Türkçe Metinlerin Sınıflandırılmasında Metin Temsil Yöntemlerinin Performans Karşılaştırılması

A Comparison of Text Representation Methods for Turkish Text Classification

M.Fatih AMASYALI[1], Sümeyra BALCI[1], Esra Nur VARLI[1], Emrah METE[1]

[1]Elektrik Elektronik Fakültesi, Bilgisayar Mühendisliği Bölümü
Yıldız Teknik Üniversitesi
mfatih@ce.yildiz.edu.tr, sumeyrabalci@gmail.com, esranurvarli@gmail.com, emrahmete@gmail.com

Özet

Bir metnin sınıfına metnin hangi özelliklerine bakılarak karar verilebilir? Sınıflandırma probleminin türünün (metnin yazarını, yazarın cinsiyetini, yazarın ruh halini, metnin konusunu, metnin olumlu ya da olumsuz ifadeler içerdiğini tanıma) bu soruya verilecek cevaba etkisi nedir? Bu sorulara çeşitli cevaplar vererek, metin dosyalarının otomatik sınıflandırılması için uzun zamandır çalışmalar sürmektedir. Bu çalışmada çeşitli türdeki 6 adet Türkçe sınıflandırma veri kümesi üzerinde 17 adet özellik grubunun etkisi incelenmiştir. Çıkarılan özellik gruplarına örnek olarak; cümle, kelime, ek sayıları, ngramlar, kelimeler, kelime grupları ve saklı anlam indeksi verilebilir. Türkçe için bugüne kadar yapılmış en kapsamlı karşılaştırma çalışması sunulmuştur. Sonuçlarda n-gramların genel olarak diğer temsil yöntemlerinden daha başarılı sonuçlar ürettiği görülmüştür.

Anahtar kelimeler: Doğal Dil İşleme, Metin Sınıflandırma, Metin Özellikleri, Metin Temsil Yöntemleri

Abstract

Which features are the most important for text classification tasks? How does the type of text classification problem (authorship attribution, gender identification, mood identification, topic identification, sentiment analysis) affect the answer of this question? By giving various answers to these questions, the automatic text classification studies are ongoing for a long time. In this study, 17 text representation methods are compared over 6 different Turkish text classification tasks. Frequencies of the words, stem words, word phrases, n-grams, tokens, and word clusters, Latent Semantic Indexing are examples of the extracted text features. To the best of our knowledge, the most comprehensive study for Turkish text classification is presented. In general, n-grams were produced more successful results than the other text representing methods.
Keywords: Natural Language Processing, Text Classification, Text Categorization, Text Features, Text Representation Methods

1. Giriş

Bir metnin yazarı, konusu, yazarının cinsiyeti, yazarının ruh hali otomatik olarak tahmin edilebilir mi? Bu soruya cevap arayan çalışmalar otomatik metin sınıflandırma adıyla anılmaktadır. Bu çalışmalarda önceden sınıfı belli metinlerin çeşitli özelliklerinin incelenmesiyle yeni gelen bir metnin hangi sınıfa ait olduğu tahmin edilmektedir. Ancak bunun yapılabilmesi için metinlerin hangi özelliklerinin inceleneceğine diğer bir deyişle metinlerin nasıl temsil edileceğine de karar vermek gerekir. Bununla birlikte, sınıflandırma probleminin türünün (yazar, konu, cinsiyet vb.) temsil yöntemine etkisinin olup olmadığı da ayrı bir araştırma konusudur.

Literatürde çeşitli metin sınıflandırma problemleri için birçok başarılı çalışma yapılmıştır. Bunlara örnek olarak yazar tanıma [1, 2], metin konusunu belirleme [3, 4], e-posta sınıflandırma [5], metnin yazarının cinsiyetini belirleme [1] verilebilir. Bu çalışmalarda her problem türü için çeşitli metin temsil yöntemleri önerilmiştir. En çok kullanılan yöntemler; kelime / kelime grubu frekansları [3], ngram frekansları [5, 6], kelime kümeleme [3], saklı anlam indeksleme [7] ve fonksiyonel kelimelerdir [8].

Bununla birlikte literatürde farklı problem türleri için, metin temsil yöntemlerinin detaylı bir karşılaştırılması bulunmamaktadır. Bu çalışmada, bu boşluğu doldurmak amacıyla literatürdeki birçok yöntem ve kendi önerdiğimiz (çok boyutlu ölçekleme ile anlamsal uzay [9], sınıf bilgisine dayalı kelime kümeleme) yöntemler, çeşitli veri kümeleri üzerinde karşılaştırılmıştır.

Veri kümesi olarak yazıdan ruh hali tanıma, film yorumlarında yönelim tespiti, köşe yazarlarını tanıma, yazarın cinsiyetini tanıma, şiirin yazarını tanıma, haberin türünü tanıma olmak üzere toplam 6 metin sınıflandırma koleksiyonu üzerinde çalışılmıştır. 2. bölümde bu veri koleksiyonları tanıtılmıştır. 3. bölümde karşılaştırılan metin temsil yöntemleri detaylı olarak anlatılmıştır. 4. bölümde deneysel sonuçlarda her bir veri koleksiyonu üzerinde hangi

metin temsil yönteminin (metin özellikleri) daha başarılı olduğu, genelde en başarılı yöntemler tartışılmıştır. Son bölümde ise sonuçlar tartışılmıştır.

2. Veri Kümeleri

Bu bölümde farklı uygulama alanlarına ait çeşitli veri kümeleri tanıtılacaktır. Tablo 1'de çalışmada kullanılan veri kümeleri verilmiştir.

Tablo 1: Çalışmada kullanılan veri kümeleri

Veri kümesinin ismi	Sınıf say.	Top. örn. say.	Her bir sınıf. örn. say.	Ort. kelime say.	Ort. cümle say.	Rast. Başarı %
Ruh hali	4	157	38-40	247,1	28,1	25,48
Film yorumları	3	105	35	49,3	4,8	28,57
Köşe yazarı	18	630	35	398,5	45,3	5,56
Cinsiyet	2	105	50-55	377	54	52,38
Haberler	5	1150	230	204,8	17	20
Şiir	7	140	20	79,5	7,4	14,29

Tablo 1'in son sütunundaki rastgele başarı yüzdesi verilerin hepsinin en çok örneği olan sınıfa atandığında elde edilen doğru sınıflandırma yüzdesidir. Kullanılan veri kümelerine www.kemik.yildiz.edu.tr/?id=28 adresinden erişilebilir.

Veri kümelerinin her birinin nasıl oluşturulduğu ve sınıf bilgilerinin içerikleri aşağıda anlatılmıştır.

Ruh hali: Çeşitli blog sitelerinden toplanan ve elle blog yazarının ruh halinin etiketlenmesiyle oluşturulmuştur. 4 sınıf içermektedir. Sınıf etiketleri neşeli, hüzünlü, sinirli, karışık şeklindedir.

Film Yorumları: Çeşitli sinema sitelerinden filmlere yapılmış yorumların toplanmasıyla ve beğeni yönleriyle etiketlenmesiyle oluşturulmuştur. 3 sınıf içermektedir. Sınıf etiketleri pozitif, negatif, tarafsız şeklindedir.

Köşe Yazarları: Çeşitli gazetelerin web sayfalarından 18 adet köşe yazarının 35'er köşe yazısının toplanmasıyla ve yazar isimleriyle etiketlenmesiyle oluşturulmuştur. Sınıf etiketleri yazarların isimleridir.

Cinsiyet: 5'er adet erkek ve kadın köşe yazarının 50-55'şer köşe yazısının toplanmasıyla ve yazarlarının cinsiyetleriyle etiketlenmesiyle oluşturulmuştur. Erkek ve kadın olmak üzere 2 sınıf etiketi içermektedir.

Haberler: Gazetelerin kategorilere ayrılmış haber sayfalarında 5 farklı konuda toplanmış 1150 haber içeren bir veri kümesidir. Sınıf etiketleri Ekonomi, Magazin, Sağlık, Siyasi, Spor şeklindedir.

Şiir: 7 şaire ait 20'şer şiirin toplanması ve şair adlarıyla etiketlenmesiyle oluşturulmuştur. Sınıf etiketleri şairlerin isimleridir.

3. Karşılaştırılan Metinlerin Temsil Yöntemleri

Bu bölümde literatürde sıklıkla kullanılan ve bizim önerdiğimiz metin temsil yöntemleri anlatılmaktadır.

3.1. Frekans Hesaplamalarında Kullanılan Metotlar

Bu bölümde anlatılacak olan Kelime Kökleri, Kelime Türleri, Ngramlar, Fonksiyonel Kelimeler, Kelime Ekleri, Kavram Genelleştirme özellik grupları için frekans hesaplamasında kullanılmak üzere TF, TFIDF, binary, log, normalize1 ve normalize2 olmak üzere 6 farklı metot kullanılmıştır [10]. i indisli özelliğin j indisli metnindeki TFIDF değeri Eşitlik 1'deki gibi hesaplanır.

$$TFIDF(i,j) = TF(i,j) \cdot \log \frac{|TR|}{|TR(i)|} \quad (1)$$

Eşitlik 1'de *TF(i,j)*, *i* indisli özelliğin, *j* indisli metninde yer alma sayısıdır. Toplam metin sayısı *TR* ile, içinde en az bir kere *i* özelliği geçen metin sayısı *TR(i)* ile gösterilmiştir. Örneğin *i* özelliği "gr" 2gramı ise, *TF(i,j)*, "gr" ifadesinin *j* metninde yer alma sayısı olacaktır. *i* özelliği "isim" kelime türü ise *TF(i,j); j* metninde geçen isim türündeki kelime sayısı olacaktır.

Binary(i,j): i indisli özellik j indisli metin yer alıyorsa değeri 1, yer almıyorsa 0'dır.

Log(i,j): Eşitlik 2 ile hesaplanır.

$$Log(i,j) = \log_{10}(TF(i,j)+1) \quad (2)$$

Normalize1(i,j): *TF(i,j)*'leri, metindeki toplam kelime sayısıyla normalize edilmesiyle Eşitlik 3'teki gibi hesaplanır.

$$N1(i,j) = \frac{TF(i,j)}{\left(\sum_j TF(i,j)^2\right)^{1/2}} \quad (3)$$

Normalize2(i,j): *TF(i,j)*'leri, kelimenin toplam geçiş sayısıyla normalize edilmesiyle Eşitlik 4'teki gibi hesaplanır.

$$N2(i,j) = \frac{TF(i,j)}{\left(\sum_i TF(i,j)^2\right)^{1/2}} \quad (4)$$

3.2. Sayılar Özellik Grubu

Metinleri içerdikleri noktalama işaretleri sayıları, kelime sayıları, cümle sayıları, devrik cümle sayıları, harf sayıları, ek sayıları, cümlelerdeki ortalama kelime ve harf sayıları, kelimelerdeki ortalama harf ve ek sayılarıyla ifade eden 19 adet özellikten oluşan özellik grubudur.

3.3. Kelime Kökleri

Metinlerin içerdikleri kelimelerle ifade edildikleri, literatürde en yaygın kullanılan özellik grubudur. Metinler tüm metinlerde en az bir kere geçen farklı kelime sayısı boyutlu bir uzayda ifade edilir. Türkçe gibi eklemeli

dillerde, kelimelerin kendilerini kullanmak farklı kelime sayısını çok fazla arttırmaktadır. Örneğin insan, insanlık, insanlığa, insanlığım kelimeleri farklı kelimeler olarak ele alınırsa metinlerin ifade edildikleri uzayın boyutu çok büyük olmaktadır.

Literatürde bu probleme çözüm olarak sadece kelime köklerinin kullanılması önerilmektedir [11]. Bu sayede hem aynı anlama işaret eden kelimelerin birleştirilmesi (böylelikle metinler arası benzerliğin daha iyi ifade edilmesi sağlanmakta), hem de özellik boyutunun azaltılmasıyla işlem karmaşıklığının azaltılması sağlanmaktadır. Kelimelerin köklerinin bulunması için Zemberek [12] kütüphanesi kullanılmıştır.

3.4. Kelime Türleri

Metinleri içerdikleri kelimelerin türlerinin (isim, sıfat, zamir, edat vb.) frekanslarıyla ifade eden özellik grubudur. Kelimelerin türlerinin bulunmasında Zemberek kullanılmıştır. Toplam 15 tane kelime türü vardır. Buna göre her metin içerdiği isim, sıfat, vb. türündeki kelime sayılarıyla ifade edilmektedir.

3.5. Harf ve Kelime Ngramları

Metinlerin içerdikleri ngramlar ile ifade edildiği özellik grubudur. Ngramlar N boyutlu karakter ya da kelime çerçeveleridir. Örneğin "gitmek" metni için:

Harf 2 gramları: gi – it – tm – me – ek
Harf 3 gramları: git – itm – tme – mek şeklindedir.

Metinlere ait 2 ve 3 harf gramları, 2'li kelime ngramları çıkarılmıştır. Çok seyrek frekans matrisleri üretiyor olmaları sebebiyle, daha büyük pencere boyutuna sahip ngramlar kullanılmamıştır.

3.6. Fonksiyonel Kelimeler

Fonksiyonel kelimeler esasen tek başlarına anlamları olmayan, ancak yazarların üsluplarının belirlenmesinde önemli oldukları düşünülen bağlaç ve edat türündeki kelimelerdir [8] (ve, daha, gibi, de, için vb.). Metinler bu özellik grubunda, önceden belirlenmiş 620 fonksiyonel kelimeyle ifade edilmektedirler.

3.7. Kelime Ekleri

Metinlerin içerdikleri kelimelerin aldıkları eklerin türlerine göre ifade edildikleri özellik grubudur. Türkçe eklemeli bir dil olduğundan anlamın oluşmasında eklerin önemi büyüktür. Kelimelerin eklerinin belirlenmesinde Zemberek kullanılmıştır. Zemberek toplam 126 farklı ek türü çıkarmakta ve dolayısıyla metinler bu 126 ek türünü içermelerine göre sayısallaştırılmaktadırlar.

3.8. Kelime Kümeleme

Bu özellik grubunda, metinlerde geçen yakın anlamlı kelimeler birleştirilip tek bir küme, tek bir özellik haline getirilmektedir [3]. Bu sayede hem boyut sayısı azaltılmakta hem de benzer anlama sahip kelimeler tek bir kavram olarak kullanıldığından metinler arası ilişkilerde artmaktadır. Örneğin hayvan kavramına ait kelimeler (at, kedi, deve vb.) aynı kümede yer alırlarsa, sadece "at" kavramını içeren ve sadece "kedi" kavramını içeren iki metin aynı özelliğe "hayvan" sahip olacaklardır. Bir kelime kümesinin bir metindeki frekansı, küme içindeki kelimelerin o metindeki geçiş sayılarının toplamıdır.

Kelime kümelerinin belirlenmesinde 4 kaynak kullanılmıştır: TF matrisi, TFIDF matrisi, Cooccurance (birlikte geçme) matrisi ve sınıf bilgisi. TF ve TFIDF matrisi 3.1. bölümde, Cooccurance (birlikte geçme) matrisi 3.9. bölümde, sınıf bilgisi 3.14. bölümde anlatılmıştır. Kelimelerin bu kaynaklar kullanılarak kümelenmesi için 3 algoritma kullanılmıştır.

1. Hiyerarşik Kümeleme: Her bir adımda birbirine en çok benzeyen iki kümenin birleştirildiği algoritmadır [13]. Başlangıçta her bir örnek ayrı bir kümeyi ifade eder. Kümelerin birbirine benzerliklerinin ölçülmesinde 3 yol izlenmektedir. A) En Yakın: Kümeler birbirine en yakın elemanları kadar yakındır. B) Ortalama: İki kümenin her bir elemanının diğer kümenin her bir elemanıyla arasındaki mesafelerin ortalaması ile yakınlık ölçülür. C) En uzak: Kümeler birbirine en uzak elemanları kadar yakındır.

2. K-means: Veri dağılımını en iyi temsil edebilecek küme merkezlerini bulunmaya çalışır [14]. Başlangıçta küme merkezleri rastgele atanır. Küme merkezlerinin kendi kümelerindeki elemanlara olan ortalama mesafesi iterasyonlar ilerledikçe azalır.

3. SOM: Bu algoritmada veri dağılımını en iyi temsil edebilecek küme merkezlerini bulunmaya çalışır. Küme merkezleri başlangıçta yine rastgele atanır. Buna ek olarak merkezler birbirlerine rastgele bağlanırlar. Her bir küme merkezinin güncellenmesinde ona bağlı olan diğer küme merkezleri de güncellenir. Bu sayede hem verileri ifade edecek küme merkezleri hem de verinin topolojisi bulunmaktadır [15].

3.9. Birlikte Geçme Matrisi

Birlikte geçme matrisi metinlerdeki farklı kelime sayısı boyutlu simetrik bir matristir. Matrisin i,j hücresinin değeri, i. kelime ile j. kelimenin eğitim metinlerinden kaçında birlikte geçtiğini ifade etmektedir. Birlikte geçme matrisi kelime kümelemede (Bölüm 3.8) ve anlamsal uzayda (Bölüm 3.12) kullanılmaktadır.

3.10. Kelime Filtreleme

Her metinde geçen kelimelerin ayırt ediciliği azdır. Buna karşın çok az metinde geçen kelimeler de gereksiz yere boyut sayısını arttırabilirler. Bu nedenlerle metinlerin ifadesinde kullanılan kelimeler, frekansa dayalı bir filtreden geçirilmişlerdir. Kelimeler belirlenen minimum ve maksimum geçiş sayılarına göre filtrelenmiştir.

3.11. Saklı Anlam İndeksleme

Bu yöntem, sınıf bilgisini kullanmadan metin-kelime matrisi (A) üzerinde Eşitlik 5'teki Tekil Değer Ayrıştırma (Singular Value Decomposition) işlemini yaparak metinlerin boyut sayısını azaltır [7].

$$A_{m\times n} = U_{m\times k}.S_{k\times k}.V^{T}_{n\times k} \quad (5)$$

Bu işlem sonunda S matrisinde, özvektörlerin özdeğerleri diagonalde büyükten küçüğe sıralanmış olur. Seçilen boyut (*k*) adedi işleme alınarak, A matrisi m*n'lik bir matristen (m→metin sayısı, n→ farklı kelime sayısı), m*k'lık (k→ metinlerin yeni boyut sayısı) bir matrise dönüştürülür. Tekil Değer Ayrıştırma için JAMA kütüphanesi [16] kullanılmıştır.

3.12. Anlamsal Uzay

Bu özellik grubunda, metinlerin anlamsal bir uzaydaki koordinatları bulunmaktadır [17]. Bunun için önce metinleri oluşturan kelimelerin anlamsal uzaydaki koordinatları bulunup, daha sonra, metinlerin koordinatları; içerdikleri kelimelerin koordinatlarının ortalaması alınarak hesaplanmaktadır. Kelimelerin sayısal koordinatları, kelimelerin birbirlerine anlamsal yakınlıklarıyla uyumlu olarak bulunmaktadır. Buna göre birbirine anlamsal olarak yakın 2 kelimenin bulunan koordinatları arasındaki Öklid mesafesi de küçük olmaktadır. 2 kelimenin birbirine anlamsal yakınlığının ölçütü olarak Harris'in [18] yaklaşımı kullanılmıştır. Buna göre 2 kelime ne kadar çok birlikte kullanılıyorsa birbirlerine anlamsal olarak o kadar yakındır. Kelimelerin bu yakınlıklarını ölçmek için birlikte geçtikleri metin sayıları hesaplanmış ve Bölüm 3.9'da anlatılan Birlikte Geçme Matrisi'ne yazılmıştır.

Literatürde aralarındaki 2'li uzaklıklar bilinen kavramların bu uzaklıklarla uyumlu koordinatlarının bulunması için Çok Boyutlu Ölçekleme (Multi Dimensional Scaling) [19] yöntemi kullanılmaktadır. Bu yöntem uzaklıklar üzerinde çalıştığından bir yakınlık ölçütü olan Birlikte Geçme Matrisi Eşitlik 6'daki gibi tersine çevrilerek uzaklık ölçütüne dönüştürülmüştür.

$$dis(i,j) = \frac{1}{sim(i,j)} \quad (6)$$

Eşitlik 6'da, *dis(i,j);* i. ve j. kelimeler arasındaki uzaklığı, *sim(i,j) ;* i. ve j. kelimelerin birlikte geçtikleri metin sayısını göstermektedir. Bu yeni matrise Çok Boyutlu Ölçekleme uygulanarak kelimelerin koordinatları bulunmuştur. Çok Boyutlu Ölçekleme için MDSJ Kütüphanesi [20] kullanılmıştır. Metinlerin koordinatları ise daha önce de belirtildiği gibi içinde geçen kelimelerin bu anlamsal uzaydaki koordinatlarının ortalaması alınarak bulunmuştur.

3.13. Kavram Genelleştirme

Kelimelerin anlamlarına göre hiyerarşik bir yapıda düzenlendikleri Wordnet, Conceptnet gibi birçok çalışma mevcuttur. Türkçe için ise benzer yapıda hazırlanmış Türkçe Wordnet [21] bulunmaktadır. Bu özellik grubunda, bu hazır, sabit, eğitim kümesinden bağımsız veri kaynakları, kelimeleri bir üst kavramlarıyla ifade etmede kullanılmıştır. Bu sayede kelime kümelemede olduğu gibi, hem metinlerin boyut sayısı indirgenmiş, hem de metinlerin içerdikleri kavramlar daha anlamsal bir şekilde ifade edilmiş olmaktadırlar.

Kelimeleri bir üst kavramlarıyla ifade etme (genelleştirme) işlemi için 2 veri kümesi kullanılmıştır. İlki Türkçe Wordnet'ten çıkarılmış 15018 adet kavram-üst kavram ikilisidir. İkincisi ise Zemberek'in kütüphanesinden alınan özel isimler listesidir. Eğer bir kelime kavram–üst kavram ikilileri listesinde kavramlar içinde yer alıyorsa onun yerine karşılık gelen üst kavram kullanılmıştır. Eğer bir kelime özel isimler listesinde yer alıyorsa onun yerine "insan" kavramı kullanılmıştır.

3.14. Sınıf Bilgisiyle Kelime Kümeleme

Bizim önerdiğimiz bu yöntemde, öncelikle kelimelerin her sınıftaki frekansları bulunarak kelimeler sınıf sayısı boyutlu bir uzayda birer noktaya dönüştürülmektedir. Daha sonra kelimeler, bu boyutları üzerinde 3.8. bölümde anlatılan kümeleme metotlarıyla sınıf sayısı adet kümeye ayrılmaktadır. Bu işlem sonunda her bir kelime sınıf sayısı adet kümeden birine ait olmaktadır. Metinler ise, sınıf sayısı adet kümenin frekanslarıyla ifade edilmektedirler. Bir kümenin bir metindeki frekansı, o kümede bulunan tüm kelimelerin o metindeki frekanslarının toplamıdır. Bu sayede metinler kelime sayısı adet boyut yerine, sınıf sayısı adet boyutla ifade edilmiş olmaktadır.

4. Deneysel Sonuçlar

Bölüm 2'de tanıtılmış olan 6 veri kümesi için, Bölüm 4'te anlatılan özellik gruplarının çeşitli konfigürasyonlarıyla arff'ler üretilmiştir. Özellik gruplarının her birinin çeşitli konfigürasyonları bulunmaktadır. Tablo 2'de özellik gruplarının (metin temsil yöntemlerinin) adları ve yazının bundan sonraki kısmında kullanılan kısaltmaları verilmiştir.

Tablo 2: Özellik gruplarının isim ve kısaltmaları

Kısaltma	Açıklama
Say	Sayılar Özellik Grubu (Bölüm 3.2)
2G	Harf 2 gramları (Bölüm 3.5)
3G	Harf 3 gramları (Bölüm 3.5)
K2G	Kelime 2 gramları (Bölüm 3.5)
KE	Kelime ekleri (Bölüm 3.7)
KT	Kelime Türleri (Bölüm 3.4)
FK	Fonksiyonel kelimeler (Bölüm 3.6)
KGI	Kavram genelleştirme isim tabanlı (Bölüm 3.13)
KGO	Kavram genelleştirme özel isim tabanlı (Bölüm 4.13)
KK	Kelime kökleri
KKHAL	Hiyerarşik kelime kümeleme (küme benzerlikleri ortalamaya göre) (Bölüm 3.8)
KKHCL	Hiyerarşik kelime kümeleme (küme benzerlikleri en uzak elemanlara göre) (Bölüm 3.8)
KKHSL	Hiyerarşik kelime kümeleme (küme benzerlikleri en yakın elemanlara göre) (Bölüm 3.8)
KKK	Kmeans ile kelime kümeleme (Bölüm 3.8)
KKS	SOM ile kelime kümeleme (Bölüm 3.8)
MDS	Birlikte geçme matrisi tabanlı anlamsal uzay (Bölüm 3.12)
LSI	Saklı Anlam İndeksleme (Bölüm 3.11)

Tablo 3'te kelime Türleri, 2 gram, 3 gram, Fonksiyonel kelimeler, kelime 2 gramları, kelime ekleri, kavram genelleştirme isim, kavram genelleştirme özel isim, kelime kökleri olmak üzere toplam 9 özellik gruplarının her biri için kullanılan 6 frekans hesaplama yöntemi ve yazının bundan sonraki bölümlerinde kullanılacak olan kısaltmaları verilmiştir.

Tablo 3: Frekans hesaplama için kullanılan yöntemler

Yöntem	Açıklama
Binary	Bir kavram metinde geçiyorsa 1 geçmiyorsa 0 (Bölüm 3.1)
Log	Eşitlik 2.
N1	Eşitlik 3.
N2	Eşitlik 4.
TF	Bir kavramın bir metindeki geçiş sayısı (Bölüm 3.1)
TFIDF	Eşitlik 1.

Tablo 4'te kelime kümelemede, KKHAL, KKHCL, KKHSL, KKK, KKS olmak üzere toplam 5 metodun her biri için kelimelerin kümelenmesinde kullanılan 4 matrisin isimleri ve kısaltmaları verilmiştir.

Tablo 4: Kelime Kümeleme için kullanılan matrisler

Yöntem	Açıklama
Cooccurance	Kelime kümele Birlikte geçme matrisine göre (Bölüm 3.8 ve 3.9)
Snf	Kelime kümele sınıf bilgisine göre (Bölüm 3.14)
mTF	Kelime kümele TF matrisine göre (Bölüm 3.8)
mTFIDF	Kelime kümele TFIDF matrisine göre (Bölüm 3.8)

Kelimelerin kümelenmesinde küme sayısı 50 olarak belirlenmiştir. LSI ve MDS için metinlerin ifade edileceği boyut sayısı 50 olarak belirlenmiştir. MDS'te ve birlikte geçme matrisinin kullanıldığı her yöntemde yakınlıktan uzaklığa geçiş için Eşitlik 6 kullanılmıştır. Say özellik grubu için değişken bir parametre kullanılmamıştır.

Üretilen tüm metin temsil yöntemleri WEKA [22] ile birlikte kullanılabilmeleri için arff formatında kaydedilmiştir. Sonuç olarak 6 veri kümesi her biri için (9*6) + (5*4) + 3 (LSI, MDS, Say) =77 arff, toplamda 77*6 = 462 arff üretilmiş ve her arff üzerinde 5'li çapraz geçerlemeyle WEKA kütüphanesinde yer alan 5 adet sınıflandırıcının (en yakın komşu-1NN, karar ağacı-C4.5, destek vektör makineleri-SVM, Naive Bayes-NB, Random Forest-RF) performansı ölçülmüştür. Özellik gruplarının sınıflandırma performanslarını gösteren tablolardaki (Tablo 5-10) tüm değerler 5'li çapraz geçerlemenin ortalama değerleridir.

Üretilen arff'lerin özellik sayısı (boyutu) 5000'den fazla olanlarda zaman ve hafıza problemlerinden ötürü özellikler önce InfoGain [23]'lerine göre sıralanmış daha sonra en yüksek IG'e sahip 100 özellik seçilerek arff bu haliyle kaydedilmiştir. Sonuç tablolarında (Tablo 5-10), kullanılan 5 algoritmadan en yüksek performansa sahip olanının adı ve başarı yüzdesi verilmiştir. Özellik sayısı 101 olan kayıtlarda orijinal özellik sayısı 5000'i aştığından bilgi kazancına göre özellik seçimi yapılmıştır. *arff* dosya formatında sınıf bir özellik olarak yer almaktadır. Buna göre özellik sayısı *d* olan bir veri kümesinde, *d-1* adet özellikle sınıf etiketi tahmin edilmektedir.

4.1. Şiir Veri Kümesi Denemeleri

Tablo 5'te bir şiirin yazarını tahmin etme problemi üzerinde yaptığımız denemeler verilmiştir.

Tablo 5: Şiir veri kümesinde her bir özellik grubunun en başarılı olduğu konfigürasyon ve başarı yüzdeleri, rastgele başarı % 14,29

Özellik Grubu	Konfi gürasyon	Özellik sayısı	Başarı yüzdesi	Sınıflandırıcı
3G	Binary	101	75,29	NB
KKHCL	Snf	8	67,86	1NN
2G	N1	1670	63,86	SVM
KKK	Snf	8	62,29	RF
Say		20	53,29	RF
KGI	Binary	1829	48	NB
KGO	Binary	1672	43,71	NB
KKS	Snf	8	41,86	NB
LSI		51	41,71	RF
KK	Binary	441	40,57	NB
FK	N1	410	36,14	RF
KE	TF	102	33,57	NB
KT	Log	16	32,29	SVM
KKHAL	mTFIDF	51	30,71	RF
KKHSL	Snf	8	29,14	1NN
MDS		51	28,71	SVM
K2G	TFIDF	21	18,86	C45

Tablo 5 incelendiğinde bir şiirin yazarını tahmin etmede en başarılı metin temsil yönteminin (özellik grubunun) 3gramlar'ı binary olarak kodlamak (3G-Binary) olduğu görülmektedir. Veri kümesi için üretilen tekil 3gramların sayısı 5 bin'i geçtiğinden özellik seçimi yapılmış ve bilgi kazancına göre en iyi 100 adet 3gram metinlerin temsilinde kullanılmıştır. 7 şaire ait 20'şer şiirle elde edilen sonuçlara göre bir şiirin yazarı % 75,29'luk doğrulukla tahmin edilebilmektedir. En başarılı 2 yöntem (3G-Binary, KKHCL-Snf) arasında oldukça büyük bir fark bulunmaktadır.

4.2. Köşe Yazarı Veri Kümesi Denemeleri

Tablo 6'da bir köşe yazısının yazarını tahmin etme problemi üzerinde yaptığımız denemeler verilmiştir.

Tablo 6: Köşe Yazarı veri kümesinde her bir özellik grubunun en başarılı olduğu konfigürasyon ve başarı yüzdeleri, rastgele başarı % 5,56

Özellik Grubu	Konfigürasyon	Özellik sayısı	Başarı yüzdesi	Sınıflandırıcı
2G	Log	3817	93,78	SVM
3G	Log	100	86,76	SVM
KK	Binary	800	85,68	SVM
FK	Log	534	81,49	SVM
KGI	Log	470	79,78	SVM
KGO	Log	448	79,68	SVM
Say		20	73,4	RF
KKK	Snf	19	65,49	RF
LSI		51	63,59	RF
KKHCL	Snf	19	63,17	RF
KE	Log	114	61,05	SVM
MDS		51	56,41	NB
K2G	Binary	427	53,94	NB
KKS	mTFIDF	43	53,87	RF
KT	Log	16	50,51	NB
KKHAL	mTFIDF	51	47,17	SVM
KKHSL	Co	51	35,14	RF

Tablo 6 incelendiğinde bir köşe yazısının yazarını tahmin etmede en başarılı metin temsil yönteminin (özellik grubunun) 2gramlar'ı Log olarak kodlamak (2G-Log) olduğu görülmektedir. 18 köşe yazarının 35'şer yazısıyla elde edilen sonuçlara göre bir köşe yazısının yazarı % 93,78'lik doğrulukla tahmin edilebilmektedir. En başarılı 2 yöntem (2G-Log, 3G-Log) arasında büyük bir fark bulunmaktadır.

Şiirlerin yazarlarının doğru tahmin yüzdesi % 75,29 iken, köşe yazılarının yazarlarının yüzdesi % 93,78'dir. Üstelik köşe yazarı veri kümemizde 18 yazar(sınıf) varken, şiir veri kümemizde 7 şair (sınıf) vardır. Sınıf sayısının artmış olmasına rağmen, başarının da artmış olması beklenmedik bir durumdur. Bu duruma 2 açıklama getirilebilir. İlki sınıflara ait örnek sayılarıdır. Veri kümelerinde şairlere ait 20'şer şiir varken, köşe yazarlarına ait 35'er örnek vardır ki bu köşe yazarlarının daha başarılı tahmin edilebilmesine olanak sağlamış olabilir. İkinci açıklama ise şiirlerde, köşe yazılarından çok daha fazla söz sanatına başvuruluyor olmasıdır. Ve eğer bir şairin üslubu kullandığı söz sanatı türlerine göre belirlenebiliyorsa ve çıkarılan özelliklerde bu söz sanatları yer almadığından şiirlerin yazarlarını köşe yazarları kadar iyi tahmin edemiyor olabilir.

4.3. Haberler Veri Kümesi Denemeleri

Tablo 7'de bir haber metninin konusunu tahmin etme problemi üzerinde yaptığımız denemeler verilmiştir.

Tablo 7: Haberler veri kümesinde her bir özellik grubunun en başarılı olduğu konfigürasyon ve başarı yüzdeleri, rastgele başarı % 20

Özellik Grubu	Konfigürasyon	Özellik sayısı	Başarı yüzdesi	Sınıflandırıcı
2G	N1	3698	94,54	SVM
KK	N1	864	92,63	RF
KGO	Log	719	91,13	RF
KGI	N1	724	90,85	RF
3G	N1	101	90,47	RF
MDS		51	89,84	RF
KKK	Snf	6	87,51	RF
KKHCL	Snf	6	87,25	RF
KKS	mTFIDF	47	81,63	RF
LSI		11	73,48	RF
KKHAL	mTFIDF	51	73,44	RF
KE	Log	120	68,02	SVM
K2G	TFIDF	101	65,84	RF
FK	N1	534	62,61	RF
Say		20	57,08	RF
KT	Log	16	56,21	SVM
KKHSL	Co	51	47,06	C45

Tablo 7 incelendiğinde bir haberin türünü tahmin etmede en başarılı metin temsil yönteminin (özellik grubunun) 2gramlar'ı Eşitlik 3'teki gibi normalize edip kodlamak (2G-N1) olduğu görülmektedir. 5 haber türüne ait 230'ar haber metniyle elde edilen sonuçlara göre bir haberin türü %94,54'lük doğrulukla tahmin edilebilmektedir. En başarılı 2 yöntem (2G-N1, KK-N1) arasında küçük bir fark bulunmaktadır.

4.4. Cinsiyet Veri Kümesi Denemeleri

Tablo 8'de bir köşe yazısının yazarının cinsiyetini tahmin etme problemi üzerinde yaptığımız denemeler verilmiştir.

Tablo 8: Cinsiyet veri kümesinde her bir özellik grubunun en başarılı olduğu konfigürasyon ve başarı yüzdeleri, rastgele başarı % 52,38

Özellik Grubu	Konfigürasyon	Özellik sayısı	Başarı yüzdesi	Sınıflandırıcı
2G	Binary	101	99,62	SVM
3G	Log	101	98,67	1NN
LSI		51	98,1	RF
KK	Binary	101	95,81	SVM
KGI	Binary	95	95,81	SVM
KKK	Snf	3	95,62	1NN

KGO	TFIDF	93	95,43	SVM
KKHSL	Snf	3	95,24	C45
KKHCL	Snf	3	93,52	1NN
K2G	TFIDF	83	87,62	NB
FK	Log	42	81,14	SVM
Say		20	79,24	RF
KE	Log	11	77,33	NB
KKS	mTF	49	76,57	SVM
KKHAL	Co	6	71,05	1NN
MDS		6	70,67	NB
KT	N1	3	64,19	NB

Tablo 8 incelendiğinde bir metnin yazarının cinsiyetini tahmin etmede en başarılı metin temsil yönteminin (özellik grubunun) 2gramlar'ı binary olarak kodlamak (2G-Binary) olduğu görülmektedir. Veri kümesi için üretilen tekil 2gramların sayısı 5 bin'i geçtiğinden özellik seçimi yapılmış ve bilgi kazancına göre en iyi 100 adet 2gram metinlerin temsilinde kullanılmıştır. Her bir cinsiyetteki yazarlara ait 50-55'şer yazı ile elde edilen sonuçlara göre bir metnin yazarının cinsiyeti % 99,62 gibi yüksek bir doğrulukla tahmin edilebilmektedir. En başarılı 2 yöntem (2G-Binary, 3G-Log) arasında küçük bir fark bulunmaktadır.

4.5. Film Yorumları Veri Kümesi Denemeleri

Tablo 9'da bir filme yapılmış yorumun duygusal yönünü tahmin etme problemi üzerinde yaptığımız denemeler verilmiştir.

Tablo 9: Film Yorumları veri kümesinde her bir özellik grubunun en başarılı olduğu konfigürasyon ve başarı yüzdeleri, rastgele başarı % 28,57

Özellik Grubu	Konfi gürasyon	Özellik sayısı	Başarı yüzdesi	Sınıflandırıcı
KKK	Snf	4	96,38	1NN
3G	Binary	101	88,76	NB
KKHCL	Snf	4	75,81	SVM
LSI		6	75,62	RF
KKS	Snf	4	65,14	SVM
KGI	Binary	21	64,76	NB
2G	Binary	93	59,62	NB
KGO	TF	21	54,86	NB
KK	Binary	23	54,29	SVM
KKHAL	mTFIDF	6	52,57	C45
FK	Log	54	48,76	RF
K2G	TFIDF	6	48,19	C45
KE	N1	10	46,86	SVM
MDS		51	46,29	NB
KKHSL	Co	51	46,1	NB
KT	Binary	16	44,19	C45
Say		20	38,48	NB

Tablo 9 incelendiğinde bir filme yapılan yorumun negatif, pozitif ya da nötr olduğunu tahmin etmede en başarılı metin temsil yönteminin (özellik grubunun) kelime köklerini Kmeans ile Sınıf bilgisine göre kodlamak (KKK-Snf) olduğu görülmektedir. 3 gruba ait 35'er yorumla elde edilen sonuçlara göre bir yorumun yönü %96,38'lik doğrulukla tahmin edilebilmektedir. En başarılı 2 yöntem (KKK-Snf, 3G-Binary) arasında büyük bir fark bulunmaktadır.

4.6. Ruh Hali Veri Kümesi Denemeleri

Tablo 10'da bir yazının yazıldığı anda yazarının içinde bulunduğu ruh halini tahmin etme problemi üzerinde yaptığımız denemeler verilmiştir.

Tablo10: Ruh Hali veri kümesinde her bir özellik grubunun en başarılı olduğu konfigürasyon ve başarı yüzdeleri, rastgele başarı % 25,48

Özellik Grubu	Konfi gürasyon	Özellik sayısı	Başarı yüzdesi	Sınıflandırıcı
3G	Binary	101	85,23	NB
KKK	Snf	5	80,67	1NN
KKHCL	Snf	5	78,73	1NN
KGI	TFIDF	100	67,29	RF
LSI		6	66,21	RF
KK	Binary	101	64,98	NB
KGO	Binary	101	64,19	NB
2G	N1	101	61,63	RF
K2G	TFIDF	101	56,84	C45
KKHAL	mTF	51	54,94	RF
KKS	Snf	5	49,73	RF
FK	N1	534	48,49	RF
MDS		51	46,76	RF
KE	Log	118	45,83	RF
KT	Log	16	41,12	RF
KKHSL	Co	6	39,26	RF
Say		20	38,06	RF

Tablo 10 incelendiğinde bir blog yazarının blog yazıldığı andaki ruh halini tahmin etmede en başarılı metin temsil yönteminin (özellik grubunun) 3gramlar'ı binary olarak kodlamak (3G-Binary) olduğu görülmektedir. Veri kümesi için üretilen tekil 3gramların sayısı 5 bin'i geçtiğinden özellik seçimi yapılmış ve bilgi kazancına göre en iyi 100 adet 3gram metinlerin temsilinde kullanılmıştır. 4 ruh halinin her birine ait 38-40'ar blogla elde edilen sonuçlara yazarın ruh hali % 85,23'lük doğrulukla tahmin edilebilmektedir. En başarılı 2 yöntem (3G-Binary, KKK-Snf) arasında büyük bir fark bulunmaktadır.

4.7. Deneme Sonuçlarının Birlikte Değerlendirilmesi

Sonuç tabloları (5-10) en başarılı sınıflandırıcılar yönünden incelendiğinde (sınıflandırıcıların tabloların son sütunlarında yer alma sayıları) Tablo 11'deki sonuçlar elde edilmiştir. Tablo 11'deki sayılar sınıflandırıcının 102 (17 özellik grubu * 6 veri kümesi) arff dosyası üzerinde kaçında en başarılı algoritma olduğunu göstermektedir.

Tablo 11: En Başarılı sınıflandırıcıların dağılımı 102 arff dosyası içinde kaç kere en başarılı oldukları

Sınıflandırıcı	Kaç kere en başarılı olduğu
RF	39
SVM	24
NB	23
1NN	9
C45	7

Tablo 11 incelendiğinde en başarılı sınıflandırıcının Random Forest (RF) olduğu, onu Destek Vektör Makineleri (SVM) ve Naive Bayes (NB)'in takip ettiği görülmektedir. En yakın komşu (1NN) ve karar ağaçları (C4.5) algoritmalarının pek başarılı olamadıkları görülmektedir.

Tablolar (5-10) Özellik gruplarının en başarılı oldukları konfigürasyonlar yönünden incelenmesiyle (ikilinin tabloların ilk 2 sütununda yer alma sayıları) ilgili sonuçlar Tablo 12'de verilmiştir. Örneğin 2G özellik grubu, 6 veri kümesinin üçünde N1 ile, ikisinde Binary ile birinde ise Log ile en başarılı olmuştur.

Tablo 12: Özellik gruplarının en iyi oldukları konfigürasyonlar

Özellik Grubu	En iyi olduğu Konfigürasyon ve Sayıları
2G	3 N1, 2 Binary, 1 Log
3G	3 Binary, 2 Log, 1 N1
FK	3 N1, 3 Log
K2G	5 TFIDF, 1 Binary
KE	4 Log, 1 N1, 1 TF
KGI	3 Binary, 1 Log, 1 N1, 1 TFIDF
KGO	2 Binary, 2 Log, 1 TF, 1 TFIDF
KK	5 Binary, 1 N1
KKHAL	4 mTFIDF, 1 mTF, 1 Cooccurance
KKHCL	6 Snf
KKHSL	4 Co, 2 Snf
KKK	6 Snf
KKS	3 Snf, 2 mTFIDF, 1 mTF
KT	4 Log, 1 Norm1, 1 Binary

Tablo 12'ye göre K2G özellik grubunun TFIDF ile, KK özellik grubunun Binary ile ağırlıklandırılması, kelime kümeleme yaparken ise KKHCL ve KKK metotlarında Snf kullanılması daha iyi sonuçlar üretmiştir.

Özellik gruplarının genel olarak ne kadar başarılı oldukları incelemek için Tablo 5-10'larda özellik grubunun her bir veri kümesinde en başarılı kaçıncı oldukları (1-17 arası) yazılmış ve bu sayıların ortalamaları alınmıştır. Elde edilen sonuçlar Tablo 13'te verilmiştir.

Tablo 13: Özellik gruplarının ortalama ve her veri kümesindeki başarı sıralamaları (17 özellik grubu arasında, 6 veri kümesinde ortalama kaçıncı oldukları)

Özellik Grubu	Ortalama Başarı Sıralaması	Şiir	Köşe yazarı	Haberler	Cinsiyet	Film yorumları	Ruh hali
2G	3,5	3	1	1	1	7	8
3G	2,2	1	2	5	2	2	1
FK	10,5	11	4	14	11	11	12
K2G	12,3	17	13	13	10	12	9
KE	12,5	12	11	12	13	13	14
KGI	5	6	5	4	5	6	4
KGO	6,3	7	6	3	7	8	7
KK	5,7	10	3	2	4	9	6
KKHAL	12,7	14	16	11	15	10	10
KKHCL	5,8	2	10	8	9	3	3
KKHSL	14,7	15	17	17	8	15	16
KKK	4,7	4	8	7	6	1	2
KKS	10,2	8	14	9	14	5	11
KT	15,3	13	15	16	17	16	15
LSI	6,7	9	9	10	3	4	5
MDS	12,8	16	12	6	16	14	13
Say	12,2	5	7	15	12	17	17

Tablo 13 incelendiğinde kullanılan 6 veri kümesi için en başarılı 3 özellik grubunun sırasıyla 3G, 2G ve KKK olduğu görülmektedir.

Harf 2gramları (2G) film yorumları ve ruh hali veri kümeleri haricinde çok başarılı sonuçlar elde etmiştir. Harf 3gramları (3G) ise haberler haricinde çok başarılı sonuçlar elde etmiştir. Fonksiyonel kelimeler en yüksek başarılarını köşe yazarlarını tanımada göstermişlerdir. Düz metinlerdeki üslubun belirlenmesinde fonksiyonel kelimelerin etkin bir şekilde kullanılabileceği buradan görülmektedir. Kelime 2gramları (K2G), Kelime türleri (KT) ve kelime ekleri (KE) hiçbir veri kümesinde başarı gösterememişlerdir. Kavram genelleştirmenin başarılı sayılabilmesi için kelime köklerinden (KK) daha başarılı sonuçlar üretmesi gereklidir. Buna göre kavramların bir üst kavramlarıyla ifade edilmesi (KGI), şiir, film yorumları ve ruh hali veri kümelerinde kelime köklerinden (KK) daha iyiyken, özel isimlerin insan kavramına dönüştürülmesi (KGO) sadece şiir veri kümesinde kelime köklerinden daha iyi sonuçlar üretebilmiştir. Kelime kökleri (KK) şiir ve film yorumları haricinde başarılıdır.

Kelime kümeleme metotları (KKHAL, KKHCL, KKHSL, KKK, KKS) arasında en başarılısı kmeans ile kümeleme (KKK) olmuştur. Tablo 12'deki değerlere bakıldığında KKK'nın en iyi Snf ile birlikte çalıştığı gözükmektedir.

Anlamsal uzaylarda LSI, MDS'e göre oldukça başarılı sonuçlar üretmiştir. 6 veri kümesinde, MDS'in LSI'dan iyi olduğu sadece 1 veri kümesi (haberler) bulunmaktadır. Bu veri kümesini diğerlerinden ayıran özelliği her bir sınıfa ait çok sayıda örneğin oluşudur. Buna göre MDS ile iyi sonuçlar almak için örnek sayısının çok olduğu veri kümelerinin kullanımı önerilebilir. Sayılar (Say) özellik grubu şiir ve köşe yazarı veri kümeleri haricinde çok kötü sonuçlar almıştır.

Sonuçlar veri kümelerinin türlerine göre incelenmesiyle Tablo 14 elde edilmiştir.

Tablo 14: Veri kümelerinde en başarılı sonuçlar alan konfigürasyonlar

Veri kümesi	Ort. Kelime Sayısı	Ort. Cümle sayısı	Rastgele Başarı	Elde edilen en yüksek başarı	Özellik sayısı	Özellik grubu ve konfigürasyonu
Ruh hali	247,1	28,1	25,48	85,23	101	3G-Binary
Film yorumları	49,3	4,8	28,57	96,38	4	KKK-Snf
Köşe yazarı	398,5	45,3	4,76	93,78	3817	2G-Log
Cinsiyet	377	54	52,38	99,62	101	2G-Binary
Haberler	204,8	17	20	94,54	3698	2G-N1
Şiir	79,5	7,4	14,29	75,29	101	3G-Binary

Tablo 14 incelendiğinde, en yüksek başarı elde edilen veri kümesi yazarın cinsiyetini belirlediğimiz veri kümesidir. En zor veri kümesinin şiirlerin yazarlarını bulma olduğu görülmektedir. Bununla beraber tüm veri kümelerinde en az % 75'lik bir başarı elde edilmiş olması otomatik metin sınıflandırma konusunda çok çeşitli görevlerde başarının yakalandığını göstermektedir.

En başarılı metin temsili yöntemlerine bakıldığında, harf ngramlarının başarısı göze çarpmaktadır.

Ngramların ağırlıklandırılmasında oldukça popüler olan TF ve TFIDF yerine Binary, Log ve N1 ağırlıklandırma yöntemlerinin daha başarılı oldukları görülmektedir.

Metinlerin boyutlarıyla başarı oranları birlikte incelendiğinde, kısa metinlerde de uzun metinlerde de yüksek başarıların elde edilebildiği görülmüştür.

5. Sonuçlar ve Tartışma

Sınıflandırma uygulamalarında örneklerin nasıl temsil edileceği performansa en çok etki eden parametredir. Bunun doğal sonucu olarak uygun özelliklerin seçimi sınıflandırma performansını arttırmaktadır. Bu çalışmada çeşitli türdeki metin sınıflandırma problemleri için özellik gruplarının performansa etkileri araştırılmıştır.

Kullanılan 6 farklı metin veri kümesinin her biri için 17 özellik grubunun çeşitli konfigürasyonlarından oluşan 77 farklı temsil yöntemi denenmiş ve bu işlem sonucu oluşan 462 (6*77) adet arff dosyası 5 adet sınıflandırma metoduyla sınıflandırılmıştır. Sınıflandırma performansları 5'li çapraz geçerleme ile ölçülmüştür.

Ayrıntıları 5.bölümde verilen sonuçlara göre genel olarak n-gramların diğer metin temsil yöntemlerinden daha başarılı olduğu görülmüştür. Ngramların her türlü dile uygulanabilir olması, herhangi bir dile özel ön işlem gerektirmiyor olması ngramların diğer olumlu yanları olarak sayılabilir. Ngramların ağırlıklandırılmasında oldukça popüler olan TF ve TFIDF yerine Binary, Log ve N1 ağırlıklandırma yöntemlerinin daha başarılı oldukları görülmüştür.

Tarafımızdan geliştirilmiş olan sınıf bilgisine dayalı kelime kümeleme yöntemi, Film Yorumları veri kümesinde en başarılı temsil yöntemidir. Film yorumları veri kümesi, Internet üzerindeki kullanıcı yorumların olumlu ya da olumsuz şekilde etiketlenmesi üzerine bir veri kümesidir ki bu konu günümüzde ticari değeri yüksek olduğundan oldukça sıcak bir araştırma alanıdır.

Gelecek çalışmalar olarak anlamsal uzayda yakınlıktan uzaklığa dönüşümde yeni yöntemlerin kullanılması ve yöntemlerin İngilizce metinler üzerinde denenmesi düşünülmektedir.

6. Teşekkür

Özellik çıkarımına ve deneylerin çalıştırılmasına büyük emek harcayan Feruz DAVLETOV, Arslan TORAYEW, Ümit ÇİFTÇİ, Burhanettin İRMİKÇİ'ye ve değerli yorumlarıyla çalışmamıza katkıda bulunan Kemik Doğal Dil İşleme Grubu (www.kemik.yildiz.edu.tr) üyelerine teşekkür ederiz. Bu çalışma, bir Ericcson şirketi olan Bizitek'in 3110278 no'lu TÜBİTAK-TEYDEB projesi kapsamında desteklenmiştir.

.

Kaynaklar

[1] Amasyalı, M. F., Diri, B., "Automatic Turkish Text Categorization in Terms of Author, Genre and Gender", 11th International Conference on Applications of Natural Language to Information Systems-NLDB2006, LNCS Volume 3999, 2006.

[2] Türkoğlu, F., Diri, B., Amasyalı, M. F., "Author Attribution of Turkish Texts by Feature Mining", Third International Conference on Intelligent Computing, ICIC 2007, Qingdao, China, LNCS Volume 4681, 2007.

[3] Bekkerman R., Ran El-Yaniv, Naftali T., Yoad W., "Distributional Word Clusters vs. Words for Text Categorization", Journal of Machine Learning Research, 1-48, 2002.

[4] Song, F., Liu, S., Yang, J., "A comparative study on text representation schemes in text categorization", Pattern Anal Applic (8), 199–209, 2005.

[5] Çiltik, A. ve Güngör, T., "Time-Efficient Spam E-mail Filtering Using N-gram Models", Pattern Recognition Letters 29(1), 19-33, 2008.

[6] Ciya L., Shamim A., Paul D., "Feature Preparation in Text Categorization", Oracle Text Selected Papers and Presentations, 2001.
[7] Özel, B., "Küresel k-Ortalamalı Gruplama Yöntemi ile Metinlerin ve Terimlerin Saklı Anlam İndekslenmeleri", ASYU, Istanbul, Turkey, 223-227, 2004.
[8] Holmes, D. I., "The Evolution of Stylometry in Humanities Scholarship", Literary and Linguistic Computing, 13 (3): 111-117, 1998.
[9] Amasyalı, M. F., Davletov, F., Torayew, A, Çiftçi, Ü, "text2arff: Türkçe Metinler İçin Özellik Çıkarım Yazılımı", SİU, Diyarbakır, 2010.
[10] Berry, M. W., Browne, M., "Understanding Search Engines: Mathematical Modeling and Text Retrieval", Society for Industrial and Applied Mathematics, Philadelphia, 2005.
[11] Torunoglu, D., Cakirman, E., Ganiz, M.C., Akyokus, S., Gurbuz, M.Z., "Analysis of preprocessing methods on classification of Turkish texts", INISTA, Istanbul, Turkey, 112 - 117, 2011.
[12] http://code.google.com/p/zemberek/
[13] Alpaydın, E., "Introduction to Machine Learning", The MIT Press, 146-148, 2004.
[14] Alpaydın, E., "Introduction to Machine Learning", The MIT Press, 135-139, 2004.
[15] Alpaydın, E., "Introduction to Machine Learning", The MIT Press, 282-283, 2004.
[16] JAMA: A Java Matrix Package, http://math.nist.gov/javanumerics/jama
[17] Amasyalı, M. F., Beken, A., "Türkçe Kelimelerin Anlamsal Benzerliklerinin Ölçülmesi ve Metin Sınıflandırmada Kullanılması", SİU, 2009.
[18] Haris Z. S., "Mathematical structures of language", Wiley, 12, 1968.
[19] Alpaydın, E., "Introduction to Machine Learning", The MIT Press, 121-124, 2004.
[20] Multidimensional Scaling for Java, University of Konstanz, Department of Computer & Information Science, Algorithmics Group, http://www.inf.uni-konstanz.de/algo/software/mdsj/
[21] Bilgin, O., Çetinoğlu, Ö., Oflazer, K., "Building a wordnet for Turkish", Romanian Journal of Information Science and Technology, (7), 163-172, 2004.
[22] Witten, I. H., Frank, E., "Data Mining: Practical machine learning tools and techniques", 2nd Edition, Morgan Kaufmann, San Francisco, 2005.
[23] Hall, M. A., "Correlation-based Feature Selection for Machine Learning", Doktora Tezi, Hamilton, NZ: Waikato University, Department of Computer Science, 1998.