

NEFRET SÖYLEMİ

Güncel sorunlar ve
tartışmalar



HDV
YAYINLARI

HRANT DİNK VAKFI

Hrant Dink'in, 19 Ocak 2007'de, gazetesi Agos'un önünde öldürülmesinden sonra, benzer acıların yeniden yaşanmaması için; onun daha adil ve özgür bir dünyaya yönelik hayallerini, dilini ve yüreğini yaşatmak amacıyla kuruldu. Etnik, dinî, kültürel ve cinsel tüm farklılıklarıyla herkes için demokrasi ve insan hakları talebi, vakfın temel ilkesidir.

Vakıf, ifade özgürlüğünün alabildiğine kullanıldığı, tüm farklılıkların teşvik edilip yaşandığı, yaşatıldığı ve çoğaltıldığı, geçmişe ve günümüze bakışımızda vicdanın ağır bastığı bir Türkiye ve dünya için çalışır. Hrant Dink Vakfı olarak 'uğruna yaşadığımız davamız', diyalog, barış, empati kültürünün hâkim olduğu bir gelecektir.

NEFRET SÖYLEMİ: GÜNCEL SORUNLAR VE TARTIŞMALAR

ISBN 978-625-97168-1-7



editörler

İdil Engindeniz, Zeynep Arslan

çevirmen

Defne Orhun

yayın koordinatörleri

Başak Can, Elif Erol, Buket Kapısız, Yasemin Korkmaz,
Pelin Önal, Tuğba Özsoy, Elif Yazarbaşı

tasarım

Yasemen Cemre Gürbüz

grafik uygulama

Selin Uluer

baskı

Sena Ofset Ambalaj Sanayi ve Ticaret. Ltd. Şti.
Yakuplu Mh. 194. Sk. 3. Matbaacılar Sit. N:1 D:465
Beylikdüzü İstanbul/Türkiye
T: (212) 613 38 46

İstanbul, Mart 2025



© Hrant Dink Vakfı Yayınları

Anarad Hıçutyun Binası Papa Roncalli Sokak No: 128
Harbiye, 34373 Şişli, İstanbul
T: 0212 240 33 61
F: 0212 240 33 94
info@hrantdink.org
www.hrantdink.org

Dijital Teknolojileri Kullanarak Nefret Söylemi ve Ayrımcılıkla Mücadele projesi, Avrupa Birliği ve Friedrich Naumann Vakfı tarafından desteklenmektedir. Raporda yer alan görüşler destekçi grupların görüşlerini yansıtmamaktadır.



FRIEDRICH NAUMANN
FOUNDATION For Freedom.
Türkiye



HRANT DINK VAKFI
HRANT DINK FOUNDATION
ZİPİLA SÖĞE ZINTIPI



Sabancı
Universitesi

NEFRET SÖYLEMİ

Güncel sorunlar ve
tartışmalar



HDV
YAYINLARI

İÇİNDEKİLER

Önsöz	7
-------------	---

NEFRET SÖYLEMİ VE DİĞER İLGİLİ KAVRAMLARIN TESPİTİ VE KATEGORİLEŞTİRİLMESİ	11
---	-----------

Nefret söyleminin farklı tanımları	13
---	-----------

Yasemin İnceoğlu

Ayrımcılık ile nefret söylemi arasındaki ilişki	17
--	-----------

Ayşecan Terzioğlu

Tehlikeli söylem: Zarar veren dili ayırt edebilmeye yönelik bir çerçeve	19
--	-----------

Susan Benesch

Saldırgan dil: Değişkenlik ve güçlükler	23
--	-----------

Tommaso Caselli

Kesişimsellik ve nefret söylemi:

Türkiye'deki LGBTİ+ yurttaş ve göçmenlerin durumu	27
--	-----------

Eser Selen

NEFRET SÖYLEMLERİNİN TESPİTİ: POLİTİKALAR, YÖNTEMLER VE GÜÇLÜKLER	35
--	-----------

Devlet politikaları, düzenlemeler ve politika analizi	37
--	-----------

Tirşe Erbaysal Filibeli

Sosyal medya platformlarının nefret söylemine ilişkin mevcut politikaları	39
--	-----------

Tirşe Erbaysal Filibeli

Nefret söylemi tanımlarındaki çeşitliliğin veri setleri üzerindeki etkileri	51
--	-----------

Arzucan Özgür

Nefret söyleminde insan eliyle ayıklama: Faydalar ve güçlükler	53
---	-----------

Ayşecan Terzioğlu, Didar Akar

Nefret söylemi tespit modellerinde etiketleyicinin kimliğinin önemi	55
--	-----------

Claudia von Vacano

Doğal Dil İşleme ile nefret söylemi tespiti için mevcut kategoriler	59
Roser Morante	
Araştırmacıların veriye erişebilirliği	65
Houda Bouamor, Onur Varol, Berrin Yanıkoğlu	
Nefret söyleminin otomatik tespitine yönelik araçlar geliştirilmesi konusunda mevcut sorunlar ve olası çözümler	69
Tommaso Caselli	
Sosyal medya şirketlerinin yapay zekâ teknolojilerini etik şekilde uygulaması	75
Zeynep Özarslan	
NEFRET SÖYLEMİYLE MÜCADELE: POLİTİKA VE DÜZENLEMELERİN ÖTESİ	80
Nefret söylemini zayıflatmak için karşıt söylem yaklaşımları	83
Susan Benesch	
Aktivizm aracılığıyla nefret söylemiyle mücadele: İki başarılı örnek	85
İlayda Ece Ova	
Pozitif etkileşimi teşvik etmek için alternatif yaklaşımlar	91
Peter Adams, Metin V. Bayrak, Alex Mahadevan	
Sonuç yerine	101

Bu rapor, dijital ortamlarda artan nefret söylemi, dezenformasyon ve zararlı söylemlerle mücadele etmeyi amaçlayan bir yıllık ortak çalışmanın sonucudur. Bu girişim, Hrant Dink Vakfı tarafından Avrupa Birliği ve Friedrich Naumann Vakfı'nın finansal desteğiyle yürütülen **Dijital Teknolojileri Kullanarak Nefret Söylemi ve Ayrımcılıkla Mücadele**¹ başlıklı proje kapsamında ortaya çıktı. Boğaziçi Üniversitesi ile Sabancı Üniversitesi ortaklığında yürütülen proje, dijital ortamlarda nefret söylemi ve ayrımcılık ile mücadeleye yönelik disiplinlerarası bir çaba başlatmayı amaçlıyor.

Bu rapor, hayatı boyunca diyalog ve karşılıklı anlayışı savunan Ermeni gazeteci Hrant Dink'e yönelik suikastın ardından 2007 yılında kurulan Hrant Dink Vakfı'nın süregiden çalışmalarının bir parçası olarak hazırlandı. Hrant Dink medyada nefret söyleminin bizzat hedefi olmuş ve bu söylemler suikastına uzanan süreçte önemli bir rol oynamıştı. Vakıf, diyalogu teşvik ederek, barışı destekleyerek ve ayrımcılığa katkıda bulunan toplumsal sorunlarla mücadele ederek Hrant Dink'in mirasını sürdürmek amacıyla kuruldu. Bu ilkeler doğrultusunda bilhassa nefret söylemi ve ayrımcı dil konusuna eğilen Vakıf, çeşitli girişimler yoluyla ayrımcılıkla mücadele etmek ve bir arada yaşamı güçlendirmek üzere çalışmalar yürütüyor.

2009 yılından beri yürütülen **Medyada Nefret Söyleminin İzlenmesi**² projesi Vakfın Türkiye'de nefret söylemi ve ayrımcılık konusunu ele almak üzere inşa ettiği köşe taşlarından biridir. Bu projenin amacı medya üzerinde sivil denetimin önemini gözeterek, kimlik özellikleri nedeniyle hedef alınan kişi ve gruplara yönelik ayrımcı ve ötekileştirici söylemlerin tespiti amacıyla Türkiye'deki yazılı basının taranıp izlenmesini amaçlıyor. Ulusal ve yerel yazılı basın sistematik bir şekilde izlenip raporlanarak, bu konuda farkındalık uyandırmak ve kapsayıcı söylemleri teşvik etmek hedefleniyor.

1 **Hrant Dink Vakfı.** (2022). *Dijital Teknolojileri Kullanarak Nefret Söylemi ve Ayrımcılıkla Mücadele*. <https://hrantdink.org/tr/asulis/faaliyetler/projeler/dijital-teknolojileri-kullanarak-nefret-soylemi-ve-ayrimcilikla-mucadele>

2 **Hrant Dink Vakfı.** (2016). *Medyada Nefret Söyleminin İzlenmesi*. <https://hrantdink.org/tr/asulis/faaliyetler/projeler/medyada-nefret-soylemi>

Vakıf 2016 yılında, Medyada Nefret Söyleminin İzlenmesi projesi yoluyla geliştirilen uzmanlığı bir adım daha ileri taşıyarak, Türkiye’de söylem çalışmaları alanına özgü ilk araştırma merkezi olan **ASULİS Dil, Diyalog, Demokrasi Laboratuvarını**³ kurdu. Bu merkezin adı Ermenicede *asel* (söylemek) ve *İsel* (dinlemek) fiillerinin birleşiminden oluşuyor. ASULİS ayrımcılıkla mücadele etmek, söylem üzerine araştırmalar yapmak ve bu alandaki çalışmaları desteklemek üzere kurulmuş ilk sosyal bilimler laboratuvarıdır. Nefret söylemi ve ayrımcı dil konusunda yürütülecek araştırmalar, kavramsal tartışmalar ve faaliyetler için disiplinlerarası bir alan yaratmayı amaçlayan merkez, Vakfın çalışma kapsamını daha da geniş bir alana taşıyor.

2021 yılında ise **Nefret Söylemi Dijital Arşivi**⁴ kullanıma açıldı. Herkesin erişimine açık olan bu dijital arşiv 2009 yılından beri Türkiye’deki yazılı basında yer alan nefret söylemlerinin 10 yıldan uzun bir süre boyunca sistematik bir şekilde taranmasıyla oluşturuldu. Kategorilere ayrılmış binlerce örnek barındıran bu arşiv, araştırmacılara, aktivistlere ve genel anlamda kamuoyuna medyada nefret söylemi biçimlerini ve bunların etkilerini anlamak bakımından kapsamlı bir kaynak sunuyor.

Vakıf, dijital platformların kamusal söylem üzerinde yarattığı etkinin gitgide artmasından hareketle, 2021 yılında çalışmalarını sanal ortamlarda nefret söylemi ve dezenformasyon konusunu ele alacak şekilde genişletti. “Dijital Teknolojileri Kullanarak Nefret Söylemi ve Ayrımcılıkla Mücadele” projesi kapsamında sanal ortamdaki zararlı dil ve anlatıları tespit edip ele almak üzere etkili yöntemlerin araştırılması amaçlanıyordu.

Proje boyunca Türkiye içinden ve dışından 20 uzman –akademisyen, araştırmacı, sivil toplum temsilcisi ve aktivist– bir araya gelerek söylem çalışmaları, bilişim, dilbilim, ve sivil toplum gibi farklı alanlardan oluşan bir ağ kurdular. Disiplinlerarası bir niteliğe sahip olmakla birlikte, nefret söylemi konusunda ortak bir bakış açısına sahip olan grubun her bir üyesi, yürütülen tartışmalara kendi alanındaki uzmanlık ve anlayışı dahil etti. Bu ağ, Mart 2023 ile Şubat 2024 tarihleri arasında düzenli olarak tematik toplantılar kapsamında buluşarak dijital okuryazarlık, sosyal medyada nefret söylemi ve dezenformasyon, dijital aktivizm ve zararlı dil- le mücadeleye yönelik yenilikçi araçlar gibi çeşitli konuları ele aldılar. Bu toplantıların her biri, hem yerel hem de küresel perspektiflerle aydınlatılan bir dayanışma, deneyim paylaşımı ve bilgi alışverişi mekanına dönüştü.

3 **Hrant Dink Vakfı.** (2016) *ASULİS Dil, Diyalog, Demokrasi Laboratuvarı*
<https://hrantdink.org/tr/asulis>

4 **Hrant Dink Vakfı.** (2021). *Hrant Dink Vakfı Arşivi*. Hrant Dink Vakfı.
<https://archive.hrantdink.org/?l=tr>

Bu rapor, söz konusu toplantılarda ortaya çıkan tartışma ve anlayışları belgelemeyi amaçlıyor. İçinde bulunduğumuz dijital çağda nefret söylemi ve dezenformasyonun yarattığı sorun ve sıkıntıları anlayıp konu edinmekle ilgilenen herkes için kapsamlı bir kaynak oluşturmayı hedefliyor. Bu yayın sadece bu ağdaki katılımcıların katkılarına yer vermekle kalmayıp, ağın öğretilerini daha geniş bir kitleye yayarak, aktivistler, sivil toplum kuruluşları, araştırmacılar ve politika yapımında rol oynayan herkes için bir kaynak oluşturma niyeti taşıyor. Rapordaki uzman analizleri ve tavsiyeler bu alandaki acil meselelerin üstesinden gelebilmek için bir kılavuz işlevi görüyor.

Raporun ilk kısmında nefret söyleminin nasıl tespit edilip sınıflandırılacağı incelendikten sonra ayrımcılık, tehlikeli söylem ve saldırgan dil gibi konuyla ilgili kavramlara ilişkin farklı tanımlar ve anlayışlar sunulurken bir yandan da nefret söyleminin hedef aldığı grupların kesişimsel kimlikleri irdeleniyor. İlerleyen bölümlerde ise nefret söylemiyle ilgili politikalar, otomatik tespit modelleri ve nefret söylemiyle mücadele konusunda yasal düzenlemeler dışında geliştirilebilecek stratejiler ele alınıyor. Çeşitli disiplinlere mensup uzmanların katkıları disiplinlerarası bir diyalog yaratarak raporu zenginleştirirken; tartışmaların önemli noktalarını öne çıkartıp okurlara ek bağlam sunan editoryal kutularla tamamlanıyor.

Bu ağ içerisinde yer alan ve raporu kaleme alan uzmanlara, yürüttükleri canlı tartışmalar ve alana yaptıkları değerli katkılar için teşekkürlerimizi sunuyor ve bu yayının dijital ortamda pozitif söylemleri, barışı ve sosyal dayanışmayı geliştirme yönünde süregiden çabaları büyütmesini umuyoruz.

NEFRET SÖYLEMİ VE DİĞER İLGİLİ KAVRAMLARIN TESPİTİ VE KATEGORİLEŞTİRİLMESİ

Ağ üyeleri öncelikle nefret söyleminin tanımından başlayarak, bunu önlemeye yönelik yöntemlere dair bir tartışma başlattılar. İster manuel yöntemlerle ister yapay zekâ aracılığıyla, yani başvurulmuş önleme metodolojisi ne olursa olsun, nefret söylemini ortadan kaldırmak ve önlemek için ne yapılması gerektiğini tespit edip harekete geçmek için önce nefret ve ayrımcılık içeren söylemlerin tanımlanması şarttır. Alanda bu kavramın tanımına ilişkin olarak yürütülen tartışmalar, raporun ilerleyen bölümlerinde tespit, önleme ve müdahale konusunda daha hareketli bir tartışmaya temel hazırlayacak.

Nefret söylemi kavramını tanımlarken, Yasemin İnceoğlu bu tanımdaki sınırlılıklar ve tanım etrafında yürütülen tartışmalar hakkında bir özet sundu.

Nefret söyleminin farklı tanımları

Yasemin İnceoğlu

Öncelikle, “nefret söylemi” terimini oluşturan iki sözcüğü parçalarına ayırarak incelememiz gerek: “nefret” ve “söylem”.

“Nefret” sözcüğünü bir duygu, yaratım, sendrom, hatta –psikoanalitik bir yaklaşımdan hareketle– ruhsal bir anormallik olarak açıklamaya çalışan araştırmacılar var. Ancak bu rapor bağlamında ve Hrant Dink Vakfı tarafından yürütülen diğer çalışmalar açısından bakıldığında bu sözcük bütün bir (siyasi-sosyal) sistemin veya yaklaşımın sonucu olarak değerlendirilir.

“Söylem” ise dilbilimsel açıdan genellikle konuşma kalıpları ve dil kullanımını kapsar; bu bağlamda belirli bir topluluk içerisinde kullanılan lehçeleri ve toplumsal kabul gören ifadeleri de içerir.

Söylem/diskur toplumsal kökenlere dayanan ve dil içinde kodlanan bir ideoloji olarak işlev gösterir. Van Dijk (2008) eleştirel söylem analizi yönteminde, söylemi kontrol etmenin veya üretmenin, toplum üzerinde zihinsel kontrol sağlamak bakımından kaçınılmaz olduğunu savunur. Ancak söylemin öznel ve psikolojik bağlamlardan etkilendiğini teslim etmek gerekir; bu bağlamlar söyleme dahil olan kişileri, onların niyetlerini ve durumsal faktörleri de içerir. Van Dijk (2008, s. 107-108), söylemi kontrol etmenin en temel koşulunun, onun bağlamsal unsurlarını yönetmek olduğunu vurgular.

“Nefret söylemi” hukuk, politika yapımı ve akademi çevrelerinde yaygın olarak kullanılmakla birlikte, bu terimin uluslararası insan hakları hukuku veya ulusal hukuk çerçevesinde genelgeçer bir tanımı bulunmuyor. Bu kavram, özellikle düşünce ve ifade özgürlüğü, ayrımcılık yasağı ve eşitlik gibi temel ilkelerle kesiştiği noktalarda halen bir tartışma konusu olmayı sürdürüyor.

Nefret söylemi genel anlamda, konuşmacının ırk, din, ten rengi, cinsel kimlik, toplumsal cinsiyet kimliği, etnik köken, engellilik durumu veya ulusal köken gibi faktörlere dayanarak belirli bir grubu veya birey kategorisini aşağıladığı, küçük gördüğü veya onlara yönelik nefreti kışkırttığı herhangi bir ifade biçimi olarak anlaşılabilir. Bu noktada, söz konusu kategorilerin zamanla değişip genişlediğini vurgulamak gerekir. Aşağıda bu kavramın temel unsurlarını örnekleyen (her ne kadar

bu amaçla hazırlanmış olmasalar da) bazı uluslararası belgeleri kronolojik sırayla gözden geçireceğiz.

Medeni ve Siyasi Haklara ilişkin Uluslararası Sözleşme'nin (MSHS)⁵ 20(2). maddesi doğrudan bir nefret söylemi tanımı içermemekle birlikte, bu kavrama dair temel bir ilke addedilir. Bu maddeye göre, "Ulusal, ırksal ya da dinsel nefretin ayrımcılık, düşmanlık ya da şiddete kışkırtma şeklini alacak biçimde savunulması yasalarla yasaklanır." 1966 yılında kabul edilmiş olan bu MSHS hükmü, sadece din, ırk veya uyruk temeline dayalı nefreti tanımladığı için nefret söyleminin yakın tarihli tanımlarına nazaran çok daha dar bir kapsama sahiptir.

Avrupa Konseyi (AK) nefret söylemine ilişkin tavsiye kararında bu kavramı şöyle tanımlamıştır: "[I]rkçı nefreti, yabancı düşmanlığını, Yahudi düşmanlığını veya azınlıklara, göçmenlere ve göçmen kökenli insanlara yönelik saldırgan milliyetçilik ve etnomerkezcilik, ayrımcılık ve düşmanlık şeklinde ifade bulan hoşgörüsüzlük de dahil olmak üzere hoşgörüsüzlüğe dayalı başka nefret biçimlerini yayan, kışkırtan, teşvik eden veya meşrulaştıran her türlü ifade biçimi." (1997)

Birleşmiş Milletler İnsan Hakları Yüksek Komiserliği (İHYK) tarafından düzenlenen bölgesel uzman çalıştayları kanalıyla hazırlanan 2012 tarihli Rabat Eylem Planı uzmanlar tarafından benimsenen sonuç ve tavsiyeleri içeriyordu. "Ayrımcılığa, düşmanlığa yahut şiddete tahrik eden ulusal, ırksal ya da dinsel nefret savunuculuğunun yasaklanmasına ilişkin Rabat Eylem Planı [...] nefretin ağırlık derecesini tartmaya yönelik olası unsurlar arasında, savunulan beyanın ya da zararın acımasızlığı veya kastı, sıklığı, niceliği ve ne ölçüde yaygınlaştırılmış olduğu"nu belirtiyor (İHYK, 2012). Ayrıca, ceza hukuku kapsamında suç sayılan ifadeler için aşağıdaki altı kriterden oluşan bir eşik kıstası öneriyor:

- Bağlam
- Konuşmacı
- Niyet
- İçerik ve biçim
- İfade fiilinin derecesi
- Zamansal olarak yakınlık da dahil olmak üzere, zarar oluşturma ihtimali

5 Medeni ve Siyasi Haklara ilişkin Uluslararası Sözleşme, 16 Aralık 1966 tarihli, 2200A (XXI) sayılı Genel Kurul Kararı ile kabul edilmiş ve imza, onay ve katılmaya açılmıştır. 23 Mart 1976 tarihinde yürürlüğe girmiştir. 21 Temmuz 2003 tarihli ve 25175 sayılı Resmî Gazete'de yayımlanmıştır. <https://www.resmigazete.gov.tr/eskiler/2003/07/20030721.htm#2>

Birleşmiş Milletler'in (BM) 2019 tarihli "Nefret Söylemine Yönelik Strateji ve Eylem Planı"nda daha yeni bir tanım yer alıyor. Bu belgede nefret söylemi terimi; konuşma, yazma ve davranış içerisinde bir kişi veya grubun kim olduğuna, başka bir ifadeyle dinine, etnisitesine, uyruğuna, ırkına, rengine, soyuna, toplumsal cinsiyetine veya diğer kimlik unsurlarına dayanarak saldırıda bulunulan ya da aşağılayıcı veya ayrımcı bir dil kullanılan her türlü iletişim olarak algılanıyor (BM, 2019). Bu tanım iki açıdan son derece geniş: Birincisi "veya diğer kimlik faktörleri" ifadesini içererek çok çeşitli gruplara atıfta bulunuyor, ikincisi ise "aşağılayıcı" terimi geniş bir söylem yelpazesini kapsıyor.

Bu rapor boyunca göreceğimiz üzere, nefret söylemine dair genelgeçer bir tanıma ulaşma çabalarının ötesinde, sosyal medya şirketleri, sivil toplum kuruluşları ve araştırmacılar nefret söylemine ilişkin çeşitli tanımlar benimseyebiliyor. Ancak farklı tanımlar ortaya atmadan veya nefret söylemi kavramının yarattığı boşlukları doldurmak için bu alanda yeni beliren kavramlar hakkında konuşmadan önce, meseleyi çevreleyen en önemli tartışmalardan birine değinmek istiyoruz: söylem ile eylem arasındaki ilişki. Yasemin İnceoğlu bu tartışmaya şu iki soruyla alan açıyor:

"Söylem sadece kelimelerin teatisinden mi ibaret ve bu kelimeleri seslendiren insanlar sadece kendi kişisel eylemlerinden mi sorumlu? Yoksa kimi söylemler ayrımcılığı tetikleyen ve nihayetinde şiddeti meşrulaştıran propaganda amaçlı bir unsur olarak mı değerlendirilmeli?"

Bu sorulara yanıt olarak, aşağıdaki yazıda Ayşecan Terzioğlu ayrımcılık ile nefret söylemi arasındaki ilişkiyi izah ediyor.

Ayrımcılık ile nefret söylemi arasındaki ilişki

Ayşecan Terzioğlu

Nefret söylemi ile ayrımcılık birbiriyle yakından ilişkilidir; öyle ki, siyaset bilimci Katharine Gelber (2021) nefret söyleminin, ilgili herkes için eşit fırsat ve haklar ilkesine çoğu zaman zarar veren söylemsel bir ayrımcılık fiili olarak anlaşılabilirliğini vurgular. Gelber, nefret söylemi ve ayrımcılığın, siyasi, ekonomik ve toplumsal bağlama göre insanlara doğrudan veya dolaylı olarak zarar verebileceğini anlatır. Bu bağlamsal izahın bir de zamansal boyutu vardır; nefret söylemi ve ayrımcılık zaman içerisinde ayrı ayrı veya birbirine eşlik ederek tekrar eder. Nefret söylemi ve ayrımcılık bazen birbirine dönüşüp belirli grupları hedef alan benzer söylem ve pratikler de üretebilir. Fakat ayrımcılık sadece söylemsel değil, eylemsel ve kurumsal boyutları da içerdiği, sözlü olmayan yollarla da ifade edilip uygulandığı için nefret söylemine nazaran çok daha geniş bir kavramdır.

Hem nefret söylemi hem de ayrımcılık, bazen ayrı ayrı, bazen de birbiriyle kesişimli bir şekilde iç içe geçerek çoğu zaman ırk/etnisite, sınıf ve toplumsal cinsiyet temeline dayandırılır (Crenshaw, 2005). Fakat insanların yaşı, medeni durumu, mesleği, eğitim seviyesi, siyasi görüşü, dini inancı ve yaşadığı yer gibi başkaca demografik ve sosyoekonomik faktörler de belirli bağlamlar içerisinde onları nefret söyleminin ve/veya ayrımcılığın hedefi haline getirebilir. Ayrımcılık bireysel eylemlerde de ortaya çıkabilir; örneğin Almanya’da katılımcıların araç paylaşımı tercihlerine ilişkin yapılan ankette, Türk ve İtalyanlar gibi belirli göçmen topluluklara karşı Alman toplumunda var olan önyargılar yüzünden, örneğin sürücünün etnik kökenine ilişkin **algının** araştırmaya katılan insanların seçimlerinde önemli bir kriter oluşturduğu saptanmış (Liebe & Bayer, 2021).

Ayrımcılık fiillerinin ve nefret söyleminin nasıl su yüzüne çıktığını göstermek için tipik bir örnek olarak, sahip oldukları toplumsal cinsiyet, etnisite, ırk veya vatan-daşıklıktan ötürü eğitim, barınma, sağlık hizmetleri veya iş imkânlarından mahrum bırakılan insanlardan bahsedebiliriz. Bu tür mahrumiyetlerde nefret söyleminin, “onlar (...) oldukları için bu haklardan yararlanmayı hak etmiyor” denilerek bir “mazeret” olarak kullanıldığını görüyoruz. Bu durumda ayrımcılığı “meşrulaştırmak” amacıyla kullanılan nefret söyleminin sistematik bir şekilde tekrarlanması halinde, nefret söylemi ile ayrımcılık fiilleri arasında bir kısır döngü oluşma riski bu-

lunuyor. İşte bu yüzden ki, akademisyenler, politika yapımında rol oynayanlar ve içerik üreticileri gibi konuyla ilgili tarafların, toplumda ve toplumsal bağlamlarda nefret söyleminin normalleştirilmesi riski taşıyan bu örüntülere karşı bilhassa temkinli olması gerekir. Dil felsefesi ve epistemoloji alanında çalışmalar yürüten akademisyen Kindermann (2023), birtakım gruplardan dışlamak ve bu gruplara mensup insanlarla ilişkilerini bozmak suretiyle insanlara zarar verse dahi, ayrımcı söylemlerin tüm biçimlerinin ahlaki olarak reddedilmeyip hukuki düzenlemeye tabi tutulmayacağını savunuyor. Kindermann’a göre, bu türden bir yaptırım ve düzenlemeye konu edilebilmek için gerekli olan “yeterince zararlı” seviyesinde daha düşük bir kıstas tanımlamak oldukça zordur, zira sözümona “en ufak” ve ağır ayrımcılık fiilleri bile aynı temel niteliğe sahip olup, aynı süreç ve ortamın parçasıdır.

Bilişim, bilgisayar mühendisliği ve sosyal psikoloji alanlarında çalışan Fortuna ve Nunes (2018), şakalar da dahil olmak üzere zımni ayrımcılık fiillerinin bile nefret söylemi sayılması gerektiğini, zira bunların da kalıplaşmış genellemelere ve olumsuz yargılara dayandığını ve insanlar üzerinde olumsuz psikolojik etkiler yaratabildiğini öne sürüyor. Ayrımcılık insanların gündelik yaşamsal eylemlerinin ötesine taşınıp davranış kalıplarına dönüştüğünde kurumsal hale gelerek kurumların politika ve prosedürlerini şekillendirmeye başlayabilir. Örneğin bazı şirketlerin işe alım ve terfi modelleri toplumsal cinsiyete ve/veya cinsel yönelime dayalı ayrımcılık sergileyebiliyor; öyle ki kadınlar ve LGBTİ+ bireyler işe alınabilmek ve/veya terfi edilebilmek için görünmez engellerle mücadele etmek zorunda kalabiliyorlar (Manzi & Heilman 2021).⁶

6 Okurlar, nefret söylemi ve LGBTİ+ göçmenler bağlamında gündelik hayattaki etkileri hakkında daha ayrıntılı bir tartışma için bu raporun Eser Selen tarafından kaleme alınan kısmını incelemeye davet edilmektedir.

Tehlikeli söylem: Zarar veren dili ayırt edebilmeye yönelik bir çerçeve

Susan Benesch

Nefret söylemi için olduğu gibi, yine onunla bağlantılı toksik veya saldırgan söyleme dair de birbiriyle yarışan pek çok tanım mevcuttur. Bunun başlıca sebeplerinden biri de, söz konusu tanımlara dahil edilen hakaret veya saygısızlık gibi pek çok zararın öznel olmasından ve çeşitli kişi ve kültürler arasında değişkenlik arz etmesinden kaynaklanıyor. Ayrıca hakaret ve saygısızlık, kamusal söylemde izin verilmesi gereken eleştiri gibi kimi söylem biçimleriyle de yer yer kesişebilir veya kafa karışıklığı yaratabilir. Dolayısıyla, nefret söylemi kategorisinin sınırlarında dolaşırken, bilhassa kültürler ve normatif gruplar arasında hangi içeriğin bu kategoriye girdiğini tespit etmek hususunda bir fikir birliğine varmak güçleşiyor.

Öte yandan, benzer bir başka söylem kategorisini daha tutarlı bir şekilde ve daha az görüş ayrılığı yaratarak tanımlamak mümkün olabilir: tehlikeli söylem, yani insanları kendinden başka bir gruba mensup insanlara yöneltilen şiddete göz yummaya ve hatta bizzat böyle bir şiddete başvurmaya teşvik eden içerik. Tehlikeli söylem “gösterge” olarak anılan çarpıcı benzerlikler veya retorik kalıplar içerir. Bu tehlikeli söylem göstergeleri arasında insandılaştırma (canavarlaştırma), topluluk içindeki kadın ve kız çocuklarına yönelik saldırı iddiaları, topluluğun safliğine veya bütünlüğüne yönelik bir sözde tehdit anlatısı ve aslında konuşmacı dış topluluğa saldırmayı planlarken, bu saldırının haklı ve hatta yiğitlik sayılacağı algısı yaratmak için iç topluluğun mensuplarına dış topluluğun kendilerine saldırma planı yaptığını söyleyerek kitleyi kendi planladığı saldırının savunma amaçlı olacağına inandırmak anlamına gelen “yansıtılmalı suçlama” yer alır.

Neredeyse tüm tehlikeli söylemlerin taşıdığı bir başka özellik ise, insanları başka bir topluluktan korkmaya ve o topluluğu ya da üyelerini ciddi bir tehdit olarak görmeye ikna ederek şiddete karşı var olan olağan toplumsal bariyerleri zayıflatmaktır. Korku duygusu, grupları şiddet noktasına taşıyacak şekilde birbirinden uzaklaştırarak belki de nefretten çok daha güçlü ve belirgin bir etken oluşturabilir (Leader Maynard & Benesch, 2018).

Tehlikeli söylem, uzmanın *başka* insanların davranışı hakkında bir tahminde bulunmasını gerektiren analitik bir çerçeve kullanılarak tespit edildiği için, nefret söylemi kadar öznel bir tanıma sahip değildir.

Son olarak, tehlikeli söylemin kategorik sınırları herhangi bir kimlik grubunu dahil ederek veya hariç tutarak oluşturulmaz. Böylece, çoğu zaman belirli toplulukları konu edinen nefret söylemi tanımları arasındaki büyük bir tutarsızlık kaynağı tehlikeli söylem açısından aşılmış olur. Yukarıda belirtildiği üzere, son zamanlarda önerilen nefret söylemi tanımları da, İkinci Dünya Savaşı sonrasında hazırlanan insan hakları hukuku belgelerinde etnisite, ırk, din ve uyrukla sınırlı tutulan eski tanımlara nazaran daha geniş çeşitlilikte toplulukları içeriyor.

Yeni tanımlar — teknoloji şirketlerinin içerik moderasyonu için kullandıkları tanımlar da dahil olmak üzere — sosyal sınıf, göçmenlik statüsü ve cinsiyet kimliği gibi kimlik belirteçlerini de içeriyor. Ancak bu tanımlara da genellikle toplumsal cinsiyet, siyasi görüş/mensubiyet ve istihdam durumu dahil edilmiyor. Bu nedenle, örneğin gazetecilere yönelik zehir zemberek saldırılar, aslında nefret ve ayrımcılığı körüklediği halde, çoğu zaman nefret söylemi sayılmıyor. Oysa bunların tehlikeli söylem oluşturma ihtimali kesin ve çoğu zaman da gerçekten tehlikeli söylem teşkil ediyorlar.

Tehlikeli söylem ile nefret söyleminin bir diğer farkı da, nefret söylemlerinin tümü topluluklar arası şiddet riskini arttırmazken, tehlikeli söylemlerin hepsinin de nefret yüklü olmamasıdır. Yani, Gelber'in (2019) sistematik ayrımcılık analizinde savunduğu üzere , nefret söylemi ayrımcılık gibi bazı zararları her zaman artırıyor olabilir, fakat (sayıları yetersiz olsa da) kimi kitleler şiddete başvurmak veya göz yummak konusunda oldukça temkinli davrandığı içindir ki, şiddeti her zaman artırdığı söylenemez. Öte yandan, insanları başka bir topluluktan nefret etmeye değil ama *korkmaya* sevk ederek, şiddete göz yummaya ikna etmek mümkündür.

Şekil 1: Nefret Söylemi ile Tehlikeli Söylem Kesişim Kümesi



Kısacası, nefret söylemi ile tehlikeli söylem Şekil 1'deki Venn şemasında gösterilen bir kesişime sahiptir.

Tehlikeli söylem çoğu zaman asılsızdır; bu hiç de şaşırtıcı değildir, zira bir grup insan hakkında korkunç tanımlamalarda bulunur. Ne yazık ki insanlar yanlış bilgi (asılsız iddialar) veya dezenformasyon (bilerek veya kasten yayılan asılsız iddialar) yoluyla kolayca kandırılabilir. Üstelik bu yanlış bilgi ne kadar ürkütücüyse, insanların bunları doğru olup olmadığı konusunda emin olmasalar bile yayma ihtimalleri artar. Bu gibi durumlarda insanlar abartılı veya asılsız haberleri benimsemeye hazırdır (Leader Maynard & Benesch, 2018).

Akademisyenler ve özellikle de teknoloji şirketlerinin söylem regülatörleri tehlikeli söylem kavramını birkaç nedenden ötürü kullanışlı buluyor. Birincisi, yukarıda belirtildiği üzere, bu kavram topluluklar arasında şiddet riskini artırma kapasitesi üzerinden tutarlı ve nesnel bir şekilde tanımlanıyor. İkincisi, toplumların çoğu kitlesel şiddete karşı olmak ve bunu önlemenin değeri konusunda katı bir görüş birliğine sahip. Üçüncüsü, her topluluk tehlikeli söyleme hedef olabileceğinden, hangi kimlik grubunun bu tanıma dahil edileceğini tartışmaya gerek yoktur. Tehlikeli söyleme dair birçok ülke ve zamanda yürütülen çalışmalar, bu söylemin çok sayıda topluluğu ve topluluk türünü hedef aldığını tespit etmiştir.

Saldırgan dil: Değişkenlik ve güçlükler

Tommaso Caselli

Rencide edici, *saldırgan dil* terimini tanımlamak kolay iş değildir. “Bir şeyin rencide edici olması için birilerinin rencide olması lazımdır” (O’Driscoll, 2020, s. 11). Buradan hareketle (varsa) hangi kelimelerin rencide edici olduğu, bu durumun nasıl ve ne zaman tetiklendiği (yani, kültürel geçmiş ve deneyimler de dahil olmak üzere, belirli bir mesajın dillendirildiği ve muhataplara iletildiği durum) ile ilgili çeşitli sorular doğar.

Biz burada, söylenen şeyin ardında yatan niyetten ziyade, varsayımsal bir söylem durumunda kelime ve cümlelerin muhataplar üzerinde yaratabileceği etkilere odaklanacağız. O’Driscoll’un (2020) yaklaşımında saldırgan dil ölçeklendirilebilir bir olgu olarak değerlendirilir: Söylenen şeyin ne ölçüde rencide edici olarak algılandığına dair hayali bir sürekliliğin takibi yapılır ve bu ölçek nesnel değildir. Saldırgan dilin rencide ediciliği esasen öznel bir olgudur: Muhatabın algısına dayanır.

Saldırgan dil olgusuna farklı disiplinlerden (dilbilim, medya çalışmaları, iletişim çalışmaları, doğal dil işleme, felsefe, psikoloji, hukuk) ilgi olmakla birlikte, bu olgunun tek ve müşterek bir tanımı yoktur. Aşağıdaki Tablo 1’de bu araştırma konusunu nitelendirmek için İngilizcede kullanılan en yaygın terimlerin bir kısmını kullanarak Google Akademik⁷ üzerinden yapılmış, sistematik olmayan bir ince-

Tablo 1. “Saldırgan dil”den bahsederken İngilizcede sıklıkla kullanılan bazı ifadelerin kullanım sıklığına ilişkin Google Akademik sonuçları.

Terim	Google akademik sıklığı
Toxic language (Toksik dil)	3.720.000
Abusive language (Tacizkâr dil)	2.180.000
Offensive language (Saldırgan dil)	2.090.000
Taboo language (Tabu olan ifadeler)	1.500.000
Obscene language (Yakışsız/müstehcen ifadeler)	624.000
Insulting language (Aşağılayıcı ifadeler)	522.000
Swearing (Küfür)	220.000

7 Araştırma 11 Şubat 2024 tarihinde yapılmıştır.

lemenin sonuçları yer alıyor. Sıklık, bu terimlerle beslendiğinde arama motoru tarafından alınan girişlerin sayısını ifade eder.

Bu anlamda en çok kullanılan terim “toksik dil” olurken, “saldırgan dil” ifadesinin üçüncü sırada yer alması oldukça şaşırtıcıdır. Öncelikle, bu terimlerin her birine dair gelen sonuçlarla ilişkilendirilen metin parçacıklarında yer alan tanımların bir kısmı incelendiğinde, bunların O’Driscoll (2020) tarafından da işaret edildiği üzere birbirinin yerine geçecek şekilde kullanıldığını görüyoruz. Bu durum sadece bir ifadenin diğerine tercih edilmesinden ibaret olmayıp, saldırgan dilin tam olarak ne içerdiğine dair yazarlar arasında oldukça yaygın bir fikir ayrılığı bulunduğuna işaret ediyor. Saldırgan dilin anlam ve niteliğini ortaya koymak için kullanılan tanımların bir kısmını alıp karşılaştırmaya başladığımızda saptadığımız ortaklıklar ve genel örüntü, bu dilin son derece geniş kapsamlı olup, çok daha hususi ve çirkin ifadeleri (örn. nefret söylemi) içine alan geniş bir şemsiye terim işlevi gösteren genel bir olgu olarak kullanıldığına işaret ediyor (bkz. Tablo 2).

Tablo 2. “Saldırgan dil” için önerilen mukayeseli tanımlar

Açıklama	Kaynak
(...) “toksiklik” terimi, kişinin bir tartışmayı terk etmesini muhtemel kılan kaba, saygısız ve abes dil olarak tanımlanır.	Perspective API ⁸
(...) muhatap olan kişinin benlik algısı ve/veya esenliği üzerinde olumsuz bir etkisi olan veya olabilecek, yani onları orta veya aşırı derecede rahatsız eden ve/veya aşağılanmış hissettiren ve/veya inciten ve/veya korkutan ya da bu sonuçlara yol açma ihtimali olan herhangi bir kelime veya kelimeler dizisi.	O’Driscoll (2020, p. 16)
“Biz ne türden olursa olsun kabul edilemez bir dil (sövgü) veya zımnı yahut doğrudan nitelikte hedefli bir saldırı içermesi halinde bir gönderiyi saldırgan olarak etiketliyoruz. Hakaretler, tehditler ve kaba bir dil veya küfür içeren gönderiler bu kategoriye giriyor.”	Zampieri et al. (2019)
“Müstehcen sözler, başka türden saldırgan dil kullanımlarının aksine, iletme niyeti taşıyor olabileceği diğer tüm mesajlar bir yana, içerdiği belirli bazı sözcükler yüzünden dinleyiciyi bütünüyle sarsar.”	Feinberg (1983)

8 Perspective API. (t.y.). *API Hakkında – Tavırlar ve Diller*
<https://developers.perspectiveapi.com/s/about-the-api-attributes-and-languages>

Her ne kadar kavramı isimlendirmek için kullanılacak terim üzerinde mutabakat sağlanamasa da (toksiklik/saldırgan dil/müstehcenlik), Tablo 2’deki tüm tanımlar saldırgan dilin özünü oluşturan, sarf edilen ifadelerin muhatap üzerinde yarattığı olumsuz etkilere odaklanıyor.

Saldırgan dilin (mizojini, tehlikeli söylem, homofobi vs.) başka olgularla nasıl ilişkilendiğini sistematik bir şekilde düzenleyip ilişkilendirmek kolay bir şey değil; ve tekrar belirtmek gerekirse, bu konuda görüş birliği yok. Konuyla ilgilenen okurlar öncelikle Poletto vd. (2021) kaynağını okuyabilir.

Nefret söylemi ile ayrımcı söylem tespit edilirken en önemli faktörler arasında söylemin üreticisi ile söylemin hedef aldığı birey veya gruplar yer alır. Söylemin üretildiği bağlam da nefret söylemi ile ayrımcı söylemin değerlendirilmesi bakımından önem taşır. Nefret söyleminin hedefi olan grupların temel özellikleri neredeyse her tanımda benzer şekilde dile getirilmekle birlikte, bu grupların sabit ve değişmez olmadığı, bağlama, zamana ve başka birçok değişkene bağlı olarak yeni grupların hedef alınabileceğini hatırlamak gerekir. Tanımların birçoğunda cinsiyet ve/veya toplumsal cinsiyet hedef alınan bir grup özelliği olarak anılmakla birlikte, Avrupa Konseyi Gençlik Biriminin yaptığı bir araştırmaya göre kadınlar hala görünürlüğü düşük bir hedef grup olmayı sürdürüyor (Avrupa Konseyi, 2016). Bu anket çalışmasında, kadınların sıklıkla sanal ortamda nefret söyleminin hedefi olduğunu ve bunun da çoğu zaman ırksal veya dinsel nefret söyleminden daha az ciddiye alındığını gösteriyor. Ankete katılan kişiler cinsiyetçi nefret söylemini, kadınlara yönelik çok daha geniş ayrımcılık biçimlerine çanak tutan ve yapısal eşitsizliklerden beslenen toplumsal cinsiyete dayalı şiddetin yaygın bir türü olarak görüyor. Bu çalışma ayrıca, toplumsal cinsiyet eşitliği ile ifade özgürlüğünün birbiriyle çatışan değerler olarak tarif edilmesinin kapsayıcı kamusal söylem önünde engel oluşturduğunu vurguluyor. Aslında toplumsal cinsiyet eşitliği ile ifade özgürlüğü birbirini pekiştirir, zira kimi topluluklar nefret söylemiyle sistematik bir şekilde ötekileştirildiğinde gerçek ifade özgürlüğünün varlığından bahsetmek mümkün değildir.

Kesişimsellik ve hedef alınan gruplar açısından ise, LGBTİ+ birey ve mültecilerin son yıllarda Türkiye’de hem basında hem de sosyal medyada nefret söylemine en yoğun şekilde maruz kalan gruplar arasında olduğunu söylemek abartı olmaz. Bu durumu, bilhassa seçim dönemlerinde siyasi partilerin ürettiği söylemler bağlamında genel siyasi ortamın bir yansıması olarak değerlendirmek de mümkündür. Bu noktadan hareketle, Eser Selen nefret söylemi ile şiddet arasında doğrudan bir bağlantı kurup, Türkiye’deki LGBTİ+ mültecilerin nefret söylemine ne şekillerde maruz kaldığını ve bundan nasıl etkilendiğini ele alıyor.

Kesişimsellik ve nefret söylemi: Türkiye'deki LGBTİ+ yurttaş ve göçmenlerin durumu

Eser Selen

İrk, etnisite, din, dil ve sınıf gibi toplumsal sınıflandırmaların cinsel yönelim ve cinsiyet kimliği ile nasıl kesiştiğini ve ayrımcı söylemin bir parçası olarak aktif şiddeti körükleyen yoğun nefret söylemi deneyimlerine yol açtığını (Chun vd., 2013) anlamak bakımından kesişimsellik kavramı (Crenshaw, 1989; Acker, 2012) büyük önem taşıyor. Yabancı düşmanlığı, ırkçılık ve diğer önyargı ve damgalama biçimlerini de içine alan nefret söylemi LGBTİ+⁹ bireylere yönelik ötekileştirme ve şiddeti devam ettirerek bu kişilerin topluma entegrasyonuna ve tam anlamıyla katılımına engel olur (Faloppa vd., 2023). Türkiye’de LGBTİ+ yurttaş ve göçmenler kesişen kırılğanlıklarla dolu karmaşık bir ortamda yollarını bulmaya çalışıyor. Nefret söylemi bu kişilerin sadece cinsel yönelimleri ve cinsiyet kimliklerinden ötürü değil; ırk, etnisite, din, dil, sınıf ve ulusal köken ya da göçmenlik statüsü gibi başka faktörlerden de kaynaklı olarak karşılaştığı güçlükleri içeriyor. Örneğin, Türkiye’deki eşcinsel bir Suriyeli mülteci uyuğu, dili, dini ve cinsel yönelimi sebebiyle maruz kalabileceği önyargılardan ötürü nefret söylemine ve diğer şiddet biçimlerine karşı daha kırılğan hale gelebilir.

İnsan hakları savunucuları, politika yapımında rol oynayanlar ve araştırmacılar, nefret söylemi konusunu ele alırken, bunlarla sınırlı olmamak kaydıyla etnisite, ırk, din, toplumsal cinsiyet ve cinsel yönelim gibi kesişimsel faktörleri de göz önünde bulundurmalıdır, zira bu faktörler söz konusu söylemlerin yarattığı zararı ağırlaştırabilir. Nefret söyleminin kapsamını tam olarak saptayabilmek için kesişen kimliklere sahip bireyler üzerindeki etkisini anlamak şarttır. Ayrıca sanal veya fiziki ortamlarda nefret söylemine karşı durup fiili bir yanıt verebilmek için nefret söylemini sınıflandırabilmek gerekir. Bilhassa öznelliklerinden ötürü karmaşık güçlükler yaşayan LGBTİ+ göçmenler açısından bu husus hayati önem taşır. Farklı gruplar nefret söylemini kendilerine özgü şekillerde tecrübe ediyorlar ve bunların ayrıntılı şekilde sınıflandırılması bu söylemleri ele almak için özel yaklaşımlar geliştirilmesine imkân sağlar (Yuying vd., 2013). Ancak nefret söylemi her zaman açık ve sesli bir şekilde sergilenmez; mikroagresyon, yani esprili görseller

9 Türkiye’de aktivistler ve topluluk mensupları kendilerini tanımlamak amacıyla LGBTİ+ kısaltmasını kullandığı için ben de bu rapor için bu kısaltmayı benimsedim.

(mem) ve daha nükteli ifade biçimleri de dahil olmak üzere çeşitli yöntemlerle tezahür edebilir (Nadal, 2008; Nadal vd., 2011a; 2011b).

Nefret söylemini tanımlamanın yer yer zor olabileceği ve bağlamına göre değişiklik gösterebileceği tekrar vurgulanmalıdır (Russell, 2020). Ancak ayrımcı söylemin içeriği “kabul gören” tanımlardan hiçbirine girmiyorsa bile, kişi yine de söylemin zararlı ve saldırgan bir nitelik taşıması sebebiyle bunu sadece nefret dolu bir iletişim olarak değil, nefret söylemi olarak algılama hakkına sahiptir. LGBTİ+ bireylerin toplumsal söylem ve eylemler karşısındaki algıları ve bunlara verdiği yanıtlar bu kişilerin cinsiyet kimliklerini ve cinsel yönelimlerini çevreleyen yaygın fobiden ziyadesiyle etkileniyor. Bu fobi, ayrımcılık kalıpları üzerinde kritik bir faktör olarak işlev görür ve bireylerin toplumsal tutumlar, davranışlar ve duygusal tepkilerle nasıl başa çıktığını şekillendirir (Wijaya, 2022). Ayrıca cinsiyet kimliği ve cinsel yönelim konusundaki fobinin içselleştirilme oranı da LGBTİ+ topluluğu içerisinde çeşitlilik sergilediğinden, kişinin benlik algısını ve esenliğini etkileyebiliyor (Gao vd., 2023; McLean, 2021). Kapsayıcılığın desteklenmesi açısından, içselleştirilmiş fobinin bu kişilerin sarf edilenleri nefret söylemi olarak algılayıp algılamadığı, üzerinde yarattığı etkiyi tespit etmek şarttır. Bunun yanında, lezbiyen, gey, biseksüel, trans,¹⁰ kararsız, kuir, interseks ve non-binary deneyimlerin bağlamından kopuk bir şekilde değerlendirilemeyeceğini kabul etmek de şarttır (Fassinger & Arseneau, 2007). Dolayısıyla LGBTİ+ bireylerin kimlik ve deneyimlerine ilişkin bütünlüklü yelpazeyi yakından değerlendirmek gerekir.

LGBTİ+ bireyler tüm dünyada temel sosyal ve ekonomik haklara erişim konusunda pek çok güçlük ve engelle karşı karşıyadır (Alessi vd., 2017; Williamson, 2023). Araştırmalar LGBTİ+ bireylerin –göçmen olsun olmasın– istihdam, barınma, sağlık hizmeti, medeni hukuk ve aile hukuku konularında nasıl damgalandığını ve ayrıca nefret söylemine, AIDS’le bağlantılı olarak ayrımcılığa, polis tacizine ve şiddet içeren nefret suçlarına nasıl maruz bırakıldığını gösteren sayısız örnek içeriyor (Aswad & Kaye, 2020; Edward vd., 2020; Gowin vd., 2017; Hopkinson, 2017; Morales, 2013; Piwowarczyk, 2017). Önyargılar, kalıp yargılar, yanlış ve çarpıtılmış bilgiler de LGBTİ+ yurttaş ve göçmenlerin yaşadığı sıkıntıları katbekat arttırıyor. Bunlar arasında hiç şüphesiz ki en ağır basan sorun LGBTİ+ bireylere yöneltilen şiddettir; zira toplumun ahlaki (cinsel) pusulasını oluşturan heteroseksüellik bazen ölümcül de olabilen şiddetin bahanesi –ya da gerekçesi– olarak

10 Burada “trans” terimini kendisini transseksüel, non-binary, cinsiyetsiz, akışkan cinsiyet kimliğine sahip veya başka türlü tanımlayan bireyleri kapsayan bir tanımlama olarak değil, onların toplumsal cinsiyet deneyimlerinin çeşitliliğini teyit eden kapsayıcı bir belirteç olarak kullanıyorum.

kamusal ahlak kavramına başvurur (Selen, 2020). Örneğin, Türkiye’de kadına yönelik “normal/leştirilmiş” şiddet anaakım medyada devlet yetkilileri, nüfuzlu gazeteciler ve ünlüler tarafından mütemadiyen kınanırken, hemen hepsi trans kadınlara (veya bir LGBTİ+ yurttaş) yönelik şiddeti bu söylemlerinin dışında bırakıyor (Selen, 2020). LGBTİ+ bireylere yönelik şiddetin normalleştirilmesi devlet makamlarının toplumsal cinsiyet ideolojisiyle pekiştirilince, nefret söylemini de içeren LGBTİ+ karşıtı söylem bu bireylerin yaşamına her açıdan tesir eden tehlikeli bir ortam yaratıyor (Selen, 2020).¹¹ Ayrıca, anayasada ve yasalarda yapılan ayrımcı değişikliklerle toplumsal cinsiyet ideolojisi söyleminin hukuk sistemine dahil edilme ihtimali de ciddi bir tehdit oluşturuyor (McClain & Waite-Wright, 2015). Topluma kök salmış bu mesele, böylesi şiddet fiillerine göz yuman devlet politikalarıyla daha da derinleşiyor (McClain & Waite-Wright, 2015).

Benzer biçimde, göçmen ve mültecilere yönelik nefret söylemi de bugün tüm dünyada aciliyet arz eden bir mesele ve Türkiye’de de durum maalesef farklı değil (Badalı, 2019; Eduardo vd., 2013; Filibeli vd., 2021; Özdüzen & Korkut, 2020). Göçmenlere yönelik nefret söylemi Türkiye’nin başlıca ve gitgide büyüyen meselelerinden biri. Mevcut kamusal söylem göçmenleri genellikle çözülmesi gereken bir “sorun” olarak resmediyor. Birleşmiş Milletler Mülteciler Yüksek Komiserliği (BMMYK) rakamlarına göre Türkiye, çatışma, şiddet ve zulüm sonucunda yerlerinden edilmiş insanların büyük bir kısmına ev sahipliği yaparak dünya çapında en yüksek mülteci sayısına sahip ülke pozisyonunu koruyor. Halihazırda Türkiye’de 3,6 milyon Suriyeli mülteci ve başka uyruklardan ilgi alanındaki kişi addedilen yaklaşık 320.000 kişi barınıyor (BMMYK, 2024; Kaya, 2023). Araştırmalar, dünya genelinde LGBTİ+ sığınmacı ve göçmenlere yönelik nefret söylemi de dahil olmak üzere ayrımcı dil ve davranışların ağır sonuçlarını ortaya koymaktadır. (Alessi vd., 2017; Aswad & Kaye, 2020; Edward vd., 2020; Gowin vd., 2017; Hopkinson, 2017; Morales, 2013; Piwowarczyk, 2017). Son yıllarda LGBTİ+ haklarının ciddi engellerle karşılaştığı Türkiye’de ise, LGBTİ+ göçmenlere yönelik nefret söylemi toplumda halihazırda var olan ayrımcılık ve düşmanlığı daha da alevlendirerek bu topluluğun esenliğini, sağlığını, geçim ve güvenliğini etkiliyor.

¹¹ 2013 yılındaki Gezi protestolarından beri LGBTİ+ bireylerin Türkiye’de en çok hedef alınan azınlıklardan biri haline geldiğini söylemek yanlış olmaz. LGBTİ+ bireyler sadece ülkedeki LGBTİ+ karşıtı hareket tarafından değil, toplumsal cinsiyet politikalarının bir parçası olarak üst düzey devlet yetkilileri ve siyasi parti liderlerinin yanı sıra yirmi yılı aşkın süredir nüfuzlu konumlarda bulunan gazeteciler tarafından da hedef alınıyor (bkz. Selen, 2020). Dahası, barışçıl toplantı ve gösteri yapma hakkına yönelik ihlaller de LGBTİ+ topluluğun kamusal alanda görünürliğini kısıtladığından, radyo ve televizyon yayıncılığı, sosyal medya ve sanat alanlarında sansürü daha da arttırıyor (bkz. Selen, 2012; Selen, 2020; Kılıç, 2023).

Dijital ortamda LGBTİ+ karşıtı nefret söyleminin yükselişi, bu söylemin ne kadar yaygın hale geldiğini gözler önüne sererken şu hususları da ortaya çıkarıyor:

- **Hesap Verebilirlik Eksikliği:** Lezbiyen, gey, biseksüel ve trans bireyleri hedef alan şiddet içerikli saldırılar ve cinayetlere yönelik resmi soruşturma sayılarındaki çarpıcı eksiklik koruma ve adaleti sağlamak konusunda sistematik bir acizlik olduğunu gösteriyor.
- **Zorunlu Görünmezlik:** Cinsel yönelimin açıklanmaması yönündeki baskılar bir korku iklimi yarattığından, LGBTİ+ bireyler devamlı surette güvenlik endişesi yaşıyor ve/veya geçim kaynaklarından mahrum kalıyor; üstelik bundan ötürü başvurabilecekleri hukuki bir yol da bulunmuyor.
- **İnsandışılaştıran (Canavarlaştıran) Söylem:** LGBTİ+ özelliğinin yaygın bir şekilde “sapkınlık”, “hastalık”, “terör” ve “tehdit” ile ilişkilendiriliyor olması LGBTİ+ yurttaş ve göçmenlerin varlık ve haklarının inkârına yol açıyor.

Tablo 3. Veri tabanı girdilerine örnek

Kullanıcı adı	Anahtar kelime	Platform	Gönderi	Yaş	Cinsiyet
@anonim_#1	LGBT*	Twitter (2020)	“Lgbt = sapıklık insanın bozulmuş hali imha edilmeli idam edilmeliler. Bir yerde hastalık varsa mikroplar öldürülür lgbt toplumu hasta eden mikroplardır.”	Belirtilmemiş	Belirtilmemiş
@anonim_#2	Trans*	Facebook (2019)	“Allah onların belasını versin. Rabbim onları mahvetsin, kahrolsunlar. En kısa zamanda yok olsunlar.”	28	Erkek

Sanal anlatıları inceleyip gerçek dünyadaki deneyimlerle ilişkilendirdiğimizde, nefret söyleminin bileşik yapısını ve Türkiye'deki LGBTİ+ bireylerin yaşamı üzerindeki yıkıcı etkisini gözler önüne serebiliriz (Özdüzen & Korkut, 2020). Nefret söylemiyle mücadele edip daha kapsayıcı bir toplum yaratabilmek için etkili stratejiler geliştirmek bakımından böyle bir incelemeye gerek duyuluyor.

Twitter (2002-2020) veya Facebook (2019-2020) gibi sosyal medya sitelerinden metin madenciliği yapılarak oluşturulan veri kümelerinde "LGBT*", "trans*", "homoseksüel", "biseksüel" gibi anahtar kelimeler kullanılarak yapılan bir ön Boolean araması "ahlaksız", "sapık", "sapkınlık", "hastalık", "tehdit" ve "terörist" gibi terimlerle olan ilişkilendirilmeyi ortaya çıkarıyor. Bu terimler, bize, "yok etme/edilme", "öldürme", "helak etme/edilme" ve "lanetleme" gibi nefret söylemi ve/veya tehlikeli söylemlerin yanı sıra LGBTİ+ bireylere yönelik Türkçe'deki diğer "ayrımcı söylem" biçimlerini de gösteriyor (Tablo 3).

Araştırmacılar LGBTİ+ örgütlerin temel çalışmalarını güvence altına alıp aktif bir şekilde desteklemenin yanı sıra, eşitliğin sağlanması ve sosyal ve ekonomik hakların korunması adına LGBTİ+ topluluğun hak ve meseleleri hakkında kamuoyunu erişilebilir incelemeler yoluyla bilgilendirmelidir. Nefret söylemini ortadan kaldırmak için şu tedbirler hayati önem taşıyor:

- **Nefret söylemine karşı çıkmak:** Farkındalığı arttırmak ve kapsayıcılığı teşvik etmek suretiyle dijital ortamda LGBTİ+ yurttaş ve göçmenlere yönelik nefret söylemine aktif bir şekilde karşı çıkmak.
- **Destek hizmetleri sunmak:** LGBTİ+ yurttaş ve göçmenlerin ihtiyaçları doğrultusunda psikolojik danışmanlık, hukuki yardım ile barınma ve istihdam konularında yardım gibi özel destek hizmetleri ve kaynaklar sunmak.
- **Toplulukları güçlendirmek:** Eğitim, kapasite geliştirme ve hukuki yardım yoluyla LGBTİ+ toplulukları güçlendirmek ve haklarını savunmak.
- **Yasal korumaları arttırmak:** Nefret söylemi ve ayrımcılığa karşı yasal korumaları, LGBTİ+ bireyleri de açıkça içerecek şekilde genişletip arttırmak.
- **Hesap verebilirliği sağlamak:** Nefret söylemi faillerini hukuki işlemler ve kamusal kınama yoluyla sorumlu kılmak.
- **Yasal korumaları uygulamak:** Nefret söylemine karşı getirilen yasal korumaların etkili bir şekilde uygulanmasını ve icrasını sağlamak.

Hrant Dink Vakfı'nın Türkiye yazılı basınında nefret söyleminin izlenmesine yönelik on yıllık projesi (2009-2019), hedef alınan grupların sıralanmasının değişen siyasi ve toplumsal dinamiklere bağlı olarak dalgalandığını, ancak bazı kimliklerin güncel olaylardan bağımsız şekilde her daim nefret söyleminin hedefi olduğunu ortaya çıkardı. **Ermeniler, Yahudiler, Hristiyanlar, Rumlar ve Yunanlar** yazılı basında nefret söyleminin daimi hedefi olurken, bu gruplara yöneltilen husumetin derinlere yerleşmiş tarihsel anlatılarda yatan kökleri siyasi söylemlerle pekiştiriliyor. Vakıf bu grupları **“değişmeyen ötekiler”** olarak adlandırırken, medyadaki tasvirlerinin zaman içerisinde değişmeden sürdüğünü vurguluyor. 2014 yılında Suriye'den yaşanan kitlesel göçün ardından **Suriyeliler** de hızla en çok hedef alınan topluluklar arasına girdi. Bu değişim, nefret söyleminin mütemadiyen **“yeni ötekiler”** inşa edip, geçmişten beri hedef alınan gruplara yeni tehdit olarak resmettiği başka gruplar katarak kapsamı genişlettiğini gösteriyor.

Bu gruplara yönelik nefret söylemi türlerini incelediğimizde tekrar eden örüntüler ortaya çıkıyor. Altı grup açısından en yaygın görülen nefret söylemi biçimleri *abartma*, *yükleme* ve *çarpıtma* olup, yanlış bilgiler ve olumsuz genellemeler yoluyla kalıpyargılar ve mevcut önyargılar devam ettiriliyor. Ancak her grup birbirinden farklı anlatılarla hedef alınıyor. Ermeniler “ulusal kimliğe yönelik iç ve dış tehdit” olarak tasvir edilerek orantısız bir *düşmanlık ve savaş söylemine* maruz bırakılıyor. Rumlar ve Yunanlar onları eskiden beri genellikle “düşman” olarak betimleyen haberlerle şekillenen iç içe geçmiş “husumet” anlatılarına konu ediliyor. Suriyeliler ise öncelikle “demografik ve kültürel tehdit” söylemi üzerinden hedef alınarak, medyada çıkan haberlerde ekonomik kaygı kisvesi altında “ulusal kimliğin zayıflaması, suç ve siyasi istikrarsızlık” endişeleri körükleniyor.

Son yıllarda izleme çalışmaları **LGBTİ+**lara yönelik nefret söylemlerini de içine alacak şekilde genişletilmiş olup, cinsel yönelim ve cinsiyet kimliğinin de ayrımcı söylemlere zemin oluşturmaya başladığı görülüyor. İlk başta etnik, ulusal ve dini kimlikler konu edilmekle birlikte, **LGBTİ+**ların gitgide “yeni öteki” olarak inşa edildiğine, devlet politikaları, resmî söy-

lem ve düşmanca bir siyasi iklim yoluyla hedef alındığına şahit oluyoruz. Aşırı sağcı, hükümet yanlısı anaakım medya LGBTİ+ları “sapkın” olarak betimleyip varlıklarını “toplumsal bir tehdit” olarak resmederek bu söylemin pekiştirilmesinde kilit bir rol oynuyor.

Vakfın izleme çalışmaları bu örüntüleri belgeleyip incelemeye devam ederken, Türkiye’deki medya ortamında nefret içerikli anlatıların ve ayrımcı söylemlerin ne kadar derinlere işlediğini gözler önüne seriyor. Nefret söyleminin gelişimini ve yoğunluğunu takip eden bu tespitler yoluyla, yapısal ayrımcılığa ve hedef alınan gruplara yönelik toplumsal algıyı şekillendirmek bakımından medyanın üstlendiği role dair önemli içgörüler sağlanıyor.

NEFRET SÖYLEMLERİNİN TESPİTİ: POLİTİKALAR, YÖNTEMLER VE GÜÇLÜKLER

Nefret söyleminin tespitine yönelik girişimler, hukuki çerçeve, platform politikaları ve zararlı dile dair değişen bir toplumsal anlayışla şekillenen daha geniş bir bağlam içerisinde yer alır. Nefret söylemine yönelik politika ve çerçeveler, kültürel, siyasi ve teknolojik faktörler yoluyla şekillendiğinden hukuki ve dijital ortamda farklılıklar gösterir. Devlet politikaları nefret söylemi konusunu düzenleyip cezalandırmak için genellikle yasal mekanizmalara dayanırken, sosyal medya platformları kendi ekosistemleri içinde içeriği denetlemek üzere kendi çerçevelerini oluşturur. Bu yaklaşımlar, hem nefret söyleminin tanımlanmasına ilişkin karmaşıklıkları hem de zararlı dile mücadele ile ifade özgürlüğünün korunması arasında var olan daha geniş kapsamlı gerilimleri yansıtacak şekilde, kapsam, uygulama ve temel motivasyonlar açısından farklılık gösterir.

Bu politikalar, manuel veri ayıklamadan yapay zekâ destekli yaklaşımlara kadar uzanan tespit yöntemlerinin nasıl geliştirildiğini etkiler. Bu tespit, hukuki veya platform düzeyindeki düzenlemelerde her zaman hesaba katılmayabilen teknolojik imkânlar, dilsel nüanslar ve kültürel bağlamlarla da şekillenir. Bazı durumlarda politikalar ya birtakım zararlı söylem biçimlerini yakalayamayan dar tanımlar ya da aşırı sansür ve ifade baskısına yol açan çok geniş ve muğlak kriterler getirerek tespit çabalarına mâni dahi olabilir.

Bu bölümde önce devlet ve platform düzeyinde nefret söylemine ilişkin politikalar incelenerek bunların tespit yöntemleriyle nasıl etkileşime girdiği anlatılıyor. Ardından manuel veri ayıklama, veri kümesi iyileştirme ve yapay zekâ tabanlı modeller de dahil olmak üzere nefret söyleminin tespitine yönelik farklı yaklaşımlar incelenirken, önyargı, doğruluk ve etik kaygılardan ötürü karşılaşılan güçlükler ele alınıyor. Politikalar ile tespit yöntemleri arasındaki etkileşimi anlamak, nefret söyleminin pratikte teşhis edilmesini güçleştiren karmaşıklıklara dair daha ayrıntılı ve dengeli bir tartışma yürütülmesine imkân sağlıyor.

Nefret söylemine ilişkin politikaların anlaşılması, bunların tespit çabaları üzerindeki etkilerini ve içerdikleri güçlükleri eleştirel bir bakışla değerlendirmek açısından temel önem arz eder. Tirşe Erbaysal Filibeli tarafından hazırlanan bu bölümde, devlet ve dijital platformlar tarafından uygulanan düzenleyici çerçeveler incelenerek bunların icra mekanizmaları, sınırlılıkları ve etkileri mercek altına alınıyor.

Devlet politikaları, düzenlemeler ve politika analizi

Tirše Erbaysal Filibeli

Ülkeler nefret söylemine yönelik düzenleme yaparken farklı yaklaşımlar benimsiyorlar. Bu noktada temel sorun, “nefret söylemi” ile “ifade özgürlüğü” arasındaki gri alandan kaynaklanıyor. 32 Avrupa ülkesini değerlendiren 2024 Medyada Çoğulculuğu İzleme Raporuna göre (Bleyer-Simon vd., 2024), sadece Almanya, Danimarka, Finlandiya, İsveç ve Litvanya dezenformasyon ve nefret söylemine karşı korunma konusunda düşük bir risk sergiliyor. Öte yandan Arnavutluk, Bulgaristan, Karadağ, Kıbrıs, Macaristan, Malta, Romanya, Sırbistan, Slovenya ve Türkiye kapsayıcı nitelik taşımayan, eksik geliştirilmiş nefret söylemi politikalarına sahip olduğu için medyada çoğulculuk ve çeşitlilik açısından yüksek risk yaratıyor.

Sosyal medya platformlarının büyük bir kısmı hizmet şartları bakımından ABD’deki yasa ve düzenlemeleri benimsemiş durumdadır. Ne var ki orada da nefret söyleminin yasal bir tanımı olmayıp, nefret söylemi biçimlerinin çoğu Anayasa kapsamında ifade özgürlüğü koruması kapsamındadır.

Almanya ve Fransa sanal ortamda nefret söylemiyle mücadele için özel yasalar çıkardı. Almanya Holokost’un inkârını, nefretin körüklenmesini ve Nazi propaganda yapılmasını suç sayan katı bir nefret söylemi politikası izliyor. Sanal ortamda nefret söylemiyle mücadele ve sanal ortamda verilen zararlara karşı koruma sağlamak amacıyla 2020 yılında Alman Sosyal Ağ Yasası, 2021 yılında ise Küçüklerin Korunmasına ilişkin Yasa değiştirildi (Holznagel vd., 2023). 2020 yılında Fransa, çevrimiçi platformların nefret söylemini fark edildikten sonra 24 saat içinde kaldırmasını zorunlu kılan Avia Yasası’nı kabul etti; ancak bu yasa, ifade özgürlüğü üzerindeki olası olumsuz etkileri sebebiyle endişelere yol açtı (Rebillard & Sklower, 2022). Danimarka’da nefret söylemine karşı bir mevzuat bulunmakla birlikte, çevrimiçi nefret söylemine açıkça atıfta bulunulmamıştır (Simonsen, 2023). Litvanya ise işbirliğine dayalı eşsiz bir sistem işletiyor; nefret suçları ve nefret söylemi kamu kurumları ve STK’lar tarafından sanal platformlara bildirilebiliyor (Balčytienė & Juraitė, 2022).

Türkiye, 18 Ekim 2022’de Basın Kanunu ile Bazı Kanunlarda Değişiklik Yapılmasına dair 7418 sayılı Kanun’u çıkarttı. Bu mevzuatla Türk Ceza Kanunu’nun “Kamu Barışına Karşı Suçlar” başlıklı bölümünde yer alan 217. maddede deği-

şiklik yapılarak “halkı yanıltıcı bilgiyi alenen yayma” suçu bu hükme eklendi. Ancak kanunda dezenformasyon ve nefret söylemine ilişkin net bir tanım olmadığından, bu terimlere ilişkin anlayış yetkili mercilere bırakılmış durumda (İnceoğlu vd., 2022).¹²

Nefret söylemine ilişkin açık ve net bir tanım olmaması kanunun keyfi şekilde uygulanmasına yol açabilir. Bu durum gazeteciler, insan hakları aktivistleri, akademisyenler ve diğer savunucular açısından olumsuz etkiler yaratıyor. Dolayısıyla bu tür keyfi işlemlerin önlenmesi için kapsamlı ve kapsayıcı nefret söylemi politikaları geliştirilmesi her açıdan şart görünüyor.

¹² **Editör Notu:** 7418 sayılı Kanun’la getirilen yasal belirsizlikten faydalanılarak, Ocak 2024’te önerilen Etki Ajanlığı yasa tasarısında muğlak hukuki tanımların kapsamı daha da arttırılmış olup, bu durum söz konusu tasarının birçok sektör açısından olası etkileri hakkında endişelere yol açtı. Yasa tasarısı yabancı fonlardan destek alan kişi ve kuruluşları suçlu haline getirebileceği için ifade özgürlüğü ve sivil toplum açısından ciddi riskler içeriyor. Tasarının içindeki muğlak ifadelerin keyfi uygulamaya imkân sağlayıp gazetecileri, akademisyenleri ve STK’ları hedef alma potansiyeli taşıdığı yönünde eleştiriler dillendiriliyor. Bu tasarı yasalaşacak olursa, bağımsız sesler ve uluslararası dayanışma daha da kısıtlanmış ve böylelikle otosansüre ve hukuki açıdan belirsizliğe yol açan atmosfer daha da ağırlaşmış olacaktır.

Sosyal medya platformlarının nefret söylemine ilişkin mevcut politikaları

Tirşe Erbaysal Filibeli

Sosyal medya platformlarına yönelik düzenleme konusunda süregiden tartışmalar, bu platformların paylaşılan içerikten ne derece sorumlu olduğu ve benimsedikleri algoritmalar ile politikaların toplumsal etkileri üzerine yoğunlaşıyor. Örneğin, nefret söylemine yönelik politikalar genellikle şiddeti, ayrımcılığı, tacizi veya ırk, etnisite, din, toplumsal cinsiyet, cinsel yönelim, engellilik durumu veya koruma kapsamındaki başka kategoriler gibi belirli özelliklere dayalı nefreti destekleyen veya teşvik eden içerikleri yasaklamaya odaklanıyor. Ancak çeşitli sosyal medya platformları arasında politika farklılıkları da bulunuyor.

Statista'ya göre,¹³ Nisan 2024 itibarıyla dünya çapında en popüler olan ilk beş sosyal ağ platformu Facebook, YouTube, WhatsApp, Instagram ve TikTok idi. Türkiye'de ise en çok kullanılan sosyal medya platformları Twitter, Instagram, Facebook, WhatsApp ve TikTok oldu. Bunlar en çok kullanılan platformlar olarak, nefret söylemi ve dezenformasyonun yayılması için de en çok kullanılan mecralar olmuştur.

Sosyal medya platformlarının nefret söylemi politikaları, **kullanıcılar için güvenli bir ortam yaratırken ifade özgürlüğüyle de denge sağlamak** şeklinde ortak bir amaç güdüyor. Ancak bu platform politikalarının uygulanma biçimi ve detayları çok çeşitlilik arz ediyor.

Meta: Facebook, Instagram, Threads ve WhatsApp

Meta'nın Facebook, Instagram ve Threads için uyguladığı nefret söylemi politikası ırk, etnik köken, ulusal köken, engellilik durumu, dini inanç, sınıf, cinsel yönelim, cinsiyet, cinsiyet kimliği ve ciddi bir hastalık gibi koruma kapsamındaki özelliklere dayalı saldırıları yasaklayarak güvenli bir ortam yaratmayı amaçlıyor.¹⁴

¹³ **Statista.** (t.y.), *Aylık aktif kullanıcı sayısına (milyon) göre Nisan 2024 itibarıyla dünya çapında en popüler sosyal ağlar*, <https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/>

¹⁴ **Meta.** 7 Ocak 2025, *Nefret Dolu Davranışlar – Topluluk Standartları*. Meta Şeffaflık Merkezi. <https://transparency.fb.com/en-gb/policies/community-standards/hate-speech/>

Meta'nın nefret söylemi politikası, özgür ifade ile nefret söylemine karşı koruma arasında bir denge yaratabilmek adına iki kategori altında düzenlenmiştir:

- 1. Kategoride korunan özelliklere dayanılarak birey ve topluluklara yöneltilen şiddet içerikli söylemler veya şiddeti savunmayı ya da yüceltmeyi, insanları hayvanlara, patojenlere veya insanlık dışı başka yaşam formlarına benzeten insandışılástico söylemleri ve zararlı kalıp yargıları yayan ifadeler ciddi ihlaller olarak tanımlanır.
- 2. Kategori ise fiziksel, zihinsel veya ahlaki açıdan yetersiz olduđu iddiasıyla topluluklara yöneltilen ve temizlik, dış görünüş, zekâ, eğitim, ruh sağlığı, cinsel davranış ile ilgili aşağılayıcı terimleri, koruma altındaki özelliklere yönelik nefret, tiksinti veya dışlama içeren ifadeleri ve, romantik ayrılıklar gibi belirli bağlamlar dışında, bu özelliklerden ötürü kişileri hedef alan tahkir veya küfürleri de içeren aşağılayıcı ve küçümseyici ifadeleri kapsar. Ayrıca, korunan özelliklere dayanılarak toplulukların ayrılmasını savunan eylem çağrılarını ve beyanları veya bu grupların siyasi katılım hakkından, ekonomik haklardan veya kamusal alanlara erişimden mahrum bırakılması gibi dışlama ve segregasyon ifadelerini de bu kapsamda tanımlar.

Meta, insandışılástico dil ve zararlı kalıp yargılar içeren yasaklanmış belirli içerik türleri saptayıp nefret söylemi politikası dahilinde koruma altındaki özelliklere sahip kişilere saldıran tahkir ve tezyif içerikli kelimeleri yasaklasa da, nefret söylemini kınamak, farkındalık uyandırmak veya bu kelimelerin kişinin kendine yönelik olarak ya da kişiyi güçlendirici şekilde kullanıldığı durumları istisna kapsamında tutuyor. Ancak bunun için niyetin açıkça belirtilmiş olması şart, zira niyetin muğlaklığı içeriğin kaldırılmasına yol açabilir. Dahası, ihlal içerikli unsurları eleştiren veya bunlarla alay eden hiciv türü de kapsam dışında kalabilir.

Instagram

Instagram Topluluk Kuralları kapsamında,¹⁵ “Instagram topluluğunun diğer üyelerine saygı duyun” şeklinde bir alt-başlık yer alıyor. Buna göre “İnandırıcı tehditler veya nefret davranışı barındıran içerikler, aşağılama veya küçük düşürme amacıyla şahısları hedef alan içerikleri, birine şantaj yapma veya taciz etme amaçlı

¹⁵ **Instagram.** (2024). *Topluluk Kuralları*. Instagram Yardım Merkezi.
https://www.facebook.com/help/instagram/477434105621119/?cms_id=477434105621119

*kişisel bilgiler ve tekrar tekrar gönderilen istenmeyen mesajlar kaldırılıyor”.*¹⁶ Instagram bu bakımdan geçerli olan tehdit ve nefret söylemi tanımları için Meta’nın web sayfasına bağlantı veriyor.

Instagram’ın politikası, kullanıcıların nefret söylemine karşı çıkmak amacıyla, niyetlerini açıkça ifade etmeleri kaydıyla konuya ilişkin zararlı örnekler vermesine müsaade edilebileceğini belirtiyor.

TikTok

TikTok nefret söylemi konusundaki politikasını kendi Topluluk İlkeleri kapsamında “Güvenlik ve Nezaket” başlığı altında tanımlıyor (Aralık 2024).¹⁷ Platformun politikası, nezaket kavramının, herkesin özündeki itibarın farkında olmayı ve eylemler, ifadeler ve üslup açısından saygılı olmayı içerdiğini vurguluyor. Güvenlik ve Nezaket başlığı “Nefret Söylemi ve Nefret Davranışları” şeklinde bir alt-başlık içeriyor; burada topluluk üyelerinin çeşitliliğine verilen önem dile getirilip, nefret söylemi, davranışı ve nefret ideolojileri yasaklanıyor. Bu politika, kast, etnik köken, ulusal köken, ırk, din, kabile, göçmenlik durumu, toplumsal cinsiyet, cinsiyet kimliği, cinsiyet, cinsel yönelim, engellilik ve ciddi hastalık gibi koruma altındaki özelliklere dayalı olarak kişi ve topluluklara yönelik saldırıda bulunan ifadeleri kapsıyor.

Bu politika ırksal üstünlük, kadın düşmanlığı, LGBTQ+ karşıtlığı ve antisemitizm gibi koruma altındaki özelliklerden birine dayalı ayrımcılık yapan nefret ideolojilerini ve inanç sistemlerini de kapsamına dahil ediyor.

TikTok izin verilmeyen içerikleri şöyle tanımlıyor:

Koruma altındaki bir özelliğe dayanarak şiddet, ayırma, ayrımcılık ve başka tür zararları teşvik etme; koruma altındaki bir grup üzerinde üstünlük iddia etmek veya herhangi bir nefret ideolojisini teşvik etmek; birini koruma altındaki özelliklerinden ötürü küçük düşürmek; koruma altındaki bir özellik ile ilişkili nefret dolu bir hakaret kullanmak; koruma altındaki gruplara zarar verdiği belgelenmiş olan tarihi olayları inkar etmek veya bu olayları olduklarından daha küçük ölçekli göstermek; koruma altındaki bir özelliğe sahip bir kişinin neden olduğu kötü eylemlerden bütün bir gru-

¹⁶ **Meta.** 22 Ocak 2024, *Zorbalık ve Taciz – Topluluk Standartları*. Meta Şeffaflık Merkezi. <https://transparency.meta.com/tr-tr/policies/community-standards/bullying-harassment/>

¹⁷ **TikTok.** 2024, *Güvenlik ve Nezaket*. TikTok Topluluk İlkeleri <https://www.tiktok.com/community-guidelines/tr/safety-civility>

bu sorumlu tutma; korunan özellikleri nedeniyle insanları hor gören veya varlıklarını reddeden içerikler paylaşma; kişinin tercih ettiği kimliği yerine eski adını veya cinsiyetini kullanarak kişiyi yanlış cinsiyetlendirme; nefret söylemini veya nefret içerikli ideolojileri teşvik eden ürünlerin ticaretine olanak sağlamak.

Meta olarak, “kendine referansla yapılan hakaretler” ve “nefret söylemini eleştirmek veya raporlamak amacıyla yapılan ifadeler” gibi içeriklere izin verilmektedir. TikTok, Topluluk İlkeleri kapsamında ayrıca “Şiddet Yanlısı ve Nefret Yayan Örgütler ve Kişiler” başlığına da yer veriyor; bu başlık altında “Şiddet yanlısı radikaller, şiddet yanlısı suç örgütleri, şiddet yanlısı siyasi örgütler, nefret yayan örgütler ve kitlesel şiddete neden olan kişiler de dahil olmak üzere şiddet yanlısı ve nefret yayan örgütlerin ve kişilerin bu platformda varlığına izin verilmediği” belirtiliyor.

X (Eski Twitter)

“Kurallar ve politikalar” başlığı altında, X (eski Twitter) nefret suçunu Güvenlik ve Siber Suçlar bölümü kapsamında “Nefret Davranışı” başlığıyla ele alıyor.¹⁸

Yukarıdaki politikalarda olduğu gibi, X’in politikası da ırk, etnik köken, ulusal köken, sosyal sınıf, cinsel yönelim, cinsiyet, cinsiyet kimliği, dini inanç, yaş, engellilik durumu ya da önemli hastalık temeline dayalı saldırılara karşı kişilerin korunması amacıyla hazırlandığını belirtiyor.

Ancak platformun ifade özgürlüğünü ve çeşitli bakış açılarını içeren kamusal sohbeti desteklemeyi amaçladığı özel olarak belirtildiğinden, diğer politikalardan farklıdır. Bu politika nefret söylemini haklara yönelik bir taciz olarak görüyor ve kullanıcılar birinin X politikasını ihlal ettiği kanaatine varmışsa bunun bildirilmesi için bir bağlantı veriyor.¹⁹

X, platformun “Gönderiler veya Direkt Mesajlar içinde aşağıdaki davranışlarla [yani X’in politikasını ihlal eden davranışlarla] bir kişiyi veya bir grup insanı hedef alan hesaplara yönelik raporları inceleyip önlem alacağını” belirtiyor. Politika daha sonra bu ihlalleri, nefret içerikli göndermeler, kışkırtma, iftiralar ve kinayeler, canavarlaştırma, nefret içerikli görüntüler ve nefret içerikli profil şeklinde şöyle detaylandırıyor:

¹⁸ X (eski Twitter). (Nisan 2023). *Nefret Davranışı*. X Yardım Merkezi
<https://help.x.com/tr/rules-and-policies/hateful-conduct-policy>

¹⁹ X (eski Twitter). (Nisan 2023). *Güvende kalma ve hassas içerikler*. X Yardım Merkezi.
<https://help.x.com/tr/forms/safety-and-sensitive-content/abuse>

- **Nefret içerikli göndermeler**, korumalı bir kategorinin birincil hedef veya kurban gösterildiği, tacizi hedefleyen şiddet biçimlerine ya da şiddet eylemlerine göndermeler yapan içerikle kişi veya grupları hedef alır. Soykırımlar ve linç eylemlerine referanslar bu göndermelere girer.
- **Kışkırtma**, platform içinde veya dışında korumalı kategorilere karşı taciz veya ayrımcılığı kışkırtan davranışları içerir; buna korku uyandırmak veya korkutucu kalıpyargılar yaratmak da dahildir. Bu tanım altında platformun “Şiddet Söylemi” politikasına doğrudan bağlantı da veriliyor.²⁰
- **Hakaret ve kinayeler**, itibar zedeleyen ya da olumsuz kalıpyargıları pekiştiren içeriği kapsar.
- **İnsandışılştırma (canavarlaştırma)** din, sınıf, yaş, engellilik, ciddi hastalık, ulusal köken, ırk, etnik köken, cinsiyet, cinsel kimlik veya cinsel yönelim nedeniyle belirli grupların canavarlaştırılması anlamına gelir.
- **Nefret içerikli görüntüler**, koruma altındaki kategorilere dayalı olarak başkalarına karşı kin ve düşmanlığı destekleme amacı taşıyan logo, sembol veya resimleri içerir; bunların platformda kullanımı yasaktır. Buna geçmişte nefret gruplarıyla ilişkilendirilen semboller ve nefret içeren semboller içerecek şekilde değiştirilen görüntüler de dahildir.
- **Nefret içerikli profil**, profil resminde, başlığında ve kullanıcı bilgilerinde nefret içeren görseller veya sembollerin kullanımını içerir; birini hedef alarak taciz etmek veya korumalı kategoriye karşı nefret ifade etmek de buna dahildir.

Platformun politikası, ihlallere karşı alınan tedbirlerin ihlalin ağırlığına ve kişinin ihlal siciline bağlı olarak, içeriği daha az görünür hale getirmekten, hesapların askıya alınmasına kadar varan bir çeşitlilik sergilediğini gösteriyor. Bu politika ayrıca, X’in ihlal bildiriminde bulunulmasını şart koştuğunu söyleyerek askıya alınan hesaplar için bir itiraz süreci öngörüyor.

YouTube

YouTube’un nefret söylemi politikası yaş, toplumsal sınıf, engellilik, etnik köken, toplumsal cinsiyet veya cinsel yönelim, uyruk, ırk, göçmenlik durumu, din, cinsiyet, önemli bir şiddet olayının kurbanları ve yakınları ve gazilik durumu gibi ko-

²⁰ X (eski Twitter). (Şubat 2025). *Saldırgan söylem politikası*. X Yardım Merkezi. <https://help.twitter.com/en/rules-and-policies/violent-speech>

ruma altındaki çeşitli özelliklere bağlı olarak kişi veya gruplara yönelik şiddet ve nefreti teşvik eden içerikleri yasaklıyor.

YouTube, ihlalleri bildirmek için bir yönerge sunuyor²¹ ve kullanıcıları bu ihlalleri nasıl bildirecekleri konusunda yönlendiriyor. Diğer platformlardaki ilkelerin aksine ve kendisinin bir video paylaşım sitesi olmasıyla uyumlu olarak, YouTube nefret söylemi politikasını bir video ile tanıtıyor.

YouTube'un politikası, şiddeti teşvik eden veya yukarıda anılan koruma altındaki statülerinden ötürü kişi ya da grupları tehdit eden, onlara karşı nefreti kışkırtan veya onları insan olmayan varlıklarla karşılaştırarak insandışılaştıran içeriklerin yayınlanmasını yasaklıyor. Ayrıca, korunan gruplara yönelik şiddeti öven veya yücelten, nefreti teşvik edici hakaret veya kalıpyargılar kullanan, herhangi bir grubun üstün veya aşağı olduğunu iddia eden, nefret içeren üstünlükçülüğü veya bu tür ideolojilere sapmayı teşvik eden, belirli gruplara yönelik komplocu iddiaları yaygınlaştıran, tarihte yaşanmış büyük şiddet olaylarını veya kurbanlarını inkâr eden ya da önemsiz gösteren veya insanlara duygusal, romantik yahut cinsel cazibelerine dayalı olarak saldıran içerikleri de yasaklıyor. Bu politika, video, video açıklaması, yorum, canlı yayın ve içerikteki harici bağlantılar da dahil olmak üzere YouTube tarafından sunulan diğer her türlü ürün veya özellik açısından geçerlidir.

Tablo 4. Sosyal medya platformlarının nefret söylemi konusundaki politikalarına ilişkin özet

Platform	Politika ilkeleri	Yasaklanmış içerik	İzin verilen içerik	Uygulama	Ek notlar
Facebook	İnsandışılaştırıcı dili, zararlı kalıpyargıları yasaklar; üç kategoriye ayrılır.	1. Kategori: Şiddet içeren, canavarlaştırıcı söylem; 2. Kategori: Aşağılayıcı beyanlar	Farkındalık, nefret söyleminin kınanması veya kendine dönük kullanım istisnaları	Ağırlığa bağlı olarak kategorilere ayrılmış uygulama; içeriğin kaldırılması ihtimali.	Farklı ihlal türleri için detaylandırma yapılıyor.

21 YouTube. (2024). İçerik bildirme ve yaptırım. YouTube Yardım. <https://support.google.com/youtube/answer/2802027>

Instagram	Korunan özelliklere dayalı saldırıları yasaklar; pozitif, çoğulcu bir topluluk desteklenir.	İnandırıcı tehditler, özel şahısları hedef alan nefret söylemi, özel bilgilerin taciz amaçlı paylaşılması.	Niyeti açık olmak şartıyla nefret söylemi hakkında farkındalık uyandıracak tartışmalara izin verir.	İhlal konusu içeriğin kaldırılması; kullanıcı bildirim sistemi.	Tanımlar için Me-ta'nın genel politikasına bağlantılar.
WhatsApp	Uçtan uca şifreleme sebebiyle sınırlı izleme yetkisi; kullanıcı bildirimine odaklanır.	Kullanıcılar tarafından bildirilen veya şüpheli faaliyet sonucunda saptanan nefret söylemi yayan her türlü içerik.	Doğrudan içerik izlenmesi mümkün olmadığı için bu başlığın geçerliliği yok.	Kullanıcı bildirimlerine ve hizmet şartlarının ihlaline dayanarak hesapların yasaklanması.	Şifreli iletişim doğrudan içerik moderasyonunu sınırlandırıyor.
TikTok	Nefret içeren davranışı, söylemi ve nefret içerikli ideolojilerin teşvikini yasaklar.	Korunan özellikleri yüzünden kişi ya da topluluklara saldıran içerik, nefret içerikli ideolojiler, tarihsel olayların inkârı.	Kendine dönük olarak kullanılan hakaretler, nefret söylemi konusunda eğitim içeriği.	Nefret içerikli kuruluşları ve kişileri, şiddeti savunan radikalleri ve kitlesel şiddet faillerini yasaklar.	Belirli bir konuya bağlı olarak (örn. insan kaçakçılığı, taciz ve zorbalık) ek ilkeler içerir.
X (Twitter)	Korunan çeşitli özelliklere dayalı saldırılara karşı koruma sağlar; özgür ifadeyi destekler.	Kışkırtma, nefret içerikli göndermeler, hakaretler, canavarlaştırma, nefret içerikli görseller ve profil bilgileri.	Açıkça detaylı anlatılmamış, daha ziyade yasaklanan şeye odaklanıyor.	İçeriği daha az görünür kılmaktan başlayıp hesapların askıya alınmasına varan tedbirler.	İtirazlar ve ihlal bildirimleri için süreçler içeriyor.

YouTube	Korunan özelliklere dayalı olarak şiddeti veya nefreti teşvik eden içeriği yasaklar.	Nefreti kışkırtma, şiddet, canavarlaştırma, nefret içeren üstüncülük, önemli olayların inkârı.	Bağlamı net olarak verilen eğitim, belgesel, bilim veya sanat içerikleri.	İçeriğin kaldırılması, uyarı sistemi, kanal kapatma ve maddi yaptırım ihtimali.	Yorumlar ve canlı yayınlar da dahil olmak üzere tüm özellikler açısından geçerlidir.
----------------	--	--	---	---	--

Kendisini bir özgür ifade platformu olarak tanımlayan YouTube “net bir şekilde eğitim, belgesel, bilim ya da sanat amacı olan videolar için istisna” uyguluyor:²²

Bu tür istisnalara örnek olarak bir nefret grubuyla ilgili belgeseller yayınlanabilir. Belgeselde nefret söylemi bulunabilir ancak içeriğin belgesel amacı açıkça görülüyorsa, içerik nefret söylemini teşvik etmiyorsa ve izleyicilere belgeseli yapılan konunun ne olduğu ve neden anlatıldığıyla ilgili yeterli bağlam sağlanıyorsa içeriğe izin verebiliriz.

Bu politikanın ihlali içeriğin kaldırılması ile sonuçlanır ve içerik üreticisi durumdan eposta yoluyla haberdar edilir. İlk kez yapılan ihlallerde muhtemelen ceza verilmeden sadece uyarıda bulunulur, ancak uyarının 90 gün içinde kaldırılması için kullanıcının politika eğitimine katılması gerekir.²³ İhlaller devam ederse, kullanıcı kanalın kapatılması veya reklamlara ve diğer maddi özelliklere erişimin askıya alınması gibi daha ağır tedbirlere çarptırılır. YouTube ayrıca politikalarını ihlal etmeyen, ancak kaldırılma sınırına yakın veya bazı izleyicileri rencide edebilecek içerikler için yorumlar, önerilen videolar ve beğeniler gibi özellikleri sınırlandırabilir.

Bu raporda bahsi geçen politikalar bu platformların rapor yazıldığı tarihte geçerli olan en güncel politikalarını yansıtıyor; ancak bu yayın hazırlanırken Meta kendi politikalarını üç kez güncelledi ve YouTube ile TikTok da büyük değişikliklere

²² **YouTube.** (2024). *YouTUBE, topluluğu nefrete ve tacize karşı nasıl koruyor? Nefret söylemi politikasının uygulanmasıyla ilgili istisnalar var mı?*

<https://www.youtube.com/intl/ALL-tr/howyoutubeworks/our-commitments/standing-up-to-hate/#policy-exceptions>

²³ **YouTube.** (2024). *Nefret söylemi politikası.* YouTube Yardım.

https://support.google.com/youtube/answer/2801939?hl=tr&ref_topic=9282436&sjid=13736758284860870198-EU

gitti. Ocak 2025 başlarında Meta'nın CEO'su Mark Zuckerberg, şirkete ait platformlarda nefret söylemi ve üçüncü taraf doğruluk kontrolü konusunda önemli politika değişiklikleri yapılacağını duyurdu ve bu haber geniş tartışmalara yol açtı. Zuckerberg, Meta'nın bağımsız doğruluk kontrolü programını sürdürmeyeceğini, X tarafından benimsenen yaklaşıma benzer "Topluluk Notları" diye bir sisteme geçeceğini söyledi. Bu beyanlar uzmanlar ve sivil toplum örgütleri arasında, platformun dezenformasyon ve nefret söylemi akışını azaltma kapasitesi hakkında endişelere yol açtı. Önerilen "Topluluk Notları" profesyonel doğruluk kontrolünün hesap verebilirliğinden yoksun olabilir ve platformları nefret ve dezenformasyon kampanyalarına karşı daha savunmasız bırakabilir.

Bu gelişmelerin arka planında, Donald Trump'ın yeniden başkan olarak seçilmesinin ardından ABD'de siyasi açıdan gerginleşen atmosferin de payı var. Zuckerberg ve Elon Musk gibi önde gelen teknoloji CEO'larının yemin töreninde Donald Trump'ın arkasında dururken çekilen görüntüleri bu platformların ABD hükümetinin önceliklerine göre konumlanacağı hakkındaki kaygıları daha da arttırdı. Bu görüntüler, söz konusu platformların politikaları hakkında endişeli olan herkesin, bu platformları nihayetinde kimin yöneteceğini merak etmesine yol açıyor ve bu şirketlerin nefret söylemi ve dezenformasyona ilişkili politikaları hazırlarken ve tedbirleri uygularken ne kadar bağımsız kalabileceğine dair ciddi soru işaretleri uyandırıyor.

Nefret söylemi konusunda farklı politikaları inceledikten sonra şimdi bu söylemin tespiti ve analizinde pratikte yaşanan güçlükleri ele alacağız. Nefret söyleminin etkili bir şekilde konu edilmesi için politika tedbirlerinin anlaşılması tek başına yeterli olmayıp, öznellik, erişilebilirlik ve bağlamsal nüanslar bakımından sağlam tespit yöntemlerine ihtiyaç vardır.

Nefret söyleminin hedef aldığı grupların çeşitliliği ve bağlama göre değişkenliği düşünülecek olursa, en önemli meselelerden biri nefret söyleminin nasıl tespit edilip sınıflandırılacağı hususudur. Kapsamlı uluslararası tanımlara uygun veri kümeleriyle çalışırken bile, belli bir söylemin nefret söylemi kapsamına girip girmediğinin tespiti güç olabilir. Bu güçlük, söz konusu söylemin yaratabileceği sonuçların yanı sıra, üretildiği bağlam, yaratılmasından sorumlu olan kişinin kimliği, yansıttığı güç dinamikleri, hedef aldığı kitle, yaygınlaştırılma aracı, iletilme tonu ve nasıl algılandığı gibi birçok faktörün hesaba katılması gerekliliğiyle daha da artar. Ayrıca, söylemin üretildiği toplumun kültürel kodları ve belli kelimelerle ilintili tarihsel çağrışımlar gibi ölçülmesi zor unsurlar da nefret söyleminin tespitinde önemli rol oynar.

Nefret söyleminin analizinde farklı sınıflandırma yöntemleri kullanılsa da, tespit süreci halen insan emeğine dayanıyor. Kültürel bağlam ve güç dinamikleri hakkında bilgisi olan kişiler tarafından yapılan manuel analiz bile öznel sonuçlar verebilir. Ayrıca, sosyal medya platformlarının gitgide daha fazla kullanılması, nefret söylemine bu ortamlarda da daha sık rastlanılmasına neden oluyor. Bu platformlarda paylaşılan içeriğin hacmi düşünülürse, manuel analiz bu alanlarda çalışmak için gitgide daha yetersiz (hatta imkânsız) hale geliyor. İşte bu nedenle teknolojiye başvurma gerekliliği (veya zorunluluğu) doğmuştur. Ancak nefret söyleminin tespitininin yapay zekâ (YZ) ve makine öğrenmesi yoluyla otomatikleştirilmesiberaberinde mevcut ve yeni, zorlukları getiriyor. Örneğin, belirli bir veri kümesiyle eğitilmiş modellerin taşınırılığı, aynı tanımlar benimsense bile, farklı veri kümelerine uygulandığında daha düşük performansa yol açabilir. Bu durum veri dağılımında zaman, platform ve konu gibi faktörler açısından yaşanan farklardan kaynaklanır. Otomasyon, bu tür söylemlerin bağlama özgü ve nüanslı yapısını tam olarak kavrayamadığı içindir ki, doğal dil işleme (NLP) ve yapay zekâ araçlarını kullanan araştırmacılar sürekli olarak, bu teknolojilerde yaşanan hızlı ilerlemeye rağmen, tespit sürecini daha doğru kılacak yöntemler arıyor.

Nefret söylemi tespitinin otomasyonu, yüksek kaliteli veri kümelerine erişimin sınırlı olması nedeniyle de yavaş ilerliyor. Bu kümeler modellerin eğitilmesi açısından son derece önemli olup, doğruluğu doğrudan etkiliyor. Ne var ki, kısıtlayıcı sosyal medya politikaları ve sınırlı araştırma kaynakları nedeniyle bu verileri elde etmek gitgide zorlaşıyor. Veri kümeleri iyi etiketlenmiş ve yüksek bir doğruluk oranına sahip olsa bile, uygulandıkları bağlama göre etkileri değişebiliyor. Bir bağlamda hedef alınan bir grup, başka bir bağlamda bu şekilde tanınmayabiliyor; bu da hem doğru hem de geniş uygulama alanına sahip modeller oluşturmaya zorlaştırıyor. Modellere hakim olan önyargılar da endişe yaratan bir husustur. Eğitilmiş modeller genellikle eğitimlerine zemin oluşturan verileri yansıtır; dolayısıyla, etiketlemeye tesir eden herhangi bir eğilim –lehçe kullanımı veya hakaret içeren kelimelerin yeniden sahiplenilmesi gibi– bu yolla pekiştirilip halihazırda nefret söyleminin hedef aldığı topluluklara daha da fazla zarar verebilir. Bunun yanı sıra, çalışmaların birçoğu toplumlara egemen olan kültürel veya sosyal gruplara yönelik verilere yoğunlaştığı için ötekileştirilen topluluklar yeterince temsil edilemiyor. Verilerdeki bu noksanlıklar, tespit modellerinin nefret söylemini tüm kapsamıyla yakalayabilmesine engel oluyor.

Bir başka güçlük de açıklanabilirlik konusunda yaşıyor. Yapay sinir ağlarına bağlı yaklaşımların kararları genellikle şeffaflıktan yoksun olduğu için belirli bir mesajın neden nefret içerikli olarak etkilendiğini anlamak güçleşiyor. Özellik ataması veya dikkate dayalı mekanizmalar gibi çözümler önerilmekle birlikte, bunların uygulanmasında da halen sorunlar var. Model kararlarının şeffaflığının artırılması hem doğruluk hem de etik uygulama açısından kritik önemde olduğundan, daha fazla araştırmaya ihtiyaç duyuluyor.

Bu meseleleri ele almak amacıyla, raporun ilerleyen bölümleri nefret söylemi ve diğer ilgili kavramların tespitinde dijital teknolojilerin kullanımını özetlemeyi amaçlamaktadır.

Nefret söylemi tanımlarındaki çeşitliliğin veri setleri üzerindeki etkileri

Arzucan Özgür

Sanal sosyal platformlarda kullanılan nefret söylemlerindeki artış, nefret söylemini saptayıp bununla mücadele etmek için otomatikleştirilmiş yöntemler geliştirilmesini gerekli kılıyor, zira bu paylaşılan veri miktarı manuel yöntemlerle çalışılmasını neredeyse imkânsız hale getirdi. Otomasyon yöntemlerinin kalitesi ve güvenilirliği ise büyük oranda bu modellerin eğitilmesinde kullanılan veri kümelerine bağlıdır. Nefret söyleminin tespitinde veri kümelerini etiketlemek, bu kavramın olası öznel niteliği sebebiyle güç bir iştir.

Nefret söylemini tanımlamadaki güçlük, veri etiketleme ilkelerini saptamayı da güçleştiriyor. Net olmayan yönergeler etiketlemede öznelliği artırıyor ve yanlış veri kümeleri ile modeller oluşturulmasına yol açıyor. Dahası, etiketleme uzmanları arasında tatminkâr bir uzlaşıya varılmasını da zorlaştırıyor (Madukwe vd., 2020); ki bu da tutarsız şekilde etiketlenmiş veri kümeleri yaratıyor. Öyle ki, aynı veri kümesindeki aynı tweet'lere farklı etiketler atandığı bile görüldü (Awal vd., 2020). Bu zorluğu aşabilmek için araştırmacıların kullandığı stratejilerden biri, birden fazla sayıda etiketleme uzmanının etiket üzerinde mutabık kaldığı verilere (örn. tweetler) yönelmektir. Örneğin, sosyal medya metinlerinde nefret söylemi tespitine yönelik yakın tarihli ortak işlerde –ekiplerin aynı sorun üzerinde çalıştığı ortak çalışmaya dayalı değerlendirmeler– eğitim ve test kümelerine sadece en az üç etikete sahip tweetler dahil ediliyor (Arın vd., 2023; Uludoğan vd., 2024).

Nefret söylemi için etiketlenmiş birçok veri kümesi bulunmakla birlikte, nefret söylemi tanımlarının ve ayıklama ilkelerinin farklı olması sebebiyle, bir veri kümesi üzerinden eğitilmiş makine öğrenmesine dayalı modeller başka kümeler üzerinden test edildiğinde düşük performans gösteriyor. Yani nefret söyleminin farklı tanımları ve farklı veri ayıklama yönergeleri geliştirilen modellerin genelleştirilebilirliğini sınırlandırıyor (Swamy vd., 2019).

Nefret söyleminin insan eliyle etiketlenmesi: Faydalar ve güçlükler

Ayşecan Terzioğlu, Didar Akar

Nefret söyleminin tespiti ve etiketlenmesi detaylı ve karmaşık bir süreçtir; iletilen mesajın bağlamına ilişkin verilerin sınırlı olması, ayıklama şemalarındaki varyasyonlar, veri birleştirme yöntemlerinin çeşitliliği ve veri kalitesine ilişkin sıkıntılar gibi çeşitli faktörlerden etkilenir. Ayıklama sürecinde bir ifadenin nefret söylemi içerip içermediği veya söz konusu ifadedeki saldırganlık derecesi konusundaki anlaşmazlıklar genellikle ifadenin muğlaklığına veya etiketleme uzmanının kişisel görüşlerine, eğilimlerine ve tecrübelerine bağlanır (Novak vd., 2022).

Bu zorluklar, alandaki uzmanları hem etiketleme konusundaki uzlaşının ölçüsünü hem de tespit modelinin genel anlamda kalitesini değerlendirmek üzere performans kıstasları geliştirmeye yöneltmiş bulunuyor. Keza son yıllarda etiketleme uzmanlarının sosyoekonomik geçmişi ile etiketleme örüntüleri arasındaki ilişkiye odaklanan akademik çalışmaların sayısı da artıyor. Geçmişten getirilen bu farklılıkların etkisini en aza indirebilmek için, etiketleme uzmanlarına geniş çeşitliliğe sahip somut örnekler üzerinden pek çok eğitim veriliyor. Bu uzmanların iki kişi halinde çalışması veya tartışmalı ifadeleri ele almak üzere düzenli olarak bir araya gelmesi de diğer olası çözümler arasında yer alıyor (Kocon vd., 2021).

Etiketleme uzmanlarının mutabakatı genellikle “altın standart” olarak idealize edilse de, bu uzmanların farklı geçmişlerden gelmesi ve farklı bakış açılarına sahip olması dolayısıyla mutlak bir görüş birliğine ulaşılmasının mümkün olmadığını biliyoruz. Yeterli bağlam bilgisi bulunmaması ve uzmanların fikir ve tutumlarının zaman içerisinde değişen niteliği gibi, etiketleyiciler arasında ve hatta bir etiketleyicinin kendi içinde dahi tutarlı olamıyor olması işleri daha da karmaşıktırıyor. Dolayısıyla uyuşmazlıkların en aza indirilmesi ve etiketleme konusunda tutarlılığın artırılması için somut örneklerle desteklenen net ve ayrıntılı kılavuz ilkelere ihtiyaç bulunuyor (Aroyo ve Welty, 2015).

Nefret söylemine karşı yapay zekâya dayalı çözümler geliştirmeyi amaçlayan projelerde, etiketleme uzmanlarının uygulamaları sonuçta ortaya çıkan algoritmaların performans ve doğruluğunu doğrudan ve temelden etkiliyor, zira bu algoritmaları eğitecek veri kümesi bu uzmanların kararlarıyla oluşturuluyor. Ne

var ki etiketleme süreci hiç basit olmadığı gibi, kolay bir yöntemi de yok. Süreci karmaşıklaştıran başlıca unsurlardan biri de, etiketleme uzmanlarının kimliği ile karar verme süreçleri arasındaki ilişki.

Uzmanın kimliğinin bu süreçte hesaba katılması gereken bir boyut oluşturması şaşırtıcı değil. Tarafsız bir unsur olmadığı için nefret söyleminin nasıl algılanıp sınıflandırıldığını belirliyor. Uzmanın etnik, siyasi ve ailevi geçmişi onu belirli türden nefret söylemlerine daha duyarlı kılarken, bazı söylemleri göz ardı etmesine de yol açabilir. Benzer biçimde, mesleki veya eğitim geçmişi de karar verirken farklı hususları göz önünde bulundurmasına sebep olabilir. Örneğin yapay zekâ eğitim prosedürüne aşina bir uzman çok net olmayan bir nefret söylemi örneğini etiketlemekten imtina edebilir, zira böyle bir etiketin algoritmayı tarafsız bir söylemi nefret söylemi kategorisinde sınıflandıracak şekilde eğitmesinden ve bunun sonucunda yanlış pozitiflere yol açmaktan çekinebilir. Bu nedenledir ki, epistemik bir duruş olarak nesnellik etiketleme uzmanları açısından arzulanan bir şey olmakla birlikte, sarıh veya yeterli bağlam bilgisi olmadığına muğlak bir şekilde ifade edilmiş nefret söylemlerinin gözden kaçması gibi dezavantajları da beraberinde getirir.

Etiketleme uzmanlarının geçmişlerinden getirdikleri farklılıklar bir yana, uzmanların mutabakatını sağlamanın önündeki bir başka büyük engel de dilsel olmayan işaretleri yorumlamadaki farklardır. Bizler insan olarak bir metni yorumlarken dilsel içerikten daha fazlasını kullanırız. Görüntüler, memler, semboller, noktalama işaretleri, emojiler vs. gibi görsel öğeler de anlam yaratma sürecinde önemli bir rol oynar. Ayrıca metnin harekete geçirdiği çoklu dizinsellik seviyelerinin anlaşılmasında genel kültür, metinlerarasılık ve üstü kapalı göndermeler de devreye girer.

Bahsedilen bu güçlükler veri ayıklama ve etiketleme sürecine yönelik disiplinlerarası bir yaklaşımı gerekli kılıyor; dilbilim, sosyal bilimler ve sayısal alanlardan aktarılabilecek fikir ve anlayışlar birleştirilerek etiketleme metodolojilerinin geliştirilmesi ve model performansının iyileştirilmesi mümkün olabilir.

Nefret söylemi tespit modellerinde etiketleyicinin kimliğinin önemi

Claudia von Vacano

Sağlam ve şeffaf bir makine öğrenmesi modeli oluşturmak, veri temsili, etiketleme, yanlılık ve açıklanabilirlik gibi birçok kilit unsurun takibini içerir. Bu unsurların her biri, modelin güvenilirliğini ve adillliğini sağlamakta kritik rol oynar. Her başarılı modelin temelinde veri yatar. Bir makine öğrenmesi modelini eğitmek için kullanılan verilerin, modelin işletileceği gerçek dünyadaki senaryoları temsil edebilmesi son derece önemlidir. Bunun için modelin karşılaşması beklenen vaka çeşitliliğini ve değişkenleri doğru şekilde yansıtması gerekir. Bu temsili sağlayamayan bir veri kümesinin kullanılması yanlılığın pekişmesine yol açabilir. Bunu önlemek için, veri toplama süreci kapsayıcı bir yaklaşımla yürütülmeli, farklı demografik grupları, koşulları ve değişkenleri içeren geniş bir veri yelpazesine sahip olmalıdır. Azınlıkta kalan ve ötekileştirilen grupların verilerini daha fazla örneklem (*oversampling*) gibi tekniklerle de denge sağlamak mümkün olabilir.

Amaçlanan tasarıma uygun bir veri kümesi hazırlandıktan sonra karşılaşılan zorluk, ayıklama/etiketleme sürecidir. Etiketleme uzmanları, makine öğrenmesi modelleri için veriyi etiketlendirir, fakat bu uzmanların yorumları yanlılık içerebilir. Bu eğilimler uzmanın kültürel arkaplanı, ırkı/etnisitesi, dini, ulusal kökeni/vatandaşlık/göçmenlik durumu, toplumsal cinsiyet kimliği, cinsel yönelimi, yaşı, engellilik durumu veya dezavantajlı gruplara dair koruma altındaki başkaca kategorilerden kaynaklanabilir. Kişinin bu benzersiz kimlik özellikleri, yaşam deneyimini ve dünyayı algılama biçimini şekillendirir. Karşılığında da bu bakış açıları sosyal medya platformlarındaki belirli bir yorum gibi verileri nasıl algıladığı, yorumladığı ve etiketlediği üzerinde etkili olabilir. Örneğin, *belirli bir etiketleme uzmanı*, kişisel ve kültürel bağlamından ötürü olumsuz bir çağrışımı fark etmeyebilir. Bu değişkenlik, modelin öğrenme sürecini doğrudan etkileyebilir.

Bunu aşmak için kullanılan etkili bir metodoloji, Madde Tepki Kuramı (MTK) gibi bir çerçeveyi “Derin Öğrenme” ile birleştirmektir. Mark Wilson’ın Ölçüm İnşası yaklaşımı etiketleme uzmanlarının ürettiği verilerin analizine yönelik yapılandırılmış bir yöntem sunar. Uzmanın belirli bir öğeyi, o öğenin özelliklerine ve kendi sahip olduğu eğilimlere dayalı olarak belirli bir biçimde etiketleme olasılığını anlamaya yardımcı olur. Madde Tepki Kuramını yaygınlaştırarak, veri etiketleme sürecinde belirli bazı eğilimleri anlayıp diğerleriyle dengelemek suretiyle denet-

leyebiliriz. Örneğin bir veri kümesini etiketleyen uzmanlar daha ziyade beyaz erkeklerden oluşuyorsa, azınlıkta kalan gruplara ilişkin yorumları güçlendirebiliriz. Böylece modelin hassasiyet seviyesini arttırmış oluruz.

“Etiketleme Uzmanının Kimlik Hassasiyetini Madde Tepki Kuramıyla Değerlendirmek: Nefret Söylemi Korpusu Vaka Analizi” başlıklı araştırmada (Sachdeva vd., 2022) özel olarak nefret söyleminin tespit ve ölçümüne yönelik makine öğrenmesi algoritmalarının eğitilmesinde kullanılan bir veri kümesindeki örüntüleri ayıklayan etiketleme uzmanının kimliğinin yarattığı etkiyi inceledik. Uzmanların demografik kimliklerinin farklı nefret söylemi türlerine yönelik hassasiyetlerini modelleyip ölçebilmek için Madde Tepki Kuramından yararlandık.

Biz bu araştırma çerçevesinde, uzmanın muhakemesini yanlı olarak görmek yerine, hassasiyet olarak kavramayı öneriyoruz. Bu yaklaşım nefret söyleminin etiketlenmesi gibi işlere öznelliğin dahil olduğunu, yani uzmanların kişisel kimliklerinin içeriğe dair farklı algı ve yorumlara yol açabildiğini kabul ediyor.

Uzmanın hassasiyetini ölçmek amacıyla MTK yönteminden yararlanmak, uzmanlar arasındaki bireysel farklılıklara daha ayrıntılı bir açıdan bakılmasını sağlar. Bu metodolojik tercih, bir uzmanın doğru ve yanlış şeklindeki ikili kavramın ötesinde kimliği dolayısıyla belirli bir etiketleme yapma olasılığını değerlendirmemize imkân sağlıyor.

Çalışmamızda yaklaşık 10.000 etiketleme uzmanı tarafından etiketlenmiş 50.000’den fazla sosyal medya yorumunu inceledik. MTK kullanılarak yapılan bu araştırma, uzmanların kendileriyle özdeşleştirdikleri grupları hedef alan nefret söylemlerine karşı daha fazla hassasiyet gösterme eğiliminde olduğunu ortaya koyuyor. Hassasiyet derecesini farklı açılardan ölçmek için üç MTK tekniği kullanıldı ve uzmanların ırkının çeşitli ırksal grupları hedef alan nefret söylemi algılarıyla büyük oranda örtüştüğü görüldü. Bu analiz Afrikalı Amerikalı uzmanların, siyahları hedef alan yorumları başka ırkları hedef alan yorumlara kıyasla nefret söylemi olarak saptama ihtimalinin çok daha yüksek olduğunu gösteriyor.

Araştırmanın bulguları, makine öğrenmesi modellerinin tasarlanması ve uygulanmasında uzmanın kimliğini dikkate almanın önemini vurguluyor; zira bu kimlik özellikleri eğitim verilerine ve dolayısıyla algoritmaların davranışına tesir edebiliyor. Uzmanın hassasiyetini anlamak algoritmadaki eğilimleri en aza indirip adilliği geliştirmeye yardımcı olacaktır. Gelecek araştırmalarda uzmanların kimliğinin başka bağlamlarda başka içerik türleriyle nasıl etkileşime girdiği incelenmelidir. Otomasyon sistemlerin adilliğini ve doğruluğunu artırmak için veri kümesi yaratma ve algoritma tasarımında çeşitli uzmanların kimliklerinin daha iyi şekilde temsil edilerek göz önünde bulundurulması gerektiği kanaatindeyiz.

Hrant Dink Vakfı son on yılda gitgide daha karmaşık bir hal alan medya ortamlarına ve kamusal söylemin şekillendirilmesinde dijital platformların artan rolüne uyum sağlamak için nefret söylemi izleme metodolojisini sürekli olarak geliştirdi. İlk başta sadece yazılı basına odaklanan izleme çalışmaları, Türkiye’de nefret söylemindeki örüntü ve anlatıların anlaşılabilmesi bakımından kritik bir zemin oluştuyordu. Ancak kamusal tartışmaların ana mecrası gitgide dijital ortamlara kaydıçça, geleneksel izleme yöntemlerinin sınırlılıkları daha görünür oldu. Dijital içeriğin geniş hacmi, zararlı söylemin yayılma hızı ve nefret söyleminin değişen niteliği –örtük veya dolaylı ifadeler de dahil– nefret söylemini daha etkin bir şekilde saptayıp analiz etmek için yeni yöntem ve teknolojik araçların benimsenmesini gerektirdi.

Dijital içeriğin artan hacmi ve hızı manuel izlemeyi sürdürülemez hale getirirken, nefret söyleminin karmaşıklaşması ve süratle yayılması daha sofistike ve ölçülebilir bir yaklaşımı mecbur kıldı. Vakıf bu ihtiyaçlar doğrultusunda Boğaziçi Üniversitesi ve Sabancı Üniversitesi ile işbirliği yaparak, Türkçe ve Arapça metinlerde nefret söylemini tespit ve analiz etmek üzere makine öğrenmesi temelli **pari** aracını geliştirdi ve çalışmalarını sosyal medyayı ve yapay zekâya dayalı tespiti içine alacak şekilde genişletti.

Bu geçiş teknolojik bir değişimden ibaret olmayıp, Vakfın nefret söylemi izleme konusunda on yıllık deneyimine dayalı yoğun tartışma, araştırma ve düzenlemelerin de sonucudur. Yapay zekâya dayalı tespitte kullanılan nefret söylemi kategorileri tespit amacıyla yapılandırılmış bir çerçeve sunarken, bu söylemdeki dilsel, sosyal ve bağlamsal farkları da yakalayabiliyor. Ayrıntılı tartışmalar, disiplinlerarası araştırmalar ve insan haklarına dayalı bir yaklaşımla geliştirilen kategoriler, ifade özgürlüğüne vurguyu önceliklendirirken nefret söylemi yelpazesinin tamamını etkili şekilde kavramayı sağlıyor.

Dijital Teknolojileri Kullanarak Nefret Söylemi ve Ayrımcılıkla Mücadele projesinin bir parçası olarak geliştirilen yapay zekâ aracı **pari**, şu dört nefret söylemi kategorisini içeren ayrıntılı bir nefret söylemi izleme metodolojisiyle manuel veri etiketlenerek eğitildi:

- **Abartma, Yükleme, Çarpıtma ve Genelleme**

Olumsuz genellemeler, yanlış bilgi veya kişilerin eylemlerine ya da münferit olaylara dayanılarak tüm bir topluluğu hedef alan yüklemeler yoluyla üretilen nefret söylemi.

- **Küfür, Hakaret, Aşağılama ve İnsandışılaştırma**

Bir grubu hedef almak için kullanılan doğrudan aşağılayıcı, canavarlaştırıcı ve küçültücü dil.

- **Düşmanlık/Savaş Söylemi, Şiddet, Saldırı veya Yaralama Tehdidi**

Bir grubu tehdit veya düşman olarak gösteren, düşmanlığı besleyen ve çatışma veya zararı ima eden söylem.

- **Simgeleştirme**

Kimlik ifadesinin kendisinin, bir grubu istenmeyen bir özellik ile olumsuz şekilde ilişkilendirmek ya da kimliklerini aşağılamak amacıyla kullanılması.

Yapay zekâ yoluyla izleme insana ait uzmanlığa duyulan ihtiyacın yerine geçmez. Vakıf, tespit yöntemlerinin doğruluğunu sağlamak, bağlamı anlamak ve bu yöntemlerin etik şekilde uygulanması için manuel denetimi sürdürüyor. Nefret söyleminin etkili şekilde saptanması, izlenmesi ve analizi için geleneksel araştırma biçimlerini dijital inovasyonla birleştirerek, dijital ortamlarda hak temelli, kapsayıcı söylemleri teşvik ederken daha net belgeleme ve müdahaleye imkân sağlayan bütüncül bir yaklaşıma ihtiyaç vardır.

Doğal Dil İşleme ile nefret söylemi tespiti için mevcut kategoriler

Roser Morante

Nefret söyleminin tespiti yaygın bir Doğal Dil İşleme (NLP) işi haline gelmiş bulunuyor (Fortuna & Nunes, 2018; Jahan & Oussalah, 2024). Bu anlamda ilk kümelerden biri 2012 yılında yaratılmış olup (Warner & Hirschberg, 2012), son yıllarda yaratılan veri kümelerinin sayısı hızla arttı (Poletto vd., 2021). Bu ilk veri kümelerinin yaratılmasından bu yana çok sayıda veri kümesi oluşturulması, muhtemelen nefret ifadelerinin her geçen yıl daha da yaygınlaşarak toplumun birçok kesimine ulaşıyor olmasını yansıtıyor.

Peki nefret söylemi NLP bünyesinde nasıl tanımlanıyor? Schmidt ve Wiegand (2017), “Doğal Dil İşleme kullanılarak Nefret Söylemi Tespitine ilişkin bir Araştırma” başlıklı makalelerinde, kullanıcılar tarafından yaratılan çok çeşitli aşağılayıcı içeriğe yönelik geniş bir şemsiye terim oluşturduğu için “nefret söylemi” terimini tercih etmişlerdi. Nefret söylemini “ırk, renk, etnisite, toplumsal cinsiyet, cinsel yönelim, uyruk, din gibi kimi özelliklere veya başkaca özelliklere dayalı olarak bir kişi veya grubu aşağı gören her türlü iletişim biçimi” olarak tanımlayan Nockleby’nin (2000) yaklaşımını benimsemişlerdi. Gibert vd. (2018) İngilizce internet forum gönderilerinde cümle düzeyinde etiketlenmiş, kamuya açık ilk nefret söylemi kümesini sunan çalışmalarında, ilgili literatürün çoğunda (Nockleby, 2000; Djuric vd., 2015; Gitari vd., 2015; Nobata vd., 2016; Silva vd., 2016; Davidson vd., 2017) nefret söyleminin şu şekilde tanımlandığına işaret etmişlerdi: a) kasıtlı olan, b) belirli bir insan topluluğunu hedef alan ve c) söz konusu topluluğun gerçekten veya algısal düzeyde kimliğini oluşturan unsurlardan beslenen bir saldırı.

Nefret söylemi tespit modelleri üzerine çalışan Röttger vd. (2021) ise şu tanımları tercih ediyor: “Koruma altındaki bir topluluğu veya o topluluğun bir parçası olduğu için üyelerini hedef alan çirkin sözler.” Bu araştırma kapsamına giren koruma altındaki topluluklar bu alana ait başka belgelerde sözü edilen gruplardan farklıdır: “Koruma altındaki toplulukları yaş, engellilik, cinsiyet kimliği, ailevi durum, hamilelik, ırk, ulusal veya etnik köken, din, cinsiyet veya cinsel yönelim temelinde tanımlıyoruz.”

Avrupa Birliği Temel Haklar Ajansı, gönderi içeriğinin nefret yüklü olması temelden hareketle bir nefret söylemi tipolojisi tanımlıyor. Bu çalışma birbirini dışlamayan beş temel çevrimiçi nefret kategorisi belirlemiştir:

- **Şiddeti, ayrımcılığı veya nefreti kışkırtma:** Kitleyi şiddet eyleminde bulunmaya; korunan (veya algılanan) bir özellikten ötürü birine farklı muamele etmek anlamına gelen ayrımcı bir tarzda davranmaya; konuşma ve yazı da dahil olmak üzere nefret içeren bir şekilde hareket etmeye açıkça teşvik eden veya cesaretlendiren içerik anlamına gelir.
- **Karalama:** Bir veya daha fazla kişinin, belirli bir gruba (algılanan) mensubiyetiyle bağlantılı olarak, sıfatına, kişiliğine veya itibarına yönelik ya da Avrupa Konseyi tavsiye kararında belirtildiği üzere, “gerçek ya da atfedilen kişisel [korunan] özellikleri nedeniyle yapılan” bir saldırı.
- **Saldırgan dil:** Korunan özelliklere atıfta bulunan aşağılamalar gibi “incitici, küçültücü veya yakışsız/müstehcen” ifadeler.
- **Olumsuz kalıpyargılar:** Korunan özelliklerle ilişkili olarak “bir sosyal gruba veya üyelerine atfedilen ve olumsuz anlam yüklenen” kimi olumsuz özellik ve nitelikler.
- **Nefret içerikli başka içerikler:** Nefret içeren ideolojilere veya Holokost’un inkârına destek içerebilir.

Nefret söylemini tanımladıktan sonra şimdi nasıl kategorize edildiğine bakalım. NLP bünyesinde nefretin etiketlenmesine yönelik yaklaşımlar genel olarak şu üç eğilimi benimsiyor:

- İkili sınıflandırma (nefret içermek – içermemek)
- Nefret söylemi – aşağılayıcı veya saldırgan dil
- Farklı nefret söylemi türleri

Ancak aşağıda vereceğimiz örnekler araştırmanın amaçları doğrultusunda bu sınıflandırmaların çeşitlendirilebileceğini gösteriyor.

Warner ve Hirschberg (2012) paragraflara atfedilebilecek yedi etiket kategorisi tespit etmişti: Yahudi karşıtı, Siyah karşıtı, Asyalı karşıtı, kadın karşıtı, Müslüman karşıtı, göçmen karşıtı ve diğer nefret türleri. Nefret söylemi biçimlerinin uygulanan kalıpyargılar teşhis edilerek birbirinden ayırt edilebileceğini savunuyorlardı. Adı geçen yazarlara göre her bir kalıpyargının nefret ifadesinde kullandığı, tek kelimelik sıfatlar, deyimler, kavramlar, metaforlar ve karşılaştırmalar gibi, birtakım dilsel özellikler mevcuttu. Örneğin Hispanik karşıtı söylemde sınır geçişi veya yasal kimliklendirmeye atıf yapılabilir. Kwok ve Wang (2013) tweetleri ırkçı, siyah karşıtı olan ve olmayan diye etiketlendirirken bu yolu izlemişti.

Waseem ve Hovy (2016) yaklaşık 16.000 tweetten oluşan bir veri kümesi ırkçı, cinsiyetçi veya hiçbirini olarak sınıflandırmıştı. Musto vd. (2016) Twitter'dan coğrafi etiketli veriler toplayarak homofobik, ırkçı, cinsiyetçi, antisemitik veya engelli karşıtı beş farklı nefret söyleminin konumlarını gösteren bir Nefret Haritası çıkarmıştı. Del Vigna vd. (2017) ise altı kategori belirlemişti: din, engellilik, sosyal statü, siyaset, ırk, cinsiyet ve toplumsal cinsiyet meseleleri ve diğer. Gambäck ve Sikdar (2017) tweetleri önceden belirlenmiş şu dört kategoriden biriyle etiketlemişti: ırkçılık, cinsiyetçilik, her ikisi (ırkçılık ve cinsiyetçilik) ve nefret içermeyen söylemler. "Her ikisi" kategorisi konunun kesişimselliğini gösterdiği için oldukça önemlidir.

Gibert vd. (2018) İngilizce internet forum gönderilerinden oluşan kamuya açık bir veri kümesini nefret ifadesi taşıyıp taşımadığına dair cümle düzeyinde etiketledi. Cümleler belirtilen üç koşulu taşıyorsa NEFRET etiketi eklendi: kasıtlı olan, belirli bir insan topluluğunu hedef alan ve topluluğun kimliğini oluşturan unsurlardan beslenen saldırı.

Sharma vd. (2018) zararlı dil içeren 9000 tweetten oluşan bir veri kümesi derleyip, bunları nefret içeriği niyetinin derecesine göre etiketledi ve zararlı dilin varlığı için belirli bir şiddet derecesinin üzerine çıkılması gerektiğini savundu:

- **I. Kategori:** Çoğu zaman hiçbir olumlu amaç taşımayan, topluma veya belirli bir gruba yöneltilmiş nefret söylemi. Yazarlar bir grubu hedef alan nefret veya şiddet içerikli davranışın bireysel suçlamalardan veya şiddetten daha ağır olduğunu öne sürüyor. Bağlama bakıldığında, konuşmacının belirli "-cilik"lerden (köktencilik) doğan hassasiyetleri kaşıma ve buna karşılık şiddet içeren olası bir tepkiyi kışkırtma niyeti taşıdığı açıktır.
- **II. Kategori:** Sanal atışma (suçlama, tehdit etme ve karşı çıkmak amacıyla agresif/provokatif dil kullanma) ve sözlü düello. Bu kategoriye giren nefret I. Kategoriden daha zayıftır. Hassasiyetler kaşınır, fakat şiddet içeren bir tepkiyi kışkırtacak seviyeye ulaşmaz. Bu kategoriye giren nefret, bir ideoloji veya topluluk değil de bir kişiyi hedef aldığında son derece provokatif olabilir.
- **III. Kategori:** Çoğu zaman bir kişiye yönelik olup, bir topluluğu hedef almayabilen orta düzeyde provokatif mesajlar. Bu kategori genelde trolleme, ironi veya alay bağlamında saygısızlık veya müstehcenlik barındıran kelimeler kullanır.

Salminen vd. (2018) YouTube ve Facebook videolarında yer alan 5143 nefret içerikli ifadeyi manuel olarak etiketleyerek 13 ana kategori ve 16 alt kategoriden (toplamda 29 kategori) oluşan ayrıntılı bir sınıflandırma yarattı. Ana kategoriler

arasında dil türünü tanımlayan dört kategori (suçlamalar, şiddeti teşvik eden ifadeler, aşağılamak, küfür etmek) ve ifadelerin hedefini tanımlayan dokuz kategori yer alıyor:

- **Finansal güç:** Varlıklı insan ve şirketlere ve onlara ait imtiyazlara yönelik nefret. Manipülasyon yapma ve suç işleme niyetleriyle ilgili iddialar.
- **Siyasi meseleler:** Hükümete, siyasi parti ve hareketlere, savaşa, teröre, sistemin kusurlarına yönelik nefret.
- **İrkçilik ve yabancı düşmanlığı:** Siyahlara, beyazlara, Asyalılara yönelik ırkçı yorumlar. Kimi özellikler hakkında genellemeler ve mülteciler hakkında nefret yüklü yorumlar.
- **Din:** Nefretin hem konusu hem de hedefi olabilen, Yahudilik, Hristiyanlık, İslamiyet de dahil olmak üzere tüm dinler ve genel anlamda din hakkında.
- **Belirli ulus(lar):** Farklı ülkelere, sistemlere, insanlara (uyruk belirtilirse) ve göç, toprak ve egemenlik gibi belirli olaylara yönelik nefret.
- **Belirli kişi:** Sıradan insanlardan oluşabilecek belirli kişilere, siyasetçilere, milyonlara, ünlülere veya belirli haberlerle ilgili olarak başka kişilere yönelik nefret.
- **Medya:** Yozlaşmış medya tarafından kasten yapılan taraflı ve yalan haberler hakkında dile getirilen görüşler ve duygusal çıkışlar.
- **Silahlı kuvvetler:** Orduya, kolluk güçlerine ve bunların etik olmayan davranışlar da içerebilen faaliyet biçimlerine yönelik nefret.
- **Davranış:** Dünyaya, insanlığa, toplumun bazı kesimlerinin ahlaksız eylemlerine, cehalete, belirli eylemlerde bulunan ve belirli davranışlara sahip insanlara yönelik nefret.

Bosco vd. (2018) yaklaşık 6000 tweetten oluşan, göçmenlere yönelik nefret söylemi bakımından etiketlenmiş bir Twitter seçkisi sundu. Etiketleme şeması nefret söylemi, saldırganlık, rencide edici söylem, ironi, kalıpyargılar ve (deneysel olarak) şiddet yoğunluğu başlıklarını içeriyordu. Yazarlar nefret söyleminin tespitinde şu iki unsuru gözettiler: 1) bir topluluk veya bireysel özelliklerinden ziyade söz konusu topluluğa mensup addedilen bir kişi olarak hedef; 2) belirli bir hedefe yönelik nefretin veya şiddetin yayılması, kışkırtılması, teşvik edilmesi veya meşrulaştırılması ya da hedefin insandışılaştırılması, meşruiyetinin ortadan kaldırılması, incitilmesi veya sindirilmesini amaçlayan eylem veya ifadenin edimsel gücü.

Fortuna vd. (2019) 5668 tweetten oluşan Portekizce bir veri kümesi oluşturdu. Bu tweetler önce, uzman olmayan kişiler tarafından “nefret içerir” – “nefret içermez” olarak ayıklandı. Ardından, etiketleme uzmanları tweetleri 81 adet nefret suçu kategorisi içeren çok ayrıntılı bir hiyerarşik etiket şeması izleyerek sınıflandırdılar; bunlar arasında “şişman insanlar”, “şişman kadınlar”, “çirkin insanlar”, “çirkin kadınlar”, “erkekler”, “feministler”, “sol ideolojiye mensup insanlar” gibi, nefret söylemi sınıflandırmasında çok daha az değinilen kategoriler de mevcuttur. Bu çalışma farklı kategorilerde uzmanlar arasındaki uzlaşma derecesinin çeşitlilik gösterdiğini ortaya koyması bakımından önemlidir; bu sayede, nefret söyleminin belirli türlerinin diğerlerine nazaran daha zor sınıflandırılabileceği anlaşılmıştır.

Ortak işler için etiketlenen çok sayıda veri kümesi mevcuttur. 2018 tarihli Kaggle Toksik Yorum Sınıflandırması,²⁴ aşağıda belirtilen kategorilerde toksik davranış bakımından etiketlenmiş 150.000 Wikipedia yorumu içeren bir küme içeriyordu: toksik, ağır toksik, yakışsız/müstehcen, tehdit, aşağılama ve kimlik nefreti.

HatEval (Basile vd., 2019), SemEval değerlendirme girişimi çatısı altında örgütlenen ortak bir çalışmada, Twitter’da göçmenlere ve kadınlara yönelik nefret söyleminin çokdilli tespiti konusunu ele aldı. Bunun için tweetler kadınlara veya göçmenlere yönelik nefret ifadesi içerir ve içermez olarak etiketlenildi ve ardından, nefret içerikli olarak sınıflandırılan tweetler bu kez saldırgan olup olmadığı üzerinden etiketlenildi. Nefret Söylemi Tespiti işinin İtalyanca Twitter mesajlarında nefret içeriğinin tespitine eğilen ikinci edisyonunda tweetler nefret ifadesi taşıyan ve taşımayan olarak etiketlenirken, kullanılan kalıpyargı da işaretlendi.

2021-2024 arasında düzenlenen Sosyal Ağlarda Cinsiyetçiliğin Saptanması konulu EXISTS yarışmaları (Rodríguez-Sánchez vd., 2021, 2022; Plaza vd., 2023), tweetlerde cinsiyetçiliği tespiti odaklanmıştır. 2024 edisyonu²⁵ birçok iş/görev içeriyordu: önceki edisyonlarda olduğu gibi cinsiyetçilik tespiti (ikili sınıflandırma) ve cinsiyetçiliğin sınıflandırılması. Bu sınıflandırma ideolojik ve eşitsizlik, kalıpyargı üretme ve tahakküm, nesneleştirme, cinsel şiddet, ve kadın düşmanlığı ve cinsel olmayan şiddet şeklindeydi. Ayrıca, 2023’te olduğu gibi, yazarın niyetine (doğ-

24 **Kaggle.** (2018). *Jigsaw toxic comment classification challenge: Overview*. <https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge/overview>

25 **UNED NLP Group.** (2024). *EXIST 2024: Explainable hate speech detection*. <http://nlp.uned.es/exist2024/>

rudan, dolaylı veya yargılayıcı) dayalı olarak mesajın sınıflandırılmasını içeren kaynak niyeti görevi de dahil edildi. Son olarak, 2024'te, memlerdeki cinsiyetçiliği ele almak üzere üç yeni görev getirildi: cinsiyetçiliğin tespiti, sınıflandırılması ve memlerde kaynağın niyeti.

NLP'de nefret söylemi çalışması ikili sınıflandırmadan çok daha karmaşık ve ayrıntılı kategorizasyona doğru gelişme göstermiştir. Nefret söylemi yeni grupları hedef alıp ve farklı biçimlerde ortaya çıkıp gelişmeye devam ettikçe, etiketleme metodolojilerinin geliştirilmesi de alandaki araştırmacılar açısından bir mesele oluşturmayı sürdürecektir.

Araştırmacıların veriye erişebilirliği

Houda Bouamor, Onur Varol, Berrin Yanıkoğlu

Çevrimiçi etkileşim, içerik moderasyonu ve nefret söylemini de içeren zararlı bilgilerin yayılması konulu çalışmalar, araştırmacıların sosyal medya verilerine erişimine dayanır. Ancak bu erişim, platform politikaları, gizlilik ilkeleri ve teknik sınırlılıklar nedeniyle giderek daha da kısıtlanıyor. Sosyal medya şirketleri veri erişimini katı hizmet kullanım şartları, tescilli algoritmalar ve çoğu zaman şeffaflıktan yoksun şekilde geliştirilen politikalarla düzenliyor.

X/Twitter, Facebook, Instagram ve Reddit gibi platformlarda yapılan son değişiklikler, araştırmacıların veri toplama imkânını daha da kısıtladı. Örneğin X/ Twitter platformdan alınabilecek mesaj sayısını sınırlarken, Facebook mesajları arkadaş olmayanlara halen kapalı. Bu kısıtlılıklar, teknolojik ve düzenleyici engellerle birleşince, araştırmacıların sanal platformları etkin biçimde analiz edebilme olanakları oldukça daralıyor.

Veriye erişim zor. Araştırmacıların sosyal medyada nefret söylemini incelemek için gerekli verilere erişimini güçleştiren çeşitli etkenler var. Sosyal medya platformları genellikle gizlilik ilkelerine ve mülkiyet haklarına dayanarak kullanıcı verilerini katı şekilde koruyor. Ayrıca, bu platformlarda kullanıcılar tarafından oluşturulan içeriğin geniş hacmi, manuel veri toplama ve analizi güçleştirdiğinden, teknik açıdan karmaşık ve kaynak yoğun olabilen otomatik yaklaşımların kullanılması gerekiyor. Örneğin, bu platformlardan sadece birkaçı uygulama programlama arayüzü (API) sunarken, bunlardan bazılarının kullanımı hiç de kolay değil. API istekleri için hız sınırlaması ve aboneliğe dayalı modeller, araştırmacıları örnek veri toplamaya veya analizde kullanılan veri kümelerini nihayetinde yanlış hale getiren katı kriterler kullanmaya yönlendiriyor. 2022 yılında bir grup araştırmacı, kolektif bir çalışma yaparak, X/Twitter'ın tüm genel akışını yakalamak için hız sınırlamaları ve API kısıtlamalarına karşı geldiler (Pfeffer vd., 2023). Bu denli yoğun bir çabaya rağmen, sadece tek bir günlük veri yakalamayı başardılar; büyük ölçekte veri toplamanın zorluğunu gösteren bu örneğin ardından Elon Musk sahipliğinde yapılan API değişiklikleri bu tür girişimleri de imkânsız hale getirdi.

Veriye erişim farklı platformlarda farklı şekillerde düzenleniyor. Araştırmacıların her platformda farklı yasal gerekliliklere uyması gerekiyor. Veri erişimi ve

kullanımına ilişkin etik ilke ve düzenlemeler platformlar arasında büyük ölçüde değişiklik gösterebiliyor. Bazı platformlar kullanıcı gizliliğini korumak için daha katı politikalara sahipken, bazıları daha esnek yaklaşımlar benimseyebiliyor. Araştırmacılar açısından bu kuralları uygulama aşamasında yorumlamak ve farklarla baş etmek, hele ki karşılaştırmalı araştırmalar için farklı platformlar üzerinde çalışmak istiyorlarsa, büyük bir güçlük yaratıyor.

Veriler değişiyor. Ayrıca, bu platformlardaki sanal içeriğin dinamik yapısı dolayısıyla veriler kullanıcılar tarafından silinebilir veya platformlar belirli verileri silip erişimi kısıtlayabilir; bu da araştırmacıların istikrarlı çalışmalar yürütmesini zorlaştıran bir faktördür. Nefret söyleminin çok hızlı yayılması araştırmacıların verilere gerçek zamanlı erişimini gerektirdiği için bir başka karmaşıklık daha söz konusudur. Keza, bot hesaplar bu boşlukları kullanıp manipülatif ya da nefret yüklü içerikler yayınlıyor ve platform ya da tespit sistemlerinin müdahalesinden önce bunları siliyorlar; fakat kullanıcılar halihazırda bilgiyi tüketmiş ve bundan etkilenmiş oluyor.

Kullanıcılar platform kural ve düzenlemelerini kandırebiliyor. Platformlar kullanıcıları korumak için nefret söylemini otomatik olarak tespit edebilir veya kullanıcılara önceden tanımlanmış kriterlere göre içeriği filtreleme olanağı sunabilir. Hesaplar da diğer kullanıcı ve içerikler hakkında şikâyetle bulunabilir. Ancak, bilhassa nefret söylemi söz konusu olduğunda, hesaplar bu mekanizmaları aşabilmek için belirli kelimeleri farklı şekilde yazarak filtrelerden kaçmak veya olumsuz tepkileri azaltmak için mütemadiyen içerik yayınlayıp silmek gibi yaratıcı yöntemler bulabiliyor. Örneğin AB Genel Veri Koruma Tüzüğü'nde (GDPR)²⁶ yer alan ve kişisel verilerin belirli koşullar altında platformlardan kaldırılmasını talep imkânı tanıyan “unutulma hakkı”, sanal ortamda zararlı davranışlar sergiledikten sonra dijital izlerini silmeye çalışan kullanıcılar tarafından kötüye kullanılabilir (Xue vd., 2016). Bu durum, gizlilik hakları ile nefret söylemiyle mücadele ve kamu güvenliğini koruma ihtiyacı arasında denge gözetmeye çalışan platformlar ve düzenleyici merciler açısından büyük zorluk yaratıyor.

Sosyal medya araştırmalarının veri erişiminden dolayı değişkenlik göstermesi, özellikle manipülasyon, koordineli faaliyetler ve nefret söyleminin sistematik yayılımı konusunda tespit araştırmalarına ihtiyaç duyulan bir dönemde kaygı yaratıyor. Araştırmacılar bırakın hafifletici stratejiler geliştirebilmeyi, sorunun boyutlarını ölçmek için bile gereken verilere ulaşamıyor.

²⁶ Avrupa Birliği 2016/679 sayılı ve 14 Nisan 2016 tarihli Genel Veri Koruma Tüzüğü (yürürlük tarihi 25 Mayıs 2018), <https://www.ab.gov.tr/siteimages/resimler/Nihai-ABB-HCDB-GDPR.pdf>

AB'nin çıkardığı **Dijital Hizmetler Tüzüğü** (DSA)²⁷ sosyal medya platformlarını düzenlemek ve platform şirketlerine yükümlülükler getirmek için atılmış önemli bir adımdır. “Demokrasi ve Teknoloji Merkezi”²⁸ ve “Bağımsız Teknoloji Araştırmaları Koalisyonu”²⁹ gibi bağımsız kuruluşlar da şeffaflığı ve veri erişimini destekliyor.

Yanlılık. Önceki bölümlerde ele alındığı üzere veri etiketleme sırasındaki yanlılık meselelerinin yanı sıra, sosyal medya platformlarından veri toplama süreci de hayli yanlı olabilir. Önceden API'ler aracılığıyla veri erişimi sağlayan birçok platform olmakla birlikte, bu sistemlere içkin örnekleme stratejileri ve olası yanlılıklar bilinmezliğini koruyor. X/Twitter API'leri üzerine yürütülen çalışmalar (Morstatter vd., 2013) kamuya açık rastgele örnekleme için ücretli erişim yoluyla elde edilen verilerden farklı olduğunu göstererek veri erişimindeki tutarsızlıkları vurguluyor. Bundan başka, kullanıcı davranışı da veri toplama sürecini etkiliyor. Platformlar API erişimine hız sınırlaması getirdiğinden, geliştiriciler veri yakalamayı en üst düzeye çıkarmak için sorguları optimize ediyor ve bu da veri kalitesini etkiliyor: Farklı örnekleme stratejileri ortaya çok farklı veri kümeleri çıkartabiliyor. Veri yanlılığı ve platform şeffaflığına dair endişeler iyi bir şekilde belgelenmiştir. Yanıltıcı bilgi araştırması için ideal olarak kullanacakları veri setleri sorulduğunda, çoğu araştırmacı bu sorunları vurgulamaktadır (Pasquetto vd., 2020).

27 Avrupa Birliği 2022/2065 sayılı ve 19 Ekim 2022 tarihli Dijital Hizmetler Tüzüğü, https://www.ab.gov.tr/siteimages/resimler/hcdb_dsa_14.10.2024.pdf

28 Center for Democracy and Technology. (2024). <https://cdt.org/>

29 Coalition for Independent Technology Research. (2024). <https://independenttechresearch.org/>

Nefret söyleminin otomatik tespitiye yönelik araçlar geliştirilmesi konusunda mevcut meseleler ve olası çözümler

Tommaso Caselli

Son on yılda Doğal Dil İşleme (NLP) –insan dillerinin otomatik olarak anlaşılması ve üretilmesine yönelik araçlar geliştirmeye odaklanan Yapay Zekâ alanı– nefret söylemi ve ilişkili olguların otomatik tespiti için dil kaynakları (korpus ve araçlar) geliştirme konusuna eğiliyor. Önceki çalışmalar birçok dili kapsayan ortak iş/görev ve yarışmalarla³⁰ genişletilse de halen farklılıklar mevcut ve bunlar sadece yüzeysel olarak ele alınıyor. Önceki araştırmalara bakıldığında (Vidgen & Derczynski, 2020; Poletto vd., 2021), NLP ile ilgili olarak ele alınması gereken üç temel konu saptanıyor: tanımlar, veri kaynakları ve veri etiketleme.

Tanımlar

Genelgeçer bir nefret söylemi tanımının yokluğundan ötürü yaşanan güçlüklere önceki bölümlerde ele alınmış olduğu için bunları tekrarlamadan konuyu kısaca değerlendireceğiz.

Tartışmayı nefret söylemine referans oluşturacak bir tanıma yönlendirmek bakımından olası bir başlangıç noktası, bu kavramın öznel ve saldırgan niteliğini kabul etmek olacaktır. Saldırgan dil, bununla bağlantılı olarak nefret söylemi gibi olguların ayrıştırılmasına yardım edebilecek geniş bir şemsiye işlevi görür. NLP araçlarının değerlendirilmesine yönelik bir çalıştay serisi olan SemEval 2019 ve 2020 boyunca yürütülen iki başarılı ortak görevde önerilen saldırgan dil tanımı (Zampieri vd., 2019, 2020) bu açıdan faydalı olabilir. Bahsi geçen bağlamlarda saldırgan dil, *hedefli saldırılar, hakaretler, tehditler ve yakışıksız ifadeler içeren mesajlar da dahil olmak üzere, açık veya zımni, kabul edilemez nitelikte her türlü ifade* olarak tanımlanıyor (tam alıntı için bkz. yukarıdaki Tablo 2). Bu tanımın başlıca avantajı birçok farklı duruma uygulanmasını mümkün kılan genelliğidir. “Kabul edilemez nitelikte her türlü ifade”ye işaret etmesi bu olguların öznel niteliğini yansıttığı için önemlidir. Bir hedefin varlığını zorunlu kılmaması nefret söylemini başka olgulardan ayırmak için başlangıç olabilir. Örneğin nefret söylemi; ırk, din, cinsel yönelim, toplumsal cinsiyet, siyasi

30 **Kaggle.** (2018). *Jigsaw toxic comment classification challenge*.
<https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge>

görüş gibi içkin özelliklerinden ötürü bir kişi veya grubu hedef alan hakaret olarak tanımlanırsa, sonrasında etiket şemalarının ve otomatik araçların geliştirilmesinde kullanılabilecek ortak özelliklerin tespitine katkıda bulunur. Belirli olguları rencide edici niteliğiyle yeniden ilişkilendirme olasılığı da birleştirilmiş bir çerçeve geliştirmeye yardımcı olabilecek kilit bir öğedir.

Veri kaynakları

Nefret söylemi için Hint-Avrupa dilleri, Türkçe, Sami dilleri gibi farklı dil ailelerini kapsayan korpuslar geliştirildi.³¹ Fakat herhangi bir dilde ortak kabul gören tek bir *referans* veri kümesi veya ölçüt belirlemek mümkün olmadığı için geliştirilen araç ve yaklaşımları karşılaştırmayı güçleştirmektedir.

Veri kaynaklarındaki değişkenlik de bir sınırlama yaratıyor. Vidgen ve Derczynski (2020) tarafından vurgulandığı üzere, son 10 yıldır bu ve benzeri dil fenomenleri için veri kümeleri oluşturmak üzere kullanılan başlıca sosyal medya platformu, 2023 yılı başlarına kadar uygulama programlama arayüzünün (API) açık olması ve desteklenen en az 34 dilde 300 milyondan fazla kişi tarafından kullanılması sayesinde Twitter (X) olmuştur. Bu durum birçok dilde mukayese edilebilir veri kümeleri yaratabilen avantajlar sunmakla birlikte, alanda yürütülen araştırmalar açısından kendine özgü bir sınırlılık yaratmıştır. Mesajların azami uzunluğuyla ilgili özgün özelliklere sahip tek bir sosyal medya platformuna odaklanmak,³² okur açısından mesajların saldırganlık durumunu tetikleyebilecek dilsel ifadelerle dair çarpık veya yanlı beyanlara yol açmıştır. Nefret söylemi ve benzer fenomenlerin NLP'deki popülerliği bir yana, geliştirilmiş olan bir sistemin taşınırılığı, dil fenomeni (örn. saldırgan dil olarak etiketlenmiş veriler) değişmese bile veri kaynağı farklılaştığında (örn. Twitter/X üzerinde eğitilmiş ve Reddit'te uygulanan bir sistem) zayıf kalmıştır (Fortuna vd., 2020).

Nefret söyleminin doğal örneklerini toplamak bakımından sosyal medyanın ekolojik bir platform oluşturduğu açık, fakat bu platformlar aynı zamanda, etiketlenen verilerin açıklanmasını ve deneylerin tekrarlanabilirliğini de kısıtlıyor. Örneğin, Twitter/X

31 Zubiaga, L. (2020). *The Hate Speech Dataset Catalogue: A comprehensive community-based resource documenting datasets for hate speech detection (and related phenomena) in multiple languages*. GitHub. <https://github.com/leondz/hatespeechdata/tree/master>

32 X. (2024). *XPremium Hakkında*. X Yardım Merkezi. <https://help.x.com/tr/using-x/x-premium> Twitter'ın Elon Musk tarafından satın alınmasından beri, izin verilen mesaj uzunluğu abonelik planına göre değişiklik gösteriyor: Basic katman kullanıcıları 280 karakter, ücretli Premium ve Premium+ katman kullanıcıları ise 25.000 karaktere kadar varabiliyor.

kullanım şartları, içeriğin tek sahibi olan kullanıcıdan izin alınmadığı sürece, mesajın kimliği dışında hiçbir bilginin kamuya açıklanmasına izin vermiyor. Rehidrasyon, yani veri kümesini oluşturan mesajların platformdan tekrar indirilmesi gibi teknikler kullanılarak verilere ulaşmak mümkün olsa da, yukarıda belirtildiği gibi, bazı mesajların (ve genellikle de saldırgan içeriği fazla olanların) platform tarafından kaldırılmış veya kullanıcılar tarafından silinmiş olma riski yüksektir. Bu durum çoğu zaman veri kümelerinin “küçülmesine” yol açar: Bir rehidrasyondan diğerine boyut değiştiren veri kümeleri, sonuçların NLP sistemleri üzerinden karşılaştırılmasını imkânsız hale getirir. Bu sorunun etkisini ortadan kaldırmaya değilse bile hafifletmeye çalışan çözümler tüm kullanıcı adlarını ve URL bağlantılarını kaldırarak mesajları anonimleştirir. Yani bir mesaj sosyal medya platformundan kaldırılırsa (ve kaldırıldığında) kullanıcının söz konusu mesajın sahibini saptayabilme olasılığı ortadan kalkar ve toplanan verilerin bütünlüğü korunmuş olur. SemEval çalıştayları bağlamında düzenlenen çokdilli saldırgan ifadelerin tespiti konulu ortak çalışma OffensEval 2020 (Zampieri vd., 2020) bu çözümü benimsemiştir. Anonimleştirilip özel bir Veri Paylaşım Sözleşmesi imzalandıktan sonra açıklanan Hollanda Hakaret Korpusu – DALC (Ruitenbeek vd., 2022) için de benzer bir çözüm kullanılmıştır. Son zamanlarda toplanan veriler açısından ortaya çıkan bir mesele de, sosyal medya platformlarının sahip oldukları verileri paraya çevirmek amacıyla, API’lerinin kullanımına getirdikleri artan kısıtlamalardır. İndirilebilecek mesaj sayısı sınırlı olduğu (örn. Twitter/X bunu günde 500 mesajla sınırlamıştır), API’lere erişim ücretli olduğu ya da verilerin makine öğrenmesi veya yapay zekâ modellerini eğitmek için kullanılması açıkça yasaklandığı için (örn. Reddit) geniş ölçekte veri toplama girişimleri gitgide zorlaşıyor.

Daha önce belirtildiği gibi, veri kümeleri ve derlemeleri çeşitli coğrafi ve dilsel ortamları içine alacak şekilde genişliyor. Fakat veri toplamaya ilişkin standartlaştırılmış protokol ve yönergeler bulunmadığı için çokdilli ve dillerarası NLP araçlarının gelişimine ket vuruluyor. Etkili veri toplama metodolojileri bu süreçte öncü rol oynuyor. Wiegand vd. (2018) tarafından vurgulandığı üzere, anahtar kelimelere dayalı yaklaşımlar belirli konu veya yazarlar etrafında yoğunlaşan eğilimlere yol açabilir. Bu durumda toplumsal, etnografik ve topluluk merkezli hususlar çoğu zaman ihmal edilip yeterince temsil imkânı bulamıyor, dolayısıyla nefret söyleminin tespiti sadece gelişmiş bir örüntü tanıma ile sınırlı kalıyor. Oysa farklı sosyal gruplar neyin nefret teşkil ettiği konusunda farklı standartlar benimseyebilir. Örneğin, drag topluluğunda b-kelimesi³³, hakaret anlamında değil, daha çok bir sevgi veya şefkat terimi olarak kullanılır. Veri içerisinde bu farklı perspektiflerin

33 **Yayınca Notu:** Buradaki b-kelimesi İngilizce’de *bitch* kelimesi yerine sıkça kullanılan *b-word*’ün çevirisidir.

doğru şekilde yansıtılmaması uygunsuz sonuçlar yaratabilir. Nefret söylemine yönelik çoklu türden veri setlerinin sayısı da halen çok düşük (Hosseinmardi vd., 2015; Zhong vd., 2015; Suryawanshi vd., 2020); yazılı kip baskın olduğunda, internet ortamının gitgide daha görsel bir araç haline geldiği de düşünülürse, görüntü ve metin gibi birden fazla yöntemle çalışmanın yaratacağı sinerji potansiyelleri göz ardı ediliyor.

Veri etiketleme

Etiketlenmiş verilerin varlığı, her türlü denetimli makine öğrenmesi yaklaşımında kilit unsurdur. Ancak bu verilerin varlığı daha nitel çalışmalarda da rol oynar; aynı değilse bile benzer dil olgularının mukayesesini kolaylaştırır. Biz burada NLP bünyesinde geliştirilmiş uygulamaları ele alacağız.

Nefret söylemine ilişkin ilk çalışmalarda, mesajlar kabul gören belirli bir tanımlı karşılayıp karşılamadığı şeklinde etiketlenerek bütüncül bir yaklaşım benimsenmişti. Waseem vd. (2017) tarafından öne sürülen çok eksenli etiketleme yaklaşımının ardından daha ayrıntılı etiketler geliştirildi. OffensEval 2019 ve 2020 ortak görevlerinde ise farklılaştırılmış bir yaklaşım uygulanarak üç seviyeli hiyerarşik bir açıklama şeması geliştirildi: İlk seviyede saldırgan olan ve olmayan mesajlar arasında ikili bir ayırım yapıldı; ikinci seviyede saldırgan mesajın bir hedef içerip içermediği ayırt edildi ve son olarak üçüncü seviyede hedef bir kişi, grup veya başka bir oluşum olarak işaretlendi. Bu çözümün bir başka yararı da tüm seviyeleri birleştirerek hakaret veya nefret söylemi gibi daha çirkin dilsel olguların tespitini mümkün kılmasıydı.

Nefret söylemi genellikle münferit şekilde ayıklanıp etiketlenir: Yani mevcut veri kümelerinin hiçbirinde etiketleme uzmanlarının (veya makinelerin) incelenen mesajın (tam) bağlamına erişimi yoktur. Ana gönderiler (yani bir mesaj dizisi – *thread*– veya başka kullanıcılarla bir tartışma başlatma potansiyeli taşıyan mesajlar) hayli hatasız bir şekilde münferit mesajlar olarak etiketlenebilirken, yanıt niteliğindeki mesajlarda bağlamdan kopukluk yıkıcı bir etki yaratabilir: Nefret içeren bir mesaja yanıt olarak gönderilen gülen yüz emojisi de nefret içerir. Münferit etiketleme sadece mesajın oluşum bağlamını etkilemekle kalmayıp, mesajı yayınlayan kullanıcıya dair sosyodemografik bilgilerden de yoksundur. Bu verilere erişim, gizlilik ve başkaca etik sebeplerden ötürü, her zaman mümkün olmayabilir ve mümkün olsa bile arzu edilmeyebilir. Bu veriler ayrıca, etiketleme uzmanlarının, hakaret içermekle birlikte azınlıklar tarafından yeniden sahiplenildiği bilinen kelimeleri yanlış şekilde etiketlemekten kaçınmasına yardımcı olabilir.

Artstein ve Poesio (2008) farklı uzmanların bir etiket üzerinde mutabık kalma oranıyla sistemin performansı arasında bir bağlantı olmadığına işaret etse de, nefret söylemini etiketlemenin zor bir iş olduğu muhakkaktır. Mevcut veri kümeleri, genel olarak, uzmanları arasında orta düzeyde bir mutabakat seviyesi sergiler. Dahası, etiketleme yönergelerinin nasıl formüle edildiğine bağlı olarak başka uyumsuzluk kaynakları da doğabilir. Röttger vd. (2022) özneliği yüksek işler için etiketleme yönergeleri geliştirirken iki temel paradigmadan bahseder. *Tanımlayıcı* paradigma diye anılan birinci yaklaşım etiketleme uzmanlarının özneliğine çok kısıtlama getirmez, yani uzmanların birçok kanaati kayda geçirmesine izin verir. Etiketlerin daha ayrıntılı şekilde modellenmesine imkân sunarak büyük oranda öznel değerlendirmelerin ortaya çıktığı ve dolayısıyla uzmanlar arasında görüş ayrılıklarının bulunduğu örnekleri anlamaya yardımcı olur. Nefret söylemi açısından bakıldığında bu etiketleme yaklaşımının, farklı uzmanların farklı nefret seviyeleri algıladığı durumların tespitine yardımcı olabileceğini görüyoruz. *Kuralcı* yahut *normatif* olarak anılabilecek ikinci paradigma ise uzmanların öznel yorumları en aza indirebilmek amacıyla, öngörülen etiket yönergelerine sıkı sıkıya bağlı kalmasını gerektirir. Mevcut veri kümelerindeki etiketler açısından var olan bazı sorunlar, geliştirilen etiketleme yönergelerinin bu paradigmalardan biri gözetilerek hazırlanmış olmamasından kaynaklanıyor olabilir.

Son derece öznel bir iş olan nefret söyleminin etiketlenmesi, çok sayıda etiketleme uzmanının her bir mesaja ilişkin kanaatlerini ifade ettiği görevler olarak görünüyor; bu noktada söz konusu kişinin konuya hakim bir uzman olup olmadığına bakılmıyor. Etiketlerdeki uyumsuzluklar da genellikle göz ardı edilip ya etiketleme görevini yapanların çoğunluğunun tercihinine bağlı olarak tek bir etiket altında toplanıyor ya da bu kişilerden uyumsuzluğu çözmeleri isteniyor. Bu türden bir işteki uyumsuzlukların sadece veri gürültüsü veya hatası olarak algılanmak yerine ek bilgi kaynağı olarak değerlendirilmesi ancak yakın zamanlarda, bu konuda farkındalığın artmasıyla gözlenmeye başlandı (Plank vd., 2014; Leonardelli vd., 2021; Cabitza vd., 2023).

Sosyal medya şirketlerinin yapay zekâ teknolojilerini etik şekilde uygulaması

Zeynep Özarslan

Sosyal medya platformları ifade özgürlüğünü kolaylaştıran ortam ve araçlar sağlıyor. Ancak, Ullmann ve Tomalin'in (2020) vurguladığı üzere, bu platformlar "nefret söyleminin neredeyse anında çevrimiçi olarak ve genellikle anonim bir şekilde ifade edilebildiği süratli ve etkili yeni iletişim biçimleri yarattı" (s. 70). Başta üretken yapay zekâ olmak üzere çeşitli yapay zekâ (YZ) teknolojilerinin yaygın kullanımı, bu platformları yeni önyargı, ayrımcılık ve nefret türlerine aracılık eden ve bunları pekiştiren ortamlar olarak öne çıkarttı. Nefret söylemi sosyal medyada mem, GIF, emoji, filtre, sahte kimlik kullanımı ve ırkçılığı teşvik eden fenomenler gibi hem açık hem gizli taktiklerle yayılıyor. Sosyal medya ve yapay zekâ teknolojisi şirketlerinin kendi politika ve içerik moderasyon süreçleriyle nefret söylemini önleme çabasına rağmen, bu teknolojiler, paradoksal şekilde, nefret ve ayrımcılığın idamesinde kullanılıyor ve böylelikle baskıyı ırk, cinsiyet, cinsellik, dil, din gibi hatlarda yeniden şekillendiriyor.

Loebbecke vd. (2021) tarafından öne sürüldüğü üzere, bu platformlar dünya çapında kültürel çeşitliliğe sahip pazarlarda faaliyet gösterdiği için, hizmet şartları kapsamında küresel bir kitleyi tatmin etmek bakımından muhtelif nefret söylemi tanımlamalarının benimsenmesi gerekiyor. Genelgeçer bir dizi kural yerine, dünyanın her bir köşesinde farklı düzenlemelere gidilmesi, sosyal medya platformlarının nefret söylemi gibi yasadışı içeriklerin kaldırılmasına yönelik süreçler geliştirmesini şart kılıyor. Platformlar bunun için, kötü niyetli içeriğin tespitine yönelik içerik moderasyon sistemlerine yatırım yapıyorlar (Loebbecke vd., 2021). Genel uygulama olarak, platformlardaki saldırgan mesajlar kullanıcılardan gelen şikâyetler üzerine insan moderatörler tarafından incelenir. Söz konusu içerik nefret söylemi sınıfına giriyorsa platformdan kaldırılır; değilse, platformda erişilebilir olmayı sürdürür. Ancak saldırgan içerik silindiğinde dahi muhataba verilen zarar bu mesaja daha öncesinde maruz kalınmış olması sebebiyle devam edebilir. Ayrıca, sanal nefret söylemi içeriğinin geniş hacmi ve hızı insan moderatörleri bunaltabilir. Bu nedenle "son yıllarda, sanal ortamdaki bu türden saldırgan materyalle mücadele amacıyla yeni stratejiler geliştirmek üzere yeni nesil otomatik nefret söylemi tespit sistemleri üretilmeye başlandı" (Ullmann & Tomalin, 2020, s. 69). Bundan başka, metin tabanlı, kullanıcı tabanlı ve ağ tabanlı özellikleri ayıklayarak ve sanal zorbalıkları ve

nefretle bağlantılı anahtar kelimeleri saptayarak nefret söylemini ve hakaret dilini tespit etmek amacıyla doğal dil işleme (NLP), makine öğrenmesi, derin öğrenme ve açıklanabilir yapay zekâ (XAI) gibi yapay zekâ teknolojileri devreye sokuldu (Meh-ta & Passi, 2022). “Ne var ki, neyin nefret söylemi teşkil ettiğine dair müşterek bir anlayış olmaması bu gayretlerin başarıya ulaşmasını güçleştiriyor. Farklı tanımlar, düzenlemeler ve bağlamlar konuyla ilgili yapay zekâ sistemleri geliştirilip uygulan-masını zorlaştırıyor” (Loebbecke vd., 2021).

Yapay zekâ teknoloji ve araçları sanal nefret söyleminin tespitinde son derece fay-da sağlamakla birlikte, bunların etik şekilde uygulanması şarttır. Yeni yapay zekâ teknolojilerinin olası zararlarını azaltmak amacıyla yakın zamanda bir YZ etik yö-nerge ve ilkeler korpusu hazırlandı. Söz konusu ilke ve yönergeler, tutarlı bir “etik yapay zekâ” vizyonu oluşturularak hayata geçirilmesi için gerekli etik ve teknik standartların belirlenmesi amacıyla uzmanlar tarafından incelenip sentezlendi. Bu inceleme sonucunda örneğin Royackers vd. (2018) tekrar eden 6 tema tespit etti: “gizlilik, özerklik, güvenlik, insan onuru, adalet ve güç dengesi”. Ayrımcılık, özerklik, insan onuru ve eşitsiz güç dengesi gibi konularda yeterli denetim olma-dığını da saptadılar. Keza, Jobin vd. (2019) 84 etik YZ belgesini inceledi ve içerik analizlerinden ortaya çıkan 11 etik değer ve ilke belirledi: şeffaflık, adalet ve hak-kaniyet, kötü niyet taşımama, sorumluluk, gizlilik, iyi niyet, özgürlük ve özerklik, güven, onur, sürdürülebilirlik ve dayanışma. Bunlar arasından şeffaflık, adalet ve hakkaniyet, kötü niyet taşımama, sorumluluk ve gizlilik ilkeleri üzerinde daha yo-ğun bir fikir birliği mevcut. Yine buna paralel olarak, Morley vd. (2020) “(a) insan-lara ve çevreye faydalı ve saygılı (iyi niyet); (b) sağlam ve güvenli (kötü niyet taşı-mama); (c) insanî değerlere saygılı (özerklik); (d) adil (adalet) ve (e) açıklanabilir, hesap verilebilir ve anlaşılabilir (açıklanabilirlik)” olmayı içeren, etik ilkelere bağlı bir makine öğrenmesi tanımı geliştirdi (s. 2145). Hagendorff (2020) ise temel YZ etik ilke ve tavsiyelerini inceleyip hesap verebilirlik, gizlilik veya adillik gibi tekrar eden temaları vurguladı. Ayrıca, hesap verebilirlik, açıklanabilir YZ, adillik ve ay-rımcılık konusunda hassas veri madenciliği ve gizlilik konularında etik ölçütlere ulaşmak için muazzam teknik çabalar sergilendiğine dikkat çekti (Hagendorff, 2020, s.103). Ancak “neredeyse hiçbir yönergenin otomatikleştirilmiş propagan-da, botlar, sahte haberler, sahte videolar (deepfake), mikro hedefleme, seçim-lerde hile vs. bağlamında *yapay zekâ sistemlerinin siyasi olarak kötüye kullanılması* ihtimalini... [ve] yapay zekâ topluluğu içinde *çeşitliliğin bulunmaması* meselesini konu edinmediğini” de vurguladı (Hagendorff, 2020, s.105). Bundan başka, ya-pay zekânın toplumsal uyumun azalması ve radikalleşmenin yükselişine, akılcı kamusal söylemin zayıflamasına ve toplumsal bölünmeye çanak tuttuğunu öne

sürüyor (Hagendorff, 2020, s.110). Öte yandan Robles Carrillo (2020), temel ilke ve ortak unsurlar var olmakla birlikte, etik anlayış ve ilkelerin gelenekler, kültürler, ideolojiler, sistemler ve ülkeler açısından çeşitliliğe tabi olduğunu ve “etik” özünün değişen zaman ve toplumlarla birlikte değiştiğini savunuyor (s.3).

Yapay zekâ sistem ve araçlarının kullanımı, şirketlerin bu teknolojiiden nasıl yararlandığına bağlı olarak etik dışı bir nitelik taşıyabilir. Ancak bu tür etik olmayan uygulamaları önlemek için birtakım strateji ve teknikler mevcuttur. Sosyal medya platformları, YZ sistem veya araçlarını geliştirip kullanırken, özellikle eğitime aşamasında doğru, eksiksiz ve çeşitli veriler elde edilip işlenerek veri kümelerinde eğilim ve ayrımcılık risklerini ortadan kaldırmaya yönelik algoritmik adalet, adillik ve eşitliğe öncelik vermelidir. Gizliliğin ihlali ve ayrımcılık gibi olası zararlardan kaçınmayı içeren, kötü niyet taşımayan YZ sistemleri geliştirmeli ve/veya kullanmalıdır. Başka bir deyişle, şirketler YZ sistemlerinin yaratabileceği olumsuz etkilerin farkında olarak bu etkileri azaltacak tedbirler almalıdır. Bu minvalde Zhuo vd. (2023), eğitim verilerinin belirli gruplara yönelik yanlış beyanlar içermesi halinde geniş dil modellerinin (LLM) bu gruplara karşı haksız veya ayrımcı öngörüler içerdiğini söylüyor. Dolayısıyla, bunu önlemek için, “eğitim verilerinin çeşitlilik göstermesini ve kullanılacağı nüfusu temsil etmesini sağlamak ve verilerdeki olası önyargı ve eğilimleri birebir saptayıp ortadan kaldırmak şarttır” (Zhuo vd., 2023). Eğitim verilerinin tek bir kültür gibi nüfusun sadece bir bölümünü temsil etmesi anlamına gelen dışlayıcı normlar da etik açıdan bir başka mesele yaratır. Zira böyle bir durum modelin, örneğin farklı dil konuşanlar veya başka kültürlere mensup insanlar gibi temsil gücü düşük grupları anlayamamasına veya bunlar için içerik üretememesine yol açabilir (Zhuo vd., 2023). Bu nedenledir ki YZ sistem ve araçları mümkün olduğunca geniş bir kitleye fayda sağlayacak denli çeşitli ve kapsayıcı olmalıdır.

YZ modellerinin yalnızca bir dildeki veriler üzerinden eğitildiği tek dilli yanlışlık, modellerin başka dillerde metin oluşturmaya ve anlamaya engel olabilir; bu durumda söz konusu dili konuşmayan kişilere fayda sağlamayacağı gibi, potansiyel olarak bu gruplar hakkında yanlış veya haksız öngörülere yol açabilir. Bunu önlemek için, eğitim verilerinin büyük oranda çeşitli dil ve kültürlerden yüksek kaliteli korpuslar içermesini sağlamak gerekir (Zhuo vd., 2023). Aynı çalışmanın yazarları, modellerin zararlı veya saldırgan içerik üretme veya anlama kapasitesine atfen, geniş dil modellerinin içerdiği toksiklik potansiyelini de vurguluyor. Toksiklik, saldırgan dil içeren eğitim verilerinden kaynaklanabilir ve modelin de kullanıcı etkileşimleri sırasında saldırgan içeriği tanıyıp üretmesi anlamına gelir.

Bu riski azaltmak için, eğitim verilerinin saldırgan dil içermediğinden emin olmak ve var olabilecek her türlü saldırgan materyali önceden hareket edip kaldırmak şarttır (Zhuo vd., 2023).

YZ sistemleri açısından bir başka etik mesele de güvenilirliktir. Güvenilir bir YZ sistemi geliştirmek için, eğitim verilerinin yanlış, isabetsiz veya yanıltıcı bilgi içermemesi gerekir. Yoksa YZ sistemi tarafından yanıltıcı çıktılar üretilecektir. Bu nedenle, eğitim verilerinin güncel tutulması ve YZ sistemlerinin en doğru bilgiyi vermesini sağlamak için devamlı surette izlenip güncellenmesi şarttır. Ayrıca, Buolamwini ve Gebru (2018), otomatik yüz analizi başta olmak üzere bilgisayarlı görü teknolojilerindeki etik meseleleri inceleyip algoritmik adalet, doğruluk, şeffaflık ve hesap verebilirlik meselelerine vurgu yapıyor. Gebru vd. (2018) de makine öğrenmesi modellerinin eğitim veri kümelerine yansıyan istenmeyen toplumsal eğilimleri yeniden üretip pekiştirebileceğini söylüyor. Bu yüzden, “makine öğrenmesi topluluğu içinde şeffaflığı ve hesap verebilirliği artırma, makine öğrenmesi modellerindeki istenmeyen toplumsal eğilimleri azaltma, makine öğrenmesi sonuçlarının daha çok tekrarlanabilirliğini kolaylaştırma ve araştırmacı ve uygulayıcıların atanmış işler için daha uygun veri kümeleri seçmelerine yardımcı olma potansiyeline sahip” veri kümeleri için veri listeleri öneriyorlar (Gebru vd., 2018).

Bu bağlamda, YZ’nin etik şekilde kullanımı nefret söyleminin üretilip yayılmasını önlemek konusunda sosyal medya platformlarına nasıl yardım edebilir? Bu konuyu etkili bir şekilde ele alabilmek için aşağıdaki gibi çok katmanlı bir yaklaşım uygulanabilir:

Çeşitli ve temsili eğitim verileri: Eğitim veri kümelerinde çoklu dil, lehçe, kültür ve demografik temsillerin kapsayıcı bir şekilde kullanılması, ayrımcılığa veya nefret söylemine yol açan eğilimlerin azaltılmasına yardımcı olabilir. Başka bir deyişle, eğitim verilerine farklı dil ve kültür zeminlerinin dahil edilmesi tek-dilli yanlılığı hafifletebilir ve böylece sosyal medya algoritmalarının nefret söylemini etkili şekilde tespitine yardım edebilir. Ayrıca, demografi içerisindeki ayrımcı kalıpları tespit etmek için protokoller oluşturulması, sosyal medya platformlarında yapısal önyargıların yeniden üretilmesini önleyebilir.

Bağlamsal ipuçlarının analizi: Platformlar toksiklik potansiyelini saptayıp işaretlerken, kullanıcının geçmiş katılım ve faaliyetlerini ve ağ içerisinde diğer etkileşimleri göz önünde bulunduran bağlamsal ipuçlarını analiz eden YZ sistemleri geliştirebilir. Bağlamsal analiz için bu tür YZ sistemleri geliştirilmesi nefret söyleminin daha net şekilde tespitine ve dolayısıyla yayılmasının önlenmesine yardımcı olabilir.

Adillik ve şeffaflık standartları: Sosyal medya platformlarının algoritmaları tarafından kullanılan eğitim verisi türleri, içerik moderasyonu ve karar süreçleri hakkında açıklamalar sağlayarak kendi YZ kullanımı konusunda şeffaf olması şarttır. YZ sistemleri kullanımında şeffaflık standartları oluşturulması ve bu konuda düzenli raporlar yayınlanması, platformlarda nefret söylemini tespit için daha etkili yöntemler geliştirilmesine yardımcı olabilir.

Yapay zekâ etik standartlarına uygunluk: Sosyal medya platformları, algoritmalarında, şeffaflık, adalet, kötü niyetli olmama, sorumluluk ve gizlilik sağlayan güncel YZ etik ilke ve yönergelerini benimsemelidir. Bundan başka, sosyal medya algoritmalarını düzenli olarak denetlemek üzere etik uzmanları, sosyologlar, hukukçular, teknoloji uzmanları, kullanıcı temsilcileri, sivil toplum kuruluş ve toplulukları da dahil olmak üzere çeşitli paydaşlardan oluşan bağımsız oluşumların kurulması, etik dışı uygulamaların tespitine ve nefret söylemi filtreleme mekanizmalarının etkinliğinin değerlendirilmesine yardım edebilir. Bu disiplinler arası yaklaşım, etik meselelerin kapsamlı şekilde ele alınmasını sağlar. Böylece bu oluşumlar, sistemlerinin çeşitli etik anlayışlara uyacak şekilde özelleştirilmesi için tavsiyelerde bulunabilir ve nefret söylemini önlemeye yönelik içgörüler sunabilir.

İnsan moderatör ve kullanıcıların yapay zekâ ile güçlendirilmesi: Sosyal medya platformlarındaki etik YZ sistemleri, potansiyel nefret söylemi içeriğinin tahmin edilip işaretlenmesine yardımcı olabilir ve böylece nefret söyleminin yayılmasını önleyebilir. Bu aynı zamanda insan moderatörlerin iş yükünü azaltabilir ve moderasyon doğruluğunu artırabilir. Ayrıca, nefret söylemi dalgalarına hızlı yanıt vermek için YZ destekli kriz yönetimi protokolleri oluşturmak da nefret söyleminin yayılmasını önlemeye yardım edebilir. Ayrıca, YZ destekli filtrelerin geliştirilmesi, kullanıcıların nefret söylemini bildirip içerik maruziyetini yönetmelerini sağlayarak proaktif içerik moderasyonunu mümkün kılabilir.

Platformlar arası işbirliği: YZ etik ilke ve yönergelerinin kullanılması, doğası gereği karmaşık ve çok yönlüdür. Bu nedenle, nefret söylemine karşı tutarlı bir küresel strateji için farklı platformlar arasında anlaşmalar yapmak ve etik YZ kullanımında en iyi uygulamaları paylaşarak nefret söylemiyle mücadele çabalarını koordine etmek için platformlar arasında bağımsız komiteler oluşturarak işbirliği sağlamak, nefret söylemiyle etkili mücadeleye yönelik içgörü ve stratejilerin paylaşılması bakımından faydalı olabilir.

Eğitim kampanyaları: Sosyal medya platformları nefret söyleminin tespit edilerek ortadan kaldırılmasında etik YZ kullanımına yönelik eğitim kampanyaları başlatabilir. Ayrıca, etik YZ sistemleri nefret söyleminin etkisini zayıflatmak pozitif ve eğitime yönelik karşı anlatıları teşvik etmek ve dolayısıyla daha olumlu ve yapıcı ilişkiler kurmak için de kullanılabilir.

NEFRET SÖYLEMİYLE MÜCADELE: POLİTİKA VE DÜZENLEMELERİN ÖTESİ

Devletin hukuki düzenlemeleri ve sosyal medya platformlarının nefret söylemini yasaklayan ve hatta cezalandıran politikaları ayrımcılıkla mücadeleyi amaçlasa da, bir önceki bölümde de belirtildiği gibi, bu çabalar daima ifade özgürlüğü ile dengeli şekilde yürütülmelidir. Zira bu düzenlemelerin icrası bazen ifade özgürlüğünü tehdit edebilir. Bazı durumlarda, zararlı dil ile meşru muhalefet arasındaki çizgi bulanıklaşıp, kısıtlamaların aşırıya kaçabileceği, açık tartışma ve eleştirel düşünceyi boğabileceği yönünde endişelere sebep olur. Dahası, bu tür yasalar totaliter rejimlerin elinde muhalefeti susturmak için bir araç haline gelebilir. Örneğin Türkiye’de “kin ve düşmanlığa tahrik” gibi yasal hükümler dezavantajlı grupları korumaktan ziyade muhalefeti bastırarak şekilde uygulandığı için eleştiriliyor. Bu durum özellikle, odak noktasının ayrımcılıkla mücadeleden muhalif seslerin hedef alınmasına kaydığı siyasi ortamlarda bu düzenlemelerin kötüye kullanılma ihtimalini gösteriyor.

Sosyal medya platformlarında, nefret söyleminin yasaklanmasını hedef alan net topluluk ilkeleri olsa dahi, bu tür içeriği tespit edip sınıflandırmadaki güçlükler, yukarıda ayrıntılı bir şekilde ele alındığı üzere, tutarsız uygulamalara yol açıyor. Nefret söyleminin incelikli veya zımni biçimlerini tespit etmekte yaşanan doğal zorluklardan ötürü, bu tür içeriklerin çoğu internet ortamında dolaşmaya ve dolayısıyla zararlı anlatılar üretip daha da pekiştirmeye devam ediyor. Dolayısıyla, bu platform politikaları işlevsel olmakla birlikte, nefret söyleminin yayılmasını engellemiyor.

Temel meselenin, nefret söylemini ifade özgürlüğünü çiğnemediği tanımlamak olduğunu kabul etmek son derece önemlidir. Aşırı detaylı yasal düzenleme ve ağır uygulama koşulları korku ve otosansür ortamı yaratıp, insanların meşru şikâyetlerini dillendirmekten veya eleştirel tartışmalara girmekten çekinmesine yol açabilir.

Dahası, bu düzenleme ve yasaklamalar toplumsal önyargıların kökeninde yatan sebeplere kalıcı bir çözüm sunamayabilir. Tek başına nefret söylemini yasaklamak, yapısal eşitsizlik gibi temelde yatan ve nefret söylemine yol açan faktörlerin değerlendirilmemesi anlamına geliyor. Kapsayıcı bir ortamı kucaklamak için sa-

dece yasal düzenlemelere bel bağlamak yeterli olmayıp, otoriter örneklere baktığımızda anlaşıldığı üzere, tehlikeli olma potansiyeline de sahiptir.

Meselelerin tespiti, bu konuda veri üretilmesi ve toplum içerisinde bağlamına göre ortaya çıkabilen ayrımcılık biçimlerinin izlenmesi için bu raporda ele alınan teknolojik araçların geliştirilmesinin önemini biz de vurguluyoruz. Ancak aynı zamanda, çözüm önerilerinde bulunurken sorunun kökeninde yatan sebeplerin de ele alınması gerektiğine inanıyoruz. Nefret söylemini konu edinen her türlü girişimin köşe taşlarından biri ifade özgürlüğü olmak durumundadır; zira kapsayıcılık denen şey muhalefeti susturma pahasına desteklenemez.

Bu bölümde nefret söyleminin üretilip yaygınlaştırılmasını engellemeyi ve hedef alınan gruplar üzerindeki olası etkisini azaltmayı amaçlayan çeşitli yaklaşım ve girişimler incelenecek.

Öncelikle Susan Benesch, nefret söylemi sorununa olası bir çözüm olarak karşı söylem kavramını inceleyecek. Ardından sosyal medya kampanyalarının taşıdığı potansiyeli göstermek amacıyla vaka analizleri sunulacak. Son olarak da eğitimin hayati rolü ve çocuk ve gençlerle etkileşime geçme gerekliliği vurgulanacak. Bu bağlamda sivil toplum örgütlerinin katkıları son derece önemli olup, müspet değişimlere büyük oranda tesir etme, diyalog geliştirme konusunda hayati bir rol oynama ve eğitim ve farkındalık çalışmalarıyla kapsayıcılığı teşvik etme becerileri yüksektir. Ayrıca, nefret söylemi ve ayrımcılığın temelinde yatan sebeplerin ele alınması için bu çözümlerin hızla uygulamaya geçirilmesi gerektiğini fark etmek de son derece önemlidir.

Nefret söylemini zayıflatmak için karşıt söylem yaklaşımları

Susan Benesch

Çevrimiçi ortam da dahil olmak üzere nefret söylemi konusunda neler yapılabileceğine dair artan endişeler ışığında bazı ülkeler (bu yazı yazıldığı esnada son olarak İskoçya) nefret söylemine yönelik hukuki cezaları yükseltme yoluna gidiyor; ancak söz konusu yasalar ifade özgürlüğünü de tehdit edip, çoğu zaman azınlıkları veya muhalif sesleri susturmakta kullanılıyor. Oysa başka bir seçenek daha var: Çevrimiçi söylemin iyileştirilmesi için tabana yayılmış çabalar. Binlerce internet kullanıcısı nefret söylemine doğrudan yanıt vermek için düzenli olarak çeşitli iletişim stratejilerini kullanmak suretiyle bu tür söylemleri çürütüyor veya zayıflatıyor.

Bazıları nefret içerikli yanlış dezenformasyonları düzeltmek için olaylara dayalı gerçek bilgiler sunarak başka okurların yalan bilgilere kanmasına engel oluyor. Kimileri daha fazla kullanıcının göreceği daha geniş forumlarda gönderimde bulunarak nefret söylemi hakkında internet kullanıcılarını eğitmeye çalışıyor. Verdikleri yanıtların bir kitleye ulaşması için (çünkü insanlar bir hesabı komik buldukları için takibe alabiliyorlar) veya bu türden rahatsız edici içeriklere yanıt vermenin duygusal yükünü azaltmak için mizaha başvuranlar da oluyor. Son bir yöntem de nefret söylemini yayan kişilerle empati kurmaya çalışmaktır. Empati, nefret içerikli söylemin asıl sahibini bu davranışını veya düşüncelerini (bu çok zor bir şey olmakla birlikte) değiştirmeye yönlendirebilecek bir anlayış duygusunun aktarımını sağlayabilir. Karşıt söylem başvuranların birçoğu bunu tek başına yaparken, bazıları da bu karşılıkları koordine edip birbirlerini desteklemek için gruplar kuruyor.

Karşıt söylem yeni bir kavram değil fakat bu konudaki tartışma ve çalışmalar son yıllarda artış gösterdi. ABD’de karşı söylem kavramı genellikle, 1927 tarihli *Whitney/California* davasında, nefret içerikli söylemi sansürlemek yerine ona yanıt vermenin en iyi yöntem olduğunu öneren Yüksek Mahkeme Yargıcı Louis D. Brandeis’e gönderme yaparak açıklanıyor. Amerika Komünist İşçi Partisini kurmaya çalışmış olan Kaliforniyalı bir kadın hakkında verilen mahkûmiyet kararını onaylayan bir kararda Yargıç Brandeis şu sözleri sarf etmişti:

“Şayet yalan ve safsataları tartışma yoluyla açığa çıkarmaya, kötülüğü eğitim süreçleriyle geri çevirmeye vakit varsa, başvurulması gereken çare zorla susturmak değil, daha fazla konuşulmasına izin vermektir.”

Yargıç Brandeis karşıt söylem terimini kendisi hiç kullanmamış olsa da, ABD’li hukukçular Brandeis’in bu kanaatine sık sık karşı söylem doktrini olarak atıf yaparlar.

Karşıt söylem başlığı altında gitgide sayıları artan akademik çalışmaların çoğu karşıt söylemin etkili olup olmadığına odaklanıyor. Birçok araştırmacı bu soruyu yanıtlamaya çalışmış olsa da, etkililiğin nasıl tanımlanacağı da dahil olmak üzere birtakım güçlüklerle karşılaşılıyorlar. Karşıt söylemin, kullanılan diskur üzerinde birçok açıdan müspet bir etkisi olabilir. İnsanların kanaatlerini veya sadece davranışlarını değiştirerek zararlı dil kullanan gönderiler yapmalarını engellemek konusunda ikna edici olabilir (İnsanlar bir düşünceyi alenen ifade etmeleri halinde eleştirilmekten veya toplumsal yaptırımlara uğramaktan korktuğu için, söz konusu düşünceye halen sahip olmakla birlikte davranışlarını değiştirmeleri mümkün olabilir). Kullanılan söylem, nefret içerikli gönderilerde bulunan insanların düşüncelerinde veya çevrimiçi ifadelerinde herhangi bir değişiklik olmadan da iyileşme gösterebilir. Karşıt söylem kullanarak “kitleye”, yani yorumları okuyan insanlara nüfuz etmek mümkün olabilir. Karşıt söylem kullananların paylaştığı görüşlere halihazırda katılan, fakat kendi görüşlerini dile getirecek kadar cesur hissetmeyen kişiler açısından, karşıt söylem cesaretlendirici olabilir ve böylelikle zaman içerisinde, hiçbir düşüncede değişiklik olmasa bile, diskuru karşıt söylemde ifade edilen görüşlerden yana olacak şekilde değiştirebilir. Kitleye dahil olan diğer insanlar tartışılan konu hakkında henüz sabit bir görüş geliştirmemiş olabilir ve karşıt söylem bu insanların zararlı nitelikteki yanlış bilgilere inanmasını veya bunları paylaşmasını engelleyebilir.

Karşıt söylem, adaletsizliği, ayrımcılığı veya eşitsizliği idame ettiren egemen anlatılara ya da ideolojilere meydan okunmasını da içeren daha geniş bir strateji olan karşı anlatıdan farklıdır. Bu anlatılar genellikle tarihsel olaylara, toplumsal meselelere, kültürel kurallara veya siyasi ideolojilere dair alternatif bir yorum, analiz veya anlayış içerir.

Karşıt söylem ile karşıt anlatı zararlı ideolojilere meydan okumak bakımından benzer amaçlar taşımakla birlikte, kapsam, odak ve metodolojileri açısından farklılaşırlar. Karşıt söylem öncelikle münferit nefret söylemi örneklerini ele alarak, bu örneklerdeki belirli argüman veya mesajları çürütmeyi yahut zayıflatmayı amaçlar. Karşı anlatı kampanyaları ise sistematik baskı ve ötekileştirmeyi devam ettiren geniş kapsamlı anlatıları eleştirerek yıkmayı amaçlar.

Aktivizm aracılığıyla nefret söylemiyle mücadele: İki başarılı örnek

İlayda Ece Ova

Hem ulusal hem de uluslararası sivil toplum kuruluşları (INGO) düzeyinde nefret söylemiyle mücadele amacıyla yürütülen kurumsal ve hukuki çabalar, geniş kaynakları ve etki alanları sayesinde hayli görünür niteliktedir. BM, nefret söylemiyle mücadele konusunda özel bir takvim benimsemiş, 2019'da BM Nefret Söylemine Yönelik Strateji ve Eylem Planı'nı kabul etmiş ve 18 Haziran gününü Uluslararası Nefret Söylemiyle Mücadele Günü ilan etmiştir. AB ise çeşitli organlar aracılığıyla farklı sektör ve ülkelerde yürütülen projelere fon sağlayarak bu alanda çalışan sivil toplum örgütlerine sürekli finansal destek sunuyor. **Nefret Söylemi ve Nefret Suçuyla Mücadele Üst Düzey Çalışma Grubu**³⁴ AB'nin nefret söylemiyle ilgili gündeminin belirlendiği platform olup, **CERV programı**³⁵ da nefret söylemiyle mücadele çalışmalarına temel oluşturan AB hak ve değerlerini koruyup pekiştirmeyi amaçlayan finansal çerçeveyi sağlıyor. Kamu kurumlarına ve spor alanındaki paydaşlara teknik destek sunarak sporda nefret söylemiyle mücadele etmeyi amaçlayan **Sporda Nefret Söylemiyle Mücadele projesi**;³⁶ cinsel yönelim, cinsiyet kimliği veya ifadesine dayalı nefret içerikli ve hoşgörüsüz söylemleri, şiddet ve ayrımcılığı ele alıp önlemeyi amaçlayan **LGB-TİQ Karşıtı Şiddet ve Nefret Söylemiyle Mücadele** projesi;³⁷ nefret suçlarıyla mücadelede farklı kuruluşlar arasında işbirliğini geliştirmeye yönelik kurumlararası bir model olan **Stand-Up Projesi**;³⁸ ve raporlama araçları ve yönlendirme mekanizmaları geliştiren uluslararası ve kurumlararası bir işbirliği olan **Nefre-**

34 Avrupa Komisyonu. (2023). *Nefret Söylemi ve Nefret Suçuyla Mücadele Üst Düzey Çalışma Grubu*. <https://ec.europa.eu/transparency/expert-groups-register/screen/expert-groups/consult?lang=en&groupID=3425>

35 Avrupa Komisyonu. (2024) *Vatandaşlık, Eşitlik, Haklar ve Değerler programı (CERV)*. Avrupa Komisyonu Fon ve İhale Portalı. <https://ec.europa.eu/info/funding-tenders/opportunities/portal/screen/programmes/cerv>

36 Avrupa Konseyi. (2024). *Sporda nefret söylemiyle mücadele: Proje hakkında*. <https://pjp-eu.coe.int/en/web/combating-hate-speech-in-sport/about-the-project>

37 Avrupa Konseyi. (2024). *LGBTİQ Karşıtı Şiddet ve Nefret Söylemiyle Mücadele*. <https://www.coe.int/en/web/sogi/combating-anti-lgbtqi-violence>

38 **STAND-UP Projesi**. (2024). *STAND-UP: AB'de nefretle mücadele*. <https://stand-up-project.eu/>

te Karşı Gerçekler³⁹, bu alanda dikkat çeken proje ve programlar arasında yer alıyor. Bu projeler, amaçları ve sonuçları farklı olmakla birlikte, sektörler arası işbirliği, veri toplama iyileştirmeleri, dijital araçlar, söylemden etkilenen kişi ve topluluklara destek, spor ve sanal ortam gibi alanların yanı sıra ötekileştirilmiş topluluklar içerisinde ayrımcılıkla mücadeleye yönelik çabalara odaklanıyor.

ABD’de savunuculuk, eğitim, davalar ve toplumsal katılım yoluyla nefret söylemiyle mücadele eden başlıca kuruluşlar arasında şunlar sayılabilir: **The Anti-Defamation League**⁴⁰ (ADL – İftira ve İnkârla Mücadele Birliği) **Stop Hate for Profit**⁴¹ gibi girişimlerle radikalizm ve antisemitizm konularını ele alıyor; **Southern Poverty Law Center**⁴² (SPLC – Güney Yoksulluk Hukuk Merkezi) ise **Nefret Haritası**⁴³ aracılığıyla nefret söylemi üreten grupları takip edip farkındalık uyandırıyor. **Human Rights Campaign**⁴⁴ (HRC – İnsan Hakları Kampanyası) LGBTQ+ bireyleri hedef alan nefret söylemine odaklanırken, **American Civil Liberties Union**⁴⁵ (ACLU – Amerika Temel Haklar Birliği) özgür ifadeyi savunurken, zararlı etkilerini de azaltmayı amaçlıyor. **The Center for Countering Digital Hate**⁴⁶ (CCDH – Dijital Nefretle Mücadele Merkezi) sanal nefret ve dezenformasyonla mücadele ederken, **Faith in Action**⁴⁷ inançlararası dayanışma yoluyla toplulukların dayanıklılığını artırıyor. **NAACP**⁴⁸ (Siyahi İnsanların Gelişmesi için Uluslararası Birlik) ırksal nefret ve ayrımcılıkla mücadele ediyor, **CAIR**⁴⁹ (Amerikan-İslam İlişkileri Konseyi) ise İslam düşmanlığını ele alı-

39 Avrupa Birliği Temel Haklar Ajansı (FRA). (2024). *Nefret Karşıtı Gerçekler*.

<https://fra.europa.eu/en/promising-practices/facts-against-hate-o>

40 Anti-Defamation League (ADL). (2024). *İftira ve İnkârla Mücadele Birliği*. <https://www.adl.org/>

41 Anti-Defamation League (ADL). (2024). *Kâr için Nefreti Köruklemeyi Bırak*.

<https://www.adl.org/stop-hate-profit-o>

42 Southern Poverty Law Center (SPLC). (2024). *Güney Yoksulluk Hukuk Merkezi*.

<https://www.splcenter.org/>

43 Southern Poverty Law Center (SPLC). (2024). *Nefret Haritası*.

<https://www.splcenter.org/hate-map/>

44 Human Rights Campaign (HRC). (2021). *İnsan Hakları Kampanyası*. <https://www.hrc.org/>

45 American Civil Liberties Union (ACLU). (2024). *Amerika Temel Haklar Birliği*.

<https://www.aclu.org/>

46 Center for Countering Digital Hate (CCDH). (2022). *Dijital Nefretle Mücadele Merkezi*.

<https://counterhate.com/>

47 Faith in Action. (2024). *İnanç Hareketi*. <https://faithinaction.org/>

48 National Association for the Advancement of Colored People (NAACP). (2024). *Siyahi İnsanların Gelişmesi için Uluslararası Birlik*. <https://naacp.org/>

49 Council on American-Islamic Relations (CAIR). (2024). *Amerikan-İslam İlişkileri Konseyi*.

<https://www.cair.com/>

yor. **GLAAD**⁵⁰ (Hakarete Karşı Eşcinseller Birliği) medyadaki LGBTQ+ tasvirlerini izleyip olumlu anlatıları teşvik ederken, **Not In Our Town**⁵¹ (NIOT – Kasabamızda İzin Vermeyiz) yerel toplulukları nefrete karşı tabandan harekete geçmek üzere güçlendirirken çok yönlü bir ulusal tepki yaratıyor.

Nefret söylemi karşıtı girişimlerin ulusal ve uluslararası kuruluşlar tarafından yaygın şekilde benimsenmesi görünürlüğü oldukça arttırırken, kurumsal ve yasal çözümlerin güçlenmesine de katkıda bulunuyor. Ancak bu durum, daha katı yasal düzenlemelerin ve büyük fonlarla desteklenen programların nihai çözüm olduğu algısına yol açabilir. Bunlar faydalı olmakla birlikte, tüm dünyada sağcı ve otoriter rejimlerin yükselişi, devlet öncülüğündeki girişimlerin kırılganlığını ve kapsayıcı politikaların nasıl da tersine çevrilerdiğini gösteriyor. Devlet destekli kapsayıcı girişimler, sağcı, kadın düşmanı ve LGBTİ+ karşıtı ideolojilerin etkisi altında gerilerken, nefret söylemine karşı aktivizm giderek daha kritik ve hayati hale geliyor. Bu da sivil toplum ve yurttış öncülüğündeki tepkilerin önemini arttırıyor. Uzun vadeli toplumsal değişim için sivil toplum örgütleri, taban hareketleri ve kendini bu harekete adanmış kişilerin önderliğinde bağlama uygun, yerelleştirilmiş ve kültürel açıdan anlamlı aktivizm faaliyetlerinin yürütülmesi gerekiyor.

Nefret söylemine karşı tabandan yükselen aktivist hareketler, bağlama özgü eylemlerin duruma has zorluklarını ele alırken, uluslararası ağlar arasında bilgi alışverişini ve etkili uygulamaları da kolaylaştırdığına örnek oluşturuyor. Nefret söylemi karşıtı kampanyalar, farkındalık yaratma, hoşgörüyü teşvik ve kapsayıcı toplumlar inşa etmek gibi stratejilerde ortaklaşıyor. Uzun vadeli dayanıklılığı beslemek için eğitim girişimlerine, toplumsal katılıma ve gençlerin mobilizasyonuna odaklanıyorlar. Sosyal medya, karşı anlatıların yayılmasında kilit rol oynayarak, sanal ortamdaki çatışmaları yatıştırmak için çatışmacı olmayan tepkilere ağırlık veren birçok kampanya yürütülüyor. Nefretin yayılmasında dezenformasyonun ne kadar yer kapladığı düşünüldüğünde, doğruluk kontrolü ve yanlış bilgiyle mücadele elzemdir. Kampanyalar ayrıca mağdurlara psikolojik, hukuki ve çevrimiçi kaynaklar aracılığıyla yardım sağlarken, geliştirilmiş platform moderasyonu ve politikalarını destekler. Birçok kampanya basit ama tesir gücü yüksek mesajlar aktarmak için kültürel karşılığı olan semboller kullanıyor. Devlet, STK ve özel sektörle yapılan ortaklıklar politikalar üzerinde etki yaratmaya ve içerik denetimini güçlendirmeye yardımcı oluyor. Küresel koordinasyon ve yerel uyarlanabilirlik sayesinde farklı bağlamlarda nefret söylemiyle etkin bir şekilde mücadele edilebiliyor.

⁵⁰ GLAAD. (2024). Hakarete Karşı Eşcinseller Birliği. <https://glaad.org/>

⁵¹ Not In Our Town (NIOT). (2024). Kasabamızda İzin Vermeyiz. <https://www.niot.org/>

Bu tür yaklaşımlara örnek olarak nefret söylemiyle yenilikçi, kültürel açıdan anlamlı stratejiler üzerinden etkili bir şekilde mücadele eden iki örnek kampanyadan bahsedeceğiz.

Panzagar (Çiçek Söylemi) Kampanyası

Panzagar (“çiçek söylemi”) kampanyası⁵² Myanmarlı gençlik lideri Nay Phone Latt tarafından Müslümanlara karşı yükselen şiddete cevap olarak 2014 yılında başlatıldı. Bu kampanya sembol olarak Myanmar’da barışı temsil eden çiçekleri kullanıyordu. Sivil toplum tarafından hazırlanan raporlar, ülkede çevrimiçi nefret söyleminin öncelikle Müslümanlara yöneltildiğini gösteriyordu. Bu kampanyanın temel stratejisi, genç yerel grafik tasarımcılarla işbirliği yaparak ağızlarından çiçekler çıkan karakterlerin yer aldığı anime benzeri görseller oluşturmak ve bunları sosyal medyada yaygın şekilde dolaşıma sokmaktı.⁵³

Şekil 2: Panzagar kampanyasının ilk görsellerinden biri⁵⁴



Facebook 2014 yılında, Myanmar’da bu platformun kötüye kullanımına ilişkin bir kullanıcı bildirim mekanizması başlattı, ancak ülkenin kendine özgü demografik yapısı ve hukuki bağlamı bu hedefe yönelik ek adımlar atılmasını gerektiriyordu.

⁵² Beautiful Trouble. (2024). Çiçek söylemi kampanyası. Beautiful Trouble Toolbox. <https://beautifultrouble.org/toolbox/tool/flower-speech-campaign>

⁵³ Panzagar. (2014). Panzagar Facebook sayfası. Facebook. <https://www.facebook.com/panzagar>

⁵⁴ Support Flower Speech. (2014). Çiçek Söylemini Destekle Facebook sayfası. Facebook. <https://www.facebook.com/supportflowerspeech/>

O dönemde Myanmar dünyanın en düşük internet ve cep telefonu kullanım oranlarına sahip ülkeler arasındaydı ve devletin yaygın bir sansür politikası vardı. Bu nedenle kampanya hem çevrimiçi hem çevrimdışı bileşenlerle yürütüldü ve Müslümanların Budist çoğunluk karşısında tehdit olarak gösterilmesine karşı çıkan Budist kuruluşlardan ilhamla şekillendirildi.

Aktivist Thinzar Shunlei Yi, #FlowerSpeech kampanyasının güçlü ve zayıf yönlerini değerlendirirken, kampanyanın sosyal medya, halka açık etkinlikler, müzik ve çıkartmalar aracılığıyla geniş katılımı teşvik etmeyi başardığını ve pasif destekçileri aktif katılımcılara dönüştürdüğünü vurguluyor. Kampanya, Budizm ilkelerinden biri olan “doğru konuşma”ya atıfla, kötüye kullanım içermeyen etik iletişimi teşvik ediyordu. Başarısı da büyük ölçüde, insanların ağızda kültürel açıdan önemli bir imge (Burma’nın ulusal çiçeği *padauk*) tutmasıyla ve böylelikle mesajın kolayca anlaşılmasıyla sağlandı. Kampanya ilk başta çizimlerin cinsiyetçi öğeler içermesi, Myanmar’a özgü imgeler içermemesi ve kalıp yargılara dayanması sebebiyle eleştirilmişti. Bu eleştiriler karşısında söz konusu görseller değiştirildi. Yöneltilen diğer bir eleştiri de, ağızda çiçek tutulmasının, nefrete karşı aktif direniş yerine sessizliği teşvik ediyor olarak yorumlanabileceğiydi. Tüm bunlara rağmen, kampanya nefret söylemini zayıflatıp yapıcı diyalogu teşvik etmek açısından etkili araçlar sağladı.

#IAmHere (#Buradayım) Uluslararası Kampanyası

#IAmHere (#Buradayım) hareketi İsveç’te, İran doğumlu gazeteci Mina Dennert tarafından #JagArHar adıyla başlatıldı.⁵⁵ Dennert, sosyal medyada nefret dolu içeriklerin arttığını görünce buna karşı sakin ve yapıcı bir mücadele amacıyla bu girişimi başlatmıştı. Medyanın ilgisini çeken kampanya hızla büyüyerek İsveç’te kısa sürede 75.000 üyeye ulaştı ve İtalya, Fransa, Slovakya, Polonya ve Birleşik Krallık’ta benzer grupların kurulmasına ilham oldu. Bugün dünya çapında faaliyet gösteren on dört farklı #IAmHere grubunun müşterek misyonu, gerçeklere dayalı, çatışmadan uzak, yapıcı bir diyalog yoluyla çevrimiçi nefret söylemine karşı koymak, nefretin hedefindeki kişilere destek olmak ve daha sağlıklı dijital ortamları savunmak.

Uluslararası kampanya ise doğruluk kontrolüne ve dezenformasyonun yayılmasının önlenmesine odaklanıyor. Gönüllüler sosyal medyadaki nefret içerikli yorum-

⁵⁵ BBC News. (2019, June 10). ‘#IAmHere’: The people trying to make Facebook a nicer place (“#Buradayım”: Facebook’u daha iyi bir yer haline getirmeye çalışan insanlar). <https://www.bbc.com/news/blogs-trending-48462190>

lara itiraz etmek, nefret söyleminin hedef aldığı insanlara destek olmak ve yasal düzenlemeler getirilmesini savunmak üzere harekete geçiyor.⁵⁶ Dijital ortamda dezenformasyon genellikle zararlı anlatıları pekiştiren ve çevrimiçi ortamdaki tartışmaları manipüle eden trol ve bot ağlarının eşgüdümlü çabalarıyla yayılıyor. Askerî benzeri hiyerarşilerle çalışan trol gruplar, sohbetleri domine etmek, ılımlı sesleri bastırmak ve saldırgan bir şekilde dezenformasyonu yaygınlaştırmak amacıyla yapılandırılmış ve stratejik bir eşgüdüm mekanizmasını kullanıyorlar. #IAmHere hareketi bu taktikleri püskürtmek için benzer şekilde örgütlenen bir yaklaşım benimseyerek, üyelerini nefret söylemine ve yalan haberlere karşı çıkmak için odaklandırılmış ve stratejik bir biçimde seferber ediyor. Kampanyanın son adımı olan “nefreti bastır”, nefret içerikli yorumların görünürlüğünü azaltacak algoritma ve platform düzenlemelerini savunuyor. Ancak, daha önce belirtildiği gibi, sadece sosyal medya platformlarının politika ve algoritmaları iyileştirmelerine bel bağlamak bazı sorunlar doğuruyor. Bu nedenle, kampanyanın ifade özgürlüğüne ve doğruluğu teyit edilmiş yapıcı yanıtlara odaklanması, nefret söylemine karşı tabandan gelen ve sivil toplum öncülüğünde yürütülen çabalar için dünya çapında son derece önemli bir model oluşturuyor.

56 #IAmHere International. (2024). *Hakkımızda*. #IAmHere International.
<https://iamhereinternational.com/about-us/>

Pozitif etkileşimi teşvik etmek için alternatif yaklaşımlar

Peter Adams, Metin V. Bayrak, Alex Mahadevan

Kapsayıcı anlatılar için Influencer'lardan yararlanmak

Popüler içerik üreticileri yanlış bilgi ve nefret söyleminin yayılmasına aracılık ediyor olabilir, öte yandan influencer'lar ve tanınmış içerik üreticileri çocuklar ve gençler üzerindeki güvenilirlikleri ve geniş kitlelere ulaşma kapasiteleri dolayısıyla dijital medya okuryazarlığı ve nefret söylemiyle mücadele açısından büyük bir fırsat da oluşturuyor. Tıpkı araştırmacıların YZ kaynaklı dezenformasyonla mücadele için yapay zekâ temelli araçlar geliştirmesi gibi, STK'lar, haber kuruluşları ve diğer sivil toplum örgütleri de geleneksel dezenformasyon araçlarını kullanarak yalan yanlış bilgilerle mücadele edebilir. Influencer'lar ve içerik üreticileri YouTube, TikTok, Instagram ve Twitch gibi platformlarda geniş etkiye sahipler; dolayısıyla sosyal medya şirketlerinin bu kişilerin yanlış bilgi veya nefret söylemini yaygınlaştırmamasını sağlamak konusunda bir sorumluluğu bulunuyor. Facebook, Instagram (Meta) ve TikTok gibi platformlar doğruluk kontrolü yapan kuruluşlarla işbirliği yapıyor ve yukarıdaki bölümlerde açıklandığı üzere, tüm platformların nefret söylemi ve ayrımcılık konusunda topluluk ilkeleri var. Fakat bu ilkelerin ne denli etkili olabildiğine dair ciddi ve haklı eleştiriler yöneltiliyor. Üstelik Meta Ocak 2025 itibarıyla doğruluk kontrolü programına son verdiği için, platformlarındaki içerik moderasyonunun güvenilirliği konusunda endişeler arttı. Yine de bu noktada başarılı bir örnekten bahsetmek gerek: **Avustralya Associated Press kuruluşu TikTok üreticileriyle birlikte çalışarak**⁵⁷ yanlış bilgilerin çürütülmesi ve medya okuryazarlığı becerilerinin inşası amacıyla bu platforma yönelik müdahaleler geliştiriyor.

Influencer'ların ve içerik üreticilerinin bir başka gücü de, reklam veya kamu spotu gibi "görünmeyen" içerikler üretme becerilerinden kaynaklanıyor. Bu kişiler aynı zamanda "viral" olma potansiyeli en yüksek video ve görselleri de

⁵⁷ **Mediaweek.** (28 Şubat, 2024). *TikTok ve AAP, içerik üreticilerini yanlış bilgiyle mücadelede güçlendirmek için iş birliği yapıyor.* Mediaweek.
<https://www.mediaweek.com.au/tiktok-and-aap-partner-to-empower-creators-to-fight-misinformation/>

ustalıkla oluşturuyorlar. Poynter Medya Çalışmaları Enstitüsünün medya okuryazarlığı girişimi olan MediaWise, 2018 yılından beri tanınmış gazeteciler, YouTuber'lar, yazarlar, güzellik ve mühendislik alanındaki influencer'lar, LGBTQ+ odaklı içerik üreticileri ve sporculardan oluşan bir Elçi Programı yürütüyor. Bu kapsamda örneğin, Guatemala'dan badminton şampiyonu ve olimpiyat sporcusu Kevin Cordón, WhatsApp üzerinden sunulan bir kurs için **medya okuryazarlığı videoları** çekti.⁵⁸

Stanford Üniversitesi tarafından yapılan bir çalışma, yarattığı etki açısından, yaşlanma konularıyla ilgilenen ABD'li gazeteci ve influencer Joan Lunden'in etrafında şekillendirilen kursların, yanıltıcı haber başlıklarını ayırt etme becerisini artırdığını ortaya koyuyor.⁵⁹

Son olarak, Ulusal Medya Okuryazarlığı Eğitimi Derneği'nin müfredat oluşturulmasına katkı sunduğu YouTube'un **Hit Pause** kampanyası⁶⁰ **çocuk televizyon yıldızı Blippi** gibi influencer'larla işbirliği yaptı.⁶¹ Her bir video en az 40 milyon görüntülemeye ulaşırken, bazıları 200 milyonun üstüne çıktı.

Bu örnekler ışığında, sivil toplum örgütlerinin sosyal medyanın nefret söylemini azaltma potansiyelini yükseltebileceği birkaç strateji önerilebilir:

- **Hikâye Anlatıcılığı Kampanyaları:** Influencer'lar, nefret söylemi ve yanlış bilgidен etkilenen insanların kişisel hikâye veya anlatılarını paylaşarak bu meseleyi insani bir boyuta taşıyıp empati duygusunu teşvik edebilir. Bu tür kampanyalar etkileşim ölçütleri kullanılarak ve yorum ve paylaşımlardaki duyguların analizi yapılarak ölçülebilir.
- **Eğitim Dizileri:** Nefret söylemi ve yanlış bilgi ile ilgili karmaşık konu başlıklarını bölerek anlatan bir dizi eğitim içeriği yaratılması için influencer'lar ile işbirliği yapılabilir. Görüntülenme sayıları, dizilerin izlenip tamamlanma

58 Poynter Enstitüsü. (2024). Guatemala'da MediaWise. <https://www.poynter.org/mediawise/international/guatemala/>

59 Poynter Enstitüsü. (2020, December 10). Yeni Stanford çalışması, Poynter'in MediaWise eğitiminin insanların dezenformasyonu tespit becerilerini büyük oranda arttırdığını söylüyor. <https://www.poynter.org/news-release/2020/poynters-mediawise-training-significantly-increases-peoples-ability-to-detect-disinformation-new-stanford-study-finds/>

60 YouTube. (2023). Aileler için Dijital Esenlik. YouTube. <https://www.youtube.com/playlist?list=PL4SOO4mxq3nsE6nzCV-QDGzg6VF7cFzGh>

61 YouTube. (2023). YouTube: Komik bir esneme molasıyla durakla! <https://www.youtube.com/watch?v=xdKDcKuRs3M>

oranları ve anlayış ve tutumlardaki değişimleri değerlendirmek amacıyla kampanya öncesi ve sonrası anketler yoluyla başarı ölçümü yapılabilir.

- **İnteraktif Meydan Okumalar:** Influencer'ların küçük yarışma veya soru-cevap videoları gibi interaktif içerik kullanılarak pozitif mesajları teşvik eden veya yanlış efsaneleri çürüten meydan okumalar başlatması istenebilir. Başarı ölçütleri olarak katılım oranları, içeriğin viralleşme oranı (paylaşımlar ve beğeniler) ve yorumlardaki söylemlerin niteliği değerlendirilebilir.
- **Politika Savunuculuğu için Ortaklık:** Nefret söylemi ve yanlış bilgiye karşı politikalar geliştirilmesi savunuculuğu için de influencer'lardan yararlanılabilir. Buradaki başarı, kampanyalarda toplanan imza sayısı, bağlantılı etkinliklere katılım sayısı ve bu yolla yapılan politika değişiklikleri ile ölçülebilir.
- **Dijital Okuryazarlık Atölyeleri:** Influencer'lar ile katılımcıların yanlış bilgi ve nefret söylemini tespit edip buna karşılık vermek konusunda bir şeyler öğrenebileceği canlı oturumlar yapmak. Bu etkinliğin değerlendirmesi katılım oranları, oturumlar sırasındaki etkileşimler ve geri bildirim formları üzerinden yapılabilir.

Influencer'lar ile çalışmanın bazı riskler barındırabileceği unutulmamalıdır, zira belirli bir program dışında ne yaptıklarını veya söylediklerini kontrol etmek mümkün olmayacaktır. Hele ki streamer'lar yıllar içerisinde binlerce saat içerik oluşturmuş olabilir ve bunların bir kısmı sorgulanabilir nitelik içerebilir. İşbirliğine girişmeden önce, kutuplaştırıcı, yanlı veya önyargılı siyasi içerik üretmediklerinden emin olmak için bu kişilerin önceki çalışmalarını titizlikle incelemek gerekir. Ayrıca, içerik geliştirme konusunda influencer ile eşgüdümlü çalışılması ve ortaya çıkan materyalin aktarılacak istenen mesaja uygun ve tarafsız olması için dikkatle düzenlenmesi önerilir. İçerik üretecek kişiye önceden bir senaryo taslağı –hatta senaryonun tamamı– verilmesi de faydalı olur.

Kapsayıcılık İnşası: Gençlere yönelik eğitim yaklaşımları

Nefret söylemi, dezenformasyon ve aşırılık yanlısı retoriğin yaygınlaşması sadece hoşgörüsüzlük, nefret ve şiddetin sürmesine değil, üretken sivil söylemin yozlaşmasına ve herkes açısından temel demokratik zihniyet gelişiminin zayıflamasına yol açar. Ancak biz bu bölümde konuya özellikle çocuklar ve gençler bağlamında odaklanacağız.

Dezenformasyon –özellikle de komplocu içerik ve fikirler– nefret söylemi ve aşırılık yanlısı ideolojilerin yayılmasında asli araçlardan biridir. Aşırı sağcı topluluklar internette oldukça aktif olup bilişsel eğilimleri kötüye kullanmak, mantıksal saf-satalardan yararlanmak, kanıt uydurmak ve güçlü duygusal tepkiler yaratmayı amaçlayan bağlamsal hilelere başvurmak da dahil olmak üzere yaygın dezenformasyon taktiklerini ve komplo imalarını kullanarak genellikle gençleri hedef alıyorlar. Dezenformasyon yayan diğer kesimler gibi, aşırılık yanlıları da anlayış, bağlantı, topluluk, amaç ve eylemlilik arzusu gibi evrensel insan ihtiyaçlarından yararlanır (bunlar bazen “**itme ve çekme**” faktörleri⁶² olarak algılanır). Bu ihtiyaçlar özellikle, sosyal ilişkilere ve olası sosyal kazanımlara daha fazla odaklanma eğiliminde olan gençler ve genç yetişkinler arasında yoğun olarak gözlenir.

Ebeveyn, vasi, eğitimci ve gençlerin hayatındaki diğer tüm paydaşların onları aşırılık yanlısı mesajları ve diğer zararlı dezenformasyon biçimlerini saptayıp bunlara karşı direnç göstermeye hazırlaması hayati önem taşıyor. Bu tür etkiler gençlerin bilgi akışlarının neredeyse her kanalına girmiş durumda. Bunlar anaakım sosyal medya platformlarında, genelde kaçamak, kodlanmış veya sterilize edilmiş dil ve semboller kullanılarak kapalı mesajlaşma gruplarında; hususi mesajlaşma forumlarında ve sosyal paylaşım sitelerinde; beyaz milliyetçi müzikler yoluyla; Discord gibi ilgi alanına dayalı platformlarda ve Twitch gibi oyun akışı ve canlı sohbet platformlarında yapılan gönderi ve yorumlar üzerinden ortaya çıkıyorlar.

Gençlerin nefret söylemi konusunda farkındalığını uyandırmak için yetişkinleri güçlendirmek

Çocuklar ve gençler arasında farkındalık uyandırmak çok önemli olmakla birlikte, başta ebeveynler ve eğitimciler olmak üzere yetişkinler de nefret söylemine karşı kapsayıcı ortamların şekillendirilmesinde çok önemli bir rol oynuyor. Bu kişilerin ayrımcılık konusunu ele alma, eleştirel düşünmeyi teşvik etme ve güvenli alanlar yaratma becerileri, gençlerin zararlı anlatılarla nasıl ilişki kurduğunu ve bunlara nasıl tepki verdiğini büyük oranda etkiliyor.

Nefret söyleminin ele alınması, empati ve hoşgörünün geliştirilmesi gibi eğitim hedefleriyle paralellik arz eder. Bu değerleri anlayan yetişkinler, bunları eğitim uygulamalarına entegre ederek çocukların öğrenme deneyimlerini zenginleştirebilir.

62 TED. (2017). *Gençler aşırılık yanlısı şiddet gruplarına nasıl katılıyor – ve onları nasıl durdurabiliriz*. Youtube. <https://www.youtube.com/watch?v=HY7io88saG4>

Nefret söylemini fark edip buna direnme konusunda çocukları destekleyebilmek için bu yetişkinlerin doğru araç ve bilgilerle donanmış olması gerekir. Yetişkin eğitiminin ayrımcılıkla mücadele çabalarını güçlendirebileceği bazı kilit alanlar şunlardır:

- **Olumlu Sosyal Davranışların Teşviki:** Yetişkinlerin nefret söylemini ele almak konusunda eğitilmesi, çocukların saygı mefhumunu öğrenebileceği ve çeşitliliğe değer verebileceği ortamların yaratılmasına yardımcı olarak küçük yaşlardan itibaren olumlu davranışları teşvik eder.
- **Ruh Sağlığının Korunması:** Nefret söylemi çocukların ruh sağlığına zarar vererek anksiyete, depresyon ve özsaygı düşüklüğüne yol açabilir. Yetişkinleri bu tür söylemleri fark edip bunlarla mücadele etmeleri yönünde bilinçlendirerek çocukların ruh sağlığını korumaya yardımcı oluruz.
- **Kapsayıcı Topluluklar Oluşturmak:** Nefret söylemi bölme ve dışlamayı besler. Bunlarla mücadele becerisi güçlendirilmiş olan yetişkinler, herkesin kendini değerli hissettiği daha kapsayıcı ve uyumlu topluluklar oluşturur.
- **Eleştirel Düşünmeyi Teşvik Etmek:** Nefret söyleminin tehlikelerini fark etmek üzere eğitilmiş yetişkinler çocuklara eleştirel düşünmeyi öğretebilir. Çocukları nefret dolu mesajları sorgulayıp analiz etmeye teşvik etmek, zararlı ideolojilere karşı direnç geliştirmelerine yardımcı olur.
- **Zorbalık ve Tacizin Önlenmesi:** Nefret söylemi çoğu zaman hem çevrimiçi hem çevrimdışı zorbalık ve tacize yol açar. Yetişkinleri bu konuyu ele alacak şekilde donatmak, daha güvenli alanlar yaratılmasına yardımcı olarak zorbalık vakalarını azaltır.
- **İnsan Haklarının Korunması:** Nefret söylemi, onur ve eşitlik gibi temel insan haklarını ihlal eder. Yetişkinleri buna karşı çıkacak şekilde güçlendirmek, insan haklarını koruma ve sosyal adaleti teşvik yönündeki global çabaları destekler.
- **Çocukları Çeşitliliğe Açık Bir Dünyaya Hazırlamak:** Günümüzün iç içe geçmiş dünyasında çocukların farklı bakış açılarına saygıyla yaklaşması gerekir. Nefret söylemini ele almak üzere eğitilmiş yetişkinler, çocukların farklı kültürlerle anlayışlı bir şekilde ilişki kurmaya hazırlanmasına yardımcı olabilir.
- **Sorumlu Dijital Vatandaşlığın Teşviki:** Nefret söylemi internette yayıldıkça çocukların dijital güvenliği daha da tehdit altına giriyor. Sorumlu dijital vatandaşlığı teşvik etmeyi öğrenen yetişkinler, nefret söylemine daha az göz yumulan, daha güvenli dijital ortamlar yaratmaya yardımcı olur.

Çocuklar ve gençler arasında nefret söyleminin önlenmesi, bunun temelde yatan nedenlerini ele alan, eleştirel düşünceyi ve empatiyi besleyen ve aktif vatandaşlığı güçlendiren çok yönlü bir yaklaşım gerektirir. Eğitime, farkındalık yaratmaya ve kapsayıcı politikalara yatırım yaparak, tüm çocuk ve gençlerin ayrımcılık ve nefretten uzak ortamlarda büyüyeceği bir gelecek inşa edebiliriz.

Aşağıdaki bölümde, özellikle eğitim alanında kullanılabilecek yöntem ve benzetmelere ilişkin bir dizi öneri sunuluyor.

Gençleri nefret söylemini anlamaya yönlendirecek etnografik benzetmeler

Çöp Kutusu Benzetmesi

Etnografik bir örnekten yola çıkarak, bir grup küçük çocuğun çöp kutusunu kullanmak yerine oraya buraya çöp attığı bir çizgi film hayal edelim. Attıkları çöplerin her biri tek başına önemsiz görünebilir, fakat biriktikçe daha büyük bir sorun yaratarak çevrelerini kirletir. Aynı şekilde, arada sırada nefret söylemine başvuran kişiler de –belki bir hakaretin veya klişe sözün zararı olmayacağını düşünerek– daha geniş bir zarar atmosferine katkıda bulunurlar. Zamanla bu sözler birikerek, genellikle ilk başta fark edilmeyen şekillerde çevredekileri olumsuz etkiler.

Balon Benzetmesi

Etnografik örneklerden esinlenerek, gençlerin balon şişirdiği ve bir balon patladığında etrafındakileri ürkütüp rahatsız ettiğinin farkına varmadan sürecin tadını çıkardığı bir çizgi film hayal edelim. Balon şişirmek ilk başta zararsız, hatta eğlenceli bir şey gibi görünüyor, fakat balonlar patladıkça ortaya çıkan gürültü, rahatsızlık ve karmaşa yaratıyor. Benzer biçimde, sosyal medyada da insanlar genellikle kendi “baloncukları” içinde tartışmaya girip de dışarıdan bir bakış açısı olmayınca, paylaşılan düşünceler daha da pekişiyor. Bu yankı odaları tesadüfen oluşmaz; etkileşime öncelik veren, öfke ve düşmanlık dahil olmak üzere güçlü duygusal reaksiyonları tetikleyen içerikleri pompalayıp duran platform algoritmaları tarafından bilinçli şekilde körüklenirler. Bunun sonucunda bireyler devamlı aynı bakış açılarına maruz kalıp, ideoloji “baloncuklarını” hiçbir engelle karşılaşmadan şişirip dururlar. Ayrıca, trol ve botlar da bölücü içerikleri yapay bir şekilde körükleyerek bu dinamikleri yoğunlaştırır ve kendi içine kapalı alanlar daha da katılaşır. Bu söylemler kontrolsüz bir şekilde dolaşıma girdikçe, patlamanın eşiğine gelmiş aşırı şişik balonlar gibi aşırılaşırlar. Baloncuklar daha geniş bir kamusal alana dağıldığında

ise zararlı söylemler de yayılır, ilk baştaki grubun dışındaki birey ve toplulukları etkileyerek genellikle gerçek dünya sonuçlarına yol açar.

Dalga Etkisi Benzetmesi

Durgun bir göle atılmış bir taş hayal edin. Suya çarpmanın etkisiyle dışarı doğru genişleyerek yayılan dalgalar oluşturur ve bu dalgalar yoluna çıkan her şeye dokunur. Nefret içeren tek bir söz veya ayrımcı bir klişe, tıpkı göle atılmış bir taş gibi, ilk atıldığı anın çok ötesine yayılan dalgalar yaratır. Nefret söylemi *münferit değildir* – bireyleri, toplulukları ve nihayetinde toplumun yapısını etkileyip şekillendirir. Ayrımcı anlatılar tarih kitapları ve eğitim müfredatı yoluyla pekiştirilince, dışlama ve önyargıları meşru hale getirdiğinden, ortadan kaldırılması daha da zorlaşır.

Nefret anlatılarıyla mücadelede kullanılacak eğitim araçları

Resimli Kitap ve Masallar

Çocuk kitapları ve masalları, nefret söylemine karşı çıkmak, zararlı tarihsel anlatıları parçalayıp atmak ve önyargıların pekişmesini eğitim yoluyla önlemek bakımından güçlü araçlardır. Masallar, farklı bakış açıları getirerek, ayrımcı klişeleri konu edinerek ve eleştirel düşünceyi teşvik ederek, genç okurların zararlı dile karşı empati ve direnç geliştirmesine yardım edebilir. Eğitimin içerisine nefret söylemi ve dışlayıcı anlatılar sokulduğunda, çocukların tarih, kimlik ve çevrelerindeki ötekilere ilişkin anlayışları bu yolla şekillenmiş olur. Kapsayıcı ve eleştirel hikâye anlatımı bilinçli bir şekilde düzenlenip sergilendiğinde, eğitim, ayrımcılığı pekiştirmek yerine kendi başına güçlenmenin bir aracı haline gelebilir.

Oyunlaştırma: Nefret Söylemini Saptayıp Ona Karşı Çıkmayı Öğrenmek

Çocuklar ve gençler arasında nefret söylemi ve ayrımcılık konusunda farkındalık yaratmak ve aynı zamanda farklı kimliklere yönelik önyargıları azaltmak için kullanılabilecek değerli bir eğitim aracı da oyunlardır. Oyunlaştırma; zorluklar, ödüller ve rol oynama senaryoları gibi unsurlar yardımıyla, katılımcıların gerçek dünya meselelerini güvenli ve sürükleyici bir ortamda ele almalarını sağlar. Oyunlar, gençlerin yanlış bilgi, taraflı söylem ve tarihsel önyargıların kamuoyunu nasıl şekillendirdiğini fark etmesine yardımcı olacak şekilde tasarlanabilir. Oyunların ayrımcı anlatılar hakkında farkındalık yaratmak için nasıl kullanılabileceğine dair kavramsal bir örnek ise *Mantığı Savunmak: Nefret Söylemine Karşı Mücadele!* isimli hayali bir oyun olabilir. Bu oyunda, oyuncular sosyal medya platformları,

forumlar ve yorum kısımları gibi çeşitli sanal ortamlarda mantıksal argümanlar gibi gösterilmeye çalışılan nefret söylemleriyle karşılaşılıyorlar. Bunlar arasındaki safsataları tespit ederek yanlış bilginin nasıl yayıldığını ve tarihsel önyargıların söylem yoluyla nasıl pekiştirildiğini öğreniyorlar.

Hackathonlar – Dijital Aktivizm Atölyeleri

Hackathonlar ve dijital atölyeler, gençlerin internet ortamında nefret söylemi ve yanlış bilgiyle mücadele etmenin yenilikçi yollarını keşfetmesine imkân verir. Katılımcılar, ayrımcı söylemlere karşı stratejiler geliştirerek ve zararlı dili tespit edip çürüten dijital araçlar tasarlayarak dijital ortamların daha kapsayıcı bir hal almasında aktif rol oynayabilirler.

Yaratıcı direniş: Nefret içerikli anlatılara karşı gençleri güçlendirme aracı olarak sanat

Nefret söylemi ve ayrımcı anlatılar çocuk ve gençleri derinden etkileyip, kendilerini, başkalarına, tarihsel ve toplumsal söylemlere bakışlarını şekillendirir. Sanat, bu olumsuz etkilere karşı bir kalkan görevi görebilir, bir ifade ve dayanıklılık aracı sağlayabilir ve çevremizi şekillendiren anlatıları yeniden tanımlayabilir. Yaratıcı yöntemlerle ayrımcılığa karşı çıkmaya yönlendirilen gençler böylelikle kendi bakış açılarını ifade edebilir, daha kapsayıcı anlatılara katkıda bulunabilir ve dışlanmaya direnen bir kültürü besleyebilirler.

Resim: Anlatıların Yeniden Tasavvuru

Gençler resim yoluyla nefret söyleminin yarattığı duygusal etkileri ifade edebilir, ayrımcılığın yol açtığı yıkımı görselleştirebilir ve eşitlik ve çeşitlilik gibi değerlere dayalı bir gelecek resmedebilir. Sanat atölyeleri, gençlerin egemen anlatılara karşı çıkarak farklı tarih ve bakış açılarını yansıtan yeni anlatılar tasavvur etmesine imkân sağlayabilir.

Tiyatro: Anlatılara Meydan Okumak

Tiyatro, gençlerin farklı bakış açılarını keşfetmesi, kendisini özgürce ifade etmesi ve ayrımcılık ve nefret söylemi de dahil olmak üzere sosyal meselelere eleştirel bir gözle bakması için yaratıcı bir alan sağlar. Kendilerini farklı rol ve anlatılara teslim eden gençlerde, söz ve eylemlerin kişi ve toplulukları nasıl etkilediğine dair daha derin bir anlayış uyanır. Tiyatroya dahil olmak iletişim becerilerini geliştirir, özgüven inşa eder ve ekip çalışmasını teşvik ederek gençleri düşüncelerini ifade

etme ve karmaşık sosyal dinamiklerin üstesinden gelme konusunda yüreklendirip güçlendirir. Tiyatro, ayrımcılığa doğrudan meydan okumanın yaratacağı yükü sırtlanmaktansa, yanlışlıkları fark edip, zararlı anlatıları sorgulamaya ve doğal hissettikleri yollarla kapsayıcılığı savunmaya yönelik araçları geliştirmelerine yardımcı olur. Hikâye anlatıcılığı ve performans yoluyla kendilerini ifade etmeyi deneyebilir, empati duygularını besleyebilir ve daha fazla anlayış içeren bir söylemi şekillendirmekte diyalogun ve farklı seslerin kıymetini anlayabilirler.

Dijital Hikâye Anlatıcılığı: Anlatıları Seyre Çıkıp Yeniden Şekillendirmek

Gençlerin dünyayla ilişki kurma aracı olarak –haber almak, sosyal etkileşimde bulunmak yahut eğlenmek– dijital platformları kullanma oranı her geçen gün artıyor; bu nedenle, karşısına çıkan anlatıları eleştirel bir gözle incelerken kendi hikayelerini etkili bir tarzda anlatabilecekleri araçlara sahip olmaları gerekiyor. Dijital hikâye anlatıcılığı, günümüzün bilgi ortamında hayati önem taşıyan medya okuryazarlığı becerilerini geliştirirken, farklı bakış açılarını ifade etme, ayrımcılığa meydan okuma ve zararlı anlatılara karşı çıkma imkânı sunuyor. Dijital hikâye anlatıcılığı sayesinde gençlerin nefret söyleminin etkilerine ilişkin tartışmalara katılması, yanlış anlatıların nasıl kurulduğunu analiz etmesi ve daha kapsayıcı perspektifleri deneyimlemesi mümkün olabilir. Farklı hikâye anlatıcılığı tekniklerini keşfederek, dijital ortamlarda karşısına çıkan mesajları eleştirel bir gözle incelerken kendi düşüncelerini ifade etmelerine yardımcı araçlar edinebilirler. Dijital hikâye anlatıcılığı aynı zamanda, benzer teknikleri manipüle ederek yanlış veya zararlı anlatıları yaygınlaştıran dezenformasyon ve yanlış bilgilendirme kampanyalarını fark etme becerisi de sağlar. Gençler, medyanın kamuoyunu nasıl şekillendirdiğini inceleyerek, farkındalığı daha yüksek tüketiciler ve içerik üreticileri haline gelebilirler. Dijital hikâye anlatıcılığı atölyeleri hem yaratıcı ifadeye, hem de eleştirel katılıma alan açtığı için gençlerin sadece dijital dünyanın katılımcısı değil, onu tanımlayan anlatıların aktif şekillendiricileri haline gelmesini sağlar.

Çocukların ve gençlerin eleştirel dijital okuryazarlık becerilerini geliştirmek amacıyla Türkiye’de geliştirilmiş olan örnek niteliğinde bir el kitabı, öğrencileri dijital içeriği eleştirel bir gözle değerlendirme, yanlış bilgileri fark etme ve sorumlu sanal katılım geliştirme becerileriyle donatıp güçlendirmek için dijital okuryazarlığın eğitim müfredatına nasıl dahil edilebileceğine dair mükemmel bir emsal oluşturuyor.

Eğitimde Eleştirel Dijital Okur-Yazarlık: Öğretmenden Öğretmene⁶³ başlıklı el kitabı **Öğretmen Ağı**⁶⁴ ve **Teyit.org**⁶⁵ işbirliğiyle ve Türkiye'nin 19 farklı ilinden çeşitli dallarda hizmet veren 39 öğretmenin katkılarıyla hazırlandı. Bu el kitabı, eleştirel dijital okuryazarlık becerilerinin eğitim ortamları üzerindeki etkisini ele alırken doğrulama ile bağlantı kuruyor. Eğitimcilerin alana yönelik farkındalık ve ilgisini geliştirecek bölümlerin yanı sıra öğrenme ortamlarında uygulanabilecek etkinlik önerileri de içeriyor.

Hrant Dink Vakfı'nın Dil, Diyalog, Demokrasi Laboratuvarı ASULİS, dilin toplumsal kapsayıcılık ve eşitlik üzerindeki etkisini ele almak üzere bir **Kapsayıcı Dil Atölyesi**⁶⁶ geliştirmiş bulunuyor. Bu atölye, Vakfın 2009 yılından bu yana sürdürdüğü **Medyada Nefret Söyleminin İzlenmesi**⁶⁷ projesini temel alıyor. Atölye, katılımcıları gündelik iletişimde ayrımcı dil kalıplarını fark edip bunlara karşı çıkmak için gereken bilgi ve araçlarla donatmayı amaçlıyor. Teorik bilgileri pratik alıştırmalarla birleştiren bu atölye, dilin dışlayıcılığı nasıl sürdürebileceğini ya da kapsayıcılığı nasıl teşvik edebileceğini daha derinlemesine anlamayı sağlıyor. Atölyenin içeriği, dilbilimsel değişiklikleri takip etmenin yanı sıra söylem çalışmaları ve konuyla ilgili raporlardan elde edilen bilgilerle düzenli olarak güncelleniyor. Katılımcıları daha kapsayıcı iletişim pratiklerini benimsemeye yönlendirerek, hem kişisel hayat hem de iş bağlamında daha adil ve eşitlikçi ortamlar yaratılmasına katkıda bulunan bu atölye çalışması, Vakfın uzun zamandır sürdürdüğü diyalog ve anlayışı teşvik hedefini de güçlendiriyor.

63 **Öğretmen Ağı.** (Ekim 2023). *Eğitimde Eleştirel Dijital Okuryazarlık: Öğretmenden Öğretmene El Kitabı*. Öğretmen Ağı.

https://www.ogretmenagi.org/sites/www.ogretmenagi.org/files/publications/egitimde_elestirel_dijital_okuryazarlik.pdf

64 **Öğretmen Ağı.** (2024). *Öğretmen Ağı – Öğretmenler için işbirliğine dayalı bir öğrenme platformu*. <https://www.ogretmenagi.org>

65 **Teyit.** (2024). *Teyit – Şüpheli bilgileri inceleyen doğrulama platformu*. <https://teyit.org/>

66 **Hrant Dink Vakfı.** (2022). *Kapsayıcı Dil Atölyesi*. <https://hrantdink.org/tr/asulis/duyurular/3704-kapsayici-dil-atolyesi>

67 **Hrant Dink Vakfı.** (2016). *Medyada Nefret Söyleminin İzlenmesi*. <https://hrantdink.org/tr/asulis/faaliyetler/projeler/medyada-nefret-soylemi>

Sonuç yerine

Bu rapor, uzmanlar arasında nefret söylemini ve ayrımcı söylemi anlayıp ele almaya yönelik olarak bir yıl boyunca sürdürülen bir işbirliği sonucunda ortaya çıkan kolektif bilgi ve içgörülerin yanı sıra disiplinlerarası tartışmaları kapsıyor. Nefret söyleminin tanımlanması ve kategorizasyonu, tespitte karşılaşılan güçlükler ve bu söylemle mücadelede alınacak politika tedbirlerinin ötesine geçen stratejileri inceleyen rapor, bu meseleyle mücadelenin içerdiği karmaşık unsurlara dair kapsamlı bir perspektif sunuyor.

Rapora katkıda bulunan yazarlar, ince farklar içeren tanımlara, etik yaklaşımlara ve hukuki ve düzenleyici çerçevelerin ötesinde çok katmanlı müdahalelere duyulan ihtiyacın altını çiziyorlar. Yürütülen tartışmalar kapsayıcı ve yapıcı bir söylemin geliştirilmesi bakımından disiplinlerarası işbirliğinin, sivil toplumun katılımının ve eğitim girişimlerinin önemini vurguluyor.

Bu yayının araştırmacılar, aktivistler, politika yapıcılar ve sosyal uyum, eşitlik ve diyalogu geliştirmeye çalışan herkes için değerli bir kaynak olmasını umuyoruz. Burada dile getirilen görüş ve tavsiyeler sadece bu ağın yürüttüğü çalışmaları yansıtmakla kalmayıp, daha kapsayıcı bir kamusal söyleme yönelik sürdürülen çabalara da katkı sunuyor.

Alex Mahadevan

Alex Mahadevan, Poynter'daki MediaWise programının yapay zekâ girişimlerini ve yanlış bilgi araştırmalarını yöneten direktörüdür. Binlerce öğrenciye, yetişkinlere ve gazeteciye çevrimiçi doğrulama ve medya okuryazarlığı eğitimi vermiştir. Poynter'ın yapay zekâ etiği rehberini yazmış, YZ, Etik ve Gazetecilik Zirvesi'ni yönetmiş ve Stanford'ın *Empowering Diverse Digital Citizens Lab* ekibine liderlik etmiştir. Alex, üretken yapay zekâ, açık kaynak istihbaratı (OSINT) ve yeni teknolojiler üzerine dünya çapında atölyeler düzenlemiştir.

Ayşecan Terzioğlu

Ayşecan Terzioğlu, Sabancı Üniversitesi Kültürel Çalışmalar ve Toplumsal Cinsiyet Çalışmaları Programları'nda öğretim üyesidir. Boğaziçi Üniversitesi'nde lisans ve yüksek lisans eğitimini tamamladıktan sonra City University of New York'ta Antropoloji alanında doktora yapmıştır. Araştırmaları, Orta Doğu antropolojisi, biyopolitika, sağlık eşitsizlikleri, toplumsal cinsiyet, zamansallık ve mekânsallık üzerine sosyal çalışmalar gibi alanlara odaklanmaktadır. Çok sayıda yayını bulunan Terzioğlu, RHWG, AMEA ve JOTSA yönetim kurullarında görev yapmaktadır.

Arzucan Özgür

Arzucan Özgür, Boğaziçi Üniversitesi Bilgisayar Mühendisliği bölümünde öğretim üyesi ve TABİ Lab'ın eş-direktörüdür. Araştırmaları, biyoinformatik ve doğal dil işleme üzerine yoğunlaşmakta olup, insan dilleri ve biyolojik diller için algoritmalar geliştirmektedir. Michigan Üniversitesi'nden doktora derecesine, Boğaziçi Üniversitesi'nden yüksek lisans ve lisans derecelerine sahiptir. Daha önce İstanbul Teknik Üniversitesi'nde görev yapmış olup, AILAB üyesidir.

Berrin Yanıkoğlu

Berrin Yanıkoğlu, Sabancı Üniversitesi Bilgisayar Bilimi ve Mühendisliği Bölümü'nde Profesör ve Veri Analitiği Mükemmeliyet Merkezi (VERİM) kurucu direktörüdür. Boğaziçi Üniversitesi'nden çift anadal derecesi ve Dartmouth College'dan doktora derecesi almıştır. Sabancı Üniversitesi'ne 2000 yılında katılmadan önce Rockefeller Üniversitesi, Xerox Imaging Systems ve IBM Almaden'de çalışmıştır. Araştırma alanları arasında bilgisayarla görü ve makine öğrenmesi yer almakta olup nefret söylemi tespiti, görüntü anlama, el yazısı tanıma ve metin özetleme gibi konulara odaklanmaktadır.

Claudia von Vacano

Dr. Claudia von Vacano, nefret söylemi üzerine algoritmik adalet, şeffaflık ve açıklanabilirlik konularında uzmandır. UC Berkeley'deki D-Lab ve ilgili programların direktörü olarak, her yıl 6.000'den fazla akademisyene veri odaklı araştırma ve eğitim desteği sağlamaktadır. Eğitim ve halk sağlığı eşitliği üzerine çok kampüslü bir NSF hibesinin eş lideridir ve UC Berkeley'de Dijital Beşeri Bilimler Yaz Programı'nı geliştirmiştir. Ayrıca, Anti-Defamation League ile Online Hate Index projesini yönetmektedir. UC Berkeley'den doktora, Stanford Üniversitesi'nden yüksek lisans derecesine sahiptir.

Didar Akar

Didar Akar, Boğaziçi Üniversitesi Dilbilim Bölümü'nde Doçent olarak görev yapmaktadır. Michigan Üniversitesi'nden Dilbilim alanında doktora derecesine sahiptir. Öğretim ve araştırma alanları söylem analizi, konuşma analizi, sosyodilbilim ve edimibilimdir. Çalışmaları, dil ve toplumsal cinsiyet, dil ve kimlik konularına odaklanmakta olup son dönemde sosyal medya ve nefret söylemi üzerine yoğunlaşmaktadır.

Eser Selen

Eser Selen, Özyeğin Üniversitesi'nde İletişim Tasarımı alanında Doçent olarak görev yapmaktadır. Feminist teoriler, performans çalışmaları, kuir teori ve çağdaş sanat alanlarında uzmanlaşmıştır. Araştırmaları, *Gender, Place & Culture* ve *Women & Performance* gibi dergilerde yayımlanmıştır. *Contesting Gender and Sexuality through Performance* adlı ilk monografisini tamamlamakta olup, bu eser 2026 yılında University of Edinburgh Press tarafından yayımlanacaktır. Multimedya sanatçısı olarak performans, enstalasyon ve video çalışmaları uluslararası sergilerde yer almıştır.

Houda Bouamor

Dr. Houda Bouamor, Carnegie Mellon Üniversitesi Katar Kampüsü'nde Bilgi Sistemleri Bölümü'nde öğretim üyesi ve doğal dil işleme ile makine öğrenmesi alanlarında araştırmacıdır. Paris-Sud Üniversitesi'nden Bilgisayar Bilimleri alanında doktora derecesine sahiptir. Dr. Bouamor, doğal dil işleme ve hesaplamalı dilbilim alanlarında uzmandır. 20'den fazla konferansın program komitesinde görev almış olup, temel araştırma alanları istatistiksel makine çevirisi üzerinedir.

İlayda Ece Ova

İlayda Ece Ova, toplumsal cinsiyet eşitliği, sivil katılım ve dijital aktivizm alanlarında uzmanlaşmış bir araştırmacı ve program analistidir.

Metin V. Bayrak

Metin Bayrak, Hacettepe Üniversitesi'nde Felsefe eğitimi almıştır ve 2008'den beri uygulamalı felsefe alanına odaklanmaktadır. 2014 yılında kurduğu Opus Noesis aracılığıyla Türkiye'de uygulamalı felsefenin yaygınlaşmasını sağlamış ve Galatasaray Üniversitesi ile İstanbul Okan Üniversitesi'nde dersler vermiştir. Uygulamalı felsefe ve P4C alanında sertifika programları düzenlemektedir. 2021'de Opus Kitap'ı kurarak bu alandaki eserleri yayımlamaya başlamıştır.

Onur Varol

Onur Varol, Sabancı Üniversitesi Bilgisayar Bilimi ve Mühendisliği Bölümü'nde Doçent ve Veri Analitiği Mükemmeliyet Merkezi (VERİM) üyesidir. VIRAL Lab'ın başında olup, hesaplamalı sosyal bilimler, ağ analizi ve makine öğrenmesi üzerine çalışmalar yürütmektedir. Araştırmaları algoritmik adalet, sosyal bot tespiti ve çevrimiçi platformlardaki kullanıcı etkileşimlerinin modellenmesi gibi konulara odaklanmaktadır.

Peter Adams

Peter Adams, News Literacy Project'in Araştırma ve Tasarım Kıdemli Başkan Yardımcısıdır ve 2009'dan beri bu organizasyonda çalışmaktadır. Eğitim kariyerine New York'taki devlet okullarında öğretmen olarak başlamış, ayrıca New York City Teaching Fellows Programı'nda eğitmen, Chicago'daki devlet okullarında gençlik medya eğitmeni ve Roosevelt Üniversitesi ile Chicago City Colleges'da ders veren öğretim görevlisi olarak görev yapmıştır.

Roser Morante

Dr. Roser Morante, UNED Üniversitesi'nde araştırmacı olup, Doğal Dil İşleme ve Bilgi Erişimi grubuna bağlıdır. Tilburg Üniversitesi'nden Hesaplamalı Dilbilim alanında doktora derecesine sahiptir. 80'den fazla akademik makale yayımlamış, semantik ve söylem analizi üzerine çalışmıştır. Sosyal ağlarda cin-

siyetçilik tespiti etkinliklerinin organizasyonunda yer almış ve *Computational Linguistics Special Issue on Modality and Negation* gibi akademik dergilerin editörlüğünü yapmıştır.

Susan Benesch

Susan Benesch, şiddeti teşvik edebilecek söylemleri inceleyen ve zararları önlemeye yönelik çalışmalar yürüten *Dangerous Speech Project*'i kurmuştur ve yönetmektedir. Yale Üniversitesi'nden insan hakları hukuku eğitimi almış olup, Amerikan Üniversitesi'nde ders vermektedir ve Harvard'ın Berkman Klein Merkezi'nde öğretim üyesi olarak görev yapmaktadır.

Tirşe Erbaysal Filibeli

Doç. Dr. Tirşe Erbaysal Filibeli, Galatasaray Üniversitesi'nde Medya ve İletişim Çalışmaları alanında yüksek lisans ve doktora derecelerini tamamlamıştır. 2016 yılından bu yana Medya Çoğulculuğu İzleme Projesi'nin (Media Pluralism Monitoring Project) Türkiye ekibinde araştırmacı ve koordinatör olarak görev yapmaktadır. 2020-2023 yılları arasında Ermiscom Projesi'nde çalışmıştır. Sahte haberler, dijital medya okuryazarlığı ve çatışma dönemlerinde gazetecilik konularında ortak editörlük yaptığı kitapları bulunmaktadır. Araştırmaları nefret söylemi, popülizm ve dijital kapitalizm gibi alanlara odaklanmaktadır. 2018 yılından beri Bahçeşehir Üniversitesi İletişim Fakültesi Yeni Medya Bölümü'nün başkanlığını yürütmektedir.

Tommaso Caselli

Dr. Tommaso Caselli, Groningen Üniversitesi Dil ve Biliş Merkezi'nde (Center for Language and Cognition) Hesaplamalı Anlam Bilimi (Computational Semantics) alanında yardımcı doçenttir. Araştırmaları olay çıkarımı (event extraction), çerçeveleme (framing), nefret söylemi tespiti ve dezenformasyonla mücadele üzerine yoğunlaşmaktadır. *Computational Analysis of Storylines* (CUP, 2021) kitabının ortak editörlüğünü yapmıştır ve İngilizce ve İtalyanca için çeşitli doğal dil işleme semantik değerlendirme kampanyalarını düzenlemiştir. Akademik çalışmalarıyla birçok ödül kazanmış olan Dr. Caselli, Jantina Tammes Dijital Toplum Okulu'nda (Jantina Tammes School of Digital Society) Yapay Zekâ ve Dil teması koordinatörüdür.

Yasemin İnceoğlu

Yasemin İnceoğlu, İstanbul Üniversitesi İngiliz Dili ve Edebiyatı bölümünden mezun olmuş, Marmara Üniversitesi'nde Gazetecilik alanında yüksek lisans ve doktora eğitimini tamamlamıştır. 1999 yılında profesör unvanını almıştır. Galatasaray Üniversitesi'nde Gazetecilik Bölüm Başkanlığı ve İletişim Fakültesi Dekanlığı görevlerini üstlenmiş olup, 2016 yılında emekli olmuştur. UNESCO, ILAD ve Avrupa Konseyi gibi uluslararası kuruluşlarla iş birliği yapmış, ayrıca Avrupa Üniversitesi Enstitüsü (EUI) ve University College London (UCL) gibi kurumlarda dersler vermiştir. Medya, azınlıklar, nefret söylemi ve popülizm konularında birçok kitabın yazarıdır. Halen Londra Ekonomi Okulu (London School of Economics) Medya ve İletişim Bölümü'nde çalışmalarına devam etmektedir.

Zeynep Özarslan

Zeynep Özarslan, İletişim Bilimleri alanında doktora derecesine sahiptir ve Çukurova Üniversitesi'nde İletişim Çalışmaları bölümünde profesör olarak görev yapmaktadır. 1998 yılından bu yana film, iletişim ve yeni medya alanlarında dersler vermekte ve yüksek lisans ve doktora tezleri yönetmektedir. Araştırma ilgi alanları arasında yeni medya, film çalışmaları, iletişim sosyolojisi ve yaratıcı endüstriler yer almaktadır. Son dönem yayınları, dijital oyunlar, ünlü çalışmaları, dijital şiddet ve nefret söylemi üzerine yoğunlaşmaktadır. Akademik çalışmalarının yanı sıra, nefret söylemi, dijital şiddetle mücadele ve internet haklarının savunulması alanlarında sivil toplum kuruluşlarıyla iş birliği yapmaktadır.

Kaynakça

- Acker, J. (2012). Gendered organizations and intersectionality: Problems and possibilities. *Equality, Diversity and Inclusion: An International Journal*, 31(3), 214–224. <https://doi.org/10.1108/02610151211209072>
- Alessi, E. J., Kahn, S. ve Van Der Horn, R. (2017). A Qualitative Exploration of the Premigration Victimization Experiences of Sexual and Gender Minority Refugees and Asylees in the United States and Canada. *Journal of sex research*, 54(7), 936–948. <https://doi.org/10.1080/00224499.2016.1229738>
- Arın, I., Işık, Z., Kutal, S., Dehghan, S., Özgür, A. ve Yanıkoğlu, B. (Temmuz 2023). SIU2023-NST-hate speech detection contest. 2023 31. Sinyal İşleme ve İletişim Uygulamaları Konferansı (SIU) içinde (s. 1–4). IEEE.
- Aroyo, L. ve Welty, C. (2015). Truth is a lie: Crowd truth and the seven myths of human annotation. *YZ Magazine*, 36(1), 15–24. <https://doi.org/10.1609/aimag.v36i1.2564>
- Artstein, R. ve Poesio, M. (2008). Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4), 555–596. <https://doi.org/10.1162/coli.07-034-R2>
- Avrupa Komisyonu (2020). Nefret Söylemine yönelik Avrupa Birliği Strateji ve Eylem Planı. https://commission.europa.eu/strategy-and-policy/policies/justice-and-fundamental-rights_en
- Avrupa Konseyi. (1997). Avrupa Konseyi Bakanlar Komitesinin “Nefret Söylemi” konusunda Üye Devletlere yönelik R (97)20 sayılı Tavsiye Kararı. *Bakanlar Komitesinin Medya ve Bilgi Toplumu Alanındaki Tavsiye Karar ve Bildirileri* içinde (s. 106–108). Strazburg: Avrupa Konseyi.
- Avrupa Konseyi (Aralık 2016). Cinsiyetçi nefret söylemiyle mücadele semineri: Rapor (10–12 Şubat 2016, Avrupa Gençlik Merkezi, Strazburg). <https://rm.coe.int/16806cc316>
- Awal, M. R., Cao, R., Lee, R. K. W. ve Mitrović, S. (2020). On analyzing annotation consistency in online abusive behavior datasets. *arXiv preprint arXiv:2006.13507*. <https://doi.org/10.48550/arXiv.2006.13507>

Badali, J. J. (2019). Migrants in the closet: LGBT migrants, homonationalism, and the right to refuge in Serbia. *Journal of Gay & Lesbian Social Services*, 31(1), 89–119. <https://doi.org/10.1080/10538720.2019.1548330>

Balčytienė, A. ve Juraitė, K. (2022). *Monitoring media pluralism in the digital era: Application of the Media Pluralism Monitor in Lithuania in the year 2021*. Avrupa Üniversitesi Enstitüsü, Medyada Çoğulculuk ve Medya Özgürlüğü Merkezi (CMPF). <https://doi.org/10.2924/EJLS.2022.015>

Basile, V., Bosco, C., Fersini, E., Nozza, D., Patti, V., Rangel Pardo, F. M., Rosso, P. ve Sanguinetti, M. (2019). Task 5: Multilingual detection of hate speech against immigrants and women in Twitter. 13. *Uluslararası Semantik Değerlendirme Çalıştayı bildirileri* içinde (s. 54–63). <https://doi.org/10.18653/v1/S19-2007>

Birleşmiş Milletler. (2019). *Nefret söylemine yönelik Birleşmiş Milletler strateji ve eylem planı*. https://www.un.org/en/genocideprevention/documents/advising-and%20mobilizing/Action_plan_on_hate_speech.EN.pdf

Bleyer-Simon, K., Brogi, E., Carlini, R., Da Costa Leite Borges, D., Kermer, J. E., Nenadic, I., Palmer, M., Parcu, P. L., Reviglio Della Venaria, U., Trevisan, M., Verza, S. ve Žuffová, M. (2024). *Monitoring media pluralism in the digital era: Application of the media pluralism monitor in the European member states and in candidate countries in 2023*. EUI, RSC, Araştırma Projesi Raporu, Medyada Çoğulculuk ve Medya Özgürlüğü Merkezi (CMPF). <https://hdl.handle.net/1814/77028>

Bosco, C., Felice, D. O., Poletto, F., Sanguinetti, M ve Maurizio, T. (2018). Overview of the EVALITA 2018 hate speech detection task. *CEUR Çalıştayı bildirileri* içinde (Cilt 2263, s. 1–9).

Buolamwini, J. ve Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. Sorelle ve Wilson (der.), 2018 *Konferans bildirileri* içinde (s. 1–15).

Cabitza, F., Campagner, A. ve Basile, V. (Haziran 2023). Toward a perspectivist turn in ground truthing for predictive computing. *Yapay Zekâ konulu AAYZ Konferansı bildirileri* içinde (Cilt 37, sayı 6, s. 6860–6868).

Chun, J. J., Lipsitz, G. ve Shin, Y. (2013). Intersectionality as a social movement strategy: Asian immigrant women advocates. *Signs: Journal of Women in Culture and Society*, 38(4), 917–940. <http://www.jstor.org/stable/10.1086/669575>

Crenshaw, K. (1989). Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics. *University of Chicago Hukuk Forumu*, 139–167.

<https://chicagounbound.uchicago.edu/cgi/viewcontent.cgi?article=1052&context=uclf>

Crenshaw, K. (2005). Mapping the margins: Intersectionality, identity politics, and violence against women of color. R. K. Bergen, J. L. Edleson ve C. M. Renzetti (der.), *Violence against women: Classic papers* içinde (s. 282–313). Pearson Education New Zealand.

Davidson, T., Warmesley, D., Macy, M. ve Weber, I. (2017). Automated hate speech detection and the problem of offensive language. *Web ve Sosyal Medya konulu 11. Uluslararası AAYZ Konferansı bildirileri* içinde (s. 512–515). Yapay Zekânın Geliştirilmesi Derneği. <https://doi.org/10.1609/icwsm.v11i1.14955>

Del Vigna, F., Cimino, A., Dell’Orletta, F., Petrocchi, M. ve Tesconi, M. (2017). Hate me, hate me not: Hate speech detection on Facebook. *Siber Güvenlik konulu Birinci İtalyan Konferansı bildirileri (ITASEC17)* içinde, Venedik, İtalya, 86–95.

Djuric, N., Zhou, J., Morris, R., Grbovic, M., Radosavljevic, V. ve Bhamidipati, N. (2015). Hate speech detection with comment embeddings. 29–30. <https://doi.org/10.1145/2740908.2742760>

Eduardo Sánchez, M. ve Eduardo, M. (2013). Latino Lesbian, Gay, Bisexual, and Transgender Immigrants in the United States. *Journal of LGBT Issues in Counseling*. <https://doi.org/10.1080/15538605.2013.785467>

Edward, J. A., Kershaw, S., Grimm, B., Wehner, L. ve Murphy, D. (2020). A qualitative exploration of the integration experiences of LGBTQ refugees who fled from the Middle East, North Africa, and Central and South Asia to Austria and the Netherlands. *Sexuality Research and Social Policy*. <https://doi.org/10.1007/s13178-018-0364-7>

Erbaysal Filibeli, T. ve Ertuna, C. (2021). Sarcasm beyond hate speech: Facebook comments on Syrian refugees in Turkey. *International Journal of Communication*, 15, 2236–2259. <https://ijoc.org/index.php/ijoc/article/view/16582>

Faloppa, F., Gambacorta, A., Odekerken, R. ve van der Noordaa, R. (2023). *Study on preventing and combating hate speech in times of crisis*. Avrupa Konseyi. <https://rm.coe.int/-study-on-preventing-and-combating-hate-speech-in-times-of-crisis/1680ad393b>

Fassinger, R. E. ve Arseneau, J. R. (2007). “I’d rather get wet than be under that umbrella”: Differentiating the experiences and identities of lesbian, gay, bisexual, and transgender people. K. J. Bieschke, R. M. Perez ve K. A. DeBord (der.), *Handbook of*

counseling and psychotherapy with lesbian, gay, bisexual, and transgender clients içinde (2. baskı., s. 19–49). Amerikan Psikoloji Derneği. <https://doi.org/10.1037/11482-001>

Feinberg, J. (1983). Obscene words and the law. *Law and Philosophy*, 2(2), 139–161.

Fortuna, P. ve Nunes, S. (2018). A survey on automatic detection of hate speech in text. *ACM Bilişim Anketleri (CSUR)*, 51(4), 1–30. <https://doi.org/10.1145/3236009.3236010>

Fortuna, P., Rocha da Silva, J., Wanner, L. ve Nunes, S. (2019). A hierarchically-labeled Portuguese hate speech dataset. *Sanal Hakaret Dili konulu Üçüncü Çalıştay bildirileri içinde*, (s. 94–104). <https://doi.org/10.18653/v1/W19-3510>

Fortuna, P., Soler, J. ve Wanner, L. (2020, May). Toxic, hateful, offensive or abusive? What are we really classifying? An empirical analysis of hate speech datasets. *On İkinci Dil Kaynakları ve Değerlendirme Konferansı bildirileri içinde* (s. 6786–6794). Avrupa Dil Kaynakları Derneği. <https://aclanthology.org/2020.lrec-1.838.pdf>

Gambäck, B. ve Sikdar, U. K. (2017). Using convolutional neural networks to classify hate-speech. *Sanal Hakaret Dili konulu Birinci Çalıştay bildirileri (ALW1) içinde*, 85–90. <https://doi.org/10.18653/v1/W17-3013>

Gao, S. (S.), Brandt, S. A. ve Stults, C. B. (2023). Internalized transphobia and self-concept clarity among transgender and gender-nonconforming young adults: Characteristics, associations, and the mediating role of self-esteem. *Psychology of Sexual Orientation and Gender Diversity*. Basılı yayın öncesi çevrimiçi versiyon. <https://doi.org/10.1037/sgd0000691>

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumée, III, H. ve Crawford, K. (2018). Datasheets for datasets. *arXiv*, 1–17. <https://arxiv.org/abs/1803.09010>

Gelber, K. (2019). Differentiating hate speech: A systemic discrimination approach. *Critical Review of International Social and Political Philosophy*, 24(4), 393–414. <https://doi.org/10.1080/13698230.2019.1576006>

Gelber, K. (2021). Differentiating hate speech: A systemic discrimination approach. *Critical Review of International Social and Political Philosophy*, 24(4), 393–414. <https://doi.org/10.1080/13698230.2019.1576006>

Gibert, O., Perez, N., García-Pablos, A. ve Cuadros, M. (2018). *Hate speech dataset from a white supremacy forum*. D. Fišer, R. Huang, V. Prabhakaran, R. Voigt, Z. Waseem ve J. Wernimont (der.), *Sanal Hakaret Dili konulu İkinci Çalıştay bildirileri (ALW2) içinde* (s. 11–20). Bilişimsel Dilbilim Derneği. <https://aclanthology.org/W18-5102/>

Gitari, N. D., Zuping, Z., Damien, H. ve Long, J. (2015). A lexicon-based approach for hate speech detection. *International Journal of Multimedia and Ubiquitous Engineering*, 10(4), 215–230. <https://doi.org/10.14257/ijmue.2015.10.4.21>

GLAAD. (2024). *Promoting Positive Narratives and Monitoring Anti-LGBTQ+ Media Portrayals*. <https://www.glaad.org/>

Gowin, M., Taylor, E. L., Dunnington, J., Alshuwaiyer, G. ve Cheney, M. K. (2017). Needs of a silent minority: Mexican transgender asylum seekers. *Health Promotion Practice*, 18(3), 332–340. <https://doi.org/10.1177/1524839917692750>

Hagendorff, T. (2020). *Minds and Machines*, 30, 99–120. <https://doi.org/10.1007/s11023-020-09517-8>

Holznagel, B. ve Kalbhenn, J. C. (2023). *Monitoring media pluralism in the digital era: Application of the media pluralism monitor in the European Union, Albania, Montenegro, the Republic of North Macedonia, Serbia, and Turkey in the year 2022: Country report: Germany*. Avrupa Üniversitesi Enstitüsü. <https://hdl.handle.net/1814/75723>

Hopkinson, R. A., Keatley, E., Glaeser, E., Erickson-Schroth, L., Fattal, O. ve Nicholson Sullivan, M. (2017). Persecution experiences and mental health of LGBT asylum seekers. *Journal of Homosexuality*, 64(12), 1650–1666. <https://doi.org/10.1080/00918369.2016.1253392>

Hosseinmardi, H., Mattson, S. A., Rafiq, R. I., Han, R., Lv, Q. ve Mishra, S. (2015). Analyzing labeled cyberbullying incidents on the Instagram social network. *International Conference on Social Informatics*, 49–66. https://doi.org/10.1007/978-3-319-27433-1_4

İnanç Özgürlüğü Girişimi. (2021). *Rabat Eylem Planı*. <https://inancozgurlugugirisimi.org/wp-content/uploads/2021/02/Rabat-Eylem-Plani-Turkce.pdf>

İnceoğlu, Y., Erbaysal Filibeli, T., Ertuna, C. ve Çenberli, Y. (2022). *Monitoring media pluralism in the digital era: Application of the Media Pluralism Monitor in Turkey in the year 2022*. Avrupa Üniversitesi Enstitüsü, Medyada Çoğulculuk ve Medya Özgürlüğü Merkezi (CMPF). <https://hdl.handle.net/1814/75744>

Jahan, M. S. ve Oussalah, M. (2023). A systematic review of hate speech automatic detection using natural language processing. *Neurocomputing*, 546, 126232. <https://doi.org/10.1016/j.neucom.2023.126232>

Jobin, A., Ienca, M. ve Vayena, E. (2019). The global landscape of YZ ethics guidelines. *Nature Machine Intelligence*, 1, 389–399. <https://doi.org/10.1038/s42256-019-0088-2>

- Kaya, A. (2023). *The world's leading refugee host, Turkey, has a complex migration history*. Migration Policy Institute. <https://www.migrationpolicy.org/article/turkey-migration-history>
- Kılıç, O. (2023). Queer cybercultures and digital activism: Transformations in LGBTQ+ digital spaces. *Lambda Nordica*, 28(2-3), 902. <https://doi.org/10.34041/ln.v28.902>
- Kindermann, D. (2023). Against 'Hate Speech'. *Journal of Applied Philosophy*, 40(5). <https://doi.org/10.1111/japp.12648>
- Kocon, J., Figas, A., Gruza, M., Pulchaska, D., Kajdanowicz, T. ve Kazienko, P. (2021). Offensive, aggressive, and hate speech analysis: From data-centric to human-centered approach. *Information Processing and Management*, 58, 1-26. <https://doi.org/10.1016/j.ipm.2021.102643>
- Kwok, I. ve Wang, Y. (2013). Locate the hate: Detecting tweets against blacks. *Yapay Zekâ konulu AAYZ Konferansı bildirileri (AAAI-13)* içinde, 1621-1622. <https://doi.org/10.1609/aaai.v27i1.8539>
- Leader Maynard, S. ve Benesch, S. (2018). Dangerous speech and hate speech: Definitions, characteristics, and analysis. *International Journal of Communication*, 12, 2740-2759. <https://ijoc.org/index.php/ijoc/article/view/8200>
- Leonardelli, E., Basile, V. ve Bosco, C. (2021). Agreeing to disagree: Annotating offensive language datasets with annotators' disagreement. *Özellik, Duyarlık ve Sosyal Medya Analizine Bilişimsel Yaklaşımlar konulu 11. Çalıştay bildirileri* içinde, 167-176. <https://aclanthology.org/2021.wassa-1.17>
- Liebe, U. ve Beyer, H. (2021). Examining discrimination in everyday life: A stated choice experiment on racism in the sharing economy. *Journal of Ethnic and Migration Studies*, 47(9), 2065-2088. <https://doi.org/10.1080/1369183X.2019.1710118>
- Loebbecke, C., Luong, A. C. ve Obeng-Antwi, A. (2021). YZ for tackling hate speech. *ECIS 2021 Research-in-Progress Papers*, 10.
- Madukwe, K., Gao, X. ve Xue, B. (Kasım 2020). In data we trust: A critical analysis of hate speech detection datasets. *Sanal Hakaret ve Zararlar konulu Dördüncü Çalıştay bildirileri* içinde (pp. 150-161). Bilişimsel Dilbilim Derneği.
- Manzi, F. ve Heilman, M. E. (2021). Breaking the glass ceiling: For one and all? *Journal of Personality and Social Psychology*, 120(2), 257-277. <https://doi.org/10.1037/pspa0000260>

McClain, M. ve Waite-Wright, O. (2015). The LGBT community in Turkey: Discrimination, violence, and the struggle for equality. *Creighton International and Comparative Law Journal*, 7(1), 152–176.

McLean, L. L. (2021). Internalized homophobia and transphobia. E. M. Lund, C. Burgess ve A. J. Johnson (der.), *Violence against LGBTQ+ persons* içinde. Springer. https://doi.org/10.1007/978-3-030-52612-2_3

Mediaweek. (26 Ekim 2022). *TikTok and AAP partner to empower content creators to combat misinformation*. <https://www.mediaweek.com.au/tiktok-and-aap-partner-to-empower-creators-to-fight-misinformation/>

Mehta, H. ve Passi, K. (2022). Social media hate speech detection using explainable artificial intelligence (XAI). *Algorithms*, 15(8), 291. <https://doi.org/10.3390/a15080291>

Morales, E. (2013). Latino lesbian, gay, bisexual, and transgender immigrants in the United States. *Journal of LGBT Issues in Counseling*, 7(2), 172–184. <https://doi.org/10.1080/15538605.2013.785467>

Morley, J., Floridi, L., Kinsey, L. ve Elhalal, A. (2020). From what to how: An initial review of publicly available YZ ethics tools, methods, and research to translate principles into practices. *Science and Engineering Ethics*, 26(4), 2141–2168. <https://doi.org/10.1007/s11948-019-00165-5>

Morstatter, F., Pfeffer, J., Liu, H. ve Carley, K. (2013). Is the sample good enough? Comparing data from Twitter’s streaming API with Twitter’s Firehose. *Web ve Sosyal Medya konulu 7. Uluslararası AAYZ Konferansı bildirileri (ICWSM 2013)* içinde (s. 400–408). Yapay Zekânın Geliştirilmesi Derneği (AAAI). <https://doi.org/10.1609/icwsml.v7i1.14401>

Musto, C., Semeraro, G., Polignano, M. ve Stranisci, M. (2016). *Modeling community behavior through semantic analysis of social data: The Italian Hate Map experience*. 8. Uluslararası Bilgi Keşfi ve Bilgi Erişim Konferansı Bildirileri (KDIR 2016) içinde (s. 343–350). SCITEPRESS – Science and Technology Publications.

Nadal, K. (2008). Preventing racial, ethnic, gender, sexual minority, disability, and religious microaggressions: Recommendations for promoting positive mental health. *Prevention in Counseling Psychology: Theory, Research, Practice and Training*, 2(1), 22–27.

Nadal, K. L., Issa, M.-A., Leon, J., Meterko, V., Wideman, M. ve Wong, Y. (2011a). Sexual orientation microaggressions: “Death by a thousand cuts” for lesbian, gay,

and bisexual youth. *Journal of LGBT Youth*, 8(3), 234–259. <https://doi.org/10.1080/19361653.2011.584204>

Nadal, K. L., Wong, Y., Issa, M.-A., Meterko, V., Leon, J. ve Wideman, M. (2011b). Sexual orientation microaggressions: Processes and coping mechanisms for lesbian, gay, and bisexual individuals. *Journal of LGBT Issues in Counseling*, 5(1-2), 21–46. <https://doi.org/10.1080/15538605.2011.554606>

Nobata, C., Tetreault, J., Thomas, A., Mehdad, Y. ve Chang, Y. (2016). Abusive language detection in online user content. 25. *Uluslararası World Wide Web Konferansı (WWW 2016) bildirileri* içinde (s. 145–153). ACM

Nockleby, J. T. (2000). Hate speech. *Encyclopedia of the American Constitution*, 3, 1277–1279.

Novak, P. K., Scantamburlo, T., Pelican, A., Cinelli, M., Mozetič, I. ve Zollo, F. (2022). Handling disagreement in hate speech modelling. *Communications in Computer and Information Science*, 1602, 681–695. https://doi.org/10.1007/978-3-031-08974-9_54

O’Driscoll, J. (2020). *Offensive language: Taboo, offence and social control*. Bloomsbury Publishing.

Özdüzen, A. ve Korkut, U. (2020). ‘Refugees are not welcome’: Digital racism, online place-making and the evolving categorization of Syrians in Turkey. *Environment and Planning C: Politics and Space*, 39(5), 851–869. <https://journals.sagepub.com/doi/10.1177/1461444820956341>

Pasquetto, I. V., Priedhorsky, R., Lazer, D. ve Terveen, L. (2020). Ethical and transparency challenges in data collection for misinformation research. *Journal of Information Ethics*, 29(1), 58–77. <https://doi.org/10.1016/j.jinfo.2020.01.004>

Pfeffer, J., Zorbach, T. ve Carley, K. M. (2023). Challenges in collecting data on social media platforms: A case study of X/Twitter. 2023 *Sosyal Medya ve Toplum konulu Uluslararası Konferans bildirileri* içinde (s. 1–10).

Piwowarczyk, L., Fernandez, P. ve Sharma, A. (2017). Seeking asylum: Challenges faced by the LGB community. *Journal of Immigrant and Minority Health*, 19(4), 723–732. <https://doi.org/10.1007/s10903-016-0363-9>

Plank, B., Hovy, D. ve Søgaard, A. (2014). Learning part-of-speech taggers with inter-annotator agreement loss. 14. *Bilişimsel Dilbilim Derneği Avrupa Şubesi Konferansı bildirileri* içinde, 742–751. <https://aclanthology.org/E14-1078>

- Plaza, L. vd. (2023). Overview of EXIST 2023: sEXism identification in social networks. J. Kamps vd. (der.), *Advances in information retrieval: ECIR 2023* içinde (s. 1–15). Springer. https://doi.org/10.1007/978-3-031-28241-6_68
- Poletto, F., Basile, V., Sanguinetti, M., Bosco, C. ve Patti, V. (2021). Resources and benchmark corpora for hate speech detection: A systematic review. *Language Resources and Evaluation*, 55, 477–523. <https://doi.org/10.1007/s10579-020-09502-8>
- Poynter Enstitüsü. (10 Aralık 2020). *Poynter's MediaWise training significantly increases people's ability to detect disinformation, new Stanford study finds*. <https://www.poynter.org/news-release/2020/poynters-mediawise-training-significantly-increases-peoples-ability-to-detect-disinformation-new-stanford-study-finds/>
- Rebillard, F. ve Sklower, J. (2022). *Monitoring media pluralism in the digital era: Application of the Media Pluralism Monitor in France in the year 2021*. Avrupa Üniversitesi Enstitüsü, Medyada Çoğulculuk ve Medya Özgürlüğü Merkezi, <https://hdl.handle.net/1814/74750>
- Robles Carrillo, M. (2020). Artificial intelligence: From ethics to law. *Telecommunications Policy*. <https://doi.org/10.1016/j.telpol.2020.101937>
- Rodríguez-Sánchez, F., Carrillo-de-Albornoz, J., Plaza, L., Gonzalo, J., Rosso, P., Comet, M., ve Donoso, T. (2021). Overview of EXIST 2021: sEXism identification in social networks. *Procesamiento del Lenguaje Natural*, 67, 195–207.
- Rodríguez-Sánchez, F., Carrillo-de-Albornoz, J., Plaza, L., Mendieta-Aragón, A., Marco-Remón, G., Makeienko, M., Plaza, M., Gonzalo, J., Spina, D. ve Rosso, P. (2022). Overview of EXIST 2022: sEXism identification in social networks. *Procesamiento del Lenguaje Natural*, 69, 229–240.
- Royakkers, L., Timmer, J., Kool, L. ve van Est, R. (2018). Societal and ethical issues of digitization. *Ethics and Information Technology*, 20(2), 127–142. <https://doi.org/10.1007/s10676-018-9452-x>
- Röttger, P., Vidgen, B., Nguyen, D., Waseem, Z., Margetts, H. ve Pierrehumbert, J. (2021). HateCheck: Functional tests for hate speech detection models. *Bilişimsel Dilbilim Derneği* 59. Yıllık Toplantı bildirileri içinde (s. 41–58). Bilişimsel Dilbilim Derneği.
- Röttger, P., Vidgen, B., Hovy, D. ve Pierrehumbert, J. (2022). Two contrasting data annotation paradigms for subjective NLP tasks. 2022 *Bilişimsel Dilbilim Derneği Kuzey Amerika Şubesi Konferansı: İnsan Dili Teknolojileri bildirileri* içinde, Bilişimsel Dilbilim Derneği.

Ruitenbeek, W., Zwart, V., van der Noord, R., Gnezdilov, Z. ve Caselli, T. (2022). “Zo Grof!”: A Comprehensive Corpus for Offensive and Abusive Language in Dutch. *Sanal Hakaret ve Zararlar konulu Altıncı Çalıştay (WOAH) bildirileri* içinde (s. 40–56). Bilişimsel Dilbilim Derneği. <https://aclanthology.org/2022.woah-1.5/>

Russell, E. L. (2020). Hate in language, hate and language. K. Hall ve R. Barrett (der.), *The Oxford handbook of language and sexuality* içinde. Oxford University Press. <https://doi.org/10.1093/oxfordhb/9780190212926.013.67>

Sachdeva, P. S., Barreto, R., von Vacano, C. ve Kennedy, C. J. (Haziran 2022). Assessing annotator identity sensitivity via item response theory: A case study in a hate speech corpus. *Adillik, Hesap Verebilirlik ve Şeffaflık konulu 2022 ACM Konferans bildirileri* içinde (s. 1585–1603). Bilişimsel Dilbilim Derneği.

Salminen, J., Almerexhi, H., Milenkovič, M., Jung, S. G., An, J., Kwak, H. ve Jansen, B. J. (2018). Anatomy of online hate: Developing a taxonomy and machine learning models for identifying and classifying hate in online news media. *Web ve Sosyal Medya konulu 12. Uluslararası AAYZ Konferansı (ICWSM-18) bildirileri* içinde. AAYZ Press.

Sanguinetti, M., Poletto, F., Bosco, C., Sanguinetti, V., Stranisci, M. ve Russo, I. (2020). HaSpeeDe 2 @ EVALITA2020: Overview of the EVALITA 2020 hate speech detection task. *CEUR Çalıştay Bildirileri*, 2765, 1–15.

Schmidt, A. ve Wiegand, M. (2017). A survey on hate speech detection using natural language processing. *Sosyal Medyada Doğal Dil İşleme konulu 5. Uluslararası Çalıştay bildirileri* içinde (s. 1–10).

Selen, E. (2020). “The public immoralist”: Discourses of queer subjectification in contemporary Turkey. *International Journal of Communication*, 14, 5518–5536.

Sharma, S., Agrawal, S. ve Shrivastava, M. (2018). Degree-based classification of harmful speech using Twitter data. *arXiv:1806.04197*. <https://arxiv.org/abs/1806.04197>

Silva, L., Mondal, M., Correa, D., Benevenuto, F. ve Weber, I. (2016). Analyzing the targets of hate in online social media. *Web ve Sosyal Medya konulu Onuncu Uluslararası AAYZ Konferansı (ICWSM-2016) bildirileri* içinde.

Simonsen, S. (2023). *Monitoring media pluralism in the digital era: Application of the Media Pluralism Monitor in the European Union, Albania, Montenegro, the Republic of North Macedonia, Serbia, and Turkey in the year 2022. Country report: Denmark*. Medyada Çoğulculuk ve Medya Özgürlüğü Merkezi

- Suryawanshi, S., Chakravarthi, B. R., Arcan, M. ve Buitelaar, P. (2020). Multimo-
dal meme dataset (MultiOFF) for identifying offensive content in image and text.
Trolleme, Saldırganlık ve Siber Zorbalık konulu İkinci Çalıştay bildirileri içinde, 32–41.
<https://www.aclweb.org/anthology/2020.trac-1.6/>
- Swamy, S. D., Jamatia, A. ve Gambäck, B. (2020). Studying generalisability across
abusive language detection datasets. *Bilişimsel Doğal Dil İşleme konulu 23. Konferans
bildirileri (CoNLL)* içinde (s. 940–950). Bilişimsel Dilbilim Derneği.
- Ullmann, S. ve Tomalin, M. (2020). *Ethics and Information Technology*, 22, 69–80.
<https://doi.org/10.1007/s10676-019-09516-z>
- Uludoğan, G., Dehghan, S., Arın, I., Erol, E., Yanıkoğlu, B. ve Özgür, A. (Mart 2024).
Overview of the hate speech detection in Turkish and Arabic tweets (HSD-2Lang)
shared task at CASE 2024. *Sosyopolitik Olayların Metinden Otomatik Olarak Ayrıştırıl-
masına ilişkin Sorunlar ve Uygulamalar konulu 7. Çalıştay (CASE 2024) bildirileri* içinde
(s. 229–233). Springer. <https://aclanthology.org/2024.case-1.32.pdf>
- BMMYK. (2024). *Türkiye’deki Mülteciler ve Sığınmacılar*. [https://www.unhcr.org/tr/
kime-yardim-ediyoruz/tuerkiyedeki-muelteciler-ve-siginmacilar](https://www.unhcr.org/tr/kime-yardim-ediyoruz/tuerkiyedeki-muelteciler-ve-siginmacilar)
- Van Dijk, T. A. (2008). *Discourse and power*. Palgrave Macmillan.
- Vidgen, B. ve Derczynski, L. (2020). Directions in abusive language training data, a
systematic review: Garbage in, garbage out. *Plos one*, 15(12), e0243300.
- Warner, W. ve Hirschberg, J. (2012). Detecting hate speech on the World Wide Web.
Sosyal Medya Dili konulu İkinci Çalıştay bildirileri içinde (s. 19–26).
- Waseem, Z. ve Hovy, D. (2016). Hateful symbols or hateful people? Predictive fe-
atures for hate speech detection on Twitter. *NAACL Öğrenci Araştırması Çalıştayı
bildirileri (NAACL SRW 2018)* içinde, 88–93.
- Waseem, Z., Davidson, T., Warmesley, D. ve Weber, I. (2017). Understanding Abu-
se: A Typology of Abusive Language Detection Subtasks. *Sanal Hakaret Dili konulu
Birinci Çalıştay bildirileri* içinde, s. 78–84, Vancouver, BC, Kanada. Bilişimsel Dilbilim
Derneği.
- Wiegand, M., Siegel, M. ve Ruppenhofer, J. (2018). Overview of the germeval 2018
shared task on the identification of offensive language.
- Wijaya, H. Y. (2022). Digital homophobia. *Indonesia and the Malay World*, 50(146),
52–72. <https://doi.org/10.1080/13639811.2022.2010357>

Williamson, M. (2023). A global analysis of transgender rights: Introducing the Trans Rights Indicator Project (TRIP). *Perspectives on Politics*. Basılı yayın öncesi çevrimiçi versiyon. <https://doi.org/10.1017/S1537592723002827>

Xue, L., Zhang, H. ve Li, Y. (2016). The right to be forgotten: Users exploiting GDPR for hate speech removal. *International Journal of Information Management*, 36(3), 351–356.

Yuying, T., Nadine, N., Nadine, N., Mark, P. ve Mark, P. (2013). Borders and margins: Giving voice to lesbian, gay, bisexual, and transgender immigrant experiences. *Journal of LGBT Issues in Counseling*. <https://doi.org/10.1080/15538605.2013.785235>

Zampieri, M., Malmasi, S., Nakov, P., Rosenthal, S., Farra, N. ve Kumar, R. (Hazi-ran 2019). SemEval-2019 Task 6: Identifying and categorizing offensive language in social media (OffensEval). 13. *Uluslararası Semantik Değerlendirme Çalıştayı bildirileri* içinde, 75–86. Bilişimsel Dilbilim Derneği.

Zampieri, M., Nakov, P., Rosenthal, S., Atanasova, P., Karadzhov, G., Mubarak, H., Derczynski, L., Pitenis, Z. ve Çöltekin, Ç. (2020). SemEval-2020 task 12: Multilingual offensive language identification in social media (OffensEval 2020). *arXiv preprint arXiv:2006.07235*.

Zhong, H., Li, H., Squicciarini, A., Rajtmajer, S., Griffin, C., Miller, D. ve Caragea, C. (2015). Content-driven detection of cyberbullying on the Instagram social network. *International Joint Conference on Artificial Intelligence*, 3952–3958.

Zhuo, T. Y., Huang, Y., Chen, C. ve Xing, Z. (2023). Red teaming ChatGPT via jail-breaking: Bias, robustness, reliability and toxicity. *arXiv*. <https://doi.org/10.48550/arXiv.2301.12867>

© HRANT DİNK VAKFI YAYINLARI, 2025

a5.u.lis DİL
DİYALOG
DEMOKRASİ
LABORATUVARI

Görüş ve önerileriniz için:

