TOBB EKONOMİ VE TEKNOLOJİ ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

SOSYAL MEDYADA KULLANICI GİZLİLİĞİNİ KORUMAK İÇİN TARAF TESPİTİ GÖREVİNDE DÖNÜŞTÜRÜCÜ DİL MODELLERİNİ YANILTMA YÖNTEMLERİ

YÜKSEK LİSANS TEZİ

Dilara DOĞAN

Bilgisayar Mühendisliği Anabilim Dalı

Tez Danışmanı: Dr. Öğretim Üyesi Mücahid KUTLU

AĞUSTOS 2023

TEZ BİLDİRİMİ

Tez içindeki bütün bilgilerin etik davranış ve akademik kurallar çerçevesinde elde edilerek sunulduğunu, alıntı yapılan kaynaklara eksiksiz atıf yapıldığını, referansların tam olarak belirtildiğini ve ayrıca bu tezin TOBB ETÜ Fen Bilimleri Enstitüsü tez yazım kurallarına uygun olarak hazırlandığını bildiririm.

.

Dilara DOĞAN

ÖZET

Yüksek Lisans

SOSYAL MEDYADA KULLANICI GİZLİLİĞİNİ KORUMAK İÇİN TARAF TESPİTİ GÖREVİNDE DÖNÜŞTÜRÜCÜ DİL MODELLERİNİ YANILTMA YÖNTEMLERİ

Dilara DOĞAN

TOBB Ekonomi ve Teknoloji Üniversitesi
Fen Bilimleri Enstitüsü
Bilgisayar Mühendisliği Anabilim Dalı

Danışman: Dr. Öğretim Üyesi Mücahid KUTLU

Tarih: Ağustos 2023

Doğal dil işleme alanındaki heyecan verici yeni gelişmeler dillerin karmaşıklıklarının daha iyi anlaşılmasını, metinler üzerinden yapılan anlam çıkarımları ve analizlerle daha başarılı sonuçlar ortaya koyulmasını sağlamıştır. Doğal dil işleme modelleri için geniş veri kümeleri sunan sosyal medya platformlarının kullanımı her geçen gün artarak insanların günlük hayatlarının önemli bir parçasına haline gelmiştir. İnsanlar, sosyal medya platformları üzerinden paylaştıkları metinlerde duygularını, düşüncelerini, deneyimlerini ve kendileriyle ilgili kişisel birçok bilgiyi ifade edebilmektedir. Yapay zekâ modellerinin, bu verileri insanların takip edilmesinde kullanabilmesi, kullanıcılarda önemli gizlilik endişelerini de beraberinde getirmiştir. Bu tez çalışmasında, sosyal medya platformlarını kullanan bireylerin yapay zekâ modelleri tarafından tespit edilememeleri için yapabileceklerini araştırıyoruz. Araştırmamızda birçok konuda kullanıma açık olan taraf tespiti görevini çeşitli konulardaki Türkçe ve İngilizce veri kümeleriyle ele alıyoruz. BERT ve BERTurk tabanlı dönüştürücü modellerini, yanıltmak amacıyla yeniden ifade etme ve kasıtlı yazım hataları yapma tabanlı yöntemler öneriyoruz. Önerilen 13 farklı yöntemin modellerin performanslarını etkileme seviyelerine göre etkinliklerini araştırıyoruz.

Deneylerimiz sonucunda, yazım hataları karşısında BERT ve BERTurk tabanlı modellerin performanslarının belirgin bir şekilde düştüğü gösterilmiştir. Yazım hatalarına yönelik yöntemlerden iki dilde de en etkili yöntemlerin görsel olarak benzer karakterleri birbirleri yerine kullanma, boşluk ekleyerek kelimeyi bölme ve kelimelerdeki harflerin sıralarını karıştırma olduğu sonucuna ulaşılmıştır. Fakat bunula birlikte, yeniden ifade etme yöntemlerinin bu modellerin performanslarını etkileme konusunda başarılı olmadığı görülmüştür. Yöntemlerin uygulanmasında manuel ve otomatik olmak üzere iki farklı yöntem kullanılmıştır. Yöntemlerin otomatik uygulanması sonucunda elde edilen metinlerin hâlâ eski anlamlarını koruyarak okunabilir olması istenmiştir. İki değerlendirici tarafından bu kontroller sağlanmış olup harf sıralarını karıştırma, hashtag silme ve boşluk ekleme yöntemleri kullanılarak yapılan otomatik değişiklikler sonucunda okunurluğun azalması ve anlam değişmeleri gibi durumlar tespit edilmiştir. Bu sebeple bu yöntemlerin uygulanması konusunda daha dikkatli olunması gerektiği sonucuna ulaşılmıştır. Diğer bir nokta ise hashtag'lere dayalı yöntemlerde hashtag seçimleri oldukça önemli olup modellerin daha iyi performans göstermesine de sebep olabilmektedir. Bununla birlikte hashtag silme ve hashtag kullanılmaması çoğu durumda daha etkili sonuçlar vermiştir. Önerdiğimiz yöntemler ve elde ettiğimiz sonuçlar, bilgi ve gizliliklerini yapay zekâ modellerinden korumak isteyen kullanıcılar için yol gösterici nitelik taşımaktadır.

Anahtar Kelimeler: Taraf tespiti, Dönüştürücü modeller, Kullanıcı gizliliği.

ABSTRACT

Master of Science

METHODS OF DECEIVING TRANSFORMER LANGUAGE MODELS IN STANCE DETECTION TO PROTECT USER PRIVACY IN SOCIAL MEDIA

Dilara DOĞAN

TOBB University of Economics and Technology
Institute of Natural and Applied Sciences
Computer Engineering Science Programme

Supervisor: Asst. Prof. Dr. Mücahid KUTLU

Date: August 2023

The recent advances in natural language processing have led to a better understanding of language complexities and more successful outcomes in text analysis and comprehension models. Social media platforms, which offer large datasets for natural language processing models, have become an integral part of people's daily lives. Individuals express their emotions, thoughts, experiences, and various personal information through the text they share on social media platforms. However, the ability of artificial intelligence models to track and analyze this data has raised significant privacy concerns among users. In this thesis, we investigate what individuals using social media platforms can do to avoid being detected by artificial intelligence models. We address the task of stance detection on various topics using Turkish and English datasets. We propose methods for BERT and BERTurk-based transformer models to deceive the models by rephrasing and introducing intentional spelling errors. We investigate the effectiveness of the 13 different methods based on their impact on the models' performances. Our experiments demonstrate that intentional spelling error methods significantly reduce the performance of BERT and BERTurk-based models for stance detection. The most effective methods for spelling errors in both languages involve using visually similar characters, splitting words by adding spaces and shuffling the order of letters in words. However, paraphrasing methods are found to

be unsuccessful in affecting the models' performances. Two different approaches, manual and automatic, were used for applying the methods. The automatic application of the methods aimed to retain the readability and original meanings of the resulting texts. Two evaluators ensured these checks, and some methods were found to result in reduced readability and changes in semantics due to automatic modifications. Hence, caution is advised in applying shuffle, delete hashtag and adding space. Another point is that in methods based on hashtags, hashtag selections are very important and can cause models to perform better. However, removing or not using hashtags has been more effective in most cases. The proposed methods and the results obtained serve as a guiding reference for users who want to protect their information and privacy from artificial intelligence models.

Keywords: Stance detection, Transformer models, User privacy.

TEŞEKKÜR

Yüksek lisans eğitimim ve tez çalışmalarım boyunca tecrübeleriyle yoluma ışık tutan, desteğini ve emeğini hiçbir zaman esirgemeyen, birlikte çalışmaktan ve öğrencisi olmaktan gurur duyduğum değerli danışman hocam Dr. Öğretim Üyesi Mücahid KUTLU'ya teşekkürlerimi sunarım.

Kıymetli zamanlarını ayırarak bu tezi okuyan ve değerli görüşleriyle katkı sağlayan Prof. Dr. Muhammed Fatih DEMİRCİ ve Doç. Dr. Ahmet Murat ÖZBAYOĞLU hocalarıma ayrı ayrı teşekkürlerimi sunarım.

Lisans ve yüksek lisans eğitimimde kıymetli tecrübelerini ve desteklerini esirgemeyen TOBB Ekonomi ve Teknoloji Üniversitesi Bilgisayar Mühendisliği Bölümü öğretim üyesi Prof. Dr. Osman ABUL hocama ve üzerimde emeği geçen tüm öğretim üyelerine teşekkür ederim.

Beni TÜBİTAK ARDEB 3501 Programı'nın, 120E514 numaralı "Sosyal Medya Üzerinden Toplumsal Eğilim Tespiti" adlı projesi kapsamında destekleyen TÜBİTAK'a teşekkür ederim.

HAVELSAN'da birlikte çalışmaktan mutluluk duyduğum ve özellikle bu zorlu yüksek lisans sürecini başarmamı kolaylaştıran takım liderim Şenol Lokman ALDANMAZ'a teşekkür ederim.

TOBB Ekonomi ve Teknoloji Üniversitesi'nin hayatıma katmış olduğu güzelliklerden, lisans hayatımdan beri yanımda olan dostluk, destek ve sevgilerini her zaman yanımda hissettiğim Büşra GÜLTEKİN ve Esra ÜNAL'a teşekkür ederim.

Son ve en önemli olarak, beni bu günlere getiren, maddi manevi her konuda arkamda olan, bana daima inanan ve motive eden, destek ve sevgilerini her zaman hissettiren, hayatlarındaki en büyük yatırımı biz çocuklarına yapan, haklarını ödeyemeyeceğim, canım annem ve babama, desteği ve sevgisini her daim yanımda hissettiğim, fikirleriyle ufkumu açan, her zaman yüzümü güldüren canım kardeşime en kalbi duygularımla teşekkürlerimi sunarım.

İÇİNDEKİLER

ŞEKİL LİSTESİ

Sayfa

ÇİZELGE LİSTESİ

Sayfa

KISALTMALAR

AK Parti **:** Adalet ve Kalkınma Partisi
BERT **:** Çift Yönlü Kodlayıcı Temsilleri (Bidirectional Encoder Representations from Transformers)
BERTurk **:** Türkçe için Çift Yönlü Kodlayıcı Temsilleri **(**Bidirectional Encoder Representations from Transformers Models for Turkish)
CHP **:** Cumhuriyet Halk Partisi
GPT-3 **:** Üretken Önceden Eğitilmiş Transformatör 3 (Generative Pre-trained Transformer 3)
GPT-4 **:** Üretken Önceden Eğitilmiş Transformatör 4 (Generative Pre-trained Transformer 4)
HDP **:** Halkların Demokratik Partisi
SAADET **:** Saadet Partisi
LSTM **:** Uzun Kısa Süreli Bellek (Long Short-Term Memory)
NLTK **:** Doğal Dil İşleme Araç Kiti (Natural Language Toolkit)

1. GİRİŞ

Yapay zekânın son zamanlardaki gelişimiyle pek çok alanla birlikte doğal dil işleme alanındaki gelişmeler oldukça artmıştır. Doğal dil işleme, kişisel asistanlar [1] ve "chatbot"lardan [2] metinleri sınıflandırmaya [3], metinlerden anlam çıkarmadan [4] makine tercümesine [5], otomatik düzeltme ve tamamlama [6] sistemlerinden anket analizlerine [7], kişiselleştirilmiş reklamlardan [8] personel alım süreçlerine [9] kadar pek çok alanda aktif olarak kullanılarak günlük hayatımızda önemli bir yere sahiptir. Günlük hayatımızı etkileyen diğer bir belirgin değişim ise sosyal medya kullanımının artmasıdır. İçerik paylaşımı, kolay iletişim kurmayı sağlama, bilgi alma ve iş fırsatları sunması gibi pek çok sebeple sosyal medya kullanımı oldukça yaygınlaşmıştır. Bu sebeple, Facebook, Twitter, Instagram, LinkedIn ve daha pek çok platform, milyonlarca insanın günlük hayatının bir parçası haline gelmiştir. Kullanıcılar, bu platformlar üzerinde farklı içerikler üretir, paylaşır ve etkileşimde bulunur. Fotoğraf ve video paylaşımları, metin tabanlı gönderiler, yorumlar, beğeniler ve kullanıcıların sosyal ağları gibi pek çok veri, sosyal medya platformlarında toplanır. Platformlar üzerinden kullanıcılar hakkında pek çok veri elde edilebilmekte olup, bu büyük veri kümesi, yapay zekâ modellerinin eğitim sürecinde son derece değerlidir. Kullanıcılar tarafından kişisel bilgileri, hareketleri, arkadaş ve sosyal çevreleriyle etkileşimleri gibi pek çok veri yapay zekâ modellerini eğitmekte kullanılmaktadır. Modeller, kullanıcıların, yaş, cinsiyet [10], lokasyon [11], etnik kökenini [12] belirlemenin yanı sıra kişilerin çeşitli konulardaki görüşlerinin [13] ve hatta ruh sağlıklarındaki problemlerin tespit edilmesine [14] yönelik geliştirilebilmektedir.

Kutuplaşma analizi [15] ve doğruluk kontrolü [16] gibi konularda geliştirilen tahmin modelleri toplumsal açıdan faydalı olacak şekilde kullanılmaktadır. Bu gibi faydalı kullanım alanlarının yanı sıra, yüksek miktarda kişisel bilginin sosyal medya platformlarında yer alması, verilerin reklamcılık amacıyla kullanılması, çalınma riski, pek çok kişi tarafından takip edilip izlenebilmesi gibi sebeplerle insanların mahremiyetini tehlikeye atarak kullanıcılarda gizlilik endişesine sebep olmaktadır. Sosyal medya platformları tarafından kişisel bilgilerin toplandığı bilinmekle birlikte, bu durumu daha rahatsız edici kılan ise verilerimize herkesin erişmesidir.

Birçok kullanıcının, açık profil kullanması sonucu kullanıcılara ait veriler herkesçe erişilebilir hale gelmektedir. Bu durumda düşük seviyede kodlama ve doğal dil işleme bilgisine sahip kişiler bile BERT [17], GPT-3 [18], ve GPT-4 [19] gibi etkili dönüştürücü doğal dil işleme modellerini sosyal medya verileriyle eğiterek veya hassas ayar yaparak başarılı modeller elde edebilir. Örneğin, lokasyon tespitiyle ilgili bir modelin kullanımı kötü niyetli kişiler tarafından insan hayatlarını tehdit edebilecek şekilde suistimale açık bir konudur. Benzer şekilde taraf tespit modellerinin, ifade özgürlüğünün sınırlandırıldığı çeşitli coğrafya ve ülkelerde, kişilerin hedef gösterilmesi, çeşitli temel hak ve özgürlüklerinin ellerinden alınması konusunda kullanılması ortaya çıkabilecek olumsuz sonuçlardan bazılarıdır. Bu ve benzeri çeşitli konuların ortaya çıkardığı gizlilik endişeleri nedeniyle pek çok kullanıcı anonim kalabilmek amacıyla kendi isimleri yerine takma isimler, konum bilgilerini ise "evren, dünya, her yer" gibi gizli tutacak şekilde gizlemektedir. Ancak birçok bilgi, doğası gereği ait oldukları kişileri farklı yollarla ifşa etme riski taşıması ve yapay zekâ modellerinin açık ve örtülü ipuçlarını kullanabilmesi sebebiyle bu tarz önlemler de her zaman gizliliğin sağlanmasında etkili olmamaktadır.

Bu tez çalışmasında, kişilerin sosyal medya platformlarını kullanırken yapay zekâ modellerini aldatarak, korunmalarının hangi yöntemlerle mümkün olabileceği araştırılmıştır. Bu kapsamda kullanıcıların aleyhinde kullanılması durumunda tehdit niteliği taşıyabilecek bir görev olan taraf tespiti görevine odaklanılmıştır. Kullanıcıların gizliliklerini korumak amacıyla, İngilizce ve Türkçe dillerindeki farklı konulardaki veri kümeleri üzerinde yapay zekâ modellerini aldatmak amacıyla 13 farklı yöntem önerilmiştir. Modelleri yanıltma seviyelerine göre yöntemlerin etkinlikleri araştırılmıştır. Çalışma kapsamında odaklanılan araştırma sorularına (AS) şunlardır:

AS-1: Hangi metin değiştirme yöntemleri, metinlerde anlam değişimine neden olmadan yapay zekâ modellerini aldatmada etkili olmuştur?

Bulgu: Türkçe ve İngilizce veri kümeleri üzerinde yapılan deneyler sonucunda, yazım hatası tabanlı yöntemlerin BERTurk ve BERT tabanlı modelleri yanıltarak etkili yöntemler oldukları görülmüştür. Modelleri yanıltmada en etkili yazım hatası tabanlı yöntemler ise harf sıralarının değiştirilmesi ve boşluk eklenmesidir. Bununla birlikte İngilizce veri kümesinde metnin yeniden ifade edilmesi temelli yöntemlerin etkisiz olduğu görülmüştür.

AS-2: Modelleri aldatmaya yönelik yöntemlerin okunabilirlik ve anlam üzerindeki etkileri nelerdir?

Bulgu: Boşluk ekleme yöntemi İngilizce veri kümesi için okunabilirliği azalttığı oranda anlam değişimine neden olmuştur. İngilizce ve Türkçe veri kümelerinde ortak olarak harf sıralarının karıştırılması okunabilirliği düşürebilirken anlam değişimine sebep olmamıştır. Her iki dildeki veri kümeleri için ortak bir diğer bulgu ise hashtag silmenin anlam değişimine sebep olabilmesidir. Bunlar dışında kullanılan diğer yöntemlerin uygulanması sonucunda, okunabilirlik bozulmamış ve herhangi bir anlam değişimi olmamıştır.

Bu tez çalışmasının literatüre dört önemli katkısı bulunmaktadır. (1) Daha önce yapay zekâ modellerini aldatma üzerine yapılmış çalışmalar olmasına rağmen, bu çalışma, yapay zekâ modellerinin çalışma prensipleri hakkında hiçbir bilgi sahibi olmayan ve rastgele bir sosyal medya kullanıcısının nasıl gizliliğini koruyabileceğini araştırarak, aynı soruna farklı bir perspektiften yaklaşmaktadır. (2) İngilizce ve Türkçe dillerindeki farklı konulardaki veri kümeleri üzerinde uygulanan modelleri aldatma yöntemleri sonucunda benzerlik ve farklılıklar incelenmiştir. (3) Modelin performansını etkileyen en önemli metin parçacıklarının bulunmasına dair bir yöntem geliştirilmiştir. (4) Taraf tespiti modellerini aldatmak için belirlenen 13 farklı yöntemin etkileri araştırılmış ve bunlara göre kullanıcılara öneriler sunulmuştur.

Bu tezin organizasyonu şu şekilde yapılmıştır: Bölüm 2. ’de taraf tespiti ve farklı görevler için modelleri yanıltmak amacıyla daha önceki çalışmalarda kullanılan yöntem ve teknikler ele alınmıştır. Bölüm 3. ’de tez kapsamında yer alan kavramlarla ilgili temel bilgiler sunulmuş, Bölüm 4. ’de geliştirilen taraf tespit modelleri anlatılmıştır. Bölüm 5. ’de modelleri aldatmak için üzerinden değişikliklerin yapılacağı önemli/önemsiz kelimelerin seçilmesi için kullanılan yöntemlerden bahsedilmiştir. Bölüm 6. ’da modelleri aldatmak için kullanılan yöntemler, Bölüm 7. ’de veri kümeleri, deney sistemleri ve elde edilen sonuçlar detaylı olarak anlatılmıştır. Bölüm 8. ’de tez sonucunda elde edilen çıkarım, yorum ve tartışmalara yer verilmiştir.

2. LİTERATÜR ARAŞTIRMASI

Çeşitli görevler için çok sayıda araştırmacı düşmanca (adversarial) saldırılar ve savunma mekanizmaları üzerine çalışmıştır [20]. Önceki çalışmaların belirli bir kısmı farklı görevler için saldırı türlerini incelerken, farklı diğer bir konu olarak ise modellerin sağlamlıklarını (robustness) arttırmaya odaklanmıştır. Soru-cevap [21], diyalog oluşturma [22], makine tercümesi [23], duygu analizi [24-26] , toksiklik tespiti [27], istenmeyen e-posta tespiti [27] ve metin uygunluğu [24] görevleri için düşmanca saldırılara odaklanmışlardır. Liang vd. [28] güvenlik açıklarını keşfetmeye yönelmiştir. Jin vd. [24] ve Li vd. [26] eğitim için düşmanca örnekler üretmiş, Niu vd. [22] ve Muller vd. [29] modelleri gürültülü verilere göre iyileştirmişlerdir.

Bu tez çalışması kapsamında yapılan çoğu çalışmadan farklı olarak sosyal medya platformu kullanıcılarının potansiyel "saldırgan" olabilme ihtimalleri yerine potansiyel "kurban" durumunda olabilmeleri üzerine odaklanılarak, kullanıcıların gizliliklerini nasıl koruyabilecekleri araştırılmıştır. Modellere saldırı üzerine yapılan çalışmaların modelleri yanıltma üzerine sundukları yöntemlerle, bizim kullanıcıların gizliklerini korumak amaçlı modelleri yanıltmak için kullandığımız yöntemler bakımından benzerlik göstermesi sebebiyle uyguladığımız yöntemler daha önceki çalışmalardaki yöntemlerle karşılaştırılmıştır.

Kurita vd. [27] , duygu analizi, toksisite tespiti ve istenmeyen e-posta tespiti görevleri için Gu vd. [30] tarafından belirtilen çeşitli arka kapı (back door) saldırılarını karşılaştırmıştır. Saldırı başarılarının her doğal dil işleme görevi için değiştiğini göstermişlerdir. Yang vd. [31], yalnızca bir kelime vektörünü değiştirmenin, üzerinde değişiklik yapılmamış örnek verilerin sonuçlarında herhangi bir bozulmaya neden olmaksızın duygu analizi ve cümle çifti sınıflandırma modellerini "hack"lemek için etkili bir yöntem olabildiğini göstermiştir. Benzer şekilde, Dai vd. [25] ise eğitim verilerine tetikleyici cümleler ekleyerek gerçekleştirilen arka kapı saldırısının, LSTM tabanlı duygu analizi modeli üzerinde son derece etkili olduğunu göstermişlerdir. Chen vd. [32], belirli tetikleyicilerin (örneğin, kelimeler) kullanıldığı durumlarda modellerin başarısız olmasına neden olan arka kapı saldırılarını gerçekleştirirken, aynı zamanda

ilgili doğal dil işleme modelinin üzerinde herhangi bir değişiklik yapılmayan verilerde normal şekilde çalışabileceğini göstermiştir.

Sun vd. [33], doğal dil işleme modellerine yönelik doğal arka kapı (natural back door) saldırıları önermiştir. Önerilen bu saldırılar, insanlar ve dil düzeltme sistemleri tarafından fark edilmesi oldukça zor olan doğal yöntemlerle metin sınıflandırma problemleri için %83 gibi yüksek bir oranla başarılı olmuştur. Bu tez çalışmasında ise daha önceden eğitilmiş modellere erişimimizin olmadığını varsayarak kara kutu model olarak ele alınarak, tez kapsamında kullanılan BERT, Twitter-RoBERTa (Url-1) ve BERTurk (Url-2) modellerinin yanıltılması hedeflemiştir. Fakat geleneksel arka kapı saldırıları, yapay zekâ modellerini eğitim aşamasını etkileyebilecekleri varsayımı üzerine tasarlanmaktadır.

Birçok araştırmacı, kara kutu bir ortamda doğal dil işleme modellerinin savunmasızlıklarını inceleyip test verilerini değiştirerek bu konuda araştırma yapmıştır. Önceki çalışmada incelenen yöntemler temel olarak üç kategoriye ayrılmaktadır: 1) karakter düzeyinde değişikliklerde, karakterleri farklı şekillerde yazılmaktadır, 2) kelime düzeyinde değişikliklerde, kelimeler değiştirilir, çıkarılır veya eklenir ve 3) cümle düzeyinde değişikliklerde, yeni cümleler veya ifadeler eklenmekle birlikte mevcut cümleler ve ifadeler çıkarılabilir veya yeniden düzenlenebilir. Önceki çalışmada incelenen bu düşmanca saldırı yöntemleri örnekleriyle birlikte Çizelge 2.1'de gösterilmiştir.

Boşluk karakteri ekleme [26, 33] , kelime ortasındaki karakteri yer değiştirme [23] yöntemlerine bu çalışmada da yer verilmiştir. Dai vd. [25], Li vd. [26], Liang vd. [28] ve Morris vd. [34] bazı harfleri görünüş olarak benzerleriyle değiştirdiği gibi bu çalışmada da farklı bir dönüşüm tablosu kullanılarak benzer yöntem kullanılmıştır. Liang vd. [28], aynı karakter dönüşümlerinin yanı sıra ilgili eğitim veri setinde en sık kullanılmakta olan ifadeleri tespit ederek onları ekleyip çıkararak doğal dil işleme modellerini yanıltmaya çalışmışlardır. Bu tez kapsamında, eğitim verisindeki kelimelerin kullanılmak sıklıkları vb. dair herhangi bir analiz yapmaksızın, modeller için önemli olabildiği düşünülen hashtag'leri çıkarılıp eklenmesi şeklinde yöntemler uygulanmıştır.

Çizelge 2.1 : Önceki çalışmada kullanılan saldırılar. Kalın yazılan kelimeler eklenen kelimeleri temsil etmektedir.

	Yöntem	Örnek	İlgili Çalışmalar
Karakter Seviyesinde	Ekleme	apple → applee	[33]
	Silme	school → schol	[26, 33]
	Karakter Yer Değiştirme	hello → hlelo	[26, 33]
	Aynı/Benzer Telaffuz Edilen Farklı Kelimeler Kullanma	egg → agg	[33]
	Karakteri Klavyede Ona En Yakın Olan Karakterle Değiştirme	shy → why	[23, 26, 33, 35]
	Görsel Olarak Benzer Karakterlerle Harfi Değiştirme	foolish → fo0lish	[25, 26, 28, 34]
	Boşluk Karakteri Ekleme	school → sc hool	[26, 33]
	Bir Karakteri Yanlış Yazmak	talk → taln	[33, 36]
	Yaygın Yazım Hatası Yapmak	film → flim	[28]
	Kelime Ortasındaki Karakterleri Yer Değiştirme	noise→ nisoe	[23]
Kelime Seviyesinde	Yerine Anlamca Benzer Kelime Kullanma	awful → terribly	[22, 24, 26, 36]
	Peş Peşe Kelimeleri Yer Değiştirme	"I don't want you to go" → "I don't want to you go"	[22]
	Sık Kullanılan Kelimeleri (Stopwords) Kaldırma	Ben ate ~~the~~ carrot	[22]
	Yeni Kelime Ekleme	The Uganda Securities Exchange (USE) is the **historic** principal stock exchange of Uganda.	[28]
	Kelime Silme	The Old Harbor Reservation Parkways are three ~~historic~~ roads in the Old Harbor area of Boston.	[28]
Cümle Seviyesinde	Cümle Ekleme	The Old Harbor Reservation Parkways are three historic roads in the **Some exhibitions of Navy aircrafts were held here.**	[28]
	Cümleyi Aynı Anlamda Yeniden Yazma	"How old are you" → "What's your age"	[22]
	İfadeyi Aynı Anlamda Yeniden Yazma	the actual composer is ~~different from~~ **not** the artist	[28]
	İfadeyi Silme	~~promotion of world security~~, improvement of economic conditions	[28]
	Dilbilgisi Hataları	"He ~~doesn't~~ **don't** like cakes"	[22]

Jin vd. [24], Li vd. [26] ve Ebrahimi vd. [36], kelimelerin anlamca birbirine benzer olanlarla değiştirilmesine yönelik çalışmış olup anlamca benzerlikleri de kelime vektörlerini kullanarak tespit etmişlerdir. Kelime değişimleriyle ilgili bu çalışmada

uygulanan yöntemlerden bazıları; kelimelerin eşanlamlılarıyla değiştirilmesi ve ünlü kişilerin yaygın olmayan isimlerinin kullanılmasıdır.

Niu vd. [22], metinleri okuyup anlamak, özetlemek ve tercüme etmek gibi görevler için kullanılan dil modelinden olan "İşaretçi-Üretici Ağları" (Pointer Generator Networks) kullanarak cümleleri yeniden ifade etmiştir. Liang vd. [28], Barzilay vd. [37]'nin önerdiği yöntemi kullanarak ifadeleri yeniden yazmışlardır. Benzer bir şekilde cümlelerin yeniden yazılmasına yönelik olarak bu çalışmada da deyimlerin kullanılmasıyla birlikte elbette ki bu cümlenin tamamen yeniden yazılmasıyla birebir aynı değildir. Jia vd. [21], okuduğunu anlama sistemleri için düşmanca örnekler oluşturmak amacıyla manuel olarak seçilmiş cümleler eklemişlerdir. Bu çalışmada kullanılan tweet metinlerinin anlamlarının minimum seviyede değişmesine odaklanılması sebebiyle, cümle eklemeyle ilgili yöntemlere yer verilmemiştir.

Schiller vd. [35], taraf tespit modellerinin sağlamlığını araştırmak amacıyla üç farklı rakip saldırı kullanmıştır. Bu saldırılardan ilki her cümlenin başına "and false is not true" (ve yanlış doğru değildir) totolojisini eklemektir. Diğer bir yöntem ise karakter değiş tokuşları ve yer değiştirmelerle yazım hatalarını ortaya koymaktır. Son olarak ise geri çeviri ile yeniden ifade etmek yöntemlerinin kullanılmasıdır. Bu yöntemler sonucunda dönüştürücü modellerin eğitim verilerinin yanlılığının aşırı öğrenme sebebinden kaynaklı olarak sağlamlık konusunda ciddi problemleri olduğunu rapor etmişlerdir. Çalışma özellikle taraf tespiti görevine odaklanmış olmaları ve kullandıkları yöntemler sebebiyle bu çalışmayla benzerlik göstermektedir. Modelleri kandırmak için kullandıkları karakter değiş tokuşları ve yer değiştirmelerle yazım hataları gibi yöntemlerle çalışmamızdaki bazı yöntemlere benzerlik göstermesine rağmen boşlukları silme ve deyimlerin kullanımı gibi modelleri yanıltmak için bu tez kapsamında daha farklı yöntemler de kullanılmıştır.

Bildiğimiz kadarıyla, boşlukları silme ve deyimleri kullanma dâhil olmak üzere bazı yöntemlerimiz önceki çalışmalarda kullanılmamış yöntemlerdir.

Çalışmamız aynı zamanda, kullanıcıların verilerinin izinleri alınmaksızın kullanılması ve haklarında çeşitli etiketlemelerde bulunabilinmesi sebebiyle gizlilik endişesi taşımaları vb. gibi konular bakımından doğal dil işleme alanında etik üzerine yapılan çalışmalarla da oldukça ilişkilidir. Mieskes vd. [38], doğal dil işlemede veri toplama ve paylaşımında etik sorunları araştırarak, hassas verilerin doğrudan kullanmaları

yerine kullanıcıları anonimleştirmesi önermiştir. Ancak, Feyisetan vd. [39], anonimleştirilmiş verilerin de aslında gizlilik sorununu çözmediğini belirtmektedir. Ayrıca, NLTK (Url-3), Stanford CoreNLP (Url-4) ve SpaCy (Url-5) gibi doğal dil işleme araçlarının anonimleştirilmiş verilerde bile kişisel bilgileri etiketleyebileceğini göstermiştir [40]. Bu tez çalışmasında ise, sosyal medya platformlarını kullanırken, kullanıcıların gizliliklerini nasıl koruyabileceği konusunda odaklanılmıştır.

3. TEMEL BİLGİLER

3.1 Dönüştürücü Modeller & BERT Model

Dönüştürücü modeller, doğal dil işleme alanındaki önemli bir gelişme olup doğal dil işleme problemlerini çözmek için kullanılan derin öğrenme temelli yapılardır. Bu modeller, dil verilerini anlamak, yorumlamak ve çeşitli dil işleme görevlerini gerçekleştirmek için kullanılırlar. Bu kapsamda eğitilmiş dönüştürücü modellerden biri olan BERT, 2018 yılında Google tarafından tanıtılmıştır [41]. BERT, çift yönlü dönüştürücü yapısı sayesinde bir cümleyi hem soldan sağa hem de sağdan sola doğru işleyerek, dil verilerini daha iyi anlamak ve doğal dil işleme görevlerinde yüksek başarı elde etmek için tasarlanmıştır. Model, büyük ölçekli dil veri kümeleriyle (örneğin, BookCorpus ve Wikipedia) önceden eğitilir. BookCorpus veri kümesiyle 800 milyon kelime ve Wikipedia veri kümesiyle 2.5 milyar kelime büyüklüğünde iki temel model olan “bert_large” ve “bert_base” geliştirilmiştir. Eğitildikten sonra ise taraf tespiti, metin sınıflandırma, duygu analizi, doğal dil anlama, metin özetleme, makine tercümesi vb. gibi çeşitli doğal dil işleme görevleri için hassas ayar yapılarak kullanılabilmektedir. Twitter-RoBERTa, 58 milyon tweet ile önceden eğitilmiş olan, yazım hatalarına karşı potansiyel olarak daha sağlam bir modeldir. Bu tez çalışması kapsamında ön eğitimli BERT, Twitter-RoBERTa ve Türkçe dili için özel eğitilmiş olan BERT model olan BERTurk modelleri kullanılmıştır.

3.2 Tweet

Tweet, Twitter platformunda paylaşılan genel olarak metin tabanlı gönderilere verilen genel isimdir. Tweetler genellikle kişisel düşünceleri, haberleri, etkinlikleri duyurmayı, fikir alışverişi yapmayı veya diğer kullanıcılarla etkileşimde bulunmayı amaçlayan, kısa ve öz iletiler olarak tanımlanmaktadır. Twitter kullanıcıları, 280 karakterlik sınırlamaya sahip olan tweetlerde metin, linkler, görseller ve hashtag'ler gibi içeriği paylaşabilirler. Hashtag'ler, sosyal medya içeriğini etiketlemek ve gruplamak için kullanılır. Hashtag'ler, kullanıcıların belirli bir konu, olay, etkinlik veya ilgi alanı hakkında içerikleri kolayca bulmalarına yardımcı olur. Hashtag'lerin kullanımı, sosyal medya kullanıcıları arasında etkileşimi ve iletişimi artırır.

Twitter gibi sosyal medya platformlarında milyonlarca kullanıcılar tarafından günlük olarak oluşturulan tweetler, büyük ve çeşitlilik içeren bir veri kümesi sunmaktadır. Bu veriler, doğal işleme modellerinin eğitilmesi ve geliştirilmesi için son derece değerlidir.

Bu tez kapsamında tweet verileri üzerinde çalışılmış olup bu kavram sıkça kullanılmaktadır. Çizelge 3.1'de tweet örnekleri ve bu tweet'lerin desteklediklerini açıkça yansıtmış oldukları siyasi parti taraf etiket değerleri verilmiştir. Örneklerde görülmekte olduğu gibi tweet verileri kitaplarda veya resmi yazışmalardaki kullanılan dilden farklı olarak gündelik dille yazılmakla birlikte çok sayıda yazım ve imla hatasını içermektedir. Bunun yanı sıra genellikle politik ve taraflı veriler olmalarıyla birlikte dilbilgisi kurallarına uygun olmayan kısaltmalar, argo söylemler, emojiler ve trendleri belirlemek için önemli bir araç olan hashtag'ler kullanılmaktadır.

Çizelge 3.1: Taraf etiketleriyle verilmiş örnek tweetler

Tweet	**Taraf Etiketi**
4 Haziran seçimi kaçın toma geliyor diyenlerle, koşun tank kaçıyor diyenler arasında olacak. Biz daha son sözümüzü söylemedik DEVAM diyoruz ve bıkmadan söylüyoruz ki durmak yok yola #ERDOĞANileDEVAM	AK Parti
Chp nin maltepe'de ki şu kalabalığını gören Ak parti korkmasında ne yapsın 😂😂😂#AKPyusufyusuf	CHP
Şu an 5 oy sana başkan. 5 oy da HDP ye verdik hayırlısı neyse o olsun	HDP
İZMİRDE SİMİTE GEVREK, CUMHURBAŞKANINA MERAL AKŞENER DENİR. 😎 #EliniUzat #ErdoğanKimeCevapVeremiyor #YüzünüGüneşeDönTürkiye #İYİlerkazanacak #TürkiyeİYİOlacak	İYİ Parti
#ülkemdeki problemleri ancak ve ancak milli görüşün tek temsilcisi saadet partisi çözer NOKTA	SAADET

4. TARAF TESPİT MODELLERİ

Çalışma kapsamında, İngilizce için BERT, Twitter-RoBERTa tabanlı ve Türkçe için BERTurk tabanlı olmak üzere üç farklı taraf tespit modeli kullanılmıştır.

4.1 BERT Tabanlı İngilizce Taraf Tespit Modelleri

Veri kümesi oluşturma adımları Bölüm 7.1.1 anlatılmış olan İngilizce veri kümesiyle BERT ve Twitter-RoBERTa modelleri hassas ayar yapılmıştır. Bu modeller İngilizce için ateizm, iklim değişikliği, feminizm, Hillary Clinton ve kürtajın yasallaştırılması konuları üzerine metinlerin bu konulara karşı taraflarının tespiti edilmesi görevlerini gerçekleştirmektedir. Her konu için ayrı bir BERT modeline hassas ayar yapılmıştır. Modele girdi olarak tarafı tespit edilmek istenen bir metin verilmektedir. Model çıktı olarak, verilen metin için "DESTEKLEYEN", "KARŞIT" veya "NÖTR" etiketlerinden birini vermektedir.

4.2 BERTurk Tabanlı Türkçe Taraf Tespit Modeli

Veri kümesi oluşturma adımları Bölüm 7.1.2 'de anlatılmış olan Türkçe veri kümesiyle hassas ayar yapılmış bir BERTurk modeli eğitilmiştir. Bu model Türkçe metinlerin destekledikleri siyasi partilere göre taraf tespiti görevini gerçekleştirmektedir. Her etiket için her partinin resmi kısaltma ismi kullanılmıştır. Modele girdi olarak tarafı tespit edilmek istenen bir metin verilmektedir. Model çıktı olarak, verilen metin için siyasi parti etiketlerinden birini vermektedir.

5. ÖNEMLİ/ÖNEMSİZ KELİMELERİ BULMA YÖNTEMLERİ

İnsanlar okudukları bir metnin tarafını tespit ederken genellikle kullanılan belirli ifadelerden bunu anlayabilmektedirler. Bu kelimeleri önemli kelimeler olarak adlandırıyoruz. Önemli kelimeler, bulundukları cümle içerisinde genel olarak taraf tespiti yapabilmeyi kolaylaştıran, kişinin tarafıyla ilgili önemli ipucu içeren kelimeler olup, bazen konuyla ilgili popüler söylemler ve sloganlardan oluşabilirler. Metin üzerinde rasgele kelimeler yerine, özenle seçilmiş bu kelimeler üzerinde yapılan değişlikler modelleri aldatmakta daha faydalı olacaktır. Bu bölümde, otomatik bir şekilde metinlerdeki değişiklik yapılacak kelimeleri seçmek için önerdiğimiz denetimsiz ve denetimli iki farklı yöntemi anlatıyoruz.

5.1 FastText Tabanlı Önemli/Önemsiz Kelime Tespit Modeli

Belirli bir konu bulunduğu durumlarda, etiketli veri kullanmadan, önemli kelimeleri tespit etmek için FastText kelime vektörlerini kullandık [42]. Bu yöntemi, bir tweet'teki tüm kelimelerin, ilgili tweet'in ait olduğu veya değerlendirildiği konuyu temsil eden kelime veya kelime grubuyla (Örn. iklim değişikliği konusu) kosinüs benzerliklerine göre sıralanması şeklinde uyguladık. Yaptığımız sıralamada ilgili konu vektörüne en yakın olan kelimeyi en önemli olarak değerlendirirken, uzaklaştıkça kelimeler daha önemsiz olarak değerlendirdik. Buna göre kelimeler, en yakından en uzağa olacak şekilde sıralanmıştır. Bir cümleden, N tane kelime değiştirilecekse bu sıralamadaki ilk N kelime seçilerek belirlenmektedir.

Yöntem uygulanırken gerekli olan kelime vektörleri, kelimeleri temsil etmek için kullanılan matematiksel modellerdir. Bu vektörler, doğal dil işleme ve makine öğrenimi modellerinde yaygın olarak kullanılmakta ve metin tabanlı verilerin sayısal formata dönüştürülmesini sağlamaktadır. Kullandığımız "wiki-news-300d-1M.vec", büyük bir metin veri kümesi olan Wikipedia'nın ve haber makalelerinin kullanıldığı önceden eğitilmiş bir kelime vektörü modeli olup; bu model, 300 tane boyuta sahip vektörlerle her kelimeyi temsil etmektedir ve toplamda 1 milyondan fazla kelime içermektedir [43].

5.2 BERTurk Tabanlı Önemli/Önemsiz Kelime Tespit Modeli

Önemli/önemsiz kelimelerin tespiti için diğer bir yöntem olarak denetimli öğrenme temelli bir yöntem öneriyoruz. Denetimli öğrenme yapmak için öncelikle bir veri kümesi oluşturulması gerekmektedir. Bu veri kümesi, elimizdeki siyasi parti tweetleri kullanılarak elde edilmiştir. Bu tweetler kullanılarak öncelikle seçilen tweet'in tarafı konusunda Bölüm 4.2'de anlatılan BERTurk Tabanlı Türkçe Taraf Tespit modelinin tahmin etmesi sağlanarak tahmin değerini kaydettik. Ardından tweet'deki her bir kelimeyi teker teker çıkardık ve modelin orijinal tweet'e verdiği tahmini değiştiren kelimeleri "ÖNEMLİ" olarak etiketledik. Tahmini değiştirmeyenleri ise "ÖNEMSİZ" olarak etiketledik. Örneğin, "Vakit İNCE vaktidir." cümlesi için örnek kullanım Çizelge 5.1'de gösterilmiştir. Bu tweet üzerinden herhangi değişiklik yapılmadan model tarafından tahmin edildiğinde CHP'yi destekler şekilde tahmin edilmiştir. Bu cümlede önce "Vakit" kelimesini çıkardığımızda "İNCE vaktidir." cümlesinin hâlâ tahmininde bir değişiklik olmamıştır. Bu sebeple "Vakit" kelimesi "ÖNEMSİZ" bir kelime olarak etiketlenmiştir. "İNCE" kelimesi çıkarılıp "Vakit vaktidir." cümlesi tahmin edildiğinde ise İYİ Parti'yi destekler şekilde tahmin edilmesiyle etiket tahmini değişmiştir. Bu sebeple "İNCE" kelimesi "ÖNEMLİ" bir kelime olarak etiketlenmiştir. "vaktidir" kelimesi çıkarıldığında ise "Vakit İNCE" cümlesinin de CHP'yi destekler olarak tahminlenmesi sebebiyle "vaktidir" kelimesi de "ÖNEMSİZ" olarak etiketlenmiştir.

Çizelge 5.1: Siyasi parti veri kümesinden önemli/önemsiz veri kümesi üretme örneği

"Vakit İNCE Vaktidir." metni için üretilen durumlar	**Orijinal Tahmin Etiketi**	**Yeni Taraf Tahmin Etiketi**	**Kelimenin Önem Etiketi**
İNCE Vaktidir.	CHP	CHP	ÖNEMSİZ
Vakit Vaktidir.	CHP	İYİ Parti	ÖNEMLİ
Vakit İNCE	CHP	CHP	ÖNEMSİZ

Nihai olarak bu yöntem ile "ÖNEMLİ" ve "ÖNEMSİZ" etiketli kelime listeleri elde edilmiştir. Elde edilen bu listelerdeki dağılımı incelediğimizde, "ÖNEMLİ" etiketli verinin eğitim kümesinin sadece 4,502 adet olup %5.2'lik bir dilimine sahip olduğu geriye kalan kelimelerin "ÖNEMSİZ" olarak etiketlenmiş kelimelerden oluştuğunu gördük. Modelin, "ÖNEMLİ" kelimeleri daha iyi tespit etmesi için daha fazla

“ÖNEMLİ” etiketli veriye ihtiyaç duyulması sebebiyle tweet’ler üzerinde adlandırılmış varlık tanımlama yapılarak yer, kişi ve kuruluş olarak etiketlenen kelimelerin de “ÖNEMLİ” olabileceğini kabul ettik ve bu kelimeleri de “ÖNEMLİ” olarak etiketleyerek eğitim veri kümesinde dâhil ettik. Bu sayede “ÖNEMLİ” olarak etiketlenmiş kelime oranı %5.2’den %17’ye yükseltilmiştir. Siyasi parti verilerinden ve adlandırılmış varlık tanımlama yoluyla elde ettiğimiz kelime listelerini kullanarak BERTurk modelini hassas ayardan geçirdikten sonra elde ettiğimiz modeli önemli/önemsiz kelime tespiti için kullandık.

Bu model, girdi olarak aldığı kelimeyi, “ÖNEMLİ” ya da “ÖNEMSİZ” olacak şekilde tahmin etme şekilde çalışmaktadır. Taraf tespit modellerini yanıltmak için, bir tweet’deki her kelimeyi ayrı ayrı olarak bu modele girdi olarak veriyoruz. Modelin “ÖNEMLİ” dediği kelimeler üzerinde, Bölüm 6. ’da anlatılan yöntemleri uyguladık. Burada belirtmek isteriz ki, önerdiğimiz bu model, klasik sözlük temelli bir model değildir. Bu sebeple, eğitim kümesinde görmediği kelimeleri anlam olarak bakarak benzer kelimelere bakarak tahmin yapabilme yeteneğine sahiptir. Örneğin, “siyaset” kelimesi eğitim veri kümesinde “ÖNEMLİ” olarak etiketlenmiş olarak yer aldığı bir durumda “politika” kelimesi eğitim veri kümesinde yoksa bu kelimeyi de anlam benzerliğinden dolayı modelimiz aynı etiketi verebilecektir. Ancak, modelimiz tweet içeriğindeki bağlam bilgisini düşünmemektedir. Modelimize dair eğitim ve test veri kümesi bilgiler Bölüm 7.1.3 ‘te anlatılmıştır.

6. MODELLERİ ALDATMA YÖNTEMLERİ

Bu bölümde, taraf tespit sistemlerini aldatmak için çalışma kapsamında kullanılmış olan yöntemler ele alınmıştır. Tweetler üzerindeki taraf tespiti görevi yapan makine öğrenme yöntemlerini aldatma konusu üzerine odaklandığımız için yazım hataları eklemeyi ve tweetleri yeniden ifade etmeye yönelik yöntemleri uyguluyoruz.

Kullanıcıların kişisel bilgileri gizlenirken aynı zamanda sosyal medya üzerinden tweet ve gönderilerini nasıl yayınlayabileceklerini keşfedebilmek amacıyla, sosyal medya üzerinden yayınlamış oldukları metinlerin manuel ve otomatik olarak değiştirmesi yönünde bir çalışma gerçekleştirilmiştir. Çalışmada tweet içeriğini değiştirmekte etkili yöntemleri belirlemek amacıyla, İngilizce için SemEval Görev-6 veri kümesi [44], Türkçe için çalışma kapsamında oluşturulan ve ilerleyen bölümlerde detaylı olarak anlatılacak olan siyasi partilerle ilgili veri kümesi [45] kullanılarak BERT, Twitter-RoBERTa ve BERTurk modellerine taraf tespiti görevi için hassas ayarlama yapılmıştır. Daha sonra, modeli yanıltmak için tweet içeriklerini manuel veya otomatik olarak değiştirilerek, kullanılan yöntemlerin etkinlikleri belirlenmiştir.

Çalışmada kullanılan yöntemler kasıtlı olarak yapılan yazım hataları ve yeniden ifade etme olmak üzere temel iki başlık altında incelenmiştir. Bahsedilecek yöntemlerin uygulanma biçimleri deney düzeneklerine göre manuel veya otomatik olmak üzere iki farklı şekilde uygulanmıştır. Yöntemlerimizi daha iyi anlayabilmek için her bir yöntem için bir örnek tweetini Çizelge 6.1'de sunulmuştur.

6.1 Kasıtlı Olarak Yapılan Yazım Hataları

BERT modelleri, kendi sözcük dağarcığına dayalı olarak alt kelimeler için kelime vektörü oluşturmaktadır. Eğer karşılaşılan bir kelime, sözcük dağarcığında bulunmayan bir kelime ise, kelimeyi alt kelimelere böler ve her biri için bir kelime vektörü oluşturulmaktadır. Örneğin, kelime İngilizce "against" kelimesini "aganist" şeklinde yazmak, BERT'in "against" kelimesi yerine "ag", "-ani" ve "-st" alt kelimeleri için kelime vektörleri oluşturmasına yol açmaktadır.

Çizelge 6.1 : Modelleri aldatmak için değiştirilmiş halleriyle birlikte örnek tweetler. Değiştirilen kelimeler kalın harflerle yazılmıştır.

	Yöntemler	Orijinal Tweet	Değiştirilmiş Tweet
Yazım Hataları	Boşluk silme	also what's up with this ridiculous weather ? ? it was raining this morning and now it's like super hot ! #weather problems #lame	also what's up with this **ridiculousweather** ? ? İt **wasraining** this morning and now it's **likesuper** hot ! #weather problems #lame
	Boşluk Ekleme	breaking 911 probably she made a promise to support gun rights to one citizen , while promising to ban guns to the other	**b reaking** 911 pro bably she made a promise to **su pport g un** rights to one **cit izen** , while **pro mising** to **b an** guns to the other
	Harf Sıralarını Karıştırma	adam smith usa because clearly hillary clinton is a champion for us all	adam smtih usa because **clarely hlliary clitonn** is a **champoin** for us all
	Karakter Değiştirme	men and women should have equal rights, we are all human	men **änd w0men** should have **equä| r!ghts** , we are all **humän**
	Hashtag İşareti Ekleme	hillary clinton hillary for nh hope to see her in not cool soon	hillary clinton hillary **#for** nh **#hope** to **#see** her **#in** not **#cool** soon
Yeniden İfade Etme	Bilinenin Dışındaki İsimleri Kullanma	hillary clinton hillary for nh hope to see her in not cool soon	hillary **diane** clinton for us hope to see her in not cool soon
	Zıt Anlamlı Kelimeleri Bir Arada Kullanma	there's no more normal rains anymore always storms, heavy and flooding	contrary to **normal** there is more **abnormal** rain now always storms, heavy and floods
	Yeni Hashtag Ekleme	it's time that we move from good words to good works, from sound bites to sound solutions hillary clinton #ready for hillary	it's time that we move from good words to good works, from sound bites to sound solutions hillary clinton #ready for hillary **#usa #decision #time**
	Hashtag Silme	#fiona bruce wants a government that forces women to have children, and then refuses to financially help them #body autonomy	bruce wants a government that forces women to have children , and then refuses to financially help them #body autonomy
	Eş Anlamlısıyla Değiştirme	generate belief in quality existence for everyone especially children in that community kitti ngt on 2016	generate belief in quality existence for everyone especially **kids** in that community kitti ngt on 2016
	Deyim Kullanma	also what's up with this ridiculous weather ? ? it was raining this morning and now it 's like super hot ! #weatherproblems #lame	also what's up with this ridiculous weather ? ? it was raining this morning and now **it 's dog days** ! #weatherproblems #lame
	Kelime Silme	success hillary clinton said she 's receiving a constant barrage of attacks from the right great job , guys keep it up !	success hillary clinton said she 's receiving a constant barrage of attacks from the right great job , guys keep it up !
	Olumsuz İfadeleri Birlikte Kullanma	the irish national school system is secular under law we can reaffirm secularism by going through the courts ! humanism ireland	the irish national school system is **not non secular** under law we can reaffirm secularism by going through the courts ! humanism ireland

Kasıtlı olarak yazım hataları eklenmesindeki amacımız, BERT, Twitter-RoBERTa ve BERTurk modellerinin kelime ya da konuyla ilgisiz alt kelimeler için kelime vektörleri oluşturmasına neden olmak ve sözcük dağarcığı dışında kalan kelimelerin sayısını arttırmaktır. Kasıtlı olarak yapılan hatalar, boşluk silme, boşluk ekleme, harf sıralarını karıştırma, karakterleri değiştirime, hashtag işareti ekleme olmak üzere temel 5 yöntemden oluşmaktadır.

6.1.1 Boşluk silme

Doğal dil işleme modellerinin metni anlaması için kelimeler arasındaki uygun boşluklar önemlidir. Bu yöntemde, tweet içeriğini değiştirmeyi, belirli boşluk karakterlerini kaldırarak bitişik kelimeleri birleştirmek hedeflenmektedir. Ancak tüm boşlukları kaldırmak metni anlaşılmaz hale getirmektedir. Bu nedenle, doğru tahminler için etkili olacağı düşünülen önemli kelimeleri seçip bağlamına bağlı olarak önceki veya sonraki kelimeyle birleştirilmesi yönünde bir yol izlenmiştir. Bu işlem manuel olarak uygulandığında, tweet'in okunabilir kalmasına sağlayacak noktaya kadar devam ettirilebilmektedir.

6.1.2 Boşluk eklemek

Bu yöntemle, önemli kelimelerin harfleri arasına bir boşluk karakteri ekleyerek sözcük dağarcığı dışında kalan kelimelerin sayısını artırmayı amaçlanmaktadır.

6.1.3 Harf sıralarını karıştırma

Önceki çalışmalarda kullanılan karakter değiş-tokuşu yöntemine [26, 33] benzer şekilde, seçilen kelimelerin ilk ve son harfleri sabit bırakılması koşuluyla harf sırasını değiştirilmesi şeklinde gerçekleştirilmiştir. Bu yöntem, “Typoglycemia” adıyla bilinen bir şehir efsanesinden esinlenilerek uygulanması yönünde karar verilmiştir. Bu yöntem uygulanırken, bazı durumlarda metni anlaşılmaz hale getirilmiş olabilmektedir. Bu sebeple kelimelerin hâlâ okunabilir olmasına dikkat edilmiştir.

6.1.4 Karakterleri değiştirme

Sosyal medya platformlarında sıkça rastlanan popüler yazım stilleri yaygın olarak kullanılmaktadır. Bu stillerin birçoğu bazı harflerin benzer görünüme sahip olması veya benzer telaffuza sahip olmalarına dayanmaktadır. Bu yöntemdeki değiştirme prosedürü şu şekildedir: a→¨a, i→!, l→|, o→0, ae→æ, to→2, for→4 ve great→gr8. Değiştirilmiş tweetler hâlâ anlaşılabilir olsa da, orijinal tweetlerle karşılaştırıldığında

profesyonel görünmeyebilmektedir, bu da yöntemin gerçek hayatta kullanımını sınırlı olmasına sebep olabilir.

6.1.5 Hashtag işareti eklemek

Hashtag'ler, belirli konuların önemini göstermek için sosyal medyada sıkça kullanılmaktadır. Bu yöntemde, önemsiz kabul edilen kelimelerin önüne "#" işareti eklenmektedir. Bunun sonucunda, aldatılması amaçlanmakta olan modelin önemsiz kelimelere daha fazla dikkat ederek yanıltılması ve yanlış tahminlerde bulunması amaçlanmıştır.

6.2 Yeniden İfade Etme

Bu yöntem grubunda amaç, tweet içeriğinde önemli değişiklikler yaparken aynı zamanda anlamın korunmasıdır. Bu yöntemlerdeki temel düşünce, modellerin eğitildikleri veri kümelerinden kazandıkları içsel önyargıları kullanmaktır. Örneğin, BERT modelleri, bağlamsal kelime vektörleri üretse de Niven vd. [46] BERT'in tahminlerinin bazı kelimelerin (Örn. İngilizce'deki "not" kelimesinin) varlığından etkilendiğini bildirmiştir. Bu nedenle, eğitim veri kümesinde sıkça görünen ve belirli bir etiketle ilişkilendirilen bir kelimenin varlığı, kelime etikete doğrudan bağlı olmasa bile sonuçları etkileyebileceği yönündedir. Metinler üzerinde değişiklik yapılma çalışması sırasında karşılaşılan başlıca zorluk, başkaları tarafından yazılan metinlerin anlamlarını değiştirmeden üzerlerinde değişiklik yapılmaya çalışılmasıdır. Ayrıca, çalışmanın temelini oluşturan sosyal medya gönderilerinde, lehçeleri, şiveleri, eksik cümleleri ve dilin yanlış kullanımlarını içerebilen bir dil kullanılmaktadır. Bu durumda metinler, çok sayıda ve çeşitte dilbilgisi hatası içerebilmektedir. Bu nedenle, sonuç metnin anlamlı ve tutarlı olmasını sağlamakta zorluklar yaşanmıştır. Bazı durumlarda, dikkatli çalışılmasına rağmen, yapılan değişiklikler alışılmadık dil kullanımına yol açmış olabilmektedir. Ancak, temel amaç belirli ifadelerin sonuca olan etkilerini araştırmaktır. Tweet'leri yeniden ifade etme yöntemleri, bilinenin dışındaki isimleri kullanma, zıt anlamlı kelimeleri bir arada kullanma, yeni hashtag ekleme, hashtag silme, eş anlamlısıyla değiştirme, deyim kullanma, kelime silme, olumsuz ifadeleri birlikte kullanma üzere temel 8 yöntemden oluşmaktadır.

6.2.1 Bilinenin dışındaki isimleri kullanma

Metinlerde geçen kişilerin yaygın olarak bilinen isimlerini kullanmak yerine kişilerin bilinmeyen isimlerinin veya isimlerinin kısaltmalarının kullanıldığı yöntemdir. Örneğin, “Hillary Clinton” için “Hillary Diane Clinton” veya “HC” kullanılması.

6.2.2 Zıt anlamlı kelimeleri bir arada kullanma

Bir kelimenin zıt anlamlısını kullanmak cümledeki anlamı tersine çevirmektedir. İki zıt anlamlı kelimenin cümlede bir arada kullanılması ile modellerin cümledeki anlamı anlamasını zorlaştırmak prensibine dayanmaktadır. Örneğin, “normal” ve “abnormal” kelimelerinin bir arada aynı cümlede kullanılması. Yöntem uygulanırken cümlenin orijinal anlamının bozulmamasına da dikkat edilmektedir.

6.2.3 Yeni hashtag ekleme

Hashtagler, verilen bir tweet'in durumunu tahmin etmede faydalı olabilmektedir. Bu nedenle bu yöntem, modelleri aldatabilmek için tweet'in durumuyla “tarafsız” ilişkili olan hashtaglerin eklenmesi yöntemidir. Örneğin İngilizce veri kümesi için “#Monday” ve “#future”; Türkçe veri kümesi için “#gündem”, “#haberler” gibi nötr anlam ifade eden hashtag’leri tweet sonlarına eklenmesi şeklinde uygulanmaktadır.

6.2.4 Hashtag silme

Bu yöntem ise cümle anlamı bozmayacak şekilde, genellikle cümle sonunda bulunan hashtag’lerin kaldırılması işlemidir. Fakat aksi olan durumlarında uygulanması da mümkündür.

6.2.5 Eş anlamlısıyla değiştirme

Cümlelerdeki kelimelerin mümkün olduğunda kelimelerin eşanlamlılarıyla değiştirilmesi yöntemidir. Örneğin, İngilizce veri kümesi için “children” → “kids” şeklinde uygulanmaktadır.

6.2.6 Deyim kullanma

Deyimlerin anlamları, gerek dile dair söz sanatları içerebilmesi bakımından gerekse kültürlere ait öğeler içerebilmesi bakımından dil modelleri için bir zorluk oluşturabilmektedir. Bu durumdan faydalanarak, bu yöntemde anlamca uygunluk oluşması durumlarında deyimler kullanması olarak ifade edilmektedir. Örneğin, İngilizce’de kullanılan “brass monkey” deyiminin “çok soğuk hava” yerine

kullanılmasıdır. Yine benzer şekilde İngilizce'de kullanılan "raining cats and dogs" deyiminin "yoğun yağış" yerine koyulması şeklinde uygulanmaktadır.

6.2.7 Kelime silme

Bu yöntem, anlamı belirgin bir şekilde etkilemeyecek kelimeleri tweet cümleleri içinden kaldırılması prensibine dayanmaktadır.

6.2.8 Olumsuz ifadeleri birlikte kullanma

Olumsuz ifadeler, dil modelleri için zorluk oluşturabilir konulardan olup bunun sebebi kelimenin anlamını bir kelime ile tam tersine çevirebilir olmasıdır. Örnek olarak İngilizce'deki "not" ve "without" kelimeleri verilebilir. Bu nedenle, bu yöntemde pozitif ifadeleri olumsuzluk kelimeleri ve orijinal ifadenin zıddı ile değiştirerek, iki negatif ifade ile pozitif mana oluşturmak prensibi temel alınmıştır. Örneğin İngilizce veri kümesi üzerinde "is religious" ifadesi "is not nonreligious" şeklinde değiştirilebilir.

7. DENEYLER

7.1 Veri Kümeleri

İngilizce ve Türkçe olmak üzere iki farklı deney düzeneği için iki farklı veri kümesi kullanılmıştır.

7.1.1 İngilizce veri kümesi

Çalışma kapsamında İngilizce dili taraf tespiti görevi için SemEval 2016 Görev-6 veri kümesi kullanılmıştır [44]. Bu veri kümesi, beş konudan oluşmaktadır: Ateizm, İklim Değişikliği, Feminizm, Hillary Clinton ve Kürtajın Yasallaştırılması. Her tweet, "DESTEKLEYEN", "KARŞIT" ve "NÖTR" etiketlerinden biriyle etiketlenmiştir. Eğitim ve test verilerinin etiket dağılımı Çizelge 7.1'de gösterilmiştir.

Çizelge 7.1 : İngilizce taraf tespit veri kümesinin etiket dağılımı

	Eğitim			**Test**		
Konu/Etiket	DESTEKLEYEN	KARŞIT	NÖTR	DESTEKLEYEN	KARŞIT	NÖTR
Ateizm	92	304	117	32	160	28
İklim Değişikliği	212	15	168	123	11	35
Feminizm	210	328	126	58	183	44
Hillary Clinton	112	361	166	45	172	78
Kürtajın Yasallaştırılması	105	334	164	46	189	45
Toplam	**731**	**1,342**	**741**	**304**	**715**	**230**

7.1.2 Türkçe veri kümesi

Çalışma kapsamında Türkçe dili taraf tespiti görevi için [45]'deki 2018 Türkiye Cumhuriyeti Cumhurbaşkanlığı seçim sonuçlarını tahminlemek amaçlı kullanmış oldukları veri kümesinden 2,475 hesaba ait tweet metin içerikleri kullanılmıştır. Hesapların etiketlenme işlemi hesapların profil açıklamaları ve fotoğrafları dikkate alınarak bir değerlendirici tarafından manuel olarak yapılmıştır. Çalışma kapsamında metinler, adayların isimleriyle etiketlemek yerine bu adayları öneren partilerin etiketleriyle etiketlenmiştir. Kullanılan aday ve parti dönüşüm tablosu Çizelge 7.2'de

ifade edilmiştir. Bu tez çalışması kapsamında parti etiketleri kullanılması tercih edilmiştir.

Çizelge 7.2 : Adayların parti etiketleriyle eşleşmeleri

Aday	Parti
Recep Tayyip Erdoğan	AK Parti
Muharrem İnce	CHP
Selahattin Demirtaş	HDP
Meral Akşener	İYİ Parti
Temel Karamollaoğlu	SAADET

Çalışmada eğitim ve test verisi olarak kullanılmak üzere toplamda 2,475 adet kullanıcı bulunurken, 95,650 adet tweet kullanılmıştır. Öncelikle kullanıcıların destekledikleri partilere göre etiketlenmeleri gerçekleştirilmiştir. Ardından, kullanıcılara ait tüm tweetler, kullanıcının profilinin etiketlendiği siyasi parti etiketiyle etiketlenerek kullanılmıştır. Kullanıcıların eğitim ve test kümelerine göre dağılımları Çizelge 7.3; tweetlerin etiket, eğitim ve test kümelerinde göre dağılımları ise Çizelge 7.4'de ifade edilmiştir.

Çizelge 7.3: Kullanıcılar taraf etiketleri, eğitim ve test kümeleri için dağılımları

Kullanıcılar	Eğitim	Test
AK Parti	688	172
CHP	416	104
HDP	188	47
İYİ Parti	464	116
SAADET	224	56
Toplam	**1,980**	**495**

Çizelge 7.4: Tweetlerin taraf etiketleri, eğitim ve test kümeleri için dağılımları

Tweetler	Eğitim	Test
AK Parti	32,133	660
CHP	25,736	411
HDP	8,849	197
İYİ Parti	17,717	461
SAADET	9,262	224
Toplam	**93,697**	**1,953**

7.1.3 BERTurk tabanlı önemli/önemsiz kelime bulma veri kümeleri

Bölüm 5.2'de anlatılan BERTurk tabanlı önemli/önemsiz kelime tespit modelinin eğitim ve test veri kümesi, Çizelge 7.5'de ifade edilmiştir.

Çizelge 7.5 : BERTurk Tabanlı Önemli/Önemsiz Kelime Tespit Modeli için oluşturulan eğitim ve test kümeleri için dağılımları önem etiketleri dağılımı

Veri Kümesi Etiketler	**Eğitim**	**Test**
ÖNEMLİ	6,209	2,994
ÖNEMSİZ	30,801	12,795
Toplam	**37,010**	**15,789**

7.2 Deney Düzenekleri

Çalışmada temel olarak üç farklı deney düzeneği bulunmaktadır. Bunlar İngilizce veri, Türkçe veri, okunabilirlik ve anlam değişimleri üzerinedir. Bu bölümde deney düzenekleri ve deneylerden elde edilen sonuçların değerlendirileceği metriklerden bahsedilmiştir.

7.2.1 İngilizce veri deney düzeneği

İngilizce taraf tespiti deney düzeneği için Çizelge 7.1'de verilen veri kümesi üzerinde Bölüm 6. 'da anlatılan metin değiştirme yöntemleri manuel ve otomatik olarak uygulanmasından oluşmaktadır. Ardından taraf tahmini için kullanılan BERT ve Twitter-RoBERTa modelleri, 2e-5 öğrenme oranıyla, 16 grup boyutu (batch size) ve 11 çağ (epoch) değerleriyle hassas ayar yapılmıştır. Orijinal ve değiştirilmiş metinlerin tahmin değerleri kaydedilerek, elde edilen sonuçlar Bölüm 0 ve Bölüm 7.5'de yorumlanmıştır.

7.2.2 Türkçe veri deney düzeneği

Türkçe taraf tespiti düzeneği için Çizelge 7.4'de verilen tweet veri kümesi üzerinde Bölüm 6. 'da anlatılan metin değiştirme yöntemlerinden "Hashtag İşareti Ekleme", "Hashtag Ekleme", "Hashtag Silme", "Karakter Değiştirme", "Harf Sıralarını Karıştırma", "Boşluk Ekleme" ve "Boşluk Silme" yöntemleri otomatik olarak uygulanmasından oluşmaktadır. Ardından taraf tahmini için kullanılan BERTurk modeli, 2e-5 öğrenme oranıyla, 8 grup boyutu (batch size) ve 3 çağ (epoch) değerleriyle hassas ayar yapılmıştır. Orijinal ve değiştirilmiş metinlerin tahmin değerleri kaydedilerek, elde edilen sonuçlar Bölüm 7.5'de yorumlanmıştır.

7.2.3 Okunabilirlik ve anlam değişimi deney düzeneği

Metin değiştirme yöntemlerinin manuel yapılan deneylerde okunabilirliğin ve anlamın değişmemesinin sağlanması insanlar tarafından dikkat edilerek yöntemlerin uygulanması şeklinde yapılmaktadır. Fakat metin değiştirme yöntemlerinin otomatik olarak yapılması sonucunda okunabilirliğin ve anlam değişiminin kontrol edilmesi gerekmektedir. Bu kapsamda metinler üzerinde değişikliklerin yapıldığı veri kümesinden seçilen örnekler iki kişi tarafından okunurluk ve anlam değişimi bakımından değerlendirilmesi şeklinde gerçekleştirilmiştir.

7.2.4 Değerlendirme metrikleri

Çalışma kapsamında taraf tespiti modellerinin performanslarını ölçmek için doğruluk ve makro ortalama F1 skorları kullanılmaktadır. Veri kümelerine göre etiketlendirme değişmekte olup İngilizce taraf tespit sistemi için "DESTEKLEYEN", "KARŞIT", "NÖTR" etiket değerleri kullanılırken; Türkçe taraf tespit sistemi için parti isimlerinin resmi kısaltmaları olan "AK Parti", "CHP", "HDP", "İYİ Parti", "SAADET" etiketleri; BERTurk tabanlı önemli/önemsiz kelime tespit sisteminde ise "ÖNEMLİ", "ÖNEMSİZ" etiketleri kullanılmaktadır.

7.3 Önemli/Önemsiz Kelime Tespit Model Performansı

Türkçe veri kümeleri için kullanılmakta olan Bölüm 5.2'de anlatılan BERTurk modeline hassas ayar yapılan sistem BERTurk tabanlı önemli/önemsiz kelime tespit modeli olarak adlandırılmaktadır. Bu modelin performansına dair metrikler Çizelge 7.6' da verilmiştir. Modelin önemsiz olan kelimeleri bulmakta daha başarılı olmasına rağmen önemli kelimeleri bulmakta da başarılı olduğu görülmektedir. Bunun sebebi, modelin eğitim veri kümesindeki önemsiz etiketli verilerin yüksek yoğunlukta olmasıyla açıklanabilir.

Çizelge 7.6 : BERTurk tabanlı önemli/önemsiz kelime tespit model performansı

Etiket	F1 Skoru
ÖNEMLİ	0.617
ÖNEMSİZ	0.882

7.4 Manuel Değiştirilen Metinlerdeki Deney Sonuçları Performansı

Başlangıçta, önerilen tweet bozma yöntemlerini geliştirmek için BERT modeli tarafından doğru bir şekilde tahmin edilen tweetler kullanılmıştır. Fakat, önerilen yöntemlerin etkinliğini daha güvenilir bir şekilde değerlendirebilmek amacıyla, model tarafından yanlış sınıflandırılan tweetlerin de bulunduğu daha kapsayıcı bir örneklemeye ihtiyaç olması sebebiyle, rastgele seçilen tweetleri örnekleme yöntemiyle tercih edilmiştir. Bu seçilen tweetler üzerinde manuel değişiklikler yapılmıştır.

Manuel metin değişikliklerinin sonuçları, değişiklikleri yapan kişiye bağlı olabilmektedir. Sonuçlardaki bu önyargıyı azaltmak amacıyla, manuel değişiklikler 3 kişi tarafından yapılmıştır. Birinci kişi tarafından, Bölüm 6. 'da açıklanan yöntemleri kullanarak tweetleri manuel olarak değiştirme süreci başlatılmıştır. Her tweet için bu kişi, üç farklı yöntem uygulayarak üç değiştirilmiş versiyon geliştirmiştir. Ardından, diğer iki kişi, birinci kişinin yaptığı gibi aynı yöntemleri kullanarak tweetleri manuel olarak değiştirilmiştir. Örneğin, bir tweet için birinci kişi harf sıralarını değiştirme, hashtag ekleme ve karakterleri değiştirme yöntemlerini uygulayarak üç değiştirilmiş versiyon oluşturduğu durumda, diğer kişiler de aynı teknikleri o tweet için uygulamışlardır. Buna göre belirli örnekler için aynı yöntemi uygulamaları gereklidir, ancak nasıl uygulayacakları konusunda kişiler serbest bırakılmışlardır. Ayrıca her kişinin yaptığı değişiklikler bir diğer kişi tarafından kontrol edilmiştir. Örneğin, farklı kelimelerin karakterlerini değiştirebilir veya farklı hashtag'ler oluşturabilmektedirler. Bu yaklaşım ile aynı yöntemi uygulayarak tweetin farklı versiyonlarını oluştururken her yöntem için deneme sayısını kontrol edilmesi sağlanmıştır. En nihayetinde, toplam 738 (= 82×3×3) manuel olarak değiştirilmiş tweet üretilmiştir.

Çizelge 7.7'da manuel değişlikler için seçilen örneklemdeki konu dağılımını ve hassas ayarı yapılmış olan BERT ile Twitter-RoBERTa modellerinin tweetler üzerinde herhangi bir değişiklik yapılmadan önceki doğruluk oranlarını göstermektedir. BERT, Twitter-RoBERTa'ya göre kürtajın yasallaştırılması konusu dışında daha başarılı performans sergilemiştir.

Çizelge 7.7 : Manuel olarak değiştirdiğimiz tweet sayısı ve her konu tahmini için orijinal tweet'ler kullanıldığında hassas ayarlı BERT ve Twitter-RoBERTa modellerinin doğruluğu

Konu	**Tweet Sayısı**	**BERT**	**Twitter-RoBERTa**
Ateizm	21	0.952	0.905
İklim Değişikliği	11	0.909	0.727
Feminizm	16	0.813	0.688
Hillary Clinton	19	0.947	0.737
Kürtajın Yasallaştırılması	15	0.600	0.867
Tamamı	**82**	**0.854**	**0.793**

Çizelge 7.8'de her yöntem için yapılan deneme sayısı, BERT ve Twitter-RoBERTa modellerinin çıktısı olarak doğru/yanlış tahmin oranları gösterilmektedir. Deneme sayısı her yöntem için farklılık gösterebilmektedir. Bunun nedeni ise bazı yöntemlerin yalnızca belirli durumlara uygulanabilir olmasıdır. Örneğin, "Deyim Kullanma" yöntemini uygulamak için, deyimin kullanılabileceği belirli bir ifade olması gerekmektedir. Her yöntem için değişen deneme sayıları nedeniyle, her yöntemin deneme sayısına göre oranları ifade edilmiştir.

AS-1 ile ilgili olarak, sonuçlar tweetleri yeniden ifade etme yöntemlerinin her iki modelin tahminleri üzerinde sınırlı bir etkisi olduğunu göstermektedir. Bu durum, her ne kadar farklı kelimeler kullanılsa da her iki modelin de tweetlerin anlamını anlayabildiğini göstermiştir. Bununla birlikte, bilinenin dışındaki isimleri kullanılma yöntemi bazı durumlarda modellerin tahminlerini değiştirmekte etkili olabildiği gözlemlenmiştir. Ancak, Twitter-RoBERTa yanlış tahminleri doğruya çevirmede %19 oranında etkili olduğu görülmüştür. Beklenenin dışında, zıt anlamlı kelimeleri bir arada kullanmanın BERT tahminleri üzerinde herhangi bir etkisi olmazken, Twitter-RoBERTa'yı %22 oranında yanılttığı gözlemlenmiştir.

Aynı zamanda, Sun vd. [33] çalışmasının bulguları tekrarlanarak, her iki modelin de yazım hatalarına karşı oldukça hassas olduğu gözlemlenmiştir. Karakterleri görsel olarak benzer olanlarla değiştirildiğinde, önemli kelimelerin harfleri arasına boşluk ekleyerek bölündüğünde ve kelimelerin harf sıralarını karıştırıldığında, BERT modelini yaklaşık üçte birlik bir oranda aldatmayı başarabildiği görülmektedir. BERT modeli için boşlukları silme işleminin diğer yazım hatası temelli yöntemlere göre daha az etkili olduğu gözlemlenmiştir. Bunun nedeni, BERT belirtecinin, bazı durumlarda

hiç boşluk olmadan ardışık iki kelimeyi doğru bir şekilde anlamlandırabilmesidir. (Örn. “ridiculousweather”)

Twitter-RoBERTa, orijinal tweetlerde BERT modeline kıyasla daha düşük performansa sahip olmasına rağmen Çizelge 7.8’de, yazım hatalarına dayalı yöntemlerden, BERT modeline göre daha az etkilenmektedir.

Çizelge 7.8 : Manuel metin değişikliklerinin BERT ve Twitter-RoBERTa modellerinin performansı üzerindeki etkisi. D, Doğru anlamına gelir ve Y, Yanlış anlamına gelir. D → Y, ilgili metin değiştirme yöntemini kullanarak karşılık gelen modelin doğru tahminini yanlış bir tahminle değiştirebileceği durumların oranını gösterir. Benzer şekilde, Y→D, yanlış bir tahminin doğru bir tahminle değiştirildiği durumların sayısını gösterir. Y→Y ve D→D, tahmini hiç değiştirmeyen durumların sayısını gösterir.

	Yöntemler	**Deneme Sayısı (N)**	**BERT**				**Twitter-RoBERTa**			
			D → D	Y → Y	D → Y	Y → D	D → D	Y → Y	D → Y	Y → D
Yazım Hataları	Karakter Değiştirme	125	52%	14%	32%	2%	50%	20%	22%	7%
	Boşluk Ekleme	84	62%	7%	31%	0%	64%	10%	23%	4%
	Harf Sıralarını Karıştırma	93	52%	16%	32%	0%	66%	11%	18%	5%
	Boşluk Silme	48	69%	25%	6%	0%	75%	13%	13%	0%
	Hashtag İşareti Ekleme	90	81%	9%	10%	0%	74%	18%	4%	3%
Yeniden İfade Etme	Hashtag Silme	55	64%	11%	20%	5%	73%	16%	9%	2%
	Eş Anlamlısıyla Değiştirme	81	79%	14%	7%	0%	72%	17%	9%	2%
	Yeni Hashtag Ekleme	75	71%	16%	9%	4%	76%	24%	0%	0%
	Zıt Anlamlı Kelimeleri Bir Arada Kullanma	9	100%	0%	0%	0%	78%	0%	22%	0%
	Bilinenin Dışındaki İsimleri Kullanma	21	76%	0%	24%	0%	48%	24%	10%	19%
	Deyim Kullanma	27	96%	0%	4%	0%	93%	0%	0%	7%
	Kelime Silme	12	75%	25%	0%	0%	50%	50%	0%	0%
	Olumsuz İfadeleri Birlikte Kullanma	18	72%	17%	6%	6%	67%	22%	6%	6%

Bu, Twitter-RoBERTa'nın (tipik olarak gürültülü) tweetlerle önceden eğitildiği için yazım hatalarını daha etkili bir şekilde işleyebilme yeteneğine sahip olmasından kaynaklanmaktadır.

Hashtag'lerin BERT modeli için oldukça önemli olduğu görünmektedir. Hashtag'lerin silinmesi doğru bir tahmini yanlış bir tahmine dönüştürebildiği durum %20'lik bir oranda gerçekleşmektedir. Bununla birlikte dikkat edilmesi gereken bir nokta, hashtag'leri silinerek her iki modelin de yanlış tahminlerini doğru tahminlere dönüştürme durumunun da %5 oranında olmasıdır. Nötr yeni hashtag'ler eklemek veya bazı kelimeleri hashtag'lere dönüştürmek de BERT tahminlerinin sırasıyla %9 ve %10'unda yanlış tahminlemesine neden olmaktadır.

Buna karşılık, Twitter-RoBERTa, BERT'e göre hashtag değişikliklerine daha dayanıklı olduğu görünmektedir. Twitter-RoBERTa'nın performansı, hashtag'leri eklemekten etkilenmemekte ve hashtag işareti eklemek ve hashtag silmekten minimal seviyede etkilenmektedir.

Çeşitli konulardaki görüşleri nedeniyle izlenmek veya tespit edilmek istemeyen kişiler için, tahmin edilen taraf değerinin "NÖTR" olacak şekilde değiştirilmesi, kendi görüşüne zıt bir tarafta algılanmasından daha önemli olabilmektedir. Fakat manuel olarak değiştirdiğimiz tweetlerin hiçbirinin "NÖTR" olduğuna dair bir etiket bulunmamaktadır. Ancak tahmin için orijinal tweetleri kullandığımızda, BERT için altı adet tweet "NÖTR" olarak tahmin edilirken, Twitter-RoBERTa için bu sayı sıfırdır.

Değiştirilen tweetlerden oluşan 738 örnekten, BERT 126'sı, Twitter-RoBERTa 129'unu "NÖTR" olarak tahmin etmiştir. Değiştirilmiş tweetlerin tahminleri nötr taraf değerine değiştirmede kısmen etkili olduğunu düşündürmüştür. Bu durum, belirli konulardaki görüşleri veya inançlarının tanınmasından kaçmak isteyen bireyler için değerli bir bulgu olarak değerlendirilebilir.

Manuel yapılan değişikliklerin, tweetleri değiştiren kişilere göre taraflı olabilmesi nedeniyle, bu yöntemlerin etkinliğinin değişip değişmediği kişilere göre incelenmiştir. Çizelge 7.9 ve Çizelge 7.10 her yöntem ve tweet'leri değiştiren her kişi için BERT ve

Twitter-RoBERTa'nın tahmin değişikliklerinin sayısını göstermektedir. Değişikliği yapan kişiler K1, K2 ve K3 olarak ifade edilmiştir. Genel olarak, yöntemlerin performansının insanlar arasında benzer olduğunu ve yöntemlerin karşılaştırılmasıyla ilgili sonuçları değiştirmediği gözlemlenmiştir. Bununla birlikte, BERT için K1, K2 ve K3'ün sırasıyla 46, 51 ve 42 durumda; Twitter-RoBERTa için K1, K2 ve K3'ün sırasıyla 38, 29 ve 24 durumda doğru tahminleri yanlış tahminlere değiştirebildiğini gözlemlenmiştir, bu da yöntemlerin nasıl uygulandığının da kısmen önemli olduğunu vurgulamaktadır. Genel olarak, yeniden ifade yöntemleri, yazım hata temelli yöntemlere göre insanlar arasında daha kararlı sonuçlar vermektedir. Örneğin, "Kelime Silme" ve "Zıt Anlamlıları Bir Arada Kullanma" yöntemlerinde, tüm değiştirenlerin performansı tamamen aynıdır, muhtemelen bu durum yöntemlerin sınırlı esnekliklerinden kaynaklanmaktadır.

Çizelge 7.9 : Deneye katılan her bir kişinin (K1, K2 ve K3 olarak temsil edilmektedir.) manuel metin değişikliklerinin BERT modelinin tahminleri üzerindeki etkisi. Her tahmin değişikliği türü için örnek sayısı da gösterilmiştir. Sadeleştirmek için Y → Y sonuçlarını atılmıştır.

		D → D			D → Y			Y → D		
	Yöntemler	**K1**	**K2**	**K3**	**K1**	**K2**	**K3**	**K1**	**K2**	**K3**
Yazım Hataları	Karakter Değiştirme	24	20	21	12	14	14	1	0	1
	Boşluk Ekleme	18	15	19	8	11	7	0	0	0
	Harf Sıralarını Karıştırma	16	13	19	10	13	7	0	0	0
	Boşluk Silme	11	12	10	1	0	2	0	0	0
	Hashtag İşareti Ekleme	22	26	25	5	1	3	0	0	0
Yeniden İfade Etme	Hashtag Silme	11	13	11	4	3	4	1	1	1
	Eş Anlamlısıyla Değiştirme	20	21	23	3	2	1	0	0	0
	Yeni Hashtag Ekleme	18	16	19	2	4	1	2	1	0
	Zıt Anlamlı Kelimeleri Bir Arada Kullanma	3	3	3	0	0	0	0	0	0
	Bilinenin Dışındaki İsimleri Kullanma	6	4	6	1	3	1	0	0	0
	Deyim Kullanma	9	9	8	0	0	1	0	0	0
	Kelime Silme	3	3	3	0	0	0	0	0	0
	Olumsuz İfadeleri Birlikte Kullanma	5	5	3	0	0	1	1	0	0
Toplam		166	160	170	46	51	42	5	2	2

Çizelge 7.10 : Deneye katılan her bir kişinin (K1, K2 ve K3 olarak temsil edilmektedir.) manuel metin değişikliklerinin Twitter-RoBERTa modelinin tahminleri üzerindeki etkisi. Her tahmin değişikliği türü için örnek sayısı da gösterilmiştir. Sadeleştirmek için Y → Y sonuçlarını atılmıştır.

	Yöntemler	D → D			D → Y			Y → D		
		K1	**K2**	**K3**	**K1**	**K2**	**K3**	**K1**	**K2**	**K3**
Yazım Hataları	Karakter Değiştirme	20	21	22	11	9	8	2	1	6
	Boşluk Ekleme	16	17	21	8	8	3	2	0	1
	Harf Sıralarını Karıştırma	17	22	22	9	4	4	2	1	2
	Boşluk Silme	12	13	11	2	1	3	0	0	0
	Hashtag İşareti Ekleme	22	23	22	2	1	1	2	0	1
Yeniden İfade Etme	Hashtag Silme	13	14	13	2	1	2	0	1	0
	Eş Anlamlısıyla Değiştirme	20	19	19	2	3	2	1	0	1
	Yeni Hashtag Ekleme	19	19	19	0	0	0	0	0	0
	Zıt Anlamlı Kelimeleri Bir Arada Kullanma	2	3	2	1	0	1	0	0	0
	Bilinenin Dışındaki İsimleri Kullanma	4	2	4	0	2	0	2	2	0
	Deyim Kullanma	8	8	9	0	0	0	1	1	0
	Kelime Silme	2	2	2	0	0	0	0	0	0
	Olumsuz İfadeleri Birlikte Kullanma	3	4	5	1	0	0	1	0	0
Toplam		158	167	171	38	29	24	13	6	11

7.5 Otomatik Değiştirilen Metinlerdeki Deney Sonuçları Performansı

7.5.1 İngilizce veri kümesi için otomatik sonuçlar

İngilizce veri kümesi için önerilen manuel deney düzeneği, bir tweet alt kümesinin insanlar tarafından manuel olarak değiştirilmesini içermekteydi. Bu bölümde, tüm veri kümesine otomatik olarak uygulanan yöntemlerin etkileri anlatılmıştır. İlgili deney düzeneğinin yürütülmesi için aşağıdaki adımlar atılmıştır.

Manuel olarak yapılan deney düzeneğinde, yeniden ifade yöntemlerinin modelleri aldatmada beklenen seviyede etkili olmadığını göstermiştir. Bununla birlikte, bu yöntemleri otomatik olarak uygulanması da ayrıca bir zorluk oluşturmaktadır. Bu nedenle, bu deneylerde potansiyel olarak etkili ve kolayca otomatik olarak uygulanabilecek yöntemlere odaklanılmıştır. Bu kapsamda odaklanılan yöntemler şunlardır: “Hashtag İşareti Ekleme”, “Hashtag Ekleme”, “Hashtag Silme”, “Karakter Değiştirme”, “Harf Sıralarını Karıştırma”, “Boşluk Ekleme” ve “Boşluk Silme”.

Manuel yapılan değişikliklerde, değiştirilmesi gereken kelime sayısı üzerinde herhangi bir kısıtlama koyulmamıştır. Ancak otomatik değişiklikler için, N olarak gösterilen bir

parametre kullanılmıştır. Bu parametre, değiştirilecek kelime sayısını veya kaldırılacak/eklenecek hashtag sayısını belirlemektedir. Deneylerde N değeri 0 ile 4 arasında değiştirilmektedir.

Manuel yapılan deneylerde, üzerlerinde değişiklik yapılacak önemli kelimeler manuel olarak seçilmiştir. Otomatik olan bu deney düzneğinde, değiştirilecek kelimeleri seçmek için Bölüm 5.1'de detaylı olarak bahsedilen FastText tabanlı yöntem kullanılmıştır.

İstisnai bir durum olarak, "Boşluk Silme" yöntemini ardışık kelimelere uygulamak, tweetlerin okunurluğunu azaltabilmektedir. Bu nedenle, bu yöntem için N tane ardışık olmayan kelime seçilmesine özellikle dikkat edilmiştir. Benzer şekilde diğer bir istisnai durum da "Hashtag İşareti Ekleme" yönteminde de hashtag'e dönüştürülecek kelimeleri tespit edilmesi için "FastText" yöntemine göre konuya en uzak olan yani önemsiz kabul edilen kelimeler üzerlerinde değişiklik yapılmıştır.

"Hashtag Ekleme" yönteminde, her konu için manuel olarak bir hashtag listesi tanımlanmıştır ve bunlardan N değerine göre kullanılmıştır. Örneğin, kürtajın yasallaştırılması konusunun hashtag'leri #MondayMotivation, #goals, #opinion ve #thoughts kelimelerinden oluşmaktadır. N değeri arttıkça sırasıyla bu hashtag'ler tweet sonlarına eklenmektedir.

Ayrıca "Harf Sıralarını Karıştırma" yönteminde seçilen kelimenin 7 veya daha fazla harfi varsa sadece 2., 3., 4. ve 5. harflerin yerlerini değiştirilmektedir, bir harfin orijinal konumu ile karışık versiyonundaki konumu arasındaki mesafe en fazla üç olarak tutulmaktadır. Bu şekilde tasarlanmasının nedeni, bir harfin konumu çok değiştiğinde yüksek olasılıkla kelimelerin okunurluğunun azalmasına sebep olmasıdır. Amaç okunurluğun azalmasını minimum seviyede tutmaktır.

Bu deney düzeneğinde, her bir tweet'i yukarıda açıklandığı gibi test veri kümesini değiştirilmiş ve orijinal eğitim verilerine göre hassas ayarlanmış BERT ve Twitter-RoBERTa modellerinin performansları incelenmiştir. Sonuçlar, Şekil 7.1'de gösterilmiştir.

Elde edilen sonuçlara göre, “Hashtag Silme” diğer yöntemlere göre en az etkili yöntem gibi görünmektedir. Diğer taraftan, manuel olarak değiştirilen metinlerle yaptığımız deneylerde olduğu gibi, “Karakter Değiştirme”, “Boşluk Ekleme” ve “Harflerin Sıralarını Karıştırma” yöntemleri, N=4 olduğunda ortalama olarak beş konuda BERT modelinin performansını sırasıyla %28, %27 ve %23 azalttığı görülmektedir. Benzer şekilde, “Karakter Değiştirme”, “Boşluk Ekleme” ve “Harflerin Sıralarını Karıştırma” yöntemleri, N=4 olduğunda ortalama olarak beş konuda Twitter-RoBERTa modelinin performansını sırasıyla %20, %25 ve %20 azalttığı tespit edilmiştir.

“Hashtag Ekleme” yöntemine yönelik incelemeler sonucunda karmaşık sonuçlar üretilmiştir. Bu yöntemin çoğu durumda her iki modelin performansını hafifçe etkilediği gözlenirken, ateizm konusunda BERT modelinin performansını düşürdüğü, Twitter-RoBERTa'nın ise performansını artırdığını gözlemlenmiştir. “Hashtag İşareti Ekleme” yöntemi için de benzer bir örüntü görülmektedir.

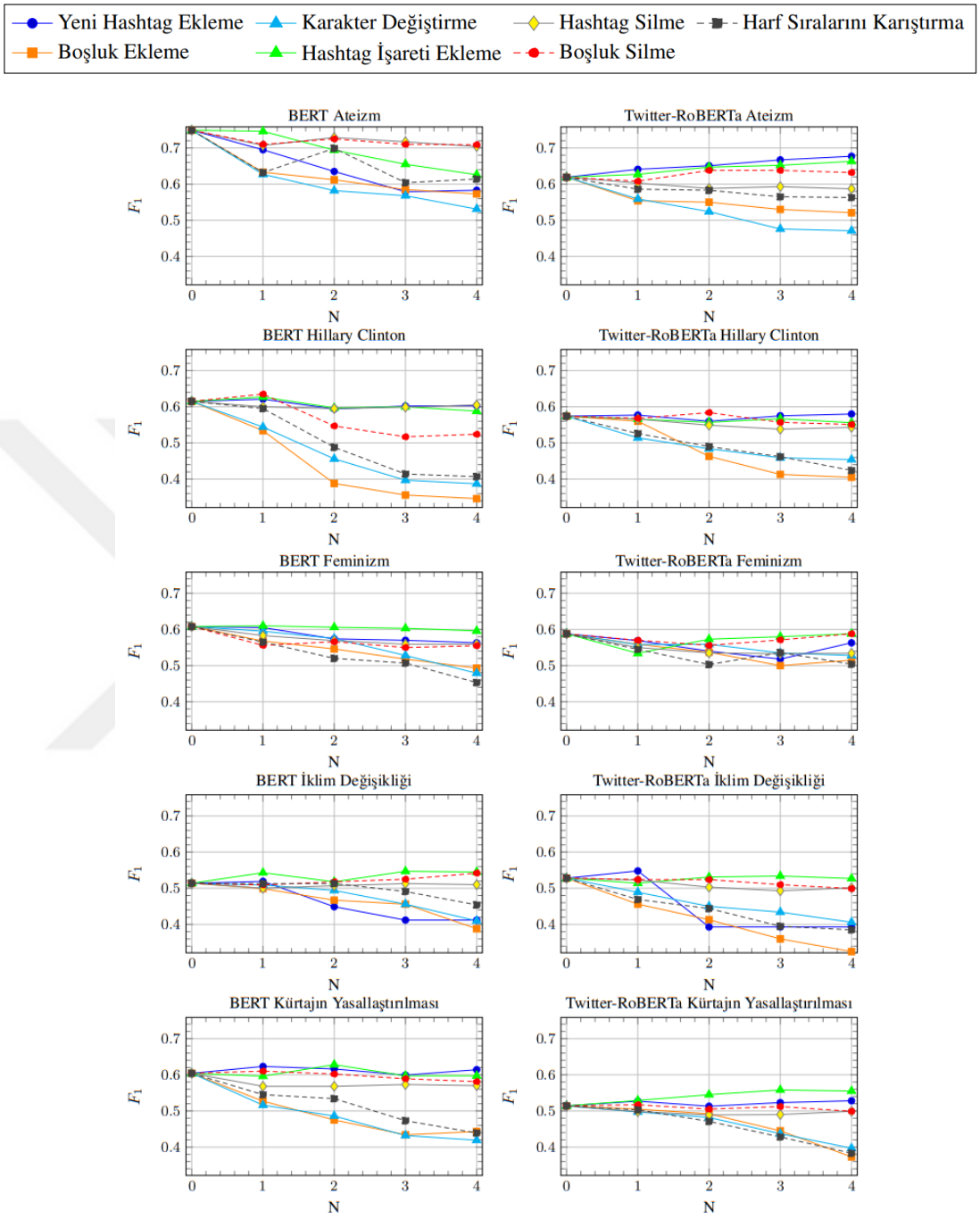
Bu sonuçlar, hashtag'lerin eğitim verilerindeki etiketlerle ilişkili olabileceği fikrini düşündürmektedir. Bu nedenle, modellerin eğitim verilerini bilmeden hashtag'leri kullanmak riskli olmaktadır.

BERT ve Twitter-RoBERTa modellerini karşılaştırma açısından, N=4 olduğunda BERT, Twitter-RoBERTa'dan daha yüksek performans sağlamaktadır. Beklenilen durumda, Twitter-RoBERTa'nın gürültülü verilerle ön eğitildiği için yapılan yazım hatalarından daha az etkileneceği yönünde olmakla birlikte, otomatik değişikliklerle yapılan deneylerde, modellerin göreceli performans değişiklikleri arasında anlamlı bir fark bulunmamaktadır.

7.5.2 Türkçe veri kümesi için otomatik sonuçlar

İngilizce veri kümesinde yapılan çalışmalar sonucunda otomatik olarak uygulanması daha zor olan yeniden ifade yöntemlerinin modelleri aldatmada beklenen seviyede etkili olmaması, yazım hatalarına dayalı otomatikleştirilebilecek yöntemlerin daha etkili olması sebebiyle, Türkçe veri kümesi üzerinde yapılan çalışmalarda otomatikleştirilmiş yöntemlerin kullanımı tercih edilmiştir. Bu kapsamda deney düzeneğinde kullanılan yöntemler şunlardır: “Hashtag İşareti Ekleme”, “Hashtag

Ekleme”, “Hashtag Silme”, “Karakter Değiştirme”, “Harf Sıralarını Karıştırma”, “Boşluk Ekleme” ve “Boşluk Silme”.



Şekil 7.1 : Değişken deneme sayıları için SemEval2016’nın ilgili veri setinin test verilerinde taraf tespiti görevinde hassas ayarlı BERT ve Twitter-RoBERTa modellerinin performansı. Örneğin N= 4, ilgili yöntemin dört kelime için uygulandığı anlamına gelir. İlk kolonda solda BERT’in F1 skoru, ikinci kolonda sağda ise Twitter-RoBERTa’nın F1 skoru gösterilmiştir.

Türkçe veri kümesi siyasi partilere yönelik kullanıcı hesaplarının taraflarının tespit edilmesi yönünde olup tek bir konu üzerine 5 farklı siyasi parti taraf etiketi bulunmaktadır.

İngilizce otomatik değişiklik deney düzeneği için kullanıldığı gibi bu sistemde de N olarak gösterilen bir parametre kullanılmıştır. Bu parametre, üzerine değişiklik uygulanacak kelime sayısını veya kaldırılacak/eklenecek hashtag sayısını belirlemektedir. Deneylerde N değeri 0 ile 4 arasında değiştirilmektedir.

Otomatik olarak kelimeler üzerinde değişikliklerin yapılması öncelikle önemli/önemsiz kelimelerin tespit edilmesi işlemi gerçekleştirilmiş. Önemli kelimelerin belirlenmesiyle ilgili yöntem detayları Bölüm 5. ’te verilmiş olup bu deney düzeneğinde kullanılan yöntem bu çalışma için üretilmiş olan BERTurk Tabanlı Önemli/Önemsiz Kelime Tespit Modelidir. Bu model kullanılarak önemli/önemsiz kelime tespiti işlemi gerçekleştirilmiştir.

Yöntemler gerçekleştirilirken N sayısınca “ÖNEMLİ” olarak etiketlenmiş kelime kullanılarak bu kelimeler üzerinde çeşitli yöntemler uygulanmaktadır. Cümlelerdeki “ÖNEMLİ” kelime sayısı N değerinden küçük olması halinde ise diğer “ÖNEMSİZ” olarak etiketlenmiş kelimeler üzerinden değişiklikler yapılmaya devam edilmektedir.

Diğer yöntemlerden farklı olarak “Hashtag İşareti Ekleme” yönteminde hashtag'e dönüştürülecek kelimeleri tespit edilmesi için öncelikli olarak “ÖNEMSİZ” olarak etiketlenmiş kelimeler üzerinde sonrasında N değerinin “ÖNEMSİZ” kelime sayısından küçük olması halinde ise ilgili değişiklikler “ÖNEMLİ” olarak etiketlenmiş kelimeler üzerinde yapılmaktadır.

“Hashtag Ekleme” yönteminde, manuel olarak bir hashtag listeleri tanımlanmıştır ve bunlardan N değerine göre kullanılmıştır. Bu kapsamda hashtag etkisinin incelenmesi için 5 farklı konuda hashtag listesi oluşturulmuştur. Bunlar Çizelge 7.11’de listelenmiştir.

Çizelge 7.11: Farklı konulardaki hashtag ekleme yöntemi için kullanılan hashtag listeleri

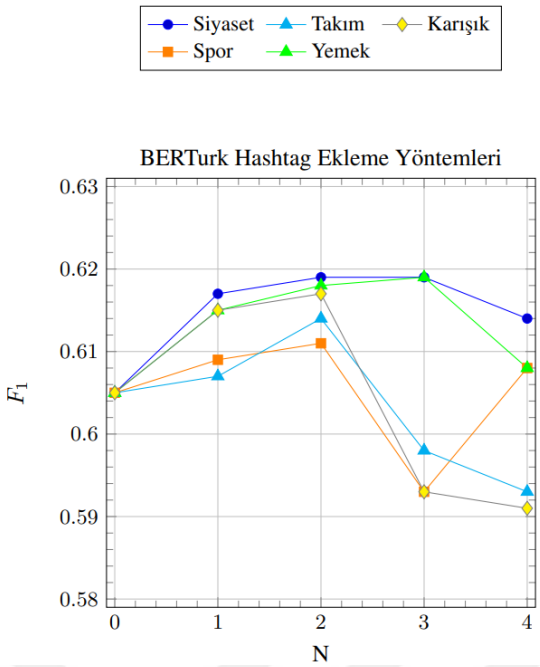
Konu	**Hashtag Listesi**
Siyaset	#seçim #gündem #haberler #siyaset
Spor	#futbol #voleybol #okçuluk #yüzme
Takım	#fb #gs #bjk #ts
Yemek	#pizza #makarna #kahve #erik
Karışık	#pizza #gündem #okçuluk #ts

Metinler üzerinde değişiklikler yapılmadan önce Bölüm 4.2'de anlatılan BERTurk tabanlı taraf tespit modelinin makro ortalama F1 skor değeri 0.605'dir. Eğer herhangi bir yapay zekâ modeli kullanılmaması, rastgele 1 ile 5 arası sayı üretme şeklinde gerçekleştirilip her bir sayı bir taraf etiket değerine karşılık gelecek şekilde test kümesindeki her bir tweet için yüzer kere gerçekleştirildiğinde ise makro ortalama F1 skor değeri 0.190 olmuştur. Modelimizin rastgele yöntemine göre oldukça başarılı olduğu görülmektedir.

"Hashtag Ekleme" yönteminde, manuel olarak tanımlanan Çizelge 7.11'de listelenmiş olan 5 farklı konulara dair hashtag'lerin eklenerek elde edilen sonuçlar Şekil 7.2'de gösterilmiştir. Elde edilen sonuçlara göre özellikle N=4 için "Karışık" ve "Takım" konulu hashtag ekleme yöntemleri daha etkili olup başarılı bir şekilde sistem performansını düşürmüştür. Bu sebeple "Karışık" konulu hashtag ekleme yönteminin seçilerek devam edilmiştir.

Bu deneyde, her bir tweet'i yukarıda ve Bölüm 4.2'de açıklandığı gibi test veri kümesi değiştirilmiş ve buna göre BERTurk tabanlı taraf tespit modelinin genel performansı incelenmiştir. Hashtag Ekleme yöntemi olarak en etkili olan hashtag listesi "Karışık" olarak adlandırılan yani "#pizza #gündem #okçuluk #ts" hashtag'lerini içeren listedir. Diğer Türkçe veri üzerinde yapılan deney çalışmalarında hashtag ekleme yöntemleri bu liste kullanılarak gerçekleştirilmiştir.

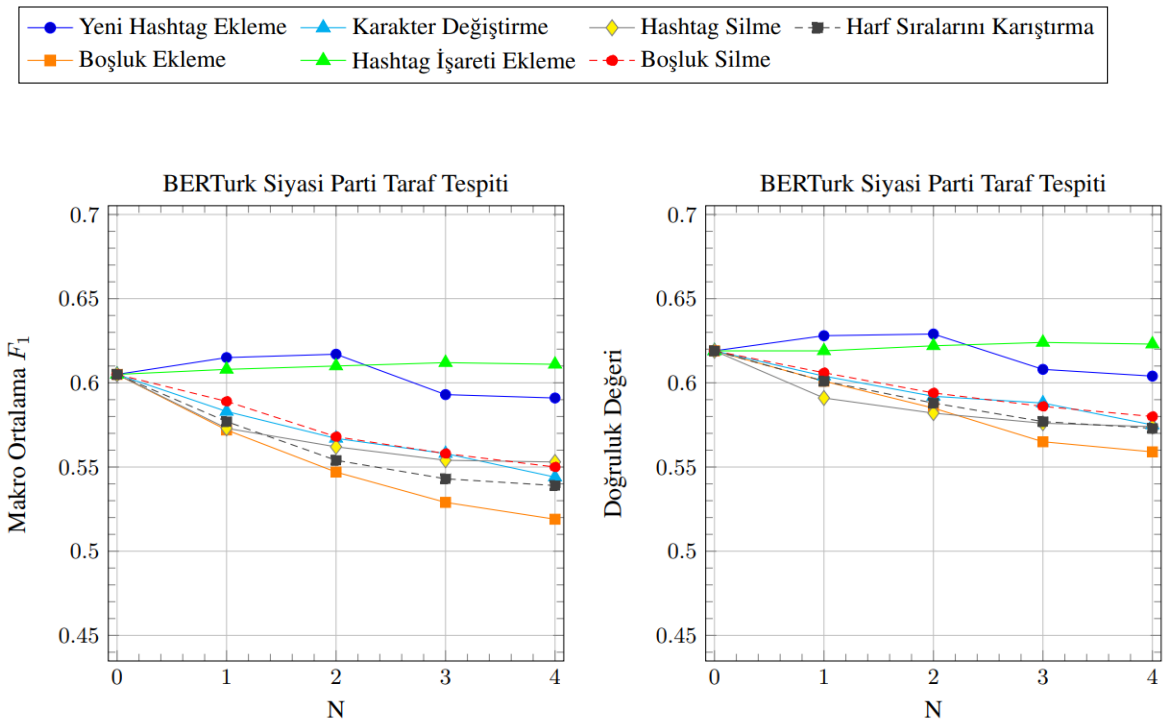
Etiket türü ayırt etmeksizin ortalama performans değerleri Şekil 7.3'de gösterilmiştir. Elde edilen sonuçlar incelendiğinde ilk olarak, "Yeni Hashtag Ekleme" ve "Hashtag İşareti Ekleme" yöntemlerinin diğer yöntemlere göre en az etkili yöntemler oldukları görülmektedir.



Şekil 7.2 : Çeşitli hashtag ekleme listelerinin çeşitli N değerlerinde uygulanması sonucunda elde edilen BERTurk performans değerleri

”Yeni Hashtag Ekleme” yönteminin N değeri arttıkça sistem performansını düşürmek yerine arttırdığı ve en etkisiz yöntem olduğu görülmektedir. Bunun sebebi hassas ayarın sosyal medya verisi üzerinden yapılması sonucunda modelin hashtag’lerle metni daha iyi anlamasına sebep olması gösterilebilir. Bununla birlikte metin üzerinde aslında herhangi bir bozulma olmayıp sadece bazı kelimeler hashtag’e dönüşmüştür. Bu kelimeler ÖNEMSİZ kelimelerden seçilmesine rağmen model tarafından metnin daha iyi anlaşılmasına sebep olmuş olabilir. “Hashtag İşareti Ekleme” yönteminde ise N=1 ve N=2 için hashtag eklendikçe modelin genel performansı amacımızın tersi olarak arttığı görülmüştür.

Bunların yanı sıra diğer yöntemlerin genellikle etkili olduğu ancak ortalama olarak en etkili yöntemlerin “Boşluk Ekleme”, “Harf Sıralarını Karıştırma” ve “Karakter Değiştirme” olduğu, N=4 için sırasıyla model başarısını %8.2, %6.1 ve %5.9 düşürdükleri gözlemlenmiştir. Her etiket için uygulanan yöntemlerin etkinlerine yönelik detaylı sonuçlar ise Şekil 7.4’de gösterilmiştir.



Şekil 7.3 : Model bozma yöntemlerinin çeşitli N değerlerinde uygulanması sonucunda elde edilen BERTurk makro ortalama F1(sağ) ve doğruluk (sol) performans değerleri

Her bir taraf etiketine göre yöntemlerden elde edilmiş performanslar incelendiğinde AK Parti ve İYİ Parti verileri için en iyi 3 yöntemin "Boşluk Ekleme", "Harf Sıralarını Karıştırma" ve "Hashtag Silme" olduğu N=4 için model performansını sırasıyla %4.9, %4 ve %3.3 düşürmüştür. CHP verisi için de en etkili yöntemler benzer olup farklı olarak "Boşluk Silme", "Boşluk Ekleme" yönteminden özellikle N=4 için model performansını sırasıyla %4.2, %3.3 düşürerek etkili bir yöntemler oldukları görülmüştür.

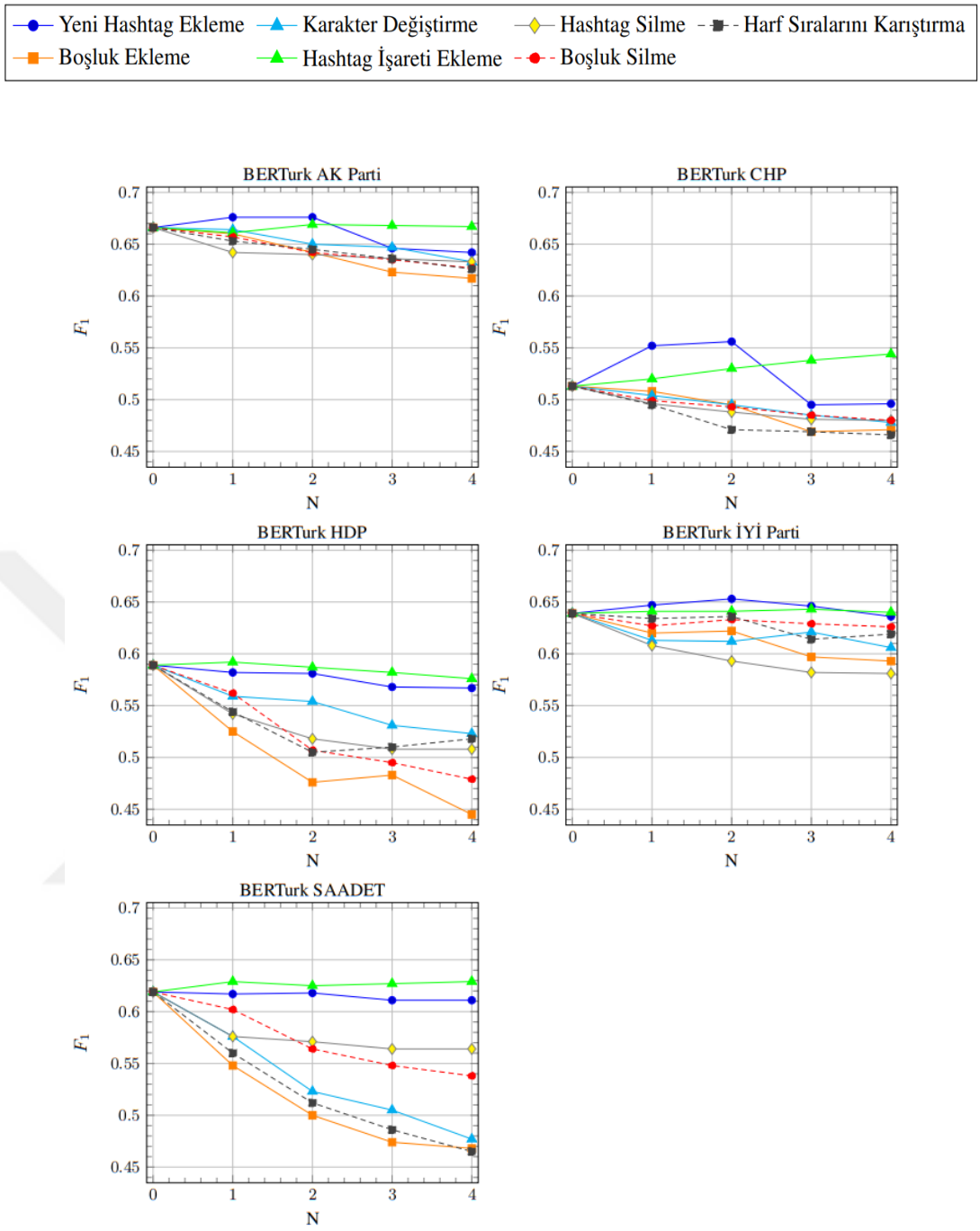
HDP verileri için yapılan deneylerde ise N=4 için "Boşluk Ekleme" %14.4, "Boşluk Silme" %11 ve "Hashtag Silme" %8.1 performansı düşürerek en etkili yöntemler iken "Harf Sıralarını Karıştırma" ve "Karakter Değiştirme" yöntemlerinin etkinlik seviyeleri birbirlerine oldukça yakın olup orta etkili yöntemler arasında yer almıştır. İYİ Parti verisi için ise "Hashtag Silme" yönteminin N=4 için %5.8 performans düşüşüyle en etkili yöntem, "Boşluk Ekleme" ve "Karakter Değiştirme"nin sonraki en etkili yöntemler olduğu, N=4 için sırasıyla %4.6 ve %3.3 başarıyı düşürdüğü görülmüştür. SAADET verilerinde ise "Harf Sıralarını Karıştırma", "Boşluk Ekleme"

ve “Karakter Değiştirme” yöntemleri sırasıyla %15.4, %15.1 ve %14.2 olacak şekilde sistem performansını düşürmüştür.

Tüm taraflara ait veriler üzerinde performansın düşürülebilmesi ve yöntemlerin en etkili oldukları taraf verileri incelendiğinde performans düşürme oranları en çoktan en aza göre sırasıyla şöyledir: SAADET’de %15.4, HDP’de %14.4, İYİ Parti’de %5.8, AK Parti’de %4.9 ve CHP’de %4.7. Yapılan çalışmalar sonucunda özellikle SAADET ve HDP tweetlerini bozmakta daha başarılı olduğu görülmektedir.

“Hashtag İşareti Ekleme”, HDP verileri dışında diğer verilerde modellerin performansını arttıracak yönde bir etkisi olmuştur. “Yeni Hashtag Ekleme” her veri grubu için oldukça farklı sonuçlar vermiştir. AK Parti, CHP ve İYİ Parti verilerinde N=2’ye kadar model performansını arttırırken sonrasında model performansını düşürmüştür. Bunun sebebi eklenen hashtag’lere bağlı olarak yorumlanabilir. “#pizza #gündem #okçuluk #ts” hashtag’lerinin ilgili N kadar eklenmesiyle uygulanılan yöntemde “#pizza” ve “#gündem” hashtag’leri eklenince model performansı artmış, “#okçuluk” ve “#ts” hashtag’leri eklenince azalmıştır. “#pizza” ve “#gündem” hashtag’leri CHP ile eşleşmiş olarak yorumlanabilir. Bu durumda aslında nötr hashtag eklemenin göründüğü kadar kolay olmadığı, beklenmedik bir şekilde model tarafından bir gruba yönelik algılanabildiğinin göstergesidir. “Yeni Hashtag Ekleme” yönteminde ise önemsiz kelimelerin hashtag yapılmasına rağmen yine benzer şekilde CHP verileri için model etkinliğini arttıran bir yöntem olmuştur. Bunun sebebi CHP’ye ait veri kümelerinde çok sayıda hashtag kullanılmasıyla hashtag sayısındaki artışın modelin CHP’ye ait olduğu fikrine kapılmasına sebep olmuş olabilir.

Her bir taraf etiketi için ortak olan şeylerden biri ise “Yeni Hashtag Ekleme” ve “Hashtag İşareti Ekleme” yöntemlerinin etkilerinin önceden tahmin edilemez seviyede ve modelleri yanıltmada en az etkili yöntemler olduğudur. Bunun sebebi ise yeni eklenen veya oluşturulan hashtag’lerin metinin anlamını değiştirme oranının oldukça düşük olması ve bazı durumlarda modelin metni daha iyi anlamasını sağladığı şeklinde değerlendirilebilir. Şekil 7.5 ve Şekil 7.6 okunabilirlik ve anlam değişimi incelendiğinde bu iki yönteme dair oluşturulan tweetlerde ne okunurluğun azaldığı ne de anlamsal bir değişikliğin olduğu görülmektedir.



Şekil 7.4 : Değişken deneme sayıları için test verilerinde taraf tespiti görevinde hassas ayarlı BERTurk modelinin Türkçe siyasi parti taraf etiketlerine göre performansı. Örneğin N= 4, ilgili yöntemin dört kelime için uygulandığı anlamına gelir.

7.6 Okunabilirlik ve Anlam Değişim Sonuçları

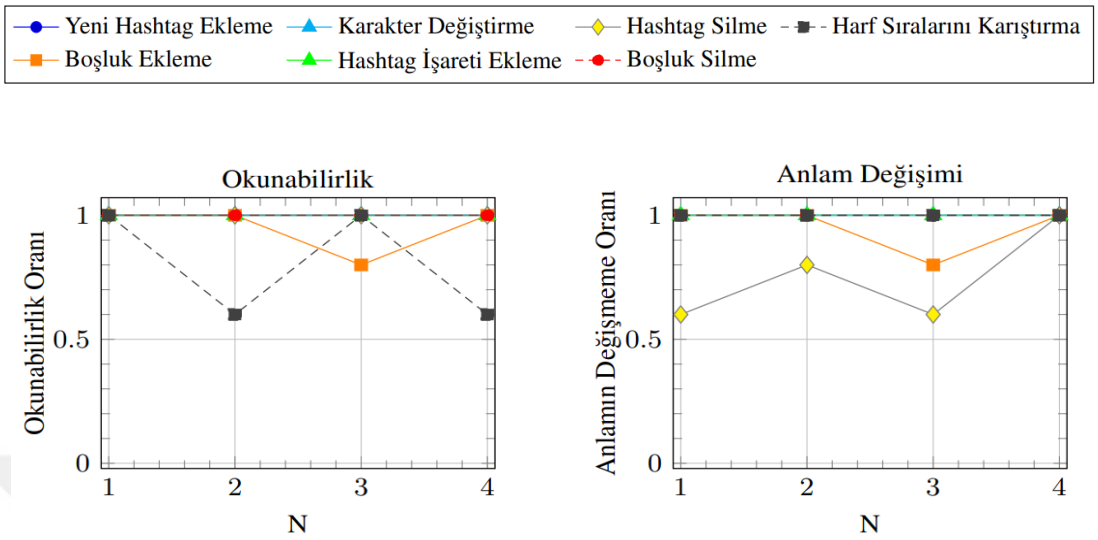
Manuel olarak insanlar tarafından yöntemlerin uygulanmasında, okunabilirliğin bozulmaması ve anlamın değişmemesi dikkate alınarak gerçekleştirilmiştir. Fakat, yöntemlerin Türkçe ve İngilizce veri kümeleri üzerinde otomatik olarak uygulanması

sonucunda değiştirilmiş olan verilerin okunabilirliğinde ve anlamlarının korunmasından emin olunacak herhangi bir otomatik yöntem bulunmayıp, otomatik yöntemlerin gerçek başarısından emin olabilmek için bu kontrollerin değerlendirici kişiler tarafından gerçekleştirilmesi gerekmektedir. Okunabilirlik ve anlam değişimi bakımında oluşturulan değerlendirmede her yöntem için N 1'den 4'e kadar olmak üzere toplamda 5'er adet değiştirilmiş örnek tweet seçilmiş olup toplamda 140 (=5x4x7) adet örnek iki kişi tarafından değerlendirilmiştir. Her iki kişi de aynı metinlerin okunabilirliği ve anlamındaki değişmeyi incelemiştir. Herhangi bir kişinin bir metne okunamaz demesi durumunda o metin diğer kişi tarafından okunabilir olsa bile okunamaz olarak kabul edilmektedir. Benzer şekilde anlam değişimi konusunda da herhangi bir kişinin anlamı değişmiş olarak metni değerlendirmesi sonucunda metin anlamı değişmiş olarak kabul edilmektedir. Yani okunabilir veya anlamı değişmemiş olarak değerlendirilen metinler herkesçe okunabilir veya anlamı değişmemiş olanlardır. AS-2 ile ilgili olarak, otomatik değişikliklerden sonra ortaya çıkan metinlerin okunabilir olup olmadığını ve orijinal tweet ile aynı/benzer anlama sahip olup olmadığı araştırılmıştır.

Şekil 7.5'de İngilizce veri kümesinde uygulanan her bir yöntem için incelediğimiz, tweet'ler arasında okunabilirliği ve orijinal tweet'lerle aynı/benzer anlama sahip tweet'lerin oranını göstermektedir. "Harflerin Sırasını Karıştırma" ve "Boşluk Ekleme" dışındaki tüm yöntemlerin okunabilir tweetler verdiği görülmüştür. "Harflerin Sırasını Karıştırma", N=2 ve N=4 olduğunda tweetlerin %40'ını okunamaz hale getirerek uygulanabilirliğini azaltmıştır. Örneğin, "*ny investing big bkaenr (banker) bdus (buds) need toratchet up their haillry (hillary) cares about the little polepe(people) propaganda*" İngilizce veri kümesine ait okunamaz olarak değerlendirilmiş bir tweettir. Tweet'teki kelimelerin doğru hallerinin parantez içinde yazılmıştır.

Anlam değişikliği açısından, "Boşluk Ekleme" ve "Hashtag Kaldırma" dışındaki yöntemlerin tweetlerin anlamını değiştirmediği gözlemlenmiştir. "Hashtag Silme" yöntemi anlam değişikliğine neden olmaktadır, çünkü insanların tweetlerde önemli kelimeler için hashtag'leri kullandığı sonucuna varılmıştır. Örneğin, *"agent 350 this isnot a fantasy this is negligence collusion with criminal cor-porations acting with*

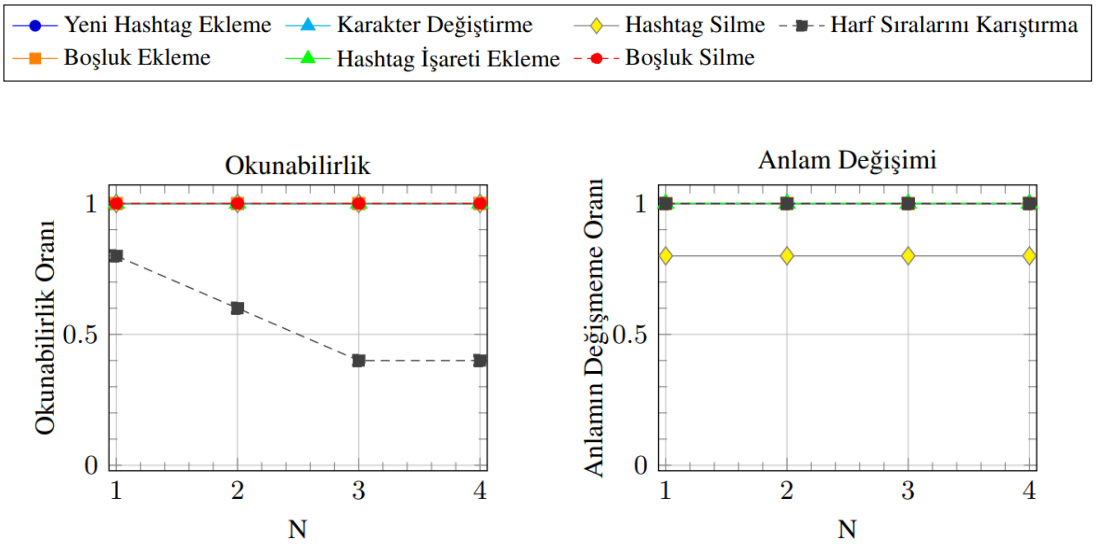
negligence to ~~#ecocide~~" tweet'i, kaldırılan hashtag (üstü çizili olarak gösterilen metin) cümlenin anlamı için önemli olup anlamı etkilemektedir.



Şekil 7.5 : İngilizce veri kümesi için değişen sayıda değiştirilen kelime için taraf tespiti görevinde okunabilirlik ve anlam değişim analizi.

"Harflerin Sırasını Karıştırma" yönteminin okunamayan tweetlerde anlam değişikliğine neden olmamasının nedeni, değiştirilen kelimelerin başka anlamlı bir kelime ifade etmemesidir. Bu nedenle, eğer bir kişi bunları doğru bir şekilde okuyabiliyorsa, anlamda herhangi bir değişiklik olmayacağını varsayılmıştır. Bununla birlikte, nitel analiz olarak, hem "Harflerin Sırasını Karıştırma" hem de "Hashtag Silme" yöntemlerinin tweetlerin okunabilirliğini ve anlamını değiştirmeden kalması için özel dikkat gerektirdiğini göstermektedir.

Şekil 7.6'de Türkçe veri kümesi için görüldüğü üzere okunabilirliğin, N değeri arttıkça "Harf Sıralarını Karıştırma" yöntemi için azaldığı fakat anlam açısından bozulmanın olmadığı gözlenmiştir. Bunun sebebi üretilen yeni kelimenin farklı bir kelime yerine koyulamayacak şekilde karıştırılmış olması ve bir kişi bunları doğru okuyabilirse anlamda da herhangi bir değişiklik olmayacağını düşüneceğini varsaydık. Bunun yanı sıra "Hashtag Silme"nin anlamdaki değişmeye yol açtığı gösterilmiştir.



Şekil 7.6 : Türkçe veri kümesi için değişen sayıda değiştirilen kelime için taraf tespiti görevinde okunabilirlik ve anlam değişim analizi.

Şekil 7.5 ve Şekil 7.6 incelenip Türkçe ve İngilizce veri kümelerinde otomatik deneyler sonucunda "Harflerin Sıralarının Karıştırılması" yönteminin dikkatli yapılması gerektiği ve özellikle Türkçe veri kümesinde yapılan değişiklikler sonucunda okunurluğu azaldığı görülmektedir. Bununla birlikte "Hashtag Silme" yönteminin her iki veri kümesinde de anlam kaybına yol açabildiği görülmüştür. Bu sebeple bu iki yöntemin uygulanması konusunda otomatik yöntemler için farklı denetim mekanizmaları veya değişiklikleri sınırlandırmak için çeşitli kurallar koyulması gerektiği şekilde yorumlanabilir.

7.7 Nitel Sonuçlar

Türkçe veri kümesi için yöntemler için okunabilirlikle ilgili veri kümesindeki örnekler Çizelge 7.12 'de verilmiştir. Çizelge 7.12'deki 1 numaralı örneğin okunurluğunu değerlendiren kişi tarafında "amreika" kelimesi "amerika" olarak ve "iasril" kelimesi "israil" olarak okunamaması sebebiyle okunamadığı yönünde karar verilmiştir. 2 numaralı örnekte "Doğu Perinçek kendisi" şeklinde okunabilmesi ve 3 numaralı örnekte de "Zeytinburnu" ve "Kurulunda" kelimeleri değerlendiriciler tarafından okunurluk yönünde herhangi sorun oluşturmadığı sebebiyle okunabilir olarak etiketlenmiştir.

Çizelge 7.12 : Yöntemlerin otomatik uygulanması sonucunda okunabilirlikle ilgili örnekler

Örnek Numarası	Yöntem	Deneme Sayısı (N)	Örnek Tweet	Okunabilirlik Değeri
1	Harflerin Sırasının Karıştırılması	4	#Iran #Müslüman #ıslam #Türkiye #katil **ırsail otardoğuyu** bulandırıyor yine göz yaşları #emparyalis **amreika** kahrolsun #BM **iasril** savaşa dur	OKUNAMAZ
2	Harflerin Sırasının Karıştırılması	3	**Dğou Peirnçek kndeisi** muhalif görünen yandaştır elden ele yayalım..lütfen kaybolsun	OKUNABİLİR
3	Karakter Değiştirme	3	Cumhurbaşkanlığı imza toplama süreci başladı...Tabi ki YSK ilk günden devrede!!! Şimdilik; 1) **Zeyt!nburnu** İlçe Seçim **Kuru|unda** Sistem çalışmıyor!! 2) Beşiktaş İlçe Seçim **Kuru|unda** sadece 1 bilgisayar işlem yapıyor!!! #ProtestoEdiyorum	OKUNABİLİR

Türkçe veri kümesi için yöntemler için anlam değişmesiyle ilgili veri kümesindeki örnekler Çizelge 7.13 ’de verilmiştir. Çizelge 7.13 ’de 1 numaralı örnek de üstü çizili ve koyu renkte gösterilmiş olan hashtag silinmiş olup silinen ifadenin anlamı hangi konsepte ait olduğu yani siyaset ya da edebiyat mı olduğunun anlaşılmasını etkilemesinden dolayı değerlendiriciler tarafından anlamı değişmiş olarak etiketlenmiştir.

Çizelge 7.13 ’deki 2 numaralı örnek de üstü çizili ve koyu renkte gösterilmiş olan hashtag silinmiş olup silinen ifadenin verilmek istenilen anlamı doğrudan etkilememesinden dolayı anlamı değişmemiş olarak değerlendiriciler tarafından etiketlenmiştir. Çizelge 7.13 ’deki 3 numaralı örnek de koyu renkte gösterilmiş olan kelimeler üzerinde karakter değişikliği yapılmış olup ifadede verilmek istenilen anlamı doğrudan etkilememesinden dolayı anlamı değişmemiş olarak değerlendiriciler tarafından etiketlenmiştir.

Uygulanan otomatik yöntemlerle yöntemlerinin modelleri yanılttığıyla ilgili örnek, değişiklikler ve BERT modelden elde edilen sonuçlar Çizelge 7.14’de gösterilmiştir. Uygulanan otomatik yöntemlerle yöntemlerinin modelleri yanıltamamasıyla ilgili örnek, değişiklikler ve BERT modelden elde edilen sonuçlar Çizelge 7.15’de gösterilmiştir. Çizelge 7.14’deki Yeni Hashtag Ekleme yöntemi modelin taraf tahminini değiştiren etkili bir yöntem örneklerindendir. Konuyla alakasız olarak

kullanılan hashtag'lerin modeli bu hashtag'lere normalde olması gerekenden daha fazla anlam yüklemesi ya da odaklanmasından kaynaklandığı şeklinde yorumlanabilir.

Çizelge 7.13 : Yöntemlerin otomatik uygulanması sonucunda anlam değişimiyle ilgili örnekler

Örnek Numarası	Yöntem	Deneme Sayısı (N)	Örnek Tweet	Anlam Değişim Değeri
1	Hashtag Silinmesi	1	Bu kitaplar Fâtih'tir, Selim'dir, Süleyman'dır; Şu mihrab Sinânüddin, şu minâre Sinân'dır; Haydi, artık uyuyan destanını uyandır! Delikanlım! işaret aldığın gün atandan! Yürüyeceksin! Millet yürüyecek arkandan! **~~#WeAreErdogan~~**	ANLAM DEĞİŞMİŞ
2	Hashtag Silinmesi	2	#Iran #Müslüman #ıslam #Türkiye #katil ısrail ortadoğuyu bulandırıyor yine göz yaşları **~~#emparyalis~~** amerika kahrolsun **~~#BM~~** israil savaşa dur 👆	ANLAM DEĞİŞMEMİŞ
3	Boşluk Ekleme	3	Cumhurbaşkanlığı imza toplama süreci başladı...Tabi ki **Y SK** ilk günden devrede!!! Şimdilik; 1) **Zey tinburnu** İlçe Seçim **Kurulund a** Sistem çalışmıyor!! 2) Beşiktaş İlçe Seçim Kurulunda sadece 1 bilgisayar işlem yapıyor!!! #ProtestoEdiyorum	ANLAM DEĞİŞMEMİŞ

Buna karşıt olarak Çizelge 7.15'deki yine N sayısıyla aynı yöntem uygulandığında modelin kararında herhangi bir değişiklik olmamıştır. Bunun sebebi bu örnekte karar vermesine yardımcı diğer güçlü ifadelerin bulunması hashtag'lere verdiği önemi düşürmesine sebep olabilir şeklinde yorumlanabilir.

Çizelge 7.14'deki Boşluk Ekleme yönteminin etkili olması gerçekten konuyla ilgili anahtar sözcüklerin yapısının bozulması veya farklı kelimeler olarak model tarafından anlaşıldığı şeklinde yorumlanabilir. Bununla birlikte aynı yöntem Çizelge 7.15'de aynı N değeriyle uygulandığında ise karar değişikliğine sebep olmamıştır. Modelin farklı kelimeleri ve heceleri farklı anlamlandırması şeklinde yorumlanabilir. Örneğin model "oğan" kelimesini "Erdoğan" la benzerliğini anlayamazken, "şener" kelimesini "Akşener" ile benzer olarak değerlendirmesi olasıdır. Karakter Değiştirme yöntemi incelendiğinde Çizelge 7.14'deki örnekte cümle başlangıcında yer alan peş peşe olan kelimelerdeki değişim modelin karar vermesini zorlaştırabilmektedir veya başka kelimelere benzetmesi de mümkündür. Çizelge 7.15'deki diğer örnekte ise yine peş peşe olan bir değişim görülmekle birlikte metnin tarafını belirlemede etkili olabilecek

bir kelime olan “HDP” kelimesinin kullanımı modelin hata yapmasını önlemiş olması seçenekler arasındadır.

Çizelge 7.14’deki Hashtag İşareti Ekleme yönteminde orijinal tweet model tarafından yanlış etiketlenmiştir. Fakat model yeni hashtag’ler eklendiğinde doğru karar vermiştir. Modelin kararını önceki kararını değiştiren şey ise yeni hashtag’lerin eklenmiş olmasıyla mevcut hashtag’lerin karar vermekteki ağırlığın kalkması ya da yeni eklenen hashtag’lerin örneğin “#Türk” kelimesi modelin eğitim kümesinde 19 kere gördüğü bir kelimedir. Bununla birlikte Çizelge 7.15’de yine aynı yöntemin aynı N değeriyle uygulanmasına rağmen modelin kararı değişmemiştir. Bunun sebebi metindeki asıl karar vermeyi etkileyebilecek olan “Recep Tayyip Erdoğan” kelimelerinde bir değişiklik olmaması ve “#Görüyorsunuz” ve “#bu” gibi özel bir anlam ifade etmeyen önemsiz kelimelerin hashtag yapılması modelin kararını değiştirmesi için yeterli olmamıştır.

Çizelge 7.14’deki Hashtag Silme yönteminde ise modelin kararını değiştirmiş olma sebebi hastag’ler silindiğinde metinde herhangi bir hashtag ifadesinin kalmaması ve insan bir değerlendirici tarafından anahtar sözcüklerden farklı bir etiket kararına verilmesine sebep verecek “soyguncu” kelimesinin, model tarafından anlaşılamaması sonucunda model tarafında farklı bir karar verilmiştir. Bununla birlikte Çizelge 7.15’de benzer örnekte ise silinen iki hashtag yerine benzer taraf değerine yönelik çok sayıda hashtag olması ve taraf tespitinde modelin kararını etkileyebilecek “Muharrem İnce” kelimelerinde herhangi bir değişme olmaması sebebiyle model kararını değiştirmemiş olabilir.

Çizelge 7.14’deki Boşluk Silme ile ilgili örnek incelendiğinde taraf tespitinde modelin kararını etkileyebilecek kelimelerden olan “Aziz Babuşçu” , “Selahattin Demritaş” kelime kalıplarının değiştirilmemiş modelin bunları artık başka birer kelime olarak anlamasına veya ek bir anlam yüklememesine sebep olurken, diğer kelimelerinde modelle ciddi bir bilgi vermemesi sebebiyle kararını değiştirmesine sebep olmuştur. Bununla birlikte Çizelge 7.15’deki örnekte boşlukların çıkarılması modelin kararını değiştirmesine sebep olmamıştır. Bunun sebebi “#yalnızdeğilsinSAADET” hashtag’inin aynen durması ve “Ankara'daSaadet” kelimesinin model tarafından ayrıştırılabilir yani hala “Saadet” kelimesinin anlaşılabilir olması olabilir.

Çizelge 7.14 : Türkçe için otomatik olarak yapılan ve etkili olan örnekler

Yöntem	**Deneme Sayısı (N)**	**Orijinal Tweet**	**Değişen Tweet**	**Orijinal Tahmin Etiketi**	**Değişiklik Sonrası Tahmin Etiketi**
Yeni Hashtag Ekleme	4	Saadetin chp ile bir olup Merhum lider Erbakan hocanın kemiklerini sızlattığına #İmzamıAtarım	Saadetin chp ile bir olup Merhum lider Erbakan hocanın kemiklerini sızlattığına #İmzamıAtarım **#pizza #gündem #okçuluk #ts**	AK Parti	SAADET
Boşluk Ekleme	3	Erdoğan Baskın seçim kararı sonrası ilk mitingini İ.Tatlıses'le birlikte İzmir'de yaptı.Anlaşılan Erdoğan çaresiz.!İzmir gibi bir ilde Mafya artığı Ve pkk yaltakçısı bir Sözde Sanatçıdan medet umuyorki;birliktenSahne alıyorlar.Şaşırmış bizim ki;Eskiden Diyarbakır'da yapardı	**Erd oğan** Baskın seçim kararı sonrası ilk mitingini İ.**Tatlıse s**'le birlikte İzmir'de yaptı.Anlaşılan Erdoğan çaresiz.!İzmir gibi bir ilde Mafya artığı Ve **pk k** yaltakçısı bir Sözde Sanatçıdan medet umuyorki;birliktenSahne alıyorlar.Şaşırmış bizim ki;Eskiden Diyarbakır'da yapardı	İYİ Parti	AK Parti
Karakter Değiştirme	4	Diploması olmayandan Devlet Başkan'ı olunursa ilk okul mezunundan doktor olup milletvekili adayı göstermiş olmaları normaldir öyle değil mi sayın Başkanım	**D!ploması o|mayandan Dev|et Bäşkan**'ı olunursa ilk okul mezunundan doktor olup milletvekili adayı göstermiş olmaları normaldir öyle değil mi sayın Başkanım	HDP	AK Parti
Hashtag İşareti Ekleme	2	Eğer Türk Milletini güldürmek istiyorsan onu tehdit et 😉 #akp #mhp #devletbahçeli #fetö #hedefbirmilyonimza	**#Eğer #Türk** Milletini güldürmek istiyorsan onu tehdit et 😉 #akp #mhp #devletbahçeli #fetö #hedefbirmilyonimza	AK Parti	İYİ Parti
Hashtag Silme	2	Dün Deniz Gezmiş'in asılması için evet oyu veren Chp, bugün Deniz Gezmiş denen soyguncuyu kullanıp edebiyatını yapıyor.. İkiyüzlülük diye işte buna deniyor.. #DenizGezmiş #6Mayıs1972	Dün Deniz Gezmiş'in asılması için evet oyu veren Chp, bugün Deniz Gezmiş denen soyguncuyu kullanıp edebiyatını yapıyor.. İkiyüzlülük diye işte buna deniyor.. ~~**#DenizGezmiş #6Mayıs1972**~~	AK Parti	CHP
Boşluk Silme	2	Aziz Babuşçu'nun "Selahattin Demirtaş içeriden çıkmalı" ifadesini duyan bir kısım vatandaşların nutku tutuldu!	**AzizBabuşçu**'nun "**SelahattinDemirtaş** içeriden çıkmalı" ifadesini duyan bir kısım vatandaşların nutku tutuldu!	SAADET	HDP
Harflerin Sıralarını Karıştırma	1	Dün Erbakan hoca dış güçler dediği zaman alaya alıp gülüyordunuz Bu gün ise bir yaprak zamansız dalından düşünce sebebi nedense dış güçler oluyor	Dün **Ebrakan** hoca dış güçler dediği zaman alaya alıp gülüyordunuz Bu gün ise bir yaprak zamansız dalından düşünce sebebi nedense dış güçler oluyor	SAADET	AK Parti

Çizelge 7.15 : Türkçe için otomatik olarak yapılan ve etkili olmayan örnekler

Yöntem	**Deneme Sayısı (N)**	**Orijinal Tweet**	**Değişen Tweet**	**Orijinal Tahmin Etiketi**	**Değişiklik Sonrası Tahmin Etiketi**
Yeni Hashtag Ekleme	4	Siz tamam deyince ses batıdan geliyor biz devam deyince ses ÜMMETTEN geliyor dünya lideri REİSLE DEVAM	Siz tamam deyince ses batıdan geliyor biz devam deyince ses ÜMMETTEN geliyor dünya lideri REİSLE DEVAM **#pizza #gündem #okçuluk #ts**	AK Parti	AK Parti
Boşluk Ekleme	3	Cumhurbaşkanı adayları lütfen diplomalarınızı noterde onaylamayınız emsal teşkil edeceğinden dolayı kötü örnek olabilirsiniz mesela MERAL AKŞENER lisans ve doktora ne gerek var böyle şeylere birde gitmiş TARIH PROF olmuş 🤣	Cumhurbaşkanı adayları lütfen diplomalarınızı noterde onaylamayınız emsal teşkil edeceğinden dolayı kötü örnek olabilirsiniz mesela **MER AL AK ŞENER** lisans ve doktora ne gerek var böyle şeylere birde gitmiş **TAR IH** PROF olmuş 🤣	İYİ Parti	İYİ Parti
Karakter Değiştirme	4	HDP Eş Başkanı Pervin Buldan seçim bildirgesini açıklıyor. HDP umuttur.	HDP Eş **Bäşkanı Perv!n Bu|dan seç!m** bildirgesini açıklıyor. HDP umuttur.	HDP	HDP
Hashtag İşareti Ekleme	2	Görüyorsunuz, bu ülkede bazılarının Erdoğan nefreti, vatanın bekasınından bile önemli. Bu ittifakda mesele vatan, Millet değil. Mesele Recep Tayyip Erdoğan.!! #DemekkiOluyormuş	**#Görüyorsunuz**, **#bu** ülkede bazılarının Erdoğan nefreti, vatanın bekasınından bile önemli. Bu ittifakda mesele vatan, Millet değil. Mesele Recep Tayyip Erdoğan.!! #DemekkiOluyormuş	AKP	AKP
Hashtag Silme	2	Türkiye'ye Güvence Muharrem İnce! TR #turkiyeyeguvencemuharremince #incedendemokrasigelecek #inceince #muharremince #TürkiyemKazanacak #BizeBirZaferGerek #BizHAZIRIZ #BizVARIZ #HazırMısınTürkiye #HaziranlarBizimdir #HerSeyGuzelOlacak #KorkmaBizVARIZ #KorkmaCHPvar	Türkiye'ye Güvence Muharrem İnce! TR #turkiyeyeguvencemuharremince #incedendemokrasigelecek #inceince #muharremince #TürkiyemKazanacak #BizeBirZaferGerek #BizHAZIRIZ #BizVARIZ #HazırMısınTürkiye #HaziranlarBizimdir #HerSeyGuzelOlacak ~~**#KorkmaBizVARIZ #KorkmaCHPvar**~~	CHP	CHP
Boşluk Silme	2	Ankara'da Saadet Partililere davadaşlarıma alçakça saldıran ülkücülükle alakaları kalmamış olan zavallılar Allah sizi ıslah etsin! Bizler sizin Saadetiniz için çalışmaya devam edeceğiz. İsteseniz de istemeseniz de bu düzeni #DEĞİŞTİR ECEĞİZ! #yalnızdeğilsinSAADET	**Ankara'daSaadet** Partililere davadaşlarıma alçakça saldıran ülkücülükle alakaları kalmamış olan **zavallılarAllah** sizi ıslah etsin! Bizler sizin Saadetiniz için çalışmaya devam edeceğiz. İsteseniz de istemeseniz de bu düzeni #DEĞİŞTİR ECEĞİZ! #yalnızdeğilsinSAADET	SAADET	SAADET
Harflerin Sırlarını Karıştırma	1	Tüm kanallar rte ve ince nin haberlerini mitinglerini canlı yayın yapsın. Ama bize gelince radyo frekansları falan. Gercekten çok özgürmüşüz.. sonra tabi oooo ince bunu demiş oooo rte bunu yapmış. Neyse ntv de falan değil ama sandıkta görüşürüz. #İyiOlucaz	Tüm kanallar rte ve **icne** nin haberlerini mitinglerini canlı yayın yapsın. Ama bize gelince radyo frekansları falan. Gercekten çok özgürmüşüz.. sonra tabi oooo ince bunu demiş oooo rte bunu yapmış. Neyse ntv de falan değil ama sandıkta görüşürüz. #İyiOlucaz	İYİ Parti	İYİ Parti

Harflerin sıralarının Çizelge 7.14'deki örnekte "Ebrakan" kelimesinin model tarafından yanlış yazılmış başka bir kelime olarak anlaşılmış olması ve diğer

kelimelerde de ciddi ayırıcı anlam ifade eden kelimelerin olmaması modelin kararını farklı şekilde yapmasında etkili olmuştur. Çizelge 7.15'deki diğer örnekte ise yapılan değişikliğin ciddi bir değişikliğe sebep olmaması, hala "rte", "ince" ve "#İyiOlucaz" ifadelerinin yer alması modelin karar değiştirmemesine sebep olduğu şeklinde yorumlanabilir.

8. SONUÇ

Bu tez çalışmasında, bireylerin sosyal medya platformlarını kullanırken gizliliklerini, yapay zekâ modellerinden nasıl koruyabilecekleri araştırılmıştır. Bu amaçla metinlerin değiştirilmesi için yeniden ifade etme ve yazım hatalarına yönelik 13 farklı yöntem ele alınmıştır. Türkçe ve İngilizce veri kümeleri üzerinde uygulanılan manuel veya otomatik deneylerden elde edilen sonuçlar incelenmiştir. Elde edilen sonuçlara göre gizliliğini korumak isteyen kullanıcılara verilecek öneriler şunlardır:

İlk olarak, İngilizce veri kümesinde oluşturulan deney düzeneğinde yeniden ifade etme yöntemlerinin modelleri aldatmada etkisiz olduğu bulunmuştur.

İkinci olarak, yazım hatalarına yönelik yöntemlerden olan görsel olarak benzer karakterleri birbirleri yerine kullanma, boşluk ekleyerek kelimeyi bölme ve harflerin sıralarını karıştırma hem Türkçe hem de İngilizce veri kümeleri için taraf tespit modellerinin performansını azaltmada en etkili yöntemlerdir. Ancak bu yöntemlerin uygulanması esnasında dikkat edilmesi gerek bir husus ise okunurluğun bozulmamasıdır. Bu yöntemlerin dikkatsiz uygulanması sonucunda metinin okunurluğunun bozulma olasılığı da bulunmaktadır.

Diğer bir önemli nokta ise hashtag'lerle ilgili hashtag silme, hashtag işareti ekleyerek yeni hashtag oluşturma, yeni hashtag ekleme yöntemlerde hashtag yapılacak veya eklenecek kelime ve kelime gruplarının seçiminde dikkatli olunması gerektiğidir. Hashtag silme orta ekili yöntemlerden olurken, hashtag eklemede eklenen hashtag kelimelerine ve hashtag işareti ekleyerek yeni hashtag olmak için seçilen kelimelere bağlı olarak metne istenmeden bir tarafa yönelik ifadeler eklenmiş olabilir hatta modellerden kaçmak yerine daha çok yakalanmaya sebep olabilmektedir. Bu sebeple hashtag'lerle ilgili değişiklikler bilinçli bir kullanım gerektirmektedir. Hashtag silme ve hashtag kullanılmaması çoğu durumda daha etkili sonuçlar vermiştir.

Sosyal medyada yapay zekâ modelleri tarafından takip edilememe ve kişisel hayatın gizliliğini koruma amacıyla önerilen bu yöntemler aynı zamanda, yanlış bilgi yaymak veya nefret söyleminde bulunmak veya toksik mesajlarının tespit edilememesi isteyen kişiler tarafından da yapay zekâ modellerini aldatmak için de kullanabilir. Fakat bununla birlikte, yapay zekâ modellerinin tehdit unsuru olabildikleri bir durumda,

çalışma kapsamında önerilen teknikler bu tehditlere karşı koruyucu zırhlar olarak kullanılabileceği ve tehditlerin bulunduğu bir ortamda bunlara karşı tedbirlerin alınması prensibi temel alınmıştır.

Çalışmamız birçok yönde geliştirilebilir olup, çok sayıda kullanıcının olduğu geniş veri kümeleriyle ırk, etnik köken ve ruh sağlığı gibi farklı görevlerde kişisel bilgilerin yapay zekâ modelleri tarafından tahmin edilememesine odaklanan, bu modelleri aldatmak için daha sofistike yöntemler geliştirildiği çalışmalar yapılabilir. Deneylerde platforma özgü oluşabilecek yanlılığın azaltılması için Twitter'ın yanı sıra farklı sosyal medya platformlarından da veri kümelerinin araştırılması düşünülmektedir. Nihai olarak farklı veri kümeleri için de geçerli olan etkili ve farklı tekniklerin yer aldığı otomatik araçların geliştirilmesi de araştırılabilecek konular arasındadır.

Geliştirilen yöntemlere yönelik örneğin, okunurluğu bozmadan karakter değiştirme yöntemini farklı dillere ait alfabelerdeki benzer görünüme sahip olan karakterlerin değiştirilmesi gibi mevcut yöntemlere yenilikçi yaklaşımların uygulanması sonucunda modelleri yanıltma başarısının incelenmesi araştırılabilir. Bu ve bunun gibi yazım hatası oluşturmaya yönelik yöntemlerin dayanıklılığının FastText tabanlı yazım hatalarına daha dayanıklı olan modellerin kullanması sonucunda, modelleri aldatma seviyelerine olan etkilerinin incelenmesi gelecekte yapılabilecek çalışmalar arasında yer almaktadır. Ayrıca, yöntemlerin ne kadar sayıda uygulanması sonucunda metin okunurluklarının bozulduğuna dair çeşitli deney düzenekleri tasarlanarak çeşitli kural ve yöntemler üretilebilir.

Bir diğer gelecek çalışma ise önerdiğimiz yöntemlerin farklı kombinasyonlarla uygulanması sonucunda etkili yöntemlerin bir arada kullanılmalarının veya etkili olmayan yöntemlerin bir arada kullanılması sonucunda bu kombinasyonların modelleri aldatmadaki etkinliklerinin araştırılmasına dair çalışmalardır.

Ayrıca taraf tespit sistemine yakalanmamak için önerdiğimiz yöntemlerin kullanıldığı durumlarda, insanların taraf eğilimlerini tespit etmek isteyen kişilerce yöntemlerimizin etkisizleştirilmesine yönelik bazı önlemler alınabilir. Örnek olarak, değiştirilmiş metinleri modellere tahminlemeleri için vermeden önce, değiştirilen metin üzerinde yazım hatası düzeltme yazılımlarının kullanılması verilebilir. Bu gibi durumlarda yöntemlerimizin etkinlik seviyelerinin nasıl değiştiğinin araştırılmasına yönelik çalışmalar yapılabilir.

Önemli/önemsiz kelimelerin tespitinde tweet'in bağlam bilgisinin de dâhil edildiği bir model oluşturulması durumunda elde edilecek performans incelemeleri gelecek çalışmalarda ele alınabilir.

KAYNAKLAR

[1] **Canbek, N.G.,Mutlu, M.E.**,(2016) On the track of artificial intelligence: Learning with intelligent personal assistants, *Journal of Human Sciences*, 13, 592-601.

[2] **Ahmad, N.A., Che, M.H., Zainal, A., Abd Rauf, M.F.,Adnan, Z.**,(2018) Review of chatbots design techniques, *International Journal of Computer Applications*, 181, 7-10.

[3] **Dogra, V., Verma, S., Chatterjee, P., Shafi, J., Choi, J.,Ijaz, M.F.**,(2022) A complete process of text classification system using state-of-the-art NLP models, *Computational Intelligence and Neuroscience*, 2022,

[4] **Kao, A.,Poteet, S.R.** *Natural language processing and text mining.* Springer Science & Business Media, 2007.

[5] **Zong, Z.,Hong, C.**,(2018) On application of natural language processing in machine translation, *2018 3rd International Conference on Mechanical, Control and Computer Engineering (ICMCCE)*,

[6] **Piotrowski, M.** *Natural language processing for historical texts.* Morgan & Claypool Publishers, 2012.

[7] **Rashidiani, S., Doyle, T.E., Samavi, R., Duncan, L., Pires, P.,Sassi, R.**,(2022) Textionnaire: An NLP-Based Questionnaire Analysis Method for Complex and Ambiguous Task Decision Support, *2022 IEEE 21st International Conference on Cognitive Informatics & Cognitive Computing (ICCI* CC)*,

[8] **Church, K.W.,Rau, L.F.**,(1995) Commercial applications of natural language processing, *Communications of the ACM*, 38, 71-79.

[9] **Nimbekar, R., Patil, Y., Prabhu, R.,Mulla, S.**,(2019) Automated resume evaluation system using NLP, *2019 International Conference on Advances in Computing, Communication and Control (ICAC3)*,

[10] **Morgan-Lopez, A.A., Kim, A.E., Chew, R.F.,Ruddle, P.**,(2017) Predicting age groups of Twitter users based on language and metadata features, *PloS one*, 12, e0183537.

[11] **Rahimi, A., Cohn, T.,Baldwin, T.**,(2018) Semi-supervised User Geolocation via Graph Convolutional Networks, July.

[12] **Preoţiuc-Pietro, D.,Ungar, L.**,(2018) User-level race and ethnicity predictors from twitter text, *Proceedings of the 27th international conference on computational linguistics*,

[13] **Küçük, D.,Can, F.**,(2020) Stance detection: A survey, *ACM Computing Surveys (CSUR)*, 53, 1-37.

[14] **Sekulic, I.,Strube, M.**,(2019) Adapting Deep Learning Methods for Mental Health Prediction on Social Media, November.

[15] **Rashed, A., Kutlu, M., Darwish, K., Elsayed, T.,Bayrak, C.**,(2021) Embeddings-based clustering for target specific stances: The case of a polarized turkey, *Proceedings of the International AAAI Conference on web and social media*, 15,

[16] **Baly, R., Mohtarami, M., Glass, J., Màrquez, L., Moschitti, A.,Nakov, P.**,(2018) Integrating Stance Detection and Fact Checking in a Unified Corpus, June.

[17] **Devlin, J., Chang, M.-W., Lee, K.,Toutanova, K.**,(2019) BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,

[18] **Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G.,Askell, A.**,(2020) Language models are few-shot learners, *Advances in neural information processing systems*, 33, 1877-1901.

[19] **OpenAI, R.**,(2023) GPT-4 technical report, *arXiv*, 2303.08774.

[20] **Ren, K., Zheng, T., Qin, Z.,Liu, X.**,(2020) Adversarial attacks and defenses in deep learning, *Engineering*, 6, 346-360.

[21] **Jia, R.,Liang, P.**,(2017) Adversarial Examples for Evaluating Reading Comprehension Systems, September.

[22] **Niu, T.,Bansal, M.**,(2018) Adversarial Over-Sensitivity and Over-Stability Strategies for Dialogue Models, October.

[23] **Belinkov, Y.,Bisk, Y.**,(2017) Synthetic and Natural Noise Both Break Neural Machine Translation,

[24] **Jin, D., Jin, Z., Zhou, J.T.,Szolovits, P.**,(2020) Is bert really robust? a strong baseline for natural language attack on text classification and entailment, *Proceedings of the AAAI conference on artificial intelligence*, 34,

[25] **Dai, J., Chen, C.,Li, Y.**,(2019) A backdoor attack against lstm-based text classification systems, *IEEE Access*, 7, 138872-138878.

[26] **Li, J., Ji, S., Du, T., Li, B.,Wang, T.**,(2018) Textbugger: Generating adversarial text against real-world applications, *arXiv preprint arXiv:1812.05271*,

[27] **Kurita, K., Michel, P.,Neubig, G.**,(2020) Weight Poisoning Attacks on Pretrained Models, July.

[28] **Liang, B., Li, H., Su, M., Bian, P., Li, X.,Shi, W.**,(2018) Deep text classification can be fooled, *Proceedings of the 27th International Joint Conference on Artificial Intelligence*,

[29] **Muller, B., Sagot, B.,Seddah, D.**,(2019) Enhancing BERT for lexical normalization, *The 5th workshop on noisy user-generated text (W-NUT)*,

[30] **Gu, T., Dolan-Gavitt, B.,Garg, S.**,(2017) Badnets: Identifying vulnerabilities in the machine learning model supply chain, *arXiv preprint arXiv:1708.06733*,

[31] **Yang, W., Li, L., Zhang, Z., Ren, X., Sun, X.,He, B.**,(2021) Be Careful about Poisoned Word Embeddings: Exploring the Vulnerability of the Embedding Layers in NLP Models, June.

[32] **Chen, X., Salem, A., Chen, D., Backes, M., Ma, S., Shen, Q., Wu, Z.,Zhang, Y.**,(2021) Badnl: Backdoor attacks against nlp models with semantic-preserving improvements, *Annual computer security applications conference*,

[33] **Sun, L., Hashimoto, K., Yin, W., Asai, A., Li, J., Yu, P.,Xiong, C.**,(2020) Adv-bert: Bert is not robust on misspellings! generating nature adversarial samples on bert, *arXiv preprint arXiv:2003.04985*,

[34] **Morris, J.X., Lifland, E., Lanchantin, J., Ji, Y.,Qi, Y.**,(2020) Reevaluating adversarial examples in natural language, *arXiv preprint arXiv:2004.14174*,

[35] **Schiller, B., Daxenberger, J.,Gurevych, I.**,(2021) Stance detection benchmark: How robust is your stance detection?, *KI-Künstliche Intelligenz*, 1-13.

[36] **Ebrahimi, J., Rao, A., Lowd, D.,Dou, D.**,(2017) HotFlip: White-Box Adversarial Examples for Text Classification, *Annual Meeting of the Association for Computational Linguistics*,

[37] **Barzilay, R.,McKeown, K.**,(2001) Extracting paraphrases from a parallel corpus, *Proceedings of the 39th annual meeting of the Association for Computational Linguistics*,

[38] **Mieskes, M.**,(2017) A quantitative study of data in the NLP community, *Proceedings of the first ACL workshop on ethics in natural language processing*,

[39] **Feyisetan, O., Ghanavati, S.,Thaine, P.**,(2020) Workshop on privacy in NLP (PrivateNLP 2020), *Proceedings of the 13th International Conference on Web Search and Data Mining*,

[40] **Silva, P., Gonçalves, C., Godinho, C., Antunes, N.,Curado, M.**,(2020) Using nlp and machine learning to detect data privacy violations, *IEEE INFOCOM 2020-IEEE conference on computer communications workshops (INFOCOM WKSHPS)*,

[41] **Kenton, J.D.M.-W.C.,Toutanova, L.K.**,(2019) Bert: Pre-training of deep bidirectional transformers for language understanding, *Proceedings of naacL-HLT*, 1,

[42] **Bojanowski, P., Grave, E., Joulin, A.,Mikolov, T.**,(2017) Enriching word vectors with subword information, *Transactions of the association for computational linguistics*, 5, 135-146.

[43] **Mikolov, T., Grave, E., Bojanowski, P., Puhrsch, C.,Joulin, A.**,(2017) Advances in pre-training distributed word representations, *arXiv preprint arXiv:1712.09405*,

[44] **Mohammad, S., Kiritchenko, S., Sobhani, P., Zhu, X.,Cherry, C.**,(2016) Semeval-2016 task 6: Detecting stance in tweets, *Proceedings of the 10th international workshop on semantic evaluation (SemEval-2016)*,

[45] **Bayrak, C.,Kutlu, M.**,(2022) Predicting Election Results via Social Media: A Case Study for 2018 Turkish Presidential Election, *IEEE Transactions on Computational Social Systems*,

[46] **Niven, T.,Kao, H.-Y.**,(2019) Probing Neural Network Comprehension of Natural Language Arguments, July.

Url-1<*https://huggingface.co/cardiffnlp/twitter-roberta-base*>, alındığıtarih:13.07.2023

Url-2<*https://github.com/stefan-it/turkish-bert*>, alındığıtarih:13.07.2023

Url-3<*https://www.nltk.org/*>, alındığıtarih:13.07.2023

Url-4<*https://stanfordnlp.github.io/CoreNLP/*>, alındığıtarih:13.07.2023

Url-5<*https://spacy.io/*>, alındığıtarih:13.07.2023