T.C.

YILDIZ TEKNİK ÜNİVERSİTESİ

FEN BİLİMLERİ ENSTİTÜSÜ

YAPAY ÖĞRENME ALGORİTMALARINI KANDIRMAK

Fatma GÜMÜŞ

DOKTORA TEZİ

Bilgisayar Mühendisliği Anabilim Dalı

Bilgisayar Mühendisliği Programı

Danışman

Doç. Dr. Mehmet Fatih AMASYALI

Temmuz, 2021

T.C.

YILDIZ TEKNİK ÜNİVERSİTESİ

FEN BİLİMLERİ ENSTİTÜSÜ

YAPAY ÖĞRENME ALGORİTMALARINI KANDIRMAK

Fatma GÜMÜŞ tarafından hazırlanan tez çalışması 01.07.2021 tarihinde aşağıdaki jüri tarafından Yıldız Teknik Üniversitesi Fen Bilimleri Enstitüsü Bilgisayar Mühendisliği Anabilim Dalı, Bilgisayar Mühendisliği Programı **DOKTORA TEZİ** olarak kabul edilmiştir.

Doç. Dr. Mehmet Fatih AMASYALI
Yıldız Teknik Üniversitesi
Danışman

Jüri Üyeleri

Doç. Dr. Mehmet Fatih AMASYALI, Danışman
Yıldız Teknik Üniversitesi _____

Doç. Dr. Sırma YAVUZ, Üye
Yıldız Teknik Üniversitesi _____

Doç. Dr. Tuğba DALYAN, Üye
İstanbul Bilgi Üniversitesi _____

Prof. Dr. Banu DİRİ, Üye
Yıldız Teknik Üniversitesi _____

Dr. Öğr. Üyesi Can EYÜPOĞLU, Üye
Milli Savunma Üniversitesi Hava Harp Okulu _____

Danışmanım Doç. Dr. Mehmet Fatih AMASYALI sorumluluğunda tarafımca hazırlanan Yapay Öğrenme Algoritmalarını Kandırmak başlıklı çalışmada veri toplama ve veri kullanımında gerekli yasal izinleri aldığımı, diğer kaynaklardan aldığım bilgileri ana metin ve referanslarda eksiksiz gösterdiğimi, araştırma verilerine ve sonuçlarına ilişkin çarpıtma ve/veya sahtecilik yapmadığımı, çalışmam süresince bilimsel araştırma ve etik ilkelerine uygun davrandığımı beyan ederim. Beyanımın aksinin ispatı halinde her türlü yasal sonucu kabul ederim.

Fatma GÜMÜŞ

İmza

Anneme

ve

babama

TEŞEKKÜR

Bu çalışma, Yıldız Teknik Üniversitesi Fen Bilimleri Enstitüsü Bilgisayar Mühendisliği Anabilim Dalında yapılan "Yapay Öğrenme Algoritmalarını Kandırmak" adlı doktora tez çalışmasını içermektedir. Metin ve konuşma sınıflandırıcıları, algoritma hakkında bilgilerin gizli olduğu kara kutu ortamda modelin girdi-karar ilişkisini keşfedecek akıllı ve uyarlanabilir bir "girdi bozma" algoritmasıyla saldırıya açık olduğu gösterilmiştir. Üretilen saldırı algoritmaları kara-kutu metin ve konuşma sınıflandırıcılarında açık oluşturabilecek noktalarını göstererek daha gürbüz ve güvenilir sistemlerin geliştirilmesinde katkı sağlaması üzere araştırmacıların kullanımına sunulmuştur.

Değerli hocam Doç. Dr. Mehmet Fatih AMASYALI'ya destekleyici danışmanlığı için minnettarım. Yardımına ihtiyaç duyduğumda hazır bulunarak bana araştırmamı mümkün olan en ideal koşullarda yürütmem için gereken özgürlüğü ve desteği verdi. Bir meslektaşı olarak bana değer verdi ve diğer araştırmacılarla da çalışarak profesyonel ağımı genişletmem için beni her zaman teşvik etti. Tez izleme komitesi üyesi hocalarım Doç. Dr. Sırma YAVUZ ve Doç. Dr. Tuğba DALYAN'a ayrıca teşekkürlerimi sunarım. Araştırma hayatımın bu ilk uzun soluklu serüveninde destekleyici rehber hocalarım olduğu için kendimi şanslı hissediyorum.

Tez araştırmalarımla ilgili beni dinlemek üzere kapıları bana hep açık olan başta Hava Harp Okulu Dekanlığı Bilgisayar Mühendisliği Bölüm Başkanı Dr. Öğr. Üyesi Ömer Özgür BOZKURT olmak üzere, değerli hocalarım Dr. Öğr. Üyesi Aytuğ BOYACI, Dr. Öğr. Üyesi Hüseyin Fehmi Selim BAYRAKLI ve tez savunma komitesi üyesi Dr. Öğr. Üyesi Can EYÜPOĞLU ve tüm çalışma arkadaşlarıma teşekkür ederim.

Bu uzun doktora eğitim maratonunda yanımda olan aileme; özellikle beni hayatım boyunca destekleyen annem Fisun GÜMÜŞ ve babam Hüseyin GÜMÜŞ'e, bir zamanlar arkamda ördek yavruları gibi dolaşırken artık yanımda birer yoldaş olan küçük kardeşlerim Mustafa, Betül ve Eyyüb Emre GÜMÜŞ'e teşekkür ederim. Bu sağlam sosyal destek sistemi olmadan araştırma sürecinde motivasyonumu korumak çok zor olabilirdi.

Fatma GÜMÜŞ

İÇİNDEKİLER

SİMGE LİSTESİ

α	bozulma çapı
θ	sınıf güven puanı
δ	takas sayısı

KISALTMA LİSTESİ

ABI+	Ability, Benevolence, Integrity and Predictability (Yetenek, İyilik, Bütünlük ve Tahmin Edilebilirlik)
BiLSTM	Bidirectional Long-Short Term Memory (Çift Yönlü Uzun-Kısa Süreli-Bellek)
CIA	Confidentiality, Integrity, and Availability (Gizlilik, Bütünlük ve Kullanılabilirlik)
CNN	Convolutional Neural Network (Evrişimsel Ağ Mimarisi)
CTC	Connectionist Temporal Classification (Bağlantıcı Zamansal Sınıflandırma)
DDİ	Doğal Dil İşleme
DTW	Dynamic Time Warping (Dinamik Zaman Bükümü)
FGSM	Fast Gradient Sign Method (Hızlı Gradyan İşaret Yöntemi)
G2P	Grapheme to Phoneme (Grapheme'den Phoneme Dönüştürme)
HCNN	Highly Confident Near Neighbor (Son Derece Güvenilir Yakın Komşu)
LSTM	Long-Short Term Memory (Uzun-Kısa Süreli-Bellek)
MFCC	Mel-frekans Cepstral Katsayıları
MLP	Multi Layer Perceptron (Çok Katmanlı Algılayıcı)
NB	Naïve Bayes
PoS	Part of Speech (Sözcük Görev Etiketleri)
RMS	Root Mean Square (Ortalama karekök)
SER	Speech Emotion Recognition (Konuşma Duygusu Tanıma)
YZ	Yapay Zeka

ŞEKİL LİSTESİ

TABLO LİSTESİ

ÖZET

Yapay Öğrenme Algoritmalarını Kandırmak

Fatma GÜMÜŞ

Bilgisayar Mühendisliği Anabilim Dalı

Doktora Tezi

Danışman: Doç. Dr. Mehmet Fatih AMASYALI

Yapay öğrenme algoritmaları, yüksek doğrulukta sınıflandırma ve tanıma sistemlerinin geliştirilmesine olanak sunarak modern yaşantıda vazgeçilmesi zor uygulamaların bir parçası olmuştur. Ancak, sistem geliştirme ve dağıtım sürecinde ortaya çıkan güvenlik açıkları hizmete veya ürüne olan güveni etkileyebilir. Dahası, sistem çıktısının sağlık gibi insan hayatı ve toplum yaşantısı üzerinde geri dönülemeyecek etkilere yol açabileceği uygulama alanlarında oluşabilecek zarar büyüktür. Yapay öğrenme odaklı hizmet ve ürünlerin başarılı bir şekilde yürütülmesini sağlamak ve zarar verecek sonuçlardan korunmak için "güvenilir yapay öğrenme" konusunun araştırılması son derece önemlidir.

Yapay öğrenme sistemlerini geliştirme aşamasında modeller sonuçların doğruluğu için optimize edilir. Yüksek doğrulukta sonuçlar elde etmek temel işlevsellik açısından model güvenilirliğini sağlarken dağıtım ortamlarında girdi üzerine yapılan müdahalelere karşı zafiyet oluşturabilir. Saldırgan, kurban modelin girdi-karar ilişkisini keşfedecek akıllı ve uyarlanabilir "girdi bozma" algoritmasıyla güvenilirlik ihlaline neden olur.

Bu tez çalışmasında kara-kutu koşullarında bir güvenilirlik çerçevesi belirlenmiş, metin ve konuşma sınıflandırma modelleri için güvenilirlik ihlaline yol açan kaçınma saldırıları geliştirilmiştir. Yüksek doğrulukla çıktı üretilen girdi örnekleri bozulma algoritmaları ile kara-kutu model ortamında zararlı örneklere dönüştürülmüştür.

Metin sınıflandırma modelleri için tasarlanan kutupluluk tabanlı küçük müdahalelerin, birbirinden çok farklı yapay öğrenme algoritmaları olan naïve Bayes ve BiLSTM modellerinin kandırılmasında oldukça etkili olduğu görülmüştür. Saldırının uygulanabilirliği gerçek bir kara-kutu olan IBM Watson doğal dil anlama servisi üzerinde doğrulanmıştır. Konuşma sınıflandırma sistemleri olarak öznitelik tabanlı çok katmanlı yapay sinir ağı ve ham sinyal tabanlı evrişimsel duygu tanıma modelleri için beyaz gürültü ve perde manipülasyonu ile zararlı örnekler üretilmiştir. Sinyal işleme yöntemleri ile ayrıntılı olarak incelenen sonuçlarda perde manipülasyonunun etkili bir saldırı yöntemi olduğu görülmüştür. Metin ve konuşma sınıflandırma saldırı deneylerinden elde edilen çıkarımlardan faydalanılarak iki-kipli sınıflandırma modeline düzenlenen bütünleşik saldırıların kolektif modelin gücünü kırabildiği gösterilmiştir. Son olarak, sistem kurulumundan önce uygun önlemleri tasarlayarak kötü niyetli aktörlere karşı sınama aşamasında sisteme entegre edilmek üzere öneriler sunulmuştur. Tez çıktılarının kara-kutu metin ve konuşma sınıflandırıcılarında açık oluşturabilecek noktalarını göstererek daha gürbüz ve güvenilir sistemlerin geliştirilmesinde literatüre katkı sağlaması dileğimizdir.

Anahtar Kelimeler: Güvenilir yapay öğrenme, çekişmeli ortamda öğrenme, metin sınıflandırma, konuşma sınıflandırma, kara-kutu çekişmeli örnek saldırıları

ABSTRACT

Deception of Machine Learning Algorithms

Fatma GÜMÜŞ

Department of Computer Engineering

Doctor of Philosophy Thesis

Advisor: Assoc. Prof. Dr. Mehmet Fatih AMASYALI

Allowing the development of high-accuracy classification and recognition systems, machine learning algorithms have now become an integral component of applications that are deemed indispensable in modern life. However, it is possible for the vulnerabilities exposed during system development and deployment to affect the trust placed in a service or product. It is also possible for the system output to cause a significant amount of damage in areas of human life and society, such as irreversible effects in healthcare. In order to ensure successful execution of machine learning-oriented services and products and to avoid harmful consequences, it is essential to investigate the issue of “trustworthy machine learning”.

During development of machine learning systems, the models are optimized for result accuracy. While high-fidelity results can provide model trustworthiness in terms of basic functionality, they can also render deployment environments vulnerable to input tampering. Attackers cause a trust violation with intelligent and adaptive "input tampering" algorithms that discover the victim model's underlying input-decision mapping.

In this thesis, we established a trustworthiness framework in black-box conditions, and developed evasion attacks that lead to reliability violations for text and speech classification models. Using distortion algorithms, we converted input samples that produced high-accuracy output into adversarial samples in a black-box model environment.

Polarity-based small interventions designed for text classification models were found to be very effective in deceiving the naïve Bayes and BiLSTM models, two machine learning algorithms with highly differing properties. The viability of attacks was validated on the IBM Watson natural language understanding service which is a true black-box. As for speech classification systems, adversarial samples were produced with white noise and pitch manipulation for feature-based multilayer artificial neural networks and raw waveform-based convolutional emotion recognition models. A detailed analysis of the results using signal processing methods revealed that pitch manipulation is an effective attack method across the models. Based on inferences from the text and speech classification attack experiments, we showed that it is possible for integrated attacks on the bimodal classification model to break the robustness of the ensemble model. Finally, we presented suggestions that can be implemented by designing appropriate measures before system deployment and integrated into the system during the phase of testing against malicious actors.

It is our wish that the thesis output will contribute to the literature in development of more robust and trustworthy systems by revealing the aspects that may create vulnerabilities in black-box text and speech classifiers.

Keywords: Trustworthy machine learning, learning in adversarial environment, text classification, speech classification, black-box adversarial sample attacks

1
GİRİŞ

Güvenilirlik, çok boyutlu ve sistem yaşam döngüsü boyunca devam eden dinamik bir kullanıcı-sistem ilişkisidir. Boyutlardan biri güvenin koşullu oluşudur. Koşulluluk, taahhüt edilen gibi sonuç verme kapasitesini ifade eder. Yapay öğrenme sistemi geliştiricileri, kullanıcılarına sistemin değişen koşullarda yüksek doğrulukta çalışacağını taahhüt eder. Ancak, yapay öğrenme tekniklerinin büyük bir kısmı, iyi huylu yürütme ortamları için tasarlanmıştır. Sistemin yüksek doğrulukta çalışmasının önüne geçmeyi amaçlayan düşmanların, ya da başka ifade ile rakiplerin, varlığı, modelin eğitildiği ve test edildiği dağılımlar arasında bir uyumsuzluktan yararlanarak "iyi huylu yürütme ortamı" ile ilgili varsayımlardan bazılarını geçersiz kılabilir. Yapay öğrenme uygulamaları karar yardımcı sistemler olarak güvenilmesi nedeniyle, saldırgan aktörler için giderek daha cazip bir hedef haline gelmiştir. Yapay öğrenme sisteminin açıklarını keşfetmek, proaktif olarak önlem alınması için önemlidir. Araştırmacı ve uygulama geliştiricileri, öğrenme algoritmalarının zayıf noktalarını keşfederek, daha gürbüz ve güvenilir çalışan sistemlerin geliştirilmesini hedeflemektedir. Yeni tehditlerin belirlenerek, sistem kurulumundan önce uygun önlemleri tasarlayarak kötü niyetli operatörlerin stratejisi öngörülmeye çalışılır. Algoritmaların gürbüzlüğünü artırmak üzere kontrollü şartlarda hazırlanmış saldırılar belirli güvenlik açıklarını ortaya çıkarabilir. Bu şekilde reaktif değil, proaktif bir önleme yaklaşımı izlenir. Böylece fiziksel ortamda yaşanabilecek kayıp ve aksaklıkların önüne geçilmiş olunur. İlerleyen alt bölümlerde bu tezin konusu olan sistem bütünlüğüne yönelik örnek saldırıları için literatür özeti, tezin amacı, orijinal katkı ve tezin bölümleri tanıtımına yer verilmiştir.

1.1 Literatür Özeti

Kötü amaçlı olarak geliştirilmiş akıllı ve uyarlanabilir girdi verisi (kötü amaçlı örnek, düşman örnek) ile düzenlenen nedensel ya da keşifsel saldırılar, yapay öğrenme

uygulamasının çalışmasını engelleyebilir ya da hatalı çalışmasına neden olabilir (Barreno ve diğ., 2006).

Sistemin eğitim aşamasında kullanılacak veri kümesine zararlı örneklerin eklenmesi *zehirleme saldırısı* olarak adlandırılır. Amaç sistem başarısını düşürerek kullanılmaz kılmak ya da bazı örneklere yanlılığın artırılması olabilir. Destek Vektör Makinelerine karşı yapılan saldırılar sınır fonksiyonu parametrelerini değiştirmeyi hedefler (Biggio, 2012; Laishram, 2016). Zehirleme saldırıları adaptif ve çevrimiçi öğrenme yapan dinamik ortamlar için özellikle tehdit oluşturur. Li ve diğ. (2016) işbirlikçi filtreleme kullanılan bir öneri sisteminde kötü amaçlı örneklerle zehirleme saldırıları gerçekleştirme senaryosunu çalışmıştır. Saldırganın amacı performans başarısını düşürmek, belli bir nesneyi daha fazla/az önermektir. Topluluk duyarlı sistemler de benzer şekilde çalışmaktadır. Topluluk içindeki bireyler sistem için sensör gibi yer alır, sağlık, kentsel ve çevre izleme, akıllı ulaşım gibi uygulama alanlarında kullanılacak girdi toplanır. Kötü niyetli operatör sistemin kullanılabilirliğini azaltarak ya da belli bir öğeye yanlılık oluşturarak sistemin doğru çalışmasını engeller (Miao ve diğ., 2018). Zehirleme saldırısına karşı savunma zararlı örnekleri tespit edip, bunları eğitim dışı bırakmak ya da düşük katsayılarla önemini azaltmak yoluyla gerçekleştirilir (Feng ve diğ., 2014; Liu ve diğ., 2016a; Steinhardt ve diğ., 2017).

Kaçınma saldırıları ise modele test sırasında gönderilen zararlı örnekler yoluyla gerçekleştirilir. Sistemde geçerli olabilecek örneklerin üzerine küçük gürültüler eklenmek yoluyla kötü amaçlı örnekler oluşturulur. Güncel çalışmaların büyük bölümünü bu tür saldırılar oluşturur. Beyaz kutu veya kara kutu ortamında deneyler yapılır. Beyaz kutuda ilgili yapay öğrenme modeli ile ilgili parametrelerin ve mimarinin bilindiği varsayılır ve bu özellikler kullanılarak saldırı gerçekleştirilir (Moosavi-Dezfooli, 2016; Papernot ve diğ., 2016a). Kara kutu saldırılarında eğitim kümesi, model parametreleri, olasılık dağılımları gibi bilgiler saldırganın elinde değildir. Sisteme gönderdiği girdilere karşılık alınan çıktıların değerlendirilmesi yoluyla zararlı örnekler üretilir. Üretilen örnekler mimarisi ve parametreleri bilinmeyen başka sistemlere de verilerek saldırı gerçekleştirilir (Papernot ve diğ., 2016b; Papernot ve diğ., 2016c; Kurakin ve diğ., 2016a; Liu ve diğ., 2016b). Kara

kutu ortamında yapılan saldırılar ile ilgili önemli bir husus *aktarılabilirlik* özelliğidir. Kötü amaçlı örneklerin aynı yapay öğrenme algoritmasını kullanan farklı mimariler arasında ortak kandırabilirliği *örnek aktarılabilirliği* (Papernot ve diğ., 2016b; Kurakin ve diğ., 2016a; Liu ve diğ., 2016b), farklı yapay öğrenme algoritmaları arasında ortak kandırılabilirlik ise *çapraz teknik aktarılabilirliği* olarak incelenmektedir (Papernot ve diğ., 2016c).

Yukarıda özetlenen araştırmaların tümü görüntü işleme uygulama alanı üzerine kurgulanmıştır ve yapay öğrenme algoritmalarını kandırma literatürünün en büyük kısmını oluşturur. Huang ve diğ. (2017) ve Lin ve diğ. (2017a) derin pekiştirmeli öğrenme alanında policy girdileri üzerinde yapılan saldırıları ele alır. Bu şekilde aktör hareketi engellenir ya da kısıtlanır, sonuçta optimum çözümden uzaklaşılması sağlanır. Jia ve Liang (2017) metinde soru cevaplama problemi üzerinde çalışmış, zararlı örnek üretme yöntemleri geliştirmiştir. Konuşma tanıma problemi üzerinde beyaz kutu ortamında araştırma yapan Carlini ve Wagner (2017), uzun konuşma bölütlerinin üzerine insan işitmesinin algılayamadığı gürültü ekleyerek sistemin hedeflenenden farklı bir konuşma tanınmasına neden olmuştur. Alzantot ve diğ. (2018a) kara kutu ortamında kısa ses komutlarının başka komutlara dönüştürülmesi üzerine çalışmıştır. Carlini ve diğ. (2016) ise kara-kutu ve beyaz-kutu saldırı ortamlarında deneyler gerçekleştirmiş, beyaz kutu deneylerinde fiziksel ortamda kısıtlı örneklerde saldırı aktarımı gerçekleştirmiştir.

Araştırmacıların deneysel olarak gerçekleştirdiği kandırma sistemleri, genel olarak sanal ortamda sistemlerden birinin saldırı üretmesi ve diğerinin çıktı üretmesi şeklinde gerçekleşir. Fiziksel ortama aktarılan saldırı düzeneği ile ilgili çalışmalar daha azdır. Kurakin ve diğ. (2016b) görüntü sınıflandırma sistemine saldırıyı fiziksel ortama aktarmak için, zararlı örneklerin kağıt çıktısından çekilen fotoğrafların sınıflandırma sistemine aktarmıştır. Çalışmada, fiziksel ortamda zararlı özelliğini kaybetmeyen örnek üretim algoritmaları tespit edilmiş, ışık, karşıtlık, bulanıklık gibi değişen kamera koşullarının etkisi incelenmiştir. Ses ve konuşma işleme problemlerinde fiziksel ortam çalışmaları, bir hoparlörden çalınan zararlı ses kayıtlarının hedef sistemde hatalı sonuç döndürmesine neden olması olarak tasarlanabilir. Bu konuda Carlini ve diğ.'nin (2016) kısıtlı gerçeklemesi

dışında tamamen başarıya ulaşmış bir çalışma, literatür araştırması kapsamında, bulunamamıştır.

Kötü amaçlı örneklerin oluşturulması ve bunlardan korunma yollarının keşfi konusu görüntü işleme uygulamaları kapsamında yoğun olarak çalışılmıştır. Koruyucu damıtma (defensive distillation) karar sınırlarına yakın bölgelere yapılan saldırılardan kaçınmak üzere iki model eğitir. Birinci model kesin etiketler ile eğitilir, çıktı katmanında elde edilen olasılıklar saklanır. İkinci model eğitilirken orijinal eğitim veri kümesinin yanında etiketler için birinci modelden elde edilen olasılık değerleri kullanılır (Carlini ve Wagner, 2017; Papernot ve diğ., 2016d; Papernot ve McDaniel, 2017). Rekabetçi öğrenme metotları savunmada başvurulan başka bir yöntemdir (Kurakin ve diğ., 2016a; Tramèr ve diğ., 2018). Eğitim sırasında orijinal veriye küçük gürültüler ekleyerek, sistemin ayırt edemeyeceği zararlı örnekler üretilir. Ardından orijinal örnekler ve üretilen zararlı örnekler ile model yeniden eğitilerek kandırılmaya daha dirençli model elde edilir. Kötü amaçlı örneklerin tespit edildikten sonra, gürültüsüz olarak yeniden inşası üretken ağları kullanan bir savunma yöntemidir (Song ve diğ., 2018; Lin ve diğ., 2017b).

Doğal dil işleme ve konuşma işleme alanlarında kandırılmaya karşı savunmanın yapay öğrenme yöntemleriyle yapılmasından ziyade, sensörlerde bloklama ve ilave kullanıcı girdisi gerektiren güvenlik katmanları kullanılmaktadır (Alzantot ve diğ., 2018). Tez çalışmasına başlarken, bu alanlarda yapay öğrenme ile saldırı ve savunma metotlarına dair ve yöntem geliştirme konularının açık olduğu görülmüştür. Son üç yılda, amacımıza paralel çalışmalar az da olsa yayınlanmıştır. Tezin ilerleyen bölümlerinde birbirinden çok farklı olan doğal dil işleme ve konuşma işleme sistemlerine dair güncel literatür, ilgili bölüm içinde verilmiştir.

1.2 Tezin Amacı

Birçok karar verme, planlama ve örüntü analiz süreci, çıktıları yaşamları ve gelirleri etkileyecek şekilde yazılı ve sesli doğal dil hizmetlerini içerir. Kötü niyetli girdilerin çok olası varlığında bile güvenilir ve kesin olmaları gerekir. Ancak bu tür hizmetler, özellikle sınıflandırıcılar, bozulma saldırılarına karşı savunmasızdır. Kara kutu ortamında bu tür saldırılara odaklanan literatürün çoğu, üretilen örnekleri aktarmak için bir aracı model (kahin) kullanır veya bilinmeyen kelime dağarcığı

önyargısını kötüye kullanır. Bu tez çalışmasında metin ve konuşma sınıflandırıcıları için, yapay öğrenme modellerini kara-kutu ortamda bir aracı model olmaksızın kandıran yöntemler geliştirilmesi amaçlanmıştır.

1.3 Orijinal Katkı

Bütünlük ihlaline yönelik örnek saldırılarıyla ilgilenen literatürün büyük kısmı görüntü işleme uygulamalarıyla ilgilenmektedir. Metin ve konuşma sınıflandırma sistemlerine yönelik test zamanında sistem bütünlüğüne yönelik saldırılar bu çalışmada incelenmiştir. Eğitilmiş bir modele tam veya sınırlı bilgi koşullarında, eğitim ve test dağılımları arasında en kötü durumdan faydalanmak için test verilerini değiştiren stratejiler tasarlanmıştır. Araştırma, bir sisteme çok zayıf erişimi olan ve kullanılan yapay öğrenme teknikleri hakkında çok az bilgisi olan bir saldırganın, bu tür sistemlere karşı güçlü saldırılar düzenleyebileceğini göstermiştir. Yapay öğrenme modellerinin zayıf genelleştirmesinin bu sistematik açıklaması, model eğitim veri dağılımından farklı dağılım seyreden örnekler üzerinde test yürütülürken tahminlere yönelik güven oranlarının istismara açık olduğunu göstermiştir. Birbirinden çok farklı veri türleri olan metin ve konuşma sinyali için kara-kutu ortamda sistem istismar noktalarını gösteren saldırılar tasarlanmıştır. Çalışmanın özet katkıları şöyle sıralanabilir:

- Bütünlük ihlaline yönelik örnek saldırı literatürünün büyük kısmı görüntü üzerine, metin ve konuşma sinyali için sınıflandırma sistemlerinde ise çoğunlukla beyaz-kutu saldırılar üzerine olmuştur. Bu çalışma hem incelenen yapay öğrenme metodu, hem de saldırı ortam koşulu açısından ilgili literatürdeki sınırlı sayıda çalışmalardan biridir.
- İkili ve çok sınıflı sınıflandırıcıların kutupluluk tabanlı kandırılma yöntemine, konuşma duygu sınıflandırıcıların ise perde manipülasyonu yönelimli olarak etkin bir şekilde kandırabildiği gösterilmiştir.
- İnsan katılımcılar ile yapılan değerlendirmede, saldırı örnekleri ile orijinal örnekler arasında metin sınıflandırma için %90, konuşma sınıflandırma için %91.66 sınıf farkı tespit edilemediği görülmüştür. Bu da zararlı örneklerin manuel olarak ayıklanmasının zor olduğunu göstermektedir.
- Bir aracı model olmadan kara-kutu saldırı yöntemleri gerçekleştirilmiştir.

- Metin Sınıflandırıcılar için;
    - Kelime ikameleriyle çekişmeli örnekler oluşturulmuştur.
    - İkili ve çok sınıflı modellere yönelik hedefli ve hedefsiz saldırılar, saldırı yönüne yönelik güveni önemli ölçüde artırmıştır.
    - İkili modellerde maksimum ortalama yanlış sınıflandırma güveni %53 artarken, çok sınıflı modellerde bu oran %45'tir.
    - Saldırının gerçek bir kara kutu modeli üzerinde uygulanabilirliğini gösterilmiştir: IBM Watson Natural Language Understanding servisi.
    - ELMo kelime gömmelerini kullanarak orijinal ve saldırı örnekleri arasında basit semantik benzerlik ölçümü ile saldırının otomatik anlamsal ölçüt ile tespit edilmesinin kolay olmadığı gösterilmiştir. Anlamsal kosinüs benzerliğinin %90 oranında korunduğu rapor edilmiştir.
- Konuşma sınıflandırıcılar için;
    - Konuşma duygu sınıflandırıcılarına saldırmak üzere genetik bir çekişmeli örnek hazırlama yöntemi geliştirilmiştir.
    - İki tür pertürbasyon analiz edilmiştir: ilave beyaz-gürültü ve perde pertürbasyonu. Özellikle mutluluk, öfke ve sakinlik duygularının tespitinden kaçınılması amaçlanmıştır.
    - Kara kutu ortamında iki kurbana saldırı deneyi yapılmıştır: önceden işlenmiş özelliklere sahip çok katmanlı bir yapay sinir ağı modeli ve örnek bir dalga biçimi evrişimli sinir ağı modeli.
    - Model girişinin kendisi küçük beyaz gürültüye karşı sağlam olduğundan, çok katmanlı algılayıcı kurbanı beyaz gürültü saldırılarına karşı daha gürbüz olduğu görülmüştür.
    - Evrişimli sinir ağı kurbanı hem beyaz gürültüye hem de perde bozulmalarına karşı savunmasız olduğu görülmüştür.
    - Akustik ve algısal özellikler açısından bozulmaların kapsamlı bir analizini sunulmuştur.

- Pertürbasyonların hesaplanması kolay ve beş iterasyondan daha az bir sürede, toplam %79.16'lık bir kandırma oranıyla yüksek bir kaçınma oranı elde edilmesiyle gerçek ortamda geçekleştirmesi son derece mümkün saldırı modelleri olduğu görülmüştür.

- İki-kipli sınıflandırıcılar için;
  - Metin ve ses kiplerini kullanan sınıflandırma mimarisi için pertürbasyon saldırısı düzenlenmiştir. Bilgimiz dahilinde, bu çalışma, metin ve ses kipleri üzerinde bütünleşik saldırı yöntemi sunan, literatürde ilk örnektir.
  - Bütünleşik saldırıda, segment değişimi ve akustik yönelimli olmak üzere iki farklı yöntem incelenmiştir. Akustik yönelimli yaklaşımın çalışma zamanı açısından saldırı için daha uygun olduğu rapor edilmiştir.
  - Akustik yönelimli bütünleşik saldırıda hedefli, ayrım gözetmeyen ve uyarlanabilir hedefli saldırı stratejileri incelenmiştir. Yapılan analizde, %88.28 saldırı başarısı ile uyarlanabilir hedefli stratejinin diğerlerine göre üstün olduğu görülmüştür.

Tezin ilerleyen bölümleri şöyle düzenlenmiştir: Bölüm 2'de yapay öğrenme için güven ve tehdit kavramları açıklanmış, Bölüm 3'te metin sınıflandırıcılarına düzenlenen kutupsallık tabanlı saldırılar verilmiş, Bölüm 4'te ise konuşma duygu sınıflandırıcılarının perde manipülasyonu ile kandırılmasına yer verilmiştir. Bölüm 5'te metin ve konuşma sinyali olarak iki-kipli çalışan bir sınıflandırıcı için yapılan çalışma sunulmuş, son bölümde ise değerlendirme ve savunma önerileri sıralanmıştır.

2
YAPAY ÖĞRENME İÇİN GÜVEN VE TEHDİT MODELİ

Güvenli yazılım geliştirme süreçleri saldırgan aktörler ve güvenlik sağlayıcılar arasında çekişmeli bir yarış halindedir. Saldırgan açıklardan faydalanmaya çalışırken, güvenlik sağlayıcı proaktif ve reaktif olarak açıkları engellemeye çalışır. Bilgisayar yazılımına dayalı tüm hizmet ve ürünlerde olduğu gibi, yapay öğrenme odaklı hizmet ve ürünlerde güvenlik konusu ele alınması zorunlu bir hususdur. Yazılım geliştirme yaşam döngüsünde köklü güvenlik zihniyeti hakimdir (Örneğin, Lipner, 2004; Mohammed ve diğ., 2017; Assal ve Chiasson, 2018). Yapay öğrenme odaklı yazılım geliştirmede ise daha yüksek doğrulukta model geliştirmeye yönelik ivme kazanan araştırma ve uygulama dünyası güvenilirlik konusunda henüz emekleme aşamasındadır. Güvenlik hususunun geliştirme sürecindeki yerinin tam belirlenmemiş olmasının önemli nedenlerinden biri, yapay öğrenme sistemlerinin geleneksel yazılım güvenliği perspektifinden önemli farklılıklarının olmasıdır. Bu bölümde güven ve tehdit, tezin konusu olan kara kutu yapay öğrenme modeli açısından incelenmiştir. Güvenirlik ve tehdit modeli için öncül tanımlardan sonra, üçüncü taraf servis kullanıcıları kaynaklı güvenlik riskleri açıklanmıştır. Amaç konunun tüm yönleriyle ayrıntılı bir çerçeve oluşturmak yerine, geliştirdiğimiz saldırılarla ilişkili yönlere dair öncül kavramları açıklamak ve ilgili literatürü sunmaktır.

2.1 Güvenilirlik

Yapay zeka (YZ) ve yapay öğrenmenin etik boyutlarına artan ilgiyle birlikte, mevcut ve gelecekteki uygulamaların güvenilirliğini sağlamanın yollarına odaklanıldı. Bu alandaki vurgu, hizmet ve ürünlerin kabulü ve başarılı bir şekilde benimsenmesini sağlamak için YZ'ya olan güveni korumanın kritik olabileceğinin kabulünü yansıtmaktadır. Güvenin nasıl oluşturulduğu, sürdürüldüğü veya aşındığı (eroded), bir bireyin veya grubun başkalarıyla etkileşimi, veriler, ortamlar, hizmetler, ürünler ve faktörler dahil olmak üzere bir bireyin güvenilirlik algısını veya başka bir şekilde şekillendirmek için bir araya gelen bir dizi faktöre bağlıdır. Güvenilirlik algıları,

yapay zekayı etkiler ve sonuç olarak, bir kişinin hizmet veya ürünle ilgili kararını ve davranışını etkiler. Teknoloji literatüründe güvene dayalı olarak, yapay öğrenme teknolojilerinin güveni nasıl artırabileceğini veya etkileyebileceğini belirlemek önemlidir. Dietz ve Den Hartog (2006), güvenilirlik yargılarının dayandığı dört temel özelliğin şunlar olduğunu öne süren ABI+ modelini geliştirdi: Yetenek; İyilik; Bütünlük ve; Tahmin edilebilirlik (Ability; Benevolence; Integrity and Predictability). Tahmin edilebilirliğin rolü göz önüne alındığında, güvenin süregelen ilişkiler yoluyla sürdürülmesinin önemine dikkat çekmiştir. Son kullanıcının, teknolojik ürün ve hizmetlerle olan ilişkisi "güven var/yok" şeklinde basit ikili sınıflandırma ile tanımlanamayacak kadar karmaşıktır. Hatta birçok durumda son kullanıcı "kararsız" bir güven sergiler; bu da insanların teknoloji ve yeniliklere "otomatik güven duyma" durumunun olmadığını gösterir.

Bir teknolojinin güvenilir olarak algılanma potansiyelini göstermek için üç tür koşul ileri sürülmüştür: beşerî, çevresel ve teknolojik nitelikler (Siau ve diğ., 2018). Beşerî ve çevresel etkiler bu tezin konusu dışında, sosyal bilimler ile incelenecek konulardır. Konumuz olan teknolojik nitelikler ise, teknolojinin kendisinin söz verildiği gibi sonucu verme kapasitesini ifade eder. Bu taahhüt çok boyutlu incelenebilir. İlk boyut, teknolojinin sonuçları kabul edilebilir bir doğruluk aralığında vermesi gerekliliğidir. Bu da performans ölçütü oluşturulması ve saha sonuçlarının söz verilen ölçü aralığında olmasını sağlaması ile olur. Bir diğer boyut, ürün/hizmet çözümleri için somut sınırlar tanımlanmasıdır. Teknolojinin kullanıcısına, teknolojinin amacını anlaması ve beklentilerini bu anlayışa dayalı olarak makul bir şekilde belirlemesi için yeterli bilgi sağlamalıdır. Bu kapsamda önemli olan bir boyut da "çıktı oluşum süreci"dir. Sonucun nasıl ede edildiği ve performansa ilişkin açıklamalar önemli bir güven faktörüdür.

Güven statik değil, dinamik bir unsurdur. Zaman içinde değişebilir (Li ve diğ., 2008). Yapay öğrenme sistemi geliştirilmesi, devreye alınması ve yürütülmesi aşamalarının herhangi birinde yaşanan gelişmeler ürüne veya hizmete olan güveni değiştirebilir. Romei ve diğ. (2014) vaka çalışmaları ile güveni çeşitli teknoloji aşamalarında incelemiştir. Önemli bulgulardan biri, güven etkisinin son aşamalarda daha belirgin

olduğudur. Yani, teknolojinin devreye alınması sürecinde gerçekleşen olaylar güveni daha fazla etkilemiştir.

Yapay öğrenme hizmeti/ürünü geliştirme süreci, genellikle, başlangıçtan sonuca sıralı bir hat izleyip sonlanan bir yapıya sahip değildir. Süreçler tekrarlanabilir, model güncellenebilir. Dahası, hizmet/ürün geliştirme yaşam döngüsünün tüm aşamalarında değişen güven, teknolojinin son kullanıcı tarafından kabul edilmesinden politika ve yönetmelik tasarımına kadar birçok kararı etkiler. Yapay öğrenme tabanlı bir hizmeti kullanıma sunmak bile, yeni önyargılar getirme veya mevcut olanları daha da kötüleştirme riskini taşıyabilir. Bir hizmetin sunuş şeklini değiştirmek bile, belirli gruplar için daha az yararlı veya etkili olduğu anlamına gelebilir. Güvenilir yapay öğrenme yönelimli çözümler üretmek için, geliştirme sürecinin ilk aşamalarından başlayarak güveni dikkate almanın önemi açıktır. Bu nedenle model dağıtım ortamına gitmeden önce başarıya dair ölçümler ve saldırılar için önlemler alınarak, devreye alma aşamasından sonra karşılaşılabilecek kötü sürprizler minimize edilebilir.

Son olarak, güvenirlik bağlamında kazalar, ya da güvenin ani çöküşlerine yol açabilecek öngörülmemiş başarısızlıkları ele alacağız. Bu tür güven hatalarının tipik örnekleri, veri kaybı/çalınması veya yapay öğrenme model sonuçlarında yanlılığın keşfedilmesidir. Yanlılık konusu özellikle son zamanlarda, modele olan güven için önemli bir unsur olmuştur. Bu konuda kapsamlı bir çalışma sunan Garrido-Muñoz ve diğ. (2021), doğal dil işleme literatüründe yanlılığı incelemiştir. İncelenen modellerde bir bireye, cinsiyet ve ırk gibi belirli bir gruba ait olma durumuna göre stereotipleme olduğu görülmüştür. Vektör uzayı yönelimli, ilişkilendirme tabanlı, veriye yönelik ve özel skorlama olmak üzere yanlılık giderme literatürünü incelemiştir. Buradaki ortak nokta, son kullanıcı güveninin yapay öğrenme modeline karşı adaletsiz, şeffaf ve hesap verilebilir olmadıklarının düşünülmesi ile algoritmaların topluma ve bireye zarar verebileceği endişesidir.

Özetle, yapay öğrenme modelleri ve algoritmaları, sonuçların doğruluğu ve sonuçları elde etme aşamasındaki verimlilik için optimize edilmiştir. Sonuçların doğruluk ve hız açısından temel işlevselliği gösterdiği düşünülebilse de bunlar her zaman diğer güven niteliklerini karşılamayabilir, ya da uyumlu olmayabilir. Bu

nedenle modeller doğruluk skoru ve hız dışında, algoritmaları kandırmak üzere yapılan çeşitli saldırılara karşı da gürbüzlükleri ile de değerlendirilmelidir. Modelin sonuçlarını saldırılara karşı değerlendirme mekanizması kurmak önemlidir. Yani algoritmaya saldırılara proaktif savunmalar üretmek gerekir. Bunun için, saldırı gerçekleşmeden saldırı noktaları araştırılmalıdır. Tez çalışmasında, algoritmalar "kara kutu" olarak kabul edilir. Böylece algoritmaya özel değil, modelden bağımsız bir saldırı tasarlamak amaçlanmıştır.

2.2 Tehdit Modeli

Yapay öğrenme yönelimli ürün ve hizmetler saldırıların hedefi olabilir. Saldırıya maruz kalan model "kurban" olarak adlandırılır. Saldırıyı yapan ise "saldırgan", "düşman", ya da "rakip" sözcükleriyle belirtilir. Saldırganın veri ve modele erişimi konusundaki çerçeve, saldırı amacı, sistem hakkında bilinenler ve sistem mimarisindeki zayıflıklar ile ilgilidir. Saldırgan, zarar vermek için hedefli bir saldırı gerçekleştirebilir veya ayrım gözetmeyen bir saldırı gerçekleştirebilir. Ayrıca saldırgan, bilgi toplama veya istihbarat amacıyla saldırıyı gizlice gerçekleştirebilir veya manipülasyon amacıyla sistemin işleyişine aktif olarak dahil olabilir.

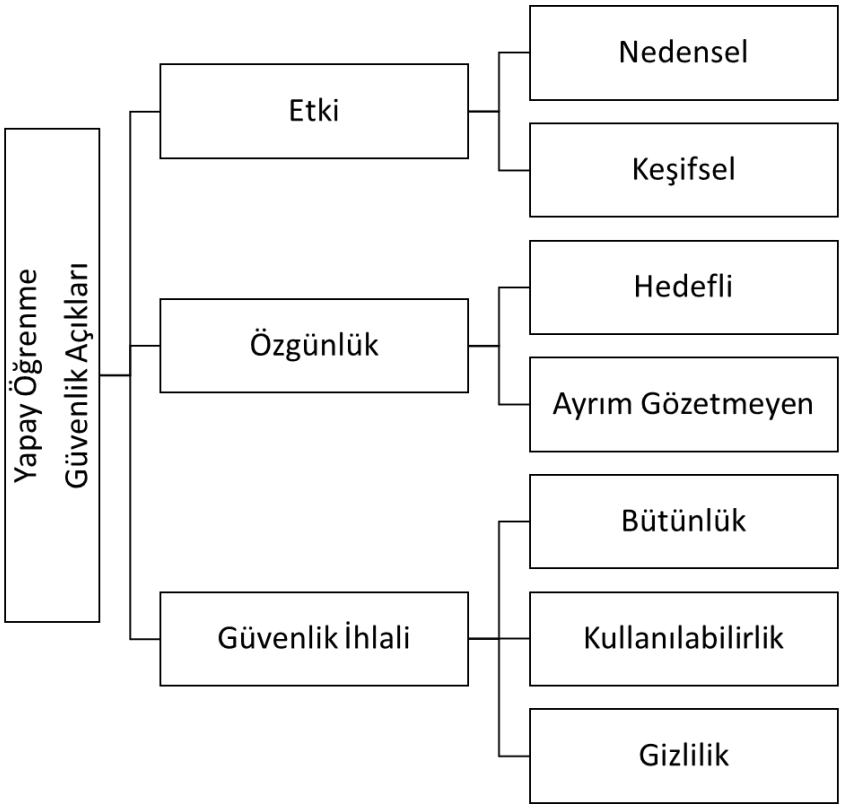
Bir yapay öğrenme modelinin güvenlik ve mahremiyet temelleri, klasik Gizlilik, Bütünlük ve Kullanılabilirlik güvenlik modelinden (CIA modeli) farklı değildir. Saklanan verilerin dikkatsizce hazırlanması, bir saldırgana bilgi sızdırabilir (Liu ve diğ., 2018a) ancak gizlilik, örneğin farklı (differentiable) mahremiyet (Abadi ve diğ., 2016) veya yapay öğrenmeye entegre kriptografi yaklaşımları (Shokri ve Shmatikov, 2015) gibi birçok yöntem ile geliştirilebilir. Bütünlük, ön işleme gibi bozulmayı önleyerek veya bozulmuş veriyi keşfederek ve onararak geliştirilebilir (Laskov ve Kloft, 2009). Ancak algoritmanın yürütülme ortamında da bütünlüğün dikkate alınması önemlidir (Liu ve diğ., 2018a).

Bu noktada kasıtlı ve kasıtsız hatalardan bahsetmek gerekir. Hedeflerine ulaşmak için sistemi bozmaya çalışan aktif bir düşmanın neden olduğu kasıtlı hatalar (Li ve diğ., 2018; Chakraborty v.e diğ., 2018), özel eğitim verilerini çıkarmak ya da temeldeki algoritmayı çalmayı amaçlayabilir. Bir yapay öğrenme sisteminin resmi olarak doğru ancak tamamen güvenli olmayan bir sonuç üretmesi ise kasıtsız hataları oluşturur (Ortega ve diğ., 2018; Amodei ve diğ., 2016).

Barreno ve diğ. (2006) çalışmasında yapay öğrenme algoritmalarının güvenlik açıkları araştırılmış, saldırı çeşitlerini gruplamış ve saldırılara karşı savunmaları sıralamıştır. Rekabetçi öğrenme konusunda yazılan makalelerde genel olarak buradaki sınıflandırmaya atıf yapılmaktadır. Çalışmada belirtilen taksonomiye göre bir saldırı aşağıdaki üç boyut ile tanımlanabilir: Etki (Influence), Özgünlük (Specificity), Güvenlik ihlali (Security violation). Güvenlik ihlali için CIA prensiplerinden gizliliği (confidentiality) de ekleyip Şekil 2.1'deki genişletilmiş taksonomiyi sunuyoruz.

Saldırının etkisi nedensel (causative) ya da keşifsel (exploratory) olabilir. Nedensel saldırı eğitim verisini etkilemek suretiyle eğitim sürecini değiştirir. Keşifsel saldırı ise, eğitim sürecini değiştirmez. Öğrenici üzerinden çevrimdışı analiz yaparak veri ya da algoritmaya ilişkin bilgi çıkarımı yapmayı amaçlar. Özgünlük, saldırı yapılması amaçlanan grup ile ilgilidir. Örneğin hedef bir sınıf dağılımına doğru bozulmaya uğratılması hedefli (targeted), herhangi bir yanlış negatife doğru bozulma da ayrım gözetmeyen (indiscriminative) bir saldırıdır. Son olarak, saldırının gerçekleştirdiği güvenlik ihlali bütünlüğü (integrity) ve/veya kullanılabilirliği etkileyebilir. Her iki kategori de yanlış sınıflandırmalar ile ilgilidir ancak kapsamları farklıdır. Saldırı örneklerinin yanlış negatifler olarak sınıflandırılması amaçlandığında bütünlük ihlali söz konusudur. Sistemin etkili bir şekilde kullanılamayacak hale gelmesi için yanlış pozitif ve yanlış negatif hatalarına yol açılması kullanılabilirlik için tehlike oluşturur.

Takip eden alt bölümlerde saldırı türleri sıralanacaktır. Tezin konusu olan pertürbasyon türü saldırılar ve diğer saldırı türleri olarak iki başlık halinde incelenmiştir.



Şekil 2. 1 Barreno ve diğ. (2006) taksonomisinden geliştirilmiş yapay öğrenme güvenlik açıkları kategorizasyonu

2.2.1 Pertürbasyon Türü Saldırılar

Pertürbasyon (Bozulma), bir nesnenin veya sistemin, normal durumundan veya yolundan, bir dış etkinin neden olduğu sapma olarak tanımlanabilir. Tez kapsamında, test kümesinden örneklerin sistemden istenen çıktıyı elde edilmek üzere bozulmaya uğratılmasını (Wei ve diğ., 2018) ifade etmek için kullanılmıştır. Örnek pertürbasyonları, erişim ihlali değil, bir girdi bütünlüğü ihlalidir. Modelin sınıflandırma performansını tehlikeye atar. Örnek saldırıları için bozulmaya uğratılan örnek "rakip örnek", "zararlı örnek" ya da "düşman örnek" olarak ifade edilebilir.

Ayrım gözetmeyen saldırıda, kurban modelde doğru sınıflandırılan bir örnek bozulur ve doğru olmayan herhangi bir sınıf sonucunun üretilmesi zorlanır. Hedefli, saldırıda çıktı olarak istenen belli bir sınıf söz konusudur. Hedefli düşman örneklerin üretilmesi, belli bir çıktı zorlandığı için, daha uzun sürebilir. Bununla beraber, her iki saldırı grubunda da zararlı örnek insan gözüyle fark edilmeyecek kadar az değişime uğramış olabilir, ya da rastgele gürültü olarak değerlendirecekleri

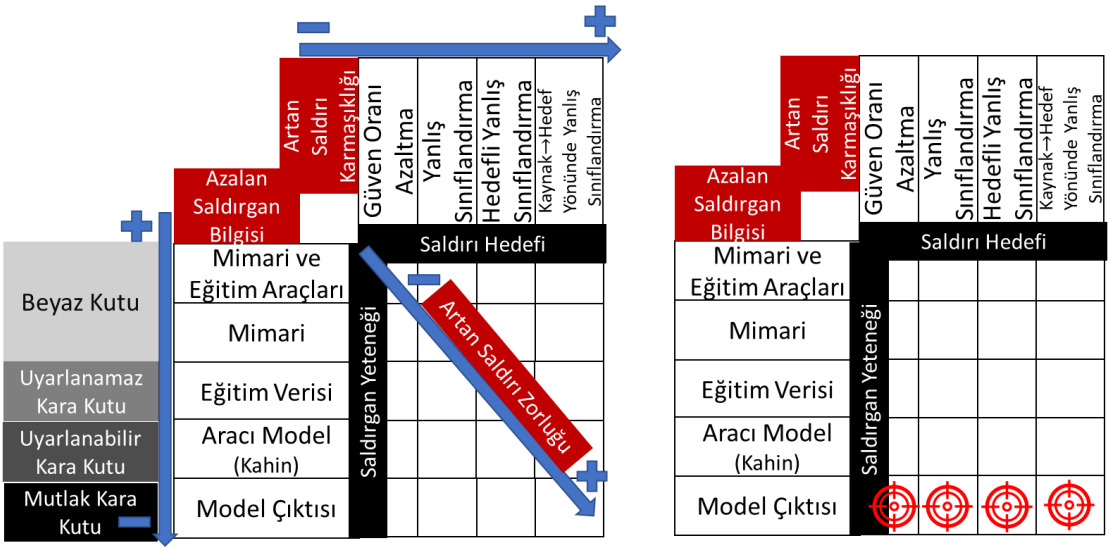
üzere az dikkat çekebilir. Modelin sınıflandırma kararlarında güven oranlarına ince bir müdahalede bulunularak, yani bariz olmayan pertürbasyon ile, güvenlik için oluşturulmuş bazı sistem uyarılarının tetiklenmesinin önüne geçilebilir.

Pertürbasyon saldırıları genel olarak birden fazla saldırı girişimini içerir. Düşman örnek, başarılı olana kadar iteratif bir şekilde kuvvetlendirilerek değişime uğratılır. Model, üretilen her zararlı örnek versiyonunda yoklanır (probing). Bir yapay öğrenme servisi için hizmet reddine (DoS) yol açabilecek saldırılara karşı sorgu miktarı ya da süresi ile ilgili önlem alınmış olabilir. Böyle durumlarda aracı bir yerel model kullanılarak, o model için elde edilen rakip örnek kurban modele gönderilebilir.

Örnek müdahaleleri, saldırının hedefi, yeteneği ve stratejisi olarak üç boyutta tanımlanabilir. Chakraborty ve diğ. (2018) beyaz kutu ve kara kutu ayrımına ek olarak, bu iki grup saldırı ortamı için alt kategoriler de tanımlamıştır. Bunu Papernot ve diğ.'nin (2016b) taksonomisiyle birleştirdik. Şekil 2.2'de hem bu birleşik taksonomi hem de çalıştığımız saldırı alanı gösterilmiştir.

Beyaz ve kara kutu ortam/model saldırılarını ayırt etmek önemlidir. Beyaz kutu saldırıda, mimari, eğitim parametreleri ve/veya eğitimle ilgili diğer hususlar hakkında bilgi kullanılır. Chakraborty ve diğ. (2018) eğitim verisinin bilinmesi durumunu ayrı tutmuş, "uyarlanamaz kara kutu" olarak adlandırmıştır. Aynı araştırmada "uyarlanabilir kara kutu" kavramı, örnek aktarılabilirliğinden yararlanılmasıdır. Bu tezde çalıştığımız ortam ise mutlak kata kutu, yani gerçek kara kutu saldırılardır. Bu grupta, saldırgan, yalnız modele verdiği girdiye aldığı cevap bilgisine sahiptir.

Saldırganın amacı yanlış negatif ve/veya yanlış pozitifler elde etmek olabileceği gibi, sınıf kararının etiketlenip etkilenmediğine bakılmaksızın gerçek etikete olan güvenin azaltılması olabilir. Şekil 2.2'de belirtildiği üzere, bu çalışmanın odak noktası mutlak kara kutu saldırgan yeteneği ile çeşitli saldırı karmaşıklık düzeylerini incelemektir. Bu alanı taksonomi "ez zor saldırı" kapsamı olarak göstermektedir.



Şekil 2. 2 Papernot ve diğ. (2016b) ve Chakraborty ve diğ.'ne (2018) göre saldırı taksonomisi (sol) ve tez çalışmasının araştırma alanı (sağ)

2.2.2 Diğer Saldırı Türleri

Bu bölümde, modele beyaz kutu olarak saldırmak üzere bazı bilgilerin çalınması amacıyla gerçekleştirilen saldırı türler ve zehirleme saldırılarının kuşbakışı haritasının çizilmesi amaçlanmıştır.

Üyelik Çıkarım (Membership Inference) Saldırısı. Kurban modeli eğitmek için bir veri örneğinin kullanılıp kullanılmadığının belirlenmesini amaçlar. Gizlilik güvenlik ihlalini oluşturur. Homer ve diğ.'nin (2008) saldırı yöntemi, bir genomik veri kümesinde belirli bir genomun varlığını tespit etmiştir. Yapay öğrenme modellerine karşı bir üyelik çıkarım saldırısı geliştiren Shokri ve diğ. (2019), bir örneğin sınıflandırıcıyı eğitmek için kullanılıp kullanılmadığını, yalnızca model çıktısına bağlı olarak anlaşılabileceğini göstermiştir.

Model Çalma (Model Stealing) Saldırısı. Saldırganlar, modeli sürekli sorgulayarak model parametreleri ya da mimarisi hakkında çıkarım yapar. Bu saldırı sonucunda elde edilen yeni (çalıntı) modelin işlevselliği, kurban modelinkiyle aynıdır. Wang ve Gong (2018) hiperparametreler, model parametreleri ve eğitim veri seti arasındaki ilişkileri, objektif fonksiyonunun gradyanı sıfıra eşitleyerek türetilen bir doğrusal

denklem sistemine kodlamak yoluyla Amazon modelleri için hiperparametrelerin başarıyla çalınabildiğini göstermiştir.

Derin Ağı Yeniden Programlama (Reprogramming deep neural nets) Saldırısı. Özel olarak hazırlanmış bir sorgu aracılığıyla, kurban model, gerçek amacından sapan bir görevi gerçekleştirmek üzere yeniden programlanabilir. Elsayed ve diğ. (2018), saldırgan tarafından seçilen bir görevi gerçekleştirmesini sağlamak için, test örneği üzerinde tek bir pertürbasyonunun yeterli olduğunu göstermiştir.

Zehirleme (Poisoning) Saldırısı. Saldırgan, sınıflandırma doğruluğunu azaltmak için eğitim örnekleri oluşturur ve modele enjekte eder. Kurban modelin eğitimi sırasında sınıflandırma karar fonksiyonunu tamamen bozarak, hedefli ya da ayrım gözetmeyen yanlış sınıflandırmalara yol açar. Chen ve diğ. (2017) hedefli gürültü enjeksiyonu ile zehirleme yoluyla gri-kutu (beyaz ve mutlak kara kutu arasındaki konfigürasyonlar) ve beyaz kutu ortamda tespitten kaçınma amacı ile (yanlış negatif) graf kümeleme ve graf yerleştirme yöntemlerine dayalı küçük topluluk saldırıları tasarlamıştır.

3
METİN SINIFLANDIRICILARINA KARŞI KUTUPLULUK TABANLI KARA KUTU SALDIRISI

Sohbet robotları ve kişisel sanal asistanlar gibi uygulamalar modern hayatı içine çekerken, IBM ve Amazon gibi şirketler, uygulama geliştiricilerin kullanımı için doğal dil işleme (DDİ) hizmetlerini sağlamaktadır. Bu hizmetler, geliştiricilerin duyarlılık analizi, metin sınıflandırması, toksik dil algılama ve içerik önerisi gibi alt görevleri gerçekleştirmek için DDİ sistemini sorgulamasına olanak tanımaktadır. Yapay öğrenme kullanan DDİ hizmetleri, eğitimde iyi performans göstermelerini sağlarken eğitim/test aşamasında doğruluklarını en üst düzeye çıkarmayı amaçlar. Bununla birlikte, doğal dil olgusu, bir yapay öğrenme modelinde tamamen temsil edilemeyecek kadar karmaşıktır. Eğitim verisi büyük olasılıkla ek karmaşıklık içerecektir; başka bir deyişle, iyimser ortamda bile bir miktar gürültü olması beklenir. Model yeterince genellenirse eğitim ve test performansı birbirine yakın olacaktır. Öte yandan, performansı hizmetin bütünlüğünü tehlikeye atacak bir düzeye indirmek için eğitim girdisi kötü niyetle manipüle edilebilir.

Ne klasik yapay öğrenme modelleri ne de derin öğrenme modelleri, sistemi kötü niyetle kullanan aracılara (insan veya makine) karşı bağışık değildir. Bu saldırganlar, sistemin bütünlüğünü veya kullanılabilirliğini tehlikeye atmak için mevcut bilgileri kullanarak saldırılar gerçekleştirir.

Önerdiğimiz tehdit modeli, düşman bilgisi ve saldırı hedefine dayalı saldırı taksonomisinde (Şekil 2.2) en zor saldırı düzeyi olan mutlak kara-kutu koşullarını için geliştirilmiştir. Yöntemimiz, saldırganın, sorgulara verdiği çıktı dışında model hakkında hiçbir bilginin olmadığı tam kara kutu ortamında sınıflandırma güvenini azaltmayı amaçlar. Hedefli ve ayrım gözetmeyen saldırılar ayrı ayrı incelenmiştir.

Bu bölümde sunduğumuz yöntem ile literatüre katkılarımız aşağıdaki gibidir:

• Bir kahin veya bir üretici model eğitmeden kara kutu sözcük pertürbasyonları gerçekleştirilmiştir. Kara kutu hakkında hiçbir bilgi olmaksızın saldırı gerçekleştirilebilir.

• Önerilen saldırı hem temel, hem de nöral yapay örenme modellerinde yüksek saldırı başarısı ile çalışır. Bu özelliği ile, genelleştirilmiş bir saldırı yöntemi olduğunu göstermiştir.

• Deneyler, nöral modelin bazı stratejiler altında daha savunmasız olduğunu göstermiştir.

• Semantik benzerliğe dayanan basit otomatik ölçümle saldırımızı tespit etmek zordur.

• IBM Watson Doğal Dil Anlama hizmetine yönelik bir kara kutu saldırısı değerlendirilmiştir.

3.1 Metin Sınıflandırma Algoritmaları

Metin verisinin elde edilmesi kolaydır ve zengin bir bilgi kaynağı olarak yapay öğrenme algoritmalarında kullanılabilir. Yapılandırılmamış doğası ise doğal dil işleme problemlerini çözmek üzere kural tabanlı çıkarım yapmayı zor ve zaman alıcı hale getirebilir. Yapay öğrenme yönelimli yaklaşım sayesinde metin verisini çeşitli doğal dil problemlerini çözmek üzere işlemek çok daha kolay hale gelmiştir.

Metin sınıflandırıcılarına karşı geliştirdiğimiz yöntem için kullanılan iki sınıflandırıcıda "kara kutu kabulü" yapılmıştır. Yani model bilgisi bulunmasına rağmen saldırıda kullanılmamıştır. Diğer sınıflandırıcı ise gerçek bir "mutlak kara kutu" olan IBM Watson doğal dil anlama servisidir.

Naive Bayes (NB) ailesindeki algoritmalar, olasılık ve istatistik teorilerinden faydalanarak örnekleri sınıflandırır. Bu çalışmadaki deneylerde çok terimli (multinominal) naive Bayes kullanılmıştır (Kibirya ve diğ., 2004). Uzun-Kısa Süreli-Bellek (LSTM, Long-Short Term Memory) sınıflandırıcısı, metni bir zaman serisi olarak ele almak suretiyle sınıflandırma yapan bir tekrarlayan yapay sinir ağıdır (Hochreiter & Schmidhuber, 1997). Çift Yönlü Uzun-Kısa Süreli-Bellek (BiLSTM, Bidirectional Long-Short Term Memory) sınıflandırıcısı ise metinde uzun veya kısa vade bağımlılıkları çift yönlü olarak modeller (Schuster & Paliwal, 1997). Alt bölümlerde bu iki sınıflandırıcı algoritmaları açıklanmıştır.

3.2 Çok Terimli Naive Bayes

Naive Bayes sınıflandırıcılar, her öznitelik arasında koşullu bağımsızlık (Eşitlik 3.2) varsayımıyla Bayes teoremini (Eşitlik 3.1) kullanır. Sınıf etiketi y ve öznitelik vektörü $(x_1, \dots, x_m)$ için naive Bayes sınıflandırıcıları için genel koşullu olasılık Eşitlik 3.1 ve 3.2'de verilmiştir. $P(x_1, \dots, x_m)$ metin girdisi için sabit alınır, böylece sonsal olasılık $\hat{y}$, Eşitlik 3.3 ile hesaplanır.

$$P(y|x_1, \dots, x_m) = \frac{P(y)P(x_1,\dots,x_m|y)}{P(x_1,\dots,x_m)} \tag{3.1}$$

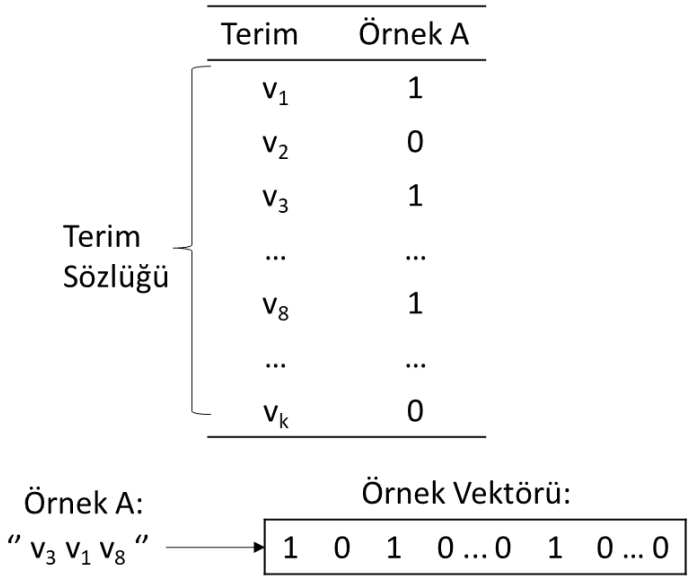
$$P(y|x_1, \dots, x_m) = \frac{P(y)\prod_{i=1}^{m} P(x_i|y)}{P(x_1, \dots, x_m)} \tag{3.2}$$

$$\hat{y} = arg\max_{y} P(y)\prod_{i=1}^{m} P(x_i|y) \tag{3.3}$$

Naive bayes sınıflandırıcı ailesindeki algoritmalar $P(x_i|y)$ değeri için farklı yaklaşımlar kullanabilir. Çok terimli naive Bayes algoritmasında girdi temel olarak sayma tabanlı sözcük kodlama ile temsil edilir. Metin içerik sayısallaştırılırken eğitim kümesinden terim sözlüğü oluşturulur ve her bir terim için tam sayı temsilleri belirlenir. Örnekler sayısallaştırılırken önce terimlere ayrılır, sonra sözlükteki sayısal karşılıkları ile örnek vektörü oluşturulur (Şekil 3.1). Sözcük frekansları ile sayısallaştırmanın bazı avantaj ve dezavantajları Tablo 3.1'de sunulmuştur.

Tablo 3. 1 Sözcük frekansları ile sayısallaştırmanın bazı avantaj ve dezavantajları

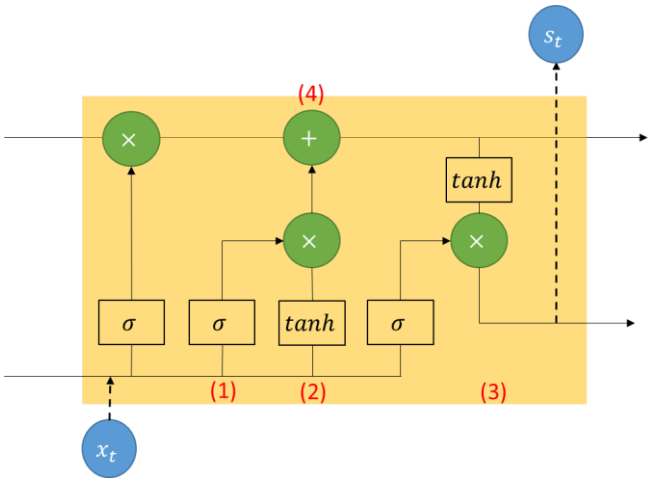
Avantaj	Dezavantaj
Bir örnekteki en açıklayıcı terimleri çıkarmak için temel bir ölçüt sunar.	Veri kümesi için seyrek (sparse) matris oluşur. Sıfırları düzeltmek üzere ek prosedürlerin kullanımı gerekebilir.
İki örnek arasındaki benzerlik kolayca hesaplanabilir.	Dağıtık kodlamaya göre anlambilim açısından zayıftır.
n>1 olmak üzere n-gramlar ile kullanıldığında, terim sırası ile ilgili bilgi de kodlanabilir.	Büyük n-gramlar için kullanışlı değildir.



Şekil 3. 1 Sayma tabanlı örnek sayısallaştırma

3.3 BiLSTM Sınıflandırıcı

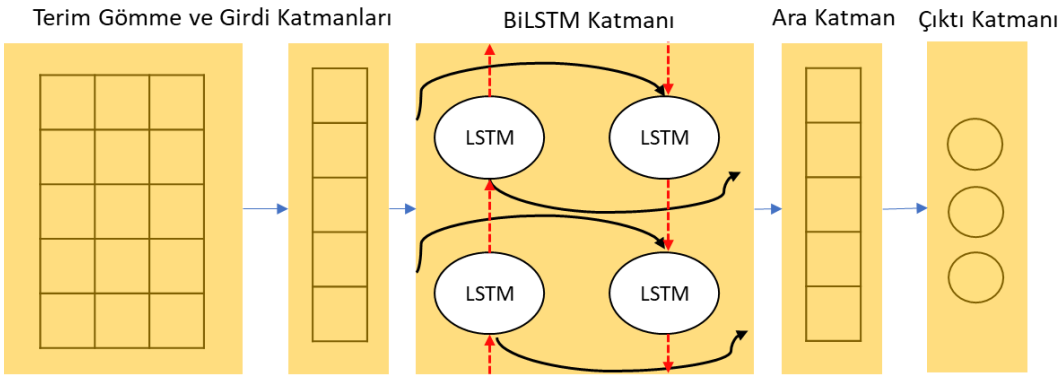
LSTM hücresi, tekrarlaya sinir ağlarında sıklıkla görülen kaybolan gradyan problemini gidermek üzere geliştirilmiş bir çözümdür. Dört kapılı bir yapı sunar. Bu kapılar (Şekil 3.2) bilgiyi yönetir: (1) hangi bilginin unutulacağı, (2) hangi bilginin saklanacağı, (3) saklanan bilginin ne zaman kullanılacağı, (4) önceki durumdan gelen bilginin sonraki duruma aktarılıp aktarılmayacağı.



Şekil 3. 2 LSTM hücresi

BiLSTM sinir ağı, odak terimin öncesindeki ve sonrasındaki bağlamı birleştirmek üzere zıt yönde çalışan iki katman LSTM'den oluşur (Şekil 3.3). BiLSTM, yinelenen

bağlam bilgilerini hafızaya almadan uzun vadeli bağımlılıkları kodlayabilir. Bu nedenle, metin sınıflandırma için yaygın olarak kullanılmaktadır. BiLSTM ağı girdi örneklerini dağıtık kodlama yöntemi olan terim gömmeleri (embedding) olarak alır. Bunlar ön eğitimli olabilir, ya da veri kümesinde eğitim gerçekleştirilirken eş zamanlı olarak güncellenme yoluyla öğrenilebilir. Gömme katmanı, sözlükteki her terim için dağıtık bir temsil öğrenir. Benzer bağlamda birlikte kullanılması bilgisi olan terimler arasında gizli ilişkileri kodlar.



Şekil 3. 3 Temel bir BiLSTM mimarisi

3.4 Metin Sınıflandırıcılarına Saldırılar

DDİ alanındaki beyaz kutu saldırılarının çoğu, kelime vektör uzayından yararlanır. Papernot ve diğ., (2016e) Jacobian matris ile kelime yerleştirmelerini hesaplayarak Hızlı Gradyan İşaret Yöntemi (FGSM, Fast Gradient Sign Method) kullanarak LSTM ağlarına saldırmıştır. Çalışmada "önemi" sözcükler Jacobian ile belirlenmiş ve pertürbasyon gerçekleştirilmiştir. Daha sonra, kelime gömme alanında bu bozulmaya yakın bir kelime ile değiştirilmiştir. Benzer şekilde, Samanta ve Mehta (2017), önemli kelimeleri belirlemek için gradyanları kullanmıştır. Takas, silme veya ekleme işlemi ile pertürbasyon gerçekleştirmiştir. Kullandıkları algoritma, orijinal örneğin anlamsal özelliklerini ve dil bilgisi yapısını korumak için tasarlanmıştır. Gradyanları kullanan diğer çalışmalar da önemli kelimeleri tespit eder ve yanlış sınıflandırmaya yönelir (Sato, 2018; Miyato, 2016; Gong, 2018; Ebrahimi, 2018). Benzer şekilde, hepsi mimari bilgisine ve türevlenebilir ağırlıklara ihtiyaç duyar. Beyaz kutu saldırıları, aktarılabilirliği kullanarak kara kutu modellerine saldırmak için aracı olarak kullanılabilir (Szegedy ve diğ., 2013). Bu

durumda başarı, aracı model ile saldırganın içeriden hiçbir bilgiye sahip olmadığı kara kutu modeliyle uyumluluğuna bağlı olacaktır. Evrensel bir çekişmeli örnek seti oluşturulmadıkça, uygun bir aracı tasarlamak çok zor olurdu.

Aktarılabilirlik kullanımı dışında, kara kutu ortamı için geliştirilen bir diğer yöntem karakter pertürbasyonlarıdır. Bu grup yöntemler, insan gözüyle kolay algılanamayacak karakter bozulmaları ile kelimelerin DDİ modellerinde tanınmamasına neden olur (Gao ve diğ., 2018; Li ve diğ., 2018). Hem beyaz kutu ortamında (Li ve diğ.,2018) hem de kara kutu ortamında (Gao ve diğ.,2018; Li ve diğ., 2018) saldırganlar da önce "önemli" terimleri tespit işlemi yapmıştır. Daha sonra, sözcüğü modelde "bilinmeyen" olarak kodlanmasını sağlamak için iki karakterin yer değiştirmesi gibi küçük karakter işlemleri gerçekleştirirler. Yanlış yazılan sözcüklerin otomatik düzelticiler veya yazım denetleyicileri tarafından düzeltilerek savunma sağlanabileceği raporlamıştır.

Geliştirdiğimiz yöntem, sözcük bozulmalarının söz konusu olduğu kara kutu modellerine yönelik hedefli ve hedefsiz saldırılara odaklanır. Kelime gömme vektörlerini kullanan saldırı yöntemlerinden farklı olarak, kutupluluk tabanlı algoritmamız yalnızca kara kutu hizmetinden döndürülen sınıf etiketi ve güven puanı bilgilerini kullanır. Saldırı, kurban modeli sorgulamaya dayalı olduğu için çalışmamız Vijayaraghavan ve Roy (2019), Alzantot ve diğ. (2018a) bazı yönlerden benzemektedir. Vijayaraghavan ve Roy (2019), hem karakter bozulmalarına hem de kelime yerleştirmelerine dayanan bir kod çözücü-kodlayıcı (decoder-encoder) modeli eğitmiştir. Alzantot ve diğ. (2018a), bir genetik algoritma aracılığıyla hedef sınıfın güvenini en üst düzeye çıkarmayı amaçlayan terim takaslarını kullanmıştır. Yöntemimizde yalnız bu çalışmalarla rekabet eden sonuçlar elde etmekle kalmamış, aynı zamanda farklı sınıflandırma yöntemlerindeki çalışmalarımızı doğrulamak için çeşitli veri kümeleri üzerinde deneyler yapılmıştır. Ayrıca, gerçek durum senaryolarında uygulanabilirliğini daha fazla doğrulamak için IBM'in Watson NLU hizmeti üzerinde gerçek bir kara kutu deneyi yürütülmüştür.

Kutupluluk tabanlı düşmanca saldırımıza ek olarak, iki temel saldırı yöntemi sunuyoruz: rastgele saldırı ve “sıcak terimler” saldırısı. Her ikisi de seçilen kelimeleri bilinmeyen bir kelime ile değiştirerek bilinmeyen kelime saldırılarının

arkasındaki fikri taklit eder (Gao ve diğ.,2018; Li ve diğ.,2019 çalışmalarındaki gibi). Rastgele saldırı, terimleri rastgele seçerken, sıcak terimler saldırısı evrensel bir $kaynak\ sınıf$a ait örnekler için frekansı yüksek terimler listesini kullanır.

3.5 Saldırı Modeli Mimarisi

Pertürbasyon stratejisi, rakiplerin amacı ve yetenekleri ile sınırlıdır. Amacımız, model hakkında mümkün olduğunca az bilgi ile kara kutu sistemlerine saldırı gerçekleştirmektir. Saldırı modelinde, test örneği harici bir kaynaktan toplanır ve saldırgan tarafından manuel etiketlenir. Ardından, test kümesi kullanılarak istenen saldırı yönlerindeki ($kaynak\ sınıf \rightarrow hedef\ sınıf$) örnekler ve ilgili etiketler kullanılarak bir kutupluluk tablosu oluşturulur. Kaynak ve hedef sınıflar "kutuplar" olarak kabul edilmek kaydıyla ve terim sözlüğündeki her sözcük için bu sınıflar arasındaki kutupluluk hesaplanır. Son olarak, çıktı etiketi ve güven puanları elde etmek için kara kutu modelini sorgulanır. Düşman stratejisi tarafından belirlenen koşullar karşılanana kadar bu adımlar tekrarlanarak saldırı örneği oluşturulur. Aşağıdaki alt bölümlerde, tehdit modeli ve pertürbasyon metodolojisi detaylı olarak tanımlanmıştır

3.6 Tehdit Modeli

Tehdit oluşturan düşmanın üç boyutu vardır: amaç, bilgi ve yetenek, strateji.

Amaç. Sistemi kandırma görevine karşı tutumunu belirtir. Düşman nihayetinde sistemin bütünlüğünü tehlikeye atarak daha yüksek yanlış sınıflandırma oranını hedefler. Bu amaca ulaşmak için, ikincil bir kaynaktan toplanan manuel etiketlenmiş örneklerle kara kutu modelinin hedeflenen ve ayrım gözetmeyen etki ile yanlış sınıflandırma olasılıklarını iteratif olarak artırır.

Bilgi ve Yetenek. İki varsayım ele alınmıştır; birincisi asgari bilgiyi, ikincisi beceriklilik derecesini belirtir.

• **Varsayım 3.1.** Düşman, görevi tanır; görevin ne tür bir problem olduğu (örneğin, duygu analizi) ve sınıf kümesini (örneğin, pozitif/negatif veya spor/sağlık/eğitim) bilir.

• **Varsayım 3.2.** Düşman, görevle ilgili neredeyse sınırsız sayıda etiketlenmemiş örneği harici bir kaynaktan (örneğin, web taraması) maliyetsiz olarak toplayabilir.

İki varsayım göz önüne alındığında, düşmanın iki bilgi durumu vardır: mükemmel ve sınırlı bilgi. Bir girdi örneği, verilen her sınıf için olasılık puanları döndürürse, düşman mükemmel bilgiye sahiptir. Çıktı, yalnızca sınıf etiketiyse, saldırgan sınırlı bilgiye sahiptir. Düşman, yeteneğin iki yönü ile de sınırlandırılmıştır. Birincisi, Varsayım 2'ye göre toplanan örnekleri etiketlemek için sonsuz ya da sınırlı bütçe olması durumudur. İkinci kısıt, "sorgu bütçesi"dir. Saldırganın kurban modele yaptığı sorguların örnek başına sınırını ifade eder.

Strateji. Düşman, sistemin tamamen gözlemlenebilir veya yarı gözlemlenebilir olabileceği kara kutu ortamda keşif saldırıları gerçekleştirir. Tamamen gözlemlenebilir bir ortamda, düşmanın bir aracı modele bile ihtiyacı yoktur. Yeterli miktarda sorgu bütçesi verildiğinde saldırıları yapılandırabilir. Yarı gözlemlenebilir ortamda ise, kötü niyetli örnekler oluşturmak ve örnek aktarılabilirlik özelliğinden yararlanmak için bir aracı modelden yararlanır. Her iki durumda da, saldırı şemasındaki tek değişen özellik, saldırıyı oluştururken geri bildirim aldığımız modeldir.

"Önemli terimler" belirleme fikrinden yola çıkarak, herhangi bir sınıflandırıcı kullanmadan terimlere saldırı yönüne göre logaritmik önem puanı yani kutupluluk puanı atanmıştır. Saldırı yönü, bir kaynak ve hedef sınıftan oluşur. Saldırı yönü değiştikçe önemli terimler değişir. Saldırı, bir sınıfın bir örneğini diğerine dönüştürürken, hedef sınıf üzerinde yüksek pozitif ölçekte yer alan terimlerin, hedef üzerinde düşük negatif ölçekte yer alan terimlerle değiştirilmesi yoluyla gerçekleştirilir. Ancak, bu tür değişiklikler orijinal örneklerde yüksek bozulmaya neden olacaktır. Bu nedenle, bu tür saldırılara karşı geliştirilebilecek savunma yöntemlerinden kaçınmak için hem puan değişimi hem de takas sayısını sınırlandırılmıştır. Ayrıca, bozulmayı daha da azaltmak için aday takas terimleri sözcük görev etiketlerine (PoS, Part of Speech) göre filtrelenmiştir.

3.7 Kutupluluk Tabanlı Pertürbasyon

Tehdit modeli, saldırının gerçekleştiği ortamın ayrıntılı bağlamının özetlenmesine olanak tanır. Rakip, Tablo 3.2'de gösterildiği gibi mükemmel veya sınırlı bilgi ve yeteneğe sahip olabilir. Özellikle aşağıdaki senaryolar dikkate alınmıştır:

- **Mükemmel ortam:** Düşman mükemmel bilgiye ve sınırsız bütçeye sahiptir.

- **Tamamen gözlemlenebilir kısıtlı ortam:** Düşmanın, dışarıdan toplanan örnekler için sınırlı etiketleme bütçesi vardır. Model, her sınıf için güven (olasılık puanları) sağlar. Düşmanın modele sorgu gönderme süresiyle ilgili herhangi bir kısıtlama yoktur.

- **Yarı gözlemlenebilir kısıtlı ortam:** Düşmanın sınırlı bilgisi vardır; model, sınıf başına güven skorlarını geri döndürmez. Toplanan örnekler için sınırsız etiket bütçesi vardır. Düşmanın modele sorgu gönderme süresiyle ilgili herhangi bir kısıtlama yoktur.

- **Gerçek kara kutu ortamı:** Kara kutu sistemi, bir kutupluluk tablosunu doldurmak ve/veya çıktıyı doğrudan gözlemleyerek çekişmeli örnekler oluşturmak için aşırı miktarda sorgulanamaz. Modelden sağlanan bir güven puanı yoktur. Bu kriter modele bağlı olmadığı için etiketleme bütçesi geçerli olabilir veya olmayabilir.

Tablo 3. 2 Tehdit modeline ilişkin senaryo durumları

<table>
<tr><td>Kaynaklar</td><td colspan="8">Evet/Hayır</td></tr>
<tr><td>Sınırsız Etiket Bütçesi</td><td>E</td><td>H</td><td>E</td><td>H</td><td>E</td><td>H</td><td>E</td><td>H</td></tr>
<tr><td>Sınırsız Sorgulama Bütçesi</td><td>E</td><td>E</td><td>H</td><td>H</td><td>E</td><td>E</td><td>H</td><td>H</td></tr>
<tr><td>Güven Değerlerinin Sağlanması</td><td>E</td><td>E</td><td>E</td><td>E</td><td>H</td><td>H</td><td>H</td><td>H</td></tr>
<tr><td colspan="9">Ortam Özellikleri</td></tr>
<tr><td>Yetenek</td><td colspan="2">Tamamen Gözlemlenebilir</td><td colspan="7">Yarı-Gözlemlenebilir</td></tr>
<tr><td>Bilgi</td><td>Mükemmel</td><td colspan="8">Sınırlı</td></tr>
<tr><td>Deney Ortamı</td><td>Mükemmel</td><td colspan="6">Sınırlı</td><td>Gerçek kara-kutu</td></tr>
</table>

Önce, mükemmel şartlar altında saldırı yöntemi genelleştirilmiştir. Sorgu bütçesi ve varsayılan olarak güven puanlarının kullanılabilirliği gibi olası sınırlamaları göz ardı edilmiştir. Kısıtlamalar daha sonraki alt bölümlerde tanıtılmıştır.

3.8 Mükemmel Ortam

Saldırganın pertürbasyon stratejisi sözcük değiş tokuşuna dayanır. Kelime takaslarını koordine etmek için, Varsayım 1 ve 2 tarafından toplanan örnekleri kullanarak bir kutupluluk tablosu doldurulur (Algoritma için bkz. Şekil 3.4). Çok yüksek düzeyde, kutupluluğa dayalı düşman stratejisi şu şekilde sıralanabilir:

1. Toplanan örneklerden elde edilen terimler için kutupluluk puanlarının ve PoS etiketlerinin bulunduğu bir kutupluluk tablosu doldurulur.
2. Herhangi bir örnek için, kutupluluk tablosuna bakılarak, aynı PoS etiketine sahip tablo satırlarında bir bozulma aralığı α içinde sınırlı sayıda terim takası yapılır.
3. Hazırlanan bu yeni saldırı örneğinin bir hedef sınıf için istenen minimum olasılığı verip vermediğini kontrol edilir Olasılık puanı iyileşmezse değişiklikler geri alınır.
4. Maksimum sayıda takas işlemi yapılmışsa, bozulma durdurulur.

Düşmanın amacı kara kutu sistemine saldırmak olduğundan, model mimarisine veya eğitim seti dahil herhangi bir parametreye sahip değildir. Düşman, Varsayım 1 ve 2 kapsamında derlediği test setindeki bilgileri kullanır. Kutupluluk tablosunun oluşturulması esas olarak saymaya dayanır (Eşitlik 3.4 ve 3.5).

$$c(w_t) = \sum_{k=1}^{|X|} \sum_{i=1}^{N_k} [w_{ki} = w_t] \tag{3.4}$$

$$c(w_t|L) = \sum_{x_k \in L} \sum_{i=1}^{N_k} [w_{ki} = w_t] \tag{3.5}$$

Eşitlik 3.4, tüm veri kümesinde bulunan bir w_t teriminin sayısını gösterir. Eşitlik 3.5'te ise yalnızca L sınıfı için sayma yapılır. Harici olarak derlenmiş test kümesi $[X, Y]$ olmak üzere, X örnek kümesi ve Y karşılık gelen etiket kümesidir. $(x_k, y_k) \in$

$[X, Y]$ için $x_k = \{x_{k1}, x_{k2}, \dots, x_{ki,\dots}\}$ örnek için terim dizisini ve y_k ilgili sınıf etiketini gösterir. Eşitlik 3.6'da N_k, x_k'deki toplam terim sayısını gösterirken, x_{ki}, x_k'deki i-inci terimdir. $c(w_t)$, test kümesinde geçen w_t teriminin sayısıdır ve $c(w_t|L)$ ifadesi, L∈Y olmak üzere, terimin L olarak etiketlenen sınıfta bulunması sayısıdır.

Ayrım gözetmeyen saldırılar için tüm benzersiz bir etiket tanımlanmıştır: "Rest". Hedeflenmemiş saldırılar da 3.5'e dahil ederek Eşitlik 3.6 elde edilir.

$$c(w_t|L) = \begin{cases} \sum_{k=1}^{|X|} \sum_{i=1}^{N_k} [w_{ki} = w_t] & , L \in Y \\ c(w_t) - \sum_{k=1}^{|X|} \sum_{i=1}^{N_k} [w_{ki} = w_t] & , L = "Rest" \end{cases} \tag{3.6}$$

Saldırmayı hedeflediğimiz etiket grubu için "hedef sınıf" terimi kullanılmıştır. Spesifik olarak, saldırılar için kullandığımız gösterim $y \rightarrow y^*$ şeklindedir, burada y kaynak ve y* hedef sınıf etiketidir. y∈Y, y≠y* ve y*∈(Y∪'Rest') olmak üzere, y kaynak sınıfından, y* hedef sınıfına bir saldırı tanımlar. $s_{w_t}(y \rightarrow y^*)$, w_t teriminin y sınıfından hedef sınıf y*'ye (Eşitlik 3.7) taşındığında kutupluluk puanıdır. $s_{w_t}(y \rightarrow y^*) = (-1)\, s_{w_t}(y^* \rightarrow y)$ olacaktır.

$$s_{w_t}(y \rightarrow y^*) = \begin{cases} tanımsız & , \; c(w_t|y^*) = 0 \\ \log \dfrac{1}{c(w_t|y^*) + 1} & , \; c(w_t|y) = 0 \\ \log \dfrac{c(w_t|y)}{c(w_t|y^*) + 1} & , \; diğer\ durumlar \end{cases} \tag{3.7}$$

Düşman örnek oluşturmak üzere hesaplanan kutupluluk tablosu, tüm olası y→y* çiftleri için terimleri, PoS etiketlerini ve ilgili kutupluluk puanlarını içerir. Takas için seçilen terimlerinin PoS etiketlerinin aynı olması gerekir. Saldırı örnekleri oluşturulurken önce PoS etiketlerini filtrelenir. Her bir takas adayı w_c'nin kutupluluk puanı α bozulma çapı olmak üzere $s_{w_t}(y \rightarrow y^*) - \alpha < s_{w_c}(y \rightarrow y^*) < s_{w_t}(y \rightarrow y^*) + \alpha$ olacak şekilde yeniden filtrelenir. Orijinal örnek terimleri ,(w_t'ler) rastgele sırada değerlendirmeye alınır. Orijinal örnek $x = \{w_1, w_2, \dots, w_t, \dots\}$ ve bozulmuş örnek $x^* = \{w_1, w_2, \dots, w_c, \dots\}$ için kaynak sınıftan hedef sınıfa doğru bir

kaymanın olup olmadığı ve varsa kaymanın yönü ve miktarı, kara kutu çıktısına göre anlaşılır. Hedef sınıf için belirlenen hedef olasılığı θ'ya ulaşana ya da çıktı etiket hedef sınıfına değişene kadar δ adet takas gerçekleştirilir. Şekil 3.4-3.6'da tüm işlemler algoritma ve blok şemalar olarak gösterilmiştir.

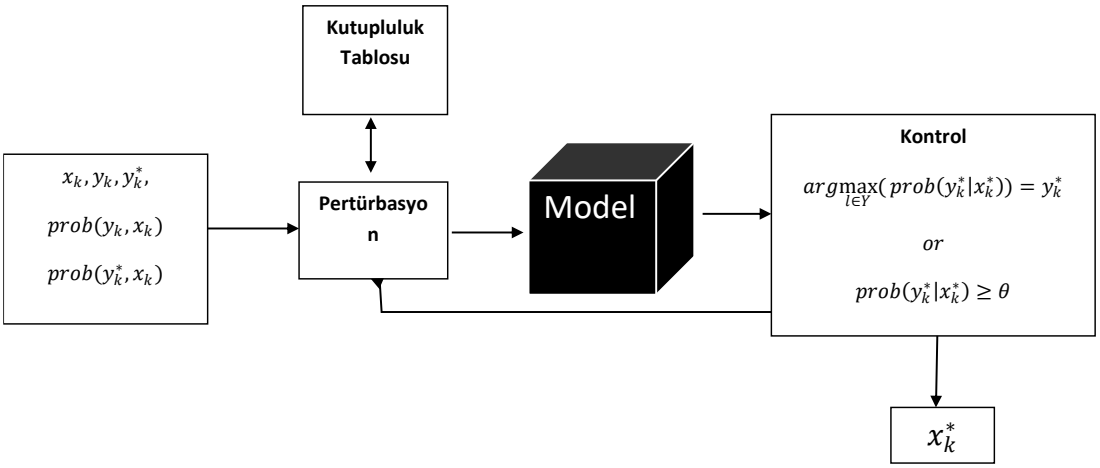
Algorithm 1 Kutupluluk Tabanlı Pertürbasyon

x: orijinal örnek
y: orijinal (kaynak) etiket
y^*: hedef etiket (ayrım gözetmeyen saldırılar için hedef 'Rest' olarak değiştirilir)
PT: $y \rightarrow y^*$ için $terim$, PoS, $skor$ alanlarından oluşan kutupluluk tablosu
α: maksimum bozulma çapı
δ: maksimum takas sayısı
θ: istenen hedef güven puanı
Conf: örnek x girdisi için bir sınıf etiketine göre güven puanını döndüren fonksiyon
tokenize_and_tag: örnek için (terim,PoS etiketi) ikilisi döndüren fonksiyon
randomize: girdiyi rastgele sıralanmış olarak geri döndüren fonksiyon

Input: $x, y, y^*, PT, \alpha, \delta, \theta$
Output: x^*

$y^*_{conf} \leftarrow Conf(y^*|x)$
$candidates \leftarrow randomize(tokenize\_and\_tag(x))$
$n \leftarrow 0$
$x^* \leftarrow x$
for all c in $candidates$ **do**
  if $n \geq \delta$ **or** $y^*_{conf} \geq \theta$ **then**
    break
  end if
  $PT_{res} \leftarrow PT.term{=}c.token$ **and** $PT.tag{=}c.tag$
  if $PT_{res} \in \emptyset$ **then**
    continue
  end if
  $lower \leftarrow PT_{res}.score - \alpha$
  $upper \leftarrow PT_{res}.score + \alpha$
  $PT_{res} \leftarrow PT_{res}[lower < PT_{res}.score < upper$
  for all r in PT_{res} **do**
    $x_{temp} \leftarrow$ swap $c.token$ with $r.term$ in $x*$
    $y^*_{temp} \leftarrow Conf(y^*|x_{temp})$
    if $y^*_{temp} > y^*_{conf}$ **then**
      $x^* \leftarrow x_{temp}$
      $y^*_{conf} \leftarrow y^*_{temp}$
      $n \leftarrow n + 1$
    end if
  end for
end for
return x^*

Şekil 3. 4 Kutupluluk tabanlı pertürbasyon algoritması



Şekil 3. 5 Kutupluluk tabanlı pertürbasyon blok şeması

Şekil 3. 6 Kutupluluk tabanlı pertürbasyon örneği

3.9 Kısıtlı Ortam Şartları

Tamamen Gözlenebilir Kısıtlanmış Ortam. Düşman topladığı tüm örnekleri etiketleyemediğinde, kutupluluk tablosu "sınırlı" hale gelecektir. Bu ortamda, rakip bulabildiği kadar çok örnek toplamıştır, ancak hepsi için etiket oluşturamamıştır. Bu, saldırganın daha az örneği inceleyeceği anlamına gelir. Etiketleme bütçesi çok düşükse, kutup puanları terimlere doğru önemi atayamaz ve rakibin hedefini kaçırmasına neden olabilir. Deneylerimizde, saldırılarda başarılı olunabilmesi için bol miktarda etiketli örneğin olmasının büyük önem taşıdığı görülmüştür.

Yarı Gözlenebilir Kısıtlanmış Ortam. Kurban model için sorgu sayısı sınırı olduğunda geçerlidir. Sorgu gerçekleştirme ya pahalıdır ya da hizmet sunucusunda şüpheli sorgulama ile işaretlenmesine yol açacak bir önlem vardır. Naive Bayes ve BiLSTM deneylerinde örnek başına (10.000) sorgu sınırı verilirken, gerçek kara kutun için miktar (100)'e düşürülmüştür.

3.10 Veri Kümeleri

Sınıflandırma için ikili ve çoklu sınıflandırma olarak, altı sınıflandırma veri kümesi kullanılmıştır. İki görev incelenmiştir: duygu sınıflandırma ve kategorizasyon. Tablo 3.3'te veri kümelerinin özelliklerini gösterilmiştir. Tüm modellerde 2^{16} öznitelik kullanılmış ve sonuçta ortaya çıkan bilinmeyen kelime oranları da Tablo 3.3'te rapor edilmiştir.

Tablo 3. 3 Veri kümeleri

Veri Kümesi	Görev	#Sınıf	#Eğitim	#Test	Sözlük Boyutu	Bir örnekte ort. sözcük sayısı	UNK Oranı
yelp	Duygu Sınıflandırma	2	560K	38K	234830	158	0.31
amazon	Duygu Sınıflandırma	2	3.6M	400K	869440	88	0.31
imdb	Duygu Sınıflandırma	2	25K	25K	75069	275	0.31
dbpedia	Kategorizasyon	14	560K	70K	649764	53	0.30
ag_news	Kategorizasyon	4	120K	7600	67463	38	0.37

3.11 Kurban Sınıflandırıcı Modeller

Deneylerde üç sınıflandırıcı kullanılmıştır: Naive Bayes sınıflandırıcı (NB), BiLSTM sınıflandırıcı ve IBM Watson sınıflandırıcı. NB ve BiLSTM sınıflandırıcıları, sözcük tabanlı modellerdir. IBM Watson Sentiment Classifier bizim için tam bir kara kutudur, ne mimarisi ne de eğitim kümesi hakkında bilgimiz yoktur; saldırılar oluşturulurken ihtiyaç duyulan sınıf etiketi ve olasılık puanının sağlandığı bir web hizmetidir.

NB sınıflandırıcı monogram TD-IDF özelliklerini kullanırken BiLSTM eğitim kümesi üzerinde kelime yerleştirmelerini öğrenir. BiLSTM sınıflandırıcı hiperparametreleri Bayesian optimizasyon ile tüm veri kümeleri için optimize edilmiştir. Kelime gömme boyutu 64, dizi uzunluğu 200, batch boyutu 128, ve BiLSTM katmanında 128 nöron bulunmaktadır. Model 0.5 dropout ve eğitimde erken durma ile düzenli hale getirilmiştir. Duygu analizi ve kategorizasyon görevleri için parametreler ve mimari, çıktı sınıflarının sayısı dışında aynıdır. Sınıflandırıcıların özellikleri ve ilgili veri kümelerindeki doğruluk puanları Tablo 3.4'te gösterilmektedir.

Tablo 3. 4 Kurban model başarısı

Sınıflandırıcı / Özellikler	**Yelp**	**Amazon**	**IMDB**	**DBPedia**	**AG News**
Naive Bayes	0.84	0.79	0.80	0.95	0.88
Bi-LSTM	0.96	0.94	0.79	0.99	0.91
IBM Watson	0.83	0.76	0.85		

3.12 Basit Saldırganlar

Kutupluluk tabanlı pertürbasyon yöntemiyle karşılaştırmak üzere iki temel saldırgan tasarlanmıştır. Her ikisi de karakter pertürbasyon modellerinin etkisini simüle eder: bir kelimeyi bilinmeyen bir simgeye dönüştürmek. Basit saldırganlar, işaretledikleri kelimeleri <unk> simgesine dönüştürür.

Rastgele saldırgan, örnekten δ rastgele kelime seçer ve hedeflenen sınıf için θ olasılık puanına ulaşılana veya seçilen tüm kelimeler değiş tokuş edilene kadar bunları <UNK> belirteci ile değiştirir.

İkinci basit saldırgan, orijinal sınıf dağılımındaki popülerliklerine dayalı olarak kelimelerin naif seçimidir. Kaynak ve hedef sınıflarına göre en iyi T kelimesi seçilirken seçerken sıralı kutupluluk tablosu kullanılmış ve bunlar "sıcak terimler" olarak adlandırılmıştır. Örneğin, DBPedia veri setinde hedef sınıf "Şirket" ve hedef sınıf "Eğitim Kurumu" için, sıralanmış "Şirketten Eğitim Kurumuna Kutupluluk" puanlarını kullanılır ve "sıcak terimler" olarak ilk T adet terim seçilir. Daha sonra, işaretlenen kelimelerin rastgele seçilen δ adet sıcak terime kadar değiş tokuş yaparak rakip örnekler oluşturulur.

3.13 Deneysel Sonuçlar

Önerdiğimiz çekişmeli üretim algoritması, çok az sayıda büyük kutupluluk değişimini hedeflemiştir. Seçilen bir terim, aday havuzundan başka bir terim ile takas edildiğinde, bozulma meydana gelir. Aday havuzu PoS etiketleri ve izin verilen maksimum kutupluluk puanı bozulması ile sınırlandırılır. Rakibin amacı, minimum sayıda değiştirme gerçekleştirerek belirli bir hedef sınıfın güven puanını artırmaktır.

Bir sınıf, bir örnek için 0.51 güvenirliğe ulaştığında, sonuç etiketi, sınıf sayısından bağımsız olarak o sınıfa ait olacaktır. Bu nedenle, tüm deneyler boyunca hem ikili duygu hem de çok sınıflı kategorizasyon problemlerinde θ=0.51 alınmıştır. Saldırgan gücünü artırmak için θ'yi daha yüksek bir güven değerine eşitlenebilir, karşılaştırma amacıyla bir kontrol değeri kullanılabilir.

Çok sınıflı modeller için, tüm test kümesindeki karışıklık matrisleri kullanılarak en çok karıştırılan sınıf yönünde etiket değişimi hedeflenmiştir. Aşağıdaki alt bölümlerde, mükemmel koşullarda, aktarılabilirlik, bütçe ve limit durumu ve son olarak gerçek bir kara kutu ortamına yönelik saldırılar incelenmiştir.

3.14 Mükemmel Koşullar Altında Saldırı

Mükemmel ortamda incelememiz sırasında, her deney stratejisi (Tablo 3.5) için aşağıdaki deneysel sorular sorulmuştur:

- **Sınırsız saldırı**: Maksimum bozulma aralığına izin verdiğimizde, düşmanın hedefini gerçekleştirmek için yapması gereken takas sayısı konusunda bir dirsek noktası var mı?
- **Kontrollü bozulma:** Düşman, temiz bir örnek üzerinde birçok küçük değişiklikle başarılı saldırı gerçekleştirebilir mi?
- **Kontrollü takas sayısı:** Düşman, temiz bir örnek üzerinde birkaç büyük değişiklikle başarılı saldırılar gerçekleştirebilir mi?
- **Sınırlı saldırı:** Düşman, temiz bir örnek üzerinde birkaç küçük değişiklik yaparak hedefine ulaşabilir mi?

Tablo 3. 5 Deney stratejileri

Strateji	**Bozulma($\propto$)**	**#Takas(δ)**
Sınırsız Saldırı	5.0	100
Kontrollü Saldırı	1.0	100
	5.0	5
Sınırlı Saldırı	1.0	5

Şekil 3.7 ve 3.8, mükemmel şartlar altında oluşturulan çekişmeli örneklerin özelliklerini özetlemektedir. Grafikler, gerçekleştirilen takas (δ) işlemlerinin ortalama sayısını ve nihai-orijinal etiketler güven değerleri arasındaki farkı ($\Delta\theta$) gösterir. Dört strateji ayrı grafiklerde gösterilmiştir. Duygu tanıma modelleri ve çok sınıflı modeller, stratejiler altında kendi grafiklerine sahiptir ve iki sınıflandırma tekniği (NB ve BiLSTM) grafik üzerinde belirtilmiştir. Her nokta bir kurban modelini temsil eder. İdeal olarak, modellerin, grafiğin sol üst alanına karşılık gelen bölgede bulunması, yani hedef sınıf için yüksek güven değişikliği ile minimum sayıda takas ile kandırılması idealdir.

Kutupsallık puanlarını hesaplamak için modellerin test kümeleri kullanılmıştır, örn. AG News eğitim kümesi ile eğitilen model, AG News test kümesinden oluşturulan

kutupluluk tablosuyla 100 adet dışarıda kalan (left out) AG News test örneğini bozmak için kullanılır. Kara kutu ortamında, düşmanın eğitim verisi hakkında hiçbir bilgisi yoktur. Bu nedenle, Varsayım 2 ile toplanan verinin, kurban modelinin eğitildiği verilerle aynı dağılımda olmayacağını beklenen bir durumdur. Duygu tanıma görevlerinin tümünün "pozitif" ve "negatif" etiketlerin ikili sınıflandırmaları olduğu düşünüldüğünde, etki alanları arası (yani farklı eğitim verisi ile eğitilmiş modeller) kutupluluk tabloları kullanılabilir. Örneğin, Amazon veri kümesi ile eğitilen modeller, 100 adet dışarıda bırakılmış Amazon test örneğini bozmak için ayrı ayrı Yelp, Amazon ve IMDB test setlerinden gelen kutupluluk tablolarıyla bozulmaya uğratılır. Ayrıca, çok sınıflı modeller için hedefli ve ayrım gözetmeyen saldırılarda geliştirilen yöntemin etkinliği incelenmiştir.

Sınırsız Saldırı: Takas işlemlerinin sayısı için bir dirsek noktası aramak üzere algoritmada uygun parametre değerlerinin ayarlanması gerekir. Kutupluluk puanları logaritmik ölçekte olduğundan, α=5.0'lık bir bozulma, takas terimleri için sınırsız bir aday havuzu sağlayacaktır. Bu konfigürasyon, δ=100 değeri ile birlikte kullanıldığında, deneysel ortam neredeyse bir 'sınırsız saldırı' ile eşdeğerdir. Şekil 3.7 (S1) ve Şekil 3.8 (M1)'de kurban modellere karşı gerçekleştirilen sınırsız saldırılar gösterilmiştir.

Kutupluluk tabanlı saldırılar için BiLSTM modelleri, NB muadillerinden belirgin şekilde daha yüksek hedef güveniyle kandırılmıştır. Tüm duygu analizi görevleri, 7 veya daha az ortalama değişiklik gerçekleştirerek başarıyla kandırılmıştır. İlgili dizi uzunluklarına karşı takas sayısının oranı çok düşüktür, bu da orijinal örnekler üzerinde "birkaç gürültülü değişiklik yapma" hedefini sağlar.

Şekil 3.7 (S1)'deki etki alanları arası kutupluluk tablolarının sonucu incelendiğinde, NB için Amazon kutupluluk tablolarının saldırı oluşturmada çok daha etkili olduğunu görülmektedir. BiLSTM modelleri IMDB kutupluluk tablosuyla daha güvenle kandırılsa da, Amazon kutupluluk skorları ile daha az takas yapılarak hedefe ulaşılabildiğini göstermiştir. Öncelikli saldırgan hedefi, daha az değişiklikle başarılı saldırılar olduğunda, Amazon kutupluluğunun daha yararlı olarak olduğu açıktır.

Amazon kutupluluk tablosuyla bir örneğin sınıfını değiştirmek için yalnız birkaç takasın gerekli olması etkileyicidir; ancak tablonun oluşturulmasında kullanılan devasa veri kaynağı (400 bin örnek) göz önüne alındığında şaşırtıcı değildir. Kutupluluk kaynağının büyüklüğü, puanların problemi ne kadar doğru temsil ettiğinin bir belirleyicisidir ve bunun dizi uzunluğundan daha önemli olabileceği görülmektedir.

Kategorizasyon görevleri için eğitilmiş BiLSTM modellerin, çok sınıflı NB'lere kıyasla daha az takas ile daha fazla güven skoru ile kandırılabileceğini görülmektedir (Şekil 3.8 (M1)). Daha iyi bir sınıflandırıcı kullanarak çekişmeli örnekler oluşturmanın daha kolay olduğu söylenebilir. Burada kolaylık, daha az sayıda takas yapılmasına karşılık gelir. AG News'in sınıflandırılacak dört kategorisi vardır ve yüksek güvenle daha az değişiklik yapılarak kandırılabilir. DBPedia modelleri on bir sınıfa ayrılır ve daha fazla sayıda takas gerektirir.

Kontrollü Bozulma: Bu deney setinde, "birçok küçük değişiklikle" hazırlanmış rakip örnekler değerlendirilmiştir. Sınırsız saldırılara benzer şekilde, BiLSTM modelleri NB'den çok daha yüksek bir güvenle kandırılır (Şekil 3.7 (S2) ve 3.8 (M2)).

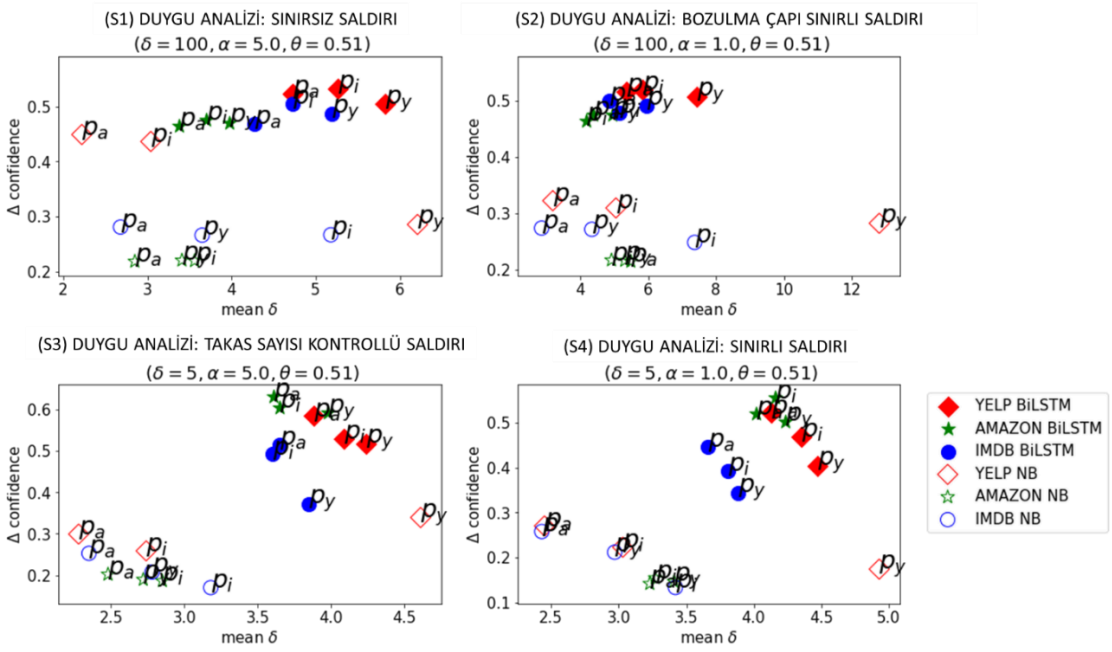
Bu saldırı stratejisinde, düşman, duygu görevi için bozuk örnekleri oluşturmak üzere çoğunlukla sekizden daha az ortalama değişikliğe ihtiyaç duymuştur. Etki alanları arası kutupluluk deneylerinde genelleştirilebilir bir model görülmemiştir, ancak Amazon kutupluluk tablosunun bu deney düzeneğinde de rakip örnekleri oluşturmak için son derece iyi bir kaynak olduğunu gözlemlenmiştir.

Çok sınıflı modeller, sınırsız deneylere benzer bir sonuç sergiler. AG News modelini kandırmak çok daha kolaydır. AG News'de daha yüksek güven farkı ile daha az takas ile ayrım gözetmeyen saldırgan örnekler elde edilebilirken, DBPedia modeli için aynı durum söz konusu değildir.

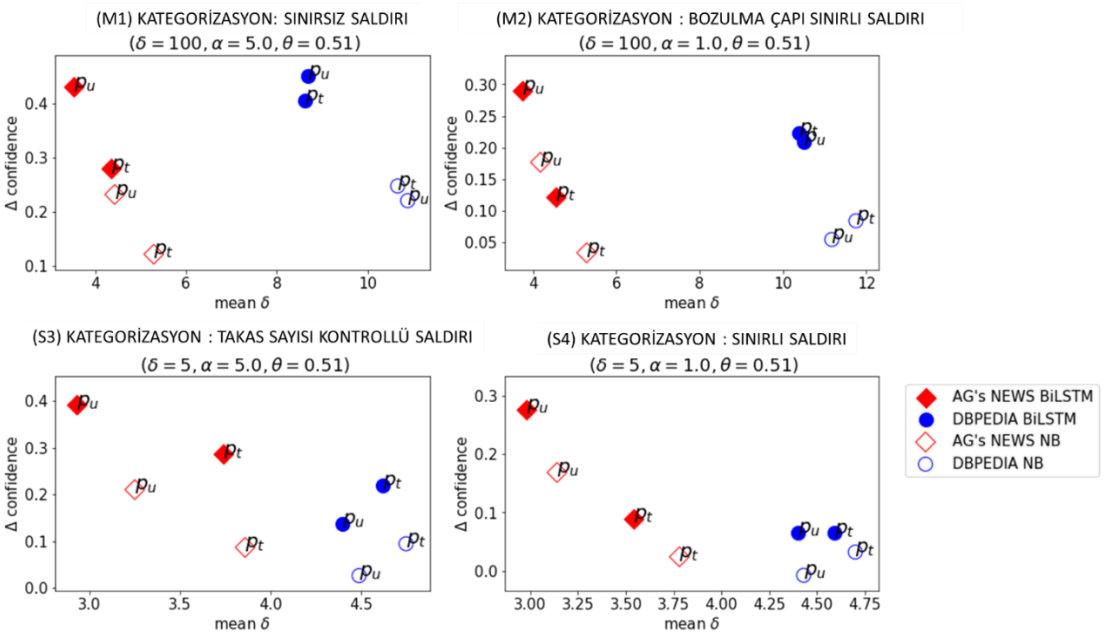
Kontrollü Takas Sayısı: Şekil 3.7 (S3), NB'nin daha az takas ile kandırılabileceğini, ancak düşmanın BiLSTM modellere daha güvenle saldırabileceğini göstermektedir. Yine, Amazon kutupluluk tablosu, duygu tanıma görevlerinde yüksek güvenle örnekleri bozmak için önemli bir kaynaktır. Çok sınıflı sonuçlarda Şekil 3.8 (M3), AG News ve DBPedia arasındaki saldırı güveni farkı burada da izlenmektedir. Her iki

kontrollü stratejide de, DBPedia modeli için saldırı güveninde ciddi bir güven değişikliği olduğunu gözlemlenmiştir.

Sınırlı Saldırı: Önceki stratejilerdeki sonuçlara benzer şekilde, BiLSTM bozulması, bu deney setinde de daha yüksek güven farkı oluşturmuştur (Şekil 3.7 (S4), 3.8 (M4)). Duygu tanıma görevlerinde, NB modeller için daha fazla terim takas maliyeti görülmüştür. Etki alanları arası kutupluluk deneylerinde, Amazon kutupluluk puanları en yüksek hedef güveniyle sonuçlanmıştır. Algoritma 1.0'lık küçük bir bozulma yarıçapı içinde beş veya daha az takas ile kısıtlandığında, duygu tanıma görevlerinde karşı sınıfa yüksek güven atayan bozuk örnekler oluşturmaya devam edilebildiği görülmektedir.

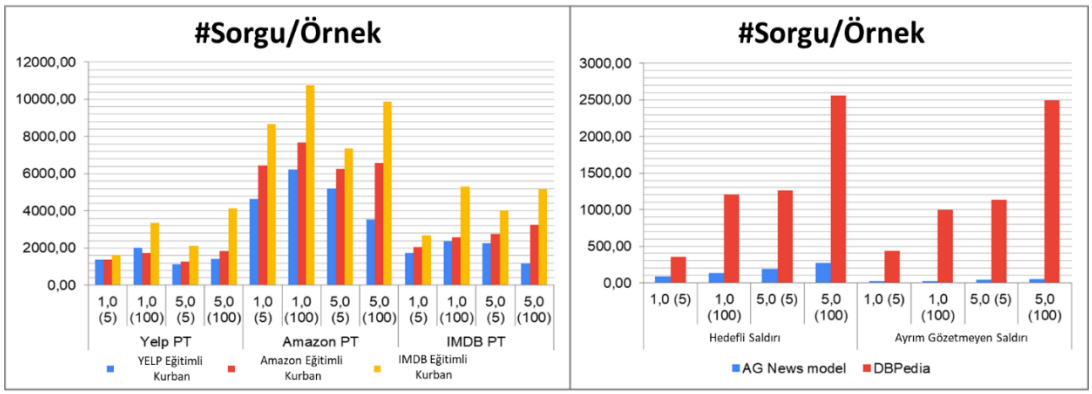


Şekil 3. 7 Duygu analizi saldırıları



Şekil 3. 8 Kategorizasyon saldırıları

Kutupluluk saldırıları, kara kutu modelinden elde edilen yüksek güven sonuçlara büyük ölçüde bağlıdır. Düşman, hizmete hem zaman hem de parasal maliyet açısından pahalı olabilen kara kutu sorgusu gönderir. Bunun, kara kutudaki herhangi bir sorgu tabanlı saldırı yöntemi için bir zayıflık olduğunu açıktır. Bu konuda daha fazla bilgi sağlamak için Şekil 3.9 sunulmuştur.



Şekil 3. 9 Örnek başına ortalama sorgu sayıları

Duygu tanıma modeli deneylerinde, saldırıların oluşturulmasında kullanılan kutupluluk tablosunun, tek bir örnek oluşturmak için gönderilen isteklerin sıklığına göre belirlendiği açıkça görülmektedir. Daha önce belirtildiği gibi, Amazon test örnekleri ile elde edilen kutupluluk puanları, daha yüksek saldırı güveni sağlar. Bunun sebebinin örneklerin zenginliği (400K) olduğu çıkarımı yapılmıştır. Daha

büyük veri kaynağı ile kutupluluk puanları daha hassas bir şekilde hesaplanabilir. IMDB kutupluluk puanları Yelp puanlarından daha iyi bir kaynaktır çünkü daha uzun dizilere sahip örneklerden oluşur. Diğer bir deyişle, kutupluluk kaynağının zenginliği, hem veri kümesindeki örnek sayısı hem de dizi uzunluğu ile belirlenebilir. Kutupluluk kaynağı zengin olduğunda, hesaplanan puanlar birbirine yakın olabilir ve bu da aşırı sorgulamaya neden olabilir. Bu nedenle, kendinden emin saldırılar ile örnek başına sorgu maliyeti arasında bir ödünleşim (trade-off) vardır. Beklediğimiz gibi, daha geniş aday arama alanı nedeniyle orijinal örneklerin dizi uzunluğu da sorgu maliyetinde belirleyici olduğu görülmektedir. Bununla birlikte, daha önemli bir rol oynayanın örnek sayısı mı yoksa dizi uzunluğu mu yoksa kaynak örneklerin kalitesi mi olduğunu analiz etmek için bu çıkarımların incelenmesi gerekmektedir.

Daha önce belirtildiği gibi, AG News modelleri kutupsallık saldırılarına karşı daha savunmasızdır. Orijinal örneklerin kaynak zenginliği ve dizi uzunluğu göz önüne alındığında, ters bir eğilim görüyoruz. AG News rakip örneklerini oluşturmak çok daha az maliyetlidir; kutupluluk kaynağı karşılaştırıldığında çok zayıf olmasına ve dizi uzunluklarının daha kısa olmasına rağmen. Bunun nedeninin sınıf sayısı olduğu düşünülmektedir. AG News modelleri sadece 4 sınıfa ayrılırken DBPedia'nın 11 kategorisi vardır.

Her iki sınıflandırma alanında da, sınırlı saldırılarda daha az sorgu maliyeti görülmüştür. Hem δ hem de α arama uzayının kapsamını belirlediğinden, bu beklenen bir sonuçtur.

ELMo yerleştirmeleri (Peters ve diğerleri, 2018), kelimelerin sözdizimsel ve anlamsal bilgilerini korur ve iki metin örneği arasındaki anlamsal benzerliği analiz etmek için kullanılabilir. Kara kutu hizmet sağlayıcısı, girişler için çekişmeli örneklere karşı anlamsal bir değerlendirme uygulayabilir. Anlamsal değerlendirmenin bir yolu, sözcük yerleştirmelerini kullanmaktır. Google'ın önceden eğitilmiş ELMo v3 veri yapısını kullanarak, orijinal ve bozuk örnekler arasındaki kosinüs benzerliği karşılaştırılmıştır. Tablo 3.6 ve 3.7, sınırsız ve sınırlı saldırı stratejileri altındaki ortalama benzerlik sonuçlarını göstermektedir. Bozuk duygu örneklerinin, anlamsal olarak orijinallere çok benzediği görülmektedir. Daha

düşük oranlarda olsa da, çok sınıflı saldırı örnekleri için de orijinalleri ile benzerlik oranı 0.90'lardadır. Sonuçlar, saldırı örneklerinde çok fazla anlamsal değişiklik olmadığını göstermektedir.

Tablo 3. 6 Duygu tanıma problemleri için orijinal ve saldırı örneklerinin benzerliği

	Kutupluluk Kaynağı		
Model	**Yelp**	**Amazon**	**IMDB**
Yelp NB, Sınırlı	0,96	0,98	0,98
Yelp BİLSTM , Sınırlı	0,97	0,97	0,97
Yelp NB, Sınırsız	0,98	0,98	0,98
Yelp BiLSTM, Sınırsız	0,97	0,97	0,98
Amazon NB, Sınırlı	0,97	0,97	0,97
Amazon BiLSTM, Sınırlı	0,97	0,97	0,97
Amazon NB, Sınırsız	0,97	0,98	0,97
Amazon BiLSTM, Sınırsız	0,97	0,97	0,97
IMDB NB, Sınırlı	0,99	0,99	0,99
IMDB BiLSTM, Sınırlı	0,99	0,99	0,99
IMDB NB, Sınırsız	0,99	0,99	0,99
IMDB BiLSTM, Sınırsız	0,98	0,99	0,99

Tablo 3. 7 Kategorizasyon problemleri için orijinal ve saldırı örneklerinin benzerliği

Model	**Benzerlik**
AG News NB, Sınırlı	0,93
AG News BİLSTM, Sınırlı	0,94
AG News NB, Sınırsız	0,90
AG News BiLSTM, Sınırsız	0,93
DBPedia NB, Sınırlı	0,90
DBPedia BiLSTM, Sınırlı	0,90
DBPedia NB, Sınırsız	0,93
DBPedia BiLSTM, Sınırsız	0,88

3.15 Aktarılabilirlik

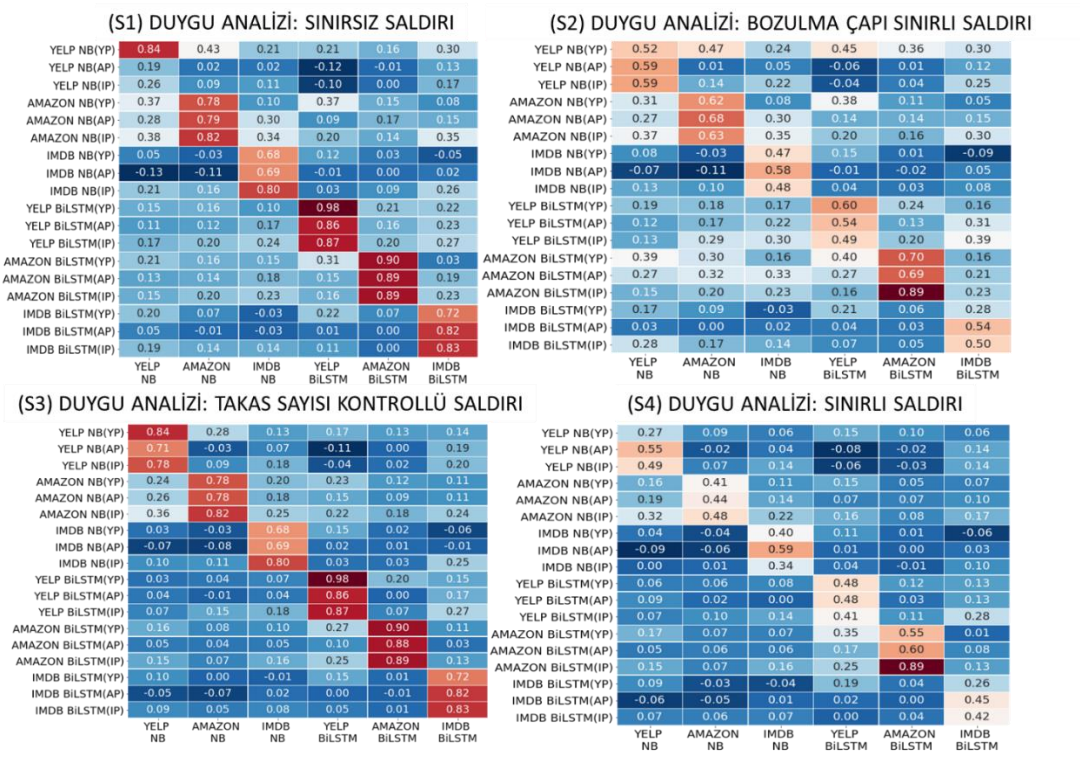
Bu bölümde, mükemmel ortam deneylerinden elde edilen saldırı için örnek ve çapraz teknik aktarılabilirliği değerlendirilmiştir. Şekil 3.10'da, etki alanları arası kutupluluk deneylerini içeren duygu analizi kurban modellerinde (sütunlar) ilgili

saldırıların (satırların) doğruluk farkını gösterir. Bu fark ne kadar yüksek olursa, saldırılarımızın başarısı da o kadar yüksek olur.

Bu alt bölümdeki tüm değerlerde dikkat çeken ilk husus, saldırıların çok aktarılabilir olmamasıdır. Kurban kara kutuya özel düşman örnekleri oluşturulduğu için, bu beklenen bir sonuçtur. Bozuk örnekler, saldırganın bunları oluştururken sorguladığı kurban modeli için son derece özelleştirilmiştir.

Deneyler boyunca, BiLSTM modellerinin daha yüksek güven seviyelerinde daha az sayıda takas ile kandırılabileceğini gözlemlenmiştir. Bu, BiLSTM modellerinde daha yüksek doğruluk farkı ile doğrulanmıştır.

İkili bir sınıflandırıcı için, doğruluktaki %50 değişiklik, bütünlükten ödün verilmesidir. Sınırsız ve kontrollü takas stratejileri üzerinde, örneklerin genel olarak bütünlüğü tehdit edecek derecede çok etkili olduğunu görüyoruz. Etki alanları arası kutupluluk (köşegenler) ile üretilen bozuk örnekler arasında NB sınıflandırıcıları arasında aktarılabilir olduğu görülmektedir. Saldırgan sınırlı strateji uyguladığında bile, sistem bütünlüğüne yönelik tehdit çoğunlukla geçerliliğini korunur.



Şekil 3. 10 Duygu analizine saldırıların aktarılabilirlik matrisi

AG News NB kurban modeli büyük oranda kutupluluk saldırılarına karşı gürbüzken, BiLSTM kurbanı hedefli saldırılara karşı çok savunmasızdır (Tablo 3.8). DBPedia, sınırsız saldırı stratejisinde çok savunmasızdır; ayrıca BiLSTM modeli, aynı NB modelinin sorgulanmasıyla oluşturulan hedefli ve hedefsiz saldırılara karşı daha savunmasızdır. Saldırıların, çoğunlukla tüm çok sınıflı modeller için sınırlı stratejide başarısız olduğu görülmektedir.

Tablo 3. 8 Kategorizasyon modellerine saldırıların aktarılabilirlik matrisi

Kurban Modeller	AG News NB (T)	AG News NB (U)	AG News BiLSTM (T)	AG News BiLSTM (U)	DBPedia NB (T)	DBPedia NB (U)	DBPedia BiLSTM (T)	DBPedia BiLSTM (U)
sınırsız (δ=100; α=5.0)								
NB	0,14	0,03	0,00	0,01	**0,47**	**0,46**	0,04	0,04
BiLSTM	0,21	0,12	**0,42**	0,20	0,13	0,14	**0,66**	**0,65**
δ–Kontrollü (δ=5; α=5.0)								
NB	0,05	0,00	0,00	0,00	0,12	0,05	0,01	0,01
BiLSTM	0,13	0,07	**0,32**	0,17	0,04	0,03	**0,26**	0,17
α-Kontrolllü (δ=100; α=1.0)								
NB	0,03	-0,03	0,00	0,00	0,10	0,06	0,02	0,01
BiLSTM	0,11	0,02	0,18	-0,01	0,04	0,02	**0,37**	**0,31**
Sınırlı (δ=5; α=1.0)								
NB	0,01	-0,01	0,00	0,00	0,03	0,00	0,01	0,00
BiLSTM	0,05	-0,02	0,10	-0,02	0,00	0,00	0,07	0,02

3.16 Basit Saldırganlarla Karşılaştırma

Her iki basit saldırgan da seçili terimleri, sözlükte bilinmeyen kelimelerin işaretlenmesi için kullanılan <UNK> sembolüne dönüştürür. Bu temel saldırganlar ile karakter pertürbasyonlarının etkileri test edilmiştir. Şekil 3.11'deki blok diyagramın akışı hala geçerlidir. Değişiklik, pertürbasyon şemasındadır.

Rastgele saldırıda, istenen güven değeri elde edilene kadar δ adet kelime rastgele olarak <UNK> ile takas edilir. "Sıcak terim" saldırganı kutupluluk tablosundan bir hedef $y \rightarrow y^*$ değişimi için en sık yüksek puanlı 1000 kelime arasından takas gerçekleştir.

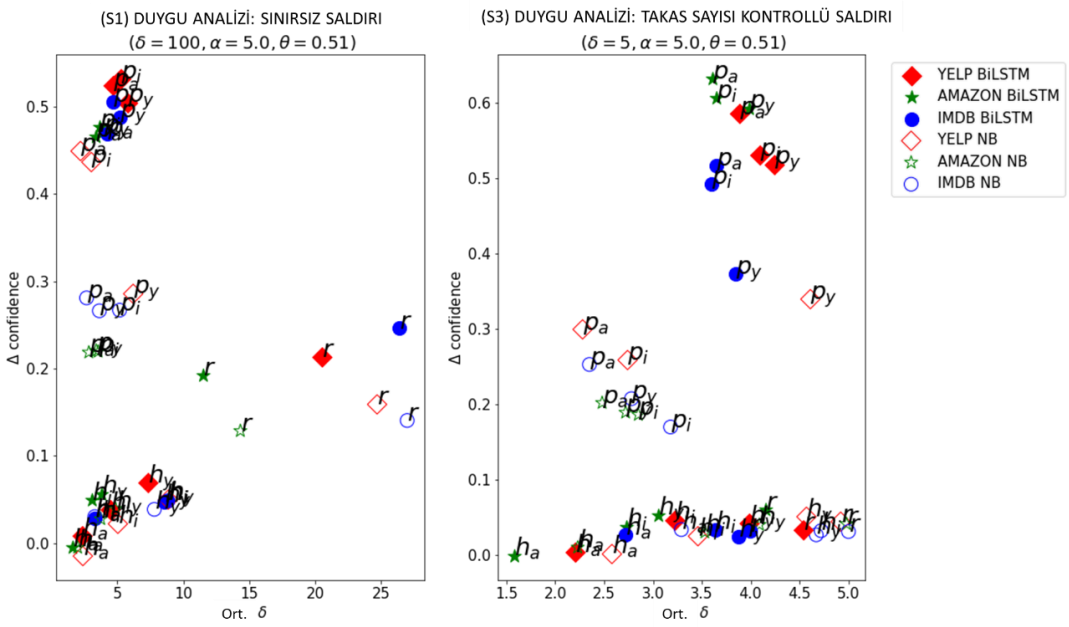
Şekil 3.11 ve 3.12, kontrollü takas sayısı ve sınırsız stratejiler altındaki sonuçları göstermektedir. Grafikler rastgele, sıcak terimler ve kutupluluk saldırılarını

gösterir. Şekil 3.11 (S1) ve 3.11 (S3) duygu analizi kurban modelleri için hazırlanan saldırı örneklerine dair sonuçları içerir:

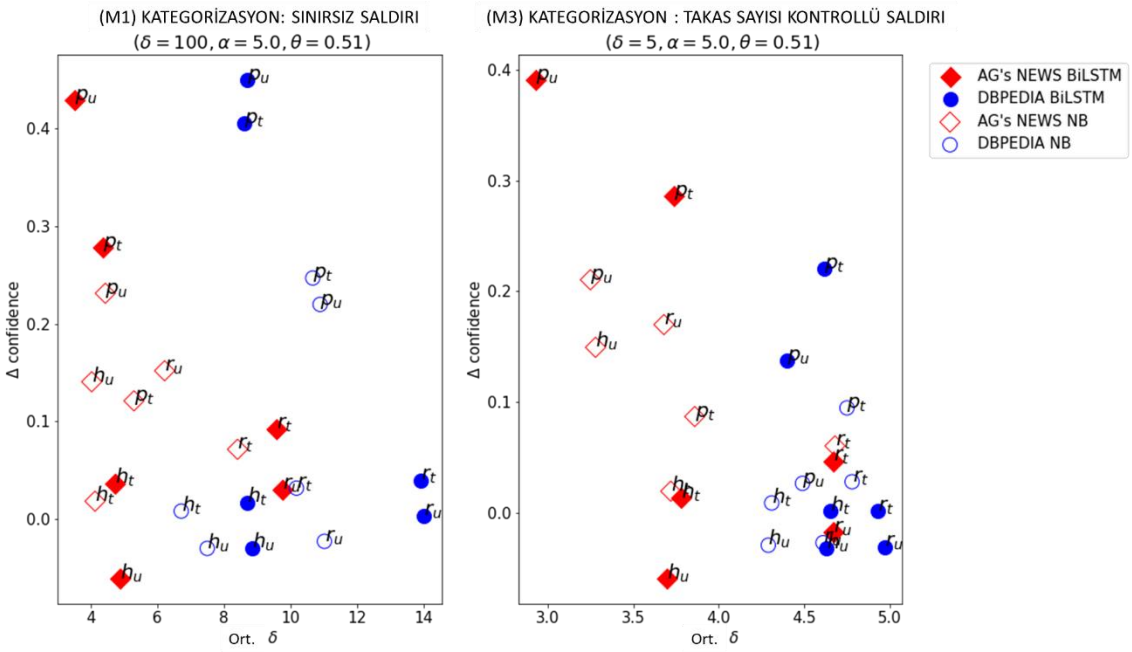
- p_{ap} ve h_{ap} : AMAZON test setinin kutupluluk tablosunu kullanan kutupluluk ve sıcak terim saldırılarını,
- p_{ip} ve h_{ip} : IMDB test setinin kutupluluk tablosunu kullanan kutupluluk ve sıcak terim saldırılarını,
- p_{yp} ve h_{yp} : YELP test setinin kutupluluk tablosunu kullanan kutupluluk ve sıcak terim saldırılarını,
- r: rastgele saldırı noktalarını gösterir.

Şekil 3.12 (M1) ve 3.11 (M3) kategorizasyon kurban modelleri için hazırlanan saldırı örneklerine dair sonuçları içerir:

- p_{ta}, h_{ta}, ve r_{ta}: hedefli kutupluluk, sıcak terimler ve rastgele saldırıları,
- p_{ua}, h_{ua}, ve r_{ua}: ayrım gözetmeyen kutupluluk, sıcak terimler ve rastgele saldırıları noktalarını gösterir.



Şekil 3. 11 Duygu tanıma problemleri için kutupluluk ve basit saldırı sonuçları karşılaştırması



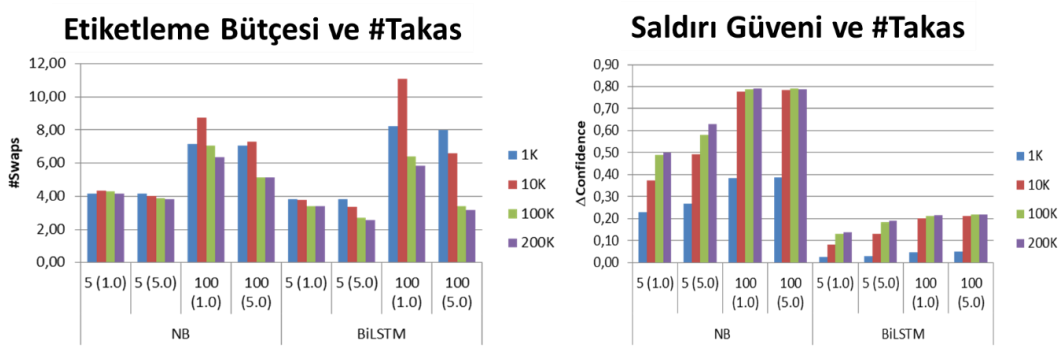
Şekil 3. 12 Kategorizasyon problemleri için kutupluluk ve basit saldırı sonuçları karşılaştırması

Kutupluluk saldırıları, basit saldırılarla karşılaştırıldığında açıkça daha yüksek güven seviyelerine sahiptir. Sıcak terim saldırılarında da saldırı yönü kullanır, ancak arama gerçekleşmez; sadece en önemli terimler saldırı doğrultusunda takas edilir. Yüksek kutupluluk puanlı terimlerin basit bir şekilde değiştirilmesinin işe yaramadığını görülmüştür. Rastgele saldırılar, ne seçimde ne de değiştirmede saldırı yönü bilgisini kullanmaz. Yine de kara kutu sistemini sorgularken, bazen sıcak terimler yönteminden daha güvenli bozuk örnekler oluşturmayı başarabildikleri görülmüştür.

3.17 Sınırlı Ortam Koşullarında Saldırı

Varsayım 3.1 ve 3.2'ye göre düşman, problem alanı hakkında bilgi sahibidir (örneğin sınıf etiketleri) ve kutupluluk puanlarını hesaplamak için sınırsız sayıda örneği maliyetsiz olarak toplayabilir. Bu örneklerin etiketlenmesi ise, maliyetli bir işlemdir.

Amazon test seti, kutupluluk tablosu oluşturmak için kullandığımız en büyük kümedir. Bu nedenle, etiket bütçesinin etkisini araştırmak için Amazon kurban modellerini kullanılmıştır. Şekil 3.13, çeşitli etiketleme limitleri ile tüm saldırı stratejileri ve sınıflandırıcılar genelinde ortalama takas sayısını ve ortalama güven değişimini göstermektedir.



Şekil 3. 13 Amazon kutupluluk tablosuyla üretilen saldırı örneklerinin takas sayılarının etiketleme bütçesi ve saldırı güveni açısından incelenmesi

Hedef, düşük sayıda takas ve yüksek güven farkı olmasıdır. Kutupluluk tablosunu oluştururken kullanılan etiketli veri miktarının saldırı güveniyle ilişkili olduğu daha önce belirtilmişti. Şekil 3.13'teki sonuç da bu çıkarımı destekler niteliktedir. Ayrıca, hem takas sayısı hem de saldırı güveni için 100K ve 200K etiket bütçeleri arasında çok az fark vardır. 200K etiket için iyileştirmeler olduğu görülmektedir, ancak etiket maliyeti ile performans iyileştirmesi arasındaki ödünleşim çok büyük olabilir.

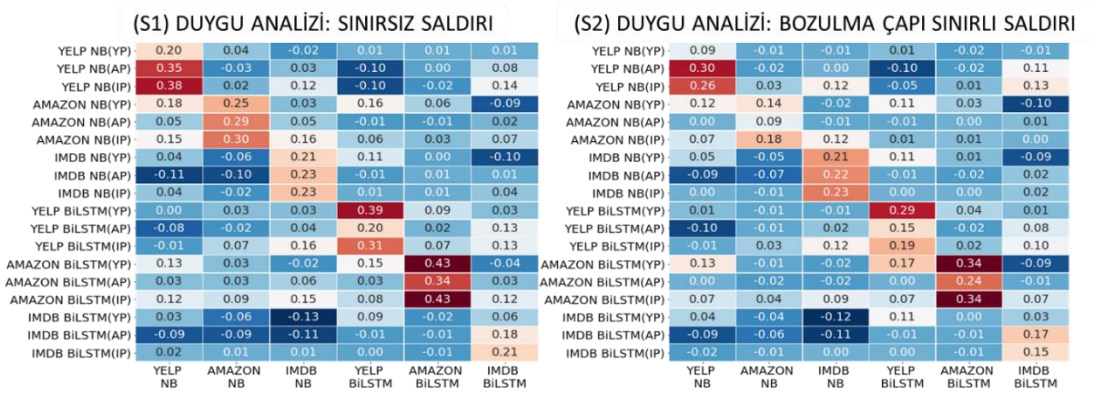
Güven puanlarının olmaması, kara kutu ortamında karşılaşabileceğimiz başka bir kısıtlamadır. Sistem bu bilgiyi istemcilere sağlamayabilir. Algoritma 1'i, güven iyileştirme kriterlerini kaldırarak bu kısıtlamaya uyacak şekilde düzenleyerek, güven artırma yerine kara kutu sorgusu hedef sınıfı döndürdüğünde algoritma durdurulma kriteri eklenmiştir (Şekil 3.14). NB ve BiLSTM kurbanları için 10000 sorgu bütçesi belirlenmiş ve δ kaldırılmıştır. Bu ortam koşullarında istenen güven eşiğini kontrol edilemez.

if $n_{query} \geq budget$ **or** $victim(x^*) = y^*$ **then**
  break
end if

Şekil 3. 14 Güven puanı kısıtında algoritma sonlandırma koşulu

Şekil 3.15 ve Tablo 3.9 güven puanları kullanılmadan oluşturulan örnekler üzerinde kurban modellerinin sınıflandırma doğruluğundaki farkı göstermektedir. Duygu analizi modelleri sonuçları mükemmel ortamda olduğu kadar başarılı olmadığı görülmektedir. Bunun nedeni, saldırıların daha kör bir şekilde oluşması; tahmin sınıfı değişmediği sürece, saldırganın adaylar arasında hangi takasın daha iyi olduğunu bilmesinin bir yolu olmayışıdır. Bu düşük başarıya rağmen, duygu analizi

kurbanları üzerinde doğruluğun azaldığı görülmektedir. Diğer ortamda olduğu gibi, burada da aktarılabiliğin yüksek olmadığı görülmektedir.



Şekil 3. 15 Güven puanı kısıtında duygu analizi için üretilen saldırı örneklerinin aktarılabilirliği

Bu senaryoda, hem sınırsız hem de α kontrollü saldırılar, çok sınıflı modeller için mükemmel ortamda olduğundan daha az başarılıdır (Tablo 3.9). Sonuçlar, güven puanı mevcut olmadığında, bozuk örneklerin çok sınıflı kurbanları kandırmakta başarısız olduğunu göstermektedir.

Tablo 3. 9 Güven puanı kısıtında kategorizasyon için üretilen saldırı örneklerinin aktarılabilirliği

Kurban Modeller	AG News NB (T)	AG News NB (U)	AG News BiLSTM (T)	AG News BiLSTM (U)	DBPedia NB (T)	DBPedia NB (U)	DBPedia BiLSTM (T)	DBPedia BiLSTM (U)
Sınırsız (α=5.0)								
NB	0,01	0,04	0,00	0,01	0,08	0,11	0,07	0,08
BiLSTM	-0,01	-0,03	0,08	0,16	0,06	0,04	0,15	0,23
α-Kontrollü (α=1.0)								
NB	0,00	0,02	0,00	0,00	0,00	0,06	0,00	-0,01
BiLSTM	-0,01	-0,03	0,02	0,05	0,00	0,00	0,07	0,11

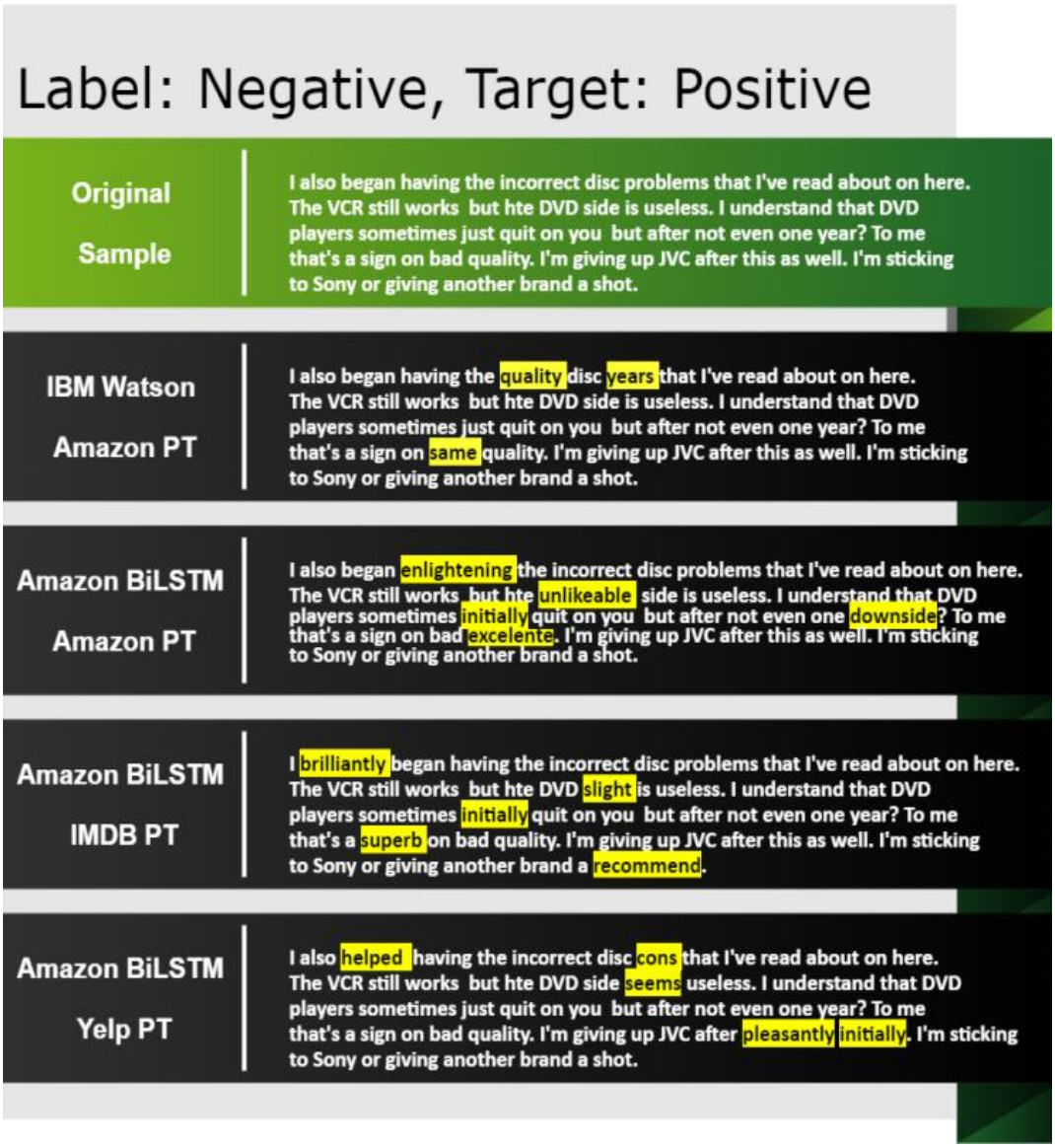
3.18 Gerçek Kara Kutu Saldırı

IBM Watson Doğal Dil Anlama hizmeti, hem etiketin hem de güven puanının döndürüldüğü belge düzeyinde duygu analizi sağlar. Tablo 3.10, bozuk örneklerin NB ve BiLSTM modelleri aracılığıyla aktarılabilirliğini gösterir.

Tablo 3. 10 Aracı model kullanarak IBM Watson'a saldırıda doğruluk farkı

Aracı Model	Kutupluluk Skoru Veri Kaynağı		
	P(Yelp)	P(Amazon)	P(IMDB)
Yelp NB	0,29	0,14	0,28
Yelp BiLsTM	0,18	0,10	0,19
Amazon NB	0,38	0,32	0,54
Amazon BiLSTM	0,26	0,13	0,25
IMDB NB	-0,02	0,16	0,30
IMDB BiLSTM	0,06	0,05	0,05

Algoritma 1, örnek başına 100'lük bir sorgu sınırı ek kısıtlamasıyla kullanılmıştır. δ=5, α=5.0 ve θ=0.51 parametreleri ile 10 seçilmiş örnek için δ kontrollü saldırılar gerçekleştirilmiştir. IBM Watson ve diğer deneysel modellerimiz tarafından bozulan bir örnek Şekil 3.16'da gösterilmektedir. Saldırı parametreleri tüm örneklerde aynıdır. Yelp kutupluluk tablosu tarafından bozulan örnek en gürültülü olanıdır. Bu sonuç, NB ve BiLSTM ile gerçekleştirdiğimiz deney sonuçları ile tutarlıdır. Üretilen diğer örnekler için, değişiklikler insan gözüne takılabilir, ancak duygu olumsuza dönüşmez. Altı katılımcı örnekleri etiketlemiştir: üç değerlendirici 10 orijinal örneği, diğer üçü IBM Watson NLU hizmetini sorgulayarak bozulan 10 saldırı örneğini etiketledi. Katılımcılara çalışma hakkında bilgi verilmeden, sadece örnekleri 'negatif' veya 'pozitif' olarak etiketlemeleri ve dikkatlerini çeken herhangi bir durumu belirtmeleri istendi. Orijinal örnek katılımcıları, orijinal etiketlerin ikisine aykırı etiketleme yapmıştır. Bazı örnekler için ikili sınıflandırmanın zor olduğunu ve bazı dilbilgisi hataları geri dönüşünü yapmışlardır. Saldırı örneği katılımcıları, 9/10 numuneye temiz numune değerlendiricileri ile aynı etiketleri atamıştır. İlginç bir şekilde, saldırı örneği değerlendiricileri de dilbilgisi ve ikili olmayan örnekler konusunda geri bildirimde bulunmuştur. Genel olarak, değerlendiriciler çoğunlukla iki grup örnek için de aynı etiketleri seçmiştir.



Şekil 3. 16 Farklı kutupluluk tablolarıyla IBM Watson servisi sorgulanarak elde edilen bir örnek

4
KONUŞMA SINIFLANDIRICILARINA KARŞI PERDE BOZULMASI TABANLI SALDIRI

Konuşma duygu tanıma, yalnızca otomatik konuşma işlemenin zor bir görevi değil, aynı zamanda pertürbasyon saldırıları alanında daha az incelenmiş olan bir alandır. Konuşma duygu sınıflandırıcılarına saldırmak için genetik bir çekişmeli örnek hazırlama yöntemi önerilmiştir. İki tür pertürbasyon analiz edilmiştir: beyaz-gürültü ve perde pertürbasyonu. Bir kara kutu ortamında iki kurbana saldırı gerçekleştirilmiştir: önceden işlenmiş özniteliklere sahip çok katmanlı bir sınıflandırıcı modeli ve bir dalga boyu verisini girdi alan evrişimli sinir ağı modeli. Öznitelikler küçük beyaz gürültüye karşı gürbüz olduğundan, çok katmanlı algılayıcı beyaz gürültü saldırılarına karşı da daha gürbüz olduğu görülmüştür. Evrişimli sinir ağı kurbanı hem beyaz gürültüye hem de perde bozulmalarına karşı savunmasız olduğu görülmüştür. Alt bölümlerde akustik ve algısal özellikler açısından bozulmaların kapsamlı bir analizini sunuyoruz.

4.1 Konuşma İşlemeye Genel Bakış

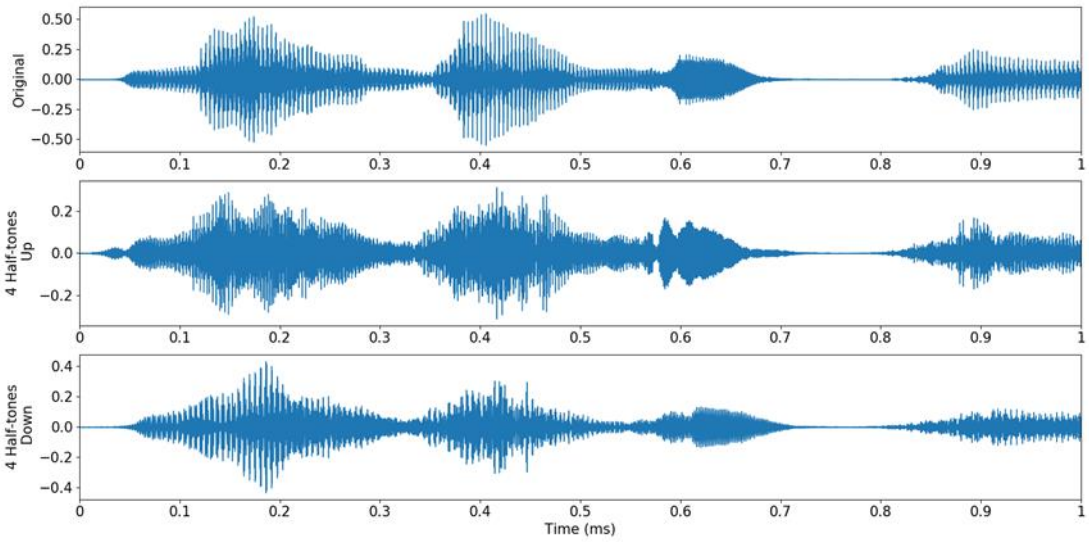
Mel ölçütü. İnsan işitsel sistemi, sesin havada oluşturduğu basınç dalgasının frekansını kulak zarındaki dalganın neden olduğu maksimum genliğin konumu üzerinden tespit eder. İnsan kulağı ses yoğunluğuna logaritmik olarak yanıt verir. "Kritik bantlar", insan kulağının bireysel frekans tonlarını ayırt etme yeteneğini ölçmek için kullanılır. Kritik bant filtrelerin tasarımı için Mel ölçütü ve Bark ölçütü kullanılır. Mel ve Bark ölçekleri insan işitsel sisteminin doğrusal olmayan doğasına dayanan, benzer şekilde, doğrusal olmayan ölçeklerdir. Ses ve konuşma işlemede en sık kullanılan Mel ölçütü perde (pitch), Bark ölçütü ise ses yüksekliği (loudness) algısını temel alır. Mel ölçütü kullanılarak üretilen filtreler ile çıkarılan öznitelikler, hem konuşma (örn. Ittichaichareon v.de diğ, 2012) hem de konuşmacı tanıma (örn. Martinez ve diğ, 2012) için ayırt edici niteliktedir.

Perde (pitch). Konuşmacının ses yolunun fiziksel özellikleri ile ilişkili olup, işitsel algıdaki en güçlü akustik özelliklerden biridir (Shen ve Souza, 2018) . Tonlama, ton,

vurgu ve ritim gibi konuşma tanımayı kolaylaştıran ve konuşmadaki duyguları ileten bilgi taşır. Perde kaydırma (pitch shifting), temel frekansın Eşitlik 4.1 ve 4.2'de gösterildiği üzere belirli bir faktörle ölçeklendirilmesinden oluşur. p yarı tonu, f temel frekansı (Hertz) ve s +/- yönde yarı ton kaydırma miktarını ifade eder. Şekil 4.1'de üst ve alt perdeye çekilmiş konuşma sinyali parçası verilmiştir.

$$p = 69 + 12 \times \log\left(\frac{f}{440}\right) \tag{4.1}$$

$$f_{son} = 2^{\left(\frac{s}{12}\right)} \times f_{ilk} \tag{4.2}$$



Şekil 4. 1 Orijinal konuşma sinyali (üst), 4 yarım-ton üst perde (orta) ve 4 yarım ton alt perde (alt)

Konuşmanın algılanması. Konuşmayı frekanslar olarak tanımlanabilir. Konuşmacılar sözcükleri telaffuz ederken ani başlangıçlar ve duruşlar, sessizlikler ve sesler meydana gelir. Konuşma, konuşmacıların sözcükleri sıralayıp, cümleler oluşturduğunda anlam kazanır. Konuşma algısı sadece fiziksel ses uyaranına değil aynı zamanda işitilenin yorumlanmasına yardımcı olan bilişsel süreçlere de bağlıdır. Örneğin, insanın dinlediği kalitesi düşük bir konuşma kaydında harf ya da kelime eksiklikleri olduğunda, bağlama göre bilişsel olarak tamamlama yapabilir. Otomatik konuşma sistemlerinin temelinde de bu mantık taklit edilmiştir.

Belirli seslerin algılanmasını/algılanmamasını etkileyen en önemli özelliklerinden birkaçı aşağıda sıralanmıştır:

- Ortamdaki diğer seslerin türü ve miktarı
- Sesteki gürültü miktarı
- İçerdiği frekanslar (çok düşük/yüksek perde)

Fonem ve fon. Konuşma sinyali yönetilebilir birimlere bölünerek işlenir. Metin temel alındığında anlamlı birimler cümle, kelime ve harf olarak sıralanabilir. Cümle, konuşma tanıma için çok büyük bir birimdir. Her dil için harf tek başına ses birimi olarak fazla bilgi taşımayabilir. Araştırmacıların büyük bölümü, konuşma algısını, "fon" (phon) olarak adlandırılan birimlerden oluşan “fonem”ler (phonem) üzerinde incelemektedir. Fon, kelimelerin anlamları için kritik olup olmadığına bakılmaksızın, farklı bir konuşma sesidir. Fonem, belirli bir dilde, başka bir fonemle değiştirildiğinde, kelimenin anlamını değiştirecek en kısa konuşma segmentidir.

Paralinguistik konuşma işleme. Paralinguistik konuşma inceleme, konuşmanın sinyalinin dilbilimsel ve konuşmacı kimliği içeriğinden farklı olan bilgileri çıkarmak için analiz edilmesini ifade eder. Bilişsel ve nörofizyolojik durumu incelemek üzere gerçekleştirilir. Ses perdesi paralinguistik sınıflandırma problemlerinde önemli bir öznitelik olarak kullanılır (Schuller ve diğ, 2013).

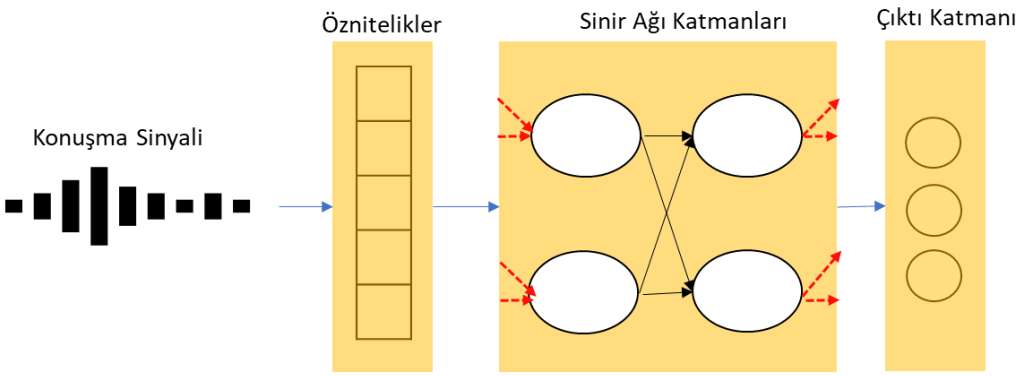
4.2 Konuşma Sınıflandırıcılar

Konuşma sınıflandırma için yapay sinir ağı tabanlı kurban modeller kullanılmıştır. Konuşma sinyalinin sayısallaştırılmış genlik gösterimine dalga formu denir. Çalışma boyunca, konuşma sinyali terimini dalga formu ile eşdeğer olarak kullanıyoruz.

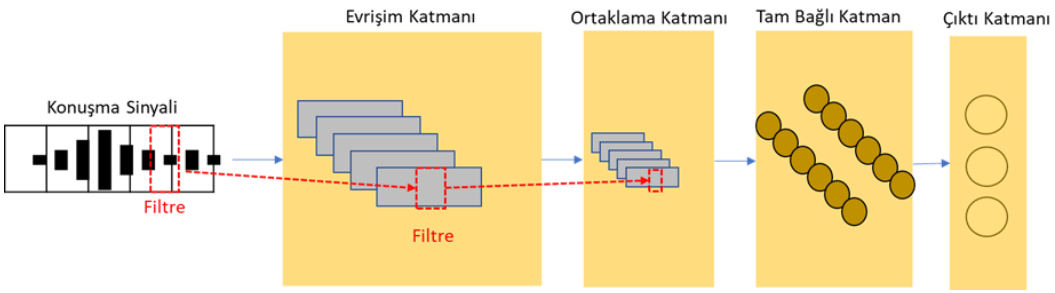
Fourier dönüşümü tabanlı işlemler ve filtreler kullanılarak konuşma sinyalinden öznitelikler çıkartılarak yapay öğrenme algoritmalarına girdi oluşturulur. Bu çalışmada çok katmanlı algılayıcı (MLP, Multi Layer Perceptron) ve evrişimsel ağ mimarisi (CNN, Convolutional Neural Network) kullanılmıştır.

MLP, önceden hazırlanmış öznitelikleri girdi alırken, CNN konuşma sinyalini herhangi bir işleme almadan girdi alır. CNN mimarileri evrişim (convolution), ortaklama (pooling) ve tam bağlantı (fully connected) katmanlarının birbirine bağlanması ile oluşturulur. Dalga formunu girdi olarak alan evrişimsel model, filtrelemeyi evrişim katmanlarında gerçekleştirerek kendi öznitelikleri öğrenerek

sınıflandırma yapar. Ortaklama katmanında boyut azaltma, tam bağlantı katmanında ise önceki katmanlardan gelen bilgi toplanır. MLP ve CNN modelleri için temel blok gösterimler Şekil 4.2 ve 4.3'te verilmiştir.



Şekil 4. 2 Temel bir MLP mimarisi



Şekil 4. 3 Temel bir CNN mimarisi

4.3 Konuşma Sınıflandırıcılarına Saldırılar

Konuşma sinyali, konuşmacı kimliği, ses yolunun fiziksel özellikleri veya bir kişinin psikolojik durumu gibi fonetik yapı başta olmak üzere çok fazla bilgi taşır. Konuşma duygusu tanıma (Speech Emotion Recognition, SER), konuşma içeriği ya da konuşmacı kimliği ile ilgili olmadığı için paralinguistik bir görevdir. Belirli bir zaman çerçevesinde bir konuşmacının kişilik özelliklerini veya duygusal durumunu tespit etmeyi amaçlar. Duygu algılanan bir özellik olduğundan, özellikle zor bir problemdir. Dilbilimsel içerik duygu hakkında ipuçları verse de özellikle kinaye tespiti gibi zor dil problemlerinin varlığında güvenilir bir kaynak değildir. Bu nedenle, son çalışmalar daha çok doğal dil ve konuşma işleme tekniklerini

birleştirmeye odaklanmaktadır. Bu çalışmada, konuşma sinyalinde öfke, sakinlik ve mutluluk tespitinden kaçınılması incelenmiştir. Bu duygusal durumların tespiti, özellikle ruh sağlığının izlenmesi (örn., France ve diğ., 2000), insan davranışı analizi (örn., Kamaruddin ve Wahab, 2010) ve çağrı merkezi kalite güvencesi (örn., Bojanić ve diğ., 2020) gibi kaçınma saldırısının finansal ve toplumsal sonuçlara neden olabileceği uygulama alanlarında oldukça önemlidir. Sistemlere sahte saldırılar yapmak, sistemin zayıflığının ortaya çıkarılmasına yardımcı olur ve önleyici tedbirlerin uygulanması için önemli bir adımdır.

Görüntü işleme ve metin analizi uygulamalarında olduğu gibi, konuşma işleme alanındaki kaçınma saldırısı çalışmalarının büyük bir kısmı, beyaz kutu ortamında model gömme katmanı ve geri yayılım ağırlıkları kullanılarak gerçekleştirilmiştir (örn. Carlini ve Wagner, 2018; Kreuk ve diğ., 2018). Mevcut konuşma saldırısı araştırmalarının çoğu, konuşmadan-metne modellere odaklanır. Yalnızca birkaçı, mimari, hiper-parametreler ve eğitim verisi dahil olmak üzere model bilgilerinin saldırgan tarafından bilinmediği tamamen kara kutu sistemlerini hedeflemiştir. Kara kutu saldırıları, sistemin iç durumlarının erişilemez olduğunu varsayar. Tipik olarak, modelin tekrar tekrar sorgulanmasıyla saldırı gerçekleştirilirler. Bir kara kutu saldırısı meşru bir örnekle başlatılır, ona küçük varyasyonlar, yani pertürbasyon uygulanır ve ardından modelin yanıtı gözlemlenir. Saldırı hedefine ulaşılana kadar iterasyonlar boyunca ek pertürbasyon uygulanır. Saldırganın model hakkında herhangi bir bilgisi olmadığı için, pertürbasyon miktarı ve bozunmanın yönü doğrudan tahmin edilemez. Bu nedenle, saldırganın yalnızca girdiyi (konuşma sinyali) ve çıktıyı (karar ve/veya güven) kullanarak bir arama gerçekleştirmesi gerekir.

Latif ve diğ. (2018), örnekler için gürültü üretmek üzere Demand veri kümesinden (Thiemann ve diğ., 2013) üç gürültü sınıfını kullanarak, ikili sınıflandırma için (pozitif/negatif duygu) eğitilmiş kara kutu modeline saldırı geliştirmiştir. LSTM-RNN duygu sınıflandırıcının doğruluğunu için önemli ölçüde azaltmayı başarmıştır. Ren ve diğ. (2020b), birden fazla hedef modele saldırmak için bir CNN saldırgan modelinde yaşam boyu öğrenmeyi kullanarak, rakip örneklerin bir kara kutu

modeline aktarılabilirliğini iyileştirmiştir. Temel olarak, CNN saldırgan, birden çok kurban model için zararlı olan gürültü temsillerini öğrenmiştir.

Çalışmamızda, kara kutu SER modelleri üzerinde genellenebilir olup olmadıklarını incelemek için iki farklı pertürbasyon stratejisi kullanılmıştır. İlk strateji toplamsal beyaz gürültü, ikincisi ise prozodik gürültüdür. Spesifik olarak, Gauss gürültüsü ekleme ve perde kaydırma uygulanmıştır. İki sinir ağı kurban model incelenmiştir: el yapımı özelliklerin bir kombinasyonuna sahip MLP ve doğrudan dalga girdili CNN. Eklemeli gürültü ve perde gürültüsünden oluşan iki ayrı mutasyon opsiyonuyla bir genetik algoritma yaklaşımı uygulanmıştır. Ardından kurbanların bu tür saldırılara karşı duyarlılıkları analiz edilmiştir. Başlıca katkılar aşağıdaki gibidir:

• Bu çalışma, çekişmeli örnek oluşturmada yalnızca toplamsal gürültünün kullanıldığı önceki çalışmalara göre bir iyileştirmedir.

• Beyaz gürültü ve perde mutasyon stratejilerine sahip bir genetik algoritma yaklaşımının çok az iterasyonda başarılı saldırı örnekleri üretebileceği gösterilmiştir. Bu, kötü niyetli üçüncü parti uygulamaların kara kutu ortamda hızlı saldırılar gerçekleştirmesini sağlayabilir.

• Sonuçlar, perde mutasyon stratejisinin beyaz gürültü seçeneğinden daha genelleştirilmiş bir saldırı olduğunu göstermiştir.

• İnsan işitme değerlendirmesi, orijinal ve hazırlanmış ses örnekleri arasında duygu algısında çok az değişiklik olduğunu ortaya koymuştur.

4.4 Veri Kümesi ve Kurban Modeller

RAVDESS (Livingstone ve Russo, 2018) ses veri seti, 24 aktörden (12 erkek ve 12 kadın) toplanan sekiz duygu kategorisi (nötr, sakin, mutlu, üzgün, sinirli, korkulu, iğrenme, şaşırmış) içermektedir. Bu veri seti iki sınıflandırıcıyı eğitmek için kullanılmıştır. Rastgele bir erkek ve bir kadın konuşmacıyı eğitimin dışında bırakılmıştır. Bu veri setinde İngilizce iki sözlü ifade vardır, Türkçe çevirisi ile: "Çocuklar kapının yanında konuşuyor" ve "Köpekler kapının yanında oturuyor". Fonetik çeşitliliğin olmaması, fonetik etkileri analiz etmeye gerek kalmadan görevin tamamen dil ötesi bir düzeyde incelenmesine yardımcı olmuştur.

Üç katmanlı MLP ve dört evrişim katmanına sahip bir CNN kurban modeller kullanılmıştır. MLP modeli için Issa ve diğ. (2020) tarafından önerilen özellikleri kullanılmıştır: Mel-frekans cepstral katsayıları (MFCC'ler), Chroma (Ellis, 2007), Mel Özellikleri, Spektral Kontrast ve Tonnetz (Harte ve diğ., 2006). MFCC ve Mel özellikleri, insan işitmesinin algısal özellikleri olan Mel filtresinden türetilen algısal bileşenlerdir. Chroma özellikleri perde sınıfı profillerdir, Tonnetz ton boşluğunu modeller ve Spektral Kontrast spektrumdaki tepeleri ve vadileri modeller. Öznitelik çıkarımı için 2048 uzunluğunda FFT penceresi ve girdi olarak çerçeveler üzerindeki öznitelik ortalaması kullanılmıştır. Daha sonra özellikler, örnek başına tek bir vektör oluşturacak şekilde birleştirilmiştir.

İkinci kurban, CNN, örnek düzeyinde ham dalga formu girişi olan bir modeldir. Model, temel bir evrişim mimarisine ve ardından havuzlama ve ReLU aktivasyonlarına sahiptir. Son katman tamamen bağlantılı bir katmana ve Softmax'a sahiptir. Model, 128 filtreli 4 evrişim katmanından oluşmakta ve tek boyutlu evrişimlere ve havuz boyutunun 8 olduğu 5 filtre boyutuna sahiptir.

MLP sakin-nötr/mutlu/sinirli bir sınıflandırıcı olarak eğitilirken, CNN erkekler ve kadınlar için ayrı olmak üzere sakin/mutlu/üzgün/sinirli/korkulu konuşma duygularını ayırt etmek üzere eğitilmiştir (5x2=10 sınıf). İki model mimari ve girdi yapısı açısından farklıdır. Tek ortak payda, RAVDESS veri seti ile eğitilmeleridir.

Bu kurbanları seçmemizin nedeni, girdide (MLP kurbanında olduğu gibi) algısal ve akustik bilgilerin olmasının ya da (CNN kurbanında olduğu gibi) olmamasının analiz etmektir. Başka bir deyişle, sisteme verilen ilk bilgilerle saldırıları başarısının değişip değişmediği, değişti ise bunun yönü araştırılması amaçlanmıştır.

Saldırı algoritmasının eğitimden bağımsız bir analizini yapmak için, dışarıda bırakılan örnekler (3x2 = 6 örnek) arasında erkek ve kadının öfke, sakinlik ve mutluluk için örnekler alınmıştır. Seçilen örnekler, her iki model için de 0.90 ve üzeri gerçek sınıf güven oranına sahiptir.

4.5 Konuşma Sinyalinin Pertürbasyonu

Bir sesin bazı dilsel sembollerle eşleştirildiği konuşma tanıma görevlerinde, fonetik segmentler birincil analiz birimleridir. Bunlar, genellikle örtüşen bir şekilde

bölümlere ayrılmış, konuşmanın kısa süreli ses bölümleridir (örneğin 25 ms). Buna karşılık, prozodi daha büyük segmentlerde gözlenir ve konuşmacının durumu ve anlatımla ilgili özellikleri hakkında bilgi taşır. Temel frekans (F0), konuşma prozodisinin akustik bir özelliğidir. Konuşmanın F0'ı, esas olarak konuşmacının ses tellerinin rezonans frekansı tarafından belirlenen fiziksel bir olgudur. Sesli konuşma sinyallerinde saniyedeki ortalama dalgalı salınım sayısıdır. F0, konuşmacının fizyolojisine ve konuşulan dilin doğal konuşma düzenine bağlıyken, perde algılanan periyodikliktir. Konuşma çerçevelerindeki perde konturu, duygusal durum hakkında önemli bilgiler sağlar; düşük perde, üzüntü gibi orta düzeyde aktivite duygularını belirtirken, yüksek perde, yüksek seviyeli duyguları ifade eder.

Konuşmanın prosodik özellikleri, duyguları ayırt etmede insan algısı için önemlidir (Tóth ve diğ., 2008) ve konuşmanın duygusu, perde konturu kalıplarıyla oldukça ilgilidir. Pierre-Yves (2003) tarafından yürütülen duygusal konuşma sentezi deneyleri, perde özellikleriyle ilgili olarak duygular yoluyla önemli farklılıklar olduğunu ortaya koymuştur. Hem mutluluk hem de öfke için perde ortalaması ve varyansının yüksek olduğu; bununla birlikte, hecelerdeki perde hatlarının yönü, mutluluk için yükseldiği ve öfke için düştüğü raporlanmıştır. Aynı çalışmada sakinliğin, yükselen kontur ve daha düşük hacim ile perdede düşük ortalama ve varyansa sahip olduğu belirtilmektedir.

Saldırı örneklerinin genetik nesli, çaprazlama işlemini iki ebeveynin noktasal ağırlıklı bir ürün toplamı olarak uygulanmıştır. Mutasyon durumunda, iki ayrı pertürbasyon şeması kullandık. İlki, çocuğa az miktarda beyaz gürültünün δ_W eklendiği toplamsal gürültüdür (Eşitlik 4.3). İkinci şema, perdenin mutasyonudur. Adım kayması, ilk önce STFT penceresini faktör Eşitlik 4.4 ile hareket ettirmek suretiyle sinyali zaman ekseninde "germe" işleminin ardından FFT ile orijinal örnekleme hızına yeniden örnekleyerek gerçekleştirilir (Robel, 2006). Bu şemada, perde 12 oktav aşağıya ($\delta_p = -1$) veya yukarıya ($\delta_p = 1$) kaydırılır, burada ortaya çıkan sinyal uzunluğu zaman ekseninde değişmez (Eşitlik 4.5).

$$x' = x + \delta_w \tag{4.3}$$

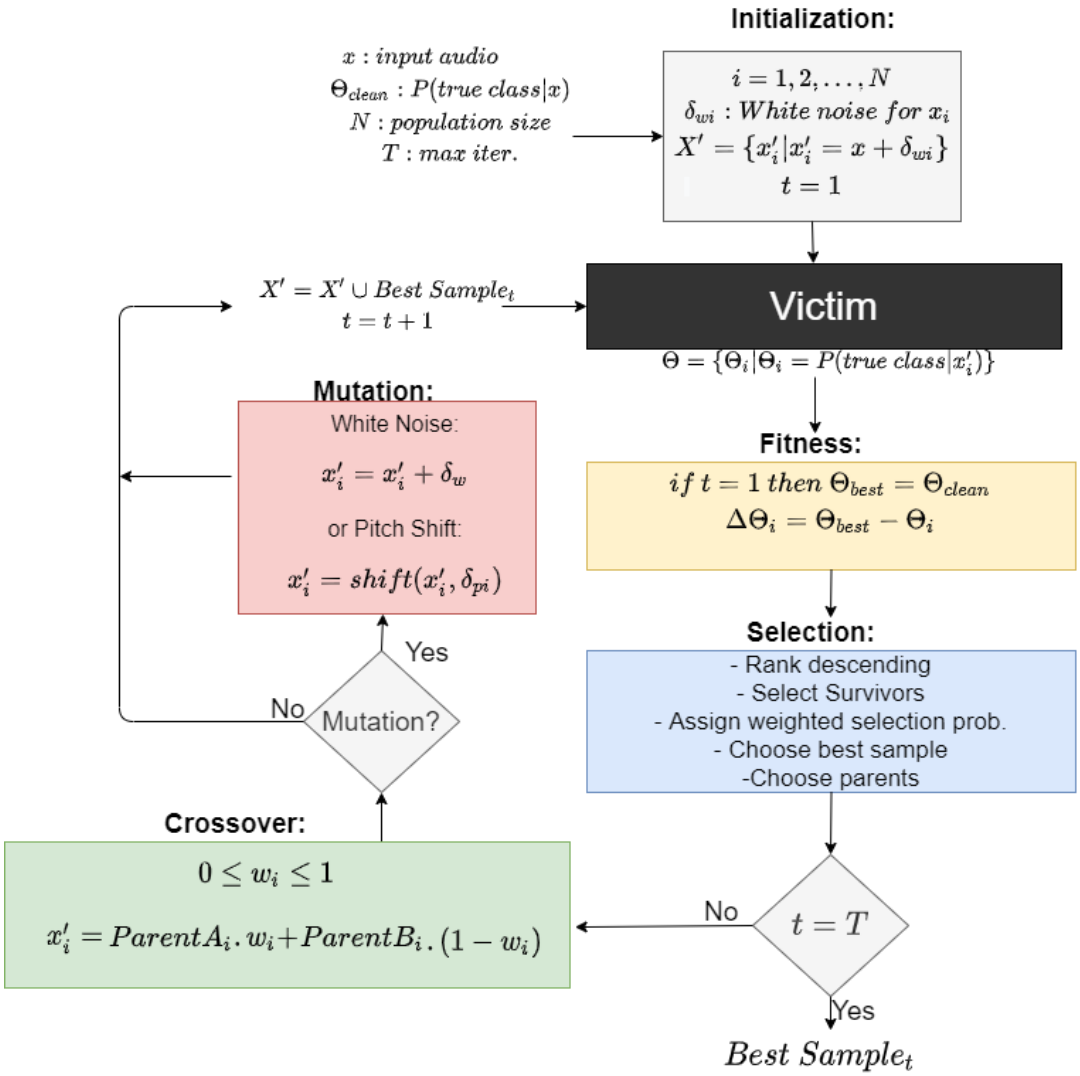
$$\frac{sample\ rate}{2^{-\delta_p/12}} \tag{4.4}$$

$$x' = shift(x, \delta_p) \tag{4.5}$$

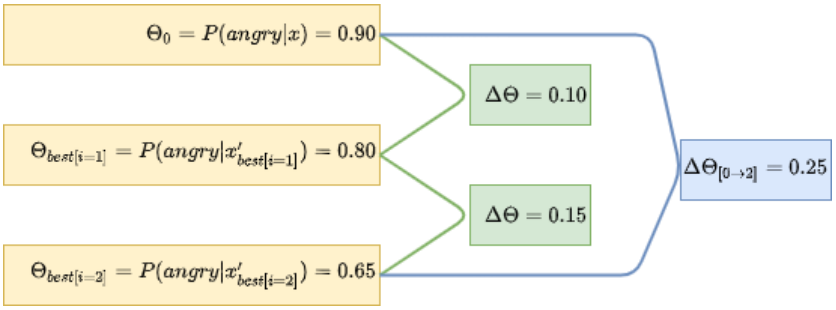
4.6 Perde Tabanlı Saldırı

Düşmanın amacı, SER sistemlerinde sinirli ve mutlu duyguların algılanmasından kaçınmaktır. Tehdit modelimiz, popülasyon oluşturma → uygunluk değerlendirmesi → seçim → çaprazlama → mutasyon → birkaç adımda ince ayarlarla sonlandırma kontrolünün standart bir genetik algoritma döngüsünü takip eder (Şekil 4.4). Popülasyonun 100 üyesi vardır ve ilk popülasyon kompozisyonu küçük ek beyaz gürültü ile gerçekleştirilir. Daha sonraki iterasyonlarda, ağırlıklı noktasal çaprazlama ve 0.1 olasılıkla mutasyon kullanılır.

Gerçek sınıf güvenini P(gerçek sınıf $\mid$ w),olmak üzere, w bir konuşma sinyali girişidir, kısaca θ ile gösterir. Uygunluk puanlarını Eşitlik 4.5'e göre değerlendirmek için, orijinal örneğin gerçek sınıf güveni ile oluşturulan popülasyondaki örnekler arasındaki farkın sonuçları azalan yönde sıralanır. Ardından, negatif değerler filtrelenir. Sistem büyük pertürbasyonlarla kandırılmak istenmediğinden, sıralanmış popülasyonun orta üyelerine daha yüksek ağırlıklı ebeveynler olarak seçilmek için rastgele ağırlıklı bir olasılık atanmıştır. Bu düzenleme Bölüm 5'te kullanılan deneylerde gözlemlenen derin vadilerin önüne geçilmesini sağlamıştır. Genetik iterasyonlara elemelere rağmen en iyi örneklerin uygunluğu, Şekil 4.5'te gösterildiği gibi, maksimum iterasyon sayısı içindeki en iyi saldırı örneğini belirlemek için hesaplanmıştır.



Şekil 4. 4 Genetik algoritma tabanlı bozulma şeması



Şekil 4. 5 İterasyonlarda örnek uygunluğu seçimi

4.7 Saldırı Başarısının Değerlendirme Kriterleri

Beş iterasyon içinde orijinal sınıflandırıcı çıktısını değiştirebiliyorsa, rakip bir örneği "başarılı" olarak değerlendirilmiştir. Girişimi başarısız görsek bile başarıya ne kadar yakın olduğunu araştırmak isteriz. Bunu başarmanın en basit yolu, ilk yinelemedeki θ_0 temiz girdi ile son yinelemedeki θ_f saldırı girdisiyle gerçek sınıfın güven puanını karşılaştırmaktır (Eşitlik 4.6).

$$\Delta\theta_{[0\to f]} = \theta_0 - \theta_f \tag{4.6}$$

Ortalama karekök (RMS), sinyal işlemede yaygın olarak kullanılan bir ortalama alma tekniğidir. Bu çalışmada, örtüşen çerçeveler üzerinde orijinal ve hazırlanmış örnekler arasındaki bozulmayı gözlemlemek için bir genlik RMS grafiği kullanıyoruz. RMS grafiği, ham dalga biçiminden daha net bir sinyal bozulması görünümü sunar. Eşitlik 4.7, N=2048 (%25 örtüşme) çerçeveleme penceresine sahip bir çerçeve i üzerinde RMS'yi gösterir.

$$RMS_i = \sqrt{\frac{|frame_i|^2}{N}} \tag{4.7}$$

Spektral düzlük, bir sinyalin beyaz gürültüye ne kadar benzediğini gösterir. Değerler [0,1] aralığındadır ve 1.0'a ne kadar yakınsa beyaz gürültüye o kadar benzer. Ayrıca konuşma sinyallerindeki duygusal durumları belirlemek için kullanılır (Přibil ve Přibilová, 2009). Sonuç olarak, çekişmeli örnek değerlendirmesinin belirlenmesi için önemli bir ölçümdür. Bu, yalnızca ek beyaz gürültünün etkisini görmeyi değil, aynı zamanda duygu algısındaki değişimi nicel olarak karşılaştırmayı da amaçlayacaktır. Spektral düzlük Eşitlik 4.8'de verilmiştir, burada $|S_k|^2$ büyüklük spektrumu ve N_{FFT} ise FFT pencere boyutudur. Değerlendirmemizde N_{FFT} =2048 kullandık.

$$S_F = \frac{\exp\left[\frac{2}{N_{FFT}} \sum_{k=1}^{N_{FFT}/2} \ln |S_k|^2\right]}{\frac{2}{N_{FFT}} \sum_{k=1}^{N_{FFT}/2} |S_k|^2} \tag{4.8}$$

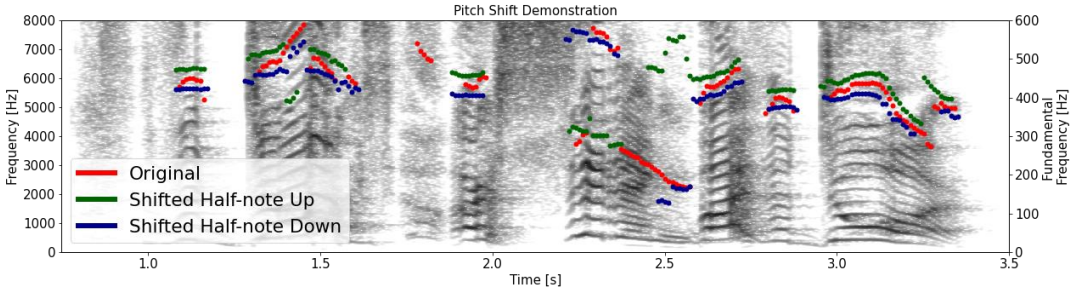
Orijinal ve rakip örnekler arasındaki ortalama perde (Hz), yerel titreşim ve yerel parıltı (dB) arasındaki farkları inceliyoruz. Titreşim, periyotlar arasındaki frekans değişimidir ve ışıltı, genlikteki değişimdir. Sıralı kareler arasında yerel titreşim ve yerel parıltı hesaplanır. Titreşim ve ışıltı, ses özelliklerini tanımlamada faydalıdır (Farrús ve diğ., 2007; Teixeira ve diğ., 2013) ve ses patolojilerini tanımlamada yaygın olarak kullanılırlar (örneğin, Teixeira ve Fernandes, 2014). Bu çalışmada, bozulmayı incelemek için orijinal ve rakip örnekler arasındaki ortalama perde, titreşim ve parıltının mutlak farkı kullanılmıştır.

Titreşim ve parıltı Eşitlik 4.9 ve 4.10'daki gibi hesaplanır; burada T_i temel frekans uzunlukları, A_i tepeler arasındaki genlik ve N çerçeve sayısıdır. Titreşim ve parıltının yerel bir ölçümü için, denklemler sırasıyla ortalama periyoda ve ortalama genliğe bölünmüştür.

$$Titreşim = \frac{1}{N-1}\sum_{i=1}^{N-1}|T_i - T_{i+1}| \tag{4.9}$$

$$Parıltı = \frac{1}{N-1}\sum_{i=1}^{N-1}|20\log(A_{i+1} - A_i)| \tag{4.10}$$

Perde hatlarının görsel özellikleri, perde konturundaki değişikliği göstermek için sonuçlarda da gösterilmektedir. Perde izleri, orijinal spektrogram üzerinde ayarlanmış perde izlerinin bir gösteriminin Şekil 4.6'te gösterildiği praat-parselmouth ile hesaplanmıştır.



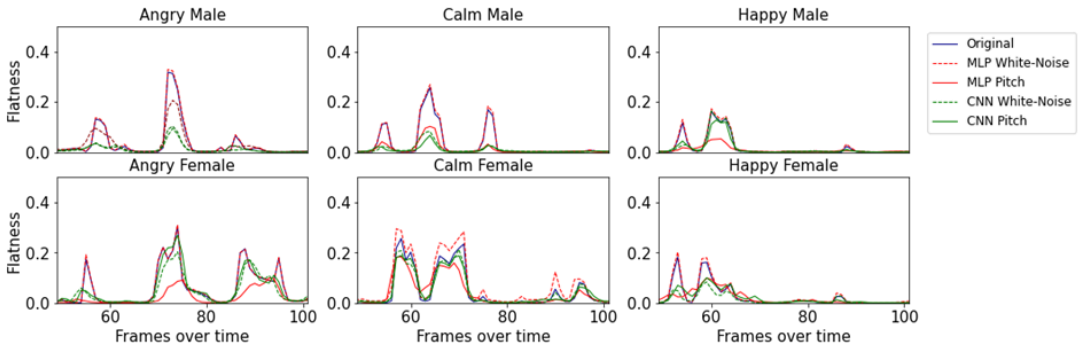
Şekil 4. 6 Perde kontür izleri

4.8 Deneysel Sonuçlar

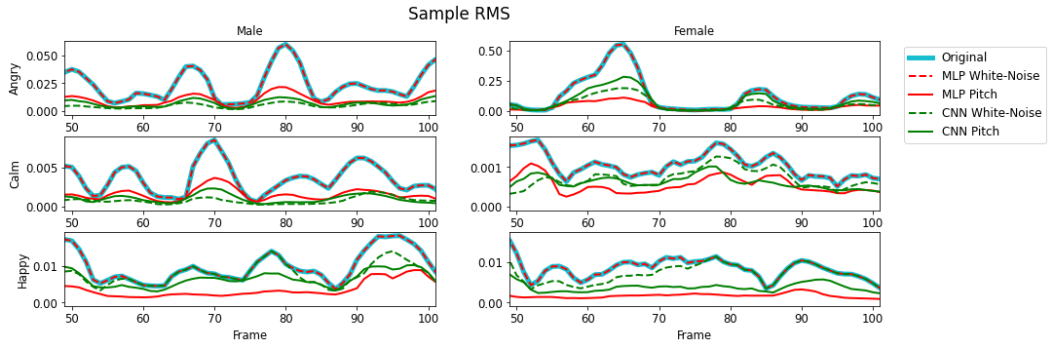
Saldırgan örneklerin genetik üretimimizde, ilk popülasyonun rastgele Gauss gürültüsü eklenerek üretildiği 100'lük bir popülasyon boyutu kullanılmıştır. Deneylerimiz boyunca Gauss gürültüsü için parametreler $\mu = 0$ ve $\sigma = 10^{-5}$ idi. İki SER kurbanı (MLP ve CNN) üzerinde beş yineleme için saldırı yöntemi de (beyaz gürültü ve perde mutasyonu) çalıştırılmıştır. Bu ortamda, hazırlanmış örnek başına maksimum istek sayısı 500 ile sınırlandırılmıştır. Pertürbasyonun türünü, miktarını ve saldırı başarısını analiz edilmiştir.

RMS'nin genliği, çerçeveleri üzerinden sesin ortalamasını gösterdiğinden, temiz ve rakip örnek arasındaki farkı tasvir etmede ham dalga biçiminden görsel olarak daha iyidir. Şekil 4.7'de, MLP modelindeki beyaz gürültü mutasyonlu örnekler, orijinal örneklerin genel şeklini korurken, iki kurbandaki perde mutasyonlarının daha büyük bir farka neden olduğu görülmektedir. Görsel netlik için grafiklerden çıkardığımız her sesin başında ve sonunda büyük seslendirilmemiş bölümler vardır. Aslında duygu durumu ancak seslendirilen bölümlerden belirlenebilir. CNN kurbanı üzerindeki beyaz gürültü mutasyona uğramış örnekler, sakin ve öfkeli erkek örnekleri dışında, orijinal dalga biçimine perde karşılığına göre daha yakındır. Perde mutasyonlarının daha "gürültülü" pertürbasyonlar olduğu açıktır; bununla birlikte, bozulma ölçeği hala düşüktür. Diğer beş örneğe göre en yüksek perde ve hacme sahip olduğundan beklenen genel bozulma, sinirli kadın sesi örneğinde (Şekil 6, sol) en yüksektir. Bu o kadar güçlü bir öfke ifadesi ki, fark edilmemek için daha yüksek derecede pertürbasyon ihtiyacı oluşmuştur.

Genlik RMS'nin hazırlanmış numuneler üzerindeki genel bozulma seviyesini yansıttığı durumlarda, spektral düzlük, bu dalgalanmaların sinyali daha rastgele (veya gürültü benzeri) yapıp yapmadığını gösterir. Şekil 4.7'de genel olarak pertürbasyonların düzlüğü artırmadığı görülmektedir. Genel olarak, MLP sistemi üzerindeki beyaz gürültü mutasyonu düzlüğü arttırır. Spesifik olarak, sakin kadın ve mutlu kadın sesi örnekleri, beyaz gürültü pertürbasyonları ile spektrumda daha düz hale geldi. Bu bulgunun ortalama düzlük değerlerinde izlenmesi (Şekil 4.8, sağ), bu örnekler için orijinal düzlük zaten yüksektir; bu nedenle, bozulmalar düzlüğe daha fazla katkıda bulunmuştur.

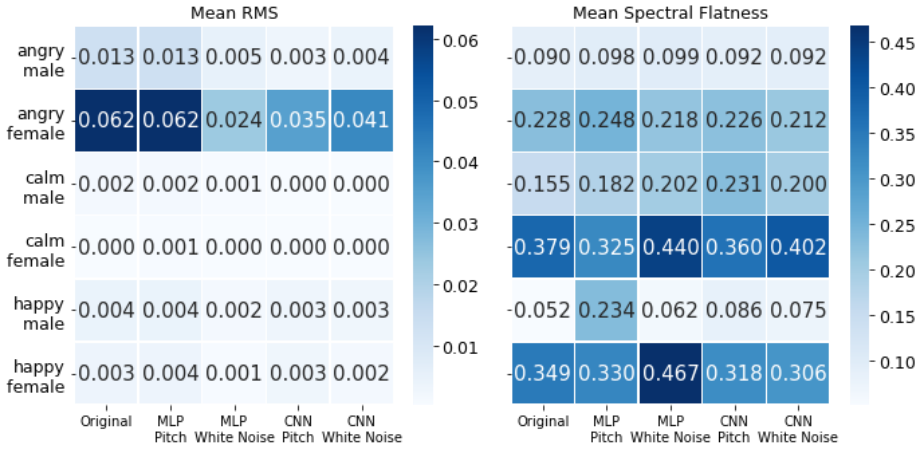


Şekil 4. 7 Saldırı örnekleri için spectral düzlük karşılaştırması

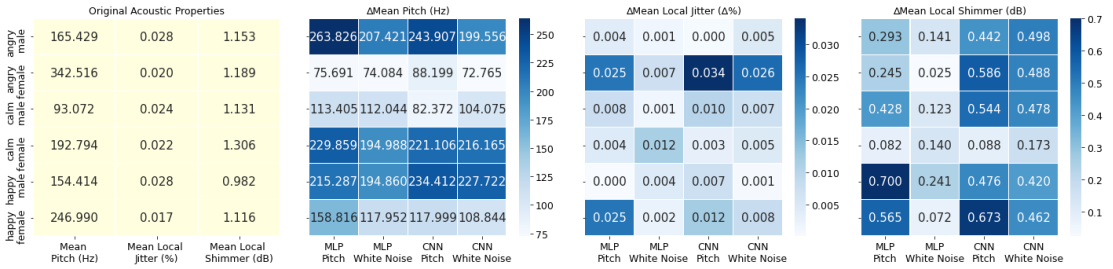


Şekil 4. 8 Saldırı örnekleri için RMS karşılaştırması

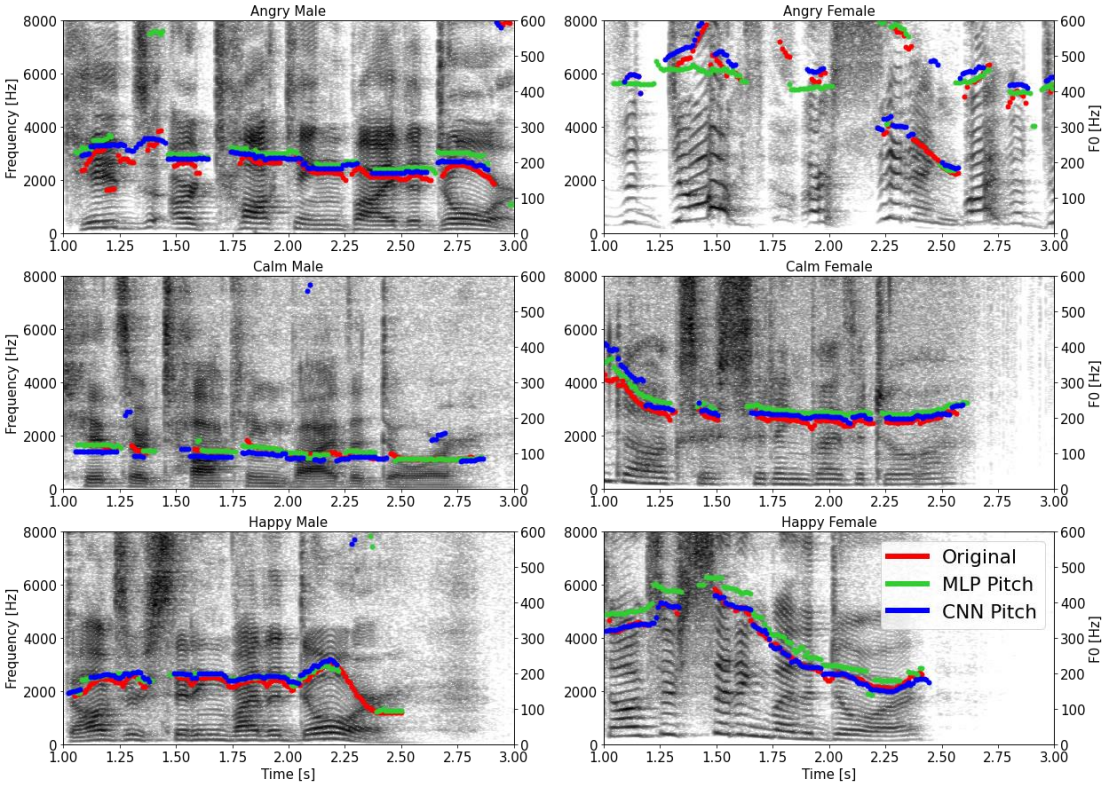
Akustik özelliklerdeki perde, titreşim ve parıltı üzerindeki mutlak ortalama değişim Şekil 4.9'da görülmektedir. Titreşim frekans değişimleridir; sinirli kadın ve mutlu kadın sesi örnekleri dışında genel değişim minimaldir. Sinirli kadın sesi örneğinin yüksek bir perdesi vardır ve mutlu kadın sesi örneği başlangıçta yüksek düzlüğü vardır; bu nedenle, örnekleri kaçırmaya doğru hareket ettirmek için hazırlanmış örneklerdeki periyodiklik farkının daha dramatik olması gerekiyordu. Işıltı, genlik değişimidir ve yaklaşık %50 mutlak değişim gözlemlenmiştir. İstisna, genlik RMS sonuçlarında belirtildiği gibi değişikliğin minimum olduğu MLP beyaz gürültü modelidir. Perdedeki mutlak fark, en yüksek sinirli ve mutlu erkek ses örneği ve sakin kadın ses örneği olmak üzere değişir. Perde farkını daha fazla araştırmak için orijinal spektrogramlar üzerinde perde izlerini çizilmiştir (Şekil 4.10). Spektrumun üzerindeki değişiklik, sinirli kadın ses örneği için çok büyüktü. Çok belirgin olmasa da, perde mutasyon şeması için perde izi farklılıkları, hazırlanmış ve orijinal örneklerin tümü için kolayca tanınabilir.



Şekil 4. 9 Akustik özelliklerdeki ortalama RMS ve ortalama spectral düzlükteki değişim

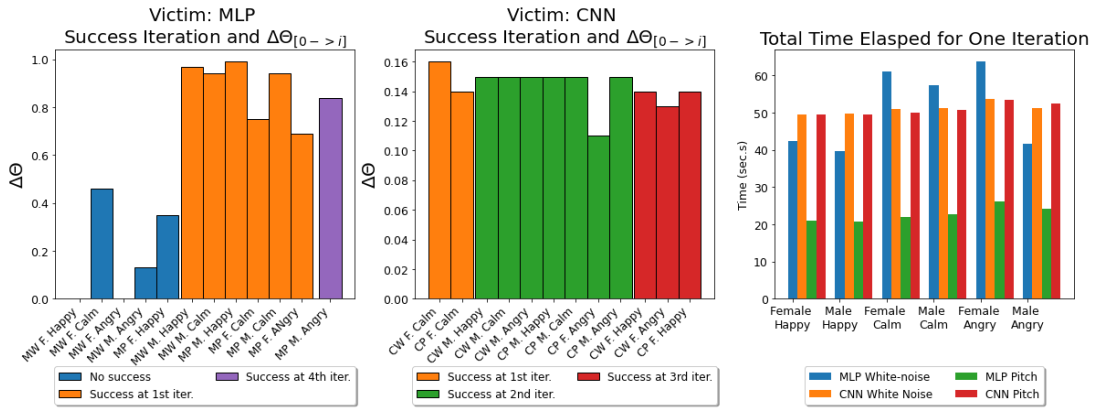


Şekil 4. 10 Akustik özelliklerdeki perde, titreşim ve parıltı üzerindeki mutlak ortalama değişim



Şekil 4. 11 Orijinal ve saldırı örnekleri için perde izleri

Daha önce, orijinal ve rakip örnekler arasındaki farkı çeşitli açılardan gözlemledik ve değişimin yüksek perdeli "sinirli kadın sesi örneği" ve daha düz "mutlu kadın sesi örneği" gibi güçlü ifadelerde daha belirgin olduğunu gösterdik. Bu sefer genel saldırı başarısını ve performansın ele alınmıştır. Her yinelemede en iyi hazırlanmış örneği topladık ve beş genetik yinelemede en erken başarıyı kaydedilmiştir (Şekil 4.12). MLP kurbanının 3 sınıflı bir ayrımcı olduğunu, CNN kurbanının ise 10 sınıflı bir problem olduğunu ve buna göre $\Delta\theta$ çubuklarını analiz etmemiz gerektiğini belirtmek önemlidir.



Şekil 4. 12 Saldırı başarıları ve süre

MLP kurbanına yapılan beyaz gürültü saldırıları, sakin erkek ve mutlu erkek örnekleri dışında başarısız oldu. Bu, MLP kurbanının genel olarak beyaz gürültü bozulmalarına karşı dayanıklı olduğunu gösterir. Bu belirli kurban modelinin girdi özellikleri, az miktarda ilave beyaz gürültüye karşı dayanıklı olduğundan, sonuç beklenen bir sonuçtur. Bir başka başarısız saldırı, MLP kurbanına karşı mutasyona uğramış mutlu kadın düşmandı. Bir miktar Δθ gözlemlenen (mavi çubuklar) üç başarısız saldırı girişimi için, güven yönünün başarıya doğru pozitif olduğunu fark ettik ve büyük olasılıkla aşağıdaki yinelemelerde başarılı oldular. 24 örnekten 5'i başarılı olmadığı için, her iki kurban modeli üzerindeki toplam kandırma oranı %79.16'dır (Tablo 4.1). Saldırıların çoğu, birinci veya ikinci yinelemelerde başarılı olmuştur. Genel olarak, CNN'ye yapılan saldırılar, CNN kurbanı için saldırı başarısı yinelemesi açısından beyaz gürültü veya perde mutasyonu arasında fazla bir fark olmadığında daha fazla yinelemeye ihtiyaç duyuyordu. MLP kurbanına yönelik öfkeli erkek saldırısının perde pertürbasyonunun, oluşturulan popülasyonun dördüncü yinelemeye kadar uygunluk açısından sıkışıp kaldığı aykırı davranış gösterdiğini düşünüyoruz.

Tablo 4. 1 Saldırı başarısı

Kurban Model	**Mutasyon**	**#Başarılı Saldırı / #Toplam Saldırı**
MLP	White-Noise	2/6
	Pitch Shift	5/6
CNN	White-Noise	6/6
	Pitch Shift	6/6
Toplam Kandırma Oranı		***19/24=79.16%***

Saldırı örneğinin üretilme süresi, kara kutunun sorgulanma sayısı kadar önemli bir husustur. Şekil 4.12'de (sağda), örnekler üzerinde bir yineleme için geçen ortalama süre, en çok zaman alan çekişmeli örnek oluşturma işleminin yineleme başına 60 saniyede gerçekleştirildiğini ve bunun başarısız bir girişim olduğunu göstermektedir. Başka bir deyişle, daha zor bir problemdi. MLP kurbanı için üretim için harcanan zaman, beyaz gürültü ve perde mutasyonları için aşırı bir farka sahiptir. Daha önce de belirtildiği gibi, MLP kurbanı, beyaz gürültü sağlam giriş özelliklerinin bir sonucu olarak beyaz gürültü saldırılarına karşı daha dayanıklıdır. Böylece, modele karşı perde pertürbasyonlu örneklerin oluşturulması daha kolaydır. Öte yandan, CNN kurbanına karşı bir saldırı türünün diğerinden daha kolay olup olmadığı sonucuna varılamaz.

İnsan değerlendirmelerinde, dört pertürbasyon varyasyonu üzerinde altı örneğin dört farklı şekilde (iki model x iki pertürbasyon yöntemi) bozulması ile elde edilen saldırı örnekleri ve orijinal altı örneği (değerlendirici başına 6x4+6=30 numune) kullandık. Örnekleri karışık sırayla sunduk ve 10 değerlendiriciden dinleyerek algıladıkları kapalı duyguyu mutlu/sakin/sinirli seçenekleri arasından seçmelerini istedik. Her değerlendiricinin temel gerçeği için orijinal örnekler için verilen cevapları kullandık. Daha spesifik olarak, 10 değerlendiricinin her biri 24 çekişmeli ve 6 orijinal numuneyi etiketledi, burada orijinal numune etiketlemesi her bir kişi için temel gerçekti. Sonuç olarak, çekişmeli örnekler için toplam 240 etiketleme yapılmıştır. İnsan değerlendirmesinin genel aldatma oranı %8.34'tür (Tablo 4.2). Değerlendiriciler, bazı örneklerin "kayıt kalitesinin daha düşük" olduğunu veya seçeneklerin daha çeşitli olmasını istediklerini, ancak bunun dışında herhangi bir şüphe belirtmemişlerdir.

Tablo 4. 2 Dinleme değerlendirmeleri sonucu

#Kandırılan	**0**	**1**	**2**	**5**	**Toplam**
Frekans (#İnsan Değerlendiriciler)	2	1	6	1	***20***
Kandırma Oranı	$(20\ kandırılan)/(24\ düşman\ örnek \times 10\ katılımcı)$				***8.34%***

5
İKİ-KİPLİ SINIFLANDIRICIDA PERTÜRBASYON SALDIRISI

İki-kipli sözlü diyalog sistemleri dil bilimsel ve akustik birimlerden oluşur. Kötü niyetli örnek saldırıları olarak da ifade edilebilen başarılı kaçınma saldırıları sonucunda sistem bütünlüğü ve/veya kullanılabilirliği azalır. Kara kutu saldırılarında, model ile ilgili bilgi ya da eğitim kümesi hakkında bilgi gerekmeden, sistem sorgulanmak suretiyle saldırı örnekleri elde edilir. Bu bölümde, Bölüm 3'te gerçekleştirilen metin pertürbasyon yöntemi genişletilmiş, metin ve konuşma sınıflandırıcına bütünleşik saldırı yöntemleri çeşitli stratejiler altında incelenmiştir.

5.1 Konuşma Sentezleme Yaklaşımı

Metin ve konuşma sinyali veri kümesi kullanılarak, veri kümesinde bulunmayan bir metinden konuşma sinyali elde etme konuşma sentezleme probleminin genel çerçevesini tanımlar. Bu problem Eşitlik 5.1 ve 5.2 ile ifade edilebilir (Tokuda, 2019). $w \notin W$ olmak üzere metin-konuşma veri kümesi (W, X) konuşma modelinin hesaplanan parametreleri $\hat{\lambda}$ kullanılarak (eğitim) x metni sentezlenir.

$$\hat{\lambda} = \underset{\lambda}{\operatorname{argmax}}\, p(\hat{\lambda}|W, X) \tag{5.1}$$

$$x \sim p(x|w, \hat{\lambda}) \tag{5.2}$$

En temel sentez modeli, tekrarlayan bir dalga formu üreten bir osilatördür. Çoğu sentez modeli, başlangıç noktası olarak bir veya daha fazla osilatör kullanır. İstenen tını elde edilene kadar sinüs dalgası kısmi değerlerini kendi genliklerinde bir araya getirmek mümkündür. Teoride, yeterli kısmi kullanılırsa, herhangi bir sesin tam eşleşmesi sağlanabilir. Gerçek dünya tonları zaman içinde harmonik içeriklerini değiştirme eğiliminde olduğundan, her bir parça kendi genlik zarfına sahip olabilir. Herhangi bir ses bu şekilde temsil edilebilir (McAulay ve Quatieri, 1995). Buna alternatif olarak, karmaşık bir dalga formu ile başlayıp ve daha sonra bazı kısımlarının çıkarılması için filtrelenebilir. Genelde iki yöntemi birlikte kullanmak

avantajlıdır. Bir yandan karmaşık dalga formu üretilirken, diğer yandan istenmeyen frekans içeriğini silen/bastıran filtreler kullanılır.

Ses perdesi, yüksekliği (loudness) ve MFCC gibi özellikler genellikle konuşma sentezleyicileri kontrol etmek için kullanılırken (Örn. Chazan ve diğ., 2000; Shao ve Milner, 2004), Dinamik Zaman Bükümü (Dynamic Time Warping, DTW) yöntemi uzunluklarının birbirine eşit olması şartı aranmaksızın iki konuşma sinyalinin benzerliğini ölçmek için kullanılır (Amin ve Mahmood, 2008).

5.2 Sözlü Diyalog Yönetim Sistemleri

Sözlü diyalog yönetim sistemleri (spoken dialog management systems), içinde doğal dil anlama ve ses işleme modüllerinin belli bir iş akışı içinde, tanımlanmış bir görevi yerine getirirler. Bu çalışma kapsamında sözlü diyalog yönetim sistemlerini, insan-makine ya da insan-insan arası sözlü iletişim kayıtlarının girdi alınarak oluşturulan çıkarımların kişi ya da kuruluşların fiziksel eylemlerine etki ettiği ekosistem olarak tanımlanmıştır.

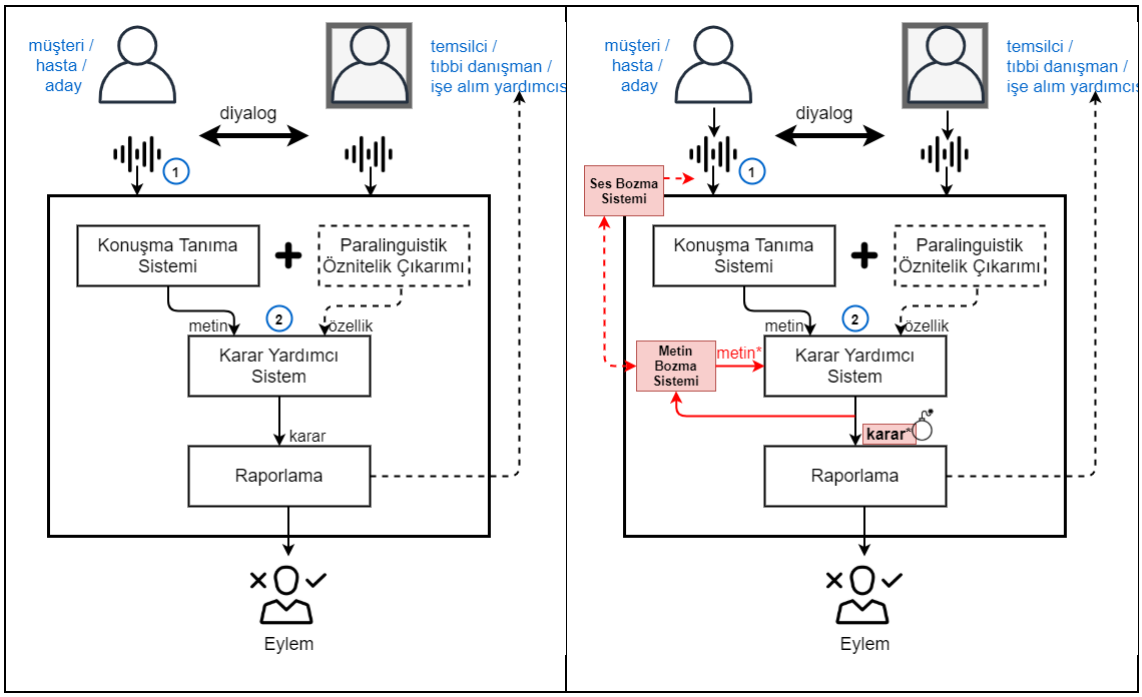
Gerçek kişi ya da sanal asistanlarla diyaloglar işlenerek pazarlama birimleri, sağlık kurumları, işe alım birimleri gibi fiziksel ortamdaki eylemleri etkileyecek çıkarımlar oluşur. Bu sistemlerin ses işleme bileşenleri, çevresel ve kullanıcı kaynaklı varyasyonlar nedeniyle meydana gelen konuşma sinyalindeki değişimlere karşı oldukça hassastır. Sistem bütünlüğünün korunması üzere, bu sistemlerin gürbüzlüğünü tehdit eden kötü niyetli müdahaleleri araştırmak son derece önemlidir.

Kötü niyetli örnek saldırıları sistem bütünlüğünü ve/veya kullanılabilirliğini azaltır. Kara kutu saldırılarında, model ile ilgili bilgi ya da eğitim kümesi hakkında bilgi gerekmeden, sistem sorgulanmak suretiyle saldırı örnekleri elde edilir.

Bu tür sistemlerin ortak modüllerini incelediğimizde, konuşma tanıma, metin işleme ve dil bilimsel olmayan (paralinguistics) sınıflandırma modellerinden oluşabilmektedir. İncelediğimiz patentlerin (Ashoori ve diğ., 2020; Hernandez, 2019) yapısına dayanarak Karar Yardımcı Sistemin şu modülleri içerdiğini varsayımında bulunduk: konuşma tanıma, metin sınıflandırma, ses duygu analizi. Bu

sistemlere çeşitli noktalarda saldırı yapılabilir. Bu çalışmada konuşma algılayıcı ve metin işleme noktalarına erişimimiz olduğu kabulü ile saldırılar düzenlenmiştir.

Sözlü diyalog yönetim sistemlerinin üç kullanım alanını (çağrı merkezi kalite değerlendirme, ruh ve sinir hastalıkları duygu örüntüleri tanıma, işe alım aday görüşme değerlendirme) değerlendirilmek suretiyle, bağımsız kara-kutu modülleri birleştirilerek temsili Şekil 5.1'de verilen diyalog sistemi oluşturulmuştur. Birleştirilen kara kutu modüleri Tablo 5.1'de gösterilmiştir. Özellikleri bilinen modüller ile ilgili bilgi saldırı formülasyonunda kullanılmamış olup, tüm modüller kara-kutu olarak ele alınmıştır. Sistem raporlama modülüne çıktılar kaydedildiğinde, orijinal örnekten farklı sonuçlar elde edildiğinde saldırı başarılı olarak tanımlanmıştır.



Şekil 5. 1 Sözlü diyalog sistemi ve saldırı noktaları

Tablo 5. 1 Sözlü diyalog sistemi modülleri

Modül	Kara Kutu Sistem
Konuşma Tanıma Sistemi	Google Cloud Speech-to-Text API[1]
Karar Yardımcı Sistem	İki kip: • Metin Sınıflandırma: BiLSTM Tabanlı Duygu Sınıflandırma • Akustik Sınıflandırma: CNN Tabanlı Duygu Sınıflandırma[2]
Raporlama	Birleştirilmiş" Karar Yardımcı Sistem" sonucu

Sistemde "sistem kullanıcısı-sistem operatörü" ikilisi, insan-insan ya da insan-makine olabilir. Diyalogları kayıt altına alınıp, bağlı modüllerde değerlendirilerek insan aksiyon alıcı (örn. yüksek düzey yönetici) ya da sistem operatörüne raporlanır. Örnek diyalog yönetim sistemi hem ses hem metin verisini değerlendirerek iki-kipli (two-modal) duygu analizi gerçekleştirir. Metin dilbilimsel olarak, konuşma sinyali de akustik özellikler olarak incelenir. Konuşma tanıma sistemi girdisi konuşma sinyali; çıktısı ise metin, sözcük zaman aralıkları ve sözcük tahmin güven değerleridir. Metin sınıflandırma, konuşma tanıma modülünden gelen metni girdi olarak alır ve iki sınıflı (pozitif/negatif) duygu analizi gerçekleştirir. Akustik sınıflandırma ise, konuşma sinyali girdisi beş duygu (sakin, mutlu, üzgün, sinirli, korkmuş) ve iki cinsiyet olmak üzere toplam on sınıfa (kadın, erkek) ayırır.

Şekil 5.1'de (1) ve (2) ile gösterilmiş iki saldırı noktası belirlenmiştir. Sözlü Diyalog Yönetim Sistemine entegre edilen üçüncü parti modüllerin bu noktaların birinden ya da her ikisinden sızarak sistem bütünlüğünü tehlikeye atması senaryosu incelenmiştir. Saldırı senaryosunda kullanılan modeller Tablo 5.1'de belirtilmiştir.

5.3 Konuşma Tanıma Sistemlerinde Kara-Kutu Saldırıları

Görüntü işleme ve metin analizi uygulamalarında olduğu gibi, ses ve konuşma işleme alanında kaçınma saldırısı çalışmalarının büyük kısmı beyaz kutu ortamında

[1] Google Cloud Speech-to-Text API, https://cloud.google.com/speech-to-text, Erişim: Haziran 2020

[2]Speech Emotion Analyzer, https://github.com/MITESHPUTHRANNEU/Speech-Emotion-Analyzer/, Erişim: Haziran 2020

model embedding katmanı ve geri yayılım ağırlıkları kullanılarak gerçekleştirilmiştir (Örn. Carlini ve Wagner, 2018; Kreuk ve diğ., 2018).

Mevcut kara-kutu saldırıları genetik algoritmalardan yararlanmaktadır. Wu ve diğ. (2019) DNN-HMM sistemine saldırı tasarlamış, HMM saklı katman geçiş bilgisini embedding görevinde kullanmıştır. Konuşma deneylerinin "yes, no, speech, attack" olmak üzere dört kısa komutla yapıldığı çalışmada, konuşma tanıma saldırılarının başarısı %20'de kalmıştır. Alzantot ve diğ. (2018a) da 10 farklı sözcükten oluşan kısa talimatlar veri kümesinde çalışmıştır. CNN üzerine yapılan genetik yarı kara-kutu saldırıda logit katmanlarına softmax uygulanmasıyla popülasyon uygunluk ölçümleri yapılmıştır. Katman bilgisine erişim, saldırının gerçek bir kara-kutu olmasının önüne geçmektedir. Modern uygulamalarda konuşma tanıma sistemleri çok büyük derlem üzerinde eğitilmiştir. Güncel kullanımda olan sistemlerde, kara-kutu olarak son katmanda derlemdeki tüm sözcükler ile ilgili çapraz-entropi ya da softmax değerleri sunulmamaktadır.

Uzun cümlelerde sözcük takasları gerçekleştiren bir genetik kara-kutu modeli olması nedeniyle Taori ve diğ.'nin (2019) çalışması bu çalışmayla daha ilgilidir. Karakter tabanlı derin konuşma tanıma sistemine yapılan genetik saldırıda, her örnekten rastgele seçilen iki sözcük yine İngilizce'de en yaygın 1000 sözcükten yapılan rastgele seçimlerle takas edilmiştir. Bağlantıcı zamansal sınıflandırma katmanından (CTC) yaklaşık gömme bilgisini elde etmiş ve %35 başarı sağlamıştır. Yine katman bilgisi kullandığı için gerçek ortamda uygulanabilir bir kara-kutu saldırısı değildir.

Çalışmamız, genetik algoritma kullanmasına rağmen, HMM durumları üzerinde üretim gerçekleştirilmediği için Wu ve diğ.'den (2019) büyük farklılık göstermektedir. Alzantot ve diğ. (2018a) konuşma tanıma modeli kadar kompleks olmayan bir CNN komut sınıflandırma modeli üzerinde çalışmıştır. Logit katmanına erişim olduğu varsayımının geçerliliği yüksek değildir. Her ne kadar Taori ve diğ. (2019) deney düzeneği olarak benzer görünse de CTC katmanı üzerinde çalışmış olduğu için saldırı yöntemi tamamen farklıdır.

5.4 Metin Bozma Sistemi

5.4.1 Kutupluluk Tabanlı Metin Bozma

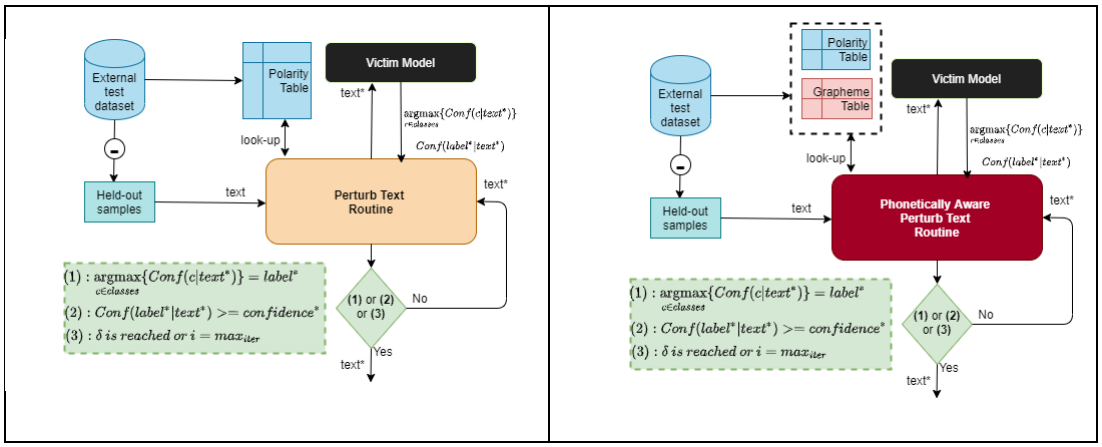
Bölüm 3'te belirtildiği üzere, sistem dışı kaynaktan toplanmış ve etiketlenmiş veri ile kara-kutu siteminin sorgulaması yoluyla hesaplanan kutupluluk skorlarından yararlanılarak, sözcük takasına dayalı başarılı saldırı gerçekleştirilebilmektedir. Tasarladığımız saldırı modeli deney sonuçlarını kısaca şöyle özetlenebilir:

- Uzman model olmadan kara-kutu metin sınıflandırma sistemlerine saldırı gerçekleştirilmiştir.
- Hedef modelin mimarisinden bağımsız, genelleştirilmiş bir saldırı modelidir.
- Neural modelin saldırıya karşı daha savunmasız olduğu görülmüştür.
- Kara-kutu sistemi sınırsız sorgulama bütçesi olduğunda %100 saldırı başarısı elde edebilmek mümkündür.
- Basit otomatik semantik benzerliğe dayalı önlemle saldırıyı tespit etmek zordur.
- Gerçek bir kara kutu saldırısı gerçeklenmiştir: IBM Watson

Bu saldırı modelinde, örnek bozulma oranı kutupluluk skorundaki değişim olarak tanımlanmıştır. Kutupluluk skoru, bir sözcüğün ikili sınıflar arası etiket değişimleri için ne kadar etkili olduğunun bir ölçümüdür. Saldırı örneğindeki bozulma oranı, takas sözcüklerindeki kutup değişimi (α) ve takas edilen sözcük sayısı (δ) ile kontrol edilmiştir.

5.4.2 Fonetik Duyarlı Kutupluluk Tabanlı Metin Bozma

Bu rapor döneminde, mevcut metin bozma sistemindeki tehdit tanımı aynı kalmak üzere, takas sözcüğü seçim stratejisine akustik benzerlik değerine bağlı sıralama eklenmiştir. Bir grapheme'den phoneme dönüştürme (Grapheme to Phoneme, G2P) sistemi ile grapheme tablosu oluşturulmuş ve zararlı örnek üretilme aşamasında fonetik benzerlik metriği kullanılarak takas adayları en yüksek fonetik benzerliğe göre sıralanmıştır. Her iki versiyonu Metin Bozma Sistemi için blok diyagram Şekil 5.2'de verilmiştir.



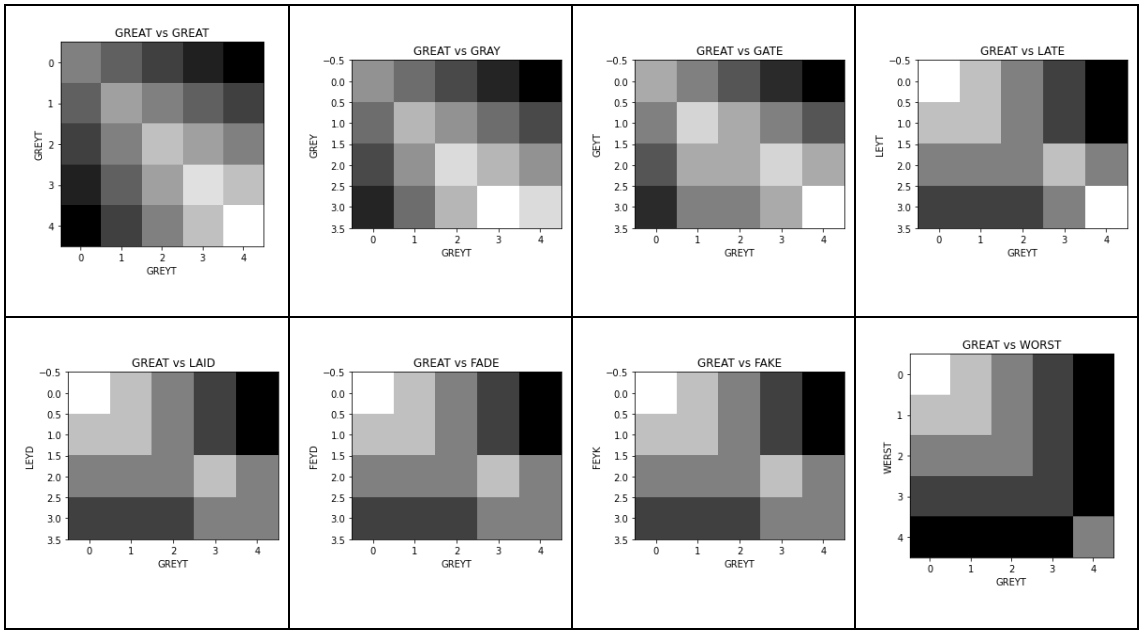
Şekil 5. 2 Metin bozma sistemi

Bu çalışmada grapheme tablosu, 256 saklı üniteye sahip 3 katmanlı CMUSphinx transformer modeli ile doldurulmuştur. G2P metin girdisini ARPABET semboller listesine dönüştürür. ARPABET, Gelişmiş Araştırma Projeleri Ajansı (Advanced Research Projects Agency, ARPA) tarafından geliştirilen bir dizi fonetik transkripsiyon kodudur. En temel haliyle İngilizce 39 fonu, farklı ASCII karakter dizileriyle temsil eder. Takas adayları temsilleri elde edildikten sonra Needleman-Wunsch (Needleman v.e diğ, 1970; Hixon ve diğ.,2013) algoritması sonucuna göre sıralanır, öncelikli olarak üst sıralı adaylar arasından seçim yapılır. Tablo 5.2'de GREAT sözcüğü için sıralı takas adayları, Tablo 5.3'te ise hizalama matrisi örneği verilmiştir.

Tablo 5. 2 GREAT sözcüğü için sıralı takas adayları

	Sözcük	**Temsil**
Orijinal	great	G R EY T
Fonetik Yakınlığa göre	gray	G R EY
	gate	G EY T
	late	L EY T
	laid	L EY D
	fade	F EY D
	fake	F EY K
	worst	W ER S T

Tablo 5. 3 Örnek hizalama matrisi



5.5 Segment Yönelimli Ses Bozma Sistemi

Konuşma tanıma sistemi girdi noktasına müdahalede, Metin Bozma Sisteminden elde edilen çıktı kullanılarak zararlı örnekler elde edilmiştir. Tehdit modeli, Metin Bozma Sistemiyle uyacak şekilde, yine üç boyutta incelenmiştir: amaç, bilgi ve kabiliyet, saldırı stratejisi. Amaç, bilgi ve kabiliyet boyutları tanımlandıktan sonra üç farklı saldırı stratejisi (Konuşmacı Uyumlu Segment Değişimi, Bağımsız Genetik Algoritma, Ortak Bozulma Optimizasyonu) açıklanmıştır.

5.5.1 Tehdit Modeli

Amaç: Saldırgan, sistemin bütünlüğünü önemli oranda azaltarak, saldırı hedefindeki modülün yanlış çıktı almasını hedefler. Belli bir çıktıyı hedefleyerek zararlı örnekleri üretir. Başarılı saldırı, hedef çıktı ya da orijinal çıktıdan sapma olarak değerlendirilir. İkincil bir kaynaktan toplanan örneklerin orijinal konuşma tanıma çıktısının, Metin Bozma Sisteminden geçtikten sonra elde edilen zararlı metin örnekleri rehberliğinde bozulan orijinal konuşma sinyali ile orijinal çıktıdan farklı sonuç elde edilmesi amaçlanmıştır.

Bilgi ve Kabiliyet: Tehdit modelinin bu boyutunda üç varsayım ele alınmıştır; birincisi asgari bilgiyi, diğer ikisi beceriklilik derecesini ifade eder.

Varsayım 5.1. Sözlü Diyalog Yönetim Sistemini oluşturan modüller ayrı kara-kutular olarak erişilebilir; yani her modüle girdi seviyesinde erişim açıktır ve çıktılar okunabilir.

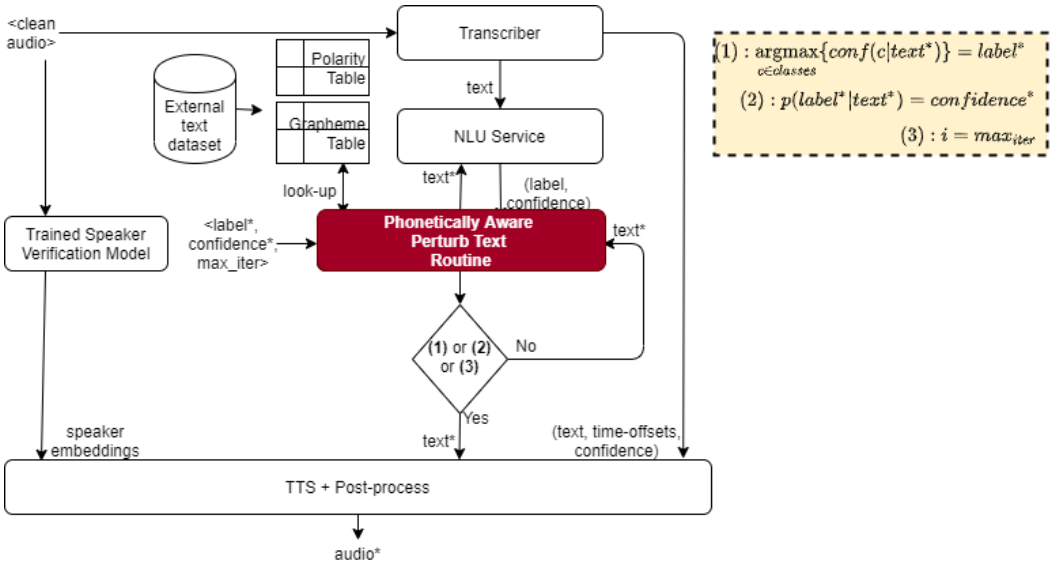
Varsayım 5.2. Saldırganının, hedef modülden bağımsız spektral sinyal manipülasyon araçlarına erişimi vardır.

Varsayım 5.3. Saldırgan, görevle ilgili neredeyse sınırsız sayıda etiketsiz örneği harici bir kaynaktan (ör. Web taraması) ücretsiz olarak toplayabilir.

Strateji: Üç saldırı stratejisi değerlendirilmiştir: Konuşmacı Uyumlu Segment Değişimi, Bağımsız Genetik Algoritma ve Ortak Bozulma Optimizasyonu. Konuşmacı Uyumlu Segment Değişimi ve Bağımsız Genetik Algoritma yalnız konuşma tanıma modülü üzerinde işlem yapar. Bunlar, Sözlü Diyalog Yönetim Sisteminin bağlı diğer modülleri yok sayarak, saldırı yalnız dilbilimsel seviyede değerlendirir. Ortak Bozulma Optimizasyonu ise tüm modüllerine eş zamanlı saldırı düzenler, sistemin toplam bütünlüğünü optimum bir şekilde bozmaya çalışır. Konuşmacı Uyumlu Segment Değişimi yeniden sentezleme, Bağımsız Genetik Algoritma ve Ortak Bozulma Optimizasyonu ise genetik algoritmalardan yararlanarak saldırı örnekleri üretir.

5.5.2 Konuşmacı Uyumlu Segment Değişimi

Bu strateji yeniden sentezleme tabanlıdır, kesin bir konuşma çıktısı elde edilmek hedeflendiğinde basit olarak kullanılabilecek ilk yöntemdir. Jia ve diğ. (2018) metinden-konuşmaya mimarisi kullanılarak Şekil 5.3'te gösterildiği üzere, konuşmacı uyumlu yeni konuşma segmenti elde edilip, sinyal üzerinde, Metin Bozma Sistemine benzer şekilde, takas yapılır.

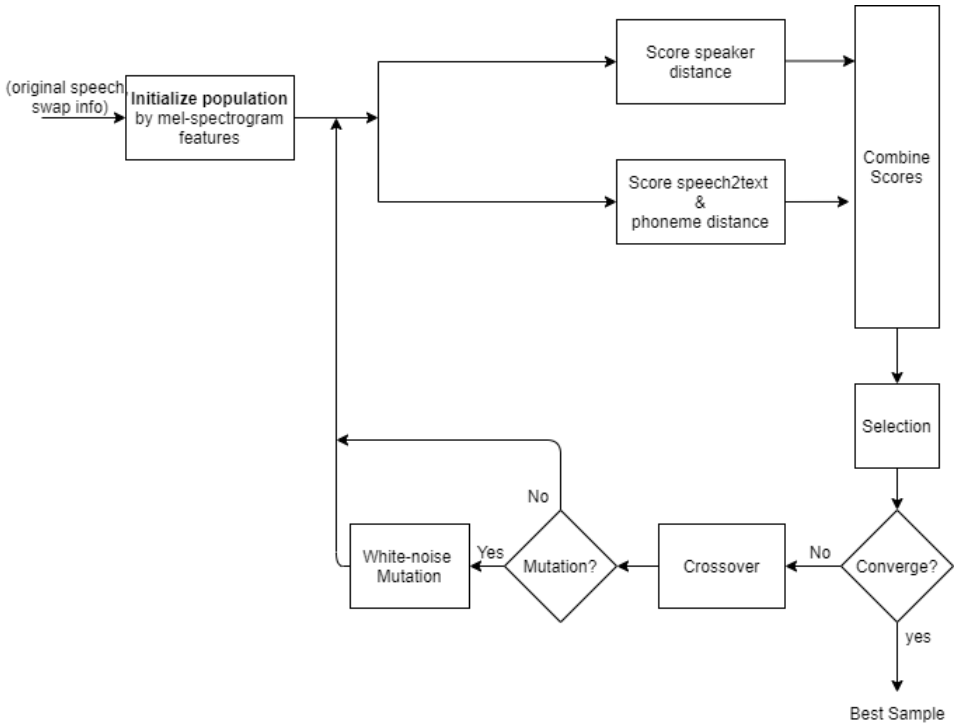


Şekil 5. 3 Konuşmacı uyumlu segment değişimi şeması

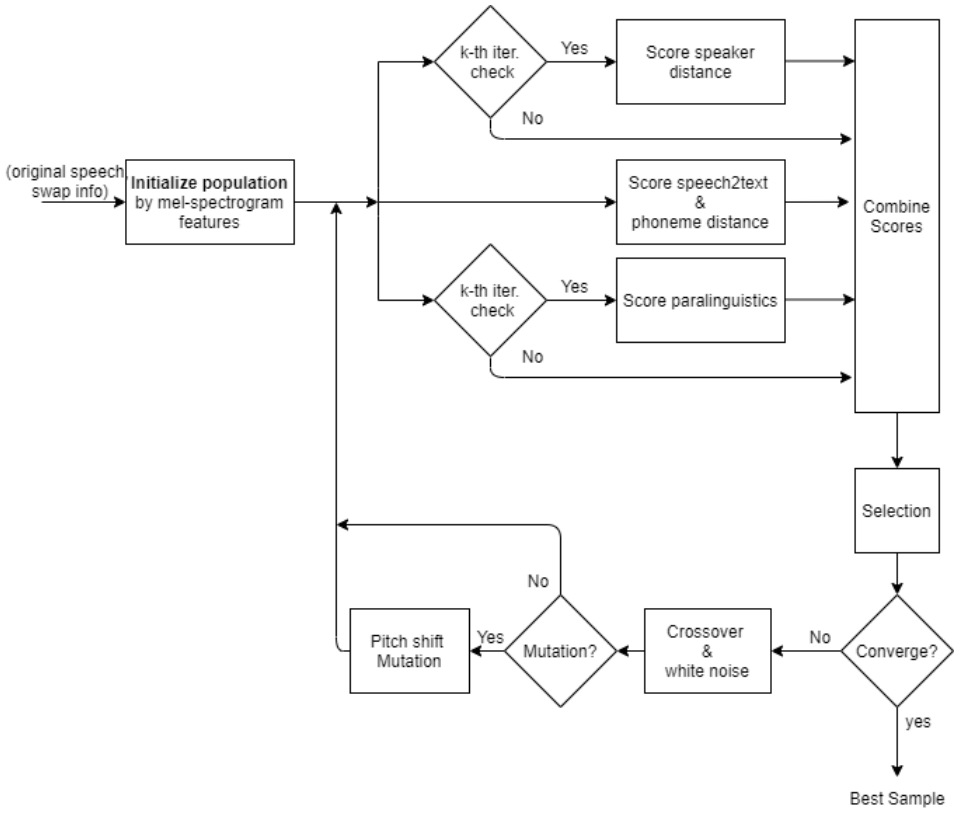
Konuşmacı uyumu, eğitilmiş derin metinden bağımsız konuşmacı tanıma modelinden çıkarılan konuşmacı temsilleri (speaker embeddings) ile sağlanır. Eğitilmiş modele test aşamasında örnek konuşmacı sinyali sağlandığında son katman konuşmacı temsillerini verir. Yeniden sentezleme prosedüründe Wan ve diğ. (2018) konuşmacı temsilleri, konuşma kodlama için Kalchbrenner ve diğ. (2018) ve sentez için Shen ve diğ. (2018) kullanılmıştır.

5.5.3 Genetik Algoritma Stratejileri

Genetik algoritma genel adımları popülasyon oluşturma, uygunluk değerlendirme, seçim, çaprazlama ve mutasyon döngüsünden oluşur. Bu adımların Bağımsız Genetik Algoritma ve Ortak Bozulma Optimizasyonu stratejilerinde nasıl farklılaştığı Şekil 5.4 ve 5.5'da görülmektedir.



Şekil 5. 4 Bağımsız genetik algoritma için akış



Şekil 5. 5 Ortak bozulma optimizasyonu için akış

İlk popülasyonun oluşturma yöntemi her iki stratejide de aynıdır. Konuşma sinyalinin değiştirilecek bölümü zaman ekseninde bellidir. İlk popülasyon üretimi için ağırlıklı Gaussian beyaz gürültü eklenir. Ağırlık kat sayısı sabittir, değişim segmentinde daha fazla gürültüye neden olmak üzere hesaplanan beyaz gürültüde ilgili segment bölümünde çarpılarak uygulanır.

Popülasyonda birey uygunluğu istenen sonuca uzaklığı değeri ifade eder, [0,1] aralığındadır. Uygunluğu aranan tüm özellikler yönünden istenen çıktı elde edilmişse sonuç sıfır olur. Her iki stratejide de popülasyon bireyleri için uygunluk hesaplanırken konuşma skoru $G(x^*, w, w^*)$ ideal konuşma çıktısına olan uzaklığı, konuşmacı skoru $E(x, x^*)$ ise değişime uğrayan sinyal konuşmacı temsilcileri ile konuşmacının orijinal konuşmacı temsilleri arasındaki mesafeyi ifade eder.

Eşitlik 5.3'te STT konuşma tanıma, SR konuşmacı temsillerinin elde edildiği konuşmacı doğrulama modelidir. (w, x) orijinal metin-sinyal ikilisi, (w^*, x^*) zararlı metin-değişen sinyal ikilisi olup, t zaman ekseninde değişim segmentini ifade eder. Konuşma tanıma modeli orijinal metin çıktısını veriyorsa 1, hedeflenen çıktıyı veriyorsa sıfır skorunu verir. Diğer durumlarda orijinal segmenten uzaklık Cosine metriği ile normalize DTW maliyet matrisi hesaplanarak, tümleyen ile orijinalden ne kadar uzaklaşıldığı kontrol edilir.

$$G(x^*, w, w^*) = \begin{cases} 1, & eğer\ STT(x^*) = w \\ 0, & eğer\ STT(x^*) = w^* \\ 1 - DTW(w^*[t], w[t]), & diğer\ durumlar \end{cases} \quad (5.3)$$

Konuşmacı uzaklığı 256-boyutlu d-vektörlerin Cosine bezerliğinin tümleyeni olarak Eşitlik 5.4'daki gibi hesaplanır.

$$E(x, x^*) = 1 - \frac{d_x \,.\, d_{x^*}}{||d_x||_2 \;\; ||d_{x^*}||_2} \quad (5.4)$$

Ortak Bozulma Optimizasyonu stratejisinde bu skorlarla birlikte, orijinal paralinguistik model çıktısından sapma skoru $C(x, x^*)$ de hesaplanır. Eşitlik 5.5, PL paralinguistik modeli ifade eder.

$$C(x, x^*) = \begin{cases} 0, & eğer\ PL(x^*) \neq PL(x) \\ p(PL(x)|x^*), & diğer\ durumlar \end{cases} \quad (5.5)$$

Toplam uygunluk skoru $F(.)$ ilgili alt skorların ağırlıklı toplamıdır. Ağırlıkların toplamının 1.0 olacak şekilde önceden belirlenmiş olması gerekir. Toplam uygunluk, Bağımsız Genetik Algoritma için Eşitlik 5.6'de, Ortak Bozulma Optimizasyonu için Eşitlik 5.7'da verilmiştir. Bağımsız Genetik Algoritma her iterasyonda iki skor bileşenini de hesaplarken, Ortak Bozulma Optimizasyonunda, hesaplama karmaşıklığını azaltıp, algoritmayı hızlandırmak için $E(x, x^*)$ ve $C(x, x^*)$ her K-iterasyonda hesaplanır.

$$F(.) = \alpha G(x^*, w, w^*) + (1 - \alpha)E(x, x^*), \qquad 0 \leq \alpha \leq 1 \quad (5.6)$$

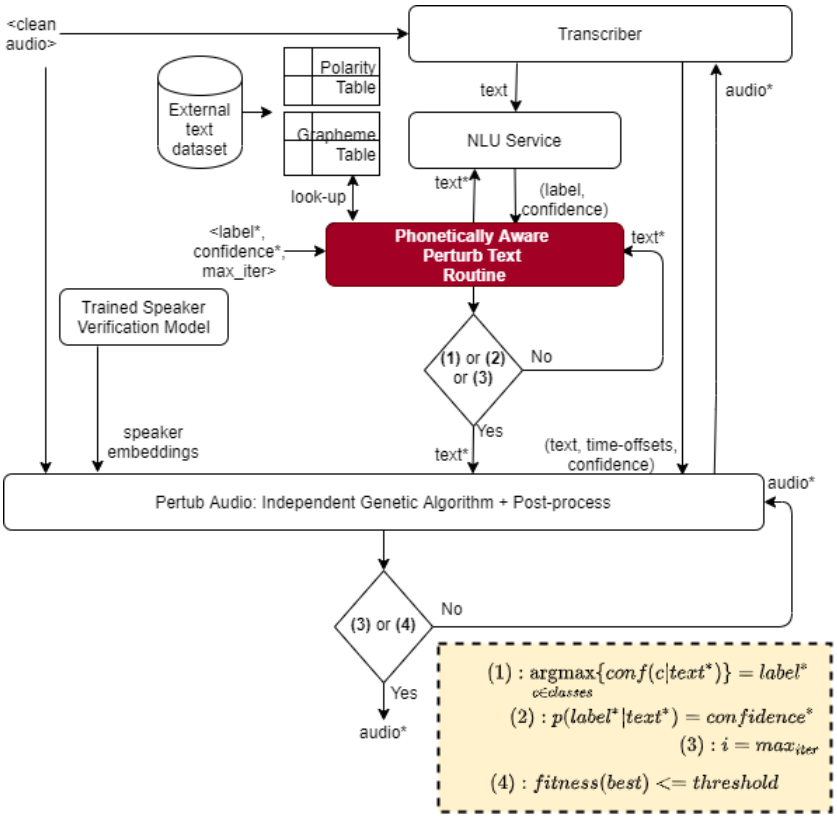
$$F(.) = \alpha G(x^*, w, w^*) + \beta E(x, x^*) + \gamma C(x, x^*), \quad 0 \leq \alpha, \beta, \gamma \leq 1, \alpha + \beta + \gamma = 1 \quad (5.7)$$

Eğer takas edilen sözcüklerin fonem uzunlukları eşit değilse zaman uzatması gerçekleştirilir. Bu işlem zaman eksininde çerçeve tabanlı olarak, ses perdesi etkilenmeden gerçekleştirilir.

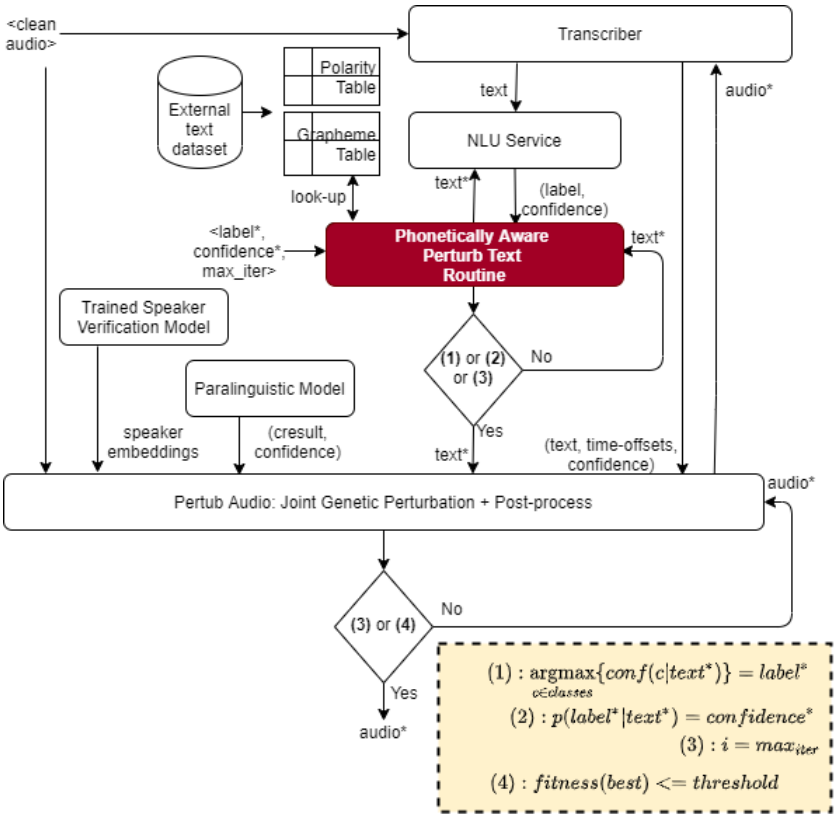
Her iki stratejilerde ebeveynler, uygunluk skoruyla doğru orantılı seçilme olasılıklarıyla, rastgele seçilirler. Tek nokta çaprazlama yapılır. Ortak Bozulma Optimizasyonunda çaprazlamadan sonra yeni popülasyona ağırlıklı beyaz gürültü eklenir. Mutasyonda ise iki strateji arasında şu değişiklik bulunmaktadır:

- Bağımsız Genetik Algoritma: Ağırlıklı beyaz gürültü ekleme.
- Ortak Bozulma Optimizasyonu: [-1, -2, 1, 2] arasından rastgele seçilmiş yarı-ton değeriyle perde kaydırma.

Her iki stratejide de algoritma, eşit uygunluk değerine ulaşılınca ya da maksimum iterasyon sayısına ulaşılınca sonlanır (Şekil 5.6 ve 5.7).



Şekil 5. 6 Bağımsız genetik algoritma



Şekil 5. 7 Ortak bozulma optimizasyonu

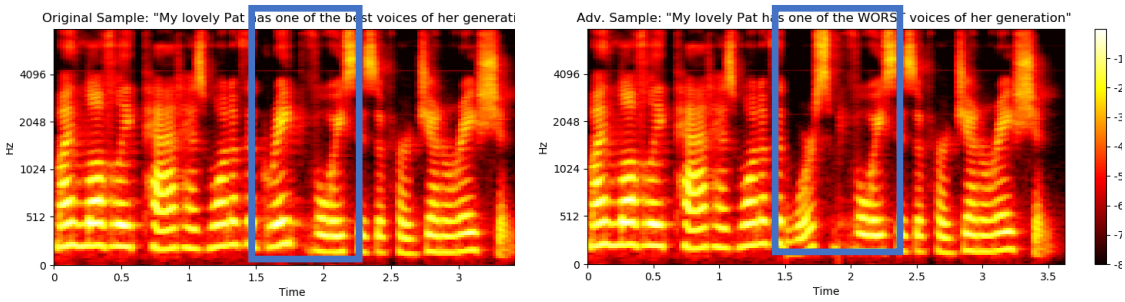
5.6 Segment Değişimi Deneysel Sonuçları

Bu bölümde, bir örnek üzerinde tüm saldırı stratejilerinin sonucu incelenmiştir. Kullandığımız orijinal konuşma dosyasının metin çıktısı şöyledir:

"My lovely Pat has one of the GREAT voices of her generation"

Orijinal örnek, GREAT kelimesinin diğer kelimelerle takası yoluyla bozulmuştur. Takasta, orijinal sözcüğe fonetik uzaktan yakına doğru sıralanmış {"worst", "late", "gate"} sözcükleri gösterilmiştir. ELSDR veri kümesinde olmayan bir konuşmacıdan elde edilmiş konuşma kaydı kullanılmıştır.

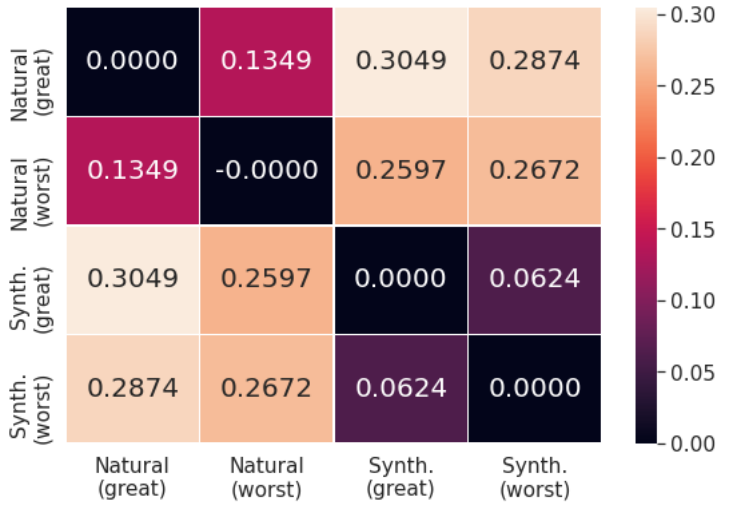
Konuşmacı Uyumlu Segment Değişimi stratejisinde konuşmacı temsilleri (embeddings) kullanılarak yeni segment sentezlenip, sinyal üzerinde takas işlemi gerçekleştirilmektedir. Sentezlemede Jia ve diğ.'nin (2018) modeli kullanılmıştır. Bu strateji en basit örnek üretme stratejimiz olup, yalnız konuşma tanıma sistemine saldırmaktadır. Şekil 5.8'de orijinal ve zararlı örneklerin mel-spektrogramları incelendiğinde birleşim noktasında kesikli olduğu gözlemlenmektedir. Bireysel dinleyici değerlendirmesinde konuşma seyrinde kopukluk olduğu doğrulanmıştır. Kesiklik algısının nedenlerinden biri artükülasyon bozukluğudur.



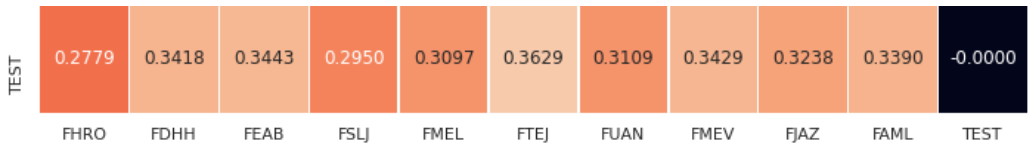
Şekil 5. 8 Original örnek mel-spectrogram (sol), zararlı örnek mel-spektrogram (sağ).

Şekil 5.9'da test konuşmacısının doğal ve sentezlenmiş konuşma sinyalleri için 256-boyutlu d-vektörlerinin Cosine uzaklığı hesaplandığında, doğal ve sentezlenmiş cümleler arasındaki konuşmacı uzaklığının yüksek olduğu görülmektedir. Farklı konuşmacı uzaklıkları ile kıyaslamak amacıyla test konuşmacısından ELSDR eğitim cümleleri kaydı alınmıştır. Şekil 5.10'de, test konuşmacısı ile aynı cinsiyette ELSDR konuşmacıları ile uzaklıklar görülmektedir. Her bir konuşmacı için eğitim

konuşmaları uç uca eklenip çıkarılan d-vektörlerin Cosine uzaklığı hesaplanmıştır. Şekiller karşılıklı değerlendirildiğinde, test konuşmacısının doğal ve sentezlenmiş konuşmaları arasında uzaklığın farklı konuşmacılar ile olan uzaklığa yakın olduğu görülmektedir[3]. Kesiklik algısının bir diğer nedeninin konuşmacı ses kimliği uyumunun çok yüksek olmamasıdır.



Şekil 5. 9 Doğal (Natural) ve sentezlemiş (Synth.) konuşma sinyali konuşmacı temsillerinin Cosine uzaklığı. "great" orijinal test cümlesini, "worst" ise test cümlesinde WORST kelimesi ile takası ifade eder



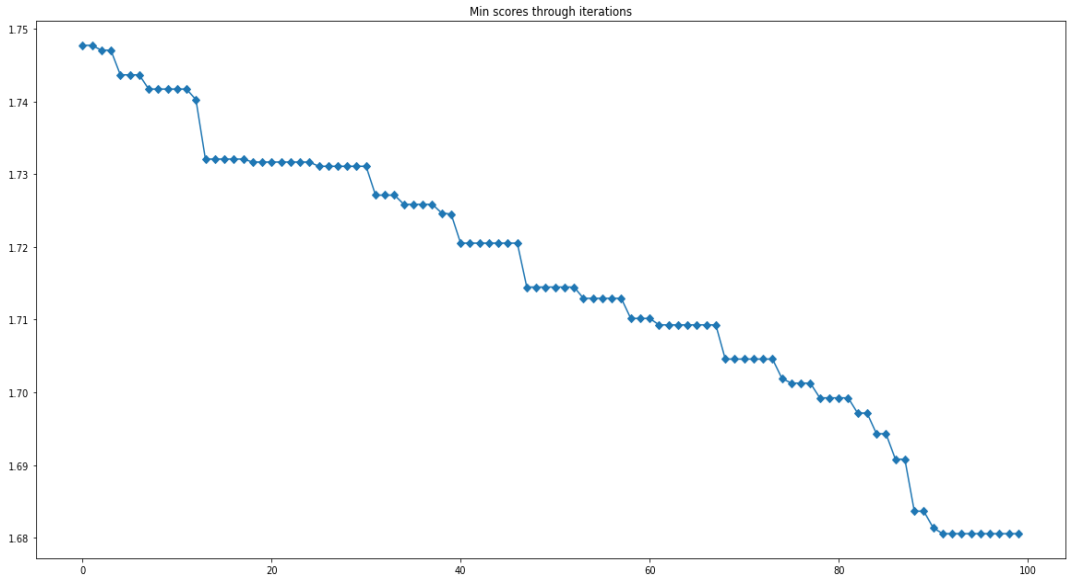
Şekil 5. 10 Test konuşma kaydı ve ELSDR konuşmacı temsillerinin Cosine uzaklığı

İki stratejiyi karşılaştırmak üzere ortak parametreler birbirine aşağıdaki gibi eşitlenmiş, iterasyon durdurma eşiği kaldırılmıştır:

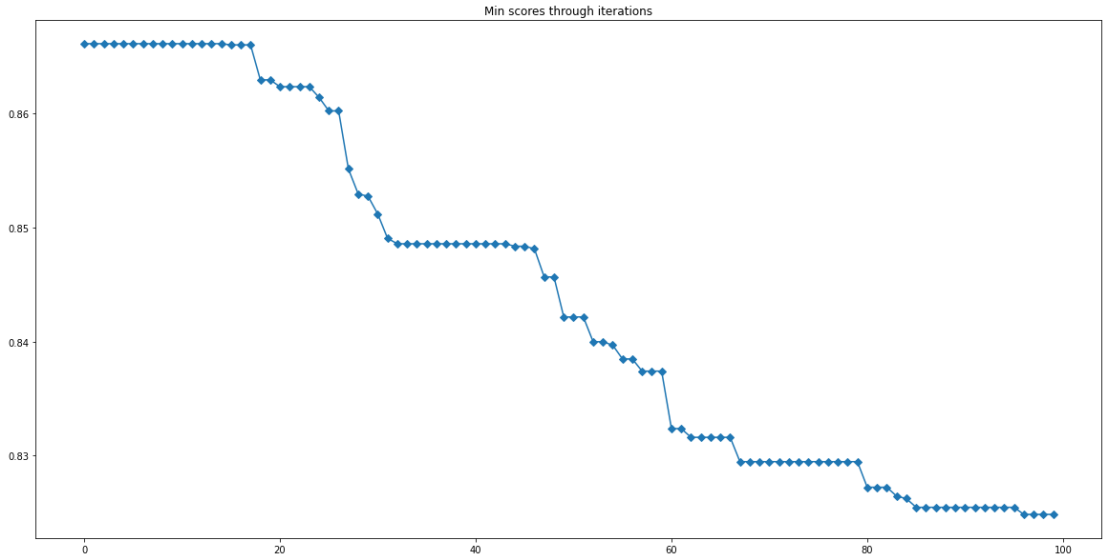
- Ağırlıklı beyaz gürültü katsayısı: 5
- Populasyon büyüklüğü: 100 ve Mutasyon olasılığı: 0.05
- Maksimum iterasyon: 100

[3] Jia v.diğ.'nin modelinden faydalanan başka araştırmacılar da sentez konuşmanın "jenerik" bir ses olduğunu raporlamıştır. (Örn. N. Müller, https://github.com/CorentinJ/Real-Time-Voice-Cloning/issues/126#issuecomment-604999848, Erişim: Haziran 2020).

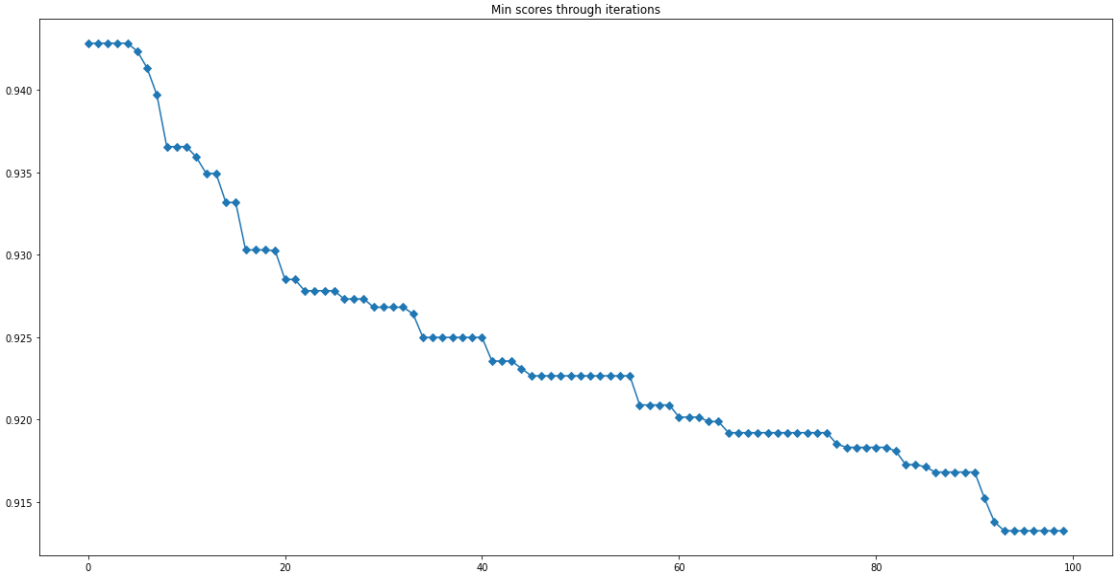
Bağımsız genetik algoritma için uygunluk fonksiyonu katsayısı $\alpha = 0.5$ ile konuşma ve konuşmacı skoruna eşit ağırlık verilmiştir. Strateji geliştirilirken, takas segmentlerinin sıfırlar dizisi ilklendirme yöntemi 3000 iterasyona kadar eksik segmente her hangi bir sözcüğü işitsel ya da $STT(x^*)$ sonucu olarak üretebilmeyi başaramamıştır. Bu nedenle, orijinal segmentin, nesiller boyu hedefe yönelik uyarlanması yoluna gidilmiştir. Şekil 5.11-5.13'te normalize edilmemiş DTW matrisi ile hesaplanan, 100 iterasyon boyu uygunluk skorları görülmektedir. Bireysel dinleme testlerinde 100 iterasyonda fonem değişimleri, yalnız GATE takası için 3000 iterasyon sonunda dinleme testi anlaşılır olmuştur. Şekillerin tümünde uzun süreli platolar görülmektedir. Bu bölgelerde nesiller boyunca iyi özellikler üretilemediği anlaşılmaktadır. Şekil 5.11'de, takas havuzunda fonetik olarak GREAT sözcüğüne en uzak ama kutupluluk olarak en yüksek olan WORST sözcük takası skoru görülmektedir. Bu nedenle ilk iterasyonlardaki uygunluk skoru da çok daha yüksektir.



Şekil 5. 11 Bağımsız genetik algoritma: great -> worst



Şekil 5. 12 Bağımsız genetik algoritma: great -> late



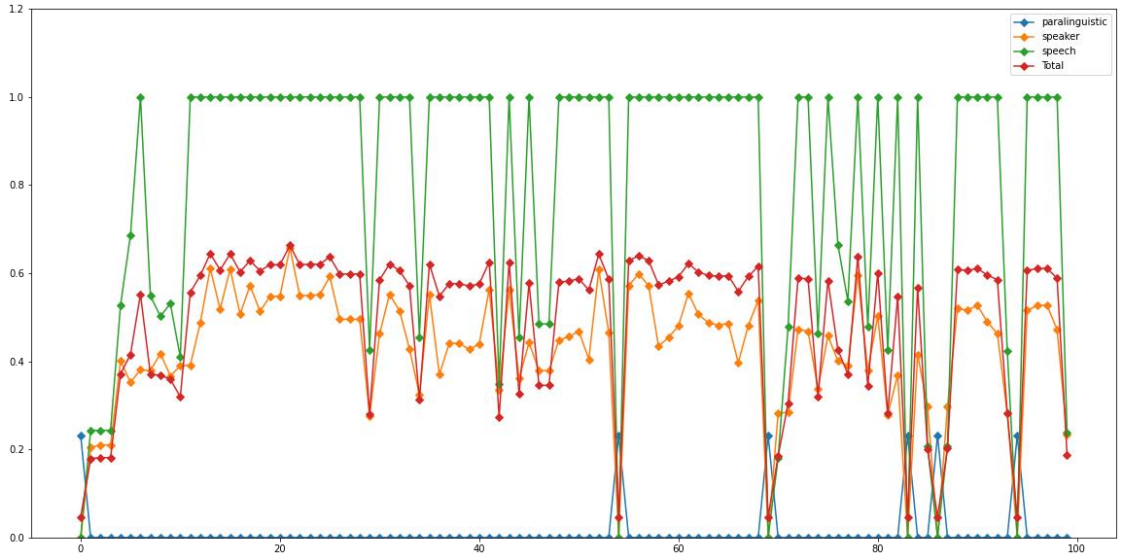
Şekil 5. 13 Bağımsız genetik algoritma: great -> gate

Ortak bozulma optimizasyonu stratejisinde uygunluk skoru için $\alpha = 0.4, \beta = 0.4$, ve $\gamma = 0.2$ ağırlıkları kullanılmıştır. Her bir alt skorun toplam uygunluk değerine etkisini değerlendirebilmek üzere Şekil 5.14-5.16'da tüm skorlar birlikte çizdirilmiştir. Her iterasyonda alt skorları izleyebilmek için K-iterasyon kontrolü kapatılmıştır.

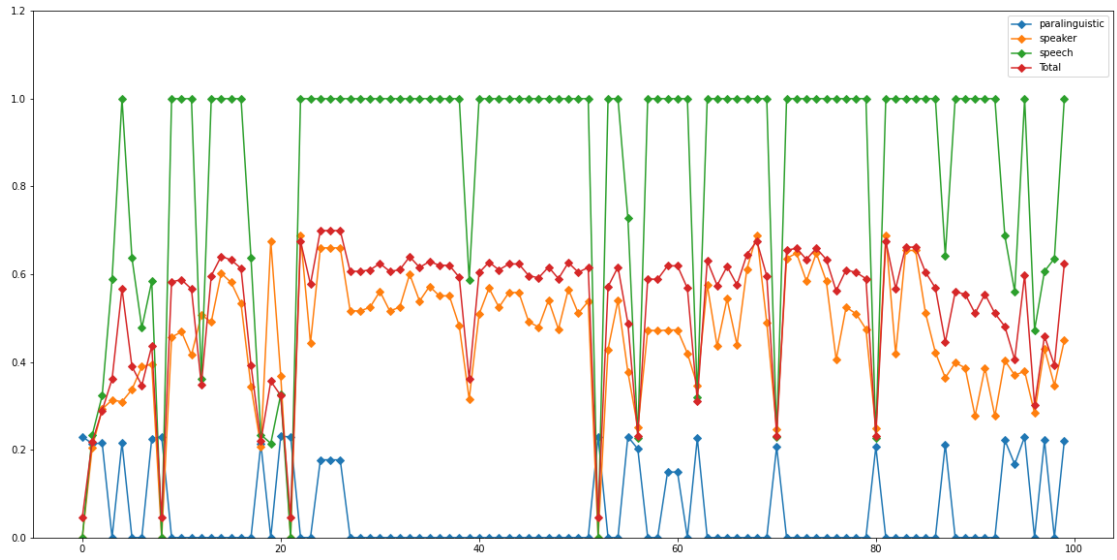
Toplam skorda geniş platolardan kaçmak için algoritmaya penaltı entegre edilmiştir. Toplam uygunluk skoru son üç nesilde $\varepsilon = 0.01$ eşiğinden fazla fark

oluşturmuyorsa, son üç nesilin en iyi bireylerine benzeyen popülasyon üyeleri agresif bir şekilde bastırılır.

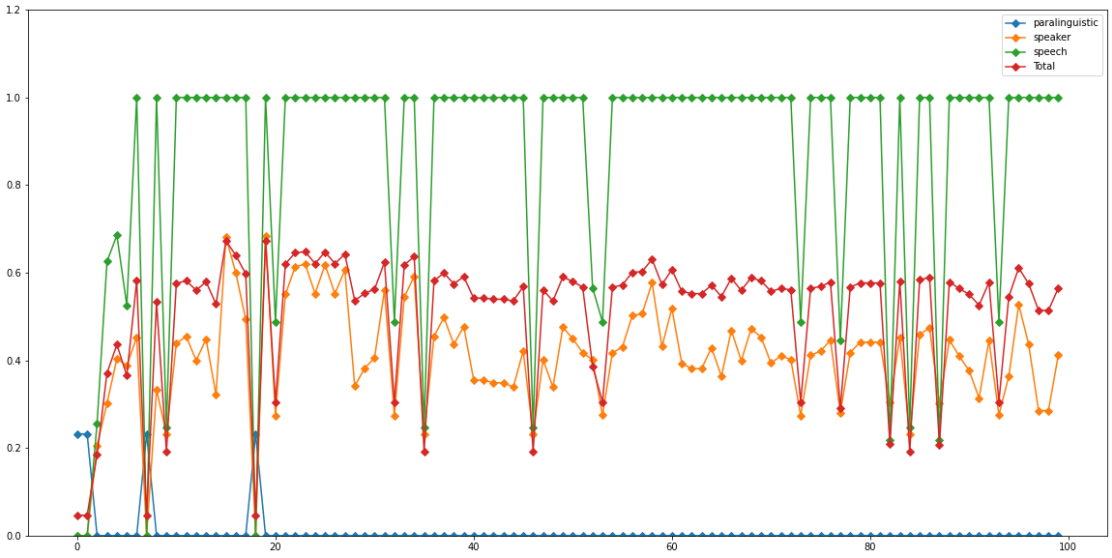
Öncelikle, tüm grafiklerde göze çarpan durum konuşma tanıma skoru $G(x^*, w, w^*)$ 'nin oluşturduğu yüksek uçurumlardır. Sinyal bozulmasına rağmen konuşma tanıma için kullanılan Google Speech API'den bağlama bağlı olarak orijinal metni dönmüştür. Dikkat çeken diğer bir nokta, paralinguistik modülünün çok daha kolay kandırılabilir görünmesidir.



Şekil 5. 14 Ortak bozulma optimizasyonu: great -> worst



Şekil 5. 15 Ortak bozulma optimizasyonu: great -> late



Şekil 5. 16 Ortak bozulma optimizasyonu: great -> gate

Tablo 5.4, K-iterasyon kontrollerinin efektifliğini göstermektedir. Deneyler Google Collabarotory hizmetinde gerçekleştirilmiştir. 100 iterasyon sonunda dinleme testlerinde takas sözcüklerinin üretildiği belirlenmiştir. Bazı nesillerde, dinleme testlerinde sözcüğün oluştuğu algılansa da Google Speech API metin çıktılarını Tablo 5.5'te örnekleri görüldüğü üzere şaşırmaktadır. Takas sözcüğü orijinal sözcükten fonetik olarak ne kadar uzaksa, orijinal cümleden o kadar farklı varyasyonlar olduğu görülmektedir. Buna göre, Bağımsız Genetik Algoritma stratejisinden farklı olarak, burada ilk 100 iterasyon içinde başarılı hedefsiz saldırılar üretilebilmektedir.

Tablo 5. 4 Genetik stratejilerin ortalama iterasyon süreleri

Strateji	**Ortalama (dk/iterason)**
Bağımsız Genetik	0.9
Ortak Bozulma Optimizasyonu (K-iterasyon kontrolü kapalı)	1.1
Ortak Bozulma Optimizasyonu (K-iterasyon kontrolü açık)	0.6

Tablo 5. 5 Kelime dönüşümünde oluşan örnek ses kayıtlarının Google Speech API çıktıları

great->worst	**great->late**	**great->gate**
1. my lovely past generation 2. my Walgreens have a generation 3. how long is a generation 4. hi lovely pattern what are the generations 5. my lovely house	1. my lovely Pat has one of the eight voices of her generation 2. my lovely cat has 180 voices of her generation 3. my lovely cat voices of her generation 4. I love it 5. find Walgreens	1. my lovely Pat has one of the great voices of her generation 2. my lovely Pat has one of the voices of her generation 3. my lovely Pat has one of the group 18 voices of her generation 4. my lovely having voices of her generation 5. I won't leave you a voice in the first generation

5.7 Segment Değişimi Stratejileri Değerlendirmesi

İki-kipli sesli diyalog yöntemlerine kaçınma saldırıları tasarlanmıştır. Kutupluluk tabanlı metin sınıflandırıcı saldırılarında, önceden tanımlı grapheme tabloları eklenerek zararlı örneklerde fonetik duyarlılık aranmıştır.

Sesli diyalog yönetim sistemi senaryosunda iki saldırı noktası belirlenmiştir: metin sınıflandırıcı girdisi ve ses tanıma girdisi. Ses tanıma modellerine karşı literatürdeki kara-kutu ve yarı-kara kutu kaçınma saldırı çalışmaları (Alzantot ve diğ., 2018; Wu v.e diğ., 2019; Kreuk ve diğ., 2018) konuşma tanıma sistemlerinin son katmanlarından mimariye dayalı özellikleri kullanmaktadır. Bu çalışmada ise daha alt seviye akustik özellik ve yöntemlerden yararlanılarak Google Cloud Speech API'ye gerçek kara-kutu saldırısı gerçekleştirilmiştir. Hedefsiz saldırılarda başarı elde edilmiş, hedefli saldırılarda da bireysel dinleme testi başarılı olmuştur.

Üç farklı strateji ile saldırı gerçekleştirilmiştir: Konuşmacı Uyumlu Segment Değişimi, Bağımsız Genetik Algoritma ve Ortak Bozulma Optimizasyonu. Segment değişimi stratejisinde, üç ayrı derin sinir ağının eğitim yükü bulunmaktadır. Yalnız konuşmacı temsillerinin eğitildiği konuşmacı kodlayıcı (speaker encoder) sisteminin Google Colab GPU eğitimi 48 saatten fazla sürmektedir. Tüm bu

hesaplama yüküne rağmen, konuşmacı uyumlu segment üretildiğinde bunun uyumunun doğal varyasyonlardan düşük olduğu görülmüştür. Bağımsız Genetik Algoritma akustik manipülasyon sonucunda 3000 iterasyonda 2/3 test örneğinde dinleme testinin başarısız olduğu gözlemlenmiştir. Genetik üretim prosedüründe paralinguistik sınıflandırıcı ekleme, çaprazlama, mutasyon ve uygunluk fonksiyonlarında değişikliğe gidilerek, 100 iterasyon sonunda örneklerin dinleme testini geçtiği görülmüş, konuşma tanıma sistemine hedefsiz saldırı örnekleri oluşmuştur.

Sonuç olarak, metin ve konuşma verisinin bozulmaya uğratılarak iki-kipli bir sisteme yapılan bütünleşik saldırıda, ses sinyali üzerinde segment değişimi yapılması maliyetinin yüksek olduğu görülmüştür. Dahası, dinleme testlerinde hedef çıktının algılanması için yüksek sayıda iterasyon gerekmektedir. Bu da konuşma segment değişimi prosedürlerinin servis sağlayıcılarının işlem yoğunluğuna ya da sıklığına yönelik aldıkları önlemlere yakalanma riskini artırır. Şekil 5.1 sistem saldırı noktalarını dikkate aldığımızda, konuşma tanıma sistemine saldırma prosedürünün atlanması halinde bile metin bozma sisteminden yapılan müdahale sayesinde son çıktının etkilenebileceği açıkça görülmektedir. Bu nedenle, Bölüm 4'te tanıttığımız, 5 iterasyonda bile yüksek başarı gösteren ses bozma sistemi bütünleşik saldırıya entegre edilmiştir (Bölüm 5.8).

5.8 Akustik Yönelimli Ses Bozma Sistemi ile Bütünleşik Saldırı

İnsan-bilgisayar etkileşiminde önemli yere sahip olan duygu tanıma problemi üzerine yapılan araştırmalar çok-kipli girdi işlemenin bireysel başarıları yükselttiğini belirtmektedir (Yoon ve diğ., 2018; Sahu, 2019; Sahu ve diğ., 2019; Tripathi ve diğ., 2018). Dilbilimsel ve akustik bozulmalar ile kaçınma saldırısı için hedef uygulama alanlarından biri olması nedeniyle, geliştirilen yöntemin uçtan uca bir kara kutuda değerlendirilmesini yapmak üzere iki-kipli çok sınıflı duygu tanıma sistemi tasarlanmıştır. Bilgimiz dahilinde, bu düzenekte iki-kipli saldırı modeli üzerine literatürde benzer çalışma yoktur. Segment değişimi yönelimli yaklaşımdan farkı, ses kipinde metin bozma sisteminde sözcük takas noktalarına denk gelen segmentlerin yeni oluşturulmuş segmentler ile değiştirilmesi yerine, konuşmacı kimliğinin göz ardı edilerek sinyal üzerinde fark edilmeyen değişikliklerin

yapılmasıdır. Segment değişimi stratejisinde görülen derin vadi sorunu genetik algoritmada en iyi örneğin sonraki nesillerde korunması sağlanarak, negatif yönde bozulmasının önüne geçilerek sağlanmıştır. İlgili yöntem Bölüm 4'te Şekil 4.4 ve 4.5 üzerinde gösterilmiştir. Bölüm 4'te elde edilen sonuçlardan yararlanılmış, ses akustik perde bozulmalarına uğratılmıştır.

5.8.1 Kurban Model

Uçtan uca daha gerçekçi bir iki-kipli duygu tanıma modeli elde etmek üzere dilbilimsel ve fonetik çeşitliliği yüksek ve daha fazla sayıda duygu arasında sınıflandırma görevi içeren IEMOCAP (Busso ve diğ., 2008) veri kümesi kullanılmıştır. IEMOCAP, 5 oturumda toplanmış, her oturumda farklı senaryo ve doğaçlama etkileşimlerden oluşan üç-kipli olarak toplanmış veri kümesidir (Tablo 5.6 ve 5.7). Her oturumda farklı iki konuşmacı (1 kadın, 1 erkek) bulunmaktadır (5 oturum: toplam 10 katılımcı). Fiziksel jestler için hareket noktaları, ses ve transkript kipleri, on sınıf (angry, happy, sad, neutral, frustrated, excited, fearful, surprised, disgusted, other) olarak, topluluk oylaması ile etiketlenmiş, transkript el ile oluşturulmuş olarak sunulmuştur. Düşük frekanslı örnekler olduğu için disgusted, fearful ve other etiketleri genellikle çalışma dışı tutulmaktadır.

Tablo 5. 6 Kullanılan IEMOCAP etiket dağılımı

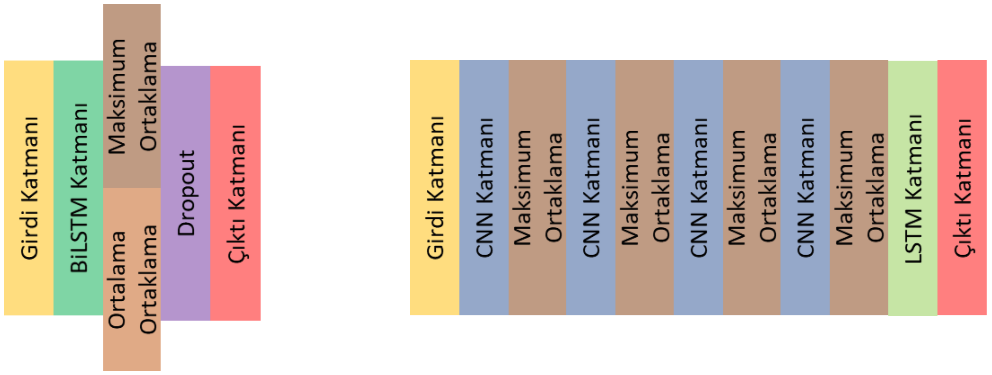
Etiket	Eğitim	Test	Toplam
Angry	876	227	1103
Excited	831	210	1041
Frustrated	1488	361	1849
Happy	479	116	595
Neutral	1362	346	1708
Sad	868	216	1084

Tablo 5. 7 Veri kümesi örnek uzunluğu

	Eğitim	Test
Maksimum (sn)	34.14	22.52
Minimum (sn)	0.73	0.58
Ortalama (sn)	4.58	4.62

Şekil 5.17'de metin ve konuşma kipleri için kurban mimari katmanları görülmektedir. Metin sınıflandırma modeli hiper-parametreleri Bayesian

optimizasyonu ile, konuşma sınıflandırma modeli hiper-parametreleri ise Zhao ve diğ.'nin (2019) aynı veri kümesi üzerinde yaptığı çalışmaya göre belirlenmiştir (Ayrıntı için bkz. Tablo A.4).



Şekil 5. 17 Metin (sol) ve konuşma (sağ) sınıflandırma kurban mimarileri

Kiplerin birleştirilmesinde erken füzyon ya da geç füzyon kullanılabilir. Erken füzyon, kip modellerinin son katmanlarının birleştirilerek eğitilmesidir. Ancak, bu durumda iki modaliteden gelen özelliklerin ön işlemeden geçirilerek iyi hizalanmaları sağlanmalıdır. Geç füzyon, kategorik çapraz entropi maliyeti üzerinden gerçekleştirilir ve iki temel yaklaşım mevcuttur: eklemeli ve çarpımsal. Liu ve diğ. (2018b) çok-kipli derin modellerin birleştirilmesi ile ilgili yaptıkları araştırmada, birden fazla derin modelin son bir ara katmanla birleştirilerek yapılan eğitimin (eklemeli yaklaşım), farklı veri tipleri girdileri ile oluşan kip modellerinin bireysel güvenilirliğinin eşit tutulması nedeniyle, farklı kiplerin güvenirliğini dikkate alan çarpımsal yaklaşımın tercih edilebileceğini belirtmiştir. Çarpımsal kombinasyon, daha güvenilir bir kipten bir karar seçer; daha zayıf kipten geri dönen hatalar bastırılır. Yalnız son sonuç değil, kipler üzerinde oluşan bireysel performans da incelenmek amaçlandığı için çalışmanın bu bölümünde çarpımsal yaklaşım ile iki-kipli model eğitimi uygulanmıştır (Eşitlik 5.8 ve 5.9).

$$q_s^y = (1 - p_t^y)^\beta,\ q_t^y = (1 - p_s^y)^\beta \tag{5.8}$$

$$p^y = -(q_s^y log\ p_s^y + q_t^y log\ p_t^y) \tag{5.9}$$

Bir örnek için gerçek sınıf olasılığı p, gerçek etiket y, metin sınıflandırma modeli t ve konuşma sınıflandırma modeli s ile gösterilmek üzere, ölçekleme faktörü q, diğer kip modelinin tahmin kalitesini temsil eder. $\beta = 0.5$ olmak üzere eşitlikler, iki-kipli sistemi oluşturan modellerden daha kötü olana daha düşük ağırlıklandıracaktır. İki kipli bu modelin doğruluk başarısı 0.6441'dir.

5.8.2 Tehdit Modeli

Amaç, bilgi ve yetenek ve varsayımlar metin bozma sistemi için Bölüm 3.6, ses bozma sistemi için Bölüm 5.5.1'de tanımlandığı üzeredir. Metin bozma stratejisinde fonetik duyarlı kutupluluk tabanlı metin bozma prosedürü (Bölüm 5.4.2) uygulanmış, kutupluluk tablosu için IEMOCAP test ve DailyDialogue (Li ve diğ., 2017) veri kümesi birlikte kullanılmıştır. Ses bozma için perde tabanlı pertürbasyon yöntemi (Bölüm 4.5) kullanılmıştır.

Üç saldırı stratejisi uygulanmıştır: hedefli, uyarlanabilir hedefli, ayrım gözetmeyen. Hedefli strateji, bir örneği saldırganın manuel belirlediği bir yanlış pozitife doğru bozulmayı amaçlar. Uyarlanabilir hedefli, örneği en olası ikinci yanlış pozitif sınıfa yönelik değiştirir. Herhangi bir hedef belirtilmediğinde ise, ayrım gözetmeyen strateji uygulanır. Her iki kip modeli için de ortak bir objektif fonksiyonu kullanılır. Objektif fonksiyonu, c sınıf etiketi belirtmek üzere, hedefli ve uyarlanabilir hedefli saldırılar için $y = c_{hedef}$ ile Eşitlik 5.9'u maksimize etmek, ayrım gözetmeyen saldırılar için tüm $y = c_{gerçek}$ ile Eşitlik 5.9'u minimize etmektir.

5.8.3 Deneysel Sonuçlar ve Değerlendirme

Mükemmel deney ortamında (Bkz. Tablo 3.2) sınırsız saldırı (Bkz. Tablo 3.5) deneyleri geçekleştirilmiştir. Test kümesinde, doğru sınıflandırılan 0.50 üzeri gerçek sınıf güveni döndüren örnekler ortak bozulma işlemine tabi tutulmuştur. Hedefli saldırılarda, hedef yönde bulunan doğru pozitifler ayrım gözetilmeyen strateji ile bozulmaya uğratılmıştır. Tüm saldırıların hava ortamında başarıyla aktarılabilirliği simülasyon ortamında doğrulanmıştır (simülasyon bilgileri için bkz. Tablo A.5).

Tüm stratejilerde metin üzerinde ortalama takas sayısı 5'ten azdır (Tablo 5.8). Kurban modele gerçekleştirilen toplam sorgularında ortalama değerler, yalnız

segment yönelimli ses bozma sisteminde (Bölüm 5.5) yapılan sorgudan ($iterasyon\ sayısı \times popülasyon$) çok daha düşüktür. En yüksek takas frekansına sahip ikililer Tablo A.1-A.3'te her saldırı stratejisi için gösterilmiştir.

Tablo 5. 8 Saldırı stratejilerinde ortalama takas ve sorgu sayısı

Hedef Sınıf	Ort. δ	Ort. Sorgu
Angry	2.7099	202.5170
Excited	2.5851	204.5296
Frustrated	3.1046	298.8008
Happy	4.2295	383.8235
Neutral	4.5372	512.5473
Sad	1.9911	155.6810
Ayrım Gözetmeyen	2.3820	150.6583
Uyarlanabilir Hedefli	2.3177	140.3203

Belirli bir sınıfa yanlış pozitif yönelimi hedeflenmiyorsa, ayrım gözetmeyen veya uyarlanabilir hedefli saldırıların ortalama takas ve sorgu maliyetlerine bakılarak birbirlerine büyük oranda üstünlükleri yoktur. Ancak saldırı başarısı açısından (Tablo 5.9) uyarlanabilir hedefli stratejinin daha avantajlı olduğu görülmektedir. Hatta tüm saldırı stratejileri ile karşılaştırıldığında uyarlanabilir hedef belirlemenin saldırı başarıyı sergilemiştir.

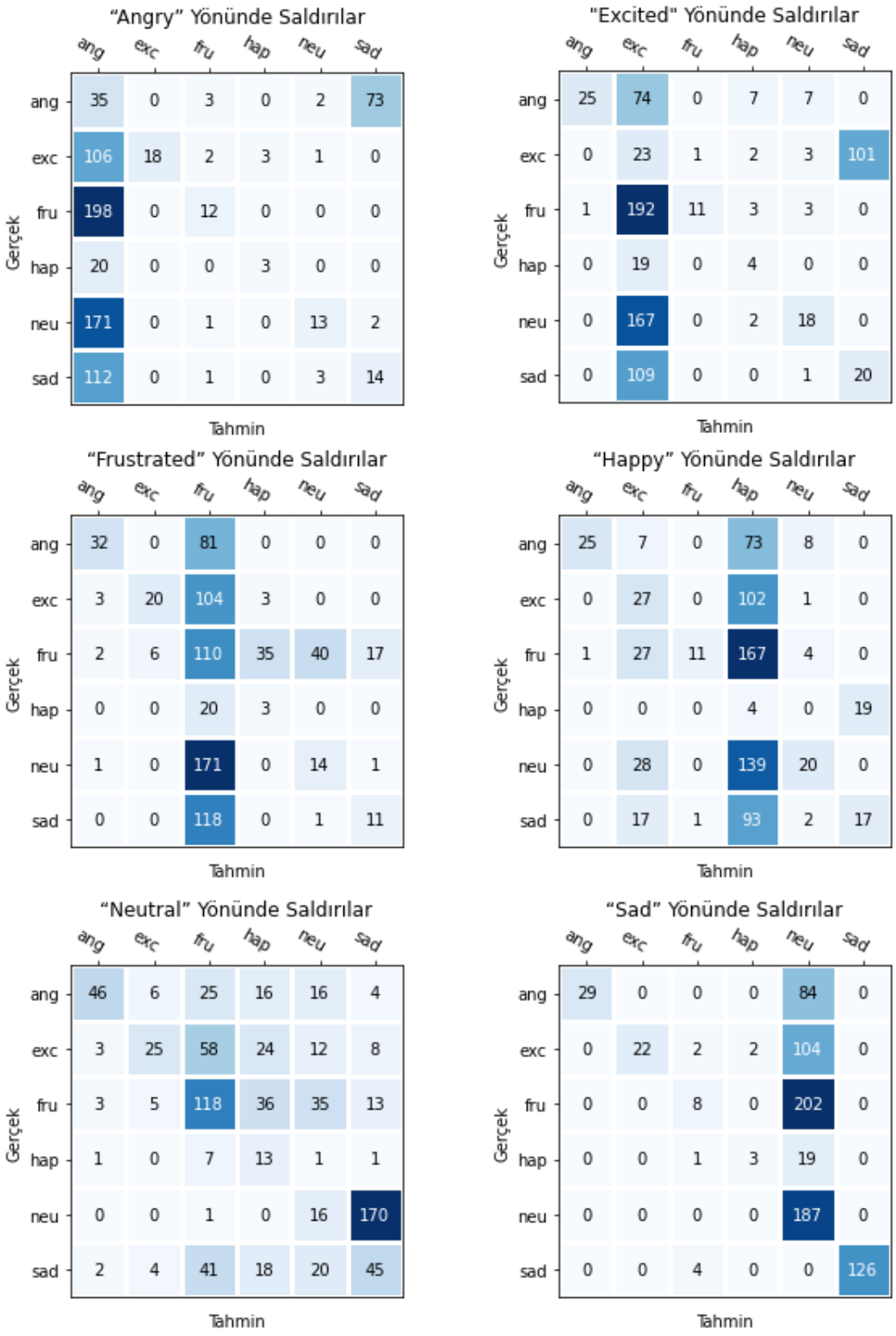
Tablo 5. 9 Saldırı stratejilerinde saldırı başarısı ve ortalama hedef sınıf güveni

Hedef Sınıf	Saldırı Başarısı	Ort. Hedef Sınıf Güveni
Angry	0.8790	0.5136
Excited	0.8715	0.5056
Frustrated	0.7594	0.3894
Happy	0.8677	0.4306
Neutral	0.6675	0.1420
Sad	0.5264	**0.6732**
Ayrım Gözetmeyen	0.7254	0.5918
Uyarlanabilir Hedefli	**0.8828**	0.5281

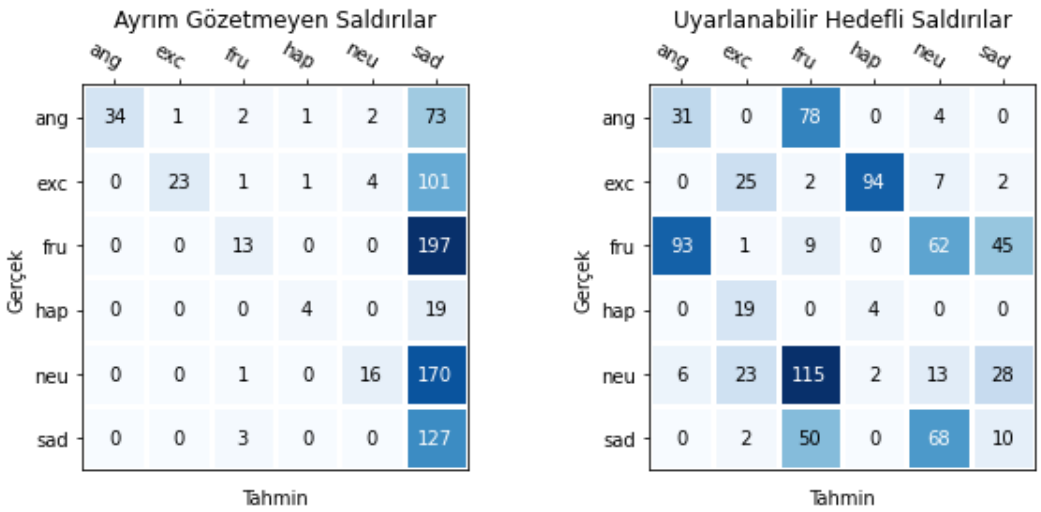
Şekil 5.18 ve 5.19 karmaşıklık matrisleri saldırı stratejilerinin bozulmanın hangi sınıflara yöneldiği hakkında sonuçları gösterir. Tablo 5.8 ve 5.9 birlikte değerlendirilerek "kolay" ve "zor" hedeflere dair çıkarım yapılabilir. Happy ve Neutral sınıflarında görülen ortalama değerler, bir örneği bu sınıflara dönüştürmenin zor olduğuna işaret etmektedir. Happy sınıfı için bu zor saldırı

yüksek başarıyla sonuçlansa da, Neutral sınıfında en düşük ikinci başarı oranı sağlanmıştır. Benzer şekilde, Sad sınıfına dönüşümlerde çok daha az maliyetli bozulmalar gerçekleştirilmiş, düşük başarı elde edilmiştir. Ortalama hedef sınıf güveni Neutral sonucuyla karşılaştırıldığında, yüksek güven oranı olduğu görülmektedir.

Hedefli ve ayrım gözetmeyen saldırılarda, gerçek pozitif sınıflandırmaların Sad etiketine doğru yönelimi görülmektedir. Ancak, Sad sınıfı örnekleri ayrım gözetmeyen yönde saldırıya karşı dirençlidir. Yine bu sınıf ile ilgili dikkat çeken diğer bir sonuç, bu yönde gerçekleştirilen hedefli saldırıların Neutral etiketine yönelmiş olmasıdır. Bu iki sınıf ile ilgili öğrenilen ortak özelliklerin yüksek olması olasıdır. Ayrıca, örnekleri hedefli olarak Neutral yönüne çevirmek zordur. Bu yöndeki saldırılarda doğru pozitiflerin Sad sınıfına yönelmesi, bu iki sınıfın ortak özelliklerinin yüksek oranda olduğuna dair kanaatimizi güçlendirmektedir.



Şekil 5. 18 Hedefli saldırı stratejilerinde karmaşıklık matrisleri



Şekil 5. 19 Ayrım gözetmeyen (sol) ve uyarlanabilir (sağ) saldırı stratejilerinde karmaşıklık matrisleri

6
SONUÇ VE ÖNERİLER

6.1 Metin Sınıflandırıcılarına Saldırı Sonuçları

Metin sınıflandırıcıları için kutupluluk değeri tabanlı hedefli ve ayrım gözetmeyen saldırı yöntemi tasarlanmış, performansı incelenmiştir. İki sınıflı modeller için çapraz kutupluluk deneyleri, en zengin kaynak olan Amazon kutupluluk tablosunun örnekleri bozmada daha iyi olduğunu göstermektedir. Daha kesin kutupluluk skorları hesaplamak için kutupluluk skorlarının hesaplandığı kaynak veri kümesi boyutunun çok önemli olduğu görülmüştür. Saldırı güveni ve sorgu maliyeti ödünleşimlidir. Kutupluluk kaynağının zenginliği, yapılması gereken sorgu sayısı için çok önemlidir. Zengin kaynaktan hesaplanan kutupluluk oranları birbirine yakınsama eğilimi gösterirler ve bu, arama uzayını büyük ölçüde genişletir.

Çok sınıflı problemler için çekişmeli örnekler oluşturmak daha zordur. Yine de, bozulma aralığından veya takas sayısından ödün verdiğimizde başarılı saldırılar üretebildiği görülmüştür.

Kutupluluk tabanlı saldırılar kurban modele göre ince ayarlanmıştır, bu nedenle kötü niyetli örnekler modeller arasında iyi bir şekilde aktarılamaz. Bununla birlikte, çapraz-kutupluluk kaynakları etkili bir şekilde aktarılabilir. Bu, düşmanın rastgele bir kaynaktan topladığı veriyi kullanarak doldurduğu kutupluluk tablosunun kara kutu saldırıları gerçekleştirmek için kullanılabileceği anlamına gelir. Ancak saldırı güveni, kaynağın zenginliğine bağlıdır; düşmanın sınıfların yeterli bir temsili olan örnek uzayına sahip olması gerekir.

Bozulmuş örneklerin orijinallerle semantik benzerliklerini koruduğu ELMo kelime gömme vektörleri ile gösterilmiştir. Saldırının, böyle semantik benzerliğe dayalı basit otomatik karşı önlem ile tespit edilmesinin kolay olmadığını göstermiştir. Ayrıca, gerçek kara kutu saldırılarındaki insan değerlendirmesinde, saldırı örneklerinin çoğunlukla orijinal örneklerle aynı kategoriye atanmıştır.

Kutupluluk saldırılarına karşı önerdiğimiz bir savunma yönü, NLU hizmetlerinin genel kutupluluk önyargılarını güncellemek için metin akışlarındaki terimlerin zamansal kutuplarını ölçmektir. Bu tür bir savunmaya ilham veren birkaç yeni çalışma vardır. AL-Sharuee (2020) duygu kalıplarını denetimsiz bir şekilde zaman içinde güncellemiştir. Bu tür model güncellemeleri, rakiplerin kutupluluk modeline ayak uydurmasını zorlaştıracaktır, çünkü topladığı harici verileri güncellemesi ve her güncelleme için büyük miktarda etiketleme bütçesi ayırması gerekecektir. Fkih ve Omri'nin (2020) çalışmaları da benzer bir şemaya dayanmaktadır. Yöntem, sınıf-sınıf kutupluluğu yerine, ilgili-ilgisiz-ilgisiz terimlerin aracı model ile öğrenilmesine dayanır. Araştırmacılar, ayrıntılı dilbilimsel özelliklerden yararlanmış ve veri üzerinde pencere kaydırması kullanmıştır. Bu tür mekanizmalar, NLU hizmet sömürüsünün azaltılması konusunda umut verici araştırma yönergelerine yol açacaktır. Bununla birlikte, eğer düşman periyodik olarak pahalı etiketleme maliyetini üstlenebiliyorsa, kutupluluk puanı yaklaşımımız bu savunmaya entegre edildiğinde hem bire-bir (hedefli) hem de bire-bir dinlenme (ayrım gözetmeyen) sınıflandırma durumunda önemli terimleri belirlemenin çok daha basit bir yoludur. Bu potansiyel savunma, patlamalı terimler için "hareket eden bir kutupluluk" tespit edilerek ve hesaplanarak aşılabilir. Bir metin akışında ani ve artan sıklıkta görülen bir terim "patlamalı" terimler, metin akışlarında konu modelleme için önemli özelliklerdir (Zhao, 2017). Gelecekteki çalışmalarda bu araştırma yönünü araştırılması önerilmektedir.

Sorgu işlemeden önce konuşlandırılacak herhangi bir savunma, ek hesaplama yükü anlamına gelir. Bu da yanıt süresini etkileyerek müşteriye olumsuz deneyim olarak yansır. Bu nedenle, önce basit bir anomali algılama şeması uygulamak, ardından şüpheli sorguları daha fazla inceleme için başka savunma modüllerine yönlendirme konusu araştırılabilir. Kolomvatsos (2018), yükleme ve yürütme süresi gibi geçmiş davranışlara dayalı bir savunma yöntemi önermiştir. Saldırımızı azaltmak için NLU hizmetleri tarafından benzer bir yaklaşım kullanılabilir. Sorgu isteği/yanıtı veya semantik olmayan meta özelliklerin belirli geçmiş davranışı, şüpheli sorgunun yakalanması ve daha fazla araştırma için özel bir modüle yönlendirilmesi gerçekleştirilebilir.

6.2 Konuşma Sınıflandırıcılarına Saldırı Sonuçları

Beyaz gürültü mutasyonu ve perde mutasyonu olmak üzere iki versiyonun bir genetik algoritmasıyla çekişmeli örnekler oluşturulmuştur. Kara kutu ortamında önceden işlenmiş özellikleri girdi alan MLP ve dalga boyu girdisi alan CNN kurban modelleri hedef alınmıştır. MLP model girdisinin kendisi küçük beyaz gürültüye dayanıklı olduğundan, MLP kurbanı beyaz gürültü saldırılarına karşı daha gürbüz performans sergilemiştir. Perde pertürbasyonları, kurban modellerinde genel olarak başarılı olmuştur. Perde bozulması, daha genelleştirilmiş bir versiyonu olarak, iki-kipli sistemlere saldırı için de kullanılmıştır.

Hesaplama maliyeti açısından, doğrudan dalga formu üzerinde gerçekleştirildiği için beyaz gürültü saldırılarını gerçekleştirmek daha kolaydır. Bu bozulma yöntemi, bazı SER modellerine saldırmanın etkili bir yolu olabilir. Ancak MLP kurban modelinin beyaz gürültüye gürbüz olduğu görülmüştür. Öte yandan, perde mutasyon stratejisi, iki kurban üzerinde de etkili olan, daha genelleştirilmiş bir saldırıdır. Bu nedenle, saldırganın SER sistemi için temel model hakkında hiçbir bilgiye sahip olmadığı kara kutu ayarlarında perde mutasyonu saldırısı seçilebilir.

Beyaz gürültü saldırısının bir diğer dezavantajı, bunlara karşı savunmanın daha kolay olmasıdır. Eklemeli beyaz gürültü, konuşma sinyalinden çıkarılması kolay olan zayıf bir saldırıdır (ör. Sambur, 1978; Frey ve diğerleri, 2001). Örneğin spektral çıkarma (Boll, 1979) gibi algoritmalar aracılığıyla Gauss beyaz gürültüsünün silinmesi, gürültünün sınıflandırma üzerindeki olumsuz etkisini azaltmak için kullanılan etkili bir yöntemdir (Huang vd., 2013). Modelin, dalga veya gradyan seviyesinde toplamsal gürültüye sahip artırılmış bir veri kümesi ile eğitildiği, toplamsal gürültü bozulmalarına karşı gürbüzlük elde etmek için bir başka yöntem de çekişmeli eğitimdir (Latif ve diğerleri, 2018; Lakomkin ve diğerleri, 2018; Bao ve diğerleri, 2019; Latif ve diğerleri, 2020; Ren ve diğerleri, 2020a). Bilgimiz dahilinde, perde değişimi için çekişmeli eğitimin etkisine dair bir çalışma bulunmamaktadır.

CNN kurbanı hem beyaz gürültüye hem de perde bozulmalarına karşı savunmasızdır. Dalga formu tabanlı CNN'ler, farklı görevler için sinyalin farklı yönlerini öğrenir (Muckenhirn ve diğerleri, 2019). Bildiğimiz kadarıyla, dalga boyu

tabanlı CNN model katmanları, SER görevi bağlamında kodladıkları bilgi için analiz edilerek savunma yönleri belirlenebilir.

Genel olarak, saldırılar sorgu sayısı ve yürütme zamanı açısından hafiftir ve herhangi bir özel işlem gücü gerektirmemektedir. Kara kutu ortamında toplam kandırma oranı %79.16 olmuştur. Ancak insan değerlendiriciler, orijinal ve rakip örnekler arasındaki duygusal farklılıkları %8.34'lük bir hata oranıyla ayırt edememiştir. Bu bölümdeki deneylerde, kurban modeller sınırlı fonetik varyasyona sahip bir veri seti üzerinde eğitilmiştir. İki kipli sistemlere saldırı çalışmalarında, fonetik varyasyona sahip kurban model üzerinde incelendiğinde de ortak bozulma sistemi için perde kaymasının etkili bir yöntem olduğu görülmüştür.

6.3 İki-kipli Saldırı Sonuçları

İki-kipli diyalog sistemlerine karşı saldırı versiyonları incelenmiştir. Metin sınıflandırma kipi için kutupluluk tabanlı dilbilimsel bozulma yönteminde ek kısıt uygulanabilirliği fonetik duyarlılık sıralaması uygulamasıyla gösterilmiştir. Ses bozma sistemi için “segment değişimi yönelimli” ve “akustik yönelimli” olmak üzere iki ayrı yöntem incelenmiştir.

Segment değişimi yöneliminde metin bozma sisteminden alınan takas bilgisine dayalı sentezleme yapılmıştır. Üç strateji ile saldırı analizi gerçekleştirilmiştir: yeniden sentezlemeye dayalı konuşmacı uyumlu segment değişimi ve iki farklı genetik üretime dayalı korpus tabanlı ardışık konuşma sentezi (bağımsız genetik algoritma ve ortak bozulma optimizasyonu). Yeniden sentezleme için müdahale edilen her akustik alt birime ayrı sentez modelleri eğitilmesi gerekir. Bu stratejinin düşük sorgu sayısıyla başarılı saldırı yapılması amaçlanan kara kutu ortamında uygun olmadığı görülmüştür. Genetik üretimde korpus tabanlı ardışık konuşma sentezi, beyaz gürültü ve perde kaydırmaya dayalı yapı kombinasyonları kullanılmıştır. Ortak bozulma optimizasyonu ile dilbilimsel ve paralinguistik paralel skor yönteminin, daha hızlı başarılı sonuca ulaştırıldığı gözlemlenmiştir.

Segment değişimi yönelimi, kurban modelin yoğun olarak sorgulanmasını gerektirmektedir. Sözlü diyalog sistemi saldırı noktalarını dikkate aldığımızda, konuşma tanıma sistemine saldırma prosedürünün atlanması halinde bile metin

bozma sisteminden yapılan müdahale sayesinde son çıktının etkilenebileceği açıkça görülmektedir. Fonetik duyarlı metin bozma sistemi ve düşük iterasyonda başarılı saldırı gözlemlenen perde tabanlı akustik saldırı yöntemi bütünleşik olarak iki-kipli duygu tanıma sistemi kurban modeli üzerinde uygulanmıştır. Hedefli, uyarlanabilir hedefli ve ayrım gözetmeyen saldırı stratejileri değerlendirilmiştir. Bazı kaynak-hedef yönlerinde muhtemel ortak özellikler nedeniyle saldırı başarısının düşük olduğu görülmüştür (Neutral ve Sad). Ayrım gözetmeyen saldırılarda Sad sınıfına yoğun bir yönelim izlenirken, uyarlanabilir hedefli saldırıların saldırı maliyeti, başarısı ve yön esnekliği nedeniyle ayrım gözetmeyen saldırılara karşı üstün olduğu görülmüştür. Nispeten düşük saldırı maliyeti ve yüksek başarı ile, bu yöntemin gerçek kara-kutu ortamda uygulanabilirliği daha yüksektir.

6.4 Savunma Önerileri

Geliştirdiğimiz saldırı algoritmalarına karşı genel bir çözüm olarak özellikle iki konunun incelenmesini öneriyoruz. Xu ve diğ. (2018) temel bir modelin çekişmeli sağlamlığını güçlendirmek için güven bilgisini ve en yakın komşu aramasını birleştiren bir çerçeve olan Son Derece Güvenilir Yakın Komşu (HCNN, Highly Confident Near Neighbor) önermektedir. Bu yaklaşım, "iyi bir modelin kendine güvenen bölgeleri iyi ayrılmalıdır" prensibine dayalıdır. Güven azaltmaya dair geliştirdiğimiz saldırılar için değerlendirilmesi önemli bir savunma yöntemidir. Jha ve diğ. (2018), pertürbasyon saldırılarına karşı gürbüzlük ile yapay öğrenme modelleri tarafından oluşturulan bireysel kararların ilişkisini incelemiştir. Rakip girdilerin ilişkilendirme alanında gürbüz olmadığını, yani birkaç özelliği yüksek ilişkilendirmeyle maskelemenin, yapay öğrenme modelinin zararlı örnekler üzerinde kararsızlığa yol açtığını belirtmiştir. Gerçek örneklerin ise, öznitelik uzayında gürbüz olduğu görülmüştür. Bu sonuç kullanılarak yapay öğrenme modeli pertürbasyon saldırılarına karşı gürbüz hale getirilebilir.

KAYNAKÇA

Alzantot, M., Balaji, B., & Srivastava, M. (2018a). Did you hear that? adversarial examples against automatic speech recognition. *arXiv preprint arXiv:1801.00554*.

Alzantot, M., Sharma, Y. S., Elgohary, A., Ho, B. J., Srivastava, M., & Chang, K. W. (2018b, January). Generating natural language adversarial examples. In *Proceedings of the 2018 conference on empirical methods in natural language processing*.

Amin, T. B., & Mahmood, I. (2008, November). Speech recognition using dynamic time warping. In *2008 2nd international conference on advances in space technologies* (pp. 74-79). IEEE.

Ashoori, M., Briggs, B. D., Clevenger, L. A., & Clevenger, L. A. H. (2020). *U.S. Patent No. 10,580,435*. Washington, DC: U.S. Patent and Trademark Office.

Assal, H., & Chiasson, S. (2018). Security in the software development lifecycle. In *Fourteenth symposium on usable privacy and security ({SOUPS} 2018)* (pp. 281-296).

Bao, F., Neumann, M., & Vu, N. T. (2019). CycleGAN-based emotion style transfer as data augmentation for speech emotion recognition. In Interspeech (pp. 2828-2832).

Barreno, M., Nelson, B., Sears, R., Joseph, A. D., & Tygar, J. D. (2006, March). Can machine learning be secure?. In *Proceedings of the 2006 ACM symposium on information, computer and communications security* (pp. 16-25).

Biggio, B., Nelson, B., & Laskov, P. (2012, June). Poisoning attacks against support vector machines. In *Proceedings of the 29th international coference on international conference on machine learning* (pp. 1467-1474).

Bojanić, M., Delić, V., & Karpov, A. (2020). Call redistribution for a call center based on speech emotion recognition. *Applied Sciences*, *10*(13), 4653.

Boll, S. (1979). Suppression of acoustic noise in speech using spectral subtraction. *IEEE Transactions on acoustics, speech, and signal processing*, *27*(2), 113-120.

Busso, C., Bulut, M., Lee, C. C., Kazemzadeh, A., Mower, E., Kim, S., ... & Narayanan, S. S. (2008). IEMOCAP: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, *42*(4), 335-359.

Carlini, N., & Wagner, D. (2017, May). Towards evaluating the robustness of neural networks. In *2017 IEEE symposium on security and privacy (sp)* (pp. 39-57). IEEE.

Carlini, N., & Wagner, D. (2018, May). Audio adversarial examples: Targeted attacks on speech-to-text. In *2018 IEEE Security and Privacy Workshops (SPW)* (pp. 1-7). IEEE.

Carlini, N., Mishra, P., Vaidya, T., Zhang, Y., Sherr, M., Shields, C., ... & Zhou, W. (2016). Hidden voice commands. In *25th {USENIX} Security symposium ({USENIX} security 16)* (pp. 513-530).

Chakraborty, A., Alam, M., Dey, V., Chattopadhyay, A., & Mukhopadhyay, D. (2018). Adversarial attacks and defences: A survey. *arXiv preprint arXiv:1810.00069*.

Chazan, D., Hoory, R., Cohen, G., & Zibulski, M. (2000, June). Speech reconstruction from mel frequency cepstral coefficients and pitch frequency. In *Proceedings of the 2000 IEEE international conference on acoustics, speech, and signal processing. (Cat. No. 00CH37100)* (Vol. 3, pp. 1299-1302). IEEE.

Chen, Y., Nadji, Y., Kountouras, A., Monrose, F., Perdisci, R., Antonakakis, M., & Vasiloglou, N. (2017, October). Practical attacks against graph-based clustering. In *Proceedings of the 2017 ACM SIGSAC conference on computer and communications security* (pp. 1125-1142).

Ebrahimi, J., Rao, A., Lowd, D., & Dou, D. (2018, July). HotFlip: White-box adversarial examples for text classification. In *Proceedings of the 56th annual meeting of the association for computational linguistics (Volume 2: Short Papers)* (pp. 31-36).

Ellis, D. (2007, April 22). Chroma feature analysis and synthesis. Retrieved November 17, 2020, from https://labrosa.ee.columbia.edu/matlab/chroma-ansyn/

Elsayed, G. F., Goodfellow, I., & Sohl-Dickstein, J. (2018). Adversarial reprogramming of neural networks. *arXiv preprint arXiv:1806.11146*.

Farrús, M., Hernando, J., & Ejarque, P. (2007). Jitter and shimmer measurements for speaker recognition. In *the Eighth annual conference of the international speech communication association*.

Feng, J., Xu, H., Mannor, S., & Yan, S. (2014). Robust logistic regression and classification. *Advances in neural information processing systems*, *27*, 253-261.

France, D. J., Shiavi, R. G., Silverman, S., Silverman, M., & Wilkes, M. (2000). Acoustical properties of speech as indicators of depression and suicidal risk. *IEEE transactions on biomedical engineering*, *47*(7), 829-837.

Frey, B. J., Deng, L., Acero, A., & Kristjansson, T. (2001). ALGONQUIN: Iterating Laplace's method to remove multiple types of acoustic distortion for robust speech recognition. In T*he seventh european conference on speech communication and technology*.

Garrido-Muñoz, I., Montejo-Ráez, A., Martínez-Santiago, F., & Ureña-López, L. A. (2021). A survey on bias in deep NLP. *Applied Sciences*, *11*(7), 3184.

Gong, Y., & Poellabauer, C. (2017). Crafting adversarial examples for speech paralinguistics applications. *arXiv preprint arXiv:1711.03280*.

Gong, Z., Wang, W., Li, B., Song, D., & Ku, W. S. (2018). Adversarial texts with gradient methods. *arXiv preprint arXiv:1801.07175*.

Harte, C., Sandler, M., & Gasser, M. (2006). "Detecting harmonic change in musical audio." In Proceedings of the 1st ACM workshop on audio and music computing multimedia (pp. 21-26). Santa Barbara, CA, USA: ACM Press. doi:10.1145/1178723.1178727

Hernandez, S. (2019). *U.S. Patent No. 10,404,859*. Washington, DC: U.S. Patent and Trademark Office.

Hixon, B., Schneider, E., & Epstein, S. L. (2011). Phonemic similarity metrics to compare pronunciation methods. In *Twelfth annual conference of the ınternational speech communication association*.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, *9*(8), 1735-1780.

Homer, N., Szelinger, S., Redman, M., Duggan, D., Tembe, W., Muehling, J., ... & Craig, D. W. (2008). Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. *PLoS Genet*, *4*(8), e1000167.

Huang, C., Chen, G., Yu, H., Bao, Y., & Zhao, L. (2013). Speech emotion recognition under white noise. *Archives of acoustics*, *38*(4), 457-463.

Huang, S., Papernot, N., Goodfellow, I., Duan, Y., & Abbeel, P. (2017). Adversarial attacks on neural network policies. *arXiv preprint arXiv:1702.02284*.

Issa, D., Demirci, M. F., & Yazici, A. (2020). Speech emotion recognition with deep convolutional neural networks. *Biomedical signal processing and control*, *59*, 101894.

Ittichaichareon, C., Suksri, S., & Yingthawornsuk, T. (2012, July). Speech recognition using MFCC. In *International conference on computer graphics, simulation and modeling* (pp. 135-138).

Jia, R., & Liang, P. (2017, September). Adversarial examples for evaluating reading comprehension systems. In *Proceedings of the 2017 Conference on empirical methods in natural language processing* (pp. 2021-2031).

Jia, Y., Zhang, Y., Weiss, R., Wang, Q., Shen, J., Ren, F., ... & Wu, Y. (2018). Transfer learning from speaker verification to multispeaker text-to-speech synthesis. In *Advances in neural information processing systems* (pp. 4480-4490).

Kalchbrenner, N., Elsen, E., Simonyan, K., Noury, S., Casagrande, N., Lockhart, E., ... & Kavukcuoglu, K. (2018). Efficient neural audio synthesis. *arXiv preprint arXiv:1802.08435*.

Kamaruddin, N., & Wahab, A. (2010, June). Driver behavior analysis through speech emotion understanding. In *2010 IEEE Intelligent vehicles symposium* (pp. 238-243). IEEE.

Kibriya, A. M., Frank, E., Pfahringer, B., & Holmes, G. (2004, December). Multinomial naive bayes for text categorization revisited. In *Australasian joint conference on artificial intelligence* (pp. 488-499). Springer, Berlin, Heidelberg.

Kreuk, F., Adi, Y., Cisse, M., & Keshet, J. (2018, April). Fooling end-to-end speaker verification with adversarial examples. In *2018 IEEE International conference on acoustics, speech and signal processing (ICASSP)* (pp. 1962-1966). IEEE.

Kurakin, A., Goodfellow, I., & Bengio, S. (2016a). Adversarial machine learning at scale. *arXiv preprint arXiv:1611.01236*.

Kurakin, A., Goodfellow, I., & Bengio, S. (2016b). Adversarial examples in the physical world. In Yampolskiy, R.V. (Ed.), *Artificial intelligence safety and security*, (pp. 99-112). doi: 10.1201/9781351251389

Laishram, R., & Phoha, V. V. (2016). Curie: A method for protecting svm classifier from poisoning attack. *arXiv preprint arXiv:1606.01584*.

Lakomkin, E., Zamani, M. A., Weber, C., Magg, S., & Wermter, S. (2018, October). On the robustness of speech emotion recognition for human-robot interaction with deep neural networks. In *2018 IEEE/RSJ International conference on intelligent robots and systems (IROS)* (pp. 854-860). IEEE.

Latif, S., Rana, R., & Qadir, J. (2018). Adversarial machine learning and speech emotion recognition: Utilizing generative adversarial networks for robustness. *arXiv preprint arXiv:1811.11402*.

Latif, S., Rana, R., Khalifa, S., Jurdak, R., & Schuller, B. W. (2020). Deep architecture enhancing robustness to noise, adversarial attacks, and cross-corpus setting for speech emotion recognition. *arXiv preprint arXiv:2005.08453*.

Li, B., Wang, Y., Singh, A., & Vorobeychik, Y. (2016). Data poisoning attacks on factorization-based collaborative filtering. *Advances in neural information processing systems*, 1893-1901.

Li, Y., Su, H., Shen, X., Li, W., Cao, Z., & Niu, S. (2017, November). DailyDialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the eighth international joint conference on natural language processing (Volume 1: Long Papers)* (pp. 986-995).

Li, J., Ji, S., Du, T., Li, B., & Wang, T. (2019, January). TextBugger: Generating adversarial text against real-world applications. In *26th Annual network and distributed system security symposium*.

Lin, Y. C., Hong, Z. W., Liao, Y. H., Shih, M. L., Liu, M. Y., & Sun, M. (2017a, August). Tactics of adversarial attack on deep reinforcement learning agents. In *Proceedings of the 26th international joint conference on artificial intelligence* (pp. 3756-3762).

Lin, Y. C., Liu, M. Y., Sun, M., & Huang, J. B. (2017b). Detecting adversarial attacks on neural network policies with visual foresight. *arXiv preprint arXiv:1710.00814*.

Lipner, S. (2004, December). The trustworthy computing security development lifecycle. In *20th Annual computer security applications conference* (pp. 2-13). IEEE.

Liu, C., Li, B., Vorobeychik, Y., & Oprea, A. (2016a). Robust high-dimensional linear regression. *arXiv preprint arXiv:1608.02257*.

Liu, Q., Li, P., Zhao, W., Cai, W., Yu, S., & Leung, V. C. (2018a). A survey on security threats and defensive techniques of machine learning: A data driven view. *IEEE access*, *6*, 12103-12117.

Liu, K., Li, Y., Xu, N., & Natarajan, P. (2018b). Learn to combine modalities in multimodal deep learning. *arXiv preprint arXiv:1805.11730*.

Liu, Y., Chen, X., Liu, C., & Song, D. (2016b). Delving into transferable adversarial examples and black-box attacks. *arXiv preprint arXiv:1611.02770*.

Livingstone, S. R., & Russo, F. A. (2018). The ryerson audio-visual database of emotional speech and song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. *PloS one*, *13*(5), e0196391.

Martinez, J., Perez, H., Escamilla, E., & Suzuki, M. M. (2012, February). Speaker recognition using Mel frequency Cepstral Coefficients (MFCC) and Vector quantization (VQ) techniques. In *Conielecomp 2012, 22nd International conference on electrical communications and computers* (pp. 248-251). IEEE.

McAulay, R. J., & Quatieri, T. F. (1995). *Sinusoidal coding* (No. MS-11427). Massachusetts Inst Of Tech Lexington Lincoln Lab.

Miao, C., Li, Q., Xiao, H., Jiang, W., Huai, M., & Su, L. (2018, June). Towards data poisoning attacks in crowd sensing systems. In *Proceedings of the eighteenth ACM international symposium on mobile ad hoc networking and computing* (pp. 111-120).

Miyato, T., Dai, A. M., & Goodfellow, I. (2016). Adversarial training methods for semi-supervised text classification. *arXiv preprint arXiv:1605.07725*.

Mohammed, N. M., Niazi, M., Alshayeb, M., & Mahmood, S. (2017). Exploring software security approaches in software development lifecycle: A systematic mapping study. *Computer standards & interfaces*, *50*, 107-115.

Moosavi-Dezfooli, S. M., Fawzi, A., & Frossard, P. (2016). Deepfool: a simple and accurate method to fool deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2574-2582).

Muckenhirn, H., Abrol, V., Magimai-Doss, M., & Marcel, S. (2019). Understanding and visualizing raw waveform-based CNNs. In *Interspeech* (pp. 2345-2349).

Papernot, N., & McDaniel, P. (2017). Extending defensive distillation. *arXiv preprint arXiv:1705.05264*.

Papernot, N., McDaniel, P., & Goodfellow, I. (2016b). Transferability in machine learning: from phenomena to black-box attacks using adversarial samples. *arXiv preprint arXiv:1605.07277*.

Papernot, N., McDaniel, P., Goodfellow, I., Jha, S., Celik, Z. B., & Swami, A. (2016c). Practical black-box attacks against deep learning systems using adversarial examples. *arXiv preprint arXiv:1602.02697*, *1*(2), 3.

Papernot, N., McDaniel, P., Jha, S., Fredrikson, M., Celik, Z. B., & Swami, A. (2016a, March). The limitations of deep learning in adversarial settings. In *2016 IEEE European symposium on security and privacy (EuroS&P)* (pp. 372-387). IEEE.

Papernot, N., McDaniel, P., Swami, A., & Harang, R. (2016e, November). Crafting adversarial input sequences for recurrent neural networks. In *MILCOM 2016-2016 IEEE Military communications conference* (pp. 49-54). IEEE.

Papernot, N., McDaniel, P., Wu, X., Jha, S., & Swami, A. (2016d, May). Distillation as a defense to adversarial perturbations against deep neural networks. In *2016 IEEE symposium on security and privacy (SP)* (pp. 582-597). IEEE.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*.

Pierre-Yves, O. (2003). The production and recognition of emotions in speech: features and algorithms. *International journal of human-computer studies*, *59*(1-2), 157-183.

Přibil, J., & Přibilová, A. (2009). Spectral flatness analysis for emotional speech synthesis and transformation. In *Cross-modal analysis of speech, gestures, gaze, and facial expressions* (pp. 106-115). Springer, Berlin, Heidelberg.

Ren, Z., Baird, A., Han, J., Zhang, Z., & Schuller, B. (2020a, May). Generating and protecting against adversarial attacks for deep speech-based emotion recognition models. In *ICASSP 2020-2020 IEEE International conference on acoustics, speech and signal processing (ICASSP)* (pp. 7184-7188). IEEE.

Ren, Z., Han, J., Cummins, N., & Schuller, B. W. (2020b). Enhancing transferability of black-box adversarial attacks via lifelong learning for speech emotion recognition models. *Proc. Interspeech 2020*, 496-500.

Robel, A. (2006). Signal modifications using the STFT. *Summer 2006 lectrue on analysis, modeling, and transformation of audio signals, institute of communication science TU-Berlin*, 1-73.

Sahu, G. (2019). Multimodal speech emotion recognition and ambiguity resolution. *arXiv preprint arXiv:1904.06022*.

Sahu, S., Mitra, V., Seneviratne, N., & Espy-Wilson, C. Y. (2019, September). Multi-modal learning for speech emotion recognition: An analysis and comparison of asr outputs with ground truth transcription. In *Interspeech* (pp. 3302-3306).

Samanta, S., & Mehta, S. (2017). Towards crafting text adversarial samples. *arXiv preprint arXiv:1707.02812*.

Sambur, M. (1978). Adaptive noise canceling for speech signals. *IEEE Transactions on acoustics, speech, and signal processing*, *26*(5), 419-423.

Sato, M., Suzuki, J., Shindo, H., & Matsumoto, Y. (2018, January). Interpretable adversarial perturbation in input embedding space for text. In *27th International joint conference on artificial intelligence, IJCAI 2018* (pp. 4323-4330).

Schuller, B., Steidl, S., Batliner, A., Burkhardt, F., Devillers, L., MüLler, C., & Narayanan, S. (2013). Paralinguistics in speech and language—State-of-the-art and the challenge. *Computer speech & language*, *27*(1), 4-39.

Schuster, M., & Paliwal, K. K. (1997). Bidirectional recurrent neural networks. *IEEE transactions on signal processing*, *45*(11), 2673-2681.

Shao, X., & Milner, B. (2004, May). Pitch prediction from MFCC vectors for speech reconstruction. In *2004 IEEE International conference on acoustics, speech, and signal processing* (Vol. 1, pp. I-97). IEEE.

Shen, J., & Souza, P. E. (2018). On dynamic pitch benefit for speech recognition in speech masker. *Frontiers in psychology*, *9*, 1967.

Shokri, R., Strobel, M., & Zick, Y. (2019). On the privacy risks of model explanations. *arXiv preprint arXiv:1907.00164*.

Song, Y., Kim, T., Nowozin, S., Ermon, S., & Kushman, N. (2018, February). PixelDefend: Leveraging generative models to understand and defend against adversarial examples. In *International conference on learning representations*.

Steinhardt, J., Koh, P. W., & Liang, P. (2017, December). Certified defenses for data poisoning attacks. In *Proceedings of the 31st international conference on neural information processing systems* (pp. 3520-3532).

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., & Fergus, R. (2013). Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*.

Taori, R., Kamsetty, A., Chu, B., & Vemuri, N. (2019, May). Targeted adversarial examples for black box audio systems. In *2019 IEEE Security and privacy workshops (SPW)* (pp. 15-20). IEEE.

Teixeira, J. P., & Fernandes, P. O. (2014). Jitter, Shimmer and HNR classification within gender, tones and vowels in healthy voices. *Procedia technology*, *16*, 1228-1237.

Teixeira, J. P., Oliveira, C., & Lopes, C. (2013). Vocal acoustic analysis–jitter, shimmer and hnr parameters. *Procedia technology*, *9*, 1112-1122.

Thiemann, J., Ito, N., & Vincent, E. (2013). The diverse environments multi-channel acoustic noise database: A database of multichannel environmental noise recordings. *The Journal of the acoustical society of america*, *133*(5), 3591-3591.

Tokuda, K. (2019). Interspeech 2019 conference keynote presentation. URL: *https://www.sp.nitech.ac.jp/~tokuda/ınterspeech2019.pdf* Erişim: Haziran 2020.

Tóth, S. L., Sztahó, D., & Vicsi, K. (2008). Speech emotion perception by human and machine. In *Verbal and nonverbal features of human-human and human-machine interaction* (pp. 213-224). Springer, Berlin, Heidelberg.

Tramèr, F., Boneh, D., Kurakin, A., Goodfellow, I., Papernot, N., & McDaniel, P. (2018). Ensemble adversarial training: Attacks and defenses. In *6th International conference on learning representations, ICLR 2018-conference track proceedings*.

Tripathi, S., Tripathi, S., & Beigi, H. (2018). Multi-modal emotion recognition on iemocap dataset using deep learning. *arXiv preprint arXiv:1804.05788*.

Vijayaraghavan, P., & Roy, D. (2019, September). Generating black-box adversarial examples for text classifiers using a deep reinforced model. In *Joint european conference on machine learning and knowledge discovery in databases* (pp. 711-726). Springer, Cham.

Wan, L., Wang, Q., Papir, A., & Moreno, I. L. (2018, April). Generalized end-to-end loss for speaker verification. In *2018 IEEE International conference on acoustics, speech and signal processing (ICASSP)* (pp. 4879-4883). IEEE.

Wang, B., & Gong, N. Z. (2018, May). Stealing hyperparameters in machine learning. In *2018 IEEE Symposium on security and privacy (SP)* (pp. 36-52). IEEE.

Wei, W., Liu, L., Loper, M., Truex, S., Yu, L., Gursoy, M. E., & Wu, Y. (2018). Adversarial examples in deep learning: Characterization and divergence. *arXiv preprint arXiv:1807.00051*.

Wu, X., Jang, U., Chen, J., Chen, L., & Jha, S. (2018, July). Reinforcing adversarial robustness using model confidence induced by adversarial training. In *International conference on machine learning* (pp. 5334-5342). PMLR.

Wu, Y., Liu, J., Chen, Y., & Cheng, J. (2019, November). Semi-black-box attacks against speech recognition systems using adversarial samples. In *2019 IEEE International symposium on dynamic spectrum access networks (DySPAN)* (pp. 1-5). IEEE.

Yoon, S., Byun, S., & Jung, K. (2018, December). Multimodal speech emotion recognition using audio and text. In *2018 IEEE Spoken language technology workshop (SLT)* (pp. 112-118). IEEE.

Zhao, J., Mao, X., & Chen, L. (2019). Speech emotion recognition using deep 1D & 2D CNN LSTM networks. *Biomedical signal processing and control*, *47*, 312-323.

A
EK TABLOLAR

Tablo A. 1 Hedefli metin pertürbasyonlarında takas frekansları

Orijinal	Takas Edilen	Frekans
"Angry" Yönünde Saldırılar		
İs	does	64
not	indeed	48
know	are	43
do	are	38
are	am	38
"Excited" Yönünde Saldırılar		
is	wants	64
not	forward	48
is	does	41
am	are	36
what	who	36
"Frustrated" Yönünde Saldırılar		
is	takes	52
are	am	38
know	am	34
do	am	32
not	everywhere	28
"Happy" Yönünde Saldırılar		
do	are	87
is	wants	81
am	are	73
not	far	66
is	makes	65
"Neutral" Yönünde Saldırılar		
is	does	57
are	am	48
what	who	47
is	takes	44
am	are	33
"Sad" Yönünde Saldırılar		
is	gets	67
what	who	51
not	definitely	39
not	absolutely	34
do	are	31

Tablo A. 2 Ayrım gözetmeyen metin pertürbasyonlarında takas frekansları

Orijinal	Takas Edilen	Frekans
is	does	90
not	alone	71
are	am	48
what	who	44
is	takes	30

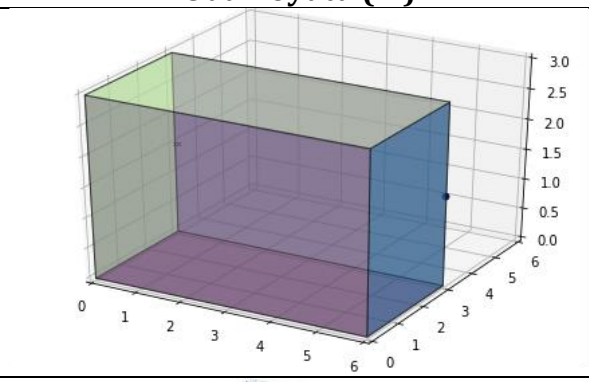
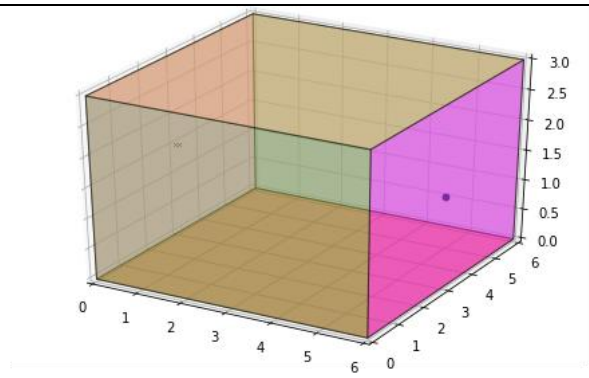
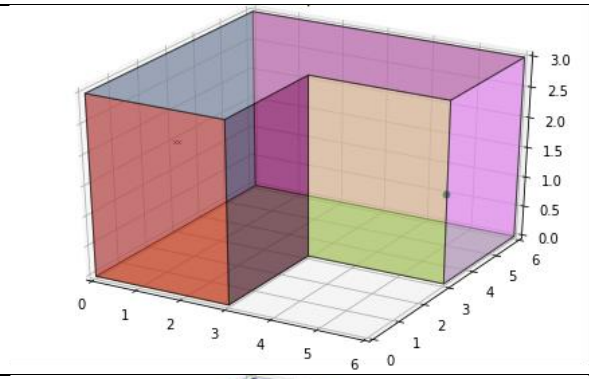
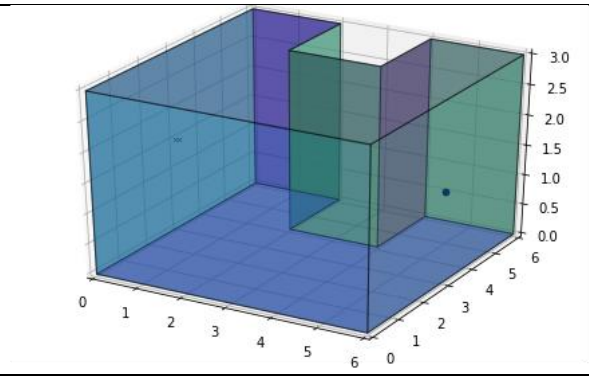
Tablo A. 3 Uyarlanabilir hedefli metin pertürbasyonlarında takas frekansları

Orijinal	Takas Edilen	Frekans
do	are	30
am	are	29
what	who	28
is	makes	28
is	says	25

Tablo A. 4 İki-kipli duygu tanıma kurban modeli hiper-parametreleri

Metin Duygu Sınıflandırma	Konuşma Duygu Sınıflandırma
Sözlük boyutu: 27593 Maksimum örnek uzunluğu: 82 Batch büyüklüğü: 49 Gömme uzunluğu: 127 BiLSTM hücre sayısı: 20 Dropout: 0.55 Epoch: 47	Girdi boyutu: 128000 CNN Katman Boyutları: 64,64,128,128 LSTM Hücre Sayısı: 256 CNN kernel boyutu ve stride: 3 ve 1 Ortaklama kernel boyutu ve stride: 4 ve 4 Batch momentum=0.99 Batch epsilon=0.001 ELU alpha=0.9

Tablo A. 5 Oda özellikleri

Oda Boyutu (m)	Ortak Özellikler
	**Oda Yüksekliği** 3 m **Mikrofon Konumu (Sol ve Sağ)** (0, 3, 1.5) ve (1, 3, 1.5) **Kaynak (Hoparlör) Konumu** (6, 3, 1.5) **Sinyal Gecikmesi** 0 sn **Duvar Materyali Sinyal Emilimi** 0.35
	
	
	

TEZDEN ÜRETİLMİŞ YAYINLAR

Makale

1. Gümüş, F., & Amasyalı, M. F. (2021, Just Accepted). Exploiting Natural Language Services: A Polarity Based Black-box Attack. *Frontiers of Computer Science*. Retrieved from https://journal.hep.com.cn/fcs/EN/10.1007/s11704-021-0198-y