Classification of New Titles by Two Stage Latent Dirichlet Allocation

Conference Paper · December 2018

DOI: 10.1109/ASYU.2018.8554027

CITATIONS
6

READS
353

3 authors:

Zekeriya Anıl Güven
İzmir Bakırçay University
15 PUBLICATIONS **60** CITATIONS
SEE PROFILE

Tolgahan Cakaloglu
Walmart Labs
17 PUBLICATIONS **47** CITATIONS
SEE PROFILE

Banu Diri
Yildiz Technical University
174 PUBLICATIONS **2,732** CITATIONS
SEE PROFILE

İki Aşamalı Gizli Dirichlet Ayırımı ile Haber Başlıklarının Sınıflandırılması

Classification of New Titles by Two Stage Latent Dirichlet Allocation

Zekeriya Anıl Güven
Bilişim Sistemleri Mühendisliği
Recep Tayyip Erdoğan Üniversitesi
Rize, Türkiye
zekeriyaanil.guven@erdogan.edu.tr

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

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

***Özetçe*—Internetin hızlı gelişimi ile her gün değişik kanallardan binlerce farklı habere ait dokümanlar bizlere sunulmaktadır. Bu kadar haberin, özellikle medya sektöründe, insan emeği olmadan sınıflandırılarak arşivlenmesi önemli bir sorundur. Bu çalışmada, haber sitelerinden toplanan geniş içerikli haber başlıklarının hangi haber türüne ait olduğunun tespit edilmesi amaçlanmıştır. Bunun için konu modellemede kullanılan klasik Gizli Dirichlet Ayırımı (GDA) algoritması referans alınarak geliştirilmiş iki aşamalı bir yöntem önerilmiştir. Geliştirilen iki aşamalı GDA yöntemiyle, klasik GDA'nın karşılaştırılması yapılmıştır. Ardından konulara ait kelime ağırlıklarından arff uzantılı dosya oluşturularak Weka içerisindeki makine öğrenme yöntemlerinin başarısı ölçülmüştür.**

Anahtar Kelimeler — Konu Modelleme, Gizli Dirichlet Ayırımı, Doğal Dil İşleme, Haber Analizi, Makine Öğrenmesi.

***Abstract*—With the rapid development of the Internet, thousands of different news reports from different channels are presented to us. So much news, particularly in the media sector, is an important question to be categorized and archived without human effort. In this study, it is aimed to be able to determine which news item belongs to large news headlines collected from news sites. For this, a two stage method is proposed, which is based on the classical Latent Dirichlet Allocation (LDA) algorithm used in the model. With the developed two stage LDA method, comparison of the conventional LDA was made. Then, by creating a file with an arff extension from the word weights of the topics, the success of the machine learning methods in Weka was measured.**

Keywords — Topic Modelling, Latent Dirichlet Allocation, Natural Language Processing, New Analysis, Machine Learning.

I. GİRİŞ

Konu modelleme, metin belgesinin anlamsal yapısını belirleyen bir makine öğrenmesi yöntemi olup, doğal dil işleme araştırma alanıdır. Konu modelleme yöntemleri ile yüksek içeriğe sahip metin belgesi organize edilebilir ve özetlenebilir [1]. Konu modelleme, otomatik belge indeksleme, belge sınıflandırma, konu keşfi gibi birçok alanda başarıyla uygulanabilmektedir [2]. Metin belgesi konu modelleme ile konuların birleşimi olarak gösterilebilir. Konular, kelimeler üzerinde bir olasılık dağılımı olarak hesaplanırken; metin belgeleri de konular üzerinde bir olasılık dağılımı olarak hesaplanmaktadır [3].

Gizli Dirichlet Ayırımı (GDA) ile yapılan literatür araştırmalarında; metin madenciliğiyle alakalı makaleler araştırılmıştır. Çelikyılmaz ve diğerleri [4], GDA modelinin soru cevaplama sistemine uygulanmasını incelemiştir. Kullanıcının sorduğu soru ile aday cevaplar arasındaki benzerlik ölçütlerinin bulunup sıralanması GDA ile yapılmıştır. Bir diğer çalışma metin sınıflandırma için GDA'nın kullanılmasıdır. Bu çalışmada bir dokümandaki kelime sıklığına (tf), kelimenin birden fazla dokümanda geçme sıklığına (idf) göre sözlüğe eklenmesini kontrol eden kelime özellik modelleriyle GDA'nın karşılaştırması yapılmıştır [5]. Çelikyılmaz ve diğerleri [6], konuşma anlama üzerine semantik bir işlem uygulamışlar ve konuşma anlama sisteminde semantik yapıyı öğrenmek için gizli n-gram kümeleme ve yarı denetlenmiş GDA kullanmışlardır. Geliştirilen GDA yöntemiyle elde edilen konu semantik yapı için öğrenme modeline ek kısıtlama getirmiştir. Ürün özelliklerinin çıkarılmasında da konu modelleri kullanılmaktadır. Titov ve diğerleri [7], geliştirdikleri Çok Tanecikli GDA ile çıkarılan lokal konuları oylanan özelliklerin, global konuların ise ürün özelliklerinin çıkartılmasında kullanmışlardır. Lin ve diğerleri [8], geliştirdikleri denetimsiz ve GDA tabanlı bir yöntem olan Joint Sentiment/Topic Model ile sinema yorumlarından ürün özelliklerini ve duygu ifadelerini eş zamanlı olarak çıkarmışlardır. Lee ve diğerleri [9], tüketici yorumlarının içerisinden değerli olan bilgileri çıkarmak amacıyla algısal bir harita ve radar diyagramı oluşturarak farklı firmalara ait ürünlerin karşılaştırılmasını yapmak için Mining Conceptual Map isimli bir model tasarlamışlardır. Çalışmada sanal dokümanlar oluşturulup, ağırlıklı GDA ile ürünlerin özellikleri çıkarılmıştır. Chatterjee ve diğerleri [10], Twitter verilerinden konuları çıkarmak amacıyla GDA'nın gelişmiş versiyonları FB-LDA ve RCB-LDA ile duygu ifadelerinin sınıflandırılması için de SentiStrength ve yarı denetimli Destek Vektör Makinelerini kullanılmışlardır. Poria ve diğerleri [11], GDA'ya kavramlar arası ilişki bilgisini de dahil ederek sözdizimsel bir GDA'dan anlamsal bir GDA'ya geçiş yaparak, istatistiksel bir yöntem yerine kelimeler arası anlamsal ilişkiden yararlanarak başarılı bir kümeleme işlemi yapmışlardır. Feuerriegel ve diğerleri [12], GDA'yı finans haberlerindeki konuları çıkartarak bu konuların Alman borsasını nasıl etkilediğini belirlemek için kullanmışlardır.

Türkçe ve İngilizce haber başlıklarının hangi haber türüne ait olduğunu tespit etmek için birçok yöntem kullanılmaktadır. Bu çalışmada en önemli konu modelleme tekniği olan GDA algoritması kullanılmıştır. Algoritma iki aşamalı olacak şekilde geliştirilmiştir. Bu doğrultuda, GDA'ya dayalı olarak temsil edilen haberlerle iki aşamalı geliştirilen GDA yönteminin karşılaştırılması yapılmıştır.

978-1-5386-7786-5/18/$31.00 ©2018 IEEE

Çalışmanın ikinci bölümünde, kullanılan yöntem, veri seti ve üzerinde yapılan ön işlemlerden bahsedilmiştir. Üçüncü bölümde haber verileri üzerinde gerçekleştirilen konu analizi çalışmalarına yer verilmiş ve yöntemler arası çalışmaların sonuçları gösterilmiştir. Dördüncü bölümde ise çalışmanın değerlendirme ve sonuçlarına yer verilmiştir.

II. MATERYALLER

A. Gizli Dirichlet Ayırımı Algoritması

GDA, olasılık tabanlı bir konu modelleme yöntemidir. Model bir dizi dokümandan kelime ağırlığına dayalı olarak konuları oluşturmaktadır. GDA yönteminde, metin belgesi konuların birleşimi olarak temsil edilmektedir. Yöntemin temelinde, konular kelimeler üzerinde bir olasılık dağılımı, metin belgeleri de konular üzerinde bir olasılık dağılımı olarak temsil edilmektedir. Her bir konu ise sabit kelime seti üzerinde bir dağılım olarak modellenmektedir [13]. Model, gözlemlenen veriye dayalı olarak ağırlıklandırma ile temel konu yapısını belirlemeyi amaçlar. Dokümanlardaki kelimeler, sistemde gözlemlenen verilerdir.

GDA eğitimsiz öğrenme algoritması olup, önceden tanımlanmış kelimelere ihtiyaç duymamaktadır. Modelde konu sayısı belirleme işleminden sonra, sınıflara göre konulara etiket atanmaktadır. Geliştirilen GDA yönteminde, dokümanlar konulara iki aşamalı olarak atanmaktadır. Öncelikle, metin belgesinde yer alan her bir kelime geçici olarak bir konuya atanmaktadır. Sonrasında, dokümandaki kelimelerin olasılık dağılımlarına göre, doküman belli bir konuya atanmaktadır [14].

GDA her doküman için dokümandaki kelimelere rastgele konu ataması yapar. Her doküman için konu atama işlemi gerçekleştirildikten sonra bu bilgiyi kullanarak çeşitli istatistikler çıkartılır. Yerel istatistik, her dokümandaki konulara kaç adet kelime atandığını gösterirken, global istatistik ise tüm doküman için her kelimenin her konuya kaç kere atandığını göstermektedir. İstatistiksel bilgiler elde edildikten sonra her doküman için her kelimenin yeniden konu ataması gerçekleştirilir. Bunun için mevcut kelime bilgileri güncellenmelidir [15].

$$\frac{n_{ik} + \alpha}{N_i - 1 + K\alpha} \tag{1}$$

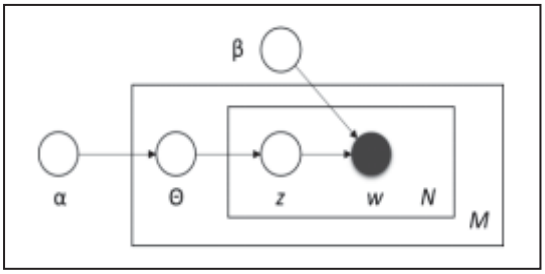
Kelimeler, yeni konulara atanırken ilk önce mevcut dokümanın konular ile hangi oranda ilişkili olduğuna bakılır. Denklem (1)'de n_{ik}, i. haberde k. konuya atanan kelime sayısını göstermektedir. N_i ise dokümanda yer alan toplam kelime sayısıdır. Değerden 1 çıkartılmasının nedeni kullanılan kelimenin yok sayılmasıdır. α değeri; konuların dokümanlardaki dağılımını vermektedir. K değeri de belirlediğimiz konu sayısıdır [15].

$$score(v_i, v_j, \varepsilon) = \log\frac{D(v_i, v_j) + \varepsilon}{D(v_j)} \tag{2}$$

K konu sayısı; konu modelleme ölçütü olan tutarlılık (coherence) değerinin hesaplanmasıyla belirlenmektedir. Tutarlılık değeri; kelimelerin birbirine benzerliğini ölçerek seçilecek konu sayımız hakkında bize bilgi vermektedir. Hesaplanan tutarlılık değerleri arasından en yüksek çıkana ait K değeri, konu sayısı olarak seçilmektedir [15]. Sistemde tutarlılık değeri UMass ölçütü ile hesaplanmaktadır. UMass ölçütü, bu sayıları harici bir derlem yerine konu modellerini eğitmek için kullanılan orijinal derlem üzerinden hesaplar. Denklem (2)'de $D(v_i, v_j)$, v_i ve v_j sözcüklerini içeren dokümanların sayısını, $D(v_j)$ ise v_j içeren dokümanların sayısını saymaktadır. ε ise düzeltme faktörü olan sabit bir değerdir ($\varepsilon = 1$). Dokümandaki tüm v_i ve v_j ikilileri için değerler elde edilir. Değerlerin toplamı tutarlılık değerini vermektedir [16].

$$\frac{n_{word,,k} + \beta}{\sum_{w \in V} n_{w,k} + V\beta} \tag{3}$$

İkinci olarak, mevcut kelimenin konular ile ne kadar ilişkili olduğu hesaplanır. Hesaplamayla kelimenin, verilen konu altında ne kadar kullanıldığının bilgisi çıkarılır. Denklem (3)'de; $n_{word,k}$ geçerli kelimenin k. konuya tüm dokümanda kaç kere atandığını gösterir. β değeri; kelimelerin konulardaki dağılımını vermektedir. V ise veri setindeki tüm kelimelerden oluşturulan sözlüğün boyutudur. Denklem (1) ve (3)'ten elde edilen sonuçlar çarpılarak geçerli kelimenin k. konuya atanma olasılığı hesaplanmaktadır. Tüm doküman sayısı boyunca değerler tekrar hesaplanıp, en yüksek değere ait olan konu kelimenin yeni konusu olarak belirlenir. Veri setindeki tüm dokümanlara ait kelimeler için aynı işlemler uygulanarak yeni konular bulunur ve sistemde belirlenen iterasyon sayısına kadar güncelleme devam eder. Kelimelere konu dağılımı ataması yapıldıktan sonra sistemin modelini çıkarmak için doküman-terim matrisi oluşturulmaktadır. Bu matris ile kelime ağırlıkları hesaplanarak, kelimelerin konulardaki ağırlıkları ortaya çıkarılır [17].



Şekil 1. Gizli Dirichlet Ayırımı süreci [18]

Şekil 1'de GDA süreci grafiksel olarak özetlenmiştir. Rastgele olan değişkenler düğümler ile gösterilmektedir. Düğümler arasındaki olası bağlantılar ise kenarlar kullanılarak temsil edilmiştir. Gösterimde;

- α doküman başına konu dağılımını veren öncelikli Dirichlet parametresidir.
- β konu başına kelime dağılımını veren öncelikli Dirichlet parametresidir.
- Θ belli doküman için konu dağılımıdır.
- z her bir kelime için atanan konulardır.
- w gözlemlenen kelimelerdir.

Şekil 1'de ki yapıdan da anlaşılacağı gibi α ve β parametreleri, sistemin oluşturulması sırasında bir kez örneklenmektedir. Θ ise sistemdeki her bir doküman için örneklenmektedir [18].

B. Veri Seti

Milliyet, Mynet gibi sitelerden faydalanarak Türkçe haber başlıklarından oluşan bir veri seti oluşturulmuştur. Her bir haber başlığı maksimum 30 kelimeden oluşmaktadır. Haber başlığının genişletilmiş olmasının sebebi, bazı haber başlıklarının çok az kelimeden oluşması ve bu durumun da haber türünün belirlenmesini engellemesidir. Bu yüzden html içerisinde tanımlama etiketinde yer alan alt başlık içerisinde yer alan bilgide başlığa dahil edilmiştir. Veri seti; ekonomi, magazin, siyaset, spor, sağlık, teknoloji ve yaşam olmak üzere 7 farklı türde haberden oluşmaktadır. Her haber türüne

ait 600 adet haber başlığı toplanmıştır. Her biri 3, 5 ve 7 sınıf etiketine sahip, 3 farklı veri seti hazırlanmıştır. 3 sınıf için ekonomi, yaşam ve spor; 5 sınıf için ekonomi, yaşam, spor, magazin ve siyaset; 7 sınıf için ise ekonomi, magazin, siyaset, spor, sağlık, teknoloji ve yaşam haber türlerinden oluşmaktadır. Veri setinde üç haber sınıfı için 1800, 5 haber sınıfı için 3000 ve 7 haber sınıfı için de 4200 adet haber başlığı yer almaktadır.

İngilizce veri seti içinde yine haber başlıklarından oluşan Uci-news ve spor sitelerinden oluşturulan ikinci bir veri seti kullanılmıştır [19]. Veri seti; ekonomi, magazin, sağlık, teknoloji ve spor olmak üzere her birinden 1000 adet haber başlığına sahip olan 5 farklı türden oluşmaktadır. Denemeler için veri seti 3 ve 5 sınıf olmak üzere sırası ile 3000 ve 5000 veriden oluşan 2 farklı veri seti oluşturulmuştur. 3 sınıf için magazin, sağlık ve teknoloji; 5 sınıf için ise ekonomi, magazin, sağlık, teknoloji ve spor haber türleri kullanılmıştır. Her iki dil için oluşturulan veri setlerinde veri grubunun %80'i eğitim, %20'si de test için kullanılmıştır.

C. Ön İşleme

Türkçe veri setinde yer alan metinlerde ilk önce noktalama işaretleri temizlenmiştir, sonrasında tüm veri seti küçük harfe dönüştürülmüştür. Türkçe karakterlerde problem yaşandığı için İ, Ö, Ç gibi İngilizcede olmayan harfler kod içerisinde küçük harfe çevrilmiştir. Haberin türünü tespit etmede önemi olmayan etkisiz kelimeler de (stopwords) haber başlıklarının içerisinden çıkarılmıştır. Ayrıca, haberler için anlam taşımayan fiillerden bir liste oluşturulmuş ve bu listedeki kelimeleri içeren başlıklara ayrıştırma işlemi uygulanmıştır.

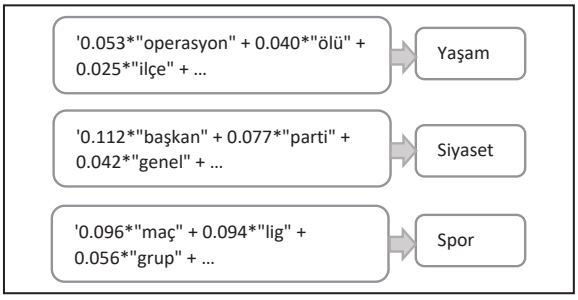
Kelimelerin kökünü bulmak için 3 farklı yöntem kullanılmış ve her biri için VS-Z, VS-S ve VS-5 ismi verilen veri setleri oluşturulmuştur;

- VS-Z: Zemberek kütüphanesi kullanılarak kelimelerin kökleri elde edilmiş ve isim, fiil ve kısaltma içeren kelimelerden oluşan veri seti oluşturulmuştur [20].
- VS-S: Snowball stemmer kütüphanesi kullanılarak kelimelerin kökleri elde edilmiş ve kök uzunluğu 8 karakterden uzun olanlar için ilk 5 harf kök olarak kabul edilmiştir [21]. Diğer kelimelerin aynısı alınarak veri seti oluşturulmuştur.
- VS-5: Veri seti içerisindeki kelimelerin ilk 5 harfi kök kabul edilerek, üçüncü bir veri seti oluşturulmuştur.

İngilizce veri setinde ise noktalama işaretlerinin ve etkisiz kelimelerin çıkarılması için bir ön işleme yapılmıştır. Kök bulma işlemi için de İngilizce kök bulma kütüphanesi olan Porter Stemmer (UCI-P) kullanılmıştır [22].

III. DENEYSEL ÇALIŞMALAR

Haber başlıklarının konularını tespit etme işlemi hem Türkçe hem de İngilizce veri seti için ayrı ayrı gerçekleştirilmiştir.



Şekil 2. Türkçe haber için GDA ile konuların etiketlenme örneği

Şekil 2'de konulara GDA yöntemi sonucunda, kelimelerin ağırlıkları kullanılarak ilgili sınıfın etiketlenme işlemi gösterilmiştir. Kelimelerin ağırlıkları ve kelimelere bakılarak, konuya en uygun sınıf etiketi belirlenmektedir. Tüm konular için sınıf etiketi belirlendikten sonra veri setindeki haberlerin konuya atanması gerçekleşmektedir. Haber, tüm kelimelerin her konudaki ağırlıkları toplanarak en yüksek değere sahip konuya atanmaktadır.

Literatürde kullanılan veri kümeleri erişilebilir olmadığından, geliştirilen model sadece klasik GDA ile karşılaştırılmıştır. Ayrıca, veriler etiketli olsaydı sistem bu etiketleri referans alarak makine öğrenme yöntemleri ile modellenebilirdi. Ancak gerçek dünyada her zaman etiketlenmiş veri olmadığından, böyle bir durum için GDA algoritması kullanılmıştır. Bununla birlikte, GDA denetimsiz bir yöntemdir ve sistemin başarısını arttırmak için n-aşamalı yöntem geliştirilmiştir. Geliştirilen yöntemin sonuçları, makine öğrenme yöntemlerine verilmiştir.

A. Türkçe Veri Seti için Çalışmalar

Gerekli ön işlem adımları uygulandıktan sonra farklı kök bulma yöntemleri kullanılarak oluşturulmuş olan VS-Z, VS-S ve VS-5 veri setinin 3, 5 ve 7 sınıf için tutarlık değerleri hesaplanmıştır. 3, 5 ve 7 sınıflı veri setlerinde konu sayısını bulmak için sınıf sayısı kadar artacak şekilde her birinde 10 tutarlık değeri hesaplanmıştır. Tutarlık değeri en yüksek çıkan değerin konu sayısı, bizim sistemimizi eğitmek için kullanacağımız değer olarak kabul edilmiştir. Örnek olarak VS-Z için sınıflara göre tutarlık değeri ve bulunan konu sayıları Tablo I'de belirtilmiştir.

TABLO I. Sınıf sayısına göre en yüksek değerdeki konu sayısı (VS-Z için)

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

Bulunan konu sayıları üzerinde GDA algoritmasıyla elde edilen sistemde başarı oranı, sınıf sayısı arttıkça beklenildiği gibi azalmaktadır. Sistemin 3, 5 ve 7 sınıf için GDA ile elde edilen başarısı Tablo II'de gösterilmektedir.

TABLO II. Kullanılan kök bulma araçlarına göre GDA'nın başarısı

Sınıf \ Araç (%)	VS-5	VS-Z	VS-S
3	73.5	**81.38**	67.78
5	65.3	**69**	67
7	**53**	52.3	48

Haber başlığı içerisinde yer alan farklı kelime sayısını azaltmak amacıyla GDA algoritması için iki aşamalı bir yöntem önerilmiştir. Uygulanan işlem; bir konu için atanmış kelimelerin toplam ağırlık değerinin, toplam kelime sayısına oranlanması sonucu ortaya çıkan değeri eşik seviyesi olarak kabul etmek ve bu değerden yüksek olan kelimeleri kullanmaktır. Bu işlem sonucunda sözlükteki toplam kelime sayısı yaklaşık olarak 1/4 oranına düşmektedir. Yeni elde

edilen kelime sözlüğü ile tüm sınıflar için tutarlık değeri ve konu sayısı tekrardan hesaplanmıştır. Sistem iki aşamalı GDA (2-GDA) ile modellendiğinde alınan başarı sonuçları Tablo III'te gösterilmektedir.

TABLO III. Geliştirilen 2-GDA yönteminin sistemdeki başarısı

Sınıf \ Araç (%)	VS-5	VS-Z	VS-S
3	**91.66**	90.28	81.38
5	74.83	**76.5**	70.33
7	57.61	**57.62**	50.35

Tüm veri seti için 2-aşamalı GDA'dan yararlanarak her haberin tüm konularına ait ağırlık değerlerinden oluşan ve sınıf etiketi de olan 'arff' uzantılı bir dosya elde edilmiştir. Tüm konular için ağırlık değeri, veri setindeki her habere ait kelimelerin 2-GDA modelindeki konularda bulunan kelime-ağırlıklarına bakarak hesaplanmıştır. Haberdeki kelimelerin tek tek konulara ait ağırlıkları hesaplandıktan sonra cümlenin konulara ait toplam ağırlık değerleri elde edilmiştir. Dosya oluşturulurken 2-GDA'da genelde en iyi sonucu veren VS-Z veri seti kullanılmıştır. Bu dosyanın, Weka içerisinde yer alan Multinomial Naive Bayes (MNB), Random Forest (RF), Support Vector Machines (SVM) ve Multilayer Perceptron (MP) makine öğrenmesi yöntemleri ile 10-katlı çapraz doğrulama kullanılarak 3, 5 ve 7 sınıf için başarıları ölçülmüştür. 2-GDA ile özel oluşturulan arff uzantılı dosyanın sınıflandırma algoritmalarındaki başarısı Tablo IV'te gösterilmiştir.

TABLO IV. 2-GDA ile oluşturulan dosyanın sınıflandırmalardaki başarısı

Sınıf \ Araç (%)	3	5	7
MNB	89.33	80.3	67.59
RF	**95.22**	**88.06**	**83.19**
SVM	88.77	79.96	70.5
MP	92.61	82.73	75.33

B. İngilizce Veri Seti için Çalışmalar

Veri setine ön işlem adımları uygulandıktan sonra 3 ve 5 haber sınıfı için tutarlık değerleri hesaplanmış ve çıkan en büyük değere ait konu sayısı GDA modelimiz için seçilmiştir. Seçilen konu sayısına göre sistemin başarısı Tablo V'te gösterilmiştir.

TABLO V. Bulunan konu sayısına göre GDA'nın başarısı

Sınıf	Tutarlık Değeri	Konu Sayısı	Başarı (%)
3	0.4868	15	**67.16**
5	0.462	15	**59.4**

Türkçe veri setinde olduğu gibi İngilizce veri setinde de sistem için geliştirilen 2-GDA algoritması uygulanmıştır. Elde edilen yeni tutarlık değerleri ve konu sayılarına göre sistem yeniden modellenerek sistemin başarısı ölçülmüştür. Sistemin başarısı Tablo VI'da gösterilmiştir.

TABLO VI. GDA ve geliştirilen 2-GDA'nın karşılaştırılması

Sınıf	GDA (%)	2-GDA (%)
3	67.16	**71.83**
5	59.4	**63.7**

Türkçe veri setinde olduğu gibi 2-aşamalı GDA yöntemi yardımıyla elde edilen konulara ait kelime–ağırlık değerlerinden, her haberin tüm konularına ait ağırlık değerlerinden oluşan ve sınıf etiketi de içeren 'arff' uzantılı bir dosya elde edilmiştir. Weka içerisinde yer alan Multinomial Naive Bayes (MNB), Random Forest (RF), Support Vector Machines (SVM) ve Multilayer Perceptron (MP) makine öğrenmesi yöntemleri ile sistemin başarısı 3 ve 5 sınıf için ölçülmüştür. Yöntemlerde 10-katlı çapraz doğrulama kullanılmıştır. 2-GDA ile özel oluşturulan dosyanın sınıflandırma algoritmalarındaki başarısı Tablo VII'de gösterilmiştir.

TABLO VII. -GDA ile oluşturulan dosyanın sınıflandırmadaki başarısı

Sınıf \ Araç (%)	3	5
MNB	75.8	75.48
RF	**92.63**	**92.74**
SVM	82.16	80.04
MP	86.8	83.68

IV. SONUÇ

Çalışmada haber başlığından yola çıkarak haberlerin hangi türe ait olduğu konu modelleme algoritması olan GDA ile tespit edilmiştir. İki aşamalı olarak geliştirilen GDA ile klasik GDA yöntemi arasında %4 ile %18 arası bir başarı artışı gözlemlenmiştir. Bunun en önemli nedeni tüm belgede kullanılan kelime sayısının, az ağırlığa sahip kelimelerin silinmesinden dolayı azalmasıdır. Ayrıca, oluşturulan arff uzantılı dosyanın Weka'daki makine öğrenme yöntemlerinde, her sınıf için Türkçe ve İngilizcede en başarılı yöntem Random Forest olmuştur. Türkçe veri seti için en yüksek başarı üç sınıf için %95.2, beş sınıf için %88 ve yedi sınıf için ise %83.2 olarak alınmıştır. İngilizce veri setinde ise üç sınıf için %92.63 ve beş sınıf için %92.74 başarı sağlanmıştır. Sonuç olarak, konu modelleme yöntemi GDA ile geliştirilen yöntemin sisteme olumlu yönde katkısı olduğu gözlenmiştir.

Gelecek konu modelleme çalışmalarımızda 2-GDA algoritmasını; müzik türünü, sosyal medyada yazılan mesajların içerdiği duyguyu, yazılan metinin hangi yazar tarafından yazıldığını, soru cevap sistemlerinde doğru cevabı tespit etmek için kullanacağız. Geliştirilen yöntem ile kelimelere daha doğru bir ağırlıklandırma yapıldığından sisteme olumlu yönde bir katkı sağlayacağını düşünmekteyiz.

KAYNAKÇA

[1] D. M. Blei, "Probabilistic topic models", Communications of the ACM, 55(4), 77-84, 2012.

[2] A. Daud, J. Li, L. Zhou ve F. Muhammad, "Knowledge discovery through directed probabilistic topic models: a survey", Frontiers of Computer Science in Chine, 4(2), 280-301, 2010.

[3] M. Steyvers ve T. Griffiths, "Probabilistic topic models", Laurence Erlbaum, 1 Eylül 2007, New Jersey, 2007.

[4] A. Çelikyılmaz, G. Tur ve D. Tur, “LDA Based Similarity Modeling for Question Answering”, Proceedings of the NAACL HLT 2010 Workshop on Semantic Search, Haziran 2010, Los Angeles, 2010.

[5] L. Li ve Y. Zhang, “An empirical study of text classification using Latent Dirichlet Allocation”, 2009.

[6] A. Çelikyılmaz, G. Tur ve D. Tur, “Latent semantic modeling for slot filling in conversational understanding”, IEEE International Conference on Acoustics, Speech and Signal Processing, 26-31 Mayıs 2013, Vancouver, 2013.

[7] I. Titov ve R. McDonald, “Modeling Online Reviews with Multi-grain Topic Models”, In Proceedings of International Conference on World Wide Web, 21-25 Nisan 2008, Beijing, 2008.

[8] C. Lin ve Y. He, “Joint sentiment/topic model for sentiment analysis”, In Proceedings of ACM International Conference on Information and Knowledge Management, 2-6 Kasım 2009, Hong Kong, 2009.

[9] A. J. T. Lee, F-C. Yang, C-H. Chen, C-S. Wang ve C-Y. Sun, “Mining perceptual maps from consumer reviews”, Decision Suppoprt Systems, 82, 12-25, 2016.

[10] R. Chatterjee ve S. Agarwal, “Twitter Truths: Authenticating Analysis of Information Credibility”, In 2016 International Conference on Computing for Sustainable Global Development, 16-18 Mart 2016, New Delhi, 2016.

[11] S. Poria, I. Chaturvedi, E. Cambria ve F. Bisio, “Sentic LDA: Improving on LDA with Semantic Similarity for Aspect-Based Sentiment Analysis”, In 2016 International Conference on Neural Networks, 24-29 Temmuz 2016, Vancouver, 2016.

[12] S. Feuerriegel, A. Ratku ve D. Neumann, “Analysis of How Underlying Topics in Financial News Affect Stock Prices Using Latent Diirichlet Allocation”, 49th Hawaii International Conference on System Sciences, 5-8 Ocak 2016, Koloa, 2016.

[13] D. M. Blei, A. Y. Ng ve M. I. Jordan, “Latent Dirichlet Allocation”, Journal of Machine Learning Research, 3, 993-1022, 2003.

[14] A. Onan, “Türkçe Twitter Mesajlarında Gizli Dirichlet Tahsisine Dayalı Duygu Analizi”, Akademik Bilişim, 8-10 Şubat 2017, Aksaray, 2017.

[15] Z. A. Guven, B. Diri ve T. Cakaloglu, “Classification of TurkishTweet emotions by n- stage Latent Dirichlet Allocation”, Electric Electronics, Computer Science, Biomedical Engineerings' Meeting (EBBT), 18-19 Nisan 2018, İstanbul, 2018.

[16] 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, 12-14 Temmuz 2012, Jeju Adası, 2012.

[17] https://rstudio-pubs-static.s3.amazonaws.com/79360_850b2a69980c4488b1db95987a24867a.html (2017).

[18] http://www.wikizero.net/index.php?q=aHR0cHM6Ly9lbi53aWtpcGVkaWEub3JnL3dpa2kvTGF0ZW50X0RpcmljaGxldF9hbGxvY2F0aW9u (2017).

[19] https://www.kaggle.com/uciml/news-aggregator-dataset/data (2017).

[20] http://www.java2s.com/Code/Jar/z/Downloadzemberektum20jar.htm (2017).

[21] http://snowball.tartarus.org/download.html (2017).

[22] http://snowball.tartarus.org/algorithms/porter/stemmer.html (2017).