T.C. İSTANBUL TİCARET
ÜNİVERSİTESİ

FEN BİLİMLERİ ENSTİTÜSÜ

KONU MODELLEME YÖNTEMLERİNİN BELGE SINIFLANDIRMA ÜZERİNE KULLANIMI

Süleyman ÖZDEMİRCİ

Danışman
Dr. Öğr. Üyesi Metin TURAN

YÜKSEK LİSANS TEZİ
BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI
İSTANBUL - 2021

KABUL VE ONAY SAYFASI

Süleyman Özdemirci tarafından hazırlanan **"Konu Modelleme Yöntemlerinin Belge Sınıflandırma Üzerine Uygulanması"** adlı tez çalışması 0505/0707/20212021 tarihinde aşağıdaki jüri üyeleri önünde başarı ile savunularak, İstanbul Ticaret Üniversitesi Fen Bilimleri Enstitüsü **Bilgisayar Mühendisliği Anabilim Dalı**'nda **YÜKSEK LİSANS TEZİ** olarak kabul edilmiştir.

Danışman	**Dr.Öğr.Üyesi Metin TURAN** İstanbul TicaretTicaret Üniversitesi	.....
Jüri Üyesi	**Dr. Öğr. Üyesi Murat ORHUN** İstanbul Bilgi Üniversitesi	.....
Jüri Üyesi	**Dr. Öğr. Üyesi ORHUNFeyza Merve HAFIZOĞLU** İstanbul BilgiTicaret Üniversitesi	.....

Onay Tarihi : 29.07.2021

İstanbul Ticaret Üniversitesi, Fen Bilimleri Enstitüsünün 29.07.2021 tarih ve 2021/317 numaralı Yönetim Kurulu Kararının 2. maddesi gereğince, ders yüklerini ve tez yükümlülüğünü yerine getirdiği belirlenen "Süleyman ÖZDEMİRCİ" adlı öğrencinin mezun olmasına oy birliği ile karar verilmiştir.

Prof. Dr. Necip ŞİMŞEK
Enstitü Müdürü

AKADEMİK VE ETİK KURALLARA UYGUNLUK BEYANI

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

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

beyan ederim.

20.07.2021

Süleyman ÖZDEMİRCİ

İÇİNDEKİLER

ÖZET

Yüksek Lisans Tezi

KONU MODELLEME YÖNTEMLERİNİN BELGE SINIFLANDIRMA ÜZERİNE UYGULANMASI

Süleyman ÖZDEMİRCİ

İstanbul Ticaret Üniversitesi
Fen Bilimleri Enstitüsü
Bilgisayar Mühendisliği Anabilim Dalı

Danışman: Dr.Öğr.Üyesi Metin TURAN
2021, 30 sayfa

Konu modellemenin doküman sınıflandırma, konu kümeleme, belge etiketleme, geniş belge koleksiyonlarında özellik çıkarma gibi çok sayıda uygulaması vardır. Bu çalışmada, Latent Dirichlet Allocation konu modellemenin pratik keşif yöntemi, Bidirectional Encoder Representations from Transformers ve Terim Frekansı - Ters Belge Frekansı yöntemi deneysel belge setine ayrı ayrı uygulanmıştır. Bu veri seti lisansüstü öğrenciler tarafından internetten toplanan toplam 801 adet spor ve eğitim makalelerini içermektedir. Bu çalışmanın amacı, konu modellemesine hangi yöntemin en uygun olduğunu gözlemlemek ve mümkünse bu yöntemler topluluğu ile doğruluk oranını arttırmaktır. Çalışmada, en iyi yöntemlerin güçlü özellikleri birleştirilerek yeni bir yöntem önerilip önerilemeyeceği gibi soruların cevapları aranmıştır. Bu çalışma sonucunda BERT'in bazı dezavantajları olsa da doğru konuya sahip belgeleri ortalama %92.6 başarı oranıyla sınıflandırdığı, diğer yöntemlerden daha başarılı olduğu görülmüştür.

Anahtar Kelimeler: BERT, Konu Modelleme, LDA, Sınıflandırma, TF-IDF.

ABSTRACT

M.Sc. Thesis

KONU MODELLEME YÖNTEMLERİNİN BELGE SINIFLANDIRMA ÜZERİNE UYGULANMASI

Süleyman ÖZDEMİRCİ

Istanbul Commerce University
Graduate School of Applied and Natural Sciences
Department of Computer Engineering

Supervisor: Assist. Prof. Dr. Metin TURAN
2021, 30 pages

Topic modeling has numerous applications like text categorization, topic clustering, document tagging, feature extraction on wide document collections. In this study, practical exploration method of topic modeling of Latent Dirichlet Allocation, transformers based machine learning method Bidirectional Encoder Representations from Transformers and Term Frequency — Inverse Document Frequency method were applied to the experimental document set separately. It includes sport and education articles collected from internet by graduate students, 801 number totally. The purpose of this study is to observe which method best suits to the topic modeling and if it is possible, increase the accuracy rate via the combination of these methods. In the study, the answers to the questions such as whether a new method can be proposed by combining the strong features of the best methods were sought. Although BERT has some disadvantages, it was observed that classifying the documents with the correct topic was achieved with a %92.6 success rate, overwhelming the other methods.

Keyword: BERT, Classification, LDA, TF-IDF, topic modeling,

TEŞEKKÜR

Bu araştırma için beni yönlendiren, karşılaştığım zorlukları bilgi ve tecrübesi ile aşmamda yardımcı olan değerli Danışman Hocam Dr. Öğr. Üyesi, Metin TURAN'a teşekkürlerimi sunarım.

Tezimin her aşamasında beni yalnız bırakmayan başta eşim Tuğçe Ceren ÖZDEMİRCİ ve aileme sonsuz sevgi ve saygılarımı sunarım.

Süleyman ÖZDEMİRCİ
İSTANBUL, 2021

ŞEKİLLER

ÇİZELGELER

SİMGELER VE KISALTMALAR

ALBERT	A Lite BERT for Self-supervised Learning of Language Represention s
BERT	Bidirectional Encoder Representations from Transformers
BOW	Bag of Words
DDİ	Doğal Dil İşleme
IDF	Inverse documentFrequency
LDA	Latent Dirichlet Allocation
RoBERTa	Robustly Optimized BERT Pretraining Approach
SCiBERT	A Pretrained Language Model for Scientific Text
SOM	Self-organizing maps
TF	Terim Frekansı

1. GİRİŞ

Günümüzde neredeyse tüm bilgiler dijitalleşme yolunda veya elektronik olarak bir veri tabanında saklanmaktadır. Bu veriler her yıl çok hızlı büyükmekte ve 2025 yılı gelindiğinde 160 zettabaytı geçeceği öngürülmektedir (Eravcı, 2020). Veritabanlarındaki bu büyük miktarda veriden, kimi zaman bir makale için literatür taraması yapılması, kimi zaman şirketler için bir işin geliştirilmesi safhasında araştırma yapılması veya öğrencilerin bilgiye ulaşmak için kaynak araştırması yapmaları gerekmektedir. Büyük veri ve doküman yığınlarından, insanların istedikleri bilgileri elde etmelerini kolaylaştırmak için bu verilerin kategorize edilmesi kullanışlı olacaktır. Bu verileri analiz etmek isteyen bu kişiler, sınıflandırma, bilgi çıkarma ve ilgili belgeleri kümeleme için araçlara ihtiyaç duymaktadır. Bu araçların geliştirilmesi Doğal Dil İşleme (DDİ) ve Metin Madenciliği (MM) gibi bilim alanlarının kullanımı gerekli kılmaktadır. Bu amaçla, özellikle DDİ konu modelleme yöntemlerinin kullanımı önemli bir yer tutmaktadır.

Konu modelleme, bu alandaki uygulamaları geliştirmek için çalışılan bir araştırma alanıdır. Sınıflandırma, özetleme, özellik çıkarma ve benzerlikler gibi gerekli görevler için belgelerin temel istatistiksel ilişkilerini bulmayı amaçlamaktadır (Wu vd., 2017). Bu strateji, gizli anlambilim yapılarını ve metin külliyatından anlamları belirlemek için kullanılır (Zhang vd., 2020). Bu çalışmada konu modelleme yöntemlerinin sınıflandırma üzerine uygulanmasına yönelik çalışmalar yürütülmüştür.

Konu modellemede kullanılan birçok yöntem vardır. Bunlardan en önemlileri olan Latent Dirichlet Allocation (LDA), Terim Frekansı(TF)-Ters Doküman Sıklığı(IDF) ve cümle gömme yöntemi olan Bidirectional Encoder Representations from Transformers (BERT) bu çalışmada kullanılmıştır. Her yöntem, konuları bulmak için farklı bir yol izler ve herbirinin kendi güçlü ve zayıf noktaları vardır. Bu tez çalışmasının amacı, belge sınıflandırması için bu yöntemleri karşılaştırmak ve bu yöntemlerin daha iyi sonuçlar elde etmek için birlikte kullanılıp kullanılamayacağının araştırılmasıdır.

Çalışmada ilk olarak, DDİ alanında kullanılan en yaygın yöntem olan TF-IDF ile veri seti üzerinde çalışılmıştır. TF-IDF'in bize sunduğu en önemli bakış açısı, bir belgede en sık kullanılan kelimelerin istatiksel karşılığını bize vermesidir. Bu, belgenin sınıflandırılması için bize faydalı bilgiler sunabilmektedir ancak sadece kelimenin frekansı penceresinden bakmak bize her zaman istediğimiz doğrulukta sonuçlar vermediği gibi bazı durumlarda bizi yanıltabilmektedir. Örneğin spor akademilerindeki eğitim hakkında bilgi veren bir doküman içerisinde çok sayıda spor ile ilgili kelime geçebilmektedir. TF-IDF'in bu dokümanla ilgili bize vereceği bilgi, bu dokümanın bir eğitim dokümanı değil spor ile ilgili bir doküman olduğu yönünde olacaktır ve çalışmamız için yanlış bir yönlendirme yapmış olacaktır.

Çalışmada bir diğer yöntem olarak LDA incelenmiştir. LDA, her belgenin bir konu koleksiyonu olarak kabul edildiği ve belgedeki her kelimenin konulardan birine karşılık geldiği bir modelleme yöntemidir (Blei vd., 2003). Burada, TF-IDF'den farklı olarak, belgenin sınıfını belirten bir dizi kelime elde edilmiştir. LDA bize dizide bulunan her kelimenin konuya yakınlığını gösteren sayısal bir değer vermektedir. Böylelikle TF-IDF'den daha fazla bakış açısına sahip olunmuştur. Bu bakış açısı sayesinde sınıflandırmamızı daha doğru hale getirebilmek mümkün olabilmektedir. Ancak LDA birçok çalışmada (Ekinci vd., 2020), (Onan vd., 2020), (Yıldıztepe ve Uzun, 2018), (Ventura vd., 2019) farklı veri türlerinde kullanılmasına ve daha geniş bir bakış açısı sunmasına rağmen, hala bize kelime persfektifinden sonuçlar sunmaktadır. Her ne kadar konu modelleme konusunda bizi daha doğru bir yöne götürse de kelime bakış açısından çıkmamız henüz mümkün olmamaktadır. Anlamsal düzeyde bilgi taşıyan, kelime bakış açısının dışında yöntemlerin kullanıması önemlidir.

Anlamsal bilginin ne olduğunu açıklamakta fayda vardır. Kelimeler bir araya geldiğinde cümlelerin, cümleler biraraya gelerek paragrafların oluştuğunu biliyoruz. Eğer cümle düzeyinde işlem yapabiliyor olsaydık, bu cümleler bize belgelerinin konusu hakkında daha fazla bilgi veriyor olacaktı. Paragraf düzeyi daha büyük miktarda veri taşımasına rağmen, genellikle işlemsel açıdan taşıdığı zorluktan ötürü pek tercih edilmemektedir. Üçüncü yöntem BERT bize belgelere kelime bakış açısından bakma şansı vermektedir. BERT kullanarak belgelerdeki

cümlelerden vektörler elde edilebilmektedir (Devlin vd., 2018). LDA ve TF-IDF dışında, belgeye cümle vektörleri aracılığıyla bakmak, konular ve belgeler arasında anlamsal benzerlikler bulmamızı da sağlamaktadır (Peinelt vd., 2020). LDA ve TF-IDF ile belgelere kelime bakış açısından bakarken, BERT kullanarak belgeleri cümle persfektifinden inceleyip anlamsal ilişkileri bulmaya çalışmak mümkündür. Ancak her yöntemin olduğu gibi BERT'inde zorlukları bulunmaktadır. BERT konu özelinde olduğu için inceleyeceğimiz belge seti değiştiğinde BERT modeli üzerinde her defasında ayarlamalar (batch büyüklüğü, öğrenme oranı, epoch sayısı vb. gibi) yapmamız gereklidir. Bunun yanında BERT modelini eğitmek diğer yöntemleri eğitmeye göre çok daha maliyetlidir. Çalışmamızda tüm bu yöntemlerin özellikleri incelenmiş ve sınıflandırma için en iyi yöntem seçilmeye çalışılmıştır.

TF-IDF ve LDA kelime bakış açısından dokümanlara bakabilmemize olanak sağlamakta, BERT ise dokümana cümle persfektifinden baktığı ve anlamsal ilişkiler sunduğu için doküman sınıflandırma konusunda bize farklı bir yol sunmaktadır. Çalışmamızda BERT'in sunduğu bu anlamsal cümle vektörleri yaklaşımı ve LDA'nin sunduğu olasılıksal kelime vektörleri yaklaşımını birleştirilmesi ile yeni bir yöntem deneyerek doküman sınıflandırma konusunda başarım oranı arttırılmaya çalışılmıştır. Her bir çalışmanın güçlü yönlerini ortaya çıkarmak üzere, birlikte kullanılmak suretiyle doğruluk oranı arttırılmaya çalışılmıştır.

Bu tezde incelenen veriler, spor ve eğitim konularını içeren İngilizce yazılmış 801 belgeden oluşmaktadır. Sporla ilgili bazı metinler eğitim ile ilgili bilgiler içerirken, bazı eğitim metinleri, konu modelleme yöntemlerinin çeşitlilikle nasıl başa çıktıklarını değerlendirmek ve onların sınıflarını test etmek için sporla ilgili bilgiler içermektedir. Bu yöntemleri uygulamadan önce, noktalama ve metin için anlam ifade etmeyen sözcüklerinin kaldırılması, isimlerin bulunması ve sözcüklerin köklerinin bulunması gibi veri ön işleme aşamaları uygulanmıştır.

1.1. Çalışmanın Amacı

Bu çalışmada büyük veri yığınları için, konu modelleme yöntemleri kullanılarak doküman sınıflandırma problemine çözümler aranmaktadır. En yaygın kullanılan konu modelleme yöntemleri olan TF-IDF, LDA ve BERT kullanılarak veri seti üzerinde doküman sınıflandırma uygulanmıştır. Çıkan en iyi sonuçlara göre, başarılı yöntemler birarada kullanılarak literatüre katkı sağlamak amaçlanmıştır.

2. LİTERATÜR ÖZETİ

Latent Dirichlet Allocation (LDA), belgeler ve metin külliyatları gibi ayrık veriler için olasılıksal bir model olarak tanıtıldı. Yazar, bu algoritmanın ana fikrinin, çok sayıda metin derlemesinin istatistiksel ilişkilerini koruyan kısa bir açıklamasını bulmak için sınıflandırma, özetleme, metin benzerliği ve yenilik tespiti için kullanılan, temel bir istatistiksel ilişki elde etmek olduğunu belirtmiştir (Zhang vd., 2020).

LDA, başka bir araştırmada verilen konuların özetleri olarak, konuları oluşturmak için de uygulandı. Burada yazarlar, her kelime temel bir konudan türetilir yaklaşımı benimsemiştir (Ritter vd., 2011).

Başka bir çalışmada, LDA'yı hangi yazarların hangi konuları kullandığını arayan yazar modeli ile genişletmişlerdir. Bu çalışmada, her yazar, konular üzerinde çok terimli bir dağılımla ve her konu, kelimelere göre çok terimli dağılımla ilişkilendirilir. Birden çok yazarı olan bir belge, yazarlarla ilişkilendirilen dağıtımların bir karışımı olan konular üzerinden bir dağıtım olarak modellenir (Rosen-Zvi vd., 2012).

LDA algoritması, Endonezya'daki büyük Twitter verilerinde; Ekonomik, Askeri, Spor, Teknoloji gibi dört ana konu oluşturmak için bir model üretmek için de kullanıldı. Yazarlar, metinler arasındaki benzerlikleri çıkarmak ve 5 kelimelik konu kümelerinin hangi metinleri temsil ettiğini bulmak için LDA'yı uyguladılar (Negara vd., 2019).

Malezya'da başka bir çalışmada 1973 yılından 2019 yılına kadar olan 1824 adet makalenin üzerine LDA ile konu modelleme çalışması yapıldı. Bu çalışmada temel amaç, makalelerden gelişmekte olan pazarlarda innavasyon araştırmasını desteklemek için 10 adet konu belirlenmeye çalışmaktı (Hilmi vd., 2020).

LDA ile yapılan başka bir çalışmada, Arapça belge sınıflandırmasında kullanılan ve sinir ağlarına dayanan yeni bir yöntem, LDA ile desteklenerek yapıldı. Bu

araştırmada, tüm metinlerden önemli konular çıkarıldı. Her konu farklı bir tanımlayıcı dağılımı ile tanımlandı ve ardından her metin belirli konuların vektörleri üzerinde tasvir edildi. Çalışmada yazarlar, önerilen çözümün Arapça metin sınıflandırma görevi için% 85.11 doğruluk oranı ile yüksek verimlilik sağlayabildiğini ortaya koydular (Alharbi vd., 2020).

Sosyal medyadaki nefret söylemini analiz etmeyi amaçlayan başka bir çalışmada, LDA ve denetimsiz bir öğrenme çeşidi olan self-organizing maps (SOM) ve K-means kümeleme uygulamasını bir arada kullanan bir metedoloji önerildi. Yazarlar, bu yöntem ile LDA'dan alınan kümeler, SOM ve K-means için bir girdi oluşturdu. Buradan alınan değerler birleştirilerek nefret söylemi analizi için bir yöntem sundular (Saini vd., 2020).

LDA, sadece konu modelleme ve doküman sınıflandırılmada kullanılmadı. Tıp alanında yapılan bir çalışmada, sağlıklı dokular ve kanser dokularında gen ekspresyon modellerini keşfetmek için denetimsiz bir konu modeli olan LDA kullanıldı. Burada, meme kanseri ve sağlıklı dokudan alınan haberci RNA (mRNA) sekans verileri, LDA'nın sağlıklı dokuya karşı kanser sınıflandırmasına uygulanması için etkili bir yaklaşım belirlemek için kullanıldı (Naderan ve Zaychenko, 2020).

Bidirectional Encoder Representations from Transformers (BERT) yeni bir dil temsil modülü olarak Google tarafından tanıtıldı. Yazarlara göre; BERT, özellik tabanlı ve ince ayar gibi aşağı akış görevleri için önceden eğitilmiş bir dil temsilidir ve bu çalışmada da kullanılan, dokümanlara cümle bakış açısından yaklaşmamızı sağlayan cümle yerleştirme vektörleri üretmektedir (Devlin vd., 2018).

Başka bir çalışmada, BERT ile dokümanlar arasında anlamsal benzerlikler bulunarak konu modellemeye çalışılmıştır. Ayrıca LDA ve BERT'nin birlikte çalıştıklarında daha iyi sonuçlar verdiği yazarlar tarafından vurgulanmıştır (Peinelt vd., 2020).

Tek bir dilde yazılan makalelerde veya çok dilli olarak birden fazla farkı dillerde yazılan makalelerde, tensör benzerliği yerleştiren LDA olasılık değeri ve BERT cümle vektörleriyle konu değerlendirmesini bulmak için, bu iki yöntem birleştirilmiştir (Xie vd., 2020).

BERT ile yapılan başka bir çalışmada, sosyal medyada üretilen içeriklerin, insanlar tarafından mı yoksa botlar tarafından mı üretildiğini anlamak için bir model geliştirilmiştir. Bu makalede yazarlar tarafından, sosyal medya bot algılama modeli için konudan bağımsız özellikleri tanımlamak üzere, tweetlerin duyarlılık sınıflandırması için BERT kullanan yeni bir model sunulmuştur. Yeni bot algılama modeli için konudan bağımsız özellikler türetmek üzere bir DDİ yaklaşımı kullanmanın, bu çalışmayı önceki bot algılama modellerinden ayırdığına değinilmiştir. Yazarlar çalıştıkları veri setinin içeriğini bir bot veya insan tarafından oluşturulmuş olarak sınıflandırarak %94 doğruluk değerini yakaladıklarını belirtmişlerdir (Heidari ve Jones, 2020).

Belge sınıflandırma üzerine yapılan bir diğer çalışmada, BERT üzerinde ince ayarlamalar uygulanarak, Reuters, AAPD ve IMDB veri setleri üzerinde çalışılmıştır. Ayrıca yazarlar tek katmanlı bir sistemin etkinliğini artırmak için BERT modelleri tarafından öğrenilen bilgileri kullanan LSTM modelini bilgileri damıtmak (distillation) için kullandıklarını belirttiler. Bu yöntem ile yazarlar BERTlarge'a göre 40 kat daha hızlı çıkarım yapabildiklerinden bahsetmişlerdir (Adhikari vd., 2019).

Konu modelleme için kullanılan bazı DDİ yöntemleri vardır. Bunlardan en önemlisi terim frekansı-tersine çevrilmiş dokümanfrekansıdır (TF-IDF). TF-IDF, bir belgedeki bir terimin önemini belirten istatistiksel bir yöntemle hesaplanan ağırlık faktörü olarak tanımlanır. TF-IDF yöntemi ilk olarak Salton tarafından tanıtıldı. TF-IDF, çeşitli konularla ortak özellikleri paylaşabilen belgelerde daha sık kullanılan ilgili konuyu ve ilişkili terimleri verir (Salton ve Buckley, 1988).

Konu modelleme yöntemleri ile ilgili bir araştırmada, ise sınıflandırma modellerini eğitmek için sadece TF-IDF ile elde edilen ağırlıklar kullanılmıştır.

Bu çalışmada yazarlar, kullanıcıları sınıflandırmada %90 doğruluk değeri yakaladıklarını vurgulamışlardır (Hong ve Davison, 2010).

Ayrıca araştırmacılar tweet'leri karakterize etmek için konu modellemesi kullanmış, TF-IDF ve LDA'nın birlikte kullanılmasının daha doğru sonuçlar verdiğini belirtmişlerdir (Ramage vd., 2010).

TF-IDF'nin bir başka konu bazlı uygulaması, tıbbi raporlardaki belirsizlik seviyesini tahmin etmek amacıyla, 5 genel konu sınıfını belirlemek üzere sınıflandırıcıların eğitimi için yürütülmüştür (Wang vd., 2019).

3. YÖNTEM

3.1. Veri Seti

Çalışmada veri seti olarak lisans öğrencileri tarafından toplanan 801 adet doküman kullanılmıştır. Bu dokümanların yarısını eğitim diğer yarısını ise spor ile ilgili konuları içeren dokümanlar oluşturmaktadır. Eğitim ile ilgili dokümanlardan bazıları spor eğitimi hakkında da bilgi vermektedir. Burada amaç geliştirilecek modelin başarı sınırlarını karmaşıklığı artırılmış bir veri seti ile test etmektir. Aşağıdaki çizelge 3.1'de dokümanların konulara göre dağılımları belirtilmiştir.

Çizelge 3.1. Dokümanların dağılımı

Doküman Tipi	Doküman Adedi
Spor	400
Eğitim	401
Total	801

Eğitim ve spor dokümanlarından örnek iki adet doküman başlığı ve içeriği aşağıdaki çizelge 3.2'de listelenmiştir.

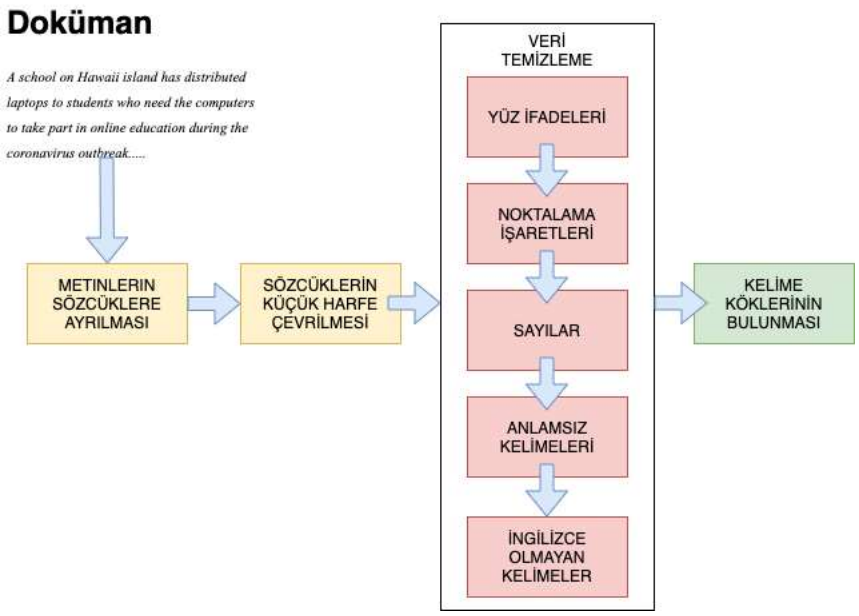
Çizelge 3.2. Örnek Eğitim ve Spor Dokümanları

Konu/No: Eğitim/710 **Başlık:** Hawaii School Provides Laptops For Home Online Education **İçerik:** A school on Hawaii island has distributed laptops to students who need the computers to take part in online education during the coronavirus outbreak. Hawaii News Now reported Wednesday that Pahoa High and Intermediate School distributed about 200 Google Chromebooks during three drive-thru collection events. The public school's students signed release forms allowing them to borrow the laptops. The school's online learning program began this week and staff said teachers and students are adapting to the changes
Konu/No: Spor /369 **Başlık:** Badminton Star aiming for Olympic **İçerik:** Badminton star aiming for Olympic gold 2May 2019 Badminton star PV Sindhu became the first Indian woman to win an Olympic silver medal in 2016. As she builds towards Tokyo 2020, BBC Sport meets the 23-year old who was named by Forbes as the seventh highest-earning sportswoman of 2018. BBC World News is featuring some of the highest profile Indian sportsmen and women in Sport Today, including Commonwealth javelin champion Neeraj Chopra, Olympic badminton silver medallist PV Sindhu and Indian football's highest international goalscorer, Sunil Chhetri.

3.1.1. Veri Ön İşleme

Veri seti işlenmeye başlamadan önce bazı ön işleme adımları uygulanmaktadır. Veri setinin içerisinde bulunan dokümanlarda verinin doğrulanması, hataların giderilmesi ve normalleştirilmesi gerekmektedir. Bu adımları uygulamaktaki temel amaç, veri setinin çalışılmasında oluşabilecek hataların, yanlış yönlendirmelerin ve çalışmada kullanılmayacak olan gereksiz içeriklerin temizlenmesidir.

Bütün bu durumlar ışığında öncelikle veri setinin içerisinde bulunan her bir doküman tokenlarına (sözcüklere) ayrılır. Sonrasında bu sözcüklerin tamamı normalizasyon amacı ile küçük harflere çevrilir. Bu adımın ardından bütün sözcüklerin yazım hataları giderilir. Sözcükler içerisinde yer alan sayılar, yüz ifadeleri (emojiler), noktalama işaretleri ve en önemlisi anlamsız kelimeler (stop words) sözcük listesinden temizlenir. Son olarak sözcüklerin kökleri bulunarak normalizasyon işlemi (sayısal hesaplamada benzerlik açısından) tamamlanır. Aşağıda şekil 3.1'de uygulanan bütün adımlar gösterilmektedir.



Şekil 3.1. Veri ön işleme adımları

Yukarıdaki şekilde verilen tüm adımların uygulandıktan sonra örnek verilen spor ve eğitim dokümanlarının incelemeye hazır halleri aşağıdaki çizelge 3.3'te gösterilmiştir.

Çizelge 3.3. Dokümanların veri ön işleme sonrası gösterimi

Konu/No: Eğitim/710 **Başlık:** Hawaii School Provides Laptops For Home Online Education **İçerik: [**'school', 'laptop', 'home', 'online', 'educate', 'school', 'hawaii', 'island', 'laptop', 'student', 'compute', 'part', 'educate', 'outbreak', 'news', 'wednesday', 'school', 'phrasebook', 'collect', 'event', 'school', 'student', 'release', 'form', 'laptop', 'school', 'program', 'week', 'staff', 'teacher', 'student', 'change']
Konu/No: Spor /369 **Başlık:** Badminton Star aiming for Olympic **İçerik:** ['badminton', 'star', 'olympic', 'gold', 'badminton', 'star', 'woman', 'olympic', 'silver', 'medal', 'toward', 'tokyo'', 'sport', 'forbes', 'sportswoman', 'world', 'news', 'profil', 'sportsmen', 'women', 'sport', 'today', 'commonwealth', 'javelin', 'champion', 'badminton', 'silver', 'medallist', 'footbal', 'goalscorer']

Kelimeleri ayrıştırma(tokenizasyon) işlemi gerçekleştirilirken nltk.tokenize sınıfının, WordTokenizer modülü kullanılmıştır. Daha sonra yazım hatalarını düzeltmek için symspellpy kütaphanesinin SymSpell modülü kullanılmıştır. İngilizce olmayan kelimelerin ayrıştırılması için language_detecter sınıfından yararlanılmıştır. Kelimeleri fiilerden ayrıştırmak için nltk kütüphanesinin postaggin özelliği kullanılmıştır. Durak kelimelerinin temizlenmesi için stop_words python kütüphanesinden yararlanılmıştır. Bu işlem esnasında veri seti defalarca incelenerek, çalışma için gereksiz görülen, tarih, lokasyon, kişi isimleri vb. gibi kelimeler anlamsız kelimeler olarak görülmesine yarayan regex yapıları geliştirilerek çıkarılmıştır.

3.2. Terim Frekansı(TF)-Tersine Çevrilmiş Doküman Frekansı(IDF)

TF-IDF'nin amacı, temelde bir belgedeki kelimenin önemini belirlemek için DDİ alanlarından biri olan bilgi alma ve bilgi çıkarma alt görevidir (Zhang vd., 2011).

Öncelikle TF değeri hesaplanır.

$$tf(w,d) = \log(1 + f(w,d)) \tag{3.2}$$

Sonrasında IDF değeri hesaplanır.

$$idf(w, D) = \log(\frac{N}{f(w,D)}) \quad (3.3)$$

- N very setindeki toplam doküman sayısını temsil etmektedir.
- d verilen dokümanı temsil etmektedir.
- D, tüm belgelerin koleksiyonudur.
- w bir belgedeki belirli bir sözcüktür.

Li ve Wu tarafından yapılan araştırmada, K-means kümeleme yöntemi TF-IDF değerleri kullanılarak gerçekleştirilmiştir. Bu çalışmada TF-IDF hesaplamasından elde edilen veriler K-ortalamalı kümeleme yöntemine girdi olarak kullanılmış ve dokümanlara en uygun sınıflandırma yöntemleri çıkartılmıştır (Li vd., 2012).

3.3. Latent Dirichlet Allocation(LDA)

Konu modellemeye yönelik LDA yöntemi, kelime kümelerinin, belgelerin bir temsili olmasıdır şeklinde ifade edilmiştir (Ha-Thuc vd., 2018). Bir LDA modeli oluştururken, bir sözlüğe ve kelime çantasına (Bag of Words -BOW) ihtiyacımız bulunmaktadır (Chehal vd., 2020). Açık kaynak kodlu bir python kütüphanesi olan Gensim kütüphanesi, külliyat ve sözlük oluşturmak için kullanılır (vRehruvrek ve Sojka, 2010). Önceden ön işlemesi yapılmış veriler Gensim kütüphanesine girilir, ardından her bir sözcük için bir indeks çıktı olarak üretilir. Sözlük, indekslerine sahip bu sözcüklerden oluşur ve BOW corpus, bu sözcüklerin indeksinden ve sıklığından oluşur. Modeli oluşturduktan sonra 10 konu grubu elde edilir. Model üzerinde farklı sayılarda konu grubu ayrımı denenmiş olup, en iyi doğruluk oranını 10 adet olarak belirlendiğinde alınmıştır. Herbir konu grubu içerisinde skorlarına ve birbirleriyle ilişkilerine göre sözcükler bulunmaktadır. Dokümanlar bu konu gruplarındaki sözcük skorlarına yakınlıklarına göre dağılır. Bu konu gruplarının yarısı sporu, diğer yarısı ise eğitimi temsil etmektedir.

3.4. Bidirectional Encoder Representations from Transformers (BERT)

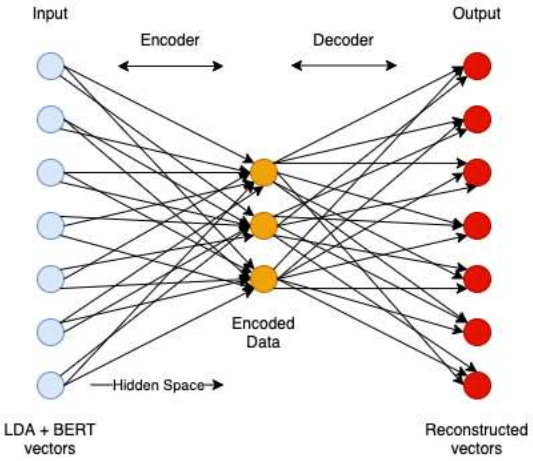
Çalışmada, BERT'in cümle vektörü üretme özelliğini kullanmak için açık kaynaklı bir model olan cümle-dönüştürücü (sentece-transformers) reposundan alınan bert-base-nli-mean-tokens kullanılmıştır. Bu repo, model eğitiminin metin ve cümle gömme vektörleri oluşturmasına ve kullanmasına izin verir (Reimers ve Gurevych, 2019). Cümle gömme vektörleri, belgelerdeki metinden, vektörlerin cümlelerin bağlamsal tanımını yakaladığı bir vektör uzayında oluşturulur. Bu cümle yerleştirme vektörlerinden 10 set konu kümesi elde edilir. Yine LDA'da olduğu gibi, bu kümeler eğitim ve spor konularını içerir.

3.5. BERT ve LDA'nın Birleştirilmesi

Çalışmada, daha önce başarımını test ettiğimiz yöntemlerin güçlü yönlerini birleştirerek daha iyi sonuçlara ulaşılmak hedeflenmiştir. Bu hedeften yola çıkılarak, BOW, LDA kullanılarak bulunur ve bağlamsal bilgi BERT kullanılarak elde edilir. LDA ile elde edilen olasılık vektörleri ve BERT ile elde edilen cümle gömme vektörleri bir girdi oluşturmak için birleştirilir. Burada, iki vektör arasındaki önemi dengelemek için LDA lehine bir ağırlıklandırma parametresi kullanılır. Otomatik kodlayıcı, verilerdeki daha önemli bilgileri ortaya çıkarmak için kullanılır.

3.5.1. Otomatik Kodlayıcı

Otomatik kodlayıcılar, verileri doğrusal olmayan bir şekilde gizli bir alana yansıtabilir ve bu, verilerin en göze çarpan özelliklerini çıkarabilir (Zhang vd., 2021). Otomatik kodlayıcıların en iyi avantajı, veri kümesinin etiketlenmesini gerektirmeyen denetimsiz bir model olmasıdır (Naderan ve Zaychenko, 2020). Otomatik kodlayıcı ile elde edilen daha fazla bilgi içeren daha düşük boyutlu temsiller, kümeleme modeli için girdi olarak kullanılır. Çalışmada üç düğümlü bir otomatik kodlayıcı kullanılmıştır. Aşağıdaki şekil 3.2'de kullanılan otomatik kodlayıcının yapısı gösterilmiştir.



Şekil 3.6. Otomatik kodlayıcı

Çalışmada input layer LDA ve BERT'in vektörlerinden oluşmaktadır. Giriş katmanı ile hidden space arasındaki kısma encoder adı verilir. Encoder çok boyutlu datayı az boyutlu dataya dönüştürmemizi sağlar. Şekilde gösterilen decoder kısmı ise sıkıştırılmış hidden space'in boyutunu artırarak girdiyi yeniden inşa etmeye çalışır. Çalışmada otomatik kodlayıcının ana amacı LDA ve BERT vektörlerinin içinden önemli input özelliklerini çıkarmaktır.

3.6. Karar Matrisi ile Çoğunluk (Majority) Analizi

Çalışmada her bir yöntemden dokümanlar için sonuçlar alındı. Bu yöntemlerin işaretlediği dokümanların sonuçları bir karar matrisinde birleştirildi. Ve ağırlıklı olarak doküman hangi sınıftaysa dokümanın sınıfı bu yönde işaretlendi. Aşağıdaki çizelge 3.4'te örnek bir karar matrisi listelenmiştir.

Çizelge 3.4. Karar matrisi

Docs ID	LDA-BERT	TF-IDF	BERT	LDA
0	1	1	1	0
1	1	1	1	1
21	1	0	0	0
29	1	0	1	1
51	1	0	0	1
249	1	0	1	1

Çizelge bütün yöntemler listelenmiştir ancak sadece TF-IDF, BERT ve LDA yöntemlerinin sonuçları kullanılmıştır. Bunun sebebi, LDA ve BERT

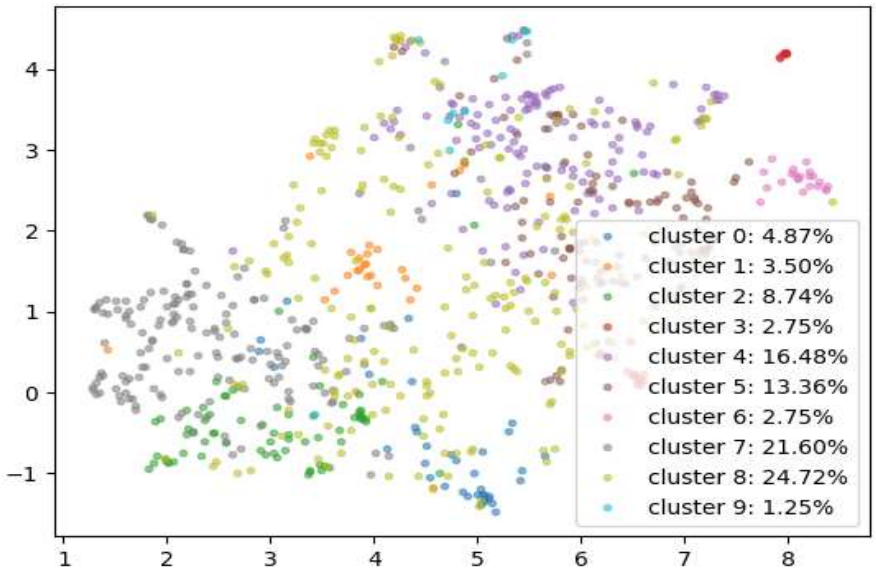
yöntemlerinin birleştilmiş halinin, LDA ve BERT'in bireysel uygulamalarınında sonuçlarını kapsayabileceği ihtimalidir. Çizelge 3.4'de 1'ler doğru sınıflandırmayı temsil ederken, 0'lar hatalı sınıflandırmayı temsil etmektedir. Bir doküman için 1'ler çoğunlukta ise o dokümanın sınıfı ilgili sınıf olarak işaretlenmiştir.

4. SONUÇ VE ÖNERİLER

Çalışmada veriler bir ön işlemeden geçirildi. Sonrasında sırasıyla TF-IDF, LDA, BERT, LDA ve BERT vektörlerinin birleştirilmesi ve Çoğunluk Yaklaşımı (Majority Vote) ile, ön işlemesi yapılmış veri üzerinde çalışıldı. Her bir yöntem için küme setleri oluşturuldu. Bu küme setlerinin yarısı eğitim diğer yarısı spor dokümanlarını temsil etmektedir. Yöntemlerden alınan sonuçlara göre sınıflandırma doğruluğu hesabı yapılmıştır.

4.1. TF-IDF Yönteminin Sonuçları

Sınıflandırma sonuçları bize eğitimle ilgili belgelerde çok sayıda yanlış eşleşme yapıldığını gösterdi. TF-IDF ile yapılan çalışmada spor sınıflandırması yüksek düzeyde elde edilmesine rağmen eğitim belgeleri için iyi sonuçlar vermemiştir. Bu, TF-IDF ile doküman sınıflandırmanın eğitim gibi daha belirsiz konularda yanlış sonuçlar verebileceğini ve belki de daha iyi sonuçlar elde etmek için büyük külliyat gerektirdiğini vurgulamaktadır. Aşağıda şekil 4.1'de görüleceği üzere konuların neredeyse yarısı 7 ve 8 numaralı kümelere dağılmıştır. Bu bir doküman sınıfının ağırlıklı olarak bulunduğunu göstermektedir.



Şekil 4.1. TF-IDF ile küme setlerinin dağılımı

Aşağıdaki çizelge 4.1'de listelenen sonuçlar spor dokümanlarının 7 ve 8 numaralı sınıflarda öbeklendiği göstermektedir. TF-IDF ile yapılan sınıflandırmada 4 ve 5 numaralı kümelerde eğitim dokümanlarını sıfır hata ile sınıflandırmıştır.

Çizelge 4.1. TF-IDF ile kümelerin dağılımı

Küme numarası	**Spor**	**Eğitim**	**Doğru**	**Hatalı**	**Hata Oranı**
0	23	16	23	16	%41,03
1	18	12	18	12	%40,00
2	52	12	52	12	%18,75
3	1	20	20	1	%4,76
4	0	131	131	0	%0,00
5	0	108	108	0	%0,00
6	2	20	20	2	%9,09
7	142	33	142	33	%18,86
8	161	39	161	39	%19,50
9	1	10	10	1	%9,09
Toplam	400	401	685	116	%15.5

Aşağıdaki çizelge 4.2'de listelenen sonuçlarda açıkça görüldüğü üzere TF-IDF yöntemi spor dokümanlarını neredeyse hatasız olarak bulmuştur

Çizelge 4.2. TF-IDF yönteminin sonuçları

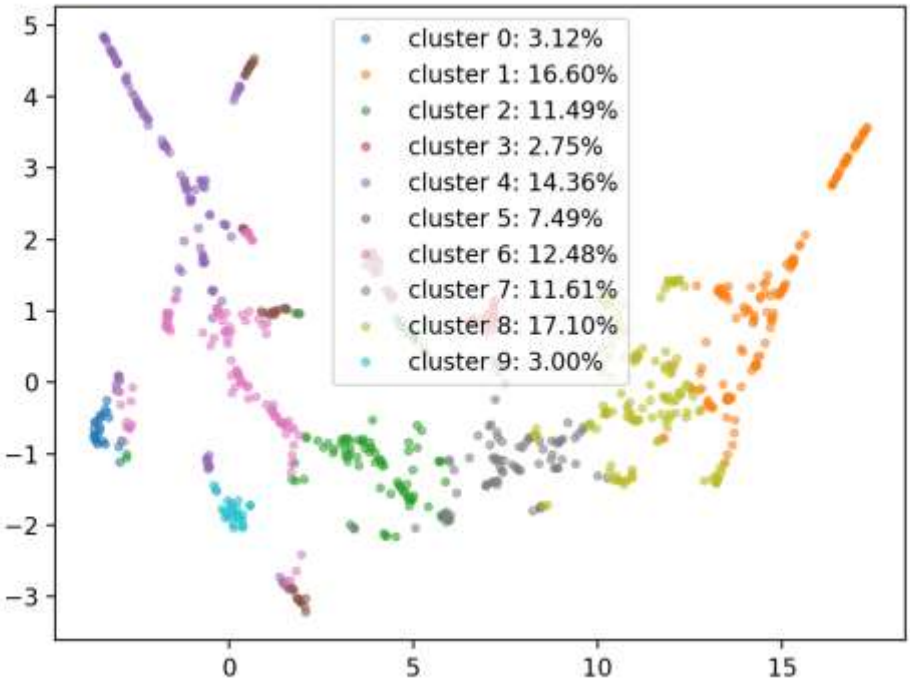
Sınıf	**Doküman Sayısı**	**Doğru Sınıflandırma**	**Hatalı Sınıflandırma**	**Başarım Oranı**
Eğitim	401	289	112	%72
Spor	400	396	4	%99
Toplam	801	685	116	%85.5

Spor dokümanlarını sınıflandırılmasında TF-IDF yöntemi %99 gibi çok yüksek bir oran sunmaktadır. 400 doküman içerisinde sadece 4 adet dokümanda yanılmıştır. Ancak eğitim dokümanlarında sonuçlar spor dokümanlarından çok uzaktır. 289 adet doküman doğru sınıflandırılırken, 112 adet doküman hatalı sınıflandırılmıştır. Doğruluk oranı olarak %72 gibi düşük sayılabilecek bir değer elde edilmiştir. Burada temel sonuç, spor gibi daha dar bir alanda kelimelerin frekansları doğru bir sınıflandırmaya olanak verdiği yönündedir. Eğitim dokümanları içerisinde spor eğitimi gibi birçok farklı daldan bahsedildiği için

sınıflandırma oranı düşük kalmıştır. TF-IDF bütün dokümanların sınıflandırılma oranına bakıldığında bize %85.5 oranında bir başarım sunmaktadır.

4.2. Latent Dirichlet Allocation Yönteminin Sonuçları

Çalışmadaki ikinci deney olan LDA yöntemi uygulandığında oldukça kabul edilebilir sınıflandırma sonuçlarına ulaşılmıştır. LDA, TF-IDF yöntemi ile karşılaştırıldığında, küçük boyutlu külliyat kullanarak bile daha iyi sınıflandırma yapabildiği ve karmaşık konuları ele alma becerisine sahip olduğu sonucuna varılabilmektedir. Aşağıda şekil 4.2'te konu kümelerinin dağılımı gösterilmiştir.



Şekil 4.3. LDA ile küme setlerinin dağılımı

Aşağıda çizelge 4.3'de LDA ile elde edilen sonuçlar listelenmiştir.

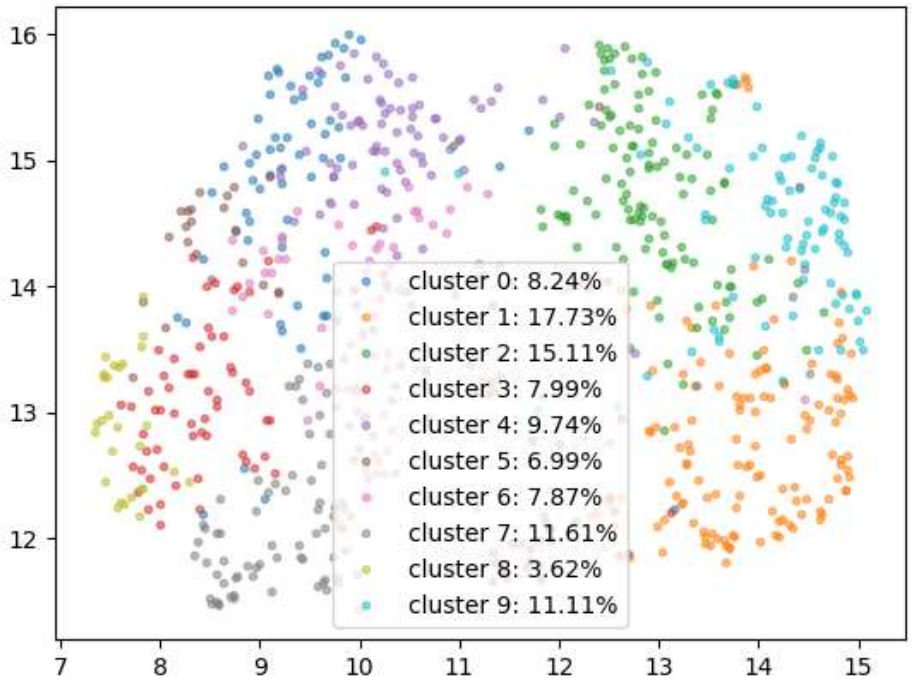
Çizelge 4.3. LDA yönteminin sonuçları

Sınıf	**Doküman Sayısı**	**Doğru Sınıflandırma**	**Hatalı Sınıflandırma**	**Başarım Oranı**
Eğitim	401	363	38	%90.5
Spor	400	333	67	%83.3
Toplam	801	696	105	%86.7

TF-IDF yöntemine kıyasla eğitim dokümanlarında ciddi bir başarım oranı yakalanmaktadır. 401 adet doküman içerisinden 363 adet dokümanı başarı ile sınıflandırırken, 38 adet dokümanda hatalı sınırlandırma yapılmıştır. Başarım oranı ise %90.5 olarak kayıt edilmiştir. Spor dokümanlarındaki başarım oranı ise TF-IDF'e kıyasla düşüktür. Bunun sebebi olarak LDA konuların belirli ya da belirsiz olmasına bakmaksızın hangi kelimenin hangi dokümanda hangi konuyu temsil ettiğini arayan bir konu modelleme yöntemi olmasıdır. LDA, spor dokümanlarında 400 adet doküman içerisinde 333 adet dokümanı doğru olarak sınıflandırırken, 67 adet dokümanda hatalı sınıflandırma gerçekleştirmiştir. Spor dokümanlarındaki başarım oranı %83.3 olarak kaydedilmiştir. Genel ortalama olarak LDA bize %86.7 gibi bir başarım oranı sunmaktadır.

4.3. BERT Yönteminin Sonuçları

Bir diğer yöntem olan BERT denenmiş ve beklendiği gibi giriş bölümünde anlatılan cümle yoluyla bağlama göre daha fazla bilgi getirmiş ve aşağıdaki çizelgede görüleceği üzere en iyi sonuçları vermiştir. TF-IDF ve LDA ile kelime bakış açısı üzerinden bir sonuçlar elde edebilmiştik. BERT bize anlamsal olarak cümlelerin ilişkisine bakmanın da doküman sınıflandırma için çok şey söylebileceğini göstermiştir. Aşağıda şekil 4.3'de bu 10 kümenin dağılımı gözükmektedir. Bazı konu kümelerinde çok net bir ayrım yapılabilmişken, diğer kümeler fazlasıyla birbine girmiş bir görüntü sergilemektedir. Ancak BERT bu kümleri ayrımada da başarılıdır. Diğer modellerin aksine BERT daha iyi ortalama dağılım yapmıştır.



Şekil 4.3. BERT ile küme setlerinin dağılımı

Diğer modellerde bazı küme gruplarının temsil ettiği dokümanlar %20'nin üzerindedir. Aşağıdaki çizelgede her kümede hangi doküman tipinin tempsil edildiği ve hatalı olarak kaç adet dokümanın işaretlendiğine yönelik bilgiler listelenmiştir. BERT en düşük başarım gösterdiği 8 numaralı kümede %20 lik bir hata oranı ile dokümanları doğru sınıfa koyabilmiştir. Yine kümedeki doküman sayısından bağımsız olarak maximum 10 hata ve mininum 1 hatalı sınıflandırma yapabilmiştir.

Çizelge 4.4. BERT yönteminin kümelerin dağılımı

Küme numarası	**Spor**	**Eğitim**	**Doğru**	**Hatalı**	**Hata Oranı**
0	4	63	63	4	%5,97
1	131	1	131	1	%0,76
2	112	5	112	5	%4,27
3	7	63	63	7	%10,00
4	6	74	74	6	%7,50

5	41	6	40	6	%12,77
6	7	65	65	7	%9,72
7	7	90	90	7	%7,22
8	6	24	24	6	%20,00
9	79	10	79	10	%11,24
Toplam	400	401	742	59	

Aşağıdaki çizelge 4.5'te BERT yönteminden alınan sonuçlar listelenmiştir.

Çizelge 4.5. BERT yönteminin sonuçları

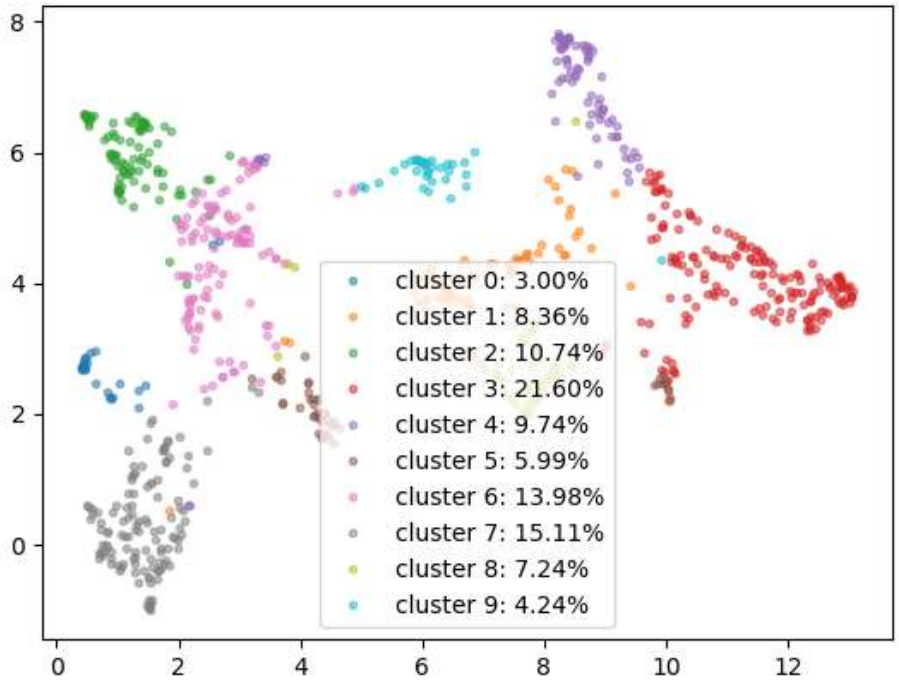
Sınıf	**Doküman Sayısı**	**Doğru Sınıflandırma**	**Hatalı Sınıflandırma**	**Başarım Oranı**
Eğitim	401	380	21	%94.7
Spor	400	362	38	%90.5
Toplam	801	742	59	%92.6

BERT, TF-IDF ve LDA yöntemlerinin aksine eğitim dokümanlarını sınıflandırmada daha başarılı sonuçlar elde etmiştir. 401 adet eğitim dokümanının 380 adedini başarılı olarak sınıflandırabilmiş ve 21 adet dokümanda hatalı sınıflandırma yapmıştır. BERT için eğitim dokümanlarını sınıflandırma %94.7 gibi bir başarım oranı kaydedilmiştir. Spor dokümanlarında da BERT'in başarısı gözükmektedir. 400 adet doküman içerisinde 362 adet dokümanı başarılı olarak sınıflandırırken 38 dokümanda hatalı sınıflandırma yapmıştır. BERT tüm dokümanların başarım oranı ölçüldüğünde bize %92.6 gibi bir oran sunmaktadır. Bu durum cümle vektörlerinin anlamsal ilişkisinin doküman sınıfı belirleme konusunda ne denli önemli olduğunu göstermektedir.

4.4. LDA ve BERT Yöntemlerinin Birleştirilmesi ile Elde Edilen Sonuçlar

BERT ve LDA deneyleri daha iyi sonuçlar verdiğinden dolayı, son olarak sınıflandırma için her iki yöntemi bir araya getirerek yeni bir yöntem aranmaya çalışılmıştır. BERT ile oluşturulan anlamsal bilgi içeren cümle vektörlerin ve LDA

kullanılarak oluşturulan olasılık vektörlerinin birleştirilmesiyle bu yöntem elde edilmiştir. Aşağıda şekil 4.4'te konu kümelerinin dağılımı gösterilmiştir.



Şekil 4.4. LDA + BERT ile küme setlerinin dağılımı

Aşağıdaki çizelge 4.6'da LDA-BERT birleşik yönteminin küme bazında dağılım sonuçları listelenmiştir. Yöntem toplam 81 adet dokümanı hatalı olarak sınıflandırmıştır. Bu hatalı sınıflandırılan dokümanların 53 tanesinin 2,3 ve 7 numaralı kümelerde öbeklendiğini aşağıdaki çizelgeden görebilmekteyiz.

Çizelge 4.6. LDA-BERT birleşik yöntemi ile kümelerin dağılımı

Küme numarası	**Spor**	**Eğitim**	**Doğru**	**Hatalı**	**Hata Oranı**
0	5	19	19	5	%20,83
1	61	4	61	4	%6,15
2	17	70	70	17	%19,54

3	24	153	153	24	%13,56
4	64	6	64	6	%8,57
5	46	2	46	2	%4,17
6	100	4	100	4	%3,85
7	12	116	116	12	%9,38
8	64	0	64	0	%0,0
9	7	27	27	7	%20,59
Toplam	400	401	720	81	%10,11

Aşağıdaki çizelge 4.7'de bu yeni yöntemin sonuçları listenmiştir.

Çizelge 4.7. LDA-BERT birleşik yönteminin sonuçları

Sınıf	**Doküman Sayısı**	**Doğru Sınıflandırma**	**Hatalı Sınıflandırma**	**Başarım Oranı**
Eğitim	401	385	16	%96
Spor	400	335	65	%83.7
Toplam	801	720	81	%89.8

İki yöntemin güçlü özelliklerinin birleştirilmesiyle elde edilen yöntemde, en iyi yöntemimiz olan BERT'e göre, eğitim dokümanlarında kayda değer bir sınıflandırma başarısı sağlanmıştır. Ancak spor dokümanlarında LDA'ya göre gelişme sağlanırken, BERT'in verdiği değerlerden daha düşük bir başarım elde edilmiştir. Yöntem bize eğitim dokümanlarında 401 adet dokümandan, 385 adedini başarılı olarak sınıflandırırken, 16 adet dokümanı hatalı olarak sınıflandırmıştır. Eğitim alanında başarım oranı %96 olarak kaydedilmiştir. Spor dokümanlarında ise 400 adet dokümandan 335 adedi doğru olarak sınıflandırılırken 65 adet dokümanda hatalı sınıflandırma yapılmıştır. Spor dokümanları için sınıflandırma başarım oranı %83.7 olarak ölçülmüştür. Yöntemde, tüm dokümanların başarım oranına bakıldığında 801 adet dokümandan 720 adet doğru 81 adet yanlış sınıflandırmaile %89.8 gibi bir başarım elde edilmiştir.

4.5. Coğunluk Yaklaşımı (Majority Vote) ile Yöntemlerin Çoğunluk Analizi

Son olarak, tüm bu deneysel modeller bir toplu yöntem olarak kabul edildi. Diğer bir deyişle, bu modellerin çıktı sınıfı etiketleri çoğunlukta ele alınırsa, sonuçları iyileştirmek mümkün olabilir mi bu sorunun cevabı arandı. Böylece sınıflandırma için TF-IDF, LDA ve BERT modelleri etiketleri çoğunluk tekniği kullanılarak değerlendirilmiştir. Aşağıdaki çizelge 4.8'te çoğunluk analizinin sonuçları listelenmiştir.

Çizelge 4.8. Karar matrisi çoğunluk analizi yönteminin sonuçları

Sınıf	**Doküman Sayısı**	**Doğru Sınıflandırma**	**Hatalı Sınıflandırma**	**Başarım Oranı**
Eğitim	401	361	40	%90
Spor	400	379	21	%94.7
Toplam	801	740	61	%92.4

TF-IDF, LDA ve BERT'i çoğunluk analizi için seçmekteki ana amaç, çalışmadaki saf yöntemler olmasıdır. Eğer LDA ve BERT'in birleşiminin çıktısını kullansaydık bu aynı zamanda bazı dokümanlar için ayrı ayrı LDA ve BERT'in sonuçlarını da kapsayabilecekti. Bu da çalışma için çok faydalı bir veri oluşturmayacaktı. Karar matrisinden aldığımız çoğunluk analizinin başarım oranı hesaplandı. En iyi yöntemimiz olan BERT'e çok yakın değerler vermesine rağmen genel başarım oranında BERT'in başarım oranının altında kalmıştır. Eğitim dokümanlarında 401 adet doküman içerisinde 361 dokümanı doğru sınıflandırmışken, 40 adet dokümanda hatalı sınıflandırma yapıldığı tespit edilmiştir. Bu yöntemin eğitim dokümanlarındaki başarım oranı %90 olarak ölçülmüştür. Spor dokümanlarında ise TF-IDF'in bireysel çalışmasında aldığımız baskın sonuçların bu yönteme etki ettiğini görmekteyiz. 400 adet doküman içerisinde, 379 adet dokümanı doğru olarak sınıflandırırken, 21 adet dokümanda hatalı sınıflandırma yapılmıştır. Spor dokümanlarında başarım oranı %94.7 olarak ölçülmüştür. Yöntemin genel başarım oranı ise %92.4 olarak bulunmuştur.

4.6. Sonuçların Değerlendirilmesi

801 belgeden oluşan veri setine sırasıyla TF-IDF, LDA, BERT, LDA ve BERT birleştirilmesi ve Çoğunluk matrisi ile analiz olmak üzere farklı konu modelleme yöntemleri uygulanmıştır. Tüm sonuçlar dikkate alındığında, BERT'in ortalamada %92,6 ile en iyi skoru elde ettiği ölçülmüştür.

Esasen, TF-IDF ve LDA, kelime temelli yaklaşımları nedeniyle iyi tanımlanmış konuları ayırmada başarılı olsalar da, daha belirsiz konularda başarı oranının düşük olduğu gözlemlenmiştir. Ayrıca BERT, anlamsal çıkarımlar yapmamıza izin verdiği ve metin hakkında daha fazla bilgi içerdiği için belirsiz konularda LDA ve TF-IDF'den daha başarılı sonuçlar vermiştir. Öte yandan, BERT'in dezavantajları bağlam temelli olması ve her seferinde farklı konular için ince ayar gerektirmesi ve bunun yanında eğitim süresi diğerlerinin iki katı olmasıdır. Ek olarak, TF_IDF, LDA ve BERT gibi tüm yöntemler kullanılarak işaretledikleri dokümanların çoğunluğu ile elde edilen sonuçların BERT sonuçlarının üst sınırları ile sınırlı olduğu görülmüştür.

Gelecekte BERT yerine RoBERTa (Liu vd., 2019), SciBERT (Beltagy vd., 2019), ALBERT (Lan vd., 2020) vb. modeller kullanılabilir. Ayrıca, LDA vektörlerini seçmek için kullanılan ağırlık değeri üzerinde çalışmalar yaparak daha çok bilgi içeren vektörlerin seçilmesi sağlanabilir. Son olarak, gizli katmanlardan değerli bilgiler içeren vektörleri çıkarmak için kullandığımız, otomatik kodlayıcıdaki düğüm sayısı arttırılarak daha çok bilgi içeren yeni vektörler yeniden inşa edilebilir. LDA ve BERT ile elde edilen başarı değeri bu çıktı vektörleri kullanılarak artırılabilir. Yine gelecekte İngilizce veri seti üzerine yapılan bu çalışma, Türkçe veri setleri üzerine uygulanarak başarısı sınanabilir.

KAYNAKLAR

Adhikari, A., Ram, A., Tang R., Lin, J., 2019. DocBERT: BERT for Document Classification, Erişim Tarihi: 12.12.2020. https://arxiv.org/abs/1904.08398

Alharbi, A., Alharbi S., Aljaedi A., Akanbi O., 2020. Neural Networks Based on Latent Dirichlet Allocation For News Web Page Classifications. 2020 IEEE 2nd International Conference on Artificial Intelligence in Engineering and Technology (IICAIET), 1-6, Kota Kinabalu, Malaysia.

Beltagy, I., Lo, K., Cohan, A., 2019. SciBERT: A Pretrained Language Model for Scientific Text. Erişim Tarihi: 15.12.2020. https://arxiv.org/abs/1903.10676

Blei, D., Ng, A., Jordan, M., 2003. Latent dirichlet allocation Journal of Machine Learning Research, 993-1022.

Chehal, D., Gupta, P., Gulati, P., 2020. Implementation and comparison of topic modeling techniques based on user reviews in e-commerce recommendations. Journal of Ambient Intelligence and Humanized Computing.

Devlin, J., Chang, M., Lee, K., Toutanova, K., 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. Erişim Tarihi: 05.12.2020. https://arxiv.org/abs/1810.04805

Ekinci, E., Omurca, S., Kırık, E., Taşçı, Ş., 2020. Tıp Veri Kümesi için Gizli Dirichlet Ayrımı. Dokuz Eylül Üniversitesi Mühendislik Fakültesi Fen ve Mühendislik Dergisi, 67-80.

Eravcı, D., 2020. KURUMLARIN DİJİTAL DÖNÜŞÜMÜ: BÜYÜK VERİ. Çalışma İlişkileri Dergisi, 90 – 112.

Ha-Thuc, V., Srinivasan, P., 2008. Topic Models and a Revisit of Text-related Applications. 25-32.

Heidari, M., Jones, J., 2020. Using BERT to Extract Topic-Independent Sentiment Features for Social Media Bot Detection, 11th IEEE Annual Ubiquitous Computing, Electronics & Mobile Communication Conference, 0542-0547, New York City.

Hilmi, M., Mustapha, Y., Che Omar, M., 2020. Innovation in an Emerging Market: A Bibliometric and Latent Dirichlet Allocation Based Topic Modeling Study, International Conference on Decision Aid Sciences and Application 882-886, Sakheer, Bahrain.

Hong, L., Davison, B., 2010. Empirical Study of Topic Modeling in Twitter. In Proceedings of the First Workshop on Social Media Analytics, 80–88.

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., Soricut, R., 2020. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. Erişim Tarihi: 09.12.2020. https://arxiv.org/abs/1909.11942

Li, Y., Wu, H., 2012. A Clustering Method Based on K-Means Algorithm. Physics Procedia, 1104-1109.

Li, X., Zhang, R., Wang, Q., Zhang, H., 2021. Autoencoder Constrained Clustering With Adaptive Neighbors, IEEE Transactions on Neural Networks and Learning Systems, 443-449.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V., 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach, Erişim Tarihi: 09.12.2020. https://arxiv.org/abs/1907.11692

Lossio-Ventura, J., Morzan, J., Alatrista-Salas, H., Hernandez-Boussard, T., Bian, J., 2019. Clustering and topic modeling over tweets: A comparison over a health dataset, IEEE International Conference on Bioinformatics and Biomedicine, 1544-1547, San Diego.

Naderan, M., Zaychenko, Y., 2020. Convolutional Autoencoder Application for Breast Cancer Classification, IEEE 2nd International Conference on System Analysis & Intelligent Computing, 1-4, Kyiv, Ukraine.

Negara, E., Triadi, D., Andryani, R., 2019. Topic Modelling Twitter Data with Latent Dirichlet Allocation Method, 2019 International Conference on Electrical Engineering and Computer Science (ICECOS), 386-390, Batam Island, Indonesia.

Onan, A., Yalçın, A., Atİk, E., 2020. Üniversite Bilgi Yönetim Sistemi Servis Destek Taleplerinin Konu Modelleme Tabanlı Analizi, Avrupa Bilim ve Teknoloji Dergisi, 389-397.

Peinelt, N., Nguyen, D., Liakata, M., 2020. tBERT: Topic Models and BERT Joining Forces for Semantic Similarity Detection, Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 7047-7055.

Ramage, D., Dumais, S., Liebling, D., 2010. Characterizing Microblogs with Topic Models, Proceedings of the International AAAI Conference on Web and Social Media.

Reimers, N., Gurevych, I., 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. Erişim Tarihi: 09.12.2020. https://arxiv.org/abs/1908.10084

Rosen-Zvi, M., Griffiths, T., Steyvers, M., Smyth, P., 2004. The Author-Topic Model for Authors and Documents, Proceedings of the 20th Conference in Uncertainty in Artificial Intelligence, Banff, Canada.

Ritter, A., Clark, S., Etzioni, O., 2011. Named entity recognition in tweets: An experimental study, Proceeding Conference Empirical Methods Natural language Process, 1524-1534.

Saini, Y., Bachchas, V., Kumar, Y., Kumar, S.,2020. Abusive Text Examination Using Latent Dirichlet Allocation, Self Organizing Maps and K Means Clustering, 4th International Conference on Intelligent Computing and Control Systems, 1233-1238, Madurai, India.

Salton, G., Buckley, C., 1988. Term-weighting approaches in automatic text retrieval,Information Processing & Management, 513-523.

vRehruvrek, R., Sojka, P., 2010. Software Framework for Topic Modelling with Large Corpora. Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks, 45-50.

Wang, Y., Sohn, S., Liu, S., Shen, F., Wang, L., Atkinson, E. J., Liu, H., Amin, S., 2019. A clinical text classification paradigm using weak supervision and deep representation. BMC Medical Informatics & Decision Making.

Wu, Z., Lei, L., Li, G., Huang, H., Zheng, C., Chen, E., Xu, G., 2017. A topic modeling based approach to novel document automatic summarization. Expert Systems With Applications, 12–23.

Xie, Q., Zhang, X., Ding, Y., Song, M., 2020. Monolingual and multilingual topic analysis using LDA and BERT embeddings. Journal of Informetrics.

Yalamanchili, H., Kho, S., Raymer, M., 2017. Latent Dirichlet Allocation for Classification using Gene Expression Data, IEEE 17th International Conference on Bioinformatics and Bioengineering, 39-44, Washington.

Yıldıztepe, E., Uzun, V., 2018. Olasılıksal Yöntemler ile Türkçe Metinlerin Anlamsal Benzerliğinin Belirlenmesi, Sinop Üniversitesi Fen Bilimleri Dergisi, 66-78.

Zhang, F., Gao, W., Fang, Y., Zhang, B., 2020. Enhancing Short Text Topic Modeling with FastText Embeddings, 2020 International Conference on Big Data, Artificial Intelligence and Internet of Things Engineering (ICBAIE), 255-259, Fuzhou, China.

Zhang, W., Yoshida, T., Tang, X., 2011. A comparative study of TF*IDF, LSI and multi-words for text classification. Expert System Application. 2758-2765.

ÖZGEÇMİŞ

Adı Soyadı : Süleyman ÖZDEMİRCİ

Eğitim Durumu

Lise : Antalya Anadolu Lisesi, 2008

Lisans : Orta Doğu Teknik Üniversitesi, Kuzey Kıbrıs Kampusu, Bilgisayar Mühendisliği Bölümü, 2015

Yüksek Lisans : İstanbul Ticaret Üniversitesi, Fen Bilimleri Enstitüsü, Bilgisayar Mühendisliği Anabilim Dalı, 2021

Yayınları

Özdemirci, S., Turan, M., 2021. Case Study on well-known Topic Modeling Methods for Document Classification, 2021 6th International Conference on Inventive Computation Technologies (ICICT), 2021, pp. 1304-1309