Avrupa Bilim ve Teknoloji Dergisi
Özel Sayı, S. 473-479, Eylül 2020

Araştırma Makalesi

www.ejosat.com ISSN:2148-2683

European Journal of Science and Technology
Special Issue, pp. 473-479, September 2020

Research Article

Kavramlar Arası WordNet Tabanlı Anlamsal Benzerlik Değerlerinin Farklı Metriklerle Değerlendirilmesi

Mustafa Özgür Cingiz[1*]

[1] Bursa Teknik Üniversitesi, Mühendislik ve Doğa Bilimleri Fakültesi, Bilgisayar Mühendisliği Bölümü, Bursa, Türkiye (ORCID: 0000-0003-4469-1440)

(1st International Conference on Computer, Electrical and Electronic Sciences ICCEES 2020 – 8-10 Ekim 2020)

(**DOI:** 10.31590/ejosat.819599)

ATIF/REFERENCE: Cingiz, M. Ö. (2020). Kavramlar Arası WordNet Tabanlı Anlamsal Benzerlik Değerlerinin Farklı Metriklerle Değerlendirilmesi. *Avrupa Bilim ve Teknoloji Dergisi*, (Özel Sayı), 473-479.

Öz

Kelimelerin anlam belirsizliği giderilmesi metin madenciliği, bilgi erişimi, doğal dil işleme gibi alanlarda yüksek doğruluklu başarı elde edilmesi için önemli bir adımdır. Kelimelerin bağlam içerisinde yer alan doğru anlamı belirlemek için sözlük tabanlı yaklaşımlar, eğiticili- eğiticisiz öğrenmede kullanılan etiketli-etiketsiz külliyatlar, kelime gömme gibi yeni yaklaşımlar sıklıkla kullanılmaktadır. Çalışmamız kapsamında ekonomi, teknoloji ve spor kategorilerine ait RSS haberleri haber sağlayıcılarından elde edilmiştir. Çalışma kapsamında RSS haber beslemelerindeki kelimeler kategorilere göre terim frekansı- ters doküman frekansı (tf-idf) ağırlandırması gerçekleştirilmiştir. Kelimeler arasındaki anlamsal benzerliklerin belirlenmesi için elle etiketlenmiş hiyerarşik çizge tabanlı sözlük olan WordNet tabanlı yaklaşımlar kullanılmıştır. İlk adımda tf-idf ağırlıklarına göre belirlenen kelimeler WordNet tabanlı Wu-Palmer, Lin ve Jiang – Conrath anlamsal benzerlik yaklaşımlarına göre tekrar sıralanmıştır. Aynı kategoride yer alan tf-idf değeri en yüksek elli kelimenin Kategorik Anlamsal İlişki Değeri (KAİD) hesaplanarak kelimelerin kategorilere ait anlamsal ilişki değerleri belirlenmiş. En yüksek KAİD değerine sahip 3, 5, 10 ve 20 kelime tüm kategoriler için çıkartılmıştır. Elde edilen kelimeler elle etiketlenmiş ve tf-idf ağırlıkları kullanılarak sıralanmış kelimelerle karşılaştırılmıştır. Karşılaştırma sonuçlarına göre iki katmanlı eleme ile anlamsal ilişkileri çıkarılan kelimeler ile insan tarafından belirlenen kelimelerin benzerlik oranının yüksek olduğu sonucu elde edilmiştir. WordNet tabanlı yöntemlerle elde edilen ve sıralanan kelimeler aynı zamanda tf-idf ağırlıklandırmasıyla elde edilen ve sıralanan kelimelerle de karşılaştırılmıştır. Sonuçlara göre ağırlıklandırma ile sıralanan kelimelerde örtüşme oranı insan algısıyla elde edilen kelimelerden daha düşük çıkmıştır. İki katmanlı değerlendirme ile oluşturulan kelimelerin anlamsal ilişki değerleri kategori uzayında görselleştirilerek anlamsal ilişki değerlerinin başarısı değerlendirilmiştir. İleriki çalışmalarda iki katmanlı değerlendirmeyle elde edilen kelimeler bilgi edinimi, metin özetleme, metin sınıflandırma alanında kullanılması hedeflenmektedir.

Anahtar Kelimeler: Anlamsal benzerlik, En kısa yol ölçütleri, Derinlik ölçütleri, WordNet, Metin madenciliği, Bilgi sistemleri ve uygulamaları

Evaluation of WordNet Based Semantic Similarity Values Between Concepts with Different Metrics

Abstract

Word sense disambiguation is an important step in text mining, information retrieval, natural language processing to obtain more accurate results. Dictionary- and knowledge-based, supervised, unsupervised and word embedding methods are used to discover the correct sense of words in the context. We retrieve RSS feeds ,whose categories are economy, technology and sport, to utilize in our study. After data retrieval, we used data preprocessing steps of text mining and we applied term frequency- inverse document

[*] Mustafa Özgür CİNGİZ: Bursa Teknik Üniversitesi, Mühendislik ve Doğa Bilimleri Fakültesi, Bilgisayar Mühendisliği Bölümü, Bursa, Türkiye, ORCID: 0000-0003-4469-1440, mustafa.cingiz@btu.edu.tr

frequency(tf-idf) for term weighting. WordNet is a large lexical database in which sense of words are kept in hierarchical network. In the first step, the words determined according to tf-idf weights were ranked according to the WordNet based semantic similarity measures Wu-Palmer, Lin and Jiang - Conrath. We used the top fifty ranked words ,which are obtained from tf-idf scores, to calculate Categorical Semantic Relationship Value (CSRV) of each word for each category. We determined the top 3, 5, 10 and 20 words due to CSRV for each category. Semantic ordered words are compared with tf-idf weighting based words and hand-labeled words which are determined according to semantic relationship by humans. The similarity rate is high between words are determined by two tier semantic structure based words and human labeled words. This similarity rate is lower between words are determined by two tier semantic structure based words and words which are ordered by tf-idf values. We also visualize the semantic similarity values in class dimension space to evaluate the success of the system. We intend to use two tier semantic structure in information retrieval, text summarization and text classification projects as future works.

Keywords: Semantic similarity, Shortest path measures,· Information content measures, WordNet, Text mining, Information Systems and Applications

1. Giriş

Son yıllarda sosyal ağlar üzerinden üretilen yazılı, görsel veriler ve basılı yayınların dijital ortama aktarımıyla birlikte internet üzerinde yer alan verilerin büyüklüğü üssel olarak artmıştır. İnternet kullanıcıları tarafından oluşturulan ve değerli içerik yoğunluğu düşük olan bu verilerden anlamlı bilgi çıkarımı önem kazanmıştır. Araştırmacılar bilgi erişimi, metin madenciliği, doğal dil işleme gibi çalışma alanlarında yer alan yaklaşımlar birlikte kullanılarak bilgi keşfinin yüksek doğrulukla yapılmasını hedeflemektedir.

İnsanlar metinde veya dinlediği içerekte yer alan kavramları hafızasında yer alan kavramlarla ilişkilendirebilme yetisine sahiptir. İnternette yer alan içeriklerin yüksek doğruluklu işlenmesi için de kavramlar arasındaki ilişkilerin bilinmesi gerekir. Metinlerde yer alan kelimelerin birden fazla anlama gelmesi kategorik değerlendirmede anlam belirsizliklerine yol açmaktadır. İngilizce içeriklerde "interest" kelimesi ilgi, alaka anlamında kullanıldığı gibi faiz anlamında da kullanılabilmektedir. Kişilerin günlük yaşamında hobilerini, ilgi alanlarını ifade ederken oluşturdukları içeriklerde "interest" kelimesi ilgi, alaka anlamında kullanılırken ekonomi, finans haberlerinde yer alan "interest" kelimesinin faiz ile ilgili olma olasılığı daha yüksektir.

Literatürde kelimelerin anlam belirsizliğini gidermek için sözlük gibi sözcüksel anlamları kullanan yaklaşımlar (Chen et al., 2005; Dang et al., 2002) sözcüklerin anlamlarını içeren etiketli külliyatlar kullanan yaklaşımlar (Mihalcea, 2007), sözcüklerin bağlamdaki yerlerini ve ilişkilerini kullanan eğiticisiz yaklaşımlar (Seo et al., 2004) ve yarı eğiticili yaklaşımlar (Pham et al., 2005) kullanılmaktadır. Son yıllarda kelime gömme (word embeddings) yaklaşımlarının kullanımıyla kelimelerin anlam belirsizliğinin gideriminde başarılı sonuçlar alınmaktadır (Simov et al., 2002). Kelimelerin anlam belirsizliğini gidermede kullanılan yaklaşımlar genel olarak sözcüklerin aynı bağlamda yer alan diğer kelimelerle olan ilişkilerini incelemekte ve kelimelerin bağlam içerisindeki gerçek anlamını belirlemeye çalışmaktadır.

Kelimelerde anlam belirsizliğinin giderimi için kullanılan bir diğer kaynak ise WordNet'tir. WordNet (Miller, 1998) , İngilizce için oluşturulmuş hiyerarşik sözcüksel bir veri tabanıdır. Bu elektronik sözlükte isimler, fiiller, sıfatlar ve zarflar kavramsal eşanlamlı kümeler (synset) ile gruplandırılmıştır. WordNet'in son sürümü olan 3.1 versiyonunda 155,287 kelime 117,659 eşanlamlı kümede yer almaktadır. WordNet'de kelimelerin farklı tüm anlamları(glosses) verilmekle birlikte her bir anlam(sense) farklı bir eşanlamlı kümede (synset) yer alabilmektedir. Eşanlamlı kümeler aynı insan zihninde olduğu gibi birbirleriyle anlamsal bağlar ile ilişkilendirilmektedir. Çalışmamızda kullandığımız isimler anlamsal olarak üst kavram (hypernym: elma-meyve), alt kavram (hyponym:meyve-karpuz), sıralı terimler (coordinate term : köpek- tilki), bölümün bütünü (holonym :vagon-tren) ve bütünün üyesi (meronym: vites- araba) gibi ilişkilerle birbirleriyle ilişkilendirilebilir. Fiiler ise üst kavram (hypernym: uçmak fiilini üst kavramı seyahat etmek), bir fiili farklı şekil yapılması (troponym: kekelemek), gereklilik (entailment: horlamak için uyuma gerekliliği) gibi ilişkilerle birbirine bağlıdır. Bu ilişkileri kullanarak birbiriyle ilişkili olan kelimeler, kavramlar insan algısında olduğu gibi işaretlenebilmektedir.

WordNet hiyerarşisinde iki kavramın birbirlerine yakınlığı hiyerarşide yer aldıkların yere göre belirlenir. Bu hiyerarşide kelimeler bir çizge tabanlı ağda gösterilmektedir. Çizgede yer alan her bir düğüm kelimelere karşılık gelmekte ve düğümler arasında geçiş (graph traversing) algoritmalarıyla kelimeler arasındaki yakınlık belirlenebilmektedir. Kavram ilişkilerinin belirlenmesiyle ilgili pek çok çalışma yapılmıştır. Budanitsky ve Hirst (Budanitsky ve Hirst, 2006). kavramsal ilişkilerinin WordNet hiyerarşisi üzerinde belirlenmesi için üç ölçüt kullanımından bahsetmiştir. Kavram benzerliği için ilk ölçüt kavramlar arasındaki en kısa uzaklığı veren "uzunluktur". İki kavram arasındaki uzunluk kavramları gösteren iki düğüm arasındaki köşelerin sayılmasıyla bulunur. İkinci ölçüt kavramların kök düğüme olan uzaklığı yani "derinliktir". Son ölçüt ise iki kavrama kapsayan en yakın üst kavramdır (the least common subsumer- LCS). Wu& Palmer (Wu ve Palmer, 1994), Lin (Lin, 1998), Leacock & Chodorow (Leacock ve Chodorow, 1998) sadece WordNet üzerindeki kavramların birbirlerine olan uzaklık, derinlik ve LCS ölçütlerine bakarak ilişki seviyelerini belirlemeye çalışmışlardır. Resnik, Lin, Lord, Jiang & Conrath gibi araştırmacılar ise derinlik, uzunluk ve LCS ölçütlerinin yanında kavramların derlemde geçme sıklıklarını kavramlar arası ilişki belirlemede kullanmışlardır. Bilgi içeriği ve köşe sayma yaklaşımlarının dışında Adapted Lesk yaklaşımı gibi özellik tabanlı yaklaşımlarda benzerlik çalışmalarında kullanılmaktadır (Oliver, 2020; Kolajo et al., 2020). WordNet kullanılarak elde edilen kavramlar arasındaki bilgi çıkarımı (Iqbal et al., 2019), metinlerde anlamsal ilişki belirlenmesi (Hasan et al., 2020),, sosyal ağlarda duygu analizi ve etkileyici lider keşfinde ve metin sınıflandırmada özellik seçimi (Zhu et al., 2019),, metin özetlemede (Jain et al., 2019), kullanılmaktadır.

Çalışmamız kapsamında haber sitelerinden elde edilmiş RSS haber beslemeleri kullanılarak sınıf etiketi belirlemede önemli kelimeleri belirlenerek bu kelimeler arasındaki anlamsal benzerlik ilişkileri WordNet ile çıkarılmıştır. Çıkarılan bu ilişkilerin terim

ağırlıklandırmasıyla elde edilen terimler ve elle işaretlenmiş anlamsal ilişkileri belirlenmiş kelimelerle karşılaştırılması yapılmıştır. WordNet ile belirlenen anlamsal ilişkilerin insan algısı ve kelime ağırlıklandırmayla örtüşmesi incelenmiştir.

Çalışmamızın ikinci bölümünde veri kümesi ve sistem tasarımı anlatılmıştır. Üçüncü bölümde elde edilen sonuçlar sunulmuş ve bulgular değerlendirilmiştir. Son bölümde ise sonuçlar özetlenmiş ve gelecek çalışmalara değinilmiştir.

2. Materyal ve Metot

Bu bölümde çalışmamızda kullandığımız veri seti, WordNet benzerlik çıkarım yaklaşımları ve sistem tasarımı anlatışmıştır.

2.1. Veri Kümesi

Çalışmamızda spor, teknoloji ve ekonomi kategorilerine ait 1561 RSS haber beslemesi BBC, CNN gibi haber içeriği sağlayıcılardan elde edilmiştir. Bu haberlerin 543 tanesi spor, 548 tanesi ekonomi ve 470 tanesi ise teknoloji kategorilerine ait haberlerdir. RSS haberleri haber başlığı ve haber içeriği birleştirilerek çalışmamızda kullanılmıştır.

2.2. WordNet ile Kavramlar Arası Benzerlik Çıkarım Yaklaşımları

WordNet çizge tabanlı ağ yapısında olup kavramlar düğümler halinde bu ağ üzerinde gösterilmektedir. Çalışmamızda Lin, Wu-Palmer (WP) ve Jiang Conrath (JC) benzerlik ölçüm yöntemleriyle kavramlar arasındaki yakınlıklar belirlenmiştir.

Wu-Palmer benzerliği, iki kavramın hiyerarşik bir yapıdaki yakınlıklarına bağlı olarak hesaplanır. Kenar merkezli bir benzerlik yaklaşımıdır. C1 ve C2 arasındaki yakınlık aşağıdaki denklem 1'de gösterilmiştir. C1 ve C2'yi kapsayan en yakın üst kavram LCS ile gösterilmiştir. En yakın ortak kavramın kök düğüme olan uzaklık derinlik fonksiyonuyla verilmiş, düğüm sayarak elde edilen C1-C2 arasındaki uzaklıklık ise uzunluk fonksiyonu ile aşağıdaki denklemde gösterilmiştir.

$$\mathrm{WP}(\mathrm{C1},\mathrm{C2}) = \frac{2*\mathrm{derinlik}(\mathrm{LCS}(\mathrm{C1},\mathrm{C2}))}{\mathrm{uzunluk}(\mathrm{C1},\mathrm{C2})+2*\mathrm{derinlik}(\mathrm{LCS}(\mathrm{C1},\mathrm{C2}))} \quad (1)$$

Lin benzerliğine göre hiyerarşik yapıdaki iki kavramın kökleri ne kadar genel bir kavramda kesişiyorsa bu iki kavram o ölçüde benzerdir. Lin bunu ortaklık olarak nitelendirmiştir ve ortaklık aşağıdaki gibi ifade edilir. Buradaki IC bilgi içeriğini (information content) ifade etmektedir. Denklem 2'de C1 ve C2 arasındaki Lin benzerliğinin hesaplaması gösterilmiştir.

$$\mathrm{Lin}(\mathrm{C1},\mathrm{C2}) = \frac{2*\mathrm{IC}(\mathrm{LCS}(\mathrm{C1},\mathrm{C2}))}{\mathrm{IC}(\mathrm{C1})+\mathrm{IC}(\mathrm{C2})} \quad (2)$$

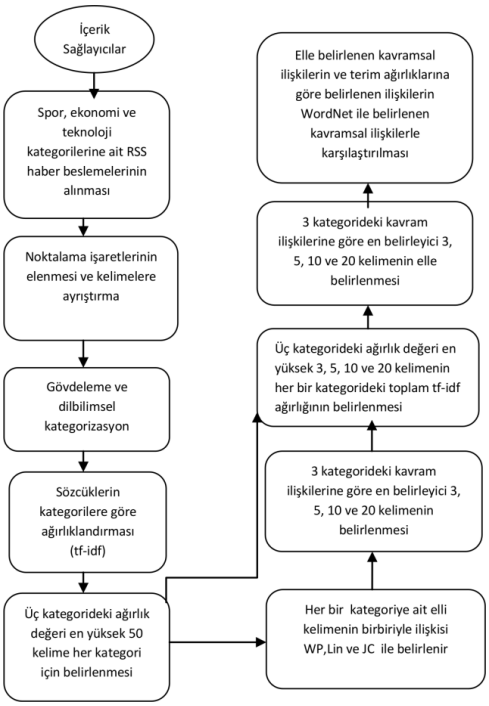
Jiang – Conrath benzerliği de düğüm merkezli bir benzerlik modeli olup bir kavramın üst kavramına olan anlamsal uzaklığını koşullu olasılıkla değerini, P, belirleyerek hesaplar.

$$\mathrm{JC}\ (\mathrm{A},\mathrm{B}) = \frac{1}{2\log\mathrm{P}(\mathrm{LCS}(\mathrm{A},\mathrm{B}))-(\log\mathrm{P}(\mathrm{A})+\log\mathrm{P}(\mathrm{B}))} \quad (3)$$

2.3. Sistem Tasarımı

Şekil 1'de çalışmamızın temel adımları gösterilmiştir. Üç kategoriye ait İngilizce RSS verileri haber kaynaklarından elde edildikten sonra metin madenciliğinin ön adımları Stanford NLP kütüphanesi (Manning et al., 2014) kullanılarak gerçekleştirilmiştir. Noktalama işaretleri atıldıktan sonra RSS beslemeleri kelimelerine ayrıştırılmıştır. Özellik sayısının azaltılması ve aynı kelimenin farklı yapılarının tek bir özellik olarak gösterilmesi için gövdeleme işlemi yapılmıştır. WordNet üzerinde kavramsal benzerliklerine bakacağımız kelime türleri isimler olduğu için Stanford NLP ile sözcüklerin türleri belirlenmiş ve isim türündeki gövdeler elde edilmiştir.

RSS beslemelerindeki bir terimin her bir kategoriye ait terim frekansı ilgili kategorideki RSS beslemelerindeki geçme sıklığı kadardır.Bir terimin ters doküman frekansı ise tüm kategorilere ait RSS beslemelerinin bir kategoriye ait geçtiği RSS beslemelerine oranın logaritmik değeridir. Bu iki değerin çarpılmasıyla terim frekansı-ters dokuman frekansı (tf-idf) ağırlıklandırma değerleri elde edilmiştir. Çalışmamız kapsamında her kategoriyi en iyi yansıtan 50 kelime tf-idf terim ağırlığına göre belirlenmiştir. Böylece ekonomi, teknoloji ve spor kategorileriyle ilişkili terim ağırlık değeri en yüksek ilk 50 kelime belirlenmiştir.



Şekil 1. Sistem tasarımı

Üç kategoriyle ilişkili en yüksek ağırlık değerine sahip 50 terim kendi aralarındaki anlamsal benzerlik değerleri NLTK (Loper ve Bird, 2002) kütüphanesi kullanılarak hesaplanmıştır. Tablo 1'de gösterildiği gibi kategorileri yansıtan her kelimenin aynı kategorideki diğer kelimelerle anlamsal benzerlikleri Lin, Wu-Palmer ve Jiang-Conrath yaklaşımlarıyla ayrı ayrı hesaplanmakta ve sonunda tüm kelimelerle olan anlamsal benzerlikleri toplanarak her bir kelimenin "Kategorik Anlamsal İlişki Değerleri" (KAİD) üç ayrı benzerlik yaklaşımı için belirlenmektedir.

Tablo 1. Kategorik Anlamsal İlişki Değerleri

	Benzerlik Yaklaşım Değerleri				
	1	*2*	.....	*50*	*Toplam*
1	-	*0,3*	...	*0,04*	$\sum_{i=1}^{50} f1_i$
2	*0,3*	-		*0,71*	
. . .			-		. . .
50	*0,04*	*0,71*		-	$\sum_{i=1}^{50} f50_i$

Her bir kategorideki kelimeler için belirlenen kategorik kavramsal ilişki değerlerinden en yüksek değere sahip 3,5,10 ve 20 kelime belirlenerek insan tarafından elle belirlenen kategorik değeri en yüksek 3,5, 10 ve 20 kelime ile karşılaştırılmıştır. Benzer şekilde her bir kelimenin tf-idf değerleriyle ilgili kategoriyi yansıtma değerleri yine WordNet ile elde edilen benzerlik yaklaşımıyla karşılaştırılmıştır. Bunun için tf-idf değeri en yüksek 3, 5, 10, 20 kelime ile Wu-Palmer, Lin ve Jiang-Conrath tarafından belirlenen en yüksek 3,5,10 ve 20 anlamsal ilişki değeri skoruna sahip kelimeler karşılaştırılmıştır. Çalışmamızda kavramlar arasındaki benzerlik ilişkilerinin insan algısıyla ve terim ağırlıklandırmasıyla karşılaştırılması gerçekleştirilmiştir.

3. Araştırma Sonuçları ve Tartışma

İlk karşılaştırmada kelime ağırlıklandırma değeri her bir kategori için belirlenen 50 kelime kullanılarak en yüksek KAİD değeri olan 3, 5, 10 ve 20 kelime belirlenmiş ve aynı 50 kelime kullanılarak insan algısıyla oluşturulmuş en yüksek 3, 5, 10 ve 20 kelime belirlenerek KAİD ile insan algısı tarafından ortak belirlenen kelimelerin sayısı Tablo 2'de gösterilmiştir. Çalışmamızda insan tarafından (elle) belirlenen kelimelerin anlamsal ilişki değerleri için birden fazla öğrenci kullanılmıştır.

Tablo 2. Kelimelerin KAİD ve Elle Etiketlemedeki Anlamsal Değerlerine Göre Sıralamasının Karşılaştırılması

	En yüksek 3, 5, 10 ve 20 KAİD olan kelimeler											
	JC				**Lin**				**WP**			
Kategori	**3**	**5**	**10**	**20**	**3**	**5**	**10**	**20**	**3**	**5**	**10**	**20**
Ekonomi	3	4	6	10	3	4	7	10	0	1	2	9
Spor	1	2	4	11	2	2	3	12	0	1	5	14
Teknoloji	0	0	3	6	0	0	2	6	0	1	5	6

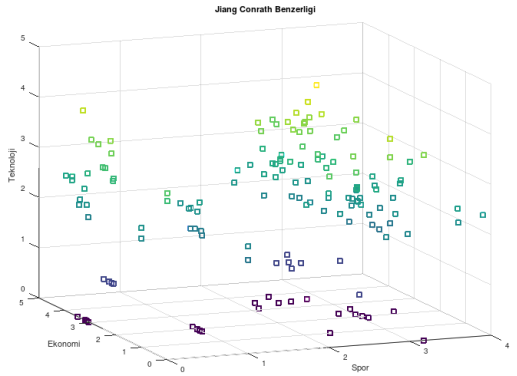
Elle etiketleme sonucunda kategorisini en iyi yansıttığı düşünülen 3, 5, 10 ve 20 kelime ile JC, Lin ve WP yaklaşımlarıyla elde edilen 3, 5, 10 ve 20 en yüksek değere sahip olan kelimelerin örtüşmesi sırasıyla Tablo 2'de her bir kategori için gösterilmiştir. Elde edilen sonuçlara göre insan algısıyla WordNet tabanlı yaklaşımların en az benzerlik gösterdiği kelimelerin teknoloji kategorisine ait olduğu gözlemlenmektedir. Bu durumun olası en büyük nedeni teknolojiyle ilgili yeni içeriklerin RSS beslemelerinde statik WordNet'e göre daha hızlı değişmesidir. Ekonomi ve spor kategorilerindeki insan algısıyla ve WordNet yaklaşımlarındaki oranlar genel olarak birbirine yakın sonuçlar elde edilmiştir. Bu iki kategori için insan algısı değeriyle KAİD benzerlik değerleri en yüksek 20 kelime için %50'lerin üzerinde değerler elde edilmiştir. Örneğin Wu-Palmer KAİD ile elde edilen spor kategorisine ait en yüksek değerli 20 kelimeden insan algısıyla 14 kelime ile aynı kelimeler olduğu gözlemlenmiştir. Bir başka ilginç sonuç ise ekonomi kategorisindeki Jiang-Conrath ve Lin KAİD değeri en yüksek 3 kelimenin insan algısıyla belirlenen en yüksek değere sahip üç kelimeyle hepsinin aynı olmasıdır. Bu sonuçlara göre spor kategorisiyle insan algısının oluşturduğu en yüksek kelimelerin birbiriyle örtüştüğü gözlemlenmektedir. JC, Lin ve WP karşılatırmasını insan algısıyla elde edilen kelimeler kullanılarak yaptığımızda üç yaklaşımın da birbirine benzer şekilde insan algısıyla örtüştüğü gözlemlenmektedir.

İkinci karşılaştırmada her bir kategoride kelime ağırlıklandırma değeri en yüksek 3, 5, 10 ve 20 kelime ile kelime ağırlıklandırma değeri her bir kategori için belirlenen 50 kelime kullanılarak elde edilen kelimeler içerisinden en yüksek KAİD değeri olan 3, 5, 10 ve 20 kelime belirlenerek örtüşen ortak kelimelerin sayısı Tablo 3'te gösterilmiştir

Tablo 3. Kelimelerin KAİD ve TF-IDF Değerlerine Göre Sıralamasının Karşılaştırılması

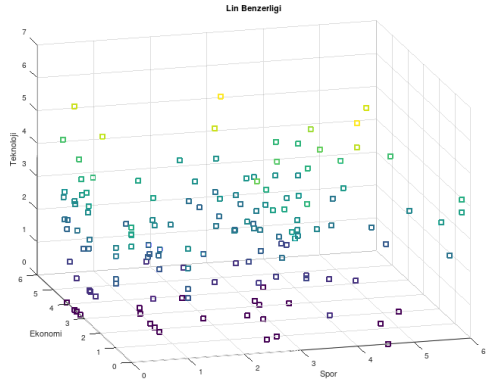
	En yüksek 3, 5, 10 ve 20 KAİD olan kelimeler											
	JC				**Lin**				**WP**			
Kategori	**3**	**5**	**10**	**20**	**3**	**5**	**10**	**20**	**3**	**5**	**10**	**20**
Ekonomi	1	2	4	8	0	0	4	9	1	1	1	9
Spor	0	0	1	7	0	0	2	6	0	0	1	5
Teknoloji	1	1	4	8	0	0	4	10	1	1	5	9

Tablo 3'te yer alan verileri incelediğimizde tüm kategoriler için tf-idf ağırlıkladırma değerleri kullanılarak sıralanan kelimeler ile WordNet tabanlı KAİD ile elde edilen kelimelerin oranın benzer olduğu sonucu gösterilmiştir. En yüksek 20 değeri olan örtüşme değerleri karşılaştırıldığında spor kategorisine ait değerlerin diğer iki kategorideki örtüşme değerlerinden biraz daha düşük olduğu gözlemlenmektedir. Tablo 2'deki değerlerde olduğu gibi elde edilen sonuçlarda JC, Lin ve WP tarafından KAİD değerlerine göre belirlenip örtüşen kelime sayısı birbirine benzer sayılarda çıkmıştır. Genel olarak Tablo 3'te elde edilen değerlerin Tablo 2'de elde edilen değerlerden daha düşük olduğu gözlemlenmektedir. Bu nedenle insan algısıyla elde edilen kavramlar arası benzerlik elle işaretlenerek hazırlanmış hiyerarşik bir sözlük olan WordNet ile daha uyumlu çıkmıştır.



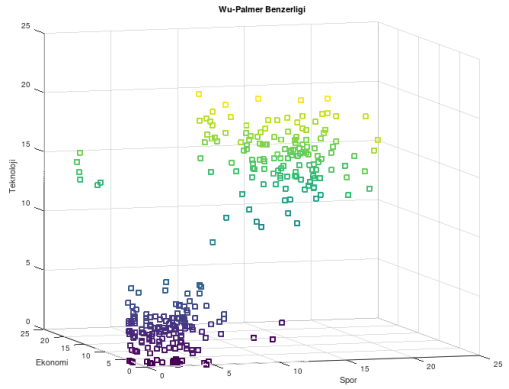
Şekil 2. Kelimelerin Jiang Conrath Benzerlik Değerleri

Şekil 2'de kelimelerin Jiang Conrath kategorik anlamsal ilişki değerleri teknoloji-ekonomi-spor kategorilerine içeren 3 boyutlu uzayda gösterilmiştir. Her üç kategori için 50 kelime olduğu için toplam 150 nokta bulunmaktadır. Mor veriler spor kategorisindeki, mavi veriler ekonomi kategorisindeki ve yeşil veriler teknoloji kategorisindeki verileri göstermektedir.



Şekil 3. Kelimelerin Lin Benzerlik Değerleri

Şekil 3'de kelimelerin Lin kategorik anlamsal ilişki değerleri teknoloji-ekonomi-spor kategorilerine içeren 3 boyutlu uzayda gösterilmiştir.



Şekil 4. Kelimelerin Wu-Palmer Benzerlik Değerleri

Şekil 4'de kelimelerin Wu-Palmer kategorik anlamsal ilişki değerleri teknoloji-ekonomi-spor kategorilerine içeren 3 boyutlu uzayda gösterilmiştir.

Şekil 2, Şekil 3 ve Şekil 4'deki veriler karşılaştırıldıklarında x-y-z eksenlerinde sıfır değerine düşen veri sayısının Lin benzerliğinde daha fazla olduğu gözlemlenmektedir. Bunun dışında verilerin en yüksek KAİD'leri hesaplama farklılığından dolayı Wu-Palmer yaklaşımıyla elde edilmiştir. Wu-Palmer ile elde edilen verilerin değer dağılımları birbirlerine yakın çıkmakla birlikte Jiang Conrath ve Lin gösterimlerinde verilerin dağılım varyasyonları üç eksende de daha fazla olduğu gözlemlenmiştir.

Doğal dil işleme, veri madenciliği, bilgi erişimi, metin madenciliği gibi alanlarda kelimelerin anlam belirsizliğinin giderimi önem taşımaktadır. Çalışmamız kapsamında kelimelerin anlam belirsizliği gidermek için kullanılan WordNet ile kavramlar arası ilişkiler belirlenmiştir. Anlamsal benzerlik değerleri çıkartılan kelimeler tf-idf ağırlıklandırma değerleri ve insan algısıyla elde edilen kelimelerle karşılaştırılmıştır. Veri kümesi olarak kullandığımız RSS beslemeleri ön işlem adımlarından geçirildikten sonra üç kategoriye ait terim ağırlıkları elde edilmiş ve ilişki değeri en yüksek kelimeler kullanılarak insan algısı, ağırlıklandırma değerleri ve WordNet kelime benzerlik yaklaşımları birbirleriyle karşılaştırılmıştır. Beklendiği üzere insan algısı ve WordNet tabanlı yaklaşım benzerlikleri terim ağırlıklandırmasıyla elde edilen kelimelerden daha benzer çıktığı gözlemlenmiştir. Bu sonuçlara göre WordNet tabanlı yaklaşımlarla kelimelerin anlam belirsizliğinin gideriminde insan algısına yakın sonuçlar çıktığı sonucu çıkarılmıştır.

Benzerlik ölçütlerini kullanarak anlam belirsizliği giderimi gerçekleştiren çalışmalar literatürde yer almasına rağmen bu işlemin doğruluğunu elle etiketlenmiş veriler üzerinden yaparak insan algısıyla benzerliği üzerine yapılan çalışma sayısı oldukça eksiktir. Çalışmamızda diğer çalışmalardan farklı olarak insan algısının dışında kategorik bazlı tf-idf ağırlık hesaplaması yaparak kategoriler için belirlenmiş en belirleyici kelimelerin sıralamasının da karşılaştırmasını gerçekleştirmiştir. Kelimeler ilk olarak kategorilerdeki tf-idf değerlerine göre sıralanarak en önemli 50 kelime belirlenmiş daha sonra da bu kategorik kelimeler arasında da WordNet tabanlı yaklaşımlar kullanılarak anlamsal olarak ilgili kategoriyi en iyi yansıten 3, 5, 10 ve 20 kelime belirlenmiştir. İki aşamalı değerlendirme yoluyla elde edilen bu kelimeler metin madenciliğinde özellik seçiminde, etiket bulutu belirlemede ve metin özetlemede kullanılabilir.

Kelimelerin kategorik anlamsal ilişki değerleri aynı zamanda kategori uzayında çizdirilmiş ve anlamsal değerleri eksen üzerine yakın bulununan veriler kontrol edilmiştir. Bununla birlikte en yüksek KAİD'e sahip WordNet yaklaşımı da belirlenmiştir. Verilerin

görselleştirilmesiyle kategorilere ait kelimelerin kategorileri farklı WordNet benzerliği yöntemlerine göre yansıtma gücü de değerlendirilmiştir.

Bundan sonraki çalışmalarda elde edilen iki aşamalı benzerlik yaklaşımıyla özellik seçimi yaparak metin sınıflandırmada WordNet'in etkisi incelenecektir.

4. Sonuç

Çalışmamız kapsamında ilk olarak tf-idf ağırlıklandırmasıyla kategorik olarak anlamlı kelimeler belirlenmiş ve daha sonra bu kelimelerinde KAİD'leri çıkartılarak iki katmanlı anlamsal ilişki değerleri yüksek kelimeler insan algısı ve sadece tf-idf ağırlıklandırmasıyla elde edilen kelimelerle karşılaştırılmıştır. Elde ettiğimiz sonuçlar ileride iki katmanlı yapının anlam belirsizliği gidermede farklı çalışma alanlarında kullanılabileceği göstermiştir.

Kaynakça

Chen, J., & Palmer, M. (2005). Towards robust high performance word sense disambiguation of english verbs using rich linguistic features. In International Conference on Natural Language Processing (pp. 933-944). Springer, Berlin, Heidelberg.

Dang, H. T., Chia, C. Y., Palmer, M., & Chiou, F. D. (2002). Simple features for Chinese word sense disambiguation. In *COLING 2002: The 19th International Conference on Computational Linguistics*.

Mihalcea, R. (2007, April). Using wikipedia for automatic word sense disambiguation. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference* (pp. 196-203).

Seo, H. C., Chung, H., Rim, H. C., Myaeng, S. H., & Kim, S. H. (2004). Unsupervised word sense disambiguation using WordNet relatives. *Computer Speech & Language*, *18*(3), 253-273.

Pham, T. P., Ng, H. T., & Lee, W. S. (2005). Word sense disambiguation with semi-supervised learning. In *Proceedings of the National Conference on Artificial Intelligence* (Vol. 20, No. 3, p. 1093). Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999.

Simov, K., Koprinkova-Hristova, P., Popov, A., & Osenova, P. (2020). A Reservoir Computing Approach to Word Sense Disambiguation. *Cognitive Computation*, 1-10.

Miller, G. A. (1998). *WordNet: An electronic lexical database*. MIT press.

Budanitsky, A., & Hirst, G. (2006). Evaluating wordnet-based measures of lexical semantic relatedness. *Computational linguistics*, *32*(1), 13-47.

Wu, Z., & Palmer, M. (1994). Verb semantics and lexical selection. *arXiv preprint cmp-lg/9406033*.

Lin, D. (1998). An information-theoretic definition of similarity. In *Icml* (Vol. 98, No. 1998, pp. 296-304).

Leacock, C., & Chodorow, M. (1998). Combining local context and WordNet similarity for word sense identification. *WordNet: An electronic lexical database*, *49*(2), 265-283.

Oliver, A. (2020). Aligning Wikipedia with WordNet: a Review and Evaluation of Different Techniques. In *Proceedings of The 12th Language Resources and Evaluation Conference* (pp. 4851-4858).

Kolajo, T., Daramola, O., Adebiyi, A., & Seth, A. (2020). A framework for pre-processing of social media feeds based on integrated local knowledge base. *Information Processing & Management*, *57*(6), 102348.

Iqbal, F., Fung, B. C., Debbabi, M., Batool, R., & Marrington, A. (2019). Wordnet-based criminal networks mining for cybercrime investigation. *IEEE Access*, *7*, 22740-22755.

Hasan, A. M., Noor, N. M., Rassem, T. H., Noah, S. A. M., & Hasan, A. M. (2020). A proposed method using the semantic similarity of WordNet 3.1 to handle the ambiguity to apply in social media text. In *Information Science and Applications* (pp. 471-483). Springer, Singapore.

Zhu, X., Xu, Q., Chen, Y., & Wu, T. (2019). An Improved Class-Center Method for Text Classification Using Dependencies and WordNet. In *CCF International Conference on Natural Language Processing and Chinese Computing* (pp. 3-15). Springer, Cham.

Jain, A., Vij, S., & Tayal, D. K. (2019). Text Summarization Using WordNet Graph Based Sentence Ranking. In *Proceedings of 2nd International Conference on Communication, Computing and Networking* (pp. 711-715). Springer, Singapore.

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J. R., Bethard, S., & McClosky, D. (2014). The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations* (pp. 55-60).

Loper, E., & Bird, S. (2002). NLTK: the natural language toolkit. *arXiv preprint cs/0205028*.