NEFRET SÖYLEMİYLE MÜCADELEDE YAPAY ZEKÂ

Etiketleme, sınıflandırma ve tespit kılavuzu

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 yaşanası davamız', diyalog, barış, empati kültürünün hâkim olduğu bir gelecektir.

NEFRET SÖYLEMİYLE MÜCADELEDE YAPAY ZEKÂ: .
ETİKETLEME, SINIFLANDIRMA VE TESPİT KILAVUZU

ISBN 978-605-71835-8-3

a5·u·lis DİL DİYALOG DEMOKRASİ LABORATUVARI

editörler
Tirşe Erbaysal Filibeli, Tunga Güngör

proje ekibi
İnanç Arın, Didar Akar, Başak Can, Somaiyeh Dehghan, Elif Erol, Burak Işık, Sıla Kartal, Buket Kapısız, Yasemin Korkmaz, Arzucan Özgür, Nural Özel, Pelin Önal, Gökçe Uludoğan, Ayşecan Terzioğlu, Murat Tercan, İrem Topçu, Tuğba Özsoy, Berrin Yanıkoğlu, Elif Yararbaş, Umut Şen

yayına hazırlayan
Başak Can, Elif Erol, Buket Kapısız, Yasemin Korkmaz, Pelin Önal, Tuğba Özsoy, Elif Yararbaş

çevirmenler
Burcu Becermen, Simon Charles Popay

tasarım ve veri görselleştirme
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

Istanbul, Mart 2025

HDV YAYINLARI

Anarad Hığutyun Binası Papa Roncalli Sokak No: 128
Harbiye, 34373 Şişli, İstanbul
T: 0212 240 33 61
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 ՀՐԱՆԴ ՏՆՔ ՀԻՄՆԱԴՐԱՄ

BOĞAZİÇİ ÜNİVERSİTESİ 1863

Sabancı Üniversitesi

NEFRET SÖYLEMİYLE MÜCADELEDE YAPAY ZEKÂ

Etiketleme, sınıflandırma ve tespit kılavuzu

HDV YAYINLARI

İÇİNDEKİLER

Giriş

Hem geleneksel hem de sosyal medyada artarak yaygınlaşan nefret söylemi özellikle kriz dönemlerinde toplumsal huzur ve barış için endişe verici boyutlara ulaşmaktadır. Geleneksel medyada mesleki kodlar, yasal düzenlemeler ve kurum içi kurallar ile sınırlı da olsa kontrol edilmeye, denetlenmeye ve kısıtlanmaya çalışılan nefret söylemi, sosyal medya platformlarında bilginin hızlı akışı ve etkileşim sayılarının yüksek olması gibi nedenlerle çok hızlı bir şekilde dolaşıma girmekte ve yaygınlaşmaktadır. Sosyal medya platformları, bazı iç düzenlemeler yapmalarına karşın, kullanıcıyı sistemde daha uzun tutmak için geliştirilen algoritmaların kullanımı sebebiyle, hem kâr hem de toplumsal fayda odaklı etkin bir kontrol mekanizması geliştirememektedir. Bu nedenle tüm dünyada sivil toplum kuruluşları ve farklı disiplinlerden akademisyenler bir araya gelerek, bir mücadele mekanizması geliştirmek için çalışmalar yürütmeye başlamıştır. Türkiye'de de dijital mecralarda nefret söyleminin üretimine, dolaşımına ve yaygınlaştırılmasına karşı çok sayıda çalışma yürüten Hrant Dink Vakfı (HDV), Boğaziçi Üniversitesi ve Sabancı Üniversitesi ile, sosyal medyada nefret söyleminin tespitine yönelik olarak bir **nefret söylemi tespit ve sınıflandırma aracı** geliştirmek amacıyla bir araya gelmiştir.

Bu rapor, iki üniversitedeki Bilgisayar Mühendisliği, Dilbilim ve Kültürel Çalışmalar bölümlerinden on üç araştırmacı ve HDV ASULIS Dil, Diyalog, Demokrasi Laboratuvarı tarafından **Dijital Teknolojileri Kullanarak Nefret Söylemi ve Ayrımcılıkla Mücadele** projesi kapsamında yapılan çalışmaları ve geliştirilen nefret söylemi tespit aracı *pari*'yi tanıtmak amacıyla hazırlanmıştır. Raporda kabul edilen nefret söylemi tanımı, geliştirilen tespit ve sınıflandırma aracını eğitmek üzere kullanılan verinin toplanması için seçilen hedef gruplar ve anahtar kelimeler açıklanmaktadır. Her nefret söylemi kategorisi örneklerle açıklanmakta, etiketleme prosedürü açısından ortaya çıkabilecek sorular ele alınmaktadır. Ayrıca toplanan veri kullanılarak nefret söyleminin tespiti, sınıflandırılması ve derecelendirilmesi için geliştirilen model sunulmaktadır.

Dijital Teknolojileri Kullanarak Nefret Söylemi ve Ayrımcılıkla Mücadele Projesi

Dijital Teknolojileri Kullanarak Nefret Söylemi ve Ayrımcılıkla Mücadele projesi 2022 yılında başlamış olup, dilbilim, bilgisayar bilimleri, sosyal bilimler, bilişim sektörü ve sivil toplum gibi farklı alanlar arasında işbirliği kurarak, dijital alanda

nefret söylemi, ayrımcılık ve dezenformasyon ile mücadele etmeyi amaçlamaktadır. Projenin temel çıktısı yapay zekâ teknolojisi kullanarak çevrimiçi nefret söylemini tespit edecek bir nefret söylemi tespit ve sınıflandırma aracının üretilmesidir.

Geleneksel yöntemlerle nefret söylemi tespiti, büyük ölçüde insan emeğine dayanan bir süreçtir. Dijital medya platformlarının kullanıcı sayılarının ve dijital medyanın kullanımının hızla artması ve bununla birlikte medya izleme çalışmalarının insan emeği ile sürdürülebilir olmaması, bu projenin ortaya çıkışında önemli bir rol oynamıştır. Sosyal medya firmaları, politika değişiklikleri, kullanıcı sözleşmelerinde yaptıkları detaylı bilgilendirmeler ve çeşitli projeler ile nefret söylemi ve dezenformasyonla mücadele etmeye başlamış olsalar da, yapmış oldukları çalışmalar ve değişiklikler ile platformun kendi çıkarlarını korumayı amaçlamaktadırlar. Bu nedenle, hem şirketlerin kendi belirledikleri nefret söylemi sınırlarından çıkılması hem de nefret söyleminin kökenlerini daha iyi anlayarak etkili ve bilimsel önlemler alınmasına zemin hazırlamak için, nefret söyleminin hak temelli çalışan sivil toplum örgütleri ve akademik kurumlar tarafından bağımsız ve nesnel bir şekilde tespit edilip incelenmesi gerekmektedir. **Dijital Teknolojileri Kullanarak Nefret Söylemi ve Ayrımcılıkla Mücadele** başlıklı proje kapsamında **etnik, dinî ve cinsiyet temelli ayrımcılıkla** mücadele etmek ve bu gruplara yönelik nefret söylemini tespit etmek için, yeni teknolojilerden ve yapay zekâdan yararlanılarak bir dijital araç geliştirilmiştir. Proje kapsamında geliştirilen otomatik nefret söylemi tespiti aracının açık kaynak olarak sunulmasıyla, nefret söylemi izleme çalışmalarına katkı sağlanması ve nefret söylemi ve ayrımcılıkla mücadelede daha etkili ve sürdürülebilir bir çözüm sunulması amaçlanmaktadır.

Türkiye'de birçok grup farklı biçimlerde nefret söyleminin hedefi olmaktadır. **Düzenli ve düzensiz göçmenler, mülteciler, sığınmacılar, yasal olarak tanınan ve tanınmayan azınlıklar, etnik ve dinî gruplar, kadınlar, LGBTİ+'lar, engelliler** ve benzeri kırılgan gruplar üzerine çalışan akademisyenler, araştırmacılar, sivil toplum örgütleri, politika yapıcılar, meslek örgütleri ve benzeri kurumlar bu projenin nihai yararlanıcılarını oluşturmaktadır.

Proje sürecinde geliştirilen araç için X[1] adlı sosyal medya platformunda bulunan paylaşımlar toplanmıştır. X'in bir kullanıcı türevli içerik platformu olması, yani kullanıcıların güncel olaylar hakkında anlık yazılı içerikler ile paylaşım yapması üzerine kurulu olması, metin bazlı sosyal medya çalışmalarının akademik araştırmalar için uygunluğu ve platformun siyasi içerikli söylemler için sıkça tercih edilmesi söz konusu platformu proje için uygun kılmaktadır.

1 2023 yılının Haziran ayında Twitter olarak bilinen sosyal medya platformunun adı X olarak değiştirilmiştir. Platformun adı bu raporda X olarak geçmektedir.

Veri toplamak amacıyla X'in akademik API'si ve kazıma (*scraping*) aracılığıyla Yahudi, Yunan, Arap, Alevi, Ermeni, Kürt, LGBTİ+ ve mülteci karşıtlığına yönelik içerikler, belirlenen hashtag ve anahtar kelimeler kullanılarak çekilmiştir. Bu hashtag ve anahtar kelimeler, Türkiye'de sıkça nefret söylemine maruz kalan gruplarla ilgilidir ve gündem düzenli olarak takip edilerek seçilmiştir. Etiketleme için toplamda **16,254 tweet kullanılmış ve her bir tweet üç kişi tarafından etiketlenmiştir**. Her etiketleyici, etiketleme öncesinde farklılıkları ortadan kaldırmak ve tutarlı bir etiketleme süreci oluşturmak için aynı eğitimden geçmektedir. Aracı geliştirmek için kullanılan veri bu kolektif çabanın sonucu olarak ortaya çıkmaktadır.

1 NEFRET SÖYLEMİ ETİKETLEME KILAVUZU

Yapay zekâ ve derin öğrenme algoritmaları için büyük veri önemli bir yere sahiptir. Proje kapsamında geliştirilen yapay zekâ aracını eğitmek için Türkçe sosyal medya içerikleri toplanmış ve nefret söylemi içerip içermedikleri, kategori ve şiddet dereceleri göz önüne alınarak etiketleme yapılmıştır. Bu kılavuzda gerçek tweet örnekleri kullanılmıştır. Kullanıcı adları ve başka tanımlayıcı ifadeler çıkartılmıştır. Ayrıca, tweetlere ek olarak, aracın geliştirilmesi için Hrant Dink Vakfı'nın yazılı basındaki nefret söylemine yönelik olarak yürütmekte olduğu "Medyada Nefret Söyleminin İzlenmesi Projesi"nin[2] on yıllık verisi örnek veri havuzu oluşturmak amacıyla kullanılmıştır. Proje kapsamında Ortadoğu ve Kuzey Afrika (ODKA) bölgesi hedeflendiği için daha ufak kapsamda aynı çalışmalar Arapça[3] içerikler için de yapılmıştır.

Bu kılavuz, X paylaşımları üzerinden yapılan çalışma kapsamında, toplanan verilerin etiketleme yöntemlerini açıklamak ve gelecekteki araştırmalara örnek oluşturmak amacıyla hazırlanmıştır. Başka platformlardan gelen veriler için kullanılması durumunda gerekli değişiklikler yapılmalıdır.

2 https://hrantdink.org/tr/asulis/faaliyetler/projeler/medyada-nefret-soylemi/256-medyada-nefret-soyleminin-izlenmesi

3 Bu proje kapsamında aracın eğitilmesi için hem Türkçe hem de Arapça tweetler etiketlenmiştir. Ancak kılavuzda sadece Türkçe örnekler kullanılmaktadır.

1. SÖYLEMİN NEFRET SÖYLEMİ OLUP OLMADIĞINA NASIL KARAR VERİYORUZ?

Nefret söyleminin evrensel olarak kabul görmüş, değişmez bir tanımı yoktur. Bu nedenle nefret söylemi üzerine olan çalışmalar farklı nefret tanımları üzerine kurulmaktadır. Bu proje kapsamında Avrupa Konseyi Bakanlar Komitesi'nin 1997 tarihli nefret söylemi hakkındaki tavsiye kararında ortaya koyduğu tanım temel alınmıştır:

"Nefret Söylemi kavramı, ırkçı nefreti, yabancı düşmanlığını, Yahudi düşmanlığını veya azınlıklara, göçmenlere ve göçmen kökenli insanlara yönelik saldırgan ulusalcılık ve etnik merkezcilik, ayrımcılık ve düşmanlık şeklinde ifadesini bulan, dinsel hoşgörüsüzlük dahil olmak üzere hoşgörüsüzlüğe dayalı başka nefret biçimlerini yayan, teşvik eden, savunan veya meşrulaştıran her tür ifade biçimini kapsayacak şekilde anlaşılacaktır."[4]

Tanıma ek olarak, bir söylemin nefret söylemi olup olmadığına karar vermek için söylemin üretildiği toplumun özellikleri, bağlam, gündem, dilin özellikleri gibi pek çok farklı sebep de göz önünde bulundurulmalıdır. Söylemin nefret söylemi olup olmadığına karar verilmesi için öncelikle üç temel soru önemlidir:

- Söylemde hangi gruptan ya da kimlikten bahsediliyor?
- Söylem, bu gruba veya kimliğe nasıl yaklaşıyor?
- Söylemin olası etkileri ve sonuçları ne olabilir? (İnsan hakları ihlallerine yol açabilir mi?)

Benimsenen nefret söylemi tanımı doğrultusunda aşağıdaki yönlendirmeler belirlenmiştir:

- Tweet'te doğrudan bir ulusal, etnik, dinî kimliği veya cinsiyet kimliğini hedef alan düşmanlaştırıcı, ayrımcı ve kutuplaştırıcı bir söylem varsa **nefret söylemi** olarak etiketlenmelidir. Doğrudan bir ulusal, etnik, dinî kimliği veya cinsiyet kimliğini hedef almayan söylemler ifade özgürlüğünü de gözeterek **nefret söylemi değil** olarak etiketlenmelidir.
- Tweet'te *örtülü* bir nefret söylemi varsa, yani tweet içeriğinde nefret söylemi görünmese de, etiketleyen kişi bağlamı bildiğinden yazılanın nefret söylemi içerdiğini anlıyorsa, **nefret söylemi** olarak etiketlenmelidir.

4 Avrupa Konseyi. 1997. *Bakanlar Komitesi'nin Üye Devletlere Yönelik "Nefret Söylemi" Konulu R (97) 20 No'lu Tavsiye Kararı. Medya ve Bilgi Toplumu Alanında Bakanlar Komitesi Tavsiye ve Bildirgeleri*, 106–108. Strazburg: Avrupa Konseyi.

- Verinin doğruluğu açısından, nefret söylemi olup olmadığına karar verilemeyen söylemler **emin değilim** olarak işaretlenmelidir. (Emin olunmayan tweet'ler toplanıp tekrar değerlendirilmektedir.) Nefret söylemi yok ya da emin değilim şeklinde işaretleme yapıldığında, yine de etiketleme çalışmasının diğer kısımları **boş bırakılmamalıdır**.

Sosyal medyada paylaşılan içeriklerde görsellerin, emoji'lerin, hashtag'lerin, etiketlerin, devrik cümlelerin, kısaltmaların ve sarkastik cümlelerin vb. kullanımı içeriğin nefret söylemi olup olmadığına karar vermeyi zorlaştırabilmektedir. Bu noktada paylaşımın bağlamı önem kazanmaktadır. Yönlendirmelerin detayları ve zorlayıcı örneklerin (örtülü nefret söylemi gibi) etiketlenmesinde aldığımız kararlar, ilerleyen bölümlerde daha ayrıntılı şekilde ele alınmıştır. Kabul edilen nefret söylemi tanımı ve yukarıdaki üç temel soruya dayanarak etiketleme arayüzü oluşturulmuş ve aşağıda adım adım açıklanmıştır.

2. ETİKETLEME ARAYÜZÜ

2.1. Hedef grubun belirlenmesi

Veri seti oluşturmak için çekilen tweet'ler, sıklıkla nefret söyleminin hedefi olan kimlikler için kullanılan belli anahtar sözcükler ve hashtag'ler üzerinden seçilmiştir. Hedef grubun belirlenmesi, arayüzde **"Genel Tutum ve Duruş"** başlığı altında yer almaktadır. Birden fazla hedef grup bulunduğu için arayüzde de bu başlıktaki kimlik ibaresi veri setindeki kimliğe göre değiştirilmektedir. Örneğin, "#birgeceansızıngelebiliriz", "#Yunankaşınıyor" ve "denize dökmek" gibi anahtar sözcük ve hashtag'ler Yunan kimliğine karşı yapılan nefret söylemi tespiti için kullanılmaktadır. Buna göre çekilen tweet'lerin veri setinde "Genel Tutum ve Duruş" kısmında "Yunan karşıtı" ibaresi geçmektedir. Diğer ibareler "Ermeni karşıtı", "LGBTİ+ karşıtı", "Alevi karşıtı" ve benzeri ifadeleri içermektedir. Bu kılavuzda tek bir örnek üzerinden ilerlemek için "Yahudi Karşıtı" seçilerek örneklerle açıklama yapılmıştır.

Bu kısımda karşımıza çıkan seçenekler:

- Emin değilim
- Yahudi karşıtı
- İfade özgürlüğü(nötr)[5]
- Alakasız

5 Etiketleme arayüzünde, ifade özgürlüğü kapsamına giren ifadeler için "nötr" seçeneği sunulduğundan, bu durum raporda da aynı şekilde belirtilmiştir.

Tweet'in Yahudi karşıtı olup olmadığına karar verilemiyorsa veya başka bir kimliğe yönelik bir söylemin genel tutumuna dair bir kararsızlık yaşanıyorsa, tweet'in **"Emin değilim"** olarak etiketlenmesi gerekiyor.

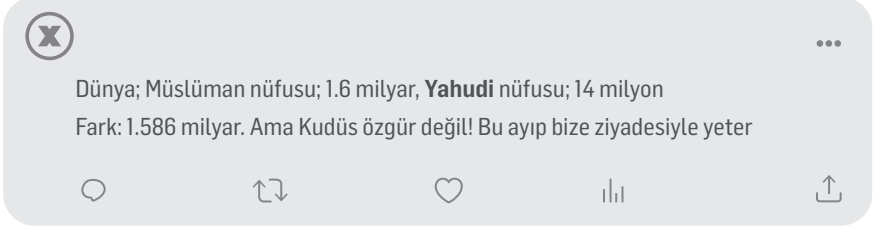
Tweet'in Yahudi kimliğine yönelik ayrımcı söylem veya nefret söylemi içerdiğine karar verildiyse, **"Yahudi karşıtı"** seçeneğinin işaretlenmesi gerekiyor. Farklı hedef gruplarına yönelik tweet'ler için etiketlemenin bu kısmında "Mülteci Karşıtı", "Yunan Karşıtı", "LGBTİ+ karşıtı" gibi seçenekler de mevcut.

Tweet'in söz konusu hedef gruba yönelik ifadesinin tarafsız olduğuna, nefret söylemi içermediğine veya söz konusu kimliğin tamamına yönelik bir söylem üretmediğine karar verildiyse, **"Nötr"** kapsamında değerlendirilmesi gerekiyor.

Tweet'in Yahudi kimliğine yönelik nefret söylemi içermediği ve tweet içeriğinin alakasız olduğu düşünülüyorsa **"Alakasız"** seçeneğini işaretlemek gerekiyor. Ayrıca, tweet'in Yahudilere yönelik nefret söylemi içermediği ancak, farklı bir kimliğe yönelik nefret söylemi içerdiğinin düşünüldüğü durumlarda "Alakasız" olarak etiketleme yapılması gerekiyor. Örnek olarak, Yahudi kimliğine dair hazırlanmış bir veri setinde etiketleme yaparken, Yahudi kimliğine yönelik nefret söylemi içermeyen ancak, LGBTİ+lara yönelik nefret söylemi içeren bir tweet, önce ilgili konudan farklı olduğu için "Alakasız" olarak etiketlenmeli, sonrasında ise etiketlemenin "hedef grup" bölümünde **"Cinsel Yönelim"** seçeneği seçilmelidir.

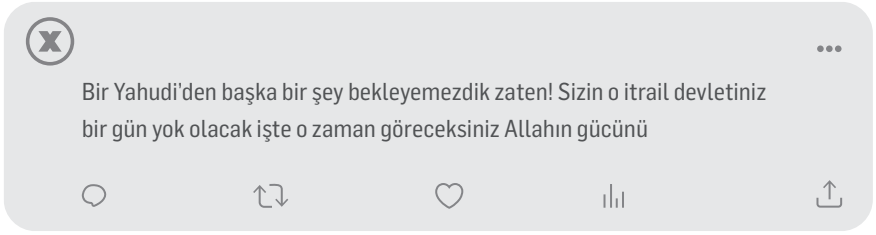
Aşağıda her genel duruş ve tutum seçeneği için birer örnek verilmiştir:

Emin değilim



Yukarıdaki örnekte Müslüman ve Yahudi nüfusları sayılar verilerek karşılaştırılmakta ve bu sayı farkı üzerinden Kudüs'ün durumuna değinilmektedir. Ancak, bu tweet'in Yahudi karşıtı olup olmadığı tam olarak anlaşılamamaktadır. Bu durumda "Emin değilim" şıkkı işaretlenmelidir.

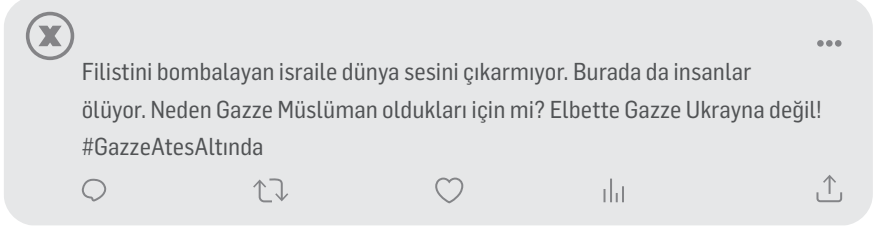
Yahudi karşıtı



İsrail-Filistin veri setinde "itrail" ifadesi karşımıza sıkça çıkan ifadelerden biridir. Ancak bu ifadeyi nasıl değerlendireceğimiz tweet'lerin içeriğine göre değişmektedir. Örneğin, bu tweet'le ilgili iki nokta bulunmaktadır. Birincisi, bu örnekte Yahudi kimliği ve İsrail bir bütün olarak görülmektedir. Bu tutum İsrail'in yaptırımlarını bütün Yahudilerle bağdaştırmaktadır. Küfür, hakaret, aşağılama ve insandışılaştırma yoluyla Yahudi kimliği hedef alınmakta ve nefret söylemi üretilmektedir. İkinci olarak Yahudi kimliğiyle bağdaştırılan olumsuz özelliklere atıfta bulunulmasıdır. Yazan kişinin gözünde acımazsızlık (ya da bunun gibi özellikler) Yahudilerin sahip olduğu içsel ve doğuştan gelen bir özellik olarak gösterilmektedir.

İfade özgürlüğü (nötr)

Bunların dışında tüm Yahudilerin hedef alındığı ayrımcılık, önyargı, düşmanlık veya nefret ifade eden ifadelerin bulunmadığı; belli bir olayın sadece haber verildiği, ya da yaşanan bir olayın veya uygulamanın eleştirildiği, bunlara karşı duyulan üzüntünün belirtildiği ifadeler "Yahudi karşıtı" olarak kabul edilmemelidir. Nefret söylemi de içermediği için çalışmanın temel amacının dışında olduklarından "nötr" olarak işaretlenmeleri gerekmektedir.

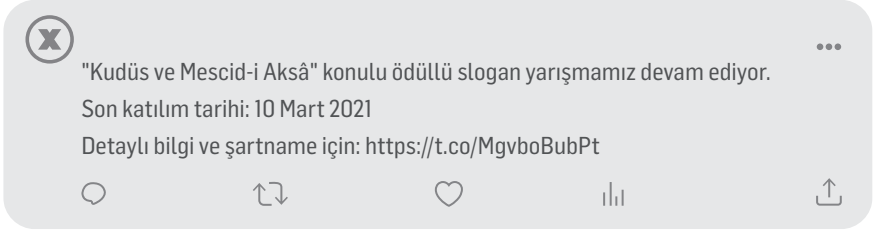


Yukarıdaki tweet'te Yahudilerin tümüne yönelik bir nefret söylemi görülmemektedir. Hamas'ın 7 Ekim 2023'te İsrail'e saldırmasının ardından İsrail'in Gazze'ye saldırması ve hâlâ devam etmekte olan savaşta sivillerin hayatını kaybediyor olmasına vurgu yapılmaktadır. Bombalamadan söz edilerek İsrail'in kullandığı orantısız güç eleştiri-

lirken dünyadaki diğer devletlerin bu faaliyete karşı sessiz kalmaları eleştirilmektedir. Ayrıca Ukrayna ve Rusya arasındaki savaşa gönderme yapılarak, Ukrayna'nın işgali ve devam eden savaşa karşı diğer devletlerin, benzer şekilde davranmadıklarına vurgu yapılarak bir ayrımcılık yapıldığı ima edilmektedir. Öte yandan Gazze'de yaşayanların Müslüman olup olmadığına yönelik yorum, ayrımcılığın sebebi olarak gösterilmektedir. Sadece belirli bir olay ve buna karşı verilen tepkideki eşitsizlik eleştirildiğinden, tüm Yahudi topluluğuna yönelik bir nefret söylemi oluşturulmadığından, bu tweet'in düşünce ve ifade özgürlüğü olarak işaretlenmesi gerekmektedir.

Alakasız

Bu yarışmanın içeriğinde İsrail ve Yahudi karşıtı ifadelerin olup olmayacağını öngöremediğimiz gibi tweet'in kendisi de herhangi bir Yahudi karşıtlığı içermemektedir. Bu yüzden "Alakasız" şıkkının işaretlenmesi gerekmektedir.



Tweet'lerde tek bir hedef grup ve birden fazla hedef grup olabileceği gibi bazı tweet'lerde hedef grup net bir şekilde anlaşılamayabilir. Bu gibi durumlarda tweet etiketlemesi yaparken:

- **Tek bir hedef grup** olan tweet'lerde **ilgili grup** seçilmelidir. **Birden fazla grup** hedefleniyorsa, **o grupların tümü** seçilmelidir.
- **Tweet'teki ifadede nefret söylemi yoksa,** ancak, tweet yine de ilgili kimlik grubuyla ilgili bir görüş bildiriyorsa, hedef grup **ifade özgürlüğü (nötr)** olarak işaretlenmelidir. Sadece nefret söylemi içeren tweet'lerde ilgili kimlik hedef grup olarak işaretlenmelidir.
- **Açıkça belirtilen bir hedef grup yoksa**, hem ifade özgürlüğünü korumak hem de "yanlış pozitif" riskini azaltmak adına **hedef grup var** işaretlenmelidir. (Örneğin, bir söylemin içinde doğrudan mülteci, Suriyeli, Afgan gibi kelimeler geçmese bile, mültecilerle ilgili çekilen tweet'lerde aslında örtülü/gizli bir hedef grup vardır. Bu nedenle hedef grup seçilmelidir.)

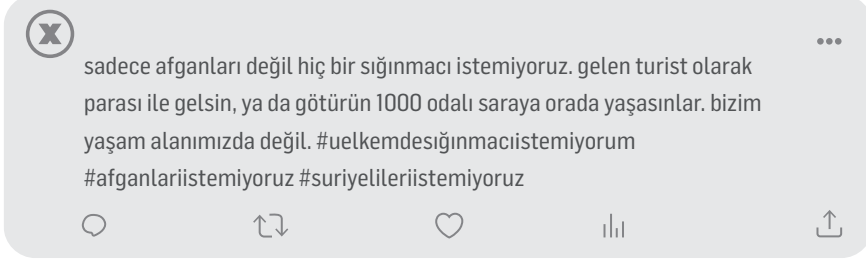
- **Demografik/Sosyo-ekonomik grup:** Bu kategori altında nefret söyleminin **ırk/etnik köken, ülke/milliyet, din, cinsiyet** veya **cinsel yönelim** özelliklerinden hangisini hedef aldığı seçilmelidir. Bu grup bahsedilen özellikler üzerinden bir grubun tümünü hedef aldığı veya bir kişi ya da grubun buradaki kimlik özelliklerinden ötürü hedef alındığı örnekleri ifade etmektedir. Birden fazla kimlik özelliğinin hedef alındığı ifadelerde ilgili özelliklerin tümü işaretlenmelidir.
- Bazı söylemlerin **hedef gruba mı yönelik yoksa düşünce grubuna mı yönelik olduğu net değildir.** Böyle durumlarda nefret söylemi kategorisini seçmek zor olabilmektedir. Örneğin, bazı durumlarda tweet'i yazan kişinin Siyonizm karşıtı mı yoksa İsrail ve/veya Yahudi karşıtı mı olduğu belirsiz gözükebilir. Bu tweet'lerde, **hedef grup belli değil** seçilmelidir.
- Hedef grup birden fazla şeklinde etiketleme yapılıyorsa demografik/sosyoekonomik kısmında ülke/milliyet ve din gibi iki farklı yer işaretlenebilir.

2. 2. Söylemin türünün belirlenmesi

Hedef grup belirlendikten sonraki aşamada, bu grupların nasıl hedef alındığı detaylandırılmaktadır. Bu nedenle arayüzde öncelikle tweet'in ayrımcılık mı yoksa nefret söylemi mi içerdiği belirlenmelidir.

Dışlama / ayrımcı söylem

Bir grubun tamamının ya da kimliği sebebiyle grubun bazı üyelerinin, topluma dahil olma, hak ve özgürlüklerden faydalanma gibi alanlarda baskın gruptan olumsuz anlamda farklı görüldüğü söylemlerdir.



Sığınmacı kimliğiyle ülkeye gelen insanların bütününe karşı yapılan bu söylem, bu gruba sığınma ve barınma gibi hakları tanımaması sebebiyle ayrımcı söylem/dışlama olarak etiketlenmelidir.

Nefret söylemi

Tweet'in nefret söylemi olması durumunda ise söylemin bir grubu ne şekilde hedef aldığına göre analiz yapmak için nefret söylemi kategorileri kullanılmaktadır. Bir tweet, proje kapsamında belirlenen ve aşağıda belirtilen nefret söylemi kategorilerine göre etiketlenmelidir. Tweet'ler birden fazla nefret söylemi kategorisine uyuyorsa, **birden fazla kategori seçilmelidir**. Etiketleyen kişi mümkün olduğu kadar bir kategori seçmeye çalışmalı fakat yine de birden fazla hedef grup olduğu durumlarda farklı kategorilere giren nefret söylemi örnekleri varsa çapraz etiketleme yapılabilir.

2.3. Nefret söylemi kategorisinin belirlenmesi

Söylemin bir grubu ne şekilde hedef aldığına göre, bu konudaki uluslararası bilimsel çalışmalardan yararlanılarak ve ülkeye özgü dil ve kültür farklılıkları dikkate alınarak belirlenen kategoriler, bir analiz birimi olarak, söz konusu metnin neden nefret söylemi içerdiğini anlamakta ve açıklamakta işlevsel bir rol oynamaktadır. HDV Medyada Nefret Söyleminin İzlenmesi Projesi'nde nefret söylemi, dört kategoriye ayrılmaktadır:

1) **abartma/yükleme/çarpıtma/genelleme:** Bir kişi ya da olaydan yola çıkarak bir topluluğa yönelik olumsuz genelleme, çarpıtma, abartma, olumsuz atıf içeren söylemler.

   Bu kategoride nefret söylemi en çok genelleme ile üretilir.

   (örn. "Suriyeliler gına getirdi", "Yunan ölüme terk etti", "Yahudi havadan saldırdı", "Eşcinsel sapkınlar dehşet saçıyor", "Ermenilerin tazminat ve toprak hayalleri suya düştü", "Hıristiyan terörünü İslam'a maletti", "Akdeniz'de Rum Gerilimi")

2) **küfür/hakaret/aşağılama:** Bir topluluk hakkında doğrudan küfür, aşağılama, hakaret içeren söylemler (örn: "Küstah Rum'a Gözdağı", "Barbar Yunan", "Hadsiz Yahudi", "Danimarkalı itler iftar basıp, Kur'an yaktı", "Barbar ve ahlaksız Fransızlar").

3) **düşmanlık/savaş söylemi:** Bir topluluk hakkında düşmanca, savaşı çağrıştıran ifadelerin yer aldığı söylemler (örn. "Rum vahşeti", "Haydut Rumlar Ateşle Oynuyor", "Rumlar yine tahrik ediyor", "Mültecilere Yunan zulmü", "Hans iyice kudurdu").

4) **simgeleştirme:** Doğal bir kimlik öğesinin nefret, aşağılama unsuru olarak kullanıldığı, simgeleştirildiği söylemler (örn. "Bizi Eurovision'da Yahudi mi temsil edecek?" "Ermeni gibi konuştular", "Yunan aynı Yunan", "Yunan artığına Atatürk cevabı", "Rum ağzıyla rapor", "Cenk Tosun'a gavur eziyeti", "Yunan askerinden mültecilere bir gavurluk daha", "İçimizdeki İsraillile").

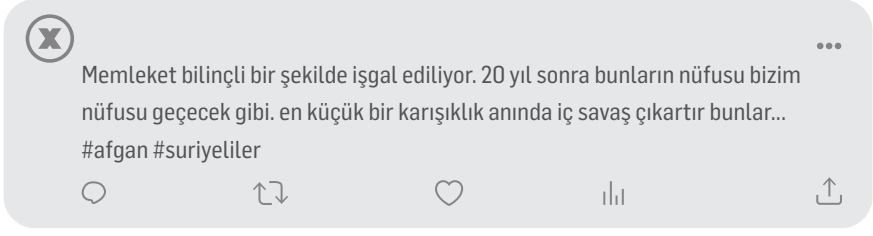
Bu kategoriler Dijital Teknolojileri Kullanarak Nefret Söylemi ve Ayrımcılıkla Mücadele Projesi kapsamında aşağıdaki şekilde genişletilerek, ayrıntılandırıldı:

- Emin değilim
- Nefret Söylemi Bulunmuyor
- Abartma, Genelleme, Yükleme, Çarpıtma
- Küfür, Hakaret, Aşağılama, İnsandışılaştırma
- Düşmanlık, Savaş, Saldırı, Öldürme, Yaralama Tehdidi
- Simgeleştirme

Aşağıda etiketleme sürecinde kullanılan kategorilerin kısa tanımlamaları ve her biri için seçilmiş örnekler sıralanmaktadır. Her kategori için üç örnek verilip, kısaca incelenmektedir. Birinci örnek söz konusu kategoriyi açıkça içerirken, ikinci örnek daha belirsiz görünse de yine de söz konusu kategoriden nefret söylemi içermektedir. Üçüncü örnek ise ilk bakışta söz konusu kategoriye ait olabilirmiş gibi görünmekle birlikte, aslında ya nefret söylemi içermemekte ya da başka kategorilere ait söylemler içermektedir.

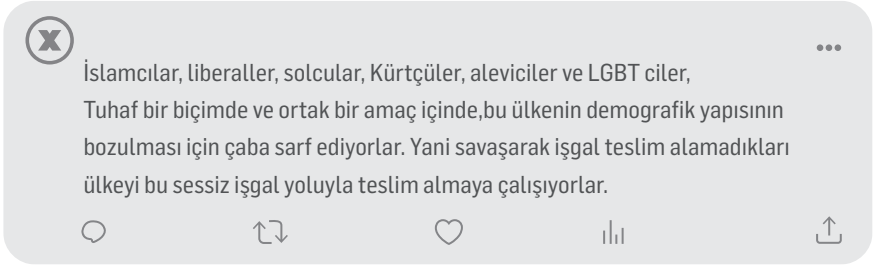
Abartma: Bir olayı, durumu ya da aksiyonu olduğundan daha büyük gösterme, gerçeğe dayanmayan sonuçlara ve çıkarımlara varma.

Net örnek:



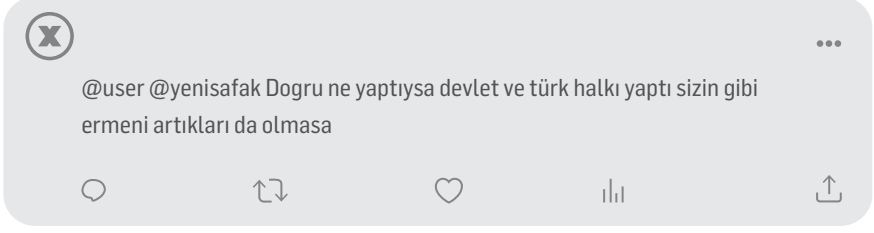
Bu örnekteki öngörü olası bir senaryoyu sergilemeyip, bir iç savaş tehdidi çıkarımı yaptığından abartı yoluyla nefret söylemi olarak etiketlenebilir.

Net olmayan örnek:



Yukarıdaki örnekte, mültecilerin ülkeye alınmasının insan hakları adı altında aslında gizli bir amaca hizmet ettiği düşüncesi vurgulanmakta ve "sessiz işgal" yoluyla Türkiye'nin demografik yapısının bozulmasının hedeflendiği belirtilmektedir. Bu örnek abartma yoluyla nefret söylemi olarak sınıflandırılmalıdır.

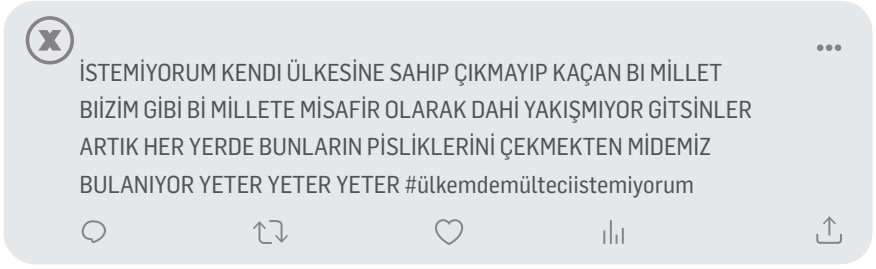
Yanıltıcı örnek:



Yukarıdaki örnekte ülke içinde yapılan "doğru" işler bir gruba ithaf edilirken bu duruma tezat oluşturduğu düşünülen olayların sorumluluğu "sizin gibiler de olmasa" diyerek Ermenilere yüklenmiştir. Noktalama işaretlerinin eksikliği tweet'in yorumlanmasında ve kategorize edilmesinde farklılık yaratabilir. Ancak, doğrudan abartma içermemektedir.

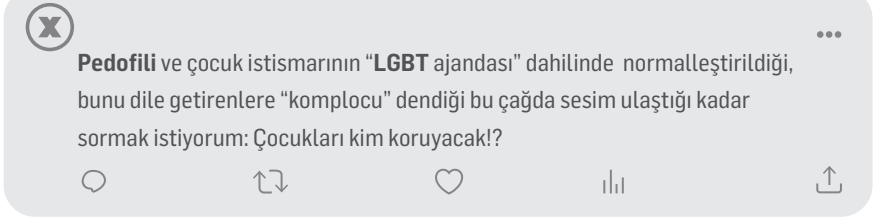
Çarpıtma: Bir olayı, durumu ya da aksiyonu gerçek verilerden saptırarak, okuyucunun algılama biçimi ve çıkarımlarını manipüle edecek biçimde, yanlış, eksik veya yanlış anlaşılmaya sebebiyet verecek şekilde aktarma.

Net örnek:



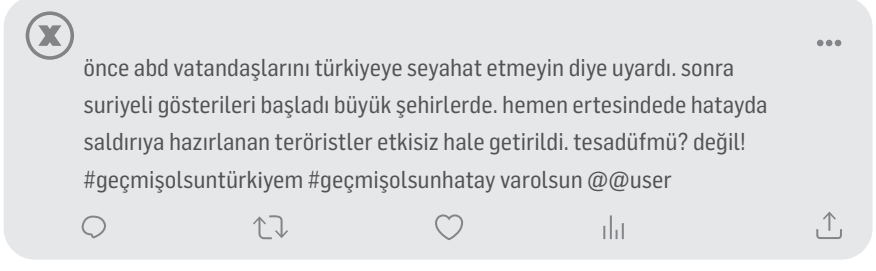
İç savaş nedeniyle göç etmek zorunda kalan Suriyelilerin, ülkelerinden kaçtıkları şeklinde yansıtılması, hem gerçeğin saptırılmasına hem de mültecilerin kendi ülkelerine yararlı olmadıkları gibi Türkiye'ye de faydalı olamayacağına dair bir algının oluşmasına neden olmaktadır. Kullanılan hashtag ile birlikte nefret söylemi oluşturulmuştur ve "çarpıtma" olarak kategorize edilmelidir.

Net olmayan örnek:



Yukarıdaki tweet pedofili karşıtı olmakla beraber aslında dolaylı olarak LGBTQ+ kimliğine sahip kişi ve grupların tecavüz ve pedofili destekçisi olduğunu iddia etmektedir. Bu sebeple çarpıtma yoluyla nefret söylemi kategorisine dahil edilebilir.

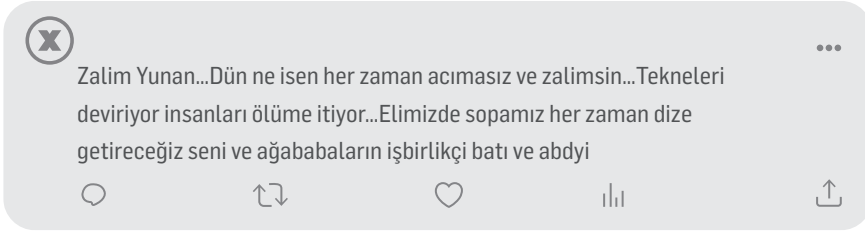
Yanıltıcı örnek:



Bu örnekte olaylar kronolojik olarak sıralanmaktadır. Tweet'i yazan kişinin olaylar arasında bağlantı olduğunu iddia etmesi gerçeği çarpıtmak değil, bireysel bir çıkarımda bulunmak olduğundan, bu tweet çarpıtma yoluyla nefret söylemi kategorisine alınmamalıdır.

Yükleme: Bir olay ya da durumun sebebi olarak bir grup veya kimliği asılsız bir şekilde öne sürme.

Net örnek:

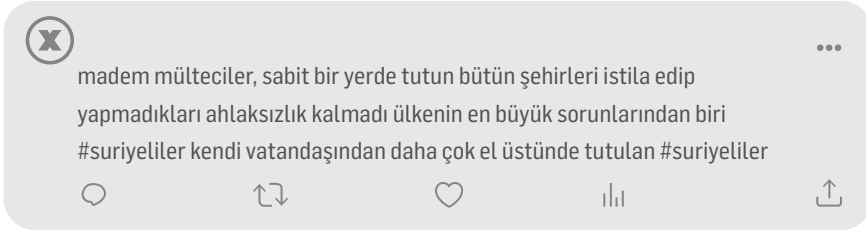


Yunan Sahil Güvenlik güçlerinin mülteci tekneleri Türk karasularına yasadışı olarak geri ittiğini söyleyen bir haberi alıntılayan bu tweet, sahil güvenlik güçlerinin aksiyonlarını bütün Yunanlara mal edip Yunanları zalimlikle suçladığından yükleme yoluyla nefret söylemine örnek gösterilebilir.

Net olmayan örnek:

Yukarıdaki tweet'te, Suriyeli mültecilerin güvenlik sorununa neden olduğuna dair net bir olay belirtilmemiştir. Ancak, bu tweet'te mültecilerin güvenlik sorunlarından sorumlu tutulduğu ve sonrasında hashtag kullanımıyla nefretin körüklendiği görülmektedir. Bu nedenle, örnek, yükleme yoluyla nefret söylemi olarak etiketlenebilir.

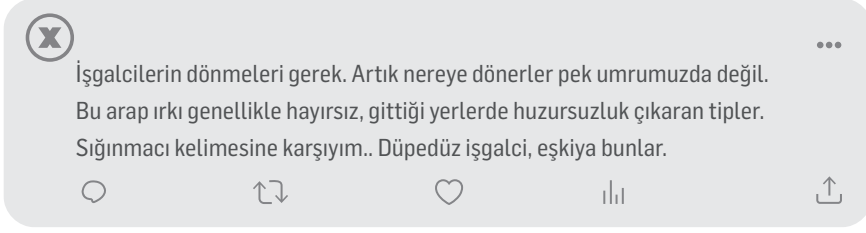
Yanıltıcı örnek:



Bu örnek, abartı ve genelleme yollarıyla Suriyeli mültecilere karşı nefret söylemi içermektedir. Belirli bir olayın ya da durumun sorumluluğunu mültecilere yüklemediği için, yükleme yoluyla nefret söylemi olarak etiketlenmemelidir. Doğru kategori abartma ve genellemedir.

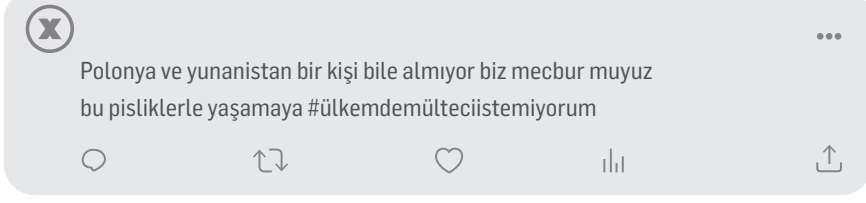
Genelleme: Bir olay, durum, özellik ya da aksiyonun kendisini ya da sonuçlarını bir kimliğin bütününe mal etme.

Net örnek:



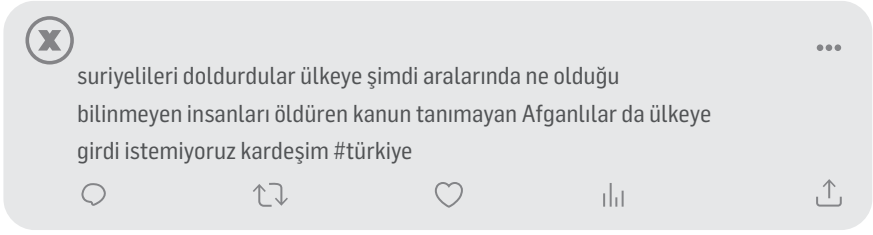
Arapların tamamı hedef gösterilerek yapılan ithamlardan dolayı, bu örnek genelleme yoluyla nefret söylemi olarak etiketlenebilir.

Net olmayan örnek:



Bu örnekte pislik sözcüğü, tweet'in devamındaki hashtag de göz önüne alındığında tüm mültecilere karşı kullanılmış bir hakaret olarak karşımıza çıkmaktadır. Küfür/hakaret/aşağılama etiketinin yanında, genelleme yoluyla nefret söylemi olarak da etiketlenmelidir.

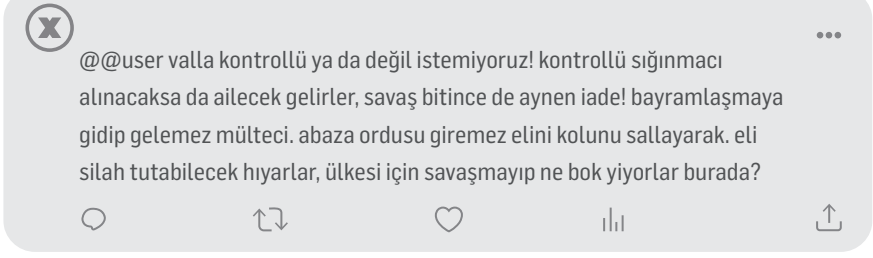
Yanıltıcı örnek:



Yukarıdaki örnekte Afgan kimlikli gruba karşı yapılan sözlü saldırı "aralarında" sözcüğünden dolayı grubun tamamına yapılan bir saldırı olarak ele alınamayacağından, bu tweet genelleme yoluyla nefret söylemi kategorisine dahil edilmemelidir.

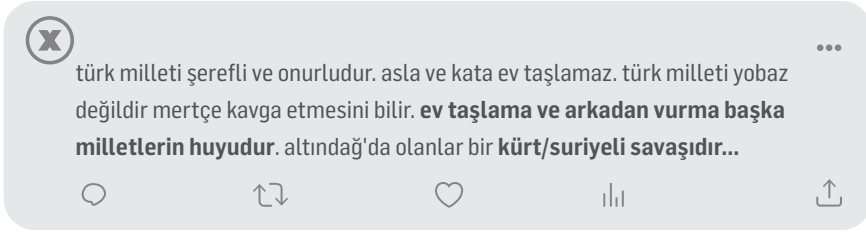
Hakaret: Bir ırka veya topluluğa, onur şeref ve saygınlığını rencide edebilecek nitelikte somut bir fiil veya olgu isnat etmektir. Hakaret, bir nitelik yakıştırmasıdır. Örneğin, hukuki olarak birine "sen hırsızsın" demek, hakaret sayılır. Bu gibi durumlarda olayın doğruluğuna bakılabilir. Eğer doğru ise, hafifletici neden kabul edilir.

Net örnek:



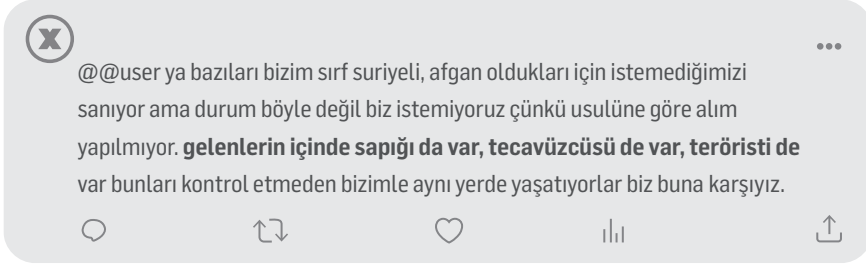
Örnekte, "abaza ordusu" ve "hıyarlar" gibi hakaret içeren kelimeler kullanılmıştır. Örneğin genelinden bu olumsuz sıfatların bahsedilen sığınmacıların tümüne yakıştırıldığı anlaşılmaktadır. Sığınmacıların tümü hedef alınarak hakaret edilmiş ve nefret söylemi oluşturulmuştur.

Net olmayan örnek:



Örnek, "Türk milleti" için olumlu sıfatlar sıralanarak başlamıştır. Örneğin devamında doğrudan olmasa da "Türk milleti" için sıralanan bu olumlu sıfatların belirtilen diğer topluluklarda olmadığı sezdirilmiştir. "Kürt" ve "Suriyeli" kelimeleri açık bir şekilde kullanılmış ve hedef gösterilmiştir. Bu yolla belirtilen topluluklara karşı nefret söylemi oluşturulmuştur.

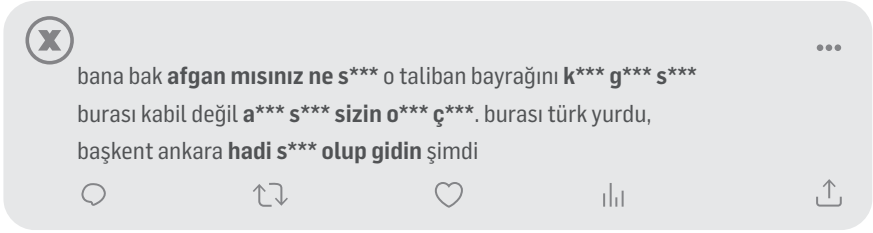
Yanıltıcı örnek:



Örnekte "sapık", "tecavüzcü", "terörist" gibi hakaret içeren ifadeler kullanılmıştır. Ancak, örneğin geneline baktığımızda bu ifadelerin "Suriyelilerin" ve "Afganların" tümüne yönelik söylenmediği anlaşılmaktadır. Gelen gruplar içerisinde suça meyilli insanlar olabileceği ve bu nedenle denetimli bir şekilde ülkeye girişlerinin kabul edilmesi gerektiği anlatılmaktadır. Bir devlet politikası eleştirisi yapılmaktadır. Bu nedenle bir nefret söylemi oluşturulmamıştır.

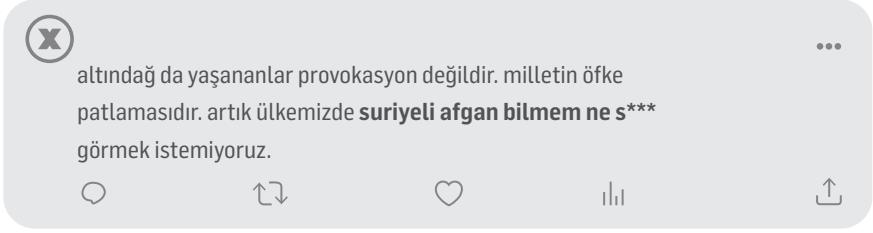
Küfür: Bir ırka veya topluluğa kültürel, geleneksel ilişkiler açısından küçültücü bir eyleme dönük bir istek belirtmek ya da varsaymaktır. Bir niteliğin yakıştırılması değildir, bir eyleme gönderme yapar. Hukuki olarak küfürün doğruluğu veya gerçekliği sorgulanmaz.

Net örnek:



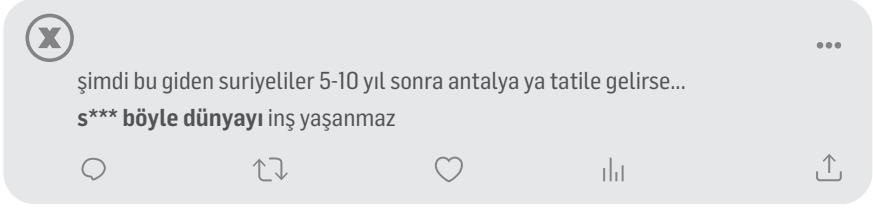
Örnekte görüldüğü gibi sansürlenmiş kısımlarda tüm "Afganlar" hedef alınarak küfür edilmiştir.

Net olmayan örnek :



Yukarıdaki bir önceki örnekte olduğu gibi doğrudan olmasa da, sansürlenmiş kısımda bir topluluğun tümü hedef alınarak küfür içeren ifadeler kullanılmıştır.

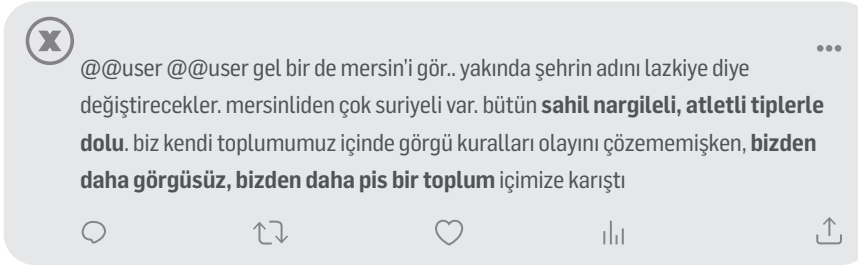
Yanıltıcı örnek:



Örnekte, sansürlenmiş kısımda küfür içeren bir ifade kullanılmış olsa da; bu ifadenin herhangi bir ırk veya topluluğu hedef almadığı, gelişigüzel söylendiği anlaşılmaktadır. Herhangi bir nefret söylemi içermemektedir.

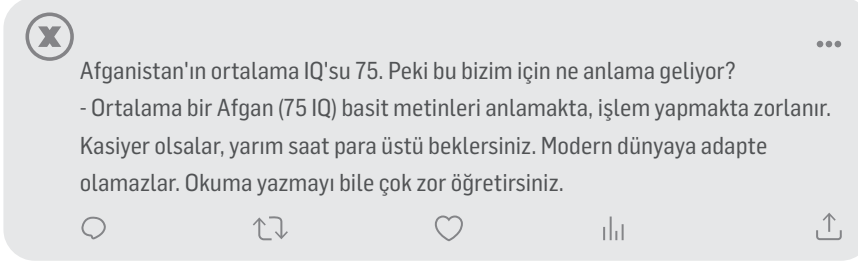
Aşağılama: Bir kişi, ırk veya topluluk için, genel kabul gören değerlerden daha azına sahip olduğu sanısı yaratmak veya bu değerler için küçük görmektir.

Net örnek:



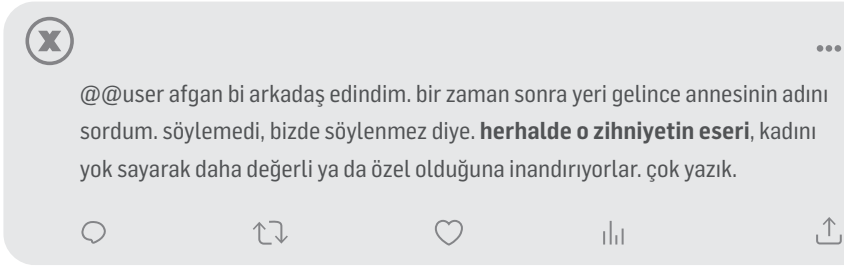
Örnekte, tüm "Suriyeliler" hedef alınarak "daha görgüsüz" ve "daha pis" oldukları iddia edilmiştir. Bu yolla bahsedilen toplum, var olan değerlerin daha azına sahip olduğu belirtilerek aşağılanmıştır. Ayrıca, örnekte geçen "nargileli, atletli tipler" ifadesi ile belli bir dış görünüş tüm topluma genellenmiş ve örneğin genelinden anladığımız kadarıyla bu dış görünüş üzerinden tüm toplum aşağılanmıştır.

Net olmayan örnek:



Yukarıdaki tweet'te daha önce bahsedilen bir araştırmanın sonuçları değerlendiriliyor. Bahsedilen araştırmadan yapılan alıntı doğru bile olsa, verilen sonuçların bir ortalama olduğu ve yapılan çalışmanın güvenilirliği göz ardı edilerek Afganların modern dünyaya adapte olabilecek bilişsel gelişime sahip olamayacakları ifade edilmiştir. Tweet'in tümüne baktığımızda sadece bir araştırmanın sonuçlarının değerlendirildiği bilimsel bir yorum gibi dursa da yapılan çıkarımlarla aşağılama yoluyla nefret söylemi oluşturulmuştur.

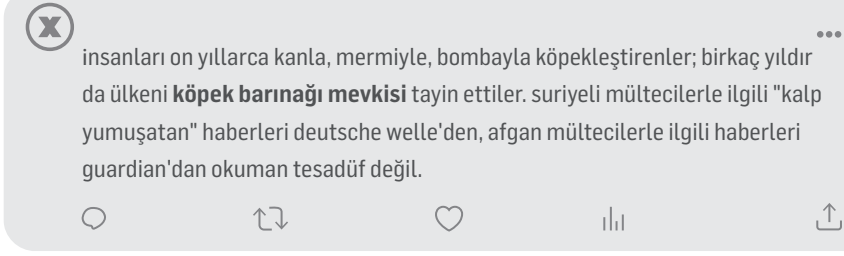
Yanıltıcı örnek:



Örnekte, belli bir zihniyetin değerleri aşağı görülmüştür. Bu aşağı görülen zihniyetin nasıl olduğuna, belli bir topluluğa mensup birinden yola çıkılarak ulaşılmışsa da; bahsedilen "o zihniyetin" tüm topluluğa ait bir değer olup olmadığı anlaşılamamaktadır. Bu zihniyet; belli grupların, örgütlerin empoze ettiği görüşler olarak okunabilir. Bu nedenle bu örnek, belli grupların empoze ettiği görüşlere karşı yapılan bir eleştiridir.

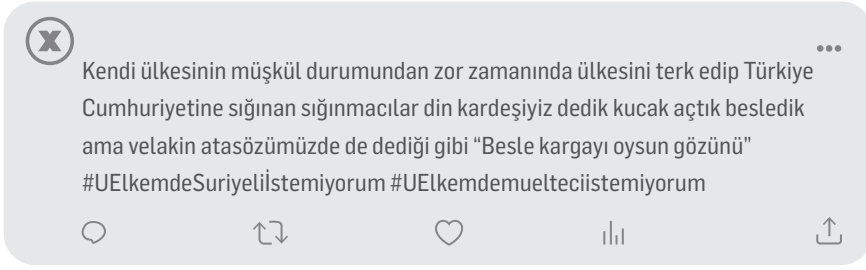
İnsandışılaştırma: Bir ırkı veya topluluğu insan dışı bir varlığa (örneğin bir hayvana) benzetme veya insan dışı varlıklara özgü eylem ve sıfatları bunlara yakıştırma yoluyla yapılan aşağılamalardır. Bu tarz kullanımların tamamı nefret söylemidir. Olumlu benzetmeler bu kapsamın dışında tutulmaktadır (örneğin, "kedi gibi uysal", "köpek gibi sadık"). Kullanılan sözcükler içinde "beslemek", "üremek" gibi fiiller bulunmaktadır.

Net örnek:



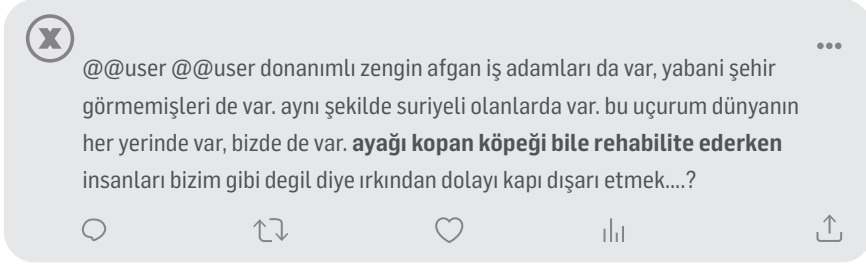
Örnekte açık bir şekilde görüldüğü gibi, "Suriyeliler"in "köpekleştikleri" söylenmiş, ülkede var oldukları için ülkenin "köpek barınağı" hâline geldiği ifade edilmiştir. "Suriyeliler" insan dışı bir varlık olan köpeğe benzetilerek aşağılanmıştır.

Net olmayan örnek:



Örnekte bulunan "beslemek" sözcüğü bazı durumlarda insanlar için de kullanılsa da; daha çok hayvanlarla bağdaştırılan bir kullanımdır. Bu tür canlılara özgü bir eylem belli bir topluluğa yakıştırılarak bu topluluk aşağılanmış ve hashtag ile desteklenerek nefret söylemi üretilmiştir.

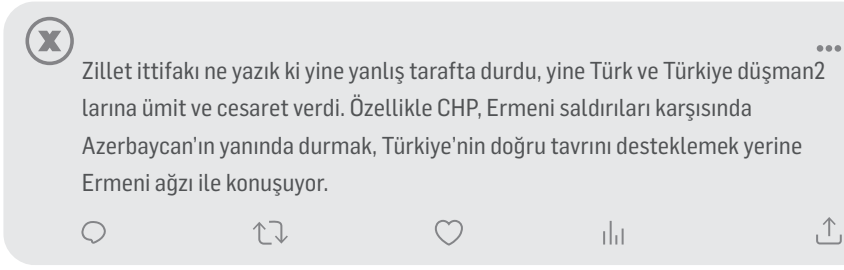
Yanıltıcı örnek:



Örnekte, "ayağı kopan köpeği bile rehabilite ederken" ifadesi ile bir karşılaştırma yapılmıştır. İfadede insan dışı varlıklarla bir ilişki kuruluyor gibi gözükse de; bahsedilen topluluğa yönelik bir yakıştırma yapılmamıştır. Ayrıca, aşağılama amacıyla kullanılmamıştır.

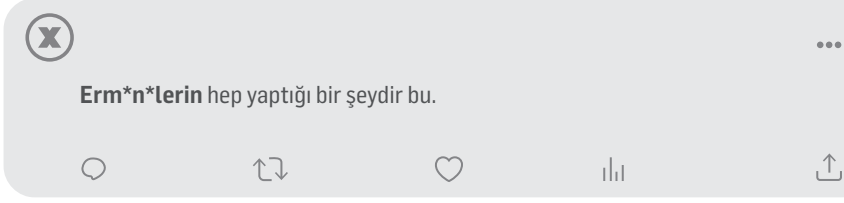
Simgeleştirme: Doğal bir kimlik öğesinin bir hakaret veya bir nefret ve aşağılama unsuru olarak kullanıldığı, simgeleştirildiği söylemlerdir. Diğer kategorilerde belli bir kimlik öğesi hedef alınırken, bu kategoride kimlik öğesinin kendisi hakaret, aşağılama veya nefret unsuru oluşturmak için kullanılır.

Net örnek:



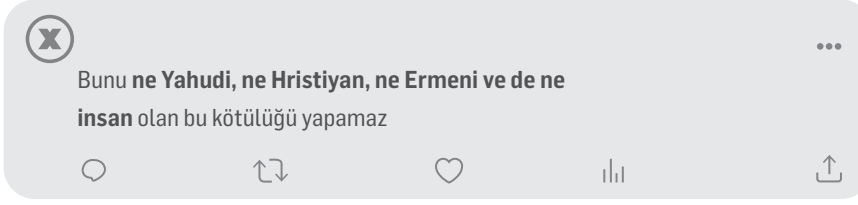
Örnekte "Ermeni ağzı" ifadesi oldukça olumsuz bir anlama geliyormuş gibi kullanılmıştır. Örnek, Ermenileri hedef alarak söylenmemiş olsa da, siyasi bir eleştiri yapılırken "Ermeni ağzı" ifadesi aşağılık bir unsurmuş gibi kullanılmıştır. Bu yolla nefret söylemi oluşturulmuştur.

Net olmayan örnekler:



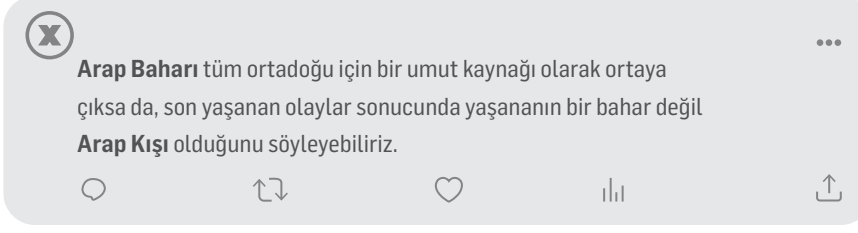
Örnekte bağlam açısından "Ermeniler" ifadesiyle hedeflenen grubun ne tür bir eylemle bağdaştırıldığı ilk bakışta anlaşılmamaktadır ve bahsedilen eylemin iyi veya kötü olduğunu bilmemiz mümkün değildir. Ancak, yıldız işaretleri kasıtlı olarak konulmuş ve "Ermeni" kelimesi küfür içeren bir ifadeymiş gibi sansürlenmiştir. Yıldız işareti sosyal medyada sıkça kullanılan ve hakaretlerle ilişkilendirilen bir ifadedir. Bu nedenle, aynı işaret bir kimlik öğesi için kullanılarak "Ermeni" kelimesi simgeleştirilip nefret söylemi oluşturulmuştur. Bu örneğin net olmayan örnek olarak kategorize edilmesinin sebebi, simgeleştirme yoluyla üretilen nefret söylemlerinde kimlik ifadesinin kullanımına dikkat edilmesinin önemini vurgulamaktır. Tweet örneğinde kimliğe atfedilen ve bu yolla üretilen bir nefret söylemi olmamasına rağmen, kimliğin kendisi simgeleştirilerek nefret söylemi oluşturulmuştur.

Örnekte, belirtilen kişinin etnik kökeninden bahsediliyormuş gibi gözükse de "iyi araştırın!" gibi ifadelerle "Ermeni" olmanın aşırı olumsuz bir şey olduğu sezdirilmiştir. Etnik kimliğin simgeleştirilmesine ek olarak siyasi bir figür olan Kemal Kılıçdaroğlu başka bir kökenden olduğu iddiasıyla hedef gösterilerek nefret söylemi oluşturulmuştur.



Örneğe baktığımızda bahsedilen ırk ve topluluklara yönelik olumlu bir ifade kullanılıyormuş gibi dursa da; örneğin genelinden bahsedilen ırk ve toplulukların insana ait değerlerin en alt sınırına sahip oldukları sezdirilmiş ve "onların bile yapmayacağı" ifade edilmiştir. Bu yolla simgeleştirilerek, bu ırk ve toplumlar aşağılanmıştır.

Yanıltıcı örnek:

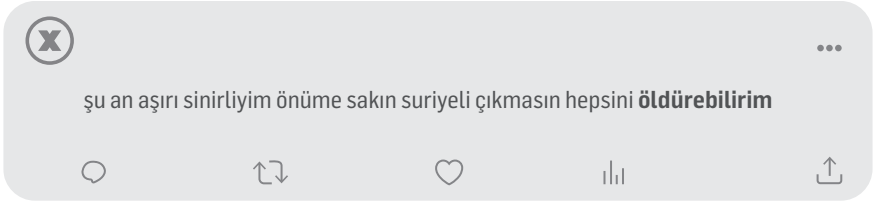


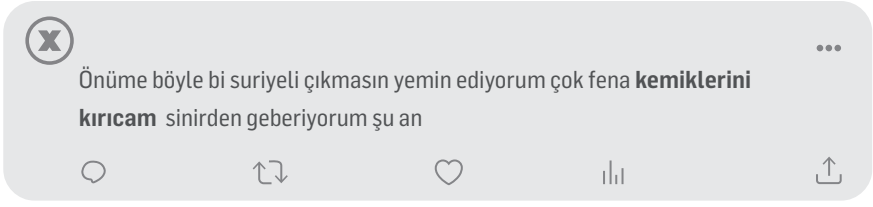
Örnekte geçen Arap baharı ve Arap kışı ifadeleri bir millete değil, toplumsal olaylara yönelik olarak kullanılmıştır. Bu nedenle nefret söylemi içermemektedir.

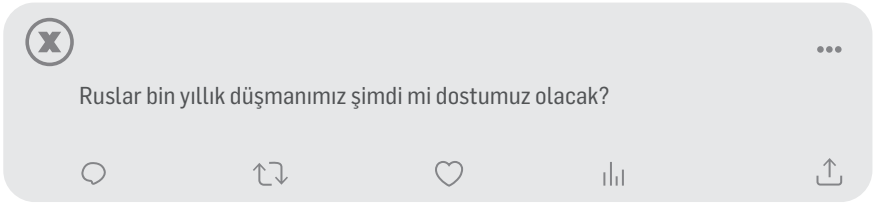
Düşmanlık ve Saldırı Tehdidi: Bir topluluğa karşı yapılabilecek fiziksel ve psikolojik saldırı eylemlerini ve bu eylemleri yapanları meşrulaştıran ya da bu eylemleri yapmayı veya bunların yapılmasını temenni eden ifadelerin bulunduğu nefret söylemi kategorisidir. Ayrıca, bir topluluğa karşı düşmanlık ifade eden veya düşmanlığı

körükleyen ifadeler de bu kategori altında değerlendirilmelidir. Diğer kategorilerde bulunan hakaret, küfür, insandışılaştırma içeren ifadeler de düşmancıl bir eylemdir. Ancak, bu kategoride sadece saldırı eylemlerinin meşrulaştırıldığı veya temenni edildiği tweet örnekleri ele alınmaktadır.

Net örnekler:

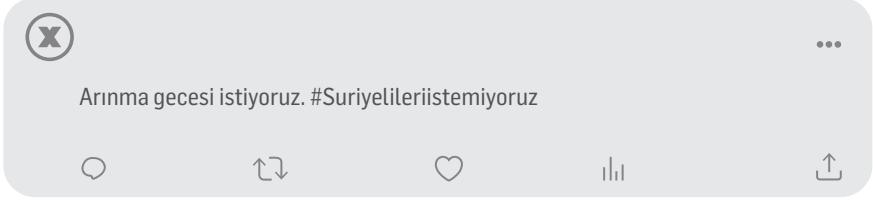






Yukarıdaki ilk iki örnekte net bir şekilde zarar verme tehdidi bulunmaktadır. Son örnekte de bir topluluğa karşı duyulan düşmanlık açık bir şekilde ifade edilmiştir. Bu nedenle bu örnekler "Düşmanlık" kategorisinde değerlendirilmelidir.

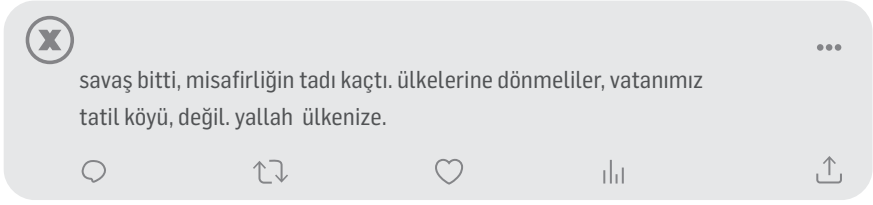
Net olmayan örnek:



Yukarıdaki tweet'te geçen "arınma gecesi" ifadesi konu hakkında bilgisi olmayan birisi tarafından anlaşılamayabilir. Bu nedenle, oluşturulmuş nefret söylemi tes-

pit edilemeyebilir. Ancak, "Arınma Gecesi" insanların bir geceliğine birbirilerini öldürmelerinin veya birbirilerine işkence yapabilmelerinin yasal olduğu bir filmin adıdır. Tweet'in sonundaki hashtag göze alınarak değerlendirildiğinde tweet'i atan kişinin filmin konusu ile benzer şekilde Suriyelilere karşı çeşitli saldırı eylemlerinin meşru kabul edilmesini temenni ettiği anlaşılmaktadır.

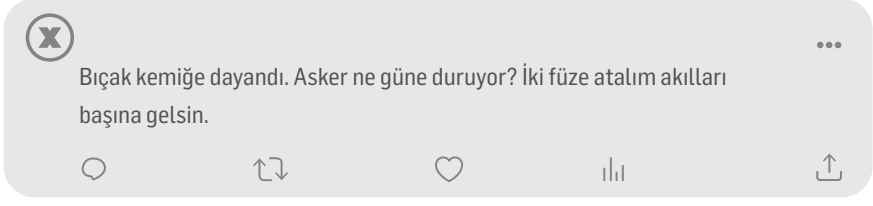
Yanıltıcı örnek:

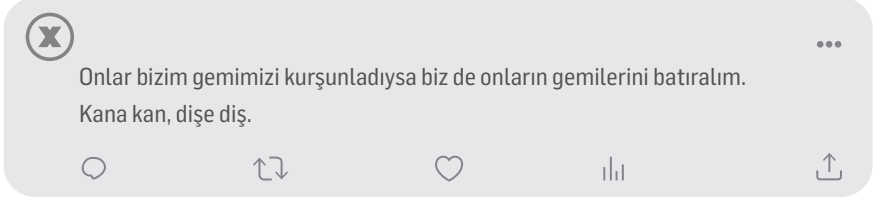


Yukarıdaki örnekte herhangi bir zarar verme isteği açık bir şekilde dile getirilmediği için bu kategoriye girmemektedir. Ancak, "ülkelerine dönmeliler" ifadesi geçici koruma statüsünde bulunan Suriyelilerin yasal statüsüne yönelik olası bir hak ihlaline işaret ettiğinden nefret söylemi olarak değerlendirilmelidir.

Savaş Söylemi: Bir topluluk hakkında savaşı çağrıştıran, savaş çığırtkanlığı yapılan, savaşa varacak askerî müdahaleleri ve mevcut bir savaşı meşrulaştırmaya çalışan ifadelerin kullanıldığı nefret söylemi kategorisidir.

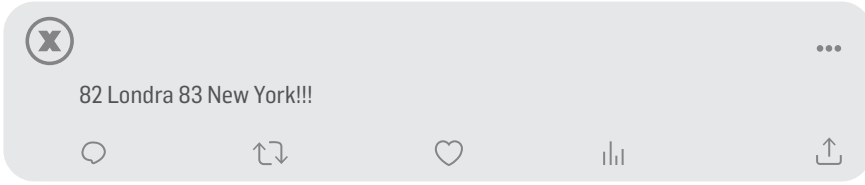
Net örnekler:

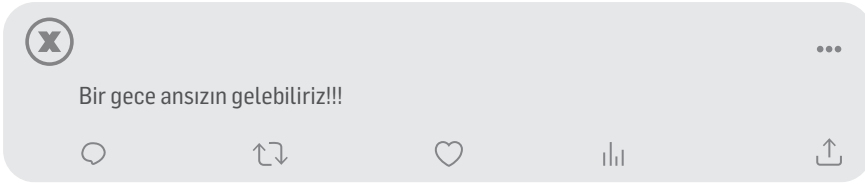


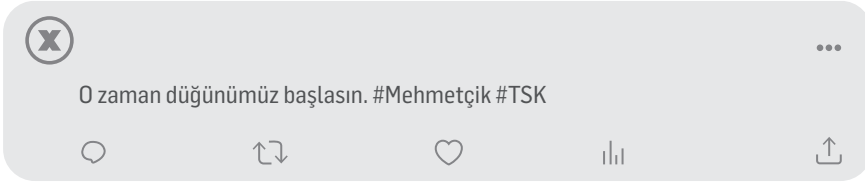


Yukarıdaki örneklerde "füze atalım" ve "gemilerini batıralım" gibi ifadelerle açık bir şekilde savaş çağrısı yapılmaktadır. Ayrıca, "bıçak kemiğe dayandı" ve "kana kan, dişe diş" gibi ifadelerle savaş isteği meşrulaştırılmaya çalışılmıştır.

Net olmayan örnekler:

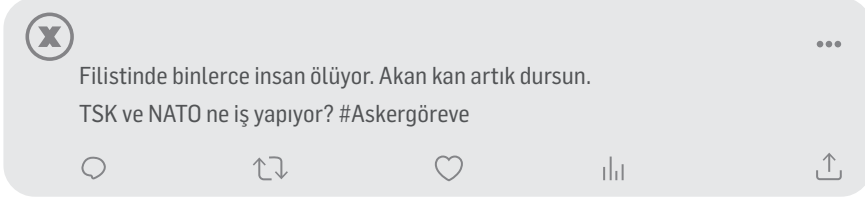






Yukarıdaki örneklerde açık bir şekilde savaş isteği ifade edilmemiştir. Ancak, "82 Londra 83 New York" ve "ansızın gelebiliriz" ifadeleriyle savaş tehdidinde bulunulmaktadır. Ayrıca, son tweet'teki "düğünümüz başlasın" ifadesi, tweet'in sonundaki hashtag'lerle birlikte değerlendirildiğinde, bir savaş çağrısı niteliği taşımaktadır.

Yanıltıcı örnek:



Tweet'te kullanılan "#Askergöreve" hashtagi bir savaş çağrısı olarak algılanabilir. Ancak, "akan kan dursun" ifadesi ile tweet yazarının çatışmaların bitmesine yönelik bir tutum içerisinde bulunduğu anlaşılmaktadır. Bu nedenle, bu tweet'in savaş söylemi olarak kabul edilmemesi gerekmektedir.

2. 4. Zorlayıcı örneklerin değerlendirilmesi

2.4.1. Hashtag ve emoji içeren tweet'ler

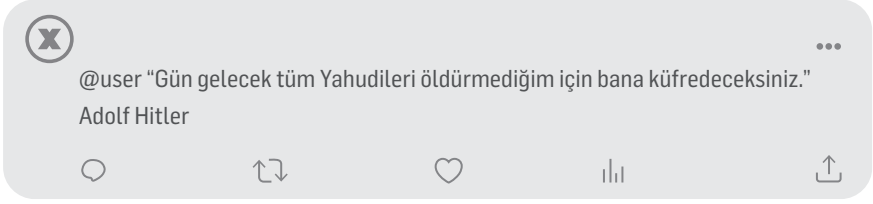
Belirlenen konulara göre tweet çekme aşamasında çeşitli hashtag'ler kullanıldığı için karşımıza çıkan tweet'ler de hashtag içermektedir. Tweet'lerde yapay zekâ modellerinin doğrudan anlamayacağı **hashtag'ler** de kullanılmaktadır. Bazı örneklerde kullanılan hashtag'ler tweet metnindeki fikri destekler nitelikteyken bazı örneklerde tweet metni konudan bağımsızdır ve hashtag sadece tweet'i üst sıralara taşımak için kullanılmıştır. Tweet'ler hashtag'lerle birlikte değerlendirilmelidir. Tweet içeriğinin nefret söylemi içermediği durumlarda, hashtag içeriyorsa ifade nefret söylemi olarak ele alınmalıdır.

Aşağıdaki tweet'lerde de görüldüğü gibi bazı örneklerde çeşitli emoji'ler kullanılmaktadır. Emoji içeren tweet'ler de etiketlemeye dahil edilmiştir. Kullanılan emoji'ler tweet metninden alakasız ya da metindeki fikri destekler nitelikte olabilir. Bu nedenle emoji'lerin kattığı anlamla ilişkili olarak tweet'ler değerlendirilmektedir.

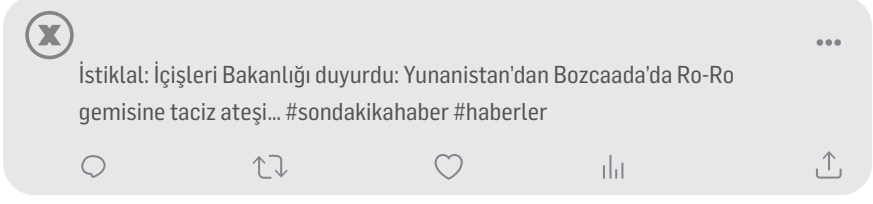
Yukarıdaki tweet örneklerinde nefret söylemi emoji'ler kullanılarak oluşturulmuştur ve emoji'ler göz ardı edilerek değerlendirildiğinde nefret söylemi tespit edilemeyecektir. Yukarıdaki ilk örneğe benzer şekilde toplumsal açıdan hakaret kabul edilen çeşitli el işareti emoji'leri nefret söylemi oluşturmak için yaygın bir biçimde kullanılmaktadır. Ayrıca, ikinci örnekte görüldüğü gibi insandışılaştırma amacıyla tweet'lerde çeşitli hayvan emoji'lerine de yer verilebilmektedir. Ek olarak, çeşitli sebze ve meyve emoji'leri cinsel organları çağrıştıracak şekilde kullanılmakta ve bu yolla hakaret edilebilmektedir. Bu nedenle, etiketleme yapılırken tweet'in içeriğinin ve emoji'lerin bir bütün olarak değerlendirilmesi gerekmektedir.

2.4.2. Başka birinin söylemine yer veren/alıntı yapan tweet'ler

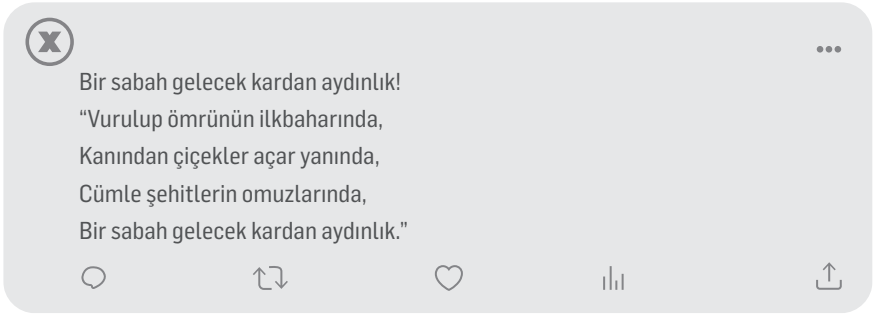
Başka birinin nefret söylemi içeren demecini alıntılayan/aktaran tweet'ler, o söylemi tekrar dolaşıma sokmaktadır ve yayılmasına neden olmaktadır. Bu gibi tweet'ler aşağıdaki örnekte olduğu gibi, eleştirel bir yorum getirilmeden paylaşılıyorsa o tweet **nefret söylemi var** olarak işaretlenmelidir.



Yukarıdaki örnekte Yahudilere yapılan soykırımı meşrulaştıran bir alıntıya yer verilmiştir. Tweet'in tümüne bakıldığında alıntıda geçen sözlerin eleştirilmediği görülmektedir. Tweet'i atan kişi, alıntıyı yeniden dolaşıma sokarak konu dışı bir olayla bağdaştırmıştır. Yahudilere karşı düşmanlık ve saldırı tehdidi içeren bu alıntı tekrar dolaşıma eleştirilmeden sokulduğu için nefret söylemi olarak etiketlenmelidir. Buna karşın, aşağıdaki örnekte alıntılanan tweet'te doğrudan bir kimlik grubu hedef alınmamaktadır ve alıntı yapılan içerikte nefret söylemi bulunmamaktadır:



Bu örnekte olduğu gibi haber değeri taşıyan, nefret söylemi bulunmayan ve alıntı yapılarak aktarılan tweet içeriklerinin etiketlemesi yapılırken genel tutum/duruş bölümünde nötr veya alakasız seçeneği işaretlenmelidir. Bir diğer örnek aşağıda gösterilmiştir:



Tweet içeriği konuyla doğrudan alakalı olmadığı için alakasız seçeneği işaretlenmelidir.

2.4.3. Sarkastik içerikler

Tweet örneklerinde sarkazm (alaycı veya iğneleyici) olduğu düşünülen içerikler incelendiğinde hepsinde sarkazm bulunmadığı anlaşılmıştır. Buna göre bir tweet'in sarkastik içeriğinin olup olmadığına karar verirken sarkazm tanımına, ne gibi ifadelerin bu tanım dahilinde değerlendirilebileceğine dikkat edilmesi gerekmektedir.

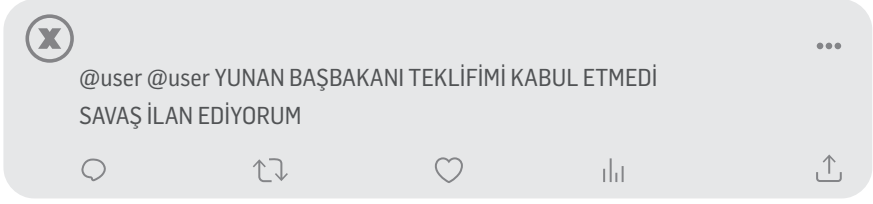
Genelde sarkastik ifadelerde söylenenin tam tersi amaçlanır. Bu durumu tespit etmemizde bize yardımcı olabilecek önemli ipuçlarından biri emoji'lerdir. Emoji'ler, yazılı ifadeyi yazan kişinin ifadeyi yazdığı ruh hâlini ifade etmesine olanak tanır. Kullanılan ifadelerle çelişki içinde olan bir emoji kullanımı, ifadenin sarkastik olduğunu belli edebilir. Örneğin, "Beş milyon mülteci kardeşimizle mutlu bir yıl dilerim." ifadesi, mültecilere iyi dileklerde bulunan bir cümledir. Ancak, bu ifadenin ardından kullanılan ve öfkeyi ifade eden bir emoji, bu cümleyi yazan kişinin mültecilerin varlığından çok rahatsız olduğunu hissettirir. Bu nedenle, bu ifadenin sarkastik bir ifade olduğu söylenebilir. Benzer şekilde, üzücü bir olayı anlatan bir cümleden sonra mutlu veya komik emoji'lerin kullanılması, bu ifadeyi kullanan kişinin olaya önem vermediğini veya küçümsediğini gösterir. Bu nedenle, emoji'lerin kullanımı, sarkastik ifadeleri tespit etmede yol göstericidir.

Bunun yanı sıra, günlük konuşmalarda ifadenin sarkastik olduğunu ima etmek için vurgu veya tonlama kullanılabilir ancak, yazılı metinlerde bu vurguları tespit etmek doğal olarak mümkün değildir. Ancak, farklı yazım tercihleri ve kullanılan işaretler bu konuda yardımcı olabilir. Örneğin, "YENİ YILINIZI BEŞ MİLYON MÜLTECİ KARDEŞİMİZ İLE KUTLARIM!" Bu örnekte, bütün kelimelerin büyük harfle yazılması, bu ifadelerin özellikle vurgulandığını gösterir. Bu nedenle, bu cümleyi yazan kişinin hem ülkedeki mülteci sayısından hem de "kardeş" ifadesinden rahatsız olduğu anlaşılmaktadır. Büyük harflerin kullanımının dışında, bazı

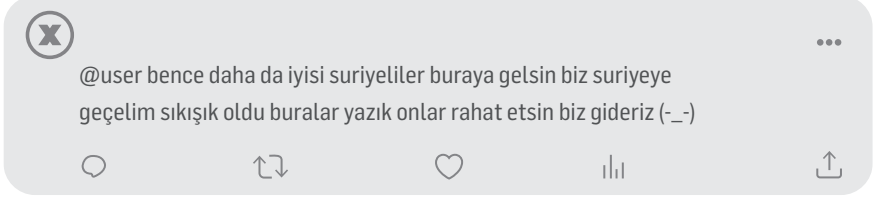
kelimeler tırnak ya da parantez içine alınabilir ve bazı kelimelerin ardından ünlem işareti konulabilir. Bu tercihler, cümleyi yazan kişinin bu kelimeleri vurgulayarak, kullanılan ifadenin tam tersini ifade etmek istediğini göstermektedir.

Ancak, yukarıda bahsedilen ipuçları olmadan da bir cümlenin sarkastik olup olmadığını belirleyebiliriz. Aynı örnekle devam edelim: "Beş milyon mülteci kardeşimizle mutlu bir yıl dilerim." Bu ifade, halkın genel görüşüne ters düştüğü için, bu görüşü benimseyen biri tarafından sarkastik olarak algılanabilir. Ancak, her zaman belirli bir görüşe sahip çoğunluğun tersine görüşlere sahip azınlık bir grup olduğu ihtimali de göz ardı edilmemelidir.

Örnekler:



Yukarıdaki tweet örneğinde "savaş ilan etme" fiilinde birinci tekil şahıs kullanımı ifadenin sarkastik olduğuna işaret ediyor. Tweet'in bağlamı bilinmese de, tweet'i yazan kullanıcının "savaş ilan etme" eylemini alay konusu hâline getirdiği ve bunu sarkastik bir yolla ifade ettiği görülmektedir. Tweet'te geçen savaş söylemi ifadesinden ötürü bu tweet "Yunan karşıtı" olarak değerlendirilmeli ve kategorilerde "düşmanlık/savaş söylemi" olarak etiketlenmelidir.



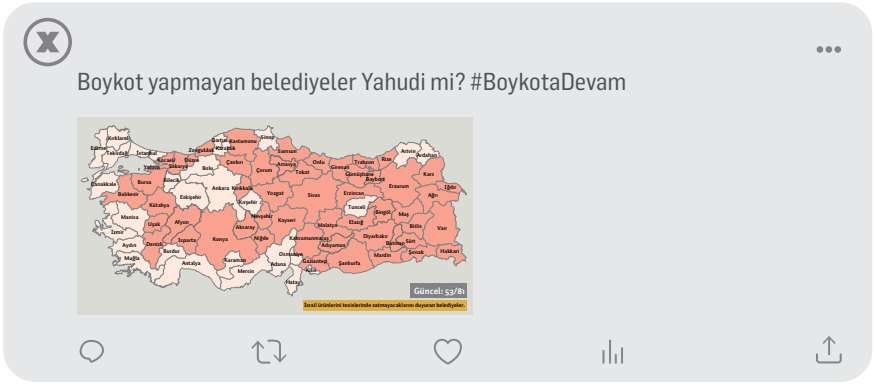
Bu örnekte "sıkışık oldu buralar" ve onlar ve biz karşıtlığıyla ifade edilen durum beklenmedik ve doğal olmayan bir durum olduğu için sarkastik içerik taşıyor. Aynı zamanda tweet'in sonundaki "yazık onlar rahat etsin biz gideriz" ifadesiyle kullanıcının Suriyelilerin günlük hayatlarında haklarına erişiminden rahatsızlık duyduğu anlaşılıyor. Dolayısıyla bu tweet "mülteci karşıtı", "ayrımcı söylem" olarak değerlendirilmelidir.

Sarkastik ifadelere dahil edebileceğimiz bir başka durum ise azınlık grupların çoğunlukla nefret söylemi bağlamında kullanılan kelimeleri dövizlerde, sloganlarda, veya grup içi iletişimlerinde kullanması olabilir. Örneğin, LGBTİ+'lara yönelik ayrımcılık ve hakaret amaçlı kullanılan "dönme" kelimesi Onur Ayı yürüyüş ve etkinliklerinde "Velev Ki Dönmeyiz" gibi sloganlarda kullanılarak bu kelime grup tarafından benimsenmiş ve asıl amacına ters bir şekilde kullanılmaya başlanmıştır. Bu tarz durumlarda genellikle nefret söylemi bağlamında kullanılan kelimelerin azınlık gruplarınca "geri kazanılması" sosyal medyada sık rastlanan bir durum olduğundan bu metodoloji kılavuzunda da yer alması önemlidir. Benzer şekilde yine LGBTİ+'lar sosyal medyada birbirleriyle şaka, ironi veya gönderme amaçlı, normal bağlamda nefret söylemi sayılacak kelime ve ifadeler kullanarak konuşabilir. Etiketleme esnasında bu gibi ifadelerin olduğu tweet'lerle karşılaşıldığında, eğer tweet bağlamından ifadelerin hakaret veya aşağılama içerdiği anlaşılıyorsa tweet'ler nefret söylemi olarak değerlendirilmelidir. Ancak, bağlamdan içeriğin anlaşılmadığı veya ifadelerin "geri kazanma" amaçlı kullanıldığı durumlarda, ifadelerin kendisi aracın doğru eğitilmesi açısından tetikleyici kelime veya küfür/hakaret olarak seçilmeli, nefret söylemi olarak değerlendirilmemelidir.

Her halükârda sosyal medyadaki ifadelerin her zaman direkt anlamlarına karşılık gelmeyeceğinin farkında olmak, nefret söylemi ile ilgili geniş ve kapsayıcı bir anlayışa sahip olmamız için önemlidir.

2.4.4. Örtülü nefret söylemi

Tweet'lerin içeriğinde kullanılan her ifade açıkça nefret söylemi içermiyor olabilir. İlk bakışta tarafsız gözüken ifadelerin kullanıldığı ancak biraraya geldiğinde okuyucunun kültürel bağlam bilgisi ile beraber nefret söylemi olacak ifadeler örtülü nefret söylemi olarak kategorize edilmektedir. Açıkça yapılan nefret söylemi kadar önemli ve zarar vericidir. Bu nedenle bu tarz tweet'lerin nefret söylemi olarak etiketlenmesi önemlidir. Aşağıda örtülü nefret söylemi için X'den bir örnek verilmiştir:



Yukarıda verilen örnekte tweet bir fotoğrafla desteklenmektedir ve bahsi geçen fotoğraf İsrail ürünlerini boykot eden ve etmeyen belediyeleri renklerle ayırıp göstermektedir. İlk bakışta yazılan cümle açık bir hakaret ya da kemikleşmiş saldırgan/ayrımcı ifadeler kullanılmadığı için nefret söylemi içermiyor olarak gözükebilir. Ancak, bağlam bilgisi ve kullanılan etnik sözcük nefret söylemine işaret etmektedir. "Yahudi" kimliğinin İsrail devletinin eylemleriyle bir tutulmasına ek olarak aynı kimlik simgeleştirilerek aşağılama unsuru olarak kullanılmıştır. Bu nedenle tweet örtük bir şekilde nefret söylemi içermektedir ve ona göre işaretleme yapılmalıdır.[6]

2. 5. Ek etiketleme başlıkları

Son bölümde, projedeki arayüzde kullanılan diğer etiketleme başlıkları bir arada sunulmuştur. Uzun bir planlama sürecinin ardından oluşturulan bu başlıklar, etiketlemeyi daha ayrıntılı hâle getirmek ve aracı daha verimli bir şekilde eğitmek amacıyla geliştirilmiştir. Bu başlıkların, gelecekteki çalışmalar için de bir başlangıç noktası oluşturması hedeflenmektedir.

2.5.1. Tweet dili

Bu kısımda "Türkçe" ve "Türkçe değil" seçenekleri bulunmaktadır. Türkçe'den farklı dilde olan tweet'ler için "Türkçe değil" seçeneği işaretlenerek **diğer kısımları etiketlemeden** tweet'in etiketlenmesi tamamlanmalıdır. Farklı dildeki tweet'lerde içerik anlaşılsa bile bu şekilde ilerlenmelidir.

6 Bu proje kapsamında fotoğraflı içerikler etiketlenmemiştir. Tweet örtülü nefret söylemine örnek olması adına verilmiştir.

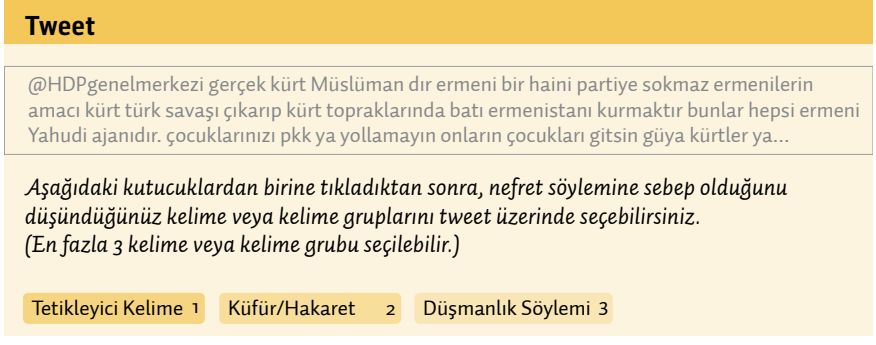
Türkçe tweet'ler içerisinde geçen ve kişinin farklı dillerden kelimeleri bir arada kullandığı durumlarda "Türkçe değil" seçeneği seçilmelidir. Bu bağlamda internet dilinde sıkça kullanılan "LOL (laughing out loud)" veya "OMG (oh my god)" gibi İngilizce kısaltmaların bulunduğu tweet'ler de "Türkçe değil" olarak seçilmelidir.

Bunun yanında, Türkçe dışındaki dillerde yazılı hashtag'lere sahip tweet'ler de, tweet Türkçe olsa bile, "Türkçe değil" şeklinde etiketlenmelidir. Bu durum oluşturulacak dijital aracın daha doğru yorumlama yapabilmesi için teknik sebeplerle tercih edilmiştir. Tweet metni Türkçe yazılmasına rağmen görsel içerikte Türkçe olmayan bir metin yer alıyorsa, "Türkçe değil" seçeneği işaretlenmelidir.

2.5.2. Nefret söylemi içeren bölümün belirtilmesi

Hem etiketleyen kişilerin yönlendirilmesi hem de farklı yapay zekâ yöntemleriyle nefret söylemi tespit algoritmalarının kullanımı için, nefret söylemi içeren mesajlarda ilgili kelimelerin işaretlenmesi de beklenmektedir. Bu doğrultuda şunlara dikkat edilmelidir:

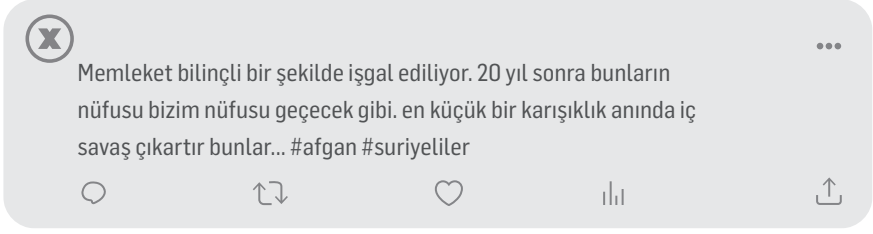
- Tweet'teki söylemi değerlendirirken **okuyucu için tetikleyici olan nefret söylemi ya da ayrımcı söylem** içeren kelime ya da kelime grupları seçilmelidir. Örneğin, uzun bir tweet'te tetikleyici olduğu düşünülen ifadelerden öne çıkanları işaretleyerek en fazla üç işaretleme yapmaya dikkat edilmelidir. Tetikleyici olduğu düşünülen ifadelerden bazıları düşmanlık söylemi kategorisine dahil olabilirken, bazıları küfür/hakaret kategorisine dahil olabilir. Her iki kategoriye ait nefret söylemi bulunan tweet'lerde, görselde görülen kategori başlıklarına tıklayarak tüm ilgili ifadeler seçilmelidir.

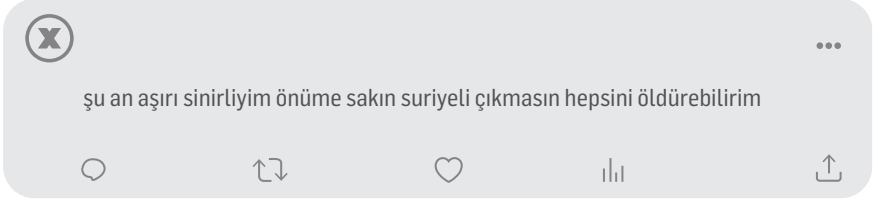


2.5.3. Nefret söylemi derecesi

Tweet'leri etiketlerken nefret söylemi derecesi seçilmelidir. **1-10** arasında bir skalada, 1 en düşük 10 en yüksek olacak şekilde, işaretleme yapılmalıdır. Nefret söylemi olmayan tweet'lerde derece 0 olarak işaretlenmelidir. İşaretleme yaparken, nefret söylemi dereceleriyle, nefret söylemi kategorileri arasında şiddet üzerinden bir ilişki kurmamak amacıyla, derece ve kategori işaretlemeleri birbirinden bağımsız ele alınmalıdır. Derecelendirme yaparken dikkat edilmesi gereken, belirtildiği gibi kategoriye göre seçim yapmaktan ziyade, kullanılan dilin içeriğidir. Derece nefret söylemi içeren sözcüklerin yoğunluğu ve sıklığı ile doğru orantılı olmalıdır.

Örnekler:





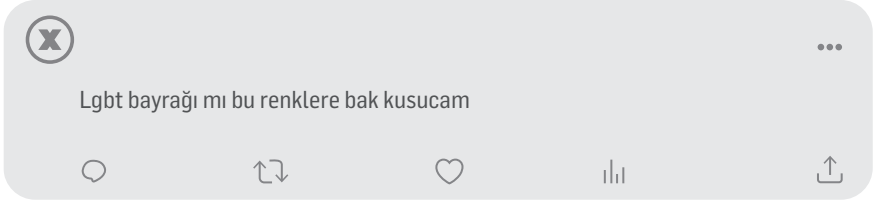
İki örnek de nefret söylemi içermektedir. Ancak, içeriklerinin yoğunluğu aynı seviyede olmadığı için farklı derecelendirme yapılmalıdır. İkinci tweet'te kullanılan açık fiziksel tehdit nefret söyleminin şiddetini birinci tweet'e göre arttırmaktadır. Bu nedenle iki tweet'in derecesi farklı olacaktır.

Son olarak, nefret söylemi var/yok seçeneğinde "Emin Değilim" işaretlendiyse burada da "Emin Değilim" seçeneği işaretlenmelidir.

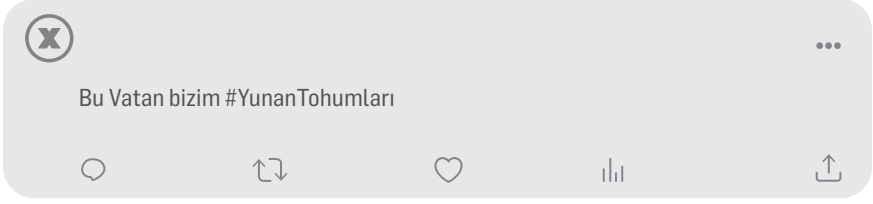
2.5.4. Saldırgan dil

Bu aşamada incelediğimiz tweet'i saldırı tehdidi ve/veya temennisi üzerinden değerlendirmemiz gerekiyor. Tweet'in içeriğinde saldırganlık taşıyan herhangi bir ifade yoksa "Yok" seçeneğini seçmeliyiz. İçerikteki saldırı tehdidi ya da temennisinin "Zayıf" mı yoksa "Şiddetli" mi olduğuna karar verirken söz konusu tehdidin büyüklüğünü ve

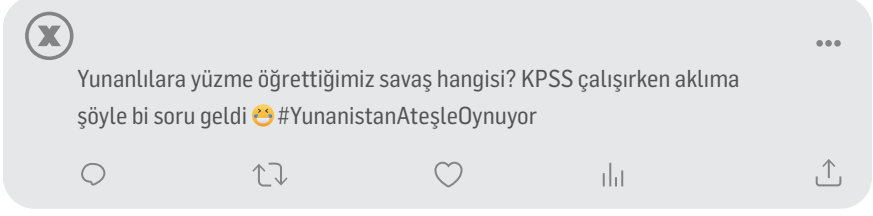
etki derecesini düşünebiliriz. Bu yüzden etiketlemenin nefret söylemi kategorisi ve şiddeti kısımlarıyla ilişkili değerlendirilmemesi bu kategori özelinde değerlendirilmesi tutarlı bir tespit yapabilmek adına önemlidir.



Bu örnekte LGBTİ+'ların kullanılan olumsuz ifadelerle hedef alındığını tespit edebiliyoruz. Ancak, LGBTİ+'lara yönelik saldırganlık içeriyor mu sorusunu düşündüğümüzde hedef gruba yönelik saldırı tehdidi ya da temennisi bulunmadığını, bu yüzden tweet'te saldırgan dil bulunmadığını söyleyebiliriz. Etiketleme yaparken bu ve benzer örnekler için "Yok" seçeneğini seçmeliyiz.

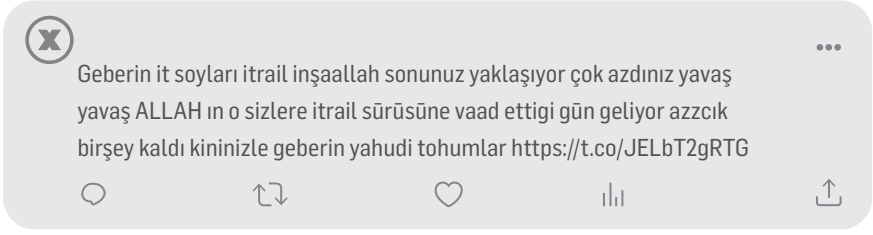


Bu örnekte Yunan kimliği simgeleştirme yoluyla hedef gösteriliyor. Tweet'in kime cevap olarak yazıldığını göremiyoruz ancak, "Bu Vatan bizim" ifadesinin hashtag'deki ifade dolayısıyla belli bir kimlik grubuna söylendiğini anlayabiliyoruz. Bu ifadede Yunan kimlik ismi aşağılama amacıyla kullanılıyor. Tweet'teki saldırgan dili değerlendirdiğimizde saldırı tehdidi ya da temennisi yer almadığı için saldırgan dil yok şeklinde etiketlemeliyiz.



Bu örnekteyse kullanılan hashtag ile ilişkili olarak tweet'in kalan kısmını değerlendirmemiz gerekiyor. Hashtag ve onun dışında kalan ifadeyi birlikte düşündüğümüzde, tarihî olaylara atıfta bulunulup nefret söyleminin desteklendiğini düşünebiliriz.

Bu hashtag'de kullanılan "ateşle oynamak" deyimi hedef alınan gruba yönelik bir tehdit ifade ettiği için bu tweet'te saldırgan dil bulunduğunu söyleyebiliriz. Ancak, bu tweet'te saldırgan dil için "Zayıf" seçeneğini işaretlemeliyiz. Her ne kadar saldırı tehdidi tespit etsek de, içerikte yer aldığı kadarıyla bu tehdidin planlı, zarar verme etkisi açısından büyük olduğunu söyleyemeyiz. Bu nedenle "Şiddetli" seçeneğini elemeliyiz.



Bu örnekteyse hedef alınan birden fazla grup olduğunu, Yahudilerin simgeleştirme yoluyla, İsraillilerin ise küfür ve hakaret içerikleriyle hedef alındığını söyleyebiliriz. İçerikteki ilk kelime ve devamındaki şiddet temennileri dolayısıyla bu tweet'teki saldırgan dili de "Şiddetli" olarak işaretlemeliyiz.

2 YAPAY ZEKÂ MODELİNİN GELİŞTİRİLMESİ

Nefret söylemi ile mücadeledeki zorlukları ele almak amacıyla, bir önceki bölümde yer alan yönergeler ışığında, Türkçe tweet içeriklerinde nefret söylemi tespiti yapabilmek için yapay zekâ tabanlı bir araç geliştirdik. Bu araç yalnızca nefret söylemini tespit etmekle kalmıyor, aynı zamanda yoğunluk derecesini değerlendiriyor, kategorilere ayırıyor ve metin içindeki konumunu saptıyor. Aracımız, Arapça tweet içeriklerinde ve Türkçe yazılı basında nefret söylemi tespiti yapmak için tasarlanmış modellerle daha da geliştirildi.[7] Bu aracın bir diğer önemli özelliği de X platformu üzerinde belirli periyotlarda gerçek zamanlı olarak izleme yapabilmesi, ilgili tweet içeriklerini derleyebilmesi ve yapay zekâ modellerimizi kullanarak bu içerikleri otomatik olarak etiketleyebilmesidir.

7 Hrant Dink Vakfı tarafından yürütülen ve 10 yılı aşkın bir süredir yazılı basına odaklanan Medyada Nefret Söyleminin İzlenmesi Projesi'nin verileri bir örneklem veri havuzu oluşturmak için kullanıldı.

1. VERİ TOPLAMA VE ETİKETLEME

Nefret söylemi tespit aracı, makine öğrenme yaklaşımı kullanılarak geliştirildi. Makine öğrenme araçlarının geliştirilebilmesi için, modellerin hem eğitilmesini hem de test edilmesini sağlayacak etiketlenmiş verilere (annotated data) ihtiyaç vardır. Bizim çalışmamız özelinde bu süreç, farklı seviyelerde ve kategorilerde nefret söylemi içeren tweet içeriklerinin ve aynı zamanda nefret söylemi içermeyen tweet içeriklerinin toplanmasını gerektiriyor. Tweet'ler, bağlama dair sınırlı bilgi içeren kısa metinlerdir. Genellikle kısaltmalar, yazım hataları ve gramer hataları dahil olmak üzere standart dışı bir dil kullanımı söz konusudur.

Bu çalışma kapsamında, Ortadoğu ve Kuzey Afrika (ODKA) bölgesinden STK'ların ve araştırmacıların ihtiyaçlarına yanıt verebilmek amacıyla, bu bölgeden anadili Arapça olan uzmanlara danışarak, Arapça dilindeki tweetler için veri toplama ve etiketleme süreçleri belirledik ve geliştirdik. Bu süreçte, nefret söylemi ağı platformumuzda yer alan uzmanlara, proje faaliyetlerine katılmış anadili Arapça olan etiketleyicilere (annotators) ve projenin başlangıç aşamasında işbirliği geliştirdiğimiz açılış etkinliği katılımcılarına danıştık.

Aşağıdaki bölümlerde veri toplama ve etiketleme süreçlerine dair ayrıntılı bilgiler yer alıyor.

1.1. Veri toplama

X platformunun akademik API ve veri kazıma (*scraping*) yöntemleri kullanılarak, belli dönemlerde, belli anahtar kelimeler ve hashtag'lerle paylaşılan tweet içerikleri indirildi. Söz konusu anahtar kelimeler ve hashtag'ler, güncel olayların düzenli takibi yapılarak ve Türkiye'de nefret söylemine sıklıkla maruz kalan gruplar dahil edilerek seçildi. Buradan hareketle, Türkçe dilindeki tweet'lerde, nefret söylemi ile en fazla hedeflenen dokuz grup (Alevi, Arap, Ermeni, Yunan, Yahudi, Kürt, LGBTİ+, mülteciler [Arapça], mülteciler [Türkçe]) ile ilgili tweet içeriklerine erişim sağlandı.

Bu noktada, bu proje ile paralel bir şekilde, Hrant Dink Vakfı tarafından yürütülmekte olan nefret söylemi izleme faaliyetlerinde hedef gruplarla ilgili herhangi bir sınırlamanın olmadığını not düşmek gerekir. Yıllardır devam eden söz konusu izleme faaliyetleri neticesinde, 100'ü aşkın grup veya kimliğin yazılı basında nefret söyleminin hedefi olduğu tespit edildi. Ancak, bu proje özelinde, veri havuzumuzu nefret söylemi içeren tweet'lere en sık maruz kalan dokuz grup üzerinde çalışacak şekilde daralttık. Bu durum, belirli anahtar kelimeler ve hashtag'ler üzerinden daha etkin veri toplayabilmemizi sağladı ve etiketleme süreçlerini kolaylaştırdı. Bununla

birlikte, algoritmamızın diğer grupları hedef alan nefret söylemini de tespit edecek şekilde genellenebilir olmasını hedefledik.

Yukarıda belirtilen kriterler ışığında, toplam 16.254 tweet etiketi toplandı. Toplam tweet etiketi sayısı ve nefret söylemi konu başlığına göre toplam tweet sayısı Tablo 1'de yer alıyor. "Mülteciler (Arapça)" haricindeki tüm konu başlıkları Türkçe tweet'lerden oluşurken, "Mülteciler (Arapça)" başlığı altındaki tüm tweet'ler ise sadece Arapça dilinde paylaşılan tweet'leri içeriyor. Bazı tweet'lerin birden fazla grubu hedef alacak şekilde nefret söylemi içermesi nedeniyle, aynı tweet'i farklı hedef grupları altında değerlendirmek durumunda kaldık.

Tablo 1. Nefret söylemi konusuna göre toplam etiket ve indirilen tweet sayısı

Konu Başlığı	En Az 3 Etiketleyici Tarafından Etiketlenmiş Tweet Sayısı	Erişim Sağlanan Tweet Sayısı
Yahudi	3720	8200
Yunan	2418	19500
Mülteci (Türkçe)	2289	4350
Mülteci (Arapça)	2999	5750
Alevi	1000	5650
Ermeni	979	3300
Arap	1005	7550
Kürt	947	18500
LGBTİ+	897	1350
Toplam	**16254**	**74150**

1.2. Veri etiketleme

Tweet içeriklerinin manuel bir şekilde etiketlenmesi, bilhassa nefret söylemine ilişkin etiketlemelerin manuel olarak yapılması, bu işlemin öznelliği, tweet'lerin kendine özgü doğası ve bağlama olan bağımlılığı nedeniyle zorlu bir görevdir. Etiketleyiciler belli bir hedef kitleye yönelik küfür veya tehdit ifadeleri gibi müstehcen dil içeren bağlamın nefret söylemi olarak etiketlenmesi konusunda genelde hemfikir olabilseler de, daha ince nüanslar içeren ayrımcı söylemin (örneğin: "Mülteciler hükümet desteği almamalıdır") ne şekilde sınıflandırılacağına dair genelde görüş ayrılığına düşebiliyorlar. Araştırmacılar bu durumu yönetebilmek amacıyla, genelde etiketleyiciler arasında uyuşmazlık bulunan örnekleri göz ardı

etme eğiliminde oluyorlar. Bu durum da veri kaybına ve modelin aşırı iyimser sonuçlar vermesine yol açabiliyor. Etiketleyici uyuşmazlıklarını ele almanın ve veri kalitesini artırmanın bir diğer yolu da ikinci bir etiketleme süreci işletilerek etiketleyicilerin bir uzlaşmaya varmasını sağlamak olabilir.

Yüksek kaliteli bir veri kümesi elde edebilmek amacıyla, bilgisayar bilimcileri, dilbilimciler, sosyal bilimciler ve sivil toplum uzmanlarından oluşan ekibimiz, yakın bir işbirliği içinde çalışarak, nefret söylemi içeren tweet'lerin etiketlenmesi için bir dizi yönerge geliştirdi (bkz. Kılavuz bölümü). Etiketleme yönergelerini (*annotator guidelines*) geliştirirken yinelemeli bir yaklaşım benimsedik ve veri etiketleme sürecindeki belirsizlikleri ve çelişkileri giderebilmek için bu yönergelere ince ayar yaptık. Örneğin, kafa karışıklığını ortadan kaldırabilmek amacıyla her bir nefret söylemi kategorisine daha fazla örnek ekledik. Bir tweet'in birden fazla gruba yönelik nefret söylemi içerdiği ya da örtülü nefret söylemi içerdiği (bir kişi ancak bağlamı biliyorsa tweet'in nefret söylemi içerdiğini tespit edebilir) belirsiz durumlarda, nasıl hareket edilebileceğine açıklık getirdik. Veri etiketleme süreçlerinde, nefret söylemine farklı yaklaşımları yansıtabilmek ve nefret söyleminin yöneldiği hedef grupları temsil edebilmek için farklı arkaplanlara sahip etiketleyicilerle çalışmak önemlidir.

Bu süreç sonunda ortaya çıkan yönergeler bir hayli kapsamlı olup nefret söylemi kategorisi, hedef grup ve –yapay zekâ modeli üzerindeki etkisini de değerlendirmek için kategorisinden bağımsız olarak– algılanan nefret söylemi derecesi için etiketler içeriyor.

Nefret söylemi kategorisi altında etiketleme yapabilmek için her bir kategorinin kapsayıcılığı ve kapsamı üzerine detaylı tartışmalar yürüttük. Sonuç olarak, "nefret söylemi yok" kategorisine ek olarak dört spesifik nefret söylemi kategorisi belirledik:

- Simgeleştirme,
- Abartma/Genelleme/Yükleme/Çarpıtma,
- Küfür/Hakaret/Aşağılama/İnsandışılaştırma,
- Düşmanlık/Savaş/Saldırı/Öldürme/Yaralama Tehdidi.

Nefret söylemi derecesini tayin etmek için 0'dan 10'a kadar farklı nefret söylemi seviyeleri tanımladık. Nefret söyleminin olmadığı durumlar 0 ile temsil edilirken, nefret söylemi yoğunluğu arttıkça seviyeler de yükseliyor.

Tablo 2'de nefret söylemi kategorilerine ilişkin açıklamalar yer alıyor. Bu kategoriler, Hrant Dink Vakfı'nın Türkçe yazılı basında nefret söylemini izlemek için kullandığı mevcut kategoriler üzerine inşa edildi.[8] Buradaki yaklaşımı iyileştir-

8 Bkz. Medyada Nefret Söylemi ve Ayrımcı Söylem 2019 raporu: https://hrantdink.org/tr/asulis/yayinlar

mek ve sosyal medya mecrası için de güncelliğini ve geçerliliğini korumasını sağlamak amacıyla, proje süresince her üç kuruluştan da araştırmacılar kendi katkılarını sundular.

Tablo 2. Nefret söylemi kategorilerine ilişkin açıklamalar

Nefret Söylemi Kategorisi	**Açıklama**
Simgeleştirme	Doğal bir kimlik öğesinin hakaret, nefret ve aşağılama unsuru olarak kullanıldığı ve kimliğin bu şekilde simgeleştirildiği söylemler
Abartma/ Genelleme/ Yükleme/ Çarpıtma	Belli bir olaydan, durumdan veya eylemden genel çıkarımlar yapan, gerçek verileri manipüle eden veya münferit olayları bir kimliğin tamamına atfeden söylemler
Küfür/Hakaret/ Aşağılama/ İnsandışılaştırma	Bir topluluk hakkında doğrudan küfür, aşağılama, hakaret içeren söylemler veya bir topluluğu genelde insan dışı varlıklarla ilişkilendirilen eylemler veya niteliklerle tanımlayan söylemler
Düşmanlık/Savaş/ Saldırı/Öldürme/ Yaralama Tehdidi	Düşmanca, savaşı çağrıştıran veya söz konusu kimliğe zarar verme arzusunu dile getiren ifadelerin yer aldığı söylemler

Proje kapsamında, çoğunlukla medya çalışmaları ve sosyoloji gibi çeşitli alanlardan üniversite öğrencilerinden oluşan bir etiketleyici ekibi oluşturuldu. Etiketleyici seçim süreci, konuya duyulan ilgi ve özgeçmiş incelemelerine dayanarak yürütüldü. Veri etiketleme süreci başlamadan önce, etiketleyicilere Hrant Dink Vakfı proje ekibi tarafından projenin metodolojisi üzerine kapsamlı bir eğitim verildi. Bu eğitim esnasında, etiketleme yönergelerimize genel bir giriş yapıldı ve her bir nefret söylemi kategorisi ve seviyesine ilişkin birkaç tweet örneği incelendi. Tweet'ler, veri etiketleme işlemi için, her biri elli tweet içeren partilere ayrıldı. Sonrasında bu tweet'ler etiketleme sunucusuna yüklendi; burada etiketleyiciler etiketleme işlemi için Label Studio arayüzünü kullandı. Etiketlerin kalitesini sağlama almak için her tweet üç farklı etiketleyici tarafından etiketlendi; bir başka deyişle, her bir parti, üç ayrı bağlantı noktasında bulunan üç kişi tarafından etiketlendi. "Mülteciler (Arapça)" konu başlığındaki gibi etiket sayısının yetersiz olduğu durumlarda, yalnızca bir veya iki etiketleyici tarafından etiketlenmiş tweet'ler de model eğitiminde kullanıldı. Tek bir tweet içeriğiyle birden fazla grubun hedef alınabilmesi nedeniyle "Hedef Grup" ve 'Nefret Söylemi Kategorisi' etiketleri için birden fazla seçime izin verildi. Böyle durumlarda, etiketleyici oylarını seçilen gruplar veya kategoriler arasında paylaştırma yoluna gittik.

Birden fazla etiketleyici tarafından etiketlenen tweet'ler için etiketleyici etiketlerinin tutarlılığını değerlendirmek amacıyla Krippendorff'un alfa katsayısı yöntemini kullandık. Bu katsayı -1 ile 1 aralığında değişmekte olup 1 tam uzlaşmayı, -1 tam uyuşmazlığı, 0 ise etiketleyici seçimleri arasında rastgele korelasyonu gösteriyor. 0,33 ile 0,67 arasındaki değerler orta derecede güvenilir (orta derecede uzlaşma) olarak kabul edilirken, 0,67'nin üzerindeki değerler yüksek derecede güvenilir (yüksek uzlaşma) olarak değerlendiriliyor. Nefret söylemi konu başlıklarına göre Krippendorff'un alfa katsayısı değerleri Tablo 3'te sunuluyor. Örneğin, hedeflenen grup Yahudi olduğunda (en üst satır) saldırgan bir dil kullanılıp kullanılmadığı konusundaki uzlaşma oranı 0,329 katsayısı ile temsil edilirken, bu da orta düzeyde bir uzlaşmaya işaret ediyor. Ancak, nefret söylemi gücü ve kategorisi sütunlarındaki katsayılar ise daha küçüktür. Bu durum, nefret söyleminin öznel doğasından ve veri etiketleme görevinin gönüllülük esasına dayanmasından kaynaklanıyor. Etiketleyici değişim oranının yüksek olması, çok sayıda etiketleyici olması (her bir etiketleyici tarafından daha az sayıda tweet'in etiketlenmesi anlamına gelir) ve nefret söyleminin farklı şekillerde yorumlanabilmesi gibi faktörler uzlaşma oranlarının daha düşük olmasına yol açabiliyor. Bu durum, bu raporun "Çalışmanın Kısıtları" bölümünde daha ayrıntılı olarak inceleniyor.

Tablo 3. Nefret söylemi konu başlıklarına göre Krippendorff'un alfa katsayısı değerleri

Konu Başlığı	**Genel Tutum ve Duruş**	**Nefret Söyleminin Gücü**	**Saldırgan Dil**	**Hedef Grup**	**Nefret Söylemi Kategorisi**
Yahudi	0,31	0,176	0,326	0,268	0,197
Yunan	0,294	0,176	0,341	0,283	0,296
Mülteciler (Türkçe)	0,502	0,064	0,296	0,416	0,281
Mülteciler (Arapça)	0,522	0,213	0,186	0,158	0,198
Alevi	0,013	0,055	0,386	0,237	0,075
Ermeni	0,284	0,068	0,343	0,179	0,143
Arap	0,306	0,158	0,327	0,377	0,211
Kürt	0,214	0,147	0,368	0,329	0,348
LGBTİ+	0,191	0,158	0,229	0,252	0,346

Bu verilerin toplanması ve etiketlenmesi için harcanan tüm çabaya rağmen, etiketleme süreçlerinin gelecek çalışmalarda daha da iyileştirilebilmesi mümkündür: i) yukarıda belirtildiği gibi, tweet'lerin çoğunluğu üç etiketleyici tarafından etiketlendirilirken, geri kalanı bir veya iki kişi tarafından etiketlendirilmiş olup üç etiketleyiciye tamamlanabilir ve ii) birden fazla etiketleyicinin süreçte yer aldığı durumlarda, etiketleyiciler arası uzlaşmazlıklar açıklıkla ele alınmamakta, daha ziyade kullanıcıların kararına bırakılmaktadır. Bununla birlikte, çoklu etiketlemeleri [*multiple annotations*] birleştirmek için farklı stratejiler öneriyoruz (örneğin ortalama oylama, çoğunluk oylaması veya ağırlıklı çoğunluk oylaması) ve bu yaklaşımları yapay zekâ sisteminin genel performansı üzerindeki etkileri açısından değerlendiriyoruz.

2. NEFRET SÖYLEMİNİ TESPİT ETMEK VE ÖLÇMEK İÇİN GELİŞTİRİLEN YAPAY ZEKÂ ARACI

Nefret söylemini işleyebilecek sağlam yapay zekâ modelleri geliştirmek amacıyla, metinleri anlama ve işleme becerisiyle bilinen son teknoloji bir dil modeli olan BERT modelini[9] kullandık. BERT modelleri, büyük veri kümeleri üzerinde eğitilmiş ve belirli görevlere uyarlanabilen (ince ayar yapılabilen) makine öğrenimi modelleridir. Türkçe dilindeki içerikler için, Türkçe web verileri üzerinden eğitilmiş bir versiyon olan BERTurk'ü[10] kullanırken, Arapça için Arapça diline uyarlanmış benzer bir model kullandık.

Bu bölümde, nefret söylemi tespiti sürecinin farklı aşamaları ve boyutları için kullanılan modeller hakkında açıklamalar sunuluyor. Buna ek olarak, veri ön işleme adımları ve medya izleme süreci açıklanıyor.

2.1. Veri ön işleme ve sözel olmayan (paralinguistik) öğeler

Sosyal medya gönderileri genellikle resmi olmayan bir tonda yazılır ve metnin semantik yapısını kavramayı zorlaştıran bahsetmeler [*mentions*], bağlantı linkleri (URL'ler) ve sözel olmayan öğeler (örneğin, emoji'ler ve hashtag'ler) içerir. Bu nedenle, söz konusu unsurların bir kısmını ortadan kaldırmak ve dilsel varyansı azaltmak amacıyla bu tür metinsel veriler yaygın olarak ön işleme sürecinden geçirilir. Biz de bu stratejiyi benimseyerek tweet'leri makine öğrenimi modellerinde kullanmadan önce ön işlemeye tabi tuttuk. Bu süreçte, bağlantı linkleri (URL'ler) ve kullanıcı adları genellikle

9 Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. "BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding." In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1*, 4171–86.

10 Schweter, Stefan. 2020. "Berturk - BERT Models for Turkish." *Zenodo*, 27 Nisan 2020. https://doi.org/10.5281/zenodo.3770924.

nefret söylemi tespiti veya sınıflandırması bakımından faydalı bilgiler sağlamadığından bunları kaldırdık. Bağlantı linkleri (URL'ler) ve kullanıcı adlarının kaldırılması kararının olumlu bir yönü de oldu. Bu sayede, nefret söylemi içeren gönderiler paylaşan bireylerin doğrudan hedef alınmamasını, böylelikle kullanıcı mahremiyetinin korunmasını ve hassas verilerin etik bir şekilde ele alınmasını sağladık.

Emojiler, duygu ve düşünceleri kısa yoldan anlatmak için kullanılan Unicode grafik sembolleridir. Grafik emoji'ler gündelik diyaloglarımızın ayrılmaz bir parçası hâline geldi. Örneğin, başparmak yukarı / başparmak aşağı emoji'si, bir konuşmacının tek bir kelime dahi telaffuz etmeden bir konu hakkındaki onayını veya görüş ayrılığını gösterebiliyor. Sosyal medya ortamında hashtag kullanımı çok önemlidir, zira hashtag'ler mesajları belirli bir konu etrafında birbirine bağlamayı sağlarlar. Dil işleme çalışmalarında, genellikle sonraki adımda modellemeyi sadeleştirmek amacıyla, hashtag'ler ön işleme esnasında kaldırılırlar. Ancak, hashtag'ler bazen bir cümlenin ortasında kelime olarak kullanılır ve bunların çıkarılması tüm cümlenin anlamını yok eder. Bu sebeple, sözel olmayan öğelerin (emoji'ler ve hashtag'ler) nefret söylemini tespit etme performansı üzerindeki etkisini inceleyebilmek amacıyla, bu öğelerin tweet metni içinde muhafaza edildiği ya da tweet metninden çıkarıldığı iki farklı senaryo için farklı makine öğrenimi modelleri oluşturduk. Bu şekilde, tweet'lerin tweet metnine ek olarak emoji belirteci, emoji metinsel karşılığı ve hashtag içerdiği veya içermediği farklı tweet yapılandırmalarıyla modelleri test ettik. Ön testlerimizde en iyi performansı (doğruluğu) "metin + emoji belirteçleri + hashtag" yapılandırmasıyla elde ettik ve bu yapıyı BERT sınıflandırıcı modellerimizi eğitmek için kullandık.

2.2. Nefret söyleminin tespiti ve sınıflandırılması

İlk deney etabında, bir tweet'in nefret söylemi içerip içermediğini tespit etmek ve Tablo 2'de yer alan kategorileri kullanarak tweet'i sınıflandırmak amacıyla makine öğrenimi modelleri geliştirdik. Nefret söylemi tespit aşamasını, "nefret söylemi yok" ve "nefret söylemi" seçeneklerini içeren 2 sınıflı bir sınıflandırma olarak kurguladık. Sonraki aşamada, nefret söylemini kategorilere ayırmak için ise iki farklı ayar kullandık. Söz konusu ayarlardan biri "nefret söylemi yok" sınıfı ve Tablo 2'de yer alan dört kategori dahil olmak üzere 6 sınıftan oluşuyor. Diğer ayarda, görevi basitleştirmek için 2. ve 3. kategorileri tek bir sınıf altında, benzer şekilde 4. ve 5. kategorileri de tek bir sınıf altında birleştirerek 4 sınıflı bir sınıflandırma modeli tasarladık.

Elimizdeki verilerin %80'ini eğitim amaçlı, %20'sini ise test amaçlı kullandık. Bu modellerin Tablo 4'te belirtilen performansını ise doğruluk metriği (doğru sınıflandırmaların toplam veri sayısına oranı) kullanarak hesapladık. Beklendiği gibi, görev basitleştikçe performansın arttığını ve 2 sınıflı (nefret söylemi içeren ve nefret söylemi içermeyen) modelde %84,87'lik bir doğruluk oranına ulaşıldığını gözlemledik.

Tablo 4. Nefret söylemi sınıflandırma modellerinin sonuçları

Model	6-sınıf	4-sınıf	2-sınıf
Doğruluk oranı	%80,46	%80,55	%84,87

2.3. Nefret söylemi gücünün tahmini

İkinci deney etabında, nefret söyleminin gücünü 1'den 10'a kadar bir skalada ölçebilmek için BERTurk kullanarak bir regresyon problemi formüle ettik. Modelimiz burada, 1,67'lik bir Kök Ortalama Kare Hatası (KOKH) (*Root Mean Squared Error - RMSE*) değeri elde etti. KOKH, tahmini ve gerçekleşen değerler arasındaki farkı ölçer ve ortalama almadan önce kare alma işlemi yaparak daha büyük hatalara daha fazla vurgu yapar. KOKH değerinin daha düşük olması tahminlerin de daha doğru olduğuna işaret eder, bu da nefret söyleminin değişen yoğunluğunu yakalama konusunda modelin etkinliğini gösterir.

2.4. Hedef grupların tanımlanması

Bir diğer önemli görev de nefret söyleminin yöneldiği genel hedef grup(lar)ının ve özel grup(lar)ın tanımlanmasıdır. Bu tür gönderi içeriklerinin yöneldiği hedef grubu ve özel grubu tanımak, çeşitli kimlik gruplarının karşılaşabileceği olası zararı değerlendirmek açısından hayati önem taşır. Nefret söylemi içeren her tweet için bir genel hedef grup kategorisi (örneğin, toplumsal cinsiyet, milliyet) ve özel grup kategorisi (örneğin, kadınlar, mülteciler, LGBTİ+'lar) tanımladık. Hedef Grup Sınıflandırması'nda, bir metin içeriğinin bir grubu hedef alıp almadığını ve hangi genel hedef grup kategorisine (örneğin toplumsal cinsiyet, milliyet) girdiğini saptıyoruz. Özel Grup Sınıflandırması'nda ise, söz konusu hedef grup kategorisi içindeki bireysel grubu belirliyoruz (örneğin, toplumsal cinsiyet altında kadınlar, milliyet altında mülteciler, cinsel yönelim altında LGBTİ+'lar).

Üçüncü deney etabında ise, dört hedef sınıfı için genel bir hedef grup tanımlama modeli geliştirdik (Ayrıntılı bilgi için yönergelerdeki "Hedef Grupların Tanımlanması" bölümüne bakınız.):

0: Hedef grup belirtilmemiş veya mevcut değil
1: Ülke/milliyet/ırk/etnik köken
2: Din
3: Toplumsal cinsiyet/cinsel Yönelim

Tablo 5. Genel hedef grupların tanımlanmasına ilişkin açıklamalar

Hedef Grupların Tanımlanması	**Açıklamalar**
0: Hedef grup belirtilmemiş veya mevcut değil	Hedef alınan/hedeflenen kimliğin belirsiz olduğu veya açıkça tanımlanmadığı söylemler
1: Ülke/milliyet/ırk/etnik köken	Birey(ler)in/grupların ülkeleri/milliyetleri/ırkları/etnik kökenleri nedeniyle hedef alındığı söylemler. Bu proje kapsamında ele alınan kategoriler şunlardır: Mülteciler, İsrail-Yahudiler, Yunanlar, Ermeniler, Kürtler, Araplar (tam liste için bakınız: **Bölüm 2.5**)
2: Din	Birey(ler)in/grupların dini kimlikleri nedeniyle hedef alındığı söylemler Bu proje kapsamında ele alınan kategoriler şunlardır: Yahudiler, Aleviler (tam liste için bakınız: **Bölüm 2.5**)
3: Toplumsal cinsiyet/cinsel yönelim	Birey(ler)in/grupların toplumsal cinsiyet ve/veya cinsel yönelim nedeniyle hedef alındığı söylemler Bu proje kapsamında ele alınan kategoriler şunlardır: LGBTİ+'lar, Kadınlar (tam liste için bakınız: **Bölüm 2.5**)

Tablo 6, her bir grup için ve ayrıca tüm veri kümesi için sınıf bazlı ve genel F1 skorları üzerinden hedef grupların tanımlanmasına ilişkin sonuçları gösteriyor. F1-skoru, kesinlik skoru (belirli bir sınıf için yapılan tahminlerin ne kadarının gerçekten o sınıfa ait olduğu) ile duyarlılık (*recall*) skorunu (ilgili sınıfa ait verilerin ne kadarının doğru tahmin edildiği) tek bir skorda birleştiren bir metriktir. Sınıflandırma problemlerinde, farklı sınıflardaki örnek sayısı önemli ölçüde farklılık gösterdiğinde F1 skoru doğruluk oranına tercih edilir. Tablo 6'da gösterildiği gibi, hedef tespit modelimiz için ortalama F1 skoru (makro ortalama) %60,0 iken, doğruluk oranı (mikro ortalaması alınmış F1 skoru ile aynıdır) %73,0'dır. Ülke, milliyet, ırk veya etnik köken sınıfında yer alan bir hedef grup daha güvenilir tanımlama sonuçları verirken, din veya cinsel yönelim sınıfındaki grupların tanımlanmasında başarı oranı daha düşüktür.

Tablo 6. Çok etiketli genel hedef grupları için tanımlama modelinin sonuçları

	F1-skoru	**Örnek sayısı (büyüklük)**
Hedef grup belirtilmemiş veya mevcut değil	0,70	870
Ülke/Milliyet/Irk/Etnik Köken	0,82	1349
Din	0,46	256
Toplumsal Cinsiyet/Cinsel Yönelim	0,43	49
Ortalama (mikro ortalama)	**0,73**	**2524**
Ortalama (makro ortalama)	**0,60**	**2524**

2.5. Özel grupların tanımlanması

Dördüncü deney etabında, aşağıda yer alan on bir kategoriden oluşan bir özel grup tanımlama modeli geliştirdik:

0: Hedef grup yok
1: Mülteciler
2: Yahudiler
3: Yunanlar
4: Ermeniler
5: Aleviler
6: Kürtler
7: Araplar
8: LGBTİ+'lar
9: Kadınlar
10: Diğer gruplar

Tablo 7, geliştirdiğimiz modelin özel grupların sınıflandırılması açısından elde ettiği sonuçları gösteriyor.

Tablo 7. Özel grupların sınıflandırılmasına ilişkin sonuçlar

11 sınıflı sınıflandırma modeli	
Doğruluk oranı	0,96

2.6. Metin aralığı tespiti

Nefret söylemi tespitinin amacı belli bir metnin nefret içerip içermediğini belirlemek iken, metin aralığı tespiti [*span detection*], nefret söylemi göstergelerinin metin içindeki yerini tam olarak saptayarak daha iyi bir içgörü sağlar.

Bu işleve yönelik bir model geliştirebilmek için, metin aralığı tespiti işlemini bir belirteçleme (tokenizasyon - metin öbeğinin küçük parçalara bölünmesi) görevi olarak formüle ettik. Her bir belirteç [*token*] –bu çalışma özelinde her bir alt kelime– nefret söylemi aralığına girip girmediğini belirten özel bir etiketle [*tag*] işaretlendi. BERTurk modelini bu amaç doğrultusunda eğitebilmek için, tweet içeriklerindeki nefret söylemi metin aralıklarının işaretlendiği veri kümesine ihtiyaç duyduk. Ancak, etiketleme işlemi tweet bazında yapıldığı için, başka bir deyişle tweet içeriğindeki nefret söylemi aralığından ziyade her bir tweet içeriğinin tamamı etiketlendiği için, nefret söylemini gösteren metin aralıkları için tweet'lere ek bir etiketleme yapmamız gerekiyordu. Bunu yapabilmek için, üç etiketleyicinin anlaşmaya vararak etiketlediği tweet'ler arasından farklı büyüklükteki grupları (Ermeni, Yunan, Yahudi, Arap, Göçmen/Mülteci, LGBTİ+, Alevi, Kürt) hedef alan tweet'leri seçtik. Sadece göçmenleri/mültecileri hedef almaları ve sayılarının sınırlı olması sebebiyle, yeterince çeşitlilik arz etmedikleri için, Arapça tweet'leri analizimizden çıkardık. Sonrasında GPT-4 büyük dil modelini işleterek, tutarsız metin aralıklarını [*hallucinated spans*] filtreleyerek ve küçük metin varyasyonlarını gidererek nefret söylemine işaret eden metin aralıklarını otomatik olarak çıkardık. Bu süreçte, iki etiketleyici GPT-4 metin aralıklarını inceledi, anlaşmazlık hâlinde üçüncü bir etiketleyici devreye girerek en uygun etiketin seçilmesini sağladı. Bu süreç sonucunda, dağılımı Tablo 8'de gösterilen 3.697 tweet elde edildi.

Tablo 8. Hedef gruba göre metin aralığı tespiti için kullanılan tweet sayısı

Hedef Grup	**Tweet Sayısı**
Yahudi	1132
Yunan	1119
Ermeni	628
Arap	337
Göçmen+Mülteci	242
Alevi	127
Kürt	63
LGBTİ+	49

Metin aralığı tanımlama modelini eğitmek amacıyla, burada yer alan verileri, üç etiketleyici tarafından fikir birliği hâlinde "nefret söylemi yok" olarak etiketlenen tweet içerikleriyle birleştirdik. Bu şekilde, model üzerinden, nefret söylemi içeren metin aralıklarının tespitinde %41'lik bir F1-skoru elde ettik. Bu da, modelimizin nefret söylemi içeren metin aralıklarının neredeyse yarısını doğru bir şekilde tespit edebildiğini, bunu yaparken de doğruluk ve tamlık arasında bir denge kurabildiğini gösterdi. Geliştirdiğimiz aracımız, bu modeli kullanarak, belirli bir metin içerisinde nefret söylemine işaret eden kelimelere etiketleme yapabiliyor. Modelin orta düzeyde bir performans sergiliyor olması, etiketlenmiş veri kümesinin küçük bir boyutta olmasıyla ve bu problemin doğası gereği zorluklar barındırmasıyla açıklanabilir.

2.7. Türkçe yazılı basında nefret söyleminin tespiti

Türkiye'de belirli grupları ve kimlikleri hedef alan nefret söyleminin varlığını sürdürdüğü mecralardan biri de yazılı basındır. Bu mecrada etkili bir içerik moderasyonu sağlamak ve nefret söylemiyle mücadele etmek amacıyla, Türkçe haberlerde nefret söyleminin farklı yönlerini analiz eden bir dizi model geliştirdik. Bu modellerden ilki nefret söyleminin varlığını saptamaya, ikincisi nefret söyleminin türünü kategorize etmeye ve üçüncüsü de özel olarak hedeflenen bir grubu belirlemeye yöneliktir. Nefret söylemi tespit modeli, bir haber veya köşe yazısının nefret söylemi içerip içermediğini tahmin eden ikili bir sınıflandırma modelidir. Nefret söylemi kategorizasyon modeli, içeriği üç kategoriye ayıran çok sınıflı bir sınıflandırma modelidir. Bu model, yazılı basın haberleri veri kümesinde yer almayan *dışlama/ayrımcı söylem* kategorisi hariç olmak üzere, (Tablo 2'de listelenen) tweet bazlı nefret söylemi kategorilerimizle uyumludur. Hedef grup tespit modeli ise, nefret söyleminin yöneltildiği özel grupları tanımlamak için kullanılmak üzere, tweet bazlı modelle uyumlu, on bir sınıflı bir sınıflandırma görevi olarak formüle edilen bir modeldir.

Söz konusu tüm modeller, Türkçe metinleri anlama kabiliyeti nedeniyle BERTurk üzerine inşa edildi ve Vakfın haber arşivi kullanılarak modellere ince ayar yapıldı. Arşiv başlangıçta görüntü tabanlı bir formatta olduğundan, yüksek kaliteli metne dönüştürmek için ön işleme gerektiriyordu. Bu süreçte, görüntüden metne dönüştürme işlemi için EasyOCR kullanıldı, sonrasında GPT-4 ile son işleme yapılarak veri kümesinin kalitesi önemli ölçüde artırıldı. Ortaya çıkan veri kümesi, etnik kökenler, milliyetler ve dinî topluluklar dahil olmak üzere farklı hedef gruplarına yönelik farklı uzunluklarda haber ve köşe yazılarından oluşuyor. Veri kümesi, hem yerel hem de ulusal basından 1.210 medya kuruluşundan elde edilen haberleri ve 200'den fazla hedef grubu içeriyor. Dengeli bir eğitim seti oluşturabilmek amacıyla, veri kümesi, nefret söylemi içeren haberlerin yanı sıra aynı zaman diliminden rastgele seçilmiş

nefret içermeyen haber içerikleriyle tamamlandı. Veri kümesinde yer alan nefret içerikli haberlerin dağılımı aşağıdaki gibidir:

10.198 haber içeriği *abartma/genelleme/yükleme/çarpıtma* kategorisinde, 1.199 haber içeriği *simgeleştirme* kategorisinde olup 121 haber içeriği her iki kategoriye de giriyor (karma kategori olarak değerlendirildi). Veri kümesinde ayrıca *düşmanlık/savaş/saldırı/öldürme/yaralama tehdidi* olarak sınıflandırılan 2.279 haber içeriği, *küfür/hakaret/aşağılama/insandışılaştırma* olarak sınıflandırılan 644 haber içeriği ve her iki kategoriye de giren (bir başka karma kategori oluşturan) 13 haber içeriği bulunuyor. Bununla birlikte, nefret söylemi kategorisi modelinde, nefret söylemi içermeyen 14.715 haber içeriği de ayrı bir sınıf olarak yer alıyor.

Söz konusu modeller değerlendirme aşamasında sağlam bir performans sergiledi. İkili sınıflandırmalı nefret söylemi tespit modeli %87,49'luk bir F1 skoru elde ederek nefret içeren metinleri nefret içermeyen metinlerden ayırt etme konusundaki güçlü kabiliyetini ortaya koydu. Hedef grup tespit modeli %82,38'lik bir F1 skoru elde ederek tweet bazlı tespit modeli muadilinden daha düşük bir performans sergiledi. Bu fark iki ana faktöre bağlanabilir: veri kümesindeki sınıflandırma işleminde dengesizlik ve tweet ile haber içeriği arasındaki yapısal farklılık. Haber içerikleri, doğası gereği daha doğrudan anlatımlı kısa ve öz tweet'lere kıyasla, birden çok hedef grupla bağlantılı olabilen daha uzun ve karmaşık metinlerden oluşur ve daha ince nüanslar içeren bir dil kullanır. Bu da haber içeriğinin sınıflandırılması görevini daha da zorlaştırır. Nefret söylemi kategorisi tahmin modeli ise, orta derece ancak makul sayılabilecek %78,57'lik bir F1 skoru elde ederken, karmaşık nefret söylemi kategorilerinin ince nüanslarını yakalamadaki zorluklara da işaret ediyor. Bu sonuçlar, Türkiye'nin yazılı basınında yer alan kutuplaştırıcı ve çoğunlukla düşmanca söylemi anlama ve analiz etme konusunda söz konusu modellerin etkinliğini ortaya koyarken, içerik moderasyonu ve analizinin otomasyonu için de önemli araçlar sunuyor.

2.8. Medya takibi ve analizi

Nefret söylemiyle mücadelede bir diğer önemli adım da olası nefret söylemi kaynaklarının gerçek zamanlı ve sürekli olarak izlenmesi ve analiz edilmesidir. Bu amaç doğrultusunda, Suriyeli ve Mülteci gibi belirli anahtar kelimeler üzerinden X platformundaki içeriği sürekli olarak izleyen bir yazılım bileşeni ekleyerek geliştirmiş olduğumuz aracı daha da zenginleştirdik. Bu bileşen, X API'sini kullanarak her doksan dakikada bir otomatik olarak söz konusu anahtar kelimeleri içeren tweet'leri yakalıyor, ardından nefret söylemi tespit modellerimizi kullanarak bunları etiketliyor, işlevlerini ortaya koyuyor ve böylelikle ilgili söylemin analiz edilme ve raporlanma sürecini kolaylaştırıyor. Bu işlemler sonrasında, elde edilen sonuçlar,

nefret içerikli tweet'lerin günlük yüzdesini gösterir bir grafik olarak (x ekseninde günler, y ekseninde yüzdeler belirtilerek) sunuluyor. Söz konusu bileşen, içerik takibi ve daha derin analizler konusunda umut vaat eden faydalı bir araç olmakla birlikte, X platformunun kota ve politikaları gereği günlük olarak erişim sağlanan tweet sayısının az olması sebebiyle kısıtlı kalıyor. Tweet içeriklerine daha fazla erişilebilirlik sağlandığı takdirde, bu aracın farklı söylemleri analiz etmede değerli olabileceğine inanıyoruz.

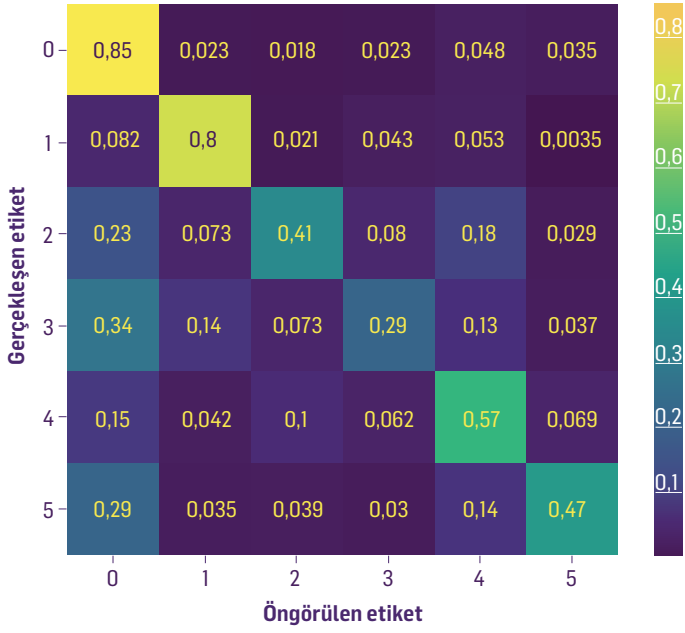
3. HATA ANALİZİ

Geliştirdiğimiz yapay zekâ aracının performansını daha iyi anlamak amacıyla, test seti üzerindeki performansını farklı açılardan analiz ettik. Bölüm 2.2'de yer alan altı sınıflı sınıflandırma denemesi üzerinde detaylı hata analizi yaptık.

Şekil 1, nefret söylemi kategorizasyonu hata matrisini ortaya koyuyor ve farklı kategoriler arasındaki yanlış sınıflandırmayı yüzdelik değer cinsinden gösteriyor. Bu noktada, her satırın toplamının 1 olduğunu not etmek gerekiyor; bir başka deyişle, satırdaki her değer, Öngörülen Etiket sınıflandırmasının Gerçekleşen Etiket'e dönüşme yüzdesini temsil ediyor. Örneğin, Gerçekleşen Etiket 2 (Simgeleştirme) ve Öngörülen Etiket 4 (Küfür, Hakaret, Aşağılama, İnsandışılaştırma) satırına karşılık gelen değerin 0,18 olması, esasında 2. Kategori'ye ait olması gereken tweet'lerin %18'inin yanlışlıkla 4. Kategori altında sınıflandırıldığına işaret ediyor. Benzer şekilde, Gerçekleşen Etiket 3 (Abartma, Genelleme, Yükleme, Çarpıtma) ve Öngörülen Etiket 3 satırına karşılık gelen değerin 0,29 olması, 3. Kategori'ye ait tweet'lerin %29'unun doğru bir şekilde 3. Kategori altında sınıflandırıldığını gösteriyor.

Genel olarak, modelin, tweet'leri 0 Kategorisi (nefret söylemi yok) altında sınıflandırma eğiliminde olduğunu gözlemliyoruz; bu kategoride doğruluk oranının %85 çıkması da bu durumu kanıtlar nitelikte. Bu gözlemimiz 2., 3., 4. ve 5. kategorilerdeki hata oranlarının yüksekliği ile, bir başka deyişle bu kategorilerde yer alması gerekirken yanlışlıkla 0 kategorisinde sınıflandırılmaları öngörülen tweet oranları ile de destekleniyor. 1. Kategori'de (Dışlama, Ayrımcı Söylem) %80 gibi nispeten yüksek bir doğruluk oranına ulaşılıyor. Bununla birlikte, diğer kategorilerde daha düşük doğruluk seviyeleri ortaya çıkıyor, 3. Kategori (Abartma, Genelleme, Yükleme, Çarpıtma) %29 ile en düşük performansı sergiliyor. Ayrıca, 2. Kategori (Simgeleştirme) ve 4. Kategori (Küfür, Hakaret, Aşağılama, İnsandışılaştırma) arasında da kayda değer bir hata oranı olduğu ortaya çıkıyor, zira her iki kategori arasında çapraz bir şekilde hatalı sınıflandırma yapıldığı görülüyor.

Şekil 1. Nefret söylemi kategorizasyonu hata matrisi



Bir sonraki adımda, doğru bir şekilde sınıflandırılan ve hatalı bir şekilde sınıflandırılan tweet'lere ilişkin etiketleyici uzlaşma oranlarını [*annotator agreement*] inceledik. Etiketleyiciler arasında uzlaşma durumunu değerlendirebilmek için bir kez daha Krippendorff'un alfa katsayısını kullandık. Bu yöntemde, değerlerin daha yüksek olması etiketleyiciler arasında daha güçlü bir uzlaşma olduğuna işaret ediyor. Tablo 9'da yer alan sonuçlar, model tarafından doğru bir şekilde sınıflandırılmış tweet'lerde daha yüksek bir etiketleyici uzlaşma eğilimi olduğuna işaret ediyor. Bu da, etiketleyicilerin kategorize etmekte zorlandığı tweet'leri doğru sınıflandırmak konusunda modelin de zorlandığını gösteriyor. Ek olarak, üç etiketleyicinin de aynı şekilde sınıflandırdığı tweet'lerin alt kümesinde nefret söylemi kategorizasyonunun doğruluğunu da ölçtük. Burada model %90,17'lik bir doğruluk oranı elde etti; bu da Tablo 4'te yer alan tüm tweet'lerdeki doğruluk oranı olan %80,46'ya kıyasla çok daha yüksek bir doğruluk oranına ulaştığı anlamına geliyor.

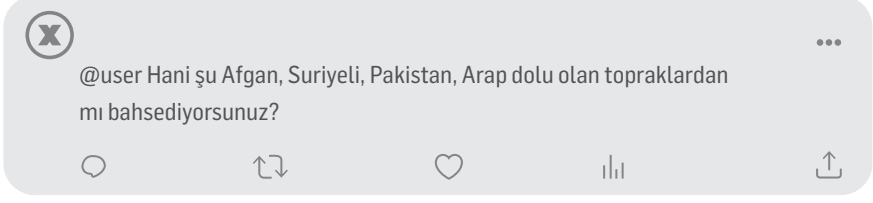
Tablo 9. Doğru ve yanlış sınıflandırılmış tweet'lere göre Krippendorff Alfa katsayıları

	Krippendorff Alfa Katsayısı
Doğru Sınıflandırılmış	0,387
Yanlış Sınıflandırılmış	0,202
Tüm Test Verileri	0,335

Tüm bu sonuç ve analizler, nefret söylemini tespit etmenin ve sınıflandırmanın, özellikle etiketleyici anlaşmazlığının olduğu durumlarda, zorlu bir görev olduğunu ortaya koyuyor.

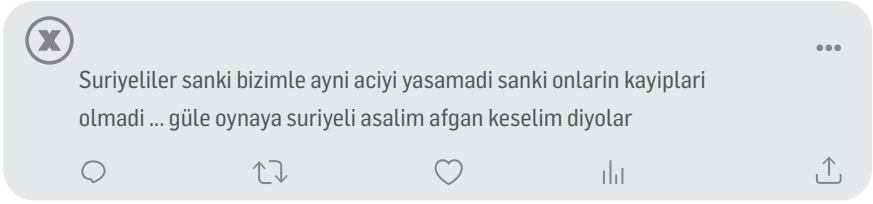
Son bir analiz olarak, model tarafından yanlış sınıflandırılmış bazı tweet'leri inceledik ve modelin bunları doğru sınıflandırma konusunda karşılaştığı olası zorluklara dair açıklamalar sunduk. Nefret söyleminin otomatik olarak sınıflandırılması zorlu bir görev olup, doğal dilin anlaşılmasındaki standart zorluklara ek olarak, tweet'lerde ve medyada nefret söylemi tespiti aşağıda özetlenen bazı başka zorluklar da barındırıyor:

- **Bağlam eksikliği:** Bir metnin yazarının niyetini tam olarak kavramak zorlu bir görevdir. Kısa tweet'lerin bağlamları/arkaplanları dikkate alınmadan incelenmesi (etiketleme işlemi sırasında veya makine tarafından kategorizasyon sürecinde bu tweet'ler tek başlarına inceleniyor), bağlam eksikliğini ana zorluk hâline getiriyor. Bu durum, önceki tweet'lere cevap olarak paylaşılan tweet'lerde daha da sorunlu bir hâl alıyor. Örneğin:

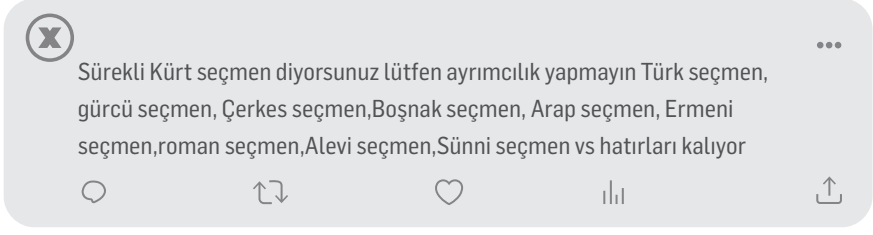


Mevcut bağlam göz önüne alındığında, ilgili tweet yazarının belirli ülkelere mi yoksa bu ülkelerden gelen insanların yaşadığı Türkiye topraklarına mı atıfta bulunduğu belirsizdir. Tweet'in nefret içerip içermediği ise bu tür ayrımlara göre değişir. Ancak, yukarıdaki tweet'te görüldüğü gibi, bağlamın olmayışı, doğru etiketleme yapılmasını karmaşıklaştıran bir faktördür.

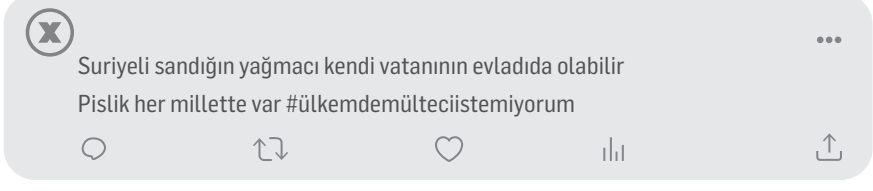
- **Gizli niyet:** Bazı tweet'ler nefret söylemi unsurları içerebilir, ancak, tweet içeriğinin kendisi nefret dolu değildir; örneğin, tweet yazarı esasında belirli bir kimliğe yönelik olarak nefret söylemi kullananları kınıyor olabilir. Bu durum tipik bir doğal dil anlama problemi olsa da, kısa tweet'ler söz konusu olduğunda kullanıcının niyetini anlamak daha da zorlaşır.



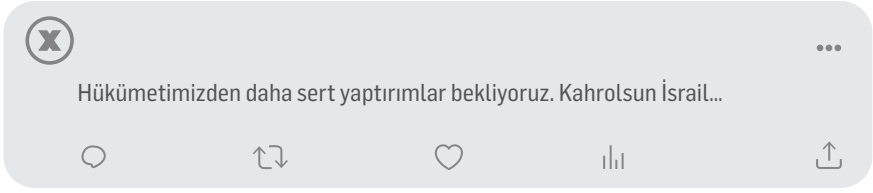
Bu durum iğnelemeler için de söz konusudur. Kullanıcılar alaycı bir dil kullanarak zarar verme niyetlerini gizleyebilir, ifadelerinin daha az agresif ve hatta mizahi görünmesini sağlayabilirler. Alaycılık, kullanıcıların sözlerini şaka veya ironi olarak sunarak hesap vermekten kaçınmalarını sağlar. Sonuç olarak, nefret söyleminin normalleşmesine katkıda bulunarak tespit edilmesini ve ele alınmasını daha zor hâle getirebilir.



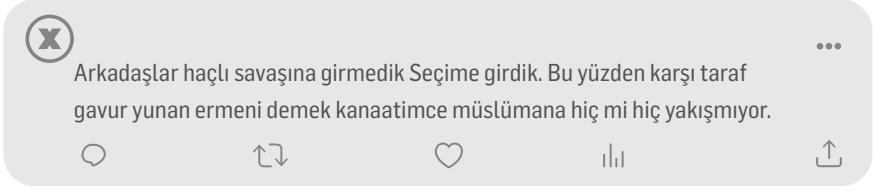
- **Karşıt metin-hashtag kullanımı:** Bazı tweet'ler nefret içerikli bir hashtag içerebilir, ancak, hashtag'in vurgu yaptığı hislere karşıt görüşler ifade ediyor olabilir. Bu gibi durumlarda etiketler muğlaklaşır, ancak, model böyle bir tweet'i nefret içerikli olarak sınıflandırma eğiliminde olabilir.



- **Hatalı etiketleme:** Etiketleme süreci her ne kadar kapsamlı ve itinalı bir şekilde yürütülse de, bazı tweet'ler etiketleyiciler tarafından hatalı etiketlenmiş olabilir. Model eğitim verilerindeki bu yanlış etiketlenmiş tweet'ler modelin kafasını karıştırabilir ve performansın düşmesine neden olabilir.



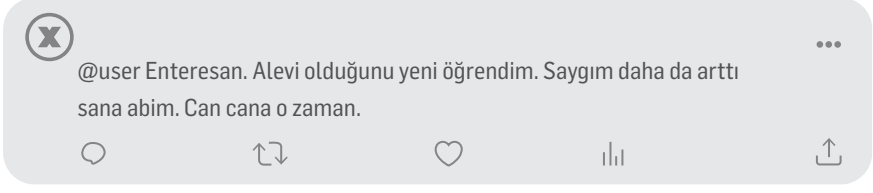
Bu tweet etiketleyiciler tarafından "Nefret Yok" olarak etiketlenirken, model aslında bu tweet için doğru kategorizasyon olan "Düşmanlık, Savaş, Saldırı, Öldürme, Yaralama Tehdidi" sonucunu vermiştir.



Bu tweet etiketleyiciler tarafından "Nefret Yok" olarak etiketlenirken, model aslında bu tweet için doğru kategorizasyon olan "Simgeleştirme" sonucunu vermiştir.

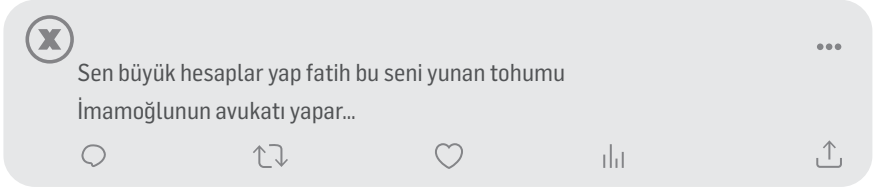
Gerçekleşen ve öngörülen etiketler arasında yaptığımız karşılaştırma çalışması kapsamında, her bir kategoride model tarafından doğru sınıflandırılmış bazı tweet örneklerini de paylaşıyoruz:

Nefret Söylemi Yok:

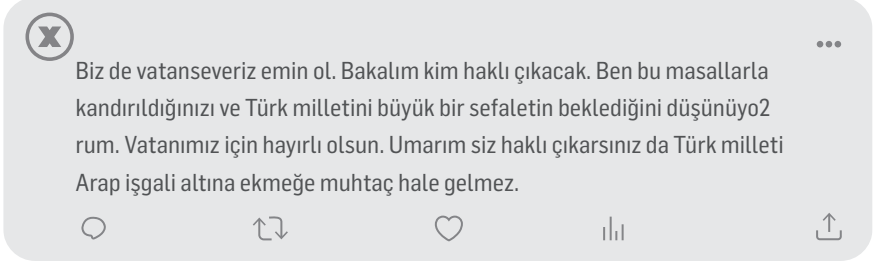


Dışlama, Ayrımcı Söylem:

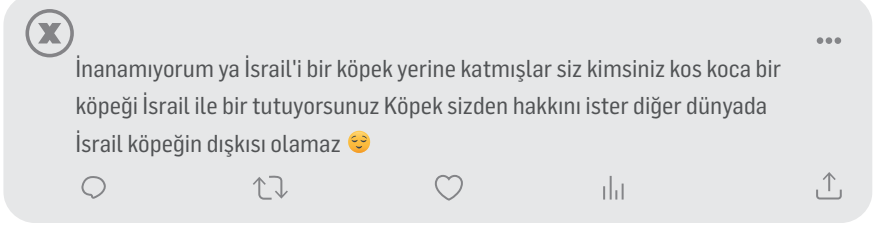
Simgeleştirme:



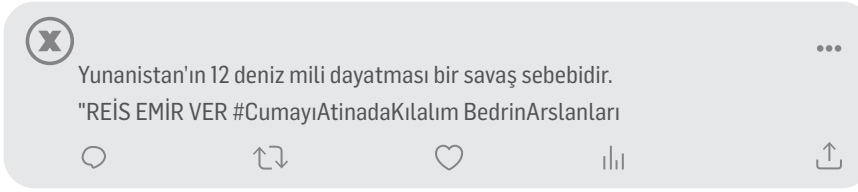
Abartma, Genelleme, Yükleme, Çarpıtma:



Küfür, Hakaret, Aşağılama, İnsandışılaştırma:



Düşmanlık, Savaş, Saldırı, Öldürme, Yaralama Tehdidi:



4. ÇALIŞMANIN KISITLARI

Nefret söylemi tespit aracı geliştirilirken, modelin performansını ve genellenebilirliğini etkileyebilecek çeşitli kısıtlarla karşılaşıldı. Önemli zorluklardan biri, büyük ölçüde X platformunun değişen politikaları ve kota kısıtlamalarından kaynaklanan veri toplama sorunları oldu. Söz konusu değişiklikler çok daha kapsamlı bir veri kümesine erişimi kısıtlarken, erişim sağlanabilecek tweet çeşitliliğini ve hacmini azalttı. Bu durum, modelin farklı nefret söylemi kategorileri arasında genelleme yapma kabiliyetini etkileyebilecek bir faktördür.

Bir diğer kısıt da etiketleme sürecinden kaynaklandı. Her ne kadar etiketleyicilere yönelik eğitim yapılmış olsa da, etiketleyiciler arasında tutarlı bir şekilde anlaşma sağlamak güç oldu. Etiketleyicilerin farklı arkaplanlara sahip olmaları, kendi bakış açılarını ve yorumlarını da beraberinde getirdiği için, nefret söyleminin nasıl tanımlandığı konusunda değişkenlik söz konusu oldu. Bu değişkenlik durumu, nefret söylemini tespit sürecinin öznel doğasına işaret ederken, eğitimli bireyler arasında dahi ortak bir anlayışa ulaşmanın ne kadar karmaşık olabileceğini de güçlü bir şekilde ortaya çıkardı.

Bunlara ek olarak, veri kümesinin bütünlüklü bir ileti dizisi [*thread*] yerine tek tek tweet'lerden oluşması, çoğunlukla bağlam eksikliği sorununa yol açtı. Bağlama dair bilgi eksikliği, etiketleyicilerin her bir tweet'in arkasındaki niyet ve nüansı doğru bir şekilde değerlendirmesini zorlaştırdı. Zira, tweet içerikleri daha geniş bir bağlam içinde verilmeden tek başlarına ele alındığında muğlak veya yanıltıcı olabiliyor. Ayrıca, bu süreçte tweet'lerin yayılımını veya viralleşme durumunu analiz etmedik; dolayısıyla, nefret söyleminin nasıl yayıldığını ve dezenformasyon gibi konularla nasıl kesiştiğini keşfetme fırsatını kaçırdık. Nefret içeren tweet'lerin yayılma biçimlerini anlamak, bu içeriklerin etkisi ve erişim alanına dair değerli bilgiler sağlayabilirdi ve kapsamlı nefret söylemi azaltma stratejileri geliştirilmesine son derece önemli katkılar sunabilirdi. Buna ek olarak, aracımızın geliştirilmesi sırasında görseller eğitim sürecine dahil edilmedi. Bazı durumlarda, bir tweet ancak beraberindeki görselle birlikte analiz edildiğinde nefret söyleminin varlığının

ortaya çıkabildiğini gözlemledik. Dolayısıyla, görsel içeriğin incelenmesinin büyük önem taşıdığını not ettik. Bunun gelecekteki araştırmalar için çok önemli bir alan olmaya devam edeceğini düşünüyoruz.

Nefret söyleminin dinamik ve sürekli evrilen doğası, özellikle bu süreçte yeni türetilen argo tabirler, şifreli dil kullanımı ve kültürel göndermeler, tespit aracı açısından süregelen bir zorluk teşkil ediyor. Modeli bu türden değişimlere ayak uyduracak şekilde uyarlamak için modeli sürekli güncellemek ve doğruluk oranlarını korumak için de yeni verilerle yeniden eğitmek gerekiyor. Bu faktörler bir bütün olarak ele alındığında, sürekli iyileştirmelere duyulan ihtiyaç ve nefret söylemi tespit aracının elde ettiği sonuçları yorumlarken daha geniş ölçekteki kısıtları anlamanın önemi daha güçlü bir şekilde ortaya çıkıyor.

3 SONUÇ

“Dijital Teknolojileri Kullanarak Nefret Söylemi ve Ayrımcılıkla Mücadele” projesi kapsamında Hrant Dink Vakfı, Boğaziçi Üniversitesi ve Sabancı Üniversitesi araştırmacıları tarafından hazırlanan bu rapor, nefret söylemine iletişim, dilbilim, kültürel çalışmalar ve bilgisayar bilimleri gibi disiplinlerin farklı bakış açılarını ve ortak çabalarını yansıtmaktadır. Etiketleme kılavuzunda yer alan örnekler ve bu örnekler üzerinden yapılan tanımlamalardan da görülebileceği gibi, nefret söylemi verileri çoğu zaman anlaşılması güç, belirsizlikler ve tutarsızlıklarla dolu olsa da, bu projedeki temel amacımız geliştirdiğimiz nefret söylemi tespit aracı sayesinde, nefret söylemini daha derinlemesine ve daha net bir şekilde anlayıp bu sorunla daha etkili bir şekilde mücadele edebilmektir. Bu yöntemin ve geliştirilen aracın bundan sonraki çalışmalarda da kullanılabileceğini ya da en azından temel bir başlangıç noktası sağlayacağını umuyoruz.

EK-1: Etiketleme arayüzü

Tweet

@HDPgenelmerkezi gerçek kürt Müslüman dır ermeni bir haini partiye sokmaz ermenilerin amacı kürt türk savaşı çıkarıp kürt topraklarında batı ermenistanı kurmaktır bunlar hepsi ermeni Yahudi ajanıdır. çocuklarınızı pkk ya yollamayın onların çocukları gitsin güya kürtler ya...

Aşağıdaki kutucuklardan birine tıkladıktan sonra, nefret söylemine sebep olduğunu düşündüğünüz kelime veya kelime gruplarını tweet üzerinde seçebilirsiniz.
(En fazla 3 kelime veya kelime grubu seçilebilir.)

Tetikleyici Kelime 1 | Küfür/Hakaret 2 | Düşmanlık Söylemi 3

Dil

Lütfen tweet'in hangi dilde olduğunu işaretleyiniz.

◉ Türkçe [4] ○ Türkçe değil[5]

Genel Tutum/Duruş

Lütfen yalnızca 1 seçeneği seçiniz.

◉ Emin değilim [6] ○ Alevi Karşıtı [7] ○ Nötr [8] ○ Alakasız [9]

Hedeflenen Grup

Hedef grup birden fazlaysa birden fazla kategori seçilebilir.

◉ Demografik/ Sosyoekonomik/ Irk/Etnik Köken [0]
○ Ülke/ Milliyet [q]
○ Din [w]
○ Cinsiyet [e]
○ Cinsel Yönelim [t]
○ Belli Görüş/Statü/ Uygulama, Mesleki Pozisyon Grubu [a]
○ Hedef grup belirgin değil veya yok. [s]
○ Hedef grup birden fazla. [d]

Nefret Söylemi Derecesi

Lütfen yalnızca 1 seçeneği seçiniz.

○ Emin değilim. [f] ◉ 0 [g] ○ 1 [z] ○ 2 [x] ○ 3 [c] ○ 4 [v]
○ 5 [b] ○ 6 [y] ○ 7 [i] ○ 8 [o] ○ 9 [p] ○ 10 [j]

○ Dışlama, Ayrımcı Söylem

Bir topluluğun hak ve özgürlüklerden faydalanma, topluma dahil olma gibi alanlarda baskın gruptan olumsuz anlamda farklı görüldüğü söylemlerdir.[k]

Nefret Söylemi Kategorisi

Birden çok kategori seçilebilir; hedef belli değil ise de metin içeriğine göre kategori seçilmeli. Genel tutum/duruş kısmında emin değilim işaretlendiyse burada da emin değilim işaretlenmeli.

○ **Emin değilim**[l]	○ **Nefret söylemi bulunmuyor**[n]	○ **Simgeleştirme** *Bir kimlik öğesinin kendisinin hakaret, nefret veya aşağılama unsuru olarak kullanıldığı, kimliğin bu yollarla simgeleştirildiği söylemlerdir.*[m]
○ **Abartma, Genelleme, Yükleme, Çarpıtma** *Bir olayı, durumu ya da eylemi olduğundan daha büyük sonuçlara ve çıkarımlara vardıran, gerçek verileri saptırarak manipüle eden veya münferit olayları öznelerinden yola çıkarak kimliğin bütününe yükleyen söylemlerdir.*	◉ **Küfür, Hakaret, Aşağılama, İnsandışılaştırma** *Bir topluluğa yönelik doğrudan küfür, hakaret, aşağılama içeren veya onları insan dışı varlıklara özgü eylem veya sıfatlara niteleyerek yapılan aşağılamaların bulunduğu söylemlerdir.*	○ **Düşmanlık, Savaş, Saldırı, Öldürme, Yaralama Tehdidi** *Bir topluluk hakkında düşmanca, savaşı çağrıştıran veya söz konusu kimliğe yönelik zarar verme isteğinin dile getirildiği ifadelerin yer aldığı söylemlerdir.*

Saldırgan Dil

Lütfen yalnızca 1 seçeneği seçiniz.

○ **Yok**

Örnekler:

- Mülteciler bu ülkeden gitmeli.
- Suriyelilerle aynı otobüse binmek istemiyorum.

◉ **Zayıf**

Örnekler:

- Aptallar yine saçmalıyor.
- @KULLANICI sen ne biliyorsun sanki gerizekalı.

○ **Şiddetli**

Örnekler:

- *Bu ucuza çalışan o***** ç****ları işimizi elimizden alıyor.*
- *@KULLANICI buradan yazmak kolay. Sokakta karşıma biriniz çıkarsa onun a**** k***cağım.*

aS·u·lis

DISCOURSE
DIALOGUE
DEMOCRACY
LABORATORY