Türkçe Haberlerin Tür Tespiti İçin Konu Modelleme Yöntemlerinin Karşılaştırılması

Comparison of Topic Modeling Methods for Type Detection of Turkish News

Zekeriya Anil Güven
Bilgisayar Mühendisliği
Ege Üniversitesi
İzmir, Türkiye
zekeriya.anil.guven@ege.edu.tr

Banu Diri
Bilgisayar Mühendisliği
Yıldız Teknik Üniversitesi
İstanbul, Türkiye
diri@yildiz.edu.tr

Tolgahan Çakaloğlu
Department of Computer Science
University of Arkansas
Arkansas, ABD
txcakaloglu@ualr.edu

***Öz*—Günümüzde Internet tabanlı belgelerin artmasıyla, bizlere işlenmesi ve değerlendirilmesi gereken birçok veri sunulmaktadır. Medya, haber ve reklam sektörü bu verilerin değerlendirildiği alanlardan bazılarıdır. Haber için özellikle medya sektöründe insanların sınıflandırma yapması, zaman açısından önemli bir sorundur. Çalışmada, haber başlıklarının hangi türe ait olduğunu tespit etmek hedeflenmiştir. Veri setimiz, 7 sınıfa ait 4200 adet Türkçe haber başlıklarından oluşmaktadır. Türleri tespit etmek için konu modellemede kullanılan klasik Gizli Dirichlet Ayırımı (GDA), Gizli Anlamsal Analiz (GAA) ve Negatif Olmayan Matris Faktörizasyonu (NMF) algoritmasından yararlanılmıştır. Ayrıca, GDA tabanlı n-GDA yöntemi de kullanılmıştır. Kullanılan tüm yöntemlerin doğrulukları ölçülmüş ve karşılaştırılmıştır. Üç sınıf için en başarılı yöntem NMF olurken, beş ve yedi sınıf için GAA olmuştur.**

Anahtar Sözcükler — Konu Modelleme, Gizli Dirichlet Ayırımı, Doğal Dil İşleme, Haber Analizi, Negatif Olmayan Matris Faktörizasyonu, Gizli Anlamsal Analiz.

***Abstract*— Today, with the increase of Internet-based documents, we are presented with many data that need to be processed and evaluated. Media, news and advertising are some of the areas where these data are evaluated. For the news, the classification of people in the media sector is an important problem in terms of time. In this paper, it is aimed to determine which types of news titles belong to. The dataset consists of 4200 Turkish new titles belonging to 7 class labels. In order to determine the types, classical Latent Dirichlet Allocation (LDA), Latent Semantic Analysis (LSA) and Non-Negative Matrix Factorization (NMF) algorithms were used in topic modeling. In addition, the LDA-based n-LDA method was also used. The accuracy of all methods used was measured and compared. NMF was the most successful method for three classes, while for five and seven classes LSA was the most successful method.**

Keywords — Topic Modelling, Latent Dirichlet Allocation, Natural Language Processing, New Analysis, Non-Negative Matrix Factorization, Latent Semantic Analysis.

I. GİRİŞ

Elektronik belge arşivlerinin çoğalmasından dolayı belgeleri yönetmek için büyük koleksiyonları otomatik olarak düzenleme, arama, indeksleme ve taramayla ilgilenen yeni teknikler veya araçlar kullanılması gerekmektedir. Günümüzde makine öğrenmesi ve istatistiksel araştırmalara dayanarak, belge koleksiyonlarındaki sözcük kalıplarını bulmak için yeni teknikler geliştirilmiştir. Bu tekniklere konu modelleri denilmektedir. Konu modelleri, görüntüler, biyolojik veriler ve anket bilgileri gibi kelimelerden ziyade konuları analiz etmek için kullanılmaktadır [1].

Konu modellemenin amacı, kelime kullanım modelleri ve benzer modelleri paylaşan belgeleri nasıl birleştirileceğini keşfetmektir. Dolayısıyla, konu modelleri, belgelerle çalışılabilecek bir terimdir. Bu belgeler konuların karışımı olarak düşünülebilir. Konu ise, bir konu üzerindeki olasılık dağılımıdır. Diğer bir deyişle, konu modelleme belgeler için üretken bir modeldir. Belgelerin üretilebileceği basit bir olasılık prosedürünü belirtmektedir [2].

Metin analizi ve metin madenciliği tarafında, konu modelleri, kelime torbası (bag of words) varsayımına dayanır. Gizli Anlamsal Analiz (GAA), Olasılıklı Gizli Anlamsal Analiz (O-GAA), Gizli Dirichlet Ayırımı (GDA), İlişkili Konu Modeli (IKM) gibi çeşitli konu modelleri olarak geliştirilmiştir. Ayrıca, konu modelleme için de kullanılan Negatif Olmayan Matris Faktörizasyonu (NMF) ile analiz edilecek veriler, negatif değildir. Böylece analiz edilen veriler düşük dereceli verilerin çelişmesini önlemek için negatif olmayan değerlerden oluşmalıdır [3].

İlgili çalışmaları incelersek; Bergamaschi ve diğerleri [4], film benzerliğine dayanan otomatik bir film öneri sistemi önermişlerdir. GAA ve GDA iki yüz bin filmden oluşan bir veritabanına uygulanmış ve kapsamlı bir şekilde karşılaştırılmıştır. Standart metriklere dayalı konu modelleri yaklaşımları sonucunda 30 kullanıcı ile performans değerlendirmesi yapmışlardır. Sonuç olarak, benzer seçimlerde GAA'nın performansının GDA'nınkinden daha iyi olduğunu göstermişlerdir. Crossno ve diğerleri [5], birden fazla metin derlem modelini görsel olarak karşılaştırmak ve araştırmak için bir uygulama sunmuşlardır. Uygulama, GAA ve GDA yaklaşımları kullanılarak oluşturulan modellere uygulanmıştır. İki yaklaşım kavramsal içerik olarak karşılaştırılmıştır. Stevens ve diğerleri [6], her konu modeli paradigmasının güçlü ve zayıf yönlerini araştırmışlardır. NMF ve GDA'nın hem özlü hem de tutarlı konuları öğrendiğini ve değerlendirmelerde benzer bir performans elde ettiğini göstermişlerdir. Bununla birlikte NMF'in, GDA ve Tekil Değer Ayrıştırma'dan (TDA) daha tutarsız konuları öğrendiğini belirlemişlerdir. Son kullanıcısının öğrenilen konularla etkileşime gireceği uygulamalar için, GDA'nın esneklik ve tutarlılık avantajlarını güçlü bir şekilde dikkate alınmasının gerektiğini açıklamışlardır. Ayrıca, GAA ile GDA kıyaslandığında GDA'nın en iyi tanımlayıcı konuları öğrenirken, GAA'nın ise belgelerin ve kelimelerin kompakt anlamsal gösterimini oluştururken en iyisi olduğunu belirtmişlerdir. Utsumi [7], NMF ve anlamsal uzayları oluşturmak için en popüler yöntem olan GAA'yı karşılaştırmıştır. Sonuç olarak, NMF'nin hiçbir testte GAA'dan daha iyi performans göstermediği bulunmuştur. Bu bulgu, NMF'nin sözcük anlamları elde etmede literatürde beklenenden daha az etkili olduğunu göstermektedir. Suri ve diğerleri [8], Twitter'dan elde edilen bu metinsel verilerden başlıkları ve haber başlıkları RSS beslemesini saptamak amacıyla, GDA ve NMF kullanmışlardır. Gözlenen

978-1-7281-3964-7/19/$31.00 ©2019 IEEE

sonuçlarda, her iki algoritmanın da metindeki konuları tespit etmede iyi performans gösterdiğini belirtmişlerdir. Ayrıca, NMF'in GDA'dan daha hızlı, ancak GDA'nın sonuçlarının daha anlamlı olduğunu göstermişlerdir. Chen ve diğerleri [9], GDA ve NMF temelli öğrenme programları arasındaki bilgileri keşfetmek için kapsamlı bir şekilde iki kısım halinde deney gerçekleştirmişlerdir. Spesifik olarak, temel GDA ve NMF, birinci kısımda halka açık kısa metin veri setleri üzerindeki farklı deneyler ile karşılaştırılmıştır. Bu durumda NMF, GDA'dan daha iyi performans göstermiştir. İkinci kısımda ise, NMF tabanlı yeni bir yöntem önermişlerdir. Önerilen yöntem ile yapılan deneylerde de NMF tabanlı yöntemler daha etkili olmuştur. Chen ve diğerleri [10], robotlarda belirsiz verilerden kelime-nesne ilişki öğrenmesi ve öğrenilen örnek seçmede etkili olması için NMF ve GDA'ya dayalı gizli konu öğrenme modelleri olan birkaç model kullanmışlardır. Bu yaklaşımların, güçlü ve zayıf yönlerini analiz etmek için bir gruptaki aynı verileri kullanarak karşılaştırma yapmışlardır.

Literatürde konuların tespiti ile ilgili birçok yöntemden bahsedilmiştir. Çalışmada haberlerin hangi türe ait olduğunu belirlemek için GDA, GAA ve NMF yöntemleri kullanılmıştır. Yöntemlerin doğruluğu hesaplanarak karşılaştırılması sağlanmıştır. Ayrıca, geçmişte önerdiğimiz GDA tabanlı n-aşamalı (n-GDA) yöntemi kullanılarak diğer üç yöntem ile karşılaştırılmıştır [11].

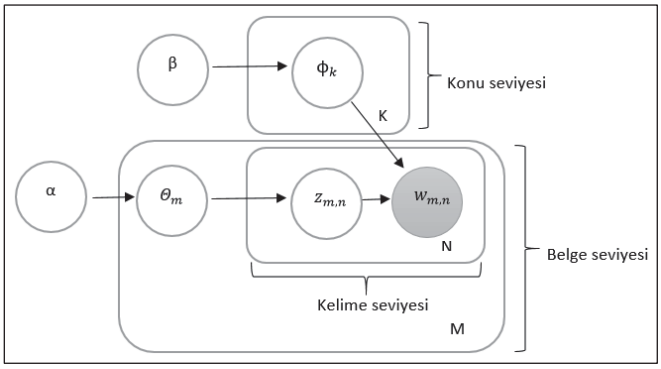
Makalenin ikinci bölümünde kullanılan yöntemler, veri seti, veri setine uygulanan ön işlemler anlatılmıştır. Üçüncü bölümde GDA, n-GDA, NMF ve GAA yöntemlerinin haber başlıkları üzerine konu tespit analizi yapılmıştır ve yöntemlerin doğruluk sonuçları gösterilmiştir. Dördüncü bölümde ise yöntemlerin sonuçları değerlendirilmiştir.

II. MATERYALLER VE METOTLAR

A. Gizli Dirichlet Ayırımı Algoritması

Gizli Dirichlet Ayırımı (GDA), metin madenciliği için istatistiksel (Bayesian) konu modeline dayanan bir algoritmadır ve çok yaygın olarak kullanılmaktadır. Böylece GDA belirlenen konular ile bir belge oluşturmaya çalışmaktadır. GDA'da her belge konuların karışımı olarak modellenmektedir ve her konu, her bir kelimenin belirli bir konuda görünme ihtimalini tanımlayan ayrı bir olasılık dağılımını göstermektedir [2]. GDA ile her belge, konuların kümelere benzeyen daha üst düzey kavramlar olarak görülebileceği konuların multinomial bir dağılımı olarak temsil edilmektedir. Yöntem, bir koleksiyondaki her belgenin, her bir konunun sunulduğu birkaç gizli başlıktan oluşturulduğu varsayımına dayanmaktadır. Bu şekilde, iki aşamalı bir süreçle belge oluşturulması tanımlanabilmektedir:

- Her belge için m ∈ M, Dirichlet dağılımı olan Dir (α)'dan bir konu oranını örneklemektedir.
- M belgesindeki her kelime yer tutucusu *n* için;
  - θ_m örneklenen konu oranlarına göre rasgele bir $z_{m,n}$ konusu seçer.
  - Önceden seçilen konunun ϕ_k multinomial dağılımından rastgele bir $w_{m,n}$ kelimesi seçer.



Şekil 1. Gizli Dirichlet Ayırımının yapısı [12]

Yukarıda belirtilen süreçte, α ve β parametreleri, tüm belgeler için bir konu dağılım seti olarak θ üzerindeki ve tüm konulardaki bir kelime dağılım seti olarak φ üzerinde Dirichlet önceliğini belirleyen hiperparametrelerin vektörleridir. Tipik olarak, olasılık dağılımının tek bir noktaya nasıl odaklandığını tanımlayan $\alpha_1 = \alpha_2 = \alpha_k = \alpha$ olan simetrik Dirichlet öncelikleri kullanılır. Tüm işlem Şekil 1'de gösterilmektedir [12].

GDA nispeten basit bir model olsa da, gizli değişkenlerin ve parametrelerin tam çıkarımını yapamamaktadır. Bu nedenle, değişken EM'den (Expectation Maximization) beklenti yayılımına ve Gibbs örneklemesine kadar değişen model parametrelerinin yaklaşık tahminlerini elde etmek için bir takım algoritmalar mevcuttur [12].

Sistemi uygun konu sayısı ile modellemek önemlidir. Konu sayısını belirlemek için genellikle tutarlılık (coherence) değeri kullanılmaktadır. Tutarlılık değeri; kelimelerin birbirine benzerliğini ölçmektedir. Hesaplanan tutarlılık değerlerinden en yüksek olana ait konu sayısı, sistemin eğitileceği konu sayısı olarak seçilmektedir [13].

B. Gizli Anlamsal Analiz

Gizli Anlamsal Analiz (GAA), kelimelerin geniş bir belge koleksiyonunda bir araya geldiği frekansların istatistiksel analizinden anlamsal uzay oluşturan bir algoritmadır. GAA'nın belge koleksiyonundan anlamsal bir uzay oluşturduğu sürece 'eğitim' denir. Eğitimden sonra, anlamsal uzay, belge koleksiyonunda karşılaşılan her bir kelimenin anlamsal özelliklerini içeren bir dizi vektör içermektedir. Anlamsal uzayda bulunan vektörler, anlamsal vektörler olarak adlandırılmaktadır. Bir eğitim derlemi içinde ne kadar fazla belge varsa, sistemin kelimeleri anlamsal olarak ayırt etmesi veya sıralaması gereken bağlamsal bilgi de o kadar fazladır [14].

GAA, kelimenin anlam gösterimi için en çok kullanılan yöntemlerden biridir. Yöntem, girdi olarak bir eğitim derlemi, yani bir belge koleksiyonu almaktadır. Doküman eş-oluşum matrisi ile bir kelime oluşturulur. Tipik olarak, kelime-doküman matrisindeki bilgi vermeyen yüksek frekanslı kelimelerin ağırlığını azaltmak için tf-idf dönüşümü uygulanmaktadır. Tf-idf dönüşümünün çıktısı, $w_{i,j}$ öğesinin, j dokümanındaki i kelimesinin ağırlığı olduğu bir W matrisidir.

$$w_{i,j} = tf_{i,j}.\log_2(\frac{D}{df_i}) \quad (1)$$

Denklem (1)'de $tf_{i,j}$, j belgesindeki i kelimesinin sıklığını, D eğitim derlemi içerisindeki doküman sayısını vermektedir.

df_i ise i kelimesinin bulunduğu doküman sayısıdır. Daha sonra, her bir doküman ağırlığı, birim uzunluğuna normalize edilmektedir. Ayrıca en büyük tekil değerlerin seçildiği Tekil Değer Ayrışımı (TDA) ile boyutsallık azaltması uygulanmaktadır. Bu yöntem, eğitimli derlemde mevcut her kelimenin düşük boyutlu bir vektörel gösterimini sağlamaktadır. GAA'nın kelimelerin gizli anlamını yakalamadaki başarısı bu düşük boyutlu haritalamadan gelmektedir [15].

C. Negatif Olmayan Matris Faktörizasyonu

NMF, negatif olmayan kısıtlamalar ile verilerin temsilini elde etmek için kullanılan bir vektör uzayı yöntemidir. Bu kısıtlamalar, parça bazında bir temsile yol açabilir, çünkü orijinal verilerin yalnızca eklenecek kombinasyonlarına izin verirler. Bu, temel bileşen analizi (PCA) gibi tekil değer ayrıştırma yöntemlerinin aksinedir. PCA ile ilgili bir problem, temel vektörlerin hem pozitif hem de negatif bileşenlere sahip olmasıdır. Bu veriler, vektörlerin pozitif ve negatif katsayılarla doğrusal kombinasyonları olarak temsil edilmektedir. Ancak, birçok uygulamada, negatif bileşenler fiziksel gerçeklerle çelişmektedir. Özellikle, metin madenciliğindeki terim frekansları negatif değildir [16].

NMF, W ve H olarak adlandırılan negatif olmayan faktörleri içeren A matrisine düşük dereceli bir yaklaşım oluşturmak için kullanılmıştır. Bir veri matrisi A'nın NMF'si, denklem (2)'deki doğrusal olmayan optimizasyon problemini çözerek oluşturulmaktadır [17].

$$\min || A_{mxn} - W_{mxk} H_{kxn} ||_F^2 \qquad (2)$$
$$W \geq 0, H \geq 0.$$

Denklem (2)'deki Frobenius normu ($||_F^2$), genellikle orijinal A matrisi ile düşük dereceli yaklaşım olan WH arasındaki hatayı ölçmek için kullanılmaktadır. Yaklaşımın derecesi k, kullanıcı tarafından ayarlanması gereken bir parametredir [17].

NMF, TDA gibi diğer düşük dereceli faktorizasyonların yerine, iki ana avantajı nedeniyle kullanılmaktadır: Depolama ve yorumlanabilirlik. Negatif olmayan bileşenler nedeniyle, NMF verinin sözde "ek parça bazında" gösterimini sağlamaktadır. Bunun sonucunda, W ve H faktörlerinin genellikle aralıklı olması, böylece TDA'nın yoğun faktörleriyle karşılaştırıldığında büyük miktarda depolama imkanı sağlamaktadır. Yöntem aynı zamanda, faktörlerin yorumlanması açısından da faydalıdır. Örneğin, bir terim belgeli veri matrisi A_{mxn}'nin faktörizasyonunu gerektiren bir metin işleme uygulamasını düşünürsek; k belge koleksiyonunda bulunan gizli konuların sayısı olarak kabul edilmektedir. Bu durumda, W_{mxk} , sütunları NMF baz vektörleri olan bir terim-konu matrisi haline gelmektedir. Aralıklı ve negatif olmayan W'nin 1 sütununun sıfır olmayan elemanları, özel terimlere karşılık gelmektedir. Bu vektördeki en yüksek ağırlıklı terimler göz önüne alındığında, kişi temel vektör 1'e bir etiket veya konu atayabilir [17].

D. n-aşamalı GDA

Modelin doğruluğunu artırmak için önerilmiş olan GDA tabanlı bir yöntem ile bulunan n değeri, veri setine göre değişiklik göstermektedir. Yöntem ile sistemin doğruluğunu olumsuz etkileyen sözlükteki kelimelerin silinmesi hedeflenmiştir. Sonuçta kelimelerin ağırlık değerleri artarak konuların sınıf etiketleri daha kolay belirlenebilmektedir. Yöntemin aşamalarını adım adım açıklamak gerekirse;

- Sözlükteki kelime sayısını azaltmak için her konuya ait eşik değeri hesaplanmaktadır. Eşik değeri, ilgili konudaki tüm kelimelerin ağırlıkları toplamının kelime sayısına oranlanması ile elde edilmektedir;

$$ed(k_i) = \frac{\sum_{j=1}^{m} w_j}{n_i} \qquad (3)$$
$$w_j \geq ed(k_i)$$

  Denklem (3)'te k_i , dokümandaki i. konuyu göstermektedir. w_j , k_i konusunda bulunan kelime ağırlıklarını, n_i ise k_i konusundaki toplam kelime sayısını ifade etmektedir (i: konu; j: konudaki kelime sayısıdır).

- Belirlenen eşik değerinden daha küçük ağırlığa sahip kelimeler konulardan silinerek model için yeni bir sözlük oluşturulmaktadır.
- Son olarak sistem, yeni sözlükle GDA algoritması kullanılarak yeniden modellenmektedir. Bu aşamalar *n* defa tekrarlanabilmektedir.

Aşama sayısının artmasıyla sözlükte kalan kelimelerin ağırlık değerlerinin değişimi Tablo I'de gösterilmiştir. Tablo incelendiğinde kelimenin ağırlık değerinin arttığı görülmektedir. Bu artışın nedeni, daha az kelimeyle sistemin modellenebilmesidir. Böylece konunun sınıf etiketi daha kolay belirlenebilmektedir.

TABLO I. SPOR HABERLERİNE AİT 'MAÇ' KELİMESİNİN N-AŞAMA ile AĞIRLIK DEĞERLERİ

Aşama	Kelimenin Ağırlığı
1-Aşama	0.025
2-Aşama	0.081

E. Veri Seti

Milliyet, Mynet gibi Türkçe haber sitelerden yararlanarak haber başlıkları veya alt başlıklardan oluşan bir veri seti oluşturulmuştur. Haber başlıkları yaklaşık olarak 30 kelimeden oluşmaktadır. Bazı haber başlıkları çok az kelime içerdiği ve bu olayın haber türünü belirlemeyi olumsuz etkileyeceğinden dolayı haber başlığı genişletilmiştir. Bu yüzden haber içeriğindeki alt başlıklar da kullanılmıştır. Veri seti; ekonomi, siyaset, magazin, spor, teknoloji, sağlık, teknoloji ve yaşam olmak üzere 7 farklı sınıf etiketine sahip haberlerden oluşmaktadır. Her haber türü için 600 adet haber başlığı olmak üzere veri seti toplam 4200 adet haber içermektedir. Her biri 3, 5 ve 7 sınıf etiketine sahip, 3 farklı veri seti hazırlanmıştır. Üç sınıf için ekonomi, spor ve yaşam; beş sınıf için ekonomi, spor, yaşam, siyaset ve magazin sınıf etiketleri kullanılmıştır. Veri setinin %80'i eğitim, %20'si de test için ayrılmıştır.

F. Ön İşlemler

Haber metinlerinde ilk önce noktalama işaretleri temizlenmiştir. Ardından kelimelerin büyük-küçük harf uyumsuzluğunu gidermek için tüm veri seti küçük harfe dönüştürülmüştür. Harf dönüştürme işleminde Türkçe karakterlerde problem olabildiği için İ, Ö, Ç gibi İngilizcede olmayan harfler kod aracılığıyla küçük harfe çevrilmiştir. Sonradan haber başlıklarının içerisinde yer alan etkisiz kelimeler (stopwords) veri setinden silinmiştir. Ayrıca, haberler için anlamı olmayan kelimelerden birliste oluşturulmuş ve bu kelimelerde başlıklardan çıkarılmıştır. Son olarak kelimelerin kökünü bulmak için Zemberek

kütüphanesi kullanılmıştır. Kütüphane aracılığıyla isim, fiil ve kısaltma içeren kelimeler ile veri seti güncellenmiştir.

III. DENEYSEL ÇALIŞMALAR

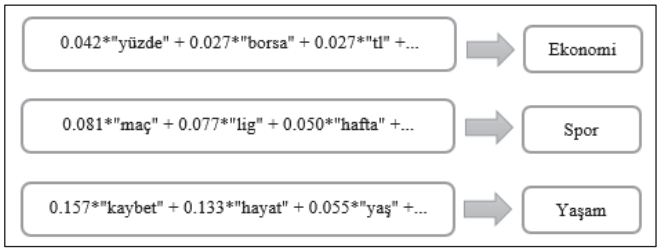
Diğer çalışmalarda kullanılan veri setlerine erişilemediği için bu çalışma için ayrı bir veri seti oluşturulmuştur. Kullanılan veri seti ile n-GDA modelinin; klasik GDA, GAA ve NMF yöntemleri ile sistem doğrulukları karşılaştırılmıştır.

Üç, beş ve yedi sınıflı veri setine ön işlemler uygulandıktan sonra her veri setinde uygun konu sayısını belirlemek için tutarlılık değerleri hesaplanmıştır. Tutarlılık değerleri arasından en yüksek değere sahip olan konu sayısı, sistemi modelleyeceğimiz konu sayısı olarak seçilmiştir. Tüm sınıflara ait tutarlılık değeri ve belirlenen konu sayısı Tablo II'de gösterilmiştir.

TABLO II. SINIFLAR için TUTARLILIK DEĞERLERİNE GÖRE BELİRLENEN KONU SAYILARI

Sınıf Sayısı	Tutarlık Değeri	Konu Sayısı
3	0.5207	12
5	0.5068	20
7	0.4618	21

Tablo II'deki konu sayıları ile sistem ayrı ayrı modellenmiştir. Model ile her veri seti için kelimeler ve ağırlıklarından oluşan bir konu listesi elde edilmiştir. Konu listesindeki konulara, en uygun sınıf etiketi kelimelerin ağırlık değerlerinden yararlanılarak atanmıştır. Şekil 2 konulara sınıf etiketi atanmasına örnek gösterilmiştir.



Şekil **2.** Konulara uygun sınıf atanması için örnek

Klasik GDA ile modellenen konuların sınıf etiketleri atandıktan sonra, haberin hangi konuya ait olduğu belirlenmiştir. Bu işlem, haber içerisindeki kelimelerin tüm konulardaki ağırlıkları ayrı ayrı toplanarak en yüksek değere atanmasıyla gerçekleşmiştir. Klasik GDA ile sistemin doğruluğu Tablo III'te gösterilmiştir. Tablo incelendiğinde, sınıf sayısı arttığında sistemin doğruluğunun düştüğü görülmektedir.

TABLO III. KLASİK GDA'NIN TÜM SINIFLAR için DOĞRULUK DEĞERLERİ (%)

Sınıf Sayısı	Klasik GDA
3	81,4
5	69
7	53

Sistemin doğruluğunu artırabilmek için sisteme n-GDA yöntemi uygulanmıştır. Yöntem ile her konu için bir eşik değeri belirlenmiştir ve eşik değerinden düşük değere sahip kelimeler sözlükten silinmiştir. Kelime sayısının azalması ile yeni konulardaki kelimelerin ağırlığı da artmıştır. Yeni oluşturulan sözlük ile tüm sınıflar için tutarlılık değerleri yeniden hesaplanmıştır. Son olarak sistem yeni konu sayıları ile iki seviyeli-GDA (2-GDA) yöntemiyle modellenmiştir. Tablo IV'te klasik GDA ve 2-GDA'nın başarısı gösterilmiştir. Sistemin başarısı 2-GDA ile beraber hedeflenildiği gibi artmıştır.

TABLO IV. KLASİK GDA ve 2-GDA'NIN DOĞRULUK KARŞILAŞTIRMASI (%)

	3	**5**	**7**
Klasik GDA	81.4	69	53
2-GDA	**90.3**	**76.5**	**57.6**

Makalenin ilerleyen bölümlerinde konu modellemede de kullanılan NMF yöntemi ile sistemin doğruluğu ölçülmüştür. Diğer yöntemler ile doğru karşılaştırma yapabilmek için NMF yönteminde de aynı konu sayları kullanılmıştır. NMF yöntemi ile konuların kelimeleri ve ağırlıkları elde edilmiştir. Diğer yöntemlerdeki gibi konulara ait uygun sınıf etiketi belirlenmiştir. Ardından sistem modellenerek doğruluğu ölçülmüştür. Tablo V'te NMF yöntemi ile modellenen sistemin doğruluğu gösterilmiştir.

TABLO V. NMF YÖNTEMLERİNİN TÜM SINIFLAR için DOĞRULUĞU (%)

Sınıf Sayısı	NMF
3	95
5	82.1
7	71.5

Son olarak sistem GAA ile modellenmiştir. Modelleme esnasında yine aynı konu sayıları kullanılmıştır. Konular, kelime ve ağırlık değerlerinden oluşmaktadır. Ancak, diğer yöntemlerin aksine konulardaki kelimelerin ağırlıkları pozitif ve negatif değerlere sahip olup, kelimelerin ağırlıklarına bakılarak uygun sınıf etiketi atanmıştır. Sonrasında eğitilen sistemin doğruluğu GAA yöntemi için hesaplanmıştır (Tablo VI).

TABLO VI. GAA YÖNTEMLERİNİN TÜM SINIFLAR için DOĞRULUĞU (%)

Sınıf Sayısı	GAA
3	90.8
5	85.3
7	81.3

Tüm yöntemlerin her sınıf için doğruluk değerleri Tablo VII'de gösterilmiştir. Tablo incelendiğinde tüm sınıflar için klasik GDA yöntemi en düşük doğruluğu vermiştir. En yüksek doğruluk değeri üç sınıf için NMF, diğer sınıflar için ise GAA olmuştur.

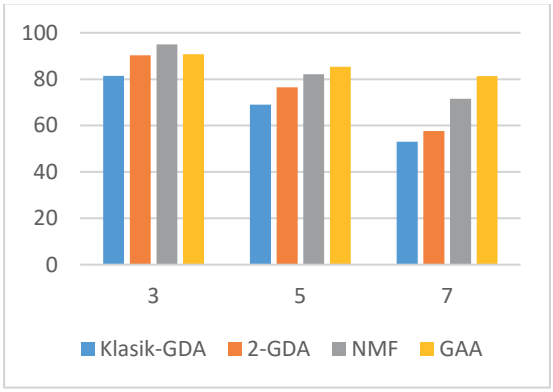
TABLO VII. TÜM YÖNTEMLERİNİN DOĞRULUK TABLOSU (%)

	3	**5**	**7**
Klasik GDA	81.4	69	53
2-GDA	90.3	76.5	57.6
NMF	**95**	82.1	71.5
GAA	90.8	**85.3**	**81.3**

IV. SONUÇ VE TARTIŞMA

Haber başlıklarının hangi türe ait olduğunu tespit etmek için konu modellemede kullanılan GDA, GAA ve NMF yöntemlerinden yararlanılmıştır. Ayrıca, önerilmiş olan n-GDA yöntemi de diğer yöntemler ile karşılaştırılmıştır. Kullanılan n-GDA yöntemi bu çalışmada 2 aşamalı olarak modellenmiştir. 2-GDA yöntemi, klasik GDA yöntemi ile karşılaştırıldığında tüm sınıflar için %4 ile %9 aralığında

doğruluk artışı sağlamıştır. Bunun nedeni olarak, sisteme olumsuz etki eden düşük ağırlıklı kelimelerin eşik değerine göre silinmesi gösterilebilir. Böylece, kalan kelimelerin ağırlığının artmasıyla daha kolay sınıf etiketi verilebilmektedir. Önerilen n-GDA yönteminin sisteme olumlu katkısı gösterilmiştir. Ardından sistem NMF ile modellenmiştir. NMF, önceki iki yöntemle karşılaştırıldığında tüm sınıflar için daha iyi doğruluğa ulaşmıştır. NMF'in doğruluğu, 2-GDA yöntemine göre %5 ile %14 arasında artmıştır.



Şekil **3.** Tüm yöntemlerin doğruluk grafiği (%)

Sistem son olarak GAA ile modellenerek doğruluğu ölçülmüştür. Tüm yöntemler ile kıyaslandığında GAA yöntemi, beş ve yedi sınıf için sırasıyla %82.1 ve %71.5 ile en yüksek doğruluğa ulaşan yöntem olmuştur. Ancak üç sınıf için, NMF yöntemi %95 ile daha iyi doğruluk vererek daha başarılı olmuştur. Tüm yöntemlerin doğruluk grafiği Şekil 3'te de gösterilmiştir. Şekil 3 incelendiğinde en başarısız yöntemin klasik GDA, en başarılının ise GAA olduğu gözükmektedir.

Gelecek çalışmalarımızda n-GDA yöntemini; metinle ilgili çalışmalarda diğer konu modelleme yöntemleri ile karşılaştırarak kullanabiliriz. Konu olarak metin özetlemenin yöntemlere etkisi, müziğin türünü belirleme veya kişinin duygusunu tespit etme seçebiliriz.

KAYNAKÇA

[1] D.M. Blei, ve J. D. Lafferty, "Dynamic Topic Models", Proceedings of the 23rd International Conference on Machine Learning, Pittsburgh, PA, 2006.

[2] R. Alghamdi, ve K. Alfalqi, "A Survey of Topic Modeling in Text Mining", International Journal of Advanced Computer Science and Applications, vol. 6, no. 1, pp. 147–153, 2015.

[3] M. W. Berry, M. Browne, A. N. Langville, V. P. Pauca, ve R. J. Plemmons, "Algorithms and Applications for Approximate Nonnegative Matrix Factorization", Computational Statistics & Data Analysis, vol. 52, no. 1, pp. 155–173, 2007.

[4] S. Bergamaschi, L. Po, ve S. Sorrentino, "Comparing Topic Models for a Movie Recommendation System", Proceedings of the 10th International Conference on Web Information Systems and Technologies, 2014.

[5] P. J. Crossno, A. T. Wilson, T. M. Shead, ve D. M. Dunlavy, "TopicView: Visually Comparing Topic Models of Text Collections", 2011 IEEE 23rd International Conference on Tools with Artificial Intelligence, 2011.

[6] K. Stevens, P. Kegelmeyer, D. Andrzejewski, ve D. Buttler, "Exploring Topic Coherence over many models and many topics", Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pp. 952–961, Jul. 2012.

[7] A. Utsumi, "Evaluating the performance of nonnegative matrix factorization for constructing semantic spaces: Comparison to latent semantic analysis", 2010 IEEE International Conference on Systems, Man and Cybernetics, 2010.

[8] P. Suri and N. R. Roy, "Comparison between LDA & NMF for event-detection from large text stream data", 2017 3rd International Conference on Computational Intelligence & Communication Technology (CICT), 2017.

[9] Y. Chen, H. Zhang, R. Liu, Z. Ye, and J. Lin, "Experimental explorations on short text topic mining between LDA and NMF based Schemes", Knowledge-Based Systems, vol. 163, pp. 1–13, 2019.

[10] Y. Chen, J.-B. Bordes, and D. Filliat, "An experimental comparison between NMF and LDA for active cross-situational object-word learning", 2016 Joint IEEE International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob), 2016.

[11] Z. A. Guven, B. Diri, and T. Cakaloglu, "Classification of New Titles by Two Stage Latent Dirichlet Allocation", 2018 Innovations in Intelligent Systems and Applications Conference (ASYU), 2018.

[12] M. Pavlinek and V. Podgorelec, "Text classification method based on self-training and LDA topic models", Expert Systems with Applications, vol. 80, pp. 83–93, 2017.

[13] Z. A. Guven, B. Diri, and T. Cakaloglu, "Classification of Turkish Tweet emotions by n-stage Latent Dirichlet Allocation", 2018 Electric Electronics, Computer Science, Biomedical Engineerings Meeting (EBBT), 2018.

[14] P. J. Kwantes, N. Derbentseva, Q. Lam, O. Vartanian, and H. H. Marmurek, "Assessing the Big Five personality traits with latent semantic analysis", Personality and Individual Differences, vol. 102, pp. 229–233, 2016.

[15] E. Altszyler, S. Ribeiro, M. Sigman, and D. F. Slezak, "The interpretation of dream meaning: Resolving ambiguity using Latent Semantic Analysis in a small corpus of text", Consciousness and Cognition, vol. 56, pp. 178–187, 2017.

[16] M. W. Berry and M. Browne, "Email Surveillance Using Non-negative Matrix Factorization", Computational and Mathematical Organization Theory, vol. 11, no. 3, pp. 249–264, 2005.

[17] A. N. Langville, C. D. Meyer, R. Albright, J. Cox, and D. Duling, "Algorithms, Initializations, and Convergence for the Nonnegative Matrix Factorization", arXiv, 2014.