

**T.C.
FIRAT ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ**

**ENFORMASYON SİSTEMLERİNDE SAKLI MARKOV
MODELİ VE BAYES TABANLI SINIFLANDIRICILAR İLE
BİLGİ MODELLERİNİN GELİŞTİRİLMESİ**

Adem DOĞANER

**Doktora Tezi
İstatistik Anabilim Dalı**

**Danışman: Doç. Dr. Sinan ÇALIK
OCAK-2015**

**T.C.
FIRAT ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ**

**ENFORMASYON SİSTEMLERİNDE SAKLI MARKOV MODELİ VE BAYES
TABANLI SINIFLANDIRICILAR İLE BİLGİ MODELLERİNİN
GELİŞTİRİLMESİ**

DOKTORA TEZİ

Adem DOĞANER

(102133201)

Anabilim Dalı: İstatistik

Programı: İstatistiksel Bilgi Sistemleri

Danışman: Doç. Dr. Sinan ÇALIK

Tezin Enstitüye Verildiği Tarih: 5 Ocak 2015

OCAK-2015

**T.C.
FIRAT ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ**

**ENFORMASYON SİSTEMLERİNDE SAKLI MARKOV MODELİ VE BAYES
TABANLI SINIFLANDIRICILAR İLE BİLGİ MODELLERİNİN
GELİŞTİRİLMESİ**

DOKTORA TEZİ

Adem DOĞANER

(102133201)

Tezin Enstitüye Verildiği Tarih : 5 Ocak 2015

Tezin Savunulduğu Tarih : 30 Ocak 2015

Tez Danışmanı :
Diğer Jüri Üyeleri :

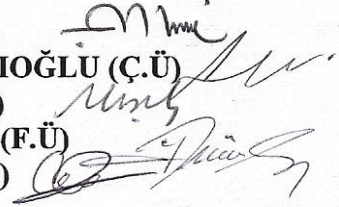
Doç. Dr. Sinan ÇALIK (F.Ü)

Prof. Dr. Sadullah SAKALLIOĞLU (Ç.Ü)

Doç. Dr. Mahmut IŞIK (F.Ü)

Doç. Dr. Mehmet GÜRCAN (F.Ü)

Doç. Dr. Cemil ÇOLAK (İ.Ü)



OCAK-2015

ÖNSÖZ

Tez konusunun belirlenmesi ve yürütülmesi aşamalarında, her türlü bilgi, yardım ve desteğini esirgemeyen değerli danışmanım Doç. Dr. Sinan ÇALIK'a, tezin teorik aşamalarında bilgilerini paylaşan ve desteklerini esirgemeyen bölüm hocalarıma, tez çalışmaları boyunca desteğini benden esirgemeyen eşim Gül TUZ DOĞANER'e, tecrübelerini ve desteğini benden esirgemeyen Yrd. Doç. Dr. Gülbahtiyar DEMİREL'e, her zaman yanımda destek olan DOĞANER ve TUZ ailesine teşekkür eder, sonsuz saygılarımı sunarım.

Adem DOĞANER
ELAZIĞ-2015

İÇİNDEKİLER

	<u>Sayfa No</u>
ÖNSÖZ.....	II
İÇİNDEKİLER.....	III
ÖZET.....	VI
SUMMARY.....	VII
ŞEKİL LİSTESİ.....	VIII
TABLO LİSTESİ.....	IX
SEMBOLLER LİSTESİ.....	XI
1. GİRİŞ.....	1
1.1. Çalışmanın Amacı.....	3
1.2. İlgili Çalışmalar.....	4
2. MATERYAL VE METOT.....	8
2.1. Veri Modelleri.....	8
2.1.1. Hiyerarşik Model.....	8
2.1.2. Ağ Modeli.....	10
2.1.3. İlişkisel Model.....	11
2.1.4. Nesneye Yönelik Veri Modeli.....	12
2.2. Bilgi Modelleme.....	12
2.3. Bilgi Modelleme Yaklaşımları.....	16
2.3.1. Varlık İlişki Modelleme.....	17
2.3.2. Nesne-Rol Modelleme (ORM).....	18
2.3.3. Birleşik Modelleme Dili (UML).....	21
2.3.4. Bilgi Modelleme Yaklaşımlarının Farklılıkları.....	25
2.4. Veritabanı ve Varlık İlişki Diyagramı.....	26
2.4.1. İlişkisel Model.....	28
2.5. Soyutlama Mekanizmaları.....	29
2.5.1. Sınıflandırarak Soyutlama.....	29
2.5.2. Kümeleyerek Soyutlama.....	29
2.5.3. Genelleyerek Soyutlama.....	30
2.6. Varlık İlişki Modelinin Kalitesi.....	30
2.6.1. Normal Formlar.....	30
2.6.2. Normalleştirilmiş ve Normalleştirilmemiş İlişkiler.....	31
2.6.3. Normalizasyon Kuralları.....	31
2.6.4. Veritabanı Oluşturma İşlemleri.....	32
2.7. Metin Madenciliği.....	32
2.7.1. Bilgi Erişimi.....	33
2.7.2. Bilgi Çıkarımı.....	34
2.7.3. Metin Madenciliğinde Kullanılan Yöntemler.....	34
2.7.4. Metin Özetleme.....	34
2.7.5. Konu Temsili Yaklaşımı.....	35
2.7.6. Frekans Odaklı Yaklaşımlar.....	35
2.7.7. Gizli Anlambilimsel Analiz.....	35

	<u>Sayfa No</u>
2.7.8. Bayescil Konu Modeli.....	36
2.7.9. Cümle Kümeleme.....	36
2.7.10. İçerik Etkisi.....	36
2.7.11. Makine Öğrenme Yöntemi.....	37
2.8. Metin Madenciliğinde Sınıflandırma Yöntemleri.....	37
2.8.1. Olasılıksal Sınıflandırma.....	37
2.8.2. Kural Tabanlı Sınıflandırma.....	38
2.8.3. Doğrusal Sınıflama.....	38
2.8.4. Destek Vektör Makinesi.....	38
2.8.5. Yapay Sinir Ağları.....	39
2.8.6. Regresyon Yöntemi.....	39
2.8.7. Karar Ağaçları.....	40
2.9. Markov Zincirleri.....	40
2.9.1. Geçiş Zamanı.....	43
2.10. Saklı Markov Modeli.....	44
2.10.1. Saklı Markov Modelinin Temelleri.....	45
2.10.2. Saklı Markov Modeli Tahmin Algoritması.....	47
2.10.3. Viterbi Algoritması.....	48
2.10.4. Baum-Welch Algoritması.....	49
2.10.5. İleri Algoritması.....	51
2.11. Bayes Tabanlı Sınıflandırıcılar.....	52
2.11.1. Bayes Teorisi.....	53
2.11.2. Naive Bayes Sınıflandırıcılar.....	53
2.11.3. Bayes Ağları.....	55
3. UYGULAMA.....	56
3.1. Modelin Geliştirilmesi.....	56
3.1.1. Türkçe’de Kelimeler ve Cümle Dizilimi.....	56
3.1.2. İşlem Aşamaları.....	56
3.1.3. Kelimelerin Anlam Değiştirmeyen Eklerinden Ayrılması.....	57
3.1.4. Saklı Markov Modeli ve Bayes Tabanlı Sınıflandırıcılar ile Kelimelerin Sınıflandırılması.....	60
3.1.5. Varlık İlişki Modeli ile Enformasyon Çıkarım Alanlarının Düzenlenmesi.....	61
3.1.6. Verilerin Veritabanına Aktarılarak Enformasyon Çıkarımının Gerçekleştirilmesi.....	62
4. BULGULAR.....	64
4.1. Kelimleri Anlam Değiştirmeyen Eklerden Ayırma İşlemlerinin Metin Belgesine Uygulanması.....	64
4.2. Saklı Markov Modeli ve Bayes Tabanlı Sınıflandırıcılar ile Kelimleri Sınıflandırma İşlemlerinin Metin Belgesine Uygulanması.....	66
4.3. Varlık İlişki Modeli ile Enformasyon Çıkarım Alanlarını Düzenleme İşlemlerinin Metin Belgesine Uygulanması.....	72

	<u>Sayfa No</u>
4.4. Verilerin Veritabanına Aktarılarak Enformasyon Çıkarımı Gerçekleştirme İşlemlerinin Metin Belgesine Uygulanması.....	74
5. SONUÇLAR ve TARTIŞMA.....	82
6. ÖNERİLER.....	87
KAYNAKLAR.....	88
EKLER.....	96
ÖZGEÇMİŞ.....	104

ÖZET

Bilgi teknolojilerinin gelişmesi ile birlikte bilgisayar ve internet ortamında çok sayıda metin yığını oluşmuştur. Kullanıcılar, bu yoğunluktaki metin yığınları karşısında istedikleri bilgiyi elde etmek için zorluklar ile karşılaşmaktadırlar. Kullanıcılar çoğu kez bu verileri incelemek için zaman bulamamaktadır. Mevcut enformasyon sistemleri ise bu yoğunluktaki kitlelerden enformasyon sağlamada yetersiz kalmaktadır.

Çalışmada bilgisayar ve internet ortamındaki metin içeriklerinden otomatik olarak enformasyon sağlayacak model geliştirilerek büyük boyutlardaki metin kitlelerinden, bilgisayar ortamında otomatik olarak enformasyonun elde edilmesi amaçlanmaktadır. Çalışmada aşamalı işlemlerden oluşan kompleks bir model geliştirilmiştir. Modelde Türkçe'nin sondan eklemeli dil yapısı dikkate alınarak kelimeleri eklerinden ayırma işlemleri için algoritma oluşturulmuştur. Modelin teorik yapısını oluşturan Saklı Markov modeli ve Bayes tabanlı sınıflandırıcılar ile kelimeler türlerine göre sınıflandırılmıştır. Kelimeler ile bu kelimelere ilişkin veriler ve alanlar, varlık ilişki modeli ile incelenerek veritabanına aktarılmış ve yapısal sorgulama dili (SQL) ile enformasyon çıkarım işlemleri gerçekleştirilmiştir.

Modelin internet ortamından elde edilen 584 kelimelik Türkçe metin içeriğine uygulanması sonucunda, kelimeleri türlerine göre sınıflandırma başarısı % 90,8 olarak elde edilmiştir. Ayrıca model 584 kelimelik metin kitlesinden otomatik olarak 13 kelimelik anlamlı ve metni temsil yeteneği yüksek bir enformasyon sağlamıştır.

Türkçe Dili'nde yazılmış metin içeriklerinden enformasyon sağlamak için geliştirilen model, kelimeleri türlerine göre sınıflamada ve otomatik enformasyon elde etme konusunda başarılı sonuçlar sağladığı gözlemlenmiştir.

Anahtar Kelimeler: Saklı Markov Modeli, Bayes Sınıflandırıcılar, Bilgi Sistemleri, Metin Madenciliği, Veri Madenciliği, Varlık İlişki Modeli

SUMMARY

Development of Information Models with the Hidden Markov Model and Bayes-Based Classifiers in Information Systems

Along with the developments in information technologies, a large number of texts accumulated in the computer and Internet environments. The users are faced with difficulties in accessing the information they seek in the face of such dense accumulation of texts. Mostly, the users cannot find the time to go through such data. The existing information systems remain inadequate in providing information from such dense masses of data.

The present study aims to develop a model that will automatically provide information from text content in the computer and Internet environments, thereby enabling automatic access to information from large-scale masses of texts in the computer environment. The study resulted in the development of a complex multiphase model. The model was prepared by taking into consideration the agglutinative structure of Turkish and through the development of an algorithm for the separation of words from suffixes. The Hidden Markov Model and Bayes-based classifiers, that constituted the theoretical structure of the model, were used to classify the words with respect to word types. Words and their associated data and fields were examined through the entity-relationship model and subsequently transferred to the database. This was followed by the extraction of information through the Structured Query Language (SQL).

Once the model was applied to the content of a Turkish text of 584 words retrieved from the Internet environment, a success rate of 90.8% was achieved in the classification of words by type. In addition, the model enabled a body of meaningful and representative information of 13 words to be obtained automatically from the textual mass of 584 words.

Developed with the aim of providing information from textual content in Turkish, the model was observed to offer successful results in classifying words by type and obtaining information automatically.

Key Words: Hidden Markov Model, Bayes Classifiers, Information Systems, Text Mining, Data Mining, Entity-Relationship Model.

ŞEKİL LİSTESİ

	<u>Sayfa No</u>
Şekil 2.1.	Varlık ilişki model diyagramına ilişkin örnek..... 17
Şekil 2.2.	Nesne-rol model diyagramına ilişkin örnek..... 19
Şekil 2.3.	Birleşik modelleme dili diyagramına ilişkin örnek..... 22
Şekil 2.4.	Dışsal kısıtlama için karşı örnekleme şeması..... 24
Şekil 2.5.	Veritabanı diyagramına ilişkin örnek..... 27
Şekil 3.1.	Ekleri ayırma ve ek harflerin sayılarak gruplara ayrılmasına ilişkin algoritma..... 59
Şekil 4.1.	Saklı Markov modeline ilişkin diyagram..... 68
Ek Şekil 3.1.	Harput ve iklim enformasyon alan kelimelerinin sorgulanması için düzenlenen SQL kodları..... 98
Ek Şekil 3.2.	Harput ve sanayi enformasyon alan kelimelerinin sorgulanması için düzenlenen SQL kodları..... 98
Ek Şekil 3.3.	Harput ve tarım enformasyon alan kelimelerinin sorgulanması için düzenlenen SQL kodları..... 99
Ek Şekil 3.4.	İklim ve sanayi enformasyon alan kelimelerinin sorgulanması için düzenlenen SQL kodları..... 99
Ek Şekil 3.5.	İklim ve tarım enformasyon alan kelimelerinin sorgulanması için düzenlenen SQL kodları..... 100
Ek Şekil 3.6.	Elazığ, Harput, sanayi, iklim ve tarım enformasyon alan kelimelerinin sorgulanması için düzenlenen SQL kodları..... 100
Ek Şekil 3.7.	Elazığ, Harput, iklim enformasyon alan kelimelerinin sorgulanması için düzenlenen SQL kodları..... 101
Ek Şekil 3.8.	Elazığ, Harput, iklim ve sanayi enformasyon alan kelimelerinin sorgulanması için düzenlenen SQL kodları..... 101

TABLO LİSTESİ

	<u>Sayfa No</u>
Tablo 2.1. Veritabanında bulunan tablolara ilişkin örnek.....	15
Tablo 3.1. Kelime türlerinin gruplandırılması.....	58
Tablo 3.2. Ek harf sayılarının gruplandırılması.....	58
Tablo 3.3. Durum sembolleri ve gözlem sembolleri.....	60
Tablo 4.1. Metin belgesindeki kelimeler ve kelime sıra numaraları.....	65
Tablo 4.2. Birinci bölümdeki işlemlerin metin belgesine uygulanması ile elde edilen veriler.....	65
Tablo 4.3. Saklı Markov modelinde durum ve gözlem sembollerinin temsili.....	66
Tablo 4.4. Saklı Markov modeli için düzenlenen R kodları.....	66
Tablo 4.5. Başlangıç olasılıkları vektörü.....	67
Tablo 4.6. Durumlar arası geçiş olasılıkları matrisi.....	67
Tablo 4.7. Emisyon olasılıkları matrisi.....	67
Tablo 4.8. Baum-Welch algoritması için düzenlenen R kodları.....	69
Tablo 4.9. Baum-Welch algoritması ile elde edilen başlangıç olasılıkları vektörü.....	69
Tablo 4.10. Baum-Welch algoritması ile elde edilen durumlar arası geçiş olasılıkları matrisi.....	69
Tablo 4.11. Baum-Welch algoritması ile elde edilen emisyon olasılıkları matrisi.....	70
Tablo 4.12. “Elazığ” kelimesine ilişkin gözlem frekansları ve emisyon olasılıkları.....	70
Tablo 4.13. İkinci bölümdeki işlemlerin metin belgesine uygulanması ile elde edilen veriler.....	71
Tablo 4.14. Enformasyon alan kelimelerinin belirlenmesi.....	73
Tablo 4.15. Üçüncü bölümdeki işlemlerin metin belgesine uygulanması ile elde edilen veriler.....	74
Tablo 4.16. Veritabanında alanların tanımlanması.....	75
Tablo 4.17. Veritabanına aktarılan verilere ilişkin bir kesit.....	75
Tablo 4.18. Elazığ ve Harput enformasyon alan kelimelerinin sorgulanması için düzenlenen SQL kodları.....	76
Tablo 4.19. Elazığ ve Harput kelimelerinin birlikte kullanıldığı cümleler ve fiiller.....	76
Tablo 4.20. Elazığ ve Harput enformasyon alanlarından elde edilen yeni cümlelerin kelimeleri.....	77
Tablo 4.21. Elazığ ve iklim enformasyon alan kelimelerinin sorgulanması için düzenlenen SQL kodları.....	77
Tablo 4.22. Elazığ ve iklim kelimelerinin birlikte kullanıldığı cümleler ve fiiller.....	77
Tablo 4.23. Elazığ ve iklim enformasyon alanlarından elde edilen yeni cümlelerin kelimeleri.....	78

Tablo 4.24.	Elazığ ve tarım enformasyon alan kelimelerinin sorgulanması için düzenlenen SQL kodları.....	78
Tablo 4.25.	Elazığ ve tarım kelimelerinin birlikte kullanıldığı cümleler ve fiiller.....	79
Tablo 4.26.	Elazığ ve tarım enformasyon alanlarından elde edilen yeni cümlelerin kelimeleri.....	79
Tablo 4.27.	Sanayi ve tarım enformasyon alan kelimelerinin sorgulanması için düzenlenen SQL kodları.....	79
Tablo 4.28.	Sanayi ve tarım kelimelerinin birlikte kullanıldığı cümleler ve fiiller.....	80
Tablo 4.29.	Sanayi ve tarım enformasyon alanlarından elde edilen yeni cümlelerin kelimeleri.....	80
Tablo 4.30.	Fiillerin ekler ile tamamlanması.....	81
Tablo 4.31.	Metin belgesinden model ile doğrudan elde edilen enformasyon.....	81
Tablo 5.1.	Viterbi algoritması ile geliştirilen modelin sınıflama başarısının karşılaştırılması.....	83

SEMBOLLER LİSTESİ

O_i	:Gözlem Sembolü
S_i	:Durum Sembolü
HMM	:Saklı Markov Modeli
P	:Olasılık
ORM	:Nesne Rol Modelleme
UML	:Birleşik Modelleme Dili
ERM	:Varlık İlişki Modeli
TF*IDF	:Terim Frekansı* Ters Belge Frekansı
SDD	:Sınıflandırma Durum Değeri
SQL	:Yapısal Sorgulama Dili

1. GİRİŞ

Teknolojinin gelişimi, insanların hayatına pek çok farklılık kazandırdığı gibi yaşamı kolaylaştıran kazanımlar sağlamıştır. Teknolojik gelişmeler, dünyada yeni olguların ortaya çıkmasına imkan tanımaktadır. Teknolojik alanlardaki ilerlemenin en somut örneklerinden biri de şüphesiz ki bilgiye ve bilgisayara bağlı teknolojilerdir. Bilgisayar teknolojilerindeki gelişmeler, yirminci yüzyılda bilgisayar mimarilerinin geliştirilmesi ve ilk bilgisayarların ortaya çıkması ile büyük bir ivme kazanmıştır. Bireysel kullanıma sunulan bilgisayarlar oldukça yaygınlaşmıştır. Mimari özellikleri gereği veriyi saklayabilme ve işlem yapabilme gibi özellikleri sayesinde tüm veriler bilgisayarda tutulmaya ve burada işlenmeye başlanmıştır. Verinin bilgisayarda saklanabilmesi ve işlem yapılabilmesi, bilgi üretiminde yeni bir sayfa açmıştır. Fakat bilgisayarların ortaya çıktığı ilk zamanlarda ağ teknolojisinin gelişmemiş olmasından dolayı bilgisayarlarda tutulan bilgiler farklı kullanıcılar tarafından paylaşılamamış ve yeni bilgi üretimi sınırlı kalmıştır. Bilgi üretimindeki esas gelişim internet teknolojisinin ortaya çıkışı ile yaşanmıştır.

İnternetin yirminci yüzyılın sonlarına doğru kullanılabilmesi ile bilgisayarlar arasında ağ oluşturarak bağlantı kurabilen bilgisayarlar, bilginin üretilmesi ve iletilmesi adına önemli kazanımlar elde etmiştir. İnternet ile dünyanın herhangi bir yerindeki bilgisayarda bulunan bilgi, anında dünyanın herhangi bir yerindeki başka bir bilgisayar ile paylaşılabilir. Bilginin bu hız ile iletimi, bilginin anında farklı kitlelere ulaşımını sağlarken, kitleler tarafından işlenen bilgiler sonucunda yeni bilgiler elde edilmesini kolaylaştırmaktadır. Yeni bilgilerin üretimi yine bu işlemler silsilesi aracılığı ile sağlanmaktadır.

Bilginin üretimi internetin ortaya çıkışına dek düşük ivmeli bir artış ile gelişmekte iken, internet ile büyük bir sıçrama yaşamıştır. İnternet sonrasında yoğun bir bilgi birikimi gerçekleşmiştir. Bilgi yığınları günümüzde de güncel bir şekilde hızla artmaya devam etmektedir.

Sayısal ortamdaki bilgi yığınlarının tamamının incelenmesi mümkün değildir. Bu sorun, bilgisayar bilimlerinde yeni alanlar doğmasını sağlamıştır. Her kullanıcı ihtiyacı doğrultusunda arzu ettiği bilgiye erişmek istemektedir. Kullanıcının istediği bilgiyi, bilgi yığınlarını inceleyerek bulması zaman açısından mümkün değildir. Kullanıcıların bu ihtiyacını gidermek üzere, büyük boyutlardaki bilgi yığınlarından arzu edilen bilginin

çekilmesini sağlayacak enformasyon sistemleri geliştirilmiştir. Bu enformasyon sistemleri istenilen bilgileri, veritabanlarından sorgulama, filtreleme gibi yöntemler ile elde edebilmesinin yanı sıra, büyük metin kitlelerinden metin madenciliği yöntemleri ile elde edebilmektedir.

Uygulanan yöntemler sonucunda elde edilecek sonuç, veri, enformasyon veya bilgi değeri taşımaktadır. Veri, enformasyon ve bilgi kavramları esasında birbiri ile yakından ilişkili kavramlar olmasına rağmen, anlamsal olarak birbirinden farklıdır [1]. Veri, enformasyon ve bilginin herhangi birinin tanımını yapmak için diğer kavramlara da ihtiyaç duyulmaktadır. Enformasyonun oluşumu için veriye ihtiyaç duyulduğu ve bilginin oluşumu için ise enformasyona gereksinim duyulduğu belirtilerek, verinin enformasyona ve enformasyonunda bilgiye dönüştüğü süreçler olduğu ifade edilmektedir [2]. Bu ifadeler dikkate alındığında veriden yola çıkarak bilgiye doğru uzanan bir işlem süreci ortaya çıkmaktadır. Bu işlem sürecinin başlangıç aşaması olarak veri kavramı ön plana çıkmaktadır.

Veri, içeriğe sahip olmayan ve bu haliyle anlamı olmayan yapılar olarak ifade edilmektedir [3]. Bu ifade ile birlikte verinin, enformasyonun ve bilginin hammaddesi olduğu söylenebilmektedir. Verinin yorumlardan ve değerlendirmelerden bağımsız olarak sunulması nedeniyle anlam ifade etmemektedir. Bu nedenle veri, enformasyon ve bilgi formuna dönüştürülmeden bilgi sağlama açısından güvenli bir yapı oluşturmamaktadır [4].

Enformasyon, bilginin oluşumundaki ikinci aşamadır. Verilerin bazı yöntem ve işlemler ile düzenlenmiş formal yapısı, enformasyon olarak nitelendirilmektedir. Enformasyonda veriler nispeten anlam kazanmıştır ve düzenlenen veriler birbirleri ile ilişkilidir. Veriler genellikle belirli işlemlerden geçirilerek enformasyona dönüştürülür ve enformasyona dönüştürüldükten sonra amaca uygun olarak karar aşamalarında ve çözüm üretmede önemli rol üstlenmektedirler.

Bilgi sürecindeki son adım ise bilginin oluşumudur. Birçok çalışmada bilgi ve enformasyon aynı anlamda kullanılmasına karşın anlam ve yapısal olarak birbirinden farklıdır. Enformasyon bilgi oluşumundaki bir aşama olarak değerlendirildiğinde bilgi, enformasyonun bireyde oluşturduğu anlamsal kavram olarak nitelendirilmektedir. Bilginin, bireylerin çevresini doğru olarak algılamasını sağlayan enformasyon olarak belirtilmesi, enformasyonun bilgi açısından önemini ortaya koymaktadır [5].

Veri, enformasyon ve bilginin her ne kadar aynı yapılar olmadığını söylemek mümkün ise birbirinden tamamen soyutlamak, yanlış bir değerlendirme olacaktır. Üç kavram birbiri ile ilişkili ve tamamlayıcı kavramlardır.

1.1. Çalışmanın Amacı

Bilgisayarların ve internetin yaygınlaşması sonucunda sayısal ortamda veri yığınları oluşmuştur. Kullanıcılar bu veri yığınları karşısında istedikleri bilgiyi elde etme açısından büyük zorluklar ile karşılaşmaktadırlar. Öte yandan tüm verileri gözden geçirecek zaman ve enerji bulunmamaktadır. Bazı durumlarda kullanıcılar, istedikleri küçük boyutlu bir bilgiyi elde etmek için yüzlerce hatta binlerce sayfadan oluşan belgeleri gözden geçirmek durumunda kalmaktadırlar. Mevcut enformasyon sistemleri, bu gibi durumlar için kısmi çözümler sağlamaktadır. Nitekim arama motorlarındaki mantığın dayandırıldığı, web sayfalarındaki aranan konu ile ilgili web sayfalarını listeleyen bilgi erişim modelleri, metin yığınlarından sözcükleri sınıflayan veya filtreleyen metin madenciliği yöntemleri bu duruma örnek olarak verilebilmektedir. Mevcut modeller lokal sorunlar için başarılı sonuçlar üretebilirken, kompleks yapıdaki sorunların çözümlerinde istenilen sonucu sağlamada sınırlı kalmaktadırlar. Kullanıcılar sadece aradığı konuya ilişkin web sayfalarına veya metin belgelerine erişmekten ziyade, elde ettikleri bu belgeleri okumadan ve zaman kaybetmeden belgelere ilişkin enformasyon elde etmek istemektedirler. Öte yandan metin belgelerinden bilgi çıkarımı elde etmek için geliştirilen modeller genelde İngilizce dilini esas almaktadır. İngilizceyi esas alarak geliştirilen modeller Türkçe metin belgeleri üzerine uygulandığında yeterince başarılı sonuçlar alınamamaktadır. Türkçe'nin sondan eklemeli bir dil olması ve detaylı bir yapısal içeriğe sahip olması nedeniyle, mevcut modeller ile bilgi çıkarımını güçleştirmekte ve metinlerin stabil yapı içeren bir model ile değerlendirilmesini mümkün kılmamaktadır.

Enformasyon sistemlerinde kullanılan bazı bilgi modelleri, kompleks yapılarda enformasyon elde edilmesinde kullanıcının isteklerine yeterince cevap verememektedir.

Bu çalışmada, stokastik yöntemlerden Saklı Markov modeli ve Bayes tabanlı sınıflandırıcılar kullanılarak enformasyon sistemlerindeki bilgi modellerinin geliştirilmesi amaçlanmaktadır. Bu modelin geliştirilmesi ile;

- Büyük boyutlu metin belgelerindeki binlerce sayfayı incelemeye gerek duyulmadan kullanıcının arzu ettiği ilgili enformasyonun otomatik olarak elde edilmesi amaçlanmaktadır.
- İlgili modelin sadece metin belgelerinde değil web sayfalarında da uygulanabilir yapıda olması amaçlanmaktadır.
- Doğrudan Türkçe dilbilgisi esas alınarak geliştirilen model, Türkçe'nin sondan eklemeli dil yapısı dikkate alınarak başarılı enformasyon sağlaması amaçlanmaktadır.
- Modelin uygulanması ile elde edilen enformasyonların metin özetine temel teşkil etmesi amaçlanmaktadır.
- Geliştirilen modelin incelediği metin belgelerindeki sözcükleri veritabanına aktarması amaçlanmaktadır.
- Veritabanındaki kelimeleri türlerine göre sınıflandırarak, Türkçe metinlerdeki kelimelerin tasnifinde kullanılması amaçlanmaktadır.
- Kullanıcılar açısından zaman kaybının önlenerek iş hayatında büyük pratiklik sağlaması amaçlanmaktadır.
- Teorik stokastik modelleri baz alarak geliştirilen algoritmik modelin, metin madenciliği ve enformasyon bilim alanlarında Türkçe metinler için farklı bir model olarak kazandırılması amaçlanmaktadır.
- Modelin iş dünyasında raporlama konusunda büyük kolaylık sağlaması amaçlanmaktadır.
- Bilgi ve bilgisayara bağlı bilim alanlarında farklı bir model kazandırılması amaçlanmaktadır.

1.2. İlgili Çalışmalar

Enformasyon sistemlerine ilişkin model geliştirme çalışmalarının başlangıcı, tarihsel olarak yakın geçmişe dayanmasına rağmen konunu önemi ve pratikte kullanım yoğunluğu nedeniyle pek çok bilim insanının ilgi konusu olmuştur. Enformasyon sistemlerine ilişkin pek çok alanda modeller geliştirilirken veri ve bilgileri düzenlemeye ilişkin yöntemler geliştirilmiştir.

Bilgi modelleme üzerine, bilgi çıkarımı ve metin madenciliği alanında yöntemler geliştirmek için, farklı algoritma ve modeller geliştirilmiştir. Geliştirilen bu model ve algoritmalar pek çok çalışmada uygulanmıştır. He ve Ling [6] çalışmalarında, ontoloji tabanlı modellerine varlık ilişki modelini uygulayarak meta veri ve meta bilginin sınıflandırılmasında kullanmışlardır. Bu şekilde içeriğin semantik çözümlemesini entegre şema üzerinde sağlamışlardır. Alhajj [7] varlık ilişki modelini kullanarak belgeler üzerinde çıkarım gerçekleştirebilecek bir algoritma üzerine çalışmıştır. Arslan [8] çalışmasında veri madenciliği yöntemlerinden yararlanarak Türkçe metinler için anlamsal bilgi çıkarımı üzerine çalışmıştır. Çalışmada ilişkilerin tanımlanması aşamasında destek vektör makinası algoritmasından yararlanmıştır.

Genellikle varlık-iliski bilgi modeline dayanarak tasarlanan sistemler, veritabanı oluşturmak için uygulanmaktadır. Veritabanına aktarılan veri setinden, bir sonraki aşamada verilerin özetlenmesi, filtrelenmesi gibi işlemler sonucunda enformasyonlar elde edilmektedir. Bu işlemler genel olarak veri madenciliği olarak nitelendirilmektedir. Öte yandan metin belgelerinde de filtreleme, sorgulama, özetleme gibi işlemler uygulanabilmektedir. Bu işlemler sonucunda metin belgelerinden ilgili bilgiye doğrudan erişim, metin filtreleme, metin özetleri ve bilgi çıkarımı sağlanabilmektedir. Bu işlemler ise metin madenciliği olarak ifade edilmektedir. Metin madenciliği yöntemleri pek çok araştırmacı tarafından dikkate alınmıştır. Tsai ve Chang [9] metin sınıflandırma için örnek seçimlerini dikkate alan modelinde destek vektör makinası algoritmalarından yararlanmıştır. Metin sınıflama için çok sayıda model uygulanabilmektedir. Ghiassi ve arkadaşları [10] otomatik metin sınıflandırma için dinamik yapay sinir ağları modeli üzerine çalışmışlardır. Chen ve Chen [11] çalışmalarında metin sınıflandırması için benzerlik ölçüsünden ve ki-kare istatistiğinden yararlanmıştır. Jiang ve arkadaşları [12] özellik çıkarım tabanlı grafik madenciliği yöntemini kullanarak metin sınıflandırma üzerine çalışmışlardır. Klose ve arkadaşları [13] belge benzerlikleri üzerine metin erişimi konusunda çalışmışlardır. Bayer [14] metin madenciliği yöntemlerini kullanarak kelimeleri köklerine ayırmış ve kelimelerin anlam bilgisini tahmin etmeye yönelik uygulamalarda bulunmuştur. Bazı çalışmalarda metin madenciliği yöntemleri tıbbi belgelerin incelenmesinde uygulanmıştır [15,16]. Metin madenciliği, uygulama sahası ve kullanılan yöntemler bakımından geniş bir çeşitlilik sunmaktadır. Uygulama alanları bakımından gerek sağlık bilimleri alanlarında gerekse de sosyal bilimler alanlarında metinlerin çözümlenmesinde sıklıkla kullanılabilmektedir. Metin madenciliği, pek çok model ve

algoritma kullanarak çözüm sağlamaktadır. Bu yöntemler istatistiksel teknikleri içerebildiği gibi makine öğrenme tekniklerini de içerebilmektedir. Bazı çalışmalarda metin madenciliği yöntemleri farklı sahalarda uygulanarak çözümler elde edilmiştir [17-24].

Metin madenciliği kendi içerisinde sıklıkla kullanılan bilgi erişimi ve bilgi çıkarımı gibi alt alanlara ayrılmaktadır. Temel olarak bilgi erişimi ve bilgi çıkarımı konuları birbirlerinden farklılık göstermektedir. Bilgi erişimi var olan bilgiye belirli ipuçları ve sorgular eşliğinde ilgili belgeleri bularak listelemektedir. Bu işlemler web tabanlı arama motorlarının teorik yapısı ile yakından ilgilidir.

Bilgi erişim modelleri, internet ve bilgisayar bilimlerindeki güncel gelişmelere bağlı olarak gelişim göstermektedir. Bilgi erişimine ilişkin yeni yöntem ve algoritmalar ileri sürülmektedir. Bu yöntemlerden biri terim frekansı-ters belge frekansı (TF*IDF) yaklaşımıdır. Kocabaş [25] çalışmasında istatistiksel tekniklere bağlı olarak bilgi erişim modellerindeki indeks terim ağırlıklandırma yöntemi üzerine çalışmıştır. Salton ve Buckley [26] otomatik metin erişimi üzerine terim ağırlıklandırma yaklaşımı konusunda çalışmışlardır. Can ve arkadaşları [27] Türkçe metinler üzerine bilgi erişimi konusunda çalışmışlardır. Ceylan [28] çalışmasında bilgi erişimine fonetik işaretlerin etkisini araştırmıştır. Sbattella ve Tedesco [29] çalışmalarında doğal dil ile yazılmış belgelere üç katmanlı alan modeli tabanlı bilgi erişim sistemini uygulamıştır. Bilgi erişim modellerinin uygulandığı pek çok alan bulunmaktadır. Bu modeller özellikle web tabanlı belgeler için başarılı sonuçlar sağlamaktadır [30-34].

Bilgi çıkarımı da tıpkı bilgi erişimi gibi metin madenciliğinin önemli alanlarından biridir. Bilgi çıkarımında enformasyon, metin belgesinden veya içerikteki yığınlar içerisinde ilgili terimler aracılığı ile elde edilmektedir. Bilgi çıkarımı kavramsal açıdan bilgi erişiminden farklıdır. Bilgi erişiminde ilgili koşulları sağlayan tüm belgeler listelenirken, bilgi çıkarımında metin veya içerik yığınlarından ilgili terimler aracılığı ile enformasyon sağlanmaktadır. Bilgi çıkarımı üzerine farklı modeller geliştirilmiştir. Birçok bilgi çıkarımı yöntemi algoritmalar ile sonuç elde etmektedir. Bu algoritmalar metin yığınlarından bilgi sağlamada başarılı sonuçlar elde etmektedir. Ko ve Seo [35] çalışmalarında istatistiksel tekniklerden yararlanarak metin özetleme ve cümle çıkarımı gibi işlemlerde başarılı sonuçlar elde etmişlerdir. Hobbs [36] tıbbi metin belgelerinden bilgi çıkarımı gerçekleştirmiştir. Downey ve arkadaşları [37] olasılıksal model analizleri ile web ortamından denetimsiz bilgi çıkarımı üzerine çalışmışlardır. Mykowiecka ve arkadaşları [38] klinik verilerden bilgi çıkarımı sağlamışlardır. Bu işlemi kural tabanlı bilgi

ıkarım yntemleri ile elde etmiřlerdir. Bilgi ıkarım yntemleri Trke Dili'nde yazılmıř metin belgeleri zerinde uygulanabilmektedir. Birok arařtırmacı bu yntemleri kullanarak Trke metinlerden bilgi ıkarımı yapacak algoritma ve modeller zerinde alıřmıřtır [39-42].

Metin madenciliėi yntemlerinde sıklıkla bařvurulan yntemlerden biride stokastik modellerdir. Farklı alıřmalarda bilgi saėlama adına stokastik modellerden sıklıkla yararlanılmıřtır [43,44].

2. MATERYAL VE METOT

2.1. Veri Modelleri

2.1.1. Hiyerarşik Model

Hiyerarşik veri modeli tüm verilerin hiyerarşik bir düzende oluşturulmasını esas almaktadır. Hiyerarşik model yapısı üst ve üste bağlı ast ilişkisi ile tasvir edilmektedir. Bu yapı çocuk-ata ilişkisi ile de temsil edilebilmektedir. Bir firma, personeller açısından alt kademe personeller ve bu alt kademenin bağlı olduğu üst kademe yöneticilerden oluşmaktadır. Üst kademe yöneticiye bağlı birden fazla alt kademe personel bulunmakta iken, her alt kademe personel sadece bir üst yöneticiye bağlıdır [45].

Hiyerarşik veri modelinde kayıtlar, ağaç modelinde tutulmaktadır. Üst hiyerarşide kök bulunmakta iken alt hiyerarşiye inildiğinde dallanmalar artmaktadır. Alt hiyerarşideki her birim bir bağlantı noktasından üst birimlere bağlanmaktadır. Hiyerarşik veri modelinde alt birimlere bağlı dallar yönlendirilmemiştir. Fakat hiyerarşik yapı yukarıdan aşağıya doğru yönelmektedir. İlişkilerin kurulması esnasında varlıkların tek bir varlığa bağlanması gerekmektedir. Bilgi taraması yapılmak istendiğinde hiyerarşinin en üst yapısından başlanarak alt birimlere doğru ilerlenmektedir. Kayıtların bağlı olduğu dallardan birinin silinmesi durumunda, ilgili dala bağlı tüm kayıtlar silinmektedir [47]. İlişkilerin ayrı olarak belirtilmesinde ve ilişkilerde veri tekrarı gerçekleştirmeyerek veritabanında fazla alan kaplanmasının önüne geçilmesi ile hiyerarşik modelde çoklu ilişkilerin optimal şekilde ifade edilmesi sağlanmaktadır [48]. Hiyerarşik kayıtlar, adres ve adres bilgisinin tutulduğu işaretçiler aracılığı ile birbirine bağlanmaktadır. Kayıtlar iki veya daha fazla işaretçi alanı içermektedir. İşaretçi alanının biri alt birime bağlı iken diğeri ise bağlantılı kayıta aittir. Hiyerarşik modelin çok hızlı şekilde tarama işlemini gerçekleştirmesi en büyük artılarından [49].

Veritabanında bilgiler, dosyalarda, kayıtlarda ve alanlarda düzenlenmiştir. Personellere ilişkin dosyalar, firmanın tüm personelleri hakkında tüm bilgileri içermektedir. Her personelin personel kayıtları bulunmaktadır, bu nedenle personel dosyaları bireysel personel kayıtlarını içermektedir. Dosyadaki her kayıt benzer bir yöntem ile düzenlenmektedir. Bunu örnek üzerinden açıklamak gerekirse, personele ait herhangi bir

bilginin her kayıt için aynı yerde düzenlenmesi beklenmektedir. Benzer şekilde personele ait diğer bilgilerinde her bir kayıta aynı yerde bulunması beklenmektedir. Bu durum, kayıtlarda “alan” olarak adlandırılmaktadır. Her bir personel bilgisi, tüm kayıtlarda alan olarak ifade edilmektedir. Alt kademe personelin kayıtları, alanların tüm türlerini içerebilmektedir. Hiyerarşik modelde ayrıca kayıtlar ile bağlantılı ilişkiler söz konusudur. Personel veya firma ile bağlantılı birimlerin alanları ve kayıtları da hiyerarşik modelde desteklenebilmektedir. Bu durumda iki ayrı dosya ambarı mevcuttur. Bu dosya ambarlarından biri personele ilişkin kayıtları diğeri ise firma ile bağlantılı birimlerin kayıtlarına aittir. İki dosya ambarı arasındaki yapı “ilişki” olarak adlandırılmaktadır. Firma ile bağlantılı birimler, personele bağımlı olmak zorundadır. Öte yandan personelin, firma ile bağlantılı birimlere bağımlı olma zorunluluğu bulunmamaktadır [45,65].

Veritabanı modellerinin tümünde ilişkiler yapısal kısıtlama olarak nitelendirilmektedir. Yapısal kısıtlama, önemlilik ve opsiyonellik olmak üzere iki notasyona sahiptir. Önemlilik bir kayıt türünden diğer kayıt türlerine kaç tanesi ile bağlantılı veya birkaç kayıt türünden bir kayıt türüne bağlantılı olduğu ile ilgilidir. Bu durumda önemlilik birden çok (1 to M), veya çoktan çok (M to M) bağlantılı olabilmektedir. Opsiyonellik ise bir kaydın diğer dosyalardaki kayıtlara karşılık gelen bir kaydın var olup olmadığı ile ilgilidir. Örnek ile ifade edilirse, personelin firma bağlantılı birimlere karşı ilişkisi opsiyonel veya kısmi olarak ifade edilmektedir. Firma bağlantılı birimlerin personele olan ilişkisi ise zorunlu ve tam olarak ifade edilmektedir. Yani personelin firma bağlantılı birimler ile ilişkili olma zorunluluğu bulunmaz iken, firma bağlantılı birimlerin bir personel ile ilişkili olması gerekmektedir. Personelden firma bağlantılı birimlere ilişki birden çok (1 to M) önemliliğe sahiptir [45].

Hiyerarşik modelde kayıtlar arasındaki tüm ilişkiler birden çok (1 to M) veya birden bire (1 to 1) önemliliğe sahiptir. Hiçbir zaman çoktan bire (M to 1) ve çoktan çok (M to M) ilişki gerçekleştirmezler [45].

Hiyerarşik veritabanındaki temel yöntem, üst ve alt kademelerdeki kayıtların bağlantılarına bazı fiziksel yolların seçimini içeren yöntemin uygulanmasıdır. Personel-firma bağlantılı birim ilişkisini uygulamanın bir yolu, personel kayıt noktasının firma bağlantılı birimlerin kayıtlarına ve firma bağlantılı birimlerin kayıt noktasının sonraki firma bağlantılı birimlere sahip olması gereksinimini doğurmaktadır [45].

Alt ve üst kademeye ait kayıtları bağlamak için bağlantılı liste yaklaşımında sistemin avantajları ve dezavantajları bulunmaktadır. Her personel sadece bir noktada iken firma

bağlantılı birimlerin bağlantı listesinin boyutunun teorik olarak sınırsız olması önemli bir avantajdır. Öte yandan herhangi bir firma bağlantılı birime ilişkin kayıtların zarar görmesi zincirin kırılmasına neden olmakta ve bu sistemin kırılabilirliğini gösteren bir dezavantaj olarak ortaya çıkmaktadır. Alt kademeye ilişkin herhangi bir kayıtların bulunması için birçok kayıtların incelenmesi gerekmektedir [45,66].

Alt ve üst kademe kayıtlarının bağlantılarının gerçekleştirilebilmesi için çok sayıda yöntem mevcuttur. Her metodun farklı avantaj ve dezavantajları bulunmaktadır. Herhangi bir alt kademe kaydı için üst kademe kayıtları arandığında bağlantılı liste sistemi zorluklar ile karşı karşıya kalmaktadır, bu nedenle temel veritabanı yazılımı altında bazı sistemleri uygulamak için seçim yapılması gerekmektedir. Tüm alt kademe kayıtlardan üst kademeye bağlanmak için kullanılan bağlantı yolları, bir durum için yetersiz ve ilgisiz olabilmektedir. Bu bağlamlar kapsamında hiyerarşik veri modelinin üç dezavantajı bulunmaktadır.

1.Tüm durumlar birden çok (1 to M) veya alt kademe üst kademe formatına uygun olmayabilmektedir.

2.Etkili performans için bağlanan yol seçimleri hem pozitif hem de negatif katkı sağlayabilmektedir.

3.Alt ve üst kayıtların bağlantısı fiziksel olarak gerçekleştirilmektedir. Firma bağlantılı birimlere ilişkin dosyalar tekrar düzenlendiğinde, tüm kayıt noktalarının boşaltılması gerekmektedir [45].

2.1.2. Ağ Modeli

Ağ modeli hiyerarşik modelin geliştirilmiş versiyonu olarak ortaya çıkmıştır. Ağ modeli çoktan çok (M to M) ilişki veya çoktan bire (M to 1) ilişkinin kabul gördüğü modeldir. Hiyerarşik modelde yer alan ilişkilere ek olarak yeni bir katılım söz konusudur. Hiyerarşik model üzerinden bu katılımın uygulanması durumunda yeni katılan yapı ilişkilerini belirtmek için gereksiz çok sayıda dosya oluşumu ortaya çıkacaktır. Her bir yeni katılımda mevcut hiyerarşik modeldeki var olan yapılar için oluşturulan listeler bir veya birkaç kez tekrar oluşturulmak durumunda kalacaktır. Ağ veritabanında çoktan çok ve çoktan bire ilişkilerin sağlanması ile bu gereksiz tekrarlardan kaçınılmaktadır. Modelde bağlantılar kolaylıkla sağlanabilmektedir [45].

Ağ modelinde farklı iki düzenleme aracı söz konusudur. Kayıt aracı ve bağlantı aracı sırasıyla varlık tipi ve ilişki tipini ifade etmektedir. Bağlantılar arasında kısıtlama bulunmamaktadır. Fakat modelde çoklu ilişkilerin kurulmasına izin verilmemektedir [48].

Ağ veri modelinde, hiyerarşik veri modelinde kullanılan kayıtların işaretçi ile bağlanması yöntemi geçerlidir. Ağ modelinde erişim yolları, önce veri yapısının tanımlanması esnasında ortaya konulmaktadır. Ağ modelinde veri sınırlandırması, daha önceden belirlenen yolların kullanılması ile ortaya çıkmaktadır. Bu sayede bazı durumlarda işlemlerin hızlanması yönünde olumlu katkılar sağlanmaktadır [49].

Ağ modelinin avantajlarının yanı sıra hiyerarşik modelde olduğu gibi dezavantajları da bulunmaktadır. Yol seçimlerinin performansla olumsuz etkileri olabilmektedir. Alt ve üst kademe bağlantıların fiziksel olarak yapılması nedeni ile kayıt noktalarının boşaltılması gerekebilmektedir. Veritabanı ağ modelinde daha karmaşık bir yapıya bürünmektedir. Bu nedenle bağlantı yolları ve bakım sorunlarını yönetmek daha zor hale gelmektedir [45].

2.1.3. İlişkisel Model

E. Codd 1970'lerde hiyerarşik ve ağ modelinin içerdiği dezavantajları sunmayan ilişkisel modelde bir veritabanı tanıtmıştır. Codd'un sunduğu bu model, veri dosya yollarını iki boyutlu, sırasız tablolar içerisine düzenli veriler yerleştirir ve bağlantılarını oluşturur. Sayısal sorgular, mantıksal modelin fiziksel gerçekleştirilmesi yerine sadece veri odaklıdır. Codd'un sunduğu ilişkisel model, mantıksal gerçeklerin fiziksel olarak depolanması ile ilgilidir. Aksine veri setleri, iki boyutlu basit ve sıralanmamıştır. Bunların tamamı Codd'un önerdiği normal formlara ulaşmayı amaçlamaktadır [89,67].

İlişkisel veri modelinin sağladığı en büyük olumlu katkılardan biri de tekrarlı verilerin kullanılmayarak esnek bir modelin oluşturulmasıdır. İlişkisel veri modelini esas alan veritabanlarında bazı uygulamalara imkan tanınmaktadır. Bu uygulamalardan biride yapısal sorgulama dili vasıtası ile verilerin sorgulanarak filtrelenebilmesidir. İlişkisel veri modelini kullanan veritabanları, tablolama raporlama gibi pek çok avantajın yanı sıra birçok uygulama programlarını kullanabilmektedir. Öte yandan ilişkisel model çoklu bağlantıya izin vermektedir [49].

2.1.4. Nesneye Yönelik Veri Modeli

Nesneye yönelik veri modeli, nesne tabanlı programlama dilleri ile elde edilen program ve veritabanlarının temelini oluşturan veri modelidir. Programlama mantığına dayanan nesneye yönelik veri modeli kompleks yapılarda başarılı sonuçlar elde etmektedir. Nesneye yönelik veri modeli uygulamada istatistiksel tekniklere sıklıkla başvurmaktadır.

Nesneye yönelik veri modelinde varlıklar, nesne olarak kabul edilmektedir. Nesnelere bağlı tanımlayıcılar bulunmaktadır. Bu veri modelinde alt sınıf ve üst sınıf yapısı bulunmaktadır. Mevcut sınıf, yeni bir sınıf üretebilir ve üretilen bu sınıf kalıtsal olarak üst sınıfın özelliklerini taşıyabilmektedir. Üretilen bu sınıf, alt sınıf olarak nitelendirilmektedir. Kalıtsal özelliklerin taşınması sayesinde model içerisinde veri tekrarı gerçekleşmeyerek performans ve zaman kaybı yaşanmamaktadır [50].

2.2. Bilgi Modelleme

Günümüz dünyasının gerektirdikleri dikkate alınarak bir yaşam düzeni ortaya konacak ise temel gereksinimlerden biride şüphesiz ki bilgiye olan ihtiyaçtır. Birey, hayatta başarılı bir yaşam sürebilmesi için beslenmeye, uyumaya, dinlenmeye ihtiyaç duyuyor ise hayatta mücadele edebilmesi içinde bilgiye ihtiyaç duymaktadır. 21. yüzyılda internet ve bilgisayarlar bilginin önemini artıran temel unsurlar olmuştur. Eski teknikler ile bilginin ulaşılması, gelişmesi, yorumlanması iletişim kanallarının eksikliğinden bilginin yavaş ilerlemesine neden olmuş buda yeni bilginin oluşması ve enformasyonun işlenmesini bir o kadar zor kılmıştır. Bilgisayarların gelişmesinin yanı sıra internet ile birlikte bilgi transferi ve bilgi akışı katlanarak artmıştır. Günümüzde yapılan her işlemde bilgi kullanılmaktadır. Bilgi olmadan sektörlerin gelişmesi ve iş yaşamının başarılı bir şekilde ilerlemesinden bahsedilmesi mümkün değildir.

Bilginin bu denli önemli olduğu günümüzde bilginin eksikliği veya yanlış bilginin oluşturduğu olumsuz sonuçlar telafi edilmesi mümkün olmayan büyük kayıplara neden olabilmektedir. Yanlış veya eksik bilgi, bir insanın ölümüne neden olabileceği gibi mali değeri milyarlarca lira olan bir projenin heba edilmesine kadar birçok olumsuz sonuç doğurabilmektedir [46].

Bilginin yanlış olarak yorumlanması, eksik veya hatalı bilginin oluşması her zaman insan kaynaklı olmamaktadır. Bu hataların kaynağı bazen bilgisayarlar olabilmektedir.

Bilgisayarlar program ve özel yazılımlarda donatılmış makinalardır. Bilgisayarlar hissetme veya algılama gibi yetenekleri olmadığından bu eksikliklerin giderilmesi için yazılımlar vasıtası ile gerçek hayattaki olguları modelleyerek bilgisayarlara bazı özellikler kazandırılmak istenmektedir. Oluşturulacak veritabanı ve modelleme teknikleri ile bilgisayarlar bilgi üretimi yapabilecek özelliklere sahip olabilmektedir. Bu nedenle bilgi sentezi ve bilgi modelleme işlemleri için veritabanı mantığının algılanması ve modelleme yaklaşımlarının doğru kullanımı önem taşımaktadır [46].

Veritabanı kullanılarak bilginin modellenmesinde dikkat edilmesi gereken önemli hususlar bulunmaktadır. İleri düzeydeki algılama yaklaşımı, doğada gerçeğe odaklanmanın özünü oluşturmaktadır. Nesne tabanlı modelleme veya geleneksel yaklaşımlar kullanılarak veritabanı modellemesinde gerçeğe uyumlu nesnelerin bulunması, modellemenin başarısını daha da artırmaktadır. Veritabanını modelleyenlerin programcı mantığı ile yaklaşması, anlamsal yaklaşım ile birlikte veritabanını tasarlamada güçlü ve doğal bir yöntem sunmaktadır [46,70].

Veritabanı esasen ilişkili verilerin toplamından oluşmaktadır. Veritabanı bireyler tarafından ilişkili bilgiler kümesi olarak algılanmaktadır. İşletme alanlarının altında yatan bilgi türlerinin keşfedilmesi ve bilgilere kuralların uygulanması ilgi çekici ve açıklayıcıdır. Veritabanı tasarımının kalitesi açısından bilgilerin ve kuralların kullanılması kritik önem taşımaktadır. İyi bir mimari ve tasarıma sahip veritabanı, bilgi sağlamada ve doğru sonuca erişimde büyük kolaylıklar sağlamaktadır. Veritabanının tasarımına yöneltilen problem dönüşleri ve sürecin olumsuz şekilde ilerlemesi, veritabanının kötü tasarlanmasından kaynaklanmaktadır [46].

İşletme alanlarındaki her bir veritabanı modeli, bilgi dünyasının belirli bir kısmını oluşturmaktadır. Kütüphanelerde kullanılan veritabanları buna örnek teşkil etmektedir. Kütüphanedeki mevcut değişiklikler veritabanına yansıtılarak veritabanı güncel tutulmaktadır. Güncelleme işlemi manuel veya otomatik olarak gerçekleştirilebilmektedir. Veritabanının otomatik veya manuel olarak kullanılması, programlama dilleri ile tasarlanmış özel amaçlı yazılımlar ile denetlenmesi mümkündür [46,71].

Veritabanı uygulamaları, veritabanı yönetim sistemleri kullanılarak geliştirilmektedir. Veritabanı yönetim sistemleri, veritabanını çalıştırmak ve sorguları yanıtlamak için geliştirilen yazılım sistemleridir. Bir veritabanı yönetim sistemi birden fazla veritabanında çalışabilmektedir [46].

Tipik uygulamalar, sürekli verileri depolamak için geçici verileri veritabanlarında tutmada bellek nesne modelini kullanmaktadır. Kullanıcı dostu ara yüze sahip model, veritabanına giriş ve erişim açısından daha kullanışlıdır. Alanların altında yatan anlamları açıkça ortaya çıkaran bilgi modelleri, çıkarım yapan bilgiler üzerine çalışmaktadır. Bazı araçlar ise sadece otomatik çalışan veritabanı üretmek için bilgi modellerini kullanmanın yanı sıra nesne modeli ile kullanıcı ara yüzü elde etmek içinde kullanılabilmektedir [46].

Uygulamalarda veri erişimi ve veritabanı bakımının gerektiği durumlar için veritabanı yönetim sistemi, manuel kayıt tutma üzerine birçok avantaj sunabilmektedir. Veriler uygun şekilde elektronik ara yüzler aracılığı ile tutularak hızlı bir şekilde işlendikten sonra kompakt disklerle depolanabilmektedir. Bazı veri hataları, otomatik olarak fark edilerek sistem tarafından erişim izinlerinin kısıtlanması mümkündür. Veritabanlarının diğer bir avantajı, bireylerin bilgisayarlarda yapması gereken rutin işlerden kurtararak, daha fazla yaratıcı tasarımlara zaman ayırmasına imkan tanımasıdır. Yazılım mühendislerinin geliştirdiği uygulama yazılımları veritabanı kullanımında kolaylık sağlayabilmektedir [46,70].

Çalışan gruplar açısından 19.yy sonlarına doğru tarım çağının yerini alan endüstri çağı günümüzde yerini bilgi çağına bırakmaktadır. Endüstriyel mekanizasyon ve sürekli büyüyen bilgi birikimi, bilgiye dayalı sektörlerde çalışan personel sayısında ciddi artışa yol açmıştır. İşletmeler günümüzde bilgi teknolojilerinden yararlanarak başarı ve kazanımlarını önemli derecede artırmaktadır. Bilgi sistemleri, sadece istihdama katkısının yanı sıra, bilgi içeriklerinden destek alarak işletmenin verimli etkileşim yeteneği kazanmasına olanak tanımaktadır [46].

Pek çok çalışan açısından bilgi teknolojilerine olan ihtiyacın bilinmesine karşın, çalışanlara göre bilgi yönetimindeki görevlerin karmaşıklığındaki bilgi boyutları bakımından farklılıklar söz konusu olmaktadır. İlk zamanlar pek çok bilgisayarın teknik işlemleri, bilgisayar programcısı ve sistem analisti olarak bilgisayar uzmanları tarafından gerçekleştirilmiştir. Piyasaya kullanıcıların kolay kavrayabileceği yazılımların çıkışı ve bu yazılımların ilgi görmesi ile bireysel bilgisayarlar üretilmiştir. Böylelikle işlem gücü yeni bir dağılım göstermiştir. Bu kullanımın yaygınlaşması ile beraber elektronik tabloların hazırlanması, buna benzer profesyonellik gerektiren uygulamaların kullanımı ve bilgi yönetim görevlerinin kullanımı yaygın biçimde gerçekleştirilmiştir [46].

Teknoloji ve bilgisayar alanındaki gelişmelerin hızla ilerlemesi bilgisayar alanındaki uzmanlara olan ihtiyacı artırmıştır. C# ve java gibi programlama dillerini bilen

programcılara olan ihtiyaç artmıştır. Bunun yanı sıra kompleks bilgi sistemi modelleyebilen ileri derecede kabiliyeti olan bireylere olan talepte artmıştır [46].

Bilgi sistemleri mühendisliği, gereksinim analizi, veritabanı tasarımı, kullanıcı ara yüzü ve rapor yazımı gibi alt disiplinleri içermektedir. Veritabanı tasarımı aşamasında ilgili bilgileri saklamak için yapının temelini oluşturan konu seçimi büyük önem taşımaktadır. Veritabanında tutulan bilgiler tablolar halinde sunulmaktadır [46].

Tablo 2.1. Veritabanında bulunan tablolara ilişkin örnek [46].

Alan 1	Alan 2	.	Alan n
Veri 1	.	.	.
Veri 2	.	.	.
.	.	.	.
.	.	.	.
.	.	.	.
Veri m.	.	.	.

Tablo 2.1.'de belirtilen şema boş bir veritabanından elde edilen sonuç tablosudur. Sütunlarda alanlar ve sütunlardaki her hücrede alana ilişkin veriler tutulmaktadır. Veritabanından elde edilen tablolar belirli sorgular sonucu elde edilmektedir. Bazı hücrelerde aynı veriler bulunabilmektedir. Bu özelliklere birbirinden ayırt edilebilmesi için her özelliğe benzersiz bir numara atanmaktadır. Tablolar, verilerin daha açık ve pratik şekilde yorumlanmasını sağlamaktadır. Verilerin veritabanında depolanma biçimleri ve yöntemleri birbirlerinden farklılık gösterebilmektedir. Bazı veritabanları bir hücrede birden fazla veri tutulabiliyor iken, ilişkisel veritabanında her hücrede sadece bir veri tutulmaktadır. Veritabanının oluşturulma ve veri girişi aşamasında dikkat edilmesi gereken önemli hususlar bulunmaktadır. Sütunlarda boş hücre bırakılmaması ve her veritabanının isimlendirilmesi bu hususlardan bazılarıdır [46].

Veritabanı tasarımının bilimsel yönünün yanı sıra sanatsal yönü de bulunmaktadır. İyi bir metot ile desteklenen tasarım süreci, entelektüel, tatmin edici ve kullanıcılarına yarar sağlayan kaliteli bir veritabanı oluşmasına imkan tanıyan bir yapıyı ortaya çıkarmaktadır [46,68,69].

2.3. Bilgi Modelleme Yaklaşımları

Veritabanı tasarımı oluşturulurken ilk olarak dikkat edilmesi gereken nokta, veritabanının modelinin ortaya konulmasıdır. Belirli bir işletme alanı için veritabanı tasarlandığında, ilk olarak onun modeli ortaya konulmalıdır. İşletme alanı söylem evreni olarak adlandırılmakta ve modellenmektedir. Söylem evreni veya işletme alanı bilgi dünyasının tipik bir kısmını oluşturmaktadır. İyi bir modelin geliştirilmesi, modellenecek dünyanın iyi bir şekilde anlaşılmasına bağlıdır. Modelin geliştirilmesindeki esas zorluk, söylem evreninin açıkça ve tamamen tanımlanmasında yaşanmaktadır [46].

Yazılım geliştirme aşamasından bir sonraki aşamaya kadar yapılan filtreleme boyunca hatalı girişlere karşı dikkat edilmelidir. Hataların daha sonra fark edilmesine karşı ileri aşamalarda bu hataların ortadan kaldırılması pahalıya mal olabilmektedir [46].

Söylem evrenini modelleyen kişiler modelleyici olarak adlandırılmaktadır. Modelleme yapılacak alan iyi bir şekilde tanımlandığında bireysel modelleme yapılabilmektedir. Bilinen bir alanda modelleme yapılmayacağı durumlarda, ilgili alanda bilgi birikimi yüksek bireylerden destek alınması gerekmektedir. Bu bireyler alan uzmanı veya öznel konu uzmanı olarak adlandırılır. Modelleme, modelleyici ile alan uzmanı arasındaki ortaklaşa gerçekleştirilen bir aktivitedir [46].

İnsanların doğal olarak kelimelerle, resimlerle ve örnekler ile iletişim kurmasından dolayı, söylem evreni hakkında net bir tanımlama yapılmasının en iyi yöntemi, sezgisel diyagramlar, doğal dil ve örneklerin kullanımınıdır. Modelleme işlemlerinin kolaylaştırılması için mümkün olan en küçük birimdeki bilgiler incelenmektedir. Model, ilk olarak kavramsal seviyede ifade edilmiştir. Bu durumda model, doğal dillerden kaldırılarak kavramsal olarak belirtilmektedir. İlişkisel veritabanı yapıları mantıksal veri modeli seviyesindedir. Veritabanı yönetim sistemi diğer mantıksal veri modellerinden en az biri ile aynı seviyededir. Spesifik olarak kavramsal tasarım taslağı bulunmaktadır. Modelleyici bu aşamada uygulama endişelerinden bağımsız olarak çalışmalıdır. Söylem evreninin doğru bir şekilde geliştirilmesi modelleyici açısından zorluk içeren bir işlemdir [46].

Başlangıçta kavramsal tasarım oluşturulduktan sonra, herhangi bir veri modeli, mantıksal tasarım için haritalandırılabilir. Bu esneklik, veritabanı yönetim sistemi üzerindeki birden fazla türdeki aynı uygulamaların daha kolay işleme alınmasını ve korunmasını sağlamaktadır. Pek çok uygulama verilerin yanı sıra süreçleri de içermesine rağmen uygulamalarda veriler üzerinde odaklanılmaktadır. Bu nedenle süreçler verilere

bağlı olarak gelişmektedir. Bilgi modelleme konusunda genelde üç yaklaşım üzerinde durulmaktadır. Bu yaklaşımlar, varlık ilişki modelleme, bilgi tabanlı modelleme ve nesne tabanlı modelleme olarak ifade edilmektedir. Herhangi bir modelleme yaklaşımı, model inşa etme amacı ile notasyonların kullanımında notasyonların yanı sıra prosedürleri de içermektedir. Bilimsel yöntemlerdeki veri modellerinin temelini oluşturmak ve sistemi yönetmek için veri türlerine ihtiyaç bulunmaktadır. Sistem içerisinde kullanılan veri olguları olmasına bağlı olarak olgularda kullanılan veri örnekleri sistem içerisinde aranmaktadır. Bu örnekler bazı yöntemlerde bilgi sunan, çıktı raporları, giriş ekranları, formlar olabilmektedir. Bazı örnekler manuel olarak yüklenmiş olabilir veya bilgisayar kayıtları olarak önceden var olabilir. Bazı zamanlarda uygulamalar yeni yöntemler içerebilmektedir, bu nedenle çözüm geliştirilmesi ve uyarlanması gerekebilmektedir. Modelleyici ile alan uzmanları birlikte uygulama alanına ilişkin yeni örnekler oluşturabilmektedir [46,73].

2.3.1. Varlık İlişki Modelleme

Varlık ilişki modeli 1976 yılında Peter Chen tarafından geliştirilmiştir. Halen en yaygın biçimde kullanılan veri modelleme yaklaşımı olarak bilinmektedir. Varlıklar açısından dünya resimlendiğinde ilişkide katılımcılar ve özelliklerin olduğu gözlemlenmektedir. Zamanla varlık ilişki diyagramının bazı versiyonları ortaya çıkmıştır. Varlık ilişki modeli tek değildir fakat standart bir varlık ilişki notasyonu bulunmaktadır. Varlık ilişki diyagramının farklı versiyonları farklı kavramları destekleyebilir ve farklı semboller aynı kavramlar için kullanılabilir. Şekil 2.1.'de Oracle tarafından geliştirilen Case araçları tarafından desteklenen yaygın varlık ilişki diyagramları temsil edilmektedir [46].



Şekil 2.1. Varlık ilişki model diyagramına ilişkin örnek [46].

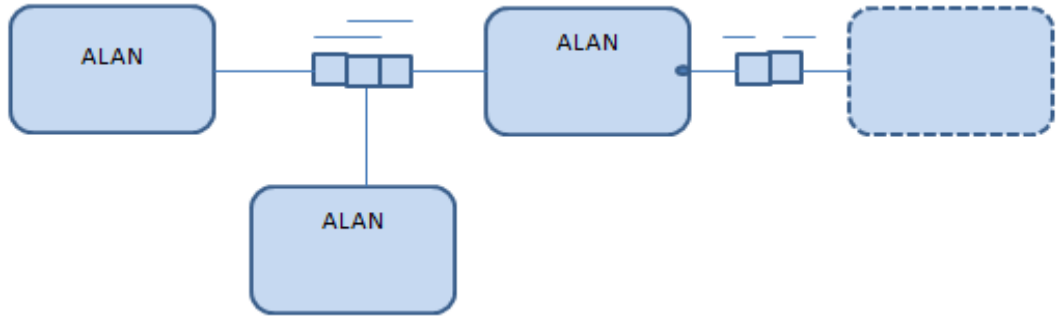
Şekil 2.1.'de varlık ilişki diyagramında kullanılan varlıklar arasındaki ilişki ve bağlantı yapısına örnek verilmiştir. # (Ototrop) varlık türünün ilk tanımlayıcılarının bileşen özelliğini göstermektedir. *(yıldız) özelliğin zorunlu olduğunu göstermektedir. ... (üç nokta) diğer özellikleri göstermektedir. İlişkiler, varlık türleri ve isimler arasında çizgiler halinde tanımlanmaktadır. Yalnızca ikili ilişkilere izin verilmektedir. İlişkinin bir yarısı zorunlu (sürekli) çizgileri diğer yarısı ise opsiyonel (kesikli) çizgileri göstermektedir. İlişki, varlık türü için ilk tanımlayıcının bileşenleridir. Varlık ilişki diyagramındaki ördek ayağı olarak bilinen ilişkinin bir ucu varlık türlerinin bazı örneklerinin ilişkinin diğer ucundaki aynı varlık türleri ile bağlantılı olabileceğini göstermektedir. Ördek ayağının birinin eksik olması, bir ucundaki en az bir varlık örneğinin diğer uçtaki herhangi bir varlık örneği ile bağlantılı olduğunu göstermektedir.

Varlık ilişki diyagramı, hedef yazılım platformunun bağımsız olması için yoldaki alanları tasvir etmektedir. Deneyimli bir varlık ilişki modelleyici, modellemeyi gerçekleştirebilmek için gerekenleri zamanında fark ederek varlık türlerini belirler. Bu işlemlerin gerçekleştirilmesi, tecrübesiz bir modelleyici için zor olabilmektedir [46].

2.3.2. Nesne-Rol Modelleme (ORM)

Nesne-rol modeli 1970'lerin ilk yıllarında anlamsal modelleme yaklaşımında nesneler açısından dünyaya bakış açısında rol oynamaktadır. Nesne-rol modeli, doğal dil, bilgi analizi metodu gibi çeşitli formlardan ortaya çıkmıştır [46].

Alan uzmanları verilerin nasıl kullanıldığına bakılmaksızın, doğal dil cümlelerindeki bilgi içeriğini, anlamı ile ifade etme konusunda yeteneklidirler. Alan uzmanları tarafından anlaşılır hale getirilen informal ifadeleri formal ifadelere dönüştürme, modelleyicinin sorumluluğundadır. Bu ifadeler alan uzmanı ve modelleyici tarafından dönüştürülmektedir. Bu adımlar nesne-rol modellemenin kavramsal analiz prosedürü adımlarını içermektedir. Bu aşamalarda veri örnekleri, bilgi örnekleri olarak ifade edilmekte ve bilgi türleri özetlenmektedir [46,71].



Şekil 2.2. Nesne-rol model diyagramına ilişkin örnek [46].

Modelleyici olarak tanımlayıcılar tarafından her nesnenin tanımlanması iki temel cümlelerin yeniden ifade edilmesi ile gerçekleşmektedir. Alan uzmanının bu ifadeyi kabul etmesi nedeniyle bilgi örneklerinden bilgi türlerine özetleme gerçekleştirilmiş olmaktadır. Nesne rol modeline ilişkin nesne, alan ve rol bağlantılarının belirtildiği diyagram örneği Şekil 2.2.'de belirtilmiştir [46,72].

Nesne-rol modelinde bilgi türleri, çözümün önemli bir parçasını oluşturmaktadır. İlişki, örnek rol oynayan nesne türünün her bağlantısında bir veya daha fazla rol kutusunun isim zinciri halinde gösterilerek belirtilmektedir. Varlık ilişki diyagramlarının aksine, nesne-rol modeli temel modellerinde, özelliklerin kullanımına izin verilmemektedir. Varlıklar ve değerlerinin oynadığı rol açısından tüm bilgiler temsil edilmektedir. Bu durum geniş diyagramlar oluşmasına neden olmasına rağmen, özelliklerin kullanılmadığı yaklaşım, kavramsal analiz için, sadelik, kolay doğrulama ve istikrar sağlama yönünden avantaj oluşturmaktadır. Nesne-rol modeli herhangi bir rol numarasının ilişkisine izin vermektedir [46].

Her bir bilgi türü, bilgi tablosunda örnek bir popülasyon ile birlikte kısıtlamaları onaylamak için eklenebilmektedir. Bilgi tablosundaki her bir sütun bir rol ile bağlantılıdır. Şekil 2.2.'deki diyagramda bulunan rol kutuları yanındaki çizgiler benzersiz içsel kısıtlamaları temsil etmektedir. Rol ve rol kombinasyonlarını gösteren benzersiz girişler olmalıdır. Nesne-rol modeli şemaları, diyagramlar ve metinler ile temsil edilmektedir. Bazı nesne rol modeli araçları iki temsil arasında otomatik dönüşüm gerçekleştirebilmektedir. Modeller iki temel yöntem ile alan uzmanları tarafından doğrulanabilmektedir [46,71].

Somut örnekler, alan uzmanlarına kuralların gerçekliği hakkında karar vermede yardımcı olmaktadır. Bu doğrulama adımı, mantıksal terimlerin yanlış kullanımından kaynaklanan sorunlarda, dili kontrol eden alan uzmanları açısından yararlı olmaktadır.

Varlık ilişki diyagramları ile nesne rol modellerine ilişki diyagramları karşılaştırıldığında kompakt görüntüleme açısından kıyaslama yapıldığında, varlık ilişki diyagramlarının daha anlaşılır olduğu söylenebilmektedir. Varlık ilişki diyagramlarının doğal dilden çıkarım yapmasına rağmen alan uzmanları tarafından çıkarımın kavramsallaştırılması daha karmaşık bir işlemdir. Varlık ilişki diyagramı notasyonu, kuralları ve kısıtlamaları belirleme yönü ile nesne rol modeli açısından daha az açıklayıcılığa sahiptir. Varlık ilişki diyagramlarında kavramsal analiz aşamasında ilişkinin önemi hakkındaki kararlar teşvik edilmektedir. Bazı durumlarda bu durum avantaj olarak görülebilmektedir. Bazı durumlarda ise ilişki üzerindeki ayrımlar önemli dezavantajlar oluşturabilmektedir. Tercih seçimleri diğer bilgi türlerine bağlı olmaktadır. Bilgisizlik altında karar alabilmek oldukça zor bir durumdur. Diğer bilgilerin kayıt altına alınması zorunludur. Aksi durumda modelin bu kısmının değişmeye ihtiyacı olacaktır.

Özelliklerin nitelik olarak modellenerek ilgili kayıtların daha sonra ortaya çıkarılacağı durumlarda varlık tiplerini ve ilişkilerini yeniden modellemek zor olacaktır. Bunun nedeni, özelliklerin ilişkilere katılamaması olarak gösterilmektedir. İleriye yönelik bilgiyi öğrenmenin çok zor olmasından dolayı özellik tabanlı model istikrarsız olacaktır. Model özellikleri değiştiğinde, modeli kullanan uygulamaların tekrar kodlamaya ihtiyacı bulunmaktadır. Nesne-rol modelinin esasında değişikliklere karşı bağımsızlığı bulunması nedeniyle anlamsal olarak istikrar sunmaktadır [46].

Nesne-rol modelleri, ifade etmeyi kolaylaştırmaktadır. Özellikler, örnek veri popülasyonunu kullanarak biçimsizliği ortadan kaldırmaktadır. Opsiyonel özellikler ile eksiklikleri doldurularak karmaşıklığı önlemeye yönelik sıfır değerleri tanıtmıştır. Bu hususlar dikkate alındığında nesne-rol tabanlı modelleme yaklaşımı en az varlık ilişki modelleme yaklaşımı kadar kavramsal analiz açısından avantaj sağlamaktadır. Nesne rol modelleme genelde, kavramsal aşamanın başlangıç kısmını oluştururken varlık ilişki modeli tasarım görünümünün elde edilmesine fırsat sağlamaktadır. Model için varlık ilişki modeli kullanılmaya karar verildiğinde, nesne rol modeline ilişkin notasyonların kullanımı gerekmemektedir. Modelleme aşamasında uyarlama ve uygulama için nesne rol modelinin kavramsal şeması, varlık ilişki notasyonlarına uyarlanabilmektedir [46,71].

2.3.3. Birleşik Modelleme Dili (UML)

Nesne tabanlı program kodu tasarımında kullanılan bazı modeller veritabanı tasarımı içinde kullanılmaktadır. Pek çok nesne tabanlı yaklaşım bulunmaktadır. Bu yaklaşımlardan en çok kullanılanlardan biride nesne yönetim grubu tarafından adapte edilen birleşik modelleme dilidir. Birleşik modelleme dili birçok diyagram yapısı arasında, statik veri yapısını belirtmek için sınıf diyagramlarını içermektedir. Modelde sınıf diyagramları, operasyonları belirtmenin yanı sıra nesne tabanlı kodları belirtmek için düşük seviyeli tasarım kararlarında kullanılabilir. Uygulama detaylarından arındırıldığı durumda birleşik modelleme dili sınıf diyagramları, varlık ilişki modelinin genişletilmiş versiyonu olarak ilişkilendirilebilmektedir. Mantıksal problemlerin çözümünde varlık ilişki diyagramı başarılı olabilmektedir. Birleşik modelleme dilinin nesne odaklı olması nedeniyle sınıflar için kavramsal tanımlayıcı şema gerektirmemektedir. Bunun yerine varlık örneklerinin, iç nesne tanımlayıcılar tarafından tanımlanacağı varsayılmaktadır [46,71].

Sistem geliştirme süreci, programın işleyişi ve verimliliği açısından çözüm aşaması ve problemin belirlenmesini gerekli kılmaktadır. Sistem geliştirme ve modelleme aşamasında kullanılan birleşik modelleme dili nesneye yönelik bir modelleme dilidir. Birleşik modelleme dili, sistem tasarımı ve sistem incelemelerinde farklı modellerin kullanımına imkan tanımaktadır. Bu modeller;

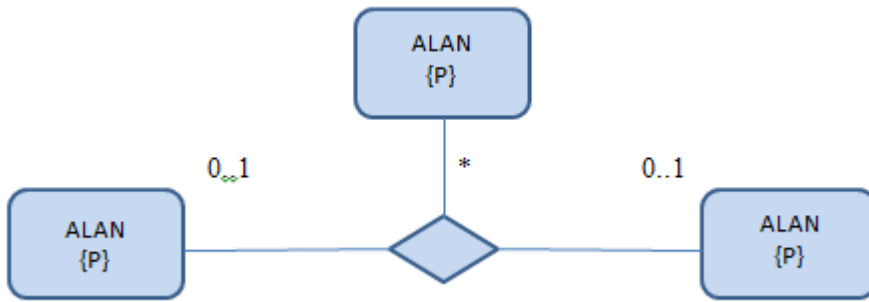
- Kullanıcı senaryoları modeli
- Gerçekleştirim modeli
- Etkileşim modeli
- Yaygınlaştırma modeli
- Durum modeli
- Sınıf modeli

olarak belirtilmektedir [50].

Birleşik modelleme dilinde nesne üzerinde yapılan işlemlere göre ara birim ve çok biçimli olarak ifade edilen yapılar söz konusudur. Arabirim, nesneye ilişkin gerçekleştirilmesi muhtemel işlemleri ifade etmekte iken çok biçimlilik ile farklı nesnelerin ortak bir arabirimden farklı işlemlerin gerçekleştirilmesini ifade eden sınıflar belirtilmektedir [51].

Birleşik modelleme dili yapısı itibarı ile bir programlama dili gibi görünse de esasen ilişki modelleme dilidir. Modelleme dili gerçekleştirilirken pek çok diyagram oluşturulmaktadır. Bu özelliğinin, birleşik modelleme dili kullanımı açısından pek çok olumlu katkısı bulunmaktadır. Tasarım ve süreç aşamasının kolaylaştırılması, oluşabilecek hataların asgari düzeye indirgenmesi, kullanılan belleğin azaltılması, maliyetlerin düşürülmesi bu katkılardan bazılarıdır [51,52].

Birleşik modelleme diline ait sınıflar için benzersiz olabilecek özellik değerlerini belirtmede kullanılan standart notasyonları bulunmamaktadır. Buna rağmen birleşik modelleme dili, herhangi bir dilde parantez veya dipnotlarda kullanıcı tanımlı kısıtlama eklemeye izin vermektedir. Bu sembollerin standart ve taşınabilir olmamasından dolayı ilk benzersizlik için {p} alternatif benzersizlik için {u1} eklenmektedir. Özellikler varsayılanlar tarafından zorunlu kılınmaktadır. Birleşik modelleme dilinde bağlantı rolleri sıralanmadığı için kullanılamamaktadır. Aynı sınıf içerisinde birden fazla rol oynadığında durum daha da karmaşıklaşmaktadır. Birleşik modelleme dili isim vermek için bağlantı rollerine gereksinim duymaktadır. Fakat rol isimleri doğal dilde her zaman sıralanan cümleyi düzenlememektedir. Birleşik modelleme dilini temsil eden bir diyagram örneği Şekil 2.3.'te verilmiştir. Şekildeki diyagram örneğinde alanlar ve alanlar arasındaki ilişki yapısı belirtilmiştir. Birleşik modelleme dili türetim kuralları ve kısıtlama ifadesinde bilgilerin belirtilmesi bakımından zayıf konumdadır [46,72].

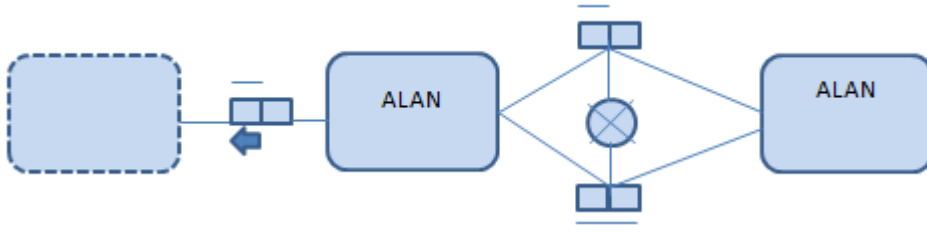


Şekil 2.3. Birleşik modelleme dili diyagramına ilişkin örnek [46].

Birleşik modelleme dilinin tanımlaması, kuralların düzenli bir ifadesi için nesne kısıtlama diline gereksinim duyulmaktadır. Nesne kısıtlama dili, teknik bilgisi olmayan alan uzmanları tarafından onaylanması için kullanılan basit matematiksel yapıdır. Genelde yüksek dereceli diller, nesne modelleme dili için otomatik olarak nesne kısıtlama diline dönüştürüldükten sonra dizayn edilmektedir. Nesne modelleme dilindeki ifadenin yetersiz

olmasından dolayı örnek popölasyon ile doğrulama denenmektedir. Özellik tabanlı notasyonlar çoklu örnekler için yarar sağlamamaktadır. Esas olarak popölasyondaki sıfırları belirtmektedirler [46].

Nesne modelleme dili, tekli sınıf örneği özellikleri hakkında konuşma yetkisi veren nesne diyagramları sağlamaktadır. Fakat tekli sınıf örneği, çoklu sınıf örnekleme yardımcı olmamaktadır. Temelde nesne modelleme dilindeki ikili bağlantıların örneğini oluşturmak için bilgi tablosu ortaya konulmaktadır. Fakat model, ikili olmayan bağlantı örnekleri için uygulanmamaktadır. İkili bağlantı olmadıkça okuma yönü belirlenmemektedir. Nesne-rol modelinde olduğu gibi bilgi satırları ve bağlantı rolleri arasındaki bağıntı yeterince açık değildir. Her rolün adlandırılması ve rol isimlerinin bilgi tablo başlıkları olarak kullanılması önemli bir çözümdür. Bağlantı rollerinin doğrusal olmayan düzeni nedeni ile bilgi satırlarının görsel bağlantıları sınıf diyagramının zayıf olmasına neden olmaktadır. Nesne modelleme dili, varlık ilişki modeli ve nesne rol modelinden farklı olarak uygulama tasarımının değer yönlerini belirtmek için daha fazla yeteneğe sahiptir. İşletme sürecini belirtmek için farklı diyagramlama tekniklerini içermektedir. İşletme alanları ihtiyaçlarının tüm özellikleri diyagramın dinamik yönüne hitap etmektedir. Uygulamada nesne tabanlı kodlar uygulanacak ise, birleşik modelleme dili, programlama kodu yapısını belirlemek için daha kesin tanımlama yetkisi sunmaktadır. Veriler üzerindeki işletme kısıtlarının ifade kapasitesi varlık ilişki modeli ve birleşik modelleme diline kıyas ile nesne rol modelinin notasyonları anlamlı derecede daha zengindir. Nesne rol modeli değişikliklere karşı daha etkili ve daha dik bir duruş sergilemektedir. Chen'in orjinal varlık ilişki notasyonunun aksine birleşik modelleme dili, çizgiler aracılığı ile tipik olarak çizimlenmektedir. Temelde birleşik modelleme dili ikili bağlantılar için alternatif notasyon kullanımına izin vermemektedir, fakat uygulamada bu durum nadiren de olsa görülebilmektedir. Her rolün bir tablo sütunu ile bağlantılı olduğu, bir veya birden fazla kural zinciri olarak nesne-rol modelinin ilişki tasviri hem ifade hem de örnek popölasyonu tarafından doğrulanan ve kolaylaştıran düzgün bir genel notasyon sağlamaktadır [46,74].



Şekil 2.4. Dışsal kısıtlama için karşı örnekleme şeması [46].

Şekil 2.4.'te örnek bir şema verilmiştir. Şemada bulunan yön işareti işlem yönünü belirtmektedir. Birleşik modelleme dilinden kaynaklanan kural eksikliği, yönetme ve yorumlama bağlantılarını kapsayan eş-roller arasındaki kısıtlama, X döngüsü ile birlikte grafiksel olarak tutulmaktadır. Bu durum dışsal kısıtlama olarak ifade edilmektedir. Alan uzmanı ile birlikte kuralları doğrulamak için karşı örnekler sağlanmalı ve kurallar ifade edilmelidir [46].

Alan uzmanlarının farklı etmenler ile çalıştığı ve başarılı olduğu hususlar vardır. Fakat tüm alan uzmanları somut örnekler ile çalışmada başarılı olamamaktadırlar. Diyagramları görmek, alan uzmanı açısından gerekli olmamasına rağmen, diyagram üzerinde herhangi bir rolü doğrudan örneklendirmek, modelleyici olarak kurallar hakkında açıkça düşünmek için kolaylık sağlamaktadır [46].

Birleşik modelleme dili, genel dışsal kısıtlama için grafik notasyonuna sahip olmamasına rağmen, uygun model elemanlarına eklenilen notlarda doküman kısıtlamasına izin vermektedir. Kavram modellemenin bir parçası gibi düşünüldüğünde, modeli düşünmek daha kolay olmaktadır. Dışsal kısıtlama notasyonunun birleşik modelleme dilinde geliştirilememesinden dolayı, modelin gelişiminde kısıtlama kaybı gerçekleşmektedir. Aynı ilişki, varlık ilişki diyagramı içinde geçerli olmaktadır. Nesne-rol modeline ilişkin modelleme prosedürleri, kısıtlamalar ile ilgili düşünceleri harekete geçirerek düzenli kurallar içerisinde görselleştirmek ve tutmak için izin vermektedir [46,71,72].

2.3.4. Bilgi Modelleme Yaklaşımlarının Farklılıkları

Varlık ilişki modeli, nesne-rol modeli ve birleşik modelleme dili, bilgi modelleme ve veritabanı tasarımı konusunda pek çok ortak özelliğe sahip olduğu gibi birbirlerinden oldukça farklı yönleri de bulunmaktadır. Bu farklılıklar veritabanı tasarımında avantaj sağladığı gibi daha etkin bir modelin kullanımında başarı sağlayabilmektedir [46].

Kavramsal nesne tipleri veya anlamsal alanlar, kapsamlı bir resimde uygulama modelindeki çeşitli bileşenleri birbirine bağlayan kavramsal bir bağlayıcı görevini üstlenmektedir. İlişkisel veri modelinin seviyesi düşük olsa dahi, ilişkisel modelin kurucusu E.E. Codd alanların ilişkisel veritabanını bir arada tutan model olduğunu öne sürmektedir. Nesne rol modelindeki diyagramlarda bulunan nesne türleri, anlamsal alanlardır. Bu nedenle modelin bağlılığı transparandır. Nesne-rol modelinin bu özelliği kavramsal sorgular için önemli avantajlara sahiptir. Kullanıcılar, yönlendirme boyunca nesne türlerinden kurulu ilgili bağlantılara, doğrudan kavramsal modeli sorgulayabilmektedir [46,67].

Varlık ilişki modeli ve birleşik modelleme diline ilişkin diyagramlar ilgili kısıtlamaları veya özellikleri ifade etmede başarısız kalmaktadır. Birçok varlık ilişki modelinin aksine, birleşik modelleme dili sadece bağlantılar arasında olmak üzere XOR kısıtlaması sağlamaktadır [46].

Nesne-rol modeli ile birleşik modelleme diline ilişkin diyagramlarda anlamsal alanların yazılışları farklılık göstermektedir. Birleşik modelleme dilinde alan adları özelliklere eklenerek açık konuma getirilmektedir. Modelleme sürecindeki farklı aşamalarda ilgili bilgiyi tek bir yerde görmek, modelleyici açısından fayda sağlamaktadır. Diğer nesne-rol modeli özellikleri alt tipler için esnek biçimde destek olmaktadır. Formal alt tip, tanımlamalara dayalı çoklu kalıtım içermektedir [46].

Genelde, sınırsız sayıda kısıtlama türü bulunması nedeniyle metinsel kısıtlama dili, diyagram gereksiniminde tamlık için gerekli olabilmektedir. Bu, varlık ilişki diyagramları, nesne-rol modeli ve birleşik modelleme dili için geçerli olmaktadır. Varlık ilişki diyagramları ve nesne rol modeli diyagramları, model üzerinde dönüşüm gerçekleştirmek veya kapsamlı bir model geliştirmek için kullanılan bazı önemli nesne rol modeli kısıtlamalarında standart notasyonlar içermektedir [46].

Alternatif temsiller arasından belirlenen bir temsildeki kısıtlamaları sağlamak için nesne rol modelinde ciddi düzeyde kontroller gerçekleştirilmektedir. İlgili kısıtlama

görselleştirildiğinde formal teorinin arkasındaki dönüşümler daha kolay uygulanabilmektedir.

Nesne rol modeli, birleşik modelleme dili ve varlık ilişki modelinden farklı olarak dil tabanından oluşturulmuştur. Alan uzmanı tarafından iletişim ile doğrulama için söylevden ve popülasyondan fayda sağlamaktansa, akıl ile tasarlanan bir dilden fayda sağlamak daha verimli olmaktadır. Nesne rol modelinin öğrenmesi diğer modellere göre daha kolay gerçekleşmektedir [46].

Alan uzmanı, kavramsal modeli onaylandıktan sonra uygulamalar için veritabanı yönetim sistemi haritası veya program koduna ihtiyaç duymaktadır. Veritabanı uygulamaları için tablo yapısını görmek, yabancı anahtar ilişkisi ve diğer ilişkilerde ilişkisel veya nesne tabanlı modeller, varlık ilişki modeli ve birleşik modelleme dilinde olduğu gibi kompakt bir görüş sunmaktadır. Modellerin her özelliği ilişki olarak göstermesinden dolayı nesne rol modeli özellik tabanlı modelden daha fazla alan kaplamaktadır. Mantıksal tasarım aşamasında özellik tabanlı modeller nesne rol modelinden daha fazla yarar sağlamaktadır. Birleşik modelleme dili ise veri modeli üzerine detaylı uygulamalar içermesi ve model yaklaşımından dolayı nesne tabanlı kodların fiziksel ve mantıksal tasarımı için uygundur [46,71].

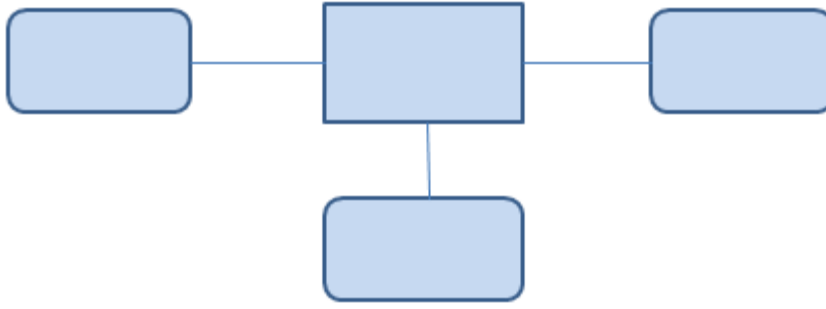
2.4. Veritabanı ve Varlık İlişki Diyagramı

Enformasyon sistemlerinin tasarımında veritabanı, veritabanı mantığı kullanılarak geliştirilen her sistemin özünü teşkil etmektedir.

Varlık-özellik-ilişki yaklaşımı 1976'da P.P. Chen [74] tarafından geliştirilmiş ve veri modellemede kullanımı yaygınlık kazanmış bir yöntemdir. İlişkisel veri modelinin tasviri için farklı nesne ve araçlar kullanılmaktadır. Bu tasvirler gerçekleştirilirken gerçek dünyadan benzetim esas alınmaktadır. İlişkisel veritabanında kullanılan nesne türleri, özellikler, varlıklar ve ilişkilerdir [53,75,76].

Varlığın gerçek ile düşüncedeki gerçeklik arasındaki farklılık olarak ifade edilmesi mümkündür. Varlık, diğer nesnelerden ayıran fiziksel veya soyut nesnel bir oluşumdur. Özellikler ise varlıkların karakteristik isimlerini oluşturmaktadır. Özellikler, varlıklar üzerinde çok sayıda işlem yürütebilmektedirler. Özellikler, varlıkları tanımlamanın yanı sıra varlıkların sınıflandırılması veya diğer yapılarda ilişki kurmasını sağlamaktadır. Çok sayıda özellik tarafından varlıkların tanımlanması mümkündür. Her bir özellik, değerler

içeren bir kümeye sahiptir. Özellik değeri ile özellik adı birbirinden farklı değerlerdir. Varlık türleri, aynı özellikler tarafından karakterize edilen varlık sınıflarıdır. Böylelikle varlık türlerinin varlık oluşumları sınıfının soyutlaması ile ilgili olduğu gözlemlenmektedir. İleri dereceli nesneler tüm muhtemel varlık oluşumlarını temsil etmektedir. Tasarım kararlarına ilişkin kayıtlar ve aksiyon sonuçları uygun grafik notasyonları ile temsil edilmektedir. Varlık-özellik-ilişki analizi şeffaftır ve kıyaslaması, kullanıcılar açısından kolay grafik formlar ile gerçekleştirilmektedir. Şekil 2.5.'te varlık ilişki modelini temsil eden diyagram örneği verilmiştir. Varlık, özellik ve ilişkinin grafiksel olarak belirtilmesinde, köşeli dikdörtgen kutucukta varlıklar, yuvarlanmış dikdörtgen kutularda özellikler ve çizgiler de varlık ile özellikler arasındaki ilişkiler belirtilmektedir [53].



Şekil 2.5. Veritabanı diyagramına ilişkin örnek

Varlık-ilişki-özellik modelleme esasen gerçek dünyayı belirli oluşumlardan ziyade soyutlama ve türler aracılığı ile algılamaktadır. Çok sayıda özelliğin varlığı karakterize edebileceği gibi geliştiriciler, hedef sistem nesnelerinden uygun olanları seçmelidir. Dikkat edilmesi gereken husus, varlık bütünlüğü ilkelerine göre seçimde ki uygun bulunan özellikler, model ile uyumlu olmak zorundadır. İlkeler göre varlık türleri için her varlık oluşumu, özellikleri tarafından tanımlanabilir olmalıdır. Özellik sayısının, bu kriteri yerine getirecek kadar büyük olması gerekmektedir [53].

Varlık türlerinin iki oluşumu hakkında tüm özellik değerlerinin listelenmesi ile her zaman ayırt edilebileceği varsayılmaktadır. Çok sayıda özellik değerinin kullanılması veritabanını ağırlaştırabileceği gibi verimliliği de düşürebilmektedir. İyi bir veritabanı ihtiyaç bulunan minimum düzeyde özellik kümesi kullanarak varlık oluşumlarını eşit şekilde tanımlamaktadır. Her varlık oluşumunun kendi sınıfı içerisinde benzersiz

tanımlanmasını sağlayan özellikler aday anahtar olarak nitelendirilmektedir. Bileşik anahtar durumunda minimal gereksinimler geçerlidir. Varlık türlerinin birden fazla aday anahtara sahip olması mümkündür. Bu durumda tasarımcı tarafından birincil anahtar belirlenir. Birincil anahtar her bir varlık için benzersiz tanımlayıcılığa sahip olmalıdır. Varlık düzenlerinde nesnelerin ayrıştırılması, modellenmesi ve birbirleri ile olan etkileşimi için tanımlanması gerekmektedir. Nesnelerin ayrışımı, dünyadaki olguları model olarak algıladığından, tıpkı dünyadaki diğer olgular gibi doğal bir olgudur. İlişki genel olarak varlıklar arasında etkileşim ve bağlantının semantik durumu olarak kabul edilmektedir. İlişki iki veya daha fazla varlık türünü içermektedir. Bir varlık oluşumundan bir sınıfa ve yine bir varlık oluşumundan diğer sınıflara bağlantı gerçekleştirebilmektedir. İlişki genellikle üç farklı biçimde gerçekleştirilebilmektedir. Bu ilişkiler, birden bire ilişki (1 to 1), birden çok ilişki (1 to M) ve çoktan çok ilişki (M to M) olarak ifade edilmektedir. İlişkide belirli sınıfların tüm elemanları tanımlanmış ise ilişkinin zorunlu olduğu söylenebilmektedir. Diğer durumlarda ilişki opsiyoneldir [53].

2.4.1. İlişkisel Model

İlişkisel veri modelinin güçlü bir teorik yapısı bulunmaktadır. İlişkisel veri modelinde üç temel ilke esas alınmıştır;

Veritabanının fiziksel ve mantıksal yönleri arasındaki belirgin ayrımın veri bağımsızlığı olarak nitelendirilmesi,

Yapısal sadelik ile kullanıcı ve programcılar rahatlıkla anlayabileceği model ve verilerin diğer veriler ile kolay biçimde bağlantılı olabilmesi,

Küme odaklı sürecin işletilerek tek ifade ile çoklu küme işlemlerinin ifade edilebilmesidir [53,77].

İlişkisel veri modeli, yapısal kısım, davranışsal kısım ve manipülatif kısım olmak üzere üç bileşenden oluşmaktadır. İlişkisel modeldeki en küçük birimler özelliklerdir. Model ile özelliklere uygun alanlardaki değerler çizilmektedir. Alanlar, belirli özellikler için mümkün değerleri içermektedir. İşlemler açısından alan kavramı, veri türleri için eşdeğer kavram olabilmektedir. Alan ve varlık işlemlerinin birbirinden ayrılması daha iyi anlaşılması bakımından gereklilik arz edebilmektedir. Alan prosedürlerinin tamamen işlemlere kapalı olmasından dolayı veritabanlarının tüm alanlarının tamamen belirtilmesi

zorluk oluşturmaktadır. Bunun dışında, özelliklere ilişkin değerlerin alanlarda uygun olmayan veri tipinde yazılması hatalara neden olmaktadır [53].

İlişkinin içerisinde dizilere ilişkin numaralar bulunmaktadır. Her dizi belirli sırada listelenmiş özellik numaraları içermektedir. Özellikler gerçek dünya hakkında önermeler ifade eden yapılar da ve bazı ilişkilerin içerisinde bulunabilmektedir. Her dizi varlık oluşumuna karşılık gelebilmekte iken her ilişki de varlık türüne karşılık gelmektedir.

İlişkisel modelde temel kavram anahtardır. Özellik sezgisel olarak, her ilişki içerisindeki belirli dizileri benzersiz olarak tanımlayabilmektedir. Modelde veri çiftlerinin bulunması ihtimaline karşılık en az bir anahtar bulunması gerekmektedir. Özellikler benzersizlik durumunu bozduğu durumlarda anahtar vasfından çıkarılabilmektedir [53].

İlişkilerde birden fazla aday anahtar bulunabilmesi mümkündür. Bu anahtarlardan biri dizileri tanımlayan birincil anahtar, diğerleri ise alternatif anahtar olarak veritabanı tasarımcısı tarafından ifade edilmektedir. Birincil anahtarda bulunan özellikler birincil özellikler, diğer özellikler ise birincil olmayan özellikler olarak ifade edilmektedir. Anahtarı içeren özellikler kümesi, özelliklerin alt kümesi olarak süper anahtar niteliği ile ifade edilmektedir [53].

2.5. Soyutlama Mekanizmaları

2.5.1. Sınıflandırarak Soyutlama

Varlık ilişki modelinin temel üç kavramı sınıflandırarak soyutlama uygulamaları olarak geliştirilmiştir.

- Özellik, varlıkların veya ilişkilerin atomik özelliklerini temsil eden değer sınıfıdır.
- İlişki, iki veya daha fazla varlık arasındaki bağlantının gerçek atomik sınıfıdır.
- Varlık, ortak özellikler ile birlikte gerçek sınıfıdır [54,78].

2.5.2. Kümeleyerek Soyutlama

Varlık ilişki modelini kümeleyerek karakterize eden üç durum söz konusudur.

- Varlık, özelliklerin kümelenmesidir.
- İlişki, varlık ve özelliklerin kümelenmesidir.
- Bileşik özellik, özelliklerin kümelenmesidir [54,78].

2.5.3. Genelleyerek Soyutlama

Genelleyerek soyutlama genelleştirilmiş hiyerarşik veya alt kümeler tarafından belirtilmektedir. Genellikle varlıklara uygulanmaktadır. Genelleyerek soyutlama bazı model eklentilerinde ilişki veya modellemede kullanılabilmektedir [54,78].

2.6. Varlık İlişki Modelinin Kalitesi

Varlık ilişki modeli veritabanını geliştirme adına diğer modeller ile kıyaslandığında pek çok avantaj sunmaktadır. Sunduğu bu avantajların başında kavramsal üstünlüğü, sadeliği ve minimum kapasite sunması gelmektedir. Bazı durumlarda ise bu avantajları olumsuz algılar oluşturabilmektedir. İyi bir varlık ilişki modelinin niteliklerini ifade etmek sistemin ihtiyaçlarına göre değişebilmektedir. Varlık ilişki modeli açıkça ifade edilebildiği gibi anlamlılığı yönünden de oldukça başarılıdır.

Varlık ilişki diyagramı basit bir model değildir, özelliklerin anlaşılabilir kullanımı, özelliklerin tanımlanması ve önemlilik açısından zorluklar içermektedir. Varlık ilişki modelinin n’li ilişkileri desteklemesi, ikili model ile sıklıkla kıyaslanmaktadır. Varlık ilişki diyagramları görünüşte yoğun olmalarına rağmen yapısal olarak en düşük yoğunluğa sahip modellerdir. Varlık ilişki diyagramında tanımlanan tüm yapılar grafik ile gösterilebilmektedir. Bu diyagramlar yorumlama ve okuma açısından sadelik ve kolaylık içermektedir.

Varlık ilişki modeli, güçlü bir veritabanının temelini teşkil etmenin yanı sıra, basit sade ve kullanıcı açısından yararlı bir model sunmaktadır [54].

2.6.1. Normal Formlar

Birinci normal form tablolarda iki boyutlu veri gerektirmektedir.

İkinci normal form tablolardaki verilerin tablo anahtarlarına bağımlı olmasını gerektirmektedir. Kısmi bağımlılığa izin verilmemektedir.

Üçüncü normal form tablodaki verilerin tablonun birincil anahtarına bağlı olmasını gerektirmektedir [45].

2.6.2. Normalleştirilmiş ve Normalleştirilmemiş İlişkiler

İlişkisel model, işlemler ve verilerin birlikte tutulduğu yapısal kümeleri barındıran formal bir sistem olarak tasarlanmıştır. Temel düşünce, önceden tanımlanmış işlemlerin uygulanarak yapısal olarak düzgün ve serbest alt yapılar oluşturulmasıdır. Bu yapılar, başlıklar ve alt başlıklar olarak yapısal bölünmeler gerçekleştirmektedir. Tablo, tamamlayıcı elemanları ile ilgili çeşitli ara sonuçları içermektedir. Bu işlem manuel olarak süreçten geçirilmekte ve işlem, yazı-tablo içeriğine sahip olmaktadır. Yazı tablo modeli bazı işlem sonuçlarını kolaylaştıran ve hızlandıran bir kullanım sunması yönünde fayda sağlamaktadır. Veritabanı tasarımında istenilen sonuçlar bilgisayarda işlenebilir olmalı ve sonuç ile işlemler tablo satırlarında tutulabilmelidir. Bu nedenle veritabanı ile ilişkili tüm yapılar erişilebilir olmalıdır. Veritabanı ilişkilerinde dönüşüm, saf olarak kabul edilen normalleştirme anlamı taşıyan işlemlerdir. Normalleştirme bir dönüşüm işlemidir [53].

2.6.3. Normalizasyon Kuralları

Veritabanına aktarılabacak verilerin özelliklerinin belirlenmesi veritabanı tasarımının önemli bir aşamasıdır. Bu aşamada özelliklerin belirlenmesinde kullanılan kurallar, normalizasyon kuralları olarak ifade edilmektedir. Bu kurallar kısaca şu şekilde ifade edilmektedir [48].

Birincil anahtarı olmayan alanlar, tablo içerisinde birincil anahtar olan alanlara bağlanmaktadır.

Birincil anahtar alanları ile birincil anahtar olmayan alanlar arasında birden çok (1 to M) ilişki gerçekleşmemektedir.

- Anahtarı olmayan alanlar arasında bağlantı gerçekleşmemektedir.
- Tablolar olabildiğince minimum bölümlere ayrılmaktadır.
- Bir satır içerisinde birden fazla bilgi bulunmamaktadır [48].

2.6.4. Veritabanı Oluşturma İşlemleri

Veritabanı oluşturulması başlı başına uzun bir işlem sürecini ve doğru adımların atılmasını gerektiren prosedürleri içermektedir. Veritabanı tasarlanırken oluşturulacak küçük bir hata, yanlış veri tanıtımı, modelleme hatası gibi pek çok unsur bilgi sisteminin problemlili oluşturulmasına neden olmaktadır [48].

Veritabanı oluşturulurken dikkat edilmesi gereken hususlar bulunmaktadır. Bu hususlar şu şekilde ifade edilmektedir.

Veritabanında kullanılması öngörülen nesneler ve bu nesnelere bağlı özellikler belirlenir.

Nesneler için isimler belirlenerek tablolar oluşturulur. Tabloların içerisinde tutulacak veriler belirlenir.

Tablolardaki verileri birbirinden ayırt edebilme ve her bir verinin benzersiz bir niteleyiciye kavuşması için birincil anahtar ataması gerçekleştirilir.

Nesnelerin özellikleri için yeni sütunlar oluşturulur.

Veri tekrarlarından kaçınmak için tasarım tekrar gözden geçirilerek gerekli işlem ve normalizasyon prosedürleri uygulanır.

Tablodaki alanlar ilişki içerisinde, fakat bazı alanların anahtar alan ile bağlantılı olmadığı durumlar söz konusudur. Bu durumda bu alanlar farklı tablolara kaydırılmalıdır.

Veritabanı tek bir tablo veya tek bir alandan oluşmamaktadır. Tablolar veritabanında ilişki içerisinde, Bu ilişkiler tanımlanarak tablolar arasındaki ilişki bağlantıları oluşturulmaktadır [48].

2.7. Metin Madenciliği

Yirminci yüzyıl, dünyanın değişiminde önemli bir rol oynamıştır. Bu değişimde en büyük katkıyı bilgi ve bilgiye bağlı teknolojilerin gelişmesi sağlamıştır. Bilgisayarların ortaya çıkışı, bilginin üretilmesinden bilginin iletilmesine kadar pek çok alanda çığır açmıştır. Bunun yanı sıra internetin ortaya çıkışı ve bilginin veritabanlarında tutulmaya başlaması yoğun miktarda bilgi birikimine yol açmıştır.

Yoğun miktardaki bilgiyi kullanma ve içerisinden doğru bilgiyi çekme açısından bireyler büyük zorluklar ile karşılaşmışlardır. Bu durum bilgi bilimlerinde yeni bir alanın

ortaya çıkışını hazırlamıştır. Büyük veri kitlelerinden ilgili bilginin elde edilmesi ve bu işlemin bilgisayar ortamında yapılması zaman ve pratiklik kazandırmaktadır.

Metin madenciliği, bilgi yığınlarından analiz araçları vasıtası ile kullanıcı-bilgi yığını etkileşimli bilgi elde etme süreci olarak ifade edilebilmektedir. Metin madenciliği de veri madenciliği gibi veri yığınlarından keşif ve tanımlama ile amacına uygun bilgi çıkarmayı hedeflemektedir. Metin madenciliğinin veri madenciliğinden önemli bir farkı bulunmaktadır. Veri madenciliğinde veriler, veritabanlarında düzenli bir şekilde bulunmaktadır. Metin madenciliği için veri kaynakları yapılandırılmamıştır [59].

Metin madenciliği ile veri madenciliği arasında farklılıklar bulunmasına karşın sistem mimarileri kıyaslandığında birbirlerine benzerlik göstermektedirler. Veri madenciliği yapısal olarak verilerin normalize edilmesi ve tablo oluşturma gibi işlemleri gerçekleştirmektedir. Metin madenciliği ise doğal dildeki belgeler için tanımlama ve çıkarım işlemlerini yürütmektedir.

Metin madenciliği istenilen bilgileri elde etmek için pek çok prosedür ve algoritma ile işlem yapmaktadır. Bu yöntem ve metodolojilere göre metin madenciliğinin farklı alanları bulunmaktadır. Bu alanlar, bilgi erişimi, bilgi çıkarımı ve derlem tabanlı bilişsel dilbilimdir [59].

2.7.1. Bilgi Erişimi

İnternet ortamında son yıllarda artan web sitesi sayısı ile çok büyük miktarda bilgi ve içerik sayfası bulunmaktadır. Bilgi erişimi internet ortamında bulunan belgelerin, bazı içerik, kelime ve ifadeler vasıtasıyla eşleşmeye dayanan ve eşleşen belgeleri ekrana getiren sorgulama esasına dayanmaktadır.

Arama motorlarına benzer yapılar ile sunulan bilgi erişim modelleri, arama motoruna girilen, aranan belgeyi tanımlayabilecek ilgili kelimeler ile en iyi eşleşmeyi sağlayan belgeleri sıralamaktadır. Süreç genelleştirilerek ilgili aranan kelimeler, tüm belgeyi temsil etmektedir. Girdi belgesi, yığındaki tüm belgeler ile eşleştirilerek en iyi uyumu veren belgelere erişilmektedir.

Bilgi erişim modelleri, eşleşmelerde için ve ilgili belgeleri bulmada benzerlik ölçüsünden yararlanmaktadır. Arama motoruna yazılan ilgili kelimeler ile aranan tüm belgelerdeki benzerlikler karşılaştırılmaktadır. En yakın komşu yöntemi, bilgi erişiminde kullanılan yöntemlerden biridir [60].

2.7.2. Bilgi Çıkarımı

Bilgi çıkarımı, büyük veri kitlelerinden temsil yeteneği kuvvetli örnek bilgi alt dizisinin elde edilme işlemleridir. Bilgi çıkarımı, iyi bir temsil yeteneğine sahip olmalıdır. Elde edildiği kitleden anlamca aykırılık taşımamalıdır. Bilgi çıkarım yöntemleri, sağladığı avantajlar neticesinde pek çok disiplinde uygulama sahası bulmuştur.

Bilgi çıkarımı bilgisayar programları tarafından otomatik biçimde gerçekleştirildiği için bu programlar bazı özel yöntemler ile geliştirilmiştir. Bilgi çıkarımında kullanılan en yaygın yöntemler, kural tabanlı öğrenme yöntemi, sınıflama tabanlı model yöntemi, ardışık etiketleme tabanlı yöntemler bu yöntemlerden bazılarıdır. Bilgi çıkarımı iki aşamadan oluşmaktadır. İlk aşama çıkarım aşaması, ikinci aşama ise eğitim aşamasıdır [61].

2.7.3. Metin Madenciliğinde Kullanılan Yöntemler

Metin madenciliği, içerisinde pek çok matematiksel ve istatistiksel tekniği barındıran bilgi biliminin alt disiplinidir. Uygulama sahası bakımından pek çok alanda kullanılan metin madenciliği, veri madenciliği gibi bilgisayar mühendisliği alanının temel çalışma alanlarından biridir. Böylesine kompleks bir yapıya sahip çok disiplinli bir alan olan metin madenciliği içerisinde farklı çözümler sunan yöntemler mevcuttur. Bu yöntemlerden bazıları metin sınıflandırma, metin kümeleme, metinden bilgi çıkarımı, bilgi erişimi, frekans odaklı yaklaşımlar ve anlam bilimsel analiz olarak belirtilmektedir.

2.7.4. Metin Özetleme

Metin özetleme, büyük paragraf veya uzun metin yığınlarından anlamsal içeriği bozmayan sadeleştirilmiş ve kısaltılmış bilginin elde edilmesini amaçlamaktadır.

Metin özetleme yaklaşımları, çıkarım teknikleri üzerine yoğunlaşmış gibi görünse de yapısal olarak derin farklılıklar içermektedir. Metin özetlemede kullanılan yaklaşımlar avantaj kabul edilebilecek farklılıklara sahip olabilmektedir. Genel olarak metin özetleme sistemleri üç ayrı işlemi içermektedir. Bu işlemler ara temsil oluşturulması, cümlelerin puanlanması ve birkaç cümleden oluşan özet seçimi olarak ifade edilmektedir [62].

2.7.5. Konu Temsili Yaklaşımı

Konu temsili yaklaşımı metin özetleme metotları içerisinde performansı ile yaygın olarak kullanılan bir yöntemdir.

Konu sözcüğü metin özetleme yaklaşımlarında ön çalışma olarak belgedeki kelimeleri tanımlamak için sıklık eşiği olarak kullanılmaktadır. Belgelerden sık olarak kullanılan kelimeler belirlenmekte ve Log-olabilirlik testi uygulanmaktadır [62,81].

2.7.6. Frekans Odaklı Yaklaşımlar

Konu sözcüğünün belirlenmesi için kelimelerin ağırlıklandırılması, belgedeki kelimelerin belirli ağırlıklara sahip olması, konuyu daha iyi açıklayabilen kelimelerin daha yüksek ağırlık katsayısına sahip olması konuyla ilişkili kelimelerin belirlenmesinde yararlı olmaktadır. İkili değerler yerine sürekli değerleri kullanmak daha başarılı sonuçlar verebilmektedir. Bu konuda ileri sürülen tekniklerin biri de terim frekansı*ters belge frekansı (TF*IDF) yaklaşımıdır [62,82].

2.7.7. Gizli Anlam Bilimsel Analiz

Gizli anlam bilimsel analiz yöntemi gözlenen eş anlamlı kelimelere bağlı olarak metinlerin gizli anlam bilimsel çözümlemesini gerçekleştiren denetimli olmayan güçlü bir yöntemdir. Gizli anlam bilimsel analiz, çözümlemede lineer cebir tekniklerini esas almaktadır. Gizli anlam bilimsel analiz yönteminde her belge bir vektörü, kelimeler n boyutlu satırları, cümleler ise m boyutlu sütunları oluşturmak üzere matrisin çözülmesi amaçlanmaktadır. Cümlelerin kelimeleri içermesi durumunda kelimelerin ağırlığı TF*IDF katsayısına eşittir. Kelime içermeyen cümlelerin ağırlık katsayısı sıfırdır. Elde edilen matrise tekli değer ayrışımı yöntemleri uygulanarak ilgili değerler elde edilmektedir. Bu yöntem ile gizli anlam bilimsel analiz başarılı sonuçlar üretmektedir [62].

2.7.8. Bayescil Konu Modeli

Bayescil konu yaklaşımı metin özetlemede gittikçe yaygınlık kazanan başarılı bir model olarak ön plana çıkmaktadır. Bu modelin en önemli özelliği birkaç dökümanı birden özetleyebilmesidir. Denetimsiz bir model olan Bayescil konu modeli birçok modelin ortaya çıkaramadığı bilgileri ortaya çıkarabilmektedir. Bayescil konu modeli Kullback-Lieber ayrışımından yararlanmaktadır. Kullback-Lieber ayrışımı aynı olayların sahip oldukları olasılıklar arasındaki uyumsuzlukları ortaya koymaktadır [62,79,80].

2.7.9. Cümle Kümeleme

Cümle kümeleme yaklaşımı birden fazla belge içeren veya birden fazla konuya ilişkin dökümanların birbirine benzerlik gösteren bilgilerini kümeleme işlemlerini içermektedir. Cümlelerin farklı belgelerde birebir bulunması mümkün olmayabilir. Fakat içerdiği bilgi bakımından benzerlik gösterebilmektedir. Özetleme işleminde vektörler arasındaki benzerliği belirlemede kosinüs benzerlik ölçüsü kullanılarak vektör ile temsil edilen cümlelerin benzerlikleri hesaplanmaktadır [83]. Kosinüs benzerlik ölçüsü vektörler arasındaki benzerliği Bu yaklaşım ile birlikte cümle benzerliklerinin kümelenmesi işlemi, konunun temsili olarak kabul edilmektedir. Temsil cümlelerinden bir seçim yapmak, özet çıkarmak için iyi bir yöntemdir. Bu seçim, kümelerden gerçekleştirilmektedir [62].

2.7.10. İçerik Etkisi

Metin özetleme yöntemleri sadece belgeleri özetlemek için geliştirilmemiştir. Milyonlarca bilgi içeren belgelerden önemli bilgileri özetlemede yararlı olabilmektedir. Bu bağlamda internet ortamında sunulan web sayfaları ve bu sayfaların bağlantılı olduğu diğer sayfalardan özetleme yapmak, metin özetlemedeki içerik etkisinin temel kazanımı olmaktadır. İçerik etkisi yaklaşımı, özetleyeceği içeriğe göre farklı çözümler öne sürmektedir. Buna göre web sayfalarını özetlemede, bilimsel çalışmaları özetlemede, sorgu odaklı özetlemede ve e-mail özetlemede uygun özetleme modeli belirlendikten sonra özetleme işlemi gerçekleştirilmektedir [62].

2.7.11. Makine Öğrenme Yöntemi

Özetleme yaklaşımı için makine öğrenme yöntemlerinin uygulanması Edmunson'un çalışmalarında söz konusu olmuştur [84,85]. Edmunson ele alınan konuların tek olarak temsil edilmesinden ziyade bazı önemli göstergelerinin birleştirilmesini önermiştir.

Özetlemede denetimli yöntemler, ele alınmış konularda özeti olan ve olmayan tüm cümleler ikili sınıflama problemi olarak temsil edilen önemli cümlelerin seçiminde kullanılmaktadır. Cümlelerdeki öz bilgiler, istatistiksel sınıflama yöntemleri ile ayrıştırılarak önemliliğe göre listeler ile temsil edilmektedir.

Cümlenin olabilirliği, özet sınıfı veya sınıflayıcının güveni özet içerisinde cümlenin puanıdır. Sınıflayıcının seçimi cümle puanı fonksiyonu açısından önem taşımaktadır. En yüksek puana sahip cümleler özetleme için seçilmektedir. Seçilen bu cümleler birleştirilerek özet oluşturulmaktadır [62].

2.8. Metin Madenciliğinde Sınıflandırma Yöntemleri

2.8.1. Olasılıksal Sınıflandırma

Olasılıksal sınıflandırma yöntemi genel olarak Bayes teoremi üzerine kurulmuştur. Bu yöntemde sınıflandırma durum değeri $SDD(d, c)$ d belgesindeki c sınıfının $P(c|d)$ olasılığını hesaplamak için,

$$P(c|d) = \frac{P(d|c)P(c)}{P(d)} \quad (2.1)$$

eşitliğinden yararlanmaktadır.

Marjinal olasılıkları tüm sınıflar sabit bir değer içermesinden dolayı her koşulda hesaplamak mümkün olmamaktadır. $P(d|c)$ hesaplaması için ise bazı varsayımlara ihtiyaç duyulmaktadır. $d = (w_1, w_2, \dots)$ olmak üzere,

$$P(d|c) = \prod_i P(w_i|c) \quad (2.2)$$

eşitliği Naive Bayes sınıflayıcısı olarak ifade edilmektedir. Naive Bayes sınıflandırıcıları kelimenin ve belge içeriklerinin sınıflandırılmasında olasılıksal temellere dayanan başarılı bir yöntemdir [59,86].

2.8.2. Kural Tabanlı Sınıflandırma

Kural tabanlı sınıflandırmada, özellik kümesi altında yatan koşullar, sınıf etiketlerinin bulunduğu kurallar kümesi ile veri uzayı modellenmektedir. Kurallar kümesi eğitim verilerinden üretilen modeldir [62].

Kural tabanlı yöntemler, bazı optimal kriterlere göre mümkün tüm kurallar arasından en iyi kuralı seçme alternatifini kullanmaktadırlar. Kurallar ayrık normal formu ile yakından ilgilidir. Ayrık normal form kuralları aşağıdan yukarı yönlü dizilişe sahiptir. Başlangıç sınıflandırıcıları eğitim kümesinden elde edilmektedir. Kural öğrenciler, özel yöntemlerden optimal kriterlerde farklılık göstermektedir. Bu konuda geliştirilmiş algoritmalarından biri de RIPPER algoritmasıdır. RIPPER algoritması yüksek hassasiyet, başarı ve güçlü hafıza gibi özelliklere sahiptir [87-89]. Kural tabanlı sınıflama yöntemi pek çok çalışma alanında uygulama sahası bulmuş ve halen güncelliğini koruyabilen bir yöntemdir [59].

2.8.3. Doğrusal Sınıflama

Doğrusal sınıflama yöntemi, lineer tahmincilerin sonuçları $P = \bar{A} * \bar{X} + b$ ve $\bar{X} = (\alpha_1, \alpha_2, \dots, \alpha_n)$ normalize edilmiş belge sözcüğü sıklığı vektörü, $\bar{A} = (a_1, \dots, a_n)$ özellik uzayı ve b reel sabit olmak üzere aynı boyuttaki vektörlerdir. $P = \bar{A} * \bar{X} + b$, sınıflar arasındaki ayrılmış düzlemlerdir [62]. Doğrusal sınıflama yöntemlerinin farklı modelleri bulunmaktadır.

2.8.4. Destek Vektör Makinesi

Destek vektör makinası, metin sınıflandırmada etkili ve başarılı sonuçlar üreten bir algoritmadır. Geometrik anlamda ikili destek vektör makinası sınıflayıcı kategorilerin temsil açısından olumlu örneklerin olumsuz örneklerinden ayrıldığı özellik uzayıdaki bir

düzlem olarak görmek mümkündür. Düzlemdeki sınıflama eğitim esnasında benzersiz düzlemdeki bilinen pozitif örnekleri, bilinen negatif örneklerden maksimum marj ile ayırma işlemi olarak bilinmektedir. Pozitif ve negatif kümelerin komşu noktaları arasındaki uzaklık marj olarak ifade edilmektedir. Destek vektör düzlemi eğitim örneklerinin küçük alt kümeleri tarafından belirlenmektedir. Diğer eğitim verilerinin sınıflandırmaya katkısı bulunmaktadır. Destek vektör makinaları sınıflama açısından sağladığı kolaylıklar ile avantajlı konumdadır. Ayrıca sağladıkları performans açısından oldukça başarılıdır [59].

2.8.5. Yapay Sinir Ağları

Yapay sinir ağları yöntemi ile sınıflandırma, nöronların analog test hesaplama yöntemlerinin geliştirilmesinde nörobilimciler tarafından sunulan modeldir. Sinir ağları bağlantılı giriş çıkış kümesini içeren ve her bağlantının belirli bir ağırlığa sahip olduğu sistemi içermektedir. Öğrenme aşamasında ağın öğrenmesi, girişteki bilgilerin sınıf etiketlerini doğru tahmin edebilmek için ağırlıklar tarafından ayarlanarak gerçekleştirilmektedir.

Yapay sinir ağları kompleks modelleri öğrenmede başarılı sonuçlar verebilmektedir. Ağın öğrenmesi uzun bir eğitim süresi gerektirse de uygulamada başarılı sonuçlar alınmasını sağlamaktadır. Bu başarıyı, sağlam ve iyi dizayn edilmiş ağ yapısı ile sağlamaktadır.

Yapay sinir ağları ile sınıflama pek çok avantajı birlikte sunmaktadır. Az sayıda bilgi ve özellik arasındaki ilişkiyi ortaya koymada kullanılabilir. Sürekli değerlere sahip giriş çıkış modelleri için uygun olmaktadır. Pek çok alanda, tanımlama, sınıflama, tıbbi tetkikler gibi alanlarda başarılı sonuçlar üretmektedir.

Yapay sinir ağlarının farklı algoritma ve modelleri bulunmaktadır [63].

2.8.6. Regresyon Yöntemi

Regresyon modeli metin sınıflandırmada bir dizi üzerinde bulunan noktaları bilginin değerlerini kullanarak gerçek değerli fonksiyona yaklaşık değerler elde etmek için kullanılmaktadır. Bu işlemler ile sınıf belirlemeyi amaçlamaktadır. Regresyon modelinde

atama fonksiyonu, sürekli gerçel değerler fonksiyonu üyesi olarak düşünülmektedir. Böylelikle regresyon, sınıflayıcı olmaktan ziyade sınıflayıcı oluşturmak için kullanılmaktadır. Lineer en küçük kareler metodu regresyon modelinde metin sınıflama için kullanılan temel yöntemdir [90]. Sınıflandırıcıyı oluşturma aşamasında eğitim verileri için matris oluşturularak lineer en küçük kareler yöntemi ile eğitim verilerindeki hatalar minimize edilmesinde A herhangi bir matris ve F Frobenius normu olmak üzere,

$$\|A\|_F = \sqrt{\sum A_{ij}^2} \quad (2.3)$$

eşitliği kullanılmaktadır. Regresyon yönteminde matrisin elde edilmesi için tekil değer ayrışımı kullanılmaktadır [59].

2.8.7. Karar Ağaçları

Karar ağaçları sınıflandırma işlemlerinde hiyerarşik bir düzen izlemektedir. Karar ağaçları, karar düğümleri ve yapraklardan oluşmaktadır. Karar ağacı sınıflandırma yönteminde ele alınan belge tekrarlı işlemler vasıtasıyla terimlerin ağırlık değerleri ile kıyaslanmaktadır.

Kıyaslama işlemi ağaç yaprağında son bulur ve ilgili yaprak ele alınan belgenin sınıfını belirler [64].

Bu yöntemin hiyerarşik bir yapıda olması nedeni ile test için alınan belge, işlem sürecine kökten başlayarak yaprak ucuna ulaşıncaya dek devam etmektedir.

2.9. Markov Zincirleri

Belirsizliğin modellenmesinde, olasılık dağılımlarına ait kantitatif olası sonuçlar kümesi tanımlanmaktadır. Modelleme esnasında sürecin anlaşılması için dağılım özellikleri önem kazanmaktadır. Stokastik süreç t anında X durumundaki rasgele değişkenlerin toplamıdır [55].

$\{x, t \geq 0\}, t \in T$ olmak üzere T sınırlıdır ve sayılabilir stokastik süreç olarak ifade edilebilmektedir. İndislerin kesikli veya sürekli olduğu varsayılmaktadır. Rasgele yürüyüş,

Markov zincirleri, Wiener süreçleri, stokastik diferansiyel eşitsizlikler stokastik süreçlerin farklı türleri olarak bazı uygulamalarda kullanılabilmektedir [55].

Markov zincirleri, 1907 yılında Rus Matematikçi Markov tarafından geliştirilmiş [105], sosyal bilimlerden bilgisayar bilimlerine pek çok alanda uygulama sahası bulmuştur. Markov zinciri kısaca ifade edilmek istenirse iki durumlu bir Markov zinciri geçiş olasılıkları matrisi üzerinde,

$$P = \begin{bmatrix} p & 1-p \\ 1-p & p \end{bmatrix}$$

olarak ifade edilmektedir.

Markov zincirlerinin matematiksel ifadesi,

$$P(X_n = j | X_{n-1} = i) \quad p_{ij} \geq 0 \quad i = 1, 2, \dots, n \quad j = 1, 2, \dots, n \quad (2.4)$$

$$\sum_j^n p_{ij} = 1 \quad (2.5)$$

olarak belirtilmektedir [55].

P_{ij} zamandan bağımsız olmak üzere,

$$P(X^{(n+1)} = i | X^n = j, X^{n-1} = i_{n-1}, \dots, X^{(0)} = i_0) = P_{ij} \quad n \geq 0 \quad (2.6)$$

$$i, j, i_0, i_1, \dots, i_{n-1} \in M$$

eşitliği, Markov zinciri süreci olarak ifade edilmektedir. $X^{(n+1)}$ gelecek durumun koşullu dağılımı için geçmiş durumlar $X^0, X^1, X^2, \dots, X^{n-1}$ ve mevcut durum X^n olmak üzere X^{n+1} geçmiş durumlardan bağımsızdır ve sadece mevcut duruma bağlıdır.

P_{ij} olasılığı sürecin i durumundan j durumuna geçişini temsil etmektedir. Geçişlerin toplam olasılığı,

$$P_{ij} \geq 0 \quad \sum_{i=0}^{\infty} P_{ij} = 1 \quad j = 0, 1, \dots \quad (2.7)$$

eşitliği ile ifade edilmektedir. Markov zincirlerinde n adım geçiş olasılıkları P_{ij}^n ile temsil edilmektedir. P^n n adım geçiş olasılığı matrisi ve p bir adım geçiş olasılığı olmak üzere P_{ij}^n i durumundan j durumuna n geçiş olasılığı sürecini ifade etmektedir.

$P^n = P * P * P * ... * P$ olmak üzere,

$$P_{ij}^{(n+1)} = \sum_{k \in M} P_{ki}^{(n)} P_{jk}^{(1)} = \sum_{k \in M} P_{ki}^n P_{jk} = [P^{n+1}]_{ij} \quad (2.8)$$

eşitliği sağlanmaktadır.

Markov sürecinin durumları $n = 0$ süreç i durumunda a_i ve $i = 0, 1, 2, \dots$ olmak üzere i durumunda ve n geçiş sonrasında j durumunda olması $P_{ji}^{(n)} = [P^n]_{ji}$ ve P_{ji} i durumundan j durumuna bir adım geçiş olasılığı olmak üzere,

$$\sum_{i=0}^{\infty} P(X^{(0)} = i) * P_{ji}^{(n)} = \sum_{i=0}^{\infty} a_i * [P^n]_{ji} \quad (2.9)$$

eşitliği ile gösterilmektedir.

$$X^{(n)} = (\check{X}_0^{(n)}, \check{X}_1^{(n)}, \dots) \quad (2.10)$$

Markov zinciri sürecinin n . geçişteki durumların olasılık dağılımıdır.

$X_i^{(n)}$ i durumunda n geçiş sonrası olasılık değeri,

$$\sum_{i=0}^{\infty} \check{X}_i^{(n)} = 1 \quad (2.11)$$

olarak ifade edilmektedir. Buradan,

$$X^{(n+1)} = P X^{(n)} \quad (2.12)$$

$$X^{(n+1)} = P^{(n+1)} X^{(0)} \quad (2.13)$$

elde edilmektedir [56].

P^T matrisi bir durumdan diğer durumlara T adımda geçiş olasılıklarını gösterebilir. Geçiş olasılıkları adım sayısının büyük olması durumunda geçiş olasılıkları, ergodik geçiş olasılıkları ve kararlılık olasılığı olarak isimlendirilmektedir. Buna göre,

$$\pi_{ij} = \lim_{t \rightarrow \infty} p_{ij}(t) \quad (2.14)$$

$$\pi = \lim_{t \rightarrow \infty} p^t \quad (2.15)$$

ve Chapman-Kolmogorov matrisinin uygulamasının çarpma formülü,

$$\begin{aligned} \pi &= \lim_{t \rightarrow \infty} p^t \\ &= \lim_{t \rightarrow \infty} p p^{t-1} \\ &= p \lim_{t \rightarrow \infty} p^{t-1} \\ &= p \pi \end{aligned} \quad (2.16)$$

eşitliği ile sağlanmaktadır. Bu eşitsizlikler için lineer eşitsizlik sistemleri kullanılarak ergodik olasılıkların hesaplanması,

$$\sum_{i=1}^n \pi_i = 1 \quad \pi_i \geq 0 \quad (2.17)$$

ile sağlanmaktadır [55].

2.9.1. Geçiş Zamanı

Genellikle belirli bir durumun sağlanması için zaman belirlenmelidir. Varsayalım ki i durumunda i durumundan j durumuna n adımda ilk geçiş olasılığı $f_{ij}(n)$ ile gösterilsin. Bir adım geçiş için geçiş matrisi bu olasılığı vermektedir.

$$f_{ij}(1) = p_{ij}(1) = p_{ij} \quad (2.18)$$

$$f_{ij}(2) = p_{ij}(2) - f_{ij}(1)p_{ij} \quad (2.19)$$

olmak üzere genel olarak ilk geçiş olasılığı

$$f_{ij}(n) = p_{ij}(n) - [f_{ij}(1)p_{ij}(n-1) + f_{ij}(2)p_{ij}(n-2) + \dots + f_{ij}(n-1)p_{ij}] \quad (2.20)$$

eşitliği ve j durumları bağlantılı ise geçiş süresi,

$$\mu_{ij} = \sum_{n=0}^{\infty} n f_{ij}(n) \quad (2.21)$$

ile ifade edilmektedir. Ortalama ilk geçiş süresi $k = j$ ve $\mu_{kj} = \mu_{kk} = 0$ olmak üzere,

$$\mu_{ij} = 1 + \sum_{k \neq j} p_{ik} \mu_{kj} \quad (2.22)$$

eşitliği ile elde edilmektedir [55,91].

2.10. Saklı Markov Modeli

Saklı Markov modeli Markov zincirinden genel olarak farklı bir yapıda geliştirilmiştir. Saklı Markov modeli Markov zincirinden farklı parametrelerle modellenen Markov zincirinin ek özellikler kazanmış formu olarak ifade edilmektedir.

Saklı Markov zincirindeki amaç modeldeki gözlemlere ilişkin emisyon olasılıkları aracılığı ile gözlemlenemeyen durum geçişlerinin tahmin edilmesine dayanmaktadır. Saklı Markov modelinde “saklı” ifadesinin verilmesinin nedeni durumların doğrudan gözlenememesi ve gözlenen veriler vasıtası ile tahmin edilmek istenmesidir. Bir sonraki duruma geçişin sadece mevcut duruma bağlı olması ve durumlar arasında geçiş olasılıkları gizli sürecin doğal bir Markov olduğunu göstermektedir. Gözlemler veya durumlar dizisi sınırlı durum uzayı ile Markov zincirini takip etmektedir. Herhangi bir zamandaki gözlemin olasılık dağılımı, sadece Markov zincirindeki mevcut duruma bağlıdır [55].

Bilgisayar bilimlerindeki hızlı gelişmeler neticesinde dönemin bilim insanları gerçek sorunlara çözüm sağlamak amacıyla bazı algoritmalar için deterministik ve stokastik otomatlar ile çalışmışlardır. Lineer filtre teorisi, klasik filtre teorisinde etkin iken, zamanla

stokastik filtre ve lineer olmayan filtreler daha önemli pozisyona gelmiştir. 50'li yıllarda Luenberger-observer Wiener filtreleri ve Kalman filtreleri öne sürülmüştür [92-94].

Saklı Markov modelini anlamak uygulamak ve oluşturmak için algoritma geliştirme tarihindeki önemli unsurlar dikkate alınmalıdır [94].

2.10.1. Saklı Markov Modelinin Temelleri

Saklı Markov modeli sınırlı öğrenen modellerdir. Saklı Markov modelinde iki stokastik süreç söz konusudur. Bu süreçlerden biri çok boyutlu olasılık dağılımları ile bağlantılı sınırlı durum kümesidir. Farklı durumlar arasındaki geçişler geçiş olasılıkları olarak adlandırılan olasılık kümesi tarafından düzenlenmektedir. Diğer süreç ise gözlemlerden oluşmaktadır. Gözlem dizisi doğrudan gözlenebilmekte iken durumlar gizlidir. Bu nedenle gözlem dizisine bağlı olarak gizli durumlar arasındaki geçişler olasılıksal olarak tahmin edilmek istenmektedir. Durumların doğrudan gözlenememesi bu modele Saklı Markov modeli adının verilmesinde etkili olmuştur. Her bir Saklı Markov modeli, geçiş olasılıkları matrisi, başlangıç olasılıkları vektörü ve emisyon olasılıkları matrisinden oluşmaktadır [95,97,99-104].

Saklı Markov modeli 5 madde ile ifade edilebilmektedir.

1. Modelin N durumu vardır ve durumlar $S = \{S_1, S_2, S_3, \dots, S_N\}$ ile ifade edilmektedir.
2. M gözlem sembolü ve her durum $V = \{v_1, v_2, v_3, \dots, v_M\}$. Gözlemler sürekli ise M sınırsızdır.
3. Durum geçiş olasılıkları dağılımı $A = \{a_{ij}\}$ ve $t + 1$ anında S_j durumuna t anındaki S_i durumundan geçiş olasılığıdır. Stokastik matrisin bu yapısı modelin yapısal bağlantıları olarak tanımlanmaktadır. a_{ij} katsayısının sıfır olduğu durumlarda eğitim süreci boyunca sıfır olmaya devam edecek ve hiçbir zaman S_i durumundan çıkamayacaktır. q_t mevcut durum olmak üzere,

$$S_j. a_{ij} = p\{q_{t+1} = j | q_t = i\} \quad 1 \leq i, j \leq N \quad (2.23)$$

$a_{ij} \geq 0$ ve $1 \leq i, j \leq N$ olmak üzere,

$$\sum_{j=1}^N a_{ij} = 1 \quad 1 \leq i, j \leq N \quad (2.24)$$

olarak ifade edilmektedir.

4. Gözlem sembol olasılıkları dağılımı her durum için $B = \{b_i(k)\}$ ve $b_j(k)$, v_k sembolünün S_j durumundaki olasılığıdır.

$$b_j(k) = p\{o_t = v_k | q_t = j\} \quad (2.25)$$

Stokastik kısıtlama ile $b_j(k) \geq 0 \quad 1 \leq j \leq N \quad 1 \leq k \leq M$ olmak üzere

$$\sum_{k=1}^M b_j(k) = 1 \quad (2.26)$$

eşitliği elde edilmektedir. Gözlemler sürekli ise, kesikli olasılıklar yerine sürekli olasılık yoğunluk fonksiyonu kullanılmaktadır. c_{jm} = ağırlık katsayısı, μ_{jm} =ortalama vektörü, Σ_{jm} =kovaryans matrisi olmak üzere,

$$b_j(o_t) = \sum_{m=1}^M c_{jm} N(\mu_{jm}, \Sigma_{jm}, o_t) \quad (2.27)$$

stokastik varsayım $c_{jm} \geq 0$, $1 \leq j \leq N$ ve $1 \leq m \leq M$ olmak üzere,

$$\sum_{m=1}^M c_{jm} = 1 \quad 1 \leq j \leq N \quad (2.28)$$

ile gösterilmektedir.

5. Saklı Markov modelinin başlangıç durum dağılımı $\pi = \{\pi_i\}$, Model S_i durumunda ve $t = 0$ anında,

$$\pi_i = p\{q_1 = i\} \quad 1 \leq i \leq N \quad (2.29)$$

eşitliği ile ifade edilmektedir. Kesikli parametrelili Saklı Markov modeli genel olarak,

$$\lambda = (A, B, \pi) \quad (2.30)$$

notasyonu ile ifade edilmektedir. Sürekli parametrelili Saklı Markov modeli ise,

$$\lambda = (A, c_{jm}, \mu_{jm}, \Sigma_{jm}, \pi) \quad (2.31)$$

notasyonu ile ifade edilmektedir [94].

2.10.2. Saklı Markov Modeli Tahmin Algoritması

Bazı süreçlerde gözlem dizileri Markov süreçleri ile olasılıksal olarak ilişkilidir. Gözlenen durumların sayısı gizli durumların sayısından farklı olabilmektedir. Saklı Markov modeli iki durum kümesi ve üç olasılık kümesi içermektedir. Modele ilişkin bazı tanımlamalar şu şekilde belirtilmiştir.

Saklı Durumlar; Sistemin durumları Markov süreci tarafından tanımlanabilmektedir.

Gözlenen Durumlar; Sürecin durumları gözlenebilir.

Pi-Vektörü; Başlangıç zamanında gizli durumlardan bir tanesinde olasılıkları içeren modeldir.

Durum Geçiş Matrisi; Bir durumdan diğer duruma geçiş yapan gizli durumların olasılıklarını içermektedir.

Emisyon Olasılıkları Matrisi; Saklı durumlara bağlı olarak ortaya çıkan gözlemlerin durumlara göre gözlenme olasılıklarını içermektedir.

Saklı Markov modelleri ölçülebilir durumlar ile bu durumların gizli durumları arasındaki olasılıksal ilişkilerden oluşan güçlü bir Markov süreci sunmaktadır. Markov özelliği ile bir önceki zamandaki durumların değerleri tarafından durumların geçmiş bilgilerinin özetlenebileceği varsayılmaktadır. Sistemin güçlü bir hafızası yoktur, Gözlemler önceki durumlara bağlı değildir. Toplam olasılık,

$$P(y^T, x^T) = p(y_1) \prod_{t=2}^T p(y_t | y_{t-1}) \prod_{t=2}^T p(x_t | y_t) \quad y^T = (y_1, y_2, \dots, y_T) \quad (2.32)$$

ile ifade edilmektedir. Saklı Markov modeli üç matris ile karakterize edilmektedir,

$$A_{i,j} = p(y_t = i | y_{t-1} = j) \quad (2.33)$$

$$B_{t,i} = p(x_t | y_t = i) \quad (2.34)$$

$$\pi_{t,i} = p(y_t = i) \quad (2.35)$$

Üç matris bileşeni Saklı Markov modelini oluşturmaktadır. Tanımlamalardan yol çıkarak,

$$\gamma_{t,i} = p(y_t = i | x^T) \quad (2.36)$$

$$\Omega_{t,ij} = p(y_t = i, y_{t-1} = j | x^T) \quad (2.37)$$

$$\alpha_{t,i} = p(y_t = i | x^t) \quad (2.38)$$

$$\kappa_t = p(x_t | x^{t-1}) \quad (2.39)$$

$$\beta_{t,i} = \frac{p(x_{t+1}, \dots, x_T | y_t = i)}{P(x_{t+1}, \dots, x_T | x^t)} \quad (2.40)$$

eşitlikleri elde edilmektedir [55].

2.10.3. Viterbi Algoritması

$O = o_1, o_2, \dots, o_T$ gözlem dizisi model $\lambda = (A, B, \pi)$ olmak üzere muhtemel en iyi durum dizisi ortaya konmak istenmektedir. Bazı yaklaşımlarda en iyi durum dizisinin tahmini için tüm durumların ardışık olarak bağlanması önerilmektedir. Fakat bu gibi yaklaşımlar her durumda başarılı sonuçlar üretememektedir.

Viterbi algoritması, gözlem olasılıklarından yararlanarak en iyi durum dizisini tahmin etmeye çalışmaktadır. Durum dizisini maksimum olasılıkları elde ederek sağlamaktadır.

Mevcut i durumunda $t = t$ iken verilen en yüksek olasılıklı kısmi gözlem dizisi ve durum dizisi için yardımcı değişken tanımlanmaktadır.

$$\delta_{(i)} = \max_{q_1, q_2, \dots, q_{t-1}} p\{q_1, q_2, \dots, q_{t-1}, q_t = i, o_1, o_2, \dots, o_{t-1} | \lambda\} \quad (2.41)$$

Buradan,

$$\delta_{t+1}(j) = b_j(o_{t+1}) [\max_{1 \leq i \leq N} \delta_t(i) a_{ij}], \quad 1 \leq i \leq N \quad 1 \leq t \leq T - 1 \quad (2.42)$$

$$\delta_1(j) = \pi_j b_j(o_1), \quad 1 \leq j \leq N \quad (2.43)$$

elde edilmektedir. En iyi muhtemel durum dizisini hesaplamak için hesaplamaya $\delta_T(j)$ $1 \leq j \leq N$ 'den başlanmaktadır. j durumu sonuçları $j = \underset{1 \leq j \leq N}{argmax} \delta_T(j)$ olarak belirtilmektedir. Algoritma bu durumdan diziyi oluşturmaya başlar ve durumlar dizisinin geri izlerindeki her durumu gösterir [94,97].

2.10.4. Baum-Welch Algoritması

Saklı Markov modelinin cevap aradığı temel üç problemten biride O gözlemler olmak üzere $p\{O|\lambda\}$ ' yı maksimize edecek model parametresinin belirlenmesidir. Parametrenin elde edilmesi ile modelin öğrenmesi sağlanarak problem çözümüne destek sağlanmaktadır. Model, öğrenmenin sağlanması için bazı optimizasyon kriterleri kullanmaktadır. Bu kriterlerden biride maksimum olabilirlik kriteridir. Bu kriterde model parametrelerini belirleyebilmek için Baum-Welch algoritmasından yararlanılmaktadır [94,98-99].

Bir w sınıfının Saklı Markov modeli λ_w modelinin parametrelerine karşılık gelen O^w gözlemler dizisinin olasılıklarını maksimize etmek için ifade edilen olabilirlik fonksiyonu,

$$L_{tot} = p\{O^w | \lambda_w\} \quad (2.44)$$

ile ifade edilmektedir. Sadece w sınıfı dikkate alındığı için indisler göz ardı edildiğinde,

$$L_{tot} = p\{O | \lambda\} \quad (2.45)$$

elde edilmektedir. Modelin analitik çözüme kavuşturulmasında iteratif işlemleri içeren Baum-Welch algoritması kullanılmaktadır.

Maksimize işlemlerinde,

$$Q(\lambda, \bar{\lambda}) = \sum_q p\{q|O, \lambda\} \log[p\{O, q, \bar{\lambda}\}] \quad (2.46)$$

$\bar{\lambda}$ için elde edilmektedir. İleri ve geri değişkenleri için iki yardımcı değişken gerekmektedir. Değişkenlerden biri,

$$\xi_t(i, j) = p\{q_t = i, q_{t+1} = j|O, \lambda\} \quad (2.47)$$

olarak ifade edilmektedir. Bu eşitlik ayrıca,

$$\xi_t(i, j) = \frac{p\{q_t = i, q_{t+1} = j, O|\lambda\}}{p\{O|\lambda\}} \quad (2.48)$$

ileri ve geri değişkenlerinin kullanılması sonucunda,

$$\xi_t(i, j) = \frac{\alpha_t(i) a_{ij} \beta_{t+1}(j) b_j(o_{t+1})}{\sum_{i=1}^N \sum_{j=1}^N \alpha_t(i) a_{ij} \beta_{t+1}(j) b_j(o_{t+1})} \quad (2.49)$$

elde edilmektedir. Sonsal olasılık olan ikinci değişken,

$$\gamma_t(i) = p\{q_t = i|O, \lambda\} \quad (2.50)$$

eşitliği ile, ileri ve geri değişkenlerde,

$$\gamma_t(i) = \left[\frac{\alpha_t(i) \beta_t(i)}{\sum_{i=1}^N \alpha_t(i) \beta_t(i)} \right] \quad (2.51)$$

ile ifade edilmektedir. $\gamma_t(i)$ ve $\xi_t(i, j)$ arasındaki ilişki,

$$\gamma_t(i) = \sum_{j=1}^N \xi_t(i, j) \quad 1 \leq i \leq N \quad 1 \leq t \leq M \quad (2.52)$$

ile ifade edilmektedir. Yeniden tahminleme formülleri ile Saklı Markov modeli parametreleri,

$$\bar{\pi}_i = \gamma_1(i) \quad 1 \leq i \leq N \quad (2.53)$$

$$\bar{a}_{ij} = \frac{\sum_{t=1}^{T-1} \xi_t(i, j)}{\sum_{t=1}^{T-1} \gamma_t(i)} \quad 1 \leq i \leq N \quad 1 \leq j \leq N \quad (2.54)$$

$$\bar{b}_j(k) = \frac{\sum_{t=1}^T \gamma_t(j)}{\sum_{t=1}^T \gamma_t(j)} \quad 1 \leq j \leq N \quad 1 \leq k \leq M \quad (2.55)$$

olarak güncellenmektedir [94].

2.10.5. İleri Algoritması

$\lambda = (A, B, \pi)$ parametrelili ve $O = o_1, o_2, \dots, o_T$ gözlem dizisi için $p\{O|\lambda\}$ tahmin edilmek istenmektedir. Bu tahminlerin gerçekleştirilmesi için bazı olasılıksal yöntemler sonuç olarak başarılı değildirler.

İleri değişken $\alpha_t(i)$ ve gözlem dizisi $o_1, o_2, \dots, o_T$ olmak üzere,

$$\alpha_t(i) = p\{o_1, o_2, \dots, o_t, q_t = i|\lambda\} \quad (2.56)$$

ile gösterilebilmektedir. Tekrarlı ilişki,

$$\alpha_{t+1}(j) = b_j(o_{t+1}) \sum_{i=1}^N \alpha_t(i) o_{ij} \quad 1 \leq j \leq N \quad 1 \leq t \leq T-1 \quad (2.57)$$

ve

$$\alpha_1(j) = \pi_j b_j(o_1) \quad 1 \leq j \leq N \quad (2.58)$$

$\alpha_T(i)$ $1 \leq i \leq N$ için yineleme kullanılarak tekrar hesaplanmaktadır. Olasılık,

$$p\{O|\lambda\} = \sum_{i=1}^N \alpha_T(i) \quad (2.59)$$

eşitliği ile elde edilmektedir. Bu yöntem ileri algoritması olarak adlandırılmaktadır. $\beta_t(i)$ geri değişkeni olmak üzere,

$$\beta_t(i) = p\{o_{t+1}, o_{t+2}, \dots, o_T | q_t = i, \lambda\} \quad (2.60)$$

i durumunda geri değişkeni $\beta_t(i)$ ve gözlem dizisi $o_{t+1}, o_{t+2}, \dots, o_T$ olmak üzere,

$$\beta_t(i) = \sum_{j=1}^N \beta_{t+1}(j) a_{ij} b_j(o_{t+1}) \quad 1 \leq i \leq N, \quad 1 \leq t \leq T-1 \quad (2.61)$$

$$\beta_T(i) = 1 \quad 1 \leq i \leq N \quad (2.62)$$

$$\alpha_t(i) \beta_t(i) = p\{O, q_t = i | \lambda\} \quad 1 \leq i \leq N \quad 1 \leq t \leq T \quad (2.63)$$

eşitlikleri elde edilir. Son olarak $p\{O | \lambda\}$ için,

$$p\{O | \lambda\} = \sum_{i=1}^N p\{O, q_t = i | \lambda\} = \sum_{i=1}^N \alpha_t(i) \beta_t(i) \quad (2.64)$$

ile geri algoritması ifade edilmektedir [94].

2.11. Bayes Tabanlı Sınıflandırıcılar

Bayes tabanlı sınıflandırıcılar önsel olasılık aracılığı ile sonsal olasılığı tahmin etmeyi amaçlayan istatistiksel bir sınıflama yöntemidir. Bayes tabanlı sınıflandırıcıların amacı belirli bir sınıfa ait üyelik olasılığını belirlemektir.

Bayes tabanlı sınıflandırıcılar Bayes teorisini esas alarak geliştirilmiştir. Naive Bayes sınıflandırıcısı olarak bilinen Bayes tabanlı sınıflandırıcılar veritabanlarında ve bilgi sistemlerinde diğer sınıflama yöntemleri gibi yaygın bir şekilde kullanılabilmektedir.

Naive Bayes sınıflandırıcısı verilen bir sınıfın özellik değerinin etkilerinin diğer özellik değerlerinden bağımsız olduğunu varsaymaktadır. Bu duruma sınıfın koşullu bağımsızlığı adı verilmektedir. Bayes ağları, Naive Bayes sınıflandırıcılardan farklıdır. Bu ağlar özelliklerin alt kümeleri boyunca bağımlılıkları gösteren grafiksel modellerdir. Bayes ağları, Naive Bayes sınıflandırıcıları gibi sınıflandırmada kullanılabilmektedir.

2.11.1. Bayes Teorisi

Bayes teorisi adını teoremi geliştiren 18. Yüzyılda karar teorisi ve olasılık üzerine çalışan Thomas Bayes tarafından öne sürülmüştür.

$P\{H|X\}$ X olasılığına bağılı olarak H olasılığını belirlemeye çalışan sonsal olasılık olarak adlandırılmaktadır. $P(H)$ H'nin önsel olasılığı olarak ifade edilmektedir. $P(X)$ ise X'in önsel olasılığı ile ifade edilmektedir. Buna göre $P(X|H)$ H koşulunda X'in sonsal olasılığı olarak ifade edilmektedir.

Bayes teoremi,

$$P(H|X) = \frac{P(X|H)P(H)}{P(X)} \quad (2.65)$$

eşitliği ile ifade edilmektedir [57].

2.11.2. Naive Bayes Sınıflandırıcılar

1. D , sınıf etiketleri ile bağlantılı eğitim kümesi olmak üzere her küme, n boyutlu özellik vektörü $X = (x_1, x_2, \dots, x_n)$ ile temsil edilsin. n sayıda ölçüm n özellikten üretilen kümeler sırası ile $A_1, A_2, \dots, A_n$ ile gösterilmektedir.

2. $C_1, C_2, \dots, C_m$ olmak üzere X veri seti verildiğinde, sınıflama X koşulunda X 'e ait olan en yüksek sonsal olasılığı tahmin etmektedir. Naive Bayes sınıflayıcı C_i sınıfı için tahmini,

$$P(C_i|X) > P(C_j|X) \quad 1 \leq j \leq m \quad j \neq i \quad (2.66)$$

ile ifade edilmektedir. $P(C_i|X)$ olasılığı Bayes teoremi kullanılarak

$$P(C_i|X) = \frac{P(X|C_i)P(C_i)}{P(X)} \quad (2.67)$$

eşitliği ile elde edilmektedir.

3. $P(X)$ tüm sınıflar için sabit olup sadece $P(X|C_i)P(C_i)$ için maksimize edilmektedir. Eğer sınıfın önsel olasılığı bilinmiyorsa tüm sınıf olasılıkları $P(C_1) = P(C_2) = P(C_m)$ eşit varsayılmaktadır. $P(X|C_i)$ bu durumda maksimize edilir. $|C_{i,D}|$ D'deki sınıfların eğitim kümesi sayısı olmak üzere sınıfların önsel olasılığı,

$$P(C_i) = \frac{|C_{i,D}|}{|D|} \quad (2.68)$$

ile tahmin edilmektedir.

4. Bazı özellikler ile verilen veri kümesi, $P(X|C_i)$ 'nin hesaplanması için pahalı olabilmektedir. $P(X|C_i)$ değerlendirmelerinin hesaplamalarını azaltmak için koşullu sınıf bağımsızlığına ilişkin Naive varsayımı elde edilmektedir. Bu varsayım özellik değerlerinin diğer koşullu bağımsızlığıdır. Böylece,

$$P(X|C_i) = \prod_{k=1}^n P(X_k|C_i) = P(x_1|C_i) * P(x_2|C_i) * ... * P(x_n|C_i) \quad (2.69)$$

eşitliği ile ifade edilmektedir. Eğitim veri kümelerinden $P(x_1|C_i), P(x_2|C_i), \dots, P(x_n|C_i)$ olasılıkları tahmin edilebilmektedir. x_k X kümesinin A_k özellik değerini belirtmektedir. Her özelliğin değerinin sürekli veya kesikli olup olmadığına dikkat edilmelidir.

$P(X|C_i)$ için A_k kesikli olmak üzere D'deki C_i sınıfının küme sayıları $P(x_k|C_i)$ ve A_k için x_k değerine sahiptir. A_k sürekli olmak üzere ortalaması μ ve standart sapması σ olan eşitlik,

$$g(x, \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad (2.70)$$

$$P(x_k|C_i) = g(x_k, \mu_{C_i}, \sigma_{C_i}) \quad (2.71)$$

ile ifade edilmektedir.

5. X'in sınıf etiketlerini tahmin edebilmek için $P(X|C_i)P(C_i)$ her C_i sınıfı için değerlendirilir. Sınıflayıcı C_i sınıfının tahmini için,

$$P(X|C_i)P(C_i) > P(X|C_j)P(C_j) \quad 1 \leq j \leq m \quad j \neq i \quad (2.72)$$

eşitsizliği kullanılır ve her $P(X|C_i)P(C_i)$ için sınıf etiketi tahmincisi C_i sınıfı maksimumdur [57].

2.11.3. Bayes Ağları

Naive Bayes sınıflandırıcı verilen veri kümesinin sınıf etiketi için özellik değerlerini birbirinden koşullu bağımsız olduklarını varsayarak sınıfların koşullu bağımsızlık varsayımlarını gerçekleştirmektedir. Naive Bayes sınıflandırıcısı uygulama açısından sade bir yöntemdir. Varsayımlar doğru olduğunda, Naive Bayes sınıflandırıcı birçok sınıflandırıcıdan daha başarılı sonuçlar üretmektedir. Değişkenler arasında bağımlılıklar görülebilmektedir. Bayes ağları ortak koşullu olasılık dağılımlarını belirlemektedir. Bayes ağları sınıfların koşullu bağımsızlıklarını değişkenlerin alt kümeleri arasında tanımlanmasına olanak tanımaktadır. Bayes ağları ilişkiler için grafiksel bir model sağlamaktadır. Öte yandan Bayes ağları sınıflama için başarılı bir şekilde kullanılabilir.

Bayes ağları iki bileşenden oluşmaktadır. Yönlü döngüsel olmayan grafikler ve koşullu olasılık tablosu Bayes ağlarının iki temel bileşenini teşkil etmektedir. Yönlü döngüsel olmayan grafikteki her nodda rasgele değişkenler belirtilmektedir. Bu değişkenler sürekli veya kesikli parametrelili olabilir. Bu değişkenler ilişkiyi düzeltmek için veri kümesi veya gizli değişkenler olarak gerçek özelliklere karşılık gelebilir.

$X = (x_1, \dots, x_n)$ veri seti sırası ile $Y_1, Y_2, \dots, Y_n$ değişkenleri veya özellikleri tarafından tanımlanmaktadır. Değişkenlerin ortak olasılık dağılımının tanımlanması için,

$$P(x_1, x_2, \dots, x_n) = \prod_{i=1}^n P(x_i | \text{Ata}(Y_i)) \quad (2.73)$$

eşitliği ile birlikte ağ sağlanmaktadır. $P(x_1, \dots, x_n)$, X değerlerinin özel kombinasyonlarının olasılıkları ve $P(x_i | \text{Ata}(Y_i))$ Y_i için koşullu olasılık tablosuna girişleri ifade etmektedir.

Çok sayıda algoritma Bayes ağlarına uygulanabilmektedir. Bu algoritmalar sınıflama işlemlerinde başarılı sonuçlar üretmektedir [57].

3.UYGULAMA

3.1. Modelin Geliştirilmesi

Bilgisayar ortamında bulunan metin belgelerinin yoğunluğu ve boyutları dikkate alındığında, her birini incelemenin çok güç olması nedeniyle anlamlı bilgi çıkarımı yapabilen, metindeki kelimeleri sınıflandırabilen ve kompleks yapılarda enformasyon sağlayabilen başarılı bir model geliştirilmesini zorunlu hale getirmiştir. Çalışmada Türkçe Dili'nde hazırlanmış metin belgelerinden enformasyon sağlanması amaçlanmıştır. Bu nedenle geliştirilen modelin Türkçe Dil özelliklerine göre tasarlanması önem taşımaktadır. Model ile incelenmesi düşünülen tüm metinler doğal dil ile yazılmış olacağı dikkate alındığından, Türkçe Dil yapısı ve kuralları modelin geliştirilmesinde etkili olmaktadır.

3.1.1. Türkçe 'de Kelimeler ve Cümle Dizilimi

Türkçe, dünyada en fazla konuşulan diller arasındadır. Türkçe içeriği bakımından zengin bir dildir. Yapısal olarak sondan eklemeli bir dildir. Bu nedenle bir kelimenin sonuna ekler gelerek yeni anlamlar kazandırabildiği gibi sadece kelimenin türünü de değiştirebilmektedir. Bu nedenle kelimeye gelen ek, kelimeyi oldukça etkileyebilmektedir.

Türkçe 'de isim, sıfat, zamir, zarf, edat, bağlaç, fiil ve ünlem olmak üzere sekiz kelime türü bulunmaktadır [106,107]. Türkçe' de cümle dizilimi ise çoğunlukla özne, nesne, yüklem biçiminde sıralanmaktadır.

3.1.2. İşlem Aşamaları

Geliştirilen model dört ana bölüm altında, aşamalı işlemler dizisinden oluşmaktadır.

Birinci bölüm: Kelimelerin anlam değiştirmeyen eklerinden ayrılması.

İkinci bölüm: Saklı Markov modeli ve Bayes tabanlı sınıflandırıcılar ile kelimelerin sınıflandırılması.

Üçüncü bölüm: Varlık ilişki modeli ile enformasyon çıkarım alanlarının düzenlenmesi.

Dördüncü bölüm: Verilerin veritabanına aktarılarak enformasyon çıkarımının gerçekleştirilmesi.

3.1.3. Kelimelerin Anlam Değiştirmeyen Eklerinden Ayrılması

Modelin ilk bölümünü oluşturan kelimelerin ek harflerden ayrılması 7 aşamada gerçekleştirilmektedir. Bu bölümde modelin doğru ilerlemesini sağlayacak en önemli aşama, kelimelerin hangi eklerden nasıl ayrılacağına ilişkindir. Kelimelerin eklerden ayrılma işlemi bir algoritma ile gerçekleştirilmektedir. Bu algoritma kelimeleri öncelikle tüm ekleriyle beraber işleme almaktadır. İşleme alınan kelime, kelimedeki anlamı değiştirmeyen eklerinden arındırılıncaya kadar döngü devam etmektedir. Kelimelerin kökleri ve kelimenin anlamını değiştiren ekleri, açık kaynak kodlu Türkçe doğal dil işleme kütüphanesi Zemberek 'ten [108] ve dil uzmanlarından destek alınarak belirlenmiştir.

Modelde kelimelerin yalnız anlam değiştirmeyen eklerinden arındırılması yerine, tüm eklerinden arındırılarak sadece köklerin işleme alınmamasının nedeni, Türkçe 'de kök halinde bulunan eş anlamlı kelimelerin oldukça fazla olmasından kaynaklanmaktadır. Buda yanlış enformasyona yol açacaktır. Örneğin “ilerleme” kelimesinin kökü “il” kelimesidir. Eğer kelime, model için tüm eklerinden arındırılmış olsaydı, şehir, kent anlamı taşıyan “il” sözcüğü dikkate alınacak ve yanlış enformasyona neden olacaktı. Böylelikle, sadece kelimenin anlamını değiştirmeyen eklerin arındırılmasıyla “ilerleme” kelimesinden “ilerle” kelimesi elde edilerek anlam kaybına uğranmayacaktır.

Modelde kelime türlerinin sınıflandırılması için iki kategorili yöntem benimsenmiştir. Türkçe Dili'nde sekiz farklı kelime türü bulunmaktadır. Fakat bazı kaynaklarda, bu kelime türleri üç grup altında veya iki grup altında incelenmektedir. Modelde kelimelerin aldıkları anlam değiştirmeyen ekler dikkate alınarak kelime türleri iki kategoride incelenmiştir. Bu kategoriler;

İsim grupları: İsim, sıfat, zarf, zamir, edat, bağlaç, ünlem

Fiil Grupları: Fiil

Modelin bu bölümündeki işlem süreci, 7 aşamada ifade edilmektedir.

1. Enformasyon çıkarımı yapılması hedeflenen metin belgesi ve metine ilişkin kelimeler getirilir. Metin belgesi web ortamından elde edilebilir veya farklı dosya türlerinde bilgisayarlarda kayıtlı tutulan belgelerden elde edilebilir.

2. Metin belgesinde bulunan tüm kelimeler, metin belgesi içerisinde kullanım sırasına göre sıralanır ve liste halinde sunulur. Metin belgesindeki sayı ile ifade edilen tarih, zaman gibi sayısal değerler de işlemlere dahil edilir.

3. Sıralanan ve liste halinde bulunan metin belgesindeki her kelime algoritmaya tabi tutularak ilgili kelimedeki anlam deęiřtirmeyen tüm eklerinden ayrıştırılırlar. Algoritma sonucunda her kelimededen ayrılan eklerin toplam harf sayıları tespit edilir.

4. Metin belgesindeki tüm kelimeler ve algoritma sonucu hesaplanan tüm kelimelerin ayrıştırılan harf sayıları listelenir.

5. Sıralanan ve liste halinde bulunan eklerinden ayrılmış kelimeler, kelime türüne göre isim grupları veya fiil grupları kategorisine atanmaktadır. Kelimelere ilişkin kategoriler sıra ile listelenir. Kelime türlerinin grupları Tablo 3.1.'de belirtilmiştir.

Tablo 3.1. Kelime türlerinin gruplandırılması

Kelime Türü	Kelime Türü Grupları
İsim, Sıfat, Zamir, Zarf, Edat, Ünlem, Bağlaç	İsim Grubu
Fiil	Fiil Grubu

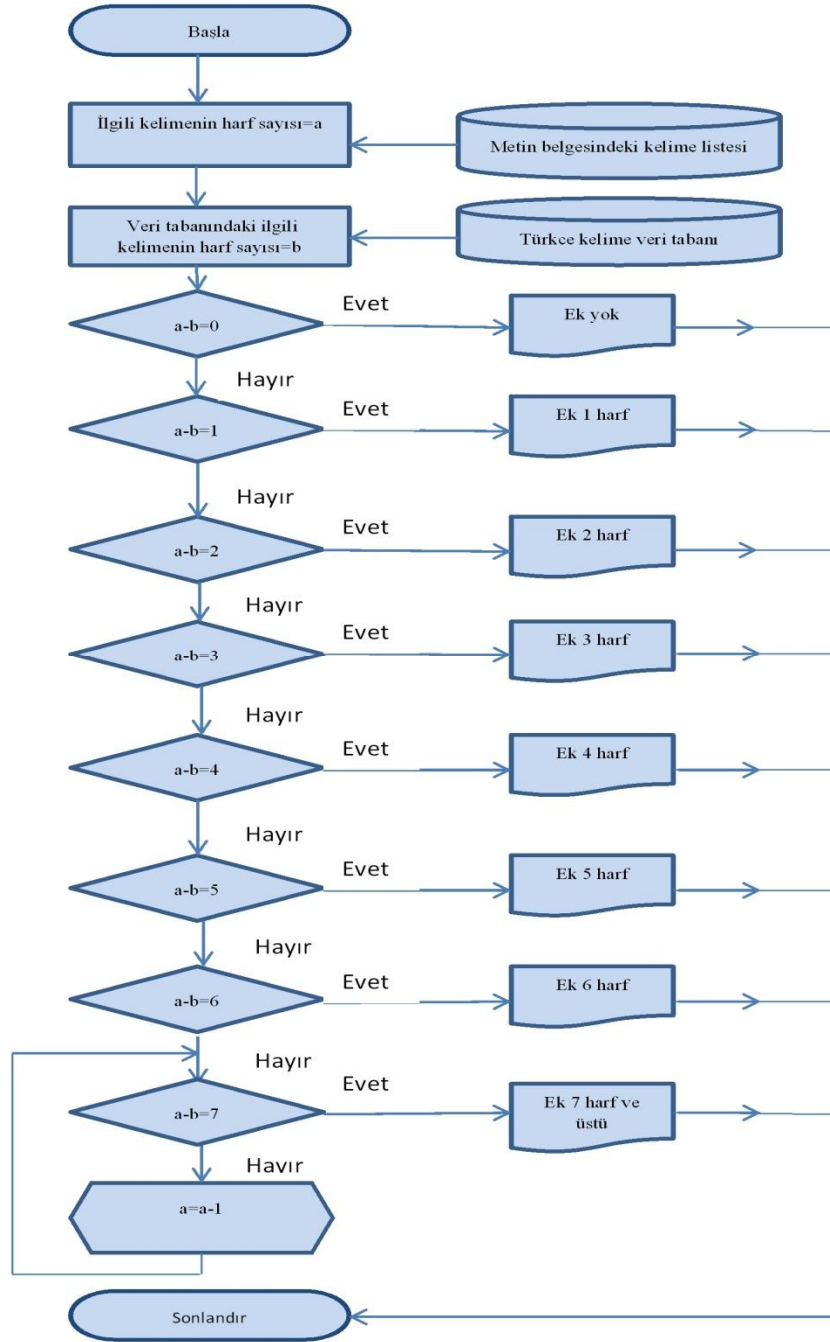
6. Algoritma ile her kelimenin anlam deęiřtirmeyen eklerinin harf sayıları 3. aşamada tespit edilmiştir. Bu verilere göre eklerin harf sayıları kategorize edilmiştir. Kelimelerin almış oldukları anlam deęiřtirmeyen eklerin harf sayılarına göre ek harf gruplarından biri atanır ve listelenir. Anlam deęiřtirmeyen eklerin harf sayılarına göre ek harf grupları Tablo 3.2.'de verilmiştir.

Tablo 3.2. Ek harf sayılarının gruplandırılması

Kelimelerin Anlam Deęiřtirmeyen Eklerinin Harf Sayıları	Ek Harf Grupları
0	Ek harf yok
1	Ek 1 harf
2	Ek 2 harf
3	Ek 3 harf
4	Ek 4 harf
5	Ek 5 harf
6	Ek 6 harf
7+	Ek 7 harf ve üstü

7. Metinde kullanılan kelimeler için kelime numarası atanır ve listelenir. Metin içerisinde aynı kelime birden fazla yerde kullanılmış olabilir. Aynı kelimeler için tek kelime numarası kullanılmaktadır.

Bu işlemler sonucunda metin içerisindeki kelimeler için kelime sıra numarası, kelime türü grupları, kelime ek harf grupları ve kelime numarası değişkenleri elde edilmektedir. Kelimeleri eklerinden ayırma ve eklerin harf sayılarının belirlenerek kategorize edilme işlemleri için geliştirilen algoritma Şekil 3.1.'de belirtilmiştir.



Şekil 3.1. Ekleri ayırma ve ek harflerin sayılarak gruplara ayrılmasına ilişkin algoritma

3.1.4. Saklı Markov Modeli ve Bayes Tabanlı Sınıflandırıcılar ile Kelimelerin Sınıflandırılması

Modelin ikinci bölümünde kelimelerin stokastik modeller ile kelime türü gruplarının tahmin edilmesi amaçlanmaktadır. Kelimelerin sınıflandırılmasında Saklı Markov modeli, Baum-Welch algoritması ve Bayes sınıflandırma yöntemleri kullanılmaktadır. Saklı Markov modeli gözlem dizileri aracılığı ile gözlenemeyen gizli durumların geçiş olasılıklarını belirlemede kullanılmaktadır. Model için belirlenen gözlem dizileri kelimelerin almış oldukları anlam değiştirmeyen eklerinin harf sayılarıdır. Modelin ilk bölümünde kelimelerin almış oldukları anlam değiştirmeyen eklerinin harf sayıları “ek harf yok”, “bir ek harf”, “iki ek harf”, “üç ek harf”, “dört ek harf”, “beş ek harf”, “altı ek harf” ve “yedi ve üstü ek harf” olarak 8 grupta kategorize edilmiştir. Metin belgesindeki her kelime için bir gözlem bulunmaktadır. Bu gözlemler aracılığı ile gizli durumlara ilişkin olasılıklar elde edilmesi amaçlanmıştır. Modeldeki gizli durumlar ise kelimelerin türleridir. Bilgisayarların doğrudan bir kelimenin hangi türde olduğunu bilmesi mümkün değildir, bu nedenle gizli olan kelime türü gözlemler aracılığı ile tahmin edilmektedir. Tablo 3.3.’te gözlemler ve durumlara ilişkin bilgi verilmiştir.

Modelde Saklı Markov modelinin sağladığı en büyük avantaj metin içerisindeki ve cümle içerisindeki kelimelerin dizilişini tahmin etmede kullanılabilmesidir. Geliştirilen bu model, metin içerisindeki kelimelerin dizilişi ve kelime türlerinin dizilişini ortaya koyarak elde etmektedir. Bunu yaparken ise gözlemlerden yola çıkmaktadır. Örnek ile belirtmek gerekirse, çok sayıda ek harf alan bir kelimenin türü, Saklı Markov modeli ile “fiil grubu” olarak tahmin edildiği varsayalım. Kelime türü “fiil grubu” olan bir kelimedenden sonra metin içerisinde hangi kelime türünün gelebileceği geçiş olasılıkları matrisi aracılığı ile tahmin edilebilmektedir. Bu şekilde metin içerisindeki kelimelerin dizilişi hakkında önemli tahminler sağlanabilmektedir.

Tablo 3.3. Durum sembolleri ve gözlem sembolleri

Durum Sembolü	S ₁				S ₂			
Durumlar	İsim Grupları				Fiil Grupları			
Gözlem Sembolü	O ₁	O ₂	O ₃	O ₄	O ₅	O ₆	O ₇	O ₈
Gözlemler	Ek harf yok	1 ek harf	2 ek harf	3 ek harf	4 ek harf	5 ek harf	6 ek harf	7 ve üstü ek harf

Modelin ikinci bölümü olan Saklı Markov modeli ve Bayes tabanlı sınıflandırıcılar ile kelimelerin sınıflandırılması, belirli işlem adımlarını takip etmektedir. Bu adımları kısaca şu şekilde ifade etmek mümkündür.

1. Başlangıç olasılıkları vektörü oluşturulur.
 2. Durumlara ilişkin geçiş olasılıkları matrisi oluşturulur.
 3. Gözlemlere ilişkin emisyon olasılıkları matrisi oluşturulur.
 4. Saklı Markov modeli oluşturulduktan sonra Baum-Welch algoritması kullanılarak modelin parametre değerleri tahmin edilir.
 5. Her kelime numarasına göre Baum-Welch algoritması ile elde edilen parametre değerleri kullanılarak kelimelerin modele göre kelime türlerini belirlemek üzere kelime numarası bulunan kelimelerin isim grubu Bayes olasılığı ve fiil grubu Bayes olasılığı belirlenir.
 6. Bayes sınıflandırıcılar ile isim grubu Bayes olasılığı ve fiil grubu Bayes olasılığı belirlenen kelimelerde, Bayes olasılığı yüksek olan kelime türü grubu seçilir. Seçilen kelime türü grubu ilgili kelimenin türü olarak ataması yapılır ve böylece kelime türü tahmin edilir.
 7. Her kelime numarası için atanan kelime türleri listelenir.
- Bu işlemler sonucunda Saklı Markov modeli parametreleri ile kelimelerin kelime türleri tahmin edilmektedir.

3.1.5. Varlık İlişki Modeli ile Enformasyon Çıkarım Alanlarının Düzenlenmesi

Modelin üçüncü bölümü enformasyon çıkarmaya yönelik işlem aşamalarını içermektedir. Bu bölümde metin yığınının metine ilişkin enformasyon çıkarımında bulunacak alanlar belirlenerek tüm kelime ve kelimelere ilişkin verileri, veritabanına aktarımına hazır hale getirilmesi amaçlanmaktadır. Enformasyon çıkarımı yapılacak alan sayısının belirlenmesinde, elde edilmesi istenen enformasyona ait kelime sayısının metindeki kelime sayısına oranı esas alınmıştır. Bu oran, enformasyonun metni iyi bir temsil yeteneğine sahip olması ve minimum kelime sayısı ile enformasyonun metni ifade etmesi amaçlanarak belirlenmiştir. Enformasyondaki kelime sayısının metindeki kelime sayısına oranı yaklaşık 1/50 olarak dikkate alınmıştır. Metindeki yaklaşık her 50 kelimeyi 1 enformasyon kelimesi temsil etmektedir. 1/50 oranının elde edilmesi için enformasyon

alan sayısının belirlenmesi adına simülasyonlar gerçekleştirilmiştir. Yapılan simülasyon denemelerinde enformasyon kelime sayısı/metin kelime sayısı oranı 1/50'nin yaklaşık olarak elde edilebilmesi için enformasyon alan sayısının belirlenmesinde;

$$\text{Enformasyon Alan sayısı} = \frac{\text{Metindeki kelime sayısı}}{125} \quad (3.1)$$

eşitliği elde edilmiştir.

Bu işlem aşamaları kısaca ifade edilmek istenirse;

1. Metindeki toplam kelime sayısı 125'e bölünerek enformasyon çıkarımı yapılacak alan sayısı belirlenir.
2. Enformasyon çıkarımı yapılacak kelime sayısı kadar metinde en çok kullanılan kelimeler belirlenir ve bu kelimeler enformasyon çıkarım alanları olarak atanır.
3. Metinde cümle içerisinde enformasyon çıkarım alanları ile birlikte kullanılan tüm kelimeler ilgili enformasyon çıkarım alanı altında listelenir.
4. Metindeki tüm cümlelere *cümle_no* ataması gerçekleştirilir ve her kelimenin bağlı bulunduğu *cümle_no*'ları listelenir.
5. Cümle içerisindeki kelimelere cümledeki diziliş sırasına göre *cümle_içi_sıra_no* ataması yapılır ve her kelimenin cümle içindeki sıra numarası listelenir.
6. Tüm kelime numaralarına ait isim grubu Bayes olasılıkları ve fiil grubu Bayes olasılıkları listelenir.

Bu bölümdeki işlemler sonucunda enformasyon çıkarım alanları belirlenir ve böylece kelimelere ilişkin tüm veriler veritabanına aktarılmaya hazır konuma getirilir.

3.1.6. Verilerin Veritabanına Aktarılarak Enformasyon Çıkarımının Gerçekleştirilmesi

Geliştirilen modelin son bölümünde veritabanı üzerinde varlık ilişki modeli ile çalışılmış ilgili enformasyonların elde edilebilmesi için yapısal sorgulama dili (SQL) kullanılmıştır. Sorguların ardından enformasyonlar Türkçe yapısına uygun bir şekilde düzenlenerek ilgili enformasyonun sağlanması hedeflenmiştir. Bu bölümün işlem adımları, aşağıda gösterildiği gibi sıralanır.

1. Metin belgesindeki kelimelerden elde edilen tüm veriler varlık ilişki diyagramında modellenerek Access veritabanına aktarılır.

2. Cümle içerisinde en az iki enformasyon çıkarım alanı bulunan cümleler yapısal sorgulama dili (SQL) ile belirlenir.

3. İlgili cümledeki enformasyon çıkarım alanları cümle içerisindeki kullanım sırasına göre dizilir.

4. İlgili cümlede fiil grubu olarak atanan kelimelerden en yüksek Bayes olasılığına sahip kelime, dizilmiş olan enformasyon çıkarım kelimelerinin sonuna getirilir. Eğer en yüksek Bayes olasılığına sahip birden fazla kelime var ise cümle içerisinde en son kullanılan en yüksek Bayes olasılığına sahip fiil grubu kelimesi seçilir ve yeni bir cümle elde edilir.

5. Elde edilen yeni cümlelerin fiil grubu kelimesinin sonuna öğrenilen geçmiş zaman ve bildirme eki getirilir.

6. Elde edilen cümleler art arda dizilerek enformasyon oluşturulur.

4.BULGULAR

Geliştirilen model, Türkçe metin belgelerinde enformasyon çıkarımında ve sınıflandırmada başarısını test etmek amacıyla gerçek bir metin belgesi üzerinde test edilmiştir. Test edilmek üzere seçilen metin, web ortamından elde edilmiştir. Metin, web ortamında vikipedi web sayfasından seçilmiş olup, Elazığ ilinin tanıtımına ilişkin; Etimoloji, iklim ve ekonomi başlıkları altındaki metin içeriğinden elde edilmiştir [109]. Metin belgesi 584 kelimeden oluşmaktadır.

Metin belgesindeki kelimelerin kökleri ve eklerinin değerlendirmeleri ile incelemeleri açık kaynak kodlu Türkçe doğal dil işleme kütüphanesi olan Zemberek [108] ve Türk Dili alanındaki uzmanlardan destek alınarak gerçekleştirilmiştir.

Metin belgesindeki kelimelerin Saklı Markov modeli, Viterbi ve Baum-Welch algoritmaları ile değerlendirmelerinde R 2.15.3 programından yararlanılmıştır. Ayrıca Bayes olasılıklarının elde edilmesinde MS Excel Programından yararlanılmıştır.

Elde edilen verilerin varlık ilişki diyagramında incelenmesi ve yapısal sorgulama dili (SQL) ile sorgulanmasında MS Access veritabanından yararlanılmıştır. Metin belgesinin incelenmesi, analizi ve enformasyon çıkarımında modeldeki işlem aşamaları takip edilmiştir.

4.1. Kelimleri Anlam Değiştirmeyen Eklerden Ayırma İşlemlerinin Metin Belgesine Uygulanması

Uygulamada ilk olarak, geliştirilen modelin kelimelerin anlam değiştirmeyen eklerinden ayrılması bölümündeki işlem adımları uygulanmıştır. Uygulamada web sitesinde bulunan metin içeriği getirilerek metindeki kelimeler, diziliş sırasına göre listelenmiştir. Listelenen kelimelere, kelime sıra numarası atanmıştır. Tablo 4.1.'de metin belgesinden elde edilen ve sıralanan kelimeler ile kelime sıra numaraları gösterilmiştir.

Tablo 4.1. Metin belgesindeki kelimeler ve kelime sıra numaraları

Kelime Sıra No	Kelimeler
1	Elazığ
2	ili
.	.
.	.
.	.
292	yaşanan
293	iklimin
.	.
.	.
.	.
583	sanayi
584	kuruluşlardır

Kelimelerin sıralanmasının ardından her kelime, ayrı ayrı sıra ile algoritmaya alınmıştır. Algoritma 584 defa çalışmıştır. Böylece 584 kelimenin ek harf grupları belirlenmiştir. Kelime türü grupları, her kelime için belirlenmiş ve listelenmiştir. Algoritmanın çalışması sonucunda elde edilen ek harf grupları listelenmiştir. Metin içerisindeki kelimelere, kelime numarası atanmıştır. Bu bölümdeki işlemler sonucunda elde edilen veriler Tablo 4.2.'de belirtilmiştir.

Tablo 4.2. Birinci bölümdeki işlemlerin metin belgesine uygulanması ile elde edilen veriler

Kelime Sıra No	Kelimeler	Kelime No	Ek Harf Grubu	Kelime Türü Grubu
1	Elazığ	1	Ek harf yok	İsim grubu
2	ili	2	Ek 1 harf	İsim grubu
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
292	yaşanan	180	Ek 3 harf	Fiil grubu
293	iklimin	142	Ek 2 harf	İsim grubu
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
583	sanayi	195	Ek yok	İsim Grubu
584	kuruluşlardır	344	Ek 6 harf	İsim Grubu

4.2. Saklı Markov Modeli ve Bayes Tabanlı Sınıflandırıcılar ile Kelimeleri Sınıflandırma İşlemlerinin Metin Belgesine Uygulanması

Metin belgesine ilişkin Saklı Markov modelinin oluşturulmasında R 2.15.3 programından yararlanılmıştır. Modelin oluşturulmasında kodlamada kolaylık sağlaması açısından durumlar ve gözlemlerin isimleri kısa semboller ile belirtilmiştir. Kodlamada kullanılan semboller Tablo 4.3.'de belirtilmiştir.

Tablo 4.3. Saklı Markov modelinde durum ve gözlem sembollerinin temsili

Durum Sembolü	S ₁			S ₂				
Durumlar	İsim grupları			Fiil Grupları				
Kodlamada Kullanılan Sembol	İsim			Fiil				
Gözlem Sembolü	O ₁	O ₂	O ₃	O ₄	O ₅	O ₆	O ₇	O ₈
Gözlemler	Ek harf yok	Ek 1 harf	Ek 2 harf	Ek 3 harf	Ek 4 harf	Ek 5 harf	Ek 6 harf	Ek 7 harf ve üstü
Kodlamada Kullanılan Sembol	A	B	C	C	E	F	G	H

R programında Saklı Markov modelinin parametrelerinin oluşturulması için Tablo 4.4.'deki kodlar yazılarak çalıştırılmıştır.

Tablo 4.4. Saklı Markov modeli için düzenlenen R kodları

```
hmm=initHMM(c("isim","fiil"),c("A","B","C","D","E","F","G","H"),  
c(.85,.15),matrix(c(.84,.89,.16,.11),2),matrix(c(.61,.02,.08,.02,.10,.18,.09,.14,.07,.23,.04,.05,.002,.25,.008,.11),2))  
  
print(hmm)
```

Modelin ikinci bölümünü oluşturan kelimelerin stokastik yöntemler ile sınıflandırılması işlemlerinde ilk olarak metindeki kelimelere ilişkin kelime türleri ile başlangıç olasılıkları vektörü elde edilmiştir. Başlangıç olasılıkları vektörü, kelime türü grupları olasılıkları ile sağlanmıştır. Elde edilen başlangıç olasılıkları vektörü Tablo 4.5.'te belirtilmiştir.

Tablo 4.5. Başlangıç olasılıkları vektörü

	S ₁	S ₂
p	0.85	0.15

Saklı Markov modelinin parametrelerinin biride durumlar arası geçiş olasılıkları matrisidir. Durumlar arası geçiş olasılıkları matrisi ile bir kelime türü grubundan diğer kelime türü gruplarına geçiş olasılıkları elde edilmiştir. Metindeki kelime gruplarına ilişkin geçiş olasılıkları matrisi Tablo 4.6.'da belirtilmiştir. Matrise göre isim grubundaki bir kelimeden sonra yine isim türü bir kelimenin gelme olasılığı 0.84, fiil grubu bir kelime gelme olasılığı ise 0.16 olarak elde edilmiştir. Fiil grubu bir kelimeden sonra isim grubu kelime gelme olasılığı 0.89, fiil grubu bir kelime gelme olasılığı ise 0.11 olarak elde edilmiştir.

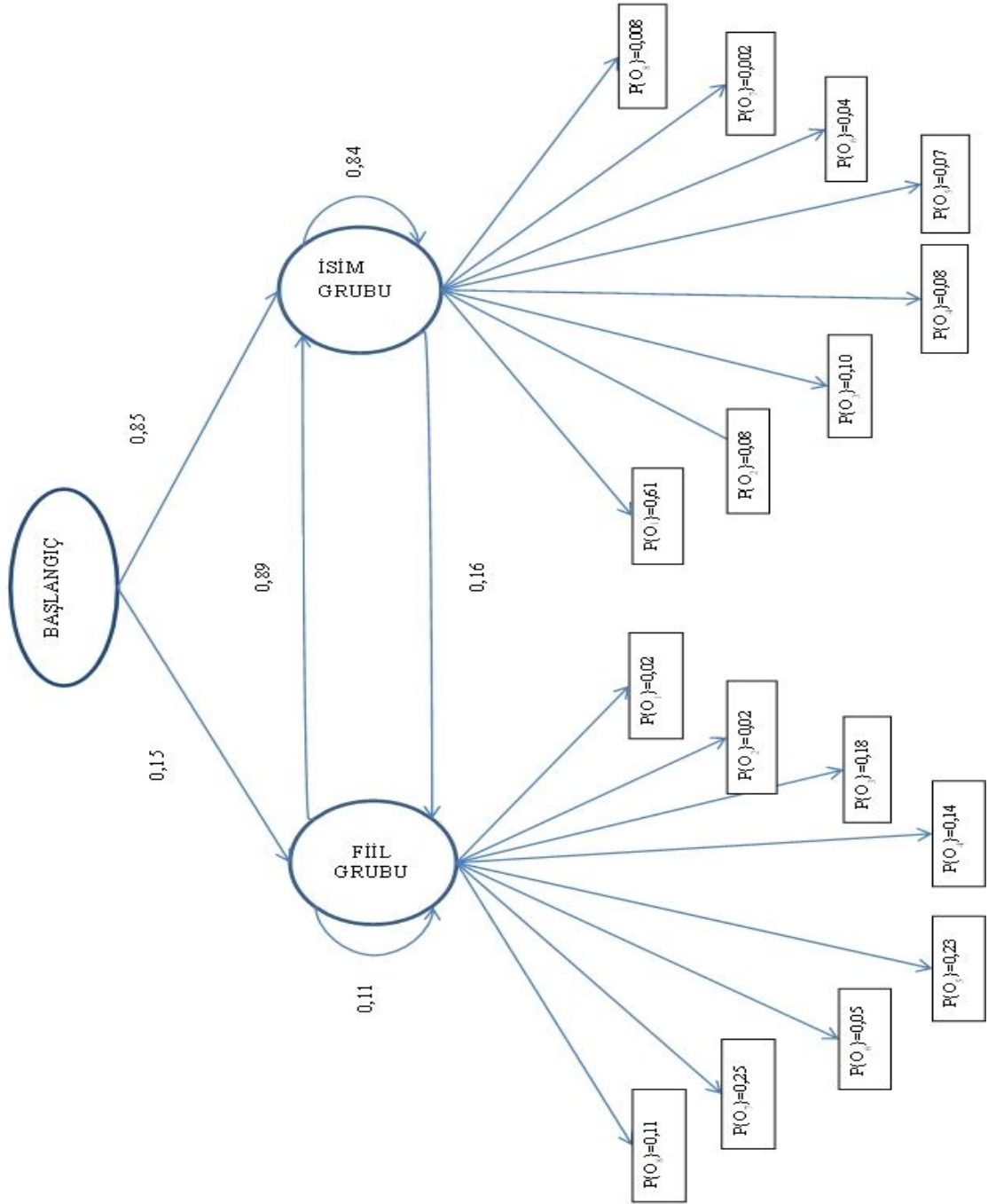
Tablo 4.6. Durumlar arası geçiş olasılıkları matrisi

	S ₁	S ₂
S ₁	0.84	0.16
S ₂	0.89	0.11

Gözlemlerin durumlara göre olasılıksal dağılımı emisyon olasılıkları matrisi ile elde edilmiştir. Emisyon olasılıkları matrisi Tablo 4.7.'de belirtilmiştir. Emisyon olasılıklarına göre hiç ek harf almayan bir kelime 0.61 olasılık ile kelime türü isim grubu olacağı tahmin edilmektedir. 6 ek harf alan bir kelimenin 0.25 olasılık ile kelime türü fiil grubu olacağı tahmin edilmektedir.

Tablo 4.7. Emisyon olasılıkları matrisi

	O ₁	O ₂	O ₃	O ₄	O ₅	O ₆	O ₇	O ₈
S ₁	0.61	0.08	0.10	0.09	0.07	0.04	0.002	0.008
S ₂	0.02	0.02	0.18	0.14	0.23	0.05	0.25	0.11



Şekil 4.1. Saklı Markov modeline ilişkin diyagram

Saklı Markov modeline ilişkin diyagram Şekil 4.1.'de belirtilmiştir. Saklı Markov model parametrelerinin elde edilmesinin ardından, model parametre tahminlerini gerçekleyecek algoritma olan Baum-Welch algoritması çalıştırılmıştır. Baum-Welch algoritmasının oluşturulması ve parametre tahminlerinin elde edilmesi için Tablo 4.8.'deki kodlar yazılarak çalıştırılmıştır.

Tablo 4.8. Baum-Welch algoritması için düzenlenen R kodları

```
hmm=initHMM(c("isim","fiil"),c("A","B","C","D","E","F","G","H"),c(.85,.15),matrix(c(.84,.89,.16,.11),2),matrix(c(.61,.02,.08,.02,.10,.18,.09,.14,.07,.23,.04,.05,.002,.25,.008,.11),2))

print(hmm)

x=sample(c(rep("A",800),rep("B",700),rep("C",600),rep("D",500),rep("E",400),rep("F",300),rep("G",200),rep("H",100)))

y=sample(c(rep("A",700),rep("B",600),rep("C",500),rep("D",400),rep("E",300),rep("F",200),rep("G",100),rep("H",800)))

z=sample(c(rep("A",600),rep("B",500),rep("C",400),rep("D",300),rep("E",200),rep("F",100),rep("G",800),rep("H",700)))

v=sample(c(rep("A",500),rep("B",400),rep("C",300),rep("D",200),rep("E",100),rep("F",800),rep("G",700),rep("H",600)))

t=sample(c(rep("A",400),rep("B",300),rep("C",200),rep("D",100),rep("E",800),rep("F",700),rep("G",600),rep("H",500)))

u=sample(c(rep("A",300),rep("B",200),rep("C",100),rep("D",800),rep("E",700),rep("F",600),rep("G",500),rep("H",400)))

m=sample(c(rep("A",200),rep("B",100),rep("C",800),rep("D",700),rep("E",600),rep("F",500),rep("G",400),rep("H",300)))

w=sample(c(rep("A",100),rep("B",800),rep("C",700),rep("D",600),rep("E",500),rep("F",400),rep("G",300),rep("H",200)))

observation=c(x,y,z,v,t,u,m,w)

bw=baumWelch(hmm, observation,10)

print(bw$hmm)
```

Algoritmanın çalıştırılması sonucunda model parametreleri tahmin edilmiştir. Baum-Welch algoritması ile tahmin edilen başlangıç olasılıkları vektörü Tablo 4.9.'da, durumlar arası geçiş olasılıkları matrisi Tablo 4.10.'da ve emisyon olasılıkları matrisi Tablo 4.11.'de belirtilmiştir.

Tablo 4.9. Baum-Welch algoritması ile elde edilen başlangıç olasılıkları vektörü

	S_1	S_2
p	0.85	0.15

Tablo 4.10. Baum-Welch algoritması ile elde edilen durumlar arası geçiş olasılıkları matrisi

	S_1	S_2
S_1	0.698	0.302
S_2	0.627	0.373

Tablo 4.11. Baum-Welch algoritması ile elde edilen emisyon olasılıkları matrisi

	O ₁	O ₂	O ₃	O ₄	O ₅	O ₆	O ₇	O ₈
S ₁	0.184	0.178	0.145	0.145	0.123	0.155	0.010	0.060
S ₂	0.002	0.015	0.084	0.082	0.130	0.063	0.365	0.259

Baum-Welch algoritması ile model parametreleri tahmin edilmesinin ardından her kelimenin sahip olduğu gözlem ve bu gözlemlere ilişkin isim grubu ve fiil grubu emisyon olasılıkları belirlenmiştir. Herhangi bir kelime metin içerisinde birden fazla yerde kullanılmış olabilmektedir. Bu nedenle kullanıldığı yerlerde farklı sayıda ek harf almış olabilmektedir. Tüm kelimeleri emisyon olasılıklarına göre sınıflandırma yerine emisyon olasılıkları kullanarak her kelimenin sahip olduğu bir veya birden fazla gözlem vasıtası ile kelime türlerine göre Bayes olasılıkları belirlenmiş ve Bayes olasılığı yüksek kelime türü ilgili kelimenin kelime türü olarak atanmıştır. Yapılan işlemler, bir kelime ele alınarak ifade edilmek istenirse, Bayes teoremi,

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{\sum_j P(B|A_j)P(A_j)} \quad (4.1)$$

eşitliği olmak üzere Metinde kullanılan “Elazığ” kelimesinin isim türü Bayes olasılıkları ve fiil türü Bayes olasılıkları elde edilmek istensin. “Elazığ” kelimesi metin içerisinde 15 defa kullanılmıştır. “Elazığ” kelimesinin gözlem frekansları, 9 “ek harf yok”, 2 “ek 1 harf” ve 4 “ek 2 harf” olmak üzere toplam 15 defa metinde geçmektedir. “Elazığ” kelimesinin metin içerisindeki frekansı, ve emisyon olasılıkları Tablo 4.12.’de belirtilmiştir. Her gözlem sembolü için Bayes olasılıkları elde edilmiştir.

Tablo 4.12. “Elazığ” kelimesine ilişkin gözlem frekansları ve emisyon olasılıkları

	O ₁	O ₂	O ₃	O ₄	O ₅	O ₆	O ₇	O ₈
Kelime Frekansı	9	2	4	0	0	0	0	0
S ₁	0.184	0.178	0.145	0.145	0.123	0.155	0.010	0.060
S ₂	0.002	0.015	0.084	0.082	0.130	0.063	0.365	0.259

$$P(S_1|O_i) = \frac{(0,184 * \frac{9}{15}) + (0,178 * \frac{2}{15}) * (0,145 * \frac{4}{15})}{\left((0,184 * \frac{9}{15}) + (0,178 * \frac{2}{15}) * (0,145 * \frac{4}{15}) \right) + \left((0,002 * \frac{9}{15}) + (0,015 * \frac{2}{15}) + (0,084 * \frac{4}{15}) \right)} = 0.87$$

$$P(S_2|O_i) = \frac{(0,002 * \frac{9}{15}) + (0,015 * \frac{2}{15}) + (0,084 * \frac{4}{15})}{\left((0,184 * \frac{9}{15}) + (0,178 * \frac{2}{15}) * (0,145 * \frac{4}{15}) \right) + \left((0,002 * \frac{9}{15}) + (0,015 * \frac{2}{15}) + (0,084 * \frac{4}{15}) \right)} = 0.13$$

Bu sonuçlar eşliğinde “Elazığ” kelimesi için Bayes yöntemi ile kelime türü olasılıkları elde edilmiştir. Bu olasılıklara göre Elazığ kelimesinin S_1 durumunda yani kelime türü isim grubu olma olasılığı 0.87, S_2 durumunda yani fiil grubu olma olasılığı 0.13 olarak tahmin edilmiştir. Modelde ilgili kelime, Bayescil olasılık değeri yüksek olan kelime grubuna atanacağından “Elazığ” kelimesinin kelime türü grubu, olasılık değerlerine bakıldığında S_1 (isim grubu) durumu olma olasılığı $0.87 > 0.13$ olduğundan isim grubu olarak sınıflandırılmıştır. Bu işlem tüm kelimeler için uygulanmış, kelime türleri ve Bayes olasılık değerleri elde edilmiştir. Bu bölümdeki işlemlerin sonunda Tablo 4.13.’deki veriler elde edilmiştir.

Tablo 4.13. İkinci bölümdeki işlemlerin metin belgesine uygulanması ile elde edilen veriler

Kelime Sıra No	Kelimeler	Kelime No	İsim grubu Bayes olasılıkları	Fiil grubu Bayes olasılıkları
1	Elazığ	1	0.87	0.13
2	ili	2	0.77	0.23
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
292	yaşanan	180	0.77	0.23
293	iklimin	142	0.81	0.19
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
583	sanayi	195	0.99	0.01
584	kuruluşlardır	344	0.21	0.79

4.3. Varlık İlişki Modeli ile Enformasyon Çıkarım Alanlarını Düzenleme İşlemlerinin Metin Belgesine Uygulanması

Metindeki kelimelere ilişkin kelime türü Bayes olasılıklarının belirlenmesi ve kelime türü sınıflandırmasının tamamlanmasının ardından, bu bilgiler eşliğinde metin belgesinden enformasyon çıkarımı sağlayacak işlemler gerçekleştirilmiştir. İlk olarak metin belgesindeki kelimelerden enformasyon alanları belirlenmiştir. Enformasyon çıkarımı sağlayabilmek için metni en iyi şekilde tanıtan ve gerçekleştirebilecek minimum boyutta enformasyon çıkarılması için metindeki kelime sayısına orantılı olarak enformasyon çıkarım alan sayısı tespit edilmiştir. Modelin üçüncü bölümünde yapılan simülasyon çalışmaları sonucunda eşitlik 3.1 elde edilmiş, metindeki her 125 kelime için bir enformasyon çıkarım alanı belirlenmiştir. Enformasyon çıkarımı için alan sayısının belirlenmesinde,

$$\text{Alan sayısı} = \frac{\text{Metindeki kelime sayısı}}{125}$$

eşitliği kullanılmıştır. Bu eşitliğe göre metin belgesindeki enformasyon çıkarım alan sayısı,

$$\text{Alan Sayısı} = \frac{584}{125} = 4.67 \quad (4.3)$$

olarak elde edilmiştir. Alan sayısının belirlenmesinde ondalıklı değer elde edildiği için ondalıklı kısım bir üst tamsayıya tamamlanacaktır. Bu nedenle metin belgesinden 5 enformasyon çıkarım alanı elde edilmiştir. Metin belgelerinde başarılı bir enformasyon çıkarımı sağlanması için belirlenecek enformasyon alanlarının metni temsil yeteneği oldukça önem taşımaktadır. Bu nedenle enformasyon alanı olarak belirlenecek kelimeler için koşullar oluşturulmuştur. Bu koşullardan biri enformasyonun kendi başına anlam taşıyan bir kelime olmasıdır. Yani enformasyon alanı olarak belirlenecek kelime, kavram, varlık veya nesne olmalıdır. Bu nedenle enformasyon kelimesi edat, bağlaç, sıfat, zamir, zarf, ünlem ve fiil olmamalıdır. Diğer bir koşul ise enformasyon çıkarım alanlarının belirlenmesinde metin içerisinde en çok kullanılan kelimelerin enformasyon çıkarım alanı olarak belirlenmesidir. Fakat bu aşamada ilk koşul dikkate alınarak işlem yapılmaktadır. İlk koşula uygun olmayan kelimeler metin içerisinde en çok kullanılan kelimelerden biri

olsa dahi enformasyon çıkarım alanı olarak kabul edilemez. Metin içerisinde en çok tekrar eden kelimelerin frekansları belirlenmiştir. Bu verilere göre metin içerisinde en fazla kullanılan kelimelerden enformasyon alan sayısı kadar kelime, enformasyon alanları olarak tayin edilmiştir. Metin içerisinde en fazla frekansa sahip kelimeler Tablo 4.14.'te belirtilmiştir.

Tablo 4.14. Enformasyon alan kelimelerinin belirlenmesi

Kelime No	Kelime	Frekans	Enformasyon çıkarım alanı uygunluğu	Açıklama
35	Ve	17	Uygun değil	“ve” bağlaç olduğundan seçilemez.
1	Elazığ	15	Uygun	enformasyon alanı seçilebilir.
48	Bir	12	Uygun değil	“bir” sıfat olduğundan seçilemez.
195	Sanayi	9	Uygun	enformasyon alanı seçilebilir.
7	Harput	8	Uygun	enformasyon alanı seçilebilir.
29	Ol	8	Uygun değil	“ol” fiil olduğundan seçilemez.
142	İklim	8	Uygun	enformasyon alanı seçilebilir.
196	Tarım	7	Uygun	enformasyon alanı seçilebilir.
123	Yap	7	Uygun değil	“yap” fiil olduğundan seçilemez.
4	Yıl	7	Uygun	Enf. alanı için 5 kelimelik kota dolduğundan seçilemez.
.	.	.	.	.
.	.	.	.	.

“Elazığ”, “Sanayi”, “Harput”, “İklim,” “Tarım” kelimeleri metin belgesinin enformasyon alan kelimeleri olarak belirlenmiştir. Metinde enformasyon çıkarım alanlarının geçtiği cümlelerin belirlenmesi ve cümle içerisindeki diğer kelimeler ile ilişkisinin ortaya konulması enformasyon çıkarımı açısından önem taşımaktadır. Bu amaçla, enformasyon çıkarım kelimesinin kullanıldığı tüm cümleler belirlenmiştir. Cümlelerin belirlenmesinin ardından enformasyon çıkarım alanlarının içerisinde kullanıldığı cümleler ve cümlenin içerisindeki diğer kelimeler ilgili enformasyon çıkarım alanları altında listelenmiştir. Örnek ile belirtmek gerekirse “Elazığ” enformasyon alanının içinde geçtiği tüm cümleler ve bu cümlelerin içerisindeki diğer kelimeler “Elazığ” enformasyon alanının altında listelenmiştir. Öte yandan metin içerisindeki her cümleye, cümle numarası ve cümlelerin içerisindeki kelimelere cümledeki diziliş sırasına göre cümle içi sıra numarası atanmıştır. Varlık ilişki modelinin oluşturulabilmesi için birincil anahtar atanarak metin içerisinde kullanılan her kelimeye birbirinden farklı bir tanımlayıcı numara verilmesi gerekmektedir. Modelin geliştirilme aşamasında her kelime için belirlenen kelime sıra numarası alanı birincil anahtar olarak tayin edilmiştir. Bu bölümün sonunda yapılan işlemler birkaç kelime örnek ile Tablo 4.15. ile belirtilmiştir.

Tablo 4.15. Üçüncü bölüm işlemlerinin metin belgesine uygulanması ile elde edilen veriler

Kelime Sıra No	Kelime No	Kelime	Cümle No	Cümle içi Sıra No	Elazığ	Sanayi	Harput	İklim	Tarım
1	1	Elazığ	1	1	Elazığ	-	Elazığ	-	-
2	2	ili	1	2	ili	-	ili	-	-
.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.
292	180	yaşanan	20	8	yaşanan	-	-	yaşanan	-
293	142	iklimin	20	9	iklimin	-	-	iklimin	-
.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.
583	195	Sanayi	43	37	-	Sanayi	-	-	-
584	344	Kuruluş lardır	43	38	-	Kuruluş lardır	-	-	-

4.4. Verilerin Veritabanına Aktarılarak Enformasyon Çıkarımı Gerçekleştirme İşlemlerinin Metin Belgesine Uygulanması

Metin belgesindeki kelimelere ilişkin tüm verilerin ve alanların elde edilmesinin ardından bu veri ve alanlar Access veritabanına taşınmıştır. Veritabanına aktarılan alanlar Tablo 4.16.'da belirtilmiştir. *Kelime_sıra_no* alanı, birincil alan olarak tanımlanmış ve her kelime tanımlayıcı niteliği taşıyan birbirinden farklı bir numaraya sahip olmuştur. *Kelime_no* ile metin belgesindeki birbirinden farklı kelimelere numaralar atanmıştır. *Kelime_türü_grubu*, kelimelerin sınıflandırma sonucu belirlenen kelime türünü belirtmektedir. *Cümle_no*, metindeki her cümleye numara atanması ile elde edilmiştir. *Cümle_içi_sıra_no*, cümle içerisindeki kelimelerin sıra numarasını yani cümle içi diziliş numarasını belirtmektedir. Elazığ, Harput, iklim, sanayi ve tarım enformasyon alanlarıdır. Kelime ile belirtilen alan, metindeki kelimelerin anlam değiştirmeyen harflerden arındırılmış halini listelemektedir. *Bayes_isim_pr* alanı, ilgili kelimenin Bayes sınıflandırması ile isim grubu olasılığını belirtmektedir. *Bayes_fiil_pr* alanı ise aynı şekilde ilgili kelimenin Bayes sınıflandırması ile elde edilen fiil grubu olasılığını ifade etmektedir.

Tablo 4.16. Veritabanında alanların tanımlanması

BİLGİ	Alan Adı	Veri Türü
	Kelime_sıra_no	Sayı
	Kelime_no	Sayı
	Kelime_türü_grubu	Metin
	Cümle_no	Sayı
	Cümle_içi_sıra_no	Sayı
	Elazığ	Evet/Hayır
	Harput	Evet/Hayır
	İklim	Evet/Hayır
	Sanayi	Evet/Hayır
	Tarım	Evet/Hayır
	Kelime	Metin
	bayes_isim_pr	Sayı
	bayes_fiil_pr	Sayı

Veritabanına aktarılan veriler ise Tablo 4.17.'de belirtilmiştir. Tabloda alanlara ilişkin veriler listelenmiş veri boyutunun yüksek olması nedeniyle tabloda sadece belirli bir kısmı gösterilebilmiştir.

Tablo 4.17. Veritabanına aktarılan verilere ilişkin bir kesit

BİLGİ	Kelime_sıra_no	Kelime_no	Kelime_türü	Cümle_no	Cümle_içi_sıra_no	Elazığ	Harput	İklim	Sanayi	Tarım	Kelime	bayes_isim	bayes_fiil
	560	325	isim grubu	43	14	☐	☐	☐	☒	☐	fosfat	0.99	0.01
	561	326	isim grubu	43	15	☐	☐	☐	☒	☐	kireç	0.99	0.01
	562	327	isim grubu	43	16	☐	☐	☐	☒	☐	plastik	0.99	0.01
	563	328	isim grubu	43	17	☐	☐	☐	☒	☐	boru	0.99	0.01
	564	329	isim grubu	43	18	☐	☐	☐	☒	☐	tüpgaz	0.99	0.01
	565	330	isim grubu	43	19	☐	☐	☐	☒	☐	imalat	0.92	0.08
	566	35	isim grubu	43	20	☐	☐	☐	☒	☐	ve	0.99	0.01
	567	331	isim grubu	43	21	☐	☐	☐	☒	☐	dolum	0.99	0.01
	568	332	isim grubu	43	22	☐	☐	☐	☒	☐	kâğıt	0.99	0.01
	569	333	isim grubu	43	23	☐	☐	☐	☒	☐	tekstil	0.99	0.01
	570	334	isim grubu	43	24	☐	☐	☐	☒	☐	meşrubat	0.99	0.01
	571	335	isim grubu	43	25	☐	☐	☐	☒	☐	matbaacılık	0.99	0.01
	572	279	isim grubu	43	26	☐	☐	☐	☒	☐	mermer	0.89	0.11
	573	336	isim grubu	43	27	☐	☐	☐	☒	☐	ayçiçek	0.99	0.01
	574	337	isim grubu	43	28	☐	☐	☐	☒	☐	yağ	0.99	0.01
	575	338	isim grubu	43	29	☐	☐	☐	☒	☐	ayakkabı	0.99	0.01
	576	339	isim grubu	43	30	☐	☐	☐	☒	☐	mobilya	0.99	0.01
	577	340	isim grubu	43	31	☐	☐	☐	☒	☐	sabun	0.99	0.01
	578	341	isim grubu	43	32	☐	☐	☐	☒	☐	tıbbi	0.99	0.01
	579	342	isim grubu	43	33	☐	☐	☐	☒	☐	malzeme	0.99	0.01
	580	343	isim grubu	43	34	☐	☐	☐	☒	☐	fabrika	0.92	0.08
	581	260	fiil grubu	43	35	☐	☐	☐	☒	☐	başlıca	0.48	0.52
	582	14	isim grubu	43	36	☐	☐	☐	☒	☐	büyük	0.99	0.01
	583	195	isim grubu	43	37	☐	☐	☐	☒	☐	sanayi	0.99	0.01
	584	344	fiil grubu	43	38	☐	☐	☐	☒	☐	kuruluş	0.20	0.80

Verilerin veritabanına aktarılmasının ardından yapısal sorgulama dili (SQL) ile ilişkiler ve sorgular oluşturulmuştur. Anlamlı enformasyon sağlama amacı ile cümle içerisinde bulunan iki enformasyon alanı olan cümleler sorgulanmıştır. Bu aşamada yapılmak istenen cümlelerin incelenerek içerisinde iki enformasyon alanı olan cümleleri belirlemektir. Örnek olarak bir cümle içerisinde “Elazığ” ve “Sanayi” enformasyon alan kelimelerinin ikisi de kullanılmış ise bu cümle seçilen cümlelerden biri olacaktır. Bu cümlelerin seçimi için yapısal sorgulama dili (SQL) komutları kullanılmıştır.

İlk olarak “Elazığ” ve “Harput” enformasyon alanlarının birlikte kullanıldığı cümleler sorgulanmıştır. Sorgu, Tablo 4.18.’de belirtilmiştir. Sorgu sonucunda elde edilen bulgular Tablo 4.19.’da verilmiştir.

Tablo 4.18. Elazığ ve Harput enformasyon alan kelimelerinin sorgulanması için düzenlenen SQL kodları

```
SELECT
BİLGİ.Kelime_sıra_no,BİLGİ.Kelime_türü_grubu,
BİLGİ.Cümle_no, BİLGİ.Cümle_içi_sıra_no, BİLGİ.Elazığ,
BİLGİ.Harput, BİLGİ.İklim, BİLGİ.Sanayi, BİLGİ.Tarım,
BİLGİ.Kelime, BİLGİ.bayes_isim_pr, BİLGİ.bayes_fiil_pr

FROM BİLGİ

WHERE (((BİLGİ.Kelime_türü_grubu)="Fiil grubu") AND
((BİLGİ.Elazığ)=Yes) AND ((BİLGİ.Harput)=Yes));
```

Tablo 4.19. Elazığ ve Harput kelimelerinin birlikte kullanıldığı cümleler ve fiiller

Kelime_sıra_no	Kelime_türü	Cümle_no	Cümle_içi_sıra_no	Elazığ	Harput	İklim	Sanayi	Tarım	Kelime	bayes_isim	bayes_fiil
5	fiil grubu	1	5	☒	☒	☐	☐	☐	kur	0.20	0.80
6	fiil grubu	1	6	☒	☒	☐	☐	☐	san	0.49	0.51
8	fiil grubu	1	8	☒	☒	☐	☐	☐	kent	0.49	0.51

Bulgulara göre, “Elazığ” ve “Harput” enformasyon alan kelimelerin birlikte kullanıldığı bir cümle bulunmuştur. Bu cümlede ilk kullanılan enformasyon alan kelimesi “Elazığ” ikincisi ise “Harput”tur. Modele göre oluşturulacak yeni cümlede son olarak fiil belirlenmektedir. Sorguda üç fiil grubu kelime belirlenmiştir. Fiilin seçiminde en yüksek

Bayes olasılığına sahip fiil grubu kelimesi tayin edilmektedir. Buna göre sorguda 0.80 olasılık ile “kur” fiil grubu kelimesi yeni cümlelerin fiili olarak belirlenmiştir. İlgili kelimeler Tablo 4.20.’de belirtilmiştir.

Tablo 4.20. Elazığ ve Harput enformasyon alanlarından elde edilen yeni cümlelerin kelimeleri

1.Enformasyon Alan Kelimesi	2.Enformasyon Alan Kelimesi	Fiil Grubu
Elazığ	Harput	kur

Bir sonraki sorgulamada “Elazığ” ve “İklim” enformasyon alan kelimeleri sorgulanmıştır. Bu sorgulamada kullanılan kodlar Tablo 4.21.’de ve sorgulama sonucunda elde edilen bulgular Tablo 4.22.’de belirtilmiştir.

Tablo 4.21. Elazığ ve iklim enformasyon alan kelimelerinin sorgulanması için düzenlenen SQL kodları

```
SELECT
BİLGİ.Kelime_sıra_no,BİLGİ.Kelime_türü_grubu,
BİLGİ.Cümle_no,BİLGİ.Cümle_içi_sıra_no,BİLGİ.Elazığ,
BİLGİ.Harput,BİLGİ.İklim,BİLGİ.Sanayi,BİLGİ.Tarım,
BİLGİ.Kelime,BİLGİ.bayes_isim_pr,BİLGİ.bayes_fiil_pr
FROM BİLGİ
WHERE (((BİLGİ.Kelime_türü_grubu)="Fiilgrubu") AND
((BİLGİ.Elazığ)=Yes) AND ((BİLGİ.İklim)=Yes));
```

Tablo 4.22. Elazığ ve iklim kelimelerinin birlikte kullanıldığı cümleler ve fiiller

Kelime_sıra_no	Kelime_türü	Cümle_no	Cümle_içi_sıra_no	Elazığ	Harput	İklim	Sanayi	Tarım	Kelime	bayes_isim	bayes_fiil_p
210	fiil grubu	13	5	☒	☐	☒	☐	☐	sür	0.49	0.51
212	fiil grubu	13	7	☒	☐	☒	☐	☐	yap	0.34	0.66
214	fiil grubu	13	9	☒	☐	☒	☐	☐	yap	0.34	0.66
223	fiil grubu	13	18	☒	☐	☒	☐	☐	geçiş yap	0.20	0.80
260	fiil grubu	17	7	☒	☐	☒	☐	☐	yüz	0.20	0.80
262	fiil grubu	18	2	☒	☐	☒	☐	☐	Yap	0.34	0.66
266	fiil grubu	18	6	☒	☐	☒	☐	☐	belir	0.20	0.80
272	fiil grubu	18	12	☒	☐	☒	☐	☐	geçiş yap	0.20	0.80
286	fiil grubu	20	2	☒	☐	☒	☐	☐	yap	0.34	0.66

Sorgu sonucunda “Elazığ” ve “İklim” enformasyon alan kelimelerinin birlikte kullanıldığı dört cümle belirlenmiştir. Bu cümlelerde var olan dokuz adet fiil grubu kelime bulunmuştur. Yeni cümlenin oluşumunda kullanılacak fiilin belirlenmesi için en yüksek Bayes olasılıklı fiil grubu kelimesi belirlenmektedir. En yüksek Bayes olasılığı 0.80’dir. 0.80 olasılığına sahip üç fiil grubu kelime bulunmaktadır. Bu durumda cümle sıra ve cümle içi sıra numarası en son olan fiil grubu belirlenir. Bu işlem sonucunda “geçiş yap” fiil grubu kelimesi yeni cümlenin fiili olarak atanmıştır. “Geçiş yap” birleşik bir fiil olduğundan tek bir fiil olarak dikkate alınmaktadır. İlgili kelimeler Tablo 4.23.’de belirtilmiştir.

Tablo 4.23. Elazığ ve iklim enformasyon alanlarından elde edilen yeni cümlenin kelimeleri

1.Enformasyon Alan Kelimesi	2.Enformasyon Alan Kelimesi	Fiil Grubu
Elazığ	iklim	geçiş yap

“Elazığ” ve “Tarım” enformasyon alanlarının birlikte kullanıldığı cümlelerin belirlenmesi için kullanılan sorgu Tablo 4.24.’de ve sorgulamanın sonucunda elde edilen bulgular Tablo 4.25.’de belirtilmiştir.

Tablo 4.24. Elazığ ve tarım enformasyon alan kelimelerinin sorgulanması için düzenlenen SQL kodları

```
SELECT
BİLGİ.Kelime_sıra_no,BİLGİ.Kelime_türü_grubu,
BİLGİ.Cümle_no, BİLGİ.Cümle_içi_sıra_no,
BİLGİ.Elazığ, BİLGİ.Harput, BİLGİ.İklim, BİLGİ.Sanayi,
BİLGİ.Tarım, BİLGİ.Kelime, BİLGİ.bayes_isim_pr,
BİLGİ.bayes_fiil_pr
FROM BİLGİ
WHERE (((BİLGİ.Kelime_türü_grubu)="Fiilgrubu")
AND ((BİLGİ.Elazığ)=Yes) AND ((BİLGİ.Tarım)=Yes));
```

Tablo 4.25. Elazığ ve tarım kelimelerinin birlikte kullanıldığı cümleler ve fiiller

Kelime_sıra_no	Kelime_türü	Cümle_no	Cümle_içi_sıra_no	Elazığ	Harput	İklim	Sanayi	Tarım	Kelime	bayes_isim	bayes_fiil
422	fiil grubu	30	5	☒	☐	☐	☐	☒	yanış	0.49	0.51
426	fiil grubu	30	9	☒	☐	☐	☐	☒	geç	0.45	0.55

“Elazığ” ve “Tarım” enformasyon alan kelimelerinin birlikte kullanıldığı bir cümle tespit edilmiştir. Bu cümle içerisinde iki fiil grubu kelime mevcuttur. Bu kelimelerden 0.55 Bayes olasılık değeri ile diğer fiil grubu kelimesinden daha yüksek olan “geç” fiil grubu kelimesi yeni cümlelerin fiili olarak atanmıştır. İlgili kelimeler Tablo 4.26.’da belirtilmiştir.

Tablo 4.26. Elazığ ve tarım enformasyon alanlarından elde edilen yeni cümlelerin kelimeleri

1.Enformasyon Alan Kelimesi	2.Enformasyon Alan Kelimesi	Fiil Grubu
Elazığ	tarım	geç

“Sanayi” ve “Tarım” enformasyon alanlarının birlikte kullanıldığı cümlelerin belirlenmesi için kullanılan sorgu Tablo 4.27.’de ve sorgulamanın sonucunda elde edilen bulgular Tablo 4.28.’de belirtilmiştir.

Tablo 4.27. Sanayi ve tarım enformasyon alan kelimelerinin sorgulanması için düzenlenen SQL kodları

```
SELECT
BİLGİ.Kelime_sıra_no,BİLGİ.Kelime_türü_grubu,
BİLGİ.Cümle_no,BİLGİ.Cümle_içi_sıra_no,
BİLGİ.Elazığ,BİLGİ.Harput,BİLGİ.İklim,BİLGİ.Sanayi,
BİLGİ.Tarım,BİLGİ.Kelime,BİLGİ.bayes_isim_pr,
BİLGİ.bayes_fiil_pr
FROM BİLGİ
WHERE (((BİLGİ.Kelime_türü_grubu)="Fiilgrubu")
AND ((BİLGİ.Sanayi)=Yes) AND ((BİLGİ.Tarım)=Yes));
```

Tablo 4.28. Sanayi ve tarım kelimelerinin birlikte kullanıldığı cümleler ve fiiller

Kelime_sıra_no	Kelime_türü	Cümle_no	Cümle_içi_sıra_no	Elazığ	Harput	İklim	Sanayi	Tarım	Kelime	bayes_isim	bayes_fiil
335	fiil grubu	24	3	☐	☐	☐	☒	☒	yapılması	0.49	0.51
344	fiil grubu	24	12	☐	☐	☐	☒	☒	kal	0.20	0.80
346	fiil grubu	24	14	☐	☐	☐	☒	☒	alan	0.33	0.67
347	fiil grubu	24	15	☐	☐	☐	☒	☒	azal	0.49	0.51
348	fiil grubu	24	16	☐	☐	☐	☒	☒	paralel	0.49	0.51
350	fiil grubu	24	18	☐	☐	☐	☒	☒	canlan	0.20	0.80
353	fiil grubu	25	3	☐	☐	☐	☒	☒	gelir	0.49	0.51
363	fiil grubu	25	13	☐	☐	☐	☒	☒	ed	0.49	0.51

Sorgulama sonucunda “Sanayi” ve “Tarım” enformasyon alanlarının birlikte kullanıldığı iki cümle tespit edilmiştir. Bu iki cümlede sekiz adet fiil grubu kelimesi tespit edilmiş ve en yüksek olasılığa sahip olan iki fiil grubu kelimesinden cümle numarası ve cümle içi sıra numarasın en son olan “canlan” fiil grubu kelimesi yeni cümlelerin fiili olarak belirlenmiştir. İlgili kelimeler Tablo 4.29.’da belirtilmiştir.

Tablo 4.29. Sanayi ve tarım enformasyon alanlarından elde edilen yeni cümlelerin kelimeleri

1.Enformasyon Alan Kelimesi	2.Enformasyon Alan Kelimesi	Fiil Grubu
Sanayi	tarım	canlan

Yapılan diğer sorgularda iki enformasyon alan kelimesinin, üç enformasyon alan kelimesinin, dört enformasyon alan kelimesinin ve beş enformasyon alan kelimesinin birlikte kullanıldığı cümle tespit edilememiştir. Böylece enformasyon çıkarımı için toplam dört cümle elde edilmiştir. Oluşturulan bu yeni cümlelerin fiillerine öğrenilen geçmiş zaman eki ve bildirme eki eklenmiştir. Ayrıca metin içeriğinde ilgili fiil, kelime sıra numarasından tespit edilmiş, model tarafından etken veya edilgen yapıda olup olmadığı dikkate alınarak metinde edilgen yapıda kullanılan fiillere bu aşamada da edilgenlik eki eklenmiştir. Ekleme işlemi Tablo 4.30.’da belirtilmiştir.

Tablo 4.30. Fiillerin ekler ile tamamlanması

Kelime _sıra no	Fiil	Metindeki orjinal kullanımı	Etken/ edilgen	Fiil	Edil gen eki	Öğrenilen geçmiş zaman eki	Bildir me eki
5	Kur	Kur du	Edilgen	Kur	ul	muş	tur
272	Geçiş yap	Geçiş yapmıştır	Etken	Geçiş yap	-	miş	tır
426	Geç	Geçtiği	Etken	Geç	-	miş	tır
350	Canlan	Canlanmıştır	Etken	Canlan	-	miş	tır

Son olarak enformasyon çıkarımının tamamlanması için elde edilen tüm cümleler sırasına göre birleştirilerek enformasyon çıkarımı tamamlanmıştır. Elde edilen enformasyonda kelimeler yalın olarak eksiz bir şekilde kullanılmıştır. Tablo 4.31.'de modelde belirtilen işlemler sonucunda metin belgesinden elde edilen enformasyon belirtilmiştir.

Tablo 4.31. Metin belgesinden model ile doğrudan elde edilen enformasyon

Elazığ Harput kurulmuştur. Elazığ iklim geçiş yapmıştır. Elazığ tarım geçmiştir. Sanayi tarım canlanmıştır.

Enformasyonda, fiiller haricindeki kelimelerin ek kullanılmadan belirtildiği için tam olarak bir dil bilgisi bütünlüğü sağlanması mümkün değildir. Fakat elde edilen enformasyon okuyucuya büyük oranda metin içeriği hakkında bilgi sağlamaktadır. Modelin sağladığı enformasyon, metin içeriğini homojen bir şekilde temsil yeteneğine sahiptir.

5.SONUÇLAR VE TARTIŞMA

Bu çalışmada metin belgelerinden veya metin içeriklerinden enformasyon sağlanması amacı ile Saklı Markov modeli ve Bayes tabanlı sınıflandırıcılar ile bilgi modellerinin geliştirilmesi üzerine çalışılmıştır.

Enformasyon sağlamak için geliştirilen model 4 bölümden oluşturulmuştur. Her bölüm kendi içerisinde çok sayıda işlem basamağından oluşmaktadır. Modelin ilk bölümünde metindeki kelimelerin anlam değiştirmeyen eklerinden arındırılmasına yönelik işlem basamakları yer almaktadır. Bu işlemler bir algoritma eşliğinde yürütülmüştür. Modelin ikinci bölümünde ilk bölümden elde edilen veriler Saklı Markov modeli ve Bayes tabanlı sınıflandırıcılar ile kelimelerin türlerini sınıflandıran işlem basamakları yer almaktadır. Modelin üçüncü bölümünde ikinci bölümden elde edilen veriler ile varlık ilişki modeli oluşturularak alanlar belirlenmesine dair işlemler mevcuttur. Son bölüm ise tüm alan ve verilerin veritabanına aktararak sorgular eşliğinde enformasyon çıkarımını gerçekleştiren işlem basamakları yer almaktadır.

Model, web ortamında bulunan metin içeriğinden elde edilen metin belgesi üzerine uygulanmıştır. Modele alınan metin içeriği, 584 kelimedenden oluşmak üzere, sözel ve sayısal karakterler içermektedir. Dört bölümden oluşan modelin ilk bölümü işlem aşamalarına göre metine uygulanarak kelimeler listelenmiş, kelimelerin almış olduğu anlam değiştirmeyen eklerden arındırılması ise algoritma ile sağlanmıştır. Algoritma işlemlerinin sonucunda metindeki her kelime, almış oldukları ek harf sayılarına göre kategorize edilmiştir.

Saklı Markov modelinin oluşturulması aşamasında metindeki kelimelerin almış oldukları anlam değiştirmeyen ek harf sayıları gözlemler dizisi ve kelimelerin türleri durumlar olmak üzere Saklı Markov modeli kurulmuştur. Saklı Markov modelinin kurulması ile başlangıç olasılıkları vektörü, durumlar arası geçiş olasılıkları matrisi ve emisyon olasılıkları matrisi elde edilmiştir. Elde edilen olasılıklar ile Saklı Markov modelinin parametrelerinin tahmini için Baum-Welch algoritması çalıştırılmıştır. Saklı Markov modelinin parametre tahminleri Bayes olasılıklarının elde edilmesinde kullanılmıştır. Böylelikle her kelimenin isim grubu Bayes olasılığı ve fiil grubu Bayes olasılığı elde edilmiş ve tüm kelimelerin kelime türleri sınıflandırılmıştır.

Metindeki kelimelere ilişkin birçok verinin elde edilmesinden sonra enformasyon çıkarımını sağlayacak varlık ilişki modeli ve alanlar oluşturulmuştur. Enformasyon çıkarım alanları belirli örnekleme yöntemleri ve en yüksek frekansa sahip kavram veya varlık ifade eden kelimelerden belirlenmiştir. Bu aşamada metin içerisinde en yüksek frekansa sahip ve enformasyon kelimesi olmaya uygun “Elazığ”, “Harput”, “İklim”, “Sanayi” ve “Tarım” kelimeleri enformasyon çıkarım alanı olarak belirlenmiştir. Bunun yanı sıra varlık ilişki modelinin kurulması için, birincil anahtar tayini yapılmış ve cümlelere ilişkin yeni alanlar listelenmiştir.

Modelin son bölümünün uygulanmasında tüm veriler veritabanına aktarılarak, yapısal sorgulama dili (SQL) ile veritabanında cümle içerisinde en az iki enformasyon alanı bulunan cümleler tespit edilmiştir. Son olarak enformasyon alanları ve belirlenen fiiller ile yeni cümleler oluşturularak enformasyon çıkarımı sağlanmıştır.

Uygulama sonucu olarak modelin her aşamasında elde edilen bulgular, metin belgelerinin sınıflandırılmasında ve bilgi modellerinin geliştirilmesinde Saklı Markov modeli ve Bayes sınıflandırıcılarının başarılı bir şekilde uygulandığını göstermektedir. Saklı Markov Modeli ve Bayes tabanlı sınıflandırıcıların kelime türlerini sınıflandırmada başarılı sonuçlar verdiği gözlemlenmiştir. Modelin sınıflandırmadaki başarısını karşılaştırmak amacı ile modelin doğru sınıflandırma oranı, Viterbi algoritmasının doğru sınıflandırma oranıyla karşılaştırılmıştır. Viterbi algoritması optimum durum dizisini tahmin etmektedir. Gerçek durum dizisi ile viterbi durum dizisi eşleştirilerek sınıflandırma sonuçları Tablo 5.1.’de belirtilmiştir. Modelin ve Viterbi algoritmasının durumları yani kelime türlerini sınıflaması ile kelimelerin gerçek kelime türleri ile oransal olarak kıyaslaması Tablo 5.1.’de belirtilmiştir.

Tablo 5.1. Viterbi algoritması ile geliştirilen modelin sınıflama başarısının karşılaştırılması

		Kelime Türleri			
		Doğru sınıflandırma		Hatalı sınıflandırma	
		n	%	n	%
Viterbi	İsim grubu	489	98.6	7	1.4
	Fiil grubu	31	35.2	57	64.8
	Toplam	520	89.0	64	11.0
Model	İsim grubu	468	94.4	28	5.6
	Fiil grubu	62	70.5	26	29.5
	Toplam	530	90.8	54	9.2

Sınıflandırma sonuçları incelendiğinde, Viterbi algoritmasının oluşturduğu optimum durum dizisindeki kelime türleri ve metindeki kelimelere ilişkin gerçek kelime türleri karşılaştırılmıştır. Tabloya göre Viterbi algoritması, isim grubu kelimelerin % 98.6'sını, fiil grubu kelimelerin ise % 35.2'sini doğru sınıflandırmıştır. Viterbi algoritması isim grubu kelimelerin sınıflandırılmasında başarılı bir sınıflandırma sunarken fiil grubu kelimelerin doğru sınıflandırılmasında düşük bir oran elde etmiştir.

Geliştirilen model ise, isim grubu kelimelerin % 94.4'ünü, fiil grubu kelimelerin ise % 70.5'ini doğru sınıflandırmıştır. Model, isim grubu kelimelerin sınıflandırılmasında yüksek oranda başarılı bir sınıflandırma gerçekleştirirken, fiil grubu kelimelerin sınıflandırılmasında başarılı düzeyde bir sınıflandırma ortaya koymuştur. Metindeki tüm kelimelerin sınıflandırma başarısı dikkate alındığında Viterbi algoritması % 89.0 oranında doğru sınıflandırma başarısına sahip iken, modelin doğru sınıflandırma başarısı % 90.8 olarak belirlenmiştir. Model sınıflandırmada başarılı bir sonuç ortaya koymaktadır.

Türkçe Dil yapısında kelimelerin sondan eklemeli olması, kelimelerin türlerine göre sınıflandırılmasında farklı bir avantaj sunmuştur. Kelimelerin almış olduğu eklerin harf sayıları kelimelerin türlerinin belirlenmesinde yardımcı olmaktadır. Çalışmada Saklı Markov modeli ile saklı durumdaki kelime türlerinin tahmin edilebilmesi için kelimelerin almış olduğu eklerin harf sayıları gözlem dizisi olarak belirlenmiş ve bu gözlem dizisine göre değerlendirmeler gerçekleştirilmiştir. Metin belgesindeki kelimelerin incelenmesi sonucu elde edilen emisyon olasılıklarına göre kelime türü isim grubu olan kelimeler 0.61 olasılık ile ek harf almamaktadır, 0.08 olasılık ile ek 1 harf, 0.10 olasılık ile ek 2 harf, 0.09 olasılık ile ek 3 harf, 0.07 olasılık ile ek 4 harf, 0.04 olasılık ile ek 5 harf, 0.002 olasılık ile ek 6 harf ve 0.008 olasılık ile ek 7 harf veya daha fazla sayıda harf aldığı gözlemlenmiştir. Kelime türü fiil grubu olan kelimelerin ise 0.02 olasılık ile ek harf almadığı, yine 0.02 olasılık ile ek 1 harf, 0.18 olasılık ile ek 2 harf, 0.14 olasılık ile ek 3 harf, 0.23 olasılık ile ek 4 harf, 0.05 olasılık ile ek 5 harf, 0.25 olasılık ile ek 6 harf, 0.11 olasılık ile ek 7 veya daha fazla harf aldığı gözlemlenmiştir.

Durumlar arası geçiş olasılıkları ile metin içerisindeki bir kelime türünden diğer kelime türlerine geçiş olasılıklarını belirtmektedir. Kelime türü isim grubu olan bir kelimeden sonra 0.84 olasılık ile yine kelime türü isim grubu olan bir kelime, 0.16 olasılık ile kelime türü fiil grubu olan bir kelime geleceği tahmin edilmiştir. Öte yandan kelime türü fiil grubu olan bir kelimeden sonra 0.89 olasılık ile kelime türü isim grubu olan bir kelime, 0.11 olasılık ile kelime türü fiil grubu olan bir kelime geleceği tahmin edilmiştir. Baum-Welch

algoritması ile Saklı Markov modelinin parametrelerini tahmin etmek için Baum-Welch algoritması ile parametre olasılık değerleri elde edilmiştir. Saklı Markov modelinin metinde bulunan kelimelere ilişkin kelime türlerinin olasılıklarının belirlenmesinde başarılı bir yöntem olduğu gözlemlenmiştir.

Model içerisinde kelimelerin sınıflandırılmasında Bayes tabanlı sınıflandırıcılar kullanılmıştır. Bu yöntemde metindeki her kelimeye ilişkin kelime türlerinin Bayes olasılıkları elde edilmiş ve nihai sınıflandırma Bayes tabanlı sınıflandırma yöntemleri ile sağlanmıştır. Bayes tabanlı sınıflandırma yöntemleri ve Bayes teorisi metin belgesindeki kelimeleri türlerine göre sınıflandırmada başarılı sonuçlar vermektedir.

Metin belgesinden enformasyon sağlaması için varlık ilişki modeli ile alanlar oluşturulmuştur. Enformasyon çıkarımında kullanılacak enformasyon alan kelimeleri belirlenmiştir. Tüm verilerin elde edilmesi sonucunda veriler ve alanlar veritabanına aktararak yapısal sorgulama dili (SQL) ile sorgular oluşturulmuştur. Sorgular ve modeldeki diğer işlem adımları ile enformasyon oluşumunda kullanılacak cümleler elde edilmiştir. İşlemler sonucunda 4 cümle elde edilmiştir. Bu cümleler “Elazığ Harput kurulmuştur”, “Elazığ iklim geçiş yapmıştır”, “Elazığ tarım geçmiştir”, “Sanayi tarım canlanmıştır” olarak elde edilmiştir. Toplam 13 kelimeden oluşan cümlelerin ilk iki kelimesi enformasyon alan kelimesi diğer kelime ise fiil gruplarından elde edilmiştir. Elde edilen cümlelerdeki kelimeler fiiller haricinde hiçbir ek kullanılmadan aktarılmıştır.

Model, başlangıcından son aşamasına kadar doğrudan insan müdahalesi olmadan bilgisayar ve yardımcı programlar aracılığı ile metin kitlelerinden enformasyon çıkarımı yapacak şekilde tasarlanmıştır. Uygulama dikkate alındığında 584 kelimelik metin belgesinden 13 kelimelik bir enformasyon elde edilmiştir. 13 kelimelik enformasyon 4 cümleden oluşmaktadır. Cümlelerin tamamı metin belgesindeki içeriği homojen bir şekilde yansıtmaktadır. Cümlelerde kullanılan kelimeler, doğrudan insan müdahalesine yer verilmemesi ve modeldeki eklerinden arındırılmış kelimelerden elde edilmesi nedeniyle cümlelerde yalın olarak belirtilmiştir.

Metinden elde edilen enformasyondaki cümleler anlamsal olarak incelendiğinde, cümlelerdeki kelimelerin hiçbir ek taşımasına rağmen anlamsal bütünlüğünü koruduğu gözlemlenmiştir.

Elde edilen ilk cümle olan “Elazığ Harput kurulmuştur”, Elazığ’ın kuruluşu ve kurulduğu bölge hakkında Harput ile ilgili olduğu hususunda bir enformasyon sağladığı görülmektedir.

“Elazığ iklim geiş yapmıřtır” cümlesinde Elazığ’ın ikliminin deėiřtiėi ve eski iklim düzeninden yeni bir iklim düzenine geiş yaptıėı hususunda enformasyon sağladıėı görölmektedir.

“Elazığ tarım gemiřtir” cümlesinde Elazığ tarım sektöründe kıyaslandıėı bařka bir sektör veya bařka konu olduėu ve bu konuda bir üstünlük sağlandıėı hususunda enformasyon sağladıėı gözlemlenmektedir.

“Sanayi tarım canlanmıřtır” cümlesinde sanayi ve tarım sektörlerinde canlanma ve hareketlilik olduėu hususunda enformasyon sağladıėı gözlemlenmektedir.

Cümlelerin sağladıėı anlamlar, metin belgesindeki içerik ile homojen bir uyum sağlamaktadır. Anlam kaybı minimum düzeyde olduėu ve dikkate deėer bir anlam sapması olmadıėı görölmektedir. Öte yandan elde edilen enformasyonun metin belgesini tamamını homojen bir řekilde temsil ettiėi gözlemlenmektedir. Uygulamada incelenen metin belgesi “Elazığ” hakkında etimolojik bilgi, iklim bilgisi ve ekonomi bilgisi ile ilgilidir. Enformasyondaki birinci cümle Elazığ’ın etimoloji ile ilgilidir. İkinci cümle Elazığ’ın iklimi ile ilgilidir. Üüncü ve dördüncü cümle Elazığ’ın ekonomisi ile ilgilidir.

Sonuç olarak alıřmada Saklı Markov modeli ve Bayes tabanlı sınıflandırıcılar kullanılarak bilgi modelleri geliştirilmiřtir. Geliřtirilen bilgi modeli metin kitlelerinden otomatik olarak enformasyon ıkarmıř ve metindeki kelimeleri bařarılı bir řekilde sınıflandırmıřtır. Geliřtirilen modelin Türke metinlerden otomatik olarak enformasyon elde edilmesinde bařarılı sonuçlar verdiėi gözlemlenmiřtir.

6.ÖNERİLER

Çalışmada Saklı Markov modeli ve Bayes tabanlı sınıflandırıcılar ile enformasyon elde etmeyi amaçlayan bilgi modeli geliştirilmiştir. Kompleks bir yapıya sahip bilgi modelinde Türkçe Dili'ndeki metin belgelerinden doğrudan enformasyon çıkarımı gerçekleştirilmiştir.

Türkçe Dili'nin yapısına bağlı olarak bir kelime birden fazla anlamda kullanılabilmektedir. Bu nedenle kullanılan kelimenin anlamını otomatik olarak tahmin edilebilmesi üzerine yapılacak çalışmalar, önemli katkılar sağlayabilir.

Türkçe Dili üzerine metin madenciliği yöntemleri ile çalışmalarda Türkçe tüm kelimeleri içeren ve alabileceği tüm ekleri ihtiva eden geniş bir Türkçe kelime veritabanı üzerine çalışmalar gerçekleştirilmesi, çalışmaların başarısına katkı sağlayabilir.

Kelime türlerinin sınıflandırılmasında bazı kelime türlerine ait kelimeler özelliklerine göre birbirlerine oldukça benzerlik gösterebilmektedir. Bu kelimeler birden fazla kelime türü içerisinde de kullanılabilmektedir. Bu kelimelerin tasnifinde farklı özellikler kullanılması çalışmaların hassasiyetini artırabilir.

Modelde kelimelerin anlam değiştirmeyen eklerinden ayrıştırılmış olması nedeni ile enformasyon cümlelerinin elde edilmesinde kelimeler yalın halde kullanılmıştır. uygulanmaktadır. Bu kelimelere uygun ekler getirilerek enformasyon, dilbilgisi kurallarına uygun hale getirilebilir. Fakat bu işlemin gerçekleştirilmesinde tamamen bilgisayar tarafından otomatik olarak gerçekleştirilmesinin sağlanmasına önem verilmelidir.

Geliştirilen model Türkçe metin belgelerinden enformasyon elde etme ve kelimelerin sınıflandırılması hususunda başarılı sonuçlar elde etmiştir. Bu model farklı disiplinlerde de kullanılabilir.

Modelin büyük boyutlardaki metin kitlelerinde uygulanması, kullanıcıları hem zaman hem de iş yükü bakımından önemli yararlar sağlayacaktır.

Modelin kullanılması, işletme dünyasında ve birçok sektörde büyük avantajlar kazandıracaktır.

Modelin arama motorları gibi yapılarla uyumlu hale getirilmesi ile arama motorları ile istenilen belgelerdeki metin kitlelerinden zaman ve enerji harcamaksızın arzu edilen enformasyonun sağlanmasında kullanılabilir.

KAYNAKLAR

- [1] **Yılmaz, M.**, 2009. Enformasyon ve bilgi kavramları bağlamında enformasyon yönetimi ve bilgi yönetimi. *Ankara Üniversitesi Dil Tarih-Coğrafya Fakültesi Dergisi* **49**(1),95-118.
- [2] **Jensen, P.E.**, 2005. A contextual theory of learning and the learning organization. *Knowledge and Process Management*. **12**(1),53-64.
- [3] **Misra, D.C. Hariharan, R. and Khaneja, M.**, 2003. E-knowledge management framework for government organizations. *Information System Management*. **20**(2),38-48.
- [4] **Durna, U. Ve Demirel, Y.**, 2008. Bilgi yönetiminde bilgiyi anlamak. *Erciyes Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi*, **30**,129-156.
- [5] **Barutçugil, İ.**, 2002. Bilgi Yönetimi. Kariyer Yayıncılık. İstanbul.
- [6] **He, Q. and Ling, T.W.**, 2006. An ontology based approach to the integration of entity-relationship schemas. *Data&Knowledge engineering*, **58**(2006),299-326.
- [7] **Alhajj, R.**, 2003. Extracting the extended entity-relationship model from a legacy relational database. *Information Systems*. **28**(2003),597-618.
- [8] **Arslan, A.A.**, 2011. Türkçe metinlerden anlamsal bilgi çıkarımı için bir veri madenciliği uygulaması. *Yüksek Lisans Tezi*, Başkent Üniversitesi Fen Bilimleri Enstitüsü. Ankara.
- [9] **Tsai, C.F. and Chang, C.W.**, 2013. SVOIS: Support vector oriented instance selection for text classification. *Information Systems*. **38**(2013),1070-1083.
- [10] **Ghiassi, M. Olschimke, M. Moon, B. and Arnaudo, M.P.**, 2012. Automated text classification using a dynamic artificial neural network model. *Expert Systems with Applications*. **39**(2012),10967-10976.
- [11] **Chen, Y.T. and Chen, M.C.**, 2011. Using chi-square statistics to measure similarities for text categorization. *Expert Systems with Applications*. **38**(4),3085-3090.
- [12] **Jiang, C. Coenen, F. Sanderson, R. and Zito, M.**, 2010. Text classification using graph mining-based feature extraction. *Knowledge-based systems*. **23**(4),302-308.
- [13] **Klose, A. Nürnberger, A. Kruse, R. Hartmann, G. and Richards, M.**, 2000. Interactive text retrieval based on document similarities. *Physics and Chemistry of the earth, Part A: Solid Earth and Geodesy*. **25**(8),649-654.

- [14] **Bayer, H.**, 2011. Veri madenciliğinde bir metin madenciliği uygulaması. *Yüksek lisans tezi*. Beykent Üniversitesi Fen bilimleri Enstitüsü. İstanbul.
- [15] **Soysal, E. Çiçekli, İ. ve Baykal, N.**, 2010. Design and evaluation of on ontology based information extraction system for radiological reports. *Computers in Biology and Medicine*. **40**(11-12),900-911.
- [16] **Deniz, O.**, 2012. Ontology based text mining in Turkish radiology reports. *Master Thesis*, Middle East Technical University. The Graduate School of Natural and applied sciences. Ankara.
- [17] **Yang, Y.**, 1999. An evaluation of statistical approaches to text categorization. *Information Retrieval*. **1**(1-2),69-90.
- [18] **Sebastiani, F.**, 2002. Machine learning in automated text categorization. *ACM computing Surveys*. **34**(1),1-47.
- [19] **Varol, M.**, 2011. Metin madenciliği yöntemlerini kullanarak Türkçe dökümanlarda tür ve yazar tanıma. *Yüksek Lisans Tezi*. Süleyman Demirel Üniversitesi Fen Bilimleri Enstitüsü. Isparta.
- [20] **Gieger, C. Deneke, H. and Fluck, J.**, 2003. The future of text mining in genome-based clinical research. *Biosilico*. **1**(3),97-102.
- [21] **Erhardt, R.A. Schneider, R. and Blaschke, C.**, 2006. Status of text mining techniques applied to biomedical text. *Drug Discovery Today*. **11**(7-8),315-325.
- [22] **Karaca, M.F.**, 2012. Metin madenciliği yöntemi ile haber sitelerindeki köşe yazılarının sınıflandırılması. *Yüksek Lisans Tezi*. Karabük Üniversitesi Fen Bilimleri Enstitüsü. Karabük.
- [23] **Taha, S.M.**, 2011. Metin madenciliği ile doküman demetleme. *Yüksek Lisans Tezi*. Gazi üniversitesi Bilişim enstitüsü. Ankara.
- [24] **Pilavcılar, İ.F.**, 2007. Metin madenciliği ile metin sınıflandırma. *Yüksek Lisans Tezi*. Yıldız Teknik Üniversitesi Fen bilimleri Enstitüsü. İstanbul
- [25] **Kocabaş, İ.**, 2011. Bilgi erişim sistemlerinde istatistiksel bağımsızlık esasında indeks terim ağırlıklandırma. *Doktora tezi*. Ege Üniversitesi Fen bilimleri Enstitüsü. İzmir.
- [26] **Salton, G. and Buckley, C.**, 1988. Term-weighting approaches in automatic text retrieval. *Information Processing and Retrieval*. **24**(5),513-523.
- [27] **Can, F. Kocberber, S. Balcik, E. Kaynak, C. Ocalan, H. C. ve Vursavas, O. M.**, 2008. Information Retrieval On Turkish Texts. *Journal of the American Society for Information Science and Technology*, **59** (3),407-421.

- [28] **Ceylan, N.M.**, 2010. Effects of diacritics in Turkish Information retrieval. *Yüksek Lisans Tezi*. Dokuz Eylül Üniversitesi. Fen Bilimleri Enstitüsü. İzmir.
- [29] **Sbattella, L. and Tedesco, R.**, 2013. A novel semantic information retrieval system based on a three-level domain model. *Journal of Systems and Software*. **86**(5),1426-1452.
- [30] **Yoo, D.**, 2012. Hybrid query processing for personalized information retrieval on the semantic web. *Knowledge-based Systems*. **27**(2012),211-218.
- [31] **Lu,W.H. Lin, R.S. Chan, Y.C. and Chen, K.H.**, 2008. Using web resources to construct multilingual medical thesaurus for cross-language medical information retrieval. *Decision Support Systems*. **45**(3),585-595.
- [32] **Li, Q. Chen, Y.P. Myaeng, S.H. Jin, Y. and Kang, B.Y.**, 2009. Concept unification of terms in different languages via web mining for Information Retrieval. *Information Processing and Management*. **45**(2),246-262.
- [33] **Liaw, S.S. and Huang, H.M.**, 2006. Information retrieval from the World Wide Web: a user focused approach based on individual experience with search engines. *Computer in Human Behavior*. **22**(3),501-517.
- [34] **Rieh, S.Y. and Xie, H.**, 2006. Analysis of multiple query reformulations on the web: The interactive information retrieval context. *Information Processing and Management*. **42**(2006),751-768.
- [35] **Ko, Y. and Seo, J.**, 2008. An effective sentence-extraction technique using contextual information and statistical approaches for text summarization. *Pattern Recognition Letters*. **29**(9),1366-1371.
- [36] **Hobbs, R.**, 2002. Information extraction from biomedical text. *Journal of Biomedical Informatics*. **35**(4),260-264.
- [37] **Downey, D. Etzioni, O. and Soderland, S.**, 2010. Analysis of a probabilistic model of redundancy in unsupervised information extraction. *Artificial Intelligence*. **174**(11),726-748.
- [38] **Mykowiecka, A. Marciniak, M. and Kupsc, A.**, 2009. Rule based information extraction from patients clinical data. *Journal of Biomedical Informatics*. **42**(5),923-936.
- [39] **Tur, G. Tur, D.H. and Oflazer, K.**, 2003. A statistical information extraction system for Turkish. *Natural Language Engineering*, **9**(2),181-210.
- [40] **Tatar, S. and Çiçekli, İ.**, 2011. Automatic rule learning exploiting morphological features for named entity recognition in Turkish. *Journal of Information Science*. **37**(2),137-151.

- [41] **Tatar, S.**, 2011. Automating information extraction task for Turkish texts. *Doktora tezi*. Bilkent Üniversitesi, Mühendislik ve Fen Bilimleri Enstitüsü. Ankara.
- [42] **Adalı, Ş.** 2009. An integrated architecture for information extraction from documents in Turkish. *Doktora Tezi*. İstanbul Teknik Üniversitesi. Fen ve Teknoloji Bilimleri Enstitüsü. İstanbul.
- [43] **Vlad, M.O. Moran, F. and Ross, J.**, 2003. Lifetimes and on-off distributions for single- molecule kinetics. Stochastic approach and extraction of information from experimental data. *Chemical Physics*. **287**(1-2),83-90.
- [44] **Hromic, H. Atkinson, J.**, 2012. Stochastic web-based natural language dialog planning for web information filtering. *Expert systems with Applications*. **39**(10),9420-9426.
- [45] **Bagui, S. and Earp, R.**, 2003. Database design using entity relationship diagrams. Auerbach Publications, Florida.
- [46] **Halpin, T. and Morgan, T.**, 2008. Information modeling and relational databases; second edition. Morgan Kaufmann Publisher, USA.
- [47] **Çölkesen, R.**, 2010. Veri Yapıları ve algoritmalar, Papatya Yayıncılık Eğitim, İstanbul.
- [48] **Burma, Z.A.**, 2009. Veritabanı Yönetim Sistemleri. Seçkin Yayıncılık, Ankara.
- [49] **Gökçen, H.**, 2007. Yönetim Bilgi Sistemleri, Palme Yayıncılık, Ankara.
- [50] **Baransel, C. ve Mumcuoğlu, A.**, 2003. Web tabanlı üç katmanlı bilişim mimarileri. Sas Bilişim Yayıncılık, Ankara.
- [51] **Güngören, B.**, 2005. UML ile nesne tabanlı çözümleme ve tasarım. Seçkin Yayıncılık, Ankara.
- [52] **Uymaz, S.A.**, 2007. UML ile sanal kampüs modellemesi, *Yüksek Lisans Tezi*, Selçuk Üniversitesi, Fen Bilimleri Enstitüsü, Konya.
- [53] **Stanezyk, S.**, 1993. Theory And Practice of relational databases. Ucl Press, London.
- [54] **Batini, C. Ceri, S, and Navathe, S. B.**, 1992. Conceptual database design, an entity relationship approach. Benjamin Cummings Publishing, California.
- [55] **Bhar, R. and Hamori, S.**, 2004. Hidden Markov models, application to financial economics, Kluwer Academic Publishers, Dordrecht.
- [56] **Ching, W.K. and Ng, M.K.**, 2006. Markov Chains: Model algorithms and applications. Springer Science+Business Media, USA.

- [57] **Han, J. and Kamber, M.,** 2006. Data Mining: Concepts and techniques, second edition. Morgan Kauffman Publishers, San Francisco.
- [58] **Mitcell, T.M.,** 1977. Machine Learning. McGraw-Hill Science/Engineering/Math, Portland.
- [59] **Feldman, R. and Sanger, J.,** 2006. The Text Mining handbook. Cambrigde University Press, New York.
- [60] **Weiss, S.M., Indurkha, N., and Zhang, T.,** 2010. Fundamentals of predictive text Mining. Springer-Verlag, London.
- [61] **Prado, H.A., and Ferneda, E.,** 2008. Emerging Technologies of text mining: techniques and applications. Information Science Reference, Hershey.
- [62] **Aggarwal, C. and Zhai, C.,** 2012. Mining text data. Springer Science+Business media, New York.
- [63] **Han, J. and Kamber, M.,** 2006. Data Mining; Concept and techniques. Second Edition. Morgan Kauffman Publishers, San Francisco.
- [64] **Güven, A.,** 2007. Türkçe Belgelerin anlam tabanlı yöntemler ile madenciliği. *Doktora tezi.* Yıldız Teknik Üniversitesi, Fen Bilimleri Enstitüsü, İstanbul.
- [65] **Date, C.,** 1995. An Introduction to Database Systems, 6th edition. Addison-Wesley, Reading.
- [66] **Maier, D.,** 1983. The Theory of Relational Databases. Computer Science Press, Rockville.
- [67] **Codd, E.,** 1970. A Relational Model for Large Shared Data Banks, *CACM*. **13(6)**,377-387.
- [68] **Simsion, G. and Witt, G.,** 2005. Data Modeling Essentials, 3 rd edition. Morgan Kaufmann, San Francisco.
- [69] **Muller, R.,** 1999. Database Design for Smarties, Morgan Kaufmann, San Francisco.
- [70] **Connolly, T., Begg, C. and Strachan, A.,** 1999. Database Systems: A practical approach to design implementation and management, 2 nd edition, Addison-Wesley.
- [71] **Halpin, T. and Bloesch, A.,** 1999. Data Modeling in UML and ORM: A Comparison, *Journal of Database Management*, **10(4)**, 4-13.
- [72] **Booch, G. Rumbaugh, J. and Jacobson, I.,** 1999. The Unified Modeling Language User Guide. Addison-Wesley, Reading.

- [73] **Kent, W.**, 2000. Data and Reality, 2 nd Edition, The 1st Books Library.
- [74] **Chen, P.**, 1976. The Entity-Relationship Model-Toward a Unified View of Data. *ACM Transactions on Database Systems*, 1(1), 9-36.
- [75] **Vetter, M. and Maddison, R.N.**, 1981. Data base Design Methodology. Prentice/Hall International.
- [76] **Tsichritzis, D.C. Lochovsky, F.H.**, 1982. Data Models. Prentice-Hall, Englewood Cliffs, New Jersey.
- [77] **Codd, E.F.**, 1982. The 1981 ACM Turing Award Lecture: Relational Database – A Practical Foundation for Productivity, *CACM*, 25(2),109-117.
- [78] **Smith, J.M. and Smith, D.C.P.**, 1977. Database Abstraction: Aggregation and Generalization. *ACM Transactions on Database Systems*. **2**(1977),105-33.
- [79] **Daume, III H. and Marcu, D.**, 2006. Bayesian query-focused summarization. *In Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, 305–312.
- [80] **Haghighi, A., and Vanderwende, L.**, 2009. Exploring content models for multi-document summarization. *In Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 362–370.
- [81] **Dunning, T.**, 1994. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, **19**(1),61–74.
- [82] **Gupta, S. Nenkova, A. and Jurafsky, D.**, 2007. Measuring importance and query relevance in topic-focused multi-document summarization. *In Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, Demo and Poster Sessions*, 193–196.
- [83] **Salton, G. Singhal, A. Mitra, M. and Buckley, C.**, 1997. Automatic text structuring and summarization. *Information Processing and Management*, **33**(2),193–207.
- [84] **Edmundson, H.P.**, 1969. New methods in automatic extracting. *Journal of the ACM*, **16**(2),264–285.
- [85] **Kupiec, J. Pedersen, J. and Chen, F.**, 1995. A trainable document summarizer. *In Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 68–73.
- [86] **Domingos, P. and Pazzani, M.**, 1997. On the Optimality of the Simple Bayesian Classifier under Zero-One Loss. *Machine Learning* **29**,103–130.

- [87] **Cohen, W.W.**, 1995a. Learning to Classify English Text with ILP Methods. (Raedt, L. D. ed.) In *Advances in Inductive Logic Programming*. IOS Press, Amsterdam.
- [88] **Cohen, W.W.**, 1995b. Text Categorization and Relational Learning. In *Proceedings of ICML-95, 12th International Conference on Machine Learning*. Lake Tahoe, NV, Morgan Kaufmann Publishers, San Francisco.
- [89] **Cohen, W. W. and Singer, Y.**, 1996. Context-Sensitive Learning Methods for Text Categorization. In *Proceedings of SIGIR-96, 19th ACM International Conference on Research and Development in Information Retrieval*. Zurich, ACM Press, New York.
- [90] **Yang, Y. and Chute, C. G.**, 1994. An Example-Based Mapping Method for Text Categorization and Retrieval. *ACM Transactions on Information Systems*, **12**(3), 252–277.
- [91] **Tapeiro, C. S.**, 1998, *Applied Stochastic Models and Control for Finance and Insurance*, Kluwer Academic Press, Dordrecht.
- [92] **Föllinger, O.**, 1992. *Regelungstechnik*, 7. Auflage, Hüthig Buch Verlag Heidelberg.
- [93] **Kalman, R.E.**, 1960. A New Approach to Linear Filtering and Prediction Problems. *Transactions of the ASME-Journal of Basic Engineering*. **82**(D),35-45.
- [94] **Koumeou, G.L.**, 2011. History and Theoretical Basics of Hidden Markov Models, (Ed. Dymarski, P.), *Hidden Markov Modes, Theory and applications*, Intech, Croatia.
- [95] **Baum, L. E., Petrie, T., Soules,, G. and Weiss, N.**, 1970. A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *The Annals of Mathematical Statistics*, **41**(1),164-171.
- [96] **Sahani, M.**, 1999. Latent variable models for neural data analysis, *PhD Thesis*, California Institute of Technology. California.
- [97] **Viterbi, A.J.**, 1967. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm, *IEEE Transactions on Information Theory*. **13**(2),260-269.
- [98] **Baum, L.E. and Petrie, T.**, 1966. Statistical inference for probabilistic functions of finite state Markov chains. *The Annals of Mathematical Statistics*, **37**(6). 1554-1563.
- [99] **Rabiner, L. R.**, 1989. A tutorial on Hidden Markov Models and selected applications in speech recognition, *Proceedings of the IEEE*, **77**(2), 257-286.

- [100] **Huang, X.D. and Jack, M.A.,** 1990. Semi-continuous hidden Markov models for speech recognition, *Ph.D. thesis*, Department of Electrical Engineering, University of Edinburgh.
- [101] **Huang, X. D. Ariki, Y. and Jack, M. A.,** 1990. Hidden Markov Models for Speech Recognition. Edinburgh University Press.
- [102] **Kouemou, G.,** 2010. Radar Technology, (Ed. Kouemou, G.), Intech, Croatia.
- [103] **Rabiner, L.R. Wilpon, J.G. and Juang, B.H.,** 1986. A segmental k-means training procedure for connected word recognition, *AT&T Technical Journal*, **65**(3),21-31.
- [104] **Warakagoda, N.,** 2009. Hidden Markov Models, <http://jedlik.phy.bme.hu/~gerjanos/HMM/node2.html> 14 Ağustos 2010.
- [105] **Markov, A. A.,** 1908. *Wahrscheinlichkeitsrechnung*, B. G. Teubner, Leipzig, Berlin.
- [106] **Gül, M.,** 2008. Türkiye Türkçesinde Sözcük Türlerinin Sınıflandırılması. *Yüksek Lisans Tezi* Mersin Üniversitesi Sosyal Bilimler Enstitüsü, Mersin.
- [107] **Ergün, K.,** 2012. Metin madenciliği yöntemleri ile ürün yorumlarının otomatik değerlendirilmesi. *Doktora Tezi*. Sakarya Üniversitesi Fen Bilimleri Enstitüsü, Sakarya.
- [108] <https://code.google.com/p/zemberek/> Zemberek. 15.07.2014.
- [109] <http://tr.wikipedia.org/wiki/Elazığ> Elazığ. 17.09.2014.

EKLER

Ek 1. Viterbi Algoritmasının Modele Uygulanması İçin Kullanılan Kodlar

```
hmm=initHMM(c("isim","fiil"),c("A","B","C","D","E","F","G","H"),c(.85,.15),matrix
(c(.84,.89,.16,.11),2),matrix(c(.61,.02,.08,.02,.10,.18,.09,.14,.07,.23,.04,.05,.
002,.25,.008,.11),2))
> print(hmm)
observations=c("A","B","C","F","G","E","A","E","E","E","A","C","F","A","G","A",
"A","E","C","C","A","C","A","B","F","A","A","C","A","A","C","G","A","D","A","A",
"A","C","A",
+"G","A","F","D","D","C","C","A","A","C","C","E","E","B","H","A","C","D","F",
+"E","A","A","A","A","A","E","A","A","A","F","A","D","A","A","D","G","E","D",
+"B","A","D","A","E","E","D","G","A","A","A","A","E","A","C","A","A","A","G",
+"A","A","E","F","D","D","A","F","A","A","A","A","D","B","B","G","A","E","F","D",
+"C","A","A","E","A","A","E","A","A","D","G","A","A","E","C","A","D","C","C","B",
+"A","E","A","C","B","G","A","E","A","A","A","A","D","C","A","C","A","A","D",
+"B","A","A","B","G","A","B","A","H","A","D","C","A","E","G","A","D","C",
+"A","E","C","B","A","E","A","F","A","B","A","A","A","D","C","A","B","D","C",
+"C","A","A","D","E","A","C","E","E","A","E","E","G","C","A","C","A","E","A","E",
"A","H","A","F","C","A","A","A","B","A","G","A","C","E","C","E","A","A",
+"A","A","A","C","A","D","A","A","A","H","A","B","C","A","C","H","C",
+"H","E","A","A","D","D","D","B","C","A","B","C","H","A","E","D","E","A",
+"G","A","A","A","B","A","G","C","A","A","D","D","A","A","D","A","A","A","C",
+"A","E","D","C","A","A","D","D","C","E","A","A","A","C","C","A","A","A","A","C",
"A","B","D","A","A","B","A","A","A","A","B","B","A","A","A","F","F","A","A",
+"F","B","A","C","A","A","A","B","C","A","D","E","A","B","C","F","A","B","A",
+"C","H","A","H","E","E","A","G","A","A","E","B","A","B","A","A","B",
+"A","F","C","E","A","A","A","B","A","D","A","E","A","A","A","F","A","E","C",
+"A","F","E","A","F","A","A","A","A","A","A","A","A","A","A","A","A",
+"A","A","A","A","A","A","A","A","C","H","D","D","A","A","A","A","A",
+"C","B","D","G","D","A","F","C","E","A","A","B","E","A","D","E","C","D","A","A",
+"A","A","A","A","H","A","A","D","A","A","A","A","A","A","A","A","A","E",
+"A","A","E","A","A","C","A","A","E","A","A","A","D","A","A","A","H","E",
+"A","C","D","C","A","B","E","A","D","D","A","A","F","A","A","A","G","A","C",
+"A","F","A","A","D","A","A","A","H","E","A","C","B","A","A","G","C","C",
+"A","A","A","B","D","A","A","E","A","A","D","A","A","B","D","C","A","A","D",
+"E","A","A","C","C","G","A","D","A","D","A","A","B","H","B","G","H","B",
+"A","F","A","A","A","A","A","A","B","A","A","A","A","A","A","A","A",
+"A","A","B","A","A","A","A","A","A","A","A","A","A","A","A","A","B",
+"A","A","A","G")
viterbi=viterbi(hmm,observations)
> print(viterbi)
```

Ek 2. Viterbi Algoritması Sonucunda Tahmin Edilen Durum Dizisi

[illegible]

Ek 3. Uygulamada Kullanılan Diğer SQL Kodları

Ek Şekil 3. 1. Harput ve iklim enformasyon alan kelimelerinin sorgulanması için düzenlenen SQL kodları

```
SELECT  
  
BİLGİ.Kelime_sıra_no, BİLGİ.Kelime_türü_grubu,  
BİLGİ.Cümle_no, BİLGİ.Cümle_içi_sıra_no,  
BİLGİ.Elazığ, BİLGİ.Harput, BİLGİ.İklim,  
BİLGİ.Sanayi, BİLGİ.Tarım, BİLGİ.Kelime,  
BİLGİ.bayes_isim_pr, BİLGİ.bayes_fiil_pr  
  
FROM BİLGİ  
  
WHERE (((BİLGİ.Kelime_türü_grubu)="Fiil grubu")  
AND ((BİLGİ.Harput)=Yes) AND  
((BİLGİ.Sanayi)=Yes));
```

Ek Şekil 3. 2. Harput ve sanayi enformasyon alan kelimelerinin sorgulanması için düzenlenen SQL kodları

```
SELECT  
  
BİLGİ.Kelime_sıra_no, BİLGİ.Kelime_türü_grubu,  
BİLGİ.Cümle_no, BİLGİ.Cümle_içi_sıra_no,  
BİLGİ.Elazığ, BİLGİ.Harput, BİLGİ.İklim,  
BİLGİ.Sanayi, BİLGİ.Tarım, BİLGİ.Kelime,  
BİLGİ.bayes_isim_pr, BİLGİ.bayes_fiil_pr  
  
FROM BİLGİ  
  
WHERE (((BİLGİ.Kelime_türü_grubu)="Fiil grubu")  
AND ((BİLGİ.Harput)=Yes) AND  
((BİLGİ.Sanayi)=Yes));
```

Ek Şekil 3.3. Harput ve tarım enformasyon alan kelimelerinin sorgulanması için düzenlenen SQL kodları

```
SELECT  
  
BİLGİ.Kelime_sıra_no,BİLGİ.Kelime_türü_grubu,  
BİLGİ.Cümle_no, BİLGİ.Cümle_içi_sıra_no,  
BİLGİ.Elazığ, BİLGİ.Harput, BİLGİ.İklim, BİLGİ.Sanayi,  
BİLGİ.Tarım, BİLGİ.Kelime, BİLGİ.bayes_isim_pr,  
BİLGİ.bayes_fiil_pr  
  
FROM BİLGİ  
  
WHERE (((BİLGİ.Kelime_türü_grubu)="Fiilgrubu")  
AND ((BİLGİ.Harput)=Yes) AND ((BİLGİ.Tarım)=Yes));
```

Ek Şekil 3.4. İklim ve sanayi enformasyon alan kelimelerinin sorgulanması için düzenlenen SQL kodları

```
SELECT  
  
BİLGİ.Kelime_sıra_no,BİLGİ.Kelime_türü_grubu,  
BİLGİ.Cümle_no, BİLGİ.Cümle_içi_sıra_no,  
BİLGİ.Elazığ, BİLGİ.Harput, BİLGİ.İklim, BİLGİ.Sanayi,  
BİLGİ.Tarım, BİLGİ.Kelime, BİLGİ.bayes_isim_pr,  
BİLGİ.bayes_fiil_pr  
  
FROM BİLGİ  
  
WHERE (((BİLGİ.Kelime_türü_grubu)="Fiilgrubu")  
AND ((BİLGİ.İklim)=Yes) AND ((BİLGİ.Sanayi)=Yes));
```

Ek Şekil 3.5. İklim ve tarım enformasyon alan kelimelerinin sorgulanması için düzenlenen SQL kodları

```
SELECT  
  
BİLGİ.Kelime_sıra_no, BİLGİ.Kelime_türü_grubu,  
BİLGİ.Cümle_no, BİLGİ.Cümle_içi_sıra_no,  
BİLGİ.Elazığ, BİLGİ.Harput, BİLGİ.İklim, BİLGİ.Sanayi,  
BİLGİ.Tarım, BİLGİ.Kelime, BİLGİ.bayes_isim_pr,  
BİLGİ.bayes_fiil_pr  
  
FROM BİLGİ  
  
WHERE (((BİLGİ.Kelime_türü_grubu)="Fiilgrubu")  
AND ((BİLGİ.İklim)=Yes) AND ((BİLGİ.Tarım)=Yes));
```

Ek Şekil 3.6. Elazığ, Harput, sanayi, iklim ve tarım enformasyon alan kelimelerinin sorgulanması için düzenlenen SQL kodları

```
SELECT  
  
BİLGİ.Kelime_sıra_no, BİLGİ.Kelime_türü_grubu,  
BİLGİ.Cümle_no, BİLGİ.Cümle_içi_sıra_no,  
BİLGİ.Elazığ, BİLGİ.Harput, BİLGİ.İklim, BİLGİ.Sanayi,  
BİLGİ.Tarım, BİLGİ.Kelime, BİLGİ.bayes_isim_pr,  
BİLGİ.bayes_fiil_pr  
  
FROM BİLGİ  
  
WHERE (((BİLGİ.Kelime_türü_grubu)="Fiilgrubu")  
AND ((BİLGİ.Elazığ)=Yes) AND ((BİLGİ.Harput)=Yes)  
AND ((BİLGİ.İklim)=Yes) AND ((BİLGİ.Sanayi)=Yes)  
AND ((BİLGİ.Tarım)=Yes));
```

Ek Şekil 3.7. Elazığ, Harput, iklim enformasyon alan kelimelerinin sorgulanması için düzenlenen SQL kodları

```
SELECT BİLGİ.Kelime_sıra_no,  
BİLGİ.Kelime_türü_grubu, BİLGİ.Cümle_no,  
BİLGİ.Cümle_içi_sıra_no, BİLGİ.Elazığ, BİLGİ.Harput,  
BİLGİ.İklim, BİLGİ.Sanayi, BİLGİ.Tarım, BİLGİ.Kelime,  
BİLGİ.bayes_isim_pr, BİLGİ.bayes_fiil_pr  
  
FROM BİLGİ  
  
WHERE (((BİLGİ.Kelime_türü_grubu)="Fiilgrubu")  
AND ((BİLGİ.Elazığ)=Yes) AND ((BİLGİ.Harput)=Yes)  
AND ((BİLGİ.İklim)=Yes));
```

Ek Şekil 3.8. Elazığ, Harput, iklim ve sanayi enformasyon alan kelimelerinin sorgulanması için düzenlenen SQL kodları

```
SELECT BİLGİ.Kelime_sıra_no,  
BİLGİ.Kelime_türü_grubu, BİLGİ.Cümle_no,  
BİLGİ.Cümle_içi_sıra_no, BİLGİ.Elazığ, BİLGİ.Harput,  
BİLGİ.İklim, BİLGİ.Sanayi, BİLGİ.Tarım, BİLGİ.Kelime,  
BİLGİ.bayes_isim_pr, BİLGİ.bayes_fiil_pr  
  
FROM BİLGİ  
  
WHERE (((BİLGİ.Kelime_türü_grubu)="Fiilgrubu")  
AND ((BİLGİ.Elazığ)=Yes) AND ((BİLGİ.Harput)=Yes)  
AND ((BİLGİ.İklim)=Yes) AND ((BİLGİ.Sanayi)=Yes));
```

Ek 4. Modelin Uygulandığı Metin İçeriği

Elâzığ ili, MÖ-3000'li yıllarda kurulduğu sanılan, Harput kentinin ovadaki devamıdır. Bu nedenle yıllarca Harput denmiştir. Büyük Selçuklu hakimiyetinin Anadolu'ya kayması ile Harput'un Türk yurdu olmasında en önemli savaşın Malazgirt Meydan Muharebesi olduğuna şüphe yoktur. Nitekim Harput ve çevresi Malazgirt muharebesinden sonra Türklerin eline geçmiş olup yörede Büyük Selçuklu Devleti'ne bağlı olarak Çubukoğulları Beyliği kurulmuştur (1085). Harput'un Türkler tarafından alınmasına kadar sadece müstahkem bir kale hüviyetinde kalan bu yer Türklerle beraber büyüyen bir şehir haline gelmiştir. Çubukoğulları Beyliği'nin ömrü uzun sürmemiş (1110) yılında Artukoğulları dönemini başlatmıştır. Bir müddet sonra Harput Artukluları diye bilinen bağımsız bir beylik kurulmuştur. Harput, (1230) yılında Moğolların eline geçmiş (1234) yılından itibaren ise Anadolu Selçuklu Devleti'nin hakimiyeti altına girmiştir. (1507) yılında Safevilerin eline geçen Harput, 1515 yılında ise Çaldıran zaferinden sonra Osmanlı hakimiyetine girmiştir (1516). Mezra denilen bugünkü yerleşim yerine 1834'de taşınan Elâzığ'a 1862 yılında Sultan Abdulaziz'in tahta çıkışının 5. yılında Hacı Ahmet İzzet Paşa devrinde buraya tayin edilen Vali İsmail Paşa'nın teklifi ile *Mamuret-ül-Aziz* ismi verilmiştir. Fakat telaffuzu güç olduğundan halk arasında kısaca *El'aziz* olarak söylenegelmiştir. Zaman içinde bölgeye eyalet merkezliği yapan şehre 1937 yılında Atatürk tarafından tahıl ambarı bolluk ve bereket anlamına gelen *El'azık* adı verilmiş olup, zamanla Türkçe ses uyumuna uygunluğu ve söyleniş kolaylığı nedeniyle *Elâzığ* olarak kullanılır olmuştur.

Geçmişte karasal iklimin hüküm sürdüğü Elâzığ, yapılan ve yapılmakta olan barajların etkisi ile ılıman bir iklime geçiş yapmıştır. Bu sürecin sonucunda özellikle önceleri çok soğuk ve yoğun kar yağışlı geçen kışlar nispeten daha ılıman geçmektedir. Elâzığ, toprağı verimli bir ovaya kurulmuştur. Elâzığ'ın dağlarında Karadenizdeki gibi geniş ağaçlıklar yoktur. Bunun nedeni Elâzığ'ın karasal iklimi yaşaması yüzündendir. Ancak Yapılan barajlar nedeniyle Elâzığ belirtildiği gibi ılıman bir iklime geçiş yapmıştır. Normalde Doğu Anadolu Bölgesi'nde yazlar çok sıcak, kışlar ise çok soğuk geçer. Ancak yapılan barajlar Elâzığ'ın Doğu Anadolu Bölgesi'nde yaşanan iklimin aynısını değil, (Yani karasal iklimin aynısı değil çünkü Doğu Anadolu Bölgesi karasal iklimi vardır.) ancak tam tersi hafif ılıman bir iklim yaşar. Elâzığ'a her zaman kar yağmamasının kışların çok soğuk geçmemesinin sebebi budur.

Ekonomisi sanayi, tarım ve ticarete dayanır. Keban Barajı'nın yapılmasından sonra tarıma elverişli toprakların bir kısmı su altında kaldığından, tarım alanlarının azalması paralelinde sanayi canlanmıştır. Gayri safi gelirinin %30'u sanayi, %10'u ticaret ve %25'i tarım sektöründen elde edilir. Toprak altı ve üstü çok zengindir.

Tarım: Ovaları az, fakat çok verimlidir. Bol suları bulunan büyük akarsuların suladığı bu ovalarda, buğday, arpa, pirinç, şekerpancarı, tütün, fasulye, nohut, mercimek, fiğ, burçak, soğan, sarımsak, pamuk, üzüm, elma, armut, kayısı, ceviz, badem ve dut yetişir.

Yetiştirilen ürünler arasında lahana, kavun ve çilek önemli gelir kaynağı haline gelmiştir.

Madencilik: Elâzığ madenciliğin tarımla yarıştığı ve hatta tarımı geçtiği bir yerdir. Toprakları madenle doludur. Bakır, krom, simli kurşun ve betonit başlıcalarıdır. Ergani Bakır İşletmesi'nde blister bakır, sülfürik asit ve prit tüvenan cevher istihsal edilir. Diğer maden işletmeleri Guleman Krom İşletmesi, Ferro Krom Tesisleri ve Elâzığ Betonit Fabrikası'dır. Alacakaya ve Arıcak ilçelerinde çıkarılan mermer dünyaca meşhurdur. Kendine has özelliği bulunan Elâzığ mermerini işlemek üzere son senelerde birçok mermer işleme fabrikası kurulmuştur.

Sanayi: Elâzığ'ın maden bakımından zengin ve Türkiye'nin en büyük hidroelektrik santrallerinden birinin bu ilde oluşu ile sanayi gelişmiştir. İrili ufaklı 1200 sanayi iş kolu vardır. Elâzığ sanayi alanında Doğu Anadolu Bölgesi'nde önemli bir yere sahiptir. Özellikle Organize Sanayi Bölgesi'nin kurulması ile fabrika sayısı hızla artmıştır. 49 fabrikalık sanayi bölgesinde 20 fabrika inşaatı tamamlanarak üretime geçmiştir. Diğerlerinin inşaatı devam etmektedir.

Un, deri, şeker, çimento, pamukyağı, pamuk ipliği, kiremit, yün, süt, yem, azot, süper fosfat, kireç, plastik boru, tüpgaz imalatı ve dolum, kâğıt, tekstil, meşrubat, matbaacılık, mermer, ayçiçek yağı, ayakkabı, mobilya, sabun, tıbbi malzeme fabrikaları başlıca büyük sanayi kuruluşlardır [109].

ÖZGEÇMİŞ

1985 yılında Adana’da doğdum. İlk, orta ve lise öğrenimimi Adana’da tamamladım. 2004-2008 yılları arasında Muğla Üniversitesi Fen Edebiyat Fakültesi İstatistik Bölümünde lisans öğrenimimi tamamladım. 2009 yılında Fırat Üniversitesi Fen Bilimleri Enstitüsü İstatistik Anabilim Dalı İstatistiksel Bilgi Sistemleri Bilim Dalında yüksek lisans öğrenimine başladım. 2011 yılında Fırat Üniversitesi Fen Bilimleri Enstitüsü İstatistik Anabilim Dalı İstatistiksel Bilgi Sistemleri Bilim dalında doktora eğitime başladım. 2009 yılında Fırat Üniversitesi Fen Fakültesi İstatistik Bölümüne araştırma görevlisi olarak atandım. Halen aynı bölümde araştırma görevlisi olarak görev yapmaktayım.