BAŞKENT ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ
BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİMDALI
BİLGİSAYAR MÜHENDİSLİĞİ TEZLİ YÜKSEK LİSANSPROGRAMI

DERS VİDEOLARININ İÇERİK TABANLI ERİŞİMİ

HAZIRLAYAN

VEYSEL SERCAN AĞZIYAĞLI

YÜKSEK LİSANS TEZİ

ANKARA – 2020

BAŞKENT ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ
BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİMDALI
BİLGİSAYAR MÜHENDİSLİĞİ TEZLİ YÜKSEK LİSANSPROGRAMI

DERS VİDEOLARININ İÇERİK TABANLI ERİŞİMİ

HAZIRLAYAN

VEYSEL SERCAN AĞZIYAĞLI

YÜKSEK LİSANS TEZİ

TEZ DANIŞMANI

PROF. DR. HASAN OĞUL

ANKARA – 2020

BAŞKENT ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

Bilgisayar Mühendisliği Anabilim Dalı Bilgisayar Mühendisliği Tezli Yüksek Lisans Programı çerçevesinde Veysel Sercan Ağzıyağlı tarafından hazırlanan bu çalışma, aşağıdaki jüri tarafından Yüksek Lisans Tezi olarak kabul edilmiştir.

Tez Savunma Tarihi: 24 / 08 / 2020

Tez Adı: Ders Videolarının İçerik Tabanlı Erişimi

Tez Jüri Üyeleri (Unvanı, Adı - Soyadı, Kurumu) **İmza**

Doç. Dr. Aydın Kaya, Çankaya Üniversitesi ……..

Prof. Dr. Hasan Oğul, Başkent Üniversitesi ……..

Doç. Dr. Mustafa Sert, Başkent Üniversitesi ……..

ONAY

Prof. Dr. Faruk Elaldı
Fen Bilimleri Enstitüsü Müdürü
Tarih: … / … / ……..

BAŞKENT ÜNİVERSİTESİ
FEN BİLİMLER ENSTİTÜSÜ
YÜKSEK LİSANS / DOKTORA TEZ ÇALIŞMASI ORİJİNALLİK RAPORU

Tarih: 30 / 08 / 2020

Öğrencinin Adı, Soyadı : Veysel Sercan Ağzıyağlı
Öğrencinin Numarası : 21810100
Anabilim Dalı : Bilgisayar Mühendisliği
Programı : Bilgisayar Mühendisliği Tezli Yüksek Lisans
Danışmanın Unvanı/Adı, Soyadı : Prof. Dr. Hasan Oğul
Tez Başlığı : Ders Videolarının İçerik Tabanlı Erişimi

Yukarıda başlığı belirtilen Yüksek Lisans tez çalışmamın; Giriş, Ana Bölümler ve Sonuç Bölümünden oluşan, toplam 77 sayfalık kısmına ilişkin, 30 / 08 / 2020 tarihinde tez danışmanım tarafından Turnitin adlı intihal tespit programından aşağıda belirtilen filtrelemeler uygulanarak alınmış olan orijinallik raporuna göre, tezimin benzerlik oranı

%5 'dir.

Uygulanan filtrelemeler:
1. Kaynakça hariç
2. Alıntılar hariç
3. Beş (5) kelimeden daha az örtüşme içeren metin kısımları hariç

"Başkent Üniversitesi Enstitüleri Tez Çalışması Orijinallik Raporu Alınması ve Kullanılması Usul ve Esaslarını" inceledim ve bu uygulama esaslarında belirtilen azami benzerlik oranlarına 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.
Öğrenci İmzası:…….

ONAY
… / … / 20…
Prof. Dr. Hasan Oğul

TEŞEKKÜR

Yazar, bu çalışmanın gerçekleşmesinde katkılarından dolayı, aşağıda adı geçen kişi ve kuruluşlara içtenlikle teşekkür eder.

Sayın Prof. Dr. Hasan Oğul'a (tez danışmanı), çalışmanın sonuca ulaştırılmasında ve karşılaşılan güçlüklerin aşılmasında her zaman yardımcı ve yol gösterici olduğu için…

ÖZET

Veysel Sercan AĞZIYAĞLI
DERS VİDEOLARININ İÇERİK TABANLI ERİŞİMİ
Başkent Üniversitesi Fen Bilimleri Enstitüsü
Bilgisayar Mühendisliği Anabilim Dalı
2020

İnternet teknolojisinin ve içerik sağlayıcıların artmasıyla birlikte tüm dünyada ders videolarında da diğer videolarda olduğu gibi büyük bir artış gerçekleşmiştir. Covid 19 salgınının dünyayı etkisi altına alması hem çevrimiçi eğitim içeriğinin artmasına hem de uzaktan eğitimin hızlı bir şekilde artmasına sebep olmuştur. Video sayılarında bu yüksek artış hızı öğrencilerin video içeriklerine erişimini çok zorlaştırmıştır. Bu çalışmada önerilen yöntemler videoların içerik tabanlı erişimini sağlamak üzerinedir. Ders videoları metinsel, işitsel ve görsel içeriğe sahiptir. Bu çalışma için 110 videoluk bir ders videosu veri kümesi oluşturulmuştur. Veri kümesindeki ders videolarının metinsel içerikleri Optic Character Recognition (OCR) teknolojisi ile çıkarılmıştır. Bu çıkarılan içerikler üzerinden 3 adet geleneksel makine öğrenimi yöntemi ve 1 adet derin öğrenme yöntemi ile sınıflandırma gerçekleştirilmiştir. Kullanılan geleneksel makine öğrenimi yöntemleri Support Vector Machine, Naive Bayes ve Random Forest yöntemleridir. Kullanılan derin öğrenme yöntemi ise Long Short Term Memory yöntemidir. Bu çalışma, makine öğrenme yöntemlerinin ve derin öğrenme yönteminin ders videolarının içerik tabanlı erişimde kullanılabilmesi için bir yaklaşım önermektedir.

ANAHTAR KELİMELER: Ders Videolarının İçerik Tabanlı Erişimi, Support Vector Machine, Naive Bayes, Random Forest, Long Hort Term Memory

ABSTARCT

Veysel Sercan AĞZIYAĞLI
CONTENT BASED LECTURE VIDEO RETRIEVAL
Baskent University Institute of Science and Engineering
Department of Computer Engineering
2020

By the development of the internet technology and increasing internet providers have risen the amount of lecture videos as well as the other type of contents. While the impact of Covid 19 Pandemic around all over, that also changed the road map of education. Both the number of online educational content and the distance learning source and demand have increased alot. This rate of increase in the content and providers made it difficult to reach exact contents at its finest. The methods suggested in this study aim content based access to videos. Lecture videos have textual, audio and visual content. In order to illuminate this study, a lecture video dataset with 110 videos was created and the textual contents of the lecture videos in the Data Set were extracted by Optical Character Recognition (OCR) technology. Classification has done by three traditional machine learning methods and one deep learning method. Traditional machine learning methods applied in this study are Support Vector Machine, Naive Bayes and Random Forest methods. The deep learning method applied is the Long Short Term Memory method. In this study it is intended using machine learning and deep learning approaches to reach content based lecture videos.

KEYWORDS: Content Based Lecture Video Retrieval, Support Vector Machine, Naive Bayes, Random Forest, Long Hort Term Memory

İÇİNDEKİLER

TABLOLAR LİSTESİ

Sayfa

ŞEKİLLER LİSTESİ

Sayfa

SİMGELER VE KISALTMALAR LİSTESİ

LSTM	Long Short Term Memory
Bi-LSTM	Bidirectional Long Short Term Memory
SVM	Support vector Machine
RF	Random Forest
NB	Naive Bayes
EBA TV	Eğitim Bilişim Ağı Televizyonu
OCR	Optic Character Recognition
ASR	Automatic Speech Recognition
BIC	Bayesian Information Criterion
ML	Machine Learning
www	World Wide Web
TED	Technology, Entertainment, Design
CNN	Convolutional Neural Network

1. GİRİŞ

İnternet teknolojileri geçtiğimiz yıllar içerisinde çok hızlı bir şekilde gelişim göstermiştir. İnternet teknolojilerinin bu kadar bu kadar hızlı gelişmesi medya platformlarının artması sonucunu da paralel olarak yanında getirmiştir. Şu an dünya çapında ya da bölgesel olmak üzere farklı ülkelerde farklı medya platformları bulunmaktadır. Birçok içerik sağlayıcı da sayısız medya platformlarına içerik sağlamak için sürekli üretim yapmaktadır. Medya platformlarının ve içerik sağlayıcıların internet teknolojisi ile paralel hızda artması ile birçok kaynaktaki ders videoları artmıştır. Üniversitelerde, konferanslarda ve birçok benzer kurumda dersler kaydedilmekte ve sonrasında internetten erişime açılmaktadır. Hem özel sektörde hem de kamu kurumlarında benzer şekilde çevrimiçi ders videoları kullanımı gerçekleşmektedir.

Çevrimiçi öğrenme ve uzaktan eğitim toplumumuz için yeni kavramlar değildir. Bunlar zaten internet teknolojisinin kamuya ulaşmasıyla birlikte hayatımızın bir parçası olmaya başlamış kavramlardır. Her ne kadar çevrimiçi öğrenme ve uzaktan eğitim yeni kavramlar olmasalar da son yıllarda insanların bu derslere ulaşma talebinde bir artışa meydana gelmiştir. Udemy, Coursera, O'Reilly, Cambly, Preply vb. platformların artması bu talep artışının bir sonucudur. Covid-19 salgını sırasında çevrimiçi eğitim farklı ülkelerde yükseldiği gibi bizim ülkemizde de yükselmiştir. EBA TV üç televizyon kanalı ile yayın yapmaktadır. Televizyon yayıncılığının yanında internet üzerinden de hizmet sağlamaktadır. Covid-19 salgını görünüşe göre web tabanlı araçlar yoluyla öğrenme faaliyetlerini destekleyecek teknolojilerin önemini haklı çıkarmıştır [1-3].

Covid-19 bir sağlık problemi olması ve alınan tedbirlerinde sağlık açısından olmasına karşın ekonomik ve sosyal hayatı çok kuvvetli düzeyde etkileyerek değiştirmiştir. Bugün birçok üniversite, okul ve kurs eğitimlerini uzaktan yapmaktadır. Öyle görülüyor ki bu hastalık hayatımızı kalıcı olarak değiştirecektir [3].

Video derslerin ve konferansların çevrimiçi olarak ulaşılabilir olmasının çeşitli artıları bulunmaktadır. Bu artıların bir kısmı zamandan ve konumdan bağımsız içeriğe erişimdir. İzleme kolaylığı ve esnek arama gibi geleneksel sınıf derslerinde gerçekleştirilemeyecek faydaları mevcuttur. Kullanıcılar eğer isterlerse ders adı, açıklama, eğitmen ve müfredat gibi üstveri tabanlı arama yapabilirler. Dersleri ve konuları istenen üstverilere göre sınıflandırabilirler. Sadece bu üstverilerden yapılan aramalar ve sınıflandırmalar dahi fazlası ile yararlıdır.

Üstveri özellikleri bir dereceye kadar çevrimiçi arama yeteneği sunabilmesine rağmen, kullanıcılar veritabanı yöneticileri tarafından sağlanan üstverilerle sınırlıdır. Bu üstveriler şu an kullanıcılar tarafından sistemlere girilmektedir. Bunları sağlamak ayrıca işgücü gerekmektedir. Gerekli işgücü kolaylıkla sağlanabilse dahi girilen üstveriler öznellik barındıracaklardır. Bu konularda eğitim almamış kimseler tarafından videoların yanlış üstverilerle etiketlenmesi muhtemel bir problemdir.

Videolar üstveri açıklamalarıyla kolayca temsil edilemeyen görüntü, metin ve konuşma gibi zengin içeriklere de sahiptir. Ortaya çıkan bir ihtiyaç daha ihtimamlı aramayı kolaylaştırmak için ders videolarına otomatik olarak açıklama ekleyecek araçlar geliştirmektir. Bu yüksek lisans tezinde ders videolarını içeriğe dayalı olarak daha belirgin bir şekilde sınıflandırmak için metinsel bir yaklaşım ve bu metinsel yaklaşım ile kullanılabilecek çeşitli algoritmalar önerilmektedir.

Ders videoları görsel, işitsel ve metinsel olarak farklı türlerde bilgileri kendi içlerinde barındırabilmektedir. Bir Tıp Dersindeki Anatomi çizimi görsel bir bilgi kaynağıdır. Herhangi bir derste ya da konferansta anlatan kimsenin söyledikleri işitsel veri kaynağı sınıfındadır. Ders videolarında kullanılan slaytlar ve tahtaya yazılan yazılar metinsel veri kaynağı olarak değerlendirilir.

Bu tezde kullanmak üzere toplam 110 adet videodan oluşan bir veri kümesi oluşturulmuştur. Bu veri kümesi toplam 110 videodan oluşmanın yanında farklı hiyerarşik seviyelerde farklı sınıflarda video içeriği barındırmaktadır. Farklı hiyerarşik seviyelerde farklı içerikler kullanılarak hangi yöntemlerin ne tip veri kümelerinde daha başarılı olduğu irdelenmiştir.

Bu tezde kullanılmış olan yaklaşım metinsel veriye dayalıdır. Bu tezde ders videolarından metinsel veriyi elde etmek için Optic Character Recogntion (OCR) teknolojisi kullanılmıştır. OCR teknolojisi videolardaki metinsel içeriği işlenebilir sayısal hale getirmektedir. OCR işlemi sonrasında iki ayrı ön işleme algoritması kullanılmıştır. İki ayrı ön işleme algoritmasından bir tanesi kelimelerin frekans vektörlerine dayanmaktadır. Bu vektörler üç ayrı makine öğrenimi yöntemi ile sınıflandırılmıştır. Burada kullanılan makine öğrenmesi yöntemleri Naive Bayes[18], Random Forest[20] ve Support Vector Machine[19] yöntemleridir. Bu üç ayrı geleneksel makine öğrenme yöntemi de frekans vektörleri kullanılarak eğitilmiş ve test edilmiştir.

İki ayrı ön işleme algoritmasından diğeri sözcük sıralama vektörleri oluşturmaktadır. OCR tarafından çıkarılan kelimeler burada sıralamalı olarak vektör haline getirilir.

Bu vektörler sıralamalı olarak derin öğrenme yöntemlerinden bir tanesi olan Bidirectional Long Short Term Memory (Bi-LSTM) algoritmasını beslemek için kullanılır.

Önerilen yöntemleri test etmek için farklı hiyerarşilerde farklı anlam seviyelerinde farklı sayılarda videolar kullanılmıştır. Ön işleme algoritmaları sonrasında dört ayrı sınıflandırma yöntemi test edilmiştir. Toplam kullanılan video sayısı 110 adettir. Bu 110 adet video hiyerarşik olarak farklı anlam seviyelerine bölünmüştür. Sistemin başarımı farklı düzeylerde sınıflandırma doğruluğu temel alınarak değerlendirilmiştir.

1.1. Önceki Çalışmalar

Çevrimiçi video ders içerikleri kullanıcılar açısından daha popüler hale gelmiştir. Video miktarı özellikle WWW tabanlı ders video dosyaları hızla artmaktadır. Organizasyonların, üniversitelerin ve araştırma odaklı kurumların birçoğunun önceliklerinden birisi bu tür çevrimiçi ders videoları oluşturmaktadır [8, 11]. Bu durum internetteki video verilerinde büyük artışın sebeplerinden biridir. Son kullanıcılar için internetten ya da internet erişimine açık olmayan video arşivlerinden ilgili videoyu bulmak video sayılarının çok hızlı artmasından dolayı çok zordur. Bu nedenle verimli video arşivleme ve videoya erişim yöntemlerine ihtiyaç bulunmaktadır.

Her ne kadar bu tezde olduğu gibi farklı ders videolarını metinsel veriler üzerinde farklı algoritmaları test eden bir çalışma literatürde bulunmasa da Chand and Oğul [4] ders videoları üzerinde arama işlemi ile ilgili bir analiz yayını yayınlamışlardır. Chand and Oğul [4] 'un çalışmalarında içerik tabanlı arama kullanıldığında olabilecek faydalar ortaya koyulmuştur. Ayrıca içerik tabanlı aramdaki zorluklar ve veri kaynakları anlatılmaktadır.

Ders videolarından sınıflandırma yapabilmek ya da ders videoları içerisinde arama yapabilmek için ders videolarından üstveriler çıkarılması gerekmektedir. Ders videolarında yer alan metinleri, yansı metinleri, yansılardaki şekiller ve yansıdaki matematiksel ifadeler videonun görüntü verisinden çıkarılabilecek üstveri kaynaklarıdır [4]. Ders videolarının ses dosyaları ders ile ilgili üstveriler barındırmaktadır. Ders videolarının ses dosyaları da bu yönü ile ders videolarının görsel yönü kadar üstveri barındırmaktadır.

Chand and Oğul [4]' un çalışmasında analiz edildiği gibi ders videolarının içerisinden gerekli görsel ve işitsel üstverileri OCR ve Automatic Speech recognition (ASR) teknolojileri ile çıkardıktan sonra çıkarılan verilerin bir dizi matematiksel ve istatistiksel işleme tabi tutulması gerekmektedir. Ders videolarından çıkarıldıktan sonra önişleme tutulmuş veriler hem özgül bir konu ile ilgili ders videosu bir arşivde bulmak için hem de ilgili videonun hangi kısmında aranan bilginin olduğunu indeksleyebilmek için çok önemlidir.

İçerik tabalı arama ve içerik tabanlı video alma sistemleri kendi içlerinde zorluklar barındırmaktadır. İçerik tabanlı video erişim ve videoda arama sistemleri veritabanlarında içerikle ilgili üstveriler barındırmalıdır. Bu tür verileri insan eliyle çıkarmak hem zaman alıcı hem de zor şeylerdir. Video üretiminin çok hızlı gerçekleştiği bir platformda insan eliyle çıkarılan video üstverileri bir standartta çıkarmak çok zordur. Bu etiketlemeyi yapan kimseler öznel seçimler yapabilirler. Ders videolarından karelerde oluşacak farklılıklara göre yani kare geçişlerine göre üstveri çıkarmak zor bir işlemdir. Bunun sebebi ders videolarının genellikle tek kamera ile sabit açıdan çekilmesi ayrıca ders videolarında kareler arasında farklılıkların az olmasıdır. Genellikle tek konuşmacı sabit açıdan dersi anlatmaktadır. Ders video karelerinin aynı olması kareler arasında farkı bulmayı zorlaştırmaktadır. Bu durumda farklı karelerden veri metinsel veri çıkaracak OCR yazılımının başarısını düşürmektedir. Ders videolarının genellikle düşük çözünürlüğe sahip düşük kaliteli videolar olması da OCR' ın başarısını olumsuz etkilemektedir. Yansıların parlaklığının ve biçiminin değişken olması OCR başarısını olumsuz etkileyebilir. Konuşmacının hareketli olması kamera ile yansı arasına girmesi OCR' ın videodan çıkaracağı metinsel veriyi olumsuz etkilemektedir.

Yang and Meinel [10] tarafından yapılan çalışmada büyük ders video arşivlerinde otomatik video dizinleme ve video arama için yeni bir yaklaşım sunulmaktadır. Geleneksel video erişim sistemlerinin kullandığı yaklaşımlar ders videoları gibi sahne geçişi hemen hemen hiç olmayan ya da çok az olan videolar için kullanılamamaktadır. Çükü geleneksel yöntemler daha çok videoları oluşturan kareler arasındaki farklardan beslenmektedir. Ders videoları genelde tek bir kameradan çekilmektedir ve diğer videolara oranla düşük görüntü kalitesine sahiptir.

Çevrimiçi ders videoları günümüzde artık tek bir görsel karede iki görüntünün birleşiminde oluşmaktadır. Bir yanda konuşmacı diğer yanda ise üzerine konuşmakta olduğu yansı yer almaktadır. Ders videolarının içeriğini anlamak için metinsel veriler çok anlamlı

olmaktadır. Bu sebepten anahtar kareler çıkarıldıktan sonra OCR teknolojisi ile sayısal hale getirilecek verilerin kullanımı modern sistemlerde önemli yer alacaktır. OCR teknolojisi tarafından tanınan metinsel verilerdeki kelimelerdeki hatalar metin içeriğini anlamada sorun yaşatabilir. Bu sebepten de çıkarılan kelimelerin doğru olup olmadığı kontrol gerektirir.

Yang and Meinel [10]'ın önerdikleri sistemde OCR teknolojisi ile birlikte bir de ASR teknolojisini kullanmıştır. Metinsel ve işitsel kaynaklardan elde edilen verilerin anahtar kelime olarak puanlanması yaklaşımını getirmişlerdir. Bu sayede metinsel ve işitsel kaynaklardan elde edilen veriler alaka düzeylerine göre puanlanarak sistemde yer almıştır. Yapılan çalışmada OCR her üç saniyede bir defa kareler arasındaki ya da karelerden ayrılmış parçalar arasındaki farkı kontrol eder. Üç saniyeden daha az bir sürede yansılarda meydana gelmiş olabilecek değişim sistem tarafından ihmal edilmiş olur.

Yansı geçişlerinde ilk olarak bir önceki başlık ile bir sonraki başlık kontrol edilir. Eğer başlık değişmişse yansıda farklı bir sayfaya geçildiği anlaşılır. Başlıkların aynı olduğu parçalarda ilk ve son cümlelere analiz edilir. Bu yöntem yansının içinde video var ise uygun olmayabilir. Bu problem için SVM sınıflandırıcı kullanılmıştır. Burada SVM yansılar arasındaki histogram özelliklerine göre sınıflandırma yapar. Çalışma deneysel olarak test edilmiştir. Çalışmada farklı konuşmacıların yer aldığı yirmi ayrı video kullanılmıştır. Çalışmada SVM sınıflandırıcının eğitimi için 2597 yansı görüntü parçası ve 5224 yansı olmayan görüntü parçası kullanılmıştır. Görüntü parçalarındaki metinsel verilerin sayısallaştırılarak işlenebilir hale getirilmesi için açık kaynak kodlu tesseract programı kullanılmıştır. Tesseract progmaı harfleri %92 oranında doğru tanımıştır ve kelimeleri %85 oranında doğru tanımıştır.

Optic Character Recognition (OCR) teknolojisi uygulaması sonrasında ders ses parçalarına Automatic Speech Recognition (ASR) teknolojisi uygulanmıştır. Automatic Speech Recognition teknolojisi sayesinde ses dosyaları da işlenebilir metinsel üst verilere dönüştürülmüştür. OCR ve ASR transkriptinin yanı sıra algılanan slayt metni satır türleri, içeriğe dayalı video tarama ve arama için hem video hem de segment düzeyinde anahtar kelimelerin çıkarıldığı anahtar kelime çıkarma için bu yaklaşım benimsenmiştir. OCR ve ASR tanıma teknolojilerinden çıkan verilerdeki yanlış metinlerin sınıflandırıcıda açabileceği sorunlardan dolayı anahtar kelimeler çıkarılarak sınıflandırma bu kelimeler üzerinden tamamlanmıştır. Anahtar kelimelerin çıkarımı yaklaşımında en fazla geçen kelimeler esas alınmıştır.

N. Radha [11] ‘nın çalışmasında son kullanıcının arama yöntemini basitleştirmek ve ilgili videoyu almak için büyük çevrimiçi ders video arşivlerinde otomatik video indeksleme ve video arama için bir yaklaşım önermektedir. N. Radha [11] ‘nın çalışmasında otomatik video indeksleme ve ders video veritabanları için video alımının analiz süreci sunulmaktadır. Video geri alma sistemi videoların bilgilerini görsel ve ses parçaları olmak üzere iki ana veri olarak çıkarır. Kullanılan videolar ders videolarıdır ve iki ana bölümünden üst verileri çıkarır. Görsel ekrandan, öncelikle slayt geçişleri tespit edilir ve her bir benzersiz slayt çerçevesi video parçası olarak kabul edilir. Slayt çerçevelerinden gelen metinsel üst verilerin bulunduğu parçalar çıkarılır ve daha sonra video OCR tekniği kullanılarak tanınır. OCR sonuçlarına dayanarak, video bilgilerindeki ilgili metin kaydedilir. İkinci olarak konuşma metni analizi videonun ses parçalarından elde edilir. Sfenks konuşma tanıma modelleri, konuşmadan metne dönüştürme işleminin tanınması için kullanılır. Önerilen çalışmada yansı çerçevesinden çıkarılmış metin bilgileri ve videodan çıkarılmış ses parçaları videoyu ders video veritabanından almak için kullanılmıştır. Bu birleşik video alma sisteminin performansı, video sistemlerinde sırasıyla ses ve metin kullanılarak oluşturulan ayrı bir sistemle karşılaştırıldığında performansta önemli bir iyileşme göstermektedir. Çalışma için 60 video toplanmıştır ve deneysel olarak kullanılmıştır. Bu deneysel çalışmada otomatik olarak çıkarılan görsel yansı metninin, otomatik olarak çıkarılan sözlü metne kıyasla daha yüksek hassasiyette video alımı sağladığı açıkça belirtmektedir. Otomatik olarak çıkarılan görsel yansı metni verilerini ve ASR ile çıkarılan işitsel metni birleştirmek performansta artış sağlamıştır. İçeriğe dayalı video erişim sistemlerinin iyileştirilmesine yönelik araştırmalar önemli kayda değer ilerlemeler kaydetmiştir. Video erişim sisteminde yapılan çeşitli geliştirmeler anahtar çerçeve seçimi, özellik çıkarma, sınıflandırma, dizin oluşturma, sorgu taraması sonuçları ve kullanıcı arayüzü geliştirmeleridir. Deneysel değerlendirme çeşitli ders videoları için test edilmiştir. Bu sistemin performansı yine görsel ve işitsel verilerin birlikte kullanımı ile daha verimli sonuçlar vermiştir. Birleşik sesli ve görsel metin bilgileri yüksek tanınma doğruluğu sağlamıştır.

Haubold and Kender [14] ‘ın çalışmasında videoları ses ve görsel üst verilere göre parçalara ayırma, görselleştirme ve dizinleme yöntemleri üzerinde çalışılmıştır. Ses verisi konuşmacı tarafından parçalara ayrıldıktan sonra ASR yöntemi ile üst veriye çevrilmektedir. Video parçası, görsel farklılıklar ve konuşmacı hareketlerindeki değişikliklere göre bölümlere ayrılır ve anahtar kareler çıkarılır. Etkileşimli kullanıcı arayüzü ses, video, metin ve anahtar

karelerin görsel sunumunu birleştirir. Bu durum kullanıcının sunum videolarında arama yapmasına olanak tanır. Çalışma 7,5 saatlik öğrenci sınıf sunum videosu ile yapılmıştır. Toplam 32 sunum adet sunum kullanılmıştır. Bu sunumlara 176 farklı öğrenci yer almaktadır. Bu 176 öğrenci farklı seviyede bilgi birikimine sahiptir. Karakteristik olarak sınıf sunumları birkaç öğrenci tarafından gerçekleştirilir ve bilinen bir yapıyı takip eder. Sınıf sunumları çeşitli kritik açılardan ders videolarından farklıdır. Bu videolar geleneksel ders videolarından daha uzundur ve sınıf sunumları birden fazla öğrenci tarafından gerçekleştirilir. Ses kalitesi önemli ölçüde değişir ve sunumlar arasında anahtar kelimelerin tekrarlayan bir yapısı vardır.

Haubold and Kender [14] 'ın çalışmasında tartışılan yöntemler ve araçlar sunumlarda yer alan iki kitlenin ihtiyaçlarını ele almaktadır. Bu iki kitle eğitmenler ve öğrencilerdir. Sınıf sunumları eğitmenler tarafından ekiplerin ya da bireysel öğrencilerin performansını değerlendirmek için kullanılır. Sınıf sunumlarından öğrenciler için olan faydası da kendilerini izleyerek sunum yeteneklerini geliştirme şeklindedir.

Bu tür sınıf sunumu videoları için video analizi gereklidir. Haubold and Kender [14] 'ın çalışmasında görsel ve işitsel verinin parçalara ayrıştırılması ve dijitalleştirilmesi yöntemi kullanılmıştır. Genelde videodaki işitsel verinin ayrıştırılması konuşmacı tabanlıdır. Yani konuşmacı değiştiği zaman videodaki işitsel verinin parçalanması gerçekleştirilir. Bu çalışmada konuşmacı farklılıklarını anlamak için Bayes Bilgi Kriteri (Bayesian Information Criterion) istatistiki yöntemi kullanılmıştır. Ses parçası düzenli aralıklarla örneklenmiştir ve her bir ses örneği seti için frekans vektörleri örneklenmiştir. İki pencereli bir yaklaşım kullanılarak, BIC bu aralığın her bir bölümü için hesaplanır. BIC değerleri arasında net bir pozitif maksimum varsa konuşmacı değişikliği var demektir, aksi takdirde aralık ek ses örnekleri ile uzatılır. Bu yöntemle işitsel verilerin parçalanmasını çok uygundur. Hiç yanlış negatif ile karşılaşılmamıştır ve az sayıda yanlış pozitif ile karşılaşılmıştır. Yani sistemin tespit edemediği hiçbir konuşmacı değişikliği olmamıştır.

Videoda işitsel olduğu gibi görsel olarak da parçalara ayrılmıştır. Burada görsel değişiklikler sunumlardaki yansı geçişleri ya da konuşmacının ani hareket geçişleridir. Bu çalışmada konuşmacıların pozisyonları ve vücut hareketleri de dikkate alınmıştır. Ardışık kareler arasındaki histogram değişikliklerini hesaplama ve zaman içindeki değişim derecesini karşılaştırarak uzun vadeli değişiklikleri tespit etme yöntemlerini uygulanmıştır. İki ile dört saniyelik video kareleri arasında karşılaştırmalar yapılmıştır ve video karelerinden alınan

kesitler arasındaki sapma önemli ölçüde kare geçişlerini tespit etmiştir. İki ile dört saniyelik pencereler arasındaki sapmayı ölçmek için deneysel olarak türetilmiş bir eşik değeri kullanılmıştır. Eşik değerinin üzerinde olan sapmalarda sunum yansılarında ya da konuşmacının vücut hareketlerinde değişiklikler tespit edilmiştir. Bu yöntemin sunum slaytlarındaki değişiklikleri tespit etmede sağlam bir yöntem olduğunu bulunmuştur. Daha ilginç bir şekilde bu yöntem aynı zamanda iki konuşmacı arasındaki karakteristik hareket modellerini ayırarak konuşmacı değişikliklerini de algılamıştır. Bu çalışmada hareketlerin anlamları ile değil, öğrenciler arasındaki fark ve olağandışı hareketin ortaya çıkmasıyla ilgilenilmeyen farklı bir yaklaşım kullanılmıştır. Sezgisel olarak bu tür sistemler sunum sırasında gerçekleşen ilginç anları tanımlama eğilimindedir. Bu nedenle bu ölçüm görsel bir grafik olarak kullanıcı arayüzüne dahil edilmiştir. Ayrıca ham aktivite grafiğinden yüksek derecede görsel değişiklik ile görsel parça geçişlerini anlamak kolaylaşmıştır.

Görsel ve işitsel veri parçalara ayrılarak ayrıştırıldıktan sonra değişimlerin yalnızca görsel veya işitsel sahne değişiklikleri ile değil, ikisinin bir kombinasyonu ile tanımlandığı bir yeni değerin tanımına yol açmıştır. Bazı durumlarda iki yöntemden yalnızca birinin kullanılması olumsuz sonuçlara yol açmıştır. Konuşmacı ayrıştırması tek başına bir öğrencinin gerçekleştirdiği uzun bir sunum için hiç video kesiti üretmez veya çok az ayrı kesit üretir. Sadece görsel verinin ayrıştırılması özellikle iki konuşmacının aynı slaydı kullandığında diyalog değişikliklerini kaçırabilir. Görsel ve işitsel verinin birlikte kullanılması konuşmacı, vücut hareketleri ve görsel yardım değişiklikleri de dahil olmak üzere önemli ses ve görsel değişiklikleri yakalayabilir. Her iki veri türünü birlikte kullanmak ayrı ayrı kullanmaktan daha doğru sonuçlar vermektedir. Toplamı parçalarından daha anlamlıdır. Bu da ayrı bölümlemelerin mantıksal olarak birleştirilmesi gerektiğini gösterir.

Haubold and Kender [14] tarafından yapılan çalışmada öğrencilere verilen görevler arasında öğrenciler tarafından bilinen ve bilinmeyen çeşitli sunum bölümlerinin aranması ve ek açıklamalar verilen sunumların özetlenmesi yer alıyordu. Görsel içerik ve indeksleme yöntemlerimiz ön arayüzde videoların içeriği hakkında çeşitli bilgilere sahip 176 öğrenci ile değerlendirildi. Sonuçlara göre erişim benzer sistemlerle karşılaştırıldığında %20 daha hızlı yanıtlar vermiştir.

Medida and Ramani [15] tarafından yapılan çalışma Makine Öğrenmesi (ML) metin sınıflandırma algoritmasına dayalı olarak ders videolarının verimli aranması ve alınması için bir

metodoloji sunmaktadır. Metin içeriği sadece video derslerinden çıkarılan ses içeriğinden üretilir. Bu içerik makine öğrenmesi metin sınıflandırma modelinin eğitimi için kullanılan özet ve anahtar kelime çıkarımı için kullanılır.

Metin içeriğe dayalı verileri almak için sıklıkla kullanılan üstün bir anlamsal özelliktir. Ders videolarında yer alan metinler ders için bir taslak görevi görür ve içeriği anlamak için çok önemlidir. Ders videolarının metin bilgisi videolardan OCR ve ASR yöntemleri kullanılarak oluşturulabilir. Medida and Ramani [15] tarafından yapılan çalışmada konuşma içeriği üzerinden eğitilmiş metin sınıflandırma modeline dayalı olarak ders videolarını alan bir sistem önerilmiştir. Medida and Ramani [15] tarafından yapılan çalışma ASR' ye dayalı olarak işitsel verileri metin haline getirmektedir. ASR' ye dayalı olarak çıkarılan üst veriler metinlerin sınıflandırmasında kullanılmıştır. Belgeler sabit sayıda önceden tanımlanmış kategorilerde sınıflandırılmıştır. Bir tekniğin performansı sadece kullanılan algoritmaya değil, aynı zamanda kullanılan verinin özelliklerine de bağlıdır. Bu nedenle belirli bir verinin ve belirli bir görevi için doğru tekniği tanımlamak önemlidir. Bu nedenle metni sınıflandırırken farklı tekniklerin nasıl performans gösterdiğinin değerlendirilmesi çok değerlidir. Video alımı sırasında arama tüm veritabanından geçmek yerine sınıflandırıcı modeli tarafından yapılan tahminler üzerine kurulmuştur. Bu modeli özet ve anahtar kelimeler konusunda eğitmek için Naive Bayes, Destek Vektör Makinesi ve Lojistik Regresyon algoritmalarını uygulanmıştır. Bu tekniklerin her biri tarafından elde edilen sınıflandırma etkinliği karşılaştırılmıştır.

Değişken uzunlukta ve farklı iç yapıdaki ders videoları farklı internet kaynaklarından toplanmıştır. Toplanan video dosyalarının formatları MP4 biçimindedir. Ses dosyası verilen ders video dosyasından çıkarılır ve veritabanında saklanır. Bu nedenle oluşturulan veritabanı, arama işlemi sırasında içeriğin etkili bir şekilde alınmasını sağlayacak şekilde eklenen ders video içeriği ile iyi organize edilmiştir. Video ders kaydından ses parçasının çıkarılması, Python yazılımı ile Hızlı İleri Hareketli Resimler Uzman Grubu (FFmpeg) kullanılarak gerçekleştirilir. Ses parçası WAV formatında çıkarılır. Çıkarılan WAV dosyası yeniden örneklemeye tabi tutulmuştur.

Bu çalışmada çıkarılan ASR çıktılarına otomatik özetleme tekniği kullanılmıştır. Otomatik özetleme tekniği metinden bir veya daha fazla önemli cümleyi çıkararak verilen metni otomatik olarak özetler. Otomatik anahtar kelime çıkarma tekniğini tanımlar. Bu nedenle otomatik özet ve anahtar kelime çıkarma teknikleri orijinal metin içeriğinin temel niteliklerini

korurken belgenin karmaşıklığını ve uzunluğunu azaltacaktır. Metin sınıflandırma modeli doğrudan metin transkript verileri üzerinde eğitilmek yerine model özet ve çıkarılan anahtar kelimeler üzerine eğitilmiştir. Bu adım modelin verimliliğini etkilemezken metin sınıflandırma modelini eğitmek için gereken süreyi azalttığı için önemlidir. Python'un Gensim kütüphanesi metin döküm belgelerinden özet ve anahtar kelimeler oluşturmak için kullanılmıştır. Özet ve anahtar kelime çıkarımı için Gensim uygulaması "TextRank" algoritmasına dayanmaktadır. Çıktı özeti ve anahtar kelimeler, metin sınıflandırma modelinin eğitimi için kullanılan veri kümesine öznitelik olarak eklenir. Çok sınıflı sınıflandırma gerçekleştirmek için her örnek için önceden tanımlanmış bir kategori eklenir. Veri kümesi video dosyalarını atanan kategoriye göre sınıflandıran modeli eğitmek için kullanılır.

Bu çalışmada denetimli bir öğrenme modeli kullanılmıştır. Eğitimli ML modeline göre optimize edilmiş bir arama elde edilir. Sistemin performansı sistem eğitimi için Naive Bayes, Support Vector Machine ve Logistic Regression algoritmaları kullanılarak karşılaştırılır. Performans değerlendirmesi, sınıflandırıcıların her biri için araştırmanın kesinliği, hatırlanması, F-skoru ve doğruluğu ile yapılmıştır. Naive Bayes sınıflandırma algoritması üzerine eğitilen sistemin hem zaman açısından hem de arama sonuçlarının alaka düzeyi açısından daha iyi performans sağladığı görülmektedir.

Masneri and Schreer [7] tarafından yapılan çalışmada SVM tabanlı bir yaklaşım sunulmuştur. Bu çalışmada amaç SVM kullanarak veri kümesinde olan üniversite derslerini ve konferansları sınıflandırabilmektir. Semantik kavramlara dayalı olarak videonun zamansal bir ayrıştırılmasını gerçekleştirilebilmektedir. Videolar içerikleri dört farklı kategoride sınıflandırılabilmektedir. Bunlar tahta, konuşma, slayt ya da karışık olarak dört sınıfa ayrılırlar. Sistem ayrıca slayt geçişlerini, animasyonları ve sunum içindeki videolar gibi dinamik içerikleri algılamak için sunum parçalarını analiz eder.

Geliştirilen yaklaşımda SVM sınıflandırıcıyı eğitmek için birkaç yüz saatlik videonun iki farklı veri kümesinden çeşitli renk ve yüz özellikleri kullanılmıştır. Bu veri kümeleri TED konuşmalarından ve VideoLectures internet sitesinden alınmış videolardan oluşmaktadır. TED konuşmaları ile VideoLectures ders videoları arasında belirgin farklılıklar bulunmaktadır. TED konuşmaları oldukça kısa videolardır ortalama olarak 16 dakika civarındayken VideoLectures ders videoları ortalama olarak bir saatten daha uzundur. TED konuşmalarında çok sayıda açı geçişi ya da sahne geçişi bulunmasına karşın ders videolarında böyle bir geçiş bulunmamaktadır.

TED videoları daha yüksek çözünürlüklere sahip olmasına karşın VideoLectures ders videoları düşük kaliteye sahiptir.

Sistem sınıflandırma için yüz tanıma ve renk tanıma tabanlı iki özellik çıkarımı kullanmıştır. Yüz tanıma kısmen diğerine göre kolay olmasına karşın renk özelliklerini çıkararak sınıflandırma daha zordur. Renk siyahsa yazı yazılan kara tahtaya mı ait ya da konferans sırasında pembe rengin tonları gibi özellikler çıkarılmıştır. Renk tabanlı 48 özellik çıkarılmışken yün tanıma tabanlı 3 özellik çıkarılmıştır. Bu sayede örnekleme yapılan her karede 51 boyutlu bir özellik vektörü elde edilmiştir. Sistem SVM kullanarak sınıflandırmayı kare kare gerçekleştirir ve önceden hesaplanmış sahne geçişi bilgisi gerektirmez. Bu çalışmada Masneri and Schreer [7] 'in önerdiği en büyük farklılık bu noktadadır. Sahne geçişi bilgisine gerek olmaksızın yaklaşımlarını kare bazında geliştirmişlerdir. Sonuçlar her 50 karede bir tek bir sınıfta birleştirilmiştir. Sunulan sonuçlar algoritmanın sağlamlığını ve doğruluğunu kanıtlamaktadır. Yaklaşımın SVM gibi konu özelinde olmayan beynelmilel bir yaklaşım olduğu göz önüne alındığında, sistem farklı ders veri kümelerine kolayca uyarlanabilir.

Adcock and Cooper [8] bir web tabanlı arama motoru tasarımı ve uygulaması geliştirmişlerdir. Videoda bulunan yansılar algoritma tarafından tespit edilir ve OCR teknolojisi ile içeriği tespit edilir. Slayt analizinin temel amacı kullanıcıların videolar arasında gezinmesinde kullanılması ve video dizine görsel bilgi eklemek için yararlı anahtar kareleri ayıklamaktır. Bunun için de ders videolarındaki slaytların ve slayt değişikliklerinin tespit edilmesi gerekir. Anahtar kareler kullanıcılara görsel bir özet ve tam video içeriğine giriş noktaları sağlamak üzere internet arayüzünde bulunmaktadır. Slaytlar kullanıcılara herhangi bir ses dinleme veya videoyu izleme gerektirmeden basit görsel inceleme ile ders içeriği için anında bağlam sağlar. İkinci olarak da slaytlar genellikle OCR tarafından çıkarılabilen ve bir konuşmaya veya görüşmeler topluluğuna metin tabanlı arama için bir dizin oluşturmak üzere kullanılabilen metin içerir.

Adcock and Cooper [8] 'un çalışmasında yaklaşık 200 ders videosu ve 100 ders videosu olmayan videodan oluşan bir test seti kullanılmıştır. Her bir videonun ilk 5 dakikasından alınan karelerle bir SVM sınıflandırıcı ders ve ders dışı materyal arasında ayrım yapmak üzere eğitilmiştir. Burada ikili bir sınıflandırma algoritması kullanılmıştır. Kullanılan özellikler şunlardır: sabit içeriğin toplam süresi, sabit video parçalarının sayısı, sabit video parçalarının ortalama uzunluğu ve dikey kenarların izdüşümünün minimum ve ortalama entropisi. Karelerin

sabit olup olmaması ile ilgili %1’ lik piksel değişimi eşik değer olarak alınmıştır. Sınıflandırma %95’ lik başarı ile sınıflandırma işlemini gerçekleştirmiştir.

Adnan Yazıcı et all. [6] akıllı bir akıllı bir çoklu ortam sistemi tasarlamıştır. Bu çoklu ortam sistemi en basit tanımlaması ile iki kısımdan oluşmaktadır. Bu iki kısımda makine öğrenmesi ve veri tabanı teknolojilerini barındırmaktadır. Bu çalışmadaki ana fikir anlamsal içerik çıkarımı sistemi ile depolama ve alma sistemi birlikte çalıştırmaktır. Adnan Yazıcı ve ark. [6] ‘nın tasarladığı sistemde görsel, işitsel ve metinsel tüm içerik çıkarılmaktadır.

Nesne özelliklerinin çıkarımı iki aşamadan oluşmaktadır. Bunlar aday nesne tespiti ve nesne sınıflandırma kısımlarıdır. Aday nesne tespiti için önce anahtar kareler belirlenir ve sonrasında beblirlenmiş anahtar karelerin parçaları çıkarılır. Bu çalışmada aday nesne tespiti sınıflandırması için SVM kullanılmıştır. Nesne kategorileri sınıfa özgü özeliklere göre tanımlanmıştır. Nesne kategorilerine göre sınıfa özgü özellikleri değerleri farklılık gösterir. Mesela bir araba nesnesi için renk önemli bir özellik değildir. Her aday nesne için sınıflandırıcıdan her bir nesne için 0 ile 1 arasında bir değer döner. Sonuç olarak videodaki her bir nesnenin ekranda ne kadar süre kaldığı ve pozisyonu bilgisi çıkarılır. Elde edilen anlamsal içerikler ve üyelik değerleri puan olarak sonraki işlemler için füzyon modülüne sunulur.

Nesne özelliklerinin çıkarımı yapıldığı gibi bunun yanında videodan işitsel verilerin de çıkarımı yapılır. Konuşma, müzik ve ortam sesi gibi farklı sesler olabileceği için ses tanıma işi zordur. Önce sessizlik analizi yapılır. 10 ms uzunluğundaki ses parçalarını analiz edilerek ortam sessiz mi bakılır. Sesin 30 ms’ lik bölümleri 10’ ar ms uzunluğunda analiz edilir. Ses sınıflandırmasında SVM kullanılır. Sınıflandırma sonrası bir düzleştirme işlemi yapılarak veri füzyon modülüne iletilir.

Üçüncü olarak da videodan metinsel verinin çıkarımı işlemi geçekleştirilir. Burada metin vurgulayıcı görüntüden metni çıkarma ve sonrasında adlandırılmış varlık tanıma işlemini yapar. Kişi, konum ve kuruluşların isimlerini tanır. Bunları anlamlı bir metin halinde sunar.

Metin çıkarımı ve sonrasında yazım denetimi gerçekleştirilir. Adlandırılmış varlık tanıma işleminde metinsel kaynaklar ve örüntü tabanlı işlemler kullanılır. Metinsel kaynaklar iyi bilinen yerlerin, kişilerin ve organizasyonların isimlerini tutar. Örüntü tabanı yer ve organizasyon listesi için kural listeleri içerir. Adlandırılmış varlık tanıma metinsel kaynaklar ile eşleşen kelimeleri görüntü tabanı kurallarına göre cümle haline getirir.

Füzyon modülünün amacı farklı yöntemlerle elde edilen verileri bütünleştirmektir. Füzyon sayesinde herhangi bir parçaya olan bağımlılık azalır. Böylece genel sistem bireysel yöntemlerin hatalarından daha az etkilenir. Füzyon için genel bir yöntem olan SVM seçilmiştir. SVM ortam bağımsız bir yöntemdir. Özelleşmiş yöntemler popüler olmasına karşın veri seti bağımlıdır. Anlamsal kavramlar arasındaki bağlantıyı öğrenme görevi sınıflandırıcıya verilmiştir. Anlamsal içeriği tahmin etmek için sınıflandırıcı her bir anlamsal kavram için eğitilir. Anlamsal içerik ve alt düzey içerik bilgileri nesne yönelimli veritabanında tutulur.
Çoklu ortam verileri için özellikle tasarlanmıştır. Videolar arası ve kavramlar arası ilişkiyi saklar ve döndürür. Bulanık sorgular için kesin olmayan ölçümleri tutabilir. Bulanık bilgi tabanı ve bulanık etkileme motoru videolar kategorisini yönetir. Anlamsal bulanık mantık kuralları nesnelerin özelliklerinden yeni bilgiler çıkarır. Sistemin avantajı anlamsal içeriğe göre arama yaparken alt seviye tanımlayıcıları da kullanabilmesi. Video klipler birden fazla modalite barındırır ve bu çalışma tüm bu modaliteleri kullanabilmektedir. Füzyon işlemi yapabilmektedir ve yeni videolardan yeni kavramlar çıkarabilmektedir. Sistem performansı birçok açıdan muadilleri ile kıyaslanmıştır ve sistem testleri başarılı tamamlamıştır.

1.2. Motivasyon ve Tezin Katkısı

Covid-19 salgının tüm dünyayı etkilemesi ile bütün sektörler gibi eğitim de aynı ölçüde etkilenmiştir. Hatta görünen o ki en çok etkilenen sektörlerden birisi olmasının yanında eğitim sektöründeki Covid-19 etkileri kalıcı olacaktır. Birçok sektör gibi eğitim sektörü de uzaktan eğitim ve evden çalışma gibi uygulamalara geçmiştir. TRT bünyesinde EBA TV kalıcı kanal olma özelliği kazanmıştır.

2023 itibari ile Türkiye Cumhuriyeti Milli Eğitim Bakanlığı internet üzerinden alınan kredilerin sınıf dersleri yerine de sayılacağını ve bu dersleri vererek öğrencilerin mezun olup diploma alabileceğini ifade etmiştir. Buna benzer etkiler çevrimiçi öğrenme alışkanlığının ülkemizde çok hızlı bir şekilde artacağını göstermektedir.

Bu tezde yapılan çalışma da ders videoları üzerine yapılmıştır. Önceki çalışmalarda çok sayıda video analiz, video sınıflandırma ve video geri alma sistemleri incelenmiştir. Bu sistemlerin birçoğu ders videolarına ve bir kısmı da her türlü videoya göre tasarlanmış sistemlerdir. Çeşitli ders videoları üzerinde yapılan çalışmalarda çeşitli sınıflandırma yöntemleri kullanılmıştır. Videoları nesnel, metinsel ve işitsel olarak irdeleyen sistemler mevcuttur. Ders

videoları üzerine OCR ile yapılan çalışmalarda birçok sınıflandırma yöntemi yerine tek yöntem kullanılmıştır.

Ders videoları üzerine birden fazla sınıflandırma yöntemi kullanılan çalışmalarda ise metinsel veri yerine işitsel veri ön planda kullanılmıştır.

Bu tezde literatüre kazandırılan yenilik, farklı veri setlerinde OCR ile elde edilen veriler üzerinden birden fazla sınıflandırma yöntemi kullanılmasıdır. Daha önceki çalışmalarda bu noktada geleneksel makine öğrenme yöntemleri kullanılırken bu tezde üç adet geleneksel makine öğrenimi yöntemiyle birlikte bir adet de derin öğrenme yöntemi deneysel olarak uygulanmıştır.

2. VERİ KÜMESİ

Bu tez çalışması için yeni bir veri kümesi oluşturuldu. Veri kümesinde toplam 110 adet ders videosu bulunmaktadır. Veri kümesindeki videolar 1920x1080 ya da 1280x720 ekran oranlarına sahiptir yani tüm videolar 16:9 videolardır. Tüm videolar 30 fps ve progresive (tek geçişli) yapıdadır. Veri kümesindeki videolar .mp4 uzantılıdır. Video codecleri AVC 'dir.

Veri kümesi 3 ayrı seviyeden oluşmaktadır. Birinci seviyede 4 sınıf, ikinci seviyede 3 sınıf ve üçüncü seviyede 4 sınıf bulunmaktadır.

Birinci seviyede toplam 110 adet ders videosu bulunmaktadır. Bunlar 4 sınıfa ayrılmaktadır. Bu sınıflar Eğitim, Tıp, Sosyal Bilimler ve Mühendislik sınıflarıdır. Eğitim sınıfında 10 adet, Tıp sınıfında 10 adet, Sosyal Bilimler sınıfında 10 adet ve Mühendislik sınıfında 80 adet video bulunmaktadır. Videoların uzunlukları değişkenlik göstermektedir.

Bu tez çalışmasında kullanılan videoların ikinci seviyesi Mühendislik sınıfına ait videolardan oluşmaktadır. Mühendislik sınıfında toplam 80 adet ders videosu bulunmaktadır. Bu sınıfta Elektronik Mühendisliği, Makine Mühendisliği ve Bilgisayar Mühendisliği sınıfları bulunmaktadır. Elektronik Mühendisliği sınıfında 20 adet, Makine Mühendisliği sınıfında 20 adet ve Bilgisayar Mühendisliği sınıfında 40 adet ders videosu bulunmaktadır.

Bu tez çalışmasında kullanılan videoların ikinci seviyesi Bilgisayar Mühendisliği sınıfına ait videolardan oluşmaktadır. Bilgisayar Mühendisliği sınıfında toplam 40 adet ders videosu bulunmaktadır. Bu sınıfta Yapay Zeka, Veri Tabanı, Algoritma ve Bilgisayar Ağları sınıfı bulunmaktadır. Yapay Zeka sınıfında 10 adet, Veri Tabanı sınıfında 10 adet, Algoritma sınıfında 10 adet ve bilgisayar Ağları sınıfında 10 adet ders videosu bulunmaktadır. Veri kümesi ile ilgili ayrıntılı bilgi Tablo 2.1'de bulunmaktadır.

Farklı seviyelerde farklı sınıflardan oluşan bir veri kümesinin oluşturulmasının temel motivasyonu algoritmaların farklı anlamsal içeriklerde farklılık gösterip göstermeyeceği ve hangi anlamsal içerikler için hangi algoritmaların daha iyi sonuç verebileceğidir.

Tablo 2.1. Veri Setinin Seviyelere ve Sınıflara Göre Dağılımı

Seviye 1	**Seviye 2**	**Seviye 3**	**Video Sayısı**
Eğitim	-	-	10
Tıp	-	-	10
Sosyal Bilimler	-	-	10
Mühendislik	Elektronik Mühendisliği	-	20
	Makine Mühendisliği	-	20
	Bilgisayar Mühendisliği	Yapay Zeka	10
		Veri Tabanı	10
		Algoritma	10
		Bilgisayar Ağları	10

3. SINIFLANDIRMA YÖNTEMLERİ

3.1 Naive Bayes

Naive Bayes geleneksel makine öğrenimi yöntemlerindendir. Naive Bayes oldukça basit istatistiki temelli bir yöntemdir ve etkili bir metin madenciliği için kullanılmaktadır. Naive Bayes sınıflandırıcısı Bayes teoreminin bağımsızlık önermesiyle basitleştirilmiş halidir. Bayes teoremi aşağıdaki denklemle ifade edilir.

$$\mathrm{P(A|B)} = \frac{\mathrm{P}(B|A)P(A)}{\mathrm{P(B)}} \tag{3.1}$$

P(A|B) : B olayı gerçekleştiği durumda A olayının meydana gelme olasılığıdır.
P(B|A) : A olayı gerçekleştiği durumda B olayının meydana gelme olasılığıdır.
P(A) : A olayının olma olasılığıdır.
P(B) : B olayının olma olasılığıdır.

Naive Bayes Sınıflandırması Makine Öğreniminde denetimli öğrenme alt sınıfındadır. Daha açık bir ifadeyle sınıflandırılması gereken metinlerin ya da örnek verilerin hangi sınıflara ait olduğu öğretilmelidir. Bu tezde yapıldığı gibi videolardan çıkarılan görsel içeriklerin sınıflandırılması buna örnek verilebilir. Bu örnekte çıkarılan metinlerin hangi sınıflara ait olduğu sınıflandırıcıya öğretilir. Burada sınıf sayısı iki ya da daha fazla olabilir. Yeni videolardan çıkarılacak yeni metinler eski metinlerin sahip oldukları sınıf özelliklerine göre istatistiki olarak sınıflandırılacaklardır. Örnek 1' de dördüncü metnin hangi sınıfa ait olduğu Naive Bayes sınıflandırma yöntemi ile bulunmuştur.

Örnek 1:

Döküman	Metin	Sınıf
1	Lacivert Lacivert Sarı	f
2	Lacivert Lacivert Beyaz	f
3	Siyah Beyaz	b
4	Lacivert Sarı	?

P (f) = 2 / 3

P (b) = 1 / 3

P (Lacivert | f) = (4 + 1) / (6 + 1) = 5 / 7

P (Sarı | f) = (0 + 1) / (2 + 1)= 1 / 3

P (Lacivert | b) = (0 + 1) / (2 + 1) = 1 / 3

P (Sarı | b) = (0 + 1) / (2 + 1) = 1 / 3

Sınıf seçme: P (f | d_4) = (2 / 3) x (5 / 7) x (5 / 7) = 0,34

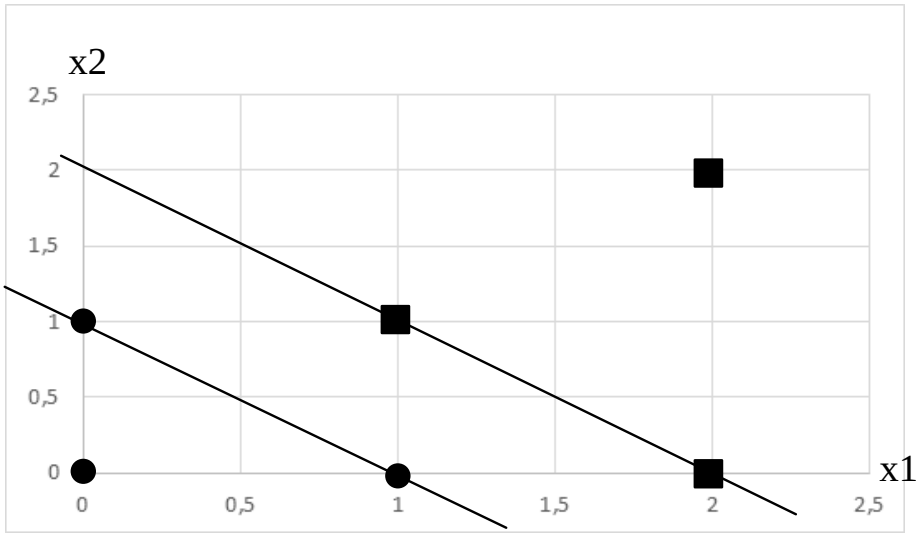
P (b | d_4) = (1 / 3) x (1 / 3) x (1 / 3) = 0,03

P (f | d_4) > P (b | d_4) ➔ Dördüncü doküman f sınıfına aittir.

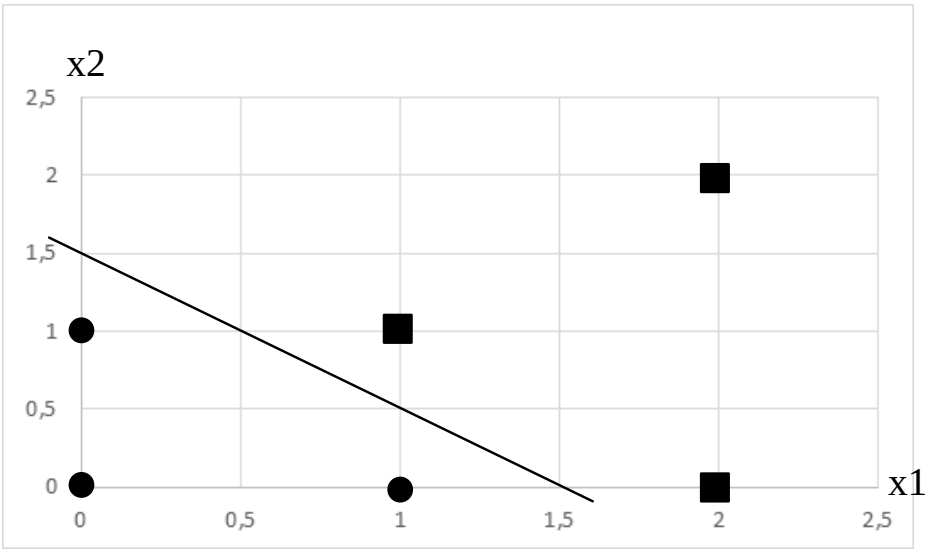
3.2 Support Vector Machine (Destek Vektör Makinesi)

Support Vector Machine geleneksel makine öğrenimi yöntemlerindendir. Sınıflandırma için kullanılan oldukça basit ve etkili yöntemlerden birisidir. Sınıflandırma için bir düzlemde bulunan iki grup arasında bir sınır çizilerek iki grubu ayırma yöntemini uygular.

Eğitim verilerindeki herhangi bir noktadan en uzak olan iki sınıf arasında bir karar sınırı bulur. Bu sınırın çizileceği yer ise iki grubun da üyelerine en uzak olan yer olmalıdır. Support Vector Machine bu sınırın nasıl çizileceğini belirler. Bu işlemin yapılması için iki gruba da yakın ve birbirine paralel iki sınır çizgisi çizilir ve bu sınır çizgileri birbirine yaklaştırılarak ortak sınır çizgisi üretilir. Şekil 3.1'de iki sınıfı ayıran sınır çizgileri yer almaktadır. Şekil 3.2'de her iki sınıfı ayıran karar sınırı gözükmektedir.



Şekil 3.1. Support Vector Machine Sınır Çizgileri



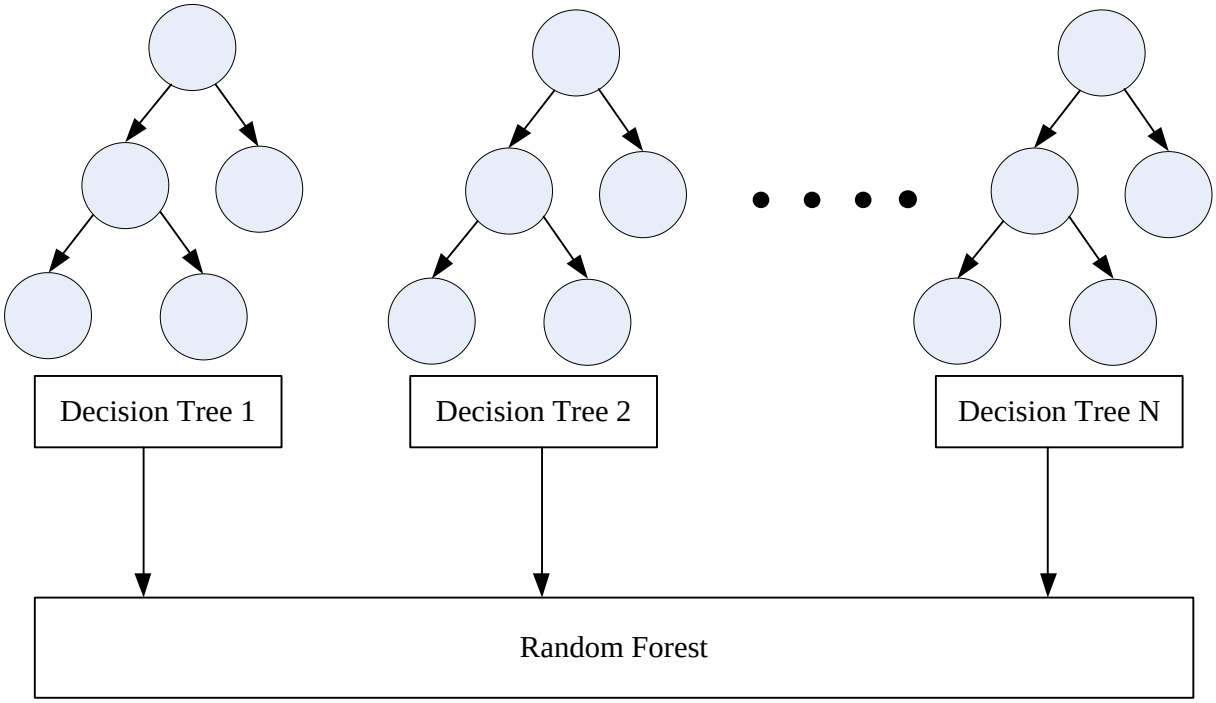
Şekil 3.2. Support Vector Machine Karar Sınırı

3.3 Random Forest (Rastgele Orman)

Random Forest geleneksel makine öğrenimi yöntemlerindendir. Günümüzde Random Forest algoritması, sınıflandırmada çok iyi performans sergilediği için toplu öğrenme yöntemleri içinde sıklıkla tercih edilmektedir [12]. Random Forest ağaç tipi sınıflandırıcılar topluluğudur. Bagging (torbalama) yönteminin daha gelişmiş bir şekli olarak kabul edilir. Hızlı

ve belirli bir kalıpta olmayan bir yöntem olarak nitelendirilir. Random Forest yönteminde ne kadar ihtiyaç varsa ya da ne kadar istenirse o kadar ağaç çalışır.

Random Forest' ta torbalama yöntemi rastgele özellik seçimi ile ilgilidir. Bagging yönteminin seçilmesinin iki önemli nedeni vardır. Birinci özellik bagging yönteminde rastgele özellik kullanılmasıdır. Bu durum doğruluğu arttırır. İkincisi bagging dışında kalan hataların hesaplanmasıdır. Rastgele özellik seçimi için öncelikle gerçek veri kümesinden yeni bir eğitim veri kümesi oluşturulur. Sonrasında rastgele özellik seçimi kullanılır. Yeni eğitim setinden bir ağaç türetilir. Türetilen ağaçlar bu yöntemde budanmaz. Budama metodunun seçiminin ve özellik seçim ölçütlerinin olmamasının ağaç tabanlı sınıflandırıcıların performansını etkilediğini belirtmektedir. Budamanın olmaması Random Forest' ı diğer karar ağacı yöntemlerinden daha avantajlı hale getirmektedir. Şekil 3.3'te Random Forest yöntemi şeması gözükmektedir.

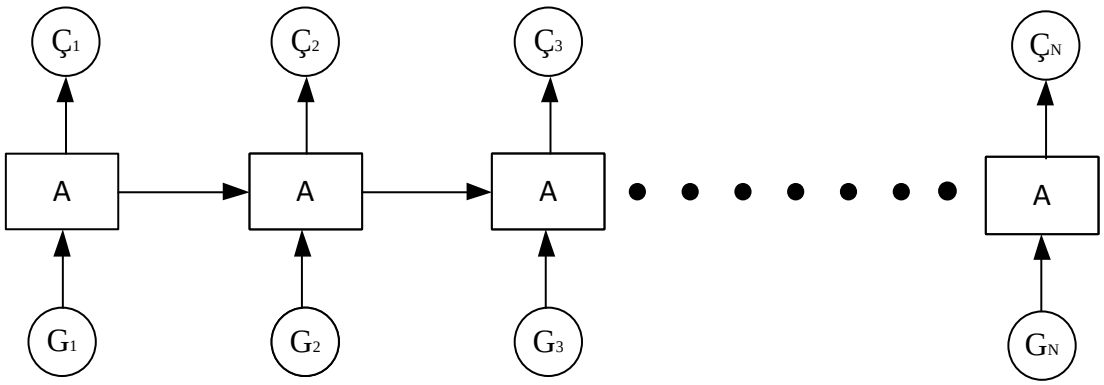


Şekil 3.3. Random Forest Yöntemi Şeması

3.4. Long Short Term Memory (Uzun Kısa Süreli Bellek)

Long Short Term Memory (LSTM) bir derin öğrenme yöntemidir. Tekrarlayan sinir ağları mimarisinin bir alt kümesidir. Standart ileri beslemeli sinir ağlarının aksine, LSTM' nin geri bildirim bağlantıları vardır. LSTM birçok Doğal Dil İşleme uygulamasında kullanılmaktadır. Örneğin nefret söylemlerini algılama, niyet sınıflandırması ve haber makalelerinin düzenlenmesi gibi sürekli olarak büyüyen veri kaynaklarında aktif olarak kullanılmaktadır.

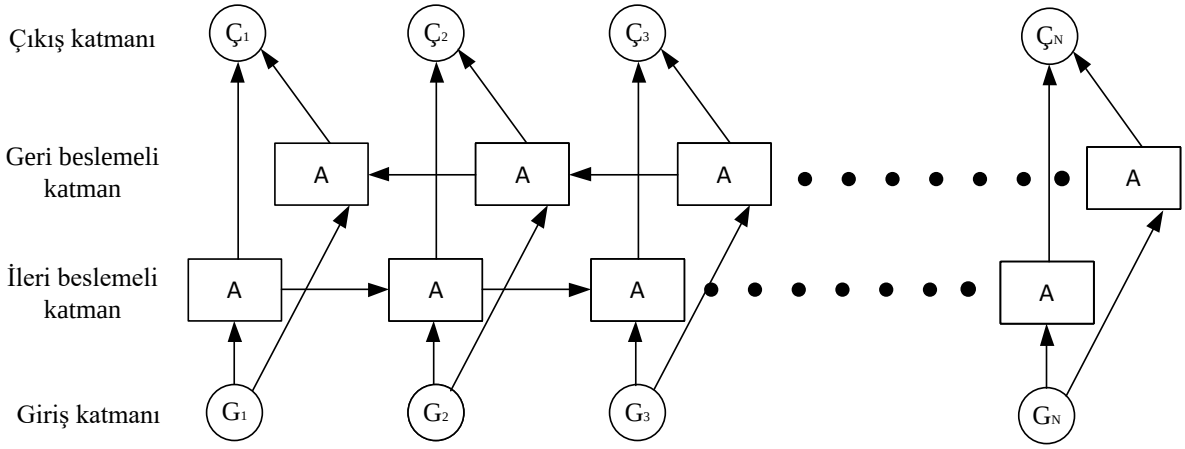
LSTM' nin en büyük farkı sıralı olan gelen girdileri sıralarına göz önüne alarak sınıflandırmasıdır (Şekil 3.4). Şekil 3.4' te "A" ileri beslemeli bir sinir ağını sembolize etmektedir. "G" girdileri, "Ç" ise çıktıları göstermektedir. Örneğin: "Siyahlı adam anahtar getirdi" ve "Siyahlı anahtar adam getirdi". Bu iki cümleyi ele aldığımız zaman sadece kelime frekansına bakmak bize anlamlı bir sınıflandırma sonucu vermeyebilir. İkinci cümle, "Siyahlı anahtar adam getirdi" her ne kadar dilbilim açısından anlamsız bir cümle olsa da kelime frekansları olarak değerlendirildiğinde bir sınıflandırma değeri taşıyacaktır. LSTM gibi kendisinden önceki kelimelere de bağlı bir yöntem bu tip durumlarda sınıflandırma avantajı sağlayacaktır.



Şekil 3.4. Long Short Term Memory Gösterimi

Bu tezde videolardan metinsel verilerin çıkarımı sonrasında bu metinsel veriler üzerinden sınıflandırma yapılmıştır. Klasik LSTM yerine Bidirectional LSTM kullanılmıştır. Bu sayede kelimelerin hem kendinden sonraki kelimelerle hem de kendinden önceki kelimelerle sıralama özellikleri göz önüne alınarak bir sınıflandırma gerçekleştirilmiştir. Çoğu uygulamada

Bidirectional LSTM klasik tek yönlü LSTM' ye oranla daha iyi sonuç vermektedir. Şekil 3.5'te bu tezde kullanılmış olan Bidirectional LSTM' nin gösterimi yer almaktadır. Şekil 3.5' te "A" bir LSTM hücresini, "G" girdileri ve "Ç" çıktıları sembolize etmektedir.



Şekil 3.5. Bidirectional LSTM

4. UYGULAMA VE YÖNTEM

4.1. Genel Mimari

Bu tezde 110 videoluk bir veri kümesi 4 ayrı sınıflandırma yöntemi ile kullanılmıştır. Tüm videoların 30 karede bir yani her 1 saniyede OCR çıktıları alınmıştır. OCR çıktıları sayesinde elde edilen metinsel veriler 3 adet geleneksel makine öğrenmesi yöntemi ve 1 adet de derin öğrenme yöntemi ile sınıflandırılmıştır. Sistemin genel mimarisi ile ayrıntılı bilgi Şekil 4.1 'de bulunmaktadır.

Videolardan OCR çıktıları alınması sayesinde görüntülerden metinsel veriler elde edilmiştir. Metinsel verilerin çıkarılması ile ilgili daha fazla bilgi "4.2. Metinsel Verilerin Videolardan Çıkarımı" bölümünde yer almaktadır.

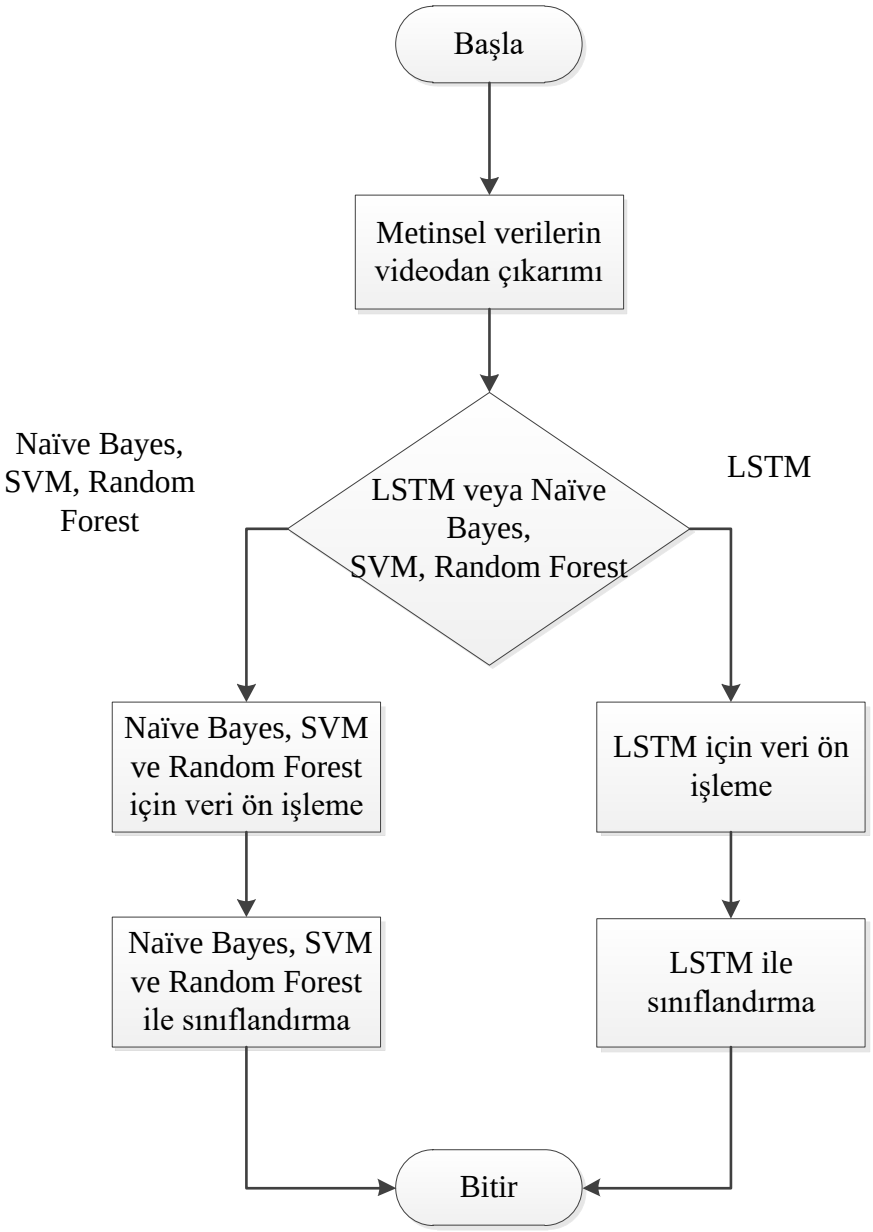
OCR çıktıları alınması sonrasında çıkarılmış olan metinsel verilerin sınıflandırma yöntemlerine uygun hale getirilmesi için sınıflandırma yöntemlerine bağlı olarak sıralı ön işlemler uygulanmıştır. Tüm metinsel veriler toplamda 4 ayrı sınıflandırma yöntemi ile sınıflandırılmıştır. Kullanılan sınıflandırma yöntemlerinden 3 tanesi geleneksel makine öğrenme yöntemlerindendir. Bu tezde kullanılan geleneksek makine öğrenimi yöntemleri Naive Bayes, Support Vector Machine ve Random Forest yöntemleridir.

OCR çıktıları sayesinde elde edilen metinsel veriler 3 adet geleneksel makine öğrenmesi yöntemi haricinde 1 adet de derin öğrenme yöntemi ile sınıflandırılmıştır. Bu tez için seçilen derin öğrenme yöntemi Bidirectional Long Short Term Memory (Bi-LSTM) yöntemidir.

Geleneksel makine yöntemleri için metinsel verileri benzerlik oranına göre karşılaştırma, verileri küçük harfe çevirip gereksiz kelimeleri atma, kelimeleri İngilizce sözlükte kontrol etme, kelime frekansını bulma, kelime frekansını normalleştirme ve veriyi .arff formatına çevirme ön işlemleri uygulanmıştır. Geleneksel makine yöntemleri için yapılan veri önişlemleri ilgili daha fazla bilgi "4.4. Naive Bayes, SVM ve Random Forest için veri önişleme" bölümünde yer almaktadır. Geleneksel makine öğrenim yöntemleri için yapılan önişlemler frekans vektörleri oluşturmaktadır. Bu noktası ile derin öğrenme yöntemi için yapılan önişlemlerden farklıdır.

Geleneksel makine öğrenme yöntemleri ile yapılan sınıflandırma Weka programında gerçekleştirilmiştir. Weka ile yapılmış sınıflandırma ile ilgili daha fazla bilgi "4.5. Naive Bayes, SVM ve Random Forest ile sınıflandırma" bölümünde yer almaktadır.

Derin öğrenme yöntemi için metinsel verileri benzerlik oranına göre karşılaştırma, verileri küçük harfe çevirip gereksiz kelimeleri atma, kelimeleri İngilizce sözlükte kontrol etme ve kelimeleri sıralama vektörlerine çevirme veri önişlemleri uygulanmıştır. Derin öğrenme yöntemleri için yapılan önişlemler ile ilgili daha fazla bilgi "4.6. LSTM için veri önişleme" bölümünde yer almaktadır. Derin öğrenme yöntemi için yapılan önişlemlerde metin sıralama vektörlerine çevrilmektedir. Bu noktası ile geleneksek makine öğrenme yöntemleri için yapılan önişlemlerden farklıdır. Kelimeleri sıralama vektörlerine çevirmek için metni kelime kelime ayırarak her kelimeye sıklığına göre sayısal bir değer verme, takviye etme ve kırpma önişlemleri uygulanmıştır. Bu üç yöntemle ilgili daha fazla bilgi "4.6.1. Metni kelime kelime ayırmak (tokenization) ve her kelimeye sayısal bir değer vermek", "4.6.2. Takviye etme" ve "4.6.3. Kırpma" bölümlerinde yer almaktadır.



Şekil 4.1. Genel Mimari Akış Şeması

4.2. Metinsel Verilerin Videolardan Çıkarımı

Şekil 3.1' de ayrıntılı olarak görüldüğü gibi bu tez kapsamında tüm videolardan metinsel veri çıkarımı gerçekleşmiştir. Metinsel veri çıkarımı için Tesseract Optic Character Recognition yazılımı kullanılmıştır [21]. Tesseract Optic Character Recognition yazılımı açık kaynak kodlu olduğu ve Python yazılımı ile uyumlu olarak kullanılabildiği için seçilmiştir. Bunların yanında Tesseract Optic Character Recognition yazılımının gelişimi şu an Google tarafından karşılanmaktadır. Açık kaynak kodlu olarak erişilebilecek en doğru optic charater recognition yazılımlarından bir tanesidir.

Kullanılan videolar Veri Kümesi kısmında anlatıldığı gibi saniyede 30 karedir. Bu tez çalışmasında tüm Tesseract çıktıları 30 karede bir defa alınmıştır. Metinsel veri çıkarımı saniyede bir defa gerçekleşmiştir. Bir saniyenin altında olan slayt ya da yazı geçişleri ihmal edilmiştir.

Videoların codecleri Veri Kümesi bölümünde anlatıldığı gibi AVC' dir. Video dosyaları ses dosyalarına oranla daha büyük oldukları için videolardan metinsel veri çıkarımı ses dosyalarından metinsel veri çıkarımına oranla daha fazla zaman almaktadır. Uygulama Python yazılımında multithreading (çoklu kullanım) olarak geliştirilmiştir. Bu sayede metinsel veri çıkarımı işlemi merkezi işlemci ünitesindeki çekirdek sayısı ile doğru orantılı olarak artmıştır.

4.3. LSTM veya Naive Bayes, SVM, Random Forest

Videolardan çıkarılan metinsel veriler Naive Bayes, Support Vector Machine ve Random Forest algoritmaları için frekans vektörlerine dönüştürüldüler. Bunun yanında Long Short Term Memory algoritması sıralamalı vektörlere dönüştürüldüler. Verilerin hangi ön işleme yöntemine doğru yönlendirileceği ile ilgili seçim bu aşamada yapıldı.

4.4. Naive Bayes, SVM ve Random Forest için Veri Önişleme

Tezin Naive Bayes, SVM ve Random Forest için veri önişleme aşamasında çok çeşitli önişlemler yapılmıştır. Yapılan işlem en geniş tanımı ile ders videolarından Tesseract yazılımı ile elde edilmiş olan metinsel verileri Naive Bayes,SVM ve Random Forest algoritmalarının uygulanabilmesi için frekans vektörlerine çevirmektir.

Burada ilk olarak metinsel veriler birbirleri ile benzerlik oranlarına göre ayrıştırıldı. Bu ayrıştırma işleminde birden fazla eşik değer uygulandı. Bu tezde uygulanan eşik değerler %50,

%75 ve %90 eşik değerleridir. Art arda gelen iki kareden çıkarılmış metinsel verilerin benzerlik oranı bu eşik değerlerine göre değerlendirildi. Bu eşik değerin üzerinde kalan kareler atıldı. Bu sayede tekrar eden veriler sınıflandırma aşamasına gelmeden elenmiş oldular.

Daha önce de metinsel verilerin videodan çıkarımı anlatılırken bahsedildiği gibi her 1 saniyede 1 defa metinsel veri çıkarılmakta. Ders videolarında her saniyede slayt görüntüleri ve tahta yazıları değişmemekte ya da belli oranlarda değişmektedir. Verilerin uzun süre sabit kalması benzerlik oranına göre veri elenmesi konusunda yapılması gereken ön işlemeyi zaruri kılmaktadır.

Bu noktada Python yazılımının difflib modülü kullanılmıştır. Difflib modülüne ait olan SequenceMatcher kullanılmıştır. SequenceMatcher algoritmik olarak Gestalt Pattern Matching algoritmasını kullanmaktadır. Gestalt Pattern Matching 1980' lerin sonlarında Ratcliff ve Obershelp tarafından temellendirilmiştir [23]. Algoritma karakter dizilerindeki benzer karakterlerin iki katını dizilerdeki toplam karakterlere böler. Bu sayede karakterlerin benzerlik oranına göre 0 1 arasında bir değer döndürür [24].

Örnek:
Kelime 1: Başkent
Kelime 2: Beykent

Aynı olan karakterler: “B”, “k”, “e”, “n”, “t”
İki kelimedeki toplam karakterler: “B”, “a”, “ş”, “k”, “e”, “n”, “t”, “B”, “e”, “y”, “k”, “e”, “n”, “t”

$$\text{Benzerlik Oranı} = \frac{\text{2 x (Aynı olan karakterler)}}{\text{İki kelimedeki toplam karakterler}} \quad (4.1)$$

$$\text{Benzerlik Oranı} = \frac{2 \text{ x } 5}{14} = 0{,}714$$

Bu benzerlik oranına göre karelerin silinmesi sonrasında geriye kalan veride ilk olarak kelimelerin küçük harfe çevirme işlemi gerçekleştirilmiştir. Kelimelerin büyük ya da küçük

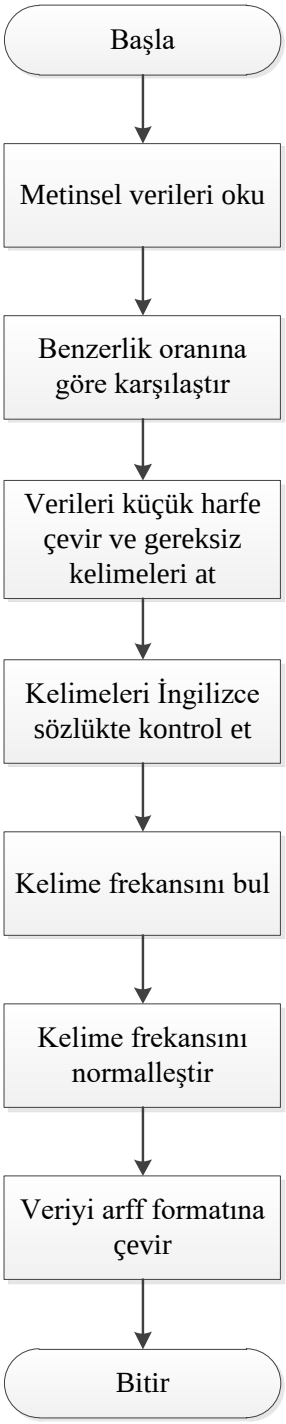
harfte olması anlamsal açıdan pek bir şey ifade etmemesine karşın bilgisayar mimarisi açısından farklı değerlerdir. Farklı ASCII kodlarına bağlıdır ve bu farklılık sınıflandırma kısmında farklı özellikler olarak gözükebilecektir.

İkinci olarak da gereksiz kelimeler çıkarılmıştır. Burada gereksiz kelimelerden kasıt İngilizcede cümle içinde pek anlam barındırmayan. Bu gereksiz kelimeler cümlelerden çıkarıldığında anlamsal olarak bir şey feda edilmeyen kelimelerdir. Bu gereksiz kelimeler çıkarılmıştır.

Üçüncü aşamada kalan kelimelerin İngilizce sözlükte olup olmadığı kontrol edildi. İngilizce sözlükte olmayan kelimeler atıldı. Tesseracttan gelen kelimelerin bazıları anlamsız hatta İnglizce' de de olmayan kelimelerdi. Bu aşmada bu şekilde bir denetim gerçekleştirilerek anlamsız kelimeler atılmıştır.

Bu üç aşamadan sonra kalan kelimelerin tümü okunarak bir kelime havuzu oluşturulmuştur. Kelime havuzunun eleman sayısını belirleyen temel faktör benzerlik oranıdır. Tüm akış uygulandığında benzerlik oranına göre benzer kareleri atma, gereksiz kelimeleri atma ve İngilizce' de olmaya anlamsız kelimeleri atma gibi kelime havuzunu daraltan işlemler uygulanmıştır. Kelime havuzu oluşturulduktan sonra videodaki her bir kelimenin o videoda kaç defa geçtiği ile frekansı elde edilmiştir. Sonrasında da kelime sayısına söre frekans normalleştirilmiştir.

Son olarak da veri Waikato Environment for Knowledge Analysis (Weka) uygulamasında sınıflandırma yapabilmek için arff formatına çevrilmiştir [22]. Ayrıntılı bilgi Şekil 4.2' de bulunmaktadır.



Şekil 4.2. Naive Bayes, SVM ve Random Forest için veri önişleme

4.5. Naive Bayes, SVM ve Random Forest ile Sınıflandırma

Naive Bayes, SVM ve Random Forest sınıfandırma için Weka programı kullanılmıştır. Naive Bayes, SVM ve Random Forest için veri önişleme kısmında çıkarılan arff dosalar Weka için uygun halde tasarlanmıştır. Sadece format olaraktan ziyade özelliklerin dizilimi olarak da Weka makine öğrenimi programında sınıflandırmaya uygun olarak tasarlanmıştır.

Weka ile sınıflandırma 9 ayrı veri kümesi için de yapılmıştır. Bunlar %90, %75 ve %50 benzerlik oranlarına göre ayrıştırılmış birinci seviye, ikinci seviye ve üçüncü seviye veri kümeleridir. Veri kümeleri ayrıntılı olarak Veri Kümesi başlığı altında tarif edilmiştir.

4.6. LSTM için Veri Önişleme

Tezin LSTM için veri önişleme aşamasında OCR ile çıkarılan metinsel verilere benzerlik oranına göre karşılaştırma, verileri küçük harfe çevirip gereksiz kelimeleri atma, kelimeleri İngilizce Sözlük' te kontrol etme ve kelimeleri sıralama vektörlerine çevirme veri önişlemleri uygulanmıştır. Bu aşamada yapılan işlem en geniş tanımı ile daha önce bahsedilen ders videolarından Tesseract yazılımı ile elde edilmiş olan metinsel verileri LSTM algoritmasının uygulanabilmesi için sıralama vektörlerine çevirmektir.

LSTM için veri önişleme işleminde ilk olarak metinsel veriler birbirleri ile benzerlik oranlarına göre ayrıştırıldı. Bu ayrıştırma işleminde birden fazla eşik değer uygulandı. Bu tezde uygulanan eşik değerler %50, %75 ve %90 eşik değerleridir. Eşik değerleri bu tez için rastgele olarak seçilmiştir. Art arda gelen iki kareden çıkarılmış metinsel verilerin benzerlik oranı bu eşik değerlerine göre değerlendirildi. Bu eşik değerin üzerinde kalan kareler atıldı. Bu sayede tekrar eden veriler sınıflandırma aşamasına gelmeden elenmiş oldular.

Daha önce de metinsel verilerin videodan çıkarımı anlatılırken bahsedildiği gibi her 1 saniyede 1 defa metinsel veri çıkarılmaktadır. Ders videolarında her saniyede slayt görüntüleri ve tahta yazıları değişmemekte ya da belli oranlarda değişmektedir. Verilerin uzun süre sabit kalması benzerlik oranına göre veri elenmesi konusunda yapılması gereken ön işlemeyi zaruri kılmaktadır.

Bu noktada Python Naive Bayes, Support Vector Machine ve Random Forest algoritmaları için yapıldığı gibi difflib modülünün SequenceMatcher algoritması kullanılmıştır. Bölüm 4.4' te SequenceMatcher uygulamasının algoritması tarif edilmiştir.

Bu benzerlik oranına göre karelerin silinmesi sonrasında geriye kalan veride ilk olarak kelimelerin küçük harfe çevirme işlemi gerçekleştirilmiştir. Kelimelerin büyük ya da küçük harfte olması anlamsal açıdan pek bir şey ifade etmemesine karşın bilgisayar mimarisi açısından farklı değerlerdir. Farklı ASCII kodlarına bağlıdır ve bu farklılık sınıflandırma kısmında farklı özellikler olarak gözükebilecektir.

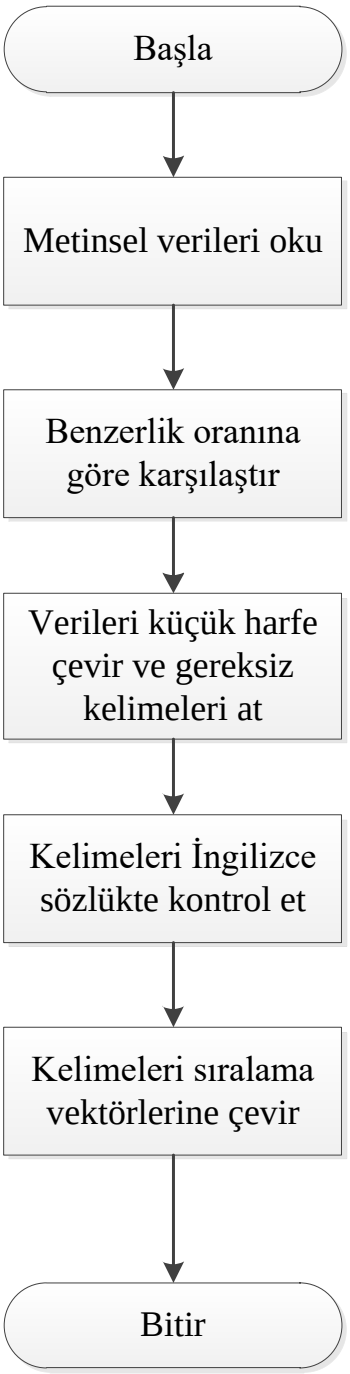
İkinci olarak da gereksiz kelimeler çıkarılmıştır. Burada gereksiz kelimelerden kasıt İngilizcede cümle içinde pek anlam barındırmayan kelimlerdir. Bu gereksiz kelimeler cümlelerden çıkarıldığında anlamsal olarak bir şey feda edilmeyen kelimelerdir. Bu gereksiz kelimeler çıkarılmıştır.

Üçüncü aşamada kalan kelimelerin İngilizce sözlükte olup olmadığı kontrol edildi. İngilizce sözlükte olmayan kelimeler atıldı. Tesseracttan gelen kelimelerin bazıları anlamsız hatta İngilizce' de de olmayan kelimelerdi. Bu aşmada bu şekilde bir denetim gerçekleştirilerek anlamsız ve manasız kelimeler atılmıştır.

Bu 3 aşamadan sonra verinin LSTM için ayrışan ön işleme kısmı gelmektedir. Bu aşamada veriler frekans vektörlerinden farklı olarak sıralama vektörlerine dönüştürülürler. Bu dönüştürme işleminde ilk olarak tüm kelimelerden bir kelime havuzu oluşturulur ve her kelimeye bir değer verilir. Bu değer cümleleri matematiksel olarak ifade etme aşamasında kullanılır.

Test kümesinde daha önceden görülmemiş ve kelime havuzunda bir matematiksel değere sahip olmayan kelimelerin var olma olasılığı vardır. Eğitim kümesinde olmayan kelimelerin test kümesinde olmasına karşı bu kelimelere genel geçer daha önce görülmediğini ifade edecek bir değer verilir. Bu sayede cümlelerin sıralama vektörleri daha doğru sonuçlar elde edebilecek şekilde işleme tutulmuş olur. LSTM gibi sıra vektörleri ile sınıflandırma yapacak bir algoritma için bu hayati bir konudur. Aksi takdirde mesela daha önceden eğitilmemiş kelimelerin göz ardı edildiği yaklaşımlarda farklı anlamlara sahip olabilecek vektörler oluşabilir. Bu durum da sınıflandırmanın başarısını anlamlı ölçüde düşürmektedir.

Bir video metninin kaç kelime ile ifade edip sınıflandırılacağı da bir diğer önemli konudur. Birbirlerinden farkı uzunluğa sahip videoların farklı uzunluğa sahip metinleri olması olasıdır. Bu durum eşit uzunluğa sahip farklı öğretim araçlarının kullanıldığı videolarda da gerçekleşebilir. Bu sebepten burada kullanılacak kelime sayısı veri kümesine bağımlı bir hal almaktadır. LSTM için veri önişleme işlemlerinin genel şeması Şekil 4.3' te bulunmaktadır.



Şekil 4.3. LSTM için veri ön işleme

Kelimelerin sıralama vektörlerine çevrilmesi için üç modelleme tekniği uygulanmıştır. Bunlar metni kelime kelime ayırarak (tokenization) her kelimeye sayısal bir değer verme, takviye etme (padding) ve kırpma (truncating) modelleme teknikleridir.

4.6.1. Metni kelime kelime ayırmak (tokenization) ve her kelimeye sayısal bir değer vermek

Metni kelime kelime ayırma ve her kelimeye sayısal bir değer verme işlemleri için Kerasın Tokenizer Sınıfı veri önişlemede kullanılmıştır [25]. Keras'ın Tokenizer Sınıfı bir metin derlemini vektör haline getirmek için kullanılmıştır. Tokenizer Sınıfı metodları metinleri vektör haline getirme işlemini sözcüklerin metindeki sayılarına ve term frequency–inverse document frequency **(**TF-IDF) istatistiksel ölçme yöntemine göre yapmaktadır [25-26].

Kullanılan sınıfta "0" rakamı takviye etme işlemi ve Out of Vocabulary (OOV) kelimeler için ayrılmıştır. Onun sonrasındaki rakamlarda kelimenin metin içinde kaç defa geçtiğine göre belirlenmektedir. En fazla geçen kelimeler en düşük rakamları en az geçen kelimelerde en yüksek rakamları almaktadır. Örnekte metni kelime kelime ayırma ve her kelimeye sayısal bir değer verme yöntemleri kullanılmıştır.

Örnek:

Aşağıda sınıflandırma için kullanılacak 3 adet cümle vardır.

Cümle 1: başkent üniversitesi bilgisayar mühendisliği

Cümle 2: başkent üniversitesi lisansüstü eğitimi

Cümle 3: başkent üniversitesi bilgisayar mühendisliği ana bilim dalı

Tablo: Her bir kelimenin karşılık geldiği sayısal bir değer.

Kelime	Matematiksel Değer
OOV	0
başkent	1
üniversitesi	2
bilgisayar	3
mühendisliği	4
lisansüstü	5
eğitimi	6
ana	7
bilim	8
dalı	9

Kelimelerin matematiksel ifadesi aşağıdaki şekilde olur.

Cümle 1: [1, 2, 3, 4]

Cümle 2: [1, 2, 5, 6]

Cümle 3: [1, 2, 3, 4, 7, 8, 9]

4.6.2. Takviye etme

LSTM sınıflandırma yöntemi için her bir sıralama vektörü eşit uzunluğa getirilmiştir. Bunun için takviye etme yöntemi uygulanmıştır. “4.6.1. Metni kelime kelime ayırmak (tokenization) ve her kelimeye sayısal bir değer vermek” bölümünde verilen örnekte Cümle 1 ve Cümle 2 eşit uzunluğa sahip iken Cümle 3 daha uzundur. Sonraki sayfadaki örnekte takviye etme yöntemi ile kelimeler eşit uzunluğa getirilmektedir.

Örnek:

En uzun cümleye göre takviye etme yöntemi kullanılmıştır.

Cümle 1: [1, 2, 3, 4, 0, 0, 0]

Cümle 2: [1, 2, 5, 6, 0, 0, 0]

Cümle 3: [1, 2, 3, 4, 7, 8, 9]

4.6.3. Kırpma

Videolardan OCR tanıma ile elde edilen metinsel veriler çok uzundur. Bu sebepten LSTM öncesi bu veriler kırpma yöntemi ile kısaltılmıştır. Örnekte "4.6.2. Takviye etme" bölümünde takviye etme yöntemi ile oluşturulmuş cümleler kırpılmıştır.

Örnek:

En uzun cümle 4 kelimeden oluşacak şekilde kırpma yöntemi uygulanmıştır.

Cümle 1: [1, 2, 3, 4]

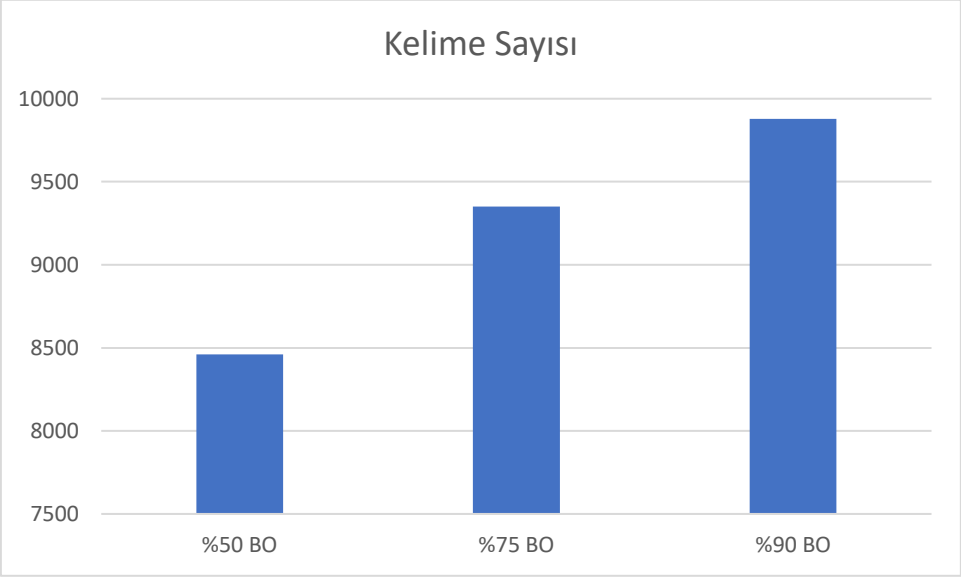
Cümle 2: [1, 2, 5, 6]

Cümle 3: [1, 2, 3, 4]

5. SONUÇLAR

Veri kümesi kısmında anlatıldığı gibi toplam 110 videodan oluşan 3 ayrı seviyede veri kümeleri vardır. Bu veri kümeleri benzerlik oranlarına göre veri önişlemeye alındıklarında toplam dokuz adet ayrı metinsel veri kümesi meydana gelmektedir. Bu dokuz ayrı veri kümesi üç adet geleneksel makine öğrenme yöntemi ile ve bir adet derin öğrenme yöntemi ile sınıflandırılmıştır. Yapılan 36 sınıflandırmanın tümü 10 kat çaprazlama yöntemi ile yapılmıştır.

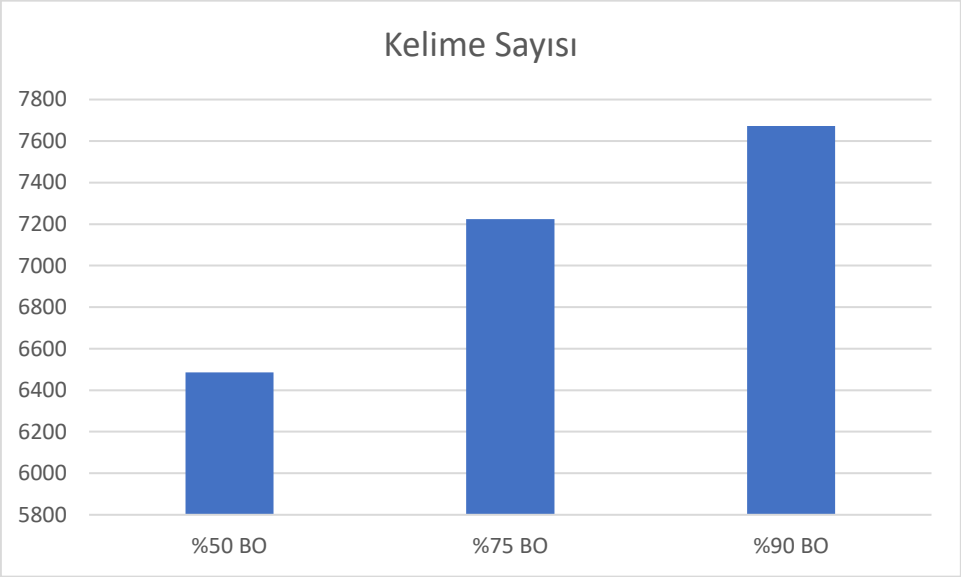
Seviye 1' de Eğitim, Tıp, Sosyal Bilimler ve Mühendislik olmak üzere dört sınıf ders videosu bulunmaktadır. Videolardaki benzersiz kelime sayısı veri önişleme kısmında kullanılan benzerlik oranına göre değişim göstermektedir. Kullanılan benzerlik oranları %50, %75 ve %90 olmak üzere üç adettir. Şekil 5.1' de seviye 1 için elde edilen benzersiz kelime sayıları yer almaktadır.



Şekil 5.1. Seviye 1 için benzersiz kelime sayıları

Seviye 2' de Elektronik Mühendisliği, Makine Mühendisliği ve Bilgisayar Mühendisliği olmak üzere üç sınıf ders videosu bulunmaktadır. Videolardaki benzersiz kelime sayısı veri önişleme kısmında kullanılan benzerlik oranına göre değişim göstermektedir. Kullanılan

benzerlik oranları %50, %75 ve %90 olmak üzere üç adettir. Şekil 5.2’ de seviye 2 için elde edilen benzersiz kelime sayıları yer almaktadır.



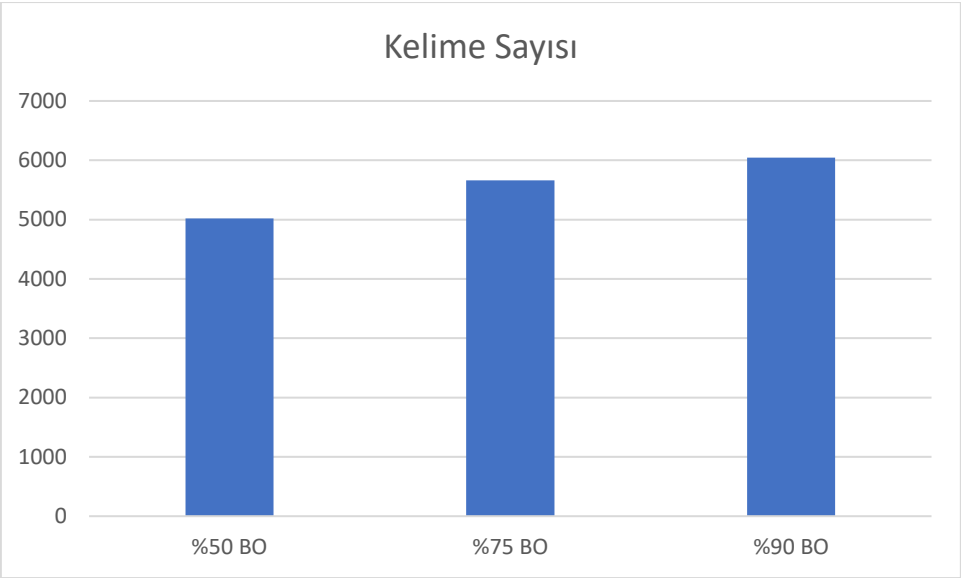
Şekil 5.2. Seviye 2 için benzersiz kelime sayıları

Seviye 3’ de Yapay Zeka, Veri Tabanı, Algoritma ve Bilgisayar Ağları olmak üzere dört sınıf ders videosu bulunmaktadır. Videolardaki benzersiz kelime sayısı veri önişleme kısmında kullanılan benzerlik oranına göre değişim göstermektedir. Kullanılan benzerlik oranları %50, %75 ve %90 olmak üzere üç adettir. Şekil 5.3’ te seviye 3 için elde edilen benzersiz kelime sayıları yer almaktadır.

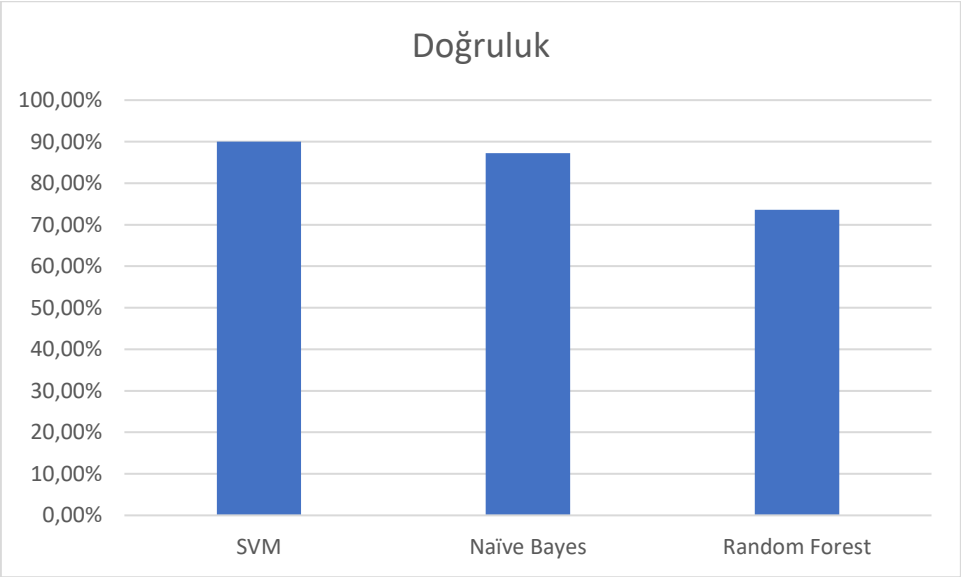
Grafiklerden de görüleceği gibi veri kümesindeki video sayısının artması ve benzerlik oranı sınıflandırma için kullanılacak kelime sayısını en çok etkileyen faktörlerdir.

5.1. Geleneksel Makine Öğrenimi Yöntemleri ile Elde Edilen Sonuçlar

110 elemanlı ve 4 sınıflı Eğitim, Tıp, Sosyal Bilimler ve Mühendislik veri setinde %50 benzerlik oranı ile yapılan sınıflandırmada en yüksek doğruluk oranı SVM algoritması ile elde edilmiştir (Şekil 5.4).

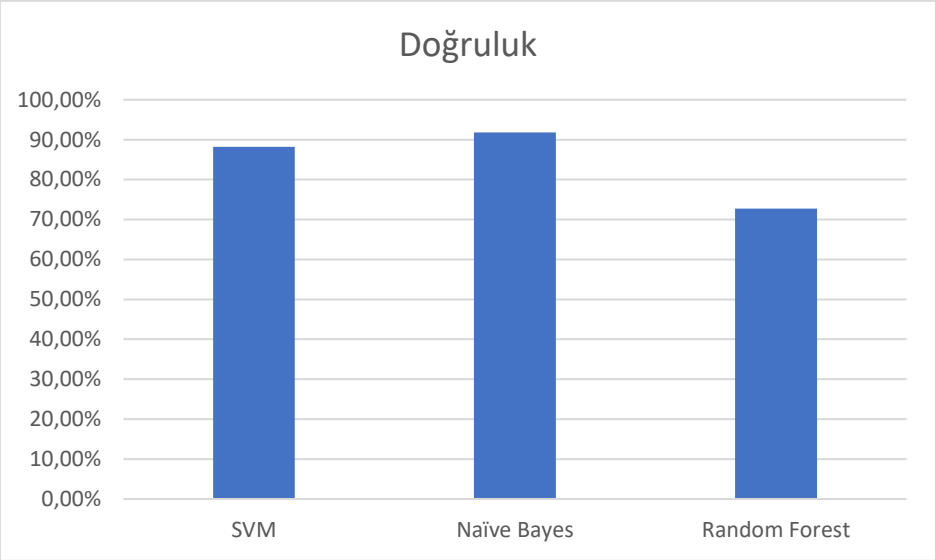


Şekil 5.3. Seviye 3 için benzersiz kelime sayıları



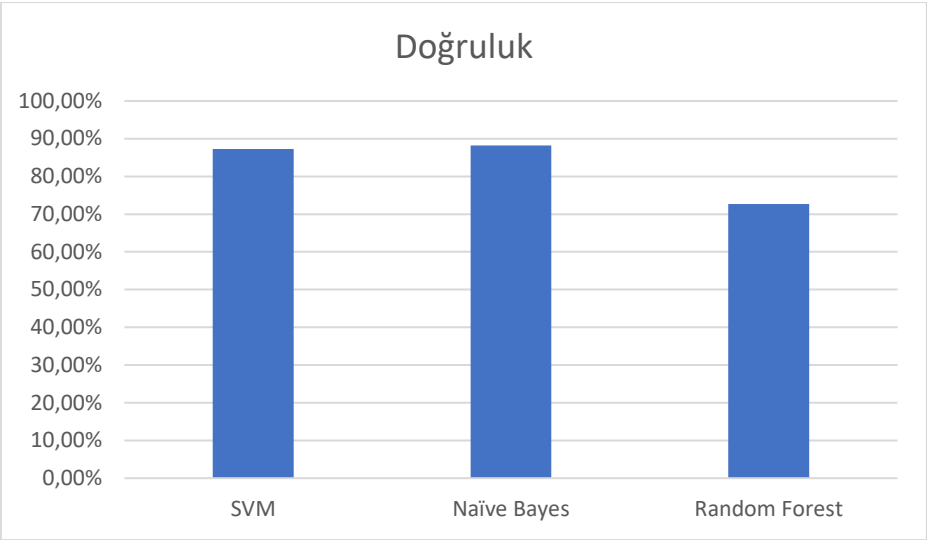
Şekil 5.4. Seviye 1 veri kümesinde %50 benzerlik oranı için sınıflandırma sonuçları

110 elemanlı ve 4 sınıflı Eğitim, Tıp, Sosyal Bilimler ve Mühendislik veri setinde %75 benzerlik oranı ile yapılan sınıflandırmada en yüksek doğruluk oranı Naive Bayes algoritması ile elde edilmiştir (Şekil 5.5).



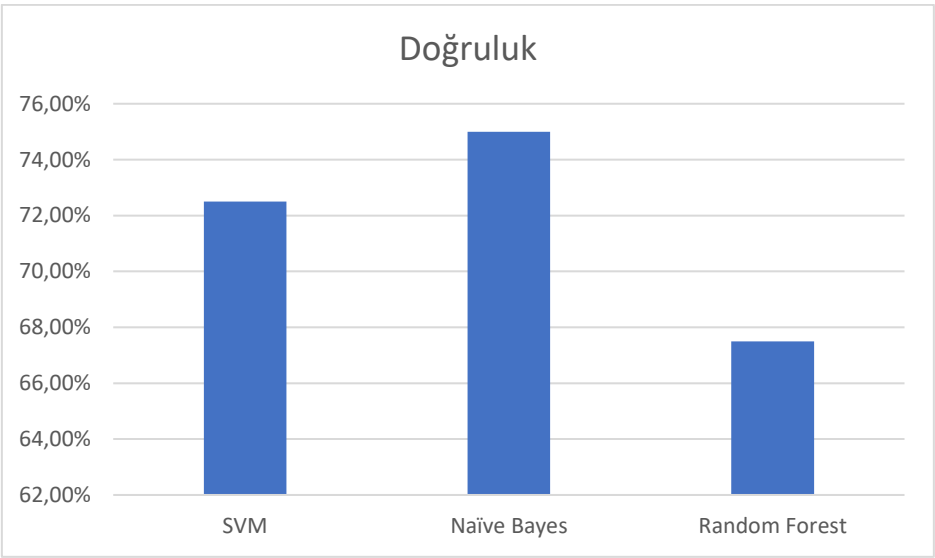
Şekil 5.5. Seviye 1 veri kümesinde %75 benzerlik oranı için sınıflandırma sonuçları

110 elemanlı ve 4 sınıflı Eğitim, Tıp, Sosyal Bilimler ve Mühendislik veri setinde %90 benzerlik oranı ile yapılan sınıflandırmada en yüksek doğruluk oranı Naive Bayes algoritması ile elde edilmiştir (Şekil 5.6).



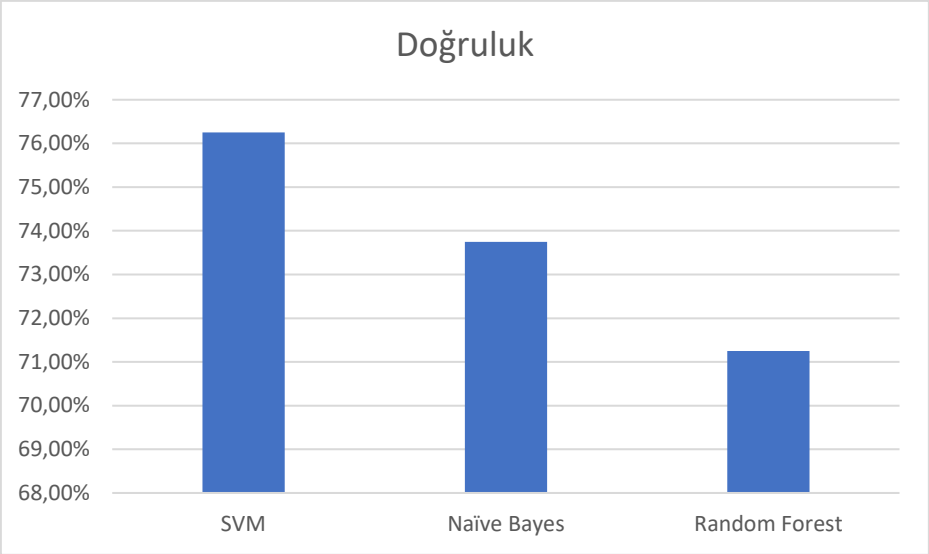
Şekil 5.6. Seviye 1 veri kümesinde %90 benzerlik oranı için sınıflandırma sonuçları

80 elemanlı ve 3 sınıflı Elektronik Mühendisliği, Makine Mühendisliği, ve Bilgisayar Mühendisliği veri setinde %50 benzerlik oranı ile yapılan sınıflandırmada en yüksek doğruluk oranı Naive Bayes algoritması ile elde edilmiştir (Şekil 5.7).



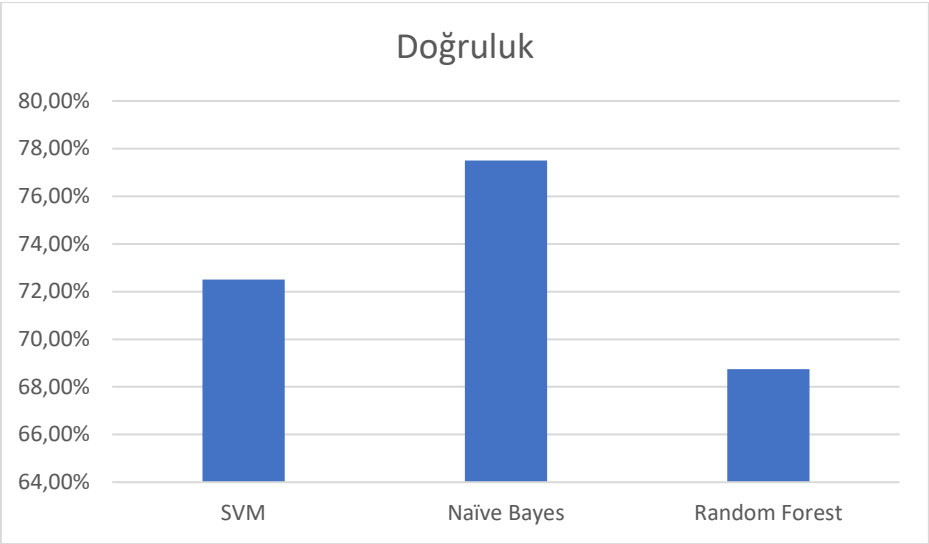
Şekil 5.7. Seviye 2 veri kümesinde %50 benzerlik oranı için sınıflandırma sonuçları

80 elemanlı ve 3 sınıflı Elektronik Mühendisliği, Makine Mühendisliği, ve Bilgisayar Mühendisliği veri setinde %75 benzerlik oranı ile yapılan sınıflandırmada en yüksek doğruluk oranı SVM algoritması ile elde edilmiştir (Şekil 5.8).



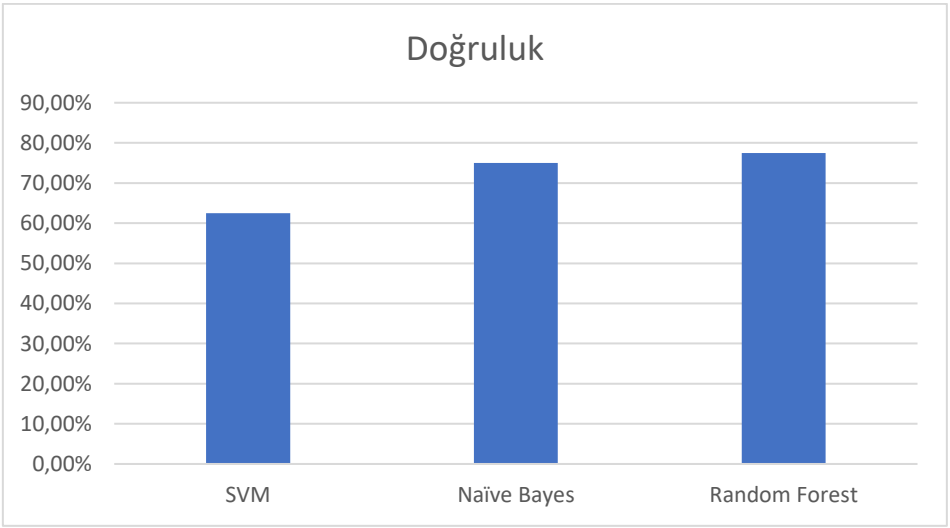
Şekil 5.8. Seviye 2 veri kümesinde %75 benzerlik oranı için sınıflandırma sonuçları

80 elemanlı ve 3 sınıflı Elektronik Mühendisliği, Makine Mühendisliği, ve Bilgisayar Mühendisliği veri setinde %90 benzerlik oranı ile yapılan sınıflandırmada en yüksek doğruluk oranı Naive Bayes algoritması ile elde edilmiştir (Şekil 5.9).



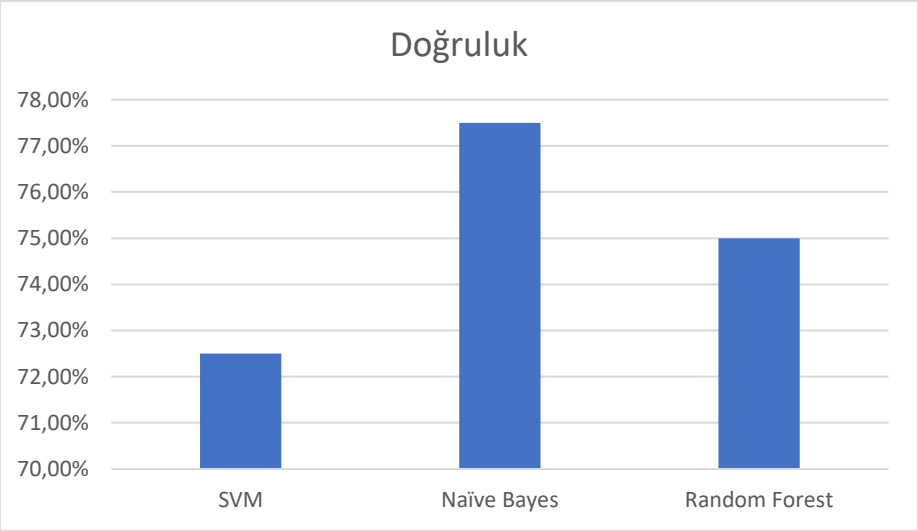
Şekil 5.9. Seviye 2 veri kümesinde %90 benzerlik oranı için sınıflandırma sonuçları

40 elemanlı ve 4 sınıflı Yapay Zeka, Veri Tabanı, Algoritma ve Bilgisayar Ağları veri setinde %50 benzerlik oranı ile yapılan sınıflandırmada en yüksek doğruluk oranı Random Forest algoritması ile elde edilmiştir (Şekil 5.10).



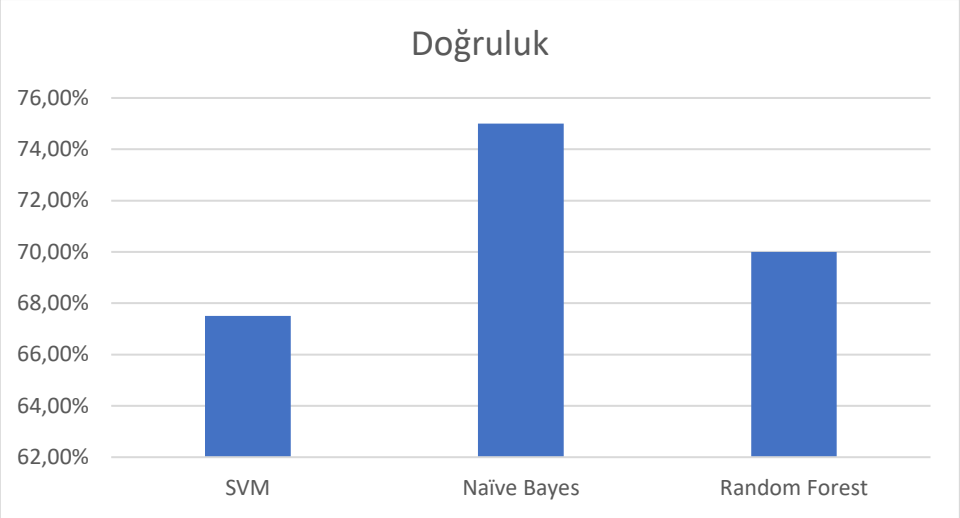
Şekil 5.10. Seviye 3 veri kümesinde %50 benzerlik oranı için sınıflandırma sonuçları

40 elemanlı ve 4 sınıflı Yapay Zeka, Veri Tabanı, Algoritma ve Bilgisayar Ağları veri setinde %75 benzerlik oranı ile yapılan sınıflandırmada en yüksek doğruluk oranı Naive Bayes algoritması ile elde edilmiştir (Şekil 5.11).



Şekil 5.11. Seviye 3 veri kümesinde %75 benzerlik oranı için sınıflandırma sonuçları

40 elemanlı ve 4 sınıflı Yapay Zeka, Veri Tabanı, Algoritma ve Bilgisayar Ağları veri setinde %90 benzerlik oranı ile yapılan sınıflandırmada en yüksek doğruluk oranı Naive Bayes algoritması ile elde edilmiştir (Şekil 5.12).



Şekil 5.12. Seviye 3 veri kümesinde %90 benzerlik oranı için sınıflandırma sonuçları

Sınıflandırmada kullanılan 9 ayrı veri kümesinden 3 tanesi Seviye 1 veri kümeleridir Seviye 1' de yer alan 3 veri kümesi için sınıflandırma sonucu elde edilen doğruluk oranlarının toplu gösterimi Tablo 5.1' de bulunmaktadır.

Tablo 5.1. Seviye 1 veri kümeleri için sınıflandırma doğruluk oranları

Eğitim, Tıp, Sosyal Bilimler ve Mühendislik			
Benzerlik Oranı	**SVM**	**Naïve Bayes**	**Random Forest**
50%	90%	87,27%	73,63%
75%	88,18%	91,81%	72,72%
90%	87,27%	88,18%	72,72%

Sınıflandırmada kullanılan 9 ayrı veri kümesinden 3 tanesi Seviye 2 veri kümeleridir. Seviye 2' de yer alan 3 veri kümesi için sınıflandırma sonucu elde edilen doğruluk oranlarının toplu gösterimi Tablo 5.2' de bulunmaktadır.

Tablo 5.2. Seviye 2 veri kümeleri için sınıflandırma doğruluk oranları

Elektronik Mühendisliği, Makine Mühendisliği ve Bilgisayar Mühendisliği			
Benzerlik Oranı	**SVM**	**Naïve Bayes**	**Random Forest**
50%	72,50%	75,00%	67,50%
75%	76,25%	73,75%	71,25%
90%	72,50%	77,50%	68,75%

Sınıflandırmada kullanılan 9 ayrı veri kümesinden 3 tanesi Seviye 3 veri kümeleridir. Seviye 3' de yer alan 3 veri kümesi için sınıflandırma sonucu elde edilen doğruluk oranlarının toplu gösterimi Tablo 5.3' te bulunmaktadır.

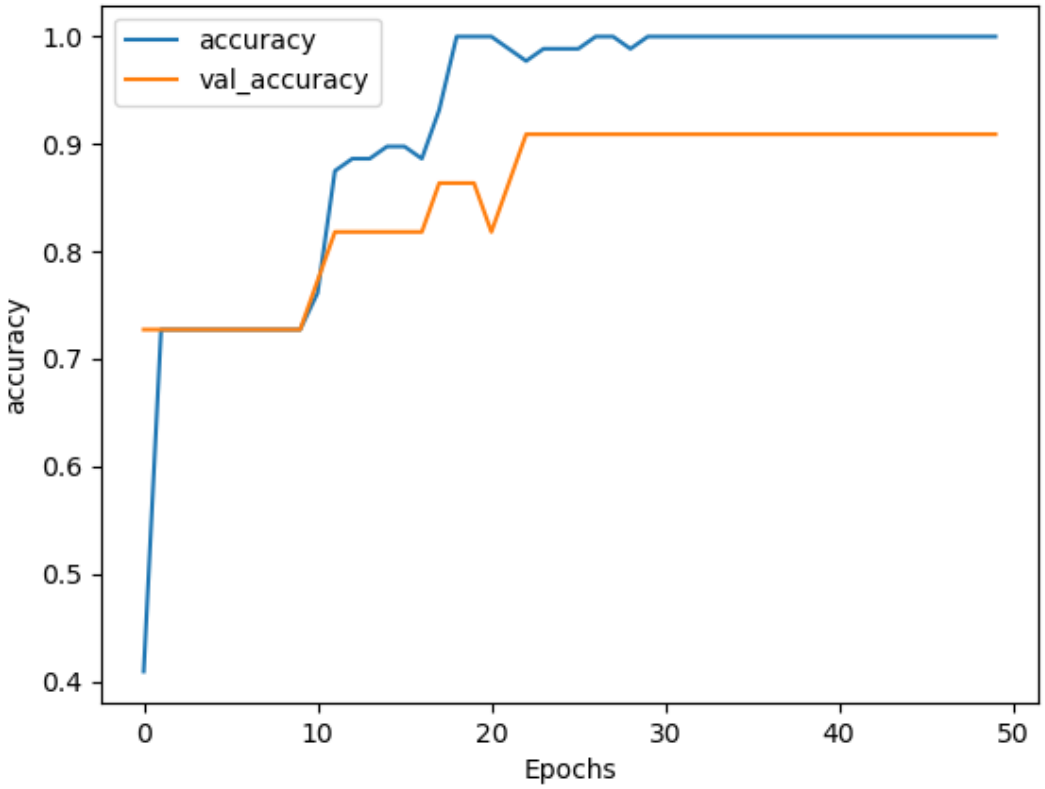
Tablo 5.3. Seviye 3 veri kümeleri için sınıflandırma doğruluk oranları

Yapay Zeka, Veri Tabanı, Algoritma ve Bilgisayar Ağları			
Benzerlik Oranı	**SVM**	**Naïve Bayes**	**Random Forest**
50%	62,50%	75,00%	77,50%
75%	72,50%	77,50%	75,00%
90%	67,50%	75,00%	70,00%

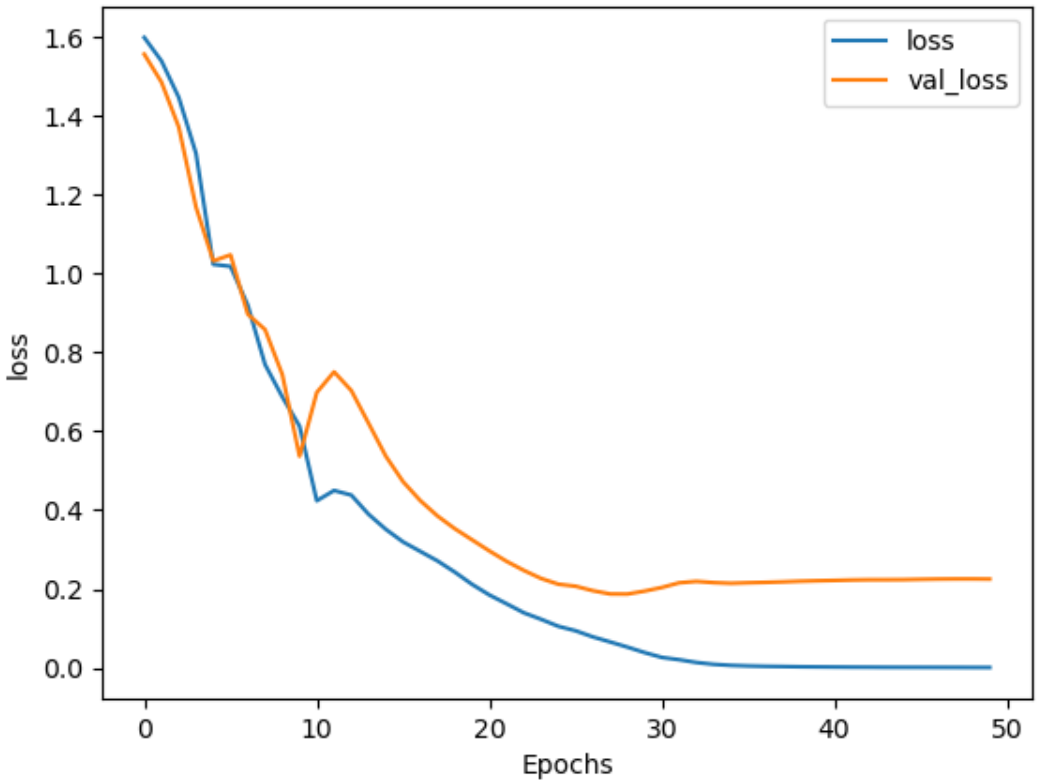
5.2 LSTM ve Geleneksel Makine Öğrenimi Yöntemleri ile Elde Edilen Sonuçlar

Ders videolarından oluşan 9 adet veri kümesinde geleneksel makine öğrenimi yöntemleri ile birlikte derin öğrenme yöntemlerinden LSTM de kullanılmıştır. LSTM yöntemi çift yönlü olarak uygulanmıştır. Hem ileri beslemeli hem de geri beslemeli bir LSTM mimarisi ile sınıflandırma gerçekleştirilmiştir.

LSTM sınıflandırma yöntemi kullanılırken geleneksel makine öğrenimi yöntemlerinde olduğu gibi 10 kat çaprazlama ile sınıflandırma sonuçları elde edilmiştir. Sınıflandırma işlemi öğrenme gerçekleştiğinde doğrulama kaybının (validation loss) artmadığı parametreler ile gerçekleştirilmiştir (Şekil 5.13, Şekil 5.14).

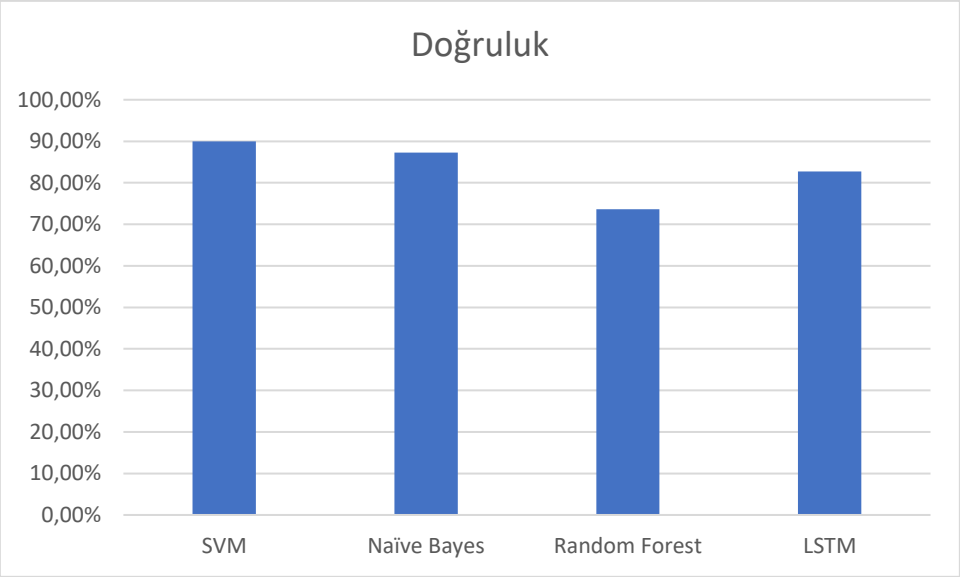


Şekil 5.13. Eğitim ve doğrulama doğruluğu



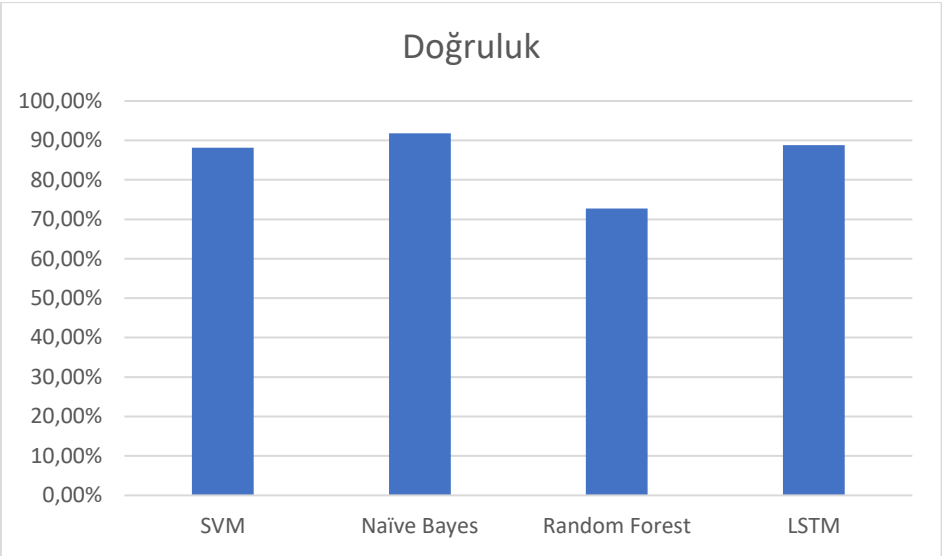
Şekil 5.14. Eğitim ve doğrulama kaybı

110 elemanlı ve 4 sınıflı Eğitim, Tıp, Sosyal Bilimler ve Mühendislik veri setinde %50 benzerlik oranı ile yapılan sınıflandırmada en yüksek doğruluk oranı SVM algoritması ile elde edilmiştir (Şekil 5.15).



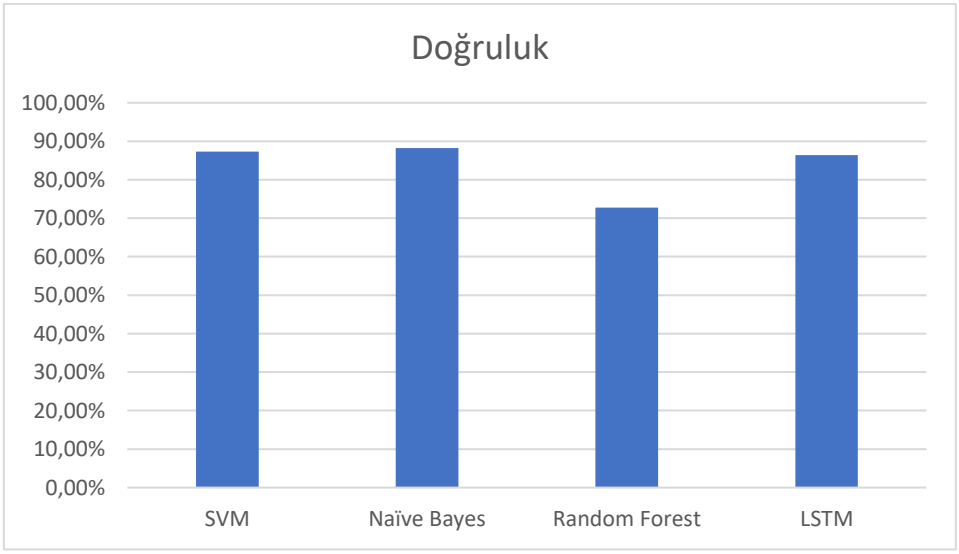
Şekil 5.15. Seviye 1 veri kümesinde %50 benzerlik oranı için sınıflandırma sonuçları

110 elemanlı ve 4 sınıflı Eğitim, Tıp, Sosyal Bilimler ve Mühendislik veri setinde %75 benzerlik oranı ile yapılan sınıflandırmada en yüksek doğruluk oranı Naive Bayes algoritması ile elde edilmiştir (Şekil 5.16).



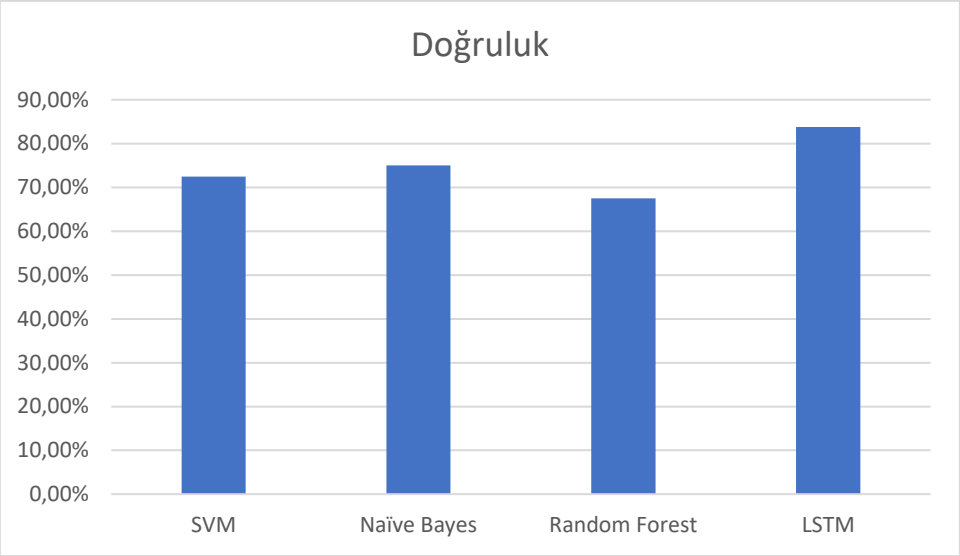
Şekil 5.16. Seviye 1 veri kümesinde %75 benzerlik oranı için sınıflandırma sonuçları

110 elemanlı ve 4 sınıflı Eğitim, Tıp, Sosyal Bilimler ve Mühendislik veri setinde %90 benzerlik oranı ile yapılan sınıflandırmada en yüksek doğruluk oranı Naive Bayes algoritması ile elde edilmiştir (Şekil 5.17).



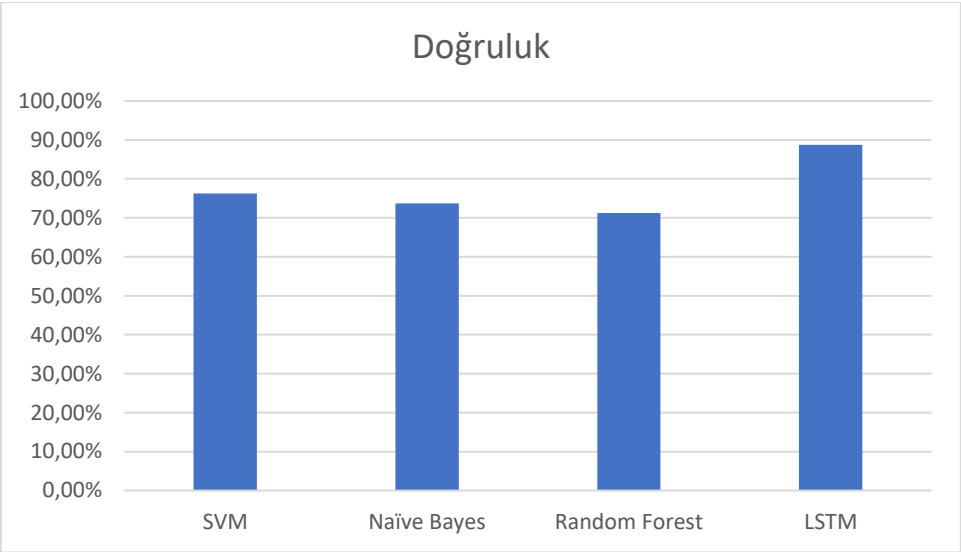
Şekil 5.17. Seviye 1 veri kümesinde %90 benzerlik oranı için sınıflandırma sonuçları

80 elemanlı ve 3 sınıflı Elektronik Mühendisliği, Makine Mühendisliği, ve Bilgisayar Mühendisliği veri setinde %50 benzerlik oranı ile yapılan sınıflandırmada en yüksek doğruluk oranı LSTM algoritması ile elde edilmiştir (Şekil 5.18).



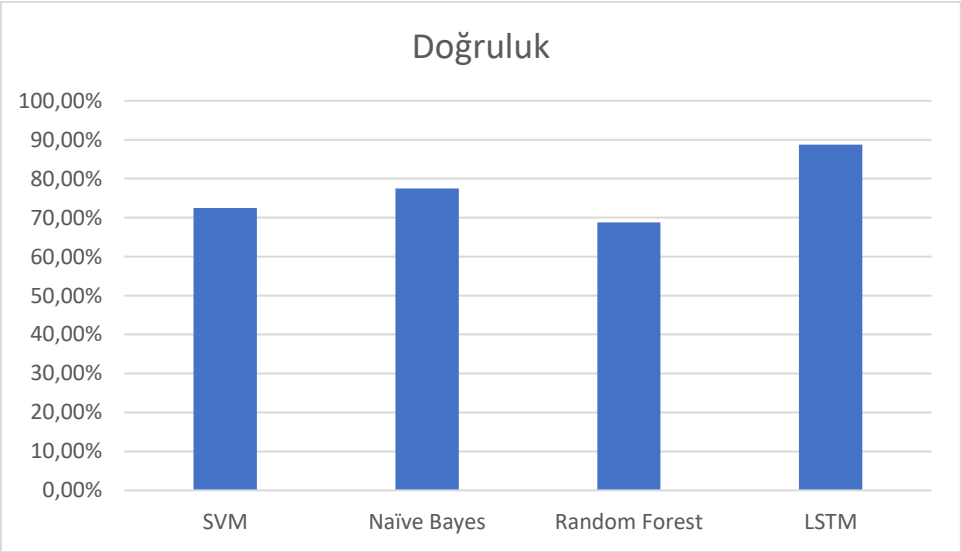
Şekil 5.18. Seviye 2 veri kümesinde %50 benzerlik oranı için sınıflandırma sonuçları

80 elemanlı ve 3 sınıflı Elektronik Mühendisliği, Makine Mühendisliği, ve Bilgisayar Mühendisliği veri setinde %75 benzerlik oranı ile yapılan sınıflandırmada en yüksek doğruluk oranı LSTM algoritması ile elde edilmiştir (Şekil 5.19).



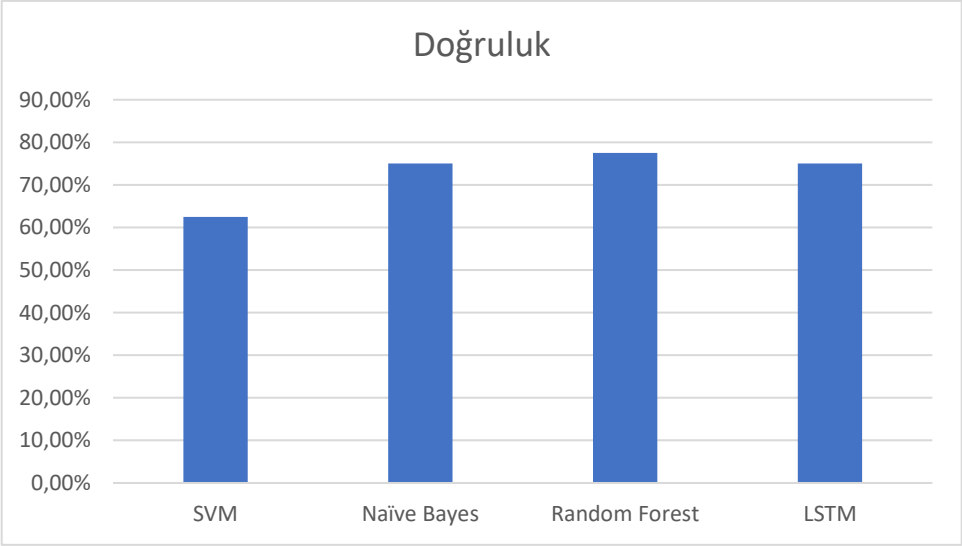
Şekil 5.19. Seviye 2 veri kümesinde %75 benzerlik oranı için sınıflandırma sonuçları

80 elemanlı ve 3 sınıflı Elektronik Mühendisliği, Makine Mühendisliği, ve Bilgisayar Mühendisliği veri setinde %90 benzerlik oranı ile yapılan sınıflandırmada en yüksek doğruluk oranı LSTM algoritması ile elde edilmiştir (Şekil 5.20).



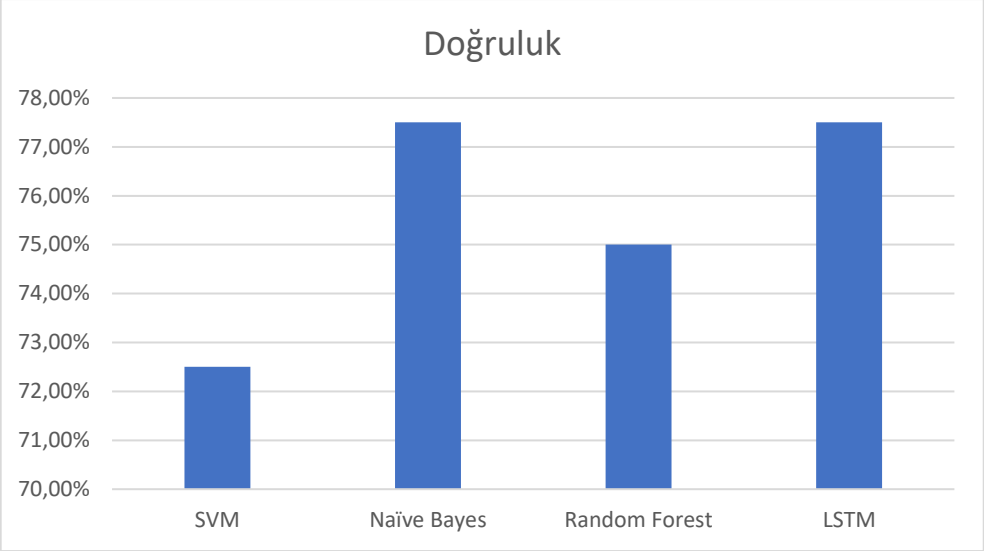
Şekil 5.20. Seviye 2 veri kümesinde %90 benzerlik oranı için sınıflandırma sonuçları

40 elemanlı ve 4 sınıflı Yapay Zeka, Veri Tabanı, Algoritma ve Bilgisayar Ağları veri setinde %50 benzerlik oranı ile yapılan sınıflandırmada en yüksek doğruluk oranı Random Forest algoritması ile elde edilmiştir (Şekil 5.21).



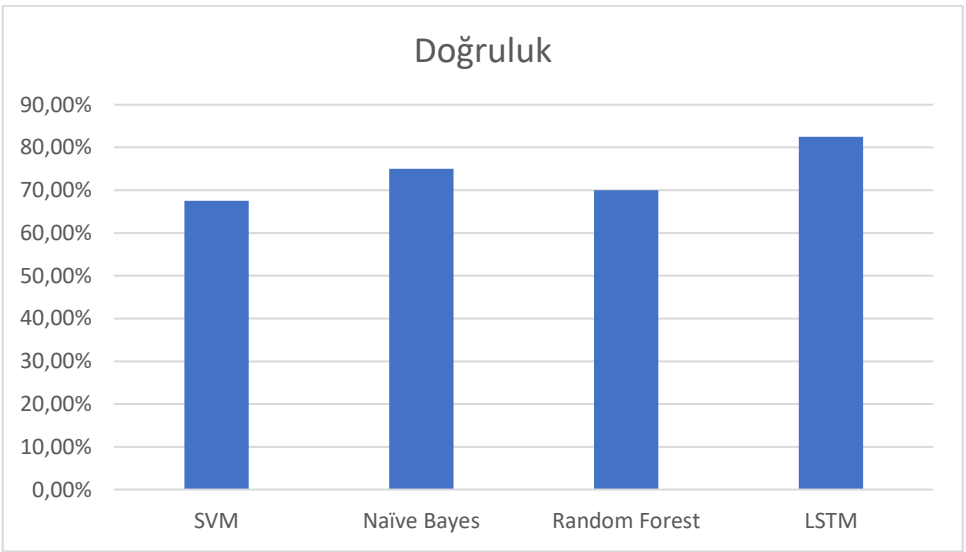
Şekil 5.21. Seviye 3 veri kümesinde %50 benzerlik oranı için sınıflandırma sonuçları

40 elemanlı ve 4 sınıflı Yapay Zeka, Veri Tabanı, Algoritma ve Bilgisayar Ağları veri setinde %75 benzerlik oranı ile yapılan sınıflandırmada en yüksek doğruluk oranı Naive Bayes ve LSTM algoritmaları ile elde edilmiştir (Şekil 5.22).



Şekil 5.22. Seviye 3 veri kümesinde %75 benzerlik oranı için sınıflandırma sonuçları

40 elemanlı ve 4 sınıflı Yapay Zeka, Veri Tabanı, Algoritma ve Bilgisayar Ağları veri setinde %90 benzerlik oranı ile yapılan sınıflandırmada en yüksek doğruluk oranı Naive Bayes ve LSTM algoritmaları ile elde edilmiştir (Şekil 5.23).



Şekil 5.23. Seviye 3 veri kümesinde %90 benzerlik oranı için sınıflandırma sonuçları

Geleneksel makine öğrenimi algoritmaları ve LSTM sınıflandırma yönteminin uygulandığı 9 ayrı veri kümesinden 3 tanesi Seviye 3 veri kümeleridir. Seviye 1' de olan 3 veri kümesi için doğruluk oranları Tablo 5.4' te bulunmaktadır.

Tablo 5.4. Seviye 1 veri kümeleri için sınıflandırma doğruluk oranları

Eğitim, Tıp, Sosyal Bilimler ve Mühendislik				
Benzerlik Oranı	**SVM**	**Naïve Bayes**	**Random Forest**	**LSTM**
50%	90%	87,27%	73,63%	82,73%
75%	88,18%	91,81%	72,72%	88,18%
90%	87,27%	88,18%	72,72%	86,36%

Geleneksel makine öğrenimi algoritmaları ve LSTM sınıflandırma yönteminin uygulandığı 9 ayrı veri kümesinden 3 tanesi Seviye 2 veri kümeleridir. Seviye 2'de olan 3 veri kümesi için doğruluk oranları Tablo 5.5' te bulunmaktadır.

Tablo 5.5. Seviye 2 veri kümeleri için sınıflandırma doğruluk oranları

Elektronik Mühendisliği, Makie Mühendisliği ve Bilgisayar Mühendisliği				
Benzerlik Oranı	**SVM**	**Naïve Bayes**	**Random Forest**	**LSTM**
50%	72,50%	75,00%	67,50%	83,75%
75%	76,25%	73,75%	71,25%	88,75%
90%	72,50%	77,50%	68,75%	88,75%

Geleneksel makine öğrenimi algoritmaları ve LSTM sınıflandırma yönteminin uygulandığı 9 ayrı veri kümesinden 3 tanesi Seviye 3 veri kümeleridir. Seviye 3' de olan 3 veri kümesi için doğruluk oranları Tablo 5.6' da bulunmaktadır.

Tablo 5.6. Seviye 3 veri kümeleri için sınıflandırma doğruluk oranları

Yapay Zeka, Veri Tabanı, Algoritma ve Bilgisayar Ağları				
Benzerlik Oranı	**SVM**	**Naïve Bayes**	**Random Forest**	**LSTM**
50%	62,50%	75,00%	77,50%	75,00%
75%	72,50%	77,50%	75,00%	77,50%
90%	67,50%	75,00%	70,00%	82,50%

5.3 LSTM ve Geleneksel Makine Öğrenimi Yöntemleri ile Elde Edilen Duyarlılık (Recall) Değerleri

Seviye 1, Seviye 2 ve Seviye 3 veri kümelerinden OCR ile çıkarılan metinsel veriler %50, %75 ve %90 benzerlik oranları ile önişleme tabi tutulmuştur. Veri önişlemleri sonrasında çıkarılan dokuz ayrı veri seti üç adet geleneksel makine öğrenme yöntemi ve bir adet derin öğrenme yöntemi ile sınıflandırılmıştır. Tüm sınıflandırmalar doğruluk değerleri açısından ölçülmüştür (Tablo 5.4, Tablo 5.5, Tablo 5.6).

Bu bölümde sınıflandırma sonuçlarının Duyarlılık (Recall) ölçümlerine yer verilmiştir. Duyarlılık sınıflandırma sonunda pozitif olarak sınıflandırılması gereken verilerin ne kadarının pozitif olarak sınıflandırabildiğini ölçen metriktir.

Örnek:

Gerçekleşen			
Bilgisayar Mühendisliği	Elektronik Mühendisliği	Makine Mühendisliği	Beklenen
34	6	0	Bilgisayar Mühendisliği
5	15	0	Elektronik Mühendisliği
8	3	9	Makine Mühendisliği

Yukarıdaki tabloya göre Bilgisayar Mühendisliği sınıfına ait olan 40 veriden 6 tanesi Elektronik Mühendisliği olarak sınıflandırılmıştır. Duyarlılık ölçümü doğru pozitif değerlerin doğru pozitif ve yanlış negatiflere bölünmesi ile bulunur.

$$\text{Duyarlılık} = \frac{\text{Doğru Pozitif}}{\text{Doğru Pozitif} + \text{Yanlış Negatif}} \tag{4.2}$$

Duyarlılık denklemine göre ilgili sınıfların ölçümleri;

Bilgisayar Mühendisliği = 34 / (34 + 6 + 0) = 0,85

Elektronik Mühendisliği = 15 / (5 + 15 + 0) = 0,75

Makine Mühendisliği = 9 / (8 + 3 + 9) = 0,45

Geleneksel makine öğrenimi yöntemlerinin ve LSTM sınıflandırma yönteminin uygulandığı 9 ayrı veri kümesinden 3 tanesi Seviye 1 veri kümeleridir. Seviye 1 ’de olan 3 veri kümesi için duyarlılık oranları Tablo 5.7’ de bulunmaktadır.

Tablo 5.7. Seviye 1 veri kümeleri için duyarlılık oranları

Eğitim, Tıp, Sosyal Bilimler ve Mühendislik				
Benzerlik Oranı	**SVM**	**Naïve Bayes**	**Random Forest**	**LSTM**
50%	90,00%	87,30%	73,60%	82,70%
75%	88,20%	91,80%	72,70%	88,20%
90%	87,30%	88,20%	72,70%	86,40%

Geleneksel makine öğrenimi yöntemlerinin ve LSTM sınıflandırma yönteminin uygulandığı 9 ayrı veri kümesinden 3 tanesi Seviye 2 veri kümeleridir. Seviye 2 ’de olan 3 veri kümesi için duyarlılık oranları Tablo 5.8’ de bulunmaktadır.

Tablo 5.8. Seviye 2 veri kümeleri için duyarlılık oranları

Elektronik Mühendisliği, Makie Mühendisliği ve Bilgisayar Mühendisliği				
Benzerlik Oranı	**SVM**	**Naïve Bayes**	**Random Forest**	**LSTM**
50%	72,50%	75,00%	67,50%	83,80%
75%	76,30%	73,80%	71,30%	88,80%
90%	72,50%	77,50%	68,80%	88,80%

Geleneksel makine öğrenimi yöntemlerinin ve LSTM sınıflandırma yönteminin uygulandığı 9 ayrı veri kümesinden 3 tanesi Seviye 3 veri kümeleridir. Seviye 3 ’de olan 3 veri kümesi için duyarlılık oranları Tablo 5.9’ da bulunmaktadır.

Tablo 5.9. Seviye 3 veri kümeleri için duyarlılık oranları

Yapay Zeka, Veri Tabanı, Algoritma ve Bilgisayar Ağları				
Benzerlik Oranı	**SVM**	**Naïve Bayes**	**Random Forest**	**LSTM**
50%	62,50%	75,00%	77,50%	75,00%
75%	72,50%	77,50%	75,00%	77,50%
90%	67,50%	75,00%	70,00%	82,50%

5.4 LSTM ve Geleneksel Makine Öğrenimi Yöntemleri ile Elde Edilen Kesinlik (Precision) Değerleri

Seviye 1, Seviye 2 ve Seviye 3 veri kümelerinden OCR ile çıkarılan metinsel veriler %50, %75 ve %90 benzerlik oranları ile önişleme tabi tutulmuştur. Veri önişlemleri sonrasında çıkarılan dokuz ayrı veri seti üç adet geleneksel makine öğrenme yöntemi ve bir adet derin öğrenme yöntemi ile sınıflandırılmıştır. Tüm sınıflandırmalar doğruluk değerleri açısından ölçülmüştür (Tablo 5.4, Tablo 5.5, Tablo 5.6).

Bu bölümde sınıflandırma sonuçlarının Kesinlik (Precision) ölçümlerine yer verilmiştir. Duyarlılık sınıflandırma sonunda pozitif olarak sınıflandırılan değerlerin kaç tanesinin gerçekten pozitif olarak sınıflandırabildiğini ölçen metriktir.

Örnek:

Gerçekleşen			
Bilgisayar Mühendisliği	Elektronik Mühendisliği	Makine Mühendisliği	Beklenen
34	6	0	Bilgisayar Mühendisliği
5	15	0	Elektronik Mühendisliği
8	3	9	Makine Mühendisliği

Tabloya göre Bilgisayar Mühendisliği olarak sınıflandırılan 47 veriden 34 tanesi gerçekten Bilgisayar Mühendisliği sınıfına aittir. Kesinlik ölçümü doğru pozitif değerlerin doğru pozitif ve yanlış pozitiflere bölünmesi ile bulunur.

$$\text{Kesinlik} = \frac{\text{Doğru Pozitif}}{\text{Doğru Pozitif} + \text{Yanlış Pozitif}} \tag{4.3}$$

Kesinlik denklemine göre ilgili sınıfların ölçümleri;
Bilgisayar Mühendisliği = 34 / (34 + 5 + 8) = 0,72
Elektronik Mühendisliği = 15 / (6 + 15 + 3) = 0,62
Makine Mühendisliği = 9 / (0 + 0 + 9) = 1

Geleneksel makine öğrenimi yöntemlerinin ve LSTM sınıflandırma yönteminin uygulandığı 9 ayrı veri kümesinden 3 tanesi Seviye 1 veri kümeleridir. Seviye 1 ’de olan 3 veri kümesi için kesinlik oranları Tablo 5.10’ da bulunmaktadır.

Bu tez çalışmasında çok sınıflı veri kümeleri ile çalışılmıştır. Random Forest algoritması Seviye 1 veri kümesi sınıflandırmalarında Kesinlik ölçümünün yapılamayacağı ölçüde hatalı sınıflandırmada bulunmuştur. Tablo 5.10’ da Random Forest algoritmasının Kesinlik değerleri bu sebepten hesaplanamamıştır.

Örnek:

Gerçekleşen			
Bilgisayar Mühendisliği	Elektronik Mühendisliği	Makine Mühendisliği	Beklenen
0	20	20	Bilgisayar Mühendisliği
0	15	5	Elektronik Mühendisliği
0	3	17	Makine Mühendisliği

Tabloya göre Bilgisayar Mühendisliği ait olan hiçbir sınıflandırma doğru yapılamamıştır yani Doğru Pozitif değeri sıfırdır. Ayrıca herhangi bir Yanlış Pozitif sınıflandırması da yapılmamıştır. Bu sebepten kesinlik hesaplanamaz.

$$\text{Kesinlik} = \frac{\text{Doğru Pozitif}}{\text{Doğru Pozitif} + \text{Yanlış Pozitif}} \tag{4.4}$$

Kesinlik denklemine göre ölçüm;
Bilgisayar Mühendisliği = 0 / 0 = Belirsiz

Tablo 5.10. Seviye 1 veri kümeleri için kesinlik oranları

Eğitim, Tıp, Sosyal Bilimler ve Mühendislik				
Benzerlik Oranı	**SVM**	**Naïve Bayes**	**Random Forest**	**LSTM**
50%	91,20%	91,50%	-	82,30%
75%	88,00%	93,90%	-	88,20%
90%	88,30%	92,20%	-	88,20%

Geleneksel makine öğrenimi yöntemlerinin ve LSTM sınıflandırma yönteminin uygulandığı 9 ayrı veri kümesinden 3 tanesi Seviye 2 veri kümeleridir. Seviye 2 ’de olan 3 veri kümesi için kesinlik oranları Tablo 5.11’ de bulunmaktadır.

Tablo 5.11. Seviye 2 veri kümeleri için kesinlik oranları

Elektronik Mühendisliği, Makie Mühendisliği ve Bilgisayar Mühendisliği				
Benzerlik Oranı	**SVM**	**Naïve Bayes**	**Random Forest**	**LSTM**
50%	76,80%	76,70%	74,50%	86,70%
75%	77,70%	75,80%	79,20%	89,30%
90%	73,10%	78,30%	79,20%	88,90%

Geleneksel makine öğrenimi yöntemlerinin ve LSTM sınıflandırma yönteminin uygulandığı 9 ayrı veri kümesinden 3 tanesi Seviye 3 veri kümeleridir. Seviye 3 'de olan 3 veri kümesi için kesinlik oranları Tablo 5.12' de bulunmaktadır.

Tablo 5.12. Seviye 3 veri kümeleri için kesinlik oranları

Yapay Zeka, Veri Tabanı, Algoritma ve Bilgisayar Ağları				
Benzerlik Oranı	**SVM**	**Naïve Bayes**	**Random Forest**	**LSTM**
50%	79,60%	84,60%	83,40%	76,50%
75%	84,00%	84,70%	82,20%	83,20%
90%	82,70%	82,70%	78,20%	86,90%

5.5 LSTM ve Geleneksel Makine Öğrenimi Yöntemleri ile Elde Edilen F1 Skoru Değerleri

Seviye 1, Seviye 2 ve Seviye 3 veri kümelerinden OCR ile çıkarılan metinsel veriler %50, %75 ve %90 benzerlik oranları ile önişleme tabi tutulmuştur. Veri önişlemleri sonrasında çıkarılan dokuz ayrı veri seti üç adet geleneksel makine öğrenme yöntemi ve bir adet derin öğrenme yöntemi ile sınıflandırılmıştır. Tüm sınıflandırmalar doğruluk değerleri açısından ölçülmüştür (Tablo 5.4, Tablo 5.5, Tablo 5.6).

Bu bölümde sınıflandırma sonuçlarının F1 skoru ölçümlerine yer verilmiştir. F1 skoru Duyarlılık ve Kesinlik değerlerinin Harmonik Ortalaması alınarak bulunur. Aritmetik Ortalama yerine Harmonik Ortalama alınmasının sebebi uç örnekleri göz ardı etmemektir.

Bu tezdeki Seviye 1 ve Seviye 2 veri kümleri eşit dağılıma sahip değildir. Eşit dağılıma sahip olmayan veri kümelerinde F1 Skor değeri Doğruluk sonuçlarından daha açıklayıcı sonuçlar vermektedir. Bu noktası ile bakıldığında F1 skoru bu tezdeki veri kümelerinin dengesiz dağılıma sahip olduğundan sebep azami öneme sahiptir.

Örnek:

Gerçekleşen			
Bilgisayar Mühendisliği	Elektronik Mühendisliği	Makine Mühendisliği	Beklenen
34	6	0	Bilgisayar Mühendisliği
5	15	0	Elektronik Mühendisliği
8	3	9	Makine Mühendisliği

$$\mathrm{F1} = 2 \text{ x } \frac{\text{Duyarlılık x Kesinlik}}{\text{Duyarlılık + Kesinlik}} \tag{4.5}$$

Yukarıdaki tabloya ait olan ve aşağıdaki hesaplama için kullanılan Duyarlılık ve Kesinlik değerleri "5.3 LSTM ve Geleneksel Makine Öğrenimi Yöntemleri ile Elde Edilen Sonuçlar Duyarlılık (Precision) Değerleri" ve "5.4 LSTM ve Geleneksel Makine Öğrenimi Yöntemleri ile Elde Edilen Sonuçlar Kesinlik (Precision) Değerleri" bölümlerinde hesaplanmıştır.

F1 Skoru denklemine göre ilgili sınıfların ölçümleri;
Bilgisayar Mühendisliği = 2 x 0,72 x 0,85 / 0,72 + 0,85 = 0,78
Elektronik Mühendisliği = 2 x 0,62 x 0,750 / 0,62 + 0,75 = 0,68
Makine Mühendisliği = 2 x 1 x 0,45 / 1 + 0,45 = 0,62

Geleneksel makine öğrenimi yöntemlerinin ve LSTM sınıflandırma yönteminin uygulandığı 9 ayrı veri kümesinden 3 tanesi Seviye 1 veri kümeleridir. Seviye 1 'de olan 3 veri kümesi için kesinlik oranları Tablo 5.13' te bulunmaktadır.

Bu tez çalışmasında çok sınıflı veri kümeleri ile çalışılmıştır. Random Forest algoritması kullanıldığında Seviye 1 veri kümesinde "5.4 LSTM ve Geleneksel Makine Öğrenimi Yöntemleri ile Elde Edilen Sonuçlar Kesinlik (Precision) Değerleri" bölümünde anlatıldığı gibi

kesinlik değeri hesaplanamayacak sınıflandırma sonuçları elde edilmiştir. Bu sebepten Tablo 5.13' te Random Forest algoritması için F1 Skoru da hesaplanamamıştır.

Tablo 5.13. Seviye 1 veri kümeleri için F1 skoru oranları

Eğitim, Tıp, Sosyal Bilimler ve Mühendislik				
Benzerlik Oranı	**SVM**	**Naïve Bayes**	**Random Forest**	**LSTM**
50%	88%	88,20%	-	79,90%
75%	87,20%	92,30%	-	87,60%
90%	86,50%	89,00%	-	86,40%

Geleneksel makine öğrenimi yöntemlerinin ve LSTM sınıflandırma yönteminin uygulandığı 9 ayrı veri kümesinden 3 tanesi Seviye 2 veri kümeleridir. Seviye 2 'de olan 3 veri kümesi için kesinlik oranları Tablo 5.14' te bulunmaktadır.

Tablo 5.14. Seviye 2 veri kümeleri için F1 skoru oranları

Elektronik Mühendisliği, Makie Mühendisliği ve Bilgisayar Mühendisliği				
Benzerlik Oranı	**SVM**	**Naïve Bayes**	**Random Forest**	**LSTM**
50%	71,60%	75,20%	63,60%	84,00%
75%	75,70%	74,10%	67,30%	88,70%
90%	72,40%	77,70%	64,00%	88,80%

Geleneksel makine öğrenimi yöntemlerinin ve LSTM sınıflandırma yönteminin uygulandığı 9 ayrı veri kümesinden 3 tanesi Seviye 3 veri kümeleridir. Seviye 3 'de olan 3 veri kümesi için kesinlik oranları Tablo 5.15' te bulunmaktadır.

Tablo 5.15. Seviye 3 veri kümeleri için F1 skoru oranları

Yapay Zeka, Veri Tabanı, Algoritma ve Bilgisayar Ağları				
Benzerlik Oranı	**SVM**	**Naïve Bayes**	**Random Forest**	**LSTM**
50%	62,80%	75,90%	77,80%	75,30%
75%	70,90%	77,40%	75,20%	77,40%
90%	64,70%	76,10%	71,00%	83,40%

6. TARTIŞMA

Bu tezde 110 videoluk bir veri kümesi bulunan videolardan her 30 karede bir yani her 1 saniyede 1 defa OCR çıktıları alınmıştır. Burada açık kaynak kodlu Tesseract OCR yazılımı kullanılmıştır. Tesseract OCR yazılımı yerine daha yüksek tanıma oranına sahip bir OCR uygulaması kullanılsaydı daha fazla metinsel veri elde edilebilirdi.

OCR ile metinsel veri çıkarımı sonrasında Naive Bayes, Support Vector Machine ve Random Forest sınıflandırma yöntemleri için veri önişleme gerçekleştirilmiştir. Bunlar çıkarılan kareleri benzerlik oranına göre karşılaştırarak eşik değer üzerinde kalan kareleri göz ardı etmek, metinsel verileri küçük harfe çevirerek gereksiz kelimeleri atmak, kelimeleri İngilizce sözlükte kontrol ederek OCR tarafından yanlış hatalı çıkarılmış kelimeleri atmak, kelime frekansını bulmak ve kelime frekanslarını normalleştirme işlemleridir. Bu işlemler sonrasında metinler .arff formatına çevrilerek Weka programı ile sınıflandırılmıştır. Tüm sınıflandırmalar 10 kat çaprazlama yöntemi ile gerçekleştirilmiştir.

OCR ile metinsel veri çıkarımı sonrası Long Short-Term Memory sınıflandırma yöntemi için veri önişleme gerçekleştirilmiştir. Bunlar çıkarılan kareleri benzerlik oranına göre karşılaştırarak eşik değer üzerinde kalan kareleri göz ardı etmek, metinsel verileri küçük harfe çevirerek gereksiz kelimeleri atmak, kelimeleri İngilizce sözlükte kontrol ederek OCR tarafından yanlış hatalı çıkarılmış kelimeleri atmak ve kelimeleri sıralama vektörlerine çevirmektir. Metinlerin sıralama vektörlerine çevrilmesi işlemi cümleleri kelime kelime ayırarak her kelimeye sayısal bir değer vermek, takviye etme ve kırpma işlemleridir. Metinlerin sıralama vektörlerine çevrilme işlemleri 4.6.1, 4.6.2 ve 4,6.3 bölümlerinde ayrıntılı olarak görülebilmektedir.

Metinlerin sıralama vektörlerine çevrilme işlemleri sonrasında sınıflandırma Long Short-Term Memory sınıflandırma yöntemi ile gerçekleştirilmiştir. Kullanılan uygulama Python yazılımın dilinde geliştirilmiştir. Tüm sınıflandırmalar 10 kat çaprazlama yöntemi ile gerçekleştirilmiştir.

Metinlerden çıkarılan kareleri benzerlik oranına göre karşılaştırarak eşik değer üzerinde kalan kareleri göz ardı etmek, metinsel verileri küçük harfe çevirerek gereksiz kelimeleri atmak

ve kelimeleri İngilizce sözlükte kontrol ederek OCR tarafından yanlış hatalı çıkarılmış kelimeleri atmak veri önişlemleri dört sınıflandırma yöntemi için de aynıdır.

Metinsel veri önişlemelerde farklı olan kısım geleneksel makine yöntemleri için frekans vektörlerinin kullanılması bunun yanında derin öğrenme yöntemi için sıralama vektörlerinin kullanılmasıdır.

6.1. SVM Sonuçları

Seviye 1 veri kümeleri %50, %75 ve %90 benzerlik oranlarına göre ayrılarak SVM sınıflandırma yöntemi ile sınıflandırılmıştır. Seviye 1 veri kümeleri en fazla sayıda video dosyasına sahip olan veri kümeleridir. Bu sınıflandırma sonunda en yüksek doğruluk oranı %50 benzerlik oranı kullanılan veri kümesinde elde edilmiştir. En düşük doğruluk oranı %90 benzerlik oranına sahip veri kümesinde elde edilmiştir (Tablo 5.4).

Seviye 2 veri kümeleri %50, %75 ve %90 benzerlik oranlarına göre ayrılarak SVM sınıflandırma yöntemi ile sınıflandırılmıştır. Seviye 2 veri kümeleri en az sayıda farklı sınıfı sahip olan veri kümeleridir. Seviye 1 ve Seviye 3 veri kümelerinde 4' er adet sınıf varken Seviye 2 veri kümelerinde 3' er adet farklı sınıf bulunmaktadır. Bu sınıflandırmada en yüksek doğruluk oranı %75 benzerlik oranı kullanılan veri kümesinde elde edilmiştir. %50 ve %90 benzerlik oranı kullanılan veri kümelerinde eşit doğruluk oranı elde edilmiştir (Tablo 5.5).

Seviye 3 veri kümeleri %50, %75 ve %90 benzerlik oranlarına göre ayrılarak SVM sınıflandırma yöntemi ile sınıflandırılmıştır. Seviye 3 veri kümesi en az sayıda video dosyasına sahip olan veri kümeleridir. Bu sınıflandırma sonunda en yüksek doğruluk oranı %75 benzerlik oranı kullanılan veri kümesinde elde edilmiştir. En düşük doğruluk oranı %50 benzerlik oranına sahip veri kümesinde elde edilmiştir (Tablo 5.6).

SVM sınıflandırma yöntemi ile yapılan sınıflandırmalar doğruluk oranları benzerlik oranları açısından incelendiğinde aralarında direk bir ilişki bulunmamaktadır. SVM sınıflandırma yöntemi ile yapılan sınıflandırmalarda farklı veri kümelerinde farklı doğruluk oranları ile karşılaşılmıştır. Veri kümesinden bağımsız olarak en iyi benzerlik oranı bulunmamaktadır.

Benzerlik oranları sabit tutulup Seviye 1, Seviye 2 ve Seviye 3 veri kümelerinde elde edilen doğruluk oranları karşılaştırıldığında veri kümeleri küçüldükçe sınıflandırma başarısının

düştüğü görülmektedir (Tablo 5.4, Tablo 5.5, Tablo 5.6). Aralarında paralel ilişki bulunmaktadır.

SVM sınıflandırma yöntemi ile yapılan sınıflandırmaların duyarlılık oranları benzerlik oranları açısından incelendiğinde aralarında direk bir ilişki bulunmamaktadır. SVM sınıflandırma yöntemi ile yapılan sınıflandırmalarda farklı veri kümelerinde farklı duyarlılık oranları ile karşılaşılmıştır. Veri kümesinden bağımsız olarak en iyi benzerlik oranı bulunmamaktadır.

Benzerlik oranları sabit tutulup Seviye 1, Seviye 2 ve Seviye 3 veri kümelerinde elde edilen duyarlılık oranları karşılaştırıldığında veri kümeleri küçüldükçe ve veri kümeleri daha dengeli bir dağılıma sahip oldukça duyarlılık oranlarının görülmektedir (Tablo 5.7, Tablo 5.8, Tablo 5.9). Aralarında paralel ilişki bulunmaktadır.

SVM sınıflandırma yöntemi ile yapılan sınıflandırmaların kesinlik oranları benzerlik oranları açısından incelendiğinde aralarında direk bir ilişki bulunmamaktadır. SVM sınıflandırma yöntemi ile yapılan sınıflandırmalarda farklı veri kümelerinde farklı kesinlik oranları ile karşılaşılmıştır. Veri kümesinden bağımsız olarak en iyi benzerlik oranı bulunmamaktadır.

Benzerlik oranları sabit tutulup Seviye 1, Seviye 2 ve Seviye 3 veri kümelerinde elde edilen kesinlik oranları karşılaştırıldığında daha az sınıfa sahip Seviye 2 veri kümesinde kesinlik oranının düştüğü görünmektedir. (Tablo 5.10, Tablo 5.11, Tablo 5.12).

SVM sınıflandırma yöntemi ile yapılan sınıflandırmaların F1 skoru oranları benzerlik oranları açısından incelendiğinde aralarında direk bir ilişki bulunmamaktadır. SVM sınıflandırma yöntemi ile yapılan sınıflandırmalarda farklı veri kümelerinde farklı F1 skoru oranları ile karşılaşılmıştır. Veri kümesinden bağımsız olarak en iyi benzerlik oranı bulunmamaktadır.

Benzerlik oranları sabit tutulup Seviye 1, Seviye 2 ve Seviye 3 veri kümelerinde elde edilen F1 skoru oranları karşılaştırıldığında veri kümesi küçüldükçe ve dengeli bir dağılıma sahip oldukça F1 skoru oranının düştüğü görünmektedir. (Tablo 5.13, Tablo 5.14, Tablo 5.15).

6.2. Naive Bayes Sonuçları

Seviye 1 veri kümeleri %50, %75 ve %90 benzerlik oranlarına göre ayrılarak Naive Bayes sınıflandırma yöntemi ile sınıflandırılmıştır. Bu sınıflandırma sonunda en yüksek doğruluk

oranı %75 benzerlik oranı kullanılan veri kümesinde elde edilmiştir. En düşük doğruluk oranı %50 benzerlik oranına sahip veri kümesinde elde edilmiştir (Tablo 5.4).

Seviye 2 veri kümeleri %50, %75 ve %90 benzerlik oranlarına göre ayrılarak Naive Bayes sınıflandırma yöntemi ile sınıflandırılmıştır. Bu sınıflandırmada en yüksek doğruluk oranı %90 benzerlik oranı kullanılan veri kümesinde elde edilmiştir. En düşük doğruluk oranı %75 benzerlik oranına sahip veri kümesinde elde edilmiştir (Tablo 5.5).

Seviye 3 veri kümeleri %50, %75 ve %90 benzerlik oranlarına göre ayrılarak Naive Bayes sınıflandırma yöntemi ile sınıflandırılmıştır. Bu sınıflandırma sonunda en yüksek doğruluk oranı %75 benzerlik oranı kullanılan veri kümesinde elde edilmiştir. %50 ve %90 benzerlik oranına sahip veri kümelerinde eşit doğruluk oranları elde edilmiştir (Tablo 5.6).

Naive Bayes sınıflandırma yöntemi ile yapılan sınıflandırmalar doğruluk oranları benzerlik oranları açısından incelendiğinde aralarında direk bir ilişki bulunmamaktadır. Naive Bayes sınıflandırma yöntemi ile yapılan sınıflandırmalarda farklı veri kümelerinde farklı doğruluk oranları ile karşılaşılmıştır. Veri kümesinden bağımsız olarak en iyi benzerlik oranı bulunmamaktadır.

Benzerlik oranları sabit tutulup Seviye 1, Seviye 2 ve Seviye 3 veri kümelerinde elde edilen doğruluk oranları karşılaştırıldığında veri kümeleri ile doğruluk oranları arasında bir ilişki tespit edilmemiştir (Tablo 5.4, Tablo 5.5, Tablo 5.6).

Naive Bayes sınıflandırma yöntemi ile yapılan sınıflandırmaların duyarlılık oranları benzerlik oranları açısından incelendiğinde aralarında direk bir ilişki bulunmamaktadır. Naive Bayes sınıflandırma yöntemi ile yapılan sınıflandırmalarda farklı veri kümelerinde farklı duyarlılık oranları ile karşılaşılmıştır. Veri kümesinden bağımsız olarak en iyi benzerlik oranı bulunmamaktadır.

Benzerlik oranları sabit tutulup Seviye 1, Seviye 2 ve Seviye 3 veri kümelerinde elde edilen duyarlılık oranları karşılaştırıldığında aralarında bir ilişki tespit edilememiştir (Tablo 5.7, Tablo 5.8, Tablo 5.9).

Naive Bayes sınıflandırma yöntemi ile yapılan sınıflandırmaların kesinlik oranları benzerlik oranları açısından incelendiğinde aralarında direk bir ilişki bulunmamaktadır. Naive Bayes sınıflandırma yöntemi ile yapılan sınıflandırmalarda farklı veri kümelerinde farklı kesinlik oranları ile karşılaşılmıştır. Veri kümesinden bağımsız olarak en iyi benzerlik oranı bulunmamaktadır.

Benzerlik oranları sabit tutulup Seviye 1, Seviye 2 ve Seviye 3 veri kümelerinde elde edilen kesinlik oranları karşılaştırıldığında aralarında bir ilişki tespit edilememiştir (Tablo 5.10, Tablo 5.11, Tablo 5.12).

Naive Bayes sınıflandırma yöntemi ile yapılan sınıflandırmaların F1 skoru oranları benzerlik oranları açısından incelendiğinde aralarında direk bir ilişki bulunmamaktadır. Naive Bayes sınıflandırma yöntemi ile yapılan sınıflandırmalarda farklı veri kümelerinde farklı F1 skoru oranları ile karşılaşılmıştır. Veri kümesinden bağımsız olarak en iyi benzerlik oranı bulunmamaktadır.

Benzerlik oranları sabit tutulup Seviye 1, Seviye 2 ve Seviye 3 veri kümelerinde elde edilen F1 skoru oranları karşılaştırıldığında en yüksek F1 skorunun Seviye 1 veri kümesinde görüldüğü tespit edilmiştir (Tablo 5.13, Tablo 5.14, Tablo 5.15).

6.3. Random Forest Sonuçları

Seviye 1 veri kümeleri %50, %75 ve %90 benzerlik oranlarına göre ayrılarak Random Forest sınıflandırma yöntemi ile sınıflandırılmıştır. Bu sınıflandırma sonunda en yüksek doğruluk oranı %50 benzerlik oranı kullanılan veri kümesinde elde edilmiştir. %50 ve %90 benzerlik oranına sahip veri kümelerinde eşit doğruluk oranları elde edilmiştir (Tablo 5.4).

Seviye 2 veri kümeleri %50, %75 ve %90 benzerlik oranlarına göre ayrılarak Random Forest sınıflandırma yöntemi ile sınıflandırılmıştır. Bu sınıflandırmada en yüksek doğruluk oranı %75 benzerlik oranı kullanılan veri kümesinde elde edilmiştir. En düşük doğruluk oranı %50 benzerlik oranına sahip veri kümesinde elde edilmiştir (Tablo 5.5).

Seviye 3 veri kümeleri %50, %75 ve %90 benzerlik oranlarına göre ayrılarak Random Forest sınıflandırma yöntemi ile sınıflandırılmıştır. Bu sınıflandırma sonunda en yüksek doğruluk oranı %50 benzerlik oranı kullanılan veri kümesinde elde edilmiştir. En düşük doğruluk oranı %75 benzerlik oranına sahip veri kümesinde elde edilmiştir (Tablo 5.6).

Random Forest sınıflandırma yöntemi ile yapılan sınıflandırmalar doğruluk oranları benzerlik oranları açısından incelendiğinde aralarında direk bir ilişki bulunmamaktadır. Random Forest sınıflandırma yöntemi ile yapılan sınıflandırmalarda farklı veri kümelerinde farklı doğruluk oranları ile karşılaşılmıştır. Veri kümesinden bağımsız olarak en iyi benzerlik oranı bulunmamaktadır.

Benzerlik oranları sabit tutulup Seviye 1, Seviye 2 ve Seviye 3 veri kümelerinde elde edilen doğruluk oranları karşılaştırıldığında veri kümeleri ile doğruluk oranları arasında bir ilişki tespit edilmemiştir (Tablo 5.4, Tablo 5.5, Tablo 5.6).

Benzerlik oranları sabit tutularaktan veri kümesinin sahip olduğu sınıf sayısına bakıldığında en az sınıf sayısı olan Seviye 2 veri kümesinde Random Forest yöntemi en düşük doğruluk oranına sahip olmuştur (Tablo 5.4, Tablo 5.5, Tablo 5.6). Seviye 1 ve Seviye 3 veri kümelerinde daha başarılı sınıflandırma sonuçları tespit edilmiştir.

Random Forest sınıflandırma yöntemi ile yapılan sınıflandırmaların duyarlılık oranları benzerlik oranları açısından incelendiğinde aralarında direk bir ilişki bulunmamaktadır. Random Forest sınıflandırma yöntemi ile yapılan sınıflandırmalarda farklı veri kümelerinde farklı duyarlılık oranları ile karşılaşılmıştır. Veri kümesinden bağımsız olarak en iyi benzerlik oranı bulunmamaktadır.

Benzerlik oranları sabit tutulup Seviye 1, Seviye 2 ve Seviye 3 veri kümelerinde elde edilen duyarlılık oranları karşılaştırıldığında aralarında bir ilişki tespit edilememiştir (Tablo 5.7, Tablo 5.8, Tablo 5.9).

Random Forest sınıflandırma yöntemi ile yapılan sınıflandırmaların kesinlik oranları benzerlik oranları açısından incelenememektedir. Random Forest algoritması ile yapılan sınıflandırmada kesinlik oranları Seviye 1 veri kümelerinde hesaplanamamıştır (Tablo 5.10). Bu konu "5.4 LSTM ve Geleneksel Makine Öğrenimi Yöntemleri ile Elde Edilen Sonuçlar Kesinlik (Precision) Değerleri" bölümünde ayrıntılı açıklanmıştır.

Random Forest sınıflandırma yöntemi ile yapılan sınıflandırmaların F1 skoru oranları benzerlik oranları açısından incelenememektedir. Random Forest algoritması ile hesaplanamayan kesinlik oranları F1 skoru hesaplanmasını da engellemektedir.

6.4. LSTM Sonuçları

Seviye 1 veri kümeleri %50, %75 ve %90 benzerlik oranlarına göre ayrılarak LSTM sınıflandırma yöntemi ile sınıflandırılmıştır. Bu sınıflandırma sonunda en yüksek doğruluk oranı %75 benzerlik oranı kullanılan veri kümesinde elde edilmiştir. En düşük doğruluk oranı %50 benzerlik oranına sahip veri kümesinde elde edilmiştir (Tablo 5.4).

Seviye 2 veri kümeleri %50, %75 ve %90 benzerlik oranlarına göre ayrılarak LSTM sınıflandırma yöntemi ile sınıflandırılmıştır. Bu sınıflandırmada en yüksek doğruluk oranı %75

ve %90 benzerlik oranları kullanılan veri kümelerinde elde edilmiştir. En düşük doğruluk oranı %50 benzerlik oranına sahip veri kümesinde elde edilmiştir (Tablo 5.5).

Seviye 3 veri kümeleri %50, %75 ve %90 benzerlik oranlarına göre ayrılarak LSTM sınıflandırma yöntemi ile sınıflandırılmıştır. Bu sınıflandırma sonunda en yüksek doğruluk oranı %90 benzerlik oranı kullanılan veri kümesinde elde edilmiştir. En düşük doğruluk oranı %50 benzerlik oranına sahip veri kümesinde elde edilmiştir (Tablo 5.6).

LSTM sınıflandırma yöntemi ile yapılan sınıflandırmalar doğruluk oranları benzerlik oranları açısından incelendiğinde aralarında direk bir ilişki bulunmaktadır. LSTM sınıflandırma yönteminde eşik değeri %50 olduğu sınıflandırmalarda en düşük doğruluk oranlarını vermektedir (Tablo 6, Tablo 7, Tablo 8). Eşik değerinin düşmesi kullanılacak verideki kaybı arttırmıştır. Bu sebepten LSTM yöntemi en düşük başarıyı %50 benzerlik oranında vermektedir.

Benzerlik oranları sabit tutularaktan veri kümelerinde olan sınıf sayılarına göre değerlendirildiğinde LSTM yöntemi en az farklı sınıf olan Seviye 2 veri kümelerinde en başarılı sonuçları almıştır. Seviye 1 ve Seviye 3 veri kümelerinde daha düşük doğruluk oranları elde edilmiştir (Tablo 5.4, Tablo 5.5, Tablo 5.6).

LSTM sınıflandırma yöntemi ile yapılan sınıflandırmaların duyarlılık oranları benzerlik oranları açısından incelendiğinde en düşük duyarlılık oranları %50 eşik değerinde elde edilmiştir (Tablo 5.7, Tablo 5.8, Tablo 5.9). Eşik değerinin düşmesi kullanılacak verideki kaybı arttırmıştır. Bu sebepten LSTM yöntemi en düşük başarıyı %50 benzerlik oranında vermektedir.

Benzerlik oranları sabit tutulup Seviye 1, Seviye 2 ve Seviye 3 veri kümelerinde elde edilen duyarlılık oranları karşılaştırıldığında en yüksek duyarlılık oranı en az sınıfa sahip olan Seviye 2 veri kümesinde tespit edilemiştir (Tablo 5.7, Tablo 5.8, Tablo 5.9).).

LSTM sınıflandırma yöntemi ile yapılan sınıflandırmaların kesinlik oranları benzerlik oranları açısından incelendiğinde en düşük kesinlik oranları %50 eşik değerinin olduğu sınıflandırmalarda görülmüştür (Tablo 5.10, Tablo 5.11, Tablo 5.12). Eşik değerinin düşmesi kullanılacak verideki kaybı arttırmıştır. Bu sebepten LSTM yöntemi en düşük başarıyı %50 benzerlik oranında vermektedir.

Benzerlik oranları sabit tutulup Seviye 1, Seviye 2 ve Seviye 3 veri kümelerinde elde edilen kesinlik oranları karşılaştırıldığında en yüksek kesinlik oranları en az sınıfa sahip Seviye 2 veri kümelerinde tespit edilmiştir (Tablo 5.10, Tablo 5.11, Tablo 5.12).

LSTM sınıflandırma yöntemi ile yapılan sınıflandırmaların F1 skoru oranları benzerlik oranları açısından incelendiğinde en düşük F1 skoru oranları %50 eşik değeri olan sınıflandırmalarda görülmüştür (Tablo 5.13, Tablo 5.14, Tablo 5.15). Eşik değerinin düşmesi kullanılacak verideki kaybı arttırmıştır. Bu sebepten LSTM yöntemi en düşük F1 skorunu %50 benzerlik oranında vermektedir.

Benzerlik oranları sabit tutulup Seviye 1, Seviye 2 ve Seviye 3 veri kümelerinde elde edilen F1 skoru oranları karşılaştırıldığında en yüksek F1 skorunun en az sınıfa sahip Seviye 2 veri kümelerinde olduğu görülmüştür (Tablo 5.13, Tablo 5.14, Tablo 5.15).

6.5. SVM, Naive Bayes, Random Forest ve LSTM Sonuçları

Frekans vektörlerinin kullanıldığı geleneksel makine öğrenimi yöntemleri arasında veri kümelerinin tümü göze alınarak yapılan aritmetik ortalama değerlendirmesinde en yüksek başarı oranı %80,11 doğruluk ile Naive Bayes yönteminde elde edilmiştir (Tablo 18).

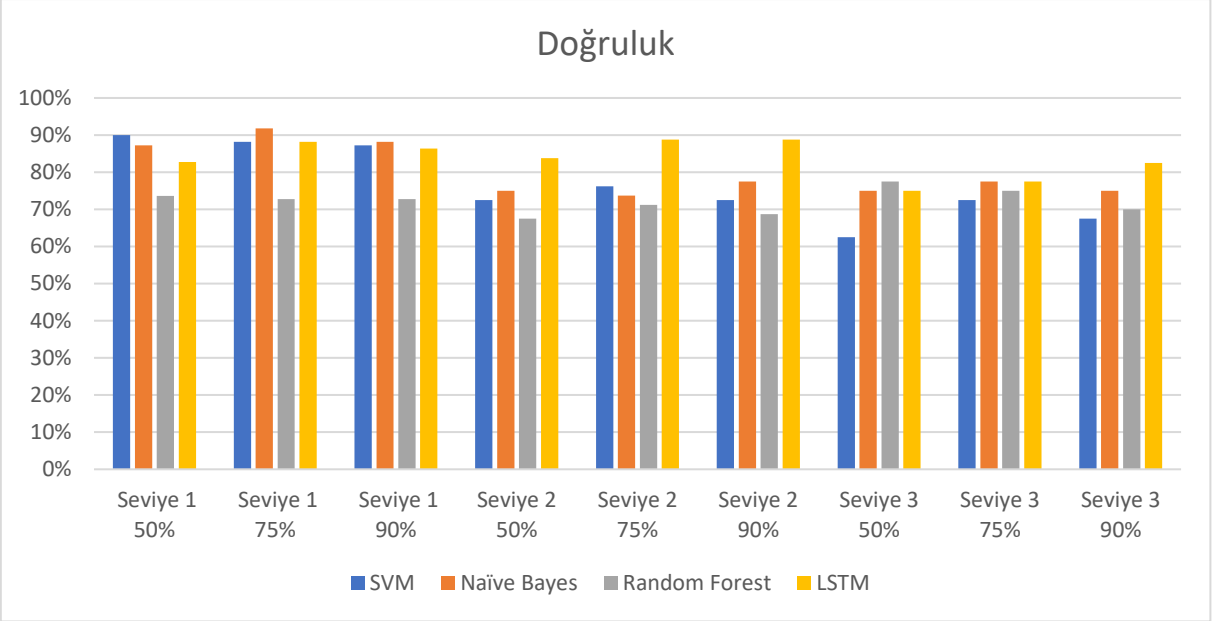
Derin öğrenme yöntemi olan veri sıralama vetörlerinin kullanıldığı LSTM için de aynı aritmetik ortalama hesaplanmıştır. LSTM ortalaması %83,72 ile 4 sınıflandırma algoritması içerisinde en yüksek değerdedir (Tablo 5.16).

Tablo 5.16 Sınıflandırma yöntemlerinde elde edilen doğruluk oranlarının aritmetik ortalaması

SVM	Naïve Bayes	Random Forest	LSTM
76,58%	80,11%	72,12%	83,72%

Doğruluk açısından dört sınıflandırma yöntemi karşılaştırıldığında Support Vector Machine ve Naive Bayes yöntemlerinin Seviye 1 veri kümelerinden eşik değeri %50 olan sınıflandırmada daha başarılıdırlar.

LSTM yöntemi %50 eşik değeri kullanılan sınıflandırmalarda daha az başarılı olmasına karşın %75 ve %90 eşik değerlerinin kullanıldığı veri kümelerinde daha başarılıdır. LSTM yöntemi daha az sınıfın bulunduğu Seviye 2 sınıflandırmalarında geleneksel makine öğrenim yöntemlerinden daha yüksek kesinlik oranına sahip olmuştur (Şekil 6.1).

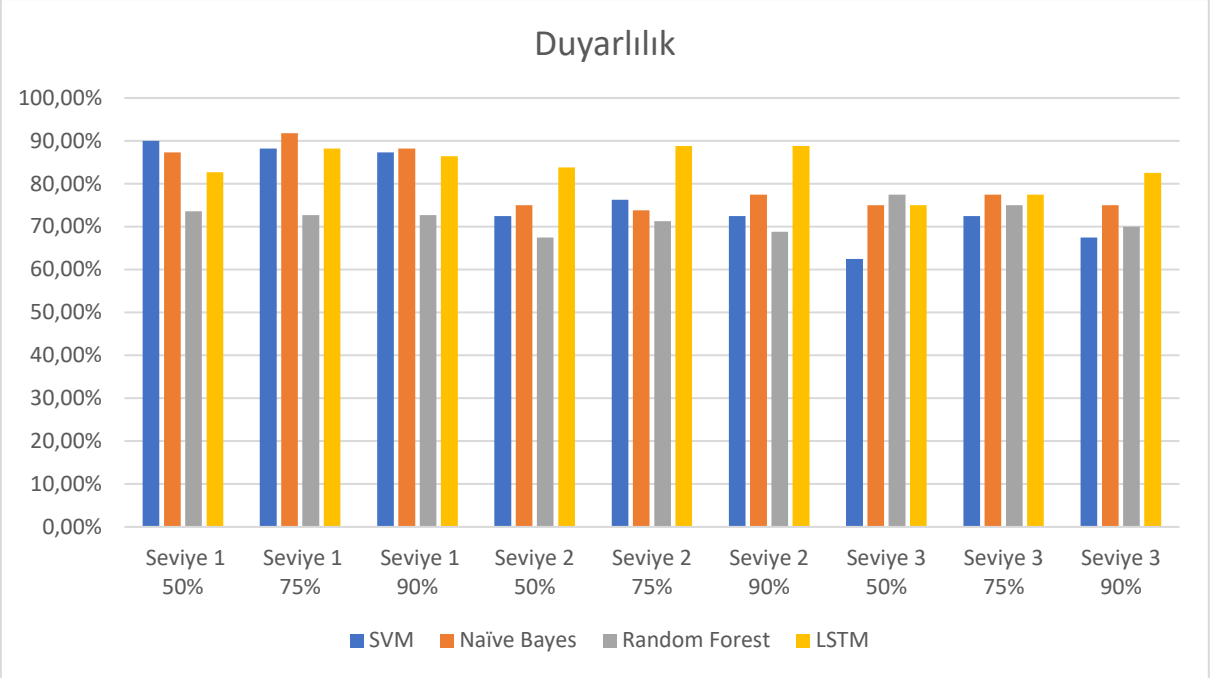


Şekil 6.1. Sınıflandırma yöntemleri için doğruluk oranları

Bu tez kapsamında Seviye 1, Seviye 2 ve Seviye 3 olan veri kümelerinden benzerlik oranı eşik değerlerine göre toplam dokuz adet veri kümesi elde edilmiştir. Seviye 1 ve Seviye 2 veri kümelerinde sınıflara ait eleman sayıları aynı değildir. Bu sebepten bu dört yöntemi karşılaştırırken duyarlılık, kesinlik ve F1 skoru değerlerimiz özelikle dengeli olmayan veri kümelerinde çok önemlidir.

Duyarlılık açısından dört sınıflandırma yöntemi karşılaştırıldığında Support Vector Machine ve Naive Bayes yöntemlerinin Seviye 1 veri kümelerinden eşik değeri %50 olan sınıflandırmada daha başarılı oldukları görülmektedir.

LSTM yöntemi %50 eşik değeri kullanılan sınıflandırmalarda daha az başarılı olmasına karşın %75 ve %90 eşik değerlerinin kullanıldığı veri kümelerinde daha başarılıdır. LSTM yöntemi daha az sınıfın bulunduğu Seviye 2 sınıflandırmalarında geleneksel makine öğrenim yöntemlerinden daha yüksek duyarlılık oranına sahip olmuştur (Şekil 6.2).

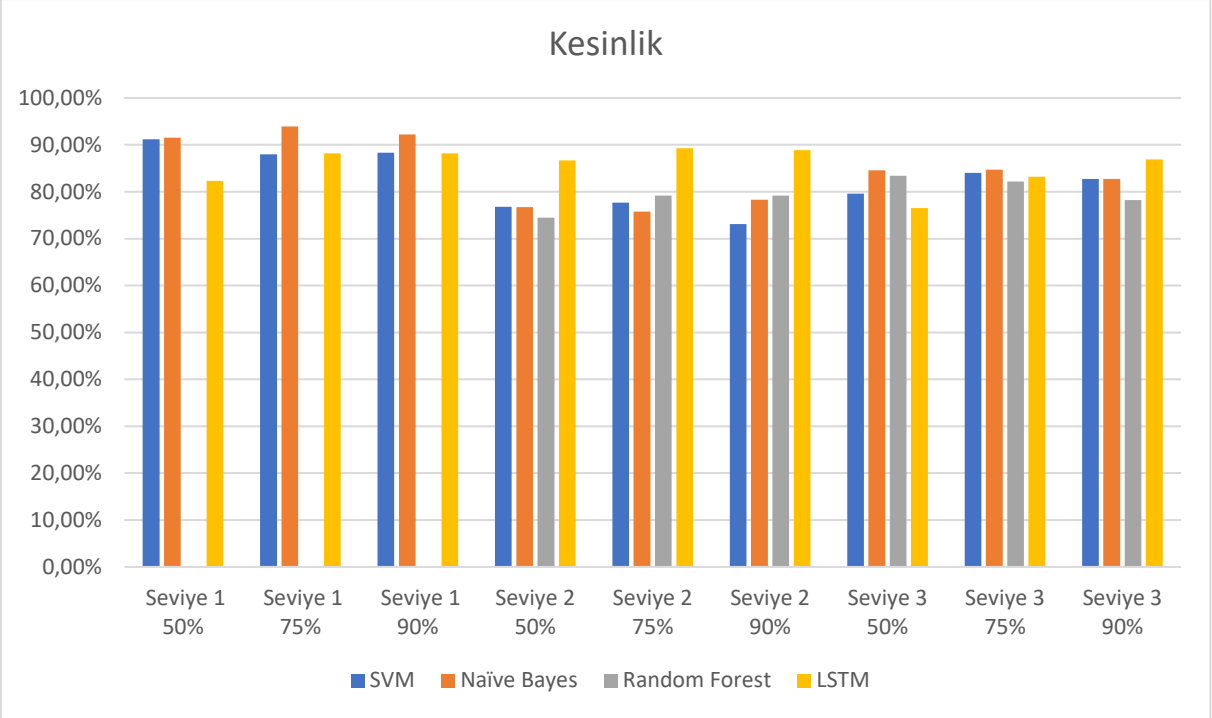


Şekil 6.2. Sınıflandırma yöntemleri için duyarlılık oranları

Kesinlik açısından dört sınıflandırma yöntemi karşılaştırıldığında Support Vector Machine ve Naive Bayes yöntemlerinin Seviye 1 veri kümelerinden eşik değeri %50 olan sınıflandırmada daha başarılı oldukları görülmektedir.

Random Forest yöntemi için en az dengeli veri kümelerinin bulunduğu Seviye 1' de kesinlik hesabı yapılamamıştır. Bu konu "5.4 LSTM ve Geleneksel Makine Öğrenimi Yöntemleri ile Elde Edilen Sonuçlar Kesinlik (Precision) Değerleri" bölümünde ayrıntılı olarak anlatılmıştır.

LSTM yöntemi %50 eşik değeri kullanılan sınıflandırmalarda daha az başarılı olmasına karşın %75 ve %90 eşik değerlerinin kullanıldığı veri kümelerinde daha başarılıdır. LSTM yöntemi daha az sınıfın bulunduğu Seviye 2 sınıflandırmalarında geleneksel makine öğrenim yöntemlerinden daha yüksek kesinlik oranına sahip olmuştur (Şekil 6.3).

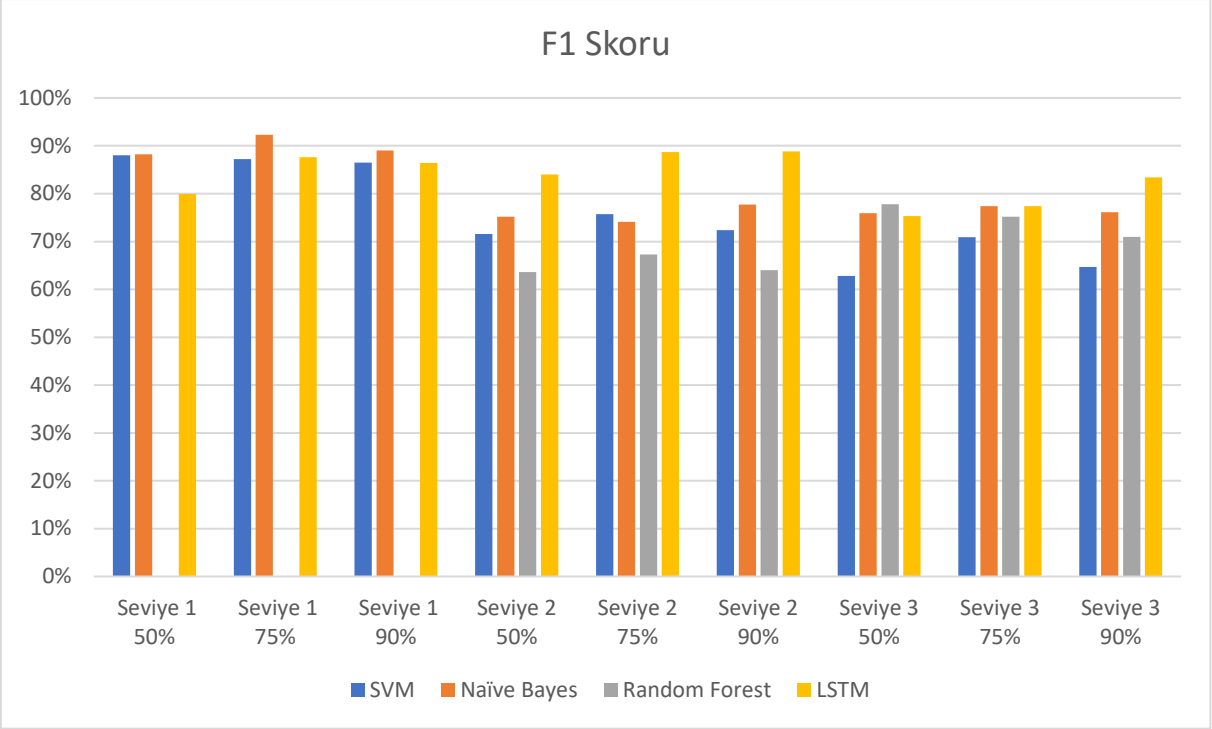


Şekil 6.3. Sınıflandırma yöntemleri için kesinlik oranları

F1 skoru açısından dört sınıflandırma yöntemi karşılaştırıldığında Support Vector Machine ve Naive Bayes yöntemlerinin Seviye 1 veri kümelerinden eşik değeri %50 olan sınıflandırmada daha başarılı oldukları görülmektedir.

Random Forest yöntemi için en az dengeli veri kümelerinin bulunduğu Seviye 1' de kesinlik hesabı yapılamamıştır. Bu sebepten F1 skoru da hesaplanamamıştır. Kesinlik hesabının yapılamayışının sebebi "5.4 LSTM ve Geleneksel Makine Öğrenimi Yöntemleri ile Elde Edilen Sonuçlar Kesinlik (Precision) Değerleri" bölümünde ayrıntılı olarak anlatılmıştır.

LSTM yöntemi %50 eşik değeri kullanılan sınıflandırmalarda daha az başarılı olmasına karşın %75 ve %90 eşik değerlerinin kullanıldığı veri kümelerinde daha başarılıdır. LSTM yöntemi daha az sınıfın bulunduğu Seviye 2 sınıflandırmalarında geleneksel makine öğrenim yöntemlerinden daha yüksek kesinlik oranına sahip olmuştur (Şekil 6.4).



Şekil 6.4. Sınıflandırma yöntemleri için F1 skoru oranları

Geleneksel sınıflandırma yöntemlerinin F1 skoru, kesinlik, duyarlılık ve doğruluk sonuçlarının aritmetik ortalamaları alınarak karşılaştırıldığında Naive Bayes en başarılı geleneksel makine öğrenim yöntemidir.

Derin öğrenmem yöntemi olan LSTM F1 skoru, kesinlik, duyarlılık ve doğruluk sonuçlarının aritmetik ortalamaları açısından geleneksel makine öğrenim yöntemleri ile karşılaştırıldığında tüm ölçüm çeşitlerinde en başarılı sonuçları vermektedir (Tablo 5.17).

	SVM	**Naïve Bayes**	**Random Forest**	**LSTM**
F1 Skoru	75,53%	80,66%	69,82%	83,50%
Kesinlik	82,38%	84,49%	79,45%	85,58%
Duyarlılık	76,59%	80,12%	72,12%	83,74%
Doğruluk	76,58%	80,11%	72,12%	83,72%

Tablo 5.17. Sınıflandırma yöntemleri başarı oranlarının aritmetik ortalamaları

7. SONUÇ VE GELECEK ÇALIŞMALAR

Çevrimiçi öğrenme yöntemleri tüm dünyada internetin ve içerik sağlayıcıların artması ile hızla artmaktadır. Tüm dünyada şu an salgın bir hastalık olan Koronavirüs (Covid 19) hastalığı tüm iş sektörlerini etkilediği gibi eğitim sektörünü de çevrimiçi öğrenme açısından etkilemiştir. Bu etki çevrimiçi öğrenmenin çok hızlı bir şekilde artacağını göstermektedir.

Ders videolarının doğan ihtiyaçtan dolayı çok hızlı artması öğrencilerin istedikleri videoyu bulmalarını çok zorlaştırmaktadır. Bu tezde, bu ihtiyaca yönelik olarak üç ayrı geleneksel makine öğrenme yöntemi ve bir derin öğrenme yöntemi kullanılmıştır.

Bu tez çalışmasında üç ayrı seviyede veri kümesi bulunmaktadır. Bu üç ayrı seviyedeki veri kümelerinden OCR ile metinsel veriler çıkarılmıştır. Çıkarılan veriler üç ayrı benzerlik oranı eşik değerine göre önişlemlere tutulmuştur. Üç ayrı benzerlik oranı eşik değerinin kullanılması toplamda dokuz ayrı veri kümesi elde edilmesini sağlamıştır. Veri kümelerinde kullanılan algoritmalar veri kümelerine bağımlı olarak farklı doğruluk oranları vermektedir.

Elde edilen dokuz ayrı veri kümesinden Seviye 1 ve Seviye 2 veri kümeleri daha az dengeli veri kümeleridir. Naive Bayes ve Support Vector Machine yöntemleri doğruluk, duyarlılık, kesinlik ve F1 skoru oranlarına bakıldığında %50 benzerlik oranı eşik değerine sahip Seviye 1 veri kümesinde daha başarılı olmuşlardır.

Random Forest algoritması kesinlik ve F1 skoru oranlarına bakıldığında Seviye 1 veri kümelerinde hesaplanamaz durumdadır. Seviye 1 veri kümesi dört sınıftan oluşmaktadır ve bu seviyedeki kliplerin %76' sı tek bir sınıfa aittir. Seviye 1 veri kümesinde Random Forest algoritması kesinlik ve F1 skoru açısından ölçülebilecek bir sınıflandırma yapamamıştır.

Long Short Term Memory yöntemi doğruluk, duyarlılık, kesinlik ve F1 skoru oranları açısından değerlendirildiğinde en az başarı %50 benzerlik oranı eşik değerinin kullanıldığı veri kümelerinde gözlemlenmiştir. Bunun sebebi eşik değerinin düşmesi ile kaybedilmiş olan metinsel verilerdir.

Naive Bayes yöntemi geleneksel makine öğrenimi yöntemleri ile doğruluk, duyarlılık, kesinlik ve F1 skoru oranları açısından elde edilen sonuçların aritmetik ortalamaları alınarak yapılan değerlendirmede en başarılı geleneksel makine öğrenme yöntemi olmaktadır.

Long Short Term Memory yöntemi dört makine öğrenimi yöntemi arasında doğruluk, duyarlılık, kesinlik ve F1 skoru oranları açısından elde edilen sonuçların aritmetik ortalamaları alınarak yapaılan değerlendirmede en başarılı sınıflandırma yöntemi olmuştur. Tüm ölçüm yöntemlerinin aritmetik ortalamaları alınarak yapılan karşılaştırmalarda Long Short Term Memory yöntemi daha başarılı sonuçlar vermiştir.

Gelecekteki çalışmalarda videolardan elde edilen metinlerin sınıflandırılması için bir diğer derin örenme yöntemi olan CNN kullanılabilir. Gelecekteki çalışmalarda LSTM sınıflandırma yöntemi videoların sınıflandırılması için ASR teknolojisi ile çıkarılmış veriler ile kullanılabilir. Gelecekteki çalışmalarda LSTM sınıflandırmaları için OCR' den ve ASR 'den gelen veriler birlikte kullanılabilir. İşitsel ve metinsel içeriklerin birlikte kullanıldığı bir sınıflandırma daha iyi sonuçlar verebilir.

KAYNAKLAR

[1] Robert Connor Chick, Guy Travis Clifton, Kaitlin M. Peace, Brandon W. Propper, Diane F. Hale, Adnan A. Alseidi, and Timothy J. Vreeland, "Using Technology to Maintain the Education of Residents During the COVID-19 Pandemic" Journal of Surgical Education Volume 00 /Number 00 / Month 2020

[2] Sandeep Krishnamurthy, "The future of business education: A commentary in the shadow of the Covid-19 pandemic", Journal of Business Research 117 (2020) 1–5

[3] Cathy Mae Toquero, "Challenges and Opportunities for Higher Education amid the COVID-19 Pandemic: The Philippine Context", Pedagogical Research 2020, 5(4), em0063, e-ISSN: 2468-4929

[4] Dipesh Chand and Hasan Oğul, "Content-Based Search in Lecture Video: A Systematic Literature Review", 2020 3rd International Conference on Informtation and Computer Technologies (ICICT)

[5] B. S. Daga, Dr A. A. Ghatol and V.M.Thakare, "Semantic Enriched Lecture Video Retrieval System Using Feature Mixture and Hybrid Classification" Society For Scıence And Educatıon, Volume 5, Issue 3, ISSN 2054-7412

[6] Adnan Yazici, Murat Koyuncu, Turgay Yilmaz, Saeid Sattari, Mustafa Sert and Elvan Gulen, "An intelligent multimedia information system for multimodal content extraction and querying", Springer Science+Business Media New York 2017, Multimed Tools Appl (2018) 77:2225–2260 DOI 10.1007/s11042-017-4378-6

[7] Stefano Masneri and Oliver Schreer, "SVM-based Video Segmentation and Annotation of Lectures and Conferences", Image Processing Department, Fraunhofer Heinrich Hertz Institut, Einsteinufer 37, 10587 Berlin, Germany

[8] John Adcock, Matthew Cooper, Laurent Denoue, Hamed Pirsiavash, Lawrence A. Rowe, "TalkMiner: A Lecture Webcast Search Engine", MM'10, October 25–29, 2010, Firenze, Italy. Copyright 2010 ACM 978-1-60558-933-6/10/10

[9] Stephan Repp, Andreas Groß, and Christoph Meinel, Member, IEEE, “Browsing within Lecture Videos Based on the Chain Index of Speech Transcription”, IEEE Transactions On Learning Technologies, Vol. 1, No. 3, July-September 2008

[10] Haojin Yang and Christoph Meinel, Member, IEEE “Content Based Lecture Video Retrieval Using Speech and Video Text Information”, IEEE Transactions On Learning Technologies, Vol. 7, No. 2, April-June 2014

[11] N. Radha, “Video Retrieval Using Speech and Text InVideo”, Conference: 2016 International Conference on Inventive Computation Technologies (ICICT)

[12] A. Özdarıcı Ok, Ö. Akar, O. Güngör, “Rastgele Orman Siniflandirma Yöntemi Yardimiyla Tarim Alanlarindaki Ürün Çeşitliliğinin Siniflandirilmasi”, Conference: TUFUAB 2011 VI. Teknik SempozyumuAt: Antalya

[13] Gang Liu, Jiabao Guo, “Bidirectional LSTM with attention mechanism and convolutional layer for text classification”, https://doi.org/10.1016/j.neucom.2019.01.078, Neurocomputing 337 (2019) 325–338, 0925-2312/©2019ElsevierB.V.

[14] Alexader Haubold and John R. Kender, “Augmented Segmentation and Visualization for Presentation Videos”, MM’05, November 6–11, 2005, Singapore. Copyright 2005 ACM 1-59593-044-2/05/0011

[15] Lakshmi Haritha Medida and Kasarapu Ramani, “An Optimized E-Lecture Video Retrieval based on Machine Learning Classification”, International Journal of Engineering and Advanced Technology (IJEAT), ISSN: 2249–8958, Volume-8, Issue-6, August 2019

[16] Asad Abdia, Siti Mariyam Shamsuddina, Shafaatunnur Hasana and Jalil Piranb, “Deep learning-based sentiment classification of evaluative text based on Multi-feature fusion”, https://doi.org/10.1016/j.ipm.2019.02.018, Information Processing and Management 56 (2019) 1245–1259

[17] Chunting Zhou, Chonglin Sun, Zhiyuan Liu and Francis C.M. Lau, “A C-LSTM Neural Network for Text Classification”, https://arxiv.org/abs/1511.08630v2

[18] Vijaykumar B., Vikramkumar, Trilochan, “Bayes and Naive-Bayes Classifier”, Computer Science & Engineering, Rajiv Gandhi University of Knowledge Technologies Andhra Pradesh, India, https://arxiv.org/abs/1511.08630v2

[19] Corinna Cortes, Vladimir Vapnik “Support-Vector Networks”, Machine Leaming, 20, 273-297 (1995) , Kluwer Academic Publishers, Boston.

[20] Yangchang Zhao “Chapter 4 - Decision Trees and Random Forest”, R and Data Mining, Examples and Case Studies 2013, Pages 27-40.

[21] Filip Zelic, Anuj Sableangchang “A comprehensive guide to OCR with Tesseract, OpenCV and Python”, https://nanonets.com/blog/ocr-with-tesseract/

[22] “Weka (machine learning)”, https://en.wikipedia.org/wiki/Weka_(machine_learning)

[23] John W. Ratcliff and David Metzener, “Pattern Matching: The Gestalt Approach”, Dr. Dobb's Journal, page 46, July 1988.

[24] “Gestalt Pattern Matching”, https://en.wikipedia.org/wiki/Gestalt_Pattern_Matching#cite_note-PY21-1.

[25] “Text data preprocessing”, https://keras.io/api/preprocessing/text/.

[26] “tf.keras.preprocessing.text.Tokenizer”, https://www.tensorflow.org/api_docs/python/tf/keras/preprocessing/text/Tokenizer.