

T.C.
TRAKYA ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

MAKİNE ÖĞRENMESİ İLE
TÜRKÇE HABER METİNLERİNDE
ANAHTAR İFADE ÇIKARIMI

Mustafa ÇETİNGÖZ

Yüksek Lisans Tezi

Bilgisayar Mühendisliği Anabilim Dalı

Danışman: Yrd. Doç. Dr. Erdem UÇAR

EDİRNE 2011

T.C.
TRAKYA ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

MAKİNE ÖĞRENMESİ İLE TÜRKÇE HABER METİNLERİNDE
ANAHTAR İFADE ÇIKARIMI

MUSTAFA ÇETİNGÖZ

YÜKSEK LİSANS TEZİ
BİLGİSAYAR MÜHENDİSLİĞİ ANA BİLİM DALI

Bu Tez 29/06/2011 Tarihinde Aşağıdaki Jüri Tarafından Kabul Edilmiştir.

Doç. Dr. Yılmaz KILIÇASLAN
Üye

Doç. Dr. Tahir ALTINBALIK
Üye

Yrd. Doç. Dr. Erdem UÇAR
Danışman

EDİRNE

2011

Yüksek Lisans Tezi
Trakya Üniversitesi Fen Bilimleri Enstitüsü
Bilgisayar Mühendisliği

**MAKİNE ÖĞRENMESİ İLE
TÜRKÇE HABER METİNLERİNDE
ANAHTAR İFADE ÇIKARIMI**

ÖZET

Sayısal kaynakların çokluğu ve genel ağda (internet) yaşanan gelişmeler, üretilen bilgi miktarında artış yaşanması sonucunu doğurmaktadır. Bilgi erişimi, ulaşılmak istenen bilgi kaynaklarının düzenlenmesi, istenildiğinde bilgiye en kısa zamanda ve en kolay şekilde ulaşılmalarının sağlanması ile ilgilenen bir araştırma konusudur.

Metin işleme; bilgi erişiminin önemli konuları arasındadır. Ayrıca, ses ve görüntü gibi bilgi kaynaklarına erişim konuları da ilgili çalışmalar arasında yer almaktadır. Metin işlemenin alt konuları arasında ise, metnin özetlenmesi, metni karakterize edecek anahtar ifadelerin belirlenmesi gibi uygulamalar bulunmaktadır.

Anahtar ifade, bir metnin içeriğini özetleyen anlamsal kelime ya da kelimeler topluluğudur. Anahtar ifade ile, bir metnin tamamının okunmasına gerek kalmayacak şekilde metnin içeriği hakkında fikir sahibi olunması amaçlanmaktadır. Otomatik anahtar ifade çıkarımı ise, insan eliyle çıkarılan anahtar ifadelere en yakın anahtar ifadelerin bulunması işlemidir.

Otomatik anahtar ifade çıkarımı alanında hâlen açık kaynak kodlu yazılım olan KEA (Keyphrase Extraction Algorithm) algoritması ile İngilizce, İspanyolca ve Fransızca dilleri için yapılmış yazılımlar sıkça kullanılmaktadır. Kea algoritması, bir metin üzerinde sözcüksel yöntemler kullanılarak aday anahtar ifadelerin belirlenmesi,

her aday ifade için özellik deęerlerinin hesaplanması ve ardından makine öğrenme yöntemlerini kullanarak aday ifadeler arasından uygun olanın seçimi adımlarından oluşur.

Bu çalışmada da Türkçe haber metinlerinden elde edilen eğitim ve test verileri kullanılarak, KEA algoritması ile ve ilave bir özellik eklenerek oluşturulan KEA-SPR algoritması ile uygulama geliştirilmiş, ilave edilen özellik için performans karşılaştırılması yapılmıştır.

Anahtar kelimeler: Anahtar Kelime, Anahtar İfade, Türkçe Haber Metni, Makine Öğrenmesi, KEA.

Master Thesis
Trakya University Graduate School of
Natural and Applied Sciences
Department of Computer Engineering

KEYPHRASE EXTRACTION FOR TURKISH NEWS TEXT WITH MACHINE LEARNING METHODS

ABSTRACT

The abundance of digital sources and developments in the general network (internet) have resulted in the increase in the amount of produced information. Information retrieval is a research subject dealing with the organization of information sources to be reached and with providing information easily and as soon as possible if required.

Especially text processing and additionally studies about the subjects concerning the access to information sources such as sound and video have taken part among the subjects of information retrieval. In addition, implementations such as summarization of text, determination of the keyphrases to characterize the text have taken part among the sub-questions of text processing.

Keyphrase is a semantic word or phrase summarizing the content of a text. By using keyphrases, it is aimed to have an opinion about the content of a text without reading it completely. Automatic keyphrase extraction is deducing the most proximate keyphrases to the one inferred by people.

It is still possible to encounter frequently the use of software designed for English, Spanish and French languages with the open source software KEA (Keyphrase Extraction Algorithm) on the subject of automatic keyphrases extraction. The Kea algorithm consists of determining the candidate keyphrases on a text by using lexical

methods, calculating the feature value for each candidate and then choosing the most convenient candidate among the candidate phrases by using machine learning methods.

In this study, the application has been developed through the KEA algorithm and KEA-SPR algorithm constituted with an additional feature by using the data of education and test obtained from Turkish news texts and also the performance comparison has been made for the additional feature.

Key words: Keyword, Keyphrase, Turkish News Text, Machine Learning, KEA.

ÖNSÖZ

Dijital kaynakların bu denli çok olması ve üretilen bilginin her geçen gün hızla artması, bu kaynaklara erişim için uygulanan yöntemlerin de önemini arttırmaktadır.

Bu noktada bir metnin tamamının okunmasına gerek kalmadan, metni karakterize eden anahtar ifadelerin otomatik olarak elde edilmesini daha önemli hale getirmektedir.

Bu çalışma ile günlük bilgi üretiminin fazla olduğu haber portalları üzerinden elde edilen haber metinleri kullanılmış, daha önce karşılaşılmamış haber metinlerinden otomatik olarak anahtar ifade elde edilmesini sağlayan yeni bir yöntem denenerek bir uygulama geliştirilmiştir.

TEŞEKKÜR

Bu tezin yazımında, başından sonuna kadar emeđi geen saygıdeđer hocam ve danıřmanım Sayın Yrd. Do. Dr. Erdem UAR'a, tım katkılarından ve desteđinden dolayı teřekkır ederim.

Tez alıřmam sırasında ihtiya duyduđum her zaman, bilgi ve tecrubesini esirgemeyen deđerli hocam Bilgisayar Mřhendisliđi Břlım Bařkanı Sn. Yılmaz KILIASLAN'a da tım katkı ve desteđinden dolayı teřekkır ederim.

Ayrıca test alıřmam sřresince desteđini eksiltmeyen deđerli dostlarım Öğretim Görevlisi Sn. M. Ozan AKI'ya ve Arařtırma Görevlisi Sn. Edip Serdar GÜNER'e de teřekkır bir bor bilirim.

İÇİNDEKİLER

ÖZET	I
ABSTRACT	III
TEŞEKKÜR	VI
İÇİNDEKİLER	1
ŞEKİLLER TABLOSU	3
TABLolar	4
1. GİRİŞ	5
2. TÜRKÇENİN DİL YAPISI	9
2.1. Türkçenin Ek Yapısı	9
2.1.1. Yapım ekleri	10
2.1.2. Çekim ekleri	12
2.2. Türkçede Kompozisyon Yapısı	13
2.2.1. Başlık	14
2.2.2. Konu	14
2.2.3. Paragraf planı	15
2.3. Gövdeleme	16
3. ANAHTAR İFADELER İLE İLGİLİ YAPILMIŞ ÇALIŞMALAR ve ÇEŞİTLİ UYGULAMA ALANLARI	21
3.1. Önceki Çalışmalar	21
3.1.1. Türkçe için yapılmış bazı çalışmalar	21
3.1.2. İngilizce için yapılmış çalışmalar	22
3.2. Bazı Uygulama Alanları	25
3.2.1. Üstverilerde (metadata) kullanım	25
3.2.2. Vurgulama işlemlerinde kullanım	26
3.2.3. Dizinlendirmede kullanım	27
3.2.4. Etkileşimli sorgulamada	27
3.2.5. Web Log analizlerinde kullanım	27
3.2.6. Haber portallarında kullanım	28
4. KEA-TR ve KEA-SPR ALGORİTMASI	30
4.1. Aday İfade Seçimi Aşaması	30
4.1.1. Giriş dosyasının temizlenmesi	33

4.1.2. Aday ifadelerin belirlenmesi	33
4.1.3. Büyük-Küçük harf dönüşümü ve gövdeleme	34
4.2. Özellik hesaplama	35
5. YAZILIMIN GERÇEKLEŞTİRİLMESİ	40
5.1. Giriş Dosyasının Temizlenmesi	41
5.2. Adayların Belirlenmesi	43
5.3. Sıklık, İlk Konum ve Dağılım Değerlerinin Tespiti.....	47
5.4. Anahtar İfade Seçimi.....	50
6. KEA-SPR TEST AŞAMASI.....	52
6.1. Spread Özelliğinin Doğrudan Etkisi	52
6.2. Spread Özelliğinin Anahtar İfade Sayısı Artışına Göre Etkisi.....	53
7. SONUÇLAR ve TARTIŞMA	55
KAYNAKLAR	56
ÖZGEÇMİŞ	59
EKLER.....	60

ŞEKİLLER TABLOSU

Şekil 3-1 Microsoft Word 2003 ile otomatik özet formu	26
Şekil 3-2 Posta Gazetesi internet sitesinde 17.05.2011 tarihinde yayımlanan haber.....	29
Şekil 4-1 Eğitim Aşaması	31
Şekil 4-2 Çıkarım Aşaması.....	32
Şekil 5-1 Borland Delphi 7 programlama arayüzü.....	40
Şekil 5-2 Örnek bir eğitim dosyası	42
Şekil 5-3 Ön işlemde geçirilmiş olan eğitim dosyası	42

TABLULAR

Tablo 2-1 “gözlükçülerdeki” kelimesi için zemberek program çıktısı.....	19
Tablo 5-1 Gövdelenmiş Kelime Listesi.....	43
Tablo 5-2 Durdurma Listesindeki Kelimeler Örneği	44
Tablo 5-3 Aday ifade listesi.....	45
Tablo 5-4 Genel Ağ Tabanlı Derlem Dosyası Örneği	46
Tablo 5-5 Sıklık değerinin sınıflandırılması.....	48
Tablo 5-6 İlk konum değerinin sınıflandırılması.....	48
Tablo 5-7 Dağılım değerinin sınıflandırılması.....	48
Tablo 5-8 Tüm olasılık değerleri örneği.....	50
Tablo 5-9 Anahtar İfade Son Seçim Örneği	51
Tablo 6-1 KEA ve KEA-SPR Algoritmaları karşılaştırma tablosu.....	53
Tablo 6-2 Anahtar İfade Sayısı Artışına Göre Karşılaştırma	54

1. GİRİŞ

Dijital kaynakların bu denli çok olması ve üretilen bilginin her geçen gün hızla artması, bu kaynaklara erişim için uygulanan yöntemlerin de önemini git gide arttırmaktadır. Bilgi erişimi, ulaşılmak istenen bilgi kaynaklarının düzenlenmesi, istenildiğinde bilgiye en kısa zamanda ve en kolay şekilde ulaşılmalarının sağlanması ile ilgilenen bir araştırma konusudur.

Bilgi erişiminde; ses ve görüntü gibi bilgi kaynaklarına erişim ile ilgilenilmesinin yanı sıra, daha çok metin belgeleri üzerinde yapılan çalışmalara rastlamak mümkündür. Metin işleme alt konuları arasında, metnin özetlenmesi, metni karakterize edecek anahtar ifadelerin belirlenmesi gibi uygulamalar yer almaktadır.

Anahtar ifade; bir metnin içeriğini özetleyen anlamsal kelime ya da kelimeler topluluğudur. Diğer bir deyişle “En kısa ve doğru bir şekilde, bir metnin konusunun ya da içeriğinin ifade edilmesidir.” (Feather and Sturges, 1996) Burada yaygın olarak kullanılan anahtar kelime ya da anahtar ifade terimlerinden bahsetmek gerekirse; aslında anlamsal olarak benzer ifadeler olmalarıyla birlikte, anahtar kelimenin genelde bir kelime için, anahtar ifadenin ise birkaç kelimenin yan yana gelmesi ile oluştuğu söylenebilir.

Anahtar kelime ya da anahtar ifadeler bilgi erişim sistemlerine tanımlayıcı ifadelerle sorgulama yapılmasında, dokümanların indekslenmesinde, sınıflandırmada, özetlenmesi aşamalarında ve belge kümeleme gibi alanlarda kullanılmaktadır. Ayrıca kullanıcılara tüm dokümanın içeriğinde önem arz eden noktaların gösterilmesinde, doküman taramasında görsel olarak önemli ifadelerin vurgulanmasında ve etkili bir doküman benzerlik ölçümünde fayda sağlar. (Witten, 1999; Gutwin, 1998) Doğal olarak arama esnasında anahtar ifade listesinde yer alan bir ifadenin bulunması için harcanan zaman ile, o ifadenin tüm doküman üzerinde tarama yapılarak elde edilmesi arasında büyük fark olacaktır. İşte bu noktada otomatik anahtar ifade belirleme sistemleri, el ile işaretlemenin yerini almaktadır. Aslında nihai amaç, dokümanlar üzerinde yapılacak bir işlem vasıtasıyla, dokümanı anlatacak, içeriği hakkında fikir sahibi olunmasını

sağlayacak anahtar ifadelerin otomatik olarak çıkarılmasıdır. Buradaki ölçüt; insan eliyle oluşturulan anahtar ifadelerle en yakın anahtar ifadelerle ulaşılmasıdır.

Dokümanlar yüksek oranda anahtar ifade içermemekle birlikte, içeren dokümanlardaki anahtar ifadelerin seçimi ise genellikle manüel yapılmaktadır. (Witten, 1999) Birçok akademik dokümana ise yazarı tarafından anahtar ifadeler yine manüel olarak eklenmektedir. Profesyonel dizin oluşturucular tarafından da, konuya bağlı olarak ve önceden tanımlanmış bir sözlük aracılığı ile yine manüel yapılır. Ancak bu işlem için konuyu önceden bilmek önem arz eder ve büyük bir işlem hacmi gerektirir. El ile anahtar ifade belirlemenin sıkıcı ve zorlu bir süreç olması sebebiyle, bu sürecin otomatik hale getirilmesinin sağlayacağı fayda tartışılmazdır.

Günlük hayatta internet üzerinde bilgi taraması yapmak, bilgisayar kullanıcılarının genel faaliyeti haline dönüşmüştür. İnternet sayesinde; sınırsız bilgi kaynaklarına ulaşma imkânına kavuşulmasının yanında, bu bilgi kalabalığı içerisinde istenilen bilgiye ulaşmak da giderek zorlaşmaktadır. Düzgün ve doğru arama terimlerinin belirlenmesi, elektronik taramanın en önemli ölçütüdür. Kütüphane kataloglarında, kullanıcılar kontrollü bir sözlükle desteklenirler. Bahsi geçen bu sözlüklerin amacı ise içeriği anlatacak anahtar kelime ya da ifadeler vasıtasıyla dokümanları etiketlenmiş gruplara ayırmaktır. Bu anahtar ifadelerin seçimi, profesyonel dizin oluşturucular tarafından el ile gerçekleştirilir. Anahtar ifadenin önemi ve el ile işaretlemenin zorluğu göz önüne alındığında, otomatik anahtar ifade belirlemenin önemi bu alanda da ön plana çıkmaktadır. Ancak pratik uygulamalarda otomatik anahtar ifade çıkarma ve tahsisi yaklaşımlarının başarısı maalesef yüksek değildir. (Medelyan, 2005)

Bir metin dosyasından özet bilgi üretiminde veya çıkarımında birkaç yöntem önerilmektedir. (Johnson, 1993) Anahtar kelimeler söz konusu olduğunda iki farklı yaklaşım söz konusudur. Bunlar; Anahtar İfade Tahsisi ve Anahtar İfade Çıkarımı yöntemleridir. Her ikisinde de Makine Öğrenme metotları kullanılmakta, aynı zamanda anahtar ifadelerle sahip eğitim dokümanlarına ihtiyaç duyulmaktadır. Anahtar İfade Tahsisinde; dokümanı en iyi biçimde tanımlayacak ifadeler kontrollü bir sözlükten seçilmeye çalışılır. Eğitim verisi sözlükteki her bir ifade için birleştirilir ve sınıflandırıcılar oluşturulur. Yeni doküman her sınıflandırıcı için işlenir ve pozitif örnekler ile sınıflandırılan model vasıtasıyla anahtar ifade tahsisi yapılır. Burada da

tahsisi yapılan anahtar ifadeler, önceden eğitim verisinde bulunan değerlerden elde edilir.

Anahtar ifade çıkarımı yaklaşımında ise; kontrollü bir sözlük kullanmak yerine, kendi kendine metinden anahtar ifade seçimi kullanılmıştır. Bu yöntemde sözcüksel ve bilgi elde edimi tekniklerinden faydalanılmıştır. Burada eğitim verisi, anahtar ifade çıkarımının ayar parametresi olarak kullanılmaktadır. (Turney, 1999)

Hâlen açık kaynak kodlu yazılım olan KEA (Keyphrase Extraction Algorithm) algoritması ile İngilizce, İspanyolca ve Fransızca dilleri için yapılmış olan yazılımın sıkça kullanımına rastlamak mümkündür. Kea algoritması, bir metin üzerinde sözcüksel yöntemler kullanılarak aday anahtar ifadelerin belirlenmesi, her aday ifade için özellik değerlerinin hesaplanması ve ardından makine öğrenme yöntemlerini kullanarak aday ifadeler arasından uygun olanların seçimi adımlarından oluşur. Makine öğrenmesi yapılacak model ise, anahtar kelimeleri belirlenmiş metin dosyaları kullanılarak oluşturulan şemadan yararlanılarak oluşturulur.

KEA'nın etkililiğinde; makine öğrenmesi gerçekleştirilecek yapı oluşturulurken anahtar kelimelerin doğru belirlenmiş olmasının rolü büyüktür. KEA; basit ve etkili kullanım imkânı sunulmakla birlikte, Naïve Bayes makine öğrenme algoritması kullanılır. KEA'nın orijinal sürümüne "<http://www.nzdl.org/>" adresinden ulaşılabilir. Ancak KEA'da Türkçe için dil desteği bulunmamaktadır.

Bu projede KEA algoritma mantığı ve standart özellikleri kullanılarak Türkçe için anahtar kelime çıkarma amacıyla KEA-TR oluşturulacak, ardından standart özelliklere ek olarak Spread özelliği ilave edilerek KEA-SPR geliştirilecek ve oluşturulan bu iki algoritmanın performans karşılaştırması yapılacaktır.

Türkçe haber metinlerinden yararlanılarak anahtar ifade çıkarımı yapılmaya çalışılması sebebiyle, ikinci bölümde Türkçenin dil yapısından bahsedilmiştir. Bu kısımda aynı zamanda kelimelerin gövdeleme işlemine tabi tutulmasının gerekmesi sebebiyle gövdeleyiciler hakkında bilgi verilmiştir. Ardından üçüncü bölümde anahtar ifade konusunda gerek Türkçe için, gerekse diğer dillerde yapılmış çalışmalardan bahsedilmiş ve anahtar ifadelerin uygulama alanları ile ilgili bilgi verilecektir.

Dördüncü bölümde standart KEA algoritması, Türkçeye uygulanmış hali olan KEA-TR algoritması ve dağılım (spread) özelliği eklenerek oluşturulan KEA-SPR algoritması tanıtılmıştır. Beşinci bölümde gerçekleştirilen yazılım anlatılmış ve altıncı bölümde uygulanan test aşamasından söz edilmiştir.

2. TÜRKÇENİN DİL YAPISI

Türkçe haber metinleri üzerinde yapılacak bu çalışma için, öncelikle Türkçenin dil yapısından bahsedilecektir.

Türkçe; eklemeli dillerden olan Ural-Altay grubunun Altay kolunda yer alan eklemeli bir dildir. Bu dil ailesinde ünlü harf sayısı fazla olup, sözcük yapımı ve çekiminin, kelimenin sonuna eklenen ekler aracılığı ile yapıldığı görülmektedir. Yine belirgin diğer bir ortak özellik olarak, cümle yapısında yüklem sona olması göze çarpmaktadır.

Türkçenin; eklemeli dil yapısına sahip olması sebebiyle, kelimelerin çoğunlukla bir kök ve kökün sonuna eklenmiş ekler tarafından oluşturulan bir yapısı vardır. Ek'in bir kökün sonuna eklenerek yeni bir kelime türetilmesi durumunda, kelimeye farklı bir anlam kazandırmaktadır.

Bu çalışmada Türkçede yapısal bazda önem arz eden ve yapım ve çekim ekleri olarak ayrılan Türkçenin Ek Yapısından bahsedilecek, ardından metnin içeriğinin anlaşılmasını sağlayacak olan kompozisyonun temellerine değinilecektir.

2.1. Türkçenin Ek Yapısı

Diğer dillerde farklı kelimelerle ifade edilen anlam, Türkçede ekler ile ifade edilebilmektedir. Bu durum, eklemeli bir dil olmasının yanı sıra, yapım ve çekim eklerinin sayısının da oldukça fazla olmasından kaynaklanmaktadır.

Türkçe; eklemeli yapısı, kurallarının kesinliği ve bu kuralların uzun yıllardan beri korunmuş olması sebebiyle her zaman Doğal Dil İşleme ile ilgilenenlerin dikkatini çekmiştir. (Kesgin, 2007) Türkçede ekleri genel olarak yapım ve çekim ekleri olarak ayırmak mümkündür. Yapım eki; bir kelimenin sonuna eklendiğinde yeni bir anlam kazandırmakla birlikte, çekim ekinde yeni bir kelime türetilmesinden ziyade, kök

kelime ile türetilen kelime arasında teklik-çokluk, sahiplik, zaman, kişi gibi ilgiler kurulması sağlanır.

Normal sıralamada yapım eki, çekim ekinden önce gelmekle birlikte, yapım eklerinden sayılan “ki” eki bunun dışında tutulmaktadır. (Ağaçtaki) Kalıplaşmış çekim eklerinden sonra yapım eki geldiği durumlar da mevcuttur. (Gündelik).

Eklerin köke eklenmesiyle ilgili diğer bir husus ise, ünlü ve ünsüz kuralına göre gerçekleşmesidir. Bu kuralda, a ve ü ünlüleri kalınlık-incelik durumuna, ı, i, u, ü ünlüleri kalınlık-inceliğe ilaveten düzlük-yuvarlaklık durumuna göre de eklendiği kökle uyumludur. Sert ünsüzlerin (f,s,t,k,ç,ş,h,p) biriyle biten sözcüklere c,d,g, yumuşak ünsüzlerinden biriyle başlayan bir ek getirildiğinde, bu eklerin başındaki c, ç’ye d, t’ye g, k’ye dönüşür. Ünsüz uyumu kuralında, eğer bir sözcük sonunda p, ç, t, k harflerinden birini bulundurunken ek alması durumunda, son harfi yumuşayarak b, c, d, ğ harflerine dönüşür. (Kitap-kitabı, ağaç-ağacı, kağıt-kağıdı, balık-balığı)

Türkçede ekleri genel olarak yapım ve çekim ekleri şeklinde iki kısma ayırmak mümkündür.

2.1.1. Yapım ekleri

Yapım ekleri, kelimenin sonuna eklenerek kelimenin anlamını değiştirip yeni bir kelime türeten eklere verilen genel isimdir. İsim yapan yapım ekleri ve Fiil yapan yapım ekleri olarak iki başlık altında toplanabilir.

İsim yapan yapım ekleri ise; isimden isim yapan ve fiilden isim yapan olmak üzere iki kısma ayrılır. İsimden isim yapan yapım eklerinde, adından da anlaşılacağı üzere kök ismin sonuna ek getirilerek, yeni isim türetilir.

-lik, (başlık, kulaklık)

-ce, (karaca, kokarca)

-ci, (gazeteci, şakacı)

-li, (evli, nişanlı)

-siz, (akılsız, bilgisiz)

-ti, (tıkırtı, homurtu)

-ki, (geceki, bugünkü)

Ayrıca, -cil, -er, -cik, -nci, -leyin, -daş, -gil gibi eklerde mevcuttur (otçul, dörder, kızılılık, sonuncu, sabahleyin, vatandaş, teyzemgil).

Fiilden isim yapan ekler ise, fiil kök ve gövdelerinden isim türetirler. Bunlardan bazıları;

-mek, (yemek, ekmek)

-me, (bölme, kavurma)

-iş, (alış, çıkış)

-en, (gezen, bozan)

-esi, (kırılması, bakılması)

-mez, (geçmez, çıkmaz)

Ayrıca, -r, -dik, -miş, -gi, -ken, -tı (yazar, beklenmedik, dolmuş, saygı, atılğan, karartı) gibi ekler de örnek olarak gösterilebilir.

Fiil yapan yapım ekleri ise; isimden fiil yapan ve fiilden fiil yapan olarak iki kısma ayrılır. İsimden fiil yapan yapım eklerinde, adından da anlaşılacağı üzere kök ismin sonuna ek getirilerek, yeni fiilin türetilmesidir.

-le, (çınla, terle)

-el, (daral, düzel)

- l, (kısal, dođrul)
- e, (yařa, tüne)
- se, (garipse, önemse)
- mse, (benimse, özümse)
- len, (canlan, hoşlan)

Fiilden Fiil yapan ekler ise, bir fiilin sonuna ek getirildiđinde yeni bir fiilin türetilmesiyle oluşur. Bunlardan bazıları;

- t, (okut, ağlat)
- dir, (açtır, çizdir)
- n, (sürün, taşın)
- ř, (görüş, selamlaş)
- l, (yazıl, çizil)
- ele, (silkele, tepele)

2.1.2. Çekim ekleri

Eklendikleri kelimelerden yeni bir kelime türetmemelerine rağmen, kelimenin sahipliđi, çokluđu, zamanı, şahsı gibi nitelikleri üzerinden kelimelere işlerlik kazandırırılar. İsim çekim ekleri ve fiil çekim ekleri olmak üzere temelde 2 kısımda incelenirler.

İsim çekim ekleri;

İsimler ile diđer isim ve fiiller arasında durum ilgisini kuran durum ekleri (belirtme, bulunma, çıkma, eşitlik, ilgi, vasıta, yönelme),

Bir isme eklendiğinde, O ismin ait olduğu şahsı ya da nesneyi gösteren Sahiplik Ekleri (Birinci-İkinci-Üçüncü tekil kişi, Birinci-İkinci-Üçüncü çoğul kişi),

Bir isme eklendiğinde, O isme çokluk katan ve bazı durumlarda da topluluk ve saygı gibi anlamlar kazandıran Çoğul Eki (-ler, -lar),

Bir isme eklendiğinde, O isim ile ilgili soru oluşturulmasını sağlayan Soru Ekinden (-mi, -mu) oluşurlar.

Fiil çekim ekleri ise; Fiillerin kök veya gövdesine eklenerek, fiillerin zamanını, yapılış biçimini ve şahsını belirtirler.

Haber kip ekleri ve Dilek kip ekleri olarak sınıflandırılan Zaman ve Şekil ekleri,

Fiili şahsa bağlayan, fiildeki işi, hareketi yapmanı, oluşa ve duruma sahip olanı bildiren Şahıs ekleri,

Fiillerin kök ve gövdelerine gelerek, yapılmadığını, işin olmadığını belirten olumsuzluk eki,

Bir fiile eklendiğinde, genelde soru amaçlı kullanılabilen, bazen de soru anlamının dışında seçenek, koşul, şaşırma, korkutma gibi durumların bildirildiği soru eklerinden oluşmaktadır.

2.2. Türkçede Kompozisyon Yapısı

Ayrı ayrı parçaların, öğelerin bir araya getirilip uyumlu bir bütünün oluşturulmasına kompozisyon denilmektedir. (Tülek, 2007) Diğer bir deyişle her hangi bir konu veya alanla ilgili ayrı ayrı malzemeyi en uygun şekilde bir araya getirip, birleştirip düzenlemek olarak da tanımlanabilir. Fıkra, makale, sohbet, tenkit (eleştiri), deneme, roman, hikâye, masal, tiyatro, hatırat, seyahatname (gezi yazıları), mektup, dilekçe (eski: arzuhâl), şiir, nutuk, konferans, röportaj, hayat hikâyesi (biyografi) edebiyatla ilgili birer kompozisyonudur. (Korkmaz, 2001) Üzerinde söz söylenebilecek

veya yazı yazılabilecek bir duygu, bir düşünce, bir haber, bir sorun, bir eşya, bir olay kompozisyon için konu olabilir.

Bir kompozisyon yazımında önemli olan kısımlar; başlık, konu, paragraf planı olarak sınıflandırılabilir.

2.2.1. Başlık

Başlık, yazıda anlatılmak istenen fikir olarak tanımlanabilir. Tabiattaki bulunan her varlığın bir ada sahip olduğu gerçeğinden hareketle, oluşturulan yazının da bir başlığının olması gerektiği aşikârdır. Başlığın etkili olabilmesi için; kısa olması, ana düşünceyi ve konuyu etkili bir biçimde yansıtması, akılda kalıcı olması, merak uyandırıcı ve dikkat çekici olması önemlidir. Eğer kapsam bakımından geniş bakış açısını yansıtır, kitap, bölüm ya da yazının başında kullanılıyor ise büyük başlık, büyük başlıkların altında yer alan ve konuya özel bir kısımdaki başlıklara alt başlık ya da küçük başlık denilmektedir.

2.2.2. Konu

Bir yazıda ele alınan olay, düşünce, fikir ya da durum olarak tanımlanabilir. Hakkında yazı yazılabilecek düşünce, duygu, sorun, olay ya da haber bir kompozisyon için konu olabilir. Konu da başlık gibi bir yazının olmazsa olmazlarından. Anlatılmak istenen durumun, yazıya başlamadan önce sınırlarının çizilmesi ve konunun net tespit edilmesi önemlidir. Konunun üç önemli unsuru olarak;

Konunun incelenecek yönünü ve özünü oluşturan, açıklanmaya çalışılan şeyin cevabı olan *konunun maddesi*,

Konunun maddesini içine alarak işleneceği yönü belirleyen ve sınırlayan, aynı zamanda rastgele işlenmesine engel olan *konunun görüş noktası*,

Konunun maddesi belirlenip sınırlarının çizilmesinin ardından, işlenip geliştirilmesine *konunun şekli*

olarak listelemek mümkündür.

2.2.3. Paragraf planı

Bir yazıda bulunan, bir satır başı ile diğer satır başı arasında yer alan ve bir yardımcı düşünceyi işleyip geliştiren cümle ya da cümle grubuna verilen genel isime paragraf denir. Paragraflar; bir konunun ya da yazının belirli bölümünü ifade etmelerine rağmen, kendi içinde anlam bütünlükleri olan cümle gruplarıdır. Bir yazıda paragraf kullanılmaz ise okuyucu yorulur, yazıda anlatılmak istenen farklı ana fikirler de birbirine karışacağından istenildiği gibi nakledilemez. Bu yüzden iyi bir paragraf planı ile yazıda anlatılmak istenen şeyin karşı tarafa aktarılması daha kolay olur.

Bir paragrafta, öncelikle aktarılmak istenen düşüncenin özünü oluşturan bir cümle mutlaka bulunmalıdır. Temel cümle olarak adlandırılan bu cümle, doğrudan ya da dolaylı yollardan aktarılabilir, paragrafın özeti olarak değerlendirilebilir ve yazının değişik yerlerinde yer alabilir.

Temel düşüncenin yardımcı düşüncelerle uygun bir şekilde açılması, konu ve düşünce bütünlüğü sağlanması yardımcı fikirler ile sağlanır. Bu aşama; tanımlama, örnekleme, karşılaştırma, benzetme ya da tasvir gibi yöntemlerle gerçekleştirilir. Aynı zamanda, temel düşünceye zıt fikirler ele alınarak ve örnekler vasıtasıyla konu açılabilir.

Yine paragraf yazımında düşünceden sapılmaması, farklı ana fikirlere dağıtılmamasına paragraf birliği denilir. Konuda bahsedilmek istenen bir yön ele alınmalı, farklı bir yön ele alınacaksa yeni bir paragrafta başlanmalıdır.

Paragrafta işlenen konunun özelliğine göre, temel cümlelerin, yardımcı fikirlerin ve detayların mantıklı bir sırada yer almasına paragraf düzeni denilmektedir. Konuda

bahsedilen fikir, bakış açısı, tarz ve zaman, düzenlemede dikkat edilecek önemli unsurlar arasında yer alır.

Bir yazının yazılması esnasında, yazının özelliğine bağlı olarak beş çeşit paragraf uygulamak mümkündür;

Başlangıç paragrafı; her yazıda olmamakla birlikte genelde uzun yazılarda bulunan, yazıya girmeden önce doğrudan konuyla ilgisi olmayıp aydınlatma amaçlı uygulanan paragrafa verilen isimdir.

Giriş paragrafı; yazıyı okuyanı konuya hazırlamak, eğer bir olay yazısı ise yeri, ya da düşünceleri tanıtmak ve okuyucuyu okumaya yönlendirmek amacıyla kullanılan paragraftır. Bir anlamda yazının vitrini olarak düşünülecek bu paragraf iyi düzenlenmelidir.

Geçiş paragrafı; Uzun yazılarda kullanılan ve paragraflar veya bölümler arasında ilgi kurmak amaçlı kullanılmaktadır. Genelde iki paragraf arasında bir kopukluk olduğunda, aradaki ilgiyi göstermek amaçlı kullanımına rastlamak mümkündür.

Gelişme paragrafı, işlenen konuya bağlı olarak, bir düşünceyi geliştirmek, çeşitli yönleriyle açıklamak amaçlı kullanılır. Genelde giriş bölümünden sonra yer alır.

Sonuç paragrafı, üst paragraflarda anlatılan düşüncelerin özetlendiği ve ana düşüncenin altının çizildiği paragrafa verilen isimdir. Bazı durumlarda sonuca okuyucunun karar vermesi amacıyla bu paragraf kullanılmamakla birlikte, bazen de okuyucunun yanılmasını engellemek adına bu paragraf özellikle düzenlenebilmektedir.

2.3. Gövdeleme

Bir kök kelimenin sonuna yapım eki getirilip yeni kelimeler elde edilmiş haline gövde, kelimenin kendisine eklenmiş çekim eklerinden arındırılarak gövdenin ya da kökün elde edilmesine ise gövdeleme denilir.

Bilgi erişiminde bir kelime metin içerisinde ne kadar çok geçiyor ise, o kelimenin metnin konusu ile ilgili olması olasılığı da o derece yükselmektedir. Dolayısı ile metin içerisinde kelimelerin tekrar sıklığının sağlıklı bir şekilde tespit edilmesi oldukça önemlidir.

Kök veya gövde şeklinde bulunan kelimeler çekim eki aldıklarında yazılış şekilleri değişmelerine rağmen genellikle anlamlarında değişiklik olmaz. Aynı kelimenin farklı çekim eki aldığı durumlar göz önüne alındığında aynı anlamı göstermelerine rağmen farklı kelimeler olarak sayılacaklardır. Örneğin bir metnin içinde cümle kelimesinin; “cümle”, “cümle”, “cümle” şekilleriyle geçtiğini varsayalım. Eğer bir gövdeleyici uygulanmaz ise bu üç kelime aynı anlamı göstermelerine rağmen farklı kelimeler olarak değerlendirileceklerdir. Bu da terim sıklığının önemli olduğu bir anahtar ifade algoritmasında, sonuçların hatalı yansıyacağı anlamı taşımaktadır. Dolayısıyla kelimelerin tekrar sıklığını sağlıklı bir şekilde tespit etmek için, sağlıklı bir gövdeleyici kullanmak oldukça önemlidir.

Gövdeleme işlemi dillere göre farklılık göstermektedir. Örneğin analitik bir dil olan İngilizce için, ek kullanımının az olması sebebiyle sadece ekler sözlüğü vasıtasıyla gövdeleyici geliştirmek mümkündür. Türkçe gibi eklemeli dillerde; kök kelimelere birçok yapım eki, oluşturulan gövdeye ya da doğrudan kök kelimelere birçok çekim eki eklenme ihtimali dolayısıyla, iyi bir gövdeleyici geliştirmek daha karmaşık bir işlemdir. Biçimbirimsel çözümleme kullanan ve biçimbirimsel çözümleme kullanmayan olmak üzere iki tip gövdeleyici kullanılabilir.

Biçimbirimsel çözümleme kullanmayan gövdeleyiciler; sözcüğün ait olabileceği kökü bulmak için sondan başa doğru eksiltme yapılarak bir sözlükten araştırma yapılması temeline dayanır.

Biçimbirimsel Çözümleme kullanan gövdeleyicilerde; en basit ifadeyle bir sözcüğe ait olabilecek tüm kök ve eklerin bulunması olarak tanımlanabilir. Temel olarak dört temel adımdan oluşmaktadır.

Önişlemeden geçirme ile; sözcüğün çözümlene işlemine hazır hale getirilmesi için, küçük harfe çevrilme, “tire” karakterinin kaldırılması, aksanların çıkarılması gibi işlemler uygulanır. Dilin alfabesinde bulunmayan harf için işlem durdurulur.

Uygun kök seçici kullanımı ile, kök adayları tespiti yapılır.

Ekleme aşaması ile; tespit edilen her kök adayı için, hedefteki sözcük inşa edilene kadar gerekliyse özel durumları uygulayıp, herhangi bir ek seçeneği kalmayana dek olası eklerin köke eklenmesine devam edilir. Her ek üretiminin ardından, girdi sözcük ile karşılaştırma yapılır. Tam, ASCII toleranslı ve hata toleranslı sözcük karşılaştırıcılardan yararlanır. Eğer üretilen ek başarılı ise, sözcüğün yeni hali bir yığında tutulur. Eşleşen ek kalmaz ise işlem son aşamaya gelmiş olacaktır.

Son işlem aşamasında ise, girdi sözcükle eşleşen tüm kelimelere ulaşılmış olacaktır.

Biçimbirimsel çözümlene yöntemi esasına dayanan açık kaynak kodlu yazılım olarak Zemberek isimli yazılım oldukça sık kullanılmaktadır. Doğal Dil İşleme işlemlerini yapabilen Türkçe ve Türkî diller için yazılmış olup, yazım denetimi ve sözcük üretme gibi temel işlevleri gerçekleştirebilen bir yazılımdır. Kullanılan bu yazılım ile;

- Yazım denetimi,
- Biçimbirimsel çözümlene,
- Gövdeleme,
- Sözcük üretimi,
- ASCII karakterden Türkçe karaktere ve Türkçe karakterden ASCII karaktere dönüştürme,
- Heceleme,
- Sözcük Önerme işlemleri yapılabilmektedir. (Akın, 2001)

Zemberek kütüphanesi temel olarak iki kısımdan oluşur. Bunlar; dil yapı bilgisi ve Doğal Dil İşlemleri. Aslında çekirdek kütüphane Türkî diller için gerçekleştirilmiştir, ancak dillerden herhangi birine özel uygulama içermemektedir. Bir dilde bu bilgiler, genel olarak ve temel doğal dil işleme işlevleri üzerinden kullanılır. Yazılım ara yüzü

kolay ve kullanışlıdır. Zemberek’te dil geliştiriciler için, bazı dil verilerini harici metin dosyalarıyla kullanma imkânı sunulur. (Akın, 2001)

Zemberek’te harf ve alfabe bilgisi, kullanılacak her dil için ihtiyaç duyulan temel yapıtaşları olup, basit olarak bir metin dosyasında tutulmaktadır.

Ekler tüm Türkî diller için diğer önemli yapı taşını temsil edip, kümelenmiş olarak XML dosyasında depolanırlar. Eklerin ardından hangi ek ya da ek kümelerinin kullanılacağı, eklenme şekli, ek ağacı başlangıcı da bir dosya da tutulur. Kullanımında fayda sağlayan ön ekler ve yumuşama, sertleşme ve ünlü düşmesi gibi özel durumlar için özel durum üretim yöntemi de tutulmaktadır. Uygulamanın çalışması esnasında kullanılmak üzere, uygulama başlangıcında tüm ek dosyası ek nesnelere transfer edilir. (Akın, 2001)

Herhangi bir ek içermeyen sözcük, kök sözcük olarak adlandırılır. Zemberek’te desteklenen diller için kök sözcükler, sözcüğün tipi ve özel durumlar metin dosyasında tutulur. “gözlükçülerdeki” kelimesine ait Zemberek tarafından çözümlenmiş bir örnek çıktı, Tablo 2-1’de gösterilmiştir.

Çözümlenen satır (kök)	Çözümlenen satır (ekler)
{Icerik: gözlükçülerdeki Kok: gözlük tip:IS}	Ekler:ISIM_YALIN_BOS + ISIM_ILGI_CI + ISIM_COGUL_LER + ISIM_KALMA_DE + ISIM_BULUNMA_KI
{Icerik: gözlükçülerdeki Kok: göz tip:IS}	Ekler:ISIM_YALIN_BOS + ISIM_BULUNMA_LIK + ISIM_ILGI_CI + ISIM_COGUL_LER + ISIM_KALMA_DE + ISIM_BULUNMA_KI

Tablo 2-1 “gözlükçülerdeki” kelimesi için zemberek program çıktısı

Biçimbirimsel çözümlenme yapıldığında, hangi kökün ya da gövdenin seçilmesi gerektiği karşımıza diğer önemli problem olarak çıkmaktadır. Bu konu ile ilgili ortaya

çeşitli tezler sürülmüş olsa da, bazı durumlarda bu kararı insan gözüyle bile vermek mümkün olmamaktadır. Bazı durumlarda, cümlenin, paragrafın hatta metnin tamamının okunması bile gerekebilmektedir. Örneğin “kalemin” kelimesinin bir metinde geçtiğini varsayalım. “kale” ya da “kalem” olarak iki anlamsal olarak farklı kelime gerçek kök olabilir. Ancak bunun kararını, kelimenin geçtiği metni okumadan vermek çok güçtür. Bu tezde gövdeleyici olarak zemberek’ten yararlanılacağından, zemberek çıktıları içerisinde, karakter sayısı en fazla olan kelimenin kök kelime olarak seçilmesine karar verilmiştir.

3. ANAHTAR İFADELER İLE İLGİLİ YAPILMIŞ ÇALIŞMALAR ve ÇEŞİTLİ UYGULAMA ALANLARI

Genel ağın (internet) gelişmesi, her geçen gün üretilen sayısal bilgi miktarının da artması sonucunu doğurmaktadır. Bu da, anahtar ifadeler konusunda yapılan çalışmalar hakkında çeşitli gelişmelerin yaşanmasını sağlamıştır. Bu bölümde, hem Türkçe için hem de diğer dillerde yapılan çeşitli çalışmalardan bahsedilecektir. Ardından, anahtar ifadeler ile ilgili uygulama alanlarına değinilecektir.

3.1. Önceki Çalışmalar

Özellikle İngilizce için otomatik anahtar ifade çıkarımı konusunda birçok çalışmaya rastlamak mümkün olmakla birlikte, Türkçe için yapılmış çalışmaların sayısı oldukça azdır. Bu kısımda, önceki çalışmalar 2 ana başlık altında toplanacaktır.

3.1.1. Türkçe için yapılmış bazı çalışmalar

Türkçe için yapılmış çalışmaların sayısı oldukça azdır. Bahsedilecek çalışmalardan ilki “Turkish Keyphrase Extraction with KEA” isimli çalışmadır. (Pala, 2007) Bu çalışmada KEA algoritması Türkçeye uyarlanmış olmakla birlikte, Orijinal KEA’ya göre Türkçe kök bulucu ve yine Türkçe için durdurma listesi (stopwords) kullanılmıştır. Ayrıca bağıl uzunluk (relative length) adı altında yeni bir özellik eklenmiştir. Bağıl uzunluk değeri; bir aday ifadedeki karakter sayısının, en uzun karakterli aday ifadedeki karakter sayısına oranıdır.

Yapılan diğer bir çalışma; “Turkeyx: Turkish Keyphrase Extractor” isimli çalışmadır. (Kalaycılar, 2008) Bu çalışmada, KEA algoritmasında olduğu gibi yazılı derlem dosyası gerektirmez. TurKeyX yöntemi ad öbekleri ve ad öbek başlıklarını

istatistiksel olarak kullanılmaktadır. Aynı zamanda da KEA'dan ve Turney'in GenEx (Turney, 2000) algoritmasından bazı özellikler kullanılmıştır. (Ozdemir, 2009)

Bahsedilecek çalışmalardan sonuncusu "Turkish Keyphrase Extraction Using Multi-Criterion Ranking" isimli çalışmadır. (Ozdemir, 2009) Bu çalışmada, yukarıda bahsedilen ikinci çalışmaya benzemekle birlikte, farklı derecelendirme stratejileri uygulanmıştır.

3.1.2. İngilizce için yapılmış çalışmalar

Gerek İngilizce için gerekse Türkçe haricinde diğer dillerde otomatik anahtar ifade çıkarımı konusunda birçok çalışma yapılmıştır. Çeşitli araştırmacılar tarafından Eğitmenli (supervised) ve Eğitimsiz (unsupervised) anahtar ifade çıkarım yöntemleri sunulmuştur.

HaCohen-Kerner et al; eğitmenli öğrenme ve bunun birleşiminden oluşan bir model önermiştir. Özellik birleşimi için J48 algoritmasını ve türevi olan C4.5 karar ağaçlarını uygulamışlardır. (HaCohen-Kerner, 2003)

Hulth et al tarafından hiyerarşik bir şekilde eş anlamlı kelimelerin organize edilmesi ve frekans analizini birleştiren bir anahtar ifade çıkarım algoritması önerilmiştir. Frekans analizi ve eş anlamlıları birleştirmek için, tümevarımlı mantık programlama kullanılmıştır. GenEx ve KEA gibi makine öğrenme tekniklerinden faydalanılmaktadır. Ancak diğerlerinden farklı olarak, çıkarılan anahtar ifadelerin uzunluğunda bir sınır konulmamıştır. Bu sistemde kök bulmak için Porter algoritması kullanılmıştır. (Porter, 1980) Yapılan denemelerde, isim öbek aday ifadelerine uygulanan çeşitli tahmin modellerinin bir arada kullanılması ile olumlu sonuçlar elde edilmiştir. Hulth tarafından dört özellik kullanılmıştır. Bunlardan ikisi KEA'da tek özellik olarak kullanılan aday ifadenin doküman içerisindeki tekrarlama sıklığı (Term Frequency) ve Eğitim dokümanlarındaki kullanım sıklığını gösteren değerdir. (Inverse Document Frequency) Diğer özellikler ise ilk kullanım pozisyon bilgisi (First Occurrence) ve POS-tag değeridir. Hulth; POS-tag değerini özellik olarak kullanmanın

belirgin bir şekilde anahtar ifade çıkarım performansını arttırdığını göstermiştir. Ancak bu çalışmada KEA ve Extractor ile bu üç sistemin doğrudan bir karşılaştırması yapılmamıştır. (Hulth, 2003)

Y. Matsuo, Y. Ohsawa, M. Ishizuka tarafından grafik temelli yöntem denenmiştir. Bu yöntemde; doküman bir grafik, ifadeler düğüm şeklinde, kenarlar da terimlerin tekrarlanması şeklinde gösterilmiştir. Bir terimin anahtar ifade olup olmadığına, grafikteki dağılımı hesaplanarak karar verilmektedir. (Matsuo, 2001)

Sinir Ağları temelli yaklaşımda geleneksel tekrarlama frekansı (TFxIDF-Term Frequency Inverse Document Frequency) ve terimin ilk konum değeri kullanılmıştır. Bu yöntemle aday ifadeler; anahtar ifadedir ya da değildir diye eğitilirler. (Wang, 2005)

GenEx; P. Turney tarafından Kanada Ulusal Araştırma Konseyinde anahtar ifade çıkarımı için hibrit genetik algoritma olarak geliştirilmiştir. Genitor ve Extractor adı altında iki kısma sahiptir. Extractor; derecelendirilmiş Anahtar ifade listesini oluşturmak için, sembolik sezgisel dizi seti ile birleştirilir. Turney'in aday ifadeleri üç sözcüğe kadar olan sözcük gruplarıdır. Aday kelimeler budama işlemine tabi tutulur. İçerisinden filtreleme yapabilmek için, her biri frekansı ve metin içerisindeki konumuna göre derecelendirilerek değerler verilir. Adayların bu değerleri birden fazla kelime içeren ifadeler, bir kelime içeren adaylara göre daha az değere sahiptir. Aynı kayıtların elenmesinin ardından, gövdeleyici işlemi uygulanmış adaylar içerisinden en sık tekrarlananlar seçilir. Extractor, en yüksek derecedeki adayları listeler. Extractor 12 nümerik parametre ve bayrak değere sahiptir. Kararlı durum genetik algoritma Genitor, eğitim verilerinden en iyi parametre ayarlarını belirlemek için uygulanır. (Turney, 1999) Bu sistem geliştirildikten ve test edildikten sonra patenti alınmış, ticari kullanıma sunulmuştur.

KEA; Waikato Üniversitesinde makine öğrenmesi çalışmalarını yürüten bir araştırma grubu tarafından geliştirilmiş olup basit ve etkili bir metottur. Birinci kısımda metinsel olarak noktalama işaretleri, yeni satırlar ve numaralara göre sembolik olarak ayrılırlar. Anahtar kelime olamayacak kelimeler listesi burada durdurma listesi (stopwords) olarak ifade edilecektir. Her bir kısımda anahtar ifadeler, durdurma listesi dikkate alınarak belirlenir. Burada durdurma listesi elde ayrı bir liste şeklinde tutulur ve

aday anahtar ifadeler bu kelimelerle başlayamayacak veya bitemeyecek şekilde seçilir. Ardından Lovins kelime kökü bulma algoritması ile kelimelerin kökleri tespit edilir. (Lovins, 1968) Filtreleme aşamasında da KEA tarafından iki özellik için değer hesabı yapılır. Bu iki değer; TFxIDF ve ilk konum (First Occurrence) değerleridir. TFxIDF (Term Frequency Inverse Document Frequency) değeri; anahtar ifadenin tekrarlama sıklığı ile eğitim verilerindeki tekrarlama sıklığına oranıdır. Bunun seçilme sebebi; eğer bir ifade doküman içerisinde tekrar ediliyorsa ve bu tüm eğitim dokümanları içerisinde de sık kullanılmamışsa, konuyla ilgisi olabileceği kabul edilir ve anahtar ifade olma ihtimalini yükseltir. İlk konum değeri ise dokümanın başlangıcından itibaren kaçınıcı kelimedede ilk defa kullanıldığı bilgisidir. Buradaki mantık ise; kelimenin başlarda kullanılması konu hakkında fikir vermesi açısından önem arz etmesi ve anahtar kelime olma ihtimalini arttırmasıdır. Yine Kea tarafından bu iki özellik ve eğitim verilerinde dokümanın yazarı tarafından belirlenmiş anahtar ifadeler için Naïve Bayes algoritması kullanılır. (Domingos, 1997) Filtreleme aşamasında ise her aday ifade için olasılık hesabı yapılır. Aday ifadeler olasılıklarına göre sıralanır ve üst sıradaki değeri taşıyan n adet ifade anahtar ifade olarak belirlenir. Burada n, kaç adet anahtar ifadenin listede yer alması isteniyor ise O değeri işaret eder. (Frank, 1999)

Başka bir çalışmada da eğitmenli ve eğitmensiz olarak her iki şekilde kullanılabilir bir algoritma sunulmuştur. Algoritma'da her doküman için semantik bir ağ vardır. Bu ağların yapısal dinamikleri, Anahtar ifadeleri ayıklamak üzere anahtar düğümleri tanımlamak için kullanılabilir. Sunulmuş olan algoritma daha fazla dijital kitaplara yoğunlaşırken, diğer anahtar çıkarma algoritmaları daha fazla kâğıtlara ve web sayfalarına yoğunlaşmıştır. Deneyler, sunulan anahtar ifade çıkarım algoritmasının, Extractor ve KEA'ya göre karşılaştırıldığında eğitim gerektirmediğini göstermiştir. Yine bu çalışmalarda, dijital kaynaklarda daha iyi performans elde edildiği görülmüştür. Ancak kısa dokümanlarda, ifadelerdeki tekrar eksikliği nedeniyle iyi bir performans sergilenmemektedir. (Huang, 2006)

3.2. Bazı Uygulama Alanları

Anahtar ifadelerin birçok kullanım alanına rastlamak mümkündür. Bu kısımda yer alan 6 önemli kullanım alanına değinilecek olup, pratikte de yaygın kullanımına rastlanmaktadır. (Turney, 1999)

3.2.1. Üstverilerde (metadata) kullanım

İnternet ve kurumsal intranetlerin git gide büyümesi neticesinde, doküman yönetimi de her geçen gün zorlaşmaktadır. Bir çok araştırmacıya göre tanımlayıcı ifade olarak adlandırabileceğimiz “metadata” kullanımı ile, doküman yönetim problemlerine çözüm getirilebileceği düşünülmektedir. Bu tür dokümanların yönetimindeki en çok kullanılan çözümlerden biri olarak kullanılmaktadır.

Bir veritabanında bulunan hücre, satır ve sütunların tanımı, bu hücrelerde bulunan elementlerin fiziksel özellikleri teknik metadadır. Tanımlayıcı metadata ise, daha çok bu yazılımları kullanan kullanıcılarla ilgilenmektedir. Örneğin, aynı veritabanındaki bu satır ve sütunların anlamını veren doküman ise tanımlayıcı metadata olarak adlandırılır.

Aslında her bir üstveri standardında bahsi geçen tanımlayıcı ifadeler ile bu tezde anlatılan anahtar ifadeler arasında bir fark bulunmamaktadır. Üstverinin oluşturulmasını sağlayan bazı yazılımlar mevcuttur. Örneğin Microsoft Word ile üstveri bilgileri tutma imkânı bulunmakla birlikte, otomatik olarak anahtar ifade üstverisi oluşturmak da mümkündür. Şekil 3-1’de, Microsoft Word 2003 ile Dosya Özellikleri görüntülenmiş form gösterilmektedir. Bu form üzerinde, anahtar kelimeler otomatik olarak yer alır. Ayrıca araçlar menüsünden “Otomatik Özet..” seçeneği ile otomatik özet bilgisi oluşturmak da mümkündür. Ancak gerek anahtar kelimeler gerekse Otomatik Özet kullanımı, Türkçe için desteklenmemektedir.

Royal wedding.doc Özellikleri

Genel **Özet** İstatistikler İçindekiler Özel

Başlık: Royal wedding: William and Kate set for Abbe

Konu:

Yazar: Ayata

Yönetici:

Şirket:

Kategori:

Anahtar sözcükler: prince wed royal abbey middleton

Açıklamalar: Royal wedding
Royal Wedding timings
Prince greets fans before wedding

Köprü tabanı:

Şablon: Normal.dot

Önizleme resmini kaydet

Tamam İptal

Şekil 3-1 Microsoft Word 2003 ile otomatik özet formu

3.2.2. Vurgulama işlemlerinde kullanım

Bir dokümana hızlıca göz gezdirmek istendiğinde önce başlığa ardında da anahtar ifadeler bakılır ise, az çok dokümanda anlatılmak istenen konu hakkında fikir sahibi olunabilir. Burada anahtar kelimeler üzerinde vurgu yapılması pratik bir uygulama olup, konunun tamamının okunmasına gerek kalmayacak şekilde anlaşılması sağlanır. Vurgulama işlemi, ilgili kelime ya da ifadelerin koyulaştırılması, altına çizilmesi, farklı bir fontla ya da renkle gösterilmesi şeklinde olabilir.

3.2.3. Dizinlendirmede kullanım

Bir doküman ya da uzun bir dokümandan alınan anahtar ifadelerinin alfabetik listesi aslında bir dizinlendirme (index) olarak değerlendirilir. Özellikle canlı kütüphanelerde bulunan dokümanların çoğunda yazarı tarafından eklenmiş anahtar ifadeler bulunmamaktadır. Bu tür dokümanlara erişim süresinin kısaltılması için, öncelikle sınıflandırılma ve dizinlendirilme işlemleri kullanılabilir.

3.2.4. Etkileşimli sorgulamada

Arama motorlarının kullanımı yinelemeli bir işlemdir. Kullanıcı sorguyu girer, sonuçlar içerisinde değerlendirme yapar ve tekrar başka bir sorgu gönderir. Birçok arama motoru yinelemeli aramayı desteklemez. Bu uygulamadaki yaklaşım genellikle kullanıcının sorgusunun alınması, bu sorgunun ilk sonuçlarının anahtar ifadeleri belirlenerek listelenmesi şeklindedir. Ardından ilk sorgunun önerilmiş anahtar ifadeleri yinelemeli sorgularda kullanılır.

3.2.5. Web Log analizlerinde kullanım

Web sitesi yöneticileri sitelerinde ziyaretçilerin neleri aradığını bilmek isterler. Yine birçok sunucunun olay günlüğü kayıt etme özelliği ile; ziyaretçi bilgi kayıtları, istemci makinelere ait talep dosyaları ve bu taleplerin zamanları ile ilgili bilgiler tutulmaktadır. Bu olay günlüğü dosyalarını analiz edebilen bazı yazılım araçları mevcuttur. Genellikle bu araçlar ile, genel trafik durumu, popüler siteler ve dosyaların hangileri olduğu ile ilgili bilgi elde edilebilir. Bir web log analiz programı ile anahtar ifadelerden yararlanılarak trafiği daha detaylı inceleme imkânı bulunmaktadır. Aynı zamanda popüler dosyalardan ziyade hangi anahtar ifadelerin daha çok sorgulandığı

yüksek öneme sahiptir. Bu sayede web sitesi yöneticileri, sitelerinde hangi başlık ya da hangi anahtar ifadelerin daha çok sorgulandığı bilgisine ulaşılabilir.

3.2.6. Haber portallarında kullanım

Bazı gazetelere ait web sitelerinde ya da internet üzerinde hizmet veren haber sitelerinde “haber etiketleme” adı altında uygulanan yöntem de aslında bir anahtar ifade uygulamasıdır.

Bu yöntemle bir habere ait haberi karakterize eden anahtar kelimeler oluşturulmakta ve sayfanın içerisinde vurgulanmakta ya da sayfanın içerisinde başka bir form üzerinde listelenmektedir. Bu etiketler genelde önceden arşivlenmiş haberlerle ilişkilendirilmektedir. Dolayısı ile bir haberin içerisinde yer alan bir etikete tuşlandığında; arşivden o anahtar ifadenin yer aldığı haberlere ulaşılabilir.

Şekil 3-2’de Posta gazetesinin internet sitesinden alınan 17.05.2011 tarihli bir haber örneği gösterilmektedir. Bu habere ait etiketler de yine aynı sayfanın sağ alt köşesinde “Haberin Etiketleri” adlı kutucukta gösterilmiştir.

Sevgilisine kızdı, Yeşilirmak'a atladı

Amasya'da, sevgiliyle tartıştıktan sonra bunalıma giren 26 yaşındaki G.D., Yeşilirmak Nehri'ne atlayarak intihar girişiminde bulundu

17 Mayıs 2011 - 21:14

Yazı Boyutu: T T T

« Önceki Haber Sonraki Haber »

Genç kız, çevredekilerin çabasıyla kurtularak hastaneye kaldırıldı. Olay, Yalıboyu Mevkii'nde bugün akşam saatlerinde meydana geldi. Caddede üzerinde gezen G.D. ve erkek arkadaşı 29 yaşındaki A.T., henüz bilinmeyen bir nedenle tartışmaya başladı. Tartışma sonrası bunalıma giren genç kız, ağılayarak sevgilisinin yanından ayrılarak Yeşilirmak Nehri'ne atladı.

Yaklaşık 200 metre suyun üzerinde sürüklenen G.D., suya atlayan 3 kişi tarafından boğulmak üzereyken kurtarıldı. A.T.'nin ise olay yerinden kaçması ve ölümden dönen sevgiliyle ilgilenmemesi dikkat çekti.

Olay yerine gelen Sivil Sivil Savunma Arama ve Kurtarma Birliğine bağlı ekipler, ırmağa indirdikleri sedye ile genç kıza koyuna çıkarttı. Ambulansa bindirilen G.D. Amasya Şerefeddin Sabuncuoğlu Devlet Hastanesi'ne kaldırılarak tedavi altına alındı.

Genç kızın sağlık durumunun iyi olduğu belirtildi. Polisin olayla ilgili başlattığı soruşturmanın sürdüğü kaydedildi.

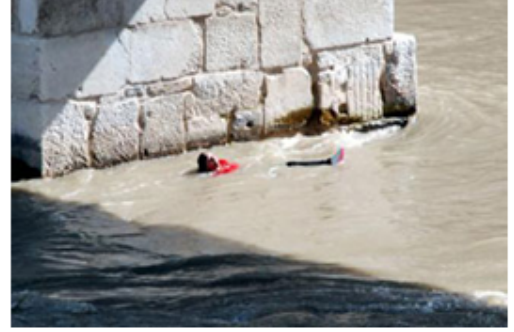
Savaş TUTAK/AMASYA, (DHA)



iPhone kullanıyor musunuz?
iPhone Posta uygulamasını indirin, tüm haberler ücretsiz, kolayca cebinize gelsin!

İlgili Haberler

- Doktor, kalfanın kulağını koparttı
- Kart borcu fuhuşa sürüklemiş
- Dönerlerde en dekonte: 2.5 lit



Tavsiye et

Arkadaşların arasında bunu tavsiye eden ilk sen ol.

Bu habere

Yorum yaz

Bu haberi

Arkadaşına gönder

Yazdır

Paylaş

facebook

Google

Yahoo

Myspace

friend feed

Mix

Digg

StumbleUpon

Del.icio.us

reddit

Twitter

Haberin Etiketleri

Amasya sağlık intihar cadde boğulmak

Şekil 3-2 Posta Gazetesi internet sitesinde 17.05.2011 tarihinde yayımlanan haber.

4. KEA-TR ve KEA-SPR ALGORİTMASI

Bu kısımda anahtar ifade çıkarımında KEA adlı projede nasıl bir algoritma kullanıldığından bahsedilip (Witten, 1999), ardından Türkçe'ye uyarlanmış şekli olan KEA-TR ve yeni özellik olarak eklenerek geliştirilen KEA-SPR'den bahsedilecektir. Normal anlatımda KEA'dan bahsedilecek, yeri geldiğinde Türkçe için yapılan değişiklikler ve geliştirilen KEA-SPR için uygulanan yöntemlere değinilecektir. KEA, temel olarak iki aşamadan oluşmaktadır.

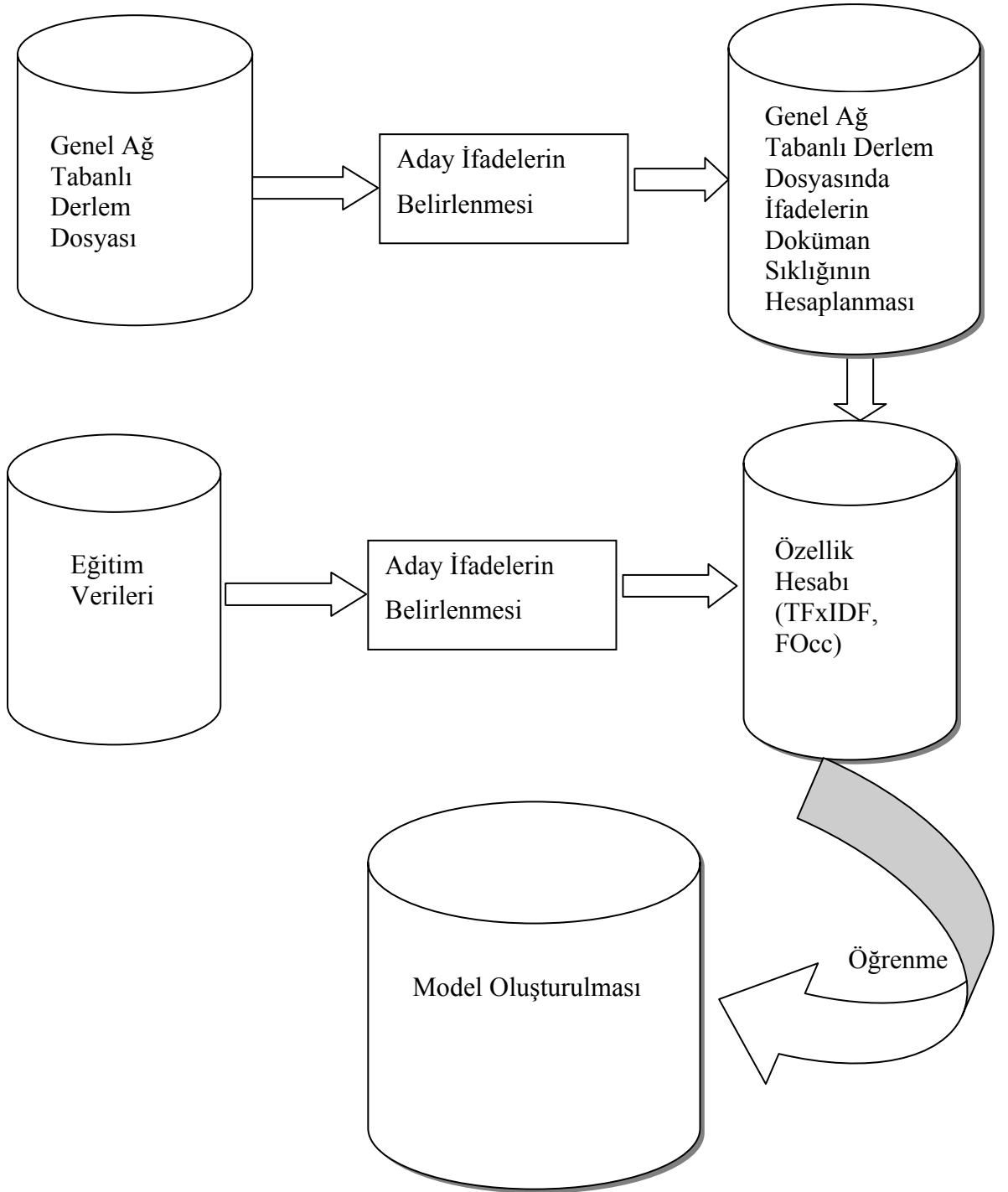
1- Eğitim (Training) : Yazarı tarafından anahtar ifadeleri işaretlenmiş eğitim dokümanları aracılığıyla bir model oluşturulur.

2- Çıkarım (Extraction) : Daha önce karşılaşılmamış bir dokümandan, eğitim aşamasında elde edilen model aracılığıyla anahtar ifade seçimi yapılır.

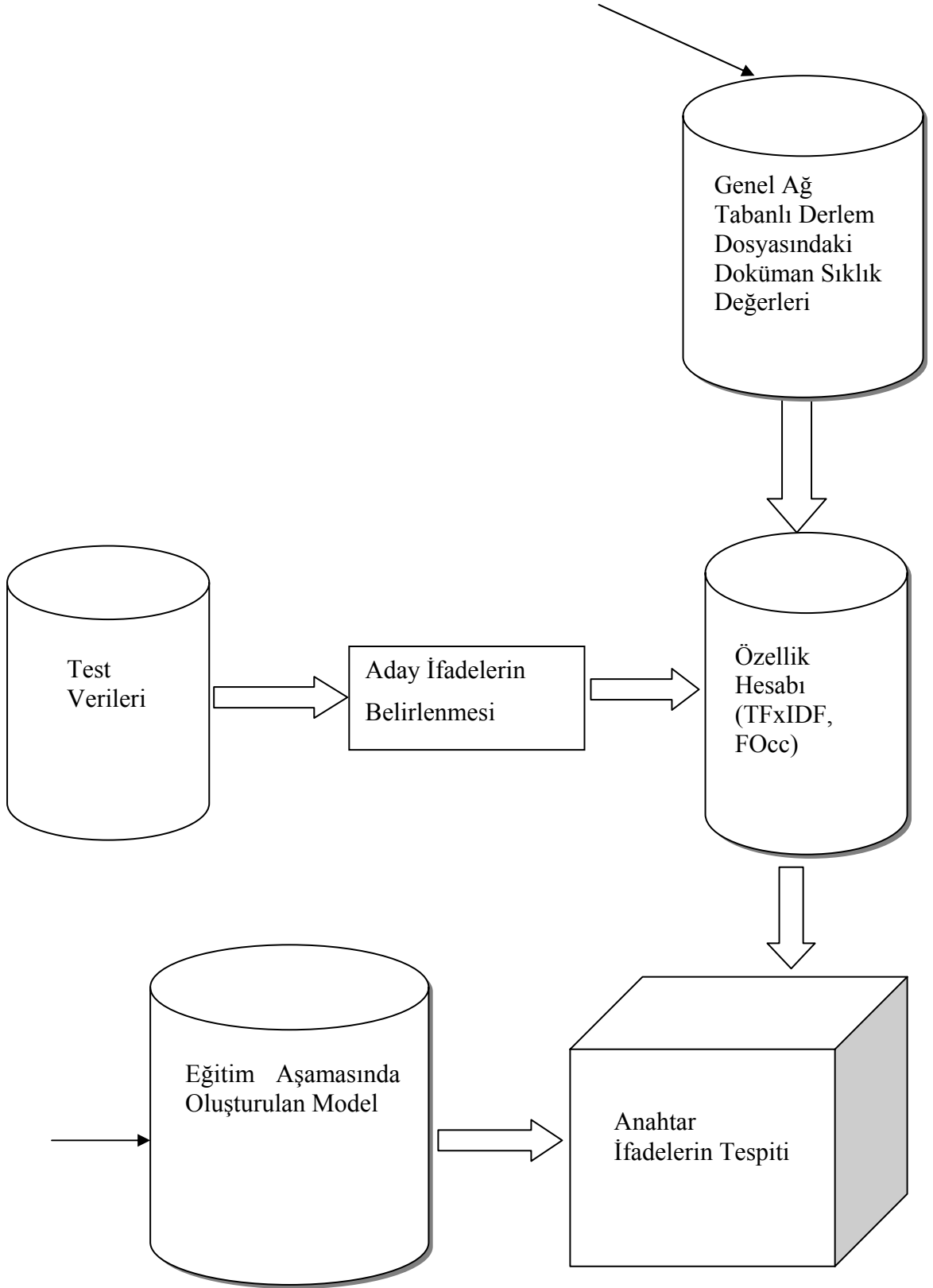
Bu aşamalar Şekil 4-1 ve Şekil 4-2'de gösterilmiştir. Eğitim aşaması aslında tek aşama gibi görünüyorsa da, eğitim model dosyasının işlenmeye başlanmasından önce genel ağ tabanlı derlem dosyaları vasıtasıyla ifadelerin terim sıklığı bilgisinin oluşturulması gerekmektedir. Bu işlem KEA'da 100 adet doküman kullanılarak oluşturulmuştur. Bu projede de, KEA-TR ve KEA-SPR için de 100 adet Türkçe doküman kullanılması tercih edilmiştir. Aday ifade seçimi ve özellik değerlerinin hesaplanması hem eğitim hem de çıkarım aşamalarında kullanılırlar. Bu sebeple öncelikle aday ifade seçimi ve özellik hesaplama adımlarından bahsedilecektir.

4.1. Aday İfade Seçimi Aşaması

Aday ifade seçimi temel olarak üç aşamada gerçekleştirilir. Bunlar Giriş dosyasının temizlenmesi, aday ifadelerin belirlenmesi ve aday ifadelerin büyük-küçük harf dönüşümlerinin sağlanarak gövdeleyici kullanarak ait olduğu kökün veya gövdenin belirlenmesi işlemi.



Şekil 4-1 Eğitim Aşaması



Şekil 4-2 Çıkarım Aşaması

4.1.1. Giriş dosyasının temizlenmesi

ASCII formatında oluşturulan giriş dosyasına bir takım filtreleme işlemleri uygulanır.

- Noktalama işaretleri, boşluklar ve rakamlar, ifade sınırı ile değiştirilir.
- Kesme işaretleri kaldırılır.
- Tire ile ayrılmış kelimeler iki kısma ayrılır.
- Geriye kalan anlamlandırılmayan ve harf içermeyen karakterler silinir.

Sonuç satırlardan oluşmalıdır ve her bir kısım en az iki harf içerecek şekilde olmalıdır.

4.1.2. Aday ifadelerin belirlenmesi

KEA'da temizleme işleminde elde edilen sonuca göre uygun aday ifadelerin seçimi gerçekleştirilir. Bu seçim işleminin kuralları genişletilebilmekle birlikte, bu çalışmada aşağıda sıralanan iki kural uygulanmıştır.

- Aday ifadeler üç kelime ile sınırlandırılır.
- Aday ifadeler durdurma listesindeki (stopwords) kelimeler ile başlayamaz ya da bitemez.

KEA'nın orijinal sürümünde aday ifadelerin özel isim olmama durumları dikkate alınmıştır. Bu özellik seçime bağlı olup, bu tezde uygulanan projede bu özellik kullanılmamıştır.

Aday ifadelerin belirlenmesi sırasında her giriş satırındaki bitişik ifadeler yukarıdaki kurala göre değerlendirilir ve sonucunda aday anahtar ifade listesi oluşturulur. Örneğin “Tahakkuk ve Tahsilat İşleri Müdürlüğü” ifadesinde Tahakkuk,

Tahsilat, İşleri, Müdürlüğü, Tahakkuk ve Tahsilat, Tahakkuk ve Tahsilat İşleri, Tahsilat İşleri, İşleri Müdürlüğü, Tahsilat İşleri Müdürlüğü aday ifadeler olarak seçilir. Aday ifade listesinin oluşturulması esnasında, eğer ifade durdurma listesinden bir kelime içermiyorsa yukarıdaki kural uygulanarak üç kelime ile sınırlandırılmıştır. Eğer içerisinde herhangi bir durdurma listesine ait kelime barındırıyorsa aday ifadedeki kelime sayısı dörde çıkarılmıştır.

Türkçe için durdurma listesindeki kelimeler (<http://www.ranks.nl/stopwords/turkish.html>, 2011)'dan elde edilmiştir. Bahsi geçen listedeki kelimeler 114 adet olup, bağlaç, zamir, sıfat gibi anahtar ifade olarak seçilmemesi gereken kelimelerden oluşmaktadır (acaba, bazı, bir, bu, v.b). KEA'nın orijinal sürümünde durdurma listesi 9 adet sözdizimsel gruba ait 425 adet İngilizce kelimedenden oluşmaktadır.

4.1.3. Büyük-Küçük harf dönüşümü ve gövdeleme

Son adım öncesinde, belirlenen aday ifadeler küçük harfe dönüştürülür ve bu ifadelerin ait olduğu kök ve gövdeler belirlenir. Kök veya gövde belirleme işlemi için zemberek isimli yazılımdan yararlanılmıştır. (<http://code.google.com/p/zemberek>) Bu işlem; kelimelerin ek almış farklı hallerinin aynı durumu ifade edip etmediğinin tespit edilmesine ve sonucunda da değerlendirmenin daha doğru yapılmasına olanak sağlar. Örneğin “başarı” ve “başarısı” kelimelerinin bir metnin içerisinde kullanıldığı varsayıldığında, aslında her ikisi tarafından aynı köke ait “başarı” kelimesinin kastedildiği tespit edilebilecektir. Eğer kök veya gövde belirleme işlemi uygulanmaz ise ayrı ayrı kelimeler olarak değerlendirilecekleri için yapılan işlem doğru olmayacaktır. Burada gerçekleştirilen işlemlerde genel ağ tabanlı derlem dosyalarında kullanılan anahtar ifadelerin de gövdeleyici işlemi uygulanmış şekilde oldukları varsayılarak değerlendirilmişlerdir. Aynı zamanda performans hesaplanırken de, ait olduğu kök ya da gövdesinin bulunmuş olduğu varsayılmıştır.

Bu işlemler yapılırken kullanıcıya gösterilmek üzere kelimelerin gövdeleme işleminden önceki hali de tutulmaktadır. Orijinal halinin tutulma sebebi, kelimenin metin içerisinde en sık kullanılan şeklinin kullanıcıya gösterilmesinin hedeflenmiş olmasıdır.

4.2. Özellik hesaplama

Geleneksel Makine Öğrenmesinde, elde edilen aday anahtar ifadeler kendi başlarına kullanışsızlardır ve bunlara ait bazı özelliklere ihtiyaç bulunmaktadır. İlk etapta bir ifade içerisinde kelime sayısı, karakter sayısı, ifadenin dokümandaki yeri, v.b gibi özellikler akla gelmektedir. KEA'nın orijinal sürümünde ve bu çalışmada kullanılan KEA-TR'de, eğitim ve çıkarım aşamalarında kullanılmak üzere temel olarak 2 özellik kullanılmıştır. Bunlar tekrarlama sıklığını gösteren TFXIDF değeri ve İlk konum değeridir. (First Occurrence) TFXIDF (Term Frequency Inverse Document Frequency) olarak adlandırılan değer; bir ifadenin metin içerisindeki tekrarlama sıklığı ile, genel ağ tabanlı derlemdeki kullanım değeri dikkate alınarak hesaplanır. İlk konum değeri ise; aday ifadenin metnin başlangıcından itibaren kaç kelimedenden sonra ilk olarak kullanımı ile ilgili bilgidir. Bu iki ifade şu şekilde hesaplanmaktadır;

Sıklık (TFxIDF) değeri;

Bu özellik ile aday ifadenin eğitim metni içerisindeki geçme sıklığı araştırılır ve yine aday ifadenin daha önce oluşturulmuş olan genel ağ tabanlı derlem dosyalarında kaç adet dokümanda ne sıklıkta kullanıldığına bakılır. Eğer eğitim dokümanında az bulunuyorsa, ifadenin anahtar kelime olarak belirlenme olasılığı yüksektir. KEA'nın İngilizce sürümünde genel ağ tabanlı derlem dosyası olarak 100 adet doküman kullanılmış olup, bu çalışmada gerçekleştirilen KEA-TR ve KEA-SPR için de aynı sayıda doküman kullanılması tercih edilmiştir. Tabii ki her aday ifadenin yukarıdaki bahsedilen yöntemler uygulanarak, kök ya da gövdelenmiş son şekilleri kullanılmıştır. KEA-TR de, KEA'da kullanılan özellikler aynı şekilde kullanılmış olup, KEA-SPR için ilave olarak dağılım (spread) özelliği de kullanılmıştır. Bu özellik daha önce bir çok farklı özellik ile

birlikte, Medelyan tarafından da kullanılmıştır. (Medelyan, 2009) Dağılım özelliğinin kullanılma sebebi olarak, bir dokümanda eğer bir ifade gerek baş tarafta gerekse sonuç paragrafında geçiyorsa, o metinde anlatılan konuyla ilgisi olma ihtimali artar. Dağılım değerinin bu katkısı düşünülerek özellik olarak ilave edilmiştir.

Bir eğitim dosyası ele alındığında; ifadeye P ve Dokümana da D denildiğinde Sıklık (TFxIDF) değeri;

$$TFxIDF = \frac{\text{freq}(P, D)}{\text{size}(D)} \times -\log_2 \frac{\text{df}(P)}{N}$$

şeklinde hesaplanır. Burada;

freq(P,D); D Dokümanında aday ifadenin sıklığı

size(D); D Dokümanındaki kelime sayısı

df(P); Genel ağ tabanlı derlem verisi içerisinde kaç adet dokümanda P ifadesinin geçtiği,

N; Genel ağ tabanlı derlem verisinin boyutunu ifade eder.

Eşitliğin ikinci tarafındaki yer alan kısımda; aday ifadenin genel ağ tabanlı derlem dosyasında bulunma olasılığının logaritmik değeri ile çarpılır. Burada negatif değer ile çarpılmasının sebebi olasılığın 1'den az olacağı düşünüldüğünde negatif değer alacağı için tekrar (-1) ile çarpılarak pozitif değere dönüştürülmesi hedeflenmiştir. Burada eğer doküman genel ağ tabanlı derlem dosyasında bulunmuyor ise, $\text{df}(P)$ ve N değerlerinin her ikisine birden 1 ilave edilir.

İlk Konum Değeri (First Occurrence)

İkinci özellik olan ilk konum (First Occurrence) değeri; aday anahtar ifadenin metnin başlangıcından itibaren kaç kelimedenden sonra ilk olarak görüldüğü araştırılır. Burada elde edilen sayı toplam kelime sayısına bölünür. Yine elde edilen değer 0 ile 1 arasında olacaktır.

KEA-SPR algoritmasına ilave edilen dağılım (spread) özelliği ise; aday anahtar ifadenin metnin ilk ve son rastlandığı konum bilgisi birlikte araştırılır. Burada bulunan değerlendirmede son konum değerinden ilk konum değeri çıkarılarak, metinde bulunan kelime sayısına bölünür.

Ayrıklaştırma (Discretization);

Yukarıda elde edilen değerlerin reel sayı olmaları sebebiyle makine öğrenmesi için nominal veriye dönüştürülmeleri gerekmektedir. Eğitim işlemi sırasında ayrıştırma tablosu her bir özellik için eğitim verisinden türetilmelidir. Bu tabloda her özellik için nümerik aralıklar ihtiva edilir ve eğitim verisindeki değerler bu nümerik aralıklara göre güncellenmelidir. Bu işlem için (Fayyad, 1993)'de tanımlanan eğitimli ayrıştırma metodu (supervised discretization method) kullanılır.

Model oluşturulması

Eğitim aşaması; anahtar kelimeleri bilinen dokümanlar kullanılarak gerçekleştirilmelidir. Tüm eğitim dokümanları için aday ifadeler belirlenir ve yukarıda bahsi geçen özellik değerleri hesaplanır. Burada eğitim setinin boyutunu sınırlandırmak amacıyla, dokümanda yalnızca bir defa geçen ifadeler göz ardı edilmiştir. El ile belirlenen anahtar ifadelerin her biri, “anahtar ifadedir” ya da “anahtar ifade değildir” şeklinde veri setinde güncellenir. Veri setinde kullanılan bu özellik makine öğrenmesinde sınıf özellik olarak kullanılır.

Bu şema vasıtasıyla özellikleri bilinen değerler temel alınarak sınıf özellik değerinin belirlenmesine yardımcı olacak bir model oluşturulur. Bu konuda değişik makine öğrenmesi şemaları mevcut olup, KEA'da basit oluşu ve olumlu sonuç vermesi nedeniyle Naïve Bayes tekniğinden yararlanılmıştır. Bu şemada yukarıda bahsedilen ayrıştırılmış değerlerden faydalanılarak iki nümerik ağırlıktan yararlanılarak öğrenme gerçekleştirilir. Anahtar ifade olanlar “yes”, diğerleri ise “no” olarak uygulanır.

Çıkarım (Extraction) Aşaması

Kea'da anahtar ifade seçiminde; öncelikle bir dokümandaki aday anahtar ifadeler belirlenir ve bunlara ait özellik değerleri hesaplanır. Ardından bu değerler öğrenme

yapılarak oluşturulan model dosyasına uygulanır. Bahsedilen modelden yararlanılarak, her aday ifade için toplamda anahtar ifade olma olasılığı hesaplanır. Ardından da bu aday anahtar ifadeler arasında en yüksek olasılığa sahip olanlar bir işlem dâhilinde seçilir.

Aday ifadeler içerisinden; Sıklık (TFxIDF) değeri için t , ilk konum (distance) değeri için d ve KEA-SPR'de uygulanan dağılım (spread) değerleri için s olarak varsayılın. Bu durumda Naïve Bayes modelinin uygulanması sonucunda aşağıdaki formüllere göre hesaplama yapılır;

KEA ve KEA-TR algoritmaları için;

$$P[\text{yes}] = \frac{Y}{Y + N} P_{TFxIDF} [t|\text{yes}] P_{\text{distance}} [d|\text{yes}]$$

$$P[\text{no}] = \frac{N}{Y + N} P_{TFxIDF} [t|\text{no}] P_{\text{distance}} [d|\text{no}]$$

Y ; eğitim dosyasındaki yazar tarafından belirlenen pozitif ifadelerin sayısını,

N ; eğitim dosyasında aday ifade olarak seçilmiş ama yazar tarafından seçilmemiş olan negatif ifadelerin sayısını ifade eder.

Sıfır olasılıklar için Laplace Estimator kullanılır. Basitçe ifade etmek gerekirse Y ve N , $Y+1$ ve $N+1$ ile değiştirilir.

Toplam olasılık ise;

$$p = \frac{P[\text{yes}]}{P[\text{yes}] + P[\text{no}]}$$

şeklinde hesaplanır.

KEA ve KEA-SPR algoritmaları için;

$$P[\text{yes}] = \frac{Y}{Y + N} P_{TFxIDF}[\text{t}|\text{yes}] P_{\text{distance}}[\text{d}|\text{yes}] P_{\text{spread}}[\text{s}|\text{yes}]$$

$$P[\text{no}] = \frac{N}{Y + N} P_{TFxIDF}[\text{t}|\text{no}] P_{\text{distance}}[\text{d}|\text{no}] P_{\text{spread}}[\text{s}|\text{no}]$$

Y; eğitim dosyasındaki yazar tarafından belirlenen pozitif ifadelerin sayısını

N; eğitim dosyasında aday ifade olarak seçilmiş ama yazar tarafından seçilmemiş olan negatif ifadelerin sayısını ifade eder.

Sıfır olasılıklar için Laplace Estimator kullanılır. Basitçe ifade etmek gerekirse Y ve N, Y+1 ve N+1 ile değiştirilir.

Toplam olasılık ise;

$$p = \frac{P[\text{yes}]}{P[\text{yes}] + P[\text{no}]}$$

şeklinde hesaplanır.

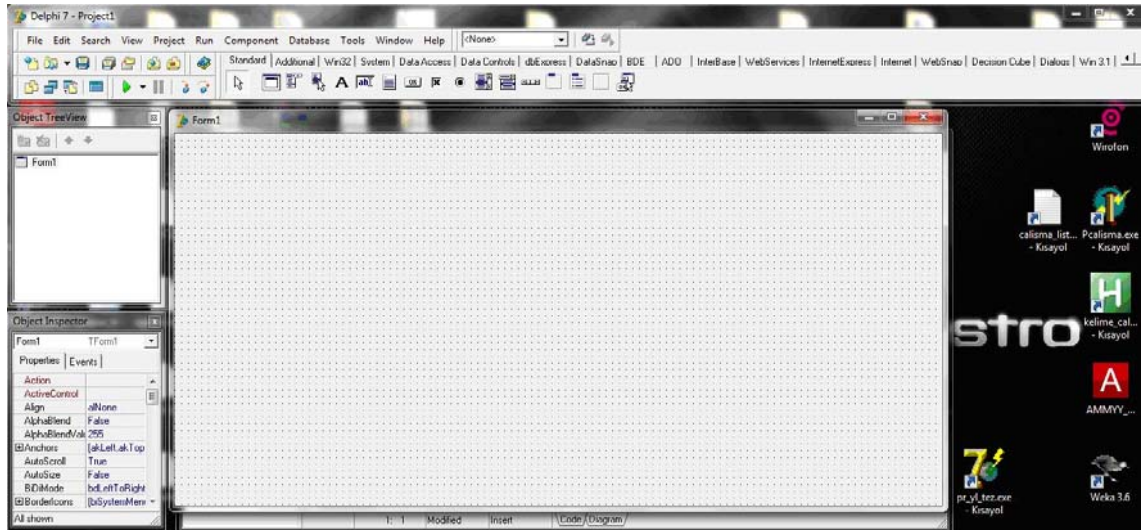
Aday anahtar ifadeler toplam olasılık değerine sıralanır ve sonuçlar üzerinde bazı işlemler uygulanır. Ayırıklaştırma işleminin yapılması nedeniyle toplam olasılığı eşit olan aday ifadelere sıkça rastlamak son derece doğal olacaktır.

Bu durumda TFxIDF değerinin ayırıklaştırma işlemi öncesi durumuna göre değerlendirme yapılarak seçim yapılmıştır. İkinci olarak eğer bir aday ifade daha yüksek olasılığa sahipse, o ifadenin alt aday ifadeleri listeden kaldırılır. Listenin kalan hali üzerinde; yüksek dereceden düşük dereceye doğru istenilen anahtar ifade sayısınca seçim yapılarak son olarak “n” adet anahtar ifade listesine ulaşılır.

5. YAZILIMIN GERÇEKLEŞTİRİLMESİ

Yazılımın geliştirildiği Delphi 7, Borland firmasına ait nesneye yönelik programların geliştirildiği kullanıcı dostu bir derleyicidir. Programcılara hem standart win-32 uygulamaları, hem de istemci/sunucu mimarisinde yazılım geliştirme imkânı sunulur. Temeli Pascal diline dayanmakla birlikte, özellikle nesne yönelimli programlama anlayışıyla yapılandırılmış, Turbo Pascal dilinin görsel sürümü olarak nitelendirilebilir.

Delphi kullanımında; formlar dâhil programa eklenen tüm nesnelere ait kodlar “Unit” ler içerisinde barındırılır. Çalışmada kullanılan Borland Delphi 7’ye ait yazılım arayüzü Şekil 5-1’de gösterilmiştir.



Şekil 5-1 Borland Delphi 7 programlama arayüzü

Türkçe haber metinleri üzerinde otomatik olarak anahtar ifade tespit çalışmasında kullanılmak üzere, internet üzerinden yayın yapan değişik haber siteleri üzerinden 54 adet haber metni elde edildi. Elde edilen bu haber metinlerini “.txt” uzantılı olarak kaydettikten sonra, her bir haber metni için anahtar ifadeler belirlendi. Belirlenen anahtar ifadeler aynı dosya adıyla fakat “.key” dosya uzantısıyla kaydedildi. (Ör: Dosya Adı : 1.txt, Anahtar İfade Dosya Adı : 1.key) “.key” uzantılı olarak oluşturulan bu dosyalar, eğitim verisini oluşturma amaçlı olarak kullanılacaktır.

Ardından ikinci olarak genel ağ tabanlı derlem (corpus) verisini oluşturmak amaçlı 100 adet haber metin dosyası internet üzerinden elde edildi. Bu veriler metin dosyası olarak “.txt” biçemi ile kaydedildi. Programın ileriki aşamalarında bu verinin nasıl oluşturulacağına tekrar değinilecektir.

Program, 4 aşamada gerçekleştirilmiştir. Bunlar;

- Giriş dosyasının temizlenmesi,
- Adayların belirlenmesi,
- TFXIDF, First Occurrence ve Spread değerlerinin tespiti,
- Anahtar ifade seçimi

aşamalarıdır.

Öncelikle eğitim verisi oluşturulacağı için elde edilen bu dosyalar ve el ile oluşturulan anahtar ifadeler vasıtası ile eğitim verisi oluşturulacaktır. Bunun için proje üzerinde “Eğitim Dosyasını Güncelle” isimli bir kontrol kutucuğu vasıtası ile bir kontrol mekanizması kurulmuştur. Eğer bir dosya eğitim dosyasını güncelleme amaçlı kullanılacaksa, program tarafından öncelikle dosya ile aynı isimli ama uzantısı “.key” olan anahtar ifade dosyasının varlığı kontrol edilecektir. Bahsi geçen kontrol kutucuğu programa bu amaçla eklenmiştir. Şimdi programdaki aşamalar ele alınacaktır;

5.1. Giriş Dosyasının Temizlenmesi

Öncelikle dosyanın aslının ve anahtar ifade dosyasının yüklenmesinin ardından, dosyalar ön işlemeden geçirilirler. Öncelikle asıl dosya üzerinde;

- Noktalama işaretleri, boşluklar ve rakamların bulunduğu yerler, ifade sınırı ile değiştirilir.
- Kesme işaretleri kaldırılır.

- Tire (-) ile ayrılmış kelimeler iki kısma ayrılır.
- Geriye kalan anlamlandırılmayan karakterler ve harf içermeyen izler silinir.

Şekil 5-2’de dosyanın aslı, Şekil 5-3’de ise önışlemeden geçirme işleminin sonrasında elde edilen çıkış dosyası görülmektedir.

Ünlü cerrah Tarık Minkari 85 yaşında hayata veda etti. Tıp dünyasında “virtüöz cerrah” olarak kabul edilen Minkari’nin mesleği ile ilgili pek çok kitabı ve makalesi yayımlandı. Ancak o yayının dünyasındaki asıl ününü emekli olduktan sonra 30 kitapta topladığı anı ve gezi yazılarıyla yaptı.

TIP dünyasında 18 tekniğiyle çığır açan efsane cerrah Tarık Minkari 85 yaşında hayatını kaybetti. Evliya Çelebi kulağına “Ahirette seyahat acentası yok, biraz da dünyayı tanı” diye fısıldayınca, emekliliğinin ardından dünyayı dolaşan ve yazdığı gezi kitaplarıyla “Cerrah Çelebi” ünvanını alan Minkari bugün Teşvikiye Camii’nde kılınacak öğle namazından sonra son seyahatine çıkıyor.

“İnsanlar ikiye ayrılır; doktorlar ve diğerleri. Doktorlar ikiye ayrılır; cerrahlar ve diğerleri. Cerrahlar ikiye ayrılır; genel cerrahlar ve diğerleri; genel cerrahlar da ikiye ayrılır; Tarık Minkari ve diğerleri...”

Tıp dünyasında “Minkari teknikleri” ile anılan Erol Tarık Minkari’nin mesleki başarısı, yakın dostu Gazeteci Aydın Boşan’ın bu tarifıyla ölümsüzleşir.

Şekil 5-2 Örnek bir eğitim dosyası

[Ünlü cerrah Tarık Minkari]yaşında hayata veda etti[Tıp dünyasında]virtüöz cerrah[olarak kabul edilen Minkari mesleği ile ilgili pek çok kitabı ve makalesi yayımlandı]Ancak o yayının dünyasındaki asıl ününü emekli olduktan sonra[kitapta topladığı anı ve gezi yazılarıyla yaptı]TIP dünyasında[teknikiğiyle çığır açan efsane cerrah Tarık Minkari]yaşında hayatını kaybetti[Evliya Çelebi kulağına]Ahirette seyahat acentası yok[biraz da dünyayı tanı]diye fısıldayınca[emekliliğinin ardından dünyayı dolaşan ve yazdığı gezi kitaplarıyla]Cerrah Çelebi[ünvanını alan Minkari bugün Teşvikiye Camii kılınacak öğle namazından sonra son seyahatine çıkıyor]İnsanlar ikiye ayrılır[doktorlar ve diğerleri]Doktorlar ikiye ayrılır[cerrahlar ve diğerleri]Cerrahlar ikiye ayrılır[genel cerrahlar ve diğerleri]genel cerrahlar da ikiye ayrılır[Tarık Minkari ve diğerleri]Tıp dünyasında[Minkari teknikleri]ile anılan Erol Tarık Minkari mesleki başarısı[yakın dostu Gazeteci Aydın Boşan bu tarifıyla ölümsüzleşir]

Şekil 5-3 Ön işlemeden geçirilmiş olan eğitim dosyası

Ön işlem den geçirme aşamasında bahsi geçen temizleme işlemleri yapıp, ifade sınırı olarak “|” işareti belirlenmiştir. Yine bu aşamada zemberek programına çözümlenecek olan dosya gönderilir ve çözümlenmiş hali bir veritabanı tablosunda tutulur. Burada yer alan kelimelerin gövdelenmiş durumlarını gösterir liste Tablo 5-1’de sunulmuştur.

Kelime_Orijinal	Kelime_Govdelenmis
ardından	art
asıl	asıl
ayrılır	ayrıl
açan	aç
başarısı	başarı
cerrahlar	cerrah
diğerleri	diğer
doktorlar	doktor
dolaşan	dolaş
dostu	dost
dünyasında	dünya
dünyasındaki	dünya
dünyayı	dünya
edilen	et
emekli	emekli
emekliliğinin	emekli
evliya	evliya
fısıldayınca	fısıll
gazeteci	gazete

Tablo 5-1 Gövdelenmiş Kelime Listesi

5.2. Adayların Belirlenmesi

Aday ifadeleri seçmek için, Şekil 5-3’de görülen işlenmiş metin üzerinde yer alan her bir ifade sınırı arasında kalan kelimeler için aşağıdaki işlemler uygulanır.

- Aday ifadeler üç kelime ile sınırlandırılır. (Eğer durdurma listesinden bir kelime içeriyorsa dört kelime ile sınırlandırılır)

- Aday ifadeler durdurma listesindeki kelimeler ile başlayamaz ya da bitemez.

Bir defaya mahsus olmak üzere durdurma listesinde bulunan kelimeler alınıp bir veritabanı tablosuna eklenmiştir. Bu aşamada öncelikle bu tablonun yüklenmesi adımı gerçekleştirilecektir. 114 adet kelimedenden oluşan ve program başlangıcında yüklenen tablonun örneği Tablo 5-2’de gösterilmiştir.

SıraNo	Kelime
1	acaba
2	altı
3	altmış
4	ama
5	bana
6	bazı

Tablo 5-2 Durdurma Listesindeki Kelimeler Örneği

Ardından her iki ifade sınırı arasında kalan kelimeler alınıp aralarındaki boşluk dikkate alınarak yukarıdaki iki kural dâhilinde aday anahtar ifade listesi oluşturulur. Bu liste örneği Tablo 5-3’de görülmektedir.

Aday_ifade	Aday_ifade_govdelenmis
Ünlü	ünlü
Cerrah	cerrah
Tarık	tarık
Minkari	minkari
Ünlü cerrah	ünlü cerrah
Ünlü cerrah Tarık	ünlü cerrah tarık
cerrah Tarık	cerrah tarık
cerrah Tarık Minkari	cerrah tarık minkari
Tarık Minkari	tarık minkari

Tablo 5-3 Aday ifade listesi

Tablo 5-3’de görüleceği üzere, aday ifade listesi oluşturulurken, öncelikle üç kelimeye kadar sınırlandırılması kuralı uygulanarak liste oluşturulur. Tabii ki, durdurma listesi ile başlayamaz ya da bitemez kuralına da dikkat edilir. Aday ifadenin gövdeleyici sonucunda elde edilen gövde kısmı (küçük harf dönüşümü yapılarak) ve ifadenin orijinal hali ile birlikte elde tutulur. Orijinal halinin son aşamada kullanıcıya orijinal şekliyle gösterilmesi amacıyla tutulur.

Bu kısımda aynı zamanda yine bir veritabanı tablosu oluşturularak, bu tablonun içerisinde, her aday ifadenin tekrar sayısı, ilk konum değer bilgisi ve orijinal KEA’da olmayıp bu projede kullanılan dağılım özelliğini (spread) hesaplamada kullanılacak olan son konum bilgisi değerleri tutulmaktadır.

Bu kısımda aday ifade listesinin oluşturulabilmesinin tamamlanması aşamasında, bu bölümün başında yer alan “genel ağ tabanlı derlem verisi” oluşturulmuştur. Bunun için programa ayrı bir komut düğmesi eklenmiştir. Eklenen bu düğme çalıştırıldığında,

öncelikle elde edilen 100 adet doküman dosyaları üzerinden, sırayla aday ifadeler oluşturulur. Bu aday ifade listesi oluşturulacak genel ağ tabanlı derlem dosyasında birleştirilir ve bu dosya veri tabanı tablosu olarak saklanır. Bu dosyada tüm dokümanlardaki aday ifadeler ve her aday ifadenin kaç adet dokümanda geçtiği bilgisi tutulmaktadır. Global yazılı doküman veri dosyası ile ilgili örnek Tablo 5-4’de gösterilmiştir.

Örnekten anlaşılacağı üzere listede kelimenin gövdeleme işlemine tabi tutulmuş hali yer almaktadır. Adet kısmında gövdelenmiş olarak kaç adet dokümanda geçtiği bilgisi tutulmaktadır.

kelime_govdelenmis	Kelime_asil	adet
aile	Aileye	14
ait ol	ait olduğunu	6
akıl	Aklınıza	19
al ceza	aldığı cezanın	4
al gerek	alması gerektiğini	5
alacak	Alacaklar	14
amaç	Amaçlarına	23
anla	Anladım	16
anlam	Anlamı	21
anlaşma	Anlaşmalarını	10
anlaşma benimse	anlaşmayı benimseyeceği	2
anlaşma sağla	anlaşma sağlanamadan	2
anlat	Anlatan	22
anne	Anne	14
araba	Arabaların	9
arka	Arkadan	23
arkadaş	Arkadaşına	26
Artık	Artık	26
avrupa	Avrupa	26
avrupa birlik	Avrupa Birliği	8
Ayak	Ayak	21
aynı zaman	Aynı zamanda	9

Tablo 5-4 Genel Ağ Tabanlı Derlem Dosyası Örneği

5.3. Sıklık, İlk Konum ve Dağılım Değerlerinin Tespiti

Bu aşamada, ilk olarak aday ifadenin tekrar sıklığı değerinden TFxIDF değeri hesaplanacaktır. Bu hesaplama için; aday ifadeye P ve Dokümana da D denildiğinde Sıklık (TFxIDF) değeri;

$$TFxIDF = \frac{\text{freq}(P, D)}{\text{size}(D)} \times \log_2 \frac{df(P)}{N}$$

freq(P,D); D Dokümanında aday ifadenin sıklığı

size(D); D Dokümanındaki kelime sayısı

df(P); Genel ağ tabanlı derlem verisinde kaç adet dokümanda P ifadesinin geçtiği

N; Genel ağ tabanlı derlem verisinin boyutu

değerleri ile hesaplanacaktır.

Ardından ilk konum değer bilgisi (First Occurrence) hesaplanmıştır. Bu değer; ilk konum bilgisinde elde edilen değer ile toplam kelime sayısı oranlanarak hesaplanmaktadır.

Son olarak bu tezde uygulanan aday ifadenin dağılım değer bilgisi hesaplanmıştır. Bahsi geçen bu değeri hesaplamak için aday ifadenin son konum bilgisi ile ilk konum bilgisi arasındaki fark alınarak, toplam kelime sayısına oranlama yöntemi kullanılmıştır.

Eğer “Eğitim Dosyasını Güncelle” kontrol kutucuğu işaretli değilse elde edilen üç sınıf için değerlerin alınıp, hangi sınıfa ait olduğu hesaplanır. Bu hesaplama için eğitim dosyasının oluşturulması esnasında her güncellemenin ardından, tüm özellikler için sınıflandırma değerleri ayrı ayrı tablolarda tutulur. Bu tablolardan yararlanılarak önceden karşılaşılmamış örnek için hesaplama yapılır.

sıra	aralık	baslangic	bitis	seviye
1	-inf-0.01375	1.000,00	0,01375	1
2	0.01375-0.04285	0,01375	0,04285	2
3	0.04285-0.1102	0,04285	0,1102	3
4	0.1102-inf	0,1102	1.000,00	4

Tablo 5-5 Sıklık değerinin sınıflandırılması

sıra	aralık	baslangic	bitis	seviye
1	-inf-0.01315	1.000,00	0,01315	1
2	0.01315-0.095	0,01315	0,095	2
3	0.095-inf	0,095	1.000,00	3

Tablo 5-6 İlk konum değerinin sınıflandırılması.

sıra	aralık	baslangic	bitis	seviye
1	-inf-0.33585	1.000,00	0,33585	1
2	0.33585-0.80745	0,33585	0,80745	2
3	0.80745-0.9427	0,80745	0,9427	3
4	0.9427-inf	0,9427	1.000,00	4

Tablo 5-7 Dağılım değerinin sınıflandırılması.

Eğer “Eğitim Dosyasını Güncelle” kontrol kutucuğu işaretliyse bu sonuç bilgileri oluşturulan eğitim veritabanı tablosuna aktarılacaktır. Elde edilen değerler sayısal olduğu için, birbirine yakın olan değerler bir çeşit sınıflandırma işlemine tabi tutulacaktır. Bu işlem ayrıklaştırma (discretization) olarak adlandırılır. Gerçekleştirmek için aşağıdaki adımlar sırayla uygulanır;

Makine öğrenme aracı olan ve sıkça kullanımına rastlanan Weka isimli araçtan faydalanılmıştır. Weka'nın filtreleme özelliğinin altında bulunan ve eğitimli öğrenmenin alt seçeneklerinde yer alan sınıflandırıcı (Discretize) özelliği seçilmiştir.

Oluşturulan bu filtre vasıtası ile yukarıda hesaplanan değerler (sıklık, ilk konum ve dağılım) sınıflandırma işlemine tabi tutulmuştur.

Sınıflandırmanın gerçekleşmesinin ardından, her üç değer için elde edilen sonuçlar Tablo 5-5, Tablo 5-6 ve Tablo 5-7’de güncellenmiştir.

Son olarak Elde edilen bu üç değer vasıtasıyla tüm olasılık hesabının yapılması aşaması gerçekleştirilecektir.

Aday ifadelerin Sıklık (TFxIDF) için t, İlk konum (distance) için d, dağılım (spread) için s değerleri varsayılmıştır. Naïve Bayes modeli uygulandığında iki değer aşağıdaki formüllere göre hesaplanır.

$$P[\text{yes}] = \frac{Y}{Y + N} P_{TFxIDF} [t|\text{yes}] P_{\text{distance}} [d|\text{yes}] P_{\text{spread}} [s|\text{yes}]$$

$$P[\text{no}] = \frac{N}{Y + N} P_{TFxIDF} [t|\text{no}] P_{\text{distance}} [d|\text{no}] P_{\text{spread}} [s|\text{no}]$$

Y; eğitim dosyasındaki yazar tarafından belirlenen pozitif ifadelerin sayısını

N; eğitim dosyasında aday ifade olarak seçilmiş ama yazar tarafından seçilmemiş olan negatif ifadelerin sayısını ifade eder.

Sıfır olasılıklar için Laplace Estimator kullanılır. Basitçe ifade etmek gerekirse Y ve N, Y+1 ve N+1 ile değiştirilir.

Toplam olasılık ise;

$$p = \frac{P[\text{yes}]}{P[\text{yes}] + P[\text{no}]}$$

şeklinde hesaplanır. Hesaplanan olasılık değerleri ile ilgili elde edilen sonuç örneği tablosu Tablo 5-8’de gösterilmiştir.

ifade govd	ifade asil	tfxidf	focc	spread	prob
anı	Anı	0,06	0,272	0,64	0,477
ayrıl	Ayrılır	0,053	0,677	0,162	0,387
cerrah	Cerrah	0,298	0,015	0,802	0,985
cerrah tarık	cerrah Tarık	0,098	0,015	0,346	0,716
cerrah tarık minkari	cerrah Tarık Minkari	0,098	0,015	0,346	0,716
cerrah ve diğer	cerrahlar ve diğerleri	0,098	0,728	0,059	0,387
doktor	Doktorlar	0,049	0,684	0,029	0,387
dünya	Dünyasında	0,057	0,074	0,809	0,534
emekli	Emekli	0,069	0,235	0,265	0,387
genel cerrah	genel cerrahlar	0,098	0,772	0,037	0,387
gezi	Gezi	0,064	0,287	0,257	0,387
hayat	Hayata	0,026	0,044	0,346	0,161
iki ayrıl	ikiye ayrılır	0,196	0,669	0,162	0,943
kitap	Kitabı	0,103	0,162	0,39	0,477
minkari	Minkari	0,343	0,029	0,904	0,968
seyahat	Seyahat	0,098	0,427	0,228	0,387
tarık	Tarık	0,196	0,022	0,904	0,968
tarık minkari	Tarık Minkari	0,196	0,022	0,904	0,968
tıp	Tıp	0,077	0,066	0,809	0,534
tıp dünya	Tıp dünyasında	0,147	0,066	0,809	0,968

Tablo 5-8 Tüm olasılık değerleri örneği

5.4. Anahtar İfade Seçimi

Aday anahtar ifadeler arasında tüm olasılık bazında, en yüksek olasılıktan en düşük olasılığa doğru sıralama yapılır. Ayırıklaştırma işleminin yapılması nedeniyle toplam olasılığı eşit olan aday ifadelere sıkça rastlamak son derece doğal olacaktır. Bu

durumda TFXIDF değerinin ayrıştırma işlemi öncesi değerine göre değerlendirme yapılarak seçim yapılmıştır.

İkinci olarak eğer bir aday ifade daha yüksek olasılığa sahipse, o ifadenin alt aday ifadeleri listeden kaldırılır. Listenin son durumu üzerinde; yüksek dereceden düşük dereceye doğru istenilen anahtar ifade sayısınca seçim yapılarak son anahtar ifade listesine ulaşılır. Örnek anahtar ifade listesi Tablo 5-9'da görülmektedir.

ifade_govde	govde_asil	prob	tfxidf
hayat	hayata	0,1601	0,0257
yaş	yaşında	0,1601	0,0333
doktor	doktorlar	0,3865	0,049
ayrıl	ayrılır	0,3865	0,0525
gezi	gezi	0,3865	0,0637
emekli	emekli	0,3865	0,0685
cerrah ve diğer	cerrahlar ve diğerleri	0,3865	0,0979
genel cerrah	genel cerrahlar	0,3865	0,0979
seyahat	seyahat	0,3865	0,0979
anı	anı	0,4769	0,0597
kitap	kitabı	0,4769	0,1029
dünya	dünyasında	0,5344	0,0574
tıp	Tıp	0,5344	0,0771
cerrah tarık	cerrah Tarık	0,7161	0,0979
cerrah tarık minkari	cerrah Tarık Minkari	0,7161	0,0979
yaş hayat	yaşında hayata	0,7161	0,0979
iki ayrıl	ikiye ayrılır	0,9426	0,1958
tıp dünya	Tıp dünyasında	0,9676	0,1471

Tablo 5-9 Anahtar İfade Son Seçim Örneği

6. KEA-SPR TEST AŞAMASI

Bu bölümde 2 aşamalı bir olarak gerçekleştirilen algoritmanın testi yapılacaktır. Birinci kısımda dağılım (spread) özelliğinin etkisi araştırılacak, ikinci kısımda ise aday anahtar ifade sayısı arttırılarak, insan eliyle oluşturulan aday anahtar ifadelerle karşılaştırılacaktır.

6.1. Spread Özelliğinin Doğrudan Etkisi

Bu testte Spread özelliğinin doğrudan etkisi araştırılmıştır. KEA algoritmasının Türkçe için standart KEA özellikleri kullanılarak geliştirilen şekli KEA-TR olarak adlandırılmış, bu tezde sunulan dağılım özelliği kullanılarak oluşturulmuş versiyonu ise KEA-SPR olarak adlandırılmıştır. İnsan eliyle anahtar ifadeleri oluşturulmuş 54 adet haber metni eğitim dokümanı, 12 adet haber metni ise test dokümanı olarak kullanılmıştır.

Test dokümanları incelendiğinde; ortalama olarak 10 adet anahtar ifadeye sahip olduğu tespit edilmiştir. Her doküman için önce standart KEA-TR algoritması ile oluşturulmuş yazılım ile, ardından da KEA-SPR algoritması ile oluşturulmuş yazılım kullanılarak aday anahtar ifadeler oluşturulmuştur. Otomatik olarak oluşturulan aday anahtar ifade sayısı, ortalama sayı olan 10 ile sınırlandırılmıştır.

Elde edilen sonuçlar Tablo 6-1'de gösterilmektedir. Görüldüğü gibi KEA-TR algoritmasında ortalama olarak %20'ye (23/115) karşılık gelen aday ifadenin aynı seçildiği, KEA-SPR'de bu oranın %25'e (29/115) yükseldiği görülmektedir. Spread özelliği kullanılarak oluşturulan algoritmanın performansı olumlu bir şekilde arttırdığı görülmektedir.

doc_adi	insan_eliyle	kea-tr	kea_spr
test_doc1	8	1	3
test_doc2	12	3	3
test_doc3	9	3	4
test_doc4	9	1	2
test_doc5	8	1	2
test_doc6	8	1	1
test_doc7	11	2	2
test_doc8	11	2	1
test_doc9	9	4	3
test_doc10	10	1	3
test_doc11	10	2	2
test_doc12	10	2	3
Toplam	115	23	29

Tablo 6-1 KEA ve KEA-SPR Algoritmaları karşılaştırma tablosu

6.2. Spread Özelliğinin Anahtar İfade Sayısı Artışına Göre Etkisi

Bu aşamadaki test işlemini gerçekleştirmek için, bir önceki aşamada kullanılan 54 adet haber dokümanı eğitim verisi, 12 adet haber dokümanı da test verisi olarak yine kullanılmıştır. Ancak bu testte insan eliyle oluşturulmuş anahtar ifadeler, KEA-TR ve KEA-SPR algoritmalarıyla oluşturulan aday ifadelerin sayısı artırılarak kıyaslanmıştır. Test neticesinde elde edilen sonuçlar Tablo 6-2’de gösterilmiştir.

Bu işlem gerçekleştirilirken öncelikle KEA-TR ve KEA-SPR algoritmaları için ortalama anahtar ifade sayısı olan 10’dan başlamak üzere 20 ve 30 adet için test gerçekleştirilmiştir. KEA-TR algoritması için elde edilen değerler sırasıyla %20, %29 ve %33 iken KEA-SPR için %25, %34 ve %40 değerleri elde edilmiştir. Yine bu aşamada da dağılım (Spread) özelliği dikkate alınarak geliştirilen algoritmanın performansı olumlu yönde arttırdığı görülmüştür.

doc_adi	insan eliyle	kea- tr 10	kea_spr 10	kea-tr 20	kea_spr 20	kea- tr 30	kea_spr 30
test_doc1	8	1	3	2	3	2	3
test_doc2	12	3	3	4	5	4	5
test_doc3	9	3	4	4	5	4	5
test_doc4	9	1	2	2	2	2	3
test_doc5	8	1	2	1	2	1	2
test_doc6	8	1	1	2	2	2	3
test_doc7	11	2	2	2	2	3	3
test_doc8	11	2	1	3	2	4	4
test_doc9	9	4	3	5	4	5	4
test_doc10	10	1	3	2	5	2	5
test_doc11	10	2	2	3	4	4	5
test_doc12	10	2	3	4	4	5	4
Toplam	115	23	29	34	40	38	46

Tablo 6-2 Anahtar İfade Sayısı Artışına Göre Karşılaştırma

7. SONUÇLAR ve TARTIŞMA

Dijital dünyada yaşanan hızlı gelişmelerin paralelinde, bilgi erişimi konusunda yapılan çalışmaların popülaritesi de giderek artmaktadır. Bu çalışmada bilgi erişimi kapsamına giren anahtar ifade çıkarımı konusu ele alınmıştır. Yine yerel ve ulusal medyada ve kullanılan teknolojilerde yaşanan gelişmelerin sürekliliği sebebiyle, günlük bilgi üretiminin oldukça fazla olduğu alanlardan biri olan Türkçe haber konusunu içeren bir çalışma gerçekleştirilmiştir.

Anahtar ifade çıkarımında öne çıkan önemli özelliklerden biri de, Türkçenin eklemeli dil yapısı sebebiyle kullanılacak olan gövdeleyicinin seçimidir. Tez kapsamında esas konunun gövdeleyici tasarlamak olmaması sebebiyle, açık kaynak kodlu bir gövdeleyici kullanılmıştır. Ancak başarıyı arttırmak adına özel bir gövdeleyici tasarlanabilir.

Sıkça kullanılan KEA algoritmasına farklı özellikler eklendiğinde başarımın arttığı görüldüğünden, başka yeni özelliklerin araştırılıp eklenmesi durumunda da, daha iyi performans elde edilmesi mümkün gözükmektedir.

Farklı kişilerin aynı metni değerlendirdiklerinde, farklı anahtar ifadelerle ulaşıyor olması bu konuda yapılan çalışmaları bir nebze zorlaştırır da, elde edilen sonuçlar, taranacak doküman miktarı düşünüldüğünde yine de olumlu bulunmuştur.

Elde edilen sonuçlar düşünüldüğünde bu çalışmanın, Türkçe haber metinleri için haberlerin konulara göre ayrımı, kümelenmesi, ilişkili haberlerin tespiti ile ilgili uygulamalarda kullanılmasının fayda sağlayacağı düşünülmektedir.

KAYNAKLAR

(2011, Jun) [Online] <http://www.ranks.nl/stopwords/turkish.html>

(2011, Jun) [Online] <http://code.google.com/p/zemberek/>

Akın, A.A. and Akın, M.D., 2007. Zemberek, an open source NLP framework for Turkic languages.

Domingos, P. and M. Pazzani (1997). On the optimality of the simple bayesian classifier under zeroone loss. *Machine Learning* 29(2/3), 103–130.

Fayyad, U.M. and Irani, K.B. (1993) “Multi-interval discretization of continuous-valued attributes for classification learning.” *Proc IJCAI'93*, 1022-1027.

Feather, J. and P. Sturges (1996). *International Encyclopedia of Information and Library Science*. London & New York: Routledge.

Frank E., Paynter G., Witten I. H., Gutwin C., Nevill- Manning C., Domain-specific keyphrase extraction. In *proceeding of the sixteenth international joint conference on artificial intelligence*, 1999, San Mateo, CA.

Gutwin, C., Paynter, G.W., Witten, I.H., Nevill- Manning, C.G. and Frank, E. (1998) *Improving browsing in digital libraries with keyphrase indexes.* Technical Report, Department of Computer Science, University of Saskatchewan, Canada.

HaCohen-Kerner Y., *Automatic Extraction of Keywords from Abstracts*, In V. Palade, R. J. Howlett, L. C. Jain (eds.): *KES 2003. Lecture Notes in Artificial Intelligence*, 2003, Vol. 2773, Springer-Verlag, Berlin Heidelberg, 843 – 849.

Huang, C., Tian, Y., Zhou, Z., Ling, C. X. and Huang, T. (2006) *Keyphrase Extraction using Semantic Networks Structure Analysis*. In *proceedings of Sixth International Conference on Data Mining (ICDM '06)*, Hong Kong.

Hulth, A. (2003) Improved Automatic Keyword Extraction Given More Linguistic Knowledge. In proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2003, pp. 216–223.

Johnson, F.C., Paice, C.D., Black, W.J. and Neal, A.P. (1993) “The application of linguistic processing to automatic abstract generation.” J Documentation and Text Management.

Kalaycilar F. and Cicekli I., “Turkeyx: Turkish keyphrase extractor,” in Proceedings of the 23rd International Symposium on Computer and Information Sciences (ISCIS 2008), Istanbul, Turkey, Oct 2008, pp. 1–4.

Kesgin, F., 2007. Türkçe metinler için konu belirleme sistemi, Yüksek Lisans Tezi, İ.T.Ü. Fen Bilimleri Enstitüsü, İstanbul.

Kılıçaslan Y., E.S. Güner, and E. Uçar, "Using Formal Concept Analysis to Improve Machine Translation", Proceedings of International Symposium on Computing & Engineering (ISCSE), Kuşadası, (2010).

Kılıçaslan, Y., E. S. Güner and S. Yıldırım, "Learning-Based Pronoun Resolution for Turkish with a Comparative Evaluation", Computer Speech and Language, Volume 23, Issue 3, 311-331, (2009).

Kılıçaslan, Y. and M.A.A. Tüysüz, "Bilgisayar Ortamında Türkçe'nin Semantik Analizi," Information Technologies Congress. Pamukkale University, May 2002, (2002).

Korkmaz, Z. ve diğ., 2001. Yüksek öğretim öğrencileri için Türk dili ve kompozisyon bilgileri, Yargı, İstanbul.

Lovins, J. B. (1968). Development of a stemming algorithm. Mechanical Translation and Computational Linguistics 11(1-2), 11–31.

Matsuo Y., Ohsawa Y., Ishizuka M., KeyWorld: Extracting Keywords from a Document as a Small World, In K. P. Jantke, A. shinohara (eds.): DS 2001. Lecture

Notes in Computer Science, 2001, Vol. 2226, Springer-Verlag, Berlin Heidelberg, 271–281.

Medelyan O., Automatic Keyphrase Indexing with a Domain-Specific Thesaurus, 2005, Master thesis. Albert-Ludwigs-Universitaet Freiburg im Breisgau, Germany.

Medelyan O., Frank E., Witten I.H. Human-competitive tagging using automatic keyphrase extraction, 2009, Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, pages 1318–1327, Singapore,.

Ozdemir, B., & Cicekli, I. (2009). Turkish Keyphrase Extraction Using Multi-Criterion Ranking. In Proceedings of the 24rd International Symposium on Computer and Information Sciences (ISCIS 2009). Northern Cyprus.

Pala N. and Cicekli I., “Turkish keyphrase extraction using kea,” in Proceedings of the 22nd International Symposium on Computer and Information Sciences (ISCIS 2007), Ankara, Turkey, Nov 2007, pp. 1–5.

Porter, M. (1980) An Algorithm for Suffix Stripping, Program, 14, 130-137.

Turney, P. (1999). Learning to extract keyphrases from text. Technical report, National Research Council Canada.

Turney P., “Learning algorithms for keyphrase extraction,” Information Retrieval, vol. 2, pp. 303–336, 2000.

Tülek, M., 2007. Türkçe için metin özetleme, Yüksek Lisans Tezi, İ.T.Ü. Fen Bilimleri Enstitüsü, İstanbul.

Wang J., Peng H., Hu J.-S., Automatic Keyphrases Extraction from Document Using Neural Network., ICMLC 2005, 633-641

Witten, I. H. Paynter, G.W. Frank et al, E. KEA: Practical Automatic Keyphrase Extraction, In E. A. Fox, N. Rowe (eds.): Proceedings of Digital Libraries'99: The Fourth ACM Conference on Digital Libraries. 1999, ACM Press, Berkeley, CA , 254 – 255.

ÖZGEÇMİŞ

Sürekli Adres
Şükrüpaşa Mahallesi Şehit Lütfü
Küçükyoğurtlu C. Buse Sit. F Blok D:6
Merkez / EDİRNE
Tel: (284) 236 40 40

E-posta: mustafacetingoz@trakya.edu.tr

KİŞİSEL BİLGİLER

Uyruğu
Doğum Yeri
Doğum Tarihi
Medeni Hali

EĞİTİM BİLGİLERİ

1988–1991

1991–1995

2008-...

İş Adresi
Trakya Üniversitesi Bilgi İşlem
Daire Başkanlığı
Balkan Yerleşkesi / EDİRNE
Tel: (284) 235 85 31 / 118

E-posta: bidb@trakya.edu.tr

T.C.
Beşiktaş / İSTANBUL
26.11.1974
Evli

I. Murat Lisesi, Edirne

İstanbul Üniversitesi Elektrik Elektronik
Mühendisliği(Lisans), İstanbul

Trakya Üniversitesi, Fen bilimleri
Enstitüsü, Bilgisayar Mühendisliği
(Yüksek Lisans), Edirne

EKLER

Örnek Eğitim Dokümanı-1

Zorunlu eğitim 13 yıla çıkıyor

18. Milli Eğitim Şurası'nda, zorunlu eğitim süresinin lise dahil 13 yıla çıkarılması, ilköğretim okullarında zorunlu eğitimin eskisi gibi 5+3 şeklinde kademeli olması, ortaöğretimde sınıf geçme yerine ders geçme sistemi getirilerek okulu erken bitirme olanağı verilmesi, haftalık ders saatlerinin azaltılarak teneffüslerin süresinin uzatılması konuları komisyon kararı olarak benimsendi.

İlköğretim ve Ortaöğretimin Güçlendirilmesi, Ortaöğretime Erişimin sağlanması” konularını tartışan komisyon, raporunu tamamladı.

8 yıllık zorunlu eğitimin tartışıldığı komisyonda, “Gelişim özellikleri bakımından farklı düzeylerdeki öğrencilerin bir arada bulunmasının ortaya çıkardığı pedagojik sorunların ortadan kaldırılması için ilköğretim okullarında 8 yıllık zorunlu eğitim, öğrencilerin yaş ve gelişim özellikleri dikkate alınarak kademelendirilmeli. Fiziksel mekanlar bu kademelere göre öğrencilerin ayrı alanlarda eğitim görmelerini sağlayacak biçimde bölümlendirilmeli” önerisi kabul edildi.

Komisyon ayrıca, zorunlu eğitimin süresinin lise dahil 13 yıla çıkarılması kararı da aldı.

Komisyonda dün tartışılan “Milli Güvenlik dersi müfredatı yenilenmeli, derse branş öğretmenleri girmeli” önerisi, “Milli Güvenlik dersi müfredatı yenilenmesi ve derse öğretmenlerin girmesi için yasal düzenleme yapılması” şeklinde değiştirildi.

DERS SAATLERİ AZALIYOR TENEFFÜS SÜRESİ ARTIYOR

Komisyonda, öğrencileri sevindirecek bazı maddeler de kabul edildi. Bu çerçevede, ortaöğretimde haftalık ders saatlerinin azaltılması, teneffüs süresinin uzatılması, ortaöğretimde sınıf geçme yerine ders geçme sistemi getirilerek okulu daha erken bitirmeye imkan sağlanması” gibi konularda görüş birliğine varıldı.

Kız öğrencilerin okullaşma oranlarının artırılmasının da tartışıldığı komisyonda “kız öğrencilerin ortaöğretime devamlarına ilişkin teşviklerin artırılması, yeni yatılı liseler açılması ve kız çocuklarının okula erişimi için pozitif ayrımcılık yapılması” benimsendi.

“Ortaöğretimde forma yerine kıyafetin serbest olması” konusunun da tartışıldığı komisyonda, Milli Eğitim Bakanlığı'nın serbest kıyafet uygulaması yönünde bir çalışması bulunduğu belirtilerek, karar alınmasına gerek bulunmadığı belirtildi.

SÖZLEŞMELİ ÖĞRETMENLİK

Öğretmenin Yetiştirilmesi, İstihdamı ve Mesleki Gelişimi konulu komisyonda da sözleşmeli öğretmenlik uygulaması nedeniyle yapılan tartışma gerginliğe yol açtı. Komisyonda, öğretmen adaylarına yönelik Kamu Personeli Seçme Sınavı'nda veya bununla ilgili yapılacak sınavlarda, “Öğretmenlik mesleğiyle ilgili genel kültür sorularının yanında mezun oldukları özel alanlarla ilgili sorular da yöneltilmesi” konusu karara bağlandı.

Komisyon ayrıca, “Öğretmenlerin istihdamında kullanılan kadrolu, sözleşmeli ve ücretli gibi farklı uygulamaların kaldırılarak tek bir istihdam modeline geçilmesi, bir perspektif plan çerçevesinde özendirici yöntemlerle personelin kademeli olarak sözleşmeli hale getirilmesi” maddesi de kabul edildi ve komisyon çalışması tamamlandı.

Bu maddenin komisyondan geçmesine karşı çıkan Türk Eğitim Sen Genel Başkanı İsmail Koncuk, “Tüm öğretmenlerin sözleşmeli hale getirilmesi” uygulamasını kabul etmeyeceklerini belirtti. Koncuk'un, “Bu yaptığınız haksızlık. Milli Eğitim Bakanlığı bürokratları ve özel öğretim kurumları yöneticilerinin görüşleri doğrultusunda rapor hazırlanıyor” sözlerine, Özel Öğretim Kurumları Genel Müdürü Mehmet Küçük “Artislik yapma” şeklinde tepki gösterdi.

Küçük ve Koncuk birbirlerinin üzerine yürürken, Öğretmen Yetiştirme Genel Müdürü Ömer Balıbey ile diğer komisyon üyeleri araya girerek, sakinleştirmeye çalıştı.

İsmail Koncuk, tartışmanın ardından yaptığı açıklamada, şunları söyledi:

“Sayın Bakan söz vermişti, sözleşmeli öğretmenlerin kadrolu hale getirileceği konusunda. Ancak öğretmenliğin kademeli olarak zaman içinde sözleşmeli hale

getirilmesi teklifini kabul ettiler. Biz de buna tepkimizi ortaya koydu. Biz ilgililerle görüşme yaptık, bunu Şura Genel Kurul toplantısı sırasında kaldırmaya çalışacağız. Öğretmenlerin haklarıyla ilgili bir karar veriyorsunuz burada öğretmen yok. Özel okulların yetkilileri, özel okulların müdürleri, akademisyenler benim öğretmenimin sözleşmeli öğretmen olmasına karar veriyor. Böyle bir demokrasi anlayışına saygı duymam. Benim bunları kınama hakkım var ve kınadım.”

Koncuk, tartışmanın ardından Mehmet Küçük ile görüştiklerini belirterek, “Bunu uzatmak bize yakışmaz. Görüştük, barıştık, hatta öpüştük. İşi düzelttik. Zaten bir yanlış anlama olmuş” dedi.

TÜM ÖĞRENCİLERE ENSTRÜMAN

Sanat ve Beceri ile ilgili oluşturulan alt komisyonda da ilköğretim sonuna kadar her öğrencinin bir enstrüman çalması önerisi yapıldı. İlköğretimde resim ve müzik ders saatlerinin yeni öğretim yöntem ve teknikleri dikkate alınarak artırılmasının istendiği alt komisyonda, ders dışı eğitim faaliyetlerinin artırılması ve “Sanat İnsanı Yetiştirme Projesi” hazırlanarak uygulamaya konulması önerildi.

18. Milli Eğitim Şurası'nda bugün öğleden sonra Genel Kurul çalışmalarına geçilecek.

Örnek Eğitim Dokümanı-1'in Anahtar İfadeleri: Zorunlu eğitim, İlköğretim, Ortaöğretim, ders saati, Milli güvenlik dersi, ders, Öğretmen, sözleşmeli öğretmenlik

Örnek Eğitim Dokümanı-2: İzmir'de karaciğer hastası Sadettin Kavruk, eniştesinden yapılacak nakil, acil başka bir ameliyat nedeniyle ertelenince, beyin ölümü gerçekleşen bir kişinin bağışlanan karaciğeriyle sağlığına kavuştu.

Aydın'ın İncirliova ilçesinde yaşayan Hepatit B hastası Sadettin Kavruk'un (56), rutin kontrolleri sırasında karaciğerinde kist tespit edildi. Sağlığına kavuşması için Kavruk'a karaciğer nakli yapılmasına karar verildi. Nakil için gönüllü olan Kavruk'un eşi ve kız kardeşinin, fazla kiloları nedeniyle donör olmaları uygun bulunmayınca, kız kardeşinin eşi Mustafa Kaplan, gönüllü oldu.

Dokuların uyumlu olduğunun tespiti üzerine Kavruk ve Kaplan, Doç. Dr. Murat Kılıç ve ekibi tarafından ameliyat edilmek üzere Kent Hastanesi'ne yattı. Hazırlıkları tamamlanan Kavruk ve Kaplan, ameliyat önlüklerini giyip ameliyathaneye alınacakları sırada, Doç. Dr. Kılıç'ın, hayati tehlikesi bulunan bir bebeğin ameliyatına girmesi gerekti. Bu nedenle, Kavruk ve Kaplan'ın ameliyatlarının bir sonraki güne ertelendiği açıklandı. Ertesi gün ameliyat hazırlıklarının tamamlandığı sırada, beyin ölümü gerçekleşen bir kişinin karaciğerinin Kavruk'a uygun olduğu bilgisi geldi. Nakil yeniden ertelenerek yapılan testler sonucu, Sadettin Kavruk'a, bu kişiden alınacak karaciğerin nakline karar verildi. Kavruk, Doç. Dr. Murat Kılıç, Doç. Dr. Murat Zeytinli, Prof. Dr. Mehmet Alper ve Op. Dr. Cahit Yılmaz'dan oluşan Kent Hastanesi Karaciğer Nakli Ekibince gerçekleştirilen operasyonla sağlığına kavuştu.

“HEM ŞAŞIRDIK, HEM SEVİNDİK”

İki kez ameliyat masasından dönen bağışçı Kaplan, durumunu sormak için telefonla aradığı eniştesinin, “Ciğer lazım” demesi üzerine, “Ben veririm” yanıtı verdiğini belirtti. Yengesi ve eşinin nakle uygun bulunmaması nedeniyle şakanın gerçeğe dönüştüğünü anlatan Kaplan, “Seve seve hastaneye yattım. Pazarcılıkla uğraştığım için işimi uzun süre engelleyecekti ama yine de gönüllü oldum. İki kez engel çıktı. 'Demek ki nasip değilmiş' dedik. Bize de böyle ilginç bir hikayeyi anlatmak düştü” dedi.

Eniştesinin karaciğerini beklerken, organ bağışıyla hayata tutunan Kavruk ise yaşanan gelişmeler üzerine şaşkınlık ve mutluluğu bir arada yaşadıklarını söyledi.

Kaplan'ın fedakarlığının kendisini mutlu ettiğini vurgulayan Kavruk, şöyle konuştu: “Eniştemin seve seve karaciğerinden vermek istemesine seviniyordum ama bir yandan da benim yüzümden bıçak altına yatacak olmasından endişeliydim. İşlerine bir süre mani olacaktı. Birden böyle bir haber gelince ne kadar sevindim anlatamam. Ben de daha hasta olmadığım zamanlar, bundan 20 yıl önce tüm organlarımı bağışlamıştım. Bir gün hepimizin başına böyle bir şey gelebileceğini diye düşünmek lazım. Allah organ bağışçı yapanlardan razı olsun. Ne kadar mutluyum anlatamam.”

Örnek Eğitim Dokümanı-2'nin Anahtar İfadeleri: organ nakli, ameliyat, karaciğer nakli, beyin ölümü, operasyon, organ bağışçı.

Örnek Eğitim Dokümanı-3: Bilkent Üniversitesi Teknoparkı'nda görevli Türk mühendisler, özellikle sayısal derslerin rahat öğrenimi için kişilerin kapasitesine uygun rota belirleyen yazılım teknolojisi geliştirdi.

ABD'de benzeri 500 bin kullanıcıya ulaşan bilgisayar yazılımı sayesinde öğrenciler, görsel, işitsel içerikli pek çok alternatiften kendi öğrenme profillerine uygun olanı seçerek ders çalışabiliyor.

Bilkent Cyberpark'da faaliyet gösteren ve tamamı Türk mühendislerden oluşan Arages Bilişim Şirketi Genel Müdürü Can Gürses, AA muhabirine yaptığı açıklamada, TÜBİTAK desteğiyle yürüttükleri akıllı öğretim sistemi projesi olan Bilgi Uzayına Adapte Edilebilir Kişisel e-öğrenme Yolu (Bilirkişi) adlı bilgisayar yazılımı uygulamasını başarıyla tamamladıklarını bildirdi.

Gürses, ``Bilirkişi``nin Türkiye'de ilk kez literatürde ``Knowledge Space Theory`` olarak geçen matematiksel-bilişsel bir teorinin eğitim alanına uygulanmasına yönelik web tabanlı bir bilgisayar programı olduğunu belirtti. E-öğrenmenin, günümüzde sadece eğitim sektörünün değil, iş hayatının da vazgeçilmez bir parçası olmaya aday olduğunu vurgulayan Gürses, e-öğrenme alanındaki gelişmelerin ve çözümlerin başarıyı hedefleyen her kurum ve birey için bir ihtiyaç olduğunu söyledi.

``KONUNUN TAMAMI ÖĞRENİLİYOR``

Geliştirdikleri yazılımın bir konunun öğrenilmesinin matematiksel modellenmesini sağladığını ve konuyu öğrenmek isteyen herkese en uygun öğrenme yolunu belirlediğini ifade eden Gürses, teknolojiye ilişkin şu bilgileri verdi: ``Yazılım, öncelikle matematik, tıp gibi sayısal ağırlıklı derslerin ya da satranç gibi oyunların öğrenme uzayını ortaya çıkartan bir altyapı sunuyor. Yani bir konuda hiçbir şey bilmeme durumundan her şeyi bilme durumuna kadar, olası tüm bilgi düzeyleri ve bu bilgi düzeylerini birbirine bağlayan öğrenme yollarının bütününe kapsıyor.

Program, kişileri konunun bilgi uzayının ortaya çıkartılmasının sonrasında, kendilerine en uygun öğrenme yolundan ilerletiyor ve konunun tamamını öğretiyor. Program, bir konuyu kullanıcılara soru-cevap yöntemiyle ve kullanıcılardan elde ettiği veriler sayesinde herhangi bir konunun zorluk derecesi, optimum öğrenme zamanı gibi oldukça

önemli verileri ortaya çıkarıyor. Teknoloji, öğrencinin konuyu anlayıp anlamadığını da matematiksel hesaplarla ölçüyor ve anlayana kadar ileri bilgiler sunuyor.

Sonrasında da konunun öğrenilip öğrenilmediğini test ediyor.`` Gürses, programlarının ayrıca, bir konuyu öğrenen kullanıcılar arasında belirli tip öğrenme stiline sahip olanları ortaya koyduğunu, başarıya veya başarısızlığa götüren yolları tespit de ettiğini belirterek, ``Sistemimiz bu yönüyle oldukça önemli ve eğitim-öğretim sistemini temelden etkileyebilecek veriler sunuyor`` dedi. ALES, KPSS, ÖSS gibi sınavlara hazırlık aracı olarak da kullanılabileceğini kaydetti.

``ABD`DE 500 BİN KİŞİ KULLANIYOR``

Uygulamanın matematiksel modelinin 1980`lerde Fransız ve Belçikalı matematikçiler tarafından ortaya konulduğunu, modelin ilk uygulamasının ise ABD`de 1990`lı yıllarda yapıldığını aktaran Gürses, bu ülkedeki kullanıcı sayısının da 500 bine ulaştığını bildirdi. Gürses, yazılımın ileriki hedeflerini ise, ``Öğrencilerin öğrenme stil ve tercihlerini belirliyoruz. Belirli tipteki öğrenme stratejisine sahip insanların, belirli öğrenme yolları , kişileri başarısızlığa ulaştıran sebepler neler? Bu gibi verileri çalışmamızın ileriki aşamasında istatistiksel olarak da göstereceğiz`` sözleriyle anlattı.

Örnek Eğitim Dokümanı-3`ün Anahtar İfadeleri: Teknopark, öğrenme profili, yazılım, akıllı öğretim sistemi, proje, e-öğrenme, bilgisayar, program, öğrenme, konu, model, öğrenci

Örnek Eğitim Dokümanı-4: MEB`in, okullarda cep telefonu kullanımını velilerin iznine tabi tutmasına rağmen uzmanlar özellikle 10 yaş altı çocuklarda dikkatsizlik, uykusuzluk ya da stres nedeniyle yasaklanmasından yana.

Ailelerin güvenlik kaygısı cep telefonunu ilköğretime kadar soktu. Uzmanlar, MEB genelgesiyle yasak olmasına rağmen velilerin izniyle ilköğretimde cep telefonu kullanmaya başlayan çocukların tehlike altında olduğunu söyledi.

İKİ KAT DAHA ZARARLI

Cep telefonu kullanan ilköğretim çağındaki çocuklarda dikkatsizlik, uykusuzluk, stres, geç algılama gibi rahatsızlıkların oluşabileceğini ifade eden Akdeniz Üniversitesi Elektrik-Elektronik Mühendisliği Bölümü Başkanı Doç. Dr. Şükrü Özen, gelişim çağında oldukları için çocuklardaki beyin ve kulak yapısının da yetişkinlere oranla en az iki kat etkili olduğunu kaydetti. Kemik ve beyin yapısının 20 yaşına kadar gelişimini sürdürdüğünü vurgulayan Doç. Dr. Özen, "Bu süreden önce bilinçsiz telefon kullanımı, buna bağlı beyin tümörlerini de arttırabilir" dedi.

Cep telefonları üzerinde teknik ölçümlerle araştırma yapan ve on yılı aşkın bu alanda yürüttüğü araştırmalarla tanınan Özen, 16 yaşın altındaki çocukların cep telefonu kullanmamaları yönünde çok sayıda araştırmanın olduğuna dikkat çekerek, "İlkokul çağındaki çocukların kullanmaması gerekir. Buna karşın 10 ve 8 yaş kritiktir. 10 yaş altı için günlük 10 dakika kullanıma güvenlik kaygısı durumunda izin verilebilir" dedi.

GELİŞİMİ OLUMSUZ ETKİLİYOR

Psikiyatr Dr. Armağan Samancı ise ilköğretim çağında cep telefonu kullanımının gelişime olumsuz etki edebileceğini söyledi. Fiziksel gelişim gibi kişisel gelişimde çevreye bağlı olduğunu kaydeden Samancı, "Bireyin karakter gelişimini çevresi etkiler. Çevre derken burada kast edilen insan etkisidir. Bir çocuğun gelişimi için insanlarla olan ilişkisi çok önemlidir. Telefon gibi iletişim araçları insanlarla birebir karşılıklı iletişim kurmayı engellediği için çocuğun gelişimini olumsuz yönde etkiler. Yerinde kullanıldığı önemlidir" diye konuştu.

Okulda yasak

Milli Eğitim Bakanlığı'nca hazırlanarak 2007 yılında Resmi Gazete'de yayınlanan "Ortaöğretim Kurumları Ödül ve Disiplin Yönetmeliği"ne göre, öğrencilerin okulda cep telefonu ile iletişim kurmaları da yasak. Buna göre yasağa uymayan öğrenciler "kınama" cezasına çarptırılabilir. Öte yandan "Bilişim araçları ile yönetici, öğretmen, eğitici personel, memur, diğer görevliler ve ziyaretçiler ile öğrencileri rahatsız edici davranışlarda bulunmak" eylemini gerçekleştirilenlere ise "okuldan kısa süreli uzaklaştırılma" cezası verileceği belirtiliyor.

Reklamlar denetlenmeli

Telefon kullanımıyla ilgili özendirici yaklaşımlardan kaçınılması gerektiğini ifade eden Doç. Dr. Şükrü Özen, "Bu bir oyuncak değildir. Çocuklara bunu dayatmak yerine uyarmak bilinçlendirmek gerekir. Aileler, okul yönetimleri ve öğretmenler bu konuda bilinçlendirme yapmalı. Yasal düzenlemelerle özellikle bazı operatörlerin yaptığı gibi cep telefonu ve operatör reklamlarında çocukların oynatılması yasaklanmalı" dedi.

Dil gelişimini yavaşlatıyor

Psikolog Aynur Sayım, cep telefonu için belli bir yaş düzeyini beklemek gerektiğini ifade ederek, cep telefonu kullanımının erken yaşta başlamasının çocukların çevresindeki iletişimi engelleyeceğini dile getiriyor. Sayım, teknolojinin önemli yanlarının da olduğunu ifade ederek, "Burada dikkat edilmesi gereken doğru kullanım. Aile içi çatışmalara neden olabilir. Çocuk ailesiyle olmak yerine mesajlaşmayı tercih edebilir. Cep telefonu erken yaşta kullanıldığı takdirde dil gelişimini yavaşlatır" dedi.

Örnek Eğitim Dokümanı-4'ün Anahtar İfadeleri: çocuk, cep telefonu kullanımı, MEB, tehlike, rahatsızlık, ilk öğretim çağı, çocuk gelişimi, yasak, reklamlar, iletişim

Örnek Eğitim Dokümanı-5: SOSYAL paylaşım sitesi facebook aracılığıyla geçen ağustos ayında tanışarak geldiği Adıyaman'da, genç kızın aşkına karşılık vermemesi üzerine intihara kalkışan ve polislerce ikna edilen 25 yaşındaki Hilmi Bozo, ikinci kez görüşme isteğine olumsuz yanıt alınca tuvalette av tüfeği ile yaşamına son verdi. Hilmi Bozo'nun 13 Ekim 2010 tarihinde bayilik sistemiyle çalışan firmalardan randevu alıp gittiği işyerlerinden hırsızlık ve kapkaç yaptığı gerekçesiyle arandığı ve Facebook hesabından tespit edilerek polis tarafından düzenlenen operasyonla yakalandığı ortaya çıktı.

Erzurumlu olan Hilmi Bozo, internet aracılığıyla tanıştığı H.E. adlı Adıyamanlı genç kıza aşık oldu. Bir süre sanal alemde görüşen Bozo, geçen Ağustos ayında genç kızla görüşmek için Adıyaman'a geldi. Ancak burada H.E'nin aşkına karşılık vermeyerek görüşme talebine olumsuz yanıt vermesi üzerine Bozo, kaldığı otel odasında başına silah dayayarak intihar etmek istedi. Polislerin uzun süren ikna çabalarının ardından Bozo, intihardan vazgeçirildi.

TUVALETTE YAŞAMINA SON VERDİ

İddiaya göre, aşkından vazgeçmeyen Bozo bugün H.E. ile görüşmek için yeniden Adıyaman'a geldi. H.E. ile buluşma yeri olan bir kafede uzun süre bekleyen Bozo, genç kızın gelmemesi üzerine elinde taşıdığı torba ile tuvalete gitti. Bir süre sonra gelen silah sesi üzerine tuvalete gidenler burada Bozo'yu kanlar içerisinde buldu. İhbar üzerine kafeye gelen polisler çevrede önlem alırken, sağlık görevlileri ise yaptıkları incelemede Bozo'nun yaşamını yitirdiğini saptı. Yapılan incelemenin ardından Bozo'nun cesedi Adıyaman Devlet Hastanesi'ne, ardından ise Malatya Adli Tıp Kurumu'na gönderildi. Daha önce de facebook aşkı için intihar girişiminde bulunan Hilmi Bozo'nun ölümüyle ilgili soruşturma başlatıldı.

DAHA ÖNCEDE FACEBOOK SAYESİNDE YAKALANMIŞTI

İnternet ve gazetelerdeki ilanları takip edip, bayilik sistemiyle çalışan firmalarla iletişime geçen, firmalardan randevu alıp gittiği firmalardan hırsızlık ve kapkaç yapan zanlı polis tarafından düzenlenen operasyonla 13 Ekim 2010 tarihinde yakalandığı ortaya çıktı. Online fatura ödeme merkezi bayiliği veren ve kuran bir firmanın sahibi olan Fatih Kırgöl'le bağlantı kuran Hilmi Bozo randevu aldı. Randevu'da fatura ödeme merkezi bayiliği için Fatih Kırgöl ile anlaşmaya varan Bozo, yazılı sözleşme imzaladı.

Fatih Kırgöl'ün güvenini kazanan Bozo, bir yakınından 60 Bin TL alacağı olduğunu belirterek; 'Telefonumun şarjı bitti, sizinkini kullanabilirmiyim' dedi. Bozo Kırgöl'ün telefonunu aldıktan sonra dışarı çıktı ve kayıplara karıştı. Bir süre sonra telefonunun çalındığını anlayan Kırgöl polise haber verdi. Şikayet üzerine Asayiş Şube Müdürlüğüne bağlı Güven Timleri Büro Amirliğine bağlı polisler geniş çaplı araştırma başlattı. Polisler yaptıkları araştırmada Erzurum/Æun Şenkaya İlçesi nüfusuna bağlı Hilmi Bozo'nun (25), Facebook'tan bir kızla tanıştığı, Adıyaman'da bulunduğu, ancak kızın tekrar onu görmek istememesini bahane ederek, bir otel odasında intihar teşebüsünde bulunduğu bilgisine ulaştı. İşyerinde dolandırılan Fatih Kırgöl, Hilmi Bozo'yu internette çıkmış haberlerdeki fotoğrafından teşhis etti.

GÜVEN TİMLERİ FACEBOOK'TA

Bunun üzerine polis ekipleri, bir sosyal paylaşım ağı olan Facebook'ta Hilmi Bozo'nun bir hesabı olduğunu öğrendi. Güven Timlerine bağlı bir kadın polis Bozo'ya arkadaşlık teklif etti. Kadın polisle ile Bozo bir süre internette ve telefonda görüştü. Gaziosmanpaşa'da bulunan bir cafede buluşma teklifi yapan Hilmi Bozo, cafeye gelir gelmez orada bekleyen sivil polisler tarafından gözaltına alındı. Gaziosmanpaşa Polis Merkezi'ne götürülen zanlı polisteki işlemelerinin ardından adliyeye sevk edildi. Zanlı daha sonra tutuksuz yargılanmak üzere serbest bırakılmıştı.

Örnek Eğitim Dokümanı-5'in Anahtar İfadeleri: facebook, tanışma, internet aracılığı, intihar, hırsızlık, polis, yakalama.

Örnek Test Dokümanı-1: Japonya'daki depremde zarar gören Fukuşima Daiçi nükleer santralında alarm seviyesi en yüksek olan 7 seviyesine çıkartıldı. Böylece Fukuşima, 1986 tarihli Çernobil felaketindeki tehlike düzeyine ulaşmış oldu. DEPREM ve tsunamiye bağlı olarak nükleer radyasyon tehlikesi yaşayan Japonya, dünyayı giderek daha çok korkutuyor. Fukuşima Daiçi'deki nükleer santralından yayılan radyasyon oranı, depremin ilk günlerinde öngörülenden çok daha yüksek çıktı. Japonya Nükleer Güvenlik Kurumu da bu nedenle, santraldaki nükleer sızıntısının tehlike derecesini "geçici olarak" en üst seviyeye çıkardı. Deprem ve tsunami dalgalarının tetiklediği kazanın meydana geldiği 11 Mart gününden bu yana 5'nci tehlike seviyesinde tutulan durum, artık 7'nci alarm seviyesinde. Bu düzeyde bir radyasyon uyarısı, en son 1986'daki Çernobil felâketi sırasında yapılmıştı. Santralden şu âna kadar sızan toplam radyasyonun miktarıysa, Çernobil'dekinin yüzde 10'u dolaylarında. Radyasyonu tahmin edilenden daha geniş bölgeye yayıldığı gerçeğini dün kabul eden TEPCO temsilcisi Dzuniti Matsumoto, "Neticede Çernobil'den daha ciddi felaket ile karşılaşabiliriz" dedi.

Rus uzman: Üstleri örtülmeli

Dün yaşanan bu gelişme, Fukuşima'daki sızıntısının başlamasından çok kısa bir süre sonra dünyayı, 'Çernobil'den bile kötü olabilir' diye uyarın Rus uzmanları haklı çıkardı. Rusya'daki Nükleer Teknolojilerin Güvenliği Enstitüsü Müdürü Leonid Bolşov, "Artık gizlenecek bir şey kalmadı. Japonya, bir yandan arızalı reaktörlerle işlenmiş

yakıt depolarını soğutmaya devam ederken, diğer taraftan da santralin yaydığı radyasyon partiküllerini zaptedecek bir yapı düşünmesi gerekir. Dört reaktörün üstü örtülmedikçe Fukuşima ölüm saçmaya devam edecektir” dedi.

En büyük pompa iş başında

ABD’den kargo uçağıyla getirilen dünyanın en büyük inşaat pompası bir kamyonu monte edilerek 4 numaralı santralde taşma aşamasına gelen radyasyonlu suyu boşaltmaya başladı. Pompa daha sonra beton dökme çalışmalarında da kullanılacak.

Bir çay içmediği kaldı

Japon bakanlardan Yukio Edano, halkı santral bölgesinde yetiştirilmiş tarım ürünlerini tüketmeye teşvik etmek için Fukuşima yöresinin domates ve çileklerinden yedi. Çernobil zamanında da bir Türk bakan, çay içerek şov yapmıştı.

Örnek Test Dokümanı-1’in Anahtar İfadeleri:

insan eliyle 1-10	kea_spr 1-10	kea_spr 2-20	kea_spr 3-30
Japonya	nükleer	rus uzman	artık
depremde	çernobil	ört	yüksek
nükleer santral	japonya	çay iç	yaşa
alarm seviyesi	santra	sızıntı	yan
felaket	tehlike	düzey	gel
radyasyon	alarm seviye	bölge	et
sızıntı	fukuşima daiçi nükleer	güven	dün
Çernobil	daiçi nükleer santra	dünya	iç
pompa	pompa	nci	devam
	reaktör	yayıl	kal

Örnek Test Dokümanı-2: Pankobirlik ve Konya Şeker Yönetim Kurulu Başkanı Recep Konuk, dünyanın en büyük kurulu hazır yemek zinciri olan Mc Donalds'ın çevre ülkelerde de Konya Şeker'e ait dondurulmuş patates ürünlerini kullanacağını bildirdi. Konuk, Türkiye'nin en büyük kurulu üretim kapasitesine sahip olan Seydibey Entegre Üretim Tesisleri'nin Türkiye'deki pazar payının yüzde 25'ine sahip olduğunu söyledi.

Tesislerde, patates ve soğandan üretilebilecek bütün ürünlerin işlenerek piyasaya sunulduğunu belirten Konuk, tesislerin kurulmasının üzerinden iki yıl geçmiş olmasına rağmen, iki yılda 3 sezon yol katettiklerini ifade etti.

Hammadde üretiminden başlayarak satış ve pazarlama faaliyetleri de dahil olmak üzere bütün aşamalarını çok sıkı takip ettiklerini dile getiren Konuk, “Bütün Ortadoğu ülkelerine, Balkan ülkelerine Türk cumhuriyetlerine ihracat çalışmalarımız devam ediyor” dedi.

PATATES VE MISIR CİPSİ ÜRETİMİ

Dondurulmuş ürünlerin yanı sıra çeşitlilik oluşturarak patates ve mısır cipsi üretimine de başlayacaklarını vurgulayan Konuk, “İlerleyen günlerde patates-mısır cipsi ve sebze-meyve kurutma tesislerimizde bu çatı altında yerini alacak” diye konuştu.

Dünyanın en büyük kurulu yemek zinciri olan Mc Donalds'ın Türkiye'deki restoranlarında Seydibey Entegre Üretim Tesisleri'nin ürünlerinin bulunduğunu vurgulayan Recep Konuk, şunları kaydetti:

“Mc Donalds ürünlerini yurt dışına da göndereceğiz. Artık çevre ülkelerdeki Mc Donalds restoranlarında da Konya Şeker'in patatesi servis edilecek. Sektördeki teknolojik ve know-how liderliğimizi, sektöre yön vermek için kullanıyor ve büyük bir ivmeyle ilerliyoruz. Bu amaçla 2011 yılı başında en son teknolojilere sahip Yan Ürünler İleri İşlem hatlarımızda devreye girdi.”

Türkiye'nin en büyük dondurulmuş ürün depolama kapasitesine sahip olduklarını dile getiren Konuk, ayrıca 10 bin ton hammadde depolanacak olan doğal depolama tesislerinin inşaatının da devam ettiğini sözlerine ekledi.

Örnek Test Dokümanı-2'nin Anahtar İfadeleri:

insan eliyle 1-10	kea_spr 1-10	kea_spr 2-20	kea_spr 3-30
patates	patates	entegre üretim tesis	sektör
Mc Donalds	konya	türkiye en büyük	bütün
Konya	büyük kurul	kapasite sahip ol	vurgula
şeker	yemek zincir ol	seydibey entegre üretim	devam
tesis	zincir ol mc	hammadde	yıl
ürün	çevre ülke	depola	getir
işleme	ol mc donalds	patates ve mısır	kullan
ihracat	recep konuk	mısır cips üretim	yan
dondurulmuş ürün	dile getir konuk	şeker	et
depolama	don ürün	ilerle	

Örnek Test Dokümanı-3: İstanbul'da Amerikan filmlerini aratmayacak bir hırsızlık gerçekleştiHabertürk Gazetesi'nin haberine göre, 1.3 milyon TL değerinde 39 kilogram külçe altın dolu zırhlı araç, Kapalıçarşı'dan Beşiktaş'taki darphaneye gidiyordu. Galata Köprüsü üzerinde "Lastik patlak" uyarısı yapan bir otomobildeki kişilerin ısrarlı uyarısı üzerine köprü çıkışında güvenlikçiler aracı durdurdu. Lastik değiştirdikten sonra zırhlı araçtaki valizlerden birinin kayıp olduğu anlaşıldı.

Altın piyasasının kalbi Kapalıçarşı'dan yüklenen 39 kilo altından 20 kilosu, sır oldu. Çarşı'nın 25 yıllık kuyumcusu Arif Üstündağ (40), 39 kilogram külçe altını önceki gün Beşiktaş'taki darphaneye sevk etmek için çalışanlarına telefonla talimat verdi.

İKİ VALİZ DOLUSU KÜLÇE ALTIN

Altınlar, komşu bir döviz bürosuna ait zırhlı araca, kuyumcu çalışanları Ahmet Kerküklü (37) ve Erol Bakış (32) tarafından iki valiz halinde yüklendi.Araç akşam

saatlerinde iki personelle yola çıktı. Karaköy'e doğru ilerleyen aracın yanına yaklaşan iki sivil otomobil, zırhlı aracın arka tekerleğinin patlak olduğunu işaret etti. Görevliler önce durmak istemedi ancak araçtakilerin ısrarı üzerine Galata Köprüsü'nün Karaköy çıkışında aracı durdurdular. Kuyumcu çalışanları, inceledikleri lastiğin gerçekten patladığını anlayınca kendilerini uyanan kişilere teşekkür ettiler. Patlayan lastiği değiştiren görevliler, darphanenin yolunu tuttu. Yolda altınları kontrol eden çalışanlar, 20 kilo külçe altın bulunan bir valizin yerinde olmadığını fark etti. Çalışanlar, olayı polise ve işyeri sahibine bildirdi. İşyeri sahibi Üstündağ, iki personeline çok güvendiğini söyleyip şikâyetçi olmazken, soyguncuların yakalanmasını istedi. Çalınan altınların sigortalarının olmadığını ve zırhlı araçla altın sevkiyatının ayın belli dönemlerinde yıllardır aynı personelle gerçekleştiği öğrenildi. Polis yaptığı teknik incelemede zırhlı aracın lastiğinin bıçakla kesildiği tespit edildi.

POLİS, KAMERALARI İNCELİYOR

Karaköy bölgesindeki işyerlerinin güvenlik kamerası kayıtlarına el koyan polis, sinyal kesiciyle aracın kapısının kapanmasına engel olan zanlıların, araç durduğunda altınları alıp kaçtığı ihtimali üzerinde duruyor. Polis ayrıca, kuyumcu çalışanlarının, altınların kaybolmasıyla ilgili ifadesine başvurarak, olaya karışmış olup olmadığını araştırıyor.

Örnek Test Dokümanı-3'ün Anahtar İfadeleri:

insan eliyle 1-10	kea_spr 1-10	kea_spr 2-20	kea_spr 3-30
İstanbul	zırh	personel	güven
hırsızlık	zırh araç	lastik değiş	kamera
külçe altın	külçe altın	aracı durdur	incele
sevkiyat	darphane	zırh aracı	yükle
zırhlı araç	galat köprü	üstündağ	uyarı
darphane	kapalıçarşı	işyeri sahip	olay
polis	kilogram külçe altın	beşiktaş	kes
kamera kayıtları	valiz	dolu	görev
	kuyumcu çalış	patlak	kişi
	karaköy	çıkış	et