Avrupa Bilim ve Teknoloji Dergisi
Sayı 23, S. 787-792, Nisan 2021

Araştırma Makalesi

www.ejosat.com ISSN:2148-2683

European Journal of Science and Technology
No. 23, pp. 787-792, April 2021

Research Article

Üretken Rakip Ağ ile Türkçe Metin Üretimi

Barış Gücük[1*], Rafet Durgut[2], Oğuz Fındık[3]

[1*] Karabük Üniversitesi, Mühendislik Fakültesi, Bilgisayar Mühendisliği Bölümü, Karabük, Türkiye, (ORCID: 0000-0002-1381-3663), barisgucuk@ogrenci.karabuk.edu.tr
[2] Karabük Üniversitesi, Mühendislik Fakültesi, Bilgisayar Mühendisliği Bölümü, Karabük, Türkiye (ORCID: 0000-0002-6891-5851), rafetdurgut@karabuk.edu.tr
[3] Karabük Üniversitesi, Mühendislik Fakültesi, Bilgisayar Mühendisliği Bölümü, Karabük, Türkiye (ORCID: 0000-0001-5069-6470), oguzfindik@karabuk.edu.tr

(İlk Geliş Tarihi 9 Ocak 2021 ve Kabul Tarihi 24 Nisan 2021)

(**DOI:** 10.31590/ejosat.857179)

ATIF/REFERENCE: Gücük, B., Durgut, R. & Fındık, O. (2021). Üretken Rakip Ağ ile Türkçe Metin Üretimi. *Avrupa Bilim ve Teknoloji Dergisi*, (23), 787-792.

Öz

Makine öğrenmesi yöntemlerinde tahmin aşamasının başarısı için kullanılan eğitim veri seti kümesi oldukça önemlidir. Doğal dil işlemede en çok karşılaşılan problemlerden birisi yeterli veri bulunamaması veya bulunan verilerin etiketsiz olmasıdır. Özellikle sınıflandırma problemlerinde belirli bir sınıftaki verinin azlığı sınıflandırmanın başarısını düşürmektedir. Bu çalışmada veri kümesinde bulunan eksik sınıfa ait metinlerin arttırılması amacı ile üretken rakip ağlar yöntemi kullanılmıştır. Haber metinleri üzerinde veri çoğalma işlemi gerçekleştirilmiştir. Elde edilen sonuçlar n-gram, destek vektör makinesi, TF-IDF ve lojistik regresyon gibi makine öğrenmesi teknikleriyle birlikte kullanılarak performansları değerlendirilmiştir. Sonuçlara göre üretken rakip ağların Türkçe metin üretimi için kullanılması sınıflandırma başarısını yaklaşık % 47 oranında arttırmıştır.

Anahtar Kelimeler: Doğal dil işleme, Üretken rakip ağlar, Metin üretimi, Sınıflandırma.

Turkish Text Generation with Generative Adversarial Network

Abstract

The training data set used for the success of the training phase in machine learning methods is very important. One of the most common problems in natural language processing is the lack of sufficient data or the unlabeled data. Especially in classification problems, the scarcity of data in a certain class reduces the success of the classification. In this study, generative adversarial network method was used in order to increase the texts belonging to the missing class in the data set. Data augmentation is performed on news texts. The results obtained were evaluated together with machine learning techniques such as n-grams, support vector machine, TF-IDF and logistic regression. According to the results, the use of generative adversarial network for Turkish text generation increased the classification success by approximately % 47.

Keywords: Natural language processing, Generative adversarial networks, Text generation, Classification.

[*] Sorumlu Yazar: barisgucuk@ogrenci.karabuk.edu.tr

1. Giriş

Makine öğrenmesi, bilgisayarın tecrübelerden çıkarım yapabilmesi için kullanılan yapay zekâ yöntemidir (Michie ve ark., 1994). Makine öğrenmesi, danışmanlı, danışmansız veya pekiştirmeli öğrenme olmak üzere olarak üç kategoriye ayrılmaktadır (Ayon, 2016). Danışmanlı öğrenmede kullanılacak modelin eğitimi için etiketli verilere ihtiyaç duyulmaktadır (Xiaojin, 2005). Etiketli verinin boyutu modelin tahmin başarısını doğrudan etkilemektedir. Verilerin etiketlenmesi konuyla ilgili uzman bir kişi tarafından yapılmalıdır. Etiketli verinin sayısı da modelin performansını etkilese de etiketli veri bulunması özellikle doğal dil işleme alanında oldukça zordur (Jun ve ark., 2008). Çünkü verilerin etiketlenmesi uzun süre alan maliyetli bir işlemdir. Etiketleme işlemini yapan kişinin uzmanlığı gereklidir. Etiketleme sırasındaki insan hataları modelin başarısını etkilemektedir. Bazı verilerde gizlilik gerekebilir bu gibi durumlarda etiketleme yapılamamaktadır.

Üretken Rakip Ağlar (ÜRA) 2014 yılında Ian Goodfellow vd. tarafından önerilmiş bir makine öğrenmesi yöntemidir (Goodfellow ve ark., 2014). Bu yöntem, iki sinir ağının birlikte çalışmasıyla meydana gelmektedir. Bir ağ üretim ile sorumlu iken diğer ağ ayırt edici olarak çalışmaktadır. Bir dengede çalışan bu iki ağ eğitilmesi sonrasında eğitim setinden farklı özgün yeni resimler, sesler üretilebilir. Ne kadar dengeli bir sistem kurulursa gerçeğe o kadar yakın sonuçlar elde edilir.

Oldukça yeni bir yöntem olmasına karşın üretken rakip ağlar üzerindeki çalışmaların sayısı hızla artmaktadır. Yapılan çalışmalarda gerçeğe çok yakın görüntüler üretilmiştir (Zhang ve ark., 2019). Buradaki başarıdan yola çıkarak, üretken rakip ağlar ile kurabilecek bir modelin metin üretimi gibi bir doğal dil işleme uygulamasında kullanılabileceği akla gelmektedir. Türkçe metin üretimlerinde daha çok Tekrarlayan Sinir Ağı (TSA) ve Uzun Kısa Süreli Bellek (UKSB) kullanılmaktadır. Bu yöntem ile üretilen metinlerin anlam yapısı bakımından başarısı yüksek değildir (Santhanam, 2018).

Metnin ayrık ve soyut doğası gereği üretken rakip ağlar ile metin üretimi aşamasında bazı problemler ile karşılaşılmaktadır. Burada akla gelen ilk çözümlerden biri pekiştirmeli öğrenme uygulamaktır (Wang ve ark., 2019).

Tong Che vd. tarafından 2017 yılında yayınlanan makalesinde üretken rakip ağlardaki kararsızlığı yok etmek için maksimum olabilirliği arttırılmış üretken rakip ağları (MaliGAN) önermiştir (Che ve ark., 2017). Bu yöntemde hedefi optimize etmektense ayırıcı sonuçları kullanılarak yeni bir hedef üretilir. Jiaxian Guo vd. yine 2017 yılında LeakGAN ile sızan bilgilerle uzun metin üretimi modelini yayınlamıştır (Guo ve ark., 2017). Bu modelde ayırıcı üreticiye daha fazla bilgi yönlendirmesine izin verilerek daha başarılı ve anlamını kaybetmeyen uzun metinler üretilebileceği ortaya koyulmuştur. SeqGAN eğim tedbirli bir sıra oluşturucudur. Lantao Yu vd. tarafından 2017 yılında yayınlanan makalesinde üretken rakip ağın eğitiminde pekiştirmeli öğrenme uygulamıştır (Yu ve ark., 2017). Üreticinin sınıflandırma problemlerini atlayıp direk olarak eğim tedbirli bir şekilde eğitimine devam etmektedir. Kevin Lin vd. 2017 yılında yaptığı çalışmada RankGAN ile ayırıcıyı sınıflandırmak için eğitmek yerine sıralama ve bir referans grubu oluşturacak şekilde değiştirmiştir (Lin ve ark., 2017). Daha sonrasında puanlandırma sistemi ile daha iyi değerlendirme yaptığını göstermiştir. William Fedus vd. 2018 yılında kelimelerin bir önceki kelimeye göre koşullandırılması yerine maksimum olasılık yöntemi ile eğitilir (Fedus ve ark., 2018). MaskGAN metodu ile üretici daha kaliteli sonuçlar üretilebileceği gösterilmiştir. Üstteki yöntemler performans olarak başarılı olsalar dahi optimizasyon aşamalarında zorluk yaşamaktadır. Zhang vd. 2017 yılında yaptığı çalışmada TextGAN diğer modellerin aksine pekiştirmeli öğrenme içermeden metin üretimi yapmıştır (Cao ve ark., 2017). Matt Kusner vd. 2016 yılında yayınladığı çalışmasında üretken rakip ağ üzerinde Gumbel-softmax dağılımı kullanarak parametrelerin farklılaşması engellenmiştir (Kusner ve ark., 2016). Böylelikle eğitimin hızı ve istikrarı artmıştır.

Türkçe üzerinde doğal dil işleme ile alakalı birçok çalışma bulunmaktadır. Bu çalışmalardan sınıflandırma konusunu ele alan çalışmalar incelendiğinde;

2019 yılında Tarık Ş. vd. doğal dil işleme alanında kullanılmak üzere Türkçe veri seti oluşturma çalışması yayınlamıştır (Tarık ve ark., 2019). Metin B. tarafından yapılan çalışmada köşe yazarlarının ait olduğu yazarın tahmini için bigram ve trigram ile LZW algoritmasından yararlanılmıştır (Metin, 2019).

Üretken rakip ağlar ile veri çoğaltmayla ilgili çalışmalar incelendiğinde;

Antreas Antoniou vd. 2017 yılında yayınladığı çalışmasında veri artırmak için üretken rakip ağ kullanmıştır. Çalışmada kullandığı DAGAN ile sınıflandırmada başarı artışı gözlenmiştir (Antreas ve ark., 2018). Georgios Douzas vd. 2018 yılında normal dağılımlı olmayan veri seti üzerinde sınıflandırma problemini çözmek için üretken rakip ağ kullanmıştır. Çalışmada kullandığı cGAN ile sınıflandırma başarısını artırmıştır (Georgios ve ark., 2018).

Bu çalışmada üretken rakip ağlar kullanılarak normal dağılımlı olmayan bir veri seti üzerinde eksik sınıfa yönelik Türkçe metin üretimi işlemi yapılmıştır.

2. Materyal ve Metot

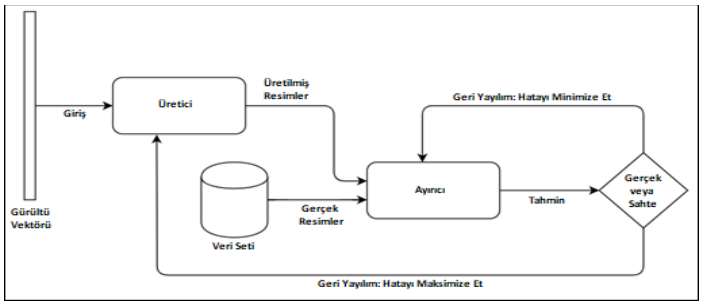
2.1. Üretken Rakip Ağlar

Üretken rakip ağlar bir makine öğrenmesi yöntemidir. İlk olarak 2014 yılında Ian Goodfellow vd. tarafından tanıtılmıştır (Goodfellow ve ark., 2014). Yapısı iki sinir ağının birbirine karşı çalışacak şekilde oluşturulmasından meydana gelmektedir. Bu iki ağ arasında sıfır toplamlı oyun Nash Dengesi (Maskin, 1999) vardır. Bunu bir ağın kazanması için diğerinin kaybetmesi gerekliliği gibi düşünülebilir. Her iki ağ aynı anda kazanamaz. Bu durumda toplam kazanç ve toplam kayıp toplamı sıfır çıkmalıdır. Zero-sum game problemleri min-max teoremi ile çözülmektedir (Gharesifard ve ark., 2013).

Modelin çalışma mantığından bahsetmek gerekirse, bir sinir ağı üretici diğer sinir ağı ise ayırıcı olarak adlandırılmaktadır. Bir ön eğitimden sonra üretici eğitim setinden öğrenmeye başlar ve bu öğrendiklerinden tahminlerde bulunur. Ayırıcı ise bu çıktıları eleyerek üreticiyi istenilen gerçeğe yakın çıktılara yönlendirmeye çalışır. Bunu sıcak soğuk oyunu gibi düşünürsek istenilen çıktıdan uzaklaşıldığında ayırıcı üreticiye soğuk şeklinde geri bildirimde bulunur. Bu durumda üretici geri yayılım kullanarak hatasını minimalize etmeye çalışır ve ağırlıklarını günceller. Diğer durumda ayırıcı üreticiden gelen çıktıların gerçeğe yaklaştığı fark eder ve bu durumda sıcak şeklinde geri bildirimde bulunur. Üretici ise bu durumda kazancını maksimize etmeye çalışarak en

iyi sonuçları üretmek için geri yayılım kullanarak ağırlıklarını günceller.

Üretici gerçeğe yakın sonuçlar üretmeye başladığında ayırıcı üretilenler ve gerçek olanlar arasında ayırım yapmakta zorlanır. Bu durumda ayırıcı da kendi ağırlıklarını geri yayılım algoritması kullanarak günceller. Böylece üretilenler ve gerçek olanları ayırt etmede daha yetenekli olur.



Şekil 1. Üretken rakip ağlar için akış diyagramı.

Üretken rakip ağların yapısı algoritma ve akış diyagramından anlaşılabileceği gibi bir gürültü vektörü ile başlar. Gürültü ile başlandığında sistemin Nash dengesine gelmesi uzun sürebildiğinden veri setinin bir kısmından ön eğitim yapılabilir. Bu hem öğrenme süresini hızlandırır hem de sistemin kararlılığının arttırır. Bu işlemden sonra küçük gruplar halinde çıktılar alınır ve ayırıcıya gönderilir. İki sinir ağının ağırlıkları geri yayılım ile güncellenir. Daha sonrasında sonuç olarak sistemin Nash dengesine ulaşması ve en gerçeğe yakın çıktılar üretmesi hedeflenir. Akış diyagramı Şekil 1'deki gibidir.

2.2. Uygulama

Bu çalışmada üretken rakip ağlar kullanılarak kendisine verilen haber metinlerinden yeni Türkçe metinler üretmek amaçlanmıştır. Kullanılan veri seti 2017'nin sonbaharında internet üzerindeki Türkçe haber sitelerinden toplanmış olup bu veri seti 3058 haber ve 832 bin 302 kelimeden oluşmaktadır. Veri setinden bir kesit Tablo 1'de verilmiştir. Bu haberler iki kategoriye ayırılmıştır. Bu kategoriler anlamına göre olumlu haberler ve olumsuz haberler şeklindedir. Bu kategorileme sonucunda 2 bin 949 olumlu habere karşılık 109 olumsuz haber bulunmaktadır. Tahmin edileceği üzerine veri seti üzerindeki bu eşit olmayan dağılım sınıflandırmayı negatif olarak etkilemektedir.

Tablo 1. Haber metinlerinin ve sınıflandırma sonuçlarının bulunduğu veri setinden bir kesit.

No	Sınıf	Veri
61	Olumlu	Başkan'dan yeni yıl ziyaretleri Belediye Başkanı ilçedeki bankaların yöneticilerine yeni yıl ziyaretinde bulundu. Başkan yeni yıl ziyaretleri kapsamında İlçede faaliyet gösteren Işbank Müdürü, Garanti Bankası Müdürü, Vakıfbank Müdürü ve TEB Bankası Müdürünü ziyaret ederek 201 8 yılını başarı dostluk ve mutluluk içerisinde geçmesini diledi
62	Olumsuz	Tünelden geçmedik çünkü... TRAFİK KİLİT AMA AVRASYA YİNE BOŞ M Avrasya Tiincli'nden geçişlerin beklenenin altında kalması İstanbul dünya trafik yoğunluğunda ilk 10'da yer almasına rağmen sürücüler neden tüneli tercih etmedi?' sorusunu gündeme getirdi. Gözler, tek yönde 16 TL'lik geçiş ücretine çevrildi. Tüneldeki ikinci krizi vatandaş şikayetleri ortaya koydu: M OGS'de para olmasına rağmen ceza yazıldı. M 'Plakaya ait ceza yoktur' yazısından üç gün sonra iki geçiş için 10 kat ceza kesildi. S/5
63	Olumlu	En çok dolar konuşuldu ama zirve borsanın Dünya ekonomisindeki toparlanma, 2018'de şirket karlarına, dolayısıyla borsa ve emtialara yarayacak Piyasalar, oynaklığın yüksek, sürprizlerin bol olduğu bir yılı daha geride bıraktı. Dolar çok konuşuldu ama 115 bin 840 puanla tüm zamanların en yüksek seviyesine çıkan Borsa Endeksi, yıllık yüzde47.6 ile 2017'nin getiri şampiyonu oldu. Eurodaki yükseliş yüzde 22.1 olurken dolardaki artış yüzde 7.5'te kaldı. Cumhuriyet altınındaki prim yüzde 20.9 olarak gerçekleşti. Yılın başında 1.000 TL'si olan için mevduatın getirisi 106 TL, tahvilin getirisi ise 123 TL olarak hesaplandı. Küresel ekonominin

Üretken rakip ağlar ile metin üretimi yapmak için birçok farklı metot kullanılabilir. Bu yöntemlerin uygulanan problem ve istenilen sonuca göre başarı oranları değişmektedir. Bu çalışmada sınıflandırma başarısını arttırmak ve test etmek için aşağıdaki metotlar kullanılmıştır.

Tensorflow kurulan yapay sinir ağının anlaşılırlığını arttırarak tur sonrasında sonuçları incelemeye yardımcı olan basit bir matematik kütüphanesidir (URL-1, 2020). Bu çalışmada üretken rakip ağların kurulmasında Keras API kitaplıkları kullanılmıştır. Ayrıca büyük veri setlerinin eğitimi uzun süre almaktadır. Keras GPU destekli çalışmasından dolayı eğitim aşamasını kısaltmaktadır.

Uzun kısa süreli bellek (UKSB) özel bir TSA türüdür. Uzun kısa süreli bellek bilgileri uzun süre hatırlamak için kullanılır. UKSB hücreleri, lojik kapılar ile neyin saklanacağına neyin unutulacağına karar verir. Ağ şekli TSA biçimindedir. Her adımda bir cümledeki sonraki kelimeyi tahmin ederek eğitilirler. Yapısından dolayı zaman serili verilerde sınıflandırma ve tahmin konularına uygundur (URL-2, 2020).

N-gramlar unigram, bigram, trigram ve ngram olarak n adet kelime ya da harfin oluşturduğu sıralı dizilerdir. N-gramlar ile tekrar oranı bulunabilir. Çalışma mantığı istatistik ve olasılığa dayanmaktadır. Kendisinden önceki n-1 kelimeye bakarak Markov zincirlerinden yararlanır ve ardından gelecek kelimeyi belirlemeye çalışır (URL-3, 2020). Doğal dil işleme çalışmalarında n-gram modelleri yaygın olarak kullanılmaktadır. Bu çalışmada bigram ve trigram olarak ngram yöntemi analiz ve test aşamalarında kullanılmıştır.

Terim frekansı veri seti içerisinde bulunan kelimelerin sıklık oranlarını incelemek için kullanılır (URL-4, 2020). Ters belge frekansı yönteminde ise bu sıklık oranlarına göre bağlaçlar tespit edilip çıkartılarak daha başarılı eğitim amaçlanmıştır. Bu sayede en olumlu kelimeler ve en olumsuz kelimeler gibi bir sıralama yapılabilir.

Destek vektör makinesi (DVM) sınıflar arasına bir karar vektörü oluşturur. Destek vektör makinelerinde çekirdek fonksiyonları sayesinde birçok duruma uygun karar çizgileri çizilebilir. Burada veri seti olumlu ve olumsuz olmak üzere iki kategoriden oluşmaktadır. Olumlu ve olumsuz gruplar DVM ile bir optimal karar çizgisi ile ayrılabilir (URL-5, 2020).

Belirli bir sınıfın ya da durumun olasılığını modellemek için lojistik regresyon kullanılabilir (URL-6, 2020). Lojistik regresyonda her ögeye 0 ile 1 arasında toplamı bir olacak şekilde bir olasılık atanır ve bu ağırlığa göre verinin sınıflandırması yapılır.

Zemberek bir doğal dil işleme aracıdır. Türkçe üzerine özelleşmiştir ve metinlerin normalizasyon aşamasında kullanılır (URL-7, 2020).

Bu çalışmada kurulan üretken rakip ağ Python üzerinde hazırlanmıştır. İlk olarak veri seti hazırlık işlemleri yapılmıştır. Veri çiftleri ve bilgisi eksik hücreler var mı diye incelenmiştir. Haber verileri çekilirken hücrelere istenmeyen ögeler eklenip eklenmediğine bakılmıştır.

Öncelikle veri seti Sklearn kütüphanesi yardımı ile eğitim ve test veri seti olarak iki gruba ayrılmıştır. Test için % 30'luk kısım eğitim için % 70'lik kısım kullanılmıştır. Sınıflandırma sonuçlarının herhangi bir işlem yapılmadan önceki karmaşıklık matrisi Tablo 2'deki gibi dağılmıştır. Burada ana köşegende doğru sınıflandırılmış sonuçlar bulunmaktadır.

Tablo 2. Karışıklık matrisi dağılımı.

N = 918	Tahmin: 0	Tahmin: 1
Gerçek: 0	15	16
Gerçek: 1	4	883

Görüldüğü gibi test veri seti 918 haber metninin olumlu haberlerin doğru sınıflandırma oranı % 99,54904 iken olumsuz haberlerin doğru sınıflandırma oranı % 48,3871'de kalmaktadır. Buradaki olumlu haberlerin doğru sınıflandırılmasının olumsuz haberlerin doğru sınıflandırılmasına göre başarı oranındaki büyük farkın sebebi veri setinin normal dağılım göstermemesidir. Olumlu haberlerin sayısının olumsuz haberlere göre çok yüksek olması sınıflandırmayı doğrudan etkilemektedir.

Üretken rakip ağ modelinde üretici kısmında UKSB kullanılarak 128 düğüm içeren 3 katman oluşturulmuştur. Ayırıcı bölümünde ise sınıflandırma için DVM ve TF-IDF'ten yararlanılmıştır.

Eğitimde turlar sonrası en başarılı ağırlıklar not edilip bu ağırlıklardan daha iyiye yönlendirmeye çalışılmıştır. Üreticinin ürettiği bu metinler arasından veri setinin başarısızlığının ana sebebi olan olumsuz haberlerin sayıca eksikliği giderilmesi sağlanmıştır. Yani üreticinin ürettiği metinler ayırıcıya yönlendirilip olumlu ya da olumsuz olduğu incelenmiştir. Olumsuz metinler ayrılarak bir kenarda saklanmıştır. Bu yeni üretilen olumsuz metinler Zemberek (URL-7, 2020)aracı ile normalizasyon çalışması geçirmiştir. Daha sonra bu metinler orijinal veri setine eklenerek eğitime devam edilmiştir.

3. Araştırma Sonuçları ve Tartışma

İlk olarak yeni üretilmiş 50 olumsuz haber orijinal veri setine eklenmiştir. Üretilen bütün bu olumsuz haberler normalizasyon işlemi gördükten sonra herhangi bir kullanıcı düzenlenmesi olmadan veri setine eklenmiştir. Eklenmesi ile toplam haber sayısı 3108'e ulaşmıştır. Karar destek makinesi çekirdek fonksiyonu olarak 'linear' seçilip C katsayısı bir olarak tutulmuştur. Bunun üzerine alınan sınıflandırma başarı sonuçları not edilerek oluşturulan Tablo 3'te verilmiştir. Bu sınıflandırma işleminde cümlelerin içerdiği bütün kelimeler değerlendirmeye alınmış olup, bunların tamamı göz önüne alınarak cümlelerin sınıflandırma işlemi tamamlanmıştır.

Tablo 3. Elli olumsuz haberin eklenmesi sonucu karışıklık matrisi.

N = 933	Tahmin: 0	Tahmin: 1
Gerçek: 0	26	19
Gerçek: 1	3	885
Olumsuz Tahmin Yüzde	57,77778	
Olumlu Tahmin Yüzde	99,66216	

Yukarıdan da görüldüğü gibi 50 adet olumsuz haberin eklenmesi olumsuz tahmin başarı yüzdesinde yaklaşık % 9'luk bir artışa sebep olmuştur. Olumlu haber tahmin başarı yüzdesi ise % 0,11312'lik düşüş göstermiştir.

Yine kullanılan metot ile eğitime devam edilip yeni üretilen olumsuz haberler orijinal veri setine eklenmeye devam edilmiştir. Her yapılan eklemeden sonra alınan sonuçlar kaydedilmiştir. Toplamda 250 yeni olumsuz haberin eklenmesiyle oluşan karışıklık matrisi Tablo 4'teki gibidir.

Tablo 4. İki yüz elli olumsuz haberin eklenmesi sonucu karışıklık matrisi.

N = 996	Tahmin: 0	Tahmin: 1
Gerçek: 0	104	22
Gerçek: 1	1	869
Olumsuz Tahmin Yüzde	82,53968	
Olumlu Tahmin Yüzde	99,88506	

Eğitime devam edilip toplamda 250 yeni olumsuz haberin eklenmesinden sonra olumsuz haber tahmininde başarı oranı %82 seviyelerine ulaşmıştır. Olumlu haberlerde ise başarılı tahmin oranları devam etmektedir.

Eğitime devam edilip toplamda sırasıyla 750 ve 1750 yeni olumsuz haberin eklenmesiyle oluşan karışıklık matrisleri Tablo 5 ve Tablo 6'da verilmiştir.

Tablo 5. Yedi yüz elli olumsuz haberin eklenmesi sonucu karışıklık matrisi.

N = 1158	**Tahmin: 0**	**Tahmin: 1**
Gerçek: 0	260	33
Gerçek: 1	1	864
Olumsuz Tahmin Yüzde	88,7372	
Olumlu Tahmin Yüzde	99,88439	

Tablo 6. Bin yedi yüz elli olumsuz haberin eklenmesi sonucu karışıklık matrisi.

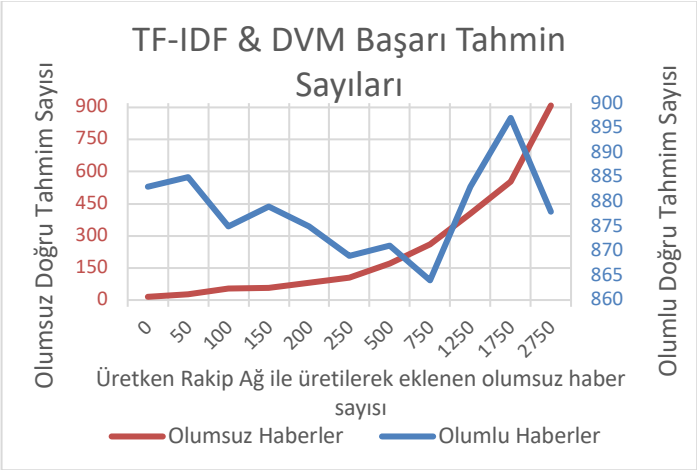
N = 1491	**Tahmin: 0**	**Tahmin: 1**
Gerçek: 0	553	40
Gerçek: 1	1	897
Olumsuz Tahmin Yüzde	93,25464	
Olumlu Tahmin Yüzde	99,88864	

Toplamda 2750 yeni olumsuz verinin üretilip eklenmesi sonucunda orijinal veri setindeki sınıf dengesizliği giderilmesi sağlanmıştır. Bu durumda oluşan yeni veri seti yaklaşık 6100 haberden oluşmaktadır. Eğitim sonucu alınan sonuçlarla oluşturulan karışıklık matrisi Tablo 7'de verilmiştir.

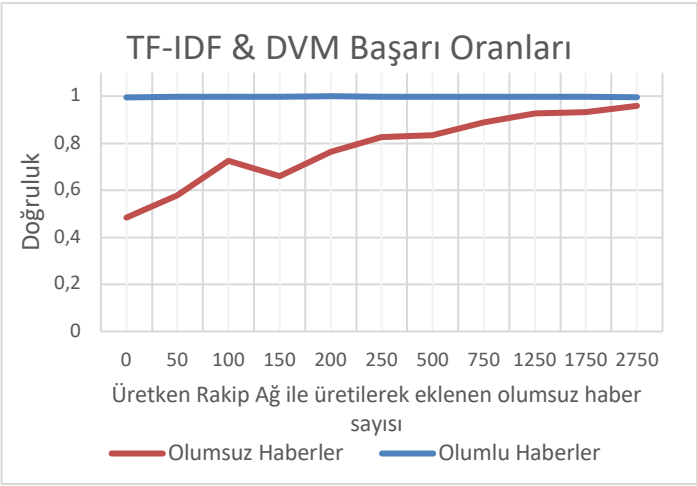
Tablo 7. İki bin yedi yüz elli olumsuz haberin eklenmesi sonucu karışıklık matrisi.

N = 1830	**Tahmin: 0**	**Tahmin: 1**
Gerçek: 0	909	39
Gerçek: 1	4	878
Olumsuz Tahmin Yüzde	95,88608	
Olumlu Tahmin Yüzde	99,54649	

Verilen çizelgelerden görüldüğü gibi veri setindeki eksik sınıfa ait metinlerin üretken rakip ağlar yöntemiyle üretilip sayıca sınıf eşitliği durumunun sağlanmasından sonra başarı oranı % 95,88608'e ulaştığı gözlenmiştir. Eğitimin her adımında alınan sonuçlar kaydedilmiş olup, başarılı tahmin sayıları ve başarılı tahmin oranları Şekil 2 ve Şekil 3'de verilmiştir.



Şekil 2. TF-IDF & DVM metodu ile eklenen olumsuz haber sayısına göre doğru sınıflandırılan haberlerin karşılaştırılması.



Şekil 3. TF-IDF & DVM metodu ile eklenen olumsuz haber sayısına göre doğru sınıflandırma oranları.

Görüldüğü gibi eğitim sonrasında olumsuz haberlerin doğru tahmin etme oranı yaklaşık % 47 artar iken olumlu haberlerin doğru tahmin etme oranında gözle görülür bir azalış olmamıştır. Üretken rakip ağların metin üretimindeki başarısı sonuçta olumlu bir etkiye sahip olmuştur.

Aşağıda üretken rakip ağ ile üretilen metinlerden birkaç örnek verilmiştir:

"buna hakkınız da yok yaşlı adam parayı cumhuriyet altında bir kapat kardı konusu bir yapılan bu karşılığı konusu bir yönetim ve istanbul bankasından alacakları ile"

"reddeddi istanbul 4 asliye hukuk mahkemesinde bulunan şubesi ve iş bankası da krediyi yakın izlemeye aldı garanti bankasının geri alındı"

"konu devlet meselesi değil memleket merkezi ve markaralara bir gidin başkanı belirtileri dolan yer yaptığını bir türk heyeti alacakları ile"

"büyük darbeyi borsa vurdu rekor üstüne de kuruluyor bankalar bankası bir yıldan uzun yüzde 1,5 artış ile"

Olumlu ve olumsuz sınıfa ait cümleler incelenmiş olup TF-IDF metodu ile sıklık oranları belirlenip en olumlu ve en olumsuz kelimeler sıralanabilir. Bu sıralamada veri setinin hazırlandığı 2017 sonbaharındaki haberlerin belirleyici etkisi olmuştur. Örnek olarak en olumlu beş kelime ve en olumsuz beş kelime aşağıda verilmiştir.

En olumlu beş kelime;

"finansal, türkiye, dijital, yeni, en"

En olumsuz beş kelime;

"telekom, karar, yakın, müdürü, bankası"

4. Sonuç

Türkçe doğal dil işleme çalışmalarındaki problemlerden birisi de etiketli veri kümesi bulma zorluğudur. Bunun yanı sıra bulunan veri kümelerinde sınıflarının eşit dağılımlı olmaması da olağandır. Bu çalışmada üretken rakip ağlar ile Türkçe metin üretimi süreci yapılmıştır. Üretken rakip ağlar ile Türkçe dilinde başarılı şekilde Türkçe metinler üretilebileceği görülmüştür. Burada üretilen metinler normal dağılım göstermeyen bir veri seti üzerinde uygulanarak sınıflandırma başarısını arttırılmıştır.

Hiçbir ekleme yapılmadan ki durumdan ve üretken rakip ağlar ile eksik sınıfa ait verilerin üretilip eklenmesiyle dengelenme durumunda sınıflandırma başarısı yaklaşık % 47 oranında arttırılmıştır. Bu da üretilen metinlerin başarılı olduğunu göstermektedir. Üretken rakip ağlar ile üretilen metinlerin çıktılarının bir normalizasyon işlemi görmesi sonrasında gerçek konuşma diline yakın anlamlı cümleler oluşabildiği görülmüştür.

Kaynakça

Michie D., Spiegelhalter D. J. & Taylor C. C. (1994). Machine learning. *Neural and Statistical Classification*, (13.1994), 1-298.

Ayon D. (2016). Machine learning algorithms: a review. *International Journal of Computer Science and Information Technologies,* (7.3), 1174-1179.

Xiaojin Z. (2005). Semi-Supervised Learning Literature Survey. *CS Technical Reports University of Wisconsin-Madison*.

Jun S. & Hideki I. (2008). Semi-Supervised Sequential Labeling and Segmentation using Giga-word Scale Unlabeled Data. *Proceedings of ACL-08 HLT*, 665–673.

Goodfellow I. J., Pouget-Abadie J., Mirza M., Xu B., Warde-Farley D., Ozair S., Courville A. & Bengio Y. (2014). Generative Adversarial Nets. *Advances in Neural Information Processing System (NIPS)*, 2672-2680.

Zhang H., Xu T., Li H., Zhang S., Wang X., Huang X. & Metexas D. N. (2019). StackGAN++: Realistic Image Synthesis with Stacked Generative Adversarial Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* (41.8), 1947-1962.

Santhanam S. (2018). Context Based Text-Generation Using Lstm Networks. *Conference: Artificial Intelligence International Conference A2IC*.

Wang W. Y., Singh S. & Li J. (2019). Deep Adversarial Learning for NLP. *Proceedings of NAACL HLT*, 1-5.

Che T., Li Y., Zhang R., Hjelm R., Li W., Song Y. & Bengio Y. (2017). Maximum-Likelihood Augmented Discrete Generative Adversarial Networks.

Guo J., Lu S., Cai H., Zhang W., Yu Y. & Wang J. (2017). Long Text Generation via Adversarial Training with Leaked Information. *Association for the Advancement of Artificial Intelligence*.

Yu L., Zhang W., Wang J. & Yu Y. (2017). Long Text Generation via Adversarial Training with Leaked Information. *Association for the Advancement of Artificial Intelligence*.

Lin K., Li D., He X., Zhang Z. & Sun M. (2017). Adversarial Ranking for Language Generation. *Advances in Neural Information Processing System (NIPS)*.

Fedus W., Goodfellow I. & Dai A. (2018). Maskgan: Better Text Generation Via Filling In The _____. *International Conference on Learning Representations (ICLR)*.

Cao Y., Zhou Z., Zhang W. & Yu Y. (2017). Unsupervised Diverse Colorization via Generative Adversarial Networks.

Kusner M. & Hernandez-Lobato J. (2016). GANS for Sequences of Discrete Elements with the Gumbel-softmax Distribution.

Şahin T., Demir Ö. & Yıldız K. (2019). Doğal Dil İşleme Uygulamaları İçin Türkçe Veri Seti Oluşturulması. *International Periodical of Recent Technologies in Applied Engineering*, (1.2), 51-57.

Bilgin M. (2019). Türkçe Metinlerin Sınıflandırma Başarısını Artırmak İçin Yeni Bir Yöntem Önerisi. *Uludağ University Journal of The Faculty of Engineering*, (24.1), 125-136.

Antreas A., Amos S. & Harrison E. (2018). Data Augmentation Generative Adversarial Networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).*

Georgios D. & Fernando B. (2018). Effective data generation for imbalanced learning using conditional generative adversarial networks. *Expert Systems with Applications*, 464-471.

Maskin E. (1999). Nash Equilibrium and Welfare Optimality. *The Review of Economic Studies*, (66.1), 23-38.

Gharesifard B. & Cortes J. (2013). Distributed convergence to Nash equilibria in two-network zero-sum games. *Automatica*, (49.6), 1683-1692.

URL-1, (2020). Neden TensorFlow. Erişim Tarihi: 15.12.2020. Erişim Adresi: https://www.tensorflow.org/about .

URL-2, (2020). Long short-term memory. Erişim Tarihi: 15.12.2020. Erişim Adresi: https://en.wikipedia.org/wiki/Long_short-term_memory .

URL-3, (2020). N-gram. Erişim Tarihi: 15.12.2020. Erişim Adresi: https://en.wikipedia.org/wiki/N-gram .

URL-4, (2020). TF-IDF/Term Frequency Technique. Erişim Tarihi: 15.12.2020. Erişim Adresi: https://medium.com/analytics-vidhya/tf-idf-term-frequency-technique-easiest-explanation-for-text-classification-in-nlp-with-code-8ca3912e58c3 .

URL-5, (2020). Support vector machine. Erişim Tarihi: 15.12.2020. Erişim Adresi: https://en.wikipedia.org/wiki/Support_vector_machine .

URL-6, (2020). Logistic regression. Erişim Tarihi: 15.12.2020. Erişim Adresi: https://en.wikipedia.org/wiki/Logistic_regression .

URL-7, (2020). Zemberek-NLP. Erişim Tarihi: 15.12.2020. Erişim Adresi: https://github.com/ahmetaa/zemberek-nlp .

Not

Bu çalışmada 2020 yılında Barış Gücük tarafından sunulan "Üretken Rakip Ağlar ile Türkçe Metin Üretimi" isimli tezden üretilmiştir.