve id2word parametreleri soru, cevap ve başlıklara göre şekillenmektedir. Passes parametresi kullanılan corpusun üzerinden kaç defa geçilmesi gerektiğini belirlemektedir. Çok sayıda doküman varsa passes parametresinin yüksek seçilmesi model oluşturma aşamasının çok uzun sürmesine neden olacaktır. Passes parametresi bu çalışma için 25 olarak belirlenmiştir.

5.3.1. Tutarlılık ve Karmaşıklık Değerleri

Uygun tutarlılık değerine sahip modeli bulmak için 5 ve 30 konu sayısı arasında modeller oluşturulup tutarlılık değerleri karşılaştırılmıştır.

Şekil 5'teki soru cevap başlıklar ile oluşturulmuş 30 konunun tutarlılık değeri verilmektedir.



Şekil 5. Veriler İçin Modellerin Tutarlılık Değerler

Diğer modeller de incelendikten sonra anahtar kelimelerin ağırlıkları ve konular arasındaki dağılımı göz önünde bulundurulduğunda en uygun modelin 12 konu sayılı model olduğu kanısına varılmıştır.

5.4. Birliktelik Analizi Kurallarının Çıkarılması

Etiketler üzerinde yapılan birliktelik analizi ile ilişkiler ortaya çıkarılmıştır. Birliktelik kurallarının çıkarılmasında Apriori algoritması kullanılmıştır. Şekil 6'da birliktelik kurallarının uygulanması ile ortaya çıkan sonuçlar verilmiştir. Şekil 6'daki sonuçlara verilere göre;

Antecedent support; birinci etiketin tek başına görülme olasılığını ifade eder.

Consequent support; ikinci etiketin tek başına görülme olasılığını ifade eder.

Support; etiketlerin birlikte görülme olasılığını ifade eder.

Confidence; ilk etiket kullanıldığında ikinci etiketin kullanılma olasılığını ifade eder.

Lift; ilk etiket kullanıldığında ikinci etiketin kullanılma olasılığının kaç kat arttığını ifade eder.

5.4.1. Analiz Sonuçları

18.satır için: Kullanılan etiketlerde hadoop'un tek başına görülme olasılığı %25 (antecedent support), bigdata'nın tek başına görülme olasılığı %100. (consequent support) 100 sorunun 25'inde mutlaka hadoop ve bigdata etiketleri beraber kullanılmıştır. (support)

79.satır için: Kullanılan etiketlerde hive'in tek başına görülme olasılığı %9, hadoop ve bigdata'nın tek başına görülme olasılığı %25. 100 sorunun 6'sında hive, bigdata ve hadoop etiketleri beraber kullanılmıştır. (support) Hive etiketini kullanan biri %60 gibi bir ihtimalle hadoop ve bigdata etiketini kullanmaktadır (confidence) ve bu etiketlerin kullanım oranını 2.39 artırmaktadır. (lift)

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction	zhangs_metric
7	(apache-spark)	(bigdata)	0.140273	1.00000	0.140273	1.000000	1.000000	0.000000	inf	0.000000
14	(database)	(bigdata)	0.047326	1.00000	0.047326	1.000000	1.000000	0.000000	inf	0.000000
18	(hadoop)	(bigdata)	0.256030	1.00000	0.256030	1.000000	1.000000	0.000000	inf	0.000000
22	(hdfs)	(bigdata)	0.040509	1.00000	0.040509	1.000000	1.000000	0.000000	inf	0.000000
23	(hive)	(bigdata)	0.097798	1.00000	0.097798	1.000000	1.000000	0.000000	inf	0.000000
25	(java)	(bigdata)	0.083901	1.00000	0.083901	1.000000	1.000000	0.000000	inf	0.000000
28	(mapreduce)	(bigdata)	0.058731	1.00000	0.058731	1.000000	1.000000	0.000000	inf	0.000000
30	(mysql)	(bigdata)	0.041689	1.00000	0.041689	1.000000	1.000000	0.000000	inf	0.000000
37	(pyspark)	(bigdata)	0.041951	1.00000	0.041951	1.000000	1.000000	0.000000	inf	0.000000
38	(python)	(bigdata)	0.133325	1.00000	0.133325	1.000000	1.000000	0.000000	inf	0.000000
40	(r)	(bigdata)	0.078133	1.00000	0.078133	1.000000	1.000000	0.000000	inf	0.000000
41	(scala)	(bigdata)	0.044966	1.00000	0.044966	1.000000	1.000000	0.000000	inf	0.000000
42	(sql)	(bigdata)	0.050996	1.00000	0.050996	1.000000	1.000000	0.000000	inf	0.000000
48	(hive)	(hadoop)	0.097798	0.25603	0.060042	0.613941	2.397922	0.035003	1.927088	0.646166
49	(mapreduce)	(hadoop)	0.058731	0.25603	0.045621	0.776786	3.033959	0.030584	3.332984	0.712227
77	(hive, hadoop)	(bigdata)	0.060042	1.00000	0.060042	1.000000	1.000000	0.000000	inf	0.000000
78	(hive, bigdata)	(hadoop)	0.097798	0.25603	0.060042	0.613941	2.397922	0.035003	1.927088	0.646166
79	(hive)	(bigdata, hadoop)	0.097798	0.25603	0.060042	0.613941	2.397922	0.035003	1.927088	0.646166
82	(bigdata, mapreduce)	(hadoop)	0.058731	0.25603	0.045621	0.776786	3.033959	0.030584	3.332984	0.712227
83	(hadoop, mapreduce)	(bigdata)	0.045621	1.00000	0.045621	1.000000	1.000000	0.000000	inf	0.000000
84	(mapreduce)	(bigdata, hadoop)	0.058731	0.25603	0.045621	0.776786	3.033959	0.030584	3.332984	0.712227

Şekil 6. Apriori Algoritması Sonrasında Elde Edilen Sonuçlar

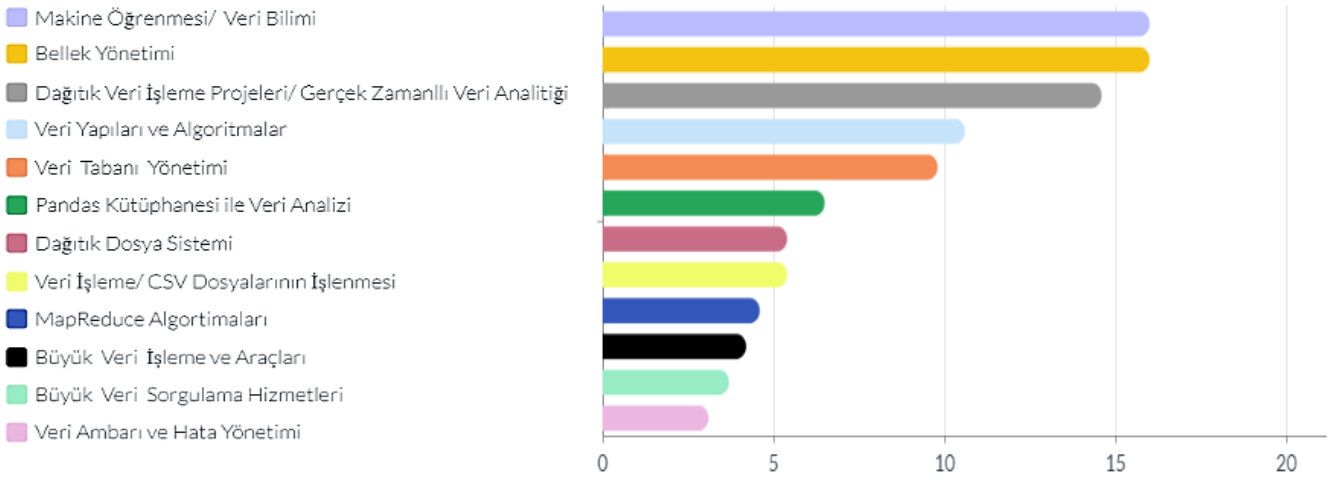
6. Bulgular

Bu projede StackOverflow platformundan 2023 yılına kadar toplanan toplamda 7598 adet çekilen verinin soru, cevap ve başlıkları ile konu modellemesi gerçekleştirilmiştir. Veriler ön işleme aşamalarına tabi tutulmuş ve ardından uygun bir LDA modeli geliştirilerek konu ataması yapılmıştır.

Soru cevap platformunda oluşturulan modele göre belirlenen ideal 12 konudan en fazla konuşulan konuların %16'lık oran ile makine öğrenmesi/veri bilimi ve bellek yönetimi olduğu belirlenmiştir. Belirlenen konu sayısı içerisinde en az konuşulan konunun %3.6 ile veri ambarı ve hata yönetimi olduğu ortaya koyulmuştur. Ayrıntılı sonuçlar aşağıdaki Tablo 1' de ve Şekil 7'de yer almaktadır.

Tablo 1. Konu Dağılımları ve Oranları

Konu	Kelimeler ve Ağırlıkları	Etiket	Oran
0	0.030*"data" + 0.012*"model" + 0.012*"dataset" + 0.011*"use" + 0.009*"train" + 0.009*"algorithm" + 0.008*"cluster" + 0.007*"memory"+ '0.006*"problem" + 0.005*"matrix" + 0.005*"r" + 0.005*"variable" + 0.005*"learn" + 0.004*"machine" + 0.004*"test"	Makine Öğrenmesi/Veri Bilimi	%16
1	0.013*"data" + 0.012*"use" + 0.012*"gb" + 0.012*"query" + 0.011*"memory" + 0.009*"index" + 0.008*"size" + 0.007*"user" + 0.007*"ram" + 0.007*"document" + 0.006*"run" + 0.006*"error"+ 0.006*"file" + 0.006*"graph" + 0.006*"try"	Bellek Yönetimi	%16
2	0.021*"stream" + 0.021*"spark" + 0.017*"kafka" +0.016*"use" + 0.013*"message" + 0.013*"apache" + 0.011*"api" + 0.010*"process" + 0.010*"data" + 0.009*"event" + 0.007*"window" + 0.007*"storm" + 0.006*"flume" + 0.006*"topic" + 0.006*"application"	Dağıtık Veri İşleme Projeleri/Gerçek Zamanlı Veri Analitiği	%14.6
3	0.014*"array" + 0.014*"data" + 0.013*"use" + 0.011*"memory" + 0.011*"number" + 0.010*"list" + 0.010*"function" + 0.009*"time" + 0.007*"object" + 0.007*"value" + 0.007*"python" + 0.007*"element" + 0.006*"problem" + 0.006*"string" + 0.005*"size"	Veri Yapıları ve Algoritmalar	%10.6
4	0.035*"data" + 0.030*"table" + 0.016*"query" + 0.013*"row" + 0.011*"partition" + 0.010*"index" + 0.010*"column" + 0.009*"use" + 0.009*"time" + 0.009*"record" + 0.009*"database" + 0.007*"key" + 0.007*"user" + 0.006*"create" + 0.006*"store"	Veri Tabanı Yönetimi	%9.8
5	0.024*"column" + 0.020*"dataframe" + 0.019*"row" + 0.015*"data" + 0.014*"use" + 0.010*"panda" + 0.009*"time" + 0.008*"function" + 0.008*"follow" + 0.008*"id" + 0.008*"try" + 0.007*"group" +0.007*"dataset" + 0.006*"create" + 0.006*"result"	Pandas Kütüphanesi ile Veri Analizi	%6.5
6	0.024*"file" + 0.021*"hadoop" + 0.018*"run" + 0.017*"cluster" + 0.016*"hdfs" + 0.013*"node" + 0.013*"job" + 0.012*"use" + 0.010*"data" + 0.009*"spark" + 0.009*"nod" + 0.007*"task" + 0.006*"error" + 0.006*"directory" + 0.006*"try"	Dağıtık Dosya Sistemi	%5.4
7	0.069*"file" + 0.023*"data" + 0.019*"read" + 0.016*"use" + 0.012*"csv" + 0.011*"column" + 0.008*"memory" + 0.007*"process" + 0.007*"gb" + 0.006*"format" + 0.006*"row" + 0.006*"line" + 0.006*"spark" + 0.006*"r" + 0.006*"try"	Veri İşleme/CSV Dosyalarının İşlenmesi	%5.4
8	0.023*"output" + 0.021*"mapper" + 0.020*"reducer" + 0.019*"map" + 0.017*"key" + 0.015*"mapreduce" + 0.014*"reduce" + 0.012*"input" + 0.012*"sort" + 0.012*"use" + 0.011*"hive" + 0.011*"bucket" + 0.010*"job" + '0.008*"hadoop" + 0.007*"number"	MapReduce Algoritmaları	%4.6
9	0.066*"data" + 0.014*"database" + 0.011*"use" + 0.011*"store" + 0.010*"big" + 0.010*"table" + 0.010*"query" + 0.007*"hadoop" + 0.007*"process" + 0.006*"time" + 0.006*"sql" + 0.005*"question" + 0.005*"system" + '0.005*"tool" + 0.005*"storage"	Büyük Veri İşleme ve Araçları	%4.2
10	0.035*"table" + 0.029*"query" + 0.021*"column" + 0.016*"join" + 0.012*"row" + 0.010*"result" + 0.009*"use" + 0.009*"index" + 0.008*"bigquery" + 0.008*"id" + 0.006*"document" + 0.006*"hive" + 0.006*"delete" + 0.006*"sql" + 0.006*"create"	Büyük Veri Sorgulama Hizmetleri	%3.7
11	0.029*"hive" + 0.024*"error" + 0.017*"spark" + 0.014*"table" + 0.014*"use" + 0.014*"log" + 0.011*"create" + 0.009*"run" + 0.008*"try" + 0.007*"data" + 0.007*"issue" + 0.006*"follow" + 0.006*"version" + 0.006*"get" + 0.006*"exception"	Veri Ambarı ve Hata Yönetimi	%3.1



Şekil 7. Konuların Başlıklara Olan Genel Bar Dağılım Grafiği

Bu araştırma bağlamında, etiketlerin analizine dayanarak, StackOverflow'da 2009-2022 yılları arasındaki gönderilerden elde edilen verilere ilişkin ayrıntılı bilgiler Tablo 2'de mevcuttur.

Tabloda görüldüğü gibi StackOverflow üzerinde 2009-2022 yılları arasındaki gönderilerde toplam 30329 etiket kullanılmış olup bu etiketler içerisindeki tekil (tekrar etmeyen) etiket sayısı 2282 olarak hesaplanmıştır.

StackOverflow'da en çok kullanılan ilk 20 etiket, toplam etiket sayısının yaklaşık %57.8'ini oluştururken, en çok kullanılan ilk 50 etiket, toplam etiket sayısının yaklaşık %67.3'ünü oluşturmaktadır. Big Data etiketi ile kullanılan etiketlerden en çok kullanılanların sırasıyla Hadoop, Apache-spark ve Python olduğu belirlenmiştir.

Tablo 2. 2009-2022 Yıllarına Ait "Big Data" Etiketlerine Dair Veriler

Etiket Kategorileri	Elde Edilen Veriler
Farklı (tekil) etiket sayısı	2282
Toplam etiket sayısı	30329
Yıllık ortalama etiket sayısı	2166
En fazla kullanılan ilk 20 etiketin toplam sayısı	17500
En fazla kullanılan ilk 20 etiketin toplam etiketlerin sayısına oranı	57.80
En fazla kullanılan ilk 50 etiketin toplam sayısı	20426
En çok kullanılan ilk 50 etiketin toplam etiketlerin sayısına oranı	67.35
Sadece 1 kez kullanılan farklı (tekil) etiket sayısı	1096
En fazla 5 kez kullanılan farklı (tekil) etiket sayısı	1874
En fazla kullanılan ilk üç etiket (Sıralı- Big Data hariç)	Hadoop, Apache-spark, Python

Tablo 3 ve Tablo 4'te, yıllara göre toplam etiket sayısı ve farklı(tekil) etiket sayılarının dağılımları sunulmaktadır. Tabloda belirtilen sonuçlara göre 2022 yılı sonunda toplam etiket sayısı 30329, tekil etiket sayısı ise 5731 olmuştur.

Tablo 3. Yıllara Göre Tekil Etiket Sayıları ve Toplamı

Yıllar	N
2009	14
2010	15
2011	81
2012	214
2013	431
2014	511
2015	644
2016	731
2017	716
2018	637
2019	552
2020	298
2021	419
2022	468
Toplam	5731

Tablo 4. Yıllara Göre Toplam Etiket Sayıları ve Toplamı

Yıllar	N
2009	22
2010	24
2011	165
2012	678
2013	2008
2014	2821
2015	4023
2016	5026
2017	4355
2018	3379
2019	2872
2020	1047
2021	1718
2022	2191
Toplam	30329

Tablo 5'te, StackOverflow üzerinde 2009-2022 arasındaki gönderilerde, yıllara göre en fazla kullanılan ilk 10 etiketin dağılımı ve bu etiketlerin toplam sayıları sıralanmıştır. Big data etiketi üzerinde işlem yapıldığından dolayı 7594 etiket ile en çok kullanılandır. İkinci sırada ise 1912 etiket ile Hadoop yer almaktadır.

Tablo 5. En Fazla Kullanılan İlk 10 Etikete Dair Veriler

Etiketler	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022	Toplam
Big data	4	6	44	177	526	721	1027	1247	1096	846	718	259	417	506	7994
Hadoop	-	1	3	47	150	238	406	442	279	171	122	16	50	17	1942
Apache- spark	-	-	-	-	2	42	129	249	160	155	130	26	67	115	1075
Python	-	1	4	10	45	65	81	132	137	121	131	78	93	129	1027
Hive	-	-	-	10	33	51	109	133	123	86	94	18	48	37	742
Java	-	-	6	16	57	86	97	126	93	60	56	6	17	20	640
R	-	2	5	26	68	64	79	62	68	56	56	29	24	50	589
Mapreduce	-	-	1	18	47	61	108	76	59	39	18	4	11	5	447
Sql	-	1	3	8	30	35	35	40	31	37	53	15	42	44	374
Database	2	1	7	7	39	50	39	60	49	37	30	14	11	20	366

7. Tartışma

Bu çalışmada, StackOverflow platformundan elde edilen 7598 veri seti üzerinden gerçekleştirilen konu modellemesi; soru, cevap ve başlıkların analiziyle şekillenmiştir. Literatürde çeşitli konu modelleme teknikleri mevcut olmakla birlikte, üretken bir model olması ve konu seçiminde deterministik sonuçlar üretmesi sebebiyle LDA yöntemi seçilmiştir (Ozyurt ve Ozyurt, 2023). Bir sorunun başlığının, soru metninin ve soru cevabının ayrı ayrı analiz edilmesi yerine bütünsel bir metin parçası düzeyinde analiz edilmesi gerektiğinin daha sağlıklı sonuç vereceği düşünülüp her bir soruya ait kelimeler tek bir metin olarak ele alınmıştır. Ayrı ayrı bir veri kümesi olarak kullanma kullanıcıların büyük veri hakkındaki ilgi alanlarını ve zorluklarını incelemek için yeterli temsili bir set olmayabilir. LDA tabanlı analiz sonucunda her konunun 15 açıklayıcı anahtar kelimeyle tanımlandığı 12 konu belirlenmiştir. Konu başlıkları, konuların anahtar kelimelerinin ortak temalarına dayalı olarak manuel olarak belirlenmiştir. Belirlenen 12 konu arasında, en çok konuşulan konuların %16'lık oranla makine öğrenmesi/veri bilimi ve bellek yönetimi olduğu görülmüştür.

Yazılım geliştirme ve teknoloji alanındaki trendler, konuşulan konuların öncelik sıralamasını etkiler. Kullanıcılar, geliştiriciler ve endüstriler arasındaki ihtiyaç ve talep değişimleri, konuşulan konuların öncelik sıralamasında etkili olabilir. Özellikle bir konunun çözülmesi veya bir başka konunun öne çıkması, bu değişimleri tetikleyebilir.

Bagherzadeh ve Khatchadourian (2019), çalışmalarında büyük veri popülerliğine bakıldığında; elde ettikleri analiz sonucu en çok konuşulan konular arasında "Makine Öğrenmesi" alt sıralarda iken 4 yıllık değişimle günümüzde ilk sırada yerini almıştır. Bunun nedeni algoritmaların ve yazılım araçlarının geliştirilmesiyle yakından ilişkilidir. Yeni ve daha etkili makine öğrenimi modellerinin ortaya çıkması, bu alandaki ilgiyi artırmıştır. Büyük veri setlerinin artması, makine öğrenimi algoritmalarının daha etkili ve doğru sonuçlar üretmesine olanak tanır. Daha fazla veri, öğrenme modellerinin daha iyi eğitilmesini sağlar. Makine öğrenimi modelleri genellikle büyük veri setlerinde eğitildiği için, büyük veri kaynaklarının artması ve işlem gücündeki artış, daha karmaşık modellerin oluşturulmasına ve kullanılmasına olanak tanımıştır.

Çalışmamızda en fazla kullanılan ilk 10 etikete dair verilere bakıldığında Hadoop etiketinin yıllara göre dağılımında azalışın nedeni olarak, Hadoop'un yerini daha spesifik ve özel amaçlı çözümlerin aldığı düşünülmektedir. Hadoop MapReduce modeli, bazı durumlarda performans ve işlem hızı açısından sınırlamalara sahiptir. Apache Spark gibi daha yüksek seviyeli ve kullanımı kolay çözümler, geliştiricilerin daha hızlı ve etkili bir şekilde büyük veri işleme yapmalarını sağlayabilir. Büyük veri işleme alanındaki odak, sadece veri depolama ve işlemeden ziyade makine öğrenimi ve yapay zeka konularına da kaymıştır. Bu nedenle Hadoop yerine alternatif çözümlere yönelme durumu söz konusudur.

StackOverflow üzerindeki etiket analizi ve büyük veri çalışmalarının detaylı incelenmesiyle elde edilen sonuçlar, "Big Data"nın betimsel analizini ortaya koymuştur. Farklı 2282 etiketten en çok kullanılan ilk 50 etiket, toplam etiket sayısının yaklaşık %67.3'ünü oluşturmaktadır. Bu kapsamda, ilk 50 etiket üzerinden elde edilecek sonuçlar ve yapılan tartışmaların, genel veri setini temsil ettiği düşünülebilir.

8. Sonuçlar

Yürütülen çalışmada LDA yöntemi kullanılarak StackOverflow platformu üzerinde "Big Data" etiketli soruların ve bu soruların cevaplarının anlamsal analizi ile büyük verinin karakteristiği ortaya çıkarılmış ve büyük veri hakkında en çok konuşulan konuların %16'lık bir oran ile makine öğrenmesi/veri bilimi ve bellek yönetimi olduğu sonucuna varılmıştır. StackOverflow gönderilerindeki "big data" ile ilgili etiketler için bir veri seti oluşturulmuş ve bu etiketlerin birliktelik analizi yapılmıştır. Bu aşamanın asıl amacı Apriori algoritması kullanarak görülemeyen ilişkileri ortaya çıkarmaktır. Örneğin kullanılan etiketlerde hadoop'un %25 olasılıkla tek başına kullanıldığı, bigdata'nın %100 olasılıkla tek başına kullanıldığı görülmektedir. 100 sorunun 25'inde mutlaka hadoop ve bigdata etiketleri beraber kullanılmış veya etiketlerde hive'in tek başına görülme olasılığı %9, hadoop ve bigdata'nın tek başına görülme olasılığı %25. 100 sorunun 6'sında hive, bigdata ve hadoop etiketleri beraber kullanılmıştır. Hive etiketini kullanan biri %60 gibi bir ihtimalle hadoop ve bigdata etiketini kullanmaktadır ve bu etiketlerin kullanım oranını 2.39 artırmaktadır. Elde edilen bütün veriler sonucunda da "Big Data"nın betimsel karakteristikleri ortaya konulmuştur.

Çalışma bir bütün olarak ele alındığında, son zamanlarda bilgisayar teknolojileri içerisinde önemli bir yer tutan "Big Data" ile ilgili tartışmaların analizi, alanda çalışanlara, eğitim görenlere ve bu alana uzmanlaşmak isteyenler başta olmak üzere, birçok paydaşa önemli katkılar sunması beklenmektedir. Çalışmanın veri bilimi ve analizi başta olmak üzere, alandaki teknolojiler, uygulamalar ve programlama dilleri gibi farklı kategorilerin gelişimi ve değişimini de ortaya koyması bu alandaki eğilimlerin belirlenmesi açısından önemlidir. Öte yandan çalışmanın sonuçlarının bu alanda ileride yapılacak çalışmalara temel oluşturması bakımından değerli olduğu düşünülmektedir.

Teşekkür

Bu çalışma 2023 yılında TÜBİTAK 2209-A Üniversite Öğrencileri Yurt İçi Araştırma Projeleri Destek Programı tarafından 1919B012221064 numaralı proje kapsamında TUBİTAK tarafından desteklenmiştir.

Araştırmacıların Katkısı

Çalışmada Ayça YURTSEVEN veri setinin oluşturulması, literatür taraması, veri analizi ve sonuçların elde edilmesi ve makalenin yazımı konularında katkı sağlamıştır. Adile GENÇ veri setinin oluşturulması, verilerin çekilmesi, sonuçların elde edilmesi ve makale yazımı konularında katkı sağlamıştır. Hacer ÖZYURT, kavramsal çerçeve, yöntem ve veri analizi konularında katkı sağlamıştır. Özcan ÖZYURT araştırmanın organize edilmesi, veri analizi ve son düzenleme konularına katkı sağlamıştır.

Çıkar Çatışması

Hazırlanan makalede araştırma ve yayın etiğine uyulmuş olup yazarlar tarafından herhangi bir çıkar çatışması beyan edilmemiştir.

Kaynaklar

Alan, M. A. & Yeşilyurt, C. (2019). Birliktelik Kuralları Madenciliği İle Yatan Hasta Profilinin Çıkarılması. *Atatürk Üniversitesi Sosyal Bilimler Enstitüsü Dergisi, 23*(4), 1917-1926.

Altınbaş, V., Albayrak, M. & Topal, K. (2021). Topic modeling with latent dirichlet allocation for cancer disease posts, *Journal of the Faculty of Engineering and Architecture of Gazi University, 36*(4), 2183–2196.

Altunışık, R. (2015). Büyük veri: fırsatlar kaynağı mı yoksa yeni sorunlar yumağı mı?. *Yildiz Social Science Review, 1*(1), 45-76.

Atalı, L. (2018). Sporda büyük veri kullanımının incelenmesi" bigdata. 16. Spor Bilimleri Kongresi Tam Metin Bildiri Kitabı, S: 1997-2000, Antalya.

Bagherzadeh, M. & Raffi, K. (2019). "Going big: a large-scale study on what big data developers ask." Proceedings of the 2019 27the ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering, Tallinn Estonia.

Bakir, C., Hakkoymaz, V, Diri, B. & Güçlü, M. (2020). Dağıtık veritabanlarında saldırı önleme metotları. *Gümüşhane Üniversitesi Fen Bilimleri Enstitüsü Dergisi, 10*(2), 425-441.

Doğan, B., Erol, B. & Buldu, A. (2014). Sigortacılık sektöründe müşteri ilişkileri yönetimi için birliktelik kuralı kullanılması. *Marmara Fen Bilimleri Dergisi, 26*(3), 105-114. doi: https://doi.org/10.7240/mufbed.56489

Ekinci, E. & Omurca, S. İ. (2017). Ürün özelliklerinin konu modelleme yöntemi ile çıkartılması. *Türkiye Bilişim Vakfı Bilgisayar Bilimleri ve Mühendisliği Dergisi, 9*(1), 51-58.

Eravcı, D. B. (2010). Kurumların dijital dönüşümü: büyük veri . *Çalışma İlişkileri Dergisi, 11*(1), 90-112.

Favaretto, M., De Clercq, E., Schneble, C. O., & Elger, B. S. (2020). What is your definition of Big Data? Researchers' understanding of the phenomenon of the decade. *PloS one, 15*(2), e0228987.

Gürcan, F. & Özyurt, Ö. (2021). Stackoverflow gönderilerinde tartışılan trend konuların kelime frekans analizi ile belirlenmesi. *Gümüşhane Üniversitesi Fen Bilimleri Dergisi, 11*(2), 357-368. doi: https://doi.org/10.17714/gumusfenbil.811123

Güven, Z. A. , Diri, B. & Çakaloğlu, T. (2018). Classification of turkish tweet emotions by n- stage latent dirichlet allocation, 2018 Electric Electronics, Computer Science, Biomedical Engineerings Meeting (EBBT). doi: https://doi.org/10.21541/apjes.459447

Güven, Z. A. , Diri, B. & Çakaloğlu, T. (2020).

Comparison of n-stage Latent Dirichlet Allocation versus other topic modeling methods for emotion analysis. *Journal of the Faculty of Engineering and Architecture of Gazi University, 35*(4), 2135-2146. doi: https://doi.org/ 10.17341/gazimmfd.556104

Hoş, S. (2020). Veri analizi nedir, büyük veri analizi nasıl yapılır? Erişim adresi: http://www.hosting.com.tr/blog/buyuk-veri-analizi/

Jelodar, H., Wang, Y., Yuan, C., Feng, X., Jiang, X., Li, Y. & Zhao, L. (2019). Latent dirichlet allocation (LDA) and topic modeling: models, applications, a survey. Latent dirichlet allocation (LDA) and topic modeling: models, applications, a survey. *Multimedia Tools and Applications, 78*(11), 15169-15211. doi: https://dl.acm.org/doi/10.1007/s11042-018-6894-4

Kaya, A. & Gülbandılar, E. (2022). Konu modelleme yöntemlerinin karşılaştırılması. *Eskişehir Türk Dünyası Uygulama ve Araştırma Merkezi Bilişim Dergisi, 3*(2),46-53. doi: https://doi.org/10.53608/estudambilisim.1097978

Ma, Y., Zhou, Q., Tag, B., Sarsenbayeva, Z., Knibbe, J. & Goncalves, J. (2023). "Hello, fellow villager!": perceptions and impact of displaying users' locations on weibo. In IFIP Conference on Human-Computer Interaction (pp. 511-532). doi: https://dl.acm.org/doi/abs/10.1007/978-3-031-42286-7_29

Ouni, A., Saidani, I., Alomar, E. & Mkaouer, M. W. (2023). An empirical study on continuous integration trends, topics and challenges in stack overflow. In Proceedings of the 27th International Conference on Evaluation and Assessment in Software Engineering (pp. 141-151). doi: https://doi.org/10.1145/3593434.3593485

Özyurt, O. & Özyurt, H. (2023). A large-scale study based on topic modeling to determine the research interests and trends on computational thinking. *Education and Information Technologies, 28*(3), 3557-3579. doi: https://dl.acm.org/doi/abs/10.1007/s10639-022-11325-9

Rosen, C. & Shihab, E. (2016). What are mobile developers asking about? a large scale study using stack overflow. *Empirical Software Engineering, 21*(3), 1192-1223. doi: https://dl.acm.org/doi/10.1007/s10664-015-9379-3

Stackoverflow (t.y.). Who We Are. Erişim adresi: https://stackoverflow.co/

Steyvers, M. & Griffiths, T. (2007). Probabilistic topic models. *Handbook of latent semantic analysis, 427*(7), 424-440.

Syam, G., Lal, S. & Chen, T. (2023). Empirical Study of the Evolution of Python Questions on Stack Overflow. *e-Informatica Software Engineering Journal, 17*(1).

Yang, X. L., Lo, D., Xia, X., Wan, Z. Y. & Sun, J. L. (2016). What security questions do developers ask? a large-scale study of stack overflow posts. *Journal of Computer Science and Technology, 31*, 910-924. doi: https://doi.org/10.1007/s11390-016-1672-0

Zhang, P. (2019). What topics do developers concern? An analysis of java related posts on stackoverflow. In 2019 2nd International Conference on Artificial Intelligence and Big Data (ICAIBD) (pp. 362-368). IEEE. doi: https://doi.org/ 10.31590/ejosat.702949