Avrupa Bilim ve Teknoloji Dergisi
Sayı 16, S. 863-869, Ağustos 2019
© Telif hakkı EJOSAT'a aittir
Araştırma Makalesi

www.ejosat.com ISSN:2148-2683

European Journal of Science and Technology
No. 16, pp. 863-869, August 2019
Copyright © 2019 EJOSAT
Research Article

Sosyal Medya Platformu Üzerinde Gizli Anlam Analizi

Volkan Altintaş[1*], Kamil Topal[2], Mehmet Albayrak[3]

[1] Süleyman Demirel Üniversitesi, Fen Bilimleri Enstitüsü, Bilgisayar Mühendisliği Anabilim Dalı, Isparta, Türkiye (ORCID: 0000-0002-1560-9017)
[2] Balıkesir Üniversitesi, Mühendislik Fakültesi, Bilgisayar Mühendisliği Bölümü, Balıkesir, Türkiye (ORCID: 0000-0002-0266-7365)
[3] Isparta Uygulamalı Bilimler Üniversitesi, Uzaktan Eğitim MYO, Bilgisayar Teknolojileri Bölümü, Isparta, Türkiye (ORCID: 0000-0002-7089-122X)

(İlk Geliş Tarihi 11 Temmuz 2019 ve Kabul Tarihi 9 Ağustos 2019)

(**DOI:** 10.31590/ejosat.590521)

ATIF/REFERENCE: Altıntaş, V., & Topal, K., & Albayrak, M. (2019). Sosyal Medya Platformu Üzerinde Gizli Anlam Analizi. *Avrupa Bilim ve Teknoloji Dergisi*, (16), 863-869.

Öz

Günlük hayatımızın vazgeçilmez bir parçası haline gelen Internet ve sosyal medya alanındaki gelişmeler ile birlikte, bilgisayar ve mobil cihaz kullanıcıların farklı mecralardaki yorumlarında büyük artış yaşanmaktadır. Bu büyük veri miktarında artış nedeniyle, kullanıcı paylaşımlarımda konu başlıklarını ve özelliklerinin doğru ve otomatik olarak çıkarılması önemli bir problem haline gelmiştir. Çeşitli platformlarda paylaşılan kullanıcı metinleri, ilişkisel olmayan ve düzensiz verilerdir. Bu verileri sınıflandırmak, büyük veri işleme ve yapay zekâ çalışma alanlarından biri olan doğal dil işleme için önemli bir konudur. Doğal dil işlemenin kullanım amaçları arasında, ilişkisel olmayan düzensiz metinlerden, anlamlı veriler elde etmek önemli bir çalışma konusudur. Buradan hareketle; iki insanın karşılıklı anlaştığı doğal bir dili anlayıp, cevap verme, özet çıkarma, gibi doğal bir insan zekasının yapabildiğini çok daha hızlı yapabilmek büyük bir önem taşımaktadır. Doğal dil işlemenin alt çalışma alanlarından biri olan konu modelleme, birçok belgenin hangi konuları içerdiğini ve bu konuların önemli özelliklerini ortaya koyar. Günümüzde birçok içerik sağlayıcılar, takipçilerine, anlık içeriklerin önerilmesi işleminde, konu modelleme yapılarını kullanarak, veri akışını doğru kişilere, çok hızlı bir şekilde yönlendirebilirler. Daha önceden etiketlenmiş eğitim setine gerek duymayan Gizli Anlam Analizi (Latent Semantic Indexing - LSI) algoritması bu çalışmada kullanılmıştır. Bu çalışmada, Türkçe kullanıcı girdilerinin yer aldığı Ekşisözlük platformunda, "Apple", "Samsung" ve "Microsoft" başlıklı tartışmalar elde edilerek ve bu tartışmaların alt konu başlıkları "Gizli Anlam Analizi" yöntemi ile modellenmiştir. Toplanan verilerden alt konu başlıkları bulunarak, elde edilen konu başlıkları ile kategoriler karşılaştırılmış, karşılaştırma sonucunda F-Score ile doğruluk oranı ölçülmüştür. Elde edilen F-Score değeri, %74 doğruluk oranı ile bu veri seti ve bu algoritma için sınıflandırma yapıldığını göstermiştir.

Anahtar Kelimeler: Doğal Dil İşleme, Gizli Anlam Analizi, Metin Madenciliği.

Latent Semantic Analysis on Social Media Platform

Abstract

There is a dramatic rise in the number of comments in Internet, which is an indispensable tool for our daily lives. Modelling topics and their features have become more important because of this high volume. Social media users' texts shared in various social media websites are unstructured and not relational data. Clustering this data is one of the most important study area of Natural Language Processing which is a crucial branch of Artificial Intelligence.The purpose of NLP is to get information from unstructured data and react in proper way just like two human being understand each other via a natural language. NLP can do such these jobs faster than natural intelligence. Topic Modelling produces the results of existing topics and their features from large collection of documents. Content providers may distribute the information to the right people fast by using topic modelling tools. In this study, we use Latent Semantic Indexing (LSI) algorithm since it does not need annotated data to train the model. We collect Turkish entries related to "Apple", "Samsung" and "Microsoft" in EkşiSözlük and find the subtopics of the discussions. Sub topics were found and the topics were compared with the categories and F-Score was measured for accuracy. The obtained F-Score showed 74% accuracy rate for this data set and this algorithm.

Keywords: Natural Language Processing, Latent Semantic Analyses, Text Mining.

[*] Sorumlu Yazar: Süleyman Demirel Üniversitesi, Fen Bilimleri Enstitüsü, Bilgisayar Mühendisliği Ana Bilim Dalı, Isparta, Türkiye, ORCID: 0000-0002-1560-9017, volkanaltintas@gmail.com

1. Giriş

Web dünyasının genişlemesi ve farklı alanlarda kullanılmasıyla birlikte, erişilebilen yapısal olmayan veri miktarı da artmıştır. Bilgi geri getirimi ve doğal dil işleme gelişen web teknolojileri sayesinde öne çıkan çalışma alanları haline gelmiştir. Elektronik ortamdaki dokümanlar, kullanıcı geri bildirimleri ve Twitter, Facebook gibi sosyal medya platformlarının sağladığı veriler/yorumlar ile doğal dil işlemeye yeni uygulama alanları katılmıştır.

Son birkaç yılda yapılan çalışmalar ve uygulamalar incelendiğinde; yapısal olmayan veriler/veritabanı, başka bir deyişle her bir bilginin belirli/ilgili alanda yer aldığı ilişkisel veritabanı yapısı yerine, tüm bilgilerin karmaşık ve düzensiz yer aldığı metinlerden istatistiksel ve matematiksel yöntemler kullanarak anlamlı bilgiler çıkartmak amacı ile doğal dil işlemenin sıklıkla kullanıldığı göze çarpmaktadır. Doğal dil işlemenin üzerinde çalıştığı öne çıkan alt alanlar; yazar tanıma, otomatik özet çıkarma, konu modelleme, sınıflandırma ve kümeleme işlemidir [1].

Kelime anlamı temsillerinde sayma tabanlı ve tahmine dayalı yöntemler kullanılmaktadır. Gizli Anlam Analizi (GAA) [2], sayma tabanlı metotlar kelime anlamı temsil etme işleminde çok kullanılan yöntemlerdendir [3][4]. Dokümanlardan oluşan veriler gizli anlam analizine girdi olarak verilir. Her dokümanda bulunan her terim için doküman-terim sıklık matrisi oluşturulur. Daha sonra bu matrise boyut dönüştürme işlemi için "Tekil Değer Ayrıştırma (TDA)" uygulanır. Her bir kelime için vektör temsilleri elde edilir. Anlamsal benzerliğin ortaya çıkarılması için iki kelime vektörü arasındaki açının kosinüsü hesaplanır [5]. Bu çalışmada Ekşisözlük [6] adlı Türkçe içerikli sosyal medya platformundan "Microsoft", "Samsung" ve "Apple" firmaları hakkında konuşulan girdiler alınarak, alt konu başlıkları GAA ile bulunmuştur. Alt konu başlıklarına göre kullanıcı girdilerinin GAA ile sınıflandırma işlemi sonunda doğruluk oranı F-Score yöntemi ile hesaplanmıştır.

Literatürde anlam analizi üzerinde farklı algoritmalar kullanılarak yapılmış konu modelleme ve anlam analizi çalışmaları, ağırlıklı olarak İngilizce olmak üzere ve farklı platformlarda yoğunlaşmaktadır. Elberrichi vd., Liu & Singh, AlSumait vd. çalışmalarında; WordNet ile kelimelerin sözlük anlamları üzerinden eşler oluşturma yöntemi, Latent Dirichlet Allocation (LDA) algoritması, bilgi veritabanı olan ConceptNet ile anlam analizi üzerine çalışmışlardır [7][8][9].

Merchant ve Pande, doğal dil işleme tekniklerini ve GAA algoritmasını kullanarak, uzun metinlerden kısa ve faydalı özetler çıkarmak için dava dosyaları üzerinde çalışma yapmışlardır. Yapılan çalışmada; ceza ve hukuk mahkemeleri türünde dosyaları incelenmiş, oluşturulan model ile ROGUE-1 skoru 0,58 oranına ulaşılmıştır. Yapılan çalışma İngilizce metinler üzerinde uygulanmıştır [10].

Altszyler vd., GAA ve Word2Vec modelini TASA derlemi (korpusu), DreamBank derlemi ve UkWaC derlemi içerisinden oluşturulan küçük dokümanlar üzerinde karşılaştırmışlardır. GAA ile küçük veri setlerinde daha iyi sonuçlar elde edildiğini göstermişlerdir. Kelime sayısı 106'dan fazla olduğunda Word2Vec modelinin benzerlik bulma konusunda daha başarılı olduğu gözlemlenmiştir [11].

Hatipoğlu ve Omurca, Türkçe'nin yapısal özelliklerine göre istatistiksel olarak puanlandırılması ve gizli anlam analizi yöntemlerini sezgisel olarak birleştirerek, cümle seçimi yapan melez bir model sunmuşlardır. Gerçekleştirdikleri çalışma, metin özetleme sorunu üzerine dayanmaktadır. Özet cümlelerin seçimi, özetlenecek metinlerin Türkçe'nin dil özelliklerine dayalı istatistiksel puanlandırılması ve anlamsal puanlandırılması yöntemlerinin melez şekilde değerlendirilmesi ile gerçekleştirilmiştir. Özetlenecek metinlerde yer alan cümlelerin özet cümle adaylığı için aldıkları puanlar, yapısal ve anlamsal özelliklerin sezgisel bir ağırlıklandırma yöntemi ile birleştirilmesi ile belirlenmiştir. Çalışma kapsamında veri ön işleme, yapısal olarak istatistiksel puanlandırma, GAA analiz ve melez cümle seçimi aşamaları Türkçe yazılmış metinler üzerinde başarıyla gerçekleştirilmiştir. Elde edilen sonuçların değerlendirilmesi için özetleme sistemi geliştirilmiştir. Geliştirilen sistem ve farklı kullanıcılara, aynı metinler verilmiş ve kullanıcıların önerileri karşılaştırılmıştır. Bu karşılaştırma sonucunda "Güneş Sistemi" metninin özeti ile kullanıcıların bu metinden seçtiği cümleler %77.5, "Charles Bukowski" metninin özeti ile kullanıcıların bu metinden seçtiği cümleler % 82 oranında eşleşmiştir [12].

Kherwa ve Bansal, TDA tabanlı GAA üzerine bir çalışma yapmışlardır. Çeşitli doğal dil işleme uygulamalarının araştırma makalelerinden oluşan bir veri setinde, terimlerin birbirleri ile ilişkilerini bulmak için GAA yöntemini kullanmışlardır. Çalışma sonunda GAA'nın TDA ile aynı anlamlı birden fazla terimi azalttığını, birden çok anlamı olan terimleri tanımlayabileceğini ve düşük boyutlu kavramsal alandaki belgeleri temsil ettiğini göstermektedir [13].

Yıldıztepe ve Uzun yaptıkları çalışmada, olasılıksal gizli anlam analizi ve gizli Dirichlet ataması yöntemleri üzerine çalışmışlardır. Farklı haber ajanslarında bulunan Türkçe haber metinlerinin anlamsal benzerliklerine göre kümeleme uygulaması oluşturulmuş ve uygulamadan elde edilen sonuçlar incelenmiştir. Elde edilen sonuçlara göre iki yöntemle de aynı konudan bahseden haber metinleri başarılı bir şekilde sınıflandırılmış ve anlamsal olarak yakın haberler belirlenmiştir [14].

Ünaldı ve Kırkgöz yaptıkları çalışmada, anadili İngilizce olan üniversite öğrencileri ile anadili Türkçe olan üniversite öğrencileri tarafından oluşturulan metinleri GAA algoritması kullanarak karşılaştırmışlardır. Karşılaştırmanın yapılabilmesi için anadili Türkçe olan öğrencilerin, İngilizce olarak yazdığı metinlerden bir derlem oluşturulmuş ve bu derlem anadili İngilizce olan üniversite öğrencileri tarafından yazılmış metinleri içeren başka bir derlemle karşılaştırılmıştır. Tümce, paragraf ve metin geneli olmak üzere 3 farklı terim kullanılmıştır [15].

Yapılan çalışmanın, EkşiSözlük platformu üzerinde kullanıcılar tarafından yapılan girdiler üzerinde 3 farklı konudaki alt başlıkları bulması, Türkçe dilinde olması, kullanıcı girdilerinde herhangi bir boyut kısıtlaması olmamasından dolayı literatürde yapılan çalışmalara göre farklılık göstermektedir.

2. Materyal ve Metot

Bu bölümde GAA, TDA yöntemleri ve kullanımları, veri setinin hazırlanması, veri seti üzerinde doğal dil işleme aşamaları ve kelime vektörlerinin çıkarılması işlemleri anlatılmıştır.

2.1. Gizli Anlam Analizi

Konu modelleme; verilen dokümanlardan alt konuları otomatik olarak bulmak için kullanılan bir istatiksel makine öğrenmesidir. Bu yöntemle alt konuların önemli özellikleri ve her bir dokümanın hangi alt konuya ait olduğunu bulunabilir. Verilen belgedeki anahtar sözcük grubunu bulmak için kullanılan, denetimsiz öğrenen bir metin analizidir. İşlem sonucu ortaya çıkan kelime grubu (özellikler), alt konuyu temsil etmektedir. Denetimsiz bir öğrenme şekli olduğu için, bulunan konuların bir uzman tarafından değerlendirilmesi gerekebilir. Ayrıca çoğu zaman kaç farklı alt konunun bulunacağı, önceden bilinmesi gerekmektedir.

Konu modellemede kullanılan başlıca modellerden birisi de Gizli Anlam Analizidir. GAA; anlaşılması ve uygulanması kolay olan bir yöntemdir. Diğer metotlara göre daha hızlıdır. Çünkü sadece doküman terim matrislerine göre işlem yapmaktadır.

Gizli Anlam Analizi için bir doküman-terim matrisine ihtiyaç vardır. Bu matrisin değerleri genel olarak Terim (Kelime) Sıklığı - Ters Dokuman Sıklığı (TF-IDF) ağırlıkları ile oluşturulur. TF-IDF, her bir dokümanın içinde yer alan kelimelere birer ağırlık oluşturur. Bu ağırlıklar, kelimelerin o doküman için ne kadar sık geçtiğine ve o kelimenin diğer dokümanlarda ne kadar geçip geçmediğine bakılarak hesaplanır. Bunun için önce Terim Sıklığı (TF) hesaplanır. Bu işlem her bir kelimenin bir doküman içinde kaç kere geçtiğini hesaplar. Daha sonra Ters Doküman Sıklığı (IDF) aşağıdaki formülle hesaplanır:

$$IDF(t) = \log(\frac{N}{|\{d \in D : t \in d\}|}) \quad (1)$$

t terim (kelime), N doküman sayısı, D tüm dokuman seti, d tek bir dokumanı temsil eder. $|\{d \in D : t \in d\}|$ ifadesi t teriminin, tüm dokümanlarda, kaçının içinde yer aldığını bulur. Eğer bir terim, çok sayıda dokümanda geçiyorsa, payda büyüyecek ve logaritmik ölçekte IDF değeri küçülecektir. Ya da bir kelime az sayıda dokümanda geçiyorsa, o kelime ilgili doküman için ayırt edici ve önemli bir kelime olur. IDF tüm kelimelerin doküman zıtlığıdır. Son olarak, TF ve IDF ağırlıkları çarpılarak, TF-IDF ağırlıkları bulunup TF-IDF matrisi oluşturulur. Doküman Sıklığı ve Ters doküman sıklığı özelliklerinin çarpımıyla elde edilen TF-IDF matrisinin her bir satırı bir dokümanı, her bir sütunu ise kelimeleri temsil eder.

TF-IDF matrisinden yararlanılarak, Tekil Değer Ayrıştırma (Singular Value Decomposition) ile işlemi yapılabilir. Bu ayrıştırma işlemi ile satırlardaki dokümanlar ve sütunlardaki kelimelerin gruplandırılması hedeflenir. Bu gruplandırma işlemi yapılırken TDA'nın aşağıdaki formülü ile elde edilen matrisler yorumlanır:

$$A = U\Sigma V^T \quad (2)$$

A TF-IDF ağırlıklarının olduğu $m \times n$ boyutundaki orijinal matristir. m döküman sayısı, n ise tüm dokümanlardan elde edilen sözcük sayısıdır. U, $m \times m$ boyutunda dik açılı (ortogonal) sol tekil değer matrisidir. Bu matriste dokümanlar ile ilgili ağırlıklar yer almaktadır. Σ matrisi $m \times n$ boyutunda köşegen bir matristir. Köşegende A matrisinin özdeğerleri (eigen values) büyükten küçüğe doğru yer alır. V^T ise $n \times n$ boyutunda dik açılı sağ tekil değer matrisidir. Bu matriste de terimler ile ilgili ağırlıklar yer alacaktır.

Σ matrisinin köşegeninde $\sigma_{11} > \sigma_{22} > \cdots > \sigma_{mm}$ değerleri yer almaktadır. Belirlenecek bir k sayısı ile bu köşegenin ilk k değeri alınır. Bu değer, kaç farklı konu gösterilmek istendiğidir. Eğer k değerinin ne olacağı bilinmiyorsa, sıralı özdeğerler arasındaki en büyük boşluğa sahip yer k olarak seçilir. Σ matrisinin yeni boyutu $k \times k$ olacaktır. Dolayısıyla U matrisinin ilk k sütununu $m \times k$ ve V^T matrisinin ilk k satırını $k \times n$ seçilmesi gerekir. Bu üç matrisin çarpımı orijinal A matrisine yakınsayacaktır.

Şimdi, U matrisinin ilk sütunu, ilk konunun doküman ağırlıklarını vermektedir. Yani, ilk sütundaki en yüksek değerler, ilk konunun ağırlığı en yüksek dokümanı olacaktır. Aynı şekilde V matrisinin ilk sütunundaki değerler, ilk konunun terim ağırlıklarını gösterecektir. Ağırlıkları yüksek olan kelimeler, o konunun açıklayıcı kelimeleri olacaktır. Bu işleme k. konuya kadar devam edilir. TDA yapıldıktan sonra U ve V matrislerinde negatif değerler yer alacaktır. Ancak A matrisi tamamen pozitif değerlerden oluşmaktadır. U, Σ ve V^Tmatrislerinin çarpımı A'yı vereceği için ve Σ'da negatif değer yer almadığı için U'da negatif bir değer varsa, V^T'de negatif olmak zorundadır. Dolayısıyla negatif değerlerde olsa bile, mutlak değerlerinin alınması sonucu pozitife dönüşmesi U ve V^T analizi için göz önünde bulundurulmalıdır.

Bu çalışmada Python dilinde "gensim" kütüphanesinde bulunan "lsimodel" modülü kullanılmıştır [16].

2.2. Veri Setinin Hazırlanması

Bu bölümde veri setinin hazırlanması ve veri seti üzerinde yapılan işlemler anlatılmıştır. EkşiSözlük (EksiSozluk), her türlü konu ve kavram hakkında, kayıtlı yazarların yorumlarını içeren katılımcı sözlük tarzında bir platform ağ olup, web sitesi Türkiye'deki katılımcı sözlükler arasında en fazla tanımlama (girdi/entry) yapılan sitedir. Kayıtlı yazarlar tarafından yapılan girdiler, paylaşılan bilgiler, yöneticiler ve "gammaz" adı verilen gönüllü kullanıcılar tarafından denetlenmekte uygun olmayanlar silinmektedir. Platformda kayıtlı olan tüm yazarlar gammaz özelliğine sahiptir. Yazar alımı sürekli yapılmamaktadır. Kısa süreli başvurular ile yazar alınmaktadır. Her yazar alınma dönemine "nesil" denilmektedir [17].

Bu çalışmada, Ekşisözlük platformunda kullanıcıların Apple, Samsung ve Microsoft firmaları için yaptığı girdiler incelenmiş, ilgili alanda yer alan girdiler analiz edilerek, kullanıcıların bu firmalar üzerinde en çok hangi konu hakkında konuştukları belirlenmiştir. Kullanılan veri setinin özellikleri şunlardır:

- Ekşisözlük platformunda kullanıcıların Apple, Samsung ve Microsoft firmaları üzerine yaptığı yorumlar yer almaktadır.
- Platformda yapılan yorumlar Türkçe olarak paylaşılmıştır.
- Çalışmada seçilen konu kapsamı özel bir alan olduğu için, konu ile alakasız paylaşım sayısı azdır.
- Literatürde Twitter vb. sosyal platformlar üzerine yapılan çalışmalarda karşılaşılan kelimelerin kısaltılması, değiştirilmesi, emoji kullanımı gibi metin analizini zor hale getiren bir durum veri setinde gözlemlenmemiştir.
- Ekşisözlük platformunda yapılan paylaşımlarda platform kuralları gereği büyük harf olmadığından bütün girdiler küçük harftir. Veri setinde harfleri küçültme ile ilgili bu yüzden herhangi bir işlem yapılmamıştır.

Veri seti hazırlanırken EkşiSözlük platformundan toplanan veriler Apple, Samsung ve Microsoft firmaları hakkında yapılan girdilerdir [18][19][20]. Verileri platformdan almak için Python programlama dilinde request kütüphanesi ile yazılan web crawler hazırlanmıştır [21]. Hazırlanan crawler ile yapılan yorumlar ve yorumun yapılış tarihi toplanarak, veritabanında saklanmıştır. Bu konu başlıkları ile ilgili Apple firması için 357 web sayfasından toplam 3568 adet girdi, Samsung firması için 141 web sayfasından 1410 adet girdi, Microsoft firması için 100 web sayfasından 997 adet girdi bulunmaktadır. Veri setinde üç firma için farklı sayıda girdi bulunmaktadır. Modelleme esnasında sistemin doğruluk oranı, modelin aynı oranda veri ile oluşturulması adına üç firma için ilk 800 girdi veri setinde kullanılmıştır.

2.2. Veri Seti Üzerinde Doğal Dil İşleme Adımları

- Elde edilen yorumlar satırlar halinde tutulmaktadır. Satırlar bölümlere (token) ayrılmıştır. Her bir bölüm bir kelimeyi ifade etmektedir.
- Veri seti oluşturulurken bağlaçlar, zamirler gibi anlamsal değeri olmayan ve metin analizinde kullanılmayacak sözcükler elde edilen verilerden çıkarılmıştır. Türkçe dilinde en çok kullanılan yaklaşık 250 adet Türkçe durak kelimesi (stop words) listesi oluşturulmuş ve bu listedeki kelimelere veri setinden çıkarılmıştır.
- Kelimeler arka arkaya eklenmiş, kelimelerden oluşan derlem (corpus) oluşturulmuştur.

2.3. Kelime Vektörlerinin Çıkartılması

Doğal dil işleme çalışmalarında, kelimelerin semantik anlamları için önemli bir konu olan kelimelerin vektör olarak temsil edilmesi kullanılmaktadır. Kelimeler sayısal değer içeren vektörlerle eleştirilmektedir. Bu eşleşme sayesinde sistem, kullanılan kelime hakkında bilgi (sayısal değer) sahibi olmaktadır.

Bu çalışmada, doküman terimleri matrisi elde edilmiştir. Doküman terimleri matrisinde belgede bulunan tüm terimler bulunmaktadır. Doküman matrisi iki boyutlu bir matristir. İlk boyutunda terimler, diğer boyutunda ise her terimin belgede geçme sıklığını göstermektedir.

3. Araştırma Bulguları

Bu bölümde yapılan işlemlerden elde edilen deneysel sonuçlar yer almaktadır. Veri setinde bulunan her firma için 800'er adet olmak üzere toplamda 2400 adet girdi bulunmaktadır. Gerçekleştirilen veri temizleme işlemi sonucunda konu ile ilgisi olmadığı tespit edilen girdiler elenerek, Apple konu başlığında 739, Microsoft konu başlığında 732, Samsung konu başlığında ise 772 girdi kaldığı görülmüştür. Veriler ilgili firmaların EkşiSözlük'te bulunan kullanıcılar tarafından oluşturulan sayfalarından elde edildiği ve ilgili firma ile ilgili yorumlarını barındırması nedeniyle firma için etiketlenmiş olarak kabul edilmektedir. Etiketlenmiş olarak bulunan veriler ile GAA yöntemi ile üç başlığa ayrılan konu başlıkları karşılaştırılmıştır. EkşiSözlük kullanıcılarının girdilerinden oluşan veri setinin özellikleri Tablo-1'de gösterilmektedir.

Tablo 1. Kullanılan Veri Seti

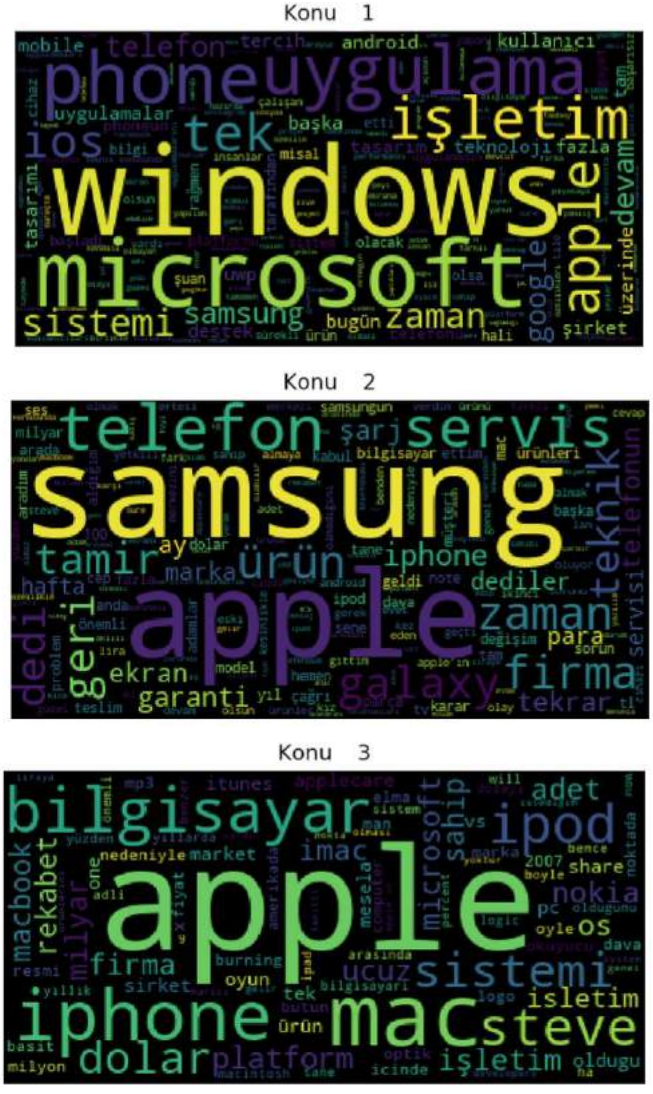
Adı	Toplam Örnek Sayısı	Alınma Tarihi	Konu
EkşiSözlük	2400	17.06.2019	Apple, Samsung, Microsoft

Veri setinde her bir konu için örnek girdiler Tablo-2'de gösterilmektedir.

Tablo 2. Konuların Örnek Girdileri

Konu	**Girdi**
Apple	steven wozniak ve steven jobs'un yarattıklari ve 1976'da apple computer'i kurdulari bilgisayar devi, diger rakip firmalar hala ipod'un karizmasina ulasamadi ve powerbook g5'in gucune yetisemediler. adamlar ustalar hoca, onlar yapsin biz alalim.
Samsung	şimdiye kadar cep telefonunda vazgeçmediğim marka fakat fotoğraf çekmek için yan tarafa tuş koymamakta ısrar ederse her türlü vazgeçeceğim marka. varsa takip eden yetkili ağızlar, bilginiz olsun bir kullanıcı olarak rahatsızız ve sadece ben değilim bu konuda şikayetçi olan.
Microsoft	1995'te yazilim dunyasinda bir devrim yapmiş, beta testini yaptirdigi 'u parayla satmiş biri. yazilim dunyasinin "daha iyisini yapamiyorsan satin al" devi.

Yapılan çalışma sonucu elde edilen konular, GAA yöntemi kullanılarak, belirlenen konu Şekil-1'de etiket bulutu olarak gösterilmektedir. Veritabanında bulunan kullanıcı girdileri içerisinden başlıca bahsedilen 3 adet konu belirlenmiştir. Belirlenen konu başlıkları incelendiğinde kullanıcıların ilgili başlık olan Microsoft, Samsung ve Apple firmaları hakkında konuştukları gözlemlenmektedir. Konu1'de öne çıkan kelimeler Microsoft, Windows, işletim, sistemi gibi konu geneli ile ilgili kelimelerdir. Konu2'de Samsung, telefon, Galaxy, servis gibi konu ile ilgili fakat Apple kelimesi firma farklı olsa da ağırlıklı olarak firma ile ilgili olan kelimelerden dolayı konu başlığı ile uyuşmaktadır. Konu3'de öne çıkan kelimeler Apple, Iphone, Mac, Dolar, Ipod, bilgisayar kelimeleri Apple firması ile ilgili konuya karşılık gelmektedir.



Şekil-1 GAA Yöntemi ile Elde Edilen Konular

GAA yöntemi kullanılarak elde edilen sınıflandırmanın başarım değerlendirmesi için gerçek ve tahmin edilen sınıfı içeren karışıklık matrisi kullanılmıştır. Üç sınıftan oluşan bir sınıflandırma örneği için karışıklık matrisi örneği Tablo-3'de gösterilmektedir.

Tablo-3 Üç Sınıflı Eğitim İçin Karışıklık Matrisi

Tahmin Edilen	**Gerçek Değer**				
		1	2	3	Toplam
	1	D11	Y12	Y13	TT1
	2	Y21	D22	Y23	TT2
	3	Y31	Y32	D33	TT3
	Toplam	GT1	GT2	GT3	

Karışıklık matrisi yardımıyla F1 ölçütü, kesinlik hassasiyet değerleri ile başarım hesabı yapılmıştır. Kesinlik getirilen bilgideki doğru sonuçların, getirilen bilginin tamamına oranı olarak hesaplanır. Hassasiyet ise getirilen doğru sonuçların, getirilmesi gereken doğru sonuçlara oranı ile hesaplanır. Kesinlik ve hassasiyet değerlerinin hesaplanma şekilleri aşağıda gösterilmektedir:

$$Kesinlik = \frac{ilgili\ veri\ getirimi\ \cap bütün\ veri\ çıkarımı}{bütün\ veri\ çıkarımı} \quad (3)$$

$$Hassasiyet = \frac{ilgili\ veri\ getirimi\ \cap bütün\ veri\ çıkarımı}{ilgili\ veri\ çıkarımı} \quad (4)$$

F1 ölçütü bu değerlerin harmonik ortalamasıdır:

$$F1 = \frac{2*Kesinlik*Hassasiyet}{Kesinlik+Hassasiyet} \quad (5)$$

F-Score değeri 0 ile 1 arasındadır. 1'e yaklaştığı zaman doğruluk oranının artmaktadır. 0'a yaklaştıkça kurulan modelin doğruluk oranı azalmaktadır.

Yapılan analizdeki tüm doğru sayısı 1674, bütün sınıflandırılan veri sayısı 2243'tür. Toplam doğruluk değeri 0.74 olarak hesaplanmıştır.

Tablo-4'de veri kümesinin GAA yöntemi ile sınıflandırma sonuçlarına göre oluşan karışıklık matrisi verilmiştir.

- Tablo-4'den Samsung firmasına yapılan girdilerin, GAA yöntemiyle en fazla oranda doğru sınıflandırıldığı görülmektedir. Bununla beraber Samsung ile yapılan yorumların kesinliği diğer iki konudan düşüktür. Tip 1 hatanın (Yanlış sınıflandırma) en fazla olduğu konu Samsung'dur. Başka bir deyişle, sistem Samsung ile ilgili yazılan girdileri yüksek bir doğrulukla bulabilirken, yine Samsung olarak etiketlediği girdilerin büyük bir kısmını yanlış olarak da etiketlemektedir.
- Kesinlik değeri en yüksek olan Apple firması ile ilgili bilgilerin en düşük doğruluk ile bulunduğu görülmektedir. Bununla birlikte Apple, Tip 2 hatanın (ilgili girdileri kaçırması) en çok görüldüğü başlıktır. Sistemin Apple diye etiketlediği verilerin %85'e yakın oranda doğru bulduğu, ancak tüm Apple girdilerinin sadece %54'ünü yakalayabildiğini söyleyebiliriz.

Her bir konunun F1 skorlarına bakıldığında Microsoft 0.79; Samsung 0.78 ve Apple 0.66 olarak görülmektedir.

Tablo-4 Sınıflandırma Karışıklık Matrisi

Tahmin Edilen	**Gerçek Değer**				
		Microsoft	Samsung	Apple	**Toplam**
	Microsoft	607	75	113	**795**
	Samsung	80	669	228	**977**
	Apple	45	28	398	**471**
	Toplam	**732**	**772**	**739**	**2243**

4. Sonuç

Bu çalışmada, bir doğal dil işleme uygulaması olan konu belirleme problemi üzerinde çalışılmıştır. Ekşisözlük platformundan elde edilen kullanıcı girdileri üzerinde veri temizleme ve diğer veri ön işleme adımları sonucunda, terim doküman matrisinin oluşturulması ve gizli anlam analizi işlemleri yapılarak, konu belirleme işlemi kullanıcılar tarafından Türkçe olarak yapılan girdiler üzerinde uygulanmıştır. Elde edilen konu başlıkları ile veri toplama esnasında sınıflandırılmış olan başlıklar karşılaştırılmıştır. Modelin başarım değerlendirmesi için F1 değeri hesaplanmıştır. Çalışma sonucu belirlenen konu başlıkları, kurulan modelin sonucundan da görüldüğü gibi, çalışılan başlık ile uyumlu olduğu ortaya çıkmıştır. Belirlenen üç konu başlığı içerisinde öne çıkanlar incelendiğinde, başlıkların konu başlıkları ile doğrudan ilintili olduğu görülmüştür. Ayrıca denetimsiz olarak öğrenen GAA, belirli amaçlarda görece başarılı istatistiksel sonuçlarla çalışmaktadır.

Çalışmanın Türkçe dilinde ve Türkçe bir platform üzerinde toplanan veriler ile yapılması konun özgün değerini ortaya koymaktadır. Çalışmanın bir başka özgün boyutu da Ekşisözlük platformunda kullanıcıların yaptıkları girdilerde paylaşımlarını diledikleri uzunlukta yapabildikleri için, kullanıcı girdilerinin boyutları sınırlandırılmamış olmasıdır.

Çalışmanın devamında, farklı konu belirleme algoritmaları ile yapılacak analizler karşılaştırılarak, aradaki farklılıklar incelenebilir. Ayrıca, derin öğrenme algoritmaları ile GAA başarıları karşılaştırılması hedeflenmektedir.

Kaynakça

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

[2] Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Harshman, R., "Indexing by Latent Semantic Analysis". Journal of the American Society for Information Science, 41(6):391–407, 1990.

[3] Harris, Z., "Distributional Structure", Word, 23(10), 146–162, 1954.

[4] Landauer, T. K., Dumais, S. T., "A Solution to Plato's Problem: The Latent Semantic Analysis Theory of Acquisition, Induction, and Representation of Knowledge", Psychological Review, 104(2), 211–240, 1997.

[5] X., Cai, Z., Wiemer-Hastings, P., Graesser, A., McNamara, D., Strengths, "Limitations, and Extensions of LSA". Handbook of Latent Semantic Analysis, 401–426, 2007.

[6] Ekşi Sözlük, 1999. https://eksisozluk.com/

[7] Elberrichi, Z. Rahmoun, A. and Bentaallah, M. A., "Using WordNet for Text Categorization", The International Arab Journal of Information Technology, s. 16- 24, 2008.

[8] Liu, H. and Singh, P., "ConceptNet-A Practical Commonsense Reasoning ToolKit", BT Technology Journal, s. 211-226, 2004.

[9] AlSumait, L. Barbará, D. Gentle, J. and Domeniconi, C., Topic Significance Ranking of LDA Generative Models, Machine Learning and Knowledge Discovery in Databases, s. 67-82, 2009.

[10] Merchant, K., Pande, Y., 2018. "NLP Based Latent Semantic Analysis for Legal Text Summarization", International Conference on Advances in Computing, Communications and Informatics (ICACCI), 2018.

[11] Altszyler, E., Sigman, M., Ribeiro, S., D. F. Slezak, D. F., "Comparative Study of LSA vs Word2vec Embeddings in Small Corpora: A Case Study in Dreams Database", arXiv: 1610.01520, 2016.

[12] Hatipoğlu, A., Omurca, S., "Türkçe Metin Özetlemede Melez Modelleme". Dokuz Eylül Üniversitesi Mühendislik Fakültesi Fen ve Mühendislik Dergisi 17: 95-108, 2015.

[13] Kherwa, P., Bansal, P., "Latent Semantic Analysis: An Approach to Understand Semantic of Text", International Conference on Current Trends in Computer, Electrical, Electronics and Communication (ICCTCEEC-2017): 870-874, 2017.

[14] Yıldıztepe, E, Uzun, V., "Olasılıksal Yöntemler ile Türkçe Metinlerin Anlamsal Benzerliğinin Belirlenmesi". Sinop Üniversitesi Fen Bilimleri Dergisi, 3 (2), 66-78, 2018.

[15] Ünaldı, İ., Kırıkgöz, Y., "Latent Semantic Analysis: An Analytıcal Tool for Second Language Writing Assessment". Mustafa Kemal University Journal of Social Sciences Institute, Volume: 8, Issue: 16, s. 487-498, 2011.

[16] Gensim, 2009. http://radimrehurek.com/gensim/ models/lsimodel.html

[17] WikiPedia, 2001. https://tr.wikipedia.org/wiki/ Ekşi_ Sozlük

[18] EkşiSözlük Apple, 1999. https://eksisozluk.com/apple--55201

[19] EkşiSözlük Samsung, 1999. https://eksisozluk.com/samsung--90291

[20] EkşiSözlük Microsoft, 1999. https://eksisozluk.com/microsoft--31834

[21] Python Request, 2001. https://pypi.org/project/requests/