Journal of the Faculty of Engineering and Architecture of Gazi University 35:4 (2020) 2135-2145

Mühendislik Mimarlık Fakültesi Dergisi
Journal of The Faculty of Engineering and Architecture of Gazi University

Elektronik / Online ISSN : 1304 - 4915
Basılı / Printed ISSN : 1300 - 1884

Comparison of n-stage Latent Dirichlet Allocation versus other topic modeling methods for emotion analysis

Zekeriya Anıl Güven[1]*, Banu Diri[2], Tolgahan Çakaloğlu[3]

[1]Computer Engineering, Ege University, Izmir, 35100, Turkey
[2]Computer Engineering, Yildiz Technical University, Istanbul 34220, Turkey
[3]Department of Computer Science, University of Arkansas, Arkansas, 72762, ABD

Highlights:

- Proposing a new method, using the Latent Dirichlet Allocation algorithm
- The effect of proposed method on emotion detection
- The proposed approach is constantly superior to other topic modeling algorithms

Keywords:

- Sentiment analysis,
- topic modelling,
- social network analysis,
- natural language processing,
- social media

Article Info:

Research Article
Received: 19.04.2019
Accepted: 26.05.2020

DOI:

10.17341/gazimmfd.556104

Correspondence:

Author: Zekeriya Anıl Güven
e-mail: zekeriya.anil.guven@ege.edu.tr
phone: +90 506 562 3572

Graphical/Tabular Abstract

Understanding of human emotion is an essential skill to be applicable in various fields of rapidly growing technology. Social media is one of the critical technological fields. We proposed a new method, namely n-stage Latent Dirichlet Allocation (n-LDA), to improve the emotion detection with an application to Turkish tweets. We, further, compared our approach with other topic modeling algorithms, such as Latent Dirichlet Allocation (LDA), Latent Semantic Analysis (LSA) and Probabilistic-Latent Semantic Analysis (P-LSA) were used to determine the emotions of individuals from Turkish tweets.

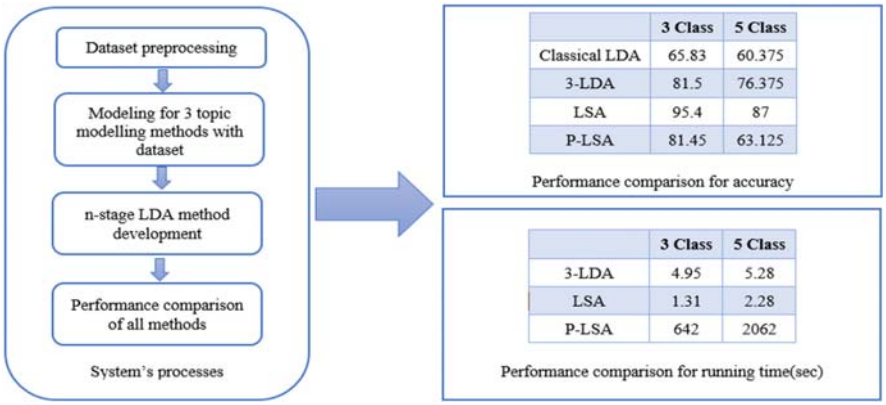


Figure A. System process and performance comparison

Purpose: This study aims to develop a Lda-based method for better enhancing emotion detection from tweets. The proposed method identifies the words, that have low weights, and eliminates them so that they do not contribute any negativity to the overall results. Additionally, we compare the performance of the proposed method with other topic modeling methods.

Materials and Methods:
The dataset consists of 4000 tweets from 5 different types of emotions, including such as anger, fear, happiness, sadness, and surprise. Since the popular benchmark datasets are not accessible from Turkey, we had to collect tweets and were able to create our dataset. We, first, applied preprocessing steps such as removing punctuation marks, converting all content to lowercase, removing stopwords, and applying a lemmatisation. Then, we executed other topic modeling algorithms—LDA, LSA, and P-LSA— using the preprocessed tweet data. Additionally, the proposed n-stage LDA method is implemented to extend the performance of the system further. The steps of the n-LDA can be listed as follows: Calculating the threshold value for each topic, creating a dictionary from words that have weights higher than the threshold value, and modeling the system based on LDA, using the newly created dictionary. Finally, all method's successes and running times were measured and.compared.

Results:
The proposed n-stage LDA is constantly superior— in terms of running time and accuracy— to LDA and P-LSA. It is important to note that LSA showed the best performance compared to other methods.

Conclusion:
We experimented that removing the words from the dictionary created a positive impact on the system. We showed that the proposed n-stage LDA performed better than P-LSA. Additionally, we also observed that LDA was even not as successful as P-LSA.

Journal of the Faculty of Engineering and Architecture of Gazi University 35:4 (2020) 2135-2145

Mühendislik Mimarlık Fakültesi Dergisi
Journal of The Faculty of Engineering and Architecture of Gazi University

Elektronik / Online ISSN : 1304 - 4915
Basılı / Printed ISSN : 1300 - 1884

Duygu analizi için n-aşamalı Gizli Dirichlet Ayırımı ile diğer konu modelleme yöntemlerinin karşılaştırılması

Zekeriya Anıl Güven[1]*, Banu Diri[2], Tolgahan Çakaloğlu[3]
[1]Ege Üniversitesi, Mühendislik Fakültesi, Bilgisayar Mühendisliği Bölümü, 35100, İzmir, Türkiye
[2]Yıldız Teknik Üniversitesi, Mühendislik Fakültesi, Bilgisayar Mühendisliği Bölümü, 34220, İstanbul, Türkiye
[3]University of Arkansas, College of Engineering, Department of Computer Science and Computer Engineering, 72762, ABD

ÖNEÇIKANLAR

- Gizli Dirichlet Ayırımı tabanlı n-aşamalı GDA yöntemi geliştirme
- Geliştirilen yöntemin duygu tespit etme üzerindeki etkisi
- Geliştirilen yöntem ile diğer konu modelleme algoritmalarının performans karşılaştırması

Makale Bilgileri
Araştırma Makalesi
Geliş: 19.04.2019
Kabul: 26.05.2020

DOI:
10.17341/gazimmfd.556104

Anahtar Kelimeler:
Duygu analizi,
konu modelleme,
doğal dil işleme,
sosyal ağ analizi,
sosyal medya

ÖZET

Sosyal medyada yer alan paylaşımlara ait duyguları anlamak, insanların düşüncelerini öğrenmede kilit bir rol almaktadır. Gelişen teknoloji ile insanın duygusunu bilmek, çeşitli alanlarda yarar sağlamaktadır. Örneğin medya, pazarlama ve reklam gibi alanlar insanların kullanımlarına ve fikrine özgü içerikleri kullanıcıya yansıtabilme imkanı tanımaktadır. Çalışmamızda konu modelleme algoritmalarından Gizli Dirichlet Ayırımı (GDA), Gizli Anlamsal Analiz (GAA) ve Olasılıksal-Gizli Anlamsal Analiz (O-GAA) Türkçe tivitlerden kişilerin duygularını belirlemede kullanılmıştır. Ayrıca, GDA algoritmasının geliştirilen n-aşamalı halinin duygu analizindeki başarısı mevcut yöntemlerle de karşılaştırılmıştır. Kullanılan veri seti kızgın, korku, mutlu, üzgün ve şaşkın olmak üzere 5 farklı duyguya ait 4000 tivitten oluşmaktadır. Tüm konu modelleme yöntemleri 3 ve 5 sınıflı veri seti için modellenerek başarıları ve çalışma süreleri ölçülmüştür. Geliştirilen n-aşamalı GDA yönteminin, GDA ve O-GAA'ya göre çalışma süresi ve performansı açısından başarı sağladığı gözlemlenmiştir. En başarılı ve en hızlı modellenen yöntem ise GAA olmuştur.

Comparison of n-stage Latent Dirichlet Allocation versus other topic modeling methods for emotion analysis

HIGHLIGHTS

- Developing a new method with Latent Dirichlet Allocation
- The effect of proposed method on emotion detection
- Performance comparison of other topic modeling algorithms with developed method

Article Info
Research Article
Received: 19.04.2019
Accepted: 26.05.2020

DOI:
10.17341/gazimmfd.556104

Keywords:
Sentiment analysis,
topic modelling,
social network analysis,
natural language processing,
social media

ABSTRACT

Understanding the emotions of sharing in social media plays a key role in learning people's thoughts. Knowing the emotion of human being with developing technology provides benefit in various fields. For example, media, marketing and advertising areas allow people to reflect on their use and idea specific content. In our study, Latent Dirichlet Allocation (LDA), Latent Semantic Analysis (LSA) and Probabilistic-Latent Semantic Analysis (P-LSA) were used to determine the emotions of individuals from Turkish tweets. In addition, the success of the developed n-stage state of the LDA algorithm in the emotion analysis was compared with the existing methods. The dataset consists of 4000 tweets of 5 different emotions, including angry, fear, happiness, sadness and surprise. All topic modeling methods were modeled for 3 and 5 class datasets and their successes and running times were measured. It has been observed that the developed n-stage LDA method achieves success in terms of running time and performance according to LDA and P-LSA. The most successful and fastest modeled method was LSA.

*Sorumlu Yazar/Corresponding Author: *zekeriya.anil.guven@ege.edu.tr, diri@yildiz.edu.tr, txcakaloglu@ualr.edu / Tel: +90 506 562 3572

1. GİRİŞ (INTRODUCTION)

Günümüz modern bilgi sistemleri, firmaların çok büyük miktarda verilere sahip olmasını sağlamaktadır. Bu verilerin çoğu, geleneksel veri tabanı yazılımı kullanılarak analiz edilebilecek şekilde yapılandırılmış verilerdir. Bununla birlikte, metinsel veriler gibi büyük miktarlardaki verilerde yapılandırma sorunu olup, verilerin elle analizi de giderek zorlaştığından, verilerin analizini otomatik bir şekilde yapmak için metin madenciliği yöntemleri geliştirilmiştir. Metin madenciliği, bilgi sistemleri disiplinindeki konu keşfinin, müşteri ilişkileri yönetiminin ve hedef reklamcılığın merkezini tanımlamak için kullanılmaktadır. Ayrıca, görüntülerden gelen metinsel nesneleri tespit etmek için de kullanılmaktadır. Metin verilerinin analiz edilmesi ve anlaşılması ihtiyacı göz önüne alındığında, metin madenciliği yöntemlerinin incelemesi ve alanlara uygulanması önemlidir [1]. Metin madenciliğinde kullanılan önemli yöntemlerden biri konu modelleme algoritmalarıdır. Konu modelleme, metin belgesinin anlamsal yapısını belirlemeyi sağlayan bir makine öğrenmesi yöntemidir. Konu modelleme yöntemleri yüksek içeriğe sahip metinlere, özetleme, sınıflandırma ve kümeleme gibi birçok metin madenciliği alanında başarılı şekilde uygulanmaktadır. Konu modelleme yöntemlerinin uygulandığı en önemli alanlardan biri de duygu analizidir. Duygu analizi en sık kullanılan sosyal medya izleme yöntemidir. Kullanıcıların Facebook, Twitter, Blog vb. gibi sosyal medya ortamlarında paylaştığı herhangi bir konu ile ilgili duygu ve düşüncesini analiz etmeyi sağlamaktadır [2]. Birçok çalışma bu konu modellerinin ümit verici performansını belirtmektedir. Ancak, Türkçe duygular için konu modellerini ve performanslarını karşılaştıran çalışma sayısı yok denecek kadar azdır.

Literatürü incelediğimizde; Lee vd. [1], Gizli Anlamsal Analiz (GAA), Olasılıksal Gizli Anlamsal Analiz (O-GAA), Gizli Dirichlet Ayırımı (GDA) ve İlişkili Konu Modeli'nin avantaj ve sınırlamalarını tanımlamak için çalışmalarında bu yöntemleri incelemişlerdir. Dört yöntemi karşılaştırarak çeşitli metotların uygulanacağı optimum koşullara değinmişlerdir. Haidar vd. [3], konuşma tanıma için yeni bir bağlam tabanlı Olasılıksal Gizli Anlamsal Analiz (O-GAA) dil modeli önermişlerdir. Geliştirilen modeli geçmişte önerilen düzensiz bigram O-GAA ile karşılaştırmışlardır. Mazarura vd. [4], Dirichlet Multinomial Mixture Model'in (GSDMM) kısa metinlere uygulandığında GDA'dan daha iyi performans göstermesi gerektiği hipotezinin geçerliliğini araştırmışlardır. Bu hipotezi araştırmak için her iki modelin de performansı, hem uzun bir metin derlemine hem de iki farklı kısa metin derlemine uygulanarak, konu tutarlılığı ve kararlığına dayalı performansı ölçülerek değerlendirilmiştir. Blei vd. [5], olasılıksal model olan Gizli Dirichlet Ayırımı (GDA) modelini unigram ve olasılıksal GAA modelinin bir karışımıyla karşılaştırarak, belge modelleme, metin sınıflandırma ve işbirlikçi filtreleme de raporlamışlardır. Kakkonen vd. [6], Otomatik Deneme Değerlendiricisi (AEA)'nde GAA, O-GAA ve GDA yöntemlerini doküman karşılaştırmaları için kullanmışlardır. Sonuçlar, sınıflandırma modeli için eğitimde öğrenme materyallerinin kullanılmasının, k-NN yöntemlerinden daha iyi performans verdiğini göstermiştir. Ek olarak, GAA kullanmanın O-GAA ve GDA'dan daha başarılı olduğunu tespit etmişlerdir. Lu vd. [7], doküman kümeleme, metin sınıflandırma ve özel amaçlı getirim dahil olmak üzere üç temsili metin madenciliği görevini kullanarak O-GAA ve GDA'nın sistematik incelemesini gerçekleştirmişlerdir. Böylece konu modellerinin daha derinlemesine anlaşılmasını ve tipik görevler için konu modellerinin performansının nasıl optimize edileceğine dair birçok faydalı çıkarımlar sunmuşlardır. Chien vd. [8], kelime dizisinin modellenmesi için GDA tabanlı yeni bir Gizli Dirichlet Dil Modeli (LDLM) geliştirmişlerdir. Deneylerde sürekli konuşma tanıma için LDLM uygulamışlar ve O-GAA tabanlı dil yönteminden daha iyi başarı sağlamışlardır. Chen [9], blog aramasının performansını analiz etmek için GAA, GDA ve O-GAA konu modelleme yöntemlerini kullanmıştır. Sistemin performansını geliştirmek için yayınlar arasındaki zaman ilişkisini değerlendiren Latent Time Relationship (LTR) isimli yöntem önermiştir. Önerilen sistem ile performans önemli ölçüde artmıştır. Xiong vd. [10], kısa metin incelemeleri için ortak bir duygu-konu modeli olan Word-Pair Sentiment-Topic Model (WSTM) önermişlerdir. Yöntem ile metinden aynı anda duyguları ve konuları tespit etmişlerdir. Ayrıca GAA ve Joint-Sentiment Topic (JST) gibi yöntemler ile WSTM'yi karşılaştırarak daha iyi sonuç verdiğini gözlemlemişlerdir.

Yaptığımız çalışmada ilk olarak, konu modelleme algoritmalarından yaygın olarak kullanılan Gizli Dirichlet Ayırımı (GDA) ile Türkçe tivitlerin hangi duyguya ait olduğunu tespit etmek amaçlanmaktadır. Sonrasında, klasik GDA yöntemini n-aşamalı kullanılabilecek şekilde bir sistem önerilerek klasik GDA yöntemi ile karşılaştırılması yapılmıştır.

Çalışmamızda ikinci olarak, konu modelleme yöntemlerinin kullanımını daha geniş bir araştırmacı kitlesine açmak için önerdiğimiz yöntem de dahil olmak üzere dört konu modelleme yöntemi, Türkçe tivitler için incelenmiştir. Gizli Anlamsal Analiz (GAA), Olasılıksal Gizli Anlamsal Analiz (O-GAA), Gizli Dirichlet Ayırımı çalışmamızda kullanılan yöntemler olup, aynı konu sayısı ve veri seti ile modellenerek performanslarının karşılaştırılması yapılmıştır. Böylece, literatüre katkı sağlamak amacıyla n-aşamalı GDA yöntemi önerilmiştir. Ayrıca Türkçe metinler için literatürde az bulunan konu modelleme yöntemlerinin uygulanmasına örnek sunulmak istenmiştir. Makalenin ikinci bölümünde veri seti, ön işlem adımları ve kullanılan yöntemler detaylı şekilde anlatılmıştır. Üçüncü bölümde, deneysel çalışmalar ve bunların duygu analizindeki sonuçlarına yer verilmiştir. Dördüncü bölümde ise deney sonuçlarının karşılaştırılması yapılmıştır ve gelecekteki çalışmalar hakkında bilgi verilmiştir.

2. MATERYALLER VE METOTLAR (METHODS AND MATERIALS)

2.1. Gizli Dirichlet Ayırımı (Latent Dirichlet Allocation)

Gizli Dirichlet Ayırımı, bir derlemdeki gizli konuları tanımlamak için ortak oluşum teriminin kullanıldığı olasılıksal üretici bir modeldir. GDA tarafından tanımlanan bir konu, konuya atanan bir kelimenin olasılığını gösteren bir olasılık dağılımı olarak modellenmektedir. Bir doküman ise, her bir konuyu ifade etme olasılığını gösteren bir olasılık dağılımı olarak ifade edilmektedir [11]. Model, konunun yapısını, gözlenen veri setinden elde edilen kelime ve kelime ağırlık değerleri ile belirlemeyi amaçlamaktadır.

GDA, konu tespiti için kullanılan etkili, denetimsiz bir öğrenme yöntemidir. Yöntem, etiketlenmemiş çok sayıda dokümanlardan oluşan bir koleksiyon kullanmaktadır [12]. Bu yöntem, her dokümanı her bir konunun kelimeler üzerindeki çok terimli dağılımına sahip bir konu karışımı olarak modellemektedir. GDA tarafından modellenen dokümanın konu ve kelime-ağırlık dağılımları doküman için en iyi konuları göstermektedir [13].

GDA ile dokümandaki tüm kelimelere rastgele konu ataması yapılmaktadır. Bu bilgiden yararlanarak belirli istatistikler çıkartılır. Yerel istatistik, her dokümandaki konulara kaç adet kelime atandığını belirtmektedir. Global istatistik ise tüm dokümanda yer alan her kelimenin konulara kaç kere atandığını göstermektedir. İstatistiksel bilgiler ile her doküman için her kelimeye yeniden konu ataması yapılmaktadır [14].

$$\frac{n_{ik}+\alpha}{N_i-1+K\alpha} \quad (1)$$

Eş. 1'de kelimeler, konulara atanırken dokümanın konular ile ilişkisi hesaplanmaktadır. n_{ik}, i. haber için k. konuya atanan kelime sayısıdır. N_i ise dokümandaki toplam kelime sayısını vermektedir. Kullanılan kelime hesaba katılmadığından değerinden 1 çıkartılmaktadır. α, konuların dokümanlardaki dağılımını göstermektedir. K değeri ise belirlenen konu sayısıdır.

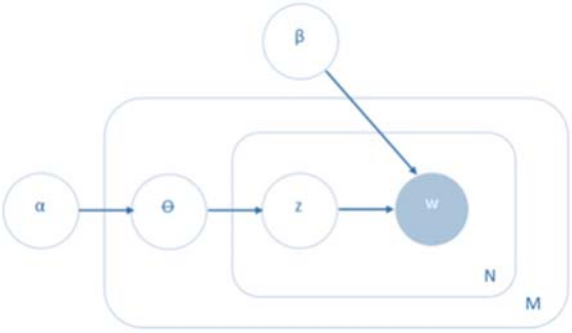
GDA için K konu sayısı, konu modelleme kriteri olan tutarlılık değeri ile belirlenir. Tutarlılık değeri kelimelerin benzerliğini ölçmektedir. En yüksek sonuca sahip olan K değeri, modelleyeceğimiz sistemin konu sayısı olarak seçilmektedir [14].

$$\frac{n_{word,k}+\beta}{\sum_{w\in V} n_{w,k}+V\beta} \quad (2)$$

Yöntem sonrasında, her kelimenin konular ile ilişkisi belirlenmektedir. Hesaplama ile kelimenin, ilgili konu ile ağırlığı hakkında bilgi çıkarımı yapılmaktadır. Eş. 2'de; $n_{word,k}$ geçerli kelimenin k. konuya tüm dokümandaki atanma sayısıdır. β değeri; kelimelerin konular üzerindeki dağılımıdır. V ise tüm kelimelerden oluşan sözlüğün boyutudur. Eş. 1 ve Eş. 2'den alınan sonuçlar çarpılarak geçerli kelimenin k. konuya atanma olasılığı bulunmaktadır. Tüm doküman sayısı boyunca değerler tekrar hesaplanmaktadır. Kelime için en yüksek değere ait olan konu, yeni konu olarak belirlenmektedir. Veri setindeki tüm dokümanlara ait her kelime için aynı işlemler uygulanmaktadır. Kelimelerin konu güncellemeleri belirlenen iterasyon kadar yapılmaktadır. En son olarak sistemin modelini çıkarmak için doküman-terim matrisi oluşturulur. Matris ile kelimelerin ağırlıkları hesaplanır ve konulardaki ağırlıkları belirlenir [14].

GDA yapısı Şekil 1'de gösterilmiştir. Rastgele değişkenler düğümlerle gösterilir. Düğümler arasındaki muhtemel bağlantılar, kenarlar kullanılarak temsil edilir. Şekil 1'de;

- α, belge başına konu dağılımını temsil eder.
- β, konu başına kelime dağılımını gösterir.
- Θ, verilen bir belgenin konu dağılımını gösterir.
- z, her kelime için atanmış bir konudur.
- w, gözlenen bir kelimedir.



Şekil 1. Gizli Dirichlet Ayırımı yapısı [15]
(Structure of latent dirichlet allocation)

Şekil 1'deki yapıda, a ve β parametreleri sistem başlatıldığında bir kez örneklenir. Sistemdeki her doküman için Θ parametresi örneklenir [15].

2.2. Gizli Anlamsal Analiz (Latent Semantic Analysis)

Gizli Anlamsal Analiz, kelimelerle metin parçaları arasındaki benzerlik ilişkilerini ortaya çıkaran ve kelimelerin veya bölümlerin anlamını çıkaran metin analiz yöntemidir. Yöntem, metin derlemini matrise dönüştürür ve boyut küçültme tekniği uygular. Böylece, verileri yalnızca temel boyutlara karşılık gelen katsayıları alarak yeniden yapılandırmak kolaylaşmaktadır. Yöntem için bu işlem metin derleminin temel bileşenlerini alabilmenin kilit noktasıdır. Bunun yanı sıra, verileri yeniden oluşturduktan sonra her kelime ve geçiş, bu vektörlerin korelasyonu temelinde birçok uygulamaya izin veren vektör olarak temsil edilebilir [16].

Sözlükler, bilgi tabanları, gramerler veya sözdizimsel ayrıştırıcılar kullanılmadığı için GAA'nın geleneksel doğal dil işlemeden tamamen farklı olduğunu belirtmek önemlidir. Yalnızca girdi olarak anlamlı geçişler olarak ayrılan ham metinleri almaktadır [16].

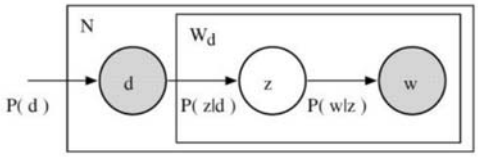
Şekil 2'de GAA yönteminin aşamaları gösterilmiştir. GAA, büyük metin derleminden her bir sıranın ayrı bir terim ve her bir sütunun da ayrı bir dokümanı temsil ettiği terim-doküman eşleşme matrisini oluşturmaktadır. Bu terim-doküman matrisinin hücreleri, frekans sayıları ile doldurulmaktadır. Bu matris temsili sadece terim-doküman birlikte oluşumunu hesaba kattığından, dokümandaki kelime sırası göz ardı edilmektedir. Yani, her doküman bir "sözcük torbası" olarak temsil edilmektedir. Matris oluşumundan sonra ham terim frekanslarıyla çalışmak yerine, terim-doküman matrisindeki hücre sayıları kullanılmaktadır. Bu bağlamda ilk adım, her hücre sayısını o hücre sayısının logaritmik haline dönüştürmektir. Daha sonra, hücre girişleri entropiye bölünmektedir. Bu ters entropi ölçüsü, belirli bir kelime ile ne kadar anlamsal bir ilişki içinde olduğunu göstermektedir [17]. GAA'da, tipik bir derlemdeki doküman sayısı göz önüne alındığında, oluşturulan terim-doküman matrisinin çok yüksek boyutlu vektörler olması muhtemeldir. GAA, matris üzerinde gerçekleştirilen düşük kademeli tekil değer ayrışmasının (SVD) sonucunda, boyut azalması ile diğer vektör uzay modellerinden ayrılmaktadır. SVD, bir matrisin ana bileşenlere ayrıştırılmasında kullanılan genel bir yöntemdir. İstenilen boyut sayısını belirten bir k parametresi ile çağrılmaktadır. K-boyutlu vektörün değerlerini belirleyen dokümanlar arasında sadece bir terimin kendi dağılımı yoktur. Aksine, SVD bir terimin tam olarak hangi dokümanlarda gerçekleştiğini en iyi tahmin eden vektörleri üretmek için elindeki tüm doğrusal ilişkileri kullanmaktadır. Son olarak, SVD ile oluşturulan vektör terimleriyle kelimeler arasındaki ilişkiyi hesaplamak için iki vektör arasındaki kosinüs benzerliği hesaplanmaktadır [17].

İdeal boyut sayısı seçimi, GAA'yı saf vektör uzay modelinden ayırmaktadır. Eğer tüm boyutlar kullanılıyorsa, orijinal matris yeniden yapılandırılacak ve değiştirilmemiş vektör uzay modeli daha sonraki işlemler için temel teşkil edecektir. Mevcut sıfırdan küçük olmayan değerlerden daha küçük boyutlar kullanılıyorsa, orijinal vektör uzayına yaklaşılmaktadır. Böylece, orijinal matristeki doğal ilgili yapı bilgisi yakalanır ve sözcük kullanımındaki gürültü ve kararsızlık azaltılmaktadır [18].

Doküman veya terim vektörlerinin benzerliğinin nasıl ölçüldüğü başka bir etkileyici parametre sınıfı oluşturmaktadır. Hem seçilen benzerlik ölçüsü hem de benzerlik ölçüm yöntemi sonuçları etkilemektedir. GAA'da çeşitli korelasyon ölçümleri uygulanmaktadır. Bunlara basit çapraz çarpım, Pearson korelasyonu ve Spearman'ın Rho'su örnek verilebilir [18].

2.3. Olasılıksal Gizli Anlamsal Analiz
(Probabilistic Latent Semantic Analysis)

Olasılıksal-Gizli Anlamsal Analiz, çıkarım uygulamalarında terim eşleşmesi için GAA'dan daha iyi sonuçlar verdiği gösterilen, veriler için gizli değişken modelidir. Doküman d'den ((d ∈ D = $\{d_1, \dots, d_j\}$)) oluşan w sözcüğü (w ∈ W = $\{w_1, \dots, w_i\}$), her gözlemle (d, w) gözlemlenmemiş bir sınıf değişkeni olan z'yi (z ∈ Z = $\{z_1, \dots, z_k\}$) ilişkilendirir. Dokümandaki her bir sözcük, karışım bileşenlerinin gizli konuların gösterimi olarak görünebilen çok terimli rastgele değişkenlerin olduğu bir karışım modeli örneğidir. Bir doküman, karışım bileşenleri için karışım oranlarının bir listesi olarak temsil edilmektedir. Yani doküman sabit gizli sınıf kümesi üzerinde olasılık dağılımına indirgenmektedir [19].



Şekil 3. O-GAA algoritmasının yapısı [20]
(Structure of O-GAA algorithm)

Şekil 3'te O-GAA algoritmasının yapısı gösterilmiştir. Üretken bir model olarak, Şekil 3'teki yapı aşağıdaki gibi tanımlanabilir:

- P (d) olasılıklı bir doküman seçilir,
- P (z | d) olasılıklı gizli bir z sınıfı seçilir,
- P (w | z) olasılığı olan bir kelime üretilir.

Her gözlem çifti (d, w) için ortaya çıkan olabilirlik ifadesi Eş. 3'te verilmiştir:

$$P(d,w) = P(d)P(w|d),$$
$$P(w|d) = \sum_{z\in Z} P(w|z)P(z|d) \quad (3)$$

Doküman d ve w kelimesi, gözlemlenmemiş konu olan z göz önüne alındığında şartlı olarak bağımsız kabul edilir.

Şekil 2. Gizli Anlamsal Analiz yönteminin aşamaları [17] (Processes of latent semantic analysis)

Maksimum olabilirlik ilkesiyle karışım bileşenleri ve karışım oranları, olasılık fonksiyonunun maksimize edilmesiyle belirlenir:

$$L = \sum_{d \in D} \sum_{w \in W} n(d,w)\big(logP(d,w)\big) \quad (4)$$

Eş. 4'te n(d, w) ifadesi frekans terimini, yani d dokümanındaki w sözcük sayısını göstermektedir [19].

Gizli değişkenlerin yer aldığı olasılık fonksiyonunu en üst düzeye çıkarmak için standart yöntem Beklenti Maksimizasyonu (EM) algoritması kullanılır. EM, her yinelemenin iki basamaktan oluştuğu, gizli sınıflar z için önceki olasılıkların hesaplandığı bir beklenti adımından oluşan yinelemeli bir algoritmadır. Ayrıca yöntem, gizli sınıfların önceki olasılıklarını ve verilen parametrelerin koşullu olasılıklarını güncelleyen bir maksimizasyon basamağıdır. Beklenti ve maksimizasyon adımları birbirini izlerken, biri logaritma olasılığının yerel maksimumunu tanımlayan bir birleşme noktasına varır. Algoritmanın çıktısı, karışım bileşenleri ve her eğitim dokümanı için bileşenler üzerindeki karışım oranlarıdır (P (w | z) ve P (z | d)) [21].

2.4. n-aşamalı GDA (n-stage LDA)

Gizli Dirichlet Ayırımı algoritmasıyla modellenen sistemin başarısını arttırmak için n-aşamalı bir yöntem geliştirilmiştir. Önerilen sistemi n aşamalı olarak adlandırmamızın nedeni, sistemde kullanılan veri setinin büyüklüğüne veya tivitlerdeki konu ile ilişkilli kelime miktarına göre dinamik olmasıdır. n değeri 1 ise klasik GDA'dır. n değeri, 2'den başlayarak doğrusal olarak artırılmaktadır. Şekil 4, yöntemin adımlarını göstermektedir.

Şekil 4'te gösterilen adımları aşağıdaki gibi açıklayabiliriz. Öncelikle, konulardaki kelimelerin ağırlıklarından yararlanarak her bir konu için eşik değeri hesaplanmaktadır. Eşik değer, konuya ait kelimelerin ağırlık değerleri toplamının toplam kelime sayısına oranı ile elde edilmektedir:

$$ed(k_i) = \frac{\sum_{j=1}^{m} w_j}{n_i}\ w_j \geq\ ed(k_i) \quad (5)$$

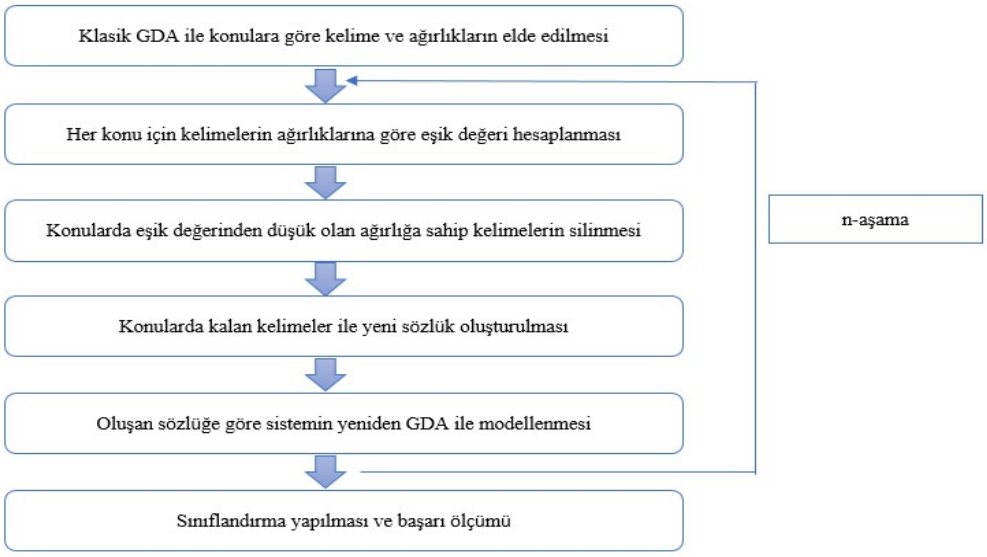
Eş. 5'te k_i, i. konuyu göstermektedir. w_j, k_i konusundaki kelime ağırlıklarını, n_i ise k_i konusundaki toplam kelime sayısını belirtmektedir (i:konu sayısı; j:konudaki kelime sayısı). Bu eşik değeri her konu için hesaplanmaktadır. Daha sonra her konu için ağırlık değeri, eşik değerinden düşük olan kelimeler sözlükten silinmektedir. Böylece kalan kelimelerle yeni sözlük oluşturulur. Son olarak, yeni sözlüğe göre sistem yeniden GDA ile modellenmektedir.

Ayrıca, bu yöntem ile tüm dokümanın sözlüğündeki kelime sayısını azaltmak amaçlanmıştır. Modelde düşük ağırlıklı kelimelerin yanlış sınıflandırmaya neden olabilmesinden dolayı sözlükteki kelimelerin sayısı azaltılmaktadır. Tablo 1, tüm veri setleri için sözlükteki kelimelerin sayısını göstermektedir. Sözlükteki kelime sayısı, aşama sayısı arttıkça beklendiği gibi azalmaktadır.

Tablo 1. Aşamaya göre veri setindeki kelime sayısı
(Word count in dataset by stage)

	1-aşama	2-aşama	3-aşama
Kelime Sayısı	2208	359	309

Aşama sayısı arttıkça, sözlükteki kelime sayısı azalır. Bundan dolayı, sözlükte kalan kelimelerin ağırlık değerleri değişmektedir. Tablo 2, aşama sayısı arttıkça örnek bir kelimenin ağırlık değerindeki değişimi göstermektedir. Tablo 2 incelendiğinde kelimenin ağırlık değerinin sürekli olarak arttığı belirlenmiştir. Bu artış, daha az kelimeyle sistemin modellenmesinden kaynaklanmaktadır. Sonuç olarak konunun hangi sınıf etiketine atanacağı daha kolay belirlenebilmektedir.



Şekil 4. Geliştirilen n-aşamalı GDA'nın adımları (Developed n-stage LDA's steps)

Tablo 2. Korku duygusuna ait 'kork' kelimesinin aşama arttıkça ağırlık değerleri
(Weight values of 'kork' word in the sense of fear by stage)

Aşama	Kelimenin Ağırlığı
1-aşama	0,132
2-aşama	0,343
3-aşama	0,688

2.5. Veri Seti (Dataset)

Türkçe tivitlerle ilgili mevcut veri seti bulunmadığı veya erişilir olmadığı için kendi veri setimiz oluşturulmuştur. Veri seti, Twitter aracılığıyla toplanan Türkçe tivitlerden oluşmaktadır. Tivit içerisinde en az bir duygu ifadesi içeren kelime olmasına dikkat edilmiştir. Veri setinde; mutlu, üzgün, şaşkın, korku ve kızgın olmak üzere beş farklı duygu etiketi kullanılmıştır. Tivitler manuel ve ortak görüş alınarak etiketlenmiştir. Her duygu sınıfı için 800 tivit olmak üzere toplamda 4000 adet tivit toplanmıştır. Eğitimde kullanılmak üzere 3 ve 5 sınıflı iki farklı veri seti kullanılmıştır. Üç sınıf için kızgın, korku ve mutlu etiketleri rastgele seçilerek 2400 tivit kullanılmıştır. Veri setinin %80'i eğitim, %20'si test için ayrılmıştır. Ayrıca, veri seti Zemberek kütüphanesiyle düzenlenmiştir.

2.6. Ön İşlemler (Preprocessing)

Ön işlem aşamasında ilk olarak, veri setindeki tivitlerden noktalama işaretleri ayrıştırılmıştır. Ardından, veri kümesindeki tüm tivitler küçük harfe dönüştürülmüştür. Türkçe karakter dönüşümü hatalı olabildiği için, İngilizce'de olmayan harfler kodda küçük harfe çevrilmiştir. Türkçe için kullanılan etkisiz kelimeler tivitlerden silinmiştir. Ayrıca, duygular için anlam ifade etmeyen kelimelerden oluşan bir liste hazırlanmış ve bu kelimeler tivitlerden silinmiştir. Ön işlemin aşamaları Şekil 5'te de gösterilmektedir.

Son olarak kelimelerin köklerini bulmak için Zemberek kütüphanesi kullanılmıştır. Kütüphane aracılığıyla veri setindeki tipi isim, sıfat ve fiil olan kelimelerden yeni bir veri seti oluşturulmuştur [22].

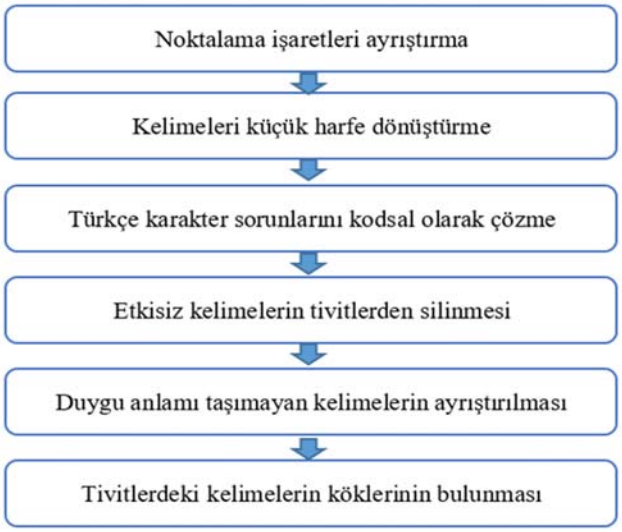
2.7. Programlama Dili (Programming Languages)

Uygulama Python (https://www.python.org) ortamında geliştirilmiştir. Ön işlemler, veri setlerini okuma, konu modelleme yöntemlerinin geliştirilmesi, dosyaya yazma gibi tüm işlemler Python programlama dilinde yazılmıştır. Tüm yöntemler için gerekli kütüphaneler Python ortamına eklenmiştir. Ardından kütüphaneler üzerinde konu modelleme yöntemleri gerçekleştirilerek Visual Studio ortamında çalıştırılmıştır.

Ek olarak, sadece veri setindeki kelimelerin köklerini ve biçimbirimsel analizini tespit için Zemberek kütüphanesi, Java platformunda kullanılmıştır.

3. DENEYSEL ÇALIŞMALAR (EXPERIMENTAL RESULTS)

Türkçe dilinde duygu verileriyle yapılan çalışmaların sayısı oldukça azdır. Veri kümeleri, ilgili çalışmalarda sıklıkla olumlu, olumsuz ve nötr olarak etiketlenmektedir. Ayrıca, literatürde kullanılan veri setlerine çoğunlukla erişim bulunmamaktadır. Bu nedenle, geliştirilen model klasik GDA ve diğer konu modelleme algoritmaları ile karşılaştırılmıştır. Klasik GDA sisteminin başarısını artırmak için n aşamalı GDA yöntemi önerilmiştir.



Şekil 5. Ön işlem aşamaları (Preprocessing stages)

Veri setlerine ön işleme adımları uygulandıktan sonra konu sayısını belirlemek için 3 ve 5 sınıflı veri setlerine ait 10 tane tutarlılık değeri hesaplanmıştır. En yüksek tutarlılık değerine sahip konu sayısı, sistemimizi eğiteceğimiz konu sayısı (K) olarak belirlenmiştir. Belirlenen konu sayısı, karşılaştırmanın doğru yapılabilmesi için tüm konu modelleme algoritmalarında aynı alınmıştır. Tablo 3'te, klasik GDA yönteminde 3 ve 5 sınıf için belirlenen konu sayısı ve tutarlılık değeri gösterilmiştir.

Tablo 3. Klasik GDA için sınıflara ait tutarlılık değeri ve konu sayısı
(Coherence value and topic number of class for classical LDA)

Sınıf	Tutarlılık değeri	Konu sayısı
3	0,4998	9
5	0,484	20

Sistem seçilen konu sayısı ile modellenirken, her konu kelimeler ve ağırlık değerlerine sahip bilgilerden oluşmaktadır. Konuya ait en uygun sınıf etiketi kelimelere ve kelimelerin ağırlıklarına göre atanmaktadır. Tablo 4'te konulara atanan sınıf etiketlerine örnek gösterilmektedir. On dört numaralı konu için kelime ve ağırlık değerlerini incelersek mutlu duyguyu ifade eden kelimeler belirgindir. Bu yüzden bu konu sayısına sınıf etiketi "mutlu" olarak atanmıştır. Tüm konuların ilgili sınıf etiketleri benzer şekilde

belirlenmiştir. Sonrasında klasik GDA yöntemi ile modellenen sistemin başarısı ölçülmüştür. Test edilen veri setindeki her tivite ait her konu için toplam ağırlık değerleri hesaplanmıştır. Test veri setindeki ilgili tivitin, en yüksek ağırlığa sahip konuya ait olan duygu, tivitin de duygu durumu olarak belirlenmiştir. Belirlenen duygu, test veri setindeki tivitin etiketli gerçek duygusuyla aynı ise doğru olarak ifade edilmiştir. Tablo 5'te, 3 ve 5 sınıflı veri seti için klasik GDA'nın başarısı gösterilmektedir. İncelendiğinde klasik GDA ile modellenen sistemin başarısı, sınıf sayısının artmasıyla azalmaktadır.

Tablo 4. GDA için konulara sınıf etiketleri atanmasına dair örnek
(Example of assigning class labels to topics for LDA)

Konu	Kelimeler ve Ağırlıkları	Etiket
2	'0.132*"kork" + 0.052*"nefret"+ 0.032*"korku"+ 0.016*"hediye" +…	Korku
4	'0.091*"sinir" + 0.076*"kafa" + 0.044*"irkil" + 0.042*"yiyecek" +…	Kızgın
6	'0.235*"mutsuz" + 0.113*"hüzün"+ 0.031*"hasta" + 0.023*"tatlı" + …	Üzgün
14	'0.161*"yaşa" + 0.103*"günü"+ 0.058*"doğum" + 0.046*"kutlu" + …	Mutlu
17	'0.201*"hayret" + 0.188*"şaşır"+ 0.162*"şaşkın" + 0.051*"aaa" + …	Şaşkın

Tablo 5. Klasik GDA'nın Türkçe duygu tespiti için başarısı (%) (Success of classical LDA for Turkish emotion detection)

	3	5
Klasik GDA	65,83	60,375

Sistemin başarısını arttırmak için GDA algoritmasına önerilen n-aşamalı yöntem uygulanmıştır. Bu yöntemi uygulamadaki amaç, sözlükteki kelime sayısını her aşamada azaltmaktır. Ayrıca, yöntem ile her aşamada sistemi olumsuz etkileyen kelimeler de silinmektedir. Yöntemdeki n değeri, veri kümesinin boyutuna göre değişiklik gösterebilir. Her n değerine uygulanan işlem, her konunun eşik değeri ile belirlenmektedir. Her konu için eşik değerinden daha fazla ağırlığı sahip olan kelimelerle yeni bir sözlük oluşturulmuştur. Yeni sözlükteki kelime sayısı, klasik GDA'daki sözlüğe göre büyük ölçüde azalmıştır. Sistemi tekrar modellemek için 3 ve 5 sınıfa ait tutarlılık değerleri tekrardan hesaplanmıştır. Her iki sınıf için tutarlılık değerleri ile belirlenen konu sayılarıyla 2-GDA'lı sistemin başarısı ölçülmüştür. Ardından 3-GDA için, daha önce olduğu gibi 2-GDA'daki konulardan faydalanarak yeni sözlük oluşturulmuştur. Böylece yeni sözlükteki kelime sayısı yine azaltılmıştır. 3 ve 5 sınıf için konu sayıları, yeniden hesaplanan tutarlılık değerleri ile belirlenmiştir. Tablo 6'da, 2-GDA ve 3-GDA modelinin başarısı karşılaştırma amaçlı olarak klasik GDA ile beraber gösterilmiştir. Tablo 6'da görüldüğü üzere aşama sayısı arttıkça her iki sınıfta da sistemimizin başarısı artış göstermektedir.

Tablo 6. GDA'nın aşama olarak başarısı (%)
(Success of classical LDA and developed n-stage method)

	3	5
Klasik GDA	65,83	60,375
2-GDA	80,83	70,5
3-GDA	81,5	76,375

Geliştirilen yöntemi diğer konu modelleme algoritmalarıyla kıyaslamak için en yüksek başarıyı veren 3-GDA modeli seçilmiştir. Tutarlılık değeri hesaplanarak 3-GDA yönteminde 3 sınıf için konu sayısı 15, 5 sınıf içinse 20 olarak belirlenmiştir. Tüm konu modelleme algoritmaları, aynı şartlarda değerlendirilmek için 3-GDA yönteminin konu sayısı ile modellenmiştir. Böylece doğru bir karşılaştırma imkanı sunulmuştur. İlk olarak Gizli Anlamsal Analiz (GAA) yöntemi ile sistem belirlenen konu sayıları ile modellenmiştir. Oluşan konularda, kelime ve kelime ağırlıkları yer almaktadır. Ancak, GDA'nın aksine konulardaki kelimelerin ağırlıkları toplamı 1 olmamaktadır. Ağırlıklar ayrıca pozitif ve negatif değerlere sahiptir. GDA örneği için gösterilen Tablo 4'teki gibi kelimelerin ağırlıklarına bakılarak konulara en uygun sınıf etiketi atanmaktadır. Ardından eğitilen sistemin başarısı GAA yöntemi için ölçülmüştür. Tablo 7'de geliştirdiğimiz yöntem ile beraber GAA'nın başarısı gösterilmektedir. Geliştirilen yöntem ile sistemin başarısı GAA'ya yaklaşmasına rağmen, GAA'nın gerisinde kalmıştır.

Tablo 7. Geliştirilen 3-GDA ile GAA modelinin başarısı (%)
(Success of the LSA model with developed 3-LDA)

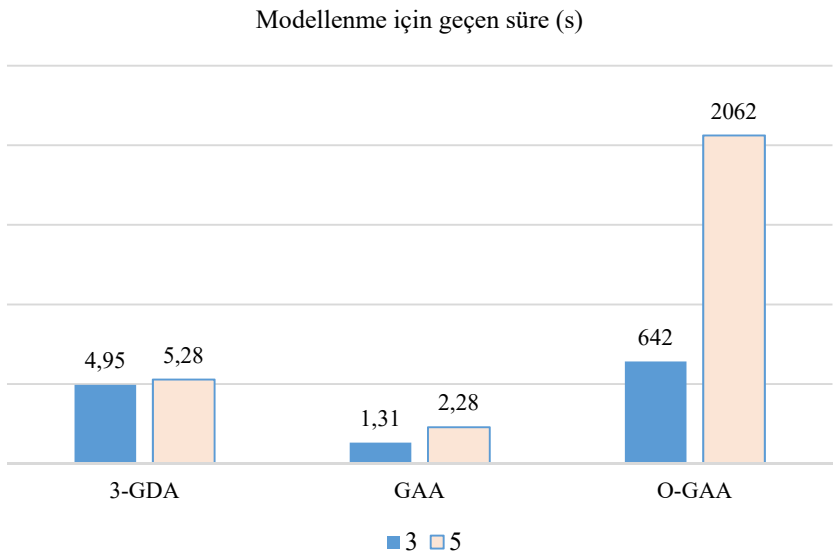
	3	5
3-GDA	81,5	76,375
GAA	95,4	87

Ardından diğer bir konu modelleme algoritması olan Olasılıksal Gizli Anlamsal Analiz (O-GAA) ile sistem modellenmek istenmiştir. Sınıflar için yine aynı konu sayıları kullanılmıştır. Benzer şekilde, kelime ve kelime ağırlıklarına sahip konulara en uygun sınıf etiketi atanmıştır. Son olarak eğitilen sistemin başarısı O-GAA için ölçülmüştür. Tablo 8'de tüm yöntemlerin başarısı gösterilmektedir.

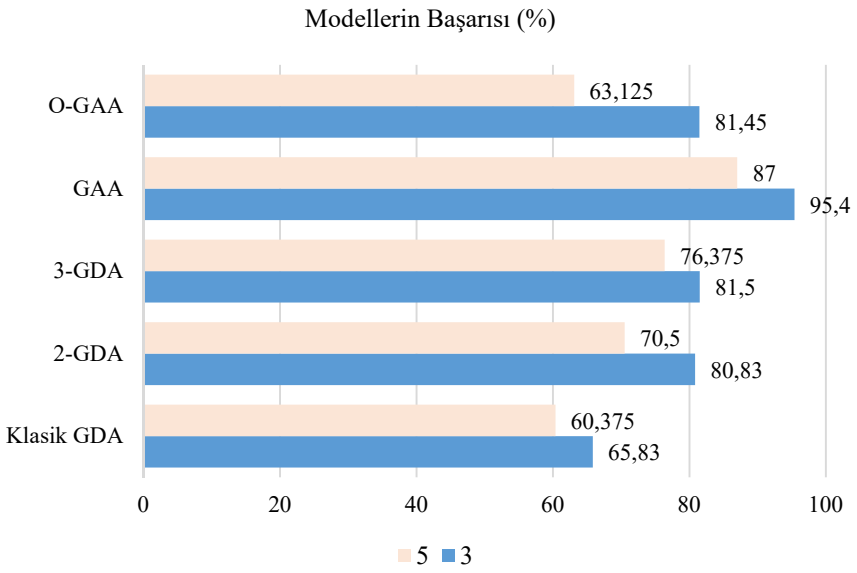
Tablo 8. Tüm yöntemlerin Türkçe tivit veri seti için başarısı (%)
(Success of all methods for Turkish tweet dataset)

	3	5
Klasik GDA	65,83	60,375
3-GDA	81,5	76,375
GAA	95,4	87
O-GAA	81,45	63,125

Geliştirilen 3-GDA yönteminin başarısı O-GAA yönteminden daha iyi sonuç vermiştir. Klasik GDA'da sistemin başarısı tüm yöntemler arasında en düşük olarak gözükürken, geliştirilen yöntem sayesinde O-GAA'nın



Şekil 6. Yöntemlerin modellenmesi için geçen süreler (Running times of all methods)



Şekil 7. Tüm yöntemlerin başarı grafiği (Success graph of all methods)

başarısının üzerine çıkılmıştır. Geliştirilen yöntem ile sistemin başarısı GAA'ya yaklaşmasına rağmen onun seviyesine ulaşamamıştır. Şekil 6'da yöntemler modellenirken geçen süreyi de karşılaştırabilmek için grafik ile gösterilmiştir.

Şekil 6 incelendiğinde, modellenmesi en uzun süren yöntem O-GAA olmuştur. Diğerlerine göre çok uzun sürmesinin nedeni EM fonksiyonunu adım adım kullanması gösterilebilir. Yöntem çok zaman harcamasına rağmen 3-GDA ve GAA'ya göre başarısız olmuştur. GAA yöntemi ise en kısa sürede modellenmiştir. Kısa sürede modellenmesine rağmen sistem en başarılı yöntem olmuştur. 3-GDA'nın modellenmesi ise GAA'ya göre biraz daha fazla zaman almıştır, ancak başarı bakımından GAA'yı geçememiştir.

4. SONUÇLAR (CONCLUSIONS)

Çalışmada, herhangi bir konuda paylaşılmış olan tivitlerin duygularını tespit etmek için konu modelleme algoritmaları kullanılmıştır. Klasik GDA yönteminin başarısını artırmak için geliştirilen n-aşamalı GDA yöntemi de mevcut konu modelleme algoritmalarına dahil edilerek karşılaştırmalar yapılmıştır. Klasik GDA yöntemine göre iki aşamalı GDA'nın sistem başarısı, %10 ile %15 arasında artış göstermiştir. Üç aşamalı GDA yöntemi uygulandığında ise

başarı oranı iki aşamalı GDA 'ya göre %1 ile %6 arasında artmıştır. Başarıdaki artışın en önemli sebebi, sözlükteki kelime sayısının azaltılmasıdır. Bu işlemde, eşik değerinden daha düşük ağırlığa sahip olan kelimeler sözlükten çıkartılmaktadır. Böylece aşama sayısı arttıkça duyguyla ilgili kelimelerin ağırlıkları da artmaktadır. Sonuç olarak konulara kolay şekilde duygu etiketi atanabilmektedir. Geliştirilen yöntemdeki n değeri, veri kümesinin boyutuna göre artırılabilir. Eğer tivitlerde duygusal anlamdan yoksun olan kelimeler fazlaysa, n aşama değerini arttırmak mantıklıdır.

Çalışmanın son bölümünde geliştirilen GDA yönteminin başarısı, diğer konu modelleme algoritmalarıyla karşılaştırılmıştır. Yöntemleri karşılaştırırken deney ortamları için aynı şartlar sağlanmıştır. Aynı veri setini kullanan sistem, her sınıf için aynı konu sayısı ile modellenmiştir. GAA yönteminin başarısı incelendiğinde 3 sınıf için %95,4 iken 5 sınıf için ise %87'dir. Klasik GDA yönteminin başarısı GAA ile karşılaştırıldığında oldukça düşüktür. Geliştirdiğimiz 3-GDA yöntemi ise GAA'nın başarısına yaklaşmasına rağmen GAA'nın başarısına ulaşamamıştır. Bir diğer konu modelleme algoritması olan O-GAA'nın başarısı 3 sınıf için %81,5 iken 5 sınıf için %63,1'dir. Klasik GDA yöntemi, O-GAA ile karşılaştırdığında daha başarısız iken, geliştirilen 3-GDA yöntemiyle modelin başarısı O-GAA algoritmasının başarısını geçmiştir. 3-GDA, O-GAA yöntemine göre 3 sınıf için %0,05 ve 5 sınıf için ise %13,25'lik başarı artışı sağlamıştır. Süre olarak incelendiğinde ise en hızlı ve başarılı yöntem GAA olmuştur. 3-GDA da performans bakımından GAA'dan sonra gelmiştir. Şekil 7'deki grafikte de tüm yöntemlerin başarısı ayrıntılı olarak gösterilmiştir.

Bir sonraki konu modelleme çalışmalarımızda, geliştirilen yöntem ile özetlemenin konu modelleme üzerine etkisini araştırmayı, yazılan metnin kime ait olduğunu bulmayı, şarkıların hangi türe ait olduğunu belirlemeyi, sosyal medyadaki ürünlere yapılan yorumların ürün üzerindeki etkilerini tespit etmeyi ve soru cevap sistemlerinde doğru cevabı ilgili kişiye verebilen sistem tasarlamayı düşünmekteyiz.

KAYNAKLAR (REFERENCES)

1. Lee S., Baker J., Song J., Wetherbe J.C., An Empirical Comparison of Four Text Mining Methods, 2010 43rd Hawaii International Conference on System Sciences, 5-8 Ocak, 2010.
2. Guven Z.A., Diri B., Cakaloglu T., Classification of New Titles by Two Stage Latent Dirichlet Allocation, 2018 Innovations in Intelligent Systems and Applications Conference (ASYU), 4-6 Ekim, 2018.
3. Haidar M.A., Oshaughnessy D., Comparison of a bigram PLSA and a novel context-based PLSA language model for speech recognition, 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, 2013.
4. Mazarura J., Waal A.D., A comparison of the performance of latent Dirichlet allocation and the Dirichlet multinomial mixture model on short text, 2016 Pattern Recognition Association of South Africa and Robotics and Mechatronics International Conference, 2016.
5. Blei D.M., Ng A.Y., Jordan M.I., Latent Dirichlet Allocation, Journal of Machine Learning Research, 3, 993-1022, 2003.
6. Kakkonen T., Myller N., Sutinen E., Timonen J., Comparison of Dimension Reduction Methods for Automated Essay Grading, Journal of Educational Technology & Society, 11 (3), 275–288, 2008.
7. Lu Y., Mei Q., Zhai C., Investigating task performance of probabilistic topic models: an empirical study of PLSA and LDA, Information Retrieval, 14 (2), 178–203, 2010.
8. Chien J.-T., Chueh C.-H., Latent dirichlet language model for speech recognition, 2008 IEEE Spoken Language Technology Workshop, 2008.
9. Chen L.-C., An effective LDA-based time topic model to improve blog search performance, Information Processing & Management, 53 (6), 1299–1319, 2017.
10. Xiong S., Wang K., Ji D., Wang B., A short text sentiment-topic model for product reviews, Neurocomputing, 297, 94–102, 2018.
11. Eddy B.P., Kraft N.A., Gray J., Impact of structural weighting on a latent Dirichlet allocation-based feature location technique, Journal of Software: Evolution and Process, 30 (1), 2017.
12. Bolelli L., Ertekin Ş., Giles C.L., Topic and Trend Detection in Text Collections Using Latent Dirichlet Allocation, Lecture Notes in Computer Science Advances in Information Retrieval, 776–780, 2009.
13. Momtazi S., Unsupervised Latent Dirichlet Allocation for supervised question classification, Information Processing & Management, 54 (3), 380–393, 2018.
14. Guven Z.A., Diri B., Cakaloglu T., Classification of Turkish Tweet emotions by n- stage Latent Dirichlet Allocation, 2018 Electric Electronics, Computer Science, Biomedical Engineerings Meeting (EBBT), 2018.
15. Wikipedia. Latent Dirichlet allocation. http://www.wikizeroo.net/wiki/en/Latent_Dirichlet_allocation. Erişim tarihi Ekim 10, 2018.
16. Slomovitz G., Latent Semantic Analysis (LSA) for syslog correlation, 2017 International Conference on Electronics, Communications and Computers, 2017.
17. Ryan J.O., A System for Computerized Analysis of Verbal Fluency Tests, Yüksek Lisans Tezi, University of Minnesota Twin Cities, Institute for Health Informatics Center for Cognitive Sciences, Minnesota, 2013.
18. Wild F., Stahl C., Investigating Unstructured Texts with Latent Semantic Analysis, Studies in Classification, Data Analysis, and Knowledge Organization Advances in Data Analysis, 383–390, 2007.

19. Hennig L., Topic-based Multi-Document Summarization with Probabilistic Latent Semantic Analysis, International Conference RANLP 2009, 144–149, 2009.

20. Xu J. Topic Modeling with LSA, PSLA, LDA & lda2Vec. Medium. https://medium.com/nanonets/topic-modeling-with-lsa-psla-lda-and-lda2vec555f f65b0b05. Yayın tarihi May 25, 2018. Erişim tarihi Kasım 5, 2018.

21. Hofmann T., Probabilistic latent semantic indexing, Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR 99), 1999.

22. Java2s. Download zemberek-tr-2.1.jar:zemberek « z « Jar File Download. http://www.java2s.com/Code/Jar/z/Downloadzemberektr21jar.htm. Erişim tarihi Eylül 20, 2018.