İLERİ MÜHENDİSLİK ÇALIŞMALARI VE TEKNOLOJİLERİ DERGİSİ

Birliktelik Kuralı Temelinde Kısaltma Genişletme

Saadet Aytaç ARPACI*[1], Banu DİRİ*[2]

*Yıldız Teknik Üniversitesi, Elektrik Elektronik Fakültesi, Bilgisayar Mühendisliği Bölümü, 34220, İstanbul, Türkiye

Araştırma Makalesi, Geliş Tarihi: 31.05.2020, Kabul Tarihi: 09.08.2020

Özet

Metinlerde yaygın olarak kullanılan kısaltmaların açık karşılıklarının bulunması, bilginin elde edilmesi ve anlaşılması açısından önemli bir gerekliliktir. Metinde kullanılan kısaltmaları, herkesin bildiği düşünülüyorsa metinde bu kısaltmaların açık karşılıklarına yer verilmeyebilir. Bununla birlikte bazen kullanılan kısaltma birden fazla açık karşılığa sahip olabilir ve bu durum anlaşılabilirliği zorlaştırır. Kısaltmalardan doğru açılımın oluşturulabilmesi halen üzerinde çalışılan bir konu olarak farklı yöntemlerle incelenmektedir. İncelenen literatürde Apriori Algoritması'nın kısaltma açılımının bulunmasına yönelik kullanımına rastlanmaması nedeniyle, bu çalışmada *PubMed* özetlerinde bulunan kısaltmaların açık karşılıklarının elde edilmesi için Birliktelik Kuralı temelinde bir yöntem önerilmiştir. İncelenen veri kümesi ve kısaltmalar dahilinde kısaltmanın birden fazla açık karşılığı olsa dahi uygulanan yöntem, ortak bir minimum destek değeri ile %87,5 farklı minimum destek değerleri ile %87,5'ten daha yüksek doğrulukla kısaltmanın açılımını bulabilmektedir.

Anahtar Kelimeler: Birliktelik kuralı, Apriori, Kısaltma genişletme.

Abbreviation Expansion on the Basis of the Association Rule

Abstract

Finding clear expansion of abbreviations commonly used in the texts is an important requirement for obtaining and understanding the information. If the abbreviations used in the text are thought to be known to everyone, these abbreviations might not be used with clear expansion in the text. However, sometimes used the abbreviation can have more than one clear equivalents and this makes understanding difficult. To be able to create correct expansion from abbreviations is still examined by different methods on as a subject being studied. In the literature reviewed, since the use of the *Apriori* algorithm for abbreviation expansion is not encountered, in this study, a method is proposed based on the association rule to obtain the clear expansion of the abbreviations in *PubMed* abstracts. Within the examined data set and abbreviations, even if the abbreviation has more than one clear expansion, the applied method can find the clear expansion of the abbreviation with 87.5% accuracy by a common minimum support value or more than 87.5% accuracy by different minimum support values.

Keywords: Association rule, Apriori, Abbreviation expansion.

[1]Sorumlu yazar saadeta99@gmail.com, [2]diri@yildiz.edu.tr

1. GİRİŞ

İngilizcede "*acronym*" kelimesinin anlamına karşılık gelen "Birkaç kelimenin baş harflerinin veya ilk hecelerinin bir araya gelmesi ile oluşan kelime" bu çalışmada kısaltma olarak anılacaktır. Kısaltmalar, bilimsel makaleler, klinik notlar ve kullanıcı sorguları gibi özellikle biyomedikal literatürde yaygın olarak kullanılmaktadır (Jin, Liu, ve Lu, 2019). Kısaltmaların açık karşılıklarıyla birlikte metin içindeki yerleşimleri Liu, Lussier ve Friedman (2001)'e göre farklılık göstermekle birlikte, çalışmamızda kısaltmanın açılımı devamında parantez içerisinde kısaltmanın olduğu özetler kullanılmıştır. Örneğin, *Prosthetic Joint Infection (PJI)*. Metin içinde bir kez açık karşılığıyla birlikte verilen kısaltmalar daha sonra metin içerisinde sadece kısaltma hali ile kullanılmaktadır. *DNA* vb. gibi çoğunlukla açık karşılıklarının bilindiği düşünülen kısaltmaların bazen metin içerisinde açılımları verilmez. Bu durum yazının anlaşılabilirliğini azaltabilir. Ayrıca, kullanılan kısaltmanın *PDA (Patent Ductus Arteriosus, Posterior Descending Artery)* (Moon, Pakhomov, ve Melton, 2012) gibi birden fazla açık karşılığının olması durumunda ise yazının anlaşılabilirliği zorlaşabilir.

Kısaltmaların belirlenmesine yönelik yapılan çalışmalardan; (Yu, Hripcsak ve Friedman, 2002) çalışmalarında bir kısaltmanın açık biçim karşılığının makalede bulunması için bir dizi kural ile çalışan bir uygulama geliştirilmiş ve kısaltmaların dört genel kısaltma veri tabanından herhangi birinde karşılığının olup olmadığı araştırılmıştır (*GenBank LocusLink*, *SWISSPROT*, *LRABR* ve *BioABACUS*). Çalışmada, tanımlanmış kısaltmalar %95 kesinlik ve %70 hassaslık ile tespit edilmiştir. Bir kelimenin anlamının, bağlamıyla uyumluluğunu belirlemek için yapılan çalışmalardan (Moon, Pakhomov, ve Melton, 2012), belirli bir pencere boyutu ve yönelimi dikkate alarak elde ettiği çevre kelimelerin anlamsal ilgisini bulmak için kullandığı denetimli öğrenme yöntemleriyle yaklaşık %90 doğruluk elde etmiştir. (Stevenson, Guo, Amri ve Gaizauskas, 2009) çalışması, kısaltmaların doğru olan genişletilmiş halini belirlemek için problemi kelime anlam ayrımı (*Word Sense Disambiguation*) sorunu olarak görmüşler ve önerdiği sistemde belirsiz kelimenin çevresindeki kelimelerden çıkardığı özellikler ile *Vektör Uzay Modeli* (*VUM*), *Naive Bayes* (*NB*) ve *Destek Vektör Makinesi* (*DVM*) gibi denetimli bir öğrenme yaklaşımıyla problemi incelemişlerdir. Çalışma, %99'a varan bir doğrulukla kısaltma açılımını belirlemiştir. (Wu, Xu, Zhang, ve Xu, 2015) çalışması ise, tıbbi metinlerde yer alan kısaltmaların açılımlarında özellik olarak sinir kelime gömmelerini (*neural word embedding*) kullanarak iki yeni kelime yerleştirme özelliği önermiş ve %95,79'luk başarı elde etmiştir.

(Li, Ji, ve Yan, 2015) çalışması, kısaltmaları belirlemek için iki kelimeden oluşan gömme modeli önermiştir. Sözcük gömme işlemi, sözcükleri sürekli ve çok boyutlu bir vektör uzayında temsil etmek olup, çalışma sözcükler arasındaki vektörel uzaklığı hesaplayarak kelime benzerliklerini farklı iki veri seti üzerinde %93 ile %95'lik bir tutarlılıkla elde etmiştir.

(Zheng, Xiao, Wang, Zhu, ve Yang, 2019) çalışması, Çince kısaltmaları tanımak için *Evrişimsel Sinir Ağı-Çift Yönlü Uzun Kısa Süreli Bellek-Şartlı Rasgele Alanlar*'a (*CNN-BLSTM-CRF*) dayanan bir sinir ağı modeli önermiştir. Alınan sonuçlar, makine öğrenmesi yöntemlerine göre daha iyi sonuç vermiştir. Derin öğrenme modellerinin farklı kombinasyonları ile yaklaşık %75 ile %79 arasında başarı elde edilmiştir. (Jin, Liu, ve Lu, 2019) çalışması, belirsiz kısaltmaların anlamlarının kesin tanımlarını bulabilmek için sadece *Çift Yönlü Uzun Kısa Süreli Bellek* modelini kullanmış ve beş farklı veri seti üzerinde yapılan testlerde en yüksek başarı macro-f1 ile 98,3 ± 3,5 olarak alınmıştır.

Birliktelik Kuralı algoritmaları (Mahgoub, Rösner, Ismail, ve Torkey, 2008; Reátegui, ve Ratté, 2019) gibi çalışmalarda, metinden çıkarılan kelimelere kelimeler arası ilişkileri bulmak için uygulanmıştır. Bu çalışmalar gibi farklı konularda da Birliktelik Kuralının geniş bir uygulama alanı bulunmaktadır. İncelenen literatürde *Apriori* algoritmasının kısaltmaların açık karşılıklarının elde edilmesine yönelik kullanımına rastlanılmadığı için çalışmamızın katkısı, Birliktelik Kuralının bu yönüyle de incelenmiş olması ve kısaltmaların açık karşılıklarının bulunması işleminde öğrenme algoritmaları haricinde de yeni bir yaklaşımın kullanılmasıdır.

2. VERİ KÜMESİ

Çalışmada kullanılmak üzere yedi farklı kısaltma ve sekiz farklı açılım (Tablo 1) belirlenmiş ve her kısaltma için içerisinde geçtiği en az altı adet doküman toplanarak bir veri seti hazırlanmıştır. Dokümanların hazırlanmasında Tablo 2'deki web adreslerinden (web adreslerinin her biri *https://www.ncbi.nlm.nih.gov/pubmed* ile başlamaktadır) 46 adet özet tıp metinleri alınmıştır. Metinlerin içerisinde en az bir adet veya daha fazla kısaltma, açılımlarıyla birlikte yer almaktadır.

Tablo 1. Çalışmada incelenen kısaltmalar

Kısaltma	Açık Karşılığı	Kısaltma	Açık Karşılığı
DASH	Disabilities of the Arm, Shoulder and Hand	CA	Carbohydrate Antigen
TEA	Total Elbow Arthroplasty	PDA	Posterior Descending Artery
PJI	Prosthetic Joint Infection	POLST	Physician Orders for Life Sustaining Treatment
ORIF	Open Reduction Internal Fixation	PDA	Patent Ductus Arteriosus

Tablo 2. Veri kümesinin elde edildiği web adresleri

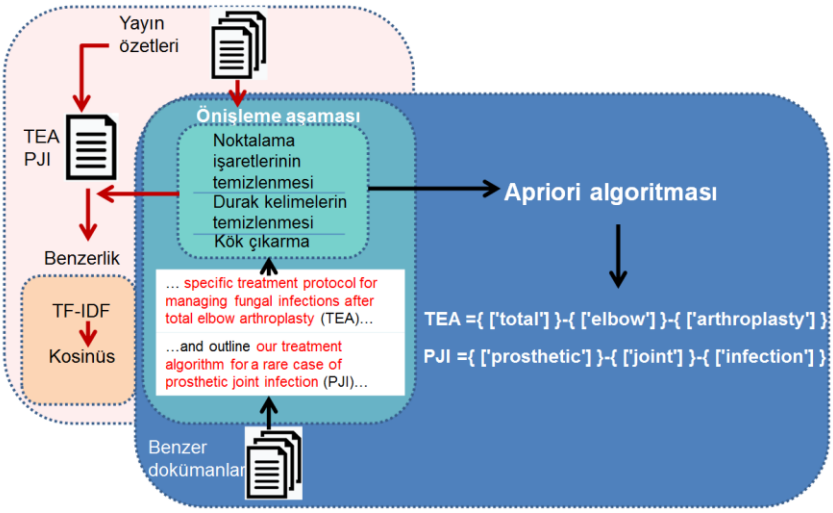
#	Web Adresi	Kısaltma	#	Web Adresi	Kısaltma	#	Web Adresi	Kısaltma
1	/31942319	DASH	16	/32191561	PJI	31	/21853470	PDA
2	/31686586	DASH, ORIF	17	/32110908	ORIF	32	/30725892	PDA
3	/31272266	DASH	18	/32158578	ORIF	33	/10642773	PDA
4	/31382027	DASH	19	/32282417	ORIF	34	/12910941	PDA
5	/30257773	DASH, TEA	20	/31949231	ORIF	35	/29298109	POLST
6	/29813168	DASH	21	/32106732	ORIF	36	/27696173	POLST
7	/31565456	TEA, PJI	22	/32057623	ORIF	37	/25441841	POLST
8	/31149971	TEA	23	/31472701	CA	38	/23865958	POLST
9	/30784119	TEA	24	/31593132	CA	39	/16948957	POLST
10	/24145266	TEA	25	/17933139	CA	40	/24743101	POLST
11	/23818030	TEA	26	/31826182	CA	41	/32211435	PDA
12	/32049563	PJI	27	/32090248	CA	42	/32175342	PDA
13	/31965585	PJI	28	/32226491	CA	43	/32143511	PDA
14	/31877146	PJI	29	/26197905	PDA	44	/32011298	PDA
15	/31857290	PJI	30	/23365974	PDA	45	/31983349	PDA
						46	/31724543	PDA

Çalışmada kullanılan kısaltmalar 2 ile 6 arasında karakterden ve büyük harflerden oluşmaktadır. Ayrıca, veri seti içerisinde *PDA* kısaltmasının iki farklı açılımına yer verilmiştir. *PDA*'nın ilk açılımı olan *Posterior Descending Artery* [29-34] dokümanlarında, ikinci açılımı olan *Patent Ductus Arteriosus* [41-46] dokümanlarında bulunmaktadır. İleriki çalışmalarda veri kümesi aynı kurallar dikkate alınarak genişletilebilir.

3. ANALİZ METODU

Önerilen yöntemin ilk aşamasında, veri kümesinin içerisindeki herhangi bir dokümanda yer alan kısaltmalar belirlenerek hangi kısaltmaların açık karşılıklarının bulunacağı bilgisi alınmıştır. (Medical Abbreviations, 2020)'deki İngilizce tıbbi kısaltmalar sözlüğünde, küçük harf içeren *kg=kilogram, syr.=syrup, Se=selenium* gibi kısaltmalara göre, *MRI=Magnetic Resonance Imaging, DI=Diabetes Insipidus* gibi çoğunlukla hastalık veya teşhis cihazlarını tanımlayan ifadeler sözlüğün yaklaşık %70'ini oluşturduğundan, bu tür yaygın ifadelere odaklanmak için kısaltmaların seçiminde 2 ile 6 karakter uzunluğunda ve büyük harflerden oluşmuş olmaları kriter olarak belirlenmiştir.

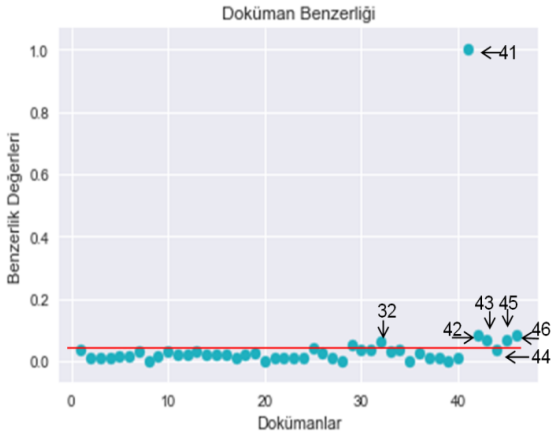
İkinci aşamada, kısaltmaları elde edilen kaynak dokümanın veri kümesindeki diğer dokümanlarla benzerlikleri ölçülmüştür. Bu işlem öncesi dokümanlar ön işlem adımı ile noktalama işaretlerinden, metin içinde sıkça tekrarlanan ve tek başına kullanıldıklarında anlam taşımayan durak kelimelerden (a, an, the, vb.) temizlenir, kök çıkarma (*lemmatizasyon*) işlemi ile kelimelerin köklerinin elde edilmesi sağlanır.



Şekil 1. Sistemin akış şeması

Çalışmamızda kelimenin ön veya son ekini atarak kök haline çeviren gövdeleme (*stemming*) işlemine (örneğin, studies, was, studying→studi, wa, studi) kıyasla, kelimeleri sözlükteki kökenine göre elde eden kök çıkarma işlemi (örneğin, studies, was, studying→study, be, study) tercih edilmiştir. Ön işlem aşamasında *NLTK* kütüphanesinden (Bird, Tan, Garrette, Ljunglöf, Nothman, Korobov, ve Dimitriadis, 2020) yararlanılmıştır. Benzerliği ölçülecek tüm dokümanlar için ön işlemin yapılmasının amacı, dokümanları kendilerine özgü kısaltma ve açık karşılıkları açısından ayırabilmektir. Önerilen sistemin genel akışı Şekil 1'de verilmiştir.

Çalışılan veri kümesi içinde "*DASH*, *TEA*, *PJI*, *ORIF*, *CA*, *PDA*, *POLST*" olmak üzere 7 kısaltma yer almaktadır. Her kısaltmanın belirli bir konuya özgü olması, aynı kısaltmaların bulunduğu dokümanları benzerliklerine göre ayırmakta tamamen olmasa da yardımcı olmaktadır. Özellikle *PDA* gibi birden fazla açık karşılığı olan kısaltmalarda doküman benzerliklerinin bulunması, işlemi kolaylaştırmaktadır. Dokümanların benzerliği, *tf-idf* değeri ve *kosinüs* açısının hesaplanmasıyla elde edilmiştir. Terim frekansı (*tf*), bir terimin bir dokümandaki geçiş sıklığı olup, bir dokümanda çok sayıda olan bir kelimenin o doküman için belirgin bir değere sahip olduğu söylenebilir. Ancak, diğer dokümanlarda da bu kelimenin geçmesi kelimenin değerini ölçmek açısından daha önemlidir. Ters doküman frekansı (*idf*) hesaplanarak kelimenin tüm dokümanlar bazında değeri elde edilir. Bu iki ölçütün birleşimiyle elde edilen *tf-idf* değeri, kelimenin dokümanların tümünde geçmesi durumunda en düşük değerini alır. Dokümanların *tf-idf* vektörel değerleri kullanılarak dokümanlar arasındaki benzerlik, *kosinüs* benzerliği ile hesaplanır. Dokümanlar aynı ise, aralarındaki açının sıfır olması nedeniyle benzerlik değeri 1, benzer ise bire yakın olur. Dokümanların benzerliğinin bulunması için gerekli *tf*, *idf*, *kosinüs* değerlerinin hesaplanmasında *scikit-learn* kütüphanesinden (Cournapeau vd., 2020) faydalanılmıştır. *PDA* kısaltmasının bulunduğu 41. dokümanın diğer dokümanlarla olan benzerlik sonucu Şekil 2'de sunulmaktadır. Şekil 2'de grafiğin altında 41. dokümana en benzer olan dokümanlardan en az benzer olanlara doğru dokümanların sıralaması da sunulmuştur.



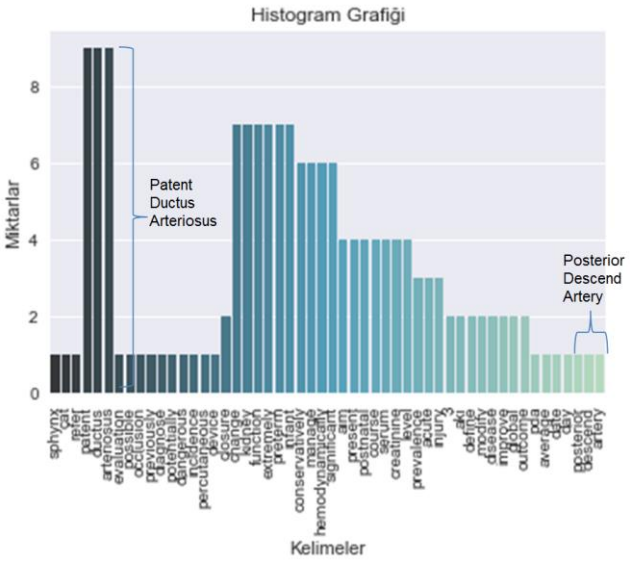
41 nolu dokümana en benzer olanlardan az benzer olanlara doğru sıralama:
41, 42, 46, 45, 43, 32, 29, 25, 31, 34, 44, 30, 1, 13, 33, 10, 7, 19, 26, 36, 16, 14, 11, 15, 12, 18, 6, 5, 9, 37, 40, 2, 3, 21, 27, 4, 17, 24, 22, 38, 23, 8, 20, 28, 35, 39
'PDA' kısaltması için alınan dokümanlar: 41, 42, 46, 45, 43, 32

Şekil 2. 41. Dokümanın diğer dokümanlarla olan benzerliğinin grafiksel gösterimi

Şekil 2'de de görüldüğü gibi *PDA=Patent Ductus Arteriosus'* un bulunduğu 41. doküman 44. dokümanı daha benzer olarak bulması gerekirken 32. dokümanı benzer bulmuştur. Bu durum gibi diğer kısaltmalar için de yapılan benzerlik işlemi, farklı kısaltmaya sahip olan dokümanların benzer olarak algılandığı bir sonuç oluşturabiliyordu. Aranılan kısaltmaya sahip dokümanlar sıralandığında çoğunlukla ilk 25 içinde yer almıştır. Bu nedenle aranılan kısaltmanın bulunduğu dokümanların seçimi genel olarak bu sınır içerisinde yapılmıştır. Bu sınır değerin verilmesi veri kümesinin daha büyük olması durumunda bir avantaj sağlamaktadır. *PDA* kısaltmasının birden fazla açık karşılığı olduğu için ilk 25 dokümandan seçim yapmak, diğer açık karşılığa sahip dokümanların da alınmasına neden olacağından bu kısaltmaya sahip dokümanların ilk 6'sı için inceleme yapılmıştır. *Apriori* algoritmasının temel mantığına göre kelime sıklığının yüksek olması, bu kelimelerin oluşturacağı alt kelime kümelerinin de görünme ihtimalini yükseltir, şeklindedir. Bu anlamda algoritmaya verilecek verilerin elde edilmesi için aranılan kısaltmaya özgü benzer dokümanlar ile devam eden işlemler yapılmıştır. Veri kümesindeki dokümanlarda kısaltma ve açık karşılıkları Şekil 1'de görüldüğü gibi parantez içinde bir kısaltma ifadesi ve parantezin sol tarafında da kısaltmaya ait açık karşılıkları verilmiş şekildedir. Doküman içindeki kısaltmaların bu açık karşılıklarına ulaşabilmek adına "(" işareti öncesi 11 eleman alınıp bu kelime grupları ön işlem basamağı ile noktalama işaretlerinden, durak kelimelerden temizlenmiş, kök çıkarma işlemi ile kelimelerin sözlükteki köklerinin elde edilmesi sağlanmıştır. Kelime grupları ön işlem basamağı ile daha belirgin kelime grupları haline getirilmiştir. *PDA* kısaltmasının bulunduğu 41, 42, 46, 45, 43 ve 32. dokümanlardan elde edilmiş kelimelere ait histogram grafiği Şekil 3'te sunulmaktadır.

Her benzer doküman için elde edilen bu kelime grupları temelinde Birliktelik Kuralı uygulanmıştır. Şekil 4'te genel olarak ifade edilmiş sözde koda dayanan *Apriori* algoritması çalıştırılmıştır. İncelenen kısaltmalar için algoritmanın üç adım ilerlemesi yeterli olmuştur.

Birliktelik Kuralına göre kelime-1 varsa kelime-2 ile birlikte ortaya çıkacağı, *kelime-1 → kelime-2* ifadesi ile belirtilir. Eşik destek değerinin üstündeki öğe alt kümelerinin elde edildiği algoritmada güçlü birliktelik kuralları oluşturmak için güven değerinin de yüksek olması beklenir. Destek ve güven değerleri için denklem 1 ve denklem 2 kullanılmıştır.



Şekil 3. 41, 42, 46, 45, 43 ve 32. dokümanlardan elde edilmiş kelimelere ait histogram grafiği

k=adım
L_k=k. adımdaki sık geçenler kümesi
Basla
L_1={sık geçenler-1 kümesi}
k←2
while L_{k-1}!=∅ do
C_k←aday kelime kümesi (L_{k-1})
C_k.frek←aday kelime kümesi frekansı (C_k)
L_k←C_k yı eşik desteğe göre ayırma (C_k.frek, mindestek)
k←k+1
return L
Son

Şekil 4. Apriori algoritması sözde kodu

$$Destek\ (kelime-1 \rightarrow kelime-2) = \frac{\alpha}{\varepsilon} \quad (1)$$

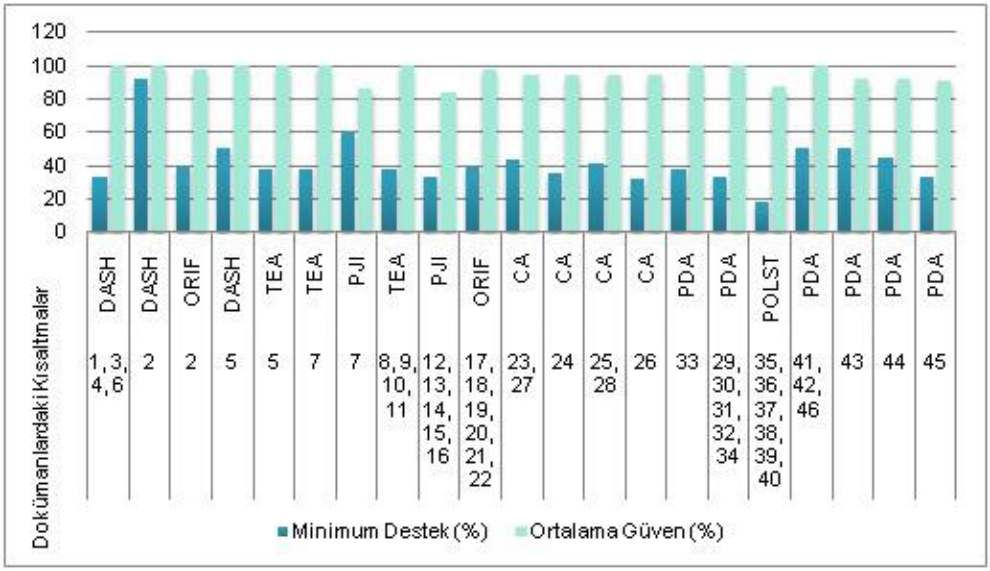
$$Güven\ (kelime-1 \rightarrow kelime-2) = \frac{\alpha}{\varphi} \quad (2)$$

α, kelime-1 ve kelime-2'nin bir arada bulunduğu küme sayısı; ε, kaynak kısaltmayı içeren benzer dokümanlardan elde edilen toplam küme sayısı; φ, kelime-1'in bulunduğu küme sayısını ifade etmektedir.

4. DENEYSEL SONUÇLAR

Tablo 3. Kısaltmaların çalışmadan elde edilen açık karşılıkları

Doküman No	Kısaltma ve Açık Karşılığı
1, 2, 3, 4, 5, 6	DASH={['disability']}-{['arm']}-{['shoulder']}-{['hand']}
5, 7, 8, 9, 10, 11	TEA={['total']}-{['elbow']}-{['arthroplasty']}
7, 12, 13, 14, 15, 16	PJI={['prosthetic']}-{['joint']}-{['infection']}
2, 17, 18, 19, 20, 21, 22	ORIF={['open']}-{['reduction']}-{['internal']}-{['fixation']}
23, 24, 25, 26, 27, 28	CA={['carbohydrate']}-{['antigen']}
29, 30, 31, 32, 33, 34	PDA={ ['posterior']}-{['descend']}-{['artery']}
35, 36, 37, 38, 39, 40	POLST={['physician']}-{['order']}-{['life']}-{['sustain']}-{['treatment']}
41, 42, 43, 44, 45, 46	PDA={['patent']}-{['ductus']}-{['arteriosus']}



Şekil 5. Her kısaltma için gereken minimum destek ve ortalama güven değerleri

Metin içindeki kısaltmanın açık karşılığının kural tabanlı bulunmasına yönelik yapılan çalışmamızda elde edilen kısaltma açılım sonuçları Tablo 3'te sunulmaktadır. Sistem tüm kısaltmaların açık karşılığını farklı minimum destek değerleriyle elde edebilmiştir. Tüm kısaltmalar için ortak kullanılan %50 minimum destek değeri ile açık karşılıkları aranan 8 kısaltmanın (*POLST* kısaltması hariç) 7'si için sonuca ulaşılmıştır. Bu anlamda sistemin genel doğruluğu %87,5'tir. *POLST* kısaltması için %18'lik bir minimum destek değeri kullanılmıştır.

Apriori algoritması ile kısaltmaların elde edilmesi için gereken minimum destek değerleri ve oluşan kurallar dahilinde elde edilen güven değerlerinin ortalaması da Şekil 5'teki grafikte sunulmaktadır.

Çalışmamızda 2 harfli (*CA*), 3 harfli (*TEA, PJI, PDA*), 4 harfli (*DASH, ORIF*), 5 harfli (*POLST*) kısaltmaların, aynı dokümandaki birden fazla kısaltmanın (örneğin, *DASH-ORIF*, *DASH-TEA*, *TEA-PJI*), aynı harf dizilimine fakat farklı açılımlara sahip (örneğin, *PDA*) olan kısaltmanın açık karşılıkları bulunmuştur. Dokümanlardan elde edilen kelime gruplarının *Apriori* algoritmasına verilmeden önce ön işlem aşamasında çıkarılan noktalama ve durak kelimeleri *DASH* kısaltmasının açık karşılığında (Disabilities *of the* Arm, Shoulder *and* Hand) yer alan kelimelerin işleme alınmamasına neden olmuştur.

Aynı doküman (örneğin, 2, 5, 7) içerisindeki iki kısaltmanın açılımlarının elde edilmesinde minimum destek α değeri her iki kısaltmanın açık karşılıklarının bulunabilmesine göre ortak verildiğinden grafikte 2 ve 5 nolu dokümanlardaki *DASH* kısaltması ile 7 nolu dokümandaki *PJI* kısaltması daha yüksek minimum destek değeri ile işleme alınmıştır.

Aynı harf diziliminde farklı açılımlara sahip kısaltmaların (örneğin, *PDA*) bulunduğu dokümanlar benzerlik işlemiyle ayrıldığında kısaltmanın aynı açılımına sahip diğer dokümanların yanında farklı açılımına sahip dokümanlarda işleme eklenebilmiştir bu durum minimum destek değerini artırma gerekliliğine neden olmuştur. Aynı kısaltmayı içeren benzer dokümanlar üzerinden işlem yapılması nedeniyle Şekil 5'te görüldüğü gibi *ORIF* kısaltmasında aynı sabit minimum destek değerine göre ortak sonuç alınmıştır, farklı dokümanların işleme eklenmesi ile *CA* kısaltmasında farklı minimum destek değerinin verilmesi gerekmiştir.

Kısaltmaların açık karşılıklarının bulunması işleminde, elde edilen ortalama güven değerleri *PJI* ve *POLST* için %80 üzerinde, *CA* ve *ORIF* için %94 üzerinde, *PDA*, *DASH* ve *TEA* için %100 olarak bulunmuştur.

Çalışmada tüm veri seti içinden dokümana benzer ilk 25 dokümana yönelik bir sınır değer üzerinden işlemlerin devam etmesi, veri seti içine eklenecek farklı kısaltmaların (özellikle birden fazla açık karşılığa sahip) bulunmasında bir kısıt olarak görülebilir, bu durumda tüm veri seti için en uygun sınır değeri/değerleri belirlenerek işlemler yapılabilir.

5. SONUÇ

Bu çalışmada *PubMed* özetlerinde bulunan kısaltmaların açık karşılıklarının elde edilmesine yönelik Birliktelik Kuralı temelinde bir yaklaşım sunulmuştur. İki ile altı karakter arasında ve büyük harften oluşan kısaltmaların açık karşılıkları, bir dokümanda bir veya birden fazla kısaltma olması durumunda bulunabilmiştir. Birden fazla açık karşılığa sahip olması nedeniyle açılımının bulunması sorun olan kısaltmaların doğru açılımları *Apriori* algoritması ile yapılabilmiştir. Önerilen yöntem tüm kısaltmaların açık karşılıklarını farklı minimum destek değerleriyle %100'lük bir başarı ile elde edebilmiştir. Ortak belirlenen bir minimum destek değerine göre işlem yapıldığında elde edilen başarı %87,5 olmuştur. Kısaltma açılımlarının bulunması için kullanılan öğrenme algoritmalarının eğitim için uzun zaman gerektirmesi nedeniyle önerilen bu yöntem ile daha kısa sürede sonuca ulaşıldığından alternatif olarak önerilebilir.

KAYNAKLAR

Bird, S., Tan, L., Garrette, D., Ljunglöf, P., Nothman, J., Korobov, M., ve Dimitriadis, A. (2020). Natural Language Toolkit. 31 Mayıs 2020 tarihinde https://www.nltk.org/ adresinden erişildi.

Cournapeau, D., Brucher, M., Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V.,… Boisberranger, J. (2020). scikit-learn. 31 Mayıs 2020 tarihinde https://scikit-learn.org/stable/ adresinden erişildi.

Jin, Q., Liu, J., ve Lu, X. (2019). Deep Contextualized Biomedical Abbreviation Expansion. 31 Mayıs 2020 tarihinde https://arxiv.org/pdf/1906.03360.pdf adresinden erişildi.

Li, C., Ji, L., ve Yan, J., (2015). *Acronym Disambiguation Using Word Embedding*. Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence (ss. 4178–4179).

Liu, H., Lussier, Y. A., ve Friedman, C., (2001). *A study of abbreviations in the UMLS*. Proc AMIA Symp (ss. 393-397).

Mahgoub, H., Rösner, D., Ismail, N., ve Torkey, F. (2008). A Text Mining Technique Using Association Rules Extraction. 31 Mayıs 2020 tarihinde http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.212.8624&rep=rep1&type=pdf adresinden erişildi.

Medical Abbreviations, (2020). 31 Mayıs 2020 tarihinde https://www.tabers.com/tabersonline/view/Tabers-Dictionary/767492/all/Medical_Abbreviations adresinden erişildi.

Moon, S., Pakhomov, S., ve Melton, G.B., (2012). *Automated Disambiguation of Acronyms and Abbreviations in Clinical Texts: Window and Training Size Considerations.* AMIA Annu Symp Proc. (ss. 1310–1319).

Reátegui, R., ve Ratté, S., (2019). *Analysis of Medical Documents with Text Mining and Association Rule Mining*. International Conference on Information Technology & Systems (ss.744–753).

Stevenson, M., Guo, Y., Amri, A., ve Gaizauskas, R. (2009). Disambiguation of Biomedical Abbreviations. 31 Mayıs 2020 tarihinde https://dl.acm.org/doi/10.5555/1572364.1572374 adresinden erişildi.

Wu, Y., Xu, J., Zhang, Y., ve Xu, H., (2015). *Clinical Abbreviation Disambiguation Using Neural Word Embeddings.* Proceedings of the 2015 Workshop on

Biomedical Natural Language Processing (BioNLP 2015) (ss.171–176).

Yu, H., Hripcsak, G., ve Friedman, C. (2002). Mapping Abbreviations to Full Forms in Biomedical Articles. *Journal of the American Medical Informatics Associatio n*, *9*(3), 262–272.

Zheng, J., Xiao, X., Wang, B., Zhu, Y., ve Yang, L. (2019). A New Method for Abbreviation Prediction viaCNN-BLSTM-CRF. 31 Mayıs 2020 tarihinde https://iopscience.iop.org/article/10.1088/1742-6596/1267/1/012001 adresinden erişildi.