Doğal Dil İşleme Teknikleri Kullanılarak Disiplinler Arası Lisansüstü Ders İçeriği Hazırlanması

Araştırma Makalesi/Research Article

Ahmet ALBAYRAK*

Bilgisayar Mühendisliği, Düzce Üniversitesi, Düzce, Türkiye
ahmetalbayrak@duzce.edu.tr
(Geliş/Received:04.04.2020; Kabul/Accepted:11.08.2020)
DOI: 10.17671/gazibtd.714447

Özet—Bu çalışmada lisansüstü seviyede açılan düşünülen disiplinler arası bir dersin içeriğinin hazırlanması için veri madenciliği tekniklerinden doğal dil işleme yöntemleri kullanılmıştır. Lisansüstü ders, Veri Bilimi ve Uygulamaları adını taşımaktadır. Veri bilimi temelde istatistik ve bilgisayar bilimlerini içine alan disiplinler arası bir kavramdır. Dersin benzer bir ad ile literatürde yeri yoktur. Veri bilimi yaklaşımı veriyi öncelikleyen ve oldukça fazla alanda uygulanan bir yaklaşımdır. Uygulama alanı çok geniş olduğundan derse Veri Bilimi ve Uygulamaları adı verilmiştir. IEEE'nin yıllardır düzenlediği bir konferansta basılan bildiriler ders içeriğinin belirlenmesinde veri seti olarak kullanılmıştır. *Data Science and Advanced Analytics* adındaki konferansın bu yıl 7. si düzenlenecektir. 2015, 2016, 2017 ve 2018 yıllarında konferansa kabul edilen bildiriler veri setinde kullanılmıştır. Bildirilerin başlık kısımları ve anahtar kelimeler doğal dil işleme teknikleri ile analiz edilerek ders içeriği belirlenmiştir. Bu çalışmada ilk olarak veri seti hazırlandıktan sonra, veri üzerinde veri temizleme işlemi yapılmış ardından bildiri başlıkları sözcüklere ayrılmıştır. Sözcüklere ayrılan veri seti içinde sözcüklerin frekansları bulunarak frekansa göre ilk yirmi sözcük seçilmiştir. Doğal dil işleme sürecinde Apache Spark NTK paketi kullanılmıştır. Seçilen 20 sözcük atomik olduğundan tümevarım yöntemi ile ana konu başlıkları belirlenmiştir.

Anahtar Kelimeler- veri bilimi, doğal dil işleme, ders içeriği hazırlama, veri bilimcisi, konu modelleme

Preparing Interdisciplinary Graduate Course Contents Using Natural Language Processing Techniques

Abstract— In this study, natural language processing methods, one of the data mining techniques, were used to prepare the content of an interdisciplinary course that is planned to be opened at graduate level. The graduate course is called Data Science and Applications. Data science is an interdisciplinary concept that includes statistics and computer science. The course has no place in the literature with a similar name. Data science is an approach that prioritizes data and is applied in many fields. Since the application area is very wide, the course is called Data Science and Applications. Papers published at a conference organized by IEEE for years were used as a data set in determining the course content. The conference called Data Science and Advanced Analytics will be held for the 7th time this year. Papers accepted to the conference in 2015, 2016, 2017 and 2018 were used in the data set. The title texts and keywords of the papers were analyzed with natural language processing techniques and the course content was determined. In this study, after the first data set was prepared, data-cleaning process was performed on the data, and then the title of the papers was divided into words. The frequencies of the words are found in the data set devoted to the words and the first twenty words are selected according to the frequency. Apache Spark NTK package was used in the natural language processing process. Since the 20 words chosen are atomic, the main topic titles are determined by the induction method.

Keywords— data science, natural language processing, course content preparation, data scientist, topic modeling

1. GİRİŞ (INTRODUCTION)

Veri bilimi, özellikle büyük verinin ortaya çıkmasıyla birlikte önemli hale gelmiş ve büyük verinin analiz edilmesindeki zorlukları çözmek için gerekli olan bir dizi disiplini barındırmaktadır. Veri biliminde ana unsurlar veri, teknolojiler ve insanlardır. Günümüzde veri her yerde bulunmakta ve artmasıyla birlikte büyük veri sorunuyla başa çıkmak için teknolojiler geliştirilmektedir. İnsanlar özellikle sosyal medya ve akıllı telefon teknolojileri ile birlikte hem veri üreticisi hem de veri tüketicisi haline gelmiştir[1].

Bilgi işlem teknolojilerindeki gelişmeler ve verinin hızlı artması (akıllı telefonlar ve sosyal medya ile birlikte) veri bilimi kavramını öne çıkarmıştır. Büyük veriler, ticari sorunları çözmek için hangi verilerin kullanılabileceğini/entegre edilebileceğini, sorunları çözmenin yeni yollarını ve daha önce çözemediğimiz yeni sorunları düşünme imkanı vermektedir. Büyük verileri yenilikçi bir şekilde kullanmanın bu yeni yoluna veri güdümlü paradigma denilmektedir [2]. Büyük veri çağındaki veri sorunlarını ve veri güdümlü paradigmayı anlamak için veri bilimi şu faktörleri içermelidir: büyük veri altyapısı, büyük veri analizi yaşam döngüsü, veri yönetimi becerileri ve davranışsal disiplinlerdir. Büyük veri altyapıları arasında Hadoop ekosistemleri, NoSQL(Not Only SQL) veritabanları, bellek içi hesaplama gibi büyük veri teknolojilerinin yanı sıra bulut bilişim gibi büyük veri etkinleştirme teknolojileri bulunur. Büyük veri analizi yaşam döngüsü iş veriyi anlama, veriyi hazırlama ve entegre etme, model oluşturma, değerlendirme, dağıtım ve izleme dahil olmak üzere veri analizinin tüm aşamalarını içerir. Veri yönetimi becerileri geleneksel veri modelleme ve ilişkisel veritabanı bilgisini içerir. Davranış disiplinleri, eleştirel düşünme, üretken sorular sorma, alan uzmanlarıyla iletişim kurma (veri yönetimi hakkında hiç bilgisi olmayan veya çok az bilgi sahibi olan) ve proje sonuçlarını iş ile ilgili hale getirme gibi insanlar ve iş ile ilgili becerileri içerir[3].

"Veriler" ilişkisel veriler ve ilişkisel olmayan veriler (sosyal medya verileri ve web verileri gibi yapılandırılmamış ve yarı yapılandırılmış veriler vb.) ve algılayıcı verileri gibi çeşitli alanlardan gelebilir. "Teknolojiler" ise bellek hesaplama, veri madenciliği, makine öğrenimi, bulut bilişimdeki Hadoop ekosistemlerini ve NoSQL teknolojilerini içermektedir. "İnsanlar", bilgisayar bilimcileri, istatistikçi, alan uzmanları, veri bilimcileri ve iş analizcileri şeklinde tanımlanmaktadır[1].

Günümüzde birçok bilim adamı kendi veri bilimi tanımlarını yapmışlardır. Bunlardan, Dhar (2013) 'ın tanımı "veri bilimi, bilgiden genelleştirilebilir bilginin çıkarılması çalışmasıdır" şeklindedir. Burada veri biliminin ana unsuru olan veri dikkate alınır [4]. Stanton (2012), veri bilimini "büyük bilgi koleksiyonlarının toplanması, hazırlanması, analizi, görselleştirilmesi, yönetimi ve korunması ile ilgili ortaya çıkan bir çalışma alanı" olarak tanımlamıştır [5]. Bu tanım, büyük veri yaşam döngüsünü de dikkate almakta ve tüm yönlerini daha geniş bir şekilde kapsamaktadır. Provost ve Fawcett (2013) ise "veri bilimi, analiz yoluyla ana unsurları anlamak için prensipler, süreçler ve teknikler içerdiğini" söylemişlerdir [6].

Veri bilimi ayrıca davranışsal disiplinlerle yakından ilişkili ve bilgi olarak uzman bir insan tarafına (ihtiyaç analizi, kullanıcı arayüz tasarımı, modellerin doğrulanması ve alan uzmanlarıyla iletişim kurma) sahiptir. Veri, teknolojiler ve insan unsurundan en önemlisi insanlardır. Büyük verileri etkin bir şekilde işlemek için daha fazla bilgisayar, depolama aracı satın alınabilir, ancak insan kabiliyeti/yeteneği ölçeklenemez. Veri bilimcileri olarak adlandırılan kişileri eğitmek, büyük veri çağının ortaya çıkardığı zorluklar ile mücadele etmenin anahtar özelliğidir[4].

1.1. Veri Bilimcisi (Data Scientist)

Veri bilimcisi, büyük veri sorunlarıyla başa çıkabilen kişileri ifade etmek için kullanılan yeni ortaya çıkmış bir iş unvanıdır. Veri bilimcilerinin başlıca görevleri şunlardır:

- ✓ Sorunları çözmek için verilerden eyleme geçirilebilir bilgiler çıkarmak.
- ✓ İş hedefleri ve metrikleri ile sonuçları akıldaki sorulara göre değerlendirmek için doğru soruları sormak.
- ✓ İstenen gereksinimleri karşılamak. Sorun ile ilgili verileri tanımlamak ve verileri kullanmak / yeniden kullanmak/ birleştirmek.
- ✓ Doğru teknolojileri ve araçları seçmek.
- ✓ Çözüm alanlarını önceden belirlenmiş bir amaç olmadan keşfetmek.
- ✓ Alan uzmanlarıyla birlikte çalışmak.
- ✓ Analiz yapmak, değerlendirmek ve görselleştirmek. Veriye dayalı karar vermeyi otomatik hale getirmek şeklindedir.

Bu adımlar aslında verinin ortaya çıkmasından bir sistem içinde yönetilebilir hale gelinceye kadar ki tüm aşamalarını içermektedir. Bu verinin yaşam döngüsüdür[3].

Kimler veri bilimcisi olabilir sorusunun cevabı, "verinin yaşam döngüsü içinde kullanılabilecek olası teknolojiler, teknikler, yaklaşımlar, analiz yöntemleri vb. birçok işlemi yapan kişiler" şeklinde verilebilir. Bilgisayar bilimi, istatistik ve matematik eğitimi alan kişiler, gerekli bilgi ve uzmanlığa sahip oldukları sürece veri bilimcisi olabilirler. Bu bilgi/uzmanlık seviyesi verinin ortaya çıktığı andan itibaren ne amaç ile değerlendirilecek ise amaca uygun olarak yapılandırılmasını sağlamalıdır. Bu bilgi/uzmanlık seviyesi insanların tek başlarına ya da kısa sürede erişebilecekleri bir seviye değildir [7]. 2011 yılında, McKinsey 2018 yılına kadar, sadece ABD'de derin

analitik becerilere sahip 1,5 milyon yönetici ve veri analisti insana ihtiyaç duyulduğunu belirtmiştir [8]. Böyle bir ihtiyacı karşılamak için çeşitli düzeylerde kurslar veya eğitimler almak önemlidir.

Veri bilimi yaklaşımı eğitim olarak dünyada çeşitli üniversiteler bünyesinde lisans, yüksek lisans ve doktora düzeyinde ders olarak, hatta program olarak verilmektedir. ABD'de veri bilimi eğitimi dört kategoriye ayrılır: lisans programları, sertifika programları, yüksek lisans programları ve doktora programlarındaki uzmanlıklar şeklindedir. Lisans programları henüz çok yaygın değildir ve sadece birkaç üniversitede (Ohio State Üniversitesi, Washington Üniveritesi) bulunmaktadır. Sertifika programları çoğunlukla çevrimiçi olarak sunulmakta ve genellikle bir yıldan az bir sürede (bir yarıyıl) tamamlanmaktadır. Sertifika programları çoğunlukla Columbia Üniversitesi'nin veri bilimlerinde mesleki başarı sertifikası gibi lisansüstü düzeyde verilmektedir [9].

Master programları ABD'de veri bilimi eğitiminde en popüler olandır. Yüksek lisans programlarının süresi esnektir, bir yıldan iki yıla kadar değişmektedir. Veri bilimi Yüksek lisans programlarında kapsamlı programlar olarak sunulmaktadır. Bazı programlar çevrimiçi olarak sunulmakta ancak büyük ölçüde New York Üniversitesi'nin veri bilimindeki yüksek lisans programları gibi yüz yüze programlar tercih edilmektedir. Yüksek lisans seviyesinde veri bilimi dersinin verilmesi öğrenme açısından daha anlamlı olmaktadır. Bunun nedeni temel lisans eğitimi almış ve verinin çeşitli formları ile tanışmış kişilerin daha başarılı olacağı düşünülmektedir. Doktora programları veri bilimindeki eğitimler arasında dört program türü arasında en nadir olanlardır. Washington Üniversitesi tarafından sunulan tek bir doktora programı bulunmaktadır. Gelecekte, birçok veri bilimcisi, veri bilimi yaklaşımını lisans programlarında öğrenebilir ve daha fazla sayıda üniversitenin bu programları açabileceği düşünülmektedir. Bu resmi eğitim programları gelecekte veri bilimcilerinin yetiştirilmesinde ana kaynak olacaktır [3].

2. BENZER ÇALIŞMALAR (RELATED WORKS)

Doğal dil işleme yaklaşımı metinsel verinin yoğun olduğu tüm alanlarda kullanılabilecek teknikler içermektedir [10]. Literatürde bilgi sistemlerinde güncel içeriği takip etme, çeşitliliğe hâkim olma ve ilgili alanların etkisini öğrenme gibi çeşitli amaçlarla, 2003-2017 yılları arasında önde gelen altı BS dergisinde yayınlanan 2962 makaleye incelenmiştir. Makalelerde yazarlar tarafından kullanılan anahtar kelimeler dikkate alınarak konu modelleme çalışması yapılmıştır. Web of Science ortamındaki makaleler başlık, anahtar kelimeler, dergi adları yayınlanma yılı gibi parametreleri ile değerlendirilmiş ve veri seti hazırlanmıştır. Konular metin madenciliği yöntemleri ile belirlendikten sonra yıllara bağlı olarak popülariteleri çıkarılmıştır. Bunun sonucunda son yıllarda sosyal medya, tasarım bilimi ve çevrimiçi topluluklar gibi konuların popülaritesinin ivme kazandığı ifade edilmiştir [11].

Literatürde büyük miktarda metinsel bilginin olduğu diğer bir araştırma alanı ise yöneticilere politika önerilerinin yapıldığı plaformlardır. Buradaki kayıtlar daha çok e-dilekçeler ile e-devlet üzerinden paylaşılan istek, dilek ve şikayetlerden oluşmaktadır. Bu verilerin hepsinin okunarak analiz edilmesi oldukça zordur. Doğal dil işleme teknikleri ile verilerin analiz edildiği bir çalışmada aciliyeti olan 30 konu belirlenmiştir. Belirlenen konular üzerinden politikacılar yönlendirilebilmektedir [12]. Tıp alanında büyük miktarda elektronik klinik veri, serbest metin formatında önemli bilgileri içerir. Tıbbi karar almada yardımcı olabilmek için metnin verimli bir şekilde işlenmesi ve kodlanması gerekir. Yapılan bir çalışmada Acil Servis bilgisayarlı tomografi raporlarının sınıflandırılmasını sağlamak için doğal dil işleme teknikleri kullanılmıştır. Önerilen sistemde, doğal dil işleme teknikleri hasta raporlarından yapılandırılmış çıktılar üretmek için kullanılmıştır. Çalışmada doğal dil işlemenin ham metin sınıflandırma sonuçlarını iyileştirdiğini göstermektedir [13].

Hizmet sektöründe müşteri memnuniyeti çevrimiçi platformlar üzerinde paylaşılan ilgili mesajların analiz edilmesi ile ölçülebilmektedir. Hizmet kalitesi, sayısal derecelendirmelerden ve ilişkili ağırlıklardan kaynaklanan yönleriyle doğru bir şekilde ölçülmesi oldukça zor olan çok boyutlu bir yapıdır. Müşteri memnuniyetini doğru biçimde ölçmek için sayısal ve metinsel özellikler birleştirilmiş ve genel memnuniyetteki varyasyonlar çıkarılmıştır. Bu bağımlılıklarıyla birlikte baskın memnuniyet unsurlarının tespit edilmesini sağlamıştır. Çalışmada doğal dil işleme teknikleri kullanılmıştır. Havayolu şirketi üzerinde denenen yaklaşım, diğer tüm hizmet kalitesi boyutları düşünüldüğünde havayolu rekabetinin düşük maliyetli yönünün daha önemli olduğunu göstermiştir [14].

Günümüzde bilgisayar ve bilgi teknolojilerinin hızlı ilerlemesiyle birlikte, çevrimdışı olduğu kadar çevrimiçi olarak da çok sayıda araştırma makalesi yayınlanmakta ve yeni araştırma alanları sürekli ortaya çıkmaktadır. Özellikle çevrimdışı makalelerin kullanıcılar tarafından bulunmasında sıkıntılar yaşanmaktadır. Ayrıca bu makalelerin dokümante edilerek kategorilere ayrılması da oldukça zordur. Yapılan bir çalışmada araştırma makalelerinin benzer konulara sahip olma olasılığı araştırılmış ve sınıflara ayırabilecek bir sınıflandırma sistemi geliştirilmiştir. Önerilen sistem, her bir makalenin ve konuların özetlerinden temsili anahtar kelimeler çıkarmaya dayanmaktadır. Ardından her bir makalenin Terim Frekansı-Ters Belge Frekansı (TF-IDF) değeri referans alınarak makaleler, benzer konulardaki araştırma makaleleri ile sınıflandırmak için K-means kümeleme algoritması kullanılmıştır [15].

TF-IDF ile ilgili yapılan başka bir çalışmada, güncel terim ve konuları bulmak için terimleri, kelime segmentasyonunda bölünmemesi gereken birleşik terimler olarak ele almışlardır. Burada ayrılmış terimleri bulmak için segment içindeki konum bilgisi ve yeni terimlerin sıklık terimi birlikte ele alınmıştır [16]. TF-IDF tekniğinin kullanıldığı bir başka çalışmada belge sınıflaması için TF-IDF, LDA (Gizli Dirichlet Tahsisi) ve Doc2Vec teknikleri birlikte kullanılmıştır. Sınıflandırma sonuçlarına göre üç tekniğin birlikte kullanılması sayesinde değişen parametrelere karşı duyarlılığının daha üst seviyede olduğu görülmüştür [17]. İnsanların gerçek görüşleri ile sosyal medya platformları üzerindeki içerikler için paylaştıkları görüşlerin ne kadar örtüştüğünün araştırıldığı çalışmada, TF-IDF tekniği kullanılmıştır. Burada yorumların içerik ile olan korelasyonu bazı içerikler için yüksek iken bazılarında ise düşük olduğu görülmüştür. TF-IDF bu çalışmada küçük veri kümesi üzerinde yanlı sonuçlar vermiş, nispeten büyük veri kümelerinde ise daha doğru sonuçlar verdiği sonucuna varılmıştır [18].

Computers & Education dergisinde yayınlanan bir çalışmada yapısal konu modellemesi kullanarak eğitim teknolojilerindeki gizli konular ve eğilimleri tespit edilmiştir. Bunun için derginin kırk yılı aşkın süredir yayınlanan makaleleri incelenmiştir. Çalışmada "Bilgisayar ve Eğitim akademik topluluğu hangi araştırma konuları ile ilgileniyor?", "Araştırma konuları zaman içinde nasıl değişmiştir?" ve "Araştırmacıların temel araştırma kaygıları nelerdir?" gibi soruların cevapları aranmıştır. Bu amaçla 1976-2018 yılları arasında Computers & Education dergisinde yayınlanmış 3963 makale analiz edilmiştir. Makalelerin yıllara göre atıf sayıları, makale sayıları, hangi ülkelerden yayınların olduğu, temel istatistiksel metotlar ile analiz edilmiştir. Türkiye bu sıralamada 11. durumdadır. Türkiye'den 117 çalışma yayınlanmış, toplam 3679 atıf yapılmış ve yayınların toplam üretkenlik indeksi ise 34 olarak belirlenmiştir. Ülkemizden giden akademik çalışmalar metodolojiler ile ilgili deneysel çalışmalar, veri madenciliği ve öğretmen eğitimi konularında olmuştur. Bu çalışmada probleme yaklaşım veri bilimi yaklaşımı olarak ifade edilmektedir [19]. Bu çalışmada da veri bilimi yaklaşımı benimsenmiş ve veri seti üzerinde temel istatistiksel analizler yapılmıştır.

Doğal dil işleme, bir bilgisayarın insanın doğal dilini anlamasını amaçlayan bir yapay zeka (AI) işlemidir. NLP, metin işlemeyi basit sözdizimsel işlemenin ötesinde, insanın doğal yeteneği olan büyük ve kritik anlamsal işlemeye genişletir [9]. NLP uygulamaları tarafından insan konuşmasının amaçlanan anlamını anlama (çıkarımsallaştırma) görevini gerçekleştirmek için kullanılan birkaç farklı yapay zekas yaklaşımı vardır. Geleneksel bir kural tabanlı yaklaşım sözdizimini uygun bir semantiğe eşlemek için önceden tanımlanmış ölçütleri (koşulları) kullanan çıkarım kurallarını içerirken, bağlantıcı yaklaşım bir haritalama (veya sınıflandırma) yöntemi geliştirmek için bir öğrenme stratejisi kullanır [20, 21, 22]. Makine öğrenimi uygulamalarındaki ilerlemelerle, bir bilgisayarın bir konuşmacının amacını keşfetmesini sağlayan daha esnek, sezgisel öğrenme algoritmaları tanımlanmıştır. NLP, bilgisayarın girdi düzenleri ve çıktı sınıfları arasındaki ilgili korelasyonları otomatik olarak keşfetmesini sağlamayı amaçlamaktadır [23].

Bu çalışmada ülkemizde henüz yeni olan veri bilimi yaklaşımı temelinde lisansüstü ders önermek için veri madenciliği teknikleri kullanılmıştır. Veri bilimi kapsam olarak ilişkili olduğu alanlar belli olsa da, ders içeriğinde nelerin olması gerektiğini belirlemek disiplinler arası bir ders olduğu için zordur. Bu amaçla IEEE'nin *Data Science and Advanced Analytics(DSAA)* adındaki konferansı referans alınmış ve 2015 yılından 2019 yılına kadar konferansta sözlü sunum yapılan bildiriler incelenmiştir. Bu çalışmanın literatürdeki benzerlerinden farkı, doğal dil işleme tekniklerinin ders içeriği oluşturmak için kullanılmış olması ve izlenen metodolojinin veri bilimi yaklaşımı içermesidir. Çalışmanın geri kalanında DSAA'nın literatürdeki yeri ve veri setinin nasıl oluşturulduğu, veri setinin doğal dil işleme teknikleri ile analiz edilmesi, dersin içeriğinin oluşturulması ve elde edilen sonuçlar açıklanmaktadır.

3. MATERYAL VE METOT (MATERIAL AND METHOD)

Bu çalışmada Düzce Üniversitesi Fen Bilimleri Enstitüsü bünyesinde açılan Veri Bilimi ve Uygulamaları dersinin içeriğinin belirlenmesinde izlenen yaklaşım açıklanmaktadır. Veri bilimi ve uygulamaları dersi 3 saatlik ders olup, AKTS kredisi 7,5'tir. Ders Bilgisayar Mühendisliği A:B.D.'ında açılmıştır.

3.1. Veri Seti (Data Set)

Dersin içeriğinin belirlenmesi için IEEE'nin *Data Science and Advanced Analytics(DSAA)* adındaki konferansında sunulan bildiriler analiz edilmiştir. DSAA 2020 yılı itibariyle yedincisi düzenlenecektir. Veri seti olarak IEEExplore'da taranan 2014-2018 yılları arasındaki bildiriler dikkate alınmıştır. Bu yıllara ait bildiri sayıları, IEEExplore atıf sayıları ve konferansın etki faktörü Tablo 1'de verilmektedir.

Tablo 1. 2014-2018 yılları arası DSAA konferans bilgileri
(DSAA conference information between 2014-2018)

Yıl	Ekti Faktörü	Toplam Atıf Sayısı	Bildiri Sayısı
2014	0.79	295	92
2015	1.19	785	131
2016	2.18	566	86
2017	2.85	186	83
2018	3.74	149	79

Tablo 1'e göre en fazla bildiri 2015 yılında sunulmuştur. En fazla atıf da 2015 yılındaki bildirilere yapılmıştır. Yıllık ortalama 92 bildiri sunulmuştur. Bildiri sayılarının 2014-2018 yıllarına göre standart sapması 21,6 iken varyansı 466,9'dur. Tablo 2'de en çok atıf yapılan on bildiri verilmektedir.

Tablo 2. En çok atıf alan bildiriler
(Most cited papers)

Yıl	Atıf Sayısı (IEEExplore)	Bildiri Adı
2014	28	Detecting stock market manipulation using supervised learning algorithms
2014	88	SAR target recognition based on deep learning
2015	21	Tracking the evolution of community structures in time-evolving social networks
2015	154	Deep feature synthesis: Towards automating data science endeavors
2015	139	Anomaly detection in ECG time signals via deep long short-term memory networks
2015	132	Explaining Explanations: An Overview of Interpretability of Machine
2015	25	Time series contextual anomaly detection for detecting market manipulation in stock market
2016	106	Anomaly Detection in Automobile Control Network Data with Long Short-Term Memory Networks
2016	53	Robust Online Time Series Prediction with Recurrent Neural Networks
2016	30	Uncovering the Bitcoin Blockchain: An Analysis of the Full Users Graph

Tablo 2'de verilen bildiriler IEEExplore'un önem derecesine göre listelemesi sonucu seçilmiştir. Bildirilerin anahtar kelimeleri (yazar ve IEEE tarafından belirlenen anahtar kelimeleri) seçilmiş ve kullanılmak üzere kaydedilmiştir. Bu şekilde toplam 10732 kelime seçilmiştir. Bu kelimeler Veri Bilimi ve Uygulamaları ders içeriğinin belirlenmesinde veri madenciliği teknikleri ile analiz edilmiştir. Veri ön işlemede tüm bildirilerin başlık kısımları ve IEEE ve yazar anahtar kelimelerinden oluşan veri seti csv formatında kaydedilmiştir. Veri setinden bir kesit Tablo 3'te verilmektedir.

Tablo 3. Veri setinden bir kesit
(A section from the dataset)

ID	Bildiri Adı	Anahtar Kelimeler
pp_1	Adaptive Threshold for Outlier Detection on Data Streams	Outlier detection, threshold setting, one class learning, auto encoder, LOF
pp_2	Citizen contributions and minor heritage: feedback on modeling and visualising an information mash-up	Spatio-historical data modelling, Citizen Science, Information visualization, Knowledge Discovery, Research Methodologies, Minor Heritage
pp_3	Estimating Causal Effects On Social Networks	Causal Inference, Interference, Spillovers, Bayesian Inference, Social Impact
pp_4	Estimating Heterogeneous Causal Effects in the Presence of Irregular Assignment Mechanisms	Machine learning, Causal inference, Causal trees, Instrumental variable, Application to social science, Policy evaluation.
pp_5	Explaining Explanations: An Overview of Interpretability of Machine Learning	Machine learning theories, Models and systems, Deep learning and deep analytics, Fairness and transparency in data science

Tablo 3'teki veri seti incelendiğinde, bildiri başlığının birden fazla sözcükten oluştuğu ve anahtar kelimeler gibi ayrı ayrı anlamlı bütün halinde parçalanması gerektiği anlaşılmaktadır.

3.2. Doğal Dil İşleme Teknikleri ile Ana Konu Başlıklarının Belirlenmesi (Determination of Main Topics with Natural Language Processing Techniques)

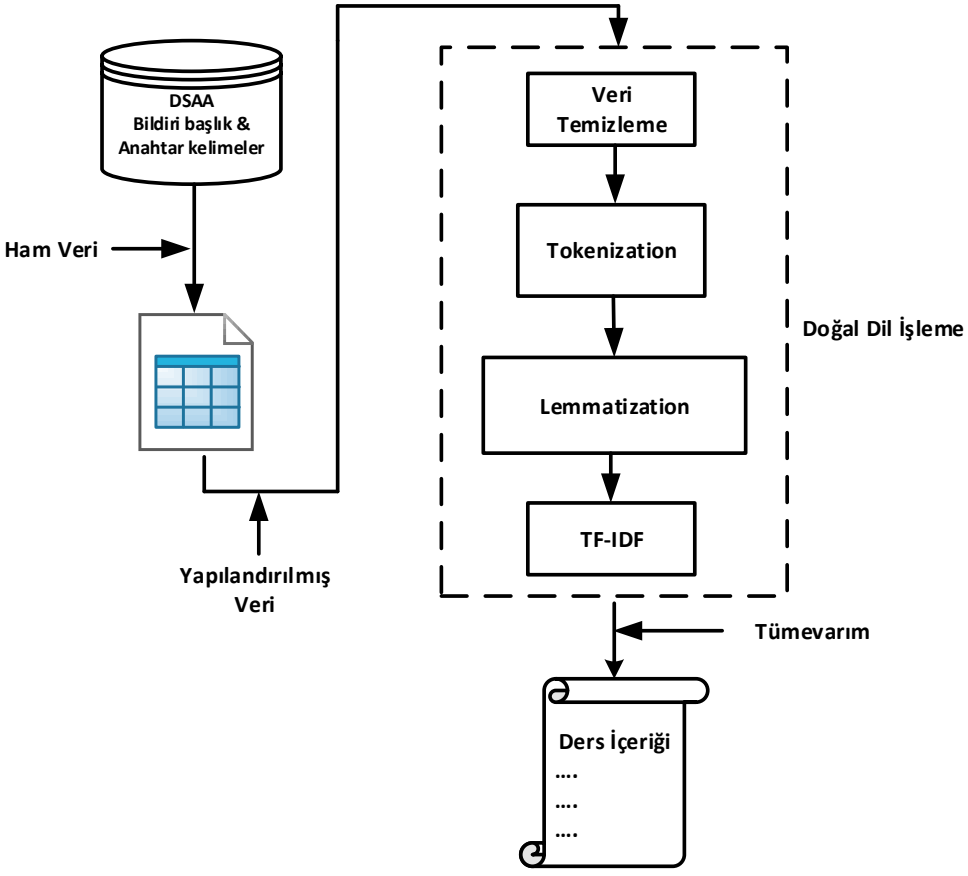
Bu çalışmada ilk olarak DSAA konferansında sunulan tüm bildiriler IEEExplore'da bulunarak, her bir bildiri bilgisi (başlık, anahtar kelimeler ve atıf sayısı) excel dosyasına alınmıştır. Şekil 1'de izlenen metodolojik yaklaşım verilmektedir. Bu veriler ham veriyi oluşturmaktadır Ham veriler id verilerek csv formatında kaydedilmiştir. Bu aşamadan sonra veriler yapılandırılmış veri formundadır.

Yapılandırılmış veriler üzerinde ilk olarak doğal dil işleme tekniklerinden veri temizleme işlemi uygulanmıştır. Veri temizleme ile verilerde bulunan ",?,- vb." özel karakterler temizlenmiştir. Ardından tokenization işlemi yapılmıştır. Tokenization işlemi cümle içindeki her bir sözcüğün indekslenmesini sağlamaktadır. Burada bildiri başlıkları birer cümle gibi olduğundan sözcüklere ayrılmıştır. Lemmatization

işlemi ile her bir sözcüğün varsa farklı biçimleri elde edilmiştir. Lemmatization sözcüklerin kökleri üzerine inşa edilen bükülmüş biçimli formlarını bulma amacıyla kullanılmaktadır [23]. Bu işlemin amacı, anahtar kelimelerin bazı araştırmacılar tarafından farklı biçimlerde kullanılmasıdır. NLP uygulamalında sıklıkla kullanılan Stemming (sözcüğün kökünü bulma) işlemi ise bu çalışmada kullanılmamıştır. Şekil 1'de de verildiği gibi çalışmanın son aşamasında sözcüklerin sözcük veri kümesindeki frekansı bulunmuş ve ilk 20 sözcük seçilmiştir. Bu yirmi sözcük tümevarım yaklaşımı ile ana başlıkları bulunarak ders içeriği belirlenmiştir.

Doğal dil işleme (NLP), bir bilgisayar programının insan dilini konuşulduğu gibi anlama yeteneğidir. NLP yapay zekanın (AI) bir bileşenidir. NLP uygulamalarının geliştirilmesi zordur, çünkü bilgisayarlar geleneksel olarak insanların kendileriyle kesin, açık ve yüksek derecede yapılandırılmış bir programlama dilinde veya sınırlı sayıda açık bir şekilde seslendirilmiş ses komutları aracılığıyla "konuşmalarını" gerektirir [24]. Bununla birlikte, insan konuşması her zaman kesin değildir - genellikle belirsizdir ve dilsel yapı argo, bölgesel lehçeler ve sosyal bağlam dahil olmak üzere birçok karmaşık değişkene bağlı olabilir. NLP'yi uygulamak için bazı araçlar bulunmaktadır [25]. Bunlar;

1. Stanford grubundan CoreNLP; bu teknoloji daha çok sözcüklerin sözdizimsel bağımlılıklar ve sözcüklerin temsil ettikleri duygusal bağlamlar ile ilgilidir.
2. Python ile en çok kullanılan NLP kütüphanesi NLTK; bu paket sözcükleri sınıflandırma, etiketleme, ayrıştırma, anlamsal çıkarımlarda bulunmak için gerekli Python fonksiyonlarına sahiptir.
3. TextBlob, kullanıcı dostu ve sezgisel bir NLTK arayüzü; bir Python kütüphanesi olup duygu analizi için kullanılmaktadır.
4. Gensim, belge benzerlik analizi için bir kütüphane; bu kütüphane metinsel bilgiyi matris formuna dönüştürerek duygu analizi yapmak için kullanılmaktadır.
5. SpaCy, performans için üretilmiş endüstriyel güçte bir NLP kütüphanesi; bu kütüphane endüstriyel uygulamaların web ortamında kullanılması için kullanılmaktadır.



Şekil 1. İzlenen metodolojik yaklaşım
(Methodological approach)

Bu çalışmada Apache Spark NLTK aracı kullanılarak sözcükler analiz edilmiş ve konu başlıkları belirlenmiştir. Apache Spark dağıtılmış sistemler üzerinde yüksek performans sunmaktadır[26]. Apache Spark açık kaynaklı dağıtılmış genel amaçlı küme bilgi işlem çerçevesidir. Spark, örtülü veri paralelliği ile tüm kümeleri programlamak için bir arayüz sağlamaktadır. Başlangıçta Kaliforniya Üniversitesi, Berkeley'in AMPLab'ında geliştirilen Spark kod alt yapısı, daha sonra onu Apache Yazılım Vakfı'na geçmiştir. Bu amaçla;

1. Veri setindeki gereksiz işaretlerin (?,:,-) kaldırılması için veri temizleme işlemi yapılmıştır.
2. Bildiri başlıklarının sözcüklere ayrılması.
3. Her bir sözcüğün varsa farklı biçimleri elde edilmiştir.
4. Sözcük tekrarına bağlı olarak her bir sözcüğün frekansının bulunması.
5. En yüksek frekansa sahip sözcüklerin listelenmesi ve ilk 20 tanesinin seçilmesi.

işlem adımları gerçekleştirilmiştir. İlk olarak veri seti içindeki gereksiz karakterler kaldırılmıştır. Herhangi bir doğal dil işleme görevinde, ham metin verilerinin temizlenmesi önemli bir adımdır. Bu işlem daha iyi özellikler elde etmeye yardımcı olurken istenmeyen kelime ve karakterlerden veri setini temizler. Veri temizleme işlemi yapılmaz ise, gürültülü ve tutarsız verilerle çalışmak zorunda kalınmaktadır. Bu işlemin amacı noktalama, özel karakterler ve sayılar gibi terimler kaldırılarak daha az bilgi veren gürültüyü temizlemektir. Bu amaçla yazılan fonksiyon aşağıda verilmektedir. Bu işlem için NLTK paketi Anaconda ortamında Jupiter için aktif edilmiştir.

```
#sözcük temizleme
def gereksiz_sil(metin, silinecek_desen):
    r = re.findall(silinecek_desen, metin)
    for i in r:
        metin = re.sub(i," ", metin)
    return metin
```

Bu fonksiyon aldığı metin içinde belirlenmiş deseni aramakta ve bulduğunda desenin yerine boşluk yerleştirmektedir. Gereksiz karakterler veri setinden kaldırıldıktan sonra bildiri başlıkları sözcüklere ayrılmıştır. Bu işlem için aşağıdaki kod ile işlem gerçekleştirilmiştir.

```
#sözcük ayırıcı
def kelime_ayirici(ifade):
    lower_ifade = ifade.lower()
    return nltk.word_tokenize(lower_ifade)
words = data.flatMap(kelime_ayirici)
print words.collect()
```

Bu fonksiyonda ilk olarak ifade eğer büyük harfli ise küçük harfe çevrilmiş ve ardından sözcüklere ayrılmıştır. Bildiri başlıkları kelimelere ayrıldıktan sonra anahtar kelimeleri de eklenerek bir sonraki adıma geçilmiştir. Üçüncü adımda aynı kökte başka sözcüklerin olup olmadığını tespit etmek için bir sözcüğün çekimsel biçimleri değerlendirilmiş ve ortak bir temel biçime indirgeme işlemi yapılmıştır. Bu işlem için geliştirilen kod aşağıda verilmektedir [13]. Bu işlem adımının amacı yazarların sözcükleri farklı kullanması durumunda sözcüğün kökü üzerinden işlem yapmaktır.

```
#sözcük lemmatization
def kokler(sozcuk):
    lemmatizer = WordNetLemmatizer()
    return lemmatizer.lemmatize(sozcuk)
lem_words = filtered_data.map(kokler)
print lem_words.collect()
```

Köklerin de bulunması ile tüm veri seti artık atomik sözcüklerden oluşmaktadır. Veri kümesi içinde her bir sözcüğün tekrarının-frekansının ne olduğunun bulunması dördüncü işlem adımıdır. Burada yöntem olarak TF(Terim Frekansı)-IDF(Ters Belge Frekansı) yöntemi seçilmiştir [26]. TF– IDF değeri, bir sözcüğün belgede kaç kez göründüğüyle orantılı olarak artar ve bazı sözcüklerin genel olarak daha sık göründüğü gerçeğinin belirlenmesine yardımcı olan tekniktir [27]. TF– IDF günümüzde en popüler terim ağırlıklandırma yaklaşımlarındandır [18]. 2015 yılında yapılan bir araştırmada, dijital kütüphanelerdeki metin tabanlı tavsiye sistemlerinin %83'ünün TF– IDF kullandığını göstermiştir [14]. Sözcüklerin frekansının bulunması için aşağıdaki fonksiyon kullanılmıştır.

```
#sözcük sayimi
def sozcukSayimi(sozcuk):
    tfidf = TfidfVectorizer(sozcuk, min_df=2,
max_features=None, stop_words='english')
    tfidf = tfidf_vectorizer.fit_transform(sozcuk)
print(tfidf.shape)
```

Bu işlem adımındaki fonksiyon da frekansı 2 ve daha az olan sözcükler dikkate alınmamıştır. Veri bilimi ve Uygulamaları adındaki dersin içeriği için 14 haftalık konu başlıklarına ihtiyaç vardır. Sözcüklerin frekansı bulunarak frekansa bağlı olarak ilk 20 kelime listelenmiştir. Tablo 4'te 20 sözcük verilmektedir.

Tablo 4'deki sözcükler dikkate alınarak tümavarım yöntemi ile ana konu başlıkları belirlenmiştir. Veri seti üzerinde yapılan denemelerde farklı büyüklüklerde seçimler yapılmış, terim frekansına bağlı olarak artan biçimde sıralandığında 20 sözcük yeterli olmuştur. Tablo 4'teki bazı sözcüklerin Tablo 5'te birebir karşılığı yoktur. Bu sözcükler Social Media, Python ve Twitter'dır. Python programlama dili olarak Veri Bilimi ders içeriğinde uygulamaların yapıldığı dildir. Ders içeriğinde yapılan uygulamalarda kullanılan veri setleri genellikle sosyal medya ve Twitter içeriklerinden oluşmaktadır.

Tablo 4. En yüksek frekansa sahip ilk 20 sözcük
(Top 20 words with the highest frequency)

Sıra No	Kelime/Sözcük	Sıra No	Kelime/Sözcük
1	Machine learning	11	Regression
2	Statistics	12	Supervised learning
3	Unsupervised learning	13	Model validation
4	Deep learning	14	Support vector machine
5	Data science	15	Logistic regression
6	Data analytic	16	Sentiment analysis
7	Neural network	17	Principal component analysis
8	Detection Anomaly/Outlier	18	Python
9	Social media	19	KNN algorithm
10	Heterogeneous	20	Twitter

3.3. Ders İçeriğinin Oluşturulması (Creating Course Content)

Araştırmalar özellikle yükseköğretim kurumlarında disiplinlerarası ders içeriği oluşturulurken öğretim üyelerinin zorluklar yaşadığını göstermektedir [24]. Bu zorluklar arasında ders içeriğinin özelleştirilmesi, öğretim formatı, ödev ve çalışma yoğunluğu, kurallar ve zorunluluklar, değerlendirme şekil ve sıklığına ilişkin sorunlar haricinde etkili kaynakların olmayışı da bulunmaktadır [25]. Ders içeriği oluşturma, bir derste olması gereken unsurlar hakkında sürekli bilgi toplama ve karar vermeyi içeren bir süreçtir [15].

En yüksek frekansa sahip anahtar kelimeler/sözcükler dikkate alınarak tümevarım yaklaşımı ile dersin ana konu başlıkları belirlenmiştir. Anahtar kelimeler/sözcükler tek başlarına bir dersin konusu olamayacak kadar alt parçaya ayrılmış/atomik olduğundan bu yaklaşım benimsenmiştir [19]. Tablo 5'te Veri Bilimi ve Uygulamaları dersinin konu başlıkları verilmektedir. Tablo 5 hazırlanırken sezgisel olarak davranılmış, dersin akışının olmasına dikkat edilmiştir. Ders kapsamındaki bazı başlıklar bilgi tecrübeye bağlı olarak yerleştirilmiştir.

Tablo 5'te verilen konu başlıkları günümüzde veri bilimi alanında sertifika veren veya yüksek lisans eğitimi veren üniversitelerin ders içerikleri ile benzerlik göstermektedir. Farklı olan yanları tamamen uygulamaların ve istenen ödevlerin Python dilinde olması ve kapsamlı konular içermesi şeklindedir. Columbia University Data Science Institute bünyesinde yürütülen master programında açılan dersler Tablo 6'da verilmektedir.

Tablo 6'daki dersler ve içerikleri Tablo 5 ile karşılaştırıldığında büyük oranda örtüştüğü söylenebilir. Columbia University Data Science Institute Master Programı tek bir dersten oluşmadığı için çok daha detaylı ve kapsamı geniş konuları da içermektedir.

Tablo 5'te *Python ile Veri Bilimini Anlama* ilk hafta işlenecek ders içeriğidir. Burada Python, veri biliminde en çok Anaconda ortamında Jupiter aracı ile kullanıldığından dersin tamamında Jupiter aracının kullanılması tercih edilmiş ve numpy, scikit-learn, panda gibi Python paketleri verilmektedir. Ayrıca neden Python'ın veri biliminde tercih edilmesi gerektiği kavramsal ve uygulamalı olarak açıklanmaktadır.

Günümüzde çeşitli web platformlarında yoğun bir şekilde veri setleri paylaşılmaktadır. Uygulama yapılırken ya da akademik çalışma için bu veri setleri kullanılabilmektedir. İkinci hafta- *Olasılık, İstatistik ve Keşifsel Veri Analizi* ile bir veri setinin nasıl hazırlanabileceği ve veri setinden daha faydalı bilgilerin istatistiksel olarak nasıl çıkarılabileceği verilmektedir. Burada temel düzeyde olasılık ve istatistik de verilmektedir.

Üçüncü hafta ders içeriğinde ise *Regresyon, Lojistik Regresyon ve Özellik Mühendisliği* bulunmaktadır. Regresyon ve lojistik regresyon günümüz veri bilimcilerinin yoğun biçimde kullandığı tekniklerdir. Özellik mühendisliği kapsamında değişkenlerin nasıl fonksiyonlara dönüştürüleceği hangi durumlarda bu işlemlerin yapılacağı uygulamalı olarak verilmektedir.

Ders planında dördüncü hafta *Veri Görselleştirme Teknikleri* için kullanılan araçlar tanıtılmaktadır. Burada Python dilinde kullanılan araçlar tanıtılmaktadır. Ayrıca bir verinin görselleştirilirken hangi grafiklerin tercih edilmesi (borsa verileri için scatter grafiği vb.) gerektiği açıklanmaktadır.

Dersin bu aşamasından sonra (5-9. haftalar, 8. hafta hariç) *makine öğrenmesi* kapsamındaki algoritmaların anlatılması ve uygulamalar yapılması tercih edilmiştir. Beşinci hafta *sınıflama algoritmaları* (destek vektör makineleri), altıncı hafta *kümeleme algoritmaları* (K-means, KNN), yedinci hafta *danışmanlı öğrenme* algoritmaları ve dokuzuncu hafta da ise *danışmansız öğrenme* algoritmaları verilmektedir. 8. hafta ara sınav olduğundan ders planında boş geçilmiştir.

Onuncu hafta- *Model Doğrulama Teknikleri ve Parametre Azaltma* bir makine öğrenmesi modelinin doğruluğunun nasıl tespit edileceği, hangi yöntemlerin hangi durumlarda kullanılabileceği uygulamalı olarak

verilmektedir. Ayrıca büyük veri ile birlikte artık neredeyse zorunlu hale gelen veri setindeki değişken sayısının nasıl azaltılabileceği (Temel bileşenler analizi) verilmektedir. Tablo 5'te on birinci haftada- *Tensorflow ve Keras* Tensorflow ve üzerine geliştirilen Keras örneklerle anlatılmaktadır. On iki ve on üçüncü haftada ilk olarak *Sinir Ağları* ardından *Derin öğrenme* verilmektedir. Burada var olan sinir ağı mimarileri de açıklanmaktadır. Var olan bir sinir ağının hangi durumlarda kullanılmasının gerektiği ve parametre ayarlamalarının nasıl yapılacağı verilmektedir.

On dördüncü haftada günümüzde akademik olarak çok çalışılan *Evrişimsel Sinir Ağı* mimarisi verilmekte ve uygulamalar yapılmaktadır. Derse öğrencinin temel bilgileri bilerek gelmesi öğrenme açısından kolaylık sağlamaktadır. Ayrıca her hafta ödev çalışmaları ile öğrenme süreci pekiştirilmektedir.

Tablo 5. Doğal dil işleme ve tümevarım yöntemi ile belirlenen Veri Bilimi ve Uygulamaları dersi konu başlıkları
(Data science and applications course topics determined by natural language processing and induction method)

Hafta	Konu Başlığı	İşleniş Şekli
1	Python ile Veri Bilimini Anlama	Anlatım ve Uygulama
2	Olasılık, İstatistik ve Keşifsel Veri Analizi	Anlatım ve Uygulama
3	Regresyon, Lojistik Regresyon ve Özellik Mühendisliği	Anlatım ve Uygulama
4	Veri Görselleştirme Teknikleri	Anlatım ve Uygulama
5	Makine Öğrenmesi: Sınıflama Algoritmaları	Anlatım ve Uygulama
6	Makine Öğrenmesi: Kümeleme Algoritmaları	Anlatım ve Uygulama
7	Makine Öğrenmesi: Danışmanlı Öğrenme Algoritmaları	Anlatım ve Uygulama
8	Vize Sınavı	Klasik Sınav
9	Makine Öğrenmesi: Danışmansız Öğrenme Algoritmaları	Anlatım ve Uygulama
10	Model Doğrulama Teknikleri ve Parametre Azaltma	Anlatım ve Uygulama
11	Tensorflow ve Keras	Anlatım ve Uygulama
12	Sinir Ağları ve Derin Öğrenme	Anlatım ve Uygulama
13	Sinir Ağları ve Derin Öğrenme	Anlatım ve Uygulama
14	Evrişimsel Sinir Ağları	Anlatım ve Uygulama

Tablo 6. Columbia university data science institu master programı dersleri [17]
(Columbia university data science institute master program courses)

Ders Adı	İçerik
Introduction to Data Science	R ile olasılık temelli makine öğrenimi
Computer Systems for Data Science	Veri analizi, temizlenmesi, Apache Spark vb.
Machine Learning for Data Science	Denetimli makine öğreniminin temel istatistiksel ilkeleri
Algorithms for Data Science	Veri düzenleme, lineer cebir vb.
Probability & Statistics for Data Science	Olasılıksal modeller, rasgele değişkenler, istatiksel çıkarımlar
Exploratory Data Analysis & Visualization	Veri görselleştirmenin temelleri ve model görselleştirme
Topics in Computer Science: Applied Machine Learning	Veri hazırlama, model seçimi ve değerlendirme de dahil olmak üzere SVM'ler, Random Forest ve Gradient boosting
Topics in Computer Science: Applied Deep Learning	Sinir ağları (DNN'ler, CNN'ler ve RNN'ler) temelleri ve TensorFlow
Topics in Computer Science: Data-Analytics	Veri oluşturma süreçleri, verilerin nasıl depolanacağı, analizin ve hesaplama yeteneklerinin bu depolamanın üzerine nasıl oluşturulacağı
Computational Models of Social Meaning	Sentiment Analysis, Emotion and Mood Analysis, Argumentation Mining vb.
Topics in Quantitative Finance: Big Data in Finance	Finanstaki gerçek büyük veriler üzerine uygulamalar
Bitirme projesi	Tez çalışması

4. SONUÇ VE ÖNERİLER (CONCLUSION AND RECOMMENDATIONS)

Bu çalışmada lisansüstü düzeyde verilen Veri Bilimi ve Uygulamaları dersinin içeriğinin oluşturulması için izlenen metodolojik yaklaşım anlatılmaktadır. Bu yaklaşıma göre IEEE'nin bu yıl yedincisini düzenleyeceği veri bilimi/analitiği (DSAA) konferansı incelenmiştir. Konferansa kabul edilen bildiriler başlıkları ve anahtar kelimeleri ile ele alınmış ve veri seti hazırlanmıştır. Ardından veri seti doğal dil işleme teknikleri ile önce bildiri başlıkları sözcüklere ayrılmıştır. Ardından tüm sözcükler biçimsel olarak analiz edilmiş ve başka çekimleri olup olmadığı incelenmiştir. Sözcüklere ayrılan başlık bilgileri ve anahtar kelimeler birleştirilerek frekans analizi yapılmıştır. Frekans analizi sonucunda ilk yirmi sözcük seçilmiştir. Seçilen yirmi sözcük tümevarım yöntemi ile üst başlıkları belirlenerek dersin konu başlıkları belirlenmiştir. Konu başlıkları dersin işlenişi dikkate alınarak sıralanmıştır. Konu başlıkları belirlenirken sezgisel hareket edilmiştir. Veri bilimi ile tamamen özdeşleşen Python, dersin ilk haftası ders içeriğine konmuştur. Bu çalışmada Apache Spark ortamında çalışan NLTK paketi kullanılmıştır. Bu platformun kullanılmasının nedeni performansının yüksek olması ve geliştirilecek olası uygulamaların hızlı bir biçimde dağıtılabilir hale getirilebilmesidir.

Lisansüstü ders önermek ve özellikle de disiplinler arası ders önermek akademisyenlerin oldukça zamanını almaktadır. Zira birçok kaynak taramak ve ilişkili alanları iyi analiz etmek gerekmektedir. Bu çalışmada pratik bir yaklaşım sunulmuştur. Columbia University Data Science Institute bünyesindeki dersler ile yapılan çalışma sonuçları örtüşmektedir. Bu şekilde özellikte hızlı değişen teknolojik koşullar dikkate alınarak geliştirilen bu prosedür, ders içeriklerinin güncellenmesi amacıyla kullanılabilir. Bu çalışmada DSAA konferansına ait tüm bildiriler indirilmiş ve tek tek incelenerek veri seti oluşturulmuştur. Bu oldukça zaman alıcı bir yöntem olmuştur. Sonraki çalışmada ise verilerin otomatik çekilmesi ve analizin daha hızlı yapılması düşünülmektedir. Bunun için çekilecek veriler (bildiri, makale, kitap vb.) belirlenmiş konu başlıkları (Tablo 4) sınıflarına atanacaktır. Makine öğrenimi teknikleri ile sınıflama çalışması yapılacaktır. Elde edilen sonuçlar istatistiksel olarak analiz edilecek ve geliştirilecek uygulamanın NLTK paketi üzerinden kullanılabilir olması sağlanacaktır. Ülkemizde veri bilimi kavramı ve kapsamı henüz yerleşmemiştir. Ancak günümüz teknolojileri düşünüldüğünde enstitüler bünyesinde Veri Bilimi programlarının açılması yerinde olacaktır.

KAYNAKLAR (REFERENCES)

[1] G. Strawn, "Data Scientist", *IT Prof.*, 18(3), 55–57, 2016.

[2] M. Kim, T. Zimmermann, R. Deline, and A. Begel, "Data scientists in software teams: State of the art and challenges", *IEEE Trans. Softw. Eng.*, 44(11), 1024–1038, 2018.

[3] C. Costa and M. Y. Santos, "The data scientist profile and its representativeness in the European e-Competence framework and the skills framework for the information age", *Int. J. Inf. Manage.*, 37(6), 726–734, 2017.

[4] V. Dhar, "Data science and prediction", *Commun. ACM*, 56(12), 64–73, 2013.

[5] F. W. Spaid and J. C. Frishett, "Incipient separation of a supersonic, turbulent boundary layer, including effects of heat transfer", *AIAA Journal*, 10(7). 1972.

[6] F. Provost and T. Fawcett, "Data Science and its Relationship to Big Data and Data-Driven Decision Making", *Big Data*, 1(1) 51–59, 2013.

[7] H. Hu, Y. Luo, Y. Wen, Y. S. Ong, and X. Zhang, "How to Find a Perfect Data Scientist: A Distance-Metric Learning Approach", *IEEE Access*, 6, 60380–60395, 2018.

[8] Internet: McKinsey, Big data: the next frontier for innovation, competition, and productivity,athttp://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_ innovation, 21.01.2020.

[9] Internet: Columbia University, Data Science Institue. Columbia University, https://datascience.columbia.edu/master-of-science-in-data-science,15.02.2020.

[10] E. Saquete, D. Tomás, P. Moreda, P. Martínez-Barco, and M. Palomar, "Fighting post-truth using natural language processing: A review and open challenges," *Expert Syst. Appl.*, 141, 112943, 2020.

[11] E. Saquete, D. Tomás, P. Moreda, P. Martínez-Barco, and M. Palomar, "Fighting post-truth using natural language processing: A review and open challenges," *Expert Syst. Appl.*, 141, 112943, 2020.

[12] M. Pejic-Bach, T. Bertoncel, M. Meško, and Ž. Krstić, "Text mining of industry 4.0 job advertisements", *Int. J. Inf. Manage.*, 50, 416–431, 2020.

[13] M. Giménez, J. Palanca, and V. Botti, "Semantic-based padding in convolutional neural networks for improving the performance in natural language processing. A case of study in sentiment analysis", *Neurocomputing*, 378, 315–323, 2020.

[14] F. Salo, M. Injadat, A. B. Nassif, A. Shami, and A. Essex, "Data mining techniques in intrusion detection systems: A systematic literature review", *IEEE Access*, 6, 56046–56058, 2018.

[15] K. A. Renn and E. R. Jessup-Anger, "Preparing new professionals: Lessons for graduate preparation programs from the national study of new professionals in student affairs", *J. Coll. Stud. Dev.*, 49(4), 319–335, 2008.

[16] I. Y. Song and Y. Zhu, "Big data and data science: what should we teach?", *Expert Syst.*, 33(4), 364–373, 2016.

[17] Y. Zhang, M. Chen, and L. Liu, **A review on text mining**, Proc. IEEE Int. Conf. Softw. Eng. Serv. Sci. ICSESS, 681–685, 2015.

[18] I. Yahav, O. Shehory, and D. Schwartz, "Comments Mining With TF-IDF: The Inherent Bias and Its Removal", *IEEE Trans. Knowl. Data Eng.*, 31(3), 437–450, 2019.

[19] X. Chen, D. Zou, G. Cheng, and H. Xie, "Detecting latent topics and trends in educational technologies over four decades using structural topic modeling: A retrospective of all volumes of Computers & Education", *Comput. Educ*., 151, 2020.

[20] L. Yao, Y. Zhang, Q. Chen, H. Qian, B. Wei, Z. Hu, "Mining coherent topics in documents using word embeddings and large-scale text data," *Eng. Appl. Artif. Intell.*, 64, 432–439, 2017.

[21] M. Pejic-Bach, T. Bertoncel, M. Meško, Ž. Krstić, "Text mining of industry 4.0 job advertisements", *Int. J. Inf. Manage*., 50, 416–431, 2018.

[22] M. Giménez, J. Palanca, and V. Botti, "Semantic-based padding in convolutional neural networks for improving the performance in natural language processing. A case of study in sentiment analysis," *Neurocomputing*, 378, 315–323, 2020, doi: 10.1016/j.neucom.2019.08.096.

[23] F. Salo, M. Injadat, A. B. Nassif, A. Shami, and A. Essex, "Data mining techniques in intrusion detection systems: A systematic literature review," *IEEE Access*, 6, 56046–56058, 2018.

[24] K. A. Renn, E. R. Jessup-Anger, "Preparing new professionals: Lessons for graduate preparation programs from the national study of new professionals in student affairs," *J. Coll. Stud. Dev.*, 49(4), 319–335, 2008.

[25] E. Ustun., "Learning Analytics and Applications in Higher Education", *Bilişim Teknolojileri Dergisi,* 13(3), 2020.

[26] H. Polat, M. Korpe, "Extracting Close Meaning Concepts from GNAT Parliamentary Minutes", *Bilişim Teknolojileri Dergisi*, 11(3), 2018.

[27] Ç. Aci, A. Çirak, "Turkish News Articles Categorization Using Convolutional Neural Networks and Word2Vec", *Bilişim Teknolojileri Dergisi*, 12(3), 2019.