Türkçe Etiketli Metin Derlemi
Turkish Labeled Text Corpus

Seçil Öztürk*, Bülent Sankur*, Tunga Güngör†
Elektrik Elektronik* ve Bilgisayar† Mühendisliği Bölümleri
Boğaziçi Üniversitesi
{secil.ozturk, bulent.sankur, gungort}@boun.edu.tr

Mustafa Berkay Yılmaz
Bilgisayar Bilimi ve Mühendisliği Bölümü
Sabancı Üniversitesi
berkayyilmaz@sabanciuniv.edu

Bilge Köroğlu, Onur Ağın, Mustafa İşbilen, Çağdaş Ulaş, Mehmet Ahat
Ar-Ge Bölümü
Yapı ve Kredi Bankası A.Ş
{bilge.koroglu, Onur.Agin, mustafa.isbilen, Cagdas.Ulas, Mehmet.Ahat}@yapikredi.com.tr

***Özetçe* —Türkçe makalelerin başlık, özetçe ve anahtar kelimelerinden oluşan sınıf etiketli bir metin derlemi oluşturulmuştur. İçeriğinde 35 adet farklı disiplinden, konu başına 200 adet belge bulunmaktadır. Bu çalışmada, metin derleminin toplanması ve içeriği sunulmuştur. Metin derleminde, TF-IDF ve Saklı Dirichlet Ataması'nda konu olasılıkları özniteliklerinin sınıflandırma başarımları karşılaştırılmıştır. Metin derlemi akademik amaçlı doğal dil işleme uygulamalarında kullanılabilmesi için açık kaynak olarak paylaşılmıştır.**

***Anahtar Kelimeler*—Derlem, Türkçe, Makale, Özetçe, Doğal Dil İşleme, NLP, Sınıflandırma, Saklı Dirichlet Ataması, Gizli Dirichlet Tahsisi, TF-IDF.**

***Abstract*—A labeled text corpus made up of Turkish papers' titles, abstracts and keywords is collected. The corpus includes 35 number of different disciplines, and 200 documents per subject. This study presents the text corpus' collection and content. The classification performance of TF-IDF and topic probabilities of Latent Dirichlet Allocation features are compared for the text corpus. The text corpus is shared as open source so that it could be used for natural language processing applications with academic purposes.**

***Keywords*—Corpus, Turkish, Paper, Abstract, Natural Language Processing, NLP, Classification, Latent Dirichlet Allocation, TF-IDF.**

I. GİRİŞ

Doğal dil işleme uygulamaları için üzerinde çalışılacak metin verisine ihtiyaç vardır. Türkçe dilinde, özellikle metin sınıflandırmada kullanılmak üzere açık kaynaklı, geniş kapsamlı ve sınıf etiketli bir derlemin eksikliği hissedilmektedir. Bu çalışma, bu eksikliği kapatmak için yapılmıştır ve belirtilen açılardan bilgimiz dahilinde literatürde ilk olacaktır.

Elektronik ortamda paylaşılmış, hem yazılı, hem sözlü olarak Türkiye Türkçesi dilini içeren derlemler mevcuttur. Türkçe Ulusal Derlemi, çeşitli konu alanlarını kapsayan yazılı ve konuşmaların yazıya aktarılmasıyla elde edilen sözlü veriden oluşmaktadır [9]. TS Corpus, metin verisini dilbilimsel ve biçimsel olarak iki ayrı düzlemde etiketlenmiş biçimde sunmaktadır [10]. ODTÜ Türkçe Derlemi ve ODTÜ-Sabancı Türkçe Ağaç Yapılı Derlemleri, biçimbirimsel ve sözdizimsel olarak işaretlenmiş derlemlerdir [11]. ODTÜ-MEDİD derlemi, çeşitli söylem özellikleriyle işaretlenmiştir [12]. Bahsedilen derlemlerin içerinde, metinlerin konularına göre sınıflandırma etiketlerini içeren bir derlem bulunmamaktadır.

Türkçe Etiketli Metin Derlemi oluşturulurken, hem fen bilimleri, hem sosyal bilimlerin farklı alanlarından, ve çeşitlilik yaratmak adına mümkün olduğunca farklı kaynaklardan Türkçe makalelerin başlık, özetçe ve anahtar kelimeleri derlenmiştir. Makalede, veritabanının içeriği, kaynakları, derlenme yöntemi, formatı, kelime, belge sayısı gibi istatistikler, ve örnek uygulama olarak derlemin bir kısmına uygulanmış sınıflandırma çalışmasının sonuçları paylaşılacaktır.

II. TÜRKÇE MAKALE DERLEMİ

Derlemin içeriğinde 35 konu ile işaretlenmiş, konu başına 200 belgeden, toplam 7000 adet belge bulunmaktadır. Her belgenin başlığı, özetçesi, (varsa) anahtar kelimeleri, kaynak ismi, ve etiketi bulunmaktadır. Makalelerin tümü yerine özetçe kısımlarının alınmasının sebebi, bu kısımların en yoğun bilgi içeren ve en fazla anahtar kelimeye sahip kısımlar olduğunun düşünülmesidir. Anahtar kelimeler ise, metin özetleme, ve anahtar kelime çıkarımı gibi alanlarda kullanılabileceği düşünülerek alınmıştır.

Derlem, farklı alanlarda Türkçe dergi ve konferansların bildiri kitaplarından derlenmiştir. Derlemde yer alan makale özetçelerinin tamamı elektronik ortamda herkese açık şekilde paylaşılan özetçelerdir. Ayrıca derlem içerisinde özetçelerin alındığı kaynak belirtilmektedir. Metinler .html ve .pdf formatındaki bildiri ve makalelerden kopyalanmıştır. Kopyalamadan kaynaklı yazım hataları, oluştuğu durumlarda düzeltilmiştir. Etiketleme için Tübitak ULAKBİM Ulusal Veri Tabanları (UVT)'nın kaynaklara göre [1] sınıflandırması temel alınmıştır.

İçerikteki sınıflar, ULAKBİM Ulusal Veri Tabanları'nın 5 ana kategorisinden de örnekler içermektedir.

Sosyal ve Beşeri Bilimler Veri Tabanı'ndan "Antropoloji", "Arkeoloji", "Coğrafya", "Dil Bilim", "Dini Araştırmalar", "Eğitim Bilimleri", "Ekonomi", "İşletme", "Felsefe", "İletişim", "Kütüphanecilik", "Siyasal Bilimler", "Sosyoloji",

978-1-4799-4874-1/14/$31.00

"Tarih", "Turizm", "Borsa-Bankacılık" olmak üzere 16 sınıf toplanmıştır.
Tıp Veri Tabanı'ndan alınan belgeler, "Harici Tıp", "Dahili Tıp", "Temel Tıp" bilimleri olmak üzere 3 sınıfa toplanmıştır. Ayrıca "Eczacılık" da ayrı bir konu olarak eklenmiştir.
Yaşam Bilimleri Veri Tabanı'ndan "Biyoloji", "Çevre Bilimleri", "Gıda Bilimleri" , "Hayvancılık", "Veteriner Hekimlik", "Spor Bilimleri" olmak üzere 6 sınıf toplanmıştır.
Mühendislik ve Temel Bilimler Veri Tabanı 'ndan "Sinyal İşleme", "Elektronik İletişim", "Endüstri Mühendisliği" , "İnşaat Mühendisliği", "Makine Mühendisliği", "Mimarlık", "Biyomedikal Mühendisliği", "Jeoloji Mühendisliği" olmak üzere 8 sınıf toplanmıştır. Hukuk Veri Tabanı'ndan "Hukuk" da ayrı bir konu olarak eklenmiştir.

Derlem oluşturulurken, farklı amaçlara yönelik olarak verinin istenen kısmının kolayca kullanılabilmesi için, literatürde sıkça görülen şekilde xml formatı kullanılmıştır [2] [3]. Derlemde, her sınıf etiketine ait belgeler tek bir .xml dosyasında toplanmıştır. Her bir .xml belgesi içeriğinde 200 adet belge vardır. Belgeler <makale> </makale> xml etiketlerinin arasına yerleştirilmiştir. Makale başlıkları <Başlık> </Başlık>, özetçeler <Özetçe> </Özetçe>, anahtar kelimeler <Anahtar> </Anahtar>, makaleye kaynak olan konferans veya derginin ismi <Kaynak></Kaynak> etiketleri arasına yerleştirilmiştir.
Örnek bir belge Şekil 1. deki gibi saklanmaktadır.

```
<makale>
<Etiket>Sınıf Etiketi</Etiket>
<Başlık>Makale Başlığı</Başlık>
<Özetçe>Bu özet, Türkçe Etiketli Metin
Derlemi formatına örnek olması için
oluşturulmuştur. Bütün alanlardaki veriler
birer örnektir. </Özetçe>
<Anahtar>Metin Derlemi, Türkçe</Anahtar>
<Kaynak>Metin Derlemi Konferansı 2014
Bildiriler Kitabı</Kaynak>
</makale>
```

Şekil 1: Örnek Belge

Derlem istatistikleri Tablo 1. 'deki gibidir.

İstatistik	Sayı
Toplam Sözcük	1.092.241
Konu Sayısı	35
Toplam Belge Sayısı	7000

Tablo I: İstatistik Tablosu

III. DERLEM ÜZERİNDE ÖRNEK ÇALIŞMA

Derlemin bir kısmından rastgele örneklenen 18 sınıf ile iki farklı öznitelik çıkarma yönteminin sınıflandırma başarımı test edilmiştir.
Kullanılan bu öznitelikler, Saklı Dirichlet Ataması (Latent Dirichlet Allocation, LDA) üzerinde sonsal konu olasılıkları, ve 1-gramlar için hesaplanan TF-IDF katsayılarıdır.
Saklı Dirichlet Ataması'na literatürde Gizli Dirichlet Tahsisi de denmektedir. Bu çalışmada SDA olarak kısaltılacaktır.
Sözcük 1-gramları için TF-IDF'lerin hesaplanmasının üzerine, gereksiz öznitelikleri elemek ve öznitelik uzayını küçültmek için bir Bilgi Kazancı (Information Gain) öznitelik seçimi yöntemi kullanılmıştır.
SDA konu olasılıklarının hesaplanmasında MALLET [5], 1-gramlar için TF-IDF hesaplanmasında ve özniteliklerin Bilgi Kazancı kullanılarak seçilmesinde Weka Veri Madenciliği kütüphaneleri kullanılmıştır.
Destek Vektör Makinaları kullanılarak sınıflandırmanın yapılmasında da WEKA Veri Madenciliği kütüphanesi kullanılmıştır. [4].

A. Weka Kütüphanesi ile TF-IDF Özniteliklerinin Hesaplanması ve Bilgi Kazancı Öznitelik Elemesi

Java ile yazılmış, GNU (General Public License) lisansına sahip otomatik öğrenme ve veri madenciliği kütüphanesi Weka [6], bu projede TF-IDF özniteliklerinin hesaplanmasında, TF-IDF özniteliklerinin bilgi kazancı (Information Gain) kullanılarak seçilmesinde kullanılmıştır.
WEKA, her veri kümesini özel bir formatta saklar. Bir veritabanını ifade eden belgeleri, belgelerin öznitelik vektörlerini, öznitelik vektörlerinin alabileceği değerleri ve belgelerin sınıf etiketlerini saklayan bu formata .arff ismi ve uzantısı verilmektedir. WEKA, .arff formatına uygun olarak bir metin dosyasında düzenlenmiş veriler üzerinde kolayca çalıştırılabilir.
WEKA, arff formatında olmayan, metin halinde, sınıflara göre klasörlenmiş haldeki veriyi weka.core.converters.TextDirectoryLoader fonksiyonunu kullanarak dizgi (string) olarak yükleyebilir. Bu çalışmada, makale veritabanı, TF-IDF özniteliklerinin hesaplanmasından önce Weka'ya bu fonksiyon kullanılarak yüklenmiştir.
Bu belgelerden basit TF-IDF özniteliklerinin çıkarılması için weka.filters.unsupervised.attribute.StringtoWordVector fonksiyonu kullanılmaktadır. Bu fonksiyon, dizgi halindeki verilerden, sözcüklerin dizgilerin içinde bulunmalarına göre hesaplanan öznitelikleri çıkarır. Yani, düzensiz veriler olan metin verisinden, matematiksel anlam içeren sözcük (veya n-gram) - frekans vektörlerini çıkarır. Bu fonksiyon çalıştırılırken, TF-IDF algoritmasının nasıl çalışacağını ayarlayan birçok parametresi bulunmaktadır. Bu parametrelerin anlamları ve bu çalışmada kullanılan değerleri aşağıda açıklanmaktadır:

- IDFTransform : True ("True" olarak seçildiğinde bir j belgesindeki i sözcüğünün frekansları , ters belge frekansı ile çarpılır.

$$f_{ij} = f_{ij} \times log\left(\frac{\text{toplam belge sayısı}}{\text{i sözcüğünü içeren toplam belge sayısı}}\right) \quad (1)$$

- TFTransform: True (True olarak seçildiğinde bir j belgesindeki i sözcüğünün frekansları , denklemdeki gibi dönüştürülür:

$$f_{ij} = log(1 + f_{ij}) \quad (2)$$

- attributeIndices: first-last (Hangi özniteliklerin kullanılıp hangilerinin kullanılmayacağını belirler. First-last ile bütün öznitelikler kullanılmaktadır.)
- doNotOperateOnPerClassBasis: True (Belirlenen maksimum sözcük sayısının ve minimum terim frekansının, her bir sınıf için değil, veri topağının tamamına uygulanacağını gösterir. Örneğin, maksimum sözcük sayısı 1000 olarak belirlendiyse ve bu değer "True" ise, tüm veri kümesinden en sık görülen toplam 1000 adet sözcük öznitelikler olarak belirlenmektedir.
- lowercaseTokens: True (Bütün sözcükler öznitelik sözlüğüne eklenmeden önce küçük harfe dönüştürülür.)
- minTermFreq: 1 (Bir sözcüğün öznitelik olarak sayılabilmesi için, veri kümesinde en az kaç defa geçmesi gerektiğini gösterir.)
- normalizeDocLength: normalize all data (sözcük frekansları belge boyutuna göre normalize edilir. Test ve eğitim topakları için ayrı seçenekler yapılabilmektedir.
- outputWordCounts: true ("True" seçildiğinde, öznitelikler, sözcüklerin belgelerde bulunmaları sayılarak hesaplanır. "False" seçildiğinde, sözcüklerin belgelerde bulunup bulunmamaları 0 veya 1 ile gösterili ve kaç defa bulundukları sayılmaz.
- stemmer: Nullstemmer (Sözcük köklerini bulabilen otomatik fonksiyonlar bulunmaktadır. Türkçe için düzgün çalışan bir algoritma WEKA'da bulunmadığından kök bulma (stemming) yapılmamıştır.)
- stopwords: Etkisiz kelimeler listesi özniteliklerin dışında tutulmaktadır. Elle derlenen ve Türkçe'deki en sık kullanılan kelimeleri içeren bir etkisiz kelimeler listesi kullanılarak, gereksiz ve sık rastlanan, öznitelik olarak sınıf belirleyici olmayacak kelimeler elenmiştir.
- tokenizer: alphabetic tokenizer (Metni sözcüklere bölme işleminin nasıl yapılacağını belirler. Alphabetictokenizer seçildiğinde, harfler harici karakterler içeren sözcükler gözardı edilir.
- useStoplist: true (Etkisiz sözcükler listesinin kullanılıp kullanılmadığını belirler.)
- wordstoKeep: 1000 (Toplamda tutulacak öznitelik sayısıdır. En yüksek frekanslı 1000 sözcük öznitelikler olarak belirlenir ve diğerleri elenerek öznitelik uzayı küçültülür.)

Özniteliklerin hesaplanmasının ardından, yaptığımız testlere göre metin sınıflandırmada iyi sonuç veren öznitelik eleme filtresi olarak weka.filters.supervised.attribute.AttributeSelection fonksiyonunun bilgi kazanımına (information gain) göre öznitelikleri seçen varyasyonu kullanılmıştır. Bu yöntemle, sınıflar için bilgi kazanımını enbüyüten öznitelikler seçilir ve diğerleri elenir. Bu adımı uygulama için kullanılan filtrenin parametre ayarlarında, değerlendirici ("evaluator") olarak "InfoGainAttributeEval", arama yöntemi ("search method") olarak sıralayıcı ("ranker") seçilmiş ve eşik değeri ("threshold") 0'a ayarlanmıştır.

$$\text{InfoGain}(\text{Sınıf}, \text{Öznitelik}) = H(\text{Sınıf}) - H(\text{Sınıf}|\text{Öznitelik}) \quad (3)$$

Bir X rastgele değişkeni için entropi (düzensizlik) H(X) ise şöyle hesaplanmaktadır:

$$H(X) = -\sum_{i=1}^{n} p(x_i) log p(x_i) \quad (4)$$

B. Mallet Kütüphanesi ile Saklı Dirichlet Ataması (Latent Dirichlet Allocation) Sonsal Konu Olasılıkları Özniteliklerinin Hesaplanması

Saklı Dirichlet Ataması (SDA) bir konu modelleme (topic modelling) yöntemidir. Konu modelleme yöntemleri, bir metin veritabanında hem bir belge içerisinde geçen sözcükleri, hem de farklı belgelerde sözcüklerin birlikte bulunduğu diğer sözcükleri inceleyerek, her belgedeki metni bir veya birden fazla konudan oluşacak şekilde sentezleyen modeli üretir. Saklı Dirichlet Dağıtımını belge sınıflandırmada kullanmak için bir yöntem, SDA sonucunda elde edilen her belgenin SDA algoritmasının saptadığı konular üzerine olasılık dağılımını bir öznitelik vektörü olarak kullanmaktır. Sayısı kullanıcı tarafından belirlenen konular sözcükler üzerinde bir olasılık dağılımı oldukları için, sözcüklere göre daha üst düzeyde bir anlam taşırlar. SDA yöntemiyle belgeler, veri kümesinden özütlenen ve dağarcıktaki tüm sözcüklere göre sayıları çok daha küçük olan, en belirleyici anahtar sözcüklerden oluşan birer alt küme olan konular ile ifade edilebilmektedir.
Bu çalışmada, belgelerin SDA sonsal konu olasılıkları öznitelik vektörlerinin hesaplanmasında, MALLET kütüphanesi kullanılmıştır. [5]
MALLET girdi olarak belgelerin sınıflara ait klasörlerin içinde bulunduğu bir ana veri dosyasını alır. Çıktı olarak aşağıdaki dosyaları vermektedir:

- Konu anahtar sözcükleri (keys) ve bunların frekansları
- Bir konu için olasılığı en yüksek anahtar sözcükler. İstenen sözcük sayısı girdi olarak belirtilmektedir.
- Her belgenin konular üzerinde dağılımını gösteren sonsal olasılıklar.

Mallet'ın hesapladığı konular üzerindeki olasılık dağılımı, bir öznitelik vektörü olarak kullanılmaktadır. Sınıflandırmada WEKA kullanıldığı için, MALLET'ın çıktısı olan sonsal olasılıklar bir algoritma ile .arff formatında ifade edilmektedir.

C. Weka Kütüphanesi ile Destek Vektör Makineleri Kullanarak Sınıflandırma

Sınıflandırma aşamasında WEKA'nın Destek Vektör Makineleri (Support Vector Machine) algoritması olan, weka.classifiers.functions.SMO kullanılmıştır [7]. Yüklemede gelen öntanımlı (default) parametreler değiştirilmemiştir. Bu fonksiyon, John Platt'ın sıralı minimal eniyileme algoritmasını [8] gerçeklemektedir. Çok sınıflı durumlarda, bire karşı hepsi (1 vs all) yöntemini kullanmaktadır. Parametreleri için WEKA'nın SMO (Sequential Minimal Optimization) için ön tanımlı değerleri kullanılmıştır. Bu öntanımlı değerler şöyledir:

- Kernel olarak, weka.classifiers.functions.supportVector.PolyKernel

(Polynom Kerneli)
kullanılmaktadır.

$$K(x, y) =< x, y > p \quad (5)$$

$$\text{veya } K(x, y) = (< x, y > +1)^p \quad (6)$$

- Karmaşıklık sabiti C=1'dir.
- Tolerans parametresi L= 1.0e-3 'dir.
- Yuvarlama hatası için epsilon P = 1.0e-12 ' dir.

Sınıflandırma başarımları hesaplanırken 10 katlı çapraz geçerleme yapılmıştır.

D. Sınıflandırma Başarımları

Makale veritabanında, sınıflandırma başarımını denemek için 18 sınıf seçilmiş ve sınıf başına 85 belge olacak şekilde örnekleme yapılmıştır.
Örneklenen sınıfların isimleri şöyledir: antropoloji, egitimbilimleri, arkeoloji, ekonometri, biyoloji, elektronikiletisim, borsa, endustrimuh, cevremuh, felsefe, dahilitip, hukuk, dilbilim, iletisim, diniarastirmalar, sporbilimleri, eczacilik, turizm.

WEKA'da TF-IDF özniteliklerinin Bilgi Kazancı ile elenmesinden sonra sınıflandırma başarımı % 63.268 olmuştur.
Karışma matrisi incelendiğinde, en sık karışan sınıfların bazıları beklenene uygundur:
- antropoloji, dilbilim, dini araştırmalar
- biyoloji, çevre mühendisliği
- borsa, ekonometri
- dahili tıp, eczacılık
- endüstri mühendisliği, ekonometri
- iletişim, dilbilim

Bazı sınıf karışmaları ise sürprizler içermektedir:
- hukuk, ekonometri, dilbilim
- felsefe, eczacılık

MALLET kullanarak, 18 sınıflı verinin SDA sonsal olasılık öznitelik vektörlerini çıkarmak için, 6, 18, 30, 50 ve 100 konulu SDA modelleri eğitilmiştir. Çıkan olasılık dağılımı öznitelikleri .arff formatına çevrilmiş ve WEKA SMO fonksiyonu ile sınıflandırılmıştır.

Konu Sayısı	Sınıflandırma Başarımı
6	38.5 %
18	51.6 %
30	66.1 %
50	62.6 %
100	67.4 %
200	65.8 %

Tablo II: Makale Veritabanında WEKA SDA Öznitelikleri ve Destek Vektör Makineleri Sınıflandırıcısı ile SDA'ya Atanan Sınıf Sayısına Göre Başarımın Değişimi

Başarımın 100 konu sayısına kadar arttığı, sonrasında tekrar azalmaya başladığı gözlenmiştir.

Bu karışmada antropoloji - dilbilim, borsa - ekonometri, dahili tıp - eczacılık, dilbilim - felsefe beklenene uygun olsa da, dahili tıp - ekonometri, biyoloji - ekonometri, antropoloji – ekonometri , dilbilim -ekonometri karışması beklenmeyen bir sonuçtur. Pek çok sınıfın beklenmeyen şekilde ekonometri ile karışmakta olduğu görülmektedir. Başarımı iyileştirmek için F5 (her sözcüğün ilk 5 karakterini alma) yöntemi de denenmiş sonuç değişmemiştir. Ayrıca Destek Vektör Makinaları parametreleri çok kapsamlı olmasa da bir ızgara üzerinde taranmış, iyileşme bulunmamıştır.

IV. DERLEME ERİŞİM

Derlem, akademik amaçlı doğal dil işleme uygulamalarında kullanılabilmesi için açık kaynak olarak bit.ly/trderlem adresinde paylaşılmıştır. Derleme erişim sağlayan kullanıcı, derlemi araştırma dışında hiçbir amaç için kullanmayacağını, hiçbir şekilde çoğaltmayacağını, dağıtmayacağını, ticari amaçlı ve gelir getirebilecek bir çalışmada kullanmayacağını ve bu derlemi kullanarak yaptığı araştırmanın sonuçlarını referans göstererek paylaşacağını taahhüt etmiş sayılmaktadır. Sorular için trderlem@gmail.com ile iletişime geçilebilir.

V. SONUÇ VE GELECEK ÇALIŞMALAR

Derlem, doğal dil işleme alanında etiketli Türkçe veritabanı eksiğini başarılı şekilde gidermektedir. Sınıfların çeşitli disiplinlerden alınması derlemde yeterli bir çeşitlilik yaratmaktadır. Birbirine oldukça yakın sayılabilecek disiplinlerin bulunması ise, metin sınıflandırmanın zorlu problemlerine yaklaşımları test etme açısından faydalı olacaktır.
Gelecekte, birincil hedef derlemin 50 sınıfı kapsayacak şekilde genişletilmesidir.
Uygulama olarak anahtar kelimeleri ve başlıkları özetçe metinlerinden tahmin etme, metin özeti çıkarma yönünde çalışmalar yapılacaktır.

TEŞEKKÜR

Bu çalışma, TÜBİTAK TEYDEB tarafından Yapı ve Kredi Bankası Arge Bölümünün 3120918 nolu Akıllı İçerik Analiz Aracı projesi kapsamında desteklenmiştir.

KAYNAKÇA

[1] http://uvt.ulakbim.gov.tr/dergiler/ Erişim tarihi 10.02.2014, 22:00

[2] N. Fuhr, M. Lalmas, A. Trotman, "The Wikipedia XML Corpus", Comparative Evaluation of XML Information Retrieval Systems, V. 4518, 2007

[3] T. Ohta, Y. Tateisi, J.-D. Kim, "The GENIA corpus: an annotated research abstract corpus in molecular biology domain", HLT '02, 2002

[4] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, I. H. Witten, "The WEKA data mining software: an update", SIGKDD Explor. Newsl. 11-1, 2009

[5] http://mallet.cs.umass.edu/ (Erişim: 06.03.2014 19:35)

[6] http://www.cs.waikato.ac.nz/ml/index.html (Erişim: 06.03.2014 19:30)

[7] http://weka.sourceforge.net/doc.dev/weka/classifiers/functions/SMO.html (Erişim: 23.03.2014 22:30)

[8] J. Platt: Fast Training of Support Vector Machines using Sequential Minimal Optimization. In B. Schoelkopf and C. Burges and A. Smola, editors, Advances in Kernel Methods - Support Vector Learning, 1998.

[9] http://www.tnc.org.tr/index.php/tr/ (Erişim 24.03.2014 01:18)

[10] http://tscorpus.com/tr (Erişim 24.03.2014 01:18)

[11] http://ii.metu.edu.tr/corpus (Erişim 24.03.2014 01:44)

[12] http://medid.ii.metu.edu.tr/ (Erişim 24.03.2014 01:46)