Hacettepe Üniversitesi Sosyal Bilimler Enstitüsü

Bilgi ve Belge Yönetimi Anabilim Dalı

HABER METİNLERİNİN KATEGORİZASYONUNDA VARLIK İSİMLERİ VE KONU BAŞLIKLARI İLİŞKİSİ

İpek Şencan

Yüksek Lisans Tezi

Ankara, 2014

HABER METİNLERİNİN KATEGORİZASYONUNDA VARLIK İSİMLERİ VE KONU BAŞLIKLARI İLİŞKİSİ

İpek Şencan

Hacettepe Üniversitesi Sosyal Bilimler Enstitüsü

Bilgi ve Belge Yönetimi Anabilim Dalı

Yüksek Lisans Tezi

Ankara, 2014

KABUL VE ONAY

İpek Şencan tarafından hazırlanan "Haber Metinlerinin Kategorizasyonunda Varlık İsimleri ve Konu Başlıkları İlişkisi" başlıklı bu çalışma, 6 Haziran 2014 tarihinde yapılan savunma sınavı sonucunda başarılı bulunarak jürimiz tarafından yüksek lisans tezi olarak kabul edilmiştir.

Prof. Dr. Yaşar Tonta (Başkan)

Prof. Dr. İlyas Çiçekli

Doç. Dr. Hasan Oğul

Doç. Dr. İrem Soydal (Danışman)

Yrd. Doç. Dr. Gülten Alır

Yukarıdaki imzaların adı geçen öğretim üyelerine ait olduğunu onaylarım.

Prof. Dr. Yusuf Çelik

Enstitü Müdürü

BİLDİRİM

Hazırladığım tezin/raporun tamamen kendi çalışmam olduğunu ve her alıntıya kaynak gösterdiğimi taahhüt eder, tezimin/raporumun kağıt ve elektronik kopyalarının Hacettepe Üniversitesi Sosyal Bilimler Enstitüsü arşivlerinde aşağıda belirttiğim koşullarda saklanmasına izin verdiğimi onaylarım:

- ■ Tezimin/Raporumun tamamı her yerden erişime açılabilir.
- ☐ Tezim/Raporum sadece Hacettepe Üniversitesi yerleşkelerinden erişime açılabilir.
- ☐ Tezimin/Raporumun …… yıl süreyle erişime açılmasını istemiyorum. Bu sürenin sonunda uzatma için başvuruda bulunmadığım takdirde, tezimin/raporumun tamamı her yerden erişime açılabilir.

06.06.2014

_____

İpek Şencan

Değerli anneme, babama ve ağabeyime…

TEŞEKKÜR

Bu çalışmada yardım ve desteğini gördüğüm pek çok kişiye teşekkür borçluyum. En başta, tez yazma sürecinde bilgi, fikir ve yorumlarını benimle paylaşan değerli danışmanım ve hocam Doç. Dr. İrem Soydal'a sonsuz teşekkür ederim.

Başım her sıkıştığında kapısını çaldığım, dibe vurduğum anlarda beni cesaretlendiren ve her zaman desteğini hissettiğim değerli hocam Doç. Dr. Umut Al'a verdiği emek için ne kadar teşekkür etsem azdır.

Yoğun çalışma temposu içerisinde zaman ayırarak çalışmamı okuyup geri bildirim veren ve jürimde yer alan değerli büyüğüm ve hocam Prof. Dr. Yaşar Tonta'ya büyük bir teşekkür borçluyum. Tez sürecimin en başında bana zaman ayırarak konumu netleştirmemde bana yol gösteren ve jürimde yer alarak görüşlerini paylaşan sevgili hocam Yrd. Doç. Dr. Gülten Alır'a çok teşekkürler. Çalışmamı okuyarak yorumlarını benimle paylaşan değerli jüri üyeleri Prof. Dr. İlyas Çiçekli ve Doç. Dr. Hasan Oğul'a da teşekkürlerimi sunmak isterim.

Araştırmanın temelini oluşturan verileri kullanma konusunda gerekli izinleri sağlayan başta Prof. Dr. Fazlı Can olmak üzere Bilkent Bilgi Erişim Grubu'na teşekkürü bir borç bilirim.

Araştırmanın çeşitli aşamalarında gazete ve habercilik konusunda bilgi ve deneyimlerine başvurduğum Hüseyin Ünal, Sinan Tartanoğlu, Necati Serhat Hürkan, Ersin Bal, Mahmut Güner, Kemal Göktaş ve Gökçer Tahincioğlu'na sonsuz teşekkürler.

Teknik konularda yardımlarına ve fikirlerine başvurduğum değerli hocalarım Dr. Güven Köse, Uzman Orçun Madran ve Uzman Enver Güneş'e çok teşekkürler. Ayrıca teknik desteğinden dolayı Hamid Ahmadlouei'ya da teşekkürler.

Araştırmam süresince ilgi ve desteklerinden dolayı değerli bölüm hocalarıma teşekkürlerimi sunarım.

Sevgili arkadaşlarım Zehra Taşkın ve Sümeyye Akça'ya güler yüzleriyle bu süreçte hep yanımda oldukları için çok teşekkür ederim. Sevgili oda arkadaşım Güleda Doğan'a da verdiği fikir ve önerilerin yanı sıra her an desteğini hissettirdiği ve anlayışını esirgemediği için sonsuz teşekkür borçluyum.

Sevgili dostum, zor zamanlarımın destekçisi sevgili Müge Akbulut, sana olan minnetimi kelimelerle ifade edebilmem mümkün değil. Ne kadar şanslıyım güzel kalbinle her daim yanımda olduğun için. Seninle birlikte gidecek çok yolumuz var…

Bugünlere gelmemde emeği olan ve haklarını asla ödeyemeyeceğim sevgili aileme, her zaman varlıklarını hissettirdikleri ve bu süreçte güçlü kalmamı sağladıkları için ne kadar teşekkür etsem azdır.

Bu araştırma Türkiye Bilimsel ve Teknolojik Araştırma Kurumu (TÜBİTAK) tarafından desteklenmiştir (Proje numarası 111K030). TÜBİTAK'a ve Hacettepe Üniversitesi Bilimsel Araştırma Projeleri Koordinasyon Birimi'ne katkılarından dolayı teşekkürlerimi sunarım.

ÖZET

ŞENCAN, İpek. *Haber Metinlerinin Kategorizasyonunda Varlık İsimleri ve Konu Başlıkları İlişkisi*, Yüksek Lisans Tezi, Ankara, 2014.

Metin kategorizasyonu ile büyük ve kirli veri yığınlarının içerisindeki bilgiler düzenlenerek bilgiye erişim kolay ve pratik hale gelmektedir. Metin kategorizasyonu ayrıca, bilgiye ihtiyaç duyan kişilerin istedikleri bilgiye erişmelerinde zaman kazandırmak açısından da son derece önemlidir. Haber metinleri gibi hızlı artış potansiyeline sahip olan yapılar metin kategorizasyonuna ihtiyaç duyulan önemli uygulama alanlarından biridir.

Bu çalışmada, BilCol-2005 Türkçe haber derleminden sağlanan 5834 haber kullanılarak, haber metinlerinin kategorizasyonunda varlık isimleri (named entities) ve konu başlıkları ilişkisinin incelenmesi amaçlanmıştır. Buna yönelik olarak 5834 haber yedi farklı varlık ismi (kişi, kurum, konum, tarih, zaman, para ve yüzde) ile etiketlenmiştir. Etiketlenen haberler IPTC (International Press Telecommunications Council) temel düzey konu başlıkları taksonomisine göre kategorize edilmiş ve derlem 13 farklı IPTC konu başlığı ile tanımlanmıştır. Bu doğrultuda gerçekleştirilen analizler ile etiketli ve etiketsiz kelimelere ilişkin sıklık ve oran değerleri bazı istatistiksel testlerden de (Mann-Whitney U testi) yararlanılarak belirlenmiştir.

Çalışma sonucunda elde edilen bulgulardan, derlemdeki haberlerin etiketlenmesinde en baskın varlık isminin “Kişi”, en pasif varlık isminin “Zaman” olduğu, derlemdeki haberlerin etiketlenme sayılarının IPTC konu başlıklarına göre farklılık gösterdiği, tüm konu başlıkları için “Kişi”, “Kurum” ve “Konum” varlık isimlerinin ön planda olduğu ve konu başlıklarının kavramsal içeriğini varlık isimlerinden çok etiketlenmemiş kelimelerin yansıttığı anlaşılmış ve ilgili hipotezlerimizin tamamı desteklenmiştir.

Bu çalışma, Türkçe bir haber derlemi üzerinde uluslararası alanda standart olarak kabul edilen IPTC’nin temel düzey konu başlıkları ile varlık isimlerinin bir arada uygulanarak aralarındaki bağlantıların sorgulandığı Türkçe literatürdeki ilk çalışma olması açısından önemlidir. Bu çalışmada elde edilen sonuçların gazetecilere, haber metinlerinin kategorizasyonu üzerine çalışanlara, haber metinlerine hızlı ve doğru erişim ihtiyacı duyan kullanıcılara yardımcı olacağı düşünülmektedir.

Anahtar Sözcükler

Metin kategorizasyonu, Haber metinlerinin kategorizasyonu, Varlık isimleri, BilCol-2005, IPTC, IPTC konu başlıkları taksonomisi

ABSTRACT

ŞENCAN, İpek. *Relationship between the Named Entities and the Subject Titles in Categorization of News Items*, Master's thesis, Ankara, 2014.

With text categorization it is possible to access information within a large pile of impure data. It also helps people to save time who wants to have information more easily and practically. One of the most important practical research areas in terms of text categorization is news, as it has a potential of rapid increase.

This thesis aims to investigate the connection of subject codes with named entities, in terms of text categorization, by using 5834 news texts which were obtained from BilCol-2005 news corpus. To address this, 5834 news were tagged with seven different named entities (person, organization, location, date, time, money and percentage). Tagged news were classified under 13 different subject codes of IPTC's (International Press Telecommunications Council) main subject taxonomies. The investigation was based on tagged and untagged words and their relations with the IPTC news codes. Key findings were revealed with the frequency and percentage values along with some statistical tests (e.g. Mann Whitney U test, Chi-square test).

Findings showed that the most and the least dominant named entitity in the corpus was "Person" and "Time", respectively. Besides, it was also revealed that the number of tagged words differed according to subject codes where "Person", "Organization" and "Location" named entities were prominent among all subjects. Moreover, it was seen that the conceptual content of the subject codes were reflected more by untagged words than the ones tagged with named entities. These findings supported our hypothesis.

This study is important because it was the first in the related Turkish literature in which the connection between an international standard news taxonomy (IPTC) and named entities was investigated. The findings are believed to be useful for journalists, for the news taxonomists and for those who need to access the news texts fast and accurately.

Keywords

Text categorization, Categorization of news items, Named entities, BilCol-2005, IPTC, IPTC news subject codes taxonomy

İÇİNDEKİLER

TABLOLAR DİZİNİ

ŞEKİLLER DİZİNİ

1. BÖLÜM

GİRİŞ

1.1. KONUNUN ÖNEMİ

Bilginin boyutu her geçen saniye katlanarak artmakta, bu durum ihtiyaç duyulan bilgiye, erişim problemini de beraberinde getirmektedir. Özellikle günlük yaşamın bir parçası haline gelmiş olan İnternet ile birlikte herhangi bir fiziksel formattan bağımsız hale gelen bilgi, günlük yaşamın bir parçası olarak geçmişe oranla daha hızlı ve kolay biçimde hem üretilmekte hem de kullanılmaktadır. Öte yandan kullanıcılar, kendileri için gerekli olduğunu düşündükleri bilgiyi ararken çoğu zaman bu devasa bilgi yığınının içerisinde kaybolmaktadırlar.

İnternet'in yaşamın her alanını etkilemeye başladığı günümüzde, bu etkiden nasibini alan alanlardan biri de haber platformlarıdır. Önceleri çoğunlukla gazete, radyo ve televizyon gibi geleneksel haber kaynakları ile sınırlı olan haber aktarımı şimdilerde yerini giderek Web ortamındaki haber platformlarına bırakmaktadır.

Haber organları genellikle, abone oldukları haber ajanslarından çektikleri haberleri kimi zaman değişiklik yapmadan, kimi zaman da başlık ve spot[1] bilgilerini değiştirerek kendi web sayfalarından yayımlamaktadırlar. Dahası, haber organları İnternet ortamında sunulan haberlere anında müdahale edebilmekte, haber metinlerine ekleme yapabilmekte ve başka sayfalarla bağlantı oluşturabilmektedirler. Ayrıca haberlere erişim kolaylaşmış, pek çok haber kaynağı haber içeriklerine ücretsiz erişim imkânı tanımıştır (Gürcan ve Batu, 2001). İnternet'in hızla yayılması haber değiş tokuşunun karmaşıklığını ortadan kaldırmış, böylece hem haber kaynaklarında hem de ortalama bir okurun başvurduğu haber unsurlarının (fotoğraf, video, vb.) sayısında çok ciddi bir artış söz konusu olmuştur. Buna karşılık bu aşırı artış, okurların çok fazla bilgiye maruz kalmasına neden olmuş, haberin erişilebilirliğini ve kullanılabilirliğini düşürmeye başlamıştır. Bu kadar büyük boyutlu bir haber akışı içerisinden istenilen unsurlara erişim için tüm içeriğin kullanılmasının zor olduğu da düşünüldüğünde içerik

[1] *Haber spotu*, "Haberin 15-20 sözcüklük özet ifadeleri ya da haberin okunmasını sağlayacak ilgi çekici anlatımlar" olarak ifade edilmektedir (Gürcan, Yüksel, Vural, Çetintaş ve Banar, 2012, s. 72).

tanımlaması ve üst veri (metadata) kullanımı daha önemli hale gelmiştir (Bacan, Pandzic ve Gulija, 2005).

Öte yandan, haber alış verişinde önemli noktalardan bir tanesi de haberlerin uluslararası olma boyutudur. Özellikle bu niteliğe sahip haberlerin farklı ülkelerle etkileşiminin sağlanması oldukça önemlidir. Bu durum, uluslararası geçerliliğe sahip bir haber tanımlama ve kategorizasyon standardının kullanılması gerekliliğini ortaya koymaktadır.

Haber erişimi ile ilgili bahsi geçen zorlukları ortadan kaldırmaya yardımcı olacak bazı yaklaşımlar ön plana çıkmaya başlamıştır. Bu yaklaşımlardan bir tanesi de metin kategorizasyonudur. Önceden oluşturulmuş bir kategorizasyon şemasına uygun olacak şekilde metnin etiketlenmesi ya da farklı kategoriler altında gruplanması şeklinde ifade edilen metin kategorizasyonu (Güven, Onur ve Sağıroğlu, 2008, s. 161) ile bunun alt alanlarından biri olan ve büyük miktardaki verilerden veriyi madenleme veya bilgi çıkarma (Han ve Kamber, 2006, s. 5) olarak tanımlanan veri madenciliği bilgi yığınının düzenlenmesi, erişim için bilgilerin daha uygun bir forma sokulması açısından bir çözüm olarak değerlendirilmektedir. Bu yaklaşımlar aynı zamanda erişim performansını önemli oranda artırmaktadır. Metin kategorizasyonu için oluşturulan etiketler ya da kategoriler çok farklı alanlarda bilgi süzme/filtreleme, bilgi çıkarımı gibi değişik amaçlarla kullanılabilmektedir. Metin kategorizasyonu konusunda gerçekleştirilen çalışmalarda çoğunlukla kategorizasyonun etkisini ve performansını artırmaya yönelik çalışmalar dikkat çekmektedir. Bu çalışmalar literatürde daha ziyade, elle ya da otomatik olarak gerçekleştirilen kategorizasyon, kategorizasyondaki başarımı artıran varlık isimlerinin (named entities) (kişi, kurum, konum, vb.) kategorizasyonda kullanılması gibi bazı yöntemler biçiminde yer bulmaktadır (Güran, Akyokuş, Bayazıt ve Gürbüz, 2009; Amasyalı ve Diri, 2006; Hayes, Knecht ve Cellio, 1988; Yang, Ault, Pierce ve Lattimer, 2000; McNamee, Mayfield ve Piatko, 2011). Öte yandan, metin kategorizasyonunun yaygın olarak kullanıldığı alanların başında haberler gelmektedir (Hayes, Knecht ve Cellio, 1988; Jo, 1999; Gui, Gao, Li ve Yang, 2012). Her türlü bilginin haber olma potansiyeline sahip olmasının da etkisiyle, her saniye artan haberler doğru olarak kategorize edildiklerinde bu haberlerin erişilmeleri ve bilgi ihtiyacını karşılamaları daha kolay gerçekleşmektedir.

Metin kategorizasyonunda kullanılan etkili yöntemlerden biri varlık isimleridir. Bir içeriğin varlık isimleri ile tanımlanması işlemi, önceden belirlenmiş kelime dizilerinin

metinde ilgili kelimelere atanmasından oluşmaktadır (McNamee, Mayfield ve Piatko, 2011, s. 33). Varlık ismi tanımlama ve bir metni varlık isimlerine göre sınıflama aslında temel olarak Kim, Ne, Nerede, Ne zaman, Neden ve Nasıl sorularına yanıt veren kişi, kurum ve yer isimleri ile tarih, saat, para, yüzde gibi sayısal ifadelerin metin içerisinde saptanması anlamına gelmektedir. Özellikle bilgi çıkarımı[2] konusunda önemli bir alt alan olan varlık ismi tanımlama ve sınıflama, pek çok araştırmacı tarafından soru cevaplama, düşünce çıkarımı,[3] olaylar/kavramlar arasındaki ilişkilerin belirlenmesi gibi değişik amaçlarla kullanılmaktadır (Nadeau ve Sekine, 2007, s. 3; Marrero, Urbano, Sanchez-Cuadrado, Morato ve Gomez-Berbis, 2012, s. 7).

Varlık isimlerinin kullanımı özellikle haber metinleri üzerine yapılan Konu Tespit ve Takip (KTT) (Topic Detection and Tracking - TDT) çalışmalarında da sıkça başvurulan yaklaşımlardan biridir. KTT, 1990'lı yılların sonuna doğru oldukça yeni ve önemli bir araştırma alanı olarak ortaya çıkmıştır. KTT çalışmalarının amacı, çeşitli haber yayım ortamlarındaki çok dilli, haber odaklı metinsel belgelerin aranması, düzenlenmesi ve yapılandırılmasına yönelik teknolojilerin geliştirilmesidir. Söz konusu çalışmalar, pek çok farklı haber kaynağından (haber yayınları, çevrimiçi gazeteler gibi) toplanan işe yarar her bir bilgiyi kısıtlı bir zaman diliminde tek tek izlemek, dinlemek veya okumak yerine teknolojiyi kullanarak bu bilgileri otomatik tekniklerle değerlendirmede büyük kolaylık sağlamaktadır (Fiscus ve Doddington, 2002, s. 17; Wayne, 1998).

KTT çalışmalarının temel adımlarından birini oluşturan ve habercilikte "yeni olay" olarak bilinen kavram, daha önce herhangi bir şekilde rapor edilmemiş, yeni olma özelliği taşıyan haber hikâyeleri için kullanılmaktadır (Kuo, Zi ve Gang, 2007, s. 215). Habercilik alanında haber ögeleri olarak bilinen ve bir haberde temel olarak yanıtlaması beklenen sorular olan Ne, Nerede, Ne zaman, Neden, Nasıl ve Kim (5N1K) sorularının haberlerde yeni olay tespit etmede başarım sağladığı bilinmektedir (Gürcan ve diğerleri, 2012; MEGEP, 2007; Makkonen, Ahonen-Myka ve Salmenkivi, 2003). Öte yandan, kişi, kurum, yer adı, tarih, zaman, yüzde, para gibi varlık isimlerinin kullanımı, KTT çalışmalarının bir diğer temel adımını oluşturan ve iki haber hikâyesinin aynı konuda olup olmadıklarını tespit etme olarak tanımlanan "hikâye bağlantı algılama"

[2] *Bilgi çıkarımı (information extraction)*, doğal dile dayalı bir metinde belirli bir ilişki veya olay grubuna ait örneklerin tanımlanması ve bu ilişki veya olaylarla ilgili parametrelerin çıkarımı olarak ifade edilmektedir (Grishman, 1997, s. 10).

[3] *Düşünce çıkarımı (opinion extraction)*, düşüncelerin makale, gazete, blog vb. kaynaklardan kelime, cümle ya da belge düzeyinde araştırılarak bulunmasıdır (Ku, Liang ve Chen, 2006).

(story link detection) görevi üzerinde de oldukça etkilidir (Zhang, Wang ve Chen, 2008, s. 436).

Haber metinlerindeki ilgili kelimelerin varlık isimleri ile tanımlanması yaklaşımının yanı sıra haber metinlerinin kategorize edilmesinde, düzenlenmesinde, kullanıcıya aktarılmasında, saklanmasında ve arşivlenmesinde dünyanın önde gelen haber ajansları (Thomson Reuters, The New York Times, BBC Monitoring, ANSA vb.) tarafından IPTC (International Press Telecommunications Council) adlı bir standart (haber kodları ve haber başlıları taksonomisi de içeren) kullanılmaktadır (IPTC members, 2014). IPTC, haber ajansları, gazeteler ve haber sistemi üreticilerinden oluşan, haberlerin ucuz, kolay, tam ve doğru paylaşımını teşvik etmek amacıyla bazı teknik şartnameler geliştiren ve yayımlayan bir konsorsiyumdur. Dünyadaki pek çok gazete ve haber web sitesi en az bir veya iki IPTC standardı kullanmaktadır. Bu standart, metin ya da fotoğraf, video gibi görsel-işitsel veriler içeren çoklu ortam (multimedia) kaynaklara ilişkin üst veri kodları içermekte ve farklı dillere adapte edilebilmektedir (IPTC membership Q&A, 2012).

Haber derlemleri üzerinde farklı yaklaşımlarla gerçekleştirilen metin kategorizasyonu çalışmaları haber yazım tekniklerinin iyileştirilmesi, birbirleriyle bağlantılı konu ve hikâyelerin belirlenmesi, bunları otomatik olarak yüksek doğruluk oranında başarabilmesi gibi konulara odaklanmaktadır. Bu sayede haberlerin hem arşivlenmesi hem de haberlere erişimde daha sağlıklı yöntemler geliştirilebileceği düşünülmektedir.

1.2. ARAŞTIRMANIN AMACI VE KAPSAMI

Çalışmanın amacı, haber metinlerinin kategorizasyonunda varlık isimleri ile konu başlıkları ilişkisinin incelenmesi ve varlık isimlerinin haberlerin konu başlıklarının belirlenmesinde etkili olup olmadığının saptanmasıdır. Bu ilişkinin belirlenmesi ile haberlerin dâhil olduğu temel konu başlıkları içerisindeki daha baskın ve daha pasif olan varlık isimlerinin tespit edilmesi hedeflenmektedir. Çalışmanın veri setini, "Türkçe Haber Benzerliklerinin Belirlenmesinde Varlık İsimlerinin Hikâye Bağlantı Algılama Görevinin Başarımına Etkisi" başlıklı TÜBİTAK Projesi (Proje No: 111K030) kapsamında etiketlenmiş olan BilCol-2005[4] haber derlemindeki 5834 haber

[4] BilCol-2005, Bilkent Üniversitesi Bilgisayar Mühendisliği Bölümü tarafından bilgi erişim çalışmalarında kullanılmak üzere oluşturulmuş haber derlemidir. Derlem ile ilgili ayrıntılar Bölüm 2'de verilmektedir.

oluşturmaktadır. Etiketleme işleminin ardından, çalışmanın amacı gereği IPTC haber konu başlıkları taksonomisine dayanan toplam 17 ana konu başlığı içerisinden derlemdeki haberlere en uygun olan 13 konu başlığı ilgili haberlere atanmıştır. Çalışmadaki analizler yukarıda anılan proje kapsamında yapılan etiketleme işlemi ve 13 farklı IPTC konu başlığına göre yapılan sınıflamaya dayalı olarak gerçekleştirilmiştir.

Bu çalışma, uluslararası alanda standart olarak kabul edilen IPTC'nin temel düzey konu başlıkları ile varlık isimlerinin bir arada belirli bir derlemdeki haberlere uygulanarak başlıklar ve varlık isimleri arasındaki bağlantıları sorgulaması açısından Türkçe literatürde gerçekleştirilen ilk çalışmadır. Ayrıca çalışmada IPTC konu başlıklarına yer verilmesi, bu standart yapıya ilişkin farkındalık yaratılması açısından önem taşımaktadır. Bu çalışma neticesinde ortaya çıkan bulguların haber metinlerinin kategorizasyonu ile uğraşanlara, habercilik sektöründe çalışanlara ve haber metinlerine hızlı ve doğru erişim ihtiyacı duyan kullanıcılara ışık tutacağı düşünülmektedir.

1.3. ARAŞTIRMA SORULARI VE HİPOTEZLER

Haber niteliği taşıyan belirli bir olayın gelişimi sırasında yeni üretilen haberlerin birbirleri ile ve dâhil oldukları genel konu başlığı ile bağlantılarının kopmaması gereği, haberin yayılma hızı ve bu haberlere hızlı ve doğru şekilde erişim ihtiyacı düşünüldüğünde giderek daha da önem kazanmaktadır. Birçok haberin uluslararası önem taşıdığı da göz önüne alınırsa ajansların birbirleri ile ve dünyadaki diğer ajanslarla "konuşabilen" sistemler geliştirmesi gerektiği açıktır. Çalışma kapsamında 24.04.2013 - 27.02.2014 tarihleri arasında Türkiye'de bulunan bazı büyük haber ajansı, haber arşivlerinin yetkilileri ve gazetecilik vakfında eğitim veren gazeteciler (Anadolu Ajansı, Cumhuriyet, Milliyet, Akşam, Vatan, Posta, Um:ag) ile çeşitli görüşmeler yapılmış, haberlerin arşivlenmesi ve haberlere erişimde uluslararası standartların tam ve sistematik şekilde kullanılmadığı, kurumların genelde kendi çözümlerini kendilerinin ürettikleri anlaşılmıştır.

Bu çalışma doğrultusunda yanıtlanması hedeflenen temel araştırma soruları şunlardır:

- Derlemdeki haberlerin etiketlenmesinde en baskın ve en pasif varlık isimleri[5] hangileridir?

[5] Haberlerin etiketlenmesinde en baskın ve en pasif varlık isimleri ile ifade edilmek istenen, belirli varlık isimlerinin haberlerin etiketlenmesinde fazla ya da az kullanılmış olmasıdır.

- Derlemdeki haberlerin etiketlenme sayıları IPTC konu başlıklarına göre farklılık göstermekte midir?
- Her bir konu başlığı için en belirleyici varlık ismi/isimleri hangileridir?
- Varlık isimleri ile etiketlenen kelimeler konu başlıklarını kavramsal olarak yansıtmada ne kadar etkilidir?

Araştırma soruları doğrultusunda ortaya çıkan hipotezler:

- Derlemdeki haberlerin etiketlenmesinde en baskın varlık ismi “Kişi”, en pasif varlık ismi ise “Zaman”dır.
- Derlemdeki haberlerin etiketlenme sayıları IPTC konu başlıklarına göre farklılık göstermektedir.
- Tüm konu başlıkları için “Kişi”, “Kurum” ve “Konum” varlık isimleri ön plandadır.
- Konu başlıklarının kavramsal içeriğini varlık isimlerinden çok etiketlenmemiş kelimeler yansıtmaktadır.

Bu sorulara yanıt aranırken, konu başlıklarındaki etiket dağılımının derlemin genelinden ne kadar farklılık gösterdiği, her bir konu başlığında yer alan haber başlıkları ile içerdikleri haber sayılarının benzerlik gösterip göstermediği, konu başlıklarının içerdiği tekil etiket oranları, her bir konu başlığı altında sık kullanan etiketli ve etiketsiz kelimeler ile ilgili de veri setine yönelik çeşitli tanımlayıcı bulgular da elde edilmiş ve çalışmanın ilgili bölümlerinde bunlara da yer verilmiştir.

1.4. ARAŞTIRMA TASARIMI

Çalışmada, BilCol-2005’ten “Türkçe Haber Benzerliklerinin Belirlenmesinde Varlık İsimlerinin Hikâye Bağlantı Algılama Görevinin Başarımına Etkisi” başlıklı TÜBİTAK Projesi (Proje No: 111K030) kapsamında yer alan 5834 haber veri seti olarak kullanılmıştır. Proje sürecinde “Kişi”, “Konum”, “Kurum”, “Tarih”, “Para”, “Yüzde” ve “Zaman” olmak üzere yedi farklı varlık ismi ile etiketlenmiş olan ifadeler ile haberlerde geçen ancak herhangi bir kategoriye girmediği düşünüldüğü için etiketlenmeyen kelimeler tez çalışmamızda analiz kapsamına alınmıştır. Etiketli ve etiketsiz kelimeler üzerinde yaklaşık altı ay boyunca devam eden oldukça yoğun bir veri temizleme işlemi gerçekleştirilmiştir.

İkinci aşamada, haberler IPTC haber konu başlıkları taksonomisine dayalı 13 konu başlığına göre kategorize edilmiştir. Sonrasında çalışmanın amacına uygun olan

analizler gerçekleştirilmiştir. Konu başlıklarına göre etiket dağılımlarının görselleştirilmesinde ve sunumunda radar grafiklere yer verilmiştir.

Çalışmamızın bu konu üzerine çalışacak diğer araştırmacılara yardımcı olacağını düşünmemizden dolayı araştırma tasarımı ile ilgili bilgiler detaylı olarak Bölüm 3'te sunulmuştur.

1.5. ARAŞTIRMANIN DÜZENİ

Çalışmamız beş bölümden oluşmaktadır:

İlk bölümde çalışmamıza temel oluşturan konunun öneminden, araştırmanın amacından ve kapsamından bahsedilmekte, buna dayanılarak oluşturulmuş araştırma soruları ile hipotezler belirtilmekte ve araştırma tasarımı hakkında kısaca bilgi verilmektedir.

İkinci bölümde literatür değerlendirmesi yer almaktadır. Bu bölümde araştırmanın kavramsal arka planını oluşturan konular hakkında bilgi verilmekte, metin kategorizasyonu, varlık isimleri ile IPTC üzerine yapılan araştırmalardan bahsedilmektedir.

Üçüncü bölümde araştırmanın gerçekleştirileceği verilerin toplanması, haberlerin etiketlenme süreci ve IPTC konu başlıklarına atanması ile verilerin temizlenerek analize uygun hale getirilmesi sürecinde başvurulan yöntemler ve çalışma sürecinde karşılaşılan sınırlılıklar hakkında detaylı bilgi verilmektedir.

Dördüncü bölümde konu başlıklarına ve varlık isimlerine göre kategorize edilen 5834 haber ve yedi farklı varlık ismi ile ilgili elde edilen bulgu ve yorumlara yer verilmektedir. Bulgulara yönelik tartışma kısmı da yine bu bölümde sunulmaktadır.

Son bölümde çalışma kapsamında elde edilen sonuçlar yer almakta ve bu sonuçlar doğrultusunda birtakım öneriler getirilmektedir.

1.6. KAYNAKLAR

Araştırma dâhilinde daha önceden yapılmış çalışmaları belirleyip erişebilmek amacıyla bazı çevrimiçi veri tabanları ile İnternet arama motorları üzerinden tarama yapılmıştır. Bu amaçla kullanılan kaynaklardan bazıları şunlardır:

ACM Digital Library

Dissertations and Theses – Proquest (1997-)

IEEE Xplore (1970-)

EBSCOhost (1969-)

Google Books (books.google.com)

Google Scholar (scholar.google.com)

SAGE (1995-)

ScienceDirect (1997-)

Springer LINK / Kluwer (1993-)

ULAKBİM TÜBİTAK Destekli Projeler Veri Tabanı (1966-)

ULAKBİM Mühendislik ve Temel Bilimler Veri Tabanı (1992-)

ULAKBİM Sosyal Bilimler Veri Tabanı (2002-)

Wiley Online Library (1997-)

YÖK Tez Merkezi (https://tez.yok.gov.tr/UlusalTezMerkezi/)

Tezin yazımında Hacettepe Üniversitesi Sosyal Bilimler Enstitüsü "Tez ve Rapor Yazım Yönergesi" kullanılmıştır.[6]

[6] Hacettepe Üniversitesi Sosyal Bilimler Enstitüsü Tez ve Rapor Yazım Yönergesi'ne http://www.sosyalbilimler.hacettepe.edu.tr/belgeler/Tez_ve_Rapor_Yazim_Yonergesi.pdf adresinden erişilmiştir.

2. BÖLÜM

LİTERATÜR DEĞERLENDİRMESİ

2.1. GİRİŞ

Bu bölümde araştırma kurgumuzun dayandırıldığı kavramsal arka planı oluşturan unsurlar anlatılmaktadır. Öncelikle metin kategorizasyonu konusu ve ilgili çalışmalar hakkında bilgiler verilmiş, daha sonra varlık isimlerinin kullanıldığı çalışmalar ve konu tespit ve takip sistemleri için yapılan araştırmalardan bahsedilmiştir. Bunların yanı sıra IPTC konu başlıklarından yararlanılan çalışmalardan örneklere ve bu araştırmaya temel oluşturan BilCol-2005 derlemi ile bu derlem kullanılarak literatürde daha önce yapılmış çalışmalara da bu bölümde yer verilmektedir.

2.2. METİN KATEGORİZASYONU

Metin kategorizasyonu, belgelerin içeriklerine göre önceden belirlenmiş olan kategorilere atanması olarak tanımlanmaktadır (Güran ve diğerleri, 2009, s. 369; Amasyalı ve Diri, 2006, s. 221; Lewis, 1991, s. 312). Bu işlem çoğunlukla otomatik olarak gerçekleştirilmektedir. Metin kategorizasyonu, belirli bir ilgi grubuna göre haberlerin filtrelenmesi, istenmeyen e-posta (spam) filtreleme ve web sayfalarının kataloglanması işlemleri sırasında sıkça karşımıza çıkmaktadır (Mathew, 2006).

1988 yılında haber metinlerinin kategorizasyonuna ilişkin gerçekleştirilen bir çalışmada, geniş konu kategorileri içerisindeki haber hikâyelerinin kategorizasyon problemine yönelik doğal dil işleme tekniklerinin kullanıldığı ticari bir pilot uygulama üzerinde durulmuştur (Hayes, Knecht ve Cellio, 1988). Çalışma sonucunda metin kategorizasyonunun farklı başlıklar içeren, bilgi tabanlı kurallar çerçevesinde uygulanan doğal dil işleme tekniklerini kullanan metinler için uygun olduğu belirlenmiştir. Bu tip bir otomatik metin işlemenin, haber hikâyelerinin, elektronik mesajların veya diğer çevrimiçi metinlerin formlarının arşivlenmesi ve iletilmesinde pek çok potansiyel uygulamaya sahip olduğu belirtilmiştir. Bir diğer çalışmada (Yang ve diğerleri, 2000), istatistiksel metin sınıflamadaki temel zorluklardan biri olan kronolojik olarak sıralanmış haberlerden olayların otomatik olarak takip edilmesi üzerinde durulmaktadır. Çalışmada, haberlerin izlenmesinde farklı iki yaklaşım (k-en yakın

komşu ve Rocchio) kullanılmıştır. Çalışma sonucunda, kullanılan yöntemlerin ağırlıklı hata oranını %71'e kadar düşürdüğü ve başarılı bir gelişme gösterdikleri belirtilmiştir. Ayrıca bu yöntemlerin birleştirilmesiyle, farklı veri derlemleri üzerindeki olay takip sisteminin performansındaki değişkenliğin büyük ölçüde azaltıldığı ifade edilmiştir. Bir başka çalışmada (Amasyalı ve Diri, 2006), Türkçe için ilk kapsamlı metin kategorizasyon modeli ortaya konmuştur. Çalışma, Türkçe belgelerin yazarlarının kimliklerinin saptanması, metin türüne göre belgelerin sınıflanması ve yazara ilişkin cinsiyetin otomatik olarak belirlenmesi olarak üç farklı kısımda gerçekleştirilmiştir. Çalışma sonucunda metin yazarının, metin türünün ve yazar cinsiyetinin belirlenmesinde sırasıyla %83, %93 ve %96'lık bir başarı elde edildiği belirtilmiştir.

Haberlerin doğru ve etkin şekilde kategorize edilmesi içerik analizi çalışmalarına da kolaylık sağlayacaktır. Haber içeriklerinin analiz edildiği çalışmalar değerlendirildiğinde, incelenen konulardaki eksikliklerin tespit edilmeye çalışıldığı ve bu eksikliklerin giderilmesine yönelik önerilerin ortaya konulduğu, ayrıca bazı konulardaki eğilimlerin belirlenmeye çalışıldığı görülmektedir. Örneğin, bu çalışmalardan birinde trafik haberlerinin nasıl sunulduğu incelenmiştir. Çalışmada, Hürriyet, Sabah, Zaman ve Cumhuriyet gazetelerinde 2000 yılı süresince yayımlanan toplam 859 trafik haberine içerik analizi uygulanmıştır. Çalışma sonucunda, trafik haberlerinin genelde ana konular arasında değil de doğal afetler, trajik olaylar gibi alt kategoriler içerisinde yer aldığı, haberlerin öznesini sıradan insanların oluşturduğu ve bu konudaki haberlerin bu çalışmaya dek pek araştırmaya değer bulunmadığına dikkat çekilmiştir (İrvan ve Çınarbaş, 2002). Berkant ve Cömert'in (2013) çalışmasında günlük gazetelerde yer alan eğitim haberleri belge analizi yöntemiyle incelenmiştir. 17.04.2012 - 06.05.2012 tarihleri arasında belirlenen dört gazetede yer alan eğitim haberleri değerlendirme kapsamına alınmıştır. Haberler, eğitim haberlerine yönelik konulara bölünmüş ve haberlere ilişkin içerik analizi haber içerik kodları ve bunlara ait sıklık ile yüzde değerleri üzerinden yorumlanmıştır. Çalışma sonucunda, incelenen gazetelerdeki eğitim haberlerinde ortaöğretim, üniversite, üniversite sınavı ve sınav sistemi gibi konuların yoğunlukta olduğu, muhabirler tarafından oluşturulmuş haberlerin çoğunlukta olduğu, bu konulardaki haberlerin pek çoğunda bilgi kaynağının Milli Eğitim Bakanlığı olduğu ve haftanın bazı günlerinde eğitim haberlerine yer verilmediği gibi tespitlerde bulunulmuştur. Bir diğer çalışmada ise sağlık haberlerine ilişkin içerik analizi gerçekleştirilmiştir. Sağlık ile ilgili haber başlıklarının içerik ile uyumu ve içeriğin tıbbi açıdan bilimsel veriler ile uyumunun değerlendirilmesine yönelik olarak Nisan-Aralık 2010 tarihleri arasında 11 gazetede yayınlanmış toplam 344 sağlık haberi

değerlendirme kapsamına alınmıştır. Haberler; haber içeriğinin başlık ile uyumu, içeriğin tıbbi açıdan doğruluğu, haber konusu ile ilgili ana bilim dalı, haber kaynağının belirtilmesi ve haber içeriğinde hekim/araştırmacı ismi belirtilmesi gibi kriterlere göre değerlendirilmiştir. Çalışma sonucunda, haber başlığı ile içeriğin tam uyumlu olduğu haber oranı %55 olarak bulunmuş, haberlerin yaklaşık %51'inin doğru bilgiler içerdiği tespit edilmiştir. Öte yandan, haber içerisinde hekim adı geçen haberlerin içeriğinin tıbbi açıdan %58 oranında doğru olduğu ve hekim adı geçen haberlerin geçmeyen haberlere göre doğruluk oranının istatistiksel olarak anlamlı derecede (p= 0,018) yüksek olduğu belirlenmiştir. Çalışma neticesinde, sağlık haberlerinde belirli kural ve standartlara bağlı kalınmasının önemi vurgulanmıştır (Hayran ve Özdemir, 2011). Bu noktada haberlere uygulanan içerik analizi yönteminin haber içeriklerine yönelik nicel ve nitel bilgiler elde etme açısından önemli olduğu anlaşılmaktadır.

Literatürde metin kategorizasyonunun bilgi erişime olan etkisinin artırılmasına yönelik çalışmaların ağırlıklı olduğu görülmektedir. Metin kategorizasyonu konusunda gerçekleştirilen çalışmaların bu konunun özellikle bilgi erişim probleminin çözümüne yardımcı olmada en önemli yollardan biri haline gelmesine katkı sağlaması açısından son derece değerlidir.

2.2.1. Varlık İsimleri

Yapılandırılmış, yarı yapılandırılmış ya da yapılandırılmamış metinlerde geçen ve belirli bir anlam içermekte olan, çoğunlukla isim türü ifadelerin varlık ismi (named entity) olarak etiketlenmesi, soru yanıtlama (QA-question answering), özetleme, bilgi erişim ve bilgi çıkarımı gibi Doğal Dil İşleme uygulaması içeren pek çok konuda uygulanmaktadır (Sekine ve Nobata, 2004, s. 1977).

"Varlık isimleri" yapısının çıkış noktası MUC-6 (Message Understanding Conference-6)'dır. Bu konferansın temel amacı, bilgi çıkarımı için geliştirilen teknolojiler içerisinde kullanımı pratik ve büyük ölçüde alandan (domain) bağımsız olarak kısa sürede yüksek doğruluk oranı çalışabilecek, otomatik fonksiyonların tanımlanmasını sağlamaktır. Bu amacı gerçekleştirebilmek için MUC Komitesi tarafından bir metinde geçen tüm coğrafi yer, kurum ve insan isimlerini tanımlamayı içeren varlık isimleri görevi geliştirilmiştir (Grishman ve Sundheim, 1996, s. 467). MUC, tıpkı gazete makaleleri gibi yapılandırılmamış metinler içerisinde yer alan savunma ve şirket faaliyetlerine ilişkin bilgi çıkarımı görevi üzerine odaklanmıştır. Bu görevin tanımlanmasında, kişi, kurum,

yer isimleri ile zaman, tarih, para ve yüzde gibi sayısal ifadelerden oluşan bilgi birimlerini ayırt etmenin önemli olduğuna dikkat çekilmiştir. Herhangi bir metinde bu varlıklara (kişi, kurum, konum, zaman, tarih, para, yüzde) yönelik ifadelerin tanımlanması "Varlık İsmi Tanıma ve Sınıflama – Named Entity Recognition and Classification (NERC)" olarak adlandırılmış ve bilgi çıkarım konusunun önemli alt alanlarından biri olarak kabul edilmiştir (Nadeau ve Sekine, 2007, s. 3).

Belgelerde yer alan varlık isimlerinin etiketlenmesi aşamasında çoğunlukla anlamsal veya bağlamsal bilgi kullanılmaktadır. Bu durumda karşılaşılan temel sorunların başında veri seyrekliği gelmektedir. İsimler, gazete metinleri veya web sayfaları gibi düzenli güncellenen belgelerde oldukça sık geçmektedir. Bunlar, yaygın olarak kullanılan isimlerden daha fazla çeşitlilik göstermekle birlikte sürekli değişmektedir. Genellikle bu tür bir etiketleme gerçekleştirebilmek adına derlem tabanlı yaklaşımlar tercih edilmektedir (Shinyama ve Sekine, 2004).

Varlık isimleri görevinin amacı, ham bir metinde yer alan ifadelerin çeşitlilik sınırlarını otomatik olarak belirlemek, sonrasında ise tanımlanmış olan bu ifadeleri kategorize etmektir (Palmer ve Day, 1997, s. 190). Varlık isimlerine yönelik "TIMEX", "NUMEX" VE "ENAMEX" şeklinde üç ana kategori tanımlandığı görülmektedir (Sundheim, 1995). "TIMEX" tarih ve zaman gibi geçici ifadeler için, "NUMEX" yüzde ve para gibi sayısal ifadeler için, "ENAMEX" ise kişi, kurum, konum gibi özel isimler için kullanılmıştır. İlk etapta varlık türüne (kişi, kurum, konum, zaman, tarih, para ve yüzde olmak üzere yedi tür) ait sayıların sınırlı olmasının başlıca nedeni, değerlendirmedeki hedef uygulamanın öncelikle iş (business) faaliyetlerine yönelik bilgi çıkarımı olmasından kaynaklanmıştır (Sekine, Sudo ve Nobata, 2002).

Varlık isimleri ile ilgili yapılan çalışmalara bakıldığında, özellikle varlık isimlerinin farklı dillerde kullanımının etkilerinin ölçüldüğü dikkati çekmektedir. Bu çalışmalardan birinde, birkaç dilde (Çince, İngilizce, Fransızca, Japonca, Portekizce ve İspanyolca) erişilebilir olan derlem üzerinde varlık isimleri görevinin istatistiksel profili ortaya konmuştur (Palmer ve Day, 1997). Bir başka çalışmada Japonca'da varlık isimlerinin de ele alındığı değerlendirmeye dayalı Bilgi Erişim ve Bilgi Çıkarımı Alıştırması (Information Retrieval and Information Extraction Exercise-IREX) başlıklı bir proje yürütüldüğünden bahsedilmektedir (Sekine ve Isahara, 2000). Proje kapsamında, MUC'ta belirlenmiş olan yedi tür varlık ismine "artifact" (örneğin, bir kitabın başlığı ya da bir ürünün ismi) yeni bir tür olarak eklenmiştir (Sekine ve Isahara, 2000). CoNLL (Computational

Natural Language Learning) 2002 ve 2003 Konferansları çerçevesinde ise İngilizce, Almanca, Flemenkçe ve İspanyolcada varlık isimleri görevlerinin işleyişi ele alınmıştır (Sang, 2002; Sang ve Meulder, 2003).

Gerçekleştirilen çalışmalar neticesinde isimleri daha detaylı varlık isimleri ile belirtme ihtiyacının ortaya çıkması, varlık ismi hiyerarşisi oluşturma ve genişletme çalışmalarını beraberinde getirmiştir. Sekine, Sudo ve Nobata (2002) çalışmalarında, günlük gazete makalelerinde görünen varlık isimlerinin çoğunu kapsayacak nitelikte bir varlık isimleri hiyerarşisi tasarlamışlardır. Yaklaşık 150 varlık ismi türü içeren bu hiyerarşinin geliştirilmesi üç aşamada gerçekleşmiştir. İlk aşamada, derleme (gazete makaleleri), tanımlara ve WordNet, Roget gibi kavram dizinlerine (thesaurus) dayalı olmak üzere farklı yöntemlerden yararlanılarak üç hiyerarşi oluşturulmuştur. Sonrasında bu üç hiyerarşi birleştirilmiş, son aşamada tanımlamalar kullanılarak derlem etiketlenmiş ve hiyerarşi yaratılmıştır. Bu çalışmanın sonuçlarının, diğer varlık ismi hiyerarşisi oluşturma ve genişletme çalışmalarının yanı sıra gazete alanlarında bilgi çıkarımı, soru cevaplama gibi doğal dil işleme uygulamalarında da yararlı olabilmesi hedeflenmiştir. Sekine ve Nobata'nın (2004) sonraki çalışmalarında, bazı uygulamalarda eksiklikler gözlendiği ifade edilerek geliştirmiş oldukları bu varlık ismi hiyerarşisindeki varlık ismi türü sayısı 200'e çıkarılmıştır. Çalışma kapsamında hiyerarşideki her kategori için gazetelerden, Web'den ve diğer kaynaklardan elle yaklaşık 130.000 civarında örnek toplanmıştır. Öte yandan özel kategorilerin de listelendiği yaklaşık 50.000 cins isim örneğinden oluşan bir sözlük geliştirilmiştir. Geliştirilen sözlük ve etiketlemede sözlükten yararlanılamayan durumlarda kullanmak üzere geliştirilen örnek-tabanlı kurallardan yola çıkılarak bir varlık ismi etiketleyici (named entity tagger) geliştirilmiştir. Genişletilen varlık isimleri hiyerarşisinin temel gazete haber alanlarını kapsayacak nitelikte olması ve bu sayede belirtilen alanlardaki uygulamalarda genişletilen hiyerarşinin rahatlıkla kullanılabilir olması çalışmanın temel amacı olmuştur.

Türkiye'de varlık isimleri ile ilgili olarak gerçekleştirilen çalışmalara bakıldığında dünya geneline oranla bu çalışmaların daha sınırlı sayıda olduğu dikkat çekmektedir. Köse 2004 yılında gerçekleştirdiği çalışmasında olay modeli yaklaşımını esas alarak haber benzerliklerinin saptanmasında "Kim" etiketinin ne ölçüde etkili olduğunu incelemiştir. Çalışma sonucunda sadece "Kim" etiketinin tek başına kullanımının yeterli olmadığı, bunun "Ne", "Nerede", "Ne zaman" etiketleriyle bütünleştirilerek incelenmesi gerekliliğini ortaya koymuştur. Bir başka çalışmada, Türkçe için kural tabanlı bir sistem tasarlanmıştır. Bu sistem, belgede yer alan yer ve kurum isimleri ile özel isimleri

etiketleyerek çıkarabilecek şekilde geliştirilmiştir. Çalışma sonucunda, en düşük başarı oranı (%80) kişi isimlerinin tanınmasında elde edilmiş ve kişi isimlerinin tespiti için farklı kurallar geliştirilmesi önerilmiştir (Dalkılıç, Gelişli ve Diri, 2010). Tatar ve Çiçekli'nin 2011 yılında gerçekleştirdikleri çalışmada doğal dildeki metinlerde yer alan varlık isimlerinin tanımlanmasında girdi metninin farklı özelliklerinden yararlanan bir otomatik kural öğrenme yöntemi tanımlanmaktadır. Aynı zamanda Türkçe metinlerde varlık isimlerinin çıkarımı için morfolojik özelliklerin kullanımı da araştırılmıştır. Geliştirilen tekniğin otomatik varlık ismi tanıma görevine başarıyla uygulanabilir olduğu çalışma sonucunda ortaya konmuştur. Öte yandan, morfolojik özelliklerden yararlanmanın sondan eklemeli bir dil olan Türkçede varlık ismi tanıma işlemini önemli ölçüde geliştirebileceği ifade edilmiştir.

Varlık isimleri kullanılarak gerçekleştirilen araştırmalarda kullanılan yöntemlerden biri de Konu Tespit ve Takip'tir (KTT). Bu çalışmaların temel hedefi sürekli gelen haber metinleri ve bu metinlerde tartışılan olayların düzenlenmesinde bilgi erişim yöntemlerini araştırmaktadır.

Konu Tespit ve Takip (KTT) programı ilk olarak 1996-1997 tarih aralığında pilot bir proje ile başlatılmış ve program sonraki çalışmalar için bir laboratuvar ortamı görevi görmüştür. Programın amacı, geniş bir haber ortamından çok dilli, haber odaklı metinsel materyalleri arayan, düzenleyen ve yapılandıran teknolojiler geliştirmek olarak tanımlanmıştır (Fiscus ve Doddington, 2002, s. 17). KTT kapsamında beş temel görev tanımlanmaktadır. Bunlar; olayları birbirinden bağımsız başlıklara ayırma işini ifade eden "hikâye bölümleme (story segmentation)", sisteme dâhil edilen haberin daha önceki hikâyelerden bağımsız olduğunun belirlenmesini sağlayan "ilk hikâye algılama (first story detection)", yeni gelen haberlerin hangi hikâye kümesine ait olduğunun tespit edilmesine yarayan "küme algılama", belirlenen hikâyelere yeni haberler eklenip eklenmediğinin takip edilmesini ifade eden "izleme (tracking)" ve sisteme dâhil olan haberlerin birbirinden farklı konuların parçası olup olmadığına bakan "hikâye bağlantı algılama (story link detection)" olarak adlandırılmaktadır (Allan, 2002, s. 3-4; Topic Detection and Tracking, 2008). KTT uygulamaları sürekli gelişmekte olan metin ve diğer çoklu ortam türü hikâyelerdeki konuları takip etme ve büyük miktardaki derlemleri düzenlemede kolaylıklar sağlamaktadır (Topic Detection and Tracking, 2008). Hikâye sınırlarının bulunması, hangi hikâyenin diğeriyle birlikte ilerlediğinin belirlenmesi, ne zaman yeni bir olay olduğunun saptanması gibi konularda veriyi otomatik olarak

ayrıntılarıyla planlayabilen, gösterebilen bir algoritma olması nedeniyle bilgi erişim konusunda kullanıcılara oldukça kolaylık sağlayabilmektedir (Wayne, 1998).

Ülkemizde KTT çalışmaları konusunda başı çeken Bilkent Bilgi Erişim Grubu, oluşturdukları BilCol-2005 derlemi ile birçok çalışmaya imza atmıştır (Kardaş, 2009; Can ve diğerleri, 2009; Can ve diğerleri, 2010). Bunun yanı sıra yine BilCol-2005 derlemi kullanılarak ülkemizde farklı yöntemlerle KTT çalışmaları gerçekleştirilmiştir (Aksoy, Can ve Koçberber, 2012; Soydal ve Al, 2014). Bu çalışmalar BilCol-2005 Haber Derlemi alt başlığında daha ayrıntılı olarak sunulmuştur.

Literatürde Konu Tespit ve Takip alanında varlık isimleri yaklaşımından yararlanılarak gerçekleştirilen çalışmalardan birinde KTT görevlerinin yerine getirilmesini destekleyici kullanıcı arayüzü boyutu üzerine odaklanmıştır. Çalışmada belge temsilinde ve kullanıcı arayüzünde varlık isimleri yaklaşımı kullanılmıştır. Çalışma sonunda KTT görevlerinin gerçekleştirilebileceği bir arayüz kurulumu önerisi sunulmuştur (Bashaddadh ve Mohd, 2011). KTT çalışmaları çerçevesinde, haber benzerliklerinin belirlenmesi ve haber metinlerinin yansıtılmasında varlık isimlerinden yararlanıldığı görülmektedir (Shah, Croft ve Jensen, 2006).

Varlık isimlerine yönelik gerçekleştirilen çalışmaların geneline bakıldığında, varlık ismi etiketleme işleminin üzerinde çalışılan alana ve kullanıcı ihtiyaçlarına paralel olarak şekillendiği ve buna bağlı olarak yeni çalışmalar yapılmasının önemi anlaşılmaktadır.

2.2.2. International Press Telecommunications Council (IPTC– Uluslararası Basın Telekomünikasyon Konseyi) Konu Başlıkları

1965 yılında Londra'da kurulan IPTC, dünyanın gözde haber ajanslarını, yayıncılarını ve bu sektördeki haber sağlayıcıları bir araya getirmiş haber dünyasına uluslararası standartlar sağlayan bir kuruluştur. IPTC'nin temel amacı, bilginin dağıtımını kolaylaştırmaktır. Bu doğrultuda, içerik sağlayıcılar, aracılar ve kullanıcılar arasındaki bilgi alış verişini ve bilgi yönetimini geliştirmek için etkili teknik standartlar geliştirmiş ve bunları desteklemiştir. IPTC, açık standartlar sağlamakta ve bunları üyelerine ücretsiz olarak erişilebilir kılmaktadır (IPTC about, 2014).

IPTC standartları, haber ajansları, haber sağlayıcılar ve haber yayıncıları için kurumlararası haber değişimini sağlamayı hedeflemekte ve farklı formattaki haberleri desteklemektedir. Örneğin, çoklu ortam türü haberlerin değişimi ve düzenlenmesi için

NewsML 1, Spor haberlerinin paylaşımı için SportsML, gazete makalelerinin içerik ve yapılarının tanımlanması için NITF gibi farklı IPTC standartlarına başvurulmaktadır (IPTC membership Q&A, 2012).

IPTC standartlarının tercih edilme nedeni büyük ölçüde üyeler için ücretsiz oluşu ve çoğunlukla kullanılan XML'e dayalı üst veri yapısının pek çok farklı formata dönüştürülebilmeye imkân tanımasıdır. Diğer taraftan, dünya çapında kabul gören standart yapısı sayesinde metinsel, görsel ya da görsel-işitsel haberleri etiketlemeyi ve paketlemeyi olanaklı kılmaktadır. Haber içeriklerinin etiketlenmesinde, hem insanlarca hem de makinalarca anlaşılabilir yapıda haberleri tanımlamaya yarayan zengin üst veri terimlerinden oluşan IPTC haber konu kodları (IPTC News Codes) kullanılmaktadır (IPTC membership Q&A, 2012).

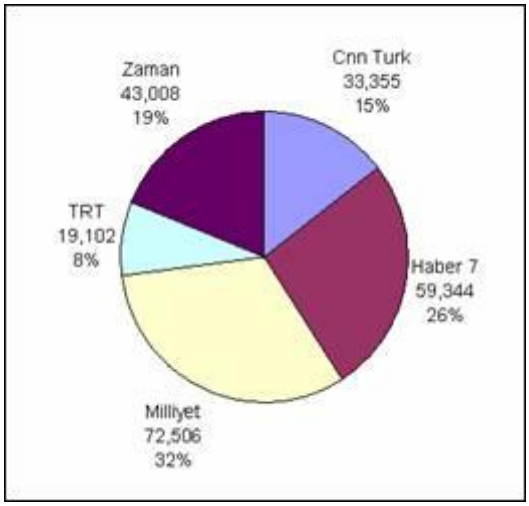
Literatürde IPTC konu başlıkları taksonomisinden yararlanılarak yapılan pek çok çalışma bulunmaktadır. Bacan, Pandzic ve Gulija tarafından 2005 yılında gerçekleştirilen çalışmada, IPTC standart konu başlıkları kategorilerine dayanarak Hırvat Haber Ajansı'ndan sağlanmış yaklaşık 2700 haber elle kategorize edilmiş ve bu haberler üzerinde vektör uzayı modeli kullanılarak otomatik haber kategorizasyonu test edilmiştir. Test sonucunda %85 oranında başarım elde edilmiş ve otomatik kategorizasyon kullanımı önerilmiştir. Troncy'nin 2008 yılındaki çalışmasında ise, farklı üst veri standartlarının basit bir bilgi ortamında karşılıklı işlerliğine olanak sağlayacak IPTC haber mimarisi için bir OWL (Ontology Web Language) tasarlanmış ve diğer çoklu ortam üst veri standartları ile bağlantılı hale getirilmiştir. IPTC haber kodları Simple Knowledge Organization System (SKOS) kavram dizinine dönüştürülmüş ve semantik web içerisinde haber üst verilerinin doğal dil işleme ve çoklu ortam analizi ile nasıl daha zengin hale getirilebileceği ortaya konmaya çalışılmıştır. Çalışma sonucunda semantik bir arama sistemi ve haberleri tarama için çeşitli keşfedici arayüzler sunulmuştur. Bir diğer çalışmada, MD Info Şirketi büyük ölçüde iş yaşamına ilişkin bilgilerle tüketicilere ilişkin haberlerden oluşan bilgileri üç düzeyli (ana başlıklar, alt başlıklar ve konular) taksonomi yapısına göre düzenlemiştir. Şirketin kullandığı taksonominin sağlanan haber metinlerini otomatik olarak kategorize edememesinden dolayı çalışma kapsamında IPTC konu taksonomisi yardımıyla MD Info taksonomisinin otomatik kategorizasyonu gerçekleştirip gerçekleştiremeyeceği test edilmiştir. Başlık bazında ve kelime bazında gerçekleştirilen testler sonucunda MD Info taksonomisi ile IPTC taksonomisi arasında doğrudan bir bağlantı kurulamayacağı ve MD Info

taksonomisi ile otomatik kategorizasyon gerçekleştirilemeyeceği tespit edilmiştir (Temmink, 2010).

Literatürdeki çalışmalarda da görüldüğü gibi IPTC standartları, haber kodları, haber konu taksonomisi gibi farklı çalışmalarda, farklı amaçlarla kullanılabilmektedir. Farklı formatlardaki haberlerin karşılıklı değişimini ve etkileşimini sağlayan bu standartların kullanımı haber kullanıcıları, haberciler, yayıncılar ve haber sektöründeki diğer kurum ve kişilerin çalışmalarını büyük ölçüde kolaylaştırmaktadır.

2.3. BilCol-2005 HABER DERLEMİ

Türkçeye yönelik yeni olay belirleme ve takip sistemlerinin etkinliğinin ölçülmesinde kullanılabilecek BilCol-2005 deney derlemi, Web'deki beş ayrı haber kaynağından indirilen tarih, saat ve dakika bilgilerini de içeren 2005 yılına ait 200 bini aşkın haber metni içermektedir (Can ve diğerleri, 2007). Derlemdeki haberler zaman ve haber sayısı açısından karşılaştırıldığında söz konusu haberlerin bu konuda önceden gerçekleştirilmiş çalışmalardan daha kapsamlı olduğu vurgulanmıştır (Can ve diğerleri, 2007). İndirilen haberlerin %32'si Milliyet gazetesinden, %19'u Zaman gazetesinden, %15'i CNN Türk'ten, %26'sı Haber 7'den, %8'i ise TRT'den sağlanmıştır (bkz. Şekil 1) (Bilkent yeni olay belirleme, 2013).



Şekil 1. BilCol-2005 Derlemindeki haberlerin kaynaklara göre dağılımı (Bilkent yeni olay belirleme, 2013)

Haber kaynaklarından haberler indirilirken hemen hemen her kaynak için farklı bir yöntem kullanılmıştır. Bu durumun haber kaynaklarının geçmiş tarihli haberleri arşivlerken farklı yaklaşımlar izlemesinden kaynaklandığı görülmüştür. Haber 7 ve TRT'den haber indirilirken, her iki haber kaynağı da haberlere haber numarası (ID) verdiğinden, 2005 yılına ait ilk ve son haberlerin numaraları temel alınarak iki numara arasında kalan tüm haberler indirilmiştir. Zaman gazetesi web sayfasında ise haberlere günlük olarak erişilmiş olup, bir tarih aralığı belirlenmiş ve haberler bu tarih aralığına göre indirilmiştir. CNN Türk web sayfasında her bir habere ait ayrı numara vardır. Haberleri indirmek için 2005 yılına ait haberlerin linklerini elde etmeye yönelik bir yazılım geliştirilmiş ve haberler dört adımda indirilmiştir. Milliyet gazetesindeki haberler, gazetenin haberlere atadığı tarih ve iki basamaklı haber numarasından yola çıkılarak başlangıç ve bitiş tarihlerinin esas alınması yoluyla indirilmiştir (Can ve diğerleri, 2007). İndirilen haberlere ilişkin HTML dosyalarındaki etiketler ayıklanmıştır. Yeni olayların ve onları izleyen haberlerin saptanması için ETracker (Event Tracker) adlı bir uygulama hazırlanmış ve derlemdeki haberler dizinlenmiştir. Bu tür derlem oluşturma yaklaşımı verimli bulunduğu için benzer bir yöntemin yenilik belirleme ve izlemeye yönelik başka deney derlemlerinin oluşturulmasında da kullanılabileceği önerilmiştir (Can ve diğerleri, 2007).

Oluşturulan BilCol-2005 derlemine ek olarak Bilkent Üniversitesi Bilgisayar Mühendisliği Bölümü Bilgi Erişim Grubu tarafından Bilkent Haber Portalı tasarlanmış ve etkinleştirilmiştir (Öcalan, 2009). Çalışmada, portalın tasarımı ve oluşturulma süreci, mimari yapısı, veri yapılarının yanı sıra bu haber portalının geliştirilmesinde kullanılacak bazı unsurların bilgi erişime etkileri de ölçülmüş, buna yönelik olarak, kök bulma yöntemlerinin, belge ve sorgu uzunluklarının etkileri araştırılmıştır. Araştırma sonuçlarına göre, orta uzunluktaki sorguların ve daha uzun belgelerin Türkçe bilgi erişimde daha etkin sonuçlar sağladığı vurgulanmıştır.

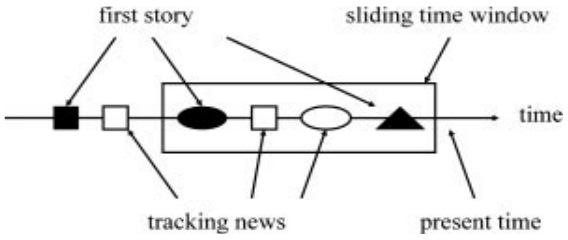
Haber portalı geliştirme ve buna benzer çalışmaların etkinliğini ve sürekliliğini sağlayabilmek amacıyla Türkçe haberlerden oluşan deney derlemlerinden yararlanılmış ve uygulamalar bunlara dayanılarak geliştirilmiştir (Can ve diğerleri, 2009). Haber portalındaki hedeflerden bir tanesi belirli bir konudaki haberler içinde yenilik içeren haberlerin belirlenmesi ve bir zaman çizgisi doğrultusunda kullanıcıya aktarılmasıdır. Buna yönelik geliştirilen algoritmalar Türkçe haber deney derlemi kullanılarak test edilmiş ve çalışmayı gerçekleştiren proje ekibi tarafından geliştirilen kapsama katsayısı kavramı kullanılmıştır (Can ve diğerleri, 2009). Çalışma sonucunda,

geliştirilen haber portalının proje ekibine yenilikçi yaklaşımların denenmesine olanak veren bir laboratuvar ortamı sağladığı vurgulanmıştır.

Uyar (2009), çalışmasında Tweezer adlı algoritma ile tekrarlanan (birbirinin aynısı) haberleri ayıklamanın yollarından biri olan varlık isimleri kullanımını ele almıştır. Bu algoritmada varlık isimlerine karşılık gelen kelimeler ve bu kelimelerin öncesinde ve sonrasında gelen kelimelerden yararlanılarak belge imzası oluşturulmuştur. Aynı imzayı taşıyan haber belgelerinin yaklaşık olarak aynı olduğu kabul edilmiştir. Varlık isimlerinin belirlenmesinde TuNER (Turkish Named Entity Recognizer - Türkçe Adlandırılmış Nesne Tanıyıcı) yöntemi kullanılmıştır. Tweezer'ın değerlendirilmesinde ise Bilkent Haber Portalı'ndan elde edilen haberlerden oluşan bir belge seti oluşturulmuştur. Çalışmada Tweezer, IDF (inverse document frequency) değeri kullanılarak belgelerin imzalarını çıkaran I-Match ile karşılaştırılmış ve Tweezer'ın I-Match'ten daha iyi olduğu maliyet fonksiyonu (yanlış ikaz ve kaçırma oranı olasılıklarını birleştiren) ve F ölçütü (anma ve duyarlığı birleştiren) kullanılarak istatiksel olarak ortaya konmuştur (Uyar, 2009).

Kardaş'ın (2009) çalışmasında anında yeni olay belirleme ve izleme konusuna bağlı problemler BilCol-2005 derlemi kullanılarak araştırılmıştır. Çalışmada, Türkçe yeni olay belirleme ve izleme işlemlerinde bazı benzerlik ölçümleri için kelimelerin ilk 5-6 harfinin o kelimenin kökü olarak kullanılmasının dilin morfolojik yapısına dayanılarak oluşturulan kök bulma yaklaşımı ile yarışabileceği belirtilmiştir. Bunu yanı sıra kelimelere yönelik durma kelimeleri listesi (stopword list) kullanımının sistem başarımını artırdığı ve iki farklı benzerlik hesaplama yönteminden elde edilen güven skorlarının birleştirilmesiyle sistemin etkinliğinin artırılabileceği vurgulanmıştır (Kardaş, 2009).

2010 yılında gerçekleştirilen Türkçe'deki yeni olay belirleme ve konu izleme problemlerinin araştırıldığı çalışmada ise BilCol-2005 deney derleminden yararlanılarak yeni olay belirleme için "kayan zaman penceresi" (sliding time window) kavramı kullanılmıştır (bkz. Şekil 2) (Can ve diğerleri, 2010).



Şekil 2. Sliding time window (Can ve diğerleri, 2010, s. 803)

Çalışmada, yeni olay belirleme süresince en yeni hikâye zaman penceresi hikâyeleri ile karşılaştırılmış, eğer en yeni hikâye zaman penceresi hikâyelerinden yeterince farklı ise bu durum yeni olay belirleme olarak ifade edilmiş ve yeni bir olayın ilk hikâyesi kabul edilmiştir. Konu tespiti sırasında ise belirlenen bir konuda yeni gelen hikâye konu açıklama vektörü ile karşılaştırılmış, eğer yeterince benzerse o hikâyenin belirtilen konuda izleyen haber olduğu kabul edilmiştir. Konu tespiti esnasında her konu ayrı ayrı ele alınmıştır. Çalışmada deneysel değerlendirme içinse BilCol-2005 deney derlemi eğitim ve test kümesi olarak ikiye ayrılmıştır. Bu amaçla, ilk sekiz ve son dört aylık haberler sırasıyla eğitim ve test derlemi olarak kullanılmıştır (Can ve diğerleri, 2010). Çalışmada Türkçe konu tespit ve takibinde gövdelemenin etkileri de incelenmiştir. Belgeleri tanımlamada kullanılan vektörlerin saptanmasında üç farklı gövdeleme yöntemi (no-stemming, fixed prefix stemming, lemmatizer-based stemming) kullanılmıştır (Can ve diğerleri, 2010). Çalışmada varlık isimlerinin kullanımı da test edilmiştir. Bu amaçla, Uyar'ın (2009) tezinde de bahsedildiği üzere 60.267 başlıktan oluşan varlık isimleri derlemi oluşturulmuştur. Bu derlemde kişi isimleri Türk Dil Kurumu web sayfasında sunulan kişi isimleri sözlüğünden ve Bilkent Üniversitesi personel, öğrenci ve lise öğrencilerine ait bilgilerin yer aldığı veri tabanından üretilmiştir. Yer isimleri içinse yine Bilkent Üniversitesi'nin veri tabanından yararlanılmış ve şehir, ilçe, semt isimleri de bu veri tabanına eklenmiştir. Kurum isimleri TRT, TÜBİTAK, MEB gibi sık kullanılan kurum isimlerinden elle türetilmiştir. Bunlara ek olarak varlık ismi tanımada büyük harfle başlayan kelimelerin içeriğine bakma gibi sezgisel yollardan da yararlanılmıştır (Can ve diğerleri, 2010). Uygulama aşamasında belge vektörlerinin kullanımında varlık isimleri dışındaki kelimeler, sadece varlık isimleri, tüm kelimeler olmak üzere üç farklı benzerlik skoru ve üçgenleştirme yaklaşımı (triangularization approach) kullanılmıştır. Uygulama sonucunda, sadece tüm kelimelerin kullanıldığı durumda en iyi performansın alındığı gözlenmiştir. Varlık isimleri ile ilgili bu sonucun nedeni, kullanılan test derlemindeki konu hikâyelerinin varlık isimlerinin kullanımında iletken olmaması olasılığına bağlanmıştır (Can ve diğerleri, 2010). Çalışma sonuçlarına

dayanarak Türkçe'de benzer konu tespit ve takip uygulamalarında, kosinüs benzerliği (cosine-similarity) ölçümünün ve dizinlemede kök çözümleme tabanlı (lemmatized-based) gövdeleme yaklaşımının kullanımı önerilmiş, yeni olay belirleme çalışmalarında durma kelimeleri listesi oluşturmanın sistem etkinliği açısından önemli olduğu ortaya konmuştur. Çalışmada ayrıca iki farklı benzerlik ölçümüne ait güven skorlarının daha yüksek etkililik için basit şekilde birleştirilebileceği de vurgulanmıştır (Can ve diğerleri, 2010).

Aksoy, Can ve Koçberber'in 2012 yılında gerçekleştirdikleri çalışmalarında belirli bir konuya yönelik haber takibinde yenilik tespiti sorunları ele alınmış, bu amaçla BilCol-2005'e dayanarak BilNov-2005 adı verilen bir başka Türkçe deney derlemi oluşturulmuştur. Çalışmada, yenilik tespitinde kosinüs benzerlik ölçümü, dil modeli gibi yöntemlerinin kullanımı önerilmiştir. Ayrıca kategoriye dayalı eşik öğrenme (category-based threshold learning) yöntemi yeni olay tespitine yönelik literatürde ilk kez kullanılmıştır. Araştırma sonuçları, dil modeline dayalı yeni olay tespit yönteminin kullanılan diğer yöntemlerden daha üstün performans sergilediğini göstermiştir. Bunun yanı sıra, kategoriye dayalı eşik öğrenmenin, genel eşik öğrenme (general threshold learning) ile karşılaştırıldığında daha başarılı sonuçlar ortaya koyduğu gözlenmiştir (Aksoy, Can ve Koçberber, 2012).

"Türkçe Haber Benzerliklerinin Belirlenmesinde Varlık İsimlerinin Hikâye Bağlantı Algılama Görevinin Başarımına Etkisi" başlıklı TÜBİTAK Projesi (Proje No: 111K030) kapsamında, KTT programı içerisinde tanımlı Hikâye Bağlantı Algılama görevinin Türkçe bir derlem üzerinde farklı erişim fonksiyonları ve bunların kombinasyonları denenerek başarımın test edilmesi ve optimum anma/duyarlık değerlerini sağlayacak kombinasyonun bulunması amaçlanmıştır. Bu amaca yönelik olarak ilk adımda derlem varlık isimleri ("Kişi", "Kurum", "Konum", "Tarih", "Zaman", "Para" ve "Yüzde") ile etiketlenmiş, varlık isimleri ile tanımlanamayacağı düşünülen kelimelere ise "Unknown" etiketi eklenmiştir. Bu aşamanın ardından, test senaryoları oluşturulmuş ve son adımda gerekli yazılımlar geliştirilerek testler uygulanmıştır. Projedeki testler BilCol-2005 haber derleminde ilgililik değerlendirmeleri yapılmış olan haberler üzerinde gerçekleştirilmiştir. Çalışma sonucunda, haber benzerliklerindeki anma, duyarlık ve f-ölçü değerlerinin belirlenmesinde vektör uzayı modelinin varlık isimleri ile etiketlenmiş haberler üzerinde yüksek başarım sağladığı görülmüştür. Diğer yandan, haberlerdeki tüm varlık isimlerinin vektörlerle ifade edildiği ve bu varlık isimlerinin vektör uzayı modeli yöntemi

kullanılarak gerçekleştirilen testlerde en yüksek başarım 0,67 f-ölçü değeri ile “unknown” etiketli varlık isimlerinde elde edildiği belirtilmiştir (Soydal ve Al, 2014).

BilCol-2005 haber derlemine yönelik gerçekleştirilen çalışmalar değerlendirildiğinde, derlemin gerçekten de Türkiye’de gerçekleştirilen pek çok bilgi erişim çalışmasına aracı olduğu anlaşılmaktadır. Özellikle KTT alanında farklı yöntem ve tekniklerin denenmesi ve geliştirilmesi açısından etkili olan bu derlem çalışmamızın verilerini sağlaması açısından da önemlidir.

3. BÖLÜM

ARAŞTIRMA TASARIMI

3.1. GİRİŞ

Bu çalışmada, haber metinlerinde etiketlenen varlık isimlerinin (named entities) haberlerin dâhil oldukları temel düzey IPTC (International Press Telecommunications Council) konu başlıkları (Bilim ve teknoloji, Siyaset, Sanat, kültür ve magazin, vb.) ile ilişkisinin incelenmesi ve hangi varlık isminin/isimlerinin haberin konu başlığını belirlemede daha etkili olduğunun ortaya konulması amaçlanmaktadır. Söz konusu amaca yönelik olarak bu bölümde, verilerin toplanması, analize uygun hale getirilmesi ve analizi aşamalarının yanı sıra haberlerin etiketlenmesi ve IPTC konu başlıklarına atanması sürecinde uygulanan yöntemler konusunda bilgiler verilmektedir.

Çalışma sürecinde verilere yönelik gerçekleştirilen çalışma adımları özetle şu şekildedir:

VARLIK İSİMLERİ

- Etiketleme işlemi
    - Etiketlerin belirlenmesi
    - Etiketleme yazılımının oluşturulması
    - 5834 haberin tek tek okunarak içeriklerinde yer alan varlık isimlerinin etiketlenmesi
- Etiketli derlemin çekilmesi
    - XML formatındaki haberlerden etiketlerin ve etiketlenen kelimelerin çekilerek Microsoft Excel 2010 yazılımına aktarılması
    - Format değişikliğinden ve elle etiketlemeden kaynaklanan hataların tespit edilmesi ve bunların düzeltilmesi için makro ve formüller yardımıyla veri temizliği yapılması (bkz. Ek 2)

- Analiz aşaması
    - Analiz için saydırma makrosunun uygulanması (bkz. Ek 3, Ek 4 ve Ek 5)

IPTC KONU BAŞLIKLARI

- BilCol-2005 haber derleminde yer alan 80 haber başlığı (Sahte rakı, Universiade 2005, Londra metrosunda patlama, Türkiye'de kuş gribi, vb.) uzman gazetecilere gösterilerek haberlerin ilgili IPTC haber konu başlığına atanmasının sağlanması (bkz. Ek 1)

Özetlenen bu adımlar aşağıdaki alt başlıklarda detaylandırılmaktadır.

3.2. VERİLERİN TOPLANMASI

Araştırmamızın amacı doğrultusunda üzerinde deneyler gerçekleştirilecek veri seti olarak "Türkçe Haber Benzerliklerinin Belirlenmesinde Varlık İsimlerinin Hikâye Bağlantı Algılama Görevinin Başarımına Etkisi" başlıklı TÜBİTAK Projesi (Proje No: 111K030) kapsamında etiketlenmiş olan BilCol-2005 haber derlemi kullanılmıştır. Bilkent Üniversitesi'nde geliştirilen ve literatürde farklı çalışmalarda yararlanılan BilCol-2005 (Can ve diğerleri, 2010) haber derlemi KTT (Konu Tespit ve Takip) çalışmalarından esinlenerek hazırlanmıştır. Bu derlem 209.296 gazete haberinden oluşturulmuştur. Derlem içerisinde geçen haberlerden 5872 tanesi BilCol-2005 kapsamında önceden belirlenmiş 80 farklı haber başlığı ile ilişkilendirilmiştir. Bu çalışmayı gerçekleştiren araştırmacılar (Can ve diğerleri, 2010) kalan tüm haberlerin bu haber başlıkları ile ilgisiz olduğunu kabul etmiştir. Bu nedenle belirtilen proje kapsamında yalnızca bu 80 haber başlığı altında yer alan 5872 haber üzerinden testler yapılmıştır. Araştırmamız kapsamında da bu veri seti kullanılmıştır.

KTT, haber metinleri içerisinde ifade edilen olaylar (events) ile doğrudan ilgilidir ve KTT çalışmalarında bir olay; özel bir mekânda, belirli kişi ya da kurumların katılımı ile belirli bir zaman diliminde gerçekleşen eylemler olarak tarif edilmektedir (Shah, Croft ve Jensen, 2006). Bu nedenle, KTT çalışmalarında bir haber içeriğinin gösteriminde varlık isimlerine yer verilmesi, KTT içerisindeki olay kavramının tanımı ile eşleşmesi bakımından bir zorunluluk gibi görünmektedir (Soydal ve Al, 2014). Buradan yola

çıkılarak proje kapsamında etiketleme işlemi gerçekleştirilmiştir. Bahsi geçen proje, metin içerisindeki varlık isimlerinin otomatik yöntemlerle çıkarılmasını sağlayan makine öğrenme yöntemleri Türkçe söz konusu olduğunda yetersiz olabildiği ve varlık isimlerinin belge benzerliklerinin belirlenmesindeki etkilerine odaklandığı için bu haberler elle etiketlenmiştir (Soydal ve Al, 2014).

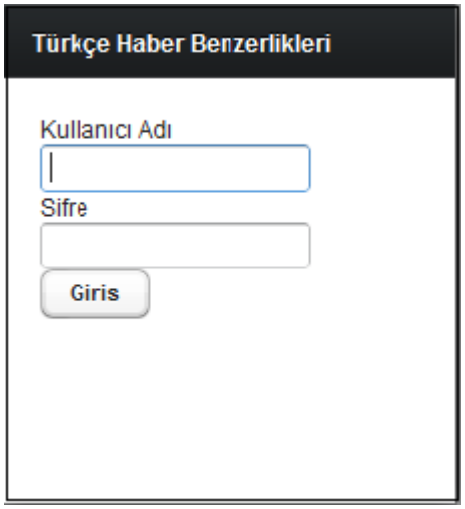
3.3. VERİ GİRİŞİ

Bu bölümde veri girişi aşamasında gerçekleştirilmiş olan derlem etiketleme ve derlemdeki haberlerin IPTC konu başlıklarına atanması işlemlerinden bahsedilmektedir.

3.3.1. Etiketleme

Sunulan haberin içeriğine bağlı olarak, metin tabanlı haberlerde geçen ifadelerden temel olarak “Ne”, “Ne zaman”, “Nerede”, “Kim” sorularına yanıt olabilecek kelimelerin işaretlenmesi işlemi etiketleme olarak ifade edilmektedir. Bu işlem sırasında, BilCol-2005 derlemi içerisinden alınmış ve başlıkları net olarak belirlenmiş olan 5872 haberin okunması ve haber metni içerisindeki kelimelerin özenle seçilerek doğru bir şekilde işaretlenmesi gerekmiştir.
Bu etiketleme çalışmasını gerçekleştirmek amacıyla Java tabanlı bir Web uygulaması geliştirilmiş ve proje çalışmalarına katılan Hacettepe Üniversitesi Bilgi ve Belge Yönetimi Bölümü 3. sınıf ve yüksek lisans öğrencilerinin bu uygulamayı kullanarak hızlı ve etkin bir biçimde etiketleme yapabilmeleri sağlanmıştır.

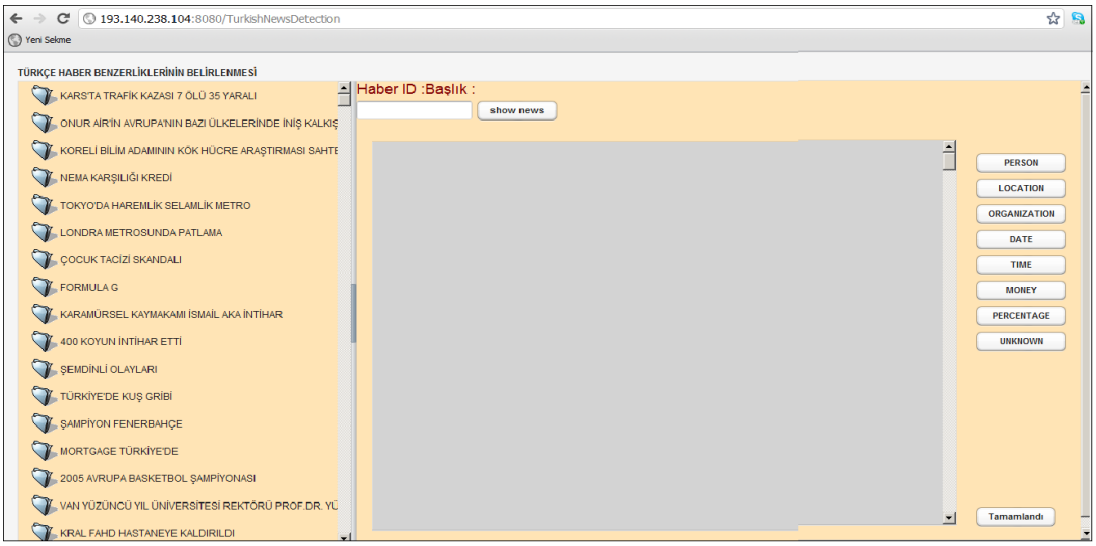
Hazırlanan Etiketleme Programı, Hacettepe Üniversitesi sunucuları üzerine uzaktan erişilebilecek şekilde yerleştirilmiştir. Uygulama çalıştırıldığında ilk olarak, etiketleme yapan veya yetkili kullanıcının giriş yapmasını sağlamak üzere oluşturulmuş, “Giriş Penceresi” açılmaktadır (bkz. Şekil 3).



Şekil 3. Etiketleme sistemi "Giriş Penceresi"

Etiketleme yapan kullanıcı veya yetkili kullanıcı bu pencereyi kullanarak "Kullanıcı Adı" ve "Şifre" bilgilerini yazdıktan sonra sisteme girebilmektedir. Sistem, uygulamaya giriş yapan kişinin etiketleme yapan kullanıcı mı yoksa yetkili kullanıcı mı olduğunu anlayarak giriş yapan kişiye özel bilgileri ekrana getirmektedir.

Yetkili kullanıcı arayüzünde ise etiketleme yapılacak bütün haberlerin başlıkları ve bu başlıklar ile ilgili haberler görüntülenmektedir (bkz. Şekil 4).

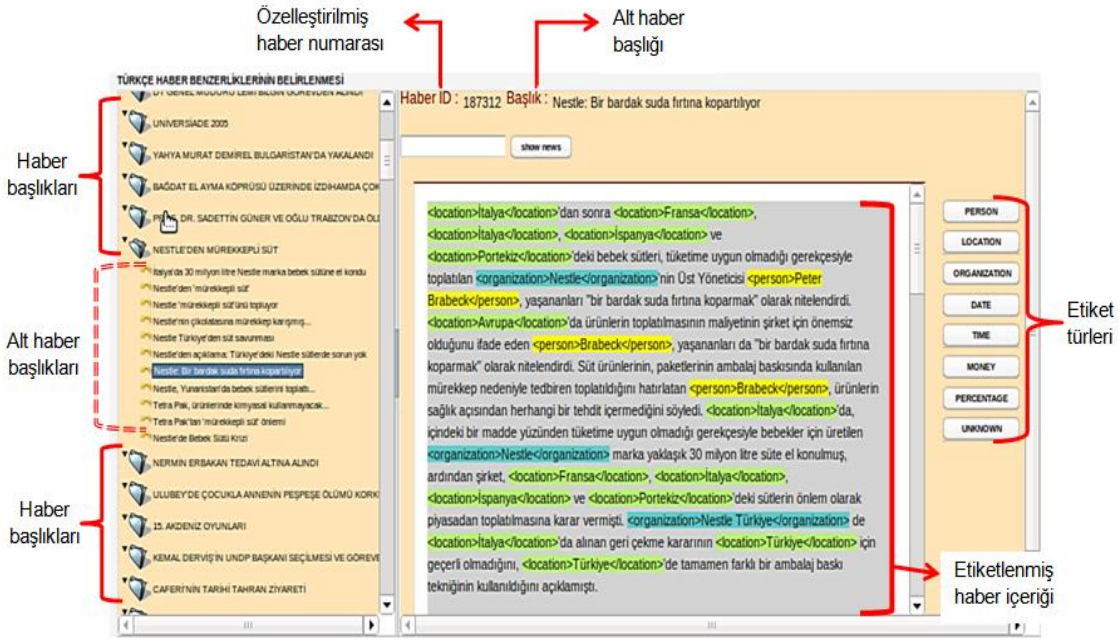


Şekil 4. Yetkili kullanıcı arayüzü

Yetkili kullanıcı bu arayüz aracılığı ile hangi haber başlığında ne kadar etiketleme yapıldığının kontrolünü gerçekleştirebilmekte ve istediği habere müdahale edebilmektedir.

Etiketleme yapan kullanıcı, sisteme giriş yaptığında sol menüde kendisine atanan haber başlıklarını görmekte ve etiketleme yapacağı başlığı seçip etiketleme işlemine başlayabilmektedir. Haber içeriğinin gösterimi ile etiketleme yapan kullanıcı haberi okuyabilmekte ve gerekli gördüğü kelimeleri sağ tarafta bulunan butonları kullanarak etiketleyebilmektedir.

Kullanıcı, etiketleme işlemini, gerekli gördüğü kelimeyi seçtikten sonra uygun olan etiket butonuna basarak gerçekleştirebilmektedir (bkz. Şekil 5).



Şekil 5. Etiketlenmiş haber örneği

Seçilen haber için etiketleme işlemi tamamlandığında, ekranın sağ alt kısmında yer alan “Tamamlandı” butonu kullanılarak etiketleme işlemi sonlandırılmaktadır. Gerekli durumlarda ilgili habere geri dönerek düzenlemeler yapmak mümkündür.

Verilerin sağlandığı proje kapsamında değerlendirilen çalışmalara dayalı olarak varlık isimleri *“kurum (organization)”, “kişi (person)”, “konum (location)”, “tarih (date)”, “zaman (time)”, “yüzde (percentage*)”, “*para (money)*” ve *“bilinmeyen (unknown)”*[7] olarak etiketlenmiştir (Shah, Croft ve Jensen, 2006; Bikel, Schwartz ve Weischedel, 1999;

[7] 111K030 No.lu TÜBİTAK projesi kapsamında etiket eklenmesine gerek olup olmadığına karar verilemeyen ifadeler “unknown” olarak etiketlenmiştir. Bu etiketler çalışmamızda etiketsiz metnin kapsamına dâhil edilmiştir.

Kumaran ve Allan, 2004; Dalkılıç, Gelişli ve Diri, 2010; Tür, Hakkani-Tür ve Oflazer, 2003; Bayraktar ve Taşkaya-Temizel, 2008; Küçük ve Yazıcı, 2009a; Küçük ve Yazıcı, 2009b; Küçük ve Yazıcı, 2010). Bu sayede hem bu proje kapsamında belirlenen yöntemler test edilebilmiş hem de oluşturulan etiketlenmiş derlemin çok daha geniş bir akademik çevre tarafından kullanılabilmesi hedeflenmiştir (Soydal ve Al, 2014). Haberlerde yer alan varlık isimleri aşağıdaki örneklerde görüldüğü şekilde etiketlenmiştir.

“Deniz Baykal”
<person> Deniz Baykal </person>

“İzmir”
<location>İzmir</location>

“Ziraat Bankası”
<organization>Ziraat Bankası</organization>

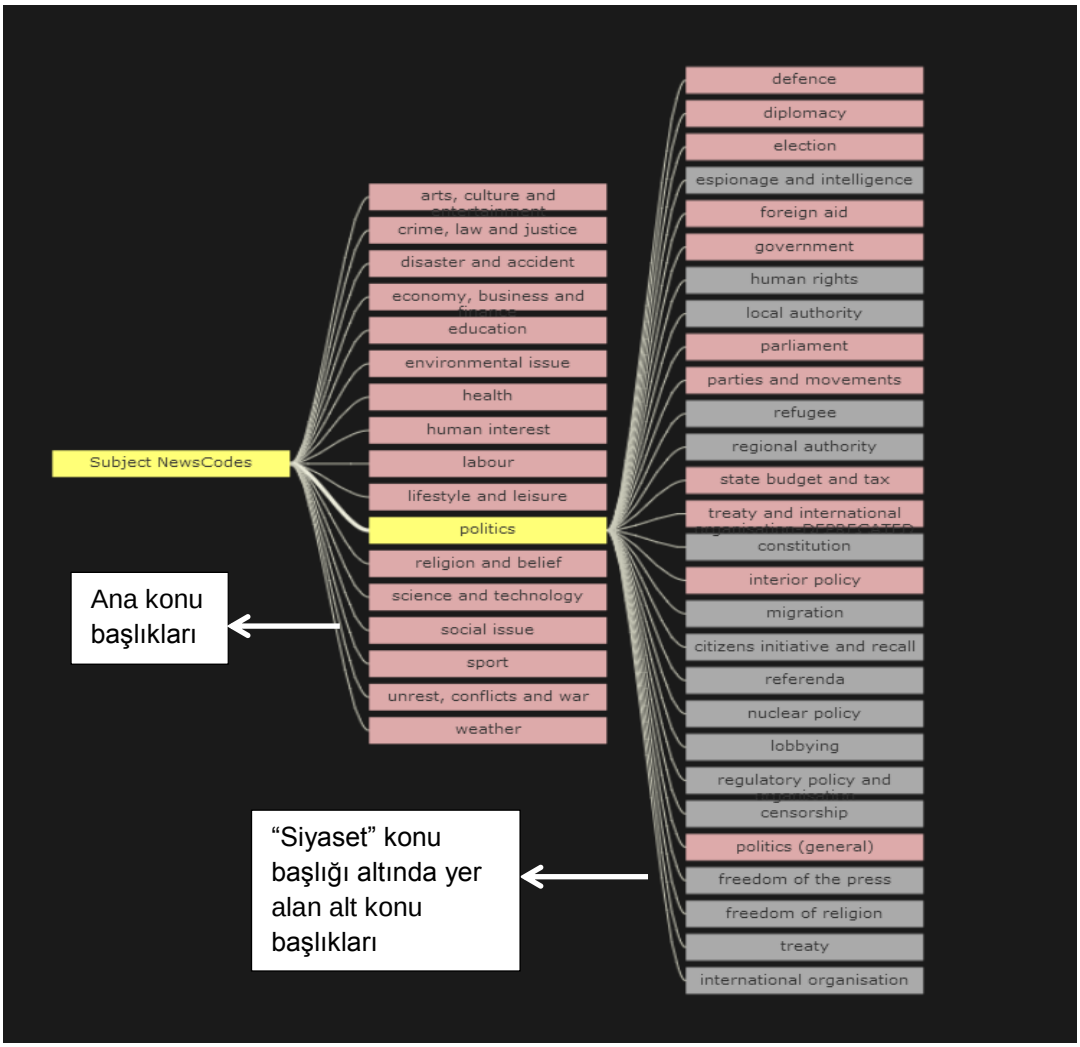
“saat 15:30”
<time>15:30</time>

Bu kapsamda, kişi adı geçen ifadelerin “Kişi” (person), üniversite, hastane, okul benzeri isimlerin “Kurum” (organization), ülke, şehir, eyalet, ilçe isimlerinin “Konum” (location) olarak etiketlenmesi gibi birtakım kurallara bağlı kalınarak BilCol-2005 derleminde haber başlıkları bilinen 5872 haber, etiketleme çalışmasında görev yapan öğrencilere paylaştırılmış ve öğrencilerin ilgili yazılım üzerinde varlık isimlerini belirlemeleri sağlanmıştır. İlk etiketleme işlemi bittikten sonra haberlerde etiketlenen varlık isimleri ikinci bir kontrolden geçirilerek derleme son hali verilmiştir (Soydal ve Al, 2014).

3.3.2. International Press Telecommunications Council (IPTC – Uluslararası Basın Telekomünikasyon Konseyi) Konu Başlıkları Uygulaması

IPTC ile ilgili kavramsal bilgilere Bölüm 2’de detaylı olarak yer verilmişti. IPTC’nin işlerlik sağlayan standart yapılarından biri olan IPTC haber konu başlıkları (Subject Code) diyagramı, haber ögelerinin ait olabilecekleri konuları belirtmektedir. Bu konu başlıkları IPTC’nin geliştirmiş olduğu, metne odaklanan özgün konu taksonomisine dayanmaktadır (bkz. Şekil 6). Başlıklar genelden özele giden bir yapıda üç düzeye ayrılmış yaklaşık 1400 terimden oluşmaktadır (IPTC news codes, 2014).

Çalışmamızın amacına bağlı olarak, haberlerin kategorizasyonu aşamasında haberlerin uygun konu başlıklarına atanabilmesi için belli konu başlıklarına ihtiyaç duyulmuştur. Türkiye'de yer alan gazete (Milliyet, Hürriyet, vb.) ve diğer haber kaynakları (ntvmsnbc, CNN Türk, TRT, vb.) kendi kurum politikalarına ya da kendi belirledikleri unsurlara göre sundukları haberleri belirli konu başlıkları altında aktarmaktadırlar. Her bir kaynağın kullandığı konu başlığı sayısı farklılık gösterebilmektedir ve tüm haber kaynaklarının kullanmış olduğu ortak bir konu başlığı şemasına rastlanamamıştır. Çalışma kapsamında gerçekleştirilecek haber kategorizasyonu işlemini uluslararası bir standart yapıya dayandırabilmek amacıyla IPTC haber konu başlıkları temel alınmıştır.



Şekil 6. IPTC "Siyaset" konu başlığı ekran görüntüsü[8]

[8] IPTC haber kodları taksonomisine şu adresten erişilebilir: http://show.newscodes.org/index.html?newscodes=subj&lang=en-GB

Verilerin çok detaylı kategorilere bölünmesinin (üzerinde çalışılan derlemin boyutu nedeniyle) çalışmada hedeflenen analizleri zorlaştıracağı düşünüldüğünden çalışma kapsamında amaçlanan deneylerin yapılabilmesi için IPTC konu başlıkları içerisinden sadece temel düzey başlıklar esas alınmıştır.

BilCol-2005 haber derleminde önceden saptanmış olan 80 haber başlığı IPTC tarafından belirlenmiş olan haber konu başlıklarına atanmıştır. Bu işlemin gerçekleştirilmesi aşamasında Milliyet, Vatan, Akşam ve Posta gazetelerinde görev yapan haber müdürü, muhabir ve köşe yazarı gibi uzman beş gazeteci ile 26-27 Şubat 2014 tarihlerinde görüşme yapılarak görüşleri alınmıştır. Gazetecilere 80 haber başlığı listesi ile IPTC'de belirtilen 17 konu başlığı (bkz. Ek 6) verilmiş olup, her haber başlığını 17 IPTC konu başlığından en ilgili olan bir tanesine atamaları istenmiştir. Gerçekleştirilen atamalar neticesinde ortak görüşler temel alınarak Tablo 1'deki örnekte verildiği şekilde haber başlıkları en uygun IPTC konu başlığına yerleştirilmiştir.

Tablo 1. Derlemdeki haberlerin atandığı IPTC konu başlıkları (ilk 10 haber başlığı)[9]

Haber Başlığı No.	Haber başlığı	IPTC konu başlığı
1	Kars'ta Trafik Kazası 7 ölü 35 yaralı	Kaza ve felaket
2	Onur Air'in Avrupa'nın bazı ülkelerinde iniş kalkışının yasaklanması	Ekonomi, işletme ve finans
3	Koreli bilim adamının kök hücre araştırması sahte	Bilim ve teknoloji
4	Nema karşılığı kredi	Ekonomi, işletme ve finans
5	Tokyo'da Haremlik Selamlik Metro	Sosyal konular
6	Londra metrosunda patlama	Savaş ve karışıklıklar
7	Çocuk tacizi skandalı	Suç, hukuk ve yargılama
8	Formula G	Spor
9	Karamürsel Kaymakamı İsmail Aka İntihar	Sosyal konular
10	400 Koyun İntihar etti	Yaşam ve ilgi alanları

3.4. VERİLERİN TEMİZLENMESİ VE ANALİZE UYGUN HALE GETİRİLMESİ

Tez çalışmamızın en emek yoğun kısmını oluşturan veri temizleme aşamasında, 111K030 no.lu TÜBİTAK projesinde üzerinde çalışılmış olan 5872 haber ve bunlara uygulanan etiketler gözden geçirilerek birtakım güncelleme ve düzenlemeler yapılmıştır. Veri temizleme işlemi esnasında sadece haber başlığından oluştuğu ya da bir iki cümle gibi oldukça kısa haber metinleri içerdiği tespit edilen haberler yeterli veri

[9] Tablonun tamamı Ek 1'de sunulmuştur. Derlemde yer alan başlıklar sisteme girildiği şekliyle aktarılmış, noktalama işaretleri ve imlâda herhangi bir düzeltme yapılmamıştır.

sağlamayacağı düşünüldüğü için analiz dışı bırakılmış ve araştırmamızdaki deneyler 5834 haber üzerinde gerçekleştirilmiştir.

Veri temizleme aşamasında öncelikle etiketlenmiş olan XML formatında 5834 haber metni veri tabanından çekilerek kaydedilmiştir. XML formatındaki haberler her bir haberin haber başlığı numarası (TopicID'si), kaynağı, tarihi, etiketlenen kelimeler ve varlık isimlerini içerecek şekilde Excel dosyasına aktarılmıştır. Excel dosyasında yer verilen tüm bu bilgilerin yanı sıra her haberin hangi IPTC konu başlığına karşılık geldiği bilgisi de sonradan dosyaya eklenmiştir

Tez çalışmamız kapsamında etiketler de gözden geçirilmiş olup, bunlarda da birtakım düzeltmeler yapılmıştır. "Unknown" etiketi ile etiketlenen ifadeler yeniden değerlendirilmiş ve bu ifadelerin herhangi bir etiket değeri taşımadığı tespit edilmiştir. Bu nedenle "Unknown" etiketi bu araştırma kapsamında göz ardı edilerek etiketlenmemiş metne (çalışmamız kapsamında "Ne" sorusuna yanıt verdiğini kabul ettiğimiz kısım) dâhil edilmiştir.

Bazı haberlerde etiketleme sisteminde oluşan teknik bir problemden kaynaklandığı düşünülen metin tekrarı görülmüştür. Örneğin;

Haber ID: 114845, Haber başlığı: ABD, metrolarda cep telefonunu bloke ediyor!

<location>İngiltere)</location>
daha önce saldırıldığını kaydetti. ? ABD'lilerin yüzde 25'i yeni bir terör saldırısından "çok endişeli"... ? ABD'lilerin yüzde 25'i yeni bir terör saldırısından "çok endişeli"... ABD'lilerin yüzde 25'i yeni bir terör saldırısından "çok endişeli"... yüzde 25'i yeni bir terör saldırısından "çok endişeli"... 25'i yeni bir terör saldırısından "çok endişeli"... yeni bir terör saldırısından "çok endişeli"... bir terör saldırısından "çok endişeli"... terör saldırısından "çok endişeli"... saldırısından "çok endişeli"... "çok endişeli"... endişeli"... tam teçhizatlı olarak nöbet tuttukları gözleniyor. Bazı uzmanlar, tünellerde cep telefonu hizmetinin bloke edilmesinin terörist saldırıları engelleme açısından önemli bir adım olduğunu savunurken, bazıları ise bunun teröristleri engellemeyeceğini, yerleştirdikleri bombaları cep telefonu ile uzaktan patlatamayan teröristlerin bunun yerine bomba düzeneklerine zamanlayıcı yerleştirerek saldırılarını düzenleyebileceklerini ifade ediyor.

Şekil 7. Etiketli metinde tekrarlanan kısım

Şekil 7'de görülen kısmın haber metninin içinde 185'ten fazla kez tekrarlandığı belirlenmiştir. Buna benzer durumda olan toplam üç haber tespit edilmiş ve haberlere müdahale edilerek tekrar eden cümleler ayıklanmıştır.

Veri temizleme esnasında karşılaşılan problemlerden bir diğeri de karakter hatalarıdır. Örneğin;

```
Fatal error at line 2093807 column 2 : invalid character 0xF

D:\>xml2csv deneme.txt deneme.csv DOCID,TEXT -E -Q
"A7Soft xml2csv 5.33" - xml to csv converter
```

Şekil 8. Karakter hatası görüntüsü

Verilerin, XML formatta yer aldığı dosyadan çekilerek Excel'e aktarılması aşamasında Şekil 8'deki karakter hatası ile sıkça karşılaşılmıştır. Bir başka karakter hatası da haberler çekilirken çoğunlukla haber sonlarında kalan özel karakterlerde yaşanmıştır. Bu özel karakterler (‘ tek tırnak, “” çift tırnak ve -tire) makro çalışırken makinaca bu ifadelerden sonra komut gelecekmiş gibi algılandığı için hata vermiş ve bahsi geçen karakter hataları tek tek metnin içine gidilerek düzeltilmiştir.

Metin ön işleme ve gövdeleme, bilgi erişim çalışmalarının ilk aşamalarını oluşturmaktadır. Metin ön işleme aşaması, üzerinde çalışılan veri setindeki kelimelerde yer alan boşlukların, noktalama işaretlerinin atılması gibi işlemleri içerirken, gövdeleme işlemi de kelimelerin çekim eklerinden arındırılarak kök ve gövde hallerine indirgenmesi sürecini ifade etmektedir. Gövdeleme işlemi, doğruluk ve performans açısından bilgi erişim başarısını artırması sebebiyle önem arz etmektedir (Moral, Antonio, Imbert ve Ramirez, 2014).

Çalışmamız kapsamında etiketleme işleminin ardından haberlerin XML formatında çekilmesi sonucu bazı etiketli ifadelerde veri temizleme işlemine ihtiyaç duyulduğu fark edilmiştir. Veri temizliğinin yapıldığı etiketli kelimelere ilişkin örneklerden birkaçı Şekil 9'da gösterilmektedir:

Veri temizliğinden	
Önce	**Sonra**
"CHP'de	chp
Ahmet Ocak?ı,	ahmet ocak
?Baykal	baykal
(Gülden,	gülden
Cem Duman-12'inci	cem duman
?Şemdinli?de	şemdinli

Şekil 9. Veri temizleme işleminin gerçekleştirildiği etiketli kelimelere örnekler

Veri temizleme esnasındaki işlemlerin gerçekleştirilmesinde makrolardan ve formüllerden yararlanılmıştır (bkz. Ek 2). İlk olarak tekil etiketlerin belirlenebilmesi için etiketler üzerinde şu çalışmalar yapılmıştır:

- Parantezlerin (), eşittir (=), tire (-), iki nokta (:), taksim (/), kesme (‘), tırnak (“), virgül (,), nokta (.), (=-) işaretlerinin ve kesme (‘) işaretinden sonraki eklerin atılması.
- Tüm kelimelerin küçük harf haline getirilmesi (LOWERCASE).
- Etiketlenen kelimelerin başındaki ve sonundaki boşlukların atılması (TRIM).
- Metinde iki farklı şekilde geçtiği tespit edilen “yüzde” ve “%” ifadelerinin “yüzde” şeklinde tek biçime dönüştürülmesi.

3.5. VERİLERİN ANALİZİ

Verilerin analizinde Microsoft Excel 2010 ve IBM SPSS Statistics 21 yazılımlarından yararlanılmıştır. Yukarıda bahsi geçen veri temizleme işlemleri sırasında, parantezlerin (), eşittir (=), tire (-), iki nokta (:), taksim (/), kesme (‘), tırnak (“), virgül (,), nokta (.), (=-) işaretlerinin ve kesme (‘) işaretinden sonraki eklerin atılması için metin düzeltme makrosu (bkz. Ek 2) kullanılmıştır. Tüm kelimelerin küçük harf haline getirilmesi (LOWERCASE) ile etiketlenen kelimelerin başındaki ve sonundaki boşlukların atılması (TRIM) işlemleri için ise Microsoft Excel 2010’da yer alan formüller kullanılmıştır. Uygulanan etiketleme işlemi neticesinde veri analizinin gerçekleştirilebilmesinde gerekli olan her bir haber için etiket sayısı, tekil etiket sayısı ile toplam etiket sayılarının ve her bir haberde kaçar tane etiket tipi ile etiketleme yapıldığı ve kaç farklı etiket tipi kullanıldığının hesaplanabilmesi için etiket saydırma makrosu (bkz. Ek 3)

oluşturulmuştur. Bu makrolar aracılığıyla elde edilen veriler değerlendirilerek tablolar ve radar grafikler oluşturulmuştur.

Etiketli ifadeler için gerekli analizler gerçekleştirildikten sonra metin düzeltme makrosu yeniden düzenlenerek "NE" sorusuna karşılık geldiği kabul edilen, metnin geri kalan kısmındaki (etiketli ifadeler çıkarıldıktan sonra kalan kısım) kelimeler üzerinde uygulanmıştır. Daha sonra etiketsiz metinde kaç kelime geçtiğini hesaplamaya yönelik kelime saydırma makrosu (bkz. Ek 5) çalıştırılarak kelimelerin sıklıkları elde edilmiş ve kelimeler analize uygun hale getirilmiştir.

Araştırma sorularına yönelik olarak, derlemdeki haberlerin etiketlenmesinde en baskın ve en pasif varlık isimleri ile her bir konu başlığı için en belirleyici varlık ismi/isimlerinin hangileri olduğunun belirlenebilmesi için öncelikle haberler varlık isimleri ile etiketlenmiştir. Etiketlenen haberler daha sonra en uygun IPTC konu başlıklarına atanmıştır. Konu başlıkları belirlenen haberlerde geçen varlık isimleri ve bunlarla etiketlenmiş kelimelerin sıklıkları ile oranları hesaplanarak, derlemin genelindeki en baskın ve en pasif varlık isimleri ile birlikte her bir konu başlığı için en belirleyici olan varlık isimleri elde edilmiştir. Derlemdeki haberlerin etiketlenme sayılarının IPTC konu başlıklarına göre farklılık gösterip göstermediğini belirleyebilmek için ilk olarak, konu başlıkları altında yer alan haberlere ilişkin etiket sayılarının normal dağılım durumuna bakılmıştır. Uygulanan Kolmogorov-Smirnov Testi ile verilerin normal dağılmadığı tespit edildiğinden farkların anlaşılabilmesi için Kruskal Wallis ve Mann-Whitney U Testleri yapılmıştır. Etiketli ve etiketsiz kelimelerin konu başlıklarını kavramsal olarak yansıtıp yansıtmadığını araştırdığımız araştırma sorumuz için ise bir çeşit içerik analizi ve göz kontrolü yapılmıştır. Bunun için derlemde etiketli ve etiketsiz olarak en çok geçen ilk on kelime belirlenmiş ve konu başlıkları altında tablo ile gösterilerek göz kontrolü gerçekleştirilmiştir.

3.6. SINIRLILIKLAR

Verilerin temizlenmesi ve analizi aşamasında birtakım sınırlılıklarla karşılaşılmıştır. Bahsi geçen sınırlılıkları şu şekilde sıralamak mümkündür:

- Veriler proje kapsamında temizlenmiş olmasına rağmen tez sürecinde tekrar gözden geçirilmiş ve etiketlerde birtakım hatalar olduğu fark edilmiştir. Buna bağlı

olarak sadece haber başlığından oluşan ya da bir iki cümlelik haber metninden oluşan kısa haberler yeterli veri sağlamadığı için inceleme dışı bırakılmıştır.

- Veri analizi aşamasında Windows OS, Mac OS X gibi farklı işletim sistemine sahip makinalarda çalışılmasının verilerin yapısının bozulmasına yol açtığı saptanmıştır.

  Örneğin,

Mac OS X	**Windows OS**
tŸrkiye	türkiye
gšrŸs	giriş
hŸcre	hücre
yazõlõm	yazılım

  Bu sorunla tekrar karşılaşmamak için çalışmaya sadece Windows işletim sistemi kullanılarak devam edilmiştir. İleride yapılacak bu tür çalışmalarda işletim sistemine dikkat edilmesi önerilmektedir.

- Kullanılan işletim sisteminin Türkçe ve İngilizce olarak dil açısından farklılık göstermesi, verilerde yukarıdaki örnektekine benzer şekilde Türkçe karakterlerin bozulmasına yol açtığı tespit edilmiştir.

- Veri sayısının çok olmasının yanı sıra sistem ve veri kaynaklı bazı öngörülemeyen hatalar ile karşılaşılmış olunmasından dolayı veri temizleme işlemi tahmin edilenden çok daha uzun sürmüş, bu durum bazı analizlerin tekrarlanması gerektiğinde zaman planlaması açısından sıkıntıya yol açmıştır.

4. BÖLÜM

BULGULAR VE DEĞERLENDİRME

4.1 GİRİŞ

Bu bölümde 5834 haberden oluşan derleme ilişkin genel bilgiler sunulmuş ve haberlere ait veriler değerlendirilmiştir. Öncelikle haberlerde uygulanan etiketleme işlemi neticesinde elde edilen bulgular incelenmiştir. Sonrasında ise haber metinlerindeki etiketli ifadeler atıldıktan sonra kalan ve "Ne" sorusuna karşılık geldiği kabul edilen ifadeler değerlendirilmiştir.

4.2 GENEL BULGULAR

Çalışmamıza temel oluşturan veri seti 5834 haberden oluşmaktadır. BilCol-2005'ten alınan bu haberler (%10'u (N= 578) TRT, %16'sı (N= 931) CNN Türk, %21'i (N= 1237) Haber 7, %22'si (N= 1266) Zaman, %31'i (N= 1822) Milliyet'ten olmak üzere) beş ayrı kaynaktan sağlanmıştır.[10]

Veri setini oluşturan haberler etiketlenerek Tablo 2 ile Şekil 10'da görülen varlık isimleri dağılımları elde edilmiştir.

Tablo 2. Varlık isimleri dağılımı

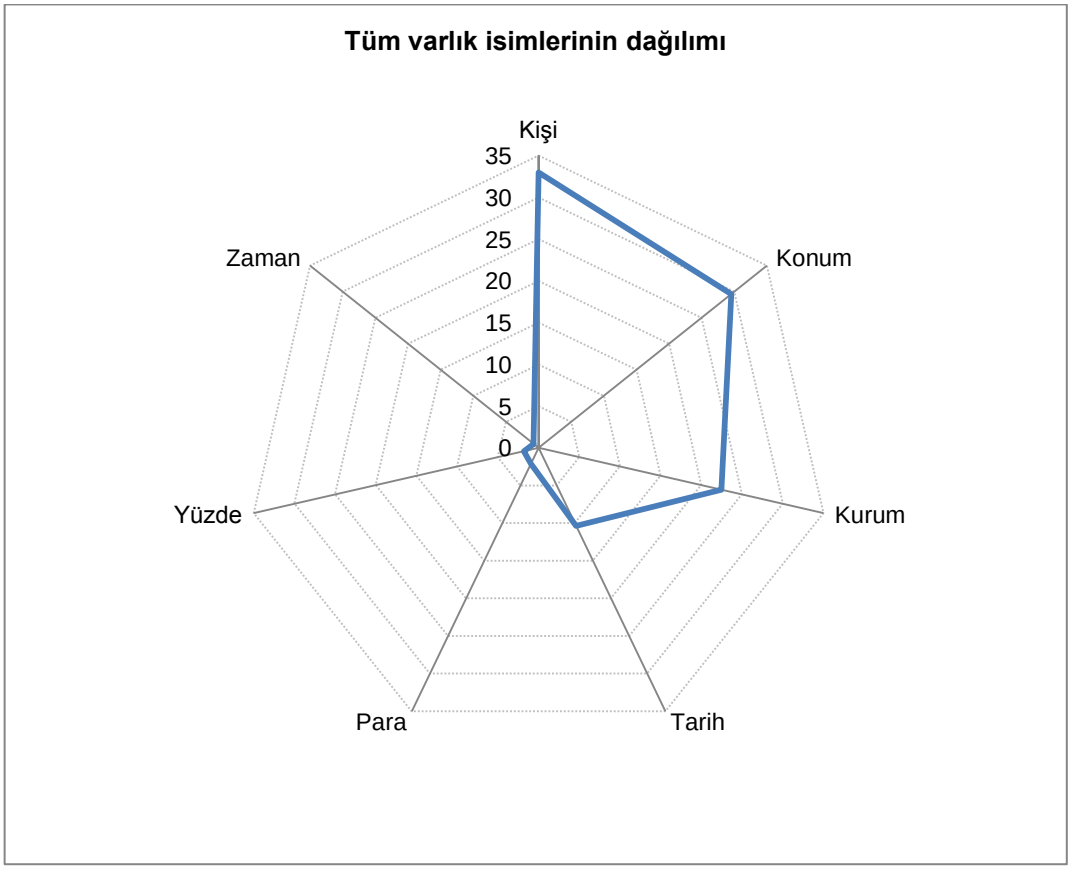
Varlık ismi	N	%
Kişi	50.969	33,0
Konum	45.608	29,5
Kurum	34.681	22,5
Tarih	16.021	10,4
Para	3.253	2,1
Yüzde	2.789	1,8
Zaman	1.154	0,8
Toplam	**154.475**	**100,1**

Not: Yuvarlama hatasından dolayı toplam %100'den farklıdır.

Haberlerin tamamında toplam 154.475 etiket bulunmaktadır. Haberler, "Kişi", "Konum", "Kurum", "Tarih", "Para", "Yüzde" ve "Zaman" olmak üzere yedi farklı varlık ismi ile

[10] Çalışmamızda kullanılan derlem 111K030 No.lu projede kullanılan verilere dayanmaktadır. Ayrıntılar için bkz. Bölüm 3.

etiketlenmiştir. Derlemde en fazla yer alan varlık ismi “Kişi” (N= 50.969, %33) etiketidir. Derlemdeki haberleri etiketlemede en az kullanılan varlık ismi ise “Zaman”dır (N= 1154, %0,8).



Şekil 10. Derlemdeki tüm varlık isimlerinin dağılımı (%)

Şekil 10’dan da görüldüğü gibi derlemdeki haberleri tanımlamada en baskın varlık isimlerinin “Kişi” (%33, N= 50.969), “Konum” (%29,5, N= 45.608) ve “Kurum” (%22,5, N= 34.681) olduğu anlaşılmaktadır. En az kullanılan varlık isimlerinin ise “Para” (%2,1, N= 3253), “Yüzde” (%1,8, N= 2789) ve “Zaman” (%0,8, N= 1154) olduğu görülmüştür.

4.3 DERLEMİN IPTC KONU BAŞLIKLARINA GÖRE KATEGORİZASYONU

Etiketleme işleminin ardından derlemdeki haberler, temel düzey IPTC konu taksonomisine dayanan 17 konu başlığı altında gruplanmak istenmiş, ancak sınıflamayı yapmaları istenilen gazetecilerin ortak görüşü ile derlemde, bu başlıklardan dört tanesinin altında sınıflanacak haber olmadığı belirlenmiştir. Gazetecilerin görüşleri dikkate alınarak derlemdeki haberler konularına en uygun olan 13 IPTC konu başlığı kullanılarak kategorize edilmiştir. BilCol-2005 derleminin oluşturulma sürecinde haber kategorizasyonu amacıyla "Olayı çağrıştıracak ve kolayca akılda kalan on kelimeden az bir cümle ya da kelimeler grubu" olarak tanımlanan (Can ve diğerleri, 2007, s. 56) 80 adet haber başlığı kullanılmıştır. Bu haber başlıkları, haber başlığı numaraları ve IPTC konu başlıkları ile birlikte Ek 1'de detaylı olarak sunulmuştur. Ayrıca haberlerin IPTC konu başlıklarına nasıl atandıkları bilgisine Bölüm 3'te (s. 30) değinilmiştir. Burada haber başlıklarının dağılımı ile haberlerin IPTC'ye göre dağılımının birbiriyle örtüşüp örtüşmediğine bakılmıştır. Bu amaçla, konu başlıklarına göre kategorizasyon ile haber sayı ve oranlarına ilişkin genel dağılım Tablo 3'te gösterilmektedir.

Tablo 3. Haberlerin IPTC konu başlıklarına göre dağılımı

IPTC Konu Başlığı	Haber başlığı		Haber		Haber başlığına düşen ortalama haber sayısı
	N	%	N	%	
Ekonomi, işletme ve finans	3	3,8	566	9,7	189
Savaş ve karışıklıklar	6	7,5	966	16,6	161
Spor	5	6,3	658	11,3	132
Sosyal konular	7	8,8	790	13,5	113
Suç, hukuk ve yargılama	10	12,5	1.088	18,6	109
Siyaset	6	7,5	407	7,0	68
Sağlık	5	6,3	322	5,5	64
İş yaşamı	1	1,3	53	0,9	53
Eğitim	1	1,3	53	0,9	53
Sanat, kültür ve magazin	9	11,3	290	5,0	32
Kaza ve felaket	20	25,0	514	8,8	26
Yaşam ve ilgi alanları	4	5,0	80	1,4	20
Bilim ve teknoloji	3	3,8	47	0,8	16
Toplam	**80**	**100,4**	**5.834**	**100,0**	**73**

Not: Bazı toplamlar yuvarlama hatasından dolayı %100'den farklıdır.

Tablo 3'ün ikinci sütununda, her bir konu başlığının 80 adet haber başlığından kaçar tanesini içerdiği verilmiştir. Örneğin, *Bilim ve teknoloji* konu başlığında üç farklı haber

başlığı (haber başlığı numaraları sırasıyla; 3, 19 ve 39) bulunmaktadır. Bu konu başlığındaki haber başlık numaraları ve ilişkili haber başlıkları Tablo 4'te görülmektedir.

Tablo 4. *Bilim ve teknoloji* konu başlığındaki haber başlık numarası ve haber başlıkları

Haber başlık numarası	Haber başlığı	IPTC konu başlığı
3	Koreli bilim adamının kök hücre araştırması sahte	Bilim ve teknoloji
19	Bill Gates türkiye'ye geldi	Bilim ve teknoloji
39	2005 Nobel Tıp Ödülü gastrit ve ülserin bakterilerden kaynaklanması	Bilim ve teknoloji

Derlemimizdeki haber dağılımına bakıldığında, derlemde çok fazla haber başlığının olması haber sayısının da fazla olduğunu göstermemektedir. Tablo 3'ten de görüldüğü üzere her dört haber başlığından biri *Kaza ve felaket* ile ilgili olmasına karşın, bu konu başlığı ile ilgili haberler çalıştığımız derlemde %8,8 oranında yer almaktadır. Derlem içinde haber sayısı en yüksek olan *Suç, hukuk ve yargılama* konu başlığı tüm haberler içinde %19 oranında temsil edilmiştir. Öte yandan, haber başlığına düşen ortalama haber sayısına göre bakıldığında ortalama 189 haber ile *Ekonomi, işletme ve finans* konu başlığının ilk sırayı aldığı görülmektedir.

Çalışmada incelenmesi hedeflenen bir diğer unsur, etiketleme işleminin ardından 13 IPTC konu başlığına atanan haberlerde bir konu başlığını tanımlamada yaklaşık kaç adet etiket kullanıldığı ile daha çok ve daha az etiketle etiketlenmiş konu başlıklarının neler olduğudur. Buna yönelik olarak, konu başlıklarına atanmış haberlere ilişkin etiket dağılımı Tablo 5'te sunulmaktadır.

Tablo 5. Varlık isimlerinin IPTC konu başlıklarına göre dağılımı

IPTC Konu Başlıkları	Kişi		Kurum		Konum		Tarih		Zaman		Para		Yüzde		Toplam	
	N	%	N	%	N	%	N	%	N	%	N	%	N	%	N	%
Suç, hukuk ve yargılama	12.843	42,0	6.683	21,8	6.354	20,8	3.631	11,9	167	0,5	831	2,7	97	0,3	30.606	100
Savaş ve karışıklıklar	6.466	22,4	4.944	17,1	13.058	45,2	4.127	14,3	156	0,5	47	0,2	86	0,3	28.884	100
Spor	7.448	42,4	4.798	27,3	3.408	19,4	1.446	8,2	188	1,1	270	1,5	23	0,1	17.581	100
Sosyal konular	5.038	30,7	3.692	22,5	6.257	38,1	1.039	6,3	101	0,6	177	1,1	131	0,8	16.435	100
Ekonomi, işletme ve finans	2.847	18,8	4.459	29,5	3.031	20,0	1.369	9,0	62	0,4	1.155	7,6	2.209	14,6	15.132	100
Siyaset	5.711	43,3	3.142	23,8	3.119	23,6	1.014	7,7	39	0,3	136	1,0	31	0,2	13.192	100
Sanat, kültür ve magazin	4.571	51,4	1.666	18,7	1.635	18,4	841	9,4	79	0,9	92	1.0	16	0,2	8.900	100
Sağlık	1.436	16,5	1.833	21,1	3.713	42,7	1.127	13,0	134	1,5	338	3,9	111	1,3	8.692	100
Kaza ve felaket	2.371	28,5	1.592	19,1	3.636	43,7	493	5,9	213	2,6	7	0,1	6	0,1	8.318	100
Bilim ve teknoloji	1.014	41,4	475	19,4	526	21,5	396	16,2	1	0,0	30	1,2	9	0,4	2.451	100
Yaşam ve ilgi alanları	761	48,2	419	26,5	274	17,3	78	4,9	8	0,5	19	1,2	21	1,3	1.580	100
Eğitim	372	23,9	542	34,9	517	33,2	94	6,0	6	0,4	10	0,6	14	0,9	1.555	100
İş yaşamı	91	7,9	436	37,9	80	7,0	366	31,9	0	0,0	141	12,3	35	3,0	1.149	100
Toplam	50.969	-	34.681	-	45.608	-	16.021	-	1.154	-	3.253	-	2.789	-	154.475	-

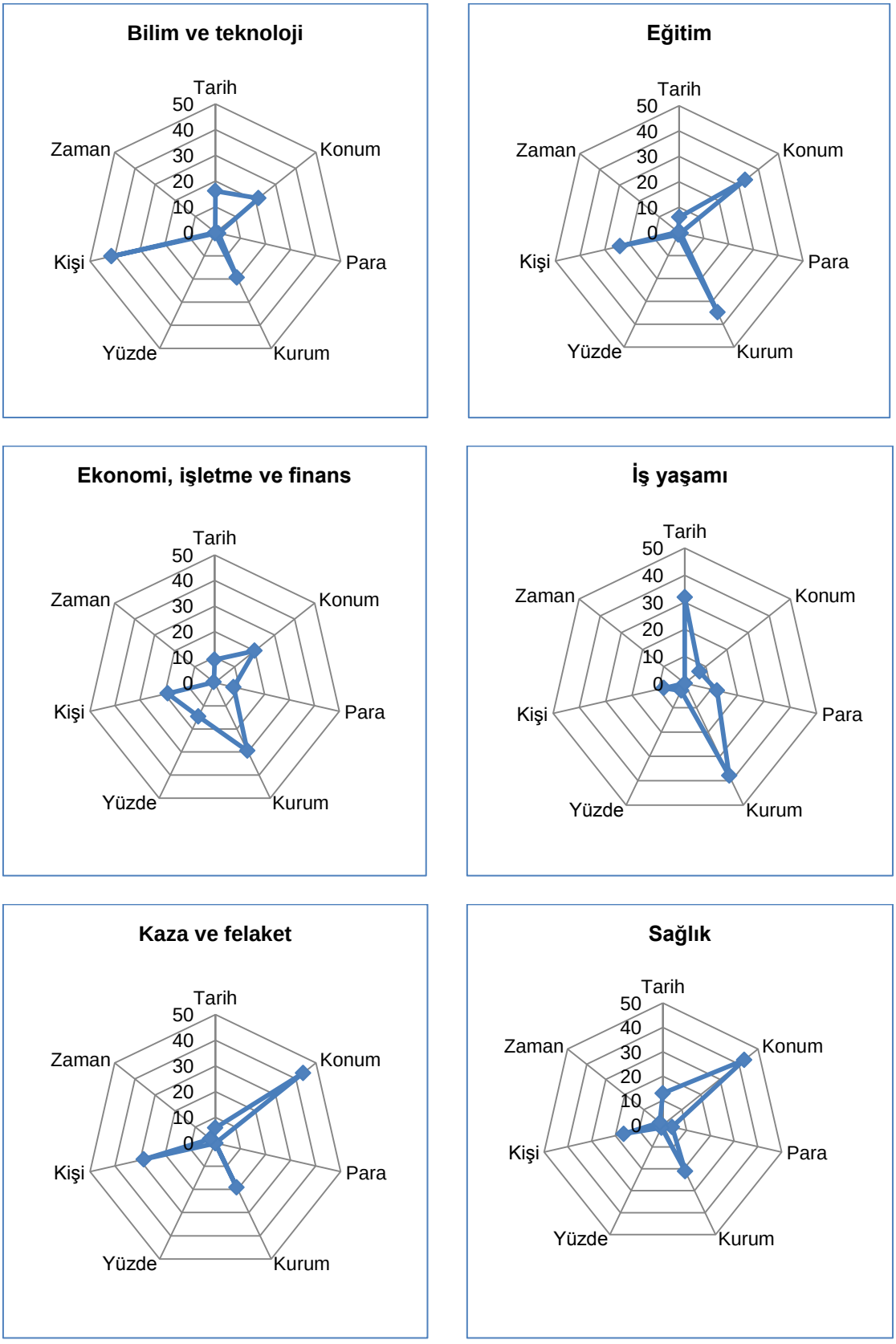
Tablo 5'te de görüldüğü gibi, derlemdeki toplam etiket sayısı 154.475'tir. En fazla etikete sahip konu başlığı 30.606 etiket ile *Suç, hukuk ve yargılama*'dır. Bunu 28.884 etiket ile *Savaş ve karışıklıklar* konu başlığı takip etmektedir. *Bilim ve teknoloji* konu başlığı altında kategorize edilmiş haberler toplam 2451 etiket içermektedir. Derlemde en az etikete sahip konu başlıkları ise *İş yaşamı* (1149 etiket), *Eğitim* (1555 etiket) ile *Yaşam ve ilgi alanları*'dır (1580 etiket).

Çalışmada, derlemdeki haberlerin etiketlenme sayılarının IPTC konu başlıklarına göre farklılık gösterip göstermediği de incelenmiştir. Yapılan Kruskal Wallis testi ile derlemdeki haberlerin etiketlenme sayılarının IPTC konu başlıklarına göre %95 güven düzeyinde istatistiksel açıdan anlamlı bir farklılık gösterdiği bulunmuştur (*H*=351,961, *SD*=12, *p*=0,000). Farklılığın hangi grup ya da gruplardan kaynaklandığını bulmak için Mann-Whitney U testi yapılmıştır. Bonferroni düzeltmesi uygulanarak tüm etkiler için anlamlılık düzeyi 0,004 olarak kabul edilmiştir. Örneğin, *Bilim ve teknoloji* ile *İş yaşamı*; *Bilim ve teknoloji* ile *Kaza ve felaket*; *Bilim ve teknoloji* ile *Sağlık*; *Bilim ve teknoloji* ile *Savaş ve karışıklıklar*; *Bilim ve teknoloji* ile *Sosyal konular*; *Bilim ve teknoloji* ile *Spor*; *Bilim ve teknoloji* ile *Yaşam ve ilgi alanları* konu başlıklarındaki etiket sayılarının birbirinden anlamlı düzeyde farklı olduğu görülmüştür (Sırasıyla *U*=558,500, *p*=0,000, *z*=-4,749, *r*=-0,47; *U*=6042,500, *p*=0,000, *z*=-5,679, *r*=-0,24; *U*=5178,000, *p*=0,000, *z*=-3,499, *r*=-0,18; *U*=15963,000, *p*=0,001, *z*=-3,441, *r*=-0,11; *U*=10748,500, *p*=0,000, *z*=-4,858, *r*=-0,17; *U*=9203,500, *p*=0,000, *z*=-4,643, *r*=-0,17; *U*=1236,000, *p*=0,001, *z*=-3,218, *r*=-0,29). Etki değerleri sadece *Bilim ve teknoloji* ile İş yaşamı konu başlıklarındaki etiket sayıları arasındaki farkın büyük olduğunu, *Bilim ve teknoloji* ile yukarıda sıralanan diğer konu başlıklarında yer alan etiket sayıları arasındaki farkın çok büyük olmadığını göstermektedir. Diğer konu başlıklarına da aynı şekilde Mann-Whitney U testi uygulanmış ve benzer bir durum ile karşılaşılmıştır (bkz. Ek 7).

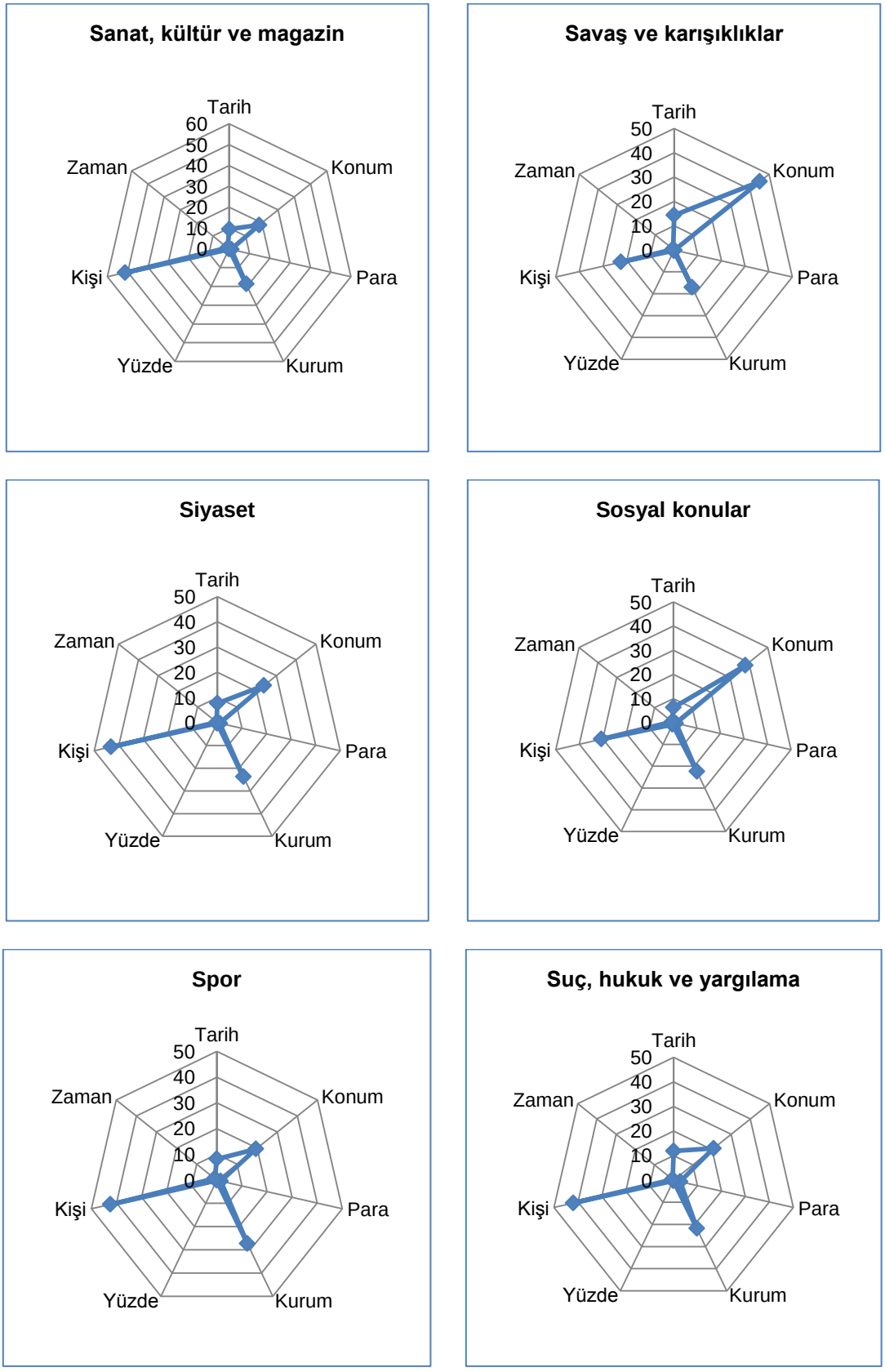
Araştırma kapsamında test edilen unsurlardan bir diğeri de varlık isimlerinin IPTC konu başlıklarına göre dağılımının farklılık gösterip göstermediği olmuştur. Bunu incelemek amacıyla Ki-kare testi uygulanmış ve varlık isimlerinin konu başlıklarına göre dağılımının %95 güven düzeyinde istatistiksel açıdan anlamlı bir farklılık göstermekte olduğu tespit edilmiştir ($\chi^2_{(72)}$= 35191,857; *p*= 0,000).

Araştırmamızın hedeflerinden biri de Tablo 5'teki verilere dayanarak, bir konu başlığı için hangi varlık isminin daha belirleyici olduğunun saptanmasıdır. "Belirleyici etiket" kavramı ile bir konu başlığındaki haberlerin en çok hangi varlık ismi ile etiketlenmiş olduğu yani bu haberlerin tanımlanmasında ve içeriğinin yansıtılmasında en çok hangi

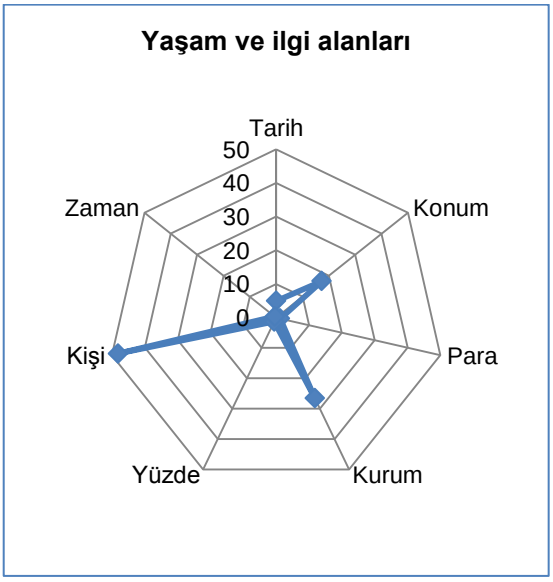
varlık ismine başvurulmuş olduğu bilgisi ifade edilmeye çalışılmıştır. Her bir konu başlığında yer alan varlık isimlerinin dağılımını ve oranlarını daha net görebilmek için varlık isimlerinin toplam değerlerini bir arada gösteren radar grafiklere başvurulmuştur (bkz. Şekil 11).



Şekil 11. Varlık ismi dağılımlarının radar grafik ile gösterimi (%)



Şekil 11. Varlık ismi dağılımlarının radar grafik ile gösterimi (%) (devam)



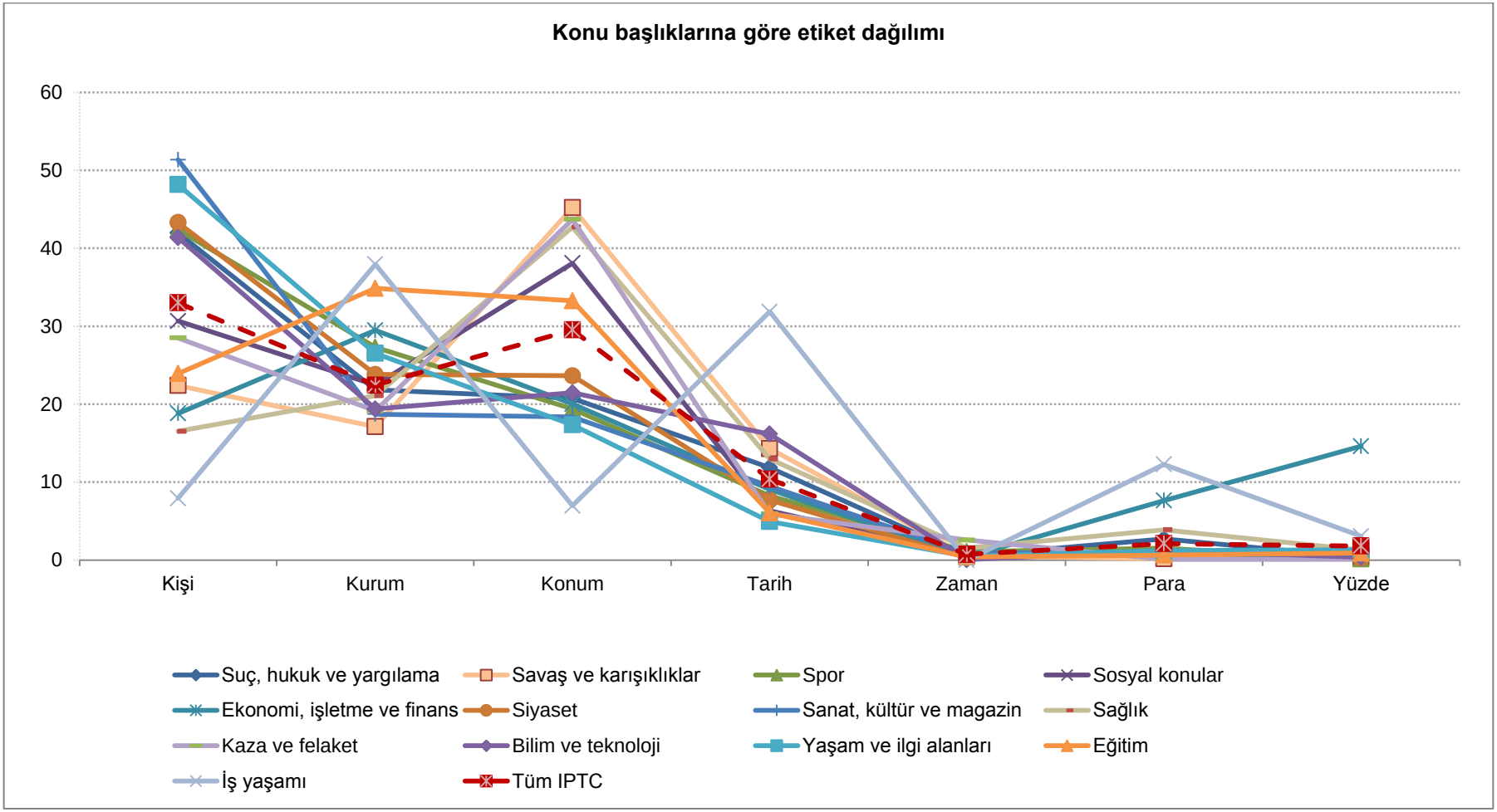
Şekil 11. Varlık ismi dağılımlarının radar grafik ile gösterimi (%) (devam)

Şekil 11'de yer alan grafikler ile konu başlıklarına yönelik varlık isimlerinin dağılımlarını gösteren Tablo 5'e göre, "Kişi" etiketinin en belirleyici olduğu konu başlıklarının *Bilim ve teknoloji* (%41,4; N= 1014), *Sanat, kültür ve magazin* (%51,4; N= 4571), *Siyaset* (%43,3; N= 5711), *Spor* (%42,4; N= 7448), *Suç, hukuk ve yargılama* (%42; N= 12.843) ile *Yaşam ve ilgi alanları* (%48,2; N= 761) olduğu anlaşılmaktadır. Bu oranlara bakıldığında, "Kişi" etiketinin en belirleyici olduğu belirtilen konu başlıklarındaki değeri ortalama olarak %45'tir. "Kişi" etiketinin çalışmamız kapsamında kullanılan IPTC konu başlıklarının yaklaşık yarısında (altı konu başlığı) en belirleyici varlık ismi olması dikkat çekicidir.

Eğitim (%34,9; N= 542), *Ekonomi, işletme ve finans* (%29,5; N= 4459) ile *İş yaşamı* (%37,9; N= 436) konu başlıklarında ise en belirleyici etiket "Kurum" dur. Bu etiketin en belirleyici olduğu belirtilen konu başlıklarındaki değeri ortalama olarak %34'tür. "Konum" etiketinin ise *Kaza ve felaket* (%43,7; N= 3636), *Sağlık* (%42,7; N= 3713), *Savaş ve karışıklıklar* (%45,2; N= 13.058) ile *Sosyal konular* (%38,1; N= 6257) konu başlıklarında baskın olduğu görülmektedir. Bu etiketin en belirleyici olduğu belirtilen konu başlıkları açısından değeri ise ortalama olarak %42'dir. Derlemin geneline bakıldığında "Kişi", "Konum" ve "Kurum" etiketlerinin belirleyici olduğu göze çarpmaktadır. Ayrıca, "Tarih" ve "Para" etiketlerinin *İş yaşamı* konu başlığında ikinci ve üçüncü derecede belirleyici olduğu görülmektedir. "Yüzde" etiketinin en fazla ön plana çıktığı konu başlığı *Ekonomi, işletme ve finans*'tır. Bu etiket diğer konu başlıklarının çoğunda oldukça az kullanılmıştır. Diğer konu başlıkları ile karşılaştırıldığında,

Ekonomi, işletme ve finans konu başlığındaki haberlerin tanımlanmasında farklı varlık isimlerinin birbirine daha yakın oranlarda kullanıldığı dikkat çekmektedir. Öte yandan, "Zaman" etiketinin konu başlıklarının hiçbirinde belirleyici role sahip olmadığı Şekil 11'deki grafiklerden ve Tablo 5'teki verilerden anlaşılmaktadır.

Şekil 11'de sunulan radar grafiklerden de anlaşılacağı gibi konu başlıklarına göre etiket dağılımlarının pek çoğu derlemin tamamına yönelik etiket dağılımı ile (bkz. Şekil 10) büyük ölçüde benzerlik göstermektedir. Konu başlıklarına göre etiket dağılımının genel etiket dağılımından farklılık gösterdiği konu başlığı sayısı sınırlıdır. Derlemin genelindeki etiket dağılımı ile derlemin genelinden en belirgin farklılık gösteren konu başlıklarının etiket dağılımını daha net görebilmek için Şekil 12'de yer alan çizgi grafik hazırlanmıştır.



Şekil 12. Tüm konu başlıklarındaki etiketler ile derlemdeki tüm etiketlerin karşılaştırılması (%)

Şekil 12'de de görüldüğü gibi *İş yaşamı* ile *Ekonomi, işletme ve finans* konu başlıklarındaki etiket dağılımı derlemin genelindeki etiket dağılımından farklı olduğu dikkat çekmektedir. İlgili IPTC konu başlığının haber sayısı en fazla konu başlıklarından bir tanesi olduğu düşünüldüğünde (bkz. Tablo 3) bu farklılığın haber sayılarıyla ilişkili olmayıp, BilCol-2005 derlemindeki haberlerin karakteristiğine ve içeriğine bağlı olduğu anlaşılmaktadır.

4.4. TEKİL ETİKETLER

Etiketleme işlemi esnasında haberlerin yedi farklı varlık ismi ile etiketlendiği daha önce belirtilmişti. Buna göre, haberlerde aynı varlık isimlerinin ne oranda tekrar ettiği çalışma kapsamında merak edilen bir diğer unsur olmuştur. Buradan yola çıkılarak, haberlerde geçen tekil etiket sayıları ve oranları incelenmiştir. Tekil etiket olarak ifade edilmek istenen, bir haberde kaç farklı varlık ismi olduğu bilgisidir.

Örneğin Şekil 13'te yer alan "Faiz yarışına Vakıfbank da katıldı" başlıklı haberde varlık isimlerinin dağılımı şu şekildedir: 6 kurum (2 tane Akbank, 2 tane İş Bankası, 2 tane Vakıfbank), 2 yüzde (2 tane yüzde 1.35) ve bir tarih (Cuma). Bu örnekte tekil etiket sayısı 5 (Akbank, İş bankası, Vakıfbank, 1.35, Cuma), tekil varlık ismi sayısı 3'tür (kurum, yüzde, tarih).

<organization>Akbank</organization> ve <organization>İş Bankası'nın</organization> konut kredisi faizlerini <percentage>yüzde 1.35'e</percentage> indirmelerinin ardından bankalar arasında hızlanan faiz düşürme yarışına <organization>Vakıfbank</organization> da katıldı. <organization>Vakıfbank,</organization> 20 yıla kadar vadede konut kredisi faiz oranlarını <percentage>yüzde 1.35</percentage> olarak belirledi. Piyasa hareketlendi <date>Cuma</date> günü <organization>Akbank</organization> ve <organization> İş Bankası'nın</organization> konut kredi faizlerinde yaptığı indirim kredi pazarını hareketlendirdi....

Şekil 13. Tekil etiket gösterimi

Bu bilgilerden yola çıkılarak Tablo 6'da haberlerde geçen toplam etiket sayıları ile toplam tekil etiket sayıları ve oranları değerlendirilmiştir.

Tablo 6. Konu başlıklarına göre tekil etiket değerleri

IPTC Konu Başlığı	Haberlerde geçen toplam etiket sayısı	Haberlerde geçen toplam tekil etiket sayısı	Haberlerde geçen tekil etiket oranı (%)
İş yaşamı	1.149	776	67,5
Eğitim	1.555	1.037	66,7
Kaza ve felaket	8.318	5.241	63,0
Sosyal konular	16.435	9.872	60,1
Yaşam ve ilgi alanları	1.580	933	59,1
Bilim ve teknoloji	2.451	1.421	58,0
Sağlık	8.692	4.839	55,7
Ekonomi, İşletme ve finans	15.132	8.316	55,0
Sanat, kültür ve magazin	8.900	4.823	54,2
Suç, hukuk ve yargılama	30.606	15.809	51,7
Spor	17.581	8.533	48,5
Savaş ve karışıklıklar	28.884	13.735	47,6
Siyaset	13.192	6.273	47,6
Ortalama	**11.883**	**6.278**	**56,5**[1]

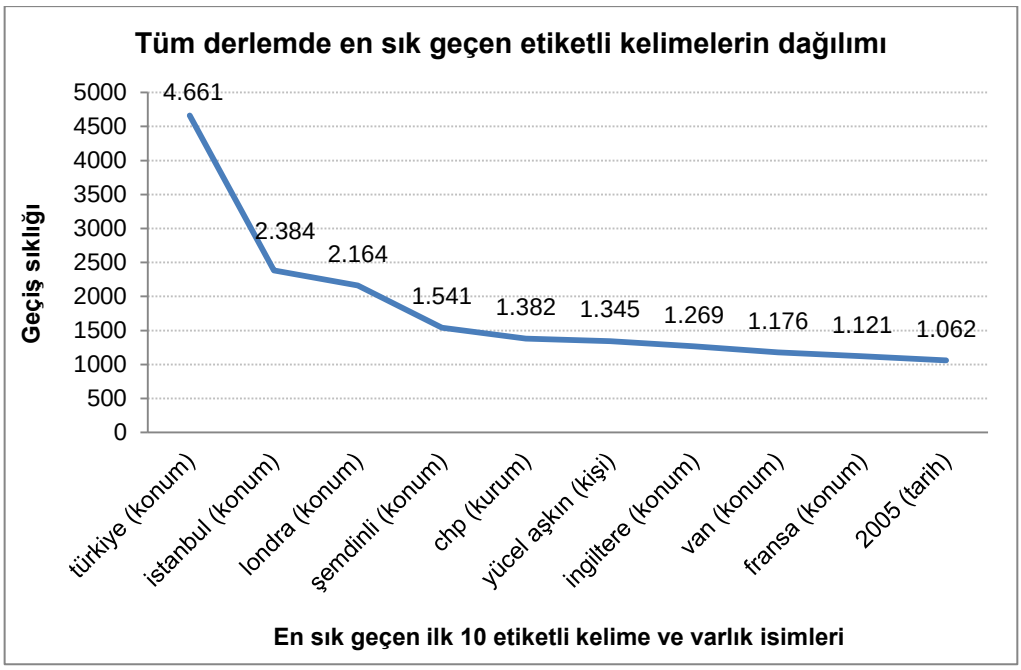
Tablo 6'da sunulan tekil etiket oranlarına bakıldığında, tekil etiket oranı en fazla olan konu başlığı %67,5 ile *İş Yaşamı*'dır. Bu konu başlığını %66,7'lik oranla *Eğitim* konu başlığı takip etmektedir. Tablo 3'te her bir konu başlığına düşen ortalama haber sayısı 449 iken bahsi geçen *İş yaşamı* ve *Eğitim* konu başlıkları 53'er haber içerdiği görülmektedir. Bu durum belirtilen konu başlığındaki haberlerin daha farklı ifadeler kullanılarak oluşturulduğu konusunda ipucu vermektedir.

Tekil etiketlenme oranı en az olan konu başlıkları *Savaş ve karışıklıklar* (%47,6), *Siyaset* (%47,6) ve *Spor* (%48,5)'dur. Konu başlıklarındaki tekil etiket dağılımları incelendiğinde, *Savaş ve karışıklıklar*, *Siyaset* ve *Spor* konularındaki haberlerde geçen kelimelerin %50'den fazlasının iki veya daha fazla kez etiketlendiği görülmektedir. Bu durum, bu konu başlıklarının varlık isimlerine karşılık gelebilecek olan daha az sayıda kelime ile tanımlanabildiğini düşündürmektedir.

[1] Bu oran, haberlerde geçen tekil etiket oranı (%) sütununda yer alan oranların ortalamasıdır.

4.5. EN FAZLA ETİKETLENEN KELİMELER VE KONU BAŞLIKLARINA GÖRE DAĞILIMLARI

Önceki bölümlerde de belirtildiği gibi BilCol-2005 derlemi, beş ayrı Türkçe haber kaynağında geçen 2005 yılına ait gündemi temsil eden haberlerden oluşan bir derlemdir. Bu durumda etiketli ifadelerin haberlerin toplandığı zaman aralığına ait gündemi yansıtmada etkili olup olmadığını incelemek amacıyla derlemde ve konu başlıkları altında en sık etiketlenen kelimeler değerlendirilmiştir. Buna yönelik olarak, Şekil 14'te tüm derlemde en sık geçen etiketlenmiş ilk 10 kelimenin dağılımı ve etiketlendikleri varlık isimleri verilmektedir.



Şekil 14. Tüm derlemde geçen etiketli kelimelerin dağılımı

Şekil 14'e göre, derlemde en sık geçen etiketlenmiş kelime "türkiye"dir (N= 4661). Bunu "istanbul" (N= 2384) kelimesi takip etmektedir. Derlemde sık geçen etiketlenmiş diğer kelimeler ise "londra" (N= 2164), "şemdinli" (N= 1541), "chp" (N= 1382), "yücel aşkın" (N= 1345), "ingiltere" (1269), "van" (N= 1176), "fransa" (N= 1121) ve "2005" (N= 1062)'tir. Varlık isimleri açısından bakıldığında, en sık etiketlenen kelimeler için "Konum", "Kurum", "Kişi" ve "Tarih" etiketlerinin ön planda olduğu göze çarpmaktadır.

Burada ortaya konan kelimelere ait varlık isimlerinin dağılımı ile Tablo 2 ve Şekil 10'da verilmiş olan tüm derlemde geçen etiketlerin dağılımı birbiriyle örtüşmektedir.

Analizlerin gerçekleştirildiği haber derleminin Türkiye'ye ait haber kaynaklarından sağlanmış olması ve çoğunlukla Türkiye ile ilgili haberler içermesinden ötürü "türkiye" kelimesinin derlemde en sık geçen etiketli kelime olması normal olarak görülmektedir. Öte yandan, derlemde geçiş sıklığı açısından onuncu sırada yer alan "Tarih" etiketli "2005" ifadesinin sık geçenler arasında yer almasının nedeni, derlemde yer alan haberlerin tamamının 2005 yılına ait haberlerden toparlanmış olmasından kaynaklanmaktadır.

Derlemde en sık geçen etiketlenmiş kelimelerin sıralaması incelendikten sonra biraz daha detaya inilerek aynı değerlendirme konu başlığı bazında da gerçekleştirilmiştir. Tablo 7'de her bir konu başlığı altında geçen en sık etiketlenmiş ilk 10 kelimenin dağılımı verilmiştir.

Tablo 7. Konu başlıklarına göre en sık geçen ilk 10 etiketli kelime

Bilim ve teknoloji		Eğitim		Ekonomi, işletme ve finans		İş yaşamı		Kaza ve felaket		Sağlık		Sanat, kültür ve magazin	
Kelime	**N**	**Kelime**	**N**	**Kelime**	**N**	**Kelime**	**N**	**Kelime**	**N**	**Kelime**	**N**	**Kelime**	**N**
nobel	172	hüseyin çelik	129	onur air	813	tbmm	60	atina	250	türkiye	693	deniz baykal	387
erdoğan	142	yök	104	türkiye	732	ocak	59	istanbul	200	ab	230	bülent ersoy	346
türkiye	108	tbmm	61	hollanda	613	1991	58	yunanistan	159	manyas	225	ata türk	278
bill gates	100	türkiye	46	şener	522	15	56	zeytinburnu	143	romanya	214	attila ilhan	264
abd	97	uşak	34	almanya	324	2005	40	kıbrıs	127	ekim	213	türkiye	239
2005	59	tekirdağ	33	2005	166	emekli sandığı	27	ankara	102	2005	145	istanbul	202
microsoft	50	burdur	31	cansızlar	156	plan ve bütçe komisyonu	26	gaziantep	101	kızıksa	140	chp	180
hwang	42	çorum	31	yıldırım	151	genel kurul	23	ırak	93	balıkesir	133	koç	141
ingiltere	26	kırşehir	29	ziraat bankası	144	salı	23	güler	85	sağlık bakanlığı	103	melih kibar	120
marshall	25	ordu	29	istanbul	136	şubat	22	maslak	81	tarım ve köyişleri bakanlığı	100	devlet tiyatroları	115

Savaş ve karışıklıklar		Siyaset		Sosyal konular		Spor		Suç, hukuk ve yargılama		Yaşam ve ilgi alanları	
Kelime	**N**	**Kelime**	**N**	**Kelime**	**N**	**Kelime**	**N**	**Kelime**	**N**	**Kelime**	**N**
londra	2.075	kemal derviş	889	malatya	807	türkiye	746	yücel aşkın	2.203	özer gürbüz	200
şemdinli	1.526	mustafa sarıgül	791	fransa	659	fenerbahçe	652	saddam hüseyin	1.124	sibel deniz	102
ingiltere	941	chp	648	istanbul	606	izmir	358	yahya murat demirel	1.051	hüseyin deniz	84
hakkari	751	türkiye	581	paris	506	ispanya	227	van yüzüncü yıl üniversitesi	876	izmir	63
11	555	deniz baykal	417	türkiye	326	2005	207	istanbul	736	central hospital	47
mısır	454	yunanistan	305	çocuk yuvası	287	galatasaray	186	türkiye	653	akp	38
yüksekova	448	fahd	296	çubukçu	193	unıversıade	172	yök	478	sağlık bakanlığı	34
türkiye	441	ydk	271	yorgancıoğlu	148	bulgaristan	170	ırak	436	türkiye	28
abd	410	ırak	186	aa	143	litvanya	169	egebank	354	sarıkamış	22
2003	384	nermin erbakan	186	tekel	136	daum	164	bulgaristan	263	aynur arslan	18

Tablo 7'ye göre, derlemde en sık geçen etiketli kelime olan "türkiye", *Sağlık* (N= 693) ve *Spor* (N= 746) konu başlıklarında da en sık geçen etiketli kelime olmanın yanı sıra *Ekonomi, işletme ve finans* (N= 732) konu başlığında ikinci, *Bilim ve teknoloji* (N= 108) konusunda üçüncü, *Siyaset* (N= 581) konu başlığında dördüncü, *Sanat, kültür ve magazin* (N= 239) konu başlığında ise en sık geçen etiketli beşinci kelime durumundadır. Şekil 14'te "türkiye" kelimesi için yaptığımız değerlendirme bu kısım için de geçerlidir. Bu kelimenin konu başlıklarının pek çoğunda sık etiketlenen kelime olma nedeni haberlerin Türkiye'ye ait haberlerden sağlanmış olmasından kaynaklanmıştır.

Derlemde en sık geçen ikinci kelime olan "istanbul" ise *Kaza ve felaket* (N= 200) konu başlığında ikinci, *Sosyal konular* (N= 606) konu başlığında üçüncü, *Suç, hukuk ve yargılama* (N= 736) konu başlıklarında beşinci, *Sanat, kültür ve magazin* (N= 202) konu başlığında altıncı, *Ekonomi, işletme ve finans* (N= 136) konu başlığında ise onuncu sırada en sık geçen kelimedir.

Tablo 7'deki her bir konu başlığında en sık geçen etiketli kelimeler daha detaylı incelendiğinde, yukarıda bahsi geçen kelimelerin dışında ilk 10 sıradaki diğer kelimelerin de haberlerin ait oldukları zaman aralığında yaşanan olaylar hakkında fikir verdiği anlaşılmaktadır. Örneğin, *Savaş ve karışıklıklar* konu başlığında en sık geçen etiketli kelimelerden "londra" sadece bu konu başlığında ön sıralarda yer almaktadır. Bunun nedeni, haberin sağlandığı dönemde Londra metrosunda bir patlama meydana gelmiş olması ve yaşanan patlama ile ilgili haberlerin söz konusu dönemde yoğunlukta olmasıdır. Bununla ilişkili olarak "ingiltere" kelimesi de *Savaş ve karışıklıklar* konu başlığında üçüncü sırada yer almaktadır. Benzer şekilde, sadece *Ekonomi, işletme ve finans* konu başlığında geçen ve bu konu başlığında ilk sırada yer alan "onur air" kelimesinin sık geçmesinin nedeni, o dönemde Onur Air havayolu firmasının Avrupa'nın bazı ülkelerinde iniş kalkışının yasaklanması sonucu bununla ilgili ortaya çıkan haber sayısının artmasıdır. Bir diğer belirgin örnek ise *Suç, hukuk ve yargılama* konu başlığında yer almaktadır. Bu konu başlığında sık geçen kelimelere bakıldığında "yücel aşkın", ve "van yüzüncü yıl üniversitesi" ifadelerinin birbiriyle ilişkili olduğu görülmektedir. O dönemde Van Yüzüncü Yıl Üniversitesi Rektörü Yücel Aşkın'ın tutuklanması ile ilgili haberlerin gündeme gelmesi *Suç, hukuk ve yargılama* konu başlığında buna ilişkin kelimelerin ön planda olmasını sağlamaktadır.

Öte yandan, her bir konu başlığında en sık geçen etiketli ilk kelimelerin IPTC konu başlığındaki geçiş sıklıkları ile derlemdeki geçiş sıklıkları incelenerek, bu kelimelerin ait oldukları konu başlığı için ne oranda ayırt edici oldukları da belirlenmeye çalışılmıştır.

Ayırt edicilik kavramı ile ifade edilmek istenen, bir kelimenin ilgili konu başlığında geçme sıklığı (terim sıklığı) yüksek ve derlemdeki diğer konu başlıklarında geçme sıklığı düşükse, o kelimenin ayırt edici özelliğinin yüksek olduğudur (Tonta, Bitirim ve Sever, 2002, s. 17). Bir başka deyişle, bir konu başlığında sık geçen kelime derlemdeki diğer konu başlıklarında da sık geçiyorsa o kelimenin ilgili konu başlığı için ayırt ediciliği düşüktür. Buradan yola çıkılarak kelimelerin IPTC konu başlıklarında geçiş sıklıkları derlemde geçiş sıklıklarına bölünerek her bir kelimenin ilgili konu başlığı için ağırlığı belirlenmiştir. Kelimelerin ağırlığı 0 ile 1 arasında yer almaktadır. Kelimenin ağırlığı 1'e yaklaştıkça o kelimenin ilgili konu başlığı için ayırt ediciliği artarken, ağırlık 0'a yaklaştıkça kelimenin ilgili konu başlığı için ayırt ediciliği azalmaktadır.

Tablo 8. Her konu başlığı için en sık geçen etiketli kelimelerin sıklıkları ve sıralamaları

IPTC Konu Başlığı	IPTC konu başlığında en sık geçen kelime	Varlık ismi	IPTC konu başlığında geçiş sıklığı	Derlemde geçiş sıklığı	*D* *	Derlemdeki geçiş sıklığı sıralaması
Ekonomi, işletme ve finans	"onur air"	kurum	813	813	1,00	17
Bilim ve teknoloji	"nobel"	kurum	172	172	1,00	130
Siyaset	"kemal derviş"	kişi	889	893	0,99	5
Suç, hukuk ve yargılama	"yücel aşkın"	kişi	2.203	2.238	0,98	6
Sosyal Konular	"malatya"	konum	807	839	0,96	16
Savaş ve karışıklıklar	"londra"	konum	2.075	2.164	0,96	3
Yaşam ve ilgi alanları	"özer gürbüz"	kişi	200	245	0,82	154
Kaza ve felaket	"atina"	konum	250	379	0,66	52
Sanat, kültür ve magazin	"deniz baykal"	kişi	387	747	0,52	20
Eğitim	"hüseyin çelik"	kişi	129	257	0,50	32
Spor	"türkiye"	konum	746	4.659	0,16	1
Sağlık	"türkiye"	konum	693	4.659	0,15	1
İş yaşamı	"tbmm"	kurum	60	427	0,14	102

* *D*= IPTC konu başlığında geçiş sıklığı/ Derlemde geçiş sıklığı

Tablo 8'de her konu başlığı için en sık etiketlenen kelimelerin ilgili konu başlığında ve derlemde geçiş sıklıkları ile sıralaması verilmiş ve veriler kelimelerin IPTC konu başlıklarında geçiş sıklıklarının derlemde geçiş sıklıklarına oranına göre sıralanmıştır.

Her bir konu başlığında en fazla etiketlenmiş kelimelerin o konu başlığı için ayırt edici olup olmadığına bakıldığında *Ekonomi, işletme ve finans* (*D*= 1,00), *Bilim ve teknoloji* (*D*= 1,00), *Siyaset* (*D*= 0,99), *Suç, hukuk ve yargılama* (*D*= 0,98), *Sosyal Konular* ile (*D*= 0,96) *Savaş ve karışıklıklar* (*D*= 0,96) konu başlıklarının derlemin genelinden daha farklı (daha kendine has) kelimelerle etiketlenme eğiliminde olduğu görülmektedir.

Ekonomi, işletme ve finans konu başlığında en sık geçen kelime olan “onur air” (N= 813) sadece bu konu başlığı altındaki haberlerde yer almakta, diğer konu başlıklarındaki haberlerin hiçbirinde geçmemektedir. Bu durum “onur air” kelimesinin derlemin oluşturulduğu dönem göz önünde bulundurulduğunda *Ekonomi, işletme ve finans* konu başlığı için ayırt edici özelliğinin olduğunu göstermektedir. Benzer şekilde *Bilim ve teknoloji* konu başlığındaki haberlerde en sık geçen “nobel” (N= 172) kelimesi derlemin geneli göz önüne alındığında sadece bu konu başlığında yer alan haberlerde geçmektedir. Dolayısıyla bu kelime de derlem kapsamında değerlendirildiğinde *Bilim ve teknoloji* konu başlığı için ayırt edici özelliğe sahiptir.

Derlemin geneline bakıldığında, yaklaşık her yedi “tbmm” kelimesinden birinin *İş yaşamı* konu başlığına ait haberlerden geldiği söylenebilir. Dolayısıyla, derlemin genelinde farklı başlıklarda da yer aldığı için, “Kurum” etiketiyle etiketlenmiş “tbmm” kelimesinin bu konu başlığı için ayırt ediciliği düşüktür (*D*= 0,14).

“türkiye” kelimesi, hem *Spor* (N= 746) hem *Sağlık* (N= 693) konu başlığında en sık geçen kelimedir. Bununla birlikte söz konusu kelimenin derlemin genelinde de (N= 4659) en sık geçen kelime olduğu daha önce vurgulanmıştı. Bu durum “türkiye” kelimesinin ne bu konu başlıkları için ne de derlemin geneli için ayırt edici özelliğe sahip olmadığını ortaya koymaktadır (Spor *D*= 0,16; Sağlık *D*= 0,15).

Etiketli kelimelere ilişkin incelenen bir diğer unsur, haberlerde geçen ortalama kelime sayıları ile ortalama etiket sayıları olmuştur (bkz. Tablo 9). Bu incelemenin neticesinde, haber metinlerinin uzun ya da kısa oluşunun etiketleme durumunu nasıl etkilediğinin ortaya konması amaçlanmıştır.

Tablo 9. Konu başlıklarına göre ortalama etiketleme değerleri

IPTC Konu Başlığı	Haberde geçen ortalama kelime sayısı	Min.	Maks.	Std. Sapma	Haberlerde geçen ortalama etiket sayısı	Min.	Maks.	Std. Sapma
Bilim ve teknoloji	342	39	1.879	404	52	4	448	72
Eğitim	254	30	576	155	29	3	77	15
Ekonomi, işletme ve finans	277	4	4.386	287	27	1	515	30
İş yaşamı	225	91	433	132	22	9	21	5
Kaza ve felaket	162	14	1.239	144	16	1	182	15
Sağlık	257	6	4.547	339	27	1	776	49
Sanat, kültür ve magazin	258	21	1.834	242	31	3	292	31
Savaş ve karışıklıklar	223	5	3.903	265	30	1	2.227	83
Siyaset	240	16	3.950	280	32	1	497	35
Sosyal konular	230	15	4.361	325	21	1	272	25
Spor	173	16	2.146	199	27	1	3.480	138
Suç, hukuk ve yargılama	248	4	1.809	260	28	1	451	27
Yaşam ve ilgi alanları	254	33	672	151	20	5	60	12
Ortalama	**242**	**23**	**2.441**	**245**	**28**	**2**	**715**	**41**

Tablo 9'a göre, *Spor* konu başlığındaki haberler en kısa metne sahip niteliktedir. Öte yandan, *Spor* konu başlığının, haberlerde geçen ortalama etiket sayısı açısından bakıldığında diğer konu başlıklarına kıyasla en yüksek değere (N= 3480) ulaştığı görülmektedir. Bu durum, *Spor* konu başlığındaki haberlerin kısa metne sahip olmasına rağmen bu metinlerde yer alan kelimelerin büyük çoğunluğunun varlık isimleri ile etiketlenmeye uygun olabileceğini göstermektedir.

Bilim ve teknoloji konu başlığı hem haberde geçen ortalama kelime sayısı açısından (N= 342) hem de haberde geçen ortalama etiket sayısı (N= 52) açısından ilk sıradadır. Haberde geçen ortalama kelime sayısı açısından ikinci sırada *Ekonomi, işletme ve finans* (N= 277) yer almaktadır. Bu konu başlıklarındaki haberlerin aktarımında varlık isimleri tarafından nitelenebilen daha fazla sayıda kelime kullanıldığı söylenebilir.

Kaza ve felaket konu başlığı hem haberde geçen ortalama kelime sayısı (N= 162) açısından hem de haberde geçen ortalama etiket sayısı (N= 16) açısından en düşük sıklıklara sahiptir. Buna dayanarak, *Kaza ve felaket* konu başlığında yer alan haberlerin varlık isimlerine karşılık gelecek daha az sayıda kelime ile ifade edildiği söylenebilir.

4.6. ETİKETSİZ KELİMELER VE KONU BAŞLIKLARINA GÖRE DAĞILIMLARI

Çalışma kapsamında analiz edilen bir başka unsur etiketsiz kelimelerdir. Etiketsiz kelimeler, bir haber metninde etiketli kelimeler çıkarıldıktan sonra kalan kelimelerdir ve bu kelimelerin haberlerde "NE" sorusuna yanıt verdiği kabul edilmektedir. Bu doğrultuda derlemdeki etiketsiz kelimeler incelenerek konu başlıklarını tanımlamada etkili olup olmadıkları değerlendirilmeye çalışılmıştır. Derlemin etiketsiz kısmında yer alan "ve, ile, değil, de, da, var, yok, için, dolayı" gibi durma listesi olarak kabul edilebilecek ifadeler ayıklanmıştır.

Çalışmanın veri setini oluşturan haberlerde toplam 847.199 etiketsiz kelime bulunmaktadır. Bu kelimelerin konu başlıklarına göre dağılımına ait oranlar Tablo 10'da sunulmaktadır.

Tablo 10. Etiketsiz kelimelerin dağılımı

IPTC konu başlıkları	Toplam kelime sayısı	Etiketli kelime sayısı	"NE" sorusuna karşılık gelen etiketsiz kelime dağılımı	
			N	%
Sosyal konular	152.290	16.435	135.855	89,2
Yaşam ve ilgi alanları	14.498	1.580	12.918	89,1
İş yaşamı	8.324	1.149	7.175	86,2
Kaza ve felaket	60.099	8.318	51.781	86,2
Ekonomi, işletme ve finans	108.623	15.132	93.491	86,1
Eğitim	10.513	1.555	8.958	85,2
Sanat, kültür ve magazin	59.693	8.900	50.793	85,1
Sağlık	57.706	8.692	49.014	84,9
Suç, hukuk ve yargılama	201.742	30.606	171.136	84,8
Siyaset	78.671	13.192	65.479	83,2
Savaş ve karışıklıklar	153.408	28.884	124.524	81,2
Bilim ve teknoloji	12.321	2.451	9.870	80,1
Spor	83.786	17.581	66.205	79,0
Toplam	**1.001.674**	**154.475**	**847.199**	**84,6**

Tablo 10'daki veriler "NE" sorusuna karşılık gelen etiketsiz kelime oranlarına göre sıralanmıştır. Derlemde en fazla etiketsiz kelime içeren konu başlığı *Sosyal konular* (%89,2)'dır. Etiketsiz kelime oranı en düşük konu başlığı ise *Spor* (%79)'dur.

Etiketleme oranının haber metinlerinin uzunluğuna bağlı olmadığı Tablo 10'da yer alan örneğin *İş yaşamı* ile *Kaza ve felaket* konu başlıklarına ait oranlardan (her ikisi için de

etiketlenmemiş kelimelerin haberlerde geçen kelimelere oranı %86,2) anlaşılmaktadır. *Kaza ve felaket* konu başlığındaki haberlerde yer alan kelime sayısı *İş yaşamı* konu başlığı altındakilerden yaklaşık yedi kat fazla olmasına rağmen etiketleme oranının birbiri ile aynı olduğu görülmektedir. Tablonun geneli için bakıldığında tüm konu başlıkları için de benzer bir durumun söz konusu olduğu, toplam kelime oranları birbirlerinden çok farklı olmasına rağmen bunların etiketlenme oranlarının birbirlerine yakın olduğu anlaşılmaktadır. Konu başlıkları altında yer alan haberlerdeki etiketsiz kelime oranlarının birbirine yakın oluşu, haberler metinlerinin kısa ya da uzun olmasının haberlerin daha fazla varlık ismi ile tanımlanmasına bir etkisi olmadığını düşündürmektedir.

Tablo 11'de her bir konu başlığındaki etiketsiz kelimelerin %33'lük ve %50'lik dağılımlarına bakılarak konu başlıklarına göre, etiketlenmemiş kelimelerin ilgili konu başlığındaki yoğunluğu ortaya konmaya çalışılmıştır. Bu tablodaki yüzdelik dilimler hesaplanırken farklı/tekil etiketsiz kelimelerin sayısı dikkate alınmıştır. Bu kelimelerin geçiş sıklıkları toplandığında yukarıdaki tabloda verilmiş olan toplam kelime sayılarına ulaşılmaktadır.

Tablo 11. Konu başlıklarına göre etiketsiz metinde geçen kelimelerin oranları

IPTC Konu Başlıkları	%33		%50		%100
	N	%	N	%	N
Suç, hukuk ve yargılama	229	1,2	615	3,3	18.452
Sosyal konular	213	1,3	542	3,2	16.753
Savaş ve karışıklıklar	168	1,1	501	3,3	15.008
Ekonomi, işletme ve finans	124	1,0	404	3,3	12.120
Sanat, kültür ve magazin	355	3,1	924	8,2	11.320
Siyaset	144	1,4	428	4,1	10.459
Spor	116	1,2	343	3,7	9.336
Sağlık	99	1,2	354	4,2	8.435
Kaza ve felaket	122	1,6	358	4,6	7.803
Yaşam ve ilgi alanları	107	3,0	289	8,1	3.554
Bilim ve teknoloji	156	4,8	419	12,8	3.261
Eğitim	39	1,8	164	7,5	2.178
İş yaşamı	36	2,2	114	6,9	1.648

Tablo 11'deki verilere göre, derlemde bulunan etiketsiz metinde 18.452 kelime ile en fazla kelime içeren konu başlığı *Suç, hukuk ve yargılama*'dır. Bu konu başlığındaki haberlerin üçte biri 229 (konu başlığı içindeki oranı %1,24) farklı etiketsiz kelime ile yarısı ise 615 (konu başlığı içindeki oranı %3,33) birbirinden farklı (tekil) etiketsiz kelime ile tanımlanmaktadır. Benzer şekilde *Sosyal konular* (N= 16.753), *Savaş ve*

karışıklıklar (N= 15.008) ile *Ekonomi, işletme finans* (N= 12.120) konu başlıklarında yer alan haberlerin de yarısı toplam tekil etiketsiz kelime sayısının yaklaşık %3,3'lük kısmı ile temsil edilmektedir. Tablodan da görüldüğü gibi az sayıdaki etiketlenmemiş kelimenin konu başlığı içindeki ağırlığı geri kalan çoğunluğa oranla daha fazladır. Bu durum, konu başlıkları altında yer alan haberlerin çoğunlukla belirli bir grup etiketsiz kelime ile ifade edildiğini ortaya koymaktadır. Bir başka deyişle, haberlerin tanımlanmasında haberlerde geçen etiketsiz kelimelerin tamamına bakmak yerine, bu kelimelerin yoğunluk oluşturduğu yüzdelik dilimdekileri dikkate almak daha anlamlı görünmektedir.

Etiketsiz kelimelerin konu başlıklarına göre dağılımlarına daha yakından bakmak ve etiketli ifadelerle ne kadar örtüştüklerini ortaya koymak amacıyla her bir konu başlığında en sık geçen ilk 10 kelime Tablo 12'de sunulmuştur.

Tablo 12. Konu başlıklarına göre etiketsiz metinde en sık geçen ilk 10 kelime

Bilim ve Teknoloji		Eğitim		Ekonomi,işletme ve finans		İş yaşamı		Kaza ve felaket		Sağlık		Sanat, kültür ve magazin	
Kelime	**N**	**Kelime**	**N**	**Kelime**	**N**	**Kelime**	**N**	**Kelime**	**N**	**Kelime**	**N**	**Kelime**	**N**
ödül	276	üniversite	511	kredi	3.238	derece	258	uçak	1.317	kuş	1.859	sanatçı	385
yıl	94	yeni	172	konut	2.418	memur	220	kişi	867	grip	1.540	bakan	302
bakan	79	tasarı	157	faiz	1.931	aylık	197	kaza	683	hayvan	892	türk	300
adam	68	bilim	155	yıl	1.305	iş	126	patlama	665	hastalık	625	yasal	296
bilim	61	yüksekokul	139	banka	1.155	tasarı	123	yolcu	538	tavuk	435	iş	253
bilgi	57	fakülte	114	uçuş	546	yasa	106	iş	346	insan	402	büyük	220
konu	52	bakan	99	aylık	340	yetim	86	rum	341	kanatlı	367	çalışan	211
bilgisayar	44	enstitü	73	vadeli	330	bağ-kur	75	açıklama	333	vaka	346	kültür	203
büyük	40	meslek	72	yasak	301	maaş	75	yasal	256	kümes	344	müdür	189
vatandaş	38	milli	64	artış	298	emekli	61	lpg	236	bebek	267	sahne	132

Savaş ve karışıklıklar		Siyaset		Sosyal konular		Spor		Suç, hukuk ve yargılama		Yaşam ve ilgi alanları	
Kelime	**N**	**Kelime**	**N**	**Kelime**	**N**	**Kelime**	**N**	**Kelime**	**N**	**Kelime**	**N**
saldırı	2.593	başkan	541	rakı	2.173	maç	1.016	rektör	1.967	bebek	374
terör	1.119	bakan	499	sahte	1.642	takım	952	mahkeme	1.260	anne	168
bomba	885	parti	492	kişi	1.199	oyun	909	dr	1.169	gebelik	151
bakan	860	ülke	469	polis	901	madalya	684	yıl	1.168	doğum	143
yasal	813	başkanlık	462	alkol	574	sporcu	542	prof	1.105	yediz	141
metro	793	kurul	456	şiddet	420	milli	534	ceza	934	yas	124
güvenlik	640	kral	344	gözaltı	415	spor	457	ifade	775	sağ	122
ingiliz	558	kalkınma	338	bakan	355	yarış	402	hakkında	732	tedavi	69
olay	379	türk	333	olay	352	konu	356	soru	709	soruşturma	68
gözaltı	366	dünya	284	metil	342	kilo	335	hapis	679	ölüm	66

Tablo 12'deki verilere bakıldığında etiketsiz metinde en sık geçen ilk 10 kelimenin ilgili konu başlığı hakkında büyük ölçüde ipucu verdiği görülmektedir. Örneğin, *Bilim ve teknoloji* konu başlığında "ödül" kelimesi en sık (N= 276) geçen kelimedir. Bu durum sözü edilen konu başlığında en sık geçen etiketli kelime olan "nobel" ile ilişkilidir (bkz. Tablo 7).

Eğitim konu başlığında ise "üniversite" kelimesi en sık (N= 511) geçen kelimedir. Bunu takiben, "tasarı", "bilim", "yüksekokul", "fakülte" gibi kelimelerin sık geçen kelimeler olması bu kelimelerin, etiketlenmemiş olsalar dahi, ilgili haberlerin Eğitim konusunda olduğu hakkında fikir verdiği görülmektedir.

Sağlık konu başlığında en sık geçen kelimelere bakıldığında, "kuş", "grip", "hayvan", "hastalık", "tavuk" gibi kelimelerin ilk sıralarda yer aldığı görülmektedir. Bu durumun büyük ölçüde haberlerin sağlandığı zaman diliminin Türkiye'de kuş gribi salgınının patlak verdiği döneme rastlamasıyla ilişkili olduğu düşünülmüştür. Benzer durum, *Savaş ve karışıklıklar* konu başlığında en sık geçen etiketsiz kelimelerin, tıpkı etiketli kelimelerde olduğu gibi, o dönemde Londra metrosunda yaşanan patlamayla ilişkili oluşu, *Sosyal konular* konu başlığında sık geçen etiketsiz kelimelerin ise o dönemde sahte rakı operasyonlarının gündeme gelmesiyle paralellik göstermesi ile belirgin şekilde ortaya çıkmaktadır.

4.7. TARTIŞMA VE YORUM

Çalışmanın temel araştırma sorularını yanıtlamak için kullanılan derlem 5834 haberden oluşmaktadır. Veri setini oluşturan haberler "Kişi", "Konum", "Kurum", "Tarih", "Para", "Yüzde" ve "Zaman" olmak üzere yedi farklı varlık ismi ile etiketlenmiştir. Haberlerin tamamında toplam 154.475 kelime varlık isimleri ile tanımlanmış, bir başka deyişle etiketlenmiştir. Ayrıca, haberlerde toplam 847.199 etiketsiz kelime bulunmaktadır.

Çalışma kapsamındaki temel araştırma sorularına yönelik elde edilen yanıtlar şunlar olmuştur:

Derlemdeki haberlerin etiketlenmesinde en baskın ve en pasif varlık isimlerinin belirlenmesine yönelik analizler, derlemde en fazla "Kişi", en az ise "Zaman" etiketi kullanıldığına yönelik tahminimizi doğrulamıştır. Gazetecilik açısından bakıldığında, haber yazımındaki 5N1K (Ne, Nerede, Ne zaman, Nasıl, Neden, Kim) kuralı çerçevesinde özellikle "Kim" ve "Ne" sorularının önemli olduğu, haber okurları

açısından ilk dikkat çeken unsurların “Kim ne yaptı?”, “Ne oldu?” gibi soruların yanıtı olduğu, bu nedenle de bir haberde önceliğin “Kim” ve “Ne” sorularını yanıtlayacak ifadelere verilmesi gerektiği vurgulanmaktadır (MEGEP, 2007). Bu çerçevede değerlendirildiğinde, derlemimizdeki varlık isimlerinde de bu durumu destekler bir dağılım görülmüştür.

Bulgular, etiketlenme oranının haber metinlerinin içerdiği toplam kelime sayısına bağlı olmadığını göstermiştir (bkz. Tablo 10). Bir konu başlığındaki haberlerin içeriğini yansıtmada en çok hangi varlık isimlerinin ön plana çıktığının anlaşılabilmesi için, her bir konu başlığında yer alan etiketlerin dağılımı ve oranları incelenmiş, *Bilim ve teknoloji*, *Sanat, kültür ve magazin*, *Siyaset*, *Spor*, *Suç, hukuk ve yargılama* ile *Yaşam ve ilgi alanları* başlıkları için en belirleyici etiketin “Kişi” olduğu anlaşılmıştır. Bu durum, sözü edilen konu başlıklarında kişilere yönelik haberlerin daha fazla yer aldığı yönünde ipucu vermektedir. Bu konu başlıklarında “Kişi” isimlerinin ön plana çıkmasının nedeninin büyük ölçüde haber içeriğiyle ilgili olduğu tahmin edilmektedir. Nitekim, *Sanat, kültür ve magazin*, *Siyaset*, *Spor* gibi konularda özellikle sanatçılar, siyasetçiler, sporcular ile onların çalışmaları, polemikleri, söylemleri gibi unsurların ön plana çıkmasının bu sonucu doğurmuş olabileceği düşünülmektedir.

Eğitim, *Ekonomi, işletme ve finans* ile *İş yaşamı* konu başlıklarında en belirleyici etiketin “Kurum” olduğu görülmüştür. Bunun nedeninin bu konu başlıkları altındaki haberlerde çoğunlukla üniversiteler, okullar, şirketler ve diğer kurumların geçmesinden kaynaklandığı görülmektedir. “Konum” etiketinin ise *Kaza ve felaket*, *Sağlık*, *Savaş ve karışıklıklar* ile *Sosyal konular* konu başlıklarında baskın olduğu belirlenmiştir. Bu da bahsi geçen konulardaki haberlerde yer bilgisinin önemini göstermektedir. Örneğin, bir kaza haberine ilişkin ilk merak edilen şey, kazanın nerede olduğu bilgisidir. Benzer şekilde, bir savaş haberinde ilk olarak savaşın nerede çıktığı, buna ilişkin olayların nerede olduğu dikkate alınır. Bu açıdan bakıldığında, “Konum” etiketinin neden bu konu başlıklarında belirleyici olduğu anlaşılmaktadır. Konu başlıklarına ayırmaksızın derlemin genelindeki etiket dağılımına bakıldığında “Kişi”, “Konum” ve “Kurum” etiketlerinin belirleyici role sahip olduğu belirlenmiştir. Konu başlıklarına göre etiket dağılımında ise *İş yaşamı* konu başlığı dışında kalan tüm konu başlıkları için “Kişi”, “Konum” ve “Kurum” etiketlerinin yine ön planda olduğu görülmektedir. Bu durum en baskın varlık isimlerinin tüm konu başlıkları için “Kişi”, “Konum” ve “Kurum” olduğuna yönelik hipotezimizi büyük ölçüde desteklemektedir. Derlemin genelinde gözlemlenen etiketler ile karşılaştırıldığında etiket dağılımları açısından farklılık göze çarpan konu

başlıkları *İş yaşamı* ile *Ekonomi, işletme finans*'tır. Farklılık oluşmasının nedeni, *İş yaşamı* konu başlığında "Kurum", "Tarih" ve "Para", *Ekonomi, işletme ve finans* konu başlığında ise "Kurum", "Para" ve "Yüzde" etiketlerinin diğer konu başlıklarına göre daha sık uygulanmış olmasıdır. *İş yaşamı* ile *Ekonomi, işletme ve finans* konularında yer alan haberlerin doğası gereği kurum, para ve yüzdelik ifade içermeleri sonucu bahsi geçen etiketlerin bu konu başlıklarında biraz daha fazla ön plana çıktığı düşünülmektedir.

Haberlerin ait oldukları konu başlığı için belirleyiciliği sağlayan unsurlardan bir diğerinin de etiketli kelimeler olduğu saptanmıştır. Örneğin, *Ekonomi, işletme ve finans* konu başlığında "onur air" sadece bu konu başlığı altındaki haberlerde yer almış, diğer konu başlıklarındaki haberlerin hiçbirinde geçmemiştir. Bu durum, "onur air" kelimesinin *Ekonomi, işletme ve finans* konu başlığı için ayırt edici olduğunu göstermiştir. Öte yandan, derlemin genelinde yaklaşık her yedi "tbmm" kelimesinden bir tanesinin *İş yaşamı* konu başlığına ait haberlerden geldiği görülmüştür. Dolayısıyla "Kurum" etiketiyle etiketlenmiş "tbmm" kelimesinin bu konu başlığı için ayırt ediciliği düşük olduğu anlaşılmıştır. Buradan, hem belirli bir konuda hem de derlem içerisinde yoğun olarak geçen kelimelerin etiketlenmesinin çok da anlamlı olmadığı görülmektedir.

BilCol-2005 derleminin 2005 yılına ait haberlerden oluşturulmuş olmasından yola çıkılarak, haberlerin büyük ölçüde o zaman aralığının gündemini yansıttığı düşünülmektedir. Bu derlemin genelinde en sık geçen kelimeler ile konu başlıkları altında en sık geçen kelimeler incelendiğinde, bunların o zaman aralığında meydana gelen olaylar hakkında ipucu verdiği anlaşılmıştır. Bu tip analizlerin mümkün olduğunca otomatikleştirilerek ayrıntılı ve daha uzun vadeli olarak yapılması sonucunda zaman içerisinde haber kaynaklarının gündeminde öne çıkan konular, kişiler, olaylar, kurumlara ait ilginç örüntüler ortaya konabileceği, bunların hem diğer ülkelerde üretilen haberlerle hem de potansiyel okuyucuların gündemi ile (örneğin sosyal medya verileri kullanılarak) karşılaştırılabileceği düşünülmektedir.

Çalışmamız sonucunda sık tekrarlanan etiketsiz kelimelerin haberlerin konusu hakkında büyük ölçüde ipucu verdiği görülmüştür. Örneğin, *Eğitim* konu başlığında "üniversite", "tasarı", "bilim", "yüksekokul", "fakülte" gibi kelimelerin sık geçen kelimeler olması bu kelimelerin etiketlenmemiş olsa dahi ilgili haberlerin *Eğitim* konusunda olduğu hakkında fikir verdiğini göstermiştir.

Zaman sınırlılığı nedeniyle haber yapısı içerisinde sık tekrarlanan ancak konu başlıklarını tanımlamada bir anlam ifade etmeyen kelimeler için bir durma listesi hazırlanamamıştır, ancak ileride bu konuda yapılacak daha ayrıntılı analizlerde Türkçe haber dilini de yansıtan bir durma kelimeleri listesi hazırlanmasının uygun olacağı düşünülmektedir.

Çalışmamızda Tablo 7 ile Tablo 12 verileri karşılaştırıldığında etiketsiz kelimelerin haberlerin ilgili olduğu konu başlığı ile ilgili ipucu vermede etiketli kelimelere göre daha önemli role sahip olabileceği görülmüştür. Örneğin, *Ekonomi, işletme ve finans* konu başlığında en sık geçen etiketli ve etiketsiz ilk 10 kelime karşılaştırıldığında bu daha net biçimde görünmektedir (bkz. Tablo 13).

Tablo 13. *Ekonomi, işletme ve finans* konu başlığında en sık kullanılan etiketli ve etiketsiz kelimeler

Ekonomi, işletme ve finans	
Etiketli kelimeler	**Etiketsiz kelimeler**
onur air	kredi
türkiye	konut
hollanda	faiz
şener	yıl
almanya	banka
2005	uçuş
cansızlar	aylık
yıldırım	vadeli
ziraat bankası	yasak
istanbul	artış

Tablo 13'te görüldüğü gibi sadece etiketli kelimelere bakıldığında konu başlığı çıkarımında bulunmak daha zor iken, etiketsiz kelimelere bakıldığında konu başlığı tahmininde bulunmak biraz daha kolaylaşmaktadır. Bu durum konu başlıklarının kavramsal olarak yansıtılmasında varlık isimlerinden çok etiketlenmemiş kelimelerin anlamlı olduğuna yönelik tahminimizi doğrulamaktadır. Etiketli ve etiketsiz kelimelerin konu başlıkları ile kavramsal ilişkisine göz kontrolü yapılarak sezgisel olarak karar verilmiş olmakla birlikte bu konunun bundan sonraki araştırmalarda daha büyük kelime setleri ile daha derinlemesine ve uzman görüşleri alınarak araştırılmaya değer olduğu ve elde edilecek ampirik verilerin daha sağlıklı yorumlar yapmaya olanak tanıyacağı düşünülmektedir.

Varlık isimleri ile tanımlanamayan kelimelerin haber konularını belirlemede daha başarılı olmasına yönelik benzer bir sonuç "Türkçe Haber Benzerliklerinin Belirlenmesinde Varlık İsimlerinin Hikâye Bağlantı Algılama Görevinin Başarımına

Etkisi" başlıklı TÜBİTAK Projesi'nde (Proje No: 111K030) de elde edilmiştir. Proje kapsamında "Kişi", "Konum", "Kurum", "Tarih", "Zaman", "Yüzde", "Para" gibi varlık isimlerinin yanı sıra hangi varlık ismi ile etiketleneceğine karar verilemeyen ama haber içeriğini yansıtmada anlamlı olduğu düşünülen ifadeleri etiketlemek için kullanılan "Unknown" etiketine de yer verilmiştir. Projede gerçekleştirilen hikâye bağlantı algılamaya yönelik etkinlik testleri sonucunda, haberlerde geçen varlık isimlerinden "Unknown" etiketli olanların en yüksek başarım gösterdiği tespit edilmiştir (Soydal ve Al, 2014). "Ne" sorusunun yanıtlanması gerekliliğini ortaya koyan bir diğer çalışma Köse'ye (2004) aittir. Bu çalışmada olay modeli yaklaşımını esas alarak haber benzerliklerinin saptanmasında "Kim" etiketinin ne ölçüde etkili olduğunu incelemiş ve neticede sadece "Kim" etiketinin tek başına kullanımının yeterli olmadığı, bunun "Ne", "Nerede", "Ne zaman" etiketleriyle bütünleştirilerek incelenmesi gerekliliğini ortaya koymuştur.

Literatürdeki örneklerde (İrvan ve Çınarbaş, 2002; Hayran ve Özdemir, 2011; Berkant ve Cömert, 2013) de belirtildiği gibi gündeme ait dönemsel özelliklerdeki değişiklikleri izleyebilmek adına bu tip çalışmaların belirli aralıklarla tekrarlanmasında yarar olacaktır. Ayrıca daha uzun ve farklı zaman dilimlerini inceleyen çalışmaların gerçekleştirilmesi hem bu zaman dilimi hakkında bilgi verecektir hem de bu çalışmaların sonuçları üzerinden haberlere yönelik dönemsel çıkarımlar yapılabilecektir. Böylelikle haberlerdeki içeriksel ve niceliksel değişimler izlenebilecek ve her döneme ait gündemdeki eğilimler belirlenebilecektir.

Çalışmada araştırma sorularının yanıtlanması sürecinde şu bulgulara da ulaşılmıştır:

Konu başlıklarına göre etiketsiz kelimeler değerlendirildiğinde, az sayıdaki etiketlenmemiş kelimenin konu başlığı içindeki sıklığının geri kalan çoğunluğa oranla daha fazla olduğu tespit edilmiştir. Bu durum, konu başlıkları altında yer alan haberlerin çoğunlukla belirleyici bir grup etiketsiz kelime ile ifade edildiğini ve etiketlemede ya da konu başlığına karar verilmesinde temel olarak bu kelimelerin dikkate alınmasının yeterli olabileceğini ortaya koymaktadır.

Yedi farklı varlık ismi ile etiketlenmiş olan haberlerde aynı varlık isimlerinin ne oranda tekrar ettiğinin ortaya konması amacıyla haberlerde geçen tekil etiket sayıları ve oranları incelenmiş ve sonuçlar *İş yaşamı* ile *Eğitim* konu başlıklarındaki haberlerin daha farklı ifadeler kullanılarak oluşturulduğunu göstermiştir. Örneğin konu başlıklarını tanımlama potansiyeli olan kelimelere ait bir otorite dizin oluşturulmak istendiğinde bazı

konularla ilgili haber metinlerinde o konulara has bazı ifadelerin yoğunluk kazanabileceği değerlendirilmelidir. Burada dikkat edilmesi ya da araştırılması gereken bir diğer nokta, sözü edilen konu başlıklarının ortaya konulan özelliklerinin derlemden derleme ya da yıldan yıla ne gibi bir değişim gösteriyor olduğudur.

Çalışmada ayrıca haberlerde geçen ortalama kelime sayıları ile ortalama etiket sayıları incelenerek haber metinlerinin uzun ya da kısa oluşunun etiketleme durumunu nasıl etkilediği ortaya konmuştur. *Spor* konu başlığındaki haberler kısa metne sahip olmasına rağmen bu metinlerde yer alan kelimelerin büyük çoğunluğunun etiketlenmeye değer olabileceği görülmüştür. *Kaza ve felaket* konu başlığı hem haberde geçen ortalama kelime sayısı açısından hem de haberde geçen ortalama etiket sayısı açısından en düşük rakamlara sahiptir. Buna dayanarak, *Kaza ve felaket* konu başlığında yer alan haberlerin daha az kelime ile ifade edildiği anlaşılmıştır.

Bu bulgular Türkçe derlemlerde etiketleme işleminin anlamlı şekilde otomatikleştirilebilmesi için haberlerin dâhil oldukları konu özelinde daha detaylı içerik analizi çalışmaları yapılmasının uygun olabileceği konusunda fikir vermektedir. Türkçe bir haber derleminin genelinde kişi, konum ve yer bilgilerine haber yazımının doğası gereği yoğun olarak yer verildiği anlaşılmıştır. Durum her ne kadar öyle olsa da konu başlıkları arasındaki farklılıkları “Yüzde” ve “Para” gibi daha az kullanılan etiketlerin ve etiketlenmemiş, yani “Ne” sorusunun yanıtı sayılabilecek kelimelerin ortaya koyduğu görülmüştür. Bu bulgular ışığında daha büyük verilerle yapılacak uzunlamasına çalışmalar sayesinde ülke gündemindeki dönemsel farklılıklar ile Türkiye’de en çok haber değeri taşıyan kişi, kurum ya da yerler ve bunların zaman içindeki değişiminin görülebileceği, Türkçe bir haber derleminde mutlaka etiketlenmesi gereken kişi, kurum, yer bilgilerine ait otorite dizinlerin elde edilebileceği, konu başlıklarını niteleyen kelimelerin tanımlanması ile haberlerin otomatik sistemler tarafından standart konu başlıkları altında sınıflanmasının mümkün olabileceği düşünülmektedir.

5. BÖLÜM

SONUÇ VE ÖNERİLER

Metnin içindeki ifadelerden bilgi çıkarımı, erişim performansını artırma, metinlerin düzenlenmesi ve sınıflanması gibi daha pek çok konuda işlevsellik sağlayan metin kategorizasyonu son yıllarda önemi giderek daha iyi anlaşılmaya başlayan konulardan biridir.

Metin kategorizasyonunun en gerekli olduğu alanların başında haber metinleri gelmektedir. Gazete, televizyon, radyo gibi geleneksel haber kaynaklarının yanı sıra haber platformlarının İnternet ortamında da yer almaya başlaması yayınlanan haber sayısında ciddi boyutlarda artışa sebep olmuştur. Buna karşılık kaynaklarda yer alan haberlerin okunması ve takibi giderek zorlaşmaktadır. Ayrıca, yayınlanan her haberin doğruluğu, güvenilirliği vb. unsurlar da devreye girdiğinde haber takibi konusunda daha seçici olunması ve içeriğin varlık isimleri ile tanımlanması, üst verilerin ve konu başlıklarının standardizasyonu gibi konuların dikkatle incelenmesi gerektiği anlaşılmaktadır. Bu sorunların literatürdeki çözümleri de değerlendirildiğinde en iyi yaklaşımın metin kategorizasyonu çalışmaları olduğu söylenebilir.

5.1. SONUÇ

Çalışmada “Türkçe Haber Benzerliklerinin Belirlenmesinde Varlık İsimlerinin Hikâye Bağlantı Algılama Görevinin Başarımına Etkisi” başlıklı TÜBİTAK Projesi (Proje No: 111K030) kapsamında etiketlenmiş olan BilCol-2005 haber derlemindeki 5834 haber incelenmiştir. Haberler, “Kişi”, “Kurum”, “Konum”, “Tarih”, “Zaman”, “Para” ve “Yüzde” olmak üzere yedi farklı varlık ismi ile etiketlenmiştir. Derlemdeki haberleri tanımlamada en etkin varlık isimlerinin “Kişi”, “Konum” ve “Kurum” olduğu anlaşılmıştır. Etkinliği en az varlık isimlerini ise “Para”, “Yüzde” ve “Zaman” oluşturmuştur. Etiketleme işleminin ardından derlemdeki haberler, IPTC haber konu taksonomisine dayanan 13 konu başlığı kullanılarak kategorize edilmiştir.

Çalışmamızdaki hipotezler ışığında elde edilen bulgulara ilişkin sonuçlar şunlardır:

- Derlemdeki haberlerin etiketlenmesinde en baskın varlık ismi “Kişi”, en pasif varlık ismi ise “Zaman”dır.

- Derlemdeki haberlerin etiketlenme sayıları IPTC konu başlıklarına göre farklılık göstermektedir.
- Tüm konu başlıkları için "Kişi", "Kurum" ve "Konum" varlık isimlerinin ön planda olduğu anlaşılmıştır.
- Konu başlıklarının kavramsal içeriğini varlık isimlerinden çok etiketlenmemiş kelimelerin yansıttığı görülmüştür.

Derlemin geneline bakıldığında "Kişi", "Konum" ve "Kurum" etiketlerinin belirleyici olduğu göze çarpmaktadır. "Tarih" ve "Para" etiketlerinin *İş yaşamı* konu başlığında ikinci ve üçüncü derece belirleyici olduğu görülmektedir. "Yüzde" etiketinin en fazla ön plana çıktığı konu başlığı *Ekonomi, işletme ve finans*'tır. Bu etiket diğer konu başlıklarının pek çoğunda oldukça az kullanılmıştır. Diğer konu başlıkları ile karşılaştırıldığında, *Ekonomi, işletme ve finans* konu başlığındaki haberlerin tanımlanmasında farklı varlık isimlerinin birbirine daha yakın oranlarda kullanıldığı dikkat çekmektedir. Öte yandan, "Zaman" etiketinin konu başlıklarının hiçbirinde belirleyici role sahip olmadığı anlaşılmaktadır.

Konu başlıklarına göre etiket dağılımlarının pek çoğu derlemin tamamına yönelik etiket dağılımı ile büyük ölçüde benzerlik göstermektedir. *İş yaşamı* ile *Ekonomi, işletme ve finans* konu başlıklarındaki etiket dağılımı derlemin genelindeki etiket dağılımından oldukça farklıdır. Bu haberlerin para, yüzde gibi diğer konu başlıklarında bulunmayan ya da daha az kullanılan varlık isimleri ile tanımlandıkları görülmüştür. Bu farklılığın BilCol-2005 derlemindeki haberlerin dâhil olduğu konu başlıklarının karakteristiğine ve içeriğine bağlı olduğu tahmin edilmektedir.

"Kişi" etiketinin en belirleyici olduğu konu başlıklarının *Bilim ve teknoloji*, *Sanat, kültür ve magazin*, *Siyaset*, *Spor*, *Suç, hukuk ve yargılama* ile *Yaşam ve ilgi alanları* olduğu anlaşılmıştır. "Kişi" etiketinin çalışmamız kapsamında kullanılan IPTC konu başlıklarının yaklaşık yarısında (altı konu başlığı) en belirleyici varlık ismi olması dikkat çekicidir. *Eğitim*, *Ekonomi, işletme ve finans* ile *İş yaşamı* konu başlıklarında ise en belirleyici etiket "Kurum" dur. "Konum" etiketinin ise *Kaza ve felaket*, *Sağlık*, *Savaş ve karışıklıklar* ile *Sosyal konular* konu başlıklarında baskın olduğu görülmüştür. Örneğin, Türkçe haberler için otomatik etiketleme sistemleri yaratılmak istenirse bu konu başlıkları için bu varlık isimlerine karşılık gelebilecek kelimelerin yakalanmaya çalışılmasına öncelik verilmesi uygun olacaktır.

Haberlerde aynı varlık isimlerinin ne oranda tekrar ettiği incelenmiş ve bunun için haberlerde geçen tekil etiket (bir haberde kaç farklı varlık ismi olduğu bilgisi) sayıları ve oranları hesaplanmıştır. Tekil etiket oranı en fazla olan konu başlıkları *İş Yaşamı* ve onu takiben *Eğitim*'dir. Bu durum belirtilen konu başlığındaki haberlerin daha farklı ifadeler kullanılarak oluşturulduğu konusunda ipucu vermektedir. Tekil etiketlenme oranı en az olan konu başlıkları ise *Savaş ve karışıklıklar*, *Siyaset* ve *Spor*'dur. Bu konu başlıklarındaki haberlerde geçen kelimelerin yarıdan fazlası iki veya daha fazla kez etiketlenmiştir. Bu durum, bu konu başlıklarında genelde aynı varlık isimlerinin kullanıldığını ve belirli etiketlerde yoğunluk olduğunu ortaya koymaktadır.

Derlemde sık tekrarlanan kelimelerden çok belirli konu başlıkları altındaki haberlerde sık kullanılmış olan kelimelerin ilgili konu başlıklarını tanımlamada daha ayırt edici olduğu görülmüştür (örneğin; *Ekonomi, işletme ve finans* konu başlığı için "onur air", *Bilim ve teknoloji* konu başlığı için "nobel"). Bu kelimeler yer aldıkları konu başlıkları için ayırt edici özelliğe sahiptir. Derlemin genelinde en sık etiketlenmiş kelime olan "türkiye" aynı zamanda *Spor* ve *Sağlık* konu başlıklarında da en sık geçen kelimedir. Bu durum "türkiye" kelimesinin hem derlemin geneli için hem de bu konu başlıkları için ayırt edici özelliğe sahip olmadığını ortaya koymaktadır.

Spor konu başlığındaki haberler en kısa metne sahip niteliktedir. Diğer yandan, bu konu başlığındaki haberlerde geçen ortalama etiket sayısı diğer konu başlıklarına kıyasla daha yüksek değere ulaşmaktadır. Bu durum, *Spor* konu başlığındaki haberlerin kısa metne sahip olmasına rağmen bu metinlerde yer alan kelimelerin büyük çoğunluğunun etiketlenmeye değer olabileceğini göstermektedir. *Bilim ve teknoloji* konu başlığı hem haberde geçen ortalama kelime sayısı açısından hem de haberde geçen ortalama etiket sayısı açısından ilk sıradadır. *Ekonomi, işletme ve finans* ise haberde geçen kelime sayısı açısından ikinci sıradadır. Bu konu başlıklarındaki haberlerin ifade edilmesinde daha fazla kelime kullanıldığı söylenebilir. Hem haberde geçen ortalama kelime sayısı açısından hem de haberde geçen ortalama etiket sayısı açısından en düşük etiket sayısına sahip konu başlığı ise *Kaza ve felaket*'tir. Buna dayanarak, bu derlemde yer alan *Kaza ve felaket* konu başlığındaki haberlerin daha az kelime ile ifade edildiği söylenebilir.

Haberlerde "NE" sorusuna yanıt verdiği kabul edilen etiketsiz kelimelerin konu başlıklarını tanımlamada etkili oldukları anlaşılmıştır. Sadece etiketli kelimelere bakılarak haberin hangi konu başlığı altında kategorize edilebileceğini tahmin etmek zor iken, etiketsiz kelimelere bakıldığında konu başlığı çıkarımında bulunmak biraz

daha kolaylaşmaktadır (Örneğin, *Eğitim* konu başlığında “üniversite” en sık geçen kelimedir. Öte yandan “tasarı”, “bilim”, “yüksekokul”, “fakülte” gibi kelimelerin de bu konu başlığında sık geçiyor oluşu etiketlenmemiş olsalar dahi bu kelimelerin ilgili haberlerin *Eğitim* konusunda olduğu hakkında fikir verebildiğini göstermektedir.)

Derlemde az sayıdaki etiketlenmemiş kelimenin konu başlığı içindeki ağırlığı geri kalan çoğunluğa oranla daha fazladır. Bu durum, konu başlıkları altında yer alan haberlerin çoğunlukla belirli bir grup etiketsiz kelime ile ifade edildiğini ortaya koymaktadır. Bir başka deyişle, haberlerin tanımlanmasında haberlerde geçen etiketsiz kelimelerin tamamına bakmak yerine, etiketsiz kelimelerin yoğunluk oluşturduğu yüzdelik dilimdeki etiketsiz kelimeleri dikkate almak daha anlamlı görünmektedir.

5.2. ÖNERİLER VE GELECEKTE YAPILABİLECEK ÇALIŞMALAR

Ülkemizde haber muhabirlerinin ve gazetecilerin habere erişmek konusunda kendi yöntemlerini geliştirdikleri ve çoğu gazete arşivinin belge ve bilgi yönetimi standartlarından uzak arşivleri olduğu gözlenmiştir. Kısa vadede bu sorunun ciddiyeti anlaşılmadığı takdirde giderek artan haber unsurlarına erişim ve haber takibinin sağlanması oldukça zorlaşacak gibi görünmektedir. Bu nedenle bu çalışmada da kullanılan IPTC gibi konu başlıkları ve haber kodları içeren uluslararası standartlara uygun haber arşivlerinin oluşturulması gerekmektedir. Ayrıca haber takibinde ve kategorizasyonunda mümkün olan en iyi çözümlerin bulunması için konu ile ilgili araştırmaların artmasına ihtiyaç vardır.

İlk etapta bu konu ile ilgili farkındalık yaratılabilmesi için gazeteci adaylarına verilen habercilik eğitimi içerisinde IPTC, haberlere yönelik kullanılan üst veriler ve basit etiketleme kurallarına ilişkin konuların anlatılması önerilmektedir. Gazeteciler hem konu uzmanı hem de kullanıcı gözüyle haberlere erişmeye çalışmaktadırlar. Bu sayede gazetecilerin daha gazetecilik eğitimi alırken bu konularda bilgi sahibi olmaları sağlanırsa, haberlerin kategorize edilmesi ve sonrasında erişilebilir olması daha kolay hale gelecektir.

Türkçe haber derlemlerinde haberlerin tanımlanması ve kategorizasyonu işlemlerinin daha kolay ve mümkün olduğunca otomatik şekilde yapılmasını sağlayacak sistemler üzerinde çalışılmalıdır. Haberlerin ilgili konu başlıkları altında otomatik kategorizasyonunun sağlanması için çalışmamız bulgularından elde edilen

ipuçlarından yola çıkılarak benzer başka araştırmaların yapılmasının önemli olduğu düşünülmektedir. Sistem tasarımında kullanılmak üzere varlık isimleri dışında kalan kelimelerden bir Türkçe haberlere yönelik otorite dizin hazırlanabilirse bunların ilgili konu başlığı ile otomatik olarak ilişkilendirilebilmesinin mümkün olabileceği düşünülmektedir. Ayrıca haberlerin tanımlanmasında, haberlerde geçen etiketsiz kelimelerin tamamına bakmak yerine, etiketsiz kelimelerin yoğunluk oluşturduğu yüzdelik dilimdeki etiketsiz kelimeler ile dur listesi dışında kalan kelimelerden bir otorite dizin oluşturulabilir. Oluşturulacak bu otorite dizin ile haberlerin dizinlenmesinde aynı anlama gelen farklı ifadeler kullanılması yerine standart bir kullanım gerçekleştirilmesinin mümkün olabileceği düşünülmektedir. Daha büyük verilerle yapılacak analizler sayesinde Türkçe haberlerde en fazla ve en az ayırt edici olan kelime ya da kelime grupları belirlenerek gerekli durumlarda Türkçe haber derlemlerine özgü durma kelimeleri listeleri (stop words) oluşturulabilir. Ancak haber konuları ve unsurlarının zaman içerisinde değişebileceği de dikkate alınmalı ve kelime otorite dizinlerinin ya da durma kelimesi listelerinin sık sık güncellenmesi gerekebileceği de düşünülmelidir.

Haber içeriklerine yönelik olarak, çalışmamıza benzer şekilde ancak daha detaylı ve büyük verilerle çok daha geniş zaman dilimini içerecek şekilde yapılacak çalışmalar sayesinde haber kaynaklarının gündeme bakış açıları ile vatandaşların gündeminin (örneğin sosyal medyada yer alan yorumlar temel alınarak) belirlenmesinin ilginç sonuçlar ortaya koyabileceği düşünülmektedir. Bir başka deyişle haber kaynaklarının gündemi ile vatandaşın gündemi arasındaki benzerlik ve farklılıkları tespit etmek sosyologlar, tarihçiler, siyaset bilimciler açısından önemli olabilir.

Bu çalışmadaki araştırma tasarımından yola çıkılarak benzer şekilde haberlerle ilgili yapılacak bilgi çıkarımı ya da içerik analizi çalışmaları belirli aralıklarla tekrarlanarak ülkenin gündemine ilişkin dönemsel özellikler ortaya konabilir. Daha sonrasında, gerçekleştirilen tüm bu çalışma sonuçları üzerinden Türkiye'de ya da çalışmanın gerçekleştirildiği bölgede en çok haber yapılan konuların, gündemin, eğilimlerin belirlenmesi sağlanarak ülke gündemine dair birtakım çıkarımlar yapılabilir.

Bu çalışmadaki analizler sosyal ağ analizi ile daha farklı boyutlarda gerçekleştirilebilir. Bu tür çalışmalar ile haberlere ilişkin günlük, haftalık, aylık, yıllık, beş yıllık, on yıllık gibi periyodik gündem bulutları oluşturulabilir. Yaratılan bu gündem bulutları bize ilk bakışta gündem hakkındaki eğilimleri söyleyecektir. Gündem bulutları özellikle günümüzde

kullanımı oldukça yoğun olan İnternet tabanlı haber portalları için kullanışlı bir yapı olabilir. Öte yandan böyle bir yapı haber takibi gerçekleştiren ve habere hızlı erişim ihtiyacı duyan kullanıcılar için de işlevsel olabilir.

KAYNAKÇA

Aksoy, C., Can, F., ve Koçberber, S. (2012). Novelty detection for topic trackin. *Journal of the American Society for Information Science and Technology*, 63(4), 777-795. 23.11.2013 tarihinde http://onlinelibrary.wiley.com/doi/10.1002/asi.21697/pdf adresinden erişildi.

Allan, J. (2002). Introduction to topic detection and tracking. *Topic Detection and Tracking: Event-based Information Organization* (Ed. J. Allan) içinde s. 1-16. Kluwer Academic Publishers: New York.

Amasyalı, M. F. ve Diri, B. (2006). Automatic Turkish text categorization in terms of author, genre and gender. *NLDB'06 11th International Conference on Applications of Natural Language to Information Systems* bildirileri içinde s. 221-226.

Bacan, H., Pandzic, I. S., ve Gulija D. (2005). Automated news item categorization. *19th Annual Conference of The Japanese Society for Artificial Intelligence* bildirileri içinde s. 251-256. Kitakyushu, Japan: Springer-Verlag.

Bashaddadh, O. M. A. ve Mohd, M. (2011). Topic detection and tracking interface with named entities approach. *Semantic Technology and Information Retrieval (STAIR), 2011 International Conference on, 28-29 June 2011, Putrajaya, Malaysia*, s. 215-219. 12.02.2014 tarihinde http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=5995791 adresinden erişildi.

Bayraktar, Ö. ve Taşkaya-Temizel, T. (2008). Person name extraction from Turkish financial news text using local grammar-based approach. *Proceedings of the International Symposium on Computer and Information Sciences 2008 (ISCIS'08)* içinde s.1-4. 21.11.2013 tarihinde http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=4717897 adresinden erişildi.

Berkant, H. G. ve Cömert, M. (2013). Günlük gazetelerdeki eğitimle ilgili haberlerin incelenmesi. *Kahramanmaraş Sütçü İmam Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi*, 3(2), 25-44.

Bikel, D.M., Schwartz, R.M. ve Weischedel, R.M. (1999). An algorithm that learns what's in a name. *Machine Learning*, 34(1-3), 211-231.

Bilkent yeni olay belirleme ve izleme deney derlemi. (2013). 22.11.2013 tarihinde http://www.cs.bilkent.edu.tr/~canf/bilcol/bilcol.html adresinden erişildi.

Can, F., Koçberber, S., Bağlıoğlu, Ö., Ercan, G., Kardaş, S., Öcalan, H. Ç. ve diğerleri. (2009). Haber portallarında yenilikçi yaklaşımlar. (Ed.) M. Akgül, E. Derman, U. Çağlayan ve A. Özgit, *XI. Akademik Bilişim Konferansı Bildirileri* içinde s. 589-595. 25.11.2013 tarihinde http://ab.org.tr/ab09/kitap/_AkademikBilisim09.pdf adresinden erişilmiştir.

Can, F., Koçberber, S., Bağlıoğlu, Ö., Kardaş, S., Öcalan, H. Ç. ve Uyar, E. (2007). Türkçe haberlerde yeni olay bulma ve izleme: Bir deney derleminin oluşturulması. (Yay. Haz.) S. Kurbanoğlu, Y. Tonta ve U. Al, *Değişen Dünyada Bilgi Yönetimi Sempozyumu 24-26 Ekim 2007, Ankara, Bildiriler* içinde s. 50-59. 23.11.2013 tarihinde http://by2007.bilgiyonetimi.net/bildiriler/can_ve_digerleri.pdf adresinden erişildi.

Can, F., Koçberber, S., Bağlıoğlu, O., Kardaş, S., Öcalan, H.Ç., ve Uyar, E. (2010). New event detection and topic tracking in Turkish. *Journal of the American Society for Information Science and Technology*, 61(4), 802–819. 25.11.2013 tarihinde http://onlinelibrary.wiley.com/doi/10.1002/asi.21264/pdf adresinden erişildi.

Dalkılıç, F.E., Gelişli, S. ve Diri, B. (2010). Türkçe kural tabanlı varlık ismi tanıma. *18. Sinyal İşleme ve Uygulama Kurultayı, Diyarbakır, 22-24 Nisan 2010,* s. 918-920.

Fiscus, J. G. ve Doddington, G. R. (2002). Topic detection and tracking evaluation overview. *J.Allan (Ed.), Topic detection and tracking: Event-based information organization* içinde s. 17–31. Norwell, MA: Kluwer Academic.

Goodfellow, N. A., Almamani, B. A., Fawwa, A. F. ve McElnay, J. C. (2013). What newspapers say about medication adherence: A content analysis. *BMC Public Health*, 13(909).

Grishman, R. (1997). Information extraction: Techniques and challenges. *Information Extraction A Multidisciplinary Approach to an Emerging Information Technology Lecture Notes in Computer Science*, 1299, s. 10-27.

Grishman, R. ve Sundheim, B. (1996). Message Understanding Conference-6: A brief history. *16th Conference on Computational Linguistics (COLING-96)* bildirileri içinde s. 466-471. Copenhagen, Denmark: Association for Computational Linguistics.

Gui, Y., Gao, Z., Li, R. ve Yang, X. (2012). Hierarchical text classification for news articles based-on named entities. *Advanced Data Mining and Applications Lecture Notes in Computer Science*, 7713, s. 318-329.

Güran, A., Akyokuş, S., Bayazıt, N. G. ve Gürbüz, M. Z. (2009). Turkish text categorization using n-gram words. *Internatonal Symposium on Innovations in Intelligent Systems and Applicatons (INISTA 2009)*'da sunulan bildiri. 29 Haziran-1 Temmuz 2009, Trabzon, Türkiye.

Gürcan, H. İ. ve Batu, Ç. (2001). İnternet haberciliğinde sanal yazı işleri ve gazetecilikte değişen roller. *Türkiye'de İnternet Konferansları VII (inet-tr 2001)*'de sunulan bildiri.

Gürcan, H. İ., Yüksel, E., Vural, A.M., Çetintaş, E. ve Banar, S. (2012). *Haberciliğin temel kavramları* (H. İ. Gürcan Ed.). Eskişehir: Anadolu Üniversitesi.

Güven, E. N., Onur, H. ve Sağıroğlu, Ş. (2008). Yapay sinir ağları ile web içeriklerini sınıflandırma. *Bilgi Dünyası*, 9(1), 158-178. 13.01.2013 tarihinde http://www.unak.org.tr/BilgiDunyasi/gorusler/2008/cilt9/sayi1/158-178.pdf adresinden erişildi.

Han, J. ve Kamber, M. (2006). *Data mining concepts and techniques*. Waltham, USA: Morgan Kaufmann Publishers.

Hayes, P. J., Knecht, L. E. ve Cellio, M. J. (1988). A news story categorization system. Proceedings of *The Second Conference on Applied Natural Language Processing (ANLC'88)* içinde s. 9-17. 21.10.2013 tarihinde http://dl.acm.org/citation.cfm?id=974235.974238 adresinden erişildi.

Hayran, M. ve Özdemir, B. (2011). Sağlık haberlerinin içerik analizi ve medya etiği. *İku*, 25, 30-36. 05.03.2014 tarihinde http://www.iku-dergisi.com/IKU/images/stories/dergi_pdf/25/saglik-haberlerinin-icerik-analizi-medya-etigi-iku25.pdf adresinden erişildi.

IPTC about. (2014). 12.03.2014 tarihinde http://www.iptc.org/site/Home/About/ adresinden erişildi.

IPTC members. (2014). 10.03.2014 tarihinde http://www.iptc.org/site/Home/Members/ adresinden erişildi.

IPTC membership Q&A. (2012). 10.03.2014 tarihinde http://www.iptc.org/site/Home/About/Membership_Q&A adresinden erişildi.

IPTC news codes. (2014). 05.03.2014 tarihinde http://www.iptc.org/site/NewsCodes/View_NewsCodes/ adresinden erişildi.

İrvan, S. ve Çınarbaş, S. (2002). Türk basınında yer alan trafik haberlerinin analizi. *Trafik Güvenliği Kongre Yayınları*. 05.03.2014 tarihinde http://www.trafik.gov.tr/SiteAssets/Yayinlar/Bildiriler/pdf/C13-47.pdf adresinden erişildi.

Jo, T.C. (1999). The categorization of news articles with informative keywords. *The Proceedings of ITC-CSCC 99, 1999* içinde s. 1136-1139. 20.10.2013 tarihinde http://tjo018.inha.ac.kr/Achievements/Research/Proceedings/1999_006.pdf adresinden erişildi.

Kardaş, S. (2009). *New event detection and tracking in Turkish*. Yayımlanmamış yüksek lisans tezi, Bilkent Üniversitesi, Ankara.

Köse, G. (2004). *Konu algılama ve izleme programında olay modeli*. Yayımlanmamış yüksek lisans tezi, Başkent Üniversitesi, Ankara.

Ku, L. –W., Liang, Y. –T. ve Chen, H. –H. (2006). Opinion extraction, summarization and tracking in news and blog corpora. *AAAI-2006 Spring Symposium on Computational Approaches to Analyzing Weblogs, March 27-29, 2006, Palo Alto, California, USA*. 20.11.2013 tarihinde http://nlg18.csie.ntu.edu.tw:8080/opinion/SS0603KuLW.pdf adresinden erişildi.

Kumaran, G. ve Allan, J. (2004). Text classification and named entities for new event detection. *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (ACM SIGIR'04)* içinde s. 297-304. Sheffield, UK: ACM.

Kuo, Z., Zi, J. ve Gang, L. W. (2007). New event detection based on indexing-tree and named entity. *SIGIR'07, Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* July 23-27, 2007, Amsterdam içinde s. 215-222. ACM: New York, USA. 21.11.2013 tarihinde http://dl.acm.org/citation.cfm?id=1277780 adresinden erişildi.

Küçük, D. ve Yazıcı, A. (2009a). Rule-based named entity recognition from Turkish texts*. International Symposium on Inovations in Intelligent Systems and Applications, Trabzon, Turkey, June 29-July 1, 2009*, sunum.

Küçük, D. ve Yazıcı, A. (2009b). Named entity recognition experiments on Turkish texts. *Proceedings of the International Conference on Flexible Query Answering Systems. Roskilde, Denmark*. T. Andreasen et al. (Ed.): FQAS 2009, LNAI 5822 içinde s. 524-535.

Küçük, D. ve Yazıcı, A. (2010). A hybrid named entity recognizer for Turkish with applications to different text genres. In *Proceedings of the 25th International*

Symposium on Computer and Information Sciences (ISCIS). London, UK. E. Gelenbe et al. (Ed.): Computer and Information Sciences, LNEE 62, s. 113-116.

Lewis, D. D. (1991). Evaluating text categorization. *Proceedings of the Workshop on Speech and Natural Language* (HLT '91) içinde s. 312-318. 13.01 2013 tarihinde http://dl.acm.org/citation.cfm?id=112471 adresinden erişildi.

Makkonen, J., Ahonen-myka, H. ve Salmenkivi, M. (2003). Topic detection and tracking with spatio-temporal evidence. *25th European Conference on Information Retrieval Research (ECIR 2003), April 14-16, 2003, Pisa, Italy* bildirileri içinde s. 251-265.

Marrero, M., Urbano, J., Sanchez-Cuadrado, S., Morato, J. ve Gomez-Berbis, J. M. (2012). Named entity recognition: Fallacies, challenges and opportunities. *Computer Standards & Interfaces*, s. 1-8. 18.01.2013 tarihinde http://www.sciencedirect.com/science/article/pii/S0920548912001080 adresinden erişildi.

Mathew, T. (2006). Text categorization using n-grams and Hidden-Markov-Models. 15.01.2014 tarihinde http://www.slideshare.net/thomas_a_mathew/text-categorization-using-ngrams-and-hiddenmarkovmodels adresinden erişildi.

McNamee, P., Mayfield, J. C. ve Piatko, C. D. (2011). Processing named entities in text. *John Hopkins Apl Technical Digest*, 30 (1), 31-40. 13.01.2013 tarihinde http://techdigest.jhuapl.edu/TD/td3001/McNamee.pdf adresinden erişildi.

MEGEP (Mesleki Eğitim Ve Öğretim Sisteminin Güçlendirilmesi Projesi). (2007). *Gazetecilik alanı haber yazma teknikleri*. Ankara: Milli Eğitim Bakanlığı.

Moral, C., Antonio, A. de, Imbert, R. ve Ramirez, J. (2014). A survey of stemming algorithms in information retrieval. *Information Research*, 19(1). 05.03.2014 tarihinde http://www.informationr.net/ir/19-1/paper605.html#.U3yyuPl_sjU adresinden erişildi.

Nadeau, D. ve Sekine, S. (2007). A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1), 3-26. 13.01.2013 tarihinde

http://www.mendeley.com/catalog/survey-named-entity-recognition-classification-4/ adresinden erişildi.

Öcalan, H. Ç. (2009). *Bilkent News Portal: A system with new event detection and tracking capabilities.* Yayımlanmamış yüksek lisans tezi, Bilkent Üniversitesi, Ankara.

Palmer, D. D. ve Day, D. S. (1997). A statistical profile of the named entity task. *Fifth Conference on Applied Natural Language Processing ANLC'97* bildirileri içinde s. 190-193. Stroudsburg, PA, USA: Association for Computational Linguistics.

Sang, E. F. T. K. (2002). Introduction to the CoNLL-2002 shared task: Language-independent named entity recognition. *Conference on Computational Natural Language Learning (CoNLL-2002)* bildiriler. Taipei, Taiwan: Association for Computational Linguistics.

Sang, E. F. T. K. ve Meulder, F. De. (2003). Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. *Conference on Computational Natural Language Learning (CoNLL-2003)* bildiriler. Edmonton, Canada: Association for Computational Linguistics.

Sekine, S. ve Isahara, H. (2000). IREX: IR and IE evaluation project in Japanese. *LREC 2000 Second International Conference on Language Resources and Evaluation* bildiriler. Athens, Greece: European Language Resources Association.

Sekine, S. ve Nobata, C. (2004). Definition, dictionaries and tagger for extended named entity hierarchy. *LREC 2004 Fourth International Conference on Language Resources and Evaluation* bildirileri içinde s. 1977-1980. Lisbon, Potugal: European Language Resources Association.

Sekine, S., Sudo, K. ve Nobata, C. (2002). Extended named entity hierarchy. *LREC 2002 Third International Conference on Language Resources and Evaluation* bildirileri içinde s. 1818-1824. Las Palmas, Canary Island, Spain: European Language Resources Association.

Shah, C., Croft, W. B. ve Jensen, D. (2006). Representing documents with named entities for story link detection (SLD). A poster presentation at the *ACM Fifteenth Conference on Information and Knowledge Management (CIKM) 2006, Arlington VA, November 6-11, 2006*.

Shinyama, Y. ve Sekine, S. (2004). Named entity discovery using comparable news articles. *20th International Conference on Computational Linguistics (COLING'04)* bildirileri içinde makale no: 848. USA: Association for Computational Linguistics.

Soydal, İ. ve Al, U. (2014). *Türkçe Haber Benzerliklerinin Belirlenmesinde Varlık İsimlerinin Hikâye Bağlantı Algılama Görevinin Başarımına Etkisi* (TÜBİTAK Sosyal Bilimler Araştırma Grubu Proje No: SOBAG 111K030). Ankara: TÜBİTAK.

Sundheim, B. M. (1995). Overview of the MUC-6 evaluation. *6th Conference on Message Understanding bildiriler*. Columbia, Maryland: Association for Computational Linguistics.

Tatar, S. ve Çiçekli, İ. (2011). Automatic rule learning exploiting morphological features for named entitiy recognition in Turkish. *Journal of Information Science*, 37(2), 137-151.

Temmink, A. (2010). *Using the IPTC taxonomy to classify articles automatically fort he MD Info taxonomy*. Yüksek lisans tezi, Erasmus School of Economics: Rotterdam.

Tonta, Y., Bitirim, Y. ve Sever, H. (2002). *Arama motorlarında performans değerlendirme*. Ankara: Total Bilişim Ltd. Şti.

Topic Detection and Tracking Evaluation. (2008). 08.01.2014 tarihinde http://www.itl.nist.gov/iad/mig//tests/tdt/ adresinden erişildi.

Troncy, R. (2008). Bringing the IPTC news architecture into the semantic web. A. Sheth, S. Staab, M. Dean, M. Paolucci, D. Maynard ve T. Finin (eds.), *7th International Conference on The Semantic Web (ISWC'08) bildirileri* içinde s. 483-498. Berlin, Heidelberg: Springer.

Tür, G., Hakkani-Tür, D. ve Oflazer, K. (2003). A statistical information extraction system for Turkish. *Natural Language Engineering*. 9(2), 181-210.

Uyar, E. (2009). *Near-duplicate news detection using named entities. [Adlandırılmış nesneler kullanarak yaklaşık-aynı haberleri saptama]*. Yayımlanmamış yüksek lisans tezi, Bilkent Üniversitesi, Ankara.

Wayne, C. L. (1998). Topic detection & tracking (TDT) overview & perspective. *Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop, February 8-11, 1998, Lansdowne, Virginia*. 05.01.2014 tarihinde http://www.itl.nist.gov/iad/mig/publications/proceedings/darpa98/html/tdt10/tdt10.htm adresinden erişildi.

Yang, Y., Ault, T., Pierce, T., ve Lattimer, C. W. (2000). Improving text categorization methods for event tracking. *ACM SIGIR Research and Development in Information Retrieval* bildirileri içinde s. 65-72.

Zhang, X., Wang, T. ve Chen, H. (2008). Story link detection based on event model with uneven SVM. *Information Retrieval Technology Lecture Notes in Computer Science*, 4993, s. 436-441.

EKLER

EK 1

Haber Başlığı Numarası	Haber başlığı	IPTC konu başlığı
1	Kars'ta Trafik Kazası 7 ölü 35 yaralı	Kaza ve felaket
2	Onur Air'in Avrupa'nın bazı ülkelerinde iniş kalkışının yasaklanması	Ekonomi, işletme ve finans
3	Koreli bilim adamının kök hücre araştırması sahte	Bilim ve teknoloji
4	Nema karşılığı kredi	Ekonomi, işletme ve finans
5	Tokyo'da Haremlik Selamlik Metro	Sosyal konular
6	Londra metrosunda patlama	Savaş ve karışıklıklar
7	Çocuk tacizi skandalı	Suç, hukuk ve yargılama
8	Formula G	Spor
9	Karamürsel Kaymakamı İsmail Aka İntihar	Sosyal konular
10	400 Koyun İntihar etti	Yaşam ve ilgi alanları
11	Şemdinli Olayları	Savaş ve karışıklıklar
12	Türkiye'de Kuş Gribi	Sağlık
13	Şampiyon Fenerbahçe	Spor
14	Mortgage Türkiye'de	Ekonomi, işletme ve finans
15	2005 Avrupa Basketbol Şampiyonası	Spor
16	Van Yüzüncü Yıl Üniversitesi Rektörü Prof.Dr. Yücel Aşkın tutuklandı	Suç, hukuk ve yargılama
17	Kral Fahd hastaneye kaldırıldı	Siyaset
18	Memurlarının bir üst dereceye çıkması	İş yaşamı
19	Bill Gates türkiye'ye geldi	Bilim ve teknoloji
20	Mısır'da üst üste patlamalar	Savaş ve karışıklıklar
21	Atilla Ilhan vefat etti	Sanat, kültür ve magazin
22	Ata Türk (24) öldürülmesi	Sanat, kültür ve magazin
23	DT Genel Müdürü Lemi Bilgin görevden alındı	Sanat, kültür ve magazin

Haber Başlığı Numarası	Haber başlığı	IPTC konu başlığı (devam)
24	Universiade 2005	Spor
25	Yahya Murat Demirel Bulgaristan'da yakalandı	Suç, hukuk ve yargılama
26	Bağdat El Ayma köprüsü üzerinde izdihamda çok sayıda insan öldü	Kaza ve felaket
27	Prof. Dr. Sadettin Güner ve oğlu Trabzon'da öldürüldü	Suç, hukuk ve yargılama
28	Nestle'den Mürekkepli Süt	Sağlık
29	Nermin Erbakan tedavi altına alındı	Siyaset
30	Ulubey'de çocukla annenin peşpeşe ölümü korkuya yol açtı	Yaşam ve ilgi alanları
31	15. Akdeniz Oyunları	Spor
32	Kemal Derviş'in UNDP Başkanı Seçilmesi ve Göreve Başlaması	Siyaset
33	Caferi'nin Tarihi Tahran Ziyareti	Siyaset
34	Gediz'de grizu patlaması	Kaza ve felaket
35	Sarıgül kendini savunacak	Siyaset
36	Paris'te göstericiler polisle çatıştı	Sosyal konular
37	Rock'n Coke	Sanat, kültür ve magazin
38	Ankara' da Tren Kazasi	Kaza ve felaket
39	2005 Nobel Tıp Ödülü gastrit ve ülserin bakterilerden kaynaklanması	Bilim ve teknoloji
40	Kayseri Erciyes Üniversitesi bebek ölümleri	Sağlık
41	Marburg virüsünden ölenler	Sağlık
42	Gamze Özçelik'in görüntülerinin internette yayınlanması	Suç, hukuk ve yargılama
43	Türkiye'nin ilk yediz bebekleri geliyor	Yaşam ve ilgi alanları
44	Yeni Turk Ceza Kanunu yururluge girdi	Suç, hukuk ve yargılama
45	Saddam Hüseyin'in Yargılanmaya Başlanması	Suç, hukuk ve yargılama
46	Beylikdüzü çöpte patlama	Kaza ve felaket
47	Endonezya'nın Bali Adası'nda eşzamanlı patlamalar	Kaza ve felaket
48	Sahte rakı	Sosyal konular
49	Hindistan'da meydana gelen patlamalar	Savaş ve karışıklıklar
50	Bülent Ersoy ve Deniz Baykal Polemiği	Sanat, kültür ve magazin
51	Tahran'da askeri uçak düştü	Kaza ve felaket
52	Sochi seferini yapan Ufuk-1 gemisi yandı	Kaza ve felaket

Haber Başlığı Numarası	Haber başlığı	IPTC konu başlığı (devam)
53	Eminönü İstanbul'da kanalizasyonda ölen isçiler YENİ	Kaza ve felaket
54	İstanbulda dünya kadınlar günü için izinsiz gösteri yapanları coplayan 3 polis açığa alındı	Suç, hukuk ve yargılama
55	Kuşadası'nda minibüsdeki patlamada beş kişi öldü	Savaş ve karışıklıklar
56	Esenboğa Havalimanı iç hatlar terminali tamamen yandı	Kaza ve felaket
57	Zeytinburnu'nda bir evde meydana gelen patlamada iki kişi öldü	Kaza ve felaket
58	Malatya çocuk yuvasında işkence	Sosyal konular
59	ABD denizaltısı ile Türk gemisi çarpıştı	Kaza ve felaket
60	Prof Dr. Kalaycı silahlı saldırı sonucu öldürüldü	Suç, hukuk ve yargılama
61	İlk Yüz Nakli	Sağlık
62	15 Yeni Üniversite Kuruluyor	Eğitim
63	Gaziantep tanker patlaması	Kaza ve felaket
64	Hakkari'de bomba patladı	Savaş ve karışıklıklar
65	Erzurum çocuk yuvası bebek ölümü	Sosyal konular
66	Kâzım Koyuncunun ölümü	Sanat, kültür ve magazin
67	Melih Kibar ın ölümü	Sanat, kültür ve magazin
68	Sarıkamış şehitleri 91. yıldönümünde de unutulmadı	Yaşam ve ilgi alanları
69	Endonezya'da yolcu uçağı düştü	Kaza ve felaket
70	Şanlıurfa'da köprü inşaatı çöktü	Kaza ve felaket
71	Japonya Osaka'da tren kazası	Kaza ve felaket
72	Manken Tuğçe Kazaz'ın din değiştirmesi	Sanat, kültür ve magazin
73	Fotoğraf sanatçısı Mehmet Gülbiz'in öldürülmesi	Suç, hukuk ve yargılama
74	Yunanistan'da Türk bayrağına çirkin saldırı	Siyaset
75	Maslak'ta Patlama	Kaza ve felaket
76	Didim'de denize uçak düştü	Kaza ve felaket
77	Rum yolcu uçağı düştü	Kaza ve felaket
78	İstiklal Caddesindeki ağaçların kaldırılması	Sosyal konular
79	Zeytinburnunda gemi battı	Kaza ve felaket
80	Bin Yıllık Yolculuk Sergisi	Sanat, kültür ve magazin

EK 2

Metin Formatlama Makrosu

```
Sub FormatText()

Dim lRowCount, lStart, lStop As Long
Dim sWord As String

    Columns("F:F").Select
    Selection.Replace What:="=-", Replacement:="", LookAt:=xlPart, _
        SearchOrder:=xlByRows, MatchCase:=False, SearchFormat:=False, _
        ReplaceFormat:=False
    Selection.Replace What:=Chr(34), Replacement:="", LookAt:=xlPart, _
        SearchOrder:=xlByRows, MatchCase:=False, SearchFormat:=False, _
        ReplaceFormat:=False
    Selection.Replace What:=Chr(39) & Chr(39), Replacement:="", LookAt:=xlPart,
        SearchOrder:=xlByRows, MatchCase:=False, SearchFormat:=False, _
        ReplaceFormat:=False
    Selection.Replace What:="(", Replacement:="", LookAt:=xlPart, _
        SearchOrder:=xlByRows, MatchCase:=False, SearchFormat:=False, _
        ReplaceFormat:=False
    Selection.Replace What:=")", Replacement:="", LookAt:=xlPart, _
        SearchOrder:=xlByRows, MatchCase:=False, SearchFormat:=False, _
        ReplaceFormat:=False
    Selection.Replace What:=":", Replacement:="", LookAt:=xlPart, _
        SearchOrder:=xlByRows, MatchCase:=False, SearchFormat:=False, _
        ReplaceFormat:=False
    Selection.Replace What:=";", Replacement:="", LookAt:=xlPart, _
        SearchOrder:=xlByRows, MatchCase:=False, SearchFormat:=False, _
        ReplaceFormat:=False
```

```
    Selection.Replace What:=",", Replacement:="", LookAt:=xlPart, _
        SearchOrder:=xlByRows, MatchCase:=False, SearchFormat:=False, _
        ReplaceFormat:=False

lRowCount = 0

While Sheets("news").Range("A1").Offset(lRowCount, 0).Value <> ""

' şu karakter için => ?
lStart = 0
lStart = InStr(1, Sheet1.Range("F1").Offset(lRowCount, 0).Value, "?", vbTextCompare)
If lStart > 0 Then
    lStop = InStr(lStart, Sheet1.Range("F1").Offset(lRowCount, 0).Value, " ", vbTextCompare)
    If lStop > 0 Then
        sWord = Mid(Sheet1.Range("F1").Offset(lRowCount, 0).Value, lStart, lStop - lStart)
        Sheet1.Range("F1").Offset(lRowCount, 0).Value =
Application.WorksheetFunction.Substitute(Sheet1.Range("F1").Offset(lRowCount, 0).Value, sWord, "")
        lRowCount = lRowCount - 1
    Else
        Sheet1.Range("F1").Offset(lRowCount, 0).Value = Left(Sheet1.Range("F1").Offset(lRowCount,
0).Value, lStart - 1)
        lRowCount = lRowCount - 1
    End If
End If

' şu karakter için => '
lStart = 0
lStart = InStr(1, Sheet1.Range("F1").Offset(lRowCount, 0).Formula, "'", vbTextCompare)
If lStart > 0 Then
    If lStart > 1 Then
        lStop = InStr(lStart, Sheet1.Range("F1").Offset(lRowCount, 0).Formula, " ", vbTextCompare)
        If lStop > 0 Then
            sWord = Mid(Sheet1.Range("F1").Offset(lRowCount, 0).Formula, lStart, lStop - lStart)
```

```
            Sheet1.Range("F1").Offset(lRowCount, 0).Formula =
Application.WorksheetFunction.Substitute(Sheet1.Range("F1").Offset(lRowCount, 0).Formula, sWord, "")

            lRowCount = lRowCount - 1

        Else

            Sheet1.Range("F1").Offset(lRowCount, 0).Value = Left(Sheet1.Range("F1").Offset(lRowCount,
0).Value, lStart - 1)

            lRowCount = lRowCount - 1

        End If

    Else

        Sheet1.Range("F1").Offset(lRowCount, 0).Formula =
Application.WorksheetFunction.Substitute(Sheet1.Range("F1").Offset(lRowCount, 0).Formula, Chr(39), "",
1)

        lRowCount = lRowCount - 1

    End If

End If

lRowCount = lRowCount + 1

Wend


End Sub
```

EK 3

Saydırma Makrosu

```
Sub TopicCount()

Dim sEnd As String

Dim lEnd, lTopic, lRowCount As Long


Sheets("Results").Activate

    ActiveSheet.Range("$A:$F").RemoveDuplicates Columns:=Array(1, 2), Header:=xlYes


    sEnd = Split(Range("B2").End(xlDown).Address, "$", , vbTextCompare)(2)

    lEnd = CLng(sEnd) + 2


    For lTopic = 80 To 1 Step -1

        Range("A" & Trim(Str(lEnd))).Value = "TOTAL_TOPIC_" & Trim(Str(lTopic))

        Range("B" & Trim(Str(lEnd))).Formula = "=COUNTIF(B2:" & Range("B2").End(xlDown).Address & ","
& Trim(Str(lTopic)) & ")"

        lEnd = lEnd + 1

    Next lTopic


    Columns("C:F").ClearContents

    Range("C1").Value = "DATE_COUNT"

    Range("D1").Value = "LOCATION_COUNT"

    Range("E1").Value = "MONEY_COUNT"

    Range("F1").Value = "ORGANIZATION_COUNT"

    Range("G1").Value = "PERCENTAGE_COUNT"

    Range("H1").Value = "PERSON_COUNT"

    Range("I1").Value = "TIME_COUNT"

    Range("J1").Value = "UNKNOWN_COUNT"

    Range("K1").Value = "UNIQUE_TAG_COUNT"

    Range("L1").Value = "TOTAL_TAG_COUNT"

    Columns("A:L").EntireColumn.AutoFit
```

```
    Range("C2").FormulaR1C1 = "=COUNTIFS(news!C1,RC1,news!C5,LEFT(R1C,LEN(R1C)-6))"
    Range("L2").Formula = "=SUM(C2:J2)"
    Range("C2").Select
    Selection.AutoFill Destination:=Range("C2:J2"), Type:=xlFillDefault
    Range("C2:L2").Select
    Selection.AutoFill Destination:=Range("C2:L" & sEnd)
    Range("C2").Select
    Range(Selection, Selection.End(xlToRight)).Select
    Range(Selection, Selection.End(xlDown)).Select
    Selection.Copy
    Selection.PasteSpecial Paste:=xlPasteValues, Operation:=xlNone, SkipBlanks:=False,
Transpose:=False

    lRowCount = 0

While Range("A2").Offset(lRowCount, 0).Value <> ""
    Range("K2").Offset(lRowCount, 0).Value = UniqueCount(Range("A2").Offset(lRowCount, 0).Value)
    DoEvents
    lRowCount = lRowCount + 1
Wend

End Sub

Function UniqueCount(sDocID As String) As Long

Dim lURowCount, lTopicCount As Long
Dim cUnique As Collection

lURowCount = 0

Set cUnique = New Collection

While Sheets("news").Range("A1").Offset(lURowCount, 0).Value <> ""
```

```
If Sheets("news").Range("A1").Offset(lURowCount, 0).Value = sDocID Then

    On Error Resume Next

    cUnique.Add Sheets("news").Range("F2").Offset(lURowCount, 0).Value, Chr(34) &
Sheets("news").Range("F2").Offset(lURowCount, 0).Value & Chr(34)

    On Error GoTo 0

End If

lURowCount = lURowCount + 1

Wend

UniqueCount = cUnique.Count - 1

End Function
```

EK 4

Etiketleri Orijinal Metinden Atma Makrosu

```
Private Sub CommandButton1_Click()
  Sayfa1.Columns(3).Clear
  Rem son satiri bul
  For I = 2 To 210000
    If Cells(I, 1) = "" Then satır = I - 1: GoTo devam
  Next I
devam:
  For I = 2 To satır
    If Cells(I, 1) <> "" Then
      ID = Cells(I, 1)
      For J = 2 To 210000
        If Sayfa1.Cells(J, 1) = ID Then
          text1 = Cells(I, 2)
          If Sayfa1.Cells(J, 3) = "" Then text2 = Sayfa1.Cells(J, 2) Else text2 = Sayfa1.Cells(J, 3)
          For K = 1 To Len(text2)
            If Mid(text2, K, Len(text1)) = text1 Then
              alsol = Left(text2, K - 1)
              alsag = Right(text2, Len(text2) - K - Len(text1))
              birles = alsol + alsag
              Sayfa1.Cells(J, 3) = birles
            End If
          Next K
        End If
      Next J
End If
  Next I    End Sub
```

EK 5

Kelime Sıklığı Saydırma Makrosu

```
Option Explicit

Sub MakeWordList()
    Dim InputSheet As Worksheet
    Dim WordListSheet As Worksheet
    Dim PuncChars As Variant, x As Variant
    Dim i As Long, r As Long
    Dim txt As String
    Dim wordCnt As Long
    Dim AllWords As Range
    Dim PC As PivotCache
    Dim PT As PivotTable

    Application.ScreenUpdating = False
    Set InputSheet = ActiveSheet
    Set WordListSheet = Worksheets.Add(after:=Worksheets(Sheets.Count))
    WordListSheet.Range("A1") = "All Words"
    WordListSheet.Range("A1").Font.Bold = True
    InputSheet.Activate
    wordCnt = 2
    PuncChars = Array(".", ",", ";", ":", """, "!", "#", _
        "$", "%", "&", "(", ")", " - ", " _", "--", "+", _
        "=", "~", "/", "\", "{", "}", "[", "]", """", "?", "*")
    r = 1

'   Bos yere kadar
    Do While Cells(r, 1) <> ""
'       covert to UPPERCASE
```

```
        txt = UCase(Cells(r, 1))
'       sil punctuation
        For i = 0 To UBound(PuncChars)
            txt = Replace(txt, PuncChars(i), "")
        Next i
'       sil
        txt = WorksheetFunction.Trim(txt)
x = Split(txt)
        For i = 0 To UBound(x)
            WordListSheet.Cells(wordCnt, 1) = x(i)
            wordCnt = wordCnt + 1
        Next i
    r = r + 1
    Loop

'   Create tablo
    WordListSheet.Activate
    Set AllWords = Range("A1").CurrentRegion
    Set PC = ActiveWorkbook.PivotCaches.Add _
        (SourceType:=xlDatabase, _
        SourceData:=AllWords)
 Set PT = PC.CreatePivotTable _
        (TableDestination:=Range("C1"), _
        TableName:="PivotTable1")
    With PT
        .AddDataField .PivotFields("All Words")
        .PivotFields("All Words").Orientation = xlRowField
    End With
End Sub
```

EK 6

IPTC news subject codes	IPTC haber konu başlıkları
Arts, culture and entertainment	Sanat, kültür ve magazin
Crime, law and justice	Suç, hukuk ve yargılama
Disaster and accident	Kaza ve felaket
Economy, business and finance	Ekonomi, işletme ve finans
Education	Eğitim
Environmental issue	Çevre
Health	Sağlık
Human interest	Yaşam ve ilgi alanları
Labour	İş yaşamı
Lifestyle and leisure	Yaşam tarzı, tatil
Politics	Siyaset
Religion and belief	Din ve inanç
Science and technology	Bilim ve teknoloji
Social issue	Sosyal konular
Sport	Spor
Unrest, conflicts and war	Savaş ve karışıklıklar
Weather	Hava

EK 7

Mann-Whitney Testi Sonuçları

IPTC Konu Başlıkları	U	Z	p	r
Bilim ve teknoloji -İş yaşamı	558,500	-4,749	0,000	-0,47
Bilim ve teknoloji - Kaza ve felaket	6042,500	-5,679	0,000	-0,24
Bilim ve teknoloji - Sağlık	5178,000	-3,499	0,000	-0,18
Bilim ve teknoloji - Savaş ve karışıklıklar	15963,000	-3,441	0,001	-0,11
Bilim ve teknoloji - Sosyal konular	10748,500	-4,858	0,000	-0,17
Bilim ve teknoloji - Spor	9203,500	-4,643	0,000	-0,17
Bilim ve teknoloji - Yaşam ve ilgi alanları	1236,000	-3,218	0,001	-0,29
Eğitim - Ekonomi, işletme ve finans	11351,000	-2,931	0,003	-0,12
Eğitim - İş yaşamı	538,000	-5,479	0,000	-0,53
Eğitim - Kaza ve felaket	5635,500	-7,037	0,000	-0,30
Eğitim - Sağlık	5565,000	-4,060	0,000	-0,21
Eğitim - Savaş ve karışıklıklar	17457,500	-3,904	0,000	-0,12
Eğitim - Sosyal konular	10887,500	-5,859	0,000	-0,20
Eğitim - Spor	9533,000	-5,498	0,000	-0,21
Eğitim - Yaşam ve ilgi alanları	1235,000	-4,069	0,000	-0,35
Ekonomi, işletme ve finans - İş yaşamı	9643,500	-4,303	0,000	-0,17
Ekonomi, işletme ve finans - Kaza ve felaket	102110,000	-8,472	0,000	-0,26
Ekonomi, işletme ve finans - Siyaset	98740,500	-3,803	0,000	-0,12
Ekonomi, işletme ve finans - Sosyal konular	175963,500	-6,699	0,000	-0,18
Ekonomi, işletme ve finans - Spor	151626,500	-5,612	0,000	-0,16
İş yaşamı - Sağlık	6221,000	-3,164	0,002	-0,16
İş yaşamı - Sanat, kültür ve magazin	4074,000	-5,442	0,000	-0,29
İş yaşamı - Savaş ve karışıklıklar	18656,000	-3,330	0,001	-0,10
İş yaşamı - Suç, hukuk ve yargılama	16602,000	-5,223	0,000	-0,15
İş yaşamı - Yaşam ve ilgi alanları	1474,000	-2,973	0,003	-0,26
Kaza ve felaket - Sağlık	65258,000	-5,153	0,000	-0,18
Kaza ve felaket - Sanat, kültür ve magazin	43741,000	-9,742	0,000	-0,34
Kaza ve felaket - Savaş ve karışıklıklar	195302,500	-6,769	0,000	-0,18
Kaza ve felaket - Siyaset	57119,500	-11,848	0,000	-0,39
Kaza ve felaket - Spor	150223,500	-3,286	0,001	-0,10
Kaza ve felaket - Suç, hukuk ve yargılama	176944,500	-11,884	0,000	-0,30
Kaza ve felaket - Yaşam ve ilgi alanları	15495,000	-3,550	0,000	-0,15
Sağlık - Sanat, kültür ve magazin	37750,500	-4,095	0,000	-0,17
Sağlık - Siyaset	50410,500	5,355	0,000	0,20
Sağlık - Sosyal konular	110556,000	-3,427	0,001	-0,10
Sağlık - Suç, hukuk ve yargılama	149931,000	-3,933	0,000	-0,10
Sanat, kültür ve magazin - Savaş ve karışıklıklar	115930,500	-4,458	0,000	-0,13
Sanat, kültür ve magazin - Sosyal konular	76882,500	-8,296	0,000	-0,25
Sanat, kültür ve magazin - Spor	67098,500	-7,291	0,000	-0,24
Sanat, kültür ve magazin - Yaşam ve ilgi alanları	9124,000	-2,925	0,003	-0,15
Savaş ve karışıklıklar - Siyaset	154796,000	-6,230	0,000	-0,17
Savaş ve karışıklıklar - Sosyal konular	330186,500	-4,864	0,000	-0,12
Savaş ve karışıklıklar - Spor	285191,000	-3,518	0,000	-0,12
Savaş ve karışıklıklar - Suç, hukuk ve yargılama	457462,500	-5,073	0,000	-0,11
Siyaset - Sosyal konular	101552,500	-10,457	0,000	-0,30
Siyaset - Spor	88935,500	-9,223	0,000	-0,28
Siyaset - Yaşam ve ilgi alanları	12032,500	-3,692	0,000	-0,17
Sosyal konular - Suç, hukuk ve yargılama	307587,500	-10,536	0,000	-0,24
Spor - Suç, hukuk ve yargılama	268754,000	-8,741	0,000	-0,21