BMBAD BJESR Mühendislik Bilimleri ve Araştırmaları Dergisi
Journal of Engineering Sciences and Researches

Twitter Platformundan Elde Edilen Türkçe Saldırgan Dil Derlemi

A Corpus of Turkish Offensive Language on Twitter Platform

[1]Şeyma ŞAHİNER YILMAZ, [2]İlyas ÖZER, [3]Hadi GÖKÇEN

[1]Bandırma Onyedi Eylül University, Management Information System Department, Bandırma, Turkey

[2]Bandırma Onyedi Eylül University, Computer Engineering Department, Bandırma, Turkey, AINTELIA Artificial IntelligenceTechnologies Company, Bursa, Turkey

[3]Gazi University, Industrial Engineering Department, Ankara, Turkey

[1]ssahiner@bandirma.edu.tr, [2]iozer@bandirma.edu.tr,
[3]hgokcen@gazi.edu.tr

Araştırma Makalesi/Research Article

ARTICLE INFO

Article history

Received : 10 September 2022
Accepted : 10 October 2022

Keywords:

Sentiment Analysis, Deep Learning, Offensive Language, Social Media, Twitter

ABSTRACT

It has been observed that content with offensive language among users' posts on social media platforms has increased significantly. The study aims to contribute to the solution of this problem in Turkish language. In this study, a data set obtained from the Twitter platform was created. This data set, consisting of 14752 Turkish tweet texts, was annotate manually by the annotator and the classification performances of LSTM (Long ShortTerm Memory) and GRU (Gated Recurrent Units) models were compared. It is the first study in which multi-classification was made for Turkish on offensive language. Here, word representations were obtained with the word2vec method. Thus, the contribution of the use of extended corpus to the classification performances was compared. The highest performance GRU model F1-score value is 94.49% with the use of extended corpus in the binary classification made in the study. The classification performance values obtained in multiclassification are 71.97% and 54.10% of the GRU F1-macro value with the contribution of the expanded corpus. The datasets and expanded corpus word vectors of the current study will be shared in order to contribute to the Turkish language literature in this field.

MAKALE BİLGİSİ

Makale Tarihleri

Gönderim : 10 Eylül 2022
Kabul : 10 Ekim 2022

Anahtar Kelimeler:

Duygu Analizi, Derin Öğrenme, Saldırgan Dil, Sosyal Medya Uygulaması, Twitter

ÖZET

Sosyal medya platformlarında kullanıcıların paylaşımlar arasında saldırgan dil barındıran içeriklerin önemli oranda arttığı gözlemlenmiştir. Çalışma Türkçe dilinde bu sorunun çözümüne katkı sağlamayı amaçlamaktadır. Bu çalışmada Twitter platformundan elde edilen bir veri seti oluşturulmuştur. 14752 Türkçe tweet metninden oluşan bu veri seti etiketleyiciler tarafından manuel olarak etiketlenmiş ve LSTM (Long ShortTerm Memory) ve GRU (Gated Recurrent Units) modellerinin sınıflandırma performansları karşılaştırılmıştır. Bilinebildiği kadarıyla saldırgan dil tespitine yönelik bu alanda yapılan çalışmalara bakıldığında çalışma Türkçe dilinde çoklu sınıflandırma yapılan ilk çalışmadır. Burada word2vec yöntemi ile kelime temsilleri elde edilmiştir. Böylelikle genişletilmiş derlem kullanımının sınıflandırma performanslarına katkısı karşılaştırılmıştır. Çalışmada yapılan ikili sınıflandırma da genişletilmiş derlem kullanımıyla en yüksek performans GRU modeli F1-makro değeri %94,49'dur. Çoklu sınıflandırmada elde edilen sınıflandırma performans değerleri genişletilmiş derlemin katkısıyla GRU F1-makro değeri %71,97 ve %54,10'dur. Bu alanda Türk dili literatürüne katkı sağlamak amacıyla mevcut çalışmanın veri setleri ve genişletilmiş derlem kelime vektörleri paylaşılacaktır.

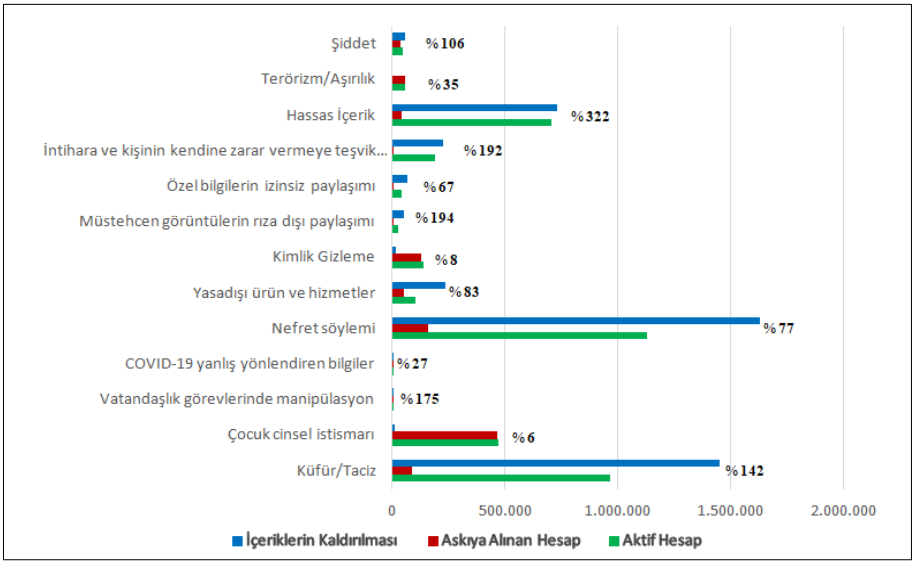
ORCID ID: [1]0000-0002-3301-9491
[2]0000-0003-2112-5497
[3]0000-0002-5163-0008

1. GİRİŞ

Twitter gibi sosyal medya mecralarının hızlı büyümesiyle birlikte daha fazla kullanıcı, görüşlerini ve fikirlerini çevrimiçi olarak paylaşmaktadır. Bu büyümenin ardından, bilgisayar bilimlerinde en verimli araştırma alanlarından biri olan duygu analizi çalışmaları artmıştır [1]. Duygu analizi çalışmalarında Twitter verileri yaygın olarak kullanılmıştır. Duygu analizi, insanların varlıklara yönelik fikirlerini, duygularını, değerlendirmelerini, tutumlarını ve bunların yazılı metinde ifade edilen niteliklerini analiz eden çalışma alanıdır [2]. Fikir, düşünce madenciliği olarak da adlandırılan duygu analizi, 2000'li yılların başından beri Doğal Dil İşlemede (DDİ) en aktif araştırma alanlarından biri olmuştur [2].

Sosyal medya insanlara görüşlerini, bilgilerini, deneyimlerini ve duygularını ifade etmeleri için çevrimiçi bir platform sağlamaktadır. Ancak sosyal medya etkileşimleri taciz edici, küfürlü ve hakaret içeren saldırgan içerikli yorumlar yapılan bir platform haline geldiğinde büyük bir sorun ortaya çıkmaktadır. Bu durumda kullanıcılar saldırgan içeriklere maruz bırakılmaktadır. Dolayısıyla kullanıcılar bu durumdan rahatsız olmaktadırlar. Saldırgan içeriklere maruz kalan kullanıcılar arasında 11-15 ergen yaş grubu azımsanmayacak sayıda bulunmaktadır [3]. Ergen yaş grubu için saldırgan içerikler psikolojik açıdan geriye dönülmesi mümkün olmayan zararlar vermektedir. Zorbalık veya nefret söylemi gibi potansiyel olarak zararlı olayların çoğu internetten önce gelmiş olsa da internetin erişimi ve kapsamı bu olaylara milyarlarca insanın hayatını etkilemek için benzeri görülmemiş bir güç ve etki sağlamıştır. Bu olayların web kullanıcıları için zihinsel ve psikolojik acı yaratmasının yanında insanları hayatları ile ilgili kararlar hakkında zorladığı ve aşırı durumlarda intihara meyillendirdiği de bilinmektedir [4]. Aynı zamanda şirketlerde sosyal medya platformlarında bu tür saldırgan içerikle çoğu kez karşılaşmaktadır [5]. Bu içeriklerle baş etmek sosyal medyanın hızı göz önüne alındığında ve kullanıcıların konuşma özgürlüğü hakkından ödün vermeden saldırgan dil kullanımına zamanında müdahale etmek oldukça zorlu bir görev olarak görülmektedir. Bu gibi sebepler ile saldırgan içeriklerin tespit edilmesine yönelik çalışmalar, gelecekte bu içeriklerin engellenebilmesi hususunda önem arz etmektedir. Çalışmada bu alana Türkçe dilinde bir veri seti kazandırılmıştır. Böylelikle derin öğrenme yöntemleri kullanarak sosyal medyada saldırgan içeriklerin tespiti alanına katkı sağlamak amaçlanmıştır.

Günümüzde sosyal ağ sitelerinde saldırgan içerik artmıştır. Bu sitelerin çoğunun, siber zorbalık, ırk, etnik köken, din, cinsiyet ve milliyet gibi korunan özelliklerine göre bireylere veya gruplara saldıran veya tehdit eden gönderileri yasaklayan nefret ve saldırgan dil söylemi politikaları vardır. Bu politikaya yönelik rapor sunan Twitter platformu 2020 rapor verilerini sunmuştur [6]. Twitter şirketinin 2020 yılı Temmuz- Aralık ayını kapsayan raporuna ilişkin veriler, Şekil 1.'de gösterilmektedir. Grafikte düşey eksen, Twitter şirketinin kendi politikaları çerçevesinde yaptırıma sebep olan içerikleri başlıklandırmıştır. Yatay eksen ise bu içerikleri üreten kişi sayısını göstermektedir. Bununla birlikte grafikte yer alan % oranlar temmuz ayından aralık ayına kadar belirtilen içeriklerin üretimindeki artışı göstermektedir. Bu artışların özellikle hassas içerik barındıran artışların bu kadar çok olması ve içeriklerin kaldırılması ve askıya alınması arasındaki farkın az olması bir içeriğin farklı kullanıcılar arasındaki potansiyel yayılma hızını göstermektedir. Sonuç olarak çevrimiçi ortamlarda bazı içerikler hızlı yayılmaktadır. Bu durum zamanında yanıt veren çözümlerin gerekli olduğunu göstermektedir. Aynı zamanda bu rapora göre sadece 6 ay da 1 milyon hesap askıya alınmıştır. Bu yaptırımların yarısından fazlası çocuk cinsel istismarı sebebiyle yapılmıştır. Tüm bunlar, otomatik yöntemlerle rahatsız edici dil algılama sorununu çözen doğru ve ölçeklenebilir bir çözüm bulma ihtiyacını doğurmaktadır. Böylelikle saldırgan dil söylemi tespiti alanında çeşitli duygu analizi araştırma çalışmaları ve TRAC, SemEval, OffensEval, GermEval gibi yarışmalar bu alana dikkat çekmektedir.



Şekil 1. Twitter yaptırımları Temmuz- Aralık 2020 raporu.

Türkçe zengin morfolojik yapıya sahip ve sondan eklemeli bir dildir. Literatürde üstün başarı gösteren İngilizce gibi dillere göre bu alanda yapılan çalışmalar Türkçe için zorlu bir görevdir. Örneğin; bir isim olan "masa" kelimesinden üretilebilecek sözcük sayısı 548 bin 390 iken "oku" kelimesinden üretilebilecek sözcük sayısının 1 milyon 461 bin 211 olduğu görülmüştür [7]. Buna ek olarak Türkçe' de bir kelimeyle anlatılan bir ifadenin İngilizce gibi dillerde bir cümleye tekâbül etmesi ve birleşik sözcüklerin kelimenin kökündeki anlamını yitirmesi gibi durumların Türkçe'nin DDİ alanında az çalışılmasına sebep olmaktadır. Ayrıca diğer dillere nazaran başarısını kısıtladığı da görülmüştür. Türkçe için bazı dillere göre kural tabanlı yaklaşımlar zorlu bir görevdir ve kaynaklar da azdır [7]. Örneğin; Türkçe için Senti-TurkNet [8] gibi kaynaklar ve 2.000 kelimeden oluşan bir kelime kümesi [9], Türkçe WordNet 15 bin kelimeden oluşan kavramsal sözlük [10] oluşturulmuştur. Bununla birlikte biçimbilimsel çözümleme [11], biçimbilimsel tekleştirme [12] grammer tabanlı çözümleme [13], ağaç yapılı derlem ile ekleri çözümleme [14] gibi kural tabanlı yöntemler bulunmaktadır. Ancak Türkçe'nin zengin morfolojik yapısı ve çok sayıda kelime oluşumunu sağlayan ek yapısı, morfolojik belirsizliklerin (kalem vb.) bulunması gibi zorluklar sebebiyle bu yaklaşımlar önerilmemektedir [15, 47]. Ayrıca sosyal medyada kullanılan ifadeler, dil bilgisi kurallarından uzak, yazım hataları, sözlükte yer almayan kelimeler ve kısaltmalar da oldukça fazladır. Bu durum duygu analizinde sınıflandırma problemlerine sebep olmuştur. Önerilen yöntem ise, alana özgü verileri kullanarak problemi çözmektir. Böylelikle Türkçe metinlerin sınıflandırılmasında başarılı sonuçlar elde edilmiştir [15]. Bu çalışmada, genişletilmiş derlem ile Türkçe dilinde birlikte kullanılan kelimelere ve cümleye göre anlam yakalamanın sınıflandırma başarımlarına etkisinin gösterilmesi amaçlanmıştır.

Türkçe dilinde yapılan çalışmaların az olmasının sebeplerinden biri de Türkçe dilinde yeteri kadar veri setinin bulunmamasıdır. Bu durum Türkçe dilinde yapılan DDİ çalışmalarını mukayeseden yoksun kılmaktadır. Böylelikle Türkçe dilinde yeteri kadar farklı veri setleri kullanılarak modellerin denenmediği ve bazı kalıplara oturtulamadığı görülmüştür. Bu sebeple Türkçe dilinde yapılan çalışmaların diğer bazı diller ile yapılan çalışmaların gerisinde kaldığı görülmüştür. Buna ek olarak literatürde yer alan Türkçe veri setlerinin bazılarının otomatik veya manuel olarak İngilizce dilinden çevrilip elde edildiği de görülmüştür [16-17]. Bu durumda çeviri metinlerde yaşanan özellikle sosyal medya platformlarında yer alan metinlerin doğası olan gayri resmi dil kullanımı, kısaltmalar, Türkçe' nin kendine özgü doğasının göz ardı edilip cümlenin anlamını ve etkisini yitirmesi [7] gibi durumlar ile karşılaşılabilinmektedir. Böylelikle bu çalışmada alana katkı sağlamasını amaçladığımız Twitter platformundan elde edilmiş yaklaşık 15 bin tweetten oluşan veri seti sunulmaktadır.

Çalışmada, alana özgü verilerin olduğu iki ayrı veri seti kullanılmıştır. Bunlardan biri alana özgü verilerden oluşan genişletilmiş derlemdir ve yaklaşık 1 milyon 866 bin 546 tweet metninden oluşmaktadır. İkinci veri seti ise 14 bin 752 tweet metninden oluşan veri setidir. Bu veri seti Python dili kullanılarak Twitter platformundan elde edilmiştir. Bu çalışmada, oluşturduğumuz veri setini etiketlemek için Zampieri ve arkadaşlarına benzer genel bir yaklaşım izlenmiştir [18-19]. Tanımlamalarımız her iki çalışmadan farklılıklar gösterse de her iki çalışmada [18-19] olduğu gibi sınıf etiketleri için hiyerarşik bir şema izlenmiştir. Bu şemanın ilk aşamasında tweet metni "saldırgan" ya da "saldırgan değil" olarak etiketlenmiştir. Tweet herhangi bir hakaret ya da aşağılama içermiyorsa saldırgan değil olarak etiketlenmiştir. Eğer metin "saldırgan" olarak tanımlanırsa, "hedefli" ve "hedefsiz" olarak etiketlenmektedir. Çalışmada "hedefsiz" olarak tanımlananlar net bir hedef olmaksızın her türlü saldırgan dil kullanımı içermektedir. Birini hedef almayan bir hakaret ya da küfür kullanımı, kişinin kendisini hedef alması veya cümlenin sonunda gelişigüzel kullanılan saldırgan ifadelerde bu etiket grubuna girmektedir. Buna ek olarak hakaret olarak algılanabilecek ama arkadaşça, şaka amaçlı kullanılan ifadeler içinde bu etiket kullanılmıştır. Bu durum özellikle çocuklar için uygunsuz olarak nitelendirilmektedir.

Eğer metin "hedefli" olarak tanımlanırlar ise grup, birey ve diğer olmak üzere etiketleme yapılmıştır. Gruba yönelik saldırgan dil toplulukları hedef almıştır. Tweet insanların aidiyetine yönelik bir hakaret ya da nefret söylemi içeriyorsa (cinsiyet, etnik kimlik, politik görüş, ya da din gibi genel ve kalıcı bir birlikteliği içeren gruplar) bu etiket grubuna girmektedir. Buna ek olarak birden fazla kişiye yönelik saldırgan ifadeler bu kategoride değerlendirilmiştir. Grup tanımına uymayan bir kişiye yönelik işlenen suç, bireyleri hedef alan bir suç olarak kabul edilmektedir. Bireye karşı yapılan saldırgan dil ise genellikle siber zorbalık, cinsel taciz suçunu içermektedir. Siber zorbalık genellikle çocuklara karşı işlenmekte ve birey üzerindeki etkileri de oldukça olumsuz olmaktadır. Cinsel taciz suçu ise cinsel arzu ve isteklere yönelik rahatsız edici sözleri kapsamaktadır [20]. Suçun hedefinin bu kategorilerden hiçbirine uymadığı durum da vardır. Bir organizasyon veya olay gibi insan olmayan bir varlığa yönelik suç buna örnek verilebilir. Bu gibi durumlarda hedef "diğer" olarak kabul edilmiştir. Kısaca saldırgan ve saldırgan değil olarak yapılan ikili sınıflandırma, ardından saldırgan değil, hedefli ve hedefsiz olarak üçlü sınıflandırma, sonrasında saldırgan değil, hedefsiz, birey, grup ve diğer olarak beşli sınıflandırma şeklinde veri seti etiketlenmiştir. Yapılan etiketleme ile ilgili bilgiler, Bölüm 3.'de ayrıntılı olarak işlenmiştir.

Çalışmada bu veri setleri kullanılarak word2vec-CBOW ile her bir tweet metninde yer alan kelimenin vektörel temsili elde edilmiştir. Bununla birlikte LSTM, GRU gibi güncel ve etkinliği kanıtlanmış derin öğrenme yöntemleri ile sınıflandırma işlemi yapılmıştır. Sınıflandırma performansını değerlendirme ölçümü olarak F1-makro değerine karar verilmiştir. Bu performans ölçümü literatürde bu alanda sıklıkla kullanılan ölçüm yöntemidir.

Yukarıda belirtilen zorluklar dikkate alındığında, bu çalışmanın temel katkıları şu şekilde sıralanabilir:

- Veri seti oluşturma kriterleri belirtilmiştir ve veri çeşitliliğinin nasıl sağlanıldığına değinilmiştir.
- Veri etiketleme kriterleri belirtilip yaşanan uyum probleminden bahsedilmiştir.

- Veri seti oluşturmak ve etiketlemenin psikolojik iyi oluşu kötü yönde etkilediği saldırgan dil metinlerine maruz kalan etiketleyiciler tarafından dile getirilmiştir.
- Popüler derin öğrenme modellerinin Türkçe saldırgan dil tespitinde başarım değerleri görülmüştür.
- Genişletilmiş derlem kullanımının sınıflandırma performanslarına olan iyileştirici etkisinin hangi oranda olduğu görülmüştür.
- Çoklu sınıflandırmada modelin performansının nasıl etkilendiği görülmüştür.

Bu çalışmada amacımız gelecekteki çalışmalara katkı sağlayacağını düşündüğümüz etiketlenmiş Türkçe veri setini alana kazandırmaktır. Buna ek olarak Türkçe dilinde saldırgan dil tespit etmek için çözümler sunmaktır. Böylelikle çoğunlukla kullanılan başarısı kanıtlanmış derin öğrenme yöntemlerinin performansları test edilerek bu modellerin Türkçe dilinde etkisi görülmüştür. Genişletilmiş derlem kullanımının bu yöntemlere iyileştirici etkisi karşılaştırılmıştır.

2. LİTERATÜR

Son on yılda, Web' de özellikle Facebook ve Twitter gibi popüler sosyal medya platformlarının ortaya çıkmasıyla birlikte kullanıcılar tarafından oluşturulan içeriklerde büyük bir artış olmuştur [21]. Artık çevrimiçi herhangi bir bilgi milyarlarca insana saniyeler içinde ulaşma gücüne sahiptir. Bu durum, yalnızca olumlu fikir alışverişiyle sonuçlanmakla kalmamış, aynı zamanda saldırgan ve potansiyel olarak zararlı içeriğin web üzerinde yaygın bir şekilde yayılmasına da yol açmıştır. Zorbalık veya nefret söylemi gibi potansiyel olarak zararlı ifadelerin çoğu internetten önce gelmiş olsa da internetin erişimi ve kapsamı bu olaylara milyarlarca insanın hayatını etkilemek için benzeri görülmemiş bir güç ve etki sağlamıştır. Bu olayların sadece web kullanıcıları için zihinsel ve psikolojik acı yaratmadığı, aslında insanları hayatları ile ilgili kararlar hakkında zorladığı ve aşırı durumlarda intihara meyillendirdiği bilinmektedir [4]. Ayrıca sözlü sohbetlerde, üniversite öğrencilerinin söylediği kelimelerin %0,5'inin küfür olduğu belirtilmektedir [22]. Twitter'da bu oran iki kattan fazla, %1,2 ve %7,7 küfür içeren tweetler olarak görünmektedir [23]. Saldırgan gönderilerin %9' undan daha büyük bir kısmının da küfür içermesi muhtemeldir [24].

Bu alandaki çalışmalarının çoğu, belirli bir saldırgan dil biçimini açıklar ve çoğu zaman, belirli bir uygulamada kullanılmak üzere tasarlanmıştır. Şimdiye kadar en yaygın uygulama, nefret söylemi algılamadır. Nefret söyleminin net bir tanımı olmamasına rağmen, tipik olarak ırk, etnik köken, cinsiyet, cinsel yönelim, sosyo-ekonomik sınıf, siyasi bağlantı veya din gibi özelliklere dayalı olarak bir grubu (veya bazen bir kişiyi) hedef alan saldırgan dili kapsar [25]. Bu çalışmalardan bazıları kapsamı belirli bir hedefe, genellikle ırk [26-27] kadınlar [26] ve mülteciler [28], belirli bir ideolojiye sahip nefret söylemi [29] veya tek bir önemli olayla ilgili nefret söylemidir [30]. Bir diğer yaygın saldırgan dil alt alanı da siber zorbalığın tespit edilmesidir [24, 31-34]. Nefret söyleminin aksine, siber zorbalığın hedefi genellikle tek bir kişidir ve bu genellikle bir çocuktur. Zorbalığın çevrimiçi versiyonu siber zorbalık, ciddi bir sağlık sorunu olarak kabul edilmektedir [35]. Bu nedenle, siber zorbalığın otomatik tespitinin tipik bir uygulaması, çocuklar için daha güvenli çevrimiçi iletişim sağlamaktır [36]. Uygulamalar farklı olsa da önemli bir örtüşme vardır. Örneğin, siber zorbalık genellikle nefret söylemi olarak kabul edilen ifadeler barındırır. Dahası, metinlerin hem dil özellikleri hem de bunları tespit etme yöntemleri benzerdir.

Genel kültürel ve ahlaki kurallar, saldırgan dili tanımlamaya fayda sağlasalar da bir ifadenin saldırgan olup olmadığı, çoğunlukla özneldir ve büyük ölçüde bağlama bağlıdır. Son zamanlarda yapılan birçok çalışmada sık sık gündeme getirilen bir nokta, saldırgan dilin tanımı ve alt kategorileri konusunda fikir birliğinin olmaması ve bunun sonucunda farklı derlemlerde etiketlemelerin uyumsuzluğudur. Birkaç istisna bir yana, saldırgan dille ilgili etiketleme görevlerinden herhangi biri üzerindeki yorumcuların aynı kararı almasındaki uyuşma durumu nispeten düşüktür. Saldırgan dil etiketlemesinde kullanıcılar arasında Wiegand ve diğerleri (2018) rapor $\kappa = 0.66$ ve Zampieri ve diğerleri (2019) %60 uyuşma bildirmiştir [18-19, 37]. Ancak saldırgan dil de net tanımlar ve sınıflandırmalar için daha nesnel olmalarını sağlama amaçlı girişimler bulunmaktadır [38-39]. Zampieri ve diğerleri (2019) tarafından yapılan çalışmaya göre saldırgan dilde hedef, bir birey veya ırkına, cinsiyetine, siyasi / ideolojik yakınlığına, dinine veya benzer bir mülke dayalı bir grup insandır. Eğer ki hedef bir bireyse; kategori genellikle siber zorbalık eylemlerini içerirken, hedef bir grup ise, muhtemelen bir nefret söylemi örneğidir. Ayrıca 2019 yılında Zampieri ve arkadaşları, hedefin bazen (açıkça) insanlar değil, örneğin bir organizasyon veya olay olduğunu da söylemektedir [18]. Hedef alınmayan hakaretler de yaygındır. Hedeflenmemiş suç, küfürlü veya müstehcen dil veya genel olarak, bir bireyi veya bir grubu rencide etme niyeti veya etkisi olmaksızın kullanılan ifadeleri içerir. Bu durum tipik olarak bir suça karşılık gelmez [18]. Bununla birlikte, çocuk kullanıcılar ve kurumsal şirket hesapları için bu dil biçimlerinden kaçınılmaktadır.

Öznel ve bağlama bağlı olduğu bilinen çeşitli saldırgan dil biçimlerinin otomatik olarak tanımlanması çalışmalarının yöntemleri ve başarı oranı da değişiklik göstermektedir. Genel olarak, başarı oranı da göreve ve veri setine bağlı olarak değişir. Çalışmaya en yakın otomatik tanımlama deneyi seti OfffensEval-2019/2020 [18, 40] ve GermEval [19] yarışmalarındaki saldırgan dil tanımlama görevleridir. Her ikisin de de birbirini izleyen görevler mevcuttur. OffensEval 2019' da ilk alt görev, saldırgan dili saldırgan olmayan dilden ayırmayı içermektedir. İkinci görev ise, verilen saldırgan belgenin hedeflenip hedeflenmediğini belirlemektir ve son olarak üçüncü alt görevler, hedef türü (grup, birey veya diğer) belirlemektir. Diğer bir benzer çalışma ise GermEval yarışmalarındaki saldırgan dil tanımlama görevidir. Bu çalışma da Struß ve arkadaşları tarafından yapılan Alman

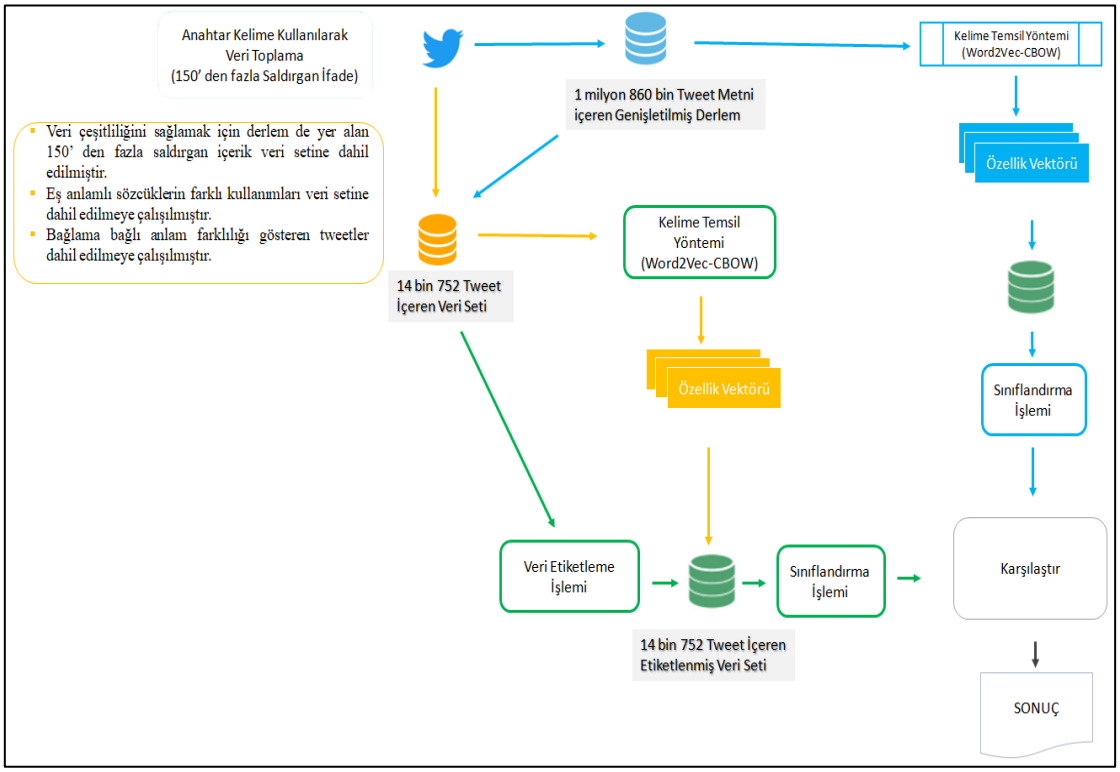
saldırgan dili tanımlama görevi, saldırgan dilin tanımlanması gibi ikili bir görevle başlar ve ardından saldırgan dilin türünü (küfür, taciz veya hakaret) tanımlamayı gerektiren ikinci bir görev izler. Alman saldırgan dili tanımlama görevinde çoklu etiket sınıflandırma söz konusudur [37]. Çalışmamıza benzer olarak her iki yarışmanın birinci görevin de ulaşılan başarı oranı yüksektir (sırasıyla; %82,9 F1 ve %76,8 F1). Diğer alt görevler de ise başarı oranı da düşmektedir (sırasıyla; %66 F1 ve %53 F1).

Saldırgan dili belirli konulara bölmeden (siber zorbalık, ırkçılık gibi) bunları bir bütün olarak gören derlemin, farklı saldırgan dil türleri arasındaki farklılık ve benzerlikleri anlaması daha olasıdır. Belirli bir konuya yönelik veri kümesinin eğitilmesi, saldırgan dil örneklerinin diğer biçimlerini belirlenen konuya yönelik saldırgan dil türünden ayırt etmekte başarısız olabilir. Örneğin, nefret söyleminin tanımlanmasını amaçlayan bir çalışma, durumu iyi temsil etmeyen bir derlem ile eğitilmişse, siber zorbalık içeren saldırgan dil biçimlerini nefret söylemi olarak karıştırabilir. Bu karışıklığın yaygın bir nedeni de bağlam eksikliğidir. Bazı tweetler bağlam olmadan saldırgan görünebilirken 'hayvan gibi güzel oynadı herif', doğru bağlamda iltifat olabilir. Diğer durumlarda, tweetler rahatsız edici görünmeyebilir. Ancak devam eden konuşmada veya dile özgü sosyal bağlamda saldırgan dil kullanılmış olabilir. '6s' in GS futbol takımına söylenmesi, ironi olarak yapılan iltifatlar veya 'insan olmayı başarabilseydin' gibi koşullu cümleler buna örnek verilebilir.

3. UYGULAMA

Çalışma üç aşamayı kapsamaktadır (Şekil 2.)

- Veri toplama aşaması
- Veri etiketleme aşaması
- Sınıflandırma aşaması



Şekil 2. Uygulama akış diyagramı.

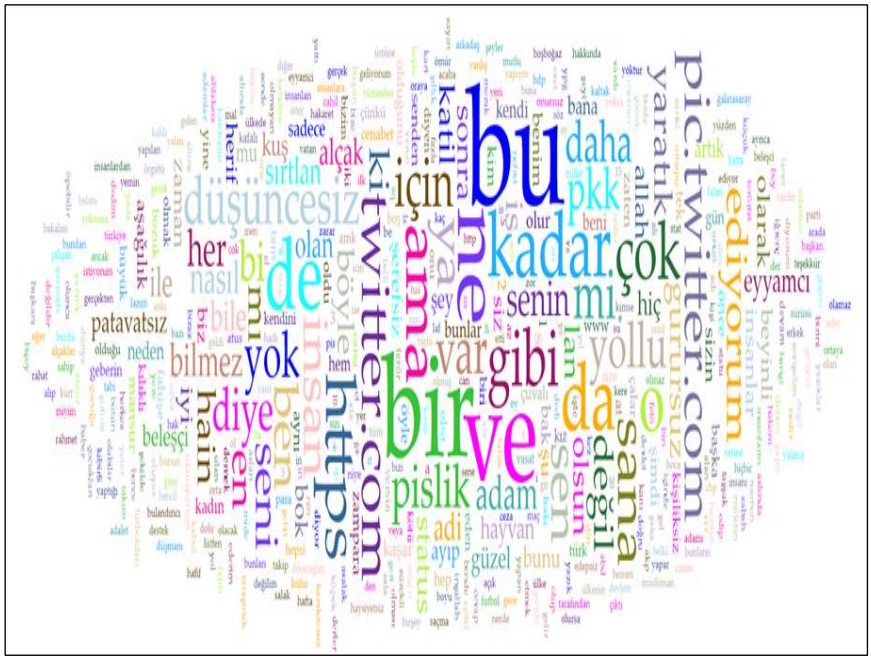
Veri toplama aşaması verilerin Twitter uygulamasından toplanma aşamasıdır. Toplanan verilerden iki ayrı veri seti elde edilmiştir. Bunlar 'genişletilmiş derlem' ve etiketlenmiş veri setidir. *Veri etiketleme aşamasında* veri setinde yer alan tweet metinlerini etiketleme işlemi yapılmıştır. Etiketleme yapılırken belirli kriterler kullanılmış ve 4 ayrı kişi tarafından etiketleme yapılmıştır. Etikete oy çokluğuna göre karar verilmiştir. Oy eşitliğinde ise ilgili tweet metni için 3 farklı kişiden görüş alınmıştır. *Sınıflandırma aşamasında* ise kullanılan veri setlerine kelime temsil yöntemlerinden word2vec-CBOW uygulanmıştır. Burada tweetler de yer alan boşluk gözetilerek her bir kelime veya ifade içeren varlık için vektörel temsil elde edilmiştir. Karşılaştırılacak olan genişletilmiş derlem ile etiketlenen veri setine LSTM, GRU sınıflandırma algoritmaları uygulanmıştır. Her bir algoritmaya uygun parametreler belirlenip model seçilmiştir. Her modelin performans değerlendirme ölçütleri elde edilmiştir. Bu ölçütler, genişletilmiş derlemin sınıflandırma algoritmalarının performansına etkisini göstermiştir. Buna ek olarak sınıflandırma algoritmalarını da karşılaştırabilmemizi sağlamıştır.

3.1. Veri Toplama

Bu başlıkta oluşturulan iki ayrı veri setinden bahsedilecektir. Sınıflandırma için oluşturulan veri setini etiketleme kriterleri örnekleriyle açıklanmıştır.

3.1.1. Veri Seti

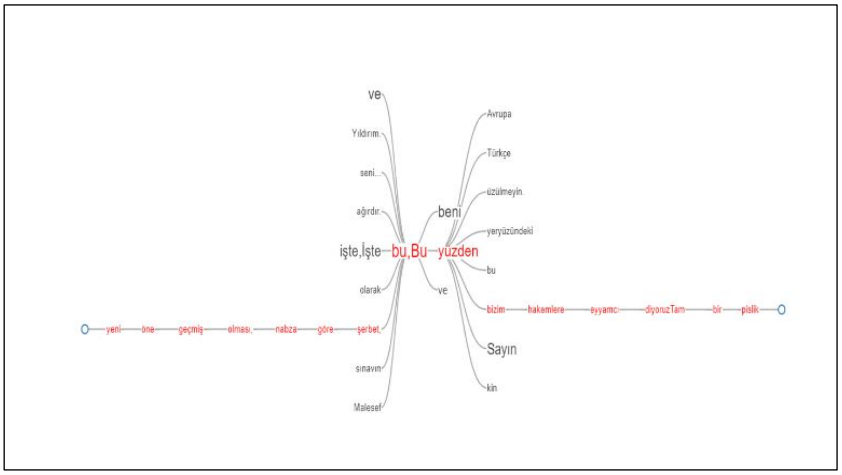
Bir veri setinin dört temel özelliği; gerçek dünyaya uygun, nispeten büyük, elektronik yollarla elde edilmiş ve alana özgü kriterleri barındırıyor olmasıdır [50]. Akademik dil, konuşma dili, sosyal medyada kullanılan gayri resmi dil, gazete yazıları, romanlar- hikayeler ve yasal davalar gibi kısa metinlerin yanı sıra metin türleri olarak bilinen çeşitli dil dizgesi içeren veri setleri oluşturulmaktadır. Öte yandan her bir dilin kendi modeline ve keskin dil kalıbına sahip olduğu görülmüştür [51]. Buna ek olarak veri setinin boyutuyla ilgili belirli kural yoktur. Ancak veri seti kullanılan hedef dil kalıplarının sistematik bir analizini yapacak kadar büyük olması gerekmektedir. Böylelikle bir kelimenin yalnızca kendisinin metne yön veremediği aynı zamanda çevresindeki kelimelerin öğrenmeye katkısı olduğu bulunmuştur. Bu da bir kelimenin anlamını ve faydasını artıran bir durumdur.
Bu çalışmada, Python programlama dili ile kendi oluşturduğumuz veri seti kullanılmıştır. Türkçe' de bilinen 150'den fazla saldırgan dil içeren kelime ve kelime öbekleri aratılarak veri seti oluşturulmuştur. 14.02.2020-17.06.2020 tarihleri arasında filtreleme yapılarak 1 milyon 860 bin tweetten oluşan bir genişletilmiş derlem elde edilmiştir. Ancak veri toplama periyodunun kısa olması tweet çeşitliliğinin yetersiz olmasına yol açabilmektedir. Bununla ilgili 2018 yılında Wiegand ve arkadaşlarının yaptıkları çalışmada Almanca olarak saldırgan dili temsilen veri seti oluşturmuşlardır ancak veri setinde mültecilere yönelik içeriklerin veri setine hâkim olduğu görülmüştür [19]. 2019 yılında bunun üstesinden gelebilmek için Strub ve arkadaşları bu veri setini iyileştirmek için kullanıcıya göre tweet metni toplamışlardır [37]. Ancak yine de belirli bir konunun veri setine hâkim olduğu görülmüştür. Bu dezavantajı gidermek için Python da geçmişe dayalı uzun bir periyodda arama kodu kullanılarak filtreleme yapılmıştır. Böylelikle genişletilmiş derlemde kullanılan tweetlerin tarihleri 5.09.2010- 17.06.2020 tarihleri arasındadır. Şekil 7. 2'de de görüldüğü gibi derlemden elde edilen kelime bulutunda herhangi bir konu yani birey grup veya olay veri setine hâkim değildir. Böylelikle bahsettiğimiz dezavantajın aksine veri setimizde konu çeşitliliğinin sağlanabildiği söylenebilir. Bu tarihler baz alınarak oluşturulan genişletilmiş derlemde 3 milyon 859 bin 322 adet eşsiz ifade bulunmaktadır. Her bir ifade için çeşitli boyutlarda vektörel temsiller elde edilmiştir. Derlem için word2vec olarak bilinen kelime temsil yöntemi kullanılmıştır. Böylelikle vektör uzayında anlamsal olarak yakın kelimeler bilinmektedir. Derlem kullanımının öğrenmeye bariz fayda sağladığı görülmüştür [41]. Çünkü derlem kullanımı, sınırlı verilere dayanmak yerine derlemde yer alan kelimelerin frekans dağılım özelliklerinin ve ilgili konunun gerçek dünyayı yansıtan dil kullanımının gözlemlenmesine ve incelenmesine izin vermektedir. Bununla birlikte belirli bir alana uygun oluşturulan derlem sözcük kalıplarına dayalı yeni duygu sözcüklerini keşfetmektedir.



Şekil 3. Genişletilmiş derlemden elde edilen kelime bulutu.

Şekil 3.'te genişletilmiş derlemde yer alan tweetler de en çok geçen 500 kelimenin kelime kullanım sıklığına göre boyutunun belirlendiği bir kelime bulutu elde edilmiştir. Örneğin bu-bir-ve-ama-için-kadar kelimeleri tweetler de oldukça fazla geçmektedir. Bu kelimeler veri temizleme işlemi yapılarak derlemden çıkarılabilir. Yapılan bazı çalışmalarda belirli kullanım sıklığının üzerinde olan kelimelerin veri ön süreç aşamasında temizlendiği

görülmüştür. Ancak çalışmada bu kelimelerden saldırgan içerik barındıran tweetler belirlenmiştir. Örneğin; veri setimizde de yer alan "bu kadar mı insansınız" tweet'i için 'bu-kadar' kelimeleri anlam ifade etmektedir. Bununla birlikte saldırgan içerikleri barındıran metinlerin dil kalıbının beklenildiği gibi gerçek dünyayı yansıtması için veri temizleme işlemi yapılmamıştır. Şekil 4'te derlemden elde edilen kelime ağacı modeli bu dil kalıbına örnektir. Burada birbirleriyle en çok kullanılan kelimelerin sırası gözetilerek bir kelime ağacı oluşturulmuştur. Şekil 4.'te de görüldüğü gibi anlamlı bir cümle elde edilebilmiştir ("*yeni öne geçmiş olması, nabza göre şerbet bu yüzden bizim hakemlere eyyamcı diyoruz Tam bir pislik* ").



Şekil 4. Genişletilmiş derlemden elde edilen kelime ağacı modeli.

Eğitimde kullanılacak olan veri seti, genişletilmiş derlemden ve Google Colab' dan 14 bin 752 tweet seçilerek elde edilmiştir. Seçilmiş olan tweetler genişletilmiş derlemden silinmiştir. Seçme işlemi yapılırken dikkat edilen durumlar:

- Birçok harfin bir araya gelerek anlamsız ifade oluşturduğu tweetler saldırgan dil için önemi belirtilmiştir ve diğer çalışmaların aksine veri setine dahil edilmiştir. Çünkü bu ifadeler ironi ve alaycı tutuma sahip metinlerde kullanılmaktadır [42].
- Veri çeşitliliğini sağlamak için derlem de yer alan 150' den fazla saldırgan içerik veri setine dahil edilmiştir.
- Eşsesli sözcüklerin farklı kullanımları veri setine dahil edilmeye çalışılmıştır. Örneğin; yollu *(i)* ve alçak *(ii)* kelimesi gibi.

i. *a. Sen yollu bir bi olduğundan mütevellit çabuk beğeniyor olabilirsin tabii, normal*- Saldırgan
b. Ucuz yoldan sonuca gitmek gibi bir hastalığa sahibiz%9x ımız – Saldırgan Değil
c. Telefonun elimdeki konumunu hiç değiştirmedigimden dolayı parmak uçlarım uyuşuyor sanki ve hafif yollu şarjım dibini çekiyor ayy ayy #TurkishArmysAlwaysLoveYouBTS – Saldırgan Değil
d. Ben de şaka yollu şeettim zaten – Saldırgan Değil
e. Orta yollu bir şey yaparlar – Saldırgan Değil

ii. *a. Pislik konuşmaya değmez aşağılık alçak* – Saldırgan
b. Ebru hanım, çok alçak gönüllü gerçekten – Saldırgan Değil
c. Yapılıyor, ancak zaman gerekmekte. Şu anda denenmiş sadece Hisar A var, o da alçak irtifa 15-20 km menzil – Saldırgan Değil

- Bağlama bağlı anlam farklılığı gösteren tweetler dâhil edilmeye çalışılmıştır *(iii)*.

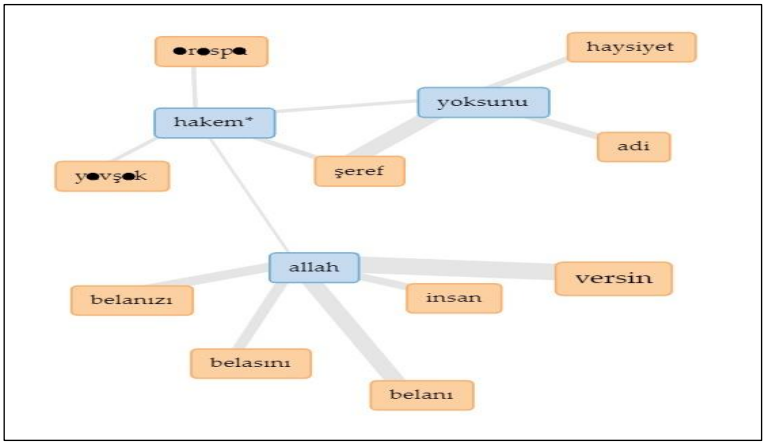
iii. *a. Hayvan gibi güzel oynadı herif* – Saldırgan Değil
b. Çık dışarı hayvan herif." B.K – Saldırgan
*c. Sen de az k*lt*k değilsin he* – Saldırgan

Etiketli veri setimizden elde edilen belirli kelimelerin anlamsal yakınlık göstermesi bazı durumlarda saldırgan dile maruz kalan grupları da göstermektedir. Örneğin Şekil 5 'hakem' kelimesiyle birlikte en çok kullanılan 5 kelimeden oluşturulmuştur. Buna göre ülkemizde hakemler saldırgan dile maruz kalan bir gruptur.

3.2. Veri Setinin Etiketlenmesi

Bu çalışmada çoklu sınıflandırma modeline uygun etiketleme işlemi gerçekleştirilmiştir. İlk olarak bir tweet metni saldırgan veya saldırgan değil olarak ayrılmıştır. Daha sonra saldırgan olan tweet metninin bir hedefe yönelik bir içerik barındırıp barındırmadığına bakılmıştır. Bir hedef belirtmeyen tweet metni hedefsiz olarak etiketlenmiştir. Hedefsiz tweetler saldırgan içerik barındırmaktadır. Bu sebeple çocuklar için filtreleme görevinde kullanılabilir. Bununla birlikte bu içerikler şirketlerin kurumsal yapısına da zarar vermektedir. Tweet metni belirli bir hedefe yönelik ise, hedefli olarak etiketlenmiştir. Hedefli olarak etiketlenen tweetler, birey, grup ve diğer olarak tweet

metninin hedef aldığı türlere göre ayrılmaktadır. Buradaki "grup" tanımı, herhangi bireyler topluluğunu (örn. mülteci, futbol takımı, güruh gibi) ya da birden çok kişiyi (örn. hainler) ifade etmektedir. Birey ise, grup tanımına uymayan kişiyi hedef alan suçlardır. Suçun hedefinin bu durumlardan herhangi birine uymadığı durumlar vardır, tipik olarak bir organizasyon veya olay gibi insan olmayan bir varlığa yönelik suçları kapsamaktadır. Bu gibi durumlarda, hedef, diğer olarak işaretlenmektedir. Burada, Zampiere ve diğerlerinin etiketleme yöntemi izlenmektedir [18]. Ancak bazı tanımlar, bu çalışmadan farklılıklar göstermektedir. Bununla birlikte saldırgan dil ile ilgili tanımlamaların kullanılan öğrenme yöntemlerin performansını iyileştirici etkisi olduğunu bilinmektedir [43].

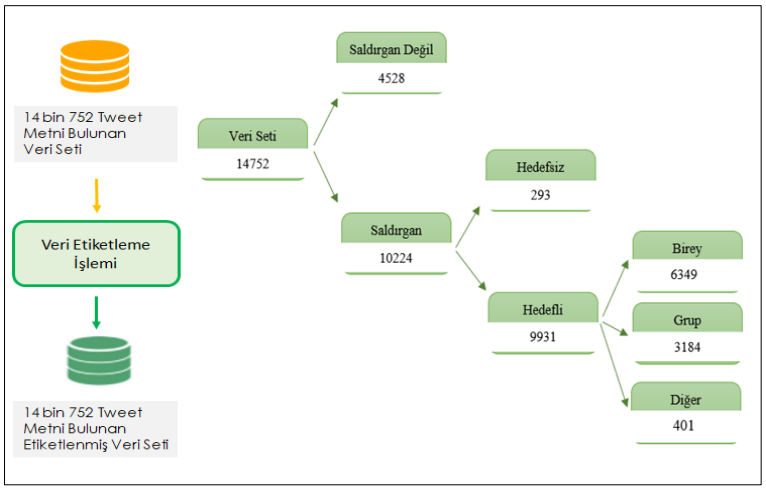


Şekil 5. '*hakem*' kelimesinin anlamsal yakınlıklarının gösterimi.

Saldırgan içeriklerin belirlenmesi öznel yargı içermektedir. Bu sebeple etiketleyiciler tarafından bir metnin saldırganlığı belirlenirken kişiye göre değişiklik göstermektedir. Zampiere ve diğerleri tarafından yapılan çalışma da etiketlemelerde kişiler arasında %60 uyum olduğu gösterilmiştir [40]. Wiegand ve diğerleri ise %66 uyuşma olduğunu belirtmiştir [19]. Bu alanda yapılan çalışmalar da bu durumun dezavantajından bahsetmiştir. Benzer olarak bu çalışmada da etiketlemecilerin saldırgan metinleri etiketlerken uyum sorunu yaşadığı görülmüştür. Örneğin; "Devleti ve milleti hedef aldı! İşte alçak saldırının faturası". Bu tweet metninde yer alan "alçak" kelimesi bir olayı nitelemesi sebebiyle saldırgandır. Ancak aynı zamanda terör olaylarını ifade etmek adına kamuoyunda kullanılan genel bir ifadedir. Bu tweet metni genel bir ifadeyi içermesi sebebiyle saldırgan değil olarak etiketlenmesine yorumlayıcılar tarafından karar verilmiştir.

Etiketleme işlemi, 4 kişi tarafından yapılmıştır. Bu kişiler en az lisans düzeyinde öğrenim görmüştür ve anadilleri Türkçe'dir. Etiketlemeler çoğunluğun kararına uygun olarak yapılmıştır. Etiketlemeciler arasında eşit oy söz konusu olduğunda ise farklı 3 etiketleyici oylamaya dahil edilerek karar verilmiştir. Bununla birlikte çalışmada yukarıda belirtildiği gibi daha net ve basit etiketleme kriterleri koyulmuştur. Böylelikle tweet metnini etiketlerken daha tutarlı olmak ve genel yargı içeren etiketlenmiş veri seti elde etmek hedeflenmiştir.

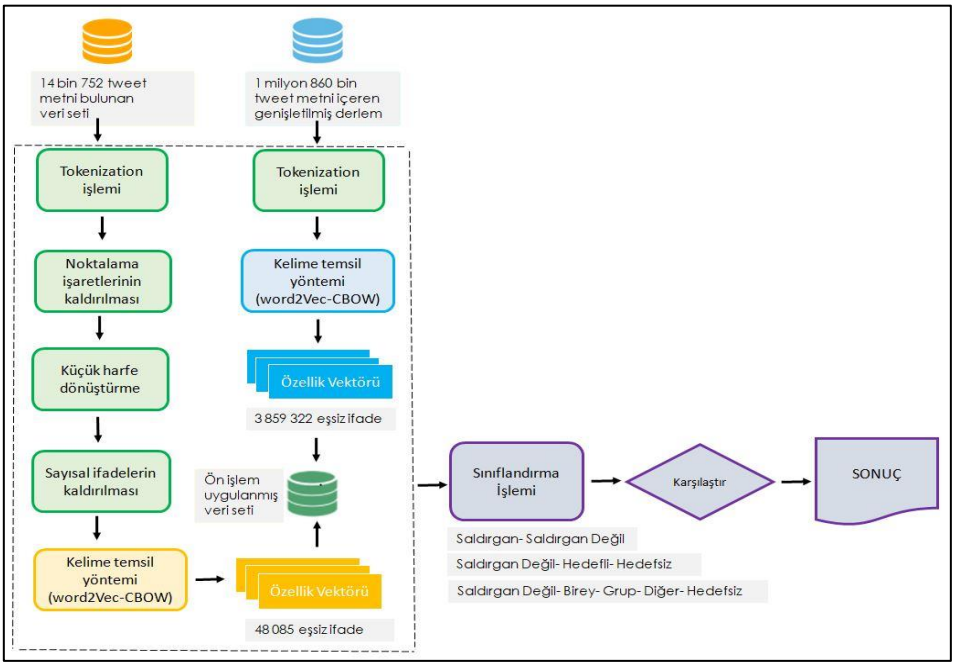
Veri setinin boyutuyla ilgili belirli kural yoktur. Ancak çalışmada kullanılan dil kalıbını öğrenmesi beklenmektedir. Çalışmada Türkçe tweetler kullanılmıştır. Veri setimizdeki tweet sayısı 14752'dir. Burada 'Saldırgan Değil' etiketli tweet sayısı 4528 iken 'Saldırgan' etiketli tweet sayısı 10224' dür. Saldırgan etiketli tweetlerin 9931'i 'Hedefli', 293'ü 'Hedefsiz' olarak etiketlenmiştir. Hedefli olarak etiketli tweetlerin 6346'sı 'Birey', 3184'ü 'Grup' ve 401'i 'Diğer' olarak etiketlenmiştir (Şekil 5).



Şekil 6. Veri etiketleme aşaması.

3.3. Veri Ön İşleme

Verilerin sınıflandırma performansını iyileştirmek için bazı ön işleme adımlarını gerçekleştirmek önemlidir. Verilerin sınıflandırma performansını iyileştirmek için bazı ön işleme adımlarını gerçekleştirmek önemlidir.Çalışmada yer alan veri seti, doğrudan Twitter' dan çıkarılan tamamen ham verilerden oluşmaktadır. Bunlardan faydalı bilgiler çıkarmak için veri önişleme aşamasına ihtiyaç duyulmaktadır. Ancak 'hashtag', 'tamamı büyük harf', 'uzatılmış', 'tekrarlanan', 'vurgu', 'sansürlü' 'emoji' içeren tweet metinleri herhangi bir ön işlem uygulamasına dahil edilmemiştir. Çünkü bunlar güçlü bir duygu veya alaycılık ve ironi ifadelerini temsil etmek için kullanılmaktadır [42]. Bu sebeple tweet metinlerine yukarıda bahsedilen ön işleme aşamaları uygulanmamıştır.Ancak sınıflandırma algoritmaları tarafından daha kolay öğrenebilir hale getirmek için sayısal forma dönüştürülmüştür. Bunun için kelime temsil yöntemlerinden olan word2vec-CBOW'u kullanarak kelime dizileri yerine gerçek sayı vektör dizileri elde edilmiştir. Bununla birlikte 14752 tweet metni içeren veri seti için Türkçe verilerde iyileştirici etkisi bilinen [52] tüm karakterlerin küçük harfe dönüştürülmesi, sayısal karakter içeren ifadelerin ve noktalama işaretlerinin çıkarılması ön işlemi uygulanmıştır (Şekil 7).



Şekil 7. Veri önişleme aşaması.

Kelime Temsil Yöntemi

Derin öğrenme modelleri veri yoğun çalışma modelini benimsemeleridir. Bu modeller de denetimli öğrenme kullanılırken çok büyük miktarda veri gerektirir. Çoğu DDİ görevi için veri setleri yalnızca birkaç bin insan etiketli örnek içermektedir. Bu, veri kümelerinde derin öğrenme modellerinin sınıflandırma performansını kısıtlamıştır. Bu sorunun çözümü, önceden eğitilmiş olarak bilinen bir yöntemle gelmiştir [44][46][48]. Bu yöntemin arkasındaki fikir, metnin genel temsilini öğrenmek ve çok miktarda etiketlenmemiş metinden oluşan genişletilmiş derlemi eğitmektir. Böylelikle alana özel söz dizimsel ve anlamsal özellikleri anlamaktadır. Etiketlenmemiş tümceler kolayca erişilebilir durumdadır ve bu nedenle daha uzun dönemler içeren ve daha kapsamlı verileri eğitmekte kullanılmaktadır. Etiketlenmemiş metinden oluşan derlemin birçok DDİ görevine katkısı görülmüştür [45]. Çalışmada etiketlenmemiş verilerden oluşan genişletilmiş derlem kullanımıyla elde edilen 1 milyon 860 bin tweetten 3 milyon 859 bin 322 eşsiz ifade elde edilmiştir. Böylelikle genişletilmiş derlem kullanımının sınıflandırma performansına katkısı görülmüştür.

4. BULGULAR

Türkçe gibi zengin morfolojik dil yapısına sahip olmak, bu alanda yapılan çalışmaların sayısını da etkilemiştir. Öyle ki Türkçe dilinde saldırgan içeriklerin tespit edilmesine yönelik çalışmanın oldukça az olduğu görülmüştür. Çalışmada bu alana katkı sağlamak amaçlanmıştır.

Çalışma da alana özgü verilerden oluşan bir genişletilmiş derlem elde edilmiştir. Çalışmanın bu bölümünde ise bu derlemden oluşturulan kelime temsillerinin başarıya katkısını gözlemlenmiştir. Buna ek olarak çalışmada hem ikili hem de çoklu sınıflandırma işlemleri yapılmıştır. Bilinebildiği kadarıyla çalışma, saldırgan içeriğin tanımlanmasına yönelik Türkçe dilinde ilk çoklu sınıflandırma yapan çalışmadır. Burada LSTM, GRU modellerinin Türkçe dilinde sınıflandırma başarısı karşılaştırılmıştır. Bu sınıflandırma işlemleri, metin sınıflandırma çalışmalarında en çok kullanılan derin öğrenme modelleri arasında yer almaktadır [49].

4.1. İki Sınıflı Verilerin Sınıflandırma Başarımlarının Değerlendirilmesi

Bu bölümde ilk olarak saldırgan ve saldırgan değil olarak sırasıyla '1' ve '0' olarak etiketlenmiş iki sınıflı verilerin sınıflandırma başarımları değerlendirilmiştir. Burada 14752 tweetten oluşan veri setinin 11802'si eğitim veri seti iken 2950'si test veri setidir. Rastgele seçilmiş olan test veri setinin 913'ü saldırgan olmayan, 2037'si saldırgan içerik barındırmaktadır. Böylelikle Tablo 1' de görülen başarım değerleri, saldırgan içeriğin tespit edilebilme performansını göstermektedir. Bununla birlikte modeller eğitilirken alana özgü elde edilen genişletilmiş derlem kullanımının modellerin sınıflandırma performansına etkisi gösterilmiştir. Burada sınıflandırma performans değerlendirme ölçütlerinden F1-makro değeri baz alınmıştır.
Veri setinden elde edilen word2vec-CBOW kelime temsilleri kullanılarak iki sınıflı verilerin sınıflandırma başarımları F1-makro değeri LSTM %85,19 iken GRU %87,99 arasındadır (Tablo 1). Genişletilmiş derlemin sınıflandırma performansını yaklaşık %10 artırabildiği görülmüştür (Tablo 1). Buna ek olarak genişletilmiş derlem kullanılması değerlendirme sonuçlarını da yaklaştırmıştır. Bu durum genişletilmiş derlem kullanımının seçilen derin öğrenme yönteminin etkisini azalttığını göstermektedir. Ayrıca GRU ve LSTM arasında sınıflandırma performanslarına bakıldığında belirgin bir üstünlük görülmemiştir.

Tablo 1. Sınıflandırma işlemi F1-makro değer gösterimi.

	Veri Seti	**Veri Seti+ Genişletilmiş Derlem**
LSTM	85,19	94,21
GRU	87,99	***94,49***

4.2. Çok Sınıflı Verilerin Sınıflandırma Başarımlarının Değerlendirilmesi

Çalışmada sınıflandırma işlemi üç aşamalıdır. İlk aşamada veri seti saldırgan ve saldırgan değil olarak etiketlenmiştir. Burada ikili sınıflandırma söz konusudur. İkinci aşamada ise saldırgan olarak etiketlenen veriler, hedefli ve hedefsiz olarak alt sınıflara ayrılmıştır. Bu aşamada saldırgan olan tweet metninin bir hedefe yönelik olup olmadığına bakılarak etiketleme işlemi yapılmıştır. Üçüncü aşama da ise saldırgan olup hedefli olarak etiketlenen veriler birey, grup ve diğer olarak alt sınıflara ayrılmıştır. Burada saldırgan içeriğin yöneldiği hedefe göre etiketleme yapılmıştır. Bu ek olarak çoklu sınıflandırma yapılırken üstün sınıflandırma başarımı sebebiyle GRU modelinin performansına bakılmıştır.
İkinci aşama sınıflar bazında başarım değerlendirilmesi:
İkinci aşamada, ilk olarak hedefsiz, hedefli ve saldırgan değil sırasıyla '0', '1' ve '2' olarak etiketlenmiş çok sınıflı verilerin sınıflandırma başarımları değerlendirilmiştir. Burada 14752 tweetten oluşan veri setinin 11802'si eğitim veri seti iken 2950'si test veri setidir. Rastgele seçilmiş olan test veri setinin 913'ü saldırgan olmayan, 2037'si saldırgan içerik barındırmaktadır. 2037 saldırgan içerik barındıran tweetten 66'sı hedefsiz, 1971'i hedefli olarak etiketlenmiş verilerdir. Çoklu sınıflandırmayla ise F1-makro değeri %56,17'e ulaşmıştır. Genişletilmiş derlem kullanımıyla F1-makro değeri %15 oranında artış göstererek %71,97'e ulaşmıştır (Tablo 2).

Tablo 2. Çoklu sınıflandırma işlemi F1-makro değer gösterimi

	Veri Seti	**Veri Seti+ Genişletilmiş Derlem**
GRU	56,17	***71,97***

Üçüncü aşama sınıflar bazında başarım değerlendirilmesi:
Üçüncü aşamada, ilk olarak hedefsiz, birey, grup, diğer ve saldırgan değil sırasıyla '0', '1', '2', '3' ve '4' olarak etiketlenmiş çok sınıflı verilerin sınıflandırma başarımları değerlendirilmiştir. Burada 14752 tweetten oluşan veri setinin 11802'si eğitim veri seti iken 2950'si test veri setidir. Rastgele seçilmiş olan test veri setinin 913'ü saldırgan olmayan, 2037'si saldırgan içerik barındırmaktadır. 2037 saldırgan içerik barındıran tweetten 66' sı hedefsiz, 1971'i hedefli olarak etiketlenmiş verilerdir. 1971 hedefli içeriğin 1233'ü birey, 646 grup ve 92'si diğer olarak etiketlenmiş verilerdir.
Üçüncü aşama çoklu sınıflandırma başarım değerleri karşılaştırılmıştır (Tablo 3.). Burada genişletilmiş derlem kullanımıyla F1-makro değerinde %14 oranında iyileşme sağlayarak %54,10 başarım değeri elde edilmiştir. Bu değer modelin yeterince iyi olmadığını göstermektedir. Burada dengesiz veri problemine yönelik yöntemlerin değerlendirme performansını artıracağı düşünülmektedir. Ancak bu yöntemlerin değerlendirilmesi bu çalışmanın dışındadır.

Tablo 3. Çoklu sınıflandırma işlemi F1-makro değer gösterimi

	Veri Seti	**Veri Seti+ Genişletilmiş Derlem**
GRU	39,49	***54,10***

5. SONUÇLAR

Çalışmada Twitter platformundan elde edilen metin verileri sınıflandırılmıştır. Sınıflandırma işlemi üç aşamalıdır. Birinci aşamada saldırgan ve saldırgan değil olarak ikili sınıflandırma yapılmıştır. İkinci aşamada saldırgan olan tweet metni hedefli ve hedefsiz olarak ayrılmıştır. Böylelikle bu aşamada hedefli, hedefsiz ve saldırgan değil olarak çoklu sınıflandırma yapılmıştır. Üçüncü aşamada ise, hedefli olan tweet metni birey, grup, diğer olarak ayrılmıştır. Bu aşamada ise saldırgan değil, hedefsiz, birey, grup ve diğer olarak çoklu sınıflandırma yapılmıştır. Ayrıca bilinebildiği kadarıyla saldırgan dil tespitine yönelik bu alanda yapılan çalışmalara bakıldığında bu çalışma Türkçe dilinde çoklu sınıflandırma yapan ilk çalışmadır. Buna ek olarak çalışmada Türkçe' de bilinen 150'den fazla saldırgan dil içeren kelime ve kelime öbekleri aratılarak veri seti ve genişletilmiş derlem oluşturulmuştur. Burada veri çeşitliliği sağlamak için geçmişe dayalı arama yapılmıştır. Veri setinde yer alan 14752 tweet metni yukarıda belirtilen sınıflandırma kriterlerine göre manuel olarak etiketlenmiştir. Burada saldırgan dil etiketleme işleminin öznel yargılar barındırmasından kaynaklı etiketleyiciler arasında uyum problemi ile karşılaşılmıştır. Bu problemin çözümünde oy sistemi ile çoğunluğun kararına bakılmıştır. Ayrıca 1 milyon 860 bin alana özgü tweet metinlerinden oluşan genişletilmiş derlem elde edilmiştir. Böylelikle genişletilmiş derlemde word2vec-CBOW yöntemi kullanılarak elde edilen kelime temsillerinin sınıflandırma performansına katkısı görülmüştür. İki sınıflı verilerin sınıflandırma başarımları F1-makro değeri LSTM %85,19 iken GRU %87,99 arasındadır. Genişletilmiş derlem kullanımı LSTM modeli sınıflandırma performansını yaklaşık %10 artırarak %94,21'e ulaştırmıştır. Bununla birlikte GRU modelinin genişletilmiş derlem kullanımı ile sınıfladnırma performansı %94,49'tür. Bu iki model arasında belirgin bir üstünlüğün olmaması ile birlikte çoklu sınıflandırma yapılırken daha üstün performans gösteren GRU modeli tercih edilmiştir. Böylelikle GRU modelinin performansı genişletilmiş derlem kullanımının etkisiyle çoklu sınıflandırmada sırasıyla %15,08 artış ile %71,97 ve 14,61 artış ile de %54,10 olmuştur. Sonuç olarak çalışmada genişletilmiş derlemin sınıflandırma performanslarına iyileştirici etkisi görülmüştür.

Yazarların Katkısı

Yazarlar çalışmaya eşit oranlı katkı sunmuşlardır.

Çıkar Çatışması

Makale yazarları, aralarında herhangi bir çıkar çatışması olmadığını beyan ederler

KAYNAKÇA

[1] E. Cambria, S. Poria, A. Gelbukh, and M. Thelwall, "Sentiment Analysis Is a Big Suitcase", IEEE Intelligent Systems, vol. 32, no. 6, pp. 74–80, 2017.

[2] B. Liu, "Sentiment analysis and opinion mining", Synthesis Lectures on Human Language Technologies, vol. 5, no. 1, pp. 160-167, 2012 .

[3] W. Craig, M. Boniel-Nissim, N. King, S. D. Walsh, M. Boer, P. D. Donnelly, and R. Van den Eijnden, "Social media use and cyber-bullying: a cross-national analysis of young people in 42 countries", Journal of Adolescent Health, vol. 66 no. 6, pp. 100-108, 2020.

[4] S. Hinduja, and J. W. Patchin, "Bullying, cyberbullying and suicide", Archiands of suicide Research, vol. 14, no. 3, pp. 206-221, 2010.

[5] C. Newberry, "36 Twitter Stats All Marketers Need to Know in 2021", https://blog.hootsuite.com/twitter-statistics/ (Erişim Tarihi: Nisan 12, 2022).

[6] Twitter, Rules Enforcement, https://transparency.twitter.com/en/reports/rules-enforcement.html#2020-jul-dec (Erişim Tarihi: Haziran 12, 2021).

[7] K. Oflazer, "Türkçe ve Doğal Dil İşleme", Türkiye Bilişim Vakfı Bilgisayar Bilimleri ve Mühendisliği Dergisi, vol. 5, no. 2, 2016.

[8] R. Dehkharghani, Y. Saygin, B. Yanikoglu, and K. Oflazer, "SentiTurkNet: a Turkish polarity lexicon for sentiment analysis", Language Resources and Evaluation, vol. 50, no. 3, pp. 667-685, 2016.

[9] R. Dehkharghani, B. Yanikoglu, D. Tapucu, and Y. Saygin, "Adaptation and use of subjectivity lexicons for domain dependent sentiment classification", In 2012 IEEE 12th International Conference on Data Mining Workshops, pp. 669-673, 2012.

[10] S. Stamou, A. Ntoulas, J. Hoppenbrouwers, M. Saiz-Noeda, and D. Christodoulakis, "Euroterm: Extending the eurowordnet with domain-specific terminology using an expand model approach", In Proceedings of the 1st International Global Wordnet Conference, 2002.

[11] K. Oflazer, "Two-level description of Turkish morphology", Literary and Linguistic Computing, vol. 9, no. 2, pp. 137-148, 1994.

[12] D. Z. Hakkani-Tür, K. Oflazer, and G. Tür, "Statistical morphological disambiguation for agglutinative languages", Computers and The Humanities, vol. 36, no. 4, pp. 381-410, 2002.

[13] K. Oflazer, and I. Kuruoz, "Tagging and morphological disambiguation of Turkish text", arXiv preprint cmp-lg/9407026, 1994.

[14] K. Oflazer, B. Say, D. Z. Hakkani-Tür, and G. Tür, "Building a Turkish treebank", In Treebanks, pp. 261-277, 2003

[15] Z. Özer, "The effect of normalization on the classification of traffic comments", Doctorate Thesis,

Karabük University, Computer Science Enstitute, Karabük, pp. 15-23., 2019.

[16] A. Safaya, E. Kurtuluş, A. Göktoğan, and D. Yuret, "Mukayese: Turkish NLP Strikes Back", Association for Computational Linguistics, pp. 846-863, 2022

[17] S. Yilmaz, and S. Toklu, "A deep learning analysis on question classification task using Word2vec representations", Neural Computing and Applications, vol. 32, no. 7, pp. 2909-2928, 2020.

[18] M. Zampieri, S. Malmasi, P. Nakov, S. Rosenthal, N. Farra, and R. Kumar, "SemEval-2019 Task 6: Identifying and Categorizing Offensive Language in Social Media (OffensEval)", In Proceedings of the 13th International Workshop on Semantic Evaluation, pp. 75–86.

[19] M. Wiegand, M. Siegel, and J. Ruppenhofer, "Oandrview of the GermEval 2018 shared task on the identification of offensiand language", In Proceedings of the GermEval 2018 Workshop at Konandns, pp. 1– 10, 2018.

[20] Ş. Sarmaşık, "İşyerinde cinsel taciz algılaması ve yönetim ilişkilerine etkisi hakkında bir araştırma", Yüksek Lisans Tezi, 2009.

[21] D. Reinsel, J. Gantz and J. Rydning, "Data Age 2025: The Evolution of Data to Life-Critical," www-content/our-story/trends/files/Seagate WPDataAge2025-March-2017 ((Erişim Tarihi: Ağustos 12, 2020).

[22] M. R. Mehl, and J. W. Pennebaker, "The sounds of social life: A psychometric analysis of students' daily social environments and natural conandrsations", Journal of Personality and Social Psychology, vol. 84, no. 4, p. 857, 2003.

[23] W. Wang, L. Chen, K. Thirunarayan, and A. P. Sheth, "Cursing in english on twitter", In Proceedings of the 17th ACM conference on Computer Supported Cooperatiand Work and Social Computing, pp. 415-425, 2014.

[24] J. M. Xu, K. S. Jun, X. Zhu, and A. Bellmore, "Learning from bullying traces in social media", In Proceedings of the 2012 conference of the North American chapter of the association for computational linguistics: Human Language Technologies, pp. 656-666, 2012.

[25] Ç. Çöltekin, "A corpus of Turkish offensiand language on social media", In Proceedings of the 12th Language Resources and Evaluation Conference, pp. 6174-6184, 2020.

[26] V. Basile, C. Bosco, E. Fersini, N. Debora, V. Patti, F. M. R. Pardo, and M. Sanguinetti, "Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter", In 13th International Workshop on Semantic Evaluation, pp. 54-63, 2019.

[27] I. Kwok, and Y. Wang, "Locate the hate: detecting tweets against blacks", Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence, pp. 1621-1622, 2013.

[28] B. Ross, M. Rist, G. Carbonell, B. Cabrera, N. Kurowsky and M. Wojatzki. "Measuring the reliability of hate speech annotations: The case of the european refugee crisis", In Proceedings of NLP4CMC III: 3rd Workshop on Natural Language Processing for Computer-Mediated Communication, vol. 17, no. 1, pp. 6–9, 2016.

[29] S. Jaki, and T. De Smedt, "Right-wing German hate speech on Twitter: Analysis and automatic detection", arXiv:1910.07518, 2019.

[30] P. Burnap, M. L. Williams, L. Sloan, O. Rana, W. Housley, A. Edwards, and A. Voss, "Tweeting the terror: modelling the social media reaction to the Woolwich terrorist attack", Social Network Analysis and Mining, vol. 4, no. 1, pp. 1-14, 2014.

[31] M. Dadvar, D. Trieschnigg, R. Ordelman, and F. de Jong, "Improving cyberbullying detection with user context" In European Conference on Information Retrieval, pp. 693–696, 2013.

[32] M. Dadvar, D. Trieschnigg, and F. de Jong, "Experts and machines against bullies: A hybrid approach to detect cyberbullies", In Canadian Conference on Artificial Intelligence, pp. 275–281, 2014.

[33] K. Dinakar, B. Jones, C. Havasi, H. Lieberman, and R. Picard, "Common sense reasoning for detection, preandntion and mitigation of cyberbullying", ACM Transactions on Interactiand Intelligent Systems (TiiS), vol. 2, no. 3, pp. 18, 2012.

[34] T. Nitta, F. Masui, M. Ptaszynski, Y. Kimura, R. Rzepka, and K. Araki, "Detecting cyberbullying entries on informal school websites based on category relevance maximization", In Proceedings of the Sixth International Joint Conference on Natural Language Processing, pp. 579-586, 2013.

[35] American Psychological Association, "APA resolution on bullying among children and youth", http://www.apa.org/about/goandrnance/ council/policy/bullying.pdf. (Erişim Tarihi: Haziran 20, 2020).

[36] Y. Chen, Y. Zhou, S. Zhu, and H. Xu, "Detecting offensiand language in social media to protect adolescent online safety", In 2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Confernece on Social Computing, pp. 71–80, 2012.

[37] J. M. Struß, M. Siegel, J. Ruppenhofer, M. Wiegand, and M. Klenner, "Oandrview of GermEval task 2, 2019 shared task on the identification of offensiand language", In Preliminary proceedings of the 15th Conference on Natural Language Processing (KONANDNS 2019), pp. 352– 363, 2019.

[38] Z. Waseem, T. Davidson, D. Warmsley, and I. Weber, "Understanding Abuse: A Typology of Abusive Language Detection Subtasks", In Proceedings of the First Workshop on Abusiand Language Online, pp. 78–84, 2017.

[39] J. Ruppenhofer, M. Siegel, and M. Wiegand, "Guidelines for IGGSA shared task on the identification of offensiand language." https://github.com/uds-lsv/GermEval-2018-Data/ blob/master/guidelines-iggsa-shared.pdf, 2018 (Erişim Tarihi: Kasım 20, 2020).

[40] M. Zampieri, S. Malmasi, P. Nakov, S. Rosenthal, P. Atanasova, G. Karadzhov, H. Mubarak, L. Derczynski, Z. Pitenis, and Ç. Çöltekin, "SemEval-

2020 Task 12: Multilingual Offensive Language Identification in Social Media (OffensEval 2020)", In Proceedings of the Fourteenth Workshop on Semantic Evaluation, pp. 1425–1447, 2020.

[41] Ş. Ş. Yılmaz, İ. Özer, ve H. Gökçen, "Türkçe Metinlerde Derin Öğrenme Yöntemleri Kullanılarak Duygu Analizi", In International Symposium of Scientific Research and Innovative Studies, vol. 22, pp. 971-982, 2021.

[42] E. Filatova, "Irony and Sarcasm: Corpus Generation and Analysis Using Crowdsourcing", In Lrec, pp. 392-398, 2012.

[43] M. Sharma, I. Kandasamy, and V. Kandasamy, "Deep Learning for predicting neutralities in Offensive Language Identification Dataset", Expert Systems with Applications, vol. 185, p. 115458, 2021.

[44] H. Qiu, Y. Zeng, T. Zhang, Y. Jiang, and M. Qiu, "FenceBox: A Platform for Defeating Adversarial Examples with Data Augmentation Techniques", ArXiv: 2012.01701, 2020.

[45] P. Bojanowski, E. Graand, A. Joulin, and T. Mikolov, "Enriching word andctors with subword information", Transactions of the Association for Computational Linguistics, vol. 5, pp. 135-146, 2017.

[46] I. Ozer, Z. Ozer, and O. Findik, "Noise robust sound event classification with convolutional neural network", Neurocomputing, vol. 272, pp. 505-512, 2018.

[47] Z. Ozer, I. Ozer, and O. Findik, "Diacritic restoration of Turkish tweets with word2vec", Engineering Science and Technology, an International Journal, vol. 21, no. 6, pp. 1120-1127, 2018.

[48] I. Ozer, Z. Ozer, and O. Findik, "Lanczos kernel-based spectrogram image features for sound classification", Procedia computer science, vol. 111, pp.137-144,2017.

[49] A. Ligthart, C. Catal, and B. Tekinerdogan, "Systematic reviews in sentiment analysis: a tertiary study", Artifitical Intelligence Revolution, vol. 54, pp. 4997–5053, 2021.

[50] L. Bowker, and J. Pearson, "Working with specialized language: A practical guide to using corpora", New York: Routledge, pp. 15-29, 2002.

[51] U. Römer, "The inseparability of lexis and grammar: Corpus linguistic perspectives", Annual Review of Cognitive Linguistics, vol. 7, no. 1, pp. 140-162, 2009.

[52] A. K. Uysal, and S. Gunal, "The impact of preprocessing on text classification", Information Processing and Management, vol. 50, no.1, pp. 104-112, 2014.