Citation Count Prediction of Academic Papers (Bilimsel Makalelerin Atıf Sayısı Tahmini)

KIZILOZ, Hakan

Avrupa Bilim ve Teknoloji Dergisi
Özel Sayı, S. 370-375, Nisan 2020

Araştırma Makalesi

www.ejosat.com ISSN:2148-2683

European Journal of Science and Technology
Special Issue, pp. 370-375, April 2020

Research Article

Bilimsel Makalelerin Atıf Sayısı Tahmini*

Hakan Ezgi Kızılöz

Türk Hava Kurumu Üniversitesi, Mühendislik Fakültesi, Bilgisayar Mühendisliği Bölümü, Ankara, Türkiye (ORCID: 0000-0002-4815-9024)

(Konferans Tarihi: 5-7 Mart 2020)

(**DOI:** 10.31590/ejosat.araconf48)

ATIF/REFERENCE: Kızılöz, H. E. (2020). Bilimsel Makalelerin Atıf Sayısı Tahmini. *Avrupa Bilim ve Teknoloji Dergisi*, (Özel Sayı), 370-375.

Öz

Bilimsel makalelerin etkisini ölçmek kolay ya da tekdüze bir süreç değildir. Makalelerin atıf sayıları, etkilerinin ölçümünde önemli bir rol oynamaktadır. Öte yandan, bir makalenin atıf sayısı, makale yayınlandığı anda elde edilebilen bir veri değildir. Atıf sayısının elde edilebilmesi için makalenin yayınlanması ve toplulukta fark edilerek atıf(lar) alması, yani uzun sayılabilecek bir süre geçmesi gerekmektedir. Bu çalışmada, atıf sayısının erişilebilir olmaması problemini basitleştirdik ve bir makalenin yayınlanmasından sonraki bir yıl içerisinde en az bir atıf alıp almayacağını tahmin eden bir derin öğrenme modeli oluşturduk. Modelimizde kelime dizileri arasındaki ilişkiyi bulabilmek adına Uzun Kısa Süreli Bellek (UKSB) kullanılmaktadır. Bunun yanı sıra, bu çalışmada modelimizin makale tam metni yerine sadece özetini kullandığımızda bu durumun performans üzerindeki etkisini de analiz ediyoruz. Deneylerimizde herkese açık veri kümelerini kullanılmıştır. Makalelerin tam metni Kaggle'da bulunan bir veri kümesinde mevcuttur. Özet, üstveri öznitelikleri ve ilk yıl atıf sayıları ise Microsoft Academic Graph'tan çıkarılmıştır. Elde edilen sonuçlar, tam metin kullanımının daha yüksek doğrulukla sonuçlandığını göstermektedir. Fakat tam metin kullanıldığında modelin eğitim süresi, özet kullanıldığındaki eğitim süresine göre çok yüksek çıkmaktadır. Ayrıca, tam metinlere kıyasla makale özetleri daha kolay erişilebilir durumdadır. Son olarak, eğittiğimiz model bu makalenin ilk yayın yılında en az bir atıf alacağını öngörmektedir.

Anahtar Kelimeler: Derin Öğrenme, Uzun Kısa Süreli Bellek, Metin Madenciliği, Denetimli Öğrenme, Atıf Tahmini

Citation Count Prediction of Academic Papers

Abstract

Even though measuring the impact of scientific papers is not a straightforward process, their citation counts play a significant role in this determination. Citation count of a paper, however, is not available until the paper gets published and a substantial amount of time passes until it spreads through the community. To overcome this issue, we relax the problem by building a deep learning model that predicts whether a paper will receive at least one citation in a one-year interval after its publication. Our model employs Long Short-Term Memory (LSTM) to capture the relationship between word sequences. In our study, we also analyze the effect of using the abstract versus full-text of papers over performance. We utilize publicly available datasets in our experiments: Kaggle for the full-text of papers, and Microsoft Academic Graph for extracting the abstract, metadata features and the initial year citation counts of papers. Our obtained results show that the use of full-text leads to higher accuracy, yet with an enormous trade-off on training time. Additionally, paper abstracts are easier to access as compared to the full-text. Finally, our model predicts that this paper will receive at least one citation during its initial year of publication.

Keywords: *Deep Learning, LSTM, Text Mining, Supervised Learning, Citation Prediction*

* Bu makale ***International Conference on Access to Recent Advances in Engineering and Digitalization (ARACONF 2020)*** de sunulmuştur.

1. Giriş

Bilimsel bir makalenin atıf sayısı, o makalenin diğer makaleler tarafından kaç kere alıntılandığını göstermektedir. Dolayısıyla, bir makale dikkat çektiktçe atıf sayısı da artmaktadır. Atıf sayıları, akademik arama motorlarının sıralama algoritmalarında da kullanılmaktadır [1]. Kullanıcıların arama sonuçlarında sadece ilk sayfadaki sonuçları incelediği, daha sonraki sayfalarda yer alan sonuçları göz ardı ettiği gözlemlenmiştir [2]. Bu durum, akademide Matthew Etkisi'nin oluşmasına yol açmaktadır [3]. Matthew Etkisi, atıf sayısı yüksek olan makalelerin arama motorlarında daha üst sıralarda çıkması, dolayısıyla daha da fazla atıf almaları olarak tanımlanabilir. Sonuç olarak, atıf sayısı, makalelerin yarattığı etkiyi ölçmekte kullanılan önemli metriklerden birisi olarak kabul edilmektedir.

Bir makalenin atıf sayısı, erişimi kolay olması sebebiyle araştırmacılar, dergiler, kurumlar, vb. tarafından sıklıkla metrik olarak kullanılmaktadır. Benzer şekilde, atıf sayısı yardımıyla hesaplanan etki faktörleri ve h-indeksi değerleri, sırasıyla dergiler ve araştırmacılar hakkında bize bilgi vermektedir. Ancak bir makalenin atıf sayısı, makale yayınlandığı anda elde edilebilen bir bilgi değildir. Atıf sayısı yardımıyla bir makalenin etkisi ölçebilmek için makale yayınlandıktan sonra uzun bir süre geçmesi gerekmektedir. Bu sebeple, alanında uzman kişiler muazzam bir efor sarf ederek yayınlanmak üzere gönderilen her bir makale metnini değerlendirmekte ve editörlere ilgili makalenin değeri hakkında bir geri bildirimde bulundurmaktadır.

Bilimsel makalelerin atıf sayısı tahmini konusu uzun yıllar boyu çalışılmış bir konudur. Bazı araştırmacılar bu problemi bir sınıflandırma problemi olarak görürken, bazıları probleme regresyon problemi olarak yaklaşmaktadır. Benzer şekilde, bazı araştırmacılar sadece makalelerin üstverisi üzerinde çalışırken, bazıları makalenin tam metnini kullanmıştır. Bu çalışmada, biz de probleme bir sınıflandırma problemi olarak yaklaştık. Çözmesi zor olan bu problemi basitleştirerek bilimsel makalelerin yayınlanmalarından sonraki bir yıl içerisinde atıf alıp almayacaklarını tahmin eden bir model geliştirdik. Geliştirdiğimiz model girdi olarak yalnızca makalelerin başlığını ve tam metnini almaktadır. Sonrasında araştırmamızı geliştirerek makale tam metni yerine özetini kullandığımızda modelin performansının nasıl değiştiğini analiz ettik. Özetlemek gerekirse, bu çalışmanın motivasyonu, yeni bir makalenin etkisini daha yayınlanmadan ölçmemize yardımcı olabilecek bir sistemi tasarlamaktır.

Makalenin devamı şu şekilde düzenlenmiştir: Bilimsel makalelerde atıf sayısı tahmini yapan çalışmalar Bölüm 2'de verilmektedir. Bölüm 3'te veri önişleme adımları ve bu çalışmada önerilen model detaylıca açıklanmaktadır. Deney ortamı ve elde edilen sonuçlar Bölüm 4'te verilmektedir. Son olarak, deneylerin sonucu, makale sonucunda elde edilen bulgular ve gelecekte yapılacak çalışmalar son bölümde verilmektedir.

2. Literatür Taraması

Bilimsel makalelerin atıf sayısı tahmini ile ilgili çok çeşitli çalışmalar vardır. Bu konuda umut verici deneysel çalışmalar olmasına karşın, makalelerin aldığı atıf sayısı makalelerin etkisini belirleyebilmek adına oldukça önemli bir etken olduğundan bu problem henüz tam anlamıyla çözülebilmiş değildir.

ACM'in her yıl düzenlediği Knowledge Discovery and Data Mining (KDD)[2] konferansı, veri madenciliği ve analizi alanında en önemli konferanslardan biridir. 1997 yılından bu yana KDD konferansı ile birlikte bir veri madenciliği yarışması düzenlenmektedir [4]. 2003 yılında yapılan yarışmanın konusu arXiv e-baskı arşivi[3] üzerine özelleşmiş ve dört farklı görevde yarışma imkanı sunan bir ağ madenciliği problemiydi. Bu görevlerden ilki, seçilmiş makalelerin gelecek üç ay boyunca alacakları atıf sayılarının değişimlerinin tahmin edilmesiydi. Yarışma veri kümesinde 30119 adet makalenin LaTeX kaynak dosyaları (dolayısıyla tam metinleri) ve bu makaleler ile ilişkilendirilmiş 719109 adet atıf bilgisi vardı. Bu yarışmaya 57 yarışmacı katıldı. Yarışmayı kazanan grup, probleme bir zaman serisi yaklaşımlı algoritma uyguladı ve yarışmayı toplam 1329 L1 mesafesi ile kazandı. Hiçbir makalenin atıf sayısında değişiklik olmayacağını, yani bütün makaleler için 0 değişiklik olacağını tahmin eden grubun yarışmada 11. sırayı alması ilginç bir anekdot olarak karşımıza çıktı.

Bilimsel yayınları incelediğimizde, McKeown vd. [5] bilimsel makalelerin tam metninden çeşitli öznitelikler çıkaran ve bu öznitelikleri kullanarak bu makalenin gelecekteki etkisini tahmin eden bir sistem ortaya koymuşlardır. Önerdikleri sistemin performansını 3,8 milyon doküman üzerinde ölçmüşlerdir. Çalışmalarının öğrenme sürecinde önce başlık, yazarlarla ilişkili öznitelikler, atıf verilen makaleler hakkında bazı öznitelikler gibi yalnızca üstveri öznitelikleri kullanmışlardır. Daha sonra, makalelerin tam metinleri üzerinde varlıklar ve ilişkiler (entities and relations), argüman bölümlemesi (argumentative zoning) ve atıf duygusu (citation sentiment) yöntemlerini kullanarak çıkardıkları yeni öznitelikleri eklemişlerdir. Sonuç olarak, tam metin öznitelikleri ile üstveri özniteliklerinin birleştirilmesinin daha iyi sonuçlar verdiği gözlemlenmiştir. Yan vd. [6] makalelerin içerik, yazar ve yayın yeri hakkında birçok öznitelik kullanarak atıf sayılarını tahmin etmek üzere benzer bir metot önermişlerdir. Çalışmalarında, Bilgisayar Bilimi alanında 1,5 milyon yayın içeren ArnetMiner'ın atıf ağı veri kümesini kullanmışlardır. Modellerini dört farklı makine öğrenmesi tekniği ile test etmişlerdir. Değerlendirme metriği olarak kararlılık katsayısı (R^2) kullanmışlar ve testlerde en yüksek 0,786 değerine ulaşmışlardır. Bir sonraki yıl modellerine yeni öznitelikler ve yeni bir makine öğrenmesi tekniği ekleyerek en yüksek R^2 değerini 0.927'ye yükseltmişlerdir [7].

[2] https://www.kdd.org

[3] https://www.arxiv.org

Chen ve Zhang [8] atıf sayısı tahmini problemine bir regresyon problemi olarak yaklaşmaktadır. Çalışmalarında tahmin sürecinde Rastgele Orman (Random Forest) ile Gradyan Artırılmış Regresyon Ağaçları (Gradient Boosted Regression Trees) tekniklerini kullanmışlardır. Bunun için öncelikle Gizli Dirichlet Ayrımı (Latent Dirichlet Allocation) ve IBM Model 1 kullanarak yazar ve içerik tabanlı öznitelikler çıkarmışlardır. Sonrasında deneylerini gerçek bir veri kümesi olan KDD'nin 2003 yılındaki yarışmada kullandığı veri kümesi üzerinde yapmışlardır. Çıkarmış oldukları bütün öznitelikleri inceledikten sonra içerik tabanlı özniteliklerin tahmin sürecinde daha etkili olduklarını belirtmişlerdir. Castillo vd. [9] yazarların önceki yayınlarının bilgisini kullanarak yeni yayınlarının ilk yıllarında kaç atıf alacağını tahmin eden bir metot önermektedir. Çalışmalarında yazar-tabanlı, bağlantı-tabanlı ve soncul-tabanlı olmak üzere üç tip öznitelik çıkarmışlardır. CiteSeer'dan elde ettikleri veri kümesi üzerinde WEKA kullanarak doğrusal regresyon (linear regression) ve C4.5 tekniklerini uygulamışlardır. Deney sonuçlarına göre atıf sayısı tahmininde makale yazarlarının itibarı oldukça etkilidir. Weihs ve Etzioni [10], yazarların gelecek 10 yıl içindeki h-indeksi değeleri ile makale atıf sayılarını tahmin etmek üzere bir gradyan artırılmış regresyon ağaçları modeli önermiştir. Çalışmalarında makale üstverileri, atıf çizgesi ve eş-yazar çizgesi kullanarak 44 adet yazar ve 63 adet makale özniteliği çıkarmışlardır. Algoritmalarının ancak belirli koşullar altında var olan algoritmalardan daha iyi performans gösterdiğini belirtmişlerdir.

Ibáñez vd. [11] atıf sayısı tahminini biyoenformatik alanında yapmışlardır. Bunun için özet, dergi ve yayın tarihi bilgilerinden seçtikleri anahtar kelimelere göre öznitelikler çıkarmışlardır. Atıf sayılarını "çok az", "biraz" ve "çok" olmak üzere üç sınıfa ayırarak probleme bir sınıflandırma problemi olarak yaklaşmış ve deneylerinde WEKA'nın sağladığı birçok sınıflandırıcıyı test etmişlerdir. Belirli kelimelerin makale özetinde yer almasının atıf sayısını etkilediğini bildirmişlerdir. Livne vd. [12] yazar, kurum, yayın yeri, kaynakça ağı ve içerik benzerliğinden oluşmak üzere beş grup öznitelik çıkarmışlardır. Bilgisayar Bilimleri, Biyoloji, Kimya, Tıp, Mühendislik, Matematik ve Fizik alanlarındaki makaleler üzerine var olan ve yeni öznitelikleri kullanarak Destek Vektör Regresyon (Support Vector Regression) tekniği uygulamışlardır. Çalışmalarında 2000 yılında yayınlanan makalelerin 2005 yılında almış olacakları atıf sayısını tahmin etmeye çalışmışlardır. Özellikle Biyoloji ve Tıp alanında ünlü yazarların yayınlarının daha çok ilgi çektiğini bildirmişlerdir.

Pobiedina ve Ichise [13] bir atıf ağı oluşturmuş ve öznitelik yaratmak için gelen bağlantılar (in-degree) ile giden bağlantıların (out-degree) sayılarını kullanmışlardır. İki adet gerçek veri kümesi üzerinde Lojistik Regresyon (Logistic Regression), Destek Vektör Makineleri (Support Vector Machines) ve Ağaçlar (Trees) tekniklerini uygulayarak bir yıllık tahmin yapmışlardır. Probleme hem bir sınıflandırma problemi hem de bir regresyon problemi olarak yaklaşmışlardır. Deney sonuçlarına göre, üretilmiş olan özniteliklerin tahmin doğruluğu yönünden yazar ve yayın yeri ile ilgili özniteliklerden daha iyi performans gösterdiğini; fakat regresyon için yazar ile ilgili özniteliklerin daha çok katkı verdiğini belirtmişlerdir. Stegehuis vd. [14] derginin etki faktörü ile makalenin bir yıl sonunda almış olduğu atıf sayısı bilgisini kullanarak yayının uzun dönemde alacağı atıf sayısını tahmin etmeye çalışmaktadır. Olasılık dağılımını tahmin etmek için kantil regresyon (quantile regression) yaklaşımını kullanmışlardır. Her iki tahmincinin de tahmin sonucuna pozitif katkı sağladığını ifade etmişlerdir.

3. Materyal ve Metot

Bu çalışmada bilimsel makalelerin atıf sayılarını tahmin etmek için derin öğrenme metotları kullanılmaktadır. Bu sebeple herkesin kullanımına açık olan iki veri kümesi kullanılmıştır. Bunlardan ilki, Open Academic Society[4] internet sitesinde bulunan Microsoft Academic Graph (MAG) veri kümesidir. MAG veri kümesi 166 milyon bilimsel makaleye ait başlık, özet, yayın yılı, yazarları, anahtar kelimeleri, referansları, atıf sayısı, yayın yeri, vb. birçok faydalı bilgiyi içermektedir. İkinci veri kümesi ise Kaggle[5] internet sitesinde yayınlanmış olan Neural Information Processing Systems (NIPS)[6] konferansında yayınlanmış olan makalelere ait bilgileri içeren bir veritabanıdır. Bu veritabanı NIPS konferansında 1987 ile 2017 yılları arasında yayınlanmış olan makalelerin başlık, yazarları, özet ve tam metnini içermektedir. Bu çalışmada MAG veri kümesi, makalelerin üstverileri ile atıf bilgilerini elde etmek için kullanılırken, diğer veri kümesi ise makalelerin tam metnine erişebilmek için kullanılmıştır.

Bu bölümün devamında ilk önce iki veri kümesindeki makaleleri eşlemek ve veriyi çalışmaya hazırlamak için gereken veri önişleme adımlarından bahsedilecek, daha sonra makalelerin başlığı ve tam metnini kullanarak ilk yıl içerisinde atıf alıp almayacağını tahmin edecek olan derin öğrenme modeli tanıtılacaktır.

3.1. Veri Önişleme

Kullanılan iki veri kümesini birleştirebilmek için önce her iki veri kümesinde bulunan makaleleri doğru bir şekilde eşleştirebilmemiz gerekmekteydi. Ancak, bu eşleme işlemi kolaylıkla çözülebilecek bir problem değildi. Bunun için, başlangıç olarak MAG veri kümesinde yayın yeri "neural information processing systems" olan makaleleri bulduk. Bu sorgunun sonucunda 7442 makale elde ettik. Diğer veri kümesinde makale sayısı 7241 idi. Her ne kadar ikinci veri kümesi NIPS konferansının 2017 yılına ait makaleleri içeriyor olsa da, MAG veri kümesi 2017 yılının ortalarında hazırlanmış olduğundan konferansın 2017 yılına ait verileri içermiyordu. Bu sebeple eşleyebileceğimiz makale sayısı 6562'ye düştü. Bu makalelerden 6058 tanesini başlık üzerinden birebir eşleştirebildik. Geriye kalan 504 makaleyi teker teker inceleyip, elle işlememiz gerekti. Bunun için ilk önce başlıkların ilk üç kelimesini henüz eşlenememiş makalelerde aratıp sonuçları teker teker inceledik. Bu yöntemle 326 makaleyi daha eşleştirebildik.

[4] https://www.openacademic.ai/

[5] https://www.kaggle.com/benhamner/nips-papers

[6] https://www.nips.cc/

Kalan 178 makalenin 172 tanesini ise başlıklarda geçen ayrıştırıcı anahtar kelimeleri aratarak tespit ettik. Geriye kalan 6 makale eşleştirilemediği için çalışmadan çıkarıldı.

Makaleler eşleştirildikten sonra her bir makalenin ilk yıl içerisinde atıf alıp almadığı bilgisine ihtiyacımız vardı. Bu sebeple MAG veri kümesinde yer alan atıf ağını detaylıca tarayarak bu bilgiyi ortaya çıkardık. Bunun için MAG veri kümesinde bulunan bütün makaleleri teker teker inceleyerek eşleştirilmiş olan makalelerimizden herhangi birisine atıfta bulunup bulunmadığını tespit ettik. Atıfta bulunan makaleleri bütün bilgileriyle birlikte filtreledikten sonra, bu makalelerin yayınlandıkları yılları inceledik. Eğer atıfta bulunan makale, NIPS konferansında bulunan ve eşleştirilmiş olan makalenin yayın yılından itibaren 1 yıl içerisinde yayınlanmışsa geçerli bir atıf olarak işlenmiştir.

Son olarak, her ne kadar MAG veri kümesinde 2016 yılında düzenlenen NIPS konferansında yayınlanmış olan makalelerin bilgileri yer alsa dahi, veri kümesi 2017 yılının ortalarında yedeklenmiş olduğundan, yani 2016 makaleleri yayınlandıktan sonra henüz bir yıllık süre tamamlanmamış olduğundan, bu yıldaki veriler de çalışmadan çıkarılmıştır. Sonuç olarak çalışmadaki deneyler 6001 makalenin verisi üzerinde test edilmiştir.

3.2. Model

Bu çalışmada, bilimsel bir makalenin ilk yıl içerisinde atıf alıp almayacağını tahmin eden bir derin öğrenme modeli önerilmektedir. Bu model, üç adet girdi almaktadır: makalenin başlığı, tam metni ve yayın yılı. Makale başlığı ile tam metni, metin bazlı öznitelikler olduklarından modelimizde kullanılabilmeri için sayısal bir gösterime dönüştürülmeleri gerekmektedir. Bu dönüşüm, bahsedilecek olan birkaç aşamadan meydana gelmektedir. Öncelikle bütün noktalama işaretleri metinlerden kaldırılmıştır. Daha sonra metinlerde geçen bütün sayılar ile etkisiz kelimeler (stop words) özel atanmış birer kelimeye dönüştürülmüştür. Bütün makalelerde kullanılan bütün kelimelerin incelenmesi gerektiğinden, toplamda üç kereden az geçen kelimelerin öğrenmeye bir faydası olmayacağı düşünülmektedir. Bu sebeple, nadir görünen kelimeler de tespit edilerek özel atanmış bir kelimeye dönüştürülmüştür. Bu aşamalar sonunda geriye kalan her bir kelime, bir sayı ile eşleştirilmiştir ve sonrasında makaleler kelime dizilerinden sayı dizilerine dönüştürülmüşlerdir. Kelimelerin sayıya dönüştürülmesi işlemi makalelerin tam metinleri için ayrı, başlıkları için ayrı şekilde yapılmıştır. Yani aynı kelimenin başlıkta ve metin içerisindeki sayısal karşılığı farklı olabilir. Bütün makalelerin sayı dizilerine dönüştürülmesi işlemi tamamlandıktan sonra makaleler ve başlıklar ortalama uzunluk değerlerine kısaltılmış ya da uzatılmışlardır. Ortalama uzunluk değeri başlıklar için 8 kelime, tam metinler için ise 4335 kelimedir.

Bu çalışmada kelime dizileri arasındaki ilişkiyi yakalayabilmek için iki adet Uzun Kısa Süreli Bellek (UKSB) ağı eğitilmiştir. Bu ağlardan birisi makale başlıklarını, diğeri ise makale tam metinlerini eğitmek için kullanılmıştır. Bu iki ağın çıktıları, makalenin yayın yılı ile birleştirilir ve üzerine toplu normalleştirme (batch normalization) işlemi uygulanır. Model, iki tam-bağlı ağ katmanı (fully-connected networks), aşırı uyum problemini engellemek için bir seyreltme (dropout) katmanı ve yine iki tam-bağlı ağ katmanı sonucunda tahminde bulunmaktadır. Bahsedilen modelin çalışması, Şekil 1 ile verilen akış diyagramında gösterilmiştir.

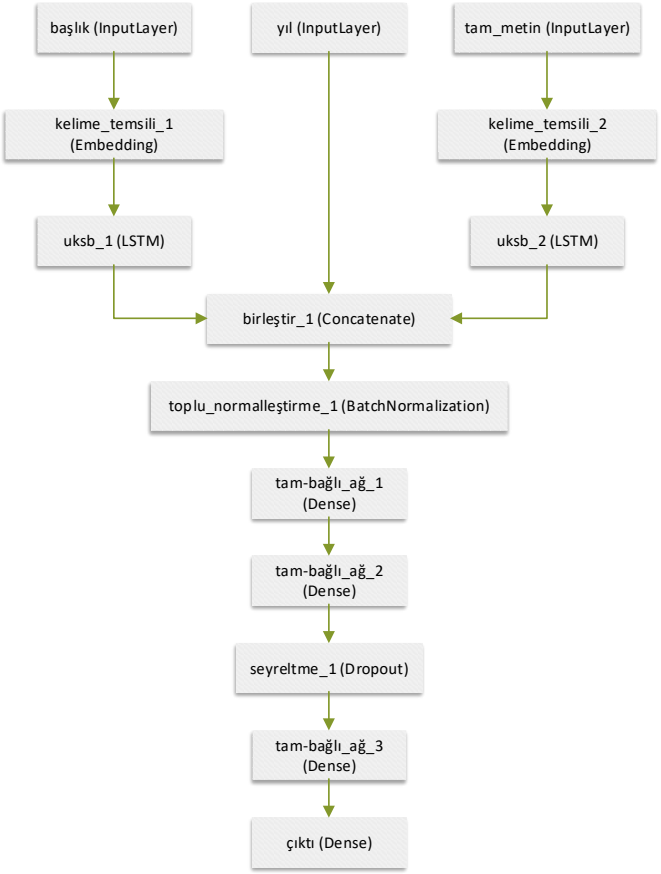
4. Araştırma Sonuçları ve Tartışma

Çalışmamızda kullanılan veriler hakkında daha fazla bilgi verebilmek adına, konferansta yayınlanan makale sayısı ile ilk yılında atıf alan ve almayan makalelerin yıllık olarak dağılım oranlarını Şekil 2'de paylaşıyoruz.

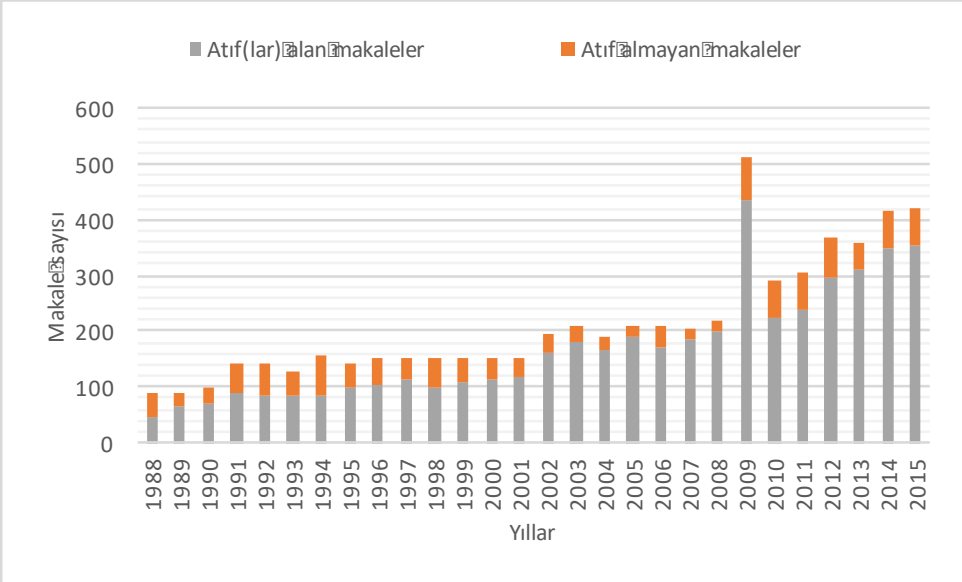
Bu çalışmanın amacı, yeni bir makalenin ilk yılında atıf alıp alamayacağını tahmin etmek olduğundan veri kümesi üzerinde çapraz geçerleme (cross validation) tekniği uygulanmamıştır. Bunun yerine, konferansın 1987 ile 2013 yılları arasını içeren ilk 26 yılında yayınlanan makaleler modelin eğitilmesi (training) için, 2014 yılında yapılan konferansta yayınlanan makaleler modelin doğrulanması için (validation) ve son olarak 2015 yılında yapılan konferansta yayınlanan makaleler modelin test edilmesi için kullanılmıştır. Sonuç olarak, bu çalışmada kullanılan eğitim, doğrulama ve test kümelerinin özellikleri Tablo 1'de detaylandırılmıştır.

Tablo 1. Çalışmada Kullanılan Eğitim, Doğrulama ve Test Kümelerinin Detayları.

	Yıllar	**Makale sayısı**
Eğitim kümesi	[1987, 2013]	5167
Doğrulama kümesi	2014	415
Test kümesi	2015	419



Şekil 1. Bilimsel Makalelerin Atıf Sayılarını Tahmin Etmek için Kullanılan Model.



Şekil 2. NIPS Konferansı'nda Yayınlanan Makale Sayılarının ve Bu Makalelerin İlk Yıllarında Atıf Alma Oranlarının Yıllara Göre Dağılımı.

Çalışmalarımız Google'ın Colaboratory servisi[7] üzerinde, ekran kartı (GPU) desteği aktifleştirilmiş şekilde çalıştırılmıştır. Bu

ayarlamalar neticesinde her bir öğrenme iterasyonunun yaklaşık 400 saniye sürdüğü gözlemlenmiştir. Test sonuçları incelendiğinde önerilen modelin bir makalenin ilk yılında atıf alabilip alamayacağını %83,15 doğrulukla tahmin edebildiği tespit edilmiştir.

Bu noktada incelemelerimizi biraz daha derinleştirerek, makalenin tam metni yerine özetini girdi olarak verdiğimizde modelin nasıl performans vereceğini gözlemlemek istiyoruz. Bunun için makale tam metni için uygulanan veri önişleme adımları makale özeti için aynen uygulanmaktadır. Bir makalenin özeti, makalenin tam metnine göre oldukça kısa olduğundan kullanılan veri miktarı epey düşmektedir. Bu düşüş, her bir öğrenme iterasyonu süresinin yaklaşık 15 saniye civarına düşmesine yol açmaktadır. Öte yandan, eğitim için kullanılan veri miktarının azalması eğitim sürecini kötü etkilemiş ve modelin tahmin performansını %82,47'ye düşürmüştür.

5. Sonuç

Bu çalışmada bilimsel bir makalenin yayınlanmasının ilk yılında atıf alıp almayacağını tahmin eden bir derin öğrenme modeli geliştirilmiştir. Bunun için herkese açık olan iki veri kümesi birleştirilmiş ve içerisinden gerekli bilgiler çıkarılmıştır. Geliştirilen modelde makalelerin yalnızca başlığı ile tam metni girdi olarak kullanılmaktadır. Daha sonra araştırmalarımız genişlemiş ve tahmin etme işini sadece başlık ve özet bilgisi, yani daha az veriyle yaptığımız durumun performansı incelenmiştir. Deney sonuçları, modelimizde tam metin kullanıldığı durumdaki tahmin performansının özet kullanıldığı durumdaki tahmin performansından bir parça daha iyi olduğunu göstermektedir. Öte yandan, özet kullanıldığında eğitim süresinin bir hayli azaldığı gözlemlenmiştir. Buna ek olarak, bir makalenin özetine ulaşmak, o makalenin tam metnine ulaşmaktan daha kolaydır. Makalelerin özetleri Microsoft Academic Graph veri kümesinde halihazırda bulunmaktadır.

Elde edilen test sonuçları, bu konuda bir ışık olduğunu göstermektedir. Bu sebeple, bu kadar az bilgiyle dahi makalelerin ilk yıllarında atıf alabilip alamayacaklarını tahmin eden böyle bir sistemin inşa edilebileceğini düşünmekteyiz.

Gelecekte modele metin öznitelikleri yanı sıra başka tipte öznitelikler de eklemeyi düşünüyoruz. Ayrıca, modelimizi kullanarak yapmış olduğumuz testleri başka konferanslarda ve dergilerde yayınlanmış makaleler üzerinde tekrarlamayı planlıyoruz.

Kaynakça

[1] J. Beel and B. Gipp, "Google Scholar's ranking algorithm: an introductory overview," in *Proceedings of the 12th International Conference on Scientometrics and Informetrics (ISSI'09)*, 2009, vol. 1, pp. 230–241.

[2] M. Jacobson. (2017) How Far Down the Search Engine Results Page Will Most People Go? [Online]. Available: https://www.theleverageway.com/blog/how-far-down-the-search-engine-results-page-will-most-people-go/

[3] R. K. Merton, "The Matthew effect in science: The reward and communication systems of science are considered," *Science*, vol. 159 (3810), pp. 58–63, 1968, American Association for the Advancement of Science.

[4] J. Gehrke, P. Ginsparg, and J. Kleinberg, "Overview of the 2003 KDD Cup," *ACM SIGKDD Explorations Newsletter*, vol. 5 (2), pp. 149–151, 2003, ACM.

[5] K. McKeown, H. Daume III, S. Chaturvedi, J. Paparrizos, K. Thadani, P. Barrio, O. Biran, S. Bothe, M. Collins, K. R. Fleischmann, and others, "Predicting the impact of scientific concepts using full-text features," *Journal of the Association for Information Science and Technology*, vol. 67 (11), pp. 2684–2696, 2016, Wiley Online Library.

[6] R. Yan, J. Tang, X. Liu, D. Shan, and X. Li, "Citation count prediction: learning to estimate future citations for literature," in *Proceedings of the 20th ACM international conference on Information and knowledge management*, 2011, pp. 1247–1252, ACM.

[7] R. Yan, C. Huang, J. Tang, Y. Zhang, and X. Li, "To better stand on the shoulder of giants," in *Proceedings of the 12th ACM/IEEE-CS joint conference on Digital Libraries*, 2012, pp. 51–60, ACM.

[8] J. Chen and C. Zhang, "Predicting citation counts of papers," in *IEEE 14th International Conference on Cognitive Informatics & Cognitive Computing (ICCI* CC)*, 2015, pp. 434–440, IEEE.

[9] C. Castillo, D. Donato, and A. Gionis, "Estimating number of citations using author reputation," in *International Symposium on String Processing and Information Retrieval*, 2007, pp. 107–117, Springer.

[10] L. Weihs and O. Etzioni, "Learning to predict citation-based impact measures," in *Proceedings of the 17th ACM/IEEE Joint Conference on Digital Libraries*, 2017, pp. 49–58, IEEE.

[11] A. Ibáñez, P. Larrañaga, and C. Bielza, "Predicting citation count of Bioinformatics papers within four years of publication," *Bioinformatics*, vol. 25 (24), pp. 3303–3309, 2009, Oxford University Press.

[12] A. Livne, E. Adar, J. Teevan, and S. Dumais, "Predicting citation counts using text and graph mining," in *Proc. the iConference 2013 Workshop on Computational Scientometrics: Theory and Applications*, 2013.

[13] N. Pobiedina and R. Ichise, "Predicting citation counts for academic literature using graph pattern mining," in *International conference on industrial, engineering and other applications of applied intelligent systems*, 2014, pp. 109–119, Springer.

[14] C. Stegehuis, N. Litvak, and L. Waltman, "Predicting the long-term citation impact of recent publications," *Journal of informetrics*, vol. 9 (3), pp. 642–657, 2015, Elsevier.

[7] https://colab.research.google.com