TİDE Academia Research
Sayı 1 / 91 -110 (2020)

Research Article/Araştırma Makalesi

ULUSLARARASI HABER RAPORLARININ RAPOR İÇERİKLERİNDE KULLANILAN İFADELERE GÖRE MAKİNE ÖĞRENMESİ YÖNTEMİYLE SINIFLANDIRILMASI VE DENETLENMESİ: METİN MADENCİLİĞİ UYGULAMASI

Firdevs DURNAGÖL[1]

Submitted/Başvuru: 19.12.2019
Revised/Düzeltme: 08.06.2020
Accepted/Kabul: 05.07.2020

Öz

Rapor verisinin miktarının çok olması ve giderek artması, artan veri yoğunluğu içerisinde raporların anlamlandırılması ve arşivlenmesi zordur. Bu zorluğun aşılması, raporların denetlenmesi, düzenlenmesi ve düzeltilmesi, Karar Destek Sistemleri yollarından biri olan Makine Öğrenme ile aşılabilir. Raporların analiz edilmesi, anlamsız veriler arasından anlamlı verilerin çıkarılması, verinin kullanımı açısından büyük kolaylık sağlamaktadır. Bu yapılan araştırma, uluslararası yayın yapan büyük bir medya organının çevrimiçi olarak dünya çapında yayınladığı haber ve bilgi raporlarının makine öğrenme algoritmaları kullanılarak sınıflandırılmasına dayanmaktadır. Uygulamanın analiz aşamasında makine öğrenmesinin bir alt dalı olan metin madenciliği kullanılmaktadır. Analiz yaparken Rastgele Orman Karar Ağacı, ZeroR, Naif Bayes yöntemleri kullanılmıştır. Bu yöntemlerin sınıflandırma başarıları birbirleri ile karşılaştırılmıştır. Bunlar arasında en iyi sonuçları veren algoritma Rastgele Orman Karar Ağacı yönteminin dayandığı algoritmada parametrik

1 Yüksek Lisans Öğrencisi, İstanbul Aydın Üniversitesi Bilgisayar Mühendisliği, eser.firdevs@gmail.com, ORCID ID: 0000-0002-5411-1924.

To cite this article: Durnagöl, F. (2020). Uluslararası Haber Raporlarının Rapor İçeriklerinde Kullanılan İfadelere Göre Makine Öğrenmesi Yöntemiyle Sınıflandırılması ve Denetlenmesi: Metin Madenciliği Uygulaması. *TİDE AcademIA Research*, 2(1), 91-110

değişiklikler ve düzenlemeler yapılması sonucu rapor sınıflandırmada sonuçlarda yüksek iyileştirmeler elde edilmiştir. Başarı oranı %91'e ve performans süresi 0.47s'e çıkmıştır. Araştırmadaki veri seti içerisinde her birinden 600 rapor olacak şekilde üç adet sınıf, uluslararası konularda raporlar, spor raporları, dergi (magazin) raporlarıdır. Veri setinin bir kısmı eğitim ve bir kısmı test kümesi olarak kullanılmış, 10-katlı çapraz doğrulama yöntemi ile algoritmik doğruluklar denetlenmiştir. Bu sayede, veri seti, hem test hem de eğitim kümesi olarak kullanılmıştır. Derleme ortamı olarak Weka veri madenciliği yazılımı kullanılmıştır.

Anahtar Kelimeler: Rapor Denetleme, Sınıflandırma, Metin Madenciliği, Makine Öğrenmesi, Rastgele Orman Algoritması.

JEL Sınıflandırması: C88, M41, M42

CLASSIFICATION AND CONTROL OF INTERNATIONAL NEWS REPORTS ACCORDING TO EXPRESSIONS USED WITHIN REPORT CONTENTS THROUGH MACHINE LEARNING: TEXT MINING APPLICATION

Abstract

It is difficult to interpretation and archive the reports with large amount of report data and increasing the density of data. This hardship can be overcome by Data Mining. Analyzing the data, extracting meaningful data from meaningless data provides great convenience in terms of data usage. This study is a classification of the news that published on the web-site of an international channel by using artificial intelligence algorithm. Text Mining, a sub-branch of machine learning, is used in the analysis of the application. Random Forest Decision Tree Algorithm, ZeroR Algorithm and Naif Bayesian Algorithm have been used in the analysis phase of the application The results of classification algorithms have been compared with each other. The algorithm that has given the best result among them is the Random Forest Decision Tree Algorithm. The success rate has been found as 91% and the duration of application has been found as 0.47 seconds. There are three classes in the data-set. These are International News (600), Sports News (600), Magazine News (600). Some of the dataset has been used as training and some has been used as test dataset. Algorithm accuracy has been checked by 10-fold cross validation method. Thus, the entire dataset has been used as both test and training dataset. Weka has been used as the compilation tool.

Keywords: Report Control, Classification, Text Mining, Machine Learning, Random Forest Algorithm.

JEL Classification: C88, M41, M42

Extended Summary

Introduction

Today, with the development of technology, reports have been moved to the computer environment, and at the same time, the data are obtained from the Internet or cloud environment. Existent datum and speed of data increase cause data confusion. In an environment where auditing is difficult, data analysis eliminates this complexity. Data with complexity does not mean anything because it cannot be understood. The meaningful classification of the data makes it easier to understand the data by the recipient. Data analysis is therefore of great importance. Data mining enables us to access meaningful data from many meaningless data. Text mining, a sub-branch of data mining, is used in the study.

While doing the text mining classification process, the data is collected first, the model is created after the data preprocessing steps are applied, the result is evaluated and the created model is applied. If the result of the applied model is good, the results are evaluated, if not, the model is reinstalled from the beginning. The data set can be obtained from many places. It is carried out in the analysis on international news texts. International news is a news event that appeals to larger audiences. International news is a news event that appeals to larger audiences. For this, news must-have some features. These features are; the news includes elements of interest to a nation, about a nation-wide known person, war, terror, world news. Some of the national news agencies are Anadolu Agency, Magnum Photos, Reuters, Syrian Arab News Agency, Uriminzokkiri.

In the study to be carried out, the English news data published on the website of an international broadcasting channel was received. Data under three categories were collected from the website. These are: International Reports (600), Sports Reports (600), Journal (Magazine) Reports (600). There are no changes in the categories, the categories are gathered under three main titles and the news under the current category names have been received. The news are taken from the news published between 2019-2020. The news used in the analysis were chosen randomly. Analyzes were made on a total of 1800 text data. The data were first subjected to text preprocessing and feature extraction. Then some of the data was used as a test and some as a training data. Random Forest Decision Tree Al-

gorithm, ZeroR Algorithm, Naive Bayes Algorithm were used in the classification process. As a result of the literature review, the most preferred algorithms were selected within the scope of text mining and analyzes were made. To find the best classification success, the success rates of the algorithms are compared with each other. It is aimed to give better results by changing parameters within the algorithm that gives the best results. 10-fold cross-validation method was used as the verification method. The reason for this is to divide the data into 10 equal parts so that all data can be used both as a test and as a training. The purpose of the analysis is to identify the best classifying algorithm and the best runtime for the data set at hand. The purpose of the study is to help businesses by providing audits and analysis on meaningful data in a shorter time without entering into the data stack.

Literature Review

When the literature study on text mining is done, it is seen that the best result is given by Random Forest Algorithm. Some investigated studies are given below.

In the study named 'The Use of Data Mining for Authority Recognition in Classical Turkish Music' by Abidin (2017), has revealed that Random Forest Algorithm has done very well the classification process.

In the study named 'Author Recognition in Text Mining' by Ünal et al. (2016), decision tree, k-nearest neighborhood, naive bayes, random forest were used as classification algorithm. The best algorithm has been found as a random forest algorithm as a result of success.

In the study named "Developing Mobile Application for Content Based Spam Filtering and Comparing Classification Algorithms" by Karasoy, (2016), Bagging, Random Forest, Random Subspace classification algorithms were used. The algorithm that gives the best success result is Random Forest.

Methodology

The text preprocessing phase is required to prepare the current data set for analysis. This stage is very important for the accuracy of the analysis results. While there were 1935 features at the beginning, 73 features remained after the preprocessing phase was over.

Within the context of text preprocessing, data transformation, expressing words as numerical matrix, feature extraction, removing unnecessary cursors from the text, word length calculation, uppercase and lowercase conversions, stop words, root finding algorithm operations have been applied.

Three algorithms were used in the study. These algorithms are ZeroR, Naive Bayes and Random Forest Algorithm. In addition, analyzes were conducted within the data mining software Weka. The success rates and runtimes of the algorithms are compared with each other. The results are given in table 1.

Table 1: Comparison of Algorithm Results

Naive Bayes	ZeroR	Random Forest	Algorithm
87	34	89	Success Rates(%)
0.06	0.02	1.78	Runtime (sec)

In text mining, the algorithm that gives the best results on the current dataset is the Random Forest Algorithm. When the classification studies on text mining are examined in the literature, it is seen that RF works better than other algorithms. However, better results can be obtained by changing the parameters of the Random Forest Algorithm.

Table 2: Random Forest Parameter Selection

Number of Trees	Feature Selection	Success Rates(%)	Runtime (sec)
100	1	90.8	1.16
100	2	90.7	0.8
100	3	90.5	0.9
50	1	90.4	0.38
50	2	90.7	0.47
50	3	90.2	0.55
25	1	89.6	0.23
25	2	89.4	0.27
25	3	89.8	0.3

In the first stage, the number of trees = 100 and feature selection = 1, while the success rate is 89% and the runtime time is 1.78 seconds. When the number of trees = 50 and feature selection = 2, it was seen that Random Forest Decision Tree Algorithm gave the best

success rate. The success rate is 90.7%.Runtime = 0.47 sec. It has classified with a success rate of approximately 0.02 and a better time of 1.31 sec. When using the Random Forest algorithm in text mining, better results are obtained if the number of trees and feature selection variables are changed among the parameters.

Conclusion

Classification was made over 1800 English news texts. There are three algorithms in the study. These algorithms are: ZeroR, Naive Bayes, Random Forest. When the success rates of the algorithms are compared, it is determined that the Random Forest Algorithm works better than the other algorithms, with a runtime of about 89% and 1.78 sec. However, this success rate and runtime can be increased by changing the parameters in Random Forest Algorithm. For this reason, "number of trees" and "feature selection" parameters have been changed. As a result of this change, the success rate was found to be 91% and the runtime was 0.47 seconds. Random Forest Algorithm in text mining gives better results than other algorithms. If the number of trees is selected as 50 and feature selection is 2, the success rate and runtime are also improved.

1. Giriş

Günümüzde teknolojinin gelişmesi ile raporlar bilgisayar ortamına taşınmış, aynı zamanda veriler de İnternet ya da bulut ortamından elde edilmektedir. Mevcut verilerin ve veri artış hızının fazlalığı veri karmaşıklığına sebebiyet vermektedir. Denetlemenin zorlaştığı bu ortamda veri analizi bu karmaşıklığı ortadan kaldırmaktadır. Karmaşıklığa sahip olan veri, anlaşılamayacağı için bir şey ifade etmemektedir. Verilerin anlamlı olarak sınıflandırılması, verinin alıcı kişi tarafından daha kolay anlaşılmasını sağlamaktadır. Veri analizi bu yüzden büyük önem taşımaktadır. Veri madenciliği birçok anlamsız veri içerisinden anlamlı veriye ulaşmamızı sağlamaktadır. Bu da zamanda tasarruf sağlamaktadır. Bir saatte yapılacak bir işlem, alakasız verilerin çokluğu, anlamlı verilerin anlamsız veriler içerisinde kaybolmasından dolayı aylar ya da yıllar sürebilmektedir. Zamandan tasarruf bütün projeler için çok önemlidir. Zamandan tasarruf etmek maliyetten de tasarruf etmek anlamına gelmektedir. Veri her yerde mevcuttur. Bu sebeple veri madenciliğinin alanı çok geniştir. Tıp, ekonomi, bankacılık, teknoloji vs. birçok alandan bu konu hakkında söz edilebilir. (Ertuğrul ve diğerleri, 2012).

Veri madenciliği yapılırken veri setine göre modeller oluşturulur. Veriyi iyi anlayıp, verinin akışına göre model oluşturulması gerekmektedir. Bu modellerin içerisinde Yapay Zeka Algoritmaları kullanılmaktadır. (Yıldız ve diğerleri, 2016).

Veri madenciliği tamamlayıcı ve tahmin edici modeller olarak ikiye ayrılmaktadır. Tamamlayıcı veri modelleri içerisinde kümeleme ve birliktelik kuralları bulunur. Tahmin edici veri modelinde ise, sınıflandırma ve regresyon analizleri bulunur. Sınıflandırma analizi, daha önceden elde edilen gruplar üzerinden yapılmaktadır. Eğitim verisinde gruplar bellidir. Test verisi kullanılarak bunların hangi sınıfa dâhil olduğu bulunmaktadır. (Aydın, 2007).

Sayısal veri ve sayısal olmayan veriler üzerinde veri madenciliği işlemler yapmaktadır. Metin madenciliği sayısal olmayan veri kategorisine girmektedir. Metin madenciliği metinlerden oluşan veri setlerinin analiz edilmesidir. Metin madenciliğinde sınıflandırma işlemi yapılırken, karşılaşılan en büyük sorun metinlerin hangi sınıfa dâhil olması gerektiğidir. Bu sorun veri madenciliğinden yararlanılarak giderilmiştir. (Tantuğ, 2012).

Veri madenciliği ve metin madenciliği arasındaki bağ şu şekilde tanımlanmaktadır; veri

madenciliği yapısal verileri analiz etmek için kullanılırken, metin madenciliği yapısal olmayan veri analizinde kullanılmaktadır. Yapısal olmayan veriler, yapısal veri haline dönüştürülerek, veri madenciliğinde yapay zeka algoritmaları aracılığıyla kullanılmaktadır. (Yıldız ve diğerleri, 2018).

Metin madenciliği sınıflandırma işlemi yapılırken, ilk olarak veri toplanır, veri önişleme adımları uygulandıktan sonra model oluşturulur, sonuç değerlendirilir ve oluşturulan model uygulanır. Eğer uygulama sonucunda uygulanan modelin sonucu iyi ise sonuçlar değerlendirilir, değil ise model baştan kurulur. Veri seti birçok yerden elde edilebilir. Uluslararası haber metinleri üzerinde analizlerde yapılmaktadır. Uluslararası haber, haber niteliği taşıyan bir olayın daha büyük kitlelere hitap etmesidir. Bunun için bazı özellikleri haberin taşıması gerekmektedir. Bu özellikler: haberin bir ulusu ilgilendiren öğeler içermesi, ulus çapında tanınan bir kişi hakkında olması, savaş, terör, dünya ile ilgili haberler. Ulusal haber ajanslarından bazıları Anadolu Ajansı, Magnum Photos, Reuters, Suriye Arap Haber Ajansı, Uriminzokkiri' dir. (Akın, 2010).

Yapılacak olan çalışmada, Uluslararası yayın yapan bir kanalın internet sitesi üzerinden yayınlanan İngilizce haber verileri alınmıştır. Üç kategori altındaki veriler, internet sitesinden toplanmıştır. Bunlar: Uluslararası Konularda Raporlar (600 adet), Spor Raporları (600 adet), Dergi (Magazin) Raporları (600 adet). Kategorilerde değişiklik yapılmamıştır, site içerisinde kategoriler üç ana başlık adı altında toplanmaktadır ve mevcut kategori isimleri altındaki haberler alınmıştır. Haberler 2019-2020 tarihleri arasında yayınlanan haberler arasından alınmıştır. Analizde kullanılan haberler rastgele seçilmiştir. Toplam 1800 adet metin verisi üzerinde analizler yapılmıştır. Veriler önce metin önişleme ve özellik çıkarımına tabi tutulmuştur. Daha sonra verilerin bir kısmı test, bir kısmı eğitim verisi olarak kullanıştır. Sınıflandırma işleminde Rastgele Orman Karar Ağacı Algoritması, ZeroR Algoritması, Naif Bayes Algoritması kullanılmıştır. Literatür incelemesi sonucu metin madenciliği kapsamında en çok tercih edilen algoritmalar seçilerek analizler yapılmıştır. En iyi sınıflandırma başarısını bulmak için, algoritmaların başarı oranları birbirleri ile karşılaştırılmıştır. En iyi sonucu veren algoritma içerisinde parametreler değiştirilerek, algoritmanın daha iyi sonuç vermesi amaçlanmaktadır. Doğrulama yöntemi olarak 10 kat çapraz doğrulama yöntemi kullanılmıştır. Bunun sebebi, veriyi 10 eşit parçaya bölerek bütün

verilerin hem test, hem de eğitim olarak kullanılmasını sağlamaktır. Analizin amacı, eldeki veri seti için en iyi sınıflandırma yapan algoritmayı ve en iyi çalışma zamanını tespit etmektir. Çalışmanın amacı ise, veri yığını içerisine girmeden anlamlı veriler üzerinde denetleme ve analizlerin daha kısa sürede yapılmasını sağlayarak işletmelere yardımcı olmaktır.

2. Literatür Taraması

Metin madenciliği konusunda literatür çalışması yapıldığında en iyi sonucun Rastgele Orman Algoritmasının verdiği görülmüştür. İncelenen bazı araştırmalar aşağıda verilmektedir.

Kılınç ve diğerleri (2018)'nin yapmış olduğu 'İstatistik Kitaplarının Metin Madenciliği Yöntemleri Kullanılarak Yazarlarının Eğitimine Göre Sınıflandırılması' adlı çalışmada, sınıflandırma algoritması olarak k-en yakın komşuluk (K-NN), destek vektör makinesi (SVM) ve Rastgele Orman (RF) kullanılmıştır. En iyi algoritma başarı sonucunu RF olarak bulmuşlardır.

Ünal ve diğerleri (2016)'nin yapmış olduğu 'Metin Madenciliğinde Yazar Tanıma' adlı çalışmada, sınıflandırma algoritması olarak karar ağacı, k-en yakın komşuluk, naif bayes, rastgele orman kullanılmıştır. En iyi algoritma başarı sonucu rastgele orman algoritması olarak bulunmuştur.

Karasoy (2016)' un yapmış olduğu 'İçerik Tabanlı İstenmeyen SMS Filtreleme için Mobil Uygulama Geliştirilmesi ve Sınıflandırma Algoritmalarının Karşılaştırılması' adlı çalışmada, Bagging, Rastgele Orman, Rastgele Alt Uzay sınıflandırma algoritmaları kullanılmıştır. En iyi başarı sonucunu veren algoritma Rastgele Ormandır.

Tekin (2018)'in yapmış olduğu 'Yazılım Geliştirme Taleplerinin Metin Madenciliği İle Sınıflandırılması Ve Önceliklendirilmesi' adlı yüksek lisans tezinde SMO, Ratation Forest, Random Forest, Naive Bayes, Naive Bayes Multional sınıflandırma algoritmasını kullanmıştır. En iyi başarı sonucunu veren algoritma Rastgele Ormandır.

Abidin (2017)'nin yapmış olduğu 'Klasik Türk Müziğinde Makam Tanıma İçin Veri Madenciliği Kullanımı' adlı çalışmada Random Forest Algoritmasının çok iyi başarı ile sınıflandırma işlemini yaptığını ortaya koymuştur.

3. Metodoloji ve Veri

3.1. Sınıflandırma Amaçlı Öğrenme Yöntemleri: ZeroR

Diğer sınıflandırma algoritmaları arasında daha ilkel olan bir algoritmadır. Çalışma mantığı çok basittir. Eğitim setinde frekansı en yüksek olanı seçer ve test verilerinin hepsini o sınıfa ait kabul eder. (Nasa ve diğerleri, 2012).

Örneğin: Bir sınıfta 40 erkek, 30 kız öğrenci vardır. Yeni gelen öğrencinin cinsiyetinin tahmin edilmesi gibi bir problemde, zeroR algoritması erkek sayısı fazla olduğu için yeni gelen bütün öğrencileri erkek kabul edecektir.

Yapılan çalışmada, 1800 adet veriden 600 adet veri doğru, 1200 adet veri yanlış sınıflandırılmıştır. Başarı oranı %34 tür. Çalışma zamanı 0.02 sn dir. Başarı oranının düşük çıkmasının sebebi, veri seti içerisindeki sınıf sayısının eşit olmasından kaynaklanmaktadır.

3.2. Sınıflandırma Amaçlı Öğrenme Yöntemleri: Naif Bayes

Bayes teoremini temel alarak çalışan bir algoritmadır. Olasılık hesaplamalarına göre çalışır. Her özelliğin sonuca etkisi üzerine olasılık değerlerinin hesaplanması ile gerçekleşir. Verinin mevcut sınıflardan hangisine ait olma olasılığını hesaplar. Özelliklerin önem derecesini hepsinde eşit almaktadır. Bu şekilde daha doğru sonuçlara ulaşmaktadır. Bütün özellikler birbirinden bağımsız olarak kabul edilmektedir. Özet olarak olasılık değeri en yüksek olan kararın değerlendirilip sonuçlandırılmasıdır. Bayes teoreminde olduğu gibi koşullu olasılıktan faydalanır ve test veri setinde bulunan üyelerin hangi sınıfa ait olduğunu bulmaya çalışır. (Kalaycı 2018; Karakoyun ve diğerleri, 2014).

Bayes Teoremi Formülü (Çalış ve diğerleri, 2013).

Yapılan çalışmada, 1800 adet veriden 1575 adet veri doğru, 225 adet veri yanlış sınıflandırılmıştır. Başarı oranı %87 dir. Çalışma zamanı 0.06 sn dir.

3.3. Sınıflandırma Amaçlı Öğrenme Yöntemleri: Rastgele Orman

Ağaçlar topluluğu olarak da tanımlanmaktadır. Regresyon ve sınıflandırma için kullanılabilir. Rastgele Orman (RF) Karar Ağacı Algoritmasında bir tane karar ağacı yoktur. Karar

ağacı n adettir ve n adet karar ağacı rastgele oluşturulur. N değeri değiştirilebilir ve n değerini kullanıcı belirleyebilir. Gini indeksini baz alarak ağaçları oluşturur. Ağaçlar oluşturulurken sadece eğitim setleri kullanılır. Test veri seti test aşamasında kullanılmaktadır. Karar ağaçlarının sonuçları tek tek ele alınır ve hangi sonuçtan daha fazla ise o sınıfa dahil edilir. (Hazım, 2018).

RF, veri setinin tek bir ağaç yerine birçok ağaç oluşturması, algoritmanın daha iyi çalışmasına sebep olmaktadır. Bir olay ya da duruma bir ağacın değil de birçok ağacın karar verip ortak kararın uygulanmasını sağlamaktadır. Parçalar bir araya gelerek daha iyi sonuç üretmektedir. Sonuç parçaların ortak kararıdır. Parçalardan bazıları yanlış olsa da, diğer parçalar doğru olacaktır. Bütüne bakıldığında karar doğru olacağından, yanlış kararlar sonucu daha az etkileyecektir. (Pervan, 2019).

Buradaki en büyük sorun varyansı dengelemektir. Eğitim veri setlerindeki ufak değişiklikler bile sonucu ciddi şekilde etkileyecektir. Eğitim seti alt ağaçlara bölünürken çantalama algoritmasını kullanmaktadır. Çantalama Algoritmasını RF'da kullanılmasının amacı, varyansı azaltarak ağaçların oluşturulmasıdır. Mevcut veri setini kullanarak n adet veri setleri üretir. Ürettiği yeni veri setleri ile n adet ağacı eğitir. Geri beslemelidir, bu sebeple bir eleman birden fazla yerde kullanılabilir. Veri setleri seçimi rastgele yapılır. (Aslan 2016).

Yapılan çalışmada, 1800 adet veriden 1633 adet veri doğru, 167 adet veri yanlış sınıflandırılmıştır. Başarı oranı %89 dur. Çalışma zamanı 1.78 sn dir.

3.4. Metin Ön İşleme ve Öznitelik Çıkarımı

Metin önişleme aşaması, mevcut veri setini analize hazırlamak için gereklidir. Bu aşama analiz sonuçlarının doğruluğu açısından çok önemlidir. Veri içerisinde analiz sonucunu etkilemeyecek kelimelerin veri setinden çıkarılması ve anlamlı kelimelerle veri analizi yapılması gerekir. Makine doğal dilden farklı olarak verileri analiz etmektedir. Bu sebeple doğal dilden veri setini ayırmak gerekir. Bunun için metin önişleme aşamaları mevcuttur. (Jivani, 2011).

Bu aşamalar tek tek ele alınarak veri seti üzerinde uygulanmıştır. Başlangıçta 1935 adet özellik varken, önişleme aşaması bittikten sonra 73 adet özellik kalmıştır. Bu özellikler

tablo 1 de özellikler verilmektedir. Bu özellikler kullanılarak metin içerisinde analizler yapılmaktadır.

Tablo 1: Ön İşlem Sonrası Veri Seti

Analizde Kullanılan Veri Seti					
Boeing	Party	Human	Sports	Season	League
Congress	President	Issue	Government	Win	Football
Department	States	Large	Groups	Workers	Game
Donald	Texas	Led	Match	Working	Standings
House	Trump	Life	Player	Cup	Title
Mexico	University	Living	World	Officials	Military
National	Washington	Main	Children	Protect	Murder
Areas	Population	Members	Citizens	Recent	Office
Islam	Institutions	Team	Community	Recently	Order
People	Tournment	War	Companies	Response	Part
Olympic	Small	Started	State	System	Country
Political	Champion	Forces	Current	Elections	Federal
Final					

Veri dönüşümü: Veriler toplandıktan sonra kullanılacak olan derleyicide çalışabilmesi için, derleyicinin desteklediği formata dönüştürülmesi gerekmektedir. Veriler toplanıp, Weka derleyicisi içerisine alınabilmesi için 'arff' dosya formatına çevrilmiştir.

Kelimelerin sayısal matris olarak ifade edilmesi: Kelimeler sayısal verilere dönüştürülürken, terim ortalamaları ve terim standart sapmaları hesaplanarak sayısallaştırma işlemi yapılmıştır. Kelime tekrarları temel alınarak hesaplamalar yapılırken, aşağıdaki formüllerden yararlanılmıştır. (Kılınç ve diğerleri, 2011).

Özellik çıkarımı: Metin içerisinde kelime ayrımı işlemi yapılmıştır. Bütün kelimeler n'li gruplara ayrılarak işlemler yapılabilir. N=1 seçilerek, tekli kelime grupları şeklinde özellik çıkarımı yapılmıştır.

Gereksiz imleçleri metin içerisinden çıkartma: Nokta, virgül, soru işaretleri, noktalama işaretleri, boşluk karakteri, özel karakterler metin içerisinden çıkartılmıştır.

Kelime uzunluğu hesaplama: Veri seti içerisinde kelime uzunlukları hesaplanmıştır. Minimum kelime uzunluğu üç ve üç' ten küçük olan kelimeler veri seti içerisinden çıkartılmıştır.

Büyük-küçük harf dönüşümleri: Metin içerisinde aynı kelimelerin tekrar etmemesi için, bu dönüşümün yapılması gerekmektedir. Metindeki tüm veriler küçük harfe çevrilmiştir. Metin içerisinde aynı kelimelerin geçmesi ve aynı kelimelerin farklı kelimeler gibi algılanması önlenmiştir.

Durdurma kelimeleri: Edat, bağlaç, zamir, sayılar, tarihler gibi sınıflandırmayı etkilemeyecek olan verilerin metin içerisinden çıkarılması gerekmektedir. Durdurma kelimesi olarak 630 adet İngilizce kelime hazırlanmıştır ve metine uygulanmıştır.

Kök Bulma Algoritması: Kök bulma kullanılmasındaki amaç, aynı köke sahip, farklı ek alan kelimeler bir arada analiz içerisinde kullanılırsa makine, iki kelimeyi farklı kelimeler olarak algılamaktadır. İki kelimenin farklı kelimeler olmadığının makineye gösterilmesi gerekmektedir. Metin içerisinde aynı kökten gelen kelimeler bulunmuştur ve metin içerisinden çıkartılmıştır. Kartopu kök bulma algoritması kullanılmıştır. Bunun sebebi kartopu algoritması kendisinden daha önce kullanılan algoritmalara göre daha iyi sonuç vermesidir. (Vijayarani ve diğerleri, 2014).

4. Analiz Sonuçları

4.1. Yöntemlerin Karışıklık Matrislerinin İncelenmesi

Tablo 2 de algoritmaların karışıklık matrisleri verilmiştir. Bu matrisler başarı oranlarının tablo ile ifade edilmesidir. Matrislere bakılarak verinin, hangi algoritma kullanıldığında kaç adet doğru-yanlış sınıflandırıldığı görülmektedir. (Kılınç ve diğerleri 2016, 92). Uluslararası Konularda Raporlar (IN), Spor Raporları (SN), Dergi (Magazin) Raporları (MN) olarak ifade edilmiştir.

Tablo 2: Algoritmaların Karışıklık Matrisleri

ZeroR				Naif Bayes				Rastgele Orman			
Sınıf	IN	SN	MN	Sınıf	IN	SN	MN	Sınıf	IN	SN	MN
IN	600	0	0	IN	510	25	65	IN	511	29	60
SN	600	0	0	SN	23	566	11	SN	30	552	18
MN	600	0	0	MN	83	22	495	MN	40	10	550

Karışıklık matrisine göre;

ZeroR Algoritmasında, 600 adet veri sadece IN kategorisinde doğru sınıflandırılmıştır. MN ve SN kategorilerindeki bütün veriler yanlış sınıflandırılarak hepsi IN kategorisinde gösterilmiştir.

Naif Bayes Algoritmasında, IN kategorisinde 510 adet doğru, 90 adet yanlış veri sınıflandırılmıştır. SN kategorisinde 566 adet veri doğru, 44 adet veri yanlış sınıflandırılmıştır. MN kategorisinde 495 adet veri doğru, 95 adet veri yanlış sınıflandırılmıştır.

Rastgele Orman Algoritmasında, IN kategorisinde 511 adet veri doğru, 89 adet veri yanlış sınıflandırılmıştır. SN kategorisinde 552 adet veri doğru, 48 adet veri yanlış sınıflandırılmıştır. MN kategorisinde 550 adet veri doğru, 50 adet veri yanlış sınıflandırılmıştır.

Tablo 3: Algoritma Sonuçlarının Karşılaştırılması

Naif Bayes	ZeroR	Rastgele Orman	Algoritma
87	34	89	Başarı Oranları(%)
1571	600	1633	Doğru Sınıflandırılan Veri Sayısı
229	1200	167	Yanlış Sınıflandırılan Veri Sayısı
0.06	0.02	1.78	Çalışma Zamanı(sn)

Metin madenciliğinde, mevcut veri seti üzerinde en iyi sonucu veren algoritma Rastgele Orman Algoritmasıdır. Literatürde de metin madenciliği ile ilgili yapılan sınıflandırma

çalışmaları incelendiğinde RF' ın diğer algoritmalara göre daha iyi çalıştığı görülmektedir.

Buna rağmen, Rastgele Orman Algoritmasının parametreleri değiştirilerek daha iyi sonuçlar da elde edilebilir. Rastgele Orman Algoritması için ormandaki ağaç sayısının değişmesi ve seçilecek olan ağaç elemanlarının seçiminin değiştirilmesi, RF başarı sonucunu ve çalışma süresini olumlu yönde etkileyecektir. (Özdarıcı ve diğerleri, 2011).

Tablo 4: Rastgele Orman Parametre Seçimi

Ağaç Sayısı	Özellik Seçimi	Başarı Oranı(%)	Çalışma Süresi(sn)
100	1	90.8	1.16
100	2	90.7	0.8
100	3	90.5	0.9
50	1	90.4	0.38
50	**2**	**90.7**	**0.47**
50	3	90.2	0.55
25	1	89.6	0.23
25	2	89.4	0.27
25	3	89.8	0.3

İlk aşamada ağaç sayısı=100 ve özellik seçimi=1 iken başarı oranı %89, çalışma süresi 1.78 sn dir. Ağaç sayısı=50 ve özellik seçimi=2 olarak seçildiğinde Rastgele Orman Karar Ağacı Algoritmasının en iyi başarı oranını verdiği görülmüştür. Başarı oranı %90.7 dur. Çalışma süresi= 0.47 sn'dir. Yaklaşık 0.02 başarı oranı ve 1.31 sn daha iyi zaman ile sınıflandırma yapmıştır. Metin madenciliğinde Rastgele Orman algoritması kullanırken parametreler ağaç sayısı ve özellik seçimi değişkenleri değiştirilirse daha iyi sonuç elde edilmektedir.

5.Sonuç

Yapılan çalışma, belirli sınıflara dahil olan haber verileri üzerinde üç farklı tip ve üç farklı kategori yapay zeka algoritmaları kullanılarak makine eğitilmesi yöntemiyle yapılan sınıflandırmada, algoritmaların sınıflandırma analizlerinin karşılaştırılması ve haber metinlerinin yönetimlerinin saptanmasıdır. Yapılan çalışma temel alınarak işletmeler kendi alanlarına göre kendi projeleri içerisine çalışmayı ekleyebilirler, analiz ve denetlemelerini daha kısa sürede yapabilirler. Örneğin; Firmalar müşteri ilişkileri uygulamalarında denetleme ve

analizler yapabilirler. Firma içi performans analizinde metin madenciliği uygulamaları kullanılabilir. Metin madenciliği uygulamasını birçok alanda kullanmak mümkündür. Çalışma, firmaların kendi veri tabanı içerisinde denetleme ve analizleri amaca göre daha doğru bir şekilde yapmasını sağlamayı hedeflemektedir. Haber verilerinin sınıflandırılmasındaki amaç, haberler yazılıp ortak havuza düştüğünde sınıflandırma işlemi sayesinde kategorizasyon ve denetleme işlemleri için, zaman ve iş gücünün kullanılmamasını sağlamaktır.

1800 adet İngilizce haber metini üzerinden sınıflandırma işlemi yapılmıştır. Çalışma içerisinde üç adet algoritma mevcuttur. Bu algoritmalar: ZeroR, Naif Bayes, Rastgele Orman'dır. Algoritmaların başarı oranları karşılaştırıldığında yaklaşık %89 oranında ve çalışma zamanı 1.78 sn olarak Rastgele Orman Algoritmasının diğer algoritmalara göre daha iyi çalıştığı tespit edilmiştir. Fakat elde edilen bu başarı oranı ve çalışma süresi Rastgele Orman Algoritması içerisindeki parametreleri değiştirerek arttırılabilir. Bu sebeple 'ağaç sayısı' ve 'özellik seçimi' parametreleri değiştirilmiştir. Bu değişim sonucunda başarı oranı yaklaşık %91 ve çalışma zamanı 0.47 sn olarak bulunmuştur. Metin madenciliğinde Rastgele Orman Algoritması diğer algoritmalara göre daha iyi sonuç vermektedir. Eğer ağaç sayısı 50 ve özellik seçimi 2 olarak seçilirse, elde edilen başarı oranı ve çalışma süresi de iyileştirilmektedir.

Çalışma sürelerinin karşılaştırılmasındaki amaç, gerçek zamanlı dinamik verilerde işlem süresi daha uzun olacaktır. İşlem süresini kısaltmaya yönelik tedbirler alınabilir. Bazı alanlarda süre, doğruluk kadar hayati önem taşımaktadır. Bu sebeple, işlem süresi değerlendirmesi projelerin olmazsa olmazlarındandır.

Gelecekte Rastgele Orman (RF) Algoritmasını geliştirmek için iki adet proje düşünülmüştür. RF içerisindeki hesaplamalar yapılırken kullanılan gini indeksi yerine verilerin ortalaması alınarak denemeler yapılması planlanmaktadır. Diğeri ise RF oluşturulurken alt ağaçları rastgele oluşturuyor, bu rastgeleliği değiştirerek bir sıralama algoritması yazmak ve bu sıralama algoritmasına göre alt ağaçları oluşturmasını sağlamaktır. Yeni oluşturulan bu rastgele orman artı adını vereceğim algoritma üzerinde denemeler yapılması hedeflenmektedir.

Finansal Destek

Yazar bu çalışma için herhangi bir finansal destek almamıştır

Kaynakça

Abidin, S., Öztürk, Ö. & Öztürk, T.Ö. (2017). Klasik Türk müziğinde makam tanıma için veri madenciliği kullanımı. Gazi Üniversitesi Mühendislik, Mimarlık Fakültesi Dergisi, 32(4), 1221–1232.

Akın, Z.O. (2010). Uluslararası haber ajanslarının Türkiye haberlerinde eşik bekçiliği uygulamaları: Reuters ve AP örneği. Yüksek Lisans Tezi, Gazi Üniversitesi Sosyal Bilimler Enstitüsü, Ankara, Türkiye.

Aslan, M. (2016). Derinlik kamerası ile yaşlılarda düşme tespiti. Doktora Tezi, Fırat Üniversitesi Fen Bilimleri Enstitüsü, Elazığ, Türkiye.

Aydın, S. (2007). Veri madenciliği ve Anadolu Üniversitesi uzaktan eğitim sisteminde bir uygulama. Doktora Tezi, Anadolu Üniversitesi Sosyal Bilimler Enstitüsü, Eskişehir, Türkiye.

Çalış, K., Gazdağı, O. & Yıldız, O. (2013). Reklam içerikli epostaların metin madenciliği yöntemleri ile otomatik tespiti. Bilişim Teknolojileri Dergisi, 6(1), 1–7.

Ertuğrul, İ., Organ, A. & Şavlı, A. (2012). Veri madenciliği uygulamasına ilişkin PAÜ hastanesinde hasta profilinin belirlenmesi. Pamukkale Üniversitesi Mühendislik Bilimleri Dergisi, 19(2), 97–103.

Hazım, L.R. (2018). Four classification methods naïve bayesian, support vector machine, k-nearest neighbors and random forest are tested for credit card fraud detection. Yüksek Lisans Tezi, Altınbaş Üniversitesi, İstanbul, Türkiye.

Jivani, A.G. (2011). A comparative study of stemming algorithms. International Journal of Computer Science, 2(6), 1930–1938.

Kalaycı, T.E. (2018). Kimlik hırsızı web sitelerinin sınıflandırılması için makine öğrenmesi yöntemlerinin karşılaştırılması. Pamukkale Üniversitesi Mühendislik Bilimleri Dergisi, 5(24), 870–878.

Karakoyun, M. & Hacıbeyoğlu, M. (2014). Biyomedikal veri kümeleri ile makine öğrenmesi sınıflandırma algoritmalarının istatistiksel olarak karşılaştırılması. Dokuz Eylül Üniversitesi Mühendislik Fakültesi Mühendislik Bilimleri Dergisi, 16(48), 30–41.

Karasoy, O. & Ballı, S. (2016). İçerik tabanlı istenmeyen sms filtreleme için mobil uygulama geliştirilmesi ve sınıflandırma algoritmalarının karşılaştırılması. Konferans bildirisi. International Artificial Intelligence and Data Processing Symposium (IDAP'16).

Kılınç, C., Bozyiğit, F., Özçitf, A., Yücelar, F. & Borandağ, E. (2011). Metin madenciliği kullanılarak yazılım kullanımına dair bulguların elde edilmesi. Konferans bildirisi. Celal Bayar Üniversitesi Hasan Ferdi Turgutlu Teknoloji Fakültesi Yazılım Mühendisliği Bölümü, Manisa, Türkiye, Ekim.

Kılınç, D., Borandağ, E., Yücalar, F., Tunalı, V., Şimşek, M. & Özçift, A. (2016). KNN algoritması ve r dili ile metin madenciliği kullanılarak bilimsel makale tasnifi. Marmara Fen Bilimleri Dergisi, 3, 89–94.

Kılınç Kan, B. & Yazarlı, Y. (2018). İstatistik kitaplarının metin madenciliği yöntemleri kullanılarak yazarlarının eğitimine göre sınıflandırılması. Turkiye Klinikleri J Biostat, 3(10), 215–223.

Nasa, C. & Suman S. (2012). Evaluation of different classification techniques for web data. International Journal of Computer Applications, 52(9), 35–40.

Özdarıcı O.K., Akar, Ö. & Güngör, O. (2011). Rastgele orman sınıflandırma yöntemi yardımıyla tarım alanlarındaki ürün çeşitliliğinin sınıflandırılması. Konferans bildirisi. ODTÜ, Jeodezi ve Coğrafi Bilgi Teknolojileri EABD, Ankara, Türkiye, Ocak.

Pervan, N. (2019). Derin öğrenme yaklaşımları kullanarak Türkçe metinlerden anlamsal çıkarım yapma. Yüksek Lisans Tezi, Ankara Üniversitesi Fen Bilimleri Enstitüsü, Ankara, Türkiye.

Tantuğ, A.C. (2012). Metin sınıflandırma. Türkiye Bilişim Vakfı Bilgisayar Bilimleri Dergisi, 6(6), 1–12.

Tekin, M.C. (2018). Yazılım geliştirme taleplerinin metin madenciliği ile sınıflandırılması ve önceliklendirilmesi. Yüksek Lisans Tezi, Maltepe Üniversitesi Fen Bilimleri Enstitüsü, Maltepe, İstanbul.

Ünal, D.İ. & Şeker, ŞE (2018). Metin madenciliğinde yazar tanıma. YBS Ansiklopedisi, 5(1), 1–6.

Vijayarani, S., Ilamathi, J. & Nitya (2014). Preprocessing techniques for text mining - An overview. International Journal of Computer Science, 5(1), 7–16.

Yıldız, B. & Ağdeniz, Ş. (2018). Muhasebe analiz yöntemi olarak metin madenciliği. Muhasebe Bilim Dünya Dergisi, 2(20), 286–315.

Yıldız, M. & Şeker, Ş.E. (2016). Veri madenciliği araçları (data mining tools). YBS Ansiklopedi, 3(4), 10–19.

Özgeçmiş

Firdevs Durnagöl 2015 yılında Karabük Üniversitesi Bilgisayar Mühendisliği bölümünden mezun olmuştur. 2020 yılında İstanbul Aydın Üniversitesi Bilgisayar Mühendisliği bölümünde yüksek lisansını tamamlamıştır. 2015-2107 yılları arasında TrtArabi'de Bilgi İşlem Uzmanı olarak çalışmıştır. 2017 yılından itibaren ise TrtWorld'de Bilgi İşlem Uzmanı olarak çalışmaya devam etmektedir.