STİLİSTİK ÖZELLİKLER KULLANILARAK YAZAR TANIMA İŞİNDE YAPAY SİNİR AĞLARININ BAŞARIMININ DEĞERLENDİRİLMESİ: TÜRKÇE KÖŞE YAZILARI

PERFORMANCE ASSESSMENT OF ARTIFICIAL NEURAL NETWORKS FOR AUTHOR ATTRIBUTION BY USING STYLISTIC FEATURES: TURKISH ARTICLES

ÖZLEM YAVANOĞLU

DOÇ. DR. EBRU SEZER

Tez Danışmanı

Hacettepe Üniversitesi
Lisansüstü Eğitim - Öğretim ve Sınav Yönetmeliğinin
Bilgisayar Mühendisliği Anabilim Dalı için Öngördüğü
YÜKSEK LİSANS TEZİ olarak hazırlanmıştır.

2017

Özlem YAVANOĞLU'nun hazırladığı "**Sitilistik Özellikler Kullanılarak Yazar Tanıma İşinde Yapay Sinir Ağlarının Başarımının Değerlendirilmesi: Türkçe Köşe Yazıları**" adlı bu çalışma aşağıdaki jüri tarafından **BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI**'nda **YÜKSEK LİSANS TEZİ** olarak kabul edilmiştir.

Prof. Dr. Şeref SAĞIROĞLU
Başkan .....

Doç. Dr. Ebru A. SEZER
Danışman .....

Doç. Dr. Suat ÖZDEMİR
Üye .....

Yrd. Doç. Dr. Gönenç ERCAN
Üye .....

Yrd. Doç. Dr. Erhan MENGÜŞOĞLU
Üye .....

Bu tez Hacettepe Üniversitesi Fen Bilimleri Enstitüsü tarafından **YÜKSEK LİSANS TEZİ** olarak onaylanmıştır.

Prof. Dr. Menemşe GÜMÜŞDERELİOĞLU
Fen Bilimleri Enstitüsü Müdürü

YAYINLAMA VE FİKRİ MÜLKİYET HAKLARI BEYANI

Enstitü tarafından onaylanan lisansüstü tezimin/raporumun tamamını veya herhangi bir kısmını, basılı (kağıt) ve elektronik formatta arşivleme ve aşağıda verilen koşullarla kullanıma açma iznini Hacettepe üniversitesine verdiğimi bildiririm. Bu izinle Üniversiteye verilen kullanım hakları dışındaki tüm fikri mülkiyet haklarım bende kalacak, tezimin tamamının ya da bir bölümünün gelecekteki çalışmalarda (makale, kitap, lisans ve patent vb.) kullanım hakları bana ait olacaktır.

Tezin kendi orijinal çalışmam olduğunu, başkalarının haklarını ihlal etmediğimi ve tezimin tek yetkili sahibi olduğumu beyan ve taahhüt ederim. Tezimde yer alan telif hakkı bulunan ve sahiplerinden yazılı izin alınarak kullanması zorunlu metinlerin yazılı izin alarak kullandığımı ve istenildiğinde suretlerini Üniversiteye teslim etmeyi taahhüt ederim.

☐ **Tezimin/Raporumun tamamı dünya çapında erişime açılabilir ve bir kısmı veya tamamının fotokopisi alınabilir.**
(Bu seçenekle teziniz arama motorlarında indekslenebilecek, daha sonra tezinizin erişim statüsünün değiştirilmesini talep etseniz ve kütüphane bu talebinizi yerine getirse bile, tezinin arama motorlarının önbelleklerinde kalmaya devam edebilecektir.)

☒ **Tezimin/Raporumun 01/04/2020 tarihine kadar erişime açılmasını ve fotokopi alınmasını (İç Kapak, Özet, İçindekiler ve Kaynakça hariç) istemiyorum.**
(Bu sürenin sonunda uzatma için başvuruda bulunmadığım taktirde, tezimin/raporumun tamamı her yerden erişime açılabilir, kaynak gösterilmek şartıyla bir kısmı ve ya tamamının fotokopisi alınabilir)

☐ **Tezimin/Raporumun ..... tarihine kadar erişime açılmasını istemiyorum, ancak kaynak gösterilmek şartıyla bir kısmı veya tamamının fotokopisinin alınmasını onaylıyorum.**

☐ **Serbest Seçenek/Yazarın Seçimi**

Selem YAVANOĞLU
20/02/2017

(İmza)

Öğrencinin Adı Soyadı

Canım oğlum Isaac Atlas'a

ETİK

Hacettepe Üniversitesi Fen Bilimleri Enstitüsü, tez yazım kurallarına uygun olarak hazırladığım bu tez çalışmasında,

- tez içindeki bütün bilgi ve belgeleri akademik kurallar çerçevesinde elde ettiğimi,
- görsel, işitsel ve yazılı tüm bilgi ve sonuçları bilimsel ahlak kurallarına uygun olarak sunduğumu,
- başkalarının eserlerinden yararlanılması durumunda ilgili eserlere bilimsel normlara uygun olarak atıfta bulunduğumu,
- atıfta bulunduğum eserlerin tümünü kaynak olarak gösterdiğimi, kullanılan verilerde herhangi bir tahrifat yapmadığımı,
- ve bu tezin herhangi bir bölümünü bu üniversitede veya başka bir üniversitede başka bir tez çalışması olarak sunmadığımı

beyan ederim.

Ö. Mi

ÖZLEM YAVANOĞLU

ÖZET

STİLİSTİK ÖZELLİKLER KULLANILARAK YAZAR TANIMA İŞİNDE YAPAY SİNİR AĞLARININ BAŞARIMININ DEĞERLENDİRİLMESİ: TÜRKÇE KÖŞE YAZILARI

Özlem YAVANOĞLU

Yüksek Lisans, Bilgisayar Mühendisliği Bölümü

Tez Danışmanı: Doç. Dr. Ebru A. SEZER

Ocak 2017, 110 Sayfa

Günümüzde internetin sağladığı olanakların başında medya kaynaklarının hızlı olması, anonimliği ve her yerden erişiminin sağlanabilmesi gelmektedir. Bireylerin web siteleri, forumlar, e-postalar gibi ortamlarda gerçek kimliklerini kullanma zorunluluğu olmadığı için bu tür ortamlar iyi ya da kötü niyetle gerçek dışı kimlik kullanımı ve aynı zamanda gerçek olan ile olmayan arasında ayrım yapabilmeyi gerektirir. Bu bazen suçluların bazen fikri hakların bazen de basitçe isim benzerliklerinin çözümü içindir. Bir metnin incelenerek, kişilerin yazma alışkanlıkları ya da stilleri (biçimleri) analiz edilerek gerçek yazar(lar)ı hakkında bilgi sahibi olmamıza yazar tanıma çalışmaları yardımcı olmaktadır. Alanyazında yazar tanıma, yazarı belli olmayan ya da yazarından şüphe duyulan bir yazının yazarını belirleme işlemi olarak ifade edilmektedir. Geçmişten günümüze bu alanda farklı çalışmalar gerçekleştirilmiştir. Yazar tanıma, bir sınıflandırma problemi olarak ele alınmakta ve potansiyel şüpheliler grubundan en uygun yazarın belirlenmesi işlemi olarak ifade edilmektedir. Bu tez kapsamında Türkçe yazar tanıma çalışmalarında kullanılmak üzere köşe yazılarından oluşan geniş bir derlem oluşturulmuştur. Oluşturulan derlem siyaset, ekonomi, yaşam ve spor

alanlarında yazılar yazan 167 köşe yazarına ait olan köşe yazılarından oluşmaktadır. Elde edilen modellerin başarısının değerlendirilmesi için farklı testler gerçekleştirilmiştir. Yapılan bu testler sonucunda elde edilen doğruluk (accuracy) değerleri 99% ile 74% arasında değişiklik göstermektedir. Ayrıca tez kapsamı içerisinde yazar tanıma çalışmasında kullanılan yazar özelliklerinin metin türü belirlemede ki başarısı değerlendirilmiş ve metin türü tanıma için farklı bir model önerilmiştir. Önerilen model bir yazının (metnin) 'Yaşam', 'Siyaset' ya da 'Ekonomi' alanlarından hangisine ait olduğunu göstermektedir. Yapılan deneyler sonucunda önerilen YSA (Yapay Sinir Ağları) modelinin doğruluk (accuracy) değeri 88% ile 70% arasında bulunmuştur. Yazar tanıma çalışma alanı için önerilen modellerin başarısının değerlendirilmesi için gözetimli öğrenme algoritmalarından biri olan k en yakın komşu (KNN) tercih edilmiştir. KNN ile yapılan deneysel çalışmalar sonucunda elde edilen doğruluk değerleri yazar tanıma için 62% olarak bulunmuştur. Bu tez kapsamında ayrıca farklı ihtiyaçlara cevap vermek ve önerilen yazar tanıma ve yazı türü tanıma modellerinin kararlılığını göstermek için hem yazar hem de yazı türü tanıyan hibrid bir YSA modeli de önerilmiştir.

Anahtar Kelimeler: Yazar Analizi, Yazar Tanıma, Metin Sınıflandırma, Yapay Sinir Ağları, Metin Özellikleri

ABSTRACT

PERFORMANCE ASSESSMENT OF ARTIFICIAL NEURAL NETWORKS FOR AUTHOR ATTRIBUTION BY USING STYLISTIC FEATURES: TURKISH ARTICLES

Özlem YAVANOĞLU
M.Sc. Thesis, Department of Computer Engineering
Supervisor: Assoc. Prof. Dr. Ebru A. SEZER
January 2017, 110 pages

One of the main opportunities that the internet provides today is the rapidity of media resources, anonymity and accessibility from anywhere. Since individuals do not have to use their real identity in places like websites, forums, and e-mails, such places require good or bad intentions to distinguish between real and non-real identity use at the same time. This is sometimes for the solutions of crimes, sometimes for conceptional rights, sometimes for simply the name similarities. By examining a text that contains a crime element and by analyzing people's writing habits or styles (forms), author identificiation efforts help us to know about the true authors of those messages. In literature, author recognition is expressed as the process of determining the author of an article whose author is not known or whose author is suspected. Different works have been carried out on this field from day to day. Author recognition is considered as a classification problem and is expressed as the process of identifying the most appropriate author from the group of potential suspects. Within the scope of this thesis, it is aimed to develop author identification models in order to respond to different needs. Different tests have been carried out to assess the success of the models obtained. Accuracy values obtained from these tests vary between 99% and 74%. In addition,

the success of the author features used in the author recognition study in determining the text type is evaluated and a different model for text type recognition is proposed. The proposed model shows whether a text belongs to the fields of 'Life', 'Politics' or 'Economy'. The accuracy of the proposed ANN (Artificial Neural Networks) models are between 88% and 70%. In this thesis, we also propose a hybrid ANN model which recognizes both writer and writing type in order to answer different needs and show the determination of the recommended author recognition and writing type recognition models.

Key Words: Author Analysis, Author Identification, Author Attribution, Text Classification, Artifical Neural Networks.

TEŞEKKÜR

Yüksek lisans tezinin ortaya çıkması, olgunlaşması, tamamlanmasında ilminden ve tecrübelerinden her zaman yararlandığım ve eğitim hayatım boyunca verdiği her türlü destek ve emek için değerli danışmanım Sayın Doç. Dr. Ebru AKÇAPINAR SEZER'e sonsuz teşekkürlerimi sunarım.

Tez savunmam sırasındaki değerli yorumları ve önerileri sebebiyle jüri üyelerim Sayın Prof. Dr. Şeref SAĞIROĞLU'na, Sayın Doç. Dr. Suat ÖZDEMİR'e, Sayın Yrd. Doç. Dr. Gönenç ERCAN'a ve Sayın Yrd. Doç. Dr. Erhan MENGÜŞOĞLU'na teşekkürlerimi sunarım.

Konu hakkında çalışmalar yapan ve bilgi birikimiyle yardımlarını esirgemeyen Dr. Oğuz ASLANTÜRK'e ve Hacettepe Üniversitesi Bilgisayar Mühendisliği bölüm başkanımız Sayın Prof. Dr. Mehmet Önder EFE olmak üzere yüksek lisans eğitimimi aldığım Hacettepe Üniversitesi Bilgisayar Mühendisliği'nin tüm akademik ve idari çalışanlarına teşekkür ederim.

Beni bugünlere getiren ve tüm eğitim hayatım boyunca sevgi ve desteklerini hiç bir zaman esirgemeyen canım aileme, hayatımın her aşamasında yanı başımda olan biricik kardeşim Özgür MİLLETSEVER ve tez çalışmam boyunca bana gösterdiği destek için sevgili eşim Uraz YAVANOĞLU'na sonsuz teşekkür ediyorum.

İÇİNDEKİLER

ÇİZELGELER

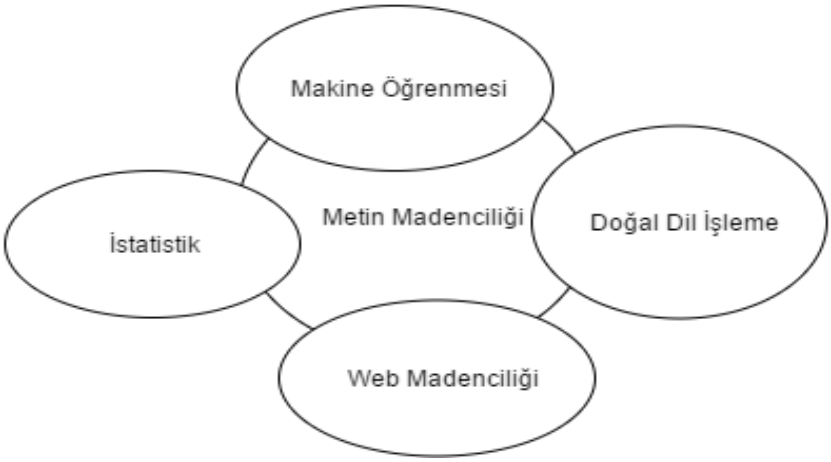
ŞEKİLLER

SİMGELER VE KISALTMALAR

ANN	Artifical Neural Networks (Yapay Sinir Ağları)
FP	False Positive (Yanlış Pozitif)
FN	False Negative (Yanlış Negatif)
k-NN	k-Nearest Neighbour (k En Yakın Komşu)
KNNDW	k-Nearest Neighbour with Distance Weighted
LDA	Linear Discriminant Analysis (Lineer Diskirimant Analizi)
ML	Machine Learning (Makine Öğrenimi)
MLP	Multi Layer Perceptron (Çok Katlı Algılayıcılar)
NB	Naïve Bayes
NLP	Natural Language Proccesing (Doğal Dil İşleme)
PCA	Principal Component Analysis (Temel Bileşen Analizi)
PCFG	Probabilistic Context-Free Grammars
RF	Random Forest (Rastgele Orman)
RMSE	Root Mean Square Error (Karesel Hata Ortalamalarının Karekökü)
RMS	Root Mean Square (Karekök Ortalama)
ROC	Receiver Operating Characteristic (Alıcı İşletim Karakteristiği)
SSE	Sum of Square Error (Kare Hata Toplamı)
SVM	Support Vektör Machine (Destek Vektör Makinesi)
TP	True Positive (Doğru Pozitif)
TN	True Negative (Doğru Negatif)
WEKA	Waikato Environment for Knowledge Analysis

1. GİRİŞ

Teknolojide yaşanan gelişmelerle birlikte özellikle Web 2.0 teknolojilerinden sonra tüm dünyada internet kullanımı hızlı bir şekilde yaygınlaşmıştır. Tüm bu gelişmelere bağlı olarak elektronik ortamlardaki veri miktarında da hızlı bir artış gözlenmiştir. Bu verilerin birçoğu yapısal olmayan verilerden oluşmaktadır. Yapısal olmayan veriler incelendiği zaman fark edilmektedir ki; bu verilerin büyük bir kısmı haberler, makaleler, araştırma bildirileri, kitaplar, sayısal kütüphaneler, e-posta iletileri ve web sayfaları gibi metinlerden oluşmaktadır. Bu metinler içerisinden anlamlı, daha önce bilinmeyen bilgilerin çıkarılması ya da metinlerin analiz edilmesi önemli bir problem haline gelmiştir [1]. Bu problem metin madenciliği çalışma alanın doğmasına sebep olmuştur. Metin madenciliği, metin şeklinde bulunan veriden daha önce bilinmeyen bilginin bilgi işlem yöntemleri kullanılarak çıkarılması olarak tanımlanabilmektedir [2]. Veri olarak metinleri kabul eden bu çalışma alanı veri madenciliğinin bir alt dalı olarak kabul edilmektedir [3]. Metin madenciliği, doğal dil işleme, yapay zekâ, istatistik, bilgi erişim (IR) gibi birçok yöntemden yararlanmaktadır. Bu yöntemlerden biri olan Doğal Dil İşleme (NLP) dünya üzerinde var olan dillerin işlenmesi ve kullanılması amacıyla çalışmalarda bulunan bir alandır. Bu dillerin işlenmesi için ilk olarak matematiksel modellerinin çıkarılmasına ihtiyaç vardır. Matematiksel modellerin elde edilmesi beraberinde metinlerin otomatik olarak çevrilmesi, otomatik konuşma ve komut anlama, metin özetleme gibi çeşitli yararları beraberinde sağlamaktadır. Metin madenciliği, metin özetleme, sınıflandırma, aynı konu hakkında yazılan metinleri bulma, birbiriyle bağlantılı olan metinlerin keşfedilmesi, konuşma tanıma, otomatik çeviri, duygu analizi, metinlerin türleri tanıma gibi birçok önemli çalışma alanına sahiptir. Veri madenciliğinin bir alt çalışma alanı olan metin madenciliği verilerin toplanması, verilerin temizlenmesi, veriler üzerinden özellik çıkarılması, makine öğrenmesi algoritmalarının kullanılması ve elde edilen anlamlı sonuçların yeniden kullanılması olmak üzere beş adımdan oluşmaktadır.



Şekil 1.1. Metin Madenciliği ve İlişkili Alanlar [2]

İnternet ortamında bulunan verilere daha etkili ve hızlı biçimde erişim sağlamak için metin madenciliği çalışma alanlarından biri olan metin sınıflandırma çalışmaları yapılmaktadır. Metin sınıflandırma, metinlerin daha önceden belirlenmiş bir alana göre sınıflandırılması işlemi olarak tanımlanabilir [3]. Yüz elli yıllık geçmişe sahip olan yazar tanıma çalışma alanı da metin sınıflandırma problemi olarak ele alınmaktadır [3]. Yazar tanıma, yazarı belli olmayan ya da yazarından şüphe duyulan metinlerin yazarlarının belirlemesi problemi olarak tanımlanmaktadır [4-7].

İnternetin kullanıcılara sağladığı anonimlik sayesinde internet ortamında bulunan bir kişi gerçekten söylediği kişi olabilir ya da farklı bir kişiymiş gibi de davranabilir. Örnekle açıklamak gerekirse dünyaca ünlü çevrimiçi (online) sanal para ticaret sistemi BitCoin en iyi örneklerden biri olabilir. Günümüzde birçok kişi Bitcoin kullanmaktadır. İnternetin sağladığı anonimlik sayesinde hiç kimse BitCoin'in gerçek yaratıcısının kim olduğunu bilmemektedir. Bu ve buna benzer durumlardan dolayı internet suçlar için ideal bir ortam sunmaktadır. Ayrıca, internetin hızlı bir şekilde gelişmesiyle birlikte günümüzde insanlar hem kişisel hem de iş hayatlarında interneti etkin şekilde kullanmaktadır. Günlük hayatımızın olmazsa olmazları arasına giren e-postalar, Facebook, Twitter yada Instagram gibi sosyal ağlar, blog (günlük), mesaj panoları, haber siteleri suçlular için ideal ortamlardır. Suçlular bu tür ortamları kullanarak gerçek kimliklerini saklayarak kişileri suistimal edebilirler. Son yıllarda bilişim suçları kapsamı altında incelenen e-posta dolandırıcılığı, çevrimiçi anonimlik suistimal gibi olaylarda artış gözlenmektedir. Ayrıca, internet devrimiyle birlikte telif hakkı ciddi bir sorun haline gelmektedir. Tüm bu

olaylar dikkate alınarak değerlendirildiği zaman yazar tanıma çalışmaları giderek daha önemli bir hal almaktadır. Uzun bir geçmişe sahip olan bu çalışma alanında onlarca çalışma gerçekleştirilmiştir [2-26]. Yapılan ilk çalışmalarda istatiksel yöntemler kullanılmasına rağmen son yıllarda yapılan çalışmalarda makine öğrenmesi yöntemleri sıklıkla kullanılmaktadır [4-40]. Ayrıca ilk çalışmalar bireylerin yoğun çalışmaları sonucunda ortaya çıkarken günümüzde bilgisayarların yüksek işlem kapasitesinden çalışmalarda sıklıkla yararlanılmaktadır. Yapılan çalışmalar incelendiğinde İngilizce ve Çince başta olmak üzere Arapça, Yunanca, Portekizce gibi farklı dillerde birçok çalışmaya rastlanmaktadır [3,4,10,27,28]. Türkçe dili için yapılan çalışmaların oldukça az olduğu fark edilmektedir [28,38,39,40]. Bundan dolayı tez kapsamında yapılan çalışma içerisinde kullanılan dil olarak Türkçe dili tercih edilmiştir. Farklı dillerde yapılan çalışmalarda yazar tanıma araştırma konusunda kullanılan çeşitli derlemlere rastlanmıştır. Bu derlemlerden bazıları; CCAT (Corporate/Industri), Enron e-posta, CHE (Chronicle of Higher Education)'dır. Türkçe dilinde yapılan çalışmalar incelendiği zaman geniş bir derleme rastlanamamıştır. Bu tez kapsamı içerisinde Türkçe dilinde geniş bir derlem oluşturalarak bu eksikliğin giderilmesi sağlanmıştır. Hürriyet, Sabah, Vatan, Milliyet, Cumhuriyet ve Posta gazetelerinde 1997/2014 tarihleri arasında köşe yazıları yazan yazarların yazıları geliştirilen yazılımlar sayesinde toplanmış ve Türkçe yazar tanıma çalışmalarında kullanılabilecek kapsamlı bir derlem elde edilmiştir.

Yazar tanıma çalışmaları içerisinde çeşitli özellikler kullanılmasına rağmen en başarılı özellik ya da özellik gruplarının neler olduğu hakkında fikir birliğine rastlanmamıştır. Bu tez içerisinde daha önce Aslantürk tarafından kullanılan Türkçe dili için elde edilen başarısı bilinen sözdizimsel ve sözcüksel özellikler kullanılmıştır [40]. Yazar tanıma çalışmasının gerçekleştirilmesi için veri olarak çalışma yapmak isteyenlerin erişiminde sorun yaşamayacağı çevrim içi gazetelerin köşe yazıları tercih edilmiştir. Köşe yazılarının tercih edilme sebeplerinden bir diğeri ise yazarların düşüncelerini dile getirirken herhangi bir kısıtlama içerisinde olmamasıdır. Elde edilen köşe yazıları 'yaşam, 'siyaset' ve 'ekonomi' olmak üzere üç alana ayrılmıştır.

Ayrıca yazar tanıma çalışmaların içerisinde kullanılan makine öğrenmesi yöntemlerinden biri olan Yapay Sinir Ağları (YSA) kullanılmıştır [11-21]. Bu tez kapsamında yazar tanıma probleminin farklı sorunlarına çözümler getirmek için çeşitli

YSA modelleri önerilmektedir. İlk olarak yazarı tartışmalı olan bir metnin gerçekten o kişi tarafından yazılıp yazılmadığının belirlenmesi için YSA modelleri önerilmektedir. Ayrıca yazarı belli olmayan bir yazının gerçekten sahibinin bulunması için farklı bir YSA modeli önerilmektedir. Önerilen bu model yazarı belli olmayan metnin yazarını belirlemek için sistemde bulunan altı yazara ait olup olmadığına değerlendirir ve bu yazarlara ait olmaması durumunda farklı bir yazara ait olduğunu ifade etmektedir. Bunun için YSA'ların genelleme özelliğinden yararlanılmaktadır.

Ayrıca tez içerisinde kullanılan biçimsel ve sözcüksel özelliklerin yazar türü belirlemedeki başarısını değerlendirilerek bir metnin yazı türü (alanı) belirlemek için farklı bir YSA modeli önerilmektedir. Önerilen bu model 'yaşam', 'siyaset' ve 'ekonomi' alanlarındaki yazıları tanıyabilme yeteneğine sahiptir. Son olarak yapılan alanyazın çalışmaları içerisinde yazar tanıma ve yazı türü belirleme için çeşitli çalışmanın yapıldığı fark edilmektedir. Yapılan bu çalışmalar içerisinde Türkçe dili için hem yazar hem de yazı türünü belirleyen hibrid bir modele rastlanamamıştır. Farklı ihtiyaçlara çözümler sunabilmek ve bir yazının hem yazarını hem de türü hakkında kullanıcılara bilgi verebilmek için hibrid bir model önerilmektedir. Önerilen YSA modeli Türkçe dili için ilk olma özelliğine sahiptir.

.

2. ALAN BİLGİSİ

2.1. Yazar Analizi

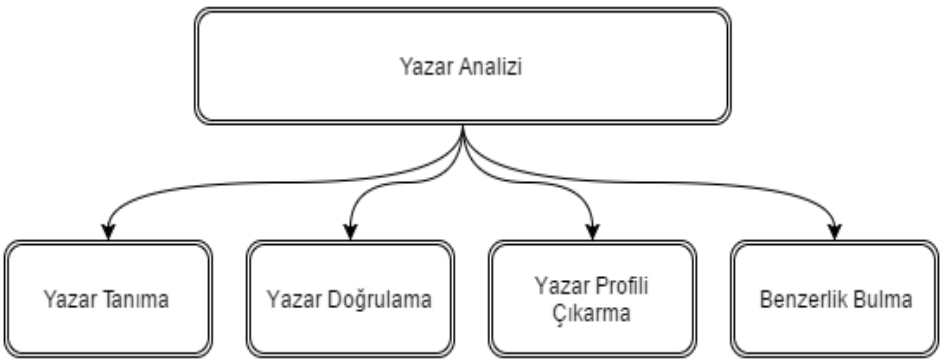
Yazarın eserlerinin incelenerek yazar hakkında bilgi edinme işlemi yazar analizi olarak ifade edilmektedir [29]. Yazar analizi çalışmaları stilometri adı verilen dilsel araştırma disiplinine dayanmaktadır. Stilometri alanında ilk yapılan çalışma 18. yüzyılda gerçekleştirilen Shakespeare'in bazı oyunlarının gerçekten kendisi tarafından yazılıp yazılmadığıyla ilgilidir [30]. 1851 yılında ise İngiliz mantıkçı Augustus de Morgan kişilerin daha uzun kelimeler kullanması yazar belirlemede kullanılabilir önermesini yapmıştır [31]. 1887 yılında ise Thomas Mendenhall tarafından yazarları tanımak için karakteristik eğri adı verilen bir yaklaşım önerilmiştir [32]. Bu çalışmadan bir yıl sonra, İngiliz matematikçi Benjamin Smith yazarların stillerini ayırt etmek için ortalama cümle uzunluklarına dayalı stil eğrisi adı verilen çalışmaları yapmıştır [33]. Profesör Lucius Sherman 1893 yılında yaptığı çalışmada ortalama cümle uzunluklarını kullanarak zamanla yazı stillerinin değişimi üzerine bir çalışma gerçekleştirmiştir [31]. Yapılan bu çalışmaları 20. yüzyılın ilk yarısında gerçekleştirilen istatiksel çalışmalar izlemiştir [34-36]. İngiliz istatistikçi olan Yule, istatiksel özellikleri kullanarak tartışmalı eserlerin yazarları üzerine çalışmalar yapmıştır [34]. Amerikalı dil bilimci Zipf, kelime frekanslarının ilk kez incelendiği bir çalışma gerçekleştirmiştir [35]. Yazar analizi alanında en etkili çalışmalardan birisi olarak kabul edilen "Federalist Papers" yazıları üzerine Mollester ve Wallace tarafından gerçekleştirilen bir çalışma yapılmıştır [36].

Geçmişten günümüze birçok çalışma yapılan yazar analizi alanında, 2013 yılında yazı analisti uzmanları Guguk Kuşu kitabının gerçek yazarının Robert Gabraith mahlasını kullanan J.K. Rowling olduğunu ortaya çıkardılar. Bu iki yazı analisti Rowling ve Robert Gabraith kelime uzunlukları dağılımlarının birbirine çok benzer olduğunu fark etmiştir [37]. Ayrıca, haklarında hüküm kararı çıkmış olan teröristlerin gerçekleştirdikleri saldırıların gerçekten onlara ait olduğu yazdıkları yazılarla doğrulanmıştır [38]. Unabomber (Üniversite ve Havayolları bombacısı) adı altında Washington Post ve New York Times gazetelerinde yayınlanan Unabomber Manifestosunun yazarı yapılan çalışmalar sonucunda Ted Kaczynski olarak bulunmuştur [38]. Dolandırıcılık, hile gibi suçlarlar ilgili bazı davalarda da yazı analizinden yararlanılmıştır. En son olarak, Paul Ceglia isimli Newyork'lu bir satıcı ile Mark Zuckberburg arasında geçen sözde e-posta

yazışmasında Facebook hisse senetlerinin yarısının kendisine ait olduğunun iddia edildiği yazışmanın gerçek olmadığı yazı analizi uzmanları tarafından ortaya çıkarılmıştır [39].

Uzun bir geçmişe sahip olan bu çalışma alanı, ünlü eserlerin ya da tartışmalı yazıların yazarlarının belirlenmesi çalışmaları yazarların birbirinden farklı ve benzersiz olan yazma özelliklerinden yararlanmıştır. Bir konu hakkında yazarlar fikirlerini ifade ederken kullandıkları dil bilgisi yapısı, kelime seçimleri gibi yapılar tamamen birbirinden farklıdır. Her yazarın kendine özgü yazma alışkanlığı bulunmaktadır. Yazma alışkanlığı olarak adlandırılan kelimelerin kullanımı, paragraf ve cümle uzunlukları, metin biçimi gibi özellikler kolayca değişmez ve her yazar için farklıdır [40]. Bu özelliklere yapılan çalışmalarda “yazar değişmezleri”, “siber parmak izi” olarak da rastlamak mümkündür [2].

Yazar analizi çalışmaları çeşitli kullanım alanlarına sahiptir. Alanyazın çalışmalarındaki metinlerin analiz edilmesi, bilgisayar program kodlarının analiz edilmesi, çevrim içi mesajların analiz edilmesi örnek olarak verilebilir. Bununla birlikte web ve sosyal ağların giderek büyümesinin sonucu olarak son yıllarda yapılan çalışmalarda forum, blog, elektronik postalar, tweet gibi yazılar üzerine çalışmalara odaklanılmıştır. Yazar analizi çalışmaları edebi, adli, akademik, ticari birçok alanda karşılığı olan bir alt problem çözümüdür. Yapılan çalışmalar dört başlık altında incelenmektedir ve Şekil 2.1’de alt disiplin dalları gösterilmiştir.



Şekil 2.1. Yazar Analizi Alt Disiplin Dalları [39].

Yazar Tanıma: Yazarı bilinmeyen ya da yazarından şüphe duyulan bir yazının yazarını bulma işlemidir. Bir başka tanımda, yazarı belli olmayan bir metinin yazarının olası yazarlar grubu arasından yazarının bulunması olarak ifade edilmektedir [40,41]. Yapılan çalışmalarda yazar tanıma sınıflandırma problemi olarak ele alınmıştır. Her yazar tanıma problemi, tüm aday yazarların bilinen yazılarını içeren eğitim derlemi ve yazarı bilinmeyen yazılardan oluşan test derleminden oluşmaktadır.

Yazar Doğrulama*:* Tek bir yazar tarafından yazılan yazılar göz önünde bulundurularak, yazarı tartışmalı olan bir yazının yazarının bu yazılar kullanılarak değerlendirilmesidir. Yani yazının gerçekten kime ait olduğunun bulunması işlemidir [42]. Yazar doğrulama çalışmalarının altında yazarın belli bir yazım stili vardır ve kısa zamanda bunu gizlemek zordur mantığı bulunmaktadır. Bu yazı stili kullanılarak yazının yazarının sorgulanan kişi tarafından kaleme alınıp alınmadığı belirlenebilir. Bu çalışma alanı çevrimiçi metinlerden oluşan çeşitli suç olaylarında oldukça yararlıdır [42-44].

Yazar Profili Çıkarma: Metnin yazarı hakkında yaş, cinsiyet, eğitim durumu gibi kişisel özelliklerinin çıkartılmasıdır. Yazarın demografik ve psikolojik özelliklerinden oluşmaktadır. Yapılan çalışmalarda yazarın cinsiyetini [45,46], yazarın yaşını [47,48], yazı dili [49], sinirsel davranışların (nörotik) seviyesi [50,51] gibi alanlar üzerinde durulmuştur.

Benzerlik Bulma: Verilen iki metnin birbirine ne kadar benzediği ya da benzemediğiyle ilgilidir. Benzerlik bulmak için çeşitli yaklaşımlar bulunmaktadır. Basit bir sınıflandırma işlemi olarak düşünüldüğü zaman mesafe esaslı, geometrik, yapısal özellik ve bilgi tabanlı yöntemlerle kategorize edilebilir [30,41].

2.2. Yazarlık Özellikleri

Kişilerin el yazıları kullanılarak yazarı belli olmayan bir yazının yazarını belirlemek mümkün iken günümüzde çevrimiçi ortamlarda yazar tanıma işlemini gerçekleştirmek için yazarın yazma alışkanlarından faydalanılmaktadır. Her yazarın kendine özgü yazma alışkanlığı bulunmaktadır [2-26]. Yazma alışkanlığı olarak adlandırılan kelimelerin kullanımı, paragraf ve cümle uzunlukları, metin biçimi gibi özellikler kolayca değişmez ve her yazar için farklılar göstermektedir [2-26]. Bu özelliklere yapılan çalışmalarda “yazarlık özellikleri”, “yazar değişmezleri”, “siber parmak izi” olarak da

rastlamak mümkündür [56]. Yazar tanıma alanında yapılan çalışmalar incelendiği zaman çok çeşitli özelliklerin kullanıldığı görülmektedir [1-65]. Yapılan bu çalışmalarda kullanılan özellikler sözcüksel özellikler, sözdizimsel özellikler, yapısal özellikler, içeriğe bağlı özellikler ve kişiye özgü özellikler olmak üzere beş gruba ayrılmıştır.

2.2.1. Sözcüksel Özellikler (Lexical Features)

Sözcüksel özellikler, karakter ve sözcük tabanlı özellikler olmak üzere ikiye ayrılmıştır. Karakter tabanlı özellikler, büyük harf sayısı, cümle başlarında kullanılan büyük harf sayısı, kelime başına ortalama karakter sayısı, cümle başına ortalama karakter sayısı gibi özelliklerden oluşmaktadır [2-24]. Sözcük tabanlı özellikler ise kelime uzunluğu dağılımları, cümle başına kelime sayısı, kelime zenginliği gibi özellikleri içermektedir. Farklı sözcüklerin sayısının toplam kelime sayısına oranı (type/token ratio), metin içerisinde yalnız bir kere geçen kelimelerin sayısı (hapax legomena), metin içerisinde yalnız iki defa geçen kelimelerin sayısı (hapax dislegomena) gibi özellikler kelime zenginliğiyle ilgilidir [6,59]. Bunların dışında aşağıdaki eşitliklerde verildiği üzere kelime zenginliğini ölçmek için çeşitli yaklaşımlar bulunmaktadır (farklı kelime sayısı X, toplam kelime sayısı Y olarak ifade edilmiştir). Bu yaklaşımlar Eşitlik 1 ile Eşitlik 7 arasında ifade edilmektedir.

$$\text{Guirad's R} = \text{X}/\sqrt{Y} \tag{1}$$

$$\text{Herdan's C} = \log_{10} X / \log_{10} Y \tag{2}$$

$$\text{Rubent's K} = \log_{10} X / \log_{10}(\log_{10} Y) \tag{3}$$

$$\text{Maas' A} = \sqrt{\log_{10} Y / \log_{10}(\log_{10} X)}/\log_{10} Y^2 \tag{4}$$

$$\text{Dugasts' U} = (\log_{10} Y)^2 \sqrt{\log_{10} Y / \log_{10} X} \tag{5}$$

$$\text{Sichel's S} = \text{Hapax Dislegomena Sayısı}/X \tag{6}$$

$$\text{L. Janenkov ve Neistoj} = 1/X^2 * \log_{10} Y \tag{7}$$

2.2.2. Sözdizimsel Özellikler (Syntactic Features)

İsim, sıfat, fiil, zarf, zamir, fiil, ünlem gibi sözcük türleri, nokta, virgül, noktalı virgül, tırnak, çift tırnak gibi noktalama işaretleri ve sözcük n-dizileri (word n-gram) gibi özellikleri içermektedir [1-15]. Ayrıca çok amaçlı işlevsel sözcüklerde bu özellik grubu altında incelenmektedir [11-16,40]. İşlevsel sözcüklerin özellik olarak kullanıldığı ilk çalışma Mosteller and Wallace tarafından "Federalist Paper" ele alınarak gerçekleştirilmiştir. Bu çalışmada işlevsel sözcüklerin etkinliğini gösterilmiştir [28]. Başka bir çalışmada ise 30-50 tane işlevsel sözcüğü yazar tanıma için kullanmıştır [28].

2.3.3. Yapısal Özellikler (Structural Features)

Daha önce bahsedilen sözdizimsel ve sözcüksel özellik grupları uygulamalardan bağımsız olan özellikleri içerirken bu özellik grubu uygulama özelliklerini yazar tanıma için dikkate almaktadır [58-60]. E-postalardaki selam, mesajlardaki veda ve girinti kullanımı gibi özellikleri içerir [58-60]. HTML formatındaki metinlerde, HTML etiketi dağıtımına ilişkin ölçümler, yazı tipi ve boyutu gibi özellikler bu grup içinde incelenmektedir [10,58].

Yapısal özellikler, yazarların metinlerinin yapısını ve düzenini nasıl organize ettiğini öğrenmemize yardımcı olmaktadır. Örneğin cümlelerin paragraflar içindeki yeri ve nasıl organize olduğu ya da metin içerisindeki paragrafların düzenlenme şekliyle ilgilidir. Başka bir önemli konu ise çok kısa metinler üzerinde yapılan yazar tanıma çalışmalarında sadece üslup özellikleri kullanılması metinleri temsil etmek için yeterli olmamaktadır. Bunun için özellik olarak hem üslup özellikleri hem de yapısal özellikler kullanılmaktadır [59].

2.3.4. İçeriğe Özgü Özellikler (Content-Specific Features)

Yazılan alan türüne bağlı olarak değişen ve o alan için diğer cümle ya da kelimelerden daha önemli olan özellikleri içermektedir [10,13, 62]. Belirli etkinlikleri, mesaj gruplarını, tartışma forumlarını kategorize etmek için kullanılmaktadır. Bilişim suçları içine giren spam, yemleme, fikri hak hırsızlığı yapan kişiler genellikle sexy (seksi), snow (kar), download (indirmek), click here (buraya tıkla) ve safe (güvenli) gibi kelimeleri kullanmaktadır. Ayrıca bir alan için oluşturulan kelime içerikleri (özellikleri) diğer bir alanda etkili değildir. Hatta aynı alan içinde kişiden kişiye de değişiklik göstermektedir.

Benzer şekilde, Zheng ve arkadaşları yazar analizi üzerinde yaptıkları bir çalışmada siber suçlar için 11 anahtar kelime (özellik) kullanmışlardır [62].

2.3.5 Kişiye Özgü Özellikler (Idiosyncratic Features)

Kişiye özgü olan özellikler, yazarlar tarafından gerçekleştirilen dil bilgisi (grammar) hataları, imla hataları gibi özelliklerden oluşmaktadır. Bu tür özelliklerin listesi kişiden kişiye göre farklılık göstermekte ve bunların kontrol edilmesi zor bir hale gelmektedir. Koppel ve arkadaşları bu özellikleri kullanan bir çalışma gerçekleştirmiştir [61]. Yazarların metinler üzerindeki bu tür hatalarını otomatik olarak tespit etmek için önce tüm veriler bir metin düzenleyici (text editor) olan MS-Word üzerinde açılmış ve spell checker (yazım denetleyicisi) kullanılmıştır. Yazım denetleyicisi tarafından bulunan hatalar önerilen en iyi öneri ile birlikte kaydedilmiştir. Fakat elde edilen sonuçlar yetersiz geldiği için yeni bir betik kodlama için kullanılmaktadır. Çalışma içerisinde kelime tekrarı, eksik kelime, uyumsuz zaman, uyumsuz tekil ve çoğul, tekrarlanan harfler, eksik tire gibi 99 özellik çıkartılmıştır [61].

2.4. Performans Değerlendirme Metrikleri

Geçmişten günümüze yazar tanıma alanında çeşitli çalışma yapılmıştır. Yapılan bu çalışmaların başarılarının nicel bir şekilde karşılaştırılması için performans değerlendirme ölçütleri kullanılmaktadır. Performans değerlendirme ölçütleri genellikle ikili sınıflandırma için oluşturulan hata matrisi kullanılarak elde edilmektedir (Çizelge 2.1). Pozitif sınıf etiketine sahip bir veri sınıflandırma sonucu pozitif olarak etiketlenmiş ise Doğru Pozitif (DP), negatif olarak etiketlendirilmiş ise Yanlış Negatif (YN) olarak, Negatif sınıf etiketine sahip bir veri sınıflandırma sonucu pozitif olarak etiketlendiyse Yanlış Pozitif (YP), negatif olarak etiketlenmiş ise (Doğru Negatif) olarak ifade edilir [65,66].

Alanyazında sınıflandırma performanslarını değerlendirmek için çeşitli performans değerlendirme ölçütleri bulunmaktadır.

Çizelge 2.1. Hata Matrisi

		Tahmin Edilen Sınıf	
		Pozitif	Negatif
Gerçek Sınıf	Pozitif	Doğru Pozitif (DP)	Yanlış Negatif (YN)
	Negatif	Yanlış Pozitif(YP)	Doğru Negatif (DN)

1. Doğruluk Oranı (Accuracy): Doğru etiketle sınıflandırılan pozitif ve negatif verilerin tüm verilere oranı olarak ifade edilmektedir.

$$Doğruluk\ Oranı = \frac{DP+DN}{DP+YP+DN+YN} \quad (8)$$

2. Hata Oranı (Error Rate): Yanlış etiketle sınıflandırılan pozitif ve negatif verilerin tüm verilere oranı olarak ifade edilmektedir (Eşitlik 2)

$$Hata\ Oranı = \frac{YP+YN}{DP+YP+DN+YN} \quad (9)$$

3. Duyarlılık (Precision) : Pozitif olarak etiketlenen pozitif verinin, tüm pozitif olarak etiketlenen verilere oranıdır (Eşitlik 3).

$$Duyarlılık\ (\pi) = \frac{DP}{DP+YP} \quad (10)$$

4. Anma (Recall): Pozitif olarak etiketlenen pozitif verinin, tüm pozitif verilere oranıdır.

$$Anma\ (\rho) = \frac{DP}{DP+YN} \quad (11)$$

5. F Ölçütü (F-Measure): Duyarlılık ve Anma performans değerlendirme ölçütlerinin harmonic ortalamasıdır.

$$F\ \text{Ölçütü} = \frac{2*\pi*\rho}{\pi+\rho} \quad (12)$$

6. Hassasiyet (Sensitivity): Pozitif olarak etiketlenen verinin, tüm pozitif verilere oranıdır. ROC eğrisi bu performans değerlendirme ölçütünün grafiksel olarak gösterimidir.

3. YAZAR TANIMA KONUSUNDA ALANYAZIN ÖZETİ

Yüz otuz yıllık geçmişe sahip olan yazar tanıma alanında geçmişten günümüze birçok çalışma yapılmıştır [1-60]. Bu alanda yapılan ilk çalışma olarak Mendenhall tarafından 1887 yılında yapılan araştırma kabul edilmektir [32]. Tez kapsamı içerisinde son yıllarda yapılan çalışmalar "İngilizce ve diğer diller için yapılan çalışmalar" ile "Türkçe için yapılan çalışmalar" olmak üzere iki gruba ayrılmış ve önemli görülen yönleriyle özetlenmeye çalışılmıştır. Yapılan çalışmalar özetlenirken kullanılan özellik grupları, veri setleri, sınıflandırma yöntemleri, performans değerlendirme ölçütleri ve elde edilen deneysel sonuçlar anlatılarak çalışma ile ilgili genel bilgi verilmek istenmiştir

3.1. İngilizce ve Diğer Diller için Yapılan Çalışmalar

Diederich ve arkadaşları 2003 yılında farklı özellik grupları kullanarak SVM sınıflandırma yönteminin yazar tanıma alanındaki başarısı üzerine bir çalışma yapmıştır. Kullanılan veri seti Alman gazetelerinde politika, ekonomi ve yerel konular hakkında yazılar yazan 7 yazara ait yazılardan oluşmaktadır. Çalışma içerisinde kullanılan özellikler kelime uzunlukları, kelime türleri, fonksiyonel kelimelerdir. Tüm deneylerde Thorsten Joachims tarafından geliştirilen SVM-light programı kullanılmıştır. Yapılan deney sonuçlarında elde edilen başarı oranı 60%-80%'dir. Ayrıca SVM yönteminin başarısını karşılaştırmak için MLP ve karar ağacı makine öğrenmesi yöntemleri kullanılmıştır. Yapılan karşılaştırma deneylerinde SVM sınıflandırıcı 100% hassasiyet değerine sahip iken karar ağacı sınıflandırma modelinden elde edilen hassasiyet değeri 22,7% dir [14].

Abbasi ve Chen 2005 yılında yaptıkları çalışmada militan grupların ya da terörist organizasyonlarının internet üzerinden yaptıkları mesajlaşmaları izleyerek, yapılması planlanan eylemleri önlemek için yazar analizi gerçekleştirmiştir. Cihatçı gruplarla ilişkilendirilmiş olan Arapça ve İngilizce web siteleri çalışma içerisinde kullanılmıştır. İngilizce veri seti Beyaz Şövalye adı verilen bir Amerikan menşeili forum sitesine ait olup politika, dinsel ve ırksal sorunlar ile ilgili yazılarından oluşmaktadır. Arapça veri seti ise, Filistin El Aksa Şehitleri ile ilişkili gruba ait yazıları içermektedir. Sözcüksel, sözdizimsel, yapısal ve içerik özgü özellik grupları olmak üzere toplam 301 özellik kullanılmıştır. Sınıflandırıcı yöntemi olarak ise SVM ve C4.5 kullanılmıştır. Farklı özellik

grupları ve bunların kombinasyonları kullanılarak yapılan deneylerde SVM sınıflandırıcısının doğruluk oranları 88% ile 97% iken C4.5 sınıflandırıcısının doğruluk oranları 61% ile 90% arasındadır. Yapılan deney sonuçları ayrıca tüm özellik gruplarının kullanılarak yapılan sınıflandırmanın en yüksek başarı oranını verdiğini göstermektedir [10].

Zheng ve arkadaşları 2006 yılında yaptıkları bir çalışmada gönderilen çevrim içi mesajın gerçekten sorgulanan kişi tarafından yazılıp yazılmadığının belirlenmesi için bir yaklaşım sunmuştur. Çalışma içerisinde kullanılan özellikler; sözcüksel, sözdizimsel, yapısal ve içerik özgü özellikler olmak üzere dört ana gruba ayrılmıştır. Önerilen yaklaşımın hem İngilizce hem de Çince dilinde gerçekleştirilmesi için otomatik özellik çıkarıcılar geliştirilmiştir. Çalışma içerisinde kullanılan veri setleri yirmi yazarın yaklaşık otuz ile kırk mesajından oluşmaktadır. Destek vektör makinası, sinir ağları ve C4.5 sınıflandırma yöntemleri kullanılarak çeşitli deneyler gerçekleştirilmiştir. Yapılan deneylerde özellik grupları arttırımsal olarak eklenerek özellik gruplarının başarısı incelenmek istenmiştir. Deney sonuçlarına göre sözcüksel özelliklerin tek başına elde ettikleri doğruluk oranları 52% ile 89% arasındadır. Tüm özellik grupları kullanılarak yapılan deneyde elde edilen doğruluk oranları 75% ile 96% arasında değişmektedir. Ayrıca kullanılan en başarılı sınıflandırma algoritması olarak SVM olarak belirlenmiştir [7].

Grieve tarafından 2007 yılında yapılan çalışmada daha önce birçok çalışmada kullanılmış olan 39 metin özelliklerin ya da özellik gruplarının yazar belirlemedeki başarıları açısından karşılaştırılması yapılmaktadır. Derlem olarak ‘Telegraph’ dergisinden yazılar yazan 40 yazara ait toplam 1600 yazı tercih edilmiştir. Bu yazarların seçiminde iyi eğitimli, İngiliz, Anglo-Saxon, orta yaşlı, muhafazakâr gibi aynı sosyal statü ve geçmişe sahip olmalarına dikkat edilmiştir. Ayrıca toplanan bu yazılar 2003-2005 tarihleri arasında yazılmıştır. Kelime ve cümle uzunlukları, kelime zenginliği, n gram, noktalama işaretleri gibi 39 metin özelliğinin başarıları ayrıntılı olarak incelenmiştir. Yapılan deneysel çalışmalar sonucunda elde edilen başarı değeri doğruluk (accuracy) olarak hesaplanmış ve 75% oranında başarı gözlenmiştir [15].

Ma ve arkadaşları 2008 yılında yaptıkları bir çalışma da adli bilişim olaylarında kullanılmak üzere Çince için yazar tanıma çalışması gerçekleştirmiştir. İngilizce ya da Hint-Avrupa dillerinin aksine, Çince kelimeler arasında kelimeleri birbirinden ayıran bir ayraç bulunmamaktadır. Bundan dolayı kelime bölütlemesi (segmentation) önemli bir sorun haline gelmektedir. Bu çalışma içerisinde sıralı desen madenciliği teknikleri kullanılmıştır. Destek Vektör Makinesi (SVM) sınıflandırıcı olarak tercih edilmiştir. Deneylerin gerçekleştirilmesi için kullanılan derlem, 3 yazar toplam 150 e-postadan oluşmaktadır. Performans değerlendirme ölçütlerinden duyarlılık (precision), anma değeri (recall), F1-ölçütü kullanılmıştır. Çalışmada minimum destek vektörü 10 olarak alındığı zaman F1 ölçüt değeri 90% olarak ifade edilmiştir [3].

Bradley ve arkadaşları 2008 yılında gerçekleştirdikleri bir çalışmada makale yazarlarının belirlenmesi için alıntılanma bilgisinden yararlanmışlardır. Çalışma içerisinde kullanılan derlemin oluşması için ilk olarak CiteSeer verisi indirilmiştir [66]. Bu veriseti içerisinde 716,772 makale bulunmaktadır. Çalışmada kullanılacak olan derlemin oluşması için alıntılanma sayısı 5'den büyük olan makaleler seçilmiştir. Bunun sonucunda 137.485 makale elde edilmiştir. Çalışmada yazar ya da yazarların belirlenmesi için, alıntılanma sıklığından ve belirli bir alt konuya dayalı makalelerde ve yapılarda alıntı grafiği oluşumundan yararlanılmıştır. İlk olarak Gizli Anlam Analizi (Latent Semantic Analysis) yaklaşımından yararlanılmıştır. Bu yaklaşım makalelerin verimli bir gösterimini sunarak yeni makaleye benzer makalelerin bulunması ya da arama terimleri ile ilgili makalelerin kolaylıkla bulunmasına olanak sağlamaktadır. Yöntem bir yazarın farklı alanlarda yazılar yazması ya da çok anlamlılık gibi özellikleri gözden kaçırılmasına neden olmaktadır. Diğer yaklaşımlar ise Latent Dirichlet Allocation, Karakteristik Vektör Modeli (Characteristic Vector Classifier) ve Rassal Yürüyüş Modeli (Random Walk Classifier)'dır. Her modelin avantajları ve dezavantajları çalışma içerisinde gözden geçirilmiştir. Gerçekleştirilen farklı sınıflandırma modellerinin başarısı (doğruluk değeri) 50%'nin üzerindedir [9].

Escalante ve arkadaşları 2008 yılında gerçekleştirdikleri bir çalışmada karakter n-gramların lokal histogramlarını kullanarak yazar tanıma çalışması gerçekleştirmişlerdir. Veri seti olarak daha önce Plakias ve Stamatatos [21] tarafından kullanılan RCV1 verisi tercih edilmiştir. Bu veri seti 10 yazardan ve her yazar için 50 eğitim ve 50 test

dokümanından oluşmaktadır. Klasik yöntemlerden biri olan kelime çantası (BoW) ve yerel ağırlıklı kelime çantası (LoWBoW) metinleri temsil etmek için kullanılmıştır. Çalışma içerisinde kullanılan özellikler kelimeler ve karakterlerdir. Sınıflandırma yöntemi olarak SVM tercih edilmiştir. Dengeli ve dengesiz veriler üzerinde yapılan deneylerde elde edilen başarılar kıyaslanmaktadır. Özellik olarak kelimelerin seçildiği deneylerde elde edilen başarı oranı (doğruluk değeri) en yüksek 82%, karakterlerin özellik olarak kullanıldığı deneylerden elde edilen başarı oranı (doğruluk değeri) 86% olarak gözlenmektedir [21].

Shaker ve Corne, 2010 yılında yaptıkları bir çalışmada evrimsel algoritma (EA) ve doğrusal analiz yöntemlerini (LDA) kullanarak Arapça dili için hibrid yazar tanıma çalışması gerçekleştirmiştir. Zamir, edat, bağlaç gibi (fonksiyonel kelimeler) özellikler kullanılmıştır. Bu kelimeler yazılan konuyla ilgisiz olup, farklı yazarlar tarafından farklı şekillerde kullanılma özelliğine sahiptir. Çalışma içerisinde kullanılan veri seti Arap yazarlar birliği web sitesinden indirilmiştir. Kullanılan veri seti altı yazara ait toplam on dört adet kitaptan oluşmaktadır. Bu kitapların kelime sayısı 13.987 ile 37.567 arasında değişmektedir. Çalışma içerisinde özellik olarak kullanılacak fonksiyonel kelimeler, kitaplarda kullanılan en sık kelimeler ve en az kullanılan kelimeler olmak üzere iki farklı özellik kümesi olarak ayrılmıştır. En az kullanılan kelimelerle yapılan deneyler daha başarılı sonuçlar vermiştir. Performans değerlendirme ölçütü olarak bu çalışma içerisinde başarı oranı kullanılmıştır. Yapılan deneyler sonucunda elde edilen başarı oranları (doğruluk değeri) 85% ile 93% arasında değişmektedir [6].

Raghavan ve arkadaşları, 2010 yılında yaptıkları çalışmada olasılık tabanlı içerik bağımsız dil bilgisini (PCFG) kullanarak yazar tanıma yöntemi önermiştir. Önerilen yöntemin başarısını değerlendirmek için futbol, gezi, şiir, ekonomi ve kriket alanlarını içeren 5 farklı veri seti oluşturulmuştur. Bu veri setleri internet üzerinden farklı dergilerden elle toplanarak elde edilmiştir. Max Ent ve bigram-I tabanlı PCFG, PCFG-E ve bigram-I tabanlı PCFG çalışma içerisinde PCFG-I olarak ifade edilmiştir. Performans değerlendirme ölçütü olarak doğruluk değeri tercih edilmiştir. PCFG yöntemi sadece futbol veri setinde istenilen başarı oranını (doğruluk değeri) 93% elde etmiştir. PCFG-E yöntemi kullanılarak elde edilen sınıflandırıcı doğruluk oranları 87%

ile 95% arasında değişmektedir. Elde edilen başarı oranı hem sözdizimsel hem de sözcüksel özelliklerin kullanılmasının başarı oranını arttırdığı göstermektedir [5].

Ma ve arkadaşları 2011 yılında yaptıkları bir çalışmada siber suç soruşturmalarında kullanılmak üzere yazar tanımaya dayalı sosyal ağ analizi gerçekleştirmiştir. Bu çalışma içerisinde kullanılan veri setleri e-postalardan ve blog yazılarından oluşmaktadır. Yapılan çalışma bilgi çıkarımı, yazar tanıma, sosyal ağ analizi ve ağ görselleştirme olmak üzere dört ana bölümden oluşmaktadır. Dilsel, yapısal ve biçimsel özellikler olmak üzere üç özellik grubu tercih edilmiştir. Dilsel özellikler kelimelerin sıklığından göre elde edilmektir. Yapısal özellikler makalenin yapısı, morfolojik ve söz diziminden oluşmaktadır. 10 yapısal özellik, Çince ve İngilizce dillerini kapsayan 30 noktalama işareti ve 12 yaygın kullanılan konuşma özellikleri çıkarılmıştır. Ayrıca makine öğrenmesi tekniklerinden biri olan SVM yazarların yazma şekillerini öğrenmek için tercih edilmiştir. Makine öğrenmesi tekniği kullanılarak bir model geliştirilmiş ve bu aşamadan sonra sosyal ağ analizi ve görselleştirme yapılmıştır [11].

Solorito ve arkadaşları 2011 yılında yaptıkları yazar tanıma çalışması için farklı yazarların yazılarını kullanarak meta özellik oluşturmaya dayalı yeni bir yaklaşım önermiştir. Sözdizimsel, sözcüksel ve stilistik yani üslupsal özellikler kullanılmış olup her model belirli bir özellikle ilgilidir. Stilistik özellikler olarak toplam kelime sayısı, her cümledeki ortalama kelime sayısı, dijital imza, sayılar, noktalama işaretleri, kullanılan kısaltma oranı, büyük harf sayısı gibi özellikler stilistik özellikler grubundadır. Sözcüksel özellikleri elde etmek içinde bi-gram yöntemi kullanılmıştır. Part-of-Speech (PoS) etiketleri ise sözdizimsel özellikleri oluşturmaktadır. Önerilen yöntemi başarısını değerlendirmek için farklı sayılarda yazarlardan oluşan veri setleri oluşturulmuştur. Ayrıca SVM, PCFG, PCFG-I ve PCFG-E gibi sınıflandırma yöntemleriyle önerilen yöntemin başarısı kıyaslanmıştır. 100 yazardan oluşan veri setinde önerilen yöntemin başarısı (doğruluk değeri) 92% iken, SVM sınıflandırma yöntemi kullanılarak elde edilen başarı (doğruluk değeri) 28%'dir [16].

Ouamour ve Sayoud 2012 yılında yaptıkları bir çalışmada eski Arap metinlerinin yazarlarının belirlenmesi için destek vektör makinesine dayanan sıralı minimal optimizasyon yöntemi önermiştir. 2011 yılında on Arap gezgine ait olan üç kitap

toplanarak derlem oluşturulmuştur. Oluşturulan derlem içerisindeki metinlerin, ortalama kelime sayısı 429 ile 529 arasında değişmektedir. Çalışma içerisinde karakterler, karakter n gramlar (1, 2, 3), kelimeler, kelime n gramlar (1, 2, 3), ve az kullanılan kelimeler özellik olarak kullanılmıştır. Sınıflandırıcı değerlendirme ölçütü olarak duyarlılık değeri kullanılmıştır. Elde edilen deney sonuçlarına göre karakter tabanlı özellikler kelime tabanlı özelliklere göre daha iyi sonuçlar vermektedir. Ayrıca en yüksek başarı veren özellikler ise; karakter tabanlı 3-gram, karakter tabanlı 4-gram'dır [4].

Brocardo ve arkadaşları 2013 yılında yaptığı bir çalışmada yazarların biçimsel özelliklerini kullanılarak gönderilen çevrim içi kısa mesajlar için yazar doğrulama sistemi geliştirmiştir. Gözetimli öğrenme (supervised learning) ve n gram analiz yöntemlerini kullanarak hibrid bir sistem tasarlanmıştır. Enerji şirketine ait e-postalardan oluşan Enron e-posta derlemi kullanılmıştır. Önerilen yöntemin başarısının değerlendirilmesi için farklı kullanıcı sayıları ve mesaj büyüklükleri seçilerek farklı deneyler gerçekleştirilmiştir. Ayrıca n gram değerleri olarak 3,4,5 seçilmiş ve yapılan deneyler sonucunda en başarılı n gram değeri n=5 olarak belirlenmiştir. Deneylerden elde edilen eşit hata oranı 17% ile 30% arasında değişmektedir [2].

Ebrahimpour ve arkadaşlarının 2013 yılında yaptıkları çalışma da çoklu diskriminant analizi (multiple discriminant analysis) ve destek vektör makinesine dayalı iki farklı otomatik yazar tanıma sistemi önerilmektedir. Fonksiyon kelimelerin kullanım sıklığı özellik olarak kullanılmıştır. Önerilen yöntemin başarısının değerlendirilmesi için farklı veri setleri üzerinde çalışılmıştır. "Federalist paper", "İbraniler mektupları" ve "English corpus" kullanılan veri setleridir. Her iki yönteminde başarısı (doğruluk değeri) 90% olarak elde edilmesine rağmen çoklu diskriminant analizi (MDA) SVM'e göre daha esnek olarak belirlenmiştir [17].

Schwartz ve arkadaşları 2013 yılında yaptıkları çalışmada geleneksel yazar tanıma çalışmalarında tercih edilen uzun metin kullanmak yerine kısa metinlerde yazar tanıma problemini ele almışlardır. Ayrıca çalışmada "flexible patterns" olarak adlandırılan yeni bir özellik önerilmiştir. Kullanılan derlem Twitter'dan toplanan tweetlerden elde edilmiştir. Sınıflandırma yöntemi olarak SVM tercih edilmiştir. Matlab yazılımın libsvm kütüphanesi kullanılarak deneyler gerçekleştirilmiştir. SVM çekirdek fonksiyonu olarak

doğrusal (linear) çekirdek fonksiyonu tercih edilmiştir. Veri seti üzerinde 10 kat çapraz doğrulama yapılmıştır. 50 yazar ve her yazar için 500 tweet kullanılarak yapılan deneylerde 50.7% doğruluk oranı elde edilmiştir. Her yazar için toplanan tweet sayısı arttırılıp 1000 yapıldığında ise elde edilen doğruluk oranı 91% olmuştur. Yazar sayısının arttırılması ile yapılan deneysel sonuçlar incelendiğinde ise 1000 yazar ve her yazar için 200 tweet ile yapılan deney sonucunda doğruluk oranı 30.3% seviyelerde kalmıştır [18].

Layton ve arkadaşları 2013 yılında yaptıkları bir çalışmada NUANCE (n-gram Unsupervised Automated Natural Cluster Ensemble) olarak adlandırdıkları otomatik ve denetimsiz kümelemeye dayalı yeni bir yaklaşım önermişlerdir. Önerilen yaklaşım dendgram oluşturmak için EAC algoritması ve dendogram kesmek için ise IPS algoritmasını kullanmaktadır. Derlem olarak birçok çalışmada da kullanılmış olan ve çeşitli dilleri içinde bulunduruan “AAAC corpus” derlemi kullanılmıştır. Bu derlem yazar tanıma çalışmaları için birtakım zor problemlerden oluşmaktadır. Bu problemlerden özellikle A ve F problemleri zor olarak görülmektedir. A probleminin dili Amerikan İngilizcesi (American English) olarak tanımlanmış on üç yazara ait yazılardan oluşurken F probleminin dili ise Orta İngilizce (Middle English) 150-1475 yılları arasında konuşulan İngilizce olarak belirlenmiş ve üç yazardan oluşmaktadır. Performans değerlendirme metriklerinden V ölçü puanı (V-measure score) kullanılarak önerilen sistemin başarısı değerlendirilmiştir. En başarılı sonuçlar karakter ve söz dizimsel özellikler birlikte kullanıldığı zaman elde edilmiştir [19].

Sapkota ve arkadaşları 2013 yılında yaptıkları bir çalışmada ortogonal benzerlik ilişkilerini kullanarak yazar tanıma problemini ele almışlardır. Çalışma içersinde iki özellik grubu oluşturulmuştur: 1. Birinci Düzey Özellik (FLF) olarak adlandırılan stilistik, söz dizimsel ve semantik özellikler 2. Yazarlar tarafından önerilen yaklaşım RMMF (Rastgele Yöntem Meta Özellik) olarak adlandırılmış ve yazarlar arasında benzerlik ilişkilerini içeren özellik kümesini oluşturmaktadır. Deneylerde kullanılan veri setleri, “The Chronicle of Higher Education” (CHE), CCAT ve Raghavan ve arkadaşları [5] tarafından internetten spor, ekonomi gibi alanlarda yazılardan oluşan veri setidir. Sınıflandırma yöntemi olarak makine öğrenmesi araçlarından biri olan Weka kullanılmış olup sınıflandırma yöntemi olarak SVM tercih edilmiştir. Yapılan deneyler sonucunda

elde edilen başarı oranları (doğruluk değerleri) CHE veri seti için 72%, CCAT veri seti için 76% ve Raghavan ve arkadaşları [5] tarafından oluşturulan veri setleri için ise 91% ile 63% arasında başarı (doğruluk değerleri) elde edilmiştir [20].

Sikos ve arkadaşları 2014 yılında yaptıkları bir çalışmada yazı stili özelliklerinin aşırı dinci yazıların yazarlarının analizinde nasıl kullanıldığını açıklamaktadır. El Kaide tarafından çevrim içi olarak internet üzerinde yayınlanan "Inspire magazine" dergisi veri seti olarak kullanılmıştır. Bu dergi içerisinde yazarların kelime seçimleri, cümle yapısı gibi biçimsel özelliklerle birlikte anlamsal ve psikolojik özellikler yazarların yazı stillerindeki benzerlikleri ve farklılıkları belirlemek için kullanılmıştır. The Linguistic Inquiry and Word Count (Dil Bilim Araştırması ve Kelime Sayısı) bir sözlük olup bilişsel ve psikolojik durumlarla ilgili olan kelime ya da kelime gruplarının tanınmasında kullanılır. İngilizce için kullanılan bu sözlüğün Arapça dili içinde kullanılması için gerekli düzenlemeler çalışma içinde yapılmıştır [13].

Okuno ve arkadaşları 2014 yılında yaptıkları bir çalışmada on bin mikro-blog kullanıcısının yazılarını tanımak için yeni bir yaklaşım önermiştir. Önerilen yöntem içerisinde ilk olarak, eğitim verisi için kombine seçim yöntemi olarak adlandırılan bir işlem gerçekleştirilmektedir. Bu işlemde eğitim verisi için belirtilen yöntem uygulanır ve belirli bir test kümesi için en uyumlu eğitim verisi seçilir. Bu seçim yapılırken belirli test kümesi için üç farklı benzerlik hesaplaması gerçekleştirilir. İkinci olarak, n gram için biased ağırlıklandırma yöntemi kullanılarak kısa metinler işlenir. Ayrıca, PoS tag-ve-n-gram kombinasyonu kullanılarak özellik azaltma işlemi yapılmıştır. Deney sonuçlarına göre önerilen yöntem 53.2% oranında yazarı belli olmayan tweetlerin yazarını belirler. Çalışmada İngilizce dilinde yazılan tweetler tercih edilmesine rağmen yazarlar takip eden çalışmalarında Japonca dili için aynı çalışmayı gerçekleştirmeyi planlamaktadır [8].

Berry ve Sazonov 2015 yılında yaptıkları çalışmada denetimsiz öğrenme algoritmalarının yazar tanıma problemi üzerindeki başarılarını incelenmiştir. Veri seti olarak mühendislik konferansına ait 23 makaleden oluşan bir derlem kullanılmıştır. 8 yazara ait olan bu makalelerden stilistik özellikler çıkartılarak dokuman kümeleme yöntemi uygulanmıştır. Çalışma içerisinde kullanılan stilistik özellikler ortalama kelime

uzunluğu, ortalama cümle uzunluğu, 3 karakterden kısa olan kelimelerin sayısı, Hapax Legomena, Hapax Dislegomena, makalelerdeki özgün kelimelerin yüzdesi ve karakter kelime frekanslarıdır. Adjusted Rand Index (ARI) kullanılmış olup bu elde edilen başarı oranı (doğruluk değeri) ise 73% olarak verilmiştir [12].

3.2. Türkçe için Yapılan Çalışmalar

2004 yılında Patton ve Can, Yaşar Kemal'in İnce Mehmed dörtleme kitap serisi için bir yazar analiz çalışması gerçekleştirmiştir. Bu çalışmada en sık kullanılan kelimeler, hece sayısı, kelime türü, cümle uzunluğu gibi özellikler tercih edilmiştir. Seçilen özelliklerin başarı oranları karşılaştırılmış ve en sık kullanılan kelimeler ve cümle uzunlukları özelliklerinin en iyi sonuçları verdiği gözlenmiştir. Yapılan deneyler sonucunda ilk iki roman birbirinden ayrılabiliyorken diğer iki roman için bir ayrım yapılamamıştır. Bunun nedeni olarak İnce Mehmed'in ilk serisinin romantik, ikinci serisi gerçekçi, son iki seri ise postmodern olarak sınıflandırılması gösterilmiştir [54].

Amasyalı ve Diri 2006 yılında gerçekleştirdikleri bir çalışmada Türkçe dili için n=1,2,3 gram modelleri kullanarak metin sınıflandırma çalışması yapmışlardır. Metinleri yazan kişinin belirlenmesi, metnin türünün belirlenmesi son olarak da yazarın cinsiyetinin belirlenmesi olmak üzere üç farklı alanda çalışılmıştır. Tür belirleme çalışması için yaşam, politika ve spor olmak üzere üç farklı alan seçilmiştir. Kullanılan veri seti 18 yazardan (14 erkek- 4 bayan) ve her yazara ait farklı 35'er yazıdan oluşmaktadır. Çalışma içerisinde Naive Bayes, SVM, C 4.5 ve RF olmak üzere dört farklı sınıflandırma yöntemi kullanılmıştır. bi-gram ve tri-gram yöntemleri kullanılarak özellikler elde edilmiştir. CFS (Korelasyon tabanlı özellik seçme) yöntemi kullanılarak elde edilen özelliklerde boyut azaltmaya gidilmiştir. Yazar tanıma alanında yapılan deneyler sonucunda bi gramlar tri gramlara göre daha başarılı olup Naive Bayes sınıflandırma metodu en iyi sonucu vermiştir. Tür belirleme deneylerinden elde edilen sonuçlara göre ise SVM sınıflandırma yöntemi diğer yöntemlere göre daha başarılı sonuçlar vermektedir. Ayrıca SVM sınıflandırma yöntemi cinsiyet belirleme çalışmasında da en başarılı sonucu vermiştir. Elde edilen başarı oranları (doğruluk değerleri) yazar tanıma için 83%, tür belirleme için 93% ve cinsiyet belirleme çalışması için 96% dır [52].

Türkoğlu ve arkadaşları 2007 yılında gerçekleştirdikleri bir çalışmalarında istatiksel, sözcüksel, dil bilgisel, fonksiyonel kelimeler ve n gram özellikleri ve bunların kombinasyonlarından farklı özellik vektörü elde etmişlerdir. Çalışma içerisinde üç farklı veri seti kullanılmıştır. Bunlar, 18 yazara ait farklı konularda yazılmış 630 metinden veri seti-2, 9 yazara ait benzer konular hakkında yazılmış 315 metinden ve veri seti-3, 9 yazara ait ve farklı konularda yazılmış 315 metinden oluşan veri setidir. Elde edilen vektörlerin performans karşılaştırılmasının yapılması için Naive Bayes, SVM, k-NN, RF ve MLP sınıflandırma yöntemleri kullanılmıştır. En başarılı yöntemler olarak SVM ve MLP olarak belirlenmiştir. Ayrıca n gram yöntemi kullanılarak elde edilen özelliklerin diğer özellik gruplarından daha başarılı sonuç verdiği gözlemlenmiştir [26].

Kaban ve Diri 2008 yılında yaptıkları bir çalışma içerisinde yapay bağışıklık sistemleri kullanarak yazar ve tür tanıması gerçekleştirmişlerdir. Çalışma içerisinde kullanılan veri seti Milliyet, Sabah ve Hürriyet gazetelerinde köşe yazıları yazan toplam on sekiz yazardan oluşmaktadır. Tür belirleme çalışması için 5 farklı alan belirlenmiştir. Önerilen yöntemin başarısını diğer sınıflandırıcı yöntemlerle de kıyaslamak için NB, DVM, RO ve k-NN sınıflandırıcı yöntemleri kullanılmıştır. Tür tanıma çalışması için yapılan deneylerde yapay bağışıklık algoritmalarından biri olan YBSP sınıflandırıcının en iyi performansı verdiğini göstermektedir. Sistemin ortalama hatası 0,8% olarak bulunmuştur. Yazar tanıma alanında yapılan deneyler incelendiğinde yapay bağışık sistemlerinin yazar tanıma başarısını arttırdığını bulmuşlardır [25].

Küçükyılmaz ve arkadaşları 2008 yılında stilistik özellikleri kullanarak 100 yazar arasında yazar tanıma çalışması gerçekleştirmiştir. Kullanılan stilistik özellikler üzerinde ayrıklaştırma (discretization) yapılmıştır. Örneğin kelime uzunluğu özelliği kısa, orta ve uzun olmak üzere kategorize edilmiştir. Ayrıca bu çalışmada peer to peer (P2P) ağlarda kişisel ve çevresel özelliklerin yazarların yazma biçimi ve kelime kullanımına etkisi gösterilmiştir [53].

2010 yılında Aslantürk ve arkadaşları yaptıkları çalışmada kaba küme sınıflandırıcı yöntemi kullanılarak homojen ve homojen olmayan yazılar üzerinde yazar tanıma gerçekleştirmişlerdir. Nokta, virgül gibi noktalama işaretleri, kelime, cümle sayısı gibi metin özellikleri ve isim, sıfat, zamir gibi kelime türleri özellik olarak kullanılmıştır. Veri

seti olarak gazetelerden 513 köşe yazısı indirilmiştir. Bu yazılar siyaset ve yaşam olmak üzere iki alanda yazılmış ve toplam 8 yazardan oluşmaktadır. Yapılan deneysel çalışmalar sonucunda aynı alanda yazılmış köşe yazılarından oluşan homojen verilerden elde edilen sınıflandırıcı başarısı farklı alanda yazılmış köşe yazılarından heterojen verilerden elde edilen sınıflandırıcı başarısından daha yüksek olduğu gözlenmiştir [40].

Aslantürk tarafından 2010 yılında gerçekleştirilen doktora tezi çalışması kapsamında yazar tanıma çalışmalarında kullanılabilecek başarılı sonuçlar veren en küçük özellik kümelerinin bulunması ve bulunan bu özelliklerin ne kadar zaman geçerli olabileceğini araştırılmıştır. Problemin çözümü için 1134 deney gerçekleştirilmiştir. Sınıflandırıcı olarak Türkçe çalışmalarda hiç kullanılmamış olan kaba küme sınıflandırıcı kullanılmıştır. Veri seti olarak siyaset ve yaşam alanında yazılar yazan toplam 8 yazara ait köşe yazıları internet üzerinden derlenmiştir. Yapılan deneyler sonucunda gelişmiş noktalama özellik grubu olarak ayrılan özelliklerin daha başarılı sonuçlar verdiği gözlenmiştir [41].

Varol 2011 yılında yaptığı yüksek lisans tez çalışmasında Türkçe şiirlerin yazarlarını ve türünü belirlemeye çalışmıştır. Kelime zenginliğine bağlı özellikler ya da karakter n gramları gibi istatiksel özellikler çıkartılmış ve Weka programında bulunan Bayes, Functions, Lazy, Meta, Rules, Trees sınıflandırıcı yöntemleri kullanılarak şair tanıma gerçekleştirilmiştir. Bu çalışma için kullanılan eğitim veri seti 7 şair ve her bir şaire ait 30 şiirden oluşmaktadır. Test veri seti ise eğitim veri setinden bulunan şairleri ve ayrıca buna ek olarak 4 şairden oluşmaktadır. Ayrıca Doğan tarafından yapılan çalışmada önerilen Ng-İnd sınıflandırma yöntemi kullanılmıştır [26]. Gerçekleştirilen deneylerin başarı karşılaştırılması yapıldığında en başarılı sınıflandırıcının 71% başarı (doğruluk değeri) ile MLP olduğu gözlenmektedir [55].

Ekinci ve Takçı 2012 yılında yaptıkları bir çalışmada e-postalar üzerinde yazar tanıma için farklı sınıflandırıcı performanslarını değerlendirmiştir. Bilişim suçları içerisinde yer alan kötü niyetli e-postaları gönderen kişinin belirlenmesi, yazar tanıma çalışmaları içerisinde değerlendirilmektedir. Kullanılan veri seti 5 kişiye ait toplam 250 adet e-postadan oluşmaktadır. Elde edilen veri seti içerisinde, her yazar için 40 e-posta eğitim,

10 e-posta ise test etmek için kullanılmıştır. Çalışma içerisinde kullanılan özellikler kelime sayısı, karakter sayısı, harf sayısı, rakam sayısı, noktalama sayısı, büyük harf sayısı, küçük harf sayısı, ortalama kelime uzunluğu, ortalama cümle uzunluğu, kelime zenginliği, fonksiyonel kelimeler, bir kere geçen kelimeler, iki kere geçen kelimelerdir. Bu özelliklere ek olarak sadece e-postalarda bulunan selamlama, imza gibi özellikler çıkarılmıştır. Deneylerin gerçekleştirilmesi için makine öğrenmesi araçlarından biri olan Weka kullanılmıştır. Yapılan deneylerin performans değerlenmesini yapmak için doğruluk, kesinlik, duyarlılık ve F-ölçümü metriklerinden yararlanılmıştır. Sınıflandırma başarıları Çizelge 3.1.'de belirtilmiştir [23].

Çizelge 3.1. Performans Değerlendirme Sonuçları [23].

Başarı Ölçümleri	Çok Katmanlı YSA (%)	SVM (%)	J48 (%)
Doğruluk	0.8	0.8	0.84
Kesinlik	0.83	0.81	0.84
Duyarlılık	0.8	0.8	0.84
F-Ölçümü	0.79	0.79	0.83

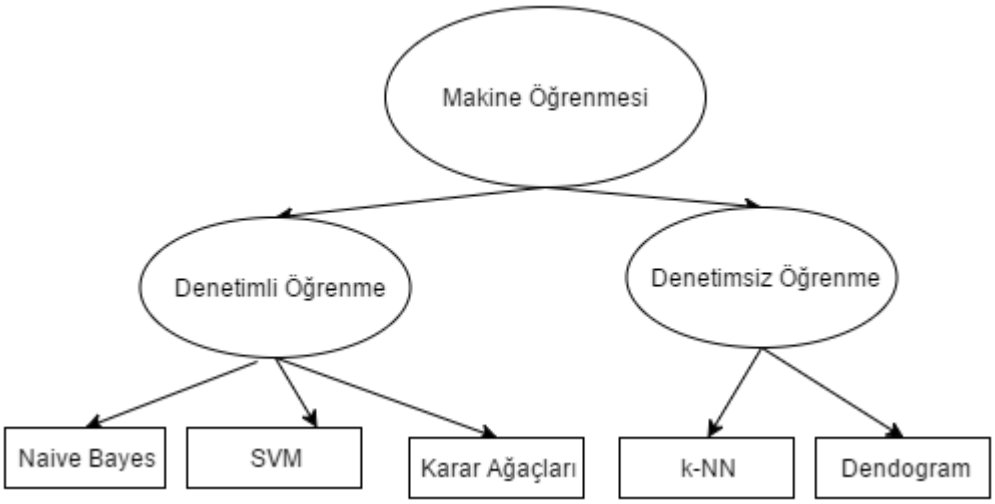
Demir tarafından 2012 yılında yapay sinir ağları kullanarak Peyami Safa, Orhan Pamuk ve Mustafa Necati Sepetçioğlu'nun romanları üzerine bir çalışma gerçekleştirmiştir. Çalışma içerisinde karakter, kelime, cümle, cümle uzunluğu, kelime uzunluğu, edat, soru işareti, virgül, üç nokta, ünlem işareti, bağlaç sayıları ve kullanılan "ve" sayısı özellik olarak çıkarılmıştır. Temel Bileşen Analizi (PCA) kullanılarak özellik azaltma gerçekleştirilmiştir. Bu aşamadan sonra sinir ağının eğitim ve test aşaması gerçekleştirilmiştir. Sınıflandırıcının başarısını değerlendirmek için aynı yazarların farklı romanları kullanılmıştır. Peyami Safa'nın "Yalnızız", Mustafa Necati Sepetçioğlu'nun "Anahtar" ve Orhan Pamuk'un "Benim Adim Kırmiızı" adlı romanlar sınıflandırıcı performansı değerlendirmek için kullanılmıştır.

Çizelge 3.2. Alanyazın Karşılaştırması

Ref	Veri Seti	Yöntem	Dil	Başarı Oranı
[14]	Germany Newspaper	SVM-Light	Almanca	ACC=60%-80%
[7]	Forum Blog	SVM	Çince	ACC= 52%-75%(S
[15]	Telegraph Newspaper	İstatiksel	İngilizce	ACC=75%
[9]	CiteSeer	Gizli Anlam Analizi. Karakteristik Vektör Modeli	İngilizce	ACC=50%
[21]	RCV1 verisi	n-gramların lokal histogram	İngilizce	ACC=82%
[6]	Arap Yazarlar Birliği	Evrimsel Algoritma ve Doğrusal Analiz	Arapça	ACC=85%
[5]	Futbol, Şiir, Ekonomi, Kriket alanında yazılar	Olasılık tabanlı içerik bağımsız dil bilgisini (PCFG)	İngilizce	ACC=85%
[11]	E-posta ve Blog	SVM	İngillizce, Çince	N/A
[16]	Blog	SVM, PCFG , PCFG-I ve PCFG-E	İngilizce	ACC=28%
[4]	Arap Gezginlerine Ait Kitaplar	SVM	Arapça	Duyarlılık= 0.6
[2]	Enron e-posta	n-gramlar	İngilizce	Eşit Hata Oranı: 17%
[18]	Twitter	SVM	İngilizce	ACC=50%
[20]	CHE, CAAT, Futbol, Şiir, Ekonomi alanında yazılar	SVM	İngilizce	ACC= 63%
[8]	Twitter	N/A	İngilizce	ACC=53.2
[12]	Makale	Adjusted Rand Index	İngilizce	ACC=53.2

4. ALANYAZIN ÖZETİNDE KULLANILAN YÖNTEMLER

Yazar tanıma alanında yapılan çalışmalar makine öğrenmesi yöntemlerinden sıklıkla yararlanmaktadır [1-55]. Bu yöntemler denetimli (supervised) ve denetimsiz (unsupervised) öğrenme olmak üzere ikiye ayrılmaktadır. Denetimli öğrenme de eğitim aşamasında kullanılan verilerin hangi sınıfa ait olduğu bilinmekte denetimsiz öğrenmede ise verilerin hangi sınıfa ait olduğu bilgisi bilinmemektedir [40]. Yazar tanıma alanında yapılan çalışmalar ayrıntılı olarak Bölüm 3'de incelenmiştir. Bu bölümde ise bu çalışmalarda sıklıkla kullanılan yöntemlerden bahsedilmiştir. Şekil 4.1'de sıklıkla kullanılan yöntemler gösterilmektedir.



Şekil 4.1. Makine Öğrenmesi Yöntemleri [32].

4.1. Navie Bayes

Bayes teoreminin temeli olasılık kuramına dayanmaktadır. Bir olay olarak A olayına koşullu B olayı için olasılık değeri ya da B olayına koşullu A olayı için olasılık değerleri farklı olmasına rağmen aralarında belli bir ilişki bulunmaktadır ve bu ilişki Bayes teorimi olarak ifade edilmektedir [67]. Bayes teorimi Eşitlik 13. ile ifade edilmektedir.

$$P(A/B) = \frac{P(B/A)P(A)}{P(B)} \quad (13)$$

Bu teorem içinde yer alan her bir terimin özel bir adı bulunmaktadır.

P(A) : A olayı için Marjinal olasılık

P(A/B):B için A'nın koşullu olasılığı

P(B/A): A için B'nin koşullu olasılığı

P(B) : B olayı için Marjinal olasılık değerlerini ifade etmektedir.

Bayes teoreminin farklı varyasyonları bulunmaktadır [67]. En sık kullanılan sınıflandırıcıdan birisi Naive Bayes (NB)'dir. NB istatiksel bir yöntem olup temeli Bayes teoremine dayanmaktadır. NB, makine öğrenmesi yöntemleri içerisinde denetimli öğrenme algoritmalarından birisidir [25-26]. Bayes teoreminin basitleştirilmiş şekli olarak ifade edilmektedir [67]. Bu yöntem özelliklerin birbirine bağımsız olma varsayımına dayanmaktadır. Naïve Bayes teorimi Eşitlik 14'de gösterilmektedir [67].

$$P(Ci \mid X) = \frac{P(X|Ci)P(Ci)}{P(X)} \tag{14}$$

Bayes teoremindeki, $P(Ci \mid X)$, x olan bir verinin *i* sınıfına ait olma olasılığıdır. $P(X \mid Ci)$, *i* sınıfın içindeki bir verinin x olma olasılığıdır. $P(Ci)$, *i* sınıfının olma olasılığı , $P(X)$ ise tüm veri kümesi içinde bir verinin X olma olasılığıdır.

NB, kolay ve anlaşılır olması nedeniyle birçok uygulama alanında kullanılmaktadır. Bu uygulama alanlarından biriside Yazar Tanıma alanıdır [25,26,52,55]. Burada özelliklerin birbirinden bağımsız olma varsayımını altında, metinlerden çıkartılan özellikler {a_1, a_2, a_3, … a_n} olarak ifade edilmekte ve yazar v olarak gösterilmektedir. Burada$a_1 a_2, a_3, \ldots a_n$ homojen ve n sabittir. Böylece metinlerin özellikleri sınıflandırma işlemi için hesaplanmaktadır. Bayes teorimi kullanıldıktan sonra Naive Bayes sınıflandırıcı olarak kullanılabilmektedir [4,15]. İşlem adımları aşağıda gösterilmektedir.

$$P(V \mid a_1 a_2, a_3, \ldots a_n) = \frac{P(V)P(a_1 a_2, a_3, \ldots a_n|\mathrm{V})}{P(a_1 a_2, a_3, \ldots a_n)} \tag{15}$$

$$P(a1, a2, a3, \ldots an \mid V) = \prod p(ai \mid V) \tag{16}$$

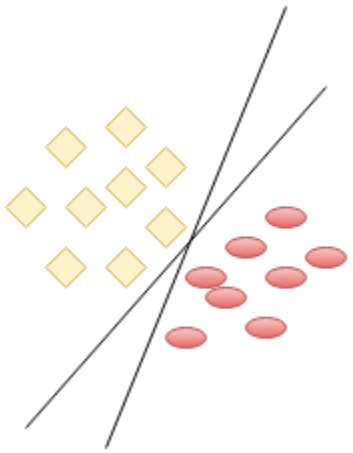
$$V = argmax_{v \in V} P(V) \prod p(ai \mid V) \tag{17}$$

P(v) eğitim veri seti içerisinde v yazarının olma olasılığıdır.

Olasılık üzerine kurulu bir sınıflandırma yöntemi olan Naive Bayes aslında verilen bir metini sınıflandırmak için kelimeler ve sınıfların olasılıklarından yararlanmaktadır [3,5,16,25-27,55]. Zhao ve Zobel 2005 yılında zarf, edat, bağlaç gibi fonksiyonel kelimeleri ile Naive Bayes ve Bayes ağları makine öğrenmesi yöntemlerini kullanarak yazar tanıma problemi üzerine bir çalışma gerçekleştirmiştir [3].

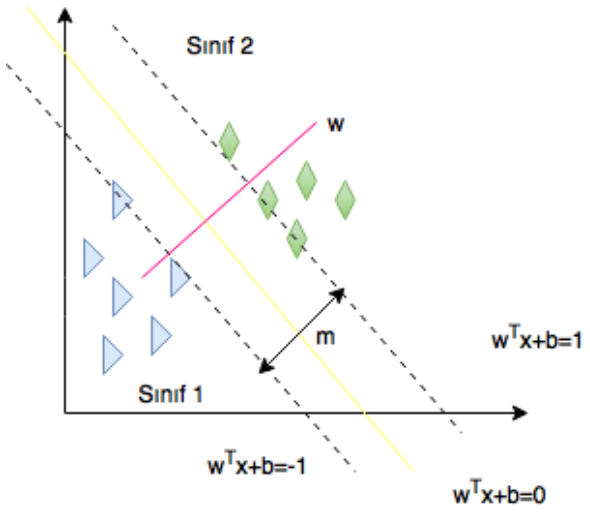
4.2. Destek Vektör Makineleri (SVM)

Veri madenciliği ya da metin madenciliği problemlerinde sıklıkla kullanılan Destek Vektör Makineleri yazar tanıma probleminin çözümünde de sıklıkla tercih edilmektedir [10-15]. Makine öğrenmesi yöntemlerinden biri olan ve Vladimir Vapnik tarafından geliştirilen bu algoritma istatiksel öğrenme teorisine ve yapısal riski minimuma getirme amacına dayanmaktadır [68]. Bio-informatik, örüntü tanıma, verilen bir metnin hangi dilde yazıldığının bulunması (dil tanıma) gibi birçok alanda kullanılmaktadır [68].



Şekil 4.2. Verilerin SVM ile Doğrusal Ayrılması [14]

SVM algoritmasının temeli aslında verileri en uygun şekilde iki sınıfa ayıran karar düzleminin belirlenmesi işlemidir [10-13]. Şekil 4.2'de farklı sınıfa ait verilerin doğrusal olarak farklı karar sınırları ile ayrılması gösterilmektedir. SVM sınıflandırma işlemi için en önemli adım verileri ayırmak için en uygun karar sınırının belirlenmesidir. Bu sınırın belirlenmesinde karar sınırının ayırdığı verilere en uzak olması değerlendirilmektedir [14].



Şekil 4.3. Karar Sınırı ve Hiper Düzlemler [14]

$\{x_1,x_2,x_3,\ldots x_n\}$ verileri ifade ediyorken $y_i \epsilon \{-1,1\}$ ise verilerin ait oldukları sınıfları göstermektedir. $y(x)=w^Tx+b$ denkleminde y sınıfı ifade ederken, bir vektör olan x kelime sayısını kadar olup w ise hiperdüzlemin normalini göstermektedir. Şekil 4.3 incelendiği zaman veriler karar sınırıyla birbirinden ayrılmıştır. Hiperdüzlem karar sınırını belirler ve bu doğruya paralel olan kesikli doğrular ise sınıflara ait olan sınırları göstermektedir. Bu sınırların üzerinde bulunan veriler destek vektörler olup SVM sınıflandırıcı sadece bu vektörleri kullanarak işlem yapmaktadır [14].

SVM makine öğrenmesi yöntemi birden fazla sınıfın bulunduğu problemlerde de kullanılmaktadır [13-15]. Bunun için çeşitli yaklaşımlar geliştirilmiş olup en sık kullanılanlar bire hepsi ya da bire-ikili sınıflandırma yaklaşımlarıdır. Bire hepsi yaklaşımında var olan tüm sınıflar için sınıflandırıcı belirlenir ve böylece ilgili sınıfa ait olan veriler diğerlerinden ayrılır ve sınıflandırma işlemi gerçekleşmiş olur. Diğer yaklaşımda ise sınıflar ikili olarak ele alınır ve karşılaştırma yapılarak ilerlenir. İlk yaklaşıma göre daha iyi sonuçlar verdiği düşünülmektedir [14].

4.3. Karar Ağaçları

Karar ağaçları çeşitli veri madenciliği ve metin madenciliği problemlerinin çözümünde kullanılmaktadır [69]. Bu problemlerden birisi de yazar tanımadır [7,10,23]. Bu sınıflandırma yöntemine anlaşılmasının kolay derlenmesi, güvenli olması ve maliyetinin düşük olması nedeniyle sıklıkla tercih edilmektedir [52]. Bu yöntem öğrenme ve

sınıflandırma olmak üzere iki aşamadan oluşmaktadır. Öğrenme aşamasında, model oluşturmak için eğitim verisi tercih edilen bir sınıflandırma algoritması kullanılarak analiz edilir. Bunun için ilk olarak entropi hesabı yapılır. Verilen bir alanın Entropi değerini bulan formül Eşitlik 18'de gösterilmektedir. Entropi değerlerine göre karar verilir ve değeri en az olan ağacın kökünü oluşturur [7,10].

$$E(C \mid A_k) = \sum_{j=1}^{Mi} p(a_k, j)\, x[-\sum_{i=1}^{N} p(c_i \mid a_k, j) \log_2 p(c_i \mid a_k, j)] \quad (18)$$

$E(C \mid A_k)$= A_k alanının sınıflama özelliğinin Entropi ölçüsü

$p(a_k, j)$= a_k alanının j deperinde olma olasılığı

$p(c_i \mid a_k, j)$= a_k alanı j. değerindeyken sınıf değerinin c_i olma olasılığı

N= Sınıf sayısı

K= Alanların sayısı

Bir veri kümesi içindeki verinin sınıfını belirlemek için eşitlik 19 hesaplanır. P_i, C_i sınıfına ayrılma olasılığıdır. Eşitlik 20'da ise Entropi hesabı verilmektedir.

$$I(S) = -(p_1 \log_2 p_1) + (p_2 \log_2 p_2) + \cdots + (p_i \log_2 p_i) \quad (19)$$

$$E(A) = \sum_{i=1}^{n} \frac{|S_i|}{|S|} * I(S_i) \quad (20)$$

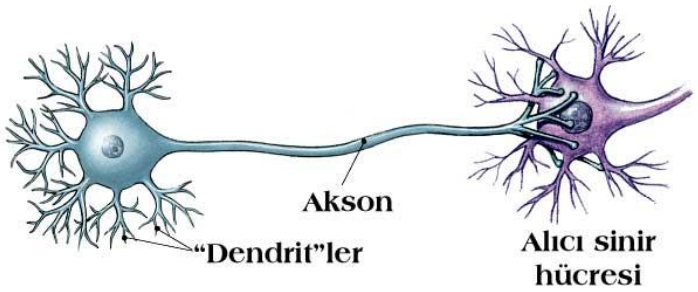
Ayrıca ID3, C4.5, CART ve CHAID gibi karar ağaçlarında kullanılan çeşitli algoritmalar alanyazında mevcuttur [7,10].

5. YAPAY ZEKA YÖNTEMLERİ

İngiliz matematikçi, bilgisayar bilimcisi ve kriptoloji uzmanı olan Alan Turing tarafından "Makineler düşünebilir mi?" sorusu üzerine ortaya çıkan bir kavram olan yapay zekâ, makinelerin öğrenme, düşünme, eski tecrübelerinden yararlanma, algılama gibi insana özgü olan özelliklere sahip olmasını hedefleyen bilim alanıdır [71]. Bu alanda yapılan ilk çalışmalardan birisi de McCullich ve Pitts'e ait olup Turing'in hesaplama kuramına dayanmaktadır [71]. Geleneksel yöntemlerle çözümü olmayan ya da hem zaman ve hem de maliyet açısından pahalı olan problemlerin çözümünde yapay zekâ yöntemlerinden yararlanılmaktadır. Yapay sinir ağları, bulanık mantık, genetik algoritmalar, bağışıklık sistemi, uzman sistemler yapay zekâ yöntemlerindendir. Tez kapsamı içerisinde yazar tanıma için kullanılan yapay sinir ağlarının yapısı ve öğrenme algoritmaları ile ilgili bilgiler bu bölümde sunulmaktadır.

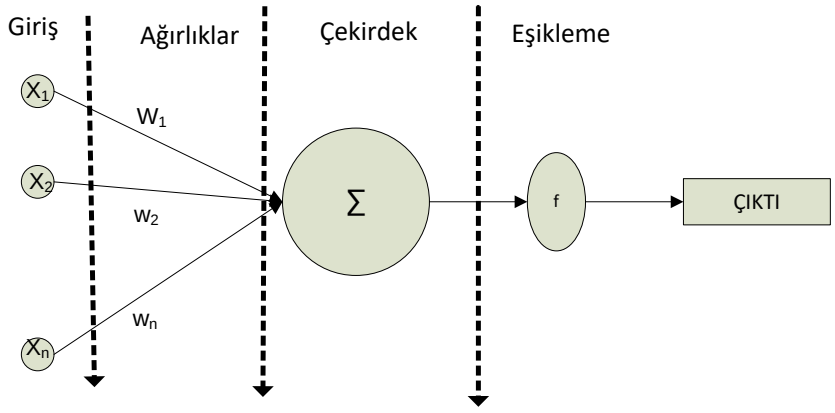
5.1. Yapay Sinir Ağları (Artifical Neural Network)

Geçmişten günümüze kadar insanoğlu karşılaştığı problemlere doğa içerisinde çözümler aramıştır. Doğadaki bazı durumlardan esinlenerek bu problemlere çözümler getirmiştir [72]. Örneğin, optimizasyon problemlerinde sıklıkla kullanılan yöntemlerden biri olan karınca kolonisi algoritması karıncaların içgüdüsel olarak yemeklerine ulaşmak için en uygun yolu seçmesinden ilham alınarak geliştirilmiştir. Evrimsel hesaplama yöntemlerinden biri olan genetik algoritma ise en uygun sonucu elde etmek için evrimden yararlanır. Yapay sinir ağları (YSA) ise insan beyninin çalışma prensibini taklit ederek burada yapılan öğrenme, genelleme yapabilme, hatırlama, tecrübe etme, bilgi depolama ve ilişkilendirme gibi işlemleri gerçekleştirmeyi amaçlayan mantıksal programlama yöntemidir [72,73]. YSA'lar biyolojik sinir sistemine benzeyen yapılardır. Haykin'in 1999 yılında Sinir Ağlarına Detaylı Bakış adlı kitabında yaptığı tanım yaygın kabul görmüş tanımlamalardan birisidir. YSA'yı, *"Bir sinir ağı, bilgiyi depolamak için doğal eğilimi olan basit birimlerden oluşan paralel dağıtılmış bir işlemcidir*." olarak tanımlar [72]. Bir diğer yaygın kabul görmüş YSA tanımlaması Zurada tarafından yapılan "*Yapay sinir sistemleri veya sinir ağları deneysel bilgiyi alan, depolayan ve kullanan fiziksel hücreli sistemlerdir*." tanımlamasıdır [73].



Şekil 5.1. Sinir Hücresi [73].

Şekil 5.1'de nöron olarak adlandırılan insan sinir hücresinin yapısı gösterilmektedir. Nöronlar, akson, dendrit ve alıcı sinir hücresi olmak üzere üç kısımdan oluşmaktadır. Dendritler farklı bir sinir hücresinden gelen bilginin alır ve alıcı sinir hücresi yardımıyla aksonlara iletir ve bir nöronda birden fazla bulunabilirler. Aksonlar ise bilgilerin yorumlandığı kısımdır. Aksonlarda yorumlanan bilgiler aksonun ucundaki dallardan başka sinir hücrelerine gider. Alıcı sinir hücresi ya da soma olarak adlandırılan kısım sinir hücresinin yaşamsal faaliyetlerini sürdürdüğü kısımdır. Ayrıca burada dendritlerden gelen sinyaller aksonlara iletilir. Sinir hücrelerinin başka sinir hücreleriyle iletişiminin gerçekleştiği yapısal ve fonksiyonel olarak özelleşmiş kısımlar sinaps olarak adlandırılır. Sinapslardaki iletim kimyasal ya da elektriksel olur. Sinir hücreleri arasındaki iletişimin elektriksel olarak gerçekleştiği sinapslar elektriksel sinapslar ve iletişimin kimyasal olarak gerçekleştirildiği sinapslarda kimyasal sinapslardır. Elektriksel sinapslarde iletişim iki yönlüyken kimyasal sinapslarda iletişim tek yönlüdür [74-77].



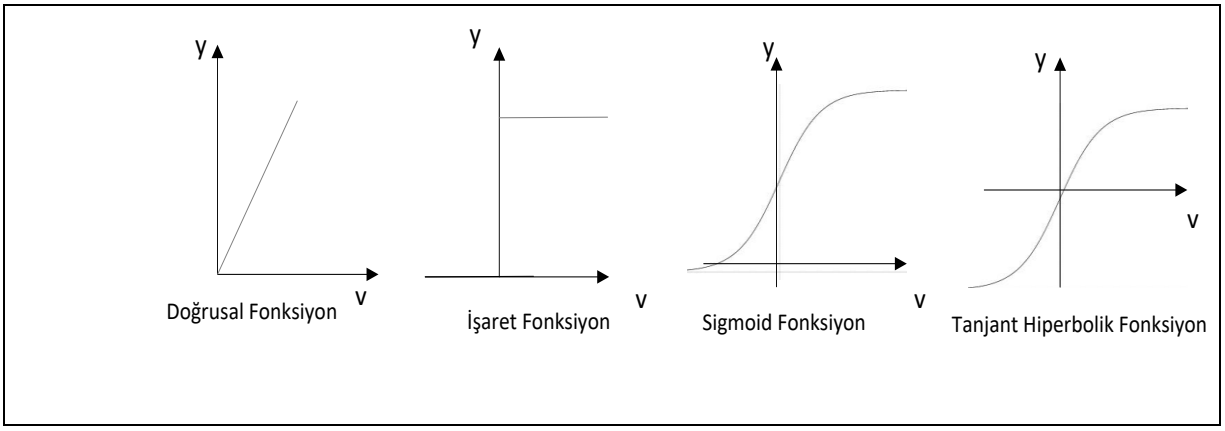
Şekil 5.2. Yapay Sinir Hücresi [76].

Yapay sinir hücresinin yapısı insan sinir hücresine oldukça benzemektedir. Bir yapay sinir hücresi (nöron), girişler, ağırlıklar, toplama fonksiyonu, transfer fonksiyonu ve çıkış olmak üzere beş kısımdan oluşmaktadır. Yapay sinir hücresinin genel yapısı Şekil 5.2'de gösterilmektedir. Bu şekilde x girişleri, y çıkışı, F fonksiyonu, n nöron giriş sayısını w ise ağırlıkları ifade etmektedir [74,75].

Giriş değerleri dış kaynaklardan ya da diğer nöronlardan gelen veriler olup bu verilerin her birinin ağırlık değerleri önemine göre değişiklik göstermektedir. YSA'da girilen giriş değerlerine göre önce toplama fonksiyonları uygulanmaktadır. Eşitlik 22' de x giriş değerleri, w ağırlık değerleri ve θ eşik (threshold) değerini göstermektedir [76].

$$v = \sum_{i=1}^{n} x_i w_{ij} - \theta_i \qquad (22)$$

Bu adımdan sonra bu değer aktivasyon fonksiyonlarından birine uygulanırsa Eşitlik 23'deki çıkış değeri elde edilmektedir. Bu başka bir yapay sinir ağının giriş değeri olabileceği gibi yapay sinir ağının çıkış değeri de olabilmektedir. Aktivasyon fonksiyonları ya da öğrenme eğrileri yapay sinir hücresinin çıkış değer aralığını istenilen değerler arasına normalize edilmektedir. Genellikle bu değerler [0-1] ya da [-1,1] arasındadır fakat bias değeri uygulanarak bu değerler üzerinde değişim yapabilmektedir. Uygulamalarda sıklıkla kullanılan aktivasyon fonksiyonların Şekil 5.3'de gösterilmektedir [76].



Şekil 5.3. Aktivasyon Fonksiyonları [76].

Doğrusal fonksiyon ya da lineer fonksiyon, nöronun giriş değerini doğrudan çıkış olarak vermektedir. YSA'nın çıkış katmanında kullanılmaktadır [77].

İşaret ya da basamak fonksiyonu, toplama fonksiyonunda elde edilen değer sıfırdan büyük ve eşit ise bir değerini alır sıfırdan küçük ise sıfır değerini alır.

Sigmoid ya da tek kutuplu fonksiyonu, çalışmalarda en sık kullanılan aktivasyon fonksiyonudur. Eşitlik 23'de verilen formüle göre hesaplanmaktadır.

$$y = \frac{1}{1 + e^{-v}} \tag{23}$$

Tanjant Hiperbolik ya da çift kutuplu (bipolar) fonksiyonu, giriş uzayının genişletilmesinde etkili bir fonksiyondur. Eşitlik 24'de verilen formüle göre hesaplama yapılmaktadır.

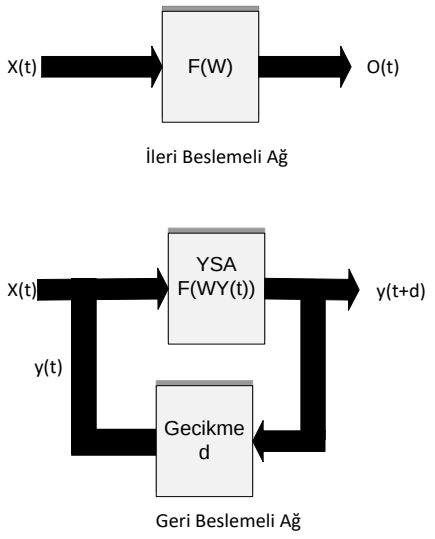
$$y = \frac{1 - e^{-2v}}{1 + e^{2v}} = \tanh(\beta v) \tag{24}$$

YSA'ları geleneksel yöntemlerden ayıran özelliklere değinilecek olursak ilk olarak yapay sinir hücrelerinin (nöronların) lineer olmaması YSA'ların doğadaki neredeyse tüm problemlere çözümler sağlamasına olanak vermektedir. Bu özellik lineer olmayan transfer fonksiyonları kullanılarak sağlanmaktadır [76]. Bir diğer özelliği ise YSA'ların öğrenme yapabilmesidir. Bir problemin YSA tarafından öğrenilebilmesi için problemin giriş verilerine ve bu değerlere karşılık gelen çıkış değerlerine ya da sadece giriş verilerine gereksinim vardır. Öğrenmenin gerçekleşmesi için yeterli sayıda giriş ve çıkış verilerinin olması ve uygun YSA modelinin düzenlemesi gerekmektedir [75-78]. Ayrıca, eğitim aşamasının ardından daha önce hiç karşılaşmadığı veriler içinde uygun çıkış değerleri üretir yani genelleme yapabilme özelliği vardır. YSA'lar sahip olduğu adaptasyon özelliği ile değişen problemlere ya da parametreleri değişen problemlere çözümler sunabilmek için tekrar eğitilebilir [75-78]. Hataya ya da gürültüye karşı duyarlılık toleransı YSA'larda oldukça güçlüdür. Oluşturulan sisteme girdi olarak verilen değerlerin bir kısmının sağlanamaması ya da bazı ağırlık değerlerinin bozulması gibi durumlarda YSA'lar girdi olarak verilen verilere uygun çıkış değerleri üretebilir. Yani girdi vektörünün ya da kendi yapısından kaynaklanan hataları tolere etme gücü oldukça fazladır [72-78].

5.2. Yapay Sinir Ağları Yapıları

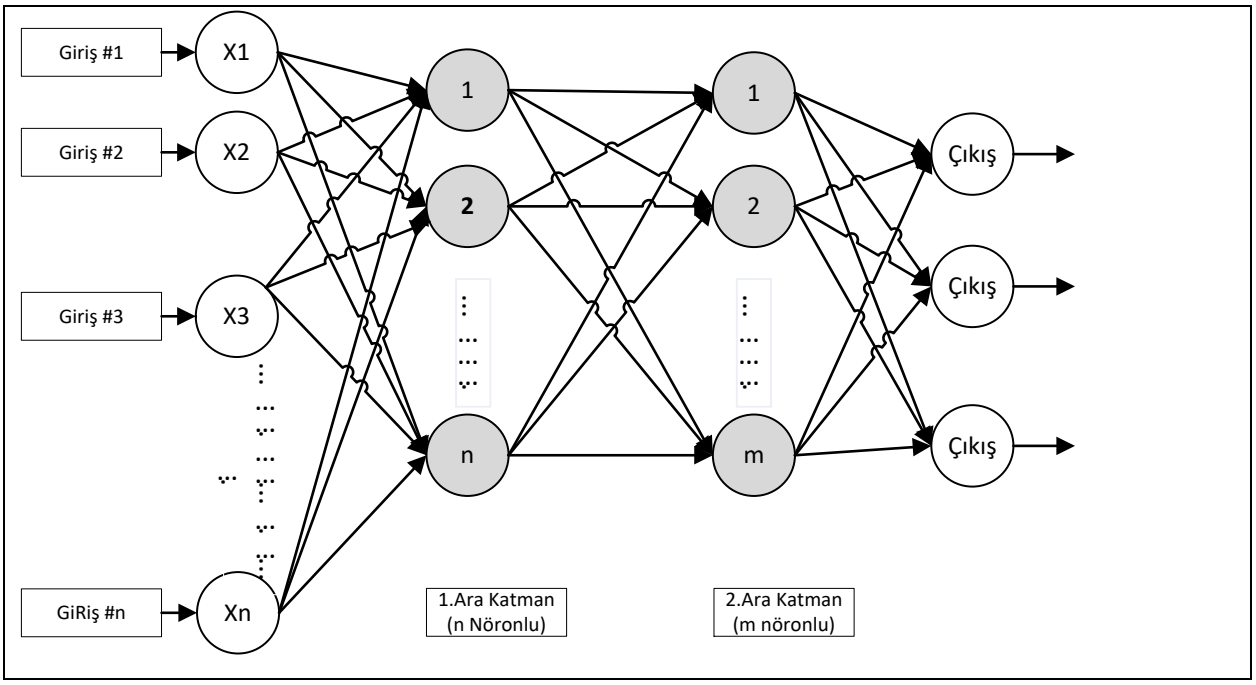
Yapay sinir ağları, yapay sinir hücrelerinin yani nöronların birbirleriyle bağlantı oluşturması sonucu oluşmaktadır. Bu bağlantıların şekli yapay sinir ağının yapısını belirlemektedir. Verilen problemin çözümünün elde edilmesi için bağlantıların nasıl değiştirilmesi gerektiğine ise öğrenme algoritmaları karar vermektedir. Seçilen öğrenme algoritması, hatayı minimize edecek şekilde ağırlıkları değiştirmektedir [76-78].

Alanyazın çalışmalarında çeşitli YSA yapıları bulunmaktadır. Yapılarına göre YSA'lar ileri beslemeli ve geri beslemeli olmak üzere ayrılmaktadır. İleri beslemeli yapay sinir ağları, giriş katmanı ile çıkış katmanı arasındaki iletişim tek yönlüdür. Nöronlar sadece başka bir katmanla bağlantı kurabilirler aynı katman içerisindeki bağlantı mümkün değildir. Geri beslemeli yapay sinir ağlarında ise çıkış veya ara katman çıkışlarının giriş katmanına ya da çıkış katmanına girdi olarak verilmesi mümkündür. Bu ağ yapıları dinamik hafızaya sahiptir. Çok katmanlı (ara katmanlı) perseptronlar (MLP), Radyal tabanlı sinir ağı (RBFNN) ya da vektör kuantamalı öğrenme (LVQ) ileri beslemeli ağlara örnek olarak verilebilir. Hopfield, Elman ve Jordan, ART (Adaptive Resonance Theory) geri beslemeli ağ yapısına sahiptir. Şekil 5.4'de ileri ve geri beslemeli ağ yapıları gösterilmektedir [74-76].



Şekil 5.4. İleri ve Geri Beslemeli Ağ Diyagramı [77].

Bu tez kapsamında kullanılan YSA yapısı ileri beslemeli ağ yapılarından biri olan çok katmanlı (ara katmanlı) perseptronlar (MLP)'dır. Bu yapının tercih edilmesinin nedeni ise, çok farklı uygulamalarda başarılı sonuçlar vermesidir. Ayrıca MLP yapısı sınıflandırma problemlerinin çözümünde de oldukça başarılı sonuçlar vermektedir [76-78]. Daha öncede bahsedildiği gibi yazar tanıma problemi aslında yazarı belli olmayan bir metinin yazarının aday yazarlar arasından bulunması problemi yani sınıflandırma problemi olmasından dolayı bu çalışmada YSA yapısı olarak MLP tercih edilmiştir. MLP'nin genel yapısı Şekil 5.5'de gösterilmektedir.



Şekil 5.5.Yapay Sinir Ağı Yapısı

Genel olarak MLP, giriş katmanı, ara katman ve çıkış katmanı olmak üzere üç ana bölümden oluşmaktadır. Katmanlarda bulunan nöronların tamamı bir sonraki katmandaki nöronların tamamına bağlıdır. İletişim tek yönlü olup geri besleme bulunmamaktadır [76]. Giriş katmanında bulunan nöronların görevi, girdi olarak alınan verilerin ara katmandaki nöronlara dağılımını sağlamaktadır. Ara katman içindeki nöronların çıkış değerleri kendisine gelen veriler ve bağlantı ağırlıklarının çarpılıp toplanması sonucudur. Bu ağ içerisindeki her bir nöronun çıkışı Eşitlik 25'deki gibi hesaplanır. Giriş katmanı içerisinde bu işlem yapılmaz [73-76].

$$y_k = f\left(\sum w_k x\right) \tag{25}$$

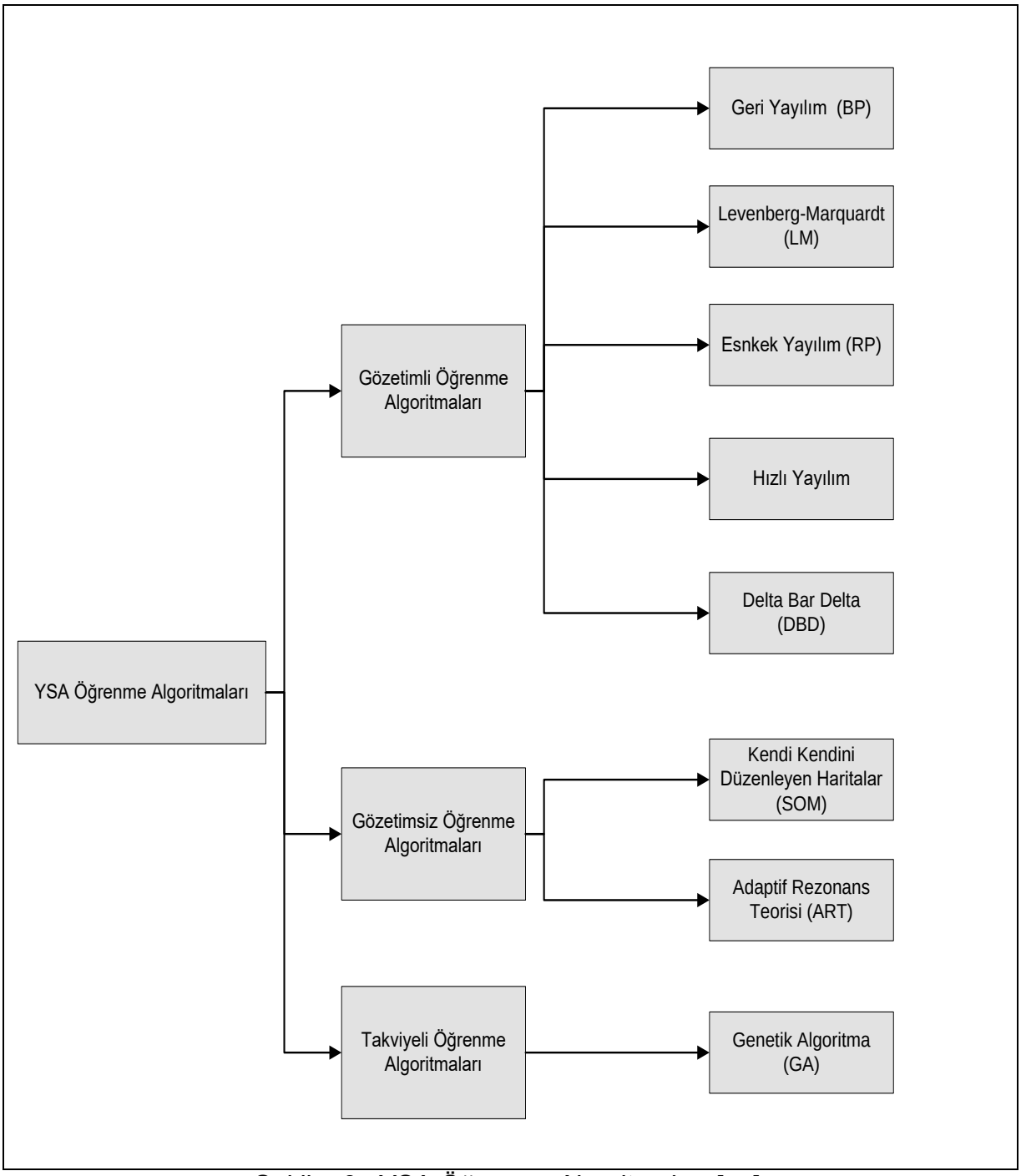
Giriş katmanındaki nöron sayısı çözülmesi istenen probleme bağlı olarak değişim göstermektedir. Kullanılacak ara katman sayısı ve ara katmandaki nöronların sayıları ise deneme yanılmaya yoluyla belirlenir. Çıkış katmanındaki nöronların sayısı da giriş katmanında olduğu gibi ele alınan probleme bağlıdır. MLP ağlarında denetimli öğrenme gerçekleştirilmektedir. YSA'a verilen girdi değerlerinin sınıfları belli olup nasıl bir sonuç üretmesi gerektiğini ağ hesaplamaktadır. Veriler giriş katmanından YSA'a uygulanır, ara katmanlarda işlenir ve çıkış katmanında ise çıkış değerleri elde edilir. Seçilen öğrenme algoritmasına göre, çıkış değeri ile asıl istenen değer arasındaki fark hata olarak adlandırılır ve geriye doğru yayılarak bu hata değeri en aza indirilene kadar sinir ağının ağırlık değerleri değiştirilir [74-76].

5.3. YSA Öğrenme Algoritmaları

Hebb 1949 yılında "*Bir nöron diğer bir nörondan giriş alıyorsa ve her iki nöronda aktif ise nöronlar arasındaki ağırlık kuvvetlendirilir*." düşüncesini ileri sürmüştür. Alanyazın çalışmalarındaki birçok öğrenme algoritması da Hebb kuralı olarak adlandırılan bu düşünceden ilham almıştır. Bu öğrenme algoritmalarının neredeyse tamamı matematiksel tabanlıdır ve ağırlıkların güncellenmesi için kullanılmaktadır. Öğrenme algoritmaları gözetimli, gözetimsiz ve takviyeli öğrenme olmak üzere üçe ayrılır. Sinir ağlarının neyi öğrenmesi gerektiğini bildiği yani çıkış değerinin belli olduğu öğrenme şekli gözetimli öğrenmedir. Bu öğrenme türünde, gerçek çıkış değeri ile sinir ağının ürettiği çıkış değeri arasındaki hataya göre nöronlar ağırlık değerlerini değiştirir ve hata minimize edilene kadar bu işlem devam etmektedir [75,76].

Gözetimsiz öğrenme algoritmasında ise, sinir ağı sınıflandırma kurallarını giriş katmanına verilen verilerden elde edilen çıkış değerine göre oluşturmaktadır. Bu öğrenme türünde asıl çıkış değerinin bilinmesine ihtiyaç yoktur. Ağırlık değerleri giriş katmanına verilen veriler arasındaki matematiksel ilişkilere göre ayarlanır. Çıkış değerleri, aynı özelliklere sahip desenlerde aynı çıkışlar kullanılırken farklı özellikte desenlere sahip olanlar için yeni çıkış değerleri yaratılır. Son olarak takviyeli öğrenme aslında danışmanlı öğrenmenin farklı bir yapısıdır. Giriş verilerine karşılık gelen çıkış verilerinin bilinmesine ihtiyaç olmamasına rağmen YSA tarafından elde edilen çıkış

değerinin giriş verilerine uygun olup olmadığını değerlendiren bir kriter bulunmaktadır [75,76].



Şekil 5.6. YSA Öğrenme Algoritmaları [77].

Bu tez kapsamında YSA modelini eğitmek için Levenberg Marquardt (LM), Gradient Descent with Momentum (GDM), Gradient Descent with Adaptive Learning Rate Back Propagation Momentum (GDA) öğrenme algoritmaları kullanılmaktadır. Bu algoritmalar ile ilgili detaylı bilgi aşağıda anlatılmaktadır.

5.3.1. Momentumlu Geri Yayılım Algoritması (BackPropagation)

Çeşitli öğrenme algoritmalarının bulunmasına rağmen Momentumlu Geri yayılım algoritması uygulamalarda en çok tercih edilen algoritmalardan birisidir. Bu algoritmanın matematiksel olarak kanıtlanması kolay ve anlaşılabilir olması tercih edilmesine sebep olmaktadır. BP hataları çıkıştan girişe doğru minimize etmeye çalışmasından nedeniyle geri yayılım adını almıştır. Çok katmanlı perseptronlar (MLP)'ın eğitmek için sıklıkla tercih edilir. Eğitme ve eğitim aşamasından sonraki test aşamasının belli adımları bulunmaktadır [72-76]. Bu algoritmaların adımları Şekil 5.7'de verilmektedir.

Eşitlik 26'de i ve j kat nöronları arasındaki ağırlık değişimi hesabı gösterilmektedir. η öğrenme katsayısı, α momentum katsayısı, δ_j ise ara ya da çıkış katmanında bulunan j nöronuna ait olan bir faktördür. Bu vektör çıkış katmanı için Eşitlik 28.'deki gibi hesaplanır ve $y_i^{(t)}$ j nöronun hedef çıkışıdır.

$$\Delta w_{ij}\,(t) = \eta \delta_j\, x_i + \alpha \Delta w_{ij}\ (\mathrm{t}-1) \tag{26}$$

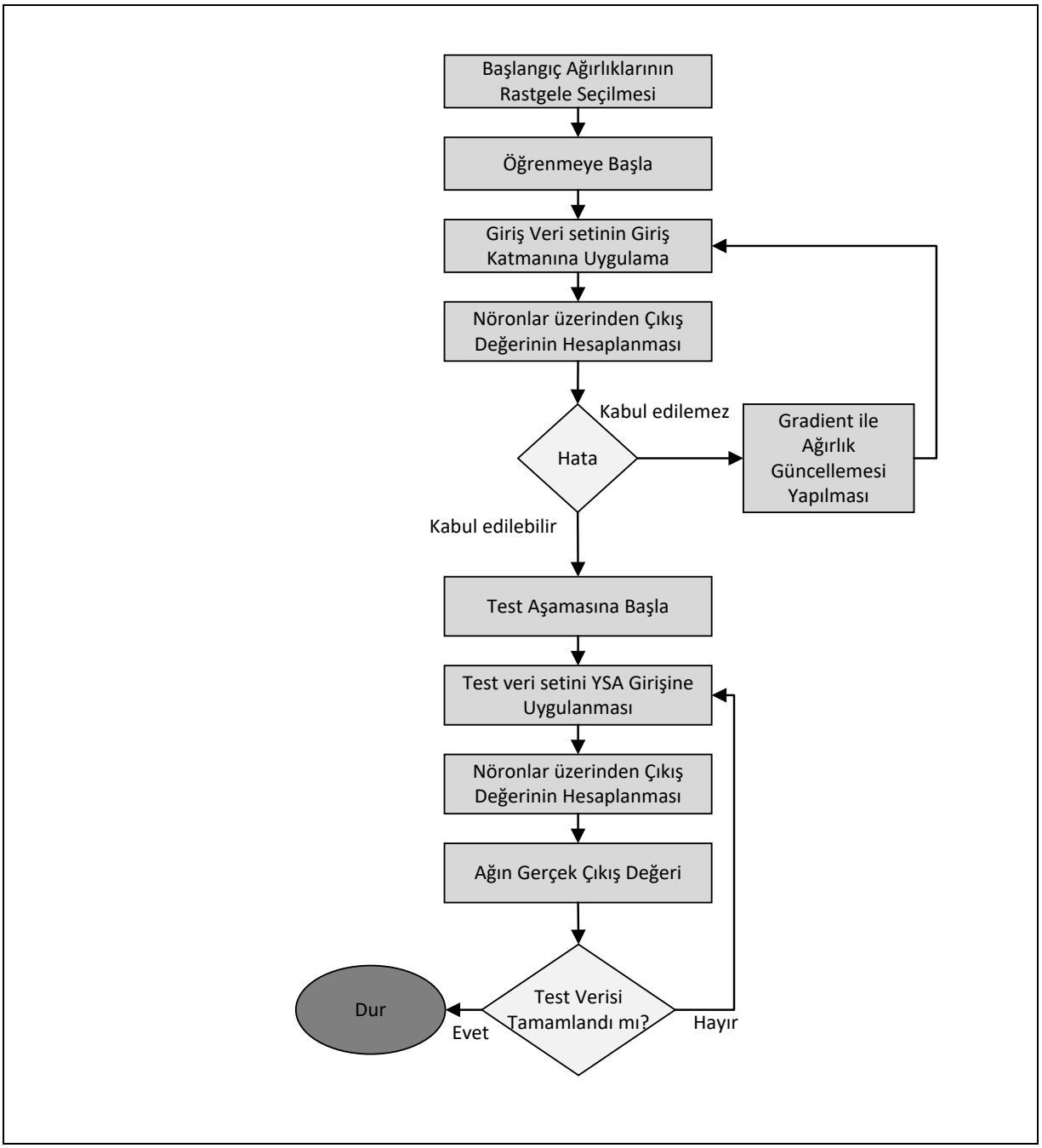
$$\delta_j\ = \frac{\partial f}{\partial net_j}\left(y_i^{(t)} - y_j\right) \tag{27}$$

$$net_j = \sum x_j\ w_{ji} \tag{28}$$

Ara katmanlar için bu vektör hesabı Eşitlik 29'da gösterildiği gibidir. Bu katmanda istenilen bir çıkış değeri olmadığından dolayı hesaplanması çıkış katmanından farklıdır.

$$\delta_j\ = \frac{\partial f}{\partial net_j} \sum w_{qi}\delta_q \tag{29}$$

Başlangıç noktası çıkış noktası olan δ_j faktörü sinir ağındaki tüm nöronlar için hesaplanır ve nöronların sahip olduğu ağırlık değerleri Eşitlik 26 kullanılarak yeniden hesaplanır.



Şekil 5.7. Momentumlu Geri Yayılım Algoritması [77].

Genellikle MLP'lerin eğitiminde kullanılan bir algoritma olan BP bir dereceli azaltma algoritması olup temel prensibi beklenen çıkış değeri ile elde edilen çıkış değeri arasındaki hatanın yani farkın ağırlıklara bağlı olarak azaltılmasına dayanır [77].

5.3.2. Levenberg-Marquardt Öğrenme Algoritması

Gauss-Newton ve Steepest-Descent algoritmalarının en iyi özelliklerinden oluşan ve bu algoritmaların kısıtlamalarını ortadan kaldıran bu öğrenme algoritmasının temeli maksimum komşuluk üzerine kurulmuş en kareler hesaplama yöntemidir. Bu yöntem yavaş yakınsama probleminden etkilenmez [74-76]. Eşitlik 30'da $E(w)$ amaç hata fonksiyonunun hesaplanması gösterilmektedir.

$$E(w) = \sum_{i=1}^{n} {e_i}^2(w) = \|f(w)\|^2 \tag{30}$$

$$e_i(w) = (y_i - yd_i)^2 \tag{31}$$

Bu eşitlikte, amaç fonksiyonu *f(.)* ve Jakobiyeni J'nin bir noktada w bilindiği varsayılır.

w (parametre vektörü)'nün $E(w)$ minimum olduğu zaman belirlenmesi Levenberg Marquardt algoritmasının amacıdır. Bu öğrenme algoritması, yeni vektör olan w_{k+1} değerini w_k'dan hesaplayarak bulur. Eşitlik 32-33'de bu adımın gerçekleştirilmesi gösterilmektedir. J_k, f yani amaç fonksiyonunun w_k değerlendirilmiş Jakobyeni iken λ Marquardt parametresi ve I ise birim ya da tanımlama matrisini ifade etmektedir.

$$w_{k+1} = w_k + \partial w_k \tag{32}$$

$$(J_K^T J_k + \ \lambda I\ )\delta w_k = (w_K) \tag{33}$$

Genel olarak LM algoritmasının adımları aşağıda gösterildiği gibidir.

1. $E(\ w_k)$ hesaplamasının yapılması.
2. λ değerinin seçilmesi
3. ∂w_k için Eşitlik 4.5 i kullan ve $E(\ w_k + \partial w_k)$ değerini hesaplanması
4. Eğer hesaplanan $E(\ w_k + \partial w_k)$ değeri $E(\ w_k)$ değerinden büyük ve eşit durumda ise λ değerini arttır ve 3. Adıma geri dön
5. Eğer hesaplanan $E(\ w_k + \partial w_k)$ değeri $E(\ w_k)$ değerinden küçük ise λ değerini azalt ve

   $w_k;\ w_k \leftarrow w_k + \partial w_k$ güncellenir ve 3. Adıma geri dönülür.

İstenilen çıkış değerine ulaşmak için LM algoritması kullanılarak ağ ağırlıklarının eğitilmesi işlemi için ilk olarak ağırlık dizisi olan w_0'a bir başlangıç değerinin verilmesi gerekmektedir. Bu işlemden sonra hataların kareleri toplamı (SSE) yani $(e_i)^2$ hesaplanır. Yani burada istenilen çıkış değerleri ile elde edilen çıkış değerleri arasındaki farkın karesi alınması işlemi yapılmaktadır. Bu işlem kullanılan veri seti için $(e_i)^2$ hata terimlerinin tümünün elde edilmesidir [74,75].

5.3.3. Esnek Yayılım Algoritması

1993 yılında Riedmiller ve Braun tarafından geliştirilen bu öğrenme algoritması türevlerin olumsuz etkilerini eğitim sürecinden uzaklaştırmayı amaçlamaktadır [74]. MLP ağ yapılarında sıklıkla tercih edilen sigmoid fonksiyonlar sinir ağlarına giriş olarak verilen sonsuz aralıktaki verileri sınırlı bir aralığa yerleştirmektedir. Gerçekleştirilen bu işlemden dolayı sigmoid fonksiyondan sıkıştırıcı (squashing) fonksiyon olarak da bahsedilmektedir. Fonksiyonun bu özelliği bias ve ağırlık değerleri optimum değere ulaşmamışken eğim değeri üzerindeki değişimin az olması karşısında, yüksek eğim azaltma ile öğrenmede sorunlara neden olabilmektedir [76].

Ağırlıkların güncellenmesi işlemi için sadece türev değerlerinin işaretlerinin kullanılması yani türev değerlerinin önemsiz olması bu algoritmayı diğer öğrenme algoritmalarından ayıran en önemli özellik olmaktadır. Ayrıca, bu özelliği sayesinde problemlerin çözümünde hızlı olmaktadır [74-76].

Ağırlık değerlerinin değişimi n hesaplanması için Eşitlik 34 kullanılmaktadır.

$$\Delta w_{ij}(k) = \begin{cases} -A_{ji}(k), eğer\ B(k) > 0, \\ +A_{ji}(k), eğer\ B(k) < 0, \\ 0, \text{ya da} \end{cases} \quad (34)$$

Eşitlikteki $B(k)$ değeri $B(k) = \frac{\partial E}{\partial w_{ij}}(k)$ ile hesaplanır.

$A_{ij}(k)$ değeri Eşitlik 35'deki gösterildiği gibi hesaplanmaktadır. Bu eşitlikteki η ve μ azaltma ve artma faktörleri olup 0< μ<1< η arasında olmaktadır.

$$A_{ij}(k) = \{\boldsymbol{\eta} A_{ij}(k-1), eğer\ B(k-1)B(k) > 0, \quad (35)$$

Ardışık iki iterasyonda, performans fonksiyonu ile türevin işaretlerinin aynı olması durumunda ağırlıklar ve bias değerleri için güncelleme için birden büyük katsayı kullanılmaktadır. Farklı işarete sahip olunması durumda ise güncelleme faktörü birden küçük katsayı faktörü kadar azaltılmaktadır. Türev değerinin sıfır olması durumunda herhangi bir değişim yapılmamaktadır. Ağırlıkların birkaç iterasyon aynı yönde değişim göstermesiyle ağırlıklardaki değişim artar. Yüksek eğim azaltma problemlerinde sıklıkla kullanılmaktadır [72-76].

5.4. YSA Modelinin Tasarlanması

Birçok problemin çözümüne olanak sağlayan YSA'ların tasarım aşaması oldukça önemlidir. YSA uygulamalarının başarısı izlenecek yaklaşımla doğrudan ilgilidir. Başarılı bir YSA modeli için karar verilmesi gereken önemli noktalar vardır. Bu noktalar,

- Ele alınan problem için uygun YSA yapısının seçimi (İleri Beslemeli YSA ya da Geri Beslemeli YSA)

Seçilen YSA yapısı için uygun olan öğrenme algoritmasının (LM, GDA, GDM gibi) belirlenmesi

- Bu öğrenme algoritması için uygun parametre seçiminin yapılması
- Seçilen yapıya uygun ve problemin çözümü için uygun girdi sayısının belirlenmesi, girdi sayısının az olması YSA modellemesinin doğru yapılmamasına neden olabileceği gibi girdi sayısının fazla olması durumunda ise YSA modeli hatalı kestirimlerde bulunabilir.
- Uygun ara katman ve ara katmandaki nöron sayısının belirlenmesi
- Aktivasyon fonksiyonun belirlenmesi
- YSA'nın öğrenme aşamasında kullanacağı eğitim verisi ve YSA'nın başarısının değerlendirileceği test veri setlerinin belirlenmesi
- Eğitim veri seti ve test veri seti için kullanılacak olan normalizasyona karar verilmesi
- YSA eğitim aşamasındaki iterasyon sayısına karar verilmesi, çok iterasyon ile eğitim süresi gereksiz uzamış olurken iterasyon sayısının az seçilmesi durumunda ise YSA modeli yeterince öğrenemez.

Bu işlemlerin tamamı zorlu ve zaman alıcı olabilmektedir. Fakat bu noktalarda doğru karar verilmemesi durumunda YSA performansı ciddi şekilde etkilenmektedir. Her problem için uygun parametrelerin seçilmesi ile oluşturulan YSA modeli kararlı bir yapı ortaya çıkaracağından dolayı elde edilen sonuçlarda buna bağlı olarak kararlı olacaktır. Kararlı bir YSA modeli oluşturmak için dikkat edilmesi gereken birçok parametre vardır. Teoride bu parametrelerden uygun olanının seçilmesi işlemi olanaklı gibi gelse de bu parametreleri öngörmek hiç kolay bir işlem değildir. Bu parametrelere karar verebilmek için daha önce yapılan çalışmalardan yararlanılması izlenmesi gereken yaklaşımdır. Daha önce yapılan çalışmalarda ele alınan problem için sunulan çözümler iyi bir başlangıç noktası olmasına rağmen YSA mimarisinin yani yapısının ele alınan problem dikkate alınarak seçilmesi gerekmektedir. Ayrıca mimari belirlenirken veri yapısının da göz önünde bulundurulması gerekmektedir. Ele alınan problem sınıflandırma problemi ise ağ yapısının MLP olarak seçilmesi iyi bir yaklaşımdır. Ağın karmaşıklığının azaltılması için en başarılı çözüm yapının değiştirilmesidir. Çünkü bir ağ gereğinden fazla nöron içeriyorsa genelleme yeteneğinde buna bağlı olarak düşmektedir.

Uygun ağ yapısının belirlenmesinden sonra öğrenme algoritmasının seçilmesi adımına gelmektedir. Bu adımda seçilen ağ yapısı öğrenme algoritmasının belirlenmesinde önemli rol oynamaktadır. Aslında, ağ yapısı içinde kullanılabilecek öğrenme algoritmaları ağın mimarisiyle doğrudan ilgilidir.

Verilerin oluşturulan YSA modeline giriş olarak verilmeden önce normalize edilmesi gerekmektedir. Bu işlem için çeşitli normalizasyon yöntemleri bulunmaktadır. Bu yöntemlerden uygun olanın seçilmesi işlemi YSA başarısını etkilemektedir. Genellikle [0,1] veya [-1,+1] aralık değerlerine veriler ölçeklendirilmektedir. Aktivasyon fonksiyonları da veri ölçeklemede kullanılabilmektedir.

[0,1] arasında ölçekleme işleminin yapılması için ilk olarak veri kümesi içinden minimum ve maksimum veri değerlerinin bulunması gerekir. Eşitlik 36'da normalizasyon formülü gösterilmektedir.

$$X_{yeni} = \frac{X - X_{min}}{X_{max} - X_{min}} \tag{36}$$

[-1,1] arasında ölçekleme işlemi Eşitlik 37'de gösterilmektedir. Normalizasyon yapılması için yukarıdaki normalizasyon yönteminde olduğu gibi minimum ve maksimum değerler bulunmaktadır.

$$X_{yeni} = 2\frac{X - X_{min}}{X_{max} - X_{min}} - 1 \tag{37}$$

Uygun YSA modeli oluşturma aşamasında dikkat edilmesi gereken bir diğer unsur ise ara katman sayısı ile bu ara katmanlarda kullanılacak olan nöron sayılarına karar verilmesidir. Bu problem için çeşitli yaklaşımlar bulunmaktadır. Bu yaklaşımlar özel problemlerin çözümünde ise yetersiz kaldığı için izlenecek en iyi yol deneme yanılma yoluyla karar vermektir. Alanyazın çalışmalarında en fazla iki ara katman kullanılması yönünde hipotezler bulunmasına rağmen ele alınan probleme göre bu değişikli göstermektedir. Uygun YSA modeli oluştururken asıl hedef en sade yapıda model tasarlanmasıdır. Yapının karmaşıklığının artması beraberinde birçok olumsuzluğu getirmektedir.

YSA modeli öğrenme aşamasında çeşitli hata fonksiyonları kullanmaktadır. Bu hata fonksiyonları öğrenme performansını etkilemektedir. Kullanılan fonksiyonlar ele alınan problemin türüne göre farklılık göstermektedir. İleri beslemeli ağ yapısında en sık kullanılan hata fonksiyonu olan karesel ortalama hata (MSE) fonksiyonudur.

$$MSE = \frac{1}{N}\sum_{İ=1}^{N}(t_i - td_i)^2 \tag{39}$$

Sık kullanılan bir diğer hata fonksiyonu ise toplam karesel hata (SSE) fonksiyonudur.

$$SSE = \sum_{İ=1}^{N}(t_i - td_i)^2 \tag{40}$$

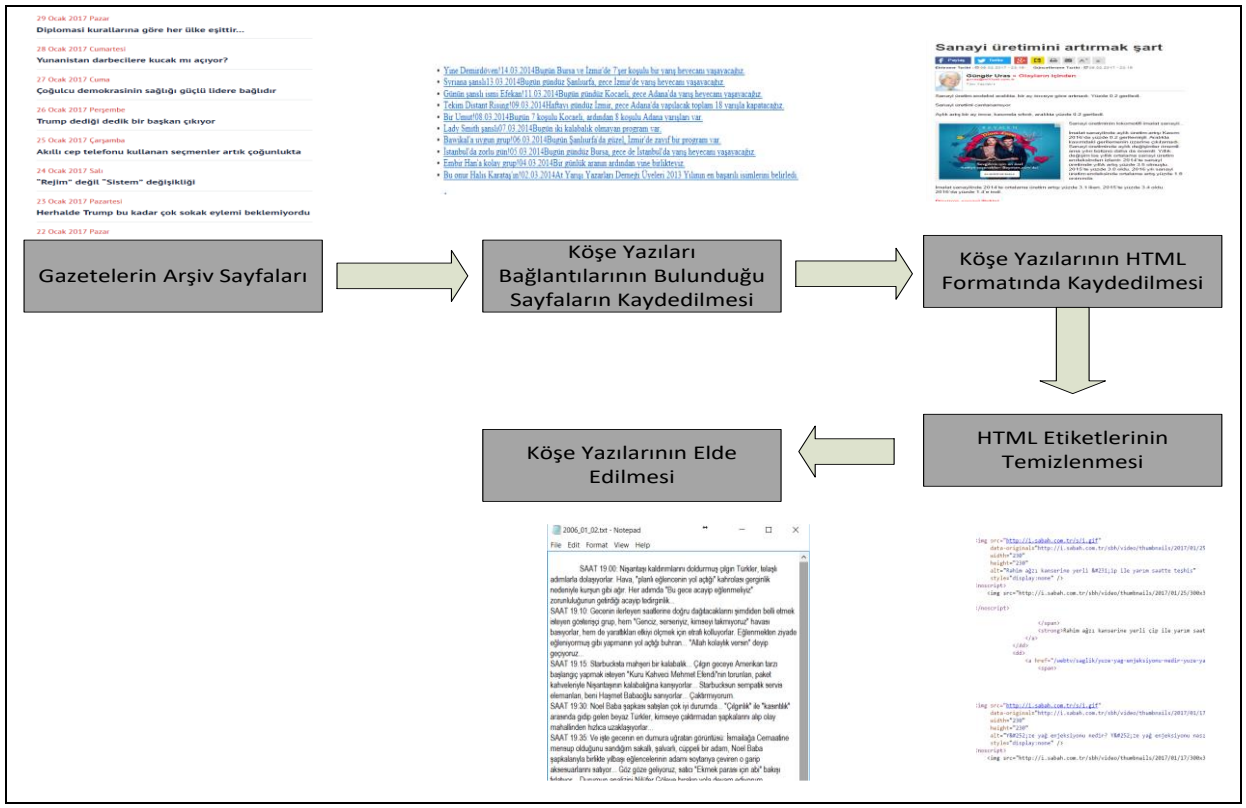
Tercih edilen fonksiyonlardan bir diğeri ise karesel ortalama hata (RMS) karekökü'dür.

$$RMS = \sqrt{\frac{1}{N}(t_i - td_i)^2} \tag{41}$$

6. KÖŞE YAZILARI VERİLERİ

6.1. Verilerin Elde Edilmesi

Yazar tanıma alanında yapılan ilk çalışmalarda derlem olarak genellikle edebi kitaplar kullanılmıştır. Günümüzde gerçekleştirilen çalışmalarda ise köşe yazıları, blog yazıları, tweetler ya da e postalar kullanılmaktadır. Başta İngilizce olmak üzere farklı dillerde gerçekleştirilen çalışmalar incelendiği zaman derlem olarak kullanılacak çeşitli veri setlerine rastlanmaktadır. Türkçe dili için böyle herhangi bir derleme rastlanmaktadır. Bu tez kapsamında Türkçe dili için gerçekleştirilecek çalışmalarda kullanılması amacıyla geniş bir kullanım alanına sahip olacak bir derlem oluşturulması amaçlanmıştır. "Türkçe Yazar Tanıma Derleminin" elde edilmesi için cevrim içi gazetelerde siyaset, ekonomi, yaşam ve spor alanlarında yayınlanan köşe yazıları tercih edilmiştir. Köşe yazılarının tercih edilmesinin sebebi kişilerin duygu düşünce ve fikirlerini özgürce kaleme aldığı bir yazı şekli olması ve köşe yazılarının belli bir rutin içinde genellikle her gün yazılmasıdır.



Şekil 6.1. Yazılımların Genel Yapısı

Hürriyet, Sabah, Milliyet, Vatan gibi günlük çıkan gazetelerin belirli bir köşe yazısına erişmek isteyen kişi ekran üzerinde bulunan yönlendirmelerle istediği köşe yazısını okuyup saklayabilmektedir. Fakat birçok köşe yazısının bu şekilde yani elle (manuel) toplanması zor ve zaman alıcı bir işlemdir. Köşe yazılarının toplanarak derlem oluşturulması için Java ortamında her gazete için farklı yazılımlar geliştirilmiştir. Bu yazılımların genel akış diyagramı Şekil 6.1'de gösterilmektedir.

Köşe yazılarının elde edilmesi için her gazetenin URL (Tekdüzen Kaynak Bulucu) yapısı incelenmiştir. Hürriyet, Milliyet, Sabah, Cumhuriyet, Vatan, Posta, Fanatik, MilliyetSpor cevrim içi gazetelerinde yazılar yazan yazarların köşe yazılarının bulunduğu sayfaların URL yapısı Şekil 6.2'de gösterilmektedir.

Çizelge 6.1. Gazetelerin URL yapıları

Gazeteler	**Arşiv URL Yapıları**
Sabah	http://www.sabah.com.tr/yazarlar/AuthorSurname/arsiv?getall=true&page=2
Vatan	http://home.gazetevatan.com/vatan2013/yazarlar-detay.asp?wid=" + YAZAR_ID + "&@@=";
Cumhuriyet	http://www.cumhuriyet.com.tr/koseyazari/Author_ID/YazarAdı.html
Milliyet	http://milliyet.com.tr//Milliyet.aspx?aType=siyasetYazarTumYazilarV4 &AuthorID="+YAZAR_ID+"&PAGE=";
MilliyetSpor	http://skorer.milliyet.com.tr/Milliyet.aspx?aType=SkorerYazarDetayTumu& AuthorName="+YAZAR_ADI+"&AuthorID=0&PAGE=1
Hürriyet	http://sosyal.hurriyet.com.tr/Yazar/"+YazarAdi"_"+YazarSoyadi"_"+YazarID+ "/tum-yazilari"
Fanatik	http://www.fanatik.com.tr/default.aspx?aType=YazarTumYazilar&AuthorID +Author_ID+"&TotalItems="+Article_Count+"&Page=";

Gazetelerin URL sayfalarının farklı yapılarda bulunmasından dolayı köşe yazılarına ulaşma işlemi oldukça zaman alıcı olmaktadır. Ayrıca her gazete için farklı olan bu yapı gazetelerin farklı zaman aralıklarında bile değişim göstermektedir. Bu zaman aralıklarının belirlenmesi ve buna uygun olarak yazılan yazılımın modifiye edilmesi ve bazı durumlarda tekrardan yazılması gerekmektedir. Bu problemler göz önünde bulundurularak Java ortamında farklı yazılımlar geliştirilmiştir. Bu yazılımlarla

gazetelerin köşe yazılarının bulunduğu sayfalara erişim ve bu sayfaların kaydedilmesi işlemi gerçekleştirilmiştir.

Köşe yazılarının linklerinin bulunduğu sayfaların ayrı ayrı kaydedilmesi işleminin ardından bu sayfalar içinden köşe yazılarının linklerinin elde edilmesi için Şekil 6.3'deki kod parçası kullanılmıştır.

```
public static Vector extractLinks(String rawPage, String page) {
          int index = 0;
          Vector links = new Vector();
          while ((index = page.indexOf("<a ", index)) != -1)
          {
              if ((index = page.indexOf("href", index)) == -1) break;
              if ((index = page.indexOf("=", index)) == -1) break;
              String remaining = rawPage.substring(++index);
              StringTokenizer st
                      = new StringTokenizer(remaining, "\t\n\r\"'>#");
              String strLink = st.nextToken();
                if (! links.contains(strLink)) links.add(strLink);
          }
          return links;}
```

Şekil 6.2. Köşeyazılarının Bağlantılarının Kaydedilmesi

Köşe yazılarının sayfalarının linkleri elde edildikten sonra bu sayfalar HTML formatında sırayla kaydedilmektedir. HTML formatında olan web sayfaları içerisinde reklam menüsü, başlık, diğer bölümlerle bağlantılarını içeren linkler ve köşe yazısının başlığı, paragraf yapısı gibi bileşenlerini düzenleyen HTML imleri (tag) bulunmaktadır. Bu HTML imlerinin yerleri gazeteden gazeteye farklı olmasının yanında aynı gazete içinde yıllara hatta bazen de aylara göre farklılık göstermektedir. Köşe yazıları sayfalarından yalnız köşe yazılarını elde etmek için her gazetenin hem farklı yıllarda hem de farklı aylarda kullanılan köşe yazısı sayfa formatının incelenmesi gerekmektedir. Bu problemi çözmek için her gazete için farklı ayrıştırıcı yazılımlar yazılmıştır. Gerekli olan durumlarda gazetelerin farklı yılları ve farklı ayları içinde sayfa yapısına göre ayrıştırıcı yazılımlar güncellenmiştir. Köşe yazılarının elde edilmesinde karşılaşılan önemli problemlerden bir diğeri ise Türkçe metinlerin UTF-8, ISO-8859-9 ya da Win-1254 gibi farklı karakter kodlamalar (encoding) ile saklanmasıdır.

Köşe yazılarının elde edilmesi işleminden sonra bu elde edilen yazıların uygun formatta kaydedilmesi gerekmektedir. Elde edilen köşe yazılarından yazar tanıma da

kullanılacak olan özelliklerin çıkarılması işlemi için bölüm 6.2'de ayrıntılı anlatılan bir yazılım kullanılmaktadır. Bu yazılımı kullanabilmek için kaydedilen köşe yazılarının belli bir formatta olması gerekmektedir. Kaydedilen metinlerin köşe yazılarının yayınlandıkları tarihe göre kaydedilmesi gerekmektedir. Bu format YYYY_AA_GG (yıl, ay, gün) şeklindendir. Tarih bilgisi de gazetelerin sayfa yapılarında farklı yerlerde bulunmaktadır. Tarih bilgisinin köşe yazısı sayfasından çıkarılması için HTML imlerinden yararlanılmaktadır. Şekil 6.3'de Vatan gazetesi için yapılan kodlama örneği gösterilmektedir.

```
String tarihKriter = "datePublished";
                    // tarihi al
                    do {
                          line = in.readLine();
                    }
                    while (line != null && !line.contains(tarihKriter));
                    int basIndis = line.indexOf(tarihKriter) +
tarihKriter.length() + 2;
                    String tarih = line.substring(basIndis, basIndis +
"01.01.2011".length()).replace('.', '_');
                    tarih = tarih.substring(6) + "_" + tarih.substring(3, 5)
+ "_" + tarih.substring(0, 2);
```

Şekil 6.3. Tarih Bilgisine Erişim İçin Kullanılan Kod parçacığı

Tüm bu işlemlerden sonra yazıları indirilen köşe yazarları derlemi yaşam, siyaset, ekonomi ve spor olmak üzere dört yazı alanına ayrılmıştır. Yazarların yazdıkları köşe yazılarının hangi alanlarda olduğunun belirlenmesi için gazetelerin yazarlar için belirledikleri alanlar kullanılmıştır. Böylece nesnel bir sınıflandırma yapılmıştır. Siyaset alanında yaklaşık 65 yazarın, yaşam alanında 35 yazarın, ekonomi alanında 15 yazarın ve spor alanında 25 yazarın köşe yazıları indirilmiştir. Yapılan çalışmalar incelendiği zaman Türkçe dili için daha önce böyle kapsamlı bir derleme rastlanılmamıştır. Yaşam alanında yazılardan oluşan derlem Çizelge 6.2'de gösterilmektedir. Siyaset, ekonomi ve spor alanlarında indirilen köşe yazıları ile ilgili bilgi Ek 1'de verilmektedir.

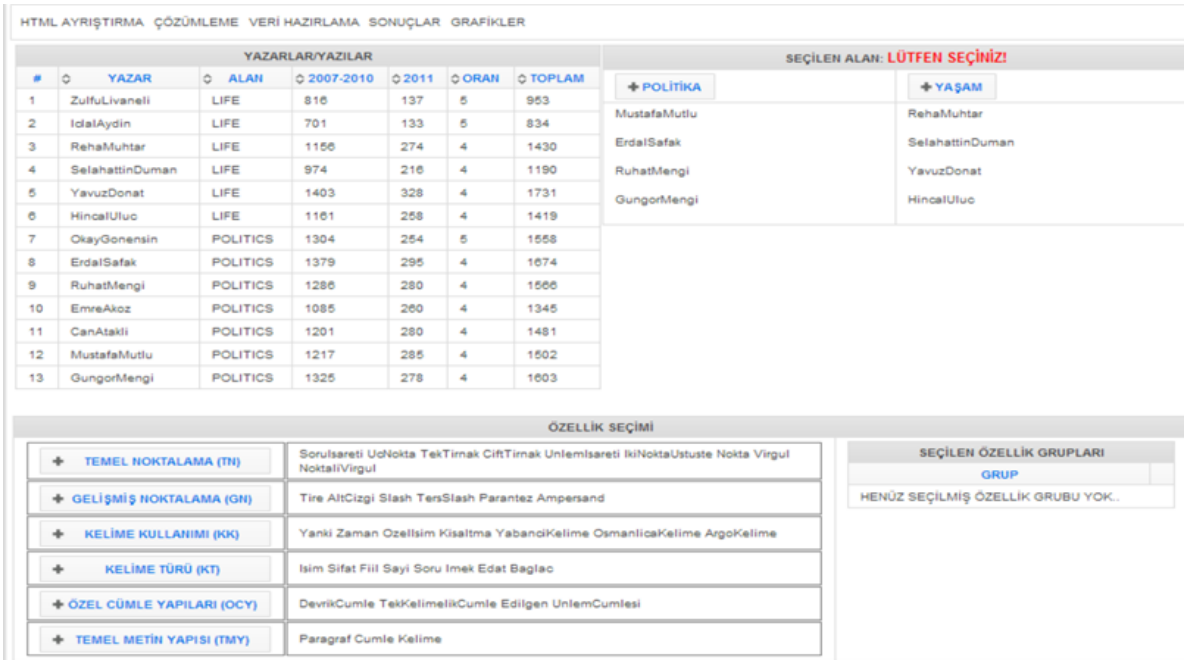
Çizelge 6.2. Yaşam Alanında Elde Edilen Köşe Yazıları

Gazete Adı	Yazar Adı Soyadı	Başlangıç		Son		ALAN
		Yıl	Ay	Yıl	Ay	
HÜRRİYET	**ArmağanCaglayan**	2004	1	2010	12	Yaşam
	Şükrü Kızılot	2002	1	2014	12	Yaşam
	Melis Alphan	2010	1	2014	12	Yaşam
	Ömür Gedik	2004	1	2014	12	Yaşam
	Gülse Birsel	2010	1	2014	12	Yaşam
	Erkan Celebi	1997	1	2005	12	Yaşam
	Doğan Hızlan	1997	1	2014	12	Yaşam
	Ayşe Arman	1998	1	2014	12	Yaşam
POSTA	**Ferhan Kaya Poroy**	2009	1	2014	12	Yaşam
	Fatih Öztürk	2009	1	2014	12	Yaşam
	Tamer Heper	2009	1	2014	12	Yaşam
	Mesut Yar	2009	1	2014	12	Yaşam
CUMHURİYET	**Deniz Kavukçuoğlu**	2008	1	2014	12	Yaşam
	Ergin Yıldızoğlu	2008	1	2014	12	Yaşam
	Işık Kansu	2008	1	2014	12	Yaşam
	İlhan Selçuk	1997	1	2010	12	Yaşam
	Mine G. Kırıkkanat	2011	1	2014	12	Yaşam
	Oktay Akbal	2008	1	2014	12	Yaşam
	Zeynep Oral	2008	1	2014	12	Yaşam
SABAH	**SevilayYukselir**	2009	1	2014	12	Yaşam
	Savaş Ay	2003	1	2014	12	Yaşam
	Yavuz Donat	2003	1	2014	12	Yaşam
	Şelale Kadak	2003	1	2014	12	Yaşam
	Şengül Balıksırtı	2003	1	2010	12	Yaşam
VATAN	**Tuğçe Boran**	2002	1	2008	12	Yaşam
	Zülfi Livaneli	2002	1	2013	12	Yaşam
	Selahattin Duman	2002	1	2013	12	Yaşam
	Reha Muhtar	2006	1	2014	12	Yaşam
	Dilek Önder	2006	1	2014	12	Yaşam
	İclal Aydın	2002	1	2014	12	Yaşam
MİLLİYET	**Ece Temelkuran**	2008	1	2010	12	Yaşam
	Güneri Civanoğlu	2008	1	2014	12	Yaşam
	Şükrü Andaç	2012	1	2014	12	Yaşam
	Aslı Aydınbaştaş	2009	1	2014	12	Yaşam
	Abbas Güçlü	2011	1	2014	12	Yaşam

6.2. Biçimsel Özelliklerin Elde Edilmesi

Her yazarın kendine özgü yazma alışkanlığı bulunmaktadır [2-6]. Yazma alışkanlığı olarak adlandırılan kelimelerin kullanımı, paragraf ve cümle uzunlukları, metin biçimi gibi özellikler kolayca değişmez ve her yazar için farklılar göstermektedir [10-15]. Bu özelliklere yapılan çalışmalarda "yazarlık özellikleri", "yazar değişmezleri", "siber parmak izi" olarak da rastlamak mümkündür [56]. Yazar tanıma alanında yapılan çalışmalar incelendiği zaman çok çeşitli özelliklerin kullanıldığı görülmektedir [20-45]. Yapılan bu çalışmalarda kullanılan özellikler sözcüksel özellikler, sözdizimsel özellikler, yapısal özellikler, içeriğe bağlı özellikler ve kişiye özgü özellikler olmak üzere beş gruba ayrılmaktadır. Bu tez kapsamında sözcüksel ve sözdizimsel özellikler kullanılmaktadır. Bu özelliklerin başarısı 2014 yılında yapılan doktora tezi kapsamında kanıtlanmıştır [41].

Tez kapsamı içinde kullanılan biçimsel özelliklerin elde edilmesi aşamasında 2014 yılında Oğuz Aslantürk tarafından doktora tezi olarak yapılan "Tamgacı: Artırımsal ve Geri Beslemeli Türkçe Yazar Çözümleme" çalışmasında geliştirilen yazılım kullanılmaktadır [41]. Şekil 6.2'de kullanılan yazılımın ekran görüntüsü gösterilmektedir.



Şekil 6.2. Kullanılan Yazılımın Ekran Görüntüsü

Aslantürk tarafından geliştirilen bu yazılım iki ana kısımdan oluşmaktadır. İlk kısımda toplanan köşe yazılarının sisteme eklenmesi gerçekleştirilmektedir [41]. Köşe yazılarının sisteme eklenmesi aşamasında kullanıcı sisteme kendi kullanıcı adı, soyadı ve şifresi ile giriş yapmaktadır. Sisteme giriş yapan kullanıcı gazete adı, yazar adı ekle, alan ekle gibi seçenekleri kullanarak eklemek istediği köşe yazarının gazetesini ve alan bilgilerini sisteme eklemektedir. Bunun nedeni, aynı anda farklı gazetelerde yazılar yazan köşe yazılarının bulunmasıdır. Bu adımdan sonra sisteme eklenecek olan yazarın ve yazılarının daha önce sistemde olup olmadığı sistem tarafından kontrol edilmektedir. Eğer eklenmesi istenen köşe yazıları daha önce sisteme eklendiyse sistem yazıların eklenmesine izin vermez. Eğer daha önce sistemde yazılar yok ise yazıların ekleme işlemi yapılır. Köşe yazılarının sisteme eklenmesi için yazıların text formatında olması ve ayrıca eklenecek olan dosyanın sıkıştırılması gerekmektedir.

Sisteme yazılar yüklendikten sonra biçimsel özelliklerin elde edilmesi için sistemden yazar tanıma için kullanılacak olan yazar özellikleri seçimi yapılır. Bu çalışma içerisinde kullanılan köşe yazılarının belirli bir formatta yazılmasından dolayı yapısal özellikler olarak adlandırılan yazı tipi, başlık gibi özellikler kullanılamaz. Bu format yazarlar tarafından değil web tasarımcısı tarafından belirlenir. Ayrıca, gramer hataları, yazım hataları gibi kişiye özgü olan özelliklerde bu çalışma içerisinde kullanılamaz. Bu tür yazarların hataları gazetelerin editörleri tarafından düzeltilmektedir. Tüm bu noktalar göz önüne alındığı zaman çalışma içerisinde kullanılacak olan özellikler sözdizimsel ve sözcüksel özelliklerdir. Kelime zenginliğine dayalı olan özellikler işlem yükü ve zamandan dolayı kullanılmamıştır. Seçilen özellikler Çizelge 6.3'de gösterilmektedir.

Çizelge 6.3. Özellikler ve Özellik Sınıfları

Özellik	**Özellik Sınıfları**
Paragraf Sayısı	Metin Bileşenleri
Cümle Sayısı	
Sözcük Sayısı	
Edilgen Cümle Sayısı	
Tek kelimelik Cümle Sayısı	
Devrik Cümle Sayısı	
Ünlem Cümlesi Sayısı	
Nokta Sayısı	Noktalama İşaretleri
Virgül Sayısı	
Soru İşareti Sayısı	
Üç nokta Sayısı	
Tek tırnak sayısı	
Çift Tırnak Sayısı	
İki Nokta Üst üste Sayısı	
Noktalı Virgül Sayısı	
Tire Sayısı	
Alt Çizgi Sayısı	
Parantez Sayısı	
Yansıma Kelime Sayısı	
Zaman Kelimesi Sayısı	
Özel İsim Sayısı	
Kısaltma Sayısı	Kelime Türleri
Yabancı Kelime Sayısı	
İsim sayısı	
Sıfat sayısı	
Fiil Sayısı	
Sayı Kelimesi Sayısı	
Soru Kelimesi Sayısı	
Edat Sayısı	
Bağlaç Sayısı	

Belirlenen özelliklerin metin içerisinden çıkarılması için kullanılan yazılımın yapısına değinilecek olunursak, ilk olarak köşe yazıları yani metinler için belirtkeleme (tokenization) yapılır. Bu aşamada cümle içerisinde bulunan terimlerin bulunması gerçekleştirilir. Bunun için Türkçe doğal dil işleme kütüphanesi olan Zemberek kullanılmıştır. Zembereğin işleyişi için bir örnek aşağıda sunulmuştur. Metin= “Akıl ve mantığın çözümleyemeyeceği mesele yoktur.” (Mustafa Kemal Atatürk) Bu metnin içindeki kelimeler zemberek ile Şekil 6.4.‘deki gibi ayrılmaktadır.

{tip=KELIME,icerik= Akıl }, {tip=DIGER, icerik=’ ‘},
{tip=KELIME,icerik= ve }, {tip=DIGER, icerik=’ ‘},
{tip=KELIME, icerik= mantığın }, {tip=DIGER, icerik=’, ‘},
{tip=KELIME, icerik= çözümleyemeyeceği }, {tip=DIGER, icerik=’ ‘},
{tip=KELIME, icerik= mesele }, {tip=DIGER, icerik=’ ‘},
{tip=KELIME, icerik= yoktur }, {tip=DIGER, icerik=’ ‘},

Şekil 6.4. Zemberekte Kelimelerin Ayrımı

Verilen metin kelimelerine ayrılır ve analiz edilebilecek kelimelerin tipi kelime olarak belirlenir. Daha sonra kelimenin kökü ve kökünün türü (isim, fiil, edat ..), cümle içerisinde aldığı ekler için bulunan bütün olasılıklar gösterilir. Bu işlem ise Şekil 6.5’de gösterilmektedir.

Akıl*:*
[Kok: Akıl, Tip:ISIM | Ekler:ISIM_KOK]
mantığın*:*
[Kok: mantığın, Tip:ISIM | Ekler:ISIM_KOK, ISIM_TAMLAMA_IN]
çözümleyemeyeceği*:*
[Kok:çözümle, Tip:FIIL | Ekler:FIIL_KOK, FIIL_EDILGENSESLI_N, FIIL_SUREKLILIK_EREK
mesele*:*
[Kok: mesele, Tip:ISIM | Ekler:ISIM_KOK]

Şekil 6.5. Kelime Türleri

Metinler içerisinde geçen yansıma, zaman, soru, kısaltma kelimeleri Zemberek kütüphanesi kullanılarak elde edilmektedir. Ayrıca isim, fiil, sıfat, zamir gibi kelime türlerinin belirlenmesi ve metinler içerisinde geçme sayılarının bulunması için de Zemberek kütüphanesinden yararlanılmıştır. Fark edileceği gibi birçok özelliğin elde edilmesinde Zemberek kütüphanesinin fonksiyonlarından yararlanılmıştır. Seçilen bu özellikler, toplanan köşe yazılarında geçme frekansını içermektedir. Biçimsel ve sözcüksel özellikler kullanılarak vektörlere dönüştürülen köşe yazıları txt dosya formatında saklanarak kullanıcıya verilmektedir. Yazar tanıma için kullanılan bu özellikler ayrıca yazı türü (alanı) belirlenmesi içinde kullanılmıştır. Çizelge 6.4.'de Vahap Munyar'ın yazılarından elde edilen biçimsel özellikler gösterilmektedir.

Çizelge 6.4. Vahap Munyar'ın Köşe Yazılarından Elde Edilen Özellikler

Köşe Yazısı ID	**Virgül Sayısı**	**Özel İsim Sayısı**	**İsim Sayısı**	**Kelime Sayısı**	**.**	**Büyük Harf Sayısı**
#1	39	0	109	527	.	320
#2	34	4	168	846	.	449
#3	88	7	253	1066	.	717
#4	34	3	142	732	.	423
#5	31	12	153	789	.	435
#6	40	10	147	838	.	446
#7	45	6	226	1017	.	620
#8	33	7	130	764	.	392
#9	33	8	139	788	.	418
.	.	.	.	.	.	.
#100	50	25	148	854	.	512

7. GELİŞTİRİLEN MODELLER

Bu tez çalışması kapsamında yazarı belli olmayan ya da yazarından şüphe duyulan bir yazının yazarının belirlenmesi için farklı yazar tanıma modelleri ile bir yazının ya da metnin türünü belirlemek için YSA modelleri geliştirilmiştir. Son olarak ise bir metnin ya da yazının hem yazarını hem de türünü belirleyen hibrid bir ANN modeli önerilmektedir. Bu tez kapsamı içerisinde yapılan çalışmalar üç bölümde incelenmektedir. İlk olarak farklı ihtiyaçlara cevap vermek için önerilen yazar tanıma YSA modelleri anlatılmaktadır. İkinci kısımda ise yazı türü tanıma için önerilen modelden bahsedilmektedir. Son bölümde bir yazının hem yazarını hem de yazı alanını tanıyan hibrid YSA modeli anlatılmaktadır.

7.1. Yazar Tanıma Modelleri

Yazar tanıma problemi tekli ve çoklu sınıflandırma problemi olarak ele alınmış olup farklı ihtiyaçları karşılamaya yönelik çeşitli yazar tanıma modelleri geliştirilmiştir. Yazar tanıma modellerinin geliştirilmesi için ilk olarak indirilen köşe yazarlarının yazıları kullanılarak bir derlem oluşturulmuştur. Önerilen yazar tanıma modellerin yazı alanından (türü) bağımsız olmasını göstermek için farklı alanlarda yazılar yazan köşe yazarları seçilmiştir. Bu kriterler göz önüne alınarak yapılan değerlendirmeler sonucunda ‘Yaşam’, ‘Siyaset’ ve ‘Ekonomi’ alanlarında yazılar yazan iki yazar belirlenmiştir. Yazar tanıma için kullanılan derlem ilgili detaylı bilgi Çizelge 7.1’de gösterilmektedir.

Çizelge 7.1. Yazar Tanıma için Oluşturulan Derlem

Yazar	Alan	Yazı Adeti	Tarih Aralığı Başlangıcı	Tarih Aralığı Sonu
Hadi Uluengin	Siyaset	100	1/6/2008	5/1/2009
Emre Aköz	Siyaset	100	1/6/2008	10/1/2009
Vahap Munyar	Ekonomi	100	1/6/2008	1/1/2009
Güngör Uras	Ekonomi	100	1/6/2008	5/1/2009
İclal Aydın	Yaşam	100	1/6/2008	9/1/2009
Ayşe Arman	Yaşam	100	1/6/2008	11/1/2009

İlk olarak yazar tanıma problemi tekli sınıflandırma problemi olarak ele alınmış olup Ayşe Arman, İclal Aydın, Hadi Uluengin, Emre Aköz, Vahap Munyar ve Güngör Uras yazarları için farklı YSA modelleri önerilmektedir.

Aslantürk tarafından geliştirilen yazılım kullanılarak elde edilen biçimsel özellikler ile yazarlar arasındaki matematiksel ilişkiyi öğrenen YSA modellerinin oluşturulması bu bölümde gerçekleştirilmektedir. Ayşe Arman, İclal Aydın, Hadi Uluengin, Emre Aköz, Vahap Munyar ve Güngör Uras köşe yazarları için farklı YSA modelleri tasarlanmıştır.

YSA modellerinin eğitimi için izlenen adımlar aşağıda belirtilmiştir.

- Uygun YSA modellerinin oluşturulması
- Eğitim veri setlerinin okunması
- Okunan verilerin normalize edilmesi
- İlk ağırlık değerlerinin verilmesi
- Eğitime belirlenen epok sayısı kadar devam edilmesi
- Elde edilen çıkış değerlerinin değerlendirilmesi

YSA Model 1- Ayşe Arman

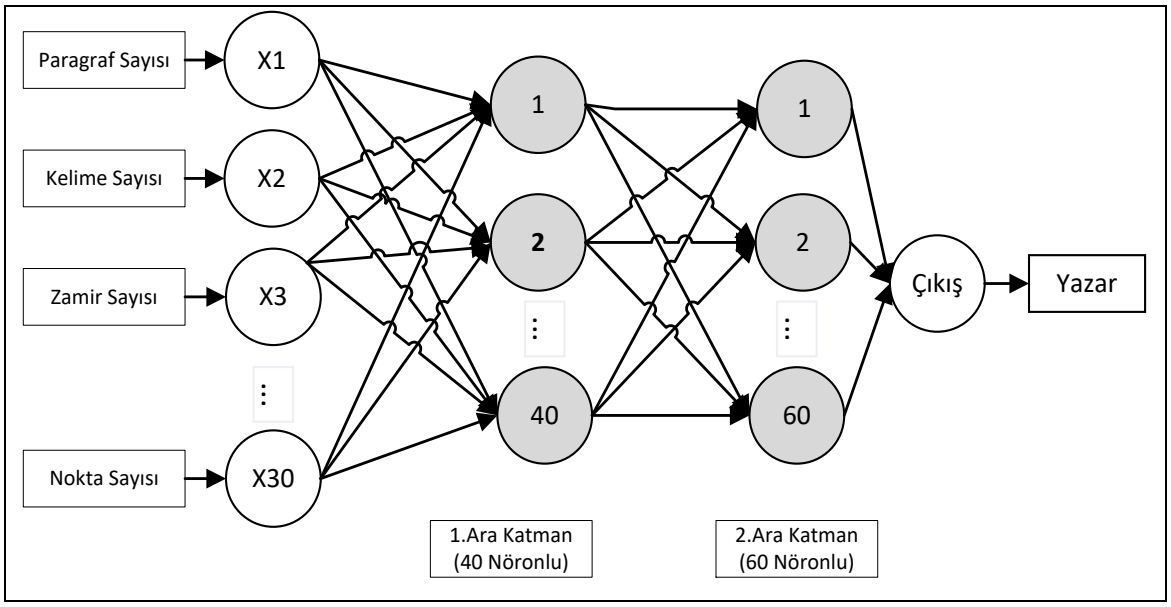
Sisteme verilen bir yazının gerçek sahibinin Ayşe Arman olup olmadığının belirlenmesi için uygun YSA modellinin geliştirilmesi gerekmektedir. Bu tez çalışması içinde kullanılan YSA modellerinin ağ mimarisi olarak ileri beslemeli ağ yapısına sahip olan MLP tercih edilmiştir. Tercih edilen bu ağ yapısının eğitim aşamasında 5 farklı öğrenme algoritması kullanılmış ve performans karşılaştırılması yapılmıştır. Ayrıca uygun YSA modelinin belirlenmesi için ara katman sayısı, katmanlardaki nöron sayıları ve geçiş fonksiyonları değiştirilerek 10 farklı YSA modeli oluşturulmuştur. Uygun YSA modelinin geliştirilmesi için Çizelge 7.2' de verildiği gibi farklı YSA yapıları deneme yanılma yoluyla test edilerek ortalama hata değerlerinin en aza indirgenmesi amaçlanmıştır. Çizelge 7.2'de en düşük hata oranını veren YSA modeli kalın (bold) olarak verilmektedir.

Çizelge 7.2. Ayşe Arman için YSA Parametrelerinin Karşılaştırılması

YSA#	AKS	HKNS	TF	ÖA	MSE
#1	2	20,40,1	S,T,LN	LM	1.07E-27
#2	2	20,40,1	S,T,LN	GDA	0.0363
#3	**2**	**40,60,1**	**S,T,LN**	**LM**	**3.47E-30**
#4	2	20,40,1	S,T,LN	GD	0.25
#5	2	25,50,1	S,T,LN	LM	3.40E-30
#6	3	20,40,60,1	S,T,T,LN	LM	5.41E-20
#7	2	20,40,1	S,T,LN	GDX	2.42E-02
#8	2	20,40,1	S,T,LN	GDM	2.50E-01
#9	2	30,60,1	T,S,LN	GD	0.25
#10	3	30,50,40,1	S,S,T,LN	LM	1.46E-22

ÖA: Öğrenme Algoritması, AKS: Ara Katman Sayısı, HKNS: Her Katmanda Nöron Sayısı, TF: Transfer Fonksiyonu,TH: Tanjant Hiperbolik, S: Sigmoid, LN: Doğrusal, MSE: Ortalama Hata Kare, LM: Levenberg-Marquardt, GD: Dereceli Azalan Geri Yayılım, GDA: Adaptif Öğrenme Oranlı Dereceli Azalan Geri Yayılım, GDX: Adaptif Öğrenme Oranlı ve Momentumlu Gradyan Azalan

Uygun YSA modelinin belirlenmesi işleminden sonra eğitim aşamasına geçilmektedir. YSA'ların eğitimi için kullanılan veri setleri yazar özelliklerinden oluşmaktadır. Bu veriler YSA'ya giriş olarak verilmeden önce normalizasyon işleminden geçirilmektedir. Böylece kullanılan verilerin belli bir aralıkta olması sağlanmaktadır. Bu tez çalışmasında max-min normalizasyon yöntemi kullanılmaktadır. Bu yöntemle veriler [0-1] aralığına getirilmektedir. Normalizasyon sonucu elde edilen veriler oluşturulan YSA yapısına giriş olarak verilir. Bu adımdan sonra, belirlenen epok sayısı kadar YSA modelinin eğitilmesi gerçekleştirilir. Her iterasyonda YSA ağırlık değerlerini değiştirerek hatayı minimize etmeye çalışır. Bu çalışma içinde YSA eğitimi için 1000 iterasyon kullanılmıştır. Son olarak, elde edilen çıktıların değerlendirilmesi yapılır. Eğitim aşaması sonucu elde edilen performans değeri 3.37e-30 olarak bulunmuştur. Şekil 7.1'de verilen YSA yapısı Ayşe Arman'ın yazılarının tanınması için oluşturulmuş bir modeldir.

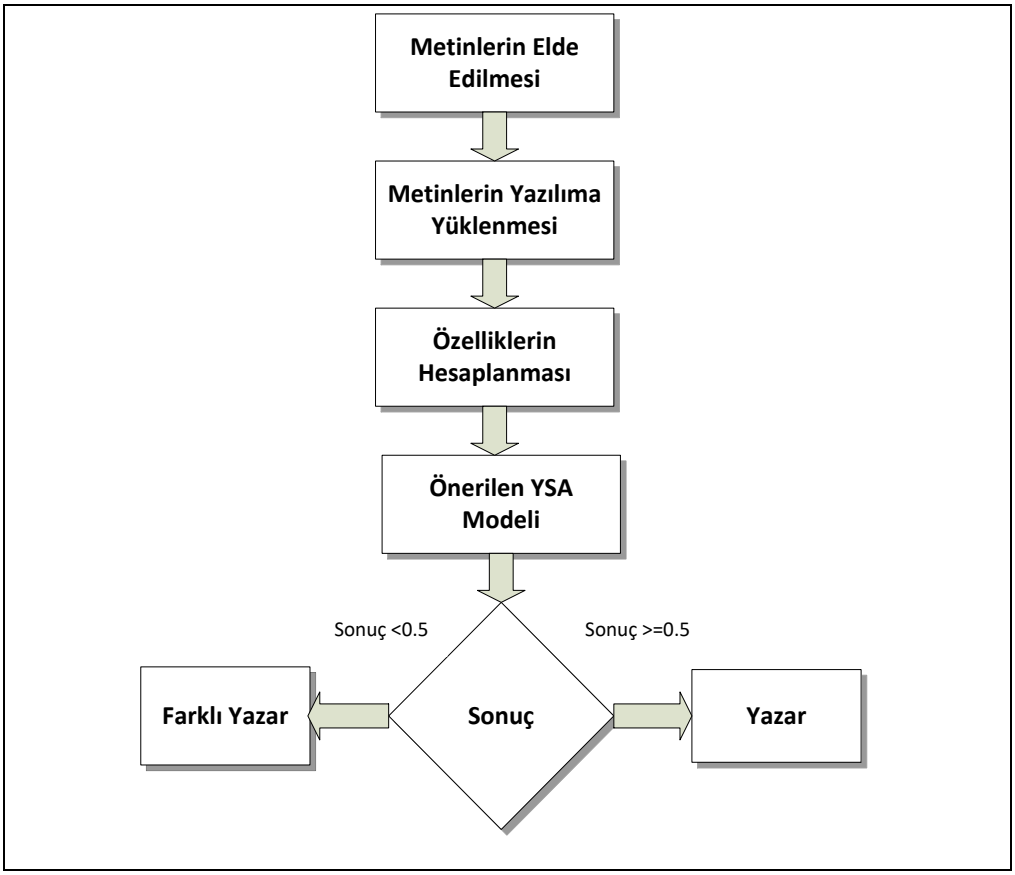


Şekil 7.1. Ayşe Arman için YSA Yazar Tanıma Modeli

Önerilen YSA modelinin yazar tanıma başarısının performans değerlendirilmesinin yapılması için çeşitli deneyler gerçekleştirilmiştir. Deneylerin gerçekleştirilmesi için Ayşe Arman'a ait 100 köşe yazısı ile Ayşe Arman'a ait olmayan 100 köşe yazısını içeren bir veri seti oluşturulmuştur. Bu veri seti içinde eğitim ve test verilerinin oluşturulmasında 5 k katlı çapraz doğrulama yapılmıştır. 5 k katlı çapraz doğrulama yönteminde veri seti 5 parçaya bölünür. Elde edilen parçalardan biri test verisi olmak üzere ayrılır geri kalan veri seti eğitim amacıyla kullanılır. Elde edilen eğitim ve test veri setleri ile sınıflandırma yapılır. Daha sonra 5 parçadan kullanılmayan başka bir parça test için seçilir ve geri kalan veri eğitim verisi olur. Bu adım tüm parçalar kullanılana kadar böyle devam eder. Yapılan deneyler sonucunda elde edilen doğruluk değerleri Çizelge 7.3'de gösterilmektedir.

Çizelge 7.3. Ayşe Arman için Önerilen YSA Modelinin Performans Değerlendirmesi

Deneyler	Eğitim Doğruluk (%)	Test Doğruluk (%)
#1	100	90
#2	100	95
#3	100	85
#4	100	90
#5	100	100
Ortalama	100	92



Sekil 7.1. Önerilen Sistemin Genel Yapısı

YSA Model 2- İclal Aydın

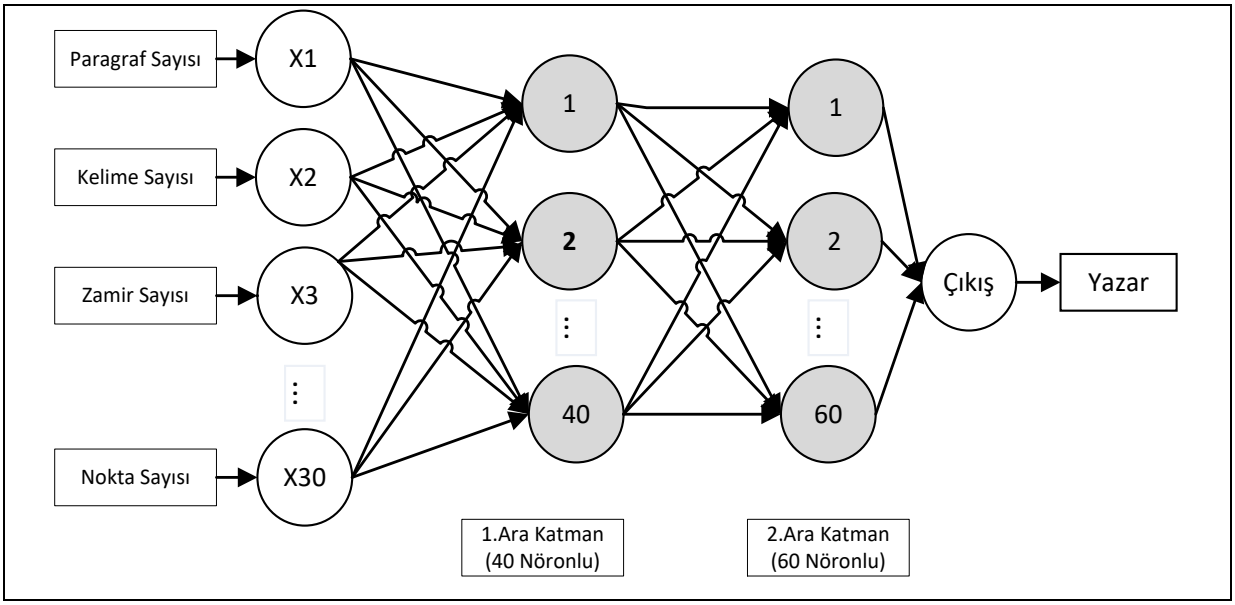
İclal Aydın'ın yazılarının tanınmasında kullanılacak olan YSA modelinin yapısının belirlenmesi için Çizelge 7.4'de gösterilen YSA yapıları oluşturulmuştur. Bu YSA yapılarında ağ mimarisi olarak ileri beslemeli ağ yapısına sahip olan MLP tercih edilmiş olup öğrenme algoritmaları olarak ise lineer, tanjant ve sigmoid transfer fonksiyonları kullanılarak 10 farklı YSA yapısı gerçekleştirilmiştir.

Çizelge 7.4. İclal Aydın için YSA Parametrelerinin Karşılaştırılması

YSA#	AKS	HKNS	TF	ÖA	MSE
#1	2	20,40,1	S,T,LN	LM	3.82E-19
#2	2	20,40,1	S,T,LN	GDA	0.000143
#3	**2**	**40,60,1**	**S,T,LN**	**LM**	**4.37E-30**
#4	2	20,40,1	S,T,LN	GD	0.00549
#5	2	25,50,1	S,T,LN	LM	2.28E-30
#6	3	20,40,60,1	S,T,T,LN	LM	1.23E-22
#7	2	20,40,1	S,T,LN	GDX	4.20E-06
#8	2	20,40,1	S,T,LN	GDM	0.00549
#9	2	30,60,1	T,S,LN	GD	0.00549
#10	3	30,50,40,1	S,S,T,LN	LM	8.18E-21

ÖA: Öğrenme Algoritması, AKS: Ara Katman Sayısı, HKNS: Her Katmanda Nöron Sayısı, TF: Transfer Fonksiyonu,TH: Tanjant Hiperbolik, S: Sigmoid, LN: Doğrusal, MSE: Ortalama Hata Kare, LM: Levenberg-Marquardt, GD: Dereceli Azalan Geri Yayılım, GDA: Adaptif Öğrenme Oranlı Dereceli Azalan Geri Yayılım, GDX: Adaptif Öğrenme Oranlı ve Momentumlu Gradyan Azalan

İclal Aydın için yapılan denemeler sonucunda oluşturulan YSA yapılarından en az hata oranını veren YSA#3 modeli tercih edilmiştir. Çizelge 7.4'de bu YSA yapısı bold (koyu renk) olarak gösterilmektedir. Seçilen YSA modeli öğrenme algoritması olarak Ayşe Arman için oluşturulan modeldeki gibi öğrenme algoritması olarak Levenberg Marquardt kullanmaktadır. Ayrıca YSA modelinin giriş katmanında 30 nöron, 1. ara katmanında 40 nöron, 2. ara katmanında 60 nöron ve transfer fonksiyonu olarak 1. ara katmanda logaritmik sigmoid fonksiyon ve 2. ara katmanda tanjant hiperbolik fonksiyonu kullanılmaktadır. Oluşturulan YSA modelinin eğitim aşamasında sıfıra hataya ulaşmak için epok değeri 1000 olarak seçilmiş olup eğitim aşaması sonucu elde edilen performans değeri 4.37e-30 olarak bulunmuştur. Önerilen YSA modeli Şekil 7.2'de gösterilmektedir.



Şekil 7.2. İclal Aydın için Oluşturulan YSA Modeli

Önerilen YSA modelinin yazar tanımda başarısının performans değerlendirilmesinin yapılması için çeşitli deneyler gerçekleştirilmiştir. Deneylerin gerçekleştirilmesi için İclal Aydın'a ait 100 köşe yazısı ile İclal Aydın'a ait olmayan 100 köşe yazısını içeren bir veri seti oluşturulmuştur. Bu veri seti içinde eğitim ve test verilerinin oluşturulmasında 5 k katlı çapraz doğrulama yapılmıştır. Elde edilen doğruluk değerleri Çizelge 7.5'de gösterilmektedir.

Çizelge 7.5. İclal Aydın için Önerilen YSA Modelinin Performans Değerlendirmesi

Deneyler	Eğitim Doğruluk (%)	Test Doğruluk (%)
#1	100	85
#2	100	100
#3	100	95
#4	100	100
#5	100	100
Ortalama	100	96

YSA Model 3- Vahap Munyar

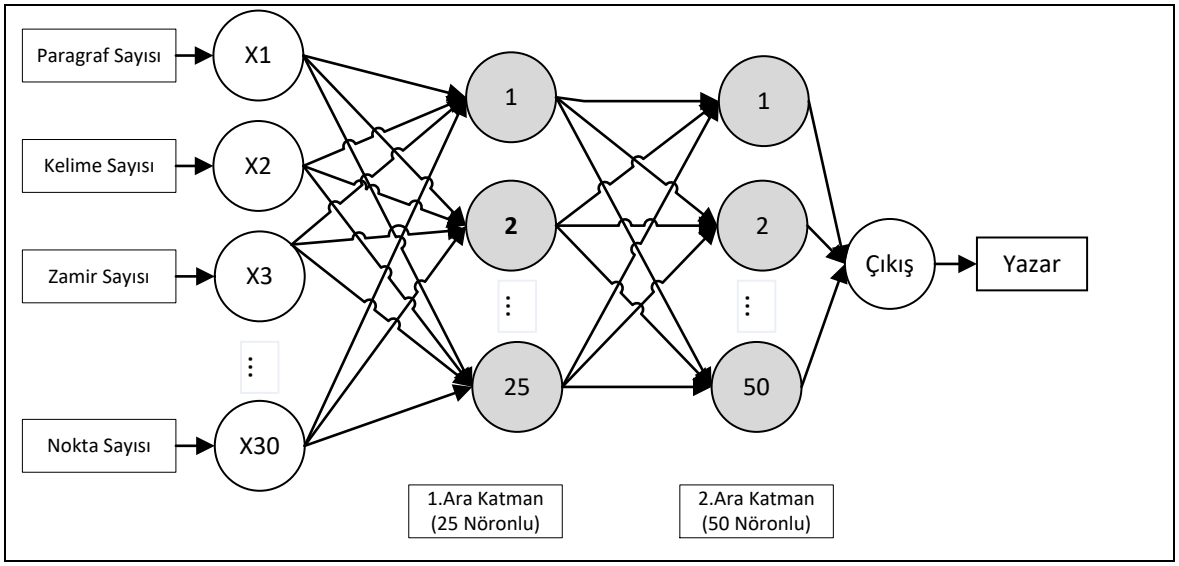
Bir yazının gerçek sahibinin Vahap Munyar olup olmadığının belirlenmesi için uygun YSA modellinin geliştirilmesi gerekmektedir. Bu tez çalışması içinde kullanılan YSA modellerinin ağ mimarisi olarak ileri beslemeli ağ yapısına sahip olan MLP tercih edilmiştir. Tercih edilen bu ağ yapısının eğitim aşamasında 5 farklı öğrenme algoritması kullanılmış ve performans karşılaştırılması yapılmıştır. Uygun YSA modelinin geliştirilmesi için Çizelge 7.6' da verildiği gibi farklı YSA yapıları deneme yanılma yoluyla test edilerek ortalama hata değerlerinin en aza indirgenmesi amaçlanmıştır.

Çizelge 7.6. Vahap Munyar için YSA Parametrelerinin Karşılaştırılması

YSA#	AKS	HKNS	TF	ÖA	MSE
#1	2	20,40,1	S,T,LN	LM	1.19E-20
#2	2	20,40,1	S,T,LN	GDA	0.00588
#3	2	40,60,1	S,T,LN	LM	8.48E-25
#4	2	20,40,1	S,T,LN	GD	0.246
#5	**2**	**25,50,1**	**S,T,LN**	**LM**	**2.19E-30**
#6	3	20,40,60,1	S,T,T,LN	LM	9.87E-20
#7	2	20,40,1	S,T,LN	GDX	4.12E-04
#8	2	20,40,1	S,T,LN	GDM	7.00E-13
#9	2	30,60,1	T,S,LN	GD	2.50E-01
#10	3	30,50,40,1	S,S,T,LN	LM	9.86E-29

ÖA: Öğrenme Algoritması, AKS: Ara Katman Sayısı, HKNS: Her Katmanda Nöron Sayısı, TF: Transfer Fonksiyonu,TH: Tanjant Hiperbolik, S: Sigmoid, LN: Doğrusal, MSE: Ortalama Hata Kare, LM: Levenberg-Marquardt, GD: Dereceli Azalan Geri Yayılım, GDA: Adaptif Öğrenme Oranlı Dereceli Azalan Geri Yayılım, GDX: Adaptif Öğrenme Oranlı ve Momentumlu Gradyan Azalan

Yapılan denemeler sonucunda minimum hata oranına sahip olan YSA#3 yapısı tercih edilmiş olup bold (koyu renk) olarak Çizelge 7.6'da ifade edilmiştir. YSA#3 modeli Şekil 7.3'de gösterilidiği gibi olup giriş katmanında 30 nöron, 1. ara katmanında 40 nöron, 2. ara katmanında 60 nöron ve transfer fonksiyonu olarak 1. ara katmanda logaritmik sigmoid fonksiyon ve 2. ara katmanda tanjant hiperbolik fonksiyonu kullanılmaktadır. Oluşturulan YSA modelinin eğitim aşamasında sıfıra hataya ulaşmak için epok değeri 1000 olarak seçilmiş olup eğitim aşaması sonucu elde edilen performans değeri 2.19e -25 olarak bulunmuştur. Önerilen bu YSA modelinde öğrenme algoritması olarak Levenberg-Marquardt kullanılmaktadır.



Şekil 7.3. Vahap Munyar için Oluşturulan YSA Modeli

Önerilen YSA modelinin yazar tanıma başarısının performans değerlendirilmesinin yapılması için çeşitli deneyler gerçekleştirilmiştir. Deneylerin gerçekleştirilmesi için Vahap Munyar'a ait 100 köşe yazısı ile Vahap Munyar'a ait olmayan 100 köşe yazısını içeren bir veri seti oluşturulmuştur. Bu veri seti içinde eğitim ve test veri setlerinin elde edilmesi için 5 k katlı çapraz doğrulama yapılmıştır. Eğitim ve testler sonucunda elde edilen doğruluk değerleri Çizelge 7.7'de gösterilmektedir.

Çizelge 7.7. Vahap Munyar için Önerilen YSA Modelinin Performans Değerlendirmesi

Deneyler	Eğitim Doğruluk	Test Doğruluk
#1	100%	95%
#2	100%	100%
#3	100%	100%
#4	100%	95%
#5	100%	75%
Ortalama		93%

YSA Model 4- Güngör Uras

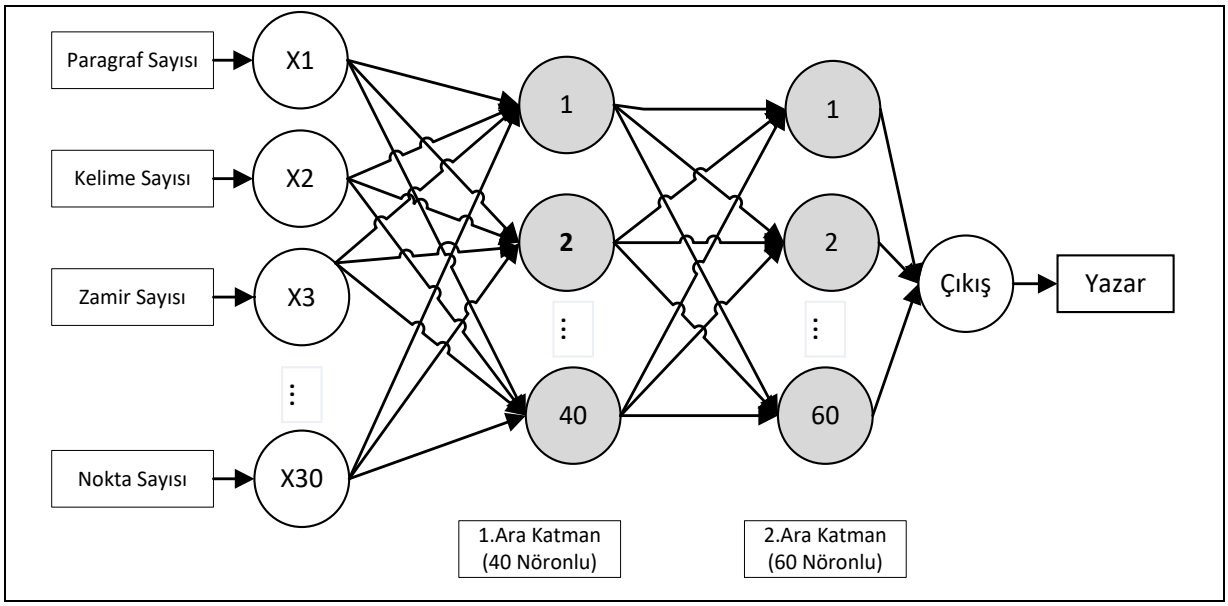
Güngör Uras için uygun YSA modelinin belirlenmesi için farklı YSA yapıları oluşturulmuştur. Bu YSA yapıları farklı öğrenme algoritmaları, farklı ara katman sayısı ile nöron sayısına ve farklı transfer fonksiyonlarına sahiptirler. Oluşturulan bu YSA yapıları ve yapılan denemelerde elde edilen hata oranı değerleri Çizelge 7.8'de gösterilmektedir.

Çizelge 7.8. Güngör Uras için YSA Parametrelerinin Karşılaştırılması

YSA#	AKS	HKNS	TF	ÖA	MSE
#1	2	20,40,1	S,T,LN	LM	6.66E-24
#2	2	20,40,1	S,T,LN	GDA	0.000323
#3	**2**	**40,60,1**	**S,T,LN**	**LM**	**5.50E-28**
#4	2	20,40,1	S,T,LN	GD	0.25
#5	2	25,50,1	S,T,LN	LM	1.19E-19
#6	3	20,40,60,1	S,T,T,LN	LM	3.86E-27
#7	2	20,40,1	S,T,LN	GDX	6.93E-07
#8	2	20,40,1	S,T,LN	GDM	0.25
#9	2	30,60,1	T,S,LN	GD	0.25
#10	3	30,50,40,1	S,S,T,LN	LM	5.46E-27

ÖA: Öğrenme Algoritması, AKS: Ara Katman Sayısı, HKNS: Her Katmanda Nöron Sayısı, TF: Transfer Fonksiyonu,TH: Tanjant Hiperbolik, S: Sigmoid, L: Doğrusal, MSE: Ortalama Hata Kare, LM: Levenberg-Marquardt, GD: Dereceli Azalan Geri Yayılım, GDA: Adaptif Öğrenme Oranlı Dereceli Azalan Geri Yayılım

Yapılan denemeler sonucunda belirlenen YSA modelinin giriş katmanında 30 nöron birinci ara katmanda 40 nöron ikinci ara katmanda ise 60 nöronu bulunmaktadır. Ayrıca bu YSA modeli öğrenme algoritması olarak Levenberg-Marquardt algoritmasını kullanmaktadır. Sigmoid, tanjant ve linear fonksiyonları ise transfer fonksiyonları olarak tercih edilmiştir. Şekil 7.4'de bu YSA modeli gösterilmektedir. Oluşturulan YSA modelinin eğitilmesi sonucunda elde edilen minimum karesel hata değeri 5.50e-28 olarak bulunmuştur.



Şekil 7.4. Güngör Uras için Oluşturulan YSA Modeli

Önerilen YSA modelinin yazar tanıma başarısının performans değerlendirilmesinin yapılması için çeşitli deneyler gerçekleştirilmiştir. Deneylerin gerçekleştirilmesi için Güngör Uras'a ait 100 köşe yazısı ile Güngör Uras'a ait olmayan 100 köşe yazısını içeren bir veri seti oluşturulmuştur. Bu veri seti içinde eğitim ve test veri setlerinin elde edilmesi için 5 k katlı çapraz doğrulama yapılmıştır. Eğitim ve testler sonucunda elde edilen doğruluk değerleri Çizelge 7.9'da gösterilmektedir.

Çizelge 7.9. Güngör Uras için Önerilen YSA Modelinin Performans Değerlendirmesi

Deneyler	Eğitim Doğruluk	Test Doğruluk
#1	100%	100%
#2	100%	100%
#3	100%	95%
#4	100%	100%
#5	100%	85%

YSA Model 5- Emre Aköz

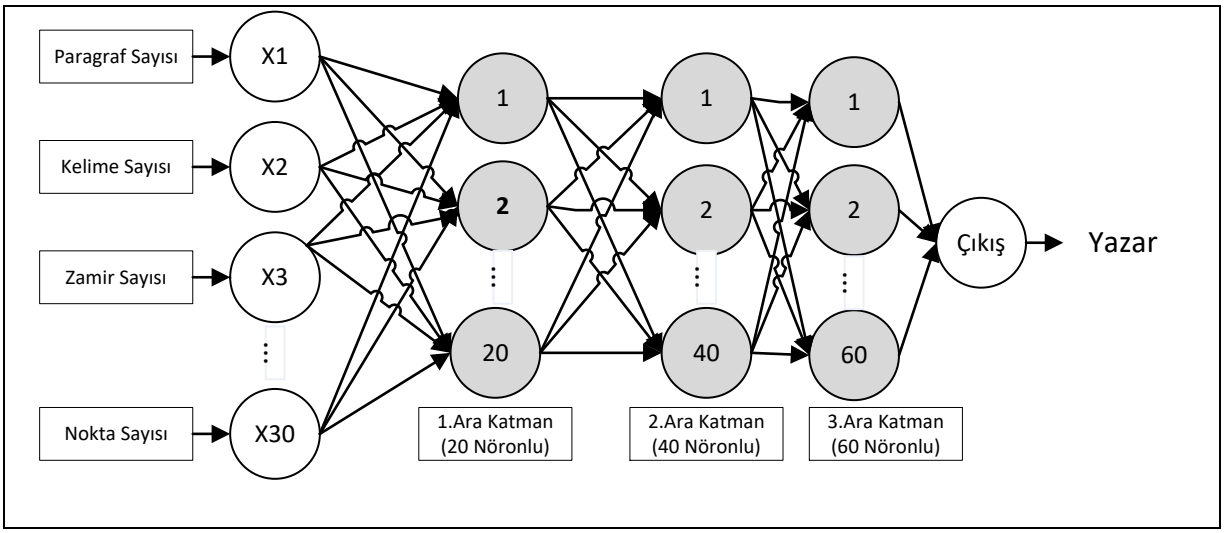
Emre Aköz köşe yazarı için uygun YSA modelinin belirlenmesi için ara katman sayısı, katmanlardaki nöron sayıları ve geçiş fonksiyonları değiştirilerek 10 farklı YSA modeli oluşturulmuştur. YSA yapılarının ağ mimarisi olarak ileri beslemeli ağ yapısına sahip olan MLP tercih edilmiş olup 5 farklı öğrenme algoritması seçilmiştir. Oluşturulan YSA yapıları Çizelge 7.10'da gösterilmektedir.

Çizelge 7.10. Emre Aköz için YSA Parametrelerinin Karşılaştırılması

YSA#	AKS	HKNS	TF	ÖA	MSE
#1	2	20,40,1	S,T,LN	LM	7.75E-21
#2	2	20,40,1	S,T,LN	GDA	0.00669
#3	2	40,60,1	S,T,LN	LM	4.02E-24
#4	2	20,40,1	S,T,LN	GD	0.25
#5	2	25,50,1	S,T,LN	LM	1.01E-23
#6	**3**	**20,40,60,1**	**S,T,T,LN**	**LM**	**2.18E-26**
#7	2	20,40,1	S,T,LN	GDX	3.28E-04
#8	2	20,40,1	S,T,LN	GDM	1.36E-14
#9	2	30,60,1	T,S,LN	GD	0.25
#10	3	30,50,40,1	S,S,T,LN	LM	3.17E-21

ÖA: Öğrenme Algoritması, AKS: Ara Katman Sayısı, HKNS: Her Katmanda Nöron Sayısı, TF: Transfer Fonksiyonu,TH: Tanjant Hiperbolik, S: Sigmoid, LN: Doğrusal, MSE: Ortalama Hata Kare, LM: Levenberg-Marquardt, GD: Dereceli Azalan Geri Yayılım, GDA: Adaptif Öğrenme Oranlı Dereceli Azalan Geri Yayılım, GDX: Adaptif Öğrenme Oranlı ve Momentumlu Gradyan Azalan

Oluşturulan 10 farklı YSA yapıları ayrı ayrı denenmiş olup bu yapıların performans kıyaslaması yapılmıştır. Yapılan denemeler sonucunda elde edilen performans değerine (MSE) göre eğitim aşamasında kullanılacak olan YSA modeli belirlemiştir. Emre Aköz yazılarının belirlenmesi için kullanılacak olan YSA modeli olarak YSA #6 belirlenmiştir. Bu modelin eğitim aşamasındaki performans değeri (MSE) 2.18E-26 olarak bulunmuştur. Diğer YSA modellerinden farklı olarak önerilen bu model 3 ara katmandan oluşmaktadır. Giriş katmanındaki nöron sayısı 30, birinci ara katmandaki nöron sayısı 40, ikinci ara katmandaki nöron sayısı 60 ve çıkış ara katmanında bir nöron bulunmaktadır.



Şekil 6.5. Emre Aköz için Oluşturulan YSA Modeli

Önerilen YSA modelinin yazar tanıma başarısının performans değerlendirilmesinin yapılması için çeşitli deneyler gerçekleştirilmiştir. Deneylerin gerçekleştirilmesi için Emre Aköz'e ait olan 100 köşe yazısı ile Emre Aköz'e ait olmayan 100 köşe yazısını içeren bir veri seti oluşturulmuştur. Bu veri seti içinde eğitim ve test veri setlerinin elde edilmesi için 5 k katlı çapraz doğrulama yapılmıştır. Eğitim ve testler sonucunda elde edilen doğruluk değerleri Çizelge 7.11'de gösterilmektedir.

Çizelge 7.11.Emre Aköz için Önerilen YSA Modelinin Performans Değerlendirmesi

Deneyler	Eğitim Doğruluk (%)	Test Doğruluk (%)
#1	100	100
#2	100	100
#3	100	90
#4	100	100
#5	100	90
Ortalama		96

YSA Model 6- Hadi Uluengin

Hadi Uluengin köşe yazarı için uygun YSA modelinin belirlenmesi için YSA yapısı olarak ileri beslemeli ağ yapısına sahip olan MLP tercih edilmiş olup ara katman sayısı, katmanlardaki nöron sayıları ve geçiş fonksiyonları değiştirilerek 10 farklı YSA yapısı oluşturulmuştur. Oluşturulan YSA yapıları Çizelge 7.12'de gösterilmektedir.

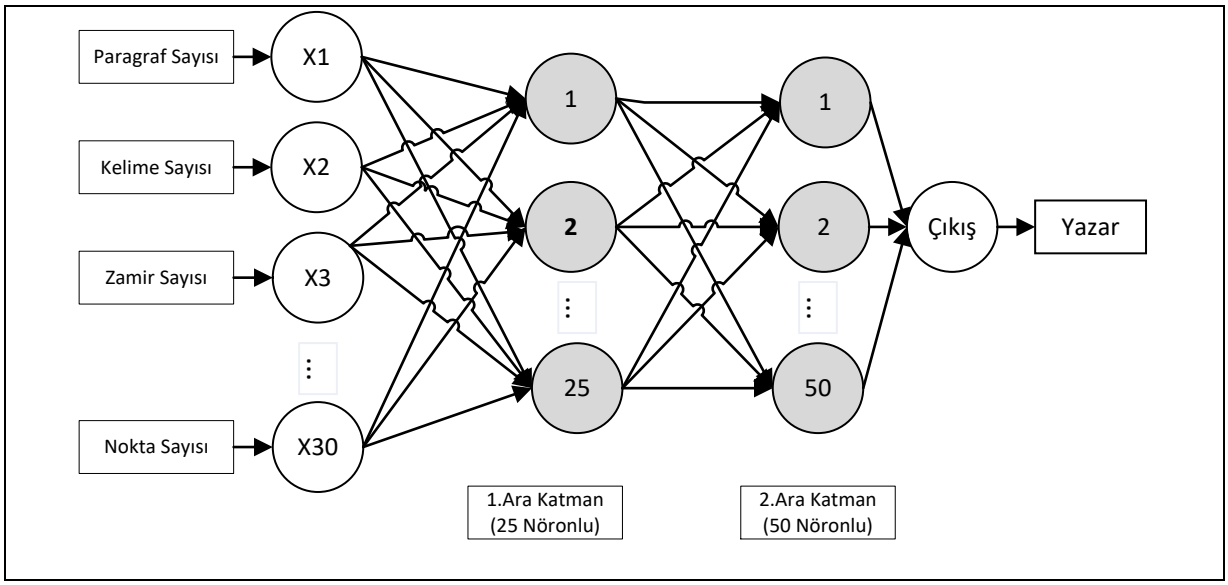
Çizelge 7.12. Hadi Uluengin için YSA Parametrelerinin Karşılaştırılması

YSA#	AKS	HKNS	TF	ÖA	MSE
#1	**2**	**20,40,1**	**S,T,LN**	**LM**	**1.24E-30**
#2	2	20,40,1	S,T,LN	GDA	0.116
#3	2	40,60,1	S,T,LN	LM	3.83E-12
#4	2	20,40,1	S,T,LN	GD	0.248
#5	2	25,50,1	S,T,LN	LM	1.17E-24
#6	3	20,40,60,1	S,T,T,LN	LM	3.29E-20
#7	2	20,40,1	S,T,LN	GDX	9.52E-02
#8	2	20,40,1	S,T,LN	GDM	2.51E-01
#9	2	30,60,1	T,S,LN	GD	0.261
#10	3	30,50,40,1	S,S,T,LN	LM	1.81E-25

ÖA: Öğrenme Algoritması, AKS: Ara Katman Sayısı, HKNS: Her Katmanda Nöron Sayısı, TF: Transfer Fonksiyonu,TH: Tanjant Hiperbolik, S: Sigmoid, LN: Doğrusal, MSE: Ortalama Hata Kare, LM: Levenberg-Marquardt, GD: Dereceli Azalan Geri Yayılım, GDA: Adaptif Öğrenme Oranlı Dereceli Azalan Geri Yayılım, GDX: Adaptif Öğrenme Oranlı ve Momentumlu Gradyan Azalan

Yapılan denemeler ile farklı YSA yapılarının performans kıyaslanması gerçekleştirilmiştir. Yapılan karşılaştırma sonucunda uygun YSA modeli olarak YSA#6 belirlenmiştir. Belirlenen YSA modeli 1. ara katmanda 20 nöron, 2. ara katmanda 40 nöron ve 3. ara katmanda 60 nöron olmak üzere üç ara katmandan oluşmaktadır. Giriş katmanındaki nöron sayısı 30 iken çıkış katmanındaki nöron sayısı ise 1 olarak belirlenmiştir. Oluşturulan YSA modeli Şekil 7.6'da gösterilmektedir. Ayrıca transfer fonksiyonu olarak 1. Ara katmanda logaritmik sigmoid fonksiyon ve 2. ara katmanda tanjant hiperbolik fonksiyonu ve 3.ara katmanda tanjant hiperbolik fonksiyonu kullanılmıştır. YSA modelinin giriş katmanındaki nöron sayısı 30 olup çıkış katmanındaki nöron sayısı ise 1 olarak belirlenmiştir. Öğrenme algoritması olarak Levenberg Marquardt kullanılmıştır. Sıfır hata oranına erişmek için modelin eğitim aşamasında 1000 epok değeri kullanılmıştır. Eğitim aşamasında epok değerlerinin

değişimi izlendiği zaman hata oranının giderek düştüğü gözlenmektedir. Gözlenen hata oranı değerleri YSA modelinin başarılı genelleme yapabildiğini göstermektedir.



Şekil 7.6. Hadi Uluengin için Oluşturulan YSA modeli

Önerilen YSA modelinin yazar tanıma başarısının performans değerlendirilmesinin yapılması için çeşitli deneyler gerçekleştirilmiştir. Deneylerin gerçekleştirilmesi için Hadi Uluengin'e ait olan 100 köşe yazısı ile Hadi Uluengin'e ait olmayan 100 köşe yazısını içeren bir veri seti oluşturulmuştur. Bu veri seti içinde eğitim ve test veri setlerinin elde edilmesi için 5 k katlı çapraz doğrulama yapılmıştır. Gerçekleştirilen öğrenme aşamaları elde edilen başarılı sonuçlardan sonra sonlandırılmıştır. Tespit edilen YSA modelleri ile eğitimler gerçekleştirilmiş olup elde edilen başarılı sonuçlardan sonra sonlandırılmıştır. Önerilen YSA modelinin başarısının test edilmesi için yapılan deneylerde elde edilen doğruluk değerleri Çizelge 7.13'de gösterilmektedir.

Çizelge 7.13. Hadi Uluengin için Önerilen YSA Modelinin Performans Değerlendirmesi

Deneyler	Eğitim Doğruluk (%)	Test Doğruluk (%)
#1	100	100
#2	100	100
#3	100	90
#4	100	100
#5	100	90
Ortalama		96

Tekli sınıflandırma problemi olarak ele alınan yazar tanıma problemi için gerekli modellerinin oluşturulmasının ardından bu modellerde kullanılan yazarların tamamı kullanılarak çoklu sınıflandırma problemi için bir model önerilmektedir. Önerilen model Ayşe Arman, İclal Aydın, Hadi Uluengin, Emre Aköz, Vahap Munyar ve Güngör Uras köşe yazarlarının yazılarını tanıyabilmektedir. Ayrıca YSA'ları genelleme yapma özelliğinden yararlanılarak önerilen model bu yazarlara ait olmayan bir yazı ile karşılaştığı zaman ise yazının seçilen bu altı yazar dışında farklı bir yazara ait olduğunu belirtmektedir.

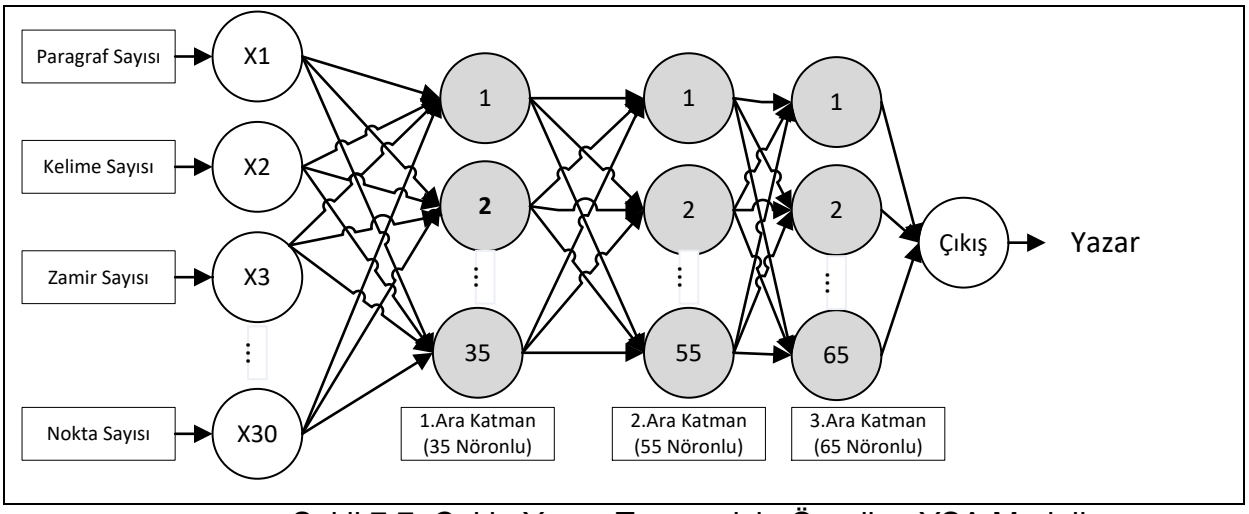
Ayşe Arman, İclal Aydın, Hadi Uluengin, Emre Aköz, Vahap Munyar ve Güngör Uras köşe yazarlarının yazıları kullanılarak oluşturulan derlem içerisinden biçimsel özelliklerin elde edilmesi için Aslantürk tarafından geliştirilmiş olan yazılım bu çalışma içerisinde de kullanılmaktadır. Yazarın parmak izi olarak isimlendirilen bu özellikler 30 adet olup nokta, virgül, noktalı virgül sayıları gibi noktalama işaretlerinden; cümle sayısı, kelime sayısı, paragraf sayısı gibi metinsel özelliklerden ve son olarak ise isim, sıfat, zamir gibi kelime türlerinden oluşmaktadır. Yazılım kullanılarak elde edilen yazar özellikleri veri seti olarak daha sonra önerilen YSA modelinin eğitilmesi için kullanılmaktadır. Oluşturulan 15 farklı YSA yapısı ayrı ayrı denenmiş olup bu yapıların performans kıyaslaması yapılmıştır.

Çizelge 7.14. Yazar Tanıma için Oluşturulan YSA Yapıları

YSA#	AKS	HKNS	TF	ÖA	MSE
#1	2	20,40,1	S,T,LN	LM	1.09E-13
#2	2	20,40,1	S,T,LN	GDA	0.966
#3	2	40,60,1	S,T,LN	LM	6.01E-19
#4	2	20,40,1	S,T,LN	GD	2.89
#5	2	25,50,1	S,T,LN	LM	1.6E-20
#6	3	20,40,60,1	S,T,T,LN	LM	0.0000505
#7	2	20,40,1	S,T,LN	GDX	0.662
#8	2	20,40,1	S,T,LN	GDM	2.07
#9	2	30,60,1	T,S,LN	LM	0.000975
#10	3	30,40,50,1	S,S,T,LN	LM	0.0795
#11	2	20,40,1	S,T,LN	GDX	0.763
#12	3	20,40,60,1	S,T,T,LN	GDX	2.66
#13	2	35,55,1	S,T,LN	LM	7.71E-30
#14	2	40,10,1	S,T,LN	LM	0.000019
#15	**3**	**35,55,65,1**	**S,T,T,LN**	**LM**	**6.86E-40**

ÖA: Öğrenme Algoritması, AKS: Ara Katman Sayısı, HKNS: Her Katmanda Nöron Sayısı, TF: Transfer Fonksiyonu,TH: Tanjant Hiperbolik, S: Sigmoid, LN: Doğrusal, MSE: Ortalama Hataların KareköküToplamı, LM: Levenberg-Marquardt, GD: Dereceli Azalan Geri Yayılım, GDA: Adaptif Öğrenme Oranlı Dereceli Azalan Geri Yayılım, GDX: Adaptif Öğrenme Oranlı ve Momentumlu Gradyan Azalan

Yapılan denemeler sonucunda elde edilen performans değerine (MSE) göre eğitim aşamasında kullanılacak olan YSA modeli belirlenmiştir. Yazar tanıma için kullanılacak olan YSA modeli olarak YSA #15 belirlenmiştir. Şekil 7.7'de önerilen YSA modeli gösterilmektedir. Bu YSA modelinin, giriş katmanında 30 nöron, 1. ara katmanda 35 nöron, 2. Ara katmanda 55 nöron, ve 3. Ara katmanda 65 nöron bulunmaktadır. Ayrıca transfer fonksiyonu olarak 1. Ara katmanda logaritmik sigmoid fonksiyon, 2. ve 3. ara katmanda tanjant hiperbolik fonksiyonu kullanılmıştır. Çıkış katmanında ise 1 nöron bulunmaktadır. Öğrenme algoritması olarak Levenberg Marquardt kullanılmıştır.



Şekil 7.7. Çoklu Yazar Tanıma için Önerilen YSA Modeli

Önerilen YSA modelinin yazar tanıma başarısının performans değerlendirilmesinin yapılması için çeşitli deneyler gerçekleştirilmiştir. Deneylerin gerçekleştirilmesi için her yazar için 100'er köşe yazısı seçilmiş olup toplam 600 köşe yazısından oluşan bir veri seti oluşturulmuştur. Bu veri seti içinde eğitim ve test veri setlerinin elde edilmesi için 5 k katlı çapraz doğrulama yapılmıştır. Gerçekleştirilen öğrenme aşamaları elde edilen başarılı sonuçlardan sonra sonlandırılmıştır. Tespit edilen YSA modelleri ile eğitimler gerçekleştirilmiş olup elde edilen başarılı sonuçlardan sonra sonlandırılmıştır. Önerilen YSA modelinin başarısının test edilmesi için yapılan deneylerde elde edilen doğruluk değerleri Çizelge 7.15'de gösterilmektedir.

Çizelge 7.15. Yazar Tanıma için Önerilen YSA Modelinin Performans Değerlendirmesi

Test	Ayşe Arman	İclal Aydın	Hadi Uluengin	Emre Aköz	Vahap Munyar	Güngör Uras	Farklı Yazar
#1	80%	70%	80%	60%	60%	70%	60%
#2	60%	90%	70%	90%	60%	80%	70%
#3	70%	90%	100%	100%	100%	100%	70%
#4	80%	60%	80%	70%	60%	90%	80%
#5	80%	60%	90%	70%	80%	100%	80%
ortalama	**74**%	**74**%	**84**%	**78**%	**72**%	**88**%	**72%**

7.2. Yazı Alanı (Türü) Tanıma YSA Modeli

Verilen bir metnin ya da dokümanın önceden belirlenen sınıflardan birine otomatik olarak atanması problemi olarak tanımlanan metin sınıflandırma çalışma konusu birçok uygulama alanına sahiptir. Yazı alanı (türü) tanıma da metin sınıflandırma çalışması alanlarından birisidir. İnternet ortamında bulunan metinlerin sayısındaki artış ile birlikte bu veriler içerisinde verimli kataloglama ve alımın yapılabilmesi için metinlerin alanlarına göre otomatik olarak sınıflandırılması oldukça önemli bir problem haline gelmektedir.

Ayrıca bulunan bir metnin yazarının belirlenmesinin zor olduğu durumlarda en azından türü hakkında bilgi sahibi olunması durumunda belirlenen tür ile ilgili yazılar yazan yazarlar arasında yazının yazarının belirlemesini yapılacak olan işlemleri kolaylaştırabilir. Bu ve buna benzer problemlere çözümler getirmek için bu tez kapsamında yazı alanı (türü) belirleyen YSA modeli önerilmektedir. Ayrıca bu çalışma içerisinde yazar tanıma için kullanılan biçimsel özelliklerin yazı alanı belirlemedeki başarısı değerlendirilmiştir. Yazı alanı tanıma için Çizelge 7.16.'de gösterilen derlem kullanılmaktadır.

Çizelge 7.16. Yazı Alanı Belirleme için Kullanılan Derlem

Alan	Yazar	Yazı Adeti	Toplam Adeti	Tarih Aralığı Başlangıcı	Tarih Aralığı Sonu
Siyaset	Hadi Uluengin	100	200	01-06-08	05-01-09
	Emre Aköz	100		01-06-08	10-01-09
Ekonomi	Vahap Munyar	100	200	01-06-08	01-01-09
	Güngör Uras	100		01-06-08	05-01-09
Yaşam	İclal Aydın	100	200	01-06-08	09-01-09
	Ayşe Arman	100		01-06-08	11-01-09

Yazı alanı (türü) tanıma için oluşturulacak uygun YSA modelinin belirlenmesinde ara katman sayısı, katmanlardaki nöron sayıları ve geçiş fonksiyonları değiştirilerek 10

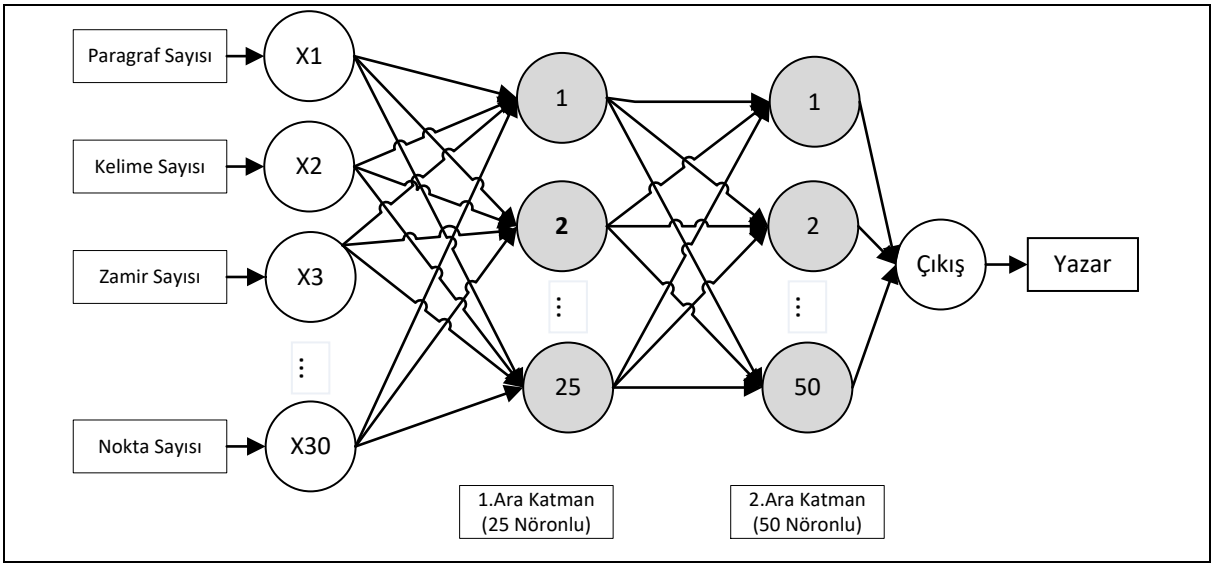
farklı YSA modeli oluşturulmuştur. Oluşturulan YSA yapıları Çizelge 7.17'de gösterilmektedir. YSA yapılarının ağ mimarisi olarak ileri beslemeli ağ yapısına sahip olan MLP tercih edilmiş olup 5 farklı öğrenme algoritması seçilmiştir. Oluşturulan YSA yapıları Çizelge 2'de gösterilmektedir. Oluşturulan 10 farklı YSA yapısı ayrı ayrı denenmiş olup bu yapıların performans kıyaslaması yapılmıştır.

Çizelge 7.17. Yazı Türü Tanıma için Oluşturulan YSA Yapıları

YSA#	AKS	HKNS	TF	ÖA	MSE
#1	2	20,40,1	S,T,LN	LM	3.30E-19
#2	2	20,40,1	S,T,LN	GDA	0.368
#3	**2**	**40,60,1**	**S,T,LN**	**LM**	**5.90E-29**
#4	2	20,40,1	S,T,LN	GD	0.55
#5	2	25,50,1	S,T,LN	LM	3.27E-24
#6	3	20,40,60,1	S,T,T,LN	LM	2.41E-22
#7	2	20,40,1	S,T,LN	GDX	0.328
#8	2	20,40,1	S,T,LN	GDM	0.559
#9	2	30,60,1	T,S,LN	GD	0.147
#10	3	30,40,50,1	S,S,T,LN	LM	1.89E-25

ÖA: Öğrenme Algoritması, AKS: Ara Katman Sayısı, HKNS: Her Katmanda Nöron Sayısı, TF: Transfer Fonksiyonu,TH: Tanjant Hiperbolik, S: Sigmoid, LN: Doğrusal, MSE: Ortalama Hataların KarekköüToplamı, LM: Levenberg-Marquardt, GD: Dereceli Azalan Geri Yayılım, GDA: Adaptif Öğrenme Oranlı Dereceli Azalan Geri Yayılım, GDX: Adaptif Öğrenme Oranlı ve Momentumlu Gradyan Azalan

Yapılan denemeler sonucunda elde edilen performans değerine (MSE) göre eğitim aşamasında kullanılacak olan YSA modeli belirlemiştir. Yazı alanı tanıma için kullanılacak olan YSA modeli olarak YSA #6 belirlenmiştir. Bu YSA modelinde giriş katmanında 30 nöron, 1. ara katmanda 40 nöron, 2. Ara katmanda 60 nöron ve transfer fonksiyonu olarak 1. ara katmanda logaritmik sigmoid fonksiyon ve 2. ara katmanda tanjant hiperbolik fonksiyonu kullanılmıştır Çıkış katmanında ise 1 nöron bulunmaktadır. Öğrenme algoritması olarak Levenberg Marquardt kullanılmıştır. Yazı alanı tanıma için önerilen YSA modeli Şekil 7.7'de gösterilmektedir.



Şekil 7.7. Yazı Alanı Tanıma için Önerilen YSA modeli

Önerilen modelin başarısının değerlendirilmesi için yapılan testler sonucunda elde edilen doğruluk değerleri Çizelge 7.18'de gösterilmektedir. Gerçekleştirilen testler sonucunda önerilen YSA modeli Yaşam alanında ortalama 70%, siyaset alanında 82% ve ekonomi alanında 88% başarı göstermektedir.

Çizelge 7.18. Yazı Alanı (Türü) Belirleme için Önerilen YSA Modelinin Performans Değerlendirmesi

Test	Yaşam Doğruluk Oranı (%)	Siyaset Doğruluk Oranı (%)	Ekonomi Doğruluk Oranı (%)
#1	60	90	85
#2	75	70	95
#3	65	85	90
#4	75	85	90
#5	75	80	80
ortalama	**70**	**82**	**88**

7.3. Yazar ve Yazı Alanı (Türü) Tanıma YSA Modeli

Alanyazın araştırmaları sonucunda yazar tanıma ve metin alanı belirleme problemleri üzerinde çeşitli çalışmaların yapıldığı fark edilmektedir. Yazar tanıma çalışması içinde kullanılan sözcüksel ve sözdizimsel özelliklerin metin alanı (türü) belirlemedeki başarısı değerlendirildikten sonra tez kapsamı içinde bir metinin hem yazarını hem konusunu (alanını) belirleyen hibrid bir modelin geliştirilmesi hedeflenmiştir. Bu model sayesinde yazarı ve yazı türü belli olmayan bir yazıyla karşılaşıldığı zaman bu yazının yazarı ve türü hakkında bilgi sahibi olunabilmesi amaçlanmıştır.

Önerilen hibrid model, 6 yazar arasında tanıma yapabilmekte ayrıca ‘Yaşam’, ‘Siyaset ve ‘Ekonomi’ olmak üzere üç yazı alanı tanıyabilmektedir. Gelişim aşamasında olan bu hibrid modelin gelecekte yapılması düşünülen çalışmalarda yazar sayısının ve alan sayısının arttırılması düşünülmektedir. Hem yazar ve hem de yazı alanı tanıma için geliştirilen hibrid modelin eğitilmesi ve test edilmesinde kullanılan veri seti yazar tanıma ve yazı alanı tanıma için kullanılan veri seti ile aynıdır. Önerilen diğer modellerinde kararlılığının gösterilebilmesi açısından bu önemlidir.

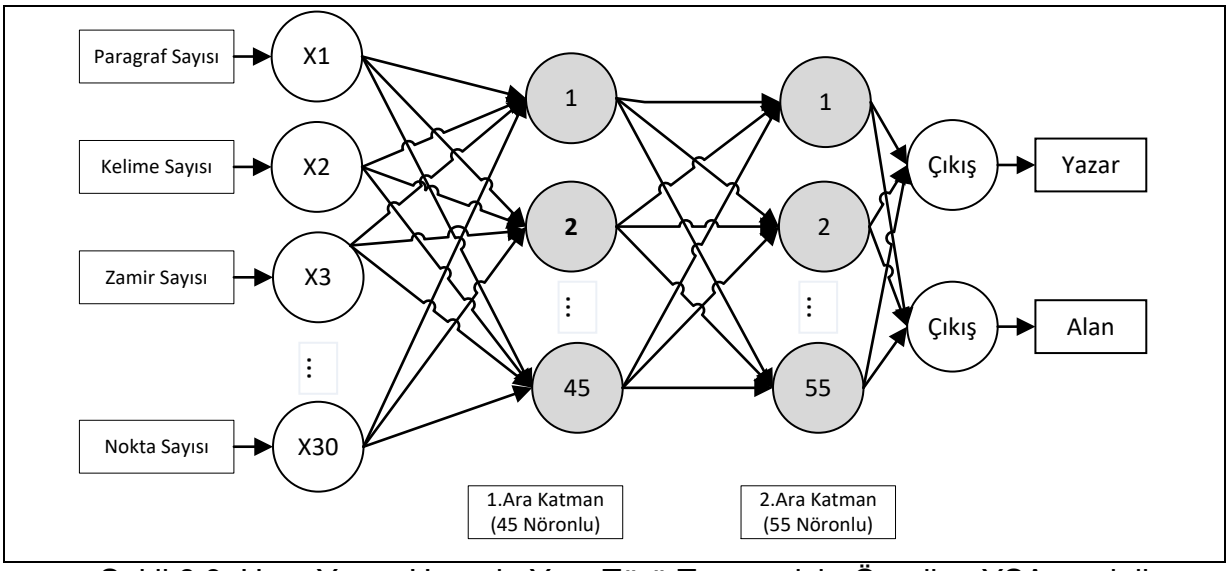
Hem yazar hem de yazı alanı tanıyan hibrid YSA modeli için YSA ağ mimarisi olarak diğer modellerde de tercih edilen ileri beslemeli ağ yapısına sahip olan MLP tercih edilmiştir. Tercih edilen bu ağ yapısının eğitim aşamasında 5 farklı öğrenme algoritması kullanılmış ve performans karşılaştırılması yapılmıştır. Ayrıca uygun YSA modelinin belirlenmesi için ara katman sayısı, katmanlardaki nöron sayıları ve geçiş fonksiyonları değiştirilerek 10 farklı YSA modeli oluşturulmuştur. Uygun YSA modelinin geliştirilmesi için Çizelge 7.19’ da verildiği gibi farklı YSA yapıları deneme yanılma yoluyla test edilerek ortalama hata değerlerinin en aza indirgenmesi amaçlanmıştır.

Çizelge 7.19. Yazar ve Yazı Alanı Tanıma için Oluşturulan YSA Yapıları

YSA#	AKS	HKNS	TF	ÖA	MSE
#1	2	20,40,1	S,T,LN	LM	7.29E-17
#2	2	20,40,1	S,T,LN	GDA	1.79
#3	2	40,60,1	S,T,LN	LM	2.18E-17
#4	2	20,40,1	S,T,LN	GD	1.8
#5	2	25,50,1	S,T,LN	LM	4.92E-19
#6	3	20,40,60,1	S,T,T,LN	LM	8.36E-21
#7	2	20,40,1	S,T,LN	GDX	1.79E+00
#8	2	20,40,1	S,T,LN	GDM	1.87E+00
#9	**2**	**45,55,1**	**S,T,LN**	**LM**	**2.89E-28**
#10	3	30,50,40,1	S,S,T,LN	LM	8.27E-28

ÖA: Öğrenme Algoritması, AKS: Ara Katman Sayısı, HKNS: Her Katmanda Nöron Sayısı, TF: Transfer Fonksiyonu,TH: Tanjant Hiperbolik, S: Sigmoid, LN: Doğrusal, MSE: Ortalama Hataların KareköküToplamı, LM: Levenberg-Marquardt, GD: Dereceli Azalan Geri Yayılım, GDA: Adaptif Öğrenme Oranlı Dereceli Azalan Geri Yayılım, GDX: Adaptif Öğrenme Oranlı ve Momentumlu Gradyan Azalan

Yapılan denemeler sonucunda minimum hata oranına sahip olan YSA#9 yapısı tercih edilmiş olup bold (koyu renk) olarak Çizelge 7.19'da ifade edilmiştir. YSA#3 modeli Şekil 7.9'da gösterildiği gibi olup giriş katmanında 30 nöron, 1. ara katmanında 45 nöron, 2. ara katmanında 55 nöron ve transfer fonksiyonu olarak 1. ara katmanda logaritmik sigmoid fonksiyon ve 2. ara katmanda tanjant hiperbolik fonksiyonu kullanılmaktadır. Oluşturulan YSA modelinin eğitim aşamasında sıfıra hataya ulaşmak için epok değeri 1000 olarak seçilmiş olup eğitim aşaması sonucu elde edilen performans değeri 2.89E-28 olarak bulunmuştur. Önerilen bu YSA modelinde öğrenme algoritması olarak Levenberg-Marquardt kullanılmaktadır.



Şekil 6.9. Hem Yazar Hem de Yazı Türü Tanıma için Önerilen YSA modeli

Önerilen YSA modelinin performans değerlendirilmesinin yapılması için çeşitli deneyler gerçekleştirilmiştir. Deneylerin gerçekleştirilmesi için oluşturulan veri seti içinde eğitim ve test veri setlerinin elde edilmesi için 5 k katlı çapraz doğrulama yapılmıştır. Gerçekleştirilen testler sonucunda elde edilen doğruluk değerleri Çizelge 7.20'de gösterilmektedir.

Çizelge 7.20. Yazar ve Yazı Alanı Tanıma için Önerilen YSA Modelinin Performans Değerlendirmesi

Test	Ayşe Arman/ Yaşam	İclal Aydın / Yaşam	Hadi Uluengin/ Siyaset	Emre Aköz / Siyaset	Vahap Munyar /Ekonomi	Güngör Uras/ Ekonomi
#1	30%	55%	50%	50%	40%	50%
#2	40%	40%	50%	40%	50%	60%
#3	45%	45%	55%	45%	60%	60%
#4	30%	50%	40%	60%	40%	40%
#5	30%	40%	50%	40%	50%	60%
Ort	35%	46%	49%	47%	48%	54%

7.4. K-NN (K Nearest Neighborhood) ile Yazar Tanıma

Makine öğrenmesi yöntemlerinden biri olan k-NN'nin (k en yakın komşu) basit ve etkili çözümler sunmasından dolayı birçok alanda kullanıldığı gibi yazar tanıma probleminin çözümünde de kullanılmaktadır [25, 27]. Bu sınıflandırma kategorik verilerden daha çok sayısal veriler üzerinde uygulama kolaylığı sağlamasından dolayı bu alanda tercih edilmektedir [70]. Birbirine yakın olan verilerin aynı sınıfa ait olması gerektiği düşüncesine dayanmaktadır. Önceden sınıflandırılmış veriler kullanılarak yeni gelen verinin özelliklerine göre hangi sınıfa ait olduğuna karar verilir. Bunun için sınıfı belli olmayan verinin sınıfları belli olan verilere olan uzaklıkları hesaplanır ve en yakın k adet içinde en fazla hangi sınıf var ise o sınıfa dâhil edilir. k değerinin belirlenme aşaması en önemli adımdır. Bu sınıflandırma gürültülü veriler üzerinde başarılı sonuçlar vermesine rağmen yeni bir verinin sınıfı belirlenirken tüm verilere olan uzaklıklarının hesaplanması zaman alıcı bir işlemdir.

Uzaklık hesabı için çeşitli uzaklık hesaplama yöntemleri kullanılmaktadır [25, 26]. Alanyazın incelemesi yapıldığı zaman sıklıkla kullanılan 11 uzaklık hesaplama yönteminin bulunduğu fark edilmiştir.

- Öklid Uzaklık

İki nokta arasındaki uzaklığı bulmak için kullanılan bir ölçümdür. Eşitlik 21'de tanımlanmaktadır.

$$d_{st}{}^2 = (x_s - y_t)(x_s - y_t)' \tag{21}$$

- Standartlandırılmış Öklid Uzaklığı

Mesafe bulma problemini optimize etmek için kullanılmaktadır. Eşitlik 22'de tanımlanmaktadır.

$$d_{st}{}^2 = (x_s - y_t)v^{-1}(x_s - y_t)' \tag{22}$$

Burada v nxn boyutlu bir matristir ve j. Di diyagonal elemanı $S(j)^2$ olarak gösterilemektedir. S ters ağırlıkları içeren bir vektördür.

- Mahalanobis Uzaklık

Mahalanobis uzaklık, nokta ile verinin dağılımı arasındaki bir ölçüttür. Eşitlik 23'de tanımlanmaktadır.

$$d_{st}{}^2 = (x_s - y_t)c^{-1}(x_s - y_t)' \quad (23)$$

C kovaryans matrisidir.

- City Block Uzaklık

İki nokta arasındaki city block uzaklığı, kartezyen koordinatlarının mutlak farkının toplamıdır. Eşitlik 24'de gösterilmektedir.

$$d_{st} = \sum_{j=1}^{n} |x_{sj} - y_{tj}| \quad (24)$$

- Minkowski Uzaklık

Minkowski, öklid uzayına dayalı mesafe bulma yöntemidir. Eşitlik 25'de gösterilmektedir.

$$d_{st} = \sqrt[p]{\sum_{j=1}^{n} |x_{sj} - y_{tj}|^p} \quad (25)$$

Minkowski uzaklık özel durum için

p=1 olduğunda Minkowski metriği city blok uzaklığını vermektedir.

p=2 olduğunda Minkowski metriği Öklid uzaklığını vermektedir.

p= ∞ olduğunda Minkowski metriği Chebychev uzaklığını vermektedir.

- Chebychev Uzaklık

İki vektör ya da nokta arasındaki uzaklığı standart koordinatlarla birlikte bulmak için kullanılan mesafe ölçümüdür. Eşitlik 26'da gösterilmektedir.

$$d_{st} = max_j\{|x_{sj} - y_{tj}|\} \quad (26)$$

Chebychev Uzaklık uzaklığı Minkowski metriğinin özel durumudur.

- Kosinüs Uzaklık

Kosinüs uzaklık, iki nokta arasındaki açının eksi kosinüs değeri kullanılarak hesaplama yapılmaktadır. Eşitlik 27'de gösterilmektedir.

$$d_{st} = 1 - \frac{x_s\, y'_t}{\sqrt{(x_s x'_s)\,(y_t y'_t)}} \qquad (27)$$

- Korelasyon Uzaklık

Korelasyona dayalı olan uzaklık, iki vektör arasındaki istatiksel bağımlılığın bir ölçüsüdür. Eşitlik 28'de gösterilmektedir.

$$d_{st} = 1 - \frac{(x_s - \bar{x}_s)\,(y_t - \bar{y}_t)}{\sqrt{(x_s - \bar{x}_s)\,(x_s - \bar{x}_s)'}\sqrt{(y_t - \bar{y}_t)\,(y_t - \bar{y}_t)'}} \qquad (28)$$

$$\bar{x}_s = \frac{1}{n}\sum j\, x_{sj}$$

$$\bar{y}_t = \frac{1}{n}\sum j\, y_{tj}$$

- Hamming Uzaklık

Farklı olan koordinatların yüzdesi olup Eşitlik 28'deki gibi hesaplanmaktadır.

$$d_{st} = \left(\frac{\#(x_{sj} \neq y_{tj})}{n}\right) \qquad (29)$$

- Jaccard Uzaklık

Jaccard değeri, eksi jaccard katsayısından hesaplanmakta ve Eşitlik 30'da hesaplanması gösterilmektedir.

$$d_{st} = \left(\frac{\#(x_{sj} \neq y_{tj}) \cap \left((x_{sj} \neq 0) \cup (y_{tj} \neq 0)\right)}{\#\left[(x_{sj} \neq 0) \cup (y_{tj} \neq 0)\right]}\right) \qquad (30)$$

Bu tez kapsamında önerilen YSA modellerinin başarısının değerlendirilebilmesi için yazar tanıma çalışmalarında başarılı sonuçlar veren K-NN tercih edilmiştir. Yazar tanıma çalışmasında kullanılan derlem kullanılarak KNN öğrenme algoritması ile çeşitli deneyler gerçekleştirilmiştir.

Yazar tanıma için gerçekleştirilen deneylerde k değeri 3,4 ve 5 olarak seçilmiştir. Çizelge 7.21'de k değeri 3 olarak seçildiğinde elde edilen doğruluk oranları gösterilmiştir. Elde edilen sonuçlar önerilen YSA modelinin başarısı ile karşılaştırıldığında ortalama doğruluk oranı önerilen YSA modeli kullanılarak gerçekleştrilen deneylerde ortalama 77% olarak elde edilmiş iken yazar tanıma da başarısı kanıtlanmış olan KNN öğrenme algoritmasında 59% olarak bulunmuştur.

Çizelge 7.21. k=3 iken Yazar Tanıma için Gerçekleştirilen Deneyler

Test	Ayşe Arman	İclal Aydın	Hadi Uluengin	Emre Aköz	Vahap Munyar	Güngör Uras	Farklı Yazar
#1	20%	30%	90%	50%	40%	80%	60%
#2	20%	30%	100%	30%	20%	30%	50%
#3	100%	30%	80%	80%	30%	40%	50%
#4	100%	50%	90%	70%	100%	50%	60%
#5	50%	80%	100%	20%	70%	40%	40%
ortalama	**58%**	**44%**	**92%**	**50%**	**52%**	**48%**	**52%**

Çizelge 7.22'de k değeri 4 seçildiği zaman elde edilen doğruluk oranları gösterilmektedir. k değeri 4 iken yazar tanıma için elde edilen ortamala doğruluk değerleri Hadi Uluengin için en yüksek iken Güngör Uras için en düşük olarak hesaplamıştır.

Çizelge 7.22. k=4 iken Yazar Tanıma için Gerçeklseştirilen Deneyler

Test	Ayşe Arman	İclal Aydın	Hadi Uluengin	Emre Aköz	Vahap Munyar	Güngör Uras	Farklı Yazar
#1	80%	70%	100%	50%	70%	40%	50%
#2	70%	50%	90%	40%	30%	40%	60%
#3	100%	40%	90%	60%	50%	60%	40%
#4	90%	50%	80%	50%	90%	50%	50%
#5	50%	80%	100%	20%	80%	60%	50%
ortalama	**78%**	**58%**	**92%**	**44%**	**64%**	**50%**	**50%**

Son olarak k değeri 5 seçilmiş ve elde edilen doğruluk değerleri Çizelge 7.23'de gösterilmektedir. Elde edilen değerler incelendiği zaman bu deney içinde (k=5) en yüksek başarı oranı Hadi Uluengin için elde edilmişken en düşük yazar tanıma oranı İclal Aydın'ın yazılarına aittir. Elde edilen ortalama yazar tanıma başarı 62% olarak elde edilmiştir.

Çizelge 7.23. k=5 iken Yazar Tanıma için Gerçekleştirilen Deneyler

Test	Ayşe Arman	İclal Aydın	Hadi Uluengin	Emre Aköz	Vahap Munyar	Güngör Uras	Farklı Yazar
#1	80%	60%	80%	50%	60%	40%	50%
#2	80%	40%	100%	50%	30%	40%	30%
#3	90%	40%	90%	80%	40%	30%	50%
#4	90%	50%	90%	60%	70%	40%	50%
#5	60%	80%	100%	30%	70%	30%	40%
ortalama	**80%**	**54%**	**92%**	**54%**	**54%**	**36%**	**44%**

Yapılan deneylerden elde edilen sonuçlar incelendiğinde görüldüğü gibi tez kapsamında önerilen YSA modelleri çoklu yazar tanıma problemi için daha başarılı sonuçlar vermektedir. Ayrıca KNN algoritması yazarı belli olmayan bir yazının yazarını hesaplarken tüm yazılara olan uzaklıklarını hesapladığından dolayı önerilen YSA modellere göre daha uzun zaman almaktadır.

8. SONUÇ

Bu tez kapsamında sözcüksel ve sözdizimsel özellikler kullanılarak Türkçe dili için farklı yazar tanıma modelleri ile yazı alanı (türü) tanıma modeli önerilmiştir. Ayrıca bu çalışmalara ek olarak farklı ihtiyaçlara karşılık vermesi için bir metnin ya da yazının hem yazarını hem de türünü belirleyen hibrid bir model geliştirilmiştir. Farklı çalışmalarda kullanılan istatiksel yöntemler yerine bu çalışma içinde yapay zekâ yöntemlerinden biri olan Yapay Sinir Ağları kullanılmış olup bu yöntemin yazar tanıma, yazı alanı tanıma problemleri üzerindeki başarısı değerlendirilmiştir.

Tez kapsamında yapılan yazar tanıma çalışması, bu alan için gerçekleştirilen istatiksel yöntemler ve makine öğrenmesi yöntemlerine ek olarak farklı bir bakış açısı katmaktadır. Alanyazın bölümünde de ifade edildiği gibi yapay zekâ yaklaşımlarının yazar tanıma tasarımında kullanılmasının doğru bir tercih olduğunu ve farklı yapay sinir ağ yapıları seçilirken tez çalışmasında kullanılan MLP yapısının Levenberg Marquardt gibi güçlü öğretim algoritması kullanılarak eğitilmesinin yazar tanıma tasarımında başarıyı arttırdığı tespit edilmiştir. Yapılan deneyler sonucunda Ayşe Arman, İclal Aydın, Güngör Uras yazılarının tanınması için önerilen YSA modelinin yapısı giriş katmanında 30 nöron, birinci ara katmanda 40nöron, ikinci ara katmanda 60 nöron ve çıkış katmanında bir nöron bulunmaktadır. Vahap Munyar için önerilen modelin YSA yapısı ise giriş katmanında 30 nöron, birinci ara katmanda 25 nöron, ikinci ara katmanda 50 nöron ve çıkış katmanında bir nöron bulunmaktadır. Hadi Uluengin için önerilen modelin YSA yapısı ise giriş katmanında 30 nöron, birinci ara katmanda 20 nöron, ikinci ara katmanda 40 nöron ve çıkış katmanında bir nöron bulunmaktadır. Son olarak Emre Aköz için önerilen YSA modeli ise diğer yazar için önerilen YSA modellerinden biraz daha farklıdır. 3 farklı ara katmandan oluşmaktadır. Giriş katmanında 30 nöron, birinci ara katmanda 20 nöron, ikinci ara katmanda 40 nöron, üçüncü ara katmanda ise 60 nöron ve çıkış katmanında bir nöron bulunmaktadır. Yazarı belli olmayan bir yazının yazarını altı yazar içinde arayan ve bu altı yazar dışında bir yazara ait olması durumunda yazının farkı bir yazara ait olduğunu ifade eden YSA modeli ise yapısı giriş katmanında 30 nöron, birinci ara katmanda 40nöron, ikinci ara katmanda 60 nöron ve çıkış katmanında bir nöron bulunmaktadır. Önerilen tüm YSA modellerinde öğrenme algoritması olarak Levenberg Marquardt kullanmaktadır. Ayrıca

transfer fonksiyonları olarak birinci ara katmanda logaritmik sigmoid fonksiyon ve ikinci ara katmanda tanjant hiperbolik fonksiyonu kullanılmıştır. Bu çalışmada kapsamında önerilen YSA modellerinin performans değerlendirilmesi için çeşitli testler gerçekleştirilmiştir. Tekli sınıflandırma problemi olarak ele alınan yazar tanıma problemi için önerilen YSA modellerinin başarısı ortalama 95% olarak bulunmuş olup, önerilen modellerin kararlı ve kesin sonuçlar verdiği görülmüştür. Elde edilen sonuçlar alanyazında yapılan diğer çalışmalarla karşılaştırıldığında oldukça yüksektir. Çoklu sınıflandırma problemi olarak ele alınan yazar tanıma için önerilen YSA modelinin başarısı ise her yazar için yaklaşık 70% olarak elde edilmiştir.

Tez içerisinde yazar tanıma için kullanılan biçimsel özelliklerin yazı alanı (türü) belirlemedeki başarısı değerlendirilmiş ve yazı alanı belirleme için farklı bir YSA modeli önerilmiştir. Bu tez kapsamında ayrıca istatiksel ve sözcüksel özellikleri kullanarak hem yazar hem de yazı alanı belirleyen model geliştirilmiştir. Oluşturulan model ‘Yaşam’, ‘Ekonomi’ ve ‘Siyaset’ alanlarında yazılar yazan 6 yazarın hem yazı alanlarını hem de yazarları tanıyabilme yeteneğine sahiptir. Önerilen modelin başarısının değerlendirilmesi için yapılan testler sonucunda elde edilen doğruluk oranı 54% olarak belirlenmiştir. Yapılan alanyazın çalışmaları incelendiği zaman bu çalışmaya benzer bir çalışmaya rastlanamamıştır. Tez kapsamı içerisinde yapılan bu çalışma alanında ilk olma özelliğine sahiptir. Oluşturulan YSA modellerini değerlendirecek olursak, Levenberg Marquardt öğrenme algoritmasının diğer öğrenme algoritmalarına göre daha yüksek başarılar verdiği gözlenmektedir.

Bu çalışmada kapsamında önerilen YSA modellerinin performans değerlendirilmesi için çeşitli testler gerçekleştirilmiştir. Tekli sınıflandırma problemi olarak ele alınan yazar tanıma problemi için önerilen YSA modellerinin başarısı ortalama 95% olarak bulunmuş olup, önerilen modellerin kararlı ve kesin sonuçlar verdiği görülmüştür. Elde edilen sonuçlar alanyazında yapılan diğer çalışmalarla karşılaştırıldığında oldukça yüksektir. Çoklu sınıflandırma problemi olarak ele alınan yazar tanıma için önerilen YSA modelinin başarısı ise her yazar için yaklaşık 70% olarak elde edilmiştir. Önerilen YSA modellerinin başarısının farklı makine öğrenmesi teknikleri ile kıyaslanabilmesi için KNN öğrenme algoritması kullanılarak çeşitli deneyler gerçekleştirilmiştir. Elde edilen sonuçlar değerlendirildiğinde önerilen YSA modellerinin ortalama başarısının daha

yüksek olduğu görülmüştür. Ayrıca KNN algoritmasının sınıfı belli olmayan bir verinin sınıfını bulurken diğer verilere olan uzaklıklarını hesaplaması nedeniyle bu algoritmanın önerilen YSA modellerine göre gerçekleştirilen testlerde daha uzun zaman harcamasına neden olmuştur. Önerilen YSA modelleri daha kısa süre içerisinde bir yazının yazarını bulabilmekte ya da yazının alanı hakkında bilgi verebilmektedir.

Bu tez kapsamında Türkçe yazar tanıma çalışmalarında kullanılmak üzere köşe yazılarından oluşan geniş bir derlem oluşturulmuştur. Oluşturulan derlem siyaset, ekonomi, yaşam ve spor alanlarında yazılar yazan 167 köşe yazarına ait olan köşe yazılarından oluşmaktadır. Yapılan alanyazın araştırmasından Türkçe dili için böyle geniş bir derleme rastlanmamıştır. Çalışma gerçekleştirilirken karşılaşılan zorluklara değinilecek olursak biri yazar tanıma probleminde kullanılacak olan köşe yazılarının elde edilmesi aşamasıdır. Gazetelerin köşe yazılarının bulunduğu arşiv web sayfalarının yapısının farklı standartlarda ve farklı formatlarda bulunmasından kaynaklanmasından dolayı köşe yazılarının içeriğine erişim sorunudur. Diğer bir zorluk ise eğitim veri setlerinin uygulanacağı YSA modellerinin belirlenmesi işlemidir. Bunun için yazar tanıma, yazı türü tanıma ve hem yazar hem de yazı türü tanıma için uygun modellerin belirlenmesi için çeşitli denemeler yapılmıştır. Önerilen her model için uygun YSA modelinin belirlenmesi için on farklı YSA yapısı oluşturulmuş olup toplamında 90 deney gerçekleştirilmiştir. Eğitim aşamasının uzun ve zaman alıcı olmasından kaynaklı bu adımda zorluk çekilmiştir. Gerçekleştirilen bu deneylerde 5 farklı öğrenme algoritması kullanılmış ve bu algoritmaların performans karşılaştırılması yapılmıştır.

KAYNAKLAR

1. Berry, M. W., “Survey of Text Mining”, Computing Reviews, 45(9), 548, **2004.**
2. Brocard M. L., Traore I., Saad S., Woungang I., “Authorship Verification for Short Messages using Stylometry”, Computer, Information and Telecommunication Systems (CITS), **2013.**
3. Ma J., Li Y., Teng G., Wang F., Zhao Y., “Sequential Pattern Mining for Chinese E-mail Authorship Identification”, The 3rd Intetnational Conference on Innovative Computing Information and Control (ICICIC), **2008.**
4. Ouamour S., Sayoud H., “Authorship Attribution of Ancient Texts Written by Ten Arabic Travelers Using a SMO-SVM Classifier”, International Conference on Communication and Information Technology (ICCIT), Digital Information Management, Hammamet, **2012.**
5. Raghavan S., Kovashka A., Mooney R., “Authorship Attribution using Probabilistic Context-free Grammars”, Proceedings of the ACL 2010 Conference Short Papers, pages 38–42,Uppsala, Sweden, **2010.**
6. Shaker K., Corne D., “Authorship Attribution in Arabic Using a Hybrid of Evolutionary Search and Linear Discriminant Analysis”, Computational Intelligence (UKCI), **2010.**
7. Zheng R., Li J., Chen H., Huang Z., “A Framework for Authorship Identification of Online Messages: Writing Style Features and Classification Techniques”, Journal of the American Society of Information Science and Technology, 57(3), 378-393, **2006.**
8. Okuno S., Asai H., Yamana H., “A Challenge of Authorship Identification for Ten-thousand-scale Microblog Users”, IEEE International Conference Big Data (Big Data), **2014.**
9. Bradley J. K., Patrick G. K., A. Roth., "Author Identification from Citations", Dept. Comput. Sci., Carnegie Mellon Univ., Pittsburgh, PA, USA, Tech. Rep**, 2008.**
10. Abbasi A., Hsinchun C., "Applying Authorship Analysis to Extremist-group Web Forum Messages", IEEE Intelligent Systems, **2005.**
11. Ma J., Teng G., Chang S., Zhang X., Xiao K., “Social Network Analysis based on Authorship Identification for Cybercrime Investigation”, In Pacific-Asia

Workshop on Intelligence and Security Informatics, pp. 27-35 ,Springer Berlin Heidelberg, **2011**.

12. Berry D., Edward S., "Clustering Technical Documents by Stylistic Features for Authorship Analysis", SoutheastCon, IEEE, **2015**.
13. Sikos J., David P., Nizar H., Reem F., "Authorship Analysis of Inspire Magazine through Stylometric and Psychological Features", In Intelligence and Security Informatics Conference (JISIC), 2014 IEEE Joint, pp. 33-40. IEEE, **2014**.
14. Diederich J., Kindermann J., Leopold E., Paass G., “Authorship Attribution with Support Vector Machines”, Applied intelligence, **2003**.
15. Grieve J., “Quantitative Authorship Attribution: An Evaluation of Techniques”, Literary and Linguistic Computing, **2007**.
16. Solorio T., Pillay S., Raghavan S., Montes-y-Gómez M., “Modality Specific Meta Features for Authorship Attribution in Web Forum Posts”, InIJCNLP, **2011**.
17. Ebrahimpour M., Putniņš T.J., Berryman M.J., Allison A., Ng B.W., Abbott D., “Automated Authorship Attribution using Advanced Signal Classification Techniques”, PloS, **2013**.
18. Rappoport R., Schwartz O., Tsur A., Koppel M., "Authorship Attribution of Micro-Messages", **2013**.
19. Layton R., Watters P., Dazeley R., “Automated Unsupervised Authorship Analysis using Evidence Accumulation Clustering”, Natural Language Engineering, 19(01):95-120, **2013**.
20. Sapkota U., Solorio T., Montes-y-Gómez M., Rosso P., “The use of Orthogonal Similarity Relations in the Prediction of Authorship”, In International Conference on Intelligent Text Processing and Computational Linguistics, Springer Berlin Heidelberg, **2013**.
21. Plakias S., Stamatatos E., “Tensor Space Models for Authorship Identification”, In Hellenic Conference on Artificial Intelligence, Springer, **2008**.
22. Escalante H.J., Solorio T., Montes-y-Gómez M., “Local histograms of Character n-grams for Authorship Attribution”, In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, **2011**.

23. Ekinci E., Takçı H., “Using Authorship Analysis Techniques in Forensic Analysis of Electronic Mails”, 20th Signal Processing and Communications Applications Conference (SIU), **2012**.
24. Demir N. M.,“Authorship Categorization With Neural Network”, Southeast Europe Journal of Soft Computing, **2012**.
25. Kaban Z., Diri B., "Genre and Author Detection in Turkish Texts using Artificial Immune Recognition Systems", 16th Signal Processing, Communication and Applications Conference. IEEE, **2008**.
26. Türkoğlu F., Diri B., Amasyalı M. F., "Author Attribution of Turkish Texts by Feature Mining", International Conference on Intelligent Computing, Springer, **2007**.
27. Mosteller F., Wallace D. L., "Inference and Disputed Authorship: The Federalist”, Addison-Wesley, Massachusetts, **1964**.
28. Burrows J. F., “Word Patterns and Story Shapes: The Statistical Analysis of Narrative Style”, Literary and Linguistic Computing, 2:61–67, **1987**.
29. Koppel M., Schler J., “Exploiting Stylistic Idiosyncrasies for Authorship Attribution”, In Proceedings of IJCAI'03 Workshop on Computational Approaches to Style Analysis and Synthesis (Vol. 69, p. 72), **2003**.
30. Baayen H., Halteren H. V., Neijt A.,Tweedie F., ”An Experiment in Authorship Attribution”, JADT, **2002**.
31. Holmes D.I., “The Evolution of Stylometry in Humanities Scholarship”, Literary and Linguistic Computing, 13(3), pp.111-117, **1998**.
32. Mendenhall T.C., ”The Characteristic Curves of Composition”, Science, pp.237-249, **1887**.
33. Mascol C., "Curves of Pauline and PseudoPauline Style", Unitarian Review, 30 Pages: 452–60, **1888**.
34. Yule, “On Sentence Length as a Statistical Characteristic of Style in Prose”, Biometrika,**1938**.
35. Zipf G.K., “Selected Studies of the Principle of Relative Frequency in Language”, Cambridge, MA.: Harvard University Press, **1932**.

36. Mosteller F., Wallace D.L., “Inference and Disputed Authorship: The Federalist”, Reading, Massachusetts: Addison-Wesley, **1964**.
37. https://www.scientificamerican.com/article/how-a-computer-program-helped-show-jk-rowling-write-a-cuckoos-calling/ (19.11.2016)
38. http://news.stanford.edu/news/2010/february1/unabomber-ethics-question-020110.html (19.11.2016)
39. http://techland.time.com/2011/06/03/how-to-write-like-mark zuckerberg/?hpt=te_bn1 (19.11.2016)
40. Aslanturk O., Sezer E. A., Sever H., Raghavan V.., “Application of Cascading Rough Set-Based Classifiers on Authorship Attribution”, In Granular Computing (GrC), pp. 656-660, **2010**.
41. Aslanturk Oğuz, “Turkish Authorship Analysis with an Incremental and Adaptive Model”, Doctoral dissertation, Hacettepe University, **2014**.
42. Brocardo M. L., Traore I., Saad S., Woungang I., “Authorship Verification for Short Messages using Stylometry”, In Computer, Information and Telecommunication Systems (CITS), **2013**.
43. Halvani O., Winter C., Pflug A., "Authorship Verification for Different Languages, Genres and Topics",Digital Investigation, **2016**.
44. Maitra P., Ghosh S., Dipankar D., "Authorship Verification: An Approach based on Random Forest", Working Notes Papers of the CLEF, **2015**.
45. Argamon S., Koppel M., Fine J., Shimoni A. R., “Gender, Genre,and Writing Style in Formal Written Texts”, **2003**.
46. Burger J. D., Henderson J. C., “An Explorationof Observable Features Related to Blogger Age”, In Computational Approaches to Analyzing Weblogs, AAAI Spring Symposium. AAAI Press, **2006**.
47. Schler J, Koppel M, Argamon S., Pennebaker J. W., “Effects of Age and Gender on Blogging”, AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs, pages 199–205, **2006**.
48. Koppel M., Schler J., Zigdon K “Determining an Author’s Native Language by Mining a Text for Errors”, In Proc. KDD, pages 624– 628, **2005.**

49. Pennebaker J.W., King L.A., “Linguistic styles: Language use as an Individual Difference”, Journal of personality and social psychology, 77(6), p.1296,**1999**.
50. Pennebaker J.W., Mehl M.R., Niederhoffer K.G., “Psychological Aspects of Natural Language Use: Our words, Our selves”, Annual review of psychology, 54(1), pp.547-57, **2003**.
51. Abbasi A., Chen H.,”Writeprints: A stylometric approach to identity-level identification and similarity detection in cyberspace”, ACM Transactions on Information Systems (TOIS), 26(2), p.7., **2008**.
52. Amasyalı M. Fatih, Diri B., "Automatic Turkish Text Categorization in terms of Author, Genre and Gender", International Conference on Application of Natural Language to Information Systems. Springer Berlin Heidelberg, **2006**.
53. Kucukyilmaz T, Cambazoglu BB, Aykanat C, Can F, “Chat mining: Predicting User and Message Attributes in Computer-Mediated Communication”, Information Processing & Management, 44(4):1448-66, **2008**.
54. Patton J. M, Can F., “A stylometric Analysis of Yaşar Kemal’s İnce Memed Tetralogy”, Computers and the Humanities, **2004**.
55. Varol M, “Metin Madenciliği Yöntemlerini Kullanarak Türkçe Dokümanlarda Tür ve Yazar Tanıma”, Süleyman Demirel Üniversitesi Fen Bilimleri Enstitüsü, Yüksek Lisans Tezi, **2011**.
56. Walter D., “Explanation in Computational Stylometry”, In Proceedings of the 14th International Conference on Computational Linguistics and Intelligent Text Processing (CICLing), Lecture Notes in Computer Science (LNCS), Springer, **2013**.
57. Mendenhall T. C.,“The Characteristic Curves of Composition”, 237–49, **1887**.
58. De Vel O., Anderson A., Corney M., Mohay G., “Mining E-mail Content for Author Identification Forensics”, SIGMOD Record, 30(4), 55-64, **2001**.
59. Teng G., Lai M., Ma J., Li Y, “E-mail Authorship Mining based on SVM for Computer Forensic”, In Proceedings of the International Conference on Machine Learning and Cybernetics, 2 (pp. 1204-1207), **2004**.
60. Zheng R., Li J., Chen H., Huang Z., “A framework for Authorship Identification of Online Messages: Writing Style Features and Classification Techniques”,

Journal of the American Society of Information Science and Technology, 57(3), 378-393, **2006**.

61. Koppel M., Schler J., Argamon S. “Computational Methods in Authorship Attribution”, Journal of the American Society for Information Science and Technology, 60(1), 9–26, **2008**.
62. Abbasi A., Chen H., “Writeprints: A stylometric Approach to Identity-level Identification and Similarity Detection in Cyberspace”, ACM Transactions on Information Systems, 26(2):1–29, **2008**.
63. Rudman J., “The State of Authorship Attribution Studies: Some Problems and Solutions. Computers and the Humanities”, 31(4), 351-365, **1998**.
64. El S M., Kassau I., “Authorship Analysis Studies: A survey”, International Journal of Computer Applications, **2014**.
65. http://pages.cs.wisc.edu/~dpage/cs760/evaluating.pdf. (19.11.2016)
66. Davis J., Mark G., "The Relationship between Precision-Recall and ROC curves." Proceedings of the 23rd international conference on Machine learning. ACM, **2006**.
67. http://citeseerx.ist.psu.edu/ (19.11.2016)
68. Murphy K. P., “Naive Bayes Classifiers”, University of British Columbia, **2006**.
69. Tong, S., & Koller, D., “Support Vector Machine Active Learning with Applications to Text Classification”, Journal of machine learning research,), 45-66, **2001**.
70. Safavian, S. R., Landgrebe D., "A survey of Decision Tree Classifier Methodology", IEEE transactions on systems, man, and cybernetics 21.3, **1991.**
71. Soucy P., Guy W. M., "A simple KNN Algorithm for Text Categorization", Data Mining, ICDM, **2001**.
72. Russell, S., Norvig, P.,“A Modern Approach”, Artificial Intelligence, Prentice-Hall, Egnlewood Cliffs,**1995**,
73. Haykin S. S., “Neural Networks and Learning Machines”, Upper Saddle River, USA, Pearson, **2009.**
74. Zurada Jacek M., “Introduction to Artificial Neural Systems”, Vol. 8. St. Paul: West, **1992.**

75. Gurney K., “An Introduction to Neural Networks”, CRC press, **1997.**
76. Anderson J. A.,“An Introduction to Neural Networks”, MIT press, **1995.**
77. Sağıroğlu Ş., Beşdok E., Erler M., “Mühendislikte Yapay Zeka Uygulamaları-1:Yapay Sinir Ağları”, Ufuk Kitabevi, Kayseri, 10-100, **2003.**
78. 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.**
79. Yavanoglu U., Colak M., Caglar B., Cakir S., Milletsever O., Sağıroğlu, Ş.,”Intelligent Approach for Identifying Political Views over Social Networks”, In Machine Learning and Applications (ICMLA), 12th International Conference on Vol. 2, pp. 281-287, IEEE, **2013.**

Ek-1

Gazete Adı	Yazar Adı Soyadı	Başlangıç		Son		ALAN
		Yıl	Ay	Yıl	Ay	
VATAN	**Ali Ağaoğlu**	2004	1	2014	12	Siyaset
	Asaf Savaş Akat	2002	1	2014	12	Siyaset
	Can Ataklı	2006	1	2014	12	Siyaset
	Mustafa Mutlu	2004	1	2014	12	Siyaset
	MutluTönbekici	2008	1	2014	12	Siyaset
	Oktay Gönensin	2002	1	2014	12	Siyaset
	Ruhat Mengi	2002	1	2014	12	Siyaset
	Ruşen Çakır	2003	1	2014	12	Siyaset
	Süleyman Ateş	2002	1	2012	12	Siyaset
SABAH	**Hasan Bülent Kahraman**	2007	1	2014	12	Siyaset
	Mahmut Ovur	2004	1	2014	12	Siyaset
	Mehmet Barlas	2012	1	2014	12	Siyaset
	Emre Aköz	2007	1	2014	12	Siyaset
	Engin Ardıç	2009	1	2014	12	Siyaset
	Okan Müderrisoğlu	2003	1	2014	12	Siyaset
	Şeref Oğuz	2007	1	2014	12	Siyaset
CUMHURİYET						
	Utku Çakırözer	2009	1	2014	12	Siyaset
	Şükran Sosner	2008	1	2014	12	Siyaset
	Sadık Çelik	2008	1	2014	12	Siyaset
	Özgen Acar	2008	1	2014	12	Siyaset
	Orhan Erinç	2008	1	2014	12	Siyaset
	Orhan Bursalı	2008	1	2014	12	Siyaset
	Oktay Ekinci	2008	1	2014	12	Siyaset
	Mustafa Balbay	2008	1	2014	12	Siyaset
	Mümtaz Soysal	2008	1	2014	12	Siyaset
	Işıl Özgentürk	2008	1	2014	12	Siyaset
	Hikmet Çetinkaya	2008	1	2014	12	Siyaset
	Güray Öz	2008	1	2014	12	Siyaset
	Erol Manisalı	2008	1	2014	12	Siyaset
	Emre Kongar	2008	1	2014	12	Siyaset
	Cüneyt Arcayürek	2008	1	2014	12	Siyaset
	Bekir Coşkun	2008	1	2014	12	Siyaset
	Bedri Baykam	2008	1	2014	12	Siyaset
	Çiğdem Toker	2008	1	2014	12	Siyaset

MİLLİYET						
	Can Dündar	2008	1	2014	12	Siyaset
	Cem Kılıç	2012	1	2014	12	Siyaset
	Çetin Altan	2011	1	2014	12	Siyaset
	Fikret Bila	2008	1	2014	12	Siyaset
	Hasan Cemal	2008	1	2014	12	Siyaset
	Hasan Pulur	2011	1	2014	12	Siyaset
	Mehmet Tezkan	2011	1	2014	12	Siyaset
	Melih Aşık	2008	1	2014	12	Siyaset
	Nihat Ali Özcan	2012	1	2014	12	Siyaset
	Sami Kohen	2008	1	2014	12	Siyaset
	Semih İdiz	2008	1	2014	12	Siyaset
	Taha Akyol	2008	1	2011	12	Siyaset
	Yaman Törüner	2008	1	2014	12	Siyaset
	Serpil Cevikan	2010	1	2014	12	Siyaset
HÜRRİYET						
	Ahmet Hakan	2005	1	2014	12	Siyaset
	Emin Çölaşan	1997	1	2007	12	Siyaset
	Ertuğrul Özkök	2005	1	2014	12	Siyaset
	Fatih Çekirge	2006	1	2014	12	Siyaset
	Erdal Sağlam	2005	1	2014	12	Siyaset
	Mehmet Barlas	2006	1	2010	12	Siyaset
	Mehmet Yılmaz	2005	1	2014	12	Siyaset
	Rauf Tamer	2006	1	2014	12	Siyaset
	Sedat Ergin	1997	1	2005	12	Siyaset
	Şükrü Küçükşahin	2003	1	2005	12	Siyaset
	Taha Akyol	2011	1	2014	12	Siyaset
	TolgaTanış	2008	1	2014	12	Siyaset
	Yalçın Bayer	2005	1	2014	12	Siyaset
	Yalçın Doğan	2002	1	2014	12	Siyaset
	Yılmaz Özdil	2007	1	2014	12	Siyaset
	Hadi Uluengin	1997	1	2005	12	Siyaset
	Cüneyt Ulsever	1998	1	2005	12	Siyaset
POSTA						
	Hakan Çelik	2009	1	2014	12	Siyaset
	Nedim Şener	2010	1	2014	12	Siyaset
	Esra Karayel	2010	1	2014	12	Siyaset
	Rauf Tamer	2009	1	2014	12	Siyaset

Gazete Adı	Yazar Adı Soyadı	Başlangıç		Son		ALAN
		Yıl	Ay	Yıl	Ay	
FANATİK	Asena Özkan	2010	1	2014	12	Spor
	Can Cobanoğlu	2007	1	2014	12	Spor
	Cem Dizdar	2007	1	2014	12	Spor
	Edip Adanır	2007	1	2014	12	Spor
	Gökhan Germen	2007	1	2014	12	Spor
	Hakan Can	2007	1	2014	12	Spor
	Hasan Ali Atasoy	2007	1	2014	12	Spor
	Hasan Tankaya	2007	1	2014	12	Spor
	Mehmet Demircan	2012	1	2014	12	Spor
	Mehmet Demirkol	2011	1	2014	12	Spor
	Nezih Alkış	2010	1	2014	12	Spor
	Oğuz Dizdar	2007	1	2014	12	Spor
	Ömer Ural Kükrer	2009	1	2014	12	Spor
	Orhan Yıldırım	2007	1	2014	12	Spor
	Serdar Dinçbaylı	2007	1	2014	12	Spor
	SerdarTatlı	2009	1	2014	12	Spor
	Serhat Demirtaş	2007	1	2014	12	Spor
	Serkan Akcan	2012	1	2014	12	Spor
	Tamer Bağlan	2007	1	2014	12	Spor
	Tunç Kayacı	2007	1	2014	12	Spor
	Yalçın Dümer	2007	1	2014	12	Spor
	Yemen Ekşioğlu	2007	1	2014	12	Spor
	Zafer Büyükavcı	2007	1	2014	12	Spor
	Necil Ulgen	2007	1	2014	12	Spor
CUMHURİYET	Arif Kızılyalın	2011	1	2014	12	Spor
	Osman Korkmazel	2011	1	2014	12	Spor
	Oğuz Tongsir	2011	1	2014	12	Spor
	Muhittin Bosat	2011	1	2014	12	Spor
	Hilmi Türkay	2011	1	2014	12	Spor
	Barbaros Talı	2011	1	2014	12	Spor
HÜRRİYET	Ateş Bakan	2012	1	2014	12	Spor
	Bayram Aydın	2011	1	2014	12	Spor
	Engin Krotzer	1997	1	2005	12	Spor
MİLLİYET	Atilla Gökce	2008	1	2012	12	Spor
	Bilal Meşe	2008	1	2012	12	Spor
	Cemal Ersen	2008	1	2012	12	Spor
	Ediz Sırapınar	2008	1	2012	12	Spor
	Ercan Güven	2008	1	2012	12	Spor

	Erkan Öncel	2008	1	2012	12	Spor
	Levent Kalkan	2008	1	2012	12	Spor
	Osman Senher	2008	1	2012	12	Spor
	Uğur Meleke	2008	1	2012	12	Spor
	Ümit Avcı	2008	1	2012	12	Spor
	Yavuz Kocaömer	2008	1	2012	12	Spor

ÖZGEÇMİŞ

Kimlik Bilgileri

Adı Soyadı : Özlem Yavanoğlu (Milletsever)
Doğum Yeri : Ankara
Medeni Hali : Evli
E-posta : milletseverozlem@gmail.com
Adresi : Dikmen Cad. 1114. Sokak 12/4, Ankara

Eğitim

Lise : 2004 – 2008 Mehmet Emin Resulzade Anadolu Lisesi
Lisans : 2008 – 2012 Gazi Üniversitesi Bilgisayar Mühendisliği
Yüksek Lisans : 2013 – 2016 Hacettepe Üniversitesi Bilgisayar Mühendisliği

Yabancı Dil ve Düzeyi

İngilizce – 73,75 (YDS)

İş Deneyimi

Yazılım Mühendisi, ICterra Bilgi ve İletişim Teknolojileri (20.12.2016 → Halen)

Deneyim Alanları

Veri Madenciliği
Makine Öğrenmesi
Doğal Dil İşleme
Yapay Sinir Ağları

Tezden Üretilmiş Projeler ve Bütçesi

Yok

Tezden Üretilmiş Yayınlar

O. Yavanoglu (Milletsever), "Intelligent Authorship Identification with using Turkish Newspapers Metadata", Proceedings of The 2016 IEEE International Conference on Big Data, Washington, DC, USA, 2016.

Tezden Üretilmiş Tebliğ ve/veya Poster Sunumu ile Katıldığı Toplantılar

----

HACETTEPE ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ
YÜKSEK LİSANS/DOKTORA TEZ ÇALIŞMASI ORJİNALLİK RAPORU

HACETTEPE ÜNİVERSİTESİ
FEN BİLİMLER ENSTİTÜSÜ
BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI BAŞKANLIĞI'NA

Tarih: 14/02/2017

Tez Başlığı / Konusu: Stilistik Özellikler Kullanılarak Yazar Tanıma İşinde Yapay Sinir Ağlarının Başarımının Değerlendirilmesi: Türkçe Köşe Yazıları

Yukarıda başlığı/konusu gösterilen tez çalışmamın a) Kapak sayfası, b) Giriş, c) Ana bölümler d) Sonuç ve e)Kaynakça kısımlarından oluşan toplam 112 sayfalık kısmına ilişkin, 14/02/2017 tarihinde şahsım/tez danışmanım tarafından *Turnitin* adlı intihal tespit programından aşağıda belirtilen filtrelemeler uygulanarak alınmış olan orijinallik raporuna göre, tezimin benzerlik oranı %10'tür.

Uygulanan filtrelemeler:

1- Kaynakça hariç
2- Alıntılar hariç/~~dâhil~~
3- 5 kelimeden daha az örtüşme içeren metin kısımları hariç

12

Hacettepe Üniversitesi Fen Bilimleri Enstitüsü Tez Çalışması Orjinallik Raporu Alınması ve Kullanılması Uygulama Esasları'nı inceledim ve bu Uygulama Esasları'nda belirtilen azami benzerlik oranlarına göre tez çalışmamın herhangi bir intihal içermediğini; aksinin tespit edileceği muhtemel durumda doğabilecek her türlü hukuki sorumluluğu kabul ettiğimi ve yukarıda vermiş olduğum bilgilerin doğru olduğunu beyan ederim.

Gereğini saygılarımla arz ederim.

14.02.2017
Tarih ve İmza

Adı Soyadı:	Özlem YAVANOĞLU (MİLLETSEVER)
Öğrenci No:	N13220149
Anabilim Dalı:	Bilgisayar Mühendisliği
Programı:	
Statüsü:	☒ Y.Lisans ☐ Doktora ☐ Bütünleşik Dr.

DANIŞMAN ONAYI

Doç. Dr. Ebru SEZER
UYGUNDUR.

(Unvan, Ad Soyad, İmza)