Metin Sınıflandırmada Yeni Özellik Çıkarımı
A New Feature Extraction Method for Text Classification

H.Kemal Yıldız[1], Murat Gençtav[2], Nurullah Usta[3], Banu Diri[4], M.Fatih Amasyalı[5]

[1,2,3,4,5] Yıldız Teknik Üniversitesi, Bilgisayar Mühendisliği Bölümü
[1]zarasoft@hotmail.com, [2,3]{muratgenctav,nurullahusta}@gmail.com,
[4,5]{banu,mfatih}@ce.yildiz.edu.tr

Özetçe

Bu çalışmada, Türkçe'nin biçimbirim yapısı kullanılarak türü bilinmeyen Türkçe bir metnin sınıflandırılması için geliştirilmiş bir özellik çıkarma yöntemi anlatılmaktadır. Önerilen yöntem, kelimelerin gövdelerini özellik olarak almakta ve özellik sayısının fazla olmasından dolayı her bir metin, sınıf sayısı kadar özellik ile ifade edilmektedir. Metne ait olan özellik değerleri, o doküman içerisinde geçen kelime gövdelerinin her sınıf içerisindeki kullanım sıklılarının toplamından oluşmaktadır. Kısacası geliştirilen yöntem bir ağırlıklandırma metodu olup, kelime gövdesi hangi sınıfta daha çok geçmiş ise ilgili metin de bu sınıfa daha fazla dahil olmaktadır. Kelime gövdelerinin kullanım sıklıklarının çıkarılması, kaç farklı metin içerisinde geçtiğinin bulunması sırasında hızı arttırmak amacıyla Trie ağaç yapısı kullanılmıştır.

Bu çalışmada ekonomi, sağlık, magazin, spor ve siyaset olmak üzere beş sınıf seçilmiştir. Elde edilen özellik vektörünün başarısı, her kelimeyi bir özellik olarak kabul eden geleneksel yaklaşımla Naive Bayes, Destek Vektör Makinesi, K-En Yakın Komşuluk, C 4.5 ve Rastgele Orman sınıflandırma yöntemleri ile karşılaştırılmıştır ve daha yüksek başarılar elde edilmiştir. En yüksek başarı Naive Bayes ile % 96.25 olarak alınmıştır.

Abstract

In this study, we have established a feature extraction process for the classification of unknown genres of Turkish texts by using Turkish morphology. The proposed method considers the features as the word stems. The fact that the number of the features exceeds the practical computing limits each document represented by a number of features as in the document classes. Each word stem in a document analyzed in different classes and the sum of the usage frequency in the document classes given the feature value of that document. To speed up the process of extracting usage frequencies of word stems and analyzing it in different document classes Trie tree structure has been used.

In this study, we have selected five different classes which are economy, healthy, magazine, sports and politics. The performance of the established method has been compared the Bag of Words approach by using Naive Bayes, Support Vector Machine, K-Nearest Neighbor, C 4.5 and Random Forest. The best performance achieved is 96.25% which has been observed using the Naive Bayes with our new feature vectors.

1. Giriş

İnternet kullanımındaki hızlı artış ve iletişim alanındaki gelişmeler, veri birikiminin artmasına neden olmuştur. Bu verilerden otomatik olarak bilginin çıkarımı önemli bir çalışma alanıdır. Bu alanda yapılan çalışmaların en önemlilerinden biri metinlerin sınıflandırılmasıdır. Metin sınıflandırma, o dokümanın özelliklerine bakarak, önceden belirlenmiş belli sayıdaki kategorilerden hangisine dahil olacağını belirlemektir. Metin sınıflandırma bilgi alma (information retrieval), bilgi çıkarma (information extraction), doküman indeksleme/filtreleme, otomatik olarak metadata elde etme ve web sayfalarını hiyerarşik olarak düzenleme gibi pek çok alanda önemli bir rol oynamaktadır [1].

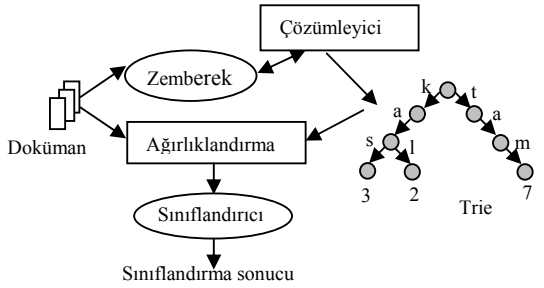
Metin sınıflandırma sistemlerinin ilk örnekleri 70'li yıllarda karşımıza otomatik doküman indeksleme olarak çıkmıştır. Belirli bir konu için özel sözlükler oluşturulmuş ve bu sözlük içerisindeki kelimeler birer kategori gibi algılanarak metinler sınıflandırılmıştır. Fürnkranz [2] n-gram (2 ve 3 uzunluğunda) özelliklerini, Tan ve arkadaşları [3] 2-gram'ları (bi-gram) kullanarak bir algoritma geliştirmiş ve doküman sınıflandırmada performansı arttırmışlardır. Çatal ve arkadaşları [4], n-gram'ları kullanarak NECL adını verdikleri bir sistem geliştirmişlerdir. Diri ve Amasyalı [5] bir metnin yazarını ve türünü belirlemede kullanılmak üzere 22 adet stil belirleyicisi oluşturmuş ve bunları kullanan bir sınıflandırma sistemi geliştirmişlerdir. Yine Amasyalı ve Diri [6] 2 ve 3-gram'ları kullanarak metnin yazarını, türünü ve yazarının cinsiyetini belirme, Amasyalı ve Yıldırım [7] Türkçe haber metinlerinin otomatik olarak sınıflandırılması üzerine çalışmışlardır.

Bu çalışmada içerisinde altı farklı kategoride 750 adet Türkçe dokümandan oluşan bir veri seti kullanılmıştır. Türkçe kelimelerin gövdeleri de özellik olarak alınarak ağırlıklandırmaya dayalı yeni bir özellik çıkarma yöntemi geliştirilmiştir. Geliştirilen sistemin başarısı makine öğrenmesi yöntemlerinden Naive Bayes (NB), Destek Vektör Makinesi (DVM), C 4.5, K-En Yakın Komşuluk (KEK) ve Rastgele Orman (RO) kullanılarak bir metnin türünün belirlenmesi gerçekleştirilmiştir. Makalenin ikinci bölümünde geliştirilen sistemin yapısı, üçüncü bölümde deneysel sonuçlar ve dördüncü bölümde de sonuç yer almaktadır.

2. Sınıflandırma Sisteminin Yapısı

Geliştirilen sistemin ilk aşamasını her metin içerisinde yer alan kelimelerin açık kaynak kodlu Zemberek programının yardımıyla [8] gövde ve eklerine ayrılması oluşturmaktadır. Bu çalışmada kelimenin sadece gövdesi kullanılmaktadır. Zemberek programından alınan gövdeler, Çözümleyici modülü yardımıyla her kelimenin her sınıfta kaç kez geçtiği ve her sınıf içerisinde kaç farklı dokümanda bulunduğu bilgisi Trie yapısında tutulur. Erişim ve hızda kolaylık sağlayan Trie ağacı, hem eğitim hem de test dokümanlarının özelliklerinin belirlenmesinde ve dokümanların içerisinde yer alan kelime gövdelerine göre ağırlıklandırılma yapılmasında kullanılmaktadır. Her dokümana ait çıkarılan özellikler ile bir

özellik vektörü elde edilir. WEKA (www.cs.waikato.ac.nz/ml/weka/) paketi içerisinde yer alan Naive Bayes, Destek Vektör Makinesi, C 4.5, K-En Yakın Komşuluk ve Rastgele Orman sınıflandırıcı metotları ile oluşturulmuş olan bu özellik vektörünün başarısı test metinleri üzerinde ölçülür. Sitemin blok diyagramı şekil 1'de ki gibi çizilebilir.



Şekil 1*: Sistemin Genel Yapısı.*

2.1. Veri Seti

Sistemin eğitim ve test aşamalarında kullanılan veri seti Hürriyet (www.hurriyet.com.tr),Vatan (www.vatanim.com.tr), Sabah (www.sabah.com.tr) gibi günlük gazetelerin ekonomi, magazin, sağlık, siyaset ve spor konularında yazan yazarlarının köşe yazılarından oluşturulmuştur. Eğitim setinde, her sınıftan 150 adet doküman alınmak şartıyla 750 adet doküman, test setinde ise her sınıftan 80 doküman olmak üzere toplam 400 doküman yer almaktadır.

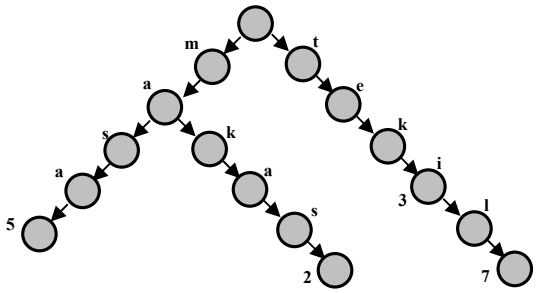
2.2. Kelimelerin Gövdelerinin Bulunması

Türkçe sondan eklemeli bir dildir. Bazen kelime tek bir gövdeden oluşurken, bazen de bu gövdeye bağlı bir veya daha fazla ekten oluşur. Bu kuraldan dolayı çok fazla sayıda yeni kelimeler oluşturulabilinir. Bu çalışmada kelimeler doğrudan alınmayıp, sadece kelime gövdeleri kullanılmıştır. Böylece eklerden doğan kelime farklılıkları ortadan kaldırılarak, frekansı bulunacak kelime sayısı azaltılmıştır. Kelimelerin gövdelerinin bulunmasında açık kaynak kodlu Türkçe Doğal Dil İşleme kütüphanesi Zemberek kullanılmıştır.

2.3. Kelime Frekanslarının Bulunması ve Trie Ağacı

Trie ağacı kullanılarak, kelime gövdelerinin frekansları, kaç dokümanda ve hangi türde kaç defa geçtiği bilgisi kelime gövdelerinden yola çıkılarak ağaç üzerinde saklanmaktadır. Trie, İngilizce re*trie*val kelimesinden gelmekte, aynı zamanda prefix tree (ön ek ağaç) olarak da adlandırılmaktadır. Trie ağacı, düğümlerinde kelimelerin harflerini sıralı bir şekilde tutar ve kelimelerin son harfini tutan düğümlerde de o kelimenin frekansı yer alır. Şekil 2'de verilen Trie ağacı, *masa* kelimesinin 5, *makas* kelimesinin 2, *tek* kelimesinin 3 ve *tekil* kelimesinin 7 kere geçtiğini göstermektedir.

Ağaca kelime eklerken kelimenin başından başlayarak harf harf ilerlenir, eğer kelimenin son harfi bir düğüme denk geliyorsa o düğümün değeri 1 arttırılır, aksi halde kalan harfler ağaca eklenerek kelimenin son harfini gösteren düğüme 1 set edilir. Bu yöntemde, ağaca yeni bir kelime eklemek ve ağaç üzerinde arama yapmak oldukça hızlıdır. Kelimenin arama karmaşıklığı kelimedeki harf sayısına eşittir. Uygulamanın basit, kelime ekleme kelime ekleme/arama hızının yüksek olması bu çalışmada kelime frekanslarının bulunmasında Trie ağacının seçilmiş olmasının sebebidir. Kelimelerin son harflerinin gösterdiği düğümler, sadece kelime frekanslarını göstermekle kalmayıp, aynı zamanda o kelimenin kaç farklı dokümanda ve her sınıfta kaç defa geçtiği bilgilerini de tutmaktadır. Böylece bize gereken tüm bilgiye hızlı bir şekilde erişim olanağı sağlanmaktadır.



Şekil 2*: Örnek Trie Ağacı.*

2.4. Özellik Vektörünün oluşturulması

Geleneksel yaklaşımda metinler kelimelerin frekanslarıyla ifade edildiğinde, her bir metin tüm metinlerdeki farklı kelime sayısı boyutunda bir vektörle gösterilmektedir. Bu metin vektörü literatürde farklı şekillerde elde edilmektedir [9]. Önerdiğimiz metotla karşılaştırma yapabilmek için en sık kullanılan Eşitlik 1 ve Eşitlik 2'deki yöntemler seçilmiştir.

Bazı kelimeler sadece belli sınıftaki metinlerde geçerken, bazıları her metinde çok sayıda geçmektedir. Tablo 1'de her metinde sıkça geçen kelimeler gösterilmiştir. Bu kelimelerin ayırt edici bir özelliği olmadığından, metnin sınıfına olan etkisini azaltılmak için literatürde Eşitlik 2'deki ağırlıklandırma (LogTFIDF [9]) kullanılmaktadır.

Tablo 1: Sık Kullanılan Kelimeler

	Geçtiği Doküman Sayısı				
Kelime	Ekonomi	Magazin	Sağlık	Siyaset	Spor
ve	138	106	146	130	134
ile	101	75	74	59	106
bir	118	115	123	120	118

Önerdiğimiz özellik vektörünün boyutu sınıf sayısına eşittir. Metin içerisinde yer alan kelime gövdelerinin her sınıf içerisindeki kullanım frekansları çıkarılıp, sınıf genelinde toplamları alındığında, her dokümanın ağırlıklandırılması yapılmış olur. Böylece özellik vektörünün bir örneği elde edilir.

Özellik vektörü bulunurken geleneksel yaklaşımdaki her bir kelimenin, her bir metindeki ağırlığı yerine (Eşitlik 1 ve 2) her bir sınıftaki bulunurken Eşitlik 3, 4, ve 5'te verilmiş olan 3 farklı metot denenmiştir.

$$m_j = tf_j \quad (1)$$

$$m_j = \log(tf_j + 0.5) * \log(D/df) \quad (2)$$

$$w_i = \log(tf_i + 0.5) * \log(D/df) \quad (3)$$

$$w_j = tf_j \quad (4)$$

$$w_i = \log(mf_i + 0.5) * \log(D/df) \quad (5)$$

D = Toplam metin sayısı
df = Kelimenin geçtiği metin sayısı
tf_i = Kelimenin i. sınıfta geçtiği metin sayısı
mf_i = Kelimenin i. sınıftaki metinlerde geçme sayısı
w_i = Kelimenin i. sınıfa göre ağırlığı
m_j = Kelimenin j. metindeki ağırlığı
tf_j = Kelimenin j. metinde geçme sayısı

Sonuç olarak her metin 5 (sınıf sayısı) boyutlu vektör ile ifade edilmiştir. "Piyasa bugünlerde düşüşte." cümlesinden oluşan bir metnin özellik vektörünün elde edilmesinde kullanılan kelimelerin, sınıflara göre dağılımları Tablo 2'de verilmiştir.

Tablo 2: Örnek Cümle

	Geçtiği Doküman Sayısı				
Kelime	Ekonomi	Magazin	Sağlık	Siyaset	Spor
Piyasa	39	3	9	2	2
bugün(*lerde*)	27	10	11	23	38
düş(*üşte*)	32	8	9	18	21

Kelimelerin frekanslarının bulunmasında kelime gövdelerinden yararlanılmış ve daha sonra her kelime ağırlıklandırılarak, her sınıf için ağırlıklar toplamı hesaplanmış ve özellik vektörü elde edilmiştir. Örnek ağırlıklandırmada Eşitlik-3'ten yararlanılmış ve sonuçlar da Tablo 3'te verilmiştir. Son olarak da vektörün bütün özelliklerinin toplamı 1 olacak şekilde normalize edilmiştir.

Tablo 3: Metnin Özellik Vektörünün Eldesi

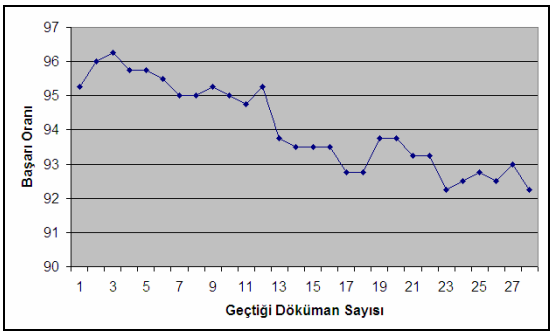
	Kelimelerin Ağırlıklandırılması				
Kelime	Ekonomi	Magazin	Sağlık	Siyaset	Spor
Piyasa	9.4	3.2	5.7	2.3	2.3
bugün(*lerde*)	5.9	4.2	4.3	5.6	6.5
düş(*üşte*)	7.2	4.4	4.6	6.0	6.3
Vektör	**22.6**	**11.8**	**14.8**	**14.0**	**15.2**
Normalize Vektör	**0.29**	**0.15**	**0.19**	**0.18**	**0.19**

3. Deneysel Sonuçlar

Önerdiğimiz özellik vektörü oluşturma yöntemi ile geleneksel yaklaşımın test setine ait karşılaştırma sonuçları bu bölümünde verilmiştir.

Farklı kelime sayısı çok fazla olduğundan tüm metinlerde sadece birkaç kez geçen kelimeler işleme katılmamıştır. Çünkü bu kelimelerin aynı sınıftaki metinlerde geçme olasılığı bile çok düşüktür. Böylece hem başarının hem de işlem hızının arttırılması amaçlanmıştır. Şekil 3, geçtiği metin/doküman sayısı belirli değerlerin altında olan kelime gövdelerinin işleme katılmaması durumunda Eşitlik 3'le ifade edilen metinlerdeki Naive Bayes metodunun sınıflandırma sonuçlarını göstermektedir.

En yüksek başarı en az 4 metin/dokümanda geçen kelimeler kullanılarak elde edilmiştir. Geliştirilen yöntem ile geleneksel yöntemin karşılaştırılması 4'ten daha az metinde geçen kelime gövdeleri çıkarıldığında geriye kalan 2456 gövde ile yapılmıştır.



Şekil 3: Belli Değerin Altında Dokümanda Geçen Kelimelerin Hesaba Katılmaması.

Önerilen özellik vektörü (ilk üç satır) ile geleneksel yaklaşımın (son 4 satır) performanslarının farklı sınıflandırma algoritmaları kullanılarak karşılaştırılması Tablo 4'te yapılmıştır. Geleneksel yaklaşımdaki 2456 boyutlu veriler üzerinde çalışma zamanı ve/veya hafıza problemleri yüzünden Rastgele Orman algoritması çalıştırılamamıştır. WEKA'nın Infogain özellik seçimi metodu kullanılarak özellik vektörünün boyut sayısı 2456'dan 350'ye düşürülmüş ve algoritmalar yeniden çalıştırılmıştır.

Tablo 4: Metin Sınıflandırma Performansları

	Doğruluk Yüzdesi				
Uygulama	**NB**	**DVM**	**KEK**	**C 4.5**	**RO**
Eşitlik 3 (5 boyut)	96.25	95.5	94.75	93.5	94.5
Eşitlik 4 (5 boyut)	94.5	95.75	91	94	95
Eşitlik 5 (5 boyut)	94.75	95.25	94.5	93.75	93.75
Eşitlik 1 (2456 boyut)	92.25	94,75	65	76	-
Eşitlik 2 (2456 boyut)	88.75	73	30.5	70.5	-
Eşitlik 1 (350 boyut)	93.75	89.75	78.25	80.5	86
Eşitlik 2 (350 boyut)	58.5	65.25	46	56.5	56.25

Geleneksel yöntemde Eşitlik 1 kullanıldığında ve özellik sayısı azaltılmış olarak çalışıldığında sınıflandırma başarısı yükselmekte ancak Eşitlik 2'de başarı düşmektedir. Yeni yaklaşımımızda, metinler daha az boyutlu bir uzayda ifade edildiğinden algoritmaların çalıştırılmasında bir problem ile karşılaşılmamıştır. Eşitlik 3-4 ve 5 kullanıldığında alınmış olan sınıflandırma sonuçları geleneksel yönteme göre oldukça yüksektir. En yüksek başarı Eşitlik 3 kullanıldığında %96.25 ile Naive Bayes' den alınmıştır.

Tablo 5: Hata Matrisi

	Tahmin Edilen Tür					
Gerçek Tür	Ekonomi	Magazin	Sağlık	Siyaset	Spor	Hata
Ekonomi	**80**	0	0	0	0	0.0
Magazin	1	**79**	0	0	0	0.01
Sağlık	4	0	**76**	0	0	0.05
Siyasi	7	0	2	**71**	0	0.11
Spor	0	0	0	1	**79**	0.01
						Ortalama: 0.038

400 test metninin birbirlerine göre sınıflandırma doğrulukları Tablo 5'te Hata Matrisi ile gösterilmiştir. Bu

sonuçlar test metinlerinin Eşitlik 3'le ifade edilip, Naive Bayes ile sınıflandırıldığında alınmış olan değerlerdir.

En doğru sınıflandırma ekonomi metinleri ile yapılmışken, 0.11 hata oranı ile yanlış sınıflandırmanın en fazla yapıldığı tür ise siyaset olmuştur. Bir metnin hatalı sınıflandırılması yapıldığında o metnin tahmin edilen sınıfı genelde ekonomi olmaktadır. Sağlık ve siyaset sınıfına ait metinlerde geçen bazı kelimeler ekonomi sınıfında da geçtiği için hatalı sınıflandırmaya neden olmaktadır. Sistemin ortalama hatası Tablo 5'de görüldüğü gibi yüzde 3.8'dir.

4. Sonuçlar

Metin sınıflandırma çalışmalarındaki problemlerden birisi bazı uygulamalarda boyut sayısının çok fazla olmasıdır. Geleneksel yaklaşımda metinler, tüm metinlerde geçen farklı kelime sayısı kadar boyuta sahip vektörlerle ifade edilmektedir. Türkçe gibi sondan eklemeli dillerde farklı kelime sayısı problemi gövdeleme işlemiyle bir ölçüye kadar çözülebilir ancak yine de boyut sayısı 1000'ler seviyesinde kalmaktadır.

Sunulan bu çalışmada metinlerin çok daha az (sınıf sayısı) boyutlu bir uzayda ifade edilmesi sağlanarak problem çözülmeye çalışılmıştır. Kelimelerin metinlerdeki ağırlıklarının yerine, sınıflardaki ağırlıkları kullanılmış ve metinde geçen kelimelerin sınıf ağırlıkları toplanıp normalize edilerek metnin yeni özellik vektörü oluşturulmuştur. Önerdiğimiz özellik vektörünün, metinleri ne ölçüde temsil edebildiğinin bulunması ve geleneksel yaklaşımla karşılaştırılması için uygulama olarak haber metinleri ekonomi, magazin, siyaset, sağlık ve spor olmak üzere 5 sınıfa ayrılmıştır. Sınıflandırma algoritmaları için WEKA paketi, kelimelerin gövdelerinin bulunmasında da Zemberek [8] paketi kullanılmıştır.

Geleneksel metotla 2, geliştirdiğimiz metotla 3 farklı şekilde metinler ifade edilmiş ve 5 farklı algoritma tarafından sınıflandırılmıştır. Önerilen metotta metinler çok daha az boyut ile gösterilmiş olmalarına rağmen, geleneksel yaklaşımdan daha yüksek bir sınıflandırma başarısı alınmış ve büyük boyutlarda çalıştırılamayan karışık algoritmalar da kolaylıkla kullanılabilmiştir. Yöntemde zaman ve yerden tasarruf etmek ve hızlı erişim için Trie ağaç modeli kullanılmıştır.

Önerilen bu özellik vektörü dilden bağımsız olarak çalışmaktadır. Türkçe metinler için yaptığımız bu çalışmada birçok sınıflandırma algoritmasında %90'nın üzerinde sınıflandırma başarısı alınmıştır. En yüksek başarı %96.25 ile Naive Bayes yönteminden alınmıştır. Gelecek çalışma olarak önerdiğimiz bu metodun diğer dillerdeki metinlerin sınıflandırılması performansının ölçülmesi düşünülmektedir.

5. Kaynakça

[1] Türkoğlu F., Diri B., Amasyalı F., "Farklı Özellik Vektörleri ile Türkçe Dokümanların Yazarlarının Belirlenmesi", Turkish Symposium on Artificial Intelligence and Neural Networks, 2006.

[2] Fürnkranz J., "A Study using n-gram Features for Text Categorization", Austrian Research Institute for Artifical Intelligence, 1998.

[3] Tan C.M., Wang Y.F., Lee C.D., "The Use of Bi-grams to Enhance", *Jo*urnal Information Processing and Management, 30(4):529-546, 2002.

[4] Çatal Ç., Erbakırcı K., Erenler Y., "Computer-based Authorship Attribution for Turkish Documents", Turkish Symposium on Artificial Intelligence and Neural Networks, 2003.

[5] Diri B., Amasyalı M.F., "Automatic Author Detection for Turkish Texts", Artificial Neural Networks and Neural Information Processing (ICANN/ICONIP):138-141, 2003.

[6] Amasyalı M.F., Diri B., "Automatic Written Turkish Text Categorization in Terms of Author, Genre and Gender", 11th International Conference on Applications of Natural Language to Information Systems, LNCS Volume 3999, 2006.

[7] Amasyalı .M.F., Yıldırım T., "Otomatik Haber Metinleri Sınıflandırma", SIU, 2004.

[8] Akın M.D., "Zemberek, Türkçe Doğal Dil İşleme Kütüphanesi ve Open Office Yazım Denetimi Eklentisi Sunumu", 2005.

[9] Liao, C., Alpha, S., and Dixon, P., "Feature Preparation in Text Categorization", ADM03 workshop (The Australian Data Mining Workshop), 2003.