Erciyes University Journal of the Institute of Science and Technology
Erciyes Üniversitesi Fen Bilimleri Enstitüsü Dergisi
ISSN 1012-2354
Cilt (Volume): 28, Sayı (Issue): 4, Temmuz/July-2012
http://fbe.erciyes.edu.tr/

Zeki doküman dili sınıflandırma ve web tabanlı çeviri sistemi

Uraz YAVANOĞLU ve Şeref SAĞIROĞLU
Gazi Üniversitesi, Mühendislik Fakültesi, Bilgisayar Mühendisliği Bölümü, Maltepe, Ankara

Anahtar Kelimeler: Dil tanıma, dil dönüştürme, web tabanlı uygulama, yapay sinir ağı

ÖZET

İletişim teknolojilerinin gelişmesine paralel olarak küresel boyutta bilginin erişimi, paylaşılması, farklı dillere çevrilmesi ve bireylerin ihtiyaçlarına uygun kullanılmasının sağlanması maksadıyla içerik dilinin bilinmesi veya internet ortamında yayınlanan kaynakların dillerinin bulunması ve bir dile çevirisinin yapılması büyük önem arz etmektedir. Bu çalışmada web tabanlı olarak geliştirilen bir yazılım platformu sayesinde, içerik dili bilinmeyen Word, PDF ve HTML dokümanlarının dil içerikleri 15 farklı dil için zeki bir sistem tarafından sınıflandırılarak otomatik olarak algılanmakta ve dil çevirisi mevcut çözümler kullanılarak 64 dile otomatik olarak yapılmaktadır. Tanımlama işlemi için, yapay sinir ağları temelli yeni bir dil tanıma yöntemleri kullanılarak bu işlemler gerçekleştirilmektedir. Geliştirilen yöntem iki farklı yaklaşım ile karşılaştırılmış, dokümanların büyüklüğüne veya içeriğin niteliğine bağlı olmadan 15 farklı dilde bulunan 3 farklı doküman tipi için yüksek bir başarım göstermiştir.

Intelligent document language classification and web based translation system

Key Words: Language identification, language translation, web based application, artificial neural network

ABSTRACT

Recent developments on information and communications technologies help globally and important to access, share, translate and the documents use easily and effectively via internet media. Language identification is an important task for web information retrieval services. Automatic language identification and translation have become increasingly important, as more and more documents are being served on internet within many languages. This study presents new methods to identify web contents, containing MS Word, PDF and HTML documents in different languages and to translate them into specified languages. The identification problem can be seen as a specific instance of the more general problem of an item classification through its attributes in a limited workspace. This novel approach is based on artificial neural network model to recognize the languages. Documents content belonging to 15 languages were used in test with a new testing methodology and translating them into 64 languages automatically for language processing. The results have shown that the approaches presented in this work are very successful to meet the expectations in real-time language identification and translation accuracy and reduce the number of letters in solution space in comparison with the available two methods.

*Sorumlu Yazar (Corresponding author) e-posta: uraz@gazi.edu.tr

1. Giriş

İnternet teknolojileri ve uygulamaları kullanırken, bir sayısal belgenin dilini tanımak pek çok açıdan yararlıdır. *Dil tanıyıcı* (DT), özellikle kelime işleme yazılımında, tüm dokümanın ya da farklı dillerden oluşan bölümlerin dillerini otomatik tanımada kullanılabilir. İnternette, dönüştürme işlemi için Microsoft, Altavista, Systran, Babylon ve Google gibi firmaların dil dönüştürme çözümleri sunulmaktadır. Bu çözümlerden hızlı bir büyüme süreci içinde olan Google firması tarafından geliştirilen Google Translator sistemi en kapsamlı çeviri çözümüdür. Bu çeviri sistemi sayesinde 64 farklı dili farklı dillere dönüştürebilme desteği vermektedir [19]. Google tarafından geliştirilen yöntem bugüne kadar kullanılan metin bilgisinden dilbilgisi öğeleri elde ederek çevirme yaklaşımını kullanmamaktadır. Google tarafından geliştirilen yöntem sayesinde bir veritabanı diller arasında önceden yapılmış ve halen yapılmakta olan birebir çevirileri cümle yapısıyla saklamaktadır. Google Translator tarafından alınan bir giriş o girişe karşılık gelen veya en çok yakınsayan kelimelere doğrudan dönüştürülmektedir.

Bu çalışmada, Almanca, Arnavutça, Fransızca, Galce, Hırvatça, İngilizce, İrlandaca, İspanyolca, İtalyanca, Letonca, Macarca, Maltaca, Portekizce, Türkçe, Vietnamca dilleri dikkate alınarak önerilen birleşim ve kesişim yöntemleriyle WORD, PDF ve HTML dokümanları dikkate alınarak dokümanların içerik dilinin otomatik olarak algılanması ve istenilen dile dönüştürülmesi için bir sistem geliştirilmiştir. Bu çalışmanın internetin daha verimli kullanılmasına, farklı dil ve kültürlerde yapılan çalışmaların kolaylıkla okunup öğrenilmesine katkılar sağlayacağı, internette karşılaşılan pek çok problemin çözümüne katkılar sağlayacağı değerlendirilmektedir. Geliştirilen zeki dil tanıma sistemi aşamaları takip eden bölümlerde açıklanmıştır. Bölüm 2'de literatürde mevcut olan dil tanıma sistemleri detaylı olarak sunulmuştur. Bölüm 3'de bu çalışmada kullanılan zeki sistem altyapısını oluşturan yapay sinir ağlarına ait genel bilgiler verilmiş ve kullanılan öğrenme algoritması kısaca açıklanmıştır. Bölüm 4'de önerilen yöntemler ve geliştirilen sistem altyapısı anlatılmıştır. Bölüm 5'de ise bu çalışmada elde edilen sonuçlar sunulmuş ve gelecek çalışmalara yön verecek önerilere yer verilmiştir.

2. Dil tanıma sistemleri

Literatürde pek çok dil tanıma metotları bulunmaktadır [1–19, 24-27]. Bu metotlar incelendiğinde, n-gram veya Bayes teoreminin kullanılması gibi istatistiksel metotların sıklıkla kullanıldığı fakat bu metotların ön işlem gibi performans kaybına neden olan süreçler içermesinden dolayı beklenen performansı veremedikleri görüldüğünden yeni metotların araştırılması ve geliştirilmesi gündemde olan bir husustur. Geliştirilen metotların kullanıldığı ve karşılaştırıldığı çalışmalar aşağıda kısaca özetlenmiştir. Padro ve Padro tarafından 2004 yılında yapılan çalışmada [1], istatistiksel yöntemlere dayalı üç farklı dil tanıma yöntemi incelenmiştir. Basit parametrelerden bazıları sınıflandırma için gerekli olan kelime miktarları ve tanınması gerekli olan dil sayısı için uygun eğitim kümesi oluşturma önem taşımamaktadır. Yapılan çalışmalar eğitim küme büyüklüğünün 50 kelimeyi geçmesinin yeterli olduğunu göstermiştir. Önemli sayılabilecek parametrelerden birisinin metin uzunluğu olduğu yine bu çalışmanın bulguları arasındadır. İncelenen yöntemlerde ancak 500 karakter ve daha uzun metinlerin tanınması %95 başarıyla gerçekleşmektedir. Yapılan testlerde metin uzunluğu 5000 ve üstüne çıkartıldığında sonuçların %99 başarılı olduğu gösterilmiştir. Başarı oranı metin boyutu kısaldıkça doğru orantılı olarak azalmaktadır. Kısa metinlerde kullanılan yöntemlerin hepsinde başarı oran bazen %60 seviyesine kadar gerileme göstermiştir [1].

Botha ve arkadaşlarının [2] 2007 yılında yaptığı çalışmada metin tabanlı 11 farklı Güney Afrika dili ele alınarak dil tanıma süreci araştırılmış ve n-gram istatistikleri sınıflandırma yapmak için kullanılmıştır [2]. Bu süreç yönetiminde özellikle destek vektör makineleri (support vector machines) ve yakın komşuluk tabanlı sınıflandırıcılar (likelihood-based classifiers) karşılaştırılmıştır. Tanıma işlemi için giriş olarak uygulanan metinlerden az sayıda kelime ile tanıma işleminin mümkün olduğu belirlenmiştir. Destek vektör makineleri genel olarak daha başarılı sonuçlar vermesine rağmen sınıflandırma için gerekli olan yüksek işlem karmaşıklığı sebebiyle ancak basit ve kararlı bir sistem tasarımı için uygun olmadığı gösterilmiştir. Yapılan n-gram testlerinde en iyi çalışma performansını veren n-gram analizi için gerekli olan n değeri 6 olarak bulunmuştur. Hollandaca, Danimarka'ca, İngilizce, Fransızca, İtalyanca ve İspanyolca dillerinin incelendiği bir çalışmada çeşitli (2-7 arası) büyüklüklerde n-gram analizlerinin ortalama hata oranları incelendiğinde 100 karakterli bir giriş seti için hata oranının %1,03 olduğu gösterilmiştir. Yapılan bu çalışmada ise 15 karakterli bir giriş seti için hata oranı %0,6 olarak bulunmuştur. Botha ve arkadaşları tarafından yapılan diğer bir çalışmada ise Güney Afrika dillerinin Avrupa dillerine göre tanınma oranının daha düşük olduğu savunulmuş olmasına rağmen hata oranları olmadığı için bu çalışma ile birebir karşılaştırma yapılamamıştır [26]. [2] nolu çalışmada yapılan testlerde daha büyük giriş setleri ile doğrusal olarak daha kesin sonuçlar elde edilmekte ve destek vektör makinelerinin karmaşıklık probleminin sistemi karasızlığa götürdüğü rapor edilmiştir. Aynı çalışmada bahsedilen diğer metotların her dil grubu için oluşturulması gereken eğitim seti büyüklüğü belirtilmiş ve elbette aynı etnik kökeni paylaşan iki farklı dil içinde geçebilecek aynı kelimeler olabileceği için yakın komşuluk tabanlı sınıflandırıcıların büyük miktarda giriş verisi ile elde edilen sonuçlar sunulmuştur [2].
El-Shishiny ve arkadaşları [3] tarafından 2004 yılında yapılan diğer bir çalışmada Arapça kökenli diller ve diğer diller arasında Arapça betiklerin kelime parçalama tabanlı bir metot ile nasıl ayrıştırabileceği araştırılmıştır. Bu yöntemin geliştirilmesi sırasında Arapça betiklere ait karakter kümesi olarak Arapça dilinde kullanılan kelimeler, ön ve son ekler ile unigramlar kullanılmıştır. Bu çalışmada internet kaynaklarından rastgele seçilen Arapça, Farsça, Urduca, Afganca ve Uygurca gibi Arapça betik tabanlı dillerden oluşan toplam 180 farklı örnek kullanılmıştır. Bu dillerden Arapça betiklerin tanınması ile %94 oranında dil tanıma başarısı elde edilmiştir. Bu her dil için mevcut dilbilgisi kurallarının çözümlenmesi ve sistemin bu kuralları kullanabilecek şekilde yapılandırılması gibi zorluklar taşımaktadır [3].

Kruengkrai ve arkadaşları [4] tarafından 2005 yılında yapılan başka bir çalışmada otomatik dil tanıma için karakter dizisi

çekirdeği (string kernels) kavramı tabanlı yeni bir yöntem önerilmiştir. Önerilen yöntemde kodlama şekline rağmen, tanınması istenilen dil metin içinden doğrudan çıkartabilmektedir. Bu işlem yapılırken dikkat edilmesi gereken husus metin bilgisinin kodlama bilgisinin bir karakter dizisinin byteları şeklinde kodlanmasıdır, çünkü mevcut kodlama örüntüleri incelendiğinde çoğu sayfa için kullanılan ISO-8859-1 kodlama standardı 19 farklı Avrupa dilini desteklemektedir. Bu genellemenin geçerliliğini kaybetmesiyle birlikte dil tanımanın bir sınıflandırma problemi olduğu görülmektedir. Sistemin kararlı bir şekilde çalışması için yapılması gereken ilk işlemin eğitim setinde her dile ait karakteristik yapının çıkartılmasının gerekliliği bu çalışmada vurgulanmıştır. Sonuçlar yeterli sayıda giriş setiyle 17 farklı dil için değerlendirilmiş ve elde edilen test sonuçlarında diller arası yakınsamanın azaltılması için eşleşen sözcük öbeklerine göre ağırlık atama yönteminin kullanılmasının gerektiği sonucuna varılmıştır. Yapılan testler dil tanıma sorunu için kararlı bir metot oluştuğunu göstermesine rağmen performans ve hız konularında sistemin uzaysal karmaşıklığı nedeniyle gerekli sonuçlara bu çalışmada yer verilmemiştir [4].

Zavarsky ve arkadaşları [5] tarafından 2005 yılında yapılan çalışmada kodlama örüntüleri karşılaştırılarak çok yüksek sayıda giriş setlerine ait dil tanıma işlemlerinin sonuçları sunulmuştur. Bu çalışmada kullanılan büyük giriş kümeleri (milyonlarca) Dil İzleme Projesinin bir ürünü olarak ortaya çıkmaktadır. Bu proje düzinelerce milyondan fazla metin tabanlı dokümandan oluşmaktadır. Önerilen yöntem tek bir Linux bilgisayarda her saniye 250 günde 20 milyon dokümanı inceleyebilmektedir. Bu sayede çok işlemcili Linux sunucularda analiz doküman sayısı günlük 100 milyona ulaşabilmektedir. Doğal dillerde yazılan metinlerin tanıma probleminde metinlerin karakter kodlama örüntülerinin belirlenmesi zor bir sorun değildir. Şayet bir doküman birden fazla dil ile yazılmamış ise yeteri kadar uzun olması kabul edilebilir bir sürede belirli sayıda dokümanın analiz edilmesini sağlamaktadır. Bu amaçla sıklıkla kullanılan yöntemler kelime tabanlı metotlardır. Elbette bu metotlar temelde karakter kodlaması bakımından etkisizdir ve daha önceden sınıflandırma bilgisi eğitilmiş herhangi bir dili tanıyabilirler. Kaydetme yapısı, tanınması beklenen dil ve kodlama örüntüsü için en uygun şekilde ve yeniden kullanılabilirlik ile çok işlemli programlama kavramı göz önüne alınarak önerilmiştir. N-gram ve kelime tabanlı araçları kullanıcılar tanıtmak istedikleri dil için eğitirler. Kullanıcı her dil için gerekli olan materyali toparlayarak eğitim sürecini gerçekleştirir. Bu araçların birçoğu sayısal dokümanlarda sıklıkla tercih edilen Avrupa dilleri ve bazı Asya dilleri için çalışmaktadır. Bu yüzden sisteme giriş olarak uygulanan dokümanların çeşitli kriterleri sağlaması gerekmektedir [5].

Peng ve arkadaşları [6] tarafından 2003 yılında yapılan çalışmada dil bağımsız ve işlem bağımsız metin sınıflandırması ile karakter tabanlı n-gram dil modelleri kullanan bir metot önerilmiştir. Bu yaklaşım en genel haliyle basit teorik ilkeler kullanan ve birçok çeşit dil sınıfında başarıyla çalışan bir yapıdır. Bu yapının kurgulanması için öznitelik vektörlerinin seçimine ya da büyük ön işlemlere gerek yoktur. Bu sistemin çalışmasını ve bağımsızlığını sınamak amacıyla Yunanca, İngilizce, Çince ve Japonca dilleri dil tanıma, metin sınıflandırma, yazar niteleme, yazı türü sınıflandırma ve konu tespiti gibi çözülmesi zor sorunlar ile test edilmiştir. Önerilen bu yaklaşım ile Asya kökenli dillerde sık karşılaşılan bir sorun olan kelime parçalama probleminin karakter seviyesinde dil modelleriyle çözülmesi sağlanmıştır. Bu çalışmada işlem bağımsız metin sınıflandırması için karakter seviyesinde n-gram modelleri kullanılmıştır. Önerilen yöntem dört farklı dil için 4 farklı metin sınıflandırma problemine uygulanmıştır. Yapılan testlerde müzik veya DNA gibi ardışık veriler ile çoklu kategori sınıflandırması problemi (Reuters–21578 veri kümesi) gibi sorunların çözümünde önerilen yaklaşımın başarısı ortaya konulmuştur. Bu çalışmanın kötü yanları ise kelimeden karakter seviyesine kadar parçalama sorunu nedeniyle işlem hızının düşük olmasıdır [6]. Nair ve arkadaşları [7] tarafından 2007 yılında yapılan çalışmada yazı dili İngilizce olmayan ülkelerin kendilerine özgü dil gruplarına karışan Latin kökenli metinlerin ayrıştırılarak gizli Markov modelleri ile bu metinlere ait dillerin analizleri incelenmiştir. Bu metinlerin resmi olmayan yazı dili Hintçe ve İngilizce anlamına gelen Hinglish olarak anılmaktadır. Bir başka örnek ise Malezya'dır. Bu ülkenin resmi olmayan yazı dili Manglish yani Maleyce ve İngilizce anlamına gelmektedir. Bu İngilizce dil gürültüsü yazı dili İngilizce olmayan ülkelere ait metinlerin ayrıştırılmasında büyük güçlüklere neden olmaktadır. Bu çalışmanın konusu gizli Markov modellerinin İngilizce olan ve olmayan metinleri ayrıştırabilmesi için bir sınıflandırma yapısının geliştirilmesidir. Bu sayede değişik dil modelleri ele alınarak oluşturulan metin editörleri ile dile bağlı renklendirme benzeri bir ayrıştırma aracı geliştirmektir. Geliştirilen yaklaşımın sözlük tabanlı metotlarla karşılaştırıldığında zaman ve uzay karmaşıklığı bakımından üstün olduğu görülmüştür. Sunulan modelin her dil için test edilememesinden dolayı modelin iyileştirilmesi gerekmektedir [7].

Ahmed ve arkadaşları [8] tarafından 2004 yılında yapılan çalışmada etkili dil sınıflandırma için n-gram tasarsız birikimli frekans ekleme metodu kullanılmıştır. Bu yeni sınıflandırma tekniği ile geleneksel Bayesian sınıflandırma metotlarına göre çok daha basit uygulanmaktadır buna rağmen sınıflandırma süreci kısa kelime parçalarında hız ve etkinlik olarak aynı performansı vermektedir. Sınıflandırma hızı öncelik sıralamalı n-gram sınıflandırma metotlarına göre 5-10 kat daha hızlı çalışmaktadır. N-gram sınıflandırma metotlarıyla dil tanıma süreci bölgesel farklılıklardan etkilenmeyen yüksek performansla çalışan literatürde yer edinmiş ve üstüne uzun araştırmalar yapılmış bir sistemdir. Bu sistemin en büyük dezavantajı sınıflandırmada kullanılan öncelik tabanlı metodun hızının düşük olmasıdır. Bu hız düşmesine neden olan şey sınıflandırma için test dokümanında frekans sayma ve sıralama değerlerinin bulunmasının gerekliliğidir. Bu çalışmada birikimli frekans ekleme metodu ile bu performans artışının sağlanması hedeflenmekteyse de uzun kelimeler ve paragraflar içinde geliştirilmesi gerekmektedir [8]. Constable [9] tarafından 2002 yılında çalışmada yeni teknik ve standartlarının geliştirilmesinin ihtiyacı üzerinde durulmuştur. Elbette yeni kullanım senaryolarına tam uyacak bir modelin geliştirilmesinde çeşitli zorluklar yaşanacaktır. Bu çalışma dil modeli ve dil içerikli kategorilerde gereksinim duyulan varlık bilimsel bir modelin eksikliğini gidermeyi hedeflemektedir. Bu noktada var olan çalışmalarda öne çıkan iki unsur dil ve yerelleştirme konunun özünü oluşturmaktadır. Mevcut problemlerin çözülebilmesi için dil tanıma sistemlerinde olması gereken bazı özelliklerin yazı sistemleri, varlık bilim, alan tanımlı veri kümesi, lehçeler ve diğer alt dil varyasyonları, üst dil bilim kategorileri, tarihsel dil değişimleri, dil bağımlı kategorileştirme ve yerelleştirme ile tüm bu sınıfların tartışıldığı konu başlıklarıyla bir inceleme yapılmıştır [9].

Adams ve arkadaşları [10] tarafından 1997 yılında yapılan çalışmada Java programlama dili platformunda tasarlanmış deneysel bir sistem tanıtılmıştır. Bu sistemin Java platformuyla geliştirilmesinin sebebi kullanıcı tarafı web göstericisinde çalışabilen doğal dil işleme sistemlerinde kullanılması için gerekli uygulama seviyesi dağıtık tabanlı bileşenlerinin Java dili içinde bulunmasıdır. Bu sayede doküman sunucusu ya da uzman sistem ajanları tarafından uyumlu çalışabilmektedir. Bu çalışmanın oluşturulmasında dil tanıma sistemlerinde sıklıkla kullanılan n-gram analizleri kullanılmıştır. N-gram analizlerde ihtiyaç duyulan etiketlendirme ve metin parçalama bilgisi dinamik olarak oluşturulmaktadır. Uygulama sonucunda elde edilmek istenilen sonuçlar dil modelleri arasında etiketlendirme performansı ile işlem bütünlüğünü arasındaki köprünün kurulmasıdır. Dil etiketleri tüm doküman için ya da doküman grupları için öğe tabanlı yapısal bir bileşen olarak ele alınmıştır. Sonuç olarak bu çalışmada karakter tabanlı etiketlendirme ile yapılan n-gram analizlerine yer verilmiştir. Bu modülasyon sayesinde internetten bağımsız çalışan bir yapı ile internet vekil sunucusu tabanlı iki sistem önerilmiştir. Bu çalışma sayesinde konuşma etiketleme, tümce tanımlama, yabancı kelime çevirisi ve konu etiketleme gibi zorlukların aşılmasıyla birlikte web tabanlı zeki arama ve gösterimleri gibi konuların işlenebileceği önerilmektedir [10].

Ölvecky [11] tarafından 2005 yılında yapılan çalışmada internetin büyümesiyle birlikte dil tanıma ve metin sınıflandırmanın öneminin artmasıyla birlikte bu alanda kullanılan n-gram varyasyonları incelenmiştir. İnternet üzerinde artan dokümantasyona karşılık olarak bu içeriklere özensizlikten meydana gelen metin ve gramer tabanlı hatalar dil tanıma işini zorlaştırmaktadır. Bu metotlar küçük değişiklikler ile metinlerden dil analizini %95 başarı oranıyla gerçekleştirebilmektedirler. Bu çalışmada tartışılan konu n-gram sınıflandırma tekniğinin bilinmeyen metinlerde sınıflandırma yeteneğinin yüksek kesinlikte bilinmesi için yapılması gerekenlerdir. Bu işlemlerden bir tanesi n-gram için kullanılan metin boyutunun azaltılmasıdır, uzun metinler eşit parçalar halinde kesilerek n-gram analiz büyüklüğü azaltılmaktadır. Bu alanda tartışılan bir başka konu ise n-gram giriş profili ile sistem eğitim büyüklüğünün ve yazım kalitesinin nasıl iyileştirilebileceği gösterilmiştir. Bu sayede tanıma performansının %10 kadar artış gösterdiği savunulmaktadır [11].

Bilcu ve Astola [12] tarafından 2006 yılında yapılan çalışmada yazınsal metinlerden dil tanıması yapabilecek yapay sinir ağı tabanlı hibrit bir metot önerilmiştir. Bu problem metinden ses birimine (fonem) bağlı bir çalışmadır. Bu problemde amaç yazılı bir metine ait harflerin kendilerine karşılık gelen ses birimlerine dönüştürülmesidir. Bu süreçte şayet dönüştürülmek istenen tek bir dil varsa dil tanıma işlemine gerek yoktur, ilgili yazılı metin direkt olarak ses birimlerine dönüştürülmesine rağmen giriş dili bilinmeyen metinlerde genellikle ilk aşama metin dilinin tespit edilmesidir. Bu sistemin geliştirilmesi sürecinde çok katmanlı geriye yayılım algoritması ve basit karar verme parametrelerine dayalı kural tabanlı bir sistem kullanılmıştır. Bu uygulama özellikle İngilizce ve Fransızca dilerinin tanınması amacıyla geliştirilmiştir. Bu sayede İngilizce veya Fransızca dillerinde yazılmış metinlerin ayrıştırılarak tanınan dil özelliklerine göre metinden ses birimlerine çevrim gibi modüler bir yapı tasarlanmıştır. Bu çalışma sayesinde iki dile ait metinlerden ses birimi hecelerine geçiş sürecinin sağlanabileceği sistemin bu doğrultuda geliştirilmesi ve mevcut sorunların çözümlenmesiyle sistemin daha kararlı bir şekil alacağı savunulmuştur [12].

Liu ve arkadaşları [13] tarafından 2005 yılında yapılan çalışmada önerilen sistem dil tanıma problemini karakterlerin resimlerinden tanımayı hedefleyen, birden çok dil barındıran dokümanlarda dil sınırlarının belirlenmesi için ilgili karakterlere ait resim verilerinin sınıflandırılması için yeni bir yöntem önerilmiştir. Bu amaçla prototip bir sınıflandırma yöntemi ve destek vektör makineleri kullanılmıştır. Büyük boyutlu eğitim veri kümesi nedeniyle her iki metot için eğitim hızını artıracak yeni bir teknik önerilmiştir. Bu metotlar kullanılarak Çince, İngilizce ve Japoncanın (Hiragana ve Katakana dahil) dillerinin tanınması hedeflenmiştir. Elde edilen sonuçlar, oldukça kararlı ve literatürde kabul görmüş yöntemlerle karşılaştırılabilecek düzeydedir [13].

Zhu ve arkadaşları [14] tarafından 2008 yılında yapılan çalışmada el yazısı ve makine çıktısı yazılardan dil tanınmasına üzerine bir yöntem önermişlerdir. Bu yöntemin gerçeklenmesi için çok sayıda gerçek içerikli dokümandan oluşan resim koleksiyonları kullanılmıştır. Bu dokümanlar 8 dilde 1500 kadardır. Bu diller Arapça, Çince, İngilizce, Hintçe, Japonca, Korece, Rusça ve Tayca ile bu dillere ait hem el yazısı hem de makine çıktısı içeriklerden oluşmaktadır. Yapılan testler sistemin yüksek karmaşıklıkta dokümanlar için iyi bir tanımlayıcı olduğunu göstermesine rağmen kod çizelgelerinin oluşturulmasında karşılaşılan uzaysal karmaşıklık gibi sorunlar nedeniyle sistemin uygulama aşamasında bazı sorunların aşılması gerekmektedir [14].

Baykan ve arkadaşları [15] tarafından 2008 yılında yapılan çalışmada internet ortamında girilen tek bir adresten dil tanımanın mümkün olduğuna ilişkin öneriler sunulmaktadır. Bu sayede belirli bir dile ait içeriklerin getirilmesini saplamayı amaçlayan internet tarama sistemleri ve arama motorları tarafından böyle bir sistemin yararlı olacağı savunulmuştur. Bu sistem gereksiz sayfaların taranmasını da engellemektedir. Sunulan, çalışmada birçok makine öğrenme algoritması değerlendirilmiş ve İngilizce, Fransızca, Almanca, İspanyolca ve İtalyanca için karşılaştırmalı testler yapılmıştır. Bu testler neticesinde en iyi metot F-ölçümü olmuştur. Bu yöntem ile tüm dillerde web tarayıcılarından elde edilen 25K boyutunda 1260 web sayfasında ortalama %90 oranında başarı elde edilmiştir. 5 K büyüklüğündeki internet arama motorlarından elde edilen başarı oranı ise %96 olmuştur. Bu çalışmada kullanılan performans kriteri F-Ölçümü'dür. Bu ölçüm makine dil tanıma başarı kriteri olarak anılmaktadır. Bu test süreçlerinde kullanılan sayfalar ODP verileridir. Bu internetin kategorilere göre taranması amacıyla geliştirilen ve dil bağımlı olarak yayınlanan bir veritabanıdır. Karşılaştırmalı algoritmalar olarak geleneksel Bayes metodu, maksimum entropi metodu kullanılmıştır. Bu çalışmada karşılaşılan en büyük sorun İngilizce gözüken web sayfası adreslerine bağlı içeriklerin İngilizce olmayan dillerde yazılmış olmasıdır. Kullanılan yöntem önceliklere ülkelere ait ccTLD kayıtlarının belirlenmesidir, bu her ülkeye ait olan ve alan adının sonuna o ülkeye ait tanımlayıcı karakter dizisinin gelmesi işlemidir. Bu aşamadan sonra her ülke grubu için 10 farklı durma kelimesi belirlenmektedir, bu durma kelimeleri her dil içinde en sık geçen kelimelerden seçilmektedir.

Bu durma kelimelerinin 8 ila 10 tanesi 1'den 10'a kadar kombinasyonlarla arama motorlarında aratılarak sonuçlar işlenmektedir. Bu sisteme son olarak 97 milyon tekrarlanmayan sonuç içeren bir internet arama motorunda 1260 rastgele sayfa eklenmektedir. Bu testlerde kişilerden ODP veri kümesinde kategorize edilmemiş 100 sayfayı tanımlamaları istenmiş ve ancak %50 başarı elde edilmiştir, sistem testlerinde ise bu başarı oranı %85 olarak gözlenmiştir [15]. Takçı ve Soğukpınar [18], harf tabanlı istatistiksel bir metot önermişlerdir. Bu yöntem ile n-gram ve ortak sözcük metodu gibi yöntemlerde gerekli olan sözcükler yerine dil tanıması için kullanılacak metin içinde geçen harflere ait bir sıklık durum analizi yapmışlardır. Bu sayede inceledikleri İngilizce, Fransızca, Almanca ve Türkçe dilleri için alfabe bazında harflerin bulunma dağılımlarını çıkartarak dil tanıma için verilen bir metin içerisindeki harfleri ait olduklara dillere göre puanlayabilecek bir sistem geliştirmişlerdir. Sağıroğlu ve arkadaşları [17], tarafından 2007 yılında yapılan çalışmada yapay sinir ağı modelinin dil sınıflandırmak amacıyla kullanılabileceği ve tanımlama başarısının sınırlı sayıda dil için literatürde kabul görmüş diğer metotlar ile aynı düzeyde olduğu gösterilmiştir. Bu çalışmada, farklı dillerdeki web sayfası içeriklerinin tanınması için yapay sinir ağları tabanlı zeki bir yaklaşım kullanılmıştır. Bir metine ait harf sıklık yüzdesinden doğal dilleri tanıyabilmek için eğitim esnasında; sisteme girişler, hangi dile ait olduğu bilinen metinlerin harf sıklık analizleri, çıkışlar ise bu dilleri temsil etmeye yarayan sayı değerleridir. Böylece sistem harf sıklık analizleri ile tanınması istenilen dil kodu arasındaki ilişkiyi Bölüm 3'te tanıtılan YSA yapısı ile öğrenmektedir. Bu zeki sistemde sunulan yöntem ile aradaki ilişkiyi öğrendikten sonra sisteme sadece tanınması istenilen dile ait metin bilgisi web sayfası üzerinden çevrimiçi olarak uygulanır, sistem öğrendiği bilgileri kullanarak metin bilgisine ait harf sıklığını modeller ve bu modelin özelliklerine göre uygulanan doğal dile ait dil sınıfını belirler. Bu sayede uygun koda karşılık gelen dil çözümlenerek açığa çıkar. Şekil 1'de geliştirilen sisteme ait akış şeması verilmiş ve detaylar sunulmuştur. Bu ilk yönteme ait çalışma uluslararası bir konferansta 2007 yılında sunulmuştur [17]. Bu bildiride geliştirilen yazılım web sayfası içeriklerini inceleyerek sayfanın dilini tespit etmekte ve ilgili sayfayı bahsedilen çeviri servisleri yardımıyla İngilizce diline dönüştürmektedir. Bu ilk çalışmada geliştirilen sistem, Türkçe, Fransızca, İtalyanca, Hollandaca ve Almanca dilleri için test edilmiştir. Geliştirilen dil tanıma sisteminin, yapay sinir ağları sayesinde tanıma performansı, literatüre göre yüksek olsa da yaklaşımın işlem karmaşıklığını arttırdığı gözlemlenmiştir. Şekil 1'de geliştirilen zeki dil tanıma sistemi blok diyagramı verilmiş ve bu sistemin çalışma prensibi akışı aşağıda verilen adımlarda özetlenmiştir.

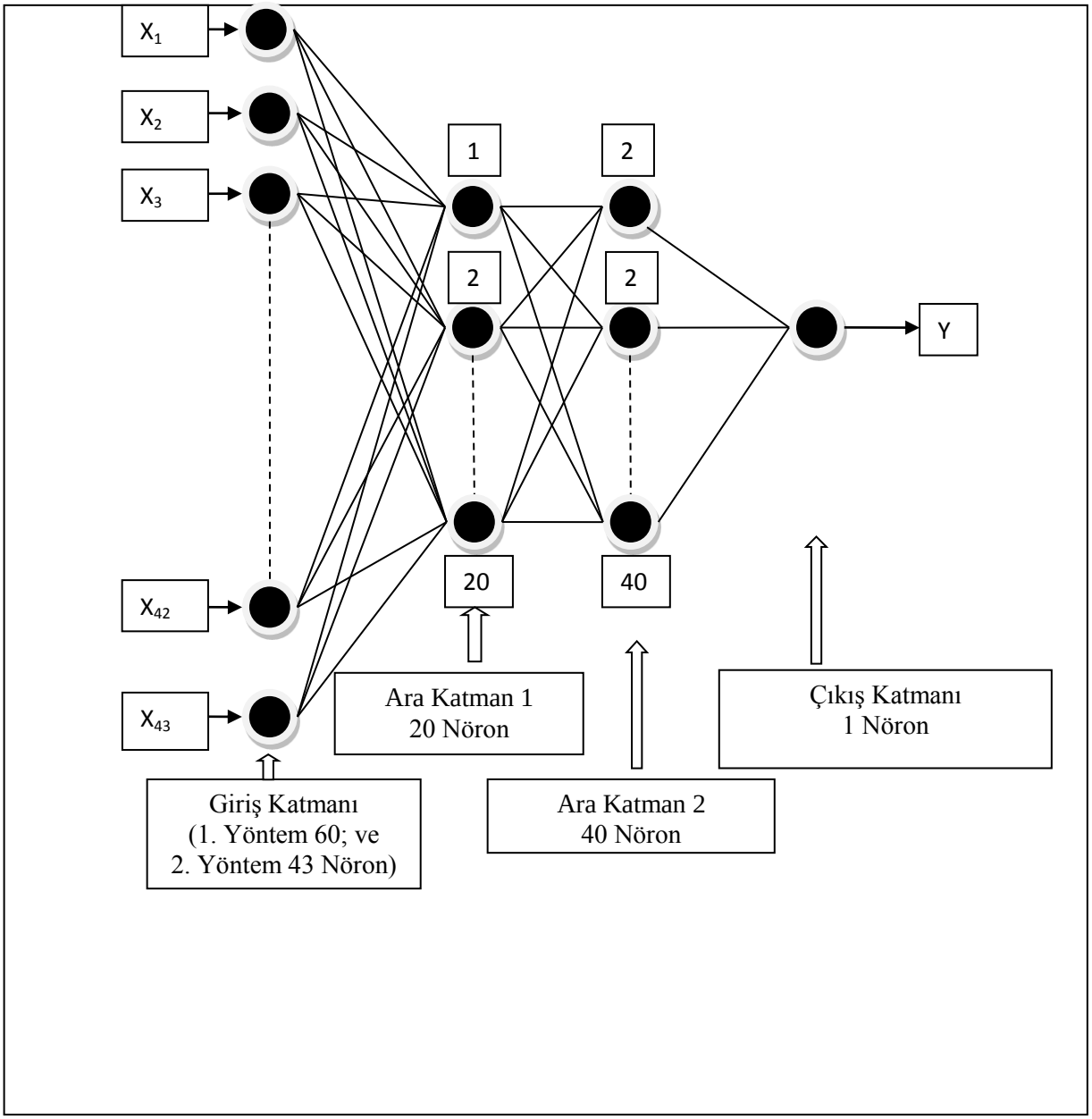
1. İçerik Getirme Modülünde Kullanıcı Girişine Web Sayfası Adresi Uygula
2. Varsayılan Karakter Kodlaması ile Sayfa Kaynak Kodunu Sisteme getir
3. Olması Gereken Karakter Kodlama Kümesi ile Sayfa Kodunu Sisteme Getir
4. Kaynak Temizleme Modülünde Sayfa Kodu
5. Üzerinde İngilizce Zorunluluk Kelimeleri ile HTML Kodlarını Temizle
6. Dil Tanıma Modülünde Önerilen Alfabe Kümesi ile Karakter Sıklık Analizi Yap
7. Karakter Frekans Analizi Sonucunu Önceden Eğitilmiş Yapay Sinir Ağına Uygula
8. Yapay Sinir Ağı Çıkışını Karşılaştır ve Dil Tanıma Sonucu Bul
9. Çeviri modülünde Web Sayfasını Tanıma Sonucundan İngilizceye Çevir

Önerilen bu yöntem sayesinde farklı dillere ait alfabelerden oluşan bir küme içerisinden eğitilen yapay sinir ağının %86 oranında düşük bir başarıyla dil tanımlama işlemi gerçekleştirmektedir [17]. Farklı içerik ve dil tanımada beklenen performansı vermemesinden dolayı yeni yöntemler geliştirmek ve kullanmak zorunluluk haline gelmiştir. Yavanoğlu ve Sağıroğlu [25], tarafından yapılan bir diğer çalışmada ise Latin alfabe grubuna dahil 15 dil için karakter frekans analiziyle çözüm uzayı karmaşıklığını göz ardı eden yapay sinir ağı temelli bir dil sınıflandırıcı önerilmiştir [25]. Bu çalışma "Web Tabanlı Otomatik Dil Tanıma ve Çeviri Sistemi Geliştirilmesi" başlıklı yüksek lisans tezinde detaylı olarak açıklanmıştır [24].

Sonuç olarak, bu alanda yapılan çalışmalar incelendiğinde hızlı cevap verebilen durağan ve karar mekanizması kararlı bir sistem tasarımına ihtiyaç bulunmaktadır. Literatür de önerilen metotlar metin bilgisinin çeşitliliği ile karmaşıklığın arttığını göstermektedir. Ayrıca, birçok platformda kullanılan sözcük bazlı metotların güncel veritabanı ihtiyaçlarının bulunduğu özellikle web tabanlı olanlarının [17], [24], [25] nolu çalışmalar hariç pratik olarak uygulamasının bulunmadığı, yazarlar tarafından geliştirilen web tabanlı yaklaşımında performansının geliştirilmesi gerektiği görülmüştür.

3. Yapay sinir ağları ve levenberg-marquardt öğrenme algoritması

Yapay Sinir Ağları (YSA), yapılarına göre ileri beslemeli (feed forward) ve geri beslemeli (feed back) olarak ikiye ayrılır. İleri beslemeli ağlarda işaretler, giriş katmanından çıkış katmanına doğru tek yönde iletilirler. Geri beslemeli ağlarda, çıkış ve ara katman çıkışları kendinden önceki katmanlara ya da girişe geri beslenir. İleri beslemeli ağlara Çok Katlı Perseptronlar (MLP), Radyal Tabanlı Fonksiyon Ağı (RBFN) ve LVQ (Learning Vector Quantization) örnek verilebilirken, ART (Adaptive Resonance Theory), SOM (Self Organizing Maps) ve Elman ve Jordan ağları geri beslemeli ağlara örnek olarak verilebilir [16]. Bu YSA yapıları Şekil 1'de verildiği gibi ileri beslemeli standart birleşmeli yapıda olabileceği gibi geri beslemeli dinamik yapıda da olabilir. Bu çalışmada, ileri beslemeli ağ yapılarından olan Şekil 2'de verilen MLP yapısı kullanılmıştır. MLP yapısının tercih edilmesinin nedeni, bilinen en eski YSA modellerinden olması çok farklı uygulamalarda başarılı olması ve sınıflandırma problemlerinde başarılı sonuçlar üretmesidir. Bunların yanında, farklı öğrenme algoritmaları ile kullanıma uygun olması MLP'nin sağladığı diğer bir üstünlüktür. Şekil 1'de bu çalışma kapsamında kullanılan MLP yapısı ve nöron sayıları örneklenmiştir. MLP modeli, bir giriş, bir veya daha fazla ara katman ve bir de çıkış katmanı içerir. Bir katmandaki bütün nöronlar bir sonraki katmandaki bütün nöronlara bağlıdır. Giriş katındaki nöronlar tampon gibi davranırlar ve giriş sinyalini ara katmandaki nöronlara dağıtırlar. Ara katmandaki her bir nöronun çıkışı, kendine gelen bütün giriş sinyallerini takip eden bağlantı ağırlıkları ile çarpımlarının toplanması ile elde edilir. Elde edilen bu toplam, çıkışın toplam bir fonksiyonu olarak hesaplanabilir.



Şekil 2. Bu çalışma kapsamında kullanılan MLP yapısı ve nöron sayıları

LM (Levenberg-Marquardt) öğrenme algoritmasında hedef, parametre vektörü *w*'nin, amaç fonksiyonu E(*w*)'yi minimum yapacak şekilde optimize edilmesidir. LM algoritmasının kullanılmasıyla yeni vektör w_{k+1}, farz edilen vektör w_k'dan Eş. 1 ve 2 yardımıyla aşağıdaki gibi hesaplanabilir.

$$w_{k+1} = w_k + \delta w_k \quad (1)$$

burada δw_k ifadesi

$$(J_k^T J_k + \boldsymbol{\lambda} I)\delta w_k = -J_k^T f(w_k) \quad (2)$$

eşitliğinden faydalanılarak hesaplanır. Eş. 2'de;

$\boldsymbol{J_k}$: f'nin w_k değerlendirilmiş Jakopyeni,
$\boldsymbol{\lambda}$: Marquardt parametresi, ve
I : birim veya tanımlama

evenberg-Marquardt algoritmasında hesaplama akışı ve diğer detaylar için [16] nolu kaynağa başvurulmalıdır.

4. Web sayfası içerik tanımlama

Bu çalışmanın amacı, sayısal dokümanların ve web sitesi içeriklerinin dilini tanımak ve tanınan dile göre içeriğin istenilen dile otomatik olarak çevrilmesini sağlamak için web tabanlı otomatik bir sistem geliştirilmesidir. Bu geliştirmede üç farklı yöntem kullanılmıştır. İlk iki yöntem daha önce sunulmuş ve bu çalışma kapsamında yeni bir yöntem geliştirilerek uygulanmıştır. Bu adımlara ait bilgiler aşağıda alt başlıklarda verilmiştir.

4.1.1 Birleşim Tespit Yöntemi

Birleşim tespit yöntemi "Web Tabanlı Otomatik Dil Tanıma ve Çevirme Sistemi" başlıklı çalışmamızda detaylı olarak sunulmuştur [25]. Bu çalışmanın motivasyonu ile hazırlanan kesişim tespit yöntemi aşağıda sunulmuştur. Birleşim tespit yöntemi ile oluşturulan yapının kesişim tespit yöntemine göre farkları ve avantajları sonraki bölümlerde sunulmuştur.

4.1.2 Kesişim Tespit Yöntemi

Bu yöntem, bu çalışma kapsamında önerilmiş olan yeni bir yöntem olup YSA'nın öğrenme performansını artırmak için geliştirilmiştir. Yöntem, tanınması istenilen dillere ait alfabelerin tek bir ortak alfabeye entegre edilme fikrine yeni bir bakış açısı getirmektedir. Bu yöntem kullanılarak oluşturulan ortak alfabe kümesinde verilen standart ve genişletilmiş Latin alfabe kümesinin tanınması istenilen 15 farklı ülke dili için analizlerden, dil içerisinde tekrar eden en az 5 harf seçilerek, kesişim tespit yöntemi eğitim ve test kümeleri oluşturulmuştur. Detay için [24] nolu kaynağa bakınız. Bu sayede bahsedilen 15 dil için her dil alfabesi içinde 5 kez ve daha fazla tekrar eden harflerden ortak alfabe kümesi oluşturulmuştur. Kesişim Tespit Yöntemi ile oluşturulan bu ortak alfabe kümesi ile eğitilen ve tasarlanan sistemin adımları aşağıda verilmiştir.

1. Tanınması istenilen dillere ait alfabe karakterleri çıkartılır.
2. Bu alfabelerin karakterlerine ait kesişim kümesi bulunur.
3. Giriş metni (web sayfası, Word, PDF vs.) sisteme giriş olarak uygulanır.
4. Sistem tarafından alfabelerin kesişim kümesinde bulunan karakterlere otomatik olarak bir sayı atanır.
5. Girilen metin içinde yer alan harfler sayaç tarafından sayılır, alfabe kesişim kümesiyle karşılaştırarak tekrar eden harflerin sayıcı değerini bir arttırılır. Bu sayede alfabe kesişim kümesinde bulunan harflerin metin içinde kaç kez geçtiği bulunur.
6. Metin analizi tamamlandıktan sonra, harf sayaç sonuçları toplanarak yinelenen toplam harf sayısı elde edilir.
7. Sistem her harfin girilen metin üzerindeki yüzdelik dağılımını hesaplar.
8. Eğitim kümesi oluşturulurken bu yüzdelik dağılımlara, incelen dile ait bir sıra numarası ağın çıkışına verilir.
9. Eğitilen ağ, yöntem neticesinde ağa uygulanan yüzdelik değerler neticesinde ağın ürettiği çıkış, eğitimde kullanılan sıra numaraları ile karşılaştırılarak test metninin hangi dile en çok yakınsadığı ve buna bağlı olan yüzdelik hatası bulunur.
10. Elde edilen sonuçlar çeviri programları, arama motorları, ofis programları gibi otomasyon yazılımlarında kullanılarak kullanıcılara sunulur.

Şekil 4'de verilen Web Sayfası Giriş Modülü ile kullanıcılar içerik dillerini bilmedikleri web sayfalarını sisteme giriş olarak uygulayabilmektedirler. Geliştirilen yazılım giriş olarak uygulanan web sayfaları için ilk etapta UTF-8 kodlaması kullanılarak içeriğin bilgisayar indirilmesi sağlamakta sonrasında ise meta etiketlerine bakarak internet yüklendikleri ülkeyi, sonrasında ise indirilen ilk içerikten kod analizi yaparak karakter kodlama kümesine erişmektedir. Bir web sitesinde, içerik bilgisinin sayfanın her metin dizisinde aynı olmaması ve HTML kodlama dili İngilizce olduğu için sayfa üzerinden erişilen ve HTML filtreleme ile temizlenen içerik bilgisinin sağlıklı yapılamaması karşılaşılan en büyük sorunlardandır [17].

4.2 Eğitim ve Test Metodolojisi

Latin alfabesi 80'den fazla farklı dil için temel teşkil etmektedir. Alfabetik birleşim kümesi bir dile ait özel karakterlere duyarlıdır. Latin alfabesi zaman içinde diller arasında gelişerek farklılık göstermektedir. Klasik Latin alfabesi a, b, c, d, e, f, g, h, i, k, l, m, n, o, p, q, r, s, t, v, x, y, z harflerinden oluşmaktadır. Bu harflere ek olarak günümüzde kullandığımız modern Latin alfabesi "j, u, w" harflerini de içermektedir. Bu çalışmaya konu olan 15 dil içerisinde İngilizce, Almanca, Fransızca, Macarca, İspanyolca dilin özel durumları dışında modern Latin alfabesindeki 26 harfi kapsamaktadır. Bu çalışmaya konu olan diğer diller ise modern Latin alfabesinin bazı harflerini içermemektedir. Bu dillere ait karşılaştırma Çizelge 1'de örnek olarak verilmiştir. Çizelge 1'den de görülebileceği gibi "Ülke" anahtar kelimesi Letoncada "valsts" kelimesine dönüşmektedir. Bu sebeple "Ülke" kelimesine karşılık bulunan dokümanlar ile "valsts" kelimesine karşılık bulunan dokümanlar kendi dil sınıfları içinde yakınsadıklarından eğitim ve test kümelerinde tercih edilmektedir.

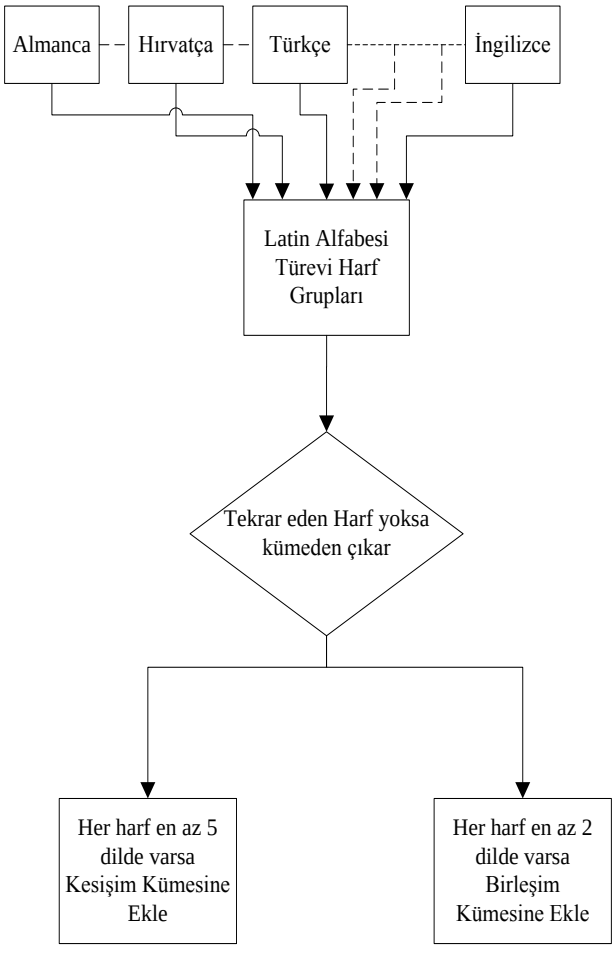
Latin alfabesinin genişletilmiş grubunda 161 harf bulunmaktadır. Bu harflerden bazıları Almanca'da bulunan "ß" gibi sadece dile özgü olmakla birlikte Galce ve Türkçe'de olan "Ü" gibi birden fazla dilde ortak olarak bulunmaktadır. Bu aşamada sadece dile özgü harflerin kullanımı YSA'nın öğrenme sürecini etkileyerek, sürecin ezbere kaymasına neden olmaktadır. Bu yüzden, Birleşim Tespit Yönteminde YSA'nın karar mekanizmasını iyileştirmek için tek dile özgü harfler kullanılmamıştır [25]. Birleşim Tespit Kümesi oluşturulurken her harfin en az 2 dilde bulunması esas alınmıştır. Bu çalışmanın konusu olan Kesişim Tespit Yönteminde ise kullanılan harfler hem klasik Latin alfabesinde hem de genişlemiş Latin Alfabesinde yer alan en az 5 alfabede ortak olan harflerdir. Bu sayede işlem uzayı küçültülerek sistemin Latin Alfabesinden türemiş her dil için çalışması esas alınmıştır.

Önişlemler sonucunda önerilen yöntemlere göre harf sıklıklarının giriş metni içinde bulunma oranları hesaplanmaktadır. YSA harf sayıları ile doğal diller arasında bir ilişki kurulmasını sağlamaktadır. Alfabetik harf sıklığından dil tanıyan sistemin teorik yapısı Şekil 3'de verilmiştir. Sistem alfabelere ait ortak veritabanı ile alfabeleri oluşturan diller arasındaki ilişki modelini öğrenir veya mevcut dil desteklerinden faydalanır. Bunun için, sisteme uygun sayıda veri kümesi oluşturulur ve sistem eğitilir. Bu bölümde önerilen yöntemlere ait detaylar açıklanmıştır.

Şekil 3'de akış diyagramı verilen yöntemler bütünü uygulanırken tanımlanması istenilen n sayıdaki latin alfabesi tek bir havuzda gruplanmaktadır. Bu havuzdan seçilen ve alfabeler içinde sadece bir kez tekrar eden harfler temizlenmektedir. Havuzda kalan ve en az 2 alfabede tekrar eden harfler birleşim tespit kümesine, en az 5 alfabede tekrar eden harfler ise kesişim tespit kümesine alınmaktadır. Bu sayede n sayıda alfabe için tekrar eden harflere göre bilimsel olarak anlamlı kümeler oluşturulmaktadır.

Çizelge.1. Anahtar kelime tablosu

	Anahtar 1	Anahtar 2	Anahtar 3	Anahtar 4	Anahtar 5	Anahtar 6	Anahtar 7	Anahtar 8	Anahtar 9	Anahtar 10
Almanca	Männlich	Frau	Mensch	Welt	Land	Wissenschaft	Kultur	Kunst	Politik	Leben
Arnavutça	marr pozitë	Femëror	njerëzor	planeti	fshatar	shkence	kulturë	zanat	Politikë	jetëdhënës
Fransızca	Homme	Femme	Humain	Terre	Pays	La Science	Culture	Art	Politique	La Vie
Galce	Dyn	Benyw	Dynol	Daear	Gwledig	Gwyddoniaeth	Diwylliant	Celfyddyd	Gwleidyddiaeth	Bywyd
Hırvatça	čovjek	Žena	čeljade	kopno	domovina	grana	gajenje	likovni	politička	osoba
İngilizce	Man	Woman	Human	Earth	Country	Science	Culture	Art	Politics	Life
İrlandaca	fear	Bean	daonna	talamh	tír	eolaíocht	cultúr	ealaín	pholaitíocht	beatha
İspanyolca	Hombre	Mujer	Humano	Tierra	País	Ciencia	Cultura	Arte	Política	Vida
İtalyanca	Uomo	Donna	Umano	Terra	Paese	Scienza	Coltura	Arte	Politica	Vita
Letonca	cilvēks	sieva	cilvēka	zeme	valsts	zinātne	kultūra	māksla	Politika	dzīvība
Macarca	ember	Nõ	emberi	szárazföld	ország	tudomány	kultúra	mûvészeti	politikai	élet
Maltaca	Ragel	Mara	Uman	Dinja	Pajjiz	Xjenza	Kultura	Arti	Pulitka	Hajja
Portekizce	Homem	Mulher	Humano	Terra	País	Ciência	Cultura	Arte	Política	Vida
Türkçe	Erkek	Kadın	İnsan	Dünya	Ülke	Bilim	Kültür	Sanat	Politika	Hayat
Vietnamca	người đàn ông	người phụ nữ	con người	trái đất	nước	khoa học	văn hóa	nghệ thuật	chính trị	cuộc sống



Şekil 3. Birleşim ve Kesişim Yöntemi Blok Diyagramı

Zeki dil tanımlama ve çeviri sistemine ait adımlar aşağıda sırayla sunulmaktadır.

1. Kesişim ve birleşim kümesinden elde edilen değerler ile hedeflenen çıkışın seçimi uygun YSA yapısının seçimi,
2. Bu yapıya uygun öğrenme algoritması ve uygulanan algoritmanın uygun parametrelerinin seçimi,
3. Seçilen yapıya uygun giriş, ara katman ve çıkış yapay nöron sayılarının belirlenmesi,
4. Seçilen nöronlarda kullanılacak olan aktivasyon fonksiyonunun belirlenmesi,
5. Karakter frekans analizleri sonucunda oluşturulan veri kümesinin birleşim ve kesişim yöntemlerine göre YSA yapısına uyarlanması
6. Birleşim ve kesişim veri kümesinin normalleştirilmesi ve YSA için giriş değerlerinin oluşturulması
7. Normalleştirilmiş frekans bilgisinin YSA giriş katmanına atanması
8. Tasarlanan YSA modellerinin LM algoritması ile 15 farklı dil için eğitilmesi
9. Ağırlık değişim döngülerinin minimum eğriye ulaşıncaya kadar tekrar edilmesi
10. Sonuç ağırlık değeriyle dil örüntüleri arasında ilişkinin kurulması
11. Sistemin 15 farklı dil HTML, PDF ve DOC dosya formatları ile test edilmesi

Önerilen yöntemler incelenen dillerin alfabelerinin farklı sayılarda birleşmesinden oluşmaktadır. Bu sebeple YSA her yöntem için farklı giriş sayılarıyla yeniden tasarlanmıştır.

Çizelge 2'de kesişim tespit yöntemi için 15 farklı dile ait alfabelerin özel karakterleri için kesişim kümeleri verilmiştir.

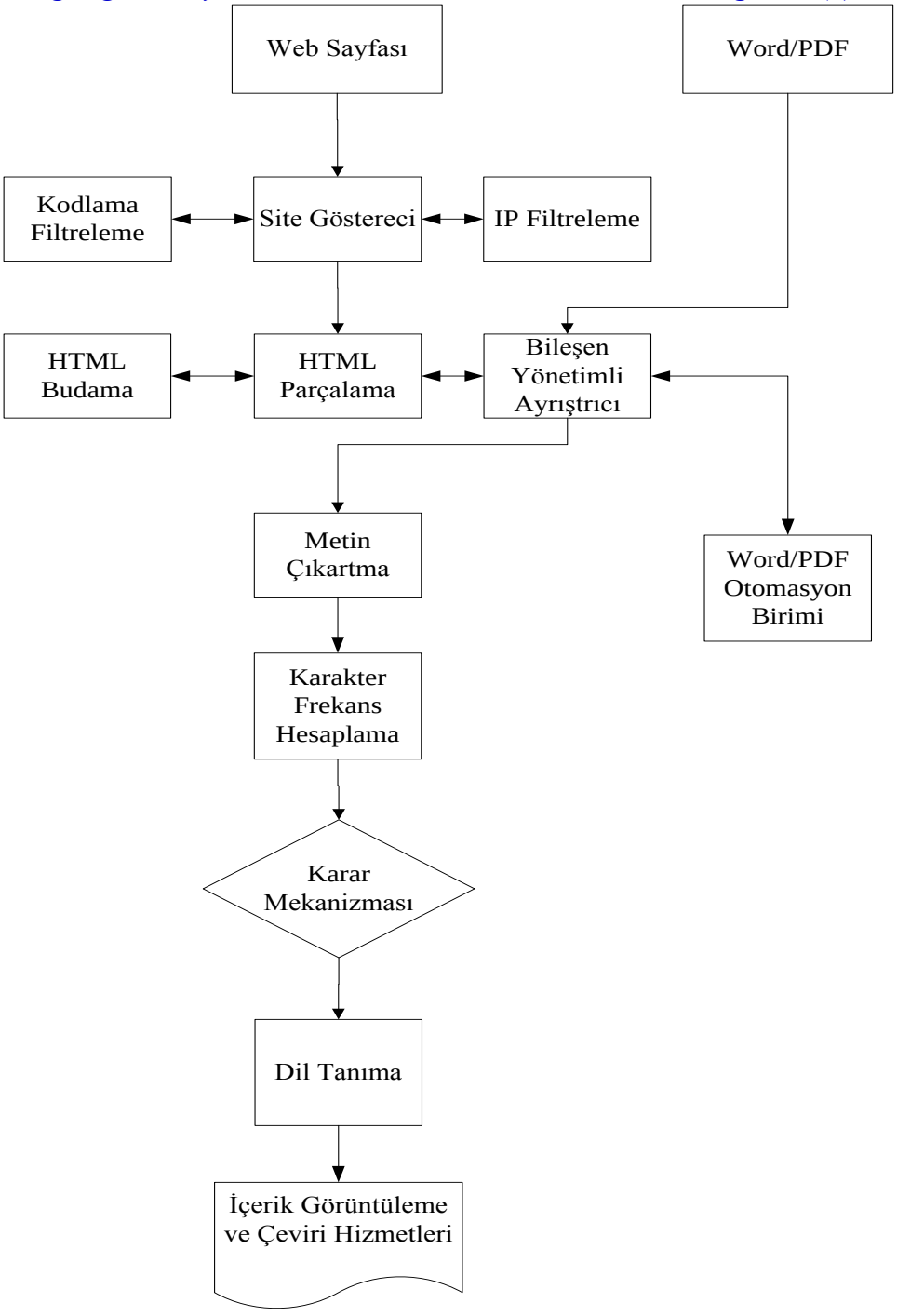
izelge 2. Alfabelere ait özel karakterler için harf kesişim kümesi [24]

	Á	**À**	**Â**	**Ç**	**É**	**È**	**Ê**	**Í**	**Î**	**Ó**	**Ô**	**Ö**	**Ú**	**Ù**	**Û**	**Ü**
	á	**à**	**â**	**ç**	**é**	**è**	**ê**	**í**	**î**	**ó**	**ô**	**ö**	**ú**	**ù**	**û**	**ü**
Almanca				Ç								Ö				Ü
Arnavutça									Î		Ô				Û	
Fransızca		À	Â	Ç	É	È	Ê		Î		Ô			Ù	Û	Ü
Galce	Á	À	Â		É	È	Ê	Í	Î	Ó	Ô	Ö	Ú	Ù	Û	Ü
Hırvatça																
İngilizce		À	Â	Ç	É	È	Ê		Î	Ó	Ô	Ö			Û	Ü
İrlandaca	Á				É			Í		Ó			Ú			
İspanyolca	Á				É			Í		Ó			Ú			Ü
İtalyanca		À			É	È			Î					Ù		
Letonca																
Macarca	Á				É			Í		Ó		Ö	Ú			Ü
Maltaca		À				È								Ù		
Portekizce	Á	À	Â	Ç	É		Ê	Í		Ó	Ô		Ú			Ü
Türkçe			Â	Ç					Î			Ö			Û	Ü
Vietnamca	Á	À	Â		É	È	Ê	Í		Ó	Ô		Ú	Ù		
Kesişim Tespit	6	7	6	5	9	6	5	6	6	7	6	5	6	5	5	8

Bu çalışma kapsamında geliştirilen yöntemin patenti tarafımızca tescil edilmiştir [27].

5. Geliştirilen yazılım platformu ve sonuçlar

Geliştirilen platforma ait akış şeması Şekil 4'de, geliştirilen ve GENIUS adı verilen yazılıma ait arayüz yazılımı Şekil 5'de verilmiştir. Bu yazılım, Latin alfabesine mensup 15 farklı dil için kullanılmıştır. Bu diller Dil Listesi (Language List) etiketi altında bulunan metin kutusunda eğitim sırasıyla verilmiştir. Latin Alfabesi 75'den fazla dile sahip çok geniş bir kitleye hitap eden yapıya sahiptir. Bu sebeple hazırlanan yazılımın esnekliği belirli bir coğrafi konumda konuşulan dillerin sisteme entegrasyon süresinin kısalmasını sağlamaktadır. Şekil 5'de verilen dil Listesi solunda yer alan 2 farklı metin kutusu Birleşim ve Kesişim metotlarını temel alarak YSA'nın sonucunu varsayılan dil çıkış sonuçları ile karşılaştırmaktadır. Bu alanlarda her metot için varılan sonuçların hata oranları hesaplanarak başarı sıraları ekrana otomatik olarak basılmaktadır. Eğitim veri kümeleri oluşturulurken Latin alfabesi grubuna dahil 15 farklı dil için giriş amacıyla kullanılacak sayısal (WORD, PDF ve HTML) dokümanları kullanılmıştır. Geliştirilen bu yaklaşımda test verilerinin üretilmesi için Çizelge 1'de verilen anahtar kelime tablosu ve Google arama motoru kullanılmıştır. Bu tablo da verilen Erkek, Kadın, İnsan, Dünya, Ülke, Bilim, Kültür, Sanat, Politika, Hayat kelimeleri bir dil içerisinde en sık kullanılan kelimeler olarak önerilmiştir. Veri kümeleri oluşturulurken bu kelimelerin tanınması istenilen diğer dillerde çevirileri yapılmıştır. Bu çeviriler bir arama motorundan rastgele seçilen sayfalarda yer alan sonuçların her dilde benzer sayısal dokümanlara ulaşmayı sağlanacağı düşüncesiyle yapılmıştır. Bu sayede farklı kişiler tarafından yazılan ama benzer özellikler taşıyan toplumsal, bilimsel, politik vb. yazılar barındıran sayısal dokümanlar veri kümelerinin oluşturulmasında kullanılmıştır. Bu anahtar kelimelerin kullanılmasıyla elde edilen sayısal dokümanların farklı dillerde aynı konuları içermesi ile önişlemler sonucunda elde edilecek eğitim veri kümelerinin rastlantısal sonuçlar içermemesini de hedeflenmektedir. Eğitim işlemi için sayısal dokümanların önişlemden geçirilmesi gerekmektedir. Bu önişlem sonucunda Latin Alfabesi grubuna ait dillerin sınıflandırılması gerçekleşmektedir.HTML kodu budama ve bileşen tabanlı HTML'den metin ayrıştırma ile Bölüm 4.1'de belirtilen kesişim ve birleşim tespit yöntemleri kullanılmıştır. Bu yaklaşımlar sayesinde farklı şekillerde çıkartılan metin bilgisi ile dil tanıma gerçekleştirildikten sonra tespit edilen dilin, istenilen dile dönüşümü yapılarak yazılım içine gömülmüş olan internet göstericisinde kullanıcı tarafından giriş olarak uygulanan web sitesinin hiç bir içerik bozulmasına uğramamış çevirisi kullanıcılara sunulmaktadır. Dil tanıma işleminde dil tanıyıcı tarafından tanınan sayısal dokümanın dili en solda bulunan Kaynak Dil (Source Language) ile etiketlenmiş ve listeden otomatik olarak seçilmektedir. Bu kısım direkt olarak Google Dil Çevirici (Google Translator) ile entegre çalışmaktadır. Kullanıcıların herhangi bir Google hesabı olması gerekmemektedir. Yazılım gerekli çeviri içeriklerini kendi alt dönüşümleriyle kullanıcılara iletmektedir. Kaynak Dil etiketinin altında Hedef Dil (Destination Language) için açılan kutu nesnesi bulunmaktadır. Erişimi yapılan dokümanın kaynak dil bilindikten sonra hedef dili seçerek Çevir (Translate) butonuna tıklandığında bir internet sayfası açılarak sisteme giriş olarak uygulanan sayısal dokümanın çevirisi bir internet sayfasında kullanıcılara sunulmaktadır. Çevir ve Kaydet (Translate and Save) butonu tıklanırsa ilgili çeviri metinleri dil tanıması ve çevirisi istenilen sayısal dokümanlar ile aynı alt klasöre HTML dosyası olarak kopyalanmaktadır. Bu sayede DOC ya da PDF uzantılı dosyalar her bilgisayar sistemi tarafından okunabilen HTML dosyasına kullanıcı tarafından anlaşılan dil çevirisi ile kayıt altına alınmaktadır.



Şekil 4. Geliştirilen sisteme ait akış diyagramı

Böylece kullanıcılar hiç dillerini bilmedikleri dokümanların dillerine ve istedikleri dile çevirilerine sadece tek bir tıklama ile ulaşabilmektedir. Bu yazılımın çalışması betimlendiği için tek tıklama işlemi farklı parçalar halinde sunulmaktadır. Bu yazılım internet sitelerini, format bozulması yaşanmadan yeni bir web göstericisi açarak kullanıcılara sunmaktadır. Geliştirilen GENIUS isimli yazılımın doküman içeriklerinin tespit yüzdelerini ortaya koymak için iki farklı test çalışması yapılmıştır. Testin birinci bölümünde Microsoft Word ve Adobe Reader tabanlı dokümanlar test edilmiştir. İkinci kısımda ise, web sayfalarının testleri yapılmıştır. Bu testler ve sonuçları farklı başlıklar altında verilmiştir. Bu testte ise HTML, PHP, JSP ve ASP tabanlı dokümanlar temel alınmıştır.

Bu yazılım platformu C# dilinde Visual Studio 2005 ortamında sayısal dokümanları işlemek amacıyla açık kaynak kodlu otomasyon yazılımları kullanılarak tasarlanmıştır [20-23]. Bu nedenle Adobe Reader PDF veya Microsoft Word DOC formatlarının işlenmesi amacıyla her iki yazılımında kullanıcının makinesinde bulunmasına ihtiyaç duymamaktadır. Yazılım genel olarak önceki bölümlerde anlatılan 6 modülden oluşmaktadır. Bu modüller Şekil1'de verildiği gibi Sayısal Doküman Analiz Modülü, Metin Otomasyon Modülü, Frekans Analiz Modülü, Zeki Dil Tanıma Modülü, Dil Geçişli Çeviri Modülü ve Çıkış Modülü oluşmaktadır. Bu yazılımın adını İngilizce deha manasına gelen genius kelimesinden almaktadır.

GENIUS (Gazi Engineering Intelligent Unified Service) bir diğer deyişle Gazi Mühendislik Zeki Birleşim Servisi uluslar arası çapta geliştirilmiş bir araçtır. Bu araç sayesinde kullanıcılar dillerini bilmedikleri internet sayfalarını, düz metinleri, Word, PDF ve HTML dokümanlarını kendi dillerine otomatik çevirebilmektedir. Bu çalışmada elde edilen sonuçlar Çizelge 3, 4, 5 ve 6'de verilmiştir. Elde edilen sonuçlar geliştirilen sistemin PDF, Word ve HTML dokümanlarını %100 başarıyla 15 dil için tanımlamış ve 64 dile dönüşümde geliştirilen yazılım platformuyla gerçekleştirilebilmektedir. Burada sonuçta en yüksek tanıma oranına sahip olan değerler %52-%99 arasında olup, geliştirilen model o dil için en yüksek değere sahip olan orandan diğer dillere doğru tanımlama oranlarıdır. Geliştirilen sistem en yüksek orana sahip dili doküman dili olarak belirlemektedir. Bu sebeple en yüksek yüzdelik puan diliminde olan dil tanımlama sonucu olarak tanımlanmaktadır. Sistemin [25] Türkçe olan dokümanlarla ilgili olarak yapılan testlerde elde edilen ortalama sonuçları Çizelge 3'de verilmiştir. Doküman bazında elde edilen sonuçlara ve diğer dillere ait verilerine ise [24] nolu kaynaktan ulaşılabilir. Elde edilen sonuçlar her iki doküman tipi ve her iki tespit yöntemi kullanılarak yapılmıştır. Elde edilen sonuçlardan Microsoft Word DOC ve Adobe PDF doküman tipleri için Kesişim Yöntemi (#K) ile %99 ve Birleşim Yöntemi (#B) ile %99 oranında bir başarı elde edilmiştir. Türkçe için her iki dosya formatı ve her iki yöntemin önceki bölümlerde bahsedilen kriterlere göre %100 başarılı olduğu görülmüştür.

Çizelge 3. Türkçe için doküman bazında ortalama başarı oranları [24]

Dil	Türkçe			
Format	DOC Dosya Tipi		PDF Dosya Tipi	
Yöntem	#K	#B	#K	#B
Almanca	0%	0%	0%	0%
Arnavutça	0%	0%	0%	0%
Fransızca	0%	0%	0%	0%
Galce	0%	0%	0%	0%
Hırvatça	0%	0%	0%	0%
İngilizce	0%	0%	0%	0%
İrlandaca	0%	0%	0%	0%
İspanyolca	24%	25%	25%	23%
İtalyanca	44%	44%	44%	43%
Letonca	59%	60%	60%	59%
Macarca	72%	73%	72%	71%
Maltaca	83%	83%	83%	82%
Portekizce	92%	92%	92%	91%
Türkçe	**99%**	**99%**	**99%**	**99%**
Vietnamca	93%	93%	93%	93%

6. Sonuç ve öneriler

Bu çalışmada web tabanlı otomatik dil sınıflandırma ve çeviri sistemi geliştirilmiştir. Geliştirilen sistem üç farklı yöntemden oluşmaktadır. İlk yöntemde geliştirilen sistem başarılı olmasına rağmen bazı dillerde düşük tanımlama oranıyla karşılaşılmıştır. YSA'nın öğrenme sürecinde başarı düşüklüğü, dil tanıma başarısının yetersizliği, sadece web siteleri ile çalışması, çeviri yeteneklerinin sınırlı oluşu, sunucu bilgisayar ve bant genişliği bağımlı olması ve tanınması istenilen dil sayısı arttıkça bu artışa paralel olarak artan eğitim veri kümesi ve YSA giriş sayısındaki büyük artış sistem karmaşıklığını yükseltip eğitim performansını düşürdüğü web sayfalarının aynı standart ve formatta bulunmamalarından kaynaklanan içeriğin doğru analiz edilememesi gibi sorunlardan ötürü üç sistem geliştirilmiştir.

Tanınması istenilen dil sayısı ile sistem karmaşıklığı orantısını kabul edilebilir düzeyde tutan hatta bazı durumlarda bu dil sayısı ve karmaşıklığı ters orantılı olarak değiştiren yöntemler ile kullanıcıların dillerini bilmedikleri sayısal dokümanları istedikleri dile otomatik olarak dönüştürebilecek bir yapı tasarlanmıştır. Daha kullanışlı bir sistem tasarımı yapılarak platform tüm dünya dillerini destekleyebilecek şu anda 64 dil desteği veren bir platforma dönüştürülmüştür. Geliştirilen sistemde kullanılan üç yöntem ile farklı teknolojiler kullanılarak geliştirilen sistemin içerik tanımlama başarısı oldukça arttırılmıştır. Bu teknolojiler bir bütün olarak entegre olduklarında önerilen alfabetik kesişim ve birleşim yöntemleri ışığında yazılım tasarım ve test süreçleri gerçekleştirilmiştir. Bu yazılım sayesinde internet ortamında sıklıkla karşımıza çıkan ve literatürde geniş kabul görmüş olan Microsoft Word ve Adobe Acrobat dosya formatları ile HTML için dil tanımlama işlemleri başarıyla yapılabilmektedir. Bu sistem kendi başına çalışır (standalone) olduğu için herhangi bir hizmet sağlayıcı sunucu veya bant genişliği kapasitesi gibi kısıtlar ortadan kaldırılmıştır. Bu çalışmada kullanılan özel bileşenler sayesinde web içeriklerinin analiz edilmesinde yaşanan betik dil analizi gibi güçlüklerde ortadan kaldırılmıştır.

Bu çalışmada yapılan web sitesi ve doküman tanıma testlerinde sistemin %95-%100 arasında başarı oranlarında kararlı ve kesin sonuçlar verdiği görülmüştür. Bazı yanlış sonuç veren testlerde coğrafi olarak yakın olmayan bölgelerin dillerinin birbirlerine yakınsadıkları gösterilmiştir. Bu dillerin kökeni ve dağılımları için araştırmalara kaynak olabilecek bir çalışmadır. Sistemin henüz kendi kendine öğrenme sürecini bulunmadığından ölçeklenebilir veritabanı ve ağ yapısının sonraki çalışmalarda kullanıcılardan gelen geri beslemeler doğrultusunda güncellemesi gerekebileceği değerlendirilmektedir.Bu çalışmada önerilen sistemin en büyük avantajı dil tanıma sistemi için kullanılan yöntemin dil grubuna özgü olmayışı ve diğer diller içinde sistemin dönüştürülme sürecinin esnekliğidir. Dil tanıma veritabanı için önerilen kesişim ve birleşim yöntemlerinin tanınması istenilen dillere ait alfabelerde tekrarlanan harf sayıları hedef alınmaktadır. Bu sebeple tanınması istenilen dil sayısındaki artış, giriş olarak uygulanan karakter frekans analizinin yapıldığı ortak alfabe kümesinde Latin alfabesindeki dillerin birbirlerine yakınsadığı coğrafyalarda tekrarlanan harf sayıları artacağından minimal düzeyde artışlara, birbirlerinden uzaklaştığı yerlerde ise tekrarlanma azalacağı için ortak alfabe kümesinde azalmalara neden olacaktır. Bu sayede sistemin hızlı çalışması, küçük veritabanı gereksinimi ve ölçeklenebilirliği ile yüksek kararlılık gerektiren endüstriyel uygulamalarda kullanılabileceği düşünülmektedir. Gerçekleştirilen sistemin test sonuçları, diğer çalışmalarla karşılaştırıldığında oldukça başarılı sonuçlar ürettiği görülmüştür.

Çizelge 4'de çalışmanın Microsoft Word dokümanları için ortalama başarı oranları, Çizelge 5'de çalışmanın Adobe PDF dokümanları için ortalama başarı oranları, Çizelge 6'da çalışmanın Web Sitesi HTML dokümanları için ortalama başarı oranları verilmiştir. Elde edilen bu başarının sonucunda LM algoritmasıyla eğitilmiş sistemin, zeki dil tanıyıcı tasarımında uygulanabilir olduğunu göstermiştir. Gerçekleştirilen çalışma literatüre, istatistiksel dil tanıma metotlarının yanında farklı bakış açısı kazandırmaktadır. Literatürde de belirtildiği gibi zeki yaklaşımların DT tasarımında kullanılmasının doğru bir tercih olduğu, farklı yapılar seçerken MLP yapısının LM gibi güçlü algoritmalarla eğitilmesiyle YSA'ların DT tasarımında başarılarını arttırdığı tespit edilmiştir. Bu çalışma sonucunda geliştirilen sistem ile kullanıcılardan Word ve PDF dokümanları ile web sayfalarını giriş olarak aldıktan sonra dil tespiti sonucunda kullanıcının dönüştürmek istediği dile yapılan tercüme yine yazılım üzerinde bulunan dahili web göstericisi arabiriminde sunulmaktadır. Bu nedenle kullanıcılar farklı dillerde yazılmış dokümanları istedikleri dile başka hiçbir araç kullanmadan zeki yöntemler ile dönüştürebilmektedirler. Bu çalışma kapsamında önceki dönem çalışmalarına göre yeni bir yöntem önerilerek dil sınıflandırıcının çözüm uzayı minimuma indirilmiştir. Çözüm uzayının küçültülmesi ve işlem yükünün azalmasıyla aynı sonuçların elde edilmesi önerilen yöntemin etkin bir algoritma tasarımı içerdiğini göstermektedir. Ayrıca, web tabanlı dil işleme sistemlerinde eğitim ve test metodolojileri için internet üzerinden arama yöntemlerine ek olarak kelime bazlı doküman sağlama yöntemi de önerilmiştir. Bu çalışmanın internetin daha verimli kullanılmasına, farklı dil ve kültürlerde yapılan çalışmaların kolaylıkla okunup öğrenilmesine ve internette karşılaşılan pek çok problemin çözümüne büyük katkılar sağlayacağı değerlendirilmektedir.

Bu sistem sayesinde, internetin geniş kitlelere ve halk gruplarına yayıldığı bir dönemde küreselleşme ile ortaya çıkan ana dil yayınlarının başka milletler tarafından takip edilmesinin önünü açacağı değerlendirilmektedir.Bu çalışma kapsamında;

1. Bu çalışmada her dil için anahtar kelimeleri barındıran 20 doküman kullanılmıştır. Platform başarımını yükseltmek için farklı konularda ve içeriklerde sayısal dokümanlar eğitime eklenmesi,
2. Hatalı sonuçların eğitime dahil edilerek sistemin iyileştirilmesi,
3. DOC, PDF ve HTML ayrıştırma ile metin çıkartma mekanizmaları iyileştirilmesi,
4. Resim formatında olan dokümanlar için OCR (Optik Karakter Tanıma) parçasının geliştirilen platforma entegre edilmesi,
5. Önerilen tespit yöntemlerinin doküman tipine ve dil grubuna olan uygunluklarının incelenmesi,
6. Tasarlanan XML (Extended Markup Language) ortak arabirim dil süzgeçlerinin her dil için yapılandırılması,

gibi çalışmalarında yapılmasıyla sunulan sistemin başarısını arttıracağı değerlendirilmektedir. Sonuç olarak,

- Literatüre, istatistiksel dil tanıma metotlarının yanında kolayca ölçeklenebilir farklı bir bakış açısı kazandırdığı,
- Geliştirilen sistem ile kullanıcılardan HTML, Word ve PDF dokümanları ile web sayfalarını giriş olarak aldıktan sonra dil tespiti sonucunda kullanıcının dönüştürmek istediği dile yapılan tercümenin yine yazılım üzerinde bulunan dahili web göstericisi arabiriminde sunularak kullanıcı dostu doküman tanımlama ve çeviri sisteminin oluşturulduğu,
- İnternetin daha verimli kullanılmasına, farklı dil ve kültürlerde yapılan çalışmaların kolaylıkla okunup öğrenilmesine ve internette karşılaşılan pek çok problemin çözümüne büyük katkılar sağlayacağı değerlendirilmektedir.

Çizelge 4. A GENIUS platformu Microsoft Word ortalama başarı sonuçları [24]

Format	Dosya Tipi "WORD"											
Yöntem	#K	#B	#K	#B	#K	#B	#K	#B	#K	#B	Tanıma Başarı Ortalaması	En Yüksek Değere göre Sonuç Başarı Oranı
Almanca	92%	83%	98%	83%	69%	98%	14%	93%	73%	12%	52%	100%
Arnavutça	62%	90%	93%	99%	99%	99%	98%	99%	0%	99%	84%	100%
Fransızca	98%	89%	87%	83%	88%	95%	76%	75%	97%	95%	88%	100%
Galce	82%	96%	99%	96%	99%	96%	96%	89%	22%	84%	86%	100%
Hırvatça	97%	94%	97%	93%	98%	99%	84%	99%	95%	89%	95%	100%
İngilizce	98%	99%	95%	97%	95%	99%	86%	94%	95%	99%	96%	100%
İrlandaca	98%	99%	98%	96%	99%	99%	99%	99%	97%	97%	98%	100%
İspanyolca	99%	97%	98%	92%	99%	98%	94%	99%	96%	98%	97%	100%
İtalyanca	96%	89%	73%	99%	94%	99%	93%	98%	94%	99%	93%	100%
Letonca	95%	99%	97%	99%	84%	99%	94%	99%	96%	99%	96%	100%
Macarca	96%	97%	97%	96%	92%	94%	99%	98%	98%	98%	97%	100%
Maltaca	91%	99%	83%	99%	99%	99%	98%	99%	99%	97%	96%	100%
Portekizce	95%	99%	61%	99%	97%	98%	90%	98%	92%	98%	93%	100%
Türkçe	87%	99%	99%	99%	99%	99%	99%	99%	99%	99%	98%	100%
Vietnamca	99%	98%	98%	59%	99%	99%	65%	84%	97%	88%	89%	100%

Çizelge 5. GENIUS platformu Adobe PDF ortalama başarı sonuçları [24]

Format	Dosya Tipi "PDF"											
Yöntem	#K	#B	#K	#B	#K	#B	#K	#B	#K	#B	Tanıma Başarı Ortalaması	En Yüksek Değere göre Sonuç Başarı Oranı
Almanca	77%	91%	96%	93%	96%	95%	96%	77%	97%	96%	72%	100%
Arnavutça	99%	99%	91%	74%	100%	100%	99%	99%	88%	52%	90%	100%
Fransızca	87%	92%	0%	72%	97%	68%	68%	99%	68%	85%	74%	100%
Galce	66%	95%	100%	100%	100%	100%	97%	94%	97%	98%	95%	100%
Hırvatça	100%	100%	97%	99%	98%	94%	35%	98%	100%	100%	92%	100%
İngilizce	99%	97%	87%	98%	96%	91%	95%	97%	96%	70%	93%	100%
İrlandaca	87%	96%	99%	87%	99%	97%	78%	98%	95%	99%	94%	100%
İspanyolca	94%	88%	95%	99%	79%	95%	74%	98%	87%	88%	90%	100%
İtalyanca	98%	99%	79%	99%	76%	97%	93%	98%	85%	99%	92%	100%
Letonca	95%	89%	90%	81%	94%	99%	83%	99%	96%	99%	93%	100%
Macarca	99%	97%	99%	98%	94%	99%	94%	98%	97%	99%	97%	100%
Maltaca	100%	100%	90%	98%	95%	98%	88%	98%	92%	99%	96%	100%
Portekizce	91%	99%	95%	98%	94%	99%	89%	98%	95%	99%	96%	100%
Türkçe	99%	99%	99%	99%	99%	99%	99%	99%	99%	99%	99%	100%
Vietnamca	99%	97%	98%	99%	98%	97%	99%	95%	99%	99%	98%	100%

#K: Kesişim Yöntemi #B: Birleşim Yöntemi

Çizelge 6. GENIUS platformu web sitesi HTML ortalama başarı sonuçları [17, 25]

Format	Web Sitesi		Başarı Ortalaması	
Yöntem	Kesişim	Bileşim	Tanıma Başarı Ortalaması	En Yüksek Değere göre Sonuç Başarı Oranı
Almanca	93%	92%	93%	100%
Arnavutça	99%	99%	99%	100%
Fransızca	91%	75%	83%	100%
Galce	95%	93%	94%	100%
Hırvatça	55%	95%	75%	100%
İngilizce	99%	86%	93%	100%
İrlandaca	99%	96%	98%	100%
İspanyolca	97%	98%	98%	100%
İtalyanca	98%	97%	98%	100%
Letonca	95%	99%	97%	100%
Macarca	95%	97%	96%	100%
Maltaca	99%	99%	99%	100%
Portekizce	99%	96%	98%	100%
Türkçe	98%	99%	99%	100%
Vietnamca	98%	99%	99%	100%

Kaynaklar

1. Padro M., Padro L., "Comparing Methods for Language Identification" *Procesamiento del Lenguaje Natural*, Barcelona, 33-35 (2004).
2. Botha G.R., Zimu V.Z., Barnard E., "Text-based language identification for the South African languages", *SAIEE Africa Research Journal*, Cape Town, 141-146 (2007).
3. El-Shishiny H., Troussov A., McCloskey DJ., Takeuchi M., Nevidomsky A., Volkov P., "Word Fragments Based Arabic Language Identification", *NEMLAR Conference on Arabic Language Resources and Tools*, Mısır, 23-26 (2004).
4. Kruengkrai C., Srichaivattana P., Sornlertlamvanich V., Isahara H., "Language Identification Based on String Kernels" *Communications and Information Technology*, Pekin, 896-899 (2005).
5. Zavarsky P., Wada S., Mikami Y.,"Language and Encoding Scheme Identification of Extremely Large Sets of Multilingual Text Documents", *The 10th Machine Translation Summit*, Puket, 354-355 (2005).
6. Peng F., Schuurmans D., Wang S.,"Language and Task Independent Text Categorization with Simple Language Models", *North American Chapter of the Association for Computational Linguistics - Human Language Technologies*, Edmonton, 110-117 (2003).
7. Nair A.S., Nair V. V., Chandra V. S. S., "Hidden Markov Model Based Identification of Transliterated Regional Language Words in Text Documents", *Twentieth International Joint Conference on Artificial Intelligence*, Haydarabad, 87-91 (2007).
8. Ahmed B., Cha S-H.,, Tappert C.,"Language Identification from Text Using N-gram Based Cumulative Frequency Addition", *Student/Faculty Research Day,* New York, 121-128 (2004).
9. Constable P.G., "Toward a Model for Language Identification", *Summer Institute of Linguistics International Working Papers", D*ublin (2002).
10. Adams G., Resnik P., "A Language Identification Application Built on the Java Client/Server Platform", *The European Chapter of the Association of Computaional Linguistics Workshop*, İspanya (1997).
11. Ölveck T., "N-Gram based Statistics Aimed. at Language Identification", *Student Research Conference in Informatics and Information Technologies*, Brastilava, 1-7 (2005).
12. Bilcu, E.B., Astola J., "A Hybrid Neural Network for Language Identification from Text", *Machine. Learning for Signal Processing Conference*, Maynooth, 253-258 (2006).
13. Liu Y-H., Chang F., Lin C-C., "Language Identification of Character Images Using Machine Learning Techniques", *International Conference on Document Analysis and Recognition*, Seul, 630-634 (2005).
14. Zhu G., Yu X.,Li Y., Doermann D., "Unconstrained Language Identification Using A Shape Codebook", *The 11th International Conference on Frontiers in Handwritting Recognition*, Montreal, 13-18 (2008).
15. Baykan E., Henzinger M., Weber I., "Web Page Language Identification Based on URLs", *International Conference on Very Large Data Bases*, Auckland, 176-187 (2008).
16. Sağıroğlu, Ş., Beşdok, E., Erler, M., "Mühendislikte Yapay Zeka Uygulamaları-1:Yapay Sinir Ağları", *Ufuk Kitabevi*, Kayseri, 10-100 (2003).
17. Sağıroğlu Ş., Yavanoğlu U., Güven E.N., "Web Based Machine Learning for Language Identification and Translation" *International Conference on Machine Learning and Applications*, Ohio, 280-285 (2007).
18. Takçı H., Soğukpınar İ. "Letter Based Text Scoring Method for Language Identification", Springer Lecture Notes in Computer Science, Vol. 3261/2005 283-290 (2004).
19. İnternet : Google Yazılım "Web Tabanlı Dil Çeviri Aracı Web Sayfası" http://translate.google.com/ (2011)
20. İnternet : Microsoft Yazılım "Visual Studio 2005 C# Windows Form Uygulaması Yazılım Geliştirme Aracı", http://msdn.microsoft.com/en-us/vstudio/default.aspx (2005).
21. İnternet : Mathworks Yazılım "Matlab R2007B Deployment Tool, Dinamik bağlantı Kütüphanesi Geliştirme Aracı", http://www.mathworks.com/products/new_products/release2007b.html (2007).
22. İnternet : Cellbi Yazılım "Microsoft Word OLE Otomasyon Bileşeni" http://www.cellbi.com/products/docframework.aspx (2008).
23. İnternet : Apache Yazılım "Adobe Reader PDF Otomasyon Bileşeni" http://incubator.apache.org/pdfbox/ (2008).
24. U. Yavanoğlu, "Web Tabanlı Otomatik Dil Tanıma ve Çeviri Sistemi Geliştirilmesi", Gazi Üniversitesi Fen Bilimleri Enstitüsü Yüksek Lisans Tezi, 2009.
25. U. Yavanoğlu ve Ş. Sağıroğlu, "Web Tabanlı Otomatik Dil Tanıma ve Çevirme Sistemi", Gazi Üniversitesi Mühendislik-Mimarlık Fakültesi Dergisi, Cilt:25, No:3, s.484-494, 2010.
26. H.P Combrinck and E.C. Botha, "Text-Based Automatic Language Identification", Proceedings of the 6th Annual Symposium of the Pattern Recognition Association of South Africa, Gauteng, South-Africa, November, 1995.
27. Patent: Web ortamında bulunan dokümanların yazı dilinin otomatik olarak tespiti ve içeriğin gerçek zamanlı olarak dönüştürülmesi yöntemi ve sistemi, Türk Patent Enstitüsü, Başvuru No: 2010/00137.