Fırat Üniversitesi Müh. Bil. Dergisi
33(2), 599-606, 2021

Araştırma Makalesi
https://doi.org/10.35234/fumbd.899917

Türkçe Metinlerde Otomatik Konu Tespiti

Galip AYDIN[1*], İbrahim R. HALLAÇ[2]
[1,2] Bilgisayar Mühendisliği, Mühendislik Fakültesi, Fırat Üniversitesi, Elazığ, Türkiye
[*1] gaydin@firat.edu.tr, [2] irhallac@firat.edu.tr

(Geliş/Received: 19/03/2021; Kabul/Accepted: 09/04/2021)

Öz: Bu çalışmada çevrimiçi kullanılabilecek bir konu tespit sistemi önerilmiştir. Gizli Dirichlet Ayırımı ile 4 farklı kategoriye ait toplam 400.000 haber dokümandan oluşan bir Türkçe derlem eğitilmiştir. Model, eğitim verisinde yer almayan, yeni gelen dokümanların konu tespitini yüksek başarı ile gerçekleştirebilmektedir. Konu modellerinin başarı değerlendirmesinde tutarlılık (coherence) değerine ek olarak sınıflandırma yöntemleri için geçerli olan kesinlik (precision), hassasiyet (recall), F-ölçümü gibi skorların elde edilmesine yönelik 2 farklı yaklaşım geliştirilmiştir. Bu yaklaşımlarda, konular ile dokümanların ait olduğu sınıfların eşleştirilmesinden yararlanılmıştır. İlk yaklaşımda, dokümanın ait olduğu sınıfa karşılık gelen konunun mevcut olup olmadığı üzerinden genel bir başarı ölçütü sunulmuştur. İkinci yaklaşımda ise modelin yüksek güven (confidence) ile gerçekleştirmediği tahminleri eleyen, "dokümanın en belirgin konusu, ait olduğu sınıftır" kabulüne göre bir eşik (threshold) değeri üzerinden değerlendirme yapılan bir ölçüt sunulmuştur. Önerilen başarı değerlendirme yöntemlerine göre sırasıyla %94.2 ve %90.9 doğrulukta konu tespiti başarısı elde edilmiştir.

Anahtar kelimeler: Konu modelleme, gizli dirichlet ayırımı, doğal dil işleme, etiketli metinlerde konu tespiti, LDA için başarı değerlendirmesi.

Automatic Topic Detection on Turkish Text

Abstract: In this study, we propose a topic modelling system that can be used online. A Turkish corpus consisting of a total of 400,000 news documents belonging to 4 different categories was trained with Latent Dirichlet Allocation algorithm. The model can successfully identify the topic of new documents that are not seen in the training data. In addition to the coherence value, in the assessment of topic models, 2 different approaches have been developed to obtain evaluation scores such as precision, recall, and F-measure, which are valid for classification methods. In these approaches, we benefit from the topic and document label matches. In the first approach, a general accuracy measure is presented based on whether the topic corresponding to the class of the document is predicted or not. Our second approach eliminates the predictions that the model does not make with high confidence. The model is evaluated over a threshold value according to the assumption that "the most significant topic of the document is the class it belongs to". Topic detection success was achieved with an accuracy of 94.2% and 90.9% according to the proposed evaluation methods respectively.

Key words: Topic modeling, latent dirichlet allocation, natural language processing, topic detection on labelled text, performance evaluation for lda.

1. Giriş

Kısaca, bir dokümanda hangi konuların geçtiğini otomatik olarak tespit eden yöntemlere doğal dil işleme (DDİ) alanında konu modelleme denir. Tespit edilen konular dokümanların özetlenmesinde kullanılabileceği gibi diğer metin analizi çalışmaları için özellik çıkartımı amacıyla da kullanılabilir [1]. Örneğin; A dokümanı hangi konuları içermekte? A ve B dokümanları ne kadar benzerdir? T konusunu araştıran bir kişi A ve B dokümanlarından önce hangisini okumalıdır? Sorularına cevap ararken ilk basamak olarak özellik çıkarımı amacıyla konu tespitinden yararlanılabilir. Daha sonra, konu tespiti yapılan dokümanlar, sadece içerdiği konular üzerinden (feature extraction model) kümelenebilir veya sınıflandırılabilir.

Gözetimsiz makine öğrenmesi yöntemlerinden yararlanılarak metin benzerliği [2], kümeleme [3], sosyal medya kullanıcılarının benzerliği [4] çözümleri geliştirilmektedir. Önceden tanımlı konu etiketleri mevcut olmadığı için, konu tespiti yöntemleri de genellikle denetimsiz makine öğrenmesi grubundaki algoritmalardan yararlanır.

Konu modelleme analizi sonunda, bir arada gruplandırılmış doküman koleksiyonları elde edilir. Gruplanmış dokümanların yanı sıra bu ilişkileri ortaya çıkarmak için tespit edilen kelime kümeleri de oluşturulur. Konu modelleme, metin kümeleme gibi önceden belirli olmayan konuların gözetimsiz olarak belirlenmesidir.

[*] Sorumlu yazar: gaydin@firat.edu.tr. Yazarların ORCID Numarası: [1] 0000-0002-9564-3329, [2] 0000-0003-0568-3114

Deterministik olmayan konu modelleme yöntemlerinde, algoritma her yürütüldüğünde farklı sonuç alınması mümkündür. Metin sınıflandırmada ise sınıflar önceden bilinir ve sabittir. Yani, belirli sayıda, daha önceden belirlenmiş sabit sınıflar söz konusudur. Çok sınıflı sınıflandırma yöntemleri dışındaki yöntemlerde, birbirinden ayrık etiketler belirlenir. Bir sınıfa ait olan doküman aynı anda başka bir sınıfa dahil edilmez. Konu modellemede ise böyle bir durum söz konusu değildir. Sınıflar birbirini dışlamaz, dokümanın hangi oranlarda hangi konuları içerdiği bulunur.

2. Konu Tespiti Yöntemleri

Literatürdeki konu modelleme yöntemleri incelendiğinde çoğunlukla kelimelerin dokümanlarda bulunma durumları üzerinden geliştirilmiş istatistiksel modeller olduğu görülmektedir [5]. Gizli Dirichlet Ayırımı (GDA), Gizli Anlamsal Analiz (GAA), çok değişkenli analiz ve lineer cebir tabanlı Negatif Olmayan Matris Faktörizasyonu (NMF) en bilinen konu modelleme yöntemlerindendir. Şekil 1'de restoran hakkında yapılan yorumlar üzerinden gerçekleştirilen otomatik konu tespiti sonucunda elde edilen dört farklı konu grubu ve bu konuları temsil eden kelimeler verilmiştir.

Konu 32	Konu 67	Konu 78	Konu 98
Kahvaltı	Salata	İçki	Pazar kahvaltısı
Kahve	Pancar	Kokteyl	Yumurta
Patates	Keçi peyniri	Akşam yemeği	Beklemek
Yumurta	Salata sosu	Liste	Mimoza
Meyve	Taraf	Martini	Omlet
Sosis	Salatalık	Alkol	Yumurtalı ekmek
Fransız tostu	Marul	Tur	Krep
Gözleme	Göğüs biftek	New York	Çılbır
İrmik	Kızarmış ekmek	Karışım	Öğleyin
Fransız ekmeği	Öneri	Likör	Somon füme

Şekil 1. Örnek konu tespiti sonucu [5].

Şekilde 1'de görüldüğü üzere bir insan tarafından yorumlandığında aynı konu ile ilişkilendirilecek kelimelerin konu modelleme yöntemleriyle semantik olarak büyük oranda doğru bir şekilde tespit edilmesi mümkündür. Fakat bu konulara bir insanmış gibi başlık üretmek için diğer DDİ yöntemleriyle birlikte ilave çalışmaların yapılması gerekir.

2.1. Gizli Dirichlet Ayırımı Yöntemi

Latent Dirichlet Allocation (LDA) algoritması Türkçe'siyle Gizli Dirichlet Ayırımı (GDA) çoğu konu tespiti modeli gibi denetisimsiz (unsupervised) bir yöntemdir. Bu yöntemde dokümanlarda bulunan kelimeler modellenir. Amaç, dokümanlar içerisinde en çok olabilirliği (likelihood) veya sonsal olasılığı (posterior probability) en iyi sağlayacak şekilde konuları çıkarmaktır [6].

Olasılıksal bir teknik olan GDA'da $P(\ kelime \mid konu)$ and $P(\ konu \mid dok\ddot{u}man)$ olasılıkları üzerinden bir en iyileme gerçekleştirmeyi amaçlar. GDA yönteminde doküman kümelemeye benzer olarak, önceden belirlenen K sayısı kadar konu dokümanlarla rasgele eşleştirilerek konu modellemesi yapılır. Her bir dokümanı en iyi temsil eden konu belirlenirken bir yandan da her konuyu temsil eden en iyi kelimelerin tespiti yapılarak bir öğrenme gerçekleştirilir.

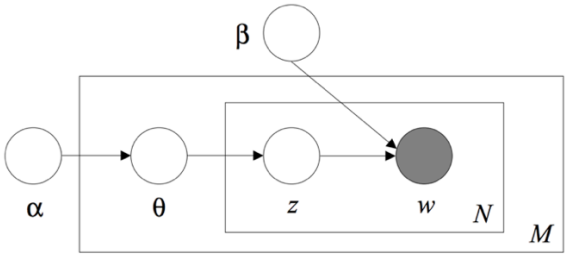
Tablo 1. GDA amaç fonksiyonu parametreleri.

Terim	Açıklama
K	Konu sayısı
n_{ik}	i. dokümanda bulunan k. konuya atanmış kelime sayısı
N_i	Dokümanda bulunan toplam kelime sayısı
a	Hangi konunun hangi dokümana atandığını gösteren parametre

İlk adımda konular ve dokümanlar rasgele eşleştirildikten sonra bu atamaların ne kadar anlamlı olduğunun hesaplanabilmesi için çeşitli istatistiklere bakılır. Bu aşamada her dokümanda her bir konu için hangi kelimelerin

ilişkilendirildiği yerel (local) istatistik ve tüm dokümanlarda her bir kelimenin konularla ilişkilendirilmesi oranı genel (global) istatistiklere bakılır [5]. Daha sonra bu bilgiler Denklem 1'e göre güncellenir. Denklemde yer alan terimlerin açıklamaları Tablo 1'de verilmiştir.

$$\frac{n_{ik}+a}{N_i-1+Ka} \tag{1}$$



Şekil 2. GDA algoritmasının levha gösterimi [1].

Levha gösterimi (plate notation) birçok değişken arasındaki bağımlılıkların ifade edilmesinde kullanılan bir yöntemdir. GDA algoritmasının grafiksel olarak gösterimi Şekil 2'de verilmiştir. Kutular, kopyaları temsil eden "plakalar" dır. Dış plaka belgeleri temsil ederken, iç plaka belge içindeki konuların ve kelimelerin tekrarlanan seçimini temsil eder. M, doküman sayısını, N bir dokümandaki kelime sayısını belirtir. Gösterimde kullanılan sembollerin açıklamaları Tablo 2'de verilmiştir.

Tablo 2. Levha gösterimi açıklamaları [7].

Terim	Açıklama
β	Her bir konu için kelime atamalarını gösteren Dirichlet parametresi
a	Her bir doküman için konu atamalarını gösteren Dirichlet parametresi
θ	θ dokümanı için konu dağılımı
z	Her bir kelime için atanan konular
w	Anlık incelenen kelime

3. Literatür Taraması

Konu modelleme genellikle yarı-denetimli (semi-supervised) veya denetimsiz (unsupervised) bir makine öğrenmesi yöntemi olarak ele alındığından dolayı önceden etiketli veriler üzerinde test edildiği çalışmalar pek görülmemektedir. Çoğunlukla bir konu için en olası kelimelerin anlamsal olarak ne kadar alakalı olduğuna bağlı olarak kümelerin doğruluğu ölçülür.

Güven ve arkadaşları, 5 farklı duygu türü içeren (kızgın, korkmuş, mutlu, üzgün, şaşkın) 4000 adet tweetin GDA algoritmasıyla duygu sınıflandırılmasını gerçekleştirmiştir [8]. Bir başka çalışmalarında yine GDA ile 7 sınıfa ait 4200 adet Türkçe haber başlıklarından oluşan veri seti üzerinde ekonomi, spor ve yaşam gibi konular için konu modellemesi algoritmalarının başarı karşılaştırmasını yapmıştır [9]. Sınıf sayısının farklı tutulduğu farklı deneylerde NMF yöntemi 3 sınıf için iyi başarıyı gösterirken, 7 sınıf için en iyi başarıyı Gizli Anlamsal Analiz (GAA) yöntemi göstermiştir.

W2E veri setinde ABD seçimleri, İngiltere Avrupa Birliği Referandumu, Yaz Olimpiyatları gibi birçok olayın meydana geldiği 2016 yılına ait Wikipedia güncel olaylar portalında yer alan dokümanlar ile oluşturulmuştur. (Wikipedia's Current Event portal -WCEP). Toplamda 5,160 olaydan oluşan veri seti 10 kategoriye ve 3083 farklı konuya bölünmüştür [10].

Jin ve arkadaşları derin öğrenme yöntemlerinin konu modelleme teknikleriyle birleştirilmesi amacıyla gerçekleştirdiği çalışmada NMF (Topic matrix factorization) algoritmasıyla LSTM modelini bir araya getirerek Amazon web sitesindeki kullanıcı yorumlarının hangi konuları içerdiğinin tespitini gerçekleştirmiştir [11].

4. Veri Seti, Yöntem ve Hesaplama Ortamı

GDA yöntemi denetimsiz bir öğrenme tekniği olduğu için normal şartlar altında doğruluk değeri, hassaslık gibi metriklerle model performansının değerlendirilme olanağı yoktur ve hangi konuların tespit edilmesi gerektiği üzerinde bir kontrol sağlamak mümkün değildir. Bu çalışmada, etiketli veriler üzerinde denetimsiz bir öğrenme gerçekleştirildikten sonra gerçekleştirilen konu modelleme başarısının dokümanların ait olduğu sınıf bilgisinden yararlanmasını sağlayan iki farklı başarı ölçütü önerilmiştir.

Veri seti olarak 4 farklı kategori için otomatik olarak toplanmış haber dokümanları kullanılmıştır. Her kategoriden 100.000 adet olmak üzere Ekonomi, Spor, Siyaset ve Kültür haberlerinden oluşan toplam 400.000 doküman bulunmaktadır. Bu veri seti ***Bigailab-5news-500K***'nın bir alt kümesidir [12].

Konu modeli performans değerlendirmesi için önerilen yöntemin açıklanması amacıyla güncel haberlerden manuel olarak elde edilen 400 adet haber dokümanı test verisi olarak kullanılmıştır. Çalışmanın uygulama kodları ve veri setlerine *GitHub:irhallac/Gensim-LDA-news-Turkish* proje dizini üzerinden erişilebilmektedir. Konu modelleme yöntemlerinin implementasyonu için yararlanılabilecek en popüler kütüphane Python Gensim paketidir [13]. Bu paketin `gensim.models.ldamodel.LdaModel(`***`parametreler`***`)` metoduna ait parametreler Tablo 3'te belirtilmiştir.

Gensim paketinin *ldamodel* kütüphanesinin kullanılması öncesinde dokümanlardan oluşan veriseti için bir takım ön işleme adımlarının gerçekleştirilmesi gerekmektedir. Uyguladığımız temel ön işleme adımları şu şekildedir:

1. Bütün harfleri küçük harflere dönüştür.
2. Noktalama işaretleri ve alfanümerik olmayan birimleri temizle.
3. Gereksiz kelime (stopwords) listesinde yer alan kelimeleri temizle.
4. Dokümanların %50'sinden daha fazlasında yer alan aşırı yüksek frekanslı kelimeleri temizle.
5. Frekansı 20'den fazla olan kelime ikilileri (bigram) ve kelime üçlülerini (trigram) bulundukları dokümanlara ilave et.
6. Doküman temsillerini oluştur.

Tablo 3. Gensim parametreleri ve açıklamaları.

Parametre	Açıklama
`corpus`	Vektörleştirilmiş dokümanların listesi
`id2word`	Kelime id'lerinden kelimeleri gösteren değişken (Mapping)
`num_topics`	Dokümanlardan çıkarılması istenen konu sayısı
`random_state`	Aynı sonuçların yeniden üretilebilmesini sağlayan random seed belirleyici
`update_every`	Model parametrelerinin hangi sıklıkta güncelleneceğini belirler
`chunksize`	Her bir eğitim iterasyonunda kullanılacak doküman sayısı Örn. 100
`passes`	Toplam eğitim iterasyon sayısı
`alpha`	Konular arasındaki seyrekliği etkileyen katsayı. Varsayılan değeri 1/konu_sayısı
`per_word_topics`	Modelin aynı zamanda her kelime için en olası konuların azalan sırasına göre sıralanmış bir konu listesi şeklinde bulunması isteniyorsa **true** değeri verilir.

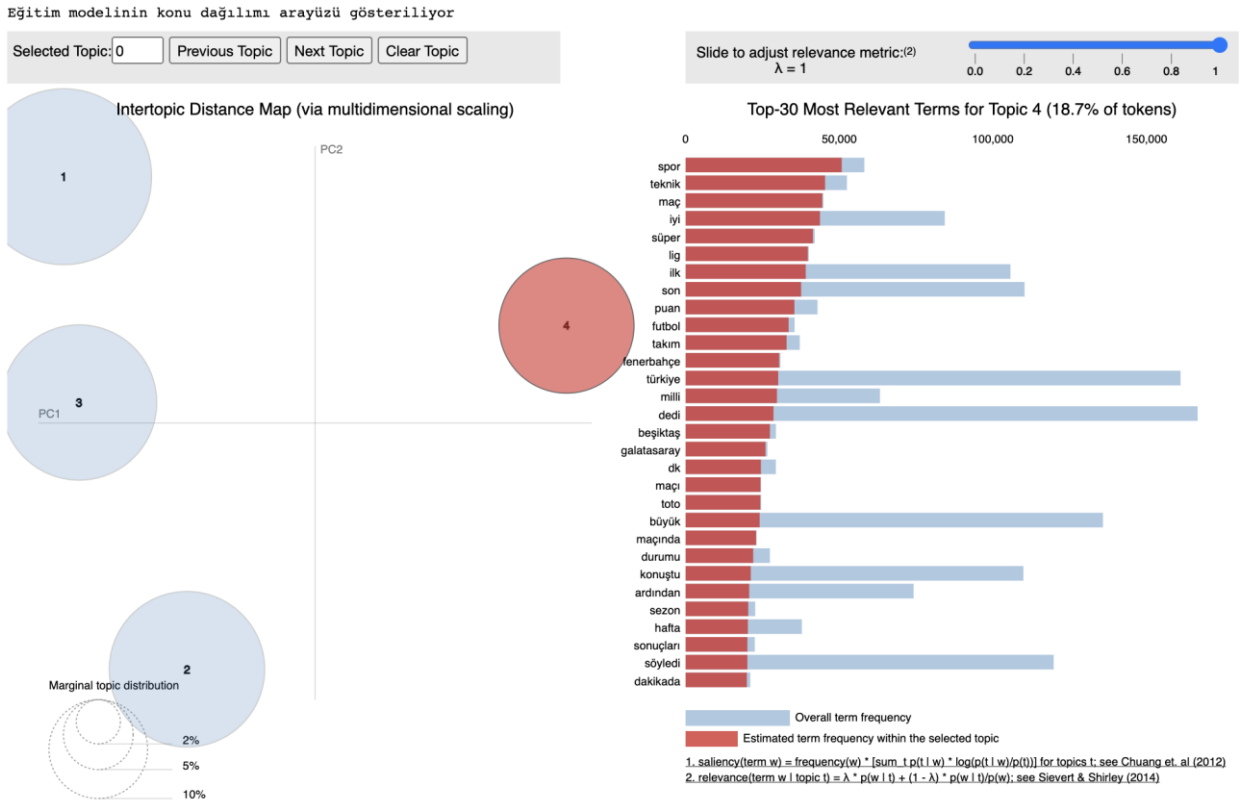
Model, dokümanları vektörleştirilmiş formatta kabul etmektedir. GDA algoritması, kelimelerin dokümanlarda ve derlem içerisinde bulunma frekansına göre hesaplanan olasılıklar üzerinden çalıştığı için en doğru yöntemin kelime frekanslarıyla metin temsili olduğu anlaşılmaktadır. Bundan dolayı kelime çantası (bag-of-words) temsilleri kullanılmıştır. Derleme ait sözlükteki belirlenen bütün kelimeleri ağırlıklandıran TF-IDF yöntemi tercih edilmemiştir.

Gesim'de, `Lda_model.show_topics(num_topics=5,num_words=10)` metoduyla modelin eğitimi tamamlandığında çıkarılan 5 adet konu, o konuları en iyi temsil eden 10'ar kelimeyle gösterilir. Konuların ne olduğu bu kelimeler üzerinden muhakeme yaparak tespit edilebilir. GDA, NMF algoritmasına kıyasla daha büyük veri setlerindeki konular için ürettiği anahtar kelimeler bakımından daha insan gözüyle ayırt edilebilir konular üretmektedir [14]. NMF algoritması için ise `sklearn.decomposition.NMF` paketi kullanılabilir [15]. Bu çalışmada sadece GDA algoritması kullanıldığı için bu paket detaylı olarak incelenmemiştir.

Uygulamalar için **48-core Intel(R) Xeon(R) CPU E5, 256GB RAM ve E5-2650** işlemcisi olan bir makine kullanılmıştır. Tablo 3'te Gensim kütüphanesinin standart konu tespit modeli sınıfı gösterilmiştir. Çalışmada, bu sınıfın paralel yürütme özelliğini taşıyan versiyonu olan `LdaMulticore` sınıfı kullanılmıştır. İş parçacıkları (thread) yardımıyla eğitimin kısmen paralelleştirilmesi mümkündür. Bu özellik kullanılırken, `workers` parametresine kullanılacak işlemci sayısı atanır. Kullanılan makine 48 çekirdeğe sahip olduğu için en fazla 48 iş parçacığı yürütebilmektedir. Model eğitiminde bu parametre 36 olarak belirlenmiştir.

5. Sonuç, Değerlendirme ve Gelecekteki Çalışmalar

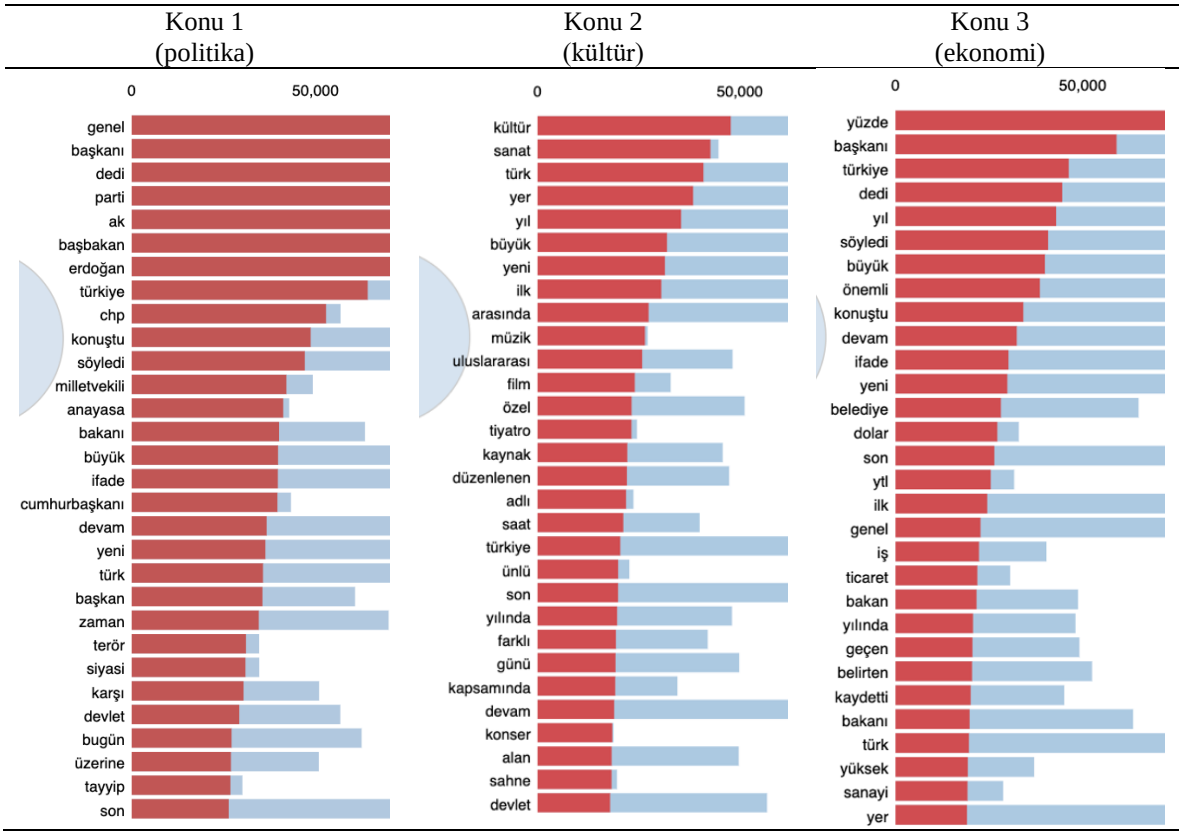
Model 400 bin haber dokümanıyla, 100 iterasyon eğitildikten sonra farklı sayılarda konu ve bu konuları iyi temsil eden kelimelerin olasılıksal sıralamaları elde edilmiştir. Şekil-3'te 4 konunun dağılımı PyDavis (*https://pypi.org/project/pydavis*) aracı ile görsel olarak sunulmuştur. Modelin ortalama-konu-uyumluluğu değeri (average-topic-coherence) −1.7443 olarak elde edilmiştir. Uyumluluk değerinin yüksek olması tercih edilir [16]. Konu id'si 4 olan grubun en alakalı 3 kelimesi "spor", "teknik" ve "maç" kelimeleridir. Bu konunun etiketli veri setinin spor haberleriyle ilgili dokümanlarına karşılık geldiği anlaşılmaktadır.



Şekil 3. Model konu dağılımı ara yüzü.

1, 2 ve 3 numaralı konular, ilişkilendirildikleri sınıflar ve en alakalı kelimelerin yer aldığı bilgiler Şekil 4'te yer almaktadır. Şekilde, Temel Bileşenler Analizi (Principal Component Analysis — PCA) ilk iki değeri PC1 ve PC2 olarak gösterilmiştir. Konuların gösteriminde [0, 50000, 100000 …] üzerinden çizdirilen çubuk grafiklerin mavi renkli alanı ilgili kelimenin bütün derlem üzerinden frekansını, kırmızı renkli alanı ise seçilen konunun tespit edildiği dokümanlar üzerinden frekansını belirtmektedir.

Bu çalışmada çevrimiçi kullanılabilecek bir konu tespit sistemi önerilmiştir. GDA ile büyük miktarda dokümandan oluşan bir derlem eğitilmiştir. Model, eğitim verisinde yer almayan yeni gelen dokümanların konu tespitini yüksek başarı ile gerçekleştirebilmektedir. Sistemin ne kadar başarılı olduğunun ölçülebilmesine olanak sağlayan 2 farklı değerlendirme hesaplaması önerilmiştir. Bu hesaplamalarla, haber verileri içerisinde konu tespiti gerçekleştirilirken etiketli verilerden yararlanılarak TP (True positive — Doğru Pozitif), FP (False positive — Yanlış Pozitif), TN (True negative — Doğru Negatif), FN (False negative — Yanlış Negatif) değerleri elde edilerek kapsamlı bir başarı ölçütü ortaya konulmuştur. Kesinlik (Precision), hassasiyet (recall), F-ölçümü (F1 skoru) değerleri hesaplanmıştır. Bunun için konu sayısının değeri sınıf sayısı olarak seçilmiştir. Konu sayısının arttırılması durumunda bir sınıf için ilişkili konu sayısı birden fazla olmakta ve uygulanan yöntemde bir probleme neden olmamaktadır.



Şekil 4. Konu ve sınıf eşleştirmeleri.

Birinci başarı testine göre, model, bir doküman içerisinde geçen konuları tahmin ettiğinde, %94.2 ortalama ile o dokümanın ait olduğu sınıfa karşılık gelen konunun mevcut olduğunu doğru bilmiştir. Bu ölçüte göre detaylı sonuçlar Tablo 4'te verilmiştir.

İkinci başarı testinde, "dokümanın en belirgin konusu, ait olduğu sınıftır" kabulüne göre bir eşik (threshold) değeri üzerinden gerçekleştiren tahminde %90.9 başarı elde edilmiştir. Bu değer 4 sınıf için elde edilen ağırlıklı kesinlik skoru ortalamasıdır. Testte kullanılan eşik değeri ile en belirgin konunun tahmininde, sadece bu değerin üzerindeki olasılıkla tahmin edilen tahminler kullanılmıştır. Yani, modelin yüksek güven (confidence) ile gerçekleştirmediği tahminler dikkate alınmamıştır. Test sonucundaki destek örnek sayısının 400'den 225'e düştüğüne dikkat edilmelidir. Eşik değeri küçüldükçe değerlendirme alınan örnek sayısı artacaktır. Bu ölçüte göre detaylı sonuçlar Tablo 5'te verilmiştir.

Tablo 4. Başarı ölçütü-1'e göre test sonucu.

	Kesinlik	Hassasiyet	F1-Skoru	Destek
Konu-1	1.000	0.850	0.919	100
Konu-2	1.000	0.850	0.919	100
Konu-3	0.769	1.000	0.870	100
Konu-4	1.000	1.000	1.000	100
Mikro ortalama	0.925	0.925	0.925	400
Makro ortalama	0.942	0.925	0.927	400
Ağırlıklı ortalama	0.942	0.925	0.927	400

Tablo 5. Başarı ölçütü-2'ye göre test sonucu.

	Kesinlik	Hassasiyet	F1-Skoru	Destek
Konu-1	1.000	0.533	0.696	30
Konu-2	1.000	0.708	0.829	48
Konu-3	0.733	1.000	0.846	77
Konu-4	1.000	1.000	1.000	70
Mikro ortalama	0.876	0. 876	0. 876	225
Makro ortalama	0.933	0.810	0.843	225
Ağırlıklı ortalama	0.909	0.876	0.870	225

Önerilen iki başarı değerlendirme sisteminde de bir dokümanda birden fazla konunun tespit edilmesi olasılığı göz önünde bulundurulmuştur. Birinci yöntemde, birden fazla konu mevcut olsa da doküman etiketine karşılık gelen konunun tespit edilip edilmediği üzerinden bir doğruluk testi yapılmaktadır. İkinci yöntemde ise, tespit edilen konular arasında yüksek olasılığa sahip konunun doküman etiketine karşılık gelen konu olması dikkate alınmıştır.

Bu çalışmanın devamında, çevrimiçi haber sitelerine ek olarak farklı kaynaklardan, farklı sınıflara ait veri setleri kullanılarak önerdiğimiz başarı değerlendirme yöntemlerinin tutarlılık değeriyle ilişkisini araştırmayı planlamaktayız. Ayrıca, konu sayısının sınıf sayısından bağımsız seçilmesi durumunda semantik eşleştirmelerin otomatik olarak yapılmasına yönelik yöntemlerin bulunması önemlidir.

Teşekkür

Bu çalışma Savunma Sanayii Müsteşarlığı, Savunma Geniş Alan (SAGA) kapsamında desteklenen Derin Öğrenme Büyük Veri Analiz Platformu (Değirmen) Projesi kapsamında gerçekleştirilmiştir.

Kaynaklar

[1] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," J. Mach. Learn. Res., vol. 3, no. Jan, pp. 993–1022, 2003.

[2] Y. Ko and J. Seo, "Automatic text categorization by unsupervised learning," 2000.

[3] A. Budiarto, R. Rahutomo, H. N. Putra, T. W. Cenggoro, M. F. Kacamarga, and B. Pardamean, "Unsupervised News Topic Modelling with Doc2Vec and Spherical Clustering," Procedia Comput. Sci., vol. 179, pp. 40–46, 2021.

[4] I. R. Hallac, S. Makinist, B. Ay, and G. Aydin, "user2Vec: Social Media User Representation Based on Distributed Document Embeddings," in 2019 International Artificial Intelligence and Data Processing Symposium (IDAP), 2019, pp. 1–5.

[5] E. Ekinci, "Dokümanların Anlamsal Benzerliklerine Dayalı Özgün Bir Konu Modelleme Yöntemi," 2019.

[6] J. D. Mcauliffe and D. M. Blei, "Supervised topic models," in Advances in neural information processing systems, 2008, pp. 121–128.
[7] D. Ramamonjisoa, "Topic modeling on users's comments," in 2014 Third ICT International Student Project Conference (ICT-ISPC), 2014, pp. 177–180.
[8] Z. A. Guven, B. Diri, and T. Cakaloglu, "Classification of TurkishTweet emotions by n- stage Latent Dirichlet Allocation," in 2018 Electric Electronics, Computer Science, Biomedical Engineerings' Meeting, EBBT 2018, 2018, pp. 1–4, doi: 10.1109/EBBT.2018.8391454.
[9] Z. A. Guven, B. Diri, and T. Cakaloglu, "Comparison of Topic Modeling Methods for Type Detection of Turkish News," in UBMK 2019 - Proceedings, 4th International Conference on Computer Science and Engineering, 2019, pp. 150–154, doi: 10.1109/UBMK.2019.8907050.
[10] T.-A. Hoang, K. D. Vo, and W. Nejdl, "W2E: A Worldwide-Event Benchmark Dataset for Topic Detection and Tracking," in Proceedings of the 27th ACM International Conference on Information and Knowledge Management, 2018, pp. 1847–1850.
[11] M. Jin, X. Luo, H. Zhu, and H. H. Zhuo, "Combining deep learning and topic modeling for review understanding in context-aware recommendation," in Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), 2018, pp. 1605–1614.
[12] I. R. Hallac, B. Ay, and G. Aydin, "Experiments on Fine Tuning Deep Learning Models With News Data For Tweet Classification," in 2018 International Conference on Artificial Intelligence and Data Processing (IDAP), 2018, pp. 1–5.
[13] R. Rehurek and P. Sojka, "Gensim--python framework for vector space modelling," NLP Centre, Fac. Informatics, Masaryk Univ. Brno, Czech Repub., vol. 3, no. 2, 2011.
[14] D. Mahapatra, M. Maddukuri, and G. Jayadev, "Topic Modelling," 2016.
[15] Sklearn, https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.NMF.html
[16] D. Mimno, H. Wallach, E. Talley, M. Leenders, and A. McCallum, "Optimizing semantic coherence in topic models," in Proceedings of the 2011 conference on empirical methods in natural language processing, 2011, pp. 262–272.