DEÜ FMD 23(67), 121-127, 2021

Dokuz Eylül Üniversitesi Mühendislik Fakültesi Fen ve Mühendislik Dergisi
Dokuz Eylul University Faculty of Engineering Journal of Science and Engineering

Basılı/Printed ISSN: 1302-9304. Elektronik/Online ISSN: 2547-958X

TF-IDF, Word2vec ve Fasttext Vektör Model Yöntemleri ile Türkçe Haber Metinlerinin Sınıflandırılması

Classification of Turkish News Text by TF-IDF, Word2vec And Fasttext Vector Model Methods

Özer Çelik [1*], Burak Can Koç [2]

[1,2] Osmangazi Üniversitesi, Matematik-Bilgisayar Bölümü, Eskişehir, TÜRKİYE
*Sorumlu Yazar / Corresponding Author *:* ozer@ogu.edu.tr

Geliş Tarihi / Received: 22.01.2020 *Araştırma Makalesi/Research Article*
Kabul Tarihi / Accepted: 21.07.2020 *DOI:10.21205/deufmd.2021236710*
Atıf şekli/ How to cite: ÇELIK, Ö.,KOÇ, B.C. (2021). TF-IDF, Word2vec ve Fasttext Vektör Model Yöntemleri ile Türkçe Haber Metinlerinin Sınıflandırılması. DEUFMD, 23(67), 121-127.

Öz

Bilgisayar ve internetin hayatımıza girmesi ile bilgiye erişmek daha kolay hale gelmiştir. İnternete ulaşımın kolaylaşması ve internet kullanıcılarının artması sonucu veri miktarı da her geçen saniye büyümektedir. Ancak doğru bilgiye erişebilmek için verilerin sınıflandırılması gereklidir. Sınıflandırma, verilerin belirli bir anlamsal kategoriye göre ayrılması işlemidir. Dijital belgelerin anlamsal kategorilere ayrılması, metnin ulaşılabilirliğini önemli ölçüde etkilemektedir. Bu çalışmada, farklı Türkçe haber kaynaklarından toplam 6 kategoride elde edilen veri kümesi üzerinde metin sınıflandırma çalışması yapılmıştır. Öncelikli olarak haber metinleri ön işlemeden geçirilmiş ve gövdelenmiştir. Ön işlemeden geçirilen metinler Tfidfvectorizer, Word2Vec ve FastText yöntemleri ile ayrı ayrı vektörize edildikten sonra Python'ın Scikit-learn kütüphanesi kullanılarak Destek Vektör Makinesi (Support Vector Machine, SVM), Naive Bayes, Logistic Regression, Random Forest ve Yapay Sinir Ağı (Artificial Neural Network, ANN) yöntemleri ile sınıflandırılmıştır. Yapılan çalışma sonucuna göre en yüksek başarı oranı %95,75 ile FastText yöntemi ve vektör modeli ile elde edilen metnin SVM ile sınıflandırılmasından elde edilmiştir.

Anahtar Kelimeler: *Metin Sınıflandırma, Türkçe Haber, TF-IDF, Word2Vec, Fasttext*

Abstract

Accessing information has become very simple with computers and internet. As the internet access is easier and the internet users increase, the amount of data is growing every second. However, in order to access correct information, data must be classified. Classification is the process of separating data according to a certain semantic category. Dividing digital documents into semantic categories significantly affects the availability of the text. In this study, a text classification study was carried out on a data set obtained from different Turkish news sources with 6 categories. After the pre-processed texts are separately vectorized with Tfidfvectorizer, Word2Vec and FastText methods, they are classified with Support Vector Machine (SVM), Naive Bayes, Logistic Regression, Random Forest and Artificial Neural Network (ANN) methods by using Scikit-learn library in Python. According to the results of the study, the highest success rate was obtained from the classification of the text gained with FastText method and vector model with 95,75% by SVM.

Keywords: *Text Classification, Turkish News, TF-IDF, Word2Vec, Fasttext*

1. Giriş

Bilgisayarın ve özellikle internetin hayatımıza girmesi ile bilgiye erişim gün geçtikçe kolaylaşmaktadır. Bununla birlikte çok büyük bir hızla üretilen veriler arasında kaybolmamak ve istenmeyen verilerden korunabilmek için verileri sınıflandırma mecburiyeti doğmuştur. Bugün günlük hayatımızda çok sık kullandığımız e-postalardan kısa mesajlara, web sitelerine kadar metin sınıflandırma önemli bir konu haline gelmiştir. İstenilen içeriğe ulaşabilmek ve istenmeyen verilerden korunabilmek için metinlerin en iyi şekilde tanımlanması gerekir.

Makine öğreniminin amacı, bilgi sistemlerindeki karmaşık sorunları tespit etmek ve onlara rasyonel çözümler sunmaktır. Bu durum, makine öğreniminin istatistik, veri madenciliği, örüntü tanıma, yapay zekâ ve teorik bilgisayar bilimi gibi alanlarla yakından ilgili olduğunu ve çok disiplinli bir çalışma gerektirdiğini göstermektedir [1].

Makine öğrenmesi yöntemleri metin sınıflandırması için etkili bir şekilde kullanılmaktadır. Metin sınıflandırmanın amacı, belgelerin belirli bir anlamsal kategoriye otomatik olarak ayrılmasıdır. Doğal dil işleme (NLP), veri madenciliği ve makine öğrenimi teknikleri, elektronik belgeleri otomatik olarak sınıflandırmak için birlikte kullanılır. Her belge kategorisiz, bir veya birden fazla kategoride olabilir. Makine öğreniminin amacı, sınıflandırıcıları, otomatik olarak kategori atayan örneklerden öğrenmektir [2][3]. Makine öğrenmesi ile metinlerin sınıflandırılabilmesi için metinleri sayısal olarak ifade etmek gerekir. Bu işlem için çeşitli yaklaşımlar geliştirilmiştir. Metinlerin vektör modellerini çıkarabilmek için kullanılan yöntemlerden Term Frequency-Inverse Document Frequency (TF-IDF), Word2Vec ve FastText yöntemlerinin literatür taramasına göre daha iyi sonuçlar verdiği gözlemlenmiştir.

1.1. Kelime Gömme (Word Embedding) Teknikleri

Aynı zamanda kelime temsili olarak da bilinen kelime gömme, kelimenin dokümandaki anlamına göre sürekli kelime vektörlerinin oluşturulmasında hayati bir rol oynamaktadır. Kelime gömme, sözcüklerin hem anlamsal hem de sözdizimsel bilgilerini yakalar ve NLP görevlerinde yaygın olarak kullanılan kelime benzerliklerini ölçmek için kullanılabilir [4].

Bu çalışmada, kelime vektörleştirme yöntemleri ile Türkçe haber metinlerini sınıflandırmada yüksek başarı oranına ulaşmak hedeflenmiştir. Her bir kategoride 2000 haberin olduğu 12000 haberlik yeni bir derlem oluşturulmuştur. Böylece TTC-3600 verisetinin yaklaşık 3.5 katı daha fazla haber içeren bir veriseti elde edilmiştir. TF-IDF, Word2Vec ve Fasttext yöntemlerinin kullanıldığı çalışmada, literatürdeki bu yöntemi kullanan çalışmalardan daha büyük bir veriseti ile daha göre daha yüksek başarı oranı alınmıştır.

1.1.1 Terim Frekansı-Ters Doküman Frekansı (Term Frequency Inverse Document Frequency, TF-IDF)

Terimden de anlaşılacağı gibi, TF-IDF, bir belgedeki her bir kelimenin değerlerini, belirli bir belgedeki kelimenin sıklığı ile kelimenin göründüğü belgelerin yüzdesinin tersiyle hesaplar. Temel olarak, TF-IDF, belirli bir belgede kelimelerin göreceli sıklığını, bu kelimenin tüm veri seti üzerindeki tersine oranına göre belirleyerek çalışır. Sezgisel olarak, bu hesaplama, belirli bir kelimenin belirli bir belge ile ne kadar alakalı olduğunu belirler. Tek veya küçük bir belge grubunda ortak olan kelimeler genel kelimelerden daha yüksek TF-IDF numaralarına sahip olma eğilimindedir [5].

$$w_{i,j} = tf_{i,j} \times log\frac{N}{df_i} \quad (1)$$

$tf_{i,j}$ = i kelimesinin j belgesinde bulunma sıklığı

df_i = i kelimesini içeren doküman sayısı

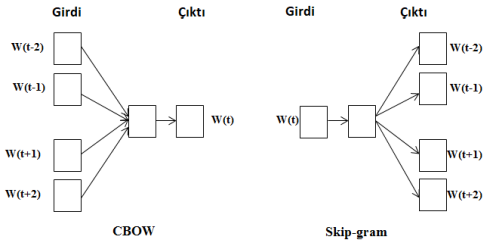
N = toplam doküman sayısı

1.1.2 Word2Vec

Word2Vec Mikolov ve arkadaşları tarafından 2013 yılında önerilmiştir [6]. Bu yöntem kelime ile kelimenin belirli pencere boyutundaki komşuları arasında yakınlık ilişkisi kurar ve anlamca yakın kelimeleri vektör uzayında birbirine yakın olacak şekilde konumlandırır. Anlam ilişkisi kurabilmek için iki farklı öğrenme mimarisi kullanır.

Bunlardan ilki Continous Bag of Words (CBoW) mimarisidir. Bu yöntemde pencere merkezindeki kelime, kelimenin pencere boyutu kadar yakınındaki komşularına bakılarak tahmin edilmeye çalışılır.

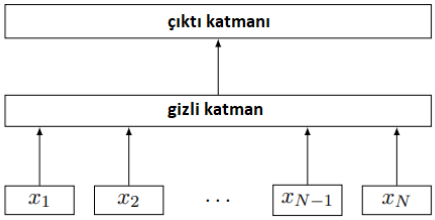
Diğer yöntem ise Skip-Gram mimarisidir. Bu yöntem de CBoW ile benzer çalışır. Ancak Skip-Gram, CBoW'un aksine pencere merkezinde konumlandırılan kelimeden, kelimenin komşularını tahmin eder. Bu yöntemin avantajı farklı anlamlara gelebilecek kelimelerin birden fazla anlamını yakalayabilmesidir.



Şekil 1. CBoW ve Skip-Gram model mimarisi

1.1.3 FastText

FastText ise 2016 yılında Facebook tarafından geliştirilen Word2Vec tabanlı bir modeldir. Bu yöntemin Word2Vec'ten farkı, kelimelerin n-gramlara ayrılmasıdır. Böylece Word2Vec ile yakalanamayan anlam yakınlığı bu yöntemle yakalanabilir [7].



Şekil 2. FastText model mimarisi

1.2 İstatistiksel Analiz

Araştırmada kullanılan veri setine makine öğrenme teknikleri uygulandıktan sonra doğruluk oranları karmaşıklık (confusion) matrisi kullanılarak hesaplanmıştır. Karmaşıklık matrisi, bir veri kümesinde doğru ve yanlış sınıflandırılmış veri gruplarının sayısını veren matristir.

Çalışmamızda çok sınıflı verilerde yaygın olarak kullanılan bir başarı değerlendirme yöntemi olan Genel Doğruluk Oranı (Overall Accuracy, OACC) kullanılmıştır [8].

Tablo 1. Karmaşıklık Matrisi

		Gerçek Sonuç			
		Sınıf 1	Sınıf 2	Sınıf 3	Toplam
Tahmin	Sınıf 1	a	b	c	j
	Sınıf 2	d	e	f	k
	Sınıf 3	g	h	i	l
	Toplam	m	n	o	N

Başarı skorları, karmaşıklık matrisi yardımıyla hesaplanmaktadır (Tablo 2). Çalışmamızda kullanılan ve karmaşıklık matrisi yardımıyla hesaplanan başarı ölçüleri ve formülleri;

Tablo 2. Karmaşıklık matrisi yardımı ile hesaplanan başarı ölçüleri ve formülleri

Formüller
N = (a + b + c + d + e + f + g + h + i)
j = a + b + c
k = d + e + f
l = g + h + i
m = a + d + g
n = b + e + h
o = c + f + i
OACC = (a + e + i) / N
Precision1 = a / j
Precision2 = e / k
Precision3 = i / l
Recall1 = a / m
Recall2 = e / n
Recall3 = i / o
Specificity1 = (e + f + h + i) / (k + l)
Specificity2 = (a + c + g + i) / (j + l)
Specificity3 = (a + b + d + e) / (j + k)

Tüm analiz ve işlemlerde, Windows 10 64 bit işletim sistemine, dört çekirdekli Intel Skylake Core i5-6500 CPU 3.2 GHz 6 MB cache ve 8 GB 2400 MHz DDR4 Ram belleğe sahip bir bilgisayar kullanılmıştır.

1.3 Literatür Taraması

Sawaf ve ark., Arapça haber metinlerini sınıflandırma amacıyla maksimum entropi ve Generalized Iterative Scaling (GIS) teknikleriyle %62,7 oranında başarılı bir sınıflandırma yapmıştır [9]. Hakim ve ark., Bahasa Endonezya'daki 15 farklı kategoride haber makalelerinin sınıflandırılmasıyla ilgili bir çalışmalarında TF-IDF ile %98,3 düzeyinde bir başarı elde etmişlerdir [10]. Dilrukshi ve ark., sosyal bir ağdaki 12 gruba ayrılan haber metinlerini sınıflandırma üzerine bir çalışmalarında, haberleri Bag of Words (BoW) yöntemi ile vektörize ederek SVM algoritması ile %75'in üzerinde bir başarı oranı elde etmişlerdir [11]. Amasyalı ve Yıldırım yaptıkları çalışmada gazetelerin web sayfalarındaki haber metinleri otomatik olarak sınıflandırılmaya çalışılmış,

metinler 5 haber sınıfına ayrılmış, en iyi sonuçlara Öğrenmeli Vektör Kuantalama (Learning Vector Quantization, LVQ) ve Naive Bayes metotları ile ulaşılmış ve %76 oranında başarı elde edilmiştir [12]. Tüfekçi ve arkadaşlarının, 5 kategorinin olduğu iki farklı haber derlemi üzerinde yaptıkları çalışmada, azaltılmış özellik vektörü kullanılarak Naive Bayes, SVM, C4.5 ve RF sınıflandırma metotlarından alınan sonuçlar genelde daha yüksek olmakla birlikte en yüksek başarı %92,73 doğruluk ile Naive Bayes algoritmasından elde edilmiştir [13]. Şen ve Yanıkoğlu yaptıkları çalışmada iki çevrimiçi haber sitesinden büyük miktarlı metin derlemleri toplanmış ve çalışmada aynı anda hem kelime vektörlerini hem de doküman sınıflandırmasını öğrenen bir yapay sinir ağı 4 kategoriye ayrılan Sabah ve 14 kategoriye ayrılan Cumhuriyet derlemleri için sırasıyla %88,28 ve %74,31 ile en iyi sonucu vermiştir [14]. Acı ve Çırak yaptıkları çalışmada, Konvolüsyonel Sinir Ağları (KSA, CNN) ve Word2Vec metodu kullanılarak Turkish Text Classification 3600 (TTC-3600) veri kümesi üzerinde metin sınıflandırma çalışması yapılmış, CNN ve Word2Vec metodu, %93,30 doğruluk ile klasik istatistiksel ve makine öğrenmesine dayalı sınıflandırma algoritmalarından daha iyi bir performans göstermiştir [15]. Erdinç ve Güran 11 farklı kategorideki 5 milyon Türkçe haber dokümanı içeren bir derlem üzerinde yaptıkları çalışmada yarı öğreticili bir teknik ile Word2Vec, Doc2Vec ve FastText algoritmalarının metin sınıflandırma problemi üzerindeki başarım değerleri kıyaslanmıştır. Haber derlemi sınıflandırılırken Naive Bayes, Destek Vektör Makineleri, Yapay Sinir Ağları, Karar Ağaçları ve Logistik Regresyon sınıflandırma algoritmaları kullanılmış, karakter n-gramların kullanılmasıyla test verisi daha iyi genelleştirilmiş ve en iyi başarım değerine %78 oranında FastText algoritması ile ulaşılmıştır [16]. Lİteratür taramasında tespit edilen çalışmların özeti Tablo 3'te verilmiştir.

Tablo 3. Literatür Özeti

Çalışma	Vektörleştirme Yöntemi	Sınıflandırma Yöntemi Başarı Oranı
Sawal ve ark. (2001)	Maximum Entropy	Generalized Iterative Scaling
Hakim ve ark. (2014)	TF-IDF	-
Dilrukshi ve ark. (2013)	Bag-of-words	SVM (%75)
Amasyali & Yildirim (2004)	LVQ	Naive Bayes (%76)
Tüfekçi ve ark. (2012)	TF	Naive Bayes (%92,73)
Sen & Yanıkoğlu (2018)	TF-IDF	ANN (%88,28)
Acı ve Çırak (2019)	Word2Vec	CNN (%93,30)
Erdinç ve ark. (2019)	FastText	SVM (%78)

2. Materyal ve Metot

Veri seti Python'ın BeautifulSoup kütüphanesi kullanılarak 6 farklı haber sitesinden derlenmiştir. Her bir kategoride 2000 adet olmak üzere, bilim-teknoloji, eğitim, kültür-sanat, sağlık, siyaset ve spor kategorilerinde toplam 12000 adet haber metni bulunmaktadır. Kategoriler haber sitelerinin belirlediği kategoriler temel alınarak belirlenmiştir.

2.1 Veri Ön İşleme

Veri setinin ham halinde makine öğrenmesi için önemli olmayan birçok karakter/kelime mevcuttur. Daha kesin sonuçlar alabilmek için, metinleri bu karakter/kelimelerden temizlemek gerekir. Bunun için ham veri üzerinde aşağıdaki işlemler uygulanmıştır.

1. Operatörler, noktalama işaretleri gibi gereksiz karakterler ve sayısal ifadeler temizlenmiştir.

2. Tüm kelimeler küçük harfe çevrilerek, birbirinin aynısı olan kelimelerin makine tarafından farklı kelimeler gibi algılanabilmesinin önüne geçilmiştir.

3. Türkçe etkisiz kelimeler (stopwords) silinmiştir.

4. Kelimeler gövdelenerek, yalnızca aldığı ekler farklı olan kelimelerin aynı kelime olarak algılanması hedeflenmiştir.

2.2 Kelime Gömme İşlemi

Veri setinin vektör modelinin elde edilmesi için TF-IDF, Word2Vec ve FastText yöntemlerinden yararlanılmıştır.

Tüm yöntemlerde tüm veri setinde en az 5 kez geçen kelimeler değerlendirmeye alınmıştır.

TF-IDF yöntemi için n-gram aralığı 1-2 olarak belirlenmiştir.

Word2Vec ve FastText yöntemlerinin her ikisinde de en iyi sonuçlar, her kelimenin vektör uzunluğu 256, pencere boyutu 5, epoch sayısı 40 olarak alındığı ve Skip-Gram algoritması kullanıldığı durumda elde edilmiştir.

FastText yönteminde minimum n-gram 2, maksimum n-gram 10 olarak belirlenmiştir. Bunun dışındaki değerler için varsayılan değerler kullanılmıştır.

3. Tartışma ve Sonuç

Bu çalışmada veri setindeki kelimelerin vektör modelinin elde edilmesi için TF-IDF, Word2Vec ve Fasttext yöntemleri ayrı ayrı uygulanmıştır. Ön işlemeden geçirilip vektör modeli çıkarılan veri seti, %70-%30 oranında eğitim ve test dokümanı olarak ayrılmış, elde edilen kelime vektörleri Python'ın Scikit-learn kütüphanesinde SVM, Naive Bayes, Logistic Regression, Random Forest ve ANN yöntemleri ile eğitilip sınıflandırılarak kıyaslanmıştır.

TF-IDF, Word2Vec ve FastText ile vektör modelleri çıkarılan verinin SVM, Naive Bayes, Logistic Regression, Random Forest ve ANN yöntemleri ile eğitiminden elde edilen başarı oranları Tablo 4'de verilmiştir.

Tablo 4. Eğitimlerin başarı oranları

	TF-IDF	Word2Vec	FastText
SVM	% 95,639	% 95,306	% 95,750
Naive Bayes	% 94,444	% 91,750	% 91,694
Logistic Regression	% 95,694	% 95,639	% 95,694
Random Forest	% 90,167	% 92,000	% 91,944
ANN	% 95,583	% 95,667	% 95,528

Kategorilere göre elde edilen başarı oranları Tablo 5'te verilmiştir. TF-IDF için en iyi başarı oranlarını veren logistic resression yönteminin karmaşiklik matrisi Tablo 6'da verilmiştir. Word2Vec için en iyi başarı oranlarını veren ANN yönteminin karmaşiklik matrisi Tablo 7'de verilmiştir. FastText için en doğru başarı oranını veren SVM yönteminin karmaşiklik matrisi ise Tablo 8'de verilmiştir.

Tablo 5. Kategorilere göre başarı oranları

	TF-IDF					Word2Vec					FastText				
	SVM	Naive Bayes	Logistic Regression	Random Forest	ANN	SVM	Naive Bayes	Logistic Regression	Random Forest	ANN	SVM	Naive Bayes	Logistic Regression	Random Forest	ANN
Bilim-Teknoloji	%93,50	%87,50	%93,16	%88,00	%92,33	%92,33	%88,33	%92,50	%91,50	%93,33	%94,00	%88,83	%92,67	%90,00	%93,83
Eğitim	%94,67	%93,17	%94,83	%90,00	%94,83	%96,00	%86,17	%95,50	%92,67	%96,33	%96,00	%85,83	%94,83	%92,00	%96,00
Kültür-Sanat	%96,17	%96,67	%96,00	%89,00	%97,00	%94,33	%96,17	%95,67	%92,67	%94,83	%94,83	%95,17	%95,83	%92,83	%94,33
Sağlık	%95,67	%95,50	%95,67	%88,83	%95,17	%95,00	%90,50	%96,33	%90,67	%96,50	%95,50	%91,50	%96,50	%90,17	%95,33
Siyaset	%95,50	%97,17	%96,33	%88,33	%96,00	%95,00	%96,67	%95,83	%91,17	%94,00	%95,00	%96,50	%96,00	%91,50	%94,67
Spor	%98,33	%96,67	%98,17	%96,33	%98,17	%99,17	%92,67	%98,00	%95,00	%99,00	%99,17	%92,33	%98,33	%95,17	%99,00

Tablo 6. TF-IDF için en iyi başarı oranlarını veren Logistic Resression yönteminin Karmaşıklık Matrisi

	Bilim-Teknoloji	Eğitim	Kültür-Sanat	Sağlık	Siyaset	Spor
Bilim-Teknoloji	559	4	12	18	6	1
Eğitim	14	569	8	2	7	0
Kültür-Sanat	5	5	576	6	6	2
Sağlık	14	5	2	574	4	1
Siyaset	8	4	1	8	578	1
Spor	0	0	5	3	3	589

Tablo 7. Word2Vec için en iyi başarı oranlarını veren ANN yönteminin Karmaşıklık Matrisi

	Bilim-Teknoloji	Eğitim	Kültür-Sanat	Sağlık	Siyaset	Spor
Bilim-Teknoloji	560	11	4	20	3	2
Eğitim	11	578	2	3	6	0
Kültür-Sanat	10	7	569	7	6	1
Sağlık	8	5	4	579	3	1
Siyaset	9	6	7	9	564	5
Spor	0	0	1	4	1	594

Tablo 8. FastText için en doğru başarı oranını veren SVM yönteminin Karmaşıklık Matrisi

	Bilim-Teknoloji	Eğitim	Kültür-Sanat	Sağlık	Siyaset	Spor
Bilim-Teknoloji	564	5	9	15	5	2
Eğitim	10	576	4	4	6	0
Kültür-Sanat	9	7	569	7	6	2
Sağlık	14	6	2	573	4	1
Siyaset	8	7	6	8	570	1
Spor	0	0	3	2	0	595

Çalışmanın sonucunda TF-IDF, Word2Vec ve FastText ile elde edilen vektör modellerinin sınıflandırma başarı oranlarının birbirine çok yakın olduğu, ancak en başarılı oranı %95.75 ile FastText ile elde edilen vektör modelinin verdiği izlenmiştir. Vektör model yöntemlerinin ve tahmin algoritmalarının genelinde en doğru tahmin edilen kategori spor kategorisi olmuştur. Literatürdeki diğer çalışmalarda da rastlanılan bu durum, spor haberlerinde, sıklıkla sadece spor haberlerinde geçen kelimeler kullanıldığını düşündürmektedir.

Bununla birlikte TF-IDF yönteminin Word2Vec ve FastText yöntemine göre bellekte daha yer kapladığı ve güçlü bilgisayarlar gerektirdiği görülmüştür. Bunun sebebi TF-IDF yönteminde, her bir haber metni için oluşturulan vektörün derlemdeki tüm kelime sayısı uzunluğunda olmasıdır. Vektör uzunluğunun azaltılması, yöntemin başarı oranını düşürmüştür. Word2Vec ve Fasttext yöntemlerinde ise haber metninin vektör uzunluğu sınırlandırılmaktadır.

Kaynakça

[1] Vapnik V. The nature of statistical learning theory. Springer, 2nd edition, 2013; New York, USA. pp: 32-40.

[2] Joachims, T. (1999, June). Transductive inference for text classification using support vector machines. In Icml (Vol. 99, pp. 200-209).

[3] Khan, Aurangzeb, et al. "A review of machine learning algorithms for text-documents classification." Journal of advances in information technology 1.1 (2010): 4-20.

[4] Liu, Y., Liu, Z., Chua, T. S., & Sun, M. (2015, February). Topical word embeddings. In Twenty-Ninth AAAI Conference on Artificial Intelligence.

[5] Ramos, J. Using tf-idf to determine word relevance in document queries. In Proceedings of the first instructional conference on machine learning 2003; (Vol. 242, pp. 133-142).

[6] Mikolov T, Chen K, Corrado G, Dean J. (2013), "Efficient estimation of word representations in vector space". Proceedings of Workshop at ICLR. Scottsdale, Arizona 2-4 Mayıs 2013.

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

[8] Osmanoglu, U.O., Atak, O.N., Caglar, K., Kayhan, H. & Can, T.C. (2020). Sentiment Analysis for Distance Education

[9] Course Materials: A Machine Learning Approach. Journal of Educational Technology & Online Learning, 3(1), 31-48.

[10] Sawaf, H., Zaplo, J., & Ney, H. (2001). Statistical classification methods for Arabic news articles. Natural Language Processing in ACL2001, Toulouse, France.

[11] Hakim, A. A., Erwin, A., Eng, K. I., Galinium, M., & Muliady, W. (2014, October). Automated document classification for news article in Bahasa Indonesia based on term frequency inverse document frequency (TF-IDF) approach. In 2014 6th International Conference on Information Technology and Electrical Engineering (ICITEE) (pp. 1-4). IEEE. Doi:10.1109/iciteed.2014.7007894

[12] Dilrukshi, I., De Zoysa, K., & Caldera, A. (2013, April). Twitter news classification using SVM. In 2013 8th International Conference on Computer Science & Education (pp. 287-291). IEEE. Doi:10.1109/iccse.2013.6553926

[13] Amasyali, M. F., & Yildirim, T. (2004, April). Automatic text categorization of news articles. In Proceedings of the IEEE 12th Signal Processing and Communications Applications Conference, 2004. (pp. 224-226). IEEE.

[14] Tüfekci, P., Uzun, E., & Sevinç, B. (2012, April). Text classification of web based news articles by using Turkish grammatical features. In 2012 20th Signal Processing and Communications Applications Conference (SIU) (pp. 1-4). IEEE.

[15] Sen, M. U., & Yanıkoğlu, B. (2018, May). Document classification of SuDer Turkish news corpora. In 2018 26th Signal Processing and Communications Applications Conference (SIU) (pp. 1-4). IEEE.

[16] Acı, Ç. İ., & Çırak, A. Türkçe Haber Metinlerinin Konvolüsyonel Sinir Ağları ve Word2Vec Kullanılarak Sınıflandırılması. Bilişim Teknolojileri Dergisi, 12(3), 219-228.

[17] Erdinç, H. Y., & Güran, A. (2019, April). Semi-supervised Turkish Text Categorization with Word2Vec, Doc2Vec and FastText Algorithms. In 2019 27th Signal Processing and Communications Applications Conference (SIU) (pp. 1-4). IEEE.