



Haber Zincirlerinde Tutarlılık ve Güvenilirlik Değerlendirmesi

Program Kodu: 3501

Proje No: 114E784

Proje Yürütücüsü: Yrd. Doç. Dr. Selma TEKİR

Bursiyerler:

Erhan SEZERER

Mustafa TOPRAK

TEMMUZ 2017

İZMİR

Önsöz

Bilgi keşfi ve veri madenciliği alanında önerdiğimiz Haber Zincirlerinde Tutarlılık ve Güvenilirlik Değerlendirmesi başlıklı projemize verdiği destek nedeniyle TÜBİTAK'a teşekkür ederiz.

İçindekiler

Önsöz	i
Şekil Listesi	iii
Tablo Listesi	iv
Özet	vi
Abstract	vii
1 GİRİŞ	1
2 VERİLEN BİR HABER ZİNCİRİNİN TUTARLILIK DEĞERLENDİRMESİ	2
2.1 Haber zinciri değerlendirme metodolojisi	2
2.2 Sonuçlar	16
3 KAFES (LATTICE) YAPISINA UYGUN HABER ZİNCİRLERİ ÜRETİLMESİ	17
3.1 Mevcut yöntem	17
3.2 Önerilen yöntem	18
3.3 Deneysel Sonuçlar	19
3.4 Sonuçlar	20
4 HABER GÜVENİLİRLİĞİNİN DEĞERLENDİRİLMESİ	21
4.1 Fikirlere/Gerçeklere dayanan haber kaynakları	21
4.2 Taraflı mı (is biased) faktörünün değerlendirmesi	28
4.3 Sonuçlar	29
5 SONUÇLAR VE TARTIŞMA	31
5.1 Proje kapsamındaki yayınlarmız	32

5.2	Konferans katılım faaliyetleri	32
5.3	Proje konusundaki lisansüstü çalışmalar	33
	Bibliography	35

Şekil Listesi

2.1	Dağımlık matrisi.	6
2.2	Doküman modeli.	7
2.3	Sözcük modeli.	8
2.4	İki parçalı doküman-sözcük yönlü çizgesi.	9
2.5	Doküman bazında aktivasyon değeri eğrileri.	10
2.6	Doküman bazında başlangıç değeri eğrileri.	11
2.7	Kısıt değerlerine karşı <i>minedge</i> değerleri.	12
2.8	Shahaf et. al. <i>minedge</i> değerleri.	13
2.9	Shahaf et. al. (tekrarlı doküman) <i>minedge</i> değerleri.	13
2.10	VAST 2006 Challenge <i>minedge</i> değerleri.	14
2.11	Rastgele haber zinciri 1 <i>minedge</i> değerleri.	14
2.12	Rastgele haber zinciri 2 <i>minedge</i> değerleri.	15
2.13	Rastgele haber zinciri 3 <i>minedge</i> değerleri.	15
4.1	Bir editorial makalesi için örnek skorlar.	23

Tablo Listesi

2.1	Metodoloji kapsamında dađınıklık katsayısı sonuçları.	7
2.2	Metodoloji kapsamında minedge metriđi deđerleri.	12
3.1	<i>Minedge</i> metrik deđerleri.	19
3.2	Dađınıklık katsayısı deđerleri.	19
4.1	Örnek kümülatif dađılıř fonksiyonu deđerleri.	22
4.2	Olumluluk ve olumsuzluk skorlarının derlem düzeyinde özet istatistikleri.	24
4.3	Varyans analizi sonuçları.	24
4.4	Scheffé testi sonuçları.	25
4.5	Cümle duygu etiketlemesi sonuçları.	27
4.6	Semantik yönelim skorlarının derlem düzeyinde özet istatistikleri	27
4.7	Semantik yönelim skorlarının varyans analizi ve Scheffé testi sonuçları	28
4.8	Derlem düzeyinde taraflı sözcük sayıları ve yüzdeleri	29
4.9	Tarafılık Scheffé testi sonuçları	29

Özet

Proje kapsamında haber derlemlerinden tutarlı haber zincirleri üretilmesi ve haber kaynaklarının güvenilirliğinin ölçülerek değerlendirilmesi hedeflenmiştir. İlk olarak bir haber zinciri değerlendirme metodolojisi kurulmuştur. Önerilen metodoloji literatürde mevcut birbirinden farklı ve bağımsız iki metriğin aynı çatı altında değerlendirilmesini sağlamaktadır. Tutarlı haber zincirleri üretmek üzere kafes tabanlı bir yaklaşım benimsenmiştir. Yaklaşım, haber zincirindeki dokümanlar arasındaki yumuşak geçişlere benzer şekilde kavram kafesindeki kavramlar arasında yumuşak geçişleri garanti etmeye çalışan sezgisel bir algoritmaya dayanmaktadır. Bu yöntemle üretilen zincirler yine projede önerilen metodoloji ile değerlendirilmiştir.

Haber güvenilirliği ise haber kaynağı güvenilirliği üzerinden ele alınmıştır. Güvenilirlik ölçümünde "fikirlere mi gerçeklere mi dayanmakta" ve "tarafı mı (is biased)" faktörleri baz alınmıştır. "Fikirlere mi gerçeklere mi dayanmakta" ölçümü için yeni bir gözetimsiz olasılıksal sözlük-tabanlı fikir madenciliği yaklaşımı geliştirilmiştir. Deneysel tasarımda güvenilirlik niteliği açısından doğası gereği farklı olan Editoriallar, New York Times makaleleri ve Reuters makaleleri kullanılmıştır. Elde edilen sonuçlar deneysel hipotezimiz olan bu kaynakların güvenilirlik niteliği açısından birbirinden farklı olduğu savını desteklemektedir. Ayrıca bu hesaplamaların yapılması için bir kaynak kod altyapısı da oluşturulmuştur.

Anahtar Kelimeler:

Bilgi keşfi ve veri madenciliği, haber zinciri, güvenilirlik, tutarlılık, fikir madenciliği, kafes.

Abstract

Constructing coherent news chains out of news corpora and measuring and evaluating the credibility of news sources are the main aims of this project. In the first place, a news chain evaluation methodology is established. The proposed methodology provides the assessment of two different, independent metrics under the same framework. In order to generate coherent news chains, a lattice-based approach is utilized. The proposed approach is based on the intuition that we should guarantee smooth transitions between concepts of a lattice similar to smooth transitions between the documents of a news chain. The constructed news chains in this manner are evaluated by the use of proposed evaluation methodology.

As for news credibility, it's addressed on the basis of news source credibility. Credibility measurement is performed based on the factors of "is factual or opinionated" and "is biased". To be specific, we proposed an unsupervised probabilistic lexicon-based opinion mining approach to describe a news source as "being factual or opinionated". In order to test the effectiveness of the approach, three different news sources are chosen. They are editorials, New York Times articles, and Reuters articles, which differ in their characteristic of credibility. As an experimental hypothesis we claim that these three sources are different with respect to their credibility. Our experimental results prove this claim statistically. Additionally, a source code infrastructure is established to support all these measurements.

Keywords:

Knowledge discovery and data mining, news chain, credibility, coherence, opinion mining, lattice.

1. GİRİŞ

Proje kapsamında haber derlemlerinden tutarlı haber zincirleri üretilmesi ve haber kaynaklarının güvenilirliğinin ölçülerek değerlendirilmesi hedeflenmiştir. Bu hedefleri gerçekleştirmek üzere proje iş-zaman planında belirtilen iş paketleri ile uyumlu bir şekilde ilerleme kaydedilmiştir. Bu kapsamda Bölüm 2’de, verilen bir haber zincirinin kalitesinin ölçülmesi için oluşturulan haber zinciri değerlendirme metodolojisi açıklanmıştır. Ardından Bölüm 3, kafes tabanlı bir yaklaşım ile haber zinciri üretilmesi ve üretilen zincirlerin değerlendirmesini ele alır. Bir sonraki bölüm haberin güvenilirliği üzerinedir. Güvenilirliğin ”gerçeklere mi fikirlere mi dayanmakta” ve ”tarafı mı (is biased)” faktörleri üzerinden ölçümünü ve farklı haber derlemleri için değerlendirmesini sunar. Son olarak da projenin genel bir değerlendirmesi verilmektedir.

Bundan sonraki bölümlerde her bölüm kendi içerisinde alt bölümlere ayrılmış olup her alt bölüm kendi içinde bir bütünlük içerecek şekilde sunulmuştur.

2. VERİLEN BİR HABER ZİNCİRİNİN TUTARLILIK DEĞERLENDİRMESİ

Literatürde, verilen bir haber zincirinin kalitesini (tutarlılığını) ölçmek üzere geliştirilmiş iyi kurulmuş bir metodoloji bulunmamaktadır. Mevcut iki çalışma birbirinden farklı ve bağımsız mekanizmalar önermektedir ve bir metodolojinin parçası olarak aynı çatı altında deneysel olarak doğrulanıp kıyaslanmamıştır. Ayrıca bu alanda kesin referans veriseti eksikliği vardır.

2.1 Haber zinciri değerlendirme metodolojisi

İlk olarak kesin referans (ground-truth) olarak kullanılacak haber zincirleri bulunmaya çalışılmıştır. İş Paketi 1'den de hatırlanacağı üzere Shahaf et al. 2010 [1] çalışmasında örnek olarak verilen zincir, zincirdeki haber başlıklarının The New York Times portalında aratılması sonucu oluşturulmuştur. Sözkonusu zincir sırasıyla aşağıda başlıkları verilen haber belgelerinden oluşmaktadır:

10991630.xml "Talks Over Ex-Intern's Testimony On Clinton Appear to Bog Down"
21040010.xml "CLINTON ADMITS LEWINSKY LIAISON TO JURY; TELLS NATION 'IT WAS WRONG,' BUT PRIVATE"
31066201.xml "G.O.P. Vote Counter in House Predicts Impeachment of Clinton"
41070964.xml "HE FACES A SENATE TRIAL, 2D IN HISTORY; VOWS TO DO JOB TILL TERM'S 'LAST HOUR'"
51084766.xml "Excerpts: Senators Talk About Their Votes in the Impeachment Trial"
61118570.xml "Aides Say Clinton Is Angered As Gore Tries to Break Away"
71222222.xml "As Election Draws Near, the Race Turns Mean"
81250985.xml "GORE ASKS PUBLIC FOR PATIENCE; BUSH STARTS TRANSITION MOVES; ADMINISTRATION WITHHOLDS AID"

Bu zincir, Shahaf et. al. haber zinciri olarak isimlendirilmiştir. Bu haber zincirindeki 4. sıradaki belgenin üç kopyası da zincire yerleştirilerek üretilen bir diğer zincire de Shahaf et. al. (tekrarlı doküman) haber zinciri (4. sıradaki belge üç defa tekrarlanmaktadır) ismi verilmiştir.

Bir başka kesin referans haber zinciri olarak VAST 2006 Challenge kapsamında sunulan "Alderwood hikayesi"[2] kullanılmıştır. Veriseti kapsamında bir haber zincirini oluşturan sıralı doküman listesi verilmemektedir ancak kurgulanmış senaryo dahilinde gerçekleşen olaylar tarih sırası ile verilmektedir ve bu olaylarla ilişkilendirilmiş doküman adları da verilmektedir (Birkaç doküman birden fazla olayla ilişkilendirilmiştir). Bu bilgiler kullanılarak 19 dokümandan oluşan bir haber zinciri elde edilmiştir.

Bu haber zincirlerine ek olarak The New York Times Corpus[3] derleminden eşit uzunlukta üç adet rastgele haber zinciri üretilmiştir.

Özetle haber zinciri değerlendirme metodolojisini oluşturmak üzere deneysel tasarım kapsamında aşağıda listelenen kesin referans (ground truth) haber zincirleri kullanılmıştır:

- Shahaf et. al. [1] haber zinciri
- Shahaf et. al. haber zinciri (4. sıradaki belge üç defa tekrarlanmaktadır).
- VAST 2006 Challenge [4] haber zinciri
- Rastgele haber zinciri 1
- Rastgele haber zinciri 2
- Rastgele haber zinciri 3

Haber zinciri değerlendirme metodolojisinin bir başka bileşeni zincirin kalitesini ölçecek yöntemler ve bu yöntemler sonucu hesaplanacak metriklerdir. Bu yöntemlerden ilki haber zincirinin tutarlılık değerlendirmesini (tutarlı bir bütün olup olmadığı) Shahaf et al. 2010 [1] çalışmasında ifade edilen doğrusal program üzerinden gerçekleştirir [5]. Bu doğrusal program $kTrans$ ve $kTotal$ kısıt değerlerini kullanarak $minedge$ metrik değerini üretir. Dolayısıyla $minedge$ metrik değeri haber zincirinin kalitesini ölçeceğimiz ilk metriktir.

Literatürde Soergel uzaklığını kullanan dağımlık katsayısı (dispersion coefficient) da haber zincirlerinin kalitesini ölçmek üzere önerilmiştir ([6], [7]).

İki doküman arasındaki Soergel uzaklığı, sözcük kümesindeki her bir sözcüğün bu iki dokümandaki ağırlık değerleri arasındaki farkın toplamının yine her sözcüğün bu iki dokümandaki maksimum ağırlık değeri toplamına bölünmesi ile hesaplanır:

$$D(d_1, d_2) = \frac{\sum_t |w_{t,d_1} - w_{t,d_2}|}{\sum_t \max(w_{t,d_1}, w_{t,d_2})} \quad (2.1)$$

Bir sözcüğün bir dokümandaki ağırlık değeri ise aşağıdaki formül kullanılarak hesaplanmaktadır [8]

$$w_{t,d} = \frac{(1 + \log(tf_{t,d}))(\log \frac{N}{df_t})}{\sqrt{\sum_{j=1}^{n_d} ((1 + \log(tf_{j,d}))(\log \frac{N}{df_j}))^2}} \quad (2.2)$$

Formülde $tf_{t,d}$, t sözcüğünün d dokümanındaki frekansı, df_t , t sözcüğünü içeren doküman sayısı, n_d , d dokümanındaki sözcük sayısı, N ise zincirdeki toplam doküman sayısıdır. Bu hesaplama, bir çeşit TF-IDF modeli cosine normalizasyonudur.

Soergel uzaklığı aynı dokümanlar arasında 0 değerini döndürürken 1 değerini hiç ortak sözcük içermeyen dokümanlar için üretebilir.

Proje kapsamında Soergel uzaklığı gerçekleştirilmiştir. Haber zincirinin alternatif kalite değerlendirilmesi için haber zincirini oluşturan dokümanların ikili Soergel uzaklığına dayalı olarak çizdirilen dağımlık (dispersion) grafikleri kullanılmaktadır. Bu grafikler için hesaplanan dağımlık katsayısının 1 değeri ideal bir haber zincirini ortaya koyarken, elde edilen 0'a yakın değerler ideal olmayan haber zincirlerinin göstergesi olarak kabul edilir [6].

Dağımlık katsayısı hesabına dayalı, önerilen bu yöntemin sezgisi Swanson'ın [9] CBD (complementary but disjoint) hipotezine dayanmaktadır. Yöntem, haber zincirindeki ardışık ve ardışık olmayan doküman ikililerinin Soergel uzaklıklarını değerlendirerek bir tutarlılık sonucu elde etmeye çalışır. İyi bir haber zincirinde doküman ikililerinin zincirdeki sıraları arasındaki fark arttıkça ilgili doküman ikilileri arasındaki Soergel uzaklığının da artması beklenir. Örneğin zincirde 0 ve 2 nolu

pozisyonlarda yer alan dokümanlar arasındaki Soergel uzaklığı 0 ve 3, 0 ve 4, 0 ve 5 vb. doküman ikilileri arasındaki uzaklıktan daha küçük olmalıdır. Bu yüzden dağımlık katsayısı metriği hesabında sıraları arasındaki fark artarken aralarındaki uzaklık azalan doküman ikilileri skoru düşürücü bir etki yapar.

Dağımlık katsayısı aşağıdaki formül üzerinden hesaplanmaktadır:

$$V = 1 - \frac{1}{n-2} \sum_{i=0}^{n-3} \sum_{j=i+2}^{n-1} disp(d_i, d_j) \quad (2.3)$$

Formülde, iç içe toplam sembolleri içerisinde yer alan $disp$ dağımlık değeri, doküman ikilisi arasındaki açığı belli bir eşik değerden büyük (uzaklık belli bir eşik değerden küçük) olduğunda pozitif aksi takdirde 0 olur. Pozitif olduğu durumlarda da doküman ikilisi arasındaki sıra farkı büyükse daha büyük değer alarak etki oluşturur, bir başka ifade ile dağımlık katsayısı değerini daha büyük ölçüde düşürür:

$$disp(d_i, d_j) = \begin{cases} \frac{1}{n+i-j}, & \text{if } D(d_i, d_j) > \Theta \\ 0, & \text{otherwise} \end{cases}$$

Dağımlık katsayısı hesabını daha da detaylandırmak gerekirse, n dokümandan oluşan bir zincirde n 'in 2 'li kombinasyonları sayısından n dokümanın ardışık doküman ikililerinin sayısı olan $n-1$ değerinin farkı kadar eleman üzerinden hesaplama yapılmaktadır. Bir başka ifade ile ardışık doküman ikilileri arasındaki Soergel uzaklıkları dikkate alınmaz.

Aşağıdaki örnekte $n = 8$ dokümandan oluşan bir zincirde dağımlık katsayısı hesabına dahil olan doküman ikilileri ve son değere katkı değerleri verilmektedir:

$d(0, 2) : 0.1666666$	$d(0, 3) : 0.2$	$d(0, 4) : 0.25$	$d(0, 5) : 0.3333333$	$d(0, 6) : 0.5$	$d(0, 7) : 1.0$
$d(1, 3) : 0.1666666$	$d(1, 4) : 0.2$	$d(1, 5) : 0.25$	$d(1, 6) : 0.3333333$	$d(1, 7) : 0.5$	
$d(2, 4) : 0.1666666$	$d(2, 5) : 0.2$	$d(2, 6) : 0.25$	$d(2, 7) : 0.3333333$		
$d(3, 5) : 0.1666666$	$d(3, 6) : 0.2$	$d(3, 7) : 0.25$			
$d(4, 6) : 0.1666666$	$d(4, 7) : 0.2$				
$d(5, 7) : 0.1666666$					

Tablo incelendiğinde, her sütunda yer alan doküman ikilileri arasındaki indis farkının dolayısıyla dağımlık değerinin sabit olduğu görülmektedir. Tabloda soldan sağa gidilirken sütunlardaki doküman ikililerinin indisleri arasındaki farkların ve aynı şekilde dağımlık değerlerinin de arttığı gözlemlenmektedir (indisler arasındaki fark sırasıyla 2, 3, 4, 5, 6 ve 7 ve ilgili dağımlık değerleri 0.1666666, 0.2, 0.25, 0.3333333, 0.5 ve 1'dir.).

Shahaf et. al. haber zinciri için Soergel uzaklığının eşik değeri 0.2 olarak belirlenerek üretilen dağımlık matrisi aşağıda verilmektedir:

```
IDs 1250985 1222222 1118570 1084766 1070964 1066201 1040010
1222222
1118570 0.0
1084766 0.0 0.0
1070964 0.25 0.2 0.0
1066201 0.0 0.0 0.0 0.0
1040010 0.0 0.3333333333333333 0.0 0.0 0.0
991630 0.0 0.0 0.0 0.0 0.2 0.0
```

Şekil 2.1: Dağımlık matrisi.

Bu matris kullanılarak yapılan örnek hesaplama ve elde edilen dağımlık katsayısı değeri verilmektedir:

$$V = (1-(1/6)*(0.25+(2*0.2)+0.3333333333333333))$$

$$\text{Dağımlık katsayısı (V)} = 0.8361111111111111$$

Bu şekilde Soergel uzaklığının eşik değeri sırasıyla 0.22 ve 0.25 olarak belirlenerek deneysel ortamda yer alan her bir haber zinciri için dağımlık katsayısı ölçümü yapılmıştır. Ölçüm sonuçları aşağıdaki tabloda sunulmaktadır:

Tablo 2.1: Metodoloji kapsamında dađınıklık katsayısı sonuçları.

	Zincir büyüklüğü	Eşik deęer 0.22	Eşik deęer 0.25
Shahaf et. al.	8	0.836111	0.483333
Shahaf et. al. (tekrarlı doküman)	10	0.984375	0.615625
VAST 2006 Challenge	19	0.862132	0.571804
Rastgele haber zinciri 1	8	1.0	0.394444
Rastgele haber zinciri 2	8	0.883333	0.316666
Rastgele haber zinciri 3	8	1.0	0.711111

Elde edilen dađınıklık katsayısı deęerlerinden mevcut haber zincirlerinin kalite deęerlendirmesine ilişkin tutarlı bir açıklama getirmek mümkün görünmemektedir. Örneęin, deneysel kontrol olarak düzenlenen tekrarlı doküman içeren Shahaf et. al. haber zinciri orijinaline göre daha iyi bir haber zinciri skoru almaktadır.

Bir dięer metrik olan minedge metrięi için girdi olarak verilen haber zincirinin tutarlılık deęerlendirmesi, Shahaf et al. 2010 [1] çalıřmasında ifade edilen doęrusal program gerçekteřtirilerek yapılmıřtır [5]. Bu doęrusal programa hazırlık olarak haber zincirindeki dokümanlar ön işlemeden geçirilerek uygun veri yapıları oluřturulmuřtur.

İlk olarak doküman kümesi metin analizi teknikleri ile işlenerek doküman ve sözcük modelleri oluřturulmuřtur. Oluřturulan doküman modeli, doküman ID'lerinin anahtar olarak belirlendięi hashmap veri yapısındadır. Deęer alanı ise dokümanda yer alan sözcüklerin frekans, aktivasyon ve bařlangıç deęerlerinin tutulduęu bir treemap'tir. Modelde, doküman içerisinde yer alan sözcükler frekans deęerlerine göre büyükten küçüęe sıralanmıř olarak tutulmaktadır. řekil 2.2'de doküman modelinin bir bölümü örnek olarak verilmektedir:

```
"3_1066201"
{
  "republican"  { :freq 7/312, :act 0,      :init 0N },
  "would"       { :freq 1/48,  :act 0,      :init 0N },
  "perjury"     { :freq 1/48,  :act 0,      :init 0N },
  "house"       { :freq 1/48,  :act 215/49584, :init 215/49584 },
  "clinton"     { :freq 1/48,  :act 23/49584, :init 23/49584 },
  "who"         { :freq 1/52,  :act 0,      :init 0N },
}
```

řekil 2.2: Doküman modeli.

Oluřturulan sözcük modelinde (řekil 2.3) ise sözcüklerin anahtar olduęu bir hashmap içerisinde deęerler hashmap'i yer almaktadır. Sözcük modelinden bir kesit ařaęıda verilmektedir:

```

"republican"
{:docs
 {"1_0991630" {:freq 0,           :act 1/1889,           :init 0},
 "2_1040010" {:freq 1/1889,      :act 12596/1184403,   :init 12596/1184403},
 "3_1066201" {:freq 7/312       :act 0,               :init 0N},
 "4_1070964" {:freq 1/403,      :act 0N,              :init 0N},
 "4_1071380" {:freq 1/403,      :act 0N,              :init 0},
 "4_1071481" {:freq 1/403,      :act 3713/2563886,   :init 3713/2563886},
 "5_1084766" {:freq 25/6362,    :act 0,               :init 0},
 "6_1118570" {:freq 0,         :act 1/1331,         :init 1/1331},
 "8_1250985" {:freq 1/1331,    :act 0,               :init 0}}}

```

Şekil 2.3: Sözcük modeli.

Önerilen doğrusal program; haber zincirinin tutarlılık değerlendirmesini sözcüklerin birbirini takip eden doküman ikililerindeki frekans değişimi (aktivasyon değeri) ve birbirini takip eden iki doküman ikilisindeki başlangıç değeri (ikililerdeki aktivasyon değeri arasındaki fark) üzerinden yapmaktadır. Aşağıda verilen aktivasyon (4.1) ve başlangıç (4.2) değeri hesaplama formüllerinden görüleceği gibi aktivasyon değeri hız, başlangıç değeri ise ivme olarak yorumlanabilir:

$$act = freq(i + 1) - freq(i) \quad (2.4)$$

$$init = act(i + 1) - act(i) = freq(i + 2) - 2freq(i + 1) - freq(i) \quad (2.5)$$

Haber zincirinin tutarlılık değerlendirmesini yapacak doğrusal program bu aktivasyon ve başlangıç değerlerini temel alarak sözcük seçimi yapmaktadır. Bir başka ifade ile doğrusal programın kısıt değişkenleri aktivasyon ve başlangıç değişkenlerine göre belirlenmiştir. İlk kısıt değişkeni $kTotal$ zincirin tamamında sözcüklerin başlangıç değeri toplamını sınırlandırır. Diğer kısıt değişkeni $kTrans$ ise birbirini takip eden doküman ikililerinde sözcüklerin aktivasyon değeri toplamını sınırlandırır. Bu iki kısıt için kullanılan formüller aşağıda verilmektedir:

$$i : document, w : word, \sum_{w,i} init_{w,i} \leq kTotal \quad (2.6)$$

$$i : document, w : word, \forall i \sum_w act_{w,i} \leq kTrans \quad (2.7)$$

Doğrusal programın *minedge* amaç fonksiyonunun hesaplanmasında bir sözcüğün bir doküman-daki aktivasyon değeri o sözcüğün o doküman ile zincirde dokümana ardışık diğer dokümanı bağlama

derecesi (ardışık dokümanların birbirine bağlanmasında spesifik olarak o sözcüğün rolü) ile ağırlıklandırılarak tüm sözcüklere ait aktivasyon toplamı elde edilir (2.8). Hedef bu değeri maksimize etmektir.

$$\forall i \text{ minedge} \leq \sum_w act_{w,i}.influence(d_i, d_{i+1}|w) \quad (2.8)$$

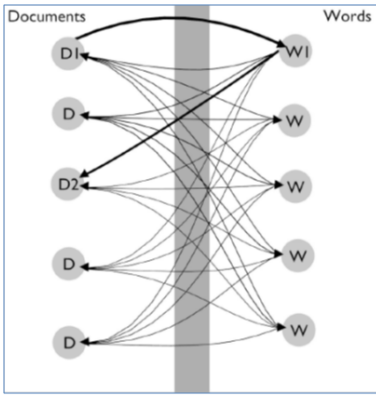
$$\forall w, i \quad act_{w,i}, init_{w,i} \in [0, 1]$$

Diğer bir deyişle global sözcük sayısı belli bir değerin altında tutulduğunda (2.6), her ardışık bağdaki aktif sözcük sayısı sınırlandırılarak (2.7) (kısıtları minimumda sağlayan) her bir sözcüğü bağ kurma potansiyeli en yüksek olan bağ kuracak şekilde uygulayarak haber zinciri üzerinde ölçüm yapılmasıdır.

Ardışık dokümanların spesifik sözcükler üzerinden bağlanma derecesini hesaplama işi için doküman-sözcük iki parçalı (bipartite) yönlü çizgesi oluşturulmuştur. Bu çizgede dokümanlar ve sözcükler düğümler, sözcüklerin normalize edilmiş frekans değerleri ise kenar ağırlıklarıdır. Bir d_1 dokümanının d_2 dokümanına w sözcüğü üzerinden bağlanma derecesi, d_1 'i d_2 'ye w üzerinden bağlayan patika kullanılarak aşağıdaki formül ile hesaplanır:

$$influence(d_i, d_{i+1}|w) = p(w|d_i).p(d_{i+1}|w) \quad (2.9)$$

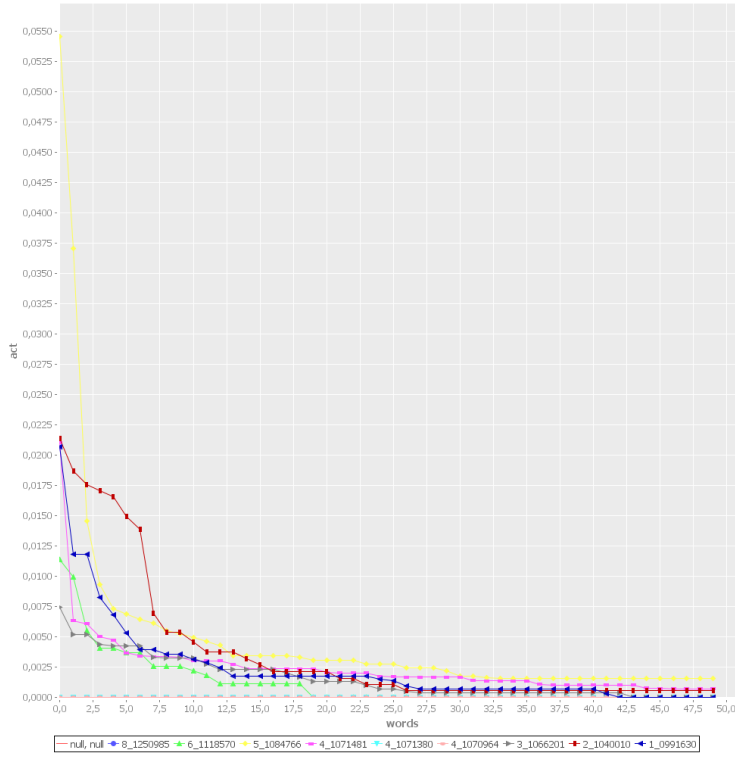
Şekil 2.4'de verilen iki parçalı yönlü çizge üzerinde ilgili patika işaretlenmiştir.



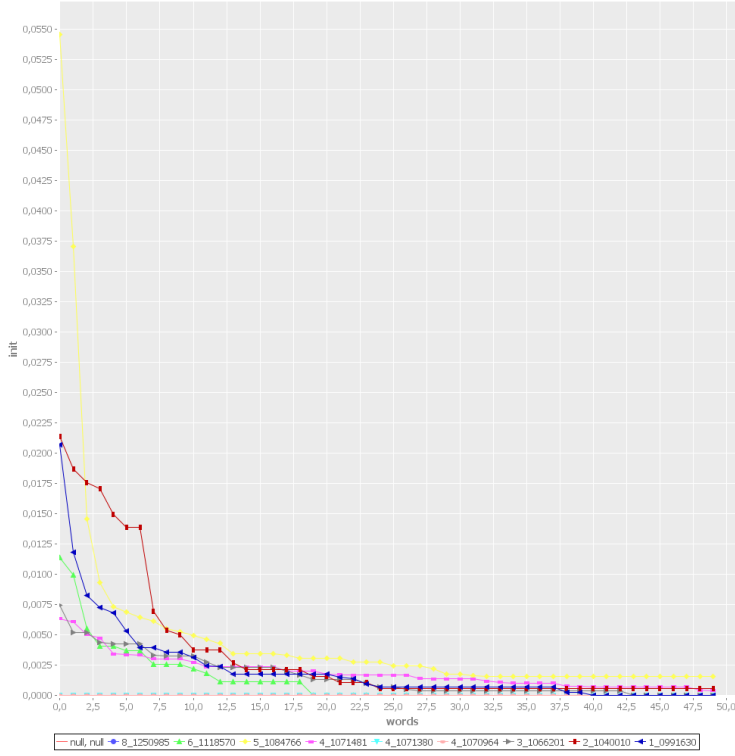
Şekil 2.4: İki parçalı doküman-sözcük yönlü çizgesi.

Kullanılan doğrusal programlama yaklaşımının anahtar noktalarından biri $kTotal$ ve $kTrans$

değerlerinin *minedge*'i maksimize edecek şekilde ayarlanmasıdır. Bu amaçla en uygun *kTotal* ve *kTrans* değerlerinin seçimi için haber zincirinde geçen sözcüklerin her bir dokümandaki aktivasyon ve başlangıç değerleri sıralanarak çizdirilmiştir. Grafiklerden (Şekil 2.5, 2.6) de görüleceği üzere aktivasyon ve başlangıç değer eğrileri birbirine benzerdir. Doküman bazında aktivasyon ve başlangıç değerleri toplamı yaklaşık olarak 0.2'dir. *kTrans* doküman bazında aktivasyon değerleri toplamını sınırlandırmak için kullanıldığından 0.2 ile ilişkilendirilebilir. *kTotal* ise zincir bazında değerlendirildiğinden doküman bazındaki başlangıç değerleri toplamının doküman sayısı ile çarpılması sonucu elde edilen 1.8 değeri ile ilişkilendirilebilir.



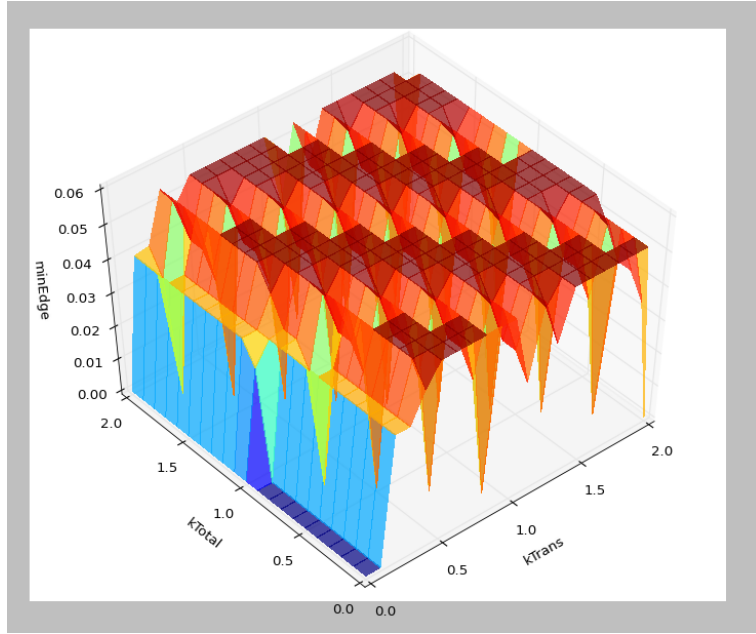
Şekil 2.5: Doküman bazında aktivasyon değeri eğrileri.



Şekil 2.6: Doküman bazında başlangıç değeri eğrileri.

Diğer yandan, $kTrans$ ve $kTotal$ kısıtları birbirinden bağımsız olarak ele alınmayacağından daha isabetli bir değerlendirme $kTotal$ ve $kTrans$ değerlerinin $minedge$ değeri ile birlikte üç boyutlu grafiği (Şekil 2.7) incelenerek yapılabilir. Çalışmada en uygun $kTotal$ ve $kTrans$ değerleri, ilgili grafikte de görüleceği şekilde $minedge$ 'i maksimize eden koyu kırmızı çapraz alanları sağlayan $kTotal$ ve $kTrans$ değerleri olarak belirlenmiştir.

Burada önemli nokta, verilen zincir için elde edilebilecek maksimum $minedge$ değerinin belirlenmesidir çünkü zincirin tutarlılık ölçüsü olarak bu değer kullanılacaktır. Verilen haber zinciri için bu tutarlılık ölçüsü değerinin 0.053624521797934685 olduğu grafikten gözlemlenmektedir.



Şekil 2.7: Kısıt değerlerine karşı *minedge* değerleri.

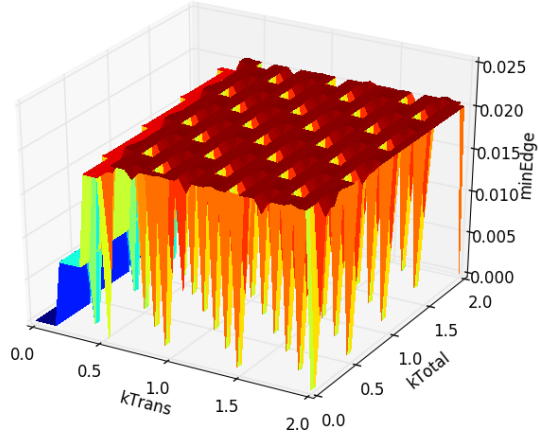
Geliştirilen uygulama parametre olarak sırasıyla $kTrans$, $kTotal$ değerlerini ve haber zincirini oluşturan doküman dizisinin bulunduğu klasörü almaktadır. Örneğin 0.03 0.3 "resources" parametre dizisi için uygulamanın döndürdüğü *minedge* değeri 0.008292161219451767 iken 0.5 1.5 "resources" girdisi için ilgili değer 0.053624521797934685'tir.

Benzer şekilde, $kTrans$ ve $kTotal$ kısıt değerlerine karşı üretilen *minedge* değerlerine ait üç boyutlu grafikler deneysel tasarımda kullanılan her bir haber zinciri için çizdirilmiştir ve sırasıyla aşağıda verilmektedir:

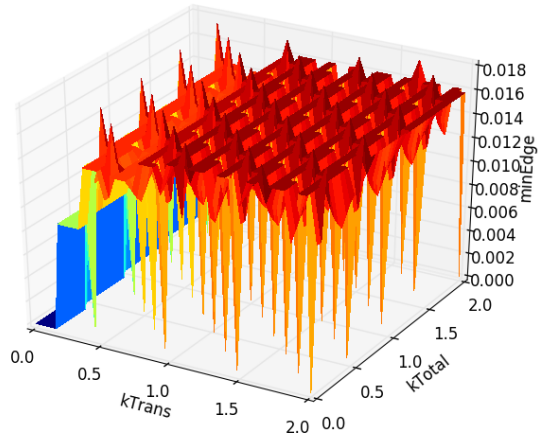
Haber zincirleri için hesaplanan *minedge* değerleri de aşağıdaki tabloda görülmektedir. Değerler incelendiğinde 0.02'den 0.0023'e doğru azalarak giden değerler *minedge* metriğinin güvenilirliği ve haber zinciri değerlendirme metodolojisinin bir parçası olarak kullanılabilirliği üzerine olumlu bir göstergedir.

Tablo 2.2: Metodoloji kapsamında *minedge* metriği değerleri.

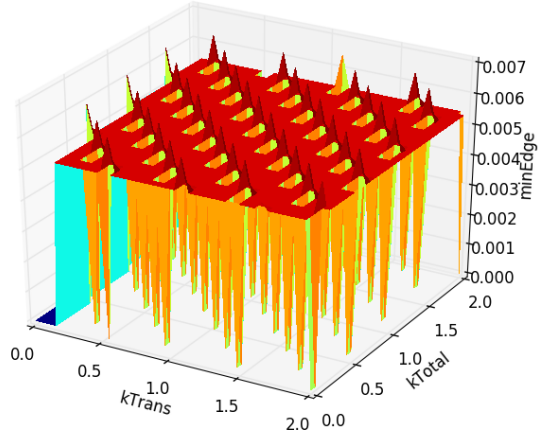
	Zincir büyüklüğü	Minedge değeri
Shahaf et. al.	8	0.02
Shahaf et. al. (tekrarlı doküman)	10	0.018
VAST 2006 Challenge	19	0.0064
Rastgele haber zinciri 1	8	0.0055
Rastgele haber zinciri 2	8	0.0036
Rastgele haber zinciri 3	8	0.0023



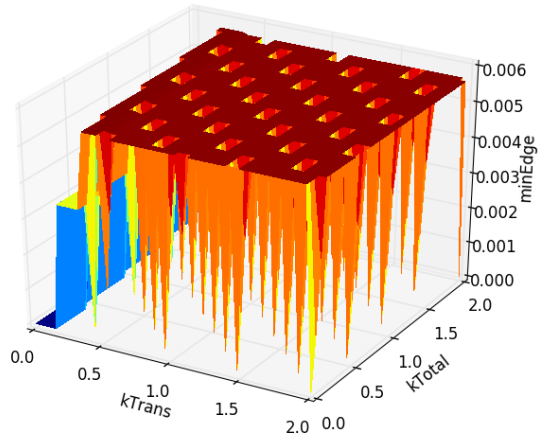
Şekil 2.8: Shahaf et. al. *minedge* değerleri.



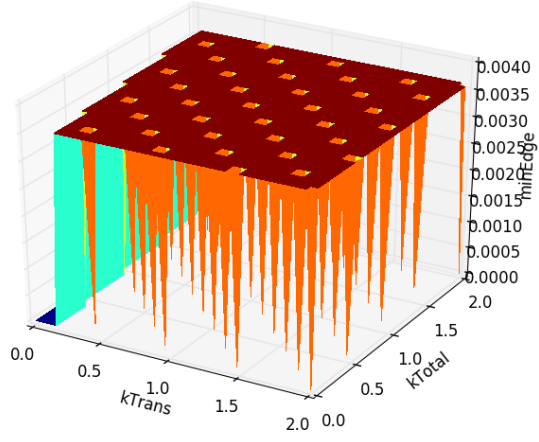
Şekil 2.9: Shahaf et. al. (tekrarlı doküman) *minedge* değerleri.



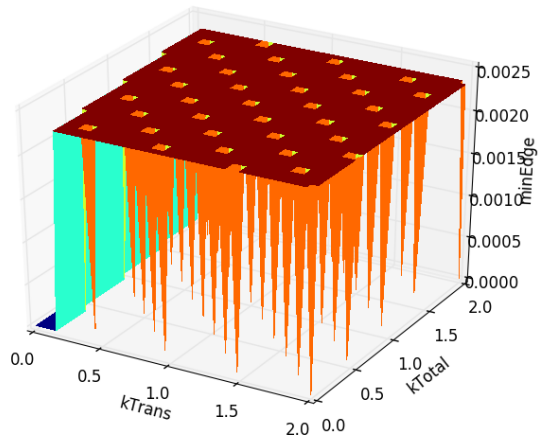
Şekil 2.10: VAST 2006 Challenge $minEdge$ değerleri.



Şekil 2.11: Rastgele haber zinciri 1 $minEdge$ değerleri.



Şekil 2.12: Rastgele haber zinciri 2 *minedge* değerleri.



Şekil 2.13: Rastgele haber zinciri 3 *minedge* değerleri.

2.2 Sonular

Bu b6l6mde elde edilen sonuların bir kısmı altta detayları verilen uluslararası konferansta inceleme ařamasındadır:

- M. Toprak, O. Ozkahraman, & S. Tekir, “A News Chain Evaluation Methodology along with a Lattice-based Approach for News Chain Construction”, Natural Language Processing meets Journalism EMNLP 2017 Workshop, Copenhagen, Denmark, 2017, under review.

3. KAFES (LATTICE) YAPISINA UYGUN HABER ZİNCİRLERİ ÜRETİLMESİ

Bu bölümde haber zinciri üretilmesi için mevcut kafes tabanlı bir yaklaşım ile projemizde önerdiğimiz kafes tabanlı haber zinciri üretme tekniği ortaya konmaktadır. Projenin başlangıcındaki çıkış noktamız kafes yapısının kısmi bir sıralama bağıntısı ile oluşturulması ve kısmi sıralı düzenin tam sıralı düzenin daha genel hali olması nedeniyle tam sıralı bir düzene sahip haber zincirlerinin daha genel bir temsil üzerinden üretilbileceği fikriydi. Aynı sezgi üzerinden yola çıkmasa da literatürdeki yöntem gerçekleştirilip değerlendirilmiştir.

3.1 Mevcut yöntem

Bu bölümde öncelikle Hossain et. al. [6] çalışmasında önerilen haber zinciri üretim yöntemi gerçekleştirilmiştir. Yöntem, dokümanların ters dizininden bir kavram kafesi üretir ve her kapalı sözcük kümesini tek bir kavram ile temsil eder. Her kavram dış özellik (extent) olarak sözcükleri, iç nesne (intent) olarak dokümanları içerir. Bu kavram kafesi yapısının üretilmesi için CHARM-L algoritması [10] kullanılmıştır.

Oluşturulan kafesten potansiyel aday zincirler üretmek için bir başlangıç dokümanı belirlenir. Algoritma bu başlangıç dokümanını iç nesne kümesinde içeren dış özellik kümesi en büyük olan kavramı arayarak devam eder. Bulunan kavram içerisinde yerel komşuluk tabanlı arama ile aday zincirler aranır. Bu sezgisel arama prosedürü iki ölçüt üzerinden tanımlanır. Klik büyüklüğü her aşamada değerlendirilecek maksimum komşu sayısını belirlerken uzaklık eşik değeri belirlenen komşular arasından seçim yapmakta kullanılır.

Algoritmayı kesin referans zinciri bilinen VAST 2006 Challenge veriseti üzerinde çalıştırdık. Başlangıç dokümanı olarak VAST 2006 Challenge kesin referans zincirinin başlangıç dokümanını seçtik. Seçilen dokümana iyi bir takipçi kümesi bulmak üzere klik büyüklüğü olarak 3 ve klik aday sayısı olarak 10'u belirledik. Özetle toplamda 10 3-klik ile çalıştık. Sonuç olarak VAST verisetinden bütün

aday zincirleri oluřturduk.

3.2 Önerilen yöntem

Önerilen yöntemde doküman kümesinde en sık tekrarlanan on sözcük seçilmiş ve kavram kafesi oluřturulmasında minimum destek eřiđi (minimum support threshold) olarak 0.4 belirlenmiřtir. Seçilen tüm sözcükleri kapsayan kavram bařlangıç noktası olarak kabul edilmektedir. Bařlangıç kavramında yer alan her bir doküman için bir haber zinciri oluřturulmuřtur.

Bařlangıç kavramından bir sonraki kavrama geçiřte bir sözcük eksiltilir bir bařka deyiřle seçilen kavram önceki kavramdaki bir sözcük dıřında diđer sözcükleri içermektedir. Hedefimiz bu sözcüğün seçilen kavramın tanım kümesinde yer almamasına rađmen kavramda yine de sık olarak görüldüğünün garanti edilmesidir. Bu sezgisel yaklařım, kavramlar arasındaki geçiřlerin yumuřak (smooth) olmasını sađlar. Haber zincirlerinde de zincirdeki haber dokümanları arasındaki geçiřlerin sözcük kümesi üzerinden yumuřak (smooth) olması aranan bir özelliktir. Bu řekilde aday kavram listesinden eksiltelen sözcüğü en fazla içeren kavram belirlenir ve yeni bařlangıç kavramı olarak kabul edilir. Algoritmanın bir varyantı olarak eksiltelen sözcüğü en fazla içeren yerine en çok sayıda dokümanda içeren kavram bir sonraki kavram olarak belirlenmektedir. Metot özyinelemeli olarak tekrarlanır ve kafesteki son kavrama ya da hiyerarřide altında kavram barındırmayan kavrama kadar devam eder. Haber zincirleri doküman zincirleri olduđundan oluřturulan bu kavram patikasında doküman seviyesine inilmesi gerekmektedir. Yöntemde belirlenen her kavramdan bir doküman haber zincirine dahil edilir (dolayısı ile bu metotla üretilen haber zincirlerinin uzunluđu en fazla kafesin derinliđi kadar olabilir). Bu iřlem iki farklı yöntem ile sınanmıřtır. İlk yöntem, bařlangıç kavramından bir sonraki kavrama geçerken eksiltelen sözcüğün bir sonraki kavramdaki dokümanlar arasında en sık geçtiđi dokümanı almaktadır. İkinci yöntem ise, bir sonraki kavramda var olan dokümanlar arasından bařlangıç dokümanına en yakın olan dokümanı almaktadır.

Algoritmayı kesin referans zinciri bilinen VAST 2006 Challenge veriseti üzerinde çalıřtırdık. Bu verisetinde en sık geçen on sözcüğün hepsini içeren kavram toplamda iki sözcük içermektedir. Dolayısıyla toplamda iki haber zinciri oluřturulmuřtur. Yukarıda da ifade edildiđi üzere algoritmanın varyantı da çalıřtırılmıřtır ve iki haber zinciri de bu varyant üzerinden üretilmiřtir.

Tablo 3.1: *Minedge* metrik deęerleri.

	Zincir uzunluęu	Minedge deęeri
Shahaf et. al.	8	0.02
Shahaf et. al. control	10	0.018
VAST 2006 Challenge	19	0.0064
Rastgele haber zinciri 1	8	0.0055
Rastgele haber zinciri 2	8	0.0036
Rastgele haber zinciri 3	8	0.0023

Tablo 3.2: Daęınlık katsayısı deęerleri.

	Zincir uzunluęu	Eşik: 0.22	Eşik: 0.25
Shahaf et. al.	8	0.836111	0.483333
Shahaf et. al. control 1	10	0.984375	0.615625
Shahaf et. al. control 2	10	0.829167	0.600893
VAST 2006 Challenge	19	0.862132	0.571804
VAST lattice all avg.	7.5	0.873	0.401
VAST lattice proposed method 1	7	0.8	0.543333
VAST lattice proposed method 2	7	0.9	0.543333
Rastgele haber zinciri 1	8	1.0	0.394444
Rastgele haber zinciri 2	8	0.883333	0.316666
Rastgele haber zinciri 3	8	1.0	0.711111

3.3 Deneysel Sonular

Deneysel tasarımıımızda yer alan bütün haber zincirleri için kalite metriklerimizi hesaplattık. Tablo 3.1’de, *minedge* metrik deęerleri verilmektedir. Elde edilen deęerlere bakıldığında kesin referans zincirlerin rastgele üretilenlere göre yüksek deęerlere sahip olduęu görölmektedir. Ayrıca, Shahaf et. al. control 1 orijinal zincirden daha düşük bir deęer almıştır. Dolayısıyla sonular *minedge* metrik deęerlerinin verilen haber zincirlerinin tutarlılıęını ölçmede doęru ve tutarlı davrandıęı savını desteklemektedir.

Deneysel alıřmanın bir dięer parası olarak Soergel uzaklıęı eřik deęerini 0.22 ve 0.25 olarak sabitleyerek bütün zincirler için daęınlık katsayısı deęerlerini hesaplattık (Tablo 3.2).

Elde edilen deęerlerden daęınlık katsayısının haber zincirlerinin kalite deęerlendirmesinde iyi alıřmadıęı görölmektedir. İlk olarak Shahaf et. al. control 1, Shahaf et. al.’dan daha yüksek bir skora sahiptir. Shahaf et. al. control 1 dördüncü sıradaki dokümanın üç kopyasını içermektedir ve daha düşük bir daęınlık katsayısı deęerine sahip olması istenir zira aynı haber dokümanının aynen tekrarlanması zincirin tutarlılıęına katkı yapmayacaktır. Bununla birlikte, daęınlık katsayısı yaklařımı dokümanların tekrar etmesini semantik bir parametre olarak dikkate almaz sadece, ardışık olmayan ve uzaklık eřięi şartını gereęinden fazla yerine getiren ikilileri cezalandırır. Bununla birlikte, verilen bir zincir için ardışık olmayan dokümanların birbirine ne kadar yakınlıkta olduęunu önceden bilmek mümkün deęildir. Bu fikri doęrulamak üzere, bu kez dördüncü sıradaki dokümanın en bařta, ortada ve en sonda tekrar edildięi ikinci bir kontrol zinciri oluřturduk (Shahaf et. al. control 2). Bu

zincir için daha düşük dağımlık katsayısı değeri elde edildi çünkü ardışık olmayan aynı doküman ikilileri 0 uzaklık değerine sahiptir ve dokümanların indis farkı büyüdüğünden cezalandırma da büyüyecektir. Üstelik rastgele zincirler (Rastgele haber zinciri 3) bu metrik için kıyas götürür yüksek değerler elde edebilir.

Haber zinciri oluşturulmasında kafes tabanlı mevcut yaklaşım ve önerdiğimiz yöntemin performans değerlendirmesi için dağımlık katsayısı değerlerini hesaplattık. Tablo 3.2’de VAST lattice all avg. kafes tabanlı mevcut yaklaşım ile üretilen zincirlerin skor ortalamasını, VAST lattice proposed method 1 ve 2 ise önerdiğimiz yöntem kullanılarak elde edilen zincirlerin skorlarını vermektedir. Elde edilen skorların VAST 2006 Challenge kesin referans zincirinin skorundan önemli ölçüde farklılık göstermediği gözlemlenmektedir.

Daha önce de belirtildiği üzere dağımlık katsayısının haber zincirlerinin kalitesini ölçmede kullanımı kabul görmüş değildir. Çünkü metrik hesaplamasında ardışık doküman ikililerini gözardı eder. Ardışık olmayan doküman ikilileri üzerinden cezalandırma yapılması bir zincir tanımı için gerekli ama yeterli bir koşul değildir. Aynı zamanda güçlü ikili ilişki garanti edilmelidir.

Minedge metrik değeri doğru ve tutarlı sonuçlar vermesine rağmen girdi zinciri değiştiğinde parametrelerin ayarlanması gerekmekte ve uygun parametre seçilmediğinde olağandışılıklar (exceptions) gözlemlenmektedir.

Özetle, kafes tabanlı haber zinciri üretme algoritmalarının performans doğrulaması için daha kapsamlı deneysel değerlendirmeye ihtiyaç vardır. Bu değerlendirme olgunlaştırılmış metrikler ve kesin referans zincirleri bilinen verisetlerini içeren bir deneysel ortamda yapılmalıdır.

3.4 Sonuçlar

Bu bölümde elde edilen sonuçların bir kısmı altta detayları verilen uluslararası konferansta inceleme aşamasındadır:

- M. Toprak, O. Ozkahraman, & S. Tekir, “A News Chain Evaluation Methodology along with a Lattice-based Approach for News Chain Construction”, Natural Language Processing meets Journalism EMNLP 2017 Workshop, Copenhagen, Denmark, 2017, under review.

4. HABER GÜVENİLİRLİĞİNİN DEĞERLENDİRİLMESİ

Proje kapsamında haber güvenilirliği ölçülerek değerlendirilmeye çalışılmıştır. Bu değerlendirmede bir haber kaynağını güvenilir yapan faktörler baz alınmıştır. Gazetecilik literatüründe verilen bu faktörlerden "fikirlere mi gerçeklere mi dayanmakta" ve "tarafı mı (is biased)" kapsamıştır. Literatürde güvenilirlik kavramını bu faktörler üzerinden ölçüp değerlendiren başka bir çalışma bulunmamaktadır. Bölüm 4.1'de "fikirlere mi gerçeklere mi dayanmakta" faktörünün ölçümü detaylandırılmıştır. Bölüm 4.2'de ise "tarafı mı (is biased)" faktörünün değerlendirmesi verilmiştir. Son olarak faktörler birlikte ele alınmıştır.

4.1 Fikirlere/Gerçeklere dayanan haber kaynakları

Haberin güvenilirliği haber kaynağının güvenilirliğinden ayrı bir şekilde değerlendirilemez. Haber kaynağının güvenilirliği iki temel boyutta ele alınmaktadır: Haber kaynağına duyulan güven ve kaynağın o konudaki uzmanlığı [11]. Haber kaynağına duyulan güvenin ölçülmesi için haber içeriği üzerinde çeşitli alt faktörler kullanılarak değerlendirme yapılabilir. Bu alt faktörler aşağıda listelenmektedir [11]:

- Adil mi (is fair),
- tarafı mı (is biased),
- hikayenin tamamını aktarıyor mu,
- doğru mu (is accurate),
- gerçeği fikirden ayırıyor mu,
- fikirlere mi gerçeklere mi dayanmakta.

Bu alt başlıkta "Gerçeği fikirden ayırıyor mu" ve "Fikirlerle mi gerçeklere mi dayanmakta" alt faktörleri baz alınmıştır. Spesifik olarak, gerçekleri fikirlerden ayırmak üzere yeni bir gözetimsiz (unsupervised) olasılıksal sözlük-tabanlı fikir madenciliği yaklaşımı geliştirilmiştir. Geliştirilen yaklaşım, sözcüklerin duygu sözlüğü içindeki duygu değerlerinin kümülatif dağılışı üzerinden bir hesaplama yapmaktadır. Sözcüklere ait duygu değerlerinin kümülatif dağılışının kullanılmasında görelilik kuramından esinlenilmiştir. Görelilik bakış açısının bu bağlamda ele alınması iki temel değerlendirme ortaya koyar. İlki, tarafsızlık (objectivity) kendi başına değerlendirilebilecek bir duygu tipi değildir, yerine olumlu (positive) ya da olumsuz (negative) duygu tipi eksikliği olarak ele alınmalıdır. Dolayısıyla tarafsızlık değeri hesabı olumlu ya da olumsuz duygu skorlarına göreli olarak yapılmalıdır. İkincisi, bir sözcüğün olumluluk ve olumsuzluk skoru ortalama bir olumluluk ya da olumsuzluk skoruna sahip bir sözcükten ne kadar daha olumlu ya da olumsuz olduğuna bakılarak elde edilmelidir. Duygu tipi değerlerinin kümülatif dağılış fonksiyonları, sözcüklerin ortalama olumluluk ya da olumsuzluk skorları hakkında bilgi verir.

Haber makalelerindeki her bir sözcük için sözcüğün olumlu ya da olumsuz skorunu o sözcüğe ait skorun eşleştiği kümülatif skor değeri olarak almaktayız. Tablo 4.1'de 0 ve 1 arasındaki skor değerlerinin eşleştiği kümülatif fonksiyon değerleri bir başka ifade ile eşleştikleri gerçek olasılık değerleri listelenmektedir:

Tablo 4.1: Örnek kümülatif dağılış fonksiyonu değerleri.

SentiWordNet Score	Objective prob. (%)	positive prob. (%)	negative prob. (%)
0	0	0	0
0.025	0.01	19.36	11.07
0.05	0.14	82.30	71.09
0.1	0.28	85.53	84.66
0.25	1.26	91.44	89.62
0.75	17.58	99.85	99.67
0.9	24.78	99.98	99.86
0.95	58.02	99.98	99.98
0.975	94.58	99.99	99.99
1	100	100	100

Önerilen yöntem; olumlu, olumsuz ve tarafsız duygu değerlerini ilgili kümülatif dağılış fonksiyonlarını kullanarak olasılık değerlerine dönüştürdüğünden normalizasyon problemi ortadan kalkmaktadır. Olumlu, olumsuz ve tarafsız duygu skorları kendi içinde değerlendirilmektedir. Daha önce de belirtildiği gibi tarafsızlık skoru kendi başına bir anlam taşımaz, yerine olumlu ya da

olumsuz duygu noksanlığı şeklinde yorumlanır. Şekil 4.1’de bir editorial makalesi için cümle bazında ortalama tarafsızlık, olumluluk ve olumsuzluk skorları görülmektedir. Yöntemde ”Fikirlere mi gerçeklere mi dayanmakta” değerlendirmesi olumluluk ve olumsuzluk skorları üzerinden yapılmaktadır.

sentence#	PROBobj	PROBpos	neg
0	0.5949975601116173	0.4258451992679612	0.11640963834838666
1	0.7525949213995927	0.17867204687870258	0.15020986072468154
2	0.5112588414582524	0.3210754748794848	0.5247922644868454
3	0.6877835763320991	0.2476279978927689	0.2354951923673008
4	0.8112437250719077	0.13717709958150384	0.11630007508593872
5	0.7975776919901642	0.15786326589952143	0.2385232975994704
6	0.801029794113388	0.22739703187155164	0.0014225031814382677
7	0.6323490416250851	0.32330170648883383	0.2934749655570052
8	0.5481623276660005	0.40603582579655645	0.35323102840882153
9	0.8238456182406715	0.1607040840868805	0.19412627730197307
10	0.664686714644967	0.17076392175170296	0.29577424879619313
11	0.6605699622479214	0.19375887311253176	0.31317219296400023
12	0.915766640559973	0.08796555802399862	0.007767328742646904
13	0.7346572876374331	0.19006957937488078	0.19574718573628672
14	0.7040599186586698	0.2614821941076791	0.1828483795031544
15	0.5861978872795361	0.31724264494127874	0.30428120978665657
16	0.7246013194271127	0.21996259089895562	0.20539710203335537
17	0.7971997067685221	0.18326850799869865	0.23022186847381335
18	0.9791556126932818	0.0	0.05404316172343508
19	0.7050950895739797	0.1711815262272514	0.3440493517629498
20	0.7419761982423527	0.26479573264089085	0.1251340265810301
21	0.39127465362416947	0.385982386490779	0.4907052004759759
22	0.40807576658602046	0.608856819216122	0.5797586152201131
23	0.6034547027700722	0.35713410052489036	0.25666568216282754
24	0.3942473012219325	0.5637371816802229	0.43900320324343584
25	0.6140484806362966	0.3184721252324393	0.3318824920913935
26	0.3642569768289492	0.5650534099214249	0.4872460816241161
27	0.6413267853705323	0.10932881851295158	0.4316430363999355
28	0.6508056727534094	0.21613414210437204	0.24866170277090283
29	0.6448693732196003	0.31332247103148886	0.26098628977479565
30	0.5879334586387084	0.45721251031188176	0.0
31	0.6851962940422063	0.2585233615649211	0.25323802396386386
32	0.7302647581538634	0.1578130613908898	0.26916525474274633
33	0.5300385753163805	0.421216481527795	0.34408566058131074

Şekil 4.1: Bir editorial makalesi için örnek skorlar.

Çalışma kapsamında gerçekleştirilen deneysel tasarımda üç farklı haber kaynağından gelen haber dokümanları üzerinde değerlendirme yapılmıştır. Bu haber kaynakları editoriallyar, New York Times makaleleri ve Reuters makaleleridir. Doğaları gereği ”Fikirlere mi gerçeklere mi dayanmakta” faktörü açısından farklı nitelik taşırlar. Editoriallyar fikir yazıları, New York Times makaleleri sıradan haber makaleleri, Reuters makaleleri ise içeriği fikirden ziyade gerçeğe dayanan haber ajansı metinleridir. Dolayısıyla hipotezimiz, ”Fikirlere mi gerçeklere mi dayanmakta” alt faktörü açısından fikirlere dayanma derecesine göre editorial-New York Times-Reuters sıralamasının yapılabileceğidir. Elde edilen deneysel sonuçlar istatistiksel açıdan bu savı doğrulamaktadır.

Deneysel tasarımıımızda yer alan derlemlerin ”Fikirlere mi gerçeklere mi dayanmakta” alt faktörü açısından derecelendirmesini yapmak üzere derlemlerdeki her cümle için olumluluk ve olumsuzluk skorlarını hesaplayıp derlem düzeyinde özet istatistiklerini çıkardık. Tablo 4.2’da derlem düzeyinde ortalama, ortanca, standart sapma ve çeyrekler açıklığı (interquartile range) değerleri gösterilmektedir. Bilindiği üzere istatistiksel değerlendirmelerde ortalama-standart sapma ve ortanca-çeyrekler

açıklığı ikilileri birlikte kullanılmaktadır.

Tablo 4.2: Olumluluk ve olumsuzluk skorlarının derlem düzeyinde özet istatistikleri.

	Editorials	NYT	Reuters
Mean of Positive Scores	0.32783	0.28363	0.22636
Median Positive Scores	0.32787	0.25194	0.23227
SD of Positive Scores	0.05323	0.09969	0.10252
IQR of Positive Scores	0.06816	0.12195	0.12154
Mean of Negative Scores	0.26938	0.20059	0.17295
Median of Negative Scores	0.26262	0.20885	0.17648
SD of Negative Scores	0.04395	0.08377	0.09523
IQR of Negative Scores	0.04807	0.10634	0.12238

Tablodan olumluluk değerlerine bakacak olursak deneysel hipotezimizi doğrulayacak şekilde ortanca değer editoriallarda yüzde 32.78 iken New York Times makalelerinde yüzde 25.19'a, Reuters makalelerinde ise yüzde 23.22'ye gerilediğini görmekteyiz. Aynı gözlem olumsuzluk skorları için de geçerlidir. Derlemlerde ortalama bir cümlenin olumsuz olma olasılığı sırasıyla editoriallarda yüzde 26.26, New York Times makalelerinde yüzde 20.88, Reuters makalelerinde ise yüzde 17.64'tür.

Derlem düzeyindeki olumluluk ve olumsuzluk skorları arasındaki farkın istatistiksel önemde olup olmadığını sınamak üzere varyans analizi yöntemi kullanılmıştır. Varyans analizi sonucunda Tablo 4.3'de verilen F istatistiği değerleri elde edilmiştir. Olumsuzluk duygusu için elde edilen F istatistik değeri 2370.54, olumluluk için ise 8765.38'dir. Her iki durumda da elde edilen p değeri 2×10^{-16} 'dır. Çok küçük olan bu p değeri, 0 hipotezini reddetmemizin çok küçük olasılıkla hatalı olabileceğini ifade etmektedir. Bir başka deyişle 0 hipotezini (editoriallar, New York Times makaleleri ve Reuters makaleleri tarafsızlık duygu değeri açısından aynıdır.) güvenle reddedebiliriz.

Tablo 4.3: Varyans analizi sonuçları.

	F	p-value
Positive scores	8765.38	2×10^{-16}
Negative Scores	2370.54	2×10^{-16}

Bunun yanında derlemler arasındaki ikili farkların da istatistiksel önemde olup olmadığını sınamak üzere varyans analizinde grup deneysel hipotezini koruyup ikililer arasındaki karşılaştırmalarda güvenilebilecek Scheffé testi uygulanmıştır. Bu test sonucunda elde edilen p değerleri de yüzde 99 güven aralığında derlemlerdeki ikili farkların önemli olduğunu söylemektedir (Tablo 4.4).

Tablo 4.4: Scheffé testi sonuçları.

Comparison	F	p-value
Negative(Editorials vs. NYT)	7.99	$3.38x10^{-4}$
Negative(Editorials vs. Reuters)	15.65	$1.59x10^{-7}$
Negative(NYT vs. Reuters)	2355.56	0
Positive(Editorials vs. NYT)	2.82	$5.98x10^{-2}$
Positive(Editorials vs. Reuters)	14.85	$3.55x10^{-7}$
Positive(NYT vs. Reuters)	8751.78	0

Haber güvenilirliğini "Fikirlere mi gerçeklere mi dayanmakta" faktörü açısından ölçerek haber kaynaklarını bu nitelik açısından birbirinden ayırabilen yaklaşımımızın yeni bir duygu analizi ve fikir madenciliği yöntemi olarak performansını ve kullanılabilirliğini değerlendirmek için bu alandaki mevcut son yöntemlerle karşılaştırılması gerekmektedir. Bu amaçla, karşılaştırılması gereken son çalışmaları, yakın tarihli bir survey makaleyi [12] baz alarak belirledik. Bu makalede mevcut çok dilli (multilingual) duygu analizi yöntemleri gerçekleştirilip iki popüler derlem (ürün inceleme ve film inceleme) üzerinde performans ölçümü yapılmıştır. Performans sonuçları, SO-PMI-IR [13] gözetimsiz semantik yönelim (unsupervised semantic orientation) algoritmasının doğruluk-hassasiyet (%65) açısından en üst sırada yer aldığını ve bu algoritmayı SVM (Support Vector Machines) algoritmasının takip ettiğini ortaya koymaktadır. SVM algoritması diğer gözetimli algoritmaların açık ara önündedir. Yine aynı çalışmada gözetimli Naive Bayes algoritmasının SVM ile birlikte duygu analizi alanındaki en iyi sınıflandırıcılar olduğu belirtilmektedir. Sonuç olarak SO-PMI-IR, SVM ve Naive Bayes algoritmaları ile karşılaştırmalı sonuçlar elde edilecektir. SO-PMI-IR ile karşılaştırma yapılması iki yönden önem arz etmektedir: İlki SO-PMI-IR algoritmasının en iyi sonuçları vermesidir. İkincisi ise bu algoritmanın önerdiğimiz algoritma gibi gözetimsiz olmasıdır.

İlk denememizi sözkonusu survey makalede karşılaştırmalı analiz için kullanılan ürün inceleme (product review) derlemi [14] üzerinde yaptık. Yöntemimiz %52'lik doğruluk-hassasiyet değerine karşılık gelen aşağıdaki çapraz çizelgeyi üretti:

	<i>POS</i>	<i>NEG</i>
<i>POS</i>	835	37
<i>NEG</i>	803	89

Elde edilen sonuç, yöntemimizin metnin duygu içerip içermediğini doğru bir şekilde bulduğunu

ancak duygunun olumlu mu olumsuz mu olduğunu tespit etmede daha düşük başarı ile çalıştığını göstermektedir.

Bu sonucun temel gerekçesinin bu ürün inceleme derleminde sıklıkla görülen argo ve küfür içeren ifadeler olduğunu düşünmekteyiz. Zira sözlük tabanlı yaklaşımımız sözlükte bulamadığı sözcükleri ihmal ederek cümledeki geri kalan sözcükler üzerinden değerlendirmesini yapmaktadır. Derlemede özellikle bu ifadeler duygu tipini tanımlamaktadır.

Bu ürün inceleme derleminde literatürdeki en iyi gözetimsiz yöntem olan SO-PMI-IR % 62'lik bir performans elde etmektedir.

SO-PMI-IR algoritması, bir metin dokümanının duygu değerini bu dokümandan seçilmiş sözcük öbeklerinin (phrases) semantik yönelim skorlarının toplamı olarak belirler. Sözcük öbeklerinin seçiminde spesifik sözcük türleri (part-of-speech) temel alınmaktadır. Semantik yönelim hesaplamasında PMI (Pointwise Mutual Information) metriği kullanılmaktadır. PMI; iki sözcüğün birlikte görülme olasılıklarının tek tek görülme olasılıkları çarpımına oranının logaritmasıdır (4.1). Bu logaritma değeri, bir sözcüğün varlığının diğerini de gözlemlediğimiz sırada bize verdiği bilgi miktarı olarak yorumlanabilir.

$$PMI(w, w') = \log \frac{p(w, w')}{p(w)p(w')} \quad (4.1)$$

Seçilen sözcük öbeklerinin semantik yönelimi ise seçilen sözcük öbeğinin, en yüksek olumlu semantik yöneline sahip "excellent" referans sözcüğü ile PMI'sının, en yüksek olumsuz semantik yöneline sahip "poor" referans sözcüğü ile PMI'sı arasındaki fark olarak hesaplanır (4.2).

$$SO(phrase) = PMI(phrase, "excellent") - PMI(phrase, "poor") \quad (4.2)$$

PMI formülünde olasılık değerinin bir ölçüsü olarak "phrase * excellent" ve "phrase * poor" sorgusuna bir arama motorundan dönen hit sayısı kullanılmaktadır.

Son olarak, cümle etiketlemesi yolu ile bir değerlendirme de yapabiliriz. Bir önceki raporda varolan deneysel tasarımıımızdaki cümle etiketlemesi çalışmasını ve elde edilen sonuçları aşağıdaki şekilde özetleyebiliriz:

Herhangi bir ön bilginiz bulunmadığında bir cümlemin olumlu, olumsuz ya da tarafsız olma olasılığı yüzde 33'tür. Yöntemimizde kullandığımız şekli ile olumluluk ve olumsuzluk duygu değerleri için cümle bazında yaptığımız hesaplamalara dayanarak skor yüzde 33'ün üzerinde ise cümleyi o duygu değeri ile etiketledik. Bu etiketleme sonuçlarını her bir derlem için ayrı olarak elde ettik. Tablo 4.5'de verilen sonuçlar yorumlanmak istenirse olumlu ya da olumsuz etiketli cümlelerin oranının mevcut derlemlerde sırasıyla (editorial, New York Times, Reuters) yüzde 76.58, yüzde 60.12 ve yüzde 43.90 olduğu görülmektedir.

Tablo 4.5: Cümle duygu etiketlemesi sonuçları.

	Editorials	NYT	Reuters
Sentence Count	632	2345221	8491252
Positive Sentence Count	301	880094	2314192
Negative Sentence Count	183	529823	1413685
Average Number of Sentences	21.067	28.004	10.568
Average Number of Positives	10.033	10.510	2.881
Average Number of Negatives	6.100	6.326	1.759
Ratio of Tagged (either Positive or Negative) Sentences	0.7658	0.6012	0.4390

SO-PMI-IR [13] algoritması ile yapılabilecek bir başka karşılaştırma aynı deneysel tasarımda (editoriallar, New York Times makaleleri ve Reuters makaleleri) SO skorları üzerinden varyans analizi hesabının yapılmasıdır. İkili karşılaştırmalar için de benzer şekilde Scheffé testi uygulanabilir. Deneysel tasarımıımızdaki semantik yönelim skorlarına ilişkin özet istatistikleri Tablo 4.6'de verilmektedir. Varyans analizi ve Scheffé testi sonuçları ise Tablo 4.7'te sunulmaktadır.

Varyans analizi sonuçları, gözetimsiz semantik yönelim yönteminin editoriallar, New York Times makaleleri ve Reuters makalelerini birbirinden ayırmakta istatistiksel önemde başarılı olduğunu ortaya koymaktadır. Scheffé testi sonuçları ise metodun editoriallar ve New York Times makalelerini ayırt etmekte başarılı olmadığını göstermektedir. Bu yönden projede önerdiğimiz relativistic yöntemin daha iyi olduğunu söyleyebiliriz.

Tablo 4.6: Semantik yönelim skorlarının derlem düzeyinde özet istatistikleri

	Editorials	NYT	Reuters
Mean of SO Scores	11.06667	11.98192	3.012828
Median of SO Scores	11	6	1
SD of SO Scores	8.270985	17.17534	5.08865
IQR of SO Scores	9.5	17	4

Tablo 4.7: Semantik yönelim skorlarının varyans analizi ve Scheffé testi sonuçları

		F	p-value
Analysis of Variance	SO scores	60088	2×10^{-16}
Scheffé Tests	SO(Editorials vs. NYT)		0.784695
	SO(Editorials vs. Reuters)		0
	SO(NYT vs. Reuters)		0

4.2 Tarafılı mı (is biased) faktörünün değerlendirilmesi

Literatürde bir metnin tarafılık niteliğini ele alan en son çalışma Recasens et. al. [15] tarafından gerçekleştirilmiştir. Çalışma, kontrollü bir düzenleme ortamı olan Wikipedia'yı veri seti olarak kullanmaktadır zira Wikipedia düzenlenirken tarafılığın kaldırıldığı versiyonlama sürecinden geçmektedir. Bu süreçteki farklı versiyonlar tarafılık niteliği için kesin referans olarak kullanılabilir.

Tarafılık niteliği Wikipedia kapsamında üç ayrı görev için ele alınabilir:

- Bir makalenin tarafılı olup olmadığının tespit edilmesi.
- Tarafılık yaratan sözcük ve sözcük gruplarının bulunması.
- Tarafılık oluşturan sözcüklerin başka sözcüklerle değiştirilmesi.

Bahsedilen çalışma ikinci görevi hedeflerken önerdiğimiz yöntem ilk göreve odaklanmaktadır. Yöntemimizde tarafılığın hesaplamak için öznitelik olarak lemmas, hedges, factive verbs, implicative verbs ve report verbs kullanılmıştır. Bu seçilmiş sözcük ya da sözcük gruplarının makalede var olduğu durumlarda tarafılık skoru artırılmıştır.

Literatürdeki çalışmadan özniteliklerin seçimi konusunda temel farkımız makaledeki kelimelerin birkaç sözcük öncesine ve sonrasına kadar olan bağlama dayalı öznitelikleri dikkate almayışımızdır. Zira onların yöntemindeki gibi tarafılık yaratan sözcük ve sözcük gruplarının bulunması, analiz edilmesi yerine, makalenin tamamında yapılan bir analiz zaten bu bilgiyi ortaya çıkarmaktadır.

İkinci farklı noktamız ise bahsedilen çalışmadaki yöntem tarafılık oluşturan kelimelerin etiketlenmiş olduğu bir veriyi kullanmaktadır. Fakat yöntemimizde kullanılan haber makalelerinin tarafılı olup olmadığına dair bir bilgi bulunmamaktadır. Dolayısıyla, literatürdeki makaleden farklı olarak makine öğrenmesi içermeyen bir yöntem kullanılmıştır. Bunun sonucu olarak da, makine öğrenmesi yöntemlerine uygun öznitelikler kullanılmamıştır (Örneğin POS bilgisi, kelimenin cümledeki konumu vs.).

Sonuç olarak bir makalenin taraflı olup olmadığının analizi, yukarıda bahsedilen özniteliklerin (lemma, hedges, vs.) makaledeki bulunma sayıları ve bulunma sıklıkları hesaplanarak yapılmıştır. Her bir makale için bu sözcüklerin sayıları ve bulunma yüzdeleri bulunmuş ve haber kaynağı bazında ortalamaları alınmıştır. Tablo 4.8 elde edilen bu sonuçları göstermektedir.

Tablo 4.8: Derlem düzeyinde taraflı sözcük sayıları ve yüzdeleri

	Taraflı sözcük sayısı (doküman başına ortalama)	Taraflı sözcük yüzdesi
Editorials	44.966667	8.006409
NYT	34.414216	6.079687
Reuters	13.057370	5.461282

Deneysel tasarımda yer alan haber kaynaklarının taraflılık niteliği açısından diğerlerinden ve ikili olarak birbirlerinden farklı olup olmadığını tespit etmek üzere varyans analizi ve Scheffé testi uygulanmıştır (Tablo 4.9). Elde edilen sonuçlar taraflılık açısından varolan farkların editorial-NYT ikilisi dışında istatistiksel önemde olduğunu ortaya koymaktadır.

Tablo 4.9: Taraflılık Scheffé testi sonuçları

	F	p-value
Editorials vs. NYT	2.18	0.1399
Editorials vs. Reuters	141.64	1.1698×10^{-32}
NYT vs. Reuters	400.80	3.8507×10^{-890}

4.3 Sonuçlar

Haber güvenilirliğinin ””Fikirlere mi gerçeklere mi dayanmakta” ve ”taraflı mı (is biased)” faktörleri açısından ölçümü deneysel tasarımımda yer alan haber kaynaklarının seçilen bu faktörler açısından farklı olduğunu ortaya koymaktadır. Her iki faktör de birbirini doğrulayıcı sonuçlar vermektedir. Güvenilirlik ölçümünde bu faktörlerin birbirini tamamlaması beklenmektedir. Kullandığımız haber kaynaklarında doküman bazında hesaplattığımız semantik yönelim ve taraflılık skorlarının doğrusal bir ilişkiye sahip olup olmadığını test ettik ve 0.6 civarında bir korelasyon değeri elde ettik. Bu noktada gelecek çalışmalarda sözkonusu skorların güvenilirlik çatısı altında birlikte kullanımının daha detaylı irdelenmesi gereği görülmektedir.

Bu bölümde elde edilen sonuçların bir kısmı altta detayları verilen uluslararası konferansa sözlü bildiri olarak kabul edilmiştir:

- E. Sezerer, S. Tekir, “A Relativistic Opinion Mining Approach to Detect Factual or Opionated News Sources”, DaWaK 2017: 19th International Conference on Big Data Analytics and Knowledge Discovery, Lyon, Fransa, Ağustos 2017.

5. SONUÇLAR VE TARTIŞMA

Proje kapsamında haber zincirlerinin otomatik bir şekilde üretilip kalite değerlendirmesinin yapılmasına ve haberin ya da haber kaynağının güvenilirliğinin ölçülüp değerlendirilmesine odaklanılmıştır. Haber zincirleri konusundaki temel problem üretilen zincirlerin doğrulanması problemidir. Zira zincirin kullanışlı olup olmadığı temelde haberi kullanan insanın tanımlayabileceği bir şeydir. Bu şekilde manuel olarak hazırlanmış, doğrulanmış haber zincirleri konusunda önemli eksikler vardır. Kabul görmüş bir haber zinciri kalite değerlendirme metodolojisinin olmaması bu alandaki bilimsel araştırmaları yavaşlatmaktadır zira önerdiğiniz yöntemin işe yararlığını deneysel olarak göstermediğiniz sürece bilimsel katkınız sınırlı kalmaktadır. Projede böyle bir metodoloji ortaya koymak üzere birbirinden farklı ve bağımsız iki metrik aynı çatı altında toplanmıştır fakat geliştirilen metodolojinin olgunlaştırılmasına gereksinim vardır. İlk olarak *minedge* metriği doğru ve tutarlı sonuçlar vermesine rağmen haber zinciri değiştikçe parametrelerin ayarlanması gerekmekte ve gerekli ayar yapılmadığında olağandışlıklar (exceptions) gözlemlenmektedir. Bu olağandışlıklar nedeniyle üretilen zincirlerin *minedge* metriği hesabında aksaklık yaşanmıştır. Dolayısıyla yazılan kodun bu parametrelerin ayarlanması konusunda eniyileştirilmesine ihtiyaç vardır. Bu amaçla ilgili kod olgunlaştırılmaya çalışılacaktır. Öte yandan bir diğer metrik olan dağınıklık katsayısı haber zinciri kalite ölçümünde ilgili bölümde de detaylı bir şekilde açıklandığı üzere pek uygun bulunmamıştır. Haber zinciri değerlendirme metodolojisi tam anlamıyla olgunlaşmamış olsa bile böyle bir metodolojinin ilk defa önerilmesi ve metodoloji kapsamında kullanılan metriklerin detaylı bir şekilde ele alınmış olması önemli katkılardır.

Haberin güvenilirliğinin ölçümü konusunda belirtilen faktörleri dikkate alan ilk çalışma bizim çalışmamızdır. Ayrıca gözetimsiz algoritmaların zor bulunduğu fikir madenciliği konusunda önerilen gözetimsiz olasılıksal sözlük-tabanlı algoritma duygu analizi ve fikir madenciliği alanına bilimsel bir katkı sunmuştur. Elbette yüksek seviyeli bir nitelik olan güvenilirlik mümkünse tüm faktörleri üzerinden ele alınmalı ve faktörlerin entegrasyonu problemi de dikkate alınmalıdır.

5.1 Proje kapsamındaki yayınlarımız

Projede elde edilen sonuçların bir kısmı altta detayları verilen uluslararası konferansa sözlü bildiri olarak kabul edilmiştir:

- E. Sezerer, S. Tekir, “A Relativistic Opinion Mining Approach to Detect Factual or Opinionated News Sources”, DaWaK 2017: 19th International Conference on Big Data Analytics and Knowledge Discovery, Lyon, Fransa, Ağustos 2017.

Projede elde edilen sonuçların bir kısmı altta detayları verilen uluslararası konferansta inceleme aşamasındadır:

- M. Toprak, O. Ozkahraman, & S. Tekir, “A News Chain Evaluation Methodology along with a Lattice-based Approach for News Chain Construction”, Natural Language Processing meets Journalism EMNLP 2017 Workshop, Copenhagen, Denmark, 2017, under review.

5.2 Konferans katılım faaliyetleri

Projenin yürürlükte olduğu sürede proje ekibinin katıldığı konferans faaliyetleri altta verilmiştir.

Proje ekibinden Mustafa Toprak çalışmasının ilk sonuçlarını sunmak üzere Eylül 2016’da Ankara’da Yapay Öğrenme ve Bilgi İşlemede Yeni Teknikler Yaz Okulu’na katılmış ve Haber Zinciri Oluşturulması İçin Kafes Tabanlı Bir Yaklaşım isimli poster bildirisini sunmuştur.

Haber güvenilirliğine dair ilk sonuçlar ACL 2016 konferansına gönderilmiş ancak oldukça rekabetçi olan bu konferanstan kabul alınmamıştır. Bunun üzerine alandaki son gelişmeleri takip etmek üzere proje yürütücüsü Yrd. Doç. Dr. Selma Tekir ve proje ekibinden Erhan Sezerer Almanya’nın Berlin şehrinde düzenlenen The 54th Annual Meeting of the Association for Computational Linguistics (ACL 2016)’e katılmışlardır.

Önceden de belirtildiği üzere “A Relativistic Opinion Mining Approach to Detect Factual or Opinionated News Sources” adlı bildirimiz DaWaK 2017: 19th International Conference on Big Data Analytics and Knowledge Discovery’de sözlü bildiri olarak kabul edilmiştir ve bildiriye sunmak üzere sözkonusu konferansa katılım sağlanacaktır.

Belirtmek isteriz ki üstte belirtilen proje ekibinin bildirilerinde ve sunumlarında TÜBİTAK’a verdiği destekten ötürü teşekkür edilmiştir.

5.3 Proje konusundaki lisansüstü çalışmalar

Proje konusunda akademik kategoride "Linking dots together" başlıklı bir lisans bitirme tezi [5] tamamlanmıştır. Tez kapsamında verilen bir haber zincirinin tutarlılık ölçümü için önerilen *minedge* metriği gerçekleştirilmiştir.

Proje kapsamında bursiyer olarak desteklenen Erhan Sezerer "News Story Analysis with Credibility Assessment by Opinion Mining" başlıklı tezini başarı ile tamamlamış ve proje iş-zaman planında belirtildiği üzere haberin güvenilirliğinin ölçülüp değerlendirilmesi işinde görev almıştır.

Yine proje kapsamında bursiyer olarak desteklenen yüksek lisans öğrencisi Mustafa Toprak'ın tez başlığı "A Lattice-based Approach for News Chain Construction" olup projede kafes tabanlı haber zinciri üretilmesi konusunda katkı vermiştir.

Her iki öğrenci de eğitimlerine doktora ile devam etmektedir.

Kaynakça

- [1] Shahaf, D. and Guestrin, C., “Connecting the dots between news articles,” in Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washington, DC, USA, July 25-28, 2010, 2010, pp. 623–632. [Online]. Available: <http://doi.acm.org/10.1145/1835804.1835884>
- [2] Kang, Y., Görg, C., and Stasko, J. T., “Evaluating visual analytics systems for investigative analysis: Deriving design principles from a case study,” in Proceedings of the IEEE Symposium on Visual Analytics Science and Technology, IEEE VAST 2009, Atlantic City, New Jersey, USA, 11-16 October 2009, part of VisWeek 2009, 2009, pp. 139–146. [Online]. Available: <http://dx.doi.org/10.1109/VAST.2009.5333878>
- [3] Sandhaus, E., “The New York Times Annotated Corpus,” Linguistic Data Consortium, Philadelphia, vol. 6, no. 12, 2008.
- [4] Whiting, M., North, C., Endert, A., Scholtz, J., Haack, J., Varley, C., and Thomas, J., “Vast contest dataset use in education,” in Visual Analytics Science and Technology, 2009. IEEE VAST 2009., 2009, pp. 115 –122.
- [5] Ozkahraman, O., “Linking dots together,” Bachelor Thesis, Izmir Institute of Technology, June 2015.
- [6] Hossain, M. S., Andrews, C., Ramakrishnan, N., and North, C., “Helping intelligence analysts make connections.” Scalable Integration of Analytics and Visualization, vol. 11, p. 17, 2011.
- [7] Hossain, M. S., Butler, P., Boedihardjo, A. P., and Ramakrishnan, N., “Storytelling in entity networks to support intelligence analysts,” in Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ser. KDD

- '12. New York, NY, USA: ACM, 2012, pp. 1375–1383. [Online]. Available: <http://doi.acm.org/10.1145/2339530.2339742>
- [8] Hossain, M. S., Gresock, J., Edmonds, Y., Helm, R., Potts, M., and Ramakrishnan, N., “Connecting the dots between PubMed abstracts,” *PLoS ONE*, vol. 7, no. 1, p. e29509, 2012.
- [9] Swanson, D. R., “Complementary structures in disjoint science literatures,” in *Proceedings of the 14th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR '91. New York, NY, USA: ACM, 1991, pp. 280–289. [Online]. Available: <http://doi.acm.org/10.1145/122860.122889>
- [10] Zaki, M. J. and Hsiao, C.-J., “Efficient algorithms for mining closed itemsets and their lattice structure,” *IEEE Trans. on Knowl. and Data Eng.*, vol. 17, no. 4, pp. 462–478, Apr. 2005. [Online]. Available: <http://dx.doi.org/10.1109/TKDE.2005.60>
- [11] Gaziano C, M. K., “Measuring the concept of credibility,” *Journalism Quarterly*, vol. 63, pp. 451–462, 1986.
- [12] Dashtipour, K., Poria, S., Hussain, A., Cambria, E., Hawalah, A. Y. A., Gelbukh, A., and Zhou, Q., “Multilingual sentiment analysis: State of the art and independent comparison of techniques,” *Cognitive Computation*, vol. 8, no. 4, pp. 757–771, 2016. [Online]. Available: <http://dx.doi.org/10.1007/s12559-016-9415-7>
- [13] Singh, V. K., Marisha, P. W., Piryani, R., and Uddin, A., “Sentiment analysis of textual reviews: Evaluating machine learning, unsupervised and sentiwordnet approaches,” 2013, Conference Paper.
- [14] Blitzer, J., Dredze, M., and Pereira, F., “Biographies, bollywood, boomboxes and blenders: Domain adaptation for sentiment classification,” in *In ACL*, 2007, pp. 187–205.
- [15] Recasens, M., Danescu-Niculescu-Mizil, C., and Jurafsky, D., “Linguistic models for analyzing and detecting biased language,” in *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, ACL 2013, 4-9 August 2013, Sofia, Bulgaria, Volume 1: Long Papers*, 2013, pp. 1650–1659. [Online]. Available: <http://aclweb.org/anthology/P/P13/P13-1162.pdf>

TÜBİTAK
PROJE ÖZET BİLGİ FORMU

Proje Yürütücüsü:	Yrd. Doç. Dr. SELMA TEKİR
Proje No:	114E784
Proje Başlığı:	Haber Zincirlerinde Tutarlılık Ve Güvenilirlik Değerlendirmesi
Proje Türü:	3501 - Kariyer
Proje Süresi:	24
Araştırmacılar:	
Danışmanlar:	
Projenin Yürütüldüğü Kuruluş ve Adresi:	İZMİR YÜKSEK TEKNOLOJİ ENS.
Projenin Başlangıç ve Bitiş Tarihleri:	01/05/2015 - 01/05/2017
Onaylanan Bütçe:	92970.0
Harcanan Bütçe:	66035.88

TÜBİTAK

<p>Öz:</p>	<p>Çok hızlı ve büyük miktarda haber akışının olduğu günümüzde haber analizi büyük bir gereksinimdir. Haberi takip etmek, doğruluğunu denetlemek, yorumlamak özellikle kurumsal bazda çok önemlidir. Bunun yapılabilmesi bir bilgi işleyiş döngüsünün çalıştırılması ile mümkündür. Bilginin toplanması, hedefler doğrultusunda işlenip analiz edilerek işe yarar bilgiye dönüştürülmesi beklenmektedir.</p> <p>Projenin amacı haber güvenilirliğinin ölçülüp değerlendirilmesine yönelik bir yaklaşım geliştirmektir. Haber güvenilirliği haberalma faaliyetinin olmazsa olmazları arasındadır. Kurumsallaşmış medya kuruluşları (BBC, The New York Times vb.) hâlihazırda çok büyük miktarda yapısal veri sunmaktadır. Haberi doğrulama, kaynak geçerliliğini denetleme gereksinimi had safhadadır.</p> <p>Projede bir haber zinciri üzerinde güvenilirlik ölçümü ve değerlendirmesi yapılacaktır. Projede haber zinciri, noktaları birleştirilim (connecting the dots) yaklaşımı ile karşılanmaktadır. Noktaları birleştirilim yaklaşımı, haber zincirini biri başlangıç diğeri bitiş noktası olarak tanımlanabilecek iki haber belgesini tutarlı bir şekilde birleştiren haber belgeleri dizisi olarak tanımlanmaktadır. Güvenilirlik değerlendirmesi ele alınacak haber zincirinin tutarlılık değerlendirmesi ile birlikte gerçekleştirilecektir. Haber güvenilirliği; haberin doğruluğu, fikirlerden ziyade somut gerçeklere dayanmasıdır. Haber kaynağının güvenilirliği ise iki temel boyutta ele alınmaktadır: Haber kaynağına duyulan güven ve kaynağın o konudaki uzmanlığı [21]. Haber güvenilirliğinin ölçümünde ?Gerçeği fikirden ayırıyor mu? ve ?Fikirler mi gerçeklere mi dayanmakta? faktörleri baz alınacaktır. Fikir madenciliği kullanılarak gerçekler fikirlerden ayırılmaya çalışılacaktır. Haber zincirini oluşturan belgelerin gerçeği fikirden ayırıp ayırmadıkları, belge yapısındaki gerçek/fikir tümcelerinin organizasyonu irdelenecektir. Belgelerdeki gerçek/fikir yapılanmasına ek olarak gerçek/fikir oranı da tespit edilip haber zincirini oluşturan dokümanların bu açıdan bütünsellik taşıyıp taşımadığı sıvanacaktır. Ayrıca ardışık dokümanlar arasındaki geçişin gerçek/fikir bilgisine dayalı bir değerlendirme mekanizması kurulacaktır. Güvenilirlik kavramının ölçümüne ilişkin benzer bir yaklaşım mevcut olmadığından proje oldukça yenilikçidir.</p> <p>Haber zincirinin güvenilirlik değerlendirmesi tutarlılık değerlendirmesinden bağımsız değildir. Haber zinciri bağlamında tutarlılığı karşılamak üzere yöntemler mevcuttur ancak yeni arayışlara gereksinim vardır. Proje kapsamında tutarlı haber zinciri elde etmeyi sağlayacak yeni bir yöntem geliştirilecektir. Geliştirilecek yöntem tutarlı haber zincirinin kafes (lattice) yapısı ile iyi temsil edileceği sezgisine dayanmaktadır. Kafes yapısının düğümleri, haber belgelerinde geçen sözcükler ve bu sözcüklerin geçtiği haber belgeleri ikilileri ile temsil edilecektir. Tutarlı haber zincirlerine ait kafeslerin tam kafes niteliğine sahip olması beklenmektedir. Zaki ve Ramakrishnan [17]?ın kapalı tanım kümesi kafesi (closed description set lattice) oluşturma algoritması bu amaçla kullanılarak iyi haber zincirleri elde edilip edilmeyeceği sıvanacaktır. Önerilen yöntem güvenilirlik değerlendirmesi yapılacak haber zincirlerini üretmesi açısından önemli ve aynı ölçüde özgündür.</p> <p>Proje tamamlandığında elde edilecek çıktıların bilgi keşfi ve veri madenciliği alanında bilimsel katkı sunması beklenmektedir. Önerilen tekniklerin olgunlaşması bunların yeni teknolojilerde kullanılmasını mümkün hale getirecektir. Ayrıca haberalma toplumun sosyo-ekonomik yapısında etkili olan bir işlemdir. Özellikle sosyal medya bu alanı yeniden şekillendirmektedir. Bu alanda, alınan haberlerin doğru şekilde işlenmesi ve haber güvenilirliği konusunda farkındalığın artırılması büyük önem arz etmektedir.</p>
<p>Anahtar Kelimeler:</p>	<p>Bilgi keşfi ve veri madenciliği, haber zinciri, güvenilirlik, tutarlılık, fikir madenciliği, kafes.</p>
<p>Fikri Ürün Bildirim Formu Sunuldu Mu?:</p>	<p>Hayır</p>