Makine Öğrenmesi Destekli Anlamsal Ürün Arama Motoru
Machine Learning Supported Semantic Product Search Engine

Simge Sarıçayır

***Özetçe*— Günümüzde e-ticaret siteleri oldukça yaygın şekilde kullanılmaktadır. İnsanların bu siteleri kullanarak alışveriş yapması sebebiyle sitenin sunduğu arama motorunun ne kadar iyi sonuçlar verdiği kilit noktayı oluşturmaktadır. Listelenen ürünlerin anlamsal yakınlıkta olması ve bu ürünlerin doğru şekilde sıralanması sitenin tercih edilmesinde önemli rol oynamaktadır. Dolayısı ile satış yapılmasında da belirleyicidir. Bu bildiri ile kelime vektörleri ve makine öğrenmesi teknikleri kullanılarak anlamsal yakınlıkta ürünleri listeleyen bir sanal market arama motoru sunulmaktadır. Deneysel sonuçlarda bir ürünü aramak için kullanılan kelimelerinin anlamsal benzerlikleri yanı sıra kullanıcı tercihlerine dayanan makine öğrenmesi teknikleri kullanılması ile daha olumlu ürün sıralaması yapılabildiği görülmektedir.**

Anahtar Kelimeler — Word2Vec; cbow; kelime vektörleri; sınıflandırma; makine öğrenmesi.

***Abstract*— Nowadays, e-commerce sites are widely used. The performance of search engines used by these sites is the key in people's choice to use these sites. The semantic proximity of the listed products and the correct ranking of these products play an important role in this choice. Therefore, it is also decisive in making sales. This paper presents a virtual market search engine that lists semantically close products using word vectors and machine learning techniques. In the experimental results, it is seen that better product rankings can be obtained by using machine learning techniques based on user preferences besides semantic similarities of words used in a search.**

Keywords — Word2Vec; cbow; Word vectors; classification; machine learning

I. GİRİŞ

Günümüzde internetten alışveriş her geçen gün artmaktadır. E-ticaret sitelerinde alışverişin kolay olması ve daha çok seçenek sunması sebebiyle insanlar alışverişlerinde bu siteleri tercih etmektedir. Arama motorları ise alışveriş esnasında doğru ürünü doğru sırada listeleme konusunda iyi olmadığında sitenin tercih edilebilirliği düşmektedir. Bu çalışmada geliştirilen arama motoru ile listelenen ürünlerin anlamsal yakınlık içermesi ve kullanıcıların tercihlerine dayalı listeleme yapılması hedeflenmiştir.

Bu çalışmada Word2Vec kelime vektörleri ve makine öğrenmesi teknikleri kullanılarak anlamsal yakınlıkta ürünlerin doğru sıralama ile listelendiği bir arama motoru geliştirilmiştir. Kullanıcıların alışveriş geçmişi bilgisi kullanışmıştır. Yaptıkları arama sonucu satın aldıkları ürünler akıllı sıralama yapma noktasında kullanılmıştır.

II. KELİME VEKTÖRLERİ

Kelime vektörleri yöntemi, kelimeleri n boyutlu bir uzayda birer vektör olarak temsil etmek ve bu yol ile kelimeler arası uzaklıkları hesaplayarak aralarındaki anlamsal benzerliği tespit etme amacıyla kullanılmıştır [1].

$$\text{Benzerlik (A, B)} = \frac{A.B}{||A|| * ||B||} \quad (1)$$

Kosinüs benzerliği, n boyutlu iki vektör arasındaki benzerliği iki vektör arasındaki açının kosinüsü ile ifade eder. A ve B vektörlerinin kosinüs benzerliği değeri (Denklem 1), A ve B'nin skaler çarpımının, A ve B'nin mutlak değerinin çarpımına bölünmesi ile elde edilir.

A. Word2Vec Cbow Modeli

Word2Vec kelimeleri vektör uzayında ifade etmek için kullanılan sinir ağı temelli bir yaklaşımdır [2]. Büyük bir metin kümesi ile eğitilen bu model yüksek boyutlu uzayda her bir kelime için benzersiz bir vektör oluşturur. Oluşturulan bu benzersiz vektörlerin özelliği veri kümesindeki benzer anlamdaki kelimelerin birbirlerine yakın vektörler oluşturmasıdır. Word2Vec birer adet girdi, çıktı ve gizli katmandan oluşmaktadır. Kelime vektörlerini oluştururken

pencere genişliği, vektör boyutu gibi parametreler kullanmaktadır.

Cbow ve Skip-gram literatürde sıklıkla kullanılan iki word2vec metodu olarak öne çıkmaktadır [2]. Cbow bir kelimeyi tahmin etmek için kelimeyi çevreleyen bir bağlamı kullanırken, Skip-gram sabit bir pencere boyutuna sahip kelimeleri çevreleyerek kelimeyi tahmin etmeye çalışır. Bu çalışmada Cbow modeli kullanılmıştır. Bunun sebebi ise tahmin edilmek istenen kelimenin eğitim yapılan cümleler içerisindeki kelimeler ile ilişkilendirilmek istenmesidir. Kelimenin geçtiği içerikte her kelime için berlirlenen pencere boyutuna göre kelimeleri ilişkilendirmektedir.

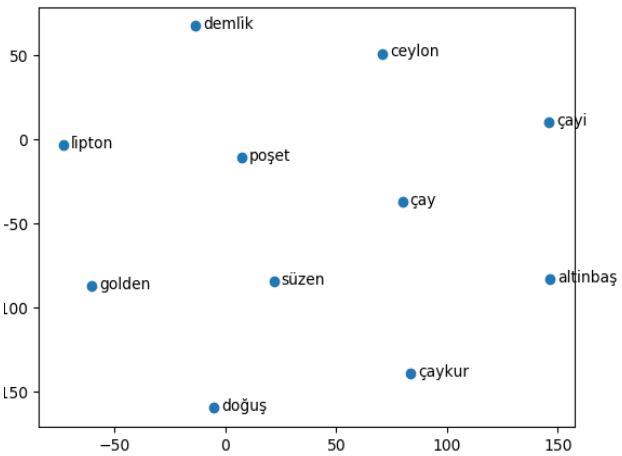
B. Veri Kümesi

Geliştirilen arama motoru için 6528 adet ürün adı kullanılmıştır. Word2Vec modelini çalışmanın amacına uygun şekilde eğitmek için kullanılan ürünler ile alakalı içeriğe sahip veri kullanılması gerekmektedir.

Çalışmalara ilk olarak iki farklı e-ticaret sitesine ait veri kümelerinin eğitim için kullanılması ile başlandı ancak geliştirilmekte olan arama motoruna ait ürünler ile içerik uyuşmazlığı yaşandığından kullanılmadı. Aynı sebeple internette var olan türkçe veri setleri de yetersiz kalmaktadır. Bunun için googlesearch ve beautifulsoup kütüphanesi kullanılarak ürünlerin google'da aratılması ile gelen ilk 5 sayfadan içerik toplanmıştır.

C. Model Eğitimi

Doğal Dil İşleme problemlerinde metinler üzerinde çeşitli ön işlemler gerçekleştirilmelidir. Bu işlemler yapılan uygulamanın doğruluğunu artırmaya yardımcı olduğundan önemlidir. Bunun için internetten toplanan veri setindeki sayılar, noktalama işaretleri ve kısaltmalar temizlenmiştir. Toplanan veri ön işleme yapılarak Word2vec modelinin eğitim için alacağı girdi formatına çevrilmiştir. Bir kelime veri kümesi içerisinde beşten az sayıda geçiyorsa eğitime alınmamıştır. Parametre seçimi için; vektör boyutu olarak 300, pencere sayısı içinse 5 değeri belirlenmiştir.



Şekil 1. Çay Kelimesine Yakın Kelimelerin Gösterimi

Word2Vec modeline verilen bir kelime için en yakın kelimeler alınabilmektedir. Modele 'çay' kelimesi verildiğinde eğitilen veri setinden öğrenime göre Şekil 1'deki kelimeler bulunmuştur. Şekilde görüldüğü üzere 'çay' kelimesine yakın olan kelimeler bu alandaki markalar ve ürünler içerisinde geçen detaylardır.

III. KELİME VEKTÖRLERİ İLE ARAMA MOTORU

Eğitilen word2vec modeli kullanılarak arama motoru geliştirmiştir. Bunun için iki farklı yöntem dikkate alınmıştır. İlk yöntem olarak aranan kelimelerin word2vec modelindeki vektörleri ile ürün içindeki kelimelerin vektörlerinin yakınlıklarına göre benzerliklerini kosinüs ile ifade eden fonksiyon (Denklem 1) kullanılarak bu benzerlik değeri toplanıp ortalaması alındı. Sıralama için benzerlik değeri büyükten küçüğe doğru ürünler sıralandı.

İkinci bir yöntem olarak aranan kelimelerin Word2Vec modelindeki vektörleri ile ürün içindeki kelimelerin vektörlerinin bütün olarak karşılaştırılması ile puanlama yapıldı. Aranan kelimeler ürün içerisinde birebir eşleşiyorsa fazladan puan vererek sıralamada üst sıraya çıkması sağlandı.

İkinci yöntemde ürün bütün olarak kıyaslandığı için birinci yönteme göre daha iyi sonuçlar vermektedir. Çalışmaya ikinci yöntem ile devam edilmiştir.

A. Arama Kelimelerinin Kontrol Edilmesi

Arama yapılan kelime word2vec modelinin sahip olduğu sözcük dağarcığında olduğu durumda benzerlik karşılaştırması yapılabilmektedir. Ancak kullanıcıların yazım yanlışı yapması durumunda herhangi bir sonuç alınamamaktadır. Bu problemin önüne geçebilmek adına ilk önce sonlu durumlu otomat ile kelimelerin Türkçe- İngilizce karakter uyuşmazlığı kontrol edildi. Karakter değişimi yapılan kelime de sözcük dağarcığında yoksa ED (Edit Distance) algoritması ile en az karakter değişimi ile uygun kelime arandı. Bu kontrol yapılırken birden fazla kelime olması durumunda yazım yanlışı düzeltilen kelimenin diğer kelime ya da kelimelerle benzerlik oranı kontrol edildi. ED algoritması ile en az değişim yapılan kelime tercih edilirken bu kelimeler arasında benzerlik oranı en yüksek olan kelime alınarak olabilecek maksimum doğru kelime alındı.

Kelimelerin ürün adında geçmesi durumunda fazladan puan vererek üst sıraya çıkması durumunda geliştirme yapılarak aranan kelimelerin kökleri de kontrol edildi. Aynı kontrol ürün içerisindeki kelimenin kökünün arama kelimesi ile eşleşmesi durumu için de yapılarak kelime yapısından kaynaklı sorunların giderilmesi için eklendi.

IV. MAKİNE ÖĞRENMESİ YÖNTEMLERİNİN UYGULANMASI

Bu çalışmada makine öğrenmesi kapsamında ürün isimleri sınıf olarak kabul edilerek sınıflandırma tekniklerinin kullanılması hedeflenmiştir. Öznitelik olarak ise bu ürünü aramak için kullanılan kelimenin word2vec modelinde temsil edilen vektörünün değerleri kullanılmıştır (Şekil 2). Bir ürün için birden fazla arama kelimesi olması durumunda ürün her kelime için sınıf olarak atanmıştır. Tablo 1'de eğitim için kullanılan veri setinden örnek sınıf ve kelimeler gösterilmiştir.

TABLO I. ÖRNEK VERI GÖSTERIMI

Ürün	Aranan Kelimeler
DAMLA SODA 200ML SADE	Soda
SUPERFRESH PATATES 450 GR	superfresh patates
FAİRY SIVI BULAŞIK DETERJANI NAR 650 ML	fairy bulaşık
ETİ 16165 CİCİBEBE 172G TAHILLI	Bebe bisküvisi
SÜTAŞ CAM ŞİŞE GÜNLÜK SÜT 1LT	günlük süt
PINAR YOĞURT ORGANİK 1000GR	organik yoğurt
KOROPLAST BANYO BOY ÇÖP TORBASI 30LU	çöp poşeti

İndeks1 İndeks2 ... İndeks300 Sınıf

0.397183 0.435034…-0.245028 AKMİNA 200ML SODA SADE

-0.507915 0.245336... 0.050537 YUDUM AYÇİÇEK 5 LT

0.206975 0.333119…0.054616 NAMET DANA MACAR SALAM 150GR

Şekil 2. Makine Öğrenmesi İçin Kullanılan Veri Örneği.

İlk olarak 4200 ürün ve ürüne ait arama kelimesi verisi kullanılarak verinin %20'si test %80'i eğitim için ayrıldı. Bir sınıfa ait üye sayısı en az 10 ise eğitime katıldı. Aynı veri seti ile farklı sınıflandırma algoritmaları kullanılarak model eğitildi. Test verisine göre tahmin edilmesi gereken sınıfın bulunma sayısına göre doğruluk değerleri edildi. Algoritmalar bu değerlere göre karşılaştırıldı (Tablo 2). Sonuçlara göre Lojistik Regresyon algoritması ile en iyi sonuç alındı.

TABLO II. İLK SINIFLANDIRMA SONUÇLARI

Algoritma	Doğruluk (Accuracy)
Lojistik Regresyon	0.88
En yakın Komşu	0.83
Karar Ağaçları	0.87
Gaussian Naive Bayes	0.75
Destek Vektör Makineleri	0.79

A. BOYUT AZALTMA – PCA

Yapılan çalışmalarda kullanılan verinin çok fazla boyuta (özniteliğe) sahip olması, boyut büyüdükçe bütün süreçlerde harcanan zamanın artması gibi sorunlara sebep olmaktadır.Öznitelikler arasında yüksek korelasyon olması ve gereksiz bilgiye sahip olunmasına ve modelde aşırı uyma problemine sebep olabilmektedir. Bu sebeple eğitim yapılan veriseti arttırılmadan önce boyut azaltma teknikleri uygulanarak 300 öznitelik sayısı düşürüldü.

Principal component analysis (PCA) metodu yüksek boyutlu bir veri setinin boyutunu azaltmak için kullanılan en yaygın yöntemlerden biridir [3,4]. PCA öznitelik çıkarımı yaparak en az bilgi kaybıyla boyut küçültmektedir. Lojistik regresyon modeli ile 300 öznitelik kullanarak eğitilen modelin doğruluk oranı ve f1-puanı aynı veriseti kullanılarak PCA uygulanan modelin doğruluk oranı ve f1-puanı ile karşılaştırıldı. Tablo 3'te lojistik regresyon sınıflandırıcı ile farklı boyutlar için elde edilen doğruluk değerleri gösterilmiştir. Elde edilen sonuçlara göre en yakın sonuç 300 öznitelik sayısının 60'a indirilmesi ile elde edildi.

TABLO III. PCA ILE DOĞRULUK DEĞERLERI

Boyut Sayısı	Doğruluk (Accuracy)
20	0.75
30	0.68
40	0.68
50	0.81
60	0.87

B. VERİNİN ÇOĞALTILMASI VE SINIFLANDIRMA MODELİ İLE ARAMA MOTORU

Veri setinin temizlenmesi, veri çıkarımı ve çeşitli anket yöntemleri ile veri arttırılarak 25536'ya çıkarıldı. Arama kelimeleri kelime vektörlerine dönüştürülerek sınıf ile ilişkilendirildi. Daha önce 4200 sınıf- arama kelimesi verisi kullanılarak gerçekleştirilen modeli geliştirmek amacıyla yeni veri seti ile eğitim gerçekleştirildi. Bu veriler filtrelendi ve sınıfların her birinden eşit miktarda (10) üye alındı. Burada amaç artan sınıf sayısı ile öğrenmenin zorlaşması yanında dengesiz veri problemini ortadan kaldırmaktır. Yapılan çalışmalar sırasında verinin dengesiz olduğu durumlarda verimsiz sonuçlar elde edilmiştir.

Verinin uygun boyutta oluşturulmasından sonra PCA ile boyut azaltıldı ve eğitim 1561 sınıf ve 26711 kelime-sınıf örneği ile yapıldı. Modelin testi içinse kelimeye karşılık yalnızca bir sınıf kontrolü yapan kütüphane fonksiyonları yerine çoklu sınıf sınıflandırması yapıldığı için kelimeye karşılık tahmin edilen ilk 10 sınıfa bakıldı.

TP (True Positive): Tahmin edilmesi gereken sınıfın tahmin edilen sınıflar arasında ilk 10'da bulunması.

TN (True Negative): Tahmin edilmemesi gereken sınıfın reddedilmesi.

FP (False Positive): Tahmin edilmemesi gereken sınıfın kabul edilmesi.

FN (False Negative): Tahmin edilmesi gereken sınıfın tahmin edilen sınıflar arasında ilk 10'da bulunmaması.

Precision (P) = TP / (TP + FP)
Recall (R) = TP / (FN + TP)
Accuracy (A) = (TP + TN) / (TP + TN + FP + FN)
F1-Score = 2*P*R / (P+R)

Şekil 2. Ölçüm Değerlerinin Hesaplanması.

TABLO IV. MODELLERİN ÖLÇÜM DEĞERLERİ

	GNB	LR	LDA
P	0.43	0.67	0.60
R	0.48	0.70	0.62
A	0.57	0.81	0.77
F1	0.45	0.69	0.61

Tüm sınıflandırma algoritmaları aynı veri seti ile her sınıf için 10 adet üye alınarak denendiğinde en iyi sonuç alınan 3 algoritmanın değerlendirme metrikleri Tablo 4'te gösterilmiştir. P ve F1 değerlerinin Şekil 8'deki formüllere göre hesaplanması ile en iyi sonuç lojistik regresyon modelinden alındı.

Daha önce yalnızca Word2Vec kelime vektörlerinin anlamsal yakınlık bilgisi sağlaması ile geliştirilen arama motoru sonuçlarına ek olarak lojistik regresyon modelinden elde edilen sonuçlar eklendi. Bu sonuçların birleştirilmesi sırasında aranan kelimeler eğitim sırasında kullanılan PCA objesi kullanılarak eğitimdeki boyut azaltımına uğradılar. Daha sonra lojistik regresyon modelinin tahmin sonuçları sırayı bozmadan alındı. Birden fazla kelime aratılması durumunda tahmin sonuçlarındaki aynı sınıfa ait tahmin puanları toplanarak üst sıraya alınması sağlandı.

Lojistik regresyon modelinden gelen sonuçların kullanıcıların tercihlerini yansıtması sebebiyle arama motorunda modelden tahmin edilen sınıflara fazladan puan verildi ve bu sonuçların üst sıralara yerleşmesi sağlandı.

V. DEĞERLENDİRME

Aynı arama kelimeleri ile satın alınan ürünler (Şekil 3) veri setinden çıkarıldı ve arama motoru sonuçlarında bulunup bulunmadıkları test edildi. Bu test sırasında aynı arama kelimeleri için satın alınan ürün sayısı 5'ten küçük bir değer olduğunda tamamının arama sonuçlarında bulunması kontrol edildi. Satın alınan ürün sayısı 5'ten büyük olan arama kelimeleri içinse en az 5 tanesinin sonuçlarda bulunması kontrol edildi.

```
soda {'SIRMA SODA 200CC SADE DOĞAL.',
      'KIZILAY SODA 200 ML.',
      'ÖZKAYNAK SODA 200 ML',
      'BEYPAZARI SADE SODA 200ML',
      'DAMLA SODA 200ML SADE'}
```

Şekil 3. Arama Kelimesi ve Satın Alınan Ürün Örneği.

TABLO V. YÖNTEMLERIN KARŞILAŞTIRILMASI

	Word2Vec ile	**Word2Vec ve Sınıflandırma ile**
Tüm Test Verisi	0.69	0.81
Üye sayısı >1	0.57	0.83

Tablo 5'teki sonuçlara göre tüm test verisi kullanıldığında %81 oranında arama kelimesine karşılık bulunması gereken sayıda ürün bulundu. Üye sayısı 1 olan verileri almadan çalışma gerçekleştirildiğinde ise %83 oranında bulunma oranı elde edildi. Her iki durumda da yalnızca kelime vektörlerinin uzaklıklarına göre yapılan sıralamaya göre iyileşme sağlanmıştır.

Listelenen ürünlerin sıralamalarını kontrol etmek amacıyla, ürünlerin sıralamadaki yerlerinin ortalaması alındığında Tablo 6'da görülen sonuçlar elde edilmiştir. (Sonuçlar yaklaşık değerlere yuvarlanmıştır.)

TABLO VI. ÜRÜNLERIN BULUNMA SIRASI ORTALAMASI

	Bulunma Sırası Ortalaması
Test Verisinin Tamamı	1.66
Üye sayısı >1	2.00

Test sonuçlarına göre lojistik regresyon modelinden elde edilen tahmin sonuçları sıralamada üst sıralara taşınması sağlanmıştır.

Listelenen ürünler ve arama kelimesi arasındaki benzerliği kontrol etmek için yapılan çalışmada 100 adet arama kelimesi ile yapılan arama sonucu test edildi. Her arama için listelenen ilk 10 ürünün her birinin içerdiği kelimelerin vektörleri ile arama kelimesinin vektörü arasındaki benzerlik oranı kontrol edildi. Bu çalışmada 100 arama kelimesi için 77 tanesinin sonucunda %70 ve üzerinde benzerlik oranı yakalandı. Bu orana ulaşılamayan 23 ürün içinse listelenen 10 ürünün tamamı için değil ancak bu değere yakın sayıda ürün için benzerlik değeri bulundu.

VI. SONUÇ

Bu çalışmada anlamsal ürün arama motoru geliştirilmiştir ve sonuç olarak aranan kelime/kelimelere karşılık listelenen ürünlerin birbirleri ve arama için kullanılan kelimelerle anlamsal yakınlıkta olması sağlanmıştır. Aynı zamanda kullanıcıların tercihleri ile makine öğrenmesi uygulanması ile sıralamada iyileştirme yapılmıştır.

Günümüzde çokça kullanılan e-ticaret sitelerinde arama motorunun ne kadar iyi sonuçlar verdiği çok önemlidir. Listelenen ürünlerin anlamsal yakınlıkta olması ve bu ürünlerin doğru şekilde sıralanması sitenin tercih edilmesinde önemli rol oynamaktadır. Bu çalışmada ise kelimelerin anlamsal ilişkileri kelime vektörleri kullanılarak öne çıkarılmış ve sınıflandırma uygulanarak etkin sıralama elde edilmiştir.

KAYNAKLAR

[1] T. Mikolov, W.T. Yih, G. Zweig. Linguistic Regularities in Continuous Space Word Representations. NAACL HLT 2013.

[2] Mikolov, Tomas, et al. "Efficient estimation of word representations in vector space." arXiv preprint arXiv:1301.3781 (2013).

[3] Maitra, Saikat, and Jun Yan. "Principle component analysis and partial least squares: Two dimension reduction techniques for regression." Applying Multivariate Statistical Models 79 (2008): 79-90.

[4]. Fodor, Imola K. A survey of dimension reduction techniques. No. UCRL-ID-148494. Lawrence Livermore National Lab., CA (US), 2002.

5. Pohar, Maja, Mateja Blas, and Sandra Turk. "Comparison of logistic regression and linear discriminant analysis: a simulation study." Metodoloski zvezki 1.1 (2004): 143.

6. T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean. Distributed Representations of Words and Phrases and their Compositionality. Accepted to NIPS 2013.