Türkçe Haber Metinlerinin Konvolüsyonel Sinir Ağları ve Word2Vec Kullanılarak Sınıflandırılması

Araştırma Makalesi/Research Article

Çiğdem İnan ACI*, Adem ÇIRAK

Bilgisayar Mühendisliği Bölümü, Mersin Üniversitesi, Mersin, Türkiye
caci@mersin.edu.tr, 1702010171013@mersin.edu.tr
(Geliş/Received:06.09.2018; Kabul/Accepted:18.06.2019)
DOI: 10.17671/gazibtd.457917

Özet— Bu çalışmada, Konvolüsyonel Sinir Ağları (KSA) ve Word2Vec metodu kullanılarak Turkish Text Classification 3600 (TTC-3600) veri kümesi üzerinde metin sınıflandırma çalışması yapılmış ve aynı veri kümesi kullanılarak yapılan önceki çalışma ile kıyaslanmıştır. Çalışmada TTC-3600'ün ham ve Zemberek yazılımıyla gövdelenmiş halleri üzerinde iki farklı KSA eğitilmiş ve test edilmiştir. KSA ve Word2Vec metodu, klasik istatistiksel ve makine öğrenmesine dayalı sınıflandırma algoritmalarından daha iyi bir performans (%93,3 doğruluk) göstermiştir. Türkçe doğal dil işleme çalışmalarının azlığı ve bu alandaki özellik çıkarma yöntemlerinin limitli olması sebebiyle, kelimelerin semantik değerlerinin önceden eğitilmiş Word2Vec ağı ile sınıflandırmaya katılabilmesi KSA modellerinin doğruluk değerlerini arttırmıştır.

Anahtar Kelimeler— Türkçe metin sınıflandırma, konvolüsyonel sinir ağları, derin öğrenme, word2vec

Turkish News Articles Categorization Using Convolutional Neural Networks and Word2Vec

Abstract— In this study, a text classification study on the Turkish Text Classification 3600 (TTC-3600) dataset was conducted using Convolutional Neural Networks (CNN) and Word2Vec method and compared with the previous study using the same dataset. In the study, two different CNN s were trained and tested on the TTC-3600 raw and stuck with Zemberek software. CNN and Word2Vec method showed better performance (93.3% accuracy) than classical statistical and machine learning based classification algorithms. Due to the limited number of natural language processing operations in Turkish and the limited feature extraction methods in this area, the accuracy of the CNN models has increased by allowing the semantic values of the words to be included in the classification with the pre-trained Word2Vec network.

Keywords— Turkish text categorization, convolutional neural networks, deep learning, word2vec.

1. GİRİŞ (INTRODUCTION)

İnternet, kullanılmaya başlandığı 1981 yılından itibaren hızla büyümeye devam etmektedir [1]. İnternet kullanımındaki bu artış, çok miktarda verinin üretilmesine yol açmaktadır. International Data Corporation'ın raporuna [2] göre, yapılandırılmamış veriler, 2020 yılı itibariyle 40 Zettabayttan fazla ve yapısal verilere göre 50 kat daha geniş olacaktır.

Yapısal olmayan veriler arttıkça, bu verilerin elle kategorize edilmesi neredeyse imkansız hale gelmektedir. Doğal Dil İşleme [3], Bilgi Getirimi [4] ve Makine Öğrenmesi [5] gibi araştırma alanları, büyük verinin otomatik olarak kategorize edilmesine izin verir. Metin tipindeki veriyi analiz etme yaklaşımlarından biri olan sınıflandırma, nesnelerin veri analiz teknikleri kullanılarak önceden tanımlanmış sınıflara ayrılması işlemidir. Metin sınıflandırma konusunda literatürde birçok çalışma mevcut olmakla birlikte yapılan çalışmaların geneli İngilizce

metinler üzerinde yoğunlaşmıştır [6–15]. Türkçe metinlerin sınıflandırılmasında çok sınırlı veri kümesi, araç ve çalışma kaynağı olsa da, yapılan çalışmalar son yıllarda artmaktadır: Türkçe dokümanların sınıflandırılmasıyla ilgili yapılan ilk çalışmalardan biri Çatal ve arkadaşları [16] tarafından, n-gram'lar kullanılarak geliştirilen NECL adlı sistemdir. Amasyalı, Diri ve Türkoğlu [17] , yazarı bilinmeyen bir dokümanın önceden belirlenmiş ve yazarlık özellikleri çıkarılmış 18 farklı yazardan hangisine ait olduğunu Naive Bayes (NB), Destek Vektör Makinesi (DVM), C4.5 ve Rastgele Orman (RO) algoritmalarını kullanarak tespit etmişlerdir. Doğan [18] ise yaptığı çalışmada, dokümanın yazarını tespit etmenin yanı sıra dokümanın türünü ve doküman yazarının cinsiyetini belirlemiştir. Bunun için Türk dilinin 2, 3 ve 4'lü gramları çıkarılarak farklı boyutlarda özellik vektörleri oluşturulmuştur. Daha sonra bu özellik vektörlerinin boyutları korelasyon tabanlı özellik seçiciler kullanılarak azaltılmış ve farklı boyutlarda özellik vektörleri elde edilmiştir. N-gram modeline dayalı bu özellik vektörleri, NB, DVM, RO ve K-En Yakın Komşuluk (KYK) yöntemleri yardımıyla sınıflanmıştır. Biricik ve Diri[19] , 28567 adet Türkçe web sayfasından elde edilen 13 farklı kategorideki dokümanın özelliklerini, terimlere ağırlık vererek ve olasılık dağılımlarını göz önüne alarak çıkarmışlardır. Önerilen bu özellik çıkarma yaklaşımını beş değişik sınıflayıcı (NB, C4.5, Radyal Tabanlı Yapay Sinir Ağı (RYSA), KYK ve RO) ile test etmişlerdir. Önerilen yaklaşım tüm sınıflayıcıların performansını artırmıştır. Aşlıyan ve Günel [20], En Yakın Komşu (EYK) ve KYK metotlarının performanslarını kıyaslamak için beş kategorideki Türkçe dokümanlardan oluşan iki farklı derlem oluşturmuş ve en yüksek doğru sınıflandırma oranını EYK metodu ile %88,4 olarak belirlemiştir. Konuşma dili kullanılarak elde edilen başka bir Türkçe veri kümesi ise NB, DVM ve KYK yöntemleri kullanılarak demografik özelliklerine göre sınıflandırılmıştır [21]. Amasyalı vd. [22], çeşitli türlerdeki 6 adet Türkçe veri kümesi üzerinde 17 adet metin temsil yönteminin etkisini karşılaştırmış ve en başarılı metin temsil yönteminin harf n-gramları olduğu sonucuna ulaşmışlardır. Her bir veri kümesinde ise en az %75 sınıflama doğruluğu elde etmişlerdir. Levent ve Diri [23], Türkçe metinlerin yazarlarını tanıma problemini Yapay Sinir Ağları (YSA) ile çözmüşler ve yazar tanıma için kullanılan önceki algoritmalara yakın sonuçlar elde etmişlerdir. Hüsem [24], DMOZ (Directory Mozilla - The Open Directory Project) web dizininden elde edilen 22 bin kayıtlık Türkçe web sitelerini NB ve DVM yöntemleriyle sınıflandırmıştır. NB ve DVM algoritmalarına n-gram kelime vektörü seçimi ve bilgi kazanım oranı yaklaşımları uygulanarak performansları karşılaştırılmıştır.

Türkçe metin sınıflama çalışmalarında haber veri kümeleri sıklıkla kullanılmaktadır [25]. Haberlerin konularına göre kategorize edilmesi, anlık olarak üretilen haber metinlerini yönetmek adına elzem bir uygulama olmuştur. Türkçe haberleri sınıflayan önceki çalışmalar şöyle özetlenebilir: Amasyalı ve Yıldırım [26] gazetelerin web sayfalarından toplanan beş kategorideki haber metinlerini sınıflamak için NB, Vektör Nicemleme ve Çok Katmanlı Sınıflayıcı kullanıp ve %76 oranında başarı elde etmişlerdir. Amasyalı ve Beken [27],Türkçe haber metinlerini sınıflamak için "iki kelimenin birbirlerine anlamsal benzerliğinin, kelimelerin birlikte geçtiği doküman sayısıyla doğru orantılı" olduğunu öne süren bir hipotezden faydalanmış ve geleneksel metotlara göre daha başarılı sonuçlar etmişlerdir. Tüfekci, Uzun ve Sevinç [28], haber portallarının ekonomi, sağlık, magazin, spor ve siyaset kategorilerinden seçilmiş iki farklı veri kümesini NB, DVM, C4.5 ve RO algoritmaları ile sınıflandırmışlardır. Sınıflama için kullanılan özellik vektörü, kelime frekansına göre ağırlıklandırılmış ve özellik olarak kelime gövdeleri seçilmiştir. Bu seçim sırasında farklı uzunlukta ve farklı tipte gövdelerin seçilmesinin sınıflandırmaya olan etkileri incelenmiş ve özellik olarak isim tipinde uzun gövdeli kelimeler seçildiğinde başarının daha yüksek seviyede olduğu görülmüştür. Haber metinlerini sınıflandıran bir başka çalışmada ise DVM algoritmasında kullanılan çekirdek fonksiyonu, "yüksek seviyeli semantik çekirdek" adı verilen yeni bir çekirdek fonksiyonu ile değiştirilmiş ve sınıflama performansının arttığı görülmüştür [29]. Negatif-olmayan Matris Faktörizasyonu yöntemi kullanarak boyut azaltma yaklaşımının Türkçe haber metinlerini sınıflandırmaya yönelik etkisi ise Güran vd. [30] yaptığı çalışmada irdelenmiştir. Çalışmanın sonuçlarına göre, boyut azaltma işlemi K-Ortalama Algoritması'nın sınıflama doğruluğunu arttırmamakla birlikte sınıflama hızını arttırmıştır. Başkaya ve Aydın [31], farklı haber sitesi ve gazetelerden alınan 4 kategoride ve her bir kategoriye ait 20 haber metni bulunan bir veri kümesinin boyutunu CfsSubset Algoritması ile indirgemiş ve ardından NB, DVM, J48 ve RO yöntemleriyle indirgenen veri kümesini sınıflamışlardır. Kaynar ve Aydın ise duygu analizi için öznitelik düşürme yöntemi olarak oto kodlayıcı ve derin öğrenme ağı kullanmışlar ve diğer yaygın öznitelik düşürme teknikleri ile karşılaştırmışlardır [32].

Bu çalışmanın motivasyonu, literatürde şimdiye kadar oldukça fazla çalışılmış yöntemlerin (NB, DVM, KYK, RO vb.) dışında nispeten yeni bir yöntem olan Derin Öğrenme ve Word2Vec ön işleme metodu kullanarak Türkçe haber metinlerini sınıflandırmaktır. Bu bağlamda, Turkish Text Classification 3600 (TTC-3600) [33] olarak da bilinen haber veri kümesi, Konvolüsyonel Sinir Ağları (Convolutional Neural Networks, CNN) (KSA) [34] ile sınıflandırılmıştır. Sınıflandırma işlemi için, iki adet Derin Öğrenme tabanlı model tasarlanmış ve TTC-3600 veri kümesi bu modellerde sınıflanmadan önce ön işlemeye tabii tutulmuştur. Ön işlemenin ardından, sınıflama doğruluğunu arttırmak için bir kelime vektör temsil yöntemi olan Word2Vec [35, 36] kullanılarak TTC-3600 veri kümesi zenginleştirilmiştir. Geliştirilen her iki modelde de literatürdeki çalışmalara kıyasla daha iyi doğruluk oranlarına ulaşılmıştır.

Makalenin geri kalanı şu şekilde organize edilmiştir: Bölüm 2, kullanılan veri kümesi ve yöntemler hakkında bilgi vermekte olup, yapılan sınıflama işleminin ayrıntıları ve elde edilen sonuçlar ise Bölüm 3'te verilmiştir. Bölüm 4'te elde edilen sonuçlar önceki çalışmalarla karşılaştırılmış ve makale sonuçlandırılmıştır.

2. MATERYAL VE METOT (MATERIAL AND METHOD)

2.1. TTC-3600 Veri Kümesi (TTC-3600 Dataset)

Türkçe haberleri sınıflama çalışmalarında yaygın olarak kullanılması amacıyla hazırlanan TTC-3600 veri kümesi, Kılınç vd. [33] tarafından 2017 yılında derlenmiştir. Türkçe haber veri kümeleri arasında son yıllarda yayımlanmış, kullanımı kolay ve iyi belgelenmiş bir veri kümesi olan TTC-3600, erişime açıktır [37]. Veri kümesi, ekonomi, kültür-sanat, sağlık, politika, spor ve teknoloji olmak üzere 6 kategoriden toplam 600 haber / metin içeren 3600 belgeden oluşmaktadır. Haber metinleri, Mayıs ve Temmuz 2015 tarihleri arasında Zengin Site Özeti (Rich Site Summary, RSS) aracılığıyla ilgili haber portallarından toplanmıştır [33].

2.2. Ön İşleme (Pre-Processing)

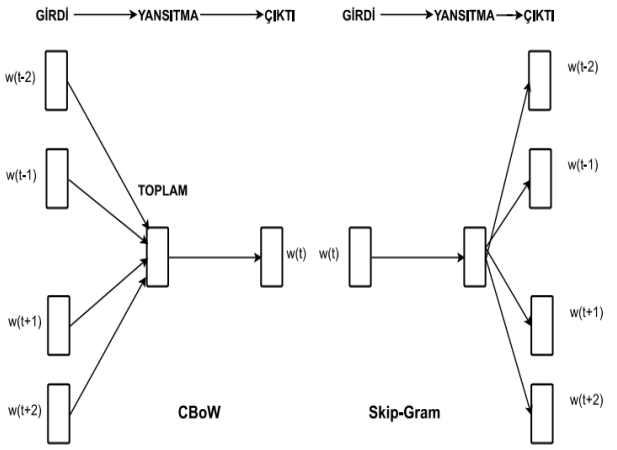
TTC-3600 veri kümesinin ön işlemeden önceki halinde, RSS beslemesinden alınan XML formatındaki haber metinleri bulunmaktadır. Sınıflama çalışmaları için elde edilen XML verilerinin <title> ve <description> başlıkları altında bulunan bilgiler dikkate alınmıştır. Bununla birlikte, bu başlıklar altında bulunan Java Script kodları, diğer HTML başlıkları (<img> , <a> , <p > , < strong > vs.), operatörler, noktalama işaretleri, yazdırılamayan karakterler, reklamlar gibi gereksiz veriler de temizlenmiştir. Gövdeleme (stemming) işlemi için boşluk karakterleri ve Zemberek [38] yazılımı ayrı ayrı kullanarak metin içerisinde bir gövdeleme işlemi yapılmıştır. Boşluk karakterleri ile gövdelenen kelimeler ham halleri ile (metinde geçtiği şekilde) kullanılmıştır. Zemberek, açık kaynak kodlu [39] Türkçe doğal dil işleme kütüphanesidir. Bu kütüphane kullanılarak kelimelerin kökleri bulunmuştur. Örneğin bir haber metninde geçen "Cumhurbaşkanı Tayyip Erdoğan'ın, koalisyon zorunluluğunun çıkmasının ardından gelenek ve teamüllere uyacağı öğrenildi." cümlesinin Zemberek ile gövdelenmiş hali "cumhurbaşkan, Tayyip, Erdoğan, koalisyon, zor, çıkmak, art, gelenek, teamül, uymak, öğrenmek" olmuştur.

2.3 Word2Vec (Word2Vec)

Kelimeler arasında anlamsal veya kavramsal açıdan ilişki kurmak için kelimelerin ne sıklıkla bir arada geçtiği verisini içeren kelime temsillerinin öğrenimi, metinlerin sınıflandırma başarısını etkilemektedir [40]. Word2Vec, Mikolov vd. [41] tarafından önerilen YSA tabanlı bir kelime vektörü temsil yöntemidir. Bu yöntem sayesinde kelimeler vektörlere dönüştürülerek aralarındaki uzaklıklar hesaplanıp kelimeler arasında analoji kurulabilmektedir [42]. Ayrıca, kelime vektörleşme işlemi sonrası elde edilen kelime vektörleri, metin içinde geçen cümleleri, vektörel değerler olarak ifade edebilmeyi sağlar [43].

Kelime koordinatlarının bulunmasında continuous bag of words (CBoW) ve skip-gram (SG) olmak üzere iki farklı öğrenme mimarisi kullanılmaktadır. CBoW mimarisinde bir kelimenin belli bir pencere boyutu içindeki komşu kelimelerine (sağındaki ve solundaki kelimeler) bakılmakta ve ilgili kelime komşu kelimelerden tahmin edilmeye çalışılmaktadır. SG mimarisinde ise tam tersi şekilde hedef kelimeye bakılarak komşu kelimeler tahmin edilmektedir [44]. Genelde SG mimarisi derlem frekansı az olan kelimeler için daha iyi kelime vektörleri üretirken, CBoW daha büyük çaptaki derlemlerde daha iyi sonuçlar üretmektedir (Şekil 1).

Word2Vec modelinin eğitimi sırasında cümle içinde yer alan belirli bir kelime girdi, cümle içerisinde o kelimeye belirli bir yakınlıkta olan kelimeler çıktı olarak seçilir. Özelleştirilmiş ve tek bir gizli katmanı bulunan YSA modeli kullanılarak bu kelime ikilileri eğitimli öğrenmeye tabi tutulur. Çok sayıda cümle ve çok sayıda kelime ikilileri kullanılarak eğitilen YSA modelinin gizli katman ağırlık değerlerine ulaşmak Word2Vec eğitim işleminin ana hedefidir. Eğitim işlemi tamamlandığında hangi kelimelerin birlikte kullanıldığı ya da hangi kelimelerin anlamsal olarak birbirine yakın olduğu istatistiksel olarak hesaplanabilmektedir. Bu sayede kelimelerin uzayda dağılımı otomatik olarak düzenlenebilmektedir. Eğitim metotları, kullanılan aktivasyon fonksiyonları ve parametreler Word2Vec modelinin kullanım amacına göre değişiklik göstermektedir [45].



Şekil 1. CBoW ve SG mimarileri (CBoW and SG architectures)

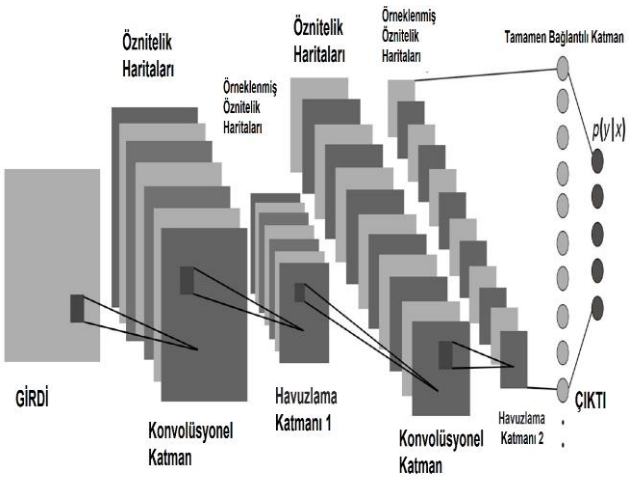
Word2Vec yöntemi son yıllarda metin sınıflaması konusunda oldukça popüler olmuştur. Birçok çalışmada [42–55] sınıflama yöntemleri uygulanmadan önce Word2Vec kullanılarak veri kümesi zenginleştirilmiş ve metin verileri vektörel hale getirilmiştir.

2.4 Derin Öğrenme ve Konvolüsyonel Sinir Ağları (Deep Learning and Convolutional Neural Networks)

Derin Öğrenme [56], son birkaç yılda popülerlik kazanmaya başlayan yeni bir makine öğrenmesi alanıdır. Derin Öğrenme, derin mimariye dayanan, gizli katmanların sayısı arttırılmış ve her katmanda probleme dair bir özniteliğin (feature) öğrenildiği YSA'lardan oluşan yöntemler bütünüdür. Bu mimaride, her katmanda öğrenilen öznitelik bir üst katmana bir girdi oluşturur. Böylelikle en alt katmandan en üst katmana doğru en basitten en karmaşık niteliğin öğrenildiği bir yapı kurulmuş

olur [57]. Derin öğrenmenin temel amacı giriş verisini çeşitli dönüşümlerle daha etkin bir öğrenme sağlayabilecek duruma dönüştürmek ve daha sonra öğrenme algoritmasını işletmektir [58].

Derin Öğrenme son yılların en önemli konuları arasındadır. Görüntü işleme [60–64], ses tanıma [55] ve doğal dil işleme[63] gibi alanlarda kendine yer bulmuştur. Derin öğrenmenin özelleşmiş bir mimarisi olan KSA'lar, özellikle görüntü işlemede oldukça başarılı olsa da son yıllarda metin sınıflandırma çalışmalarında da sıklıkla kullanılmaktadır [63–69]. Bir KSA, temel olarak konvolüsyonel katman, havuzlama katmanı ve tamamen bağlantılı katman olmak üzere üç parçada incelenebilir (Şekil 2). Konvolüsyonel katmanda giriş filtrelenir ve öznitelik haritaları elde edilir. Havuzlama katmanında öznitelik haritaları örneklenir ve ağın daha genel ve hızlı öğrenmesi sağlanır. Son olarak tamamen bağlantılı katmandaki her nöron bir önceki katmandan gelen tüm girişlere bağlı olarak çıkış üretir. Her katman bir önceki katmanın sonucuna göre öznitelik çıkarır ve tüm katmanları birleştirip eğiterek öznitelik hiyerarşisini öğrenebilir. Burada amaç, düşük seviye detaylarından başlayarak yüksek seviye detaylarına kadar etkili bir öğrenme gerçekleştirmektir.

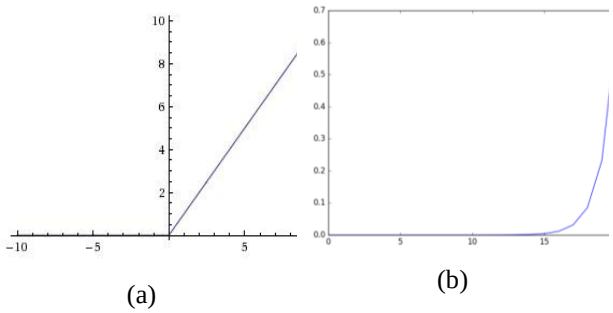


Şekil 2. Konvolüsyonel Sinir Ağı Katmanları (Convolutional Neural Network Layers)

3.TTC-3600 VERİ KÜMESİNİN SINIFLANDIRILMASI (CLASSIFICATION OF TTC-3600 DATASET)

TTC-3600 veri kümesini [33], sınıflandırma için izlenen adımlar Şekil 4'teki akış diyagramında gösterilmiştir. Akış diyagramı takip edildiğinde, ilk aşama veri kümesindeki geçersiz karakterlerin temizlenmesi işlemidir. Bu aşamadan sonra HTML etiketleri temizlenmiş ve gövdeleme aşamasına geçilmiştir. Gövdeleme aşamasında iki farklı veri kümesi oluşturulmuştur. Bunlardan ilkinde Zemberek [38] yazılımına göre, ikincisinde ise boşluk karakterlerine göre metin gövdelenmiştir. Boşluk karakterine göre gövdeleme işleminde kelimeler ham halleri ile değerlendirilmiştir. Gövdelemenin ardından metin içinde geçen edat, bağlaç ve zamir gibi çok sık kullanılan ve tek başına anlam ifade etmeyen kelimelerin metinden çıkarılması gerekir. Durak kelimeleri (stop words) de denen bu kelimeler metindeki gürültüyü azaltmak adına temizlenmiştir. Word2Vec kullanılarak kelime vektör temsili yapıldıktan sonra veri kümesi, test ve eğitim kümesi olarak rasgele ayrılmıştır. Her eğitim işlemi için %90 oranında eğitim, %10 oranında test verisi alınmıştır. Eğitim ve test işlemlerinde 10 katlı çapraz doğrulama kullanılmıştır.

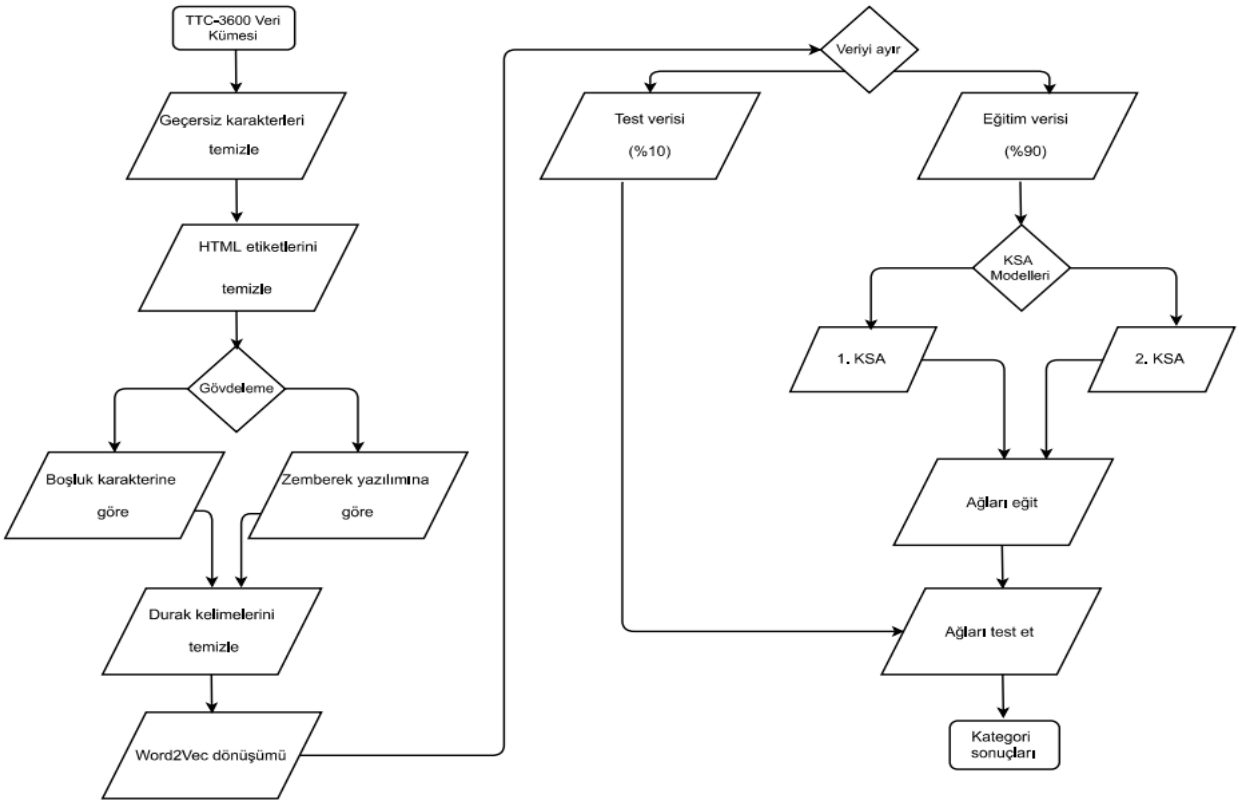
Veri kümesinin eğitiminde iki farklı derin öğrenme ağı kullanılmıştır: İlk model, 1-Boyutlu KSA kullanan 3 gizli katmanlı bir derin öğrenme ağıdır. Aktivasyon fonksiyonu olarak Doğrultulmuş Doğrusal Ünite (Rectified Linear Unit, ReLU) kullanılmıştır (Şekil 3.a). ReLU, YSA'daki düğümlerin başlangıç değerini pozitif olarak başlatmak için son yıllarda çok kullanılan bir aktivasyon yöntemidir. KSA'nın ağırlık katsayılarının güncellenmesi için gradyan tabanlı bir öğrenme algoritması olan Kök Ortalama Kare Yayılımı [70] (Root Mean Square Propagation-Rmsprop) kullanılmıştır. RMSprop, öğrenme oranını üstel olarak azalan gradyanların karesinin ortalamasına böler. Böylece, klasik gradyan öğrenmeye göre daha etkili bir öğrenme elde edilmiş olur. Bu ağda 1-Boyutlu konvolüsyonlarla zenginleştirilmiş veri kümesi üzerinde anlamlı kelime seçilimi yapılarak özellikler işlenmiştir. Çıkış katmanında ise Softmax aktivasyon fonksiyonu kullanılmıştır (Şekil 3.b).



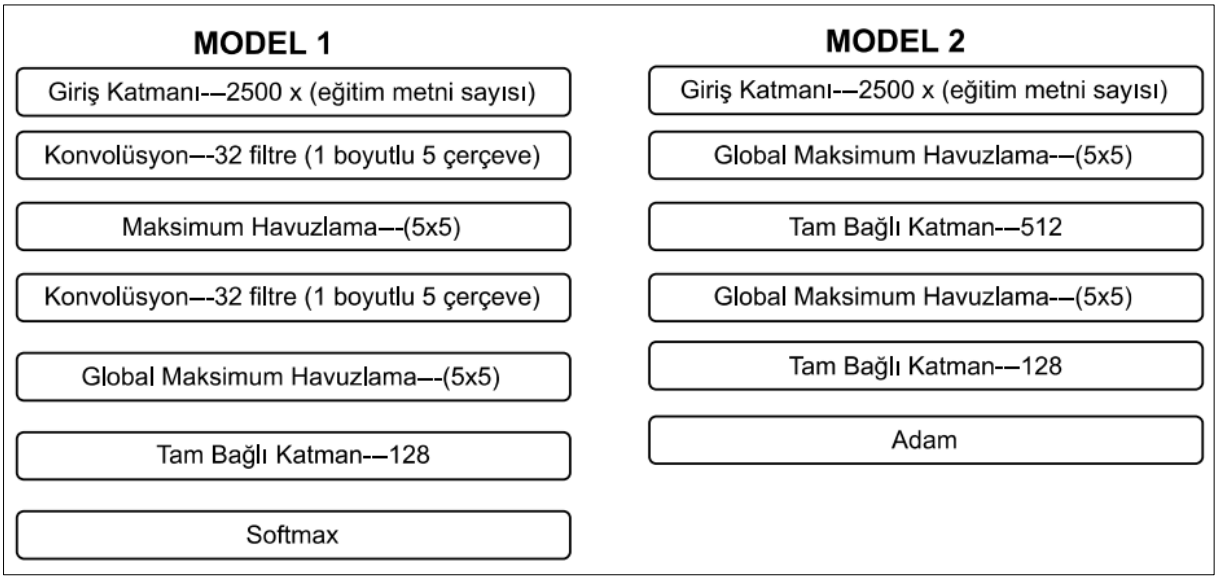
Şekil 3. Aktivasyon Fonksiyonları: (a) Doğrultulmuş Doğrusal Ünite (b) Softmax (Activation Functions: (a)Rectified Linear Unit (b) Softmax)

Eğitimde kullanılan ikinci KSA modelinde, global maksimum havuzlama (global maximum pooling) işlemi yapılmıştır. Öznitelik haritaları havuzlama işlemine tabi tutularak boyutları düşürülür ve böylece özniteliklerdeki çeşitlilik azaltır. Havuzlama işlemi genelde en büyüğü bulma (maksimum) operatörü ile yapılır. Seçilen ebatta havuzlama penceresinin içerisinde kalan özniteliklerin en büyük elemanı bulunur ve öznitelik haritalarının boyutu düşürülerek tekrar oluşturulur. Bu modelde gizli katmanlardaki aktivasyon için yine ReLU fonksiyonu, öğrenme algoritması olarak da Uyarlamalı Moment Tahmini [71] (Adaptive Moment Estimation-Adam) metodu kullanılmıştır. Adam algoritması, eğitim verilerine dayanan ağ ağırlıklarını güncellemek için klasik stokastik gradyan tabanlı yöntemler yerine kullanılabilecek bir optimizasyon algoritmasıdır. Bilinen iki gradyan tabanlı

algoritma (Uyarlamalı Gradyan Algoritması (AdaGrad) ve RMSprop) hibritlenerek Adam algoritması oluşturulmuştur. Adam algoritması, RMSProp'taki ortalama ilk moment değerini temel alan parametre öğrenme oranlarını uyarlamak yerine, aynı zamanda gradyanların ikinci momentlerinin ortalamasını da (merkezsiz varyans) kullanır. İyi sonuçlara hızlı ulaştığı için derin öğrenme alanında popüler bir algoritmadır. İkinci modelin çıkış katmanında ise Softmax aktivasyonu kullanılmıştır. Birinci ve ikinci modelin tasarımında kullanılan parametreler Şekil 5'te verilmiştir.



Şekil 4. TTC-3600 veri kümesinin sınıflandırılması (Classification of TTC-3600 dataset)



Şekil 5. Birinci ve İkinci KSA modellerinin parametreleri (Parameters of the first and second CNN models)

4. TEST SONUÇLARI VE TARTIŞMA (TEST RESULTS AND DISCUSSIONS)

Her iki KSA modeli de Python [72] dili kullanılarak, Tensorflow [73] arayüzünün Keras Kütüphanesi [74] yardımıyla kodlanmıştır. Tensorflow paralel ve dağınık olarak makine öğrenmesi alanında hesaplama arayüzü ve çerçevesi sağlarken, Keras bu katman üzerinde modellerin kolayca oluşmasını sağlamaktadır. Veri kümesinin orijinal hali ve Zemberek yazılımıyla gövdelenmiş hali ayrı ayrı test edilip Kılınç vd. [33] tarafından yapılan çalışma ile kıyaslanmıştır.

Sınıflandırma performansını ölçmek için çeşitli metotlar kullanılabilmekle birlikte bu çalışmada test sonuçları için karışıklık matrisi (confusion matrix) kullanılarak doğru sınıflandırılmış pozitif örnek (True Positive, TP), doğru sınıflandırılmış negatif örnek (True Negative, TN), yanlış sınıflandırılmış pozitif örnek (False Positive, FP) ve yanlış sınıflandırılmış negatif örnek (False Negative, FN) değerleri hesaplanmıştır. Sınıflandırmanın performansını ölçmek için doğruluk (Accuracy-ACC), Kesinlik (Precision) ve Anma (Recall) değerleri hesaplanmıştır. Bu kıstaslar, sınıflandırma performansını ölçmede yaygın olarak kullanılmakta olup Eşitlik (1), (2), (3) ve (4)'te verilen formüller ile hesaplanmaktadır:

$$Doğruluk = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

$$Kesinlik = \frac{TP}{TP+FP} \quad (2)$$

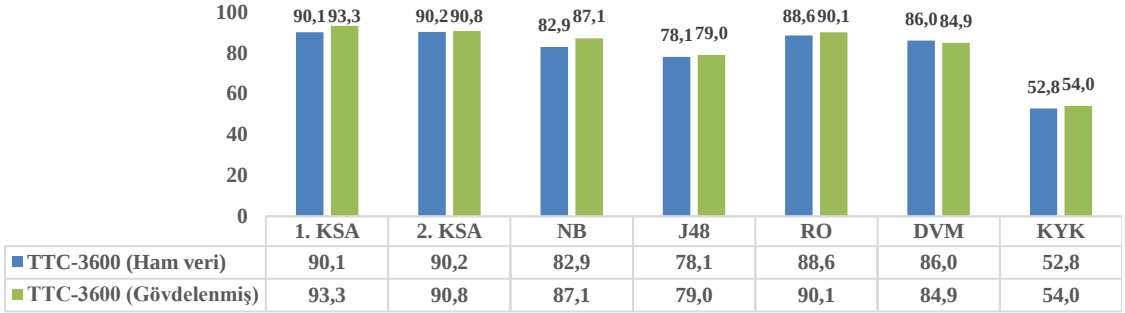
$$Anma = \frac{TP}{TP+FN} \quad (3)$$

$$F-Skoru = 2 * \frac{Kesinlik*Anma}{Kesinlik+Anma} \quad (4)$$

Her iki KSA modeli ile yapılan testler sonucunda elde edilen en iyi doğruluk yüzdeleri Şekil 6'da verilmiştir. Birinci KSA modelinde gövdelenmiş veride %93,3, ham veride ise %90,1 doğruluk değeri alınmıştır. İkinci KSA modeli ise gövdelenmiş veri üzerinde %90,8, ham veri üzerinde %90,2 doğruluk performansı göstermiştir. İki KSA modeli de gövdelenmiş veri için ham veriye kıyasla daha yüksek doğruluk yüzdeleri üretmiştir. En iyi sonuç %93,3 ile birinci KSA modelinin Zemberek ile gövdelenmiş veri kümesinde elde edilmiştir. Çalışmadan alınan sonuçlar, Kılınç vd. [33] çalışmasından elde edilen sonuçlar ile kıyaslandığında, önceki çalışmada en iyi performans gösteren RO algoritmasının (ham veride %88,6, Zemberek ile gövdelenmiş veride %90,1) test sonucundan %3'ten fazla bir doğruluk artışı sağlanmıştır.

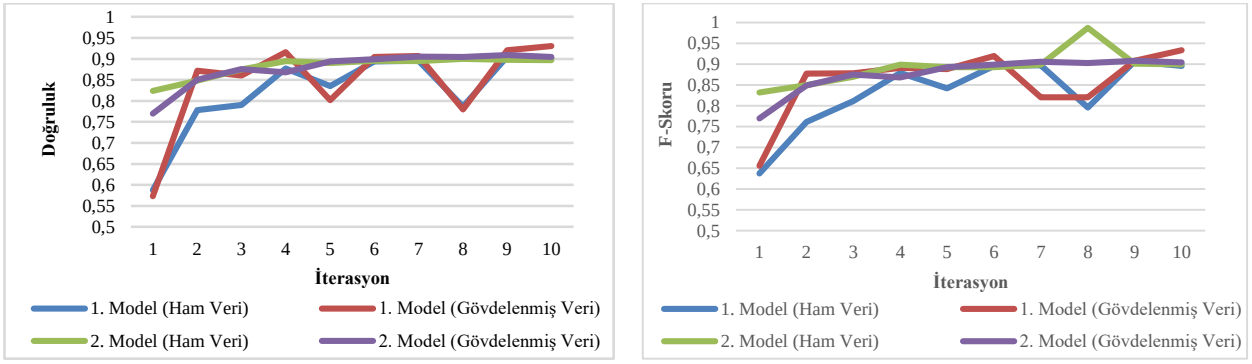
Şekil 7'de birinci ve ikinci KSA modelinin 10-katlı çapraz doğrulama esnasında elde edilen doğruluk oranları görülmektedir. Şekil 8 ise sınıflama performanslarını karmaşıklık matrisleri üzerinde ifade etmektedir. Her iki şekil de incelendiğinde, birinci model veri kümesi üzerinde yüksek performans göstermesine karşın, aşırı uyum (overfitting) göstermeye meyillidir. Şekil 7'den de görüldüğü üzere birinci model, iki veri kümesi üzerinde de 6. iterasyondan sonra düşüş göstermektedir. Fakat ilerleyen iterasyonlarda bu durum düzelmektedir. İkinci model ise birinci modele göre daha kararlı bir davranış göstermiştir. %90 doğruluk sınırlarına kısa bir iterasyonda ulaşmış ve bu sınırda sabit kalmıştır.

Sonuç olarak, literatürde son yıllarda en çok kullanılan sınıflandırma tekniklerinin kullanıldığı ve performansının doğruluk oranı ile ölçüldüğü bu çalışma, Derin Öğrenme ve Word2Vec metodunun performansının diğer metotlarla karşılaştırılması açısından bir ölçüt olmuştur. Ayrıca, çoğunlukla görüntü ve ses işleme çalışmalarında kullanılan KSA'ların, metin sınıflama uygulamalarında klasik makine öğrenmesi ve veri madenciliği algoritmalarına kıyasla daha iyi performans gösterebileceğinin ispatlanması açısından da kayda değerdir. Bu bağlamda, Word2Vec algoritmasının Derin Öğrenme mimarisinde kullanılması ile elde edilen sonuçlar, kelime temsilinin KSA'ların eğitimi için önemini vurgulamaktadır.



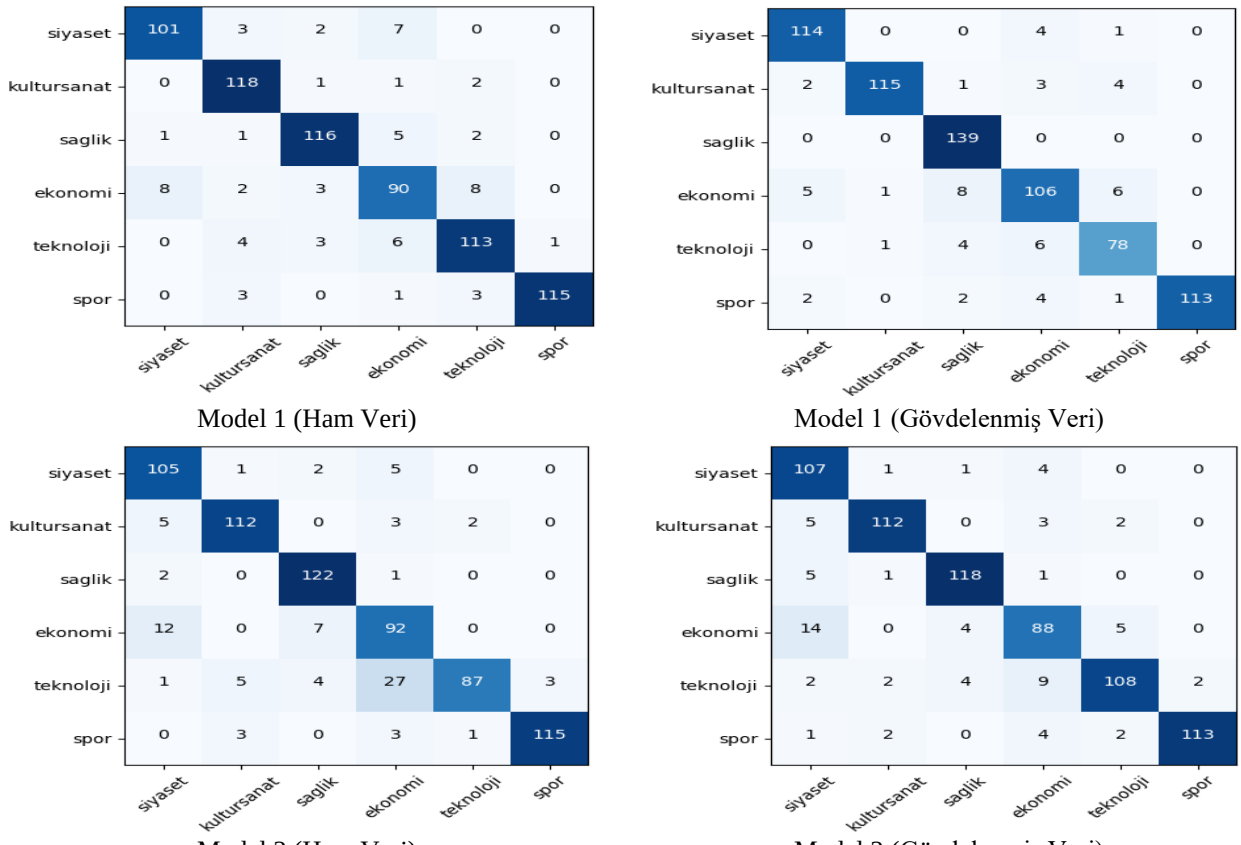
	1. KSA	2. KSA	NB	J48	RO	DVM	KYK
TTC-3600 (Ham veri)	90,1	90,2	82,9	78,1	88,6	86,0	52,8
TTC-3600 (Gövdelenmiş)	93,3	90,8	87,1	79,0	90,1	84,9	54,0

Şekil 6. Birinci ve ikinci KSA modellerinden alınan en yüksek ACC değerleri (%) ile Kılınç vd. [33] çalışmasından alınan en yüksek ACC değerlerinin (%) karşılaştırılması (The comparison of the maximum ACC values (%) obtained from the first and second CNN models with the maximum ACC values (%) obtained from the study Kılınç et al. [33])



(a) (b)

Şekil 7. 10-katlı çapraz doğrulama sonucunda elde edilen (a) doğruluk ve (b) F-Skoru oranları (The ACC (a) and F-Score (b) results of the 10-fold cross validation)



Model 1 (Ham Veri) Model 1 (Gövdelenmiş Veri)

Model 2 (Ham Veri) Model 2 (Gövdelenmiş Veri)

Şekil 8. Karmaşıklık matrisleri (Confusion matrixes)

TEŞEKKÜR (ACKNOWLEDGMENT)

TTC-3600 veri kümesini çalışmamızda kullanmamıza izin verdiği için Doç. Dr. Deniz KILINÇ ve çalışma arkadaşlarına teşekkür ederiz.

Bu çalışma, Mersin Üniversitesi Bilimsel Araştırma Projeleri Birimi (BAP) tarafından 2019-1-TP2-3436 kodlu proje kapsamında desteklenmiştir.

KAYNAKLAR (REFERENCES)

[1] Internet: World Internet Statistics. https://www.internetworldstats.com/stats.htm, 23.05.2019.

[2] Internet: International Data Corporation. https://www.idc.com/, 23.05.2019.

[3] N. Indurkhya, F.J. Damerau, **Handbook of Natural Language Processing**, Chapman & Hall/CRC, 2010.

[4] C.D. Manning, P. Raghavan, H. Schütze, **Introduction to information retrieval**, Cambridge University Press, 2008.

[5] E. Alpaydin, **Machine learning : The New AI**, The MIT Press, 2016.

[6] H. Uğuz, "A two-stage feature selection method for text

categorization by using information gain, principal component analysis and genetic algorithm", *Knowledge-Based Systems*, 24(7), 1024–1032, 2011.

[7] S. Jiang, G. Pang, M. Wu, L. Kuang, "An improved K-nearest-neighbor algorithm for text categorization", *Expert Systems with Applications*, 39(1), 1503–1509, 2012.

[8] T. Jo, "Normalized table-matching algorithm as approach to text categorization", *Soft Computing*, 19(4), 839–849, 2015.

[9] B. Tang, H. He, P.M. Baggenstoss, S. Kay, "A Bayesian Classification Approach Using Class-Specific Features for Text Categorization", *IEEE Transactions on Knowledge and Data Engineering*, 28(6), 1602–1606, 2016.

[10] M.F. Amasyali, T. Yıldırım, "Automatic text categorization of news articles", **Proceedings of the IEEE 12th Signal Processing and Communications Applications Conference**, Kusadasi, Turkey, 224–226, 28-30 April 2004.

[11] R. Johnson, T. Zhang, "Effective Use of Word Order for Text Categorization with Convolutional Neural Networks", *arXiv:1412.1058v2*, 2014.

[12] X. Zhang, J. Zhao, Y. LeCun, "Character-level Convolutional Networks for Text Classification", **Advances in Neural Information Processing Systems**, 649-657, Curran Associates Inc., 2015.

[13] G. Biricik, "Sınıf Bilgisini Kullanan Boyut İndirgeme Yöntemlerinin Metin Sınıflandırmadaki Etkilerinin Karşılaştırılması", **20. IEEE Sinyal İşleme ve İletişim Uygulamaları Kurultayı**, Muğla, Türkiye, 1–4, 2012.

[14] A. Haltaş, A. Alkan, M. Karabulut, "Metin Sınıflandırmada Sezgisel Arama Algoritmalarının Performans Analizi", *Journal of the Faculty of Engineering and Architecture of Gazi University*, 30(3), 2015.

[15] D. Kılınç, E. Borandağ, F. Yücalar, V. Tunalı, M. Şimşek, A. Özçift, "KNN Algoritması ve R Dili ile Metin Madenciliği Kullanılarak Bilimsel Makale Tasnifi", *Marmara Fen Bilimleri Dergisi*, 28(3), 89–94, 2016.

[16] Ç. Çatal, K. Erbakırcı, Y. Erenler, "Computer-based Authorship Attribution for Turkish Documents", **Turkish Symposium on Artificial Intelligence and Neural Networks**, 2003.

[17] M. F. Amasyalı, B. Diri, F. Türkoğlu, "Farklı Özellik Vektörleri ile Türkçe Dokümanların Yazarlarının Belirlenmesi", **15th Turkish Symposium on Artificial Intelligence and Neural Network,** Muğla, Türkiye, 2006.

[18] S. Doğan, B. Diri, "Türkçe Dokümanlar İçin N-gram Tabanlı Yeni Bir Sınıflandırma(Ng-ind): Yazar, Tür ve Cinsiyet", *Türkiye Bilişim Vakfı Bilgisayar Bilimleri ve Mühendisliği Dergisi*, 3, 11–20, 2010.

[19] G. Biricik, B. Diri, "Impact of a New Attribute Extraction Algorithm on Web Page Classification", **5th International Conference on Data Mining**, Las Vegas. A.B.D., 2009.

[20] R. Aşlıyan, K. Günel, "Metin İçerikli Türkçe Dokümanların Sınıflandırılması", **Akademik Bilişim Konferansı,** 659–665, 2010.

[21] İ. Türkmen, B. Diri, G. Biricik, R. Doğan, "Konuşma Dili Kullanılarak Demografik Bilgilerin Sınıflandırılması" **IEEE 19. Sinyal İşleme ve İletişim Uygulamaları Kurultayı**, Antalya, Türkiye, 2011.

[22] M.F. Amasyalı, S. Balcı, E. Mete, E.N. Varlı, "Türkçe Metinlerin Sınıflandırılmasında Metin Temsil Yöntemlerinin Performans Karşılaştırılması", *EMO Bilimsel Dergi*, 2(4), 2012.

[23] V.E. Levent, B. Diri, "Türkçe Dokümanlarda Yapay SinirAğları ile Yazar Tanıma", **15. Akademik Bilişim Konferansı**, 735–741, Mersin, 2014.

[24] S.Ş. Hüsem, **Veri madenciliği teknikleriyle Türkçe web sayfalarının kategorize edilmesi**, Yüksek Lisans Tezi, Fatih Sultan Mehmet Vakıf Üniversitesi, Mühendislik ve Fen Bilimleri Enstitüsü, 2017.

[25] Internet: Kemik - Veri Kümeleri. http://www.kemik.yildiz.edu.tr/?id=28, 23.05.2019.

[26] M.F. Amasyalı, T. Yıldırım, "Otomatik Haber Metinleri Sınıflandırma", **13.Sinyal İşleme ve Uygulama Kurultayı**, 224–226, Kuşadası, Türkiye, 2004.

[27] M.F. Amasyalı, A. Beken, "Türkçe Kelimelerin Anlamsal Benzerliklerinin Ölçülmesi ve Metin Sınıflandırmada Kullanılması", **IEEE 17. Sinyal İşleme ve İletişim Uygulamaları Kurultayı**, Antalya, 2009.

[28] P. Tüfekci, E. Uzun, "Türkçe Dilbilgisi Özelliklerini Kullanarak Web Tabanlı Haber Metinlerinin Sınıflandırılması", **21. IEEE Sinyal İşleme ve İletişim Uygulamaları Kurultayı**, Girne, KKTC, 24-26 Nisan, 2013.

[29] B. Altınel, M.C. Ganiz, B. Diri, "A novel higher-order semantic kernel for text classification", **International Conference on Electronics, Computer and Computation (ICECCO)**, Ankara, 216–219, 2013.

[30] A. Guran, M.C. Ganiz, H.S. Naiboglu, H.O. Kaptikacti, "NMF based dimension reduction methods for Turkish text clustering", *Innovations in Intelligent Systems and Applications*, 1–5, 2013.

[31] F. Baskaya, I. Aydin, "Haber metinlerinin farklı metin madenciliği yöntemleriyle sınıflandırılması", **International Artificial Intelligence and Data Processing Symposium (IDAP)**, Malatya, 1–5, 2017.

[32] O. Kaynar, Z. Aydın, Y. Görmez, "Sentiment Analizinde Öznitelik Düşürme Yöntemlerinin Oto Kodlayıcılı Derin Öğrenme Makinaları ile Karşılaştırılması", *Bilişim Teknolojileri Dergisi*, 10(3), 319 - 326, 2017.

[33] D. Kılınç, A. Özçift, F. Bozyigit, P. Yıldırım, F. Yücalar, E. Borandag, "TTC-3600: A new benchmark dataset for Turkish text categorization", *Journal of Information Science*, 43(2), 174–185, 2017.

[34] H.H. Aghdam, E.J. Heravi, **Guide to convolutional neural networks : a practical application to traffic-sign detection and classification,** Springer, A.B.D., 2017.

[35] Y. Goldberg, O. Levy, "word2vec Explained: deriving Mikolov et al.'s negative-sampling word-embedding method", *arXiv:1402.3722*, 2014.

[36] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, J. Dean, "Distributed Representations of Words and Phrases and their Compositionality", **26th International Conference on Neural Information Processing Systems**, 3111-3119, Nevada, A.B.D., 2013.

[37] Internet: UCI-Machine Learning Repository.

https://archive.ics.uci.edu/ml/datasets/TTC-3600%3A+Benchmark+dataset+for+Turkish+text+categorization, 23.05.2019.

[38] A.A. Akın, M.D. Akın, "Zemberek, an open source NLP framework for Turkic Languages", *Structure*, 10, 1-5, 2007.

[39] Internet: Google Code Archive- Zemberek, https://code.google.com/archive/p/zemberek/, 23.05.2019.

[40] A. Koç, "Eğitim Kümesi Seçiminin Kelime Temsillerine Etkisi ve Türkçe için Benzerlik Test Kümesi", **24. IEEE Sinyal İşleme ve İletişim Uygulamaları Kurultayı**, Zonguldak, Türkiye,16-19 Mayıs 2016.

[41] T. Mikolov, K. Chen, G. Corrado, J. Dean, "Efficient Estimation of Word Representations in Vector Space", *arXiv:1301.3781*, 2013.

[42] O. Karasoy, S. Balli, "Classification Turkish SMS with deep learning tool Word2Vec", **International Conference on Computer Science and Engineering (UBMK)**, Antalya, Türkiye, 5-8 Ekim, 2017.

[43] D. Ayata, M. Saraclar, A. Özgür, "Makine Öğrenmesi Ve Kelime Vektör Temsili İle Türkçe Tweet Sentiment Analizi", **25. Sinyal İşleme Ve İletişim Uygulamaları Kurultayı**, Antalya, Türkiye, 15-18 Mayıs, 2017.

[44] G. Şahin, "Word2Vec ve SVM Tabanlı Türkçe Doküman Sınıflandırma", **25. Sinyal İşleme Ve İletişim Uygulamaları Kurultayı**, Antalya, Türkiye, 15-18 Mayıs, 2017.

[45] P.U. Hatipoğlu, Y.O. Artan, A. Atvar, "Metin Madenciliği Yöntemleri ile Yazılım Gereksinim İzlenebilirliğinin Analizi", **25. Sinyal İşleme Ve İletişim Uygulamaları Kurultayı**, Antalya, Türkiye, 15-18 Mayıs, 2017.

[46] J. Lilleberg, Y. Zhu, Y. Zhang, "Support vector machines and Word2vec for text classification with semantic features", **14th International Conference on Cognitive Informatics & Cognitive Computing**, 136–140, 2015.

[47] M. Seyfioğlu, M. Demirezen, "A Hierarchical Approach for Sentiment Analysis and Categorization of Turkish Written Customer Relationship Management Data", **Federated Conference on Computer Science and Information Systems**, 361–365, 2017.

[48] L. Ge, **Improving Text Classification with Word Embedding,** Master of Science Thesis, San José State University, 2017.

[49] A. Hayran, M. Sert, "Kelime Gömme ve Füzyon Tekniklerine Dayalı Mikroblog Verileri Üzerinde Duygu Analizi", **25. Sinyal İşleme Ve İletişim Uygulamaları Kurultayı**, Antalya, Türkiye, 15-18 Mayıs, 2017.

[50] Ö. Çoban, I. Karabey, "Kelime ve Doküman Vektörleri ile Müzik Türü Sınıflandırması", **25. Sinyal İşleme Ve İletişim Uygulamaları Kurultayı**, Antalya, Türkiye, 15-18 Mayıs, 2017.

[51] E. Esen, S. Özkan, "TBMM Tutanaklarının Parti Bağdaşıklığı Açısından Analizi", **25. Sinyal İşleme Ve İletişim Uygulamaları Kurultayı**, Antalya, Türkiye, 15-18 Mayıs, 2017.

[52] O. Güngör, E. Yıldız, "Türkçe Sözcük Temsillerinde Dilbilimsel Özellikler", **25. Sinyal İşleme Ve İletişim Uygulamaları Kurultayı,** Antalya, Türkiye, 15-18 Mayıs, 2017.

[53] D. Ayata, M. Saraclar, A. Özgür, "Uzun-Kısa Süreli Bellek Özyinelemeli Ağlar ile Politik Yönelimlerin/Duyguların Twitter Verisi üzerinden Tahminlenmesi", **25. Sinyal İşleme Ve İletişim Uygulamaları Kurultayı**, Antalya, Türkiye, 15-18 Mayıs, 2017.

[54] M. Bilgin,, "Kelime Vektörü Yöntemlerinin Model Oluşturma Sürelerinin Karşılaştırılması", *Bilişim Teknolojileri Dergisi*, 12(2), 141 - 146, 2019.

[55] H. Polat, M. Körpe, "TBMM Genel Kurul Tutanaklarından Yakın Anlamlı Kavramların Çıkarılması." *Bilişim Teknolojileri Dergisi*, 11(3), 235-244, 2018.

[56] L. Deng, D. Yu, "Deep Learning: Methods and Applications", *Foundations and Trends in Signal Processing*, 7(3–4), 197–387, 2014.

[57] G. Isik, H. Artuner, "Recognition of radio signals with deep learning Neural Networks", **24. IEEE Sinyal İşleme ve İletişim Uygulamaları Kurultayı**, Zonguldak, Türkiye,16-19 Mayıs 2016.

[58] H. Yalçın, "Derin Anlama Ağları ile İnsan Aktiviteleri Tanıma", **Türkiye Robotbilim Konferansı**, İstanbul, 26 - 27 Ekim 2015.

[59] E. Cengil, A. Çınar, "A New Approach For Image Classıfıcatıon: Convolutıonal Neural Network", *European Journal of Technic*, 6(2), 2016.

[60] Y.S. Akgül, "Derin Öğrenme ile Göz Tespiti", **24. IEEE Sinyal İşleme ve İletişim Uygulamaları Kurultayı**, Zonguldak, Türkiye, 16-19 Mayıs 2016.

[61] H.K. Ekenel, "Evrişimsel Sinir Ağı Öznitelikleri ile Kişiyi Yeniden Tanıma", **24. IEEE Sinyal İşleme ve İletişim Uygulamaları Kurultayı**, Zonguldak, Türkiye,16-19 Mayıs 2016.

[62] S.E. Yüksel, "Hiperspektral Verilerin Derin Konvolüsyonel Sinir Ağlarıyla Sınıflandırılması", **24. IEEE Sinyal İşleme ve İletişim Uygulamaları Kurultayı**, Zonguldak, Türkiye,16-19 Mayıs 2016.

[63] P. Wang, J. Xu, B. Xu, C.-L. Liu, H. Zhang, F. Wang, H. Hao, "Semantic Clustering and Convolutional Neural Network for Short Text Categorization", **Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics**, 352–357, 2015.

[64] D.J. Wu, "End-to-End Text Recognition with Convolutional Neural Networks in SearchWorks catalog", **Stanford Digital Repository**, 2012.

[65] N. Kalchbrenner, E. Grefenstette, P. Blunsom, "A Convolutional Neural Network for Modelling Sentences", *arXiv:1404.2188*, 2014.

[66] C. Nogueira, D. Santos, M. Gatti, "Deep Convolutional Neural Networks for Sentiment Analysis of Short Texts", **25th International Conference on Computational Linguistics: Technical Papers**, 69–78, Dublin, 2014.

[67] D. Tang, B. Qin, T. Liu, "Document Modeling with Gated Recurrent Neural Network for Sentiment Classification", **Conference on Empirical Methods in Natural Language Processing**, 1422–1432, Lisbon, 2015.

[68] P. Wang, B. Xu, J. Xu, G. Tian, C.-L. Liu, H. Hao, "Semantic expansion using word embedding clustering and convolutional neural network for improving short text classification", *Neurocomputing*, 174(B), 806–814, 2016.

[69] B.Hu, Z. Lu, H. Li, Q. Chen, "Convolutional Neural Network Architectures for Matching Natural Language Sentences", **Advances in Neural Information Processing Systems**, 2042–2050, 2014.

[70] Internet: T. Ieleman, G. Hinton, "RMSProp, Neural Networks for Machine Learning", https://www.cs.toronto.edu/~tijmen/csc321/slides/lecture_slides_lec6.pdf, 23.05.2019.

[71] D.P.Kingma, J. Ba, "Adam: A Method for Stochastic Optimization", **3rd International Conference for Learning Representations**, San Diego, 2015.

[72] Internet: Python. https://www.python.org/doc/, 23.05.2019.

[73] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, …, "TensorFlow: A system for large-scale machine learning" **12th USENIX Symposium on Operating Systems Design and Implementation**, 265–283, Savannah, 2016.

[74] Internet: Keras. https://github.com/keras-team/keras, 23.05.2019.