TIPTEKNO'18
TIP TEKNOLOJİLERİ KONGRESİ
8-10 Kasım 2018 / Gazi Magosa / KKTC

Türkçe radyoloji raporlarının metin madenciliği kullanılarak otomatik sınıflanması
Automated categorization of Turkish radiology reports using text mining

Tuğberk Kocatekin
Bilgisayar Mühendisliği
Bahçeşehir Üniversitesi
İstanbul, Türkiye
tugberk@gmail.com

Devrim Ünay
Biyomedikal Mühendisliği
İzmir Ekonomi Üniversitesi
İzmir, Türkiye
devrim.unay@ieu.edu.tr

***Özetçe* —Metin madenciliği güvenlikten medya ve pazarlamaya kadar farklı uygulama alanları olan popüler bir araştırma konusudur. Metin madenciliği biyomedikal alanda özellikle, radyoloji uygulamalarındaki verimin artması için vazgeçilmez ancak raporların serbest metin içermesi ve biçimlerinin belirgin bir yapıya sahip olmaması nedenleriyle zorlu bir problem olan, radyoloji raporlarının sınıflandırılması problemine uygulanmıştır. Radyoloji raporlarında madencilik konusunda literatürdeki çalışmaların çoğu İngilizce metinlere odaklanmıştır; Türkçe raporları hedefleyen araştırma sayısı oldukça azdır. Dolayısıyla bu çalışmamızda rutin klinik uygulamadan elde edilen Türkçe radyoloji raporlarının otomatik sınıflandırılması için metin madenciliğini kullanmayı öneriyoruz. Raporlardaki üst ve alt bilgileri tespit ederek kaldırıyor, raporun kalan metnine frekans analizi uyguluyor, öğrenme veri kümesini kullanarak her anatomik bölgeye özel o bölgeyi diğerlerinden ayıracak terimlerden meydana gelen bir sözlük oluşturuyor ve bu sözlük temelli yaklaşım ile öğrenme kümesinde olmayan yeni raporların sekiz anatomik bölgeye göre sınıflandırılmasını gerçekleştiriyoruz. Önerdiğimiz yaklaşım 161 raporluk test kümesinde %96.3'ün üzerinde bir başarıma sahiptir.**

Anahtar Kelimeler*—*Radyoloji raporu, metin madenciliği, sözlük-temelli, anatomik sınıflandırma, Türkçe.

***Abstract*—Text mining is a popular research topic with application areas ranging from security to media and marketing. More specifically, text mining has been applied in the biomedical field for categorization of radiology reports, which is essential for improved efficacy in radiology practice but a challenging task due to the free-text and unstructured format of the reports. State-of-the-art in radiology report mining has mostly focused on English text, while studies on Turkish reports are scarce. Accordingly, in this work we propose to employ text mining for automatic categorization of Turkish radiology reports acquired from routine clinical practice. We remove header and footer of the reports, apply frequency analysis on the remaining report text, for each anatomical region construct a dictionary of discriminative words from a training set, and realize dictionary-based categorization of unseen reports into eight anatomic regions. Accuracy of the proposed solution is measured as 96.3% over a 161-reports test set.**

Keywords*—*Radiology report, text mining, dictionary-based, anatomic categorization, Turkish.

978-1-5090-2386-8/16/$31.00

I. GİRİŞ

Medikal alanda gittikçe artan miktarlarda üretilen veri (örneğin betimleme içeren kayıtlar, görüntüler, raporlar, test sonuçları vb.) etkin ve verimli depolama, sınıflama ve erişim çözümlerini gerekli kılmaktadır [1], [2]. Görüntüleme verilerinin uzman tarafından yorumlanmış metinsel açıklamalarını içeren radyoloji raporları da bu duruma istisna değildir [3], [4]. Radyoloji raporlarının hem talep eden hekim için benzer biçimli ve kullanıcı dostu olması hem de yorumlayan radyoloğun hatalarını azaltması ve veri madenciliği uygulamalarına olanak sağlaması amacıyla er ya da geç daha hasta-odaklı, yapılandırılmış ve standartlaştırılmış hale geleceği beklenmektedir [5].

Yukarıda bahsettiğimiz biyomedikal bilgi bombardımanı ve bu bilginin şifreli veri tabanlarında düzenlenmesi ile ilgili artan ihtiyaç nedeniyle etkili ve doğru metin madenciliği yapabilen çözümler gereklidir [6]. Biyomedikal alanda metin madenciliğinin başarılı uygulamaları vardır [7]–[10]. Özellikle, raporların biçimsel olarak yapılandırılmamış ve organize olmayan doğası nedeniyle zor bir problem olan radyoloji raporlarının otomatik analizi ve sınıflandırılması için metin madenciliği yaygın olarak kullanılmıştır [11]. Dolayısıyla radyoloji raporlarının madenciliği çoğunlukla doğal dil işleme yöntemleri kullanılarak gerçekleştirilmektedir [12]–[14]. Doğal dil işleme dışında yaklaşımlar içeren çözümler de bulunmaktadır [15]–[19]. Türkçe radyoloji raporlarının bilgisayar yardımı ile analizi konusunda literatürde dikte edilmiş veriden rapor oluşturma [20], serbest formatlı metin içeren raporları yapılandırılmış hale çevirmek [21], medikal kayıtların hastalık temelli sınıflandırılması [22] gibi az sayıda çalışma vardır.

Bu konuda yaptığımız önceki çalışmamızda [23] raporun üst/alt bilgilerinin ayrılması, basit bir köke dönüştürme, frekans analizi ve sözlüklerle karşılaştırma adımlarından oluşan bir çözüm önermiş ve sistemin başarımını 66-raporluk dar bir verisetinde %84 olarak ölçmüştük. Bu çalışmamızda önceki sistemimizi Zemberek doğal dil işleme kütüphanesini kullanarak daha gürbüz bir köke dönüştürme yöntemi ile geliştirdik, ve ilkeli ve kapsamlı bir analiz gerçekleştirerek

sistemin başarımını 230-raporluk daha geniş bir verisetinde ölçtük.

Radyoloji raporlarının otomatik madenciliği ve sınıflandırılması radyoloji uygulamalarının verimini artırmak için önemli ancak zor bir işlemtir. Yukarıda sunulan literatür taramasından görüldüğü üzere Türkçe radyoloji raporlarının sınıflandırılması konusundaki çözümler yok denecek kadar azdır. Ayrıca, radyoloji raporlarının – dili gözetilmeksizin - anatomik bölge temelli sınıflandırılması konusu bildiğimiz kadarıyla halen üzerinde az çalışılmış bir alandır. Dolayısıyla sunulan çalışma rutin klinik uygulamadan elde edilen Türkçe radyoloji raporlarının anatomik bölgeye göre sınıflandırılması için sözlük temelli otomatik bir sistem önermekte ve bu ihtiyacı gidermeyi amaçlamaktadır.

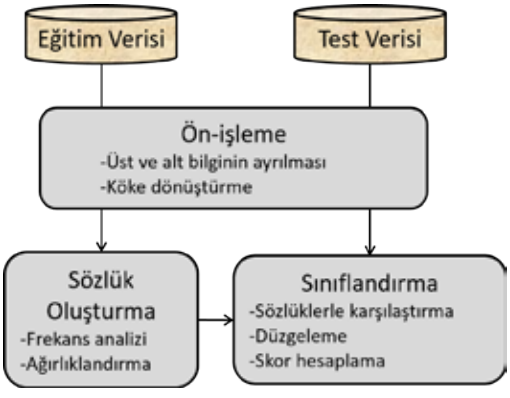
II. YÖNTEM

A. *Veri toplanması*

Bu çalışmada manyetik rezonans, bilgisayarlı tomografi, x-ray ve ultrason görüntüleri kullanılarak sekiz farklı anatomik bölgedeki bulguları özetleyen toplam 230 radyoloji raporundan faydalanılmıştır. Maltepe Üniversitesi Tıp Fakültesi'nden elde edilen bu veri kümesindeki tüm raporlar 1) hasta bilgisini tablo olarak içeren bir üst bilgi, 2) bulguları özetleyen serbest ana metin ve 3) katkı sağlayan radyologların isimlerini içeren alt bilgiden oluşacak şekilde aynı formata sahiptir. Raporların 69'u öğrenme, kalanlar ise geliştirilen sistemin başarımının ölçülmesi için kullanılmıştır.

B. *Radyoloji raporlarının sınıflandırılması*

Radyoloji raporlarının sınıflandırılması için önerdiğimiz sistem ön-işleme, sözlük oluşturma, sözlük temelli skorlama ve sınıflandırma adımlarından oluşmaktadır ve GNU/Linux'te çalışacak şekilde geliştirilmiştir (Şekil 1).



Şekil 1: Önerilen sistemin mimari yapısı

Öncelikle ücretsiz olan Antiword belge okuma uygulaması kullanılarak raporlar yalın metne dönüştürülmüştür. Sonra veri kümesindeki her rapora şu ön-işleme adımları uygulanmıştır: i) üst ve alt bilgilerin tespit edilip ana metinden ayrılması, ii) noktalama işaretlerinin kaldırılması, iii) metnin küçük harfe çevrilmesi, iv) köke dönüştürme (stemming), v) Türkçe karakterlerin İngilizce denkleri ile değiştirilmesi ve vi) boşluk tespiti yoluyla kelimelerin ayrıştırılması. Antiword üst bilgideki tabloyu kaldırıp yerine tablo içeriğini ve sonuna da "|" karakterini ekler. Bu karakter kullanılarak üst bilginin tespiti gerçekleştirilmiştir. Bunun yanında alt bilginin başlangıcı ise "saygılarımla" bitiş kelimesinin tespiti ile bulunmuştur. Bu adımlar önceki çalışmamızda ayrıntılı olarak açıklanmıştır [23].

Türkçe kelimenin anlamını değiştirebilen birçok soneke sahip sondan eklemeli dillerden olduğu için analizin önemli bir adımı köke dönüştürme (stemming) işlemidir. Ön-çalışmalarımız aslen İngilizce için tasarlanmış olan Porter'ın köke dönüştürme yönteminin [24] bizim problemimizde yetersiz kaldığını göstermiştir. Dolayısıyla Türk dilleri için geliştirilmiş açık-kaynak kodlu işletim sisteminden bağımsız bir doğal dil işleme kütüphanesi olan Zemberek [https://github.com/cbilgili/zemberek-nlp-server] kullanılmıştır. Karşılaştırma amacıyla ilk N harfi tutup kalanı kırparak çalışan en basit köke dönüştürme yönteminden de faydalanılmıştır. Ön-işlemi takiben eğitim veri kümesi kullanılarak her anatomik bölge için bir sözlük oluşturulmuştur. Bu amaçla her anatomik bölgeye ait raporlara kelime frekans analizi uygulanmıştır. Frekans analiz sonuçları kullanılarak ayırıcı kelimeler manuel olarak belirlenmiş ve her anatomik bölgeye özel bir sözlük oluşturulmuştur (bkz. Tablo I). Bu noktada sözlük oluşturma adımını manuel olarak gerçekleştirdiğimizi çünkü çalışmalarımız sonucunda literatürdeki etkisiz kelime (stopword) ve/veya kelime frekans analizi temelli otomatik çözümlerin problemimiz için yetersiz kaldığını belirtmek isteriz.

Tablo I: HER ANATOMIK BÖLGE IÇIN FREKANS ANALIZI KULLANILARAK ÖĞRENME VERISINDEN MANUEL OLARAK BELIRLENEN SÖZLÜK TERIMLERI

Anatomik Bölge	Sözlük Terimleri (Ayırıcı Kelimeler)
Göğüs	Meme, fibrokistik
Ayak	Ayak, metatars, sesamoid, metatarsofalengeal
Bilek	Karpal, fleksor, skafoid, radyokarpal, karpometakarpal, dirsek
Kafa	Ventrikul, korpus, kranial, sisternalar, serebral, kallosum, serebellar, mastoid
Omuz	Subakromial, glenoid, glenohumeral, omuz, akromion, subdeltoid, supraspinatus
Pelvis	Sakroiliak, femur, koksofemoral, sakral, pelvis, pubis, femoris, suprapubik, koksiks
Omurga	Intervertebral, noral, dural, lomber, vertebra, spinal, disk, herniasyon
Toraks	Akciger, toraks, pankreas, trakea, kardiak, abdomen

Terim frekansı, bir ya da bir grup belgede terimlerin aşırı tekrarından dolayı sistem performansının düşebileceği sebebiyle sınıflandırma için önemli ama güvenilir olmayan bir özniteliktir. Tercih edilmeyen bu yanlılığın üstesinden gelebilmek amacıyla her terimin anatomik bölgeye (sınıfa) göre önemi bir sınıftaki ilgili terimi içeren metinlerin sayısının o sınıftaki toplam metin sayısına (metin frekansı ya da DF) oranı üzerinden ölçülen bir ağırlık ile hesaba katılmıştır. TF-IDF, terim frekansı – ters metin frekansı, bir terimin veri kümesindeki metne göre önemini yansıtan popüler bir sayısal istatistiktir. Bu çalışmada TF-IDF temelli ağırlıklandırma yaklaşımı da denenmiştir. Test veri kümesindeki metinler eğitim veri kümesindekiler ile aynı ön-işlemlerden geçirilmiş ve önerdiğimiz sözlük temelli frekans analiz ve skorlama yöntemi kullanılarak ilgili anatomik bölge sınıfına otomatik olarak atanmıştır. Bu amaçla sözlük terimlerinin test metnindeki frekansları hesaplanmış ve ilgili sözlük ter-

imlerinin frekanslarının toplamı terim sayısı ile düzgelenerek her anatomik bölge sınıfına bir skor verilmiştir. Son olarak test metni en yüksek skora sahip anatomik bölge sınıfına atanmıştır.

III. DENEYSEL SONUÇLAR

Önerilen sistemin sınıflandırma başarımını sunan Tablo II'de gözlemleneceği üzere eğitim ve test veri kümelerinde benzer yüksek başarımlar görülmektedir. Sözlüklerdeki terimlerin önemini inceleyebilmek amacıyla terimler frekanslarına göre sıralanmış farklı frekans eşikleri kullanılarak değişen sayıda terim (bir-terim: en sık gözlenen; iki-terim; üç-terim; terimlerin tümü) içeren sözlükler oluşturulmuş ve bir dizi deney gerçekleştirilmiştir. Artan sayıda ayırıcı terim kullanıldığında sistem başarımında istatistiksel olarak anlamlı ($p<0.05$) bir iyileşme görülmektedir (Şekil 2-üst); sadece iki-terimli sözlük kullanılarak %90'ın üzerinde başarıma ulaşılmaktadır.

Tablo II: SISTEMIN EĞITIM VE TEST VERISINDEKI YÜZDELIK SINIFLANDIRMA BAŞARIMI

Anatomik Bölge	Eğitim Verisi (Doğruluk / F-ölçütü)	Test Verisi (Doğruluk / F-ölçütü)
Göğüs	100 / 100	100 / 100
Ayak	100 / 100	100 / 100
Bilek	100 / 100	94.1 / 97.0
Kafa	100 / 100	95.4 / 97.6
Omuz	83.3 / 90.9	100 / 100
Pelvis	93.8 / 92.3	95.7 / 97.8
Omurga	100 / 100	93.8 / 96.8
Toraks	100 / 100	100 / 100
Toplam	98.3 / 98.1	97.5 / 98.7

Şekil 2-sol alt ağırlıklandırma yaklaşımının sistem başarımına etkisini göstermektedir. Önerdiğimiz TF-DF temelli yaklaşım TF-IDF alternatifine göre daha yüksek başarım sağlamaktadır. Bunun temel sebebi "meme", "omuz", "ayak" gibi ayırıcı terimlerin metin frekanslarının çoğunlukla 1 olarak ölçülmesi ve dolayısıyla bu terimlerin TF-IDF temelli yaklaşımda hesaba katılmamasıdır.

Köke dönüştürme işleminin başarıma etkisini incelemek amacıyla Zemberek'in yanı sıra 3-6 karakterlik kırpma temelli en basit köke dönüştürme yöntemi de denenmiştir. Şekil 2-sağ altta görüldüğü üzere hem en basit köke dönüştürme yöntemi hem de Zemberek 8-sınıflık ve 230-raporluk problemimizde benzer başarım seviyelerine ulaşmaktadırlar. Bununla birlikte, daha büyük (ve/veya daha heterojen) bir radyoloji raporu veri kümesi daha çok sayıda anatomik bölgeye sınıflandırılmak istenirse köke dönüştürme için Zemberek gibi probleme uyarlanmış bir yöntem tercih edilebilir.

Önerilen sistem sözlük terimlerinin manuel olarak belirlenmesine ihtiyaç duymaktadır. Böylesi bir manuel müdahaleye ihtiyacın varlığını değerlendirmek amacıyla sözlük oluşturma adımı TF-DF temelli yaklaşım kullanılarak otomatik olarak gerçekleştirilmiş ve sonuçta elde edilen otomatik sistemin başarımı %43 olarak ölçülmüştür. Bu düşük başarımın sebebi büyük ihtimalle otomatik sistemin terim tercihini yalnızca frekansa dayalı yapmasıdır (dolayısıyla görüntüleme yöntemleri gibi düşük ayırıcılığa sahip terimler de seçilir); hâlbuki manuel müdahalede ilgili anatomik bölgeyi en iyi temsil eden terimler tercih edilir.

Sonuçların dikkatli incelenmesi sayesinde önerilen sistemin daha önce (PACS sistemine kayıt esnasında) yanlış etiketlenmiş yedi raporu doğru anatomik bölge sınıfına atadığı gözlenmiştir. Dolayısıyla manuel etiketleme sonuçlarının doğruluğunu teyit etmek ve düzeltmek konusunda da önerilen sistemin faydalı olacağı düşünülmektedir.

IV. VARGILAR

Bu çalışmada Türkçe radyoloji raporlarının anatomik bölgelerine otomatik sınıflandırılması için sözlük temelli (kural tabanlı) yeni bir sistem önerilmiştir. Rutin klinik uygulamadan alınmış 230 radyoloji raporu içeren bir veri kümesinde sistemin başarımı ortalama %98.1 duyarlılık (recall), %99.2 kesinlik (precision) ve %98.6 F-ölçütü (F-score) olarak ölçülmüştür. Bunun yanı sıra sistem yedi raporun klinik uygulama esnasında hatalı etiketlendiğini ortaya çıkarmıştır.

Önerilen sistemin başarımı öğrenme veri kümesi genişletilerek, kelime düzeltme gibi doğal dil işleme çözümlerinden yararlanılarak ve ontolojilere kodlanmış anatomik/patolojik bilgilerden faydalanılarak iyileştirilebilir. Önerilen sistem ileride (1) konuşma tanıma çözümleri ile birleştirilerek radyologların diktelerinin otomatik olarak yazıya/rapora çevrilmesinde, (2) radyoloji raporları, ontolojiler ve görüntü verisi ile birleştirilerek hastalıkların daha güvenilir teşhisinde kullanılabilir.

TEŞEKKÜR

Yazarlar veri seti için Maltepe Üniversitesi Tıp Fakültesi Radyoloji Bölümü'ne ve RadyolojiOnline'a teşekkürlerini sunar.

Bu çalışma kısmi olarak FP7-PEOPLE-2009-RG-249253 (COSeRMI, FP7 Marie Curie Actions European Re-integration Grants) fonu aracılığı ile Avrupa Komisyonu tarafından desteklenmiştir.

KAYNAKÇA

[1] J.H. Thrall, "Reinventing radiology in the digital age", Radiology, 236, 382–385, 2005.

[2] B. Reiner, "Uncovering and improving upon the inherent deficiencies of radiology reporting through data mining", J. Digital Imag., 23, 109–118, 2010.

[3] M. Bhargavan, A.H. Kaye, H.P. Forman, J.H. Sunshine, "Workload of radiologists in United States in 2006–2007 and trends since 1991–1992", Radiology, 252, 458–467, 2009.

[4] The Royal College of Radiologists, "Clinical radiology workload: guidance on radiologists' reporting figures", Technical Report, 2012.

[5] F.M. Hall, "The radiology report of the future", Radiology, 251, 313–316, 2009.

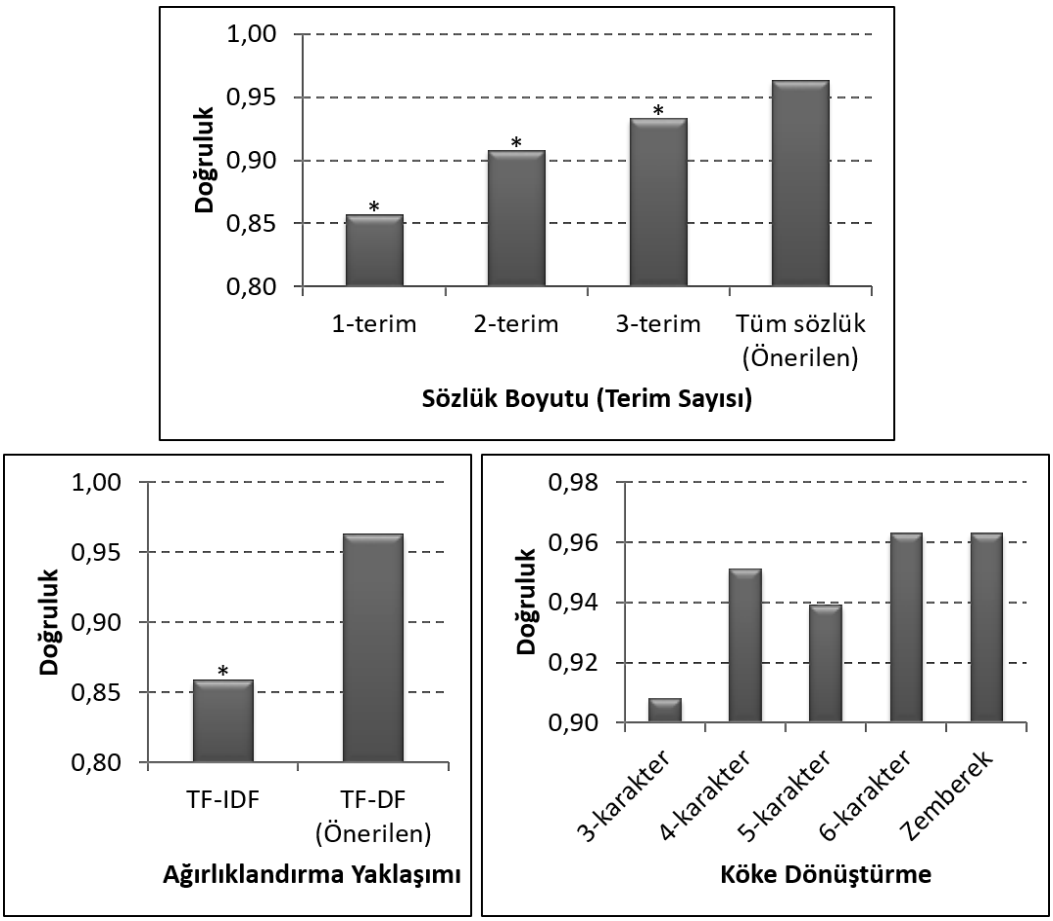
[6] A.M. Cohen, W.R. Hersh, "A survey of current work in biomedical text mining", Brief. Bioinf., 6, 57–71, 2005.

[7] E.R. Gabrieli, D.J. Speth, "Automated analysis of medical text I. clue gathering", J. Med. Syst., 14, 71–91, 1990.

[8] R. Baud, A. Rassinoux, J. Scherrer, "Natural language processing and semantical representation of medical texts", Methods Inf. Med., 31, 117–125, 1992.

[9] N. Sager, M. Lyman, C. Bucknall, N. Nhan, L.J. Tick, "Natural language processing and the representation of clinical data", J. Am. Med. Inf. Assoc., 1, 142–160, 1994.

[10] P. Spyns, "Natural language processing in medicine: an overview", Methods Inf. Med., 35, 285–301, 1996.



Şekil 2: Sözlük boyutunun (üst), ağırlıklandırma yaklaşımının (sol-alt) ve köke dönüştürme yaklaşımının sistem başarımına etkisi. * istatistiksel olarak anlamlı ($p<0.05$) sonucu belirtir.

[11] S.L. Zimmerman, W. Kim, W.W. Boonn, "Informatics in radiology: Automated structured reporting of imaging findings using the aim standard and xml", Radiographics, 31, 881–887, 2011.

[12] G. Hripcsak, J.H.M. Austin, P.O. Alderson, C. Friedman, "Use of natural language processing to translate clinical information from a database of 889,921 chest radiographic reports", Radiology, 224, 157–163, 2002.

[13] B.W. Mamlin, D.T. Heinze, C.J. McDonald, "Automated extraction and normalization of findings from cancer-related free-text radiology reports", In: AMIA 2003 Annual Symposium, Washington, USA, 420–424, 8-12 November 2003.

[14] I. Goldstein, A. Arzumtsyan, O. Uzuner, "Three approaches to automatic assignment of icd-9-cm codes to radiology reports", In: AMIA 2007 Annual Symposium, Chicago, USA, 279–283, 10-14 November 2007.

[15] C. Friedman, P.O. Alderson, J.H.M. Austin, J.J. Cimino, S.B. Johnson, "A general natural-language text processor for clinical radiology", J. Am. Med. Inf. Assoc., 1, 161–174, 1994.

[16] D.B. Aronow, F. Fangfang, W.B. Croft, "Ad hoc classification of radiology reports", J. Am. Med. Inf. Assoc., 6, 393–411, 1999.

[17] B.J. Thomas, H. Ouellette, E.F. Halpern, D.I. Rosenthal, "Automated computer-assisted categorization of radiology reports", Am. J. Roentgenology, 184, 687–690, 2005.

[18] A. Maghsoodi, M. Sevenster, J. Scholtes, G. Nalbantov, "Sentence-based classification of free-text breast cancer radiology reports", In: IEEE 2012 International Symposium on Computer-Based Medical Systems, Rome, Italy, 1–4, 20-22 June 2012.

[19] P. Lakhani, W. Kim, C. Langlotz, "Automated detection of critical results in radiology reports", J. Digital Imag., 25, 30–36, 2012.

[20] E. Arısoy, H. Dutağacı, L.M. Arslan, "A unified language model for large vocabulary continuous speech recognition of Turkish", Signal Process., 86, 2844–2862, 2006.

[21] E. Soysal, İ. Çiçekli, N. Baykal, "Design and evaluation of an ontology based information extraction system for radiological reports", Comput. Biol. Med., 40, 900–911, 2010.

[22] N.M. Ceylan, A. Alpkoçak, A.E. Esatoğlu, "An intelligent system to help on assignment of icd-10 codes to medical records'", In: 2012 Turkish Medical Informatics Conference, Belek, Antalya, 93–104, 15-18 November 2012.

[23] T. Kocatekin, D. Ünay, "Text mining in radiology reports", 2013 21st Signal Processing and Communications Applications Conference (SIU), Haspolat, 2013, pp. 1-4.

[24] M.F. Porter, "An Algorithm for Suffix Stripping", In: Readings In Information Retrieval. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 313–316, 1997.