Araştırma Makalesi

KONUYA ÖZEL WEB KAYNAKLI İNGİLİZCE OTOMATİK SÖZLÜK OLUŞTURMA

Ahmet Toprak[†], **Metin Turan**[††]
[†] İstanbul Ticaret Üniversitesi, Fen Bilimleri Enstitüsü, İstanbul, Türkiye
[††] İstanbul Ticaret Üniversitesi, Mühendislik Fakültesi, İstanbul, Türkiye
Ahmetoprak190363@gmail.com, mturan@ticaret.edu.tr

ÖZET

Dil sözlüğü alanındaki çalışmalar, otomatik sözlük oluşturma konusuna yoğunlaşmış durumdadır. Bu makalede başlangıç olarak verilen bir İngilizce kelime referans alınarak, makale konusuna ait sözlüğün otomatik oluşturulması sağlanmıştır. İlk sözlük kelimesi, sisteme başlangıç olarak verilen bu İngilizce kelimeden elde edilmektedir. Sözlüğe eklenen ilk tohum kelime ile daha sonra Azure Web Cognitive Web Search sisteminde Web araması yapılmaktadır. Arama sonucu gelen ilk dokümanın, referans dokümanına da uygulandığı üzere Helmholtz Prensibi ile anlamlı kelimeleri bulunmaktadır. Bulunan bu anlamlı kelimeler arasından, anlam değeri en yüksek olan kelime sözlüğe eklenmektedir. Böylece Web'ten elde edilen bir dokümanın işlenmesi sonucu, o dokümana ait sadece en anlamlı kelime sözlüğe eklenmektedir. Daha sonra sözlüğe eklenen bu kelime, Web'te arama işlemine tabi tutulmaktadır. Web araması sonucu elde edilen dokümanlar tekrardan sisteme sokularak, bu dokümanlara ait anlamlı kelimelerin hesaplanması sağlanmaktadır. Web'te arama döngüsü bu şekilde tekrarlanmakta, nihai olarak sözlük için istenilen kelime sayısına ulaşıldığında ise sonlanmaktadır.

Sözlüğün başarımını ölçmek üzere, Hash benzerlik değeri hesaplanmıştır. Farklı konularda verilen referans kelimelerde yapılan sınamalarda ortalama % 40.46 oranında benzerliğe sahip sözlükler oluşturulabilmektedir.

Anahtar Kelimeler: Otomatik sözlük oluşturma, web araması, helmholtz prensibi, hash benzerliği

TOPIC SPECIAL WEB RESOURCING CREATING ENGLISH AUTOMATIC DICTIONARY

ABSTRACT

Studies in the area of language dictionary are focused on automatic dictionary creation. In this article, an English word is given as a reference and an automatic creation of the dictionary of the article subject is provided. The first dictionary word, is derived from this English word which is given as a starting point for the system. Web search is then performed in the Azure Web Cognitive Web Search system by using the first seed word added to the dictionary. The first document from the search result, has meaningful words with the Helmholtz Principle as applied to the reference document. Among the meaningful words found, the word with the highest value is added to the dictionary. Thus, as a result of processing a document obtained from the Web, the most meaningful word for that document is added to the dictionary only. Then, the word added to the dictionary is searched on the Web. The documents obtained as a result of web search are put into the system and the meaningful words of these documents are calculated. The search cycle on the web is repeated in this way and finally ends when the desired number of words for the dictionary is reached.

In order to measure the performance of the dictionary, Hash similarity value was calculated. Dictionaries with a similarity of 40.46% can be created in the tests performed on the reference words given in different subjects.

Keywords: Automatic dictionary creation, web search, helmholtz principle, hash similarity

Geliş/Received	**:**	**06.05.2019**
Gözden Geçirme/Revised	**:**	**07.05.2019**
Kabul/Accepted	**:**	**31.05.2019**

1. GİRİŞ

Teknolojinin gelişmesi hem olumlu, hem de olumsuz sonuçları beraberinde getirmiştir. Bilgi erişiminin kolaylaşması, zamandan tasarruf edilmesi gibi olumlu etkilerinin yanında, elde edilen bu bilgiler içerisinden ihtiyacımız olan anlamlı bilgiyi çıkarma güçlüğü gibi olumsuz etkileri de bulunmaktadır. Bu sorunun çözümü için problem tanımına istinaden, Web üzerinden elde edilen dokümanlara ait anlamlı kelimelerin belirlenmesi sağlanmıştır. Bu doküman anlamlı kelime belirleme çalışmalarında, Helmholtz Prensibi uygulanmıştır. Dokümanlara ait anlamlı kelimeler tespit edildikten sonra, bu anlamlı kelimeler için dil sözlüğü oluşturulması sağlanacaktır. Bu sözlükler Web arama motoru optimizasyonu, konuşma sentezi, otomatik dil sözlüğü araştırmaları için kullanılabilir. Bu amaçtan yola çıkarak sisteme başlangıç olarak verilen anlamlı kelimelerden, bu kelimelere uygun dokümanlar Web ortamından elde edilmiştir. Daha sonra bu dokümanları işleyerek bu dokümanlara ait anlamlı kelimelerin belirlenmesi sağlanmış, bizim için önemli olmayan kelimeleri eleyerek istenilen amaca yönelik sözlüğün oluşturulması sağlanmaya çalışılmıştır. Böylece bilgiye daha kolay bir şekilde ulaşmamız sağlanmaktadır.

Sözlük oluşturma işlemleri, elle, yarı otomatik ve otomatik olmak üzere 3 farklı şekilde yapılabilir. Otomatik oluşturulan sözlükler dışarıdan herhangi bir müdahale olmaksızın sürekli bir büyüme içerisinde genişleyecektir. Yarı otomatik ve elle yapılan sözlükler ise statik olmakla birlikte, dışarıdan müdahaleye ihtiyaç duyarlar. Ellen R. çalışmasında elle oluşturulan sözlüklerin eksikliklerine değinmiştir (Ellen,1993). Otomatik olarak oluşturulan sözlükler amaca yönelik olarak kolay bir şekilde değiştirilebilir ve genişletilebilir. Elle yapılan sözlüklerde bu durum o kadar kolay ve etkin bir şekilde gerçekleştirilemez.

Bu çalışmada, otomatik sözlük oluşturmak için bir iş akışı önerilmiş ve sonuçlar deneysel olarak sınanmıştır. Temel yaklaşım, kullanıcının amaçladığı sözlüğe referans olarak verdiği başlangıç kelimelerden, ilişkili dokümanların elde edilmesi ve elde edilen bu dokümanların işlenerek anlamlı kelimelerin sözlüğe eklenmesidir. Akış olarak, Web ortamında bulunan ilişkili dokümanların anlamlı kelimeleri tespit edilmektedir. Bu önemli kelimelerin bulunmasında Helmholtz Prensibi uygulanmıştır. Daha sonra bu kelimeler kullanılarak, Web ortamında ilişkili yeni dokümanlar bulunur ve sözlük için diğer kelimeler elde edilir. Diğer kelimeler, sözlüğe Helmholtz Prensibi sonucu elde edilen kelimelerin en yüksek değerli olanlarından seçilir. Böylece sözlüğün kontrollü bir şekilde büyümesi sağlanmış olur. Bu işlem sözlük kelime sayısına ulaşılana kadar tekrarlanır. Böylece sözlük dışarıdan herhangi bir müdahale olmadan otomatik ve sürekli olarak büyümeye devam edecektir. Amacımız, özelleşmiş problemler üzerinde yüksek başarı oranları elde edecek genelleştirilebilir otomatik sözlük oluşturma modeli geliştirmektir.

Makalenin ikinci bölümünde literatür taraması yapılmış ve benzer çalışmalardan ya da bu çalışmada kullanılan adımları kapsayan çalışmalara değinilmiştir. Üçüncü bölümde, sözlük oluşturma için kullanılan yöntem ve çalışmalar detaylı olarak açıklanmıştır. Dördüncü bölümde, üçüncü bölümde bahsedilen veri setleri detaylı açıklanmış, son bölümde ise sonuç ve gelecek çalışmalara değinilmiştir.

2. LİTERATÜR TARAMASI

Sözlük oluşturma ile ilgili geçmişten günümüze kadar birçok çalışma yapılmıştır. İlk oluşturulan sözlüklerin çoğu elle oluşturulmuş ve basit sözlükler iken, daha sonraki süreçlerde yarı otomatik ve otomatik teknikler geliştirilmiştir. Sonuçların bu tekniklere bağlı olarak iyi yönde değiştiği gözlenmektedir. Bu çalışmaların çoğunun amacı, yığın veri içerisinden istenilen bilgiyi elde etmektir.

Helmholtz Prensibi kullanılarak önemli kelime tespit çalışmaları yapılmıştır. Agnes Desolneux ve arkadaşları çalışmalarında, Helmholtz Prensibi'ne bağlı temel algı ilkesine göre, dijital görüntüdeki geometrik yapıları önceden bilinen bir bilgi olmadan hesaplamak için kenar algılama yöntemi adında bir teoriyi anlatmışlardır. Bu teori, bir görüntüdeki kenarları ve sınırları (kapalı kenarları) parametresiz bir yöntemle tanımlamayı ve hesaplamayı sağlamaktadır. Bu çalışmanın görüntü analizi çalışmalarında ara bir katman olarak kullanılabileceği savunulmaktadır (Desolneux, Moisan ve Morel, 2001).

Raphael Khoury ve Lei Shi'da yaptıkları çalışmalarında, Helmholtz Prensibi'ni uygulamışlardır. Yapılan çalışmada, yazılım bakım maliyetlerini kolaylaştırmak, önemli bilgileri büyük bir iz üzerinden etkin olarak tanımlayabilmek için bir yaklaşım geliştirmişlerdir. Bu yaklaşım, hata ayıklama, performans analizi, özellik geliştirme gibi büyük verilerin olduğu alanlarda, bu verilerin içinden anlamlı izlerin (bilgi) elde edilmesinde Gestalt teorisi ve Helmholtz Prensibi'ne uygulanmasını ele almaktadır (R. Khoury, L. Shi ve A. Hamou-Lhadj,2016).

Bir önceki bölümde de belirtildiği şekilde, sözlük oluşturma işlemi otomatik, yarı otomatik ve elle olmak üzere 3 farklı şekilde işletilmektedirler. Bu türlerin her biri için zaman içinde çalışmalar yapılmıştır. Bu çalışmalardan biri de, Riloff Ellen'in otomatik sözlük çalışmasıdır. Riloff Ellen (R. Ellen, 1993) çalışmasında, metinden bilgi çıkarmak için otomatik olarak alana özgü kavramlar sözlüğü olan "AutoSlog" adlı bir sistem geliştirmiştir. AutoSlog'a bir metin verildiğinde, AutoSlog istenen bilgileri bu metinden çıkararak bir dizi sözlük girişi oluşturur. AutoSlog'a verilen metinler istenen bilgileri temsil ediyorsa, AutoSlog tarafından oluşturulan sözlük, önemli ölçüde başarılı sonuçlar verecektir. AutoSlog sözlüğü ile 5 kişi-saatte, terör olaylarını içeren bir sözlük oluşturulmuştur. AutoSlog sözlüğü daha sonra iki yetenekli lisansüstü öğrenci tarafından yapılan ve yaklaşık 1500 kişi-saat çaba gerektiren el yapımı bir sözlükle karşılaştırılmıştır. İki sözlük, her biri 100 metin içeren iki test seti kullanılarak değerlendirilmiştir. Sonuç olarak, AutoSlog sözlüğü, el yapımı sözlüğün performansının %98'ini sağladığı görülmüştür.

Silverman ve arkadaşlarının 1999 yılında yaptıkları (K.E. Silverman, V. Anderson ve J.R Bellegarda,1999) otomatik sözlük oluşturma çalışmasında, Apple Computer'da konuşma sentezi araştırma ve geliştirmesini desteklemek için oluşturulan Victoria sözlüğünün tasarımı ve yapısı açıklanmaktadır. Victoria sözlüğü 5 ana bölümden oluşmaktadır. Bunlar polifon, prosodik bağlam, tekrar eden konuşma, fonksiyon kelime dizileri ve sürekli konuşmadır. Bu sözlük, her biri konuşma sentezinin belirli bir yönünü kapsayacak şekilde tasarlanmıştır. Victoria sözlüğü, genel olarak ABD İngilizcesi ile oluşturulmuştur. Victoria sözlüğünün amacı, konuşma üzerinden anlamsal metinlerin toplanmasıdır. Sözlük, Apple'ın gelecek nesil text-to-speech sistemi MacinTalk 4 için süre ve adım modellerinin istatistiksel tahmininde kullanılmaktadır.

Kepuska Veton Z. ve Rojanasthien (Këpuska ve Rojanasthien, 2011) çalışmalarında, elle sözlük oluşturma işlemi yapmışladır. Bu çalışma ile film, TV dizileri ve DVD'lerden konuşma sözlüğü oluşturmak için bir veri toplama sistemi oluşturmuşlardır. Bu DVD'lerden yapılan sözlük üretimi, geleneksel bir konuşma sözlük elde etme yöntemine kıyasla, daha düşük maliyetli bir çözüm sunmaktadır. Ek olarak, verilerin toplanması ve bir sözlüğe işlenmesinin daha kısa sürdüğü belirtilmiştir.

Vijay ve arkadaşları [Vijay ve ark., 2018] çalışmalarında, son 8 yılda çevrimiçi olarak yayınlanan tweetleri kullanarak Hintçe-İngilizce kod karışık sözlük oluşturdular. Bu sözlüğü oluşturmak için, öncelikle Twitter'ın gelişmiş arama seçeneğini kullanan Twitter Python API1 kullanılarak tweetler Twitter'dan alındı. Alınan tweetler, zaman damgası, URL, metin, kullanıcı, retweetler, cevaplar, tam ad, kimlik ve beğeniler gibi tüm bilgiler json formatına dönüştürüldü. Tüm gürültülü tweet'leri kaldırmak için kapsamlı bir yarı otomatik işlem gerçekleştirildi. Tüm bu adımlarından sonra 2866 kelimelik bir sözlük oluşturuldu ve sözlük kelimeleri mutluluk, hüzün, öfke, sürpriz, nefret ve çoklu duygu olarak sınıflandırıldı. Daha sonra çevrimiçi yayınlanan Tweet'ler için duygu analizi çalışması yapılmış ve % 58,2 oranında doğruluk tespiti elde edilmiştir.

Bir diğer otomatik sözlük oluşturma çalışması da, S. Vorapatratorn, A. Suchato ve P. Punyabukkana'nın yaptıkları (Vorapatratorn ve ark., 2012) çalışmadır. Bu çalışmada, özel fonetik dağılım kullanan otomatik sözlük oluşturma yöntemi açıklanmaktadır. Genellikle, sistem bir web tarayıcısı üzerinden, İnternet'ten sürekli metin indirerek verilerini seçer. Açgözlü algoritma(covetous algorithm), uygun kelimeleri çıkarmak için metinlere uygulanır, bu işlem uygun metin sözlüğü kuruluncaya kadar devam eder. Çalışmada elde edilen sonuçlar, internetten çekilen veri sayısının hedef fonetik dağılımını gerçekleştirebildiğini ve % 99,13 oranında telefon kapsama alanı oluşturduğunu göstermektedir. Bu metin sözlüğünün, daha sonra konuşma sözlüğünü verimli bir şekilde üretmek için kullanılabileceği belirtilmiştir.

Türkçe sözlük oluşturma ile ilgili çalışmalar da yapılmaya başlanmıştır. Bunlardan biri de Türkçe WordNet çalışmasıdır. Bu çalışmada (Aktaş ve ark., 2016) , özellikle Türkçe WordNet olmakla birlikte, detaylı WordNet literatür araştırması yapılmıştır. Uzun vadede ise bugüne kadar hazırlanmış ilişkili sözlüklerle birlikte en geniş Türkçe bilişim sözlüğü hazırlanmıştır. Böylelikle büyük bilişim projelerinin zemini olan ilişkisel bilişim terimleri sözlüğü diğer sözlüklere göre ilişki ve kelime sayısı çokluğu olarak en büyük sözlük olduğu belirtilmiştir.

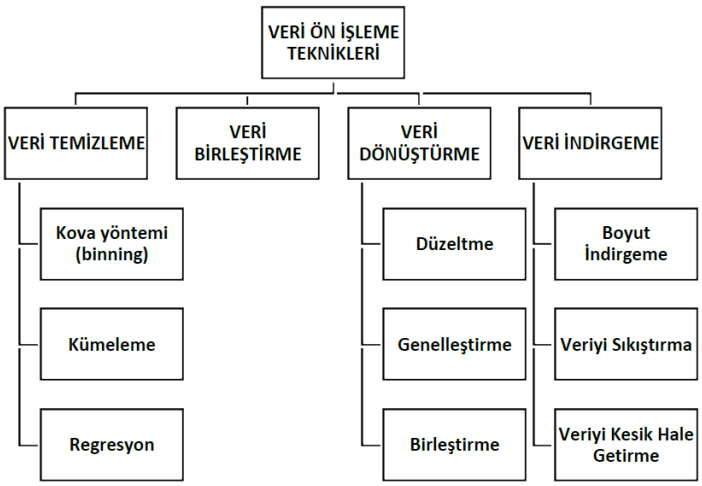
Sözlük oluşturma çalışmalarını incelediğimizde, çoğunda süreç ya elle, ya da yarı otomatik şekilde ilerlemektedir. Sözlüklerin genişlemesi için, sürekli dışarıdan bir müdahaleye ihtiyaç duyulmaktadır.

3. YÖNTEM

Başlangıç olarak sisteme verilen kelimeler Azure Cognitive Web Search işlemine tabi tutulduktan sonra, elde edilen dokümanlar ile ön işleme adımları gerçekleştirilir. Aşağıda sırasıyla uygulanan yöntemler açıklanmaktadır.

3.1. Ön İşlem

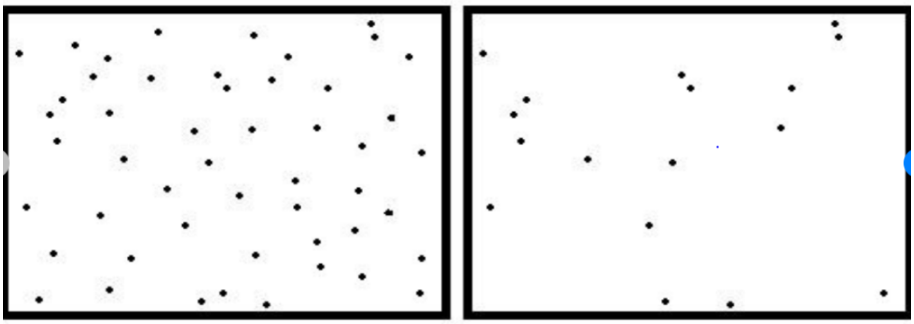
Ön işleme, metin madenciliği, doğal dil işleme (NLP) ve bilgi edinme (IR) konularında önemli bir görev ve kritik adımdır. Metin madenciliği alanında, yapılandırılmamış metin verilerinden önemsiz bilgiyi çıkarmak için veri ön işleme kullanılmaktadır (Vijayarani, 2015). Veri madenciliğinde analiz edilecek giriş verilerinin belirli bir formata sahip olması, ayrıca bozuk veya gereksiz verilerden temizlenmiş olması gerekmektedir. Metin madenciliğinin en büyük sorunu, işleyeceği veri kümesinin yapısal olmamasıdır (Omurca, S. Omurca ve ark., 2008). Genellikle doğal dil kullanılarak yazılmış dokümanlar üzerinde çalışılan metin madenciliği alanında ön işleme aşaması, veri temizlemenin yanında veriyi uygun formata getirme işlemini de gerçekleştirmektedir (R. Feldman, J. Sanger, 2006). Metin ön işlemede gürültü giderme(noise removel), sözlük normalize edilmesi(lexicon normalization), nesne standart oluşumu(object standarization) teknikleri kullanılır. D. Tanasa, B.Trousse (Tanasa ve Trousse,2004) ve V.Chitraa (Chitraa ve Davamani, 2010) çalışmalarında, ön işleme tekniklerine yer vermişlerdir. Kök bulma işlemi ile ilgili daha önceden yapılan çalışmalarda yoğun olarak kullanılması ve bu alanda İngilizce dili için geliştirilmiş algoritmalardan en yaygın olarak bilinen ve kullanımı kolay olan (C. Moral ve ark., 2014) Porter Stemmer kök bulma algoritması bu projede uygulanmıştır. Metin ön işleme süreçlerinde kullanılan adımlar Şekil 1'de gösterilmiştir.



Şekil 1. Metin ön işleme yöntemlerinin genel şeması.

3.2. Anlamlı Kelimelerin Belirlenmesi

Başlangıç olarak sisteme beslenen kelimeler Web aramasında kullanıldıktan sonra, Web'ten elde edilen dokümanlar ön işleme adımıyla tek düzene getirilir. Daha sonra tek bir forma dönüştürülen dokümanlara ait anlamlı kelimelerin tespit edilmesi gerekmektedir. Anlamlı kelimelerin tespit edilmesinde Helmholtz Prensibi kullanılmıştır. Helmholtz Prensibi, insan algı teorisine dayanmakta ve büyük veriler içerisinden anlamlı (anahtar) kelimelerin tespit edilmesinde kullanılmaktadır. İstatistiksel fizik ile doğrulanabilen bu yöntem, beklenti değeri hesaplamaları kullanılarak elde edilmektedir (Balinsky ve ark., 2011). Bu teori rastgele oluşturulmuş bir görüntü içerisinde rahatlıkla algılanabilen bir geometrik yapının şans eseri olmayacağının, bunun bir anlamı olduğunu söylemektedir. Bu durumu anlatmak için en açık örnek rastgele noktalardan oluşturulmuş Şekil 2'deki iki görüntüdür.



Şekil 2. İnsan algısında Helmholtz Prensibi (Balinsky ve ark., 2012).

Soldaki şekilde rastgele noktalara bakıldığında, insan algısı ile herhangi bir geometrik yapı algılanamıyor. Bu beklentisel bir durumdur ve görüntünün rastgele oluşturulduğu rahatlıkla söylenebilirken, sağdaki şekilde sanki görünmeyen bir çizgi üzerine konulmuş 5 nokta görülmektedir. Rastgele noktalarla oluşturulmuş bir görüntü içerisinde böyle geometrik bir yapının olma olasılığının beklenti değeri çok düşük olmasına karşın, bu olay gerçekleşmiş ve insan bunu rahatlıkla algılayabilmektedir. Öyle ise bu yapının anlamlı bir yapı olması gerekir. Çünkü beklenti değeri düşük bir olayın gerçekleşmesi, şans eseri olamayacak kadar düşük bir olma olasılığına sahiptir. Gerçekleşme olasılığı çok düşük olan bir olayın gerçekleşmiş olması bu olayın anlamlı yada önemli olduğunu göstermektedir (Balinsky ve ark., 2012).

Helmholtz Prensibi, metin madenciliğinde her bir kelime için; dokümanın paragrafında, kelimenin m kez geçmesinin olası olup olmadığının belirlenmesinde kullanılmıştır. Bu teoriye dayanan anlam değeri, aşağıdaki formüllerle hesaplanır (S. Ögtelik ve M. Turan, 2018).

$$YAS(k,P,D) = \binom{K}{m} \frac{1}{N^{m-1}} \tag{1}$$

$$Anlam(k,P,D) = -\frac{1}{m} \log YAS(k,P,D) \tag{2}$$

$$N = \frac{|D|}{|P|} \tag{3}$$

$$\log YAS(k,P,D) = \log\left(\binom{K}{m} \frac{1}{N^{m-1}}\right) \tag{4}$$

$$\binom{K}{m} = \frac{K!}{m!(K-m)!} \tag{5}$$

Helmholtz Prensibi'ne göre kullanılan bu formüllerde;

D: Dokümanın paragraflarını,
P: Paragrafta yer alan cümleler, anlamı taşımaktadır.

Çalışmada dokümanları paragraflarına ayırarak, paragraf tabanlı bir çalışma yaptığımız için bu formülü kullanırken D'yi dokuman, P'yi de dokümanda yer alan paragraf olarak ele aldık. Yaptığımız çalışma için formülde yer alan diğer terimlerin de açıklaması aşağıdaki gibidir:

k: İşlem yapılan kelime
P: Veri kümesinde yer alan paragraf sayısı
D: Veri kümesinde yer alan dokümanlar
m: Hesapladığımız kelimenin toplam kaç tane paragrafta geçtiği
K: Hesapladığımız kelimenin dokümanda toplam kaç kez geçtiğinin sayısı
N: Tüm veri kümesindeki toplam kelime sayısının, bir dokümandaki toplam kelime sayısına bölümünü temsil etmektedir.

Yukarıda belirtilen formüllerden yola çıkılarak anlam değeri (meaning value) ve Yanlış Alarm Sayısı (YAS) değeri hesaplamaları yapılmıştır.

3.3. Dökümanların Elde Edilmesi

Helmholtz Prensibi ile bulunan anlamlı kelimeler sözlüğe eklendikten sonra, bu kelimeler üzerinden Web araması yapılmaktadır. Projede Web üzerinde arama yöntemi olarak Azure Cognitive Web Search yöntemi kullanılmıştır. Azure Cognitive Web search, kullanıcıdan aldığı kelimeye uygun dokümanları istenen formatta hazırlar ve sisteme girdi olarak verir. Projede, sözlüğe eklenen kelimelerin "computer", "science", "software" olduğunu varsayarsak Web araması için kullanılacak kelime "computer science software" şeklinde kelimelerin boşluk karakteri ile birleştirilmiş hali olacaktır. Daha sonra sözlüğe eklenen tüm kelimeler için Web araması bu şekilde devam edecektir. Sözlük istenen kelime sayısına ulaşınca döngü sonlandırılacaktır.

3.4. Dökümanların Elde Edilmesi

Bu adımda, uygulama sonucunda elde edilen sözlüğün başlangıç olarak verilen kelimeler ile benzerliği tespit edilmeye çalışılmıştır. Başarı oranı tespit yöntemi olarak SimHash (Benzerlik Özeti) algoritması kullanılmıştır. SimHash algoritması, özellikle metin işlemenin yoğun olduğu, arama motoru gibi uygulamalarda dosyaların veya web sitelerinin birbirine olan benzerliğini bulmak için kullanılan bir algoritmadır. SimHash algoritması, iki dosyayı birer vektör olarak görür ve bu vektörler (yöney, vector) arasındaki cosinüs (cosine) bağlantısını bulmaya çalışır. SimHash algoritması, benzerlik tespiti amacıyla birçok farklı çalışmada kullanılmıştır. Jiang (Jiang, Qixia ve Sun, 2011) ve Pi (Pi ve ark., 2009) çalışmalarında, doküman benzerliğini elde etmek amacıyla SimHash algoritmasını kullanmışlardır.

4. VERİ KÜMELERİ

Sisteme başlangıç olarak verilen kelimeler referans alınarak, 2-gram olasılıklarına dayalı benzerlik modeli kullanan Python uygulaması yazılmıştır. Kullanıcıdan başlangıç olarak 1, 2, 3 adet şeklinde farklı sayıda kelime istenmiştir. Kullanıcıdan alınan bu kelimelerden, kelimenin konusuna göre otomatik sözlük oluşturulması sağlanmıştır. Bu sözlüklerin kelime sayısı azami 25, 50, 100 adet olacak şekilde sınırlandırılmış ve sonuçlar gözlenmeye çalışılmıştır. Aşağıda sırasıyla bu parametre değerleri için alınan sonuçlar paylaşılmıştır.

Aşağıda Tablo 1'de 25 kelimelik sözlüğün oluşması için kullanılan parametreler ve sözlük benzerlik oranı değeri verilmiştir.

Tablo 1. 25 kelimelik sözlük oluşturmak için kullanılan parametre değerleri ve sözlük benzerlik oranı.

Parametre Adı	Parametre Değeri	Parametre Değeri	Parametre Değeri
Sözlük Azami Kelime Sayısı	25	25	25
Başlangıç Kelime Sayısı	1	2	3
Sözlük Benzerlik Oranı %	**31,25**	**33,86**	**36,71**

Tablo 1'de, farklı başlangıç kelime sayılarıyla oluşturulan 25 kelimelik sözlüklerin, sözlük benzerlik oranları verilmiştir. Görüldüğü üzere, en yüksek sözlük benzerlik oranını % 36,71 oranıyla 3 başlangıç kelime ile beslenen sözlük elde etmiştir. Bu sözlüğün başarı oranının, diğer sözlüklerden daha yüksek olmasının nedeni, daha fazla kelime ile Web araması yapmış olmasıdır. Örnek vermek gerekirse, kullanıcı "Football" konusuyla ilgili bir sözlük oluşturmak istiyorsa, "sport football ronaldinho" kelimesi ile yapacağı Web araması, "sport" kelimesi ile yapacağı Web aramasından daha iyi sonuç verecektir. Çünkü kapsam daraltılarak sözlük oluşturmak istenen konuya ait dokümanlar elde edilmiş ve bu dokümanlar üzerinden anlamlı kelimeler tespit edilerek sözlük oluşturulması sağlanmıştır.

Aşağıda Tablo 2'de 50 kelimelik sözlüğün oluşması için kullanılan parametreler ve sözlük benzerlik oranı değeri verilmiştir.

Tablo 2. 50 kelimelik sözlük oluşturmak için kullanılan parametre değerleri ve sözlük benzerlik oranı

Parametre Adı	Parametre Değeri	Parametre Değeri	Parametre Değeri
Sözlük Azami Kelime Sayısı	50	50	50
Başlangıç Kelime Sayısı	1	2	3
Sözlük Benzerlik Oranı %	**30,83**	**36,02**	**40,46**

Tablo 3. Üç adet başlangıç kelimesi oluşan ile 50 kelimelik sözlük.

link	test	register	help	speaker
offer	argument	resale	type	way
structure	return	contractor	work	comment
park	month	construct	lot	route
way	year	country	mean	routine
communication	example	page	relationship	student
data	value	inform	look	question
shorelin	option	subscribe	washington	start
null	search	vacancy	suggest	cable
thank	prefer	application	century	insert

Tablo 2'de, farklı başlangıç kelime sayılarıyla oluşturulan 50 kelimelik sözlüklerin, sözlük benzerlik oranları verilmiştir. Sözlük benzerlik oranlarına bakıldığında, sonuçlar Tablo 1'de elde edilen sonuçlar ile benzerlik göstermektedir. Genel olarak, başlangıç kelime sayısı fazla verilen sözlüklerin daha yüksek sözlük benzerlik oranı elde ettiği görülmektedir. Yüksek sözlük benzerlik oranının elde edilmesindeki bir diğer faktör, başlangıç olarak verilen kelime setinin oluşturulmak istenen sözlüğün konusunu içeriyor olmasıdır.

Aşağıda Tablo 4'de 100 kelimelik sözlüğün oluşması için kullanılan parametreler ve sözlük benzerlik oranı değeri verilmiştir.

Tablo 4. 100 kelimelik sözlük oluşturmak için kullanılan parametre değerleri ve sözlük benzerlik oranı.

Parametre Adı	Parametre Değeri	Parametre Değeri	Parametre Değeri
Sözlük Azami Kelime Sayısı	100	100	100
Başlangıç Kelime Sayısı	1	2	3
Sözlük Benzerlik Oranı %	**29,34**	**33,62**	**37,98**

Tablo 4'de, farklı başlangıç kelime sayılarıyla oluşturulan 100 kelimelik sözlüklerin, sözlük benzerlik oranları verilmiştir. Oluşan sözlüklerin sözlük benzerlik oranları incelendiğinde, sonuçlar 25 ve 50 kelimelik sözlüklerin sözlük benzerlik oranından daha düşüktür. 100 kelimelik sözlüklerin, sözlük benzerlik oranının düşük olmasının nedeni, sözlüklerde belli bir noktadan sonra sapmaların oluşmasından kaynaklanmaktadır. Sapmaya neden olan faktör ise, elde edilmek istenen sözlüğün konusunun dışında bir kelimenin sözlüğe eklenmesidir. Bu kelime sözlüğe eklendiği takdirde, Web araması bu kelime üzerinden yapılacak ve oluşturulmak istenen sözlüğün konusu dışındaki dokümanlar Web üzerinden elde edilecektir. Bu dokümanların işlenip anlamlı kelimelerinin bulunması ile birlikte artık kapsam dışında olan kelimeler sözlüğe eklenmeye başlayacaktır.

5. SONUÇ

Bu çalışmada 3 farklı boyutta sözlüğün oluşması için sisteme bazı parametre değerleri uygulanmış ve ortaya çıkan sonuçlar gözlemlenmiştir. Oluşturulan sözlüklerin boyutları sırasıyla 25, 50 ve 100'dür. Bu 3 sözlüğün oluşması için sistem farklı sayılardaki başlangıç kelimeleri ile beslenmiştir. Sözlüklerin başarı oranlarını kıyasladığımızda, en iyi benzerlik oranını 3 adet başlangıç kelime ile beslenen 50 kelimelik sözlüğün elde ettiği görülmektedir. Bu sonucun elde edilmesindeki en önemli faktör, başlangıç kelime sayısının oluşturulmak istenen sözlüğün konusunu yansıtmasıdır. Buna bağlı olarak Web araması sonucu elde edilen dokümanlarda

oluşturulmak istenen sözlük konusuyla ilişkili olacaktır. Başlangıç olarak sisteme beslenen kelimeler birbiriyle ilişkili olmadığı takdirde, oluşan sözlük anlamsız kelimeler içerecektir. Bu noktalara dikkat edildiği takdirde, sözlük benzerlik oranı yüksek sözlükler elde edilebilir.

Benzer çalışmalar ile kıyaslandığında, yapılan çalışmanın ortalama sonuçlar, sözlük oluşturma ve büyüme oranı bakımından oldukça başarılı olduğu görülmektedir. Literatür taraması kısmında bahsedilen benzer [Vijay ve ark., 2018] çalışması ile kıyaslandığında, her 2 çalışmada sözlük oluşturma işleminde başarılı sonuç elde etmiştir. Ancak [Vijay ve ark., 2018] çalışmasından farklı olarak bu çalışmaya Web araması kısmı da dâhil edilmiştir. Bu çalışma, anlık ve güncel veriler üzerinden çalışması bakımından [Vijay ve ark., 2018] çalışmasından değerlidir.

Genel olarak çalışma sonuçlarına bakıldığında aşağıdaki çıkarımlar yapılabilmektedir.

- Başlangıç olarak verilen kelime seti, oluşturulmak istenen sözlük ile ilişkili olmalıdır. Eğer sistem, oluşturulmak istenen sözlük ile ilişkili olmayan kelime seti ile beslenirse, ortaya çıkan sözlük anlamsız kelimeler içerecektir.
- Başlangıç kelime seti sayısı arttırıldığında, bu duruma paralel olarak sözlük benzerlik oranı da artmaktadır.
- Anlamlı kelimeler bulunduktan sonra, bu anlamlı kelimeler içerisinden anlam değeri en yüksek olan kelime sözlüğe eklenmektedir. Sözlüğe eklenecek kelimenin, anlam değerlerinin belli bir değerin üzerinde olması için bir kısıt oluşturulursa, sözlük oluşturma hızı azalırken, ortaya çıkan sözlüğün başarısı artacaktır.
- Web araması sürecinde sürekli farklı veriler ile işlem yapılması için, sözlüğe eklenen son kelimeler ile sorgulama yapılması sözlüğün tekrarını önleyecektir. Bu nedenle sözlüğe son eklenen kelimelerin işaretlenmesi faydalı olacaktır.
- Sözlük içerisine ekleme yapılmadan önce, yukarıda 3.Bölüm'de belirtilen adımlar yapılmasına rağmen gözden kaçan bazı bozuk kelimeler olabileceği için, sözlük belli periyotlarda tekrardan ön işleme adımlarına tabi tutularak sürekli iyileşme sağlanabilir.

6. GELECEK ÇALIŞMALAR

Literatüre bundan sonraki süreçlerde katkı sağlayabileceği düşünülen aşağıdaki çalışmalar yapılabilir.

- Mevcut çalışmada, anahtar (anlamlı) kelimelerin tespit edilmesinde Helmholtz Prensibi kullanılmıştır. Bir başka çalışmada, anahtar kelimelerin tespit edilmesinde TF-IDF metrikleri kullanılıp sonuçlar gözlemlenebilir.
- Mevcut çalışmada, anahtar kelimeler tespit edildikten sonra, anlam değeri en yüksek kelime sözlüğe eklenmiştir. Çalışma yapısına benzerlik tespit yöntemlerinden biri (Wordnet gibi) dahil edilerek, sözlüğe eklenecek her kelimenin mevcut sözlük ile benzerlik oranına bakılabilir. Benzerlik oranı belli bir değerin üzerinde olan kelimeler sözlüğe eklenebilir. Böylece sözlük hem kontrollü büyümüş olur, hem de başarı oranı yüksek sözlükler elde edilir.

KAYNAKLAR

Aktaş, Y., İnce, E.Y., Çakır, A., & Kutlu, A. (2016.) Wordnet ve Bilgisayar Ağ Terimleri Sözlüğünün Oluşturulması. Akademik Bilişim 2016, Adnan Menderes Üniversitesi.

Balinsky, H., Balinsky, A., & Simske, S.J. (2011). Document sentences as a small world. 2011 IEEE International Conference on Systems, Man, and Cybernetics, 2583-2588.

Chitraa, V., & Davamani, A.S. (2010). A Survey on Preprocessing Methods for Web Usage Data. CoRR, abs/1004.1257.

Dadachev, B., Balinsky, A., Balinsky, H., & Simske, S.J. (2012). On the Helmholtz Principle for Data Mining. 2012 Third International Conference on Emerging Security Technologies, 99-102.

Desolneux, A., Moisan, L., & Morel, J. (2001). Edge Detection by Helmholtz Principle. Journal of Mathematical Imaging and Vision, 14, 271-284.

Feldman, R., & Sanger, J. (2006). The Text Mining Handbook - Advanced Approaches in Analyzing Unstructured Data.

Jiang, Q., & Sun, M. (2011). Semi-Supervised SimHash for Efficient Document Similarity Search. ACL, 93-101.

Kepuska, V., & Rojanasthien, P. (2011). Speech Corpus Generation from DVDs of Movies and TV Series.

Khoury, R., Shi, L., & Hamou-Lhadj, A. (2016). Key Elements Extraction and Traces Comprehension Using Gestalt Theory and the Helmholtz Principle. 2016 IEEE International Conference on Software Maintenance and Evolution (ICSME), 478-482.

Moral, C., Jiménez, A.D., Imbert, R., & Ramírez, J. (2014). A survey of stemming algorithms in information retrieval. Inf. Res., 19.

Omurca, S., Duru, N., Karagöz, Ş., & Sağır, M. (2008). Mühendislik & Bilgisayar, Fakültesi & Bölümü, Mühendisliği & Üniversitesi, Kocaeli. Metin Madenciliği ile Soru Cevaplama Sistemi.

Ögtelik, S., Turan, M. (2018), İngilizce Dokümanlarda Tema ve Alt Kavramlar Tespit Modeli, Düzce Üniversitesi Bilim ve Teknoloji Dergisi, 6(4), 754-764

Pi, B., Fu, S., Wang, W., & Han, S. (2009). SimHash-based Effective and Efficient Detecting of Near-Duplicate Short Messages.

Riloff, E. (1993). Automatically Constructing a Dictionary for Information Extraction Tasks. AAAI, 811–816.

Silverman, K.E., Anderson, V., Bellegarda, J.R., Lenzo, K.A., & Naik, D. (1999). Design And Collection Of a Corpus Of Polyphones and Prosodic Contexts For Speech Synthesis research and Development.

Tanasa, D., & Trousse, B. (2004). Advanced data preprocessing for intersites Web usage mining. IEEE Intelligent Systems, 19, 59-65.

Vijay, D., Bohra, A., Singh, V., Akhtar, S.S., & Shrivastava, M. (2018). Corpus Creation and Emotion Prediction for Hindi-English Code-Mixed Social Media Text. NAACL-HLT.

Vijayarani, S., Ilamathi, M., & Nithya, M. (2015). Preprocessing Techniques for Text Mining-An Overview Dr.

Vorapatratorn, S., Suchato, A., & Punyabukkana, P. (2012). Automatic online text selection for constructing text corpus with custom phonetic distribution. 2012 Ninth International Conference on Computer Science and Software Engineering (JCSSE), 6-11.