Araştırma Makalesi

# İYİLEŞTİRİLMİŞ OTOMATİK ANAHTAR KELİME ÇIKARIMI (BRAKE)

**Ahmet Sina BİRDEVRİM[†], Ali BOYACI[††], Dena Ahmed S. Al THANİ[‡]**

[†] İstanbul Ticaret Üniversitesi, Bilgisayar Mühendisliği Ana Bilim Dalı, İstanbul, Türkiye

[††]İstanbul Ticaret Üniversitesi, Bilgisayar Mühendisliği Ana Bilim Dalı, İstanbul, Türkiye

[‡] Hamad Bin Khalifa University, College of Science and Engineering, Doha, Qatar

[†]**asina.birdevrim@istanbulticaret.edu.tr,** [††]**aboyaci@ticaret.edu.tr,** [‡]**dalthani@qf.org.qa**

## ÖZET

Anahtar kelimeler ve cümleler, çok sayıda metin tabanlı materyalin analizini kolaylaştırdığı gibi istenen bilgiye hızlı ve kolay erişimin sağlanmasında da önemli rol oynarlar. Bu verileri çıkarmak için otomatik anahtar kelime algoritmaları kullanılabilir. Otomatik anahtar kelime çıkarma algoritmaları; belirli bir metinde yer alan en açık kelimeleri veya cümleleri ayıklamak olarak tanımlanabilir. Bu amaç için en sık kullanılan algoritmalar TF-IDF ve RAKE dir. Bu yöntemler pek çok metinde kullanılan aynı anlamı taşıyan farklı kelimeleri göz önüne almamaktadır.

Bu çalışmada geliştirilen algoritma ile verilen bir metindeki eşanlamlı kelimler tek bir çatı altında toplanarak bu kelimelerin sıklığı arttırılmıştır. Uygulanan bu yöntem diğer algoritmalar karşılaştırılmıştır ve sonuçlar ortaya konulmuştur.

**Anahtar Kelimeler:** Anahtar kelime çıkarımı, BRAKE, RAKE, TF-IDF

## BRAKE: BETTER RAPID AUTOMATED KEYWORD EXTRACTION

## ABSTRACT

Keywords and sentences play an important role in providing quick and easy access to desired information, as well as facilitating the analysis of a large number of text-based materials. Automatic keyword algorithms can be used to extract this data. Automatic keyword extraction algorithms extract the most explicit words or phrases in a given text. The most commonly used algorithms for this purpose are TF-IDF and RAKE. However these methods do not take into account for the different words used for the same meaning in many texts.

In this study, an algorithm is developed to gather synonymous words under a single word to increase the frequency of the terms with same meaning. This method is compared with other algorithms and the results are presented.

**Keywords:** BRAKE, keyword extraction, RAKE, TF-IDF

| | | |
|---|---|---|
| **Geliş/Received** | **:** | **25.05.2018** |
| **Gözden Geçirme/Revised** | **:** | **25.06.2018** |
| **Kabul/Accepted** | **:** | **13.07.2018** |

## 1. GİRİŞ

Dijital medyada, özellikle metin tabanlı içerikler hızla büyümektedir. Dijital metin tabanlı medyadaki en büyük zorluklardan biri, okuyucuların ilgilendiği bilgileri bulmaktır. Bilgisayarlar bu arşivleri aramak için yeterince hızlı olsa da, konuları üzerinden gruplanmış metinler üzerinde aramalar daha doğru sonuçlar üretmektedir. Bu metinleri gruplamak için genellikle anahtar kelimeler kullanılır.

Anahtar kelimeler, bir metni temsil eden en açık kelime veya kelime grubudur. Anahtar kelimeler okuyucuya aramakta olduğu veri hakkında bir ön fikir verir. Aynı zamanda anahtar kelimelerin kullanımı ile bilgiye hızlı ve doğru bir şekilde erişimi sağlar. Bu sayede okuyucu tüm metni gözden geçirmek yerine metin içeresinde geçen ve metni temsil eden anahtar kelimeler üzerinden metni okuyup okumayacağını karar verir.

Genellikle, anahtar kelimeler bir belgenin yazarları veya yayıncıları tarafından el ile seçilir. Bu durum çıkarılan anahtar kelimelerde doğruluk oranında insan faktörünü ön plana çıkararak oluşabilecek hata olasılığını arttırmakta olup okuyucunun aramakta olduğu bilgiye erişememesine sebep olabilir. Bunun yanı sıra el ile anahtar kelime çıkarma işlemleri harcanan zaman bakımından da maliyetli bir iştir.

Ayrıca, çalışmalar yazar tarafından atanan anahtar kelimelerin 19%'unun makaleye dahil olmadığını göstermiştir (Kim vd., 2010). Bu nedenlerden dolayı, veri madenciliğinin bir dalı olan verilerin otomatik olarak çıkarılması giderek önem kazanmakta ve otomatik anahtar kelime çıkarma algoritmaları ile bu problem yüksek oranda aşılabilmektedir.

Anahtar kelime çıkarma metin madenciliği alanında önemli bir görevdir. Denetlenen ve denetlenmeyen makine öğrenimi, istatistiksel yöntemler ve dilbilimsel yöntemler gibi pek çok anahtar kelime yaklaşımı bulunmaktadır (Siddiqi, 2015). Bu yöntemler kısaca şu şekilde özetlenebilir:

**Dilbilimsel Yaklaşım:** Bu yaklaşım, anahtar kelimelerin tespitinde ve metin belgelerinden çıkarılmasında kelimelerin dil özelliklerinin kullanılmasına dayanmaktadır. Temel olarak sözcüksel ve sözdizimsel analizleri içerir (Bharti & Babu, 2017). Sözcüksel analiz için tree tagger, WordNett, n-gramlar gibi kaynaklar kullanılır (Ercan & Cicekli, 2007). Sözdizimsel analiz için de isim cümlesi (NP), yığın (Parsing) kaynak olarak kullanılırlar.

**İstatistiksel Yaklaşım:** Bu yaklaşım genellikle dilsel temelli istatistiksel verileri içerir. Bu yöntem için eğitim verisi gerekmez ve bu yöntemler dilden bağımsızdır. Belgedeki kelimelerin istatistikleri, anahtar kelimeleri tanımlamak için kullanılır (Beliga, 2014). Meitndeki anahtar kelimeleri bulmak için N-gram istatistiksel verileri kullanılır. Bunun dışında, diğer teknikler arasında kelime frekansı, terim frekansı (TF) (Luhn, 1957), sözcük eşzamanlılığı, terim frekansı belge frekansı (TF-IDF) ve PAT-Tree (Chien, 1997) gibi yöntemler de vardır. Bu metotlar arasında terim sıklığı en yaygın kullanılan yöntemdir.

**Makina Öğrenimi Yaklaşımı:** Bu yaklaşım genellikle denetimli öğrenme yöntemlerini kullanır. Bu yöntemlerde, eğitim dokümanları kullanılarak anahtar kelimeler çıkartılır ve geliştirilen model farklı bir veri seti üzerinden test edilir. Uygun bir model oluşturulduktan sonra, yeni dokümanlarda anahtar kelimeleri bulmak için kullanılır. Öte yandan, denetlenen öğrenme yöntemlerinin geniş bir belge kümesi gerektirmesi nedeniyle modeli oluşturmak kolay değildir. Bu kümenin bulunmadığı durumlarda, denetimsiz ve yarı denetimsiz öğrenme metotları kullanılmaktadır (Siddiqi, 2015). Naïve Bayes, SVM (Zhang, 2006), C4.5, Bagging (Hulth, 2003), KEA, KEA++ en fazla öne çıkan denetimli öğrenme algoritmalarındandır (Medelyan & Witten, 2006).

**Hibrid Yaklaşım:** Bu yaklaşım ağırlıklı olarak Dil, İstatistik ve Makine Öğrenme yaklaşımlarının en iyi özelliklerini elde etmek için tasarlanmıştır. Anahtar kelime çıkarma görevinde, kelimelerin konumu, uzunluğu, düzen özellikleri, html etiketleri vb. gibi bazı sezgisel bilgileri kullanır (Bharti & Babu, 2017).

## 2. OTOMATİK ANAHTAR KELİME ÇIKARMA ALGORİTMALARI

Otomatik anahtar kelime çıkarma algoritmalarını, belirli bir algoritmanın en belirgin biçimini yansıtan kelimeleri veya sözcük gruplarını çıkarmak olarak tanımlanabilir.

TF-IDF ve RAKE gibi algoritmalar en fazla kullanılan algoritmalar arasındadır. Bu algoritmaların temel amacı metinlerden anlamlı, hızlı ve metni en doğru şekilde temsil edebilecek kelime veya sözcükler çıkarılmasıdır.

### 2.1 TF-IDF

TF-IDF, belgede bir terimin önemini gösteren ve kelimenin metin içinde önemli olup olmadığını belirlemek için kullanılan istatistiksel yöntemle hesaplanan bir ağırlık etkeni olup, en çok bilinen ve en yaygın kullanılan terim çıkarma algoritmalarından biridir (Guan, 2016). TF-IDF yöntemi terim frekansı ve ters belge frekansı yöntemlerinin birlikte kullanılmasıdır.

Terim frekansı (TF), belirli bir terimin bir belgenin içinde geçme sıklığı olan terim frekansını ifade eder. İstatistiksel olarak, frekans ne kadar yüksek olursa, potansiyel olarak önemli bir terim olduğu varsayılmaktadır.

Bir belgede en sık bulunan kelimeler kendi başlarına anlam ifade etmeyebilirler. Yalnızca kelime frekansı ile değerlendirilme yapıldığı durumlarda anahtar kelimeler yanlış tespit edilebilir. Bu tip terimlerin ağırlığını azaltmak için ters belge frekansı (IDF) kullanırlar. IDF, belgelerindeki yüksek sıklıkta önemsiz ifadeleri dengelemenin bir yoludur (Jing vd., 2002).

TF (t, d) aşağıdaki formülde gösterilen belge d'deki t kelimesinin sıklığını ifade eder (Adji vd., 2014).

$$tf(t,d) = \frac{f(t)}{n} \quad (1)$$

- $\boldsymbol{tf(t,d)}$ terim sıkılığını (frekansı)
- $\boldsymbol{n}$, verilen belgedeki terimlerin toplam sayısı
- $\boldsymbol{f(t)}$, terim sıklığı (frekansı).
- $\boldsymbol{d}$, doküman.

$$idf(t) = \log \frac{|D|}{|\{d : t \in d\}s|} \quad (2)$$

- ***D,*** korpus'taki tüm belge sayısıdır.
- ***d, t*** terimini içeren belgelerin miktarıdır.
- ***idf(t), t*** terimi ile ilgili ters belge frekansıdır.

$$tfidf(t,d) = tf(t) * idf(t) \quad (3)$$

TF-IDF ağırlığı, bir kelimenin bir koleksiyondaki bir belgeye verdiği önemi değerlendirir ve daha yüksek TF-IDF puanları olan kelime, belgede önemlidir ve belgeyi özetleyebilir.

Örnek olarak aşağıdaki İngilizce ile yazılmış cümleleri incelersek.

- Today, cars have two-wheel drive and four-wheel drive models. But the four-wheel drives are the most popular models on sale.
- I have a four-wheel drive car
- car prices are increasing nowadays.

İlk cümle için wheel hesaplanırsa;

- TF = 3/20 = 0,15
- IDF = log(3/2) = 0,176
- TF-IDF = 0,0264 bulunmaktadır

### 2.2 RAKE

RAKE'in geliştirilmesindeki amaç, tekil belgeler üzerinde çalışan ve özellikle belirli bir dilbilgisi kurallarına uymasına gerek olmayan, birden fazla belge türü üzerinde iyi çalışan, verimli bir anahtar kelime çıkarım yöntemi geliştirilmesi olmuştur (Rose vd., 2010).

Rake; Anahtar kelimelerin sıklıkla birden fazla kelime içerdiğini, ancak standart noktalama işaretleri veya dar anlama sahip "ve", "ile" gibi etkisiz sözcükleri (stop words) nadiren içerdiğine ilişkin gözlemine dayanır.

RAKE, sonuçlarını elde etmek için doğal dil işleme tekniklerine dayanan yöntemlerin aksine, basit bir girdi parametresi kümesi alır. Anahtar kelimeleri tek bir geçişle otomatik olarak çıkarır ve böylece geniş bir doküman ve koleksiyon yelpazesine uygun hale getirir.

RAKE algoritması genel olarak açıklanırsa, anahtar sözcükleri çıkarmak istediğimiz metindeki kelimeler bir listeye ayrılır ve listedeki etkisiz sözcükleri kaldırılır. Listedeki kalan kelimeler aday anahtar kelimeler olarak tanımlandıktan sonra, her aday anahtar kelimenin puanı hesaplanır ve ele alınan anahtar kelime veya cümledeki, kelime puanlarının toplamı olarak tanımlanır.

Kelime skorlarını hesaplamak frekans ve derece değerleri bulunur. (1) kelime frekansı(freq(w)), (2) kelime derecesi (deg(w)) ve (3) derecenin frekansa oranı (deg(w)/freq(w)). (Rose vd., 2010).

Freq: Frekansın hesaplanması, ilgili aday kelimenin çıkarılan diğer aday anahtar kelimeler listesinde kaç kez geçtiğidir.

Deg**:** Bir kelimenin derecesi, aday anahtar kelimelerdeki diğer kelimelerin birlikte oluşma sıklığını temsil eder.

Frekans ve derecesi bulunduktan sonra, her aday anahtar kelimeyi listelemek için deg(w)/freq(w) metrikleri kullanılarak hesaplanır.

Son olarak, RAKE'in basitliği ve verimliliği, anahtar kelimelerin kullanıldığı birçok uygulamada kullanılabilmesini mümkün kılar. Mevcut belgelerin çeşitliliği ve hacmine bağlı olarak, RAKE, diğer analitik yöntemler için avantajlar sağlar ve bilgi işlem kaynaklarını kısıtlamaz. RAKE kelime skoru metriği, çok sık görünen ve uzun adaylarda görünmeyen kelimeleri baskınlaştırır ve ağırlıklı olarak daha uzun aday anahtar kelimelerle ortaya çıkan kelimeleri tercih eder.

## 3. BRAKE

Bir metin içerisinde bazı kelimelerin üst üste tekrarlanması okuyucunun dikkatini dağıtıp metin kalitesini düşürebilir. Bu yüzden yazarlar tarafından sık kullanılan kelimeler eş anlamlı farklı kelimelerle değiştirilip tekrarların sayısı azaltılmaktadır.

Bu durum denetimsiz yaklaşıma uygun algoritmalarda (RAKE, TF-IDF vs.) aynı anlama gelen kelime veya cümlelerin ayrı ayrı anahtar kelime seçilmesine sebep olmakta olup, metinde geçen asıl çıkartılmak istenilen doğru anahtar kelimelerin çıkartılmamasına sebep olabilir. BRAKE algoritması ile bu sorun, aynı anlam taşıyan kelimelerin tek bir kelime altına gruplanmasıyla çözülmesi hedeflenmektedir.

BRAKE algoritması ile denetimsiz yaklaşıma uygun olup, herhangi bir eğitim dokümanına bağlı kalınmadan, dilden bağımsız tekil dokümanlarda çalışan bir anahtar kelime çıkarımı algoritması oluşturulmuştur. BRAKE algoritması bu doğrultuda RAKE algoritması çalışma prensibi üzerinden çıkarılmıştır.

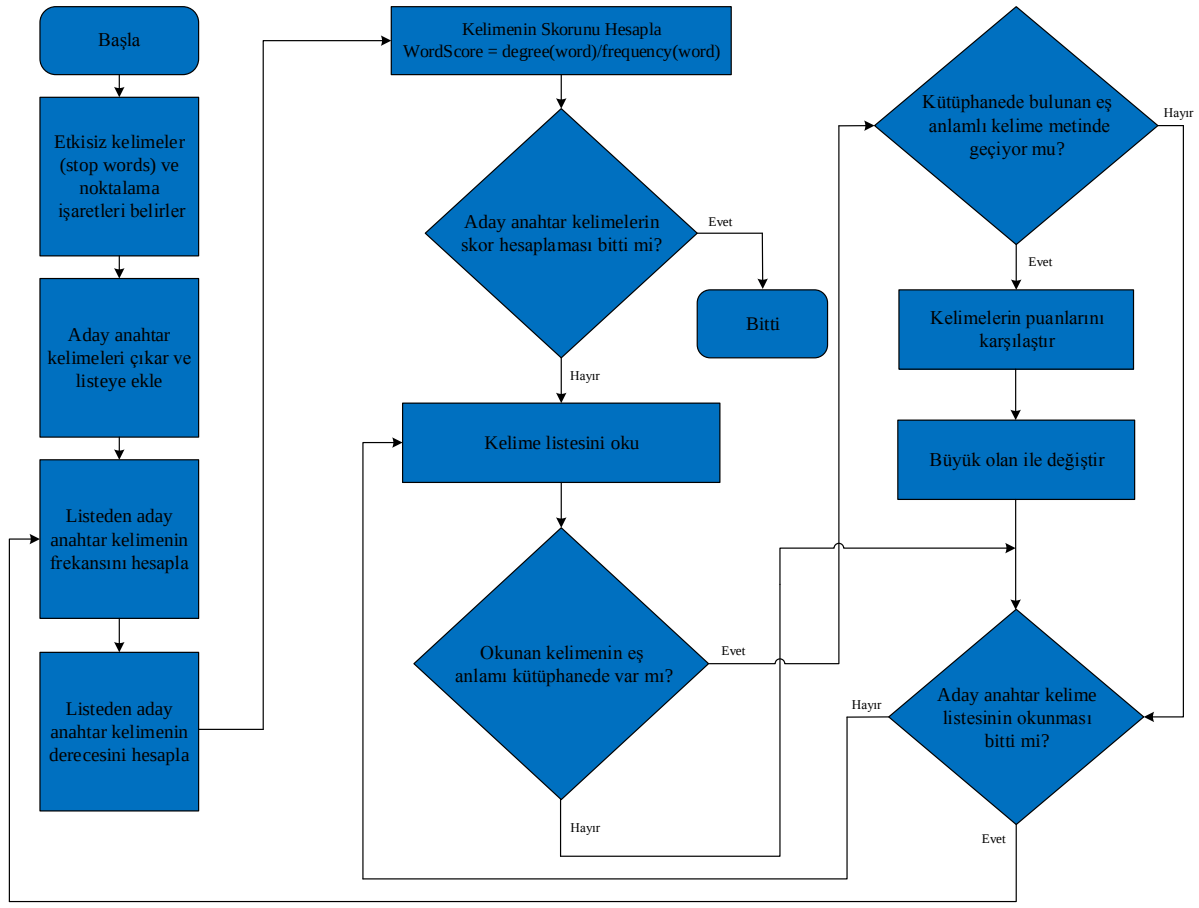
Örnek olarak, "There are two types of car engine. One is petrol and the other one is electric car engine. The one costs more is electric automobile engine[1]." İngilizce yazılmış cümlesi incelenirse:

1. Aday anahtar kelimeleri çıkarılabilmek için metnin içindeki kelimeler parçalanarak bir listede toplanır. **[there, are, two, types, of, car, engine, one, is, petrol, and, the, other, one, is, electric, car, engine, the, one, costs, more, is, electric, automobile, engine]**
2. İlgili liste içindeki önceden hazırlanmış olan stopword liste yardımı ile bunlar are, is, the gibi bağlaç veya kelimler liste içinden çıkarılır. **[types ,car, engine, petrol, electric, car, engine, costs, electric, automobile, engine]**
3. Ve oluşabilecek aday anahtar kelimeler belirlenir ve tekrardan okunarak liste tekrardan oluşturulur. **[types, car engine, petrol, electric car engine, costs, electric automobile engine]**
4. Aday anahtar kelime listesindeki her bir kelimenin puanı hesaplanır. Bu hesaplama metrik kullanarak gerçekleşir (**degree(word)/frequency(word)).**
5. Frekansın hesaplanması, ilgili aday kelimenin çıkarılan diğer aday anahtar kelimeler listesinde kaç kez geçtiğidir. Örnek: **Freq(car) = 2, Freq(engine) = 3, Freq(costs) = 1**

[1] Örneğin daha belirgin ve açıklayıcı olabilmesi için İngilizce cümle devrik olarak kurulmuştur.

6. Frekansı hesaplandıktan sonra derecesi bulunarak, kelimenin puanı çıkarılır.
7. Bir kelimenin derecesi, aday anahtar kelimelerdeki diğer kelimelerin birlikte oluşma sıklığını temsil eder. Örnek: **Deg(automobile) = 3, Deg (engine) = 8, Deg(costs) = 1**
8. word_score = **degree(word)/frequency(word)** hesabı yapılır
9. Eş anlamlı bir thesaurus listin bulunduğu kütüphaneden (db) den okunarak aday anahtar kelimelerin eş anlamları bulunur.
10. Çıkarılan eş anlamlı kelimelerin diğer aday anahtar kelimelerin içinde geçip geçmediği kontrol edilir eğer geçiyor ise işleme alınır geçmiyorsa yok sayılır.
11. İşleme alınan eş anlamlı kelimeler anlamdaş olduğu kelime ile kendisinin çıkarılan kelime puanına bakılır bu sayede dokümanda hangi anlamdaş kelime puanı yüksek ise hepsinin tek anlamda toplanacağı kelime aynı olur.
12. Belirlenen kelime mevcuttaki anlamdaşın yerine geçer ve oluşan yeni metinin frekansı ve derecesi hesaplanarak puan bulunur. **[types , automobile, engine, ,petrol, electric, automobile, engine, costs, electric, automobile, engine]**
13. Son olarak örnek puan: **word_score("automobile") = 12/4 = 3, word_score("engine") = 12/4 =3**

Bütün aday kelimelerin puanları hesaplanır ve büyükten küçüğe sıralanarak ağırlığı yüksek olan aday kelimeler anahtar kelime olarak belirlenir.




**Şekil 1.** BRAKE algoritmasına ait akış diyagramı

Adım adım gösterilen BRAKE algoritması için akış şeması Şekil 1'de gösterilmiştir.

BRAKE algoritmasının başarısı kullanılan eş anlamlılar listesine bağlıdır. Bu liste ne kadar doğru ve geniş olursa metin içerisindeki kelimelerin bulunmasında ve bulunan her bir kelimenin, doğru bir küme altında toplanmasında önemli bir etken olmaktadır.

### 3.1 RAKE ve BRAKE Algoritmasının Karşılaştırılması

RAKE ve BRAKE algoritması **"There are two types of car engine. One is petrol and the other one is electric car engine. The one costs more is electric automobile engine"** örnek cümlesi için değerlendirilerek karşılaştırıldı. Tablo 1'deki anahtar kelime çıktılarına bakıldığında, sıralamalar anahtar sözcüklerin skorlarına göre listelenir.

Burada 2. sıraya dikkat edersek, puan sırasına göre sıralanmış BRAKE algoritmasında "'automobile engine" iken RAKE algoritması "electric car engine" olur.

**Tablo1.** Araba motorları ile ilgili metinden çıkarılanlar.

| SIRA | RAKE | BRAKE |
|---|---|---|
| 1 | electric automobile engine | electric automobile engine |
| 2 | electric car engine | automobile engine |
| 3 | car engine | petrol |
| 4 | petrol | costs |
| 5 | costs | types |

## 4. YÖNTEM

TF-IDF, RAKE ve BRAKE algoritmaları, 566 adet İngilizce akademik makale kullanılmıştır. Bu makaleler çeşitli mühendislik alanlarında hakemli olarak yayın yapan dergi ve konferanslardan PDF (Portable Document Format) formatında alınmıştır. Bu makaleler kullanılarak eşleme sonuçlarına göre hassasiyet (Precision), geri çağırma (Recall) değerleri üzerinden algoritmaların performansları karşılaştırılmıştır [**Hata! Başvuru kaynağı bulunamadı.**].

Makaleler pdf formatından pdf2text aracı ile ilk olarak metin biçimine dönüştürülmüştür. Önişlemden geçen metinlerden, yazarların belirlediği anahtar kelimeler değerlendirme ölçütünde kullanılmak üzere el ile metinlerden alınmıştır. Python programlama dili kullanılarak algoritmalar gerçeklemiştir. Daha sonra, tüm makaleler bu algoritmalar ile işlenip bulunan kelimeler büyükten küçüğe puan sıralaması ile dizilmiştir. Yazar tarafından atanan anahtar kelimelerin sayısı beşten fazla ise ilk beş kelime, orijinal anahtar kelime sayısının beşten az olduğu durumlarda ise orijinal anahtar kelime sayısı kadar sözcük, makaleyi temsil eden anahtar kelime olarak belirlenmiştir..

Bu işlem sonucundaki başarı oranı değerlendirilmesinde hassasiyet (Precision), geri çağırma (Recall) ve f ölçüsü (F- Score) değerleri her makale için bulunmuştur.

Precision değeri algoritma tarafından çıkarılan doğru kelime sayısının, çıkarılan toplam kelimelere oranını verir.

$$Precision = \frac{doğru\ sayısı}{algoritmanın\ çıkarmış\ olduğu\ anahtar\ kelime\ sayısı} \quad (4)$$

Recall değeri algoritma tarafından çıkarılan doğru anahtar kelime sayısının, orijinal anahtar kelime sayısına oranına verir.

$$\text{Recall} = \frac{doğru\ sayısı}{orjinal\ anahtar\ kelime\ sayısı} \quad (5)$$

F-Score değer geri çağrım (recall) ve hassasiyet (precision) harmonik ortalamasıdır.

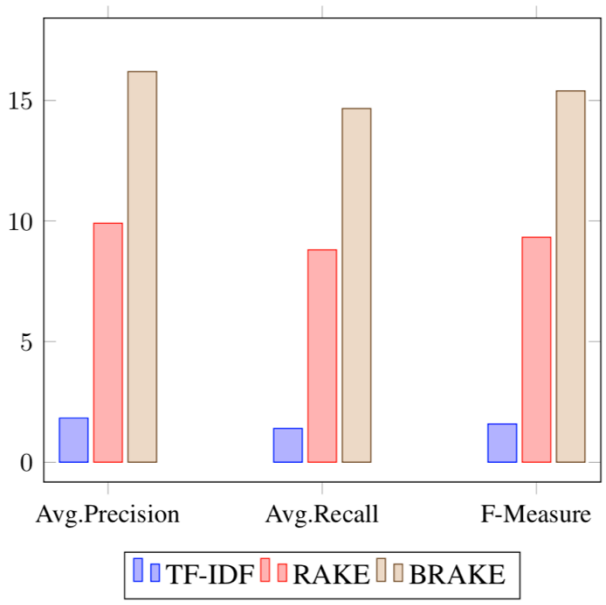
$$F - Score = \frac{2*AvgPrecision*AvgRecall}{AvgPrecision+AvgRecall} \quad (6)$$

## 5. DEĞERLENDİRME SONUCU

Yapılan değerlendirmeler sonucunda TF-IDF, Rake ve Brake algoritmaları Geri Çağırma (Recall), Hassasiyet (Precision), F-Ölçüsü (F-Measure) çıktıların ortalama değerleri Tablo 2'de yüzdelik olarak gösterilmiştir. Bu değerlendirmede orijinal anahtar kelimenin, algoritmalar tarafından çıkarılan anahtar kelime içinde geçmesi ve ya birebir aynı olması koşuluna göre hesaplanmıştır. İki durum için de eşlemenin tam olarak yapıldığı varsayılmıştır.

**Tablo 2.** Değerlendirme Sonucu

| | **Ortalama Hassasiyet** | **Ortalama Kesinlik** | **Ortalama F Ölçüsü** |
|---|---|---|---|
| **TF-IDF** | 1.8265 | 1.3965 | 1.5828 |
| **RAKE** | 9.9082 | 8.8035 | 9.3232 |
| **BRAKE** | 16.196 | 14.665 | 15.3925 |




**Şekil 2.** Ortalama Presicion, Recall ve F-Masure souçları

BRAKE algoritmasının, Tablo 2 ve Şekil 2'de gösterildiği gibi, hassasiyet (Precision), geri çağırma (Recall) ve f ölçüsüne (F-Measure) göre diğer algoritmalara göre daha başarılı sonuçlar ürettiği gözlemlenmiştir.

Bu değerlendirme yapılırken yazar tarafından atanan orijinal anahtar kelimelerin doğru olduğu varsayımı yapılmıştır.

## 6. SONUÇ VE ÖNERİLER

Bu makalede, eş anlamlı temel denetimsiz yaklaşıma dayanan BRAKE algoritması sunulmuştur. BRAKE algoritması, RAKE algoritması üzerine, eş anlama gelen farklı kelimelerin tek bir anahtar kelime alına toplanmasına dayanan bir yaklaşım ile tasarlanmıştır. Akademik makaleler üzerine, yazarların kendi seçtikleri anahtar kelimeler esas alınarak yapılan değerlendirme sonuçlarına göre BRAKE algoritması kıyaslanan diğer algoritmalara göre daha yüksek oranla eşlemeler yakalamıştır.

Yayınlanan makalelerin yapılan önişleme ile PDF formatından düz metne dönüştürmesi sırasında format ve şekillerden kaynaklı bozulmaları algoritmanın performansını olumsuz olarak etkilemektedir. Özellikle tablolar gibi anahtar kelimelerin sıklıkla geçtiği yapılardaki tekrarlar metin istatistiklerini değiştirdiği için farklı anahtar kelimelerin tespitine sebep olmuştur. Önişlemdeki başarım algoritmanın performansına doğrudan etki etmektedir.

Eş anlamlı kelimelerin tespiti sırasında, metin içerisindeki kısaltmalar göz ardı edilmiştir. Ayrıca sözcüklerin sonlarına çeşitli ekler (‘s’, ’ing’ gibi) geldiği durumlarda genellikle eşanlamlılar listesinde bir eşleme yapılamamaktadır. BRAKE algoritmasının başarısı, eşanlamlı listenin başarısına ve doğruluğuna bağlıdır. Bu sebeplerden dolayı, eklerin ayrılması ve kısaltmaların gerçek kelimeler ile değiştirilmesi ile başarı oranının artırılması mümkündür.

Ayrıca, genellikle metin içerisinde tekrar eden başlık, özet ve sonuç gibi kısımlarda geçen kelimelerin farklı şekillerde puanlanması ile algoritmanın başarısı arttırılabilir.

**KAYNAKLAR**

Adji, T. B., Abidin, Z., & Nugroho, H. A. (2014), System of negative indonesian website detection using tf-idf and vector space model. In 2014 International Conference on Electrical Engineering and Computer Science (ICEECS), 174–178.

Beliga, S. (2014). Keyword extraction: a review of methods and approaches.

Bharti, S. K., & Babu, K. S. (2017). Automatic keyword extraction for text summarization: A survey. CoRR, http://arxiv.org/abs/1704.03242

Chien, L. F. (1997). Pat-tree-based keyword extraction for chinese information retrieval. Research and Development in Information Retrieval, 20th Annual International ACM SIGIR Conference, SIGIR '97 New York, (pp. 50-58).

Ercan, G., & Cicekli, I. (2007). Using lexical chains for keyword extraction. Inf. Process. Manage., 43(6), 1705–1714.

Guan, J. (2016). A study of the use of keyword and keyphrase extraction techniques for answering biomedical questions. A thesis submitted to Macquarie University for the degree of Master of Research Department of Computing.

Hulth, A. (2003). Improved automatic keyword extraction given more linguistic knowledge. In Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing, EMNLP '03 Stroudsburg, PA, USA: Association for Computational Linguistics, (pp. 216–223).

Jing, L. P., Huang, H. K. & Shi, H. B. (2002). Improved feature selection approach tfidf in text mining. In proceedings, International conference on machine learning and cybernetics, Beijing, China, 2, 944–946.

Kim, S. N., Medelyan, O., Kan, M. Y., & Baldwin, T. (2010). Semeval-2010 Task 5: Automatic keyphrase extraction from scientific articles. In Proceedings of the 5th International Workshop on Semantic Evaluation, SemEval'10 Stroudsburg, PA, USA: Association for Computational Linguistics, (pp. 21–26).

Luhn, H. P. (1957). A statistical approach to mechanized encoding and searching of literary information. IBM Journal of Research and Development, 1(4), 309–317.

Leung, A. (2016). Evaluating automatic keyword extraction for internet reviews. Unıversıty Of Lorraıne Realself INC.

Medelyan, O., & Witten, I. H. (2006). Thesaurus based automatic keyphrase indexing. In Proceedings of the 6th ACM/IEEE-CS Joint Conference on Digital Libraries, JCDL '06 New York, NY, USA: ACM, (pp. 296-297).

Rose, S., Engel, D., Cramer, N., & Cowley, W. (2010). Automatic keyword extraction from individual documents. Text Mining: Applications and Theory, pp. 1–20.

Siddiqi, A. S. S. (2015). Keyword and keyphrase extraction techniques: A literature review. International Journal of Computer Applications, 109(2), 1–6.

Turney, P. D. (2002). Extraction of keyphrases from text: Evaluation of four algorithms. CoRR, http://arxiv.org/abs/cs.LG/0212014

Zhang, K., Xu, H., Tang, J., & Li, J. Z. (2006). Keyword extraction using support vector machine. 1, 85–96.