TBMM Genel Kurul Tutanaklarından Yakın Anlamlı Kavramların Çıkarılması

Araştırma Makalesi/Research Article

Hüseyin POLAT*, Mesut KÖRPE

Bilgisayar Mühendisliği, Gazi Üniversitesi Teknoloji Fakültesi, Ankara, Türkiye
polath@gazi.edu.tr, mesut.korpe@gazi.edu.tr
(Geliş/Received:06.03.2018; Kabul/Accepted:22.05.2018)
DOI: 10.17671/gazibtd.402468

***Özet*—**Yakın anlamlı kavramların bulunması, kavramın bir derlemdeki semantik anlamını yakalamamızı ve kavramın hangi bağlamda kullanıldığını elde etmemizi sağlar. Kelime Uzayı Modeli; anlamsal olarak benzer kelimeleri, vektör uzayında bir birine yakın dağılımla gösteren bir modeldir. Her bir kelimenin bir vektörle temsil edildiği bu modelde oluşan kelime vektörleri kelime yerleştirme (word embeddings) olarak adlandırılır. Kelime vektörleri metin analizi gerçekleştiren özellikle yapay sinir ağlarını temel alan Doğal Dil İşleme (DDİ) sistemlerinde girdi olarak kullanılır. Bu çalışmada, veri seti olarak TBMM Genel Kurul görüşme tutanakları kullanılmış, Word2vec modeli ve GloVe modeli ile kelime vektörleri çıkarılmıştır. Elde edilen kelime vektörleri kullanılarak TBMM Genel Kurul tutanaklarında geçen herhangi bir kavrama en yakın anlamlı kavramlar bulunmuştur. Literatürdeki benzer çalışmalarda iki farklı kelime yerleştirme modellerinin bir kavramı tamamen farklı bağlamda değerlendirdiği duruma rastlanılmamıştır. Bu çalışma sonucunda, Word2vec ve GloVe modellerinin çıktılarının bir kavramın farklı bağlamlarda kullanımını bulmak için değerlendirilebileceği görülmüştür. Çalışmada derleme özgü analojilerin her iki modelde de bulunabildiği görülmüştür. Bu çalışmadan elde edilen sonuçlar Bilgi Çıkarımı uygulamalarında benzer kavramların anahtar kelime olarak önerilmesi için uygundur.

***Anahtar Kelimeler*—** kelime vektörü, Word2vec modeli, GloVe modeli, sinir ağları, doğal dil işleme

Extracting Close Meaning Concepts from GNAT Parliamentary Minutes

***Abstract*—** Having close meaning concepts allows us to capture semantic meaning of a concept in a corpus and to get the context in which it is used. Vector Space Model locates similar concepts to close each other in a vector space. In this model, every word is represented by a vector and it is called as a “word embedding” in literature. Word vectors are used as an input in text analysis, especially in NLP tasks based on neural networks. In this paper, The Grand National Assembly of Turkey (GNAT) Parliamentary minutes are used as data set. Word vectors are extracted using Word2vec model and GloVe model. Using word vectors, obtained concepts which are semantically close to any concept in the GNAT Parliamentary minutes. The experimental results show that, different contexts of a concept can be extracted from corpus by both models when results assessed separately. Analogies which are specific to the corpus can be extracted by both models. Results obtained in this model are suitable for suggesting similar concepts as keywords in Information Retrieval systems.

***Keywords*—** word vectors, Word2vec model, GloVe model, neural networks, natural language processing

1.GİRİŞ (INTRODUCTION)

Benzer kavramların bulunması Bilgi Toplama (Information Retrieval) sistemlerinde bulunması amaçlanan kavram ile tam eşleşmenin bulunamadığı durumlarda bilgiye eksiksiz ulaşmak için önemlidir. Kavramın içinde bulunduğu bağlama göre anlamı değişebilmektedir. Örneğin “Asya kaplanları” kavramı hayvan bilimi ile ilgili bir bağlamda farklı bir anlamda değerlendirilebilir ama çalışmamızda kullandığımız TBMM Genel Kurul tutanak veri setinde ekonomik bir bölgeyi temsil ettiği kelime vektör değerlerinden anlaşılmaktadır. Dağılım hipotezi yakın anlamlı kelimelerin benzer bağlamda konumlanacağını anlatır [1].

Kelime vektörleri; her bir kelimenin dağılım hipotezine göre bir vektörle temsil edilmesini sağlar. Kelime vektörlerinin çok düşük boyutlu olması (low dimensitionality) ilgisiz kelimelerin bir birine yakın dağılımına ve semantik anlamın kaybolmasına, çok fazla boyut olması (high dimensitionality) hesaplama karmaşıklığının artmasına sebebiyet verebilir. Kelime yerleştirme; kelimelerin vektör temsillerinin semantik anlamını kaybetmeden, boyutun indirgenmiş olarak temsil edilebilmelerine imkân verir. Bilgi çıkarımı, Doküman sınıflandırma, Soru-Cevap sistemleri, Varlık ismi tanıma sistemleri gibi Doğal Dil İşleme uygulamalarında kelime yerleştirme sıkça kullanılmaktadır.

Kelime anlamı temsillerinde sayma tabanlı ve tahmine dayalı yöntemler kullanılmaktadır. Gizli anlam analizi (Latent Semantic Analysis-LSA), kelime anlamı temsillerinde en çok kullanılan sayma tabanlı yöntemlerden birisidir [2,3]. Gizli anlam analizi, dokümanlardan oluşan derlemi girdi olarak alır. Bu yöntemde her bir dokümanda bulunan her bir kelime için doküman-kelime sıklık matrisi oluşturulur. Daha sonra bu matrise Tekil Değer Ayrıştırma (Singular Value Decomposition-SVD) uygulanarak boyut düşürme işlemi yapılır ve her bir kelimenin vektör temsili bulunur. Anlamsal benzerlik bulunması için de 2 kelime vektörü arasındaki kosinüs açısına göre hesaplama yapılır. Yapay sinir ağları ise kelime anlamı temsillerinde en çok kullanılan tahmine dayalı bir yöntemdir [4,5]. Kelime yerleştirme için yapay sinir ağları ilk olarak Bengio [6] tarafından önerilmiştir. İleri Beslemeli Sinir Ağı Dil Modeli (Feed Forward Neaural Network Language Model-FFNNLM) doğrusal bir iz düşüm ve doğrusal olmayan bir saklı katmandan oluşmaktadır. Model, verilen kelimelerden, bağlamdan hedef kelimeyi tahmin etmeye çalışır. Mikolov [5] bu modeldeki hesaplama karmaşıklığına çözüm bulduğu Word2vec modelini tanıtmıştır. Yapay sinir ağları kelime vektörlerinin çıkarılmasında gizli anlam analizi gibi klasik yöntemlerin yerini almaya başlamıştır. Özellikle Word2vec gibi tahmine dayalı yöntemler bu sayede popüler olmuştur. Li [7] ve arkadaşları çalışmalarında Word2vec modeli ile elde ettikleri kelime vektörlerini yerel bilgi olarak değerlendirmişler ve kelimenin içinde geçtiği dokumandan elde ettikleri vektörle (global bilgi) birlikte değerlendirerek "çoklu-bağlamsal karışık yerleştirme" olarak adlandırdıkları kelime vektörleri önermişlerdir. Boyut indirgeme için oto kodlayıcılar kullanılmaktadır. Kaynar ve Aydın [8] boyut düşürme tekniklerini karşılaştırdıkları çalışmalarında boyut düşürmek için derin öğrenme tabanlı oto kodlayıcıları kullanmışlardır. GLoVe [9] modeli ise hem sayma tabanlı sistemlerin hem de tahmine dayalı yöntemlerin avantajlarını bir araya getirdiğini iddia etmektedir. Altszyler ve arkadaşları çalışmalarında [10] küçük doküman derleminde LSA ve Word2vec modelini karşılaştırmışlar ve LSA'nin küçük veri setleri için daha iyi sonuçlar verdiğini göstermişlerdir. Çalışmada kelime sayısı 10^6 dan fazla olduğunda, tahmin tabanlı Word2vec modelinin, sayma tabanlı LSA modeline göre benzerlik bulma konusunda daha başarılı olduğu gösterilmiştir.

Levy ve arkadaşları çalışmalarında [11] Word2vec modelinin kelime benzerliği konusunda GloVe'dan daha başarılı olduğunu göstermişlerdir.

Naili ve arkadaşları çalışmalarında [12], Word2vec CBOW modelinin sık geçen kelimeler için, Word2vec Skip-gram modelinin seyrek geçen kelimeler için daha başarılı olduğunu göstermişlerdir.

Semantik sözlükler bir kelimenin eş anlamı, üst anlamı, alt anlamı gibi bilgileri içerir. Faruqui ve arkadaşları çalışmalarında [13] kelime vektörlerinin kalitesini artırmak için vektör uzayı kelime temsillerinden çıkan sonuçları WordNet [14], FrameNet [15] ve ParaPharase [16] gibi semantik sözlüklerle birlikte değerlendirerek "retrofitting" adını verdikleri kelime vektörlerini elde etmişlerdir.

Bu çalışmada, TBMM Genel Kurul tutanaklarındaki birbirine yakın anlamlı kavramlar çıkarılırken yapay sinir ağları temelinde modellenen Word2vec modeli, global birlikte geçme matrisi temelinde modellenen GloVe modeli kullanılmıştır. Çıkan sonuçlar Türk Dil Kurumu güncel sözlüğü ile karşılaştırıldığında birbiri ile yakın anlamlı kavramların ve analojilerin Word2vec ve GloVe modeli ile başarılı bir şekilde elde edildiği görülmüştür.

2. MATERYAL VE METOD (MATERIAL AND METHOD)

Bu çalışmada, TBMM Genel Kurul görüşme tutanaklarındaki benzer kavramların çıkarılması için öncelikle kelime vektörlerinin oluşturulması gerçekleştirilmiştir. Kelime vektörlerinin oluşturulmasında kullanılan veri setinin özellikleri şunlardır;

- Veri seti TBMM parlamento çalışma alanıyla ilgilidir ve dili Türkçedir.
- TBMM genel kurul tutanakları, konuşma dilinin doğrudan yazıya aktarılması şeklinde olduğu için Türkçenin bölgelere göre farklılığını yansıtır. Türkçede kullanılan bütün kelimeler, deyimler dokümanlarda mevcuttur.
- Tutanaklar yazıya geçirilirken TBMM tutanak uzmanları tarafından düzeltildiği için kelimelerin yanlış yazımı, kısaltılması gibi metin analizini zorlaştıran durumlar söz konusu değildir.
- TBMM Genel Kurul oturumlarında yazılı dokümanların (Tasarı, önergeler v.d.) kâtip üye tarafından okunması tutanağa geçirildiği için hem konuşma dilinin hem de yazı dilinin dokümanlarda bulunması bu anlamda bir çeşitlilik oluşturur.
- Özellikle literatürde çok fazla çalışmanın olduğu Twitter, Facebook gibi sosyal medya mecralarında kelimelerin kısaltılması, yazarken değiştirilmesi gibi metin analizini zorlaştıran durumlar bu veri setinde gözlenmez.

Birleşim, TBMM Genel Kurulunun belirli bir günde açılan toplantısıdır[1]. Oturum, bir birleşimin ara ile bölünen kısımlarından her birisidir. Bu çalışmada kullanılan veri seti, 01.09.1994 ve 23.04.2017 tarihleri arasındaki 2800 TBMM genel kurul birleşimindeki tutanakları içerir. Bu birleşimlerdeki tutanaklarda toplam olarak 20200 metin sayfası bulunur. Her bir sayfa, ortalama 550 ile 600 kelimeden oluşmaktadır. TBMM Genel Kurul tutanaklarının önemli bir özelliği her sayfa ayrı bir doküman olarak düşünülemez. Sayfalar bir birinin devamıdır ve her bir konu birkaç birleşimde görüşülebilir.

2.1. Veri Setinin Oluşturulması (Construction of Data Set)

Veri seti oluşturulurken açık veri olarak TBMM İnternet sitesinde https://www.tbmm.gov.tr/tutanak/tutanaklar.htm adresinden yayınlanan TBMM Genel Kurul Tutanak metinlerinden yararlanılmıştır. Her bir birleşimin yayınlandığı html doküman ayrıştırma (html parsing) yapılarak JSON nesnesine dönüştürülmüştür. JSON nesneleri tutanak metni, sayfa numarası, birleşim tarihi gibi özellikleri içerir. JSON nesneleri yapısal çözümlenerek txt uzantılı metin dosyaları oluşturulmuştur. İki birleşim tarih aralığı parametre olarak alınarak, bu iki birleşim tarihi aralığındaki her bir birleşimin tüm tutanak sayfalarından oluşan txt uzantılı tek bir metin dosyası oluşturulmuştur. Bu dosyanın adı; birleşim tarihi, sayfa sayısı gibi birleşim özelliklerini içermektedir. Örneğin "20160105-26-1-23-22499-H-58.txt" dosya adı birleşimin 01.05.2016 tarihinde gerçekleştiği ve toplam 58 sayfadan oluştuğunu göstermektedir.

2.2. Veri Seti Üzerinde Doğal Dil İşleme Adımları (Natural Language Processing on Data Set)

Veri setinde bulunan 2800 birleşim dosyası, Doğal Dil İşleme (DDİ) metotları uygulanarak tek bir metin dosyası haline getirilmiştir. Türkçe DDİ metotlarının uygulanması aşamasında Java tabanlı Zemberek[2] kütüphanesi kullanılmıştır. Oluşturulan dosya üzerinde gerçekleştirilen DDİ adımları ve dosyanın özellikleri:

- Birleşim metinleri bölütlere (*token*) ayrılmıştır. Her bir bölüt bir kelimeyi ifade eder ve sözlükteki en küçük atomik yapıdır.
- Zemberek kütüphanesinden yaralanarak bölüt tipleri çıkarılmıştır. Tarih, Sayı, Noktalama İşareti bölüt tipine ait olan metin analizinde kullanılmayan sözcükler filtrelenerek derlem dışında tutulmuştur.
- Veri seti oluşturulurken bağlaçlar, zamirler gibi anlamsal değeri olmayan ve metin analizinde kullanılmayacak sözcükler derlemden çıkarılmıştır. Türkçe dilinde en çok kullanılan yaklaşık 300 çeşit Türkçe durak kelimesi (stop words) listesi oluşturulmuş ve bu listedeki kelimelere bir yazılımla derlemden temizlenmiştir.
- Bölütlerden oluşan TBMM birleşim dosyaları bir araya getirilerek art arda eklenmiş kelimelerden oluşan yeni bir metin dosyası elde edilmiştir. Bu dosyanın boyutu 718 MB'dır.

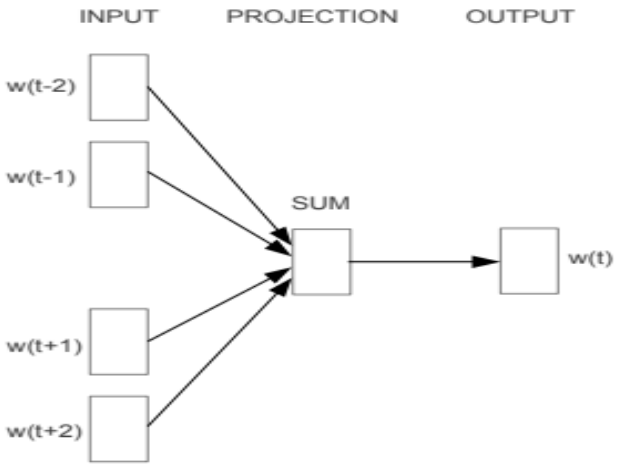
2.3. Kelime Vektörlerinin Çıkarılması (Extraction of Word Vectors)

Kelimelerin vektör olarak temsili, semantik anlamını içerdiği için DDİ çalışmalarında sıkça kullanılır. Kelime vektörleri içinde bulunduğu bağlamsal anlamı içerir ve daha az boyutlu bir temsil sağlayarak hesaplama karmaşıklığı maliyetinden kazanç sağlar.

2.3.1. Word2vec[3]

Mikolov ve arkadaşlarının Word2vec modeli [5] basitçe Harris [1]'in anlam olarak benzer kelimelerin benzer bağlamlarla birlikte bulunacağı mantığına dayanır.

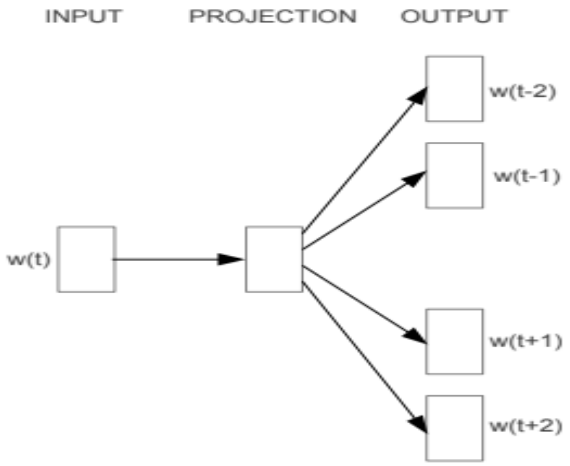
Bu çalışmada kelime vektörlerinin oluşturulması için Word2vec modeli kullanılmıştır. Word2vec modeli, CBOW (Continous Bag of Words) ve Skip-gram olarak adlandırılan iki farklı dil modelinden oluşur. Her iki modelde de bir pencere içindeki kelimeler derlem (corpus) boyunca kayar, her bir adımda yapay sinir ağı pencere içindeki kelimelerle eğitilir.



Şekil 1. Mikolov ve arkadaşlarının çalışmalarında [5] Şekil 1'de tanıttıkları Word2vec CBOW modeli
(Word2vec CBOW model from Mikolov at al.[5] in Figure 1)

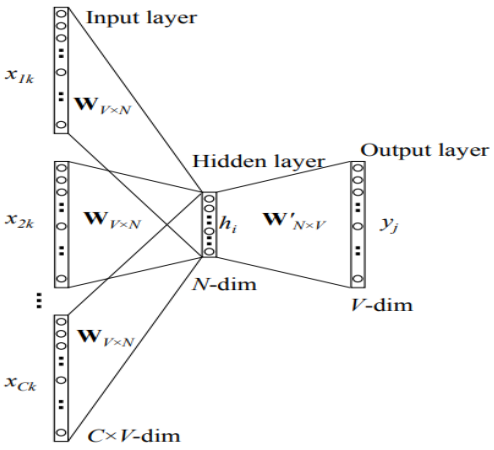
Şekil 1'de gösterilen CBOW modeli derlem boyunca her bir kelimenin bu pencerenin ortasında bulunup bulunmadığını tahmin eder. CBOW modelinde, derlemdeki her bir kelime için; sinir ağına girdi olarak verilen bağlamın (Şekil 1'de V_{t-1}, V_{t-2}, V_{t+1}, V_{t+2}) ortasındaki kelime olan V_t olma ihtimali hesaplanır ve sistem doğru kelimeyi bulması için eğitilir.

[1] https://www.tbmm.gov.tr/psozluk.htm
[2] https://github.com/ahmetaa/zemberek-nlp
[3] https://code.google.com/archive/p/word2vec/

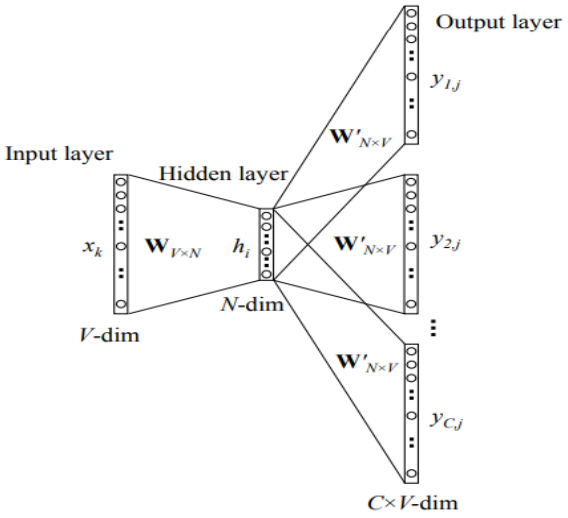


Şekil 2. Mikolov ve arkadaşlarının çalışmalarında [5] Şekil 1'de tanıttıkları CBOW ve Skip-gram modeli
(Word2vec Skip-gram model from Mikolov at al.[5] in Figure 1)

Şekil 2'de gösterilen Skip-gram modelinde ise pencerenin ortasındaki kelimeden, etrafındaki kelimeler (bağlamı) tahmin edilmeye çalışılır. Skip-gram Modeli CBOW modelinin tersidir. Sinir ağına verilen girdi V_t vektöründen bağlam (V_{t-1}, V_{t-2}, V_{t+1}, V_{t+2}) tahmin edilmeye çalışılır. Derlemdeki her bir kelimenin etrafında, bağlamdaki kelimelerin (V_{t-1}, V_{t-2}, V_{t+1}, V_{t+2}) olma ihtimali hesaplanır ve sistem doğru bağlamı bulması için eğitilir. Mikolov ve arkadaşları [17] çalışmalarında Skip-gram modeli için eklentiler sunmuşlardır. Eğitim hızını artırmak için sık geçen kelimeler için alt örneklem seçermişlerdir. "Hiyerarşik softmax" yerine negatif örneklem adını verdikleri alternatif bir yöntem kullanarak kelime vektörlerinin kalitesini artırmışlardır.



Şekil 3. Xin'in Word2vec modelini açıkladığı çalışmasında [18] Şekil 2'de CBOW modelinin sinir ağı yapısı
(Neural Network structure of CBOW model explained in Xin [18] in Figure 2)



Şekil 4. Xin'in Word2vec modelini açıkladığı çalışmasında [18] Şekil 3'de Skip-gram modelinin sinir ağı yapısı
(Neural Network structure of Skip-gram model explained in Xin [18] in Figure 3)

Word2vec yalnızca bir tane saklı katmandan oluşan derin olmayan yapay sinir ağı yapısındadır. Şekil 3'de CBOW modelinin Şekil 4'de ise Skip-gram modelinin sinir ağı yapısı gösterilmiştir. Boyutu derlemdeki sözcük sayısı (V) kadar olan vektörlerden, C adeti (pencere boyu) girdi vektörü olarak kullanılır. Pencere boyutu bağlam olarak da düşünülebilir. Bir kelimeyi temsil etmesi için isteğe bağlı olarak seçilen ve yapay sinir ağı için bir hiper-parametre olan N kelime vektörlerinin boyutunu belirler. N boyutlu bir kelime vektöründe N tane öznitelik olduğu düşünülebilir. Saklı katmanda N tane nöron vardır. Girdi katmanı ve saklı katman arasında [VxN] boyutlu ağırlık matrisi W; saklı katman ve çıktı katmanı arasında [NxV] boyutlu $W^{'}$ ağırlık matrisi bulunur. Çıktı katmanın sonucunda derlemdeki her bir kelimenin hedef kelime olma ihtimalini gösteren değerlerden oluşan V boyutlu vektör elde edilir. Bu vektör hedef kelime ile karşılaştırılarak sinir ağı geri yayılımlı olarak eğitilir.

Girdi katmanında V boyutla temsil edilen kelime vektörleri, ağırlık matrislerinde boyutları indirgenerek, N boyutla olarak temsil edilebilir hale gelir.

Yapay sinir ağının eğitilmesi sonucunda boyutu NxV olan ağırlık matrisleri elde edilir. CBOW modelde gizli katman ve çıktı katman arasındaki $W^{'}$ ağırlık matrisinden kelime vektörleri elde edilir. Skip-gram modelde girdi katmanı ve gizli katman arasındaki W ağırlık matrisinden kelime vektörleri elde edilir. Bu kelime vektörleri anlam benzerliğine göre vektör uzayında bir birine yakın konumlanırlar ve semantik bilgi içerirler.

TBMM Genel Kurul tutanaklarında anlamsal yakınlık elde etmek için Word2vec modeli, python dilinde Tensorflow[4] uygulama çatısı ile gerçekleştirilmiştir.

Uygulamada, yapay sinir ağına girdi olarak TBMM birleşimlerinden oluşan ve DDİ adımları uygulanmış metin dosyası verildi. Yapay sinir ağının eğitilmesi sonucunda çıkan kelime vektörlerinin boyutu 100'dür. Derlemde 5'den az sayıda geçen kelimeler göz ardı edilmiştir.

2.3.2. GloVe[5]

GloVe-GlobalVectors [9], Word2vec'ten sonra en çok kullanılan kelime yerleştirme algoritmasıdır. Pennington ve arkadaşları[9], kelime yerleştirme algoritmalarını LSA gibi global matris ayrışması ve Word2vec gibi lokal bağlam penceresi metodu olmak üzere ikiye ayırmışlardır. Pennington'a göre LSA istatistiksel veriyi verimli şekilde değerlendirirken analojileri ortaya çıkarmada Word2vec'e göre daha az başarılıdır. Word2vec ise analojilerde başarılıdır ama kelimelerin bir derlemde birlikte geçmesini gösteren global birlikte geçme sayısı gibi istatiksel bilgiyi ancak dahili olarak kullanır, yeteri kadar verimli kullanmaz. Bunun yerine derlem ayrı bir lokal pencere bağlamında eğitilir. GloVe modeli bu iki modelin iyi yanlarını birleştirdiğini iddia eder. GloVe, global istatistiksel veri üzerinde çalışır ama Word2vec ile analojiler gibi aynı semantik yapıları ortaya çıkarır. Model global kelime-kelime birlikte geçme sayısı ile eğitildiği için istatistiki veriyi harici olarak açık bir şekilde kullanır.

GloVe modelindeki benzerlik, kelimelerin birlikte geçme ihtimalinden bulunur. Pennington ve arkadaşları[9] bunu güzel bir örnekle açıklamışlardır. *Buz (ice)* ve buhar *(steam)* hedef kelimelerinin 6 milyar bölütten oluşan derlemden seçilen bağlam kelimeleri katı *(solid), gaz (gas), su (water), biçim (fashion)* ile ilişkisi değerlendirilmiştir.

Tablo 1. Pennington ve arkadaşlarının makalesinden[9] kelimelerin birlikte geçme ihtimallerini gösteren değerler
(Word co-occurrence probabilties from Pennington at al.[9] in Table 1)

Probability and Ratio	k = solid	k = gas	k = water	k = fashion
P(k | ice)	$1{,}9 \times 10^{-4}$	$6{,}6 \times 10^{-5}$	$3{,}0 \times 10^{-3}$	$1{,}7 \times 10^{-5}$
P(k | steam)	$2{,}2 \times 10^{-5}$	$7{,}8 \times 10^{-4}$	$2{,}2 \times 10^{-3}$	$1{,}8 \times 10^{-5}$
P(k | ice) / P(k | steam)	8,9	$8{,}5 \times 10^{-2}$	1,36	0,96

Buz ve *buhar* arasındaki ilişki k adet bağlam kelimelerden her biri ile birlikte geçme olasılıklarının oranları ile bulunmuştur. Örneğin maddenin halini belirten *katı* kelimesi *buz* ile *buhara* göre çok daha yakından ilişkilidir ve Tablo 1'in üçüncü satırındaki oran çok büyük çıkmıştır. Maddenin halini belirten g*az* kelimesi ise *buhar* ile buza göre yakından ilişkilidir ve oran çok küçük çıkmıştır. Bağlam kelimeleri incelenen kelimelere anlamsal olarak eşit yakınlıkta veya ilgisiz ise oran bire yakın olacaktır. *Su* kelimesi *buz* ve *buhar* kelimesi ile anlamsal olarak yaklaşık aynı anlamda olduğu için oran bire yakındır, yine *biçim* kelimesi *buz* ve *buhar* ile ilgisiz olduğu için bu oran bire yakın olacaktır. Oran bir biri ile ilgili kelimeleri (katı ve gaz) ilgisiz kelimelerden (su ve biçim) ayırır, ayrıca iki ilişkili kelime arasında da ayrım yapabilir.

Birlikte geçme ihtimallerinin oranını yukardaki örnekteki gibi *buz, buhar* kelime vektörlerini ve k bağlam kelimelerini (*katı, gaz, su, biçim*) girdi olarak olan bir F fonksiyonu ile tanımlarsak

$$F\left(w_i, w_j, \tilde{w}_k\right) = \frac{p_{ik}}{p_{jjk}} \quad (1)$$

F Fonksiyonunda w_i ve w_j kelime vektörlerini, $\tilde{w}_k$ ise bağlam kelime vektörünü belirtir. $\frac{p_{ik}}{p_{jk}}$ Oranı w_i kelimesinin $\tilde{w}_k$ bağlamında geçme olasılığının w_j kelimesinin $\tilde{w}_k$ bağlamında geçme olasılığına oranıdır. Bu oranda temsil edilen bilgiyi kelime vektör uzayında göstermek için vektörlerin doğrusal yapısından dolayı çıkarma işlemi yapılır.

$$F\left(w_i - w_j, \tilde{w}_k\right) = \frac{p_{ik}}{p_{jjk}} \quad (2)$$

F Fonksiyonun girdileri vektörel iken denklemde eşitliğin sol tarafı, fonksiyonun sonucu eşitliğin sağ tarafı skaler büyüklüktür. Bu durumdan kurtulmak için iç çarpım işlemi uygulanır.

$$F\left(\left(w_i - w_j\right)^T \tilde{w}_k\right) = \frac{p_{ik}}{p_{jjk}} \quad (3)$$

Birlikte geçme matrisi düşünüldüğünde bir kelime ve bir bağlam kelimesi arasındaki ayrım isteğe bağlıdır ve bu iki rolü denklemde değiştirebiliriz. Tutarlılık için değiştirme işlemi sadece kelime vektörleri $w \leftrightarrow \tilde{w}_k$ arasında değil kelimelerin derlemde geçme sıklığı $X \leftrightarrow X^T$ arasında da yapılmalıdır. Kelime sıklık matrisleri simetriktir ve F

[4] https://www.tensorflow.org/
[5] https://github.com/stanfordnlp/GloVe

Fonksiyonun simetrik olması için benzer biçimlilik gereklidir.

$$F\left(\left(w_i - w_j\right)^T \tilde{w}_k\right) = \frac{F\left(w_i^T \tilde{w}_k\right)}{F\left(w_j^T \tilde{w}_k\right)} \quad (4)$$

Denklemi çözüldüğünde,

$$F\left(w_i^T \tilde{w}_k\right) = P_{ik} = \frac{X_{ik}}{X_i} \quad (5)$$

Sonucu elde edilir. Yukardaki denklemde X_i, i indeksindeki kelimenin derlemde toplam geçme sayısını, X_{ik} ise i indeksindeki kelimenin k bağlamında birlikte geçme sayısını gösterir. F fonksiyonu üstel kabul edilirse denklem aşağıdaki gibi olur.

$$w_i^T \tilde{w}_k = \log P_{ik} = \log X_{ik} - \log X_i \quad (6)$$

X_i Bağlama bağlı değildir bu yüzden yerine b_i terimi, denklemin simetriyi koruması için $\tilde{b}_k$ terimi eklenir.

$$w_i^T \tilde{w}_k + b_i + \tilde{b}_k = \log X_{ik} \quad (7)$$

Yukardaki denkleme göre optimize edilecek Hata Fonksiyonu J bulunur.

$$J = \sum_{i,j=1}^{V} f\left(X_{ij}\right)\left(w_i^T \tilde{w}_k + b_i + \tilde{b}_k - \log X_{ik}\right)^2 \quad (8)$$

Denklemdeki $f\left(X_{ij}\right)$ manuel olarak tanımlanan ağırlık fonksiyonudur. V sözlüğün boyutudur.

TBMM Genel Kurul tutanaklarında anlamsal yakınlık elde etmek için GloVe modeli, python dilinde Tensorflow uygulama çatısı ile gerçekleştirilmiştir.

Uygulamada, TBMM birleşimlerinden oluşan ve DDİ adımları uygulanmış metin dosyası kullanılarak GloVe modeli ile elde çıkan kelime vektörlerinin boyutu 50'dir. Derlemde 5'den az sayıda geçen kelimeler göz ardı edilmiştir. GloVe modeli ile 226516 kelime vektörü elde edilmiştir.

3. BULGULAR (RESULTS)

Bu çalışmada, TBMM Genel Kurulu Tutanaklarından elde edilen derlem üzerinden Word2vec modeli ve GloVe modeli ile kelime vektörü çıkarılmıştır. Vektörler arasındaki benzerlik kosinüs benzerliği ile bulunmuştur. Kosinüs benzerliği iki vektör arasındaki açının kosinüs değerine göre hesaplanır. Dağılım hipotezine[1] göre vektör uzayında bir birine yakın konumlanan vektörler yakın anlamlıdır. Vektör uzayında bir birine yakın konumlanan vektörlerin arasındaki açı küçük olacaktır. Kosinüs açısının değeri [-1,1] Aralığında gerçekleşir. Word2vec ve GloVe modellerinde kelime vektörleri pozitif olur ve birim çemberde birinci bölgede konumlanır. Bu yüzden kosinüs benzerliğinin değeri [0,1] Aralığında gerçekleşir. Kosinüs benzerliği kavramlar arasındaki yakınlığın ölçüsü olarak kullanılmıştır.

Kelime vektörleri üzerinden yapılan analizde anlam olarak benzer kavramların vektör uzayında bir birine yakın dağıldığı görülmüştür.

3.1. Kelime vektörleri ile bulunan yakın anlamlı kavramlar

Tablo 2'de "güzel" kelimesi için bulunan yakın kavramlar ve bu kavramların vektör uzayında "güzel" kelimesine olan kosinüs mesafeleri gösterilmiştir. Türk Dil Kurumu[6] sözlüğünde "güzel" kelimesi "iyi" ile eşanlamlı, "mükemmel" sözcüğü "çok güzel" ile eşanlamlı olarak belirtilmiştir ve çalışmamızda bulunan sonuçlarla uyumludur.

Word2vec ve GloVe modeli ile bulunan sonuçların karşılaştırılması için gerekli veri tabanları Türkçede mevcut olmadığı için değerlendirme manuel olarak yapılmıştır.

Tablo 2. "güzel", "ergenekon" ve "obama" kelime vektörleri için Word2vec ve GloVe ile bulunan anlam olarak yakın kavramlar
(Similar conceptsand cosine similarities for "güzel", "ergenekon" and "obama" words extracted by Word2vec and Glove)

Kavram	**Yakın anlamlı kavramlar ve Kosinüs Benzerliği**	
	Word2vec	**GloVe**
güzel	**Mükemmel** 0.73329819	Alıkoymayı 0.78158297
	gayet_güzel 0.69489960	Oh 0.74073260
	İyi 0.67330687	**İyi** 0.73799612
	Güzeldir	**Mükemmel**

[6] http://www.tdk.gov.tr/index.php?option=com_gts&view=gts

Kavram	Yakın anlamlı kavramlar ve Kosinüs Benzerliği	
	Word2vec	**GloVe**
	0.66299429	0.71741312
	Faydalı 0.62949656	Hakikaten 0.71409937
	Muhteşem 0.59376149	**Muhteşem** 0.70438485
	Anlamlı 0.58875867	İlim 0.68872337
	Hakikaten 0.58893529	Anlattı 0.68638473
	Hayırlı 0.58875867	Güzeldir 0.68550220
	Dörtlük 0.58875867	Gerçekten 0.68392926
ergenekon	**Balyoz** 0.89680099	**balyoz** 0.89991216
	ergenekon_balyoz 0.87643147	**Hrant** 0.83259876
	Kck 0.85306293	**Kck** 0.82672301
	balyoz_ergenekon 0.84119211	**Dink** 0.82174639
	oda_tv 0.82395034	Suikast 0.78199251
	Kumpas 0.81534251	Gezi 0.76754247
	ergenekon_davası 0.81474625	Cinayet 0.74112025
	balyoz_davasında 0.80395877	Feneri 0.73158987
	askeri_casusluk 0.79091995	Gazeteci 0.72490641
obama	**Bush** 0.88371944	**Bush** 0.84964052
	Clinton 0.88318982	**Washington** 0.81586959
	Putin 0.83955807	**Clinton** 0.79838075
	Davos 0.83339022	Esad 0.76817758
	Merkel 0.82871606	Putin 0.76419517
	Davutoğlu 0.82650493	pyd 0.72723679
	dışişleri_bakanının 0.82461460	Barzani 0.72666409
	Talabani 0.80264673	Amerika 0.71037485
	Washington 0.79649092	Beşar 0.70136243
	Esad 0.76653920	Başbakanla 0.69448643

Tablo 2'de incelenen kavrama anlam olarak yakın olduğu değerlendirilen kavramlar koyu renklidir. "Güzel" kelimesi için Word2vec modelinin benzerlik bulduğu en yakın kavramlardan ilk üçü "mükemmel", "gayet_güzel", "iyi" yakın anlamlı iken GloVe modeli için ancak üçüncü kavram "iyi" yakın anlamlıdır.

"Ergenekon" kavramı için hem Word2vec modeli hem Glove modeli ile yakın anlamlı kavramlar başarılı bir şekilde bulunmuştur. Kavram, TBMM Genel Kurul tutanaklarına uygun şekilde "balyoz", "askeri_casusluk", "hrant" gibi kavramlarla yakın anlamlı bulunmuştur.

"Obama" kelime vektörü ile benzerlik analiz edildiğinde her iki modelde de bir önceki ABD başkanları her iki modelde de bulunmuştur.

Tablo 3. Örnek kavramların Word2vec ve GloVe modellerinin bulduğu anlamsal olarak yakın kavramlar
(Sematically similar concepts for sample words extracted by Word2vec and Glove)

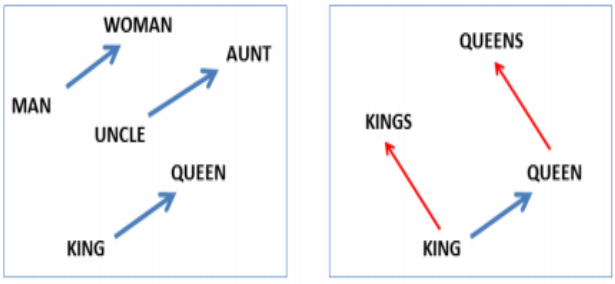
Kavram	Yakın anlamlı kavramlar	
	Word2vec	**GloVe**
suriye	**'ırak' 'kuzey_ırak'** 'pyd' 'irak_suriye' 'suriye_irak'	**'ırak'** 'ermenistan' 'kuzey' 'esad' 'kerkük'
oecd	**'eurostat' 'dünya_bankasının' 'g20'** 'oecd_ülkelerinde'	'sıradayız' 'ortalaması' 'birliğinde' 'kıyaslandığında' 'düzeyindedir'
paris	'viyana' 'new_york' 'berlin' 'atina' 'londra'	'york' 'cenevre' 'new' 'roma' 'viyana'
merkel	**'obama' 'putin' 'sarkozy'** 'chirac' 'clinton'	'başbakanı' **'berlusconi' 'schröder' 'sarkozy' 'lipponen'**
galatasaray	**'fenerbahçe' 'trabzonspor'** 'dünya_şampiyonu' 'maçında'	**'hacettepe' 'baysal'** 'üniversitesi' **'erciyes' 'gaziosmanpaşa'**
fenerbahçe	**'galatasaray' 'trabzonspor'** 'şampiyonlar_ligi' **'beşiktaş'**	**'uefa'** 'takımını' **'beşiktaş'** 'basketbol'
brics	**'sahra_afrika' 'yükselen_ekonomiler' 'brezilya_rusya' 'hindistan_brezilya'**	'salatalıkta' 'tüketicilerini' 'standardındaki' 'masumla' 'gelişmişliklerini'
dolar	'dolar_civarında' 'doların_üzerinde' 'dolarlık' 'doları' 'dolara_yakın'	'dolarlık' 'doları' 'dolara' 'dolarla' 'dolardan'
pisa	**'timss' 'öss'** 'lisans_yerleştirme' **'oecd'** 'üniversiteye_giriş'	'yaparlardı' **'okulöncesi'** 'sınavlarda' **'öss'** 'gerilere'

Kavram	Yakın anlamlı kavramlar	
	Word2vec	GloVe
şubat	'aralık' 'kasım' 'temmuz' 'mart' 'nisan'	'haziran' 'ocak' 'kasım' 'nisan' 'mart'
mavi	**'siyah'** 'boğazı' 'koyu' 'kuşlar' 'boğaz'	**'akım'** 'hattı' 'gazı' 'körfez' 'nabucco'
elma	'domates' 'soğan' 'sebze' 'patates' 'karpuz'	'patates' 'üzüm' 'narenciye' 'kayısı' 'meyve'

Örnek kelime vektörleri ve bu vektörlere yakın kelime vektörlerinin kosinüs benzerlik değerleri Tablo 3'de gösterilmiştir. Bazı kavramlar için bulunan en yakın 10 kavram arasından doğrudan ilişki kurulabilir kavramlar Tablo 3'de gösterilmiştir. "brics" ve "oecd" kavramı için Word2vec anlamlı sonuçlar bulmuştur fakat GloVe anlamlı sonuçlar bulamamıştır. Word2vec modelinin daha başarılı olduğu örneklerle görülmekle birlikte iki modelin bazı kavramları farklı bağlamlarda değerlendirdiği görülmüştür. "Galatasaray" kavramı incelendiğinde Word2vec modeli ile bulunan en yakın 10 kavramın hepsi spor ile ilgilidir, GloVe modelinde ise en yakın 10 kavram üniversite ile ilgilidir. "Galatasaray" kavramı "Galatasaray Spor Kulübü" olarak düşünüldüğünde Word2vec, "Galatasaray Üniversitesi" olarak düşünüldüğünde GloVe başarılı sonuçlar bulmuştur. Aynı şekilde "mavi" kavramını Word2vec renk ile ilgili kavramlarla birlikte değerlendirmiş, GloVe ise bir boru hattı projesi olan "Mavi Akım" projesi ile birlikte değerlendirmiştir. Özellikle Bilgi Toplama uygulamaları düşünüldüğünde iki model bazı kavramlar için bir birinin eksikliklerini tamamlar özelliktedir ve sonuçları birlikte kullanılabilir.

3.2. Kelime Vektörleri ile Çıkarılan Analojiler (Analogies extracted by Word Vectors)

Mikolov ve arkadaşları yaptıkları çalışmada[19], kelime vektörleri ile basit cebirsel işlem yaparak sadece kelimeler arasındaki benzerliği değil analoji adı verdikleri karmaşık semantik ilişkiler bulmuşlardır.



Şekil 5. Mikolov'un çalışmasında [19] Şekil 2'de gösterdikleri kelime çiftlerinin vektör ofsetleri
(Vector offsets of Word vectors from Mikolov at al. [19] in Figure 2)

Şekil 5'de kelime çiftlerinin vektör uzayında dağılımından anlamlı sonuçlar bulunabilir. Şekilde cinsiyet ilişkisini gösteren kelime çiftlerinde iki kelime vektörü arasındaki uzaklık eşittir. Sağda ise kelimelerin tekil çoğul ilişkisi vektörlerin arasındaki mesafeden bulunabilir. *man*, *woman* ve *uncle* kelime vektörlerini vektör uzayındaki konumlarından *aunt* kelimesinin *uncle* kelimesi ile cinsiyet ilişkisi bulunabilir.

Çalışmalarında 2 kelime çiftinden oluşan kayıtlardan bir veri seti oluşturmuşlardır ve sonuçları bu veri seti ile değerlendirmişlerdir. Mikolov, Vector("King")-Vector("man") + Vector("woman") cebirsel işleminin Vector("queen")'e en yakın sonucu verdiğini göstermiştir [16]. Böylece kelime vektörlerinin vektör uzayında bulunduğu konumlar arasındaki mesafeye göre analojiler bulunabilmektedir. Analojiler TBMM Genel Kurul Tutanakları veri seti üzerinde de araştırılmıştır. Aynı örnek Türkçe olarak yapılmış ve tutanak veri setine özgü benzer örnekler de çalışılmıştır.

Mikolov'un çalışmasında kullandığı örnek analoji, TBMM tutanak veri setinde de araştırılmıştır. Vektör uzayında "kral" ve "erkek" arasındaki analojinin "kadın" ve "kraliçe" arasında da olup olmadığı araştırılmıştır. Sonuç olarak Tablo 4'de gösterildiği gibi en yakın analoji "kraliçe" olarak bulunamamış Word2vec için "mandela", Glove için "dikta" bulunmuştur. Ancak TBMM tutanak veri setinde bulunabilecek analojiler araştırıldığında Tablo 4'de görüldüğü gibi anlamlı sonuçlar ortaya çıkmıştır.

Vector("King")-Vector("man")+Vector("woman") =Vector("queen")

Word2vec: Vektör("kral") -Vektör("erkek") +Vektör("kadın") = Vektör("mandela")

GloVe: Vektör("kral") -Vektör("erkek") +Vektör("kadın") = Vektör("dikta")

Tablo 4. TBMM tutanaklarında bulunan analojiler
(Analogies extracted from TBBM Parliamentary minutes)

Analoji	Word2vec	GloVe
'kral' + 'kadın' - 'erkek'	'mandela' 0.23787996	'dikta', 0.24811934
	'yeltsin' ,0.23674818	'diktatör' 0.24671753
	'padişah' 0.23628864	'ses' 0.24630152
'obama' + 'almanya' - 'abd'	'merkel' ,0.26317449	'merkel' 0.32031258
	'clinton' 0.26194161	'lipponen' 0.29804642
	'sarkozy' 0.24770099	'sarkozy' 0.2925140
	'cumhurbaşkanıyla' 0.24436413	'clinton' 0.28724876
'rusya' + 'berlin' - 'moskova'	'rusya_federasyonu' 0.23262998	'fransa' 0.25353153
	'hırvatistan' 0.23214797	'almanya' 0.24692050
	'ukrayna'	'batı'

Analoji	Word2vec	GloVe
	0.22970476	0.24322375
'rusya' + 'şam' - 'moskova'	'irak_suriye' 0.27572644	'suriye' 0.32036156
	'saddam' 0.27500082	'yemen' 0.31062516
	'suriye' 0.27361400	'irak' 0.31006093
'bahçeli' + 'chp' - 'mhp'	'baykal' 0.29765808	'özal' 0.27253096
	'genel_başkanınız' 0.29068364	'baykal' 0.27210178
	'başbakanınız' 0.28135544	'ecevit' 0.27121085
'almanya' + 'yunan' - 'alman'	'yunanistan' 0.29196318	'yunanistan' 0.3152612
	'bulgaristan' 0.28863894	'fransa' 0.3074294
	'ürdün' 0.26909002	'bulgaristan' 0.27350110

Tablo 4'de, Word2vec ve GloVe modellerinin her ikisinde de "obama" ve "abd" arasındaki ilişkinin "almanya" ve "merkel" arasında da bulunduğu görülecektir. Word2vec modelinde "bahçeli" ile "mhp" arasındaki ilişki "chp" ve "baykal" arasında bulunmuştur. Analoji 1994 yılı ve 2016 yılı arasındaki Genel Kurul tutanakları veri setinde kullanıldığı düşünüldüğünde anlamlıdır. Ancak GloVe modelinde "baykal" kelimesi ikinci en yakın kavram olarak bulunmuştur. "Yunanistan" ve "yunan" arasındaki ilişki "almanya" ve "alman" arasında her iki model için de başarı ile bulunmuştur.

4. SONUÇLAR (CONCLUSION)

Bu çalışmada Word2vec modeli ve GloVe modeli ile TBMM Genel Kurul tutanaklarından oluşturulan derlemden kelime vektörleri çıkarılmış ve anlamsal olarak bir birine yakın kavramlar bulunmaya çalışılmıştır. Sonuçlar değerlendirildiğinde Word2vec modelinin daha iyi sonuçlar bulduğu gözlemlenmiştir.

Benzer çalışmalarda kelime vektörü modellerinin bir kavramı bir birinden tamamen farklı bağlamlarda değerlendirdiği örneklere rastlanmamıştır. Bu çalışmada Word2vec ve GloVe modellerinin her ikisi de kullanılarak; aynı kavramın farklı bağlamlarda değerlendirildiği sonuçlar elde edilmiştir. Bu bir kavramın kullanıldığı farklı bağlamların bulunabilmesi için Word2vec modelinin ve GloVe modelinin sonuçlarının birbirinden ayrı değerlendirilmesi gerektiğini göstermiştir. Örneğin "galatasaray" kavramını, Word2vec "spor" bağlamında kullanılan kavramlarla birlikte değerlendirilmiş, GloVe ise "üniversite" bağlamında kullanılan kavramlarla birlikte değerlendirilmiştir.

Çalışmada, her iki modelin bir kavram için bulduğu belli sayıdaki en yakın kavramlar değerlendirilerek, kavramın ilişkili olduğu farklı bağlamların bulunabileceği görülmüştür. Bu sonuçlar, özellikle Bilgi Çıkarımı uygulamaları için iki modelin çıktılarının birlikte kullanılması gerekliliğini göstermiştir.

Bu çalışmada her iki modelle de kavramlar arasındaki analojiler tespit edilmek istenmiştir. Literatürde sıkça karşılaşılan (ülke-başkent), (ülke-konuşulan dil), (ülke-devlet başkanı) ikilileri ile bulunabilecek analojiler TBMM Genel Kurul Tutanakları derleminde her iki model ile de bulunmuştur. Bu derleme özgü (siyasi parti-genel başkanı) gibi ikililere ait analojiler için ise Word2vec modeli, GloVe modeline göre daha iyi sonuçlar göstermiştir.

Word2vec ve Glove modelleri kullanılarak TBMM Genel Kurul Tutanaklarından elde edilen kelime vektörlerinin bir kavramın semantik anlamını başarı ile temsil ettiği görülmüştür. Her iki modelin de TBMM Genel Kurul Tutanaklarının sınıflandırılması, bilgi çıkarımı gibi doğal dil işleme uygulamalarında kullanılabileceği düşünülmektedir.

KAYNAKLAR (REFERENCES)

[1] Z. Harris, "Distributional structure", *Word*, 23(10), 146–162, 1954.

[2] Thomas K. Landauer , Susan T. Dumais, "A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge", *Psychological Review*, 104(2), 211–240, 1997.

[3] X. Hu, Z. Cai, P. Wiemer-Hastings, A. Graesser, D. McNamara, **Strengths, limitations, and extensions of LSA. Handbook of Latent Semantic Analysis**, 401–426, 2007.

[4] R. Collobert , J. Weston, "A Unified Architecture for Natural Language Processing: Deep Neural Networks with Multitask Learning", **Proceedings of the 25th International Conference on Machine Learning**, Helsinki, Finlandiya, 20(1), 160–167, 2008.

[5] T. Mikolov, K Chen, G Corrado, J Dean, "Efficient estimation of word representations in vector space", *arXiv:1301.3781*,2013.

[6] Y. Bengio, R. Ducharme, P. Vincent, C. Janvin, "A neural probabilistic language model", *Journal of Machine Learning Research*, 3, 1137-1155, 2003.

[7] L. Jianqiang, L. Jing, F. Xianghua, M.A. Masud, J.H. Huang, "Learning distributed Word representation with multi-contextual mixed embedding", *Knowledge-Based Systems*, 106, 220-230, 2016.

[8] O. Kaynar, Z. Aydın, Y. Görmez. "Sentiment Analizinde Öznitelik Düşürme Yöntemlerinin Oto Kodlayıcılı Derin Öğrenme Makinaları ile Karşılaştırılması", *Bilişim Teknolojileri Dergisi*, 10(3), Temmuz 2017.

[9] J. Penington, R. Socher, C.D. Manning, "GloVe:Global Vectors for Word Representation", **Emprical Methods in Natural Language Processing (EMNLP)**, 1532-1543, 2008.

[10] E. Altszyler, M. Sigman, S. Ribeiro, D. F. Slezak. "Comparative study of LSA vs Word2vec embeddings in small corpora: a case study in dreams database", *arXiv: 1610.01520*, 2016.

[11] L.O,Goldberg, Y,Dagan, "Improving distributional similarity with lessons learned from Word embeddings", *Transactions of the Association for Computational Linguistics*, 3, 211-225, 2015.

[12] M. Naili, A. H. Chaibi, H. H. B. Ghezala, "Comparative study of word embedding methods in topic segmentation", *Procedia Computer Science*, 112, 340-349, 2017.

[13] M. Faruqui, J. Dodge, S. K. Jauhar, C. Dyer, E. Hovy, N. A. Smith, "Retrofitting word vectors to semantic lexicons", **In Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics**, Colorado, ABD, Human Language Technologies, 1606– 1615, 2015.

[14] G. A. Miller, "Wordnet: a lexical database for english", *Communications of the ACM*, 38(11), 39-41, 1995.

[15] C. F. Baker, C. J. Fillmore, J. B. Lowe, "The Berkeley FrameNet Project", **Proceedings of the 17th International Conference on Computational Linguistics**, Volume 1, Montreal, Quebec, Kanada, 86-90, 1998.

[16] J. Ganitkevitch, B. Van Durme, C. Burch, "PPDB: The paraphrase database", **Proceedings of NAACL**, 758-764, Haziran, 2013.

[17] T. Mikolov,I. Sutskever, K. Chen, G. Corrado, J. Dean, "Distributed Representations of Words and Phrases and their Compositionality", **Proceedings of the 26th International Conference on Neural Information Processing Systems**, Volume 2, Nevada, ABD, 3111-3119, 2013.

[18] X. Rong, "Word2vec Parameter Learning Explained", *arXiv:1411.2738*, 2014.

[19] T. Mikolov, W. Yih, G. Zweig, "Linguistic Regularities in Continuous Space Word Representations", **Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies,** 746-751, 2013.