Araştırma Makalesi
Zeki Sistemler Teori ve Uygulamaları Dergisi 8(2) (2025) 115-129
DOI: 10.38016/jista.1580712

Türkçe Sosyal Medya Paylaşımlarında Doğal Dil İşleme ve Makine Öğrenmesi ile Kadına Yönelik Şiddet Tespiti

Merve Kavut[1], Amina Dzafic[1], Ulya Bayram[1*]

[1] Elektrik-Elektronik Mühendisliği, Mühendislik Fakültesi, Çanakkale Onsekiz Mart Üniversitesi, Çanakkale, Türkiye

mervekavut5@gmail.com, aminadzafic1@hotmail.com, ulya.bayram@comu.edu.tr

Öz

Ülkemizde kadına yönelik fiziksel ve duygusal şiddet her geçen gün artmaktadır. Fakat bu artışı engelleyici mekanizmaların üretim ve geliştirilmesi aynı hızı yakalayamamaktadır. Gelişen yapay zekâ teknolojisinden faydalanarak kadına yönelik şiddetin önüne geçebilmek için atılabilecek ilk adımlardan biri, sosyal medya paylaşımlarından kadına yönelik şiddeti olumlayıp destekleyenleri tespit edip, bu kişileri sosyal medya mecralarından engellemektir. Bu makale, bahsedilen ilk adımın atılması amacıyla Türkçe sosyal medya paylaşımları üzerine gerçekleştirilmiş bir doğal dil işleme (DDİ) çalışmasını anlatmaktadır. Öncelikle Türkiye'de geçmiş yıllardan beri halen yaygın olarak kullanılan bir sosyal medya forumu veri kaynağı olarak seçilmiş, sonrasında ise kadına yönelik şiddet olumlaması içeren beşten fazla sayıdaki başlık altındaki paylaşımlar toplanıp işaretlenerek yeni bir Türkçe veri seti oluşturulmuştur. Veri seti farklı yöntemlerle analiz edildikten sonra, DDİ literatüründe sık kullanılan öznitelik çıkarma yöntemleriyle paylaşımlar modellenip kelime çantası, Random Forest, Gradient Boosting gibi çeşitli makine öğrenmesi yöntemleriyle şiddet olumlaması tespiti deneyleri gerçekleştirilmiştir. Bulgulara göre sosyal medya ortamlarında kadına yönelik şiddet içeren paylaşım sayılarının, kadınları savunan paylaşımlardan daha az olduğu tespit edilmiş, var olan şiddet paylaşımlarının da psikolojik şiddet ve aşağılama gibi içeriklerden oluştuğu görülmüştür. Model değerlendirme sürecinde hassasiyet, geri çağırma, F1 ve AUC (Area Under Curve) metrikleri kullanılmıştır. Elde edilen sonuçlara göre, kadına yönelik şiddet içerikli paylaşımların %76 AUC ve %77 geri çağırma oranlarıyla tespit edilebildiği ortaya çıkmıştır. Bu bulgular, sosyal medyada kadına yönelik şiddet içeren paylaşımların otomatik tespit edilip engellenmesi gibi hassas çözümlerin ülkemizde uygulanabilirliğini göstermiştir.

Anahtar kelimeler: doğal dil işleme, sosyal medya, makine öğrenmesi, kadına yönelik şiddet, Türkçe dili

Detecting Violence Towards Women from Turkish Social Media Posts with Natural Language Processing and Machine Learning

Abstract

There is an increase in physical and emotional violence towards women in Turkey. However, the development of new mechanisms to prevent this increase cannot catch up. One of the first steps that can be taken to prevent violence towards women using the technological progress in artificial intelligence is to detect the social media posts that support such violence, and then to ban them. This article presents a natural language processing (NLP) study conducted with the mentioned goal in Turkish social media. After selecting a popular social media platform that has been widely used in Turkey for many years, more than five subject titles are selected and the posts below them are scraped, then labeled, constructing a novel Turkish data collection. Following data analyses with various techniques, popular NLP feature extraction techniques and several machine learning models such as bag of words, Random Forests, Gradient Boosting are used to detect posts that support violence towards women. According to the findings, the number of posts containing violence towards women in Turkish are less than those that defend women, and the existing violent posts contain psychological violence and humiliation towards women. During the model evaluation, precision, recall, F1, and AUC (Area Under Curve) metrics are utilized. According to the results, 76% AUC and 77% recall rates can be obtained in detecting violence towards women from social media posts. These findings demonstrate the possibility of applying real-life sensitive measures on social media such as the detection and blocking of emotional violence towards women in Turkey.

Keywords: Natural language processing, social media, machine learning, violence towards women, Turkish language

* Sorumlu yazar.
E-posta adresi: ulya.bayram@comu.edu.tr

Alındı : 6 Kasım 2024
Revizyon : 10 Şubat 2025
Kabul : 18 Mayıs 2025

1. Giriş (Introduction)

Kadına yönelik şiddet, dünya çapında ciddi bir insan hakları ihlali ve bir halk sağlığı sorunudur. Dünya Sağlık Örgütü'nün (WHO, 2021) raporuna göre dünya genelinde her üç kadından biri hayatında en az bir defa şiddete maruz kalmıştır ve bu saldırıların çoğu eş veya partnerlerden gelmektedir. Aynı rapora göre kadına yönelik şiddet vakalarının ortalama %42'si yaralanma, %38'i ise kadın cinayeti içermektedir. Türkiye'de yaşanan kadına yönelik şiddet vakaları da Dünya geneline paralel olarak artıştadır. 2014 ve 2015 yılları arası yapılan bir çalışma, bir sene içinde şiddet olaylarında %90'lık bir artış raporlamıştır (Aytaç vd., 2016). Ülkemizde gerçekleştirilen çeşitli anket çalışmaları da kadına yönelik şiddetin Türkiye'de endişe verici seviyelere geldiğini onaylamaktadır (Sen ve Bolsoy, 2017). Türkiye İstatistik Kurumu'nun (TÜİK) 2023 yılı raporuna göre ülkemizde popülasyonun %49,9'unu kadınlar oluşturmaktadır ve ülkemizde kadınların %27,4'ü gece yalnız yürürken güvende hissetmemektedir (TİK, 2023). Bu sosyal gerçeği anlamak ve buna çözüm yolları bulmak için, kadına yönelik şiddetle ilgili düşüncelerin toplumda nasıl şekillendiğini ve hangi sıklıkta dile getirildiğini incelemek önemlidir. Ancak anket çalışmaları gibi geleneksel yöntemler, katılım gönüllülük esasına dayandığından ve yalnızca belirli bir kesimin görüşlerine ulaşılacağı için Türkiye'de kadına yönelik şiddet düşüncesini anlamaya yeterli olmayacaktır. Onun yerine, sosyal medya platformlarında yapılmış paylaşımları kullanarak geniş popülasyonu kapsayan veri setleri oluşturmak, toplumun genel düşünce yapısını daha iyi anlamak için gereklidir.

Sosyal medya kullanıcıları, kimliklerini gizleyen kullanıcı isimleri altında gerçek düşüncelerini özgürce paylaşmaktadırlar. Doğal olarak, kadınlara yönelik şiddet ve nefret içeren bireyler, bu düşüncelerini sosyal medya ortamlarında paylaşmaktadırlar. Toplumdaki kadına yönelik şiddeti anlamak ve önüne geçebilmek için sosyal medya paylaşımlarından kadına yönelik şiddeti destekleyenleri tespit etmek, önemli bir aşamadır. Özellikle de yapay zekâ alanındaki teknolojik gelişmelerden faydalanarak sosyal medya paylaşımlarından şiddet düşüncesi tespiti mekanizmalarının geliştirilmesi ve bu tip paylaşımlar yapanların bu mecralardan engellenmesi, atılacak ilk adım olabilir. Bu vizyonu paylaşan çeşitli araştırmacılar, İspanyolca dilinde yapılmış X (Twitter) paylaşımları üzerine DDİ yöntemleri uygulayarak Meksika'da kadınlara yönelik şiddeti anlamak için kullanmış (Castorena vd., 2021; Gonzales ve Cantu-Ortiz, 2021), Hindistan'da Hintçe paylaşımlara DDİ uygulamış (Tommasel vd., 2018), ayrıca İngilizce sosyal medya paylaşımlarında da kadına yönelik şiddet düşüncesini tespit çalışmaları yapmışlardır (Guo vd., 2023; Trinh vd., 2022; Rodriquez vd., 2021). Fakat, ülkemizde Türkçe sosyal medya paylaşımları üzerine, kadına yönelik şiddet tespiti amacıyla yapılan çalışma sayısı, bu konuda bir adım atılması açısından yetersizdir.

Bu makalede, literatürdeki kadına yönelik şiddeti tespit etme konusundaki yetersiz sayıdaki çalışma boşluğunu doldurmak ve sosyal medya ortamlarında uygulanabilir bir adım atılması için, Türkçe sosyal medya paylaşımlarında kadına yönelik şiddet düşüncesinin tespitine yönelik bir çalışma sunulmaktadır. Bu çalışmada amaç, sosyal medya verilerinde kadına yönelik şiddet içeren paylaşımların otomatik tespiti yapılarak engellenmeleri, böylece sosyal medyada kadına yönelik şiddetin ve nefret içeriklerinin yükselişinin önüne geçilmesidir. Çalışmada, uzun yıllardır kullanıma açık olan ve Türkiye'nin önde gelen sosyal medya platformlarından biri olan Ekşi Sözlük'teki paylaşımlar incelenmiştir. Literatürde ilk kez, Türkçe sosyal medya paylaşımlarından kadına yönelik şiddet odaklı bir veri seti oluşturulmuş ve paylaşımlar iki araştırmacı tarafından, şiddeti olumlama veya karşı çıkma yönünde işaretlenmiştir. Ayrıca, bu çalışmada ilk defa Türkçe dilinde kadına yönelik şiddet düşüncesi üzerine DDİ analizleri gerçekleştirilmiş; ardından, şiddeti olumlayan paylaşımları otomatik olarak tespit edebilmek için makine öğrenimi yöntemlerini içeren modeller geliştirilmiştir. Ek olarak, DDİ literatüründe sıkça kullanılan ancak Türkçe dilinde sınırlı uygulanan Word2Vec yöntemiyle kadına yönelik şiddet düşüncesini tespit üzerine analizler yapılmıştır.

Bu çalışma, literatüre pek çok yenilik katmasının yanı sıra, gelecekte Türkçe dilinde yapılacak araştırmalar için de kadınlar, çocuklar ve çeşitli gruplara yönelik şiddet düşüncesinin analiz ve tespitinde önemli bir adım olacaktır. Bu sayede, ülkemizin daha sağlıklı bir toplum olmasının önünde engel teşkil eden, şiddeti onaylayan ve öven paylaşımlar belirlenip önlenebilecektir.

2. Literatür (Literature)

Literatürde kadın odaklı çalışmalar az olmakla beraber, genel olarak siber zorbalık konusunda çok sayıda çalışma yapılmıştır. Şiddetin dijital bir türü olarak ortaya çıkan siber zorbalık, teknoloji aracılığıyla birini taciz etmek anlamına gelir ve son yıllarda internet kullanımıyla birlikte oldukça yaygınlaşmıştır. 2011 yılına ait bir çalışma, "Formspring.me" mecrasında yapılmış İngilizce paylaşımları toplayıp, Amazon Mechanical Turk aracılığı ile zorbalık içeren paylaşımları gönüllülere işaretletmiş, daha sonra işaretli verilere Karar Ağacı (Decision Tree) gibi makine öğrenmesi yöntemleri uygulamıştır (Reynolds vd., 2011). Bu çalışmada, test seti olarak ayırdıkları paylaşımlarda siber zorbalık içerenleri %78,5 doğruluk oranı ile tespit edebilmişlerdir. Bir başka çalışma da İngilizce Twitter paylaşımları üzerine kelime gömme

(word embedding), kelime çantası (BoW) ve Lineer Destek Vektör Makineleri (L-SVM) kullanarak siber zorbalık tespiti yapmıştır (Zhao vd., 2016).

Cinsiyet odaklı siber zorbalık tespiti yapan çalışmalar arasında biri SVM ve Yineleyen Sinir Ağları (RNN) yöntemlerini birleştirerek Facebook paylaşımlarından %86,10 hassasiyet (precision) ve Twitter paylaşımlarından %86,31 geri çağırma (recall) elde etmiştir (Tommasel vd., 2018). Benzer başka bir çalışma ise tf-idf özniteliklerini SVM ve Lojistik Regresyon (LR) yöntemlerini eğitmede kullanarak %74 F1 skoru elde etmiştir (Perera ve Fernando, 2021). DDİ'deki teknolojik gelişmelerin sonucu olarak ortaya çıkan "Bidirectional Encoder Representations from Transformers" (BERT) yönteminden faydalanan çalışmalar da bulunmaktadır (Desai vd., 2021). Bu yöntem, literatürde çok sayıda metin sınıflandırma çalışmasında başarılı sonuçlar alınmasını sağlamıştır (Bayram vd., 2022; Zhou vd., 2024).

Diğer bir psikolojik şiddet türü de hakaret, ayrımcılık ve nefret söylemleridir. Bu tip şiddet durumlarını tespit etmek için de çok sayıda çalışma yapılmıştır. 2018 yılında ait bir çalışma, RNN yöntemini kullanarak Twitter paylaşımlarından cinsiyetçilik ve ırkçılığı tespit etmeyi amaçlamış, %84 Eğri Altındaki Alan (AUC) skoru elde etmiştir (Pitsilia vd., 2018). Bir başka çalışma, Konvolüsyonel Sinir Ağları (CNN), Uzun Kısa Süreli Bellek (LSTM) ve bu yöntemin iki yönlüsü olan BiLSTM gibi derin öğrenme yöntemlerini kullanarak ırkçılık ve cinsiyetçilik tespitinde yüksek performans elde etmiştir (Kapil vd., 2020). Güney Afrika'daki İngilizce Twitter paylaşımları üzerine yapılan bir çalışma n-gram öznitelikleri ile SVM, LR, RF, Gradient Boosting (GB) yöntemlerini kullanarak nefret söylemlerini tespit etmiş ve %89 hassasiyet skoru elde etmiştir (Oriola ve Kotze, 2020). YouTube yorumlarını kullanarak şiddet tespiti yapmış bir çalışma ise CNN ve LSTM yöntemleri, GloVe ve fastText kelime gömme modellerini kullanmıştır (Ashraf vd., 2020). En yüksek başarıyı BoW ve BiLSTM ile %94 AUC ve %85 F1 skoru bularak elde etmiştir. Arapça Twitter paylaşımları üzerinde BERT kullanılarak yapılan bir çalışma da fiziksel şiddet tehditlerini %59,60 F1 skoru ile tespit edebilmiştir (Alsehri vd., 2020).

Psikolojik şiddetin de ötesinde, fiziksel şiddeti tespit etmek için de DDİ çalışmaları yapılmıştır. Evlilik içi şiddeti sosyal medya paylaşımlarından tespit etmek için yapılmış bir çalışma, Evrensel Cümle Kodlayıcı (USE) ve RF, SVM, LR, Naive Bayes (NB) gibi makine öğrenmesi yöntemlerini kullanarak %89 başarı oranı raporlamıştır (Trinh Ha vd., 2022). Reddit paylaşımlarını kullanan bir araştırma, aile içi şiddeti tespit için 2010-2021 yılları arasında dört subreddit'te yapılmış paylaşımlara SVM ve BERT'in bir çeşidi olan RoBERTa yöntemlerini uygulamıştır (Guo vd., 2023). #MeToo etiketi taşıyan Twitter paylaşımlarını toplayan bir çalışma cinsel şiddet yaşayanların anlatımlarını BoW, tf-idf, LSVM, LSTM gibi yöntemlerle modelleyerek, cinsel şiddet yaşayanları %80,4 hassasiyet ve %83,4 geri çağırma performansı ile tespit etmiştir (Hassan vd., 2020).

Dünyadaki literatüre rağmen, ülkemizde Türkçe dilinde şiddet tespitini hedefleyen çalışmaların sayısı yeterli değildir. Örneğin, karar ağaçları ve istatistiksel analizler kullanarak Türkiye'deki kadına yönelik şiddet vakalarının sosyal ve psikolojik boyutlarını hesaplayan çalışmalar yapılmıştır (Ensari vd., 2022; Cihan ve Karakaya, 2017). Az sayıdaki çalışmalardan biri de arama motoru sorgularını kullanarak Türkçe küfür tespitine odaklanan, n-gram, LR, SGD, L-SVM, Multinomial NB, K-En Yakın Komşuluk (K-NN), RF, XGBoost ve BERT'in bir versiyonu olan Electra yöntemlerini kullanıp, %0,93 F1 skoruna ulaşan çalışmadır (Soykan vd., 2022). Alternatif bir çalışma da insanların kadınlara yönelik şiddet konusunda görüşlerini belirlemek amacıyla Latent Dirichlet Allocation (LDA) adlı bir başlık tespiti yöntemini kullanmış ve farklı bakış açılarını tespit etmiştir (Okkali vd., 2021). Türkçe tweetlerden oluşan kadınlara yönelik siber nefreti otomatik tespit etmeyi amaçlayan bir çalışma da literatürdeki diğer çalışmalar gibi tf-idf ve SVM, DT, RF, NB gibi yöntemleri kullanmıştır ve yüksek doğruluk ama düşük geri çağırma skoru elde etmiştir (Şahi vd., 2018).

Bu makalede anlatılan çalışma, literatürdeki çalışmalardan farklı olarak ilk defa Ekşi Sözlük verilerinden, makine öğrenmesi yöntemleri ile kadına yönelik şiddet tespiti gerçekleştirmektedir. Ayrıca, literatüre bir katkı olarak, ilk defa Word2Vec yöntemini de içeren detaylı bir Türkçe kadına şiddet düşüncesini analiz çalışması sunulmaktadır.

3. Yöntemler (Methods)

3.1. *Veri toplama ve işaretleme (Data collection and labeling)*

Ekşi Sözlük, Türkiye'nin en popüler sosyal medya forumlarından biri olup, kullanıcıların çeşitli konular hakkında yasalar çerçevesinde özgürce yorum yapmalarına olanak tanımaktadır. Başlangıcı 1999 yılına uzanan bu forum, Türkiye'deki ilk uzun soluklu sosyal medya platformlarından biridir ve bu web sitesinde 1999 senesinden günümüze kadar yazılmış çeşitli başlıklara ve ilgili yorumlara ulaşmak mümkündür. Bu özellikleri sayesinde değerli bir veri seti kaynağı olan bu forum, bu çalışmanın veri kaynağı olarak seçilmiştir.

Veri kaynağını seçtikten sonra, kadına yönelik şiddete dair düşünceler içerebilecek paylaşımların bulunduğu konu başlıkları seçilmiştir. Bu seçim yapılırken, öncelikle Ekşi Sözlük mecrasında "kadın" veya "kız" kelimelerini içeren başlıklar sıralanmış, aralarında yakın zamanlarda paylaşım yapılmış olanlar seçilmiştir. Daha sonra, üç kadın araştırmacı, başlıkların

isimlerine bakarak, her birinde kadına yönelik şiddeti olumlayan paylaşımlar olabileceği konusunda hemfikir olmuştur. Seçilen başlıklar Tablo 1'de görülmektedir.

Tablo 1. Ekşi Sözlük platformundan seçilen başlıklar, her başlıkta kadına yönelik şiddeti olumlayan ve buna karşıt paylaşımların sayıları (Headings selected from the Ekşi Sözlük environment with their numbers of posts that are pro and against violence towards women)

Başlık	Kısaltma	Olumlu	Karşı
Dayak yiyen kadına yardım etmeyen tavlacı dayılar	DAYI	40	252
Güzel ama psikolojisi bozuk kadın	GÜZEL	21	151
Hiç kadın şair olmamasının sebepleri	ŞAİR	34	1089
Kadınlar ısrar eden erkeklere bayılırlar	ISRAR	50	1053
Türk kızının en yetenekli olduğu konu	YETENEK	381	1445
Diğer Başlıklar	DİĞER	17	1167
Toplam	543		5156

Seçilenler arasında yeterli sayıda paylaşım içeren beş ana başlık bulunmaktadır. Ek olarak, daha az sayıda paylaşım içeren beş başlık daha seçilmiş ve içerikleri birleştirilerek "Diğer Başlıklar" ismi ile veri setine eklenmiştir. Diğer başlıklara dahil edilen başlık isimleri de şu şekildedir: "40 yaşında problemli bekar müdür kadın," "feminist vegan yogacı 30 yaş üstü kadınlar," "kocasına kahvaltı hazırlamayan kadın," "türk kadınının temel sorunu, türk kızları neden gülümsemiyor sorunsalı." Böylece toplamda on başlık altında yapılan paylaşımlar kullanılmıştır. Seçili başlıklardan paylaşımları toplama işlemi, Python programlama dili ve web kazıma ile ilgili kütüphaneler (BeautifulSoup, requests) kullanılarak 7 Aralık 2023 tarihinde gerçekleştirilmiştir. Bu nedenle veri setindeki paylaşımlar bu tarih ve öncesine aittir. Seçilen başlıklardaki konular şu şekilde özetlenebilir:

- "Dayak yiyen kadına yardım etmeyen tavlacı dayılar" başlığındaki paylaşımlar 2021 yılında yaşanmış bir kadına yönelik şiddet olayına şahit olan esnafın tepkisizliğini eleştiren veya esnafa hak veren karşıt görüşlü içeriklere sahiptir.
- "Güzel ama psikolojisi bozuk kadın" başlığında kadınlara yönelik olumsuz genellemelerde bulunan paylaşımlar ve bunlara karşıt görüşler yer almaktadır.
- "Hiç kadın şair olmamasının sebepleri" başlığı altında da olumsuz bir genelleme ve buna karşıt görüşler yer almakta, bazı paylaşımlarda psikolojik şiddet ve kadınları aşağılamaya varan içerikler bulunmaktadır.
- "Kadınlar ısrar eden erkeklere bayılırlar" başlığı da benzer şekilde hem kadınları destekleyici hem de psikolojik şiddet ve hakaret edici paylaşımlar içermektedir.
- Veri setinde kadınlara yönelik şiddeti destekleyici en fazla sayıdaki paylaşım ise "Türk kızının en yetenekli olduğu konu" başlığında tespit edilmiştir.

Paylaşımların işaretlenmesinde iki lisans mezunu kadın mühendis görev almıştır ve her bir paylaşımı "şiddeti olumlayan," "şiddete karşıt," "nötr" ve "kararsız" olmak üzere dört etiketten en uygun olanı ile işaretlemişlerdir. Üçüncü kadın araştırmacı da işaretleri kontrol edip, doğruluklarını onaylamıştır. "Nötr" ve "kararsız" etiketlerine sahip paylaşım sayıları, diğer iki etikete göre oldukça azdır. Çünkü seçilen on başlık altında yapılan paylaşımlar genel olarak ya şiddete net olarak karşı çıkmış ya da şiddeti açıkça desteklemiştir.

3.2. Öznitelik çıkartma (Feature extraction)

Literatür özetine göre çeşitli dillerde sosyal medya verilerinden siber zorbalık ya da genel olarak şiddet düşüncesi tespiti yapan çalışmalar n-gram veya tf-idf özelliklerini sık kullanmışlardır. N-gram özellikleri unigram, bigram, trigram gibi alt özniteliklerden oluşup, en fazla n sayıdaki kelimenin peş peşe kullanım istatistiğini ifade eder. Kelime çantası (BoW) olarak da bilinen unigram, her kelimenin her paylaşımda kaç defa kullanıldığını, bigram da peş peşe yer alan iki kelimenin paylaşımlarda hangi sıklıkta birlikte yer aldığını ifade eder. Trigram yani üç kelimenin peş peşe gelme istatistiğinin katkısı unigram ve bigram kadar olmadığı için çoğu çalışma n=2'yi yeterli bulmuştur (Pestian vd., 2020).

Kelime çantası yönteminde her kelimenin her metinde var olma sıklığı hesaplanır ve öznitelik olarak kullanılır. Bu sayede, her kelimenin paylaşımlarda ne kadar sık geçtiği belirlenir ve kelime frekansları öznitelik uzayında yer alır. Örneğin, "şiddet" kelimesinin oluşturulan veri setindeki paylaşımlarda sık geçmesi beklenir. Bu durumda, "şiddet" kelimesinin öznitelik değeri yüksek olacaktır, çünkü kelime sık bir şekilde belge içinde yer almaktadır. Bu çalışmada, kelime çantası öznitelikleri kullanılmıştır. Eğitim setinde 5 veya daha az defa var olan kelimeler, anlamsal olarak katkı sağlamayacakları için kelime çantasından çıkarılmıştır.

3.3. *Makine öğrenmesi (Machine learning)*

Makine öğrenmesi yöntemleri geniş bir skalaya sahiptir ve DDİ alanında kullanılabilen çeşitli teknikler sağlar. Bu makalede anlatılan çalışmada, üç aşamada makine öğrenmesi yöntem skalasından faydalanılmıştır. Bunlar konu modelleme, kelime gömme ve paylaşım sınıflandırmadır.

3.3.1 Konu modelleme (Topic modeling)

Konu ya da başlık modelleme olarak bilinen ve geniş metin koleksiyonlarından aydınlatıcı bilgiler sağlayan yöntemlerden biri olan LDA, her bir metni kelime istatistiklerinden oluşan bir yapı olarak görür ve kelimelerin sıralamasını hesaba katmadan sadece bu istatistiklerle ilgilenir (Okkalı vd., 2021; Deng vd., 2016). LDA algoritması temel olarak iki matematik varsayımı içerir: 1) Otomatik tespit edilecek konu ya da başlıklar, kelimeler üzerinde bir olasılık dağılımıdır ve 2) her metin bu konu ya da başlıklar üzerinde bir olasılık dağılımına sahiptir. Bu iki varsayıma dayanarak şu formül izlenir (Blei vd., 2003):

$$p(k \mid d) = \sum_t p(k \mid t) \cdot p(t \mid d) \qquad (2)$$

P(k|d) koşullu olasılığı bir *d* paylaşımının *k* konusuna ait olma olasılığını gösterir. Burada *p(k|t)* *t* kelimesinin *k* başlığında bulunma olasılığını ve *p(t|d)* de *d* paylaşımında *t* kelimesinin bulunma olasılığını ifade eder. Formül, bir belgede verilen bir konunun olasılığının, belgede verilen her bir kelimenin olasılıklarının ağırlıklı toplamı olarak hesaplanabileceğini göstermektedir.

LDA yöntemi büyük veri setlerindeki konuları belirlemek için faydalı olsa da her yöntem gibi önemli bir dezavantaja sahiptir, o da konu sayısı parametresini kullanıcıdan istemesidir (Bayram, 2022). Fakat büyük veri setlerinde var olabilecek konu sayısını tahmin edebilmek mümkün değildir. Bu zorluğun üstesinden gelmek için kullanılabilecek yöntemlerden biri farklı konu sayıları ile denemeler yapmak ve her deneme sonunda elde edilen konuların kendi içinde anlamlı olup olmadığını otomatik ölçmektir. Ölçme metrikleri arasından tutarlılık (coherence) ve karışıklık (perplexity) metrikleri seçilmiştir.

3.3.2 Kelime gömme (Word embedding)

Kelime gömme değeri hesaplama fikri, Word2Vec algoritması ile popülerleşmiş, büyük veri setlerindeki kelimeler arası ilişkileri bir vektör uzayında temsil ederek anlamsal ilişkiler kurma yöntemidir (Mikolov vd., 2013a). Bir gizli katmana sahip bir yapay sinir ağını eğitip, öğrenilen ağırlıklar vektörünü bir kelimeyi ifade etmede kullanan bu yöntem, iki farklı şekilde eğitim gerçekleştirebilir: Continuous Bag of Words (CBOW) ve Skip-gram. CBOW, bir cümlenin bağlamında eksik olan kelimeyi tahmin etmeye çalışır ve matematiksel olarak şu fonksiyonla ifade edilir:

$$P(w_t \mid w_{t-n}, \ldots, w_{t+n}) \qquad (3)$$

Bu formülde, w_t tahmin edilmek istenen kelimeyi, koşullu olasılık kısmı ise o kelimenin *n* kelime öncesi ve sonrasındaki kelimeleri ifade eder. Böylece verilen bir bağlamdan, alakalı kelimeyi tahminlemeyi öğrenir.

Skip-gram modeli ise verilen bir kelimeden yola çıkarak çevresindeki bağlam kelimelerini tahmin etmeye çalışır ve şu şekilde formüle edilir (Mikolov vd., 2013b):

$$P(w_{t-n}, \ldots, w_{t+n} \mid w_t) \qquad (4)$$

Eğitimin sonunda, benzer bağlamlarda geçen kelimeler, vektör uzayında birbirine yakın konumlanarak, anlamsal benzerlik gösteren kelimelerin matematiksel olarak da yakın olmasını sağlar.

3.3.3 Sınıflandırma (Classification)

Literatür taramasına göre çeşitli dillerde siber zorbalık veya şiddet düşüncesinin otomatik tespitinde makine öğrenmesi ile metin sınıflandırması sıklıkla tercih edilmiştir. Bu yaklaşımdan ilham alarak, makaledeki çalışmada çeşitli makine öğrenmesi yöntemleri ile Türkçe paylaşımlarda kadına yönelik şiddet düşüncesinin otomatik tespiti üzerine modeller geliştirilmiştir. Deneylerde kullanılan sınıflandırma yöntemleri aşağıda sıralanmıştır.

LR: Basit sınıflandırma yöntemlerinden biri olan LR, öznitelik değerlerine bağlı olarak doğrusal bir fonksiyon oluşturur (Pedregosa vd., 2011). Bu doğrusal fonksiyonun sonucunu, sigmoid fonksiyonu ile 0 ile 1 arasında bir olasılığa dönüştürerek her örneğin belirli bir sınıfa ait olma olasılığını hesaplar. Eğitim sürecinde, sınıfları en iyi şekilde ayırmak için ağırlıklar güncellenir. Test aşamasında ise, model bu olasılığa göre verileri sınıflandırır. Bu çalışmada az parametre istemesi ve daha önceki çalışmalarda ispatladığı başarısı nedeniyle seçilmiştir (Bayram vd., 2019; Bayram ve Benhiba, 2021). Bu yöntemde lbfgs optimize yöntemi ve 100 iterasyon kullanılmıştır.

NB: Eğitim setinde ait olduğu sınıf verilen örneklerin özniteliklerinin birbirlerinden olasılıksal olarak bağımsız olduğu varsayımına dayanan bu yöntem, sınıflandırmak için Bayes teoremini uygular (Pedregosa vd., 2011). Avantajlarından biri, az sayıda eğitim verisiyle ve kullanıcıdan parametre istemeden başarılı sonuçlar dönebilmesidir.

L-SVM: Lineer kernel içeren SVM yöntemidir. LR ile benzerlikler taşısa da destek vektörleri sayesinde eğitilmesi sayesinde bazı hatalı sınıflandırma hatalarını tolere edebilir (Pedregosa vd., 2011). Çok sınıflı sınıflandırma problemlerinde ikişerli sınıflar arasında sınıflandırma gerçekleştirerek genel sonucu hesaplar. L-SVM modelinin C parametresi 1 ve iterasyon sayısı 1000 olarak kullanılmıştır.

GB: Her adımda bir kayıp fonksiyonunun negatif eğimine göre karar ağaçları ekleyerek modelini geliştiren bir algoritmadır (Pedregosa vd., 2011). Her adım, bir önceki modelin hatalarını azaltmaya çalışır ve bu sayede daha yüksek doğruluğa ulaşır. İkili sınıflandırma, yalnızca tek bir regresyon ağacının oluşturulduğu özel bir durumdur. Bu çalışmada 0.1 öğrenme oranı ve 100 alt model kullanılmıştır.

RF: GB yöntemi gibi, karar ağaçları oluşturarak modelleme yapan bir yöntemdir. Veri setinin çeşitli parçaları üzerine bir dizi karar ağacı sınıflandırıcısı oluşturur ve bu sınıflandırma sonuçlarının ortalaması ile en son kararı verir (Pedregosa vd., 2011). Çok parametre

içerdiği için eğitim setinden ayrılan parçalar üzerinde optimize edilerek en uygun parametreler bulunup otomatik kullanılmıştır. Bu parametreler arasında ağaç sayısı 100, 250, 500, 750 ve 1000 değerleri arasında, en uzun derinlik sayısı 3, 5 ve 15 değerleri arasında, yapraklarda olmasına izin verilen en düşük veri sayısı ise 2, 5 ve 8 değerleri arasında optimize edilmiştir.

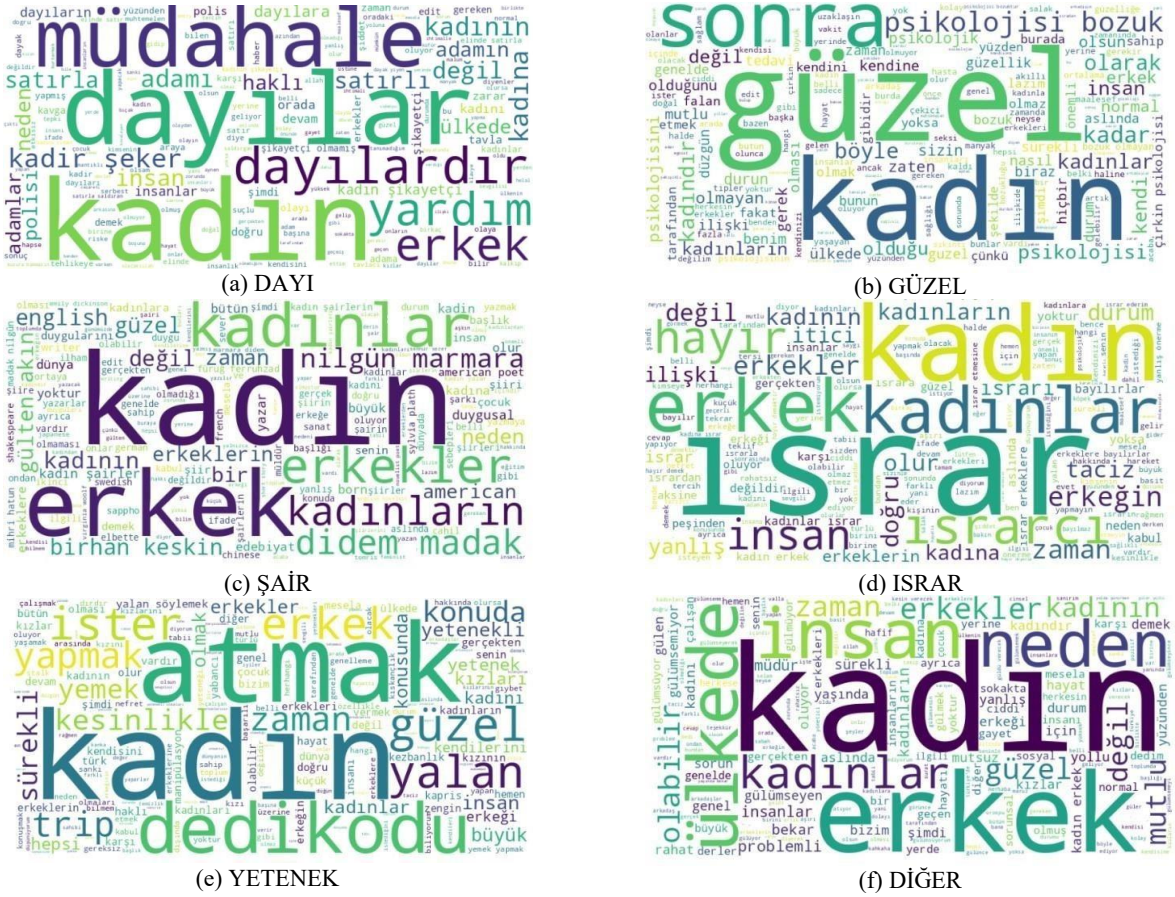
Sınıflandırmanın önemli bir aşaması da test seti üzerinde performans ölçümü yapmaktır. Bu makalede anlatılan çalışmada, dört adet metrik seçilmiştir. Bu metriklerin hesaplanmasında dört adet ölçüm kullanılır. İlk ölçüm doğru sınıflandırılmış pozitif sınıf sayısı, yani TP değeridir. Diğeri yanlış sınıflandırılmış negatif sınıfa ait veri sayısı, yani FP'dir. FN ise yanlış sınıflandırılmış pozitif sınıf verisi sayısıdır. TN ise doğru sınıflandırılmış negatif sınıf verisi sayısıdır. Bu değerler kullanılarak hesaplanan metriklerden biri hassasiyettir ve şu şekilde formüle edilir:

$$Hassasiyet(precision) = \frac{TP}{TP+FP} \qquad (5)$$

Hassasiyet skoru, pozitif sınıf olarak tahmin edilenlerin gerçekten ne kadarının bu sınıfa ait olduğunu gösterir. Geri çağırma metriğinin formülü de şöyledir:

$$Geriçağırma(recall) = \frac{TP}{TP+FN} \qquad (6)$$

Geri çağırma ise pozitif sınıfa ait verilerin ne kadarının pozitif sınıf olarak sınıflandırıldığını ifade eder. Bu iki metriğin harmonik ortalaması ise F1 skorunu oluşturur ve genel olarak sınıflandırma başarısını gösterir. Tüm bu metriklerden bağımsız alternatif bir metrik de AUC olarak bilinir ve pozitif sınıfa ait olup doğru sınıflandırılmış veri oranını negatif sınıfa ait olup da yanlışlıkla pozitif olarak sınıflandırılmış veri oranına-karşı çizer ve altındaki alanı performans ölçütü olarak kullanır (Tomak ve Bek, 2010).



(a) DAYI (b) GÜZEL

(c) ŞAİR (d) ISRAR

(e) YETENEK (f) DİĞER

Şekil 1. Veri setindeki başlıklardan oluşturulmuş kelime bulutları (Word clouds generated from the headings present in the dataset)

4. Bulgular (Results)

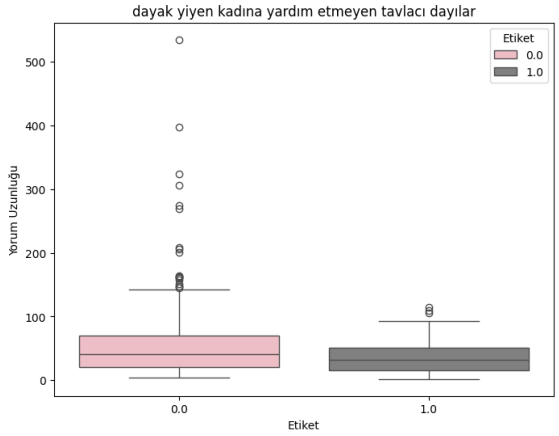
Çalışmada dört çerçevede deneyler gerçekleştirilmiştir. İlk çerçevede literatürde ilk defa oluşturulan Türkçe sosyal medya paylaşımlarında kadına yönelik şiddeti tespit veri seti analiz edilmiştir. İkinci çerçevede ise konu modelleme deneyleri gerçekleştirilerek başlıkların altında tartışılan düşünceler benzerliklerine göre gruplandırılmıştır. Üçüncü çerçevede kelime gömme yöntemi ile deneyler gerçekleştirilerek Türkçe dilinde kadına yönelik şiddet düşüncesine dair kelimesel bağlamlar araştırılmıştır. Dördüncü çerçevede ise sosyal medyada kadına yönelik

şiddet içeren paylaşımları otomatik tespit ve engelleme modelleri denenmiştir.

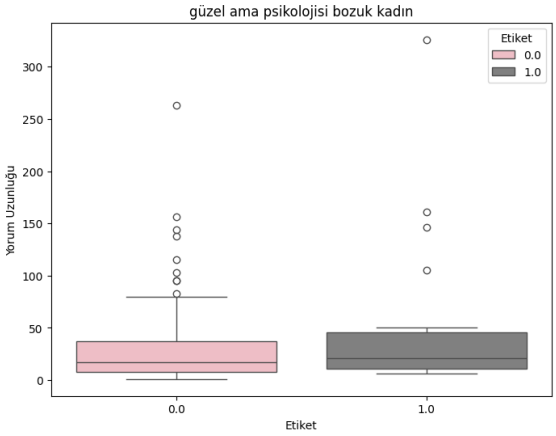
4.1. Veri analizi (Data analysis)

Binlerce sosyal medya paylaşımı içeren veri setinin içeriğini anlayabilmek için uygulanacak ilk analiz yöntemlerinden biri kelime bulutları oluşturmaktır. Şekil 1'de görülen kelime bulutları hem sık kullanılmış kelime ve deyimleri içermekte ve bunların popülerliklerini tespit etmeyi sağlamaktadır. Örneğin "DAYI" başlığında en sık kullanılan kelimeler arasında "müdahale" ve "kadir şeker" görülmektedir. İnternet taraması ile ulaşılan haberlere göre bu isme sahip kişi, 5 Şubat 2020 tarihinde bir kadının şiddet gördüğünü düşünüp çifti ayırmaya çalışırken erkek kişiyi yaralayıp ölümüne sebep olmuş ve hapis cezasıyla yargılanmıştır. Buradan da anlaşılmaktadır ki, başlıkta sözü geçen dayıların kadına yardım etmemesi ve duruma müdahale etmemesine hak vermek için bu haberdeki olay sıklıkla örnek verilmiştir. Benzer durum "GÜZEL" başlığının bulutundaki "çirkin psikolojisi" kelime öbeğinde görülmektedir.

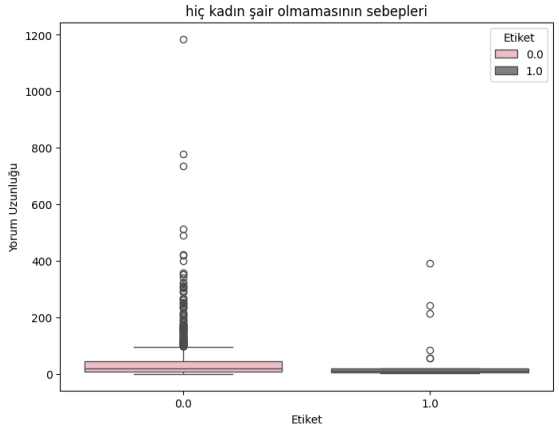
Şekil 1'de görülen "ŞAİR" başlığında ise "birhan keskin," "didem madak" ve "nilgün marmara" isimleri görülmektedir. Üç adet ünlü Türk kadın şairlere ait bu isimlerin popülerliği göstermektedir ki bu başlık altındaki paylaşımlar, başlıktaki önermeyi çürütmek amacıyla kullanılmıştır. "YETENEK" başlığının bulutunda da en sık kullanılan kelimeler arasında "trip," "yalan," "dedikodu" ve "yemek" görülmektedir. Türk kızlarının en yetenekli olduğu konuları araştıran bu başlıkta bu kelimelerin sıklıkla kullanılması ise genel olarak bu başlık altına yazanların kadına yönelik psikolojik şiddete katkı sağladıklarını göstermektedir.



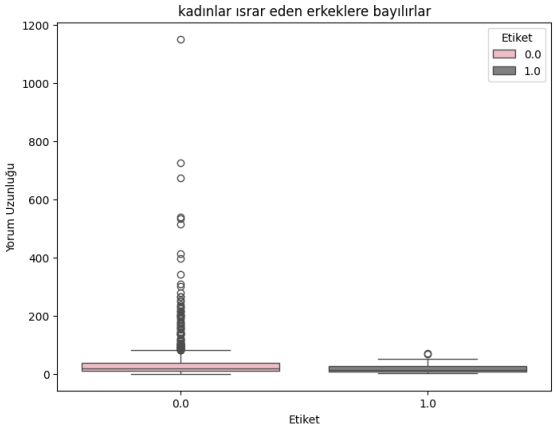
a) DAYI



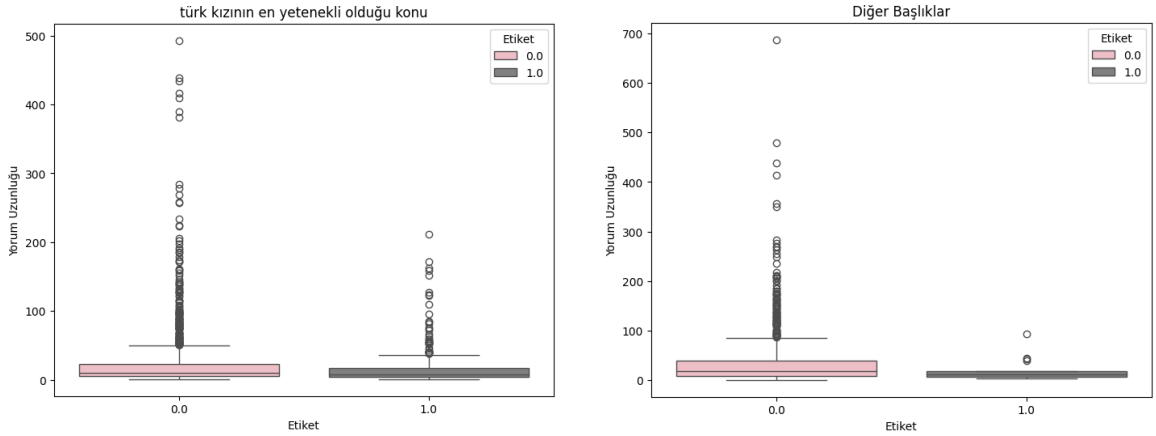
b) GÜZEL



c) ŞAİR



d) ISRAR



e) YETENEK f) DİĞER

Şekil 2. Başlıklardaki paylaşımların kelime sayısı olarak uzunluklarının dağılımları 0: kadına şiddete karşıt, 1: şiddet yanlısı (Distribution of the word counts per heading where label 0 is against violence towards women, 1 is supportive of violence)

Kelime bulutlarından sonraki analiz aşaması ise kadına yönelik şiddeti olumlamasına veya buna karşı çıkmasına göre işaretlenmiş paylaşımların kelime sayısı bağlamındaki uzunluklarına odaklanmaktadır. Bu analizde amaç bu iki karşıt düşünce şekline sahip bireylerin ne kadar uzun ya da kısa paylaşımlar yaptığını tespit etmektir.

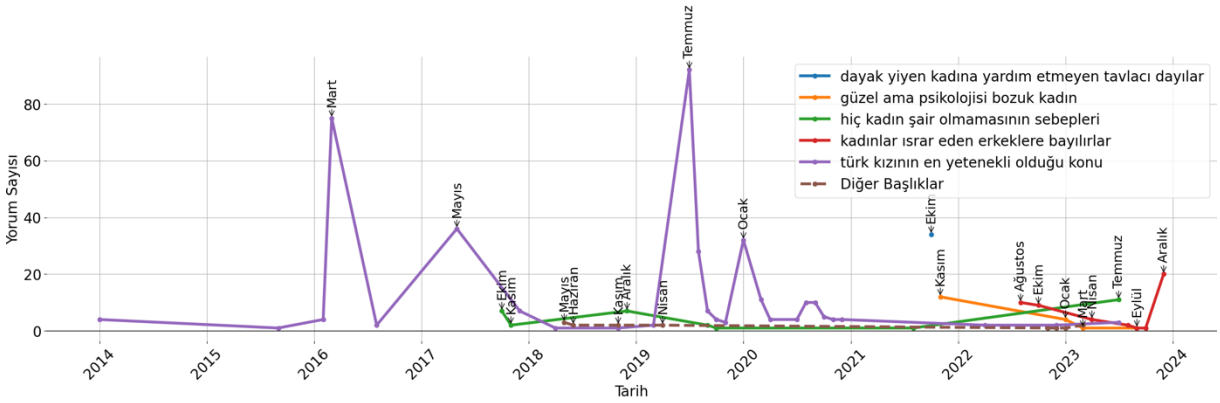
Şekil 2'de görülen dağılımlar göstermektedir ki bazı başlıklarda iki karşıt görüşten paylaşım uzunlukları ortalama olarak benzer kelime sayılarına sahiptir, bazı başlıklarda ise kesin tezatlar bulunmaktadır. Örneğin, "YETENEK" başlığında şiddet karşıtı ve yanlısı yorumlar benzer ortalama uzunluklara sahiptir. "GÜZEL" başlığında ise ortalama kadın düşmanı paylaşımların uzunlukları diğerlerinden fazladır. Bu istatistik göstermektedir ki seçilmiş tüm başlıklar arasında şiddet yanlısı düşünen kullanıcılar güzel ama psikolojisi bozuk başlığında düşüncelerini uzun metinlerle paylaşmayı tercih etmişlerdir. Diğer yandan, kalan "DAYI" ve "ŞAİR" gibi başlıklarda şiddet karşıtı paylaşımların uzunlukları diğer sınıftan fazladır. Bu da bu başlıklara daha çok başlıkta yapılan negatif önerileri çürütme amaçlı paylaşımlar yapıldığını göstermektedir. Bu başlıklarda şiddeti olumlayan paylaşımların etkileşim almak ve başlıkları gündemde tutmak için kısa metinlerle yapıldığını iddia etmek mümkündür.

Şekil 2'de dikkat çeken bir diğer bulgu da paylaşım uzunluklarının standart sapmalarıdır. "YETENEK" başlığında her iki sınıfa ait paylaşım uzunluklarının standart sapmaları yüksektir. Diğer yandan, geriye kalan başlıklardaki yüksek standart sapmalar sadece kadınlara yönelik şiddete karşı çıkan paylaşımlarda görülmektedir.

Şekil 3'te görülen grafikte, her başlıkta şiddet içeren paylaşımların sayısı ve popülerliğinin zaman içerisinde değişimini görmek mümkündür. Aralarında ilk paylaşımı en eskiye dayanan "YETENEK" başlığıdır. Bu başlığa 2016 yılının mart ayında, ertesi sene mayıs ayında, 2019 yılı temmuz ayında ve sonraki senenin başında tekrar tekrar yoğun sayıda kadına şiddeti onaylayan paylaşımlar eklenmiştir. Devamında oluşturulan başlık "ŞAİR" başlığı olmuştur ve seneler boyu paylaşım almaya devam etmiş ve 2024 yılına doğru paylaşım alma hızı artmıştır.

Genel olarak kadınları eleştiren başlıklardan bir diğeri olan "GÜZEL" de 2021 yılında oluşturulmuş ve paylaşım almıştır. "ISRAR" başlığı ise "GÜZEL"den sonra oluşturulmuş olsa da 2024 yılına doğru popülerleşmiş ve daha fazla paylaşım almaya başlamıştır. "DAYI" başlığı da bu başlığın ilk ortaya çıkmasına sebep olan olayın gerçekleştiği zaman olan 2021 yılı Ekim ayında paylaşımlar almıştır ve başlığa sonraki zamanlarda paylaşım gelmemiştir. Belki tepki çekmek ve paylaşım kazanmak amaçlı yapılan troll paylaşımlar, belki de kişilerin gerçek düşüncelerini içeren bu paylaşımların sayıları, ülkemizde kadına yönelik şiddet düşüncesinin yayılmasına ve gerçek hayatta yaşanan psikolojik ve fiziksel şiddet olaylarına karşı duyarsızlaşmaya sebep olmaktadır.

Özellikle kadınların ısrar eden erkeklere bayılması genellemesi ve hiç kadın şair olmaması önermesi başlıklarının yıllar içinde ve özellikle de son yıllarda tekrar çok sayıda kadın düşmanı paylaşım alması, ülkemizde artışta olan kadına yönelik şiddet olayları ile paralellik göstermektedir. Bu bulgu gösteriyor ki, ülkemizde kadın düşmanı söylemlerin ve kalıp yargıların yayılması, sosyal medyada yapılan otomatik tespit çalışmaları ile belirlenip engellenmesi gereken kritik bir konu haline gelmiştir. Dolayısı ile bu makalede anlatılan çalışma, bu açıdan büyük önem taşımaktadır.



Şekil 3. Her başlıkta kadına yönelik şiddet içeren paylaşımların zaman içerisindeki değişimleri (Time-domain changes per heading among the posts that support violence against women in social media)

4.2. Konu modelleme sonuçları (Topic modeling results)

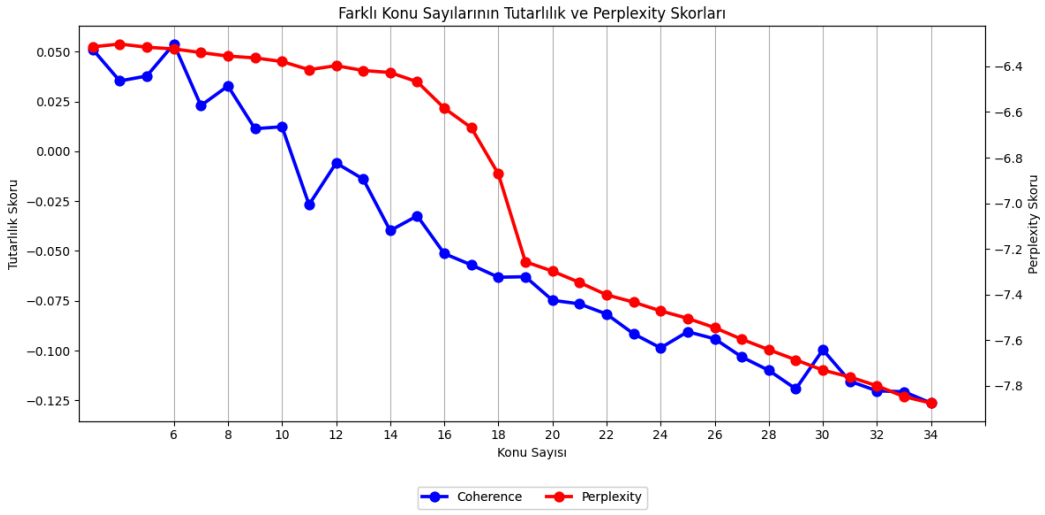
Seçilen başlıklarda tartışılan konuları makine öğrenmesi ile otomatik tespit için LDA yöntemi seçilmiştir. Kullanıcıdan konu sayısını istemesi probleminin üstesinden gelinmesi için öncelikle 2 ile 35 arasındaki her bir olası konu değeri ile LDA sonuçları elde edilmiş, daha sonra her sonucun tutarlılık ve karmaşıklık skorları ölçülmüştür. Deneyler ve skorlar Şekil 4'te görülmektedir.

Konu modelleme çalışmalarında tespit edilen konulardaki tutarlılık skorunun yüksek olması idealdir. Dolayısı ile Şekil 4'teki sonuçlara göre 24 konu ve sonrasında çok düşük değerlere sahip olan tutarlılık skorları, bu değer ve üstünün veri setinde var olan konulardan daha büyük olduğunu göstermektedir. Fakat tek bir skor değerine göre konu sayısı belirlemek, skorlama yönteminin hata yapabilme ihtimali düşünüldüğünde doğru olmayacaktır. Bu nedenle ikinci metrik olarak karmaşıklık skoru değerlendirilmiştir ve bu değerin düşük olması idealdir. Şekil 4'e göre hem tutarlılık değerinin yüksek hem karmaşıklık değerinin düşük olduğu en uygun nokta konu sayısının 19 olarak uygulandığı nokta olarak görünmektedir. Bu noktada, veri setinin toplam 10 adet başlıktan oluştuğu ve her başlıkta kadına yönelik şiddeti destekleyen ve buna karşı çıkan görüşlerin birbirinden ayrılacağı göz önüne alınırsa 20'ye yakın LDA konusu tespit edilmesi, bu yöntemin başarısını göstermektedir. En uygun konu sayısı 19 olarak tespit edildikten sonra, bu konu sayısı ile LDA deneyi tekrarlanıp, sonuçlar üzerine analizler gerçekleştirilmiştir.

19 adet farklı düşünce grubunu daha iyi analiz edebilmek için iki aşamalı çalışma yapılmıştır. İlk aşamada 19 farklı konunun tespitinde büyük rol oynamış kelimeler ağırlıkları ile görselleştirilmişlerdir. İkinci aşamada ise, konular ile veri setindeki Ekşi Sözlük başlıklarının kesişim yüzdeleri hesaplanmıştır.

Her konunun kelimeleri ve LDA ile hesaplanan ağırlıkları Şekil 5'te görülmektedir ve bu sonuçlardan çeşitli tespitler yapmak mümkündür. Örneğin LDA'nın tespit ettiği ilk konu, "trip, ego, mutsuz, yemek" kelimelerini yüksek ağırlıkla içermektedir. Bu kelimeler ile kelime bulutlarının benzerliği sebebiyle "YETENEK" başlığının bu konuda yer aldığı düşünülebilir. İkinci konuda ise "dayı, olay, sokak" kelimelerine göre "DAYI" başlığının modellendiği düşünülebilir. Üçüncü konu da "öl, şiddet, yardım, olay" kelimelerinin varlığı "DAYI" başlığını kapsamaya devam ediyor denilebilir. Benzer şekilde dördüncü konu da "yalan, dedikodu" kelimeleri ile ilk konuya denk gelen "YETENEK" başlığını kapsıyor olabilir. Beşinci konuda ise "ısrar, taciz, naz" kelimelerinden "ISRAR" başlığını kapsadığı anlaşılmaktadır. Altıncı konuya denk gelen "hayat, zaman, çocuk, sosyal, dünya" kelimelerinden net olarak hangi başlığa denk geldiği tahmin edilememektedir. Yedinci konuda ise "akın, gülten" isimlerinden şairlerle ilgili başlığı modellediği görülmektedir. Sekizinci konu da altıncı konu gibi, "keskin, çalış, iş, dönem, anne" kelimelerinden oluşmakta ve başlığı tespit edilememektedir. Bu konuların hangi başlık veya başlıkları kapsadığını tespit edebilmek için ikinci aşamaya başvurulacaktır.

Dokuz numaralı konuda "gülümse, yaş, iş, müdür, hayat, bekar" kelimelerinden "DİĞER" başlığı kapsamına denk gelen "40 yaşında problemli bekar müdür kadın" başlığını kapsadığı anlaşılmaktadır. Onuncu konudaki "satır, müdahale, polis, olay" kelimeleri tekrar "DAYI" başlığının modellendiğini göstermektedir. Devamındaki konu da "türk, kız, psikoloji, bozuk, güzel" kelimelerinden "GÜZEL" başlığını işaret etmektedir. On ikinci konuda ise "gül, yüz, ruh, yasa, cahil, ilişki, saçma" kelimelerinden "DİĞER" başlığı altında toplanmış olan "türk kızları neden gülümsemiyor sorunsalı" başlığını modelliyor olabilir. On üçüncü konuda da başlık tespit edilememektedir. On dördüncü konuda "ısrar" kelimesi "ISRAR" başlığını işaret edebilir. Sonraki başlık ise edebiyatla ilgili "ŞAİR" başlığı göstermektedir.



Şekil 4. LDA yönteminde en uygun konu sayısını bulmak için gerçekleştirilen deneylerin tutarlılık ve karışıklık perplexity skorları (Coherence and perplexity scores obtained during the LDA experiments with different numbers of topics)

On altıncı başlıkta "aşk, sev, aşık, acı, duygusal, seviş" kelimeleri ile hangi başlığı modellediği anlaşılamasa da ilişkileri anlatan paylaşımların konularını tespit edildiği anlaşılabilir. On yedinci konu da kadın şairlerle ilgili başlığı göstermektedir. Son iki konuda ise yine genel olarak ilişkilerin ifade edildiği paylaşımlar modellenmiştir.

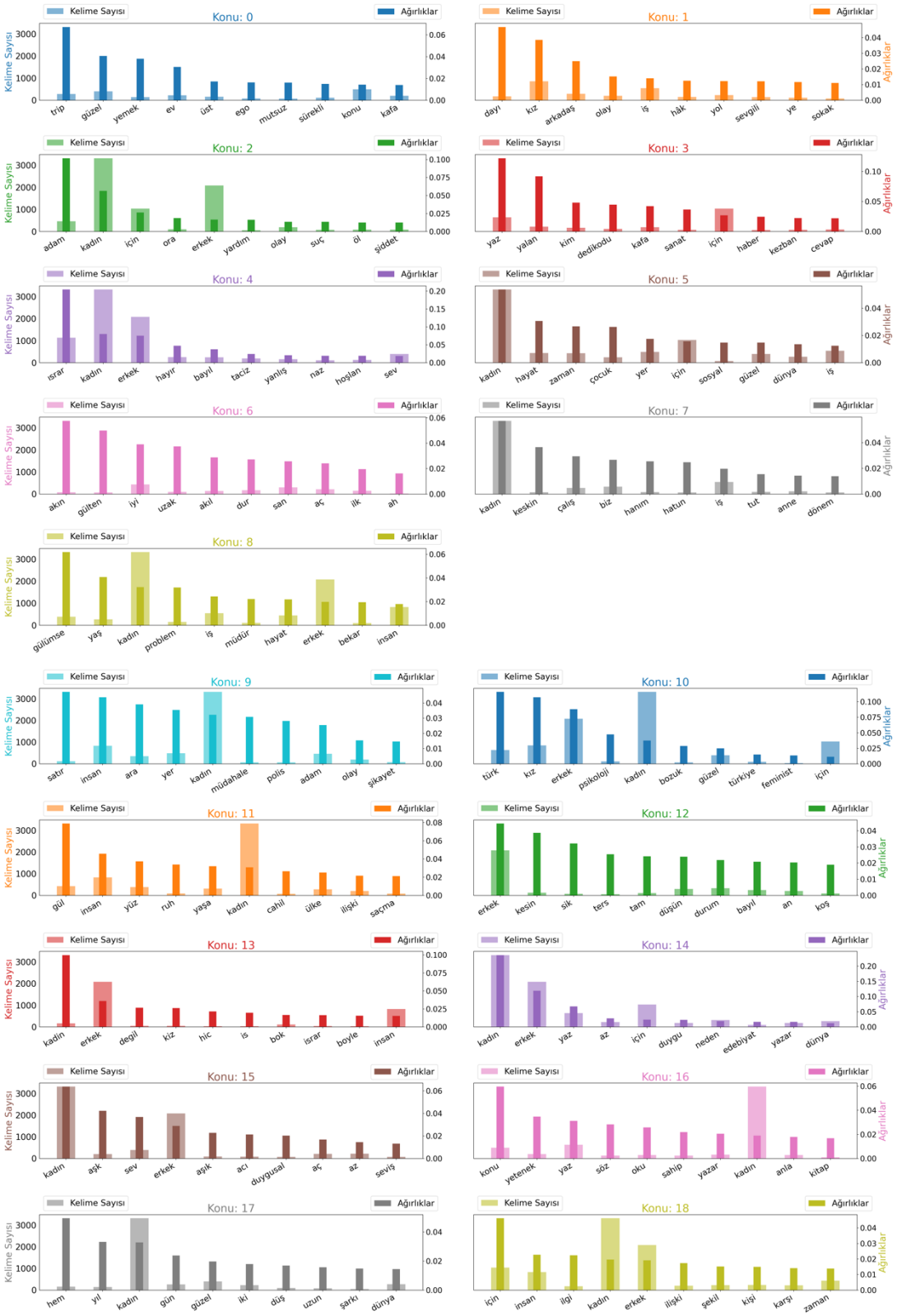
Kelimelerden elde edilmiş konular arası benzerlikler göz önüne alındığında, tespit edilen 19 konunun ne kadar çakıştığının daha detaylı analiz edilmesi gerekmektedir. Bu nedenle LDA sonuçlarının analizinin ikinci aşaması paylaşımların denk geldiği LDA konuları ile o paylaşımların ait oldukları başlıklar arasındaki kesişim yüzdelerini hesaplamıştır. Sonuçlar Tablo 2'de görülmektedir.

Tablo 2'ye göre 0 numaralı ilk konunun %36,86'sı "YETENEK" başlığına ait paylaşımlardır. Aynı sonuçlara göre 3 numaralı konunun %52,88'i de aynı başlığa ait paylaşımlardır. Diğer yandan konu 0'ın %49,48'i "ŞAİR" başlığına ait paylaşımları içermektedir. Fakat, 1 ve 2 numaralı konuların "DAYI" başlığını modellediği, ilk aşama tespiti yanıltmıştır, çünkü daha önceki kelime ağırlıklarından başlığı tespit edilemeyen 7 numaralı konunun büyük oranı bu başlığa ait çıkmıştır. Tabloya göre bu konular çoğunlukla "YETENEK'' başlığındaki paylaşımları içermektedir.

Tablo 2'deki yüzdelerden daha fazla bulgu elde etmek mümkün olsa da sayfa limiti nedeniyle bu aşama okuyucuya bırakılmıştır.

Tablo 2. Konulara denk gelen paylaşımların başlıklara aidiyet yüzdeleri (The percentage match between each heading and topic in terms of normalized number of posts per heading)

Konu	DAYI	GÜZEL	ŞAİR	ISRAR	YETENEK
0	0	2,58	**49,48**	3,87	**36,86**
1	1,39	2,78	6,15	6,55	**16,47**
2	2,44	5,16	**26,07**	**17,26**	**23,93**
3	0,96	9,62	13,46	14,42	**52,88**
4	1,41	4,93	4,93	**32,39**	**37,32**
5	1,69	2,25	**53,93**	5,62	**27,53**
6	3,00	5,00	**19,00**	**37,50**	**23,50**
7	**32,99**	4,16	7,43	**17,53**	**21,25**
8	0,95	0,95	10,48	8,57	**34,29**
9	0	2,90	6,28	7,73	**59,90**
10	1,95	**20,78**	1,95	**18,18**	**50,00**
11	2,59	5,70	**15,03**	9,84	46,63
12	1,18	3,53	11,76	10,00	**55,29**
13	0,47	3,30	5,19	**26,89**	**58,96**
14	0,40	2,82	**47,98**	10,48	**25,81**
15	0,61	0,61	**20,25**	14,11	**56,44**
16	0,50	7,46	14,93	6,47	**62,19**
17	0,19	2,27	11,17	**67,23**	13,64
18	4,23	11,27	11,27	16,90	**42,25**



Şekil 5. LDA yöntemi ile tespit edilmiş her konuya denk gelen kelimelerin ağırlıkları (Weights of the words per topic according to the LDA method)

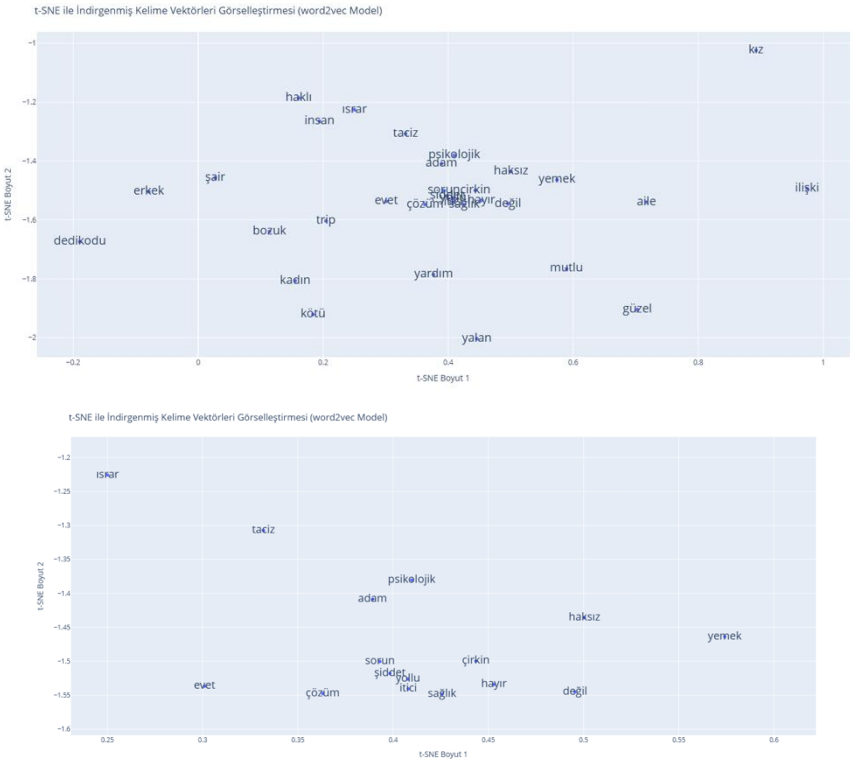
4.3. Kelime gömmesi sonuçları (Word embedding results)

Türkçe dilinde kadına yönelik şiddet düşüncesini anlamanın yöntemlerinden biri son yıllarda başarısını ispatlamış olan derin öğrenme tabanlı kelime gömme

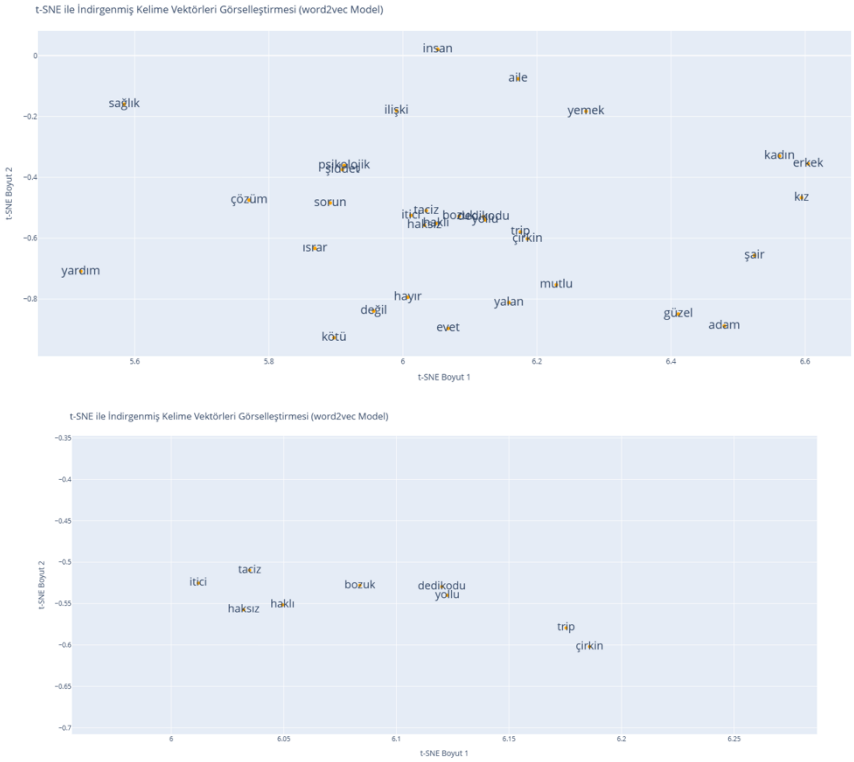
yöntemleridir. Türkçe dilinde yazılmış metinlerden eğitilmiş Word2Vec modeli, Türkçe kelimeler arasındaki anlamsal ilişkileri yakalayabilecek ve bu ilişkilerin görselleştirilmesi ile de kelimelerin birbirlerine yakınlığı veya uzaklığı sayesinde fikir edinilebilecektir. Birbirine yakın kelimeler, eğitim verisinde bu kelimelerin anlamsal olarak benzer bağlamlarda kullanıldıkları anlamına gelir. Uzak kelimelerde ise ilişki yoktur. Bu amaçla, Türkçe Wikipedi metinleri üzerinde eğitilmiş hazır bir Word2Vec modeli kullanılarak (Köksal, 2018), 32 adet Türkçe kadına şiddetle ilgili olabilecek kelimeler TSNE (t-Distributed Stochastic Neighbor Embedding) yöntemi ile görsellenmiştir (Srinivasa-Desikan, 2018). Şekil 7'de görülen bu kelimeler ve görselin en yoğun olduğu kısmın büyütülmüş halinden çeşitli bulgular tespit edilmiştir. Örneğin "şiddet, yollu, itici, sorun" kelimeleri, bu modelde bağlamsal olarak bir arada, birbirine yakın bulunmaktadır. Cinsiyet belirten kelimelere bakıldığında ise "erkek" kelimesi "şair" kelimesine ve "dedikodu" kelimesine yakındır. "Kadın" kelimesi ise "bozuk, yalan, trip" gibi kelimelere daha yakında bulunmaktadır.

Wikipedia verilerinde eğitilmiş Word2Vec, sosyal medyada yer alan düşünceler ve Türkçe yazı tarzını modelleyemeyecektir. Sosyal medyadaki kelimeler arası bağlamı yakalayabilmek için, hazır Word2Vec modelinin üzerine sadece kadına yönelik şiddeti onaylayan paylaşımlar eklenerek yeniden eğitildiğinde var olan kelime gömme vektörleri ve aralarındaki ilişkiler yenilenmiştir. Bu yenilenme sonrasında aynı 32 kelimenin ilişkisi görsellendiğinde Şekil 8 elde edilmiştir. Bu yeniden eğitme aşamasının faydası ise şekilde görüldüğü üzere "kadın, erkek, kız" kelimelerinin bir araya gelmesi olmuştur. Böylece Türkçe dilinde cinsiyetleri öğrenememiş olan hazır Word2Vec modeli, kısa bir eğitimle bu eksiğini gidermiştir.

Şekil 8'de görülen "şair" kelimesi, hem bir arada bulunan cinsiyet belirten kelimelere hem de "adam" kelimesine eşit mesafededir. "Psikolojik" ve "şiddet" kelimelerinin birbirlerine yakınlığı da bu iki kelimenin veri setinde sıklıkla aynı konularda kullanıldığını gösterir. Orta kısımda üst üste gelmiş kelimelere odaklanıldığında ise "trip" ve "çirkin" kelimelerinin, "dedikodu" ve "yollu" kelimelerinin, "taciz" ve "itici" kelimelerinin birbirlerine yakınlığı görülmektedir. Kadına yönelik şiddeti olumlayan paylaşımlarla eğitilmiş modelin gösterdiği bu bulgular, paylaşımlardaki bu alakasız kelime ikililerinin nasıl aynı konularda kullanıldığını, Türkçe dilinde bu kelimelerin kadınları aşağılamak için tercih edildiğini göstermiştir.



Şekil 7. Türkçe Wikipedi verileri ile eğitilmiş Word2Vec modelindeki 32 kelimenin ilişkilerinin görselleştirilmesi ve bu görselin en yoğun olduğu bölgenin yakınlaştırılmışı (Visualization of the relationships of 32 words in the Word2Vec model trained with Turkish Wikipedia data and zooming in on the region where the visual is most dense)



Şekil 8. Türkçe Word2Vec modelinin sosyal medya verileriyle yeniden eğitilmiş halindeki 32 kelime ilişkisi ve görselin en yoğun olduğu bölgenin yakınlaştırılmışı (32 words' embedding relations in Turkish Word2Vec model retrained with social media data and a zoom of the region where the image is most intense)

4.4. Sınıflandırma sonuçları (Classification results)

Beşli çapraz doğrulama (cross-validation) tekniği ile veri setindeki işaretli bütün paylaşımlarda kadına yönelik şiddet içerenleri otomatik tespit edecek bir model geliştirmek için beş farklı makine öğrenmesi yöntemi, kelime çantası öznitelikleri ile eğitilip test edilmiştir. Beşli çapraz doğrulamanın ortalama ve standart sapma değerleri Tablo 3'te görülmektedir.

Tablo 3. Makine öğrenmesi yöntemleri ve 5'li çapraz doğrulamayla elde edilmiş Türkçe kadına yönelik şiddeti tespit sonuçları (Turkish violence against women classification results obtained from machine learning and 5-fold cross-validation)

Yöntem	Hassasiyet	Geri çağırma	F1	AUC
LR	**67,08** ±1,02	66,09 ±0,87	**68,4** ±2,06	**76,05** ±2,07
GB	53,63 ±2,44	**77,73** ±6,00	53,18 ±1,41	71,12 ±1,68
L-SVM	48,48 ±0,92	**72,05** ±25,31	50,45 ±0,45	71,69 ±2,29
NB	60,11 ±1,36	59,15 ±1,26	61,69 ±1,58	61,69 ±1,58
RF	53,39 ±0,49	56,00 ±0,41	**66,00** ±1,26	**73,66** ±1,22

Sınıflandırma sonuçlarına göre beş yöntem arasından iki tanesi diğerlerinden daha başarılı olmuştur. Bunlar basit bir makine öğrenmesi yöntemi olan LR ve çok fazla karar ağacının ortalamasını hesaplayarak sınıflandırma yapan RF yöntemleridir. Düşük standart sapma değerleri de göstermektedir ki beşli çapraz doğrulamada farklı test verileri denense de her seferinde benzer sonuçlar elde edilmiştir. Bu bulgular, Türkçe dilinde kadına yönelik şiddet düşüncesinin sosyal medya verilerinden %76 başarı ile tespit edilebildiğini ve düşük standart sapma ile bu sonuçların genellenebilir olduğunu göstermektedir.

5. Sonuçlar (Conclusions)

Sosyal medya platformları, kullanıcıların düşüncelerini ve duygularını ifade etmeleri için önemli bir mecra olsa da kadına yönelik şiddet düşüncesinin Türkiye çapında yayılmasına sebep olmakta, halihazırda var olan problemi körüklemektedir. Bu makalede anlatılan çalışmada elde edilen bulgular, Türkçe sosyal medya verilerinden kadına yönelik şiddet düşüncesinin detayları göstermiş, daha da önemlisi, makine öğrenmesi yöntemleri ile %76 başarı oranı ile kadına yönelik şiddet ve nefret içeren paylaşımların otomatik tespit edilebildiğini göstermiştir.

Elde edilen bulgular sayesinde Türkçe dilinde kadın, çocuk ve çeşitli gruplara yönelik şiddet düşüncesini tespit için veriler toplanacak ve yapay zekâ içeren

sistemler geliştirilebilecektir. Sağlıklı bir toplum olmanın önünde engel teşkil ederek şiddeti onaylayan ve öven paylaşımların engellenmesi, hem sosyal medya ile büyüyen nesillerin şiddete meyilli yetişmesini önleyecek, hem de toplumdaki şiddet eğilimi azalacaktır.

Çalışmanın ilerleyen aşamalarında daha karmaşık derin öğrenme ve dönüştürücü yöntemlerinden faydalanılarak daha yüksek kadına yönelik şiddeti tespit performansı elde etme üzerine deneyler gerçekleştirilecektir.

Teşekkür (Acknowledgment)

Bu çalışma TÜBİTAK 2209A programı kapsamında desteklenmiştir. Proje numarası: 1919B012303496.

Kaynaklar (References)

Alshehri, A., Nagoudi, E. M. B., Abdul-Mageed, M., 2020. Understanding and detecting dangerous speech on social media. *arXiv:2005.06608*.

Ashraf, N., Mustafa, R., Sidorov, G., Gelbukh, A., 2020. Classification of individual and group violence threats in online discussions. *Companion Proceedings of the Web Conference*, 629–633.

Aytaç, S., Eteman, F. S., Aydın, G. C., Reçber, B., Sezen, H. K., 2016. Kadına yönelik şiddetin dünü, bugünü, yarını: Kestirim tabanlı bir araştırma. *Istanbul Journal of Sociological Studies*, (54), 275–297.

Bayram, U., 2022. Uncovering the impacts of the pandemic by analyzing COVID-19 related news articles using machine learning and network analysis. *Journal of Information Technologies*, 15(2), 209–220.

Bayram, U., Benhiba, L., 2021. Determining a person's suicide risk based on short-term tweet history. *Proceedings of the Seventh Workshop on Computational Linguistics and Clinical Psychology: Improving Access*, 81–86.

Bayram, U., Benhiba, L., 2022. Emotionally-informed models for detecting moments of change and suicide risk levels in longitudinal social media data. In *Proceedings of the Eighth Workshop on Computational Linguistics and Clinical Psychology*, 219–225.

Bayram, U., Pestian, J., Santel, D., Minai, A. A., 2019. What's in a word? Detecting party affiliation from word usage in congressional speeches. *International Joint Conference on Neural Networks (IJCNN)*, 1–8. IEEE.

Blei, D. M., Ng, A. Y., Jordan, M. I., 2003. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3, 993–1022.

Castorena, C. M., Abundez, I. M., Alejo, R., Granda-Gutiérrez, E. E., Rendón, E., Villegas, O., 2021. Detection of gender-based violence in Twitter messages using deep neural networks. *Mathematics*, 9(8), 807.

Cihan, Ü., Karakaya, H., 2017. Kadın-erkek kavramları bağlamında şiddet ve şiddetle mücadelede sosyal hizmetin rolü. *Bolu Abant İzzet Baysal Üniversitesi Sosyal Bilimler Enstitüsü Dergisi*, 17(4), 297–324.

Deng, K., Bol, P. K., Li, K. J., Liu, J. S., 2016. Unsupervised analysis of domain-specific Chinese texts. *Proceedings of the National Academy of Sciences*, 113(22), 6154–6159.

Desai, A., Kalaskar, S., Kumbhar, O., Dhumal, R., 2021. Detecting cyberbullying on social media using machine learning. *ITM Web of Conferences*, 40, 03038. EDP Sciences.

Ensari, T., Ensari, B., Dağtekin, M., 2022. Violence detection with machine learning: a sociodemographic approach. *Journal of European Science and Technology*, (44), 104–107.

González, G. A. R., Cantu-Ortiz, F. J., 2021. Sentiment analysis and unsupervised learning approach for digital violence against women: the Monterrey case. *4th International Conference on Information and Computer Technologies (ICICT)*, 18–26. IEEE.

Guo, Y., Kim, S., Warren, E., Yang, Y.-C., Lakamana, S., Sarker, A., 2023. Automatically detecting victims of intimate partner violence from social media. *AMIA Summits on Translational Science Proceedings*, 254.

Hassan, N., Poudel, A., Hale, J., Hubacek, C., Huq, K. T., Santu, S. K. K., Ahmed, S. I., 2020. A step towards automated sexual violence reporting tracking. *Proceedings of the International AAAI Conference on Web and Social Media*, 14, 250–259.

Kapil, P., Ekbal, A., Das, D., 2020. A study of deep learning approaches for hate speech detection on social media. *arXiv:2005.14690*.

Köksal, A., 2018. Turkish-Word2Vec. GitHub. https://github.com/akoksal/Turkish-Word2Vec

Mikolov, T., 2013a. Efficient estimation of word representations in vector space. *arXiv:1301.3781*.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., Dean, J., 2013b. Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems*, 26.

Okkalı, E., Atamtürk, H., Kilimci, Z. H., 2021. Assessing public response to violence against women in Turkey via Twitter using topic modeling. *Kocaeli Journal of Science and Engineering*, 4(2), 103–112.

Oriola, O., Kotzé, E., 2020. Evaluation of machine learning techniques for detecting aggressive and hate speech in South African tweets. *IEEE Access*, 8, 21496–21509.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ..., Duchesnay, É., 2011. Scikit-learn: machine learning in python. *The Journal of Machine Learning Research*, 12, 2825–2830.

Pestian, J., Santel, D., Sorter, M., Bayram, U., Connolly, B., Glauser, T., DelBello, M., Tamang, S., Cohen, K., 2020. A machine learning approach to detect suicidal language. *Suicide and Life-Threatening Behavior*, 50(5), 939–947.

Perera, A., Fernando, P., 2021. Accurate detection and prevention of cyberbullying on social media. *Procedia Computer Science*, 181, 605–611.

Pitsilis, G. K., Ramampiaro, H., Langseth, H., 2018. Using deep learning to detect aggressive language in tweets. *arXiv:1801.04433*.

Reynolds, K., Kontostathis, A., Edwards, L., 2011. Using machine learning to detect cyberbullying. *10th International Conference on Machine Learning and Applications and Workshops*, 2, 241–244. IEEE.

Rodríguez, D. A., Díaz-Ramírez, A., Miranda-Vega, J. E., Trujillo, L., Mejía-Alvarez, P., 2021. A systematic review

of computer science solutions to address violence against women and children. *IEEE Access*, 9, 114622–114639.

Sen, S., Bolsoy, N., 2017. Violence against women: prevalence and risk factors in the example of Turkey. *BMC Women's Health*, 17, 1–9.

Soykan, L., Karsak, C., Elkahlout, I. D., Aytan, B., 2022. Comparison of machine learning techniques for offensive language detection in Turkish. *Proceedings of the Second International Workshop on Resources and Techniques for User Information in Abusive Language Analysis*, 16–24.

Srinivasa-Desikan, B., 2018. Natural language processing and computational linguistics: a practical guide to text analysis with python, gensim, spacy, and keras. Packt Publishing Ltd.

Şahi, H., Kılıç, Y., Sağlam, R. B., 2018. Automatic detection of hate speech against women on Twitter. *3rd International Conference on Computer Science and Engineering (UBMK)*, 533–536. IEEE.

Tommasel, A., Rodriguez, J. M., Godoy, D., 2018. Text-based aggression detection: with deep learning. *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*, 177–187. Association for Computational Linguistics.

Trinh Ha, P., D'Silva, R., Chen, E., Koyutürk, M., Karakurt, G., 2022. Detecting intimate partner violence from free text descriptions on social media. *Journal of Computational Social Science*, 5(2), 1207–1233.

Türkiye İstatistik Kurumu, 2023. İstatistiklerle kadın. https://data.tuik.gov.tr/Bulten/Index?. Erişim tarihi: 28 Temmuz 2024.

World Health Organization, 2021. Global and regional estimates of violence against women: prevalence and health effects of intimate partner violence and non-partner sexual violence. W. H. O.

Zhao, R., Zhou, A., Mao, K., 2016. Automatic detection of bullying in social networks. *Proceedings of the 17th International Conference on Distributed Computing and Networking*, 1–6.

Zhou, C., Li, Q., Li, C., Yu, J., Liu, Y., Wang, G., ..., Sun, L., 2024. A comprehensive survey on pretrained foundation models: a history from bert to chatgpt. *International Journal of Machine Learning and Cybernetics*, 1–65.