T.C.
GEBZE TEKNİK ÜNİVERSİTESİ

Bilgisayar Mühendisliği Bölümü

MAKİNE ÖĞRENMESİ DESTEKLİ ANLAMSAL ÜRÜN ARAMA MOTORU

Simge SARIÇAYIR

Danışman
Prof. Dr. Fatih Erdoğan SEVİLGEN

Ocak, 2020
Gebze, KOCAELİ

T.C.
GEBZE TEKNİK ÜNİVERSİTESİ

Bilgisayar Mühendisliği Bölümü

MAKİNE ÖĞRENMESİ DESTEKLİ ANLAMSAL ÜRÜN ARAMA MOTORU

Simge SARIÇAYIR

Danışman
Prof. Dr. Fatih Erdoğan SEVİLGEN

Ocak, 2020
Gebze, KOCAELİ

Bu çalışma ..../..../200.. tarihinde aşağıdaki jüri tarafından Bilgisayar Mühendisliği Bölümü'nde Lisans Bitirme Projesi olarak kabul edilmiştir.

Bitirme Projesi Jürisi

Danışman Adı	Prof. Dr. Fatih Erdoğan SEVİLGEN	
Üniversite	Gebze Teknik Üniversitesi	
Fakülte	Mühendislik Fakültesi	

Jüri Adı	Prof. Dr. Erchan APTOULA	
Üniversite	Gebze Teknik Üniversitesi	
Fakülte	Mühendislik Fakültesi	

Jüri Adı	Doç. Dr. Hasari ÇELEBİ	
Üniversite	Gebze Teknik Üniversitesi	
Fakülte	Mühendislik Fakültesi	

ÖNSÖZ

Bu projenin gerçekleştirilmesinde yol gösterici olan Sayın Prof. Dr. Fatih Erdoğan SEVİLGEN hocama ve bu çalışmayı destekleyen Gebze Teknik Üniversitesi'ne içten teşekkürlerimi sunarım.

Ayrıca eğitimim süresince bana her konuda tam destek veren aileme ve bana hayatlarıyla örnek olan tüm hocalarıma saygı ve sevgilerimi sunarım.

Ocak, 2020 **Simge SARIÇAYIR**

İÇİNDEKİLER

ŞEKİL LİSTESİ

TABLO LİSTESİ

KISALTMA LİSTESİ

ED : Edit Distance
TP : True Positive
FP : False Positive
FN **:** False Negative
TN **:** True Negative
GNB **:** Gaussian Naïve Bayes
LR : Logistic Regression
LDA : Linear Discriminant Analysis

SEMBOL LİSTESİ

P	**:** Precision
R	**:** Recall
A	**:** Accuracy

ÖZET

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.

Gebze Teknik Üniversitesi 2020 lisans bitirme projesi olarak kelime vektörleri ve makine öğrenmesi teknikleri kullanılarak anlamsal yakınlıkta ürünleri listeleyen bir sanal market arama motoru geliştirilmiştir. Projede kullanıcıların bir ürünü satın almak için kullandıkları arama kelimelerinin makine öğrenmesi için kullanılması ile kullanıcı tercihlerine dayanarak ürün sıralaması yapılmıştır.

SUMMARY

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.

Gebze Techical University 2020 as a bachelor’s degree this project 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.

1. GİRİŞ

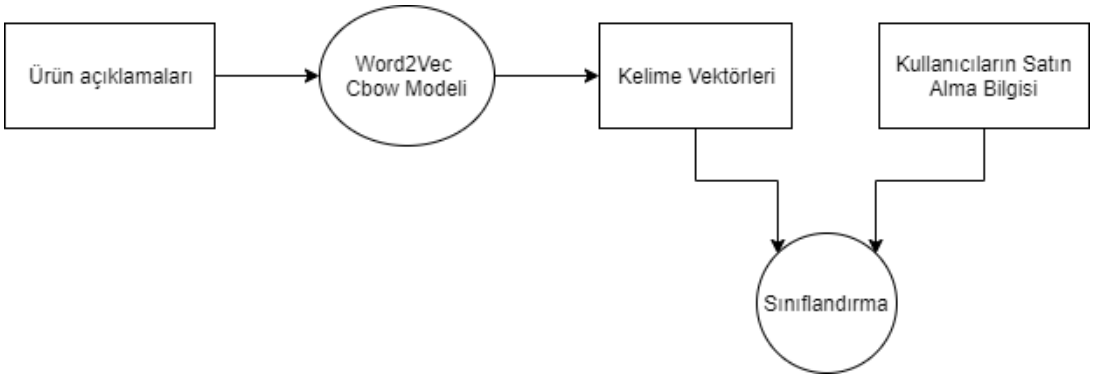
Günümüzde internetten alışveriş her geçen gün artmaktadır. E-ticaret sitelerinde alışverişin kolay olması ve daha çok seçenek sunması sebebiyle insanlar alışverişlerinde bu siteleri tercih etmektedir. Arama motorları ise alışveriş esnasında doğru ürünü doğru sırada listeleme konusunda iyi olmadığında 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

1.1 PROJE TANIMI

Aranan kelime/kelimelere karşılık listelenen ürünlerin birbirleri ve arama için kullanılan kelimelerle anlamsal yakınlıkta olması projenin genel amacı içerisinde yer almaktadır. Diğer bir amaç ise kullanıcıların ürünü satın almak için aradığı kelimelerin kullanılarak makine öğrenmesi ile eğitilen modelden elde edilen sonuçların sıralama yapılırken kullanılması ile etkili bir sıralama yapmaktır.

Arama motoru için kullanılacak ürün isimleri ve açıklamaları ise bir sanal market sitesinin kullandığı ürünlere aittir. Kullanıcıların sepete eklediği ürünlere karşılık arattığı kelime bilgisi ise aynı sanal marketin sağladığı bilgiler ve gerektiği noktada anket yolu ile toplanan veriler ile sağlanmıştır. Bu verilerin sağlandığı sanal market ve sağlayıcı hakkında gizlilik anlaşmaları sebebi ile bilgi verilmeyecektir.



Şekil 1. Sistem girdilerinin genel gösterimi

İlk olarak kelime vektörlerinin oluşturulacağı Word2Vec modelinde eğitim yapmadan önce etkili bir öğrenme yapabilmek için verilerin uygun şekilde ön işlemeden geçirilmesi gerekir. Şekil 1'de görülen ilk sistem girdisi olarak görülen ürün açıklama verileri ön işleme yapılarak geçersiz karakter, noktalama işaretleri ve kısaltamalardan temizlendikten sonra Word2Vec Cbow modeli ile eğitildi.

Cbow modelinden elde edilen kelime vektörlerinin ve kullanıcıların arattığı kelimelerle hangi ürünü satın aldığı bilgisi kullanılarak makine öğrenmesi tekniklerinden sınıflandırma modeli geliştirildi.

1.2 PROJENİN NEDEN VE AMAÇLARI

Projenin başlatılma nedeni olarak bir yazılım şirketine gelen e-ticaret sitesinin arama motoru sonuçlarını iyileştirme isteği kaynak alınmıştır. Arama sonuçlarında kullanıcıların arattığı kelimelere karşılık sonuç listelenmesi istendiğinde makine öğrenmesinin kullanıldığı bir arama motoru geliştirme fikri oluşmuştur.

Bu proje sayesinde:

- Arama motorunu kullanan kişi arattığı kelimelerle bulmak istediği ürünü üst sıralarda görebilecek.
- Satın almak istenen ürünün satışı olmaması durumunda arama yapılan kelimelerle anlamsal yakınlıktaki ürünlerin bulunması sağlanacak.
- Kullanıcıların satın alma geçmişinden öğrenen bir sistem olması sebebiyle listelenen sonuçlar yalnızca anlamsal yakınlık içeren ürünlerden oluşmayacak, aynı zamanda benzer deneyimlere göre ürün listelemesi sağlanmış olacak.

2. EYLEM RAPORU

Bu başlık altında proje boyunca yapılan çalışmaların ayrıntılı açıklamaları bulunmaktadır.

2.1 PROJE GEREKSİNİMLERİ

Bu projede başarılması gerekenler:

- Word2Vec modeli eğitebilmek için içerik uyuşmasının sağlandığı eğitim verisinin toplanması.
- Word2Vec Cbow modeli kullanılarak kelime vektörleri oluşturulması.

- Word2Vec modeli verilen metinin içeriğine bağlı öğrenme yaptığından, Word2Vec kullanarak eğitilen model ile anlamsal yakınlıktaki kelimeler elde edilmesi.
- Kelime vektörleri arasındaki mesafeye göre anlamsal yakınlık kontrolü yapılarak arama motoru geliştirilmesi.
- Aranan kelimelerin sözcük hazinesinde bulunmaması durumunda ED algoritması kullanılarak kelimedeki yazım yanlışlarının kontrol edilmesi.
- Arama kelimelerine karşılık satın alınan ürün bilgisi kullanılarak makine öğrenmesi yöntemi uygulanması.
- Makine öğrenmesini gerçeklemek için Word2Vec kelime vektörleri kullanılması.
- Kelime vektörleri kullanılarak geliştirilen arama motoru sınıflandırma modeli ile iyileştirilmesi.
- Geliştirilen arama motoru sınıflandırma modeli için kullanılan satın alınan ürün bilgisi ile test edilecektir.
- Hem anlamsal yakınlıkta ürünlerin listelenmesi sağlanacak hem de kullanıcı tercihlerine göre sıralanma yapılması sağlanacak.

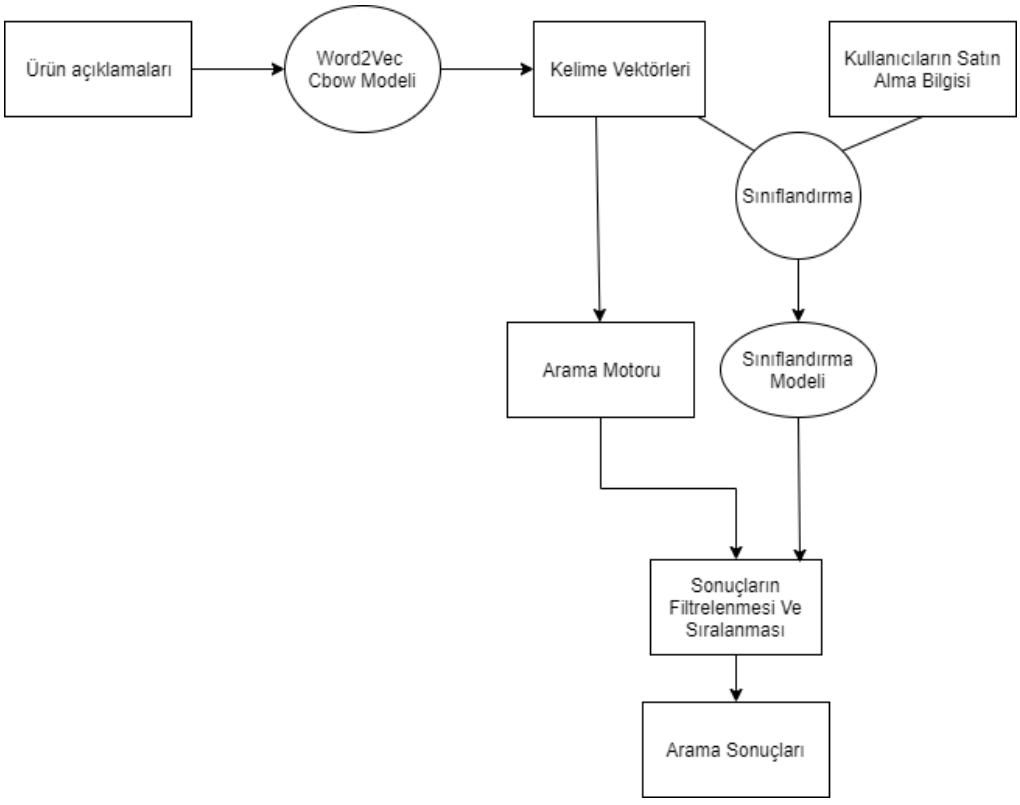
Bunların sağlanması için gerekli ihtiyaçlar:

- Arama motorunun gerçekleneceği ürün isimleri verisi.
- Ürünler hakkında bilgi içeren veri toplanması için GoogleSearch ve beautifulsoup kütüphaneleri.
- Gensim Word2Vec kütüphanesi.
- Kelimelerin morfolojik kökleri için Türkçe kelime köklerini içeren veri seti.
- Makine öğrenmesi ile kullanıcıların satın almayı tercih ettikleri ürünleri öğrenmek için arama kelimelerine karşılık satın alınan ürünün adı bilgisi.
- Pycharm entegre geliştirme ortamı.
- Makine öğrenmesi için sklearn kütüphanesi.

2.2 SİSTEM MİMARİSİ

Sistem mimarisi Şekil 2 ve aşağıdaki açıklamalar üzerinden açıklanmıştır:

- Projeyi gerçekleştirebilmek için ürünler hakkında bilgi içeren verinin toplanması ve işlenerek geçersiz karakter ve noktalama işaretlerinden temizlenmesi.
- Düzenlenmiş verinin uygun formata dönüştürülerek Word2Vec Cbow modelinin eğitilmesi.
- Word2Vec modelinden elde edilen kelime vektörleri ile bu vektörlerin uzaklıklarının kullanılması ile anlamsal ürün arama motoru geliştirilmesi.
- Kullanıcıların hangi kelimeleri kullanarak ürünü satın aldığı bilgisi kullanılarak bu kelimelerin kelime vektörleri ile sınıflandırma yapılması.
- Sınıflandırma modelinden elde edilen sonuçların daha önce geliştirilen arama motoru sonuçlarını iyileştirmek için kullanılması.
- Sonuçların başarı kriterleri doğrultusunda test edilmesi.



Şekil 2. Sistem Mimarisi

3. KELİME VEKTÖRLERİ

Bu bölümde proje boyunca kullanılan kelime vektörlerinin neden seçildiği ve yapısı üzerine bahsedilecektir.

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].

Benzerlik fonksiyonu(Formül 1**)** iki vektör arasındaki benzerliğin bulunmasında kullanılmaktadır.

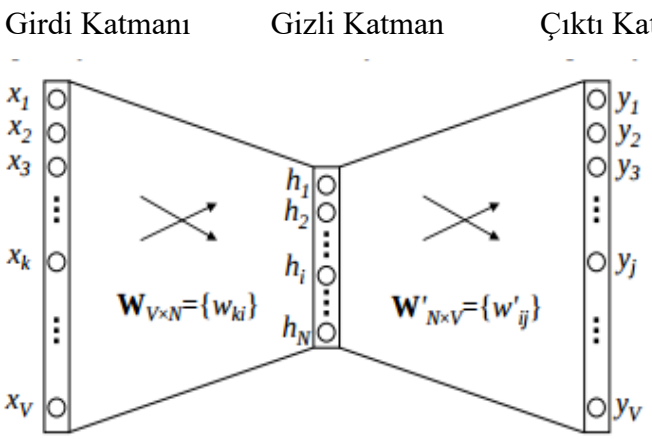
$$Benzerlik\ (A,B) = \frac{A \cdot B}{\|A\| \times \|B\|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \times \sqrt{\sum_{i=1}^{n} B_i^2}}$$

Formül 1: İki Vektör Arasındaki Benzerliği İfade Eden Fonksiyon

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

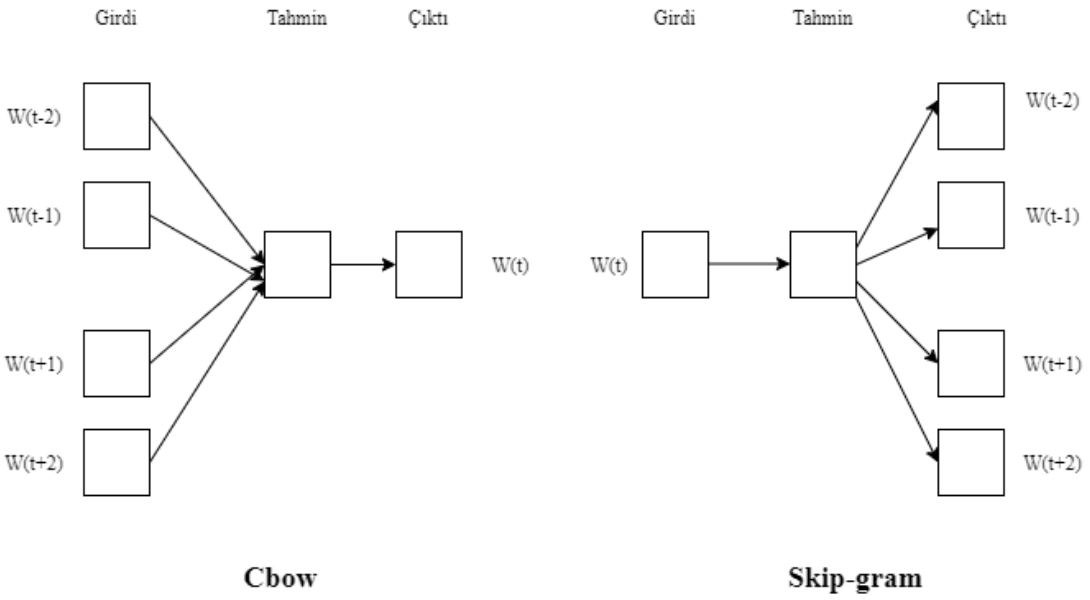
3.1 WORD2VEC CBOW MODELİ

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 (Şekil 3). Kelime vektörlerini oluştururken pencere genişliği, vektör boyutu gibi parametreler kullanmaktadır.



Şekil 3. Word2vec Katmanları

Cbow ve Skip-gram literatürde sıklıkla kullanılan iki word2vec metodu olarak öne çıkmaktadır [2] (Şekil 4). Cbow bir kelimeyi tahmin etmek için kelimeyi çevreleyen bir bağlamı kullanırken, Skip-gram sabit bir pencere boyutuna sahip kelimeleri çevreleyerek kelimeyi tahmin etmeye çalışır.



Şekil 4. Word2Vec Modelleri

Proje geliştirilirken 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.

3.2 VERİ KÜMESİ

Geliştirilen arama motoru için 6528 adet ürün adı kullanılmıştır.

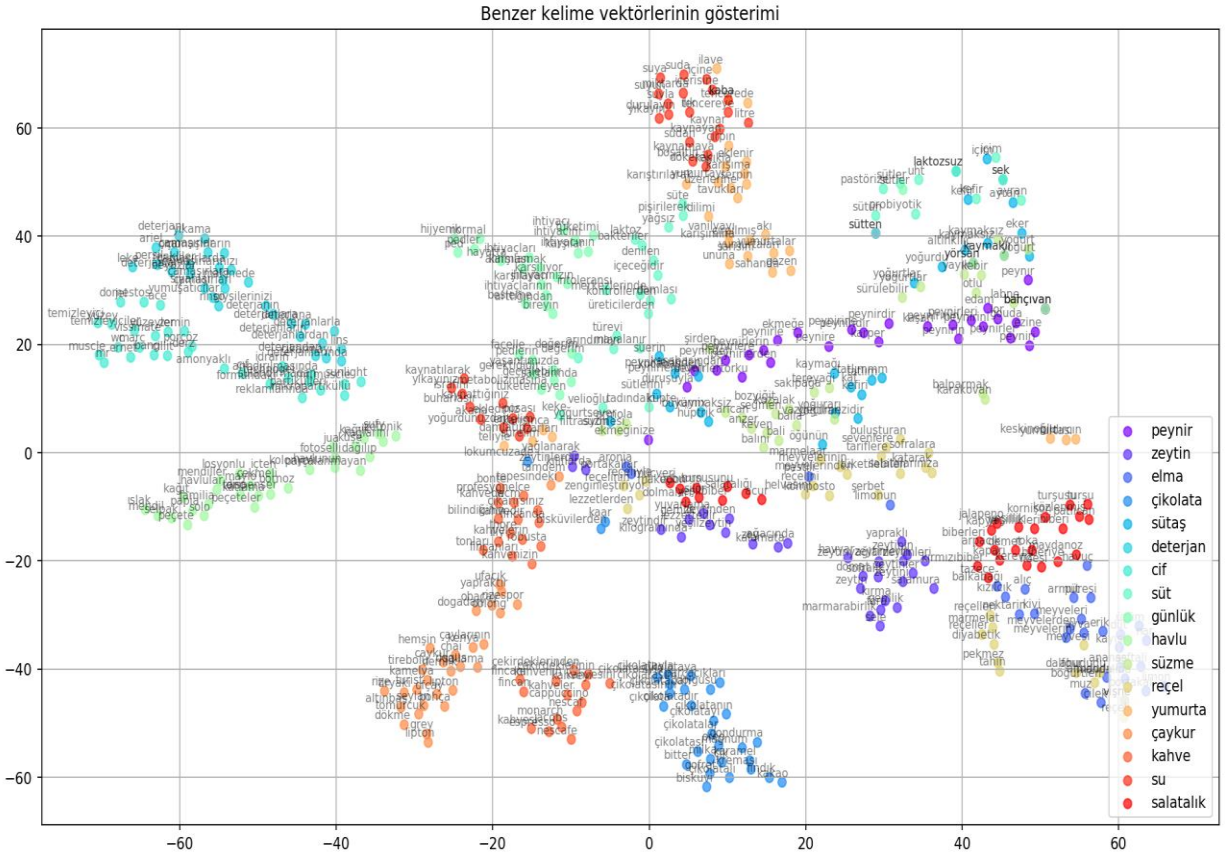
Word2Vec modelini projenin amacına uygun şekilde eğitmek için proje geliştirilerken kullanılan ürünler ile alakalı içeriğe sahip veri kullanılması gerekmektedir. Çalışmalara ilk olarak iki farklı e-ticaret sitesine ait veri kümelerinin eğitim için kullanılması ile başlandı ancak geliştirilmekte olan 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 toplandı.

3.3 MODEL EĞİTİMİ

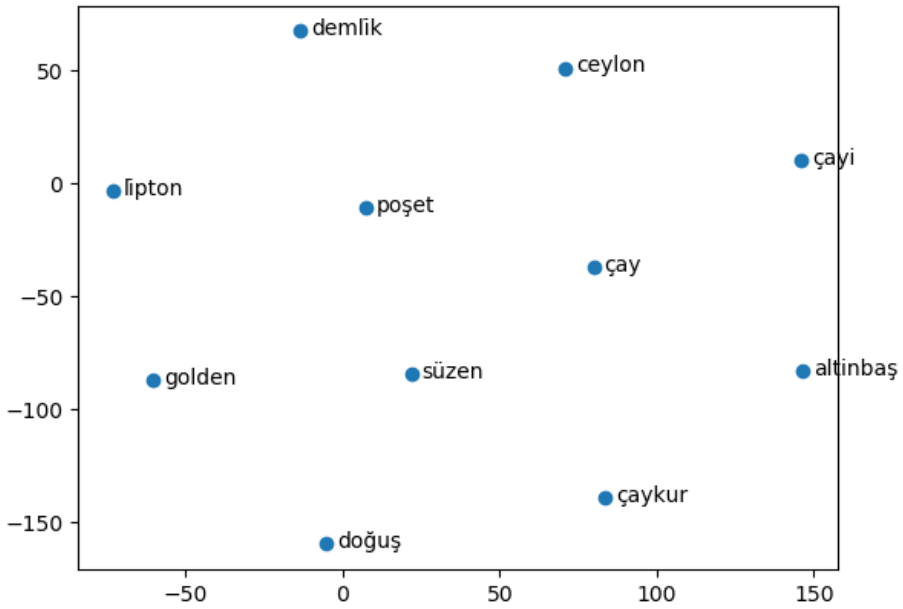
Doğal Dil İşleme problemlerinde metinler üzerinde çeşitli ön işlemler gerçekleştirilmelidir. Bu işlemler yapılan uygulamanın doğruluğunu artırmaya yardımcı olduğundan önemlidir. Bunun için internetten toplanan veri setindeki sayılar, noktalama işaretleri ve kısaltmalar temizlendi. Toplanan veri ön işleme yapılarak Word2Vec modelinin eğitim için alacağı girdi formatına çevirildi.

Parametre seçimi için; vektör boyutu olarak 300, pencere sayısı içinse 5 seçildi. Eğer bir kelime veri kümesi içerisinde 5'ten az sayıda geçiyorsa eğitime alınmadı.

Eğitilen Cbow modeli ile anlamsal yakınlıktaki kelimelerin gruplaşmış görünümü Şekil 5'de görülmektedir.



Şekil 5. Eğitilmiş Word2Vec Modelinde Örnek Kümelenme



Şekil 6. 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 6’daki kelimeler bulunmaktadır. Şekilde görüldüğü üzere ‘çay’ kelimesine yakın olan kelimeler bu alandaki markalar ve ürünler içerisinde geçen detaylardır.

3.4 KELİME VEKTÖRLERİ İLE ARAMA MOTORU

Eğitilen Word2Vec modeli kullanılarak arama motoru geliştirildi. Bunun için iki farklı yöntem denendi. İlk olarak aranan kelimelerin Word2Vec modelindeki vektörleri ile ürün içindeki kelimelerin vektörlerinin yakınlıklarına göre benzerliklerini kosinüs ifade eden fonksiyon 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 çalışma ürünü bütün olarak kıyasladığı için yapılan testlere göre için birinci çalışmadan daha iyi sonuç alındı ve projeye bu versiyon ile devam edildi.

3.4.1 ARAMA KELİMELERİNİN KONTROL EDİLMESİ

Arama yapılan kelime Word2Vec modelinin sahip olduğu sözcük dağarcığında olduğu durumda benzerlik karşılaştırması yapılabildiği için yazılım yanlışı yapılması durumunda herhangi bir sonuç alınamadı.

Bu sebeple ilk önce fst yazılarak kelimelerin türkçe-ingilizce 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.

4. MAKİNE ÖĞRENMESİ YÖNTEMLERİNİN UYGULANMASI

Makine öğrenmesi tekniklerinden sınıflandırma kullanıldı. Ürün isimleri sınıf olarak, özellik olarak ise bu ürünü aramak için kullanılan kelimenin Word2Vec modelinde temsil edilen vektörünün değerleri kullanılmıştır (Şekil 7). Bir ürün için birden fazla arama kelimesi olması durumunda ürün her kelime için sınıf olarak atanmıştır. Tablo 1'de eğitim için kullanılan verisetinden örnek sınıf ve kelimeler gösterilmiştir.

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

Tablo 1. Örnek Veri Gösterimi

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
0.472770	0.255918	... -0.139991	7DAYS KRUVASAN KAKAO KREMALI 300GR

Şekil 7. 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 veriseti ile farklı sınıflandırma algoritmaları kullanılarak model eğitildi ve sonuçlar doğruluk değerlerine göre karşılaştırıldı (Tablo 2). Sonuçlara göre Lojistik Regresyon algoritması ile en iyi sonuç alındı.

ALGORİTMA	DOĞRULUK ORANI
Lojistik Regresyon	**0.88**
En yakın Komşu	**0.83**
Karar Ağaçları	**0.87**
Gaussian Naive Bayes	**0.75**
Destek Vektör Makineleri	**0.79**

Tablo 2. İlk Sınıflandırma Sonuçları

4.1 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.

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

Tablo 3. Pca ile Doğruluk Değerleri

4.2 VERİNİN ÇOĞALTILMASI VE SINIFLANDIRMA MODELİ İLE ARAMA MOTORU

Daha önce eğitilen 4200 sınıf- arama kelimesi verisi kullanılarak model eğitimi gerçekleştirilmişti. Veri setinin temizlenmesi, veri çıkarımı ve çeşitli anket yöntemleri ile veri arttırılarak 25536'ya çıkarıldı. Arama kelimeleri kelime vektörlerine dönüştürülerek sınıf ile ilişkilendiridi. Bu sırada bir ürün için birden fazla kelime varsa bu sınıf için kelimelerin her biri tek tek ilişkilendirilerek sklearn kütüphanesinin modeli eğitmek için aldığı girdi formatına uyduruldu.

Bu veriler filtrelendi ve eğer bir sınıfa ait üye sayısı en az 10 ise eğitime alındı. Bu 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: Tahmin edilmesi gereken sınıfın tahmin edilen sınıflar arasında ilk 10'da bulunması.

TN: Tahmin edilmemesi gereken sınıfın reddedilmesi.

FP : Tahmin edilmemesi gereken sınıfın kabul edilmesi.

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

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

Şekil 8. Ölçüm Değerlerinin Hesaplanması

	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

Tablo 4. Sınıflandırma Modellerinin Ölçüm Değerleri

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 (Precision) ve F1 değerlerinin Şekil 8'deki formüllere göre hesaplanması ile en iyi sonuç Lojistik regresyon modelinden alınmıştır.

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 (Principal Component Analysis) 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.

5. BAŞARI KRİTERLERİ

Bu projeye başlanırken üç tane başarı kriteri belirlenmiştir.

1. Arama kelimelerine karşılık listelenen ilk 10 ürünün test ve eğitim için kullanılan verilerdeki aynı arama kelimesine karşılık satın alınan ürünlerin en az 5'ini içermesi.

Aynı arama kelimeleri ile satın alınan ürünler (Şekil 9) veri setinden çıkarıldı ve arama motoru sonuçlarında bulunup bulunmadıkları test edildi. Bu test sırasında arynı arama kelimeleri için satın alınan ürün sayısı 5'ten küçük bir değer olduğunda tamamının arama sonuçlarında bulunması kontrol edildi. 5'ten büyük satın alınan ürün bilgisi içeren arama kelimeleri içinse başarı kriterinde belirtildiği gibi 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 9. Arama Kelimesi ve Satın Alınan Ürün Örneği

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

Tablo 5. Birinci Başarı Kriteri Değerlendirme Sonuçları

Tablo 5'teki sonuçlara göre tüm test verisi kullanıldığında %81 oranında arama kelimesine karşılık bulunması gereken sayıda ürün bulunmuştur.Üyesi sayısı 1 olan verileri almadan test gerçekleştirildiğinde ise %83 oranında bulunma oranı elde edilmiştir.

Her iki durumda da yalnızca kelime vektörlerinin uzaklıklarına göre yapılan sıralamaya göre sonuçlar iyileşmiştir.

2. Kullanıcıdan toplanan verilere bakıldığında arama kelimesine karşılık satın alınan ürünlerin listelemedeki sıralarının toplamının ürünün sepete eklenme sayısına bölündüğünde 5'ten küçük bir değer elde edilmesi.

Birinci başarı kriterinin testlerine göre bulunanan ürünlerin sıralamadaki yerlerinin ortalaması alındığında Tablo 6'te görülen sonuçlar elde edilmiştir.(Sonuçlar yaklaşık değerlere yuvarlanmıştır.)

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

Tablo 6. İkinci Başarı Kriteri Değerlendirme Sonuçları

Test sonuçlarına göre lojistik regresyon modelinden elde edilen tahmin sonuçları sıralamada üst sıralara taşınmıştır ve belirlenen başarı kriterine ulaşılmıştır.

3. Listelenen ilk 10 ürün adının içerdiği kelimelerden en az birinin arama kelimesi ile word2vec modelindeki benzerlik oranının %70'in üzerinde olması.

Bu kriter için 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 kontol edildi.

100 arama kelimesi için 77 tanesinin sonucunda istenilen başarı kriterine ulaşıldı. Ulaşılamayan 23 ürün ise listelenen 10 ürünün tamamı için değil ancak bu değere yakın değerler ile başarı kriterine yaklaşmıştır. %77 oranında kritere ulaşıldı.

6. SONUÇ

Bu projede 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.

EKLER

Basılı: Ciltlenmiş bitirme projesi raporu

Cd: Proje dosyası