Sinop Üniversitesi Fen Bilimleri Dergisi *Sinop Uni J Nat Sci 3 (2): 66-78 (2018)*
ISSN: 2536-4383

Araştırma Makalesi

Olasılıksal Yöntemler ile Türkçe Metinlerin Anlamsal Benzerliğinin Belirlenmesi

Engin Yıldıztepe*, Volkan Uzun
Dokuz Eylül Üniversitesi Fen Fakültesi İstatistik Bölümü, İzmir.

Öz

Metin madenciliğinde, yapısal olmayan metin verilerinden matematiksel ve istatistiksel yöntemler ile anlamlı bilgiler çıkartmak amaçlanır. Metin sınıflandırma, kümeleme, duygu çözümleme, özetleme, anlamsal benzerlik bulma ve yazar tanıma, başlıca metin madenciliği çalışma alanlarıdır. Bu çalışmanın konusu olan anlamsal benzerlik analizi, metinler arasındaki anlamsal yakınlığı belirlemeye çalışır. Olasılıksal gizli anlam analizi ve gizli Dirichlet ataması, metinler arasındaki anlamsal benzerliğin belirlenmesinde kullanılan olasılıksal yöntemlerdir. Bu çalışmada, olasılıksal gizli anlam analizi ve gizli Dirichlet ataması yöntemleri detaylı olarak incelenmiştir. Ayrıca, farklı haber ajanslarından seçilen Türkçe haber metinlerini anlamsal benzerliklerine göre kümelemek için yapılan bir uygulama tartışılmıştır. Uygulamada R istatistiksel programlama dili ve Matlab kullanılmıştır.
Anahtar kelimeler: Anlamsal benzerlik; Olasılıksal gizli anlam analizi; Gizli Dirichlet ataması; Metin madenciliği

Determination of the Semantic Similarity of Turkish Texts Using Probabilistic Methods

Abstract

Text mining is the process of deriving useful information from unstructured text data. During this process, text mining uses statistical and mathematical methods. Major text mining tasks include text categorization, text clustering, concept extraction, document summarization, semantic similarity and author identification. In this study, semantic similarity issues have been examined. Semantic similarity analysis aims to determine semantic similarity between texts. Probabilistic latent semantic analysis and latent Dirichlet allocation are probabilistic methods to determine semantic similarity between texts. In this study, semantic analysis methods using probabilistic latent semantic analysis and latent Dirichlet allocation are examined in detail. In addition, an application for clustering Turkish news texts chosen from different news agencies according to semantic similarities is discussed. R statistical programming language and Matlab are used in the application.
Keywords: Semantic similarity; Probabilistic latent semantic analysis; Latent Dirichlet Allocation, Text mining

Giriş

Web dünyasının genişlemesiyle ve farklı alanlarda kullanılmasıyla birlikte erişilebilen yapısal olmayan veri miktarı da artmıştır. Bilgi geri getirimi ve metin madenciliği gelişen web teknolojileri

* Sorumlu Yazar
e-mail: engin.yildiztepe@deu.edu.tr
Received: 10.11.2017
Accepted: 07.12.2018

sayesinde öne çıkan çalışma alanları haline gelmiştir. Elektronik dokümanların, kullanıcı geri bildirimlerinin ve Twitter gibi sosyal medya ortamlarının sağladığı veriler metin madenciliğine yeni uygulama alanları katmıştır. Metin madenciliğinde, yapısal olmayan metinlerden matematiksel ve istatistiksel yöntemler ile anlamlı bilgiler çıkartmak amaçlanır. Otomatik özet çıkarma, yazar tanıma, konu modelleme, sınıflandırma ve kümeleme metin madenciliğinin başlıca çalışma alanlarıdır [1]. Metin kümelemede, metni meydana getiren dokümanların konu benzerliklerine göre kümelere ayrılması amaçlanır. Bu amaç için metinler arası benzerliği belirleyen yöntemler kullanılır [1].

Gizli anlam analizi-GAA (latent semantic analysis-LSA), olasılıksal gizli anlam analizi-OGAA (probabilistic latent semantic analysis-PLSA) ve gizli Dirichlet ataması-GDA (latent Dirichlet allocation-LDA) gibi yöntemler, metinlerin konularına göre kümelenmesinde ve benzerliklerinin belirlenmesinde kullanılabilecek yöntemlerin başında gelir.

OGAA, 2000'li yılların başında metinlerin anlamsal benzerliklerinin belirlenebilmesi için önerilmiş bir yöntemdir [2]. Metinlerin ve kelimelerin hangi olasılıklarla hangi konulara ait olduğunu belirlemeye çalışır. 1990 yılında geliştirilen GAA yönteminin [3] olasılıksal bir yorumu olarak ifade edilebilir. GAA yönteminde metinleri temsil eden matris, alt matrislere ayrıştırılır ve ayrıştırılan bu alt matrisler boyut indirgemesi yapıldıktan sonra daha düşük boyutlu yeni matrisin hesaplanmasında kullanılır. Elde edilen matris üzerinde yapılan hesaplamalar (korelasyon katsayısı vb.) ile metinlerin arasındaki benzerlik yorumlanabilir.

GDA, 2003 yılında Blei ve ark. tarafından önerilmiştir [4]. GDA'daki temel fikir, dokümanların, kelimeler ile belirlenen gizli konuların rastgele bir karışımı olarak temsil edilmesidir. GDA, Dirichlet dağılımını kullanarak metinlerin ve kelimelerin konulara ait olma olasılıklarını belirlemeye çalışır. İlgili literatürdeki uygulamalar incelendiğinde, özellikle son beş yılda, GDA temelli yaklaşımların yaygın olarak kullanıldığı görülmektedir.

OGAA ve GDA, metin madenciliği ile ilgili literatürde, konu modelleme ve metin kümelemede kullanılan olasılıksal yöntemler olarak yer almaktadır [5]. Konu modelleme yöntemlerinin gelişimini ve özelliklerini konu eden güncel bir çalışma için [6] incelenebilir.

GDA yöntemi sadece metin kümeleme değil, duygu analizi ve metin

sınıflama uygulamalarında da kullanılmaktadır. Onan ve ark. tarafından yapılan bir çalışmada temel sınıflama algoritmalarının, metin temsilinde GDA yöntemi kullanıldığı durumdaki performansları incelenmiştir [7]. Song ve ark. tarafından yayımlanan çalışmada, olasılıksal yöntemlerin kısa metinlerin (twitter mesajları, kısa mesaj, başlık vb.) sınıflandırılması için kullanımı incelenmiştir [8]. Kütüphane ve bilgi bilimi alanındaki bir başka çalışmada ise, bilimsel araştırmalardaki popüler konuları tespit etmek için GDA'dan yararlanılmıştır [9].

Son yıllarda popüler bir metin sınıflama çalışma alanı haline gelen duygu analizinde de GDA kullanıldığı görülmektedir. Ekince ve Omurca tarafından Türkçe otel yorumları kullanılarak yapılan duygu analizi çalışmasında konu modellemede GDA kullanılmıştır [10]. Onan tarafından yapılan bir başka duygu analizi çalışmasında, Türkçe Twitter mesajlarının temsilinde GDA kullanılmış ve beş farklı sınıflama algoritması karşılaştırılmıştır. [11].

Bu çalışmada OGAA ve GDA yöntemleri hakkında detaylı bilgi vermek ve Türkçe metinler üzerinde anlamca yakın konuları belirlemek amacıyla yapılan uygulamayı paylaşmak amaçlanmıştır. Takip eden bölümlerde OGAA ve GDA hakkında detaylı bilgi verilmiş, yöntemlerin sınırlılıklarına değinilmiştir. Uygulama bölümünde, OGAA ve GDA'nın farklı haber ajanslarından derlenen Türkçe haber metinleri üzerinde yapılan uygulamalarına yer verilmiştir. Son bölümde, elde edilen sonuçlar ve öneriler sunulmuştur.

Olasılıksal Gizli Anlam Analizi

OGAA yönteminde metinlere, konulara ve kelimelere dair rasgele önsel (prior) olasılıklar belirlenir. Bu önsel olasılıklar yardımıyla bir ara değer hesaplanır ve bu değer en büyüklenerek ilgilenilen olasılıklar güncellenir. Bu aşama en büyükleme işlemi öncesi ve sonrası olasılıklar arasındaki fark arzu edilen değere ulaşıncaya kadar veya belirlenen iterasyon tamamlanıncaya dek devam ettirilir. Sonunda elde edilen olasılıklara göre metinlerin arasındaki benzerlik belirlenebilir.

OGAA adımları

OGAA yöntemini kullanabilmek için, metinlerin (dokümanların) matris halinde temsil edilmesi ve konu sayısının bilinmesi gerekmektedir. OGAA'da kullanılan olasılıkların tanımları aşağıdaki gibidir:

$P(d_i|z_j)$: *i.* metnin *j.* konuya ait olma olasılığı.

$P(z_j)$: *j.* konunun (gizli sınıfın) olasılığı

$P(w_t|z_j)$: *t.* kelimenin *j.* konuya ait olma olasılığı *i:1,...,d j: 1,...,k t: 1,...,w*

Burada, d metin, k konu ve w kelime sayısını belirtir. Analize başlamadan önce, metin, gizli sınıf ve kelime olasılıkları belirlenir. P(z) ve P(w|z) için değerler başlangıçta rasgele belirlenir. P(d|z) için ise metin sayısına göre eşit olasılıklar atanır.

OGAA, Beklenti Ençoklama-BE (Expectation Maximization-EM) algoritmasını kullanır. BE algoritması gözlenen verilere göre oluşturulan olabilirlik fonksiyonundan yola çıkar ve bu olabilirlik fonksiyonlarını yineleme yoluyla en büyükleyerek parametreleri tahminler [12]. BE algoritması iki aşamadan oluşur; beklenti (B) adımı ve ençoklama (E) adımı. B adımında parametrelerin geçerli kestirim değerleri için olasılıklar hesaplanır. E adımında ise parametre kestirimleri B adımında hesaplanan olasılıklara göre güncellenir. Bu işlem B ve E adımlarında belirlenen P(z) olasılıkları arasındaki fark istenilen değere düşünceye ya da belirlenen iterasyon sayısına ulaşıncaya kadar tekrarlanır.

OGAA için beklenti adımında her bir gizli sınıf için ayrı ayrı hesaplama yapılır. İşlem sonucunda her gizli sınıfın olasılıkları $P(z|d,w)$ oluşur. $P(z|d,w)$ olasılıkları ençoklama adımında kullanılır.

Beklenti adımı;

$$P(z|d,w) = \frac{P(z)\times P(d|z)\times P(w|z)}{\sum_i P(z_i)\times P(d|z_i)\times P(w|z_i)} \quad (1)$$

Ençoklama adımında 3 ayrı olasılık hesaplanır. Bunlar;

$$P(w|z) = \frac{\sum_j n(d_j,w)\times P\left(z|d_j,w\right)}{\sum_j \sum_i n(d_j,w_i)\times P\left(z|d_j,w_i\right)} \quad (2)$$

$$P(d|z) = \frac{\sum_i n(d,w_i)\times P(z|d,w_i)}{\sum_j \sum_i n(d_j,w_i)\times P\left(z|d_j,w_i\right)} \quad (3)$$

$$P(z) = \frac{\sum_j \sum_i n(d_j,w_i)\times P\left(z|d_j,w_i\right)}{\sum_k \sum_j \sum_i n(d_j,w_i)\times P\left(z_k|d_j,w_i\right)} \quad (4)$$

Ençoklama adımında, beklenti adımında hesaplanan $P(z|d,w)$ olasılıkları kullanılarak, $P(z)$, $P(w|z)$ ve $P(d|z)$ olasılıkları yeniden hesaplanır. Yeni hesaplanan $P(z)$, $P(w|z)$ ve $P(d|z)$ olasılıkları beklenti adımı öncesindeki değerleri ile karşılaştırılır. Aradaki fark ε değerinden büyükse, BE adımları, ε değerine ulaşana ya da belirlenen iterasyon sayısı tamamlanıncaya kadar tekrarlanır [2].

ISSN: 2536-4383

OGAA yöntemindeki problem, parametre sayısının *(kw+kd)* doküman sayısı ile doğru orantılı olarak artmasıdır. OGAA, parametre sayısının artması durumunda aşırı uyum gösterebilir [4]. Daha gelişmiş bir model olan GDA'da parametre sayısı doküman sayısı ile artmaz.

Gizli Dirichlet Ataması

Dirichlet dağılımı, Beta dağılımının çok değişkenli genelleştirilmiş halidir. GDA, dokümanların sınıflandırılmasında, özet çıkarmada ve benzerliklerinin hesaplanmasında kullanılabilir. GDA yöntemi uygulanırken üç durum göz önünde bulundurulur:

- Eğer bir metnin konusu biliniyorsa, o konuya ait diğer metinlerde de benzer kelimelere rastlanır.
- Eğer metinlerin konuları hakkında hiçbir bilgi yoksa metinlerin konulara ait olma olasılıkları eşittir.
- Eğer kelimelere hangi konuda daha fazla rastlanacağı hakkında bilgi yoksa kelimelere konularda eşit olasılıkla rastlanır.

GDA adımları

C, k farklı "konudan" (topic) bahseden, *V* sayıda benzersiz kelimeden ve *M* sayıda "dokümandan" (document) oluşan bir "derlem" (corpus) olsun. Buna göre;

C : Derlem $C=(D_1,D_2,D_3,\ldots,D_M)$

D : Döküman $D=(w_1,w_2,w_3,\ldots,w_{Nd})$

w_i : dokümandaki *i*. kelime, $i=1,\ldots, N_d$

N_d : *d*. dökümandaki kelime sayısı, $d=1,\ldots,M$

olarak tanımlanır.

GDA, derlemdeki her bir doküman için aşağıdaki varsayımlarda bulunur.

1. $N \sim Poisson\ (\xi)$
2. $\theta \sim Dirichlet(\boldsymbol{\alpha})$, burada $\boldsymbol{\alpha}$ Dirichlet dağılımının (konu dağılımının) *k* boyutlu parametre vektörüdür.
3. her bir w_i için
   a. $z_i \sim Multinomial(\theta)$
   b. z_i. konudan w_i kelimesini seçme olasılığı $p(w_i|z_i,\beta)$ dir.

θ parametresi *k* boyutlu Dirichlet rassal değişkendir. θ_d, *d.* doküman için konu karışımını belirler. θ_d parametresi $\boldsymbol{\alpha_i}$ 'ye göre belirlenir. *z* değişkeni konuyu temsil eder. z_{di}, *d.* dokümandaki *i*. kelime için konuyu gösterir ve θ_d parametresi ile Multinomial dağıldığı varsayılır. β, $k \times V$ boyutunda bir matristir ve sözlükteki bir kelimenin verilen bir konuya ait olma olasılığını belirler. *k* boyutlu Dirichlet rassal değişken θ'nın olasılık yoğunluk fonksiyonu aşağıda verilmiştir.

$$p(\theta|\alpha) = \frac{\Gamma(\sum_{i=1}^{k}\alpha_i)}{\prod_{i=1}^{k}\Gamma(\alpha_i)}\theta_1^{\alpha_1-1}\dots\theta_k^{\alpha_k-1} \quad (5)$$

Verilen α ve β parametrelerine göre, θ'nın bileşik dağılımı aşağıda verilmiştir. Bu parametreler kullanıcı tarafından belirlenir. α ve β, genellikle tüm konular için aynı, 1 veya daha küçük bir sayı olarak belirlenir. Dokümanlardaki konu sayısı azaldıkça α daha küçük seçilir. Bir konudaki kelime sayısı azaldıkça β daha küçük seçilir (örneğin konuda sadece birkaç kelime varsa 0.001 seçilebilir).

$$p(\theta, \mathbf{z}, D|\alpha, \beta) = p(\theta|\alpha)\prod_{i=1}^{N} p(z_i|\theta)p(w_i|z_i, \beta) \quad (6)$$

Bir dokümanın marjinal dağılımını belirlemek için aşağıdaki işlem yapılır.

$$p(D|\alpha, \beta) = \int p(\theta|\alpha)\left(\prod_{i=1}^{N}\sum_{z_i} p(z_i|\theta)p(w_i|z_i, \beta)\right) d\theta \quad (7)$$

Her dokümanın marjinal dağılımlarını çarparak, derlem için olasılık modeli belirlenir [4]:

$$p(C|\alpha, \beta) = \prod_{d=1}^{M} p(D_d|\alpha, \beta) = \prod_{d=1}^{M}\int p(\theta_d|\alpha)\left(\prod_{i=1}^{N_d}\sum_{z_{dn}} p(z_{di}|\theta_d)p(w_{di}|z_{di}, \beta)\right) d\theta_d \quad (8)$$

Verilen bir dokümanın *z* konusuna ait olma olasılığı;

$$p(\theta, \mathbf{z}|D, \alpha, \beta) = \frac{p(\theta, \mathbf{z}, D|\alpha, \beta)}{p(D|\alpha, \beta)} \quad (9)$$

Bu olasılık aşağıdaki işlemler ile yaklaşık olarak belirlenir;

1. Başlangıçta dokümanlardaki her bir kelimenin ait olduğu konu (z_{di}) rasgele belirlenir.
2. *d. d*okümandaki *i.* kelimenin *t.* konuya ait olma olasılığını $p(\mathbf{z_{di}} = \mathbf{t}|D, \alpha, \beta)$ bulmak için, tüm diğer kelime-konu atamalarının doğru olduğu varsayılarak her bir kelime ve *t=1,…,k* için, $p\left(\mathbf{z_{di}} = \mathbf{t}\middle|\mathbf{z_{-(di)}}, D, \alpha, \beta\right)$ olasılığı bulunur. z_{di} değeri, en yüksek olasılığın elde edildiği *t* değeri ile güncellenir.
3. 2.adım konu belirlemeleri değişmeyinceye veya istenilen tekrar sayısına ulaşılıncaya kadar tekrar edilir [4].

GDA yöntemi, konu belirleme ve doküman kümeleme uygulamalarında başarılı sonuçlar vermektedir. Ancak bu yöntemin de, varsayımlarından kaynaklanan, iki önemli kısıtı bulunmaktadır. İlki, konu sayısının sabit olması, bir başka deyişle önceden bilinmesi gerekliliğidir. Bu kısıt, en uygun konu sayısını belirlemek için geliştirilen algoritmalar sayesinde aşılabilmektedir. 2009 yılında Cao ve ark. tarafından

önerilen yinelemeli bir algoritma çapraşıklık (perplexity) kriterini kullanarak en uygun konu sayısını tahmin edebilmektedir [13].

İkinci kısıt dokümanı oluşturan konuların bağımsızlığı ile ilgilidir. Derlemdeki bir doküman birden çok konudan oluşabileceği gibi aynı konu birden çok dokümanda da geçebilir. Ancak GDA yaklaşımında, dokümanlardaki konu oranları Dirichlet dağılan rassal değişkenlerdir ve bağımsız olmaları gerekir [14]. Uygulamada dokümanı oluşturan konuların ilişki düzeyini belirlemek oldukça güçtür [15]. Bu sınırlılıkların giderilmesi amacıyla Teh ve ark. tarafından 2006 yılında daha gelişmiş bir yaklaşım olan Hiyerarşik Dirichlet Süreçleri-HDS (Hierarchical Dirichlet Processes) önerilmiştir [16]. HDS yaklaşımında konu sayısının önceden bilinmesine gerek yoktur ve dokümanlar ortak konulara sahip olabilir.

Kelime frekansları ile çalışan bu yöntemlere getirilen bir başka eleştiri, kelimelerin birbiriyle olan anlamsal ilişkisini dikkate almamalarıdır. 2018 yılında yayımlanan bir çalışmada, bu sorunu gidermek için bilgi-tabanlı yeni bir yaklaşım (knowledge-based hierarchical topic model - KHTM) önerilmiştir [17].

Uygulama

Bu bölümde, OGAA ve GDA yöntemlerinin kullanıldığı bir uygulamaya yer verilmiştir. Uygulamalarda kullanılan veriler haber ajanslarının web sitelerinde yayımlanan haber metinlerinden derlenmiştir. Dört ayrı haber ajansından (HA), üç ayrı konuda toplam on iki adet haber metni kullanılmıştır. Uygulamada kullanılan haber metinleri çalışmanın ekinde verilmiştir. İstenilen sonuçları gösterebilmek amacıyla, kullanılan konular birbirinden oldukça farklı (aynı kelimeleri içermeyen) seçilmiştir.

Uygulamada kullanılan haber metinlerindeki tek başına bir anlam ifade etmeyen kelimeler (edat, bağlaç vs.) metinlerden ayıklanmıştır. Daha sonra gövdeleme (ek-kök ayrıştırması) işlemi yapılmıştır. Tüm kelimeler eklerden arındırılmış sadece kökler analize dâhil edilmiştir. Bu işlemin amacı eklerle farklılaşmış ama aynı kök ve aynı anlama sahip kelimelerin belirlenmesi ve farklı kelimeler gibi anlaşılmasının önüne geçmektir. Gövdeleme işlemi için Zemberek-NLP kütüphanesinden faydalanılmıştır [18]. Gövdeleme ve anlamsız kelimelerin ayrıştırılması işleminden geçen kelimelerin yer aldıkları haber metinlerine göre frekanslarını gösteren doküman-terim frekans matrisi

oluşturulmuştur. Matris oluşturulurken frekansı iki ve daha fazla olan kelimeler dikkate alınmıştır. Bu işlemdeki amaç matrisi küçültmek ve daha verimli sonuçlar elde etmektir. OGAA ve GDA uygulamalarında elde edilen bu terim frekans matrisi kullanılmıştır. Boyutları nedeniyle, çalışmada kullanılan terim frekans matrisine çalışmada yer verilmemiştir.

OGAA uygulaması

OGAA uygulaması MATLAB ile gerçekleştirilmiştir [19]. Analiz, 136 iterasyon sonucunda istenilen ε değerine (1e-10) ulaşmıştır. OGAA sonuçları Tablo 1'de verilmiştir.

Tablo 1'de, aynı konudaki metinlerin olasılıkları (toplamları 1'e çok yakın) görülmektedir. Bu tabloda sıfıra çok yakın çıkan değerler sıfır olarak gösterilmiştir. Aynı konuda olmalarına rağmen farklı kelimeler de içeren metinler için elde edilen olasılık değerleri farklılık göstermektedir.

Tablo 1. OGAA sonuçları.

Haber Kaynağı	Konu-1	Konu-2	Konu-3
HA-1_1	0.235	0	0
HA-2_1	0.260	0	0
HA-3_1	0.226	0	0
HA-4_1	0.277	0	0
HA-1_2	0	0.269	0
HA-2_2	0	0.256	0
HA-3_2	0	0.243	0
HA-4_2	0	0.230	0
HA-1_3	0	0	0.276
HA-2_3	0	0	0.180
HA-3_3	0	0	0.234
HA-4_3	0	0	0.308

GDA uygulaması

GDA yöntemi derlemdeki, frekansı üç ve daha fazla olan, 53 farklı kelime, 3 farklı konu, 12 doküman için uygulanmıştır. Dokümanlardaki kelime sayısını belirten N vektörü;
N={21, 25, 20, 24, 10, 13, 11, 14, 24, 16, 20, 26} olarak bulunmuştur [19].

GDA yöntemi R programlama dilindeki "topicmodels" paketindeki [20] fonksiyonlar kullanılarak uygulanmış ve Tablo 2'teki sonuçlara ulaşılmıştır.

Tablo 2. GDA sonuçları.

Haber Kaynağı	Konu-1	Konu-2	Konu-3
HA-1_1	**0.9981**	0.0009	0.0009
HA-2_1	**0.9984**	0.0008	0.0008
HA-3_1	**0.9980**	0.0010	0.0010
HA-4_1	**0.9984**	0.0008	0.0008
HA-1_2	0.0020	**0.9961**	0.0020
HA-2_2	0.0015	**0.9970**	0.0015
HA-3_2	0.0018	**0.9964**	0.0018
HA-4_2	0.0014	**0.9972**	0.0014
HA-1_3	0.0008	0.0008	**0.9984**
HA-2_3	0.0661	0.0012	**0.9327**
HA-3_3	0.0010	0.0010	**0.9980**
HA-4_3	0.0008	0.0008	**0.9985**

Tablo 2'deki sonuçlara göre GDA yöntemi haber metinlerini başarıyla üç konuya ayırmıştır. GDA sonuçlarında haber metinlerinin konulara ait olma

olasılıkları görülmektedir. Yüksek olasılık değerleri metnin ilgili olduğu konuyu göstermektedir. Bu sonuçlar ile metinlerin anlamsal benzerliği de yorumlanabilir. Aynı konu için yüksek olasılık değeri elde edilen metinlerin anlamsal olarak da birbirine yakın olduğu söylenebilir.

Sonuç

Bu çalışmada, anlamsal benzerlik için kullanılabilen OGAA ve GDA yöntemleri anlatılmış ve bu yöntemlerinin Türkçe haber metinleri üzerindeki bir uygulamasına yer verilmiştir. OGAA yönteminde, haber metinlerinin ait oldukları konular için 0.180 ile 0.308 arasında değişen olasılık değerleri hesaplanırken diğer konular için olasılıklar sıfıra çok yakın bulunmuştur. GDA yönteminde ise haber metinlerinin ait oldukları konular için bire yakın olasılık değerleri elde edilmiştir. Elde edilen sonuçlara göre iki yöntemle de aynı konudan bahseden haber metinleri başarıyla diğer metinlerden ayrılmıştır. Böylece anlamsal olarak yakın metinlerin belirlenmesi sağlanmıştır.

Yapısal olmayan verilerin benzerliklerinin belirlenmesi, anlamca yakın olan metinlerin kümelenmesi güncel bir çalışma alanıdır. Ancak Türkçe dili için bu konuda yeterince örnek çalışma olmadığı görülmüştür. Bu çalışmadaki örneklerin bu konuda katkı sağlayacağı düşünülmektedir. Konuyla ilgilenen araştırmacılar için çalışmada kullanılan haber metinleri ekte verilmiştir, talep edilmesi durumunda uygulamada kullanılan kodlar paylaşılabilir.

Anlamsal olarak benzer metinlerin aranması gelişen Web teknolojileriyle birlikte önemi daha da artan bir yaklaşımdır. OGAA, GDA ve HDS gibi yöntemler dokümanların anlamsal yakınlığının araştırılmasında kullanılabilir. Ancak, bu yöntemlerin verimli çalışabilmesi için gövdeleme ve anlamsız kelimelerin ayrıştırılması işlemlerinin doğru yapılması gerekmektedir. Gövdeleme konusunda İngilizce için Porter tarafından geliştirilen gövdeleme algoritması yaygın bir kullanıma sahiptir ve bir standart haline gelmiştir [21]. Türkçe için morfolojik gövdeleme, doğal dil işleme araştırmacıları için aktif bir çalışma konusudur. Bu konuda geliştirilen başarımı yüksek algoritmalar Türkçe metinlerin anlamsal analizi ile ilgili çalışmaların da sayısını ve niteliğini arttıracaktır.

Bu çalışmada, incelenen yöntemlerin sınırlılıklarına da yer verilmiştir. Metin kümeleme çalışmaları için en temel parametre dokümanların içerdiği konu sayısıdır. OGAA ve GDA yöntemlerinin doğru sonuç verebilmesi

için konu sayısının tam olarak bilinmesi gerekmektedir. Ancak uygulamada, değerlendirilen dokümanların kaç farklı konu ile ilgili olduğu bilinmeyebilir. Gelecek çalışmalarda, konu sayısının bilinmediği durumlar için önerilen güncel yaklaşımlar ve kelimelerin bağımsız değil anlamsal olarak ilişkili olduğunu kabul eden KHTM algoritması incelenebilir.

Kaynaklar

[1] Aggarwal CC, Zhai C, 2012. An Introduction to Text Mining. In: Aggarwal CC, Zhai C, editors. Mining text data, New York: Springer, p. 1-10.

[2] Hoffman T, 2001. Unsupervised Learning by Probabilistic Latent Semantic Analysis. Machine Learning, 42 (1-2): 177-196.

[3] Deerwester S, Dumais ST, Furnas GW, Landauer TK, Harshman R, 1990. Indexing by latent semantic analysis. Journal of the American Society for Information Science, 41(6): 391-407.

[4] Blei DM, Ng AY, Jordan MI, 2003. Latent dirichlet allocation. Journal of Machine Learning Research, 3: 993-1022.

[5] Aggarwal CC, Zhai C, 2012. An Introduction to Text Mining. In: Aggarwal CC, Zhai C, editors. Mining text data, New York: Springer, p. 259-295.

[6] Sharma D, Kumar B, Chand S, 2017. A Survey on Journey of Topic Modeling Techniques from SVD to Deep Learning. International Journal of Modern Education and Computer Science, 9 (7): 50-62.

[7] Onan A, Korukoglu S, Bulut H, 2016. LDA-based Topic Modelling in Text Sentiment Classification: An Empirical Analysis. International Journal of Computational Linguistics and Applications, 7 (1): 101-119.

[8] Song G, Ye Y, Du X, Huang X, Bie S, 2014. Short text classification: A survey. Journal of Multimedia, 9 (5): 635-643.

[9] Yan E, 2015. Research dynamics, impact, and dissemination: A topic-level analysis. Journal of the Association for Information Science and Technology, 66 (11): 2357-2372.

[10] Ekinci E, Omurca SI, 2016. Ürün Özelliklerinin Konu Modelleme Yöntemi ile Çıkartılması. Türkiye Bilişim Vakfı Bilgisayar Bilimleri ve Mühendisliği Dergisi, 9 (1): 51-58.

[11] Onan A, 2017. Türkçe Twitter Mesajlarında Gizli Dirichlet Tahsisine Dayalı Duygu Analizi, 19. Akademik Bilişim Konferansı, Aksaray, Türkiye.

[12] Dempster AP, Laird NM, Rubin, DB, 1977. Maximum likelihood from incomplete data via the EM algorithm. Journal of the royal statistical society: 1-38.

[13] Cao J, Xia T, Li J, Zhang Y, Tang S, 2009. A density-based method for adaptive LDA model selection. Neurocomputing, 72 (7-9): 1775-1781.

[14] Blei D M, Lafferty J D, 2007. A correlated topic model of science. The Annals of Applied Statistics, 1 (1): 17-35.

[15] Yau C K, Porter A, Newman N, Suominen A, 2014. Clustering Scientific Documents with Topic Modeling. Scientometrics, 100 (3): 767-786.

[16] Teh YW, Jordan M, Beal MJ, Blei DM, 2006. Hierarchical Dirichlet Processes. Journal of the American Statistical Association, 101 (476): 1566-1581.

[17] Xu Y, Yin J, Huang J, Yin Y, 2018. Hierarchical topic modeling with automatic knowledge mining. Expert Systems with Applications, 103: 106-117.

[18] Zemberek NLP, http://zemberek-web.appspot.com/ [erişim 03/2014].

[19] Uzun V, 2014. Semantic text mining and an application in Turkish documents. Yayınlanmamış Yüksek Lisans Tezi, Dokuz Eylül Üniversitesi Fen Bilimleri Enstitüsü.

[20] Hornik K, Grün B, 2011. topicmodels: An R package for fitting topic models. Journal of Statistical Software, 40 (13): 1-30.

[21] Porter MF, 1980. An algorithm for suffix stripping. Program, 14 (3): 130-137.

Ek – Uygulamada kullanılan haber metinleri

Haber metni-1

HA-1_1	Russell Crowe, yönetmen koltuğunda oturduğu ve başrolünü oynadığı 'The Water Diviner'ın İstanbul çekimlerinin ardından Fethiye'ye geldi. Filmin son sahnelerinin çekileceği Kayaköy, Crowe için film platosu haline getirildi. Crowe, tarihi kilise ve evlerde çekimleri tamamlanacak film için yapılan son hazırlıkları yerinde inceledi.
HA-2_1	Gladyatör filminin ünlü oyuncusu Avustralyalı Russell Crowe, The Water Diviner filminin çekimlerine İstanbul'un ardından Muğla'nın Fethiye ilçesinde devam ediyor. Çanakkale Savaşı'nda iki oğlunu kaybeden ve sağ kalan diğer oğlunu bulmak için büyük bir mücadeleye girişen Avustralyalı bir çiftçiyi canlandıran Crowe, Kayaköy'de atlı sahnelerin yanı sıra kilisedeki çekimlerini gerçekleştirirken bu sahnelerde Yılmaz Erdoğan da eşlik ediyor.
HA-3_1	Avustralyalı aktör Russell Crowe'un yönetmenliğini ve başrolünü üstlendiği "The Water Diviner" filminin Türkiye'deki sahnelerinin bir bölümünün Fethiye'nin tarihi Kayaköy ören yerinde

ISSN: 2536-4383

	çekilmesinin ilçenin tanıtımına katkı sağlayacağı bildirildi. Çanakkale Savaşı'nda çocuklarını kaybeden ve onların ardından Türkiye'ye gelen bir babanın hikâyesini anlatan "The Water Diviner" filminin İstanbul'daki çekimleri, 26 Şubat'ta başladı.
HA-4_1	Oscar ödüllü oyuncu Russell Crowe, yönetmen koltuğunda oturduğu ve başrolünü oynadığı 'The Water Diviner'ın İstanbul çekimlerinin ardından Fethiye'ye geldi. Filmin son sahnelerinin çekileceği 5 bin yıllık Kayaköy, Crowe için film platosu haline getirildi. Ünlü yönetmen, tarihi kilise ve evlerde çekimleri tamamlanacak film için yapılan son hazırlıkları yerinde inceledi
Haber metni-2	
HA-1_2	Malezya Havayolları'na ait 239 kişi taşıyan yolcu uçağı Tho Chu adası yakınlarında denize çakıldı. Kuala Lumpur - Pekin uçuşunu yapmakta olan MH370 sefer sayılı Boeing 777-200 tipi yolcu uçağıyla bağlantı yerel saatle gece 2.40'ta (TSİ 20.40) kesilmişti.
HA-2_2	Kuala Lumpur'dan Pekin'e gitmekte olan Malezya Havayolları'na ait yolcu uçağı, Vietnam yakınlarında denize düştü. Uçakta 227'si yolcu, 12'si mürettebat toplam 239 kişinin bulunduğu öğrenildi. Uçaktan sinyal alınan bölgeye arama kurtarma ekiplerinin sevk edildiği açıklandı
HA-3_2	Malezya Havayollarına ait 227 yolculu MH 370 sefer sayılı yolcu uçağından haber alınamıyor. Malezya Havayollarından yapılan açıklamada, Kuala Lumpur'dan Pekin'e gitmek üzere havalanan 12 mürettebat ve 227 yolcusu bulunan uçağın kaybolduğu bildirildi.
HA-4_2	Malezya havayollarına ait Boeing B777-200 tipi uçakla iletişim kesildi. Kuala Lumpur'dan Pekin'e hareket eden uçak ile bağlantı kesildi. Uçakta 239 yolcu 12' de kabin görevlisi bulunuyordu. Yolculardan 160'ının Çin vatandaşı olduğu belirtildi.
Haber metni-3	
HA-1_3	Galatasaray, Şampiyonlar Ligi 2. tur ilk maçında İngiltere'nin Chelsea takımı ile 1-1 berabere kaldı. Türk Telekom Arena'da oynanan maçta Galatasaray'ın golünü 64. dakikada Aurelien Chedjou atarken, konuk ekip Chelsea'nin golünü ise 9. dakikada Fernando Torres kaydetti.
HA-2_3	Şampiyonlar Ligi 2. Tur ilk maçında Galatasaray, Türk Telekom Arena'da İngiliz ekibi Chelsea'yi ağırladı. 1-1 biten karşılaşma sonunda

ISSN: 2536-4383

	düzenlenen basın toplantısında iki takımın teknik direktörleri açıklamalarda bulundu.
HA-3_3	UEFA Şampiyonlar Ligi 2. tur ilk maçında Galatasaray, İngiliz ekibi Chelsea ile 1-1 berabere kaldı. Çeyrek finale kalacak takımı rövanş karşılaşması belirleyecek. 64. dakikada Galatasaray beraberliği sağladı. Sneijder'in soldan kullandığı korner atışında altıpasta uygun durumda topla buluşan Chedjou, meşin yuvarlağı ağlarla buluşturdu: 1-1
HA-4_3	Şampiyonlar Ligi 2. tur ilk maçında Galatasaray, konuk ettiği İngiltere Premier Ligi'nin güçlü ekibi Chelsea ile 1-1 berabere kaldı. Türk Telekom Arena'da oynanan ve İspanyol hakem Carlos Velasco Carballo'nun düdük çaldığı maçta konuk Chelsea'nin golünü 9. dakikada Torres attı. Galatasaray'ın beraberlik golünü ise 60. dakikada Chedjou kaydetti.