Türkçe Haber Benzerliklerinin Belirlenmesinde Varlık İsimlerinin Hikaye Bağlantı Algılama Görevinin Başarımına Etkisi

The Impact of Named Entities on the Performance of Story Link Detection Task Using a Turkish Corpus of News Items

Hamid AHMADLOUEI

Prof. Dr. Hayri SEVER

Tez Danışmanı

Hacettepe Üniversitesi

Lisansüstü Eğitim-Öğretim ve Sınav Yönetmeliğinin

Bilgisayar Mühendisliği Anabilim Dalı için Öngördüğü

YÜKSEK LİSANS TEZİ olarak hazırlanmıştır.

2014

Hamid AHMADLOUEI'nın hazırladığı "**Türkçe Haber Benzerliklerinin Belirlenmesinde Varlık İsimlerinin Hikaye Bağlantı Algılama Görevinin Başarımına Etkisi**" adlı bu çalışma aşağıdaki jüri tarafından **Bilgisayar Mühendisliği Anabilim Dalı**'nda **YÜKSEK LİSANS TEZİ** olarak kabul edilmiştir.

Prof. Dr.

Başkan İlyas ÇİÇEKLİ

Prof. Dr.

Danışman Hayri SEVER

Doç. Dr.

Üye Ebru AKÇAPINAR SEZER

Doç. Dr.

Üye Süleyman TOSUN

Yrd. Doç. Dr.

Üye Erhan MENGÜŞOĞLU

Bu tez Hacettepe Üniversitesi Fen Bilimleri Enstitüsü tarafından **YÜKSEK LİSANS TEZİ** olarak onaylanmıştır.

Prof. Dr. Fatma SEVİN DÜZ

Fen Bilimleri Enstitüsü Müdürü

ETİK

Hacettepe Üniversitesi Fen Bilimleri Enstitüsü, tez yazım kurallarına uygun olarak hazırladığım bu tez çalışmasında,

- Tez içindeki bütün bilgi ve belgeleri akademik kurallar çerçevesinde elde ettiğimi,
- Görsel, işitsel ve yazılı tüm bilgi ve sonuçları bilimsel ahlak kurallarına uygun olarak sunduğumu,
- Başkalarının eserlerinden yararlanılması durumunda ilgili eserlere bilimsel normlara uygun olarak atıfta bulunduğumu,
- Atıfta bulunduğum eserlerin tümünü kaynak olarak gösterdiğimi,
- Kullanılan verilerde herhangi bir tahrifat yapmadığımı,
- Ve bu tezin herhangi bir bölümünü bu üniversitede veya başka bir üniversitede başka bir tez çalışması olarak sunmadığımı

Beyan ederim.

29/12/2014

HAMID AHMADLOUEI

ÖZET

Türkçe Haber Benzerliklerinin Belirlenmesinde Varlık İsimlerinin Hikaye Bağlantı Algılama Görevinin Başarımına Etkisi

Hamid AHMADLOUEI

Yüksek Lisans, Bilgisayar Mühendisliği Bölümü

Tez Danışmanı: Prof. Dr. Hayri SEVER

Aralık 2014, 72 sayfa

Tez, Konu Tespit ve Takip (Topic Detection and Tracking - TDT) programında tanımlı Hikaye Bağlantı Algılama (Story Link Detection - SLD) görevinin Türkçe bir derlem üzerinde farklı benzerlik fonksiyonları ve bunların kombinasyonlarını Varlık İsimler üzerinde kullanılarak başarımının test edilmesini ve anma/duyarlık uygun değerlerini sağlayacak kombinasyonun bulunmasını amaçlamaktadır. Bu kapsamda, TDT içerisinde başarısı kanıtlanmış olan Vektör Uzayı Modeli (Vector Space Model) temel yöntem olarak kabul edildi ve bu yöntemle birlikte Varlık İsimlerinin (Named Entity) kullanılmasının başarım üzerindeki etkileri değerlendirildi. Tezde tanımlanan yöntemlerin test edilebilmesi için BilCOL-2005 derlemi haberlerde kim, nerede, ne zaman vs. gibi sorularına yanıt verecek varlık isimlerin sekiz farklı (kişi, konum, zaman, kurum, para, yüzde ve belirsiz) etiketlerle işaretlenerek kullanıldı.

Anahtar Kelimeler: Konu Tespit ve Takip, Vektör Uzay Modeli, Varlık İsimler, Haber Benzerlikleri Tespiti

ABSTRACT

The Impact of Named Entites on the Performance of Story Link Detection Task Using a Turkish Corpus of News Items

Hamid AHMADLOUEI

Master of Science, Department of Computer Engineering

Supervisor: Prof. Dr. Hayri SEVER

December 2014, 72 pages

This thesis aims to test the performance of the Story Link Detection (SLD) task as part of the Topic Detection and Tracking (TDT) program using different similarity functions and test their combinations performance on named entities, and find the one that provides the optimum precision/recall values. To do this, we used the Vector Space Model (VSM) as the main method which their performance is proven in TDT studies, and evaluate the impact of Named Entities on the VSM performance. In order to test the performance of methods, we used the BilCOL-2005 corpus after tagged named entites which were used to respond to who, where, when and etc questions with eight (who, where, when, organization, Money, percentage, date, unknown) different labels.

Keywords: Topic Detection and Tracking, Vector Space Model, Named Entities, Story Link Detection.

TEŞEKKÜR

Yüksek lisans eğitimim sürecinde, mühendislik ve akademik mesleğine ve hem de hayata yaklaşımıyla, nasıl yaşamamız, çalışmamız ve elde etmek istediklerimizi nasıl kazanabileceğimiz, bizim hayatımızı, geleceğimizi, kişiliğimizi ve sorumluluklarımızı tam anlamıyla kavramamızı ve doğru kararları tartışmamızı konusunda bizlere bir örnek olan, eğitimimde emeği çok geçmiş, motivasyon ve eğitimimin devamı için, gerekli olan çalışma ortamının devamlılığını sağlayan ve yardımlarını hiç bir zaman esirgemeyen, tez danışmanım sayın Prof. Dr. Hayri SEVER'e en içten teşekkürlerimi sunarım.

İÇİNDEKİLER

ŞEKİLLER

TABLOLAR

SİMGELER VE KISALTMALAR

Simgeler

f	f-ölçüsü (f-measure)
d	duyarlık (precision)
a	anma (recall)
m	mikro
e	eşik değer (threshold)
k-m	k-means
c-m	c-means

Kısaltmalar

TDT	Topic Detection Tracking (Konu Tespit ve Takip)
IR	Information Retrieval (Bilgi Erişimi)
TF	Term Frequency
IDF	Inverse Document Frequency
TF-IDF	Inverse Document Frequency
ML	Machine Learning (Makine Öğrenimi)
SLD	Story Link Detection (Hikaye Bağlantı Algılama)
RM	Relevance Model (İlgi Modeli)
EM	Event Model (Olay Model)
NE	Named Entity (Varlık İsimler)
VSM	Vector Space Model

VUM	Vektör Uzay Model
NLP	Natural Language Processing (Doğal Dil İşleme)
KYK	K-en Yakın Komşu
K-NN	K-Nearest Neighbor
VİVM	Varlık İsimler Vektör Modeli
VİKM	Varlık İsimler Kesişim Modeli
VİBKM	Varlık İsimler Birleşim Kesişim Modeli
VİNM	Varlık İsimler Normalizasyon Modeli
TP	True Positive
FP	False Positive
FN	False Negative
TN	True Negative

1 Giriş

1.1 Problem Tanımı

Günümüzde yayınlanan online haber öğelelrin sayısı üstel büyüme sebebyle neredeyse takip edilemeyecek kadar büyümektedir, ayrıca bu sayısının üstel büyümesi ve çeşitli kanalların kullanımı, online çağında bilgi erişim sistemlerin bilgi aramalarını zorlaştırır, dolayısıyla bü büyük haber artışını ve büyük grup haberleri yönetmek artık vazgeçilmez bir ihtiyaca dönüşülmüştür. Özellikle, internet dunyasında verinin inanılmaz hızla büyümesi, kullanıcıların tam istedikleri bilgiye ulaşmaları artık zaman alıcı soruna dönüşmektedir. Sürekli olarak artan bu bilgi havuzunda erişimi kolaylaştırmak adına ve bu bilgilere kolaylıkla doğru şekilde ve kısa zamanda ulaşılabileceğini sağlamak için ayrıca her türlü blginin kolaylıkla doğru şekilde sanal ortama ekelenebilmesi ve belgelerin içeriklerine göre önceden belirlenmiş olan sınıflara atanması için Bilgi Erişim Sistemleri ortaya çıkmış. Bu alan üzerindeki akademik çalışmalar son yıllarda ağırlıklı olarak Konu Tespit ve Takip (Topic Detection and Tracking - TDT) programı üzerinde yoğunlaşmıştır [1].

1.2 Çalışma Konusu ve Kapsamı

Tez genelde Bilgi Erişim Sistemleri (Information Retrieval Systems) kapsamına giren çalışmaları içermekle birlikte özelde Konu Tespit ve Takib alanında araştırmalar içermektedir. Tez çalışmaları, TDT programı içerisinde tanımlanmış olan ve iki farklı haberin birbirleri ile ne kadar benzer olduklarını belirlemeyi hedefleyen Hikaye Bağlantı Algılama (Story Link Detection) görevinin gerçekleştirilmesinde kullanılacak olan yeni yaklaşımları kapsamaktadır. TDT içerisinde varlık isimlerinin haber benzerliklerinde kullanıldığı çalışmalar literatürde olay modeli (event model) olarak isimlendirilmektedir [1]. Tezin özel olarak TDT içerisindeki hikaye bağlantı algılama görevindeki başarımını artırmak için olay modelini kapsayacağını söylemek yanlış olmayacaktır. Buna göre literatürde bu görevin gerçekleştirilmesinde başarıları kanıtlanmış olan vektör uzayı modeli yöntemi temel alarak varlık isimleri yaklaşımı ile (ya da olay modeli) sistem başarımının artırılması üzerinde deneyler gerçekleştirildi. Bununla birlikte öngörülen deneylerin gerçekleştirilebilmesi için ihtiyaç duyulan etiketlenmiş bir Türkçe derlem bulunmadığı için BilCol-2005 derleminin “kim“, “nerede“, “ne zaman“ vs. gibi etiketlerle işaretlenmesi tezin kapsamı içerisinde yer almaktadır.

1.3 Amaç

TDT alanında gerçekleştirilen akademik çalışmalar, özellikle hikaye bağlantı algılama görevinin gerçekleştirilmesinde kullanılan yöntemler içerisinde özellikle erişim fonksiyonu bacağında farklı yöntemler birlikte kullanılarak erişim başarımının artırılıp artılamayacağı bilim insanları tarafından merak edilerek araştırılmıştır [2][3]. Farklı yöntemlerin birleştirilmesi konusunda yapılan çalışmalar genellikle sistemin anma (recall) değerlerini artırırken aynı zamanda ilgisiz pek çok belgenin de getirilmesini sağlamakta ve sistemin duyarlık (precision) değerinin dolayısıyla başarımın düşmesine neden olmaktadır. Bu nedenle, bu tür farklı erişim fonksiyonlarının birlikte kullanılacağı çalışmalarda sistem başarımını en üst seviyeye çıkarabilmek için anma ve duyarlık arasındaki dengeyi gözetecek modellerin geliştirilmesi son derece önemlidir. Kısaca bu tür sistemlerin, ideal olarak, derlemdeki tüm ilgili belgelere erişim sağlamasını aynı zamanda da ilgisizlerin dışarıda bırakılmasını sağlayacak şekilde uygun stratejileri desteklemesi gerekmektedir.

Bu çerçevede önerilen tezin amacı; Hikaye Bağlantı Algılama (Story Link Detection) görevinin Türkçe bir derlem üzerinde farklı erişim fonksiyonları ve bunların kombinasyonları kullanılarak başarımının test edilmesini ve optimum anma/duyarlık değerlerini sağlayacak kombinasyonun bulunmasını sağlamaktır.

1.4 Motivasyon ve Özgün Değer

Önerilen tez içerisinde, SLD içerisinde haber benzerliklerinin ve farklılıklarının belirlenmesinde varlık isimlerinin kullanılacak olması, Türkçe derlemler üzerinde bu kapsamdaki çalışmaların çok sınırlı olması nedeni ile tezin özgün içeriğini oldukça kuvvetlendirmekte. Ele alınan problemin önemi ve güncelliği, uygulanması planlanan yöntemlerin özgünlüğü, ayrıca Türkçe üzerinde şimdiye kadar bu tür çalışma yapılmadığını göz önünde bulundurarak, tez sonuçlarının değerini daha arttırır. Çalışma ile ilgili bir karşılaştırmalı değerlendirme düzeneği sağlanacağından, araştırma sonuçları bundan sonra yapılacak benzer çalışmalar için önem taşımaktadır. Diğer taraftan, tez kapsamında vektör uzayı ve varlık ismi karşılaştırma yöntemlerinin bağımsız başarımlarının belirlenmesi, elde edilen bağımsız başarımların OR gibi mantıksal işleçlerle birleştirilmesi, elde edilen başarım ile karşılaştırılması, vektör uzayı varlık ismi karşılaştırmasında ilgisiz bulunan sonuçların çıkarılması ile elde edilecek başarımın değerlendirilmesi gibi çalışmalar, gerçekleştirildi. Bu tez önerisinin konusunu oluşturan TDT programında Hikaye Bağlantı Algılama (Story Link Detection - SLD) görevinin gerçekleştirilmesinde farklı doküman gösterim yöntemlerinin ve elde edilen sonuçların farklı kombinasyonlarının test edilmesi konusu literatürde çalışılan bir konu olmasına rağmen seçilen ve özellikle Türkçe bir derlem üzerinde uygulanacak yöntemler açısından özgün değer taşımaktadır. Tezin bu yönü literatürde daha önce bu tür bir araştırma hiç yapılmamış olmasından dolayı oldukça yenilikçidir. Bu kapsama yakın bir çalışma [5] TDT derlemi üzerinde gerçekleştirilmiş ve sadece haberlerdeki asıl aktörlere (kim) bakılarak iki haberin aynı konuda olmadığı ile ilgili güçlü bir karar verilebileceği tezi savunulmuştur.

2 Önbilgi ve Literatür

2.1 Bilgi Erişim Sistemi

Bilgi erişim sistemleri, farklı ortamlarda bulunan belgeler içerisindeki bilginin bulunarak onunla ilgilenen kullanıcılara sunulmasını amaçlayan sistemlerdir [6]. Bir bilgi erişim sistemi: belgelerin bulunduğu derlem, kullanıcı sorguları ve kullanıcıların sorgu cümlelerindeki terimlerle derlemdeki belgelere verilen terimleri karşılaştırarak ilgili belgeleri belirlemek için kullanılan bir erişim fonksiyonundan oluşur. Bu noktada bilgi erişim sisteminin temel işlevi, kullanıcıların bilgi ihtiyaçlarını karşılaması muhtemel derlemdeki ilgili (relevant) belgelerin tümüne erişmek, ilgili olmayanları da ayıklamaktır [7].

2.2 Konu Tespit ve Takip

TDT çalışmalarının amacı; gazete, radyo ya da televizyon haberleri ile ilgili hikâyelerin organize edilmesi, belirlenen bazı hikayelerin tespit edilmesi ve zaman içerisinde bunların takip edilebilmesini sağlayacak teknolojilerin geliştirilmesini sağlamaktır [8]. Bu hedefi gerçekleştirmek için, TDT çalışmaları, sisteme ulaşan haber yayınlarını her biri bağımsız bir olayı tartışacak şekilde ayırmayı amaçlayan

“Hikaye Bölümleme (Story Segmentation)“, sisteme ulaşan haberin daha önce karşılaşılmamış yeni bir hikaye olduğunu belirlemeyi amaçlayan “İlk Hikaye Algılama (First Story Detection)“, sisteme ulaşan haberin hangi konu kümesine ait olduğunu belirlemeyi amaçlayan “Küme Algılama (Cluster Detection)“, belirlenen bir haberin sistem tarafından takip edilmesini amaçlayan “Hikaye İzleme (Topic Tracking)“ ve sisteme ulaşan iki bağımsız haberin aynı konuyu tartışıp tartışmadıklarını anlamayı amaçlayan “Hikaye Bağlantı Algılama (Story Link Detection)“ isimleri altında beş temel göreve bölünmüştür.

2.3 Hikaye Bağlantı Algılama

Hikaye Bağlantı Algılama görevi, geleneksel bilgi erişim sistemlerinde iki farklı dokümanın aynı konuyu tartışıp tartışmadığı belirlenmeye çalışılmaktadır. Önerilen tezde Türkçe bir derlem üzerinde Varlık İsimler kullanarak Hikaye Bağlantı Algılama görevin üzerinde ayrıntılı deneyler yapılarak (kişi, konum, zaman, ne vs. terimleri incelenerek) ilginç sonuçlar elde edildi.

2.4 Başarı Ölçütü Terimleri

Tez çalışmasında elde edilen deney sonuçlarının tüm bilim insanları tarafından kabul görmüş başarım hesaplama ölçütleri kullanılarak ifade edilmesi gerekmektedir. Tezde elde edilen bulguların başarımını değerlendirirken kullanılan temel kavramlar, duyarlık, anma ve F-ölçütüdür. Sistemin başarısı, haberlerin doğru konuya ait olup olmadıklarını belirlemede, doğru ve yalnış örnek sayısiyla ölçülür. Bu başarım ölçütlerin detayları ve hesaplanma yolları alt başlıklarda açıklanmıştır.

2.4.1 Duyarlık Değeri (Precision)

Sistem başarımının ölçülmesinde duyarlık ve ya system kesinlik değeri kullanılan en poüler ve basit ölçü. Bu değer hesaplanırken erişilen belgeler arasında doğru konu atanmış örnek sayısının (TP), toplam erişilen haber sayısına (TP+FP) oranıdır (Eşitlik 2.1). Bu değer her zaman 0-1 Aralığında değişmektedir.

$$Duyarlık = \frac{TP}{TP + FP}$$

Eşitlik 2.1 Duyarlık

2.4.2 Anma Değeri (Recall)

Anma değeri hedefi vurma oranı olarak bilinmektedir. Yani erişilen gereken bilgilere ne oranda ulaşılmış olduğunu gösteren bir değerdir. Bu değer hesaplanırken belirlenen doğru ilişkili pozitif haber sayısının (TP), derlemdeki toplam ilgili belgelerin sayısına (TP + FN) oranını verir. (Eşitlik 2.2). Bu değer her zaman 0-1 arasında değişmektedir.

$$Anma = \frac{TP}{TP + FN}$$

Eşitlik 2.2 Anma

2.4.3 Doğruluk (Accuracy)

Bu sistemde, başarı performansı tanımı için, konu belirlemeler ve atamaların ne kadar doğru olduğu soylemek yalnış değildir. Doğruluk değeri yapılan analizin gerçek değere ne kadar yakın olduğunu gösterir. Bu değer hesaplanırken doğru haber konusu atanmış haber sayısının toplamı (TP + TN), konu atanmış verilerin tümünün sayısına (TP + FP + TN + FN) bölünmesiyle elde edilir (Eşitlik 2.3).

$$Doğruluk = \frac{TP + TN}{TP + FP + TN + FN}$$

Eşitlik 2.3 Doğruluk

2.4.4 F1-Ölçümü (F1-Measure)

Önceki bölümlerde açıklanan Duyarlık ve Anma değerliklerinin harmonik ortalamaları (Eşitlik 2.4) hesaplanarak bu değer elde edilir. Bu değer 0 ile 1 değerleri arasında olur. Yapılmış olan testin doğruluğunu ifade eden bir değerdir. Veri madenciliği ve makine öğrenimi alanlarında yapılmış çalışmalar için en yaygın kullanılan başarı ölçütüdür.

$$F1 = 2 * \frac{Duyarlık * Anma}{Duyarlık + Anma}$$

Eşitlik 2.4. F1 Ölçümü

2.5 Literatür

Bu bölümde geçmişde en sık kullanılmış yöntemlerden bahsedilmektedir. Geleneksel bilgi erişim sistemlerinden TDT programında kullanıcı sorgularının yerini derlemdeki belgelerle ilgili olup olmadığı bilinmeyen yeni belgeler almaktadır. Bu kapsamda hikaye bağlantı algılama görevinin gerçekleştirilmesinde bazen erişim fonksiyonu bazında sorgu-belge yerine belge-belge eşleştirmesi yapmak zorundadır. Bu eşleştirmeler için kullanılan erişim fonksiyonları geleneksel bilgi erişim sistemlerinde kullanılan yöntemlerle benzerlikler göstermektedir. Bu yöntemlerden bazıları; Boole modeli, vektör uzayı modeli, olasılıksal modeller, dil modeli ve ilgi modeli olarak karşımıza çıkmaktadır [9]. Literatürdeki çalışmalara baktığımızda, “Hikaye Bağlantı Algılama görevinin”, TDT çalışmalarında kritik bir öneme sahip olduğu belirtilmiştir. Buna göre, sisteme verilen iki bağımsız hikayenin aynı konuda olup olmadığını anlamayı hedefleyen Hikaye Bağlantı Algılama görevinin başarıyla gerçekleştirilmesi halinde, TDT için pek çok problemin de beraberinde çözülebileceği öngörülmektedir [10].

Hikaye bağlantı algılama görevinin gerçekleştirilmesinde kullanılan pek çok yöntem, karşılaştırılan iki hikaye arasında nekadar fazla sayıda kelimenin örtüştüğünü araştırır. Karşılaştırılan iki hikaye arasında nekadar fazla sayıda örtüşen kelime varsa, bu iki hikayenin aynı konuyu tartışma olasılığının da o kadar yüksek olduğu kabul edilir. Bu yaklaşım, vektör uzayı modellerinden [11] başlayıp, istatistiksel dil modellerinekadar geliştirilen bütün yöntemlerin temelini oluşturmuştur. Pek çok bilgi erişim sisteminde olduğu gibi, çoğu araştırmacı, hangi kelimelerin seçileceği, bu kelimelerin nasıl ağırlıklandırılacağı ve ağırlıklandırılmış olan bu kelimelerin en etkili biçimde nasıl karşılaştırılacakları konularına odaklanmışlardır.

2.5.1 Vektör Uzay Modeli

Önerilen tez kapsamında kullanılacak olan vektör uzayı modeli (vector space model) yöntemine kısaca bakmakta yarar vardır. Vektör uzayı modeli, klasik bilgi erişim sistemleri tarafından erişim fonksiyonu olarak sıkça kullanılan ve 1960'ların sonlarında geliştirilmiş olan ve günümüzde de hâlâ yoğun olarak kullanılan oldukça popüler bir yaklaşımdır. Bu yöntemi kullanan bilgi erişim sistemlerinde, sorgular ve belge koleksiyonunda bulunan her bir belge, koleksiyonda bulunan *t1, t2, ... tn,* gibi *n* adet tekil kelimeden oluşan bir vektör gibi gösterilir. Belgenin vektör biçiminde gösterilmesinde kullanılan *t1, t2, ... tn,* katsayılarının değerleri, ilgili koleksiyon kelimesinin *(ti),* belge veya sorgu içerisinde bulunup bulunmamasına ya da kaç kez bulunduğuna göre belirlenir. Vektör uzayı modelinde, terim ağırlıkları idf-ağırlıklı kosinüs katsayısı olarak tanımlanır ve *tf.idf* (*term frequency * inverse document frequency*) olarak gösterilir. TDT çalışmalarında karşılaştırılması gereken iki belge olduğu için burada her bir belge için birer doküman vektörü oluşturulur ve belgeler arasındaki benzerlik (Eşitlikte 2.5) gibi hesaplanır

$$sim(a,b) = \frac{\sum_{w=1}^{n} tf_a(w).tf_b(w)}{\sqrt{\sum_{w=1}^{n} tf_a^2(w)}.\sqrt{\sum_{w=1}^{n} tf_b^2(w)}}$$

Eşitlik 2.5 Benzerlik Hesaplama

tfa(w), w kelimesinin *a* belgesi içerisindeki sıklığı, *tfb(w), w* kelimesinin *b* belgesi içerisindeki sıklığını ifade etmektedir.

2.5.2 Varlık İsimler Modeli

TDT, haber metinleri içerisinde ifade edilen olaylar (events) ile doğrudan ilgilidir ve bu program içerisinde bir olay; özel bir mekanda, belirli kişi ya da organizasyonların katılımı ile belirli bir zaman diliminde gerçekleşen eylemler olarak tarif edilmektedir

[12]. Bu kapsamda TDT içerisinde, haber metninin gösteriminde varlık isimlerinin (named entity) kullanılması, program içerisindeki olay (event) kavramının tanımı ile eşleşmesi açısından bir zorunluluk gibi görünmektedir. TDT çalışmaları içerisinde tanımlı olan Story Link Detection (SLD) görevinin gerçekleştirilmesi amacıyla haber benzerliklerinin belirlenmesinde varlık isimlerinden yararlanmışlardır [12]. TF.IDF ağırlıklandırma yöntemi baz alarak, bu yöntemin başarımı varlık ismi tabanlı TF.IDF, ağırlıklandırılmamış, varlık ismi genişletme yöntemi ve ağırlıklandırılmış varlık ismi genişletme yöntemleri ile karşılaştırılmıştır. Testler esnasında TDT3 ve TDT4 derlemleri kullanılmıştır. Bu çalışmada varlık isimleri kullanılarak uygulanan ilk yöntemde (TFIDF on entities) varlıklar otomatik olarak tespit edilmiş ve haber metinlerinde geçen diğer kelimeler (isimlendirilmiş varlıklar dışındakiler) atılmıştır. Sonraki aşamada, her bir doküman için belirlenen varlık isimleri kullanılarak doküman vektörleri oluşturulmuştur. Doküman benzerliklerinin belirlenmesinde vektör uzayı modeli kullanılmıştır. Bu yöntemde en büyük problem, bazı dokümanların sağlıklı bir karşılaştırma yapacak kadar varlık ismine sahip olmamasıdır. Bu problemi gidermek için varlıklar arasındaki ilişkileri gösteren çizgeler oluşturulmuş ve aynı haberde bir kez birlikte geçen varlık isimleri ilişkili olarak kabul edilmiştir. Bu yaklaşımda, doküman vektörleri oluşturulurken sadece doküman içinde geçen varlıklar değil bunlarla ilişkili diğer varlıklar da kullanılmıştır (unweighted expansion). Uygulanan son yöntemde ise çizge üzerinde birbiri ile ilişkili varlık isimlerine ilişki derecelerine göre bazı ağırlıklar verilmiş ve yeni doküman vektörleri bu ağırlıklar göz önüne alınarak oluşturulmuştur. Testler sonucu elde edilen veriler SLD görevinde haber benzerlikleri belirlenirken varlık isimlerinin kullanılmasının sistem başarımı üzerinde anlamlı bir maliyet düşüşü sağladığını göstermiştir [12]. Varlık isimlerinin TDT programında “New Event Detection – NED” görevi için kullanıldığı diğer önemli bir çalışmada, NED görevinin gerçekleştirilmesinde varlık isimlerinin kullanılmasının, belirli konularda başarım üzerinde olumlu etkisi olduğunu göstermektedir [13]. Bu çalışmanın devamında Türkçe bir derlem üzerinde NED görevinin gerçekleştirilmesinde varlık isimlerinin sistem başarımı üzerindeki etkilerini araştırmıştır [14]. Araştırmada doküman vektörleri oluşturulurken dört farklı yöntem kullanılmıştır. Bu yöntemler; 1) varlık ismi dışındaki tüm kelimelerin alınması 2) sadece varlık isimlerinin alınması 3) tüm kelimelerin alınması yaklaşımıdır [13]. Buna göre, dokümanlar içerisindeki tüm kelimelerin kullanıldığı vektör gösterimi yaklaşımı en başarılı yöntem olarak rapor

edilmiştir [14]. Bu çalışmada dokümanlar içerisindeki varlık isimlerinin belirlenmesinde otomatik çıkarsama yöntemleri kullanılmıştır.

2.5.3 Olay Modeli (Event Model)

Doküman gösterimleri (document representation) hem geleneksel bilgi erişim sistemleri hem de TDT görevleri için son derece önemli bir aşamadır. Çalışılan alanlara bağlı olmak koşulu ile doküman gösterimi için kelime tabanlı yöntemler, dil modelleri ve çizge (graph) tabanlı yöntemler [15] kullanılmaktadır. Doküman gösterimi ile ilgili olarak kullanılan yöntemlerden bazıları konudan bağımsız olarak geniş bir kullanım alanı bulurken diğer bazı yöntemler sadece sınırlı alanlarda kullanılabilmiştir. TDT çalışmaları da doğası gereği doküman gösteriminin kritik bir öneme sahip olduğu alan olarak karşımıza çıkmaktadır. TDT programı içerisinde doküman gösterimi [15] ve farklı erişim fonksiyonlarının sonuçlarının birleştirilmesi [2][3][4] ile ilgili çalışmalar geçmişten günümüze aktif olarak araştırılmış ve günümüzde hâla popülerliğini korumaktadır.

Geleneksel bilgi erişim sistemlerinde kulanılan doküman gösterme yöntemlerinin aslında TDT için yetersiz kaldığı ve olay tabanlı bu alanda destekleyici farklı yöntemlerin kullanılması gereğine literatürde sıkça vurgu yapılmıştır [16][17][18]. Bu bakış açısı ile TDT içerisindeki dokümanları klasik terim vektörleri ile ifade etmek yerine hikâyeler içerisindeki isimleri, yerleri, zamanı ve konuyu adresleyen verlık isimler ve ya olay vektörlerinin (event vector) kullanılmasının daha anlamlı olacağı fikri destek görmüştür [19]. Buna göre bir olay vektörü; olaya katılan aktörleri ifade eden kişiler (who), olayın gerçekleştiği zamanı ifade eden zaman (when), olayın gerçekleştiği mekânı ifade eden konum (where) ve olayın eylemini ifade eden konu (what) vektörlerinden oluşacak biçimde ifade edilebilir.

NED ile ilgili olarak gerçekleştirilen sonraki bir çalışmada [20] yine varlık isimleri kullanılarak iki farklı hikâyenin karşılaştırılması için isimler, konular ve tam metinler dikkate alınarak bazı deneyler yapılmıştır. Daha önce gerçekleştirilen çalışmada [13] vektörler, varlıkların türüne bakılmaksızın belirlenen tüm varlık isimleri kullanılarak oluşturulmuştu. Yeni çalışmada ise [20], TDT içerisindeki olay (event) tanımından yola çıkarak bir hikâyenin kişiler (who), yerler (where), zaman (when) ve eylemi belirleyen (what) kelimeler kullanılarak ifade edilebileceğini söylemiştir, eğer iki farklı hikâye aynı konuda ise bu hikâyelerin aynı varlık isimlerini ve konu terimlerini paylaşmaları gerekir. Diğer taraftan, eğer iki hikâye birbirine yakın ancak farklı konularda ise varlık isimleri ya da konu terimleri arasında bir eşleşme olsa da muhtemelen her ikisi birden eşleşmeyecektir [20]. Bu çalışmada, varlık isimleri kullanılarak gerçekleştirilen sınıflandırma yöntemlerinin vektör uzayı modeli temel alınarak gerçekleştirilen temel sınıflandırma modelinden anlamlı olarak daha başarılı sonuçlar elde edildiğini rapor etmişlerdir.

Benzer bir çalışmada [21] araştırmacılar, haberlerde geçen isim, yer ve zaman bilgilerini ayrı ayrı vektörlerle ifade etmişlerdir. Bu çalışmada isim, yer ve zaman gibi varlık isimleri otomatik çıkarsama yöntemleri ile elde edilmiş ve doküman içerisinde bunlar dışındaki terimlerin haberin konusunu (what) ifade edeceği belirtilmiştir. Yazarlar, varlık isimlerinin kullanılmasının yeni haber tespit etme probleminde önemli bir başarım artışı sağladığını raporlamışlardır [21]. Yine aynı konudaki çalışmalarında [19] TDT için sadece doküman terimleri kullanılarak gerçekleştirilen doküman gösterimlerinin yeterli olmadığını ve etkili bir sistem için varlık isimleri kullanılması gerektiğini vurgulanmıştır. Çalışmada özellikle yer ve zaman karşılaştırmaları için kesişime dayanan benzerlik metrikleri önerilmiştir.

TDT görevlerinin gerçekleştirilmesinde varlık isimlerinin kullanılmasının literatürde genellikle başarım üzerindeki olumlu etkilerinden bahsedilmekle birlikte bunun tersinin savunulduğu çalışmalarda vardır [22], Korece haberlerden oluşturulmuş olan derlem üzerinde gerçekleştirdikleri çalışmalarında zaman (when) bilgisinin konu takibi (topic tracking) için gerçekleştirilen deneylerde başarımı anlamlı bir oranda artırmadığını ifade etmişlerdir.

Türkçe için gerçekleştirilen benzer çalışmalar ağırlıklı olarak metinlerden varlık isimlerinin (isim, yer, zaman, organizasyon v.b.) otomatik olarak çıkarılmasını sağlayan makine öğrenme yöntemleri üzerine yoğunlaşmıştır [23][24][25][26]. Bilgi erişimin bir parçası olarak varlık isimlerinin erişim fonksiyonu ya da bunu destekler nitelikte kullanıldığı çalışmalar ise oldukça sınırlıdır [27][28][29]. Türkçe derlemler üzerinde varlık isimlerinin kullanılması ile elde edilecek erişim etkinliği konusunda sınırlı çalışmalara dikkat çekilmiş [14] ve bu konuda daha derinlemesine çalışmalar yapılması gerektiğini vurgulamışlardır.

2.6 Yöntem

TDT içerisinde başarısı kanıtlanmış olan Vektör Uzayı Modeli (Vector Space Model) ve daha önce Türkçe üzerinde yapılan çalısmamızın sonuçlarından faydalanarak[3][4], Türkçe haberler üzerinde vektör uzay modeli, yüksek başarı elde etmiştir [3]. Dolayısıyla tez çalısmaları devamında sadece vektör uzay modelin temel yöntem olarak kabul edip, bu yöntemle birlikte Varlık İsimlerinin (Named Entity) kullanılmasının başarım üzerindeki etkileri değerlendirildi.

2.6.1 Yöntem Detaylı Tanımı

Sistem testleri esnasında yöntem olarak seçilen vektör uzayı, geçmişten günümüze, bilgi erişim sistemleri ile ilgili çalışmalarda erişim fonksiyonu olarak genellikle tek başına kullanılmıştır. Vektör uzayı mdelinin arkasında yatan felsefeye bakıldığında, vektör uzayı yöntemi karşılaştırılan belgelerdeki terim çakışmalarına göre benzerlik değerlerini hesaplıyor, dolayısıyla vektör uzayı yönteminin kaçırdığı konuyla ilgili belgelerin varlık isimleri yöntemiyle yakalama şansı olabilir. Bu bağlamda, SLD görevinin gerçekleştirilmesinde vektör uzayı ve varlık isimler yöntemin vereceği bağımsız kararların OR (YA) mantıksal operatörü ile birleştirilmesi sonucu sistemin anma (recall) değerinin oldukça yüksek çıkması, diğer bir deyişle ilgili belgelerin büyük bir çoğunluğuna erişilmesi sağlanacaktır. Diğer taraftan bu tür bir birleştirme muhtemelen ilgili belgelerin yanında ilgisizleri de getireceği için duyarlık (precision) düşecektir. Bununla birlikte vektör uzayı ve olay modelinin vereceği bağımsız kararların AND (VE) mantıksal operatörü ile birleştirilmesi ile elde edilecek sonuç yöntemlerin birlikte verdikleri ilgililik kararlarının yorumlanması, diğer bir deyişle bir yöntemin diğerinden farklı olarak verdiği doğru kararların belirlenmesi açısından açıklayıcı olacaktır. Bu bağlamda who (kim), where (nerede), when (ne zaman) vs. gibi etiketlerle işaretlenmiş varlık isimlerinin teker teker ve tümü bir arada değerlendirilerek haber benzerlikleri üzerindeki etkileri açık bir şekilde ortaya konuldu. Tez kapsamındaki varlık isimleri ile ilgili çalışmaların iki boyutta incelendi. Bunlardan birincisinde; haberler içerisindeki varlık isimleri teker teker (kim, konum, ne zaman vs.) ve ikincisinde bu varlıkların birlikte kullanılarak haber benzerlikleri üzerindeki etkileri araştırıldı.

2.6.2 Ağırlandırma (TF-IDF)

TF-IDF ağırlandırma yöntemi, IR alanında terim normalizasyon yöntemidir. Yani derlemdeki dokumanlar içerisinde terimlerin önemini belirlemektedir. Dolayısiyla derlemde sıkça geçen kelimelerin etkisini azaltmak için, derlemdeki terimler üzerinde normalzasyon işlemi gerçekleştiriyor. Cümlelerde yer alan kelimelerin ne sıklıkla kullanıldığına yani frekansına ve diğer dokümanlarda geçme sıklıklarını birlikte hesaplayarak, kategoriler için en önemli kelimelerin tespitini sağlar. Aslında durma kelimeleri tespit etmede çok başarılı bir yöntem olarak kullanılmaktadır.

Term Frequency *(TF), t* teriminin bir dokumanında geçme sıklığı *(f),* bölü ilgili dokumanda toplam terim sayısı *(df)* . Inverse Document Frequency *(IDF), N* değeri toplam doküman sayısı, *df* ise kelimenin tüm derlemde toplam kaç dokümanda geçtiği. *IDF* değerinin yüksek çıkması terimin ilgili kategorinin belirlenmesi için önemli olduğunu gösterir.

$$TF = \frac{f}{df}$$

Eşitlik 2.6 TF Hesaplanması

$$IDF = \log\left(\frac{N}{df}\right)$$

Eşitlik 2.7 IDF Hesaplanması

$$TF - IDF = TF * IDF$$

Eşitlik 2.8 *TF-IDF* Hesaplanma Yöntemi

2.6.3 Filtreleme (ZEMBEREK)

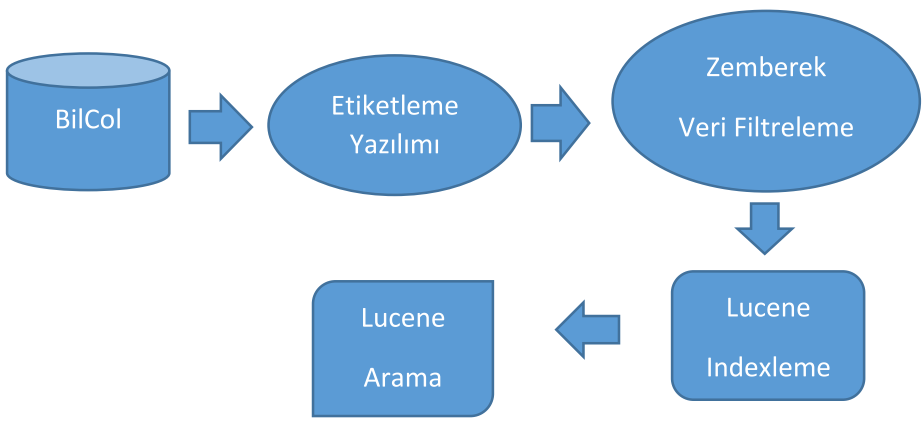
Zemberek 2005 yılında Tübitak projesi kapsamında, ilk ve tek açık kaynak kodlu Türkçe doğal dil işleme kütüphanesidir [30]. Zemberek doğal dil işleme (NLP) alanında sıkça kullanılan uygulamalardandır. Genel olarak içinde kullanılacağı dile ait kelimeleri ve bu kelimelerin köklerini barındırır. Bu tür uygulamaların amacı analiz yapılacak olan metin verilerinin en doğru bir biçimde değerlendirilmesini sağlayabilmektir.

Bu tezde kok bulma işlemi için, Zemberek kütüphanesi kullanılmıştır. Açık kaynak kodlu olduğu için çeşitli yazılım kütüphanelerinde bu uygulamaya erişmek mümkündür. Derlemdeki haberler üzerinde Zemberek aracın kullanarak gövdeleme ve filtreme işlemi tamamlandıktan sonra, artık dokuman TF-IDF hesaplama ve ardından dokuman vektörlerin hazırlanması için gerekli veri seti hazır olmuş oluyor.

2.6.4 Indeksleme (Apache Lucene)

Apache Lucene, Java tabanlı metin arama işlemini sağlayan açık kaynak kodlu kütüphanedir [31]. Aslında Zemberek kütüphanesinden çıkan veri, girdi olarak Lucene veriliyor ve indexleme işlemi veri seti üzerinde gerçekleşiyor. Indexleme esnasında her dokuman için TF-IDF değerleri hesaplandıktan sonra dokuman vektörleri elde edilir. Indexleme ve vektör aşamasından sonra Lucene Kütüphanesi indekslenmiş veri üzerinde arama yapılmasını sağlamaktadır.

2.6.5 Sistem Genel Mimari



Şekil 2.6.1 Sistem Mimari

3 Veri Hazırlama

3.1 Derlem Etiketleme Literatürü

Literatürde konu tespit ve takip sistemleri ve farklı bilgi erişim modelleri (vektör uzayı modeli, vb.) ile gerçekleştirilmiş çeşitli araştırmalar bulunmaktadır. Aşağıda değerlendirilmiş olan çalışmaların çoğunda bir derlem üzerinde çalışılmış ve sonuçlar değerlendirilirken anma ve duyarlık değerleri de dikkate alınmıştır. Literatürde yer alan çalışmalara baktığımızda, haber derlemleri üzerinde gerçekleştirilmiş çalışmalarda, örneğin bir metinde yakalanabilen olay, zaman ve zamansal ilişkilere ait ayrıntılı bir bilgi şeması tanımlamaktadır [32].

Yeni olay algılamaya yönelik bir çalışmada ise, isimler, konular ve tam metinlere dayalı üç kosinüs benzerliği bulunarak iki hikaye karşılaştırılmıştır. Üzerinde durulan sınıflayıcı modelde, tüm koleksiyonlarda vektör uzayı modeli çizgisi üzerinde istatistiksel olarak anlamlı gelişmeler olduğu görülmüştür [20]. Çalışmada isimlendirilmiş varlıklar; Olay, Jeopolitik varlık, Dil, Konum, Milliyet, Kurum, Kişi, Tarih, Zaman gibi etiketlerle belirlenmiştir. Bu özelliklerin kullanılmasında olayların bir grup terimle tanımlanmış ve her olayın kişi, yer gibi isimlendirilmiş varlıklarla karakterize edilmiş olduğu düşüncesi hakim olmuştur. Eğer iki hikaye de aynı konudaysa, konu terimlerinin yanı sıra isimlendirilmiş varlıklarını da paylaşırlar. Eğer benzer konulara sahiplerse isimlendirilmiş varlıklar ya da konu terimleri (ikisinden biri) eşleştirilecektir. Deney için TDT2, TDT3, TDT4 ve TDT5 derlemleri kullanılmıştır.

Haberlerdeki tekrarlara yönelik 2009 yılında yapılmış tezde [27], tekrarlı haber bulmak için varlık isimlendirmeyi kullanma şeklinde değişik bir yaklaşım ileri sürülmüştür, bu yaklaşım Türkçe haber belgeleri ile değerlendirilmiştir ve Bilkent Haber Portalı'ndan sağlanan haberlerden oluşan veri seti kullanılmıştır.

3.2 BilCol (Veri Seti)

Deneysel çalışmaların gerçekleştirilebilmesi amacıyla, Bilkent Üniversitesi'nde geliştirilen ve benzer makale çalışmalarında kullanılan BilCol-2005 [14] haber derlemin kullanıldı. BilCol-2005 deney derlemi TDT çalışmalarından esinlenerek hazırlanmıştır. Derlem 209.305 doküman ve seksen tanesi insanlar tarafından etiketlenmiş olaylardan oluşmaktadır.

3.3 Veri hazırlama yöntemi

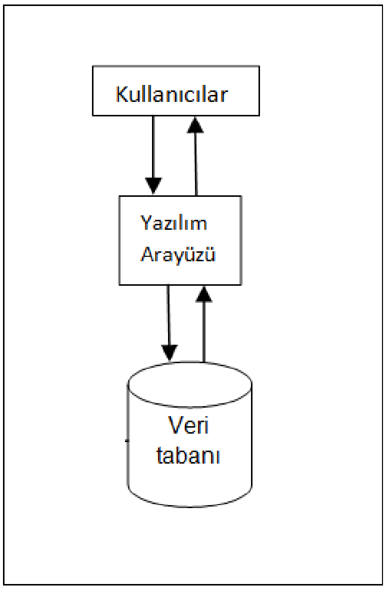
Tez varlık isimlerinin doküman benzerliklerinin belirlenmesindeki etkilerine odaklandığı için BilCol-2005 derleminin etiketlenmesinde, otomatik yöntemlerle varlık isimlerin çıkarılmasını sağlayan makine öğrenme yöntemleri yerine varlık isimlerin etiketlenmesi için manual yöntem kullanıldı. Bu kapsamda sistem testlerinin gerçekleştirilmesini sağlayacak sayıdaki belge, haberlerde geçen kim(who), nerede (where), ne zaman (when) vs. gibi sorularına yanıt verecek etiketlerle işaretlendi ve testler bu derlem üzerinde gerçekleştirildi.

Derlem etiketleme, tez kapsamında tanımlanan yöntemlerin uygulanabilmesi için öncelikli olarak gerçekleştirilmesi gereken adımdır. Bu adımın temel hedefi BilCol-2005 derleminin varlık isimlerini içerecek biçimde etiketlenmesini sağlamaktır. Ancak bu adımı gerçekleştirmek için derlemin en az hata ile etiketlenmesini sağlayacak web tabanlı bir etiketleme yazılımının geliştirilmesi ihtiyacı doğmuştur. Bu kapsamda etiketleme çalışmaları BilCol-2005 derleminde bulunan ve ilgililik değerlendirmeleri yapılmış olan 5881 adet haber üzerinde gerçekleştirildi.

Etiketleme Yazılımında amaçlanan, BilCol-2005 derleminde bulunan ve daha önceden belirlenmiş olan 80 konu başlığı ve bu konu başlıklarına ait 5881 haberin, tez kapsamında, kullanıcılar tarafından tek tek okunarak, haber içinde bulunan kim, nerde ve ne zaman gibi sorularına cevap veren kelimelerin etiketlenmesini sağlamaktır.

3.3.1 Etiketleme Yazılımı Mimarisi

Geliştirilen etiketleme yazılımının kullanıcıların eş zamanlı olarak İnternet üzerinden etiketleme yapabilmelerine olanak sağlayacak bir mimari yapıda olması hedeflenmiştir. Bu kapsamda üç katmanlı bir yazılım mimarisi kullanılarak ara yüz, iş mantıkları ve veri tabanı farklı katmanlarda tutulmuştur. Yazılım geliştirme sürecinde iş mantığı bileşenlerinin oluşturulmasında J2EE platformu, ara yüz tasarımlarında Vaadin Framework ve veritabanı işlemleri için MySQL Veri Tabanı Yönetim Sistemi, uygulama sunucusu olarak Apacahe Tomcat kullanılmıştır.



Şekil 3.3.1 Etiketleme Yazılımı Mimari Yapısı

3.3.2 Veri Tabanı Tasarımı

BilCol-2005 derlemi tüm haberleri içerisinde barındıran tek bir XML dosyası biçiminde oluşturulmuştur. Haberler üzerinde tek tek çalışılması gerektiği için 80 konu başlığındaki 5881 haberin ve bu haberlerle ilgili bilgi alanlarının veri tabanına aktarılması gerekmiştir. Bu kapsamda oluşturulan veri tabloları ve her bir tablonun amacı takip eden kısımda kısaca açıklanmıştır.

Haberler Tablosu: Haberler tablosu Etiketleme Yazılımında kullanılacak olan 5881 haberin hepsinin bulunduğu tablodur (Şekil 4.1) Bu tabloda aşağıdaki alanlar bulunmaktadır.

HaberID: Her haber için atanmış olan, tekil numaradır, aynı numaralı iki haber Tablo'da bulunamaz.

HaberBaşlık: Haberin başlığının tutulduğu kısım.

HaberKaynak: Haber kaynağının isminin tutulduğu kısım.

HaberTarih: Haberin yayınlama tarihinin tutulduğu kısım.

Haberİçerik: Haber metninin tutulduğu kısım.

Column Name	Data Type	Allow Nulls
HaberID	int	☐
HaberBaşlık	nvarchar(50)	☑
HaberKaynak	nvarchar(50)	☐
HaberTarih	date	☐
Haberİçerik	text	☑

Şekil 3.3.2 Veri Tabanı-Haberler Tablosu

Konular Tablosu: Haberler için belirlenmiş olan konuların tutulacağı tablodur (Şekil 4.2). Bu tabloda aşağıdaki alanlar bulunmaktadır.

KonuID: Her konu için atanmış olan, özel numaradır, aynı numaralı iki konu tabloda bulunamaz.

KonuBaşlık: Konu başlıklarının tutulduğu kısım.

Column Name	Data Type	Allow Nulls
KonuID	int	☐
KonuAdı	nvarchar(50)	☐

Şekil 3.3.3 Veri Tabanı-Konular Tablosu

Haber Konuları Tablosu: Hangi haberin hangi konuya ait olduğunun tutulduğu tablodur (Şekil 4.3). Bu tabloda aşağıdaki alanlar bulunmaktadır.

HaberID: Haber için atanmış tekil numaranın bulunduğu kısımdır.

KonuID: Haberin ait olduğu konunun tekil numarasının tutulduğu kısımdır.

Column Name	Data Type	Allow Nulls
HaberID	int	☐
KonuID	int	☐

Şekil 3.3.4 Veri Tabanı- Haber Konuları Tablosu

Etiketlenen Haberler Tablosu: Etiketleme tablosunda, kullanıcıların etiketlemeyi bitirdiği haberler tutulur (Şekil 4.4). Bu tabloda aşağıdaki alanlar bulunmaktadır.

HaberID: Etiketlenmiş olan haberin özel sıra numarası tutulur.

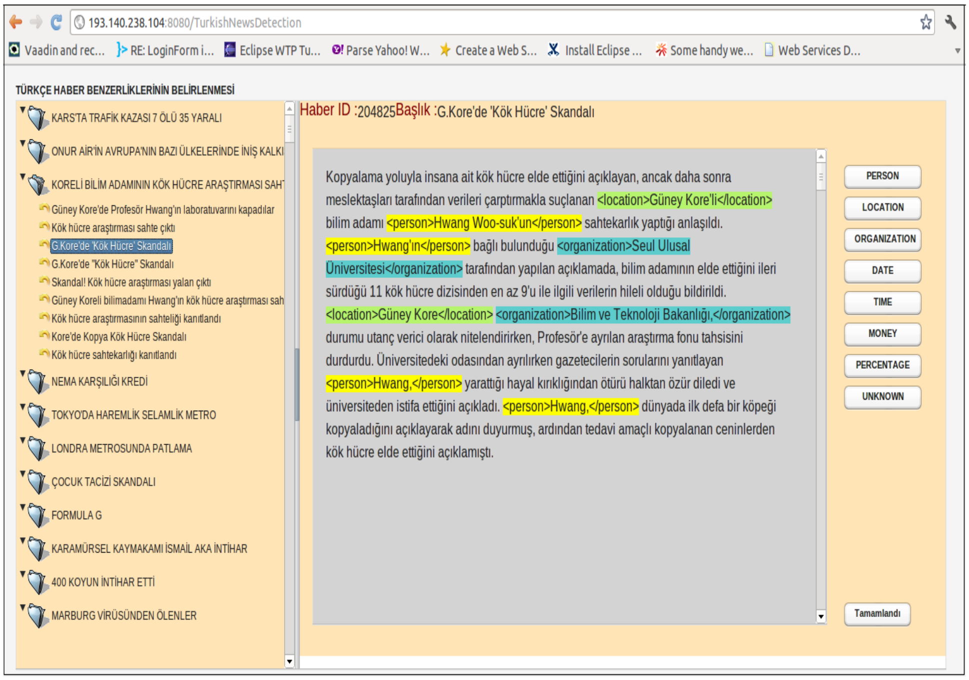
HaberEtiketlenmişİçerik: Etiketlenmiş haberin, değişmiş olan yeni metninin tutulduğu kısım.

Column Name	Data Type	Allow Nulls
HaberID	int	☐
HaberEtiketlenmişİçerik	text	☐

Şekil 3.3.5 Veri Tabanı-Etiketlenen Haberler Tablosu

3.3.3 Yazılım Kullanıcı Arayüzü

Hazırlanan Etiketleme Programı, Hacettepe Üniversitesi uygulama sunucusundan yayınlanması gerçekleştirildi. Uygulama çalıştırıldığında, kullanıcı girişi yapıldıktan sonra sayfada görünen ilk pencere, kullanıcı girişinin yapılmasını sağlamak üzere altdaki şekilde örnek olarak gösterilmiştir. Sol panelde haber konularını gözükmekte, konu seçiminden sonra o konu ile ilgili daha önceden belirlenmiş olan bütün haberler bu konu başlığı altında listelenmektedir. Haber içeriğinin gösterimi ile kullanıcı haberi okur ve gerekli gördüğü kelimeleri sağ tarafta bulunan butonları kullanarak etiketlemesini gerçekleştirir. Etiketleme işlemi, gerekli gördüğü kelimenin üstünü seçtikten sonra daha önce tanımlanmış etiketleme kurallarına göre, uygun olan etiket butonuna basarak gerçekleştirildi.



Şekil 3.3.6 Yazılım Arayüzü ve Etiketleme Yapılmış Haber Öreneği

3.3.4 Etiketleme Kuralları

Bir haber içeriğini oluşturan metinde geçen kelimelerin, nitelediği veya cevapladığı soru zamirlerine göre işaretlenmesi işlemini, “Etiketleme” olarak adlandırmaktayız. Bu işlem sırasında, BilCol-2005 derlemi içerisinden alınmış ve konuları belirlenmiş olan 5881 haberin okunması ve haber metni içerisindeki kelimelerin özenle seçilerek doğru bir şekilde işaretlenmesi gerekmektedir. 80 konu ve 5881 haber, çalışan kullanıcı sayısına göre, eşit bir şekilde dağıtılmıştır ve her kullanıcı, kendisine ayrılmış bu konuları etiketlemesini gerçekleştirdi.

Örnek:

<person>Mustafa Kemal Atatürk</person>

<location>Ankara</location>

Tez kapsamında haberler içerisindeki varlık isimleri belirlenirken “kim (who)”, “ne zaman (when)”, “nerede (where)” ve “ne (what)” sorularına cevap verecek etiketlemelerin yapılması planlanmış, ancak literatürde farklı çalışmalarda daha ayrıntılı etiketlemeler yapıldığı gözlenmiştir. Bu kapsamda tez önerisinde belirtilen ve yukarıda sıraladığımız etiketler genişletilerek varlık isimlerinin "kişi (person)", "kurum (organizaztion)", "konum (location)", "tarih (date), "zaman (time)", "yüzde (percentage)", "Para birimi"(money) ve "belirsiz"(unknown) olarak etiketlenmesine karar verilmiştir. Belirisiz etiket, aslında metinlere varlık isim özelliği taşıyan ama etiketleme kurallarına göre varlık isimlerinin etiket türlerinin hiçbirisinin altına geçmiyenler için kullanılmaktadır. Bu sayede hem bu tez kapsamında belirlenen yöntemler test edilebilecektir hem de oluşturulan etiketlenmiş derlemin çok daha geniş bir akademik çevre tarafından kullanılabilmesi sağlanacaktır. Bu kapsamda etiketleme ile ilgili olarak belirlenen bazı ön kurallar ve örnek etiketler aşağıda sunulmuştur.

Genel kurallar:

- Etiketlenecek ifadeler mümkün olduğunca en küçük parçaya bölünerek etiketlendi. Örneğin: "İzmir Atatürk Stadı" benzeri ifadeler bölünecek ("İzmir": Location, "Atatürk": Person).
- Aynı haber içinde açık adı ve kısaltması birlikte verilen kurum adları ayrı ayrı etiketlendi. Örneğin: "BM": Organization, "Birleşmiş Milletler": Organization.
- Kurum isimleri (örneğin Üniversite adları) bölünmeden (tamamı Organization olarak etiketlendi. Örneğin: "İstanbul Medeniyet Üniversitesi": Organization.
- Organization etiketi yalnızca resmi niteliği olan kurumlar için kullanıldı.
- Herhangi bir şekilde kişi adı geçiyorsa Person etiketi kullanıldı, kişi kast edilerek kullanılan mahlas ya da unvanlar (örneğin "Başbakan", "Doç.Dr." etiketlenmedi).
- Ülke kısaltmaları Location olarak etiketlendi (Örneğin: TC, ABD, UK, vb.).
- Ülke, eyalet, bölge, il, ilçe, semt, köy adları Location olarak etiketlendi.
- Mahalle, stat, spor salonu, vb. yer isimleri etiketlenmedi.
- Doğrudan "yüzde" yazıyorsa ya da "%" işareti kullanılmışsa Percentage şeklinde etiketlendi.
- Gün, ay, yıl belirtilen ifadelerin her biri ayrı ayrı olmak kaydıyla Date olarak işaretlendi.

- Saat içeren ifadeler, Time olarak işaretlendi
- Yukarıdaki kuralların hiç birisine uymuyan ama varlık isimi özelliği taşıyan, Unknown olarak işaretlendi.
- Para birimiyle birlikte geçen sayılar birlıikte Money olarak işaretlendi.
- Irk belirten ifadeler etiketlenmedi.

<Person> Etiket Örnekleri:

- Gülen
- Serratia
- Marshall Ballard
- Atatürk
- Özer Gürbüz

<Organization> Etiket Örnekleri:

- Nobel
- Galatasaray
- Fenerbahçe
- Birleşmiş Milletler
- BM

<Location> Etiket Örnekleri:

- Asya
- Avrupa
- Kuzey Afrika
- Ortadoğu
- Türkiye

<Date> Etiket Örnekleri (her bir tarih parçası ayrı ayrı etiketlendi):

- 12-13
- Nisan
- 2005
- 21
- Ekim

<Time> Etiket Örnekleri:

- 08.49'da
- 11.32.54

<Money> Etiket Örnekleri:

- 540 YTL
- 100$
- 50tl
- 100 milyon dolar
- 49 cent

<Percentage> Etiket Örnekleri:

- 75%
- Yüzde 75

<Unknown> Etiket Örnekleri:

- Dünya Kupası
- Şampiyonlar Ligi'nde
- UEFA Kupasın'da
- Müslüman
- Dünya Kadınlar Günü

3.4 Etiketli Varlık İsimlerin İstatistikleri

Sistem testlerinin gerçekleştirilebilmesi için BilCol-2005 derleminde konu başlıkları bilinen 5.872 adet haber içerisindeki varlık isimleri etiketlenmiştir. Etiketli belgeler üzerinden, etiketlenmiş varlık isimlerinin çıkarılmasıyla (Tablo 3.1) istatistiksel bilgiler elde edildi, sonraki aşamada test senaryoları oluşturulmuş.

Varlık İsim	Sayı
Person	45.201
Location	35.255
Organization	29.059
Date	10.622

Time	1.118
Money	2.708
Percentage	2.608
Unknown	10.258

Tablo 3.1 Derlem Varlık İsimler İstatikleri

4 Uygulanan Test Senaryoları

4.1 Vektör Uzay Modeli Test Senaryosu

Vektör Uzay Modelinin başarısı test edilmiştir. Bu hedefi gerçekleştirmek amacıyla yürütülen testlerde aşağıdaki test senaryoları gerçekleştirildi:

- Testler ilgililik değerlendirmesi yapılmış olan belgeler (5.872 adet) üzerinden yapıldı.
- Testlerde her konu eğitim ve test belgeleri ayırımı yapıldıktan sonra tüm belgeler (5.872 adet) üzerinde yapıldı.
- Belgelerin üçte biri eğitim seti ve kalan kısım test seti olarak belirlendi
- Vektör Uzay Modelin haber konu algılamada başarısını belirlemek için aşağıdaki adımlar gerçekleştirildi:
    - Haberlerdeki kelimelerde son ekler atıldı.
    - Haberler üzerinde Zemberek kütüphanesin kullanarak gövdeleme işlemi gerçekleştirildi.
    - Haberlerin Apache Lucene kütüphanesin kullanarak vektör modelleri oluşturuldu.
    - İlgililik değerlendirmesi yapılmış olan eğitim seti belgelerin hepsi sorgu olarak kabul edildi.
    - Her bir sorgu için derleme gönderilecek vektör uzayı modeli kullanılarak üretilen sorgu-belge eşleşme skorları belirlendi.
    - Belirlenen tüm bu skor değerleri içerisinden, sorgunun ilgili olduğu bilinen belgeler için üretilen skor değerleri çıkarıldı

ve ilgili sorgu-belge eşleşmeleri için ortalama skor değeri başlangıç eşiği olarak kabul edildi.

- Bu başlangıç eşiğine göre tüm haberler için anma/duyarlık ve f-ölçüsü değerleri hesaplandı.
- Sonraki aşamada eşik değeri belirli oranda azaltılıp-artırılarak anma/duyarlık değerleri her bir eşik için tekrar hesaplandı.
- Anma ve duyarlığın birlikte en yüksek oldukları (ya da birbirlerine en yakın oldukları) değer Vektör Uzay Modelin kesin eşik değeri olarak hesaplandı.
- Eşik değeri elde ettikten sonra, test setindeki haberlerin hepsi sorgu olarak kabul edildi ve test işlemi gerçekleştirildi.

4.2 Vektör Uzay Modeli (VUM) Test Sonuçlar

Yukarıdaki senaryoya göre vektör uzayı modeli için gerçekleştirilen test sonuçları için elde edilen tablo aşağıda sunulmuştur. Testler esnasında gövdeleme yapılmış ancak hem sorgu hem de dizin tarafında durma kelimeleri çıkarılmıştır.

Yöntem	Duyarlık(**d**)	Anma(**a**)	f-ölçü(**f**)	Eşik(**e**)
VUM	**0.61**	**0.58**	**0.59**	**0.05**

Tablo 4.1 Vektör Uzay Model Sonuç

4.3 Varlık İsimlerin Vektör Modeli (VİVM) Test Senaryosu

Varlık isimlerin haber konu bulma sistem başarısında ve temel yöntem üzerinde etkileri hedeflenmiştir. Bu hedefi gerçekleştirmek amacıyla yürütülen testlerde aşağıdaki test senaryoları gerçekleştirildi:

- Testler ilgililik değerlendirmesi yapılmış olan belgeler (5.872 adet) üzerinden yapıldı.

- Her haberde geçen tüm Varlık İsimlerin kullanarak, her haber için Varlık İsim Vektörü oluşturuldu ve Varlık İsimler Vektör Modelin haber konu algılamada başarısını belirlemek için aşağıdaki adımlar gerçekleştirildi.
    - Haberlerde tüm etiketlenen varlık isimler çıkarıldı.
    - Varlık isimler terimlerinde son ekler atıldı.
    - Varlık isimler üzerinde gövdeleme yapıldı.
    - Her haber için çıkartılan tüm varlık isimleri kullanarak Apache Lucene kütüphanesin kullanarak vektör modeli oluşturuldu.
    - İlgililik değerlendirmesi yapılmış olan 5.872 belge test ve eğitim setine ayrıldı
    - Eğim setindeki belgelerin hepsi sorgu olarak kabul edildi.
    - Her bir sorgu için derleme gönderilecek varlık isim vektör modeli kullanılarak üretilen sorgu-belge eşleşme skorları belirlendi.
    - Belirlenen tüm bu skor değerleri içerisinden, sorgunun ilgili olduğu bilinen belgeler için üretilen skor değerleri çıkarıldı ve ilgili sorgu-belge eşleşmeleri için ortalama skor değeri başlangıç eşiği olarak kabul edildi.
    - Bu başlangıç eşiğine göre tüm haberler için anma/duyarlık ve f-ölçüsü değerleri hesaplandı.
    - Sonraki aşamada eşik değeri belirli oranda azaltılıp-artırılarak anma/duyarlık değerleri her bir eşik için tekrar hesaplandı.
    - Anma ve duyarlığın birlikte en yüksek oldukları (ya da birbirlerine en yakın oldukları) değer Varlık İsimler Vektör Modelin kesin eşik değeri olarak hesaplandı.
    - Eşik değeri belirlendikten sonra test setinde geçen haberlerin hepsi sorgu olarak kabul edildi ve test işlemi gerçekleştirildi.
- Varlık İsimlerin tümünü vektör oluşumunda kullanarak konu algılamada başarım testlerin gerçekleştirdikten sonra, her varlık isim türü için (Person, Location vs) vektör oluşturup, farklı varlık isim vektörlerin testleri gerçekleştirildi. Her haber için farklı Varlık İsim Vektörlerin Modelin haber konu algılamada başarısını belirlemek için aşağıdaki adımlar gerçekleştirildi.
    - Haberlerde tüm varlık isimler çıkartarak türüne göre ayrım yapıldı.

- Her haberde geçen varlık isimlerin türleri “Person”, ”Location”, ”Organization”, ”Date”, ”Time”, ”Money”, ”Percentage” ve ”Unknown” olarak daha önce etiketleme aşamasında adlandırılmıştır.
- Varlık isimler terimlerinde son ekler atıldı.
- Varlık isimler üzerinde gövdeleme yapıldı.
- Her haber için geçen tüm varlık isimlerin türlerine göre, farklı varlık isim vektör modelleri Apache Lucene kütüphanesin kullanarak oluşturuldu.
- İlgililik değerlendirmesi yapılmış olan 5.872 belgenin içinden, etiketli olanlar eğitim ve test kümelerine ayrıldı.
- Haberlerde çıkartılan varlık isim vektörlerine göre aşağıdaki testler gerçekleştirildi.
- Haberlerde geçen “Person” varlık isim vektörün sorgu olarak kullanıldı.
- Haberlerde geçen “Location” varlık isim vektörün sorgu olarak kullanıldı.
- Haberlerde geçen “Organization” varlık isim vektörün sorgu olarak kullanıldı.
- Haberlerde geçen “Date” varlık isim vektörün sorgu olarak kullanıldı.
- Haberlerde geçen “Time” varlık isim vektörün sorgu olarak kullanıldı.
- Haberlerde geçen “Money” varlık isim vektörün sorgu olarak kullanıldı.
- Haberlerde geçen “Percentage” varlık isim vektörün sorgu olarak kullanıldı.
- Haberlerde geçen “Unknown” varlık isim vektörün sorgu olarak kullanıldı.
- Her bir sorgu için derleme gönderilecek varlık isim vektör modeli kullanılarak üretilen sorgu-belge eşleşme skorları belirlendi.
- Belirlenen tüm bu skor değerleri içerisinden, sorgunun ilgili olduğu bilinen belgeler için üretilen skor değerleri çıkarılacak ve ilgili

sorgu-belge eşleşmeleri için ortalama skor değeri başlangıç eşiği olarak kabul edildi.

- Bu başlangıç eşiğine göre tüm haberler için geçen varlık isim vektör türlerine göre anma/duyarlık ve f-ölçüsü değerleri hesaplandı.
- Sonra her varlık isim vektör türüne eşik değeri belirli oranda azaltılıp-artırılarak anma/duyarlık değerleri her bir eşik için tekrar hesaplandı.
- Anma ve duyarlığın birlikte en yüksek oldukları (ya da birbirlerine en yakın oldukları) değer her Varlık İsim Vektör türü için kesin eşik değeri olarak hesaplandı.
- Tüm varlık isim vektörlerin eşik değerleri belirlendikten sonra test işlemi test seti üzerinde gerçekleştirildi.

4.4 VİVM Test Sonuçlar

Yukarıdaki senaryoya göre varlık isimler vektör modeli için gerçekleştirilen test sonuçları için elde edilen tablo aşağıda sunulmuştur. Testler esnasında varlık terimlerden son ekler atılmış ve gövdeleme yapılmış.

VİVM	d	a	f	e
Tüm İsimler	0.65	0.56	**0.60**	0.02
Person	0.77	0.39	**0.51**	0.01
Location	0.51	0.51	**0.51**	0.25
Organization	0.46	0.46	0.46	0.01
Date	0.33	0.36	0.34	0.2
Time	0.34	0.33	0.34	0.25
Money	0.49	0.28	0.36	0.01
Percentage	0.76	0.22	0.34	0.01

Unknown	0.89	0.55	0.67	0.01

Tablo 4.2 Varlık İsim Vektör Model Sonuç

Tablo 4.2 incelededeğimizde, varlık isimlerin vektörel testinde tüm varlıkların vektör oluşumunda kullanılması daha etkin olduğu ortaya çıkıyor. Ayrıca varlık isimlerin tek tek incelediğimizde kişi ve konum en belirgin olarak anlaşılmıştır.

4.5 Varlık İsimlerin Kesişim Model (VİKM) Test Senaryosu

Bu aşamada Varlık isimlerin kesişim kontrolu haber konu bulma sistem başarısında ve temel yöntem üzerinde etkileri hedeflenmiştir. Bu hedefi gerçekleştirmek amacıyla yürütülen testlerde aşağıdaki test senaryoları gerçekleştirildi:

- Testler ilgililik değerlendirmesi yapılmış olan belgeler (5.872 adet) üzerinden yapıldı.
- Haberde tüm Varlık İsimlerin çıkardıktan sonra, her haber için Varlık İsim tabloları oluşturuldu ve Varlık İsimler kesişim Modelin haber konu algılamada başarısını belirlemek için aşağıdaki adımlar gerçekleştirildi.
    - Her haber için tüm etiketlenen varlık isimler çıkartıldı.
    - Varlık isimler üzerinde son ekler atıldıktan sonra Zemberek kütüphanesin kullanarak gövdeleme gerçekleştirildi.
    - Tüm haberler için varlık isimler tabloları oluşturuldu.
    - İlgililik değerlendirmesi yapılmış olan 5.872 belgenin içinden, tablosu oluşturanların hepsi sorgu olarak kabul edildi.
    - Haberlerde çıkartılan varlık isim tablolarına göre iki farklı adım gerçekleştirildi.
    - Birinci adımda sorgulama yapıldığında haberlerde geçen bütün varlık isimlerin kesişimi sorgu yapıldığında gerçekleştirildi.
    - İkinci adımda ise haberlerde geçen varlık isimlerin türüne göre tek tek ayrı kesişim kontrolleri gerçekleştirildi.
    - İkinci adımdaki kullanılan varlık isimler türlerin sırayla şu şekilde “Person”, ”Location”, ”Organization”, ”Date”, “Time”, “Money”, “Percentage”, “Unknown” adlandırılmıştır.

- Sonunda bu aşamada varlık isimlerin kesişiminin haber konu bulmada gerçekleşen test adımların aşağıda özetliyebiliriz:
- Tüm varlık isimlerin kesişim testi.
- “Person” varlık isimlerin kesişim testi.
- “Location” varlık isimlerin kesişim testi.
- “Organizaiton” varlık isimlerin kesişim testi.
- “Date” varlık isimlerin kesişim testi.
- “Time” varlık isimlerin kesişim testi.
- “Money” varlık isimlerin kesişim testi.
- “Percentage” varlık isimlerin kesişim testi.
- “Unknown” varlık isimlerin kesişim testi.
- Her bir sorgu için derleme gönderildiğinde varlık isim tablo kesişim modeli kullanılarak üretilen sorgu-belge kesişim sonuçları çıkarıldı.
- Çıkartılan tüm bu sonuçlar içerisinden, sorgunun ilgili olduğu bilinen belgeler için anma/duyarlık ve f-ölçüsu skor değerleri hesaplandı.

4.6 VİKM Test Sonuçlar

Yukardaki senaryoya göre varlık isimler kesişim modeli için gerçekleştirilen test sonuçları için elde edilen tablo aşağıda sunulmuştur. Testler esnasında varlık terimlerden son ekler atılmış ve gövdeleme yapılmış.

VİKM	d	a	f
Tüm isimlerin	0.23	0.98	0.37
Person	0.63	0.85	**0.72**
Location	0.25	0.97	0.39
Organization	0.39	0.88	0.54
Date	0.22	0.98	0.35
Time	0.29	0.57	0.38
Money	0.46	0.60	0.52

Percentage	0.77	0.55	0.64
Unknown	0.87	0.91	0.89

Tablo 4.3 Varlık İsimler Kesişim Sonuç

Tablo 4.3 varlık isimlerin kesişim sonucuna baktığımızda, kesişim kontrolünde en etkin varlık isim “kişi” olduğu anlaşılmıştır. Unknown varlık isimlerinin belli bir kurala daynamadıkları için şuan göz ardı ediyoruz.

4.7 Varlık İsimlerin Normalizasyon Modeli Test Senaryosu

Bu kısımda Varlık isimler üzerinde normalization işlemi yaparak benzelik fonksiyonu kullanrak, haber konu bulma sistem başarısında ve temel yöntem üzerinde etkileri hedeflenmiştir. Bu hedefi gerçekleştirmek amacıyla yürütülen testlerde aşağıdaki test senaryoları gerçekleştirildi:

- Varlık isimlerin kesişim testi gerçekleştirdikten sonra, varlık isimler üzerinde benzerlik fonksiyonu tanımlandı (varlık isim benzerlik fonksiyonu) ve bir başka yaklaşımla varlık isimlerin benzerliğini konu algılamada başarısı test edildi.
- Varlık isim benzerlik fonksiyonun baktığımızda bir basit mantıkla haberlerde geçen varlık isimler üzerinde normalizasyon yaparak elde ediliyor, aşağıda formül olarak verilmiştir.
- Varlık isimler Benzerlik Fonksiyonu *(BF)* :

BF = (D1 kesişim D2) / (D1 Birleşim D2)

Eşitlik 4.1 Varlık İsimler Kesişim Oranı

- Yukardaki formüle baktığımızda benzerlik fonksiyonun kısaca açıklarsak, *D1* ve *D2* haberde geçen varlık isimlerin kesişimi bolu *D1* ve *D2* haberlerde varlık isimlerin birleşimi, aslında bolu kısmı normalizasyon işlemini gerçekleştirmektedir.

- Tüm Varlık İsimlerin çıkardıktan sonra, Varlık İsimler benzerlik Modelin haber konu algılamada başarısını belirlemek için aşağıdaki adımlar gerçekleştirildi.
    - Geçen aşamada varlık isimler üzerinde yapılan ön işlemler, varlık isimlerin çıkarımı, son eklerin atılması ve sonunda gövdeme işlemi yine yapıldı.
    - Tüm haberler için varlık isim tabloları oluşturuldu.
    - İlgililik değerlendirmesi yapılmış olan 5.872 belge test ve eğitim setine ayrıldı.
    - Çıkartılan varlık isim tablolarına göre iki farklı benzerlik tespiti gerçekleştirildi.
    - Birinci adımda sorgulama yapıldığında haberlerde geçen bütün varlık isimlerin benzerlik tespitinde, skor hesaplamada kullanıldı.
    - İkinci adımda ise haberlerde geçen varlık isimlerin türüne göre tek tek ayrı benzerlik tespiti yapıldı.
    - İkinci adımdaki kullanılan varlık isimler türlerin sırayla şu şekilde “Person”, ”Location”, ”Organization”, ”Date”, “Time”, “Money”, “Percentage”, “Unknown” yine adlandırılmıştır.
    - Bu aşamada varlık isimlerin benzerlik tespitin haber konu bulmada gerçekleşen test adımların aşağıda sunulmuş:
        - Tüm varlık isimlerin benzerlik tespiti.
        - Sadece “Person” varlık isimlerin benzerlik tespiti.
        - “Location” varlık isimlerin benzerlik tespiti.
        - “Organization” varlık isimlerin benzerlik tespiti.
        - “Date” varlık isimlerin benzerlik tespiti.
        - “Time” varlık isimlerin benzerlik tespiti.
        - “Money” varlık isimlerin benzerlik tespiti.
        - “Percentage” varlık isimlerin benzerlik tespiti.
        - “Unknown” varlık isimlerin benzerlik tespiti.
        - Her bir sorgu için derleme gönderildiğinde varlık isim benzerlik modeli kullanılarak üretilen sorgu-belge eşleşme benzerlik skorları belirlendi.
        - Belirlenen tüm bu skor değerleri içerisinden, sorgunun ilgili olduğu bilinen belgeler için üretilen benzerlik skor değerleri

çıkarıldı ve ilgili sorgu-belge eşleşmeleri için ortalama skor değeri başlangıç eşiği olarak kabul edildi.

- Bu başlangıç eşiğine göre tüm haberler için geçen varlık isim benzerlik fonksiyonu türlerine göre tek tek anma/duyarlık ve f-ölçüsü değerleri hesaplandı.
- Eşik değeri belirli oranda azaltılıp-artırılarak anma/duyarlık değerleri her bir eşik için tekrar hesaplandı, anma ve duyarlığın birlikte en yüksek oldukları (ya da birbirlerine en yakın oldukları) değer her Varlık İsim benzerlik fonksiyonu için kesin eşik değeri olarak kabul edildi.
- Eğitim seti üzerinde eşik değerlerin çıkardıktan sonra test işlemi test seti üzerinde gerçekleştirildi.

4.8 VİNM Test Sonuçlar

Yukarıdaki senaryoya göre varlık isimler benzerlik tespit modeli için gerçekleştirilen test sonuçları için elde edilen tablo aşağıda sunulmuştur. Testler esnasında varlık terimlerden son ekler atılmış ve gövdeleme yapılmış.

VİNM	d	a	f	e
Bütün isimler	0.98	0.13	0.22	0.02
Person	0.95	0.32	0.48	0.01
Location	0.59	0.62	**0.61**	0.01
Organization	0.78	0.24	0.37	0.01
Date	0.41	0.37	0.39	0.01
Time	0.36	0.29	0.32	0.01
Money	0.96	0.30	0.46	0.01
Percentage	0.62	0.11	0.18	0.01

Unknown	0.94	0.59	0.72	0.01

Tablo 4.4 Varlık İsimler Benzerlik Fonksiyonu

Varlık isimlerin normalizasyon modelinde ise “konum” varlık isimi en etkin isim olarak anlaşılmıştır.

4.9 Varlık İsimlerin Birleşim Kesişim Model (VİBKM) Test Senaryosu

Varlık isimlerin haber konu bulma sistem başarısında ve temel yöntem üzerinde etkileri hedeflenmiştir. Bu hedefi gerçekleştirmek amacıyla yürütülen testlerde aşağıdaki test senaryoları gerçekleştirildi:

- Varlık isimlerin kesişim ve benzerlik testi gerçekleştirdikten sonra, bu aşamada haberlerde etiketlenen varlık isimleri kullanarak, varlık isimlerin farklı kombinasyonları oluşturarak kesişme modeli haber konu bulmada başarı testleri gerçekleştirilecektir.
- Literatürdeki çalışmalara baktığımızda varlık isimlerin kombinasyonlarının en yaygın ve çok kullanılanı “mekan” ve “zaman” varlık isimlerin birleşimiyle oluşmaktadır, bu varlık isimlerin birleşimi daha önce tarıtışıldığı gibi haberlerde bir olayı anlatıyor, o yüzden bu varlık isimleri kullanan yöntemlere Olay Modeli yöntemi söylenmektedir.
- Bu senaryoda Olay Modeli varlık isim birleşimlerin testi yapıldı, ve buna ek olarak daha önce geçmişteki çalışmalarda olmayan farklı varlık isimler birleşim kesişmeleri konu algılamada başarı testleri gerçekleştirildi.
- Testler ilgililik değerlendirmesi yapılmış olan belgeler (5.872 adet) üzerinden yapıldı.
- Testlerde daha önce planlanan olay kelime kombinasyonları çıkardıktan sonra, her haber için olay kelime birleşim tabloları oluşturuldu ve olay kesişim Modelin haber konu algılamada başarısını belirlemek için aşağıdaki adımlar gerçekleştirildi.
    - Her haber için tüm etiketlenen varlık isimler çıkartıldı.

- Varlık isimler üzerinde son çekim ekler atıldıktan sonra Zemberek kütüphanesin kullanarak gövdeleme gerçekleştirildi.
- Daha önceki gibi varlık isimler şu şekilde adalndırıldı,"Person", "Location", "Organization", "Date", "Time", "Money", "Percentage", "Unknown".
- Haberler için varlık isimleri kullanarak olay kelime birleşim tabloları oluşturuldu, oluşturulan varlık isim birleşimleri aşağıda sunulmuş.
- "Location" ve "Time" birleşimi.
- "Location" ve "Date" birleşimi.
- "Location", "Time" ve "Date" birleşimi.
- "Person" ve "Time" birleşimi.
- "Person" ve "Date" birleşimi.
- "Person", "Time" ve "Date" birleşimi.
- "Organization" ve "Time" birleşimi.
- "Organization" ve "Date" birleşimi.
- "Organization" , "Time" ve "Date" birleşimi.
- "Person" ve "Location" birleşimi.
- "Person", "Location" ve "Time" birleşimi.
- "Person", "Location" ve "Date" birleşimi.
- "Organization" ve "Location" birleşimi.
- "Organization" , "Location" ve "Time" birleşimi.
- "Organization" , "Location" ve "Date" birleşimi.
- "Person", "Organization" ve "Time" birleşimi.
- "Person" , "Organizaiton" ve "Date" birleşimi.
- "Person" , "Organization" ve "Location" birleşimi.

- Her bir sorgu için derleme gönderildiğinde varlık isim birleşim tablo kesişim modeli kullanılarak üretilen sorgu-belge kesişim sonuçları çıkarıldı.
- Tüm bu sonuçlar içerisinden, sorgunun ilgili olduğu bilinen belgeler için anma/duyarlık ve f-ölçü skor değerleri hesaplandı.

4.10 VİBKM Test Sonuçlar

Yukarıdaki senaryoya göre varlık isimlerin birleşim kesişim modeli için gerçekleştirilen test sonuçları için elde edilen tablo aşağıda sunulmuştur. Testler esnasında varlık terimlerden son ekler atılmış ve gövdeleme yapılmıştır.

VİBKM	d	a	f
Location-Time	0.69	0.17	0.27
Location-Date	0.44	0.80	0.56
Location-Time-Date	0.86	0.14	0.24
Person-Time	0.97	0.16	0.28
Person-Date	0.84	0.52	0.64
Person-Time-Date	0.95	0.14	0.25
Organizaiton-Time	0.79	0.16	0.26
Organization-Date	0.66	0.54	0.59
Organization-Time-Date	0.94	0.13	0.22
Person-Location	0.72	0.80	**0.76**
Person-Location-Time	0.94	0.15	0.25
Person-Location-Date	0.90	0.50	0.64
Organization-Location	0.57	0.81	0.66
Organization-Location-Time	0.98	0.16	0.27
Organization-Location-Date	0.76	0.51	0.61
Person-Organization-Time	0.96	0.15	0.25
Person-Organization-Date	0.97	0.41	0.56
Person-Organization-Location	0.89	0.64	0.74

Tablo 4.5 Varlık İsimler Birleşim Kesişim

Varlık isimlerin birleşim kesimi için, “kişi” ve “konum” birleşimi en etkin ve belirgin kombinasyon olarka ortaya çıkmış.

4.11 VUM OR VİV test senaryosu

Bu kısımda Vektör Üzay Modelin tüm kelimeler ve Varlık İsimler birleişimi haber konu bulma sistem başarısında ve temel yöntem üzerinde etkileri hedeflenmiştir. Bu hedefi gerçekleştirmek amacıyla yürütülen testlerde aşağıdaki test senaryoları gerçekleştirildi:

- Bu aşamada daha önce gerçekleşen iki yöntemin OR mantıksal birleşimiyle başarı testleri hedeflenmektedir, bunun için daha önce gerçekleşen vektör uzay modeli ve varlık isimler vektör test yöntemin OR mantıksal kombinasyonuyla haber konu bulmada başarısı test edilmiştir.
- Bu iki yöntemin kombinasyonları iki adımda gerçekleştirilmiştir.
- Birinci Adımda vektör uzay modelin, tüm varlık isimler üzerinde kombinasyon testi.
- İkinci adımda ise vektör uzay modelin, varlık isimlerin türünden oluşan tek tek vektörlerle kombinasyon testi.
- Bu iki yöntemin haber konu algılamada başarısını belirlemek için daha önce her yöntem için yapılan adımlar yine aynı şekilde gerçekleştirilmiştir.

4.12 VUM OR VİV test sonuçlar

Yukarıdaki kombinasyon senaryoya göre vektör uzay modelin ve varlık isimler vektör modelin OR birleşim kombinasyon modeli için gerçekleştirilen test sonuçları için elde edilen tablo aşağıda sunulmuştur.

VUM OR VİV	d	a	f	e
VUM OR tüm isimler	0.61	0.58	0.59	0.05
VUM OR Person	0.65	0.62	**0.64**	0.05
VUM OR Location	0.62	0.59	0.60	0.05
VUM OR Organization	0.65	0.62	**0.64**	0.05
VUM OR Date	0.63	0.60	0.62	0.05
VUM OR Time	0.61	0.58	0.59	0.05
VUM OR Money	0.61	0.58	0.59	0.05
VUM OR Percentage	0.61	0.58	0.59	0.05
VUM OR Uknown	0.62	0.58	0.59	0.05

Tablo 4.6 VUM OR VİVM

Vektör Uzay Modelin, Varlık isimler vaktör modelle birleşimlerine baktığımızda "kurum" ve "kişi" vektörlerin, temel yötem üzerinde daha etkili olduğunu anlaşılmaktadır.

4.13 VUM OR VİKM test senaryosu

Varlık isimlerin kesişim modeli, VUM birleşimi ile haber konu bulma sistem başarısında ve temel yöntem üzerinde etkileri hedeflenmiştir. Bu hedefi gerçekleştirmek amacıyla yürütülen testlerde aşağıdaki test senaryoları gerçekleştirildi:

- Daha önce sonuçlanan yöntemlerin kombinasyonlarının devamında vektör uzay model ve varlık isimler modelin OR kombinasyonları hedeflenmektedir.
- Vektör uzay modeli ve varlık isimler modelin OR kombinasyonlarının haber konu bulmada başarısını test edilmesi bu yöntemleri daha önce yapıldığında tüm ön işlemler aynı şekilde yapıldı.

- Yötemlerin ön işlemleri yapıldıktan sonra OR mantıksal birleşim testleri gerçekleştirildi.

4.14 VUM OR VİKM test sonuçlar

Yukarıdaki kombinasyon senaryoya göre vektör uzay modelin ve varlık isimler kesişim modelin OR birleşim kombinasyon modeli için gerçekleştirilen test sonuçları için elde edilen tablo aşağıda sunulmuştur.

VUM OR VİKM	d	a	f	e
VUM OR tüm isimler	0.90	0.87	**0.89**	0.05
VUM OR Person	0.75	0.72	0.73	0.05
VUM OR Location	0.83	0.80	**0.81**	0.05
VUM OR Organization	0.77	0.74	**0.76**	0.05
VUM OR Date	0.75	0.71	0.73	0.05
VUM OR Time	0.71	0.68	0.69	0.05
VUM OR Money	0.71	0.68	0.69	0.05
VUM OR Percentage	0.71	0.68	0.69	0.05
VUM OR Unknown	0.71	0.68	0.69	0.05

Tablo 4.7 VUM OR VİKM

Tablo 4.7 baktığımızda, Varlık isimlerin kesişim kontrol yöntemin VUM üzerinde, aşırı derecede etkili olduğunu goruyoruz. Tüm varlıkların kontrolü daha etkin olduğu anlaşılmıştır ve teker teker kontrollerde “konum” ve “kurum” en belirgin varlıklar olarak orta çıkmış.

4.15 VUM OR VİBKM test senaryosu

Varlık isimlerin birleşim kesişim modeli, VUM'le birleşimi haber konu bulma sistem başarısında ve temel yöntem üzerinde etkileri hedeflenmiştir. Bu hedefi gerçekleştirmek amacıyla yürütülen testlerde aşağıdaki test senaryoları gerçekleştirildi:

- Bu aşamada vektör uzay modeli ve varlık isimler birleşim kesişim modelin OR kombinasyonu hedeflenmektedir
- Vektör uzay modeli ve varlık isimler birleşim kesişim modelin OR kombinasyonlarının haber konu bulmada başarısını test edilmesini gerçekleştirmede bu yöntemlerde daha önce yapıldığında tüm ön işlemler aynı şekilde yapılıyor

4.16 VUM OR VİBKM test sonuçlar

Bu iki yöntemin kombinasyon senaryosuna göre vektör uzay modelin ve varlık isimler farklı birleşim kesişim modelin OR kombinasyon modeli için gerçekleştirilen test sonuçları için elde edilen tablo aşağıda sunulmuştur.

VUM OR VİBKM	d	a	f	e
VUM OR Location-Time	0.71	0.68	0.69	0.05
VUM OR Location-Date	0.73	0.70	0.71	0.05
VUM OR Location-Time-Date	0.71	0.68	0.69	0.05
VUM OR Person-Time	0.71	0.68	0.70	0.05
VUM OR Person-Date	0.71	0.68	0.70	0.05
VUM OR Person-Time-Date	0.71	0.68	0.69	0.05
VUM OR Organization-Time	0.71	0.68	0.69	0.05
VUM OR Organization-Date	0.71	0.68	0.69	0.05
VUM OR Organization-Time-Date	0.71	0.68	0.69	0.05

VUM OR Person-Location	0.73	0.70	**0.72**	0.05
VUM OR Person-Location-Time	0.71	0.68	0.69	0.05
VUM OR Person-Location-Date	0.71	0.68	0.70	0.05
VUM OR Organization-Location	0.71	0.68	0.69	0.05
VUM OR Organization-Location-Time	0.71	0.68	0.69	0.05
VUM OR Organization-Location-Date	0.71	0.69	0.70	0.05
VUM OR Person-Oragnization-Time	0.71	0.68	0.69	0.05
VUM OR Person-Organization-Date	0.71	0.68	0.69	0.05
VUM OR Person-Organization-Location	0.71	0.69	0.70	0.05

Tablo 4.8 VUM OR VİBKM

Varlık İsimlerin birleşim kesişim yöntemin, VUM üzerinde etkisin incelediğimizde, “kişi” ve “konum” varlıklar en etkin olarak anlaşılmıştır. Ayrıca literatür kısmında bahs ettiğimiz gibi bu varlıkların (kişi ve konum) birleşimi olay modelinde, yeni olayaları tanımlamada en temel varlıklar olarak isimlendirilmektedir.

5 Tartışma ve Değerlendirme

Yöntem	P	R	F	T
VUM	0.61	0.58	0.59	0.05
VİVM	0.65	0.56	0.60	0.02
Person	0.77	0.39	0.51	0.01
Location	0.51	0.51	0.51	0.25
Organization	0.46	0.46	0.46	0.01
Date	0.33	0.36	0.34	0.2
Time	0.34	0.33	0.34	0.25
Money	0.49	0.28	0.36	0.01

Percentage	0.76	0.22	0.34	0.01
Unknown	0.89	0.55	0.67	0.01
VİKM	0.23	0.98	0.37	0.0
Person	0.63	0.85	**0.72**	0.0
Location	0.25	0.97	0.39	0.0
Organization	0.39	0.88	0.54	0.0
Date	0.22	0.98	0.35	0.0
Time	0.29	0.57	0.38	0.0
Money	0.46	0.60	0.52	0.0
Percentage	0.77	0.55	0.64	0.0
Unknown	0.87	0.91	0.89	0.0
VİNM	0.98	0.13	0.22	0.02
Person	0.95	0.32	0.48	0.01
Location	0.59	0.62	0.61	0.01
Organization	0.78	0.24	0.37	0.01
Date	0.41	0.37	0.39	0.01
Time	0.36	0.29	0.32	0.01
Money	0.96	0.30	0.46	0.01
Percentage	0.62	0.11	0.18	0.01
Unknown	0.94	0.59	0.72	0.01
VİBKM	-	-	-	-
Location Time	0.69	0.17	0.27	0.0
Location Date	0.44	0.80	0.56	0.0
Location Time Date	0.86	0.14	0.24	0.0
Person Time	0.97	0.16	0.28	0.0
Person Date	0.84	0.52	0.64	0.0
Person Time Date	0.95	0.14	0.25	0.0
Organization Time	0.79	0.16	0.26	0.0
Organization Date	0.66	0.54	0.59	0.0
Organization Time Date	0.94	0.13	0.22	0.0
Person Location	0.72	0.80	0.76	0.0
Person Location Time	0.94	0.15	0.25	0.0
Person Location Date	0.90	0.50	0.64	0.0

Organization Location	0.57	0.81	0.66	0.0
Organization Location Time	0.98	0.16	0.27	0.0
Organization Location Date	0.76	0.51	0.61	0.0
Person Organization Time	0.96	0.15	0.25	0.0
Person Organization Date	0.97	0.41	0.56	0.0
Person Organization Location	**0.89**	0.64	**0.74**	0.0
VUM *OR VİVM*	0.61	0.58	0.59	0.05
Person	0.65	0.62	0.64	0.05
Location	0.62	0.59	0.60	0.05
Organization	0.65	0.62	0.64	0.05
Date	0.63	0.60	0.62	0.05
Time	0.61	0.58	0.59	0.05
Money	0.61	0.58	0.59	0.05
Percentage	0.61	0.58	0.59	0.05
Unknown	0.62	0.58	0.59	0.05
VUM *OR* VİKM	0.90	0.87	0.89	0.05
Person	0.75	0.72	0.73	0.05
Location	0.83	0.80	0.81	0.05
Organization	0.77	0.74	0.76	0.05
Date	0.75	0.71	0.73	0.05
Time	0.71	0.68	0.69	0.05
Money	0.71	0.68	0.69	0.05
Percentage	0.71	0.68	0.69	0.05
Unknown	0.71	0.68	0.69	0.05
VUM *OR VİBKM*	-	-	-	-
Location Time	0.71	0.68	0.69	0.05
Location Date	0.73	0.70	0.71	0.05
Location Time Date	0.71	0.68	0.69	0.05
Person Time	0.71	0.68	0.70	0.05
Person Date	0.71	0.68	0.70	0.05
Person Time Date	0.71	0.68	0.69	0.05
Organization Time	0.71	0.68	0.69	0.05
Organization Date	0.71	0.68	0.69	0.05

Organization Time Date	0.71	0.68	0.69	0.05
Person Location	0.73	0.70	0.72	0.05
Person Location Time	0.71	0.68	0.69	0.05
Person Location Date	0.71	0.68	0.70	0.05
Organization Location	0.71	0.68	0.69	0.05
Organization Location Time	0.71	0.68	0.69	0.05
Organization Location Date	0.71	0.69	0.70	0.05
Person Organization Time	0.71	0.68	0.69	0.05
Person Organization Date	0.71	0.68	0.69	0.05
Person Organization Location	0.71	0.69	0.70	0.05

Tablo 5.1 Tüm Sonuçlar

Tezde TDT programında tanımlı Hikaye Bağlantı Algılama görevinin Türkçe bir derlem üzerinde farklı benzerlik fonksiyonları ve bunların kombinasyonların kullanarak başarımının test edilmesi ve optimum anma/duyarlık değerlerini sağlayacak kombinasyonun bulunması hedeflemiştir. Ayrıca, haberlerde geçen Varlık İsimlerin hangisi ve ya hangi kombinasyonun Hikaye Bağlantı Algılama görevinde daha başarılı sonuçlar ürettiğini belirlemeside amaçlanmaktadır. Bu kapsamda sonuçları incelediğimzde alttaki sonuçlar göze çarpmaktadır

- 0.72 ve 0.89'lık F-ölçüm değeri ile sırayla "Person" ve "Unknown" varlık isimler kesişme yönteminde, en başarılı yöntem elde edilmiştir.
- 0.74 ve 0,76'lık F-ölçüm değeri ile sırayla "Person-Organization-Location" ve "Person-Location" kombinasyonları, varlık isimler birleşim kesişme yönteminde en başarılı kesişme yöntemi olarak elde edildi.
- Vektör uzay modeli ve varlık isimler kesişim yöntemlerin, OR kombinasyon modelin sonuçlarına baktığımızda, .089'lık f-ölçüsü bütün varlık isimlerin kesişime yönteminde elde edilmiştir ve en başarılı yöntem olarak Kabul edildi. Devamında tek tek varlık isimlerin kombinasyon sonuçları 0.73, 0.73, 0.76 ve 0.81'lik f-ölçü değerleri sırayla "Date", "Person", "Organization" ve "Location" varlık isimlerinden elde edilmiştir ve başarı üzerinde etkileri ilginç vaziyetde dikkat çekicidir.

- Vektör uzay modeli ve varlık isimlerin kombinasyon kesişme yöntemine baktığımızda 0.71 ve 0.72'li f-ölçü değeri sırayla "Location-Date" ve "Person-Location" birleşim kesişmelerine ait olan en başarılı sonuç olarak elde edildi.

Yukarıdaki elde edilen sonuçları göz önünde bulundurduğumuzda, vektör uzayı yönteminden daha başarılı sonuçlar elde edilmiştir. Dolayısıyla varlık isimlerin türkçe üzerinde, haber konu bulmada ve haber benzerliklerin tespitinde önemli ve system başarısı üzerinde inanılmaz dercede etkili olduğunu gormekteyiz. Örneğin vektör uzay yöntemin, varlık isimler kesişme yöntemiyle, OR kombinasyonu iyi şekilde system başarısını etkilemiştir ve gerçekleştirilen testlerde diğer yöntemler ile karşılaştırıldığında en başarılı yöntem olarak öne çıkmaktadır.

6 Sonuç

TDT içerisinde tanımlı SLD görevinin gerçekleştirilmesinde Türkçe bir derlem üzerinde vektör uzayı ve bağımsız varlık isimlerinin haber benzerlik ve farklılıklarının belirlenmesinde kullanılması gibi yöntemler tesleri gerçekleştirildi. Derlemdeki haberler, "Kişi", "Kurum", "Konum", "Tarih", "Zaman", "Para" ve "Yüzde" olmak üzere yedi farklı varlık isimi ile etiketlenmiştir. Derlemdeki haber ilişkilerini tanımlada en etkin varlık isimlerinin "Kişi", "konum" ve "Kurum" olduğu anlaşılmıştır. Ayrıca varlık isimlerin birleşim analizlerinde ise "kişi-konum" birleşimi en etkin olarak belirlenmiştir. Varlık isimlerin farklı test senaryoların Vektör Uzay Modeli üzerinde analizlerine baktığımızda, tüm varlık isimlerin kesişim kontrolü inanılmaz derecede VUM başarısın etkilemesi anlaşılmaktadır.
Tezin odaklandığı konularla ilgili olarak Türkçe derlemler üzerindeki çalışmalar oldukça sınırlıdır. Doküman-doküman eşleşmeleri üzerinde farklı ve yenilikçi pek çok yöntemin uygulanacak olması nedeniyle elde edilecek sonuçların Türkçe bilgi erişim sistemleri ile ilgili akademik çalışmalara katkısının büyük olacağını öngörmekteyiz.

Sonuç olarak, önerilen yöntemler mükemmel bir bilgi erişim sistemine ulaşmak için ihtiyaç duyulan "ilgili belgelerin tamamına erişim sağlama ilgisizleri ise dışarda bırakma" prensibine bizleri yaklaştırdı ve bazı özgün yöntemler önermektedir.

Tez kapsamında yöntemlerin başarım testlerinin uygulanabilmesi amacıyla BilCOl-2005 derlemi üzerinde varlık isimlerinin etiketlenmesi gerçekleştirildi ve bu alanda çalışan akademisyenler için önemli bir derlem ortaya konuldu.

Tezde önerilen yöntemler Türkçe bir derlem üzerinde bugüne kadar bu alanda denenmemiş ve bu açıdan bakıldığında, ortaya çıkan olan sonuçların özellikle akademik alanda faaliyet gösteren araştırmacılar, medya takip çalışmaları tetikleyici olacağını ve bu alandaki rekabeti artıracağını düşünmekteyiz.
Son olarak ise; haberler içerisindeki varlık isimleri gerek teker teker gerekse birlikte kullanılarak, iki haberin farklı konularda olup olmadıklarını belirlemede ne kadar etkili olduklarına açık pozisyon olarak gelecekteki çalışmalarda incelenebilme potansiyeline sahip olduğu ortaya çıktı.

7 Tez Kapsamında Ek Testler

Tez çalışmaları kapsamında farklı sınıflama ve kümeleme algoritmaların haber konu bulmada başarım testleri gerçekleştirildi. Testler esnasında daha önceki testlerdeki gibi Bilcol-2005 derlemin ilgilik değerlendirmesi yapılan kısmı kullanıldı.

7.1. **Sınıflama (Classification)**

Herhangi bir verinin niteliğinin diğer verilerin niteliklerine göre kıyaslama yapılarak belirlenmesi işlemine sınıflama adı verilir. Sınıflamada eğitim seti, yani önceden değerleri ve sınıfları bilinen veriler, geliştirilen yöntemin testi veya analizinin başarısının tespit edilebilmesini sağlar.

7.1.1 K-en Yakın Komşu (KYK) (k-Nearest Neighbor (k-NN))

Sınıflandırma problemlerinde etkin olarak kullanılmakta olan yöntemlerdendir. Belirlenen “k” değerine göre verilen sorgunun “k” tane en yakın (benzer) very setinde yer alan vektörleri tespit ederek sınıflama işlemini gerçekleştirir. Bu yöntemin doğru uygulanabilmesi için iyi bir eğitim kümesi oluşturulması şarttır. Eğitim kümesi bu yöntemin başarısındaki en önemli faktördür. Uygulanmasının kolay olması nedeniyle k-NN sınıflama problemlerinde sık sık kullanılmaktadır.

k-NN algoritmasının çalışma mantığı her bir sorgunun ayrı ayrı hesaplanmasını gerektirdiğinden dolayı bu yöntemin hesaplanma maliyeti çok yüksektir. En yakın komşuluk bağıntısına dayandığı için vektör uzayında ifade edilen terimlerin birbirine olan uzaklıkları Manhattan yöntemi, Euclidean yöntemi ve ya Minkowski yöntemi yardımıyla hesaplanır. Euclidean yötemi literatürde en çok kullanılan yöntemdir. k-NN eşitlikleri içinde gösterilen “k” değeri komşuluk derecesini, “x” değeri kategorinin vektörünü “y” değeriyse sonucun vektörünü temsil etmektedir. Verilen sorgudaki terim benzerlik oranı “1” değerine en yakın olan kategoriye eklenir.

$$\sqrt{\sum_{i=1}^{k}(x_i - y_i)^2}$$

Eşitlik 2.3 Euclidean Yöntemi

7.1.2 K-NN ve VUM Senaryosu

Tez çalışmaları kapsamında, Vektör Uzayı Modeli ve K-en Yakın Komşu yöntemlerin kullanılarak farklı sınıflama yaklaşımları test edildi ve testlerle ilgili çeşitli aşamalar bildiri olarak uluslararası konferansa gönderildi.

- Testler ilgililik değerlendirmesi yapılmış olan belgeler (5.872 adet) üzerinden yapılacaktır.
- Öncelikle her bir konu eğitim ve test belgeleri olmak üzere iki kısma ayrılacaktır.
    - Her bir konuda var olan belge sayısının üçte biri eğitim (1.931 adet), üçte ikisi de test belgesi (3.941 adet) olarak kabul edilecektir.
    - Eğitim belgeleri seçilirken, tarih sırasına göre derlemdeki ilk N belge seçilecektir, kalan belgeler test belgesi olarak kullanılacaktır.
- Derlemde bulunan ve ilgililik değerlendirmesi yapılmış olan 3.941 belge test belgesi olarak kullanılacaktır.
    - Test belgesi olarak belirlenen bu 3.941 belge test için sorgu olarak kabul edilecektir.

- KYK yönteminde her bir sorgu eğitim kümesine gönderildikten sonra elde edilen K belge eğitim belgesi olarak kullanılmaktadır.
- Elde edilen K belge, küme tabanı ağırlıklandırmaları çıkardıktan sonra, en ağır küme sonuç olarak kabul edilmektedir.
- VUM yönteminde ise, her bir sorgu eğitim kümesine gönderilecek ve elde edilen en yakın belgenin kümesi sonuç olarak seçilecektir.
- KYK ve VUM yöntemleri, OR mantıksal operatörü ile birleştirilecektir.
- Tüm sorgular tamamlandıktan sonra sistemin başarısı belirlenecektir.

7.1.3 KYK ve VUM Modeli Sınıflama Sonuçları

K	**KYK**	**VUM**	**KYK OR VUM**
3	92,98	93,23	94,90
4	92,42	93,23	95,20
5	92,02	93,23	95,50
6	91,41	93,23	95,70
7	90,71	93,23	95,81
8	90,30	93,23	**97,86**
9	90,35	93,23	95,81

Tablo 7.1 KYK ve VUM Sınıflama Sonuç

7.2 Kümeleme (Clustering)

Veri kümelemede yöntemlerinde, sınıflama yöntemlerinde farklı olarak gözetimsiz olarak kümeleme işlemi gerçekleşiyor, yani küme sayısı ve türü bilinmeyen verilerdeki yapılan kümelemeler için kullanılır. Kümeleme işlemi, örnekler kümesi üzerinde belirli demetleme algoritmaları kullanılarak gruplama işlemini gerçekleştiriyor. Bu aradaki uzaklığı belirlerken k-NN algoritmasında olduğu gibi

Euclidean ve benzeri uzaklık hesaplama yöntemleri kullanılır. Burada etiketleme işlemi kullanılmadığı için kategoriler oluşacak olan kümelerin sayısı kadar olacaktır. Dolayısiyla gözetimsiz kümeleme etiketlenmemiş ve sınıflanmamış veriler üzerindeki bilinmeyen yapının tespiti için kullanılır.

7.2.1 Bulanık Mantık (Fuzzy Logic)

Daha önceki bahsedilen yöntemler sorgulara her zaman iki seçenekli yani “evet”, veya “hayir” şeklinde bir kümeye ait olup olmadığını kesin olarak cevaplar vermek üzere tasarlanmıştı. Bulanık mantıkta ise verilen bir sorgunun farklı kümelere farklı uzaklık değerleriyle ait olması gibi cevaplar verilmektedir. Dolayısiyla Bulanık Mantıkta bir sorgu kesin bir kümeye ait olup olmadığı ilk adımda tespit edilmemektedir ve algoritma yeterince iterasyon yaptıktan sonra sorgunu farklı kümelere uzaklık değerini belirlemektedir.

7.2.2 K-means ve C-means 200 Kümeleme Senaryosu

Tez çalışmaları kapsamında, K-means ve C-means algoritmaların kullanarak farklı kümeleme testleri gerçekleştirildi. Bu yöntemlerin tek tek ve OR amntıksal birleşim testleri, haber verilerinde seçilen dört farklı set (200, 500, 1000 ve 1500 belge) üzerinde aynı test senaryosun kullanarak gerçekleştirildi.

- Testler ilgililik değerlendirmesi yapılmış olan 5.872 adet belgenin bir kısmı üzerinde yapılacaktır.
- Öncelikle ilgilikleri değerlendirilen belgelerden (toplam 80 küme) 10 küme seçilecektir.
- Seçilen 10 kümenin her birinden sadece ilk 20 belge seçilecektir (toplam 200 belge).
- Kümeleme yöntemlerin kullanarak 200 belge üzerine kümeleme testleri gerçekleştirilecektir.
    - K-means yönteminde K değeri 10 olarak alınıp test gerçekleştirilecek.
    - C-means yönteminde ise K değeri 10 girdiğimizde test sonuçları gerçekleştirilecek,

- Tüm testler sonunda seçilen her 10 küme için ayrı ayrı anma ve duyarlık değerleri belirlenecektir.

7.2.3 K-means ve C-means 200 belge (10 konu) Test Sonuçlar

		k-m			c-m			OR	
Konu	p	r	f	p	r	f	p	r	f
1	0.99	0.99	0.99	0.97	0.98	0.97	0.99	0.99	0.99
2	0.22	0.99	0.36	0.60	0.99	0.74	0.61	0.98	0.75
3	0.99	0.99	0.99	0.40	0.40	0.40	0.98	0.98	0.98
4	0.01	0.01	0.01	0.40	0.40	0.40	0.41	0.41	0.40
5	0.99	0.99	0.99	0.98	0.99	0.98	0.99	0.99	0.99
6	0.99	0.35	0.51	0.99	0.45	0.61	0.98	46.0	0.60
7	0.01	0.01	0.01	0.98	0.97	0.97	0.99	0.96	0.97
8	0.99	0.99	0.99	0.40	0.40	0.40	0.99	0.99	0.99
9	0.99	0.15	0.26	0.99	0.40	0.56	0.98	0.41	0.56
10	0.99	0.70	0.82	0.98	0.98	0.98	0.97	0.99	0.98
m	0.71	0.61	0.59	0.76	0.69	0.70	0.88	0.81	**0.82**

Tablo 7.2 k-means, c-means, 200 belge sonuç

7.2.4 K-means ve C-means 500 belge Test Senaryo ve Sonuçlar

Yapılan test daha önceki 200 belge üzerinde yapılan test senaryosu gibi aynen tekrarlaniyo, sadece 10 kümenin her birinden sadece ilk 50 belge seçilecektir ve toplam 500 belge üzerinde test gerçekleştirildi.

	k-m			c-m			k-m OR c-m		
Konu	P	R	F	P	R	F	P	R	F

1	0.99	0.52	0.68	0.71	0.96	0.81	0.71	0.96	0.81
2	0.99	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.98
3	0.99	0.86	0.92	0.30	0.30	0.30	0.99	0.86	0.92
4	0.99	0.50	0.66	0.93	0.96	0.95	0.93	0.96	0.94
5	0.41	0.26	0.31	0.67	0.98	0.79	0.67	0.98	0.70
6	0.99	0.90	0.94	0.97	0.96	0.95	0.99	0.90	0.95
7	0.21	0.21	0.21	0.30	0.30	0.30	0.30	0.30	0.30
8	0.97	0.80	0.87	0.94	0.98	0.96	0.94	0.98	0.95
9	0.38	0.98	0.54	0.97	0.96	0.96	0.97	0.96	0.96
10	0.99	0.22	0.36	0.30	0.30	0.30	0.99	0.22	0.36
m	0.79	0.62	0.64	0.70	0.76	0.73	0.77	0.81	**0.78**

Tablo 7.3 k-means, c-means, 500 belge sonuç

7.2.5 K-means ve C-means 1000 belge Test Senaryosu ve sonuçlar

Burdaki test senaryosu yine önceki 200 belge test senaryosu gibi yalnız bu defa 20 küme seçilip her kümeden ilk 50 belge, toplam 1000 belge seçilmiştir.

	k-m			c-m			k-m OR c-m		
Konu	p	r	f	p	r	f	p	r	f
1	0.89	0.76	0.77	0.76	0.76	0.76	0.89	0.76	0.77
2	0.92	0.54	0.68	0.48	0.48	0.48	0.92	0.64	0.68
3	0.38	0.79	0.43	0.78	0.79	0.78	0.78	0.89	0.78
4	0.89	0.32	0.42	0.20	0.20	0.20	0.89	0.32	0.42
5	0.89	0.79	0.80	0.66	0.64	0.65	0.89	0.79	0.80
6	0.80	0.38	0.48	0.75	0.72	0.73	0.75	0.72	0.73

7	0.89	0.78	0.80	0.20	0.20	0.20	0.89	0.88	0.80
8	0.89	0.32	0.45	0.53	0.79	0.63	0.53	0.79	0.63
9	0.31	0.31	0.31	0.79	0.22	0.34	0.79	0.32	0.34
10	0.84	0.52	0.61	0.49	0.79	0.60	0.84	0.52	0.61
11	0.89	0.82	0.79	0.31	0.78	0.44	0.89	0.82	0.79
12	0.31	0.31	0.31	0.20	0.20	0.20	0.31	0.31	0.31
13	0.89	0.82	0.84	0.48	0.78	0.59	0.89	0.92	0.84
14	0.96	0.34	0.50	0.79	0.22	0.34	0.96	0.34	0.50
15	0.31	0.31	0.31	0.56	0.26	0.35	0.56	0.36	0.35
16	0.89	0.60	0.70	0.79	0.22	0.34	0.89	0.70	0.70
17	0.89	0.76	0.77	0.20	0.20	0.20	0.89	0.76	0.77
18	0.89	0.78	0.77	0.79	0.79	0.79	0.79	0.81	0.79
19	0.31	0.31	0.31	0.31	0.72	0.43	0.31	0.72	0.43
20	0.89	0.64	0.68	0.78	0.79	0.78	0.78	0.80	0.78
m	0.74	0.56	0.58	0.54	0.52	0.49	0.77	0.65	**0.68**

Tablo 7.4 k-means, c-means, 1000 belge sonuç

7.2.6 K-means ve C-means 1500 Test Senaryosu ve sonuçlar

En son testde yine aynı senaryo seçilen 20 küme üzerinde yalnız her kümeden ilk 75 belge, toplam 1500 belge seçilmiştir.

	k-m			*c-m*			*k-m* OR *c-m*		
Konu	p	r	f	p	r	f	p	r	f
1	0.79	0.75	0.76	0.70	0.79	0.72	0.79	0.75	0.76
2	0.74	0.72	0.73	0.44	0.88	0.52	0.74	0.72	0.73
3	0.58	0.79	0.66	0.89	0.80	0.82	0.89	0.80	0.82
4	0.79	0.51	0.56	0.76	0.69	0.61	0.76	0.69	0.61

5	0.79	0.79	0.79	0.89	0.87	0.88	0.89	0.87	0.88
6	0.74	0.74	0.74	0.40	0.40	0.40	0.74	0.74	0.74
7	0.79	0.71	0.73	0.71	0.76	0.72	0.79	0.71	0.73
8	0.65	0.60	0.62	0.50	0.60	0.51	0.62	0.60	0.62
9	0.51	0.51	0.51	0.20	0.20	0.20	0.51	0.51	0.51
10	0.57	0.33	0.38	0.79	0.70	0.72	0.75	0.70	0.72
11	0.76	0.25	0.32	0.66	0.73	0.67	0.66	0.73	0.67
12	0.79	0.74	0.75	0.25	0.72	0.29	0.74	0.74	0.75
13	0.79	0.41	0.50	0.40	0.40	0.40	0.79	0.41	0.50
14	0.21	0.21	0.21	0.60	0.60	0.60	0.60	0.60	0.60
15	0.79	0.77	0.77	0.65	0.79	0.68	0.73	0.77	0.77
16	0.70	0.70	0.70	0.50	0.50	0.50	0.70	0.70	0.70
17	0.70	0.33	0.40	0.79	0.79	0.79	0.76	0.79	0.79
18	0.78	0.74	0.74	0.74	0.79	0.75	0.71	0.79	0.75
19	0.21	0.21	0.21	0.50	0.50	0.50	0.50	0.50	0.50
20	0.21	0.21	0.21	0.79	0.79	0.79	0.79	0.79	0.79
$\underline{m}$	0.64	0.52	0.56	0.58	0.62	0.59	0.68	0.75	**0.69**

Tablo 7.5 k-means, c-means, 1500 belge sonuç

KAYNAKLAR

[1] Allan, J. Lavrenko, V. ve Swan, R. Explorations Within Topic Tracking and Detection, Topic Detection and Tracking: Event-based Information Organization, J. Allan, ed. Kluwer Academic Publishers, pp. 197-224. **2002**

[2] Yang, Y. Carbonell, J. Brown, R. Lafferty, J., Pierce, T., & Ault, T. Multi-strategy learning for topic detection and tracking. In J. Allan (Ed.), Topic Detection and Tracking: Event-based Information Organization pp. 85-114. Norwell, MA: Kluwer Academic Publishers,**2002.**

[3] Güven Köse, Yaşar Tonta, Aydın Can Polatkan, and Hamid Ahmadlouei, Story Link Detection in Turkish Corpus, The IEEE/WIC/ACM International Conference on Web Intelligence, Nov. 17-20,Atlanta GA USA, **2013**.

[4] Güven Köse & Hamid Ahmadlouei, VSM-Based Improved Supervised News Topic Detection Model in Turkish New, 4th International Symposium on Information Management in a Changing World, Semtember 4-6, Limerick Institute of Technology, **2013**.

[5] Köse, G. Konu Algılama ve İzleme Programında Olay Modeli. Yayımlanmamış Yüksek Lisans Tezi. Başkent Üniversitesi Fen Bilimleri Enstitüsü, Ankara, **2004**.

[6] Meadow, C. T. Text Information Retrieval Systems. San Diego: Academic Press, **1992**.

[7] Tonta, Y. Bitirim. Y. ve Sever. H. Türkçe Arama Motorlarında Performans Degerlendirme. Ankara: Total Bilişim Ltd. Sti. xvi, 152 s. (ISBN 975 92923-0-0), **2002**.

[8] Allan, J, Introduction to Topic Detection and Tracking, Topic Detection and Tracking: Event-based Information Organization, J. Allan, ed. Kluwer Academic Publishers, pp. 1-16, **2002**.

[9] Lavrenko, V. ve Croft, W. B. Relevance based language models. In Proceedings of the 24th Annual international ACM SIGIR Conference on Research and Development in information Retrieval (New Orleans, Louisiana, United States). SIGIR '01. ACM, New York, NY, 120-127, **2001.**

[10] Lavrenko, V. Allan, J., DeGuzman, E., LaFlamme, D., Pollard, V. ve Thomas, S. Relevance Models for Topic Detection and Tracking, Proceedings of the Human Language Technology Conference (HLT), 104-110, 2002.

[11] Schultz, J.M. ve Liberman, M, Towards a universal dictionary for multi-language IR applications, Topic Detection and Tracking: Event-based Information Organization, J. Allan, ed. Kluwer Academic Publishers, pp. 225-239, **2002**.

[12] Shah, C. Croft, W. B. ve Jensen, D. "Representing Documents with Named Entities for Story Link Detection (SLD)," a poster presentation at the ACM Fifteenth Conference on Information and Knowledge Management (CIKM) 2006, Arlington VA, November 6-11, **2006**.

[13] Kumaran, G. ve Allan, J. Text classification and named entities for new event detection. In Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (ACM SIGIR’04) pp. 297-304. Sheffield, UK: ACM, **2004**.

[14] Can, F. Kocberber, S., Baglioglu, O., Kardas, S., Ocalan, H. C. ve Uyar, E. . "New event detection and topic tracking in Turkish." Journal of the American Society for Information Science and Technology. Vol. 61, No. 4, 2010, pp. 802-819, **2010**.

[15] Thompson, K.C. ve Callan, J. "Query expansion using random walk models." In Proceedings of the Fourteenth International Conference on Information and Knowledge Management (CIKM'05). ACM, **2005**.

[16] Qiu, J. ve Liao, L.J. Add temporal information to dependency structure language model for topic detection and tracking. Machine Learning and Cybernetics. 1575 – 1580, **2008**.

[17] Mori, M. Miura, T. ve Shioya, I. Topic Detection and Tracking for News Web Pages. WI '06 Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence. 338-342, **2006**.

[18] Jin, Y. Myaeng, S.H., Lee, M., Oh, H. Ve Jang, M. Effective Use of Place Information for Event Tracking. Lecture Notes in Computer Science. 3689, 410-422, **2005**.

[19] Chirag S. And Koji E. Improving Document Representation for Story Link Detection by Modeling Term Topicality. IPSJ Online Transactions, **2009**.

[20] Kumaran, G. ve Allan, J. Using names and topics for new event detection. In Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing (HLT '05), **2005**.

[21] Makkonen, J. Ahonen-myka, H. ve Salmenkivi, M. Applying Semantic Classes in Event Detection and Tracking. Proc. International Conference on Natural Language Processing (ICON'02). 175-183, **2002**.

[22] Xianshu Z. And Tim O, Finding News Story Chains Based on Multidimensional Event Profile.OAIR2013, Lisbon, Portugal, **2013**.

[23] Dalkılıç, F.E. Gelişli, S. ve Diri, B. "Türkçe Kural Tabanlı Varlık İsmi Tanıma", 18. Sinyal İşleme ve Uygulama Kurultayı, Diyarbakır, (22-24 Nisan) **2010**.

[24] Bayraktar, Ö. ve Taşkaya-Temizel, T. Person Name Extraction From Turkish Financial News Text Using Local Grammar Based Approach. In Proceedings of the International Symposium on Computer and Information Science (ISCIS), **2008**.

[25] Küçük, D. ve Yazıcı, A. Named Entity Recognition Experiments on Turkish Texts. In Proceedings of the International Conference on Flexible Query Answering Systems. Roskilde, Denmark. T. Andreasen et al. (Eds.): FQAS 2009, LNAI 5822, pp. 524-535, **2009**.

[26] Küçük, D. ve Yazıcı, A. A Hybrid Named Entity Recognizer for Turkish with Applications to Different Text Genres. In Proceedings of the 25th International Symposium on Computer and Information Sciences (ISCIS). London, UK. E. Gelenbe et al. (Eds.): Computer and Information Sciences, LNEE 62, pp. 113-116, **2010**.

[27] Uyar, E. (2009). Near-Duplicate News Detection Using Named Entities. Master Thesis, Computer Engineering Department, Bilkent University. 27 Şubat 2011 tarihinde http://www.cs.bilkent.edu.tr/~canf/bilir_web/theses/erkanUyarThesis.pdf adresinden erişildi, **2009**.

[28] Xianshu Z. And Tim O. Finding News Story Chains Based on Multidimensional Event Profile.OAIR2013, Lisbon, Portugal, **2013**.

[29] Letian Wang and Fang Li, Story Link Detection Based on Event Words, Springer-Verlag Berlin Heidelberg **2011**.

[30] Akin, A. and Akin, MD. Zemberek, an open source NLP framework for Turkish Languages OnlineAvailableat:https://code.google.com/p/zemberek/, **2007**.

[31] Hatcher E. Lucene in Action (1st ed.). Manning Publications.p. http://lucene.apache.org/.**2004.**

[32] Setzer, A. Temporal information in newswire articles: An annotation scheme and corpus study, University of Sheffield, UK, **2001**.

ÇİZELGELER

ÖZGEÇMİŞ

Kimlik Bilgileri

Adı Soyadı : Hamid AHMADLOUEI

Doğum Yeri : IRAN

Medeni Hali : Bekâr

E-posta : hamid2026@gmail.com

Adresi : Hacettepe Üniversitesi, Bilgisayar Mühendisliği

Eğitim

Lise : SHAHED2, URMİA, IRAN

Lisans : IRAN AZAD UNIVERSITY

Yüksek Lisans : Hacettepe Üniversitesi

Yabancı Dil ve Düzeyi

İngilizce: iyi

Türkçe: çok iyi

İş Deneyimi

2012-2014 Hacettepe Üniversitesi- Bilgisayar Mühendisliği Bölümü araştırma görevlisi

Deneyim Alanları

Bilgi Erişim Sistemleri, Makina Öğrenme, Veri Madenciliği, Yapay Zekâ, Etmen Tabanlı Modelleme

Tezden Üretilmiş Projeler ve Bütçesi

Bu araştırma Türkiye Bilimsel ve Teknolojik Araştırma Kurumu (TÜBİTAK) taradınfan desteklenmiştir (Proje numarası 111K030).

Tezden Üretilmiş Yayınlar

Güven Köse & **Hamid Ahmadlouei**, VSM-Based Improved Supervised News Topic Detection Model in Turkish New, 4th International Symposium on Information Management in a Changing World, Semtember 4-6, Limerick Institute of Technology, **2013**.

Güven Köse, Yaşar Tonta, Aydın Can Polatkan, and **Hamid Ahmadlouei**, Story Link Detection in Turkish Corpus, The IEEE/WIC/ACM International Conference on Web Intelligence, Nov. 17-20,Atlanta GA USA, **2013**.

Tezden Üretilmiş Tebliği ve/veya Poster Sunumu ile Katıldığı Toplantılar