GU J Sci, Part C, 8(2): 372-382 (2020)

Gazi Üniversitesi

Fen Bilimleri Dergisi

PART C: TASARIM VE TEKNOLOJİ

http://dergipark.gov.tr/gujsc

Classification of News according to Age Groups Using NLP

Rabia KONTUK, Metin TURAN*

Istanbul Commerce University, Department of Computer Engineering, Maltepe/İSTANBUL

Article Info:

Research article
Received: 07/02/2020
Revision 17/04/2020
Accepted: 29/04/2020

Highlights

- *Tokenization*
- *Morphology*
- *Remove Stop Words*

Keywords

News Age Group Detection
Age Group Dictionary
Zemberek
Natural Language Processing

Graphical/Tabular Abstract

In this study, the classification of news texts according to the relevant age groups was achieved by natural language processing method. A dictionary is constructed in order to use for classification. During the creation of the dictionary, tokenization, morphology, and remove stop word operations is applied to the news, respectively.

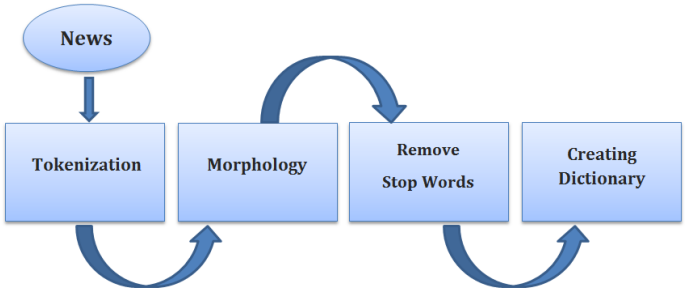


Figure A. Dictionary stages

Purpose: The purpose of this study is to help news sites' make a classification to prevent both neglecting the ethical elements of the news (sexuality, swearing, rape etc.) and being accessable by every age group. Especially in order to prevent the children from being harmed by the content published in moral and psychological terms, and not to be encouraging, the news appropriate for age groups should be readable.

Theory and Methods: Zemberek Library was used for Natural Language operations of Turkish news texts. Childhood, Youth and Adult age groups were determined by using Havighurst's Development Theory. A data set has been created using real Turkish web sites' news about the age groups. Data set divided into training and test parts. The training news was used to create a dictionary. Test news was tested applying the proposed model on the created dictionary.

Results: The developed dictionary (excluding verb) was applied to the test news and the highest success was determined in the Adult age group. The overall success of the dictionary was found to be 70% correct.

Conclusion: In this study, a dictionary is proposed to determine the age groups of the news. While creating the dictionary, Natural Language processing operations were carried out using the Zemberek Library on the data set containing the news. Then, the frequency of the term was calculated for each word in the news (except verb) and the words that were found meaningful in the relevant age groups; age group information is included in the dictionary with the number of times seen in the news. The dictionary developed was tested and 70 percent accuracy was achieved.

*Corresponding author, e-mail: mturan@ticaret.edu.tr

DOI: 10.29109/gujsc.686177

GU J Sci, Part C, 8(2): 372-382 (2020)

Gazi Üniversitesi

Fen Bilimleri Dergisi

PART C: TASARIM VE TEKNOLOJİ

http://dergipark.gov.tr/gujsc

NLP Kullanılarak Haberlerin Yaş Gruplarına Göre Sınıflandırılması

Rabia KONTUK Metin TURAN*

İstanbul Ticaret Üniversitesi, Bilgisayar Mühendisliği Bölümü, Maltepe/İSTANBUL

Makale Bilgisi

Araştırma makalesi
Başvuru: 07/02/2020
Düzeltme: 17/04/2020
Kabul: 29/04/2020

Anahtar Kelimeler

Haber Yaş Grubu Tespiti
Yaş Grubu Sözlüğü
Zemberek
Doğal Dil İşleme

Keywords

News Age Group Detection
Age Group Dictionary
Zemberek
Natural Language Processing

Öz

Bu çalışmada, Doğal Dil İşleme kullanılarak elektronik ortamlardaki haberlerin yaş gruplarına göre etiketlenmesi amaçlanmıştır. Haber sitelerinden toplanan haber veri setinin eğitim amaçlı seçilmiş olanları, NLP Zemberek Kütüphanesi kullanılarak Python dili ile işlenmiş, Havighurst'ün "Gelişim Kuramı" nın güncel duruma adapte edilmiş Çocukluk, Ergenlik ve Yetişkinlik yaş gruplarını temsil edebilecek kelime sözlüğü oluşturulmuştur (her kelimenin hangi yaş grubuna uygun olduğu). Daha sonra, bu sözlük kullanılarak haber veri setinin test amaçlı seçilmiş olanlarının sınıflarını belirlemek üzere bir sınıflandırıcı önerilmiştir. Testler sonucunda, geliştirilen sözlüğün 0.70 oranında doğru sınıfı tespit edebildiği görülmüştür.

Classification of News according to Age Groups Using NLP

Abstract

In this study, it is aimed to label the news in electronic media according to age groups by using Natural Language Processing. The selected ones for training in the news dataset collected from the news sites were processed in Python language using the NLP Zemberek Library, and a vocabulary dictionary that could represent Childhood, Adolescence and Adult age groups of Havighurst's Development Theory adapted to the current situation was created (which age group of each word as appropriate). A classifier was then proposed to determine the classes of the news dataset selected for testing using this dictionary. As a result of the tests, it was seen that the developed dictionary can detect the correct class with a success rate of 0.70.

1. GİRİŞ (INTRODUCTION)

Hayatın her alanında hızla gelişen teknoloji ve beraberinde getirdiği internet ağı çok kısa sürede her türlü bilgiye ulaşmaya olanak sağlamaktadır. Bilginin yaygın fakat kirli bir hal alması karşısında, bilginin erişebilmesinden çok doğru bilgiye ulaşma ihtiyacı önem kazanmıştır. Bilgi kirliliğinin ve sınırsız paylaşımların sanal dünyada gezinen kitlelere (özellikle de yetişkin olmayanlar için) bir tık uzaklıkta olması aslında bir sınırlandırılmanın gerekli olduğunu göstermektedir.

Her ne kadar insanlar kendilerini ve yakınlarını korumak üzere internet için servis sağlayıcılar tarafından veya doğrudan kullanıcı tarafından uygulanan kişisel filtreler olsa da, bunlar açıkça genellikle cinsel içerikli kaynaklar üzerinde yoğunlaşmıştır. Fakat internetin yaygın kullanımı beraberinde insanlara yeni alışkanlıklar da kazandırmıştır. Buna en güzel örnek, basılı gazetelerin artık yerini elektronik haber sitelerine bırakmış olmasıdır. Açıkçası bu durum ise, günlük yaşam içinde yer alan ve hiç de ahlaki olmayan veya şiddet içeren haberleri de kontrol altına alabilmenin ne kadar önemli olduğunu göstermektedir.

İnternetin haber üretimi ve sunumunu kolaylaştırıcı olanakları sayesinde, basılı gazetelere kıyasla çok daha fazla habere yer verilebilmektedir. Bununla birlikte, gerek habercilerin daha fazla ve anlık haber sunma kaygısı, gerekse yeterli mesleki bilgiye sahip olmamaları ve daha da önemlisi tıklanma sayısını

(popülerlik) artırma kaygısı haberin niteliksel ve etik unsurlarının ihmal edilmesine sebep olabilmektedir. Bu hatalar, söz konusu kullanıcının çocuk olduğu düşünüldüğünde çok daha önemli bir hal almaktadır [5].

Bu husus göz önüne alınarak, bu çalışmada doğal dil işleme ile elektronik haberlerin uygun yaş gruplarına göre sınıflandırılması istenmektedir. Yaş grupları belirlenirken Havighurst'ün "Gelişim Kuramı" temel alınmıştır (Tablo 1). Bu kuramda 6 kategoriye ayrılan yaş grupları, internet kullanım yaşı öngörüsüne göre (ilkokul çağı başlangıcı) yeniden düzenlenerek 3 kategoriye –çocukluk (6-13 yaş), ergenlik (13-18) ve yetişkinlik (18+) olmak üzere– indirgenmiştir.

Tablo 1. *Havighurst Gelişim Kuramı* [4]

Yaklaşık Yaş	Karakteristik Görevler
Erken Çocukluk (Doğum - 6 yaş)	Yürümeyi, katı yiyecek yemeyi, konuşmayı, tuvalet kullanmayı öğrenir; cinsiyet farklarının toplumsal doğrularını öğrenir. Bu evrenin sonuna doğru, daha kavramsal görevler beklenir, doğruyu yanlıştan ayırabilir, vicdan geliştirmeye başlar, okumayı öğrenmeye hazırlanır; işaretlerin (örneğin gülümseme) kelimelerin yerine geçebildiğini öğrenir.
Orta çocukluk (6-12/13 yaşlar)	Oyun oynamada gerekli fiziksel becerileri öğrenir, akranlarla birlikte olmayı öğrenir; uygun kadınsı ya da erkeksi toplumsal rolü öğrenir, okuma, yazma ve hesaplamada temel becerileri geliştirir; vicdan, ahlak ve değerler geliştirmeye devam eder, özerklik geliştirmeye başlar, temelde demokratik olan toplumsal tutumlar geliştirir.
Ergenlik (13-18 yaşlar)	Duygusal ve fiziksel olarak olgunlaşır. Her iki cinsten akranlarla olgun ilişkiler kurar, toplumsal olarak kabul edilen erkeksi ya da kadınsı toplumsal rolü öğrenir. Fiziğini kabul eder ve bedenini etkili biçimde kullanır, ana babadan ve diğer yetişkinlerden duygusal bağımsızlık kazanır, evlilik ve aile yaşamı için hazırlanır, bir mesleğe hazırlanır ve davranışlarını belirleyecek bir dizi değer ve ilke kazanır.
Erken Yetişkinlik (18-35 yaşlar)	Eş seçer, yakın bir ortakla yaşamayı öğrenir, yuva kurar, mesleğine başlar, vatandaşlık sorumluluklarını kabul eder, toplumsal ilişkiler kurar.
Orta Yaş (35-60 yaşlar)	Ergenlerin sorumlu kişiler olmasına yardım eder; toplumsal ve vatandaşlık sorumluluğu kazanır, mesleki doyum elde eder, orta yaşın fiziksel değişikliklerine uyum sağlar, yaşlanan ana babaya uyum sağlar.
İleri Olgunluk (60 yaşın ötesi)	Azalan fiziksel güce, eşin ölümüne, azalan ya da sabit kalan gelire uyum sağlar, akranlarıyla yakınlık kurar.

Doğal Dil İşleme alanında, sözlük oluşturma işlemleri üzerine günümüze kadar birçok çalışma yapılmıştır. Bu çalışmalar birçok dil için uygulansa da çoğunlukla İngilizce dili üzerine yapılmıştır. Literatürdeki sözlük çalışmalarını, Türkçe ve diğer diller olmak üzere iki grupta incelemek faydalı olacaktır.

Diğer diller için ilk çalışmalardan biri olan ve 1999 yılında ABD İngilizcesi için Silverman K., Anderson V., Bellegarda J., Lenzo K. ve Naik D. tarafından geliştirilen "Victoria" adlı sözlüğün tasarımı ve koleksiyonu tanıtılmıştır. Bu sözlük, Apple Computer'da konuşma sentezi araştırma ve geliştirmesini desteklemek amacı ile oluşturulmuştur. Çalışmada üretilen sonuçlarda ise, böyle bir sözlüğün oluşturulmasında gerçek dünya koşullarının çoğunu karakterize eden çarpıklıklar ve gürültü gibi etkenlerden dolayı daha düşük sentez kalitesine neden olacağı ve böyle bir konuşma sentezi sözlüğünün bir araya getirilmesinin çok büyük zorluklar yaratacağı kaydedilmiştir [7].

1999 yılındaki diğer bir önemli çalışma ise Riloff E. tarafından İngilizce dili için yapılan "AutoSlog" adlı sözlüktür. AutoSlog sözlüğünde amaçlanan, ilgili MUC-4 terörizm metinlerini ve bunlara ilişkin cevap anahtarları kullanılarak bilgi çıkarımı yapmaktır. Bu sözlük ile %98 oranında çok iyi bir sonuç elde edilmiştir [8].

2004 yılında Tsalidis Ch., Vagelatos A. ve Orphanos G. tarafından modern Yunanca dili için bir sözlük oluşturulmuştur. Çalışmanın yapılma amacı, modern Yunanca'nın morfolojik ve söz dizimsel seviyelerde işlenmesi için elektronik formda bir sözlüğünün olma zorunluluğudur. Çünkü modern Yunanca çok yönlü bir dil ve eski Yunanca'dan taşınan birçok karakteristik özelliği bulunan bir yazım sistemidir. Bu çalışma ile modern Yunanca'yı karakterize eden ve her türlü elektronik formdaki Doğal Dil İşlemede geçerli kılan özellikleri vurgulanmıştır [10].

2011 yılındaki bir diğer çalışmada, Këpuska V. Z. ve Rojanasthien P. İngilizce dili için konuşma grubu oluşturmuşlardır. Bu çalışmada, film ve dizilerin DVD'lerinden konuşma grubu oluşturmak için bir veri toplama sistemi sunulmuştur (konuşma tanıma, vurgu, perde, tonlama, duraksama analizi için). Çalışmanın avantajı olarak DVD'lerden sözlük üretiminin, geleneksel bir konuşma grubu elde etme yöntemine kıyasla önemli ölçüde düşük maliyetli olması verilmiştir. Buna ek olarak, verilerin toplanmasının ve sözlüğe dönüştürülmesinin de daha kısa süre alacağı kaydedilmiştir. Çalışmanın sonucunda sözlüğün konuşma grubu oluşturmak için yararlı ve çok yönlü olduğu gösterilmiştir [9].

Türkçe, her ne kadar geniş bir coğrafyada 60 milyon kişi tarafından anadili olarak konuşulan bir dil olsa da Türkçe üzerindeki Doğal Dil İşleme çalışmaları ancak son 15-20 yıl içinde hız kazanmıştır [6].

Bu çalışmaların günümüzde en başarılı örneklerinden biri, çoğu Türkçe NLP çalışmalarında kullanılan, Aktaş Y., Yılmaz İnce E. ve Çakır A.'nın çalışmalarında da yardımcı kütüphane olarak kullanılan 2007 yılında Akın A. A. ve Akın M. D. tarafından hazırlanan ve bu projede de faydalanılan "Zemberek" adlı kütüphanedir [1]. Bu çalışmaya makalenin ileriki bölümlerinde detaylandırılarak yer verilecektir.

2014 yılında Eryiğit G. tarafından geliştirilen İTÜ Türkçe Doğal Dil İşleme Yazılım Zinciri [21], Zemberek'in aksine açık kaynak kodlu olmayan NLP yazılımıdır. Bu yazılım Türkçe karakter dönüştürücü (asciifier/deasciifier), sözcük ayrıştırıcı/cümle bölücü (tokenizer/sentence splitter), yazım denetleyici (spell checker), biçim bilimsel çözümleyici/belirsizlik giderici (morphological analyzer/disambiguator), varlık ismi tanıma (named entity recognizer) ve bağımlılık çözümlemesi (dependency parser) gibi araçlardan oluşan bir platformdur. Bu platform, hem bir web ara yüzüne hem de bir uygulama programlama ara yüzüne (API) sahiptir. Böylece farklı seviyelerdeki kullanıcılar bu platformdan faydalanabilmektedir [20].

Türkçe ile ilgili çalışmalın en başarılı örneklerinden biri, 1985 yılında Princeton Üniversitesi Bilişsel Bilimler Laboratuvarı'nda Miller tarafından başlatılmış olan ve günümüzde en yaygın olarak kullanılan "WordNet" sözlüğüne [19], 2017 yılında Aktaş Y., Yılmaz İnce E. ve Çakır A., tarafından Türkçe dili için sağlanan bilişim sözlüğüdür. Çalışmada bilgisayar ağ terimlerinin ontolojik tabanlı oluşturulması işleminin otomatikleştirmesi, Türkçe dilinde kelimeler arası eş anlam yakın anlam gibi anlamsal bağlantılara sahip sözlüklerin uygun bir şekilde bir araya getirilmesi gerçekleştirilmiştir. Bu sözlükte 140.009 adet kelime bulunmaktadır. Çalışılan alanın sadece bilgisayar ağ terimleri ile sınırlandırılması çalışmanın başarısını oldukça arttırmıştır. Ontolojik bilgisayar ağları sözlüğüne veri girişi yapılan kelime sayısı arttıkça başarı artmış, anlaşılmazlık azalmıştır [11].

Bu araştırmanın motivasyonunu, haber sitelerinin anlık haber yayınlama kaygısı ile haberlerin niteliksel ve etik unsurlarını (cinsellik, küfür, argo, tecavüz, silah, şiddet) ihmal etmesi ve yayınlanan haberlerin bazı yaş grupları tarafından okunmasının sakıncası oluşturmuştur. Bu sebeple özellikle çocukların, gerek ahlaki ve gerekse psikolojik anlamda yayınlanan içeriklerden zarar görmemesi, örnek teşkil edip özendirilmemesi adına, yaş gruplarına uygun haberlerin okunabilir olması gereklidir. Bu alanda bildiğimiz kadarı ile dünyada uygulanan ilk çalışma olarak da önemlidir. Bu amaçla çalışmada mevcut probleme çözüm olmak üzere, haberlerin ilgili yaş gruplarına uygun olarak sınıflandırılmasına yönelik Türkçe haberler için bir model önerilmiş ve uygulanmıştır.

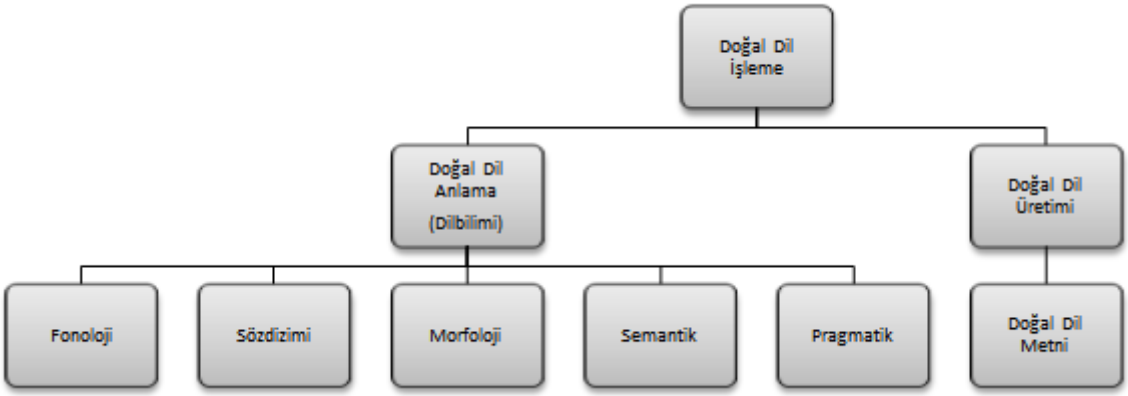
Bu çalışmanın ikinci bölümünde çalışmada kullanılan Doğal Dil İşleme, Terim Frekansı ve NLP Zemberek Kütüphanesi kavramlarından bahsedilmiştir. Üçüncü bölümde belirtilen probleme uygun olarak önerilen yöntemden ayrıntılı olarak bahsedilmiştir. Son olarak dördüncü bölümde ise önerilen yöntemle birlikte ulaşılan sonuçlara ve sonuçların değerlendirilmesine yer verilmiştir.

2. İLGİLİ KAVRAMLAR (RELATED CONCEPTS)

2.1. Doğal Dil İşleme (Natural Language Porcessing)

Doğal Dil İşleme (Natural Language Processing) son zamanlarda insan dilini hesaplamalı olarak temsil etmek ve analiz etmek için kullanılmaktadır [13]. Bu durum esasında, doğal dillerin bilgisayar veya elektronik cihazlarda temsil edilebilmesini sağlamaya yönelik çalışmalar olarak görülebilir.

Doğal Dil İşleme bilgisayarların doğal dilleri işleme sürecidir. Bu süreç insan bilgisayar etkileşimi biliminin altında veya hesaplamalı dil bilimi olarak da görülebilir. Bu dilin hesaplanması ve özelliklerinin tanınması durumudur [12]. Doğal Dil İşlemenin (DDİ) geniş sınıflandırılması aşağıdaki Şekil 1'de gösterilmektedir.



Şekil 1. *Doğal Dil İşlemenin geniş sınıflandırılması* [13]

Doğal Dil İşleme, üzerinde çalışılacak birçok konuyu da beraberinde getirmiştir. Bu konular aşağıda verilmiştir [14].

- Yazım yardımcı araçlarının geliştirilmesi
- Yazım yanlışlarının düzeltilmesi
- Bul ve değiştir
- Basılı bir metni okuma (optik olarak metin okuma) ve okuma yanlışlarını düzeltme
- Bir metnin özetini çıkarma
- Metnin içerdiği bilgiyi çıkarma
- Bilgiye erişim
- Metni anlama
- Bilgisayarla sesli etkileşim
- Bilgisayarın konuşması (metni seslendirme)
- Konuşmayı anlama (konuşmayı metne dönüştürme)
- Soru yanıt dizgeleri
- Yabancı dil okuma yardımcı araçları
- Yabancı dilde yazma yardımcı araçları
- Doğal diller arası çeviri

2.2. Terim Frekansı (Term Frequency)

Dokümanlarda en önemli özellik olan terimlerin (çoğunlukla kelime yerine kullanılır) önemi ve taşıdığı bilgi belirlenirken çeşitli metrikler kullanılır. Bunlardan en önemlisi Terim Frekansı (Term Frequency) metriğidir. Terim Frekansı (TF), metin madenciliğinde en yaygın kullanılan sayısal gösterimdir ve terim sıklığı anlamına gelmektedir. Kısaca, TF değeri bir terimin bir dokümanda görünme sayısını ifade eder [16].

Bu çalışmada veri setindeki terimlerin taşıdıkları bilgi miktarını belirlemek üzere Terim Frekansı (TF) metriği kullanılmıştır. Var olan haberlerin birinden çok bağımsız ve aykırı olmasından dolayı daha iyi sonuçlar elde etmek adına Terim Frekansı tercih edilmiştir.

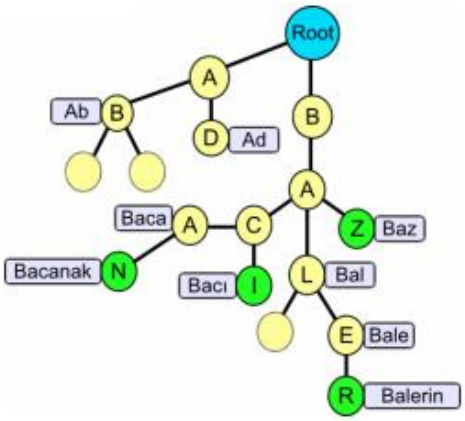
2.3. NLP Zemberek Kütüphanesi (NLP Zemberek Library)

Veri setindeki haberler üzerinde yapılacak dil bilimi işlemleri, terimlere ayrıştırma (Tokenization), terim yapı türü (Morphology) belirleme ve önemsiz terimleri (Stop Words) silme için, Ahmet Afşin Akın ve Mehmet Dündar Akın tarafından geliştirilmiş, Java tabanlı Türk dili Zemberek Kütüphanesi kullanılmıştır.

Zemberek Kütüphanesi ile çerçeve yazım denetimi, morfolojik ayrıştırma, kök bulma (Stemming), kelime oluşturma, kelime önerme, yalnızca ASCII karakterleri kullanılarak yazılan sözcükleri dönüştürme ve heceleri çıkarma gibi temel NLP işlemleri yapılabilmektedir. NLP Zemberek Kütüphanesi iki ana bölümden oluşmaktadır; dil yapısı bilgisi ve Doğal Dil İşleme işlemleri. Çekirdek kitaplık Doğal Dil İşlemeye özgü algoritmalar içerir ve dil uygulamalarına gerekli araçları sağlar. Çekirdek kütüphane özellikle Türk dilleri için tasarlanmış olsa da, herhangi bir özel dil uygulaması içermez. Bu esnekliği sağlamak için çeşitli yardımcı mekanizmalar ve soyutlamalar kullanılmaktadır. Her dil uygulaması, önceden tanımlanmış dilbilgisi gereksinimlerine uymaktan ve gerekli dil verilerini sağlamaktan sorumludur.

NLP Zemberek kütüphanesinin avantajlarından biri de açık kaynak kodlu olmasıdır [18].

Zemberek, kök sözlük tabanlı bir ayrıştırıcı kullandığından, bir kök bulma mekanizması gerektirir. Ayrıştırıcı, kök adaylarını bularak ayrıştırma işlemine başlar. Zembereğin kullandığı kök ağacı Şekil 2'de gösterilmektedir.



Şekil 2. *Doğrudan çevrimsel kelime grafiği ağacı* [1]

3. YÖNTEM (METHOD)

3.1. Veri Seti (Dataset)

Çalışma kapsamında üzerinde çalışılması gereken bir veri setine ihtiyaç duyulmuştur. Bu ihtiyaç doğrultusunda Türkiye'nin en çok okunan haber sitelerinden Hürriyet ve çocuklara, gençlere özel hazırlanan ilk ve tek haber sitesi Yumurtalı Ekmek adlı haber sitesi kullanılarak, toplamda 3925 haberi içeren bir veri seti Tablo 2'de sayısal dağılımı verildiği üzere oluşturulmuştur [2,3].

Tablo 2. *Haber sitelerinden alınan toplam haber sayıları*

Haber Sitesi	Toplam Haber Sayısı
Yumurtalı Ekmek	1313
Hürriyet	2612
Toplam	3925

Veri setinin, Havighurst'ün gelişim kuramının uyarlanmış yaş gruplarına göre etiketlendirilmesi, gönüllü Rehberlik ve Psikolojik Danışmanlar tarafından yapılmıştır. Veri setindeki haberlerin sınıflandırılmış halinde, Çocukluk (6-13 yaş) yaş grubu için 1313 adet, Ergenlik (13-18) yaş grubu için 1189 adet ve Yetişkinlik (18+) yaş grubu için 1423 adet haber bulunmaktadır. Bu veri setinin %70'i eğitim ve %30'u sınama amaçlı kullanılmak üzere iki kısma ayrılmıştır.

3.2. Veri Ön İşleme (Data Preprocessing)

Veri ön işleme, herhangi bir veri kümesindeki ilk işlemdir ve gerçek veri işleme başlamadan önce yapılan tüm işlemlerden (verinin doğruluğu, eksiklerinin giderilmesi, normalleştirme gibi işlemler) oluşur. Bu işlem ile veriler üzerinde herhangi bir analiz yapılmasını engelleyebilecek veri problemlerini çözme, verinin doğasını anlama, daha anlamlı bir veri analizi yapma ve belirli bir veri kümesinden daha anlamlı bilgiler elde etme gibi faydalar sağlayabilir [17].

Bu amaçla, veri ön işleme de öncelikle RSS ile toplanmış haberlerin, eğitim amaçlı olarak belirlenmiş olanlarının HTML etiketleri temizlenerek düz bir metin haline getirilmesi sağlanmıştır. Daha sonra haberlerde bulunan noktalama işareti, link, etiket, sembol, emoji, tarih, saat ve sayı gibi metin olmayan her türlü ifade NLP Zemberek kütüphanesi ile veri setinin içerisinden kaldırılmıştır. Metin olmayan tüm ifadelerin veri setinin içerisinden kaldırılmasından sonra var olan tüm kelimeler küçük harfe çevrilerek farklılıklar ortadan kaldırılmıştır (metin birimleri normalleştirilmiştir, aynı forma getirilmiştir).

Daha sonra haber metinleri, NLP Zemberek kütüphanesinin sözcüklere ayırma (tokenization) işlemi kullanılarak terimlere ayrılmıştır. Elde edilen her terim (kelime) için kök bulma işlemi uygulanmış ve NLP Zemberek kütüphanesinin biçim bilgisi (morphology) işlemi ile kelimenin en yalın haline ulaşılmıştır. Sözlük oluşturma işleminde kolaylık sağlaması açısından, ulaşılan her kök formunun tipi de (sıfat, zarf vb.) etiketlenmiştir.

Veri ön işlemenin son basamağı olarak, elde edilen terimlere NLP Zemberek kütüphanesi ile anlamsız kelimeleri silme (remove stop words) işlemi uygulanmış, tek başına anlam ifade etmeyen her terim kelime listesinden çıkarılmıştır.

3.3. Sözlük Oluşturulması (Creating Dictionary)

3.3.1. Terim Frekanslarının Hesaplanması (Calculating Term Frequency)

Bu aşamada sözlüğü anlamlı kelimelerle oluşturabilmek için öncelikle, her haber içinde bulunan kök formundaki terimlerin görülme sıklığı hesaplanmıştır (TF değeri). Aşağıdaki Tablo 3'teki örnekte rehberlik ve psikolojik danışman tarafından önceden Yetişkin grubu için uygun görülüp etiketlenen bir

haberin, ön işlemeden geçtikten sonra kalan terimlerinin TF değerleri görülmektedir. Bu işlem her haber için yapılmaktadır.

Tablo 3. *Haberlerin terim frekanslarının belirlenmesi*

Yaş Grubu	Haberin Terimleri	Terim Frekansları
Yetişkinlik	İlçe, uyuşturucu, esrar, uyuşturucu, komutan, ekip, ilçe, uyuşturucu, astsubay, üstçavuş, komutan uyuşturucu, madde, kılık, bahçe tabanca, tüfek, esrar	{'ilçe': 2, 'uyuşturucu': 4, 'esrar': 2, 'komutan': 2, 'ekip': 1, 'astsubay': 1, 'üstçavuş': 1, 'madde': 1, 'kılık': 1, 'bahçe': 1, 'tabanca': 1, 'tüfek': 1}

3.3.2. Eşik Değeri Bulma (Finding the Threshold Value)

Her haberin terim frekansları bulunduktan sonra haber bazında anlamlı kelimeleri seçmek üzere bir eşik değeri uygulanmıştır. Bu eşik değeri Formül 1'deki gibi ilgili haberdeki her kelimenin TF değerleri toplanarak haberde geçen toplam kelime sayısına bölünmesiyle elde edilmektedir. Böylelikle hesaplanan ortalama değerinden büyük olan TF değerli kelimeler seçilerek, sözlükteki uygun yaş grubunda ilk kez geçiyorsa haber sayısı değeri bir olarak (bir haberde görüldü anlamında) eklenir, daha önce geçmiş ise haber sayısı toplamı bir artırılarak mevcut kayıt güncellenir. Bu işlem tüm eğitim setindeki haberler için tek tek uygulanarak, sözlük oluşturmaktadır.

$$Ortalama = \frac{\sum Habere\ ait\ her\ kelimenin\ TF\ değeri}{Haberdeki\ toplam\ kelime\ sayısı} \quad (1)$$

Tablo 3'te örneği verilen haber için Formül 1 uygulandığında ortalama değer 1 olarak bulunur. 1'den büyük ve 1'e eşit geçme sıklığı olan terimler ilgili yaş grubu için Tablo 4'te görüldüğü gibi (her haber için) sözlüğe eklenmektedir.

Tablo 4. *Terimlerin Sözlüğe eklenmesi*

Yaş Grubu	Ortalamadan Büyük Olup Sözlüğe Eklenen Terimler	Haber Sayısı
Yetişkinlik	İlçe	1
Yetişkinlik	Uyuşturucu	1
Yetişkinlik	Esrar	1
Yetişkinlik	Komutan	1

3.3.3. Sözlüğün Elde Edilmesi (Dictionary)

Sözlük oluşturulurken aynı yaş grubu için başka bir haberde eşik değerinden büyük olup sözlüğe eklenecek terim aynı ise (daha önce sözlüğe eklenmiş ise) sözlükte tutulan haber sayısı değeri 1 artırılır. Böylece terimlerin yaş gruplarına uygun şekilde dağılımı görülmektedir. Tamamlanan sözlüğün küçük bir kısmı Tablo 5'te bulunmaktadır. Elde edilen sözlük ve özellik değerleri veri tabanında tutulmaktadır.

Tablo 5. *Oluşturulan sözlük*

Yaş Grubu	Ortalamadan Büyük Olup Sözlüğe Eklenen Terimler	Haber Sayısı
Yetişkinlik	İlçe	174
Yetişkinlik	Uyuşturucu	49
Yetişkinlik	Alkol	12
Çocukluk	İlçe	27
Çocukluk	Öğretmen	101
Ergenlik	Spor	82
Ergenlik	İlçe	30

3.4. Yaş Grubu Tahmini (Age Group Prediction)

Veri setinde sınama için ayrılmış her haberde önemli görülen (ortalama frekanstan yüksek) terimlerin hepsini değerlendirerek, bu haber için bir yaş grubu tahmininde bulunulması gereklidir. Bu amaçla haberin önemli terimleri göz önünde bulundurularak, 3 yaş grubu için (sözlükte yer alan geçtikleri haber sayısına uygun şekilde) ayrı ayrı puanlandırılır. Haberlerde geçme sıklığı değeri en yüksek olan yaş grubu bu haberin etiketi olarak belirlenir. Tablo 6'da görüldüğü gibi sınama kısmında var olan bir haberin terimleri yer almaktadır. Bu terimlerin oluşturulan sözlükte her yaş grubuna uygun haber sayısı değerleri toplanarak bir puan elde edilmiş olur. Tablo 6'da verilen haber örneği için Yetişkinlik yaş grubu seçilmektedir.

Sınama için ayrılan haberler için de önceden haberin uygun görülen bir (rehberlik ve psikolojik danışman tarafından manuel olarak belirlenen) yaş grubu bilgisi belirlenmiş durumdadır. Böylece sınanan her haber için olması gereken ve tahmin edilen olmak üzere iki yaş grubu bilgisi yer almaktadır. Sınama için kullanılan haber sayıları 3 yaş grubu içinde 272 tane (eşit) olacak şekilde alınmıştır.

Tablo 6. *Yaş grubu tahmini puanlama aşaması*

Yaş Grubu	kuzey	Terör	Örgüt	Sevk	Pikap	petrol	devriye	Puan
Çocukluk	35	0	12	2	0	7	0	56
Ergenlik	11	1	4	1	0	1	1	19
Yetişkinlik	122	415	370	94	1	31	31	**1064**

4. SONUÇ VE ÖNERİLER (CONCLUSIONS AND RECOMMENDATIONS)

Veri setinin sınanmaya ayrılmış kısmında bulunan haberler kullanılarak sözlük başarısı ölçülmüştür. Daha önceden belirtildiği gibi sınama verisinin önceden etiketlenen yaş grubu ve daha sonra puanlama sonucu belirlenen bir yaş grubu bulunmaktadır. Hata matrisi (confusion matrix) doldurulurken, "Gerçek" değeri için atanmış yaş grubu etiketi ve "Tahmin Edilen" değeri için ise puanlama sonucu belirlenen yaş grubu dikkate alınmıştır. Sonuçlara göre sözlüğün doğruluk (accuracy) değeri %70 bulunmuştur. Çalışmanın hata matrisi değerleri Tablo 7'de ve hata matrisinden üretilen sonuçlar ise Tablo 8'de görülmektedir.

Tablo 7. *Confusion Matrix*

		Tahmin Edilen		
		Çocukluk	**Ergenlik**	**Yetişkinlik**
Gerçek	**Çocukluk**	266	0	6
	Ergenlik	194	46	32
	Yetişkinlik	15	0	257

Tablo 8. *Çalışma sonucunda elde edilen Confusion Matrix'ten üretilen sonuçlar*

	Tutarlılık (Precision)	**Hatırlama (Recall)**	**F1-Değeri**
Çocukluk(6-13 yaş)	0.56	0.98	0.71
Ergenlik(13-18 yaş)	1.00	0.17	0.29
Yetişkinlik(18+ yaş)	0.87	0.94	0.91
Doğruluk			**0.70**
Ağırlıklı Ortalama	0.81	0.70	0.64

Tablo 8'de yer alan kavramlar şu şekilde açıklanmaktadır [15].

Tutarlılık (Precision): Tutarlılık çok sınıflı bir karışıklık matrisidir. Pozitif olarak tahmin edilen örneklerden, kaçının gerçekten pozitif olduğunu göstermektedir.

Hatırlama (Recall): Pozitif olan örneklerden, ne kadarının pozitif olarak tahmin edildiğini gösteren bir metriktir.

F1-Değeri: Tutarlılık (Precision) ve Hatırlama (Recall) değerlerinin harmonik ortalamasıdır.

Doğruluk (Accuracy): Doğru sınıflandırılan örnek sayısının toplam örnek sayısına oranı olarak hesaplanır.

Ağırlıklı Ortalama: Her üç grup için elde edilen bir metrik değerinin ortalamasıdır.

Hata matrisinden üretilen sonuçlara göre çalışmada üretilen sözlük Çocukluk (6-13 yaş) yaş grubu ve Yetişkinlik (18+) yaş grubu için daha anlamlı sonuçlar verirken Ergenlik (13-18) yaş grubu için düşük bir sonuç üretmiştir. Sonuçlar, ergenlik yaş grubu için üretilen kelimelerin, hem çocukluk hem de yetişkinlik kelimelerinden tam olarak ayrıştırılamadığını göstermektedir. Bu durum gelecek çalışmalarda ele alınması gereken önemli bir tespittir.

Gelecekte TF değerinin yanı sıra, anlamlı kelimelerin tespiti için farklı yöntemler (Helmholtz ilkesi veya Rake algoritması) kullanılabilir. Bunun yanı sıra sözlük tabanlı bir model yerine, derin öğrenme modeli kullanılarak başarım değerlerinin kıyaslanması mümkündür.

KAYNAKLAR (REFERENCES)

[1] Akın A.A., Akın M. D. Zemberek, an open source NLP framework for Turkic Languages, (2007).

[2] Hürriyet, 21 Eylül 2019, Erişim adresi: www.hurriyet.com

[3] Yumurtalı Ekmek, 21 Eylül 2019, Erişim adresi: www.yumurtaliekmek.com

[4] Çok F. Gelişim psikolojisi, kuramlar, yöntemler ve yaşamın ilk yılları (kısaltarak çeviri), Ankara Üniversitesi Eğitim Bilimleri Fakültesi Dergisi, 2 (26), s. 641-670, DOI: 10.1501/Egifak_0000000479, (1993).

[5] Fırat F. Çocuk odak'sız' habercilik: internet gazetelerinde çocuk içerikli haberlerin sunumu ve etik ihlaller, Gümüşhane Üniversitesi İletişim Fakültesi Elektronik Dergisi, 2(4), (2016).

[6] Oflazer K. Türkçe ve Doğal Dil İşleme (Turkish Natural Language Processing), Türkiye Bilişim Vakfı Bilgisayar Bilimleri ve Mühendisliği Dergisi, 2(5), (2012).

[7] Silverman K., Anderson V., Bellegarda J., Lenzo K. and Naik D. Design and Collection of a Corpus of Polyphones And Prosodic Contexts for Speech Synthesis Research and Development, (1999).

[8] Riloff E. Automatically Generating Extraction Patterns From Untagged Text, AAAI'96: Proceedings of the thirteenth national conference on Artificial intelligence, Vol. 2, s. 1044–1049, (1996).

[9] Këpuska V. Z., Rojanasthien P. Speech Corpus Generation from DVDs of Movies and TV Series, Journal of International Technology and Information Management: Vol. 20: Iss. 1, Article 4. (2011).

[10] Tsalidis Ch., Vagelatos A. and Orphanos G. An electronic dictionary as a basis for NLP tools: TheGreek case, ArXiv cs.CL/0408061 (2004).

[11] Aktaş Y., Yılmaz İnce E., Çakır A. Doğal Dil İşleme Kullanarak Bilgisayar Ağ Terimlerinin Wordnet Ontolojisinde Uyarlanması, SDÜ Teknik Bilimler Dergisi, (2017).

[12] Şeker S. E. Doğal Dil İşleme(Natural Language Processing), Yönetim Bilişim Sistemleri Ansiklopedisi, 4(2), (2015).

[13] Khurana D., Koli A., Khatter K., Singh S. Natural Language Processing: State of The Art, Current Trends and Challenges, ArXiv abs/1708.05148 (2017).

[14] Adalı E. Doğal Dil İşleme, Türkiye Bilişim Vakfı Bilgisayar Bilimleri ve Mühendisliği Dergisi, 2(5), (2012).

[15] Dev, 14 Nisan 2020, Erişim adresi: https://dev.to/overrideveloper/understanding-the-confusion-matrix-264i

[16] Binici K. Kütüphane ve Bilgi Biliminde Tema ve Yönelim, Hiper yayın, s. 41-84, İstanbul, (2018).

[17] İlhan U. Application Of K-NN and FPTC Based Text Categorization Algorithms to Turkish News Reports, (2001).

[18] GitHup, 20 Eylül 2019, Erişim adresi: https://github.com/ahmetaa/zemberek-nlp

[19] WordNet, 03 Şubat 2020, Erişim adresi: https://wordnet.princeton.edu/

[20] Uludoğan G., Özçelik R., Parlar S., Ercan G., Yıldız O. T. User Interfaces for Turkish Natural Language Processing, (2019).

[21] Eryiğit G. ITU Turkish NLP Web Service, s. 1-4, DOI: 10.3115/v1/E14-2001, (2014).