©2022 Published in 10th International Symposium on Innovative Technologies in Engineering and Science 21-23 October 2022 (ISITES2022 Bursa - Turkey)
https://doi.org/10.33793/acperpro.05.03.17783

SMS Spam Filtering for Turkish Language with Machine Learning Algorithms

Makine Öğrenmesi Algoritmaları ile Türkçe için İstenmeyen SMS Filtreleme

[1]Bekir Parlak
*[1]Mühendislik Fakültesi, Bilgisayar Mühendisliği Bölümü Amasya Üniversitesi, Türkiye

Özet

Bu çalışmada, Türkçe dilindeki kısa mesaj hizmeti (SMS=Short Message Service) istenmeyen mesajlarının filtrelenmesinde çeşitli öznitelik seçme yaklaşımlarının ve ön-işleme tekniğinin etkisi araştırıldı. Filtreleme aşamasında tüm öznitelik kümesi, kelime çantası (BoW = Bag of Words) modeliyle açığa çıkarılan özniteliklerden oluşturuldu. Kelime çantasındaki ayırt edici öznitelikler, öznitelik seçim yöntemleri kullanılarak belirlenir. Daha sonra SMS mesajlarını sınıflandırmak için yaygın olarak kullanılan sınıflandırma algoritmalarıyla beslenir. Filtreleme çerçevesi sadece Türkçe SMS veri kümesi üzerinde değerlendirildi. İlgili veri kümeleri üzerinde kapsamlı deneysel analiz, Multinomial Naïve Bayes(MNB) sınıflandırıcı ile EFS(Extensive Feature Selector) öznitelik seçim metotlarının kombinasyonlarının daha iyi sınıflandırma performansı sağladığını ortaya çıkardı. Kullanılan öznitelik seçim yöntemlerinin etkinliği, her sınıflandırıcıda biraz farklılık göstermektedir.

Anahtar Kelimeler: Öznitelik seçimi, öznitelik çıkarımı, SMS, filtreleme

Abstract

In this study, the effect of various feature selection approaches and preprocessing technique on filtering spam messages of Turkish language short message service (SMS=Short Message Service) was investigated. In the filtering stage, the entire feature set consists of the features exposed by the Bag-of-Words (BoW) model. Distinctive features in the BoW are determined using feature selection methods. It is then fed into model classification algorithms that are commonly used to classify SMS messages. The filtering framework was evaluated only on the Turkish SMS dataset. Extensive experimental analysis on the relevant datasets revealed that combinations of MNB classifier and EFS feature selection methods provide better classification performance. The effectiveness of the feature selection methods used varies slightly in each classifier.

Key words: Feature selection, feature extraction, SMS, filtering

1. Introduction

Kısa Mesaj Servisi (SMS=Short Message Service), günümüzde dünya genelinde cep telefonu kullanıcı sayısının hızla artması nedeniyle en yaygın iletişim yöntemlerinden biri haline gelmiştir. Bu hızlı artış istenmeyen (spam) göndericilerin dikkatini çekmiş ve tıpkı istenmeyen e-postalarda olduğu gibi SMS spam mesaj sorunu yaşanmasına neden olmuştur. Günümüzde cep telefonlarına gelen SMS'lerin çoğu maalesef bankaların kredi imkanları, mağazaların indirim duyuruları, iletişim servis sağlayıcılarının yeni tarifeleri gibi rahatsız edici istenmeyen mesajlardır.

*Corresponding author: Address: Faculty of Engineering, Department of Computer Engineering Amasya University, 05100, Amasya TURKEY. E-mail address: bekir.parlak@amasya.edu.tr, Phone: 0358 211 50 53-1504

SMS, kimlik doğrulama gerektiren bankacılık ya da belediye işlemleri gibi tüm önemli uygulamaların sıklıkla kullandığı önemli hizmetlerden bir tanesidir. Bu sebeple hizmetin tamamen kapatılmasından ziyade güvenilir şekilde kullanılması önemlidir. Bunun nedeni, SMS düşük maliyeti ve kolay kullanımı nedeniyle özellikle ürün veya hizmet reklamları ve duyurular için de en çok tercih edilen araçlardan biridir.

Beyaz ve kara liste yöntemlerini içeren basit teknikler, istenmeyen SMS mesajlarını sınıflandırma için yüksek performans sergileyemez. Daha da kötüsü, kara listeye eklenen bir telefon numarası, istenmeyen SMS'lerin yanı sıra normal mesajlar da gönderebilir. Örnek olarak, bir banka yeni kredi fırsatları içeren bir istenmeyen SMS mesajı ve çevrimiçi bankacılık şifresini de içeren normal bir mesaj gönderebilir. Bu durumda içerik tabanlı sınıflandırma gibi daha akıllı tekniklere başvurmak gerekir. Türkiye'de 1 Mayıs 2015 tarihinden itibaren yürürlükte olan e-ticaret yasası ile istenmeyen mesajlar filtrelenmek istenmesine rağmen, kullanıcı istekleri olmadan yurtdışı servis sağlayıcıları üzerinden istenmeyen SMS gönderimleri hala devam etmektedir. Bundan dolayı, kullanıcıları rahatsız eden istenmeyen SMS gönderilerinin tespiti ve filtrelenmesi çalışılan konular arasında yer almaktadır.

Literatürde İngilizce SMS filtreleme[1, 2] çalışmaları ile birlikte Türkçe SMS filtreleme çalışmaları bulunmaktadır[3, 4]. Ballı ve Karasoy[3], SMS tespiti için derin öğrenme tabanlı çalışma gerçekleştirmişlerdir. Farklı bölgelerden ve yaş gruplarından derlenen SMS'ler sınıflandırma performansına etkisi olabilecek öznitelik etiketleri SMS veri kümesine dahil edilmiştir. Word2Vec kütüphanesi yardımıyla oluşturulan bu veri kümesi için bir model oluşturulmuştur. Derin öğrenme tabanlı bu model aracılığıyla veri kümesindeki her SMS için yeni öznitelikler açığa çıkarılmıştır. Gerçekleştirilen deneyler sonucunda, Word2Vec ile Random Forest sınıflandırıcının en yüksek performans elde ettiği gözlenmiştir. Diğer bir çalışmada[5], istenmeyen mesajları filtrelemek için makine öğrenmesini kullanan içerik tabanlı bir sınıflandırma modeli önerilmiştir. Seçilen veri kümesinden Word2Vec kelime gömme aracı yardımıyla sınıflandırmada kullanılacak model oluşturulur. Bu model sayesinde, mesajların spam ve ham kelimelere olan mesafelerini hesaplamak için iki yeni özellik ortaya çıkıyor. Bu iki yeni özellik dikkate alınarak sınıflandırma algoritmalarının performansları karşılaştırılmıştır. Random Forest sınıflandırıcı, %99,64'lük bir doğruluk oranıyla başarılı oldu. Aynı veri kümesini kullanan diğer çalışmalara göre daha başarılı doğru sınıflandırma yüzdesine ulaşılmaktadır. Ayrıca, Ballı ve Karasoy[6], Türkçe için istenmeyen mesajları filtrelemek için makine öğrenmesi ve derin öğrenme yöntemleri kullanılarak içerik tabanlı SMS sınıflandırması yaptılar. TurkishSMS veri seti, farklı yaş gruplarından ve kişilerin bölgelerinden gelen mesajlar toplanarak hazırlanmıştır. TurkishSMS veri setinde beş farklı yapısal özellik, Word2Vec ile bulunan iki yeni özellik ve her mesajın kelime indeks değerleri ile oluşturulan 45 özellik bulunmaktadır. Toplamda 52 özellikten oluşan özellik matrisi, geleneksel makine öğrenmesi algoritmalarının yanı sıra derin öğrenme algoritmaları ile de değerlendirilmiş ve sonuçlar karşılaştırılmıştır. Sonuç olarak, evrişimli sinir ağı en yüksek sınıflandırma performansı ile en başarılı algoritma olarak bulunmuştur.

Uysal ve arkadaşları[7], SMS mesajlarını sınıflandırmak için yeni bir şema önerdiler, böylece bilgilendirici öznitelikler önce ki-kare ve bilgi kazancına dayalı iki farklı öznitelik seçim tekniği kullanılarak araştırılır ve daha sonra SMS'yi sınıflandırmak için seçilen öznitelik setleri ile birlikte Bayes temelli sınıflandırma algoritmaları kullanılır. Ayrıca, önerilen sınıflandırma şemasını

kullanan SMS spam filtrelemesi için gerçek zamanlı bir mobil uygulama geliştirildi. Ayrıca, önerilen sınıflandırma şemasını kullanan SMS spam filtreleme için gerçek zamanlı bir mobil uygulama, daha önce [8]'de tanıtılan Android uygulamasına dayalı olarak geliştirilmiştir. Filtreleme çerçevesi, yasal ve istenmeyen İngilizce SMS mesajlarını içeren geniş bir veri kümesi üzerinde değerlendirildi. Deneysel sonuçlar, SMS spam mesajlarının filtrelenmesinde oldukça yüksek doğruluk elde edildiğini açıkça göstermektedir. Uysal ve diğerleri[9], çeşitli öznitelik çıkarma ve öznitelik seçme yöntemlerinin birlikte Türkçe ve İngilizce olmak üzere iki farklı dilde SMS filtreleme üzerindeki etkilerini kapsamlı bir şekilde analizini gerçekleştirdiler. Filtreleme şemasının tüm özellik seti, kelime torbası (BoW) modelinden[10] kaynaklanan özniteliklerden ve ayrıca spam sorunu için benimsenen bir yapısal öznitelikler grubundan oluşur. Kelime torbası modeline dayalı ayırt edici özellikler, ki-kare ve Gini indeks tabanlı öznitelik seçim yöntemleri kullanılarak belirlenir. Seçilen öznitelikler daha sonra yapısal özniteliklerle birleştirilir ve SMS mesajlarını istenmeyen veya normal olarak sınıflandırmak için k-en yakın komşu ve destek vektör makinesi olmak üzere iki farklı model sınıflandırma algoritmasıyla beslenir. Filtreleme çerçevesi, sırasıyla Türkçe ve İngilizce mesajlardan oluşan iki ayrı SMS mesaj veri seti üzerinde değerlendirilir. Bu amaçla, çalışmanın bir parçası olarak, kamuya açık ilk Türkçe SMS mesaj koleksiyonu oluşturulmuş ve ayrıca İngilizce olarak mevcut bir veri seti kullanılmıştır. Her iki veri kümesi üzerinde kapsamlı deneysel analiz, tek başına kelime çantası özniteliklerinden ziyade Kelime çantası ve yapısal öznitelikler kombinasyonlarının daha iyi sınıflandırma performansı sağladığını ortaya çıkardı. Bununla birlikte, öznitelik seçme yöntemlerinin etkinliği her dilde biraz farklılık göstermiştir.

2. Veri Kümesi

Her ne kadar çok sayıda e-posta veri kümesi[11, 12] araştırmacıların kullanımına sunulmuş olsa da, literatürde yalnızca sınırlı sayıda herkese açık SMS veri kümesi bulunmaktadır. Özellikle Türkçe dili için veri kümesi sayısı oldukça azdır. Bu nedenle bu çalışma kapsamında tüm dünyada yaygın olarak kullanılan sondan eklemeli dillerden biri olan Türkçe de daha önceden oluşturulmuş bir SMS mesaj koleksiyonu kullanılmıştır. Bu, akademik literatürdeki ilk Türkçe SMS veri kümesi[4] olduğundan oldukça önemlidir. Veri kümesi, gönüllülerden toplanan 420 istenmeyen ve 430 normal mesajdan oluşuyor. Deneylerde, veri kümesinin %50'si eğitim ve %50'si test için kullanılmıştır. Ayrıca, deneylerde eğitim veri kümesi ön-işleme adımlarından önce 2086 öznitelik açığa çıkarken, ön-işleme adımlarından sonra toplam 1728 öznitelik açığa çıkmaktadır. Veri kümesinin eğitim ve test kısmında bulunan doküman sayılarını Tablo 1'de gösterdik. Normal ve istenmeyen kategorisindeki toplam doküman sayısı birbirine yakın olduğundan veri kümesi dengeli veri kümesi olarak sayılabilir.

Tablo 1. Türkçe SMS(Turkish SMS) Veri Kümesi

	Eğitim	Test
Normal	215	215
İstenmeyen	210	210

3. Öznitelik Çıkarımı ve Seçimi

İstenmeyen SMS iletilerinin tespiti, aslında istenmeyen e-posta tespit probleminin bir alt kümesidir. Bir e-posta metin, grafik ve hatta ekli dosyalar[13] içerebilirken, bir SMS mesajı yalnızca 160 karakterle sınırlı metin içerir[14]. Bu anlamda SMS, e-posta'ya göre daha sınırlı veri içermektedir. Sonuç olarak, istenmeyen mesaj ya da e-postaların tespiti, sınıfların "istenmeyen" ve "normal" olarak tanımlandığı 2 sınıflı bir metin sınıflandırma problemine karşılık gelir.

3.1. Öznitelik çıkarımı

Çok sayıda metin sınıflandırma çalışması, dokümanlardaki kelimelerin veya terimlerin tam sırasının göz ardı edildiği ancak terimlerin frekansının dikkate alındığı metin dokümanlarını temsil etmek için kelime çantası (Bag-of-Words) modelini[15] kullanır. Bir doküman koleksiyonundaki her farklı terim, sonuç olarak, tek başına bir öznitelik oluşturur. Terimlere belirli bir dokümandaki önemlerini temsil edecek şekilde belirli ağırlıklar atanır[16]. En yaygın ağırlıklandırma şeması Terim Frekansı-Ters Doküman Frekansı (TF-IDF=Term Frequency- Inverse Document Frequency), bir dokümandaki bir terimin o terimi içeren doküman sayısını dikkate alarak bir terimin frekansını belirli bir aralığa getirir[17]. Bu nedenle, bir doküman çok boyutlu öznitelik vektörü ile temsil edilir; burada vektörün her boyutu, vektör uzay modeli olarak da bilinen doküman koleksiyonundaki farklı bir kelimenin ağırlık değerine karşılık gelir[18].

İstenmeyen SMS filtreleme, geleneksel metin sınıflandırma görevi olarak ele alınabilse bile, istenmeyen mesajların yapısı resmi metinlerden önemli ölçüde farklı olabilir. Bir SMS mesajının boyutu sadece 160 karakterle sınırlı olduğundan, hem mesaj uzunluğu hem de terim sayısı büyük önem taşımaktadır. Ayrıca büyük veya küçük harf kullanımı istenmeyen mesaj göstergesi olabilir. Benzer şekilde, bazı alfasayısal olmayan karakterlere (örneğin, "!", "$") ve sayısal karakterlere (örneğin, telefon numaraları) istenmeyen mesajlarda yaygın olarak rastlanır. Ek olarak, internet(URL=Uniform Resource Loader) bağlantıları genellikle istenmeyen SMS mesajlarında da gözlenir. Bu tür öznitelikler yapısal öznitelikler olarak tanımlanır. Ancak, bu çalışmada sadece kelime çantası modelindeki özniteliklerle çalışma yapılmıştır.

Öznitelik çıkarımı sırasında ön-işleme adımları olarak alfanümerik karakterlerin silinmesi, parçalara ayırma(tokenizasyon), kök bulma(stemming), gereksiz kelimelerin kaldırılması, küçük harf dönüşümünün gerçekleştirildiğine de dikkat edilmelidir. Bununla beraber deneylerde, kök bulma işlemi yapılmadan da öznitelik vektörleri oluşturulmuştur. Bu sayede ön-işleme adımlarından biri olan kök bulma işleminin, istenmeyen SMS tespitinde performansa etkisi incelendi. Bu çalışma kapsamında sadece Türkçe veri kümesi kullanıldığından, kök bulma aşaması Türkçe diline özgüdür. Türkçe mesajlar için Zemberek algoritması kullanılmıştır[19].

Belirteçleştirme(tokenization) ya da parçalara ayırma, bir metni sözcüklere, deyimlere veya diğer anlamlı parçalara, yani belirteçlere bölme işlemidir.

3.2. Öznitelik seçimi

Filtreler, sarmalayıcılar ve gömülü öznitelik seçme yöntemleri olmasına rağmen, araştırmacılar, filtrelerin sınıflandırıcı bağımsızlığı ve nispeten düşük hesaplama süresi nedeniyle özellikle metin sınıflandırma problemlerinde ayırt edici öznitelikleri seçmek için filtre temelli yöntemleri tercih etmektedir[20]. Bu çalışmada kullanılan filtre yöntemleri, Gini Index(GI)[21], Normalized Difference Measure(NDM)[22] ve Extensive Feature Selector(EFS)[23]. Her üç yöntemin de önceki metin sınıflandırma çalışmalarında oldukça başarılı olduğu kanıtlanmıştır.

Tablo 1. Öznitelik seçim yöntemleri için ön gösterimler

Notasyon	Anlamı
$p(t\|C_j)$	C_j sınıfı mevcut olduğunda t teriminin olasılığı
$p(\bar{t}\|C_j)$	C_j sınıfı mevcut olduğunda t teriminin olmama olasılığı
$p(t\|\bar{C_j})$	C_j sınıfı mevcut olmadığında t teriminin olma olasılığı
$p(\bar{t}\|\bar{C_j})$	C_j sınıfı mevcut olmadığında t teriminin olmama olasılığı
$p(C_j\|t)$	t terimi mevcut olduğunda C_j sınıfının olma olasılığı
$p(\bar{C_j}\|t)$	t terimi mevcut olduğunda C_j sınıfının olmama olasılığı
$p(C_j\|\bar{t})$	t terimi mevcut olmadığında C_j sınıfının olma olasılığı
$p(\bar{C_j}\|\bar{t})$	t terimi mevcut olmadığında C_j sınıfının olmama olasılığı

3.2.1. Gini index(GI)

GI, orijinal olarak karar ağaçlarındaki niteliklerin en iyi alt kümesini bulmak için kullanılan yöntemin geliştirilmiş bir versiyonu olan farklı bir öznitelik seçim yöntemidir[21]. Aşağıda verildiği gibi nispeten daha basit bir hesaplamaya sahiptir:

$$GI(t) = \sum_{j=1}^{M} P\left(t\middle|\, c_j\right) * P(c_j|t) \tag{1}$$

3.2.2. Normalleştirilmiş Fark Ölçütü(NDM=Normalized Difference Measure)

Metin sınıflandırması için yeni bir öznitelik sıralama metriği olarak kullanılmaktadır[22]. NDM, dengeli doğruluk ölçüsüne (ACC2) bir düzenleyici olarak minimum doküman frekansını sunar. NDM için ACC2 ölçütünü min(tpr, fpr) ile böleriz. Matematiksel olarak, NDM şu şekilde tanımlanır:

$$NDM(t) = \sum_{j=1}^{M} \frac{\left|P(t|\,c_j) - P(t|\overline{C}_j)\right|}{\min\left(P(t|\,c_j), P(t|\overline{C}_j)\right)} \quad (2)$$

3.2.3. Kapsamlı Öznitelik Seçici (EFS=Extensive Feature Selector)

Filtre tabanlı EFS metodu[23] özniteliğin hem sınıf temelli hem de koleksiyon temelli olasılıkları kullanarak daha ayırt edici öznitelikleri seçmektedir. Formülü diğer yöntemlere göre biraz daha karmaşıktır:

$$EFS(t) = \sum_{j=1}^{M} \left(\frac{P(t|C_j)}{P(\bar{t}|C_j) + P(t|\overline{C}_j) + 1}\right) \cdot \left(\frac{P(C_j|t)}{P(\overline{C}_j|t) + P(C_j|\bar{t}) + 1}\right) \quad (3)$$

4. Sınıflandırma

4.1. Sınıflandırıcılar

MNB, metin sınıflandırma çalışmalarında en verimli sınıflandırıcılardan biridir[24]. Ayrıca MNB, Naive Bayes sınıflandırıcısının bir çeşididir. Geleneksel Naive Bayes bir metindeki ilgili öznitelikleri dahil etmeden modellerken, MNB öznitelik sayılarını kullanarak açıkça modeller. Çok değişkenli Bernoulli olay modelleri, metin sınıflandırma alanı için çok terimli modelin yanı sıra yaygın olarak kullanılmaktadır. Çok değişkenli Bernoulli olay modeli doküman frekanslarını kullanırken, MNB terim frekanslarını dikkate alır.

DVM, metin sınıflandırma çalışmalarında en verimli sınıflandırıcılardan biridir. DVM[25] sınıflandırıcının doğrusal ve doğrusal olmayan iki versiyonu vardır. DVM sınıflandırıcısının odak noktası marj kavramıdır. Sınıflandırıcılar tarafından sınıfları ayırmak için hiper düzlemler kullanılmıştır. Doğrusal çekirdek içeren SVM için LibSVM kütüphanesi kullanılır.

4.1. Deneysel çalışmalar

Deneysel çalışmada, ön-işleme, öznitelik çıkarma, öznitelik seçimi ve örüntü sınıflandırma yöntemlerinin Türkçe istenmeyen SMS mesajlarının filtrelenmesine etkileri incelenmiştir. Deneylerde ön-işleme yaparken toplam öznitelik sayısı 1728 iken, ön-işleme yapmadan 2086 öznitelik açığa çıkmıştır. Ayrıca 50, 100, 300, 500 ve 1000 olmak üzere 5 farklı boyutta performansın nasıl değiştiği gözlenmiştir.

Öznitelik kümeleri daha sonra MNB ve SVM sınıflandırıcılarıyla beslenmiştir. Veri kümesi de dengeli olduğundan (yani istenmeyen ve normal sınıflarındaki SMS mesajlarının sayısı neredeyse eşit), sınıflandırma performansını değerlendirmek için iyi bilinen F skoru [23] kullanıldı. Sınıflandırma sonuçları sırasıyla Türkçe veri kümesi için Tablo 2-5'te sunulmuştur. Sonuçlar, veri kümesini objektif olarak değerlendirmek için eğitim ve test bölümlerini %50-%50 olacak şekilde bölümlenmiştir.

En yüksek ağırlıklı ortalama F skorları göz önüne alındığında, çoğu durumda GI, NDM ve EFS metotları farklı boyutlarda ve sınıflandırıcıda en yüksek performansı göstermektedirler. Bununla birlikte ön-işleme tekniğinin uygulanmaması, MNB ve DVM sınıflandırıcılarında küçük boyutlarda daha yüksek performans elde edilmektedir. EFS metodu, hem ön-işleme yapılmadan hem de ön-işleme yapılarak MNB sınıflandırıcıda en yüksek performansı göstermiştir. Ancak, DVM sınıflandırıcıda ise, ön-işleme yapıldığında en yüksek performansı NDM metodu gösterirken, ön-işleme yapılmadığında GI metodu en yüksek performansı göstermiştir. Bununla birlikte, MNB sınıflandırıcıda, öznitelik boyutu yüksek olduğunda GI, NDM, EFS daha başarılı görünmektedir. Ancak, DVM sınıflandırıcıda ise, öznitelik boyutu düşük olduğunda GI, NDM, EFS daha başarılı görünmektedir.

Sınıflandırma algoritmalarının performansları göz önüne alındığında, MNB, DVM sınıflandırıcısından daha yüksek performans sergilemiştir. Boyutlar bazında bakıldığında DVM ve MNB farklı boyutlarda farklı öznitelik seçim metotları kombinasyonuyla en yüksek performansı sergilemektedirler. Ancak, MNB en yüksek performansı genellikle EFS kombinasyonuyla göstermiştir.

Öznitelik Seçim Metotlarına göre en iyi 10 öznitelik Tablo 6-7'de gösterilmiştir. GI ve EFS öznitelik skorları farklı olsa bile en iyi 10 öznitelik birbirinin aynısıdır. Ayrıca, NDM metodu da bu iki metoda benzer birkaç öznitelik üretmiştir. Bu benzerlik ön-işleme yapıldığında 6 iken, ön-işleme yapılmadığında 5 olmuştur.

Table 2. MNB Sınıflandırıcı ile Türkçe SMS(Turkish SMS) Veri Kümesi için F Skorları

Türkçe SMS	50	100	300	500	1000
GINI	0.918	0.920	0.951	0.957	0.957
NDM	0.934	0.934	0.953	0.957	0.953
EFS	0.920	0.927	0.953	0.957	**0.958**

Table 3. MNB Sınıflandırıcı ile ön-işleme yapılmadan Türkçe SMS(Turkish SMS) Veri Kümesi için F Skorları

Türkçe SMS	50	100	300	500	1000
GINI	0.922	0.946	0.953	0.955	0.946
NDM	0.939	0.934	0.955	0.953	0.946
EFS	0.922	0.939	0.955	**0.958**	0.946

Table 4. DVM Sınıflandırıcı ile Türkçe SMS(Turkish SMS) Veri Kümesi için F Skorları

Türkçe SMS	50	100	300	500	1000
GINI	0.891	0.924	0.941	0.936	0.922
NDM	0.924	0.924	**0.943**	0.927	0.929
EFS	0.917	0.927	0.936	0.936	0.922

Table 5. DVM Sınıflandırıcı ile ön-işleme yapılmadan Türkçe SMS(Turkish SMS) Veri Kümesi için F Skorları

Türkçe SMS	50	100	300	500	1000
GINI	0.906	0.922	**0.948**	0.936	0.943
NDM	0.929	0.910	0.934	0.934	0.936
EFS	0.910	0.927	0.943	0.936	0.943

Table 6. Öznitelik Seçim Metotlarına göre en iyi 10 öznitelik

Türkçe SMS	GINI	NDM	EFS
1	tl	indir	tl
2	com	ozel	com
3	indir	firsati	indir
4	icin	ay	icin
5	ozel	taksit	ozel
6	sadece	**hemen**	sadece
7	firsati	**gonderin**	firsati
8	ay	**cardfinans**	ay
9	taksit	com	taksit
10	ye	**yazip**	ye

Table 7. Öznitelik Seçim Metotlarına göre ön-işleme yapılmadan en iyi 10 öznitelik

Türkçe SMS	GINI	NDM	EFS
1	tl	ozel	tl
2	com	firsati	com
3	icin	indirim	icin
4	ozel	hemen	ozel
5	sadece	**hediye**	sadece
6	firsati	**gonderin**	firsati
7	indirim	**cardfinans**	indirim
8	ye	com	ye
9	tr	**yazip**	hemen
10	hemen	**ayda**	tr

Sonuçlar

Bu bildiride, özellikle Türkçe dili için istenmeyen SMS filtrelemesine ön-işleme ve öznitelik seçme yöntemlerinin etkisi, sınıflandırma doğruluğu açısından kapsamlı bir şekilde incelenmiştir. Öznitelik vektörü oluşturulurken sadece kelime çantası yaklaşımı kullanıldı. Öte yandan, kullanılan öznitelik seçme stratejilerinin etkinliği iki farklı sınıflandırıcıda test edildi. Sondan eklemeli dillerin önde gelen örnekleri Türkçe olduğundan, bu çalışmanın sonucu benzer özelliklere sahip diğer diller için de bir gösterge olabilir.

Kelime çantası yaklaşımının yanı sıra, yeni yapısal özniteliklerin incelenmesi, SMS spam filtreleme probleminde diğer öznitelik seçimi ve sınıflandırma yöntemlerinin değerlendirilmesi gelecekteki ilginç çalışmalar olarak kalacaktır. Ayrıca, farklı örüntü sınıflandırıcılar dahil edilerek istenmeyen SMS filtreleme görevine dahil edilebilir.

Kaynaklar

[1] Sjarif, N.N.A., et al., SMS Spam Message Detection using Term Frequency-Inverse Document Frequency and Random Forest Algorithm. Procedia Computer Science, 2019. 161: p. 509-515.

[2] Nagwani, N.K. and A. Sharaff, SMS spam filtering and thread identification using bi-level text classification and clustering techniques. Journal of Information Science, 2017. 43(1): p. 75-87.

[3] Karasoy, O. and S. Ballı. Classification Turkish SMS with deep learning tool Word2Vec. in 2017 International Conference on Computer Science and Engineering (UBMK). 2017. Ieee.

[4] Uysal, A.K., et al., The impact of feature extraction and selection on SMS spam filtering. Elektronika ir Elektrotechnika, 2013. 19(5): p. 67-72.

[5] Ballı, S. and O. Karasoy, Development of content-based SMS classification application by using Word2Vec-based feature extraction. IET Software, 2019. 13(4): p. 295-304.

[6] Karasoy, O. and S. Ballı, Spam SMS detection for Turkish language with deep text analysis and deep learning methods. Arabian Journal for Science and Engineering, 2021: p. 1-17.

[7] Uysal, A.K., et al. A novel framework for SMS spam filtering. in 2012 International Symposium on Innovations in Intelligent Systems and Applications. 2012. IEEE.

[8] Uysal, A.K., et al. Detection of SMS spam messages on mobile phones. in 2012 20th Signal Processing and Communications Applications Conference (SIU). 2012. Ieee.

[9] Uysal, A.K., et al., The impact of feature extraction and selection on SMS spam filtering. Elektronika ir Elektrotechnika, 2012. 19(5): p. 67-72.

[10] Parlak, B. and A.K. Uysal, The effects of globalisation techniques on feature selection for text classification. Journal of Information Science, 2020: p. 0165551520930897.

[11] Uysal, A.K. and S. Gunal, The impact of preprocessing on text classification. Information Processing & Management, 2014. 50(1): p. 104-112.

[12] Bhowmick, A. and S.M. Hazarika, E-Mail Spam Filtering: A Review of Techniques and Trends, in Advances in Electronics, Communication and Computing. 2018, Springer. p. 583-590.

[13] Venkatraman, S., B. Surendiran, and P.A.R. Kumar, Spam e-mail classification for the internet of things environment using semantic similarity approach. The Journal of Supercomputing, 2020. 76(2): p. 756-776.

[14] Roy, P.K., J.P. Singh, and S. Banerjee, Deep learning to filter SMS spam. Future Generation Computer Systems, 2020. 102: p. 524-533.

[15] Li, P., et al., Bag-of-Concepts representation for document classification based on automatic knowledge acquisition from probabilistic knowledge base. Knowledge-Based Systems, 2020. 193: p. 105436.

[16] Salton, G. and C. Buckley, Term-weighting approaches in automatic text retrieval. Information processing & management, 1988. 24(5): p. 513-523.

[17] Schütze, H., C.D. Manning, and P. Raghavan, Introduction to information retrieval. Vol. 39. 2008: Cambridge University Press.

[18] Al-Anzi, F.S. and D. AbuZeina, Beyond vector space model for hierarchical Arabic text classification: A Markov chain approach. Information Processing & Management, 2018. 54(1): p. 105-115.

[19] Akın, A.A. and M.D. Akın, Zemberek, an open source NLP framework for Turkic languages. Structure, 2007. 10: p. 1-5.
[20] Forman, G., An extensive empirical study of feature selection metrics for text classification. Journal of machine learning research, 2003. 3(Mar): p. 1289-1305.
[21] Singh, S.R., H.A. Murthy, and T.A. Gonsalves, Feature Selection for Text Classification Based on Gini Coefficient of Inequality. Fsdm, 2010. 10: p. 76-85.
[22] Rehman, A., K. Javed, and H.A. Babri, Feature selection based on a normalized difference measure for text classification. Information Processing & Management, 2017. 53(2): p. 473-489.
[23] Parlak, B. and A.K. Uysal, A novel filter feature selection method for text classification: Extensive Feature Selector. Journal of Information Science, 2021: p. 0165551521991037.
[24] Zhao, L., et al. Semi-supervised Multinomial Naive Bayes for text classification by leveraging word-level statistical constraint. in Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence. 2016. AAAI Press.
[25] Gabrilovich, E. and S. Markovitch. Text categorization with many redundant features: using aggressive feature selection to make SVMs competitive with C4. 5. in Proceedings of the twenty-first international conference on Machine learning. 2004. ACM.