Düzce Üniversitesi Bilim ve Teknoloji Dergisi

Araştırma Makalesi

İngilizce Dokümanlarda Tema ve Alt Kavramlar Tespit Modeli

Sena ÖGTELİK [a,*], Metin TURAN [b]

[a] *Bilgisayar Mühendisliği Bölümü, Fen Bilimleri Enstitüsü, İstanbul Ticaret Üniversitesi, İstanbul, TÜRKİYE*
[b] *Bilgisayar Mühendisliği Bölümü, Mühendislik Fakültesi, İstanbul Ticaret Üniversitesi, İstanbul, TÜRKİYE*
** Sorumlu yazarın e-posta adresi: senaogtelik92@gmail.com.tr*

ÖZET

Bu makalede dokümanlarda tema ve alt kavram tespiti konusunda bir model önerilmiş ve deneysel bulgular değerlendirilmiştir. Dokümanlarda tema ve alt kavramların tespiti için kullanılabilecek anlamlı sözcüklerin belirlenmesi amacıyla Helmholtz prensibi temelli Gestalt teorisi kullanılmıştır. Bu sözcüklerin girdi olduğu bir Yapay Sinir Ağı (YSA) modeli oluşturulmuş, eğitim dokümanları (140 adet) ile bu ağ eğitilmiştir. Eğitim ve sınama doküman veri seti spor ve eğitim temalarında olup, toplam 14 alt kavram seçilmiştir. YSA'nın çıktısı tema ve alt-kavram bilgilerini vermektedir. 70 adet sınama dokümanı ile farklı sayıda (5, 10, 20) anlamlı kelime seçilerek deneyler yapılmış, başarı oranının konularda yaklaşık olarak %95, alt kavramlarda ise %80 olduğu gözlemlenmiştir.

Anahtar Kelimeler: *Doğal Dil İşleme, Yapay Sinir Ağları, Helmholtz Prensibi, Sınıflandırma*.

Topic and Sub-Topic Detection Model in English Documents

ABSTRACT

In this article, a model of topic and sub topic detection is proposed in the documents and experimental findings are evaluated. The Gestalt theory based on the Helmholtz principle was used in the documents to determine the meaningful words that could be used to determine concepts and sub topic. An Artificial Neural Network (ANN) model was established in which these words were entered, and this network was trained with number of 140 training documents. The training and testing document dataset is about the sports and training topics and 14 sub-topics have been selected. The output of ANN gives the topic and sub topic information. Experiments were executed with 70 test documents with different numbers of (5, 10, 20) words. It was observed that the success rate was approximately 95% in the topic and 80% in the sub topic.

Keywords: *Natural Language Processing, Artificial Neural Networks, Helmholtz Principle, Classification.*

Geliş: 01/05/2018, Düzeltme: 23/05/2018, Kabul: 04/07/2018

I. GİRİŞ

Gelişen teknolojik ortam ile birlikte üretilen bilgilere ulaşım da daha kolay bir hal almıştır. Gelişen bu bilgi ağı içinde aradığımız bilgiyi, aradığımız nitelikleri içeren bilgiler içinde arama yapmamız durumunda istenilen bilgiye daha kısa sürede ulaşmamıza yardımcı olacaktır. Bu amaçtan yola çıkarak dokümanları sınıflandırma yaparak bizim için önemli olmayan tema veya alt kavramlara ait dokümanları eleyerek istediğimiz dokümana daha kolay bir şekilde ulaşmamız sağlanır. Y. H. Li ve A. K. Jain [1], 1998 yılında doküman sınıflandırması için dört farklı yöntem üzerinde çalışmışlardır. Çalıştıkları sınıflandırma yöntemleri, Naive Bayes sınıflandırıcı, en yakın komşu sınıflandırıcı, karar ağaçları ve bir alt uzay yöntemidir. Yaptıkları çalışmayı yedi farklı sınıf içeren Yahoo haber gruplarına (iş, eğlence, sağlık, uluslararası, politika, spor ve teknoloji) uygulamışlardır. Yaptıkları çalışma sonucu sınıflandırmaların doğruluğunda %83 başarı oranı yakalamışlardır. Metin sınıflamada en doğru alt kavramları seçmek, başarılı sonuçlar elde etmek için etkili bir yöntemdir. Yu ve Liddy [2] bu alt kavramların seçimi ile ilgili birçok yöntem hakkında çalışma yapmışlardır. 2002 yılında Ron Bekkerman, Ran El-Yaniv, Naftali Tishby ve Yoad Winter'in [3] yapmış oldukları sınıflandırma çalışmasında kullanılan veri kümelerinden biri 20 haber grubu, diğer iki veri kümesi de 21578 tane Reuters verisi ile WebKB verilerinden oluşmaktaydı. Çalışma sonucunda Reuters verilerinden %92.2, haber grubu verisinden %88.6 başarı sağlanmıştır.

2005 yılında da Man Lan ve arkadaşları [4] doküman konusu belirleme üzerine çalışmıştır. Yaptıkları deneyler sonucunda toplamda ortalama %86 ile %92 oranı aralığında başarı elde etmişlerdir. Ayrıca, bilim insanları dokümanlar üzerinde farklı amaçlar için sınıflandırma çalışmaları uygulamışlardır. Bu amaçlara örnek olarak; yazar tanıma ve metnin yazarının cinsiyetini belirleme [5,6], e-posta sınıflandırma [7], topoloji ile metin sınıflandırma[8], duygu analizi ile metin sınıflandırma [9] verilebilir.

Bu çalışma, belirli tema ve bu temalara ait alt kavramları içeren bir dokümanın, hangi sınıfa ait olduğunun tahmin edilmesi üzerinedir. Sınama amaçlı seçilmiş dokümanlardan elde edilen anlamlı kelimeler ile bir Yapay Sinir Ağları (YSA) eğitilmiş, daha sonra verilen hedef dokümanın tema ve alt kavramları tespit edilmeye çalışılmıştır. Amacımız, özelleşmiş problemler üzerinde yüksek başarı oranları elde edecek bir modelin geliştirilmesidir.

Makalenin; ikinci bölümünde tema ve alt kavram tespiti için kullanılan yöntemler anlatılmış, üçüncü bölümde çalışmaya ait kullanılan veri sertlerinden bahsedilmiş, son bölümünde ise sonuçlar açıklanmıştır.

II. YÖNTEM

A. ÖN İŞLEME

Her doğal dil işleme projesinde olduğu üzere öncelikle dokümanlar ön işlemden geçirilir. D Tanasa, B Trousse [10] ve V.Chitraa, Dr. Antony Selvdoss Davamani [11] çalışmalarında ön işleme tekniklerine yer vermişlerdir. Ön işlemede, dokümanlar metin içinde anlamı olmayan birçok kelime içermektedir ve sonlama kelimeleri (stop words) olarak adlandırılan bu kelimeler kullanıldıkları cümle içerisinde bir anlam ifade etmemektedir. Bu kelimeler çıkarıldıklarında da anlamsal bir kayba yol

açmazlar, fakat kelime frekansına göre çalışan bir modelde sonuçlara olumsuz etkileri olur. Her dilin kendine özgü sonlama kelimeleri vardır ve İngilizce de sık olarak kullanılan; bağlaçlar, imleçler, sayılar, kalıplaşmış kısaltmalar gibi içerikten bağımsız kelimeler sonlama kelimelere örnek olarak verilebilirler: “about”, “across”, ”all”, “and”, “before” ,”but”, “enough”, “everywhere”, ”over” “except”, “from”, “go”, “himself”, “make”. Bu kelimeler ayırt edici özelliğe sahip olmadıklarından, önişleme sırasında dokümanlarımızdan ayrıştırılır. Önişleme yapılırken dokümanlardaki boşluk, rakam ve noktalama işaretleri gibi anlam ifade etmeyen karakterler de elenir. Ayrıca kelimeleri düzenli bir dizilime getirebilmek amacıyla büyük harf, küçük harf uyumluluğu sağlanır.

Daha sonraki ön işleme adımı, yapım- çekim- iyelik ekleri alan kelimelerin tek bir forma dönüştürülmesidir. Frekans hesaplanmasında büyük bir önem arz eden bu çalışma için kök bulma (stemming) algoritmaları kullanılır. Bu amaçla İngilizce dili için geliştirilmiş algoritmalardan en yaygın olarak bilinen PorterStemmer kök bulma algoritması projede uygulanmıştır.

B. ANAHTAR KELİMELERİN SEÇİMİ

Bu aşamada doküman bazında her bir kelimenin, hem sınama aşaması hem de eğitim aşaması için anlam değerlerinin (meaning value) bulunmasıdır. Balinsky A., Balinsky H. ve Simske S. [12,13] yaptıkları veri madenciliği araştırmalarında Helmholtz prensibini kullanarak, anlamlı özelliklerin tespiti üzerinde çalışmışlardır. Bu çalışmada kelimelerin anlam değerlerinin bulunmasında Helmholtz prensibi uygulanmıştır.

Helmholtz Prensibi, Gestalt insan algı teorisini kullanır. Bu teori metin madenciliğinde her bir kelime için; dokümanın paragrafında, kelimenin m kez geçmesinin olası olup olmadığının belirlenmesinde kullanılmıştır. Bu teoriye dayanan anlam değeri bazı formüllerle hesaplanır. Bu formüller şu şekildedir:

$$YAS(k,P,D)=\binom{K}{m}\frac{1}{N^{m-1}} \tag{1}$$

$$Anlam(k,P,D)=-\frac{1}{m}\log YAS(k,P,D) \tag{2}$$

$$N=\frac{|D|}{|P|} \tag{3}$$

$$\log YAS(k,P,D)=\log\left(\binom{K}{m}\frac{1}{N^{m-1}}\right) \tag{4}$$

$$\binom{K}{m}=\frac{K!}{m!(K-m)!} \tag{5}$$

Helmholtz Prensibine göre kullanılan bu formüllerde;

D: Dokümanın paragrafları
P: Paragrafta yer alan cümleler, anlamı taşımaktadır.
Çalışmada dokümanları paragraflarına ayırarak, paragraf tabanlı bir çalışma yaptığımız için bu formülü kullanırken D'yi doküman, P 'yi de dokümanda yer alan paragraf olarak ele aldık. Yaptığımız çalışma için formülde yer alan diğer terimlerin de açıklaması aşağıdaki gibidir:

k:İşlem yapılan kelime
P:Veri kümesinde yer alan paragraf sayısı
D:Veri kümesinde yer alan dokümanlar
m:Hesapladığımız kelimenin toplam kaç tane paragrafta geçtiği
K:Hesapladığımız kelimenin doküman da toplam kaç kez geçtiğinin sayısı
N:Tüm veri kümesinin boyunun (Tüm veri kümesindeki toplam kelime sayısı), bir dokümanın boyuna (dokümandaki toplam kelime sayısı) bölümü

Yukarıda belirtilen formüllerden yola çıkılarak anlam değeri (meaning value) ve Yanlış Alarm Sayısı (YAS) değeri hesaplamaları yapılmıştır.

Burada anlam değerine ulaşmak için hesapladığımız Yanlış Alarm Sayısı (YAS değeri), anlam değeri ile ters orantılıdır. Başka bir ifadeyle YAS değeri ne kadar düşük çıkarsa; seçilen özelliğin (k) o sınıftaki dokümanlar (D) için anlam değeri o kadar yüksek demektir.

Anlam değerinin yüksek olması özelliklerin etkin ve verimli olduğunun göstergesidir. Burada en iyi özelliğin seçilmesi için aşağıdaki yaklaşım kullanılmıştır.

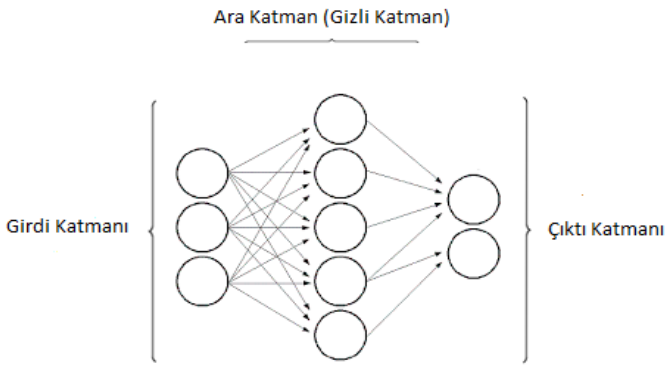
$$Anlam(k.D) = Enbüyük\left(Anlam\left(k, P, D\right)\right) \quad (6)$$

Bu formülde; çıkan en yüksek anlam değeri, hesaplamaya alınan özelliğin anlam değeri olarak kabul edilmiş ve bu metoda "EÖAS" (Eğitimli En Büyük Anlamsal Özellik Seçimi) [14] adı verilmiştir.

C.YSA'NIN EĞİTİMİ

YSA günümüzde mevcut olan birçok makine öğrenmesi yönteminden sadece bir tanesidir. İnsan beyninin öğrenme sistemini taklit ederek geliştirilmiştir ve bu sayede keşfedebilme, üretebilme, olaylar arası ilişki kurabilme ve karar verebilme gibi özelliklerin yapılması sağlanmış ve gelişim göstermeye de devam etmektedir.

Çalışmada kullanılan YSA, geri beslemeli bir yapıda olup bir adet girdi katmanı ve bir adet de çıktı katmanından oluşmaktadır (Şekil 1). Girdi sayıları, sistemde yapılan sınamalara göre değer almaktadır. Çalışmada 5, 10 ve 20 farklı anahtar kelime seçimi ile denemeler yapılmıştır. Dolayısıyla ağ girdi sayıları da sırasıyla 5, 10 ve 20 olmaktadır. Çıktı sayıları tema ve alt kavram tespitine bağlıdır. Tema tespiti için oluşturulan ağda çıktı sayısı 2, alt kavram ağı için çıktı sayısı 14 olmaktadır.



Şekil 1. *YSA yapısı*

Sistem çalışmasının tutarlı olabilmesi için, YSA'nın eğitilmesinde kullanılan eğitim veri kümesinden elde edilmiş anahtar kelime sayısı ile sınama yapılmak üzere sınama veri kümesinden elde edilmiş anahtar kelime sayısı her deneme için eşit alınmıştır. Eğitim girdi verileri, anahtar kelimelerin hangi tema ve alt kavrama ait olduğu bilgisi verilerek oluşturulmaktadır. Sınama girdi verileri, tema ve alt kavram tespiti yapacağımız sınama verilerinin anahtar kelimelerinden oluşur. 140 tane eğitim verisi kullanılarak geri beslemeli, bir adet girdi katmanı, bir adet ara katman ve bir adet çıktı katmanından oluşan ağ yapısı, tema ve alt kavramları verilerek, sırasıyla 5, 10 ve 20 farklı anahtar kelime ile ayrı ayrı eğitilmiştir. Eğitilen bu ağ, aynı sayıda anahtar kelime ile 70 adet sınama verisi için değerlendirilmiştir. Ağın tema ve alt kavramlar için sınıflandırma başarı oranları, farklı anahtar kelime sayıları ile ayrı ayrı karşılaştırılmış olup, sonuçlar 3. bölümde verilmiştir.

III. VERİ KÜMELERİ

Veri kümesi oluşturmak amacıyla, 2-gram olasılıklarına dayalı benzerlik modeli kullanan Python uygulaması yazılarak, aykırı dokümanların (farklı alt kavram içeren) ayrıştırılması sağlanmıştır. Böylece veri kümesi kaynaklı sistem hataları en aza indirilerek, deneysel ortam sonuçlarının doğruluğundan emin olunmuştur.

Çalışmada sınama veri kümesi olarak kullanılmak üzere her bir alt kavram için 5'er tane doküman olmak üzere toplamda 70 adet (Tablo 1), YSA sisteminin eğitilmesi için yine her bir alt kavramdan 10'ar tane olmak üzere toplam 140 adet (Tablo 2) doküman seçilmiştir.

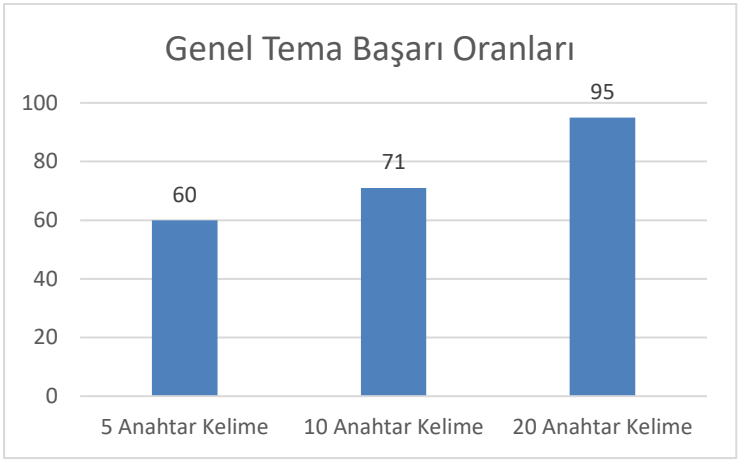
Tablo 1. *Sınama veri kümesi*

	Doküman Sayısı		**Doküman Sayısı**
Archery	5	Homeschool	5
Badminton	5	Preschool	5
Bicycling	5	Scholarship	5
Football	5	ELearning	5
Tennis	5	Language	5
Gymnastics	5	ElementarySchool	5
Swimming	5	Mathematics	5
Spor	**35**	**Eğitim**	**35**

Tablo 2. *Eğitim veri kümesi*

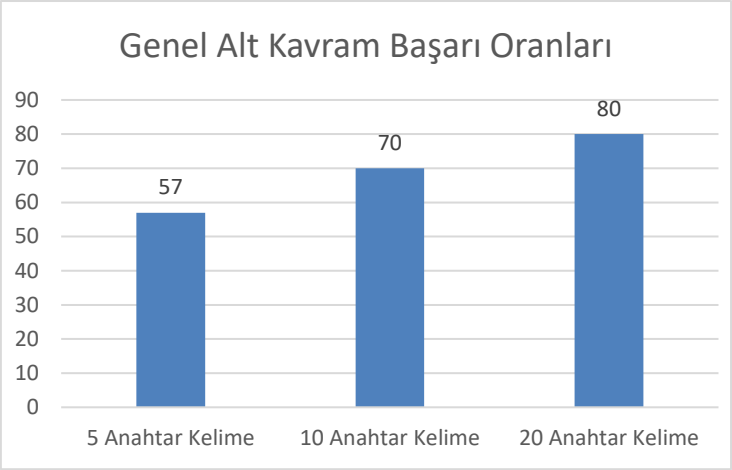
	Doküman Sayısı		**Doküman Sayısı**
Archery	10	Homeschool	10
Badminton	10	Preschool	10
Bicycling	10	Scholarship	10
Football	10	ELearning	10
Tennis	10	Language	10
Gymnastics	10	ElementarySchool	10
Swimming	10	Mathematics	10
Spor	**70**	**Eğitim**	**70**

Eğitilmiş olan YSA sistemine sınama veri kümemizdeki dokümanları soktuktan sonra çıkan sonuçlar ve başarı oranları incelenmiştir.



Şekil 2*. Genel tema başarı oranları*

Şekil 2'deki grafikte sınama veri kümesindeki dokümanların temalarının sisteme sokulan anahtar kelime sayısına bağlı olarak genel başarı oranları verilmiştir. Çıkan sonuçlara göre tema tespit başarı oranı seçilen anahtar kelime sayısıyla bağlantılı olarak arttığı gözlemlenmektedir. Bu sonuçlara göre tema tespiti için en yüksek başarı %95 oran ile 20 anahtar kelime seçimi yapılarak elde edilmiştir.



Şekil 3. *Genel alt kavram başarı oranları*

Şekil 3'deki grafik incelendiğinde ise, genel alt kavram tespitindeki başarı oranının tema tespitindekiyle benzer şekilde seçilen anahtar kelime sayısının artışına bağlı olarak arttığı gözlemlenmektedir. Alt kavram tespitindeki en yüksek başarı oranı yine 20 anahtar kelime seçiminde %80 olarak elde edilmiştir.

Tablo 3. *Hata Matrisi(Confusion Matrix) Sonuçları*

		Tahmin Sınıf		
		Olumsuz	**Olumlu**	**Toplam**
Gerçek Sınıf	**-1**	**32**	**3**	**35**
	1	**0**	**35**	**35**
	Toplam	**32**	**38**	**70**

Hata matrisinde DN (Doğru Negatif) değeri, spor verileri ele alındığında spor verileri haricinde seçilmemesi gereken verilerden kaç tanesinin doğru seçildiği sayısıdır. YP (Yanlış Pozitif) değeri, spor verileri haricinde seçilmemesi gereken verilerden kaç tanesinin seçilmediği sayısıdır. YN (Yanlış Negatif) değeri, seçilmesi gereken verilerden kaçının seçilmediği sayısıdır. DP (Doğru Pozitif) değeri, spor verilerinin seçilmesi gerekenlerden kaçının doğru seçildiğidir.

Doğruluk değeri aşağıdaki gibi hesaplanır (Formül 7):

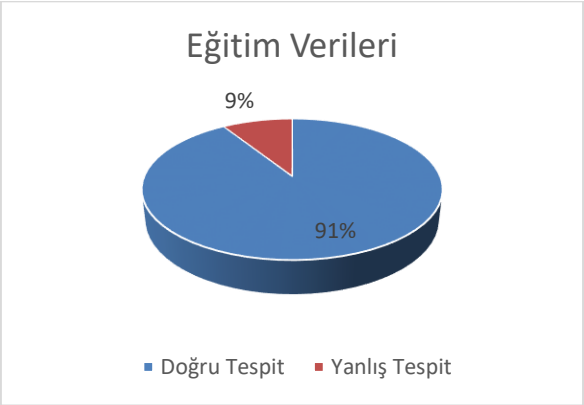
$$(DN + DP)/Toplam = (32+35)/70 = 0,95 \quad (7)$$

Hata oranı değeri ise aşağıdaki gibi hesaplanır (Formül 8):

$$(1 - Doğruluk) \quad = 1-0,95 = 0,05 \quad (8)$$

Tablo 3 incelendiğinde eğitim ve spor verilerinin temalarının bulunması sırasında %5' lik bir hata payı ile %95 başarı elde edildiği görülmektedir.

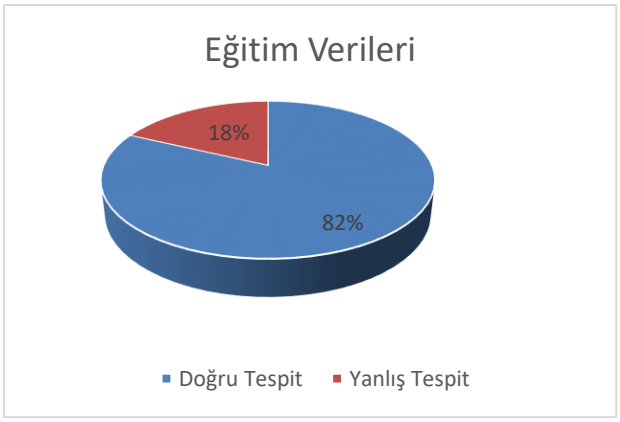
20 anahtar kelime ile toplamda 70 tane eğitim ve spor sınama verilerinin başarı oranlarına bakıldığında tema algılamada başarı oranı eğitim verileri için %91 (Şekil 4) iken, spor verileri için %100 (Şekil 5) olarak bulunmuştur. Sistemin tema algılamada genel ortalama başarı oranının %95 olduğu görülmektedir. Metin T. ve Coskun Ş. [15] yaptıkları benzer çalışmada tema algılamada başarı oranını ortalama %83 olarak bulmuşlardır.



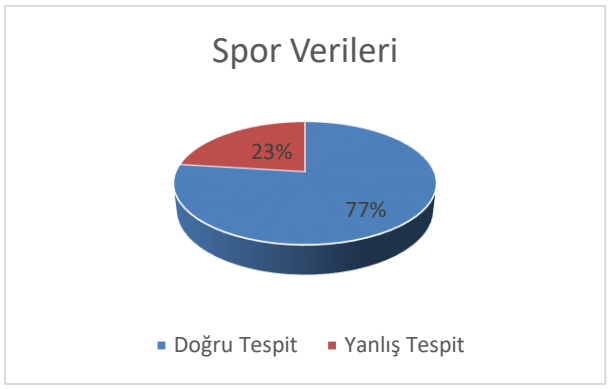
Şekil 4. *Tema tespitinde 20 anahtar kelime seçimi için eğitim verilerinin başarı oranı*

Şekil 5. *Tema tespitinde 20 anahtar kelime seçimi için spor verilerinin başarı oranı*

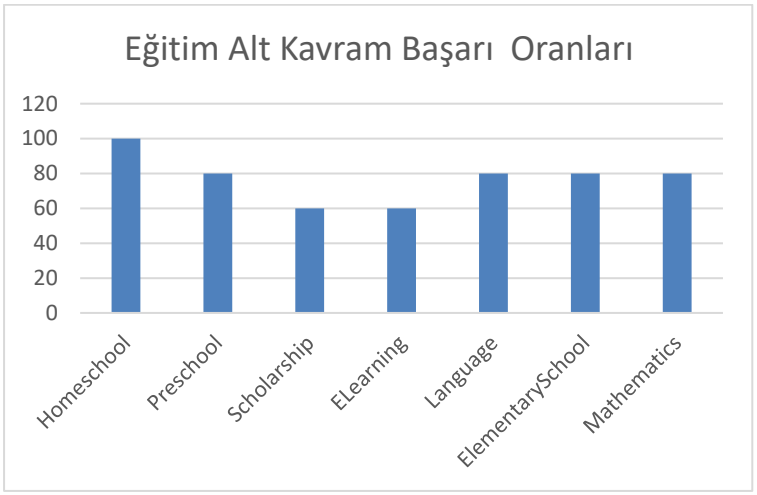
20 anahtar kelime ile alt kavram tespitinde başarı oranı %80 çıkmıştır. Eğitim verileri için %82 (Şekil 6) başarı oranı sağlanırken, spor verilerinde %77 (Şekil 7) başarı sağlanmıştır. Metin T. ve Coskun Ş.'nin [15] yaptıkları çalışmada ise alt-kavramlar için eğitim verilerin ortalama başarı oranının %61, spor verilerinde de % 72 olarak bulunduğu görülmektedir.



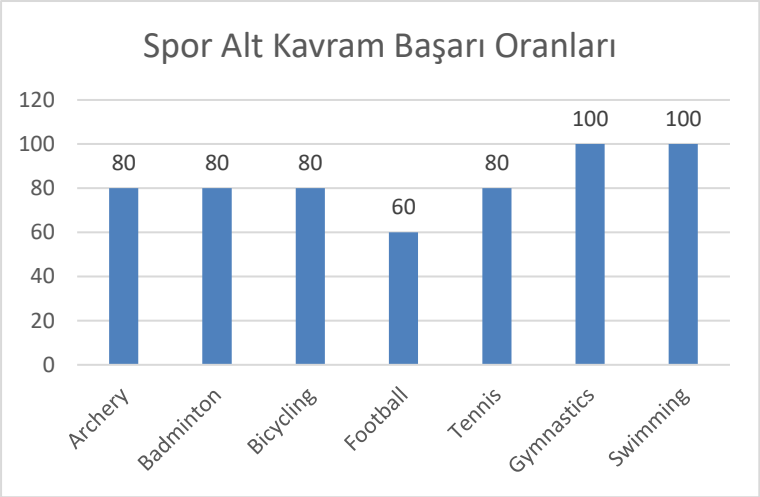
Şekil 6. *Alt kavram tespitinde 20 anahtar kelime seçimi için spor verilerinin başarı oranı*



Şekil 7. *Alt kavram tespitinde 20 anahtar kelime seçimi için spor verilerinin başarı oranı*



Şekil 8. *Eğitim alt kavramlarda başarı oranları*



Şekil 9. *Spor alt kavramlarda başarı oranları*

Her bir alt kavram tespiti için 5'er adet sınama dokümanı kullanılmıştır. Şekil 8 ve 9 incelendiğinde; Eğitim temasında Homeschool alt kavramı için %100 oranında, spor temasında Gymnastics ve Swimming alt kavramları için %100 oranında başarı elde edilmiştir.

Sistemin başarı oranı alt kavramlar için %80, tema için %95 olarak elde edilmiştir. Alt kavramlar, temalara göre daha özel bilgi içerdiklerinden dolayı doğru tespit edilmeleri daha zordur. Bu yüzden temalara göre başarı oranı daha düşük çıkmıştır.

IV. SONUÇ

Benzer çalışmalar ile kıyaslandığında, çalışmamızın ortalama sonuçlar bakımından alt kavram ve tema tespitinde oldukça başarılı olduğu görülmektedir. Benzer özellikteki [15] çalışma sonuçlarıyla bizim çalışmamızın sonuçları beraber düşünüldüğünde, her iki çalışmada da konu ve alt-konu tespitinde spor verilerinin başarı oranının eğitim verilerine göre daha yüksek olduğu görülmektedir. Bu durum kavrama dayalı sınıflamada, bazı kavramların ait oldukları sınıfları tam olarak temsil edemediklerinin sonucunu ortaya çıkarmaktadır. Çalışmamızın daha başarılı sonuçlar vermesinin ötesinde, önemli kelimeler dışında farklı özellikler kullanılarak (örneğin tema ve alt kavrama dayalı sözlük desteği) başarımın artabileceği yönündedir.

Deneyler esnasında ayrıca ağ ara katman için farklı nöron sayıları ile denemeler yapılmış, fakat ağın büyüklüğünün başarım oranı üzerinde etkili olmadığı gözlemlenmiştir. Sistemin başarı oranına etki edecek bir önemli faktörün de sistemin eğitim setinin daha büyük tutulmasıdır. Bu çok verimli bir yaklaşım olmadığından, mevcut sistemin öğrenme ile kendi sözlük yapısını oluşturmasının gelecekte en iyi çözüm olacağı düşünülebilir.

V. KAYNAKLAR

[1] Y. H. Li ve A. K. Jain, "Classification of Text Documents," *The Computer Journal*, c. 41, s. 8, ss. 537–546, 1998.

[2] E. S. Yu, ve E. D. Liddy, “Feature Selection in Text Categorization Using The Baldwin Effect,” International Joint Conference on Neural Networks, Washington, ABD, 1999.

[3] R. Bekkerman, R. El-Yaniv, N. Tsihby ve Y. Winter, “Distributional Word Clusters vs. Words for Text Categorization, ” *Journal of Machine Learning Research*, ss. 1-48, 2002.

[4] F. Song, S. Liu ve J. Yang, “A Comparative Study on Text Representation Schemes in Text Categorization,” *Pattern Analysis and Applications,* c.8, s.1-2, ss.199-209, 2005.

[5] M. F. Amasyalı ve B. Diri, “Automatic Turkish Text Categorization in Terms of Author, Genre and Gender,” *11th International Conference on Applications of Natural Language to Information Systems-NLDB 2006,* ss. 221-226, 2006.

[6] F. Türkoğlu, B. Diri ve M. F. Amasyalı, “Author Attribution of Turkish Texts by Feature Mining,” *International Conference on Intelligent Computing*, Qingdao, Çin, ss. 1086-1093, 2007.

[7] A. Çiltik ve T. Güngör, “Time-Efficient Spam E-mail Filtering Using N-gram Models,” *Pattern Recognition Letters,* c. 29, s. 1, ss. 19-33, 2008.

[8] H. Balinsky, A. Balinsky ve S. Simske, “Document Sentences As a Small World,” *2011 IEEE International Conference on Systems, Man, and Cybernetics (SMC),* Anchorage, ABD, 2011.

[9] M. Ghiassi, J. Skinner ve D. Zimbra, “Twitter Brand Sentiment Analysis: A Hybrid System Using N-gram Analysis and Dynamic Artificial Neural Network,” Expert System Applications, c. 40, s. 16, ss. 6266-6282, 2013.

[10] D. Tanasa ve B. Trousse, "Advanced Data Preprocessing for Intersites Web Usage Mining," *IEEE Intelligent Systems,* c.19, s. 2, 2004.

[11] V. Chitraa ve Dr. A. S. Davamani, “A Survey on Preprocessing Methods for Web Usage Data,” *International Journal of Computer Science and Information Security*, c.7, s.3, 2010.

[12] H. Balinsky, A. Balinsky ve S. Simske, “On the Helmholtz Principle for Data Mining,” Third International Conference on Emerging Security Technologies (EST), Lisbon, Portekiz, 2012.

[13] H. Balinsky, A. Balinsky ve S. Simske, “On Helmholtz’s Principle for Documents Processing,” *Proceedings of the 10th ACM Symposium on Document Engineering*, Manchester, İngiltere, ss. 283-286, 2010.

[14] M. Tutkan, M. C. Ganiz ve S. Akyokuş, ”Metin Sınıflandırma için Eğitimsiz Bir Anlamsal Özellik Seçimi Yöntemi,” Bilgisayar ve Biyomedikal Mühendisliği Sempozyumu, Bursa, Türkiye, 2014.

[15] M. Turan ve C. Sönmez, ”Automatize Document Topic and Subtopic Detection with Support of a Corpus,” *Procedia - Social and Behavioral Sciences,* c. 177, ss. 169-177, 2015.