Araştırma Makalesi

METİN MADENCİLİĞİ KULLANARAK İNGİLİZCE DOKÜMAN SINIFLAMA

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

ÖZET

Günümüzde metin tabanlı dokümanların sınıflandırılması özellikle kurumsal yazışmaların ve dijital dokümantasyonun çok yapıldığı durumlarda ciddi öneme sahiptir. Metin yığınlarından benzer olanları sınıflandırma üretkenliği arttıran bir faktördür. Bu makalede tema ve alt kavramı tespit edilmiş dokümanlarda benzerliğin tespiti ile ilgili bir model önerilmiş ve deneysel bulgular değerlendirilmiştir. Dokümanlarda tema ve alt kavramların tespiti için kullanılabilecek anlamlı sözcüklerin belirlenmesi amacıyla Helmholtz prensibi temelli Gestalt teorisi kullanılmıştır. Sınama doküman veri seti spor ve eğitim temalarında olup, toplam 14 alt kavram belirlenmiştir.Daha sonra doküman kümesinden rastgele seçilen dokümanların birbirlerine olan benzerlikleri hesaplanmıştır. Önceden belirlenmiş sınıflara sahip dokümanlar için Kosinüs, Jaccard ve PMI benzerlik ölçütleri karşılaştırılmıştır. Benzerlik oranı toplam doküman benzerlikleri ortalama değerin üzerinde olan dokümanların tümü baz alındığında Kosinüs benzerlik ölçütü %75, Jaccard Indeks'i %40, PMI benzerlik ölçütü ise %55 başarı sağlamıştır. Buna rağmen doğruluk değerleri baz alındığında Kosinüs benzerlik ölçütü %80, Jaccard Indeks'i %65 ve aynı şekilde PMI benzerlik ölçütü de %65 başarı sağlamıştır. Her bir dokümanın benzerlik katsayılarının ortalamaları baz alınarak yapılan sınıflama ise anlamlı kelimelerin yüzdelik değerlerine göre farklı başarımlar elde edilmiştir. Bu bakımdan PMI benzerlik ölçütü anlamlı kelime dağılımlarına adaptif bir yaklaşım sergiler iken Kosinüs benzerlik ölçütünde ve Jaccard Indeks'inde herhangi bir iyileşme gözlemlenmemiştir.

Anahtar Kelimeler: Metin sınıflandırma, noktasal karşılıklı bilgi, helmholtz prensibi, benzerlik metrikleri, kosinüs benzerlik ölçütü, noktasal ortak bilgi benzerlik ölçütü, jaccard benzerlik ölçütü

ENGLISH DOCUMENT CLASSIFICATION USING TEXT MINING

ABSTRACT

Nowadays, the classification of text-based documents is very important, especially when corporate correspondence and digital documentation are intense. Classification of text sets according to similarities is an important factor that increases productivity.In this article, a model has been proposed to determine the similarity in the documents with the concept of theme and sub and the experimental findings are evaluated. The Gestalt theory based on the Helmholtz principle was used to determine the meaningful words that can be used to determine the themes and sub-concepts in the documents. The test document data set is in the sports and educational themes and a total of 14 sub-concepts have been determined. Cosine and PMI similarity criteria were compared for documents with predetermined classes. On the basis of all of the documents with a similarity rate on average, the similarity criterion of Kosinus was 75%, Jaccard Index was 40% and PMI similarity was 55%. On the other hand, based on the accuracy values, the cosine similarity criterion was 80%, Jaccard Index was 65%, and PMI similarity was 65%. According to the averages of the similarity coefficients of each document, different performances were obtained according to the percentage of meaningful words. In this respect, while the PMI similarity criterion exhibits an adaptive approach to meaningful word distributions, no improvement was observed in the cosine similarity criterion and in the Jaccard Index.

Keywords: Text classification, pointwise mutual information, helmholtz principle, similarity metrics, cosine similarity criteria, pmi similarity, jaccard similarity criteria

Geliş/Received	**:**	**07.05.2019**
Gözden Geçirme/Revised	**:**	**08.05.2019**
Kabul/Accepted	**:**	**31.05.2019**

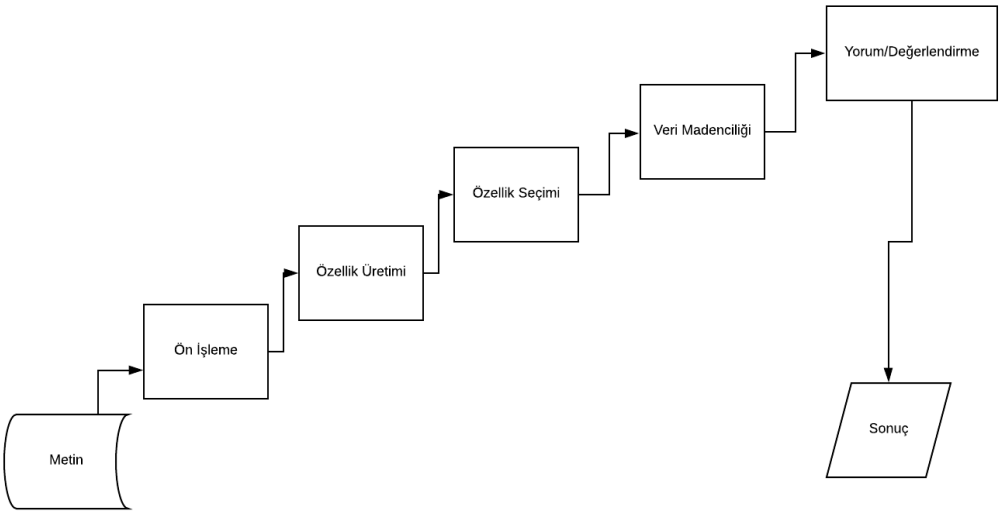
1. GİRİŞ

Günümüzde gelişen teknoloji ile birlikte üretilen veri miktarı ciddi seviyede artmıştır. Bu durum bilgiye ulaşmayı zorlaştırmakta olup, çeşitli yöntemler ile veriler anlamlandırılıp, mümkünse sınıflara ayrılıp, kullanma ve raporlama için kolaylıklar sağlanmaya çalışılmıştır. Veri madenciliği, veri yığınları içerisinden çeşitli yöntemler kullanılarak kesin olmamak ile birlikte anlamlı sonuçlar çıkarmayı amaçlar. Veri madenciliğinin bir alt dalı olan metin madenciliğinde ise yazılı dokümanlardan anlamlı bilgiler çıkarılması hedeflenir. Metin madenciliğini veri madenciliğinden ayıran fark, metin madenciliğinde çıkarımların veri tabanlarında yer alan anlamlandırılmış verileri üzerinde çalışılırken metin madenciliğinden doğal dil metinlerinden bu bilgilerin edinimidir. Yapılan çalışmalara göre, şirket bilgilerinin %80'i metin dokümanlarında tutulmaktadır (A. Akilan, 2015). Bu bakımdan metin madenciliği çalışmalarının gelişen teknoloji ile birlikte önemli bir noktaya geldiği aşikârdır. Bu çalışmada önceden belirlenmiş sınıflara sahip doküman setlerinde doğal dil işlemenin ilk adımı olan ön işlemeden geçirilmiş dokümanlar için anlamlı kelimelerin tespiti yapılmış ve verilerden yola çıkarak her bir doküman için tekrarlama değerleri yok edilip ikili sistem doğrusal vektörler oluşturulmuştur. Bu doküman vektörleri taranarak anlamlı kelimelerin dokümanlar içerisinde yer alma olasılıkları ile iki kelimenin bir doküman içerisinde birlikte geçme olasılıkları hesaplanmıştır. Bu değerler PMI (Noktasal karşılıklı bilgi) formülüne uygulanmış ve her bir doküman için PMI değerlerinin toplamı benzerlik ölçütü olarak hesaplanmıştır. Sonuçlar sınıfları önceden belirlenmiş dokümanlar için değerlendirilmiş ve benzerlik oranlarına göre önceden verilen sınıflama karşılaştırılmıştır. Bu çalışmada PMI, noktasal karşılıklı bilgi metriğinin dokümanlar arası benzerlik ölçütü için ne kadar faydalı sonuçlar verebileceği için araştırma çalışması yapılmıştır. En iyi bilinen Kosinüs ve Jaccard metrikleri ile kıyaslanmış, sonuçlar paylaşılmıştır.

1.1. Metin Ön İşleme

Metinlerden anlam çıkarımı yapılabilmesi için öncelikle verinin ön işlemden geçirilmesi gerekmektedir. Veriler bazı özel formatlar, sayı biçimleri, tarih biçimleri, ön ekler, özel ifadeler içermesinden ötürü işleme tabi tutulacak metinden ayıklanabilirler. Bu işlemleri metin madenciliğinde başarılı çıkarım yapılabilmesi için önemlidir, metin ön işleme, madencilik ve kümeleme performansını arttırır (A. I. Kadhim, Y. Cheah and N. H. Ahamed, 2014). Yapısal olmayan veriden yapısal bir metin elde edilebilmesi için bir takım dönüştürme işlemleri yapılır (Dr.S.Vijayarani, 2015).

Büyük miktardaki metinsel verilerden potansiyel olarak yararlı ve önceden bilinmeyen belirli bir önemi olan bilginin çıkarılması olarak nitelendirilen Metin Madenciliği şekilde görüldüğü gibi temelde altı adımdan oluşmaktadır. Metin Madenciliği işlemleri, Veri Madenciliğine benzer olarak Şekil 1.'deki gibi özetlenebilir.



Şekil 1. Metin işleme adımları.

1.2. Anlamlı Kelimelerin Tespiti

Doküman içerisinde geçen anlamlı kelimelerin tespiti için Helmholtz prensibi kullanılmıştır. Helmholtz prensibi, Gestalt insan algı teorisini kullanır. Gestalt teorisi, zihinsel figür gibi imgeleri nasıl düzenlediğimiz ve görüntüleri görsel, işitsel ve koku uyarıcıları gibi duyusal girdiler yoluyla nasıl algıladığımız açıklamaktadır. Tecrübelerimizi düzenli simetrik ve basit şekilde düzenleme eğiliminde olduğumuz fikrine dayanmaktadır. Bu

çalışmada Helmholtz teorisi, her bir kelime için; doküman paragrafında, kelimenin m kez geçmesinin olası olup olmadığının belirlenmesinde kullanılmıştır. Anahtar kelimeler çıkarma sorununa uygulandığında, parametre içermeyen yöntemleri kullanarak anlamlı kelimeleri tanımlamak için hızlı ve etkili araçlar sunar. Ayrıca, farklı uygulama ihtiyaçları için seçilen anahtar kelime kümelerinin boyutunun kontrol edilmesini sağlayan dokümanların anlamlılık seviyeleri tanımlanmıştır (B. Dadachev, A. Balinsky, H. Balinsky ve S. Simske, 2012).

1.3. Anlamlı Kelimelerin Veri Kümesi İçerisindeki Ağırlıklarının Belirlenmesi

Anlamlı kelimelerin bir veri kümesi içerisinde tekrarlanmasından ziyade sadece doküman içerisinde geçip geçmediğinin tespiti ile bir hesaplama yapılmak istendiği için her bir doküman için ikili sistem vektör dizileri oluşturulmuştur. Bir dokümanın "futbol, yüzme, skor, yağmur, dünya" kelimeleri ile temsil edildiğini varsayalım. Vektörün boyu temsil edilen kelime sayısı kadardır. Örneğin bu kelimelerin frekansları aşağıdaki gibi verilmiş olsun:

$$V[2, 3, 1, 4, 2]$$

İkili vektör yaklaşımında, metin verileri 1 ve 0'lar ile ifade edilmektedir. Veri içinde barındırdığı kelimelerin metin içerisindeki varlıklarına göre bu değerleri almaktadır. Veri kümesi içerisindeki kelimelerin alacağı değerler binary vektör temsilinde $V[1, 1, 1, 1, 1]$ şeklinde olmaktadır.

Vektörler oluşturulurken anlamlı kelimeler teker teker dokümanlar içerisinde taranıp doğrusal bir vektör oluşturulduğu için kelimelerin veri tabanında tutulan indeks'lerinin birebir aynı olması gerekmektedir. Bu indeks sıralaması bozulduğu takdirde vektör üzerinde tutulan ikili sistem indis'lerde kayma yaşanacağı için hatalı sonuçlar üretilecektir. Her bir doküman için oluşturulan vektörler birliktelik hesaplamalarında kullanılmıştır.

1.4. Anlamlı Kelimeler İçin Doküman İçerisinde Tekil Tekrar Olasılıklarının Hesaplanması

Doküman seti için anlamlı olarak değerlendirilen kelimelerin, doküman kümesi içerisinde geçme olasılıkları:

$$P(x) = \sum \frac{V(x)}{D} \quad (1)$$

formülüne göre hesaplanmaktadır.

Formülde kullanılan değişkenlerin açıklamaları aşağıdaki gibidir:

x: Kelime.
$\sum V(X)$: Kelimenin tekrarlanma sayısı.
D: Doküman sayısı.
$P(x)$: Bir kelimenin doküman içerisinde geçme olasılığı.

Bir kelimenin, bir doküman seti içerisindeki tekrarlanma sayısını kıyaslamaların yapılacağı toplam doküman sayısına bölerek seçilen kelimenin karşılışılma ihtimali hesaplanmıştır. Örneğin, 5 dokümandan oluşan bir doküman listesi içerisinde "futbol" kelimesinin tüm doküman vektörleri içerisinde 3 kere geçtiği düşünüldüğünde $P(x)$ değeri $3/5$ olarak hesaplanır.

1.5. Anlamlı Kelimelerin Birlikte Bulunma Olasılıklarının Hesaplanması

Bu adımda her bir anlamlı kelime çifti için birlikte bulunma olasılıkları hesaplanmaktadır. Bu değerler veri tabanına kaydedilmiş ve PMI formülünün payı olarak kullanılmıştır. Oluşturulan doküman vektörleri taranarak iki dokümanının indeksi için true(doğru) olan indislerde birlikte geçme sayısı bir arttırılarak formülün payı hesaplanmıştır. Payda ise doküman sayısı olarak belirlenmiştir.

Birlikte bulunma olasılıklarının hesaplanması için kullanılan formül:

$$P(x, y) = \frac{a}{D} \quad (2)$$

şeklindedir. Formülde kullanılan değişkenlerin açıklamaları aşağıdaki gibidir:

$P(x, y)$: Kelimelerin birlikte geçme olasılıkları.
a : Kelimelerin doküman içerisindeki birlikte geçme sayısı.
D : Doküman sayısı.

Bir kelime çiftinin doküman seti içerisinde birlikte geçme olasılıklarının hesaplanması için yine doküman vektörleri kullanılarak iki kelime içinde vektör indis değerinin 1 olduğu tüm durumlar a değişkenini bir arttırır. Bu a değişkeni hesaplama yapılan kelime çiftinin doküman vektörleri içerisinde aynı anda geçtiği durumları ifade etmektedir. Örneğin, 10 adet dokümanın kullanıldığı bir doküman seti için, futbol ve taç kelimelerini baz alırsak ve bu iki kelimenin aynı dokümanda tekrarlanma sayısının da 2 olduğunu var sayarsak. $P(futbol, taç)$ için sonuç 0.2 olarak hesaplanmaktadır.

1.6. Noktasal Karşılıklı Bilgi (PMI) Hesaplanması

Noktasal karşılıklı bilgi (PMI) veya nokta karşılıklı bilgi, bilgi teorisi ve istatistik biliminde kullanılan bir birliktelik ölçüsüdür. PMI, üzerine inşa eden karşılıklı bilginin (MI) aksine, tek olaylara atıfta bulunur, MI ise ortalama olası tüm değerleri kapsamaktadır. PMI hesaplması eşitlik (3)'de gösterildiği gibidir.

$$PMI(x, y) = \log \frac{P(x,y)}{P(x)} \cdot P(y) \quad (3)$$

Formülde kullanılan değişkenlerin açıklamaları aşağıdaki gibidir:

$PMI(x, y)$: x ve y kelimeleri için noktasal karşılık bilgi katsayısı.
$P(x, y)$: x ve y kelimelerin birlikte geçme olasılıkları.
$P(x)$: x kelimesinin dokümanlar içerisinde geçme olasılığı.
$P(y)$: y kelimesinin dokümanlar içerisinde geçme olasılığı

Yukarıdaki formül kullanılarak, örnek doküman seti içerisinde yer alan anlamlı kelimeler için kelimelerin tek başına dokümanlar içerisinde geçme olasılıkları, her bir kelime çiftinin dokümanlar içerisinde birlikte geçme olasılıkları hesaplanmış ve bulunan değerler ile tüm dokümanlar için anlamlı kelimelerin noktasal karşılıklı bilgi katsayıları hesaplanmıştır.

1.7. Benzerlik Katsayılarının Hesaplanması

PMI benzerliği ile karşılaştırılmak üzere iki yaklaşım ele alınmıştır. Kosinüs benzerliği ve Jaccard benzerliği, PM benzerliği ile aynı veri seti üzerinde hesaplanmış. Sonuçlar ağırlık(weight), doğruluk(accuracy) ve yoğunluk (intensity) metriklerine göre her bir benzerlik ölçütü için ayrı ayrı hesaplanmıştır. 1.6 bölümünde bahsedilen PMI hesaplamalarının toplamı bize her bir doküman için PMI benzerlik katsayılarını vermektedir. Bir doküman için hesaplanan PMI kombinasyonlarının toplamı PMI benzerlik katsayısını oluşturmaktadır ve aşağıdaki toplam formülü ile hesaplanmaktadır:

Benzerlik

$$\left(D_{\mathrm{l}}, D_j\right) = \sum_{i=1}^{k} \sum_{j=1}^{l} PMI(w_i, w_j) \quad (4)$$

eşitliğine göre hesaplanan tüm PMI kombinasyonları kümülatif olarak benzerlik katsayısı değişkenine eklenerek iki dokümanın birbirine benzerliği için bir değer üretilmektedir.

Bu değerlerin dokümanların sınıflandırılmasında kullanılabileceğini öne sürmekteyiz. Bu çıkarımı yapabilmek aynı doküman setlerinin, anlamlı kelimeleri baz alınarak Kosinüs ve Jaccard ölçütünde ürettiği sonuçlar kayıt altına alınmış ve bu istatistiksel bilgiler kullanılarak bir çıkarım yapılmaya çalışılmıştır.

1.8. İlişikili Çalışmalar

1990'ların başından günümüze kadar birçok alanda çalışma yapılmıştır. L. Guthrie ve E. Walker, isimli bilim insanları, makine ile doküman sınıflandırma teorisi üzerinde çalışmışlardır. Kelime frekansına dayanan matematiksel bir modeli öne sürmüş ve modelin sınıflandırma yöntemi olarak kullanılabileceğini deney sonuçlarına dayanarak raporlamışlardır (L. Guthrie ve E.Walker, 1994).

R. Agrawal ve R. Srikant yaptıkları çalışmada büyük veri tabanında birleştirme kuralları önerilmiştir. Dijital Belge Koleksiyonlarında Açıklayıcı İfade Çıkarma önerisinde, özellik vektörlerini dijital belgelerden ayıklar ve özellik vektörlerinden açıklayıcı ifadeler çıkarılmıştır (R. Agrawal ve R. Srikant, 1994). Veri tablosundaki özelliklerin destek değerine dayalı kurallar oluşturulmuştur (K. Aas ve L. Eikvil, 1999).

Bilgi Edinme (IR), Rocchio ve olasılıklı modeller (R. Baeza-Yates ve B. Ribeiro-Neto, 1999), BM25 ve destek vektör makinesi (SVM) (R. Baeza-Yates ve B. Ribeiro-Neto, 1999) bazlı filtreleme modelleri, kaba set modelleri (Y. Li, C. Zhang, ve J.R. Swan, 2000) gibi bu zorluğu çözmek için birçok terim temelli yöntem denenmiştir. Terim temelli yöntemlerin avantajları, verimli hesaplama performansı ve son birkaç on yıl boyunca IR ve makine öğrenmesi topluluklardan ortaya çıkan terim ağırlıklandırma için olgun teoriler önerilmiştir.

Ayrıca Feng Hu ve Yu-feng Zhang Ontoloji üzerine metin madenciliği tabanlı çalışmalar yapmıştır (F. Hu and Y. Zhang, 2000).

Xu, Yan & Jones, Gareth & Li, Jintao & Wang ve Bin & Sun yapmış oldukları çalışmada, Metin sınıflandırmada kullanılan ortak bilgi formunun bilgi teorisinden doğru bir şekilde türetilmediğini göstermişlerdir. Metin sınıflandırma literatüründe ortak bilgi olarak adlandırılan iki farklı ortak bilgi tabanlı özellik seçim kriteri olduğuna işaret etmişlerdir ve "noktasal karşılıklı bilgi" (PMI) olarak adlandırılması gerektiğini öne sürmüşlerdir. Ayrıca çalışmada, metin sınıflandırmadaki "karşılıklı bilgi" kavramını çevreleyen terminolojik karışıklık açıklığa kavuşturulmuş ve bilgi teorisinden doğru bir şekilde türetilmiş bir ortak bilgi yöntemini ayrıntılarıyla anlatmışlardır (Xu, Yan & Jones, Gareth & Li, Jintao & Wang ve Bin & Sun, 2007).

Metin sınıflandırması, metin belgelerinin kategorilere veya konulara göre sınıflandırılması, herhangi bir metin işleme sisteminin önemli bir bileşenidir. Doğru belge sınıflandırıcıları oluşturmak için içeriği- belgelerde görünen sözcükleri, belgelerin yapısını- ve dış kaynakları kullanan çok sayıda çalışma vardır (Olsson F. , 2009). Ek olarak, doküman sınıflandırma performansını arttırmak için dokümanlar arasındaki bağlantı yapısını kullanmaya çalışan yöntemler üzerine artan bir literatür var (R. A. Sinoara, J. Camacho-Collados, R. G. Rossi, R. Navigli, Solange O. Rezende, 2019). Metin belgeleri çeşitli yollarla birbirine bağlanabilir. En yaygın bağlantı yapısı alıntı grafiğidir: Örneğin, makaleler belgeleri ve web sayfalarını diğer web sayfalarına bağlar. Bunların hepsi birbirine bağlı bir metin belgeleri koleksiyonu oluşturmak için bir araya getirilebilir.

W.Heyong ve H.Ming yapmış oldukları çalışmada "Hebb kuralına dayalı özellik seçimi (HRFS)" olarak adlandırılan bir yöntem önermektedir. HRFS, denetlenen Hebb kuralına dayanmaktadır ve terimlerin ve sınıfların nöronlar olduğunu varsaymakta ve ilgili terimleri "heyecanlandırıyorsa" bir terimin ayırt edici olduğu varsayımıyla terimleri seçmektedir. Bu varsayım, orijinal Hebb varsayımına göre "heyecan verici" tutabiliyorsa "bir sınıf bir sınıfla yüksek oranda ilişkilidir" olarak açıklanabilir. HRFS'yi diğer yedi özellik seçim yöntemiyle karşılaştırmak için altı kıyaslama veri seti kullanılmıştır. Deneysel sonuçlar HRFS'nin, karşılaştırılan yöntemlerden daha iyi performans elde etmek için etkili olduğunu göstermektedir (W.Heyong ve H.Ming,2019).

Türkçe dili için de metin madenciliği konusunda çalışmalar yapılmıştır.

Benzerlik bulunması ile ilgili yapılan çalışmada Bunyamin Dursun ve A. Coskun Sönmez, Türkçe metin benzerliklerinin bulunması için yeni bir yöntem önermişlerdir. Çalışmada Türkçe dili için sık yapılan hatalar tespit edilmiş ve bu hataların düzeltilmesi için çözümler önerilmiştir. Uygulanan çözümlerin başarıya ulaşıp ulaşmadığının tespiti için uygulanan çözümler, Levenshtein Edit Distance benzerliği ve Jaro-Winkler benzerliği ile karşılaştırılmış ve sonuçlar değerlendirilmiştir (B.Dursun ve A.C.Sonmez, 2008).

Tuğberk Kocatekin ve Devrim Ünay ise metin madenciliği kullanarak Türkçe dilindeki radyoloji raporlarını analiz etmişlerdir (T. Kocatekin and D. Ünay, 2013).

G. İşgüder-Şahin, H. R. Zafer ve E. Adalı tarafından popüler bir Türk web sitesinden teknoloji markaları ile ilgili yorumlar toplanmış ve olumlu veya olumsuz olarak sınıflandırılmıştır (G. G. İşgüder-Şahin, H. R. Zafer and E. Adah, 2014).

A.İ.Kısayol ve M.Turan yapmış oldukları çalışmada paragraf tabanlı çıkarımsal öbekleme kullanan iki yeni yöntemi kıyaslamışlardır. Yapmış oldukları çalışmada kümeleme algoritması olarak K-Means kullanmış ve geçiş sıklıkları hesaplanan kelimelerinden en sık kullanılan 10 tanesinin en çok geçtiği paragrafların seçildiği yöntem ile Jaccard uzaklığı bakımından en yakın olan paragraflar karşılaştırılmıştır. (A.İ.Kısayol ve M.Turan, 2018)

1.9. Veri Kümeleri

Uygulamada sınanmak üzere 140 adet doküman rastgele seçim yöntemi ile programa girdi olarak verilmiştir. Dokümanların sınıfları önceden belirlenmiştir. Eğitim, Spor ana başlıkları altında 2 sınıf altında fiziksel ve mantıksal olarak tutulan kayıtlar üzerinde çalışmalar yapılmıştır. Anlamlı kelimelerin %25, %50 ve %100'ünü kullanarak veri kümesinde öbekleme algoritması çalıştırılmıştır. Çalışma için kullanılan tüm dokümanlar txt uzantısına sahip ve içerikleri HTML formatında paragraflardan oluşmaktadır.

2. YÖNTEM

Bu çalışmada Python programlama dili ve MS-SQL Server veri tabanı sunucusu kullanılarak, sınıfları önceden belirlenmiş İngilizce dokümanlarda bir benzerlik ölçütü ortaya konulmuştur. Deney veri seti olarak kullanılan HTML formatında dokümanlar öncelikle ön işleme adımlarından geçirilmiştir. Böylelikle İngilizce dilinin yazımında kullanılan fakat tek başına anlam ifade etmeyen ve metinin içerisinden çıkarıldığında anlam kaybı oluşturmayan kelimeler (stop words) atılmıştır. Bağlaç, rakam, noktalama işaretleri bu kapsamda değerlendirilmiş ve metinin içerisinden atılmıştır. Doğal dil işleme çalışmalarında sıklıkla kullanılan kök bulma algoritmalarından olan Porter Stemmer algoritması kullanılarak (B. Issac ve W. J. Jap, 2009) kalan kelimelerin kökleri bulunmuştur. Ön işleme işlemlerinin bir diğer adımı olarak da metin içerisinde yer alan anlamlı kelimeler bulunmuştur. Bu adımda Helmholtz Prensibi referans alınmıştır.

Veri setleri, dokümanlarda yer alan kelimeler ve Helmholtz prensibine dayanan kelime anlam değerleri veri tabanına kayıt edilmiştir. Bu aşamadan sonra anlamlı kelimelerin dokümanlar içerisindeki dağılımlarını hesaplayabilmek için vektör uzayları oluşturulmuştur. Her bir doküman için anlamlı kelime sayısı kadar indis içeren doğrusal vektörler oluşturulmuş ve veri tabanına kayıt edilmiştir. PMI formülünde kullanılmak üzere anlamlı kelimelerin doküman içerisinde geçme olasılıkları hesaplanmıştır. Tekil olasılıkları kayıt altına alındıktan sonra kelimelerin doküman kümesi içerisinde birlikte geçme olasılıkları hesaplanmış ve veri tabanına kayıt edilmiştir. PMI, Noktasal karşılıklı bilgi veya nokta karşılıklı bilgi, bilgi teorisi ve istatistik biliminde kullanılan bir birliktelik ölçüsüdür. Veri tabanına kayıt edilen PMI değerlerinin toplam değerleri benzerlik ölçütü olarak kabul edilmekte ve benzerlik oranlarının tahmini bu formüle göre verilmektedir.

Kosinüs, Jaccard ve PMI için bulunan sonuçlar dokümanların birbirlerine benzerliğine dair istatiksel yorum yapmayı sağlamaktadır. Bu kıyaslamalar yapılırken öznellikten uzaklaşmak için, aynı doküman setlerinin kullanılmış, aynı anlamlı kelimeler için benzerlik katsayıları baz alınmıştır. Ortalama benzerlik değerlerinin üzerindeki kayıtlar "doğru sınıflandırma", bu değerlerin altında kalan veya hesaplanamayan kayıtlar için ise "yanlış sınıflandırma" kabulü yapılmıştır.

3. SONUÇ

Bu çalışmada benzerlik ölçütü karşılaştırılması için Kosinüs, Jaccard ve Noktasal Karşılık Bilgi yöntemleri kullanılmış ve sonuçlar değerlendirilmiştir.

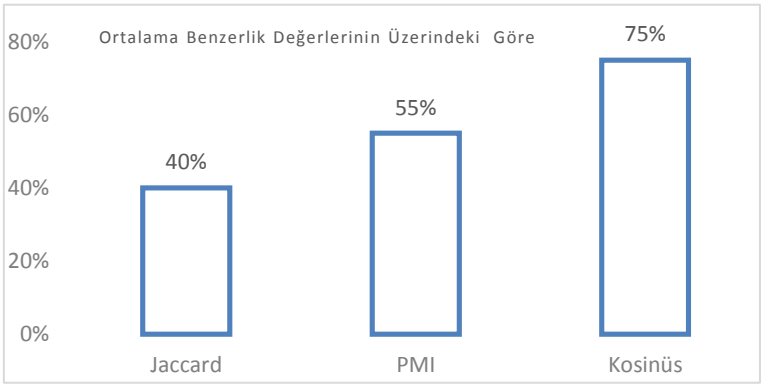
Kosinüs benzerliği, metinlerin vektör olarak ifade edildiğinde aralarındaki açıların bir benzerlik ölçütü olabileceğini bir metrik öne sürer. Kosinüs benzerliği, aralarındaki açının kosinüsünü ölçen, iç çarpım uzayının sıfır olmayan iki vektörü arasındaki benzerliğin bir ölçüsüdür. 0 ° 'lik kosinüs 1'dir ve (0, π] radyanlar arasındaki herhangi bir açı için 1'den azdır. Kosinüs benzerliği, sonucun [0,1] arasında net bir şekilde sınırlandığı pozitif alanda özellikle kullanılır. Bu ad, "kosinüs yönü" teriminden türetilir: bu durumda, birim vektörler paralel ise maksimum "benzer" ve dik (dikey) ise maksimum "farklı" olabilirler. Bu, farklar sıfır açıya döndüğü zaman birlik (en yüksek değer) olan ve kosinüse dik olduğunda sıfıra (ilişkisiz) benzerdir.

Birlik üzerine kesişim (Intersection Over Union) ve Jaccard Benzerlik Katsayısı olarak da bilinen Jaccard indeksi, sonlu örnek kümeler arasındaki benzerliği ölçer ve örnek kümelerin birliğinin büyüklüğüne bölünmesiyle kesişim ölçüsü olarak tanımlanır.

Noktasal karşılıklı bilgi (PMI), p (x, y) olaylarının belirli bir eşzamanlı oluşumunun gerçek olasılığının, bireysel olayların olasılıkları P(x), P(y) temelinde olmasını beklediğimizden ne kadar farklı olduğunun bir ölçüsüdür.

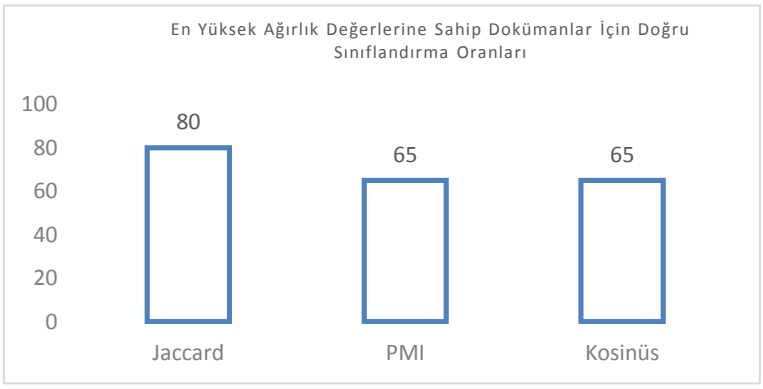
Her dokumanin diger dokumanlarla olan benzerliklerinin ortalamasi uzerinde kalanlarin ait olduklari siniflarin cogunluguna gore sinif tespiti yapildiginda elde edilen metrik degerleri Şekil 1'de görüleceği gibi Kosinüs

benzerliğinde %75, PMI benzerliğinde %55 ve Jaccard benzerliğinde %40 olmustur. Biz bu metriği ortalama üstü coğunluk olarak adlandırdık ve bu yaklasim k komsu obekleme icin neredeyse benzer bicimde uygulanmaktadir.



Grafik 1. Benzerlik ağırlığının üzerinde yer alan kayıtlar için seçilen yöntemlere göre dağılımlar.

Buna rağmen 20 dokümanın her biri için en yüksek benzerlik değerine sahip kaydın, önceden verilen sınıf değerleri ile örtüşmesi karşılaştırıldığında Şekil 2'de görülebileceği gibi Kosinüs benzerliği ile yapılan hesaplamalarda %80, Jaccard benzerliği ile yapılan hesaplamalarda %65 ve aynı şekilde PMI benzerliği ile yapılan hesaplamalarda ise %65 doğruluk olduğu tespit edilmiştir. Metrik olarak en yüksek benzerlik değerlerinin doğru olarak tespiti baz alındığında grafik aşağıdaki gibi oluşmaktadır.



Grafik 2. En yüksek ağırlık değerlerine sahip dokümanlar için doğru sınıflandırma oranları.

Dokümanlar için tespit edilen Anlamlı kelimelerin ağırlıkları değiştirilerek tüm hesaplamalar anlamlı kelimelerin %100'ü kullanılarak, %50'si kullanılarak, %25'i kullanılarak kullanılarak yeniden hesaplanmıştır. Kosinüs ve Jaccard benzerlikleri için sonuçlarda olumlu veya olumsuz hiçbir değişiklik olmazken, PMI benzerliği ile yapılan hesaplamalar iyi temsil edildikleri takdirde en iyi sonucu verebilecek iyileşmeler göstermiştir.

Tüm hesaplamaları sayısal olarak ifade eden tablolar Şekil 2, Şekil 3 ve Şekil 4'teki gibidir:

DocumentID	Önceden Verilmiş Sınıf	EducationCosine	SportsCosine	EducationJaccard	SportsJaccard	EducationPMI	SportsPMI	Kosinus Sınıf Eşleşmesi	Jaccard Sınıf Eşleşmesi	PMI Sınıf Eşleşmesi
0	Education	**0.148990**	0.119535	0.745747	**0.762551**	**44,374625**	26,469980	DOĞRU	YANLIŞ	DOĞRU
1	Education	**0.185445**	0.086015	**0.714440**	0.685026	11,166056	**11,739412**	DOĞRU	DOĞRU	YANLIŞ
2	Education	**0.155747**	0.082146	**0.715165**	0.710039	11,618104	**12,719975**	DOĞRU	DOĞRU	YANLIŞ
3	Education	**0.196831**	0.112581	0.730259	**0.750354**	15,628816	**16,296350**	DOĞRU	YANLIŞ	YANLIŞ
4	Education	**0.175772**	0.078806	0.751890	**0.778283**	**20,722472**	19,401942	DOĞRU	YANLIŞ	DOĞRU
5	Education	**0.121591**	0.082397	**0.724264**	0.711810	18,134948	**30,589308**	DOĞRU	DOĞRU	YANLIŞ
6	Education	**0.152336**	0.119817	0.766477	**0.783921**	24,366328	**29,084808**	DOĞRU	YANLIŞ	YANLIŞ
7	Education	**0.196745**	0.097079	**0.753952**	0.743736	12,881452	**18,775050**	DOĞRU	DOĞRU	YANLIŞ
8	Education	**0.085504**	0.063208	**0.763954**	0.738666	15,762737	**18,257641**	DOĞRU	DOĞRU	YANLIŞ
9	Education	**0.121551**	0.040135	**0.649838**	0.636833	9,209429	**11,369741**	DOĞRU	DOĞRU	YANLIŞ
10	Sports	**0.158364**	0.154848	0.731515	**0.741891**	6,711039	**25,841193**	DOĞRU	DOĞRU	DOĞRU
11	Sports	0.036335	**0.128739**	0.694522	**0.698535**	2,372627	**7,111401**	YANLIŞ	DOĞRU	DOĞRU
12	Sports	0.086458	**0.119428**	0.748884	**0.790770**	8,832967	**25,971739**	YANLIŞ	DOĞRU	DOĞRU
13	Sports	0.094488	**0.070117**	0.667213	**0.671085**	9,563137	**14,050180**	YANLIŞ	DOĞRU	DOĞRU
14	Sports	0.081875	**0.083002**	0.722416	**0.756575**	9,056293	**23,187546**	DOĞRU	DOĞRU	DOĞRU
15	Sports	0.051908	**0.075485**	0.708461	**0.746956**	6,882267	**16,444478**	DOĞRU	DOĞRU	DOĞRU
16	Sports	0.080408	**0.067922**	0.740780	**0.803741**	11,908392	**20,782580**	YANLIŞ	DOĞRU	DOĞRU
17	Sports	0.135863	**0.128200**	**0.757721**	0.755943	9,351473	**15,076442**	YANLIŞ	YANLIŞ	DOĞRU
18	Sports	0.074422	**0.187578**	0.764437	**0.781579**	4,567182	**16,318355**	DOĞRU	DOĞRU	DOĞRU
19	Sports	0.072597	**0.115266**	0.765270	**0.770100**	8,344630	**29,925445**	YANLIŞ	DOĞRU	DOĞRU
								Doğruluk Oranı:**0,7%**	Doğruluk Oranı:**0,75%**	Doğruluk Oranı:**0,6%**

Şekil 2. Anlamlı kelimelerin %100'ü ile yapılan hesaplama.

Şekil 2'deki tabloda verilen kolonlarda Kosinüs, Jaccard ve PMI benzerlik ölçütlerinin ortalama benzerlik değerlerinin üzerindeki kayıtlara dair sonuçlar verilmiştir. Örneğin; 0 numaralı doküman için Kosinüs benzerliği için katsayılar baz alınmış ve Sports alanına benzerliğin 0.119535 ve Education alanına benzerliğin 0.148990 olduğu görülmüştür. Bu durumda Kosinüs benzerliği kıstas alındığında dokümanın Education alınana daha fazla benzer olduğu sonucuna ulaşılmıştır. Önceden verilmiş sınıf bilgisi ile örtüşen durumlara pozitif işaretleme, uyuşmayan durumlara ise negatif işaretleme yapılmıştır. Tablodaki Kosinüs Sınıf Eşleşmesi, Jaccard Sınıf Eşleşmesi ve PMI Sınıf Eşleşmesi kolonlarının dip toplamında ise doğru tahmin edilen sınıfların tüm kolonlardaki toplamlarıyla benzerlik tahminleri için doğruluk oranları paylaşılmıştır. Şekil 2'de veriler oluşturulurken, bu doküman kümeleri için tespit edilen anlamlı kelimelerin tamamı yani %100'ü kullanılmıştır.

DocumentID	Önceden Verilmiş Sınıf	EducationCosine	SportsCosine	EducationJaccard	SportsJaccard	EducationPMI	SportsPMI	Kosinus Sınıf Eşleşmesi	Jaccard Sınıf Eşleşmesi	PMI Sınıf Eşleşmesi
0	Education	**0.148990**	0.110535	0.745747	**0.762551**	**5,417832**	1,625349	DOĞRU	YANLIŞ	DOĞRU
1	Education	**0.185445**	0.086015	**0.714440**	0.685026	**3,232620**	0,572067	DOĞRU	DOĞRU	DOĞRU
2	Education	**0.155747**	0.082146	**0.715165**	0.710039	2,788176	**3,813781**	DOĞRU	DOĞRU	YANLIŞ
3	Education	**0.196831**	0.112581	0.730259	**0.750354**	**8,035875**	3,778957	DOĞRU	YANLIŞ	DOĞRU
4	Education	**0.175772**	0.078806	0.751890	**0.778283**	**20,722472**	19,401942	DOĞRU	YANLIŞ	DOĞRU
5	Education	**0.121591**	0.082397	**0.724264**	0.711810	4,920718	**5,147165**	DOĞRU	DOĞRU	YANLIŞ
6	Education	**0.152336**	0.119817	0.766477	**0.783921**	4,177256	**5,537637**	DOĞRU	YANLIŞ	YANLIŞ
7	Education	**0.196745**	0.097079	**0.753952**	0.743736	**4,802463**	1,559565	DOĞRU	DOĞRU	DOĞRU
8	Education	**0.085504**	0.063208	**0.763954**	0.738666	**3,313573**	1,641944	DOĞRU	DOĞRU	DOĞRU
9	Education	**0.121551**	0.040135	**0.649838**	0.636833	**3,480649**	1,010427	DOĞRU	DOĞRU	DOĞRU
10	Sports	**0.158364**	0.154848	0.731515	**0.741891**	6,711039	**25,841193**	DOĞRU	DOĞRU	DOĞRU
11	Sports	0.036335	**0.128739**	0.694522	**0.698535**	0,461917	**1,906267**	YANLIŞ	DOĞRU	DOĞRU
12	Sports	0.086458	**0.119428**	0.748884	**0.790770**	1,925660	**9,567742**	YANLIŞ	DOĞRU	DOĞRU
13	Sports	0.094488	**0.070117**	0.667213	**0.671085**	**2,372607**	2,200213	YANLIŞ	DOĞRU	YANLIŞ
14	Sports	0.081875	**0.083002**	0.722416	**0.756575**	**3,586787**	0,140943	DOĞRU	DOĞRU	YANLIŞ
15	Sports	0.051908	**0.075485**	0.708461	**0.746956**	**1,486624**	1,331948	DOĞRU	DOĞRU	YANLIŞ
16	Sports	0.080408	**0.067922**	0.740780	**0.803741**	**4,704320**	-0,298912	YANLIŞ	DOĞRU	YANLIŞ
17	Sports	0.135863	**0.128200**	**0.757721**	0.755943	**1,939874**	1,450643	YANLIŞ	YANLIŞ	YANLIŞ
18	Sports	0.074422	**0.187578**	0.764437	**0.781579**	**2,185982**	1,655232	DOĞRU	DOĞRU	YANLIŞ
19	Sports	0.072597	**0.115266**	0.765270	**0.770100**	**2,522206**	1,492643	YANLIŞ	DOĞRU	YANLIŞ
								Doğruluk Oranı:**0,7%**	Doğruluk Oranı:**0,75%**	Doğruluk Oranı:**0,5%**

Şekil 3. Anlamlı kelimelerin %50'si ile yapılan hesaplama.

Şekil 3'deki tabloda verilen kolonlarda Kosinüs, Jaccard ve PMI benzerlik ölçütlerinin ortalama benzerlik değerlerinin üzerindeki kayıtlara dair sonuçlar verilmiştir. Örneğin; 2 numaralı doküman için PMI benzerliği için katsayılar baz alınmış ve Sports alanına benzerliğin 0 ve Education alanına benzerliğin 0 olduğu görülmüştür. Bu durumda PMI benzerliği kıstas alındığında dokümanın hangi sınıfa ait olduğu ile ilgili bir yorum yapılamamaktadır. Bu durumda olan tüm kayıtları N/A olarak işaretleyerek, benzerlik tahmin doğruluğunu düşürecek bir etki oluşturulmuştur. Tablodaki Kosinüs Sınıf Eşleşmesi, Jaccard Sınıf Eşleşmesi ve PMI Sınıf Eşleşmesi kolonlarının dip toplamında ise doğru tahmin edilen sınıfların tüm kolonlardaki toplamlarıyla benzerlik tahminleri için doğruluk oranları paylaşılmıştır. Şekil 3'de veriler oluşturulurken, bu doküman kümeleri için tespit edilen anlamlı kelimelerin %50'si kullanılmıştır.

DocumentID	Given Class	EducationCosine	SportsCosine	EducationJaccard	SportsJaccard	EducationPMI	SportsPMI	Kosinus Sınıf Eşleşmesi	Jaccard Sınıf Eşleşmesi	PMI Sınıf Eşleşmesi
0	Education	**0.148990**	0.110535	0.745747	**0.762551**	**7,024364**	1,264385	DOĞRU	YANLIŞ	DOĞRU
1	Education	**0.185445**	0.086015	**0.714440**	0.685026	0,000000	0,000000	DOĞRU	DOĞRU	N/A
2	Education	**0.155747**	0.082146	**0.715165**	0.710039	1,968963	**7,678957**	DOĞRU	DOĞRU	YANLIŞ
3	Education	**0.196831**	0.112581	0.730259	**0.750354**	**5,422693**	1,728771	DOĞRU	YANLIŞ	DOĞRU
4	Education	**0.175772**	0.078806	0.751890	**0.778283**	**8,222094**	3,025349	DOĞRU	YANLIŞ	DOĞRU
5	Education	**0.121591**	0.082397	**0.724264**	0.711810	**1,247646**	0,532192	DOĞRU	DOĞRU	DOĞRU
6	Education	**0.152336**	0.119817	0.766477	**0.783921**	**8,702308**	6,482892	DOĞRU	YANLIŞ	DOĞRU
7	Education	**0.196745**	0.097079	**0.753952**	0.743736	**3,455721**	2,762756	DOĞRU	DOĞRU	DOĞRU
8	Education	**0.085504**	0.063208	**0.763954**	0.738666	**4,766372**	2,593156	DOĞRU	DOĞRU	DOĞRU
9	Education	**0.121551**	0.040135	**0.649838**	0.636833	**1,727860**	0,000000	DOĞRU	DOĞRU	DOĞRU
10	Sports	**0.158364**	0.154848	0.731515	**0.741891**	2,045763	**7,095904**	DOĞRU	DOĞRU	DOĞRU
11	Sports	0.036335	**0.128739**	0.694522	**0.698535**	0,864385	**2,881285**	YANLIŞ	DOĞRU	DOĞRU
12	Sports	0.086458	**0.119428**	0.748884	**0.790770**	0,000000	**5,665787**	YANLIŞ	DOĞRU	DOĞRU
13	Sports	0.094488	**0.070117**	0.667213	**0.671085**	3,010149	**3,021622**	YANLIŞ	DOĞRU	DOĞRU
14	Sports	0.081875	**0.083002**	0.722416	**0.756575**	**5,744809**	0,081885	DOĞRU	DOĞRU	YANLIŞ
15	Sports	0.051908	**0.075485**	0.708461	**0.746956**	2,045763	**6,680866**	DOĞRU	DOĞRU	DOĞRU
16	Sports	0.080408	**0.067922**	0.740780	**0.803741**	2,319460	**3,112042**	YANLIŞ	DOĞRU	DOĞRU
17	Sports	0.135863	**0.128200**	**0.757721**	0.755943	2,477956	**7,417081**	YANLIŞ	YANLIŞ	DOĞRU
18	Sports	0.074422	**0.187578**	0.764437	**0.781579**	3,068645	**8,033831**	DOĞRU	DOĞRU	DOĞRU
19	Sports	0.072597	**0.115266**	0.765270	**0.770100**	4,091527	**9,474473**	YANLIŞ	DOĞRU	DOĞRU
								Doğruluk Oranı: **0,7%**	Doğruluk Oranı:**0,75%**	Doğruluk Oranı:**0,89%**

Şekil 4. Anlamlı kelimelerin %25'i ile yapılan hesaplama.

Şekil 4' teki tabloda verilen kolonlarda Kosinüs, Jaccard ve PMI benzerlik ölçütlerinin ortalama benzerlik değerlerinin üzerindeki kayıtlara dair sonuçlar verilmiştir. Örneğin; 2 numaralı doküman için PMI benzerliği için katsayılar baz alınmış ve Sports alanına benzerliğin 7,678957 ve Education alanına benzerliğin 1,968983 olduğu görülmüştür. Bu durumda PMI benzerliği kıstas alındığında doküman 2 için önceden verilen sınıf değeri Education olmasına rağmen PMI benzerlik ölçütü Sports alanı için daha yüksek hesaplanmaktadır. Bu durumdan ötürü benzerlik tahmini yanlış olarak işaretlenmiştir. Tablodaki Kosinüs Sınıf Eşleşmesi, Jaccard Sınıf Eşleşmesi ve PMI Sınıf Eşleşmesi kolonlarının dip toplamında ise doğru tahmin edilen sınıfların tüm kolonlardaki toplamlarıyla benzerlik tahminleri için doğruluk oranları paylaşılmıştır. Şekil 4'te veriler oluşturulurken, bu doküman kümeleri için tespit edilen anlamlı kelimelerin %25'i kullanılmıştır.

20 doküman kullanılarak yapılan hesaplamalar sonucu oluşan karmaşıklık matrisi, Kosinüs, Jaccard ve PMI benzerlik ölçütleri için Tablo 1'de ki gibi oluşmuştur.

Tablo 1. Tüm benzerlik tahminleri için karmaşıklık matrisi.

Tahmin sonucu			
Anlamlı Kelime Kullanma Oranı	Yanlış Tahmin	Doğru Tahmin	Toplam
%25	14	43	57
%50	21	36	57
%100	19	38	57
Toplam	35	79	171

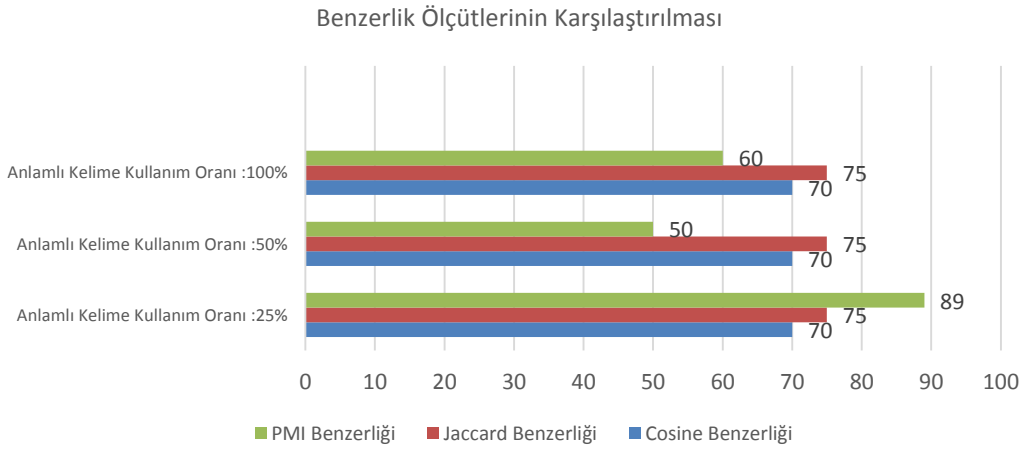
Kosinüs, Jaccard ve PMI benzerlik ölçütleri ile tüm tahminler sonucu Çizelge 6.4'teki gibi oluşan karmaşıklık matrisine göre, anlamlı kelimelerim kullanım ağırlıklarına göre ayrı ayrı sonuçlar elde edilmiştir. Tüm hesaplamalarda doğruluk oranı %46 olarak gerçekleşmiştir.

Tablo 2. PMI benzerlik ölçütü tahminleri için karmaşıklık matrisi.

Tahmin sonucu			
Anlamlı Kelime Kullanma Oranı	Yanlış Tahmin	Doğru Tahmin	Toplam
%25	14	43	57
%50	21	36	57
%100	19	38	57
Toplam	35	79	171

Sadece PMI benzerlik ölçütü baz alınarak oluşturulan karmaşıklık matrisi Tablo 2'de verilmiştir. Buna göre anlamlı kelimelerin kullanım ağırlıklarına göre ayrı ayrı hesaplanan benzerlik tahmini sonuçlarına doğruluk oranı %63 olarak gerçekleşmiştir.

Anlamlı kelimelerin dokümanın hangi oranlarda daha iyi tespit ettiğinin bulunabilmesi için sisteme farklı oranlarla hesaplamalar yaptırılmış ve aşağıdaki grafikler sonuçlar izah edilmeye çalışılmıştır.



Grafik 3. Anlamlı kelimelerin kullanım yüzdelerine göre başarı oranlarının dağılımı.

Bu çalışmada temel olarak metin benzerliği konusunda kullanılan Kosinüs ve Jaccard benzerlik ölçütleri, PMI yaklaşımı ile karşılaştırılmıştır. Sınıfları önceden belirlenmiş dokümanlar üzerinde bu metriklere göre sonuçları değerlendirilmiştir.

PMI, özellikle anahtar kelime seçim oranlarına duyarlı olduğu, bununda taşıdığı bilgi miktarıyla alakalı olduğu görülmektedir. Doğal olarak %25 anlamlı kelime oranında da diğer metriklere göre çok daha iyi sonuçları vermiştir. Anlamlı kelime kullanım oranlarına göre kullanılan 3 yöntemin başarı oranları yukarıdaki Grafik 3'te sunulmuştur. Bu durum göz önüne alındığında PMI'ın öbeklemede benzerlik ölçütü olarak kullanımının iyi sonuçlar vereceği gözlemlenmiştir. Bundan sonraki çalışmalarda bu konu irdelenecektir.

KAYNAKLAR

A. Akilan,(2015) "Text mining: Challenges and future directions," 2015 2nd International Conference on Electronics and Communication Systems (ICECS), Coimbatore, 2015, pp. 1679-1684. doi: 10.1109/ECS.2015.7124872

A. I. Kadhim, Y. Cheah and N. H. Ahamed, "Text Document Preprocessing and Dimension Reduction Techniques for Text Document Clustering," 2014 4th International Conference on Artificial Intelligence with Applications in Engineering and Technology, Kota Kinabalu, 2014, pp. 69-73.
doi: 10.1109/ICAIET.2014.21

Dr.S.Vijayarani , International Journal of Computer Science & Communication Networks,Vol 5(1),7-16

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

K. Aas and L. Eikvil, Text Categorisation, "A Survey, Technical Report Raport NR 941," Norwegian Computing Center, 1999

R. Agrawal and R. Srikant, "Fast Algorithms for Mining Association Rules in Large Databases," Proc. 20th Int'l Conf. Very Large Data Bases (VLDB '94), pp. 478-499, 1994.

R. Baeza-Yates and B. Ribeiro-Neto, "Modern Information Retrieval," Addison Wesley, 1999.

Y. Li, C. Zhang, and J.R. Swan, "An Information Filtering Model on the Web and Its Application in Jobagent," Knowledge-Based Systems, vol. 13, no. 5, pp. 285-296, 2000.

F. Hu and Y. Zhang, "Text Mining Based on Domain Ontology," 2010 International Conference on E-Business and E-Government, Guangzhou, 2010, pp. 1456-1459. doi: 10.1109/ICEE.2010.370

G. G. İşgüder-Şahin, H. R. Zafer and E. Adah, "Polarity detection of Turkish comments on technology companies," 2014 International Conference on Asian Language Processing (IALP), Kuching, 2014, pp. 136-139. doi: 10.1109/IALP.2014.6973514

T. Kocatekin and D. Ünay, "Text mining in radiology reports," 2013 21st Signal Processing and Communications Applications Conference (SIU), Haspolat, 2013, pp. 1-4.
doi: 10.1109/SIU.2013.6531400

B. Issac and W. J. Jap, "Implementing spam detection using Bayesian and Porter Stemmer keyword stripping approaches," TENCON 2009- 2009 IEEE Region 10 Conference, Singapore, 2009, pp. 1-5.
doi: 10.1109/TENCON.2009.5396056

B. Dadachev, A. Balinsky, H. Balinsky and S. Simske, "On the Helmholtz Principle for Data Mining," 2012 Third International Conference on Emerging Security Technologies, Lisbon, 2012, pp. 99-102.
doi: 10.1109/EST.2012.11

B. Dursun and A. C. Sonmez, "Türkçe metin benzerlik hesaplamasi için yeni bir yöntem," 2008 IEEE 16th Signal Processing, Communication and Applications Conference, Aydin, 2008, pp. 1-4.
doi: 10.1109/SIU.2008.4632581

Heyong Wang, Ming Hong, Supervised Hebb rule based feature selection for text classification, Information Processing & Management,Volume 56, Issue 1, 2019, Pages 167-191, ISSN 0306-4573, https://doi.org/10.1016/j.ipm.2018.09.004.

Kisayol, Ahmet & Turan, Metin. (2018). Paragraf Tabanlı Çıkarımsal Özetlemede Öbekleme Kullanan İki Yeni Yöntemin Kıyaslanması. Düzce Üniversitesi Bilim ve Teknoloji Dergisi. 6. 1047-1057. 10.29130/dubited.418453.

L. Guthrie, E. Walker. Document Classification by Machine: Theory and Practice. COLING, 1994

Xu, Yan & Jones, Gareth & Li, Jintao & Wang, Bin & Sun, Chunming. (2007). A Study on Mutual Information-based Feature Selection for Text Categorization. Journal of Computational Information Systems. 3.