ESTUDAM Bilişim Dergisi
Cilt 2, Sayı 2, 1-5, 2021

Journal of ESTUDAM Information
Volume 2, Issue 2, 1-5, 2021

(Geliş Tarihi / Received Date: 22.01.2021, Kabul Tarihi/ Accepted Date: 01.03.2021)

Türkçe Twitter Verilerinden Duygu Analizi Yapılırken Sınıflandırıcıların Karşılaştırılması

Enes Kumaş

[1]Eskişehir Osmangazi Üniversitesi, Fen Bilimleri Enstitüsü, Bilgisayar Mühendisliği. Bölümü, 26480, Odunpazarı/Eskişehir, ORCID No : http://orcid.org/0000-0002-8645-1261

Anahtar Kelimeler:
Sosyal Medya Analizi,
Veri Madenciliği,
Metin Madenciliği,
Duygu Analizi

Özet: Günümüzde sosyal medya kullanımında hızlı bir artış yaşanmaktadır. İnsanların sosyal medya platformlarında yaptığı paylaşımlar ile yüksek miktarda veriler oluşmaya başlamıştır. Bu veriler bizlere kişiler, ürünler, firmalar ve daha birçok alanda bilgi sağlamaktadır. Artan veri miktarı ile birlikte verilerin işlenip analiz edildiği çeşitli çalışma alanları ortaya çıkmıştır. Duygu analizi bu çalışma alanlarından biridir. Duygu analizi bir kişinin ya da metinin belirli bir konuya yönelik tutumunun olumlu, olumsuz ya da tarafsız olarak sınıflandırılma sürecidir. Bu çalışmada da sosyal medya platformu olarak Twitter kullanılmıştır. Türkçe Twitter verileri kullanılarak duygu analizi yapılmıştır. Twitter verilerine metin madenciliği yöntemleri uygulanarak veriler analiz edilmiştir ve Naive Bayes, KNN, Destek Vektör Makinesi, Lojistik Regresyon ve Karar Ağacı sınıflandırma algoritmaları kullanılarak pozitif ve negatif olarak sınıflandırılmıştır. Sınıflandırma sonuçları f1 skoru ile değerlendirilmiştir. Sırasıyla Naive Bayes, KNN, Destek Vektör Makineleri, Lojistik Regresyon ve Karar Ağacı sınıflandırıcılarıyla elde edilen f1 skorları %70, %65, %73, %71 ve %69 bulunmuştur.

Comparison of Classifiers While Analyzing Sentiment from Turkish Twitter Data

Keywords:
Social Media Analysis,
Data Mining,
Text Mining,
Sentiment Analysis

Abstract: Today, there is a rapid increase in the use of social media. With the posts people make on social media platforms, a large amount of data has begun to occur. These data provide us with information on people, products, companies and many other areas. With the increasing amount of data, various fields of study have emerged where data are processed and analyzed. Sentiment analysis is one of these areas of study. Sentiment analysis is the process of classifying a person's or text's attitude towards a particular subject as positive, negative or neutral. In this study, Twitter was used as a social media platform. Sentiment analysis was conducted using Turkish Twitter data. By applying text mining methods to Twitter data, the data was analyzed and classified as positive and negative using Naive Bayes, KNN, Support Vector Machine, Logistic Regression and Decision Tree classification algorithms. Classification results were evaluated with f1 score. Respectively the f1 scores obtained with Naive Bayes, KNN, Support Vector Machines, Logistic Regression and Decision Tree classifiers were found to be 70%, 65%, 73%, 71% and 69%.

1. GİRİŞ

İnternetin gelişmesiyle birlikte insanların bir konu hakkında düşüncelerini paylaşabileceği, yorum yapabileceği birçok ortam oluşmuştur. Sosyal medya uygulamaları, bloglar, e-ticaret siteleri gibi her gün artan ve yaygınlaşan yeni ortamlar oluşmaktadır. Bu sebeple internet ortamındaki verilerin kullanıldığı çeşitli alanlar oluşmaya başlamıştır. Verileri işlemek ve analiz edebilmek için çeşitli yöntemler kullanılmaya başlanmıştır.

Sosyal medya verileri yapısal olmayan veri türündendir. Twitter verileri metin tabanlı yapısal olmayan verilerdir.

Bu verilerin kullanılabilmesi için çeşitli metin madenciliği yöntemleri kullanılarak yapısal forma dönüştürülmesi gerekmektedir.

Metin madenciliği yapısal olmayan ve düzensiz haldeki metin verilerinden yapısal, kullanışlı ve düzenli veriler elde etme işlemidir. Metin madenciliği çalışmaları metni veri olarak kabul eden veri madenciliği çalışmalarıdır. Metin madenciliği metinlerin sınıflandırılması,

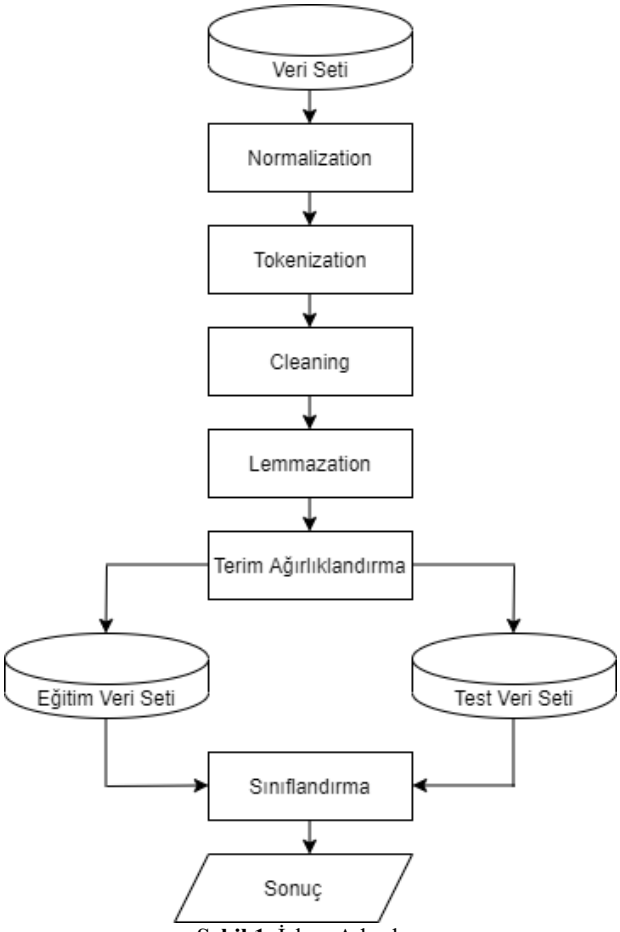
*İlgili yazar: eneskums@gmail.com

kümelenmesi, metin özetinin çıkarılması, intihal tespiti ve duygu analizi gibi çeşitli çalışmaları hedefler.

Bu çalışmada da sosyal medya platformu olan Twitterdan alınan verilere metin madenciliği yöntemleri uygulanarak duygu analizi yapmak amaçlanmıştır.

2. MATERYAL VE METOT

Çalışmadaki işlem süreçleri Şekil 1. de gösterilmiştir.



Şekil 1. İşlem Adımları

2.1. Kullanılan Araçlar ve Kütüphaneler

Proje gerçekleştirilirken geliştirme ortamı olarak pycharm, yazılım dili olarak python kullanılmıştır.

- Python, makine öğrenmesi, veri analizi ve veri işleme de kapsamlı kütüphaneler içermesi ve kullanımı kolay olduğu için tercih edilmiştir.

- Pycharm, python için oluşturulmuş yazılım geliştirme ortamıdır. Linux, macOS ve Windows işletim sistemi ile uyumlu çalışmaktadır.

Çalışmada jpype, numpy, pandas, nltk, sklearn ve zemberek kütüphaneleri kullanılmıştır.

- Jpype, python içinde java dilini kullanmayı sağlar. Java dilinde yazılan zemberek kütüphanesini kullanabilmek için kullanılmıştır.
- Numpy, temelde numpy dizilerinden oluşur. Numpy dizileri python listelerine benzer ancak hız ve işlevsellik açısından python listelerinden daha kullanışlıdır. Projede eğitim ve test verilerini tutmak için kullanılmıştır.
- Pandas, veri analizi ve ön işleme aşamasını kolaylaştırır. Projede csv dosyasında tutulan verilerin alınıp işlenmesinde ve analiz edilmesinde kullanılmıştır.
- Nltk, insan dili verileriyle çalışmak için python dilinde geliştirilen açık kaynak kodlu kütüphanesidir. Projede Türkçe durak kelimelerinin kaldırılmasında kullanılmıştır.
- Sklearn, makine öğrenmesi modelleri oluşturmak için kullanılan kütüphanedir. Projede özellik çıkarımı, sınıflandırma algoritmalarının kullanımı ve sınıflandırma sonucunun gösterilmesi işlemlerinde kullanılmıştır.
- Zemberek, açık kaynak kodlu Türkçe doğal dil işleme kütüphanesidir. Projede kelime yazım yanlışlarının düzeltilmesi ve kelime köklerinin bulunması aşamasında kullanılmıştır.

2.2. Veri Seti

Veri seti olarak Mustafa Sert'tin [20] çalışmasında kullandığı farklı konulara ait 16000 tane pozitif ve 16000 tane negatif etiketli Türkçe tweet kullanılmıştır. Tweetler pandas kütüphanesi kullanılarak negatif ve pozitif tweetleri içeren 2 adet csv dosyasına yazılmıştır.

2.3. Ön İşleme

Twitter verileri yapısal olmayan verilerdir. Bu verilerin kullanılabilmesi için metin madenciliği yöntemleri kullanılarak yapısal forma dönüştürülmesi gerekir. Yapılan çalışmada metinlerin ön işleme sürecinde gerçekleştirilen işlemler şu şekildedir;

- Normalization
- Tokenization
- Durak kelimelerini ve üç harften küçük kelimeleri kaldırma.
- Metinden url, hashtag, kullanıcı isimleri, retweetler ve emojilerin temizlenmesi
- Noktalama işaretlerinin temizlenmesi
- Sayıların temizlenmesi
- Lemmatization

Normalizasyon aşamasında zemberek kütüphanesinin metin normalizasyonu kullanılmıştır. Yanlış yazılan kelimeler düzeltilmeye çalışılmış ve kelimede tekrar eden harflerin kaldırılması amaçlanmıştır. Metindeki tüm harfler küçük harfe dönüştürülmüştür. Cümle başındaki, sonundaki boşluklar ve cümle içindeki fazladan boşluklar kaldırılmıştır. Tweetlerin normalizasyon işleminden önceki hali Şekil 2. de ve normalizasyondan sonraki hali Şekil 3. de görüldüğü gibidir.

0	RT : Ruyamda ziya bi dizide cem yilmazla basro...
1	Memleketimden uzakta ilk dogum günüm hayırlısı...
2	Pink Floyd - Wish You Were Here: http://t.co/j...
3	RT : Kızlaaaar DM'ye gelmezseniz ikinci foto g...
4	RT : abicim bizim icin biseyler ayarlarsn se...

Şekil 2. İşlenmemiş tweetler

0	rt : rüyamda ziya bir dizide cem yılmazla başr...
1	memleketimden uzakta ilk doğum günüm hayırlısı...
2	pink floyd - wish you here here : http://t.co/...
3	rt : kızlar dmye gelmezseniz ikinci foto gibi ...
4	rt : abicim bizim için bir şeyler ayarlarsın s...

Şekil 3. Normalizasyon uygulanmış tweetler

Tokanization, metinlerin istenen özelliklere göre parçalara ayrılması işlemidir. Durak kelimelerinin kaldırılması bir dilde sık geçen ve, veya gibi kelimelerin modelin doğruluğunu bozmaması için çıkarılması işlemidir. Tweetler içerisinde hashtag, kullanıcı isimleri, url adresleri, rt, emojiler, noktalama işaretleri ve sayılar gibi anlam ifade etmeyen karakterler içermektedir. Özel bir fonksiyon yazılarak tweetler bu karakterlerden temizlenmiştir. Lemmazation, metindeki kelimeleri morfolojik analizini dikkate alarak köklerine indirgeme işlemidir. Lemmazation işlemi için zemberek kütüphanesinin lemmazation fonksiyonu kullanılmıştır. Tüm bu işlemler sonucunda tweetlerin son hali Şekil 4. de görüldüğü gibidir.

0	rüya ziya dizi yılmaz başrol kıskanmak alex fe...
1	memleket uzak doğum gün hayır olmak
2	pink floyd wish her her aracılığıyla
3	kız dm gelmek ikinci foto olmak
4	abi biz şey ayarlamak cansın burcu altın umut ...

Şekil 4. Tweetlerin son hali

2.4 Terim Ağırlıklandırma

Terim ağırlıklandırma metin madenciliği çalışmalarında performansı arttırmak ve bir terimin doküman içerisinde ne kadar önemli olduğunu belirtmek amacıyla terimlere ağırlık atanması işlemidir. Bu çalışmada TF-IDF terim ağırlıklandırma yöntemi kullanılmıştır. TF-IDF, bir terimin doküman içerisindeki önemini gösteren ağırlık faktörüdür. Terim sıklığı (TF), basitçe bir belgede bir terimin kaç kez geçtiğidir. Ters doküman sıklığı (IDF), metinlerin kaçında terimin geçtiği bilgisidir. Terimin geçtiği doküman sayısının toplam doküman sayısına bölünüp logaritmasının alınmasıyla bulunur. TF-IDF değeri terim sıklığı ile ters doküman sıklığı değerlerinin çarpımı ile bulunur. Bu çalışmada terim olarak tweetlerde geçen kelimeler ve doküman olarak tweetler kullanılmıştır. TF-IDF değerinin bulunmasında sklearn kütüphanesinin tfidfvektorizer sınıfı kullanılmıştır.

2.5. Sınıflandırıcılar

Çalışmada Twitter verilerinin duygu kutuplarının belirlenmesinde beş farklı sınıflandırma algoritması kullanılmıştır. Bunlar, Naive Bayes, K En Yakın Komşu, Destek Vektör Makineleri, Lojistik Regresyon ve Karar Ağacı algoritmalarıdır. Algoritmaların kullanılması ve sınıflandırma sonuçlarının gösterilmesi işlemi Sklearn kütüphanesi ile gerçekleştirilmiştir.

2.5.1. Naive Bayes

Bayes teoremine dayanan bir sınıflandırma tekniğidir. Basit bir anlatımla naive bayes sınıflandırması bir sınıfta başka özelliklerin yapısıyla bağlantısı olmayan belirli özelliklerin olduğunu varsayar. Bu özellikler diğerleri ile bağlı ya da birbirlerinin varlığı üzerine bağlı olsalar bile, bu özelliklerin hepsi bağımsız olasılıklara katkıda bulunur. Naive bayes modeli kurulum açısından basit ve özellikle çok büyük veri kümeleri için kullanılmıştır [1].

2.5.2. K En Yakın Komşu

Sınıfları belli olan bir örnek kümesindeki verilerden yararlanılarak kullanılmaktadır. Örnek veri setine katılacak olan yeni verinin, mevcut verilere göre uzaklığı hesaplanıp, k sayıda yakın komşuluğuna bakılır. Öznitelik değerlerine göre k komşu veya komşuların sınıfına atanır[2].

2.5.3. Destek Vektör Makineleri

Destek vektör makinesi algoritmasının amacı, n boyutlu bir uzayda(n özelliklerin sayısı) veri noktalarını belirgin bir şekilde sınıflandıran bir hiper düzlem bulmaktır. İki veri noktası sınıfını ayırmak için seçilebilecek birçok olası hiper düzlem vardır. Amacımız maksimum marjı olan, yani her iki sınıfın veri noktaları arasında maksimum mesafeye sahip bir düzlem bulmaktır[3].

2.5.4. Lojistik Regresyon

Lojistik regresyon, belirli bir bağımsız değişken seti kullanarak kategorik bağımlı değişkeni tahmin etmek için kullanılır. Lojistik regresyonda bir regresyon çizgisi uydurmak yerine, iki maksimum değeri tahmin eden bir sigmoid fonksiyon kullanılır[4].

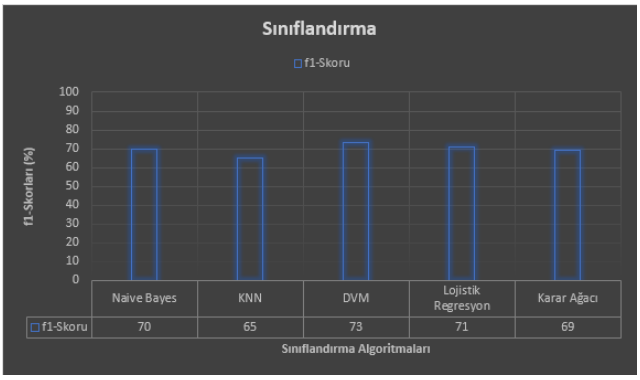
2.5.5. Karar Ağacı

Karar ağacı, dahili düğümlerin bir veri kümesinin özelliklerini, dalların karar kurallarını ve her yaprak düğümün sonucu temsil ettiği ağaç yapılı bir sınıflandırıcıdır. Bir karar ağacında karar düğümü ve yaprak düğümü olmak üzere iki düğüm vardır. Karar düğümleri herhangi bir karar vermek için kullanılır ve birden çok dala sahiptir. Yaprak düğümleri bu kararların çıktısıdır ve başka dal içermez. Kararlar verilen veri setinin özelliklerine göre gerçekleştirilir. Verilen koşullara göre bir karara ilişkin tüm olası çözümleri elde etmek için kullanılan grafiksel bir temsildir[5].

3. BULGULAR

Veriler gerekli ön işlemden geçirildikten sonra sınıflandırma aşamasına geçilmiştir. Ön işlemden geçirilen verilere tf-idf vektör dönüşümü uygulanmıştır ve sınıflandırmada özellik olarak bu dönüşüm kullanılmıştır. Veriler %80 eğitim ve %20 test olarak ayrılmıştır. Veriler

sınıflandırma algoritmalarına gönderilmiştir. Sınıflandırma aşamasında beş farklı algoritma kullanılmıştır ve elde edilen sonuçların f1 skorları Şekil 5. de gösterilmiştir. f1 skoru kesinlik ve duyarlılık değerlerinin harmonik ortalamasıdır. Sadece yanlış negatif ya da yanlış pozitif değil tüm hata maliyetlerini de içeren bir ölçme metriği olduğundan f1 skoru kullanılmıştır. Yapılan çalışmada en iyi sonuç destek vektör makinesi ile alınsa da sonuçlar genel olarak birbirine yakın bulunmuştur.



Şekil 5. Sınıflandırma sonuçları

4. TARTIŞMA VE SONUÇ

Bu çalışmada twitter kullanıcılarına ait pozitif ve negatif olarak etiketli tweet verilerine metin madenciliği yöntemleri uygulanarak duygu analizi yapılmıştır. En iyi sonuç destek vektör makinesiyle %73 olarak alınmıştır. Gelecekte yapılacak çalışmalarda, veri miktarı artırılarak ve n-gram yöntemleri kullanılarak başarı oranı arttırılabilir.

KAYNAKÇA

[1] Karaderili, Ş. 2018. Makine öğrenmesinde sınıflandırma algoritması türleri. https://medium.com/@sengul_krdrl/makine-öğrenmesinde-siniflandirma-algoritmasi-türleri-5e0f32245889 (Erişim Tarihi: 18.01.2021)

[2] Ulgen, K. E. 2017. k-En Yakın Komşuluk. https://medium.com/@k.ulgen90/makine-öğrenimi-bölüm-2-6d6d120a18e1 (Erişim Tarihi: 18.01.2021)

[3] Gandhi, R. 2018. Support Vector Machine. https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47 (Erişim Tarihi: 18.01.2021)

[4] Logistic Regression in Machine Learning. https://www.javatpoint.com/logistic-regression-in-machine-learning (Erişim Tarihi: 18.01.2021)

[5] Decision Tree Classification Algorithm. https://www.javatpoint.com/machine-learning-decision-tree-classification-algorithm (Erişim Tarihi: 18.01.2021)

[6] Saygısever, M. 2019. Metin Madenciliği Nedir?. https://medium.com/@minelsaygisever/metin-madenciliği-nedir-4a39809bfed5 (Erişim Tarihi: 16.01.2021)

[7] Metin Madenciliği Nedir?. http://www.metinmadenciligi.com (Erişim Tarihi: 16.01.2021)

[8] Demir, B. A. 2019. Duygu Analizi Nedir?. https://ardabatuhandemir.medium.com/duygu-analizi-sentiment-anaylsis-nedir-68a59a8b0142 (Erişim Tarihi: 16.01.2021)

[9] Yücel, F. Ö. 2020. Pycharm Nedir, Avantajları ve Özellikleri. https://www.mertmekatronik.com/pycharm-nedir-ozellikleri (Erişim Tarihi: 16.01.2021)

[10] JPype documentation. https://jpype.readthedocs.io/en/latest/ (Erişim Tarihi: 16.01.2021)

[11] Durna, B. M. 2019. Numpy. https://medium.com/bilişim-hareketi/veri-bilimi-için-temel-python-kütüphaneleri-1-numpy-750429a0d8e5 (Erişim Tarihi: 16.01.2021)

[12] Şirin, E. 2019. Python Pandas İle Temel İşlemler. https://www.veribilimiokulu.com/blog/python-pandas-ile-temel-islemler/ (Erişim Tarihi: 16.01.2021)

[13] Laura, M. 2021. Python Kütüphaneleri. https://tr.bitdegree.org/tutorial/python-kutuphaneleri/ (Erişim Tarihi: 16.01.2021)

[14] 2020. Zemberek. https://tr.wikipedia.org/wiki/Zemberek_(yazılım) (Erişim Tarihi: 16.01.2021)

[15] Afsina. 2018. Metin Normalizasyonu. http://zembereknlp.blogspot.com/ (Erişim Tarihi: 17.01.2021)

[16] Dayıbaşı, O. 2018. Metin Madenciliğinde Kavramlar. https://medium.com/algorithms-data-structures/metin-madenciliğinde-text-mining-kavramlar-1-e11b87b28847 (Erişim Tarihi: 17.01.2021)

[17] Demir, B. A. 2020. Duygu Analizi ve Fikir Madenciliği. https://ardabatuhandemir.medium.com/duygu-analizi-ve-fikir-madencili%C4%9Fi-3-sentiment-analysis-opinion-mining-stemming-and-39d7fd83f03c (Erişim Tarihi: 17.01.2021)

[18] Stecanella, B. 2019. What is TF-IDF?. https://monkeylearn.com/blog/what-is-tf-idf/ (Erişim Tarihi: 17.01.2021)

[19] Öğündür, G. 2019. Doğruluk, Kesinlik, Duyarlılık ya da F1 Score?. https://medium.com/@gulcanogundur/doğruluk-accuracy-kesinlik-precision-duyarlılık-recall-ya-da-f1-score-300c925feb38 (Erişim Tarihi: 18.01.2021)

[20] A. Hayran and M. Sert, "Sentiment analysis on microblog data based on word embedding and fusion techniques," 2017 25th Signal Processing and Communications Applications Conference (SIU), Antalya, 2017, 1-4. DOI:10.1109/SIU.2017.7960519

[21] Meral, M., Diri, B. 2014. Sentiment Analysis on Twitter. IEEE 22nd Signal Processing and Communications Applications Conference, 2014, 690-693. DOI:10.1109/SIU.2014.6830323

[22] Akgül, S. E., Ertano, C., Diri, B. 2016. Twitter Verileri ile Duygu Analizi. Pamukkale Üniversitesi Mühendislik Bilimler Dergisi, 22(2), 106-110. DOI:10.5505/pajes.2015.37268

[23] İlhan, N., Sağaltıcı, D. 2020. Twitter'da Duygu Analizi. Harran Üniversitesi Mühendislik Dergisi, 5(2), 146-156. DOI:10.46578/humder.772929

[24] Atan, S. 2020. Metin Madenciliği: İmkanlar, Yöntemler ve Kısıtlar. Mehmet Akif Ersoy Üniversitesi Sosyal Bilimler Enstitüsü Dergisi, 2020(31) ,220-239. DOI:10.20875/makusobed.476524