İSTANBUL TEKNİK ÜNİVERSİTESİ ★ FEN BİLİMLERİ ENSTİTÜSÜ

TÜRKÇE METİNLER İÇİN KONU BELİRLEME SİSTEMİ

YÜKSEK LİSANS TEZİ

Müh. Fatih KESGİN
(504041516)

Tezin Enstitüye Verildiği Tarih : 25 Aralık 2006
Tezin Savunulduğu Tarih : 30 Ocak 2007

Tez Danışmanı : Prof.Dr. Eşref ADALI
Diğer Jüri Üyeleri Prof.Dr. Oya KALIPSIZ (Yıldız Teknik Ü.)
Yrd.Doç.Dr. Zehra ÇATALTEPE (İ.T.Ü.)

OCAK 2007

ÖNSÖZ

Tez çalışmam sırasında desteğini esirgemeyen hocam Prof.Dr. Eşref ADALI'ya teşekkürü bir borç bilirim.

Tez çalışmam süresince beraber çalıştığım İTÜ Doğal Dil İşleme Takımı üyelerine ilgileri ve destekleri için teşekkür ederim.

Ayrıca, tüm eğitim hayatım boyunca her zaman yanımda olan sevgili aileme ve tez çalışmam süresince desteğini eksik etmeyen sevgili arkadaşım Ayşe BALTACIOĞLU'na her şey için çok teşekkür ederim.

Aralık, 2006 **Fatih KESGİN**

İÇİNDEKİLER

KISALTMALAR

BE **:** Bilgi Erişimi

SDM **:** Sonlu Durum Makinesi

BÇ **:** Biçimbirimsel Çözümleme

TABLO LİSTESİ

Sayfa No

ŞEKİL LİSTESİ

Sayfa No

SEMBOL LİSTESİ

A : a veya e harfi yerine

C : c veya ç harfi yerine

D : d veya t harfi yerine

H : ı, i, u veya ü harfi yerine

I : ı veya i harfi yerine

( ) : içerisindeki harf ek içinde yer almayabilir

TÜRKÇE METİNLER İÇİN KONU BELİRLEME SİSTEMİ

ÖZET

Bilgi erişimi (BE), bilginin temsil edilmesi, saklanması, düzenlenmesi ve gerektiği zamanda erişilebilmesini mümkün hale getirmek için yöntemlerin geliştirildiği araştırma konusudur. Genel Ağ'ın (İnternet) yaygınlaşması ile sayısal olarak saklanan ve erişilmek istenen belgelerin sayısı her geçen gün artmaktadır. Bu durum, Bilgi Erişimi'ni günümüzde en çok ilgilenilen ve araştırılan konulardan biri haline getirmiştir.

Metin işleme BE uygulamaları arasında önemli bir yer tutmaktadır. Metin işleme uygulamalarının bir alt kümesi olan Metin Sınıflandırma doğal dil ile yazılmış metinlerinin içeriklerine göre ilgili kanallara yönlendirilmesi, e-posta iletilerinin önemli önemsiz olarak ayrıştırılması, ya da metinlerin konularının belirlenmesi gibi alanlarda uygulanmaktadır.

Doğal Dil İşleme, sözlü veya yazılı dili incelemek üzere, yazılım ya da donanım olarak bilgisayar sistemleri geliştirilmesi işlemini açıklayan bir terimdir. Bilgi Erişimi alanında ele alınan metinler doğal dil ile yazılmış olduğundan, Bilgi Erişimi sistemlerinin başarımını artırmak için Doğal Dil İşleme yöntemlerinden yararlanılması gerekmektedir.

Türkçede köke yapım eki getirilerek oluşturulan yeni kelimeye gövde, bir kelimeye eklenmiş olan çekim eklerinin çıkarılması ile kelimenin gövdesinin bulunması işlemine ise Gövdeleme denilmektedir. Türkçe gibi sondan eklemeli dillerde ise gövdeleme başarımı yüksek oranda etkileyen aşamalardan biri olmaktadır.

Metin sınıflandırma, yazılı belgelerin içeriklerine bağlı olarak belirli sınıflara atanması işlemine verilen isimdir. Metin sınıflandırma işlemine örnek olarak bir kaynaktan gelen haberlerin konularına göre ayrıştırılması işlemi verilebilir.

Bu tezde, Türkçenin belirtilen özellikleri göz önüne alınarak, Türkçe bir metnin konusunun belirlenmesine yönelik algoritmalar gerçeklenen yazılımlarla birlikte tanıtılmıştır. Yapılan çalışmada, Bilgi Erişimi için gerekli olan ön çalışmalardan biri olan sözcüklerin yapım eklerinin korunarak çekim eklerinin atılması anlamına gelen gövdeleme işlemi için kullanılabilecek yöntemler karşılaştırılarak incelenmiş ve uygulanmıştır. Ön işlemlerden geçmiş olan metnin sınıflandırılması için gerekli sınıflandırma algoritmaları da incelenmiş ve uygulanmıştır.

TOPIC DETECTION SYSTEM FOR TURKISH TEXTS

SUMMARY

Information Retrieval (IR) is the research subject that deals with the representation, storage, organization and retrieval of information. With the increasing number of documents available online, information retrieval is becoming more needed and important.

Text processing is one of the main subjects in IR. Text Classification, which is a subset of text processing, has many applications such as routing, spam e-mail detection or detecting topics of texts.

Natural Language Processing (NLP) is described as developing hardware or software systems in order to analyze spoken or written natural language. In the subject of text processing, since many texts are in natural laguage, NLP is used in order to improve performance.

Turkish is a agglutinative language and every word in Turkish has a root and affixes which are added do the root. Stem is used to describe a word that is derived from a root with a derivational affix. Stemming is the process of removing inflectional affixes while keeping derivational ones. In agglutinative languages like Turkish, stemming is a very important proccess that mostly affects the overall performance.

Text classification is the process of assigning a document into one or more classes with respect to its content. A system that classifies news texts with respect to their topics can be considered as a text classification system.

In this study, a text classification system for Turkish is explained incuding developed algorithms and software. Stemming algorithms, and text classification methods are researched, compared and implemented.

1. GİRİŞ

Bilgi erişimi (BE), bilginin temsil edilmesi, saklanması, düzenlenmesi ve gerektiği zamanda erişilebilmesini mümkün hale getirmek için yöntemlerin geliştirildiği araştırma konusudur. [1]. Günümüzde BE, modelleme, belge sınıflandırma, belgeler içerisinde arama, veri görselleştirme vb. alanlarda kullanılmaktadır. 1990'lı yılların başında Genel Ağ'ın (World Wide Web) yaygınlaşmaya başlaması ile ulaşılabilen bilginin miktarı çok hızlı bir biçimde artmıştır. Bu bilgi artışı beraberinde yeni sorunlar doğurmuştur. Kullanıcılar bir bilgiye ulaşmak için Genel Ağ'da yer alan belgeleri incelemeli ve aradıkları bilgiyi içerip içermediğine bakmalıdırlar. Genel Ağ'da yer alan belge miktarı düşünüldüğü zaman bu işlemin neredeyse faydasız olduğu söylenebilir. Bu durum, Bilgi Erişimi'ni günümüzde en çok ilgilenilen ve araştırılan konulardan biri haline getirmiştir.

Bilgi erişimi konuları arasında görüntü ve ses gibi bilgi kaynakları olmakla birlikte en çok ihtiyaç duyulan alanların başında yazılı belgelere erişim gelmektedir. Sayısal ortamda saklanan yazılı belgelerin birçoğu doğal dille yazılmış olduğundan, genel veri erişim yöntemleri, örneğin verilen bir kelimenin belgelerde geçip geçmediğine bakarak arama yapmak, arama sonucunda ya çok sayıda belgenin bulunması veya hiçbir belgenin bulunamaması ile sonuçlanmaktadır. Bu durum metinlerin temsil edilmesi ve işlenmesinde yeni yöntemler geliştirilmesini zorunlu kılar.

Doğal Dil İşleme, sözlü veya yazılı dili incelemek üzere, yazılım ya da donanım olarak bilgisayar sistemleri geliştirilmesi işlemini açıklayan bir terimdir. [2] Bu terimdeki Doğal sözü, insanlar tarafından konuşma ve yazıda kullanılan dili, matematiksel yazım ya da bilgisayar dillerinden ayırmak için kullanılmıştır. Bilgi Erişimi alanında ele alınan metinler doğal dil ile yazılmış olduğundan, Bilgi Erişimi sistemlerinin başarımını artırmak için Doğal Dil İşleme yöntemlerinden yararlanılması gerekmektedir.

Türkçenin Doğal Dil İşleme'de özel bir durumu vardır. Türkçe Doğal Dil İşleme ile ilgilenen birçok araştırmacının ilgisini çekecek kadar kurallı bir dildir. Türkçenin cümle, sözcük ve hatta harf seviyesinde geçmişten günümüze kadar korunarak gelmiş olan kuralları vardır. Türkçe Ural-Altay dil grubuna dâhildir ve sondan eklemeli bir dildir. Her kelimenin bir kökü ve ekleri vardır. Her ek bir anlam taşır ve eklendiği kelimenin anlamını o yönde değiştirir.

Bilgi Erişimi kapsamına giren araştırma alanlarından biri de Metin Sınıflandırma'dır. Metin sınıflandırma, incelenen metnin bir veya daha çok sınıftan hangisine girdiğinin bulunması işlemine verilen isimdir. Bu işlem, belirli bir konuya göre ilgili ya da ilgisiz gibi ikili bir sınıflandırma olabileceği gibi, metnin hangi konulardan bahsettiğinin bulunması gibi çoklu bir sınıflandırma işlemi de olabilir. Metin sınıflandırmanın olası kullanım alanları, bir metnin konusunu belirleme, belirli bir bilgi doğrultusunda ilgili metinleri seçme, e-postalar için önemli önemsiz ayrımı yapma, bir kaynaktan gelen metinleri içeriklerine göre süzme ve ilgili birimlere yönlendirme gibi uygulamalar olabilir.

Türkçenin sondan eklemeli bir dil olması Metin sınıflandırma gibi sözcüklerin sayısına ve geçiş sıklığına dayalı işlemlerde Doğal Dil İşleme yöntemlerinin kullanılmasını zorunlu kılmaktadır.

Bu tezde, Türkçenin belirtilen özellikleri göz önüne alınarak, Türkçe bir metnin konusunun belirlenmesine yönelik algoritmalar gerçeklenen yazılımlarla birlikte tanıtılmıştır. Yapılan çalışmada, Bilgi Erişimi için gerekli olan ön çalışmalardan biri olan sözcüklerin yapım eklerinin korunarak çekim eklerinin atılması anlamına gelen gövdeleme işlemi için kullanılabilecek yöntemler karşılaştırılarak incelenmiş ve uygulanmıştır. Ön işlemlerden geçmiş olan metnin sınıflandırılması için gerekli sınıflandırma algoritmaları da incelenmiş ve uygulanmıştır.

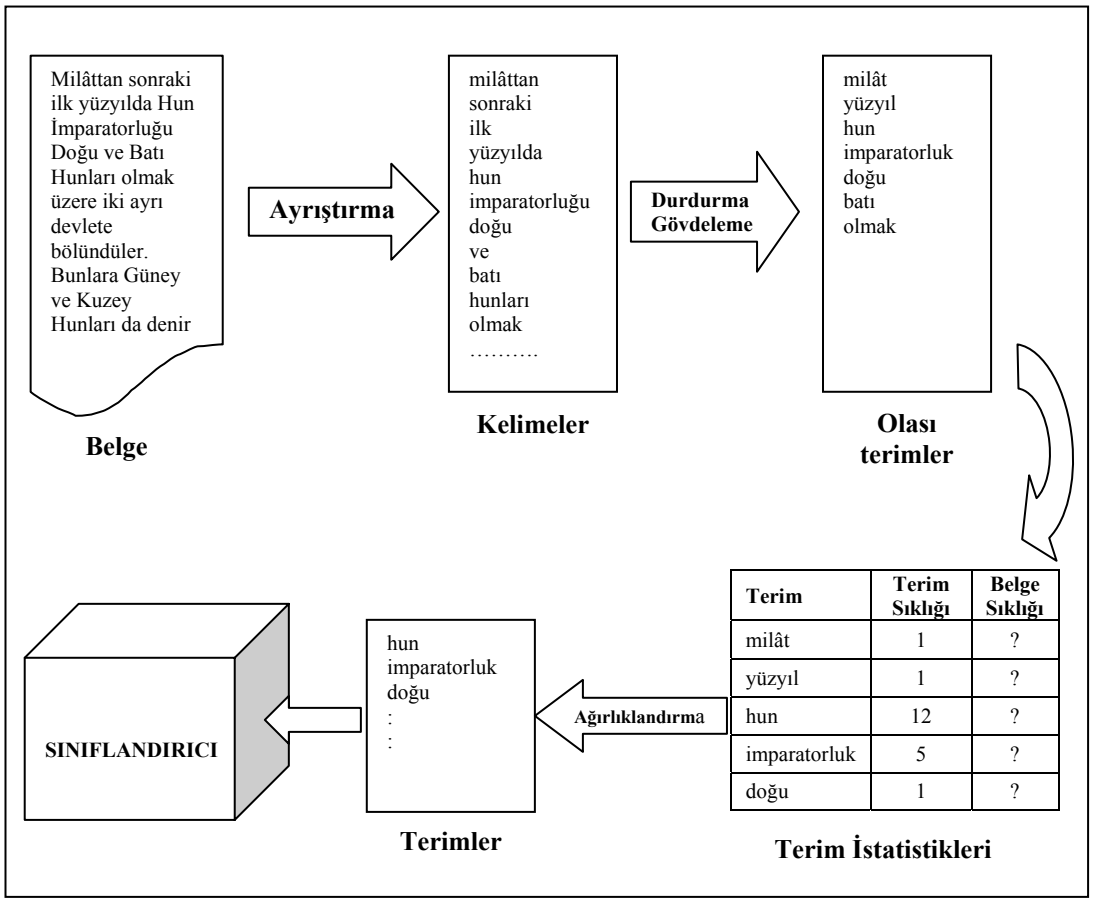
Türkçe için şimdiye kadar yapılan çalışmalar incelendiğinde, Bilgi Erişimi için kullanılan yöntemler çeşitli şekillerde uygulanmakla birlikte doğrudan konu belirleme gibi metin sınıflandırmaya yönelik bir çalışmaya rastlanmamıştır. Bu çalışma bu anlamda Türkçe için ilk çalışma olma özelliğine sahiptir.

Türkçe dışında özellikle İngilizce için yapılan çalışmaların ise Türkçenin sondan eklemeli yapısından dolayı birebir uygulanması mümkün değildir. Metin

sınıflandırma işlemi için önceden hazırlanmış bir veri kümesine ihtiyaç vardır. İngilizce için birden fazla veri kümesi kullanılabilir olduğu halde Türkçe için kullanılabilir özel amaçlı bir veri kümesi maalesef mevcut değildir.

Gerçekleştirilen çalışmanın ana adımları ve örnek veriler Şekil 1.1: *Metin Sınıflandırma Sistemi*'de gösterilmiştir.

Tez beş bölümden oluşmaktadır. İkinci bölümde Türkçenin genel kuralları incelenmiştir. Üçüncü bölümde şu ana kadar Türkçe için yapılmış olan gövdeleme çalışmaları incelenmiş ve iki farklı yöntemle geliştirilen gövdeleyiciler tanıtılmıştır. Dördüncü bölümde metin sınıflandırma sistemleri incelenmiş ve Türkçe için geliştirilen bir konu belirleme sistemi tanıtılmıştır. Son bölümde ise sonuçlar değerlendirilmiştir.



Şekil 1.1: Metin Sınıflandırma Sistemi

2. TÜRKÇE'NİN KURALLARI

Türkçe bitişken bir dil olduğu için kelimeler çoğunlukla bir kök ve köke eklenmiş ekler halinde bulunmaktadır. Her ek eklendiği köke farklı bir anlam kazandırmaktadır.

Türkçede ekler temelde ikiye ayrılır.

2.1 Türkçenin Kuralları

Türkçe Ural-Altay dil grubuna dâhil eklemeli bir dildir ve sondan eklemeli dillerin tipik bir örneğini teşkil eder. Türkçe Doğal Dil İşleme ile ilgilenenlerin dikkatini çekecek kadar kurallı bir dildir, çünkü kuralları oldukça kesindir ve yıllardan beri korunmuştur.

Türkçede köke eklenen her ek sözcüğe yeni bir anlam kazandırır. Türkçede ekler türlerine göre Yapım Ekleri ve Çekim Ekleri olmak üzere ikiye ayrılır. Çekim ekleri kalıplaşmış kullanımlar dışında yeni bir kelime türetmeyen, kelimeler arasında durum, iyelik, çokluk, kip, zaman, şahıs, gibi ilgiler kuran eklerdir. Yapım ekleri ise, eklendikleri kelimeden yeni bir kelime türeten eklerdir.

Türkçede yapım ve çekim ekleri sayıca fazladır. Bu zenginlik, diğer dillerde ayrı kelimeler olarak ifade edilen anlamların Türkçede bir ek ile ifade edilebilmesini mümkün hâle getirmiştir. Örneğin *yapmazsın* kelimesi ile *yapamazsın* kelimesi arasında sadece *a* harfi farklıdır ancak taşıdıkları anlam açısından iki kelime arasında belirgin bir ayrım vardır.

Türkçede aynı görevde yapım ekleri veya aynı çekim eki arka arkaya gelemez ancak farklı görevde yapım ekleri ve farklı çekim ekleri belirli kurallar içerisinde peş peşe gelebilir.

Türkçede çekim eki yapım ekinden sonra gelir ancak çekim ekinden sonra yapım eki gelmez. Yapım ekleri arasında sayılan *ki* eki bunun istisnasıdır. Ayrıca bazı kalıplaşmış çekim eklerinden sonra yapım eki gelebilir.

Örneğin,

Ayakkabıcı kelimesi *ayak+kap+ı+cı* şeklinde kök ve eklerine ayrıştırılabilir. Burada *ı* bir çekim eki olmasına karşın bu kelime içerisinde kalıplaşmıştır ve dolayısı ile sonrasında gelen *cı* yapım ekini alabilmiştir. Ancak *yemekkabıcı* gibi bir kullanım günümüz Türkçesinde söz konusu değildir.

Eklerin sözcüklere eklenmeleri aynı zamana ünlü ve ünsüz uyumu kuralına göre gerçekleşmektedir.

Ünlü uyumunda, ek içerisinde yer alan a ve e ünlüleri kalınlık-incelik hâline göre, ı, i, u, ü ünlüleri ise hem kalınlık-incelik hem de düzlük-yuvarlaklık durumuna göre eklendiği kök ile uyum gösterirler. Bu kurallara göre

a'dan sonra	:a, ı
e'den sonra	:e, i
ı'dan sonra	:a, ı
i'den sonra	:e, i
o'dan sonra	:a, u
ö'den sonra	:e, ü
u'dan sonra	:a, u
ü'den sonra	:e, ü

ünlüleri gelebilir. Örneğin *sepet-ler*, *sepet-çi* kelimelerinde *ler* ve *çi* eklerinde ünlüler bu kurala uygun olarak getirilmiş eklerdir.

Ünsüz uyumu kuralına göre, sözcük sonunda bulunan ünsüz eğer *p, ç, t, k, s, h, ş, f* harflerinden biri ise, *c, d, g* harfleri ile başlayan eklerin ilk ünsüzleri sertleşerek *ç, d, k* haline dönüşür. Örneğin *çarşaf-ta, yavaş-ça, kat-kı* kelimelerindeki ekler ünsüz uyumu gereği sertleşmişlerdir.

Yine ünsüz uyumu kuralına göre, sonunda *p, ç, t, k harflerinden* birini bulunduran bir sözcüğün sesli harf ile başlayan bir ek alması durumunda son harfi yumuşayarak

b,c,d,ğ harflerine dönüşür. Örneğin *kitab-ı*, *kâğıd-a*, *kazığ-ı* kelimelerinde son harfler ünsüz uyumu gereği yumuşamışlardır.

Türkçede bir kelimenin anlam ifade eden en küçük parçası kök olarak isimlendirilir. Bu köklere getirilen çeşitli sayıda ve sırada eklerle Türkçenin söz varlığı oluşmaktadır. Bu söz varlığı yeni eklemelerle zaman içinde genişlemektedir. Türkçede kökler isim ve fiil kökü olarak ikiye ayrılırlar. Yapım eki alarak bir kökten türemiş anlamına gelen *gövde* kelimeler ile kökler arasında kullanım bakımından hiçbir fark yoktur.

2.1.1 Yapım Ekleri

Yapım ekleri eklendikleri kelimenin anlamını değiştirerek yeni bir kelime türeten eklerdir. Günümüz Türkçesinde, iki yüze yakın yapım eki vardır. [3]

Yapım ekleri, isimden isim yapan, isimden fiil yapan, fiilden isim yapan, fiilden fiil yapan olmak üzere dörde ayrılırlar.

İsimden isim yapan ekler

Bir isme eklenerek yeni bir isim türeten eklerdir. Aynı ek olmamak şartı ile birden fazla isimden isim yapan ek peş peşe kullanılabilir. Bu ekler hem Türkçe kökenli hem de yabancı kökenli kelimelere getirilebilir. Bu ekler ile oluşturulmuş kelimelere örnekler şu şekildedir.

-ay, -ey : düzey, birey
-cı, -ci, cu, cü / çı, çi, çu, çü : arabacı, sucu
-lik, -lık, -luk, -lük : gözlük, kitaplık

İsimden fiil yapan ekler

İsim kök ve gövdelerinden, kökün taşıdığı anlamla ilişkili gövdeler oluşturan eklerdir. Sayıca az olan bir ek türüdür. En işlek olanı *–la, -le* ekidir.

-la, -le : kuzula-, şişmanla-, sabahla-, gözle-

Fiilden isim yapan ekler

Fiil köklerine eklenerek kalıcı veya geçici olarak isimler türeten eklerdir. Bu eklerden *–mak, -me, -ış* ekleri hem kalıcı hem geçici isimler türetebilirler.

-acak, -ecek	: açacak, gelecek
-ma, -me	: bölme, uçurtma
-mak, -mek (geçici)	: bilmek, çıkmak, duraklamak
-mak, -mek (kalıcı)	: çakmak, yemek

Fiilden fiil yapan ekler

Bu ekler eklendikleri fiilden yeni bir fiil türetebileceği gibi, eklendikleri fiilin anlamını değiştirmeden, özne veya nesnenin fiilin gerçekleşmesi ile ilgili durumunu değiştirerek çatılarını değiştirebilirler.

-ır, -ir, -ur, ür *(çatı)*	: artır-, göçür-
-ala, -ele	: kovala-, silkele-

2.1.2 Çekim Ekleri

Çekim ekleri yapım eklerinde göre daha fazla kullanılan ve eklendikleri kelimelere işlerlik kazandıran eklerdir. Eklendikleri kelimelerden yeni bir kelime türetmezler ancak kelimenin şahsı, iyeliği, zamanı, çokluğu gibi nitelikler üzerinde değişiklik yaparlar. Çekim ekleri *isim çekime ekleri* ve *fiil çekim ekleri* olmak üzere ikiye ayrılırlar.

2.1.2.1 İsim Çekim Ekleri

İsimlerin *durum*, *iyelik*, *çokluk*, *soru* ve *şahıs* bilgileri üzerinde değişiklik yapan eklerdir. Bu ek çeşitleri ve ilgili ekler şu şekildedir.

Durum Ekleri

İsimlerle, diğer isim ve fiiller arasında anlam ilgisi kuran eklerdir.

Belirtme durumu eki	: -ı, -i, -u, -ü
Bulunma durumu eki	: -da, -de / -ta, -te
Çıkma durumu eki	: -dan, -den / -tan, -ten
Eşitlik durumu eki	: -ca, -ce / -ça, -çe

İlgi durumu eki : -ın, -in, -un, -ün / -nın, -nin, -nun, -nün

Vasıta durumu eki : -la, -le

Yönelme durumu eki : -a, -e

İyelik Ekleri

Eklendikleri ismin hangi şahsa veya nesneye ait olduğunu bildirirler.

Birinci tekil kişi : -m

İkinci Tekil Kişi : -n

Üçüncü Tekil Kişi : -ı, -i, -u, -ü / -sı, -si, -su, -sü

Birinci Çoğul Kişi : -mız, -miz, -muz, -müz

İkinci Çoğul Kişi : -nız, -niz, -nuz, -nüz

Üçüncü Çoğul Kişi : -ları, -leri

Çoğul Eki

Eklendikleri isme çokluk özelliği kazandıran eklerdir. Aynı zamanda, eklendikleri kelimeye bazı özel durumlarda topluluk (*Osmanlılar*), saygı (*Vali Beyler*) gibi anlamlar da kazandırabilirler.

Çoğul Eki : -lar, -ler

Soru Eki

Eklendikleri isimle ilgili bir soru oluştururlar. Dil kuralı gereği eklendikleri kelime ile ayrı kendisinden sonra gelen kelime ile bitişik yazılırlar. Örneğin, *Evde miydiniz?*.

Soru Eki : -mı, -mi, -mu, -mü

Şahıs Ekleri

Ben : -ım, -im, -um, -üm

Sen : -sın, -sin, -sun, -sün

O : -dır, -dir, -dur, -dür / -tır, -tir, -tur, -tür

Biz : -ız, -iz, -uz, -üz

Siz : -sınız, -siniz, -sunuz, -sünüz

Onlar :-dırlar,-dirler,-durlar,-dürler/-tırlar,-tirler,-turlar,-türler

2.1.2.2 Fiil Çekim Ekleri

Fiiller ile isimler ve fiiller arasında geçici anlam ilişkisi kurmaya yarayan eklerdir. Eklendikleri fiilin anlamını kişi ve nesnelere bağlayabilir ya da, fiile *şekil*, *zaman*, *kişi* ve *soru* bilgisi eklerler. Şekil ve zaman ekleri *kip ekleri* olarak da isimlendirilirler.

Bildirme Kipleri

Görülen geçmiş zaman	: -dı, -di, -du, -dü / -tı, -ti, -tu, -tü
Duyulan geçmiş zaman	: -mış, -miş, -muş, -müş
Şimdiki zaman	: - yor
Kesin şimdiki zaman	: -mekte, -makta
Geniş zaman	: -r
Gelecek zaman	: -acak, -ecek

Tasarlama Kipleri

Emir	: *Ek kullanılmadığı zaman emir kipidir*
Gereklilik	: -meli, -malı
İstek	: -e, -a
Şart	: -se, -sa

Birleşik Zaman Çekimleri

Hikâye	: -dı, -di, -du, -dü / -tı, -ti, -tu, -tü
Rivayet	: -mış, -miş, -muş, -müş
Şart	: -se, -sa

Şahıs Ekleri

Ben	: -m, / -ım, -im, -um, -üm / -ayım, eyim
Sen	: -n, / -sın, -sin, -sun, -sün
O	: -dır,-dir,-dur,-dür /-tır,-tir,-tur,-tür /-sın,-sin,-sun,-sün
Biz	: -k / -ız, -iz, -uz, -üz / -alım, -elim
Siz	:-nız,-niz,-nuz,-nüz/-sınız,-siniz,-sunuz,-sünüz/-ın,-in, -un,-ün
Onlar	:- lar, ler / -sınlar, -sinler, -sunlar, -sünler

Sıfat Fiil Ekleri

Eklendikleri fiilden geçici sıfatlar oluştururlar. Örneğin *Okuyacağımız kitap*.

Sıfat fiil ekleri : -acak, -an, -ar, -ası, -dı, -dık, -maz, -mış, -r

Zarf Fiil Ekleri

Eklendikleri fiillerden geçici olarak zarf üreten eklerdir. Bu eklerle oluşturulan kelimeler kalıcı olmaya uygun değillerdir.

Zarf fiil ekleri :-a,-alı,-arak,-dığında,-dıkça,-ı,-ınca,-ıp,-ken,-madan

3. GÖVDELEME

Türkçede köke yapım eki getirilerek oluşturulan yeni kelimeye gövde ismi verilir. [3] Gövdeleme ise bir kelimeye eklenmiş olan çekim eklerinin çıkarılması ile kelimenin gövdesinin bulunması işlemine verilen isimdir.

Bilgi erişiminde bir metin içerisinde geçen kelimelerin sayısı o metin hakkında önemli bilgiler içerir. Metin içerisinde sıkça tekrar edilen bir kelimenin, metnin konusu ile ilgili olması oldukça olasıdır. Örneğin bu paragrafta *metin* kelimesi bu cümleden önce dört kez, *kelime, bir* ve *içerisinde* kelimeleri ikişer kez, diğer kelimeler birer kez geçmiştir. Bu kelime sayılarından yola çıkarak bu paragrafın metinlerden ve kelimelerden bahsettiğini söylemek yanlış olmayacaktır.

Bir kök veya gövde çekim eki aldığı zaman anlamı değişmese bile yazılışı değiştiği için eski kelimeden farklı yeni bir kelime haline gelir. Metinde hangi kelimenin kaç defa geçtiği sayıldığı zaman, bir kök veya gövdenin çekim eki almış hâli ile yalın hali ve başka çekim eki almış hâlleri farklı kelimeler olarak sayılacaktır. Anlam olarak düşünüldüğü zaman ise farklı yazılışları olan bu kelimeler aynı anlamı ifade etmektedir. Örneğin yukarıdaki paragrafta geçiş sayısı iki olarak verilen *kelime* sözcüğü aslında *kelimelerin* ve *kelimenin* olarak geçmektedir. Bu kelimelerin ikisinin de gövdesi *kelime* sözcüğüdür ve gerçekte de bu metinde *kelime* kavramından iki defa bahsedilmiştir.

Bu örnekten de anlaşılabileceği gibi Bilgi Erişimi algoritmalarının başarımı metnin içinde geçen kelimelerin gövdelerinin bulunmasına büyük oranda bağlıdır.

Gövdeleme işlemi, bir kelimeden çekim eklerinin çıkarılması olarak tanımlanır. Yapım ekleri eklendikleri sözcüğe yeni bir anlam kazandırdıkları için yapım eki eklenmiş her sözcük yeni bir gövde olarak kabul edilmelidir.

Örneğin *göz* kökünden türeyen kelimeleri ele alırsak, her türeyen kelimenin yine *göz* ile ilgili ancak farklı bir anlam ifade ettiği görülür.

göz	Görme yeteneği sağlayan organ
gözlük	Göz kusurlarını düzeltmeye yarayan araç
gözlükçü	Gözlük satan kişi
gözlükçülük	Gözlük satma işi

Gövdeleme işlemi dillere göre büyük farklılık göstermektedir. Örneğin İngilizce gibi eklerin kullanımının az olduğu bir dil için yalnızca ekler sözlüğüne bakılarak bir gövdeleyici geliştirmek mümkündür.[4] Bu şekilde geliştirilen bir gövdeleyicinin Türkçe gibi çok sayıda ek içeren bir dil için kullanılması mümkün değildir. Türkçe bitişken bir dil olmasından ötürü, eklerin sayısı ve eklenme çeşitleri daha detaylı bir inceleme yapılmasını zorunlu kılar. [5]

3.1 Gövdeleme Konusunda Daha Önce Yapılan Çalışmalar

Biçimbirimsel inceleyici kullanmayan yöntemler

Gövdeleme konusunda Türkçe için ilk çalışma Aydın Köksal tarafından 1981 yılında yapılmıştır[6]. Bu çalışmada tam bir gövde bulucu kullanılmamakla birlikte her kelimenin ilk beş harfi sözde gövde olarak kabul edilmiştir. Sözde gövde uzunluğu, birçok sayıda kelimenin deneysel olarak incelenmesi sonucunda beş olarak belirlenmiştir.

Bir diğer gövdeleme yöntemi ise en uzun eşleşme yöntemidir.[7] Bu yöntemde kelime önce bir sözlükte aranır, bulunamadığı takdirde kelime sonundan bir harf silinerek arama işlemi yeniden yapılır. Süreç bir gövdenin bulunması ya da kelimenin bir harf kalması durumunda sona erer. Bu yöntem uygulama açısından en basit yöntemlerden biri olmakla birlikte, kelime ile ilgisiz gövdeleri sonuç olarak bulma ihtimali vardır. Örneğin, *aksamaması* kelimesinin gövdesi bu yöntemle bulunmak istendiğinde gövde olarak *aksam* bulunmaktadır. Kelimenin gerçek gövdesi *aksa*-mak fiilinin kökü olan *aksa* kelimesidir.

Solak ve Can tarafından gerçeklenen bir başka yöntem[8], kelimenin sözlükte aranması, bulunamaması halinde sondan bir harf silinmesi ve peşinden yapısal inceleme yapılmasıdır. Eğer yapısal inceleme olumlu sonuç verirse bu gövde olası gövdeler arasına eklenmektedir. Süreç kelimenin tek bir harf kalması halinde sona ermektedir. Bu yöntem ile bir kelime için birden fazla olası gövde bulunması hâlinde bu gövdeler içinden birisi seçilmelidir.

Duran tarafından yapılan çalışmada [9] ise bir kelime soldan sağa incelenmekte ve elde edilen harf katarı sözlükte aranmaktadır. Eğer bir eşleşme bulunur ise kelimenin seçili harf katarı dışında kalan kısmı ek olarak kabul edilerek biçimbirimsel inceleme yapılmaktadır. Yapılan inceleme sonucunda doğru olarak çözümlenebilen sonuçlarda biri gövde olarak seçilmektedir.

Biçimbirimsel İnceleyici Kullanan Yöntemler

Türkçe gibi zengin bir biçimbirimsel yapısı olan dillerde gövdeleme işlemi için başvurulan yöntemlerden biri de biçimbirimsel çözümleme kullanmaktır. Biçimbirimsel çözümleyiciler, verilen bir sözcüğün tüm olası kök ve ek birleşimlerini üretirler [10]. Bu durumda gövdeleme işlemi bu olası sonuçlar içinden uygun gövdeyi seçme işlemi olmaktadır.

Biçimbirimsel çözümleyici kullanarak gövdeleme işlemi Altıntaş ve Can [10][11] tarafından gerçekleştirilen çalışmada denenmiştir. Bu çalışmada Oflazer tarafından geliştirilen biçimbirimsel çözümleyici [10] kullanılmış ve elde edilen sonuçlar içersinden *ortalama gövde uzunluğu* ve *n-gram* yöntemleri ile doğru gövde belirlenmeye çalışılmıştır.

3.2 Gerçeklenen Gövdeleyiciler

Tez kapsamında yapılan çalışmada, her iki yöntemi kullanan gövdeleyiciler gerçeklenerek denenmiştir. Biçimbirimsel çözümleyici olarak Oflazer tarafından geliştirilen [10] biçimbirimsel çözümleyici kullanılmıştır.

3.2.1 Biçimbirimsel Çözümleme Kullanarak Gövdeleme

Biçimbirimsel çözümleyiciler bir sözcüğü kök ve eklerine ayrıştıran sistemlerdir.

Türkçe için sık verilen ve özel olarak düşünülmüş bir örnek ve ayrıştırılmış hâli şu şekilde olacaktır.

OSMANLILAŞTIRAMAYABİLECEKLERİMİZDENMİŞSİNİZCESİNE

Bu tek kelimenin kök ve eklerine ayrılmış hâli aşağıdadır. Burada kelimenin kökü olan *osman* Türkçe kurallarına uygun olarak *on iki* adet ek almıştır.

OSMAN+LI+LAŞ+TIR+AMA+YABİL+ECEK+LER+İMİZ+DEN+MİŞ+SİNİZ+CESİNE

Türkçe sondan eklemeli bir dil olduğu için ekler açısından oldukça zengindir. Türkçenin biçimbirimsel kuralları karmaşık olmasına karşın, sonlu durum ile tanımlanabilmektedir. Her ek belirli bir anlam taşımakta ve eklendiği kelimeye bu anlamı kazandırmaktadır. Ekler kelimelere eklenirken ünlü ve ünsüz uyum kuralları gereğince bazı harfleri değişir. Ek içerisinde yer alan sesli harfler, kendisinden önce gelen sesli harf ile uyum içinde olmak zorundadır. Bazı belirli durumlarda köklerdeki ve eklerdeki sesli harflerin düşerek kelimeden çıkarılır. Benzer şekilde sesiz harfler de kimi zaman değişikliğe uğrayabilir ya da tamamen silinebilir. Bunun yanı sıra yabancı dillerden gelmiş olan kelimeler istisnai durumlar oluşturabilmektedir.

Biçimbirimsel çözümleyiciler tüm bu yukarda sayılan durumları da göz önünde bulundurarak kelimelerin çözümlemesin yaparlar ve belirli bir düzende sonuçları üretirler. Örneğin verilen *bakanlığı* kelimesi için biçimbirimsel çözümlemenin sonuçları Şekil 3.1'de görülmektedir.

```
İncele > Bakanlığı

1. Bakanlığı    bakanlık        +Noun+A3sg+P3sg+Nom
2. Bakanlığı    bakanlık        +Noun+A3sg+Pnon+Acc
3. Bakanlığı    bakan   +Noun+A3sg+Pnon+Nom^DB+Noun+Ness+A3sg+P3sg+Nom
4. Bakanlığı    bakan   +Noun+A3sg+Pnon+Nom^DB+Noun+Ness+A3sg+Pnon+Acc
5. Bakanlığı    bakan   +Adj^DB+Noun+Ness+A3sg+P3sg+Nom
6. Bakanlığı    bakan   +Adj^DB+Noun+Ness+A3sg+Pnon+Acc
```

Şekil 3.1: Türkçe Biçimbirimsel Çözümleyicinin Örnek Çıktısı

Bu örnekte görüldüğü üzere biçimbirimsel çözümleyiciler tüm olası sonuçları üretirler. Bu sonuçlardan hangisinin uygun gövde olduğunun belirlenmesi gereklidir.

Biçimbirimsel çözümleyici kullanıldığında uygun gövdeyi belirleme işleminde üstesinden gelinmesi gereken sorunlar ve olası çözümleri şu şekildedir.

Dilbilgisi açısından uygun gövdenin belirlenmesi

Gövdeleme işlemi yapım eklerini koruyarak çekim eklerinin çıkarılması işlemi olarak tanımlanmıştır. Biçimbirimsel çözümleyiciler tüm olası biçim bilgisini bulmak üzere hazırlandıkları için yapım eklerini de ayrıştırırlar. Bu durumda biçimbirimsel

çözümleyicinin ürettiği sonuç içerisinde yapım eklerinin ayrıştırılmamış olduğu çözümlemeleri seçmek gereklidir.

Örneğin *göz* kökü Türkçede birçok kelimenin kökü durumundadır, ancak *göz* kökünden türetilen bu kelimeler göz ile tamamen ilgisiz olmamakla birlikte yeni anlamlar taşırlar. Göz kökünden türetilmiş kelimelere örnekler şu şekilde sıralanabilir.

göz	Görme yeteneği sağlayan organ
gözlük	Göz kusurlarını düzeltmeye yarayan araç
gözlükçü	Gözlük satan kişi
gözlükçülük	Gözlük satma işi

Bilgi erişimi açısından, örneğin *gözlük* kelimesi ile arama yapan bir kullanıcı olduğu varsayılsın. Bu kullanıcı büyük ihtimalle *gözlüklerin çeşitleri*, *gözlük tarihçesi*, *gözlük fiyatları*, *markaları* gibi konularda bilgiye ulaşmak istemektedir. Bu kullanıcıya *göz* ile ilgili olan *gözün iç yapısı*, *göz renkleri*, *göz hastalıkları* gibi konularda bilgi içeren belgelerin sunulması büyük ihtimalle kullanıcının işine yaramayacaktır. Bu örnekten de anlaşılabileceği gibi, biçimbirimsel inceleme sonunda ulaşılan kelimelerden yapım eki içeren gövdelerin seçilmesi uygun olacaktır.

Anlamsal açıdan uygun gövdenin belirlenmesi

Dilbilgisi açısından uygun gövdenin belirlenmesine göre çözülmesi daha güç olan sorun anlamsal olarak doğru gövdenin seçilmesidir.Örneğin *başlar* kelimesi için biçimbirimsel çözümleyici çıktısı Şekil 3.2'de görülmektedir.

```
İncele > başlar

1. baş      baş+Noun+ A3pl+ Pnon+ Nom
2. başla    başla+Verb+ Pos+ Aor+ A3sg
3. başla    başla+Verb+ Pos+ Aor^DB+ Adj+ Zero
```

Şekil 3.2: Başlar Kelimesi için Biçimbirimsel Çözümleyici Çıktısı

Çözümleme sonuçlarından ilki *baş* köküne aittir. İsim olan baş kökünün çoğul eki –*lar* almış haline ait çözümlemedir. İkinci ve üçüncü çözümlemeler ise bir fiil olan *başla* köküne aittir.

Bu tür birden çok anlam ifade eden çözümlemelerde hangisinin kullanılmak istenen anlam olduğunu yalnızca biçimbirimsel çıktıya bakarak anlamak sadece bilgisayarla yapılan incelemelerde değil, insanlar için de mümkün değildir. Kullanılan anlamın bulunması için tüm cümlenin incelenmesi, hatta bazen paragraf veya metnin incelenmesi bile gerekmektedir.

Bu sorunun çözümü için farklı yaklaşımlar benimsenmekle birlikte, tez kapsamında en uzun gövdenin seçilmesi ve ortalama gövde uzunluğuna bakarak seçim yapılması şeklinde iki yöntem gerçeklenmiş ve sonuçları değerlendirilmiştir.

Biçimbirimsel çözümleme sonucunda en uzun kökün seçilmesi yöntemi

Bu yöntemde metin içerisinde yer alan her kelime öncelikle biçimbirimsel çözümleyici tarafından incelenir. Elde edilen çözümlemelerin kök kısımları incelenerek, harf sayısı bakımından en uzun olan kök uygun gövde olarak seçilmektedir.

Örneğin askerlik kelimesi için yapılan incelemede elde edilen sonuçlar şu şekilde olacaktır.

- askerlik askerlik+Noun+A3sg+Pnon+Nom
- askerlik asker+Noun+A3sg+Pnon+Nom^DB+Adj+FitFor
- askerlik asker+Noun+A3sg+Pnon+Nom^DB+Noun+Ness+A3sg+Pnon+Nom

Bu çözümleme sonucunda elde edilen kökler şu şekilde olacaktır.

- askerlik
- asker
- asker

Burada *askerlik* kökü *sekiz* harfli diğer kökler *beşer* harfli olduğu için gövde olarak askerlik seçilmiş olacaktır.

Bu yöntem uygulanması kolay olmanın yanı sıra, yapım eki alma hallerinde kelimenin harf sayısı yapım eki almamış haline göre uzayacağından türetilmiş kelimelerde en uzun türemiş hâlini seçmektedir.

Ortalama Kök Uzunluğuna Göre Seçim Yapma Yöntemi

Biçimbirimsel çözümleme sonuçları arasından uygun gövdeyi seçmek için uygulanabilecek yöntemlerden biri de Türkçe için ortalama gövde uzunluğunun belirlenmesi ve olası gövdeler arasında ortalama uzunluğa en yakın olan gövdenin seçilmesidir. Yapılan bir çalışmada, 1 Ocak 1997 ile 12 Eylül 1998 yılları arasında yayınlanan Milliyet Gazetesindeki metinler üzerinde yapılan incelemede, ortalama gövde uzunluğu tüm kelimeler için 4,58, tekil kelimeler için ise 6,58 olarak bulunmuştur[11].

Örneğin, tersine kelimesi biçimbirimsel çözümleyici ile incelendiğinde şu çözümlemeler elde edilir.

- Tersine ters +Adj^DB+Noun+Zero+A3sg+P3sg+Dat
- Tersine ters +Adj^DB+Noun+Zero+A3sg+P2sg+Dat
- Tersine tersine +Adverb
- Tersine tersin +Verb+Pos+Opt+A3sg

Elde edilen sonuçlar sırası ile *dört*, *yedi* ve *altı* harflidir. Ortalama uzunluklara göre seçilecek sonular ise şu şekilde olacaktır.

- 4,58 seçilmesi halinde *ters*
- 6,58 seçilmesi halinde *tersine*

Bu sonuçlardan hangisinin doğru olduğu kelimenin cümle içerisindeki kullanımına bağlıdır. Örneğin *kâğıdın tersine not aldı* gibi bir kullanımda *ters* gövdesinin

kullanımı söz konusu iken, *olumlu karar çıkması beklenirken tersine olumsuz bir karar çıktı* cümlesinde *tersine* gövdesi kullanılmıştır.

Tüm sözcükler için olan ortalama ve tekil sözcükler için ortalama uzunlukları ile yapılan deneylerde başarımın tekil sözcükler için ortalama uzunluk olan 6.58'in kullanılması halinde arttığı bildirilmiştir.[11]

3.2.2 Biçimbirimsel Çözümleme Kullanmadan Gerçekleme

Bilgi erişiminde önemli noktalardan biri de kullanıcının isteklerine en hızlı biçimde cevap verebilmektir. Bu amaca uygun olarak, gövdeme işlemlerini daha hızlı gerçeklemek üzere, bir yöntem geliştirilmiştir. Bu yöntemin aynı zamanda yukarıda sözü edilen doğru gövdeyi bulma işlemi için bir çözüm içermesi gerekmektedir. İlerleyen kısımlarda geliştirilen yöntem detayları ile açıklanmıştır.

3.2.2.1 Yöntemin Dayandığı Temeller

Uygulanması en kolay yöntemlerden biri En Uzun Eşleşme yöntemidir. Bu yöntemin algoritması Şekil 3.3'te açıklanmıştır.

1. Sözcüğü sözlükte ara
2. Eşleşen bir kök bulunursa adım 4'e git.
3. Sözcük sadece bir harf olarak kalmışsa, adım 5'e git. Aksi halde sözcüğün en son harfini çıkar ve adım 1'e git.
4. Bulunan kökü gövde olarak kabul et ve adım 6'ya git.
5. Sözcüğü bulunamayanlar listesine ekle.
6. Çık

Şekil 3.3: En Uzun Eşleşme Yöntemi Algoritması

Örneğin *eliyle* kelimesi için algoritmanın işleyişi Şekil 3.4'teki gibi olacaktır.

eliyle	→ Sözlükte Bulunamadı, tek harf değil, sondan bir harf çıkar
eliyl	→ Sözlükte Bulunamadı, tek harf değil, sondan bir harf çıkar
eliy	→ Sözlükte Bulunamadı, tek harf değil, sondan bir harf çıkar
eli	→ Sözlükte Bulunamadı, tek harf değil, sondan bir harf çıkar
el	→ Sözlükte bulundu, el kökünü gövde olarak kabul et.

Şekil 3.4: En Uzun Eşleşme Örneği

Şekil 3.4'teki örnekte gövde doğru olarak bulunmuştur. En uzun eşleşme yöntemi bazı sözcükler için anlamsız gövdeler de bulabilmektedir. Örneğin *keserek* kelimesi bu yöntemle gövdelendiğinde keser sözcüğü sözlükte bulunacağı için bu kelimenin gövdesi olarak bulunacaktır. Oysa kelimenin gövdesi *kes* fiildir.

Tez kapsamında geliştirilen yöntemde bu hataların önüne geçilmesi için bir önlem düşünülmüştür. Yeni yöntemde, en uzun eşleşme yönteminde olduğu gibi kelimenin sonundan harfler çıkarılırken sözlükte bir eşleşme olduğu anda sözcüğün kalan kısmının uygun bir ek olup olmadığı sınanır. Eğer kalan kısım uygun bir ek olabilirse sözcük gövde olarak kabul edilir.

En uzun eşleşme yönteminde, ele alınan *n* harfli bir sözcüğün, gövdesinin *p*, ek kısmının ise *t* harften oluştuğu varsayımı ile en iyi durumda *t* kez, en kötü durumda ise *n-1* kez sözlüğe erişilmesi gerekecektir. Geliştirilen yöntemde sözlüğe gereksiz erişimlerin olmaması için ek olarak kabul edilen katar öncelikle uygun bir ek olup olmadığı yönünde incelenir, eğer uygun bir ek ise sözlüğe erişilerek olası gövdenin sözlükte olup olmadığına bakılır. Bu yöntemde, en iyi durumda *bir* kez, sözcüğün sözlükte yer alan bir kelime olması hâlinde en kötü *t* kez, sözcüğün sözlükte olmaması halinde en kötü *n-1* kez sözlüğe erişilecektir. Türkçede eklerin tek başına anlam ifade etmediği düşünüldüğünde, sözlükte olmayan kelimelerin incelenmesi sırasında ek sınamasının olumsuz olarak sonuçlanacağı öngörülebilir. Örneğin *zeytin* kelimesinin sözlükte olmadığı varsayıldığında, *eksiz hal*, *-n* ve *-in* olası ekleri için sözlüğe erişilecek geri kalan olası ekler için ise sözlüğe erişilmeyecektir. Oysa en uzun eşleşme yönteminde *altı* harfli olan bu sözcük için *beş* kez sözlüğe erişilecektir.

Kelimenin harflerinin ek oluşturmayacak şekilde olması durumunda erişim sayısı düşecektir.

Yapılan bir testte, rastgele seçilen 500 kelime için, en uzun eşleşme yöntemi ile gövdelemenin 114999 ms, geliştirilen yöntemde ise 76999 ms sürdüğü gözlenmiştir. Bu sonuç, bu örnek için başarımda %50 dolayında bir başarım artışı olduğunu gösterir ki bu da yapılan öngörüyü doğrular niteliktedir.

Geliştirilen yöntemin algoritması Şekil 3.5'de gösterilmiştir.

1. Sözcüğü olası gövde ve ek olarak ayır. Ek kısmı için sınama yap. Eğer ek sınaması olumlu ise sözcüğü sözlükte ara, aksi halde adım 3'e git.
2. Eşleşen bir kök bulunursa adım 4'e git.
3. Sözcük sadece bir harf olarak kalmışsa, adım 5'e git. Aksi halde sözcüğün en son harfini çıkar ve adım 1'e git.
4. Bulunan kökü gövde olarak kabul et ve adım 6'ya git.
5. Sözcüğü bulunamayanlar listesine ekle.
6. Çık

Şekil 3.5: Geliştirilen Gövde Bulucunun Algoritması

Bu yönteme göre *keserek* sözcüğü incelendiğinde program akışı Şekil 3.6'daki gibi olacaktır.

keserek	→ Sözlükte ara, bulunamadı, sondan bir harf çıkar
kesere	→ Sözlükte ara, bulunamadı, sondan bir harf çıkar
keser	→ Sözlükte ara, bulundu, "ek" katarı için inceleme yap. Ek olamaz
	→ sondan bir harf çıkar.
kese	→ Sözlükte ara, bulundu, "rek" katarı için inceleme yap. Ek olamaz
	→ sondan bir harf çıkar.
kes	→ Sözlükte ara, bulundu, "erek" katarı için inceleme yap. Ek olabilir
	→ Sözcüğü gövde olark kabul et.

Şekil 3.6: Geliştirilen Gövde Bulucunun Örnek Çalışması

Şekil deki örnekte görüldüğü gibi bu yöntem ile *kese* ve *keser* gibi sözlükte olan kelimeler elenmekte ve doğru gövdeye erişilebilmektedir.

3.2.2.2 İsimlere Gelen Eklere Göre Sınama Yapılması

İsimlere gelebilen ekler için sınama yapılması yöntemi bu bölümde açıklanmıştır. Sınama yapılması işleminde ele alınan ek isim ekleri için tasarlanmış olan sonlu durum makinesinden [10] geçirilmektedir. Sonlu durum makinesinde her durum geçişi sırasında, geçişi sağlayan ek incelenen katardan çıkarılmaktadır. Sonlu durum makinesi çıkış durumuna ulaştığında, giriş katarı boş katar olarak kaldıysa, bir başka deyişle giriş katarında hiç harf kalmamışsa, ek geçerli ek olarak kabul edilmekte, aksi takdirde ek geçersiz olarak kabul edilmektedir.

Sonlu durum makinelerinde ve ekler tablosunda Sembol Listesinde yer alan semboller kullanılmıştır.

İsimlere eklenebilen ekler isim çekim ekleri ve ek-fiil ekleri olarak ikiye ayrılabilir.

İsim Çekim Ekleri

Türkçede var olan isim çekim ekleri Tablo 3.1: *İsim Çekim Ekleri*'de sıralanmıştır.

Tablo 3.1: İsim Çekim Ekleri

NO	EK	AÇIKLAMA	ÖRNEK
1	-lAr	Çoğul	Kalem-ler
2	-(H)m	1. Tekil kişi iyelik	Kalem-im
3	-(H)n	2. Tekil kişi iyelik	Kalem-in
4	-(s)H	3. Tekil kişi iyelik	Kalem-i
5	-(H)mHz	1. Çoğul kişi iyelik	Kalem-imiz
6	-(H)nHz	2. Çoğul kişi iyelik	Kalem-iniz
7	-lArı	3. Çoğul kişi iyelik	Kalem-leri
8	-(y)H	i hali	Kalem-i
9	-nH	i hali (3. tekil kişi iyelikten sonra)	Kalem-i-ni
10	-(n)Hn	tamlama	Kalem-in
11	-(y)A	e hali	Kalem-e
12	-nA	e hali (3. tekil kişi iyelikten sonra)	Kalem-i-ne
13	-DA	de hali	Kalem-de
14	-nDA	de hali (3. tekil kişi iyelikten sonra)	Kalem-i-nde
15	-DAn	den hali	Kalem-den
16	-nDAn	den hali (3. tekil kişi iyelikten sonra)	Kalem-in-den
17	-(y)lA	birliktelik	Kalem-le
18	-ki	İlgi	Kalem-in-ki
19	-(n)cA	görelik	Kalem-im-ce

Ek Fiil Ekleri

Ek-fiil ekleri isim soylu sözcüklere eklenebilirler. Ek-fiil eklerinin tablosu Tablo 3.2'de gösterilmiştir.

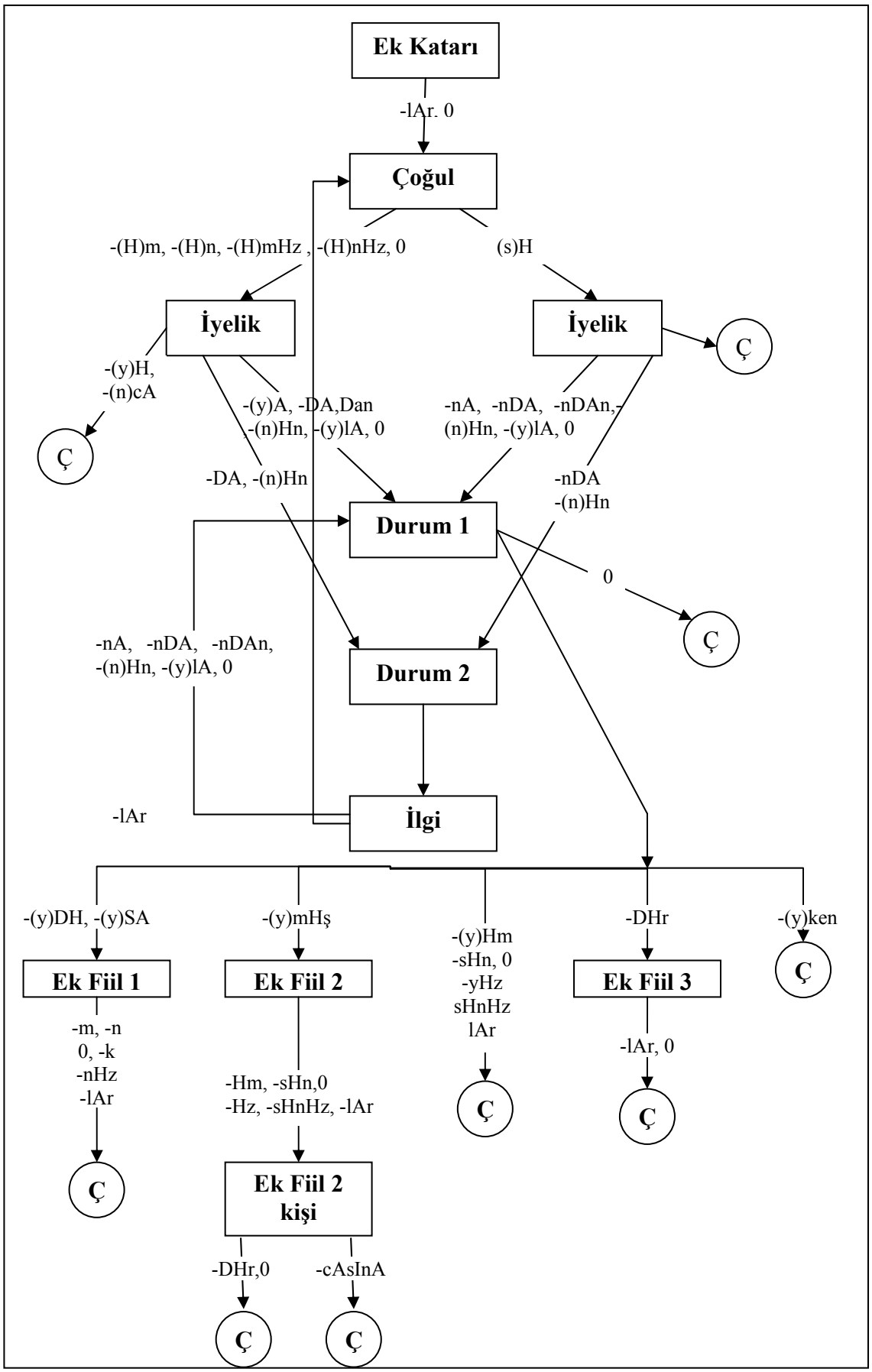
Tablo 3.2: Ek-Fiil Ekleri

NO	EK	AÇIKLAMA	ÖRNEK
1	–(y)Hm	1. tekil kişi	hasta-yım
2	–sHn	2. tekil kişi	hasta-sın
3	–(y)Hz	1. çoğul kişi	hasta-yız
4	–sHnHz	2. çoğul kişi	hasta-sınız
5	–lAr	3. çoğul kişi	hasta-lar
6	–m	1. tekil kişi ((y)DH ve (y)sA eklerinden sonra)	hasta-ydı-m
7	–n	2. tekil kişi ((y)DH ve (y)sA eklerinden sonra)	hasta-ysa-n
8	–k	1. çoğul kişi ((y)DH ve (y)sA eklerinden sonra	hasta-ysa-k
9	–nHz	2. çoğul kişi ((y)DH ve (y)sA eklerinden sonra	hasta-ydı- nız
10	–DHr	çevrik kip	hasta-dır
11	–cAsInA	tarz zarfı	hasta- ymışcasına
12	–(y)DH	di'li geçmiş zaman	hasta-ydı
13	–(y)sA	dilek-şart kipi	hasta-ysa
14	–(y)mHş	miş'li geçmiş zaman	hasta-ymış
15	–(y)ken	zaman zarfı	hasta-yken

Ek fiil ekleri *hasta-ymış-sın* örneğinde olduğu gibi yalın haldeki bir isme eklenebileceği gibi, *ev-de-ymiş-sin* örneğinde olduğu gibi çekim eki almış bir isme de eklenebilir. Burada ek-fiilden önce gelebilecek olan çekim ekleri bellidir. Bu ekler

1. -lAr, 0
2. -(H)m, -(H)n, -(H)mHz , -(H)nHz, 0
3. -(y)A, -DA,Dan ,-(n)Hn, -(y)lA, 0

eklerinin bir veya birkaç tanesinin peşpeşe eklenmesi ile oluşmuş ekler olabilir.

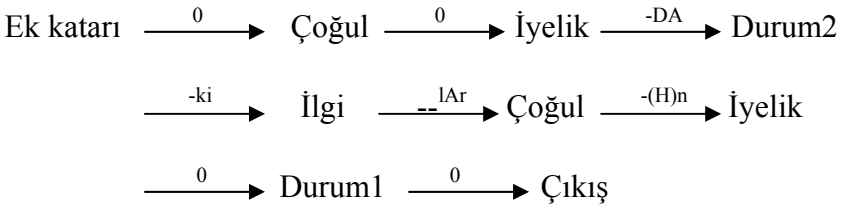


Şekil 3.7: İsimler İçin Sonlu Durum Makinesi

İsimlere eklenebilen ekler için oluşturulmuş sonlu durum makinesi Şekil 3.7'de görülmektedir. Bu sonlu durum makinesinde dikdörtgenler durumları, yuvarlaklar ise çıkış durumunu göstermektedir. Makinenin girişine geçerli bir ek olup olmadığı sınanmak istenen katar verilir. Makine katarı sol baştan başlayarak geçiş durumlarına göre inceler. Eğer geçiş durumu için yer alan ifade katar içerisinde bulunursa, uygun duruma geçilir ve geçişi sağlayan ek katardan çıkarılır. Eğer geçiş boş bir geçiş ise bu durumda sadece durum değiştirilir ancak giriş katarında bir değişiklik olmaz. Makinenin çalışması bir çıkış durumuna ulaşılana dek bu şekilde devam eder.

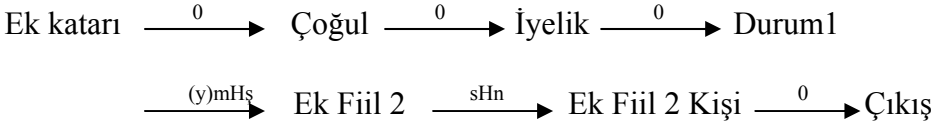
Örneğin *okuldakilerin* sözcüğünde *okul* gövdesi ve *-dakilerin* eki şeklinde bir bölünme olduğu durumda makinenin çalışması şu şekilde olacaktır.

- da - ki - ler - in



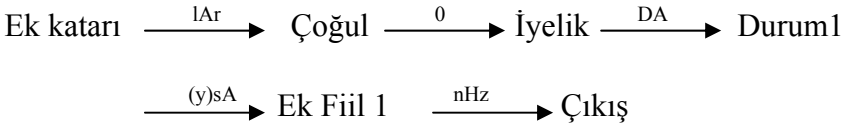
Ek-fiil eklerine örnek olarak *hastaymışsın* kelimesinin incelenmesine bakılabilir. Bu kelimenin gövdesinin *hasta*, ek kısmının *ymışsın* katarı olduğu durumda makinenin durum geçişi şu şekilde olacaktır.

-ymış-sın



Bir ismin önce çekim eki aldıktan sonra ek-fiil eki alması durumu da söz konusudur. Bu duruma örnek olarak *evlerdeyseniz* kelimesi incelenebilir. Bu kelimenin gövdesinin *ev* , ek kısmının *lerdeyseniz* katarı olduğu durumda durum geçişleri şu şekilde olacaktır.

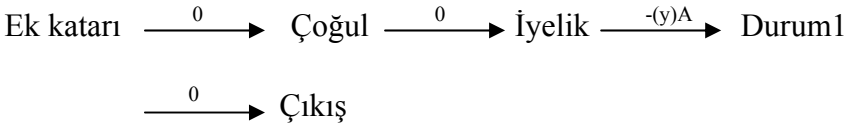
-ler-de-yse-niz



Bu sonlu durum makinesi ile isim köklü sözcüklere gelen eklerinin geçerli olup olmadığı sınanabilir. Çıkış durumuna ulaşıldığında ek katarındaki tüm harfler silinmiş ise, ek geçerlidir sonucuna varılır. Giriş katarı, boş katar haline gelmemiş ise bu durumda geçersiz bir ek sonucuna varılır. Yukarıda örneği verilmiş olan *dakilerin* katarı sonlu durum makinesinden geçtikten sonra boş katara dönüşeceği için geçerli bir ek olduğu sonucu üretilecektir.

Örneğin *aksaması* sözcüğü *aksam* gövdesi ve *–ası* eki şeklinde incelenmek istendiğinde durum geçişleri şu şekilde olacaktır.

- ası



Bu örnekte çıkış durumuna ulaşıldığında giriş katarı *–sı* olacaktır. Ekin geçerli olması için çıkış durumunda boş katara ulaşılması gerektiğinden bu ek geçerli bir isim çekim eki olarak kabul edilmeyecektir. Dolayısı ile bu sözcük için ulaşılan gövde *aksam* olamaz. Oysa *arabanın yürüyen aksamı* gibi bir kullanımda *–ı* eki sonlu durum makinesinden boş katar olarak çıkacağı için geçerli bir olarak kabul edilecek ve bu örnekteki kelimenin gövdesi *aksam* olarak kabul edilecektir.

3.2.2.3 Olası Gövdenin Sözlükte Aranması

Ek olması olasılığı olan katarın sonlu durum makinesi tarafından incelenmesi sonucunda geçerli bir ek olduğu anlaşılırsa, olası gövdenin sözlükte yer alıp almadığı sınanır. Ünsüz yumuşaması ve ünlü düşmesi sonucunda bazı gövdeler sözlükte bulunamayabilirler. Bu kuralların ele alınan gövde için geçerli olup olmadığı sınanarak, yumuşama veya ünlü düşmesine uğramış kelimeler belirlenir ve sözcük üzerinde gerekli değişiklik yapılarak bu kelimelerin gövdeleri bulunur.

Ünsüz Yumuşaması

Bölüm 2'de değinilmiş olan Türkçede ünsüz uyumu kuralı gereği gövdelerin son harflerine göre yumuşama durumu söz konusu olabilir. Ünsüz uyumu kuralına göre, sonunda *p, ç, t, k harflerinden* birini bulunduran bir sözcüğün sesli harf ile başlayan bir ek alması durumunda son harfi yumuşayarak *b,c,d,ğ* harflerine dönüşür. Örneğin *kitab-ı*, *kâğıd-a*, *kazığ-ı* kelimelerinde son harfler yumuşamıştır.

Gerçeklenen yöntemde, olası ek olan katar sonlu durum makinesinden geçirildikten sonra geçerli ek olabileceği sonucuna ulaşılmışsa, olası gövdenin son harfine bakılır. Bu harfin b,c,d,ğ harflerinden biri olması durumunda söz konusu harf, sert ünsüz olan karşılığı ile değiştirilir ve yeni gövde sözlükte aranır. Son harflerin değişimi şu şekilde olmaktadır.

- b → p
- c → ç
- d → t
- ğ → k

Örneğin *kitabı* kelimesi için *–ı* harfinin geçerli bir ek olup olmadığı sınanır ve geçerli olduğuna karar verilir. Gövde olabilecek katarın son harfi *b* olduğu için değiştirilerek *p* yapılır ve oluşan yeni *kitap* katarı sözlükte aranır. Bulunduğu takdirde gövde olarak kabul edilir.

Tek heceli kelimeler birkaç istisna dışında yumuşama kuralına uymazlar. Örneğin *tek* kelimesi *e* veya *i* ekini aldığında *teki* ve *teke* olarak yazılmakta ve okunmaktadır. Tek heceli kelimelerden bu kurala istisna oluşturan *gök*, *çok* gibi kelimeler bir istisnalar listesine eklenmiştir. Gövde olabilecek kelimenin bu istisnalar listesinde olup olmadığı sınanır.

Ünlü Düşmesi

Bazı sözcükler ek aldıkları zaman ünlü düşmesine uğrarlar. Örneğin *burun* kelimesi *burunu* olarak değil *burnu* olarak kullanılmaktadır. *Ağız*, *burun*, *göğüs*, *alın* gibi ünlü düşmesine uğrayan kelimeler için de bir istisnalar listesi oluşturulmuştur. Olası gövdeler sözlükte bulunamadıkları takdirde bu listede olup olmadığına da bakılır.

3.2.3 Gövdeleme Yöntemlerinin Değerlendirilmesi

Tez kapsamında biçimbirimsel çözümleyici kullanılarak ve kullanılmayarak iki farklı gövdeleyici gerçeklenmiştir.

Biçimbirimsel çözümleyici kullanılması halinde gövdeleme işlemi çözümleyicinin çıktısını inceleyerek uygun gövdenin seçilmesi işlemi hâlini almaktadır. Bu yöntem gerçeklemede büyük kolaylık sağlamakla birlikte elde edilen gövdelerin doğruluğu açısından da oldukça başarılıdır.

Gerçeklenmiş olan ikinci yöntemde ise amaç, hızlı ve düzgün çalışacak bir gövdeleyici gerçeklenmesidir. Biçimbirimsel çözümleyiciler harici programlar olarak kullanılmakta ve bu gövdeleme işlemi sırasında belirli bir gecikmeye sebep olmaktadır. Geliştirilen yöntemde bunun önüne geçmek için sonlu durum makineleri sadece doğrulama yapmak üzere kullanılmış ve sadece ek olabilecek sözcükler için sözlükte arama yaparak işlem olabildiğince hızlı hâle getirilmek istenmiştir.

4. METİN SINIFLANDIRMA

Metin sınıflandırma, yazılı belgelerin içeriklerine bağlı olarak belirli sınıflara atanması işlemine verilen isimdir [2]. Metin sınıflandırma işlemine örnek olarak bir kaynaktan gelen haberlerin konularına göre ayrıştırılması işlemi verilebilir.

Metin sınıflandırma kapsamına giren olası uygulama alanları şunlardır:

- *Yönlendirme*: Yönlendirme sistemleri bir kaynaktan gelen belgeleri bir veya daha çok konuya göre ayıran sistemlerdir.

- *İşaretleme*: Genellikle kütüphanecilikte kullanılan bir uygulamadır. Belirli bir sözcük haznesinden, belgelere uygun işaretçi kelimeler atanmasıdır. Bu işlemler genellikle insanlar tarafından yapılmaktadır ve belge sınıflandırma bu anlamda bir seçenek olabilir.

- *Sıralama*: Ayrıştırılmamış bir belge yığınındaki belgelerin ayrık kümelere bölünmesini sağlayan uygulamalardır. Örneğin e-posta iletileri ve günlük notlardan oluşan karışık belge yığınlarının ayrıştırılmasında kullanılabilirler.

- *Destekleme*: Bir kaynaktan gelen belgeleri uygun şekilde sıralama işlemine destek olacak sistemler tasarlanabilir. Örneğin, bir yayıncının kendisine gönderilen makaleleri konu başlıklarına göre düzenlemesi işleminde, gönderilen makalelerdeki kaynak gösterilen yayınlar, makale sahibinin önceki yayınları gibi bilgiler kullanılarak bu işlem daha kolay ve doğru hale getirilebilir.

Metin sınıflandırma sistemlerinde kullanılacak olan veriye bağlı olarak ele alınması gereken konular şunlardır:

- *Parçalanabilirlik*: Veri uzayı kaç sınıfa bölünmeli? Bu sınıflar belge uzayını nasıl bölüyorlar?

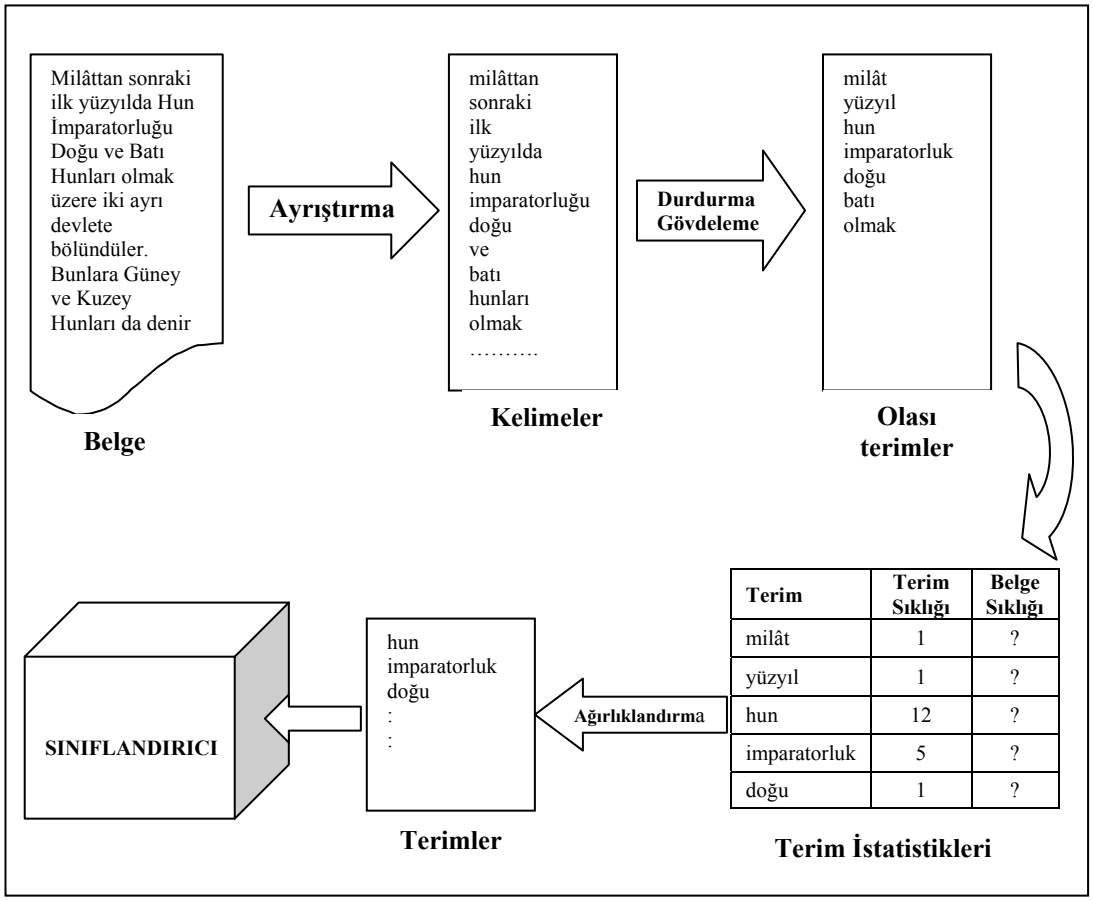
- *Boyutlandırma*: Sınıflandırma yapmak için kaç adet öznitelik kullanılacak? Tüm içerik kelimelerinin kullanılması halinde öznitelik uzayı çok büyük olacağından belgeler seyrek bir biçimde dağılmış olacaklardır.

- *Seçkinlik*: Sınıflandırma işlemi belgelerin sadece bir sınıfa ait olup olmadıklarına bakarak mı yapılacak, yoksa sınıf sayısı daha çok mu olacak?

- *Konusallık*: Ele alınan belgeler bir konudan mı yoksa birden fazla konudan mı bahsediyorlar?

Bu ve benzeri soruları göz önünde bulundurmak geliştirilecek olan sistemlerin başarılı olması için oldukça önemlidir.

Metin sınıflandırma işlemi gerçekleştirilirken Şekil 4.1'de görülen aşamalar sırası ile uygulanır.



Şekil 4.1: Metin Sınıflandırıcı Genel Yapısı

Ayrıştırma

Bir metin belgesi incelenmeden önce ön bir temizleme ve ayrıştırma işlemine tâbi tutulur. Bu işlemde öncelikle tüm harfler küçük harfe dönüştürülür ve belgede yer alan noktalama işaretleri çıkarılır. Eğer noktalama işareti kesme işareti olan ‘ ise bu işaretten sonra gelen kısım da ek olacağından metinden çıkarılır. Gerekli temizlemeler yapıldıktan sonra belge boşluklara göre kelimelere ayrılır.

Durdurma Listesi

Her dilde olduğu gibi Türkçede de çok sık kullanılan ve anlamsal olarak bir ayırıcılığı olmayan kelimeler vardır. Bu kelimeler durdurma listesi adı verilen bir listede toplanır ve metin içerisindeki her kelimenin bu listede olup olmadığına bakılır. Bu listede yer alan kelimeler sınıflandırma işlemi için bir yarar getirmeyeceğinden elenir.

4.1 Öznitelik Seçimi ve Terim Ağırlıklandırma

Metinler bir sınıflandırıcıya verilmeden önce incelenerek o belgeyi temsil eden öznitelikler belirlenmelidir. Bu işlem sayesinde sınıflandırıcılara tüm metin girdi olarak verilmez, yalnızca o belgeye ait öznitelikler gönderilir. Böylece gönderilecek olan verinin boyutunda belirgin bir azalma olmasının yanı sıra, belge içerisinde önemsiz olan, bir ayırıcılığı olmayan nitelikler elenecek böylece sınıflandırma işleminin başarımı artacaktır.

Terim Sıklığı

Bir kelimenin yer aldığı belge içerisinde kaç kez tekrarlandığı bilgisi *terim sıklığı* olarak isimlendirilmektedir. Bir belge içinde diğer kelimelere oranla daha çok kullanılan bir kelime muhtemelen o belgenin konusu hakkında ipucu verebilecek bir kelimedir. Örneğin bir metin içerisinde, *futbolcu*, *hakem*, *oyuncu* kelimeleri çok sık tekrarlanıyorsa bu belgenin konusu *futbol oyunu* olarak tahmin edilebilir.

Kelimeler, özellikle de isim soylu sözcükler çekim eki aldıkları zaman, ifade ettikleri anlam aynı kalmakta fakat yazılışları tamamen değişmektedir. Örneğin *hakem* kelimesi metin içerisinde *hakem*, *hakemler*, *hakeme*, *hakemi*, *hakemden* gibi çeşitli çekim ekleri almış halleri ile bulunabilir. Tüm bu değişik yazılışlı kelimeler aslında

hakem'den bahsetmektedir. Kelimelerin terim sıklıkları hesaplanırken bu kelimeler ayrı birer kelime olarak sayılmamalı, tüm bu kelimelerin toplam geçiş sayısı *hakem* kelimesinin terim sıklığı olmalıdır. Gövdeleme işlemi bu sebepten ötürü oldukça önemli ve başarımı büyük oranda etkileyen bir ön aşamadır.

Belge Sıklığı

Bir terimin belge koleksiyonu içerisinde yer alan belgelerde kaç kez kullanıldığına ilişkin sayıya belge sıklığı ismi verilir. Belirli bir konuya has olan kelimeler, o konudan bahseden bir belgede defalarca tekrarlanacaktır ancak o konudan bahsetmeyen belgelerde ise nadiren geçecektir. Bu durumda içerik belirten kelimeler için belge sıklığının düşük olması beklenir.

Normalizasyon

Belgelerin uzun olması durumunda terim sıklıkları kısa olan belgelere göre daha yüksek çıkacaktır. Bir belgenin daha uzun olması o belgenin daha kısa olan belgelerden daha önemli olduğu ya da bir konu ile daha ilgili olduğu anlamına gelmez. Uzun ve kısa olan belgeler arasında adil bir derecelendirme yapılması için terim sıklığı sayısı belgede geçen toplam sözcük sayısı dikkate alınarak normalize edilmelidir.

Bu üç ölçüyü birlikte kullanarak bir terimin o belge içerisindeki ağırlığı Denklem (4.1) ile hesaplanabilir.

$$a_i = \frac{TS_{i,b} \times \log(\frac{B}{BS_i})}{\sqrt{\sum_j \left[TS_{i,b} \times \log(\frac{B}{BS_i}) \right]^2}} \qquad \textbf{(4.1)}$$

4.2 Metin Sınıflandırma İçin Kullanılabilecek Yöntemler

Metin sınıflandırma işlemini yapmak üzere insan bilgisinin ve emeğinin dâhil olduğu yöntemler geliştirilebileceği gibi tamamen otomatik yöntemler de geliştirilebilir. Bu bölümde uygulanabilecek yöntemler incelenecektir.

Metin sınıflandırma için uygulanabilecek yöntemlerden ilki bilgi mühendisliği yaklaşımıdır [12]. Bu yöntemde sınıflandırma kuralları uzmanlar tarafından oluşturulur ve yeni gelen belgeler bu kurallara göre sınıflandırılabilir. Sınıflandırma kurallarının uzmanlar tarafından el ile oluşturulması yöntemi zor ve zaman alıcı bir işlem olacaktır. Editörler tarafından sınıflandırma amaçlı sorguların oluşturulması için iki günlük bir zaman dilimi gerekebilmektedir [2]. Bu sebepten bu yöntem birçok uygulama sahası için çok verimsiz ve elverişsiz olacaktır. Örneğin çok fazla sayıda sınıfın olduğu bir durumda bu sınıflar için kuralları belirlemek çok güç olabilir. Bunun yanı sıra, kendi belgelerini sınıflandırmak isteyen bir kullanıcı için uzman bilgisi var olmayacaktır. Ayrıca, sınıfların değişmesi durumunda kuralların gözden geçirilmesi ve tekrar oluşturulması gerekecektir.

Metin sınıflandırma işlemi için makine öğrenmesi yöntemlerini kullanmak tüm bu bahsedilen sorunların çözümü olabilir. Makine öğrenmesi yöntemlerinden en çok kullanılanlarından biri tümevarımsal öğrenme yöntemidir. [2] Bu yöntemde bir uygulama sınıflandırma yapmaktan çok, bir öznitelik uzayına bağlı olarak hazırlanmış ve etiketlenmiş verilerden sınıflandırma kurallarını öğrenebilir. Bu tip yöntemlere *denetimli öğrenme* ismi verilmektedir.

Denetimli öğrenme yöntemlerini kullanabilmek için bazı gereksinimler vardır. Bu gereksinimler sırası ile şöyledir:

- Belgelerin atanacağı sınıflar zaman içerisinde belirlenmiş olmalıdır.
- En basit durumda bu sınıfların birbirlerinden ayrık olması gereklidir.
- Sınıflar birbirinden ayrık değilse, n adet sınıfı, n adet alt sorun olarak değerlendirip, her alt sorunun belgeleri ilgili ya da ilgisiz olarak ayırması sağlanabilir. Bu yöntemde bir belge birden fazla sınıfa dâhil olabilir.

Makine öğrenmesi yöntemlerinden Naive Bayes Sınıflandırıcı ve En Yakın Komşu yöntemleri bu bölümde incelenecektir.

4.2.1 Naive Bayes Sınıflandırıcı

Saf Bayes olarak da isimlendirilebilecek bu sınıflandırıcıların çalışma mantığı şu şekildedir. Eğer hâlihazırda el ile ayıklanmış ve sınıflara atanmış bir miktar belge var ise, bu bilgiyi yeni gelen belgelerin sınıflandırılması için kullanabilecek yarı otomatik bir sistem kurulabilir.

Terimlerin belge içinde dağılımını hesaplayarak, yeni gelen belgeler için sınıf tahmininde bulunabilir.[2] Bu tahmini yapabilmek için iki durumun gerçekleşmiş olması gereklidir:

1. Bir sınıf verildiğinde terimlerin belge içerisinde bulunma olasılığı bilgisini, terimlerin belge içinde bulunma durumları bilindiğinde, bir sınıfa düşme olasılığına dönüştürmek gereklidir.

2. Bir belge veya sınıf ile ilişkilendirilmiş terimlerden elde edilen delillerin bir araya getirilmesi gereklidir.

Daha açık bir anlatımla bir sınıfa ait terimlerin olasılığı bilgisi olan $P(t \mid S_i)$ bilinirken bu bilgiyi bir terim için belirli bir sınıfa ait olma olasılığı bilgisi olan $P(S_i \mid t)$ şekline dönüştürmeli, daha da önemlisi B belgesinin bir sınıfta olma olasılığını hesaplamak üzere $P(S_i \mid T_B)$ bilgisine dönüştürmemiz gerekmektedir.

Öznitelikler arasında koşullu bağımsızlık olduğu varsayılarak Bayes kuralı uygulanabilir. Bu durumda denklem (4.2) elde edilecektir.

$$P(C_i \mid B) = \frac{P(B \mid C_i)P(C_i)}{P(D)} \qquad \textbf{(4.2)}$$

Bunun yanı sıra, hangi sınıflarda yer aldıkları bilinen eski belgelerden

1. Eski belgelere ait terimler

2. Yeni belgelerde görülmesi beklenen terim sıklığı bilgisi

çıkarılabilmelidir.

$B = (t_1 ....t_n)$ terim vektörü ile temsil edilen bir belge için $P(B \mid C_i)$ olasılığı denklem (4.3) ile hesaplanabilir.

$$P(B \mid C_i) = \prod_{j=1}^{j=n} P(t_j \mid S_i) \quad \textbf{(4.3)}$$

Hesaplanan bu bilgiyi kullanmak için, bir sınıfın bir belgenin hedefi olup olmadığı bilgisine ihtiyacımız vardır. Bunun için en çok rağbet edilen sınıfa daha fazla şans tanımak iyi bir yöntem olabilir.

$$P(S_i) = \frac{\text{sınıftaki eğitim belgeleri sayısı}}{\text{toplam sınıf sayısı}} \quad \textbf{(4.4)}$$

Son olarak M adet sınıf olduğu varsayımı ile bir sınıf seçme işlemi denklem (4.5) ile yapılabilir.

$$\arg\max_{S_i}[P(S_i \mid B)] = \arg\max_{S_i}[P(B \mid S_i).P(S_i)] \quad \textbf{(4.5)}$$

Naive Bayes sınıflandırıcılar, bilinen bir sınıf için terim olasılıklarının hesaplanma yöntemine göre *çok terimli (multinominal)* ve *çok değişkenli (multivariate)* olmak üzere ikiye ayrılırlar. Çok terimli yöntemde terimlerin geçiş sayıları da dikkate alınırken, çok değişkenli yöntemde terimlerin sadece var olup olmadıklarına bakılır.

4.2.2 En Yakın Komşu Yöntemi

Naive Bayes sınıflandırıcılar, sınıflandırma kurallarını eğitim verisini inceleyerek öğrenirler. En yakın komşu algoritmaları ise, eğitim verisini incelemek yerine tümevarımla öğrenirler.

En yakın komşu tabanlı sınıflandırıcılar, eğitim süresince, eğitim kümesinde yer alan tüm belgeleri belleklerinde tutarlar. Sınıflandırılmak üzere bir B belgesi geldiği zaman, sınıflandırıcı bu belgeye en yakın k adet komşu belgeyi seçer. Daha sonra komşu belgeleri dâhil oldukları sınıflara bakarak bu belgeyi bir veya daha çok sınıfa atarlar.

En yakın komşu yöntemini kullanmak için öncelikle bir uzaklık ölçüm yöntemi tanımlanmış olmalıdır. Öklid uzaklığı, ya da kosinüs benzerliği ölçüleri belge vektörü ile komşuları arasındaki yakınlığı ölçmek için kullanılabilir.

En yakın komşular bulunduktan sonra, sınıflandırma işlemi için çeşitli yöntemler kullanılabilir. Basit olarak

- En yakın komşular arasında baskın olan sınıfa atama yapmak
- Komşuların yer aldığı en iyi temsil edilen n sınıfa atama yapmak

gibi yöntemler kullanılabileceği gibi, daha karmaşık yöntemler de kullanılabilir.

Daha karmaşık bir sınıflandırma yöntemi olarak uzaklıkları ağırlıklandırma yöntemi kullanılabilir. Bu yöntemde bir komşu ne kadar yakınsa, sınıf belirlemede ağırlığı o kadar fazla olacaktır. Bir belgenin bir sınıfa göre puanı denklem (4.6) ile hesaplanır.

$$Puan(S_j, B) = \sum_{B_i \in Tr_k(B)} ben(B, B_i).\alpha_{ij} \qquad \textbf{(4.6)}$$

Denklem (4.6)'da $Puan(S_j, B)$, B belgesi için S_j sınıfının puanı, $Tr_k(B)$, B belgesinin k en yakın komşusu kümesi, $ben(B, B_i)$, benzerlik ölçüsü ve α_{ij} de B_i belgesi S_j sınıfına dâhilse 1, aksi halde 0 değerini alan bir değişkendir.

En yakın komşu yöntemine göre çalışan sınıflandırıcılarda eğitim işlemi çok kısa sürede tamamlanmakta ancak sınıflandırma işlemi görece olarak uzun sürmektedir.

4.3 K En Yakın Komşu Yönteminin Uygulaması

Tez kapsamında yapılan çalışmada sınıflandırıcı olarak En Yakın Komşu yöntemi ile geliştirilen bir sınıflandırıcı kullanılmıştır. Daha önce de belirtildiği gibi en yakın komşu yöntemini uygulayabilmek için metinleri temsil edebilecek bir yönteme ve metinler arasındaki benzerlikleri bulmak için bir uzaklık ölçüsüne ihtiyaç vardır. Tez kapsamında yapılan çalışmada metinleri temsil etmek için *Vektör Uzayı Modeli*'nden yararlanılmıştır. Bu modelde her metin *n* boyutlu uzayda bir vektör olarak temsil edilmektedir. [2] Bu modelde, iki metin arasındaki benzerlik, n boyutlu uzayda yer alan ve metinleri temsil eden iki vektör arasındaki açı ile temsil edilebilir. İki vektör arasındaki açı ne kadar büyükse vektörler arasındaki benzerlik o derece az olacaktır. İki vektörün aynı olması durumunda aradaki açı sıfır olacak ve tam benzerlik olduğu anlaşılacaktır.

4.3.1 Metinlerin Vektörler Olarak Temsil Edilmesi

Metinleri vektörler olarak temsil etmek üzere öncelikle metnin terimleri bulunur. Bir metinde geçen tüm kelimeleri terim olarak kabul etme yöntemi vektörlerin boyutunu çok büyük yapacağından karmaşıklığı artırmasının yanı sıra anlam ifade etmeyen kelimeler de vektör içerisinde yer alıp benzerliklerin bulunmasında etkili olacağından genel başarımı da düşürecektir. Bu sebeple vektörleri oluşturan terimleri bulmak için daha önce belirtildiği gibi durdurma listesi ile gereksiz kelimeler elenir ve kalan terimler ağırlıklandırılarak içerik ile ilgisi olması muhtemel kelimeler terim olarak kabul edilir.

Örneğin, "*Kitap insanın en iyi dostudur. Kitap okumayan insanlar kendilerini cahilliğe teslim ederler*" cümlesinin tüm kelimelerini içeren vektör Tablo 4.1'deki gibi olacaktır.

Tablo 4.1: Örnek Bir Metne Ait Vektör

kitap	insan	en	İyi	dost	oku	kendi	cahillik	teslim	et
2	2	1	1	1	1	1	1	1	1

Bu durumda bu metin için 10 boyutlu bir vektör oluşacaktır. Bu vektör (2,2,1,1,1,1,1,1,1,1) olarak gösterilebilir.

Tez kapsamında gerçeklenen uygulamada vektörlerin belirlenmesi şu şekilde olmaktadır:

Öncelikle tüm metinler bir temizleme, durdurma ve gövdeleme işlemine tabi tutulurlar. Bu işlemler sonucunda metinde yer alan noktalama işaretleri çıkarılır, tüm harfler küçük harfe dönüştürülür ve gövdeleme işlemi uygulanarak her kelimenin gövdesi belirlenir. İşlenmiş olan metinde geçen kelimeler belge sıklıkları tablosuna işlenirler ve işlenmiş metin bir sonraki aşamada incelenmek üzere saklanır.

Tüm metinler ön işlemden geçirildikten ve kelimeler için belge sıklıkları belirlendikten sonra her metnin kelimeleri için terim sıklıkları belirlenir. Terim sıklıkları ve daha önceden hesaplanmış olan belge sıklıkları bilgisi kullanılarak o metne ait terimlerin ağırlıkları hesaplanır. En büyük ağırlığa sahip 10 terim vektör oluşturmak üzere seçilir. Eğer metinde 10 adet terim yoksa var olan terimler vektör oluşturmak üzere kullanılır.Tablo 4.2 ve Tablo 4.3'te örnek veriler için sırası ile kelimelere ait belge sıklıkları bilgisi ve belge başına ağırlık bilgisi görülmektedir.

Tablo 4.2: Kelime-DF Tablosu

Kelime	DF
Araba	13
Molekül	2
Bakan	6
Savaş	3

Tablo 4.3: Kelime-TF-Ağırlık Tablosu

Kelime	TF	Ağırlık
Araba	28	7,4
Molekül	31	33,45
Bakan	12	7.22
Savaş	5	4.5

Belgelerin işlenmesi artımlı bir işlem olduğundan, terim uzayı sürekli olarak genişleyecektir. Yeni işlenen belge için bulunan terimler ile daha önceden karşılaşılmamış ise terim tablosunun sonuna eklenirler. Bu sayede daha önceden

oluşturulmuş olan vektörlerde bir kaydırma veya değişiklik yapılmasına gerek kalmayacaktır.

Oluşturulan vektör hangi metne ait olduğu bilgisi ile birlikte veritabanına kaydedilir. Veritabanında tutulan vektörler tablosunun bir örneği Tablo 4.4'de görülmektedir.

Tablo 4.4: Vektörler Tablosu

mid	Vektor
8	1,5563 0,7782 0,7782 0,7782 0,7782 0,7782 0,7782
9	0 0 0 0 0 0 0 0,7782 0,7782 0,4771
10	0 0 0 0 0 0 0 0 0 0 0,7782 0,7782 0,7782 0,7782
11	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0,7782 0,7782 0,7782 0,4771
12	0 0 0 0 0 0 0 0 0 0,4771 0 0 0 0 0 0 0 0 0,7782 0,7782 0,7782 0,7782

Bu işlemler sonucunda en yakın komşu algoritmalarına özgü olan her yeni girdi için öznitelikleri belleğe alma veya örnek uzayına yerleştirme işlemi yapılmış olur.

4.3.2 Vektörler Arası Uzaklıkların Belirlenmesi

İki vektör arasındaki uzaklığı belirlemek üzere *Kosinüs Benzerliği*'nden yararlanılmıştır. İki vektör arasındaki açının kosinüs değeri denklem 4.7 ile hesaplanabilir.

$$\cos(\vec{v_1}, \vec{v_2}) = \frac{\sum_{i=1}^{n} v_{1i} v_{2i}}{\sqrt{\sum_{i=1}^{n} {v_{1i}}^2} \sqrt{\sum_{i=1}^{n} {v_{2i}}^2}} \quad \textbf{(4.7)}$$

Metinlerden elde edilen vektörler ağırlık bilgisi barındırmaktadır. Bu durumda *x* ve *y* isimli iki belge arasındaki benzerlik denklem 4.8 ile hesaplanabilir.

$$Ben(x, y) = \frac{\sum_{i=1}^{n} w_{ix} w_{iy}}{\sqrt{\sum_{i=1}^{n} {w_{ix}}^2} \sqrt{\sum_{i=1}^{n} {w_{iy}}^2}} \quad \textbf{(4.8)}$$

Örneğin *kitap, kalem, defter* kelimelerinden oluşmuş üç belgelik bir sistemde aşağıdaki içerikleri verilen belgelerin olduğu varsayalım:

B_1: *"kitap kitap kitap"*

B_2: *"defter defter defter"*

B_3: *"kitap kalem kitap"*

Bu durumda bu belgelere ilişkin vektörler şu şekilde olacaktır.

d_1: *(3,0,0)*

d_2: *(0,0,3)*

d_3: *(2,1,0)*

Normalizasyon ve ağırlıklandırmayı basitleştirme amacı ile yok sayarsak belgeler arası benzerlikler şu şekilde olacaktır.

$Ben(B_1,B_2)=0$

$Ben(B_1,B_3)=6$

Örnekte de görüldüğü gibi ilk iki belge arasında herhangi bir benzerlik yoktur. Birinci ve üçüncü belgeler ise *kitap* kelimesini ortak olarak kullandıkları için aralarında benzerlik çıkmaktadır.

Gerçeklenmiş olan uygulamada, vektörler arasındaki benzerliği bulmak üzere denklem 4.8'den yararlanılmıştır. İncelenecek olan metne ait vektör belirlendikten sonra bir önceki aşamada saklanmış olan vektörler ile birer birer karşılaştırılarak benzerlikler belirlenir. Elde edilen benzerlik değerleri bir benzerlikler tablosuna, karşılaştırılan metin ve benzerlik miktarı olarak kaydedilir. Tablo 4.5'te örnek benzerlikler için benzerlik tablosu görülmektedir.

Tablo 4.5: Metinler Arası Benzerlik Değerleri

Kelime	DF
araba	13
molekül	2
bakan	6
savaş	3

4.3.3 En Yakın k komşuya İlişkin Kategorilerin Belirlenmesi

Sınıflandırılacak olan metin ön işlemlerden geçirildikten sonra, terim vektörü bulunur ve bu vektör var olan vektörler ile kıyaslanarak en yakın komşuları elde edilir. Bu aşamada en yakın komşulara ait olan sınıflar eğitim kümesinden belirlenerek yeni metin için sınıflandırma işlemi yapılır. Sınıflandırma işlemi için çeşitli yöntemler benimsenebilir.

İlk akla gelen yöntemlerden biri baskın sınıfın belirlenmesi yöntemidir. [2] Bu yöntemde en yakın komşular arasında en fazla geçen kategori incelenen metnin kategorisi olarak isimlendirilir.

Bir başka yöntem olarak, elde edilen kategori listesindeki en üst *n* adet kategorinin belirlenmesi işlemi yürütülebilir. [2] Bu işlem aynı zamanda çoklu sınıflandırma işlemini yerine getirecektir.

İlk iki yöntemden daha karışık bir sınıf belirleme yöntemi olarak metinler arasındaki uzaklığa göre sınıfların önemini belirleyen bir ölçü kullanılabilir. Bu durumda kategorileri belirlenecek olan metne en yakın metnin kategorileri en belirleyici olacaktır. İncelenen belge için bir sınıfın puanı denklem 4.9 ile belirlenebilir[2].

$$Puan(S_j, B) = \sum_{B_i \in Tr_k(B)} Ben(B, B_i)\alpha_{ij} \qquad \textbf{(4.9)}$$

Denklem 4.9'da S_j puanı belirlenecek olan sınıfı, B belgeyi, $Tr_k(B)$, D belgesinin k en yakın komşusu kümesini, $Ben(B,B_i)$ yakınlık ölçüsünü belirtir. α_{ij} değeri B_i belgesi S_j sınıfına dahilse 1, dahil değilse 0 değerini alan bir değişkendir.

5. YAZILIMIN AÇIKLANMASI

Tez kapsamında geliştirilen gövdeleyici ve sınıflandırıcı için yazılımlar geliştirilmiştir. Bu bölümde geliştirilmiş olan yazılımlar tanıtılacaktır.

5.1 Gövdeleyici Yazılımı

Gövdeleme başarımı etkileyen önemli aşamalardan biridir. Gövdeleme işlemini gerçeklemek üzere, biçimbirimsel çözümleme (BÇ) çıktısı üzerinden gövdeyi belirleme ve verilen kelimenin eklerini sonlu durum makinesi kullanarak inceleme yöntemlerine göre çalışan iki ayrı gövdeleyici tasarlanmıştır.

5.1.1 Biçimbirimsel Çözümleyici Tabanlı Gövdeleyici

Bu gövdeleyici girdi olarak Biçimbirimsel Çözümleyici [10] çıktısı kullanmaktadır. Biçimbirimsel çözümleyiciler bir liste halinde aldıkları kelimeler için tüm olası çözümlemeleri üreterek sonuçları oluştururlar. Örnek bir çıktı belgesi Şekil 5.1'de görülmektedir.

```
Dövizli     döviz +Noun+A3sg+Pnon+Nom^DB+Adj+With

askerlik    askerlik    +Noun+A3sg+Pnon+Nom
askerlik    asker +Noun+A3sg+Pnon+Nom^DB+Adj+FitFor
askerlik    asker +Noun+A3sg+Pnon+Nom^DB+Noun+Ness+A3sg+Pnon+Nom

20    20    +?

bin   bin   +Verb+Pos+Imp+A2sg
bin   bin   +Num+Card

mark  mark  +Noun+A3sg+Pnon+Nom
```

Şekil 5.1: Gövdelemede Kullanılan Biçimbirimsel Çözümleme Çıktısı

Gövdeleyici program bu girdi üzerinde çalışır. Bu çıktının sabit alanları vardır ve her kelimenin çözümlemesi bittikten sonra bir boş satır bırakıldıktan sonra diğer çözümlemeye ilişkin sonuçlar üretilir.

Çözümleme olan satırları bulmak ve kelime gövdelerini almak üzere *Düzenli İfadelerden* yararlanılmıştır. Bahsedilen işlemi yapan düzenli ifade Tablo 5.1'de görülebilir.

Tablo 5.1: BÇ Çıktısını Parçalayan Düzenli İfade

'^(.*?)\t(.*?)\t\+(.*)'

Bu düzenli ifadede ^ işareti mutlaka bir satır başı olması gerektiği şartını getirir. (.*?) öbeği herhangi bir harf öbeğini temsil etmektedir. Bu durumda yukarıdaki düzenli ifade şu şekilde okunabilir: Satır başını takip eden bir harf öbeği, bir tab karakteri, bir harf öbeği, bir tab karakteri, + karakteri ve bir harf öbeği.

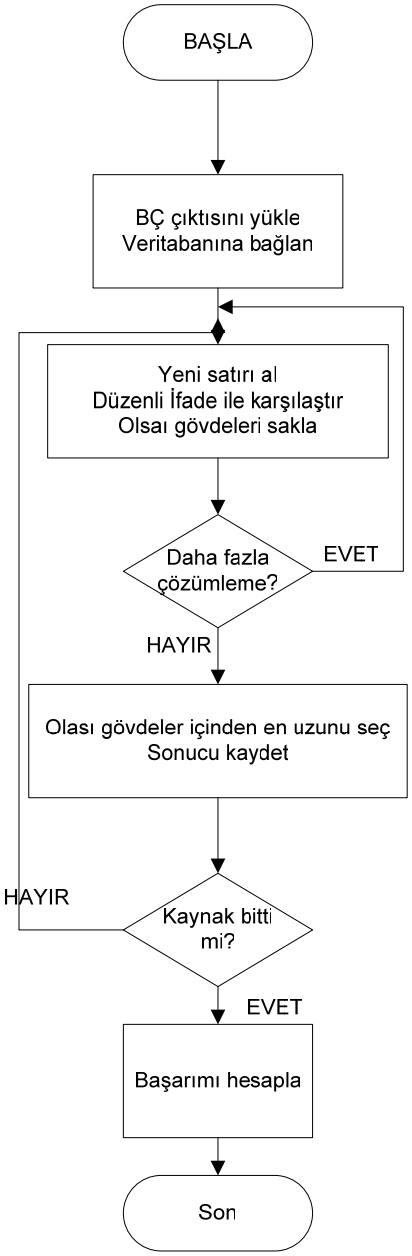
Bu düzenli ifade BÇ çıktısının her satırı için çalıştırılır ve eşleşip eşleşmediğine bakılır. Eşleşme olması durumunda () arasında yer alan harf öbekleri içerdikleri anlama göre değerlendirilir.

Örneğin Şekil 5.1'deki çıktının ilk satırı için birinci öbek `Dövizli` ikinci öbek `döviz` ve üçüncü öbek `Noun+A3sg+Pnon+Nom^DB+Adj+With` olarak belirlenecektir.

Düzenli ifadeler ile belirlenen öbeklerden gövdeye ilişkin olanları her kelime için belirlenerek geçici olarak tutulur ve içlerinde en uzun harf sayısına sahip olan katar gövde olarak seçilir.

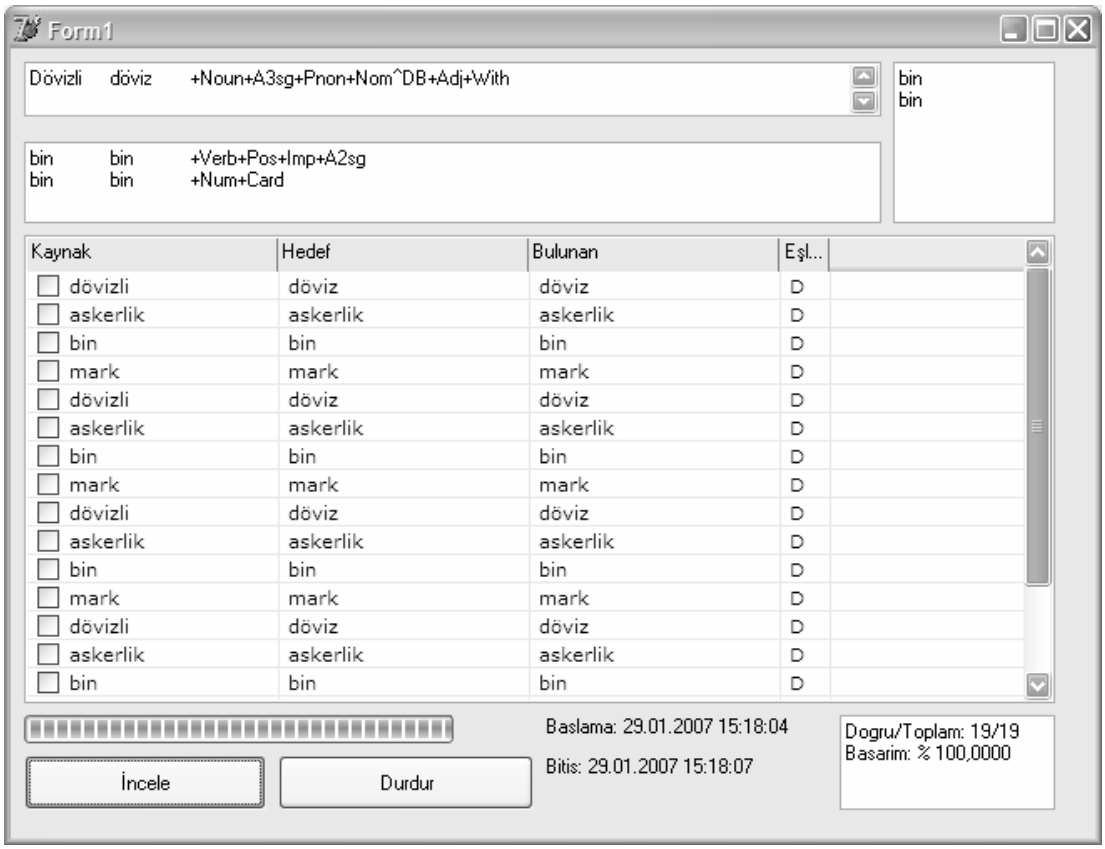
Bulunan gövdeler ile veritabanında saklı olan doğru gövdeler karşılaştırılarak doğruluk oranı belirlenir.

Bu programa ilişkin akış diyagramı Şekil 5.2'de görülmektedir.



Şekil 5.2: BÇ Tabanlı Gövdeleyici Akış Diyagramı

Biçimbirimsel çözümleme tabanlı gövdeleyiciye ait ekran görüntüsü Şekil 5.3'de görülebilir.

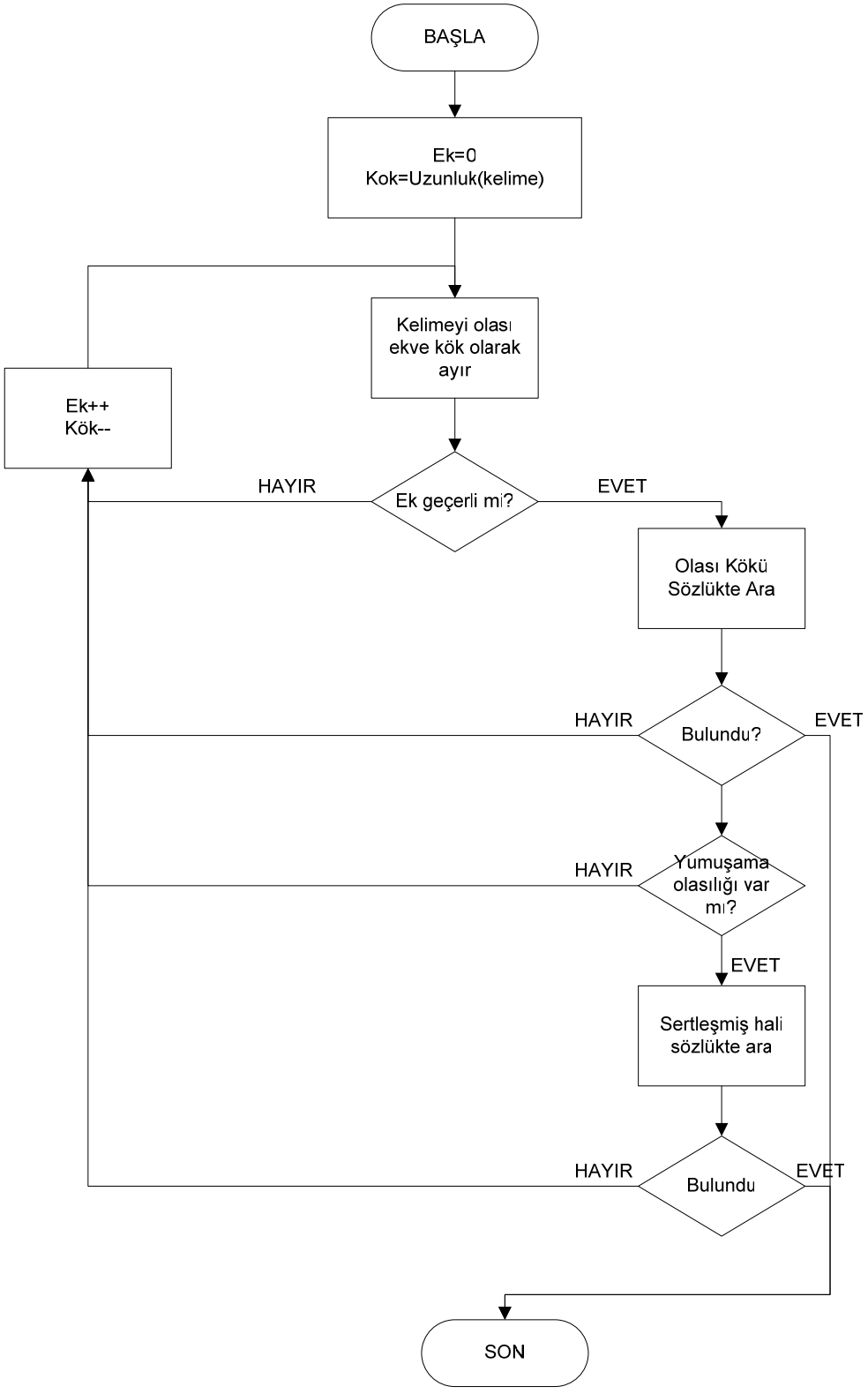


Şekil 5.3: BÇ Tabanlı Gövdeleyici Ekran Görüntüsü

5.1.2 Sonlu Durum Makinesi Tabanlı Gövdeleyici

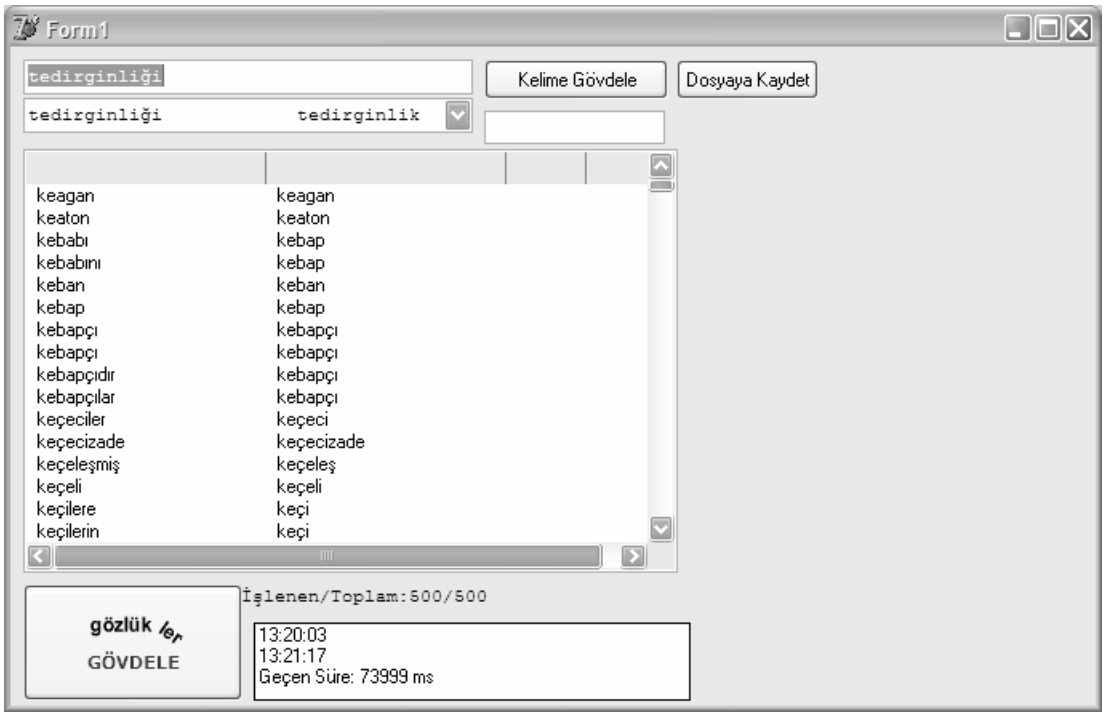
Tez kapsamında yapılan çalışmada sonlu durum makinesi ile ekler üzerinde doğrulama yapan bir program tasarlanmıştır. Bu programın düzgün çalıştığı görüldükten sonra, bu kısım gövdeleme modülü olarak ana programa eklenmiştir.

Bu gövdeleme işlemine ilişkin akış diyagramı Şekil 5.4'te görülmektedir.



Şekil 5.4: Sonlu Durum Gövdelemesi Akış Diyagramı

Sonlu durum makinesi tabanlı gövdeleyiciye ilişkin ekran görüntüsü Şekil 5.5'te görülebilir.



Şekil 5.5: Sonlu Durum Makinesi Tabanlı Gövdeleyici Ekran Görüntüsü

5.2 Sınıflandırıcı Yazılımı

Metin sınıflandırma işlemini yapmak üzere bir yazılım hazırlanmıştır. Bu bölümde bu yazılım tanıtılacaktır.

5.2.1 Yazılımın Genel Yapısı ve Tanımlar

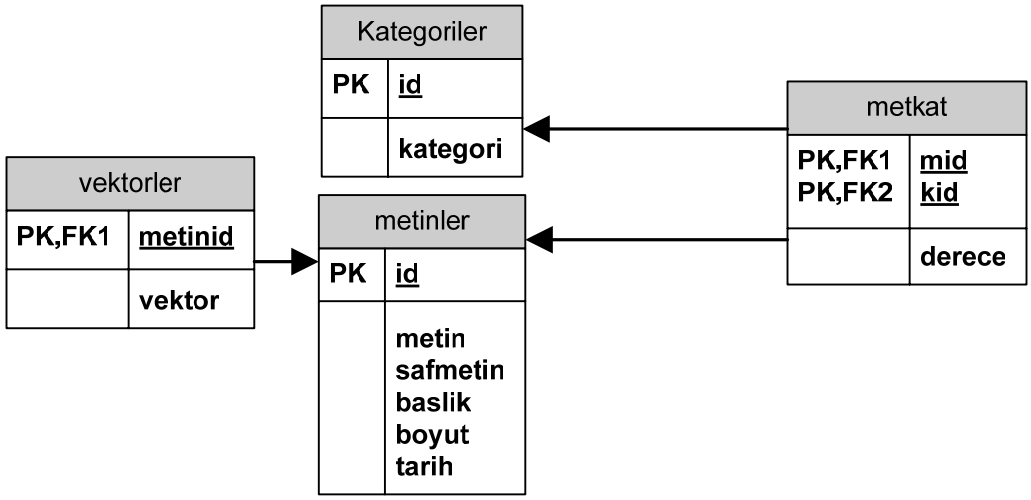
Hazırlanmış olan yazılım Borland Delphi 7 yazılım geliştirme ortamı kullanarak hazırlanmıştır. Veritabanı işlemleri için MySQL veritabanı tercih edilmiştir.

Programın ihtiyaç duyduğu bilgilerin bir kısmı metin dosyalarında, bir kısmı da veritabanında saklanmaktadır. Dosyalarda tutulan bilgiler, programın çalışmaya başlaması ise liste yapılarına alınmakta ve program süresince işlemler bu listeler üzerinde gerçeklenmektedir. Programda kullanılan listeler ve kullanım amaçları Tablo 5.2'de gösterilmiştir.

Tablo 5.2: Listeler ve Kullanım Amaçları

Liste Adı	**Liste Tipi**	**Kullanım Amacı**
durdurma	TStringList	Durdurma listesini temsil eder
yumusakkok	TStringList	Yumuşama kurallarına uymayan kökleri temsil eder

Veritabanı temelde gövdelemede kullanılan sözlüğü ve metinler ile kategoriler arasındaki ilişkileri tutmak için kullanılmıştır. Programda metinler ile ilgili işlemlerde kullanılan veritabanı yapısına ilişkin varlık ilişki diyagramı Şekil 5.6'daki gibidir.



Şekil 5.6: Veritabanı Varlık İlişki Diyagramı

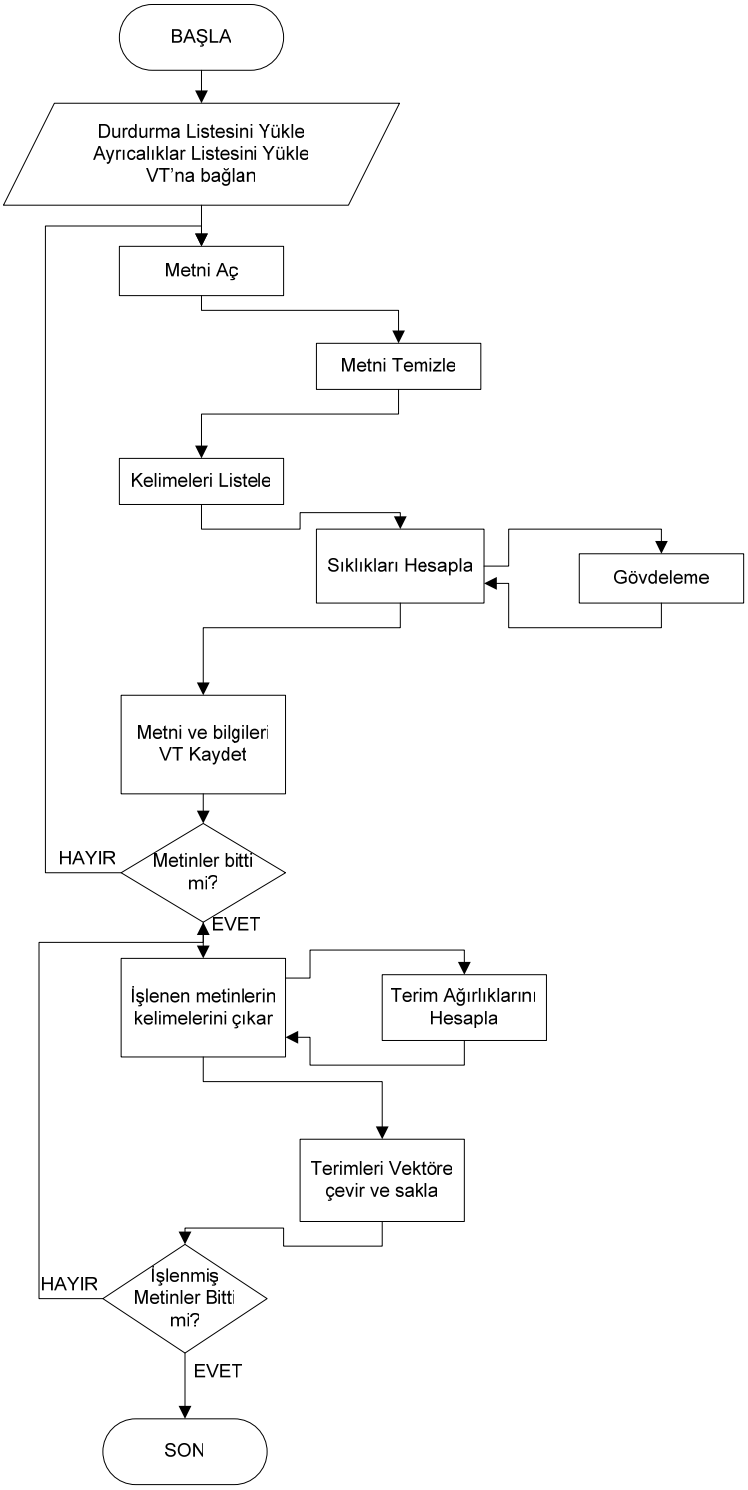
Veritabanı ilişkisel olarak tanımlanmıştır. Tablolar ve dış anahtar olarak kullandıkları alanlar Tablo 5.3'de görülebilir.

Tablo 5.3: Tablolar Arası Dış Anahtarlar

Tablo Adı	Alan Adı	Dış Anahtar
vektorler	metinid	Metinler.id
metkat	mid	Metinler.id
metkat	kid	Kategoriler.id

5.2.2 Yazılımın Akış Diyagramı

Yazılım gerek metinleri inceleme ve vektörleri oluşturma, gerekse seçilen bir metin için kategorileri belirleme konusunda çeşitli modülleri birbiri ardına çağırmaktadır. Bu modüller ve programın genel akışı Şekil 5.7'de görülebilir.



Şekil 5.7: Yazılım Akış Diyagramı

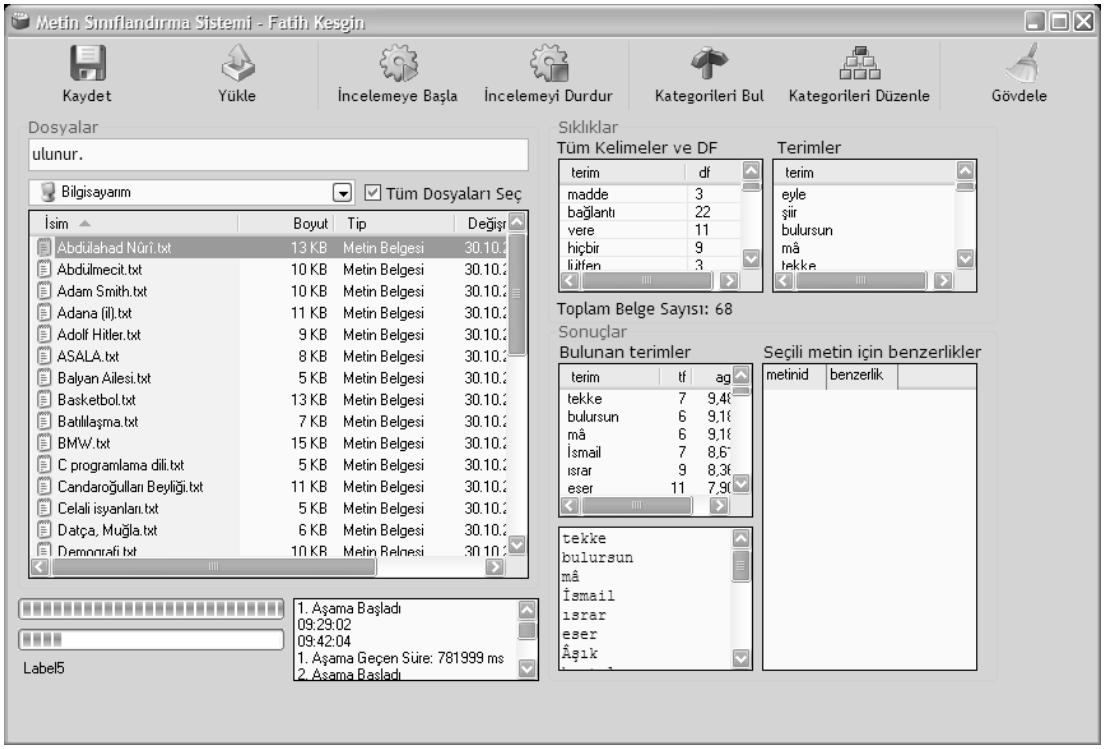
5.2.3 Yazılımın Kullanımı

Hazırlanmış olan yazılımı kullanmak için öncellikle bir eğitim kümesini oluşturan belgelerin incelenmesi ve özellikleri ile ait oldukları sınıfların belirtilmesi gereklidir. Bu işlem yazılımda istenilen belgelerin bulunduğu dizine gelindikten sonra *İncelemeye Başla* düğmesine basılarak yapılabilir.

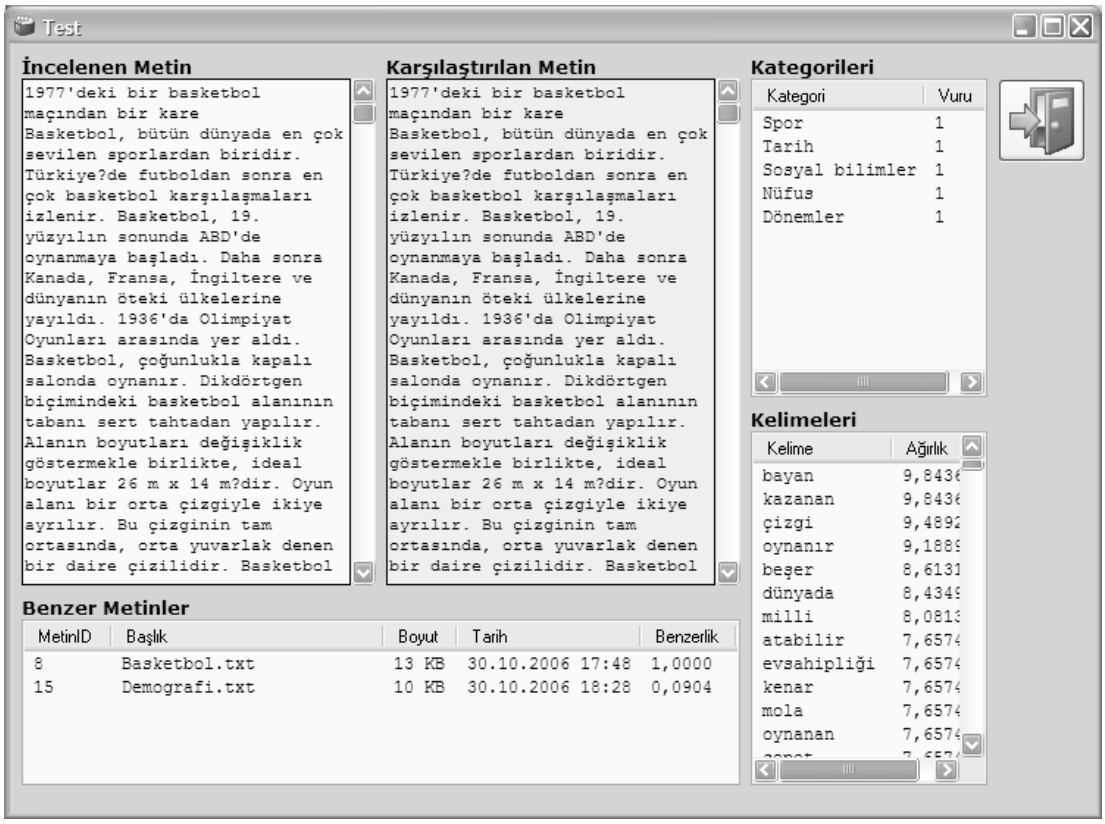
Metinler incelendikten sonra, bu metinlere ait sınıfların belirlenmesi için *Kategorileri Belirle* düğmesine basılarak kategori belirleme ekranı açılır. Bu ekran üzerinden öncelikle kategorileri belirlenecek olan metin seçilir daha sonra da ilgili kategoriler eklenerek *Kaydet* düğmesi ile bu bilgiler saklanır.

İncelenmemiş bir metni sınıflandırmak için öncelikle metin seçilir ve daha sonra *Kategorileri Bul* düğmesine basılır.Metin incelendikten sonra elde edilen sonuçlar görüntülenir.

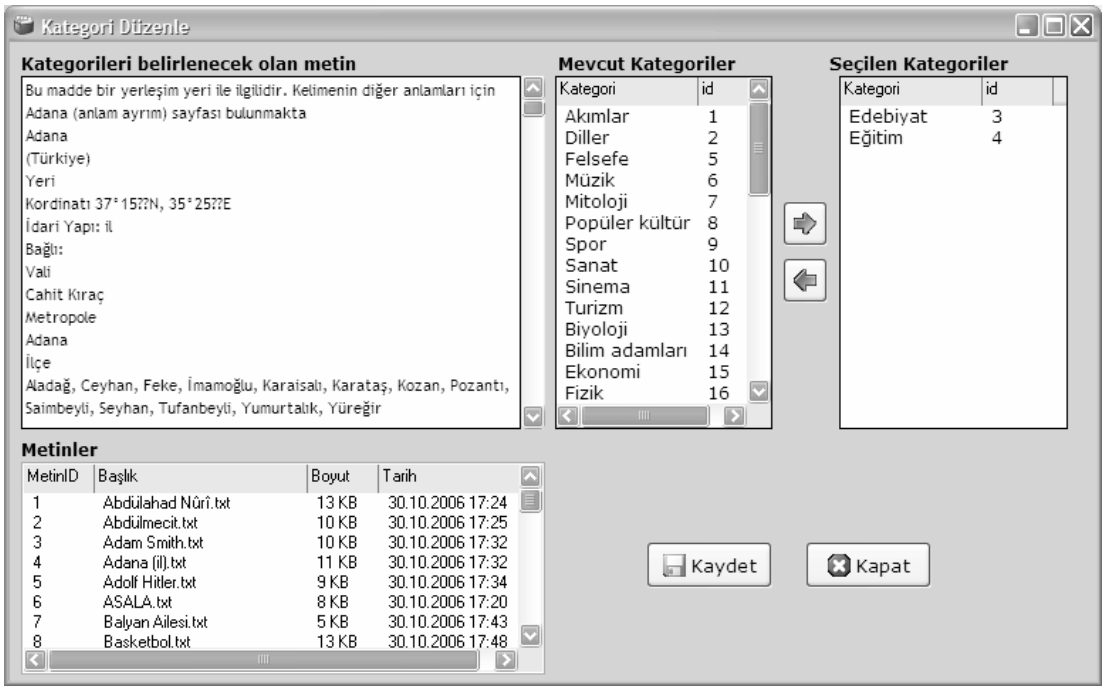
Yazılıma ait ekran görüntüleri Şekil 5.8, Şekil 5.9 ve Şekil 5.10'da görülebilir.



Şekil 5.8: Yazılım Ana Ekran Görüntüsü



Şekil 5.9: Sınıflandırma Sonucu Ekran Görüntüsü



Şekil 5.10: Kategori Tanımlama Ekran Görüntüsü

6. SONUÇLAR VE ÖNERİLER

Bilgi Erişimi günümüzde hızla artmakta olan sayısal belgeler düşünüldüğünde gittikçe önem kazanan bir araştırma alanı olacaktır. Bu tez kapsamında bilgi erişimi kapsamına giren metin sınıflandırması konusu incelenmiş ve Türkçe metinler için konu belirleme sistemi tasarlanmıştır.

Konu belirlemek için öncelikle metnin ön işlemlerden geçirilmesi gerekmektedir. Anlam olarak hemen hemen aynı ifadeye sahip sözcükler, çekim eki aldıklarında yazılışları tamamen değişmektedir. Bu durumda bu sözcüklerin gövdelerinin bulunması gerekmektedir. Türkçenin sondan eklemeli bir dil olduğu göz önünde bulundurulduğunda gövdeleme işleminin başarımı büyük ölçüde etkileyeceği söylenebilir.

Tez kapsamında gövdeleme için iki farklı yöntem denenmiş ve uygulamalar geliştirilmiştir. İlk yöntemde biçimbirimsel çözümleyici kullanılmış ve biçimbirimsel çözümleyicinin ürettiği olası gövdelerden en uygunu seçilmiştir. Biçimbirimsel çözümleyiciye bağlı kalmamak ve daha hızlı bir gövdeleyici yapmak amacı ile yeni bir gövdeleyici gerçeklenmiştir. Bu gövdeleyicide öncelikle olası bir ek katarı geçerli olup olmadığı yönünde incelenmekte, geçerli bir ek olması halinde gövde için sözlüğe bakılmaktadır. Bu sayede gövdeleme işlemi daha hızlı olarak gerçekleşmektedir.

Bu tezde yapılan çalışma, gövdeleyicinin başarımını artırmak yönünde geliştirilebilir. Sınıflandırma işlemi için de daha büyük veri kümeleri bulunmalı ve bu kümeler üzerinde incelemeler yapılmalıdır.

KAYNAKLAR

[1] **Baeza-Yates, R. and Ribeiro-Neto**, **B**., 1999. Modern Information Retrieval, Addison-Wesley, England

[2] **Jackson, P. and Moulinier, I.,** 2002**.** Natural language processing for online applications: text retrieval, extraction, and categorization, Amsterdam

[3] **Güzel, A.,** 2005. Üniversiteler İçin Türk Dili Ders Kitabı, Başkent Üniversitesi, Ankara

[4] **Porter, M.F.,** 1980. An algorithm for Suffix Stripping, *Program*, 14(3):130-137

[5] **Jurafsky, D. and Martin, J.,** 2000. Speech and Language Processing, Prentice Hall, New Jersey

[6] **Köksal A.,** 1981. Tümüyle Özdevimli Deneysel Bir Belge Dizinleme ve Erişim Dizgesi, *TBD 3. Ulusal Bilişim Kurultayı,* Ankara, 6-8 Nisan, 37-44

[7] **Alpkoçak, A., Kut, A., Özkarahan, E.**, 1995. Bilgi Bulma Sistemleri için Otomatik Türkçe Dizinleme Yöntemi, *Bilişim Bildirileri*, Dokuz Eylül Üniversitesi, İzmir, 247-253.

[8] **Solak, A., Can, F**., 1994. Effects of stemming on Turkish text retrieval, *Proceedings of the Ninth Int. Symp. on Computer and Information Sciences*. Antalya, Turkey, November 1994, 49-56.

[9] **Duran, G.,** 1997. Gövdebul: Turkish Stemming Algorithm**,** *Yüksek Lisans Tezi***,** Hacettepe Üniversitesi,Bilgisayar Mühendisliği Bölümü, Ankara, Türkiye.

[10] **Oflazer, K.** 1994. Two-level Description of Turkish Morphology, *Literary and Linguistic Computing*, Vol. 9, No:2

[11] **Altintas, K., Can, F**. 2002. Stemming for Turkish : a comparative evaluation. *Proceedings of the 11th Turkish Symposium on Artificial Intelligence and Neural Networks (TAINN)*, Istanbul / Turkey, June 2002), 181-188

[12] **Joachims, T.,** 2002. Learning to classify text using support vector machines, Kluwer Academic Publishers, Boston.

ÖZGEÇMİŞ

1982 yılında İstanbul'da doğan Fatih KESGİN, 2000 yılında İstanbul Atatürk Fen Lisesi'nden ve 2004 yılında İstanbul Teknik Üniversitesi Elektrik-Elektronik Fakültesi Bilgisayar Mühendisliği Bölümü'nden mezun olmuştur. Fatih Kesgin, 2005 yılından beri İ.T.Ü Bilgisayar Mühendisliği bölümünde araştırma görevlisi olarak çalışmaktadır.