**TOBB EKONOMİ VE TEKNOLOJİ ÜNİVERSİTESİ**
**FEN BİLİMLERİ ENSTİTÜSÜ**

# DERİN ÖĞRENME YÖNTEMLERİ KULLANILARAK TÜRKÇE DOKÜMAN SINIFLANDIRMA

**YÜKSEK LİSANS TEZİ**

**Mustafa SARI**

**Bilgisayar Mühendisliği Anabilim Dalı**

**Tez Danışmanı: Dr. Öğr. Üyesi Ahmet Murat Özbayoğlu**

**Kasım/2018**

Fen Bilimleri Enstitüsü Onayı

……..

**Prof. Dr. Osman EROĞUL**

Müdür

Bu tezin Yüksek Lisans derecesinin tüm gereksinimlerini sağladığını onaylarım.

…….

**Prof. Dr. Oğuz ERGİN**
Anabilimdalı Başkanı

TOBB ETÜ, Fen Bilimleri Enstitüsü'nün 141111052 numaralı Yüksek Lisans **Mustafa SARI**'nın ilgili yönetmeliklerin belirlediği gerekli tüm şartları yerine getirdikten sonra hazırladığı **"DERİN ÖĞRENME YÖNTEMLERİ KULLANILARAK TÜRKÇE DOKÜMAN SINIFLANDIRMA"** başlıklı tezi **28.11.2018** tarihinde aşağıda imzaları olan jüri tarafından kabul edilmiştir.

**Tez Danışmanı : Dr. Öğr. Üyesi Ahmet Murat Özbayoğlu** .....
TOBB Ekonomive Teknoloji Üniversitesi

**Jüri Üyeleri : Prof. Dr. Erdoğan Doğdu (Başkan)** .....
Çankaya Üniversitesi

**Prof. Dr. Ali Aydın Selçuk** .....
TOBB Ekonomive Teknoloji Üniversitesi

## TEZ BİLDİRİMİ

Tez içindeki bütün bilgilerin etik davranış ve akademik kurallar çerçevesinde elde edilerek sunulduğunu, alıntı yapılan kaynaklara eksiksiz atıf yapıldığını, referansların tam olarak belirtildiğini ve ayrıca bu tezin TOBB ETÜ Fen Bilimleri Enstitüsü tez yazım kurallarına uygun olarak hazırlandığını bildiririm.

.

Mustafa SARI

# ÖZET

Yüksek Lisans

DERİN ÖĞRENME YÖNTEMLERİ KULLANILARAK

TÜRKÇE DOKÜMAN SINIFLANDIRMA

Mustafa SARI

TOBB Ekonomi ve Teknoloji Üniveritesi
Fen Bilimleri Enstitüsü
Bilgisayar Mühendisliği Anabilim Dalı

Danışman: Dr. Öğr. Üyesi A. Murat Özbayoğlu

Tarih: Kasım 2018

Çalışmamızda kaleme alınmış yazıların, yazarına ve konusuna göre birbirinden ayrılması ve sınıflandırılabilmesi amaçlanmıştır. Bir gazetenin köşe yazarlarının yazılarının vektörleri oluşturulmuştur ve birbirinden ne kadar ayrılabildiğinin analizi yapılmıştır. Yazarı bilinmeyen herhangi bir yazının hangi yazara ait olduğu belirlenebilir veya birbirlerine benzer stiller gruplanarak yazar profilleri oluşturulabilir. Konusu bilinmeyen bir yazının hangi konulara ait olabileceği belirlenebilir. DeepLearning4J Java kütüphanesi ve burada yer alan Doc2Vec sınıfı kullanılmıştır. 5,10,15 ve 20 yazar içeren modeller ve yazarların yazdıkları konulara göre modeller geliştirilmiştir. Bu şekilde elde edilen benzerlik vektörleri belirli bir eşik değeri ile karşılaştırılmıştır, değişik eşik değerleri seçimine bağlı model başarımları ölçülmüştür. Elde edilen sonuçlara göre bazı yazarlar belirgin bir şekilde diğer yazarlardan ayrılmaktadır. Yazılar konularına göre başarılı bir şekilde etiketlenebilmektedir. Bu yapı özellikle yazar profili çıkarımı, yazı tespiti veya konu gruplama gibi alanlarda kullanılabilecek niteliktedir.

**Anahtar Kelimeler:** PV-DBOW, PV-DM, DL4J, Paragraf vektörleri, word2Vec, doc2Vec, Metin madenciliği.

## ABSTRACT

Master of Science

CLASSIFICATION TURKISH DOCUMENTS USING DEEP LEARNING TECHNIQUES

Mustafa SARI

TOBB University of Economics and Technology
Institute of Natural and Applied Sciences
Computer Science Programme

**Supervisor:** Asst. Prof. Ahmet Murat Özbayoğlu

Date: NOVEMBER 2018

In our study, it is aimed to distinguish and classify author profiles and text subjects with vectors which were created from authors posts. The vectors of the columnists of a newspaper were formed and analyzed for how much they could be separated from each other. Hence, author of any post, can be determined by this model. It also can group similar styles together. The DeepLearning4J Java library and the Doc2Vec class included are used during development. 5,10,15, 20 author vector models and their subject models were created according to their posts. The similarity vectors obtained in this way were compared with a certain threshold value, and the model performances based on the selection of different threshold values were measured. According to the results, some authors differed significantly from other authors. Articles can be successfully labeled according to their topics. This structure can be used especially in areas such as author profile extraction, article detection or subject grouping.



**Keywords:** PV-DBOW, PV-DM, DL4J, Paragraph vectors, word2Vec, doc2Vec, Text mining.

## TEŞEKKÜR

Çalışmalarım boyunca değerli yardım ve katkılarıyla beni yönlendiren hocam Dr. Öğr. Üyesi Murat Özbayoğlu, yüksek lisans boyunca her zaman yanımda olan ve desteğini hiç esirgemeyen sevgili annem Fatma Sarı ve babam Bekir Sarı'ya, kıymetli tecrübelerinden faydalandığım TOBB Ekonomi ve Teknoloji Üniversitesi Bilgisayar Mühendisliği Bölümü öğretim üyelerine çok teşekkür ederim.

## İÇİNDEKİLER

**Sayfa**

# ŞEKİL LİSTESİ

**Sayfa**

## ÇİZELGE LİSTESİ

**Sayfa**

## KISALTMALAR

| | |
|---|---|
| **SVM** | **:** Destek Vektör Makinesi – Support Vector Machine |
| **DL4J** | **:** Java için Derin Öğrenme - Deep Learning For Java |
| **BOW** | **:** Kelimeler Çantası – Bag of Words |
| **P** | **:** Kesinlik - Precision |
| **R** | **:** Duyarlılık - Recall |
| **PV-DM** | **:** Paragraf Vektör Dağıtılmış Bellek Modeli - Paragraph Vector Distrubuted Memory Model |
| **PV-DBOW** | **:** Paragraf Vektör Dağıtılmış Kelimeler Çantası - Paragraph Vector-Distributed Bag of Words |
| **SMC** | **:** Denetimli Anlam Sınıflandırma -Supervised Meaning Classifier |
| **ND4J** | : Java İçin Çok Boyutlu Diziler - N-Dimensional Arrays for Java |

# 1. GİRİŞ

Her insanın parmak izi gibi yazı stilleri farklıdır, el yazısını bildiğimiz birinden bir mektup geldiğinde üstünde ismi yazmasa bile kimden geldiğini kolayca anlayabiliriz. 3000 yıllık bir geçmişi olan grafoloji bilimi insanları yazı stillerinden ayırt etme ve analiz etme üzerine çalışan bilim dalıdır. Grafologlara göre dünyada parmak izinde olduğu gibi iki insanın yazı stili birbiri ile aynı olamaz [1]. Nitekim bir kişinin yazı stili sadece parmak kasları ile değil kişinin beyni ile de ilgilidir. Günümüzde neredeyse çoğu yazının elektronik ortamda yazıldığını düşünürsek, ortada bir el yazısı olmadığı için kişileri yazılarından bu yöntemle ayırt edilmesi mümkün değildir.

Mobil cihazların kullanımının yaygınlaşması ve insanların internete erişiminin kolaylaşması ile internette paylaşılan bilgi de artırmıştır. Sosyal ağlarda birçok metin ilgi çekmesi için bilindik yazarların ismi ile yayınlanmaktadır. Bu da bilgi kirliliğine neden olmaktadır. Yazarların profilinin oluşturulması ve yazarların stilinin tespit edilmesi, bir metnin ilgili yazara ait olduğunu teyit açısından önemlidir.

Bu tez çalışmasında Hürriyet gazetesi köşe yazarlarının yazılarından elde edilen veriler ile yazarların profili oluşturulmuş. Yazarların stilleri ve benzerlikleri kıyaslanmıştır. Yazarların köşe yazılarından oluşturulmuş vektörler aracılığı ile yazarı belli olmayan bir yazının hangi köşe yazarına ait olabileceği olasılığı hesaplanmıştır.

İnternette arama yapmak ve bilgiye ulaşmak için en yaygın kullanılan araçlardan birisi de Google'dır. Bir yazının alıntı olup olmadığı Google aracılığıyla taranarak kaynağından nereye veya kime ait olduğu bulunabilir. Turnitin gibi programlar aracılığı ile intihal tespiti daha önceki yazılmış yazılardan tespit edilebilir. Ancak internet ortamında veya Turnitin veritabanına girmemiş bir yazı üzerinden farklı bir yazara ait olabileceğinin tespitinin yapılması problem oluşturabilmektedir. Yazarın internet ortamında henüz yayınlanmamış bir yazısı dahi olsa, yazının stilinden intihal olasılığının tahmin edilmesi çalışmayı Turnitin gibi programlardan farklı kılan yönlerden birisidir.

Çalışmamızda doc2Vec [2] tekniği kullanılarak yazarların vektör modelleri oluşturulmuş, vektör modellerinin oluşturulmasındaki hiper parametrelerin değiştirilmesi ile sınıflandırmadaki başarım oranı değerlendirilmiştir.

Yazarların yazılarını yazdıkları konulara göre sınıflandırdığımız çalışmamızda ise konusu bilinmeyen bir metnin hangi konuya ait olabileceği tahminlemesi yapılmıştır. Ayrıca konuların çok boyutlu vektörleri oluşturulmuş ve bu çok boyutlu vektörlerin birbirleri arasındaki uzaklıkları 2 boyuta indirgenerek görselleştirilmiştir.

Bu tez çalışması şu şekilde düzenlenmiştir: Bölüm 1 de tez çalışması hakkında genel bilgiler, kullanılan araçlar ve veri kaynakları anlatılmıştır. Bölüm 2'de literatürdeki benzer çalışmalar ele alınmış ve açıklanmıştır. Bölüm 3'te kullanılan araçlar açıklanmıştır. Bölüm 4'te verilerin nasıl toplandığı ve nasıl işlendiği ele alınmış, elde edilen sonuçlar açıklanmış ve elde edilen modellerin performansları değerlendirilmiştir. Bölüm 5'te ise modellerin güçlü ve zayıf yönlerinden bahsedilmiş, gelecek çalışmalar hakkında öneriler sunulmuştur.

## 2. GEÇMİŞ ÇALIŞMALAR

Bir yazıyı konusuna veya yazarına göre ayırmak için eskiden beridir çok fazla çalışma yapılmıştır. Bir metni diğerinden ayırmada metinin stili ve metnin içeriği kullanılan iki özellik olmuştur. Bu alanda birçok dilde olduğu gibi Türkçe alanında da çalışmalar mevcuttur.

### 2.1 Metin Madenciliği Tabanlı Duygu Analizi, Yazar ve Yazı tespiti

Metin madenciliği alanında yapılan en eski çalışmalardan; Brinegar (1963) kullanılan kelimelerin uzunluğuna göre çıkarımlar yaparken, Morton [3] (1965) cümle uzunluklarını ayırt edici özellik olarak kullanarak yunan nesir yazılarında, Brainerd (1974) ise heceleri kurarak ayırt edici özellik olarak, Holmes (1992) kelimeler ve metnin uzunluğu arasında ilişki kurarak çalışma yürütmüştür [4].

Çince soruları birden fazla etiketle sınıflandırmak için word2vec kullanan Fan, Su,Liu ve Wang, yaptıkları çalışmada [5] soru cevap forumlarından elde ettikleri verileri eğitim ve test verisi olarak kullanmışlardır. Kullanılan yöntem ile elde edilen sonuçlara göre kültür sanat, sosyal, spor, bilim, ekonomi, mevzuat alanındaki soruları başarılı bir şekilde sınıflandırmışlardır.

[6] çalışmasında film yorumları eğitim seti olarak kullanılmış ve elde edilen yorumların olumlu, olumsuz ve nötr olarak sınıflandırılması çalışılmıştır. Yöntem olarak kelimeler çantası (Bag of Words - BoW), unigram ve bigram özellikleri kullanılmıştır. Fakat unigram, bigram ve BoW yöntemleri kelimelerin tekrarını ve frekansını göz önünde bulundurup, kelimenin bağlamını temsil etmedikleri için karmaşık yorumlarda başarısız olduğu belirtilmiştir. Örneğin "Filmde ki oyuncular çok iyi ve senaryo çok iyi olsa da, film bir hayli uzundu, eğlendiğimi söyleyemem"

cümlesinden bir kişi yorumun olumsuz olduğunu çıkarırken, bütüne bakılmayan bir yöntemde bu yorum olumlu olarak kabul edilebildiği sonucuna ulaşıldığı belirtilmiştir. [7]'daki çalışmasında benzer bir çalışma yapılmış, Twitter'dan elde edilen veriler olumlu olumsuz ve nötr olmak üzere üç kategoride sınıflandırılmıştır. Yazarların kullanmış olduğu unigram ve duygu-özellikeri (senti-feature) yöntemi ile %75 lik doğruluk ile başarım sağlamıştır. Elde edilen verileri gerçekler ve fikirler olmak üzere iki sınıfa ayıran bir başka çalışmada ise [8] gerçeklerin varlıkların nesnel değerlerden başka bir şey olmadığın ifade edilirken, görüşler ise kişilerin nesnelere yönelik duygularını açıklayan öznel ifadeler olduğu açıklanmıştır.

[9] çalışmasında ise Naïve Bayes ve SVM yöntemleri kullanılarak politik web günlükleri sınıflandırılmıştır. Naïve Bayes sınıflandırıcısının SVM'den önemli ölçüde daha iyi performans gösterdiğini gözlemlenmiştir.

[10] çalışmasında bir duygu-sözlüğü (senti-lexicon) oluşturularak bir restorana yapılan yorumlar üzerinde duygu analizi yapılmıştır. Naïve Bayesin geliştirilmiş iki versiyonunu önererek, bu iki versiyonu performanslarını orijinal Naïve Bayes ve SVM ile karşılaştırılmıştır. Elde edilen sonuçlardan Naïve Bayes geliştirilmiş sürümlerinin daha etkili olduğu sonucuna ulaşılmıştır.

Metinleri duygulara göre sınıflandırılan [11] çalışmasında yazarların ortaya koyduğu yeni bir yöntem öne sürülmüştür. SMC olarak isimlendirilen yaklaşımda %1 gibi çok az eğitim verisi kullanılarak farklı yaklaşımların başarısı gözlemlenmiştir. Multinomial Naïve Bayes (MNB) ve SVM gibi sınıflandırma yöntemlerinin kıyaslandığı çalışmada, SMC yönteminin daha başarılı olduğu sonucu elde edilmiştir. Eğitim verisi kümesi %70 te iken SMC yöntemi %84 başarı gösterirken SVM %69 luk başarım gösterdiği kaydedilmiştir. Çalışmanın esas konusu olan çok düşük eğitim verisindeki performans analizinde ise eğitim seti %1 de iken SMC %48 başarı sağlamış, SVM %19 başarım sağlayarak, yazarların ortaya koyduğu yöntemin daha başarılı olduğu gözlemlenmiştir.

## 2.2 Türkçe Yazılar ve Yazarlar Üzerinde Yapılan Çalışmalar

Mayda ve Yesiltepe Türkçe dokümanlarda metin madenciliği yöntemlerini kullanarak [12] yaptıkları çalışmada SVM, SMO ve unigram, bigram performans analizini gerçekleştirmişlerdir. Buna benzer başka bir çalışmada ise Deniz ve Kizilöz gazete yazarlarını [13] n-gram teknikleri kullanarak üsluplarına ve cinsiyetlerine göre kategorize etmişler ve karakter seviyesinde n-gram (character level n-gram) ve kelime seviyesinde n-gram (word level n-gram) arasındaki performansı karşılaştırılmışlardır. Elde ettikleri sonuca göre SVM, Naïve Bayes ve Random Forest yaklaşımından daha iyi performans gösterdiğini gözlemlemişlerdir. Köşe yazılarını 0.68 F-Ölçüt skoru ile ekonomi, medya ve spor kategorilerinde sınıflandırmışlardır.

Bir diğer çalışmada atılan twitlerden duygu analizi yapılmıştır. Analizde PV-DBOW ve PV-DM modelleri, Türkçe ve İngilizce dokümanlarda kıyaslanmıştır. [14]. Çalışmada duygular "pozitif", "negatif" ve "nötr" olarak sınıflandırılmıştır. Kişilerin twitter kullanıcı adları, isimleri, html etiketleri gibi veriler eğitim setinin içine katılmamıştır. Cümle sayısı arttıkça başarım oranı ölçümleri yapılmıştır. Elde edilen sonuçlara göre PV-DBOW yönteminin PV-DM yöntemine göre başarılı sonuçlar verdiğini gözlemlenmiştir.

Karasoy ve Ballı yaptıkları çalışmada word2vec yöntemi kullanarak gelen SMS leri spam olup olmadığına göre sınıflandırmıştır. Word2vec yöntemi ile birlikte rastgele orman (random forest) algoritması kullanılmış ve SMS leri başarılı bir şekilde sınıflandırmışlardır[15]. Sahin'in çalışmasında ise [16] metinleri word2vec kullanarak eğitim seti büyüklükleri farklı 7 kategoride sınıflandırmış word2vec yaklaşımının den çok daha başarılı olduğu gözlemlenmiştir.

Çoban ve Karabey yaptıkları çalışmada [17] müzik türlerini sınıflandırmış ve doc2vec ve word2vec yöntemlerini klasik BOW ve CBOW yaklaşımlarıyla kıyaslamıştır. Elde ettikleri sonuçlara göre klasik yaklaşımların müzik türlerini sınıflandırmada daha başarılı olduğunu gözlemlemişlerdir. Elde edilen modeller arasında farklılıklar

olduğunu belirtmiş, sınıflandırmada modelleri ASCII ye çevirerek oluşturmanın bu problemi ortadan kaldırabileceğini önermişlerdir.

Türkçe metinle ironi tespitinin yapıldığı [18] çalışmasında sözlü ironi, duyumsal ironi ve dramatik ironi gibi ironi türleri tespit edilmiş, ironiler kullanılan noktalama işaretleri ile tespit edilmeye çalışılmıştır. Karar ağacı, Naïve Bayes, Lojistik Regresyon, Karar Tablosu, Rastgele Orman, Çok Katmanlı Algılama (Multilayer Perception) yaklaşımları kıyaslanmış ve ironi tespitinde en başarılı algoritmaların Rastgele Orman, Çok Katmanlı Algılama olduğu tespit edilmiştir.

İntihal tespit yazılımlarının performanslarının karşılaştırıldığı bu [19] çalışmada ise metinlerde Türkçe karakterlerin olması intihal tespitinin performansını düşürdüğü tespit edilmiştir. Metinlerde zemberek yardımıyla kelime kökenini de göz önünde bulundurup benzerlik analizi yaptıklarında daha başarılı sonuçlar aldıklarını belirtmişlerdir.

Metinlerin ekonomi, magazin, sağlık, siyaset, spor alanlarında sınıflandırıldığı çalışmada [20] kelimelerin bir metin içinde geçme sıklığı baz alınarak sınıflandırma yapılmıştır. Sınıflandırmada Naïve Bayes (NB), Destek Vektör Makinesi (DVM), K-NN Yakın Komşuluk (KEK) ve Rastgele Orman (RO) yöntemlerinin performansları kıyaslanmışıtır. En yüksek başarı %96 ile Naïve Bayes'te elde edilmiştir. Sınıflandırmada en başarılı sonuçlar ekonomi alanında gerçekleştirilirken, en çok yanlış yapılan alan siyaset olduğu tespit edilmiştir. Sağlık ve siyaset alanında geçen bazı kelimeler ekonomi alanında da geçtiği için sınıflandırmanın başarımınını düşürdüğü gözlemlenmiştir.

Sabah gazetesinden seçilen yazarların köşe yazıları ile yapılan bir başka çalışmada ise yazarların profilinin oluşturulması ve köşe yazılarının hangi yazara ait olduğu çalışması yapılmıştır. Çalışmada yazarların kullandıkları noktalama işaretleri, boşluklar, alt satıra inme sayıları yazarları ayırt etmede kullanılmıştır[21]. Bazı yazarların ünlem işaretini nadiren kullandığı, bazı yazarların ise ünlem işaretini daha

sık kullandığı, bazılarının ise uzun cümleleri tercih ettiği için daha çok virgül kullandığı gibi farklılıklar tespit edilmiştir. Bu verilerle oluşturulan vektörler ile yazarların metinleri arasında benzerlik analizi çıkartılmıştır ve %86 oranında başarım sağlamışlardır.

Yazar tespit üzerine Hürriyet gazetesi üzerinde yapılan bir çalışmada ise magazin sağlık ve siyaset alanından alınan 18 farklı yazara ait metinler üzerinde çalışılmıştır[22]. Her bir yazardan 20 adet köşe yazısı seçilerek, bu yazılar üzerinden performans analizi yapılmıştır. Yazarların ortaya koyduğu yaklaşımda; her bir metindeki kelime sayısını, cümle sayısı, ortalama kelime uzunluğu, cümle uzunluğu, farklı kelime sayısı, soru işaretleri, noktalı virgül, ünlem gibi noktalama işaretleri sayısı, satır sayısı gibi özelliklerin yanı sıra cümlelerin gramer özellikleri ayırt edici parametreler olarak tespit edilmiştir. Buna göre her bir yazarın metinlerindeki ortalama isim sayısı, ortalama sıfat sayısı, ortalama yüklem sayısı gibi özelliklerle beraber 22 farklı ayırt edici stil parametreleri çıkartılmıştır. Bu 22 özellik üzerinden oluşturulmuş vektör üzerinden yazarlar kıyaslanarak %67 lik bir başarım elde edilmiştir. Yazarların ortaya koyduğu yaklaşım olan Çok Katmanlı Algılama (Multilayer Perception - MLP) yöntemi ile %60 lik bir başarım elde etmişlerdir, Radyal Taban Fonksiyonu (Radial Base Function) yaklaşımı ile ise %72 lik bir başarım elde etmişlerdir. Çalışmada yazarların birbirinden ayrılmasında en çok başarı elde edilen özellikleri şöyle sıralamışlardır: bir cümledeki ortalama kelime sayısı, kullanılan ortalama kelime uzunluğu, ortalama isim sayısı, ortalama sıfat sayısı, ortalama yüklem sayısı, ortalama bağlaç sayısı, ortalama zamir sayısı, nokta sayısı, tamamlanmamış cümle sayısının tüm cümlelere oranı, devrik cümlelerin tüm cümlelere oranı.

Borsa İstanbul 100 indeksi üzerine yapılan bir çalışmada ise [23] ekonomi haberleri kullanılarak endeksin günlük açılış yönü tahmin edilmiştir. Yapılan çalışmada finansal internet sitelerinden veriler alınarak BIST 100 endeksini hangi yönde etkileyeceği tahminlemeye çalışılmıştır. Her dokümanın özniletiklerine göre vektörleri oluşturulmuş ve Naïve Bayes sınıflandırıcısı ile sınıflandırılmıştır. Çalışmada borsa yönü %68 F-Ölçütü başarı oranı ile tahmin edilmiştir.

[24] çalışmasında anlık mesajların konusunun tespit edilmeye çalışılmıştır. Konuşmalarda geçen emojiler ve kısaltmalarında göz önünde bulundurulduğu çalışmada başlıklar spor, aşk/evlilik, eğitim, eğlence, küfür ve diğer olarak belirlenmiştir. Metinleri ayırmada Naïve Bayes, k-NN ve SVM sınıflandırma yöntemleri kullanılmıştır. En başarılı sonuç ise %92 ile küfür olarak belirlenmişlerdir. Bunun sebebi ise küfür konuşmalarının içerdiği kelimeler sebebiyle diğerlerinden farklı bir şekilde ayrılabilir olduğu belirtilmiştir.

[25] çalışmasında Hürriyet gazetesinden elde edilen veriler ile köşe yazarlarının vektörleri oluşturularak profilleri çıkarılmıştır. Paragraf Vektör yönteminin kullanıldığı çalışmada, paragraf vektör modellerinden PV-DM ve PV-DBOW modellerinin performansları karşılaştırılmıştır. Yazarların birbirinden ayrılmasında PV-DM modelinin daha başarılı olduğu çıkarımı yapılmıştır. Ayrıca birbirinden ayrılmaya çalışılan yazar sayılarının artışı ile model performanslarının düştüğü gözlemlenmiştir.

### 2.3 Metin Sınıflandırmanın Kullanıldığı Farklı Uygulama Alanları

Tıbbi dokümanların hastalıklarına göre sınıflandırıldığı çalışmada ise [26] sınıflandırma verisi olarak farklı tıbbi dergilerdeki dokümanları içeren bibliyografik bir veritabanı olan MEDLINE'nı kullanılmıştır. Doküman sayısı en yüksek olan 10 farklı hastalık sınıfı kullanılmıştır. Bu hastalıkları ayırt etmede Bayes, Karar ağacı ve Rastgele Orman algoritması olmak üzere 3 farklı yöntemin performansları kıyaslanmıştır. En başarılı sonuçları ise %68 lik oranla Bayes ile sınıflandırılan sonuçlarda alınmıştır.

Metin madenciliği tekniği ile e-ticaret sitelerinin belirlendiği [27] çalışmada kullanıcıların yapılan aramalarda e-ticaret sitelerine kolayca ulaşmaları hedeflenmiştir. K-NN ve Naïve Bayes algoritmalarının kullanıldığı çalışmada elde edilen sonuçlar karşılaştırılmıştır. 273 adet site seçilip bu sitelerin içinde geçen seçilen anahtar kelimelerle her bir sitenin vektörü oluşturulmuştur. Anahtar kelimelerin

seçiminde kullanım sıklığı göz önünde bulundurulmuş ve bu kelimeler %10 altındaysa bu kelimeler değerlendirilmemiştir. Toplam 110 kelime ile e-ticaret sitelerinin vektörleri oluşturulmuştur. Yazarlar Naïve Bayes algoritmasında %85 lik bir başarım sağlarken, k-NN en yakın komşu algoritmasında %83 lük bir başarı sağlayarak e-ticaret sitelerini tespit edebilmişlerdir.

[28] çalışmasında Twitter'dan elde edilen verilerin negatif, pozitif ve nötr olmak üzere 3 sınıfta etiketlemesi yapılmıştır. Daha önce yapılmış twitter verileri kullanılarak yapılan duygu analizi çalışmalarından farklı olarak bu çalışmada Hadoop ekosistemi kullanılarak büyük veriler üzerinden paralel işleme performansları da değerlendirilmiştir. Benzer çalışmalardan farklı olarak sözcüksel yaklaşım veya makine öğrenmesi yöntemi kullanılmamış; duygu analizi ve denetimli kümeleme (supervised clustring) modeli kullanılmıştır. Bu yöntemler konfigürasyonları farklı 3 farklı Hadoop makinesinde karılaştırılmış makinelerdeki ram artırışının performansı olumlu etkilediği gözlemlenmiştir. Ayrıca scoring ve K-means algoritmalarının performansları örneklerin ve sözlüğün düzgün seçilmesi ile artış gösterdiği belirtilmiştir.

[29] Çalışmasında sosyal medyanın kullanımının ve alanının artmasıyla işlenecek verinin büyüklüğüne dikkat çekilmiştir. Büyük bir veriye dönüşen sosyal medya içeriğinin efektif bir şekilde analizinin yapılması için Büyük Veri (Big Data) teknolojisinin gerekli olduğu ifade edilmiştir. Hızla büyüyen sosyal medya verisinin analizi problemi çözümü için geliştirilen farklı Büyük Veri mimarilerinin performansları kıyaslanmıştır.

İnsanların soru sorduğu ve diğer kişilerden cevap aldığı bir ortam olan Quora üzerinde yapılan [30] çalışmasında ise, soru başlıklarının mükerrer olup olmadığı tahminlemesi yapılmıştır. Tf-idf, word2vec yöntemlerinin özellikleri çıkartılmış ve bu özellikler açık kaynaklı gradyan artırma (Gradient Boosting) mekanizması olan XGBoost üzerinde çalıştırılmıştır. %90 lık bir oranla başarı elde edilmiştir. Kaggle yarışmasında diğer

araçlara göre (Python, H2O, R, Spark) çok iyi sonuç alınan XGBoost modelinin performansı yeni eklenen özelliklerle arttığı gözlemlenmiştir.
Veri seti olarak hastane kayıtlarının kullanıldığı [31] çalışmasında word2vec ve tf-idf yaklaşımları bir arada kullanılmış ve hastadan alınan belirtilerin ateşli hastalık olup olmadığı tespiti yapılmıştır. Tf-idf, word2vec ve word2vec tf-idf yönteminin beraber kullanımının performansı karşılaştırılmıştır. Tf-idf ile beraber kullanımda sınıflandırma başarımının bir miktar arttığı gözlemlenmiştir.

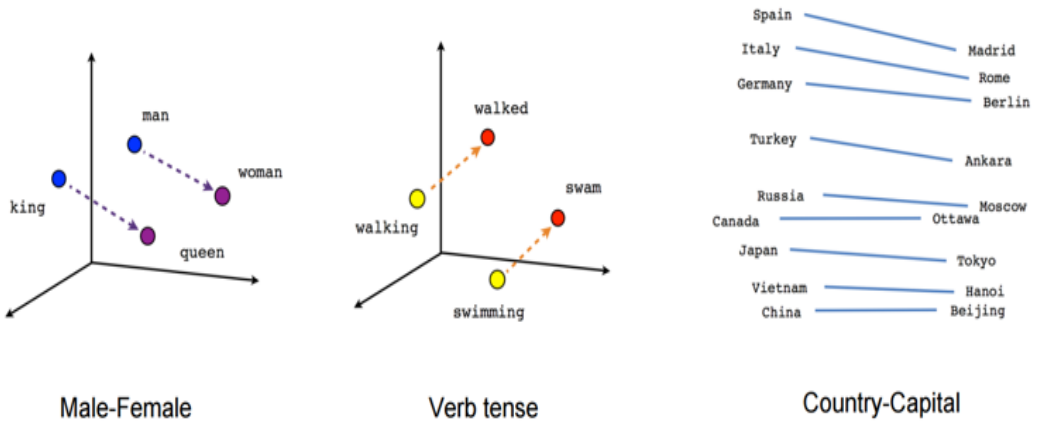
Bu çalışmamızda yazarların vektörleri 2014 yılında word2Vec genişletilmiş versiyonu olan doc2Vec yöntemi ile oluşturulmuş, gerçek zamanlı olarak yazar benzerlik tahminleri başarılı bir şekilde yapılmıştır. Çalışmamızda ayrıca yazarların yazmış oldukları konular, gazete ve Ekşi Sözlük kaynaklarından elde edilen veriler ile modelleri oluşturulmuş başarılı bir şekilde sınıflandırılmıştır. Sonraki bölümde geliştirdiğimiz model hakkında detaylı bilgiler verilecektir.

## 3. ÖNERİLEN YAKLAŞIM

Sinir ağlarını oluşturmak ve eğitmek için geliştirilmiş Java tabanlı bir kütüphane olan DL4J, hızlı bir şekilde geliştirme yapmaya olanak sağlayan açık kaynak kodlu bir kütüphanedir. DL4J bilimsel hesaplamalar yapmak, büyük boyutlu matrisler oluşturmak ve onlar üzerinde hesaplama yapmak için Java İçin Çok Boyutlu Diziler (ND4J - N-Dimensional Arrays for Java) kütüphanesini kullanır. Metin madenciliği ve metin sınıflandırmada güncel ve başarılı yaklaşım olan word2vec ve doc2Vec modelleri DL4J kütüphanesinde yer almaktadır.

### 3.1 Word2Vec

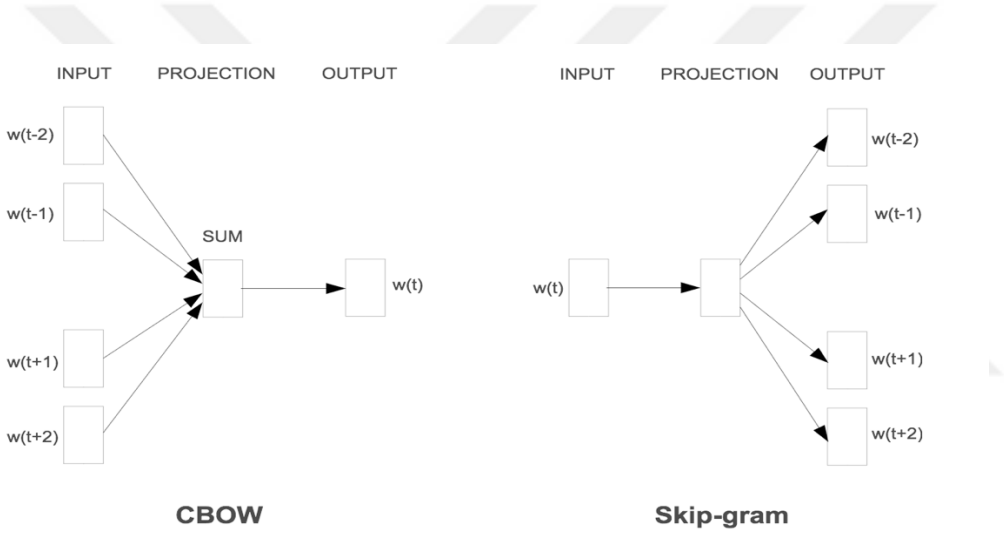
Word2Vec kelimeleri vektör uzayında ifade etmeye çalışan denetimsiz (unsupervised - no labels) ve tahmin temelli (prediction-based) bir modeldir. O dönem Google'da çalışan araştırmacı Tomas Mikolov ve ekibi tarafından 2013 yılında icat edilmiştir.[2]




Şekil 3.1 : Word2Vec kelime ilişkileri[2].

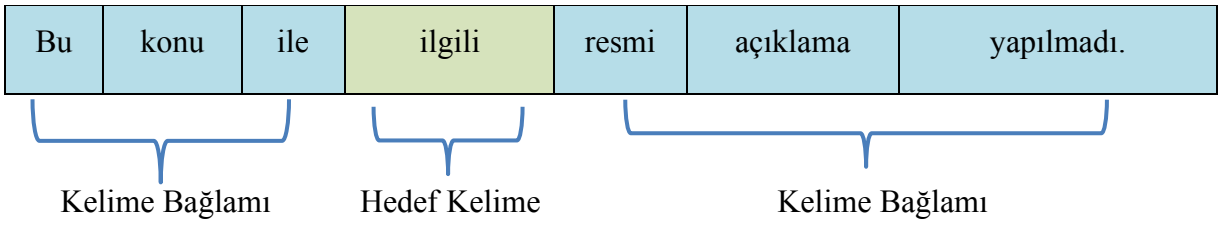
Bu vektör uzayında bir kelimeye yakın en yakın kelimeleri listelemek için kullanılabilir. Kelimelerin cümlede kullanımına göre Şekil 3.1 de görüleceği üzere farklı kelimeler arasında analojiler oluşturabiliyor ve bu vektörlerin birbirine uzaklıklarını hesaplayabilme imkânı sağlıyor. Örneğin; "Erkek" ile "Kadın" arasındaki bağlantı ile "Kral" ve "Kraliçe" arasında bağlantıyı ifade edebiliyor. Aynı şekilde yeteri kadar metin işlediğinde ülkeler ile başkentleri arasında bir bağlantı olduğu çıkarımını yapabiliyor. Word2vec kelimeler arasındaki bağlamı aritmetik işlemler veya fonksiyonlar kullanarak yapılmasına da imkân sağlıyor. Örn:

vec(Spain) – vec(Madrid) = vec(Turkey) – vec(Ankara)




Şekil 3.2 : CBOW ve Atla-gram modelinin grafiksel gösterimi[2]

Word2Vec'te atla-gram (SkipGram) ve Sürekli Kelimeler Çantası (Continuous Bag of Words) olmak üzere Şekil 3.2 görüldüğü gibi iki farklı yaklaşım vardır.




Şekil 3.3 : Pencere boyutu 3 değeri için örnek.

Word2Vec parametrelerinden pencere boyutu (window size) ortadaki kelimenin sağında ve solunda kaç kelime olabileceğini ifade eder. Şekil 3.3 te hedef kelime ve kelime bağlamına örnek gösterilmiştir. CBOW modelinde merkezdeki kelime çevresindeki kelimelerden tahmin edilmeye çalışılırken, atla-gram pencerenin merkezindeki hedef kelimeyi, çevresindeki kelimelerden tahmin etmeye çalışılır. CBOW modeli hedef kelimenin çevresindeki kelimelerden tahminleme yaptığı için, hedef kelimeyle beraber kullanılan kelimeler modelin daha çok ilgisini çeker. Örneğin; "Film gerçekten çok …" cümlesinde CBOW modeli "kötüydü" ya da "iyiydi" tahminlemesi yaparken, "sıkıcıydı" veya "korkunçtu" kelimeleri "iyi" ve "kötü" kelimelerinden daha az sıklıkta kullanıldığı için model tarafından tahminleme olasılığı düşüktür.

CBOW, atla-gram (skip-gram) göre çok daha hızlıdır ve büyük veriler için daha uygundur. Fakat atla-gram CBOW a göre tahminlemede daha iyi performans gösterir [2].

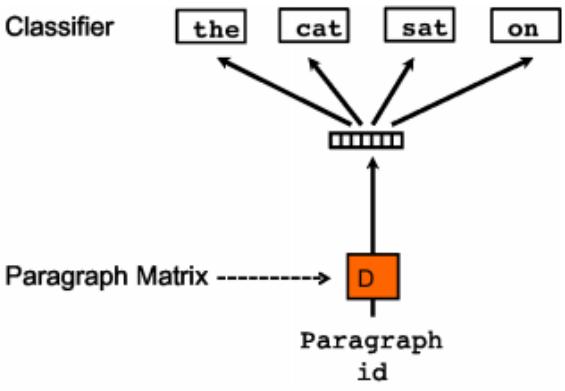
$$sim(A,B) = \cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|} \quad (1)$$

Word2Vec kelimeler arasında ki ilişkiyi hesaplamak için kosinüs benzerliği kullanır(1)[32].

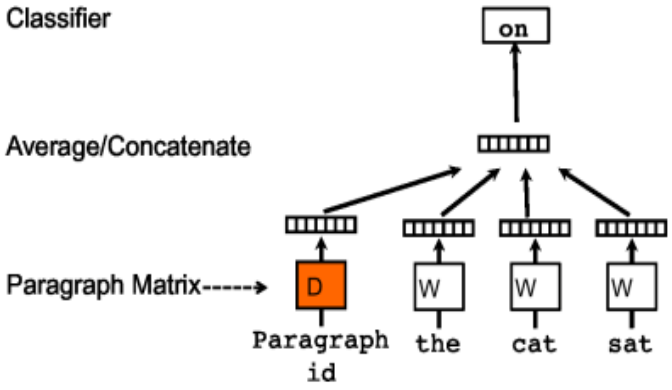
### 3.2 Doc2Vec

Doc2Vec diğer bir isimlendirme ile Paragraf Vektör, word2Vec gibi denetimsiz (unsupervised) öğrenme temelli bir algoritmadır. Word2Vec te olduğu gibi iki farklı model ile oluşur. Bu iki model Paragraf Vektör Dağıtılmış Kelimeler Çantası (Paragraph Vector Distrubuted Bag of Words) (PV-DBOW) ve Paragraf Vektör Dağıtılmış Bellek Modeli (Paragraph Vector Distrubuted Memory Model) (PV-DM) dir. Şekil 3.4 ve Şekil 3.5 te görüleceği üzere PV-DM modeli CBOW a benzerken PV-

DBOW modeli ise atla-grama(skip-gram) benzerlik gösterir. PV-DM ve PV-DBOW modellerinde CBOW ve atla-gramdan farklı olarak tahminleme için model bağlam kelimelerle beraber paragraf id değerini de kullanır.



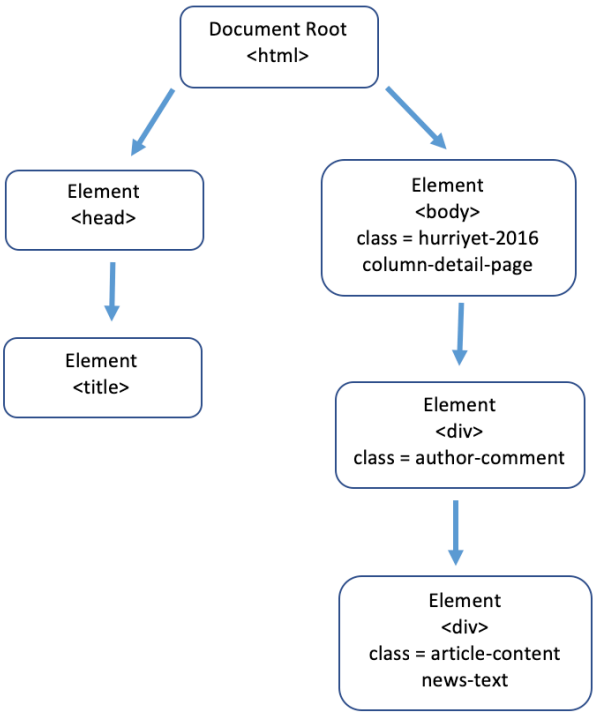

Şekil 3.4 : PV-DBOW Model[2].




Şekil 3.5 : PV-DM Model[2].

### 3.3 Jsoup

İnternet tarayıcıları girilen internet sayfalarını bir belge olarak kabul eder, içine girilen tüm elemanları ise bir nesne olarak kabul ederek Belge Nesne Modeli ile Şekil 3.6 daki gibi bir hiyerarşi yapısında gösterir.



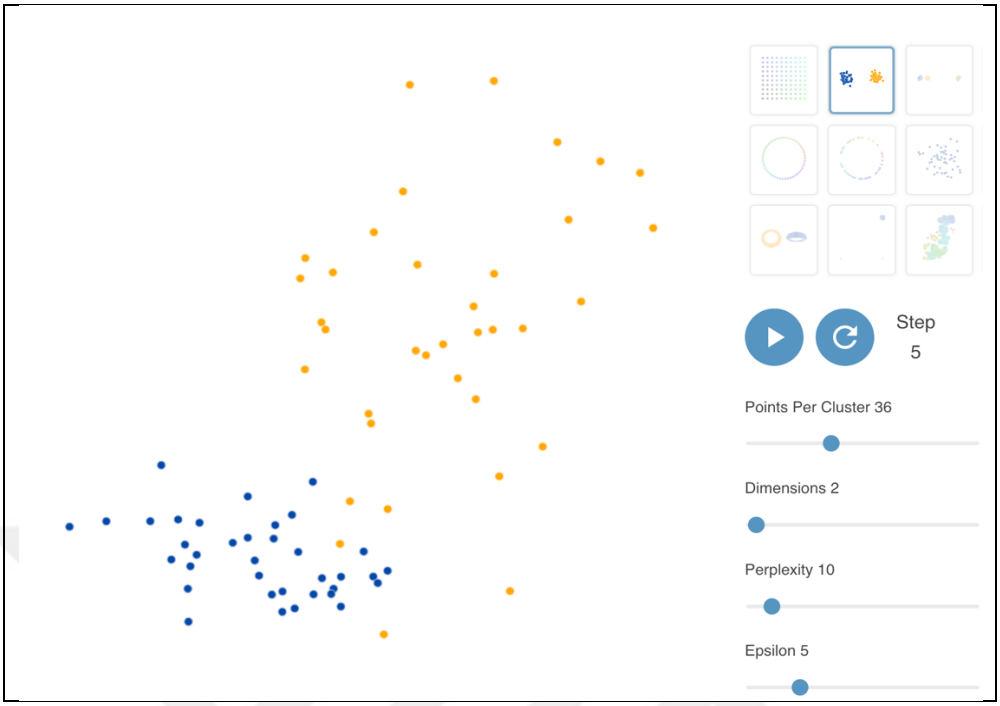

Şekil 3.6 : Html nesneleri örnek hiyerarşi.

Bir web sitesinden veri elde etmek için ise bu Belge Nesne Modeli'ni kullanarak istediğimiz nesneleri ayrıştırıp, istenilen içeriğe ulaşabiliriz. Bu işlemi kolaylaştırmak için Java da çeşitli kütüphaneler yazılmıştır. Jsoup da onlardan biridir. Jsoup, HTML Belge Nesne Modelini kullanarak ya da CSS seçiciciler ile istenilen nesneye ulaşılır [33][34].

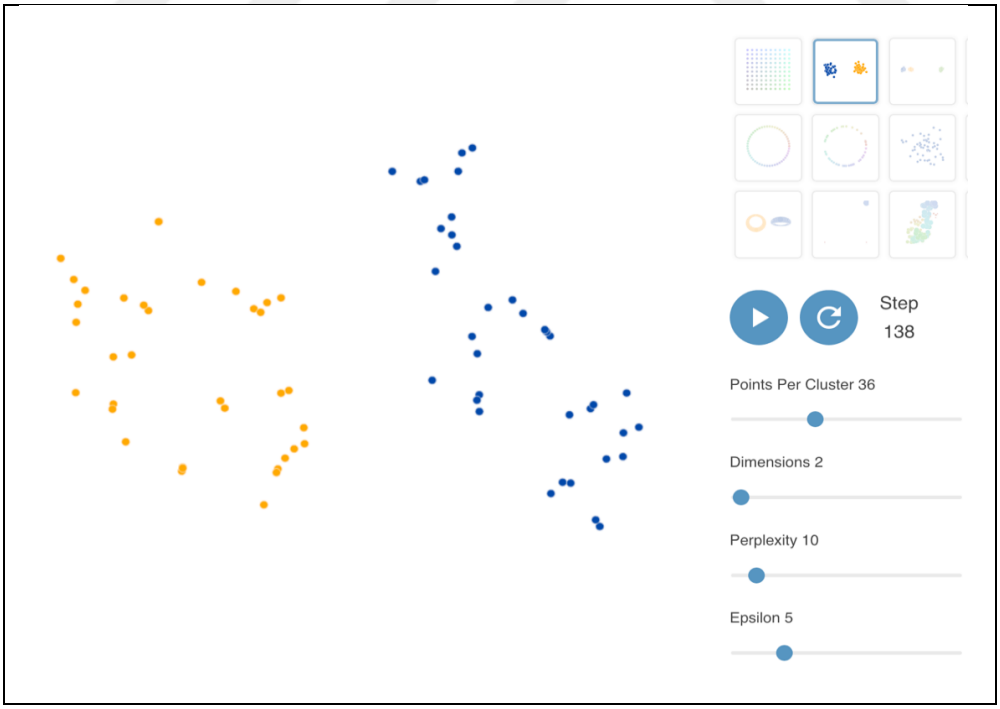
### 3.4 T-SNE

İnternet üzerinde gün geçtikçe artan ve karmaşıklaşan bir veri kümesi bulunmaktadır. Bu karmaşık veriyi çok boyutlu olarak sınıflandırabiliyoruz ve gün geçtikçe verinin boyut sayısı da artmaktadır.

Çok fazla boyutlu sınıflandırmalar bilgisayar için çok problem olmasa da insan en fazla 3 boyutu algılayabildiği için sınıflandırmayı görsel olarak algılayabilmede zorluk çekebilmektedir.




Şekil 3.7 : T-sne algoritması ile adım sayıısna bağlı oluşan iki boyutlu şekil [36]




Şekil 3.8 : T-sne algoritması ile adım sayıısna bağlı oluşan iki boyutlu şekil [36].


Şekil 3.9 : T-sne algoritması ile adım sayısına bağlı oluşan iki boyutlu şekil [36]

İnsanın algılayamayacağı çok boyutlu bir veriyi 2 veya 3 boyuta düşürmemiz Laurens van der Maaten ve Geoffrey Hinton tarafından ortaya konulan T-dağıtımlı komşu yerleştirme (T-distributed stochastic neighbor embedding - t-SNE) algoritması ile mümkün olabilmektedir[35]. Bu yöntemde farklı hiper parametrelere göre farklı 2 boyutlu grafikler elde edilebilmektedir. Bu nedenle bu parametrelerin optimize edilmesi gerekmektedir. Şekil 3.7, Şekil 3.8 ve Şekil 3.9'da aynı veriden farklı hiperparametrelerle oluşturulmuş 2 boyutlu grafik örneği paylaşılmıştır. Grafiklerden görüleceği üzere farklı adım sayılarında 3 boyutlu düzlemin 2 boyuttaki görünümü farklı olmaktadır.

# 4. SONUÇLAR VE ÖNERİLER

Bu bölümde ilk olarak metinlerin analizinde kullanılan performans metriklerinden bahsedilmiştir. Bu metriklerin değerlendirildiği veri kümelerinin ve metin vektörlerinin nasıl oluşturulduğu anlatılmıştır. Ayrıca bu veri kümelerinin analizi ile ilgili elde edilen sonuçlar paylaşılmıştır. Analizi yapılan vektörlerden oluşturulan uygulamadan bahsedilmiştir.

## 4.1 Performans Metrikleri

Sonuçlar elde edildikten sonraki aşama ise test verileri kullanılarak sonuçların başarısının değerlendirilmesidir. Başarımı değerlendirilmesi için doğruluk oranı (Accuracy), kesinlik (Precision), duyarlılık (Recall) ve F-Ölçütlerinden faydalanılır. Sonuçları değerlendirmek için sonuçlar oluşturulan hata matrisi (Confussion matrix) olarak isimlendirilen çizelge yapısıyla gösterilir [38]. Hata matrisi 4 kategoride oluşturulur: doğru pozitif (DP) (True positive), yanlış pozitif(YP) (false positive), yanlış negatif (YN) (false negative), doğru negatif (DN) (True negative). Doğru pozitif (DP) pozitif sınıfta olan örneklerin pozitif olarak doğru sınıflandırıldığını gösterir. Yanlış pozitif (YP) negatif sınıfta olan örneklerin pozitif olarak sınıflandırıldığını gösterir. Doğru negatif (DN) negatif sınıfa ait elemanların negatif olarak sınıflandırıldığı anlamına gelir. Yanlış negatif ise pozitif sınıfa ait elemanların yanlışlıkla negatif sınıfa dahil edilmesi anlamına gelmektedir.

| **Bulunduğu Sınıf/Tahmin Edilen** | **Pozitif** | **Negatif** |
|---|---|---|
| **Pozitif** | DP | YN |
| **Negatif** | YP | DN |

Çizelge 4.1 : İki sınıf için oluşturulan örnek hata matrisi

Hassaslık gerçek değeri pozitif olup pozitif değere sınıflandırılan sayısının, pozitif değere sınıflandırılanların toplamına oranıdır.

$$Hassaslık = \frac{(TP)}{(TP + FN)} \tag{1}$$

Duyarlılık, gerçek değeri pozitif olup pozitif değere sınıflandırılan sayısının, gerçek değeri pozitif olanların tümüne oranıdır.

$$Duyarlılık = \frac{(TP)}{(TP + FP)} \tag{2}$$

Yapay öğrenme uygulamalarında genellikle kullanılan ölçütlerden olan doğruluk oranı, doğru olarak sınıflandırılan örnek sayısının toplam örnek sayısına oranıdır. Doğruluk oranının hata matrisleri kullanılarak hesaplanması Eşitlik 3'de gösterilmiştir.

$$Doğruluk\ Oranı = \frac{(TP + TN)}{(TP + FP + FN + TN)} \tag{3}$$

Veri kümesinin dengesiz dağıldığı durumlarda, doğruluk oranını başarım ölçütü olarak değerlendirmek yanıltıcı olabilmektedir. Bu nedenle bir diğer değerlendirme kıstası olan F-Ölçütü performansı değerlendirmek için kullanılabilir. Kesinlik ve duyarlılık ölçütlerinin ağırlıklı harmonik ortalaması alınarak oluşturulan F-Ölçütü başarımı değerlendirmek için daha sağlıklı sonuçlar vermektedir.

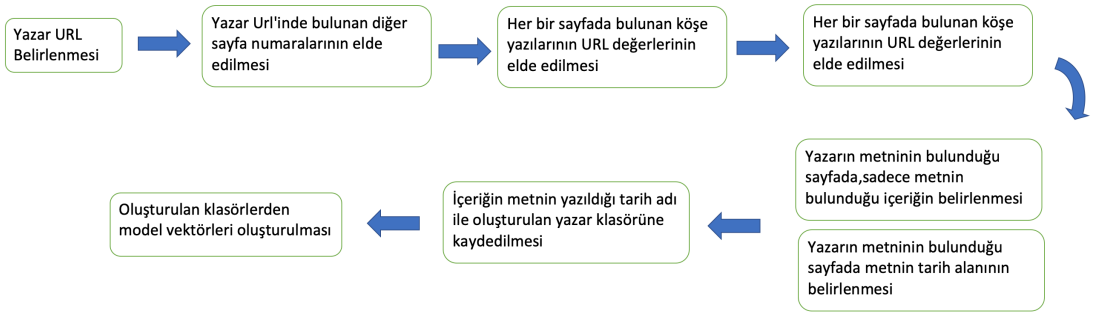
$$F - \text{Ölçütü} = \frac{2 * Kesinlik * Duyarlılık}{Kesinlik + Duyarlılık} \tag{4}$$

## 4.2 Hürriyet Yazar Tespiti

Hürriyet gazetesinde 40'a yakın köşe yazarı çalışmaktadır ve her bir yazarın 400 ila 8000 arasında yazısı bulunmaktadır. Kimi yazarın iki senelik köşe yazısı arşivi bulunurken, on seneden fazla çalışan yazarların ise bir hayli köşe yazısı bulunmaktadır. İyi bir model ortaya koymak için 1000'den fazla köşe yazısı olan yazarları dikkate aldık [39].

### 4.2.1 Veri kümesi

Hürriyet gazetesi web sitesinden elde etmek istediğimiz yazarların metinlerini hazırladığımız uygulama ile elde ettik ve Şekil 4.1 deki yolu izledik. Uygulamamızda ilk olarak yazar URL'lerini belirledik. Yazarların yazılarının bulunduğu sayfada birden fazla sayfa bulunduğu için her bir sayfanın numarasının URL değerini elde etmek için her sayfayı açıp "title" sınıf değişkeni olup olmadığını kontrol ettik.
Açılan sayfalarda yazarın yazmış olduğu yazıları verilen bağlantılar ile elde ettik.
Yazarın yazısının bulunduğu sayfada "news-detail-text" sınıf adıyla bulunan metin içeriğini kaydettik. Yine aynı sayfada bulunan "news-info-date" sınıfından çekilen yazının yazılmış olduğu tarih parametresi yazarın metnine ekledik. Kaydetme aşamasında yazıda bulunan HTML etiketleri gibi metin ile ilgisiz içerikleri eledik [40].




Şekil 4.1 : Köşe yazılarının içeriklerinin elde edilmesi.

Kaydettiğimiz dosyaya isim olarak "news-info-date" sınıfından aldığımız tarih bilgisini verdik. Fakat tarih bilgisinde bulunan ay yazı ile belirtildiği için (Örn. "20

Aralık 2018") içinde bulunan Türkçe karakterlerin uygulama tarafında dosya okuma sırasında sorunlara yol açtığını tespit ettik. Bu yüzden dosya ismindeki Türkçe karakterleri değiştirerek kaydettik ("ğ" harfini "g", "ü" harfini "u", "ı" harfini "i" gibi). Her yazarın yazısını daha sonra DL4J kütüphanesi ile modelinin oluşturulması için klasörledik.

### 4.2.2 Sonuçlar ve yorumlanması

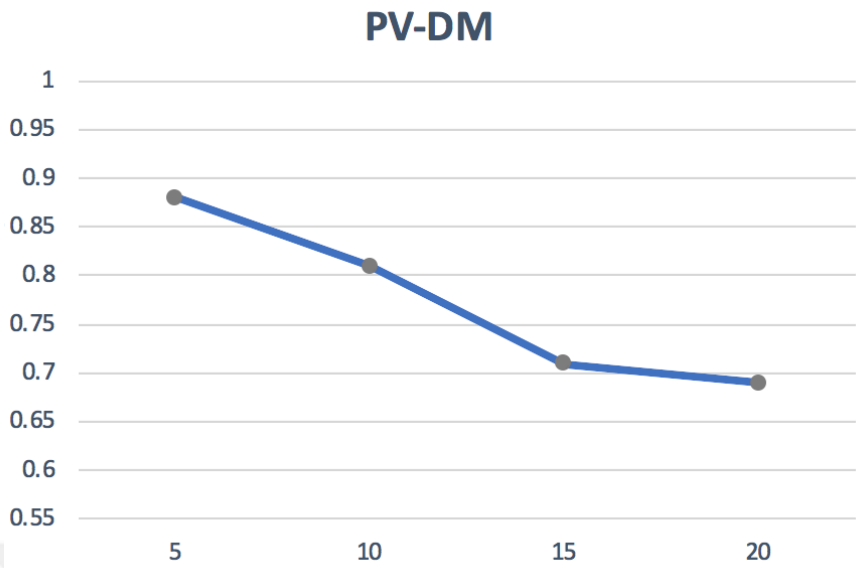
Her bir yazardan eğittiğimiz modelin içinde olmayan ve rastgele zamanlardan seçilmiş 20 yazıyı ise test için kullandık.

Seçtiğimiz rastgele 20 yazıdan, her bir yazara ait 1000 yazıdan ve toplam 20000 yazıdan oluşan modelde tahminlemelerde bulunduk. Windows size 5 olarak tuttuk ve her bir yazardan 5, 10, 15 ve 20'lik setler aldık ve bu değişken sette PV-DM ve DV-DBOW modellerinin başarımını karşılaştırdık.

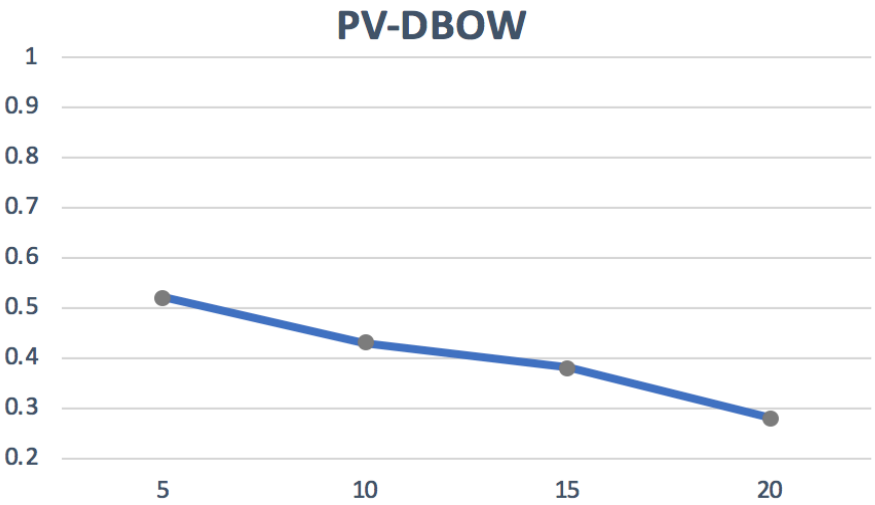
Seçtiğimiz yazının diğer yazarların yazılarına olan yakınlıklarını kıyasladık ve kosinüs benzerliği en yakın yazarı tahmin edildi olarak işaretledik. Performans analizi yapmak için F-Ölçütünü kullandık.

Bu analiz sonucunda PV-DM modeli PV-DBOW modeline göre çok daha iyi başarım göstermiştir. Bu başarımın altında yatan neden ise PV-DM modelinin kelimenin dizilişini dikkate alırken, PV-DBOW modeli almamaktadır.

Şekil 4.2 ve Şekil 4.3 te görüleceği üzere PV-DM ve PV-DBOW modellerinin performansı yazar sayısında artışa bağlı olarak düşüş göstermiştir. PV-DM yönteminin 5 yazarlı bir uzayda gösterdiği performans 0,88 iken PV-DBOW yönteminin gösterdiği performans 0,69 olmuştur. Yazar sayısı 20'ye çıktığında ise 0,52 ve 0,28 olmuştur.




Şekil 4.2 : Yazar sayısı artışına bağlı PV-DM başarı grafiği

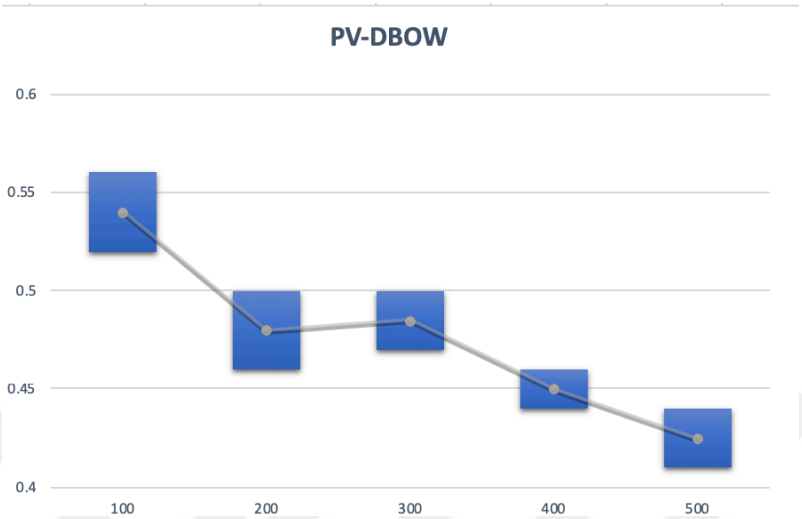



Şekil 4.3 : Yazar sayısı artışına bağlı PV-DBOW başarı grafiği

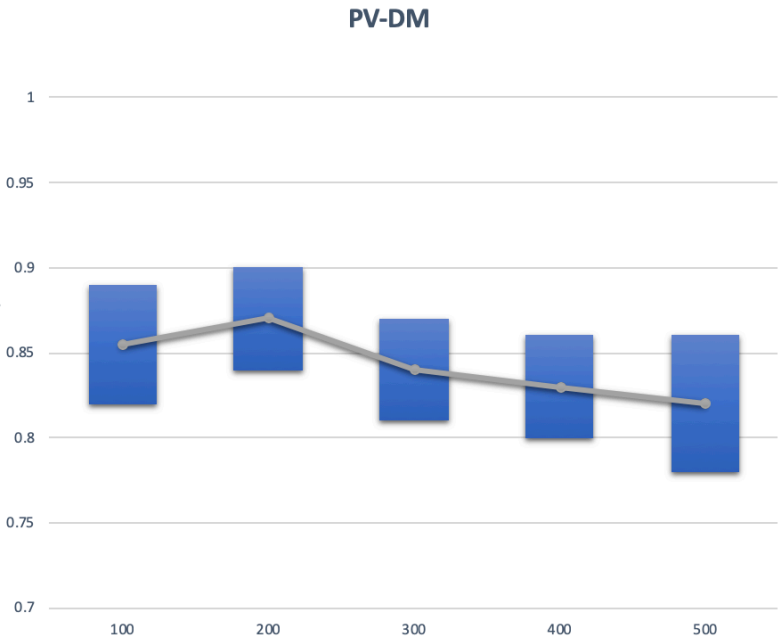
| *Yazar sayısı* | *PV-DM* | *PV-DBOW* |
|---|---|---|
| 5 | 0.88 | 0.52 |
| 10 | 0.81 | 0.43 |
| 15 | 0.71 | 0.38 |
| 20 | 0.69 | 0.28 |

Şekil 4.4 : Yazar sayısı artışına bağlı modellerin karşılaştırılması

DL4J hiperparametrelerinden olan katman sayısı (layer size) değerinin sınıflandırmadaki performansını karşılaştırmak için yaptığımız çalışmada her bir yazara ait 500 farklı köşe yazısı eğitim seti olarak seçilmiştir.




Şekil 4.5 : Katman boyutuna göre PV-DBOW performası




Şekil 4.6 : Katman boyutuna göre PV-DM performansı.

Toplam 6 farklı yazarın seçildiği bu kümede, test için 20 farklı köşe yazısı kullanılmıştır. Her bir katman (layer) boyutu için 5 deneme yapılmış ve bu denemelerde test için kullanılan 20 yazı rastgele seçilmiştir. Elde edilen sonuçlarda

seçilen pencere boyutu ile elde edilen başarımın maksimum, minimum ve ortalama değerleri çıkartılmıştır. PV-DM metodu PV-DBOW metoduna her katman boyutuna göre çok daha başarılı olmuştur. PV-DM metodunda en başarılı sonuç katman (layer) boyutu 200 olarak uygulandığında alınmıştır. Ayrıca katman (layer) boyutu arttıkça modelin bellekte kapladığı alan artarken, modeli oluşturma süreside gittikçe artmıştır. Yazarların hata matrislerinde (confussion matrix) ise benzer konularda yazan yazarların birbirine karıştırıldığı gözlemlenmiştir. PV-DBOW modelinde Cengiz Semercioğlu magazin alanında ve gündem alanında yazdığı yazılarda kullandığı kelimeler, özellikle Onur Baştürk ile karıştırılmasına neden olmuştur. PV-DBOW modelinde kelime bağlamını değerlendirirken bağlamda kullanılan kelimelerin tekrar sıklığı daha önemli olduğu için Çizelge 4.2'de Onur Baştürk'e ait 14 yazı, model tarafından Cengiz Semercioğlu'na ait olduğu tahminlenmiştir. Beş yazardan oluşturulan PV-DM modelinde ise Çizelge 4.3 da görüleceği üzere bu iki yazarın yazılarının rahatlıkla ayırt edilebildiğini görülmüştür. Bu modelde bağlamdaki sık olarak kullanılmayan kelimeleri de tahminlemesindeki başarısı ile sadece bir tane Onur Baştürk'e ait yazı Cengiz Semercioğlu olarak tespit edilmiştir. Karıştırılan bu yazıda ise Onur Baştürk ile Cengiz Semercioğlu'nun benzer gündemle ilgili yazısında benzer bağlamlar kullanmasından kaynaklandığı çıkarımı yapılmıştır. Beş yazardan oluşturulan PV-DM modelinde ise en çok karıştırılan Ayşe Arman'ın yazısı ve Cengiz Semercioğlu yazısı olmuştur, toplamda 5 Ayşe Arman yazısının Cengiz Semercioğlu olarak tahminlendiği modelde karıştırılan yazıların ortak özellikleri olarak benzer gündem ve benzer bağlam kelimelerinden (contex words) kaynaklanmaktadır.

Bir yazının hangi yazara ait olduğunu, modelde bulunan yazarlar arasından kosinüs benzerliği en çok hangisine yakınsa, yazıyı yakın olan yazarın yazısı olarak işaretlendiği belirtilmişti. 10 yazardan oluşan modelde, 5 li modellerden farklı olarak yeni eklenen yazarların kullandığı kelimeler ve bağlamlar yazıların benzetildiği yazarları da değiştirmiştir. On yazardan oluşturulan PV-DBOW modelinde (Çizelge 4.5) ise yazarların birçoğunun yazısı Ertuğrul Özkök olarak bulunmuştur. Ertuğrul Özkök yazdığı metinlerde gündem magazin gibi alanlarda birçok konuya değinmiş ve bağlamda kullandığı kelimelerin sıklığı diğer yazarlarla karıştırılmasına neden

olmuştur. Sadece bir yazısı tahmin edilebilen Kanat Atkaya'nın ise bağlamda kullandığı kelimeler PV-DBOW modeli ile yazarın profilinin oluşturulmasında yetersiz kalmıştır. On yazardan oluşturulan PV-DM modeline baktığımızda ise modelin genel başarımı iyi olduğu gözlenirken, benzer gündemlerde olan farklı yazarların birkaç yazısının karışıtırıldığı gözlemlenmiştir.

15 Yazardan oluşturulan modellerde de 10 yazardan oluşturulan PV-DBOW modeli benzerlik göstermiştir. Çizelge 4.8'de değerlendirilen yazar sayısının artmasıyla bağlamda kullanılan kelime sıklığının ön plana çıktığı PV-DBOW modelinde yazıların birçoğu model tarafından Ertuğrul Özkök etiketi ile işaretlenmiştir. PV-DM modelinde ise 5 ve 10 yazarlı PV-DM modellerinde olmayan Mehmet Y. Yılmaz yazarının eklenmesiyle Ertuğrul Özkök ve yazarının tahminleme başarımı azalmıştır. Ertuğrul Özkök'ün 8 farklı yazısı PV-DM modeli tarafından Mehmet Y. Yılmaz olarak işaretlenmiştir. Mehmet Y. Yılmaz'ın bağlamdaki kullanılan kelimelerin benzerlikleri ile ayırt edilmesinin güçleştiği gözlemlenmiştir.

20 Yazardan oluşturulan modelde ise (Çizelge 4.7 ve Çizelge 4.9), Gökhan Kimsesizcan otel isimleri gibi özel isimler ve yabancı kelimeleri sık kullanmasıyla PV-DBOW modelinde başarılı bir şekilde ayırt edilebilmiştir. PV-DBOW ile 15 yazarla oluşturulan modelde Ertuğrul Özkök ile karıştırılan Yaşar Sökmensüer, 20 yazarla oluşturulan modelde kullanılan kelimelerin bağlamının diğer yazarlar ile benzeşmesi sebebiyle diğer yazarları tahminlemede en çok karıştırılan yazar olmuştur.

Çizelge 4.2 : Beş yazarlı PV-DBOW sınıflandırma.

a=Ayşe Arman b=Cengiz Semercioğlu c= Onur Baştürk
d=Sahrap Soysal e=Vahap Munyar

| | a | b | c | d | e |
|---|---|---|---|---|---|
| a | 0 | 7 | 0 | 13 | 0 |
| b | 0 | 19 | 0 | 1 | 0 |
| c | 0 | 14 | 1 | 5 | 0 |
| d | 0 | 0 | 0 | 20 | 0 |
| e | 0 | 3 | 0 | 5 | 12 |

Çizelge 4.3 : Beş yazarlı PV-DM modeli ile sınıflandırma.

a=Ayşe Arman b=Cengiz Semercioğlu c= Onur Baştürk
d=Sahrap Soysal e=Vahap Munyar

| | ***a*** | ***b*** | ***c*** | ***d*** | ***e*** |
|---|---|---|---|---|---|
| a | 11 | 5 | 3 | 1 | 0 |
| b | 0 | 18 | 2 | 0 | 0 |
| c | 0 | 1 | 19 | 0 | 0 |
| d | 0 | 0 | 0 | 20 | 0 |
| e | 0 | 0 | 0 | 0 | 20 |

Çizelge 4.4 : On yazarlı PV-DM Modeli ile sınıflandırma.

a=Ayşe Arman b= Cengiz Semercioğlu c=Doğan Hızlan d=Erdal Sağlam
e =Ertuğrul Özkök f=Kanat Atkaya g=Mehmet Y. Yılmaz h=Onur Baştürk
i=Sahrap Soysal j=Vahap Munyar

| | ***a*** | ***b*** | ***c*** | ***d*** | ***e*** | ***f*** | ***g*** | ***h*** | ***i*** | ***j*** |
|---|---|---|---|---|---|---|---|---|---|---|
| ***a*** | 13 | 4 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 1 |
| ***b*** | 0 | 20 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| ***c*** | 0 | 1 | 19 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| ***d*** | 0 | 0 | 3 | 15 | 0 | 0 | 0 | 0 | 0 | 2 |
| ***e*** | 1 | 1 | 0 | 0 | 15 | 0 | 0 | 1 | 1 | 1 |
| ***f*** | 2 | 4 | 0 | 0 | 1 | 10 | 1 | 2 | 0 | 0 |
| ***g*** | 0 | 3 | 0 | 0 | 2 | 0 | 15 | 0 | 0 | 0 |
| ***j*** | 0 | 3 | 0 | 0 | 0 | 0 | 0 | 17 | 0 | 0 |
| ***i*** | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 19 | 0 |
| ***j*** | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 19 |

Çizelge 4.5 : On yazarlı PV-DBOW Modeli ile sınıflandırma.

a=Ayşe Arman b= Cengiz Semercioğlu c=Doğan Hızlan d=Erdal Sağlam e =Ertuğrul Özkök f=Kanat Atkaya g=Mehmet Y. Yılmaz h=Onur Baştürk i=Sahrap Soysal j=Vahap Munyar

| | ***a*** | ***b*** | ***c*** | ***d*** | ***e*** | ***f*** | ***g*** | ***h*** | ***i*** | ***j*** |
|---|---|---|---|---|---|---|---|---|---|---|
| ***a*** | 0 | 0 | 0 | 0 | 20 | 0 | 0 | 0 | 0 | 0 |
| ***b*** | 0 | 0 | 0 | 0 | 20 | 0 | 0 | 0 | 0 | 0 |
| ***c*** | 0 | 0 | 13 | 0 | 7 | 0 | 0 | 0 | 0 | 0 |
| ***d*** | 0 | 0 | 0 | 15 | 4 | 0 | 1 | 0 | 0 | 0 |
| ***e*** | 0 | 0 | 0 | 0 | 20 | 0 | 0 | 0 | 0 | 0 |
| ***f*** | 0 | 0 | 0 | 0 | 19 | 1 | 0 | 0 | 0 | 0 |
| ***g*** | 0 | 0 | 0 | 0 | 17 | 0 | 3 | 0 | 0 | 0 |
| ***h*** | 0 | 0 | 0 | 0 | 18 | 0 | 0 | 2 | 0 | 0 |
| ***i*** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 20 | 0 |
| ***j*** | 0 | 0 | 0 | 1 | 7 | 0 | 0 | 0 | 0 | 12 |

Çizelge 4.6 : Onbeş yazarlı PV-DM Modeli ile sınıflandırma.

a=Ayşe Arman b= Cengiz Semercioğlu c= Doğan Hızlan d = Erdal Sağlam e=Ertuğrul Özkök f=Gila Benmayor g=Güzin Abla h=Kanat Atkaya i=Mehmet Y. Yılmaz j=Niobe Aslı Temel k=Onur Baştürk l= Osman Müftüoğlu m=Sahrap Soysal n=Vahap Munyar o=Yaşar Sökmensüer

| | ***a*** | ***b*** | ***c*** | ***d*** | ***e*** | ***f*** | ***g*** | ***h*** | ***i*** | ***j*** | ***k*** | ***l*** | ***m*** | ***n*** | ***o*** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ***a*** | 4 | 3 | 0 | 0 | 0 | 0 | 1 | 0 | 2 | 0 | 9 | 1 | 0 | 0 | 0 |
| ***b*** | 0 | 17 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 | 0 | 0 | 0 | 0 |
| ***c*** | 0 | 1 | 19 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| ***d*** | 0 | 0 | 0 | 18 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 0 |
| ***e*** | 1 | 1 | 0 | 0 | 5 | 2 | 0 | 0 | 8 | 0 | 0 | 0 | 1 | 0 | 2 |
| ***f*** | 0 | 0 | 4 | 2 | 1 | 10 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 2 | 0 |
| ***g*** | 0 | 0 | 0 | 0 | 0 | | 17 | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 0 |
| ***h*** | 0 | 5 | 1 | 0 | 1 | 0 | 0 | 5 | 4 | 0 | 4 | 0 | 0 | 0 | 0 |
| ***i*** | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 17 | 0 | 1 | 0 | 0 | 0 | 0 |
| ***j*** | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 15 | 1 | 2 | 0 | 0 | 0 |
| ***k*** | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 19 | 0 | 0 | 0 | 0 |
| ***l*** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 20 | 0 | 0 | 0 |
| ***m*** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 20 | 0 | 0 |
| ***n*** | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 18 | 0 |
| ***o*** | 0 | 0 | 2 | 0 | 4 | 0 | 0 | 3 | 1 | 0 | 1 | 0 | 0 | 0 | 9 |

Çizelge 4.7 : Yirmi yazarlı PV-DM modeli ile sınıflandırma.

a= Abdulkadir Selvi b=Ahmet Hakan c=Ayşe Arman d = Cengiz Semercioğlu e=Doğan Hızlan f=Erdal Sağlam g=Ertuğrul Özkök h=Fatih Çekirge i=Gila Benmayor j=Gökhan Kimsesizcan k=Güzin Abla l=Kanat Atkaya m=Mehmet Y. Yılmaz n=Niobe Aslı Temel o=Onur Baştürk p=Osman Müftüoğlu q=Sahrap Soysal r=Vahap Munyar s=Yalçın bayer t=Yaşar Sökmensüer

| | ***a*** | ***b*** | ***c*** | ***d*** | ***e*** | ***f*** | ***g*** | ***h*** | ***i*** | ***j*** | ***k*** | ***l*** | ***m*** | ***n*** | ***o*** | ***p*** | ***q*** | ***r*** | ***s*** | ***t*** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ***a*** | 15 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| ***b*** | 1 | 13 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 5 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| ***c*** | 0 | 1 | 8 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 3 | 0 | 2 | 0 | 3 | 1 | 0 | 0 | 0 | 0 |
| ***d*** | 0 | 2 | 0 | 16 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| ***e*** | 0 | 0 | 0 | 1 | 19 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| ***f*** | 0 | 0 | 2 | 0 | 0 | 16 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 0 |
| ***g*** | 2 | 4 | 0 | 0 | 0 | 0 | 9 | 0 | 2 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| ***h*** | 4 | 0 | 0 | 1 | 1 | 0 | 5 | 3 | 1 | 0 | 2 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| ***I*** | 0 | 0 | 0 | 0 | 1 | 2 | 2 | 1 | 10 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 4 | 0 |
| ***j*** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 4 | 0 | 16 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| ***k*** | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 17 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 |
| ***l*** | 0 | 1 | 0 | 0 | 0 | 0 | 2 | 1 | 0 | 0 | 1 | 6 | 3 | 0 | 4 | 0 | 0 | 0 | 1 | 1 |
| ***m*** | 0 | 0 | 0 | 1 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 17 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| ***n*** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 0 | 16 | 0 | 1 | 1 | 0 | 0 | 0 |
| ***o*** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 20 | 0 | 0 | 0 | 0 | 0 |
| ***p*** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 19 | 1 | 0 | 0 | 0 |
| ***q*** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 20 | 0 | 0 | 0 |
| ***r*** | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 16 | 1 | 0 |
| ***s*** | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 17 | 0 |
| ***t*** | 0 | 9 | 0 | 0 | 1 | 0 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 5 |

Çizelge 4.8 : Onbeş yazar PV-DBOW modeli ile sınıflandırma.

a=Ayşe Arman b= Cengiz Semercioğlu c= Doğan Hızlan d = Erdal Sağlam e=Ertuğrul Özkök f=Gila Benmayor g=Güzin Abla h=Kanat Atkaya i=Mehmet Y. Yılmaz j=Niobe Aslı Temel k=Onur Baştürk l= Osman Müftüoğlu m=Sahrap Soysal n=Vahap Munyar o=Yaşar Sökmensüer

| | ***a*** | ***b*** | ***c*** | ***d*** | ***e*** | ***f*** | ***g*** | ***h*** | ***i*** | ***j*** | ***k*** | ***l*** | ***m*** | ***n*** | ***o*** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ***a*** | 0 | 9 | 0 | 0 | 11 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| ***b*** | 0 | 10 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 10 | 0 | 0 |
| ***c*** | 0 | 3 | 3 | 0 | 13 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| ***d*** | 0 | 1 | 0 | 15 | 0 | 0 | 0 | 0 | 4 | 0 | 0 | 0 | 0 | 0 | 0 |
| ***e*** | 0 | 1 | 0 | 0 | 19 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| ***f*** | 0 | 5 | 0 | 0 | 9 | 3 | 0 | 0 | 3 | 0 | 0 | 0 | 0 | 0 | 0 |
| ***g*** | 0 | 4 | 0 | 0 | 14 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 |
| ***h*** | 0 | 5 | 0 | 0 | 13 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 |
| ***i*** | 0 | 1 | 0 | 0 | 7 | 0 | 0 | 0 | 12 | 0 | 0 | 0 | 0 | 0 | 0 |
| ***j*** | 0 | 0 | 0 | 0 | 8 | 0 | 0 | 0 | 0 | 12 | 0 | 0 | 0 | 0 | 0 |
| ***k*** | 0 | 19 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| ***l*** | 0 | 2 | 0 | 0 | 2 | 0 | 0 | 0 | 1 | 0 | 0 | 15 | 0 | 0 | 0 |
| ***m*** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 20 | 0 | 0 |
| ***n*** | 0 | 1 | 0 | 0 | 12 | 1 | 0 | 0 | 3 | 0 | 0 | 0 | 0 | 3 | 0 |
| ***o*** | 0 | 3 | 0 | 0 | 13 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 3 |

Çizelge 4.9 : Yirmi yazar PV-DBOW Modeli ile sınıflandırma.

a= Abdulkadir Selvi b=Ahmet Hakan c=Ayşe Arman d = Cengiz Semercioğlu e=Doğan Hızla Sağlam g=Ertuğrul Özkök h=Fatih Çekirge i=Gila Benmayor j=Gökhan Kimsesizcan k=Güzi l=Kanat Atkaya m=Mehmet Y. Yılmaz n=Niobe Aslı Temel o=Onur Baştürk p=Osman Müfti q=Sahrap Soysal r=Vahap Munyar s=Yalçın bayer t=Yaşar Sökmensüer

| | ***a*** | ***b*** | ***c*** | ***d*** | ***e*** | ***f*** | ***g*** | ***h*** | ***i*** | ***j*** | ***k*** | ***l*** | ***m*** | ***n*** | ***o*** | ***p*** | ***q*** | ***r*** | ***s*** | ***t*** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ***a*** | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 17 | 0 |
| ***b*** | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 19 | 0 |
| ***c*** | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 18 | 0 |
| ***d*** | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 19 | 0 |
| ***e*** | 0 | 0 | 0 | 0 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 17 | 0 |
| ***f*** | 0 | 0 | 0 | 0 | 0 | 12 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 8 | 0 |
| ***g*** | 0 | 0 | 0 | 0 | 0 | 0 | 5 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 15 | 0 |
| ***h*** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 20 | 0 |
| ***i*** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 20 | 0 |
| ***j*** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 20 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| ***k*** | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 18 | 0 |
| ***l*** | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 19 | 0 |
| ***m*** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 20 | 0 |
| ***n*** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 11 | 0 | 0 | 0 | 0 | 9 | 0 |
| ***o*** | 0 | 0 | 0 | 0 | 0 | 0 | 5 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 15 | 0 |
| ***p*** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 14 | 1 | 0 | 5 | 0 |
| ***q*** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 19 | 0 | 1 | 0 |
| ***r*** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 19 | 0 |
| ***s*** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 20 | 0 |
| ***t*** | 0 | 0 | 0 | 0 | 0 | 0 | 5 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 11 | 4 |

### 4.3 Ekşi Sözlük Gönderileri Konu Tespiti

Ekşi sözlükte yazarlar belirli konu başlıkları altında yazabilmektedir. Yazarların gönderilerinin uzunluğu çok değişkenlik gösterebilmektedir. Bazı gönderilerde birkaç paragraf olabilirken, bazı gönderilerde birkaç kelime olabilmektedir. Yeni bir konu için ise yeni bir başlık açabilmektedir. Açılan yeni başlığın spor, siyaset, tarih gibi ilgili olduğu konu diğer yazarlar tarafından teyit edilmektedir. Uygulamada yer alan etiketlere tıklayarak, etiketle ilişkilendirilmiş farklı konu başlıkları

listelenebilmektedir. Yeni açılan etiketlenmemiş başlıklar ilk "başıboşlar" bölümünde yer alıp daha sonra diğer yazarlar tarafından etiketleri girilmektedir.

### 4.3.1 Veri kümesi

Ekşi sözlükte en çok etiketlenen konulardan ekonomi, sağlık, siyaset, spor, tarih, teknoloji konuları seçilmiştir. Bu konuların her birinin uzantısı tespit edilmiştir. Hazırladığımız uygulama ile tespit edilen uzantılardan ilk olarak başlıklar tespit edilmiştir. Başlıkların tespitinde DOM da bulunan "topic-list" sınıf ismi kullanılmıştır. Başlık uzantıları tespit edilmiştir. İlgili başlıkların içinde bulunan gönderiler "content" sınıf ismi ile yer aldığı belirlenmiştir. İçerikler tespit edilen sınıf ismine göre kaydedilmiştir. Kaydedilen her bir içerik elle belirlenmiş olan konusuna göre klasörlenmiştir. Oluşturulan kasörlerden her konuya ait vektörler oluşturulmuştur.

### 4.3.2 Sonuçlar ve yorumlanması

Konu bazlı sınıflandırmada ekşi sözlük yazarlarının yazdığı yazıların konularına göre tahmin edilmesi ve etiketlenmesi üzerine yapılan çalışmada toplam 6 konu üzerinden 1161 başlık ve 10417 post ile modeller eğitilmiştir. Bu model üzerinden eğitim setinin içerisinde olmayan etiketleri bilinen 20 ekşi sözlük yazısı ile test gerçekleştirildi. PV-DM ve PV-DBOW yöntemlerinin her ikisinde de başarılı sonuçlar elde edildi PV-DM yöntemi 0,98'lik oranla tahminlemede çok daha başarılı olurken, PV-DBOW oranı ise 0,8 olarak gerçekleşti.

Çizelge 4.10 : PV-DM Modeli hata matrisi.

| | *ekonomi* | *sağlık* | *siyaset* | *spor* | *tarih* | *teknoloji* |
|---|---|---|---|---|---|---|
| *ekonomi* | 18 | 1 | 0 | 0 | 0 | 1 |
| *sağlık* | 0 | 20 | 0 | 0 | 0 | 0 |
| *siyaset* | 0 | 0 | 20 | 0 | 0 | 0 |
| *spor* | 0 | 0 | 0 | 20 | 0 | 0 |
| *tarih* | 0 | 0 | 0 | 0 | 20 | 0 |
| *teknoloji* | 0 | 0 | 0 | 0 | 0 | 20 |

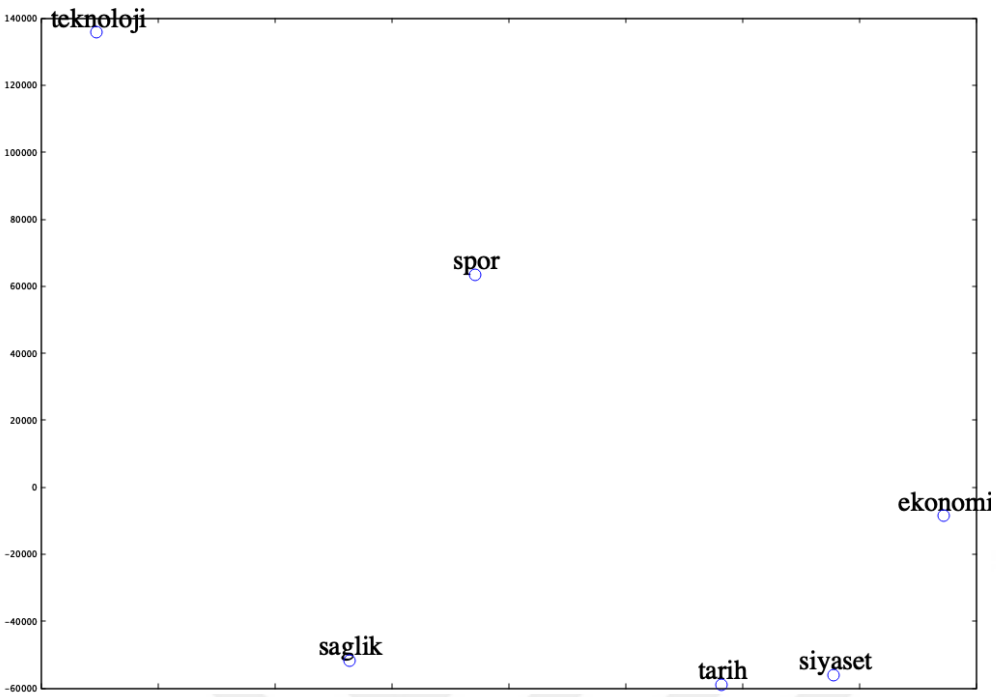
Bir konuda yazılan yazıdaki kullanılan kelimeler genelde konu ile ilgili kelimelerden oluştuğu için ekşi sözlükteki yazarların bir konuda yazmış olduğu yazılar başarılı bir şekilde birbirinden ayırt edilebilmiştir.

Konulara göre oluşturulmuş çok boyutlu vektörler görselleştirilmek için hazırlanan uygulama ile iki boyuta indirgenmiştir. İki boyuta indirgemek için t-SNE fonksiyonları kullanılmıştır.

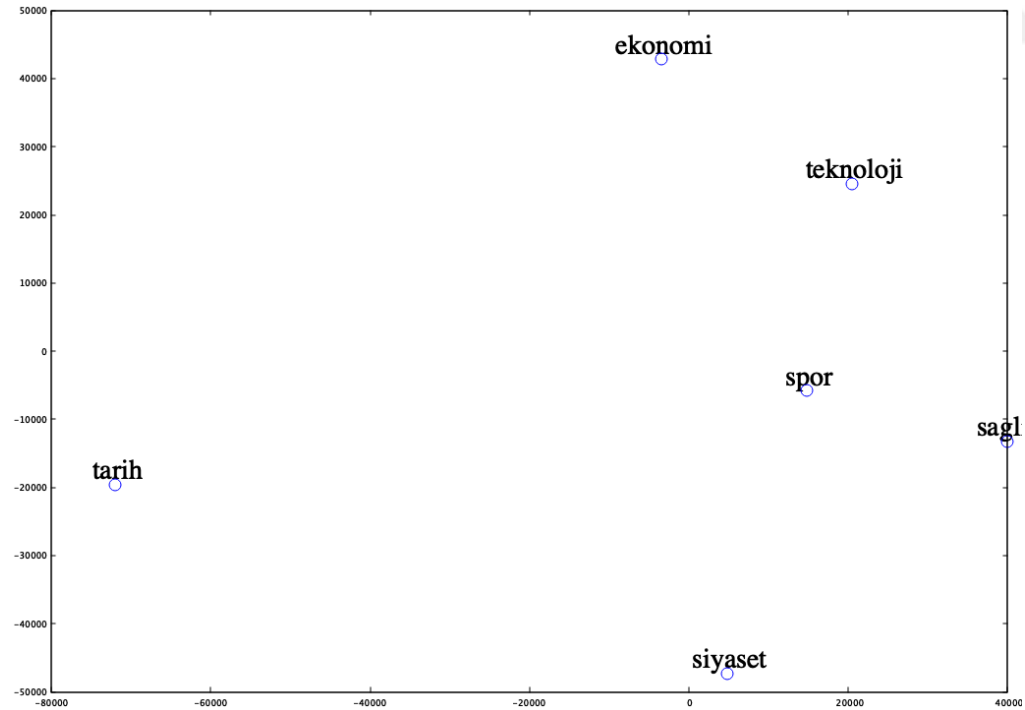
Çizelge 4.11 : PV-DBOW Modeli hata matrisi.

| | *ekonomi* | *sağlık* | *siyaset* | *spor* | *tarih* | *teknoloji* |
|---|---|---|---|---|---|---|
| *ekonomi* | 9 | 0 | 0 | 1 | 10 | 0 |
| *sağlık* | 0 | 20 | 0 | 0 | 0 | 0 |
| *siyaset* | 1 | 0 | 8 | 0 | 11 | 0 |
| *spor* | 0 | 0 | 0 | 20 | 0 | 0 |
| *tarih* | 0 | 0 | 1 | 0 | 19 | 0 |
| *teknoloji* | 0 | 0 | 0 | 0 | 0 | 20 |

t-SNE den alınan iki boyutlu vektörler gnuplot aracılığı ile [42] görselleştirilmiştir. Şekil 4.9 ve Şekil 4.10'daki verileri elde edilmiştir. Şekillerde tarih, ekonomi ve siyaset PV-DM modelinde birbirinden bir hayli uzakta kalmıştır, bu üç konunun diğer konulardan kolayca ayırt edilebildiğini söyleyebiliriz. Ancak PV-DBOW modelinde ekonomi ve siyaset konusu tarih konusu karışmıştır. Şekilde PV-DBOW modelinde bu üç konunun birbirine yakınlığı gözlemlenmiştir.



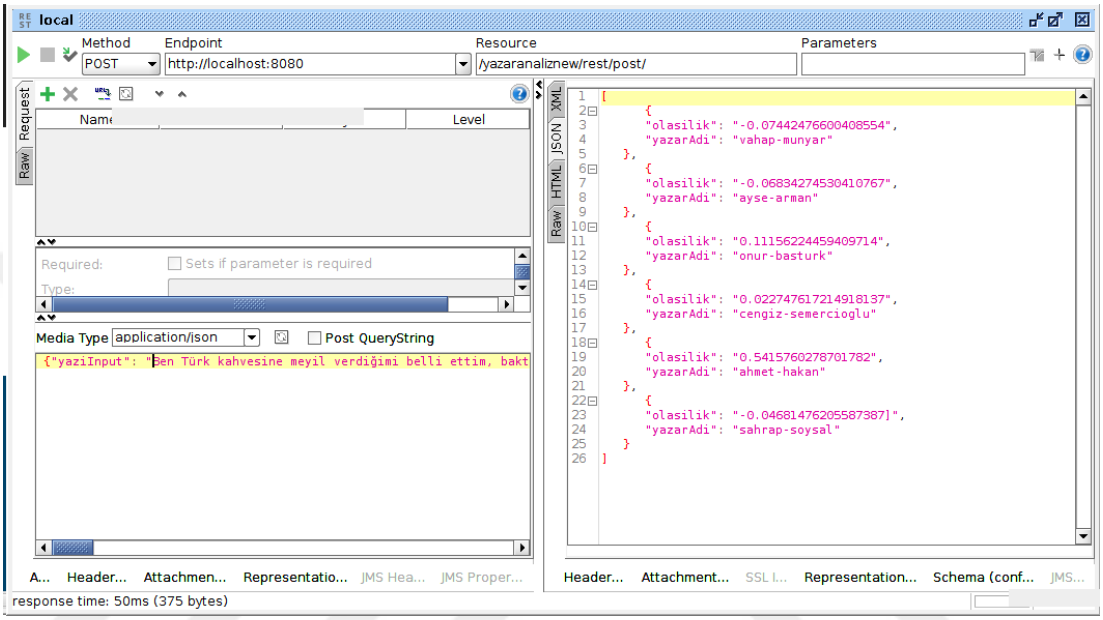

Şekil 4.7 : PV-DBOW 2 boyuta indirgenmiş grafiği




Şekil 4.8 : PV-DM 2 Boyuta indirgenmiş grafiği.

## 4.4 Yazar Tahminleme Uygulaması

Yazar analizinin gerçek zamanlı yapılması için oluşturduğumuz uygulama web-servis ve arayüz olmak üzere iki parçadan oluşmaktadır. Web servis bölümünde input değişkeni ile alınan yazıya karşılık olarak json dizisi şeklinde yazarlar ve gönderilen yazıya olan kosinüs benzerlikleri cevap dönmektedir.




Şekil 4.9 : Gerçek zamanlı yazarlara olan benzerlikleri ölçümleyen uygulama

Bu web servisi isteyen her uygulama tüketeceği gibi, kullanıcılar oluşturduğumuz arayüzden bir yazının hangi yazara ait olabileceğini sorgulayabilirler. http://yazaranaliz.msari.com adresinden çalışan uygulama aracılığıyla metinler sorgulandıktan sonra Şekil 4.10 daki gibi kullanıcılara kosinüs benzerliği olarak en çok hangi yazara yakın olduğu ve diğer yazarlara olan benzerlikleri listelenmektedir.


Şekil 4.1 : Metinlerin analiz için uygulamaya gönderilmesi

# Ahmet Hakan

Yakınlık Değerleri (Cosine similarity & Softmax sonrası)

| olasilik | yazar |
|---|---|
| 0.03688159957528114 | vahap-munyar |
| -0.1018233597278595 | ayse-arman |
| 0.12230726331472397 | onur-basturk |
| 0.15743683278560638 | cengiz-semercioglu |
| 0.4780495762825012 | ahmet-hakan |
| -0.06506098806858063] | sahrap-soysal |

# Vahap Munyar

Yakınlık Değerleri (Cosine similarity & Softmax sonrası)

| olasilik | yazar |
|---|---|
| 0.3568735420703888 | vahap-munyar |
| 0.12199679017066956 | ayse-arman |
| -0.05674559995532036 | onur-basturk |
| 0.07065824419260025 | cengiz-semercioglu |
| 0.024742664769291878 | ahmet-hakan |
| 0.013269861228764057] | sahrap-soysal |

Şekil 4.2 : Elde edilen sonuçlar ve kosinüs benzerliği en yüksek yazarın belirtilmesi

# 5. ÇIKARIMLAR VE GELECEK ÇALIŞMALAR

Bu çalışmada yazar profilini çıkarma, ekşi sözlük başlıklarını konularına göre ayırt etme probleminin çözümü için doc2vec modellerinden PV-DM ve PV-DBOW yöntemleri kullanılmış ve performansları kıyaslanmıştır. Gazeteden her bir yazardan 1000 yazı olmak üzere toplamda 20000 yazı alınmış ve PV-DM yönteminin PV-DBOW yöntemine göre 0,69 oranla çok daha iyi performans gösterdiği sonucuna ulaşılmıştır.

Ekşi sözlükten alınan metinler konularına göre ayrılmış, toplam 1161 başlık ve 10417 metin üzerinden analiz yapılmıştır. Konulardan oluşturulan vektörler model performanslarına göre kıyaslanmıştır. Başlığın konusunu belirleme de PV-DM modeli %98 oranında başarım sağlamıştır. Benzer kelime bağlamlarının farklı konularda kullanılma ihtimalini azaltması modelin başarımını artıran faktör olmuştur.

Konuların birbiri arasındaki grafik 2 boyutlu düzleme indirgenerek incelendiğinde benzer kelimelerin bir arada daha sık geçtiği tarih ve siyaset gibi alanların birbirine daha yakın çıkmıştır. Konu sayısı artırıldığında yakın içerikteki konuların birbirinden ayrılmasını zorlaşacağı çıkarımı yapılabilir.

Yapılan çalışmada yazarların farklı konularda yazması benzer kelime bağlamlarını kullanma ihtimalini artırdığı ve yazar bazlı oluşturulan modellerdeki başarımın konu bazlı oluşturulan modellere göre daha düşük olduğu gözlemlenmiştir.

Modeli oluşturulmamış bir konu ilgili yazı verildiğinde bu çalışmada bir konuya benzetilmektedir. Hürriyet gazetesi köşe yazarları için oluşturulan modelde de sistem verilen yazının modelde olan yazarlara olan uzaklıklarını belirleyip, en yakın olduğu yazarı yazının sahibi olarak tespit etme üzerine kuruludur. Gelecek çalışmalarda

verilen yazı modelde bulunan yazarlardan birine ait değilse hiçbir yazarı etiketlememesi üzerine çalışmalar yapılabilir.
Bu tespitin yapılabilmesinde öklit mesafesinden yaralanılabileceğini düşünüyoruz. Analizi yapılan yazı ve oluşturulan modeller arasında öklit mesafesi baz alınarak, belirli bir mesafeden uzaksa hiçbir konuya benzetilmemesi şeklinde sistem optimize edilebilir.

Modeldeki yazar sayısı 5 olduğunda PV-DM yaklaşımı ile 0,88 oranında başarım sağlanmış ve ayırt edilebilmiştir. Aynı yaklaşım kullanılarak yazar sayısı artırıldığında başarı oranı 0,69'a kadar düşmüştür. Doğal dil işleme yöntemi kullanılarak benzer bağlamların daha kolay çıkartılabileceğini ve daha başarılı sonuçlar alınabileceğini düşünüyoruz.

## KAYNAKLAR


[1] **İ. H. BALTACIOĞLU**, "Grafoloji Konusu, Metodu, Prensipleri," *DTCF Derg.*, vol. 12, no. 1–2, 2017.

[2] **Q. Le and T. Mikolov**, "Distributed Representations of Sentences and Documents," p. 9.

[3] **A. Q. Morton**, "The Authorship of Greek Prose," *J. R. Stat. Soc. Ser. Gen.*, vol. 128, no. 2, pp. 169–233, 1965.

[4]**E. Stamatatos, N. Fakotakis, and G. Kokkinakis,** "Automatic Text Categorization in Terms of Genre and Author," *Comput. Linguist.*, vol. 26, no. 4, pp. 471–495, Dec. 2000.

[5] **Z. Fan**, **L. Su, X. Liu, and S. Wang,** "Multi-label Chinese question classification based on word2vec," in *2017 4th International Conference on Systems and Informatics (ICSAI)*, 2017, pp. 546–550.

[6] **B. Pang, L. Lee, and S. Vaithyanathan,** "Thumbs Up?: Sentiment Classification Using Machine Learning Techniques," in *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing - Volume 10*, Stroudsburg, PA, USA, 2002, pp. 79–86.

[7] **A. Agarwal, B. Xie, I. Vovsha, O. Rambow, and R. Passonneau,** "Sentiment Analysis of Twitter Data," in *Proceedings of the Workshop on Languages in Social Media*, Stroudsburg, PA, USA, 2011, pp. 30–38.

[8] **B. Liu,** "Sentiment analysis and subjectivity," in *Handbook of Natural Language Processing, Second Edition. Taylor and Francis Group, Boca*, 2010.

[9] **K. T. Durant and M. D. Smith,** "Mining sentiment classification from political web logs," in *Proceedings of Workshop on Web Mining and Web Usage Analysis of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (WebKDD-2006), Philadelphia, PA*, 2006.

[10] **H. Kang, S. J. Yoo, and D. Han,** "Senti-lexicon and improved Naïve Bayes algorithms for sentiment analysis of restaurant reviews," *Expert Syst. Appl.*, vol. 39, no. 5, pp. 6000–6010, Apr. 2012.

[11] **M. C. Ganiz, M. Tutkan, and S. Akyokuş,** "A novel classifier based on meaning for text classification," in *2015 International Symposium on Innovations in Intelligent SysTems and Applications (INISTA)*, 2015, pp. 1–5.

[12] **I. Mayda and M. Yesiltepe,** "N-gram based approach to recognize the twitter accounts of Turkish daily newspapers," 2017, pp. 1–5.

[13] **A. Deniz and H. E. Kiziloz,** "Effects of various preprocessing techniques to Turkish text categorization using n-gram features," in *2017 International Conference on Computer Science and Engineering (UBMK)*, 2017, pp. 655–660.

[14] **M. Bilgin and I. F. Senturk,** "Sentiment analysis on Twitter data with semi-supervised Doc2Vec," 2017, pp. 661–666.
[15] **O. Karasoy and S. Ballı,** "Classification Turkish SMS with deep learning tool Word2Vec," in *2017 International Conference on Computer Science and Engineering (UBMK)*, 2017, pp. 294–297.
[16] **G. Şahİn,** "Turkish document classification based on Word2Vec and SVM classifier," in *2017 25th Signal Processing and Communications Applications Conference (SIU)*, 2017, pp. 1–4.
[17] **Ö. Çoban and I. Karabey**, "Music genre classification with word and document vectors," in *2017 25th Signal Processing and Communications Applications Conference (SIU)*, 2017, pp. 1–4.
[18] **O. Dülger,** "Türkçe Metinlerde İroni Tespiti."
[19] **M. Kaya and S. A. Özel,** "A Comparison of Text Similarity Detection Software for Turkish Documents and Investigating the Effects of Stemming and Turkish Character Usage," *Çukurova Üniversitesi Mühendis.-Mimar. Fakültesi Derg.*, vol. 29, no. 2, pp. 115–130, Dec. 2014.
[20] **H. K. Yildiz, M. Genctav,** N. Usta, B. Diri, and M. F. Amasyali, "A New Feature Extraction Method for Text Classification," in *2007 IEEE 15th Signal Processing and Communications Applications*, 2007, pp. 1–4.
[21] **H. Takçı and E. Ekinci,** "Character Level Authorship Attribution for Turkish Text Documents," *TOJSAT*, vol. 2, no. 3, pp. 12–16, Sep. 2012.
[22] **B. Diri and M. F. Amasyalı,** "Automatic author detection for turkish texts," in *Artificial Neural Networks and Neural Information Processing (ICANN/ICONIP)*, 2003, pp. 138–141.
[23] **H. Gunduz and Z. Cataltepe,** "Borsa Istanbul (BIST) daily prediction using financial news and balanced feature selection," *Expert Syst. Appl.*, vol. 42, no. 22, pp. 9001–9011, Dec. 2015.
[24] **Ö. Özyurt and C. Köse,** "Chat mining: Automatically determination of chat conversations' topic in Turkish text based chat mediums," *Expert Syst. Appl.*, vol. 37, no. 12, pp. 8705–8710, Dec. 2010.
[25] **M. Sarı and A. M. Özbayoğlu,** "Classification of Turkish Documents Using Paragraph Vector," in *2018 International Artificial Intelligence and Data Processing Symposium (IDAP)*, 2018, pp. 1–4.
[26] **A. Haltaş and A. Alkan,** "Medlıne Veritabanı Üzerinde Bulunan Tıbbi Dökümanların Kanser Türlerine Göre Otomatik Sınıflandırılması," *Bilişim Teknol. Derg.*, vol. 9, no. 2, p. 181, May 2016.
[27] **T. Kaşikçi and H. Gökçen,** "Metin Madenciliği İle E-Ticaret Sitelerinin Belirlenmesi," *Bilişim Teknol. Derg.*, vol. 7, no. 1, Oct. 2013.
[28] **M. A. Ağca, Ş. Ataç, M. M. Yücesan, Y. G. Küçükayan, A. M. Özbayoğlu, and E. Doğdu,** "Opinion mining of microblog texts on Hadoop ecosystem," *Int. J. Cloud Comput.*, vol. 5, no. 1–2, pp. 79–90, 2016.
[29] **S. Kulcu, E. Dogdu, and A. M. Ozbayoglu,** "A survey on semantic Web and big data technologies for social network analysis," in *Big Data (Big Data), 2016 IEEE International Conference on*, 2016, pp. 1768–1777.

[30] **E. Arin, M. U. Gudelek, and A. M. Ozbayoglu,** "Quora Duplicate Query Elimination."
[31] **W. Zhu, W. Zhang, G.-Z. Li, C. He, and L. Zhang,** "A study of damp-heatsyndrome classification using Word2vec and TF-IDF," in *2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 2016, pp. 1415–1420.
[32] **A. M. J. Schakel and B. J. Wilson,** "Measuring Word Significance using Distributed Representations of Words," *ArXiv150802297 Cs*, Aug. 2015.
[33] **S. Gupta, G. Kaiser, D. Neistadt, and P. Grimm,** "DOM-based content extraction of HTML documents," presented at the Proceedings of the 12thinternational conference on World Wide Web, 2003, pp. 207–214.
[34] **P. Houston,** *Instant jsoup How-to*. Packt Publishing Ltd, 2013.
[35] **L. van der Maaten and G. Hinton,** "Visualizing Data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, no. Nov, pp. 2579–2605, 2008.
[36] **M. Wattenberg, F. Viégas, and I. Johnson,** "How to Use t-SNE Effectively," *Distill*, vol. 1, no. 10, p. e2, Oct. 2016.
[37] "Deeplearning4j." [Online]. Available: https://deeplearning4j.org/. [Accessed: 20-Oct-2018].
[38] **Y. Yang,** "An evaluation of statistical approaches to text categorization," *Inf. Retr.*, vol. 1, no. 1–2, pp. 69–90, 1999.
[39] **U. Kumaresan and K. Ramanujam,** "A framework for extraction of journal information from scientific publishers web site," in *2016 10th International Conference on Intelligent Systems and Control (ISCO)*, 2016, pp. 1–5.
[40] **S. Sirsat and V. Chavan,** "Pattern matching for extraction of core contents from news web pages," in *2016 Second International Conference on Web Research (ICWR)*, 2016, pp. 13–18.
[41] **Y. Goldberg and O. Levy,** "word2vec Explained: deriving Mikolov et al.'s negative-sampling word-embedding method," *ArXiv14023722 Cs Stat*, Feb. 2014.
[42] **J. Racine,** "gnuplot 4.0: a portable interactive plotting utility," *J. Appl.Econom.*, vol. 21, no. 1, pp. 133–141, Jan. 2006.

# ÖZGEÇMİŞ


**Ad-Soyad** : **MUSTAFA SARI**

**Uyruğu** : **T.C.**

**Doğum Tarihi ve Yeri** : **11.08.1989 ÇORUM**

**E-posta** : **msari@etu.edu.tr**


## ÖĞRENİM DURUMU:

- **Lisans** : 2014, İzmir Yüksek Teknoloji Enstitüsü, Mühendislik Fakültesi, Bilgisayar Mühendisliği

## MESLEKİ DENEYİM VE ÖDÜLLER:

| Yıl | Yer | Görev |
|---|---|---|
| **2014 - …** | **TÜRKSAT** | **Bilgisayar Mühendisi** |
| **2012-2014** | **İYTE Kütüphane** | **Yarı zamanlı yazılım geliştirici** |
| **2012** | **AVATEK** | **Yarı zamanlı yazılım geliştirici** |

## YABANCI DİL: İNGİLİZCE

## TEZDEN TÜRETİLEN YAYINLAR, SUNUMLAR VE PATENTLER:

- **M. Sarı** and **A. M. Özbayoğlu**, "Classification of Turkish Documents Using Paragraph Vector," in 2018 International Artificial Intelligence and Data Processing Symposium (IDAP), 2018, pp. 1–4.