Int. J. Pure Appl. Sci. 9(1);175-187 (2023)

IJPAS
ijpas@munzur.edu.tr
ISSN: 2149-0910

Research article/Araştırma makalesi
DOI:10.29132/ijpas.1265301

GloVe Kelime Gömmeleri ve Sinir Ağları ile Haber Metinlerinin Sınıflandırılması

Hülya Hark[1]*, Meral Karakurt[2], Cengiz Hark[1], Ali Karcı[3]
[1]*İnönü Üniversitesi, Bilgisayar Mühendisliği Bölümü, Malatya, Türkiye*
[2]*Osmaniye Korkut Ata Üniversitesi, Bilgisayar Teknolojileri Bölümü, Osmaniye, Türkiye*
[3]*İnönü Üniversitesi, Yazılım Mühendisliği Bölümü, Malatya, Türkiye*
hulyahark@hotmail.com.tr, meralkarakurt@osmaniye.edu.tr, cengiz.hark@inonu.edu.tr, ali.karci@inonu.edu.tr
Makale gönderme tarihi:14.03.2023, Makale kabul tarihi:02.05.2023

Öz

Dijital haberlerin artan miktarları, istenilen türdeki haberlere doğru ve hızlı bir şekilde erişim için haber metinlerinin kategorilere ayrılmasını gerektirmektedir. Bu çalışmada, ön-eğitimli kelime gömülmelerinin, Uzun Ömürlü Kısa Dönem Bellek Ağı (Long-Short Term Memory, LSTM) ve Evrişimsel Sinir Ağları (Convolutional Neural Network, CNN) gibi derin öğrenme modelleri üzerindeki etkisi araştırılmaktadır. Global Vektör (GloVe) kelime gömülmelerinden alınan bağlamsal temsilleri girdi olarak alan LSTM ve CNN ağları kullanılarak haber metinleri sınıflandırılmıştır. Kapsamlı ve karşılaştırmalı araştırmaların eksikliği nedeniyle GloVe gömme katmanı tarafından sağlanan bağlamsal temsiller farklı sınıflandırıcılar ve veri setleri üzerinde test edilmektedir. Deneysel süreçler boyunca Türkçe Haber başlıklarından oluşan Turkish Headlines veri seti ve BBC News Classification veri setleri kullanılmıştır. Kelime gömülmelerinin ağlar üzerindeki etkisini ortaya koymak için deneysel süreçler aynı parametreler ile tekrarlanmıştır. LSTM modelinde Glove kelime gömülme yöntemi kullanıldığında modelin başarısının %81'den %91'e çıktığı gözlemlenmektedir. CNN modelinde ise Glove kelime gömülmelerinin modelin başarısının olumlu yansımadığı görülmektedir.

Anahtar Kelimeler: Yapay sinir ağları, derin öğrenme, Evrişimsel Sinir Ağları(CNN), GloVe.

Classification of News Texts with GloVe Word Embeddings and Neural Networks

Abstract

Increasing amounts of digital news require categorization of news texts for accurate and fast access to the desired type of news. This study investigates the effect of pre-trained word embeddings on deep learning models such as Long-Short Term Memory Networks (LSTM) and Convolutional Neural Networks (CNN). News texts were classified using LSTM and CNN networks, which received contextual representations from Global Vector (GloVe) word embeddings. The Turkish Headlines dataset, consisting of Turkish News headlines, was used during the experimental processes. Experimental processes were repeated with the same parameters to reveal the effect of word embeddings on networks. When the Glove word embedding method is used in the LSTM model, it is observed that the model's success increases from 81% to 91%. In the CNN model, on the other hand, it is seen that the Glove word embeddings do not positively reflect the model's success.

Keywords: Artificial neural networks, deep learning, Convolutional Neural Networks (CNN), GloVe.

GİRİŞ

Dil, insanlık tarihinde çok önemli bir yere sahiptir. İnsanların iletişim kurması, bilgilerin nesilden nesile aktarılması ve bilimin ilerlemesi dil ile mümkün olmaktadır. Bilgisayar teknolojilerinin artması ve internetin yaygınca kullanılmasıyla beraber internet üzerinden bilgi paylaşımı artmakta ve bu bilgileri çeşitli amaçlarla işlemek için birçok yapay zeka ve makine öğrenmesi yöntemi geliştirilmektedir. Yapay zeka, canlıların öğrenme sürecini matematiksel olarak ifade etmeyi amaçlayan

IJPAS
ijpas@munzur.edu.tr
ISSN: 2149-0910

Research article/Araştırma makalesi
DOI:10.29132/ijpas.1265301

bir bilim dalıdır. Makine öğrenmesi, makinelerin insan beyni gibi öğrenebilmesi ve insan gibi davranabilmesini amaçlayan bir yapay zeka kavramıdır. İnsan beyninde öğrenme işlemi, nöron adı verilen sinir hücrelerinin işlevleriyle gerçekleşmektedir. Makinelerde biyolojik nöronlar gibi işlem elemanları olarak yapay nöronlar kullanılmaktadır (Karakurt vd, 2022). Yapay sinir hücrelerinin öğrenme sürecini tamamlayıp karar verme işlemlerini gerçekleştirdiği yapının tümüne de Yapay Sinir Ağları (YSA) denilmektedir.

Doğal Dil İşleme (DDİ), bir dilde kullanılan yazılı veya sözlü ifadelerin, bilgisayarlar tarafından işlenme sürecidir (Şeker, 2015). Bilgisayarların insanlarla etkileşimini sağlayan DDİ yöntemleriyle tasarlanan teknolojik ürünler daha çok tercih edilmektedir. DDİ çalışmaları, son dönemlerde metin verilerinin artmasıyla oldukça popüler olmuştur. YSA'lar, metin üretme, metin sınıflandırma, duygu analizi, otomatik çeviri, konuşma tanıma, soru-cevap sistemlerinin geliştirilmesi, yazım yanlışlarının otomatik düzeltilmesi ve daha birçok görev için kullanılmaktadır (Adalı, 2016),(Aydoğan ve Karcı, 2019a),(Aydoğan ve Karcı, 2019b).

Yapay sinir hücrelerini modelleyen ilk çalışma McCulloch ve Pitts tarafından 1943 yılında yapılmıştır (McCulloch ve Pitts, 1943). 1958 yılında Rosenblatt, tek katmanlı bir sinir hücresi tasarlamıştır. Algılayıcı (perceptron) adını verdiği bu sinir hücresinin öğrenmesini bir "eşikten geçirme" yöntemiyle gerçekleştirmiştir (Rosenblatt, 1958). 1980'li yıllara kadar sunulan çalışmalar sadece iki seçenekli sonuçları olan doğrusal problemleri çözebilmişlerdir. Ancak gerçek hayattaki çoğu problem doğrusal olmadığından yapay zeka çalışmaları popülerlik kazanamamıştır. Doğrusal olmayan problemleri çözmek için birden fazla nöronun kullanıldığı çok katmanlı yapay sinir ağları 1980'li yıllarda ortaya atılmıştır. 1989 yılında Yann LeCun vd, yapay sinir ağlarının öğrenme sürecinde geri yayılım algoritmasını kullanmışlardır. Bir gerçek dünya problemini çözen ilk başarılı derin öğrenme yöntemi olarak bilinen çalışmalarında, el yazısı rakam tanıma işlemini gerçekleştirmişlerdir (LeCun vd, 1989). 2000'li yıllara kadar gerekli yazılım ve donanımsal alt yapının yetersiz olması nedeniyle yapay zeka çalışmaları popülerlik kazanamamıştır.

2010'lu yıllarda derin öğrenme kavramıyla tanımlanan ve ham veriler kullanılarak yapılan yapay sinir ağı çalışmaları yaygınlaşmıştır. Önceki yıllarda sunulan klasik YSA çalışmalarında verilere ait özellikler (öznitelikler) tanımlandıktan sonra ağa verilirken derin öğrenme yöntemlerinde verilere ait özellik çıkarma işlemi bizzat ağ tarafından yapılmaktadır. Tasarımlarında klasik YSA'ların da kullanıldığı karmaşık yapıya sahip birçok derin öğrenme yöntemi geliştirilmektedir.

Metin sınıflandırma probleminin çözümü için başta İngilizce olmak üzere birçok dünya dilinde metin verileri kullanılmaktadır. Türkçe metin verilerinin sayısı istenilen yeterlilikte değildir. Diri ve Amasyalı tarafından 2003 yılında sunulan ve Türkçe bir gazetede yer alan metinlerin yazarlarını ve türlerini belirleyen çalışmaları, Türkçe metin verilerinin kullanıldığı ilk çalışmalardan biridir (Diri ve Amasyalı, 2003). Yapay sinir ağlarına girdi olarak verilen metin verileri yerine bu verilerin temsillerini kullanarak ağların performans ve hızlarını arttırmada önemli bir yere sahip olan ve word2vec olarak adlandırılan kelime gömülmesi (word embedding) yöntemi, Google'de çalışan Mikolov vd tarafından 2013 yılında ortaya atılmıştır (Mikolov vd,2013).

Bu çalışmada, CNN, LSTM, GloVe CNN ve GloVe LSTM yöntemleri kullanılarak Türkçe Haber başlıklarından oluşan Turkish Headlines Dataset veri seti ile BBC-text sınıflandırılmıştır.

İLGİLİ ÇALIŞMALAR

LeCun vd (1998), gradyan tabanlı öğrenme gerçekleştiren geri yayılımlı bir CNN modeli sunmuşlardır. LeNet-5 olarak adlandırdıkları model ile MNIST veri setini kullanarak el yazısı karakter tanıma işlemi yapmışlardır (LeCun vd, 1998).

Aşlıyan ve Günel (2010), 5 sınıflı Türkçe dokümanları sınıflandırmak için En Yakın Komşu (EYK) ve K-En Yakın Komşu (KYK) yöntemlerini kullanarak yaptıkları çalışmada, %88.4 doğruluk değeriyle EYK metodunun daha başarılı olduğunu göstermişlerdir (Aşlıyan ve Günel, 2010).

Amasyalı, Diri ve Türkoğlu (2006), yazarları bilinmeyen dokümanların, yazarlık özellikleri önceden çıkarılmış 18 yazardan hangisine ait olduğunu belirlemek amacıyla N-gramlar kullanarak oluşturdukları özellik vektörlerini Naive Bayes, Rastgele Orman (RO), Destek Vektör Makinesi (DVM) ve C 4.5 yöntemleriyle sınıflandırmışlardır. Naive Bayes ve DVM yöntemlerinin başarılı olduğunu ve ayrıca N-gramların başarıyı arttırdığını göstermişlerdir (Amasyalı, Diri ve Türkoğlu, 2006).

Int. J. Pure Appl. Sci. 9(1);175-187 (2023)

IJPAS
ijpas@munzur.edu.tr
ISSN: 2149-0910

Research article/Araştırma makalesi
DOI:10.29132/ijpas.1265301

N-gramlar kullanılarak yapılan bir diğer çalışmada, Doğan ve Diri (2010), üç farklı Türkçe veri setini Ng-ind adını verdikleri yeni bir N-gram yöntemi ile yazarlarına, türüne ve yazarın cinsiyetine göre sınıflandırmışlardır. 2-, 3- ve 4-gramlar kullanarak yaptıkları çalışmayı Destek Vektör Makinesi (DVM), K-En Yakın Komşu (KYK) ve Rastgele Orman (RO) yöntemleriyle kıyaslamışlardır. Sonuç olarak cinsiyet ve tür belirlemede Ng-ind yöntemi, diğer yöntemlerden daha başarılı olmuştur (Doğan ve Diri, 2010).

Levent ve Diri (2014), Türkçe gazete köşe yazarlarına ait önceden çıkarılan yazarlık özelliklerini kullanarak tasarladıkları Yapay Sinir Ağı (YSA) ile yazar tanıma işlemini gerçekleştirmişlerdir (Levent ve Diri, 2014).

Süzen (2019), üniversitelere giriş sınavında matematik bölümündeki soruları konularına göre LSTM kullanarak sınıflandırmıştır. Türkiye'de 1981-2018 yılları arasında yapılan üniversite giriş sınavlarında 16 konudan oluşan matematik bölümüne ait 931 sorudan oluşan veri seti ile yaptığı çalışma sonunda test kümesinde ortalama %96.82 doğruluk elde etmiştir. Aynı veri seti üzerinde farklı makine öğrenmesi yöntemleri de kullanarak yaptığı kıyaslamada LSTM' nin daha başarılı olduğunu göstermiştir (Süzen, 2019).

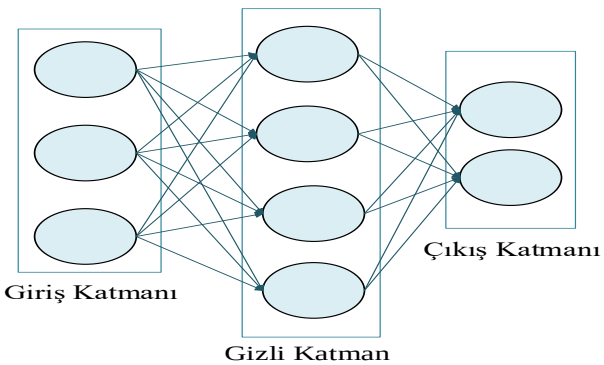
Acı ve Çırak (2019), Türkçe haber metinlerini Word2Vec metodu ile birlikte CNN kullanarak sınıflandırmışlardır. Turkish Text Classification 3600 (TTC-3600) veri setini kullanarak yaptıkları çalışmada %93.3 doğruluk değeri elde etmişlerdir. Sundukları modelin başarımını kıyaslamak için Word2Vec metodu kullanmadan ham verilerle ağı eğittiklerinde başarı değerini %90.1 olarak ölçmüşlerdir. Böylece, Word2Vec yönteminin ağın başarısını arttırdığı görülmektedir (Acı ve Çırak, 2019).

Uçkan vd. (2019), KUSH (Karci-Uçkan-Seyyarer-Hark) adını verdikleri ön işleme aracıyla metinleri bir ön işlemden geçirdikten sonra graf tabanlı bir yöntemle metin sınıflandırma yapmışlardır (Uçkan vd, 2019). Darbaş ve Karcı (2020), metin benzerliklerini ölçmek amacıyla grafların yapısal özelliklerinin kıyaslanmasıyla ölçülen yeni bir graf benzerliği yöntemi önermişlerdir. Bu yöntemle, bir cümlede yer alan bir kelimenin, diğer cümlelerde yer alıp almamasına göre bir benzerlik ölçümü yapmışlardır (Darbaş ve Karcı, 2020).

Hark (2022), sahte haber tespiti yapmak amacıyla Covid-19 sahte haber tespiti veri seti üzerinde Global Vektörler (GloVe) adı verilen ön eğitimli kelime gömülme katmanının sağladığı metin temsilleri ile birlikte Uzun Ömürlü Kısa Dönem Bellek Ağı (LSTM), Tekrarlayan Sinir Ağları (RNN), Evrişimsel Sinir Ağları (CNN) ve Çok Katmanlı Algılayıcı (MLP) yöntemlerini kullanmıştır. İlgili çalışmanın sonunda elde ettiği %91 F-skoru ile LSTM'nin en başarılı model olduğunu göstermiştir (Hark, 2022).

YAPAY SİNİR AĞLARI

Biyolojik sinir hücreleri ve bu hücrelerin birbiriyle olan ilişkilerini bilgisayar ortamında matematiksel olarak modelleyen yöntemler bütününe YSA denir. Bir YSA, tek bir nörondan oluşturulmuşsa Tek Katmanlı YSA veya algılayıcı, birden fazla nörondan ve katmandan oluşturulmuşsa da Çok Katmanlı YSA olarak tanımlanır. Temel bir Çok Katmanlı YSA, Şekil 1'de gösterildiği gibi girdi katmanı, ara (gizli) katman ve çıktı katmanından oluşmaktadır.



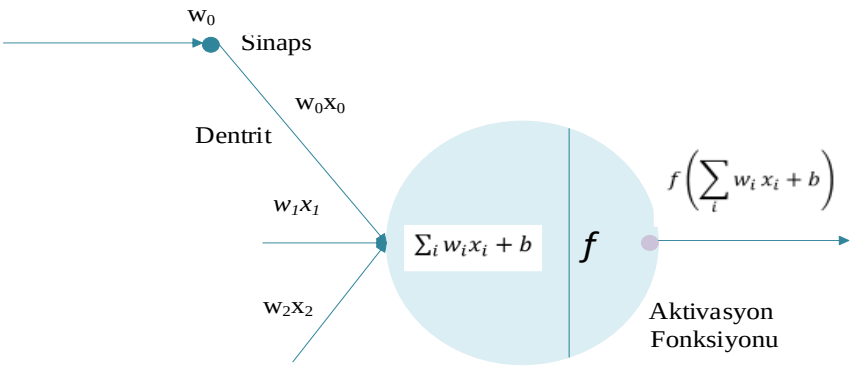
Şekil 1. Çok Katmanlı YSA'da temel katmanlar

Int. J. Pure Appl. Sci. 9(1);175-187 (2023)

IJPAS
ijpas@munzur.edu.tr
ISSN: 2149-0910

Research article/Araştırma makalesi
DOI:10.29132/ijpas.1265301

Bu katmanlarda nöronlar, ağırlıklar, bias, aktivasyon fonksiyonları kullanılarak öğrenme işlemi gerçekleştirilmektedir. Şekil 2 ile gösterilen bir nörona giren girdi değerleri ile ağırlıklarının çarpılması, devamında bu çarpımların bias değeri de eklenerek toplanması ve elde edilen bu toplam sonucunun bir aktivasyon fonksiyonundan geçirilmesiyle elde edilen sonuç o nöronun sayısal olarak çıktı bilgisidir. Aktivasyon fonksiyonu, bir nöronun çıktı değeri üretip üretmeyeceğini (nöronun aktif ya da pasif olacağını) belirleyen bir fonksiyondur (Karakurt ve İşeri, 2022). Örneğin; Sigmoid, Lineer, Softmax, Hiperbolik Tanjant, ReLU, Leaky ReLU, PReLU, Swish ve daha birçok aktivasyon fonksiyonu kullanılmaktadır.

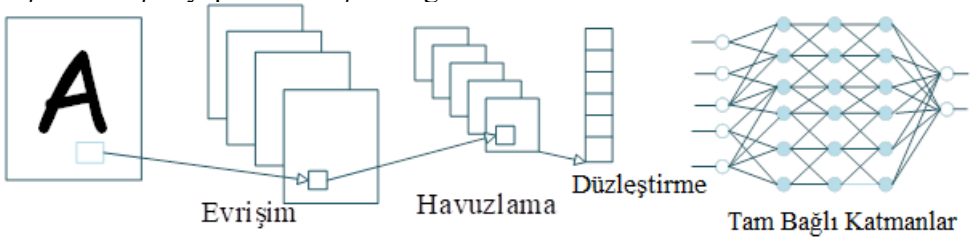


Şekil 2. Tek hücreli YSA

Biyolojik nöronlardaki sinyal bilgisine karşılık gelen yapay nöronların sayısal çıktı bilgisi Çok Katmanlı YSA'larda bir sonraki katman nöronlarına girdi olarak verilir. Böylece, girdi katmanından çıktı katmanına doğru yapılan bu işlemlere ileri yayılım (besleme) denir. İleri yayılım sonucunda elde edilen çıktının beklenen çıktı değeri ile farkı alınarak bir hata değeri hesaplanır. Eğer bu değer tolere edilemeyecek büyüklükte ise hata çıktı katmanından ara katmana doğru gönderilerek her nöronun güncellenmesi sağlanır. Ağın çıktısını beklenen çıktı değerine yaklaştırmak için yapılan bu işleme geri yayılım denir. Geri yayılım işlemi sonunda hatanın en aza düşürülmesi ve ağın başarılı bir öğrenme gerçekleştirmesi beklenir.

Evrişimsel Sinir Ağları (CNN):

CNN'ler, doğrusal değildir ve doğrusal olmayan problemlerin çözümünde başarılı sonuçlar vermektedir. Özellikle görüntü analizinde en sık kullanılan yöntemdir. Bir CNN temelde, Şekil 3'te gösterildiği gibi girdi katmanı, tam bağlı katman, bir veya daha fazla evrişim katmanı, havuzlama katmanı ve aktivasyon katmanından oluşmaktadır.



Şekil 3. Örnek bir CNN mimarisi (https://medium.com/@tuncerergin/convolutional-neural-network-convnet-yada-cnn-nedir-nasil-calisir-97a0f5d34cad)

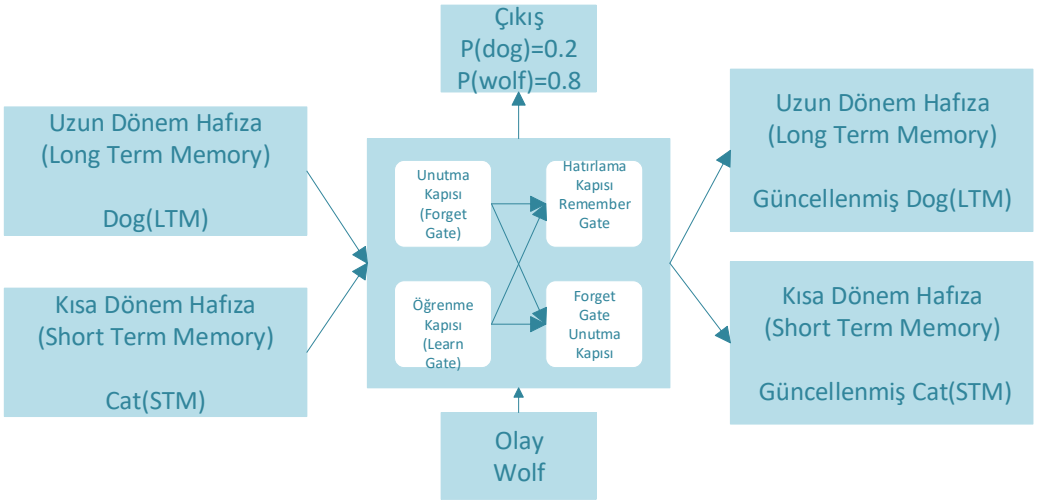
Int. J. Pure Appl. Sci. 9(1);175-187 (2023)

IJPAS
ijpas@munzur.edu.tr
ISSN: 2149-0910

Research article/Araştırma makalesi
DOI:10.29132/ijpas.1265301

Uzun Ömürlü Kısa Dönem Bellek Ağı (Long Short Term Memory – LSTM):

Literatürde Uzun Kısa Süreli Bellek, Uzun Kısa Vadeli bellek gibi isimlerle de anılan LSTM'ler, RNN'lerde oluşan bellek yetersizliği problemini çözmek amacıyla ortaya atılan bir RNN modelidir. RNN'lerde bir birim öncesi hatırlanırken LSTM'lerde daha uzun dönemlere ait bilgiler hatırlanmakta ve dolayısıyla LSTM'ler daha başarılı olmaktadır. Bir LSTM modeli temelde Şekil 4'te gösterildiği gibi giriş kapısı, çıkış kapısı, unutma kapısı ve hafıza (hatırlama) kapısından oluşmaktadır.



Şekil 4. LSTM modeli (Kaynak: Udacity)

Global Vektörler (GloVe):

Metin verileri, doğrudan (ham veriler olarak) bir yapay sinir ağının girdisi olamaz. Bu verilerle çalışabilmek için öncelikle verilerin, ağın işleyebileceği sayısal bilgilere dönüştürülmesi gerekmektedir. Ham metin verileri yerine bu verilerin anlamsal yakınlıklarına göre temsillerinin elde edildiği yöntemlere kelime gömülmesi (word embedding) denmektedir. Kelime gömülmesi yöntemlerinde, bir ağa girdi olarak verilen metin verilerinin temsillerinin elde edildiği bir ön işleme (ön eğitim) adımı gerçekleştirilmektedir. Bu yöntemler ile dile ait sözcükler veya cümleler sayısallaştırılıp birer vektör haline getirilmektedir. Böylelikle metin verileri, sayısal olarak vektör uzayında temsil edilmektedir.

Kelime gömülmesi yöntemleriyle çok büyük veri setlerinden (1.6 milyar kelimeden oluşan bir veri seti gibi), herhangi bir veri kaybı olmadan, daha küçük boyutta temsiller elde edilerek hem hesaplama maliyeti düşürülmekte hem de bu temsillerin girdi olarak verildiği ağların performansı, temsillerin kullanılmadığı ağların performansından daha iyi olmaktadır (Mikolov vd, 2013).

Kelime gömülmesi sürecinde sinir ağı, rastgele değerlerin atandığı vektörlerden temsilleri öğrenerek girdi değerleri olarak kullanılacak olan temsil vektörlerini elde etmektedir. Öğrenme adımında geri besleme yöntemiyle vektörleri güncelleyerek eğitim verisine en uygun kelime gömülme uzayını oluşturacaktır (Hark, 2022).

Bu çalışmada, literatürde var olan bir GloVe kelime gömülme yöntemi kullanılarak ön işlemeden geçen metin verilerine ait temsiller CNN ve LSTM modellerine girdi olarak verilerek sınıflandırma işlemi gerçekleştirilmiştir. Karşılaştırma yapmak için GloVe yöntemi kullanılmayan metin vektörlerini girdi olarak alan CNN ve LSTM modelleriyle de sınıflandırma yapılmıştır.

DENEYSEL SONUÇLAR

Bu çalışmada, Türkçe haber başlıklarını kategorilerine göre sınıflandırmak için literatürde var olan ön eğitimli GloVe kelime gömülmelerinin farklı derin öğrenme ağları üzerindeki performansına etkisi incelenmiştir. Çalışma kapsamındaki tüm deneyler, açık kaynak kodlu bir tümleşik dağıtım olan Anaconda IDE'sinin Jupyter notebook programında Python dili kullanılarak gerçekleştirilmiştir. Ayrıca, deneylerde Keras, TensorFlow, sklearn, numpy

IJPAS
ijpas@munzur.edu.tr
ISSN: 2149-0910

Research article/Araştırma makalesi
DOI:10.29132/ijpas.1265301

kütüphaneleri ile ön eğitimli kelime gömülmesi modeli olarak GloVe 100 boyutlu gömülme vektörü kullanılmıştır.

Veri Seti

Bu çalışmada, Türkçe Haber başlıklarından oluşan Turkish Headlines Dataset (UCI Machine Learning Repository: Turkish Headlines Dataset Data Set, 2021) veri seti (VeriSeti_1) ile BBC News Classification (Bose, 2019) veri seti (Veri_Seti2) kullanılmıştır. VeriSeti_1 Tablo 1' de gösterildiği gibi 7 sınıflı ve VeriSeti_2 Tablo 2' de gösterildiği gibi 5 sınıflı haber metinlerinden oluşmaktadır.

Tablo 1. *VeriSeti_1 özellikleri*

Sınıf Bilgisi	İçerik Sayısı
Ekonomi	600
Siyaset	600
Yaşam	600
Teknoloji	600
Magazin	600
Sağlık	600
Spor	600

Tablo 2. *Veri_Seti2 özellikleri*

Sınıf Bilgisi	İçerik Sayısı
Sport	511
Business	510
Politics	417
Tech	401
Entertainment	386

Eğitim Süreci

Çalışmada kullanılan veri setleri, eğitim için %80 ve test için %20 olarak ayrılmıştır. Eğitim verilerinin de %20'si veri doğrulama işlemi için kullanılmıştır. Her veri kümesinin miktarı otomatik veri bölme ile rastgele olarak bölünmüştür.

Performans Metriği

Sınıflandırma algoritmalarının performans değerlendirmesi, nesnel bir ölçüt olarak karmaşıklık matrisi kullanılarak ifade edilebilir. Karmaşıklık matrisindeki sütunlar gerçek değerleri, satırlar ise tahmin değerlerini göstermektedir. Karmaşıklık matrisinden elde edilen ve sistemin performansını değerlendirmek için kullanılan çeşitli ölçütler vardır. Bu çalışmada, performans metriği olarak Accuracy (Doğruluk), Precision (Kesinlik), Recall (Duyarlılık) ve F1-Score (F1-Puanı) kullanılmıştır.

Tablo 3. *Karmaşıklık matrisi*

		Gerçek Sınıfı	
		Pozitif	**Negatif**
Tahmin Sınıfı	**Pozitif**	Doğru Pozitif (TP)	Yanlış Pozitif (FP)
	Negatif	Doğru Negatif (FN)	Yanlış Negatif (FN)

Int. J. Pure Appl. Sci. 9(1);175-187 (2023)

IJPAS
ijpas@munzur.edu.tr
ISSN: 2149-0910

Research article/Araştırma makalesi
DOI:10.29132/ijpas.1265301

Tablo 4. *Performans değerlendirme ölçütleri*

Değerlendirme Ölçütü	Hesaplama Formülü	Tanımı
Doğruluk (Accurarcy)	$\frac{TP+TN}{FP+FN+TP+TN}$	Bütün örnekler içindeki doğruluk oranıdır.
Kesinlik (Precision)	$\frac{TP}{FP+TP}$	Sadece pozitif tahminleriiçinde doğru pozitif oranıdır.
Duyarlılık (Recall)	$\frac{TP}{FN+TP}$	Sadece pozitif örnekler içinde doğru pozitif tahmin oranıdır.
F1-Puanı	$2\frac{\boldsymbol{Precision} * \boldsymbol{Recall}}{\boldsymbol{Precision}+\boldsymbol{Recall}}$	Kesinlik ve duyarlılığı birlikte değerlendirmek için önerilen bir puandır.

Eğitim Modeli

Yapılan çalışmada önerilen modeller üzerinde kullanılmış olan eğitim tur sayısı (iterasyon - epoch), nöron sayısı, öğrenme oranı, kayıp fonksiyon, aktivasyon fonksiyonu, optimize edici fonksiyon değerleri Tablo 5 ile gösterilmektedir.

Tablo 5. *CNN, LSTM, RNN ve GLOVE-LSTM, GLOVE-CNN, GLOVE-RNN eğitim modelleri*

Verilerin Ağa Sunulma Biçimi	Model	Eğitim Tur Sayısı	Nöron Sayısı	Öğrenme Oranı	Kayıp Fonksiyonu	Gizli Katmanı Aktivasyon Fonksiyon	Çıkış Katmanı Aktivasyon Fonksiyon	Vektör Boyutu	Optimizasyon fonksiyonu
KELİME GÖMÜLMESİZ	LSTM	30	128	0,001	Categorical Cross Entropy	Relu	Softmax	-	Adam
	RNN	30	128	0,001	Categorical Cross Entropy	Relu	Softmax	-	Adam
	CNN	30	128	0,001	Categorical Cross Entropy	Relu	Softmax	-	Adam
GLOVE KELİME GÖMÜLMESİ	LSTM	30	128	0,001	Categorical Cross Entropy	Relu	Softmax	300	Adam
	RNN	30	128	0,001	Categorical Cross Entropy	Relu	Softmax	300	Adam
	CNN	30	128	0,001	Categorical Cross Entropy	Relu	Softmax	300	Adam

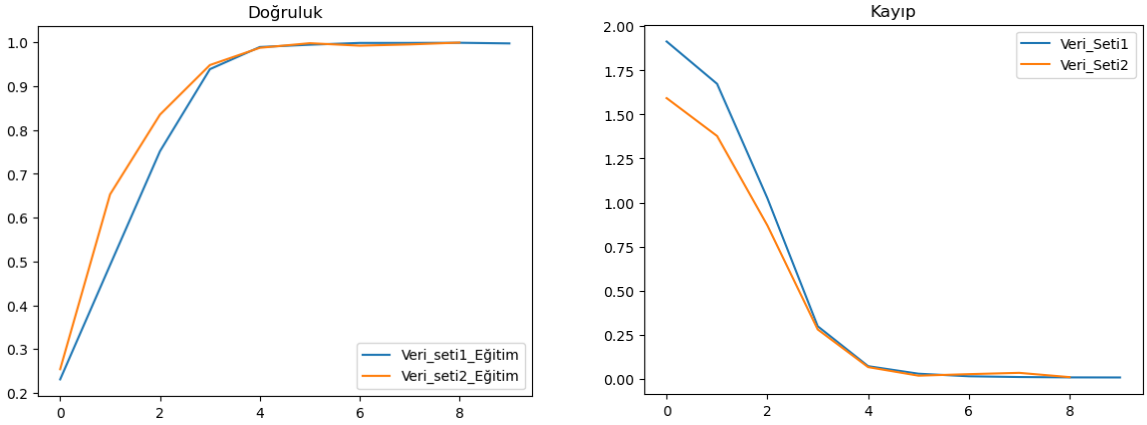
LSTM modeli

LSTM sınıflandırma modeli Tablo 5'te belirtilen hiper parametreler ile 4 katmanlı olarak tasarlanan ağdan elde edilen eğitim doğruluğu VeriSeti_1 için %87, VeriSeti_2 için %82 olarak hesaplanmıştır. LSTM modeline ait eğitim ve test verileri için doğruluk ve kayıp grafikleri Şekil 5 ile gösterilmektedir. Tablo 6' da her iki veri seti için

IJPAS
ijpas@munzur.edu.tr
ISSN: 2149-0910

Research article/Araştırma makalesi
DOI:10.29132/ijpas.1265301

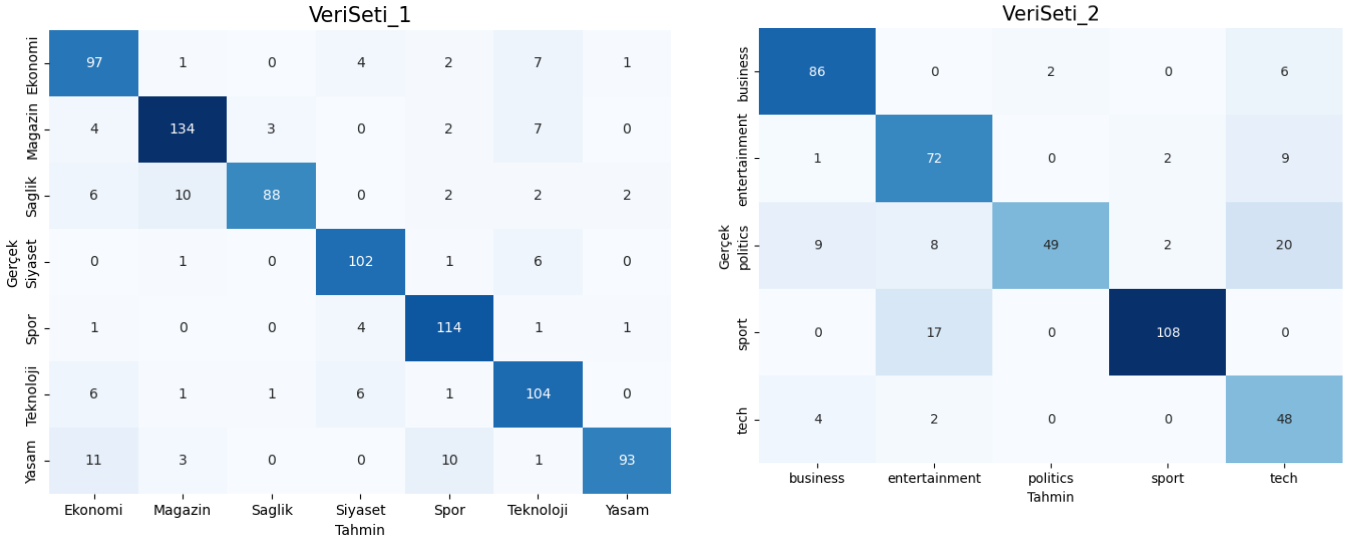
Kesinlik , Duyarlık ve F1-Puanı değerleri ve Şekil 6' da ise her iki veri setine ait LSTM modeli karmaşıklık matrisi gösterilmektedir.



Şekil 5. Modele ait doğruluk ve kayıp grafikleri

Tablo 6. *Sınıflandırma raporu*

Veri Seti	Doğruluk	Kesinlik	Duyarlılık	F1-Puanı
Veri_Seti1	0.87	0.99	0.96	0.93
Veri_Seti2	0.82	0.86	0.91	0.89



Şekil 6. Karmaşıklık matrisi

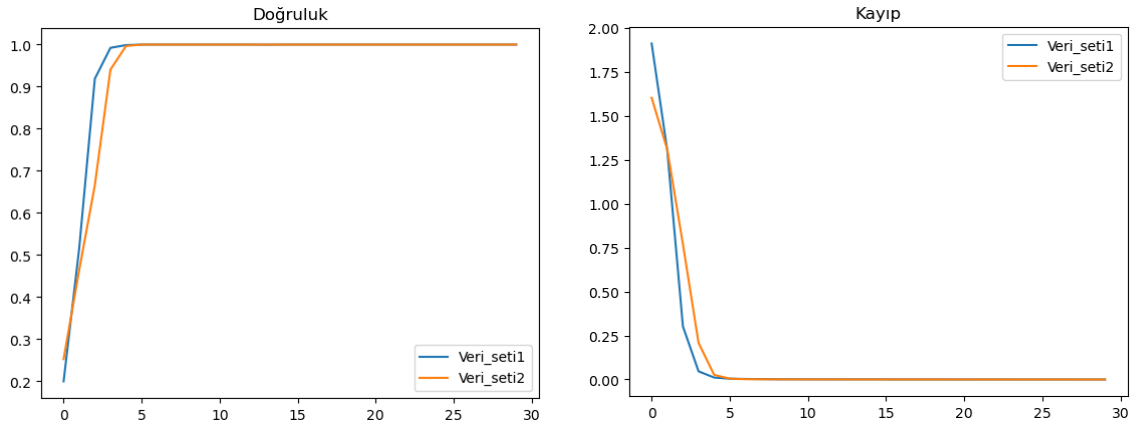
CNN

Tablo 5'te belirtilen hiper-parametreler ile 4 katmanlı olarak tasarlanan CNN modelinin eğitilmesi sonucu elde edilen doğruluk değeri; VeriSeti_1 için %96, VeriSeti_2 için 0.97 olarak hesaplanmıştır. CNN modeline ait doğruluk ve kayıp grafikleri Şekil 7 ile gösterilmektedir. Tablo 7' de her iki veri seti için Kesinlik , Duyarlık ve F1-Puanı değerleri ve Şekil 8'de ise her iki veri setine ait CNN modeli karmaşıklık matrisi gösterilmektedir.

Int. J. Pure Appl. Sci. 9(1);175-187 (2023)

IJPAS
ijpas@munzur.edu.tr
ISSN: 2149-0910

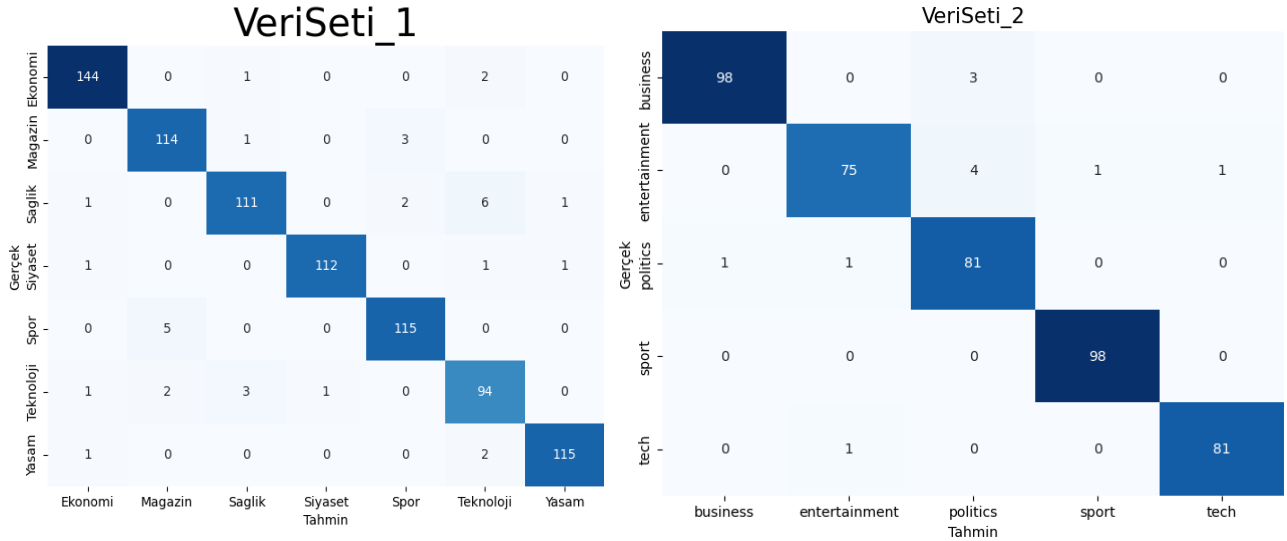
Research article/Araştırma makalesi
DOI:10.29132/ijpas.1265301



Şekil 7. Veri setleri için tasarlanan modele ait doğruluk ve kayıp grafikleri

Tablo7. *Sınıflandırma raporu*

Veri Seti	Doğruluk	Kesinlik	Duyarlılık	F1-Puanı
Veri_Seti1	0.96	0.98	0.98	0.98
Veri_Seti2	0.97	0.97	0.99	0.98



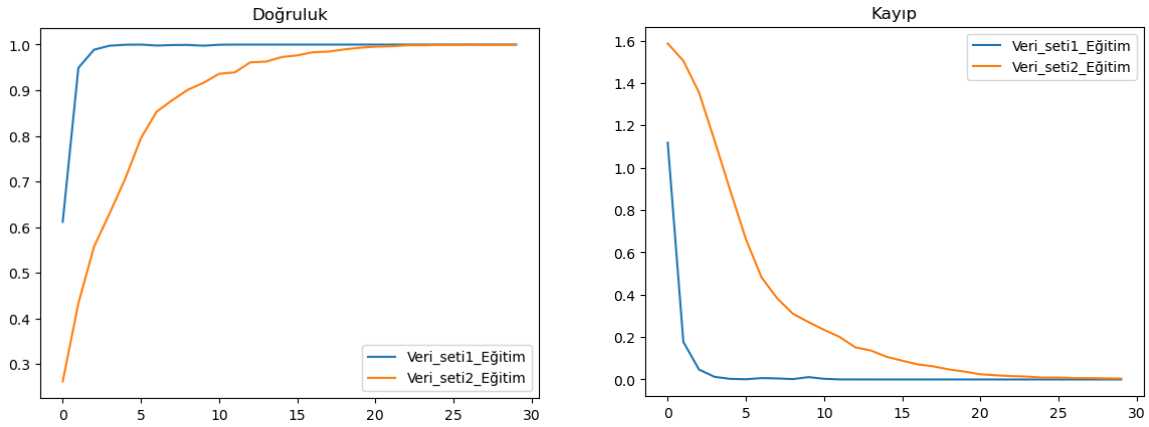
Şekil 8. Karmaşıklık matrisi

GloVe-LSTM

Tablo 5'te belirtilen hiper-parametreler ile 4 katmanlı olarak tasarlanan GloVe-LSTM modelinde 300 boyutlu GloVe kelime yerleştirmesiyle sınıflandırma testi değerlendirmesinin sonunda elde edilen doğruluk değeri VeriSeti_1 için %88 ve VeriSeti_2 için %93 olarak hesaplanmıştır. Modelin doğruluk ve kayıp grafikleri Şekil 9 ile gösterilmektedir. Tablo 8' de her iki veri seti için Kesinlik, Duyarlık ve F1-Puanı değerleri ve Şekil 10'da ise her iki veri setine ait GloVe-LSTM modeli karmaşıklık matrisi gösterilmektedir. Tablo 8 incelendiğinde GloVe kelime yerleştirme ile sistemin performansının arttığı görülmektedir.

IJPAS
ijpas@munzur.edu.tr
ISSN: 2149-0910

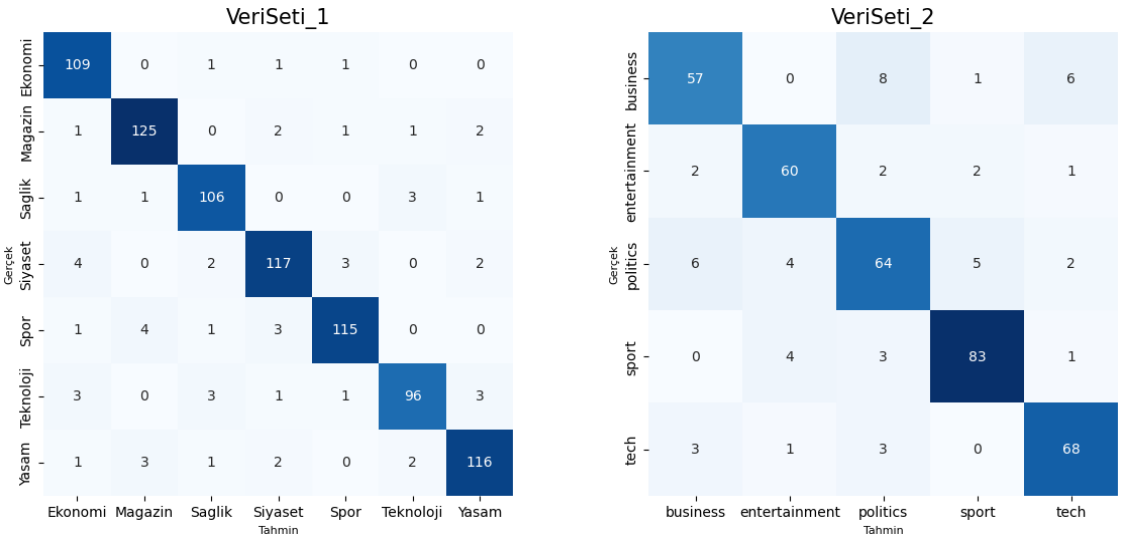
Research article/Araştırma makalesi
DOI:10.29132/ijpas.1265301



Şekil 9. GloVe-LSTM doğruluk ve kayıp grafikleri

Tablo 8. *Sınıflandırma raporu*

Veri Seti	Doğruluk	Kesinlik	Duyarlılık	F1-Puanı
Veri_Seti1	0.93	0.95	0.97	0.94
Veri_Seti2	0.86	0.91	0.91	0.91



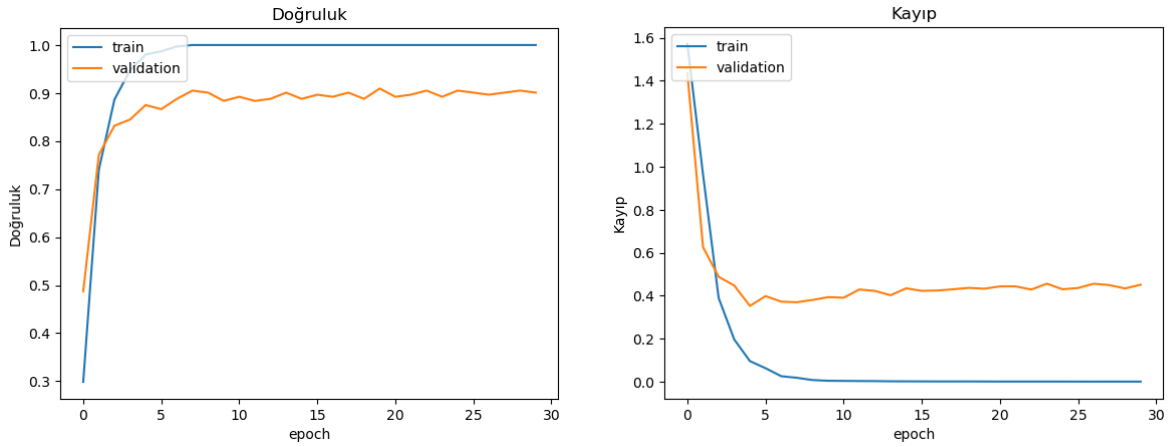
Şekil 10. GloVe-LSTM karmaşıklık matrisi

GloVe-CNN

Tablo 5'te belirtilen hiper parametreler ile 4 katmanlı olarak tasarlanan GloVe-CNN modelinde 300 boyutlu GloVe kelime yerleştirmesiyle sınıflandırma testi değerlendirmesinin performans sonuçları Tablo 9 ile gösterilmektedir. Glove-CNN modelinin eğitimi sonunda elde edilen doğruluk değeri VeriSeti_1 için %89, VeriSeti_2 için %84 olarak hesaplanmıştır. GloVe kelime yerleştirmesi ile kullanılan CNN modelinin klasik CNN modelinden daha düşük sonuçlar verdiğini göstermektedir. Ancak Şekil 11 ile gösterilen doğruluk ve kayıp eğrileri incelendiğinde Glove-CNN modelinin CNN modeline göre daha iyi sonuçlar verdiği görülmektedir. Şekil 12' de modelin karmaşıklık matrisi gösterilmektedir.

IJPAS
ijpas@munzur.edu.tr
ISSN: 2149-0910

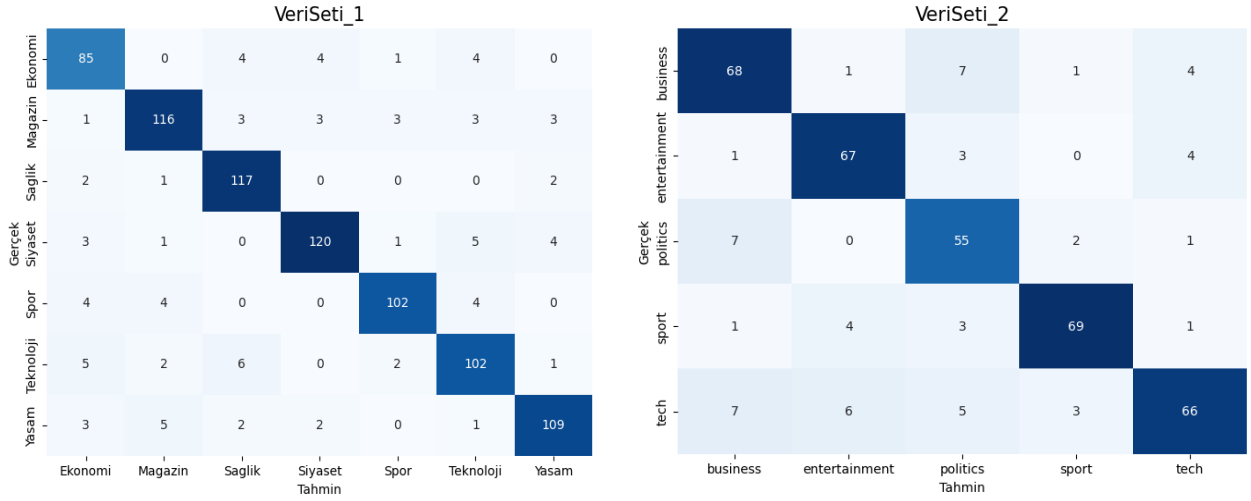
Research article/Araştırma makalesi
DOI:10.29132/ijpas.1265301



Şekil 11. GloVe-CNN doğruluk ve kayıp grafikleri

Tablo 9. *Sınıflandırma raporu*

Veri Seti	Doğruluk	Kesinlik	Duyarlılık	F1-Puanı
Veri_Seti1	0.89	0.94	0.99	0.92
Veri_Seti2	0.84	0.81	0.84	0.82



Şekil 12. GloVe-CNN doğruluk ve kayıp grafikleri

SONUÇLAR

İçinde bulunduğumuz bilgi çağında, istenilen bilgiye erişim hızlı ve doğru bir biçimde olmalıdır. Büyük miktardaki bilgilerin işlenerek anlamlı ve istenilen hale dönüştürülmesi için metinlerin kategorilerine göre sınıflandırılması gerekmektedir. Haber metinlerinin kategorilerine ayrılmasına yönelik olan bu çalışmada, derin öğrenme ağlarından faydalanılmaktadır. Yapılan çalışma, ön eğitimli kelime gömülmelerinin sistemin performansına ne tür bir katkı sağladığını göstermesi açısından büyük önem taşımaktadır.

Bu çalışmada, Türkçe haber başlıklarından oluşan veri setinin 7 farklı kategori olarak sınıflandırılması için LSTM, CNN ve GloVe kelime yerleştirmeyle tasarlanan GloVe-LSTM ve GloVe-CNN modelleri geliştirilmiş ve performans karşılaştırmaları yapılmıştır. GloVe yönteminin eklenmesiyle LSTM modelinin performansının kayda değer bir şekilde yükseldiği gözlemlenmektedir. Fakat CNN modelinde, Glove yönteminin

IJPAS
ijpas@munzur.edu.tr
ISSN: 2149-0910

Research article/Araştırma makalesi
DOI:10.29132/ijpas.1265301

eklenmesiyle sistemin başarısının düşmesine rağmen GLOVE-CNN modelinin doğruluk ve kayıp eğrilerinin CNN modelinden daha iyi olduğu gözlemlenmektedir. Geliştirilen tüm modellerde nöron sayısı, iterasyon sayısı ve yığın değeri, öğrenme paremetresi ve aktivasyon fonksiyonları aynı seçilerek her model aynı şartlarda çalıştırılmıştır. Yapılan deneylerle CNN ve GloVe-CNN modelinin tüm modellerden daha üstün performansa sahip olduğu sonucuna ulaşılmaktadır. Geliştirilen modellerin çok daha büyük veri setleri üzerinde gerçekleştirilmesi ve farklı hiper-paremetre değerleri ile eğitilerek karşılaştırılması çalışmanın sonraki hedefleri arasında yer almaktadır.

ÇIKAR ÇATIŞMASI

Yazarlar bu makaleyle ilgili herhangi bir çıkar çatışması bildirmedi.

ARAŞTIRMA VE YAYIN ETİĞİ BEYANI

Yazarlar bu çalışmanın araştırma ve yayın etiğine uygun olduğunu beyan eder.

KAYNAKLAR

Aci, Ç. ve Çirak, A. (2019). Türkçe Haber Metinlerinin Konvolüsyonel Sinir Ağları ve Word2Vec Kullanılarak Sınıflandırılması. Bilişim Teknolojileri Dergisi, 12(3), 219-228. DOI: 10.17671/gazibtd.457917.

Adalı, E. (2016). Doğal Dil İşleme. Türkiye Bilişim Vakfı Bilgisayar Bilimleri ve Mühendisliği Dergisi, 5 (2).

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

Aydoğan, M. ve Karci, A.(2019a). Turkish Text Classification with Machine Learning and Transfer Learning. 2019 International Artificial Intelligence and Data Processing Symposium (IDAP), Malatya, Turkey, pp. 1-6, doi: 10.1109/IDAP.2019.8875919.

Aydoğan, M. ve Karcı, A. (2019b). Kelime temsil yöntemleri ile kelime benzerliklerinin incelenmesi. Çukurova Üniversitesi Mühendislik-Mimarlık Fakültesi Dergisi, 34(2), 181-196.

Bose, B. (2019). BBC News Classification. Kaggle. https://kaggle.com/competitions/learn-ai-bbc

Darbaş, H. ve Karci, A. (2020). Graf Benzerliği İle Metin Kıyaslama. Computer Science, 5(2), 114-125 . Retrieved from https://dergipark.org.tr/tr/pub/bbd/issue/57870/743751.

Diri, B. ve Amasyalı, M.F. (2003). Automatic Author Detection for Turkish Texts, Artificial Neural Networks and Neural Information Processing, 138-141.

Doğan, S. ve Diri, B. (2010). 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.

Hark, C. (2022). Sahte Haber Tespiti için Derin Bağlamsal Kelime Gömülmeleri ve Sinirsel Ağların Performans Değerlendirmesi. Fırat Üniversitesi Müh. Bil. Dergisi, 34(2), 733-742.

Karakurt, M. (2019). Patoloji Görüntülerinin Derin Öğrenme Yöntemleri İle Sınıflandırılması. Yüksek Lisans Tezi, Ondokuz Mayıs Üniversitesi, Samsun.

Karakurt, M. ve İşeri, İ. (2022). Patoloji Görüntülerinin Derin Öğrenme Yöntemleri İle Sınıflandırılması. Avrupa Bilim ve Teknoloji Dergisi, (33), 192-206.

Karakurt, M., Oymak, E.A., Hark, H., Erdoğan, M.C. ve Karcı, A. (2022). Karcı Sinir Ağlarının Uygulaması ve Performans Analizi. Computer Science, Vol:7, 68-80.

LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W. ve Jackel, L. D. (1989). Backpropagation applied to handwritten zip code recognition. Neural computation, 1(4), 541-551.

LeCun, Y., Bottou, L., Bengio, Y. ve Haffner, P. (1998). Gradientbased learning applied to document recognition. Proceedings of the IEEE, 86(11), 2278-2324.

Levent, V. ve Diri, B., (2014). Türkçe Dokümanlarda Yapay Sinir ağları ile Yazar Tanıma . Akademik Bilişim (pp.1-5). Mersin, Türkiye.

McCulloch, W. S. ve Pitts, W. (1943). A Logical Calculus of the İdeas İmmanent in Nervous Activity. The Bulletin of Mathematical Biophysics, 5(4), 115-133.

Mikolov T., Chen K., Corrado G. and Dean J. (2013). Efficient estimation of word representations in vector space. Proceedings of Workshop at ICLR. Scottsdale, Arizona.

Rosenblatt, F. (1958). The Perceptron: A Probabilistic Model for Information Storage And Organization in the Brain. Psychological review, 65(6), 386.

Süzen, A.A., (2019). LSTM Derin Sinir Ağları İle Üniversite Giriş Sınavındaki Matematik Soru Sayılarının Konulara Göre Tahmini, Engineering Sciences (NWSAENS), 14(3):112-118, DOI: 10.12739/NWSA.2019.14.3.1A0436.

Şeker, S.E., (2015), Doğal Dil İşleme (Natural Language Processing), YBS Ansiklopedi, 2(4), 2015.

IJPAS
ijpas@munzur.edu.tr
ISSN: 2149-0910

Research article/Araştırma makalesi
DOI:10.29132/ijpas.1265301

UCI Machine Learning Repository: Turkish Headlines Dataset Data Set. (2021). Retrieved July 6, 2022, from https://archive.ics.uci.edu/ml/datasets/Turkish+Headlines+Dataset

Uçkan, T., Hark, C., Seyyarer E. ve Karcı A. (2019). Ağırlıklandırılmış Çizgelerde Tf-Idf ve Eigen Ayrışımı Kullanarak Metin Sınıflandırma. Bitlis Eren Üniversitesi Fen Bilimleri Dergisi, 8(4):1349-1362, doi:10.17798/bitlisfen.53122.