Pamukkale Univ Muh Bilim Derg, 24(5), 879-886, 2018

Pamukkale Üniversitesi Mühendislik Bilimleri Dergisi

Pamukkale University Journal of Engineering Sciences

Türkçe için karşılaştırmalı metin sınıflandırma analizi

A comparative analysis of text classification for Turkish language

Savaş YILDIRIM[1*], *Tuğba YILDIZ*[2]

[1,2]Bilgisayar Mühendisliği, Mühendislik ve Doğa Bilimleri Fakültesi, İstanbul Bilgi Üniversitesi, İstanbul, Türkiye.
savasy@bilgi.edu.tr, tugba.dalyan@bilgi.edu.tr

Geliş Tarihi/Received: 14.11.2017, Kabul Tarihi/Accepted: 23.02.2018 doi: 10.5505/pajes.2018.15931
* Yazışılan yazar/Corresponding author Araştırma Makalesi/Research Article

Öz

Metin Sınıflandırma Doğal Dil İşleme (DDİ) alanında önemli bir yere sahiptir. Son zamanlarda metinsel verilerin artması ve otomatik etiketlenmesi gerekliliği, metin sınıflandırma probleminin önemini artırmıştır. Geleneksel yaklaşımlardan öne çıkan kelime torbası yöntemi yıllardır metin sınıflandırmasında başarılı olmaktadır. Son zamanlarda sinir ağları dil modelleri DDİ problemlerine başarılı bir şekilde uygulanmış ve bazı alanlarda büyük başarı kaydetmişlerdir. Yapay Sinir Ağları (YSA) temelli mimarilerin en önemli avantajı daha etkili kelime ve metin gösterilimlerin oluşturmasıdır. Bu gösterilimler, geleneksel yöntemlere göre daha az boyutlu ve daha etkili bulunmuştur. Özellikle anlambilimsel ve sözdizimsel analizlerde başarılı uygulamalar yapılmıştır. Öte yandan daha uzun vektörlerle gösterilim kullanan geleneksel kelime torbası yöntemleri, metin gösterilimleri anlamında hala gücünü korumaktadır. Ancak Türkçe için bu iki yaklaşımın herhangi bir karşılaştırılması yapılmamıştır. Bu çalışmada, geleneksel kelime torbası yaklaşımı ile sinir ağı temelli yeni gösterilim yaklaşımları metin sınıflandırması açısından karşılaştırılmıştır. Bu çalışmalarda gördük ki etkili özellik seçimleri geleneksel yöntemlerinin hala yeni kuşak kelime gömme (word embeddings) yaklaşımı ile yarışacak düzeydedir. Son olarak deneylerimizi bu iki yaklaşım açısından çeşitlendirerek raporladık ve Türkçe için başarılı metin sınıflandırma mimarisini bu raporda ayrıntılı tartıştık.

Anahtar kelimeler: Metin sınıflandırma, Makine öğrenmesi, Yapay sinir ağları

Abstract

Text categorization plays important role in the field of Natural Language Processing. Recently, the rapid growth in the amount of textual data and requirement of automatic annotation makes the problem of text categorization more important. As a prominent one of the traditional methods, the bag-of-words approach has been successfully applied to text categorization problem for years. Recently, Neural Network Language Models (NNLM) have achieved successful results for various problems of Natural Language Processing (NLP). The most important advantage of the NNLM is to provide effective word and document representations. Those representations are lower dimensional and are found to be more effective than traditional methods. They have been exploited successfully for semantic and syntactic analysis. On the other hand, the traditional bag-of-words approaches that use one-hot long vector representation are still considered powerful in terms of their accuracy in document classification. However, comparing these approaches for Turkish language has not been attempted before. In this study, we compared them within a variety of analysis. We observed that the traditional bag-of-word representation utilizing an effective feature selection and a machine learning algorithm aligned with it have comparable performance with new generation vector based methods, namely word embeddings. In this study, we have conducted various experiments comparing these approaches and designated an effective text categorization architecture for Turkish Language.

Keywords: Text classification, Machine learning, Artificial neural network

1 Giriş

Geleneksel yöntemlerde metinlerin gösterilimi, o metinlerde yer alan kelimeler veya kelime öbekleri üzerinden ifade edilir. Bu ifade bir vektöre denk düştüğü için Vektör Uzay Modeli ismini alır [1],[2]. Bu vektörler, sistemlerin başarılarını artırmak için ağırlıklandırılabilir ya da var/yok şeklinde ikili olarak ifade edilir. Burada metin gösterilimi metinlerin karşılaştırılabilmesi için çok büyük ve sabit uzunlukta bir vektörle ifade edilir. Geleneksel kelime torbası yöntemi başarılı bir şekilde birçok DDİ problemine uygulanmasına karşın iki dezavantaja sahiptir. Bu yöntemde kelimelerin sırası göz ardı edilir. İkinci olarak vektör uzunluğu tüm sistemde dikkate alınan sözlük boyutu kadardır. Bir kelimelik metnin bile sözlük boyutunda seyrek bir vektörle ifade edilmesi bu yaklaşımın dezavantajını göstermektedir.

Son zamanlarda yeni uyarlamalarla derin YSA destekli dil modelleri hem dil mimarilerinin zaman karmaşıklık seviyelerinde önemli iyileştirmeler yapabildiler hem de DDİ problemlerinde önemli başarılar kaydettiler. DDİ alanındaki YSA temelli en çok dikkat çeken yaklaşım kelime gömme (word embedding) yaklaşımıdır [3],[4]. Bu yaklaşımda kelime gösterilimleri SGD (Stochastic Gradient Descent) yardımıyla ve geriye yayılım ile öğrenilir ve gösterilimler hem kısa hem yoğun olur. Böylelikle vektörlerdeki seyreklik ve veri tablolarındaki yüksek boyutluluk sorunu giderilmiş olur.

Öte yandan bu kelime gösterilimi DDİ alanındaki birçok probleme katkı sağlamış ve başarılar elde etmiştir. Kelime gösterilimleri kelimeler arasındaki ilişkilere yönelik analizler için uygun olmakla birlikte bu kelime vektörlerin ortalaması sayesinde metin gösterilimleri için de faydalı olmaktadır. Öte yandan metin gösterilimini elde etmek için kelime gömme gösterilimlerine benzer bir başka yaklaşım da mevcuttur ve bu yaklaşım bu çalışma kapsamında ele alınmıştır [5].

En popüler kelime gömme mimarisi word2vec modelidir. Burada kullanılan mimaride sadece bir katman bulunduğundan word2vec modeli derin öğrenme algoritması olarak kabul edilmez. O yüzden bu makalede bu modeli YSA modeli olarak ifade edeceğiz. Model eğitim sırasında kelime vektörleri oluşturulurken, verilen bir kelimenin vektörü bağlamındaki diğer kelimelere bakılarak tahmin edilmeye çalışılır. YSA destekli bu tahminleme sırasında kelimelerin gösterilimleri eğitilir. Amaç kelime tahmininden çok bu yan çıktı olan kelime gömme gösterilimleridir. Benzer bir yaklaşımla metin

gösterilimleri elde edilir. Her paragrafın içine bir tane de paragraf vektörü eklenir ve bu vektör diğer kelime vektörleri ile birlikte eğitilir. Burada kelime gömme vektörleri metinler arasında paylaşılırken metin gömme vektörleri sadece ilgili metne ait olduğundan paylaşılmaz. Ancak o metin içinde oluşturulan her pencere bağlamında bulunur.

Türkçe metin sınıflandırması için özgün birçok çalışma olmasına karşın, YSA temelli kelime ve metin vektörlerini kullanan bir çalışma şu ana kadar yapılmamıştır. Bu çalışmada hem YSA destekli modeller hem de geleneksel mimariler tasarlandı. Bu mimariler iki farklı veri kümesi kullanılarak ayrıntılı bir şekilde karşılaştırıldı.

2 Literatür

Literatürde Türkçe için yapılan metin sınıflandırma yöntemlerini üç ana çerçevede ele alabiliriz. Birinci grupta geleneksel yöntemleri irdeleyen çalışmalar yer almaktadır. Birçok çalışma dünya dillerinde kabul görmüş yaklaşımları Türkçe dili için başarılı bir şekilde uygulamıştır. İkinci gruptaki çalışmalar genellikle ön işleme süreçlerinin katkısını irdelemişlerdir. Son gruptaki çalışmalar ise geleneksel yöntemlerin dışındaki alternatif yaklaşımları ele almışlardır.

İlk gruptaki çalışmalarda Türkçe dili için birçok metin sınıflandırma tasarımları yapılmıştır. Bu çalışmalar metnin konusunu bulma, yazarını tespit etme, yazarın cinsiyetini bulma gibi sınıflandırma problemlerini incelemişlerdir. İlk çalışmalardan birinde n-gram ve kelime torbası yaklaşımıyla üç metin sınıflandırma problemi ele alınmıştır [6]. Metnin yazarını tanıma, türünü belirleme ve cinsiyet tespit etme problemleri için oluşturulan modellerde dört farklı makine öğrenme algoritması kullanılmış ve başarılı sonuçlar elde edilmiştir. İlk çalışmalardan biri olması açısından önemlidir. Türkçe için ilk çalışmalar arasında sayılabilecek bir çalışmada ise sadece yazar tespiti için 18 farklı yazardan alınmış 35 metin üzerinden bir sınıflandırma yapılmıştır [7]. Bu çalışmada n-gram ve kelime torbası yaklaşımları ile 10 farklı özellik bilgisi kullanılarak, 5 farklı sınıflandırma algoritması denenmiş ve sonuç olarak n-gram temsil yöntemi ile diğer özellikler birlikte kullanıldığında başarılı sonuçlar elde edildiği vurgulanmıştır. En iyi algoritma olarak YSA ve Destek Karar Makineleri (DKM) olduğu belirtilmiştir. Geleneksel yöntemlerin araştırıldığı bu alandaki en kapsamlı sayılabilecek çalışma [8] tarafından yapılmıştır. Bu çalışma metinlerin temsil edilme biçimlerinin sınıflandırma problemine katkısını 6 farklı metin sınıflandırma problemi açısından ele alan çok kapsamlı bir araştırmadır. Yazıdan ruh hali tanıma probleminde 4 farklı ruh hali, film yorumlarındaki yönelim tespitinde 3 farklı yönelim, köşe yazarlarını tanımada 18 yazar, cinsiyet tahmininde 2 sınıf, haberlerde 5 farklı kategori ve şiirin yazarını bulmada 7 farklı şair sınıflandırma problemi olarak ifade edilmiştir. Bu yaklaşımda kelime torbası metin temsil yöntemi, birlikte geçme matrisi ve saklı anlam indeksleme gibi metin kelime matrislerine dayalı geleneksel modeller kapsamlı bir şekilde ele alınmış ve başarılı sonuçlar elde edilmiştir.

Bu alanda yapılan son çalışmalardan birinde metin sınıflandırması için TTC-3600 isimli altı sınıflı örnekten oluşan Türkçe bir veri seti oluşturulmuş ve akademik amaçlı paylaşılmıştır [9]. Aynı zamanda bu çalışmada araştırmacılar kendi geliştirmiş oldukları modeli başarılı bir şekilde veri setine uygulamışlardır. Metin gösterilimi için kelime torbası, n-gram modeli ve özellik seçimi modellerini kullanmışlardır. Sınıflandırma algoritması olarak 6 farklı makine öğrenmesi algoritması uygulamışlar ve özellik seçme yöntemleri kullanarak sonuçlarda iyileştirme yapmışlardır. Zemberek kütüphanesi [22] yardımıyla kelime köklerini kullanarak, ARFS (Attribute ranking-based Feature Selection) isimli özellik seçimi ile elde edilmiş metin gösterilimlerini 6 farklı algoritma ile sınamışlardır. ARFS yöntemi özelliklerin sınıfları ayırmadaki gücünü ölçen Bilgi Kazancı gibi metrikleri kullanır. Sonuç olarak Zemberek ve ARFS ile uygulanan RO (Rastgele Orman/Random Forest) sınıflandırıcısının en uygun düzen olduğunu vurgulamışlardır.

Bazı çalışmalar özellikle önişleme süreçlerine ilişkin denemeler yapmışlardır. Bir çalışmada kelime kökleri kullanılarak Zemberek kütüphanesinin önerdiği en uzun kelime kökü ve en kısa kelime köklerinin sınıflandırmaya katkısı incelenmiş ve en uzun köklerin daha yüksek başarı sağladığı belirlenmiştir [10]. Yaklaşım açısından diğer çalışmalardan farklı bir yöntem izleyen bu çalışmada aynı zamanda kelime derlem frekansları üzerinde boyut azaltma yöntemi uygulanmıştır. Ele alınan tüm metin temsil yöntemleri Naive Bayes (NB) ve RO gibi başarılı algoritmalarla kıyaslanmıştır.

Bir başka önemli çalışma kök bulma yaklaşımının metin sınıflandırma problemi etkisini yine Zemberek Kütüphanesi [22] kullanarak ölçmüştür [11]. Ancak az sayıda sınıf ve az sayıda örnek kullanılmıştır. Yapılan deneylerde kelime köküne kelime tipini de ekleyerek model başarısı iyileştirilmeye çalışılmış ancak kelime tipinin çok büyük etkisinin olmadığı not edilmiştir. Öte yandan kelime kökü bulma algoritmasına alternatif olarak kelimenin ilk K karakterini alma yaklaşımı denenmiş ve sınıflandırma problemi açısından kök bulma algoritmalarından daha başarılı olduğu kaydedilmiştir. Burada K için en uygun değerin 5 olduğu tespit edilirken DKM ve NB gibi sınıflandırıcı algoritmalar denenmiştir.

Ön işleme açısından en kapsamlı ve önemli çalışmalardan biri [12] tarafından yapılmıştır. Bu çalışma özellikle önişleme sürecinin metin gösterilimi ve metin sınıflandırmasına katkısını irdelemiştir. Kök bulma, etkisiz kelime (stopword) elemesi ve kelime özellik ağırlıklandırma aşamaları çeşitli Türkçe veriler üzerinde denenmiştir. Zemberek kütüphanesi ve FPS7 (Fixed Prefix Stemming) gibi yöntemler kullanılarak elde edilen kelime kökünün Bilgi Kazanımı problemlerinde faydalı olduğu ancak metin sınıflandırması için herhangi bir katkısının olmadığı vurgulanmıştır. Bir başka çalışma ön işleme süreçlerinin sınıflandırma başarısına katkısını, uygulama alanı, doğal dil ve boyut indirgeme üzerinden ölçmüştür [13]. İki farklı dil ve çeşitli uygulama alanları üzerinde yaptıkları birçok denemede sistem tasarımında kullanılacak aşamaların uygun bir mimariyle bir araya getirilmesine dikkat çekmişlerdir. Böylelikle önişleme veya sınıflandırma aşamasındaki uyumsuz tasarımların başarısız sonuçlar doğurabileceği vurgulanmıştır. Önişleme çalışması, özellik çıkarımı, özellik seçimi ve sınıflandırma aşamalarının bu çerçevede ele alınmasını vurgulayan bir çalışma olmuştur. Yine kök kullanımın metin sınıflandırmasına katkısını irdeleyen bir başka çalışmada NB ve Bilgi Kazancı özellik seçimi yöntemleri kullanılmış ve kök kullanmanın metin sınıflandırmasına katkısının sınırlı olduğu vurgulanmıştır [14].

Geleneksel çalışmaların dışında alternatif yöntemleri araştıran birçok çalışma bulunmaktadır. Bu çalışmaların birinde Saklı Dirichlet Analiz/Latent Dircihlet Analysis (SDA) ve Saklı Anlamsal İndeksleme/Latent Semantic Indexing (SAİ) gibi boyut indirgeme yöntemleri ele alınmıştır [15]. Veri kümesi olarak farklı alanlarda Türkçe bildiri özetlerinden oluşturulan etiketli bir derlem kullanılmıştır. Derlemde 18 ve 34 sınıftan

oluşan veri kümeleriyle SDA konu modellemesi metin temsili için kullanılmıştır. DKM, NB ve RO algoritmaları ile kök bulma yönteminin algoritma başarılarını artırdığı belirtilmiştir. Ancak görebildiğimiz kadarıyla çalışmadaki sınıflandırıcıların performansları diğer çalışmalara göre düşük kalmıştır.

Farklı bir araştırma kelime anlamsal analizini metin sınıflandırması için kullanmıştır [16]. Bu çalışma kelimeler arası anlamsal ilişkileri kullanarak yeni bir metin gösterilimi yaratan ve metin sınıflandırma problemini bu anlamsal uzay üzerinde çözen özgün bir yaklaşımdır. Anlamsal uzayda temsil edilen kelimelerin birbirlerine olan yakınlıkları Öklid gibi uzaklık/yakınlık ölçme yöntemleriyle hesaplanarak kelime kümeleme işlemi gerçekleştirilmiştir. Böylelikle metinler daha az boyutla ifade edilebilir olmuştur. Çalışma 5 farklı kategoriyi içeren veri kümesinde, 100 boyutta Logistic Regresyon kullanarak %92.5 oranında bir başarı elde edebilmiştir.

Son olarak bir başka yaklaşım toplu öğrenme algoritmalarını metin sınıflandırması için kullanmıştır [17]. Toplu öğrenme yöntemi makine öğrenme alanında ilgi çeken konulardandır. Tek dezavantajı aşırı öğrenmeye yatkın olmasıdır. Bu çalışmada 6 sınıflı iki farklı Türkçe veri kümesi üzerinde geleneksel makine öğrenme yöntemleriyle toplu öğrenme temelli yöntemler karşılaştırılmış ve başarılı uygulamalar yapılmıştır. Çalışmada zamansal karmaşıklığı düşürmek için budama yöntemi uygulanmış ve başarından ödün vermeden oluşan karar ağacı ormanında %90.0'lık bir budama sağlanmıştır.

3 Yöntem

3.1 Geleneksel kelime torbası yaklaşımı

Geleneksel yöntemlerde kelime torbası yaklaşımı kelime ve metin için Out-of-Vocabulary (OOV) sözlük boyutunda uzun seyrek vektörler oluşturur. Buradaki temel sorun işte bu yüksek boyutluluktur. Bu sebepten dolayı bazı yaklaşımlar etkili bir şekilde bu vektörler üzerine uygulanamamaktadır. Bu yüksek boyutluluk algoritmaların zamansal karmaşıklığını etkilemekte ve hız gerektiren gerçek zamanlı projelerde uyumsuzluk oluşturmaktadır. Bu sebepten ötürü kabul gören bir çözüm kelime/özellik seçme aşamasıdır. Bu aşamada bilgi kazancı düşük özellikler (kelimeler) sistemden elenir. Bazı çalışmalar [6],[18], özellikle sık kullanılan kelimelerin diğerlerine oranla daha etkili olduğunu belirtmişlerdir. [7],[8],[14],[19],[20] referanslı çalışmalar ise Bilgi Kazancı (Information Gain), ki-kare $\chi 2$ (chi-squared) gibi seçme kriterlerinin bilgi verici özelliklerinin bulunmasından dolayı daha faydalı olduğunu belirtmişlerdir. Saklı Anlam İndeksleme (Latent Semantic Indexing/LSI) ve Saklı Dirichlet Analizi (Latent Dirichlet Analysis) gibi başka çözümler de özellikle boyut indirgeme için uygulanmaktadır [9],[21].

Kelime torbası yaklaşımı kelime seçimi sayesinde boyutu indirgenmiş vektörler üzerinde başarılı çalışabilmektedir. Birçok makine öğrenme algoritması bu indirgenmiş veri kümelerinde başarılı olabilmektedir. Bazı çalışmalar [6],[8],[18],[20] özellikle karar ağaçları için başarılı olan Bilgi Kazancı (BK) yaklaşımının metin sınıflandırmalardaki kelime seçimleri için de etkili olacağını belirtmişlerdir. Bir diğer önemli seçim yöntemi ise ki-kare yöntemidir. İki yöntem de kelimeler ve kategoriler arasında oluşan çapraz tabloyu kullanarak kelimelerin önemlerini ölçmektedir. Ayrıca PMI (Pointwise Mutual Information) ve DICE gibi seçim metrikleri de kullanılmaktadır. Esasında tüm bu seçim metrikleri ikili çapraz tablodan farklı formüller aracılığı ile bir sıralama oluşturur.

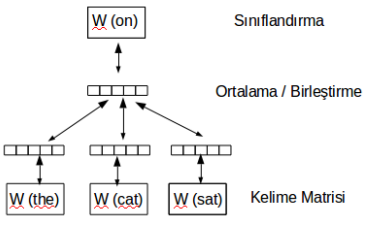
Kelime seçiminden sonra oluşan vektörlerin ağırlıklandırılması bir başka ön aşamadır. Kelimenin metin içindeki tekrarlanma sayısı, terim frekansı(tf), ile tüm derlemdeki tekrarlanma sayısı, metin frekansı (mf), kullanılabilir. Bu yönde çeşitli terim ağırlıklandırma (term weighting) yöntemleri mevcuttur: ikili (binary), kelimenin verilen metinde geçme sıklığı (terim frekansı), logaritmik sıklık (logarithmic tf function) ve verilen metinde geçme sıklığı ile tüm metinlerdeki geçme sıklığına (metin frekansı) bağlı fonksiyon (tf-idf). Buradaki idf (inverse document frequency) evrik metin frekansı anlamına gelir.

Boyut azaltmada bir diğer yöntem etkisiz kelime (stopword) elemesidir. *Ben, ve, şu, orada* gibi fonksiyonel kelimeler sözdizim açısından anlamlı ancak konu tespiti açısından etkisizdir. Hemen hemen her metinde benzer dağılıma sahiptirler. Konu bakımından bir ayırt ediciliğe sahip olmadıklarından elenebilirler. Yapılan çalışmalarda Türkçe için yaklaşık 250 civarında bir liste kullanılmaktadır. Türkçe için bir diğer önemli boyut azaltma yaklaşımı ise kelime kökünün bulunmasıdır. Türkçe sondan eklemeli dil olması sebebiyle kelime çeşitliliği sayısı diğer dillere oranla bir hayli yüksektir. Kelimenin yüzey formu yerine kökünün kullanılması hem başarıyı artırabilir hem boyutu kısaltabilir. Bu yönde çeşitli yaklaşımlar ve sonuçlar ilgili çalışmalarda tartışılmıştır.

Bu çalışma kapsamında tüm bu ön işleme yöntemleri metin sınıflandırma kapasiteleri açısından iki yaklaşım altında denenmiştir. Ortaya çıkan metin vektörleri sınıflandırma problemi açısından ele alınmıştır. Bu aşamada kullanılan sınıflandırıcılar Destek Karar Makineleri (DKM), Logistic Regression, en yakın komşu (EYK), Karar Ağacı (KA) ve çok-katlı Naive Bayes (multi-nominal NB) algoritmalarıdır. Ortaya çıkan başarı sonuçları YSA temelli dil modelleri ile kıyaslanarak raporlanmıştır.

3.2 Yapay sinir ağları temelli dil modelleri

YSA mimarisi temelli kelime gömme gösterilimleri birçok DDİ probleminde başarıyla uygulanmış ve büyük ilgi toplamıştır. Bu mimaride, başlangıç olarak her kelime belirli bir boyutta rastgele sayısal değerlere sahip bir vektör ile ifade edilir. Tek gizli katmanlı YSA verilen derlemdeki metinleri işleyerek SGD algoritması ve geriye yayılım yöntemi ile bu vektörleri sürekli günceller. Burada amaç verilen bir kelime dizisinin en sonunda yer alan kelimeyi tahmin etmektir. Şekil 1'de de gösterildiği gibi *"the", "cat" ve "sat"* kelimeleri verilmiştir ve amaç sonra gelen "on" kelimesini tahmin etmektir. Aslında bu sınıflandırma mimarisindeki en önemli çıktı derlemdeki kelimelerin kısa ve yoğun vektörlerin üretimidir. Model son kelimeyi tahmin ederken sözlükteki tüm kelimelerden her birinin olasılığını göz önünde bulundurur. Bu tıpkı çoklu-sınıf sınıflandırma problemine benzemektedir. Bu mimaride özellikle yumuşak maksimum fonksiyonu (softmax function) model eğitim süresini çok kısalttığından mimariyi de etkili kılmıştır. Sonuç olarak YSA temelli bu mimari verilen derlemde görülen her bir kelime için sabit boyutta yoğun ve kısa vektörler oluşturur.



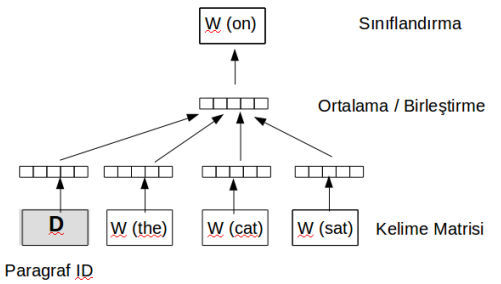
Şekil 1: Word2Vec Model Mimarisi(CBoW).

Son zamanlarda YSA temelli iki kelime gösterilim modeli başarılı bir şekilde birçok DDİ problemine uygulanmıştır. İlk olarak [3] referanslı *word2vec* isimli çalışma iki farklı eğitim mimarisinin etkili bir şekilde vektör oluşturacağını gösterdi. Bu mimariler Continuous Bag-of-Words (CBoW) ve Skip-Gram (SG) mimarileridir. Bu yöntemlerde hesaplama karmaşıklığı azaltılıp başarı oranı artırılmıştır. İkinci çalışma olan *glove* isimli model kelimelerin tüm derlemdeki kullanım sıklığını da hesaba katmaktadır. Çünkü ilk yaklaşım sadece kelimelerin ilgili bağlamlarını ele almakta ve tüm derlemdeki sıklıkları göz ardı etmekteydi. Bu iki yöntemden elde edilen vektörlerin kelimelerin arasındaki anlamsal ve sözdizimsel ilişkileri etkili bir şekilde çözdüğü gösterilmiştir.

Bu aşamadaki temel problem metin gösterilimlerinin nasıl yapılacağıdır. Kelime gösterilimleri üzerinden metin gösterilimlerini oluşturan bir yöntem o metin içerisinde geçen kelimelerin vektörel ortalamasının alınmasıdır. Öte yandan [5] referanslı çalışma metin vektörlerinin elde edilmesi için paragraf vektör isimli etkili bir yöntem önermiştir, *doc2vec*. Bu yöntem word2vec mimarisinden esinlenmiştir.

Paragraf vektörü mimarisinde öncelikle derlemde görülen tüm kelimelere önceden belirlenmiş sabit uzunlukta rastgele içeriğe sahip bir vektör atanır. Paragraflar için de aynı boyda bir vektör rastgele oluşturulur. Belirli uzunlukta bir pencere verilen bir paragrafı baştan sona tarar. Kelime vektörleriyle birlikte paragraf vektörleri de YSA temelli eğitim sürecine tabii olur ve bu pencereye her seferinde eklenir.

Şekil 2'de gösterildiği gibi son kelime olan "on" kelimesinin vektörü önceki kelimelerin vektörleri ve paragraf vektörü ile tahmin edilir [5]. Bu mimaride özgün olan husus her pencereye bir paragraf vektörünün eklenmesidir. Şekilde görüldüğü gibi paragraf-id her döngüde her pencerenin başına eklenir. Pencerenin sonundaki kelimenin vektörü diğer kelime ve paragraf vektörleri ile tahmin edilirken ortaya çıkan hata (bias) her aşamada SGD algoritması ve geriye yayılım yöntemi ile azaltılır. Dolayısıyla YSA'nın veriye yakınsamasını sağlayarak kelime vektörlerini oluşturur. Buradaki hata tahmin edilen vektör ve edilmesi gereken gerçek vektör arasında ortaya çıkan fark aracılığı ile hesaplanır. Tahmin edilen vektör en sondaki kelimenin vektörüdür. Yakınsamanın daha etkili olması açısından bir pencere için yapılan işlem birden fazla kere (epoch) tekrarlanır. Öte yandan tekrarlar bittiğinde pencere sağa doğru kayar ve yeni pencereye aynı süreç uygulanır. Bu aşamada diğer kelimelerin vektörleri birleştirilir veya ortalamaları alınarak tahmin edilen kelime vektörü ile aralarında bir öğrenme süreci oluşturulur. En sonunda derlemdeki her kelime ve her paragraf için vektörler oluşturulur. Paragraf vektörü sürekli dağıtıldığı için bu yönteme Distributed Memory Model of Paragraph Vectors (PV-DM) denir.



Şekil 2: Paragraf vektör mimarisi: PV-DM (Distributed Memory Model of Paragraph Vectors).

Aynı çalışmanın sunduğu bir başka alternatif yöntem sıralı kelime pencere yaklaşımını göz ardı ederek belirli sayıda rastgele seçilmiş kelimeler ve rastgele seçilmiş hedef bir kelime kullanır. Benzer bir şekilde hedef kelimenin vektörü diğer kelimelerin vektörleri üzerinden tahmin edilir. Tek fark PV-DM de kelimelerin sırası önemliyken burada önemsizdir. Kelime torbası yöntemine benzerliği sebebiyle bu yönteme Distributed Bag of Words-Paragraph Vector (PV-DBOW) denir. Çünkü burada kelimeler sıraları kaybolmuş bir şekilde torbaya atılır. Paragraf vektörü ise her seferinde bu torbaya atılır. Ayrıca word2vec modelindeki atla-gram (skip-gram) öğrenme mimarisine karşılık gelmektedir. Tüm kelime ve paragraf öğrenme süreçleri PV-DM ile aynıdır. Hem PV-DM mimarisinde hem de PV-DBoW mimarisinde paragraf vektörü sadece ilgili paragrafta eğitilirken kelime vektörleri metinler arası niteliğe sahip olduğundan bir diğer metne geçildiğinde aynı kelime vektörleri eğitime kaldığı yerden devam eder ve metinler arası paylaşılır.

4 Deney düzeni

Yukarıda açıklanan yaklaşımlar Kemik DDİ Grubu tarafından paylaşılan veri kümesi ve çalışma [9] tarafından paylaşılan TTC-3600 isimli veri kümesi ile sınanmıştır. Kemik grubundan elde edilen veride ekonomi, politika, spor gibi 7 farklı kategori bulunmaktadır. Bu yazı içinde ayırabilmek için bu veriye T-4900 ismini vereceğiz. Diğer veri kümesi olan TTC-3600 kümesinde 6 farklı kategori bulunmaktadır. Bu iki veri kümesi de İnternet ortamında araştırmacılar için paylaşılmıştır. İlk küme T-4900'de her bir kategori altında 700 adet metin bulunurken, ikinci veride 600 adet metin bulunmaktadır. Buradaki metinlerde Türkiye Cumhuriyeti gibi isim öbekleri bulunup etiketlenmiştir. Biçimbilimsel analiz, etkisiz kelime eleme gibi önişleme süreçleri metin sınıflandırma açısından değerlendirilmiştir.

İki derlemde de toplam kelime çeşitliliği sayısı yaklaşık 100 bin civarındadır. Bu demek oluyor ki herhangi bir eleme işlemi yapılmadığı durumda geleneksel gösterilim için ilk derlem 4900 metin x 100 bin boyutunda bir matrise, ikinci derlem 3600x100 bin boyutunda bir matrise denk gelmektedir. Bu yüzden özellik seçimi metriklerinin kullanımı çok önemlidir. Bilgi verici kelimelerin seçilmesi sırasında ayrı bir veri kümesi bu keşif için kullanılır. Öte yandan denetimli modelin eğitimi ve sınaması için iki ayrı veri kümesi daha kullanılır. Böylelikle veriyi üç kümeye ayırmış oluruz. Çünkü tüm aşamalarda aynı verinin kullanılması yapay başarı sonuçları doğurur ve yanıltıcı olur. Çalışmamızda birinci aşamada elde edilen kelimeler üzerinden metinler ifade edilmiş ve ardından 10 katlı çapraz doğrulama yöntemi ile denetimli algoritmaların başarıları hesaplanmıştır.

Özellik seçimi yöntemlerinin değerlendirilmesine ilişkin makine öğrenmesi algoritmalarının bu seçimler ile sağladığı başarılar dikkate alındı. Bu aşamada Multi-Nominal Naive Bayes (M-NB) makine öğrenmesi algoritması metin sınıflandırması için uygun görülmüştür. Çünkü yapılan son çalışmalar bu algoritmanın katlı terimli (multi-nominal) dağılımını dikkate aldığından diğer makine öğrenmesi algoritmalarından daha başarılı olduğunu kaydetmiştir [8],[20].

Bu çalışmada kelime ve metin vektörlerinin eğitilmesi ve karşılaştırılması için aynı veri setleri kullanılmıştır. Çalışmalar sırasında Python programlama dili ve onun öne çıkan gensim, nltk, sklearn, libsvm gibi yardımcı kütüphaneleri kullanılmıştır. Öte yandan R ve Weka gibi yazılımlar da çeşitli testler,

görselleştirmeler ve analizler için kullanılmıştır. Kelime ve metin vektörleri için tercih ettiğimiz boyut deney gözlemlerimiz sonucunda 300 olarak belirlendi. Paragraf Vektör yaklaşımının PV-DM ve PV-DBOW mimarilerinin ikisi de test edildi. Metin ve kelime gösterilimlerinin ardından Logistic Regression, DKM gibi lineer sınıflandırma yapabilen ve nümerik değişkenlerle uyumlu çalışabilen algoritmalar sınıflandırma açısından değerlendirilip ve raporlanmıştır.

5 Sonuçlar ve tartışmalar

T-4900 ve TTC-3600 veri setlerine geleneksel ve YSA temelli yaklaşımlar uygulandı. İlk olarak daha büyük veri kümesi olan T-4900 üzerinde denemeler yapıldı. Geleneksel yaklaşımın T-4900 kümesi üzerinden elde ettiği sonuçlar Tablo 1'de görülmektedir. Özellik seçimi metotlarının her bir makine öğrenmesi algoritmasıyla sağladığı F1 başarı değerleri bu tabloda ifade edilmiştir. Tabloda açık bir şekilde görüldüğü üzere BK ve ki-kare özellik seçimleri diğer özellik seçimlerinden belirgin bir şekilde daha başarılıdır. Makine öğrenmesi algoritmalarından en başarılısı ise katlı terimli-NB algoritmasıdır. Bu algoritma BK özellik seçimi ile birlikte çalıştığında 2K ve 4K boyutundaki özellik seçimi için sırasıyla 89.0 ve 90.0 F1 değeri elde etmiştir. Bir diğer gözlem ise kelime boyutu arttıkça algoritma başarılarının artmasıdır. Bir aşamada tepe noktaya gelmiştir. Gözlemlerimizde bu tepe noktanın 2K-4K arasında olduğu belirlenmiştir. Geleneksel metin gösteriliminde ve diğer Bilgi Getirimi (Information Retrieval) problemlerinde kullanılan bir diğer önemli teknik özellik ağırlıklandırmadır. Ancak deneylerimizde özellik ağırlıklandırmaya ilişkin istatistiksel anlamda bir farklılık gözlemlenmemiştir.

Tablo 1: Makine öğrenmesi algoritmaları ile özellik seçimi metriklerinin performans tablosu (F1).

Boyut:2K	NB	DKM	EYK	KA	Ort.
BK	89.0	80.0	56.3	72.0	74.1
χ2	89.0	80.0	54.2	72.1	74.0
Dice	87.1	78.1	47.3	70.2	69.9
Pmi	78.9	74.1	61.2	39.0	643
Ort.	85.2	77.0	55.8	59.0	69.2
Boyut:4K	NB	DKM	EYK	KA	Ort.
BK	90.0*	82.2	45.3	72.2	72.1
χ2	90.0*	82.1	44.3	71.1	72.0
Dice	88.3	81.1	42.2	68.9	70.3
Pmi	89.1	82.0	53.9	73.0	75.3
Ort.	89.1	82.0	48.1	72.0	73.0

Yaklaşımların birbirleriyle karşılaştırılması için ya aynı verilerin kullanılması ya da kullanılan verilerin ve koşulların aşağı yukarı aynı tutulması gerekmektedir. Metin sınıflandırılmasına ilişkin bazı çalışmalar karşılaştırma açısından ele alınabilir [6],[10]. Bu çalışmalarda geleneksel kelime torbası yöntemi uygulanmıştır. Çalışma [6] 3 farklı sınıf ve her biri için 200 örnek kullanarak %93.0 başarı elde etmiştir. Diğer çalışma [10] 6 kategori ve her bir kategori için 100 örnek seçmiş ve %91.7 başarı elde etmiştir. Karşılaştırma açısından kendi veri kümelerimizi ve kategori çeşitliliğimizi bu boyutlara ayarladığımızda sırasıyla %98.1 ve %90.8 başarı oranlarını gözlemledik. Dolayısıyla bulgularımız göstermiştir ki özellik seçimi metin sınıflandırma açısından önemli bir aşamadır.

YSA temelli düzeneğimizde 4 farklı yaklaşım metin sınıflandırması açısından T-4900 verisi ile değerlendirilmiştir.

- PV-DM: Paragraf Vektörü,
- PV-DBOW: Kelime Torbası temelli paragraf vektörü,
- PV-DM + PV-DBOW: İki Vektörün Birleştirilmesi,
- Kelime Gömme Vektörleri Ortalamaları.

Buradaki 3. yaklaşımda ilk iki yaklaşımda elde edilen vektörler yan yana konarak birleştirilir. Aynı metnin iki farklı mimariden elde edilen vektörleri kullanılır. Son yaklaşımda metin içinde geçen kelimelerin gömme vektörlerinin ortalaması alınır. Bu ortalama metnin vektörel ifadesini oluşturur.

Bu dört yaklaşımla elde edilen metin gösterilimleri nümerik içeriğe sahip olduğundan bu veri tipi ile verimli çalışan makine öğrenme algoritmaları seçilmiştir: Logistic Regression, SGD Sınıflandırıcı, DKM ve YSA. Elde edilen sonuçlar Tablo 2'de gösterilmiştir.

Tablo 2: YSA temelli gösterilimlerin sınıflandırıcı algoritmalar ile sağladığı performans.

F1	Log.Reg.	SGD	DKM	YSA	Ort.
Pv-DM+DBoW	89.1*	83.9	88.0	89.0*	88.1
Pv-DBoW	87.0	85.1	88.0	85.8	86.2
Pv-DM	89.3*	87.2	87.1	87.1	88.1
Vec. Ort	81.0	78.9	80.2	81.1	80.2
Ort.	87.2	84.0	86.0	86.0	85.8

Tablo 2 göstermiştir ki PV-DM mimarisi ile elde edilen vektörler PV-DBOW ile elde edilenlerden daha başarılıdır. İki mimariden elde edilen vektörlerin birleştirilmesi PV-DM'in başarısının ötesine geçememiştir. Aralarında istatistiksel anlamda bir fark tespit edilmemiştir. Kelime vektörleri ortalaması ise en düşük performans gösteren yaklaşım olmuştur. Bu anlamda paragraf vektörlerinin bu mimarilerle oluşturulmasının etkili ve anlamlı olduğu görülmüştür.

İkinci aşamada yukarıda tasarlanan mimariler TTC-3600 veri kümesine uygulanarak değerlendirildi. Geleneksel yaklaşımla tasarladığımız modelde Bilgi Kazancı özellik seçimi aracılığı ile 2K ve 4K özellik seçilmiştir. Bu seçime ilişkin sınıflandırma algoritmalarının sonuçları Tablo 3'te gösterilmektedir.

Tablo 3: Geleneksel kelime torbası yönteminin ttc 3600 verisi ile elde ettiği sonuçlar.

Boyut	NB	DKM	EYK	KA
2K	92.6	91.8	55.1	79.6
4K	93.1	92.1	53.0	80.1

Daha önceki deneylerimize benzer bir şekilde ve beklediğimiz üzere en iyi sonuçlar m-NB sınıflandırıcısı ile elde edilmiştir. Yapılan denemelerde bu sınıflandırıcı ile 2K ve 4K için sırasıyla 92.6 ve 93.1 F1 başarı skorları kaydedildi. Öte yandan TTC 3600 verisi ile elde ettiğimiz sonuçların diğer T-4900 ile elde edilen sonuçlardan daha iyi olduğu gözlemlenmiştir. Bunun en önemli sebebi sınıf sayısının daha az olmasıdır. Sınıf sayısını azalttıkça algoritmaların başarı oranlarının artması hep gözlediğimiz bir olgudur. Çünkü bu durum karmaşıklığı azaltmaktadır. Örnek olarak üç sınıf seçtiğimizde başarı 98.1 F1 olmaktadır. Tabii bu durum seçilen sınıfların kaba tanecikli (coarse-grain) ya da ince tanecikli (fine-grain) olup olmamasıyla da ilgilidir. Bu veri üzerinde diğer sınıflandırıcılardan DKM sınıflandırıcısı başarılı skorlar kaydetse de m-NB'den daha düşük değerler elde etmiştir. EYK ve KA sınıflandırıcıları bu veride de yine çok başarısız sonuçlar elde etmiştir. Öte yandan ki-kare özellik seçimi başarılı bir seçici olsa da çok az farkla Bilgi Kazancının gerisinde kalmaktadır.

YSA temelli yaklaşımların bu veri üzerindeki başarı sonuçları ise Tablo 4'te verilmiştir. Bu tablodan da görüldüğü üzere yine PV-DM mimarisi PV-DBOW mimarisine göre daha iyi sonuçlar elde etmiştir. İki mimariden elde edilen vektörlerin birleştirilmesi yaklaşımı yine herhangi bir ilerleme yaratmamıştır. Bu tabloda görüldüğü üzere en iyi model Logistic Regression sınıflandırıcının PV-DM gösterilimi ile kullanılması sonucunda elde edilmiştir. Benzer bir şekilde Logistic Regression ve DKM sonuçları birbirlerine çok yakın çıkmıştır ve aralarındaki farkın istatistiksel anlamda bir önemi yoktur. Bir diğer husus başarı oranının 7 sınıflı verimize göre daha iyi çıkmasıdır. Bu bulgu geleneksel yöntemlerde de elde edilmişti.

Tablo 4: YSA gösterilimlerin TTC-3600 kümesindeki başarısı.

F1	Log.Reg.	SGD	DKM	YSA	Ort.
Pv-DM+DBoW	92.0	89.0	92.1	90.5	91.0
Pv-DBoW	90.5	89.1	92.3	90.5	90.5
Pv-DM	**92.3**	90.2	92.1	91.0	91.4
Ort.	91.5	89.1	92.0	90.7	91.0

Bu veriyi [9] referanslı çalışma tarafından açık kaynak ortamdan araştırmacıların kullanımı için paylaşılmıştır. Aynı zamanda bu çalışma tarafından tasarlanan mimaride bu veriler kullanılmıştır. Bizim önerdiğimiz sistem ile bu çalışmanın sonuçlarının özeti Tablo 5'te verilmiştir. Aynı veri seti kullanıldığından bu karşılaştırma anlamlıdır.

Tablo 5: Karşılaştırma tablosu.

Çalışma	Model	Sonuç
Bu Çalışma	M-NB +IG	93.1 (F1)
Bu Çalışma	LogReg+PV-DM	92.3 (F1)
Diğer Çalışma [9]	RO + Zemberek	%90.1
Diğer Çalışma [9]	RO + Zemberek +ARFS	%91.3

Tablo 5'te görüldüğü üzere önerdiğimiz geleneksel yöntem ve YSA temelli model TTC-3600 veri kümesi ile sırasıyla 93.1 ve 92.3 F1 başarı skoru elde ederken, diğer çalışmadan [9] elde edilen en iyi sonuç RO+Zemberek+ARFF modeli tarafından kaydedilen %91.3 değerindedir. Veri setinde sınıf dağılımları tekdüze (uniform) olduğu durumlarda bir modelin F1 ve başarı değerleri yakın çıktığından böyle bir karşılaştırma anlamlıdır. Çünkü TTC 3600 verisinde sınıf değerleri tekdüze dağılım sergilemektedir. Burada RO sınıflandırıcısını, Zemberek kökleri bulan yapıyı ve ARFS (Attribute ranking-based Feature Selection) ise özellik seçimi ile boyut indirgemeyi sağlayan algoritmayı ifade etmektedir.

Sonuç olarak YSA temelli yaklaşımlarla oluşturulan mimariler ile özellik seçimiyle desteklenmiş geleneksel kelime torbası yaklaşımlarını karşılaştırdığımızda şu görülmektedir: m-NB + Bilgi Kazancı özellik seçimi düzeni ile PV-DM paragraf vektör oluşturma yaklaşımı metin sınıflandırma açısından yakın sonuçlar vermiştir. Geleneksel yaklaşım diğer YSA temelli kelime vektör ortalamasından üstün gelmiştir.

Özellikle derin YSA temelli yaklaşımların DDİ alanına büyük yenilikler ve avantajlar sağlamasına karşın, geleneksel yöntemler hala iyi sonuçlar vermektedir Bazı koşullarda geleneksel yöntemlerin hala daha başarılı olduğu görülmüştür. Bizim çalışmamızdaki geleneksel yöntemin kelime vektör ortalaması yaklaşımdan üstün olması bunun bir ifadesidir. Öte yandan [5] referanslı çalışma YSA temelli yaklaşımların duygu analizi ve bazı Bilgi Getirme konularında geleneksel yöntemlerden iyi çalıştığını vurgulamıştır. Ancak bu çalışmada elde ettiğimiz sonuçlarda Türkçe metin sınıflandırılması açısından iki yaklaşımın da benzer sonuçlar ürettiği görülmüştür.

6 Önişleme süreçlerinin başarıya etkisi

Kök bulma algoritmaları kelimenin yüzey formundaki çekimsel ve türetimsel ekleri atarak kelimenin kökünü bulmaya yarayan bir yaklaşımdır. Türkçe zengin eklemeli dil yapısı sebebiyle bu tip analize ihtiyaç duyar. Eklerin elenmesi birçok DDİ çalışması için önemli olmaktadır. Bu çalışmada kelime kökleri Zemberek [23] kütüphanesiyle bulunarak metin sınıflandırmasına katkısı incelenmiştir. Tablo 6 kök bulma ön işleminin geleneksel yöntemlere katkısını T-4900 verisi üzerinden elde edilen sonuçlar aracılığı ile göstermektedir. Burada en iyi sınıflandırıcılardan m-NB ve DKM algoritması seçilirken özellik seçimi için BK ve ki-kare kullanıldı. Özellik vektör boyu 2K seçildi. Bu tablo şunu gösteriyor ki geleneksel yöntemler arasındaki en iyi metin sınıflandırıcı düzenek olan BK + m-NB mimarisi kök alma önişleme sürecinden olumsuz etkilenirken bazı düzenlerde kök kullanılmasıyla daha başarılı sonuçlar elde edilmiş. Ancak bu işlem zaten çok başarılı olan ve en iyi sonuçların elde edildiği Bilgi Kazancı + m-NB yapısına herhangi bir katkı yapmamıştır. Sonuç olarak kök bulma işleminin katkısı kurulan modele göre değişiklik göstermektedir.

Tablo 6: Kök algoritmasının etkisi.

Model (F1)	KÖK	Yüzey Formu
BK + m-NB	87.5	89.1
BK + DKM	80.5	80.0
Ki-kare+ m –NB	86.6	89.0
Ki-kare + DKM	81.1	79.0

Öte yandan kök bulma yaklaşımını YSA temelli yaklaşımlar açısından ele aldık. Tablo 7 YSA temelli iki yaklaşımın yüzey formları ve kelime kökleri ile elde ettiği F1 başarı sonuçlarını göstermektedir. PV-DM mimarisinde yüzey formu ve kök kullanıldığında en iyi algoritma olan Logistic Regression sınıflandırıcısının başarısı sırasıyla 89.2 F1 ve 86.5 F1 olmuştur. Benzer bir şekilde DKM algoritması da olumsuz etkilenmiştir. Öte yandan bir tek PV-BOW ile DKM algoritması kök alma işleminden olumlu etkilenmiştir. Sonuç olarak KÖK bulma algoritmaları metin sınıflandırma açısından kurulacak düzeneğe göre farklılık gösterir. Çok büyük iyileştirmeler yapmış olmasa da [9] çalışmasında kök yaklaşımının ortalama 2 puanlık bir iyileştirme yaptığı görülmüştür.

Tablo 7: Kök bulmanın YSA algoritmasına katkısı.

Model (F1)	LogReg	DKM
Pv-DM	89.2	87.2
PV-DM + KÖK	86.5	84.3
PV-DBoW	86.6	88.2
PV-DBoW + KÖK	79.3	88.8

Başarıyı etkileyecek diğer faktörlerden terim ağırlıklandırma yöntemi üretilen metin vektörleri için vektör hücrelerindeki değerin nasıl hesaplanacağı ile ilgilenir. Temelde dört yaklaşım vardır. İkili (binary) kelimenin varlığı veya yokluğu, terim frekansı (term ferquency, tf) kelimenin verili metinde geçme sıklığı, logaritmik sıklık (logarithmic tf function) ve verili metinde geçme sıklığı ile tüm metinlerdeki geçme sıklığına (metin frekansı) bağlı fonksiyon (tf-idf). Bu ağırlıklandırma yöntemleri de kök yaklaşımda olduğu gibi kullanılan sınıflandırıcı ve özellik seçiminden etkilenmektedir. Deneylerimizde çok küçük farkla tf-idf 'in daha başarılı olduğu kaydedilmiştir. Örnek olarak bu dört ağırlıklandırmadan BK+ m-NB düzeninde en iyi sonucu %90.1 ile tf -idf elde ederken en

kötü sonucu %88.5 ile ikili ağırlıklandırma elde etmiştir. Ancak bu olgu farklı bir sınıflandırıcı kullanıldığında değişebilmektedir.

Etkisiz kelime (Stopword) eleme bir diğer ön işleme yöntemidir. Ancak burada BK gibi özellik seçimi algoritması kelimelerin etkisini ölçebildiğinden ve sınıflandırmaya katkısını hesap edebildiğinden bu işleme gerek yoktur. Kelimeler ayırt ediciliği üzerinden elendiğinden etkisi veya etkisizliği doğal bir şekilde hesaplanmaktadır. Örnek olarak 2K özellik seçimi aşamasında 226 etkisiz kelime listesinden sadece 58 tanesi özellik listesine dahil olabilmiştir. Geri kalan 168 kelime BK tarafından zaten otomatik olarak elenmiştir. Etkisiz kelime listesi kullanıldığı ve kullanılmadığı durumlarda ise başarı oranlarında herhangi bir değişim gözlemlenmemiştir. Etkisiz kelime elemesi ile m-NB algoritması yine aynı performansı göstermiştir.

7 Sonuç

Son yıllarda derin YSA temelli Mimariler DDİ problemlerine büyük yenilikler getirmiş ve başarılı çözümler sunmuştur. Bu çalışmada YSA temelli metin gösterilimleri ile geleneksel metin gösterilimleri sınıflandırma açısından karşılaştırılmıştır. İki veri kümesi kullanılarak word2vec ve doc2vec yaklaşımları geleneksel kelime torbası yöntemiyle ele alınmıştır. Geleneksel yöntemde en kritik aşama özellik seçimi adımıdır. Bu aşamada Bilgi Kazancı ve ki-kare yaklaşımı kullanıldı ve bu iki metrik çok başarılı bulundu. YSA temelli mimarilerde PV-DM, PV-DBoW, (PV-DM+ PV-DBoW) ve vektör ortalaması olmak üzere dört yaklaşım ele alınmıştır. Bu yaklaşımlardan PV-DM modeli Logistic Regression sınıflandırıcısı ile T-4900 ve TTC-3600 veri kümelerinde sırasıyla 89.0 ve 92.3 F1 başarısı elde ederken, kelime torbası temelli yaklaşımlardan Bilgi Kazancı özellik seçimli m-NB yaklaşımı sırasıyla 90.0 ve 93.1 F1 skoru ile benzer bir başarı elde etmiştir. TTC-3600 veri kümesini kullanan çalışmayla [9] karşılaştırdığımızda önerdiğimiz modellerin daha iyi sonuç verdiği görülmüştür.

YSA temelli yaklaşımın duygu analizi ve bilgi çıkarımı problemlerinde geleneksel yöntemlere göre daha başarılı olduğu vurgulamışsa da [5], bu çalışma göstermiştir ki, geleneksel yöntemler bazı alanlarda daha etkili olabilir. Bizim çalışmamızda geleneksel kelime torbası yöntemi kelime vektör ortalamasından daha başarılı olmuştur. Derin YSA temelli modellerin DDİ alanında büyük değişimler ve heyecan getirdiği bilinmekle birlikte bu yaklaşımların her zaman başarılı olma garantisi yoktur. Alana özgü deneylerin detaylı bir şekilde yapılması ve ona göre bir mimari oluşturulması gerekmektedir.

Son olarak önişleme aşamalarının metin sınıflandırma çözümüne katkısı irdelenmiştir. Tasarladığımız mimaride iki önişleme aşaması olan kök yöntemi ve etkisiz kelime elemenin belirgin bir katkısı görülmemiştir. Literatürdeki çalışmalarda bu konuda farklı sonuçlar alındığı ve farklı değerlendirmeler yapıldığı görülmektedir. Bazı çalışmalar kök bulma yöntemi ile daha iyi sonuçlar elde edebilmişlerdir. Öte yandan bazı çalışmalar ise bu iki yöntemin katkısının alana ve modele bağlı olduğunu belirtmiştir. Bu çalışmada her iki veri seti ve her iki yaklaşım açısından önişleme süreçlerinin belirgin bir katkısı görülmemiştir.

8 Kaynaklar

[1] Salton G, Wong A, Yang CS. "A vector space model for automatic indexing". *Communications of the ACM*, 18(11), 613-620, 1975.

[2] Harris, Z. "Distributional structure". *Word*, 10(2), 146-162, 1954.

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

[4] Pennington J, Socher R, Manning C. "Glove: Global vectors for word representation". *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Doha, Qatar, 25-29 October 2014.

[5] Le Q, Mikolov T. "Distributed representations of sentences and documents". *31th International Conference on Machine Learning*, Beijing, China, 21-26 June 2014.

[6] Amasyalı MF, Diri B. *Automatic Turkish text categorization in terms of author, genre and gender.* Natural Language Processing and Information Systems, Lecture Notes in Computer Science, Vol 3999, 221-226, Berlin, Heidelberg, Germany, Springer, 2006.

[7] Türkoğlu F, Diri B, Amasyalı MF. *Author Attribution of Turkish Texts by Feature Mining.* International Conference on Intelligent Computing, Lecture Notes in Computer Science, vol 4681. Springer, Berlin, Heidelberg, 2007.

[8] Amasyalı MF, Balcı S, Mete E, Varlı EN. "Türkçe metinlerin sınıflandırılmasında metin temsil yöntemlerinin performans karşılaştırılması". *EMO Bilimsel Dergi*, 2(4), 2012.

[9] Kılınç D, Özçift A, Bozyigit F, Yıldırım P, Yücalar F, Borandag E. "TTC-3600: A new benchmark dataset for Turkish text categorization". *Journal of Information Science*, 43(2), 174-185, 2015.

[10] Tüfekçi P, Uzun E, Sevinç B. "Text classification of web based news articles by using Turkish grammatical features". *20th Signal Processing and Communications Applications Conference (SIU)*, Muğla, Turkiye 18-20 April 2012.

[11] Akkuş BK, Çakıcı R. "Categorization of Turkish news documents with morphological analysis". *51st Annual Meeting of the ACL Proceedings of the Student Research Workshop*, Sofya, Bulgaristan, 5-7 August 2013.

[12] Torunoğlu D, Çakırman E, Ganiz MC, Akyokuş S, Gürbüz MZ. "Analysis of preprocessing methods on classification of Turkish texts.". *International Symposium on Innovations in Intelligent Systems and Applications (INISTA)*, İstanbul, Türkiye, 15-18 June 2011.

[13] Uysal AK, Günal S. "The impact of preprocessing on text classification". *Information Processing and Management*, 50(1), 104-112, 2014.

[14] Yıldırım S. "A knowledge-poor approach to turkish text categorization". 15th *International Conference on Computational Linguistics and Intelligent Text Processing*. Katmandu, Nepal, 6-12 April 2014.

[15] Açıkalın B, Bayazıt NG. "The importance of preprocessing in Turkish Text classification". *24th Signal Processing and Communication Application Conference*, Zonguldak, Türkiye, 16-19 May 2016.

[16] Amasyalı MF, Beken A. "Türkçe kelimelerin anlamsal benzerliklerinin ölçülmesi ve metin sınıflandırmada kullanılması". *Signal Processing and Communication Application Conference*, Antalya, Türkiye, 9-11 Nisan 2009.

[17] Toraman Ç. Text Categorization and Ensemble Pruning in Turkish News Portals. PhD Thesis, Bilkent University, Ankara, Turkey, 2011.

[18] Schütze H, Hull DA, Pedersen JO. "A comparison of classifiers and document representations for the routing problem". *18th ACM Conference on Research and Development in Information Retrieval*, New York, USA, 9-13 July 1995.

[19] Lewis D, Ringuette M. "A comparison of two learning algorithms for text categorization". *3rd Annual Symposium on Document Analysis and Information Retrieval*, Las Vegas, USA, 11-13 April 1994.

[20] Manning CD, Raghavan P, Schütze H. *Introduction to Information Retrieval*, 1st Edition, New York, USA, Cambridge University Press, 2008.

[21] Zhang W, Yoshida T, Tang X. "A comparative study of TF*IDF, LSI and multi-words for text classification". *Expert System Application*, 38(3), 2758-2765, 2011.

[22] Akin A, Akin MD. "Zemberek, an open source NLP framework for Turkic Languages". *Structure*, 10, 1-5, 2007.