YÖNETİM BİLİŞİM SİSTEMLERİ DERGİSİ
http://dergipark.ulakbim.gov.tr/ybs

Yayın Geliş Tarihi: .05.06.2016
Yayına Kabul Tarihi: 31.07.2016
Online Yayın Tarihi: 01.08.2016

Cilt:2, Sayı:3, Yıl:2016,Sayfa:29-50
ISSN: 2148-3752

METİN MADENCİLİĞİNDE ANAHTAR KELİME SEÇİMİ BİR ÜNİVERSİTE ÖRNEĞİ

Osman YILDIZ

Özet

Kurumların ileri düzey bilişim sistemlerine sahip olması ne kadar önemliyse, bu sistemlerdeki verilerin kurum yöneticilerine işlevsel, ayırt edici ve anlamlı bir şekilde sunulması da o kadar önemlidir. Bu bağlamda büyük metinlerin içerisinde yöneticiler için anahtar kelime çıkarımı oldukça önem arz etmektedir. Bu çalışmada, halen bir üniversitede kullanılan kurum ile ilgili şikayet, teşekkür, görüş ve öneri mesajlarının yazılabildiği ve bu mesajlara ilgili kurum tarafından cevap verilebildiği bir bilişim sistemine ait veriler kullanılmıştır. Veri setindeki yaklaşık 3961 mesaj, metin madenciliği teknikleri kullanılarak ön işlemden geçirilmiştir. Ön işlem sonrası elde edilen metinlerin içindeki önemli kelimeleri tespit etmek için tf-idf ve ki-kare istatistik algoritması kullanılarak anahtar kelime seçimi yapılmıştır.

Anahtar Kelimeler: Anahtar Kelime Seçimi, Metin Madenciliği, Ki-Kare, Bilgi Kazancı, TF-IDF, Yönetim Bilişim

KEYWORD SELECTION IN TEXT MINING: A UNIVERSITY SAMPLE

Abstract

It is very important for institution to have advanced level information systems. What is more, it is also important that this data in these systems served to corporate executives in functional, distinctive, and meaningful way. In this regard, keyword extraction from large texts is of great importance for administrators. In this study, data regarding a still active information system in which complaining about the institution, thanks, comments, and suggestions can be written and answered by the relevant institution has been used. Almost 3961 messages in data set have been preprocessed using text mining techniques. In order to identify the important words in the text obtained after pre-process, keyword selection has been made using tf-idf and chi-square statistics algorithm.

Keywords: Keyword Extraction, Text Mining, Chi-Square, Information Gain, TF-IDF, Management Information

GİRİŞ

Her geçen gün internet tabanlı teknolojiler hayatımızda daha fazla yer etmektedir. Aynı zamanda bu teknolojiler kullanılarak ciddi boyutlarda iş yükünden tasarruf edilmektedir. Bu bağlamda, internet teknolojileri her alanda iş yükünü azaltması açısından son derece kullanışlıdır. Örneğin web üzerinde oluşturulan bir şikayet formu ilgili kurumun hem paydaşları hem de çalışanları adına son derece yararlıdır. Bu açıdan bakıldığında kullanıcı şikayetini bir kağıda yazıp ilgili birime götürüp teslim etmek zorunda kalmaz. İlgili birimde bu formları teslim alan ve bu formları ilgili kişilere dağıtacak bir sorumlu olmasına da gerek kalmaz. Bu birimlerin çoğaltıldığı düşünüldüğünde basit bir formun elektronik ortama taşınması işi bile ilgili kurum için verimliliği artırmaktadır. Fakat birçok kurumda, gelen formlarla ilgili süreç, formun cevaplandırılması ile son bulmaktadır. Orta düzey yönetici kendi birimi ile ilgili kalite kapsamında sadece rakamsal sonuçlarla ilgilenmektedir. Üst düzey yöneticiye de bu rakamlardan oluşan raporlar sunulmaktadır. Bilgi çağında artık sadece teknolojileri basit anlamda kullanmak yeterli olmamaktadır. Vermiş olduğumuz bu örnekte bir kurum ya da iş yerinin paydaşlarını dinlemesi, anlaması ve bu anlayış üzerine yeni adımlar atması kaçınılmazdır. Bu sebeple yöneticilerin daha kullanışlı enformasyon raporlarına sahip olması gerekmektedir.

Bu çalışmada kurumun iç paydaşlarını dinleyip anlaması formların otomatik olarak bilgisayar tarafından yapısal metinlere dönüştürülmesi ile gerçekleşebilir. Bunun için metin madenciliği kullanılmaktadır. Metin madenciliği; enformasyon bombardımanı sorununu makine öğrenmesi, veri madenciliği, doğal dil işleme, bilgi yönetimi ve enformasyon elde etme gibi birçok tekniğin birleşimiyle çözmeye çalışan yeni bir bilgisayar bilimi araştırma alanıdır. Bununla birlikte metin madenciliğinin amacının sayısı milyonları bulan dokümanlar arasından enformasyonların ve desenlerin keşfedilmesi olarak ifade edilmektedir (Feldman & Sanger, 2007).

Metin madenciliğini başka bir tanımlamada; bazı amaçlara dönük metinlerden enformasyon çıkarmak için metinlerin analiz edilme süreci olarak tanımlanmaktadır (Visa, 2001). Metin madenciliği, düzensiz bir şekilde bulunan metin, resim gibi kaynaklarda bazı yöntemler kullanılarak sınıflama ve benzerlik işlemlerinin yapıldığı bir uygulama alanı olarak da tanımlamaktadır (Döven, 2013). Bununla birlikte metin madenciliği, veri madenciliğin bir alt türü olarak tanımlanarak metinsel verilerin metin madenciliği teknikleri ile yapısal anlamlı hale getirebileceğini belirtilmektedir (Çelikyay, 2010).

LİTERATÜR İNCELEMESİ

Türkiye'de metin madenciliği ile ilgili tez düzeyinde yapılan birçok çalışma mevcuttur. Bu çalışmaların bazılarında anahtar kelime seçimine yönelik bir takım metotlar kullanılmıştır.

Türkçe metinleri sınıflandırma amacıyla yapılan bir çalışmada, metin madenciliğinin Naive Bayes ve K-NN algoritmalarını kullanılmıştır. Çalışma sonuçlarına göre; Naive Bayes algoritması genel olarak K-NN algoritmasına göre daha iyi sonuçlar vermiştir. Bunun yanında doğruluk oranın anlamında K-NN algoritmasına ait TF-IDF ağırlıklandırma yönteminin K-NN algoritmasına ait diğer ağırlıklandırma yöntemlerine göre daha yüksek değere sahip olduğu belirlenmiştir (Pilavcılar, 2007).

Haber sitelerindeki köşe yazılarının sınıflandırılması ile ilgili bir çalışmada metin madenciliği yöntemleri kullanılmıştır. Bu çalışmada, metin madenciliğindeki k-NN ve Naive Bayes algoritmaları, Bit, TF, IDF ve TF-IDF ağırlıklandırma ve 37 sınıf özellik vektörü kullanılmıştır. Çalışma neticesinde; TF-IDF ağırlıklandırma ile cosine, k=7 ve Multi-Nominal birlikte kullanıldığında sınıflamada %100 başarılı olunduğu belirlenmiştir (Karaca, 2012).

Bir başka çalışmada ise metin madenciliği doküman demetlemede kullanılmıştır. Çalışmanın amacı, İngilizce ve Türkçe metinlere ait verileri bazı gruplara ayırmaktır. Bu amaç doğrultusunda, Terim Frekansı – Ters Doküman Frekansı (TF-IDF) ve Latin Semantic Index (LSI) yöntemlerinden yararlanılmıştır. LSI'deki K-Means ve K-Median algoritmalarının başarıları karşılaştırıldığında, K-Means algoritmasının kümeleme başarısının önde olduğu görülmüştür (Taha, 2011)

Makale düzeyinde yapılan çalışmaların birinde; Destekçi Vektör Makinesi (Support Vector Machine) kullanılarak yeni bir sınıflama modeli denenmiştir. Sınıflanan metin dosyalarında genişletilmiş TF yöntemi kullanımı, anahtar kelime çıkarımı başarısında önemli artış sağlamıştır (Hong & Zhen, 2012).

Patent başvuru dokümanlarında üzerinde yapılan bir çalışmaya göre anahtar kelime çıkarımı yöntemlerinden en iyi anahtar kelime seçim stratejisi; 130 anahtar kelime çıkarımı özet görünüm üzerinden TF–IDF metodudur. TF-IDF, varyans ve sıklık metotları karşılaştırılmış ve patent analizinde en iyi yöntemin TF–IDF metodu olduğu sonucu ortaya çıkmıştır. Kümeleme analizi veri seti vektörü olarak yapıldığında Boolean Expression kullanılması da en iyi sonuçları desteklemiştir (Noh, Jo, & Lee, 2015).

Nöro-bilişim alanındaki metinlerde uzmanların anahtar kelime çıkarımını gerçekleştirmek amacıyla bir yazılım aracı geliştirilmiştir. Bu yazılımda kullanılan TF–ITF–TDCF metotları geleneksel yöntemlere göre daha başarılı olduğu ortaya çıkmıştır (Usui, Palmes, Nagata, Taniguchi, & Ueda, 2007).

Bir başka çalışmada anahtar kelime çıkarımı için zor bir veri alanı olan Mikro-blog sitelerinde kullanıcı ilgi alanları anahtar kelime çıkarımı yoluyla belirlenmiştir. Frekans tabanlı yöntem ile çeviri tabanlı bir yöntem kullanılarak birlikte kullanılarak yapılan çalışmanım sonunda bu metottun kullanıcıların ilgi alanları belirlemede doğruluk ve verimliliği oranı yüksek olduğu anlaşılmıştır (Liu, Chen, & Sun, 2012).

Diğer bir çalışmada ise benzer dokümanları sınıflama için içerik tabanlı anahtar kelime yaklaşımın frekans tabanlı yaklaşımdan daha iyi sonuçları verdiği görüldü. Çalışmada gerçekleştirilen algoritmanın, özel olayların konu takibine uygun olacağını görüşü benimsenmiştir (Kang, 2003).

Bir başka çalışmada blog yazısının ana konusunu, tam metin anahtar kelime alma metoduyla belirlemenin karmaşık ve zaman alıcı bir süreç olduğundan yola çıkarak eş anahtar kelime seçimi metodu kullanılmıştır. Bu metot, tam metin anahtar kelime alma sürecinin yerine kullanılabileceğine dair pozitif sonuçlar ortaya çıkmıştır. Bunun yanında, bu çalışma için TF ve TF-IDF metodu el ile özel alanı sınıflanması gerektiği için anlamlı farklılıklar içermektedir (Chen, Lu, & Tsai, 2014).

Diğer bir çalışmada gözetimli ve gözetimsiz öğrenme görevlerinde kullanılabilen otomatik anahtar kelime çıkarım algoritması sunulmuştur. Her algoritmada her metin dokümanı anlamsal bir ağdır. Anahtar kelimelerin çıkarımı anlamsal ağın yapısal dinamiklerine göre yapılır (Huan, Tian, Zhou, Ling, & Huang, 2006).

Diğer bir çalışmada, bilimsel makalelerden otomatik anahtar sözcük çıkarma sorunun çözmek için Destekçi Vektör Makinesi (Support Vector Machine) ve Random Forests makine öğrenmesi algoritmalarının performansını artırarak doğal dil işleme yöntemleri kullanılmıştır. Değerlendirmeler uzman temelli anahtar kelime ve anahtar kelime çıkarımı algoritmalar ile yapılmıştır (Krapivin, Autayeu, Marchese, Blanzieri, & Segata, 2010)

Yakın zamandaki bir çalışmada fraktal örüntüler üzerinde bir anahtar kelime çıkarımı model kullanılmıştır. Çalışma sonucuna göre birbirine yakın metin belgesinin konusu hakkında en yakın terimler ve en önemsiz terimlerin fraktal boyut değerlerine sahip olduğu sonucuna ulaşılmıştır. Kelimelerin önemi fraktal boyutlara göre belirlemektedir (Najafi & Darooneh, 2015).

Diğer bir çalışmada bilimsel metin sınıflandırmalarında sınıflandırma algoritmalarında ile birlikte 5 istatistiksel kelime çıkma yönteminin öngörü performansını incelenmiştir. Bu 5 yöntem; en sık ölçüm (MFM), TF-ISF, eş-oluşum istatistiki bilgiler, dış merkez tabanlı ve TextRank algoritması şeklindedir. Çalışma soncunda en yüksek tahmin performansı Random Forest algoritması ile birlikte kullanılan en sık ölçüm tabanlı kelime çıkarımı yöntemi ile (%93.8) elde edilmiştir (Onan, Korukoğlu, & Bulut, 2016).

Söz konusu bu çalışmalarda; bilgi sistemlerinde, internet ortamındaki metinlerde, patent ve bilimsel dokümanlarda sınıflandırma, kümeleme, gruplara ayırma işlemleri yapılarak ile anlamlı enformasyona ulaşılmaya çalışılmıştır. Enformasyona ulaşırken anahtar kelime çıkarımında; En Sık Ölçüm (MFM), TF, IDF, TF-IDF, TF-ISF, TF–ITF–TDCF metotları ile anahtar kelime çıkarım algoritmalarının kullanıldığı görülmüştür.

YÖNTEM

Bu çalışmada anahtar kelime çıkarımı yapılmadan önce metin ön işleme sürecinden geçirilmiştir. Bu aşamadan sonra anahtar kelime çıkarımı ile ilgili literatürde sıklıkla kullanılan TF-IDF, Ki-kare ve Bilgi Kazancı yöntemlerinden faydalanılmıştır.

VeriSeti

Yıldız Teknik Üniversitesi 7/24 Yıldızlı Hat, üniversite içerisinde herhangi bir konuda görüş, öneri, şikayet ve bilgi edinmenin gerçekleştiği bir iç paydaş görüş yönetim sistemidir. Sisteme öğrenciler, akademisyenler ve personel her konuda düşüncelerini elektronik ortamdaki formu doldurarak gerçekleştirebilmektedirler. Tüm mesajlar uzman bir personel tarafından okunarak ilgili birimlere yönlendirme yapılmakta ve belirlenen zaman aralığından geri dönüşler sağlanmaktadır. Sistem içerisinde daha çok idari birimlerle ilgili mesajlar kullanılmıştır. Örneğin Kurum içerisinde bulunan AB Ofisi idari bir birim olmakla birlikte öğrencilere hizmet vermektedir. Kütüphane daire başkanlığı hem öğrencilere hem akademisyenlere hizmet veren bir birimdir. Destek hizmetler müdürlüğü, kurumun başta temizlik ve ulaşım hizmetlerini yürütmektedir. Aşağıda bu çalışma kapsamında kullanılan birimlerin listesi bulunmaktadır.

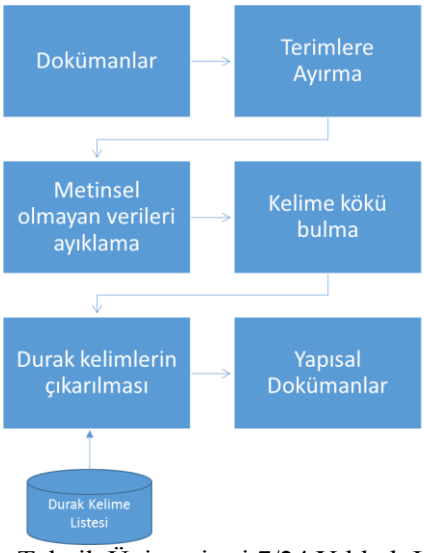
- Avrupa Birliği (AB) Ofisi
- Bakım Onarım Müdürlüğü
- Bilgi İşlem Daire Başkanlığı
- Burs Ofisi
- Destek Hizmetler Şube Müdürlüğü
- İletişim Koordinatörlüğü
- Kütüphane Daire Başkanlığı
- Öğrenci İşleri Daire Başkanlığı
- Sağlık Kültür Spor Daire Başkanlığı

Bu birimlere ait toplam içerisinde toplanan 3961 mesaj kullanılmıştır.

Metin Ön İşleme

Metin ön işleme ile ilgili adımlara ait algoritma Şekil 1 de gösterilmiştir.

Şekil 1. Metin ön işleme adımları



Dokümanlar: Yıldız Teknik Üniversitesi 7/24 Yıldızlı Hat iç paydaş görüş yönetim sistemi içerisinde yer alan mesajlar doküman olarak kullanılmıştır.

Terimlere Ayırma: Günümüzde internet ortamında yapılan yazışmalarda bazen Türkçe'de kullanılan ı,ö,ü,ç,ğ,ş harflerinin yerine çoğunlukla i,o,u,g,s harfleri kullanılmaktadır. Bu şekilde kullanım Türkçe doğal dil işleminin zorluklarından birini oluşturmaktadır. Bundan dolayı dokümanlar terimlere ayrıldıktan sonra Python programlama dili kullanılarak Türkçeleştirilmiştir.

Metinsel Olmayan Verileri Ayıklama: Dokümanlar içerisinde geçen metin olmayan !,?,/,+,☺ gibi karakterler veri setinden çıkarılmıştır.

Kelime Kökü Bulma: Açık kaynak kodlu hazırlanmış olan Zemberek kütüphanesi kullanılarak kelimeler köklerine ayrılmıştır (Akın & Akın, 2007).

Durak Kelimelerin Çıkarılması: Belgelerin içerisinde fazla sayıda geçen durak kelimeler, daha önceden hazırlanan durak kelime listesi baz alınarak çıkarılmıştır. Indeksleme sırasında etkisiz sözcüklerin kullanılmaması indeks dosyalarını küçülttüğü için bellek kullanım ve sorgu işleme verimliliğini artırmaktadır.

Yapısal Dokümanlar: Tüm bu işlemlerden sonra artık dokümanlar yapısal hale getirilmiştir. Bu aşamadan sonra öznitelik seçme yöntemleri kullanılarak anahtar kelime çıkarımı aşamasına geçilmiştir.

Terim Frekansı- Ters Doküman Frekansı (TF-IDF)

Özellik çıkarımında sıklıkla kullanılan TF-IDF yöntemi bir kelimenin doküman içinde ne sıklıkta var olduğu temeline dayanır. Veri seti içinde anahtar kelime özelliği öne çıkan ya da dokümanlar içerisinde sürekli geçen ve böylece

belirleyici özelliği bulunmayan kelimeler belirlenir (Çalış, Gazdağı, & Yıldız, 2013). Bir doküman içinde daha fazla görülen kelimeler TF ile hesaplanır. IDF ile tüm dokümanlarda nadir görülen kelimeler ile ilgili bir ölçü verilmektedir. TF-IDF skoru bir terimin sadece bir belge içindeki sıklığına bakmaz, aynı zamanda diğer belgeler içindeki durumunu da değerlendirir. Örneğin bir durak kelimesi tüm belgelerde sıklıkla geçebilir. TF-IDF yöntemi, eğer bir terim çok fazla dokümanda varsa onun önemini düşürmektedir. Çalışmamızdan bir örnek verdiğimizde; Kurumumuzda AB Ofisi bulunmaktadır. Mesajlara bakıldığında bu birim ile ilgili en çok kullanılan kelime "erasmus" kelimesidir. Bu kelime AB Ofisi birimden sıklıkla geçerken diğer birimlerde geçmemektedir. Bu durumda "erasmus" kelimesi bu birim için önemli bir kelimedir. TF-IDF çarpımı bir metinde çok bulunan ancak diğer metinlerde daha az görülen terimlerin ağırlığının fazla olduğunu göstermektedir. TF-IDF değeri aşağıda gösterilmiştir. i.döküman j.terim için ;

$$tf_{ij} = \frac{F_{ij}}{\sum_i F_j}$$

$$w_{i,d} = tf_{ij} * IDF_j$$

$$IDF_j = \log(\frac{D}{df_j})$$

Burada F:Frekans, df_j: j sözcüğünü içeren doküman sayısı, D: Doküman Sayısıdır.

Ki-Kare

Ki-Kare düşük hesaplama maliyetleri ve kolay uygulanabilir olmasından dolayı anahtar kelime çıkarımında sıklıkla kullanılan yöntemlerdendir. Ki-kare denklemi aşağıda verilmiştir.

$$Ki - Kare = N * \frac{(a * d - b * c)^2}{(a + c) + (b + d) + (a + d) + (c + d)}$$

Burada N toplam doküman sayısı, a: x terimini içeren kategori içerisindeki doküman sayısı, b: x terimini içermeyen kategori içerisindeki doküman sayısı, c: x terimini içeren kategori dışındaki doküman sayısı, d: x terimini içermeyen kategori dışındaki doküman sayısını temsil etmektedir.

Bilgi Kazancı

Bilgi kazancı hesaplanırken öncelikle alt bölümlere bölünmeden entropi bulunur, sonrasında tüm alt bölümlerin entropisi bulunarak iki değer arasındaki farkın büyük olduğu değişken en iyi kriter olarak seçilir. Bu teknik, terim azaltma işlemlerinde sıklıkla kullanılır (Cover & Thomas, 2012).

$$IG = \frac{a}{N} * log \frac{a * N}{(a + c)(a + b)} + \frac{b}{N} log \frac{bN}{(b + d)(a + b)} + \frac{c}{N} log \frac{cN}{(a + c)(c + d)} + \frac{d}{N} log \frac{dN}{(b + d)(c + d)}$$

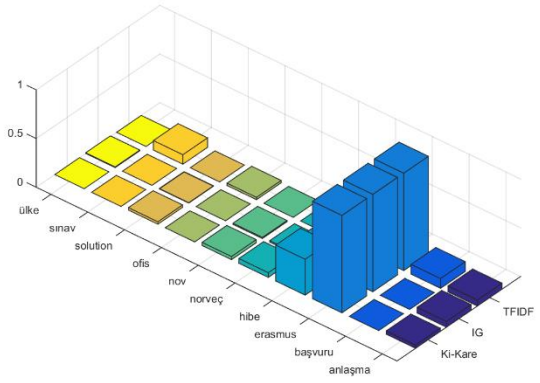
Bilgi kazancı eşitliğinde N toplam doküman sayısı, a: x terimini içeren kategori içerisindeki doküman sayısı,b: x terimini içermeyen kategori içerisindeki doküman sayısı, c: x terimini içeren kategori dışındaki doküman sayısı, d: x terimini içermeyen kategori dışındaki doküman sayısını temsil etmektedir (Akba, 2014).

BULGULAR

Ki-Kare, Bilgi Kazancı, TF-IDF Sonuçları

Şekil 2'de AB ofisi ile ilgili analiz sonuçları görülmektedir. Şekil 2'ye bakıldığında bu birimle ilgili en çok kullanılan kelimeler 3d grafik olarak gösterilmiştir. Bu grafik AB Ofisi'nde en çok "erasmus" ile ilgili mesajların olduğu, ardından "hibe" ile ilgili mesajların geçtiği görülmektedir. "Erasmus" kelimesi 3 yöntem içinde 1.sırada yer alacak skoru elde etmiştir. Ki-Kare, Bilgi Kazancı ve TF-IDF değerler aralıkları birbirinden farklıdır. Aynı grafikte ye almaları için Max-Min normalizasyon yöntemi kullanılarak, değerler 0-1 aralığına çekilmiştir.

Şekil 2. AB Ofisi anahtar kelime dağılımı

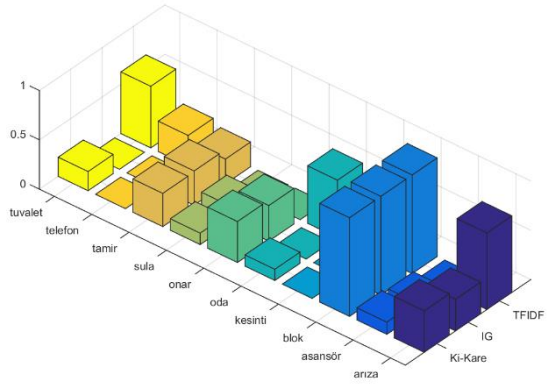


Tablo 1. AB Ofisi anahtar kelime değerleri

AB Ofisi	Ki-Kare	Bilgi Kazancı	TFIDF
arıza	0,0326	0,0580	0,0610
asansör	0,0000	0,0000	0,1001
blok	1,0000	1,0000	1,0000
kesinti	0,3680	0,2838	0,1233
oda	0,0481	0,0152	0,0000
onar	0,0321	0,0113	0,0000
sula	0,0000	0,0000	0,0237
tamir	0,0245	0,0037	0,0000
telefon	0,0000	0,0000	0,0972
tuvalet	0,0000	0,0070	0,0000

Şekil 3'de Bakım Onarım Şube Müdürlüğü mesajlarından çıkarılan anahtar kelime grafiği görülmektedir. Burdaki anahtar kelimelere bakıldığında Blok kelimesi ilk sırada yer almaktadır. Aslında Blok kelimesi tek başına bir anlam taşımamaktadır. Mesajların içerisinde "A Blok, B Blok" şeklinde kullanımları mevcuttur. Dolayısıyla bu grafikte "Blok" kelimesinden sonra en çok kullanılan kelimeler daha önemli hale gelmektedir. Örneğin "arıza", "tamir","onar" gibi kelimeler daha önemli bir hale gelmektedir.

Şekil 3 Bakım Onarım Şube Müdürlüğü anahtar kelime dağılımı



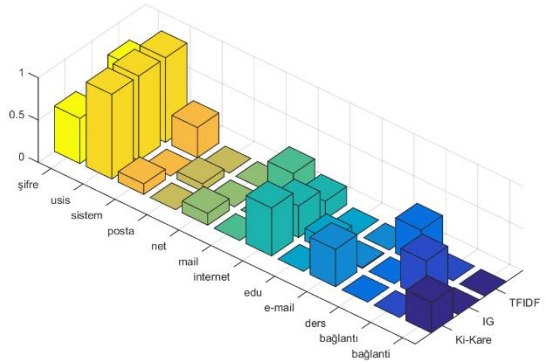
Tablo 2'de Bakım Onarım Müdürlüğü'ne ait anahtar kelime değerleri bulunmaktadır.

Tablo 2. Bakım Onarım Şube Müdürlüğü Anahtar Kelime değerleri

Bakım Onarım Md.	Ki-Kare	Bilgi Kazancı	TFIDF
arıza	0,4199	0,3205	0,7761
asansör	0,1235	0,0972	0,0000
blok	1,0000	1,0000	1,0000
kesinti	0,0000	0,0000	0,2855
oda	0,1133	0,0000	0,5828
onar	0,4155	0,3607	0,0000
sula	0,1160	0,1677	0,0000
tamir	0,3369	0,3472	0,2507
telefon	0,0000	0,0000	0,3080
tuvalet	0,1941	0,0000	0,6245

Şekil 4'te Bilgi İşlem Daire Başkanlığı'na ait anahtar kelimelere ait grafik görülmektedir. Çok açık bir şekilde ilk sırada yer alan kelime "usis" kelimesidir. Bu kelime Yıldız Teknik Üniversitesi Öğrenci Otomasyon Sistemini temsil etmektedir. Sonraki en çok kullanılan kelimelere bakıldığında "şifre", "internet","sistem" gibi kelimeler yer almaktadır.

Şekil 4. Bilgi İşlem Daire Başkanlığı anahtar kelime dağılımı



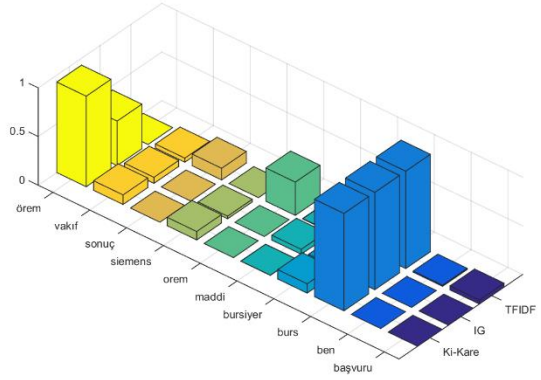
Tablo 3'de Bilgi İşlem Daire başkanlığına ait anahtar kelime değerleri bulunmaktadır.

Tablo 3 Bilgi İşlem Daire Başkanlığı anahtar kelime değerleri

Bilgi İşlem Daire Bşk.	Ki-Kare	Bilgi Kazancı	TFIDF
bağlanti	0,3592	0,0000	0,0000
bağlantı	0,0000	0,4516	0,0000
ders	0,0000	0,0000	0,4166
e-mail	0,4408	0,0097	0,0000
edu	0,0000	0,1452	0,0000
internet	0,5667	0,3727	0,2126
mail	0,0000	0,0000	0,3574
net	0,1386	0,0000	0,0000
posta	0,0000	0,0887	0,0000
sistem	0,1237	0,0000	0,3563
usis	1,0000	1,0000	1,0000
şifre	0,5344	0,3960	0,7088

Şekil 5'te Burs Ofisi'ne ait anahtar kelime grafiği yer almaktadır. "Burs" kelimesi her yöntem için birinci anahtar kelime olarak tespit edilmiştir. Sonrasında "örem" kelimesi gelmektedir. "örem" kelimesi Burs Ofisi'nin eski adı olmakla birlikte, Öğrenci Rehberlik Merkezi'nin kısaltılmasından oluşmaktadır. Tablo 4'te Burs Ofisi'ne ait anahtar kelime değerleri verilmiştir.

Şekil 5. Burs Ofisi anahtar kelime dağılımı



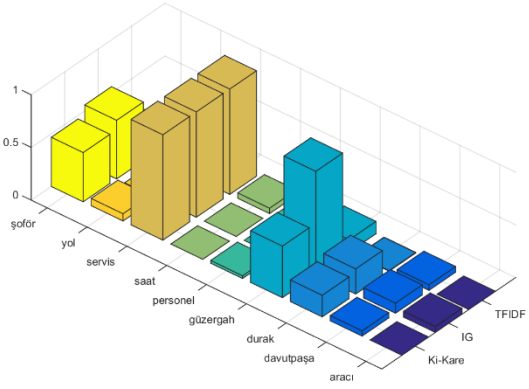
Tablo 4. Bilgi İşlem Daire Başkanlığı anahtar kelime değerleri

Ofisi	Burs Kare	Ki- Kazancı	Bilgi TFIDF
	başvuru	0,0000	0,0000	0,0406
	ben	0,0000	0,0000	0,0111
	burs	1,0000	1,0000	1,0000
	bursiyer	0,0956	0,0367	0,0000
	maddi	0,0016	0,0504	0,0047
	orem	0,0000	0,0000	0,3416
	siemens	0,0862	0,0270	0,0000
	sonuç	0,0000	0,0000	0,1196
	vakıf	0,0980	0,0617	0,0417
	örem	0,9296	0,4570	0,0000

Şekil 6'da Destek Hizmetleri Şube Müdürlüğüne ait anahtar kelimelere ait grafik görülmektedir. Başlıca görevleri arasında ulaşım ve temizlik olan birimle

ilgili anahtar kelimelere bakıldığında ulaşım ile ilgili kelimelerin ağırlıkta olduğu görülmektedir. Tablo 5'de bu birime ait anahtar kelime değerleri verilmiştir

Şekil 6. Destek Hizmetler Şb. Md. anahtar kelime dağılımı.

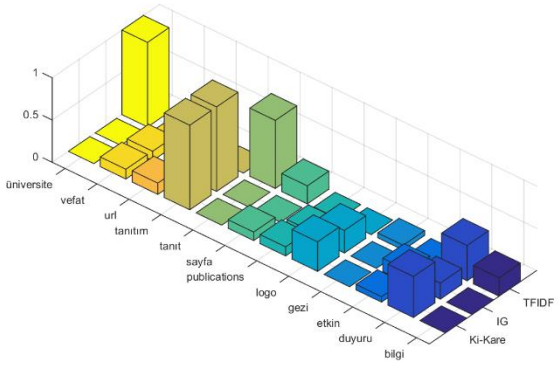


Tablo 5. İletişim Koordinatörlüğü anahtar kelime değerleri

Destek Hizmetler Şb.Md.	Ki-Kare	Kazancı	Bilgi TFIDF
aracı	0,0000	0,0550	0,0000
davutpaşa	0,0491	0,0958	0,0526
durak	0,2050	0,2385	0,0042
güzergah	0,4897	0,9818	0,1167
personel	0,0261	0,0000	0,0435
saat	0,0000	0,0000	0,0510
servis	1,0000	1,0000	1,0000
yol	0,0684	0,0000	0,0000
şoför	0,4668	0,5561	0,1222

İletişim Koordinatörlüğü'ne ait anahtar kelimelerin gösterildiği grafik Şekil 7'de görülmektedir. Başlıca görevleri arasında kurumun tanıtımı olan birime ait anahtar kelimelere bakıldığında "tanıtım" kelimesi ilk sırada yer almaktadır. Tablo 6'da birime ait anahtar kelime değerleri verilmiştir.

Şekil 7. İletişim Koordinatörlüğü anahtar kelime dağılımı

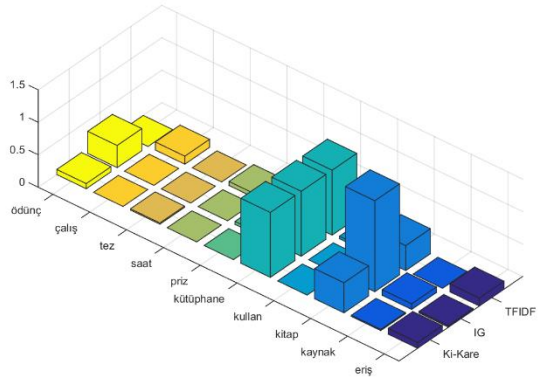


Tablo 6. İletişim Koordinatörlüğü anahtar kelime değerleri

Koord.	İletişim	Ki-Kare	Kazancı	Bilgi	TFIDF
	bilgi	0,0000	0,0000		0,2180
	duyuru	0,4847	0,1889		0,4184
	etkin	0,0640	0,1556		0,0000
	gezi	0,0000	0,0000		0,0567
	logo	0,3491	0,2667		0,0000
	publications	0,1072	0,0689		0,0000
	sayfa	0,1003	0,0000		0,2128
	tanıt	0,0000	0,0000		0,8052
	tanıtım	1,0000	1,0000		0,0000
	url	0,1331	0,0000		0,0116
	vefat	0,1152	0,1111		0,0000
	üniversite	0,0000	0,0000		1,0000

Kütüphane Daire Başkanlığı kategorisinde yapılan analizlerde ortaya çıkan anahtar kelimelere bakıldığında kütüphane ile ilgili genel kelimelerin yanında "eriş" kelimesi de seçilen anahtar kelimelerin arasına girmiştir. Bu tablodan elde edilen enformasyon yöneticiler için atacakları adımlar açısından oldukça faydalıdır. Tablo 10'da birime ait anahtar kelime değerleri görülmektedir.

Şekil 8. Kütüphane Daire Başkanlığı anahtar kelime dağılımı

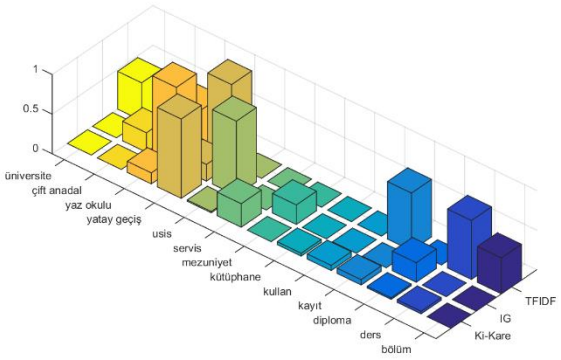


Tablo 7. Kütüphane Daire Başkanlığı anahtar kelime değerleri

Kütüphane Daire Bşk.	Ki-Kare	Bilgi Kazancı	TFIDF
eriş	0,0882	0,0251	0,1224
kaynak	0,0177	0,0858	0,0000
kitap	0,4633	1,4005	0,3842
kullan	0,0000	0,0000	0,0384
kütüphane	1,0000	1,0000	1,0000
priz	0,0000	0,0453	0,0000
saat	0,0000	0,0000	0,0619
tez	0,0189	0,0000	0,0000
çalış	0,0000	0,0000	0,1236
ödünç	0,0787	0,3406	0,0000

Öğrenci İşleri Daire Başkanlığı'na ait Şekil 11'de ki grafik incelendiğinde "yatay geçiş", "yaz okulu","usis" kelimeleri ilk sıralarda yer almaktadır. Öğrencilerin mesajları daha çok bu konularla ilgilidir.

Şekil 9. Öğrenci İşleri Daire Başkanlığı anahtar kelime dağılımı

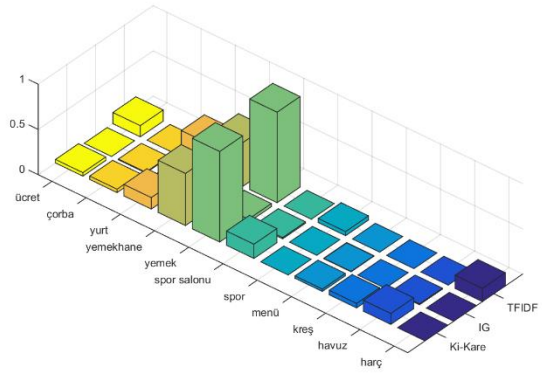


Tablo 8.Öğrenci İşleri Daire Başkanlığı anahtar kelime değerleri

Öğrenci İşleri Daire Bşk.	Ki-Kare	Bilgi Kazancı	TFIDF
bölüm	0,0000	0,0000	0,4522
ders	0,0326	0,0000	0,7428
diploma	0,0199	0,2456	0,0000
kayıt	0,0663	0,0000	0,7248
kullan	0,0641	0,0000	0,0000
kütüphane	0,0362	0,0000	0,0000
mezuniyet	0,0000	0,2462	0,0000
servis	0,2961	0,0000	0,0000
usis	0,0162	0,9378	0,0000
yatay geçiş	1,0000	0,2107	1,0000
yaz okulu	0,1390	1,0000	0,5015
çift anadal	0,0000	0,2452	0,0000
üniversite	0,0000	0,0000	0,4764

Sağlık Kültür Spor (SKS) Daire Başkanlığı birimine ait anahtar kelimeler incelendiğinde birimin sorumluluğu altında olan alanlarla ilgili kelimeler ön plandadır. Özellikle “yemek”,”yemekhane”,”spor”,”yurt” gibi kelimeler birim yöneticileri için faydalı bir enformasyondur. Şekil 10’da birime ait anahtar kelime dağılımı, Tablo 9’da ise anahtar kelime değerleri görülmektedir.

Şekil 10. Sağlık Kültür Spor (SKS) Daire Başkanlığı anahtar kelime dağılımı



Tablo 9. SKS Daire Başkanlığı anahtar kelime değerleri

Bşk.	SKS Daire	Ki-Kare	Bilgi Kazancı	TFIDF
	harç	0,0000	0,0000	0,1456
	havuz	0,1204	0,0095	0,0000
	kreş	0,0496	0,0000	0,0000
	menü	0,0262	0,0043	0,0000
	spor	0,0000	0,0000	0,0453
	spor salonu	0,1576	0,0143	0,0000
	yemek	1,0000	0,0348	1,0000
	yemekhane	0,5753	0,0583	0,4833
	yurt	0,1265	0,0000	0,2882
	çorba	0,0373	0,0083	0,0000
	ücret	0,0383	0,0000	0,1234

SONUÇ VE ÖNERİLER

Hayatın bir çok alanında sürekli veri toplanmaktadır. Her geçen gün bu veriye ait miktarlar hızla artmaktadır. Toplanan bu verilerden enformasyonlar oluşturmak, ileri düzeyde kararlara destek sağlamak, verilerden anlam çıkarmak ve süreç optimizasyonuna destek sağlamak, bulunduğu kurum için yaşadığımız bilgi çağı itibariyle kaçınılmaz bir hale gelmiştir. Yıldız Teknik Üniversitesi bünyesinde iç paydaşlar tarafından aktif olarak kullanılan 7/24 Yıldızlı Hat Yönetim Sistemi içerisinde biriken mesajlar anahtar kelime özellik çıkarımı için kullanılmıştır. Ki-Kare, Bilgi Kazancı ve TF-IDF yöntemleri ile gerçekleştirilen analiz sonuçlarında anlamlı ve faydalı anahtar kelimeler bulunmuştur. Sistem normal çalışma düzeninde yöneticilere sadece basit istatistiksel düzeyde enformasyon sağlamaktadır. Bu da kaç mesaj atıldı, kaç tanesine cevap verildi soruları ile sınırlıdır. Bu yöntemin sisteme entegre edilmesiyle üst yönetici destek sistemleri geliştirilmiş olacaktır. Beaumont ve Sutherland (1992)'a göre üst yönetim bilgi sistemlerinden beklenen en büyük yarar, planlama ve kontrol süreçlerinin yeniden yapılandırılarak etkinliğin sağlanmasıdır (Göl,1999). Bu çalışma kurumun genelinde ve bölümleri düzeyinde var olan raporlama hizmetlerini iyileştirme, yönetimle ilgili raporlama sisteminin yeniden düzenlemenin yapılabilmesine olanak sağlama kapasitesine sahiptir. Üst yönetim bilgi sistemine ilave edilecek bir modülle sonuçlar grafiksel olarak gösterildiğinden dolayı elde edilen bilgiler daha kolay anlaşılacak ve böylece yöneticinin alacağı stratejik kararların alınmasına ve uygulanmasına katkı sağlayacaktır.

Gelecek çalışmalarda Üniversite birimlerine ait web site yönetim paneline eklenecek bir bölümde, 7/24 yıldızlı hat verileri bilgi işleme ve analizi sonuçları otomatik olarak eklenebilir. İlgili birimin web yöneticisi bu bölümdeki anahtar kelimelerinden uygun gördüklerine, web sitesinde bir başlık açabilir, böylece kullanıcılar ihtiyaç duydukları bilgilere mesaj yazmadan ulaşabilirler. Bu işlem yapıldıktan sonra 7/24 Yıldızlı Hat'ta bu kelime ile ilgili mesajların yoğunluğu ile ilgili değişiklikler izlenebilir.

KAYNAKLAR

Akba, F. (2014). Duygu analizinde öznitelik seçme metriklerinin değerlendirilmesi: Türkçe film eleştirileri. *Yüksek Lisans Tezi*. Hacettepe Üniversitesi, Fen Bilimleri Enstitüsü, Ankara.

Akın, A. A., & Akın, M. D. (2007). Zemberek, an open source NLP framework for Turkic languages. Structure, 10,1-5.

Beaumont R. J. & Sutherland E. (1992). *Information recources management*, Contemporary Business Series Buttertwath - Heinemann Ltd.

Chen, Y. H., Lu, E. J., & Tsai, M. F. (2014). Finding keywords in blogs: Efficient keyword extraction in blog mining via user behaviors. *Expert Systems with Applications*, 41(2): 663-670.

Cover, T. M., & Thomas, J. A. (2012). *Elements of information theory*. Canada: John Wiley & Sons.

Çalış, K., Gazdağı, O., & Yıldız, O. (2013). Reklam İçerikli Epostaların Metin Madenciliği Yöntemleri ile Otomatik Tespiti. *Bilişim Teknolojileri Dergisi*, 6(1):1-7.

Çelikyay, E. K. (2010). Metin madenciliği yöntemiyle Türkçe'de en sık kullanılan ve birbirini takip eden harflerin analizi ve birliktelik kuralları. *Yüksek Lisans Tezi*. Beykent Üniversitesi, Fen Bilimleri Enstitüsü, İstanbul.

Döven, S. (2013). Metin Madenciliği ile Dokümanlar Arasındaki Benzerliklerin Bulunması. *Yüksek Lisans Tezi*, Bahçeşehir Üniversitesi, Fen Bilimleri Enstitüsü, İstanbul.

Feldman, R., & Sanger, J. (2007*). The text mining handbook: advanced approaches in analyzing unstructured data*. Cambridge University Press.

Göl, M. (1999). Stratejik karar alma ortamında üst yönetim bilgi sistemi ve uzman sistemler, *Dumlupınar Üniversitesi Sosyal Bilimler Dergisi*, 3:357-364.

Hong, B., & Zhen, A. (2012). An extended keyword extraction method. *Physics Procedia*, 24:1120-1127.

Huan, C., Tian, Y., Zhou, Z., Ling, C. X., & Huang, T. (2006). Keyphrase extraction using semantic network structure analysis. *In Proceedings of The Sixth International Conference on Data Mining*, 275-284.

Kang, S. S. (2003). Keyword-based document clustering. *In Proceedings of The Sixth International Workshop on Information Retrieval with Asian Languages ,Association For Computational Linguistics*, 132-137.

Karaca, M. F. (2012). Metin madenciliği yöntemi ile haber sitelerindeki köşe yazılarının sınıflandırılması. *Yüksek Lisans Tezi*, Karabük Üniversitesi, Fen Bilimleri Enstitüsü, Kocaeli.

Krapivin, M., Autayeu, A., Marchese, M., Blanzieri, E., & Segata, N. (2010). Keyphrase extraction from scientific documents: Improving machine learning approaches with natural language processing. *Lecture Notes in Computer Science*, 6102: 102–111.

Liu, Z., Chen, X., & Sun, M. (2012). Mining the interests of Chinese microbloggers via keyword extraction. *Frontiers of Computer Science*, 6(1): 76-87.

Najafi, E., & Darooneh, A. H. (2015). The fractal patterns of words in a text: A method for automatic keyword extraction. *PLoS One*, 10 (6).

Noh, H., Jo, Y., & Lee, S. (2015). Keyword selection and processing strategy for applying text mining to patent analysis. *Expert Systems with Applications*, 42(9): 4348-4360.

Onan, A., Korukoğlu, S., & Bulut, H. (2016). Ensemble of keyword extraction methods and classifiers in text classification. *Expert Systems with Applications*, 57: 232-247.

Pilavcılar, İ. F. (2007). Metin madenciliği ile metin sınıflandırma. *Yüksek Lisans Tezi*. Yıldız Teknik Üniversitesi, Fen Bilimleri Enstitüsü, İstanbul.

Taha, S. M. (2011). Metin madenciliği ile doküman demetleme. *Yüksek LisansTezi*. Gazi Üniversitesi, Bilişim Enstitüsü, Ankara.

Usui, S., Palmes, P., Nagata, K., Taniguchi, T., & Ueda, N. (2007). Keyword extraction, ranking, and organization for the neuro informatics platform. *Biosystems*, 88(3): 334–342.

Visa, A. (2001). Technology of text mining. *In International Workshop on Machine Learning and Data Mining in Pattern Recognition*, Berlin: Springer,132-137.