
GAZİ ÜNİVERSİTESİ
Gazi
1926

SOSYAL AĞ ANALİZİ VE MAKİNE ÖĞRENMESİ TEMELLİ SİBER ZORBALIK TESPİTİ

Mümin Ferhat YAKUT

YÜKSEK LİSANS TEZİ
BİLGİSAYAR MÜHENDİSLİĞİ ANA BİLİM DALI

GAZİ ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

OCAK 2024

ETİK BEYAN

Gazi Üniversitesi Fen Bilimleri Enstitüsü Tez Yazım Kurallarına uygun olarak hazırladığım bu tez çalışmasında;

- Tez içinde sunduğum verileri, bilgileri ve dokümanları akademik ve etik kurallar çerçevesinde elde ettiğimi,
- Tüm bilgi, belge, değerlendirme ve sonuçları bilimsel etik ve ahlak kurallarına uygun olarak sunduğumu,
- Tez çalışmasında yararlandığım eserlerin tümüne uygun atıfta bulunarak kaynak gösterdiğimi,
- Kullanılan verilerde herhangi bir değişiklik yapmadığımı,
- Bu tezde sunduğum çalışmanın özgün olduğunu,

bildirir, aksi bir durumda aleyhime doğabilecek tüm hak kayıplarını kabullendiğimi beyan ederim.

Mümin Ferhat YAKUT

04/01/2024

SOSYAL AĞ ANALİZİ VE MAKİNE ÖĞRENMESİ TEMELLİ SİBER ZORBALIK TESPİTİ

(Yüksek Lisans Tezi)

Mümin Ferhat YAKUT

GAZİ ÜNİVERSİTESİ

FEN BİLİMLERİ ENSTİTÜSÜ

Ocak 2024

ÖZET

Teknolojinin hızlı gelişimi, internet kullanımını hayatımızın ayrılmaz bir parçası haline getirmiştir. "We Are Social Digital 2023" raporuna göre, dünya çapında internet kullanım oranı %64 iken, Türkiye'de bu oran %84'e ulaşmaktadır. İnternetin en popüler kullanım alanlarından biri olan sosyal medya, 2023 itibarıyla yaklaşık 4,7 milyar kullanıcıya sahip olup, bu sayı dünya nüfusunun yaklaşık %57'sine denk gelmektedir. Türkiye'de ise sosyal medya kullanımı toplam nüfusun %60'ını kapsamaktadır. Bu veriler, internet ve sosyal medya kullanımının yaygınlığını açıkça göstermektedir. Ancak internet kullanımının artışı, olumlu yönlerinin yanı sıra bazı olumsuz etkileri de beraberinde getirmektedir. Bu çalışmada, bu olumsuz etkilerden biri olan siber zorbalık konusu ele alınmıştır. Siber zorbalık, bilgi teknolojileri araçlarının kullanılmasıyla gerçekleştirilen bir tür geleneksel zorbalık olarak tanımlanabilir. Bu tür zorbalık, bireylerin psikolojik ve duygusal sağlığını olumsuz etkileyebilir ve toplumda ciddi sonuçlara yol açabilir. Siber zorbalığın artışında, sosyal medya uygulamalarında kullanıcıların anonimliği ve geniş kitlelere kolayca ulaşabilmesi önemli bir rol oynamaktadır. Kullanıcıların bu platformlarda kolay etkileşim kurabilmesi, bireysel zorbalığın yanı sıra topluluklar arası zorbalığın da artmasına neden olmakta ve bu durum daha büyük olumsuzluklara yol açmaktadır. Siber zorbalığın etkilerini azaltmak ve bu davranışları tespit etmek amacıyla geliştirilen yöntemler, teknolojinin olumsuz etkilerini sınırlamada kritik bir rol oynamaktadır. Bu tez, siber zorbalığın tanımı, yaygınlığı ve etkileri üzerine odaklanarak, bu sorunun daha iyi anlaşılmasına ve çözüm yollarının geliştirilmesine katkı sağlamayı hedeflemektedir. Makine öğrenmesi ve derin öğrenme gibi yapay zekâ yaklaşımlarının, topluluk tespiti algoritmalarıyla birleştirilmesinin siber zorbalık tespitinde nasıl iyileştirmeler sağlayabileceği incelenmiştir. Yapılan deneysel çalışmada, sekiz farklı makine öğrenmesi ve derin öğrenme algoritması, çeşitli özellik çıkarım ve seçim teknikleri ile test edilmiş, en yüksek başarı oranı %90 doğrulukla Yapay Sinir Ağları algoritması ile CountVectorizer ve Ki-Kare teknikleri kullanılarak elde edilmiştir. Louvain algoritması uygulandığında, benzer konular üzerinde paylaşımlar yapan kullanıcıların aynı topluluklarda yer aldığı gözlemlenmiştir. Bu bulgu hem topluluk tespiti hem de siber zorbalık tespiti çalışmalarının derinleştirilerek sürdürülmesi için gelecek araştırmalara umut verici bir perspektif sunmaktadır.

Bilim Kodu	:	92432
Anahtar Kelimeler	:	Siber zorbalık, makine öğrenmesi, derin öğrenme, topluluk tespiti, sosyal ağ analizi
Sayfa Adedi	:	69
Danışman	:	Dr. Öğr. Üyesi Yılmaz ATAY

CYBERBULLYING DETECTION BASED ON SOCIAL NETWORK ANALYSIS AND MACHINE LEARNING

(M. Sc. Thesis)

Mümin Ferhat YAKUT

GAZİ UNIVERSITY

GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES

January 2024

ABSTRACT

The rapid advancement of technology has rendered internet usage an integral part of our lives. According to the "We Are Social Digital 2023" report, while the global internet usage rate stands at 64%, it reaches 84% in Turkey. Social media, one of the most popular facets of the internet, boasts approximately 4.7 billion users as of 2023, accounting for about 57% of the global population. As for Turkey, social media usage encompasses 60% of the total population. These data clearly illustrate the widespread adoption of the internet and social media. However, the increase in internet usage brings not only positive aspects but also some negative impacts. This study addresses one of these negative impacts: cyberbullying. Cyberbullying can be defined as a type of traditional bullying carried out using information technology tools. This type of bullying can adversely affect individuals' psychological and emotional well-being and lead to serious consequences in the society. The anonymity of users on social media applications and their ability to easily reach large audiences play a significant role in the increase of cyberbullying. The ease of interaction on these platforms not only increases individual bullying but also escalates bullying among communities, leading to more significant negative outcomes. Efforts to mitigate the effects of cyberbullying and detect such behaviors play a critical role in limiting the adverse impacts of technology. This thesis focuses on the definition, prevalence, and impacts of cyberbullying, aiming to contribute to a better understanding of the issue and the development of solutions. It is also studied that how the integration of artificial intelligence approaches, such as machine learning and deep learning, with community detection algorithms can improve the detection of cyberbullying. In the experimental study, eight different machine learning and deep learning algorithms were tested using various feature extraction and selection techniques, in which the highest success rate was achieved with a 90% accuracy using the Artificial Neural Networks algorithm in conjunction with CountVectorizer and Chi-Square techniques. When the Louvain algorithm was applied, it was observed that users sharing posts on similar topics tended to be in the same communities. This finding offers a promising perspective for future research to deepen and continue both community detection and cyberbullying detection studies.

Science Code	:	92432
Key Words	:	Cyberbullying, machine learning, deep learning, community detection, social network analysis
Page Number	:	69
Supervisor	:	Asst. Prof. Yılmaz ATAY

TEŞEKKÜR

Lisansüstü eğitim sürecince tez konusunun belirlenmesinde ve çalışmaların şekillenmesinde, yönlendirilmesinde, kontrolünde ve ihtiyaç duyulması halinde toplantılarla bilgilendirilmem konusunda yardımcı olan Dr. Öğr. Üyesi Yılmaz ATAY'a teşekkürlerimi sunuyorum.

Bu çalışmanın tamamlanmasında manevi desteğinin yanı sıra tez metnini edebi ve dil bilgisi yönünden inceleyen ve buna katkı sağlayan sevgili eşime, en büyük motivasyon kaynağım olan kızıma ve oğluma, beni yetiştiren ve hayatımın her aşamasında desteklerini hissettiren değerli aileme sonsuz teşekkürlerimi sunuyorum.

İÇİNDEKİLER

Sayfa

ÖZET iv

ABSTRACT v

TEŞEKKÜR vi

İÇİNDEKİLER vii

ÇİZELGELERİN LİSTESİ x

ŞEKİLLERİN LİSTESİ xi

SİMGELER VE KISALTMALAR xii

1. GİRİŞ 1

2. LİTERATÜR TARAMASI 7

2.1. Siber Zorbalığın Tespitinde Makine Öğrenmesi Çalışmaları 7

2.1.1. Kullanıcı ve sosyal ağ özelliklerinin sınıflandırmaya dahil edilmesi 9

2.1.2. Özellik çıkarım ve seçim tekniklerinin kullanılması 10

2.2. Siber Zorbalığın Tespitinde Derin Öğrenme Çalışmaları 11

2.3. Sosyal Ağlarda Topluluk Tespiti Çalışmaları 12

2.4. Siber Zorbalık Tespit Çalışmalarına Genel Bakış 13

3. MATERYAL VE METOT 15

3.1. Veri Setinin Hazırlanması 16

3.1.1. Veri toplama 17

3.1.2. Eğitim verisi 18

3.2. Ön İşleme Adımları 19

3.3. Özellik Çıkarım Teknikleri 20

3.4. Özellik Seçim Teknikleri 22

3.5. Sınıflandırma Algoritmaları 23

Sayfa

3.5.1. Lineer regresyon 24
3.5.2. Lojistik regresyon 26
3.5.3. SGD sınıflandırıcı 26
3.5.4. Karar ağaçları 27
3.5.5. Naive Bayes 28
3.5.6. Yapay sinir ağları 29
3.5.7. Evrişimli sinir ağları 30
3.5.8. Yinelemeli sinir ağları 31
3.5.9. LSTM 31
3.6. Topluluk Tespiti Algoritmaları 31
3.6.1. Louvain algoritması 32
3.7. Değerlendirme Metrikleri 33
3.7.1. Karmaşıklık matrisi 33
3.7.2. AUC-ROC eğrisi 35
3.8. Konu Tespiti Algoritmaları 35
3.8.1. LDA (Latent Dirichlet Allocation) algoritması 36
4. DENEYSEL ÇALIŞMALAR 37
4.1. Siber Zorbalık Tespiti ile İlgili Çalışmalar 37
4.2. Topluluk Tespiti ile ilgili Çalışmalar 41
4.2.1. Siber zorbalık skorunun oluşturulması 42
4.2.2. Ağ yapısının oluşturulması 43
4.2.3. Topluluk tespiti 45
4.2.4. Topluluk tespiti sonuçları 46
5. SONUÇ VE ÖNERİLER 51

Sayfa

KAYNAKLAR 57

EKLER..... 65

EK-1. Toplulukların konu tahminleri 66

ÖZGEÇMİŞ 69

ÇİZELGELERİN LİSTESİ

Çizelge **Sayfa**

Çizelge 2.1. Siber zorbalık çalışmalarının genel sunumu …….. 14

Çizelge 3.1. Eğitim veri setleri…… 18

Çizelge 3.2. Türkçe etkisiz kelimelerin listesi 20

Çizelge 3.3. Karmaşıklık matrisi..... 34

Çizelge 4.1. Makine ve derin öğrenme algoritmaların sonuçları 38

Çizelge 4.2. Sınıflandırılmış veri seti..... 39

Çizelge 4.3. Siber zorbalık skoru en yüksek kullanıcılar..... 43

Çizelge 4.4. Oluşturulan ağlara ilişkin istatistikler 44

Çizelge 4.5. Topluluk tespiti sonuçları 46

Çizelge 4.6. Graf3'te yer alan ilk 10 topluluk ve örnek üyeleri..... 46

Çizelge 4.7. Topluluklardaki kullanıcılara ait örnek paylaşımlar 48

Çizelge 4.8. Toplulukların tahmini konu dağılımları..... 50

Çizelge 5.1. Türkçe veri setleri ile yapılan siber zorbalık sınıflandırma çalışmaları..... 53

ŞEKİLLERİN LİSTESİ

Şekil **Sayfa**

Şekil 1.1. 2011-2018 yıllarına ait dünya genelinde siber zorbalık yayılımı 4

Şekil 1.2. Siber zorbalık platformları 5

Şekil 1.3. Google Trend verilerine göre siber zorbalık araştırmaları 5

Şekil 3.1. Siber zorbalık ve topluluk tespiti akış şeması 15

Şekil 3.2. Örnek veri seti 17

Şekil 3.3. Örnek takip/takipçi veri seti 17

Şekil 3.4. Örnek eğitim veri seti 19

Şekil 3.5. Lineer regresyon örneği 25

Şekil 3.6. Naive Bayes algoritması sözde kodu 28

Şekil 3.7. Yapay sinir ağları modeli 30

Şekil 3.8. Louvain algoritması sözde kodu 33

Şekil 4.1. Özellik çıkarım tekniklerine göre elde edilen en başarılı sonuçlar 40

Şekil 4.2. Özellik seçim tekniklerine göre elde edilen en başarılı sonuçlar 40

Şekil 4.3. Özellik çıkarım ve seçim tekniklerinin birlikte kullanım istatistiği 40

Şekil 4.4. ANN modelinin başarım grafiği 41

Şekil 4.5. Graf3 ağına ait genel görünüm 44

Şekil 4.6. Graf4 ağına ait genel görünüm 45

Şekil 4.7. Graf3 ağında yer alan ilk 10 topluluğun genel görünümü 47

Şekil 4.8. Örnek bir topluluk için LDA konu tespiti çıktısı 49

SİMGELER VE KISALTMALAR

Bu çalışmada kullanılmış kısaltmalar, açıklamaları ile birlikte aşağıda sunulmuştur.

Kısaltmalar	Açıklamalar
ANN	Yapay sinir ağları
AUC-ROC	İşlem karakteristiği eğrisi altında kalan alan
CNN	Evrişimli sinir ağları
DT	Karar ağaçları
GRU	Gated Recurrent Unit
KNN	K en yakın komşu
LDA	Latent Dirichlet Allocation
LSTM	Uzun kısa süreli bellek
LR	Lojistik regresyon
MAE	Ortalama mutlak hata
MNB	Multinomial Naive Bayes
MSE	Ortalama kare hata
NB	Naive Bayes
NLTK	Doğal dil işleme araçları kütüphanesi
RF	Rastgele orman
RNN	Yinelemeli sinir ağları
RMSE	Kök ortalama kare hata
SGD	Stochastic Gradient Descent
SVM	Destek vektör makineleri
TF-IDF	Terim Frekansı-Ters Belge Frekansı

1. GİRİŞ

Dijital platform zorbalığı, online zorbalık, sosyal medya zorbalığı ya da internet zorbalığı olarak da bilinen siber zorbalık, genel anlamda internet üzerinden yapılan farklı türdeki kötü niyetli ve zarar verici davranışları temsil eder. Bu tür zorbalıklarla mesajlaşma ortamlarında, sosyal medya platformlarında, online oyunlarda, e-posta iletişiminde veya diğer çevrimiçi ortamlarda sıklıkla karşılaşılır. Siber zorbalık, kişilere ya da kurum/kuruluşlara karşı saldırgan, aşağılayıcı, taciz edici, tehdit edici, hakaret veya argo içeren söz ve davranışları içerebilir. Bu tür bir zorbalık kişilerin, kurum/kuruluşların ve toplumların psikolojik ve duygusal durumlarını etkileyebilir. Ayrıca bazı durumlarda bu tür eylemler nefret söylemi, insanlık dışı davranışlar ve siber saldırı gibi hayati derecede önemli ve ciddi sonuçlara yol açabilir. Bu açıdan siber zorbalık farklı kavramları kapsayacak şekilde genişletilebilir. Siber zorbalığa maruz kalan bireyler genellikle haksızlıklara karşı psikolojik ve hukuksal destekler aramak ve bu tür davranışları raporlamak için farklı yöntemlere başvurabilirler. Devlet otoriteleri, sivil toplum kuruluşları, hizmet sağlayan firmalar, yazılım projeleri geliştiren kişi ve gruplar ile çevrimiçi platformlar, siber zorbalığı önlemek ve mücadele etmek için çeşitli politikalar, uyarı sistemleri ve önleyici tedbirler geliştirmeye çalışmaktadırlar. Bu amaçla bazı ülkelerde siber zorbalık ve akran zorbalığı eğitim müfredatlarına eklenmiştir.

Siber zorbalık, modern dünyada dijitalleşmeyle birlikte karşılaşılan önemli bir problemdir. Siber zorbalık, hedef alınan bireylerin psikolojik ve duygusal sağlıklarını etkileyebilir. Tehditler, hakaretler veya aşağılayıcı davranışlar, kurbanların güvenliğini ve refahını tehlikeye atabilir. Ayrıca bu tür bir zorbalık, bireyler arasında gerilim ve anlaşmazlıklara neden olabilir. Toplumsal düzeyde, bu tür davranışlar, insan ilişkilerini zedeleyebilir ve olumsuz bir atmosfer oluşturabilir. Okullarda öğrenciler arasında siber zorbalık, bu ortamlarda çok ciddi sorunlara yol açabilir. Eğitim kurumları, öğrenciler arasında sağlıklı ve güvenli bir ortam sağlamak adına bu konuda önlemler almak zorunda kalabilir.

Bir diğer önemli sorun da çevrimiçi platformlarda yaşanan siber zorbalıktır ve bu da dijital toplulukların güvenilirliğini zedeleyebilir. İnternet üzerinde güvenli bir ortamın korunması, kullanıcıların çevrimiçi etkileşimlerinde daha olumlu bir deneyim yaşamalarına yardımcı olabilir. Siber zorbalık, hukuki sonuçları olan bir davranış olarak karşımıza çıkmaktadır. Bu tür eylemler, yerel ve ulusal düzeyde suç olarak kabul edilebilmektedir. Bu nedenle siber

zorbalıkla mücadele; bireylerin, toplumların ve kurumların iş birliği içinde olması gereken çok yönlü bir çaba gerektirir. Kişisel eğitim ve farkındalık, anne ve baba eğitimleri, eğitim kurumlarından özel programlar, güvenli internet uygulamaları, gerektiğinde çevrimiçi izleme ve raporlama, online platformların güvenlik politikaları geliştirmesi, hukuki çerçevelerin çizilmesi gibi konularda çabalar siber güvenlik tehdidine karşı alınması gereken temel tedbirlerdendir. Siber zorbalıkla mücadele, sürekli bir çaba gerektirir ve teknolojik gelişmelerle birlikte güncellenmelidir. Çeşitli paydaşların bir araya gelmesi ve çözüm stratejilerini birlikte oluşturması, siber zorbalığı önleme çabalarının daha etkili olmasına katkı sağlar.

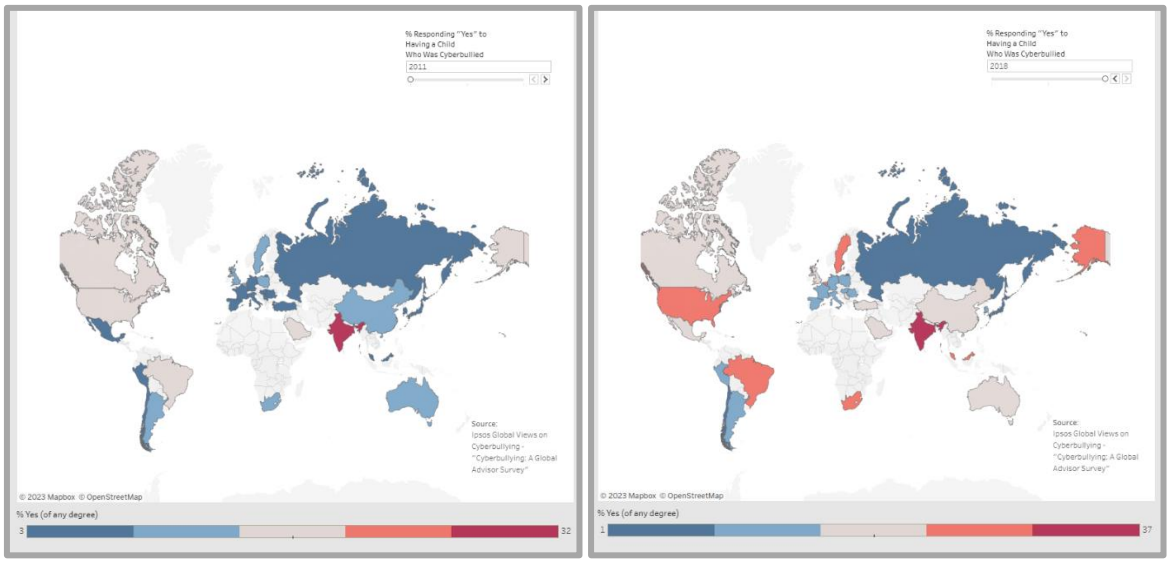
Siber zorbalık probleminin tespitine ve önlenmesine dair yapılacak çalışmaların temelini hiç şüphesiz zorbalık kavramının ve türlerinin tanımlanması oluşturmaktadır. Siber zorbalık kavramını tanımlamak için de geleneksel zorbalık konusunu ele almak gerekir. Geleneksel zorbalık, genellikle yüz yüze gerçekleşen ve bir kişi veya grup tarafından başka bir kişiye karşı sürekli ve kasıtlı olarak yapılan güç dengesizliği içeren ve tekrarlanan, saldırgan davranışları ifade eder [1]. Geleneksel zorbalığın temel özellikleri, fiziksel, sözel, psikolojik, sosyal ve siber zorbalık olarak incelenebilir. Fiziksel zorbalık; zorbanın zorbalığa maruz kalan mağdura vurma, itme, tekmeleme gibi uyguladığı fiziksel şiddeti içerir. Fiziksel zorbalık, en kolay fark edilen zorbalık türlerinden biridir çünkü genellikle somut yaralanmalara yol açar. Sözel zorbalık; hakaret, alay etme, tehdit etme, aşağılama gibi sözlü saldırıları içerir. Psikolojik zorbalık; korkutma, dışlama, dedikodu yapma, sosyal ilişkileri manipüle etme gibi davranışları içerir. Sosyal zorbalık; sosyal ilişkileri veya itibarı hedef alan, dışlama, sosyal izolasyon gibi kişinin sosyal çevresinde sorunlara neden olan davranışları içerir. Siber zorbalık ise geleneksel zorbalığın bilgi teknolojileri marifetiyle yapılması sonucunda ortaya çıkan bir zorbalık türüdür.

Kowalski ve Limber, siber zorbalığı, bir web sitesi veya cep telefonu aracılığıyla gönderilen mesajlarla yapılan zorbalık olarak tanımlarken, Hinduja ve Patchin siber zorbalığın sürekli, kasıtlı ve elektronik cihazlar aracılığıyla zarar verme amacıyla yapıldığını vurgulamıştır [2, 3]. Detaylı bir siber zorbalık tanımı bu alanda yapılacak çalışmaların temelini oluşturacağı için oldukça önemlidir. Chun, Lee J., Kim ve Lee S. çalışmalarında siber zorbalığın farklı tanımlarını incelemiş ve siber zorbalığı tanımlarken en sık kullanılan özellikleri; elektronik araçların kullanımı, tekrarlanan zarar veya davranış, kasıtlı veya bilinçli eylem, savunmasızlık, başkalarının istenmeyen bilgilerinin kullanımı ve tehdit edilmesi, taciz etmek

veya utandırmak olarak belirlemişlerdir [4]. Ülkemizde de emniyet birimleri tarafından siber zorbalık faaliyetlerinin önlenmesi amacıyla yapılan bilgilendirmelere göre, sosyal medyada bir başkasının rahatsızlık verici görüntülerini yayımlamak, hakkında yalan ya da uydurma haberler yaratmak ve bunları yaymak, özel ya da mahrem bilgilerini sızdırmak, tehdit mesajları yollamak, hakarete varan ifadeler kullanmak, anonim hesaplardan taciz etmek, rahatsızlık vermek, hesaplarını ele geçirmek bu hesaplardan uygunsuz yayınlar yapmak, ısrarla takip etmek, ya da paylaşımlarına kasıtlı olarak sürekli olumsuz yorumlar yapmak siber zorbalık olarak tanımlanmaktadır.

Gerçekleştirilme şekillerine ve sebep oldukları sonuçlara göre farklı siber zorbalık türleri vardır. Çevrimiçi platformlarda tehdit, korkutma ve taciz amaçlı rahatsız edici mesajların, paylaşımların veya yorumların sürekli olarak yapılmasını içeren *çevrimiçi taciz ve tehditler*, kişilerin özel bilgilerinin izni olmadan ve onu rahatsız edici şekilde başkaları tarafından paylaşılmasını içeren *kişisel bilgilerin ifşası (doxxing)*, kişiler hakkında yalan veya yanıltıcı bilgiler yayarak onlara zarar vermeyi hedefleyen *yalan bilgi yayma ve dedikodu*, bir başkasının kimliğini çevrimiçi olarak taklit ederek ona zarar vermeyi veya onun adına zararlı eylemlerde bulunmayı amaçlayan *kimlik hırsızlığı ve sahtekarlığı*, kişileri arkadaş ortamlarında dışlamayı, küçük düşürmeyi ve onlarla dalga geçmeyi içeren *dışlamak ve alay etmek*, kişileri zor duruma düşürerek çoğunlukla maddi kazanç sağlamayı hedeflemeyi içeren *şantaj yapmak* siber zorbalık türlerine örnek olarak sunulabilir.

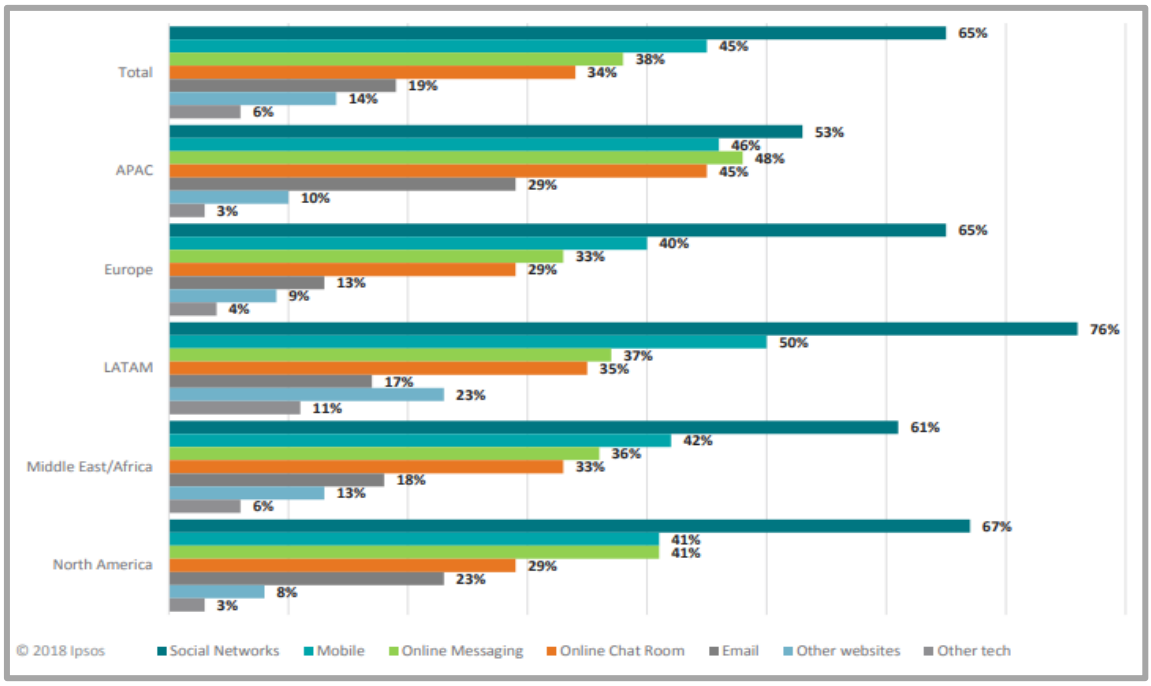
Teknolojinin günlük hayatımızdaki rolünün artması internet ve sosyal ağ uygulamalarının kullanımını tetiklemiş ve bununla birlikte siber zorbalığın kurbanlarının sayısı da paralel olarak artmıştır. Siber zorbalık davranışlarının yayılımına ilişkin yapılan çalışmalarda bu tür faaliyetlerin çok hızlı bir şekilde arttığı görülmüştür. PEW'in 2021 tarihli araştırmasına göre yetişkinlerin %40'ından fazlasının, 2022 tarihli araştırmasına göre de gençlerin %49'unun çevrimiçi ortamda siber zorbalık ve tacize maruz kaldığı görülmüştür [5,6]. Cook, dünya genelinde siber zorbalığa uğradığını ailelerine bildiren beş yaş üstü çocuğa sahip ebeveynlerin katıldığı çalışmada 2011 ve 2018 yıllarındaki değişimi Şekil 1.1'deki gibi sunmuştur [7].



Şekil 1.1. 2011-2018 yıllarına ait dünya genelinde siber zorbalık yayılımı [7]

Şekil 1.1 incelendiğinde, dünya genelinde 2011 yılında çocukları tarafından siber zorbalık şikâyeti alan ailelerin sayısının 2018'de gözle görülebilir ölçüde arttığı görülmektedir. Bu yükseliş günümüze kadar üstel bir şekilde artmaya devam etmiştir.

Siber zorbalık platformlarının belirlenmesi bu alanda yapılacak tespit ve önleme çalışmalarında çok büyük bir öneme sahiptir. Siber zorbalığın en yaygın olduğu ortamlar; sosyal medya siteleri ve uygulamaları, mobil iletişim araçları, çevrimiçi mesajlaşma uygulamaları ve konuşma odaları ile e-posta ortamlarıdır. Şekil 1.2'de bunlar arasında hem ülkemizde hem de dünya gelinde en fazla öne çıkanların sosyal medya uygulamaları ile mobil iletişim araçları olduğu görülmektedir [8]. Sosyal medya siteleri ve uygulamaları insanlara eğlenceli vakit geçirme ve bilgiye hızlı ulaşmak gibi faydalar sağlamaktadır. Ancak, bunların yanı sıra zorbalar için de kişilere ve kişiler hakkındaki özel bilgilere kolay ulaşmak, kendini kolay gizleyebilmek gibi imkânlar sağladığından siber zorbalık yapacak olan zorbalar tarafından çokça tercih edilmektedir. Bu kapsamda değerlendirildiğinde, siber zorbalık faaliyetleri için sosyal medya ortamlarının öne çıkmasının önemli sebeplerinin; bu ortamlarda kimlik belirtme zorunluluğu olmaması, sahte kimlikler ile kayıt yaptırılabilmesi ve hem siber zorbalık hem de sosyal medya siteleri ve uygulamalarında kullanıcıların gerçek kimlik bilgilerini bildirmesi ile ilgili yasal düzenlemelerin yetersiz olduğu görülmektedir.



Şekil 1.2. Siber zorbalık platformları [8]

İnternet ile birlikte siber zorbalık platformlarının (sosyal ağlar, çevrimiçi mesajlaşma uygulamaları vb.) da günlük hayatımızda giderek daha fazla yer alması, siber zorbalığın hızlı bir şekilde yayılımı ve kurbanlar üzerindeki ciddi etkileri dikkate alındığında siber zorbalığı tespit etme ve önleme çalışmaları da önem kazanmıştır. Şekil 1.3'de de sunulan Google Trends verilerine bakıldığında, dünya çapında siber zorbalık üzerine yapılan araştırmaların 2004'ten bu yana yaklaşık 20 kat arttığı görülmektedir. Bu durum siber zorbalık konusunun araştırmacılar tarafından da önemsendiğini göstermektedir.



Şekil 1.3. Google Trend verilerine göre siber zorbalık araştırmaları

Bunlara ek olarak, siber zorbalık problemini topluluklar açısından da ele almak gerekmektedir. Çünkü, siber zorbalık, bireysel eylemlerin ötesine geçerek topluluklar tarafından da gerçekleştirilebilen bir fenomen haline gelmiştir. İnternetin ve sosyal medya

platformlarının yaygınlaşması, bireylerin sadece tekil olarak değil, aynı zamanda gruplar halinde de zorbalık eylemlerine katılmalarını kolaylaştırmıştır. Bu tür, topluluk bazlı siber zorbalık faaliyetleri, genellikle forumlar, sosyal medya grupları veya çevrimiçi oyun toplulukları gibi ortak ilgi alanlarına sahip kişiler arasında meydana gelmektedir. Grup dinamikleri, bireylerin normalde yapmayacakları davranışları yapmalarına neden olabilir, çünkü grup içindeki baskı ve anonimlik hissi, bireysel sorumluluk duygusunu azaltır. Bu durum, hedef alınan bireylere karşı daha agresif ve zararlı eylemlere yol açabilir. Topluluklar halinde gerçekleştirilen siber zorbalık, hedef kişiler üzerinde daha yoğun ve yıkıcı etkilere sahip olabilir, çünkü birden fazla kişiden gelen saldırılar, kurbanın psikolojik sağlığı üzerinde daha derin izler bırakabilir.

Siber zorbalık tanımlaması yapılırken bir kişi veya grup tarafından başka bir kişi veya gruba yönelik saldırgan davranışların varlığı belirtilmiştir. Bu doğrultuda tez çalışmasının bir diğer amacı da bu grupların belirlenebilmesine yönelik katkı sağlamaktır. Deneysel çalışma sonucunda elde edilen sonuçlar gelecek çalışmalar için umut vadetmektedir.

Her bilimsel çalışmada olduğu gibi bu çalışmada da bazı sınırlılıklar vardır. Veri toplama aşamasında, sosyal ağ üzerinden verisi toplanan kullanıcı sayısının sınırlı olması ve oluşturulan ağ yapısında takip/takipçi bilgilerinin haricinde başkaca kullanıcı özelliklerinden faydalanılmamış olması en temel sınırlılıklarıdır. Ayrıca, çalışma kapsamında veri kaynağı olan X (eski adıyla Twitter), gündemde yer alan önemli konularında çokça konuşulduğu bir platformdur. Veri temin edilen tarihlerin 06.02.2023 tarihli Kahramanmaraş depremi sonrasında olmasından dolayı paylaşımların bir kısmı bu konu ile ilgilidir. Bu durum elde edilen topluluk sonuçlarında da görülmektedir. İlerleyen çalışmalarda bu tespitlerin anlık olarak yapılabileceği metotlara ve daha kapsamlı ağlar oluşturmaya yönelik çalışmalara devam ederek bu sınırlılıkların aşılması hedeflenmektedir.

Bu çalışmada, siber zorbalığın tespit edilmesinde makine öğrenmesi ve derin öğrenme algoritmaları ile topluluk tespit yöntemlerinin etkinliği araştırılmıştır. Bölüm 2’de bu alanda gerçekleştirilen literatür taraması sunulmuştur. Bölüm 3’te tez çalışmasında test edilen veri seti ve önerilen yaklaşımda yer alan makine öğrenmesi, derin öğrenme, topluluk tespit yöntemi ve performans ölçüm kriterlerine yer verilmiştir. Bölüm 4, önerilen metot ile gerçekleştirilen deneysel çalışmaların detaylarına ayrılmıştır. Bölüm 5'te ise çıkarımlar, analiz ve değerlendirme sonuçları ile potansiyel çalışma konuları sunulmaktadır.

2. LİTERATÜR TARAMASI

Siber zorbalığın tanımlanması ve etkileri 1970'lerden günümüze kadar çeşitli çalışmalara konu olmuş ancak son yıllarda internet kullanımının da yaygınlaşması ile daha popüler bir hal almıştır [3,9]. Siber zorbalığın etkilerinin ve yayılımının ölçümünü konu alan pek çok istatistiksel çalışma mevcuttur [5,6,10,11]. Tezin bu bölümünde siber zorbalığın tespit edilmesi üzerine makine öğrenmesi ve derin öğrenme algoritmalarının kullanımını ve sosyal ağ analizi teknikleri ile topluluk tespitini konu alan çalışmalar incelenmiştir.

2.1. Siber Zorbalığın Tespitinde Makine Öğrenmesi Çalışmaları

Siber zorbalığın hızlı yayılması ve kalıcı olumsuz etkilerinin olması, araştırmacıların siber zorbalık tespit problemine yönelik çalışmalar yapmak konusunda en önemli motivasyonları olmuştur. Bu alanda yapılan ilk çalışmalardan biri Yin ve arkadaşlarının chat-style (Kongregate) ve discussion-style toplulukları (Slashdot and Myspace) üzerinde makine öğrenmesi algoritmaları ile yaptıkları çalışmadır [12]. Araştırmacılar üç farklı veri seti toplamış ve özellik çıkarım tekniği olarak Terim Frekansı-Ters Belge Frekansı (TF-IDF) ve N-Gram tekniklerini test etmişlerdir. Deneysel çalışma sonuçlarına göre TF-IDF ve bağlamsal özelliklerin kullanılması sınıflandırıcı başarısını artırsa da yeterli ve ikna edici sonuçlar elde edilememiştir. Ancak, bu çalışma çevrimiçi taciz tespitine ilişkin bir veri seti oluşturulması ve bu soruna dikkat çekilmesi açısından gelecek çalışmalar için önemli bir rol oynamıştır.

Reynolds, Kontostathis ve Edwards, Formspring.me'den elde ettikleri soru-cevap veri setini incelemiş ve Weka uygulaması üzerinde C4.5, JRIP, Instance Based (IBK) ve Sequential Minimal Optimization (SMO) algoritmalarını test etmişlerdir. 296 farklı kelimenin yer aldığı kötü kelimeler listesi kullanılarak geliştirilen model metin özelliklerine odaklanmıştır. Verilerin etiketlenmesinde Amazon'un Mechanical Turk hizmeti kullanılmıştır. IBK ve C4.5 algoritmaları diğerlerine göre daha başarılı sonuçlar vermiştir [13].

Duygusal ifadelerin aktarılmasında sıkça kullanılan emojilerin ve çeşitli özellik seçimi tekniklerinin kullanılmasının sınıflandırma başarısını artırdığı gözlemlenmiştir [14,15]. Özel, Saraç, Akdemir ve Aksu, Instagram ve X (eski adıyla Twitter) üzerinden topladıkları

Türkçe metinlere Bilgi Kazanımı (Information Gain) ve Ki-Kare (Chi-Square) özellik seçme yöntemleriyle SVM, C4.5, MNB ve KNN algoritmalarını uygulamışlardır. Veri setinde yer alan emojiler ön işleme aşamasında verilerden çıkarılarak ikinci bir veri seti elde edilmiş ve bu iki veri seti üzerinde çalışmalar yapılmıştır. Elde edilen deneysel sonuçlarda emojilerin veri setine dahil edilmesi ve özellik seçme tekniklerinin uygulanması sınıflandırıcı başarısını artırmıştır. Bilgi Kazanımı tekniği Ki-Kare tekniğine göre daha başarılı sonuçlar vermiştir. Veri seti üzerinde uygulanan makine öğrenmesi algoritmalarının çalışma süreleri de değerlendirmeye dahil edildiğinde en başarılı performansı MNB algoritması göstermiştir [14].

Sintaha ve Mostakim yaptıkları çalışmada siber zorbalığın tespitine yönelik farklı yaklaşımları karşılaştırmayı amaçlamışlardır. Bunun için X (Twitter) ortamından siber zorbalığın tespiti ve önlenmesine ilişkin topladıkları veri setine NB ve SVM modellerini uygulamışlar ve veri setinde yer alan emojiler duyguları aktarmada oldukça etkili olduğundan paylaşımlardan kaldırılmak yerine belirli kodlarla eşleştirilerek analize dahil edilmiştir. Performans ölçümlerine göre SVM, NB algoritmasına göre %89,54 doğrulukla daha başarılı sonuçlar vermiştir [15].

Sahni ve Raja, Hintçe ve İngilizce olan X (Twitter) paylaşımları üzerinde siber zorbalığın tespiti için makine öğrenmesi ve duygu analizi yöntemlerini uygulamıştır. Deneysel çalışma sonuçlarına göre NB, RF ve J48 sınıflandırıcılarının üçü de %98 gibi yüksek bir başarı oranı göstermiştir [16].

Dalvi, Chavan ve Halbe, X (Twitter) API aracılığı ile elde ettikleri veri seti üzerinde Doğal Dil Araç Seti (NLTK) kullanılarak tokenizasyon yapmış ve TF-IDF tekniği ile özellik çıkarımı gerçekleştirmişlerdir. SVM ve NB sınıflandırıcılarıyla da tweetlerin olasılıkları üzerinden zorbalık içerip içermediğini tahmin etmişlerdir. Bu çalışmada, SVM'in en başarılı sınıflandırıcı olduğu gözlemlenmiştir [17]. Benzer bir çerçevede, Muneer ve Fati'nin 2020'deki araştırmasında da X (Twitter) ortamından toplanan verilerle yedi farklı makine öğrenme modelinin performansları karşılaştırılmıştır. Bu araştırmada, veri ön işleme adımında TF-IDF ile Word2Vec kullanılarak özellik çıkarımı gerçekleştirilmiştir. Bu çalışmada, Lojistik Regresyon (LR) en başarılı algoritma olarak ortaya çıkmıştır [18].

Siber zorbalığın tespiti ile birlikte zorbalığın türü ve şiddetinin ölçülmesi de ayrı araştırma konuları olarak karşımıza çıkmaktadır. Talpur ve O'Sullivan çalışmalarında, X'den (Twitter) aldıkları yaklaşık 50,000 tweet üzerinde siber zorbalığın türü ve şiddetinin belirlenmesini amaçlamışlardır. Çalışmada farklı özellik çıkarım yöntemleri (POS etiketleme, PMI, Duygusal, Sözlük vb.) ve makine öğrenmesi algoritmaları (SVM, KNN, DT, RF ve NB) kullanmışlardır. Siber zorbalığın türünün ve şiddetinin belirlenmesinde farklı siber zorbalık türlerine (cinsellik, ırkçılık, fiziksel görünüm, zekâ ve politika) ilişkin kelimelerden oluşan bir sözlük kullanmışlardır. Kullanılan sözlükteki konularla ilgili kelime zenginliğinin sınıflandırıcı başarısını etkilediği tespit edilmiştir [19].

2.1.1. Kullanıcı ve sosyal ağ özelliklerinin sınıflandırmaya dahil edilmesi

Siber zorbalığı tespit etmek ve sınıflandırmak için yapılan çalışmalarda, metinsel özellikler ile birlikte sosyal ağ veya kullanıcı özelliklerinin de kullanımı sınıflandırma başarısını artırmıştır [20-22]. Huang, Singh ve Atrey metinsel özellikler ile birlikte sosyal ağ özelliklerini kullanarak bileşik bir model önermişler ve sosyal ağ özelliklerinin kullanımının sınıflandırıcının başarısını arttırdığını gözlemlemişlerdir [21]. Dadvar, De Jong, Ordelman ve Trieschnigg Myspace platformundan elde ettikleri veri setini kullanarak siber zorbalığın tespitinde cinsiyete özgü dil özelliklerinin etkisini ölçmüş ve temel yöntemlere göre daha iyi sonuçlar verdiğini görmüştür [23]. Al-garadi, Varathan ve Ravana da siber zorbalığın tespiti için metinsel özellikler ile birlikte ağ, etkinlik, kullanıcı ve tweet içeriğini de kullanmışlardır. Çalışmada özellik seçimi için Ki-Kare, Bilgi Kazanımı ve Pearson Korelasyonu olmak üzere üç farklı algoritmayı test etmişlerdir. Buna ek olarak SMOTE ve maliyete duyarlı tekniklerle veri setinde yer alan sınıflar arasında dengeyi sağlamayı amaçlamışlardır. Sonuç olarak SMOTE ve Random Forest algoritmaları birlikte uygulandığında %94,3 AUC ve %93,6 F-Skor değerleriyle en başarılı sonuca ulaşılmıştır [24].

Balakrishnan, Arabnia ve Khan yaptıkları çalışmada #Gamergate etiketini kullanarak X (Twitter) ortamından 5453 paylaşım elde etmiş ve bunları veri seti olarak kullanmışlardır. Toplanan veri seti öncelikle uzmanlar tarafından dört sınıfa (zorba, saldırgan, spam ve normal) ayrılarak etiketlendi. Ayrıca Big Five ve Dark Triad modelleri kullanılarak kullanıcı kişilikleri tahmin edilmiş ve tahmin edilen kullanıcı kişilikleri ile siber zorbalık tespiti arasında ilişki kurulmaya çalışılmıştır. Duygu analizi teknikleri kullanılarak kullanıcıların duygu durumları da belirlenmeye çalışılmış ve kullanıcıların duygularının da siber zorbalık

üzerindeki etkisi ölçülmeye çalışılmıştır. Ancak siber zorbalık durumunun kullanıcıların duygularından daha çok kişiliklerinden etkilendiği gözlemlenmiştir. Eğitimli sınıflandırıcıların veri setindeki sınıflara göre orantısız dağılım nedeniyle bazı sınıfları daha iyi, bazılarını daha kötü tahmin etmelerini önlemek ve sınıflandırıcıların sağlamlığını arttırmak için 10 kat çapraz doğrulama tekniği (cross validation) kullanılmıştır. Elde edilen sonuçlara göre kullanıcı kişiliklerinin siber zorbalığın tespitini iyileştirdiği görülmüştür [25].

2021 yılında Bozyigit, Utku ve Nasibov'un yaptığı bir diğer çalışmada da X'den (Twitter) toplanan metinsel verilere ek olarak, kullanıcı ve gönderi özelliklerinin siber zorbalık tespitinde önemli olduğu gösterilmiştir. Ki-Kare tekniği ile sosyal medya özelliklerinin sınıflandırma sürecine olan etkisi ölçülmüştür. Bu teknikle, Retweet'ler, Favoriler, SenderFollowers ve SenderLocation özelliklerinin sınıflandırmada önemli olduğu tespit edilmiştir. Metinsel özellikleri ve sosyal medya özelliklerini içeren iki farklı veri seti hazırlanmıştır. Sonuç olarak, sosyal medya özelliklerinin sınıflandırma başarısını artırdığı belirlenmiştir [26].

2.1.2. Özellik çıkarım ve seçim tekniklerinin kullanılması

Sosyal ağlarda siber zorbalık tespiti yapılırken kullanılan sınıflandırma yöntemlerinin yanı sıra veri setinin oluşturulması, ön işleme adımlarının belirlenmesi ve kullanılan dilin özelliklerinin yapılan çalışmaya yansıtılması oldukça önemlidir. Çünkü sosyal ağlardan elde edilen veriler çok sayıda yazım yanlışı, duygusal ifadeler ve semboller barındırmaktadır. Uygulanacak ön işleme adımları ile veri setinde yer alan yazım yanlışlarının düzeltilmesinin sınıflandırıcı başarısını artırdığı görülmüştür [27, 28]. Bozyiğit, Utku ve Nasiboğlu, sosyal medya yorumlarından Türkçe içerikli bir veri seti hazırlayarak araştırmacıların kullanımına sunmuştur. Bu veri seti üzerinde veri ön işleme aşamasında, yazım hatası olan kelimelerin düzeltilmesi için 144 adet Türkçe hakaret ve küfür içeren sözlük ve Levenshtein Distance algoritması kullanılmıştır. Böylece veri setinde yer alan yazım yanlışları düzeltilmiş veya veri setinden çıkarılmıştır. Özellik çıkarımı için TF-IDF ve Bilgi Kazanımı tekniği ile birlikte Kelime Çantası (Bag of Words) tekniği kullanılmıştır. NB, SVM, RF, KNN, MNB ve C4.5 algoritmaları test edilmiş ve F-Score değeri ile performans ölçümü yapılmıştır. Sınıflandırıcıların çalışma süreleri ve F-Skor değerleri dikkate alındığında en başarılı yöntemlerin MNB, SVM ve KNN olduğu görülmüştür [27].

2019'da John Hani ve arkadaşları da siber zorbalığın tespiti ve önlenmesi için bir makine öğrenimi yaklaşımı önermiş ve çalışmanın veri ön işleme adımında tokenizasyon, gereksiz kelimelerin çıkarılması ve kelimelerin doğru yazılışlarının bulunması konularına yer verilmiştir [29].

Bu alanda yapılan çalışmalarda veri setlerinin büyük olmasından dolayı veri ön işleme ve özellik çıkarım adımları hem sınıflandırma başarısı hem de performans açısından önemli bir rol oynamaktadır [30-32]. Veri setinde yer alan bütün özniteliklerin sınıflandırma aşamasında kullanılması sınıflandırıcının çalışma süresini oldukça uzatacaktır. H. Rosa ve arkadaşlarının sunduğu çalışmada, Bullying Traces Dataset ve Formspring olmak üzere iki farklı veri seti üzerinde farklı özellik çıkarım yöntemlerinden oluşan 12 farklı senaryo test edilmiştir. İlk veri seti için en başarılı sınıflandırıcı, senaryo A: TF-IDF, senaryo D: TF-IDF + Kelime Yerleştirmeleri, senaryo H: TF-IDF + Kişilik Özellikleri + Kelime Yerleştirmeleri ile hayata geçirilen SVM modeli olmuştur. İkinci veri seti için ise en başarılı sınıflandırıcı yine "senaryo H" ile uygulanan SVM olmuştur [33].

2.2. Siber Zorbalığın Tespitinde Derin Öğrenme Çalışmaları

Büyük veri setleri ile çalışma zorunluluğu araştırmacıları çeşitli özellik çıkarma ve seçme adımlarını kullanmaya veya geleneksel makine öğrenmesi tekniklerinin yerine derin öğrenme tekniklerini kullanmaya sevk etmiştir [27]. Siber zorbalığın tespiti ile ilgili hem geleneksel makine öğrenmesi hem de derin öğrenme algoritmaları farklı çalışmalarda kullanılmıştır [17,18]. Derin öğrenme yöntemlerinin bu alandaki performansının incelendiği çalışmada, Çift Yönlü Uzun Kısa Süreli Bellek (BLSTM), GRU, LSTM ve RNN gibi derin öğrenme algoritmaları test edilmiş ve bu algoritmalar arasında BLSTM en başarılı sonuçları verdiği gözlemlenmiştir [34].

Kargutkar ve Chitre siber zorbalığı tespit etmek için CNN algoritmasını kullanmışlardır. Çalışmada tasarlanan model, veri ön işleme, CNN model katmanı ve model tahmin katmanı olmak üzere üç katmandan oluşmaktadır. CNN modelinde aktivasyon fonksiyonu olarak ReLu fonksiyonu kullanılmıştır. Geleneksel makine öğrenmesi modellerine göre daha başarılı sonuçlar elde edilmiştir [35].

Al-Ajlan ve Ykhlef geleneksel yöntemlerde kullanılan özellik çıkarma ve seçme adımlarını atlayan bir derin öğrenme yöntemini siber zorbalık tespiti için geliştirmiş ve oldukça başarılı sonuçlar elde etmişlerdir [36].

2.3. Sosyal Ağlarda Topluluk Tespiti Çalışmaları

Siber zorbalık, sadece bireyler arasında değil, aynı zamanda topluluklar içinde de endişe verici bir sorun haline gelmiştir. Topluluklar aracılığı ile gerçekleşen siber zorbalık olayları, bireysel zorbalık vakalarından daha ciddi sonuçlar doğurabilir. Bu durum sosyal ağlarda topluluk tespiti problemini de beraberinde getirmiştir. Girvan ve Newman karmaşık ağların içindeki topluluk yapılarının tanımlanması ve tespit edilmesi üzerine yaptıkları çalışmada ağda yer alan kenarların önem sırasını kullanarak topluluk yapılarını tespit eden bir algoritma sunmuşlardır [37]. Blondel, Guillaume, Lambiotte ve Lefebvre ise modülerlik skorunu optimize ederek büyük ağlardaki topluluk yapılarını tespit etmeye yarayan Louvain Algoritmasını tanıtmışlardır [38]. Burada her düğüm başlangıçta kendi başına bir topluluğun parçası kabul edilir ve ağın modülerlik skoru hesaplanır. Daha sonra her düğüm sırasıyla ele alınır ve ait olduğu topluluk değiştirilir. Eğer yapılan değişiklik modülerlik skorunu artırıyor ise bu işlem kabul edilir ve düğümün ait olduğu topluluk güncellenir.

Choi, Jeon ve Kim siber zorbalık tespiti ile birlikte karmaşık ağlardaki topluluk yapılarını ele aldıkları çalışmalarında Daum Agora platformu üzerinden topladıkları 7 farklı topluluğa ait veri seti üzerinde metin madenciliği, aşağılayıcı dil tespiti ve sosyal ağ analizi kombinasyonunu kullanmışlardır. Araştırmacılar, Losada oranı (Pozitif/Negatif paylaşım oranı) ve Cyberbullying Index (Zararlı yorumların Kullanıcı/Topluluk oranı) değerleri ile birlikte sosyal ağ analizini kullanarak sadece siber zorbalığı tespit etmekle kalmadı, aynı zamanda topluluk içindeki kullanıcıların katılım düzeyini ve davranışını ölçmeyi hedefledi [39].

Park ve Kwon X (Twitter) üzerinden topladıkları veri seti üzerinde topluluk tespiti ve metinsel analiz yöntemlerini birlikte kullanmışlar ve siber saldırı tespiti ile birlikte toplulukların siber saldırılar ile olan ilişkisini de tespit etmeye çalışmışlardır. Elde edilen sonuçlar metin analizi ve topluluk tespitinin etkinliğini ortaya koymuştur [40]. Li, Bai, Zhang, Tang ve Luo da yaptıkları çalışmada metin bilgisini ve ağ ilişkilerini bir araya getirerek görüş topluluklarını ve liderlerini tanımlama konusunda yeni bir yaklaşım

sunmaktadır. Araştırmacılar, içerik ve zaman benzerliği ile kullanıcıların topolojik yapısını kullanarak hesaplanan entegre benzerlik değerini görüş topluluklarını belirlemede kullanmışlardır. Daha sonra aynı görüş topluluğunda bulunan kullanıcılar için oluşturulan yönlü graf kullanılarak her kullanıcı için etki değeri hesaplaması yapılır ve en yüksek etki değerine sahip kullanıcı görüş lideri olarak belirlenir [41]. Hristakieva, Cresci, Martino, Conti ve Nakov da 2019 UK genel seçimleri ile ilgili twitter veri seti üzerinde kullanıcı etkinlik ve etkileşimlerine dayanan koordinasyon skoru, tweet içeriklerine dayanan propaganda skoru ve Louvain Algoritiması ile hem propaganda gruplarını oluşturmuş hem de propaganda yayılımı hakkında çıkarımlarda bulunmuşlardır [42].

2.4. Siber Zorbalık Tespit Çalışmalarına Genel Bakış

Siber zorbalığın tespit edilmesi ve önlenmesi ile ilgi yapılan literatür taramasında bu alanda yapılan çalışmaların çoğunlukla İngilizce dilinde hazırlanan veri setleri ile yapıldığı görülmektedir. Buna ek olarak, nadir de olsa Arapça, Türkçe, İspanyolca vb. dillerde de yapılan çalışmalar vardır. Bu durumun temel sebebi İngilizce veri setlerinin daha kolay erişilebilir olmasıdır. Çalışmanın yapılacağı dile ilişkin edebi ve dil bilgisi özelliklerinin de çalışma sonuçlarını etkilediği görülmektedir. Çizelge 2.1 incelendiğinde İngilizce dilinde yapılan çalışmaların başarım oranlarının Türkçe dilinde yapılanlara göre genellikle daha yüksek olduğu görülmektedir. Bu durumun sebebi hem çalışmalarda kullanılan veri setlerinin kapsamı hem de Türkçe'nin sondan eklemeli bir dil oluşudur.

Ayrıca, çalışmalarda farklı veri kaynakları kullanılsa da çalışmaların büyük çoğunluğu veriye kolay erişim sağlanabildiği için veri kaynağı olarak X'i kullanmaktadır. Ancak, bu platformlardan elde edilen veriler genellikle günlük konuşma dilinde olduğu için sınıflandırma başarısını artırmak için veri ön işleme adımları oldukça önemli bir role sahip olmaktadır. Buna ek olarak, çalışmalarda kullanılan veri setlerinin çok büyük olması sınıflandırma aşamasında performans sorunlarına sebep olmaktadır. Bu durum da özellik çıkarım ve seçim tekniklerinin kullanımının önemini göstermektedir

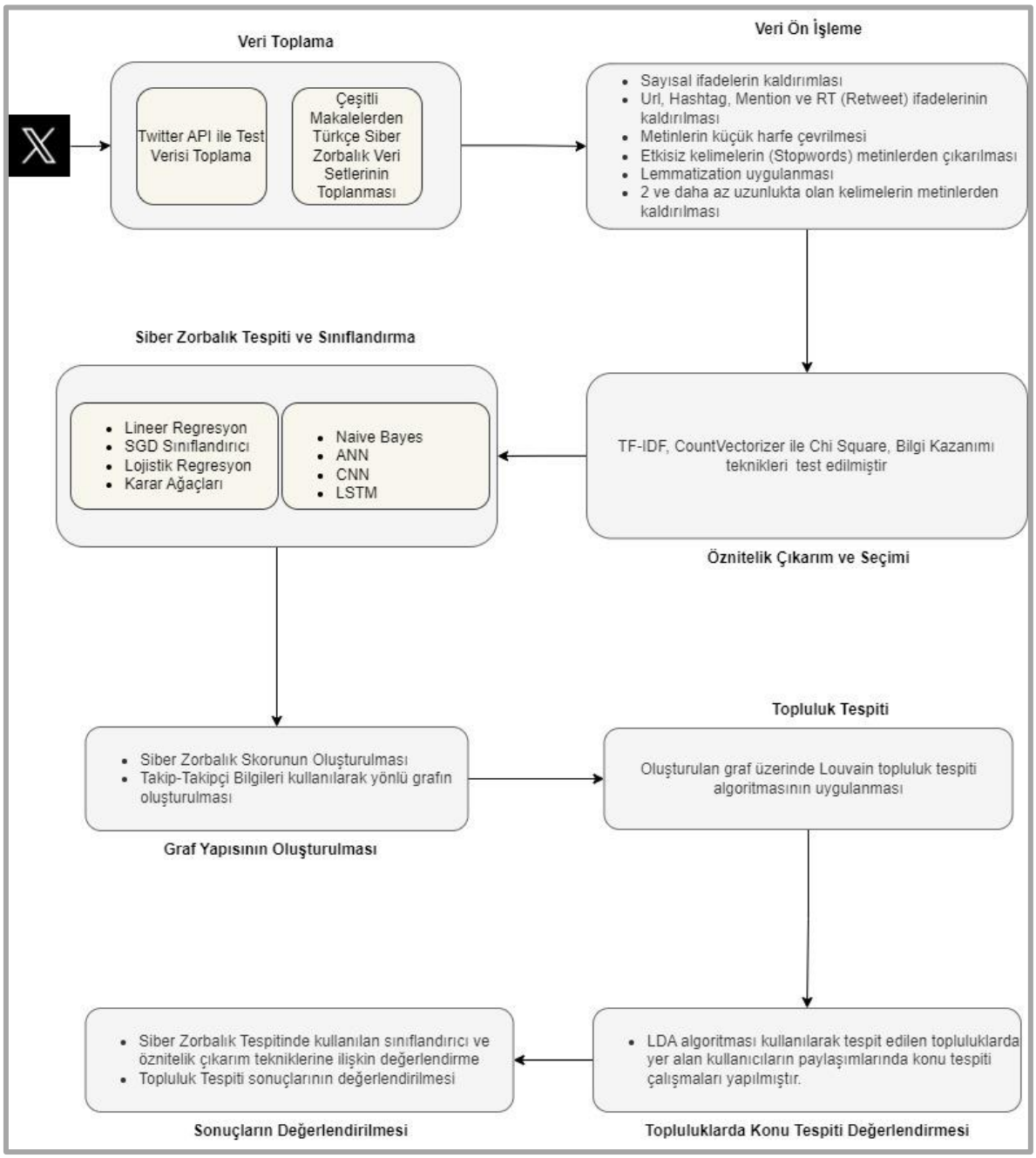
Çizelge 2.1 bu alanda yapılan çalışmalara ilişkin; kullanılan veri setleri, özellik çıkarım ve seçim teknikleri ile başarım oranlarını sunmaktadır.

Çizelge 2.1. Siber zorbalık çalışmalarının genel sunumu

Çalışma	Platform	Dil	Özellik Çıkarım/Seçim Teknikleri	Değerlendirme Ölçütü	Başarılı Sınıflandırıcı
[14]	Instagram, X	Türkçe	Chi-Square and Information Gain	F- Skor: %81	MNB
[15]	X	İngilizce	-	Doğruluk: %89,5	SVM
[16]	X	Hintçe, İngilizce	N-Gram, Sentiment Analysis	Doğruluk: %98 Kappa: %98	NB, RF, J48
[18]	X	İngilizce	TF-IDF, Word2Vec	Doğruluk: %90,5 F- Skor: %92,8	LR
[19]	X	İngilizce	POS tagging, PMI, Sentiment, Lexicon	Doğruluk: %91,1 F- Skor: %89,8	RF
[24]	X	İngilizce	Chi-Square, Information Gain and Pearson Correlation	AUC: %94,3, F- Skor: %93,6	Proposed Method
[26]	X	Türkçe	BagOfWords, TF-IDF	Doğruluk: %90,1	AdaBoost
[27]	X	Türkçe	BagofWords, TF-IDF and Information Gain	F- Skor: %91	SVM
[28]	X	Türkçe	TF-IDF, N-Gram	F-Skor: %91	ANN
[31]	Youtube	Arapça	Count Vectorizer, TF-IDF Vectorizer	F-Skor: %78,6	Count Vectorizer +LR
[36]	X	İngilizce	-	Doğruluk: %95	CNN-CB (Önerilen Yaklaşım)
[43]	X	İspanyolca	TF-IDF, N-Gram	Doğruluk: %80 F- Skor: %80	SVM
[44]	Formspring.me, Myspace	İngilizce	TF-IDF, N-Gram	F- Skor: %95	SGD
[45]	Formspring.me	İngilizce	TF-IDF, Word2Vec	F- Skor: %83,2 ROC: %92	RF
[46]	X	Türkçe	Word2Vec, FastText	F- Skor: %93	CNN

3. MATERYAL VE METOT

Bu bölümde, tez çalışmasında sunulan metodoloji ve test kapsamı ele alınmıştır. Veri setinin oluşturulması, ön işleme adımları, sınıflandırma ve topluluk tespit yaklaşımları açıklanmıştır. Siber zorbalık ve topluluk tespitine ilişkin çalışmanın ana akış şeması Şekil 3.1'de sunulmuştur.



Şekil 3.1. Siber zorbalık ve topluluk tespiti akış şeması

Çalışmada önemli konulardan biri olan veri toplama aşaması, X ortamından kullanıcılara ait paylaşımların ve takip/takipçi bilgilerinin küçük kod parçaları (*scriptler*) marifetiyle toplanması ve açık kaynaklardan Türkçe siber zorbalık tespiti için eğitim verilerinin derlenmesi adımlarından oluşmaktadır. Veri ön işleme aşamasında ise siber zorbalık tespiti öncesinde veri setinin temizlenmesi adımlarına yer verilmiştir. Üçüncü aşamada, özellik çıkarım ve özellik seçim teknikleri uygulanmıştır. Daha sonra sınıflandırma aşamasında belirlenen makine ve derin öğrenme algoritmaları ile siber zorbalık tespiti yapılmış ve elde edilen sonuçlar ışığında graf yapısının oluşturulması adımına geçilmiştir. Bu aşamada, kullanıcıların takip/takipçi bilgileri kullanılarak ağ yapısı oluşturulmuştur. Topluluk tespiti aşamasında Louvain algoritması kullanılarak oluşturulan ağ üzerinde topluluk tespiti çalışmaları yapılmıştır. Tespit edilen topluluklarda sadece takip/takipçi ilişkisi değil aynı zamanda konu birlikteliğinin de varlığının araştırılması için LDA (Latent Dirichlet Allocation) algoritması kullanılarak topluluklarda konu tespiti değerlendirmesi yapılmıştır. Son aşamada ise elde edilen sonuçlar sunularak değerlendirmelere yer verilmiştir.

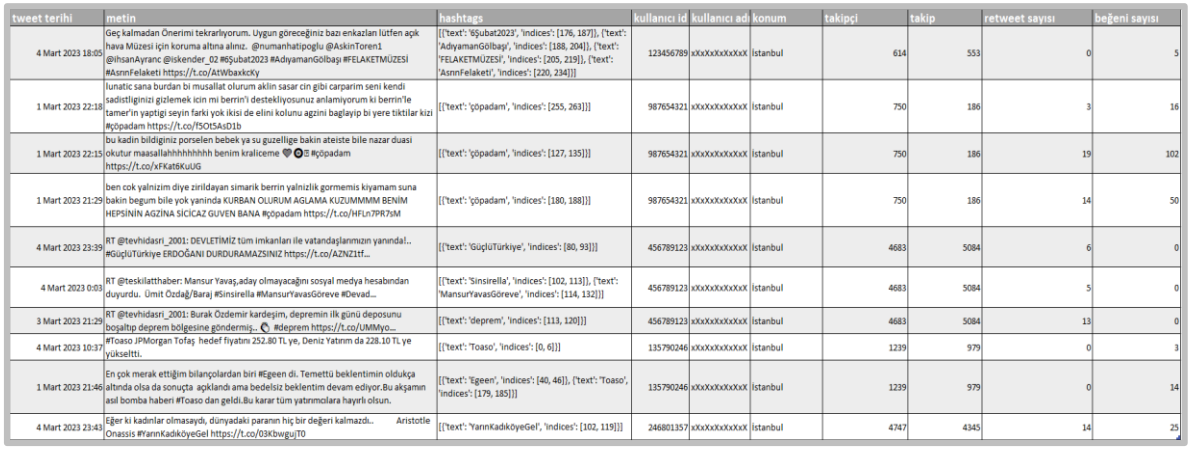
3.1. Veri Setinin Hazırlanması

Siber zorbalığı tespit ve önlemek için ilgili veri setinin toplanması ve etiketleme yapılarak eğitim verisinin oluşturulması ana gereksinimdir. Gerekli veriyi toplamak için en yaygın kullanılan yöntem çok sayıda kullanıcının yer aldığı sosyal ağlar tarafından sağlanan servislerdir. Bu servisler arasında ise Instagram API, Twitter Streaming API, twitter4j API başta gelmektedir. Ancak bazı çalışmalarda araştırmacılar özel olarak geliştirilmiş yazılım uygulamaları ile de bu sosyal ağlardan veri toplamıştır [15,27, 43]. Oluşturulan veri seti ve eğitim verisinin özelliklerinden dolayı farklı veri setlerinde farklı sınıflandırıcı başarım oranlarına ulaşılmıştır. Bu aşamada oluşturulan veri setinde kullanılan dilin özelliklerine hâkim olmak ve model geliştirirken bu özellikleri göz önünde bulundurmak oldukça önemlidir.

Veri setinin oluşturulmasında önemli olan bir diğer adımda veri etiketlemedir. Veri etiketleme, denetimli öğrenmede kullanılan veri setindeki kayıtların etiketlenme sürecidir [18]. Etiketleme genellikle uzmanlar veya kullanıcılar tarafından manuel olarak yapıldığı için maliyetli bir süreçtir [47,48]. Bu sebepten dolayı bazı derin öğrenme algoritmaları etiketleme yerine otomatik olarak öznitelikleri üretir ve öznitelik çıkarım adımlarının maliyetini düşürür [49].

3.1.1. Veri toplama

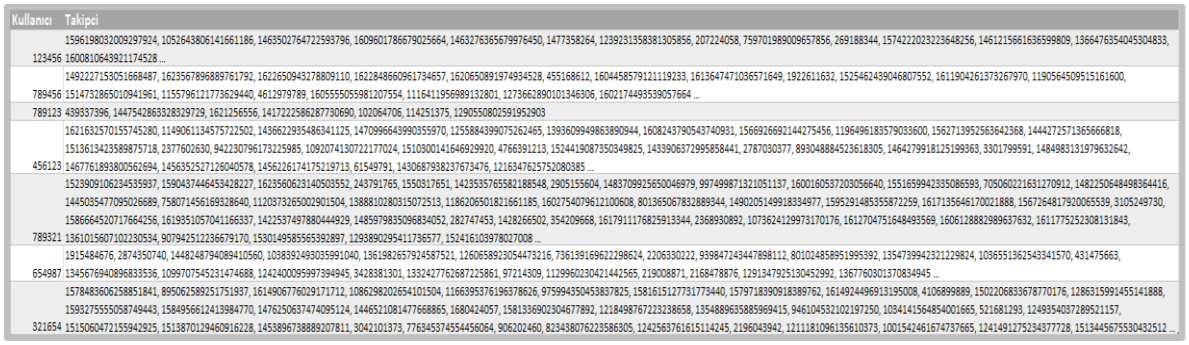
Bu çalışmada veri seti oluşturulurken “tweepy” kütüphanesi yardımıyla Twitter API (Application Programming Interface-Uygulama Programlama Arabirimi) kullanılmıştır. Geliştirilen scriptler ile 267 384 kullanıcının X (eski adıyla Twitter) hesaplarından yapmış oldukları 37 235 001 paylaşım veri setine dahil edilmiştir. Toplanan veri setinde paylaşımlar ile birlikte bazı kullanıcı özellikleri de API aracılığı ile toplanmış ve veri setine eklenmiştir. Toplanan veri setinden bir örnek Şekil 3.2’de sunulmuştur. Veri seti içerisinde yapılan paylaşıma ilişkin tarih, içerik, hashtag, mention, retweet ve beğeni sayısı gibi bilgiler yer alırken, kullanıcıya ilişkin takip-takipçi sayısı ve konum bilgisi bulunmaktadır.



tweet terihi	metin	hashtags	kullanıcı id	kullanıcı adı	konum	takipçi	takip	retweet sayısı	beğeni sayısı
4 Mart 2023 18:05	Geç kalmadan Önerimi tekrarlıyorum. Uygun göreceğiniz bazı enkazları lütfen açık hava Müzesi için koruma altına alınız. @numanhatipoglu @AskinToren1 @ihsanAyranc @iskender_02 #6Şubat2023 #AdıyamanGölbaşı #FELAKETMÜZESİ #AsnnFelaketi https://t.co/AtWbaxkcKy	[{'text': '6Şubat2023', 'indices': [176, 187]}, {'text': 'AdıyamanGölbaşı', 'indices': [188, 204]}, {'text': 'FELAKETMÜZESİ', 'indices': [205, 219]}, {'text': 'AsnnFelaketi', 'indices': [220, 234]}]	123456789	xXxXxXxXxXxX	İstanbul	614	553	0	5
1 Mart 2023 22:18	lunatic sana burdan bi musallat olurum aklin sasar cin gibi carparim seni kendi sadistliginizi gizlemek icin mi berrin'i destekliyosunuz anlamiyorum ki berrin'le tamer'in yaptigi seyin farki yok ikisi de elini kolunu agzini baglayip bi yere tiktilar kizi #çöpadam https://t.co/f5Ot5AsD1b	[{'text': 'çöpadam', 'indices': [255, 263]}]	987654321	xXxXxXxXxXxX	İstanbul	750	186	3	16
1 Mart 2023 22:15	bu kadin bildiginiz porselen bebek ya su guzellige bakin ateiste bile nazar duasi okutur maasallahhhhhhhhh benim kraliceme 🤍🧿🎖 #çöpadam https://t.co/xFKat6KuUG	[{'text': 'çöpadam', 'indices': [127, 135]}]	987654321	xXxXxXxXxXxX	İstanbul	750	186	19	102
1 Mart 2023 21:29	ben cok yalnizim diye zirildayan simarik berrin yalnizlik gormemis kiyamam suna bakin begum bile yok yaninda KURBAN OLURUM AGLAMA KUZUMMMM BENİM HEPSİNİN AGZİNA SİCİCAZ GUVEN BANA #çöpadam https://t.co/HFLn7PR7sM	[{'text': 'çöpadam', 'indices': [180, 188]}]	987654321	xXxXxXxXxXxX	İstanbul	750	186	14	50
4 Mart 2023 23:39	RT @tevhidasri_2001: DEVLETİMİZ tüm imkanları ile vatandaşlarımızın yanında!.. #GüçlüTürkiye ERDOĞANI DURDURAMAZSINIZ https://t.co/AZNZ1tf...	[{'text': 'GüçlüTürkiye', 'indices': [80, 93]}]	456789123	xXxXxXxXxXxX	İstanbul	4683	5084	6	0
4 Mart 2023 0:03	RT @teskilatthaber: Mansur Yavaş,aday olmayacağını sosyal medya hesabından duyurdu. Ümit Özdağ/Baraj #Sinsirella #MansurYavasGöreve #Devad...	[{'text': 'Sinsirella', 'indices': [102, 113]}, {'text': 'MansurYavasGöreve', 'indices': [114, 132]}]	456789123	xXxXxXxXxXxX	İstanbul	4683	5084	5	0
3 Mart 2023 21:29	RT @tevhidasri_2001: Burak Özdemir kardeşim, depremin ilk günü deposunu boşaltıp deprem bölgesine göndermiş.. #deprem https://t.co/UMMyo...	[{'text': 'deprem', 'indices': [113, 120]}]	456789123	xXxXxXxXxXxX	İstanbul	4683	5084	13	0
4 Mart 2023 10:37	#Toaso JPMorgan Tofaş hedef fiyatını 252.80 TL ye, Deniz Yatırım da 228.10 TL ye yükseltti.	[{'text': 'Toaso', 'indices': [0, 6]}]	135790246	xXxXxXxXxXxX	İstanbul	1239	979	0	3
1 Mart 2023 21:46	En çok merak ettiğim bilançolardan biri #Egeen di. Temettü beklentimin oldukça altında olsa da sonuçta açıklandı ama bedelsiz beklentim devam ediyor.Bu akşamın asıl bomba haberi #Toaso dan geldi.Bu karar tüm yatırımcılara hayırlı olsun.	[{'text': 'Egeen', 'indices': [40, 46]}, {'text': 'Toaso', 'indices': [179, 185]}]	135790246	xXxXxXxXxXxX	İstanbul	1239	979	0	14
4 Mart 2023 23:43	Eğer ki kadınlar olmasaydı, dünyadaki paranın hiç bir değeri kalmazdı.. Aristotle Onassis #YarınKadıköyeGel https://t.co/03KbwgujT0	[{'text': 'YarınKadıköyeGel', 'indices': [102, 119]}]	246801357	xXxXxXxXxXxX	İstanbul	4747	4345	14	25

Şekil 3.2. Örnek veri seti

Ayrıca, kullanıcılar arasındaki ilişkilerin belirlenmesinde kullanılmak üzere tweet bilgileri toplanmış olan kullanıcıların takip ve takipçi bilgileri de aynı API aracılığı ile toplanmıştır. Bu verinin bir örneği de Şekil 3.3’de sunulmuştur.



Kullanıcı	Takipci
123456	1596198032009297924, 1052643806141661186, 1463502764722593796, 1609601786679025664, 1463276365679976450, 1477358264, 1239231358381305856, 207224058, 759701989009657856, 269188344, 1574222023223648256, 1461215661636599809, 1366476354045304833, 1600810643921174528 ...
789456	1492227153051668487, 1623567896889761792, 1622650943278809110, 1622848660961734657, 1620650891974934528, 455168612, 1604458579121119233, 1613647471036571649, 1922611632, 1525462439046807552, 1611904261373267970, 1190564509515161600, 1514732865010941961, 1155796121773629440, 4612979789, 1605555055981207554, 1116411956989132801, 1273662890101346306, 1602174493539057664 ...
789123	439337396, 1447542863328329729, 1621256556, 1417222586287730690, 102064706, 114251375, 1290550802591952903
456123	1621632570155745280, 1149061134575722502, 1436622935486341125, 1470996643990355970, 1255884399075262465, 1393609949863890944, 1608243790543740931, 1566926692144275456, 1196496183579033600, 1562713952563642368, 1444272571365666818, 1513613423589875718, 2377602630, 942230796173225985, 1092074130722177024, 1510300141646929920, 4766391213, 1524419087350349825, 1433906372995858441, 2787030377, 893048884523618305, 1464279918125199363, 3301799591, 1484983131979632642, 1467761893800562694, 1456352527126040578, 1456226174175219713, 61549791, 1430687938237673476, 1216347625752080385 ...
789321	1523909106234535937, 1590437446453428227, 1623560623140503552, 243791765, 1550317651, 1423553576558188548, 2905155604, 1483709925650046979, 997499871321051137, 1600160537203056640, 1551659942335086593, 7050602211631270912, 1482250648498364416, 1445035477095026689, 758071456169328640, 1120373265002901504, 1388810280315072513, 1186206501821661185, 1602754079612100608, 801365067832889344, 1490205149918334977, 1595291485355872259, 1617135646170021888, 1567264817920065539, 3105249730, 1586664520717664256, 1619351057041166337, 1422537497880444929, 1485979835096834052, 282747453, 1428266502, 354209668, 1617911176825913344, 2368930892, 1073624129973170176, 1612704751648493569, 1606128882989637632, 1611775252308131843, 1361015607102230534, 907942512236679170, 1530149585565392897, 1293890295411736577, 1524161039778027008 ...
654987	1915484676, 2874350740, 1448248794089410560, 1038392493035991040, 1361982657924587521, 1260658923054473216, 7361391696229862 4, 2206330222, 939847243447898112, 801024858951995392, 1354739942321229824, 1036551362543341570, 431475663, 1345676940896833536, 1099707545231474688, 1242400095997394945, 3428381301, 1332427726887225861, 97214309, 1129960230421442565, 219008871, 2168478876, 1291347925130452992, 1367760301370834945 ...
321654	1578483606258851841, 895062589251751937, 1614906776029171712, 1086298202654101504, 1166395376196378626, 9759943504538378 25, 1581615127731773440, 1579718390918389762, 1614924496913195008, 4106899889, 1502206833678770176, 1286315991455141888, 1593275555058749443, 1584956612419984770, 1476250637474095124, 1446521081477668865, 1680424057, 1581336902304677892, 1218498767223238658, 1354889635885969415, 9461045321021972 50, 1034141564854001665, 521681293, 1249354037289521157, 1515060472155942925, 1513870129460916228, 1458896738889207811, 3042101373, 7763453745544560 64, 9062024 60, 8234380762235863 05, 1242563761615114245, 2196043942, 1211181096135610373, 1001542461674737665, 1241491275234377728, 1513445675530432512 ...

Şekil 3.3. Örnek takip/takipçi veri seti

3.1.2. Eğitim verisi

Denetimli öğrenme yöntemlerinin uygulanabilmesi için etiketlenmiş eğitim verisi gerekmektedir. Daha önceki çalışmalar incelendiğinde veri etiketleme işlemlerinin genellikle alanında uzman, veri setinin ait olduğu dilin özelliklerini bilen, zorbalık ve siber zorbalık kavramlarına hâkim kişilere yaptırılmakta olduğu görülmüştür [25,43,44]. Ancak, daha önce de belirtildiği gibi veri setinin hazırlanması ve etiketlenmesi oldukça zor, uzun ve maliyetli bir süreçtir. Dolayısıyla araştırmacıların bir kısmı da hazır veri setleri üzerinde çalışmalar yapmıştır [45].

Bu çalışmada da eğitim verisi olarak üç farklı veri seti birleştirilmiş ve tekrarlı kayıtlar veri setinden çıkarılarak kullanılmıştır. Kullanılan veri setlerine ilişkin detaylı açıklamalar Çizelge 3.1'de verilmiştir.

Çizelge 3.1. Eğitim veri setleri

Adı	Pozitif	Negatif	Toplam	Kaynak
Compherensive cyberbullying dataset including_social_media_attributes	2 500	2 500	5 000	[26]
csvnew.csv	6 115	5 004	11 119	[50]
turkish-cyber-bullying	1 497	1 503	3 000	[51]
Toplam	9 630	8 390	18 020	

Burada belirtilen veri seti adı veri setinin paylaşıldığı çalışmada veya açık kaynaklarda belirtilen isimlerdir. Pozitif etiketi metinlerin zorbalık içermediğini, Negatif etiketi ise zorbalık içerdiğini göstermektedir. Toplamda 19 119 metinden oluşan eğitim veri setinde tekrarlı kayıtlar kaldırıldığında 18 020 etiketli metin kalmıştır. Sonuç olarak oluşan eğitim veri setinin bir örneği Şekil 3.4'de sunulmuştur.

Metin	Etiket
doğa ağzımıza sıçsa hakkı var	1
anne bir sanatçıdır güzel eseri yavrusudur	0
ibrahimin oğlunu koruyan tanrı ben sokayım	1
köpeğim suratına sıçsın senin namussuz karı	1
ben söğüşledim birazda sen söğüşle	1
şerefsizlik sözde sanatçıların vazgeçemediği bir değerdir	1
kendisi bilmiyordur beyinsiz	1
aşırı bıyıklı geleneksel keko adam bi ayrılıp bi barışan boş bakışlı salak çift direnaj ertuğrul sezon dışın	1
uzun yolculuklar bile tek bir adımla başlar geleceğin mimarları gençlerimiz atılan ilk adımları var gücü	0
vallahi aralar istediğimiz tek huzur	0
chp genel başkanı kemal kılıçdaroğlu nun eşi selvi kılıçdaroğlu hazine maliye bakanı berat albayrak ın	0
beyin göçünü engellemeye çalıştığımız gençlere sosyal medyayı yasaklamak gerçekten dahiyane bir f	0
doğa ana kullanılmayan depoyu devralıyor	0
tarlasını çekerken beni çekmedin söylenenen dayının kendinin çekildiğinden habersiz oluşu	0
babanesini seven yoktur heralde anneneler kadar minnoşsa baanneler kadar çekilmez oluyo inşlah b	1
merhaba konuyla ilgili yardımcı olmak isteriz rtibat numaranızı direkt mesaj olarak bize iletir misiniz	0
nietzsche böyle buyurdu zerdüşt te önce sen kendini inşa etmelisin dimdik bir beden dimdik bir ruhla	0
hiçbir harekete geçen cehalet kadar korkunç olamaz	0
yine birileri değişimlerin tadını çıkarmakla meşgul	0
stanley parable bir oyun var abi sevebilirsin cümlesini unutamadım bir eskiden ha evet bir skyrim diyc	0
seni kitap okuyan insanlarla tanıştıracağım hayat ancak böyle insanlarla bir araya geliyorsan yaşanm	0
yaptığım başarılı aktivite uyumak	0
demiş iyilerle kötüler belli değil it itle gezer yiğit yiğitle	1
bir faydası olmayan zararı olan saray soytarısı nagehan fikrini beyan ediyor	1
ulan yanan alevi yakan alevi bugün temel aynı çanaktan yal yiyorsun olacak	1

Şekil 3.4. Örnek eğitim veri seti

3.2. Ön İşleme Adımları

Güvenilir ve kullanışlı veri setleri oluşturmakla birlikte sınıflandırıcı başarısını artırmak ve çalışma süresini azaltmak için ön işleme adımları kaçınılmazdır [28]. Literatür incelendiğinde, noktalama işaretleri, hashtagler, semboller, etkisiz kelimeler, URL'ler ve özel karakterlerin veri setinden kaldırılması ve tüm kayıtların küçük harfe dönüştürülmesi işlemlerinin en yaygın ön işleme adımları olduğu görülmüştür [17,18,52]. Kök bulma ve lemmatizasyon gibi kelimelerin değişmesine yol açan ön işleme adımları ise veri setinde kullanılan dilin özelliklerine göre tercih edilmelidir [18,31,54].

Bu çalışma kapsamında da hem eğitim veri seti hem de Twitter API aracılığı ile toplanan veriler de bazı ön işleme adımlarına tabi tutulmuştur. Veri ön işleme aşamasında;

- Öncelikle veri setinde yer alan sayısal ifadeler, Url, Hashtag, Mention, RT ve FAV gibi metnin temel anlamını etkilemeyen ifadeler kaldırılmıştır.

- Noktalama işaretleri, semboller ve sayısal ifadeler de veri setinden atılmıştır.
- Bütün veri seti küçük harflere çevrilmiştir.
- Daha sonra nltk kütüphanesinde bulunan 53 adet "turkish" etkisiz kelime de veri setinden kaldırılmıştır. Bu etkisiz kelimeler listesi Çizelge 3.2'de sunulmuştur.
- Veri setinde tokenizasyon uygulanarak geliştirilen modelin kelimeler arasındaki ilişkileri daha iyi öğrenmesi amaçlanmıştır.
- Tokenler arasında uzunluğu iki ve daha az olan ifadeler Türkçe'de anlamlı kelimeler olmamasından dolayı veri setinden atılmıştır.

Veri setinden anlamsız ve sınıflandırmayı etkilemeyecek ifadelerin kaldırılması hem sınıflandırıcı başarısını artırmada hem de daha az özellik ortaya çıkacağından doğal olarak çalışma süresini de azaltmada etkili olmuştur.

Çizelge 3.2. Türkçe etkisiz kelimeler listesi

acaba	ama	aslında	az	bazı	belki
biri	birkaç	bir şey	biz	bu	çok
çünkü	da	daha	de	defa	diye
eğer	en	gibi	hem	hep	hepsi
her	hiç	için	ile	ise	kez
ki	kim	mı	mu	mü	nasıl
ne	neden	nerde	nerede	nereye	niçin
niye	o	sanki	şey	siz	şu
tüm	ve	veya	ya	yani	-

3.3. Özellik Çıkarım Teknikleri

Makine öğrenimi uygulamaları, büyük miktarda veriyi işlemek ve bu veriden anlamlı bilgiler elde etmek için özellik çıkarımı ve seçim tekniklerini kullanmaktadır. Bu teknikler, modelin eğitim sürecini optimize etmek, performansını artırmak ve genelleme kabiliyetini geliştirmek için kritik bir rol oynar [53-55].

Ozel ve Sarac 2017'de gerçekleştirdikleri çalışmada, siber zorbalığı tespit etmede bazı özellik çıkarımı ve seçim tekniklerinin etkisini araştırdılar. Formspring.me veri setinde farklı tokenizasyon, etkisiz kelimeleri kaldırma, kök bulma ve küçük harfe çevirme tekniklerinin kombinasyonlarını, özellik seçimi için de Bilgi Kazancı ve Ki-Kare tekniklerini test etmişlerdir [56]. Elde edilen sonuçlara göre kök bulma ve etkisiz kelimelerin kaldırılması, kullanılan veri seti için sınıflandırıcının başarısını azalttı. Ancak, bir çalışma için

kullanılacak ön işleme adımlarının çalışmada kullanılan veri setinin ve bu veri setinin hazırlandığı dilin özelliklerine bağlı olarak değişebileceği açıktır.

Bu çalışmada, siber zorbalık tespiti için 8 farklı makine ve derin öğrenme algoritması TF-IDF ve CountVectorizer özellik çıkarımı teknikleri ile ayrı ayrı test edilmiştir. Bu iki teknikten kısaca bahsetmek gerekirse, TF-IDF bir kelimenin metindeki önemini, kelimenin metinde ne kadar sık kullanıldığına (TF) ve tüm veri setindeki kullanımına (IDF) dayanarak hesaplar [57, 58].

$$\mathrm{TF}(\mathrm{t}, \mathrm{d}) = \frac{\text{t kelimesinin d metnindeki frekansı}}{\text{d metnindeki toplam kelime sayısı}} \tag{3.1}$$

$$\mathrm{IDF}(\mathrm{t}, \mathrm{D}) = \log \frac{\text{Veri setinde yer alan toplam metin sayısı}}{1 + \text{t kelimesini içeren metin sayısı}} \tag{3.2}$$

Eş. 3.1'de belirli bir kelimenin bir metindeki frekansını vermektedir. Eş. 3.2 ise belirli bir kelimenin veri setindeki frekansına odaklanır. TF-IDF değeri ise bu iki ifadenin çarpımı olarak ifade edilmektedir.

CountVectorizer ise metin verisini hızlı ve kolayca sayısal bir forma dönüştürmek için kullanışlıdır. Belirli bir kelimenin bir belgede kaç kez geçtiğini sayarak belgeleri sayısal bir formata dönüştürmeyi hedefler. Öncelikle veri setinde yer alan metinlere ilişkin bir kelime dağarcığı oluşturulur ve her bir metin için dağarcığa eklenen kelimelerin bulunma durumları göz önüne alınarak metinlere ilişkin vektörler oluşturulur [59].

Örnek olarak, şu üç cümleyi ele alalım:

- "Makine öğrenmesi başladı."
- "Makine öğrenmesi algoritmaları başarılıdır."
- "Algoritmaları öğrenmeye başladı."

Bu cümleleri CountVectorizer ile işlediğimizde, aşağıdaki kelime dağarcığını elde ederiz:
{"makine", "öğrenmesi", "başladı", "algoritmaları", "başarılıdır", "öğrenmeye"}

Bu dağarcığa göre, cümlelerin sayısal temsili;

- "Makine öğrenmesi başladı." için [1, 1, 1, 0, 0, 0],
- "Makine öğrenmesi algoritmaları başarılıdır." için [1, 1, 0, 1, 1, 0] ve

- "Algoritmaları öğrenmeye başladı." için [0, 0, 1, 1, 0, 1] şekilde olacaktır.

3.4. Özellik Seçimi Teknikleri

Özellik seçimi (*feature selection*), makine öğrenimi uygulamalarında hayati bir öneme sahiptir. Özellik seçimi, bir modelin eğitimi için kullanılacak özelliklerin seçilmesi sürecini ifade eder. Özellik seçimi uygulamasının önemini vurgulayan bazı nedenler şu şekildedir.

- Model Performansını İyileştirme: Modelin çalışacağı özelliklerin sınıflandırma için daha değerli özellikler olarak belirlenmesi ve sonucu etkilemeyeceği değerlendirilen özelliklerin kaldırılması hem test hem de eğitim aşaması için modelin doğruluğunu ve performansını iyileştirebilir.
- Modelin Yorumlanabilirliğini Artırma: Daha az özellikle daha basit bir model elde edilecektir. Bu, modelin yorumlanmasını ve anlaşılmasını kolaylaştırır, bu da modelin nasıl çalıştığını ve hangi özelliklerin önemli olduğunu anlamamıza yardımcı olur.
- Aşırı Öğrenme (Overfitting) Riskini Azaltma: Aşırı sayıda özellik, modele aşırı uydurma riski getirebilir, yani model, eğitim verisine çok spesifik olarak uyarlanabilir ve yeni verilere genelleme yapmakta zorlanabilir. Özellik seçimi, bu riski azaltabilir.
- Veri Anlayışını Derinleştirme: Özellik seçimi süreci, hangi özelliklerin önemli olduğunu ve hangilerinin olmadığını belirlememize yardımcı olabilir. Bu, problemle ilgili derinlemesine bilgi edinmemize yardımcı olabilir.
- Bellek ve Depolama Maliyetini Azaltma: Daha az özellik, daha az bellek ve depolama alanı gerektirir. Bu, özellikle büyük veri kümeleriyle yapılan çalışmalarda oldukça faydalıdır [60,61].

Bu çalışmada makine ve derin öğrenme algoritmaları ile test edilen iki farklı özellik seçimi tekniği vardır. Bunlar; Ki-Kare ve bilgi kazanımıdır.

Ki-Kare testi, istatistiksel olarak iki kategorik değişken arasındaki bağımsızlığı test etmek için kullanılan yaygın bir testtir. Kısacası, bu test iki kategorik değişken arasında bir ilişkinin olup olmadığını belirlemek için kullanılır. Makine öğrenimi bağlamında, Ki-Kare testi, özellik seçimi amacıyla değişkenlerin hedef değişkene olan bağımsızlığını test etmek için kullanılır [62].

Bilgi kazanımı da makine öğrenimi uygulamalarında model performansını iyileştirmek için kullanılan entropi temelli bir tekniktir. Bir özellik sınıflardan ne kadar bağımsız ise bilgi kazanımı o kadar küçük olacaktır ve bu da özelliğin sınıflandırıcıyı çok az etkileyeceğini göstermektedir.

3.5. Sınıflandırma Algoritmaları

Makine öğrenmesi ve derin öğrenme yöntemleri yapay zekâ araştırmalarının en önemli bileşenlerindendir. Son yıllarda bu tekniklerin yaygın kullanımı ve başarılı sonuçları birçok alanda dikkat çekmektedir.

Makine öğrenmesi, bilgisayarların, özel programlama olmaksızın veriden öğrenmeyi ve öğrendikleri bilgiyle karar vermeyi öğrenmelerini sağlayan bir yapay zekâ dalıdır [63]. Denetimli, denetimsiz, pekiştirmeli ve yarı denetimli öğrenme olmak üzere 4 farklı türü vardır.

Denetimli öğrenmede, algoritma etiketlenmiş veri setinden öğrenir. Yani her örnek girişi için beklenen çıkışı da biliriz [64]. Lineer Regresyon, Lojistik Regresyon, Destek Vektör Makineleri, Karar Ağaçları ve Rastgele Ormanlar en yaygın denetimli öğrenme algoritmalarıdır.

Denetimsiz öğrenmede ise algoritma etiketsiz veri setinden yapıları veya desenleri öğrenmeye çalışır [65]. Bu alanda en yaygın kullanılan algoritmalar ise k-Ortalama, Hiyerarşik Kümeleme, DBSCAN, Bağımsız Bileşen Analizi (ICA) ve Prensip Bileşen Analizi (PCA)'dir.

Bir başka makine öğrenmesi türü olan pekiştirmeli öğrenmede, algoritma belirli bir amaca ulaşmak için ödüllendirilir veya cezalandırılır. Öğrenme süreci deneme yanılma yöntemiyle gerçekleşir [66]. Popüler pekiştirmeli öğrenme algoritmaları, Q-Öğrenme, Derin Q Ağı (DQN) ve Aktör-Kritik (AC) algoritmalarıdır.

Son olarak yarı denetimli öğrenmede hem etiketli hem de etiketsiz veriler kullanılır. Genellikle etiketsiz veri miktarı, etiketli veri miktarından çok daha fazladır [67]. En çok kullanılan algoritmalar ise yarı denetimli destek vektör makineleri ve graf-tabanlı yöntemlerdir.

Son zamanlarda oldukça yaygın bir kullanıma sahip olan derin öğrenme ise makine öğreniminin bir alt dalı olup çok katmanlı yapay sinir ağları kullanarak karmaşık veri yapılarından öğrenir. Derin öğrenme algoritmaları, geleneksel makine öğreniminden daha

karmaşık görevleri daha hızlı bir şekilde yerine getirebilen özellik çıkarım yetenekleri sayesinde son yıllarda popülerlik kazanmıştır [68]. Özellikle büyük veri setleri ile çalışan araştırmacılar derin öğrenme yöntemlerini test etmişlerdir [69,70]. Derin yapay sinir ağları, evrişimli sinir ağları, tekrarlayan sinir ağları, oto-kodlayıcılar ve derin takviye öğrenme derin öğrenme türleridir [71].

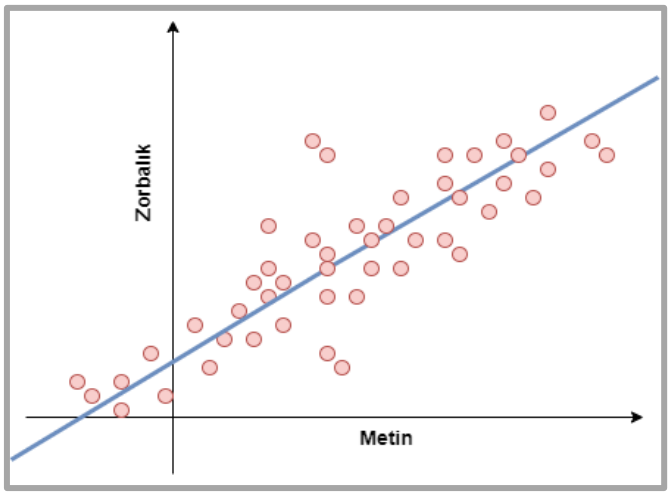
Yapılan literatür taramasında siber zorbalık tespiti, metin sınıflandırma, duygu analizi gibi konularda kullanılan popüler makine ve derin öğrenme algoritmaları belirlenerek bu çalışmada da test edilmiştir. Bu kapsamda, tez çalışmasında kullanılan makine ve derin öğrenme algoritmalarına aşağıda yer verilmiştir.

3.5.1. Lineer regresyon

Lineer regresyon, bir veya birden fazla bağımsız değişken ile bir bağımlı değişken arasındaki ilişkiyi modellemek için kullanılan istatistiksel bir yöntemdir. Bu yöntemin temel amacı, bağımlı ve bağımsız değişkenler arasında en iyi uyumu sağlayan bir doğru veya düzlem bulmaktır.

$$y = \beta_0 X_0 + \beta_1 X_1 + \cdots + \beta_n X_n + b \tag{3.3}$$

Eş. 3.3'de lineer regresyon kullanımına ilişkin eşitlik verilmiştir. Burada y bağımlı değişkeni, X_0, $X_1, \ldots, X_n$ bağımlı değişkenleri, β_0, $\beta_1, \ldots, \beta_n$ bağımsız değişkenlerin ağırlık katsayılarını, b değeri ise öteleme değerini ifade eder. Hata payı hesaplamasına göre ağırlık katsayıları ve öteleme değeri güncellenir ve en uygun eşitlik bulunur [72,73]. Şekil 3.5'de örnek olarak, bağımsız değişkenlerin ağırlıklarının ve öteleme değerinin değiştirilmesi ile bağımlı değişkene en yakın sonuçları veren doğru görselleştirilmiştir.



Şekil 3.5. Lineer regresyon örneği

Bağımlı değişkenlerin ağırlıklarının ve öteleme değerinin güncellenmesi için hata payı hesaplaması yapılmalıdır. En yaygın hata payı hesaplamaları şu şekildedir.

- Ortalama mutlak hata (Mean Absolute Error): Gerçek değerler ile tahmin edilen değerler arasındaki mutlak farkların ortalamasıdır.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - y_i^l| \tag{3.4}$$

- Ortalama kare hata (Mean Squared Error): Gerçek değerler ile tahmin edilen değerler arasındaki farkların karesinin ortalamasıdır.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - y_i^l)^2 \tag{3.5}$$

- Kök ortalama kare hata (Root Mean Squared Error): MSE'nin kareköküdür ve modelin hatalarının ölçeklenmiş bir versiyonunu sağlar.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - y_i^l)^2} \tag{3.6}$$

3.5.2. Lojistik regresyon

Lojistik regresyon, genellikle iki kategorili sonuç değişkenleri için kullanılan bir istatistiksel yöntemdir. Lineer regresyonun sınıflandırma problemlerine uygulanabilmesi için değiştirilmiş halidir.

$$\sigma(z) = \frac{1}{1+ e^{-z}} \tag{3.7}$$

Burada, z lineer fonksiyonu ifade etmektedir. Lineer fonksiyon sonucunda üretilen değer kullanılarak elde edilen σ(z) değeri gerçek değerli bir girdiyi 0-1 aralığına eşler. Bu nedenle olasılık tahmini olarak yorumlanır. Tahmin edilen olasılık belirli bir eşik değerinden büyükse, bir sınıfa; küçük ise diğer sınıfa atama yapılır [74,75].

3.5.3. SGD Sınıflandırıcı

Stochastic Gradient Descent (SGD) sınıflandırıcısı büyük ölçekli veri setlerini sınıflandırmak için kullanılan basit ama etkili bir yöntemdir. SGD, her bir eğitim örneğini tek tek alarak maliyet fonksiyonunun türevini hesaplar ve modelin ağırlıklarını veya parametrelerini bu türevin tersi yönünde günceller [76]. SGD, büyük veri setleri üzerinde çalışmak için etkili bir yöntemdir, çünkü her adımda tüm veri setini işlemek yerine yalnızca bir örneklem kullanır.

SGD algoritmasının çalışmasına bir örnek ile açıklayacak olursak;
x = [1, 2, 3, 4, 5] bağımlı değişkenleri,
y = [2, 4, 5, 4, 5] bağımsız değişkenleri ile tek bağımsız değişken olduğu için y = ax + b modeli geliştirilsin. Modelin a ve b parametrelerini en iyi şekilde tahmin edecek SGD algoritmasının adımları aşağıdaki gibidir.

1. Model parametreleri için başlangıç değerleri atanır. (a=1, b=2)
2. Veri setinden rastgele bir örneklem seçilir. (x=2, y=4)
3. Yeni a ve b değerlerini Eş. 3.8 ve 3.9'da olduğu gibi hesaplanır.

$$a = a - n * (y – (a*x + b)) \tag{3.8}$$

$$b = b - n* (y – (a *x + b)) \tag{3.9}$$

Burada; n öğrenme oranını ifade eder ve bu adımlar modelin parametreleri kararlı bir değere ulaşana kadar tekrarlanır. Sonuç olarak; a=0,98 ve b=1,97 değerleri elde edilir ve model y=0,98x+1,97 olarak belirlenir.

3.5.4. Karar ağaçları

Karar ağaçları, veri kümesini daha basit alt kümelerine bölmek için bir dizi karar kurallarını takip eder. Bu kural seti, özelliklere dayalı soruların bir hiyerarşisi şeklinde sunulur. Örneğin, bir hayvanı sınıflandırmak için "bu hayvanın kanatları var mı?" gibi bir soru sorulabilir.

Karar ağaçlarının eğitimi sırasında, hangi özelliğin veri kümesini en iyi böleceğine karar vermek için entropi, bilgi kazancı veya gini gibi ölçümler kullanılır [77,78].

Entropi, bir kümenin homojenliğini ölçmek için kullanılır ve Eş. 3.10'da belirtiği gibi hesaplanır. Tamamen homojen bir küme için entropi değeri 0'dır.

$$E(D) = -\sum_{i=1}^{c} p_i \, log_2 p_i \tag{3.10}$$

Bilgi kazancı ise bir özelliğin seçilmesiyle elde edilen saflık artışını ölçer. Yüksek bilgi kazancına sahip özellikler, karar ağacının üst seviyelerinde tercih edilir.

$$IG(D, f) = \mathrm{E(D)} - \sum_{j=1}^{m} \frac{N_j}{N} \mathrm{E}\left(D_j\right) \tag{3.11}$$

Gini saflığı da rastgele seçilen bir öğenin yanlış sınıflandırılma olasılığını ölçer. Karar ağaçları oluşturulurken düşük gini değerine sahip özellikler tercih edilir.

$$G(D) = 1 - \sum_{i=1}^{c} p_i^2 \tag{3.12}$$

3.5.5. Naive bayes

Naive Bayes algoritması, verilen bir girdi özellik seti için olası çıktıları yani sınıfları tahmin etmek için kullanılan bir sınıflandırma algoritmasıdır. Bu varsayım, gerçek dünyadaki veri setlerinde her zaman geçerli olmasa da algoritmanın etkinliğini artırır.

Bir x özellik kümesinin C sınıfına ait olma olasılığı Eş. 3.13'de olduğu gibi hesaplanır.

$$P(C|x) = \frac{P(x|C)*P(C)}{P(x)} \tag{3.13}$$

Burada; P(x|C), sınıf C olduğunda özellik setinin x olma olasılığını, P(C) sınıf C olma ve P(x) de özellik setinin x özellik seti olma olasılığını ifade eder [79]. Naive Bayes algoritmasına ilişkin sözde kod Şekil 3.6'da verilmiştir.

Naive Bayes Sözde Kodu:

Veri: Eğitim seti (X, y), Sınıf C ve test örneği x_test için;

1. Sınıf önceliklerini hesapla:
Her sınıf c ∈ C için P(c) = Sınıf c'deki örnek sayısı / Toplam örnek sayısı

2. Özellik olasılıklarını hesapla:
Her özellik f ∈ X, her sınıf c için P(f|c) hesapla:
P(f|c) = Sınıf c'deki f özelliğini taşıyan örnek sayısı / Sınıf c'deki toplam örnek sayısı

3. Test örneği için olasılıkları hesapla:
Her sınıf c için, P(c|x_test) olasılığını hesapla:
x_test f özelliğini taşıyorsa;
P(c|x_test) *= P(f|x_test)
x_test f özelliğini taşımıyorsa;
1-P(f|c)

4. En yüksek olasılığa sahip sınıfı seç:
y_pred = argmax_c P(c|x_test)

5. y_pred'i döndür.

Şekil 3.6. Naive Bayes algoritması sözde kodu

Naive Bayes'in basitliği, hızı ve genellikle yüksek doğrulukta sonuçlar vermesi, onu birçok uygulama için popüler bir seçenek haline getirir. Ancak, özelliklerin bağımsızlık varsayımının her zaman doğru olmadığı durumlarda, performansı etkilenebilir. Bu nedenle,

algoritmanın kullanılacağı veri setine ve soruna uygun olup olmadığını değerlendirmek önemlidir.

3.5.6. Yapay sinir ağları

Yapay sinir ağları, beyindeki biyolojik nöronların bilgi işleme yöntemlerini taklit etmeye çalışan bilgisayar sistemleridir. Temel yapı taşı, "nöron" adı verilen basit işlem birimidir. Bu nöronlar, girdi değerlerini alır, bu girdilere ağırlık değerleri uygular, toplam bir değer hesaplar ve bu değeri bir aktivasyon fonksiyonuna sokarak çıktıyı üretir.

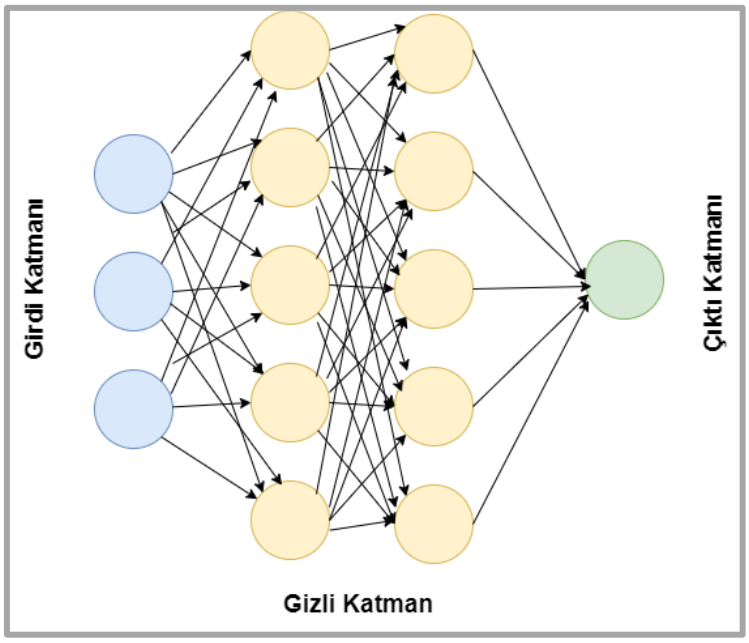
Bir nöronun matematiksel modellemesi Eş. 3.14'de gösterildiği gibidir.

$$y = \mathrm{f}(\sum_{i=1}^{n} w_i x_i + b) \tag{3.14}$$

Burada; x_i, giriş değerlerini, w_i, giriş değerlerine karşılık gelen ağırlık katsayılarını, b, öteleme değerini ve f, aktivasyon fonksiyonunu (sigmoid, tanh, RELU vb.) ifade eder.

Bu nöronlar bir araya getirilerek katmanlar oluşturur. Bir yapay sinir ağı modelinde en az üç katman bulunur. Bunlar; giriş katmanı, bir veya birden fazla gizli katman ve çıkış katmanıdır. Gizli katmanların ve bu katmanlarda yer alan nöron sayısı ağın karmaşıklığını ve kapasitesini belirler.

Eğitim sırasında, ağa sunulan girdilere karşılık gelen gerçek çıktılarla, ağın tahmini arasındaki farkı azaltmak için ağırlıklar ve öteleme değerleri sürekli olarak güncellenir. Bu güncelleme, genellikle geri yayılım (backpropagation) adı verilen bir algoritma ile gerçekleştirilir [68]. Şekil 3.7, Yapay Sinir Ağları modelinin görselini sunmaktadır.



Şekil 3.7. Yapay sinir ağları modeli

3.5.7. Evrişimli sinir ağları

Evrişimli Sinir Ağları (CNN), özellikle görüntü işleme görevleri için tasarlanmış derin öğrenme modelleridir. CNN'lerin temel özelliği, evrişim (convolution) adı verilen bir işlemi kullanarak girdi verisinden özellik haritaları (feature maps) oluşturabilmesidir.

Evrişim katmanında küçük boyutlu bir kernel veya filtre matrisi, giriş verisi üzerinde kaydırılır ve her bir pozisyonda evrişim işlemi gerçekleştirilir. Bu, girişteki yerel bilgilerin yakalanmasına yardımcı olur. Örneğin, bir görüntü üzerinde yapılan evrişim, kenarlar veya doku gibi özelliklerin saptanmasına yardımcı olabilir. Evrişim sonrasında elde edilen değerler bir aktivasyon fonksiyonundan geçirilir ve sınıflandırma yapılır. Bir diğer CNN katmanı da havuzlama katmanıdır. Bu katman, özellik haritasının boyutunu azaltırken önemli özelliklerin korunmasına yardımcı olur. En yaygın kullanılan yöntemlerden biri maksimum havuzlama (max pooling) yöntemidir. Örneğin, 2x2'lik bir pencere kullanarak her 2x2'lik alandaki maksimum değeri alarak yeni bir matris oluşturulur. Tam Bağlantı Katmanında (Fully Connected Layer) ise evrişim ve havuzlama katmanlarından elde edilen özellikler alınarak nihai sınıflandırma veya regresyon sonucunu üretir [80,81].

CNN'ler, geleneksel yapay sinir ağlarına göre görüntü tanıma görevlerinde daha üstün performans gösterirler, çünkü evrişim işlemi sayesinde giriş verisinin hiyerarşik ve yerel

özelliklerini öğrenebilirler. Bu da CNN'nin, özellikle büyük ve karmaşık görüntü veri setlerinde, yüksek doğrulukla sınıflandırma yapabilmelerini sağlar.

3.5.8. Yinelemeli sinir ağları

Yinelemeli sinir ağlarının teorik altyapısı, sıralı verilerdeki bağlamı anlama yeteneğine dayanır. Standart sinir ağları bağımsız ve aynı dağılıma sahip varsayılan veri noktaları ile çalışırken, Yinelemeli sinir ağları bu varsayımı kırmakta ve her adımda önceki adımların bilgisini kullanarak daha zengin temsiller oluşturabilmektedir.

Yinelemeli Sinir Ağları (RNN), sıralı verilerle çalışmak için tasarlanmış bir yapay sinir ağı türüdür. Temel fikir, aynı ağın tekrar tekrar kullanılması ve her adımda önceki adımın hafıza bilgisiyle birlikte yeni giriş verisinin işlenmesidir.

Yinelemeli sinir ağları, sıralı veriye sahip problemlerde, özellikle dil modelleme, metin oluşturma ve makine çevirisi gibi görevlerde etkili olabilir. Ancak, uzun bağımlılıkları öğrenmede zorluklar yaşarlar. Bu sorun, LSTM (Long Short-Term Memory) ve GRU (Gated Recurrent Unit) gibi RNN varyantlarıyla hafifletilmiştir [82].

3.5.9. LSTM

LSTM (Uzun Kısa Vadeli Bellek) ağları, Sepp Hochreiter ve Jürgen Schmidhuber tarafından 1997 yılında önerilmiş ve yinelemeli sinir ağlarının (RNN) uzun vadeli bağımlılıklarını öğrenmedeki zorluklarını aşmak için tasarlanmıştır. LSTM, standart RNN'lerin "vanishing" ve "exploding gradient" problemlerini hafifletmek için özel olarak tasarlanmış bir RNN türüdür [83,84].

3.6. Topluluk Tespiti Algoritmaları

Tez çalışmamızda temel varsayımlardan biri de siber zorbalık faaliyetlerinin topluluklar halinde hatta koordine olarak yapılabildiğidir. Bu durumun ve ilgili toplulukların tespiti ise bu çalışmanın amaçları arasındadır. Topluluklara ilişkin incelemelerin ve analizlerin yapılabilmesi için öncelikle ağ yapısının anlaşılması gerekmektedir. Aşağıda örnek bir G ağı referans alınarak tanımlara ve ağ yapısının açıklamasına yer verilmiştir.

Bir G = (V, E) grafında, V: node (düğüm), E: kenarları temsil etmektedir. g ⊆ G kuralına uyan g grafları da orijinal ağın bir parçasıdır [85]. Bir ağda yer alan topluluklar, birbirleri ile yoğun etkileşimde bulunan kullanıcıların yer aldığı g alt grafı ile ifade edilebilir ve topluluk tespiti analizleri ile bu alt graflar bulunabilir.

Topluluk tespiti analizinde kullanılan en yaygın algoritmalar arasında Louvain Algoritması bulunmaktadır [86, 87]. Louvain Algoritması, ağı optimize ederek toplulukları tanımlamayı amaçlar. Ağdaki düğümleri başlangıçta farklı topluluklara yerleştirir ve sonrasında düğümleri yeniden düzenleyerek modülerliği (modularity) maksimize eder. Sonuç olarak, daha iyi tanımlanmış topluluklar elde edilir.

3.6.1. Louvain algoritması

Louvain algoritması, büyük ağlarda topluluk tespiti için kullanılan bir algoritmadır. Bu algoritma, ağdaki her bir düğümü, en yakın komşularına göre sınıflandırır. Ardından, algoritma, bu sınıfları birleştirerek daha büyük topluluklar oluşturur.

Louvain yöntemi, modülarite optimizasyonu temelli hızlı ve kaliteli topluluk tespiti sağlar. Bu yöntem, hiyerarşik bir yaklaşım benimser. Lokal topluluk optimizasyonu ve ağın topluluklara göre birleştirilmesi olmak üzere iki adımı vardır.

Lokal topluluk optimizasyonu adımında, her düğüm kendi başına bir topluluktur. Algoritma, her bir düğümü komşu topluluklara taşıyarak modülaritenin nasıl değiştiğini değerlendirir. Modülarite artışı en yüksek olan topluluğa düğüm taşınır. Bu adım, modülarite artışı olmadığı sürece tekrarlanır.

Ağın birleştirilmesi adımında ise ilk adımda bulunan topluluklar temel alınarak yeni bir ağ oluşturulur. Bu yeni ağda, her topluluk tek bir düğüm olarak temsil edilir ve aralarındaki kenar ağırlıkları, orijinal ağdaki topluluklar arasındaki toplam bağlantı ağırlığına eşittir.
Bu iki adım, modülaritenin artmadığı bir noktaya kadar tekrar tekrar uygulanır [38]. Louvain algoritmasına ilişkin sözde kod Şekil 3.8'de sunulmuştur [88,89].

Louvain Algoritması Sözde Kodu:

1. G başlangıç ağı için
2. tekrarla
3. G'nin her düğümünü kendi topluluğuna yerleştir
4. bazı düğümler yer değiştirdikçe devam et
5. G'nin tüm düğümleri için
6. n düğümünü, modülerlik kazancını en üst düzeye çıkaran, kendi topluluğu dahil komşu topluluğuna yerleştir
7. Adım 5'i sonlandır
8. Adım 4'ü sonlandır
9. eğer yeni modülerlik, başlangıçtaki modülerlikten yüksekse
10. G = G'nin toplulukları arasındaki ağ
11. aksi takdirde
12. bitir
13. Adım 9'u sonlandır
14. Ağda herhangi bir değişim olmayana kadar devam et

Şekil 3.8. Louvain algoritması sözde kodu

3.7. Değerlendirme Metrikleri

Tasarlanan modellerin daha verimli sonuçlar almak için yeniden tasarlanması veya elde edilen başarım oranlarının görselleştirilmesi için performans ölçüm metrikleri kullanılmaktadır. Tasarlanan modellerin uygulama aşamasındaki başarımını ölçmek için en yaygın kullanılan performans ölçüm metrikleri karmaşıklık matrisi, F-1 skor ve AUC-ROC eğrisidir.

3.7.1. Karmaşıklık matrisi

Karmaşıklık ya da karışıklık matrisi (*confusion matrix*), temel olarak gerçek sınıf etiketleri ile modelin tahmin ettiği etiketler arasındaki ilişkiyi gösteren bir performans ölçüm aracıdır. Karmaşıklık matrisi, genellikle ikili sınıflandırma için kullanılsa da çok sınıflı sınıflandırma için de genişletilebilir [90]. Çizelge 3.3 ikili sınıflandırma için bir karmaşıklık matrisi örneği sunmaktadır.

Çizelge 3.3. Karmaşıklık matrisi

Tahmin Edilen / Gerçek Sınıf	Pozitif	Negatif
Pozitif	TP	FN
Negatif	FP	TN

Karmaşıklık matrisinde yer alan TP (True Positive), algoritmanın pozitif örneği doğru olarak tahmin ettiği durumların, TN (True Negative), algoritmanın negatif örneği doğru olarak tahmin ettiği durumların, FP (False Positive), algoritmanın negatif örneği pozitif olarak tahmin ettiği durumların ve FN (False Negative), algoritmanın pozitif örneği negatif olarak tahmin ettiği durumların sayısını ifade eder. Karmaşıklık matrisi, doğruluk, hassasiyet, kesinlik, F1 skoru vb. birçok sınıflandırma metriğinin hesaplanmasında temel oluşturur.

Doğruluk (*accuracy*), bir sınıflandırma modelinin tüm örnekleri doğru bir şekilde sınıflandırma olasılığını gösteren bir ölçüdür ve Eş. 3.15'deki gibi hesaplanır.
Hassasiyet, bir sınıflandırma modelinin gerçek olarak pozitif olan tüm örnekleri doğru bir şekilde sınıflandırma olasılığını gösteren bir ölçüdür. Karmaşıklık matrisinde hassasiyet, Eş. 3.16'da gösterildiği gibi hesaplanır.

Kesinlik, bir sınıflandırma modelinin gerçek olarak pozitif olarak sınıflandırdığı örneklerden kaç tanesinin doğru bir şekilde sınıflandırdığını gösteren bir ölçüdür ve Eş. 3.17'de olduğu gibi hesaplanır.

F1-Skoru, sınıflandırma problemlerinde modelin performansını ölçmek için kullanılan yaygın bir metriktir. Özellikle dengesiz veri kümeleri için oldukça faydalıdır. F1-skoru, modelin kesinlik (precision) ve hassasiyet (recall) değerlerinin harmonik ortalaması olarak hesaplanır. F1-skor hesaplaması Eş. 3.18'de sunulmuştur.

$$\text{Doğruluk} = \frac{TP+TN}{\text{TP} +\text{TN}+\text{FP}+\text{FN}} \tag{3.15}$$

$$\text{Hassasiyet} = \frac{TP}{\text{TP} +\text{FN}} \tag{3.16}$$

$$\text{Kesinlik} = \frac{TP}{\text{TP}+\text{FP}} \tag{3.17}$$

$$\text{F1-Skor} = \frac{2 * \text{Kesinlik} * \text{Hassasiyet}}{\text{Kesinlik} + \text{Hassasiyet}} \quad (3.18)$$

3.7.2. AUC-ROC eğrisi

AUC-ROC eğrisi, sınıflandırma modellerinin performansını değerlendirmek için kullanılan bir grafiktir. Eğri, modelin yanlış pozitif oranı (FPR) ve gerçek pozitif oranı (TPR) arasındaki ilişkiyi gösterir. Burada; FPR, gerçek olarak negatif olan bir örneğin yanlış olarak pozitif olarak sınıflandırılma olasılığını gösterirken, TPR, gerçek olarak pozitif olan bir örneğin doğru olarak pozitif olarak sınıflandırılma olasılığını gösterir. AUC-ROC eğrisi, 0 ve 1 arasında bir değere sahiptir. Bire yakın bir AUC değeri, modelin iyi performans gösterdiğini gösterir [91].

AUC-ROC eğrisi, özellikle dengesiz veri setlerinde modellerin performansını değerlendirmek için faydalıdır. Dengesiz veri setlerinde, bir sınıfın diğerinden çok daha fazla örneği olabilir. Bu durumda, doğruluk gibi ölçütler yanıltıcı olabilir.

3.8. Konu Tespiti Algoritmaları

Konu tespiti, büyük metin koleksiyonlarındaki gizli tematik yapıları keşfetmek için kullanılan bir makine öğrenimi tekniğidir. Metin madenciliği ve doğal dil işleme alanlarında önemli bir yere sahiptir. Bu teknik, belgelerin içerdiği konuları otomatik olarak sınıflandırmak ve böylece büyük veri setlerindeki kalıpları ve trendleri anlamak için kullanılır. Özellikle, dijital çağda, sosyal medya gönderileri, haber makaleleri, bilimsel makaleler ve müşteri geri bildirimleri gibi metinlerin hızla artmasıyla, konu tespiti araçları, verilerden anlamlı bilgiler çıkarmak için vazgeçilmez hale gelmiştir [92, 93]. Bu teknikler, veri analistlerine ve araştırmacılara, büyük metin yığınları içerisinden önemli temaları ve eğilimleri belirleme, müşteri yorumlarını analiz etme, haber akışlarından bilgi çıkarma gibi birçok alanda yardımcı olur. Bu çalışmada da tespit edilen topluluklarda bulunan kullanıcıların paylaşımlarını incelemek için LDA algoritmasından faydalanılmıştır.

3.8.1. LDA (Latent Dirichlet Allocation) algoritması

Latent Dirichlet Allocation (LDA) algoritması, belgelerin üzerinde yatan gizli konu yapılarını keşfetmek için kullanılan bir tür olasılıksal modeldir. LDA'nın temel varsayımı, her belgenin birden fazla konudan oluştuğu ve her konunun da belirli kelimelerle ilişkili olduğudur.

Örneğin; elimizde bulunan bir metin koleksiyonunda spor, ekonomi ve politika ile ilgili makaleler bulunuyor olsun. LDA modelini bu koleksiyona uygulamak için öncelikle modelin kaç konu bulması gerektiğine karar vermek gerekir.

Model başlangıçta her kelimeyi rastgele bir konuya atar. Daha sonra, her kelimenin belirli bir konuya ait olma olasılığını, o kelimenin o belgedeki sıklığı ve o kelimenin diğer belgelerde o konuyla ilişkilendirilme sıklığına dayanarak günceller. Örneğin, "gol" kelimesi sporla ilgili bir makalede sıkça geçiyorsa ve "gol" kelimesi genellikle spor konusuyla ilişkilendiriliyorsa, "gol" kelimesinin spor konusuna atama olasılığı artar.

Bu döngüsel süreç, modelin belgelerdeki konu dağılımlarını ve bu konuların kelime dağılımlarını öğrenene kadar devam eder. Sonuçta, her belge için bir konu dağılımı (bir makalenin %70 spor, %20 ekonomi, %10 politika içerdiği) ve her konu için bir kelime dağılımı (spor konusunun %30 "gol", %20 "maç", %10 "takım" gibi kelimeleri içerdiği) elde edilir. Bu dağılımlar, belgelerin ve kelimelerin altında yatan tematik yapıları yansıtır.

4. DENEYSEL ÇALIŞMALAR

Bu bölümde tez kapsamında gerçekleştirilen deneysel çalışmalar ve sonuçları hakkında bilgiler sunulmuştur. Deneysel çalışmaların ilk aşamasında siber zorbalık sınıflandırması ile ilişkili çalışmalara, ikinci aşamada ise ağ yapısının oluşturulması ve sosyal ağ analizi çalışmalarına yer verilmiştir.

4.1. Siber Zorbalık Tespiti ile İlgili Çalışmalar

Bu aşamada, modeli eğitmek için 3.1.2 başlığı altında belirtilen toplam eğitim veri seti kullanılmıştır. Daha sonra X üzerinden kullanıcılara ilişkin toplanan verilere ön işleme adımları uygulanmış ve sekiz farklı makine ve derin öğrenme algoritması farklı özellik çıkarım ve seçim teknikleri ile test edilmiştir. Çalışma sonuçları Çizelge 4.1'de sunulmuştur.

Çalışma sonuçları incelendiğinde en başarılı modelin 0,90 doğruluk oranı ile ANN sınıflandırıcısının CountVectorizer ve Ki-Kare teknikleri ile birlikte kullanılması ile oluşturulan model olduğu görülmektedir. Modellerin karşılaştırılması yapılırken doğruluk ve F1-skor metrikleri kullanılmıştır.

Makine ve derin öğrenme modellerinin test edilmesi aşamasında uygulanan adımlar sırasıyla aşağıdaki gibidir.

- Veri ön işleme adımları uygulanmıştır.
  - Sayısal ifadelerin kaldırımlası
  - Url, Hashtag, Mention ve RT (Retweet) ifadelerinin kaldırılması
  - Noktalama işaretleri, semboller ve sayısal ifadelerin kaldırılması
  - Metinlerin küçük harfe çevrilmesi
  - Etkisiz kelimelerin (Stopwords) metinlerden çıkarılması
  - Lemmatization uygulanması
  - İki veya daha az uzunlukta olan kelimelerin metinlerden kaldırılması
- Eğitim verisi, %80 eğitim ve %20 test olacak şekilde bölünmüştür.
- CountVectorizer veya TFIDFVectorizer kullanılarak metin verisi vektörize edilmiştir.
- Performansı artırmak ve gereksiz özelliklerden kurtulmak için Ki-Kare veya bilgi kazanımı teknikleri kullanılarak en iyi 5,000 özelliğin seçilmesi sağlanmıştır.
- Seçilen bu özellikler kullanılarak model eğitilmiştir.
- Eğitilen model, test verisi üzerinde çalıştırılarak tahminler elde edilmiştir.
- Sonuçlara ilişkin doğruluk ve f1-skor değerlendirme ölçütleri hesaplanmıştır.

Çizelge 4.1. Makine ve derin öğrenme algoritmalarının sonuçları

Algoritma	Özellik Çıkarımı	Özellik Seçimi	Doğruluk	F1-Score
Lojistik Regresyon	Count Vectorizer	Ki-Kare	0,851	0,829
		Bilgi Kazanımı	0,853	0,829
	TFIDF Vectorizer	Ki-Kare	0,841	0,814
		Bilgi Kazanımı	0,843	0,815
Lineer Regresyon	Count Vectorizer	Ki-Kare	0,868	0,851
		Bilgi Kazanımı	0,659	0,443
	TFIDF Vectorizer	Ki-Kare	0,851	0,822
		Bilgi Kazanımı	0,839	0,812
SGD Sınıflandırıcı	Count Vectorizer	Ki-Kare	0,861	0,838
		Bilgi Kazanımı	0,881	0,859
	TFIDF Vectorizer	Ki-Kare	0,864	0,825
		Bilgi Kazanımı	0,851	0,832
Karar Ağaçları	Count Vectorizer	Ki-Kare	0,827	0,811
		Bilgi Kazanımı	0,812	0,792
	TFIDF Vectorizer	Ki-Kare	0,813	0,805
		Bilgi Kazanımı	0,807	0,797
Multinomial Naive Bayes	Count Vectorizer	Ki-Kare	0,852	0,335
		Bilgi Kazanımı	0,829	0,801
	TFIDF Vectorizer	Ki-Kare	0,843	0,822
		Bilgi Kazanımı	0,826	0,801
Yapay Sinir Ağları	Count Vectorizer	Ki-Kare	0,901	0,887
		Bilgi Kazanımı	0,867	0,831
	TFIDF Vectorizer	Ki-Kare	0,891	0,881
		Bilgi Kazanımı	0,814	0,771
Evrişimli Sinir Ağları	Count Vectorizer	Ki-Kare	0,877	0,867
		Bilgi Kazanımı	0,834	0,819
	TFIDF Vectorizer	Ki-Kare	0,541	0,709
		Bilgi Kazanımı	0,537	0,524
Yinelemeli Sinir Ağları	Count Vectorizer	Ki-Kare	0,835	0,821
		Bilgi Kazanımı	0,829	0,791
	TFIDF Vectorizer	Ki-Kare	0,857	0,804
		Bilgi Kazanımı	0,817	0,819

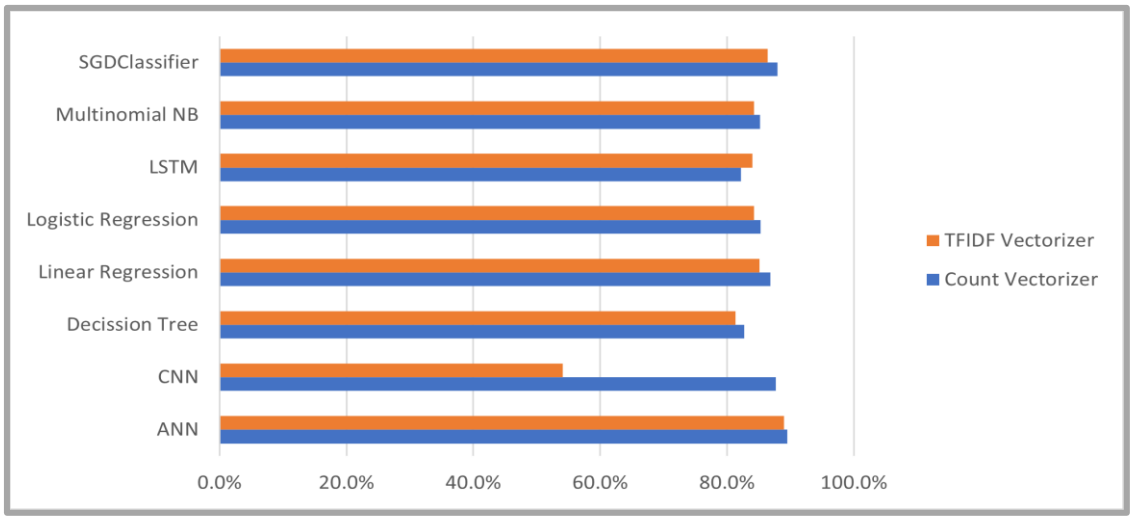
Toplanan verilerin tamamı, bu sonuçlara göre en başarılı olan model kullanılarak sınıflandırılmış ve sınıflandırma sonuçları sosyal ağ analizinde kullanılmıştır. Sınıflandırma sonucunda elde edilen verilerin bir örneği Çizelge 4.2'de sunulmuştur. Burada etiket değeri "1" olan metinler siber zorbalık içeren, "0" olan metinler ise siber zorbalık içermeyen olarak sınıflandırılmıştır.

Çizelge 4.2. Sınıflandırılmış veri seti

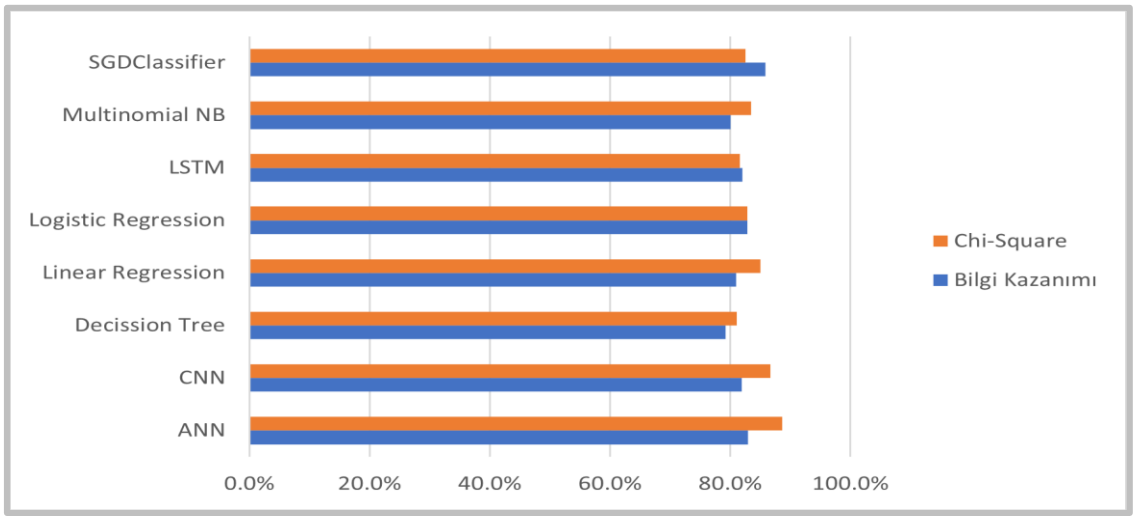
Metin	Etiket
birinden duymuştum sen dedikten sonra vay ahmet dedim	0
elimden gelir örgü işi ben katkıda bulunmak isterim	0
yönetiminden taraftarlarına tam bir terör estirdiniz utancı yaşamak yerine hala ucuz kahramanlığa soyunuyorsunuz terbiyesiz herifler	1
buyrun efendim üstelik geçen aydan kullanım var lanet olsun yaa allah belalarını versin ülkeyi bu hale getirenlerin	1
gerizekalı mal herifler bir halt bildiğiniz yok	1
arkadaşlar yardım deştiğinde bulunduğunuzda küçük tüp tencere ekleyin kuru gıda sebze pişirilemiyor	0
İyiler mutlaka kazanır derler halbuki yalan iyler hiçbir zaman kazanamaz aksine acı çeker saflığının bedelini öder	0
lanet adamın ağzına laf verdiniz helal olsun size	1
adamlar ekşi sözlük kapatmışlar ben artık dayanma sınırımı aştım bunların yobaz karaktersiz şerefsiz zihniyetine karşı geçen gün öfke doluyorum kin besler duruma geldim	1
aglamiyorum gozume sevgi kacti	0
maçın atmosferinde bunun olabileceğini söylerken tribüne karşı bir yaptığından bahsetti öyleyse herkes karşısında yaptığı konuşmayı kendisine hatırlatmak gerek	0
istifa istemi bizim hukuki hakkımızdır buna laf etmek kimseye düşmez	0
size diyoz lan yapram	1
senin sene takıldığın ligte gelip sana tek atan takıma bunu söylüyorsun köylü	1

Özel ve Saraç, siber zorbalığın tespitinde bazı özellik çıkarma ve seçme tekniklerinin etkisini değerlendirmiş ve kök alma işleminin sınıflandırıcının başarısını azalttığı gözlemlemiştir [56]. Tez çalışması kapsamında da sınıflandırma aşamasında kök alma (steeming) tekniğinin ön işleme adımlarında kullanımı test edilmiş ancak [56] ile benzer şekilde sınıflandırma başarısını etkilemediği görülmüş ve önerilen modele dahil edilmemiştir. Veri setini yapısal olarak değiştiren ön işleme adımlarının kullanılan veri setine ve dil özelliklerine göre değişkenlik göstereceği açıktır.

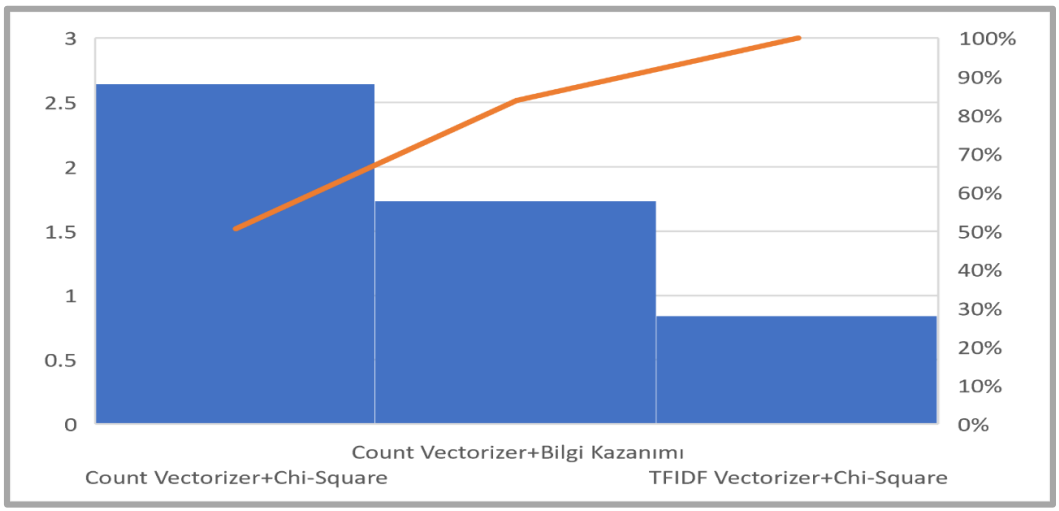
Ayrıca, Şekil 4.1 ve Şekil 4.2 kullanılan özellik çıkarım ve özellik seçim tekniklerinin birbirinden bağımsız olarak başarım oranlarına ilişkin istatistikleri sunmaktadır. Şekil 4.3 ise bu tekniklerin birlikte kullanımlarına ilişkin istatistiği sunulmaktadır. Sunulan istatistikler incelendiğinde de en başarılı sonuçların ANN algoritmasının CountVectorizer ve Ki-Kare teknikleri ile birlikte kullanımının olduğu görülmektedir.



Şekil 4.1. Özellik çıkarım tekniklerine göre elde edilen en başarılı sonuçlar



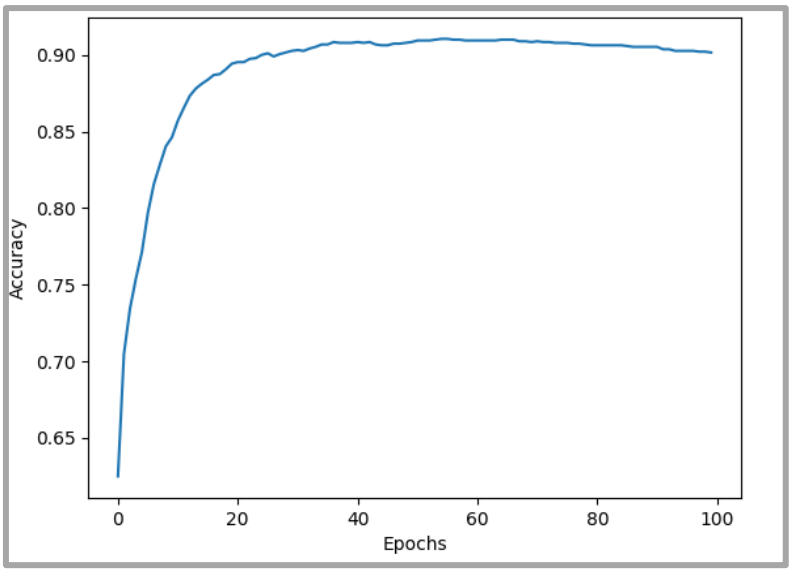
Şekil 4.2. Özellik seçim tekniklerine göre elde edilen en başarılı sonuçlar



Şekil 4.3. Özellik çıkarım ve seçim tekniklerinin birlikte kullanım istatistiği

Uygulanan modeller arasında en yüksek başarıma sahip olan ANN modelinde;

- *Eğitim veri seti %80, test veri seti %20,*
- *Özellik çıkarımı için CountVectorizer,*
- *Özellik seçimi için Ki-Kare tekniği ile ilk 5 000 özellik,*
- *Gizli katmandaki nöron sayısı 100,*
- *Öğrenme oranı 0,0001,*
- *Optimizasyon algoritması olarak 'adam',*
- *İterasyon sayısı 10 ve epok sayısı 100 olarak belirlenmiştir. Şekil 4.4, modelin başarım oranına ilişkin epok odaklı başarı bilgilerini sunmaktadır.*



Şekil 4.4. ANN modelinin başarım grafiği

4.2. Topluluk Tespiti ile İlgili Çalışmalar

Graf (çizge), nesnelerin bağlantıları veya ilişkileri ile temsil edildiği bir yapıdır. Bir graf, düğüm ve kenar olmak üzere iki temel bileşenden oluşur. Düğümler, bir ağın veya ilişki yapısının temel öğelerini temsil eder. Kenarlar ise düğümler arasındaki ilişkileri ifade eder. [94-96]. Bu ilişkilerin incelenmesi ve analiz edilmesi ise sosyal ağ analizinin konusudur. Sosyal ağ analizi bize sosyal yapıların ve grupların algılanmasını sağlar [97]. Bu bölümde siber zorbalık skoru ve graf yapısının oluşturulmasına ve topluluk tespiti için yapılan sosyal ağ analizine yer verilmiştir.

4.2.1. Siber zorbalık skorunun oluşturulması

Sosyal ağlardaki etkileşimlerin ve ilişkilerin analizi, siber zorbalığın izlenmesi ve topluluk tespiti için önemli bir adımdır. Bu çalışmada, X (Twitter) ortamından çekilen 267.384 kullanıcıya ait bilgiler, kullanıcılar arasındaki etkileşimler ve siber zorbalık sınıflandırmasında tahmin edilen veriler kullanılarak bir sosyal ağ modeli oluşturulmuştur. Bu model, kullanıcıların birbirleriyle olan bağlantılarını ve etkileşimlerini temsil eden bir graf yapısıdır. Her kullanıcı ağın bir düğümünü temsil ederken, kullanıcılara ait 'Takip', 'Takipçi' bilgileri de kenarlarla ifade edilir.

X'den (Twitter) elde edilen gönderilerin tamamı, eğitim verisi üzerinde test edilen sekiz farklı makine ve derin öğrenme algoritmalarından en başarılı sonuçları veren Yapay Sinir Ağları algoritması kullanılarak zorbalık içerip içermediğine göre sınıflandırılmıştır. Her kullanıcının siber zorbalık eğilimini belirlemek için sınıflandırma sonrasında elde edilen veri kullanılarak Eş. 4.1'e göre kullanıcılara ilişkin siber zorbalık skoru hesaplanmıştır.

$$\mu = f(x) = \begin{cases} \sum_{i=1}^{n} \frac{S_i}{T_i} * 100, & T_i \geq \alpha \\ hesaplamaya\ dahil\ etme, & T_i < \alpha \end{cases} \quad (4.1)$$

Burada; n: toplam kullanıcı sayısını, S_i: i. kullanıcıya ait zorbalık olarak etiketlenmiş paylaşım sayısını ve T_i: i. kullanıcıya ait toplam paylaşım sayısını göstermektedir. μ, her bir kullanıcının 0 ile 100 arasında olan siber zorbalık skorunu göstermektedir. Toplam paylaşım sayısı α'dan az olan kullanıcılar ağ analizi için kullanılacak olan sosyal ağa dahil edilmemiştir. α değeri yapılan testlerde 5 olarak alındığında çok nadir paylaşım yapan veya aynı paylaşımı sürekli tekrarlayan hesapların belirlendiği görülmüş ve tez çalışması kapsamında $\alpha = 5$ olarak alınmıştır. Siber zorbalık skoru en yüksek 10 kullanıcıya ait bilgiler Çizelge 4.3'de sunulmuştur.

Çizelge 4.3. Siber zorbalık skoru en yüksek kullanıcılar

Kullanıcı	Siber Zorbalık Skoru (μ)	Toplam Paylaşım Sayısı
User1	1	400
User2	1	398
User3	1	378
User4	1	200
User5	1	200
User6	1	200
User7	1	200
User8	1	199
User9	1	199
User10	1	197

Siber zorbalık skorunun ve paylaşım sayısında alt sınırlandırmanın kullanım amacı çok büyük ve karmaşık bir ağ yapısının oluşması yerine amaca yönelik ve sosyal ağ analizi algoritmalarının daha verimli çalışmasını sağlayacak daha etkin ağ yapılarının oluşmasını sağlamaktır.

Sınıflandırılan veriler incelendiğinde bazı kullanıcıların sadece bir paylaşımı olduğu ve bu paylaşımın da zorbalık olarak sınıflandırıldığı görülmüştür. Bu durumda bu kullanıcının zorbalık skoru 1 olsa da sadece bir paylaşımdan kullanıcı hakkında çıkarımlarda bulunmanın yetersiz olacağı açıktır. Bundan dolayı siber zorbalık skoru oluşturulurken toplam paylaşım sayına bir alt sınırlandırma uygulanarak bu durumun engellenmesi amaçlanmıştır.

4.2.2. Ağ yapısının oluşturulması

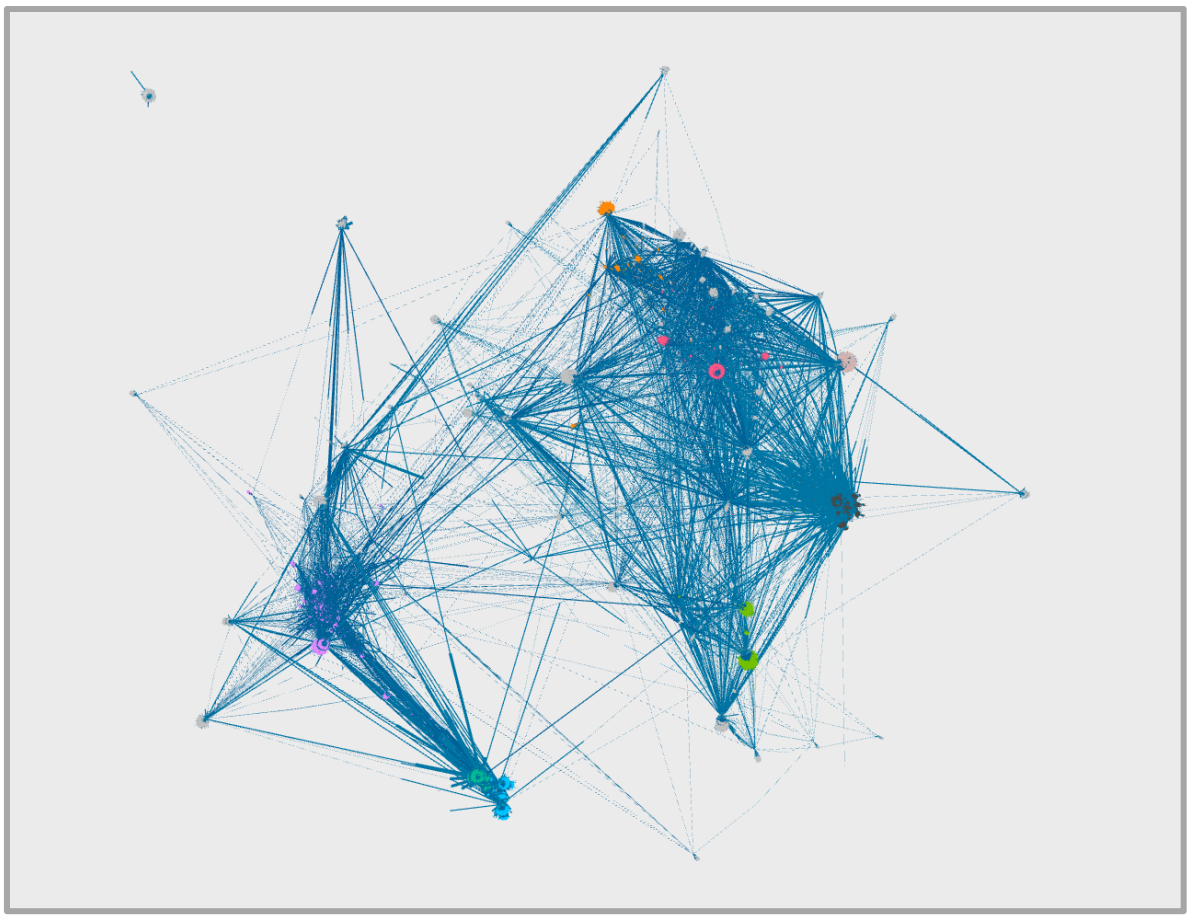
Sosyal ağ analizi için graf oluşturma aşamasında daha küçük ve amaca yönelik bir ağ oluşturmak için siber zorbalık skoruna göre grafta yer alacak kullanıcılar dört farklı şekilde seçilmiştir. İlk grafta $\mu \geq 0$, ikinci grafta $\mu \geq 0{,}25$, üçüncü grafta $\mu \geq 0{,}5$ ve dördüncü grafta $\mu \geq 0{,}75$ olan kullanıcı bilgileri ile bu kullanıcılara ait 'Takip-Takipçi' bilgileri kullanılarak dört farklı ağ oluşturulmuştur. Bu ağlarda yer kullanıcı dağılımları Çizelge 4.4'te verilmiştir.

Ağda yer alacak düğümler, siber zorbalık skorlarına göre belirtilen koşullara uyan kullanıcılardan oluşmaktadır. Kenarlar ise bu kullanıcılara ilişkin Twitter API kullanılarak elde edilen takip-takipçi bilgilerinden oluşmaktadır.

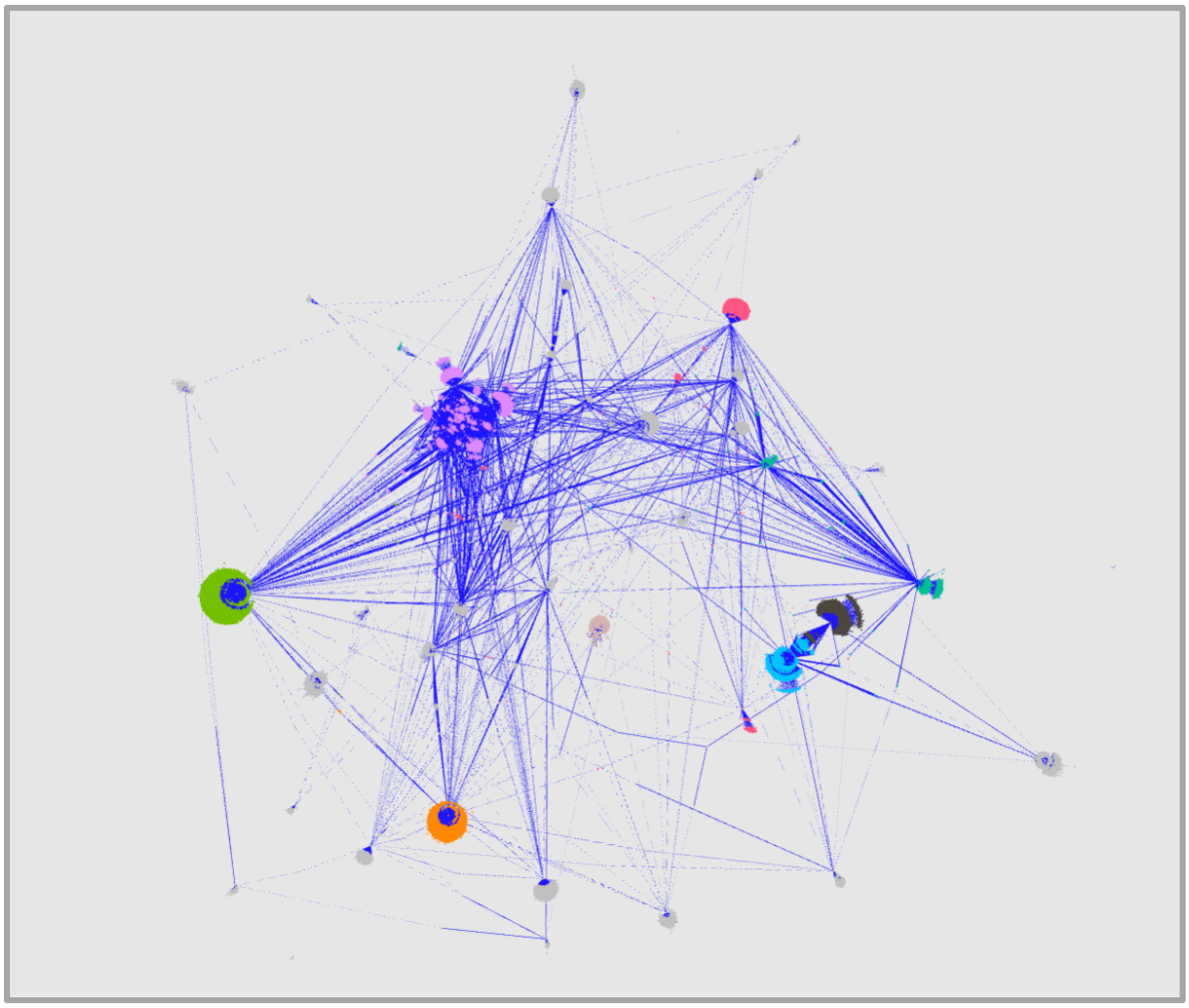
Çizelge 4.4. Oluşturulan ağlara ilişkin istatistikler

Ağ Adı	Koşul	Düğüm Sayısı	Kenar Sayısı
Graf1	$\mu \geq 0$	27 184 189	297 430 773
Graf2	$\mu \geq 0{,}25$	925 692	2 522 421
Graf3	$\mu \geq 0{,}5$	98 240	139 004
Graf4	$\mu \geq 0{,}75$	44 247	62 288

Graf1 ve Graf2 çok sayıda düğüm ve ilişkiden oluştuğundan ağların tamamının görselleştirilmesi çok anlamlı olmayacaktır. Ancak, Graf3 ve Graf4 için ağların genel görünümleri Şekil 4.5 ve 4.6'te sunulmuştur.



Şekil 4.5. Graf3 ağına ait genel görünüm



Şekil 4.6. Graf4 ağına ait genel görünüm

4.2.3. Topluluk tespiti

Topluluk tespiti, oluşturduğumuz sosyal ağların yapısını ve dinamiklerini anlamamıza yardımcı olabilir. Bu ağlar üzerinde yapılacak analizler, sosyal ağların nasıl çalıştığını ve kullanıcıların nasıl etkileşime girdiğini daha iyi anlamamıza olanak tanır. Bu çalışmada da topluluk tespiti ile hedeflenen gruplar halinde yapılan zorbalık faaliyetlerinin tespitidir. Sosyal ağ analizi için oluşturulan ağlar üzerinde en yaygın topluluk tespiti algoritmalarında biri olan modülarite tabanlı Louvain algoritması kullanılmıştır [98, 99].

Modülarite sosyal, biyolojik, ekonomik ve teknoloji gibi ağ yapılarında topluluk kalitelerini ölçmek için kullanılan bir metriktir. Modülarite, ağın topluluklara bölünmüş olmasının beklenen rastgele bağlantıya göre ne kadar iyi olduğunu ölçer. Yüksek bir modülarite değeri, ağın topluluklara iyi bir şekilde bölündüğünü gösterirken, düşük bir değer ağın rastgele

bağlantılardan çok farklı olmadığını gösterir [98, 99]. Louvain algoritması ile yapılan topluluk tespiti çalışması sonucunda elde edilen sonuçlar Çizelge 4.5'de sunulmuştur.

Çizelge 4.5. Topluluk tespiti sonuçları

Ağ Adı	Yönlü/Yönsüz	Topluluk Sayısı	Modülarite
Graf1	Yönlü	15 487 043	0,445
Graf1	Yönsüz	180	0,547
Graf2	Yönlü	308 473	0,55
Graf2	Yönsüz	88	0,703
Graf3	Yönlü	41 545	0,559
Graf3	Yönsüz	42	0,877
Graf4	Yönlü	22 817	0,457
Graf4	Yönsüz	32	0,846

Elde edilen sonuçlar, modülarite değerine göre kıyaslandığında en başarılı tespitlerin Graf3 üzerinde ilişki yönlerinin dikkate alınmadan yapılan çalışmadan elde edildiği görülmüştür. Bu çalışmada elde edilen en büyük 10 topluluk ve bu topluluklarda yer alan kullanıcılara ilişkin bilgiler Çizelge 4.6'de sunulmuştur.

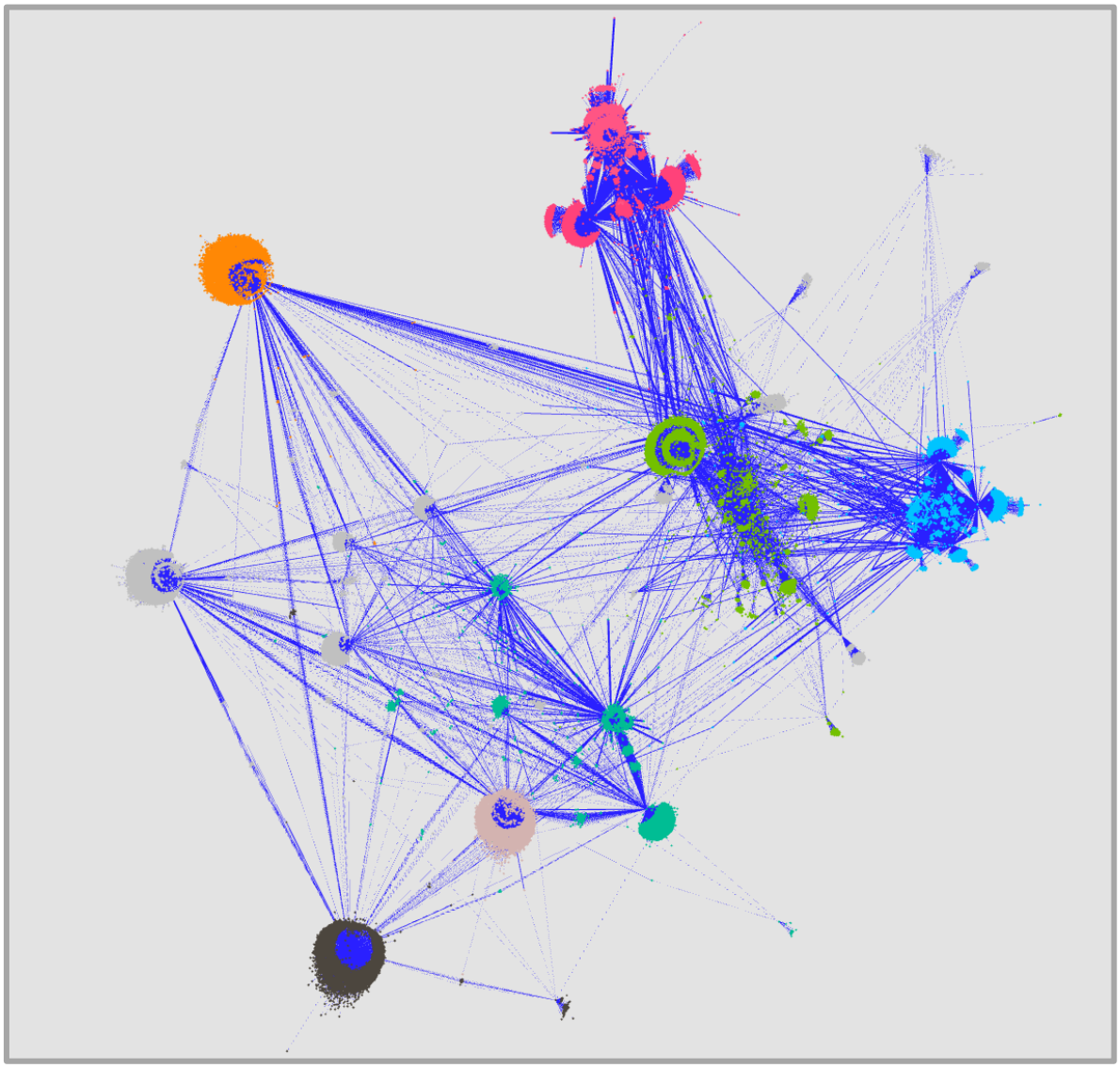
Çizelge 4.6. Graf3'te yer alan ilk 10 topluluk ve örnek üyeleri

Topluluk Numarası	Üye Sayısı	Örnek Üyeler
Grup1	8 500	36098475, 35774702146200371, 323376086537949
Grup2	8 231	73603933536739,65945371591860, 27451071328522
Grup3	7 314	8810641324, 503115602344427, 52215263072183
Grup4	5 575	46314977997393,156502959288725, 073853804230176
Grup5	5 348	110422731925581209611, 41308868, 3925949088175268001
Grup6	4 024	2466265604846796, 599029463928672, 406878307
Grup7	3 913	557404803675873,1486417033940160500, 19841807969738800
Grup8	3 909	1065696006686298100,94728997283786, 847909737899466
Grup9	3 902	18271379571575,442351679031021, 406014184777781
Grup10	3 443	7854899, 26841952, 65191844589076

4.2.4. Topluluk tespiti sonuçları

Topluluk tespiti sonuçlarında modülarite göz önüne alındığında en başarılı sonuçlar Graf3 üzerinde yapılan çalışmalarda elde edilmiştir. Bu ağ üzerinde tespit edilen toplulukların incelenmesi aşağıdaki gibidir.

Ağ üzerinde yapılan topluluk tespitinde elde edilen ilk 10 topluluğun genel görünümü Şekil 4.7'de sunulmuştur. Burada, toplulukların belirgin bir şekilde ayrıldığı görülmektedir.



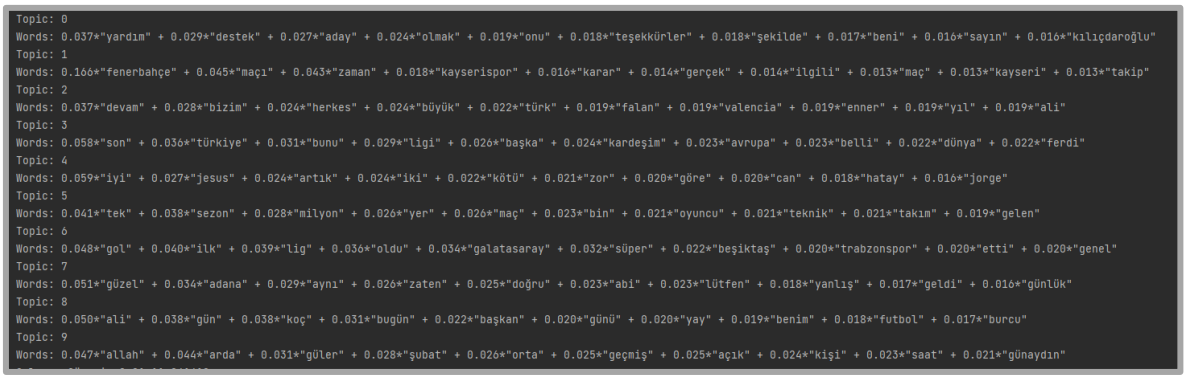
Şekil 4.7. Graf3 ağında yer alan ilk 10 topluluğun genel görünümü

Veri setinde yer alan, ağdaki kullanıcılara ait paylaşımlar incelendiğinde aynı grupta olan kullanıcıların paylaşım konularının genellikle benzer olduğu görülmüştür. Çizelge 4.7 aynı grupta yer alan farklı kullanıcıların zorbalık olarak sınıflandırılan paylaşımlarına örnekler sunmaktadır.

Çizelge 4.7. Topluluklardaki kullanıcılara ait örnek paylaşımlar

Metin	Hashtag	Kullanıcı
gollll g*tüne girsin hüseyinnnnnnn kale direkleriyle birlikteeee	'FBvSVS'	UserID1
adam öldürmeye tam teşebbüsten yargılayıp ceza vermek lazım yapılan taraftarlık değil adam bildiğin katil	'GöztepeAltay'	UserID2
bunlarda pişkinlik varkenutanırlar	'UtanacakYüzVarMI'	UserID3
kadıköy fenerbahçe taraftarlardan iktidara tepki yalan yalan yalan dolan dolan dolan sene oldu istifa ulan	'FBvKON'	UserID4
dünya yalan dolan senin sevgindir aslolan sonuna kadar fenerbahçeliyiz ulan	'FenerbahceliyizULAN', 'Fenerbahce'	UserID5
it köpek çar çakal sizin olsun bizim mız var alayınıza dert olsun	'FBvKAS'	UserID5
yavşak değil adam maç izlesen zaten ümraniyenin attığı gole yorum bile yapmazsın	'UEMRvFB'	UserID5
fener bugün yensin volkan demirel maçı sattı diyecek kemiksiz elli tane y*vşak spor yazarı sayarım	'FBvHAT'	UserID5
birazdan beş y*vşak çıkar fenerbahçeyi şampiyon yapacaklar falan der inşallah denk gelir varsa sövmezsem	'TSvFB'	UserID5
oldu paşinyan zillet miş boyle koyarlar adama	'kayserifenerbahce'	UserID6
yalan yalan yalan dolan dolan dolan	'KuralıUygulaFırsatçıTFF'	UserID6
o*ospu cocuklari	'ADSvsFB'	UserID6
yasa boyle cikarsa anayasa mahkemesinde bireysel basvuru hakkimi kullanacagim sizin adaletiniz muhalefet uyuyormu	'2000SonrasıSgklıAdaletİstiyor'	UserID6

Toplanan veriler üzerinde Latent Dirichlet Allocation (LDA) algoritması ile konu belirleme çalışmaları yapılmış ve elde edilen sonuçlar Şekil 4.8'de sunulmuştur. Konu belirleme için Çizelge 4.7'de örnek paylaşımların verildiği topluluk kullanıcılarının paylaşımları kullanılmıştır. LDA algoritması belgelerin konu dağılımlarını keşfetmek için kullanılan bir tür olasılıksal konu modelidir. LDA, her bir belgenin birden fazla konuyu içerebileceğini ve her konunun da belirli kelimelerle ilişkili olduğunu varsayar. Bu model, belgelerin ve kelimelerin olasılıksal dağılımlarını kullanarak, belirli bir belge koleksiyonunda gizli konuları ortaya çıkarmayı amaçlar.



Şekil 4.8. Örnek bir topluluk için LDA konu tespiti çıktısı

LDA algoritmasına, toplulukta bulunan kullanıcılara ait paylaşımları girdi olarak verilerek, 10 farklı konu ile bu konulardaki kelimelerin ağırlıkları belirlenmiştir. Tahmin edilen konular incelendiğinde;

- Konu 0: Siyasi ve sosyal içerikli konuşmaları temsil ediyor olabilir.
- Konu 1: "fenerbahçe", "maçı", "zaman", "kayserispor" gibi kelimelerle bu konu, Türk futbolu, özellikle Fenerbahçe'nin maçları ve belki de Kayserispor ile olan bir karşılaşmasını içeren tartışmaları yansıtıyor olabilir.
- Konu 2: "valencia" ve "enner" kelimeleri dikkate alındığında sporla ilgili olabilir.
- Konu 3: "son", "Türkiye", "ligi", "avrupa", "dünya" gibi kelimelerle bu konu, Türkiye'nin ulusal veya uluslararası spor müsabakalarındaki performansı gibi konuları içerebilir.
- Konu 4: "iyi", "jesus", "kötü", "zor" gibi kelimelerle bu konu, belki de futbol takımlarının performansları, teknik direktörlerin stratejileri veya maç sonuçları hakkında görüşleri içeriyor olabilir.
- Konu 5: "tek", "sezon", "milyon", "oyuncu", "teknik", "takım" gibi kelimelerle bu konu, futbol transferleri, sezonun genel değerlendirmesi veya teknik taktikler gibi konuları içerebilir.
- Konu 6: "gol", "ilk", "lig", "galatasaray", "süper", "beşiktaş", "trabzonspor" gibi kelimelerle bu konu, Türk Süper Lig'indeki takımların maçları, skorları ve genel lig durumu hakkında tartışmaları yansıtıyor olabilir.
- Konu 7: "güzel", "adana", "doğru", "abi", "lütfen" gibi kelimelerle bu konu, günlük sohbetler, yerel konular veya kişisel görüşler gibi daha geniş bir yelpazeyi kapsıyor olabilir.
- Konu 8: "ali", "koç", "bugün", "başkan", "futbol" gibi kelimelerle bu konu, Fenerbahçe Spor Kulübü'nün başkanı Ali Koç ve onunla ilgili güncel olaylar veya futbol dünyasındaki gelişmeleri içerebilir.
- Konu 9: "arda" ve "güler" ifadeleri dikkate alındığında sporla ilgili içerikler bulunduruyor olabilir.

Tahmin edilen konuların 0,8'i spor ile ilgili olduğu için bu toplulukta yer alan kullanıcıların ortak paydasının spor olabileceği sonucu elde edilmektedir.

Bu sonuçlar doğrultusunda, tespit edilen topluluklardan popülasyon yönünden en büyük 10 topluluğun konu dağılımları Çizelge 4.8'de sunulmuştur. Ayrıca, toplulukların konu tahminlerinin yapıldığı LDA çıktıları EK-1 olarak EKLER bölümünde sunulmuştur.

Çizelge 4.8. Toplulukların tahmini konu dağılımları

Topluluk	Konu Tahmini
Grup1	Politika ve Afet Müdahalesi
Grup2	Siyaset
Grup3	Cinsellik
Grup4	Cinsellik
Grup5	Siyaset ve Toplumsal Konular
Grup6	Cinsellik
Grup7	Cinsellik
Grup8	Cinsellik
Grup9	Siyaset ve Arkadaşlık
Grup10	Spor

5. SONUÇ VE ÖNERİLER

İnternetin ve sosyal medyanın yaygınlaşması, birçok avantajın yanı sıra bazı olumsuzlukları da beraberinde getirmiştir. Bu olumsuzluklardan biri de artan "Siber Zorbalık" problemidir. Siber zorbalık, bireylerin bilgisayar, tablet, akıllı telefon ve diğer elektronik cihazlar aracılığıyla diğer bireylere zarar vermek amacıyla gönderdikleri aşağılayıcı, tehditkâr veya kötü niyetli mesajlarla gerçekleştirdiği bir geleneksel zorbalık türüdür. Bu tür zorbalık, sosyal medya platformları, forumlar, e-postalar ve anlık mesajlaşma uygulamaları gibi çeşitli dijital ortamlarda gerçekleşebilir. Ancak, internet kullanımının yaygınlaşması ve sosyal medya ortamlarında kullanıcıların kimliklerini kolayca gizleyebilmeleri zorbaları da yoğunlukla bu platformları tercih etmeye sevk etmiştir.

Daha önceki benzer çalışmalarda çoğunlukla olduğu gibi bu çalışmada da günümüzde en yaygın sosyal medya platformlarından biri olan X (Twitter) temel veri kaynağı olarak belirlenmiştir. 267 384 kullanıcıya ait paylaşım, takip ve takipçi bilgileri Twitter API aracılığı ile toplanmıştır.

Bu çalışmada siber zorbalık sınıflandırması ve topluluk tespiti hakkında kapsamlı analizler yapılmıştır. Siber zorbalık sınıflandırma problemi ile ilgili yapılan çalışmalar, karşılaşılan zorluklar ve genel öneriler aşağıda verilmiştir.

- Verilerin toplanması: Çok sayıda kullanıcıya ilişkin paylaşım ve diğer bilgilerin toplanması Twitter API limitlerinden de kaynaklı olarak çok uzun zaman almaktadır. Bu sebepten belirli sayıda kullanıcı başlangıç olarak belirlenmiştir. Bu kullanıcıların rastgele belirlenmesi ise bu çalışmanın kısıtlarından biridir. Ayrıca, toplumsal veya doğal yaşamı etkileyen olaylar X ortamında çokça paylaşıma konu olmaktadır. Bu çalışmada kullanılan test verisinin oluşturulduğu tarihlerin de ülkemizde yaşanan 06.02.2023 tarihli Kahramanmaraş depremi ile kesişmesi veri setinde başta deprem olmak üzere birçok doğal afet ve bu doğal afetlerle mücadeleyi konu alan paylaşımların yer almasına sebep olmuştur. Tespit edilen topluluklar üzerinde uygulanan konu tespiti algoritmalarının çıktıları incelendiğinde de bu durum ile uyumlu olarak bazı toplulukların çoğunlukla deprem afetlerle mücadele ve bu alandaki siyasi ve sosyal çalışmalara ilişkin paylaşımları olduğu görülmüştür.

Buna ek olarak, siber zorbalık sınıflandırmasında Türkçe veri setleri ile çalışılan daha önceki çalışmalar incelendiğinde bu alanda etiketlenmiş bir eğitim veri setinin eksiliği göze çarpmaktadır. Bu çalışmada ise daha önceki çalışmalarda oluşturulan üç farklı eğitim verisi birleştirilerek eğitim verisinin kapsamı genişletilmeye çalışılmıştır. Ancak, bu alanda yapılacak çalışmalar için çok daha geniş kapsamlı, konularına göre kategorize edilmiş Türkçe bir veri seti hazırlanması oldukça önemlidir.

- Ön işleme adımları: Metin sınıflandırmaları ile ilgili çalışmalarda ön işleme adımları hem sınıflandırıcı başarımı hem de çalışma süresi yönünden geliştirilen modeli etkilemektedir. Eğitim veri seti ile birçok ön işleme adımı uygulanarak testler yapılmıştır. Noktalama işaretleri, hashtagler, semboller, sayısal ifadeler, etkisiz kelimeler, URL'ler ve özel karakterlerin veri setinden kaldırılması ve tüm kayıtların küçük harfe dönüştürülmesi performansı artırırken kelimelere kök alma işleminin uygulanması performansı etkilememiştir.

Ayrıca, Türkçe'de iki harfli anlamlı kelimelerin olmamasından dolayı veri setinde yer alan iki veya daha az karakterli ifadeler de kaldırılmış ve model performansının arttığı gözlemlenmiştir.

Bu çalışmada sınıflandırma yapılırken semboller, emojiler, yazım yanlışlarının düzeltilmesi işlemleri göz ardı edilmiştir. Ancak, sosyal ağ ortamlarında kullanıcılar genellikle paylaşımlarını gündelik konuşma dili ile yaptıkları için metinlerde yer alan yazım yanlışları sınıflandırma sonuçlarını oldukça etkilemektedir. Gelecek çalışmalarda test veri setinde yer alan metinlerdeki yazım yanlışlarının kelimeler arasındaki benzerlikler kullanılarak düzeltilmesi ve bu sayede sınıflandırma başarısının daha da artırılması hedeflenmektedir. Buna ek olarak, emojiler, semboller ve bazı kelimelerin yazım şekillerinin (büyük harfle yazılması, ünlem işareti kullanılması vb.) yer aldığı metnin duygusunu ortaya koyabileceği de bilinmektedir. Bu durumların da sonraki çalışmalarda kullanılması planlanmaktadır.

- Özellik çıkarım ve seçim tekniklerinin uygulanması: Bu çalışma da sekiz farklı sınıflandırma algoritması CountVectorizer ve TFIDF özellik çıkarım, Ki-Kare ve bilgi kazanımı özellik seçim teknikleri ile test edilmiştir. Sonuçlara göre CountVectorizer ve Ki-Kare tekniklerinin birlikte kullanımı bu veri seti üzerinde diğer kombinasyonlara göre daha başarılı olmuştur.

- Sınıflandırma: Bu aşamada sekiz farklı makine ve derin öğrenme algoritması farklı özellik çıkarım ve seçim teknikleri ile bir araya getirilerek oluşturulan modeller eğitim verisi üzerinde test edilmiş ve en başarılı sonuçları 0,90 doğrulukla Yapay Sinir Ağları algoritmasının CountVectorizer ve Ki-Kare teknikleri ile birlikte kullanımı vermiştir. Türkçe veri setleri kullanılarak siber zorbalık sınıflandırması yapılan çalışmalar ile bu tez çalışması kapsamında geliştirilen modele ilişkin başarım istatistikleri Çizelge 5.1'de sunulmuştur. Önerilen yaklaşıma ilişkin elde edilen sonuçlar daha önceki çalışmalarla benzer şekildedir.

Çizelge 5.1. Türkçe veri setleri ile yapılan siber zorbalık sınıflandırma çalışmaları

Çalışma	Platform	Özellik Çıkarım/Seçim Teknikleri	Değerlendirme Ölçütleri	En Başarılı Sınıflandırıcı
[14]	Instagram, X	Ki-Kare, Bilgi Kazanımı	F1-skor: %81	MNB
[26]	X	BagOfWords	Doğruluk: %90,1	AdaBoost
[27]	X	BagOfWords	F1-skor: %91	SVM
[28]	X	TF-IDF, N-Gram	F1-skor: %91	ANN
[46]	X	Word2Vec, FastText	F1-skor: %93	CNN
Önerilen Yaklaşım	X	CountVectorizer, Ki-Kare	Doğruluk: %90 F1-skor: %88,7	ANN

Topluluk tespiti ile ilgili yapılan çalışmalar, yaşanan zorluklar ve öneriler aşağıdaki gibidir.

- Siber zorbalık skorunun oluşturulması: Zorbalık yapan toplulukların çok daha hızlı ve etkin olarak belirlenebilmesi için oluşturulacak olan ağın özelleştirilmesinde kullanılmak üzere her bir kullanıcı için siber zorbalık skoru hesaplaması yapılmıştır. Sınıflandırma aşamasında 267 384 kullanıcıya ait 37 235 001 paylaşım sınıflandırılmış ve bu tahminler doğrultusunda kullanıcıların siber zorbalık skoru oluşturulmuştur. Bu çalışmada belirlenen siber zorbalık skoru her bir kullanıcı için zorbalık içeren paylaşımlarının toplam paylaşımlarına oranı olarak ifade edilebilir. Her ne kadar bu çalışmada sınıflandırma sonuçlarının tek başına kullanılması anlamlı sonuçlar vermiş olsa da siber zorbalık

skorunun hesaplanmasına paylaşımların beğeni sayıları, retweet sayıları, paylaşımı yapan veya yayılmasını sağlayan kullanıcıların popülerlik durumlarının dahil edilmesinin oluşturulacak ağın amaca uygunluğunu artıracağı değerlendirilmektedir. Gelecek çalışmalarda bu konu ele alınarak hem topluluk tespiti sonuçlarının iyileştirilmesi hem de tespit edilen zorbalığın şiddetinin ve türünün belirlenmesi hedeflenmektedir. Daha önceki çalışmalarda dikkate alındığında bu tür kullanıcı ve metin özelliklerinin çalışmalara dahil edilmesinin elde edilecek sonuçları pozitif anlamada etkileyebileceği öngörülmektedir [20-23].

- Ağın oluşturulması ve topluluk tespiti: Siber zorbalık skoru dikkate alınarak $\mu>0$, $\mu>0,25$, $\mu>0,5$ ve $\mu>0,75$ koşullarına uyan kullanıcıların takip/takipçi bilgileri kullanılarak dört farklı ağ oluşturulmuştur. Oluşan ağlarda yer alan kullanıcı ve bağlantılara ilişin istatistikler Çizelge 4.4'te sunulmuştur. Oluşan her bir ağ üzerinde Louvain algoritması kullanılarak topluluk tespiti çalışması yapılmıştır. Çalışma sonuçları, modülarite ölçümüne göre karşılaştırıldığında en başarılı topluluk tespiti 0,877 ile $\mu>0,5$ olan Graf3'te ilişkilerin yönleri dikkate alınmadığında elde edilmiştir. Graf3'te diğer ağlara göre daha yüksek modülerlikle topluluk tespitinin yapılması siber zorbalık skorunun varlığının önemini açıklamaktadır. Ancak, ağ yapısı oluşturulurken takip/takipçi bilgileri ile birlikte kullanıcılar arasındaki etkileşimler de kullanılarak ağırlıklı ağların oluşturulması daha özel grupların tespit edilmesine olanak sağlayabilir.

Bu tez çalışmasında elde edilen sonuçlar incelendiğinde, ortaya çıkarılan toplulukların çoğunlukla belirli konu başlıklarında paylaşımlara sahip olan kullanıcılardan oluştuğu gözlemlenmektedir. Bu durum belirli bir amaca yönelik topluluk halinde organize bir şekilde hareket eden kullanıcı gruplarının tespit edilmesinde ve sosyal ağlardaki kutuplaşmalar ile topluluk halinde yapılan organize faaliyetlerin tespit edilmesi açısından oldukça önemli bir sonuçtur.

Sonuçlar, siber zorbalık tespitinin yanı sıra topluluk tespiti ile kullanıcılar arasındaki bilinmeyen ilişkileri tespit etme kabiliyeti de vermiştir. Topluluklarda yer alan kullanıcıların paylaşımları incelendiğinde çoğunlukla aynı konular üzerinde paylaşımların yapıldığı büyük sosyal çevreler karşımıza çıkmaktadır. Birbirleri ile birinci dereceden bir etkileşimi olmasa da aynı konular üzerinde paylaşımlara sahip olan ve siber zorbalık skoru yüksek bir kullanıcı ile olan etkileşimlerinden dolayı aynı toplulukta yer alan kullanıcıların paylaşımlarının birbirleri ile aynı konularda olduğu gözlemlenmiştir.

Gelecek çalışmalarda, doğal dil işleme ve konu belirleme teknikleri kullanılarak, tespit edilen topluluklara ilişkin daha özel çıkarımların yapılması planlanmaktadır. Tespit edilen toplulukların ilgi alanlarının belirlenmesi yapılan zorbalığın türlerinin belirlenmesine de yardımcı olacaktır. Ayrıca, gelecek çalışmalarda kullanıcı ve metin özelliklerinin de hem sınıflandırma hem de topluluk tespiti aşamalarında kullanımı ile siber zorbalığın hedef kitlesinin daha iyi belirlenebileceği öngörülmektedir. Bu durum siber zorbalık faaliyetlerinin önlenmesi için oldukça önemlidir. Bu çalışmalara ek olarak, edebi ve dil bilgisi yönünden uzman kişilerce oluşturulacak geniş bir zorbalık kelimeler sözlüğü ve bu kelimelerin zorbalık şiddetine olan etkilerini içeren veri seti ile tespit edilen zorbalık faaliyetlerine ilişkin zorbalık şiddetine dair de çalışmalar yapılabilecektir. Bu durum da gelecek çalışmalarda ele alınabilecek önemli konulardan biridir.

Twitter API aracılığı ile toplanan verilere ek olarak paylaşımlara ilişkin metinsel özelliklerin de toplanması ve zorbalık skorunun oluşturulmasında kullanılmasının daha özelleşmiş toplulukların tespit edilmesine fayda sağlayacağı değerlendirilmektedir. Daha sonraki çalışmalarda, kullanıcıların yapmış olduğu paylaşımlara başka kullanıcılar tarafından verilen cevapların kullanılması ile siber zorbalık faaliyetinin hedefindeki kullanıcıların da belirlenmesi hedeflenmektedir.

Topluluk tespitinde, yüksek siber zorbalık skoruna sahip, popüler yani fazla sayıda takip/takipçiye ve paylaşımları çokça beğeni ve yeniden paylaşıma (*retweet*) sahip olan kullanıcılar sosyal arkadaş çevrelerinde bulunan ve siber zorbalık skoru düşük olan arkadaşlarını bulundukları topluluklara çekebilir. Bu olumsuzlukla ilgili gelecek çalışmalarda siber zorbalık skorunun popülerlik hesaplamaları ile birleştirerek topluluk bazında siber zorbalık skoru hesaplamasının yapılması planlanmaktadır. Böylece popüler kullanıcıların çalışma sonuçlarını saptırması en aza indirgenerek daha anlamlı toplulukların elde edilmesi hedeflenmektedir.

KAYNAKLAR

1. Chudal, R., Tiiri E., Klomek, A. B., Ong, S. H., Fossum, S., Kaneko, H. et al. (2021). Victimization by traditional bullying and cyberbullying and the combination of these among adolescents in 13 European and Asian countries. *European Child & Adolescent Psychiatry,* 31, 1391-1404.

2. Kowalski, R.M., Limber S.P. (2007). Electronic Bullying Among Middle School Students. *Journal of Adolescent Health*, 41(6), 22-30.

3. Hinduja, S., Patchin, J.W. (2008). Cyberbullying: An Exploratory Analysis of Factors Related to Offending and Victimization. *Deviant Behavior*, 29(2), 129-156.

4. Chun, J., Lee, J., Kim, J., Lee, S. (2020). An international systematic review of cyberbullying measurements. *Computers in Human Behavior*, 113(2),106485.

5. İnternet: Emily A. Vongels, (December, 2022). Teens and Cyberbullying 2022, *PEW Araştırma Merkezi*, Web: https://www.pewresearch.org/internet/2022/12/15/teens-and-cyberbullying-2022/, Son Erişim Tarihi: 13.11.2023.

6. İnternet: Emily A. Vongels, (January, 2021). The State of Online Harassment, *PEW Araştırma Merkezi*, Web: www.pewresearch.org/internet/2021/01/13/the-state-of-online-harassment, Son Erişim Tarihi:13.11.2023.

7. İnternet: Cook, S., (June, 2023). Cyberbullying statistics and facts for 2023. *Comparitech*, Web: https://www.comparitech.com/internet-providers/cyberbullying-statistics/, Son Erişim Tarihi:13.11.2023.

8. İnternet: Newall, M., (2018). A Global Advisor Survey: Cyberbullying, *Ipsos*, Web: https://www.ipsos.com/sites/default/files/ct/news/documents/2018-06/cyberbullying_june2018.pdf, Son Erişim Tarihi:13.11.2023.

9. Boyd, D., Ellison, N. (2007). Social Network Sites: Definition, History, and Scholarship. *Journal of Computer-Mediated Communication*, 13, 210-230.

10. İnternet: Lenhart A. (June, 2007), Cyberbullying, *PEW Araştırma Merkezi*, Web: https://www.pewresearch.org/internet/2007/06/27/cyberbullying/, Son Erişim Tarihi:13.11.2023.

11. İnternet: Justin W. Patchin, (2023). Summary of Our Cyberbullying Research (2007-2023), *Cyberbullying Research Center*, Web: https://cyberbullying.org/summary-of-our-cyberbullying-research, Son Erişim Tarihi:13.11.2023.

12. Yin, D., Xue, Z., Hong, L., Davison, B. D., Kontostathis, A., Edwards, L. (2009). Detection of harassment on web 2.0. *In Proceedings of the Content Analysis in the WEB*, 1–7.

13. Reynolds, K., Kontostathis, A., Edwards, L. (2011). *Using Machine Learning to Detect Cyberbullying*. 10th International Conference on Machine Learning and Applications and Workshops, Hawaii, 241-244.

14. Ozel, S. A., Sarac, E., Akdemir, S., Aksu, H. (2017). *Detection of cyberbullying on social media messages in Turkish*. 2017 International Conference on Computer Science and Engineering (UBMK), Antalya, 366-370.

15. Sintaha, M., Mostakim, M. (2018). *An Empirical Study and Analysis of the Machine Learning Algorithms Used in Detecting Cyberbullying in Social Media*. 21st International Conference of Computer and Information Technology (ICCIT), Dhaka, 1-6.

16. Sahni, A., Raja, N. (2018). *Analyzation and Detection of Cyberbullying: A Twitter Based Indian Case Study*. International Conference on Recent Developments in Science, Engineering and Technology REDSET 2017, 799, 484-497.

17. Dalvi, R. R., Chavan, S. B., Halbe, A. (2020). *Detecting A Twitter Cyberbullying Using Machine Learning*. 4th International Conference on Intelligent Computing and Control Systems (ICICCS), Madurai, 297-301.

18. Muneer, A., Fati, S.M. (2020). A Comparative Analysis of Machine Learning Techniques for Cyberbullying Detection on Twitter. *Future Internet*, 12, 187.

19. Talpur, B. A., O'Sullivan, D. (2020). Cyberbullying severity detection: A machine learning approach. *PLoS ONE*, 15 (10): e0240924.

20. Dadvar, M., Trieschnigg, D., Ordelman, R., De Jong, F. (2013*). Improving cyberbullying detection with user context*. Proceedings of the European conference on information retrieval 2013, Moskova, 693–696.

21. Huang, Q., Singh, V., Atrey, P. (2014). Cyber Bullying Detection Using Social and Textual Analysis. *SAM '14 2014*, 3-6.

22. Abdhullah-Al-Mamun, Akhter, S. (2018). *Social media bullying detection using machine learning on Bangla text*. 10th International Conference on Electrical and Computer Engineering (ICECE), Dakka, 385-388.

23. Dadvar, M., De Jong, F., Ordelman, R. J. F., Trieschnigg, R. B. (2012). *Improved cyberbullying detection using gender information*. Proceedings of the 12th Dutch-Belgian information retrieval workshop, Ghent, 23-25.

24. Al-garadi, M. A., Varathan, K. D., Ravana, S. D. (2016). Cybercrime detection in online communications: The experimental case of cyberbullying detection in the Twitter network. *Computers in Human Behavior*, 63, 433–443.

25. Balakrishnan, V., Khan, S., Arabnia, H. (2020). Improving Cyberbullying Detection using Twitter Users' Psychological Features and Machine Learning. *Computers & Security*, 90, 101710.

26. Bozyigit, A., Utku, S., Nasibov, E. (2021). Cyberbullying detection: Utilizing social media features. *Expert Systems with Applications*, 179, 115001.

27. Bozyigit, A., Utku, S., Nasiboglu, E. (2018). *Sanal Zorbalık İçeren Sosyal Medya Mesajlarının Tespiti*. 3rd International Conference on Computer Science and Engineering (UBMK), Saraybosna, 281-281.

28. Bozyigit, A., Utku, S., Nasiboglu, E. (2019). *Cyberbullying Detection by Using Artificial Neural Network Models*. 4th International Conference on Computer Science and Engineering (UBMK), Samsun, 520-524.

29. Mounir, J. H., Nashaat, M., Ahmed, M., Emad, Z., Amer, E., Mohammed, A. (2019). Social Media Cyberbullying Detection using Machine Learning. *International Journal of Advanced Computer Science and Applications (IJACSA)*, 10, 703-707.

30. Alam, K. S., Bhowmik, S., Prosun, P. R. K. (2021). *Cyberbullying Detection: An Ensemble Based Machine Learning Approach*. Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV), Tirunelveli, 710-715.

31. Alsubait, T., Alfageh, D. (2021). Comparison of Machine Learning Techniques for Cyberbullying Detection on YouTube Arabic Comments. *International Journal of Computer Science and Network Security (IJCSNS)*, 21(1).

32. Perera, A., Fernando, P. (2021). Accurate Cyberbullying Detection and Prevention on Social Media. *Procedia Computer Science*, 181, 605-611.

33. Rosa, H., Pereira, N., Ribeiro, R., Ferreira, P.C., Carvalho, J.P., Oliveira, S., Coheur, L., Paulino, P., Veiga Simão, A.M., Trancoso, I. (2019). Automatic cyberbullying detection: A systematic review. *Computers in Human Behavior*, 93, 333–345.

34. Iwendi, C., Srivastava, G., Khan, S., Maddikunta, P. K. R. (2020). Cyberbullying detection solutions based on deep learning architectures. *Multimedia Systems 2020*, 29, 1839-1852.

35. Kargutkar, S. M., Chitre, V. (2020). *A Study of Cyberbullying Detection Using Machine Learning Techniques*. Fourth International Conference on Computing Methodologies and Communication (ICCMC), Erode, 734-739.

36. Al-Ajlan, M. A., Ykhlef, M. (2018). Deep Learning Algorithm for Cyberbullying Detection. *International Journal of Advanced Computer Science and Applications (IJACSA)*. 9(9).

37. Girvan, M., Newman, M. E. J. (2002). Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, 99(12), 7821–7826.

38. Blondel, V. D., Guillaume, J.-L., Lambiotte, R., Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10).

39. Choi, Y., Jeon, B., Kim, H. (2021). Identification of key cyberbullies: A text mining and social network analysis approach. *Telematics and Informatics*, 56, 101504.

40. Park, J.-H., Kwon, H.-Y., (2021). Cyberattack detection model using community detection and text analysis on social media. *Information & Communications Technology Express*, 8(4), 499-506.

41. Li, C., Bai, J., Zhang, L., Tang, H., Luo, Y. (2019). Opinion community detection and opinion leader detection based on text information and network topology in cloud environment. *Information Sciences*, 504, 61–83.

42. Hristakieva, K., Cresci, S., Martino, G. D. S., Conti, M., Nakov, P. (2022). *The Spread of Propaganda by Coordinated Communities on Social Media*. WebSci '22: Proceedings of the 14th ACM Web Science Conference 2022, Barcelona, 191-201.

43. Cuzcano, X. M., Ayma, V. H. (2020). A Comparison of Classification Models to Detect Cyberbullying in the Peruvian Spanish Language on Twitter. *International Journal of Advanced Computer Science and Applications (IJACSA)*, 11(10).

44. Hussain, M. G., Mahmud, T. A., Akthar, W. (2018). *An Approach to Detect Abusive Bangla Text*. International Conference on Innovation in Engineering and Technology (ICIET), Dakka, 1-5.

45. Altay, E. V., Alatas, B. (2018). Detection of Cyberbullying in Social Networks Using Machine Learning Methods. *International Congress on Big Data, Deep Learning and Fighting Cyber Terrorism (IBIGDELFT)*, Ankara, 87-91.

46. On, E. P., Yeniterzi, R. (2020). *Cyberbullying Detection using Deep Learning and Word Embedding Analysis*. 28th Signal Processing and Communications Applications Conference (SIU), Gaziantep, 1-4.

47. Nandhini, S. B., Sheeba, I. J. (2015). Online Social Network Bullying Detection Using Intelligence Techniques. *Procedia Computer Science*, 45, 485–492.

48. Hemphill, S. A., Kotevski, A., Heerde, J. A. (2015). Longitudinal associations between cyber-bullying perpetration and victimization and problem behavior and mental health problems in young Australians. *International Journal of Public Health*, 60(2), 227–237.

49. Roh, Y., Heo, G., Whang, S. E. (2021). A Survey on Data Collection for Machine Learning: A Big Data-AI Integration Perspective. *IEEE Transactions on Knowledge and Data Engineering*, 33, 1328-1347.

50. İnternet: Beyaz M., (2021). Türkçe Sosyal Medya Paylaşımı Veri Seti. (n.d.). www.kaggle.com., Web: https://www.kaggle.com/datasets/mrtbeyz/trke-sosyal-medya-paylam-veri-seti Son Erişim Tarihi: 13.11.2023.

51. İnternet: Bozyigit A., (October, 2020). A comprehensive cyberbullying dataset including social media features, Web: https://data.mendeley.com/datasets/pgfk7h4367/1, Son Erişim Tarihi: 13.11.2023.

52. Curuk, E., Acı, C., Essiz, E. S. (2018). *Performance Analysis of Artificial Neural Network Based Classfiers for Cyberbulling Detection*. 3rd International Conference on Computer Science and Engineering (UBMK), Saraybosna, 1-5.

53. Aksangür, İ., Eren, B., Erden, C. (2022). Evaluation of data preprocessing and feature selection process for prediction of hourly PM10 concentration using long short-term memory models. *Environmental Pollution*, 311, 119973.

54. Li, J., Cheng, K., Wang, S., Morstatter, F., Trevino, R. P., Tang, J., Liu, H. (2018). Feature Selection. *Association for Computing Machinery Computing Surveys*, 50(6), 1–45.

55. Solorio-Fernández, S., Carrasco-Ochoa, J. A., Martínez-Trinidad, J. Fco. (2019). A review of unsupervised feature selection methods. *Artificial Intelligence Review*, 53(2), 907–948.

56. Ozel, S. A., Sarac, E. (2017). Effects of Feature Extraction and Classification Methods on Cyberbully Detection. *Süleyman Demirel Üniversitesi Fen Bilimleri Enstitüsü Dergisi*, 21(1), 190-200.

57. Dadgar, S. M. H., Araghi, M. S., Farahani, M. M. (2016). *A novel text mining approach based on TF-IDF and Support Vector Machine for news classification*. IEEE International Conference on Engineering and Technology (ICETECH), Coimbatore, 112-116.

58. C. -z. Liu, Y. -x. Sheng, Z. -q. Wei and Y. -Q. Yang, (2018). Research of Text *Classification Based on Improved TF-IDF Algorithm*. IEEE International Conference of Intelligent Robotic and Control Engineering (IRCE), 147, 218-222.

59. Sajid, K., Mehmoon, A., Huma, Q., Farooq, A., Marriam, N. (2023). Fake News Classification using Machine Learning: Count Vectorizer and Support Vector Machine. *Journal of Computing & Biomedical Informatics*, 4(01), 54–63.

60. Kumar, V. (2014). Feature Selection: A literature Review. *The Smart Computing Review*, 4(3), 211-229.

61. Venkatesh, B., Anuradha, J. (2019). A Review of Feature Selection and Its Methods. *Cybernetics and Information Technologies*, 19(1), 3–26.

62. Nihan, S. T. (2020). Karl Pearsons chi-square tests. *Educational Research and Reviews*, 15(9), 575–580.

63. Samuel, A. L. (1959). Some Studies in Machine Learning Using the Game of Checkers. *IBM Journal of Research and Development*, 3(3), 210–229.

64. Hastie, T., Tibshirani, R., Friedman, J. H. (2009). *The Elements of Statistical Learning*. (2 nd edition). New York: Springer, *9*.

65. Wang, D., (2001). *Unsupervised Learning: Foundations of Neural Computation*. AI Magazine, 22(2), 101–102

66. Sutton, R. S., Barto, A., (2018). *Reinforcement learning: an introduction*. (2 nd edition). Massachusetts:The Mit Press, 1-28.

67. Zhu, X., Goldberg, A.B. (2009). Introduction to Semi-Supervised Learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 3(1), 1-159.

68. Goodfellow, I., Bengio, Y., Courville, A., (2016). *Deep Learning*. *(*2 nd edition). Cambridge:MIT Press, 53-134 .

69. Murshed, B. A. H., Abawajy, J., Mallappa, S., Saif, M. A. N., Al-Ariki, H. D. E. (2022). DEA-RNN: A Hybrid Deep Learning Approach for Cyberbullying Detection in Twitter Social Media Platform. *IEEE Transactions on Social Network*, 25, 1117-1128.

70. Bharti, S., Yadav, A.K., Kumar, M., Yadav, D. (2021). Cyberbullying detection from tweets using deep learning. *Kybernetes*, 51(9), 2695-2711.

71. LeCun, Y., Bengio, Y., Hinton, G. (2015). Deep Learning, *Nature*, 521(7553), 436–444.

72. Neter, J., Kutner, M. H., Nachtsheim, C. J., Wasserman, W. (2009). *Applied linear statistical models*. (5 th edition). California: Morgan & Claypool Publishers, 33-102.

73. Maulud, D., Abdulazeez, A. M. (2020). A Review on Linear Regression Comprehensive in Machine Learning. *Journal of Applied Science and Technology Trends*, 1(4), 140–147.

74. Hosmer Jr., D. W., Lemeshow, S., Sturdivant, R. X. (2013). *Applied Logistic Regression*. (2 nd edition). Hoboken: Wiley, 1-26.

75. Böhning, D. (1992). Multinomial logistic regression algorithm. *Annals of the Institute of Statistical Mathematics*, 44(1), 197–200.

76. Bottou, L. (2010). Large-Scale Machine Learning with Stochastic Gradient Descent. *Proceedings of COMPSTAT'2010*, 34, 177–186.

77. Charbuty, B., Abdulazeez, A. (2021). Classification Based on Decision Tree Algorithm for Machine Learning. *Journal of Applied Science and Technology Trends*, 2(01), 20–28.

78. Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning*, 1(1), 81–106.

79. Özkan, S. B., Apaydin, S. M. F., Özkan, Y., Düzdar, I. (2019). *Comparison of Open Source Data Mining Tools: Naive Bayes Algorithm Example*. Scientific Meeting on Electrical-Electronics & Biomedical Engineering and Computer Science (EBBT), İstanbul, 1-4.

80. Shankar, V., Kumar, V., Devagade, U., Karanth, V., Rohitaksha, K. (2020). Heart Disease Prediction Using CNN Algorithm. *Springer Nature Computer Science*, 1, 170.

81. Zeiler, M.D., Fergus, R. (2014). *Visualizing and Understanding Convolutional Networks*, In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds) Computer Vision – ECCV 2014 (ECCV 2014) Lecture Notes in Computer Science, 8689, 818-833.

82. Cho, K., Bart van Merrienboer, Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y. (2014). *Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation*. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, 1724–1734.

83. Hochreiter, S., Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation*, 9(8), 1735–1780.

84. Amin Ali, O. M., Wahhab Kareem, S., Mohammed, A. S. (2022). Evaluation of Electrocardiogram Signals Classification Using CNN, SVM, and LSTM Algorithm: A review. *8th International Engineering Conference on Sustainable Technology and Development (IEC)*, Erbil, 185-191.

85. Sadri, Y., Afshord, S. T., Lotfi, S., Majidnezhad, V. (2022). An Extended Louvain Method for Community Detection in Attributed Social Networks. *Journal of Artificial Intelligence in Electrical Engineering*, 11(43), 11-23.

86. Chaudhary, L., Singh, B., (2018). Community Detection Using an Enhanced Louvain Method in Complex Networks. *Lecture Notes in Computer Science*, 11319, 243–250.

87. Zhang, J., Fei, J., Song, X., Feng, J. (2021). An Improved Louvain Algorithm for Community Detection. *Mathematical Problems in Engineering*, 2021, 1–14.

88. Tiv, M., Gullifer, J. W., Feng, R. Y., Titone, D. (2020). Using network science to map what Montreal bilinguals talk about across languages and communicative contexts. *Journal of Neurolinguistics*, 56, 100913.

89. Aynaud, T., Guillaume, J. L. (2010). *Static community detection algorithms for evolving networks*. In 8th International Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks IEEE, Avignon, 513-519.

90. David. (2011). Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. *ArXiv*, abs/2010.16061.

91. Sofaer, H. R., Hoeting, J. A., Jarnevich, C. S., (2019). The area under the precision-recall curve as a performance metric for rare binary events. *Methods in Ecology and Evolution*, 10(4), 565–577.

92. Blei, D.M., Ng, A.Y., Jordan, M.I., (2001). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3, 993-1022.

93. Yu, H., Yang, J., (2001). A direct LDA algorithm for high-dimensional data — with application to face recognition. *Pattern Recognition*, 34(10), 2067–2070.

94. Frank, H. (1969). *Graph Theory (on Demand Printing Of 02787)*. (1 st edition). New York: Academic Press, 2.

95. Newman, M.E. (2003). The Structure and Function of Complex Networks. *SIAM Review*, 45, 167-256.

96. Barabási, A.-L., Pósfai, M., (2016). *Network science*. (1 st edition). Cambridge: Cambridge University Press, 2.

97. Nikoletos, S., Raftopoulou, P., (2022). Employing social network analysis to dark web communities. *IEEE Xplore*, 311-316.

98. Barber, M. J., (2007). Modularity and community detection in bipartite networks. *Physical Review E*, 76(6), 066102.

99. Guo, K., Huang, X., Wu, L., Chen, Y. (2021). Local community detection algorithm based on local modularity density. *Applied Intelligence*, 52(2), 1238–1253.

EKLER

EK-1. Toplulukların konu tahminlerinin yapıldığı LDA çıktıları

Topluluk1**:**

```
Topic: 0
Words: 0.049*"chp" + 0.038*"parti" + 0.031*"oldu" + 0.031*"millet" + 0.024*"stanbul" + 0.022*"artık" + 0.019*"dünya" + 0.019*"hdp" + 0.018*"göre" + 0.018*"başkan"
Topic: 1
Words: 0.085*"erdoğan" + 0.042*"recep" + 0.036*"tayyip" + 0.032*"karşı" + 0.025*"bize" + 0.025*"size" + 0.025*"bay" + 0.025*"olsa" + 0.017*"slam" + 0.016*"ona"
Topic: 2
Words: 0.034*"tek" + 0.031*"iyi" + 0.028*"gün" + 0.027*"devlet" + 0.022*"başka" + 0.021*"aynı" + 0.021*"zaman" + 0.018*"son" + 0.015*"bizim" + 0.015*"ülke"
Topic: 3
Words: 0.045*"masa" + 0.034*"bunu" + 0.025*"siyaset" + 0.020*"real" + 0.017*"altında" + 0.017*"diyor" + 0.017*"siyasi" + 0.017*"yer" + 0.017*"altılı" + 0.016*"evet"
Topic: 4
Words: 0.056*"deprem" + 0.041*"türkiye" + 0.033*"büyük" + 0.028*"yalan" + 0.027*"yardım" + 0.019*"ilk" + 0.018*"hatay" + 0.018*"ayet" + 0.016*"devam" + 0.016*"depremzede"
Topic: 5
Words: 0.040*"türk" + 0.027*"zaten" + 0.024*"pkk" + 0.021*"olmak" + 0.021*"onu" + 0.021*"bin" + 0.021*"takip" + 0.017*"fazla" + 0.016*"terör" + 0.015*"kimse"
Topic: 6
Words: 0.047*"meral" + 0.046*"kemal" + 0.040*"kılıçdaroğlu" + 0.039*"akşener" + 0.036*"aday" + 0.035*"başkanı" + 0.027*"genel" + 0.022*"kurulu" + 0.021*"yönetim" + 0.020*"ali"
Topic: 7
Words: 0.094*"hayırlı" + 0.060*"amin" + 0.045*"allah" + 0.036*"inşallah" + 0.034*"mübarek" + 0.032*"doğru" + 0.032*"sabahlar" + 0.028*"kabul" + 0.025*"bizi" + 0.021*"dua"
Topic: 8
Words: 0.044*"devam" + 0.036*"cennet" + 0.035*"birlikte" + 0.032*"başkanımız" + 0.027*"milli" + 0.026*"onlar" + 0.026*"olmaz" + 0.024*"teşekkürler" + 0.023*"olmuş" + 0.021*"bey"
Topic: 9
Words: 0.167*"allah" + 0.053*"razı" + 0.042*"rahmet" + 0.038*"güzel" + 0.032*"rabbim" + 0.032*"teşekkür" + 0.027*"eylesin" + 0.027*"ederim" + 0.023*"versin" + 0.023*"diliyorum"
```

Topluluk2**:**

```
Words: 0.050*"eğitim" + 0.043*"kabul" + 0.040*"milli" + 0.034*"akp" + 0.033*"açık" + 0.024*"edildi" + 0.024*"gıda" + 0.023*"doğu" + 0.023*"ederiz" + 0.022*"depremzedeler"
Topic: 1
Words: 0.043*"rahmet" + 0.036*"diliyorum" + 0.027*"geçmiş" + 0.026*"hayatını" + 0.022*"allah" + 0.022*"acil" + 0.022*"gelen" + 0.021*"depremde" + 0.020*"sağ" + 0.019*"cennet"
Topic: 2
Words: 0.076*"deprem" + 0.032*"bin" + 0.027*"yardım" + 0.024*"şubat" + 0.022*"destek" + 0.019*"çadır" + 0.018*"ilk" + 0.017*"bölgesinde" + 0.015*"stanbul" + 0.014*"akşam"
Topic: 3
Words: 0.076*"devam" + 0.050*"türkiye" + 0.025*"cumhurbaşkanı" + 0.023*"ediyoruz" + 0.021*"erdoğan" + 0.020*"millet" + 0.019*"ortak" + 0.015*"halk" + 0.015*"partisi" + 0.015*"hakkında"
Topic: 4
Words: 0.033*"kılıçdaroğlu" + 0.029*"chp" + 0.029*"oldu" + 0.025*"iyi" + 0.024*"parti" + 0.023*"kemal" + 0.019*"aday" + 0.015*"doğru" + 0.015*"benim" + 0.015*"seçim"
Topic: 5
Words: 0.039*"yıl" + 0.032*"sağlık" + 0.031*"olmak" + 0.027*"türk" + 0.026*"önce" + 0.025*"üzere" + 0.024*"masa" + 0.017*"bölüm" + 0.017*"bugün" + 0.016*"arasında"
Topic: 6
Words: 0.062*"genel" + 0.043*"başkanımız" + 0.034*"başkanı" + 0.034*"sayın" + 0.030*"ziyaret" + 0.022*"belediye" + 0.021*"cumhurbaşkanımız" + 0.020*"canlı" + 0.019*"ettik" + 0.018*"hatay"
Topic: 7
Words: 0.047*"birlikte" + 0.038*"gün" + 0.028*"depremden" + 0.024*"güzel" + 0.023*"etkilenen" + 0.022*"zor" + 0.022*"teşekkür" + 0.021*"ediyorum" + 0.020*"dünya" + 0.018*"can"
Topic: 8
Words: 0.036*"tek" + 0.027*"büyük" + 0.024*"devlet" + 0.022*"ilgili" + 0.022*"altında" + 0.021*"enkaz" + 0.018*"kurtarma" + 0.018*"aynı" + 0.017*"kişi" + 0.016*"bütün"
Topic: 9
Words: 0.061*"allah" + 0.053*"akşener" + 0.034*"meral" + 0.029*"zaman" + 0.029*"artık" + 0.024*"başka" + 0.020*"bize" + 0.019*"hiçbir" + 0.017*"başladı" + 0.016*"eylesin"
```

Topluluk3**:**

```
Topic: 0
Words: 0.192*"videoya" + 0.176*"link" + 0.164*"gitmek" + 0.161*"dokunun" + 0.075*"kızı" + 0.063*"sik" + 0.054*"genç" + 0.020*"sikiyor" + 0.014*"sevgilisini" + 0.012*"turkifsa"
Topic: 1
Words: 0.095*"türbanlı" + 0.088*"ifşa" + 0.087*"azgın" + 0.080*"bayan" + 0.066*"azgınbayan" + 0.062*"çin" + 0.062*"zlemek" + 0.050*"öpüşme" + 0.050*"seks" + 0.050*"sikiş"
Topic: 2
Words: 0.162*"video" + 0.135*"türkporno" + 0.076*"azgın" + 0.068*"sakso" + 0.062*"sikiş" + 0.060*"bayan" + 0.060*"türbanlı" + 0.060*"swinger" + 0.047*"efsane" + 0.045*"çin"
Topic: 3
Words: 0.113*"sikiş" + 0.102*"türbanlı" + 0.095*"azgın" + 0.091*"öpüşme" + 0.086*"sakso" + 0.078*"pornoizle" + 0.073*"göt" + 0.061*"meme" + 0.050*"evli" + 0.030*"çin"
Topic: 4
Words: 0.224*"altyazılı" + 0.157*"turk" + 0.124*"türkçe" + 0.097*"ifsa" + 0.082*"çin" + 0.076*"zlemek" + 0.041*"türbanlı" + 0.035*"turkçe" + 0.032*"meme" + 0.020*"sikiyor"
Topic: 5
Words: 0.199*"üniversiteli" + 0.168*"kızlar" + 0.166*"türbanli" + 0.100*"türkifşa" + 0.075*"çin" + 0.072*"zlemek" + 0.072*"turkifsa" + 0.027*"azgınbayan" + 0.023*"sikiyor" + 0.016*"efsane"
Topic: 6
Words: 0.135*"sikiş" + 0.128*"sakso" + 0.112*"azgın" + 0.109*"türbanlı" + 0.099*"öpüşme" + 0.088*"pornoizle" + 0.062*"çin" + 0.057*"zlemek" + 0.032*"swinger" + 0.021*"sikiyor"
Topic: 7
Words: 0.170*"altyazılı" + 0.101*"türkçe" + 0.094*"turk" + 0.093*"turkifsa" + 0.080*"ifşa" + 0.077*"meme" + 0.059*"turkçe" + 0.057*"azgın" + 0.049*"çin" + 0.049*"zlemek"
Topic: 8
Words: 0.330*"ifşa" + 0.075*"videoları" + 0.063*"azgınbayan" + 0.060*"amcık" + 0.055*"aşkım" + 0.054*"bol" + 0.053*"azdım" + 0.049*"turkifsa" + 0.046*"devamı" + 0.041*"videosu"
Topic: 9
Words: 0.116*"altyazılı" + 0.079*"anal" + 0.069*"sert" + 0.061*"ifşa" + 0.060*"sakso" + 0.058*"meme" + 0.058*"kuzen" + 0.058*"türkçe" + 0.058*"abla" + 0.058*"turkçe"
```

Topluluk4**:**

```
Topic: 0
Words: 0.096*"isteyen" + 0.096*"cuckold" + 0.094*"varmı" + 0.094*"bunu" + 0.094*"çiftler" + 0.094*"yaşamak" + 0.094*"koca" + 0.089*"karısını" + 0.083*"şekilde" + 0.083*"izlemek"
Topic: 1
Words: 0.895*"şekilde" + 0.011*"merhaba" + 0.011*"isteyen" + 0.011*"karısını" + 0.011*"çiftler" + 0.011*"varmı" + 0.010*"bunu" + 0.010*"yaşamak" + 0.010*"izlemek" + 0.010*"cuckold"
Topic: 2
Words: 0.990*"merhaba" + 0.001*"isteyen" + 0.001*"karısını" + 0.001*"şekilde" + 0.001*"varmı" + 0.001*"cuckold" + 0.001*"yaşamak" + 0.001*"bunu" + 0.001*"çiftler" + 0.001*"koca"
Topic: 3
Words: 0.186*"isteyen" + 0.128*"karısını" + 0.094*"varmı" + 0.094*"şekilde" + 0.092*"yaşamak" + 0.092*"izlemek" + 0.090*"bunu" + 0.090*"çiftler" + 0.071*"koca" + 0.063*"cuckold"
Topic: 4
Words: 0.107*"merhaba" + 0.091*"şekilde" + 0.089*"isteyen" + 0.089*"varmı" + 0.089*"çiftler" + 0.089*"yaşamak" + 0.089*"bunu" + 0.089*"koca" + 0.089*"cuckold" + 0.088*"izlemek"
Topic: 5
Words: 0.220*"varmı" + 0.190*"bunu" + 0.190*"çiftler" + 0.159*"isteyen" + 0.159*"yaşamak" + 0.065*"karısını" + 0.004*"izlemek" + 0.003*"şekilde" + 0.003*"merhaba" + 0.003*"cuckold"
Topic: 6
Words: 0.753*"izlemek" + 0.031*"isteyen" + 0.025*"çiftler" + 0.025*"yaşamak" + 0.025*"bunu" + 0.025*"varmı" + 0.025*"cuckold" + 0.024*"karısını" + 0.024*"koca" + 0.024*"şekilde"
Topic: 7
Words: 0.801*"isteyen" + 0.130*"yaşamak" + 0.009*"varmı" + 0.008*"bunu" + 0.008*"çiftler" + 0.008*"karısını" + 0.008*"şekilde" + 0.008*"izlemek" + 0.006*"koca" + 0.006*"cuckold"
Topic: 8
Words: 0.135*"cuckold" + 0.134*"karısını" + 0.121*"varmı" + 0.121*"koca" + 0.121*"isteyen" + 0.121*"yaşamak" + 0.121*"bunu" + 0.121*"çiftler" + 0.001*"izlemek" + 0.001*"şekilde"
Topic: 9
Words: 0.172*"koca" + 0.153*"cuckold" + 0.134*"isteyen" + 0.134*"varmı" + 0.134*"bunu" + 0.134*"çiftler" + 0.134*"yaşamak" + 0.001*"izlemek" + 0.001*"şekilde" + 0.001*"karısını"
```

EK-1 (devam). Toplulukların konu tahminlerinin yapıldığı LDA çıktıları

Topluluk5**:**

```
Topic: 0
Words: 0.048*"erdoğan" + 0.048*"bin" + 0.030*"recep" + 0.028*"herkes" + 0.025*"tayyip" + 0.021*"kişi" + 0.019*"ali" + 0.018*"dedi" + 0.017*"kaç" + 0.017*"olmaz"
Topic: 1
Words: 0.065*"güzel" + 0.033*"ederim" + 0.028*"teşekkür" + 0.025*"mutlu" + 0.023*"evet" + 0.022*"olacak" + 0.021*"olsa" + 0.021*"teşekkürler" + 0.021*"ediyorum" + 0.021*"hocam"
Topic: 2
Words: 0.209*"allah" + 0.042*"inşallah" + 0.039*"rabbim" + 0.033*"rahmet" + 0.032*"razı" + 0.029*"eylesin" + 0.026*"versin" + 0.020*"olmak" + 0.020*"erdoğani" + 0.020*"cennet"
Topic: 3
Words: 0.118*"hayırlı" + 0.059*"zaman" + 0.045*"olduğu" + 0.043*"selam" + 0.038*"sabahlar" + 0.036*"günaydın" + 0.029*"onu" + 0.028*"dua" + 0.028*"aleyküm" + 0.025*"kardeşim"
Topic: 4
Words: 0.042*"türkiye" + 0.040*"önce" + 0.037*"millet" + 0.034*"tek" + 0.032*"devlet" + 0.025*"türk" + 0.022*"aday" + 0.018*"abd" + 0.018*"belli" + 0.016*"fotoğraf"
Topic: 5
Words: 0.070*"devam" + 0.045*"yalan" + 0.042*"meral" + 0.035*"nleyi" + 0.035*"akşener" + 0.030*"rei" + 0.024*"bizi" + 0.023*"geçmiş" + 0.018*"mamoğlu" + 0.016*"geri"
Topic: 6
Words: 0.051*"iyi" + 0.028*"olduğunu" + 0.027*"zaten" + 0.026*"doğru" + 0.020*"size" + 0.020*"siyaset" + 0.019*"büyük" + 0.019*"onlar" + 0.018*"yer" + 0.017*"lazım"
Topic: 7
Words: 0.030*"oldu" + 0.024*"insan" + 0.023*"sadece" + 0.023*"bütün" + 0.022*"demek" + 0.020*"altında" + 0.019*"can" + 0.018*"insanlar" + 0.017*"benim" + 0.016*"kahramanmaraş"
Topic: 8
Words: 0.065*"chp" + 0.039*"kemal" + 0.032*"parti" + 0.032*"erdoğan" + 0.030*"başkanı" + 0.025*"kılıçdaroğlu" + 0.024*"gün" + 0.024*"cumhurbaşkanı" + 0.020*"genel" + 0.020*"hdp"
Topic: 9
Words: 0.062*"deprem" + 0.029*"bunu" + 0.027*"karşı" + 0.024*"başka" + 0.020*"yardım" + 0.017*"kentsel" + 0.016*"hatay" + 0.015*"afad" + 0.014*"beni" + 0.013*"kimse"
```

Topluluk6**:**

```
Topic: 0
Words: 0.305*"sınırsız" + 0.194*"açık" + 0.178*"özele" + 0.132*"tel" + 0.119*"arasın" + 0.062*"pasifler" + 0.003*"öğrenci" + 0.003*"seven" + 0.002*"köle" + 0.002*"ensest"
Topic: 1
Words: 0.577*"seven" + 0.169*"arasın" + 0.155*"öğrenci" + 0.014*"tel" + 0.014*"özele" + 0.014*"açık" + 0.014*"pasifler" + 0.014*"sınırsız" + 0.014*"ensest" + 0.014*"köle"
Topic: 2
Words: 0.152*"açık" + 0.152*"özele" + 0.148*"tel" + 0.144*"arasın" + 0.142*"pasifler" + 0.120*"köle" + 0.055*"ensest" + 0.046*"sınırsız" + 0.022*"seven" + 0.019*"öğrenci"
Topic: 3
Words: 0.305*"pasifler" + 0.223*"arasın" + 0.145*"seven" + 0.117*"özele" + 0.116*"açık" + 0.083*"tel" + 0.003*"köle" + 0.003*"öğrenci" + 0.003*"sınırsız" + 0.003*"ensest"
Topic: 4
Words: 0.361*"arasın" + 0.360*"ensest" + 0.036*"seven" + 0.036*"özele" + 0.036*"açık" + 0.035*"köle" + 0.035*"sınırsız" + 0.035*"tel" + 0.034*"pasifler" + 0.033*"öğrenci"
Topic: 5
Words: 0.115*"açık" + 0.113*"özele" + 0.110*"arasın" + 0.108*"tel" + 0.103*"pasifler" + 0.099*"köle" + 0.089*"seven" + 0.088*"öğrenci" + 0.088*"ensest" + 0.087*"sınırsız"
Topic: 6
Words: 0.274*"açık" + 0.267*"özele" + 0.238*"tel" + 0.218*"arasın" + 0.001*"seven" + 0.001*"pasifler" + 0.001*"sınırsız" + 0.001*"ensest" + 0.001*"köle" + 0.001*"öğrenci"
Topic: 7
Words: 0.125*"açık" + 0.124*"özele" + 0.104*"arasın" + 0.102*"öğrenci" + 0.101*"pasifler" + 0.098*"tel" + 0.091*"ensest" + 0.087*"sınırsız" + 0.085*"köle" + 0.082*"seven"
Topic: 8
Words: 0.169*"pasifler" + 0.166*"öğrenci" + 0.153*"arasın" + 0.137*"açık" + 0.129*"özele" + 0.123*"tel" + 0.110*"seven" + 0.012*"köle" + 0.001*"ensest" + 0.001*"sınırsız"
Topic: 9
Words: 0.169*"sınırsız" + 0.120*"açık" + 0.120*"ensest" + 0.116*"özele" + 0.114*"tel" + 0.111*"köle" + 0.107*"arasın" + 0.101*"seven" + 0.042*"pasifler" + 0.000*"öğrenci"
```

Topluluk7**:**

```
Topic: 0
Words: 0.150*"kadın" + 0.140*"ipne" + 0.099*"hanımlar" + 0.077*"ister" + 0.076*"kimler" + 0.067*"kocalarınıza" + 0.059*"koca" + 0.049*"zamanla" + 0.049*"olmaya" + 0.048*"onlarda"
Topic: 1
Words: 0.345*"venezia" + 0.227*"you" + 0.183*"for" + 0.173*"the" + 0.039*"your" + 0.011*"week" + 0.001*"veneziada" + 0.000*"reel" + 0.000*"stanbul" + 0.000*"varsa"
Topic: 2
Words: 0.035*"kadın" + 0.021*"venezia" + 0.020*"olmaya" + 0.018*"ipne" + 0.018*"zamanla" + 0.018*"başlar" + 0.018*"onlarda" + 0.018*"giydirirseniz" + 0.018*"hanımlar" + 0.018*"davranırsanız"
Topic: 3
Words: 0.287*"veneziada" + 0.275*"the" + 0.256*"venezia" + 0.106*"your" + 0.036*"for" + 0.021*"you" + 0.002*"sınırsız" + 0.000*"stanbul" + 0.000*"yer" + 0.000*"maltepe"
Topic: 4
Words: 0.167*"reel" + 0.060*"seven" + 0.058*"varsa" + 0.057*"sadece" + 0.056*"ciddi" + 0.055*"arkadaş" + 0.055*"stanbul" + 0.055*"gaziosmanpaşa" + 0.055*"bayan" + 0.054*"arıyorum"
Topic: 5
Words: 0.077*"zevk" + 0.041*"göt" + 0.040*"veren" + 0.040*"edicek" + 0.040*"itahat" + 0.040*"küfür" + 0.039*"kaan" + 0.039*"sınırsız" + 0.038*"slm" + 0.038*"parlak"
Topic: 6
Words: 0.353*"pasif" + 0.255*"veren" + 0.013*"zevk" + 0.008*"parlak" + 0.008*"sınırsız" + 0.008*"arıyorum" + 0.008*"küfür" + 0.008*"slm" + 0.008*"atsın" + 0.008*"aşalık"
Topic: 7
Words: 0.119*"yer" + 0.103*"stanbul" + 0.102*"pasifim" + 0.101*"maltepe" + 0.100*"parlak" + 0.098*"bostancı" + 0.097*"kilsiz" + 0.077*"günaydın" + 0.077*"full" + 0.050*"erkekler"
Topic: 8
Words: 0.128*"the" + 0.094*"venezia" + 0.091*"cinens" + 0.074*"pasif" + 0.061*"arayışım" + 0.056*"haftanın" + 0.054*"slm" + 0.052*"adresi" + 0.052*"izmir" + 0.051*"filmleri"
Topic: 9
Words: 0.024*"kadın" + 0.022*"yer" + 0.021*"parlak" + 0.021*"maltepe" + 0.021*"pasifim" + 0.021*"full" + 0.021*"stanbul" + 0.020*"kilsiz" + 0.020*"bostancı" + 0.018*"günaydın"
```

Topluluk8**:**

```
Topic: 0
Words: 0.069*"gay" + 0.040*"ful" + 0.040*"selamlar" + 0.040*"sex" + 0.040*"cani" + 0.030*"yeri" + 0.021*"akşam" + 0.021*"yaşamak" + 0.021*"yazarmisin" + 0.021*"evli"
Topic: 1
Words: 0.128*"ful" + 0.108*"görüşmek" + 0.094*"varsa" + 0.082*"acık" + 0.071*"araç" + 0.036*"hemen" + 0.034*"yas" + 0.018*"gizlilik" + 0.018*"full" + 0.018*"atsın"
Topic: 2
Words: 0.160*"hemen" + 0.153*"gorusmek" + 0.093*"gay" + 0.074*"ful" + 0.043*"yeri" + 0.041*"kentteyim" + 0.040*"yas" + 0.035*"arac" + 0.020*"bakalım" + 0.020*"uyar"
Topic: 3
Words: 0.112*"arac" + 0.061*"ciddi" + 0.052*"gay" + 0.046*"merinosdayim" + 0.046*"açık" + 0.038*"görüşmek" + 0.028*"birileri" + 0.020*"varsa" + 0.018*"ful" + 0.017*"hemen"
Topic: 4
Words: 0.159*"görüşmek" + 0.151*"gay" + 0.121*"yeri" + 0.068*"ful" + 0.055*"hemen" + 0.041*"bursa" + 0.039*"gece" + 0.039*"merkezdeyim" + 0.032*"yas" + 0.031*"kentteyim"
Topic: 5
Words: 0.102*"gay" + 0.100*"ful" + 0.096*"acil" + 0.066*"kentte" + 0.056*"görüşmek" + 0.042*"sakso" + 0.036*"araç" + 0.029*"gorusmek" + 0.029*"cekmek" + 0.029*"msj"
Topic: 6
Words: 0.195*"uyumayan" + 0.094*"varsa" + 0.084*"gay" + 0.049*"çekmek" + 0.049*"sakso" + 0.034*"yazsin" + 0.026*"atarmisin" + 0.025*"heykeldeyim" + 0.013*"yer" + 0.013*"nsan"
Topic: 7
Words: 0.074*"etmek" + 0.059*"tatlı" + 0.050*"sohbet" + 0.047*"sahibi" + 0.046*"fantazi" + 0.045*"parmaklarını" + 0.045*"paylaşmak" + 0.030*"ayakalrini" + 0.030*"varmı" + 0.016*"gay"
Topic: 8
Words: 0.181*"gay" + 0.165*"araç" + 0.161*"ful" + 0.121*"gorusmek" + 0.050*"sabah" + 0.036*"hemen" + 0.025*"gece" + 0.025*"erken" + 0.025*"saatlerde" + 0.013*"heykel"
Topic: 9
Words: 0.095*"bursa" + 0.088*"yalova" + 0.076*"sohbet" + 0.062*"slm" + 0.053*"ayak" + 0.052*"fantazi" + 0.036*"gemlik" + 0.036*"orhangazi" + 0.034*"sahibi" + 0.034*"fetiş"
```

EK-1 (devam). Toplulukların konu tahminlerinin yapıldığı LDA çıktıları

Topluluk9**:**

```
Topic: 0
Words: 0.065*"azgın" + 0.051*"gizli" + 0.049*"olgun" + 0.047*"bayanlar" + 0.047*"dmden" + 0.047*"yazsın" + 0.044*"doyumsuz" + 0.030*"evli" + 0.029*"ateşli" + 0.029*"dul"
Topic: 1
Words: 0.101*"gizli" + 0.100*"bekar" + 0.099*"bayanlar" + 0.052*"olmayan" + 0.050*"yaşıyor" + 0.050*"atın" + 0.050*"olgun" + 0.050*"ankara" + 0.050*"görüşme" + 0.050*"yalnız"
Topic: 2
Words: 0.338*"özgür" + 0.020*"yavaş" + 0.020*"mansur" + 0.017*"ekrem" + 0.017*"mamoğlu" + 0.017*"dedi" + 0.016*"ilk" + 0.014*"adayı" + 0.014*"ardından" + 0.014*"depremin"
Topic: 3
Words: 0.024*"kızılay" + 0.022*"diyor" + 0.020*"millet" + 0.019*"iyi" + 0.015*"kılıçdaroğlu" + 0.014*"türkiye" + 0.014*"siyaset" + 0.014*"kemal" + 0.013*"canım" + 0.012*"istifa"
Topic: 4
Words: 0.063*"bağıra" + 0.057*"atın" + 0.053*"ağrı" + 0.042*"sert" + 0.042*"doyumsuz" + 0.042*"msj" + 0.042*"tanışalım" + 0.042*"vahşice" + 0.042*"isteyen" + 0.037*"azgın"
Topic: 5
Words: 0.046*"sesli" + 0.038*"isteyen" + 0.028*"bynlar" + 0.028*"heryastan" + 0.026*"kadın" + 0.024*"sadece" + 0.022*"yada" + 0.022*"güzel" + 0.021*"sekufurlu" + 0.021*"atsin"
Topic: 6
Words: 0.108*"çevresi" + 0.038*"tek" + 0.038*"yaş" + 0.037*"bayanlar" + 0.037*"dul" + 0.037*"stanbul" + 0.036*"farklı" + 0.035*"erkek" + 0.035*"yakası" + 0.035*"isteyen"
Topic: 7
Words: 0.068*"amini" + 0.051*"seven" + 0.049*"bayanlar" + 0.046*"azgin" + 0.045*"sesli" + 0.044*"got" + 0.044*"sert" + 0.044*"atsin" + 0.044*"deligini" + 0.044*"kufurlu"
Topic: 8
Words: 0.032*"zaman" + 0.027*"oldu" + 0.022*"kötü" + 0.020*"bende" + 0.017*"aynı" + 0.017*"beşiktaş" + 0.016*"artık" + 0.015*"sti" + 0.015*"hemen" + 0.015*"aşkım"
Topic: 9
Words: 0.083*"isteyen" + 0.051*"evli" + 0.047*"olgun" + 0.047*"dul" + 0.041*"bayan" + 0.036*"karısını" + 0.033*"bayanlar" + 0.029*"paylaşmak" + 0.028*"yaş" + 0.028*"sınırım"
```

Topluluk10**:**

```
Topic: 0
Words: 0.037*"yardım" + 0.029*"destek" + 0.027*"aday" + 0.024*"olmak" + 0.019*"onu" + 0.018*"teşekkürler" + 0.018*"şekilde" + 0.017*"beni" + 0.016*"sayın" + 0.016*"kılıçdaroğlu"
Topic: 1
Words: 0.166*"fenerbahçe" + 0.045*"maçı" + 0.043*"zaman" + 0.018*"kayserispor" + 0.016*"karar" + 0.014*"gerçek" + 0.014*"ilgili" + 0.013*"maç" + 0.013*"kayseri" + 0.013*"takip"
Topic: 2
Words: 0.037*"devam" + 0.028*"bizim" + 0.024*"herkes" + 0.024*"büyük" + 0.022*"türk" + 0.019*"falan" + 0.019*"valencia" + 0.019*"enner" + 0.019*"yıl" + 0.019*"ali"
Topic: 3
Words: 0.058*"son" + 0.036*"türkiye" + 0.031*"bunu" + 0.029*"ligi" + 0.026*"başka" + 0.024*"kardeşim" + 0.023*"avrupa" + 0.023*"belli" + 0.022*"dünya" + 0.022*"ferdi"
Topic: 4
Words: 0.059*"iyi" + 0.027*"jesus" + 0.024*"artık" + 0.024*"iki" + 0.022*"kötü" + 0.021*"zor" + 0.020*"göre" + 0.020*"can" + 0.018*"hatay" + 0.016*"jorge"
Topic: 5
Words: 0.041*"tek" + 0.038*"sezon" + 0.028*"milyon" + 0.026*"yer" + 0.026*"maç" + 0.023*"bin" + 0.021*"oyuncu" + 0.021*"teknik" + 0.021*"takım" + 0.019*"gelen"
Topic: 6
Words: 0.048*"gol" + 0.040*"ilk" + 0.039*"lig" + 0.036*"oldu" + 0.034*"galatasaray" + 0.032*"süper" + 0.022*"beşiktaş" + 0.020*"trabzonspor" + 0.020*"etti" + 0.020*"genel"
Topic: 7
Words: 0.051*"güzel" + 0.034*"adana" + 0.029*"aynı" + 0.026*"zaten" + 0.025*"doğru" + 0.023*"abi" + 0.023*"lütfen" + 0.018*"yanlış" + 0.017*"geldi" + 0.016*"günlük"
Topic: 8
Words: 0.050*"ali" + 0.038*"gün" + 0.038*"koç" + 0.031*"bugün" + 0.022*"başkan" + 0.020*"günü" + 0.020*"yay" + 0.019*"benim" + 0.018*"futbol" + 0.017*"burcu"
Topic: 9
Words: 0.047*"allah" + 0.044*"arda" + 0.031*"güler" + 0.028*"şubat" + 0.026*"orta" + 0.025*"geçmiş" + 0.025*"açık" + 0.024*"kişi" + 0.023*"saat" + 0.021*"günaydın"
```



Gazili olmak ayrıcalıktır