T.C. DOĞUŞ ÜNİVERSİTESİ

FEN BİLİMLERİ ENSTİTÜSÜ

BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI

YENİ NESİL DERİN BAĞLAMSALLAŞTIRILMIŞ KELİME GÖSTERİMLERİ VE DERİN ÖĞRENME MODELLERİYLE FİNANSAL HABERLER KULLANARAK BORSA TAHMİNLEMESİ

YÜKSEK LİSANS TEZİ

DERYA OTHAN

20172105035

DANIŞMAN: DR. ÖĞR. ÜYESİ ZEYNEP HİLAL KİLİMCİ

İstanbul, 2019

	YÜKSEK LİSANS TEZ SINAV TUTANAĞI	Doküman No	FR.1.26
		Yürürlük Tarihi	1.11.2017
		Revizyon Tarihi	1.11.2017
		Revizyon No	1
		Sayfa	1 / 1

SOSYAL BİLİMLER / FEN BİLİMLERİ ENSTİTÜSÜ

Tarih :06./12./2019

Anabilim/Anasanat Dalı : Bilgisayar Mühendisliği

Öğrencinin Adı Soyadı : Derya OTHAN

Öğrenci No : 20172105035

Tez Danışmanının Adı Soyadı : Zeynep Hilal KILIÇ

İkinci Tez Danışmanının Adı Soyadı : —

Tezin Başlığı : Yeni Nesil Derin Bağlamsallaştırılmış Kelime Gösterimleri ve Derin Öğrenme Modelleriyle Finansal Haberler Kullanarak Borsa Tahminlemesi

Doğuş Üniversitesi Lisansüstü Eğitim-Öğretim ve Sınav Yönetmeliği'nin 32.Maddesi uyarınca yapılan değerlendirmeler sonunda;

☑ **tezin kabul edilmesine** ☐ **tezde düzeltme verilmesine** ☐ **tezin reddedilmesine**

oy birliği / oy çokluğu ile karar verilmiştir.Gereği için arz olunur.

Danışman Üye
Dr.Öğr.Üyesi Zeynep Hilal KILIÇ

Üye Ramazan Duran Dr.Öğr.Üyesi

Üye Alev Mutlu Dr.Öğr.Üyesi

Üye

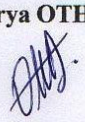
Üye

Anabilim/Anasanat Dalı Başkanı Onayı:
Dr. Öğr. Üyesi. Yusuf Kocaoğlu

YEMİN METNİ

Yüksek lisans tezi olarak sunduğum "Yeni Nesil Derin Bağlamsallaştırılmış Kelime Gösterimleri ve Derin Öğrenme Modelleriyle Finansal Haberler Kullanarak Borsa Tahminlemesi" adlı çalışmanın, tarafımdan, akademik kurallara ve etik değerlere uygun olarak yazıldığını ve yararlandığım eserlerin kaynakçada gösterilenlerden oluştuğunu, bunlara atıf yapılarak yararlanılmış olduğunu belirtir ve bunu onurumla doğrularım.

Derya OTHAN



ÖNSÖZ

Çalışmamın başından sonuna kadar yaşamış olduğum tüm sıkıntılara rağmen her zaman yanımda olan, tüm zorluklara benimle birlikte mücadele veren, değerli bilgilerini benimle paylaşan, söylemiş olduğu her kelime ile hayatıma yön veren, danışmanlığının yanında bana manevi desteğini esirgemeyen ve engin tecrübeleriyle bana yol gösteren kişiliğinden dolayı değerli danışman hocam Dr. Öğr. Üyesi Zeynep Hilal KİLİMCİ'ye sonsuz teşekkürlerimi bir borç bilirim.

Ayrıca bu günlere kadar beni yetiştiren, eğitimimde maddi manevi desteğini esirgemeyen, canım anneme, canım babama, canım kardeşime teşekkürlerimi sunarım.

Ayrıca bana her zaman yardımcı ve destek olan, yüksek lisans eğitimimi bitirmemdeki en büyük destekçim olan canım eşime sonsuz teşekkürlerimi sunarım.

İstanbul, 2019 Derya OTHAN

ÖZET

Amaç

Hisseler, ekonomik krizden etkilenen önemli bir yatırım türüdür. Bu nedenle, hisselerin yönünü tahmin etmek yatırımcılar, analistler ve araştırmacılar için önemlidir. Özellikle de yatırımcılara yapacakları yatırımların yönünü belirlemede önemli bir kaynak olmaktadır. Hisseler üzerinde yatırım yapan ve yaptıkları yatırımlar hakkında yorumlarını paylaşan kullanıcılar, hisseler hakkında analiz yapan analistler ve finansal haberlerin yayınlandığı platformlar tüm kullanıcılara bilgi paylaşımı sağlayan bir platform oluştururlar. Bu çalışmanın amacı, geleneksel derin öğrenme ve kelime gömme modellerinin yanında yeni nesil kelime gömme modellerini kullanarak insanlara BIST100'de en büyük hacime sahip olan hisselerin yönünü tahmin etmeyi ve yatırımcılara yatırımlarının yönünü belirlemede önemli bir kaynak sunmayı teklif ediyoruz. Bildiğimiz kadarıyla, BIST100'de en büyük hacime sahip olan hisseler hakkında tamamen Türkçe metinler üzerinden geleneksel kelime gömme ve derin öğrenme modellerinin yanında yeni nesil kelime gömme modelleri kullanarak analiz etmek için yapılan ilk çalışmadır.

Materyaller ve yöntemler

BIST100'de en büyük hacime sahip olan hisseler ile ilgili, bireysel ve kurumsal kullanıcı yorumları, haber sitelerinde yer alan duyurular ve yatırımcılara değerli bir kaynak olan finansal teknik analizler Türkçe metin kaynağı olarak toplandı. Bireysel ve kurumsal kullanıcı yorumları Twitter sayfalarındaki ("AKBNK", "ALBRK", "GARAN", "HALKB", "ISCTR", "SKBNK", "TSKB", "VAKBN", "YKBNK") anahtar kelimeler ile aranarak hesaplardan toplandı. Sosyal medya platform olan Twitter'daki Türkçe kullanıcı yorumlarını toplamak için Python programlama dilinde yazdığımız Selenium Crawler kullanılarak toplandı. C# dilinde kendi yazdığımız web tarayıcısı ile de, Kamuyu Aydınlatma Platformu (KAP)'ndan finansal haberler ve Mynet Finans web sitesinden kullanıcı yorumları çeşitli Türkçe metin kaynağı olarak toplanmaktadır. Big Para'dan hisselere ait analistler tarafından yapılmış finansal analizler günlük olarak toplanmıştır. Twitter, KAP ve Mynet Finans'taki veriler 01.09.2018 ile 01.09.2019 tarihleri aralığında toplanmıştır. Big Para'da geçmişe yönelik veri çekilemediğinden günlük olarak 28.08.2019 ile 15.11.2019 tarihleri arasında toplanmıştır. Bu çalışmada Word2Vec, GloVe ve FastText, kullanıcı yorumlarını, finansal analiz ve haberleri anlamsal,

bağlamsal ve sözdizimi açısından zenginleştirmek amacıyla geleneksel kelime gömme modelleri olarak kullanılmıştır. Evrişimli Sinir Ağları (CNN'ler), Tekrarlayan Sinir Ağları (RNN'ler) ve Uzun Kısa Süreli Bellek Ağları (LSTM'ler) sınıflandırma görevi için geleneksel derin öğrenme algoritmaları uygulanılmıştır. Bunların yanında yeni nesil kelime gömme modelleri olan Transformatörlerden Çift Yönlü Kodlayıcı Gösterimleri (BERT), Dil Modellerinden Yerleştirme (ELMo) ve Evrensel Dil Modeli İnce Ayar (ULMFiT) kullanılmıştır.

Deneysel Sonuçlar

Bu çalışmada, geleneksel kelime gömme modelleri, derin öğrenme algoritmaları ve yeni nesil kelime gömme modelleri kullanılarak BIST100'de büyük hacime sahip olan borsa hisselerinin yönünü tahmin etmek için kapsamlı deneyler yapılmıştır. Belirtilen tüm doğruluklar, her modelin sınıflandırma performansını ve yaptığımız çalışmanın katkısını göstermek için deneylerde kullanılan bir değerlendirme ölçütüdür. Ön işleme yöntemlerinin uygulanması ile önerilen modelin sınıflandırma performasını iyileştirme amaçlanmıştır. Kullanıcı yorumlarını içeren Türkçe metinleri sınıflandırmada yeni nesil kelime gömme modeli olan ELMo'nun ön işleme yöntemleriyle birleşimi, kullanıcıların hisselerini yönlendirmedeki hassasiyetini belirlemek ve en iyi sınıflandırma başarısı elde etmek için avantajlı bir seçim olacağı sırasıyla Twitter ve Mynet Finans'tan toplanan Türkçe veri setinden elde edilen %97.70 ve %91.55'lik doğruluk değeri ile ortaya koyulmuştur. Ancak haberler ve analizler gibi Türkçe metin içerikli veri setlerinde yeni nesil kelime gömme modellerine göre geleneksel derin öğrenme algortimaları daha iyi sonuçlar üretmiştir.

Sonuçlar

Bu çalışma, borsa hisselerinin yönünü tahmin etmek için çeşitli veri kaynaklarından toplanan metinler üzerinde geleneksel kelime gömme modelleri, derin öğrenme algoritmaları ve yeni nesil kelime gömme modellerini kullanma etkinliğini ve hisselerin yönlerini analiz ederek yatırımcılara yatırım yapacakları süreçte değerli bir katkı sağladığını göstermektedir.

Anahtar Kelimeler: Kelime gömme modelleri; Derin öğrenme; Finansal duyarlılık analizi; BERT; ELMo; ULMFiT.

ABSTRACT

Objective

Stocks are an important investment type affected by the economic crisis. Therefore, it is important for investors, analysts and researchers to predict the direction of the shares. In particular, it is an important source in determining the direction of investments to be made to investors. Users who invest in shares and share their comments on their investments, analysts analyzing shares, and platforms where financial news are published form a platform that provides information sharing to all users. The aim of this study is to propose to the people using the new generation of word embedding models as well as traditional deep learning and word embedding models to predict the direction of the largest volume of shares in BIST100 and to provide investors with an important resource in determining the direction of their investments. To the best of our knowledge, it is the first study to analyze the largest volume of shares in BIST100 using traditional Turkish embedding and deep learning models as well as new generation of word embedding models over completely Turkish texts.

Materials and Methods

Individual and corporate user reviews, announcements on news sites and financial technical analysis, which is a valuable resource for investors, have been collected as the Turkish text source. Individual and corporate user comments were collected from the accounts by searching on the Twitter pages (“AKBNK”, “ALBRK”, “GARAN”, “HALKB”, “ISCTR”, “SKBNK”, “TSKB”, “VAKBN”, “YKBNK”). . It was collected by using Selenium Crawler, which we wrote in Python programming language, in order to collect user comments on the social media platform Twitter. With our own web browser in C #, financial news from the Public Disclosure Platform (KAP) and user comments from the Mynet Finans website are collected as various Turkish text sources. Financial analyzes conducted by analysts belonging to Big Para were collected daily. The data in Twitter, KAP and Mynet Finans were collected between 01.09.2018 and 01.09.2019. Since the historical data of Big Para could not be collected, it was collected daily between 28.08.2019 and 15.11.2019. In this study, Word2Vec, GloVe and FastText are used as traditional word embedding models to enrich user interpretations, financial analysis and news in terms of semantic, contextual and syntax. Conventional neural networks (CNNs), Recurrent Neural Networks (RNNs) and Long Short Term Memory Networks (LSTMs)

have been implemented with traditional deep learning algorithms for the classification task. In addition, the new generation of word embedding models from the Transformers Bidirectional Encoder Display (BERT), Language Models Placement (ELMo) and Universal Language Model Fine Tuning (ULMFiT) were used.

Results

In this study, extensive experiments have been conducted to predict the direction of large volume stock market shares in BIST100 by using traditional word embedding models, deep learning algorithms and next generation word embedding models. All stated accuracy is an evaluation criterion used in experiments to demonstrate the classification performance of each model and the contribution of our work. With the application of pre-treatment methods, it is aimed to improve the classification performance of the proposed model. The combination of ELMo, which is a new generation word embedding model for classifying Turkish texts containing user comments, with preprocessing methods, is an advantageous choice for determining the sensitivity of the users in guiding their shares and achieving the best classification success. 97.70% and 91.55% with the accuracy value was revealed. However, traditional deep learning algorithms produced better results than the new generation word embedding models in Turkish textual data sets such as news and analysis.

Conclusions

This study demonstrates the effectiveness of using traditional word embedding models, deep learning algorithms, and new generation word embedding models on texts collected from various data sources to predict the direction of stock market shares and makes a valuable contribution to investors in the process of investing.

Keywords: Word embedding model; Deep learning; Financial sentiment analysis; BERT; ELMo; ULMFiT.

İÇİNDEKİLER

Sayfa No.

TABLO LİSTESİ

Sayfa No.

ŞEKİL LİSTESİ

Sayfa No.

KISALTMALAR

AKBNK : Akbank Bankası

ALBRK : Albaraka Bankası

GARAN : Garanti Bankası

HALKB : Halk Bankası

ISCTR : İş Bankası

SKBNK : Şeker Bankası

TSKB : T. Sanayi Kalkınma Bankası

VAKBN : Vakıf Bankası

YKBNK : Yapı Kredi Bankası

İMKB : İstanbul Menkul Kıymetler Borsası

BIST100 : Borsa İstanbulu 100 Endeksi

KAP : Kamuyu Aydınlatma Platformu

NLP : Doğal Dil İşleme

LM : Dil Modellemesi

DL : Derin Öğrenme

DWM : Derin Konvolüsyonlu Sinir Ağları

WE : Kelime Gömme

CNN : Evrişimli Sinir Ağları

RNN : Tekrarlayan Sinir Ağları

LSTM : Uzun-Kısa Süreli Bellek Ağları

GloVe : Global Vektörler

BERT : Çift Yönlü Kodlayıcı Gösterimleri

ELMo : Dil Modellerinden Yerleştirme

ULMFiT : Evrensel Dil Modeli İnce Ayar

RH : Hashtagleri kaldırma

RU : URL'leri kaldırma

STM : Stemming

AOT : Bunların hepsi

1. GİRİŞ

Dünyadaki ekonomik krizler borsaları yönlendirmektedir. İnternet ve mobil teknolojinin gelişmesiyle birlikte yatırımcılar, yorumlarını paylaşan birçok sosyal medya platformunda, haber sitelerinde ve finansal yorum sitelerinde yatırım yönünü tahmin edebiliyorlar (Young vd., 2018). Son zamanlarda, hisse senetleri hakkındaki kullanıcı görüşleri, finansal sitelerdeki haberler ve teknik analizler, borsadaki talimatlar hakkında ayrıntılı bilgi edinmek için önemli bir kaynak haline gelmiştir (Çelik & Kaya, 2010). Twitter, bilgileri olduğu gibi paylaştığı ve yaklaşık 350 milyon aktif kullanıcıyla gerçek zamanlı olarak başkalarıyla bağlantı kurduğu bilinen en popüler sosyal ağ hizmetidir (Prieto vd., 2014; Beykikhoshk vd., 2015). Bigpara, tecrübeli finansal gözden geçiricileri (Kilimci & Akyokuş, 2018; Santos vd., 2017; Gunduz vd., 2017; Zhang vd., 2018) sayesinde borsa alımlarının günlük olarak analiz edilmesi için güçlü bir platformdur. Kamuyu Aydınlatma Platformu, kamuoyunu sermaye piyasası ve değişim düzenlemelerine uygun olarak bilgilendiren bir elektronik sistemdir (Mikolov vd., 2013). Mynet Finans, son dakika finansal haberlerin, detaylı piyasa analizlerinin paylaşıldığı ve kullanıcı yorumlarını içeren kapsamlı bir sosyal ağ ve haber platformudur (Devlin vd., 2018). Bu platformlar, yatırımcıların yatırımcı deneyimini anlamalarını, haberleri paylaşmalarını ve yatırımcı hissiyatlarını anlamalarını sağlar. Bu çalışmada, BIST100'deki hisse senetlerinin yönünü analiz etmek için çeşitlendirilmiş türkçe metin kaynakları birleştirilmiştir.Son yıllarda internet ve mobil teknolojinin gelişmesiyle birlikte, sosyal medya platformları hızla büyüdü.

Son yıllarda, derin öğrenme algoritmaları, görüntü / video işleme, doğal dil işleme, örüntü tanıma gibi farklı araştırma alanlarında çok popüler olmuştur ve sıklıkla kullanılmaktadır. Bu modellerin tercih edilmesinin nedeni, geleneksel makine öğrenme algoritmalarına kıyasla daha iyi tahminler üretilmesidir. Derin öğrenme modelleri temel olarak, verilerin derin sinir ağları aracılığıyla anlamlı bir şekilde temsil edilmesini sağlamak için, minimum dış desteğe sahip karmaşık özellikleri eğiterek otomatik özellik çıkarımı sağlamak için kullanılır. Ek olarak, evrişimsel sinir ağları (CNN'ler), tekrarlayan sinir ağları (RNN'ler), uzun süreli bellek ağları (LSTM'ler) (Chatterjee vd., 2019), Word2Vec, Glove, FastText (Bataa & Wu, 2019) gibi geleneksel kelime gömme modelleri, geleneksel öğrenme yöntemleri (NLP) gibi modelleri yeni nesil kelime gömme modelleri olan Transformatörlerden Çift Yönlü Kodlayıcı Gösterimleri (BERT)

(Akcan & Kartal, 2011), Dil Modellerinden Katıştırmalar (ELMo) (Türkmen & Cemgil, 2015), Evrensel Dil Modeli İnce Ayar (ULMFiT) (Chen vd., 2015) gibi modelleri birçok alanda sınıflandırma görevlerinde kullanılmaktadır.

Bu çalışmada, bireysel ve kurumsal kullanıcı yorumlarını, haber sitelerinde yer alan duyuruları ve yatırımcılara değerli bir kaynak olarak finansal teknik analizleri analiz ederek hisse senetlerinin yönünü tahmin etmeyi öneriyoruz. Bu amaçla, Selenium web tarayıcısını kullanarak Twitter'dan bireysel ve kurumsal Türkçe kullanıcı yorumları toplanmıştır. Kendi yazdığımız web tarayıcısı ile de, Kamuyu Aydınlatma Platformu, Mynet Finans ve Big Para web sitelerinden çeşitli Türkçe metin kaynakları toplanmıştır. Metin kaynaklarını aldıktan sonra, kirli verilerin etkisini ortadan kaldırmak için çeşitli işleme teknikleri uygulanmıştır. Kullanıcı yorumları, haber bültenleri ve teknik analizleri Navie Bayes Classifier kullanarak olumsuz ve olumlu olarak etiketlenerek hisselerin yönü anlamak için etiketlenmiştir. Etiketlemede TextBlob kullanılmıştır. TextBlob metinsel verileri işlemek için kullanılan Python dilinde bir kütüphanedir. Etiketleme, isim ifade çıkarma, duygu analizi, sınıflandırma, çeviri ve daha fazlası gibi ortak doğal dil işleme (NLP) görevleri için basit bir API sağlamaktadır.

Daha sonra, hisselerin yönünü anlamlandırabilmek için derin öğrenme modelleri (CNN, RNN, LSTM), geleneksel kelime gömme yöntemleri (Word2Vec, GloVe, FastText) ve yeni nesil kelime gömme modelleri (BERT, ELMo, ULMFiT) kullanılarak temizlenmiş olan veri kümemiz eğitilmiştir. Bildiğimiz kadarıyla, bu hem derin öğrenme modellerini hem de yeni nesil kelime gömme modelini kullanarak borsa yönünü tahmin etmeye yönelik ilk girişimdir. Çalışmamızın katkısını göstermek için deneylerimizde yukarıda belirtilen çeşitli Türkçe metin kaynakları kullanılmıştır. Deney sonuçları, derin öğrenme modellerinin ve yeni nesil kelime gömme modelleri olan BERT, ELMo, ULMFiT'in dahil edilmesinin sistemin sınıflandırma başarısını geliştirdiğini göstermektedir.

Sonuçlara göre çıkarım yapmak gerekirse kullanıcı yorumları kullanılarak Türkçe veri kümelerini sınıflandırmak için yeni nesil kelime gömme modeli olan ELMo'nun ön işleme yöntemleriyle birleşiminin kullanılması, kullanıcıların hisselere yapacakları yatırımlarına yön vermede en avantajlı seçim olacaktır. Bu durum sırasıyla Twitter ve Mynet Finans'tan toplanan Türkçe veri setinden elde edilen %97.70 ve %91.55'lik doğruluk değeri ile ortaya koyulmaktadır. Ancak finansal haberler ve analizler gibi

Türkçe metin içerikli veri setlerinde yeni nesil kelime gömme modellerine göre geleneksel derin öğrenme algortimaları daha iyi sonuçlar üretmiştir.

Makalenin geri kalanı şu şekilde düzenlenmiştir: 2. Bölüm, borsa tahmini konusundaki çalışmaların bir özetini sunar. 3. Bölüm, deneylerde kullanılan malzemeleri ve yöntemleri sunar. Deneysel sonuçlar ve sonuçlar Bölüm 4 ve Bölüm 5'te verilmektedir.

2. LİTERATÜR ÇALIŞMALARI

Bu bölüm, Borsa Tahminlemesinin hakkındaki çalışmaların literatür taramasının kısa bir özetini sunar.

Bir çalışmada yazarlar, konvansiyonel önyükleme veya doğrulama validasyon yöntemi ve karşılıklı olarak bilgiye dayalı bir nitelik seçimi metodu (maksimum alaka düzeyi minimum yedeklilik - MRMR) ile toplanan özellikler kullanılır. PD veri kümesi, 195 denekten oluşan 32 kişinin (24 PD ve 8 sağlıklı) ses kayıtlarından meydana gelmiştir ve %92,9 doğruluk performansı elde edilmiştir (Sakar & Kurşun, 2010).

Bir başka çalışmada, yapay sinir ağı modellerinin İstanbul Menkul Kıymetler Borsası'nda (İMKB) hisse senetleri hakkında bilgi vererek yatırımcılara yön vermek için kullanılabileceği gösterilmiştir (Akcan & Kartal, 2011). Çalışmada genel olarak yapay sinir ağları kullanılarak yatırımcılara büyük bir bilgi kaynağı olabilecek nitelikte bir kaynak sunulmuştur. Başka bir çalışmada (Türkmen & Cengil, 2015), ABD Nasdaq Borsası'nda işlem gören hisse senedi fiyatlarının yönünü tahmin etmek için derin bir öğrenme yapısında kullanılan yığılmış otomatik kodlayıcılar kullanılmıştır. Derin Konvolüsyonlu Sinir Ağları (DVM) yöntemi, F kriterleri ve önerilen model ile en iyi performansı sağladığı kanıtlanmıştır. Chen ve diğer yazarlar (Chen vd., 2015), LSTM modeli ile Çin borsalarındaki geçmiş hisse senedi verilerini kullanarak hisse senedi getirilerini tahmin etmeye yönelik çalışma yapmışlardır. Rastgele tahmin yöntemiyle karşılaştırıldığında, LSTM modeli, hisse senedi getirilerini tahmin etmede daha başarılı olduğu gözlemlenmiştir. Evrimsel Sinir Ağı (ESA) (Gündüz vd., 2017) Gündüz ve arkadaşları tarafından elde edilen sınıflandırma performansı ki-kare özellik seçimi ve lojistik regresyon sınıflandırıcı ile elde edilenden daha yüksek doğruluk değeri üretilmiştir.

Başka bir çalışmada (Hasan vd., 2017), borsa tahmini sorununa derin öğrenme yöntemleri uygulanarak başarılı sonuçlar elde edilmiştir. Yapılan diğer bir çalışmada (Pervan & Keleş, 2019), Türkçe metinlerinde LSTM derin öğrenme yöntemi kullanılarak yapılan tahminlerin doğruluk oranı% 94.21 olarak tespit edilmiştir. Diğer bir çalışmada (Tekin & Çanakoğlu, 2019) ARSA, makine öğrenme algoritmaları ve derin öğrenme tekniği olarak LSTM, BIST30 hisse senedi fiyatlarına Borsa İstanbul'un 30 önde gelen şirketinin verilerini inceleyerek uygulanmaktadır. Hesaplamalı analiz sonucunda, ARIMA'nın LSTM'den daha iyi performans gösterdiği görülmektedir. Ek olarak, lineer

regresyon, diğer makine öğrenme tekniklerine kıyasla daha iyi performans göstermektedir. Başka bir çalışmada (Tekin & Çanakoğlu, 2018), BIST 100'ün önde gelen 25 şirketinin verileri analiz edilmiş ve çeşitli tahmin algoritmaları uygulanmıştır. Sonuç olarak, Random Forest algoritmasının en iyi sınıflandırma sonuçlarını % 57.37 doğrulukla gösterdiği görülmüştür. Diğer bir çalışmada (Gündüz vd., 2017), BIST 100'de sıklıkla işlem gören üç hisse senedinin (GARAN, THYAO ve ISCTR) günlük hareket yönü derin sinir ağları kullanılarak tahmin edilmiştir. Kestirim işlemini gerçekleştirmek için, derin sinir ağının (Konvolüsyonel Sinir Ağı) türü eğitilmiş ve sınıflandırma performansı doğruluk ve F-ölçütleri ile değerlendirilmiştir. Deneylerde GARAN, THYAO ve ISCTR hisse senetlerinin hareket yönleri sırasıyla 0.61, 0.578 ve 0.574 doğruluk oranıyla tahmin edilmiştir.

Bir başka çalışmada yazarlar derin öğrenme yaklaşımlarını kullanarak Türkçe metinlerden anlamsal çıkarımlarda bulunmuştur. Türkçe metinleri toplamak için çeşitli alanlarda ürün satışı yapan bir internet sitesini kullanmışlardır. Bu siteden toplanan metinleri analiz ederek olumlu ve olumsuz anlam içeren kelimelerin vektörel sonuçlarını öğrenerek word2vec modelinde eğitmişlerdir. Eğitilmiş olan very kümeleri kullanılarak Rastgele Orman (Random Forest- RF) modeli geliştirilmiştir. Geleneksel tekniklere ek olarak, derin öğrenme yaklaşımlarından LSTM ve CNN modelleri duygu sınıflandırma için kullanılmıştır. Sonuç olarak CNN modeline göre LSTM modelinin %94,21 daha iyi sonuç ürettiği ortaya koyulmuştur (Prashanth & Roy, 2018).

Çalışmamız, yukarıda belirtilen literatür çalışmalarına göre, hem sosyal medya platformundan hem de finansal web sitelerinden gelen metin verilerini kullanması nedeniyle farklılık göstermektedir. Ayrıca, bir başka yenilik ise, CNN, RNN, LSTM'in hisselerin yönünü tahmin etmek için derin öğrenme modelleri olarak kullanılmasının yanı sıra, yeni nesil bir kelime gömme modeli olarak BERT, ELMo, ULMFiT'in de sınıflandırma amacıyla kullanılmasıdır. Aynı zamanda bu çalışmada derin öğrenme modellerinin performansını iyileştirmek amacıylada Word2Vec, GloVe, FastText gibi geleneksel kelime gömme modelleri ile kombinasyonu da kullanılmıştır. Literatür araştırmasının aksine, yeni nesil modeller kullanarak yatırımcıların hisse senedi fiyatlarını tahmin etmesine rehberlik edecek değerli bir kaynak sunmayı öneriyoruz.

3. YÖNTEMLER, MATERYALLER VE ÖNERİLEN ÇERÇEVE

Yöntemlerin, malzemelerin ve önerilen çerçevenin bir özeti bu bölümde sunulmaktadır.

3.1. Veri Toplanması ve Önerilen Çerçeve

Bu çalışmada, Türkiye Menkul Kıymetler Borsası'nın yönünü tahmin etmek için BIST100'deki banka hisselerinin hareketlerini en yüksek işlem hacmine saip olan hisseleri kullanarak incelendi. Bu amaçla Akbank, Albaraka, Garanti, Halkbank, İş Bankası, T. Sanayi Kalkınma Bankası, Vakıfbank ve Yapıkredi bankalarının hisse senetlerinin yönünü tahmin etmeye odaklanıldı. Farklı metinlerden Türkçe metin verilerini toplandı ve harmanlandı. AKBNK, ALBRK, GARAN, HALKB, ISCTR, SKBNK, TSKB, VAKBN, YKBNK hisseleri bu amaçla verileri çekmek için Crawler'lar tarafından kullanıldı. Mynet Finans'tan (http://finans.mynet.com/) ilgili hisselerin kullanıcı yorumları, Big Para'dan (http://bigpara.com.tr/) analistler tarafından yapılan finansal analizler, Kamuyu Aydınlatma Platformu (https: //www.kap.org.tr/) (KAP) hisse senetlerine ait ek açıklamalar ve Twitter'da ilgili hisse senetlerinin etiketlerini kullanan kullanıcı yorumları, proje kapsamında C # dilinde yazılmış bir tarayıcı ile elde edildi. . Tarayıcı yazıldığında, sitelerin ilgili alanlarına odaklanmaları için anahtar kelimeler ve doğrudan bağlantı adresleri verildi. Bu şekilde, istenen veriler tarama yöntemi kullanılarak tarandı ve dosyalar halinde çıkarılabildi. Yazılan Crawler'lar ile 1 Eylül 2018 - 1 Eylül 2019 tarihleri arasında toplam 18681 veri dosyası alınmıştır. Ancak Big Para sitesinde geçmişe dönük veriler yer almadığından 28 Ağustos 2019 – 15 Kasım 2019 tarihleri arasında toplam 522 veri dosyası çekilmiştir. Selenyum tarayıcısı [18], kullanıcı yorumlarını Twitter'dan çekmek için kullanıldı. Bu, Twitter API'sinin izin verdiği sınır sorunu hakkında endişelenmeden istediğimiz kadar tweet toplamamıza izin vermektedir. "AKBNK", "ALBRK", "GARAN", "HALKB", "ISCTR", "SKBNK", "TSKB", "VAKBN", "YKBNK" etiketleri ile 1 Eylül 2018 ve 1 Eylül 2019 tarihleri arasında toplam 12720 kullanıcı yorumu toplanmıştır.

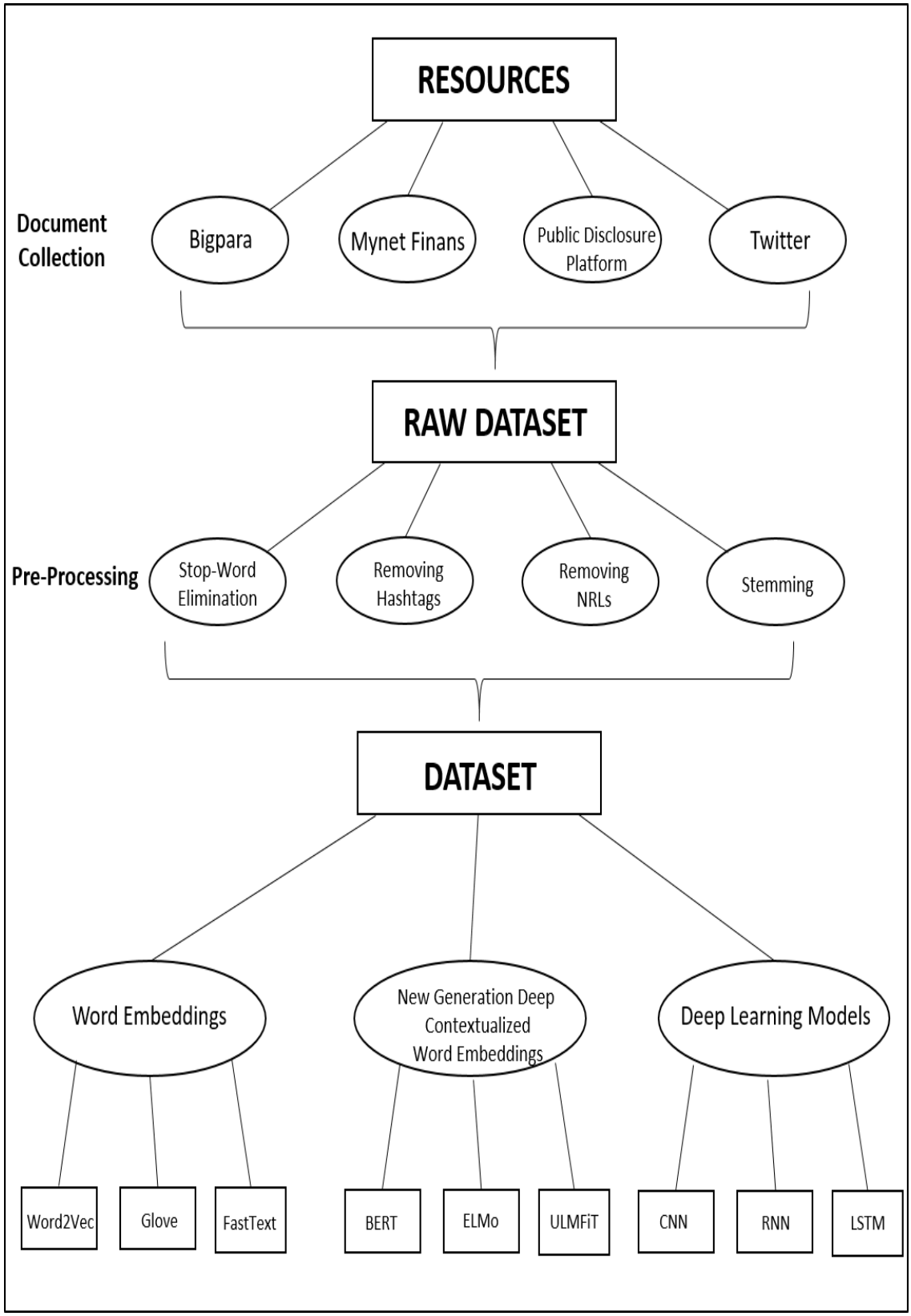
Tablo 3.1 Ön işlemesiz veri kümelerinin istatistikleri.

Veri Kümesi	#Olumlu	#Olumsuz	Toplam
Twitter	12075	645	12720
Mynet Finas	2775	1001	3776
Big Para	110	412	522
KAP	2048	137	2185

Bu çalışmada, denetimli makine öğrenme stratejisine odaklanıyoruz. Bu nedenle, her kullanıcının TextBlob kullanarak kullanıcıların ya da haberlerin tutumunu belirlemek için her bir kullanıcı yorumu, finansal haberler ve finansal analizler olumlu olumsuz olarak etiketleme ihtiyacı vardır (Loria, 2018). TextBlob, duyarlılığı belirlemek için Navie Bayes sınıflandırıcısı kullanarak her veri seti için olumlu veya olumsuz sınıf olasılığını oluşturur. Her bir veri setinden toplanan ham veri kümesi sosyal medya platformunda ve finansal sitelerde oldukça kirli. Bu nedenle, farklı ön işleme tekniklerinin uygulamasına ihtiyaç duyulmuştur. Bu çalışmada, stop-word elimination, removing hashtags, removing URLs, ve stemming teknikleri uygulanmaktadır (Bruns vd., 2009).

Sosyal medya platformları ve finansal siteler karakter kısıtlaması nedeniyle kullanıcılar yorumlarını, analistler yaptıkları finansal analizleri ve haber kaynakları haberlerin tam içeriklerini yeterince ifade edememektedirler. Bu problem gidermek için Word2Vec, GloVe ve FastText gibi gömme modelleri çalışma kapsamında kullanılmştır. Bu şekilde, her yorum kelime gömme modelleri kullanılarak anlam, içerik ve sözdizimi açısından zenginleştirilmiştir. Daha sonra, geleneksel makine öğrenme algoritmaları kullanmak yerine, evrişimli sinir ağları (CNN), tekrarlayan sinir ağları (RNN) ve uzun kısa süreli bellek ağları (LSTM) gibi üç farklı derin öğrenme algoritaları sınıflandırma amacıyla kullanılmştır. Kullanılan bu derin öğrenme algoritmalarının doğruluk performansını artırmak için de Word2Vec, GloVe ve FastText gibi kelime gömme modelleri ile kombinasyonu yapılmıştır. Bunların yanı sıra, yeni nesil kelime gömme modelleri olan Transformatörlerden Çift Yönlü Kodlayıcı Gösterimleri (BERT), Dil Modellerinden Yerleştirme (ELMo) ve Evrensel Dil Modeli İnce Ayar (ULMFiT)

modelleri Türkçe metinleri anlam, söz dizimi ve içerik açısından zenginleştirmek amacıyla kullanılmıştır. Önerilen sistemin akış şeması Şekil 3.1'de verilmiştir. Veri setimiz şekilde de gösterildiği gibi Bigpara, Mynet Finans, KAP ve Twitter'dan toplanan veriler ile oluşturulmuştur. Oluşturulan bu veri kümesi ile Raw Datasetimiz elde edilmiştir. Bu raw dataseti üzerinde pre-processing işlemleri uygulanarak Data setimiz oluşturulmuştur. Artık elimizde etiketlenmiş ve temizlenmiş olan verilerimiz word embeding, new generation deep contextualized word embeddings ve deep learning modellerine gönderilmiştir.



Şekil 3.1 Önerilen sistemin akış şeması

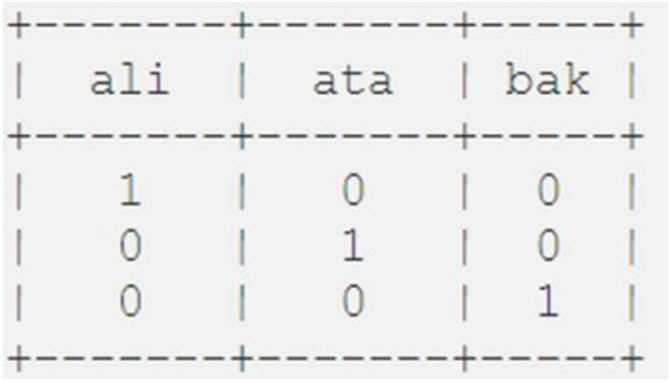
Twitter'dan veri toplamak için, Selenium web tarayıcısı kullanılarak, hisselerin hashtagleri ile arama yapıldı. Twitter'dan toplanan verileri etiketlemek için “Textblob” kütüphanesi kullanıldı. Mynet Finans ve KAP'tan veri toplamak için C # dilinde ayrı ayrı iki tane web tarayıcısı yazıldı. Big Para'dan gelen veriler günlük olarak manuel toplandı. Toplanan tüm verileri olumlu ve olumsuz olarak etiketleyebilmek için, daha önceden etiketlenmiş olan “Hepsiburada” veri kümesi kendi toplamış olduğumuz veri kümseni etiketlemek için kullanıldı. Bu veri kümesinde iki sütun bulunmaktadır. İlk sütunda kullanıcı yorumları, ikinci sütunda ise puanlar yer almaktadır. Ön işleme bu veri setinde gerçekleştirildi. Araştırmada, veri işlemede pandas ve numpy kütüphaneleri kullanılmıştır. Ön işleme için, stop-word elimination, removing hashtags, removing URLs teknikleri regex kütüphanesi kullanılarak gerçekleştirildi. Stemmig işlemini gerçekleştirmek için TurkishStemmer kütüphanesi kullanıldı. Bu kütüphane de Türkçe kelimeler yer almaktadır. Daha sonra Hepsiburada verilerinde 4-5 puan pozitif, 1-2 puan negatif verilerde ayrıldı. TextBlob aracılığı ile Navie Bayes Sınıflandırıcısını kullanarak, kendi sınıflandırılmış veri setimizi oluşturmak için tüm veri kümemiz temizlendi ve etiketlendi. Daha sonra, veri kümemiz bu veri seti üzerinden pozitif ve negatif olarak etiketlendi.

Etiketli veri setimizi modellerde kullanmak için gerekli türkçe sözlükler ve kütüphaneler indirildi. Sıralı, Word2Vec için “trmodel”, GloVe için “glove.6B.load”, FastText için “tr.vec” sözlüklerinin indirilmesi gerçekleştirildi. CNN, RNN, LSTM için kerasın bir kütüphanesi olan “sequential”, BERT için “bert-tensoroflow“, ELMo için “pandas“, ULMFiT için de “fastai“ kütüphaneleri indirilmiştir. BERT, ELMo, ULMFiT modelleri kendi içlerinde sonuçları sürekli çaprazlama işlemi yaprak tekrar tekrar çağrım yaptıkları için sonuç üretmeleri uzun sürmektedir. Bu sebeptan dolayı bu üç model bulutta çalıştırılmıştır. Modellerin bulutta çalıştırılabilmesi için Google‘ın “google.colab“ kütüphanesi kullanılmıştır.

Sonuç olarak, etiketlenmiş veri setimiz % 80 train ve % 20 test olarak ayrıldı. Ayrılan veri kümesi son olarak geleneksel kelime gömme modelleri olan Word2Vec, GloVe, FastText’e, yeni nesil kelime gömme modelleri olan BERT, ELMo, ULMFiT’e ve geleneksel derin öğrenme algoritmaları olan CNN, RNN, LSTM’e gönderilmiştir. Modellerimiz Türkçe metinleri anlam, söz dizimi ve içerik açısından zenginleştirmek amacıyla kullanılmıştır.

Çalışmamızda toplamda üç tane layer oluşturulmuştur. CNN derin öğrenme modelinde, ilk katmanda filtre sayısı 128, kernel boyutu 5, aktivasyon fonksiyonu da "ReLU" olarak belirlenmiştir. Ikinci katmanda pool "GlobalMaxPooling" olarak ayarlanmış ve max pool size 2 olarak belirlenmiştir. Üçüncü katman olan Dense Full Connectited Layer'da aktivasyon fonksiyonu olarak "Sigmoid" ayarlanmıştır. RNN derin öğrenme modelinde, ilk katmanda Bidirectional seçilerek LSTM 64 ayarlanmıştır. İkinci katmada droput katmanı overfittingleri engellemek için eklenmiştir. Üçüncü katmanda da Dense Full Connectited Layer'ı için aktivasyon fonksiyonu "Sigmoid" olarak belirlenmiştir. LSTM modeli için, ilk olarak 256 çıkışlı FC1 isimli bir katman oluşturulmuştur. Ardından ikinci katmanda aktivasyon kodu "ReLU" olarak belirlenmiştir. Üçüncü katmanda da yine droput katmanı overfittingleri engellemek için eklenmiştir. Tüm modeller için dim değeri 100 olarak belirlenmiştir. Yeni nesil kelime gömme modellerinden BERT'in oluşturulabilmesi için max batch boyutu 32, max sequence boyutu 128, öğrenme oranı 2e-5, epoch'lar 3 olarak belirlenmiştir. ELMo için, 256 girişli aktivasyon fonksiyonu "ReLU" olan kernel boyutuda 1 olarak belirlene bir katman oluşturulmuştur. Yine ELMo modeli içinde tek çıkışlı aktivasyon fonsiyonu "Sigmoid" olan bir katman oluşturulmuştur. ULMFiT modeli içinde sınıflandırma metodu olarak fit_one_cycle kullanılmış ve parametleri (1, 1e-2) olarak gönderilmiştir

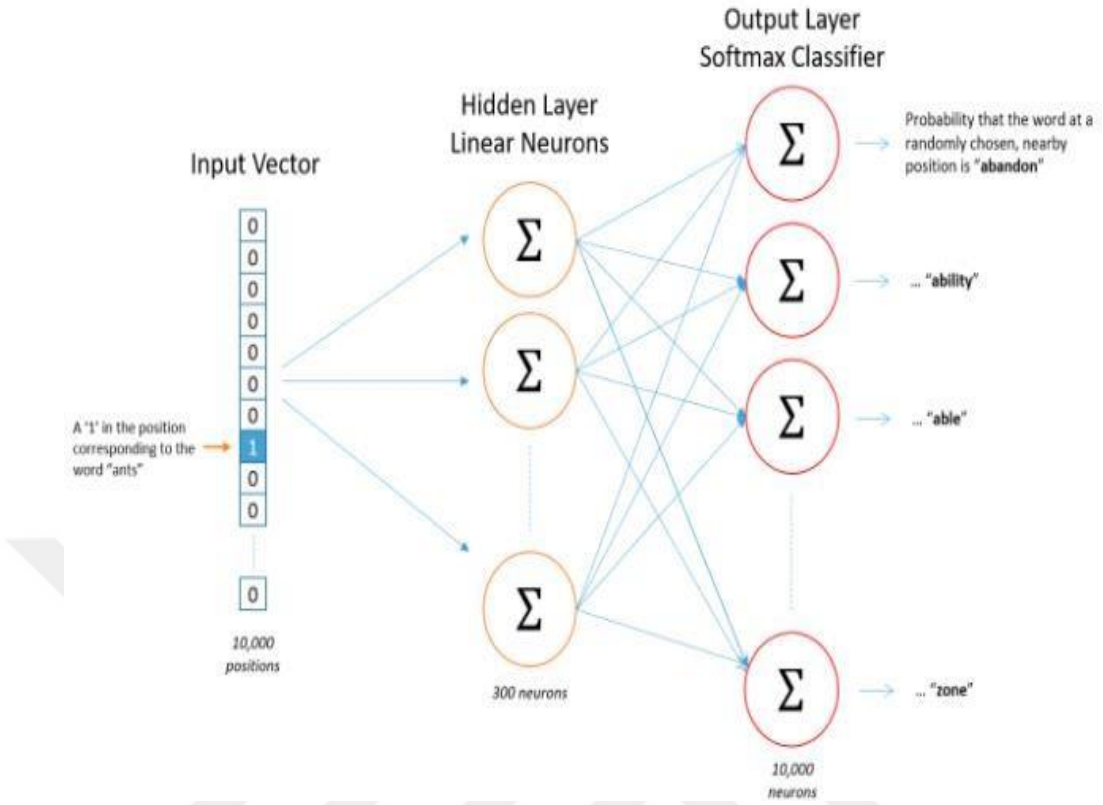
3.2. Kelime Gömme Modelleri



ali	ata	bak
1	0	0
0	1	0
0	0	1

Şekil 3.2 Word2Vec örneği.

Word2Vec, metni işleyen birer adet girdi – çıktı ve gizli katmandan oluşan iki katmanlı bir sinir ağıdır. Word2Vec derin bir sinir ağı olmasa da metni derin ağların anlayabileceği sayısal bir forma dönüştürmektedir. Kelime vektörlerini oluştururken pencere genişliği, embedding boyutu gibi değerler bulunmaktadır. Pencere genişliği hedef kelimenin sağında ve solunda kaç kelime olması gerektiğini belirtirken, embedding boyutu ise her bir kelimenin kaç boyutlu vektör olarak tanımlanacağını belirtir. Bu durum gizli katmandaki nöron sayısına karşılık gelmektedir. Şekil 3.2'de bir Word2Vec örneği verilmiştir.



Şekil 3.3 Örnek Word2Vec sinir ağı.

Şekil 3.3'te örnek olarak Word2Vec sinir ağı verilmiştir. Gizli katmanda 300 nöron olduğu şekilde belirtilmiştir. Gizli katmanda 300 nöronun bulunmasından dolayı her bir kelimenin 300 boyutlu bir vektör olarak gösterildiği aşikardır. Sözlükte 10.000 farklı kelime bulunduğundan için girdi ve çıktı boyutu 10.000 olarak oluşturulmuştur.

GloVe, kelime temsili için önerilen bir başka kelime gömme modeli olarak kaarşımıza çıkmaktadır. İstatistikler kontrol edilmeyen algoritmalara dayanmaktadır (Pennington vd., 2014). Bir belge koleksiyonu kelime oluşumlarının istatistiklerini, kelime temsillerini öğrenmek için tüm denetlenmemiş yöntemlerin kullanabileceği temel bilgi kaynağı olmaktadır. Bu durumda karşımıza çıkan birçok yöntem şu anda mevcut var olmasına rağmen, bu istatistiklerden nasıl bir yorum yapıldığına ve sonuçların nasıl ortaya çıkarıldığına dair sorular halen gelmektedir.

FastText, Word2Vec modelinin bir türevi olan başka bir kelime gömme modelidir. FastText, metin sınıflandırılması için geliştirilmiş bir kütüphane olarakta bilinmektedir. Metin veya kelimeleri herhangi bir dil, örneğin konuşma dili ile ilgili görevde kullanılabilecek sürekli vektörlere dönüştürür. Bu bağlamda, Spam postalarının tespiti en

yaygın örneklerden biri olabilir. Genel olarak, FastText sadece metin sınıflandırılması üzerine tasarlanmıştır. Diğer metin sınıflandırma yapılarına göre daha hızlı ve performaslıdır. Tek tek kelimeleri yapay sinir ağına girdi olarak vermek yerine kelimeleri birkaç harf bazlı "n-gram" halinde parçalamaktadır. N-gram ifadesinde yer alan n tekrar derecesini ifade etmektedir. Kelimenin kaçar kaçar bölüneceği buradaki n ifadesi ile sağlarak, bir kelime veya harften ne kadar olduğunu anlamamızı sağlamaktadır. FastText, son birkaç on yılda doğal dil işleme ve makine öğrenen toplulukların getirdiği en başarılı konseptlerden bazılarını birleştirmektedir. Bunlar, cümleleri kelime torbası ve n-gram torbasıyla temsil etmenin yanı sıra alt kelime bilgisini kullanma ve gizli bir temsil yoluyla sınıflar arasında bilgi paylaşımını içermektedir. Ayrıca hesaplamayı hızlandırmak için sınıfların dengesiz dağılımından faydalanan hierachical softmax yapısı kullanılmaktadır. Bu farklı kavramlar iki farklı görev için kullanıldığı gözlemlenmiştir: verimli metin sınıflandırma ve kelime vektör temsillerini öğrenme.

Derin sinir ağları kısa süre önce metin işlemede çok popüler hale gelmiştir. Bu modeller sınırlı laboratuvar uygulamalarında çok iyi performans gösterirken, çok büyük veri setlerinde kullanımlarını sınırladığından system performansı açısından eğitmek ve test etmek yavaş olabilmektedir.

FastText bu problemin önüne geçmeye yardımcı olmaktadır. Çok fazla sayıda çeşitliliğe sahip veri setlerinde etkili olmak için, farklı kategorilerin bir ağaçta organize olduğu düz bir yapı yerine hiyerarşik bir sınıflandırıcı kullanmaktadır (liste yerine ikili ağaç). Bu, metin sınıflandırma cihazlarının eğitim ve test zaman karmaşıklığını, sınıf sayısına göre doğrusaldan fonkisyona indirmektedir. FastText ayrıca, sınıfları temsil etmek için kullanılan ağacı oluşturmak için Huffman algoritmasını kullanarak sınıfların dengesiz (bazı sınıfların diğerlerinden daha sık görünen) gerçeğinden yararlanır. Bu nedenle, çok sık kategorilere giren ağacın derinliği, nadir olanlara göre daha miniktir ve bu da hesaplama işleminde performans verimliliğini arttırmaktadır.

FastText ayrıca, metinde görünen kelimelere karşılık gelen vektörleri toplayarak elde edilen düşük boyutlu bir vektörle bir metni temsil eder (Joulin vd., 2016). FastText'te, kelimelerin her bir kelimesi ile düşük boyutlu bir vektör ilişkilendirilmektedir. Bu gizli gösterim, farklı kategoriler için tüm sınıflandırıcılar arasında paylaştırılarak, bir kategori için öğrenilen kelimelerin diğer kategoriler tarafından kullanılmasına izin verilir. Kelime torbası adı verilen bu tür temsiller, kelime

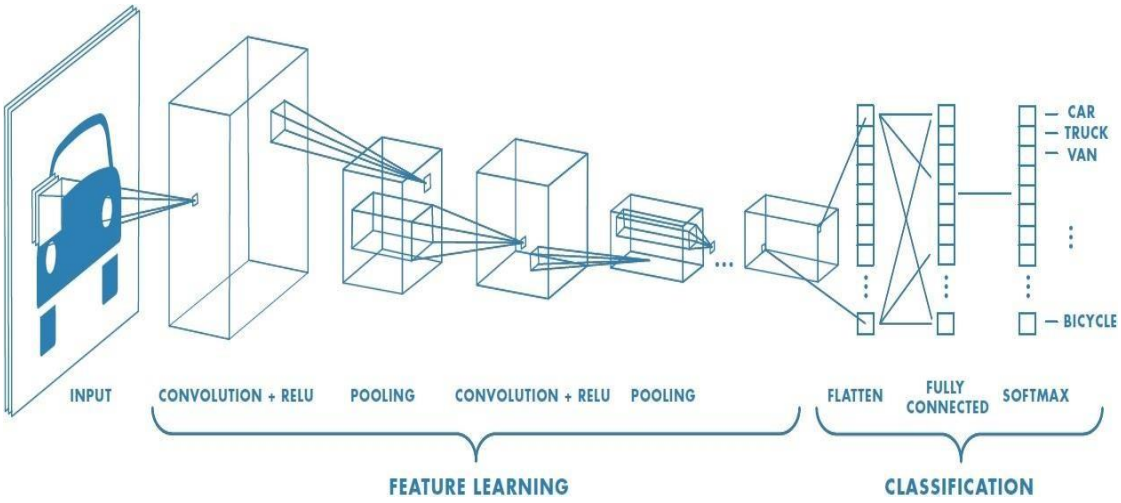
sırasını göz ardı edilmektedir. FastText'te, birçok metin sınıflandırma problemi için önemli olan yerel kelime sırasını hesaba katan kelime n-gramlarını temsil etmek için vektörleri de kullanılmaktadır.

FastText'in diğer derin öğrenme yapılarına göre daha performanslı olduğu Facebook AI Research yaptığı çalışmaları tarafından ortaya koyulmuştur. Facebook AI Research tarafından yapılan deneyde birkaç günlük eğitim süreleri ile eğitilen vektörle üzerinde yapılan araştırmalarda kendini kanıtlamış ve diğer yapılara göre doğruluk performansı açısından daha yetenekli olduğu gözlemlenmiştir.

Genel olarak FastText sadece metin sınıflandırılması üzerine tasarlanmıştır. FastText diğer metin sınıflandırma yapılarına göre daha hızlı ve performanslıdır. Metin sınıflandırmasının yanında, kelimelerin vektör gösterimlerini öğrenmek için fastText de kullanılabilmektedir. Dillerin morfolojik yapısından faydalanarak İngilizce, Almanca, İspanyolca, Fransızca ve Çekçe gibi çeşitli dillerde de kullanılmak için tasarlanmıştır. Çek gibi morfolojik açıdan zengin diller için çok iyi sonuç veren alt kelime bilgisini dahil etmenin basit ama etkili bir yolunu kullanmaktadır, dikkatlice tasarlanmış karakter tasarımı özelliklerinin kelime temsillerini zenginleştirmek için güçlü bir bilgi kaynağı olduğunu da ortaya çıkarmaktadır. FastText, popüler kelime gömme modellerinden veya diğer son teknoloji kelime gösterimlerinden daha üstün bir performans sergileyebilmektedir.

3.3. Derin Öğrenme Algoritmaları

Bu çalışmada, Evrişimli Sinir Ağları (CNN'ler), Tekrarlayan Sinir Ağları (RNN) ve Uzun Kısa Süreli Bellek ağları (LSTM) gibi geleneksel olarak kullanılan üç derin öğrenme algoritmasına odaklanmaktayız.



Şekil 3.4 CNN mimarisi modeli.

Evrişimli sinir ağları (CNN'ler) çok katmanlı bir tür algılayıcıdır (MLP). Şekil 3.4'te CNN mimarisinin modeli verilmiştir. Şekilde görme merkezindeki hücreler görüntünün tümünü kapsayacak şekilde alt bölümlere ayrılmıştır. Basit hücreler kenar benzeri özelliklere odaklanırken, karmaşık hücreler tüm alıcıya daha büyük reseptörlerle konsantre olur. İleriye dönük bir sinir ağı olan CNN, hayvanların görsel merkezinden ilham alıyor. Buradaki matematiksel evrişim süreci, bir nöronun uyarıcı kapsamından uyarıcılara olan tepkisi olarak düşünülebilmektedir. Bütün CNN katmanları tamamen birbirine bağlı olacak şekilde uyarlanmıştır ve her bir evrişim filtresi öğrenilecek maddeleri oluşturmaktadır. Bu katlamalı mimarilar, havuzlama yoluyla hem büyüklükte hem de eğitim süresinde optimizasyon sağlamaktadır.

Bu katmanlar;

- Convolutional Layer — Özellikleri belirlemek amacıyla kullanılmaktadır.
- Non-Linearity Layer — Sisteme lineer olmamanın (non-linearity) belirtilmesi için kullanılmaktadır.
- Pooling (Downsampling) Layer — Ağırlık sayısını düşürmektedir ve uygunluğu kontrol etmektedir.
- Flattening Layer — Geleneksel Sinir Ağı için verileri hazır hale getirmektedir.
- Fully-Connected Layer — Sınıflamada işleme alınan Standart Sinir Ağı olarak bilinmektedir.

Temel olarak, CNN, sınıflandırma probleminin çözüm yolu için standart Sinir Ağı kullanılmaktadır ve bilgileri bulma ve bazı özellik durumlarını belirlemek amacı ile diğer katmanlar kullanılmaktadır.

Katmanların görevlerini ve özelliklerinden bahsedecek olursak;

Convolutional katman CNN'nin ana yapısını oluşturmaktadır. Görevi resmin özelliklerini tespit etmektir. Bu katman, resme bazı fitreleri uygulamaktadır. Bunun sebebi, görüntüdeki alçak ve fazla seviyeli işlevleri ortaya koymaktadır. Örneğin, uyguladığı bu filtre kenarları tespit etmesi için bir filtre olabilmektedir. Bu filtreler genellikle boyutları çok büyük olmaktadır ve piksel değerlerini içermektedirler. (5x5x3) 5 matrisin yükseklik ve genişliğini, 3 matrisin derinliğini temsil etmektedir.

Non-linearity (doğrusal olmayan) katmanı genellikle tüm Convolutional katmanlarından sonra gelmektedir. Görüntüdeki doğrusallık sonunun sebebi, tüm katmanlar doğrusal bir metod olabildiği için Sinir Ağı tek bir algı gibi davranmaktadır. Sonuç, çıktıların lineer kombinasyonu olarak hesaplanabilmektedir. Bu katman aktivasyon katmanı (Activation Layer) olarak adlandırılmaktadır. Çünkü aktivasyon metodlarından biri kullanılmaktadır. Geçmişte, sigmoid ve tahn gibi doğrusal olmayan fonksiyonlar kullanılmıştır, ancak Sinir Ağı eğitiminin hızı konusunda en iyi performansa Rectifier (ReLu) fonksiyonu ulaştığı için artık bu metod kullanılmaya başlanmıştır.

ReLu Fonksiyonu f (x) = max (0, x).

ReLu fonksiyonunun Feature Map'a uygulandığında aşağıdaki gibi bir sonuç üretildiği gözlemlenmektedir.

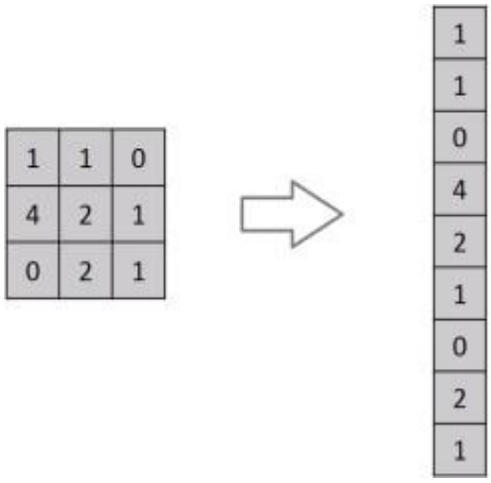
Şekil 3.5 ReLu fonksiyonunun Feature Map'a uygulanması.

Şekil 3.5'te ReLu fonksiyonunun Feature Map'e uygulandığındaki sonuçların görseli sunulmuştur. Feature Map'taki siyah olan değerler olumsuzu temsil etmektedir. Relu fonksiyonunun çalıştırılmasının ardından siyah değerlerin yerine 0 gelmektedir.

Pooling katmanı, CovNet'teki ardışık convolutional katmanları arasına sık sık eklenen bir katman olarak karşımıza çıkmaktadır. Pooling katmanın görevi, gösterimin kayma boyutunu, ağ içindeki parametreleri ve hesaplama sayısını düşürmektir. Bu katman ile birlikte ağdaki uyumsuzluk problemi kontrol altına alınmaktadır. Birçok Pooling işlemleri vardır, bunların içinde en cok kullanılan işlem max pooling'dir.

Ayrıca birçok yazar bu katmanı kullanmaktan kaçınmaktadır. Pooling katmanının yerine Convolutional katmanında daha büyük Stride (Filtreyi kaydırma işlemi) yapılabildiğinden buraya kaymalar yaşanmaktadır.

Flattening katmanının görevi olarak, son katman olan Fully Connected katmanın girişindeki verileri düzenleyip hazır hale getirmektedir. Daha genel olarak bahsedecek olursak, sinir ağları, giriş verilerini tek boyutlu bir diziden getirmektedir. Bu sinir ağındaki veriler ise Convolutional ve Pooling katmanından gelen matrislerin tek boyutlu diziye dönüştürülmüş halini sunmaktadır.



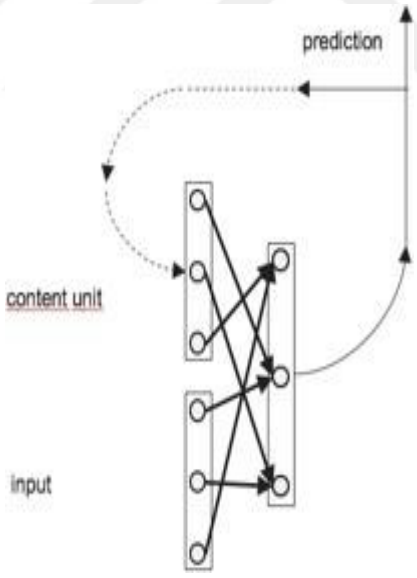
Şekil 3.6 Flattening katmanı mimarisi.

Fully-Connected katmanı, CNN'in son katmanı olarak bilinmektedir. Şekil 3.6'da Flattenning katmanının mimarisi verilmiştir. Bu katman en önemli konumdaki katman olaraka karşımıza çıkmaktadır. Verileri Flattening işleminden matris olarak almaktadır

ve Sinir ağı yoluyla öğrenme işlemini tamamlamaktadır (LeCun vd., 2015; Schmidhuber, 2015).

Bilgisayarla görme konusunda derin öğrenmenin en mühim kısmı AlexNet'tir. Bu yapıda, bırakma, ağ yapısından rastgele bilgileri ortada kaldırır ve modelin fazla takılma probleminin önüne geçmektedir. Bu bir çeşit düzenleme tekniği anlamına gelmektedir. CNN yapısının aşağıda belirtildiği gibi kullanılmasının avantajları bulunmaktadır (Krizhevsky vd., 2012; Voulodimos, vd., 2018):

- Katmanlar daha da derinleşmektedir.
- Hesaplama performansı iyileştirilmiştir (ReLU, bırakma, toplu normalleştirme).
- Ağ katmanları arasındaki bağlantılar fazlalaştıkça, backpropagation (geri yayılım) algoritması geliştirilmiştir. Geoffrey Hinton'un 80'li yıllarda backpropagation algoritmasını popülerliğini artırdığı bilinmektedir. Bu algoritma şu anda derin öğrenme uygulamalarının hemen hemen hepsinde karşımıza çıkmaktadır ve kullanılmaktadır.

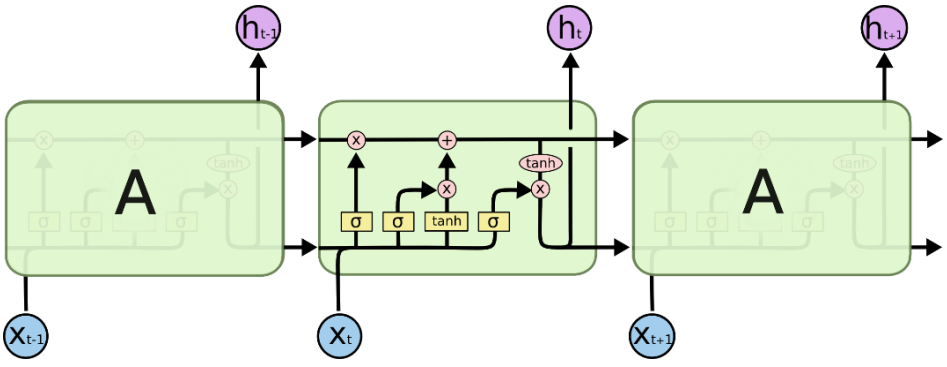


Şekil 3.7 RNN çalışma yapısı.

Şekil 3.7'de RNN çalışma yapısını gösteren görsel verilmiştir. Tekrarlayan sinir ağı (RNN), önceki adımdan gelen çıkışın mevcut aşamaya girdi olarak gelmesiyle bir tür sinir ağı olarak bilinmektedir (Elman, 1990). RNN, kelimeleri hatırlama ihtiyacı nedeni ile sunulmaktadır. Bu problem gizli katman yardımı ile çözülmektedir (Kilimci & Akyokuş

2018; Lipton vd., 2015). RNN'in en mühim özelliği, bir dizi hakkında bazı bilgileri hatırlayan gizli durum olmasındandır. RNN, hesaplananlarla ilgili tüm bilgileri hatırlayan bir "belleğe" sahiptir. RNN, diğer sinir ağlarından farklı olarak, parametrelerin karmaşıklığını azaltmaktadır. Çıktıyı üretmek için tüm girdilerde veya gizli katmanlarda aynı görevi gerçekleştirmektedir. Her giriş için aynı parametreleri kullanmak, parametrelerin karmaşıklığını azaltmaktadır (Tunali & Bilgin, 2018; Schneider, 2014).

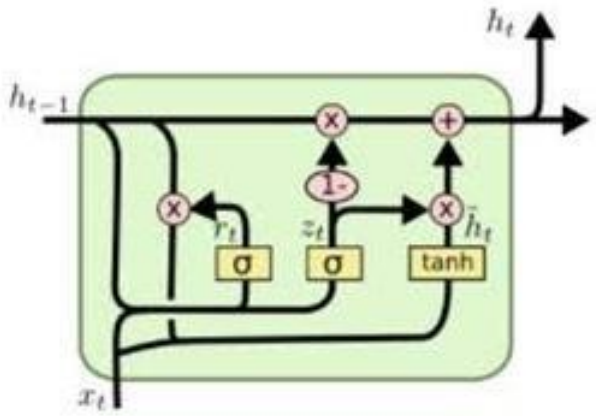
Uzun Kısa Süreli Bellek ağları genellikle "LSTM'ler" olarak adlandırılmaktadır. Uzun vadeli bağımlılıkları öğrenebilen özel bir RNN türevidir. Başlangıç noktası, derin sinir ağlarını eğitirken, geri yayılım algoritmasını kullanarak üssel hata büyümesi problemine bir çözüm sunmaktadır. Bu problemin temelinde yatan sorun, aktivasyon fonksiyonu tarafından üretilen değerlerin sürekli olarak -1, 1 aralığında olmasıdır, böylece bu değerlerin geri yayılma algoritmasına verilmesi ve sıfıra çarpılması ile çarpılmasından doğmaktadır. Bu sorundan kurtulabilmek için ve karmaşık yapılarda daha performanslı öğrenme algoritmaları çıkarabilmek için ortaya çıkarılan LSTM, uzun vadeli bağımlılıkların ve uzun vadeli bilgilerin hatırlanması gereken sorularda daha performanslı sonuçlar üretmektedir. RNN hücresine ayrıca bellek eşlik etmektedir. Her süreçte, öğrenilen hücrelerin hangilerinin ortadan kaldırılması gerektiğine ve hangilerinin yeniden oluşturulacağına karar verilmektedir. Nöral makine çevirisi için Google tarafından başarıyla kullanılan bir yapıdır. Bu çalışmada, daha önce kullanılan kelimelerin anlamlarını öğrenmek ve bu anlamlara dayalı tahminler üretebilmek için kullanılmıştır (Kilimci & Akyokuş, 2018; Zhang vd., 2018). Şekil 3.8'de LSTM ağı mimarisi verilmiştir.



Şekil 3.8 Uzun-kısa süreli hafıza ağı mimarisi.

Tüm tekrarlayan sinir ağları, bir sinir ağının yinelenen parçalarını zincirinin formuna sahip olduğu kanıtlanmıştır. Yalnızca LSTM birimi giriş olarak x_t, $h_{(t-1)}$ ve

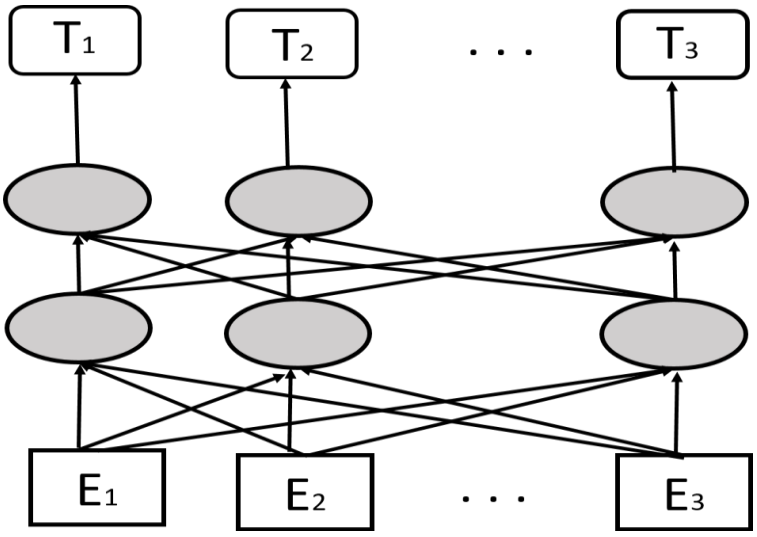
$c_{(t-1)}$ alır ve aşağıdaki bileşik yapıyı kullanarak h_t, c_t üretir. Şekil 3.9'da LSTM'in hafıza ağı katmanının ayrıntılı yapısı verilmiştir.



Şekil 3.9 Uzun- kısa süreli hafıza ağı katmanı mimarisi.

3.4. Yeni Nesil Kelime Gömme Modelleri

BERT, çift yönlü kodlayıcı gösterimleri anlamına gelen yeni nesil bir kelime gömme modelidir. BERT modelinin gömülü modellerinden farklı olarak, veri setini her iki katmanda da iki yönde önceden eğitmek ve sözcüğü hem sağ hem de sol bağlamlarda koşullandırmak için tasarlanmıştır (Devlin vd., 2018). BERT modeli, soruları yanıtlama ve dil çıkarma işleminde önemli bir görev olmadan son model modeller oluşturmak için ek bir çıkış katmanıyla ince ayar yapmak için kullanılabilir. Kavramsal olarak basit ve ampirik olarak güçlüdür (Devlinvd., 2018). Şekil 3'te, BERT modelinin mimarisi, okların bir katmandan diğerine bilgi akışını gösterdiği yerde sunulmaktadır. Üstteki □1, □2, □3 kutuları, her giriş sözcüğünün nihai bağlamsal sunumunu gösterir. Giriş kelimeleri, *1* ile *n* arasında olan *E* ile gösterilir.



Şekil 3.10 BERT modeli mimarisi.

2019 yılında gelen BERT güncellemesi uygulanan arama sorgusunun kelimelerini ayrı ayrı işlemektense bütün kelimeleri mantık olarak incelemeye alarak en uygun ve doğru çıktıları sırasıyla verir.

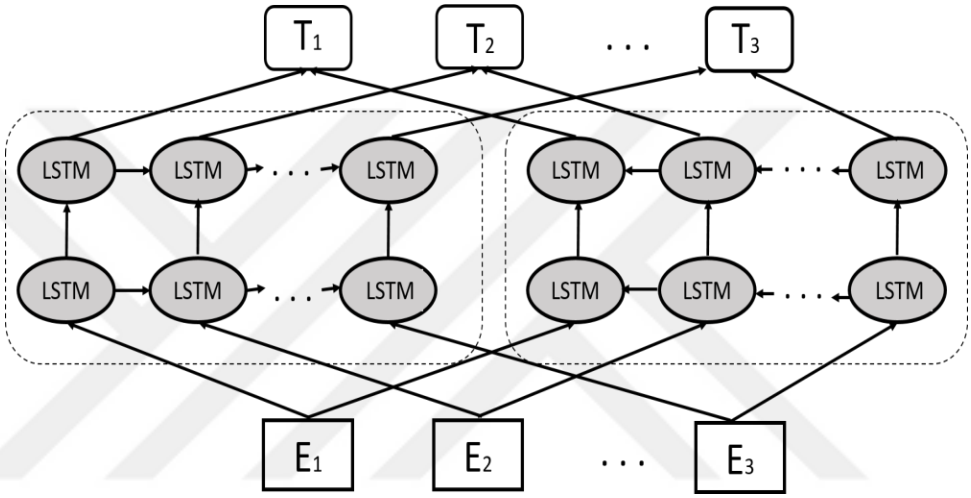
BERT modelinin temelini, önceden eğitilmiş, açık kaynak kodlu bir NPL modeli oluşturur. Bu modeli diğer modellerden ayıran özellik, arama sorgusundaki eksik olan kelimeyi bulabilmek için tüm cümleyi işlemesidir.

Makine öğrenme algoritmalarını da içerecek olan BERT, iki yönlü bir NLP özelliğini içerir. Her kelimenin diğer bir kelimeyle ilişkisini öğrenmeye odaklanır. Diğer modeller sağdan sola, soldan sağa giden yüzeysel çift yönlü bir dil işlemesi kullanırken, BERT daha karmaşık maskeli dil modeli kullanır.

ELMo (Dil Modellerinden Yerleşimler): 2018'de AllenNLP tarafından geliştirilen (Phang vd., 2018), geleneksel yerleştirme tekniklerinin ötesine geçer. Kelime gösterimi oluşturmak için derin, iki yönlü bir LSTM modeli kullanır. Dil Modellerinden Yerleştirme (ELMo), kelimeleri kullanıldığı koşulda değerlendirir. Ayrıca, modelin sözlük olmayan kelimelerin gönderimini oluşturmasını sağlayan karakter tabanlı bir modeldir. Bu, Elmo'nun kullanım şeklinin word2Vec veya fastText'ten oldukça farklı olduğu anlamına gelir. ELMo, sözcükler ve bunların vektörleri sözlüğüne bakmak yerine, metni derin bir öğrenme modelinden geçirerek anında vektörler oluşturur.

ELMo ayrıca, hem kelime kullanımının karmaşık özelliklerini hem de dilsel bağlamlarda bu kullanımların nasıl değiştiğini modelleyen tamamen bağlamsallaştırılmış

bir Kelime Sunumu olarak da adlandırılır. Bu kelime vektörleri, geniş bir metin koleksiyonunda önceden eğitilmiş, derin bir çift yönlü dil modelinin (biLM) iç durumlarının işlevleridir. Mevcut modellere kolayca eklenebilirler ve soru-cevap üretme, metin üretme ve duygu analizi de dahil olmak üzere çok çeşitli zorlu NLP problemlerinde literatür araştırmasının sonuçlarını önemli ölçüde geliştirebilirler. Şekil 3.11'de, ELMo modelinin mimarisi, okların bir katmandan diğerine bilgi akışını gösterdiği yerde sunulmaktadır. Üstteki □1, □2, □3 kutuları, her giriş sözcüğünün nihai bağlamsal sunumunu gösterir. Giriş kelimeleri, *1* ile *n* arasında olan *E* ile gösterilir.



Şekil 3.11 ELMo modeli mimarisi.

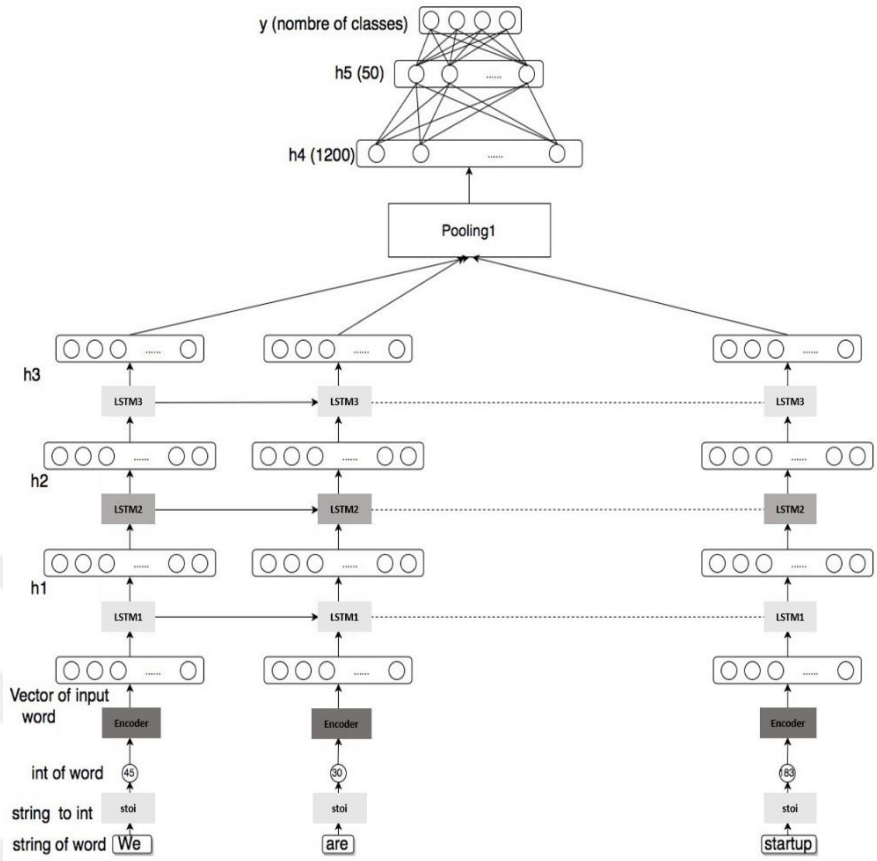
Word2vec ve GLoVe gibi geleneksel kelime yerleştirmelerinden farklı olarak, bir belirteç veya sözcüğe atanan ELMo vektörü aslında o kelimeyi içeren cümlenin tümünün bir işlevidir. Bu nedenle, aynı kelime farklı bağlamlarda farklı kelime vektörlerine sahip olabilir.

Çalışma mantığını bir örnekle ifade edecek olursak, "Dün kitabı okudum." ve "Mektubu şimdi okuyabilir misin?" cümlelerini ele alalım. Bu iki cümledeki fiiller aynı olmasına rağmen zamanları faklıdır. İlk cümlede "okumak" fiili geçmiş zamandır. Aynı fiil ikinci cümlede şimdiki zamana dönüşür. Bu, bir sözcüğün birden fazla anlama veya duyuya sahip olabileceği bir Polysemy örneğidir.

Geleneksel kelime yerleştirme modelleri, her iki cümlede de “okumak” kelimesi için aynı vektörle ortaya çıkar. Dolayısıyla, sistem çok terimli kelimeler arasında ayrım yapamaz. Bu kelime yerleştirmeleri, kelimenin kullanıldığı bağlamı kavrayamaz.

ELMo kelime vektörleri bu konuyu başarıyla ele almaktadır. ELMo kelime temsilleri giriş cümlesi kelimesini embeddings kelimesini hesaplamak için denklem içine alır. Bu nedenle, “okumak” terimi farklı bağlamda farklı ELMo vektörlerine sahip olacaktır.

Evrensel Dil Modeli İnce Ayarı (ULMFiT) aslında NLP'ye özgü farklı görevlerde de kullanılabilen bir transfer öğrenme modelidir. ULMFiT, morfolojik açıdan fakir İngilizce için iyi çalışan, ancak Türkçe gibi diller için çok büyük ve seyrek sözlüklerle sonuçlanan kelime temelli tokenizasyon kullanmaktadır. Bir alana özgü veri kümesi problemini fazla uydurmamak için çözme ihtiyacı vardır, Jeremy Howard ve Sebastian Ruder (Howard & Ruder, 2018) tarafından üç yeni yöntem önerilmiştir. ULMFiT, dil modeli ön eğitim, dil modeli ince ayar ve sınıflandırıcı ince ayar aşamalarını harmanlayarak yapılır. Dil modeli ön eğitiminde, dil modeli, bu dillerin özelliklerini farklı katmanlarda elde etmek için ortak alan koleksiyonunda eğitilmiştir. İkinci aşamada, model, ayırt edici ince ayar ve eğimli üçgen öğrenme oranları kullanarak dağılımlarını öğrenmek için veri setinde ince ayar yapılmıştır. Son katmanda, dil modelleri, büyük girdiler için gradyan yayılımını kolaylaştırmak için zaman içinde geriye yayılma ile eğitilmiştir. Bu şekilde, ULMFiT belge boyutu, numarası ve etiket türünde değişen görevleri yönetir. Aynı zamanda sadece bir mimarlık ve eğitim prosedürü kullanır. Derin öğrenme modellerinden farklı olarak özelleştirilmiş ön işleme ve özellik mühendisliği gerektirmez. Ekstra alan etiketi veya belge gerektirmez. Şekil 3.12'de, okların bir katmandan diğerine bilgi akışını gösterdiği ULMFiT modelinin mimarisi sunulmaktadır.



Şekil 3.12 ULMFiT modeli mimarisi.

ULMFiT, her NLP görevi için transfer öğrenmeyi etkinleştirme ve mükemmel sonuçlar elde etme yöntemidir. Bütün bunlar sıfırdan modeller eğitmek zorunda kalmadan.

ULMFiT, aşağıdaki gibi yeni teknikleri kullanarak son teknoloji ürünü bir sonuç elde eder:

- Ayrımcı ince ayar
- Lant Eğimli üçgen öğrenme oranları
- Kademeli çözülme

Bu yöntem, Wikitext 103 veri setinde eğitilmiş önceden eğitilmiş bir dil modelinin (LM) daha önce öğrendiklerini unutmayacak şekilde yeni bir veri setine ince ayar yapılmasını içerir. Şekil 3.13'te ULMFiT modelinin kullandığı tekniklerin akışı verilmiştir.

Şekil 3.13 ULMFiT modeli teknikleri.

Dil modellemesi (LM), bir dilin genel özelliklerini yakalar ve diğer akış aşağı NLP görevlerine verilebilecek çok miktarda veri sağlar. Dil modellemesinin ULMFiT için kaynak görev olarak seçilmesinin nedeni budur.

4. DENEYSEL SONUÇLAR

Bu çalışmada, BIST100'deki büyük hacime sahip borsa hisseleri ile ilgili belirlenen sitelerden kullanıcı yorumları, finansal analizler ve finansal haberler kullanılarak borsa tahminlemesi yapmaya yönelik geleneksel kelime gömme modelleri, derin öğrenme algoritmaları ve yeni nesil kelime gömme modelleri kullanarak analiz etmek için kapsamlı deneyler yapılmıştır. Her modelin sınıflandırma performansını ve çalışmalarımızın katkısını göstermek için deneylerde bir değerlendirme ölçütü olarak doğruluk kullanılmaktadır. Bekletme veri setinde 10 kez uygulanır. Bu yaklaşım önceki eğitim çalışmalarına benzerdir, verilerin eğitim için %80'ini ve test için %20'sini kullanır. Ön işleme yöntemleri, geleneksel kelime gömme modelleri, derin öğrenme algoritmaları ve yeni nesil kelime gömme modelleri için şu kısaltmalar kullanılmıştır: SWE: Kelime durdurma eliminasyonu, RH: Hashtagleri kaldırma, RU: URL'leri kaldırma, STM: Stemming, AOT: Bunların hepsi, CNN: Evrişimli sinir ağ, RNN: Tekrarlayan sinir ağı, LSTM: Uzun kısa süreli hafıza ağı. Elde edilen en iyi doğruluk sonuçları kalın harflerle belirtilmiştir (Kilimci & Akyokus, 2018; Mikolov, 2013).

İlk önce, geleneksel kelime gömme modellerinin, derin öğrenme algoritmalarının ve yeni nesil kelime gömme modellerinin sınıflandırma performansları veri seti bazında Tablo 4.1, Tablo 4.2, Tablo 4.3 ve Tablo 4.4'te görüldüğü şekilde hem veri setlerinin birbirleriyle hem de modellerin birbirleriyle kıyaslanması için eğitim seti yüzdeleri açısından analiz ediyoruz. Tablo 4.5, Tablo 4.6, Tablo 4.7 ve Tablo 4.8'de de derin öğrenme algoritmalarının geleneksel kelime gömme modellerinin kombinasyonu sonucunda elde edilen doğruluk sonuçları verilmiştir. Verilen son 4 tablo sayesinde derin öğrenme modellerinin geleneksel kelime gömme modelleri ile kombinasyonu sonucunda performansı iyileştirip iyileştiremeyeceğimizi kanıtlayacağız.

Tablo 4.1 Geleneksel kelime gömme modellerinin, derin öğrenme algoritmalarının ve yeni nesil kelime gömme modellerinin Twitter veri setindeki doğruluk sonuçları.

Twitter	**MODELS**								
Hisseler	**Word2Vec**	**FastText**	**Glove**	**CNN**	**RNN**	**LSTM**	**BERT**	**ELMo**	**ULMFiT**
AKBNK	94.56	95.13	92.07	96.22	96.56	96.32	94.84	98.24	96.21
ALBRK	94.30	94.44	93.26	95.48	94.92	94.82	96.04	96.72	93.22
GARAN	90.17	91.62	91.10	92.07	93.36	92.44	92.98	96.98	92.43
HALKB	94.22	93.75	92.84	94.69	94.09	94.29	93.70	98.48	94.29
ISCTR	93.09	94.16	92.75	95.35	95.52	95.15	96.72	96.85	95.14
SKBNK	94.47	95.01	93.11	96.05	96.94	96.25	97.17	98.92	95.48
TSKB	95.90	95.81	94.58	97.79	98.79	96.79	98.71	98.78	98.71
VAKBN	94.16	95.90	92.12	96.37	97.26	97.06	98.63	96.14	97.26
YKBNK	95.30	96.48	94.80	96.17	97.16	96.27	97.60	98.17	95.69
Avg.	94.02	94.70	92.96	95.58	96.07	95.49	96.27	**97.70**	95.38

Tablo 4.1‘de her bir modelden elde edilen ortalama sonuçlara göre , Twitter veri seti sonuçları incelendiğinde geleneksel kelime gömme modellerinden elde edilen oranlarda Fasttext kelime gömme modeli %94.70 en iyi sonucu üretmiştir. Sıralayacak olursak FastText>Word2Vec>Glove şeklinde olacaktır. Derin öğrenme modellerinden elde edilen oranlara baktığımızda hisselere göre en iyi ortalama sonuç %96.07 oranıyla RNN modelinden gelmiştir. Bu durumda geleneksel kelime gömme modelleri ile kıyasladığımızda RNN %96.07 doğruluk değeriyle FastText‘ten daha iyi sonucu ürettiği elde edilmiştir. Bu durumda sıralama RNN > CNN > LSTM > FastText > Word2Vec > Glove olarak şekillenmiştir. Daha sonra etiketlenmiş veri setlerinin yeni nesil kelime gömme modellerine gönderilmesiyle elde edilen sonuçlara bakıldığında ELMo modelinin tüm sonuçlara kıyasla %97.70 doğruluk değeriyle en iyi sonucu ürettiği ortaya koyulmuştur. Son durumda yeni sıralama ELMo > BERT >RNN > CNN > LSTM > ULMFiT > FastText > Word2Vec > Glove şekilde değişmiştir. Yeni nesil kelime gömme modeli olan ELMo‘nun sonuçlarını hisse bazında kontrol ettiğimizde SKBNK hissesinde

%98.92 doğruluk değeriyle en iyi sonuç, VAKBN hissesinde de %96.14 doğruluk değeriyle en düşük sonucun elde edildiği gözlemlenmiştir. Yani Twitter'dan toplanan kullanıcı yorumları veri setini yorumlamak için yeni nesil kelime gömme modeli olan ELMo'yu kullanmak sınıflandırma performansı açısında avantajlı olacaktır.

Tablo 4.2 Geleneksel kelime gömme modellerinin, derin öğrenme algoritmalarının ve yeni nesil kelime gömme modellerinin Mynet Finans veri setindeki doğruluk sonuçları.

Mynet Finans	**MODELS**								
Hisseler	**Word2Vec**	**FastText**	**Glove**	**CNN**	**RNN**	**LSTM**	**BERT**	**ELMo**	**ULMFiT**
AKBNK	77.40	79.16	75.23	80.00	76.67	82.00	80.00	93.33	73.33
ALBRK	79.55	80.72	78.94	86.67	98.92	84.95	83.33	91.42	91.77
GARAN	80.27	82.66	77.18	83.61	90.16	83.61	83.33	95.34	90.00
HALKB	68.05	70.90	65.82	82.11	84.03	70.29	82.14	89.97	71.75
ISCTR	66.14	69.71	62.37	77.27	75.76	78.79	71.87	91.95	81.25
SKBNK	71.95	70.88	70.50	86.27	90.20	76.47	72.00	96.10	76.00
TSKB	74.43	76.50	70.15	85.29	89.71	80.88	86.66	89.65	79.41
VAKBN	78.80	80.35	75.49	91.11	95.56	84.44	81.81	93.93	84.09
YKBNK	75.89	77.23	71.61	83.33	84.31	62.75	80.70	82.24	80.33
Avg.	74.72	76.46	71.92	83.96	87.26	78.24	80.20	**91.55**	80.88

Tablo 4.2'de her bir modelden elde edilen ortalama sonuçlara göre , Mynet Finans veri seti sonuçları incelendiğinde geleneksel kelime gömme modelleri, yeni nesil kelime gömme modelleri ve geleneksel derin öğrenme modellerinden elde edilen oranlardan daha düşük doğruluk değerlerine sahiptir. Bu sebeple tercih edilme sırası en alt sıralardadır. Ancak Eski nesil kelime gömme modellerinin doğruluk değerlerine göre sıralayacak olursak FastText > Word2Vec > Glove şeklinde olacaktır. Derin öğrenme modellerinden elde edilen oranlara baktığımızda hisselere göre en iyi ortalama sonuç %87.26 oranıyla RNN modelinden gelmiştir. Bu derin öğrenme modellerini yeni nesil kelime gömme modellerinden çıkan doğruluk değerleriyle kıyaslacak olursak Mynet Finans veri setini tahminlemede yeni nesil kelime gömme modellerinden ELMo'yu tercih

etmek avantajlı olacaktır. ELMo modelinin kendi alanındaki sonuçlara kıyasla %91.55 doğruluk değeriyle en iyi sonucu ürettiği ortaya koyulmuştur. Son durumda modellerin doğruluk değerlerinin sıralaması ELMo > RNN > CNN > ULMFiT > BERT > LSTM > FastText > Word2Vec > Glove şeklinde sonuçlanmıştır. Yani Mynet Finans'tan kullanıcı yorumları veri setini tahminlemek için Twitter'da da kullandığımız gibi yeni nesil kelime gömme modeli olan ELMo'yu kullanmak sınıflandırma performansı açısında avantajlı olacaktır.

Tablo 4.3 Geleneksel kelime gömme modellerinin, derin öğrenme algoritmalarının ve yeni nesil kelime gömme modellerinin Big Para veri setindeki doğruluk sonuçları.

Big Para	**MODELS**								
Hisseler	**Word2Vec**	**FastText**	**Glove**	**CNN**	**RNN**	**LSTM**	**BERT**	**ELMo**	**ULMFiT**
AKBNK	79.24	78.15	81.36	80.21	97.44	85.00	92.98	97.64	90.00
ALBRK	88.35	86.21	90.07	90.14	97.50	91.00	95.67	86.72	80.31
GARAN	88.01	87.10	89.58	80.11	96.48	92.44	90.03	86.98	89.90
HALKB	77.97	77.45	78.35	82.00	97.44	80.01	80.12	78.48	70.88
ISCTR	84.61	87.73	90.29	90.43	92.00	92.31	96.72	86.85	90.00
SKBNK	75.01	77.08	79.61	80.51	97.44	81.00	97.17	88.92	70.86
TSKB	77.84	76.54	79.27	88.19	96.47	80.56	98.71	88.78	78.79
VAKBN	75.96	76.12	77.26	80.58	97.26	81.50	90.82	86.14	75.93
YKBNK	80.02	80.59	81.34	82.60	97.44	90.00	97.70	88.17	74.11
Avg.	80.78	80.77	83.01	83.86	**96.61**	85.98	93.32	87.63	80.09

Tablo 4.3'te her bir modelden elde edilen ortalama sonuçlara göre , BigPara veri seti sonuçları incelendiğinde geleneksel kelime gömme modellerinden elde edilen oranlarda Glove kelime gömme modeli %83.01 sonucu ile en iyi doğruluk değerini üretmiştir. Bu ortalama sonucun üretilmesinde ISCTR hissesinden gelen %90.29 doğruluk değeri etkin rol almıştır. Glove modelinde en düşük doğruluk değeri %77.26 ile VAKBN hissesinde elde edilmiştir. Derin öğrenme modellerinden elde edilen oranlara baktığımızda hisselere göre en iyi ortalama sonuç %96.61 oranıyla RNN modelinden

gelmiştir. Bu durumda geleneksel kelime gömme modelleri ile kıyasladığımızda RNN %96.61 doğruluk değeriyle Glove'dan daha iyi sonucu ürettiği elde edilmiştir. Bu durumda sıralama RNN > LSTM > CNN > Glove > Word2Vec > FastText olarak şekillenmiştir. Daha sonra etiketlenmiş veri setlerinin yeni nesil kelime gömme modellerine gönderilmesiyle elde edilen sonuçlara bakıldığında ELMo modelinin kendi alanındaki sonuçlara kıyasla %93.32 doğruluk değeriyle en iyi sonucu ürettiği ortaya koyulmuştur. Ancak derin öğrenme modeli olan RNN ile üretilen sonucun üzerine çıkamamıştır. Son durumda yeni sıralama RNN > BERT >ELMo > LSTM > CNN > Glove > Word2Vec > FastText > ULMFiT şeklinde değişmiştir. Yani BigPara'dan toplanan finansal analiz veri setini yorumlamak için Twitter'da kullandığımız yeni nesil kelime gömme modeli olan ELMo'yu kullanmak yerine geleneksel derin öğrenme modeli olan CNN'i kullanmak sınıflandırma performansı açısında avantajlı olacaktır.

Tablo 4.4 Geleneksel kelime gömme modellerinin, derin öğrenme algoritmalarının ve yeni nesil kelime gömme modellerinin KAP veri setindeki doğruluk sonuçları.

KAP	**MODELS**								
Hisseler	**Word2Vec**	**FastText**	**Glove**	**CNN**	**RNN**	**LSTM**	**BERT**	**ELMo**	**ULMFiT**
AKBNK	75.32	74.80	78.36	93.23	84.52	99.18	71.45	98.86	96.72
ALBRK	72.66	73.15	75.84	92.31	84.62	84.62	69.23	98.45	82.00
GARAN	78.35	77.42	79.27	99.74	99.63	99.64	78.26	98.87	89.05
HALKB	88.04	87.91	88.63	99.36	98.36	98.26	90.00	98.30	90.00
ISCTR	81.50	80.26	83.75	99.49	99.45	99.39	94.91	90.90	81.58
SKBNK	75.48	74.90	76.32	99.64	88.24	89.24	74.07	80.00	79.63
TSKB	82.96	82.45	84.58	99.49	98.36	98.29	87.72	80.90	78.89
VAKBN	76.00	75.70	77.23	99.36	97.25	98.65	78.37	97.50	75.89
YKBNK	78.41	78.59	80.94	98.78	99.54	93.64	88.27	90.90	81.81
Avg.	78.75	78.35	80.55	**97.93**	94.44	95.66	81.36	92.74	83.95

Tablo 4.4‘te son veri setimiz olan KAP‘tan toplanmış olan finansal haberlerin sınıflandırma performansını gösteren sonuçlara yer verilmektedir. Her bir modelden elde edilen ortalama sonuçlara göre , KAP veri seti sonuçları incelendiğinde geleneksel kelime gömme modellerinden, Glove en iyi performansı %80.55 doğruluk değeriyle göstermiştir. En düşük performansta %78.35 doğruluk değeri ile FastText‘ten elde edilmiştir. Yeni nesil kelime gömme modellerinden elde edilen sonuçlar incelendiğinde de CNN modeli %97.93 doğruluk değeri ile hem kendi alanındaki modeller arasında hemde çalışmada kullanılan modeller arasında en iyi sonucu üretmiştir. Bu durumdan yola çıkarak tablonun yorumlamasını yapacak olursak KAP veri setinde toplanan text ile tahminleme yapmak için yeni nesil kelime gömme modelleri ve geleneksel kelime gömme modelleri yerine geleneksel derin öğrenme modellerinin kullanılması sınıflandırma performansı açısından avantajlı olacaktır.

Çalışmamızda ek olarak geleneksel derin öğrenme modelleri olan CNN, RNN ve LSTM‘in performasnlarını iyileştirmeye yönelik geleneksel kelime gömme modellerinin kombinasyonu da yapılmıştır. Word2Vec, GloVe ve FastText kelime gömme modellerinden çıkan vektörler CNN, RNN ve LSTM modellerine giriş olarak verilip bir doğruluk değeri üretilmiştir. Bu doğruluk değerlerinin sonuçları aşağıda sırasıyla Tablo 4.5, Tablo 4.6, Tablo 4.7 ve Tablo 4.8‘de verilmiştir. Her bir veri setinin sonuçları kombinasyonsuz sonuçlarla kıyaslanarak tabloların altında yorumlar halinde belirtilmiştir.

Tablo 4.5 Derin öğrenme algoritmalarının geleneksel kelime gömme modelleri ile kombinasyonunun Twitter veri setindeki doğruluk sonuçları.

Twitter	**MODELS**								
Hisseler	**Word2Vec+ CNN**	**Word2Vec+ RNN**	**Word2Vec+ LSTM**	**GloVe+ CNN**	**GloVe+ RNN**	**GloVe+ LSTM**	**FastText+ CNN**	**FastText+ RNN**	**FastText+ LSTM**
AKBNK	72.46	83.94	87.25	86.53	88.68	88.44	84.28	84.87	88.36
ALBRK	88.54	87.47	89.46	86.98	87.99	81.68	89.39	85.36	82.33
GARAN	87.23	85.27	78.98	85.75	81.51	80.72	82.48	80.59	78.83
HALKB	82.46	73.94	77.25	76.53	78.68	78.44	74.28	74.87	78.36
ISCTR	76.54	77.47	79.46	76.98	77.99	81.68	79.39	85.36	72.33
SKBNK	75.54	78.47	79.46	75.98	78.59	82.68	78.39	75.36	82.33
TSKB	72.46	73.92	75.20	56.53	58.68	68.44	74.21	74.85	78.34
VAKBN	75.54	76.47	79.46	77.98	78.59	80.68	78.39	79.36	82.33
YKBNK	81.23	80.87	78.98	85.75	81.51	80.72	82.48	80.59	78.83
Avg.	79.11	79.76	80.61	78.78	79.14	80.39	80.37	80.13	80.23

Tablo 4.1‘de verilen Twitter veri kümesi üzerinden elde edilen kombinasyonsuz doğruluk sonuçları Tablo 4.5‘te verilen kombinasyonlu halleri ile elde edilen doğruluk sonuçları ile karşılaştırıldığında CNN > Glove+CNN > FastText+CNN > Word2Vec+CNN olduğu elde dilmiştir. Bu sonuçtanda elde edileceği gibi kombinasyonsuz halde yani ham halde kullanılan CNN %95.58 doğruluk performansı ile daha iyi sonuçlar elde edilecektir. RNN > FastText+RNN > Word2Vec+RNN > GloVe+RNN şekilden sıralanacağından karşılaştırma sonucunda %87.26 doğruluk değeri ile kombinasyonsuz hali daha performanslı olacaktır. LSTM > Word2Vec+LSTM > GloVe+LSTM > FastText+LSTM şekilden bir sıralama ile sonuçlanacaktır. Bu modelde de yine LSTM‘in %95.49 olan doğruluk değerinden dolayı kombinasyonsuz halini kullanmak daha avantajlıdır.

Tablo 4.6 Derin öğrenme algoritmalarının geleneksel kelime gömme modelleri ile kombinasyonunun Mynet Finans veri setindeki doğruluk sonuçları.

Mynet Finans	**MODELS**								
Hisseler	**Word2Vec+ CNN**	**Word2Vec+ RNN**	**Word2Vec+ LSTM**	**GloVe+ CNN**	**GloVe+ RNN**	**GloVe+ LSTM**	**FastText+ CNN**	**FastText+ CNN**	**FastText+ LSTM**
AKBNK	87.98	88.59	70.68	88.39	89.36	72.33	84.21	84.85	88.34
ALBRK	84.67	84.77	88.85	82.66	83.82	85.20	86.53	86.68	89.44
GARAN	84.25	82.75	84.81	78.31	78.79	79.74	80.83	85.56	86.39
HALKB	77.98	78.59	80.68	78.39	79.36	82.33	74.21	74.85	78.34
ISCTR	74.67	74.77	78.85	72.66	73.82	75.20	76.53	76.68	79.44
SKBNK	74.61	74.75	78.84	72.66	73.82	75.20	76.53	78.68	78.44
TSKB	77.98	78.59	80.68	78.39	79.36	82.33	74.21	74.85	78.34
VAKBN	74.21	74.85	78.34	72.46	73.92	75.20	76.53	78.68	78.44
YKBNK	80.25	81.75	83.81	78.31	78.75	79.64	80.83	85.56	86.39
Avg.	79.62	79.93	80.62	78.03	79.00	78.57	78.93	80.71	82.62

Tablo 4.2‘de verilen Twitter veri kümesi üzerinden elde edilen kombinasyonsuz doğruluk sonuçları Tablo 4.6‘da verilen kombinasyonlu halleri ile elde edilen doğruluk sonuçları ile karşılaştırıldığında CNN > Glove+CNN > FastText+CNN > Word2Vec+CNN olduğu elde dilmiştir. Bu sonuçtanda elde edileceği gibi kombinasyonsuz halde yani ham halde kullanılan CNN %83.96 doğruluk performansı ile daha iyi sonuçlar elde edilecektir. RNN > FastText+RNN > Word2Vec+RNN > GloVe+RNN şekilden sıralanacağından karşılaştırma sonucunda %87.26 doğruluk değeri ile kombinasyonsuz hali daha performanslı olacaktır. Word2Vec+LSTM > GloVe+LSTM > FastText+LSTM > LSTM şekilden bir sıralama ile sonuçlanacaktır. LSTM modelinde Word2Vec ile kombinasyonlanması sonucunda %80.61 olan doğruluk değerinden dolayı bu model için kombinasyonlu halini kullanmak daha avantajlıdır. Ancak çalışmamızda çoğunluk olarak kombinasyonsuz halleri daha performasnlı sonuçlar ürettiğinden kombinasyonsuz derin öğrenme modelleri kullanılmıştır.

Tablo 4.7 Derin öğrenme algoritmalarının geleneksel kelime gömme modelleri ile kombinasyonun Big Para veri setindeki doğruluk sonuçları.

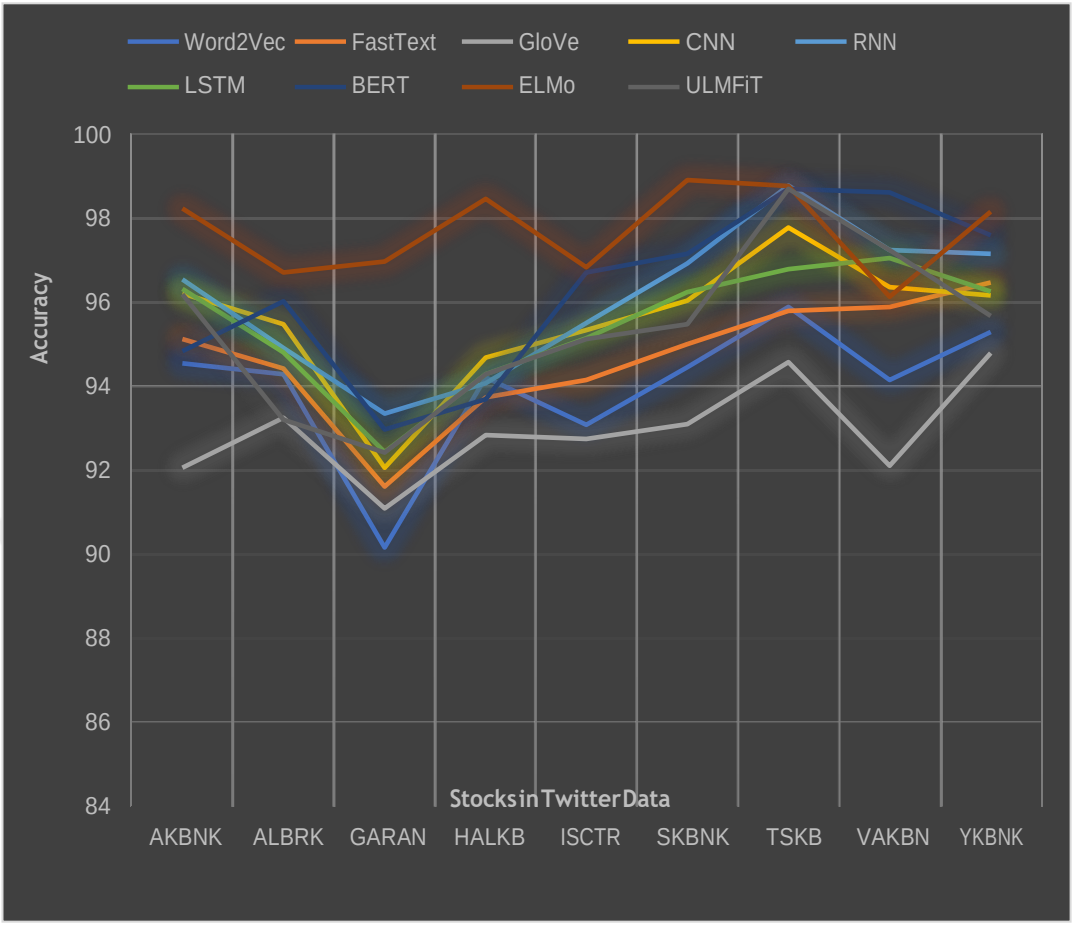
Big Para	**MODELS**								
Hisseler	**Word2Vec+ CNN**	**Word2Vec+ RNN**	**Word2Vec+ LSTM**	**GloVe+ CNN**	**GloVe+ RNN**	**GloVe+ LSTM**	**FastText+ CNN**	**FastText+ CNN**	**FastText+ LSTM**
AKBNK	90.00	90.02	90.13	89.00	87.02	86.13	80.00	80.02	70.13
ALBRK	92.02	91.07	91.12	81.02	81.07	71.12	82.02	81.07	81.12
GARAN	79.87	92.18	91.43	78.87	82.18	91.03	89.87	82.18	81.43
HALKB	80.00	97.11	70.01	78.00	95.11	81.01	78.00	87.11	80.01
ISCTR	87.43	86.00	84.31	77.43	89.00	87.31	85.43	76.00	89.31
SKBNK	77.59	95.87	82.57	79.59	85.87	82.12	77.14	85.54	72.75
TSKB	82.21	86.57	71.56	72.21	83.57	81.56	72.21	88.57	81.56
VAKBN	74.59	87.96	71.20	84.59	77.96	81.50	84.19	77.56	81.85
YKBNK	82.90	87.56	79.97	72.90	82.56	75.17	82.90	77.56	80.15
Avg.	82.96	90.48	81.37	79.29	84.93	81.88	81.31	81.73	79.81

Tablo 4.3‘te verilen Big Para veri kümesi üzerinden elde edilen kombinasyonsuz doğruluk sonuçları Tablo 4.7‘de verilen kombinasyonlu halleri ile elde edilen doğruluk sonuçları ile karşılaştırıldığında CNN > Word2Vec+CNN > FastText+CNN > GloVe+CNN olduğu elde dilmiştir. Bu sonuçtanda elde edileceği gibi kombinasyonsuz halde yani ham halde kullanılan CNN %83.86 doğruluk performansı ile daha iyi sonuçlar elde edilecektir. RNN > Word2Vec+RNN > GloVe+RNN > FastText+RNN şekilden sıralanacağından karşılaştırma sonucunda %96.61 doğruluk değeri ile kombinasyonsuz hali daha performanslı olacaktır. LSTM > GloVe+LSTM > Word2Vec+LSTM > FastText+LSTM şekilden bir sıralama ile sonuçlanacaktır. Bu modelde de yine LSTM‘in %85.98 olan doğruluk değerinden dolayı kombinasyonsuz halini kullanmak daha avantajlıdır.

Tablo 4.8 Derin öğrenme algoritmalarının geleneksel kelime gömme modelleri ile kombinasyonun KAP veri setindeki doğruluk sonuçları.

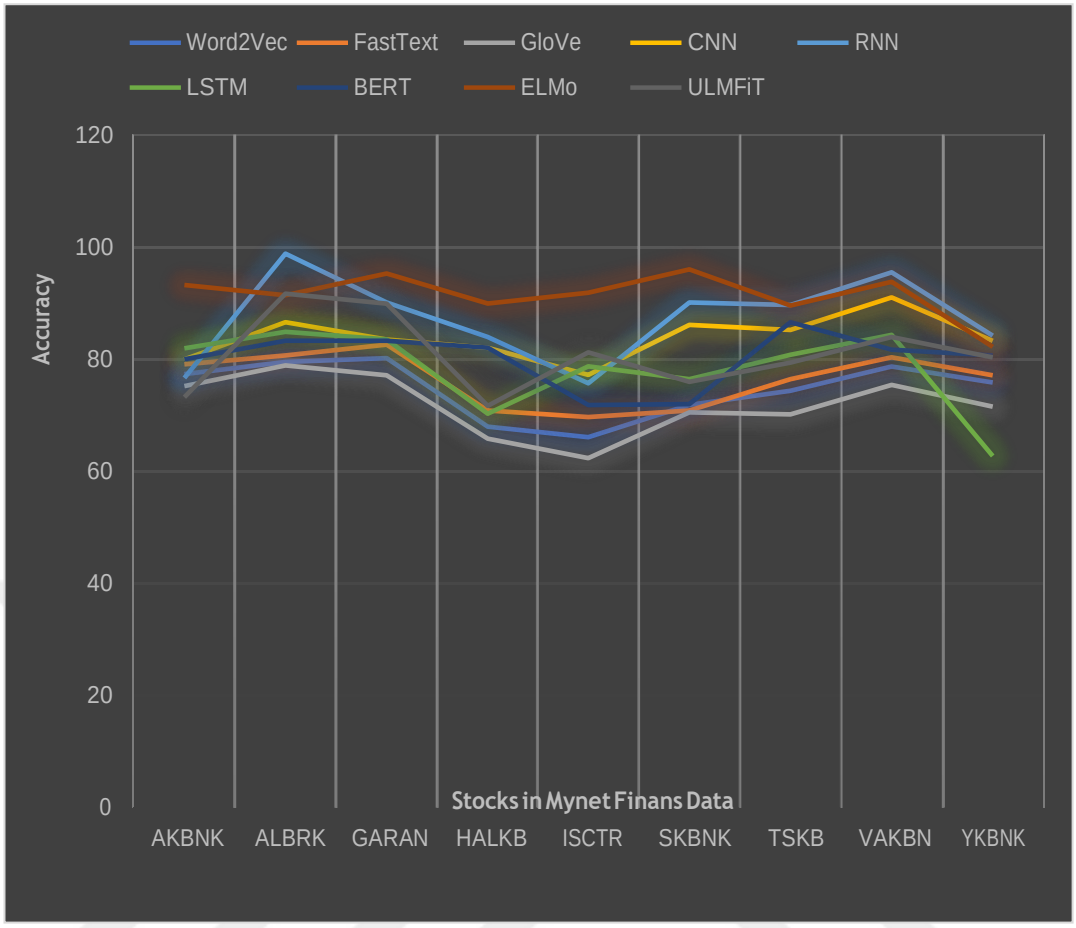
KAP	MODELS								
Hisseler	**Word2Vec+ CNN**	**Word2Vec+ RNN**	**Word2Vec+ LSTM**	**GloVe+ CNN**	**GloVe+ RNN**	**GloVe+ LSTM**	**FastText+ CNN**	**FastText+ CNN**	**FastText+ LSTM**
AKBNK	84.87	82.34	85.89	85.27	85.49	81.80	86.67	88.48	84.69
ALBRK	84.87	82.34	85.89	85.27	85.49	81.80	86.67	83.48	84.69
GARAN	81.23	80.87	78.98	87.67	82.25	81.86	78.36	76.79	75.69
HALKB	74.87	72.34	75.89	75.27	75.49	71.80	76.67	73.48	74.69
ISCTR	74.21	74.85	78.34	72.46	73.92	75.20	76.53	78.68	78.44
SKBNK	82.21	86.57	71.56	72.21	83.57	81.56	72.21	88.57	81.56
TSKB	78.56	75.42	75.33	74.21	74.85	78.34	78.99	77.89	79.73
VAKBN	64.87	62.34	65.89	75.20	71.49	71.90	76.64	73.46	74.62
YKBNK	81.23	80.87	78.98	87.67	82.25	81.86	78.36	76.79	75.69
Avg.	78.55	77.55	77.42	79.47	79.42	78.46	79.01	79.74	78.87

Tablo 4.4‘te verilen KAP veri kümesi üzerinden elde edilen kombinasyonsuz doğruluk sonuçları Tablo 4.8‘de verilen kombinasyonlu halleri ile elde edilen doğruluk sonuçları ile karşılaştırıldığında CNN > FastText+CNN > GloVe+CNN > Word2Vec+CNN olduğu elde dilmiştir. Bu sonuçtanda elde edileceği gibi kombinasyonsuz halde yani ham halde kullanılan CNN %97.93 doğruluk performansı ile daha iyi sonuçlar elde edilecektir. RNN > FastText +RNN > GloVe+RNN > Word2Vec +RNN şekilden sıralanacağından karşılaştırma sonucunda %94.44 doğruluk değeri ile kombinasyonsuz hali daha performanslı olacaktır. LSTM > FastText +LSTM > GloVe +LSTM > Word2Vec +LSTM şekilden bir sıralama ile sonuçlanacaktır. Bu modelde de yine LSTM‘in %95.66 olan doğruluk değerinden dolayı kombinasyonsuz halini kullanmak daha avantajlıdır.



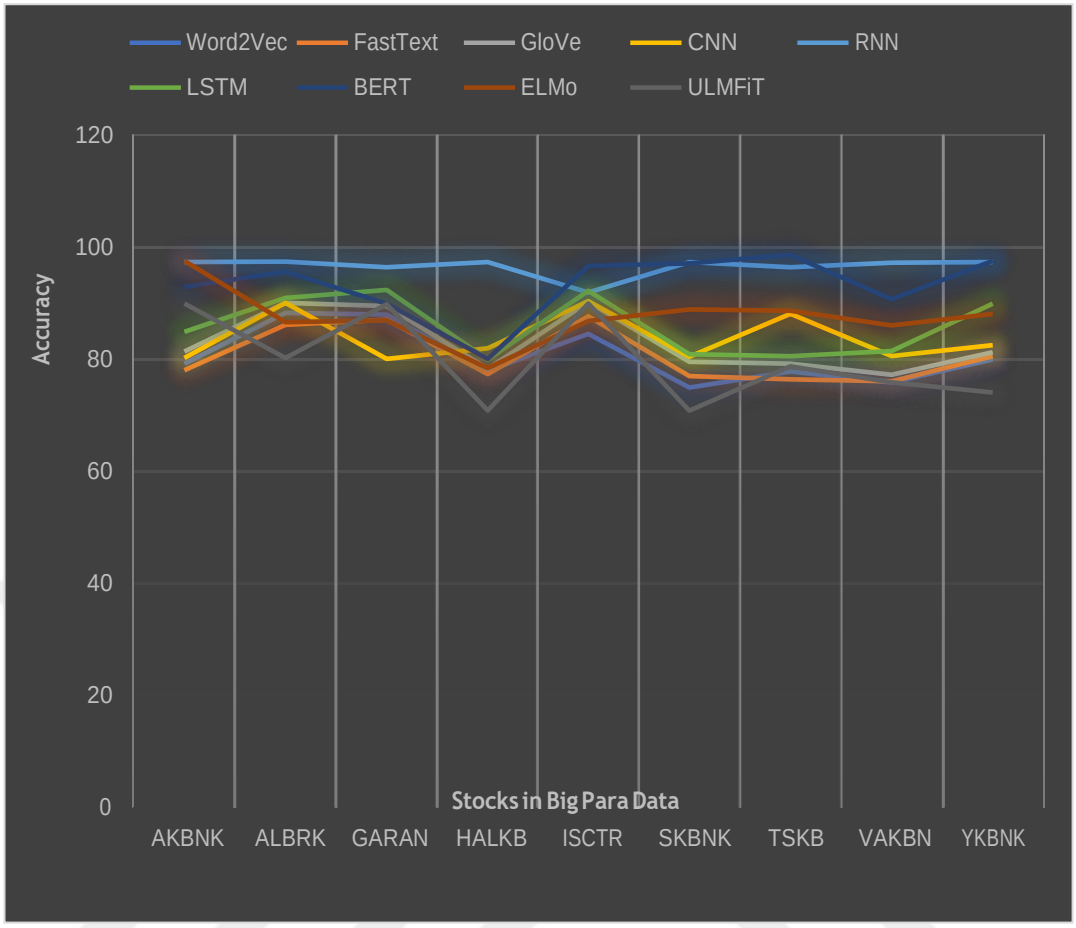
Şekil 4.1 Tüm ön işleme yöntemleri kullanılarak her kelime gömme, derin öğrenme ve yeni nesil kelime gömme modellerinin eğitim seti yüzdeleri açısından Twitter veri seti üzerinde sınıflandırma performansları.

Tüm geleneksel kelime gömme, derin öğrenme ve yeni nesil kelimde gömme modellerinin Twitter veri seti üzerindeki sınıflandırma performansları Şekil 4.1'de verilmiştir. Tabloda tüm doğruluk değerleri "AKBNK", "ALBRK", "GARAN", "HALKB", "ISCTR", "SKBNK", "TSKB", "VAKBN", "YKBNK" hisseleri için ayrı ayrı belirtilmiştir. Twitter veri seti içerisinde belirlenen hisselere ait Türkçe kullanıcı yorumlarının olduğu Bölüm 3.1'de belirtilmiştir. Tüm modellerden elde edilen sonuçlara bakıldığında geleneksel kelime gömme modelleri en düşük doğruluk performanıs sergilerken yeni nesil kelime gömme modelleri daha yüksek performans sergilemektedir. Yeni nesil kelime gömme modeli olan ELMo'nun SKBNK hissesi ile alakalı toplanan veri seti üzerinde %98 doğruluk değerinin üzerindeki sonucu ile borsa tahminlemesini yapmada en iyi performansı sergilediği kanıtlanmıştır.



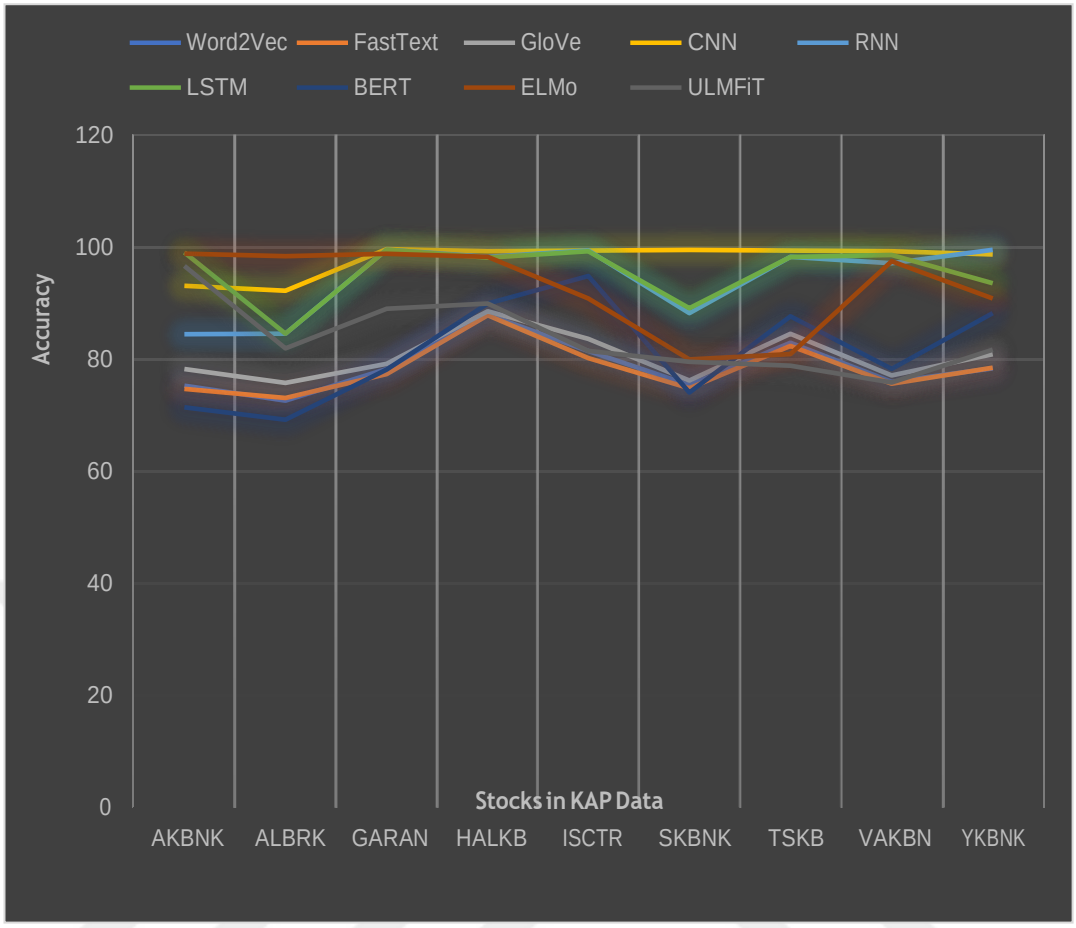
Şekil 4.2 Tüm ön işleme yöntemleri kullanılarak her kelime gömme, derin öğrenme ve yeni nesil kelime gömme modellerinin eğitim seti yüzdeleri açısından Mynet Finans veri seti üzerinde sınıflandırma performansları.

Şekil 4.2'de de bu sefer Mynet Finans'tan toplanan hisselere ait kullanıcı yorumlarının modellerden çıkan doğruluk performansları sergilenmektedir. Bu grafiğe göre de ALBRK hissesinde gelenkesel derin öğrenme modellerinden olan RNN'in yüksek performans göstermesine ragmen ortalama olarak bakıldığında yine yeni nesil kelime gömme modeli olan ELMo en iyi doğruluk değerlerine sahiptir. Mynet Finans veri kümesinde de geleneksel kelime gömme modelleri en düşük doğruluk performansını sergilemiştir. ELMo modeli yine Twitter veri setinde olduğu gibi Mynet Finans veri kümesinde de borsa tahminlemesi için en iyi doğruluk performansını SKBNK hissesinde göstermiştir.



Şekil 4.3 Tüm ön işleme yöntemleri kullanılarak her kelime gömme, derin öğrenme ve yeni nesil kelime gömme modellerinin eğitim seti yüzdeleri açısından Big Para veri seti üzerinde sınıflandırma performansları.

Şekil 4.3'e bakıldığında da Big Para'dan toplanan finansal analizlere ait veri kümesinin modellerden çıkan doğruluk performansları sergilenmektedir. Verilen değerlere göre geleneksel derin öğrenme modellerinden olan RNN'in en yüksek doğruluk performansı sergilediği görülmektedir. Yeni nesil kelime gömme modellerinin doğruluk değerinin diğer geleneksel kelime gömme ve derin öğrenme modellerine göre daha geride kaldığı gözlemlenmektedir. Türkçe metine dayalı veri seti içeren Big Para verilerininde borsa tahminlemesi yapılacağında RNN modelinin %96.61 doğruluk değeriyle en iyi performans sergilediği kanıtlanmıştır.



Şekil 4.4 Tüm ön işleme yöntemleri kullanılarak her kelime gömme, derin öğrenme ve yeni nesil kelime gömme modellerinin eğitim seti yüzdeleri açısından KAP veri seti üzerinde sınıflandırma performansları.

Şekil 4.4'te KAP sitesinden toplanan finansal haberlere ait tüm modellere gönderilmesi sonucunda çıkan doğruluk performansları sergilenmektedir. Verilen değerlere göre geleneksel derin öğrenme modellerinin diğer modellere göre daha iyi doğruluk performansı sergilediği analiz edilmiştir. Yine Big Para veri setinden olduğu gibi KAP veri setinde de Türkçe metine dayalı veriler yer almaktadır. Bu veriler üzerinde yeni nesil kelime gömme modelleri olan BERT, ELMo ve ULMFiT derin öğrenme modellerine göre daha düşük performanslar sergilerken geleneksel kelime gömme modellerine göre daha yüksek performasnlar sergilemiştir. Tüm sonuçlar göz önünde bulundurulduğunda ortalama %97.93 doğruluk oranıyla derin öğrenme modellerinden olan KAP en iyi doğruluk performansını elde etmiştir.

Sonuç olarak Tablo 4.5 ve Tablo 4.6'da da görüldüğü gibi kullanıcı yorumlarını içeren Türkçe veri kümeleri kullanarak borsa tahminlemesi yapılmak istenen çalışmalarda yeni nesil kelime gömme modellerinin kullanımının daha performanslı

olacağı saptanmıştır. Ancak Tablo 4.7 ve Tablo 4.8'de olduğu gibi Türkçe metine dayalı veri kümeleri kullanarak borsa tahminlemesi yapılacağında geleneksel derin öğrenme modelleri daha iyi performanslar sergilemiştir. Çalışmamızın sınıflandırma performansı değerlendirildiğinde, Türkçe kullanıcı yorumu içeren metinsel verilerin kullanımında, son teknolojiye kıyasla bir yenilik sunmuş olup, derin öğrenme modelleri, geleneksel kelime gömme modelleri ve yeni nesil kelime gömme modellerinin birlikte kullanımıylada son teknoloji çalışmalara kıyasla daha yenilikçi ve rekabetçi olarak değerlendirilebilir.

Bildiğimiz kadarıyla, bu çalışma borsa tahminlemesini yapmak için sosyal medya platformundan kullanıcı yorumlarını, finansal sitelerden de analizleri ve haberleri toplayarak analiz etmeye yönelik yapılan ilk girişim. Ayrıca, bu çalışmanın diğer bir yeniliği de, yeni nesil kelime gömme mollerinin geleneksel derin öğrenme ve kelime gömme modelleriyle kıyaslanarak kullanılması sonucu, borsa tahminlemesinde yatırımcılarak büyük bir kaynak sağlıyor olmasıdır.

5. SONUÇ

Bu çalışmada, borsa yönünü öngörme konusundaki son araştırmalardan farklı olarak, BIST100'de büyük hacimli stokları analiz ederek borsa yönünü belirlemek için sosyal medya platformundan, finansal analiz ve haber sitelerinden toplanan Türkçe veri setlerini kullanarak finansal duyarlılık analizine odaklanıyoruz. Bu amaçla, BIST100'ün yönünü belirlemek için kullanıcıların sosyal medya platformlarındaki finansal haberler hakkındaki yorumlarını, finansal haberlerin yönünü ve finansal analizleri anlama ve analiz etme ihtiyacı vardır. Bunu gerçekleştirmek için geleneksel kelime gömme modellerinden; Word2Vec, GloVe, FastText, derin öğrenme yaklaşımlarından; Konvolüsyonel Sinir Ağları (CNN), Tekrarlayan Sinir Ağları (RNN), Uzun Kısa Süreli Bellek Ağları (LSTM), yeni nesil kelime gömme modellerinden Çift Yönlü Kodlayıcı Gösterimleri (BERT), Dil Modellerinden Yerleştirme (ELMo) ve Evrensel Dil Modeli İnce Ayar (ULMFiT) kullanılmıştır. Ayrıca, önerilen modelin sınıflandırma performansını iyileştirmek için, aynı zamanda kelimelerin kaldırılmasını, hashtag'lerin kaldırılmasını, URL'lerin kaldırılmasını ve ön işleme yöntemleri olarak ortaya çıkmasını düşünüyoruz. Sonuç olarak, kullanıcı yorumlarını içeren Türkçe metinlerini sınıflandırmada yeni nesil kelime gömme modeli olan ELMo'nun ön işleme yöntemleriyle birleşimi, kullanıcıların stoklarını yönlendirmedeki hassasiyetini belirlemek ve en iyi sınıflandırma başarısı elde etmek için avantajlı bir seçim olacaktır. Ancak haberler ve analizler gibi Türkçe metin içerikli veri setlerinde yeni nesil kelime gömme modellerine göre geleneksel derin öğrenme modelleri daha iyi sonuçlar üretmiştir. Buradan da anlaşılacağı üzere kullanıcı yorumlarını içeren bir veri setimiz varsa yeni nesil kelime gömme modellerini tercih etmek daha avantajlı olacaktır. Türkçe metin içerikli bir veri setimiz varsa geleneksel derin öğrenme modellerini tercih etmek daha avantajlı olacaktır. Bu durum, kullanıcıların yorumlarının analizinin yatırımcılara yatırımlarını yönlendirecek bir perspektif sunacağı anlamına gelir.

Çalışmamız boyunca genel olarak yeni nesil kelime gömme modellerinin standart bir makine gücüne sahip bilgisayarda çalıştırırken uzun çalışma süresinden dolayı sıkıntı yaşadık. Bu sebeple yeni nesil kelime gömme modelleri google'ın sağlamış olduğu "colab" üzerinde çalıştırıldı.

Bildiğimiz kadarıyla, yapmış olduğumuz bu çalışma yeni nesil kelime gömme modellerini ve çeşitli Türkçe veri kümesini kullanarak borsa hisselerinin yönünü tahmin

etmede yapılan ilk çalışmadır. Bu sebeple yatırımcılara yatırımlarına yön verecek büyük bir kaynak olmaktadır. Gelecekteki bir çalışma olarak, borsa tahminlerini daha da güçlendirmek için diğer yeni nesil kelime gömme modellerini kullanarak modelimizi geliştirmeyi planlıyoruz.

KAYNAKÇA

Akcan, A., & Kartal, C. (2011). İMKB Sigorta Endeksini Olusturan Sirketlerin Hisse Senedi Fiyatlarının Yapay Sinir Ağları İle Tahmini. *Muhasebe ve Finansman Dergisi*, (51), 27-40.

Beykikhoshk, A., Arandjelović, O., Phung, D., & Venkatesh, S. (2015, August). Overcoming data scarcity of Twitter: using tweets as bootstrap with application to autism-related topic content analysis. In *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015* (pp. 1354-1361). ACM.

Bruns, A., Kornstadt, A., & Wichmann, D. (2009). Web application tests with selenium. *IEEE software*, *26*(5), 88-91.

Türkmen, A. C., & Cemgil, A. T. (2015, May). An application of deep learning for trade signal prediction in financial markets. In *2015 23nd Signal Processing and Communications Applications Conference (SIU)* (pp. 2521-2524). IEEE.

Hasan, A., Kalıpsız, O., & Akyokuş, S. (2017, October). Predicting financial market in big data: deep learning. In *2017 International Conference on Computer Science and Engineering (UBMK)* (pp. 510-515). IEEE.

Joulin, A., Grave, E., Bojanowski, P., & Mikolov, T. (2016). Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems* (pp. 1097-1105).

Voulodimos, A., Doulamis, N., Doulamis, A., & Protopapadakis, E. (2018). Deep learning for computer vision: A brief review. *Computational intelligence and neuroscience*, *2018*.

Bataa, E., & Wu, J. (2019). An Investigation of Transfer Learning-Based Sentiment Analysis in Japanese. *arXiv preprint arXiv:1905.09642*.

Beykikhoshk, A., Arandjelović, O., Phung, D., & Venkatesh, S. (2015, August). Overcoming data scarcity of Twitter: using tweets as bootstrap with application to autism-related topic content analysis. In *Proceedings of the 2015 IEEE/ACM International*

Conference on Advances in Social Networks Analysis and Mining 2015 (pp. 1354-1361). ACM.

Bruns, A., Kornstadt, A., & Wichmann, D. (2009). Web application tests with selenium. *IEEE software*, *26*(5), 88-91.

Chatterjee, A., Narahari, K. N., Joshi, M., & Agrawal, P. (2019, June). SemEval-2019 task 3: EmoContext contextual emotion detection in text. In *Proceedings of the 13th International Workshop on Semantic Evaluation* (pp. 39-48).

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Kilimci, Z. H., & Akyokuş, S. (2019, September). The Evaluation of Word Embedding Models and Deep Learning Algorithms for Turkish Text Classification. In *2019 4th International Conference on Computer Science and Engineering (UBMK)* (pp. 548-553). IEEE.

Gunduz, H., Yaslan, Y., & Cataltepe, Z. (2017). Intraday prediction of Borsa Istanbul using convolutional neural networks and feature correlations. *Knowledge-Based Systems*, *137*, 138-148.

Gunduz, H., Cataltepe, Z., & Yaslan, Y. (2017, May). Stock market direction prediction using deep neural networks. In *2017 25th Signal Processing and Communications Applications Conference (SIU)* (pp. 1-4). IEEE.

Elman, J. L. (1990). Finding structure in time. *Cognitive science*, *14*(2), 179-211.

Pennington, J., Socher, R., & Manning, C. (2014, October). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 1532-1543).

Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural networks*, *61*, 85-117.

Howard, J., & Ruder, S. (2018). Universal language model fine-tuning for text classification. *arXiv preprint arXiv:1801.06146*.

Chen, K., Zhou, Y., & Dai, F. (2015, October). A LSTM-based method for stock returns prediction: A case study of China stock market. In *2015 IEEE International Conference on Big Data (Big Data)* (pp. 2823-2824). IEEE.

Schneider, K. M. (2004, October). On word frequency information and negative evidence in Naive Bayes text classification. In *International Conference on Natural Language Processing (in Spain)* (pp. 474-485). Springer, Berlin, Heidelberg.

Kilimci, Z. H., & Akyokus, S. (2018). Deep Learning-and Word Embedding-Based Heterogeneous Classifier Ensembles for Text Classification. *Complexity, 2018*.

Kilimci, Z. H., Akyokus, S., & Omurca, S. I. (2016, August). The effectiveness of homogenous ensemble classifiers for Turkish and English texts. In *2016 International Symposium on INnovations in Intelligent SysTems and Applications (INISTA)* (pp. 1-7). IEEE.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems* (pp. 1097-1105).

Zhang, L., Wang, S., & Liu, B. (2018). Deep learning for sentiment analysis: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *8*(4), e1253.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *nature*, *521*(7553), 436-444.

Lipton, Z. C., Berkowitz, J., & Elkan, C. (2015). A critical review of recurrent neural networks for sequence learning. *arXiv preprint arXiv:1506.00019*.

Loria, S. (2018). *textblob Documentation* (pp. 1-73). Technical report.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Çelik, N., & Kaya, M. F. (2010). Uç değerler yöntemi ile riske maruz değer'in tahmini ve İstanbul Menkul Kıymetler Borsası üzerine bir uygulama. *Bankacılık ve Sigortacılık Araştırmaları Dergisi*, *1*(1), 19-32.

Pervan, N., & Keleş, Y. *Derin öğrenme yaklaşımları kullanarak Türkçe metinlerden anlamsal çıkarım yapma* (Doctoral dissertation).

Phang, J., Févry, T., & Bowman, S. R. (2018). Sentence encoders on stilts: Supplementary training on intermediate labeled-data tasks. *arXiv preprint arXiv:1811.01088*.

Prashanth, R., & Roy, S. D. (2018). Novel and improved stage estimation in Parkinson's disease using clinical scales and machine learning. *Neurocomputing*, *305*, 78-103.

Prieto, V. M., Matos, S., Alvarez, M., Cacheda, F., & Oliveira, J. L. (2014). Twitter: a good place to detect health conditions. *PloS one*, *9*(1), e86191.

Le, Q., & Mikolov, T. (2014, January). Distributed representations of sentences and documents. In *International conference on machine learning* (pp. 1188-1196).

Tekin, S., & Çanakoğlu, E. (2018, May). Prediction of stock returns in Istanbul stock exchange using machine learning methods. In *2018 26th Signal Processing and Communications Applications Conference (SIU)* (pp. 1-4). IEEE.

Sakar, C. O., & Kursun, O. (2010). Telediagnosis of Parkinson's disease using measurements of dysphonia. *Journal of medical systems*, *34*(4), 591-599.

Santos, I., Nedjah, N., & de Macedo Mourelle, L. (2017, November). Sentiment analysis using convolutional neural network with fastText embeddings. In *2017 IEEE Latin American Conference on Computational Intelligence (LA-CCI)* (pp. 1-5). IEEE.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems* (pp. 3111-3119).

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

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Young, T., Hazarika, D., Poria, S., & Cambria, E. (2018). Recent trends in deep learning based natural language processing. *ieee Computational intelligenCe magazine*, *13*(3), 55-75.

Tekin, S., & Çanakoğlu, E. (2019, April). Analysis of Price Models in Istanbul Stock Exchange. In *2019 27th Signal Processing and Communications Applications Conference (SIU)* (pp. 1-4). IEEE.

Tunali, V., & Bilgin, T. T. (2012, June). PRETO: A high-performance text mining tool for preprocessing turkish texts. In *Proceedings of the 13th International Conference on Computer Systems and Technologies* (pp. 134-140). ACM.

Prieto, V. M., Matos, S., Alvarez, M., Cacheda, F., & Oliveira, J. L. (2014). Twitter: a good place to detect health conditions. *PloS one*, *9*(1), e86191.

Kilimci, Z. H., & Akyokus, S. (2018). Deep Learning-and Word Embedding-Based Heterogeneous Classifier Ensembles for Text Classification. *Complexity, 2018*.

Kilimci, Z. H., Akyokus, S., & Omurca, S. I. (2016, August). The effectiveness of homogenous ensemble classifiers for Turkish and English texts. In *2016 International Symposium on INnovations in Intelligent SysTems and Applications (INISTA)* (pp. 1-7). IEEE.

Zhang, L., Wang, S., & Liu, B. (2018). Deep learning for sentiment analysis: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *8*(4), e1253.

ÖZGEÇMİŞ

Derya Othan 1994 yılında Kars'ta doğdu. 2017 yılında Selçuk Üniversitesi Bilgisayar Mühendisliği programından mezun oldu. 2018 yılında Doğuş Üniversitesi yüksek lisans eğitimine başladı. Haziran 2017'de Türk Telekom Şirketi'nde İş Analisti olarak görev yaptı. Şubat 2019'dan beri Türkiye Finans Katılım Bankası'nda İş Analisti olarak görev yapmaktadır. Türkiye Finans Katılım Bankası'ndaki çalışma alanları: Talepler ve iş süreçlerinin araştırılması ve geliştirilmesi, teknik analiz ile tüm toplama işlemlerinin gerçekleştirilmesi (Sistem Gereksinimleri), yazılım eksiklikleri ve yeni talepler için ilgili teknik ve teknik olmayan belgelerin hazırlanmasıdır.