Karşılıklı Bilgi Kullanılarak Metin Sınıflandırma İçin Özellik Seçimi

Feature Selection for Text Classification Using Mutual Information

İlhami SEL
Bilgisayar Mühendisliği Bölümü
İnönü Üniversitesi
Malatya,Türkiye
ilhamisel23@gmail.com

Ali KARCI
Bilgisayar Mühendisliği Bölümü
İnönü Üniversitesi
Malatya,Türkiye
ali.karci@inonu.edu.tr

Davut HANBAY
Bilgisayar Mühendisliği Bölümü
İnönü Üniversitesi
Malatya,Türkiye
davut.hanbay@inonu.edu.tr

***Özet*— Özellik seçimi, veri setini temsil edebilecek en iyi alt kümenin seçimi, yani sonucu etkilemeyen gereksiz verilerin çıkarılması olarak tanımlanabilir. Sınıflandırma uygulamalarında özellik seçimi ile boyut düşürüldüğünde sistemin verimliliği ve doğruluğu artırılabilir. Bu çalışmada Reuters haber ajansının yayınlamış olduğu "20 news group" verileri kullanılarak metin sınıflandırma uygulaması yapılmıştır. Ön işlemlerden geçen haber verileri Doc2Vec yöntemi kullanılarak vektörlere dönüştürülmüş ve veri seti oluşturulmuştur. Bu veri seti Maximum Entropy Sınıflandırma yöntemiyle sınıflandırılmıştır. Sonrasında ise özellik seçimi için Karşılıklı Bilgi yöntemi kullanılarak veri seti alt kümesi oluşturulmuştur. Oluşan veri setiyle tekrar sınıflandırma işlemi uygulanıp sonuçlar başarım oranlarına göre karşılaştırılmıştır. Özellik seçiminden önce 600 özelliğe sahip sistemin başarısı (0.9285) iken sonrasında oluşturulan 200, 100, 50, 20 özellikli modellerin başarım oranı sırasıyla (0.9454, 0.9426, 0.9407, 0.9123) çıkmıştır. Sonuçlar incelendiğinde 50 özellikli modelin başarısı başlangıçta oluşturulan 600 özellikli modelden daha yüksek çıkmıştır.**

***Anahtar Kelimeler*— Doğal Dil İşleme, Doc2Vec, Karşılıklı Bilgi, Maximum Entropy**

Abstract— **The feature selection can be defined as the selection of the best subset to represent the data set, that is, the removal of unnecessary data that does not affect the result. The efficiency and accuracy of the system can be increased by decreasing the size and the feature selection in classification applications. In this study, text classification was applied by using "20 news group" data published by Reuters news agency. The pre-processed news data were converted into vectors using the Doc2Vec method and a data set was created. This data set is classified by the Maximum Entropy Classification method. Afterwards, a subset of data sets was created by using the Mutual Information Method for the feature selection. Reclassification was performed with the resulting data set and the results were compared according to the performance rates. While the success of the system with 600 features was (0.9285) before the feature selection, (0.9285), then, the performance rates of the 200, 100, 50, 20 models were obtained as (0.9454, 0.9426, 0.9407, 0.9123), respectively. When the results were examined, the success of the 50-featured model was higher than the 600-featured model initially created.**

Key Words— **Natural Language Processing, Doc2Vec, Mutual Information, Maximum Entropy**

I. GİRİŞ

Doğal Dil İşleme (DDİ) alanındaki gelişmelerle birlikte, konuşma dilinin bilgisayar tarafından anlamlandırılması, yorumlanması ve gerektiğinde tekrar üretilmesi gibi çalışmalar yoğun bir şekilde devam etmektedir. Sosyal medya, haber siteleri, kullanıcı yorumları gibi doğal dilin temel özelliklerini taşıyan metinlerin artmasıyla oluşan büyük boyutlardaki verilerden önemli verinin çıkarılması da DDİ'nin çalışma alanlarından bir tanesidir. Metin sınıflandırma; bir metinin veya dokümanın daha önceden belirlenmiş kategorilerden birisine atanması olarak tanımlanabilir. Bu çalışma da Reuters haber ajansının yayınlamış olduğu "20 news group" isimli 20 gruptan ve 18878 dokümandan oluşan veri setinin sınıflandırılması amaçlanmıştır. Dokümanların vektöre çevrilmesi için mikolov ve ark. [1,2] geliştirmiş oldukları Doc2Vec algoritması kullanılmıştır. Sonrasında Karşılıklı Bilgi (KB) yöntemiyle Özellik Seçimi (ÖS) yapılmış oluşturulan modellerin başarım oranları karşılaştırılmıştır.

Bu çalışmanın ikinci bölümünde geçmiş çalışmalardan bahsedilmiş, üçüncü bölümde DDİ yöntemleri, dördüncü bölümde metin sınıflandırma yöntemleri açıklanmış, beşinci bölümde özellik seçimi anlatılmış, altıncı bölümde uygulama anlatılmış ve son bölümde ise sonuçlar hakkında değerlendirmeler yapılmıştır.

II. İLGİLİ ÇALIŞMALAR

DDİ ile ilgili [3], Metin Sınıflandırma [4] ve Özellik Seçimi alanlarında [5] çalışmaları ulaşabildiğimiz en güncel ve kapsamlı literatür araştırmalarıdır. Metin Sınıflandırmanın özellikleri ve yöntemleri üzerinde yapılan çalışmalar [6,8] verilmiştir.

Diğer çalışmaların yanında özellik seçimi [9], Metin Sınıflandırma da ÖS [10] ve Karşılıklı Bilgi ile ÖS [11] için çalışmalar yapılmıştır.

III. DOĞAL DİL İŞLEME

A. Veri Seti ve Ön İşlemler: Bu çalışma için Reuters haber ajansından alınan "20 news group" [12] data seti kullanılmıştır. Konu başlıkları ve haber sayıları Tablo 1 de verilmektedir.

TABLO 1: VERİ SETİ KONULARI VE HABER SAYILARI

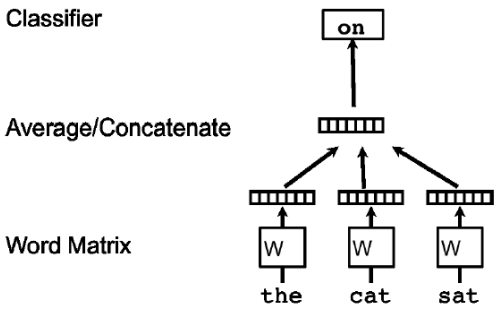
Konu Başlıkları: 20 Sınıf	*Doküman Sayısı: 18878*
comp.graphics	973
comp.os.ms-windows.misc	985
comp.sys.ibm.pc.hardware	982
comp.sys.mac.hardware	961
comp.windows.x	980
misc.forsale	972
rec.autos	990
rec.motorcycles	994
rec.sport.baseball	994
rec.sport.hockey	999
talk.politics.misc	775
talk.politics.guns	910
talk.politics.mideast	940
sci.crypt	991
sci.electronics	981
sci.med	990
sci.space	987
talk.religion.misc	628
alt.atheism	799
soc.religion.christian	997

Metinler üzerinde sınıflandırma işlemi yapılmadan önce bazı ön işlemlere ihtiyaç duyulur. Bunlar şu şekilde özetlenebilir;

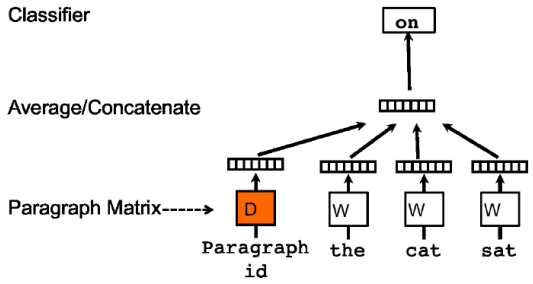
a. Küçük harf dönüşümü
b. Noktalama işaretlerinin çıkarılması
c. Tokenization (Kelimelere ayrılması)
d. Stop Words [13] (gereksiz kelimelerin çıkarılması)
e. Stemming (kök alma işlemi)

Yapmış olduğumuz uygulama için kök alma işlemi dışında diğer ön işlemler uygulanmış ve dokümanlar sınıflandırma işlemi için hazırlanmıştır. Metinlerin vektörel gösterimi için Doc2Vec algoritması kullanılmıştır.

Doc2Vec: Doğal dil işlemede kelimelerin vektör ile temsili en yaygın kullanılan yöntemlerden birisidir [1,2]. Word2Vec ve Doc2Vec yöntemi Mikolov ve arkadaşları tarafından geliştirilen ve yapay sinir ağlarını kullanarak kelimelerin vektör şekline dönüştürülmesini sağlayan algoritmalardır. Şekil-1 ve Şekil-2 de bu yöntemlerin nasıl oluşturuldukları verilmektedir.



Şekil -1: Kelime vektörünün öğrenilmesi için bir çerçeve [2]

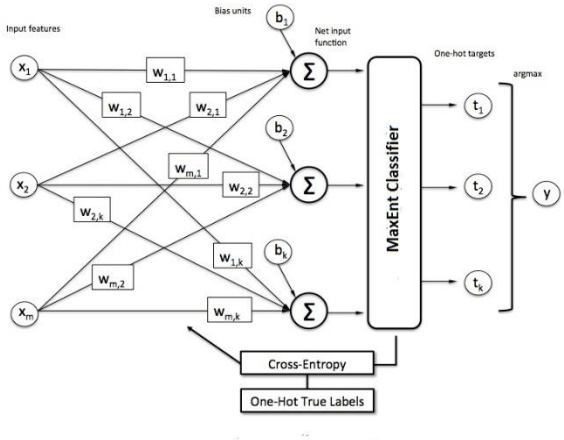


Şekil -2: Paragraf vektörünün öğrenilmesi için bir çerçeve [2]

IV. METİN SINIFLANDIRMA

Makine öğrenmesinde Karar ağaçları, Destek Vektör Makineleri, Maximum Entropy Classifier ve Naive Bayes gibi sınıflandırma algoritmaları sıkça kullanılmaktadır. Bu çalışmada sınıflandırma algoritması olarak Maximum Entropy yöntemi kullanılmıştır.

Maximum Entropy Sınıflandırıcı (MaxEnt Classifier): Maximum Entropy Sınıflandırıcı (MES) tüm sınıflandırma problemlerinde sıkça kullanılmaktadır. Multinomial Logistic Regresyon (MLR) veya softmax regresyon (Şekil-3) olarak ta bilinen sınıflandırıcı Naive Bayes gibi yöntemlerle karşılaştırıldığında [14] daha iyi sonuçlar verdiği görülmüştür.



Şekil-3: Maximum Entropy Sınıflandırıcı [15]

Bu sınıflandırıcı yaygın olarak ikiden fazla sınıfın bulunduğu problemlerde kullanılmaktadır. İkili lojistik

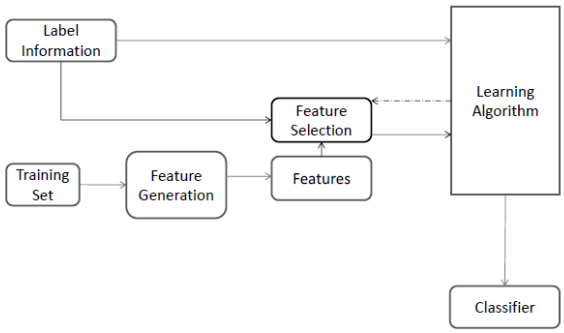
regresyon gibi, çok sınıflı lojistik regresyon da kategorik üyelik olasılığını değerlendirmek için maksimum olasılık tahmini kullanır. Matematiksel ifadesi Denklem 1,2 de verilmiştir.

$$P(y = j\,|\theta^i) = \frac{e^{\theta^i}}{\sum_{j=0}^{k} e^{\theta_k^{(i)}}} \quad (1)$$

$$where\ \theta = W_0X_0 + W_1X_1 + \cdots + W_kX_k = \sum_{i=0}^{k} W_iX_i = W^T X \quad (2)$$

V. Özellik Seçimi

Özellik seçimi, çok sayıda özellik içeren veri setlerinde sonucu etkilemeyen değişkenlerin çıkarılması olarak tanımlanabilir. Özellik seçiminin 3 temel amacı vardır. Bunlar: Sınıflandırıcıların performansını iyileştirmek, daha hızlı ve daha uygun maliyetli tahminler sağlamak ve verileri oluşturan temel sürecin daha iyi anlaşılmasını sağlamak [9]. Genel bir özellik seçim çerçevesi Şekil-4'te verilmiştir [16].



Şekil 4: Sınıflandırma uygulamalarında genel bir özellik seçim çerçevesi [16]

Karşılıklı Bilgi (KB): Bilgi teorisinin temel kavramlarından biri olan KB temeli düzensizliğe (entropy) dayanır. Düzensizlik bir rastgele değişkendeki belirsizliği ölçer. Bu ölçüm Denklem 3'tedir [17].

$$H(x) = -\sum_{i=1}^{n} p(x_i)\,\mathrm{LOG}(p(x_i)) \quad (3)$$

Burada x ayrık verilerden oluşan rastgele bir değişken olsun. x_i bu değişkendeki farklı verileri temsil etsin. $p(x_i)$ bu farklı verilerin olasılığıdır. Düzensizlik 0 ve 1 arasında değer üreten bir ölçümdür. KB ise iki rastgele değişken arasındaki paylaşılan bilginin ölçümünü verir. Bu ölçüm Denklem 2'dedir [17].

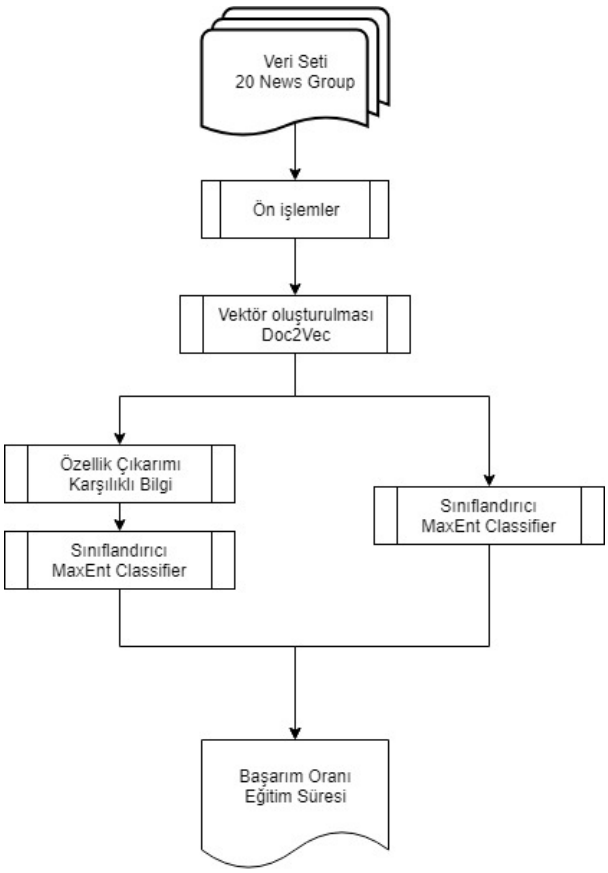
$$I(X;Y) = \sum_{x\epsilon X}\sum_{y\epsilon Y} p(x,y)\mathrm{LOG}\left(\frac{p(x,y)}{p(x)p(y)}\right) \quad (4)$$

VI. Uygulama

Bu çalışmada doğal dil işleme yöntemleri kullanılarak bir metin sınıflandırma sistemi tasarlanmıştır. Çalışma için Reuters haber ajansının "20 news group" veri seti kullanılmıştır. Metin olarak alınan veri ön işlemlerden (küçük harf dönüşümü, noktalama işaretleri ve gereksiz kelimelerin çıkarılması gibi) geçirilip Doc2Vec yöntemi kullanılarak 600 özellikli 18878 boyutunda matris oluşturulmuştur. Oluşan matris MaxEnt Sınıflandırıcı algoritması kullanılarak (%70 eğitim, %30 test) modelin başarısı ve eğitim süresi ölçülmüştür.

Sonrasında Özellik Seçimi için KB yöntemi kullanılmış KB puanı en yüksek olan özelliklerden sırasıyla 20-50-100-200 olan boyutlarda özellik alt kümeleri oluşturulmuştur. Oluşturulan her bir model için sistemin başarısı ve eğitim süreleri tekrar hesaplanarak karşılaştırma yapılmıştır. 200 özellikten sonra sistemin başarı oranı 10.000 de 1 oranında arttığından daha yüksek boyutlar seçilmemiştir. Çıkan sonuçlar tablo 5-6-7 'de gösterilmiştir.

Uygulama İntel i5 2.3Ghz işlemci, 16gb ram ve m2 ssd bulunan bilgisayarda çalıştırılmıştır. Çalışmanın kodlanmasında ise Python dili ve numpy, gensim, sklearn ve nltk kütüphaneleri kullanılmıştır. Uygulamanın genel yapısı Şekil 5'te verilmiştir.



Şekil 5: Uygulama genel yapısı

Özellik çıkarımı için kullanılan KB yönteminde her bir özelliğin ağırlığının bulunma süresi 99.49sn olmuştur. Tablo 2'de ise özellik sayılarına göre modellerin eğitim süreleri verilmiştir. 600 özellikli model ÖS uygulanmadan oluşturulan modeldir.

TABLO 2: Özellik Sayıları ve Eğitim Süreleri

600	200	100	50	20
35.53sn	12.23sn	5.85sn	3.79sn	1.02sn

Tablo 3'de oluşturulan modellerin özellik sayıları ve başarım oranları verilmektedir.

TABLO 3: ÖZELLİK SAYILARINA GÖRE MODELLERİN BAŞARIM ORANLARI

600	200	100	50	20
0.9285	0.9454	0.9426	0.9407	0.9123

Tablo 4'te modellerin precision, recall, f1-score değerleri karşılaştırmalı olarak verilmiştir.

TABLO 4: ÖZELLİK SAYILARINA GÖRE MODELLERİN PRECİSİON, RECALL, F1-SCORE DEĞERLERİ

	600	200	100	50	20
precision	0.93	0.95	0.94	0.94	0.91
recall	0.93	0.95	0.94	0.94	0.91
f1-score	0.93	0.95	0.94	0.94	0.91

VII. SONUÇ

Doc2Vec algoritmasıyla oluşturulan 600 özellikli model ve karşılıklı bilgi kullanılarak oluşturulan modellerin başarım oranları Tablo-3'te verilmiştir. Sonuçlar incelendiğinde 600 özellikli ilk model 0.9285 başarım oranı sağlarken yeni oluşturulan modeller bu başarım oranını 50 özellikle geçebilmektedir. Özellik sayısını 200 olarak belirlendikten sonra ise başarım oranı 0.9454 ile en yüksek başarım oranına ulaşmıştır.

Tüm sonuçlar incelendiğinde Doc2Vec yöntemi metinlerin vektörle temsilinde oldukça başarılı olduğu fakat oluşturulan vektörün metin sınıflandırma işlemlerinde en yüksek başarıyı sağlayamadığı görülmüştür. Karşılıklı Bilgi yöntemi kullanılarak özellik alt kümesi seçildiğinde 50 özellik boyutunda dahi sistemin modelin başarısının arttığı gözlemlenmiştir. Özellik sayısı 200 olarak belirlendiğinde ise sistemin başarı düzeyi en üst seviyeye ulaşmıştır.

KAYNAKLAR

[1] Le, Quoc, and Tomas Mikolov. "Distributed representations of sentences and documents." International conference on machine learning. 2014.

[2] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean. Distributed Representations of Words and Phrases and their Compositionality. Accepted to NIPS 2013.

[3] T. Young, D. Hazarika, S. Poria, and E. Cambria, "Recent trends in deep learning based natural language processing," in arXiv preprint arXiv:1708.02709, 2017.

[4] Khan, Aurangzeb, et al. "A review of machine learning algorithms for text-documents classification." Journal of advances in information technology 1.1 (2010): 4-20.

[5] S. Alelyani, J. Tang, and H. Liu. Feature selection for clustering: A review. Data Clustering: Algorithms and Applications, Editor: Charu Aggarwal and Chandan Reddy, CRC Press, 2013.J.

[6] Sel, İlhami, and Davut, Hanbay. "E-Mail Classification Using Natural Language Processing", 27th Signal Processing and Communications Applications Conference (SIU). IEEE, 2019.

[7] G. Şahin, "Turkish document classification based on Word2Vec and SVM classifier," in 2017 25th Signal Processing and Communications Applications Conference (SIU), 2017, pp. 1–4

[8] Yang, Yiming, and Jan O. Pedersen. "A comparative study on feature selection in text categorization." Icml. Vol. 97. No. 412-420. 1997.

[9] Guyon, Isabelle, and André Elisseeff. "An introduction to variable and feature selection." Journal of machine learning research3.Mar (2003): 1157-1182.

[10] Forman, George. "An extensive empirical study of feature selection metrics for text classification." Journal of machine learning research 3.Mar (2003): 1289-1305.

[11] Xu, Yang, et al. "A study on mutual information-based feature selection for text categorization." Journal of Computational Information Systems 3.3 (2007): 1007-1012.

[12] The 20 Newsgroups data set, http://qwone.com/~jason/20Newsgroups, E.T: 25.04.2019

[13] Stopwords, https://www.ranks.nl/stopwords , E.T: 25.04.2019

[14] Nigam, Kamal, John Lafferty, and Andrew McCallum. "Using maximum entropy for text classification." IJCAI-99 workshop on machine learning for information filtering. Vol. 1. No. 1. 1999.

[15] https://sebastianraschka.com/faq/docs/softmax_regression.html

[16] Jin, Xin, et al. "Machine learning techniques and chi-square feature selection for cancer classification using SAGE gene expression profiles." International Workshop on Data Mining for Biomedical Applications. Springer, Berlin, Heidelberg, 2006

[17] ÇELİK, Ceyhun and BİLGE, Hasan Şakir. "Ağırlıklandırılmış Koşullu Karşılıklı Bilgi İle Öznitelik Seçimi." Gazi Üniversitesi Mühendislik-Mimarlık Fakültesi Dergisi 30.4 (2015).