

T.C
FIRAT ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ



EĞİTİM ALANINDAKİ NİTEL VERİLERDEN
FİKİRMADENCİLİĞİ ALGORİTMALARI İLE
UZMAN SİSTEM TASARIMI

YÜKSEK LİSANS TEZİ
Veysel GÜNDÜZALP
(121137108)

Anabilim Dalı: Yazılım Mühendisliği Anabilim Dalı
Programı: Yazılım Mühendisliği
Danışman: Yrd. Doç. Dr. Mustafa ULAŞ
MART-2016

T.C.
FIRAT ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

EĞİTİM ALANINDAKİ NİTEL VERİLERDEN FİKİR
MADENCİLİĞİ ALGORİTMALARI İLE UZMAN SİSTEM TASARIMI

YÜKSEK LİSANS TEZİ

Veysel GÜNDÜZALP

(121137108)

Tezin Enstitüye Verildiği Tarih: 03 Mart 2016

Tezin Savunulduğu Tarih: 24 Mart 2016

Tez Danışman: Yrd. Doç. Dr. Mustafa ULAŞ (F.Ü.)
Diğer Jüri Üyeleri: Prof. Dr. Asaf VAROL (F.Ü.)
Yrd. Doç. Dr. A. Fatih KOCAMAZ (İ.Ü.)

MART-2016

ÖNSÖZ

Bu Tez çalışması, Fırat Üniversitesi Fen Bilimleri Enstitüsü Yazılım Mühendisliği Anabilim Dalı Yüksek Lisans Programı'nda hazırlanmıştır. Tezin araştırılmasında, yürütülmesinde ve oluşumunda engin bilgi ve tecrübelerinden yararlandığım ve bilgilendirmeleriyle çalışmamı şekillendiren saygıdeğer hocam ve tez danışmanım Yrd. Doç. Dr. Mustafa ULAŞ'a sonsuz teşekkürlerimi sunarım.

Ayrıca bugünlere gelmemde en büyük pay sahibi aileme, her zaman manevi desteklerini hissettiğim sevgili eşim Cemre GÜNDÜZALP'e tüm kalbimle teşekkür ederim.

Veysel GÜNDÜZALP
ELAZIĞ-2016

İÇİNDEKİLER

Sayfa No

ÖNSÖZ	II
İÇİNDEKİLER.....	III
SUMMARY	VII
TABLolar LİSTESİ	VIII
KISALTMALAR LİSTESİ	IX
ŞEKİLLER LİSTESİ	X
1. GİRİŞ.....	1
1.1. Literatür İncelemesi.....	4
2. VERİ MADENCİLİĞİ.....	6
2.1. Veri Madenciliği.....	6
2.2. Veri Ambarı.....	8
2.2.1. Veri Ambarı Mimarisi	9
2.2.2. Veri Ambarı Sürecindeki Unsurlar.....	10
2.2.3. Veri Ambarı ile Veri tabanı Arasındaki Farklar.....	11
2.3. Veri Madenciliğine Duyulan İhtiyacın Analizi	12
2.4. Veri Madenciliğinde Varolan Problemler	13
2.5. Veri Tabanında Bilgi Keşfi Süreci	14
2.5.1 Veri Tabanı Kavramı.....	14
2.5.1.1 Veri Tabanında Bilgi Keşfi Sürecinin Adımları.....	15
2.5.1.1.1 Problemin Tanımlanması.....	15
2.5.1.1.2 Verilerin Hazırlanması	15
2.5.1.1.3 Belirlenen Modelin Kurulması ve Değerlendirilmesi	16
2.5.1.1.4 Modelin Kullanılması.....	16
2.5.1.1.5 Modelin İzlenmesi	16
2.6. Veri Madenciliği Alanında Kullanılan Modeller	17
2.6.1. Tahmin Edici Modeller.....	17
2.6.1.1 Sınıflandırma ve Regresyon Modelleri	17
3. METİN MADENCİLİĞİ (TEXT MINING).....	18
3.1. Metin Veri Tabanı Oluşturma	19

3.2. Metin Önışleme	20
3.2.1. Joker Yöntemi	21
3.2.2. Veri Filtreleme ve Vektörlerin Ağırlıklandırılması.....	22
3.2.3. Kelime Değerleri	23
3.3. Benzerlik Hesaplanması	25
3.3.1. Öklit Uzaklığı	25
3.3.2. Ağırlıklı Oylama.....	25
3.3.3. Kosinüs Benzerliği	26
3.4. Metin Sınıflandırma.....	26
3.4.1. K En Yakın Komşuluk (k-NN) Algoritması	27
3.4.1.1. K En Yakın Komşuluk (k-NN) Algoritması Bit Ağırlıklandırma Algoritması	29
3.4.1.2. K En Yakın Komşuluk (k-NN) Algoritması Frekans Ağırlıklandırma Algoritması	29
3.4.1.3. K En Yakın Komşuluk (k-NN) Algoritması Tf-IDF Ağırlıklandırma Algoritması	30
3.4.2. Naive Bayes Olasılık Yöntemi	31
3.4.2.1. Multi-Variate Model.....	32
3.4.2.2. Multi-Nominal Model	33
3.5. Metin Madenciliği ve Doğal Dil İşleme Yöntemleri.....	33
3.6. Metin Madenciliğinin Uygulama Alanları	35
4. FİKİR MADENCİLİĞİ	37
4.1. Fikir Kavramı ve Bileşenleri	37
4.2. Fikir Madenciliği ile İlgili Çalışmalar	38
4.3. Fikir Madenciliği Görevleri.....	40
4.4. Fikir Sınıflandırma	43
4.5. Fikir Madenciliğinde Karşılaşılan Problemler	44
4.6. Fikir Madenciliğinin Uygulama Alanları	45
5. GELİŞTİRİLEN UYGULAMA	47
5.1. Uygulamada Gerçekleştirilen Metin Önışleme İşlemleri	48
5.2. İşbirlikli Tartışmacı Öğrenme Nesne Ambarı (İTÖNA)	49
5.2. Uygulamada Gerçekleştirilen Metin Ağırlıklandırma İşlemleri.....	52
5.3. Uygulamada Gerçekleştirilen Sınıflandırma İşlemleri	52
5.4. Uygulamanın Çalışması.....	52
5.5. Deneysel Sonuçlar	55

6. SONUÇLAR VE TARTIŞMA.....	57
KAYNAKLAR.....	59



ÖZET

İnternet bağlantılı akıllı cihazların hayatın her alanında kullanılmasıyla birlikte, üretilen verilerin boyutu da önemli şekilde artmıştır. İnternet ortamında bulunan bu veriler, genellikle sınıflandırılmamış halde bulunmaktadır. Bu veri yığınlarına yapılandırılmamış veri denmektedir. Yapılandırılmamış veri üzerinde herhangi bir teknik kullanmadan bilgi elde etmek mümkün değildir. Bu sebeple, verilerin işlenmesi ve verilerden anlamlı bilgilerin elde edilebilmesi için veri madenciliği teknikleri geliştirilmiştir. Veri madenciliği teknikleri, veriler üzerinde anlamlı bilgi çıkartmayı amaçlamaktadır. Özellikle sosyal medya ve bilgi iletişim sistemlerinin kullanımının artmasıyla birlikte büyük verilerin işlenebilmesi ve sınıflandırılabilmesi büyük bir sorun teşkil etmektedir. Çoklu ortam teknolojileri hızlı bir şekilde gelişmesine rağmen, halen internet üzerinde en fazla bulunan veri formu metindir. Bu sebeple veri madenciliğinin bir kolu olan metin madenciliği geliştirilmiştir. Metin madenciliğinin verimli bir şekilde kullanılabilmesi için doğal dillerin yapısının ve dilbilgisi kurallarının da iyi bilinmesi gerekmektedir. Dil yapıları ve dilbilgisi sistemlerindeki farklılıklardan dolayı her doğal dil için ayrı bir metin madenciliği tekniği geliştirilmelidir.

Bu tezde, veri madenciliği, metin madenciliği ve fikir madenciliği teknikleri incelenmiştir. İlk olarak bu yöntemlerle ilgili literatür taraması yapılmıştır. Tez çalışmasının uygulama aşamasında, bir uzman karar destek sistemi geliştirilmiştir. Bu sistem, kullanıcının hangi fikre sahip olduğunu otomatik tespit edebilmektedir. Geliştirilen uzman sistemi test etmek için İşbirlikli Tartışmacı Öğrenme Nesne Ambarı (İTÖNA) sisteminde bulunan kullanıcı görüşlerinden faydalanılmıştır. Geliştirilen sistemde birden fazla akıllı sınıflandırıcı kullanılmıştır. Bu sınıflandırıcıların performansları analiz edilmiştir. Kullanılan akıllı sınıflandırıcılar Naive Bayes ve k -NN (K nearest neighborhood) sınıflandırıcılarıdır.

Anahtar Kelimeler: Metin Madenciliği, Veri Madenciliği, Fikir Madenciliği

SUMMARY

Creating Expert Decision Systems with Qualitative Datasets Obtained from Education Management Software by Using Opinion Mining Algorithms

The size of produced data has significantly increased with the use of internet-connected smart devices in all areas of life. This data in the Internet environment are generally non-classified form. This dataset is called unstructured data. It is not possible to obtain information without using any technique on the unstructured data. Therefore, data mining techniques have been developed in order to process and obtain meaningful information from these dataset. Data mining techniques are intended to make meaningful information on the data. Especially with the increased use of social media and information communication system, processing and classification of big data has a major problem. Despite the rapid development of multimedia technology, the most abundant form of data on the internet is text data. Therefore, text mining that is a branch of data mining has been developed. In order to use text mining techniques in an efficient manner, the structure and grammar of natural language should be well known. A distinct text mining techniques should be developed for each natural language due to differences in language structure and grammar system.

In this thesis, data mining, text mining, opinion mining techniques has been examined. Firstly, a literature review has been done on these methods. In the implementation phase of the thesis an expert decision support system has been developed. This system can be automatically detected that the user will have the idea. In order to test developed expert system, it have been utilized the opinions of users in the Co-operative and Debater Learning Object Warehouse (CDLOW) system. In the developed system, more intelligent classifier has been used. Performance of these classifiers has been analyzed. Used intelligent classifiers are Naïve Bayes and K Nearest Neighborhood (k -NN).

Keywords: Text Mining, Data Mining, Opinion Mining

TABLÖLAR LİSTESİ

Sayfa No

Tablo 4.1. Fikir Sınıflandırma Görevine İlişkin Bir Taksonomi	42
Tablo 5.1. Geliştirilen uzman karar destek sisteminde kullanılan sınıflandırıcıların başarımı.	56
Tablo 5.2. Geliştirilen uzman karar destek uygulamasının Lima vd.' nin polarite yazılım kütüphanesiyle karşılaştırılması.	56



KISALTMALAR LİSTESİ

k	: Küme sayısı
S	: Benzerlik
n	: Nesne/Doküman sayısı
m	: Boyut sayısı
D	: Doküman koleksiyonu
w	: Terim ağırlığı
TF	: Terim frekansı
IDF	: Ters doküman frekansı
C	: Kümeleme
c	: Sınıf/kategori sayısı
GA	: Genetik Algoritma
NF	: Normalizasyon katsayısı
DDİ	: Doğal Dil İşleme
k-NN	: K En Yakın Komşu
SVM	: Destek Vektör Makinesi
NB	: Naïve Bayes
RO	: Rastgele Orman
PCA	: Temel Bileşen Analizi
YSA	: Yapay Sinir Ağları
ÖNA	: Öğrenme Nesne Ambarı
İTÖNA	: İşbirlikli Tartışmacı Öğrenme Nesne Ambarı
ACC	: Doğruluk

ŞEKİLLER LİSTESİ

Sayfa No

Şekil 2.1. Veri madenciliği süreçleri	7
Şekil 2.2. Veri ambarı mimarisi	9
Şekil 2.3. Veri ambarı süreci	10
Şekil 3.1. Verinin kümelere olan uzaklığının hesaplanması	28
Şekil 3.2. Verinin en yakın kümeye dahil olması.....	28
Şekil 5.1. Uygulamada kullanılan metin madenciliği adımları	47
Şekil 5.2. Geliştirilen uzman karar destek sisteminin blok diyagramı	48
Şekil 5.3. Uzman karar destek sistemi için geliştirilen veri tabanı	49
Şekil 5.4. İTÖNA nesne değerlendirme sayfası	50
Şekil 5. 5. Yeniden düzenlenecek nesneler sayfası	51
Şekil 5.6. Nesne geçmişi sayfası	51
Şekil 5.7. Geliştirilen uzman karar destek sisteminin giriş ekranı	53
Şekil 5.8. Anasayfa ekranı ve veri sayıları.....	53
Şekil 5.9. Nesne değerlendirme sayfası.....	54
Şekil 5.10. Değerlendirme ekranı ve sistemin ürettiği fikir çıktıları.....	54

1. GİRİŞ

Günümüzde en büyük bilgi paylaşım ortamı internettir. İnternetin en büyük avantajı çok hızlı ve çok esnek bir iletişim aracı olmasıdır. Bu iletişim aracı; bireyler, kurumlar ve kuruluşların haberleşmesinde ve bilgi paylaşımında sıkça kullanılmaktadır. Bu kullanım sıklığı nedeniyle internet üzerinde depolanan sayısal verilerin miktarı da hızlı bir şekilde artmaktadır. Yapısal olmayan bir biçimde biriken bu verilere internet kullanıcılarının kolayca erişebilmesi mümkündür. Ancak bu verilerden anlamlı bilgiler çıkarabilmek için veri madenciliği teknikleri kullanılmaktadır. Sayısal ortamda depolanan metin verilerinden bilgi elde edebilmek için metin madenciliği tekniklerinden faydalanılmaktadır.

Veri madenciliği, depolanmış veri veya veri kümelerinden anlamlı bilgileri ve ilişkileri çıkarmak için kullanılan tekniklerin genel adıdır. Veri madenciliği tekniklerinin kullanılabilmesi için veri ambarında veya veri tabanında depolanan verinin yapısal bir biçimde bulunması gerekmektedir. Ancak metin madenciliği uygulamalarında böyle bir gereksinime ihtiyaç yoktur.

Veri işlemenin en önemli sorunlarından biri farklı biçimlerde bulunmasıdır. Nitel veriler üzerinde analiz işlemleri çok kolay bir şekilde yapılabilirken nicel verilerin analizi için farklı ön işleme adımlarına ihtiyaç duyulduğu görülmektedir. Bu tez çalışmasında incelenen temel sorun, verinin sayısal formatta değil; metin formatında olması ve bu formattaki veriden bilgiye erişim sürecinin nasıl uygulanacağıdır. Metin olarak bulunan verinin anlamının ortaya çıkarılabilmesi için metin madenciliği algoritmalarında kullanılan yöntemler topluluğunun genel adı metin madenciliğidir.

Metinlerin yazılmasında standart kurallar olmaması, kısaltmaların çok sık kullanılması, özel karakter ve smiley denilen etiketlerin metinlerle beraber yer alması metin madenciliğinin karşılaştığı başlıca problemlerdendir. Ayrıca bir sözcüğün kullanıldığı cümleye göre farklı anlamlar içermesi ve yapılan edebi sanatlardan dolayı metin verisinden bilgi çıkarmak gittikçe zorlaşmaktadır. Yapısal olmayan metin verisinden bilgiyi elde etmek için kullanılan geleneksel yöntemler ise aşağıdaki gibidir.

- Anahtar kelimeler
- Mantıksal aramalar
- İstatistik ve olasılık teoremine dayalı algoritmalar
- Yapay sinir ağları

- Sözlük tabanlı örüntü eşleştirme modelleri

Yukarıda maddeler halinde yazılan yöntemler, metni oluşturan harflerin karşılaştırılması temeline oturtulmuş algoritmalarlardır. Bu sistemler frekans analizi veya örüntü eşleştirme yöntemlerini kullandıkları için anlamsal karşılaştırma yapamamaktadırlar. Anlamları elde edebilmek için doğal dil işleme (Natural Language Processing - NLP) yöntemlerinden faydalanılmaktadır. NLP tabanlı sistemler kullanılarak, karmaşık yapıda bulunan metinlerden anlamsal ifadeler çıkartılabilmektedir. Örneğin; “Deprem oldu, yüreğimde seni gördüğüm an” ile “Bugün İstanbul’da 3.5 şiddetinde deprem oldu” cümlelerinde geçen deprem sözcükleri aynı anlamda kullanılmamışlardır. NLP tabanlı sistemler bu cümlelerin farklı anlam ihtiva ettiğini tespit edebilmektedir.

Kısaca metin madenciliği; doğal dilleri kullanarak yazılmış metinlerden bilgi elde etme sürecidir ve veri madenciliğinin özelleşmiş bir alt dalıdır. Metin madenciliği iki aşamadan oluşur ve metin madenciliğini oluşturan aşamalar aşağıdaki gibi verilmiştir.

- Anahtar kelime, ifade veya içerikler metinden elde edilir.
- Elde edilen anlamlar kümeleme veya sınıflandırma algoritmaları kullanılarak sınıflandırılır veya kümeleme işlemi yapılır.

Metin madenciliği uygulamaları iki sınıf altında incelenebilir. Bu iki sınıf aşağıdaki gibi verilmiştir.

- **Metinden anlam çıkarılması ve metnin özetlenmesi:** Metinden bilgi elde etmek metin madenciliğinin temel amaçlarından biridir. Böylelikle yapısal olmayan metnin içeriği anlaşılabilir ve metin kategorize edilebilmektedir. Örneğin; sigara zammından şikâyet eden sosyal medya kullanıcıların tespit edilmesi.
- **Metin madenciliği kullanılarak modelleme yapılması:** Paylaşılan metinler vasıtasıyla bireylerin yönelimleri tespit etmek amaçlanmaktadır. Metin verisi kullanılarak elde edilen içerik veya özellik vektörü girdi olarak kullanılarak tahmini model geliştirilir [1].

İnternet kullanımının taşınabilir cihazlar aracılığıyla dahi gerçekleştirildiği günümüzde, internette depolanan veri miktarı ciddi boyutlara ulaşmıştır. Bireyler artık büyük veriyi depolamak için internet ortamını seçmektedir. Bulut hizmetlerinin bir kısmının ücretsiz olması ve bireylerin taşınabilir bellek taşıma zahmetinden kurtarması, internetin; veri depolama amacıyla kullanılmasını da cazip bir hale getirmiştir. Günde 20

milyondan fazla yeni web sitesi internete eklenmekte ve 8 milyardan fazla çevrimiçi dosya internet üzerinden paylaşılmaktadır [2]. Online Computer Library Center araştırmalarına göre internetin gelişiminin devam edeceği gösterilmiştir [3]. Son yıllarda internet kullanımının amiral gemisi haline gelen sosyal paylaşım siteleri veya sosyal medyada birçok veri metin olarak paylaşılmaktadır. Sosyal medyada paylaşılan mesajlar bireyleri, şirketleri, kurumları ve hatta devletleri etkilemektedir. İlk olarak paylaşılan metin zararsız olarak görünebilir ancak bu mesaj, metin madenciliği teknikleri kullanılarak anlamlandırılırsa risk faktörü tespit edilir. Yapısal olmayan metin verisinden, bilginin elde edildiği bu süreç doğru enformasyona ulaşmak ve buna göre hareket planı oluşturabilmek açısından çok önemli bir yere sahiptir. Bilgi iletişim teknolojileri her alana nüfuz ettiği gibi eğitim alanına da nüfuz etmiştir. Günümüzde uzaktan eğitim portalları, eğitici ve bilgilendirici web siteleri, eğitici oyunlar vb. hizmetlere sıkça rastlanmaktadır. Bu platformlarda öğrencilerin paylaştığı yorumlar veya durumlar analiz edilerek öğrencinin öğrenip öğrenmediği veya konuya olan ilgisi tespit edilebilir. Ayrıca öğrencinin derste ne kadar başarılı olduğu, derse olan ilgisi, dersin işlenişiyile ilgili görüşleri de tespit edilebilmektedir. Metin madenciliği ile gerçekleştirilen analiz çalışmalarında, ders materyallerini kullanan öğrenci değerlendirilmelerinin tespiti ve web sayfası kullanımına yönelik bilgilerin elde edilebilmesi amaçlanmaktadır [4].

Eğitimin verimi ve kalitesinin artırılması için kullanılması önerilen birçok destek yazılımı bulunmaktadır. Bu yazılımların sağladığı destekler ise yapılan çalışmalar ile ortaya koyulmuştur [5]. Ancak bu yardımcı yazılımlar sadece öğrenci ve öğretmen arasındaki iletişimi ve bilgi paylaşımını interaktif bir şekilde yapmaya yöneliktir. Bu yazılım teknolojilerinin eğitimin kalitesini artırması için yeni araştırma konularının bütünleştirilmesi gerekmektedir.

Metin madenciliği konusu diğer tüm alanlarda olduğu gibi eğitim alanında da önemli faydalar sağlamaktadır. Metin madenciliği teknikleri ile eğitim destek yazılımlarının ürettiği nitel verilerin yanında nicel verilerin de anlamlı bilgiye dönüşmesini sağlamak önemli avantajlar sunmaktadır.

Bu tez çalışmasının amacı, veri madenciliği, metin madenciliği ve fikir madenciliğiyle ilgili literatür çalışması yapmaktır. Eğitim, ticaret ve politika gibi alanlar hakkında üretilen verilerden anlamlı bilgilerin elde edilebilmesi için veri madenciliği, metin madenciliği ve fikir madenciliği teknikleri kullanılmaktadır. Tez içerisinde eğitim yönetim

yazılımlarından birinden elde edilen nitel verilerin, metin madenciliği algoritmaları ile incelenmesini kapsamaktadır.

Metin madenciliği, metinleri kategorize etmek, özetlemek, konu tespiti, kavram çıkarımı, doküman kümeleme vb. işlemleri yapmak için tekniklerin tümüdür. Fikir madenciliği ise, metin madenciliği sonucunda elde edilen bilgilerden fikir elde etmektir. Bu tez çalışması süresince incelenen makaleler aşağıda verilmiştir.

1.1. Literatür İncelemesi

Metin madenciliğinin gelişimiyle birlikte popüler web sitelerinin metinsel olarak analiz edilmesiyle ilgili birçok çalışma yapılmıştır. Werner ve Myrray [6] The Wall Street Journal gibi temel haber sitelerinde bulunan makaleleri kategorize etmek için metin madenciliği tekniklerini kullanmışlardır. Ikeda vd. [7] twitter verilerini kullanarak kullanıcı profili çıkarmışlardır ve elde ettikleri kullanıcı profillerine göre market analizi yapan bir metot önermişlerdir. Lima vd. [8] twitter mesajlarının polarite analizi için yeni bir yazılım kütüphanesi (framework) sunmuşlardır. Sosyal medyanın kullanımının artmasıyla birlikte twitter ve facebook gibi popüler sosyal paylaşım sitelerinde yapısal olmayan veri yığını artmıştır. Yapısal olmayan bu veri yığını anlamlı bilgiye çevirmek için yeni bir yazılım kütüphanesi önerilmiştir. Önerilen metodun test edilebilmesi için literatürde kullanılan 4 farklı metin veri seti kullanılmıştır. Oluşturulan yazılım kütüphanesinde 4 farklı tipte sınıflandırıcı kullanılmıştır. Bunlar Naive Bayes, Destek Vektör Makineleri, Karar Ağaçları (J48) ve k -NN (K nearest neighborhood) sınıflandırıcılarıdır. Önerilen çerçeve twitter mesajlarının polarite analizini otomatik olarak yapabilmektedir. Uysal ve Gunal [9] metin madenciliğinin ön işlem basamağının sınıflandırma üzerindeki etkisini ortaya koyan bir yöntem önermişlerdir. Önerilen bu yöntemde Türkçe ve İngilizce metin madenciliği yapılmıştır. Ön işlem basamaklarının kullanılan dile, metin uzayına (e-mail, haber vs.) ve özellik boyutuna göre değişkenlik göstermesi üzerinde durulmuştur. Li ve Xu [10] metin tabanlı duygu sınıflandırma kullanarak duygu neden çıkarımı yapan bir yöntem sunmuştur. Fleuren ve Alkema [11] metin madenciliğinin uygulamalarının biyomedikal uzayda kullanılmasıyla ilgili bir çalışma yapmışlardır. Yapılan bu çalışmada metin madenciliğinde kullanılan araçlar incelenmiş ve biyomedikal alanında metin madenciliğinin öneminden bahsedilmiştir. Özellikle ilaçlarla ilgili yorumlar ve ilaç türlerinin anlamlandırılmasında metin madenciliği tekniklerinin kullanılması önerilmiştir. Ayrıca bu makalede

askMEDLINE, XplorMed, PICO vd. web portallar metin madenciliği teknikleri kullanılarak incelenmiştir. Noh vd. [12] anahtar kelime seçimi ve işlem stratejisi metin madenciliği uygulamalarıyla patent analizi yapmak için bir çalışma önermişlerdir. Patent analizi yapabilmek için patent verileri değişik tekniklerle incelenmiştir ancak, son yıllarda internetin ve patent sahtekârlığının yaygınlaşmasıyla birlikte, patent analizi yapabilmek için metin madenciliği tekniklerinin kullanılmasının gerekliliği ortaya çıkmıştır. Bu makalede, patent analizi yapmak için metin madenciliği teknikleri kullanılmıştır. Kullanılan metin madenciliği teknikleri ise anahtar kelimelere odaklanmıştır. Anahtar kelimelerin patent verisinin hangi elemanlarıyla ilişkisinin olduğu meydana çıkarılmıştır. Bu makalede K- En yakın komşu kümeleme algoritması kullanılarak başarılı sonuçlar elde edilmiştir. Akçapınar [13] öğrencilerini intihal davranışlarını azaltmak için metin madenciliği tabanlı otomatik geribildirim yapan bir çalışma önermiştir. Benzerlik analizi yapılabilmesi için bilgisayar eğitimi gören üniversite öğrencilerinin yazdığı metinler kullanılmıştır. İntihal oranını belirlemek için öğrencilerin yazdığı metinler, diğer öğrencilerin yazdığı metinlerle ve internetteki bilgilerle karşılaştırılmıştır. İntihal davranışı gösteren öğrencilere geribildirim yapılmıştır. Bu geribildirimler sonucunda, dönem sonunda intihal davranışının görülme oranı azalmıştır. Romero vd. [14] e-öğrenme uygulamalarının analiz edilmesi için veri madenciliği ve metin madenciliği tekniklerinin kullanılmasını önermişlerdir. Moodle platformunda vaka oluşturarak veriler elde edilmiştir ve bu veriler metin madenciliği teknikleri uygulanarak analiz edilmiştir. Lin vd. [15] metin madenciliği ile çevrimiçi tartışma parçacığı türlerini keşfeden bir yöntem sunmuşlardır. İçerik yönetim sistemlerinin e-öğrenmede sıklıkla kullanılması, içerik yönetim sistemlerindeki yapısal olmayan metin verilerinin anlamlandırılması ihtiyacını doğurmuştur. Bu makalede basamaklı bir yöntem kullanılarak metin madenciliği yapılmıştır. Öğitmen tartışma formu ve öğrenci tartışma formunda var olan metinler modern metin madenciliği teknikleri kullanılarak analiz edilmiştir.

Bu tez çalışmasında yukarıda bahsedilen konularda kullanılan teknikler ve algoritmalar detaylı bir şekilde incelenip, bu konuda kullanılan tekniklerin avantajları ve dezavantajları incelenmiştir. Seçilen algoritmalar ile çalışan bir yazılım gerçekleştirilmiştir. Bu yazılımın başarımlar testleri yapılmıştır.

2. VERİ MADENCİLİĞİ

Bu bölümde veri madenciliği, veri madenciliğine duyulan ihtiyacın analizi, veri madenciliğindeki problemler, veri tabanında bulunan verilerden bilgiye erişme süreci, veri tabanı kavramı, veri madenciliği modelleri gibi bilgiler verilecektir.

Verilerin sayısal ortamda depolanmaya başlanması, akıllı cep telefonların kullanılması, hızlı işlem yapabilme kapasitesine sahip kişisel bilgisayarların artması ve internetin yaygınlaşmasına paralel olarak sayısal veri miktarı, veri tabanlarının sayısı ve depoladığı bilgi miktarı her geçen gün artmaktadır. Yüksek kapasiteli işlem yapabilme gücünün ucuzlaması ve bulutun kullanımının bir sonucu olarak; veri saklama, veriye ulaşma kolaylaşmıştır. Veri tabanlarında saklanan veriler, veri tabanı araçlarında depolanan yığınlar halinde bulunmaktadır ve ham halde herhangi bir bilgi veya anlam belirtmemektedirler. Ham veri amaca uygun şekilde işlenip analiz edilirse, anlamsız olan ham veriden belirli bir amaca yönelik olan anlamlı bilgiler elde edilebilmektedir [16].

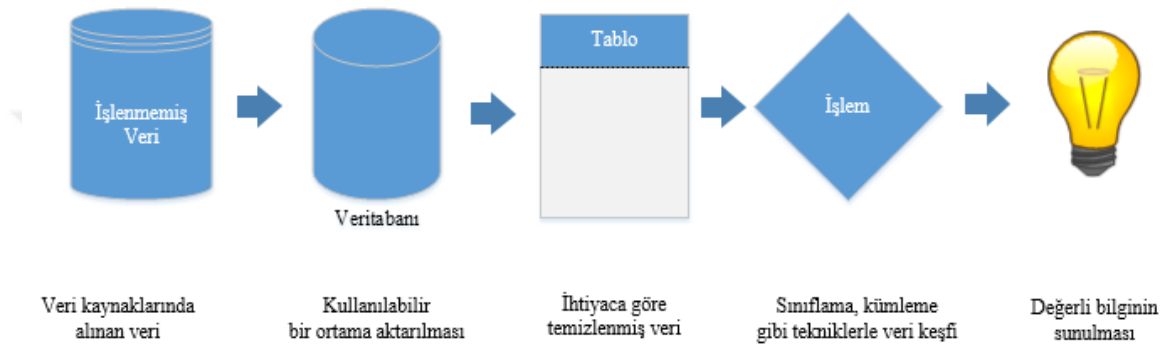
Veri madenciliği disiplininin temelinde matematik ve matematikten doğan istatistik bilimi yer almaktadır. Bu şekilde meydana gelen bu bilim dalının en sık kullanılan alt disiplinleri ise yapay zekâ, makine öğrenmesi, örüntü tanıma, web madenciliği, metin madenciliği ve veri görselleştirmedir. Veri madenciliği teknikleri eğitim bilimlerinde, uzay bilimlerinde, sağlık bilimlerinde, bankacılık ve finans sistemlerinde, sigortacılıkta vb. alanlarda sıkça kullanılmaktadır.

2.1. Veri Madenciliği

Günümüzde bilgi iletişim teknolojileri hemen hemen tüm disiplinlerde yer almaktadır. Bilgisayarların depolama kapasitelerinin ve işlem hızlarının artması, veri aktarımlarının kolay ve hızlı bir şekilde yapılması ve verileri kısa sürede işleyebilmesine olanak sağlamaktadır.

Bilgi iletişim teknolojilerinin kullanımının artmasıyla beraber büyük veri kavramı ortaya çıkmıştır ve bu büyük verinin büyük bir kısmı yapısal olmayan verilerden

oluşmaktadır. Yapısal olmayan veri yığınları, işlenmemiş ve zenginleştirilmemiş uranyuma benzemektedir. Nasıl ki ham halde bulunan uranyum bir işe yaramıyor ancak zenginleştirilmiş uranyum enerji noktasında birçok ihtiyacı giderebiliyorsa, işlenmemiş verinin de bir kıymeti yoktur, ama veri işlendikten sonra elde edilen bilgi atom bombasından daha güçlüdür. Yaşamakta olduğumuz bilgi çağında en büyük güç bilgidir. Bu güce erişmek için kullanılan temel bilim dalının adı “**Veri Madenciliği**dir” [16].



Şekil 2.1. Veri madenciliği süreçleri

Veri madenciliği, ham veriden bilgiyi elde edebilme bilimidir. Farklı bir tanımla karışık ve düzensiz veri modellerini ortaya çıkarılıp bunları karar verme sürecinde ve eylem planını oluşturmak için kullanma aşamasıdır veya veri içerisindeki gizli bilgilerin ortaya çıkarılması ve verinin karar destekli sistemlerde verinin bilgiye dönüştürülmesi aşamasıdır [16].

Literatürde yer alan veri madenciliği tanımlarından öne çıkanlar aşağıdaki gibi verilmiştir.

- Literatürde “Big Data” olarak adlandırılan büyük ölçekli verilerle çalışan ve bu büyük verilerden amaca uygun bilgi elde etmeyi planlayan bir bilim dalıdır.
- Farklı formlarda bulunan her türlü veriyi kullanarak çözüm üretmeyi amaçlayan bir sistemattir.

- Matematiği temele alan ve yapay zekâ, makine öğrenmesi, istatistik, veri tabanlarından bilgi keşfi, örüntü tanıma gibi disiplinlerin yardımcı olduğu bilgi çıkarma yöntemidir.
- Önceden keşfedilmemiş doğrulanabilen ve etkinleştirilebilen bilgileri arayan ve bu bilgileri bulmak için kullanılan bir bilim dalıdır.
- Otomatik çözüm araçları kullanarak bilgiye hızlı erişimi sağlar.
- Birçok bilim dalında ve uygulama alanında kullanılır.
- Problemlere göre değişim gösteren birden fazla çözüm aracını içinde barındırır.
- Hızlı bir şekilde büyüyen ve gelişen bir sektördür.

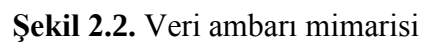
2.2. Veri Ambarı

Teknolojinin hızla gelişmesinin paralelinde bilgisayarlar ve akıllı cihazlar her alanda kullanılmaya başlamıştır. Bu gelişimle beraber veri tabanlarında bulunan veri miktarı artmaktadır. Bu hızlı ilerleme neticesinde **Veri Ambarı (Data Warehouse)** kavramı 90'lı yılların başında ortaya çıkmıştır. Günümüzde veri ambarlarında büyük veriler (Big Data) tutulmaktadır. Facebook ve Twitter'da günlük işlenen veri boyutu sırasıyla 10 ve 7 Terabyte olarak gerçekleşmektedir. Google'ın 1 trilyondan fazla sayfayı indekslediği ve günde Google'dan yapılan arama sayısının 5 milyardan fazla olduğu tahmin edilmektedir. Ancak bu büyük veri kümesinin %80'nin yapısal veri olmadığı karşımıza çıkmaktadır [2].

Veri ambarı, büyük verilerden analiz ve raporlama yapılabilmesi için kullanılmaktadır. Veri ambarının taşınması gereken özellikler aşağıda verilmiştir.

- Veri ambarları belirli bir konuyla ilgili bilgileri içermektedir.
- Veri ambarı birçok farklı kaynaktan gelen verilerin toplanmasıyla oluşmaktadır. Bu veriler yapısal veri tabanı verileri olabileceği gibi, düz metin verileri, video, ses ve fotoğraf gibi çoklu ortam verilerinden de oluşabilir.
- Veriler veri ambarına belirli aralıklarla eklenir. Örneğin saatlik, günlük, aylık veya anlık olabilir.
- Veri ambarları veri tabanları gibi sürekli olarak güncellenmez. Veriler oluşturulma tarihine göre indekslenerek veri ambarına kaydedilir.

Veri ambarı mimarisi aşağıdaki şekilde gösterildiği gibi 3 katmanlı bir yapıya sahiptir. Bu katmanlar ise alt, orta ve üst katman olarak isimlendirilmektedir.



Alt katman ilişkisel veri tabanı sistemlerinin birbirine bağlı olduğu sistemdir. Yapısal veri tabanlarından ve diğer kaynaklardan gelen veriler gateway olarak tanımlanan veri çekme uygulamaları tarafından veri ambarına aktarılır. Bu uygulamalardan en çok bilinenleri ise Sun tarafından üretilen JDBC (Java Database Connection) ve Microsoft tarafından üretilen ODBC (Open Database Connection) uygulamalarıdır.

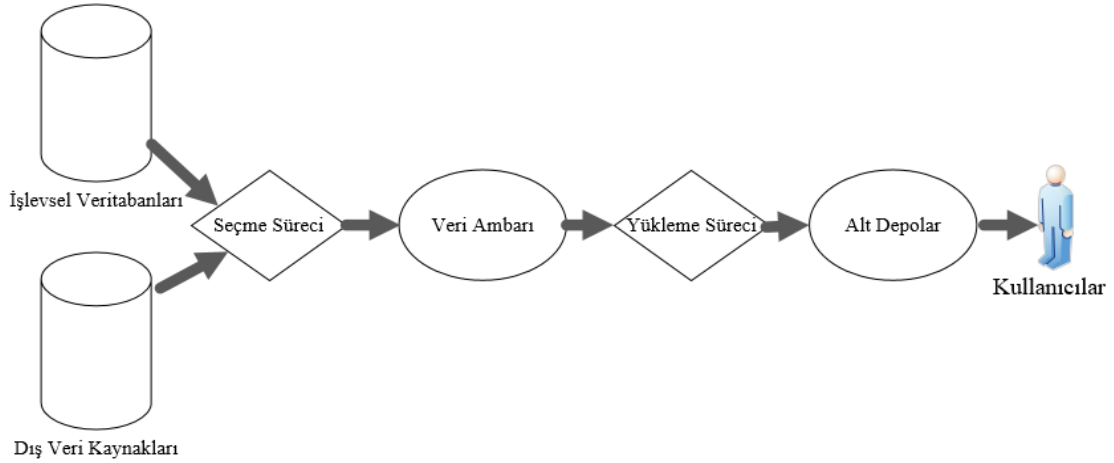
b. Orta Katman

Orta katmanda OLAP (On-Line Analytic Processing) sunucuları kullanılmaktadır. Bu OLAP sunucuları ya ilişkisel (ROLAP) veya çok boyutludur (MOLAP). Alt katmandan gelen veriler, ROLAP veya MOLAP sunucuları kullanılarak veri madenciliği için uygun hale getirilir.

c. Üst Katman

Elde edilen veriler kullanıcının anlayabileceği forma dönüştürülür. Bu katmanda sorgulama, raporlama, analiz ve veri madenciliği araçları kullanılmaktadır. Ham verinin anlamlı bilgi haline geldiği katman üst katmandır.

2.2.2. Veri Ambarı Sürecindeki Unsurlar



Şekil 2.3. Veri ambarı süreci

a. İşlevsel Sistemler

Bir işlem sonucunda elde edilen verilerin detaylı bir biçimde kaydını tutan sistemlere işlevsel sistemler denmektedir. Karar destek sistemleri için ihtiyaç duyulan verilerin çoğu bu sistemlerden elde edilmektedir.

b. Dış Kaynaklar

Veri ambarında yapılacak analizlerin doğruluk oranını arttırabilmek için sık sık dış verilere ihtiyaç duyulmaktadır. Örneğin nüfus, sosyo-ekonomik durum, ev sayısı gibi veriler.

c. Seçme/Çıkarma Süreci

Verilerin veri ambarında yapısal ve tutarlı bir şekilde saklanması gerekmektedir. Gelen verilerin filtreleme sürecidir.

d. Veri Ambarı

Karar destek sistemi için gerekli olan verilerin merkezi durumundadır. Veri ambarında alt depolar oluşturulur.

e. Yükleme Süreci

Veri ambarlarında bulunan verilerin alt depolara ayrılma sürecidir. Bu süreçte dağıtım fonksiyonları kullanılır.

f. Alt Depolar

Veri ambarının dağıtım merkezi olan alt depolar, son kullanıcıların analiz işlemlerini yapabilmesi için verinin uygun hale getirilmesini sağlar.

g. Son Kullanıcılar

Alt depolardan gelen verileri kullanıcının anlayabileceği forma dönüştüren sistemlerdir. Raporlama, analizi, görselleştirme gibi işlemler bu adımda yapılmaktadır.

2.2.3. Veri Ambarı ile Veri tabanı Arasındaki Farklar

Veri ambarı ve veri tabanı arasındaki farklar aşağıdaki gibidir.

- Veri tabanları kullanıcı tabanlı sistemler olup hızlı bir şekilde sorgu, silme, ekleme ve güncelleme yapan sistemlerdir. Veri ambarları ise yönetimsel boyutta kullanılan sistemlerdir. Veri ambarlarını müdürler, yöneticiler tarafından kullanılmaktadır.
- Veri tabanları günlük verileri kullanır ancak, veri ambarları yüksek boyutlara ulaşmış eski verileri geleceği öngörebilmek için kullanır.
- Veri tabanları bir kurumun veya bir işletmenin güncel verileriyle ilgilenmektedir. Veri ambarları ise farklı kurumlardan, farklı işletmelerden veya farklı kaynaklardan gelen verileri depolar.

- Veri tabanları okuma/yazma amaçlıdır ancak, veri ambarları sadece okuma amaçlıdır.

2.3. Veri Madenciliğine Duyulan İhtiyacın Analizi

Günümüzde bankacılık, tıp, kurumsal ve kamusal uygulamalar gibi birçok alandaki verilerin hızlı bir şekilde aktarılması ve işlenebilmesi için sayısal ortama aktarılma zorunluluğu ortaya çıkmıştır. Bu verilerin yapısal olarak depolanması için çeşitli veri tabanları kullanılmaktadır. Bunlar bir de sosyal platformdan ve uydudan elde edilen veriler eklendiğinde, sayısal ortamda bulunan veri hacminin büyük boyutlara ulaştığı bir gerçektir. Veri madenciliği, ham veriden, amaca uygun bilgiyi elde etme işlemlerinin bütünü oluşturmaktadır. Kısacası, ham ve yığınlar halinde bulunan veriden bilgi elde edebilme amacıyla kullanılan, bu veriler arasındaki ilişkiyi ve benzerliği saptayan, verileri amaç doğrultusunda kümeleyip sınıflandıran bilim dalı veya yöntemlerin tümüne veri madenciliği denmektedir. Veri madenciliğini kullanan alanlardan birkaçı aşağıdaki gibi verilmiştir [16].

- Eğitim
- Sosyal Platformlar
- Planlama ve üretim
- İstanbul Borsa
- Satış ve Pazarlama: Hedef müşteri belirleme, müşteri sınıflandırma
- Bankacılıkta örneğin kredi onaylama işlemlerinde
- Sigortacılıkta örneğin poliçe onaylama işlemlerinde
- Sistem yönetimi ve yardım masası
- Taşımacılık-ulaşım
- Konaklama
- Kurumsal uygulamalar

2.4. Veri Madenciliğinde Varolan Problemler

Veri madenciliğinin girdisi ham veri çıktısı ise bilgidir ancak, ham veriden doğru bilgiye ulaşabilmek her zaman mümkün değildir. Veri tabanlarının dinamik, eksiksiz, güncel, yeterli sayıda ve doğrulanabilir veri içermemesi, bilgiye erişim sürecinde problemlerin oluşmasına neden olmaktadır. Bu problemler aşağıdaki gibi kategorize edilmiştir.

- a) Aykırı veriler
- b) Sınırlı veri
- c) Eksik veriler
- d) Veri tabanı boyutu

a) Aykırı Veri: Toplanan veriler iletim hattında çeşitli gürültülere maruz kalabilmektedir. Gürültülü veriler bilgi çıkarım sürecini olumsuz yönde etkilemektedir, çünkü gürültü ile güvenilirlik ters orantılıdır. Gürültülü verilerden kaynaklı yanlış bilgi çıkarımları oluşabilmektedir. Verilerin gürültülerini azaltmak veya gidermek için çok sayıda algoritma önerilmiştir ancak, bu algoritmalar sadece gürültülü bölgede değil, verinin tümünde değişikliğe yol açmaktadır.

b) Sınırlı Veri: Ham veriden bilgiye erişme süreçlerinde bazen eğitim ve test basamaklarına ihtiyaç duyulmaktadır. Bu işlemleri gerçekleştirebilmek için yeterli sayıda veriye ihtiyaç duyulmaktadır. Bazen veri tabanlarında öğrenme ve sonuç çıkarma gibi işlevleri bazı özellikler ve yeterli sayıda veri bulunmayabilir.

c) Eksik Veri: Veri kümesinin büyüklüğünden veya kullanıcının eksik veri girmesinden kaynaklanmaktadır.

Eğer veri kümesinde eksik veriler bulunuyorsa aşağıda belirtilen tedbirlerin alınması gerekmektedir.

- Veri kümesinde bulunan eksik veriler silinmelidir.
- Elde edilen ortalamalar eksik verilerin yerine kullanılabilir.
- Veri kümesinde var olan tam ve tutarlı verilerden yola çıkılarak eksik verilerin yerine uygun değerdeki verilerin yerleştirmelidir.

Eksik veriler, bilgi çıkarım süreçlerinde ve istatistiksel analizlerde önemli problemlere yol açabilmektedir. İstatistiki yöntemler kullanılarak yapılan analizler ve

bu analizleri otomatik olarak gerçekleştiren hazır araçlar, verilerin eksiksiz olduğunu varsayıp geliştirilmiştir. Bu uygulamalar ve analizler, eksik veri bulunan veri kümelerine uygulandığında, elde edilen bilginin geçerliliği ve güvenilirliği düşük çıkmaktadır.

d) Veri Tabanı Boyutu: Veri tabanlarında bulunan verilerin boyutları kullanım sıklığı ve şekline göre sürekli değişmektedir. Veri tabanlarında bulunan verilerden bilgi çıkarımı yapabilmek için veri sayısını önceden belirlenmiş olması gerekmektedir. Verilerin artması, azalması veya belli bir tarafa doğru yığılması bilgi çıkarımı ve veri madenciliği süreçlerini olumsuz etkilemektedir. Veri madenciliği algoritmalarının büyük ölçekte eğitim ve test verisi kullanması doğru tahmin açısından büyük avantaj sağlasa da veri dikkatsizliğinden kaynaklanacak hatalara sebep olmaktadır.

2.5. Veri Tabanında Bilgi Keşfi Süreci

Yüksek kapasiteli ve tek başına herhangi bir anlam ifade etmeyen veri kümelerinden, anlamlı, kendi içerisinde tutarlı, doğru ve işe yarar bilgilere erişme sürecine bilgi keşfi süreci denmektedir. Bilgi keşfi sayesinde analiz ve yorumlama sistemleri birçok alanda kullanılmaya başlanmış ve bu alanlardaki verimlilik artırılmıştır. Örneğin, tarım sektöründe nem, sıcaklık, rüzgâr gibi algılayıcılar kullanılarak veriler toplanabilir ancak, bu veriler tek başına herhangi bir anlam ifade edememekte ve çiftçilikle uğraşan bireyler bu verilerden herhangi bir anlam çıkaramamaktadır. Bilgi keşfi süreci sayesinde algılayıcılardan toplanan veriler anlamlandırılarak çiftçiye tarım verimliliğini arttırmasıyla ilgili önemli ipuçları sağlayabilmektedir. Böylece tarım sektörünün verimi arttırılarak ülke ekonomisine katma değer sağlanabilmektedir.

2.5.1 Veri Tabanı Kavramı

Veri tabanı, depoladığı verilere erişim, seçim, ekleme, silme ve güncelleme gibi imkânlar sağlayan, verileri tablolarda yapısal olarak tutan bir bilgi depolama kümesidir. Veri tabanında bulunan veri kümeleri mantıksal olarak tutarlıdır, çünkü her bir tabloda o tabloyla ilişkili veriler yer almaktadır. Veri tabanlarında veri türleri, indisleme araçları ve anahtarlar kullanılmaktadır. Bu anahtarlar sayesinde veri tabanında bulunan verilerle ilgili

ilişki diyagramları oluşturmak mümkündür. Veri tabanlarında rastgele toplanmış ve belirli bir sıralama yapılmamış verilere rastlamak güçtür ve veriyi yapısal halde depolamayı hedeflemiştir [16].

Veri Tabanlarında Bilgi Bulma Süreci (VTBK) zorlu ve karmaşık bir işlemdir. Veri tabanında bulunan veri kümesine yüksek dereceli ve anlamlı açıklamalar getirebilmek için bu süreç işletilmelidir. Bu süreç basamaklı bir yapıdan oluşmaktadır. Bilgi keşfinin tüm verileri genelleyebilmesi ve yeni gelen verilere göre şekillenmesi gerekmektedir. Ayrıca bilgi keşfi süreci sonunda ulaşılan bilginin amaca hizmet edebilmesi gerekmektedir [16].

2.5.1.1 Veri Tabanında Bilgi Keşfi Sürecinin Adımları

VTBK adımları aşağıdaki gibi verilmiştir.

- a. Problemin tanımlanması
- b. Verilerin hazırlanması
- c. Modelin kurulması ve değerlendirilmesi
- d. Modelin kullanılması
- e. Modelin izlenmesi

2.5.1.1.1 Problemin Tanımlanması

Problemin tanımlanması sürecinde başarıya ulaşabilmek için öncelikle ihtiyaç analizi yapılmalıdır. İhtiyaç analizinin ardından o alandaki problemi tanımlamak gerekmektedir. Problemin sınırlarının net bir biçimde çizilmesi ve problemin anlaşılabilir bir dille ifade edilmesi sağlanmalıdır. Ayrıca erişilmek istenen başarımlar ölçütü net olarak ortaya konulmalıdır.

2.5.1.1.2 Verilerin Hazırlanması

Modelin kurulması ve değerlendirme aşamasında meydana gelebilecek problemleri çözebilmek için bu basamağa tekrar tekrar başvurmak gerekebilir. Bu aşamada veriler veri madenciliğine uygun şekle getirilir ve örüntü tanıma metotlarının ön işlemleriyle aynı

işlemleri yapmaktadır. Modele uygun verilerin oluşturulması için veri toplama, verilere değer atama, ayırık verileri birleştirme, gürültülü verilerdeki gürültüleri temizleme, örneklem uzayının oluşturulması ve dönüşüm aşamalarından oluşmaktadır.

2.5.1.1.3 Belirlenen Modelin Kurulması ve Değerlendirilmesi

Belirlenen problemin çözülebilmesinde problemi çözmemizde bize en uygun veri madenciliği modelini kullanmak gerekmektedir. Probleme en uygun modeli bulabilmek için ya önceden yapılmış çalışmalar incelenip, problemin çözümüne en uygun modeli belirlemek gerekir ya da tüm model kaba kuvvet ile tek tek probleme uygulanıp en iyi sonucun elde edildiği modelin seçilmesi gerekmektedir. Bu basamakta ortaya çıkabilecek en büyük problem ise teoride iyi sonuçlar verdiği sanılan model kullanılarak pratikte daha kötü sonuçlar elde edilebilmesidir.

Veri madenciliğinde kullanılan modeller aşağıdaki gibidir.

- Sınıflandırma ve regresyon modelleri,
- Kümeleme modelleri,
- Birliktelik Kuralları ve Ardışık Zamanlı Örüntülerini kullanan modellerdir.

2.5.1.1.4 Modelin Kullanılması

Modeller doğrudan bir uygulama olarak kullanılabileceği gibi herhangi bir uygulamanın modülü olarak da kullanılabilir. Bir amacı gerçekleştirmek için kurulan ve kullanılan modeller dolandırıcılık tespiti, risk analizi, terk etme veya kalma eğilimi gibi uygulamalarda doğrudan kullanılabileceği gibi, herhangi bir öğrenci otomasyonunda öğrencinin okulla ilgili tutumu da tespit edilebilir

2.5.1.1.5 Modelin İzlenmesi

Modelin izlenmesi, modelin test sonuçları ile hata yüzdesinin uygunluğunun yorumlanması ve kullanılan modelin farklılaşan verilere ne gibi tepkiler verdiğinin gözlenmesi süreçlerini içinde barındırmaktadır.

2.6. Veri Madenciliği Alanında Kullanılan Modeller

Veri madenciliği alanında kullanılan modeller temel olarak ikiye ayrılmaktadır. Bunlar tahmin edici ve tanımlayıcı modellerdir.

2.6.1. Tahmin Edici Modeller

Sonuçları önceden belli olan veri kümesinden yola çıkılarak farklı model geliştirilmesi ve geliştirilen bu modelden faydalanılarak sonuçları önceden belli olmayan veri kümeleri için sonuç verilerinin tahmin edilmesi amaçlanmaktadır. Öngörüle bilen modeller regresyon ve sınıflama yöntemleridir.

2.6.1.1 Sınıflandırma ve Regresyon Modelleri

Var olan veri kümesini kullanarak gelecek tahmini yapılan veri madenciliği modelidir. Eğitim ve test kümeleri oluşturularak sınıflandırma yapılmaktadır. Bu model veri madenciliğinde kullanılan en popüler modeldir. Sınıflandırma ve regresyon arasındaki en bariz fark, tahmin edilen değişkenin süreklilik göstermesidir. Verileri kategorize etmede yüksek başarıma sahip olan bu modelde karşılaşılan en büyük problem eğitim ve test verilerinin hazırlanmasıdır.

Regresyon ve sınıflandırma modellerinde kullanılan teknikler aşağıdaki gibidir;

- Yapay Sinir Ağları,
- Destek Vektör Makineleri,
- Uç Öğrenme Makineleri,
- Karar Ağaçları,
- Naive Bayes,
- Bulanık Mantık,
- Bellek Temelli Nedenleme

3. METİN MADENCİLİĞİ (TEXT MINING)

Günümüz teknolojisi her alanda çok sayıda ve çok çeşitli verinin oluşmasına neden olmakta ve bunlara kolay ulaşılmasını sağlamaktadır. Gittikçe büyüyen veri hacmi içinde önemli bilgiler kaybolup giderken, değerli bilgilere ulaşmak için özellikle yarı yapısal veriler olan belgelerin içeriğinin belirlenmesi ve buna uygun sorgulanabilmesi ihtiyacı kendini hissettirmektedir. Günlük hayatımıza baktığımızda gazeteler, makaleler, raporlar birer belgedir ve veri niteliği taşırlar. Geleneksel yöntemler bir yere kadar belge yığınlarından faydalı ve gerekli bilgileri bulmaya yardımcı olsalar da asıl olarak gereken detay ve özel bilgilere bu yöntemler ile ulaşmak zordur. Bu nedenle bu tür verilerin incelenmesinde Metin Madenciliği teknikleri öne çıkmaktadır.

Veri tabanları programların otomatik olarak işlemesi için tasarlanmaktadır. Metin ise insanların okuması için yazılır. Metinleri en azından insan kadar okuyabilen ve anlayabilen programlar yoktur [17]. Doğal dil işleme alanında gelinen olağanüstü noktalara rağmen beklenen gelecek için öyle bir program mevcut değildir. Birçok araştırmacı insanların yaptığı şekilde okuyan ve anlayan programları yazmamızdan önce aklın nasıl çalıştığını gösteren tam bir benzetime ihtiyaç olacağını düşünmektedir. Bu konuda farklı disiplinlerden çalışmalara ihtiyaç vardır. Bunlardan birisi de Metin Madenciliği çalışmalarıdır.

Metin Madenciliği, işletme dokümanları, müşteri yorumları, web sayfaları ve XML dosyalarını içeren yapısal olmayan veri kaynaklarının herhangi bir çeşidinden gelen yeni, önceden bilinmeyen, potansiyel olarak kullanışlı bilgiyi keşfetme sürecidir [18].

Başka bir deyişle Metin Madenciliği, derlemlerden yeni bilgileri türetme işlemi olarak da tanımlanabilir. Derlem (corpus-küllliyat), elektronik ortam üzerinde toplanmış yazılı ya da sözlü doğal dil seçkilerinin belli prensipler dâhilinde bir araya getirilerek dilbilimsel ve yapısal bilgi ile işaretlenmiş halinden oluşan veri tabanına denilmektedir. Metin Madenciliği çalışmalarından elde edilen bilgiyle, analiz edilmiş olan metin kaynaklarında açık olarak görülmeyen ilişkiler hipotezler veya eğilimler olduğu anlaşılır [17].

Diğer bir tanımda ise Metin Madenciliği, işletme arşivinde veya internet üzerindeki belgelerde bu belgeye benzer belgelerin olup olmadığı elle bir sınıflandırma gerekmeden benzerliği bulmaktadır. Bu benzerlik genellikle otomatik olarak çıkarılan anahtar kelimelerin tekrarı hesaplanarak yapılmaktadır [19]. Metin Madenciliği, veri

madenciliğinin genel alanının bir parçası olarak düşünülmesine rağmen alışlagelen veri madenciliğinden farklıdır. Ana farklılık, metin madenciliğinde örüntülerin olay tabanlı veri tabanlarından daha çok, doğal dil metinlerinden çıkartılmasıdır.

3.1. Metin Veri Tabanı Oluşturma

Veri tabanı bir veya daha fazla uygulamaya hizmet vermek üzere bir araya getirilmiş birbirleriyle ilişkili veriler bütünüdür. Veri tabanı sadece verinin saklanması değil, aynı zamanda o veri üzerinde değişiklik yapılmasına da imkân vermekte, bilgisayarda veri işlemek amacıyla kullanılmaktadır. Veri tabanı, değişik tiplerdeki varlıklara, bu varlıkların özniteliklerine ve bunlar arasındaki ilişkilere ev sahipliği yapan bir yapıdır. Veri tabanının özellikleri şu şekilde sıralanabilir [20]:

- Veri tabanında kayıtlar yinelenmez. Her bilgiden sadece bir tane vardır.
- Veri tabanında bulunan bilgiler standart hale getirilir.
- Veriler bir bütün halindedir.
- Veri tabanındaki bilgiler farklı amaçlar doğrultusunda farklı bakış açıları ile farklı kullanıcılar tarafından aynı anda kullanılabilir.
- Veri tabanındaki tüm bilgiler aynı yerde saklandığı için üzerinde güvenlik sınırları oluşturulur ve giriş ve çıkışlar kontrol altında tutulmaktadır. İstenmeyen veya güvenlik açısından tehdit oluşturabilecek kullanıcıların veri tabanına erişimi kısıtlanır veya engellenir.
- Beklenmeyen herhangi bir durumda veri kaybının en aza indirgenmesi için veri tabanında belirli aralıklarla yedekler alınmaktadır.

Veri tabanın en önemli özelliklerinden biri; çok büyük verilerin güvenli ve hızlı bir şekilde ihtiyaç duyulan veriye dönüştürmesidir. Veri tabanı türleri şu şekilde sıralanabilir. Nesne Yönelimli Veri Tabanları (Object-Oriented Databases), Dağıtık Veri tabanları (Distributed Databases), Hiper Medya Veri Tabanları (Hypermedia Databases), Çok Boyutlu Veri Analizi, Veri Bankaları.

Veri ambarları ise, bilgiyi elde edebilmek için gerçekleşen veri madenciliği süreçlerinin lojistik kaynağı olarak düşünülebilir. İşlenecek veriler veri ambarlarından tedarik edilir. Veri ambarlarında farklı yapıda veriler depolanır ve kullanılacak veriler tek çatı altında toplanmaktadır [21]. Ayrıca veri ambarları, işletmelerin tarihsel verilerini içerdiği için veri tabanlarındaki gereksiz detaylardan arındırılmıştır ve orada olduğu gibi

anlık olarak güncellenmezler, belli zaman aralıkları ile güncelleme söz konusudur. Veri ambarı sakladığı veriyi aynı ortamda analiz etme imkânı tanır.

Online Analytical Processing (OLAP), veri ambarı analiz etmek için geliştirilen ve en sık kullanılan yöntemdir. OLAP araçları, “BigData”yı analiz etmek için geliştirilmiştir ve SQL’ den daha üstün analiz yeteneğine sahiptir. İnteraktif veri analizini yapabilmek için de OLAP araçlarından faydalanılır. Bu araçların temel hedefi, süreci mümkün olduğu kadar otomatikleştirmektir [22].

3.2. Metin Önişleme

Metin sınıflandırma algoritmaları için gerçekleştirilmesi gereken ilk adım metin önişleme adımıdır. Veri ön işleme aşaması aşağıda belirtilen nedenlerden dolayı gerçekleştirilir.

- 1-Veri üzerinde bulunabilen problemleri çözmek,
- 2-Verinin doğal yapısını öğrenerek daha anlamlı ve kaliteli analiz yapabilmek,
- 3-Veriden daha anlamlı bilgi üretebilmek [23].

Türkçe dilinin yapısı gereği ön işleme zor olmaktadır. Çünkü Türkçe sondan eklemeli bir dildir ve bir kelimeye eklenen ekler o kelimenin anlamını daha da genişletmekte hatta değiştirmektedir. Ayrıca tek bir Türkçe kelimeden birden fazla değişik anlamda kelimeler türetilir. Sondan eklemeli yapı ve anlam genişlemesi sebebiyle, Türkçe için daha çok özel metin madenciliği tekniklerinin geliştirilmesi gerekliliği ortaya çıkmaktadır. Bu sebepten dolayı, kelimedeki var olan tüm özel karakterlerin kaldırılması ve harflerin küçük harfe çevrilmesi dışında; anahtar kelimelerin ve joker oluşturulması gibi ön işleme işlemlerinin gerçekleştirilmesi gerekmektedir.

Metin dosyalarının kategorilerini tespit edebilmek için dokümanlar kelime vektörleri olarak gösterilecektir. Vektörün boyutu daha önceden oluşturulan sözlüğün boyutuna eşit olacak, ağırlıkları ise çeşitli algoritmalarla ifade edilecektir (Binary, frekans, tf-idf gibi).

Doğal diller kullanılarak yazılan metinler matematiksel olarak bir vektör veya bir sayı dizisiyle ifade edilmezler. Böyle bir matematiksel değerin olmaması, metin madenciliği sürecinin zorlaşmasına sebep olmaktadır. Örneğin internet ortamında birçok yazılı doküman bulunmaktadır, bu metinlerde devrik yapıları cümleler, kısaltmalar, çeşitli noktalama işaretleri, özel karakterler ve hatalara da sıklıkla rastlanmaktadır. Bu sebepten dolayı metin önişleme adımı metin madenciliğinin vazgeçilmez unsurudur, çünkü metin

sınıflandırıcının anlamlandırabileceği forma dönüşmezse sınıflandırma işlemi başarısızlıkla sonuçlanır. Metin ön işleme adımları aşağıdaki gibidir.

1- Kategorizasyon aşaması: Kategoriler belirlenir ve bu kategorilerle ilişkilendirilebilecek metinler kullanılarak sözlük oluşturulur.

2- Sözlükte N adet kategori belirten kelime sayısının olduğu varsayılırsa, sözlükte tüm kelimeler tek tek incelenir ve joker olarak kullanılabilecek kelimeler tespit edilerek sözlük güncellenir. Burada sözlüğü güncellemek için doğrusal arama algoritmasının kullanılması gerekmektedir.

Örnek:

Sözlükler

Sözlükte

Sözlüğü

Sözlü* (“Sözlü” kelimesinin sonuna ne gelirse gelsin kabul et demektir, “Sözlük” sözcüğü olarak değerlendir).

3- Sözlükte oluşan tüm kelimeler vektör ağırlıklandırma yöntemi kullanılarak ifade edilir.

3.2.1. Joker Yöntemi

Metinlerde geçen sözcüklerin yerine gövdelerinin de kullanılabileceği bir önceki bölümde yer alan örnekte gösterilmişti. Bunun en temel sebebi ise, eklemeli dillerden olan Türkçe’de bulunan sözcüklerin ek alarak başka anlamları ifade etmeleridir. Örneğin “bilgisayar” kelimesi ile “bilgisayarı”, “bilgisayarda”, “bilgisayardan”, ve “bilgisayarın” kelimelerinde kesme işareti kullanılmazsa bu kelimeler farklı kelimeler olarak değerlendirilebilmektedir. Eğer ekli sözcüklerin her biri farklı kelime olarak sözlüğe eklenirse, sözlüğün boyutu çok büyür ve işlem hızı yavaşlayacaktır [24]. Joker kelime ise aynı dizilime sahip farklı ekler almış kelimelerin tek bir kelime ifade edilmesini sağlayan Joker kelime denmektedir. Joker kelimelerin tespit edilmesi için Non-deterministic Finite Automata (NFA) kullanılabilmektedir. Joker kelime, temel kelimedir ve ona gelecek eklerle anlamda büyük kaymaları yaşanmamaktadır. Kelimeden çekim ve yapım eklerini ayırıştırarak kelimenin kökü elde edilmektedir. Ancak joker kelimelerde köke indirgeme şartı yoktur. Kelime yalın halde değil, ekli halde de bulunabilir [23]. Metin kategorisinin belirlenmesinde joker kelimeler hayati önem taşımaktadır. Joker yöntemiyle anlamsal

olarak ilişkili kelimeler tespit edilebilmektedir. Örneğin “masa* ”jokerli kelimesiyle (* sembolü joker olduğunu ifade etmektedir), “masa” sözcüğünün sonuna nasıl bir ek gelirse gelsin masa sözcüğü belirtilmiş olacaktır.

Örnek: “ayakkabıcı” ve “ayakkabıcılar” kelimelerini ayrı anlamlarda düşünürsek bu kelimelerin sözlüğe eklenmeleri gerekmektedir. Bu da sözlüğün boyutunun büyümesine, durum ağacındaki durumların artmasına ve performansın düşmesine sebep olmaktadır. Bu sebepten dolayı aynı anlamı ifade edebilecek bir gövde seçilmesi gerekmektedir. Bu işlem sadeleştirme işlemi gibi düşünebilir. Bu iki sözcük yerine “ayakkabıcı*” veya “ayakkabı*” joker sözcüğünün kullanılması daha mantıklı olacaktır. Böylece ayakkabı ile başlayan tüm sözcükler aynı anlamı ifade edecektir ve kelimenin tüm harflerini analiz etmeye gerek kalmayacaktır. Türkçede sert sessizler ve sert sessizlerin yumuşaması olaylarına rastlanmaktadır. Ünsüz yumuşaması gerçekleşme olasılığına karşın iki adet joker kelime kullanılması gerekmektedir. Örneğin “kalıp*”, ”kalıb*” joker kelimeleri sözlüğe eklenecektir. Ayrıca joker kelime seçerken kelimenin köküne inmek her zaman işe yarayan bir yöntem değildir. Örneğin “balık” kelimesinin “bal*” olarak seçilirse yanlış anlamların türetilir. Bu sebepten dolayı “balık” anlamının türetilmesi için “bal*” joker sözcüğünün kullanılması yersiz olacaktır. Bir başka örnekte ise “ev*” joker sözcüğünden evren, evrim gibi sözcükler türetilir ve burada da anlam kayması yaşanmaktadır. Türkçede karşılaşılan sıkıntılardan birisi de “yor” ekinin kullanılmasıdır. Örneğin “yor” eki yok edilerek kök bulunması istenirse bazı problemler meydana gelmektedir. Örneğin “istiyor” kelimesinden “yor” eki kaldırılıp joker kelime türetilirse “isti*” olarak joker kelime elde edilir ancak, bu kelimenin yalın hali “iste”dir. Joker kelime “iste*” olarak belirlenirse, kullanılan NFA “istiyor” kelimesini tanımlayamamaktadır. Bundan dolayı kullanılan NFA’nın durum tablosunun yeniden tanımlanması gerekmektedir.

3.2.2. Veri Filtreleme ve Vektörlerin Ağırlıklandırılması

Veri filtreleme ve vektör ağırlıklandırma işlemlerinin yapılabilmesi için aşağıdaki adımların takip edilmesi gerekmektedir.

Adım 1: Metinde bulunan tüm harfler küçük harfe çevrilir ve metindeki noktalama işaretleri temizlenir.

Örnek: Türkiye’de bir ilk yaşandı.

Sonuç: türkiye bir ilk yaşandı

Örnek: Malezya'da deprem: 320 ölü...

Sonuç: malezya deprem ölü

Adım 2: Noktalama işaretleri yok edilen ve tüm harfleri küçük harfe çevrilen metin bir diziye atılır.

Dizi: (türkiye, bir, ilk, yaşandı)

Adım 3: Dizideki elemanlar sözlükteki elemanlar ile karşılaştırılır. Böylece vektörel ağırlıklar bulunur. Sözlükteki joker karakterlerin aşağıdaki gibi olduğu varsayılırsa;

başkan*

türkiye*

bir*

futbol*

Oluşturulacak vektör ise (0,1,1,0)'dır. Bu yöntemle eğitimi ve test dokümanları ağırlıklandırılabilir. Bu yöntemin dışında bit, frekans ve tf-idf gibi yöntemler de mevcuttur.

3.2.3. Kelime Değerleri

Doküman, kelimeler ve onların ağırlıkları ile gösterilir. Ağırlıklandırma ne kadar iyi yapılırsa, kategorizasyon da o kadar başarılı olur. Ağırlıklandırma önemli bir konu olduğundan, çeşitli teknikler geliştirilmiştir.

- Terim Frekansı (TF),
- Ters Doküman Frekansı (IDF),
- Terim Frekansı-Ters Doküman Frekansı (TF-IDF),
- Terim Ayırıştırma Değeri,
- Olasılıksal Terim Ağırlıklandırma,
- Tek Terim Doğruluğu,
- Genetik Algoritmalar.

Aşağıda algoritmanın kolay anlaşılabilmesi bakımından küçük boyutlu bir sözlük kullanılarak ön işleme aşamasına örnek verilmiştir.

Kelimeler (Sözlük) Jokerleri

ekonomi*

çiftçi*

futbol*

ilaç* ,

öksü*,

transfer*

Örnek eğitim dokümanları aşağıdaki gibi olursa;

1. Transfer süresinin bitmesine az kala Fenerbahçe transfer ettiği yeni futbolcuları tanıttı. (Spor)
2. Futbol dünyası yeni transfer edilen futbolcuları konuşuyor. (Spor)
3. Adam bir anda öksürmeye başlamıştı. (Sağlık)
4. Doktor hastaya üç farklı ilaç verdi. (Sağlık)
5. Çiftçiler kayısı fiyatlarındaki düşüşü protesto etti. (Ekonomi)
6. Ekonomideki düşüş birçok esnafı olumsuz etkileri. (Ekonomi)

Doğrusal arama algoritması kullanılarak sözlükteki kelimeler metin dokümanında aranacak ve o kelimeye ait jokerin belirlenmesi işlemi yapılacaktır. Eğer işlem yapılan kelimenin jokeri sözlükte mevcutsa joker bulunur. Yani “söz” sözcüğü ile beraber “söz*” jokerinin de aranması gerekmektedir [25]. Tespit edilen sözcük sayısı boyuta eklenerek boyut güncellenmelidir. Örneğin kategorisi bulunmak istenen metin aşağıdaki gibiyse:

SORGU- Transfer edilen futbolcular takımla çalışmalara başladı.

Sözlük={ ekonomi*, çiftçi* , futbol*, ilaç*, öksür*, transfer* }

Vektörlerin kelime frekansına göre ifade edilmesi;

D1=(0, 0, 1, 0, 0, 2)

D2=(0, 0, 2, 0, 0, 1)

D3=(0, 0, 0, 0, 1, 0)

D4=(0, 0, 0, 1, 0, 0)

D5=(0, 1, 0, 0, 0, 0)

D6=(1, 0, 0, 0, 0, 0)

DQ=(0, 0, 1, 0, 0, 1)

Vektörlerin bitsel olarak ifade edilmesi;

D1=(0, 0, 1, 0, 0, 1)

D2=(0, 0, 1, 0, 0, 1)

D3=(0, 0, 0, 0, 1, 0)

D4=(0, 0, 0, 1, 0, 0)

D5=(0, 1, 0, 0, 0, 0)

$$D6=(1, 0, 0, 0, 0, 0)$$

$$DQ=(0, 0, 1, 0, 0, 1)$$

3.3. Benzerlik Hesaplanması

Sınıflandırma ve doküman benzerliği bulma işlemlerinde dokümanların birbirleriyle ne kadar ilişkili olduğunun tespit edilmesi dokümanların yakınlıklarıyla ilgilidir. Dokümanların benzerliklerini bulmak, vektörel ifadeleri kullanılarak gerçekleştirilir. Benzerlik, dokümanlar arası mesafenin ölçülmesi ve değerlendirilmesidir [26], Öklid mesafe ölçümü ve kosinüs benzerliği en çok kullanılan benzerlik tespit yöntemleridir.

3.3.1. Öklit Uzaklığı

$X=\{x_1, x_2, x_3, \dots, x_n\}$ ve $Y=\{y_1, y_2, y_3, \dots, y_n\}$ iki vektör olmak üzere bu iki vektör arasındaki mesafe olan d 'nin hesaplanmasıdır. d değerinin düşüklüğü vektörlerin birbirlerine yakın olduğunu belirtir.

Öklid için bir sözcüğün iki vektörde bulunması ile iki vektörde bulunmaması aynı değerdedir. Öklid, bir sözcüğün sadece iki vektörde bulunmasıyla değil ikisinde bulunmamasıyla da ilgilenir. Dokümandaki bütün sözcükler dikkate alınarak yapılan sınıflandırma çalışmalarında düşük performanslıdır.

3.3.2. Ağırlıklı Oylama

k -NN, k komşu içerisinde en fazla tekrar eden sınıfla ilgilenir. Bu durumda komşular içerisinde çok yakın veya çok uzak eğitim dokümanı arasında değerlendirme açısından bir farklılık yoktur, ikisi de karar vermek için eşit değerdedir. Ağırlıklı oylama, uzaklık değerlerinin karelerinin çarpmaya göre tersi alınıp sınıf bazında toplanarak elde edilen en yüksek değere sahip sınıfın doküman sınıfı olarak belirlenmesi ilkesine dayanır. Bu şekilde k komşu içerisindeki en fazla tekrarlayan sınıf değil, en yüksek ağırlığa sahip sınıf doküman sınıfı olarak atanır.

3.3.3. Kosinüs Benzerliği

$X=\{x_1, x_2, x_3, \dots, x_n\}$ ve $Y=\{y_1, y_2, y_3, \dots, y_n\}$ iki vektör olmak üzere bu iki vektörün benzerliği s şeklinde Eşitlik 3.1’te formüle edilmiştir. s değerinin büyüklüğü vektörlerin birbirlerine yakın olduğunu belirtir. s değeri, iki doğru arasındaki açının kosinüsüdür. s değeri ile iki doğru arasındaki açı değeri ters orantılıdır. s benzerlik değeri skaler çarpımların iki vektörün normlarının çarpımlarına oranıyla bulunur. s , maksimum 1 değerini alır. s değerinin 1’e yakınlığı vektörlerin birbirine benzediğini gösterirken 0’a yakınlığı vektörler arasında ortak sözcüğün az olduğunu ve vektörlerin benzerliğinin düşük değerde olduğunu gösterir.

$$s = \frac{\sum_{i=1}^n x_i y_i}{\sqrt{\sum_{i=1}^n (x_i)^2} \sqrt{\sum_{i=1}^n (y_i)^2}} \quad (3.1)$$

Eşitlik 3.2’da, vektörlerin skaler çarpımlarını ifade etmektedir. Eğer bir sözcük bir vektörde var diğer vektörde yoksa skaler çarpım sonucu o sözcük için 0’dır.

$$S_C = \sum_{i=1}^n x_i y_i \quad (3.2)$$

Kosinüs, sözcüğün vektörde varlığını araştırır. İki vektörde de bulunan sözcük değerlendirmeye alınır. Bu durum, Öklid’in aksine, dokümandaki bütün sözcükler alınsa bile sözcük varlığını araştırdığı için daha doğru sonuçlar üretir.

3.4. Metin Sınıflandırma

Metin sınıflandırma, önceden belirlenmiş sınıflara doküman atamayı hedefler [27]. Sınıflandırma yapılmadan önce sınıfların belirlenmesi gerekir. Dokümanların ağırlıklandırılmış değerli vektörel ifadeleri kullanılarak elde edilen benzerlik ölçüm sonuçlarına ve uygulanan algoritmaya göre sınıflandırılması gerçekleştirilir. Metin sınıflandırma, doğal dil metinleriyle çalışan bir sınıflandırmadır [28]. Metin dokümanlarının uzunluğu metin madenciliği çalışmalarındaki en büyük sıkıntıdır. Dokümanların ön işleminden geçirilmesi ve özellik seçimi uygulanması boyutu azaltarak bu sıkıntıyı gideren işlemlerdir. Makine öğrenmesine ihtiyaç duyulmasının nedeni, el ile

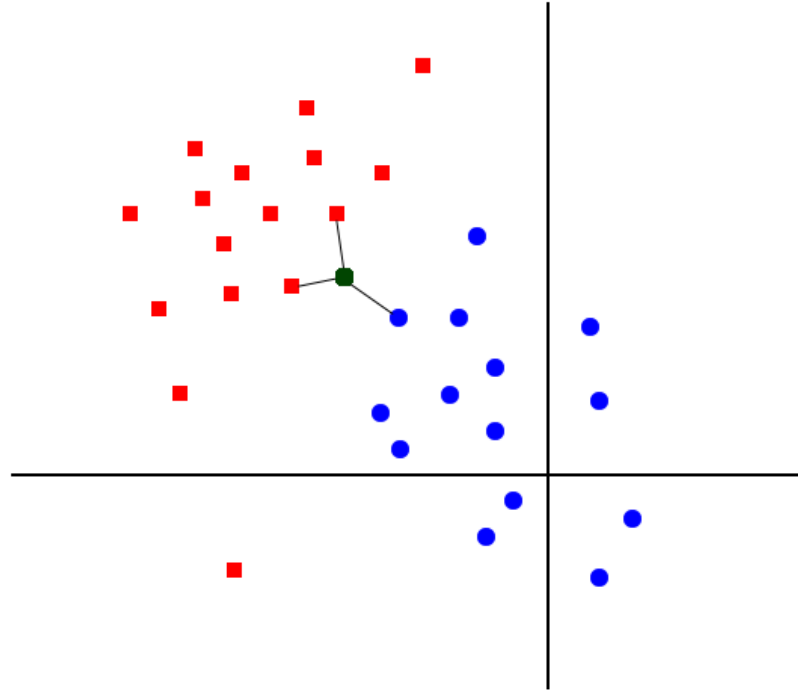
kategorizasyonun pahalı ve zaman tüketen bir iş oluşudur ki, ayrıca elle sınıflandırmada, sınıflandırmayı yapan uzmanların vermiş oldukları kararlara bağlı olarak sonuçlar da değişmektedir. Bu sebeple otomatik metotlar, algoritmalar ve büyük miktarlardaki verilerle çalışan araçlar önemli bir hale gelmiştir [29]. Metin sınıflandırma işlemlerinde eğitim dokümanları vardır. Sınıflandırma yapılırken bu eğitim dokümanları kullanılarak sonuca gidilir. Eğitim dokümanlarının sınıfları, sınıflandırma işlemlerinin karar vermesine yardımcı olur ve özellik seçimindeki sözcükler bu dokümanlardan seçilir. Kullanılan eğitim doküman sayısının azlığı [30], haber metinlerinin kısalığı, her haberin farklı konulardan bahsetmesi sınıfı yansıtan sözcüklerin tespit edilememesine neden olur ve sınıflandırma başarısını düşürür. Metin sınıflandırma işleminde k -NN, Naive Bayes, SVM ve yapay sinir ağları en çok kullanılan sınıflandırma yöntemleridir.

3.4.1. K En Yakın Komşuluk (k -NN) Algoritması

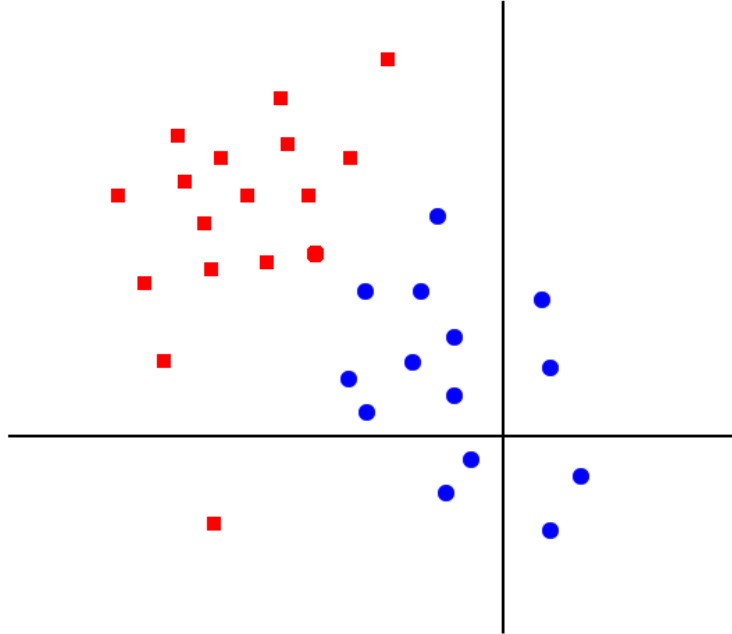
Eğiticişiz öğrenme ve kümeleme işlemlerinde sıklıkla kullanılan algoritma K en yakın komşuluk algoritmasıdır. Kategorize edilmek istenen bireyin, veri kümesinde bulunan K adet veriyle olan uzaklığının bulunması gerekmektedir. Bu algoritmanın kullandığı 3 çeşit uzaklık türü bulunmaktadır ve bunlar;

- Manhattan
- Öklit
- Minkowsky

olarak belirtilebilmektedir. Bu uzaklıklardan en sık kullanılanı ise öklit uzaklığıdır. Veri kümesine gelen yeni elemanın hangi kümeye ait olduğunun bulunabilmesi için öklit uzaklığı kullanılır ve yeni veri en yakın olduğu kümeye dahil edilir. Veri kümeye dahil edildikten sonra kümenin merkezide güncellenmektedir. Aşağıdaki şekillerde k -NN algoritması kullanılarak yeni gelen verinin kümeleme işleminin yapılması gösterilmektedir.



Şekil 3.1. Verinin kümelere olan uzaklığının hesaplanması



Şekil 3.2. Verinin en yakın kümeye dahil olması

3.4.1.1. K En Yakın Komşuluk (k -NN) Algoritması Bit Ağırlıklandırma Algoritması

Bu algoritmada aranılan metnin dokümanda hangi sıklıkla bulunduğuyla vektör ağırlıklandırması yapılmamaktadır. Bir sözcüğün ilgili metinde 2 kez geçmesi veya 5 kez geçmesi o sözcüğün ağırlığını değiştirmeyecektir. Eğer aranılan sözcük metinde varsa 1 yoksa 0 olarak ifade edilmektedir. Burada da lineer algoritması kullanılır ve aranılan kelime bulununca arama işlemi bitirilir ve diğer adımlara bakılmaz.

Eğitim ve sorgu vektörleri önışleme aşamasında ağırlıklandırılmış. Aşağıdaki örnekte bitisel ağırlıklandırılmış örnek vektörler verilmiştir.

$$D1=(0,1,0,1,0,0)$$

$$D2=(0,0,0,1,0,0)$$

$$D3=(1,0,0,0,0,0)$$

$$D4=(0,0,0,0,0,1)$$

$$D5=(0,0,1,0,0,0)$$

$$D6=(0,0,0,1,1,0)$$

$$DS=(0,0,1,0,1,0)$$

Bitisel olarak uzayda ifade ettiğimiz her vektörün sorgu vektörümüze (DS) olan uzaklığını hesaplanarak ve ilişişim denklemi kullanılarak aşağıdaki değerler hesaplanmalıdır.

$$\text{sim}(d1,q)=0$$

$$\text{sim}(d2,q)=0$$

$$\text{sim}(d3,q)=0$$

$$\text{sim}(d4,q)=0$$

$$\text{sim}(d5,q)=0.707$$

$$\text{sim}(d6,q)=0.5$$

Benzerliğı 1'e en yakın olan değeri alınır ve bu değeri $\text{sim}(d5,q)=0.707$ değeriştir. Bu sebepten dolayı verilen metin D5 kümesine aittir.

3.4.1.2. K En Yakın Komşuluk (k -NN) Algoritması Frekans Ağırlıklandırma Algoritması

Bu algoritmanın bir önceki algoritmadan en temel farkı, aranılan sözcüğün metinde geçme sıklığı yani frekansı ağırlıklandırma vektörüne etki etmektedir.

Aşağıdaki örnekte çeşitli önerilen algoritmayla ağırlıklandırılmış rastgele vektörler verilmektedir.

$$D1=(0,2,0,1,0,0)$$

$$D2=(0,0,0,1,0,0)$$

$$D3=(1,0,0,0,0,0)$$

$$D4=(0,0,0,0,0,2)$$

$$D5=(0,0,2,0,0,0)$$

$$D6=(0,0,0,1,2,0)$$

$$DS=(0,0,2,0,1,0)$$

Yukarıda verilen vektörlerin benzerlik değerleri ise aşağıda hesaplanmıştır.

$$\text{sim}(d1,q)=0$$

$$\text{sim}(d2,q)=0$$

$$\text{sim}(d3,q)=0$$

$$\text{sim}(d4,q)=0$$

$$\text{sim}(d5,q)=0.8944$$

$$\text{sim}(d6,q)=0.4$$

Benzerliği 1'e en yakın olan değer alınır ve bu değer $\text{sim}(d5,q)=0.8944$ değeridir. Bu sebepten dolayı verilen metin D5 kümesine aittir.

3.4.1.3. K En Yakın Komşuluk (k-NN) Algoritması Tf-IDF Ağırlıklandırma Algoritması

Metinleri vektörel olarak tanımlamak için kullanılan algoritmalarından birisi de Tf-idf algoritmasıdır. Tf-idf algoritması metin sınıflandırmak için k-NN ve SVM (Support Vector Machine – Destek Vektör Makineleri) ile ilişkilidir. Bu algoritmada frekans önemli bir rol oynamaktadır. Kelimenin geçme sıklığına göre metnin değeri ölçülmektedir (Tf – Terim Frekans değeri). IDF kavramı ise frekansı düşük kelimelerin ağırlıklandırılmasında kullanılmaktadır. IDF tüm eğitim verileri için hesaplanmaktadır. Bir metin belgesinde herhangi bir kelimenin frekansı çok yüksekse, o sözcüğün o metin belgesi için belirleyici olmadığı kanısına varılabilir. Örneğin telgraf metinlerinde sık sık “stop” kelimesi geçmektedir [31]. Bu algoritmada diğer algoritmalar gibi benzerlik temelli çalışmaktadır ve bu algoritmanın birden fazla versiyonu bulunmaktadır [32]. Terim ağırlığı ve frekans kimliği Eşitlik 3.3 ve 3.4’de verilmiştir.

$$w_{d,t} = tf_{d,t} \cdot idf_t \quad (3.3)$$

$$idf_t = \log\left(\frac{D}{df_t}\right) \quad (3.4)$$

tf (Dokümanda bulunan kelimenin dokümandaki görülme sıklığı)

D (Toplam doküman sayısı)

df_t (Doküman Frekansı (Kaç eğitim dokümanında ilgili kelime geçmiş))

Bu algoritmanın çalışmasında öncelikle frekans vektörleri bulunur ve frekans vektörlerine göre işlemler yapılır. Aşağıda örnek vektörler verilmiştir.

D1=(0,2,0,1,0,0)

D2=(0,0,0,1,0,0)

D3=(1,0,0,0,0,0)

D4=(0,0,0,0,0,2)

D5=(0,0,2,0,0,0)

D6=(0,0,0,1,2,0)

DS=(0,0,2,0,1,0)

D1 dokümanı için tf- idf ağırlığının hesaplanması aşağıdaki gibidir.

Tf(söz)=2

D=6

dft=1

IDF=log(D/dft)=log(6/1)=0.778

Tf*IDF=2*0,778=1,556

W1=(0,1.556,...)

Yukarıdaki işlemler yapılarak “söz” kelimesinin D1 belgesine göre ağırlığı bulunmuştur. Benzerliği 1’e en yakın olan değer alınarak işlem tamamlanır.

3.4.2. Naive Bayes Olasılık Yöntemi

İstatistiksel sınıflandırma modeli tekniklerinden biri olarak kullanılan “Naive Bayes” tekniği, olasılık hesaplama temeline dayanan bir tekniktir. Daha önceden kayıtlı olan belirli bir sınıfa yerleştirilmiş olan veriler kullanılarak, yeni verilerin mevcut olan sınıflardan hangisine dâhil olacağı olasılığının hesaplanması yöntemidir [33].

Bu teknik, herhangi bir anlaşılmaz veya belirsizlik taşıyan bir durumun modelini oluşturarak, bu durumla ilgili evrensel doğrularla ve gerçekçi gözlemler ışığında belirli sonuçlara ulaşabilmektedir. Çıkan sonuçlar üzerinde olasılık hesaplamaları yapılarak belirsizlikler giderilmeye çalışılır. Bu algoritma ile örneğin, eğitim sektöründe bir kişinin yaptığı nitel yorumun polaritesi tahmin edilebilir. Ayrıca müşterilerin harcama artışlarının belirlenmesi de bu tekniğe örnek olarak gösterilebilir. Belirli adet müşterinin yaptıkları alışveriş verilerinin yanında müşterilerin eğitimi, yaş, cinsiyet, gelir gibi verileri de müşterilerin harcama artışlarının olasılıklarının çıkarılmasında kullanılabilmektedir [34].

“Naive Bayes” tekniği, metin belgelerinin sınıflandırılmasında; metin madenciliği uygulamalarında da sıklıkla kullanılmaktadır. Büyük veri tabanları arasından hızlı olarak istenilen kuralların çıkarılmasında da başarılı sonuçlar aldığı literatür taramalarında gözlemlenmiştir.

3.4.2.1. Multi-Variate Model

Sınıflandırma işlemlerinde her sınıf ayrı olmak üzere, sınıfı oluşturan dokümanların sözcükleriyle işlem yapılır. Diğer sınıfların dokümanları ve sözcükleri hesaplama işlemlerinde dikkate alınmaz. Sözcüğün, sınıfın dokümanlarında geçip geçmediğiyle ilgilenir.

$X=\{x_1, x_2, x_3, \dots, x_n\}$ sınıflandırılması yapılacak olan vektördür. Eşitlik 3.5’de w_i sözcüğünün c_k sınıfında geçme olasılığı formülü verilmiştir [35]. $g_{s_{w_i}}$, w_i sözcüğünün c_k sınıfındaki belgelerde geçme sayısı; d_{s, c_k} sınıfındaki toplam belge sayısıdır. Eşitlik 3.5’de bulunan paydaki 1 ve paydadaki 5 ifadeleri bir kelimenin sınıf içerisinde geçmemesi durumunda sonucu 0 yapmasını engellemek içindir. Bütün sınıflardaki belge sayıları eşit ve 5 sınıf olduğu için pay 1, payda 5 verilmiştir ve bu sonucu etkilememektedir. Çeşitli sıfır olasılık sorunu gidericiler kullanılmıştır [36].

$$p(X|c_k) = \frac{1+g_{s_{w_i}}}{5+d_{s, c_k}} \quad (3.5)$$

Eşitlik 3.6’de gösterilen formül Eşitlik 3.7’deki gibi değiştirilmiştir. X vektörünün c_k sınıfında olma olasılığı ve test belgesinin sınıfı olan $C(X)$ formüle edilmiştir. En yüksek değerine sahip sınıf test belgesinin sınıfıdır [36]. wb_{w_i} , w_i kelimesinin bit olarak ağırlığını ifade etmektedir.

$$C(X) = \operatorname{argmax}_{c_k} p(X|c_k) = \operatorname{argmax}_{c_k} \prod_{i=1}^n p(w_i|c_k)^{w_{wi}} (1 - p(w_i|c_k))^{1-w_{wi}} \quad (3.6)$$

$$C(X) = \operatorname{argmax}_{c_k} p(X|c_k) = \operatorname{argmax}_{c_k} \sum_{i=1}^n \log(p(w_i|c_k)^{w_{wi}} (1 - p(w_i|c_k))^{1-w_{wi}}) \quad (3.7)$$

3.4.2.2. Multi-Nominal Model

$X=\{x_1, x_2, x_3, \dots, x_n\}$ sınıflandırılması yapılacak olan vektördür. Eşitlik 3.8’da w_i kelimesinin c_k sınıfında geçme olasılığı formülü verilmiştir [36]. $g_{s_{wi}}$, w_i kelimesinin c_k sınıfında geçme frekansını; t_s , c_k sınıfındaki toplam kelime sayısını; V , c_k sınıfındaki toplam belge sayısını belirtir.

$$p(w_i|c_k) = \frac{1+g_{s_{wi}}}{V+t_s} \quad (3.8)$$

X vektörünün c_k sınıfında olma olasılığı, her sınıfta eşit sayıda belge olduğundan Eşitlik 3.9’de gösterilen formül değiştirilerek Eşitlik 3.10’deki gibi revize edilmiştir. $p(c_k)$ sınıfının olasılığını; k ise toplam kategori sayısını göstermektedir. Olasılığı gösteren en yüksek değerli *test* belgesinin sınıfıdır ve $C(X)$ şeklinde ifade edilir.

$$C(X) = \operatorname{argmax}_{c_k} p(X|c_k) = \operatorname{argmax}_{c_k} p(c_k) k! \prod_{i=1}^n \frac{p(w_i|c_k)}{tf_{w_i}!} \quad (3.9)$$

$$C(X) = \operatorname{argmax}_{c_k} p(X|c_k) = \operatorname{argmax}_{c_k} p(c_k) k! \sum_{i=1}^n \frac{\log(p(w_i|c_k))^{tf_{w_i}}}{tf_{w_i}!} \quad (3.10)$$

3.5. Metin Madenciliği ve Doğal Dil İşleme Yöntemleri

Metin Madenciliğinde bilgi elde etmek için kullanılan en verimli yöntemlerden birisi de doğal dil işleme (NLP) yöntemleridir. Doğal dil işleme kullanılarak metinden daha anlamlı bilgiler çıkarılmaya başlanmıştır. İnsanların kullandıkları doğal dili çözümleme, anlama, yorumlama ve üretme doğal dil işlemenin ana işlevlerinden ve bu işlevi yerine getirebilmek için bilgisayar bilimlerinden faydalanan yeni bir mühendislik disiplini

ortaya çıkmıştır. İnsan- makine etkileşimi, yapay zekâ, robotik vb. çalışmalar doğal dil işleme sayesinde daha ileri düzeye çıkartılmıştır [37].

Metin Madenciliğinde, belgelerin analizi için, içeriğinin anlamını taşıyan kavramların tespit edilmesi gerekmektedir. Bu kavramlar kelimeler veya kelime grupları ile ifade edilir ve terimler olarak adlandırılır. Belge içindeki terimlerin çıkarılması başlı başına bir konudur ve doğal dil işleme çalışmaları kapsamında incelenen bir alandır. Doğal dil işlemenin belge analizi sürecindeki en önemli faydası terimlerin yani kelimelerin ayrıştırılması, eklerinden arındırılarak anlamını kaybetmeyen en kısa biçimlerine dönüştürülmesidir. Çünkü aynı anlam için kullanılan kelimeler dilbilgisi kuralları gereği farklı biçimlerde bulunabilir ve bu farklı kullanım biçimleri ortadan kaldırılmadığı takdirde farklı anlam taşıyan terimler gibi işleme alınarak, belgelerin gerçek anlamına ulaşılmasını engelleyebilirler. Doğal dil işleme çalışmaları kapsamında yürütülen girişimler dört ana grup altında toplanabilir [38].

- Biçim birimsel çözümleme (Morfolojik analiz)
- Sözdizimi çözümlemesi (Sentaktik çözümleme)
- Anlam çözümlemesi (Semantik çözümlemesi)
- Anlam kargaşasının giderilmesi

a) Morfolojik Analiz (Biçimbirim): Biçimbirim sözcüklerin yapısıyla ilgili. Türkçe için sözcüklerin türetilmesi ve ekler çok önem taşır. Her dilde iki farklı şekilde sözcük oluşturulabilir. Bunlardan biri çekim, diğeri ise türetme yöntemidir. Çekim yoluyla sözcük oluşturulurken bir sözcüğün farklı şekilleri kullanılır. Türetme ise var olan eski sözcüklere yapım ekleri eklenmesi yoluyla yeni sözcük oluşturma yöntemidir [38].

b) Sözdizimsel Analiz (Sentaktik Analiz): Var olan sözcük kümesinin anlamlı bir cümle oluşturup oluşturmadığını tespit edebilmek için sentaktik analizin yapılması gerekmektedir. Sentaktik analiz, doğal dil işleme modelinin oluşturulabilmesi için ön şarttır. Kelimelerin rastgele bir araya gelmesinden, anlamlı bir cümle oluşmayacaktır. Doğru bir cümle yapısı oluşturulmadığında, o cümleyi doğal dil işlemeyle analiz etmek yersiz olacaktır. “küçük koş mavi” kelimeler grubu anlamlı bir yapıda değildir [38].

Sözdizimsel analiz, morfolojik analiz yardımıyla cümlelerin yapısal bir tanımını oluşturan bir yöntemdir. Sözdizimsel analizin temel amacı, kelime yığınlarından cümle birimlerini elde etmektir. Cümle birimleri, tamlamalar, kelimeler ve benzeri gibi ifadeler cümle birimlerini oluşturmaktadır. Türkçede cümle yapısı özne, nesne ve yüklem

(Ö+N+Y) ögelerinden oluşmaktadır. Anlam arttıkça cümlelerde, cümlelere yer tamlayıcısı ve zaman tamlayıcısı gibi yeni ögeler yer alacaktır. Ayrıca, edat, bağlaç gibi unsurlar cümlelerin anlamlarını kuvvetlendirmek için eklenebilir.

Sınıflandırma dillerin sözdizimi açısından önem taşımaktadır. Çeşitli dillerde cümlelerin temel ögeleri olan özne (Ö), nesne (N) ve yüklem (Y) düz cümledeki temel dizilişinin Ö-Y-N, Y-Ö-N ve Ö-N-Y şeklinde olduğu görülmüştür. Ayrıca, kullanılan diğer ögeler, dolaylı tümleç (D), zarf tümleci (Z) olarak kabul edilebilir. Türkçenin anlam analizinden dolayı, cümlede vurgulanan ögeler cümlelerin yüklemine yaklaştırılır [39].

c) Anlamsal Analiz (Semantik Analiz): Herhangi bir cümlede anlatılmak istenenin anlaşılması yani cümlede anlatılmak istenen duygu veya düşüncenin analizi anlamsal olarak yapılır.

Kavram, bir nesnenin zihnindeki tasarım biçimi; kelime ise bu biçimin göstergesidir. Gösterilmek istenilen, kavram; gösteren, kelime olarak adlandırılabilir. Kelimelerin belirttiği kavramların, çevredeki her türlü olayların, hareketlerin ve hayallerin zihnindeki değerlendiriliş biçimine anlam denir. İnsanoğlu, doğada olup biten her şey için zihninde bir kavram oluşturur. Bu kavramların zihnindeki değerlendiriliş biçimleri, anlamı meydana getirmektedir. Kelimeler, insanların zihinlerinde meydana getirdikleri canlanma ile anlam kazanırlar. Burada özetle, “kelimeler yansıttıkları kavramlara göre değerlendirilirler ve anlamlı duruma gelirler” denilebilir.

Anlamsal analiz yapılırken, öncelikli olarak kelimelerin tek tek veritabanından uygun nesnelerle eşleştirilme işleminin yapılması gerekir. Bu işlem, her zaman birebir eşleme olmayabilir. Diğer bir deyişle, kelimelerin ifade ettikleri anlamlar her zaman bir tane olmayabilir. Ayrık kelimelerin bir cümledeki doğru anlamını bulma işlemine “kelime anlam berraklaştırılması” denir. Bu işlem, cümle içinde geçen bir kelimenin sözlükteki anlamlarının belirlenip bunlardan uygun olanının seçilmesidir. Cümle içinde geçen her bir kelime, diğer kelimelerin doğru anlamlarının ortaya çıkarılması için önem taşımaktadır.

3.6. Metin Madenciliğinin Uygulama Alanları

Metin madenciliği, metinlerin bulunduğu her platformda kullanılabilir. Metin madenciliğinin kullanıldığı alanlar aşağıda gösterilmiştir [36].

a. Müşteri ilişkileri yönetimi

- b.** Sahtekârlık tespiti
- c.** Sağlık alanı
- d.** Pazar arařtırmaları
- e.** Metinlerden bilgi çıkarımı
- f.** Doküman özetleme
- g.** Doküman sınıflandırma
- h.** Benzer içerikleri belirleme
- i.** Web içerikleri sınıflandırma
- j.** Yazar tanıma sistemleri
- k.** Soru-cevap sistemleri

Yukarıda belirtilen alanlarda ve birçok alanda metin madencilięi kullanılmaktadır. Özellikle sosyal medyanın yaygınlaşmasıyla beraber, bireylerin bu medya araçları aracılığıyla metin paylaşılması çoęalmıştır. İnternette paylaşılan büyük verinin yüksek bir çoęunluğu yapısal değildir. Eğer bu veri yapısal hale getirilirse paha biçilemez değerde bir bilgi altyapısına sahip olunacaktır. Bu tezde metin verisinden bilgi elde edebilmek için kullanılan yöntemler arařtırılıp literatür taraması yapılmıştır. Tez çalışmasında ise çok daha hızlı kod optimizasyonu yapılmış yeni bir metin madencilięi sistemi geliřtirmek istenmektedir.

4. FİKİR MADENCİLİĞİ

İnteraktif eğitim, sosyal medya ve internet kullanımının yaygınlaştığı günümüzde, eğilimleri, marka değerlerini, bir konunun nasıl anlaşıldığını vb. durumları tespit edebilmek için fikir madenciliğinin kullanılmasının gerekliliği ortaya çıkmıştır. Ayrıca sosyal medya üzerinden elde edilen veriler kullanılarak fikir madenciliği yapılmakta ve ülkelerde büyük çaplı olaylar meydana gelmektedir. Örneğin Mısır’da Hüsnü Mübarek’e karşı muhalifler twitter’da organize olup ve ayaklanma başlatmışlardır. Bu bölümde, “Fikir Madenciliği” ile ilgili konular incelenecektir.

4.1. Fikir Kavramı ve Bileşenleri

Fikir; kelimesi TDK’ da 1. Düşünce, 2. Düşün, 3. Kuruntu olarak tanımlanmaktadır. Fikir; a, b hakkında ne düşünüyor? Sorusuna cevap veren bir cümle parçası olarak tanımlanır. Fikirler doğrudan ifade edilmelidir. Açıklayanın kelime seçimine bakarak onun ne hissettiği hakkında birisinin yapabileceği anlam çıkarmalarını içermez. Fikirler geleceğe ait tahminler içermez [40].

Fikir öznel bir ifadedir. Bir fikir bir kanıt tarafından destekleniyor olabilir, buna rağmen insanlar aynı gerçeklerden yola çıkarak zıt fikirde bulunabilirler. Fikirler yeni kanıtlar sunulmaksızın nadiren değiştirilebilir, bu ancak gerekçeli olabilir.

Dünyada metinsel enformasyonlar geniş ölçüde iki ana kategoride sınıflandırılabilir: Gerçekler ve Fikirler. Gerçekler, olaylar ve varlıklar hakkındaki nesnel ifadelerdir. Fikirler ise, insanların olaylar ve varlıklar hakkındaki algı veya duygularını yansıttıkları sübjektif açıklamalardır. Metin enformasyon işleme üzerindeki mevcut araştırmaların çoğu, neredeyse sadece gerçek enformasyonu elde etme, web araması, diğer birçok metin madenciliği ve doğal dil işleme görevleri üzerinde yoğunlaşmıştır. Sadece yakın zamanlara kadar fikirleri işleme üzerinde küçük çalışmalar yapılmıştır. Fikirler önemlidir, birisi bir konuda karar vermek istediğinde, o konu ile ilgili diğer fikirleri bilmek ister. Bu yalnızca kişiler için değil, organizasyonlar için de geçerlidir. Amerika’da 2000 den fazla kişi üzerinde yapılan iki araştırmadan elde edilen sonuçlar bu öngörüye destekler niteliktedir ve sonuçlarından bazıları aşağıdaki şekildedir.

- İnternet kullanıcılarının % 81'i en az bir kez bir ürün üzerinde çevrimiçi araştırma yapıyor,
- İnternet kullanıcılarının % 20'si satın alma amacı olmadan da çevrimiçi araştırma yapabiliyor,
- Lokanta, otel ve çeşitli servislerin çevrimiçi yorumlarının okuyucularının %73-%87'si yorumların satın almalarında önemli bir etkiye sahip olduğunu söylüyor,
- Müşteriler %20-%99 oranında bir 5 yıldızlı ürün için, bir 4 yıldızlı ürün fiyatından daha fazlasını ödemeye istekli olduklarını söylüyorlar,
- Müşterilerin %32'si bir çevrimiçi derecelendirme sistemi aracılığıyla bir ürün, servis veya kişi hakkında değerlendirme yapmışlar ve % 30'u ürün veya servis için bir çevrimiçi yorum göndermişlerdir.

Fikrin bileşenleri aşağıdaki şekilde açıklanabilir:

Fikir sahibi: Belirli bir nesne üzerine özel bir fikre sahip kişi, grup veya organizasyondur.

Nesne: Üzerinde bir fikir ifade edilen ürün, kişi, olay, organizasyon, konu veya hatta bir fikirdir.

Fikir: Bir fikir sahibinden bir nesne üzerine bir bakış, tutum veya değerlendirmedir.

Bir fikir çoğu kez duygu kelimeleri içerir, bu fikir pozitif, negatif ve nötr kutuplar olarak sınıflandırılabilir. Örneğin; “Cemre, Veysel’in sevimsiz bir tip olduğunu söyledi” cümlesi Veysel açısından negatif kategoride değerlendirilebilecek bir cümledir.

Güncel arama motorlarında, nesneler kelimeleri ile anlatılabilir. Arama motorları fikirleri aramaz. Fikirleri birkaç kelime ile anlatmak zordur (suyla çalışan bilgisayarlar hakkında insanlar ne düşünüyorlar gibi). Geçerli arama yöntemleri fikir arama için uygun değildir. Bu olumsuzluklar fikir madenciliğinin çıkış noktasını oluşturmuştur.

4.2. Fikir Madenciliği ile İlgili Çalışmalar

Üreticiler, markaları hakkında bilgi edinmek üzere doğru olsun ya da olmasın rekabetçi bilgi, aykırı hareketler veya halkla ilişkiler aktiviteleri için her zaman medyayı

izlemeye ihtiyaç duymuşlardır. Technorati, günlük olarak yeni yaratılan 75.000 bloğu ve ürünler hizmetler üzerine tartışma içeren birçok tüketici fikirlerini (her gün 1.200.000 yeni gönderi) ortaya çıkarmaktadır. Gazete kupürü kesip ilgililere gönderen servisler, saha ajanları ve geçici araştırmalar gibi taktikler bu duruma kolayca uyum sağlayamamıştır. Bölümlere ayrılan medya ve değişen tüketici, bilinen izleme metotlarını bozmuştur.

Bu konu hakkında ne düşünüyorsunuz? Bu ifade fikir madenciliğine seslenmek için çalışan bir esastır. Bu soru bilgi değişimi için hazır bir platform sağlayan, insanların kişilik ve görüşlerini gösteren çevrimiçi ortamlarda daha önemli olur. Bireyler sosyal medya ortamlarında duygularını yazar ve paylaşır, ürünlerin olumlu veya olumsuz yönlerini anlatır, bireylerin sorunları için onların sesi olurlar. Fikir madenciliği çevrimiçi ortamlarda bulunan bilgileri daha iyi kullanmak üzere bireylere yardım edebilir ve çeşitli sorunlar için kararlarına destek olabilir. Örneğin, tüketici sorun ve memnuniyetinde fikir madenciliğinin sonuçları, yeni müşterileri etkileyecektir.

Fikir Madenciliği veya Duygu Analizi; Bilişimsel Dil Bilim, Metin Madenciliği ve Doğal Dil İşlemenin geniş bir alanına işaret eder. Genel olarak, bazı konularla ilgili olan bir konuşmacının tutumunu tanımlamaya yardım eder. Tutum onların değerlendirme, yazarken içinde bulundukları duygusal durum veya duygusal iletişim amacı olabilir.

Fikir Madenciliği, nesne hakkında fikirleri içeren doküman setinde, o nesnenin nitelik ve bileşenlerini belirlemeyi sağlar. Setteki her doküman nesne ile ilgili fikirlerin pozitif, negatif veya nötr olup olmadığını belirlemede kullanılır.

Fikir Madenciliği, herhangi bir konu üzerinde istatistiksel modeller kullanılarak fikrin kümelenmesine denir. Sınıflandırma, verilen bir kategori üzerine tarafsız, karşı fikir sahibinin tutumunu kategorize eder. Fikir madenciliği geçerliliği kanıtlanmış bir teknolojidir ve son zamanlarda savunma, kamu yönetimi ve pazarlama uygulamalarının geniş bir türünde bir dinleme, analiz ve bağlantı aracı olarak kullanılmaktadır [41].

Fikirlere ilişkin metin analizi, meşru bir araştırmacı için kıymetli olabilir. Bu araştırmacı bir ürün için sürekli bilgi veya meşru bir sonuç üzerine bakış açısı edinmek isteyebilir.

Fikir madenciliği kuruluşlar için de olumlu sonuçlar doğurabilir, onlara verdikleri hizmet ve nasıl ürün sunduklarına ilişkin uygun bir resim elde etmek isteyebilirler. Bireysel tüketiciler açısından bakıldığında; tüketicilerin bir ürün satın almak istediklerinde mevcut kullanıcıların fikirlerinin bir özetini görmeleri yararlıdır, bu şekilde bilinçli bir karar alabilirler. Bu, ürünün güçlü ve zayıf yönlerinin zihinsel bir resmini biçimlendirmek

için çok sayıdaki yorumu okumaktan daha iyidir. Kişi aynı zamanda rakip ürünlere ait fikirlerin özetini karşılaştırabilir. Fikir madenciliği aynı derecede kuruluşlar ve işletmeler için de önemlidir. Örneğin, bir ürün üreticisi için, ürünlerin ve rakiplerinin müşteri tarafından nasıl algılandığını bilmek önemlidir. Bu bilgi yalnızca pazarlama ve ürün kıyaslama için değil, aynı zamanda ürün geliştirme ve tasarımı için de yararlıdır [41].

Aşağıda fikir madenciliği konusunda konferanslarda yer almış çalışmalardan örnekler verilmiştir:

- Yasal bloglarda fikir madenciliği [42]
- Fikir madenciliği için kullanılan terimin gagesini ve oryantasyonunu belirleme [43]
- SentiWordNet: Fikir madenciliği için herkesçe erişilebilir metinsel bir kaynaktır [44]
- PageRanking WordNet Eşanlımlı Set: Fikir madenciliği için bir uygulama [45]
- Ekonometri yardımıyla fikir madenciliği: İtibar sistemleri üzerine bir durum çalışması [46]
- Bir fikir madenciliği uygulaması için internetteki niche tarayıcıları (özellikle internet üzerinden pazarlama faaliyetlerinde bulunan kişiler tarafından kullanılan ve ürün incelemeleri gibi konularda kullanıcıya kolaylık sağlayan bir araç) birbirine bağlama ve etki analizi [47]
- Müşteri görüşlerinden fikir madenciliği [48]
- Pozitif dönütlere bakarak fikirlerin yoğunlaşma ilişkileri madenciliği [49]
- Bir konuda sabit fikirleri ifade eden kelimeler vasıtasıyla blog madenciliği [50]
- Fikir Madenciliğinde ilişkilerin durumu ve değerlendirme durumu çıkarımı [51]
- HTML dokümanlarının büyük koleksiyonundan duygu analizi için sözlük inşa etme [52]

4.3. Fikir Madenciliği Görevleri

Fikir madenciliği görevleri, fikir madenciliği ile yapılabilecek işleri kapsamaktadır. Fikir madenciliği içinde çeşitli alt görevler tespit edilebilir [53].

Metnin SO (Subjective-Objective: Öznel-Nesnel) Kutbunu Belirleme: Karar vermede, verilen bir metnin gerçekçi bir doğaya sahip olup olmadığını, yani verilen bir

durum veya olayı pozitif veya negatif bir fikir belirttiğine bakmaksızın tanımlar. Bu görevde fikrin öznel ya da nesnel gruplarından hangisine dâhil olduğunun tespiti esastır.

Metnin PN (Pozitive-Negative: Pozitif-Negatif) Kutbunu Belirleme: Karar vermede, verilen bir metin eğer öznel bir ifade ise, fikir metninin pozitif veya negatif kutuplardan hangisine dâhil olduğunu belirlemek esastır.

PN Kutup Metnin Şiddetini Belirleme: Karar vermede, örneğin pozitif kutup olarak ifade edilmiş bir metnin zayıf, orta ya da yüksek derecede pozitif ifadelerden hangisine dahil olduğunu belirlemek esastır.

Metnin SO kutbunu belirleme, sık kullanılan bir görev değildir.

Fikir madenciliği görevlerini Leneve ise üçe ayırarak gruplandırmaktadır [54].

Duygu sınıflandırma: Duygu sınıflandırma, metnin duygusunu tanımlamasıdır. Kapsam belge düzeyinde veya nesneye yönlendirilmiş olabilir [55]. Duygu analizinde yapılan çalışmalar çoğunlukla blog, haber ve yorumların analizi üzerine yapılmıştır [56]. Duygu sınıflandırma, metin sınıflandırma problemi gibidir. Metni olumlu veya olumsuz olarak sınıflandırır. Örneğin, verilen bir ürün yorumunda sistem yorum ifadesinin pozitif ya da negatif gruplardan hangisine dâhil olduğunu belirler. İnsanların ne sevdiği ya da ne sevmediği hakkında hiçbir detay keşfedilmez.

Özellik tabanlı madencilik ve özetleme: Yorumcular tarafından yapılan ürün özelliklerine ilişkin tespitleri yorumlar, yorumların pozitif ya da negatif olup olmadığını belirler. Bu cümle düzeyinde yapılır. Örneğin, “Bugün hava kapalı” cümlesi, hava özelliği açısından negatif bir yorumdur.

Karşılaştırmalı cümle, ilişki çıkarma ve sınıflandırma problemi olarak sorun konu etme: Karşılaştırmalı cümle ve ilişki madenciliği ise, bir nesneyi bir veya benzer nesnelere karşı karşılaştırır. Örneğin, “X şehrinin havası Y şehrinin havasında daha soğuktur” cümlesi ile karşılaştırma ilişkisi çıkarılabilir.

Fikir madenciliği görevlerine ilişkin başka bir değerlendirmede Liu şu iki kategoriden bahsetmektedir [57].

Direk Fikirler: Bireyler, konular, vakalar, ürünler gibi bazı nesneler üzerine duygu ifadelerine direk fikirler denmektedir. Örneğin: “Bu oyunun görsel efektleri çok yüksek” (öznel)

Karşılaştırmalar: İlişkileri ifade etme, yani bir nesneden daha fazla benzerlikler ya da farklılıklar belirtmedir. Genellikle bir sıra ifade edilir. Örneğin: “A buzdolabı B

buzdolabına göre daha az elektrik tüketmektedir” (öznel veya nesnel) Wiebe ve Mihalcea, öznellik ve semantik arasındaki etkileşimi araştırdıkları çalışmalarında duyguları, “öznel, nesnel ve her ikisi” biçiminde üç gruba ayırmıştır. Bir duyguyu öznел olarak sınıflandırmanın şu şekilde tanımlamışlardır: “Duygu bir metinde veya konuşmada kullanıldığı zaman onun öznellik belirttiğini düşünüyoruz, aynı zamanda şunu da düşünüyoruz, onu içeren cümle veya cümlecik öznел olmalı.” [58].

Qiu vd. ise, görevleri cümle ve doküman, duygu sınıflandırma ve özellik tabanlı fikir özetleme olarak belirtmektedir [59].

Son görevlendirme çeşidi Tablo 4.1.’de görülmektedir.

Tablo 4.1. Fikir Sınıflandırma Görevine İlişkin Bir Taksonomi

GÖREVLER	
Kategori	Tanımlama
Sınıflar	Pozitif/negatif duygular veya öznел/nesnel metinler
Düzey	Doküman veya cümle / sınıflandırma düzeyi cümlecği
Kaynak/Hedef	Duygunun kaynağının/hedefinin bilinip bilinmemesi veya çıkarılıp çıkarılmaması
ÖZELLİKLER	
Kategori	Örnekler
Sözdizimsel	Kelime etiket n-gramları, cümlecik desenleri Noktalama
Anlamsal	Kutupluluk etiketleri, ekspertiz grupları, anlamsal yönlendirme
Link Tabanlı	Web linkleri, gönder/cevapla desenleri ve doküman tahminleri
Biçimbilim	Stilin sözcükleri ilgili ve yapısal ölçümleri
TEKNİKLER	
Kategori	Örnekler
Makine Öğrenmesi	SVM, NaiveBayes gibi teknikler
Link Analizleri	Tahmini analizler ve gönder/cevapla mesaj desenleri
Benzerlik Skoru	İbare desen eşleştirme, sıklık sayıları
ALANLAR	
Kategori	Tanımlama
Yorumlar	Ürün, film ve müzik yorumları
Web Konuşması	Forumlar ve bloglar
Yeni makaleler	Online yeni makaleler ve web sayfaları

Yukarıdaki açıklamalardan görüldüğü gibi, fikir madenciliği görevleri her ne kadar farklı yazarlar tarafından farklı şekillerde tanımlansa da, ortak bazı noktaları bulunmaktadır.

Farklılıkların sebebi daha çok amaca dayalıdır. Bu çalışma verilen bir fikrin pozitif/negatif kutbunu belirlemeye yönelik olduğundan, fikir sınıflandırma görevi aşağıda, literatürde yer alan çalışmalar ile birlikte incelenmiştir.

4.4. Fikir Sınıflandırma

Müşteri İlişkileri Yönetimi gibi faktörler, birçok organizasyon için gün geçtikçe artan bir başarı faktörü haline geliyor. Dolaylı müşteri fikirlerinin takibini yapmaya gittikçe daha fazla acil ihtiyaç duyuyoruz. Bu fikirler organizasyonlara direk olarak söylenmiyor, ama bazı üçüncü şahıs kurumlara yönlendiriliyor. Örneğin, sahip oldukları ürünler ve hizmetler konusunda müşteri fikirlerini toplayan siteler bulunmakta. Bu siteler muhtemelen legal tavsiyeler sağlıyorlar. Şikâyetler ve takdir edilen bu bilgiler genellikle arşivleniyor ve diğer tüketicilere erişilebilir tutuluyor. Daha da görünür bir şekilde kullanıcılar; haber grupları, forumlar, chat toplulukları gibi topluluklarla fikir paylaşımı içinde “özgürce ortaklık” eğilimindeler. Bu bilgi saklayıcıları içinde saklanan doküman setleri, organizasyonlar ve ürünleri için değerli birer bilgi kaynağıdır [60]. Bu nedenle analiz edilmeleri gerekir ki, bu analizlerden biri de fikir kutbunu belirlemeye yönelik fikir sınıflandırma.

Ürün yorumlarını sınıflandırma, fikir madenciliğinde genel bir problemdir ve bu problem için denetimli ve denetimsiz öğrenme içeren çeşitli teknikler kullanılmıştır. Fikir madenciliğine yönelik bazı yaklaşımlar, daha önceden pozitif ve negatif olarak tanımlanmış fikir kelimelerini gerektirir ki, bu metot Turney [61] metodu olarak bilinir. Burada, mükemmel ve zayıf kelimeleri elemek için onların PMI (Point-Wise Mutual Information)’larını hesaplamak yoluyla sözcüklerle ilgili anlamsal yönlendirmeyi tanımlamak esastır [62].

Finansal haberler alanında, bu haberlerle ilişkili iki probleme ilgi duyulmuştur: Bir haber parçasının kutbunu belirleme ve yedi puanlık bir ölçekte (çok pozitiften çok negatife) bir metni sınıflandırma [63]. Önerilen sınıflayıcı dayanağına göre, pozitif ve negatif ayrım % 46’lık bir doğruluğa sahiptir.

Dave ve diğerleri, özellikler (kelimeler veya terimler) yaratmak için çeşitli teknikler ortaya koymuş ve bir sınıflandırma görevi için eğitim setinden skorları birleştirmiştir. Bu

görev, ürün yorumları için ortak pozitif ve negatif ifadeleri incelemekten oluşur ve onların sınıflayıcıları özelliklerin skorlarını bir araya toplar [64].

Sınıflandırma ile ilişkili bir diğer problem de sınıflandırma etkisidir. Bu etki, duyguların yerine ruh hallerini sınıflandırmak için çalışır. Bu sınıflar korku, öfke, üzüntü, mutluluk vb. içerir.

Bethard ve diğerleri, “Fikir-Önerileri Çıkarma ve Fikir Sahiplerinin Sözdizimsel ve Sözcüklerle İlgili Fikirlerini Kullanma” isimli çalışmalarında tek ve çift katmanlı mimarilerden söz etmektedirler [40]. Tek katmanlı mimari bir kurucu-kurucu sınıflandırma projesidir. Cümlelerin sözdizimsel ağacında her bir kurucu fikir-öneri veya anlamsız olarak sınıflandırılır.

Son olarak Yu ve diğerleri “Politik Fikir Sınıflandırma için Fikir İfadelerinin Karakteristiklerini Keşfetme” isimli çalışmalarında, duygu sınıflamasına ilişkin iki ana yaklaşım olduğunu belirtmektedir [65].

4.5. Fikir Madenciliğinde Karşılaşılan Problemler

Fikir Madenciliğine ilişkin zorlukları Berend ve Farkas tarafında şu şekilde açıklanmıştır; “nesnelliği tanımlama, pozitif/negatif duygu yükü ve HTML dokümanlarının etkili veritabanının yapısal analizine dayanan cümlecikler şeklinde sıralamıştır”. Metin içindeki kutup ifadelerini belirleme görevine destek vermek üzere çeşitli dil kaynakları geliştirilmiştir.

Fikir madenciliği, internetin ve akıllı cihazların yaygın olarak kullanılması paralelinde doğmuş güncel bir konudur. Ancak bu alanda karşılaşılabilecek birçok problemde mevcuttur. Sosyal medyada paylaşılan yazılar ve yorumları analiz edip bunlardan fikir çıkarımı yapmak çok büyük bir sorun olarak karşımıza çıkmaktadır. Yazılan metinlerde bir standardın olmaması, farklı lehçelerle yazıların yazılması, bazı kelimelerin yerine kullanılan kısaltmalar, yazım hataları, devrik cümleler, sanatsal yazımlar vb. kullanımlar metinlerden fikir elde edilmesinin önüne ket vurmaktadır. Örneğin bir gazetede yer alan haber sosyal medyada farklı bir şekilde paylaşılmaktadır. Kişinin o habere getirdiği subjektif yorumlar, eğitim durumu ve dili kullanma yeteneği aynı haberin farklı bir biçimde yorumlanmasına sebep olmaktadır. Fikir madenciliği konusunda karşılaşılan zorluklardan biri de, belirlenmiş fikirlerin şiddeti konusunda çalışmaktır. Tespit edilen fikir

bir tartışma sırasında değişebiliyorsa, bu fikirle ilgili ortaya atılan antitezlerin daha güçlü olduğunu göstermektedir [66].

Fikir madenciliği ve duygu analizi ile ilgili gelişme, potansiyel uygulamaları nedeniyle kısmidir. Yeni düşünsel zorluklar da aynı derecede önemlidir ve araştırmacılara bu alanı sunarlar. Böylece değerlendirmeli metnin davranışını gerçek tabanlı analiz ve klasik metin madenciliğinden farklı yaparlar. Örneğin, kendi içinde imalı olmayan en iyi anahtar kelimelerin seti ile gündeme gelmek insanlar için önemli olabilir. Bu problem konu tabanlı sınıflandırmadan daha zordur. Durgunluk özelliği, kendi gözleminden öneri için herhangi bir insana uygun olamazken, verilen eğitim verileri ve onun korelasyonu ile pozitif sınıf, bir veri odaklı yaklaşım vasıtasıyla keşfedilebilir. Nitekim makine öğrenmesi tekniklerinin uygulandığı uni-gram modeller yüksek doğrulukla gerçekleştirebilir ki, bu, seçilmiş kelimeler üzerinde çalışmaktan daha iyidir.

4.6. Fikir Madenciliğinin Uygulama Alanları

Fikir Madenciliğine ilişkin uygulamalar şu şekilde özetlenebilir:

Üreticiler, şirketler, markalar ve işletmeler, insanların duygu, fikirlerini ve yönelimlerini bulmak için büyük bir miktarda para harcamaktadır (danışmanlar, anketler vb). Bu nedenle ürün/hizmet kıyaslama ve market zekâsı uygulamalarında Fikir Madenciliği yöntemleri kullanılabilir [66].

Şirketler ve devletler, bir ürün satın alma/hizmeti kullanma veya politik konular hakkında başka kişilerin fikirlerini öğrenmek istemektedirler. Bu amaçla Fikir Madenciliği yöntemleri kullanılabilir.

Kullanıcı tarafından oluşturulan içerikte reklam yatırımları amacıyla da (tek bir ürün övüldüğü zaman bir reklam yatırımı, eğer biri bir ürünü eleştirirse bir rakipten bir reklam yatırımı) Fikir Madenciliği yöntemleri kullanılabilir.

Pang ve Lee ise, fikir madenciliği ve duygu analizi uygulamalarını şu şekilde ifade etmektedir: Web sitelerindeki incelemelere ilişkin uygulamalar, bir alt bileşen teknolojisi olarak uygulamalar, iş ve devlet zekâsına ilişkin uygulamalar, farklı alanlardaki uygulamalar [66].

Fikir madenciliğinin Müşteri İlişkileri Yönetiminde kullanılması; ilk gruplandırmada “işletmeler ve organizasyonlar için” başlığı altında, ikincisinde ise “web sitelerindeki

incelemelere ilişkin uygulamalar” ve “iş ve devlet zekâsına ilişkin uygulamalar” başlıkları altında incelenebilir.

Son yıllarda kitlelerde web erişiminin artışıyla birlikte, web içeriğinin görünüşü değişmektedir. Web sayfaları gitgide, profesyonel olmayan kişiler tarafından yorumlarla doldurulmaktadır. Bunların bir kısmı “yayınlamada devrim” niteliği ile blokların fenomeni olmuştur.

Web genişleyen bir ortamdır ve bu ortamda müşteriler madencilik için olgunlaşmış olan fikirlerini belirtebilirler veya bu fikirleri bulmak için web’e girerler. Müşteri bakış açısının artan sayısı, bloglar üzerindeki gönderilerde kendini göstermektedir.



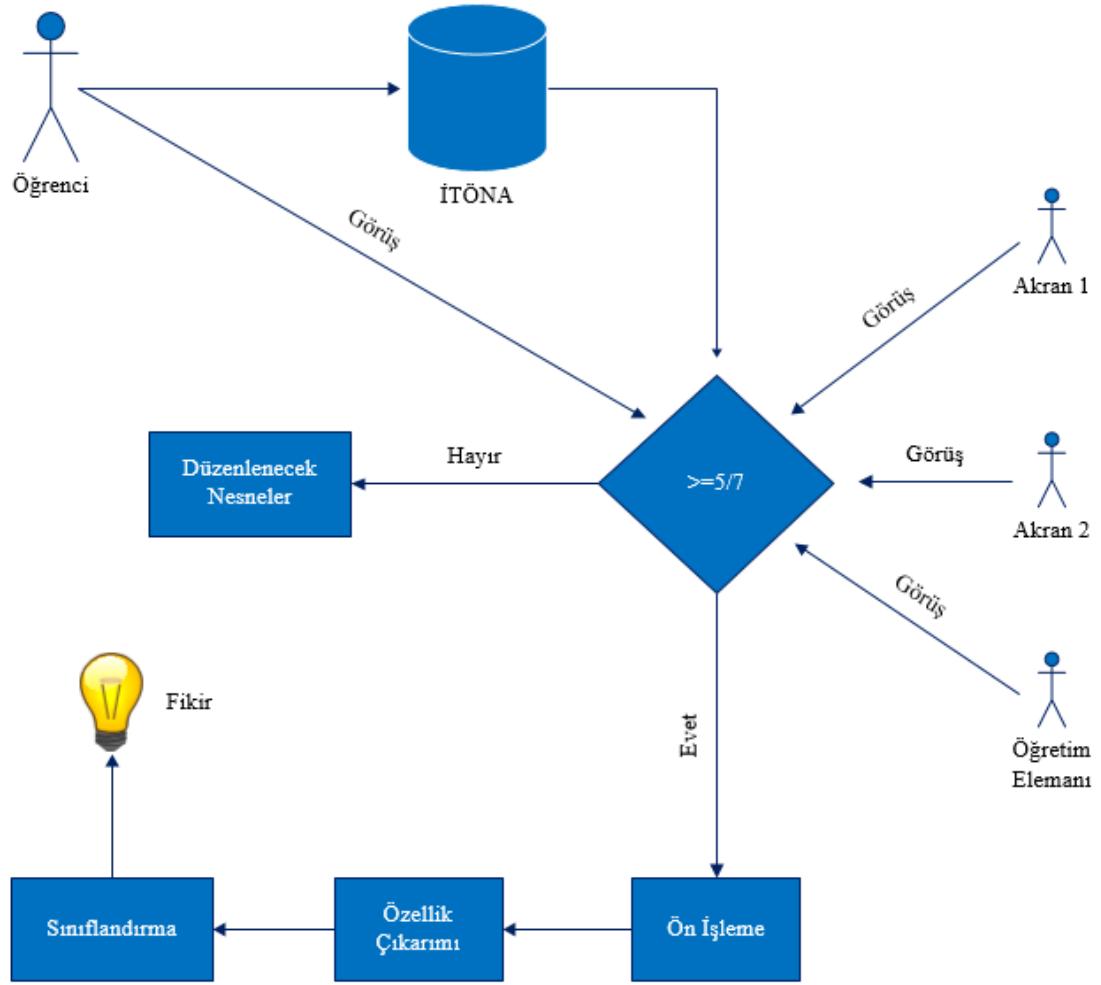
5. GELİŞTİRİLEN UYGULAMA

Bu tez çalışmasında, metin ve fikir madenciliği teknikleri kullanılarak eğitim alanında yeni bir karar destek uygulaması geliştirilmiştir. Geliştirilen uygulamanın eğitiminde, 29,272 adet anahtar kelime verisi kullanılmıştır. Bu kelime veri setinin 14,636 âdeti pozitif ve 14,636 âdeti negatif kelimelerden oluşmaktadır. Geliştirilen uygulama, Visual Studio ortamında gerçekleştirilmiştir. Uygulamayı gerçekleştirmek için C#, ASP.NET MVC ve MS-SQL dilleri kullanılmıştır. Ayrıca uygulama internet ortamından erişilebilir durumdadır [67]. Tez kapsamında geliştirilen uygulamada kullanılan veriler, İTÖNA sisteminden elde edilmektedir. Bu sisteme öğrenci, dersle ilgili görüşlerini yüklemektedir. Sisteme yüklenen görüşler, öğrencinin kendisi, 2 akranı ve bir öğretmen tarafından puanlanmaktadır. Eğer verilen puanlar belirlenen puanların ortalaması 7 üzerinden 5'i geçerse ilköğretimde kullanılabilir materyal olarak İTÖNA'ya aktarılmaktadır. Eğer belirtilen puanın altındaysa düzenlenecek nesneler arasına girmektedir. Geliştirilen sistem, var olan görüşün anlamlandırılması ve o görüşten bilgi çıkarılması safhasında kullanılacaktır. Sistem görüşleri İTÖNA veri ambarından çekmektedir. Elde edilen verilerle ilgili yapılan metin madenciliği adımları Şekil 5.1.' de gösterilmiştir.



Şekil 5.1. Uygulamada kullanılan metin madenciliği adımları

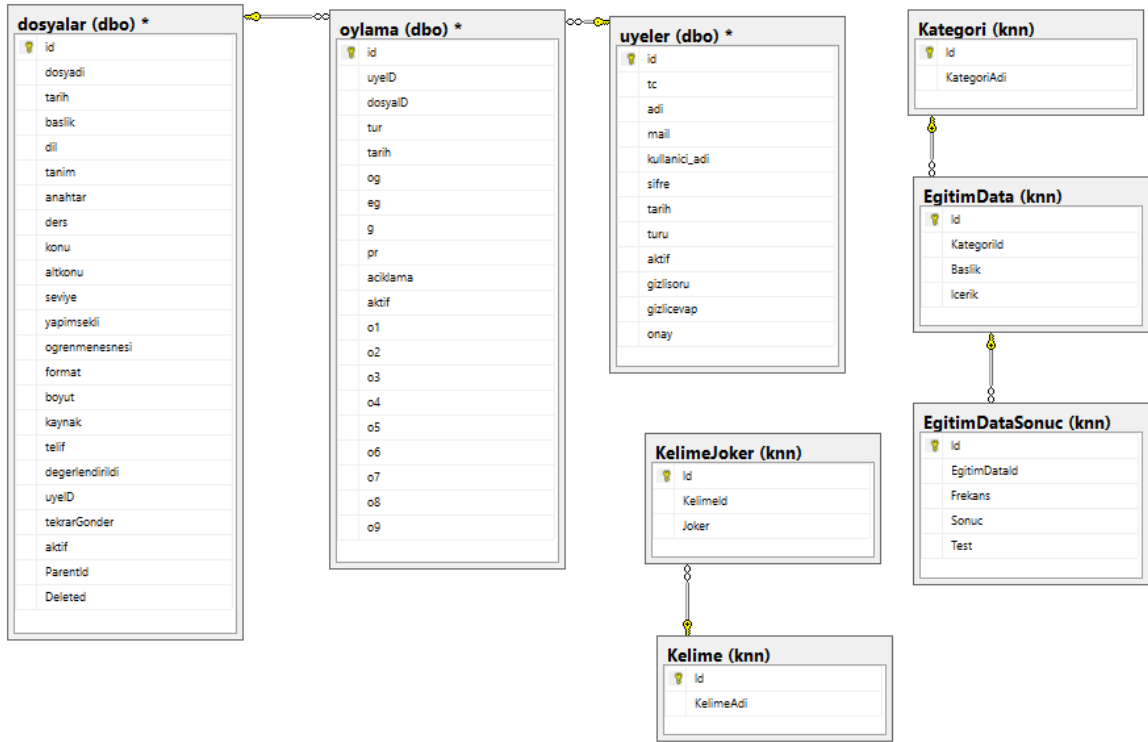
Metin madenciliği metotları ile tez kapsamında geliştirilen karar destek sistemi, öğrenci görüşlerini fikirler haline dönüştürmektedir. Böylece yapısal halde olmayan nitel öğrenci görüşleri anlamlandırılarak öğretmene dersin geribildirimi hakkında fikir vermektedir. Geliştirilen karar destek sisteminin blok diyagramı Şekil 5.2.' de verilmiştir.



Şekil 5.2. Geliştirilen uzman karar destek sisteminin blok diyagramı

5.1. Uygulamada Gerçekleştirilen Metin Önişleme İşlemleri

Geliştirilen karar destek sisteminin ilk aşamasında veri tabanı tasarımı yapılmıştır. Uzman karar destek sisteminin veri tabanı diyagramı Şekil 5.3.'te verilmiştir.



Şekil 5.3. Uzman karar destek sistemi için geliştirilen veri tabanı

Ayrıca İTÖNA gibi dış kaynaklardan alınan öğrenci görüşleri veri tabanına kaydedildikten sonra metin temizleme işlemi gerçekleştirilmektedir. Metin temizlemede noktalama ve alfa numerik karakterler yok edilir. Metin temizleme işleminin ardından küçük harf (lowercase) dönüşüm işlemleri gerçekleştirilmektedir. Bu aşamanın ardından artık metinlerden özellik çıkarılabilmektedir. Metinden fikir elde etmenin ilk aşaması önışlem aşamasıdır.

5.2. İşbirlikli Tartışmacı Öğrenme Nesne Ambarı (İTÖNA)

İTÖNA Fırat Üniversitesi tarafından oluşturulan ilköğretim seviyesinde çok sayıda dersle ilgili yaklaşık 653 öğrenme nesnesinin bulunduğu Türkçe bir nesne ambarıdır. Bu nesne ambarı 2011 yılında TÜBİTAK tarafından desteklenen bir proje kapsamında yapılmıştır. Benzetimler, konu anlatımlı sunular, resimler, videolar, alıştırma, flash animasyonlar, deneyler vb. birçok formatta nesnenin kaydedilebildiği bir nesne ambarıdır.

Bu nesne ambarı ana sayfa, nesne, dosyalar, moodle ve diğer ÖNA (Öğrenme Nesne Ambarı) kısımlarından oluşmaktadır.

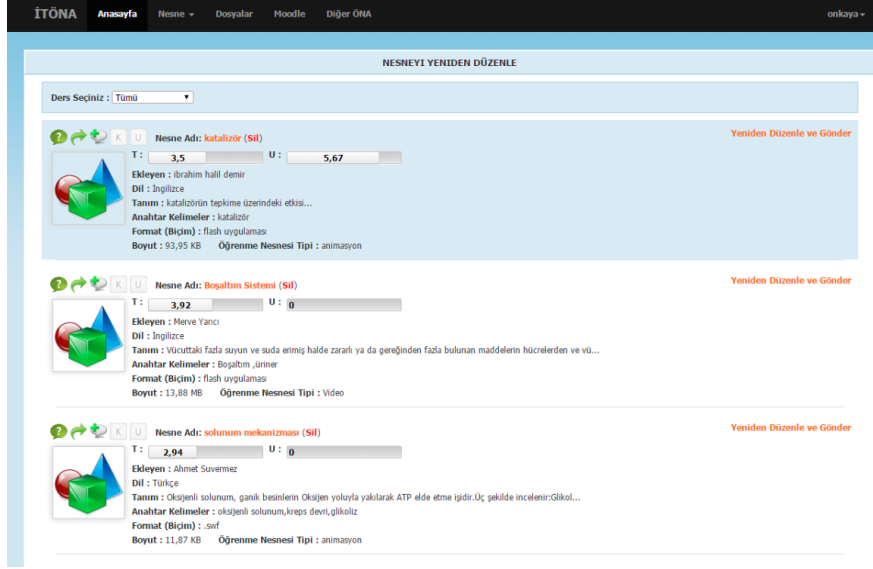
Bu nesne ambarı interaktif bir nesne ambarı olup, öğretmen değerlendirmesi, öz değerlendirme ve akran değerlendirmesine göre puanlandırılmaktadır. Puanlandırma 9 ayrı değerlendirme ölçütü kullanılarak 1 ile 7 arasında değerlendirilmektedir. Nesne değerlendirme sayfası Şekil 5.4.'te gösterildiği gibidir. Nesne ile ilgili oluşturulan nitel veriler, tez kapsamında geliştirilen uygulamanın kaynağıdır.

Öğrenme Nesnesi Değerlendirme Ölçeği		
1. İçeriğin Niteliği	: 1 ▼	Kriterler
2. Öğrenme Kazanımlarına Uygunluğu	: 1 ▼	Kriterler
3. Dönüt ve Uyum Gösterme	: 1 ▼	Kriterler
4. Motivasyon	: 1 ▼	Kriterler
5. Sunum Tasarımı (Dizayn)	: 1 ▼	Kriterler
6. Etkileşimsel Kullanılabilirlik	: 1 ▼	Kriterler
7. Ulaşılabilirlik / Kullanım Kolaylığı	: 1 ▼	Kriterler
8. Yeniden Kullanılabilme	: 1 ▼	Kriterler
9. Standartlara Uygunluk	: 1 ▼	Kriterler

Kaydet

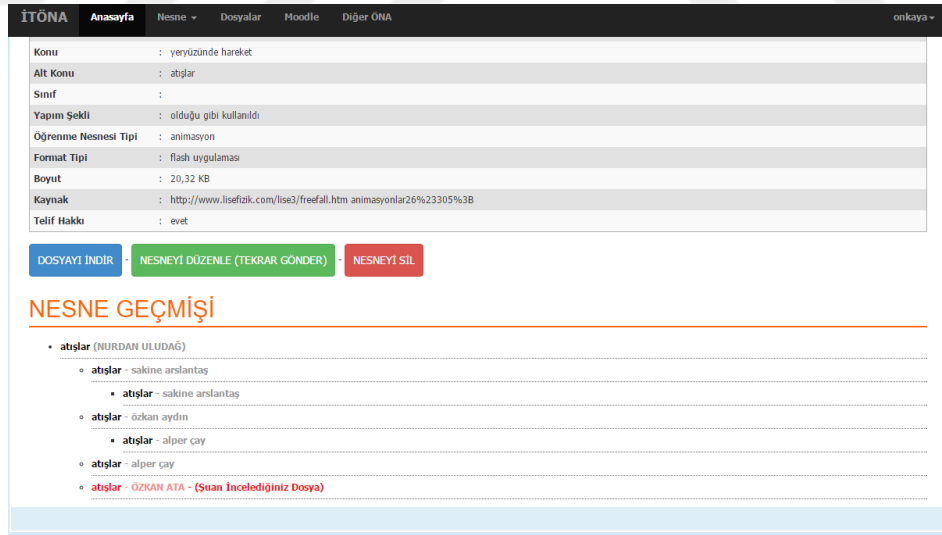
Şekil 5.4. İTÖNA nesne değerlendirme sayfası

Eğer nesnenin ortalama puanı 5 ve 5'ten büyükse İTÖNA'ya eklenmektedir. Eğer nesnenin ortalama puanı 5'ten küçükse düzenlenecek nesneler kısmına aktarılmaktadır. Düzenlenecek nesneler sayfası Şekil 5.5.'te gösterilmektedir.



Şekil 5. 5. Yeniden düzenlenecek nesneler sayfası

Düzenlenecek nesneler tablosundaki nesneleri diğer kullanıcılar düzenleyebilmektedir. Ayrıca nesnelerin hangi kullanıcılar tarafından düzenlendiği nesne geçmişinden görülebilmektedir. Nesne geçmişi sayfası Şekil 5.6.'de gösterilmiştir.



Şekil 5.6. Nesne geçmişi sayfası

5.2. Uygulamada Gerçekleştirilen Metin Ağırlıklandırma İşlemleri

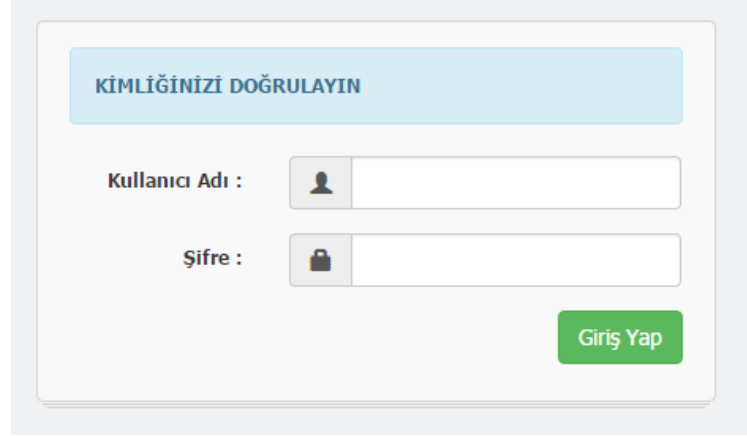
Metinleri vektörel olarak temsil edebilmek için bu aşamadaki işlemlerin gerçekleştirilmesi gerekmektedir. Metinleri vektörel olarak temsil edebilmek için önceden oluşturulmuş veri tabanının kullanılması gerekmektedir. Bu veri tabanı kullanılarak, cümlede bulunan kelimelerin frekans analizi yapılır. Frekans analizi yapıldıktan sonra cümlelerin vektörü elde edilir. Bu aşama özellik çıkarma aşamasına benzemektedir. İşlemler esnasında yapılacak en küçük hata sistemin yanlış çalışmasına sebep olacaktır. Bu aşamadan elde edilen özellikler sınıflandırıcıya gönderilir.

5.3. Uygulamada Gerçekleştirilen Sınıflandırma İşlemleri

Geliştirilen uzman karar destek sisteminin sınıflandırma aşamasında k -NN ve Naive Bayes akıllı sınıflandırıcıları kullanılmıştır. Bu sınıflandırıcıların eğitim aşamasında 29,272 adet veri kullanılmıştır. Eğitim veri seti, 14,636 adet pozitif ve 14,636 adet negatif kelimedenden oluşmaktadır. Yapılan eğitim sonucunda negatif ve pozitif kümeler oluşturulmuştur. k -NN ve Naive Bayes algoritması yeni veriden gelen veriden elde edilen değer hangi kümeye yakınsa, veri o kümeye aittir kararı verilmektedir. Ayrıca uygulamada pozitiflik ve negatiflik skoru yüzdelik olarak da kullanıcıya verilmektedir.

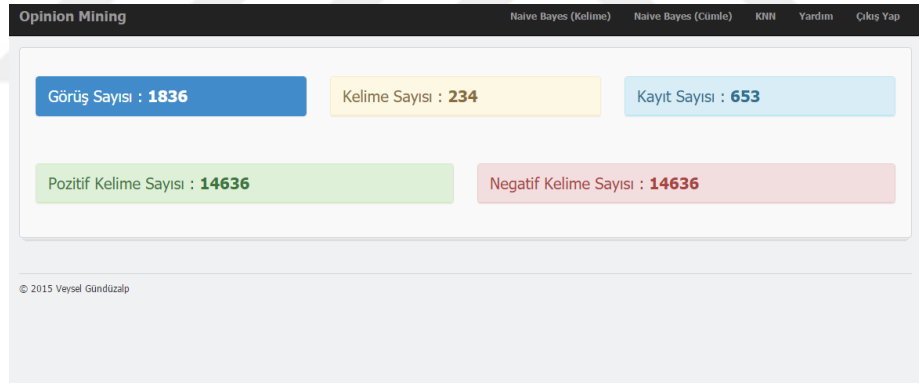
5.4. Uygulamanın Çalışması

Uygulamaya web adresinden ulaşılabilir [67]. Geliştirilen uzman karar destek sistemi kullanıcı girişi, Naive Bayes (Kelime), Naive Bayes (Cümle), k -NN ve yardım bölümlerinden oluşmaktadır. Geliştirilen karar destek sistemi, verileri İTÖNA'dan otomatik olarak çekmektedir. Geliştirilen karar destek sisteminin temel amacı, interaktif eğitim ortamlarında öğrencilerin metin olarak paylaştığı görüşlerden öğrencilerin ders, ders materyali veya öğretmenle ilgili fikirlerini otomatik olarak elde edebilmektir. Bu işlemi yapabilmek için yukarıda verilen akıllı sınıflandırıcılar kullanılmaktadır. Geliştirilen uzman karar destek sistemi metinleri pozitif, nötr ve negatif fikir olarak kategorize etmektedir. Bu tez kapsamında geliştirilen uzman karar destek sisteminin giriş ekranı Şekil 5.7.'de verilmiştir.



Şekil 5.7. Geliştirilen uzman karar destek sisteminin giriş ekranı

Uzman karar destek sistemine yönetici yetkisiyle giriş yapıldıktan sonra sistemde var olan görüş sayısı, kelime sayısı, kayıt sayısı, eğitim için var olan pozitif kelime sayısı ve eğitim için var olan negatif kelime sayısı ekrana gelmektedir. Yönetici anasayfa ekranı Şekil 5.8.' da gösterilmiştir.



Şekil 5.8. Anasayfa ekranı ve veri sayıları

İTÖNA'dan alınan görüşleri sınıflandırmak ve bu görüşler hakkında sistemin ürettiği fikirleri görebilmek için Naive Bayes (Kelime), Naive Bayes (Cümle) ve k -NN menülerinden herhangi birinin tıklanması gerekmektedir. Geliştirilen uzman karar destek sistemi, kullanıcı tarafından seçilen sınıflandırıcıyı kullanarak fikir üretmektedir. Bu menüler tıklandığında nesne değerlendirme ekranı karşımıza gelmektedir. Şekil 5.9'da nesne değerlendirme sayfası gösterilmektedir.

Opinion Mining

Naive Bayes (Kelime)

Naive Bayes (Cümle)

KNN

Yardım

Çıkış Yap

NESNE DEĞERLENDİRME								
#	Dosya Adı	Ekleyen	Özdeğerlendirme	Akran1	Akran2	Öğretim Elemanı	ORT	İşlemler
55	ince kenarlı mercekte görüntü oluşumu	Halil İbrahim Ulutaş	GÖRÜNTÜLE	GÖRÜNTÜLE	GÖRÜNTÜLE	Görüş Yok!	3,33	
56	Düşey Hareket	Halil İbrahim Ulutaş	GÖRÜNTÜLE	GÖRÜNTÜLE	GÖRÜNTÜLE	Görüş Yok!	3,67	
57	D1 Periyodik tablo	Muhammed Fatih Karabay	GÖRÜNTÜLE	GÖRÜNTÜLE	GÖRÜNTÜLE	Görüş Yok!	4	
58	KIRILMA VE YANSIMA	Halil İbrahim Ulutaş	GÖRÜNTÜLE	GÖRÜNTÜLE	GÖRÜNTÜLE	Görüş Yok!	4,67	
59	ORGANLAR	Halil İbrahim Ulutaş	GÖRÜNTÜLE	GÖRÜNTÜLE	GÖRÜNTÜLE	Görüş Yok!	3,33	
60	Virüs	Halil İbrahim Ulutaş	GÖRÜNTÜLE	GÖRÜNTÜLE	GÖRÜNTÜLE	Görüş Yok!	3,33	
61	Top yangı	Halil İbrahim Ulutaş	GÖRÜNTÜLE	GÖRÜNTÜLE	GÖRÜNTÜLE	Görüş Yok!	2,67	
62	Kanın süzülmesi	Esra ZORBOZAN	GÖRÜNTÜLE	GÖRÜNTÜLE	GÖRÜNTÜLE	GÖRÜNTÜLE	3,25	
63	Farklı gezegenlerde Kütle ve ağırlık	Muhammed Fatih Karabay	GÖRÜNTÜLE	GÖRÜNTÜLE	GÖRÜNTÜLE	Görüş Yok!	3,33	

Şekil 5.9. Nesne değerlendirme sayfası

Şekil 5.9’da gösterilen nesne değerlendirme sayfasında dosya numarası, dosya adı, ekleyen, öz değerlendirme, birinci akranın değerlendirmesi, ikinci akranın değerlendirmesi, öğretmen görüşü, verilen puanların ortalaması ve işlemler alanları yer almaktadır. İşlemler bölümünde ise indirme, nesne geçmişi ve nesne silme gibi seçenekler bulunmaktadır. Görüşleri görüntülediğimizde, nesneye verilen puanlar ve o nesne hakkında uzman karar destek sisteminin ürettiği fikir görüntülenmektedir. Görüşler ekranı ve üretilen fikir skorları ve fikirler Şekil 5.10.’de verilmiştir.

Görüş Tarihi	:	29.12.2011
İçeriğin Niteliği	:	3
Öğrenme Kazanımlarına Uygunluğu	:	5
Dönüt ve Uyum Gösterme	:	3
Motivasyon	:	6
Sunum Tasarımı (Dizayn)	:	6
Etkileşimsel Kullanılabilirlik	:	2
Ulaşılabilirlik / Kullanım Kolaylığı	:	2
Yeniden Kullanılabilirlik	:	2
Standartlara Uygunluk	:	6
Görüş cinsine göre değişken kısmında "yukarıdan bir iletkeni tıklayınız" kısmının yukarıda istediğiniz bir maddeye tıklayınız daha iyi olacaktır çünkü yukarıda ki tahta iletken değildir. renkler biraz daha canlı yapılabilir. lambanın yanıp sönmeye yerine parlak yada daha fazla/az parlak lamba olması daha uygun olacaktır. bunlar düzeltilirdi takdirde animasyon gayet güzel.		
NEGATİF : %50.06 POZİTİF : %49.94 SONUÇ: NOTR		

Şekil 5.10. Değerlendirme ekranı ve sistemin ürettiği fikir çıktıları

Şekil 5.10.' de gösterildiği gibi sistem, bireyin metin olarak ifade ettiği görüşlerden fikir üretmektedir. Geliştirilen uzman karar destek sistemi, metin madenciliği ve fikir madenciliği kullanarak kullanıcının fikrini otomatik olarak üretebilmektedir.

5.5. DeneySEL Sonuçlar

Tezin bu bölümünde, geliştirilen uzman karar destek sisteminden elde edilen sonuçlar gösterilmiştir. Geliştirilen uzman karar destek sistemi polarite analizi yaparak, bireylerin sisteme metin olarak girdiği görüşleri pozitif, nötr ve negatif görüşlere çevirmektedir. Geliştirilen uzman karar destek sistemi, İTÖNA'dan alınan 653 adet nesne ve 1836 adet görüşü değerlendirmiştir. Sınıflandırma aşamasında ise Naive Bayes ve k -NN sınıflandırıcıları kullanılmıştır. Uzman karar destek sisteminin eğitim aşamasında ise 14,636 adet pozitif kelime ve 14,636 adet negatif kelime kullanılmıştır. Oluşturulan uzman karar destek sisteminin başarımını hesaplamak için doğruluk parametresi olan ACC hesaplanmıştır. ACC'nin formülü Eşitlik 5.1.'de verilmiştir.

$$ACC = \frac{\text{Doğru tahmin edilen görüş sayısı}}{\text{Toplam görüş sayısı}} \quad (5.1)$$

Geliştirilen uzman karar destek sisteminde Naive Bayes ve k -NN sınıflandırıcıları kullanılmıştır. Ancak Naive Bayes algoritmasına iki farklı şekilde özellikler aktarılmıştır. İlk özellik çıkarma yönteminde pozitif ve negatif kelimelerin frekansları hesaplanarak metin ağırlıklandırma yapılmıştır. Elde edilen ağırlıklar Naive Bayes sınıflandırıcısı kullanılarak sınıflandırılmıştır. İkinci özellik çıkarma yönteminde ise, eğitim veri setinde yer alan negatif ve pozitif cümleler yer almaktadır. Cümle frekansları hesaplanarak özellik vektörü oluşturulmuştur ve ardından elde edilen özellik vektörü Naive Bayes sınıflandırıcısı kullanılarak sınıflandırılmıştır. Elde edilen doğruluk oranları Tablo 5.1.'de verilmiştir.

Tablo 5.1. Geliştirilen uzman karar destek sisteminde kullanılan sınıflandırıcıların başarımı.

Sınıflandırıcı	Doğruluk(ACC)
NB(Kelime)	58.032
NB(Cümle)	60.387
<i>k</i> -NN	68.425

Geliştirilen uzman karar destek sistemin başarımını diğer uygulamalarla karşılaştırmak için literatür taraması yapılmıştır ve Lima vd.'nin geliştirdiği yazılım kütüphanesinin birden fazla sınıflandırıcı kullanması ve polarite sonuçları üretmesi nedeniyle geliştirilen uzman karar destek sistemiyle karşılaştırılabileceği ortaya konulmuştur [8]. Geliştirilen uzman karar destek sistemi, Lima vd.'nin geliştirdiği polarite yazılım kütüphanesiyle karşılaştırılmıştır ve elde edilen sonuçlar Tablo 5.2.'de gösterilmiştir [8].

Tablo 5.2. Geliştirilen uzman karar destek uygulamasının Lima vd.'nin polarite yazılım kütüphanesiyle karşılaştırılması.

Doğruluk (ACC) %		
Sınıflandırıcı	Lima vd. uygulaması [8]	Geliştirilen uygulama
NB	63.011	58.032
KNN	61.535	68.425

6. SONUÇLAR VE TARTIŞMA

İnternetin kullanımının yaygınlaşması ve internet ortamındaki verilerin boyutunun büyümesi ile birlikte Büyük Veri (Big Data) kavramı ortaya çıkmıştır. Bu kavram içerisinde bulunan verilerin tek başlarına bir anlam ifade edememesi sebebi ile veri madenciliği yöntemleri kullanımı önerisi ortaya çıkmıştır. Ancak veri madenciliği tekniklerinin yapısal olan verilere uygulanması tüm verilerin işlenememesine ve bu verilerden anlamlı bilgi çıkartılamamasına sebep olmaktadır. Yapısal olmayan ve metin türünde olan verilerin işlenmesi ve bu verilerden anlamlı bilgilerin elde edilebilmesi için metin madenciliği tekniklerinin kullanılmasına ihtiyaç duyulmuştur.

Bu tez çalışmasında veri madenciliği teknikleri, veri tabanından bilgi keşfi süreci, verilerin ön işlenmesi, verilerden özellik çıkartımı ve verilerin sınıflandırılması veri madenciliği bölümünde incelenmiştir. Bu başlıklarla ilgili literatür taraması yapılmıştır. Diğer bir bölümde ise metin madenciliği teknikleri incelenmiştir. Bu başlık altında metin veri tabanı oluşturma, metin ön işleme, benzerlik hesaplaması, metin sınıflandırma, doğal dil işleme yöntemleri ve metin madenciliğinin uygulama alanları incelenmiş ve bu başlıklar ile ilgili literatür taraması yapılmıştır. Bu iki ana başlığa bağlı olarak fikir madenciliği başlığı üçüncü ana başlık olarak seçilmiştir. Fikir madenciliği bölümünde ise fikir kavramı ve bileşenleri, fikir madenciliğinin görevleri, fikir sınıflandırma, fikir madenciliğinde karşılaşılan problemler ve fikir madenciliğinin uygulama alanları ile ilgili literatür taraması yapılmıştır.

Bu tez çalışması, yapılan literatür taraması sonucunda gerçekleşmiştir. Elde edilen bilgiler ışığında metin madenciliği ve metin madenciliğinde kullanılan teknikler ele alınmıştır. Metin madenciliğinde karşılaşılan en büyük problemler sözcüğün anlamının vektörel olarak ifade edilememesi, yazım hataları, kullanılan özel karakterler ile semboller ve doğal dilin yapısıdır. Metin madenciliğinde kullanılan teknikler bu problemlerin giderilmesi hususunda yetersiz kalmaktadırlar. Ayrıca metin madenciliğinin uygulanmadığı birçok alanın var olduğu tez çalışmasında tespit edilmiştir. Tespit edilen uygulama eksikliğini giderebilmek için eğitim alanında fikir madenciliği yapan bir karar destek sistemi geliştirilmiştir. Geliştirilen karar destek sistemi, veri tabanı olarak İTÖNA'yı kullanmaktadır. İTÖNA veri tabanından elde edilen metinlere Türkçe dilbilgisi yapısına uygun olarak, önişlem, özellik çıkarma ve sınıflandırma işlemleri yapılmıştır. Tez

kapsamında gerçekleştirilen uygulamada Naive Bayes ve k -NN sınıflandırma algoritmalarını kullanarak, belirtilen görüşün pozitif, negatif ve nötr olmasıyla ilgili otomatik karar veren bir fikir madenciliği uzman karar destek sistemi geliştirilmiştir. Geliştirilen uzman karar destek sistemi sayesinde, interaktif eğitim uygulamalarında bulunan kullanıcı görüşlerini otomatik olarak değerlendirip, hızlı bir şekilde geri bildirim yapılması ve konuyla ilgili görüşlere ait fikirlerin otomatik olarak elde edilmesi sağlanmıştır. Böylece İTÖNA'da var olan nesneler hakkında belirtilen görüşlerden otomatik fikir elde edilerek kişinin verdiği puanla görüşünün uyumlu olup olmadığı ortaya çıkarılmıştır. Geliştirilen uzman karar destek sistemi Lima vd.'nin [8] önerdiği polarite analiz kütüphanesiyle karşılaştırılmıştır. Karşılaştırma için kullanılan polarite analiz kütüphanesi İngilizce dili için tasarlanmıştır. Yapılan karşılaştırmalar sonucunda, geliştirilen uzman karar destek sisteminin k -NN sınıflandırıcısında daha iyi sonuç verdiği gözlemlenmiştir. Ancak Naive Bayes sınıflandırıcısında diğer kütüphanenin daha iyi sonuç verdiği görülmüştür. Bu farklılığın ise kullanılan doğal dilin dilbilgisinden ve seçilen anahtar kelimelerden kaynaklandığı görülmektedir. Eğer önışlem ve özellik çıkarma adımlarında doğal dilin yapısına ve kullanıcı genel davranışlarına göre revize edilip yeni akıllı modeller geliştirilirse, metin madenciliği ve fikir madenciliği uzman karar destek sistemlerinin başarımı arttırılabilecektir. Tez kapsamında gerçekleştirilen uygulama ile literatür de var olan örneklerden k -NN algoritması ile daha iyi sonuç elde edildiği gözlemlenmiştir.

Gelecekteki çalışmalarda Türkçe dili için fikir madenciliği ve polarizasyon tekniklerinden oluşan bir kütüphanenin tasarlanması planlanmaktadır. Ayrıca yeni önışlem ve özellik çıkarma yöntemleri geliştirilerek, uzman karar destek sistemlerinin başarımı yükseltilebilir. Geliştirilecek yüksek başarıma sahip uzman karar destek sistemlerinin farklı alanlarda kullanılması da önerilmiştir.

KAYNAKLAR

- [1] **Dolgun, M.Ö., Özdemir, T.G., Oğuz, D.,** 2009. Veri madenciliği'nde yapısal olmayan verinin analizi: Metin ve web madenciliği, *İstatistikçiler Dergisi* 2, pp. 48-58.
- [2] **Gezer, M., Erol, Ç. ve Gülseçen, S.,** 2007. Bir Web Sayfasının Web Madenciliği ile Analizi, *AB-2007 Akademik Bilişim Konferansı*, Kütahya.
- [3] **Dean, N.,** 2000. OCLC Researchers measure the World Wide Web. Online Computer Library Center (OCLC) Newsletter.
- [4] **Daş, R., Türkoğlu, İ.,** 2010. Web tabanlı öğretim materyallerinin web kullanım madenciliği ile analiz edilmesi, *Fırat Univ. Journal of Engineering*, pp. 22 (1), 111-122.
- [5] **Rezgui, K., Mhiri, H. ve Ghédira, K.,** 2014. Extending Moodle Functionalities with Ontology-based Competency Management *Procedia Computer Science*, Volume 35, p. 570-579..
- [6] **Werner, A. ve Myrray, Z. F.,** 2001. Is all that talk just noise? The information content of internet stock message boards, *Journal of Finance*, 10, p. 1259-1294.
- [7] **Ikeda, K.; Hattori, G.; Ono, C.; Asoh, H. & Higashino, T.,** 2013. Twitter user profiling based on text and community mining for market analysis, *Knowledge-Based Systems*, 51, pp. 35-47.
- [8] **Lima, A. C. E. S.; de Castro, L. N.; Corchado, J. M.,** 2015. A polarity analysis framework for Twitter messages, *Mathematics and Computation*, 270, pp. 756-767.
- [9] **Uysal, A. K.; Günal, S.,** 2014. The impact of preprocessing on text classification, *Information Processing and Management*, 50, pp. 104-112.
- [10] **Li, W., XU, H.,** 2014. Text-based emotion classification using emotion cause Extraction, *Expert Systems with Applications*, 41, pp. 1742-1749.
- [11] **Fleuren, W.W.M., Alkema, W.,** 2015. Application of text mining in the biomedical domain, *Methods*, 74, pp. 97-106.
- [12] **Noh, H.; Jo, Y.; Lee, S.,** 2015. Keyword selection and processing strategy for applying text mining to patent analysis, *Expert Systems with Applications*, 42, pp. 4348-4360.
- [13] **Akçapınar, G.,** 2015. How automated feedback through text mining changes plagiaristic behavior in online assignments, *Computers & Education*, 87, pp. 123-130.

- [14] **Romero, C., Ventura, S., Garcia, E.,** 2008. Data mining in course management systems: Moodle case study and tutorial, *Computers & Education*, 51, pp. 368-384.
- [15] **Lin, F.-R., Hsieh, L.-S., Chuang, F.-T.,** 2009. «Discovering genres of online discussion threads via text mining, *Computers&Education*, 52, pp. 481-495.
- [16] **Özekeş, S.,** 2011. Veri Madenciliği Modelleri ve Uygulama Alanları, *İstanbul Ticaret Üniversitesi Dergisi*, 2011, pp. 65-82.
- [17] **Delen, D. ve Crossland, M.D.,** 2008. Seeding the survey and analysis of research literature with text mining, *Expert Systems with Applications*, 34, 2008, p. 1707–1720.
- [18] **Mecca, G., Raunich, S. ve Pappalardo, A.,** 2006. A new algorithm for clustering search results, *Data & Knowledge Engineering* 62.
- [19] **Oğuz, B.,** 2009. Metin Madenciliği Teknikleri Kullanılarak Kulak Burun Boğaz Hasta Bilgi Formlarının Analizi, *Akdeniz Üniversitesi Sağlık Bilimleri Enstitüsü, Yüksek Lisans Tezi*.
- [20] **Şahin, M.,** 2005. Yönetim Bilgi Sistemi, *Anadolu Üniversitesi Yayınları*, Eskişehir, p. 118-119.
- [21] **Fayyad, U., Piatetsky-Shapiro, G. ve Symth, P.,** 1996. From Data Mining To Knowledge Discovery in Databases, *AI Magazine*, Vol: 17, No: 3, p. 37-54.
- [22] **Cabena, P., Hadjinian, P., Stadler, R., Verhees, J., Zanasi, A.,** 1997. Discovering Data Mining: From Concept To Implementation, *New Jersey, Prentice Hall PTR, Upper Saddle River*, p. 195.
- [23] **İlhan, U.,** 2001. Application Of KNN and FPTC Based Text Categorization Algorithms to Turkish News Reports, *Bilkent University*.
- [24] **Amasyalı, F. ve Yıldırım, T.,** 2004. Otomatik Haber Metinleri Sınıflandırma,» %1 içinde *SUI*.
- [25] **İlhan, S., Duru, N., Karagöz, Ş., M., Sağır.,** 2008. Metin Madenciliği ile Soru Cevaplama Sistemi, *ELECO*, p. 68.
- [26] **Nanopoulos, A., Theodoridis, Y. ve Manolopoulos, Y.,** 2001. C2P: Clustering based on closest pairs, *Proceedings of the 27th International Conference on VeryLarge Data Bases*, Roma, Italy, pp. 331-340.

- [27] **Mitchell, T.M.**, 1997. Machine Learning 1st ed., New York., McGraw Hill.
- [28] **Soucy, P. ve Mineau, G. W.**, 2001. A Simple KNN Algorithm for Text Categorization, *Proceedings IEEE International Conference on Data Mining (ICDM '01)*, California, pp. 647-648.
- [29] **Lagus, K.**, 2000. Text Mining With The WEBSOM, *Helsinki University of Technology*, Finland.
- [30] **Toraman, Ç., Can, F. ve Koçberber, S.**, 2011. Developing a Text Categorization Template for Turkish News Portals, *International Symposium on INnovations in Intelligent Systems and Applications (INISTA 2011)*, İstanbul, pp. 379-383.
- [31] **Kutlu, F.**, 2001. Categorization In A Hierarchically Structured Text Database, Bilkent.
- [32] **Soucy, P. ve Mineau, W.**, 2005. Beyond TFIDF Weighting for Text Categorization in the Vector Space Model, *IEEE International Conference*, pp. 1130-1135.
- [33] **Silahtaroglu, G.**, 2008. Kavram ve Algoritmalarıyla Temel Veri Madenciliği, İstanbul, Papatya Yayıncılık.
- [34] **Argüden, Y. ve Erşahin, B.**, 2008. Veri Madenciliği: Veriden Bilgiye, Masraftan Değere, ARGE Danışmanlık Yayınları.
- [35] **Eyheramendy, S., Lewis, D. ve Madigan, D.**, 2003. On The Naive Bayes Model For Text Categorization, *In Proceedings of Artificial Intelligence and Statistics* , pp. 3-6.
- [36] **Karaca, M. F.**, 2012. Metin madenciliği yöntemleri ile haber sitelerindeki köşeyazılarının sınıflandırılması, *Karabük Üniversitesi FBE, Yüksek Lisans Tezi*, 2012.
- [37] **Delibaş, A.**, 2008. Doğal Dil İşleme ile Türkçe Yazım Hatalarının Denetlenmesi, *İstanbul Teknik Üniversitesi FBE*.
- [38] **Özbilici, A.**, 2006. Türkçe Doğal Dili Anlamada İlişkisel Ayırık Bilgiler Modeli ve Uygulaması, *Sakarya Üniversitesi FBE, Yüksek Lisans Tezi*.
- [39] **Kesgin, F.**, 2007. Türkçe Metinler için Konu Belirleme Sistemi, *İstanbul Teknik Üniversitesi Fen Bilimleri Enstitüsü Yüksek Lisans Tezi*.
- [40] **Bethard, S., Yu, H., Thornton, A., Hatzivassiloglou, V., Jurafsky, D.**, 2006. Extracting Opinion Propositions and Opinion Holders using Syntactic and Lexical Cues, *Theory And Applications, Chapter 11*, p. 125-142.

- [41] **Chen, B., Zhu, L., Kifer, D., Lee, D.,** 2010. What Is an Opinion About? Exploring Political Standpoints Using Opinion Scoring Model, *Proceedings Of The Twenty-Fourth AAAI Conference On Artificial Intelligence(AAAI 2010)*, Atlanta, GA, pp. 1007-1012.
- [42] **Conrad, J. ve Schilder, F.,** 2007. Opinion Mining in Legal Blogs, *Proceedings Of The International Conference On Artificial Intelligence And Law (ICAIL)*, New York, NY, USA, pp. 231-236.
- [43] **Esuli, A. ve Sebastiani, F.,** 2006. Determining Term Subjectivity And Term Orientation For Opinion Mining, *Proceedings Of The European Chapter Of Theassociation For Computational Linguistics (EACL)*, Sydney, Australia, p. 193-200.
- [44] **Esuli, A., Sebastiani, F.,** 2006. Sentiwordnet: A Publicly Available Lexical Resource For Opinion Mining, *Proceedings Of Language Resources And Evaluation (LREC)*, Genoa, Italy, pp. 417-422.
- [45] **Esuli, A.; Sebastiani, F.,** 2007. Pageranking Wordnet Synsets: An Application To Opinion Mining, *Proceedings Of The Association For Computational Linguistics (ACL)*, Prague, CzechRepublic, 2007, pp. 424-431.
- [46] **Ghose, A., Ipeirotis, P. ve Sundararajan, A,** 2007. Opinion Mining Using Econometrics: A Case Study On Reputation Systems, *Proceedings Of The Association For Computational Linguistics (ACL)*, Prague, CzechRepublic, pp. 416-423.
- [47] **Grefenstette, G.; Qu, Y.; Shanahan, J.; Evans, D.,** 2004. Coupling Niche Browsers And Affect Analysis For An Opinion Mining Application, *Proceedings Of Recherched 'Information Assist' Ee Par Ordinateur (RIAO)*, France, pp. 186-194.
- [48] **Hu, M. ve Liu, B.,** 2004. Mining Opinion Features in Customer Reviews, *Proceedings Of AAAI, San Jose*, p. 755-760.
- [49] **Piao, S.; Ananiadou, S.; Tsuruoka, Y.; Sasaki, Y.; Mcnaught, J.,** 2007. Mining Opinion Polarity Relations Of Citations, *International Workshop On Computational Semantics (IWCS)*, Tilburg, Netherlands, p. 366-371.
- [50] **Attardi, G. ve Simi, M.,** 2006. Blog Mining Through Opinionated Words, *Text Retrieval Conference*, Gaithersburg, Maryland.

- [51] **Kobayashi, N., Inui, K. ve Matsumoto, Y.,** 2007. Extracting Aspect-Evaluation And Aspect-Of Relations in Opinion Mining, *Joint Conference On Empirical Methods in Natural Language Processing And Computational Natural Language Learning*, Prague, p. 1065-1074.
- [52] **Kaji, N. ve Kitsuregawa, M.,** 2007. Building Lexicon For Sentiment Analysis From Massive Collection Of HTML Documents, *Joint Conference On Empirical Methods in Natural Language Processing And Computational Natural Language Learning*, Prague, Czech Republic, p. 1075-1083.
- [53] **Aytekin, Ç.,** 2011. Müşteri İlişkileri Yönetimi İçin Bloglar Üzerinde Fikir Madenciliği, Doktora Tezi, *Marmara Üniversitesi Sosyal Bilimler Enstitüsü İletişim Bilimleri Anabilim Dalı Bilişim Programı*.
- [54] **Levene, M.,** 2010. An Introduction To Search Engines And Web Navigation, Wiley Publisher, Second Edition, New Jersey, p. 391.
- [55] **Choudhury, S. ve Breslin, J.,** 2010. Focused Concept Detection And Polarity Measurement Of Web Video Using Social Context, *Proceedings Of The WebSci10: Extending The Frontiers Of Society On-Line, NC: US, Raleigh, ID Code: 378*.
- [56] **Pang, B. ve Lee, L.,** 2004. A Sentiment Education: Sentiment Analysis Using Subjectivity Summarization Based On Minimum Cuts, *In Proceedings Of The ACL*, Barcelona, p. 271-278.
- [57] **Liu, B.,** 2007. From Web Content Mining To Natural Language Processing, *ACL-2007 Tutorial*, Prague, p. 114.
- [58] **Wiebe, J., Mihalcea, R.,** 2006. Word Sense And Subjectivity, *Conference Of The International Committee On Computational Linguistics And The Association For Computational Linguistics*, Sydney, Australia, p. 1134.
- [59] **Qiu, G., Liu, B., Bu, J., Chen, C.,** 2010. Opinionword Expansion And Target Extraction Through Double Propagation, *Computational Linguistics, Vol: 1, No: 1*, pp. 1-18.
- [60] **Dini, L. ve Mazzini, G.,** 2002. Opinion Classification Through Information Extraction, *In Proceedings Of The Conference On Data Mining Methods And Databases For Engineering, Finance And Other Fields (Data Mining)*, Bologna, Italy, p. 299-310.

- [61] **Turney, P.**, 2002. Thumbs Up Or Thumbs Down: Semantic Orientation Applied To Unsupervised Classification Of Reviews, *In Proceedings Of The 40th Annual Meeting On Association For Computational Linguistics (ACL '02)*, Morristown, NJ, USA, p. 417-424.
- [62] **Saggion, H. ve Funk, A.**, 2010. Interpreting Sentiwordnet For Opinion Classification, *Proceedings Of The Seventh Conference On International Language Resources And Evaluation LREC10*, Malta, Publisher: European Language Resources Association (ELRA), p. 1129-1133.
- [63] **Devitt, A. ve Ahmad, K.**, 2007. Sentiment Polarity Identification In Financial News: A Cohesion-Based Approach, *In Proceedings Of The 45th Annual Meeting Of The Association Of Computational Linguistics, Association For Computational Linguistics*, Prague, Czech Republic, p. 984-991.
- [64] **Dave, K., Lawrence, S. ve Pennock, D.**, 2003. Mining The Peanut Gallery: Opinion Extraction And Semantic Classification Of Product Reviews, *In www '03: Proceedings Of The 12th International Conference On World WideWeb*, New York, NY, USA, p. 519-528.
- [65] **Yu, B., Kaufmann, S. ve Diermeier, D.**, 2008. Exploring the characteristics of opinion expressions for political opinion classification, *In Proceedings of the 2008 international conference on Digital government research*, Montreal, pp. 82-91.
- [66] **Pang, B. ve Lee, L.**, 2008. Opinion Mining And Sentiment Analysis, *Foundation And Trends, In Information Retrieval*, Vol: 2.
- [67] sentiment.vgunduzalp.com, Erişim Tarihi: 01.03.2016.

ÖZGEÇMİŞ

1989 yılında Çermik’te doğdu. İlk ve orta eğitimini Çermik’te yaptı. Yabancı Dil Ağırlıklı bir lise olan Siirt Fen Lisesinden 2007 yılında mezun olduktan sonra Fırat Üniversitesi Teknik Eğitim Fakültesi Bilgisayar Öğretmenliğinden 2012 yılında mezun oldu. Öğrencilik yıllarından Fırat Üniversitesinin, Bilgi İşlem Ar-Ge biriminde yarı zamanlı olarak birçok projede görev aldı. Lisans eğitimi stajını 2011 yılında Konya Teknokent’teki AE Kod Teknolojisi firmasında yaptı. Yine 2011 yılında “Ses Şiddetine Göre Yüz İfadesi Oluşturan Devre” isimli bitirme ödevini yaptı.

2013 yılında İnönü Üniversitesi Bilgi İşlem Daire Başkanlığında Uzman olarak göreve başladı. 2016 yılında “Arm Tabanlı Dijital Otoskop Tasarımı ve Kulak Hastalıkları Tespit Cihazı” isimli projesi Tübitak tarafından Bigg-1512 Teknogirişim Sermaye Destekliği Programı kapsamında desteklenmeye hak kazandı. Halen İnönü Üniversitesiden görevini sürdürmektedir.