Yazar Tanımada Köşe Yazısı ve Tweet'lerin Çapraz Kullanımı
Cross Usage of Articles and Tweets on Author Identification

İslam Mayda[1], M. Fatih Amasyalı[1]

1. Bilgisayar Mühendisliği,
Yıldız Teknik Üniversitesi
islam.mayda@stu.khas.edu.tr, mfatih@ce.yildiz.edu.tr

Özetçe

En popüler mikroblog sitesi Twitter'da kimliğini açıklamadan yaptıkları paylaşımlarla fenomen olan yazarların kimlikleri merak edilmektedir. Bir Twitter hesabından yapılan paylaşımlar bu kullanıcının kimliğini tespit etmede kullanılabilir. Özellikle daha önceden çeşitli basın yayın organlarında yazılar yazmış olan bir köşe yazarı, kimliğini açıklamasa bile sosyal medya hesabında yaptığı paylaşımlar analiz edilerek kim olduğu tahmin edilebilir. Biz de bu çalışmada, 10 köşe yazarının makaleleri ile Twitter hesabında yaptığı paylaşımları karşılaştırarak bu hesabın hangi yazara ait olabileceğini tahmin etmeye çalıştık. İlk olarak her bir tweeti birer metin olarak alarak, daha sonra belirli sayıda tweeti birleştirerek testler gerçekleştirdik. Her bir tweeti ayrı ayrı değerlendirmeye göre birleştirilmiş tweetleri kullanmanın daha başarılı sonuçlar verdiğini gördük. Ayrıca, bir Twitter hesabından yapılan paylaşımları, aday yazarların makaleleri ile karşılaştırarak bu hesabın sahibinin kim olduğunu iyi bir başarı oranıyla tahmin edebileceğimizi gördük. Metinleri sayısallaştırmada kelimelerin kendilerini, köklerini ve 3-gramları kullandık. Çeşitli sınıflandırıcılar arasından en başarılı sonuçları destek vektör makineleri ile elde ettik.

Abstract

The identities of the authors who having phenomenon with their sharings without revealing his/her identity on Twitter which is the most popular microblogging site, are wondered. The sharings of a Twitter account can be used to detect the identity of the user. Especially, a columnist who have written articles on various media organs, even if he/she does not reveal his/her identity, can be guessed. We tried to guess the author of an account by comparing the articles and sharings on Twitter accounts of 10 columnists. We performed tests firstly by taking each tweet as an individual text, and then grouping the specific number of tweets. We perceived that using the grouped tweet texts gives more accurate results than using each tweet individually. Additionally, we caught that we can guess the owner of a Twitter account with a good accuracy rate by comparing the sharings of this account and the articles of the candidate authors. We used the words themselves, their stems and 3-grams for digitizing of the texts. We achieved the most successful results with support vector machines from among several classifiers.

1. Giriş

2015 sonu itibariyle aylık 305 milyon aktif kullanıcısı olan Twitter en popüler mikroblog sitesidir[1]. Birçok ünlü isim gibi köşe yazarlarının önemli bir bölümü de aktif bir Twitter kullanıcısıdır. Yüz binlerce takipçiye sahip olan bu yazarların Twitter hesapları, en çok takip edilen hesaplar arasında üst sıralarda yer almaktadırlar.

Twitter'da kullanıcılar kim olduklarını açıklamadan takma isimlerle de paylaşım yapabilmektedirler. Yaptıkları paylaşımlarla milyonlarca takipçi toplayarak sosyal medya fenomenine dönüşen Twitter kullanıcıları da vardır. Özellikle bu tür hesapların sahiplerinin kim oldukları kamuoyunda çok merak edilmekte ve sıkça tartışılmaktadır. Bu kişiler, kendi kimliklerini açıklamasalar da yazdıkları yazılar kendilerinin kimliği hakkında birtakım tahminler yapabilmek için kullanılabilmektedir.

Doğal dil işleme alanında yazar tanıma başlığı altında bir metnin kim tarafından yazıldığını ayırt etmek üzerine çok sayıda araştırma yapılmıştır. Ancak, bu araştırmalarda genelde uzun metinler üzerinde çalışılmıştır. Uzun metinler üzerinde yüksek doğruluk oranları elde edilebildiği görülmektedir. Örneğin, Amasyalı ve arkadaşları [2] yaptıkları araştırmada 18 farklı yazara ait 35'er gazete yazısı üzerinde SVM sınıflandırıcı ile %92.5 oranında yüksek bir başarı yakalamıştır.

Twitter'da ise kullanıcılar tweet adı verilen en fazla 140 karakterden oluşan paylaşımlar yapabilmektedir. Paylaşılan metinlerin kısa olması bu metinlerin yazarlarını tanımayı zorlaştırmaktadır. Schwartz ve arkadaşları [3] karakter n-gramları ve kelime n-gramları kullanarak SVM sınıflandırıcı ile 1000 yazarın 200'er tweeti üzerinde %30.3, 50 yazarın 50'şer tweeti üzerinde %50.7, 50 yazarın 1000'er tweeti üzerinde %71.2 başarı oranı elde etmişlerdir. Schwartz ve arkadaşları yaptıkları bu çalışmada her bir tweeti bir metin olarak ele alıp, 10 kat çapraz doğrulama ile test yapmışlardır.

Green ve Sheppard [4] finans sektöründeki 12 Twitter kullanıcısı üzerinde yaptıkları çalışmada, her yazar için 120'şer tweeti üzerinde SMO (Sequential Minimal Optimization) algoritmasını kullanarak 5x2 çapraz doğrulama ile test yapmış ve %40.5 doğruluk oranı yakalamışlardır. Green ve Sheppard bu araştırmalarında her

978-1-4799-7572-3/14/$31.00 ©2016 IEEE

bir tweetin içerdiği karakterler, uzun kelimeler, boşluklar, noktalama işaretleri, köprüler, toplam karakter sayıları, toplam kelime sayıları ve bunların frekansları gibi toplam 86 özellikten oluşan öznitelik kümelerini çıkartmış ve sınıflandırmada bunları kullanılmışlardır.

Bu çalışmada ise bir köşe yazarının Twitter hesabında paylaştığı tweetlerin birbiriyle benzerliklerinin yanı sıra, bu yazarın daha önce yazdığı makaleler ile paylaştığı tweetlerin benzerlikleri iki farklı yöntemle karşılaştırılmıştır. Bu karşılaştırmalarda ilk yöntemde tweetlerin her biri ayrı birer metin olarak ele alınmıştır. İkinci yöntemde ise, tweetlerin uzunluğunun kısa olması nedeniyle, her bir tweet ayrı birer metin olarak değerlendirilmemiş, belirli sayıda tweet birleştirilerek oluşturulan daha uzun tweet metinleri üzerinde çalışılmıştır. Bu şekilde, ilk yöntemde paylaşılan tweetin hangi yazara ait olduğu araştırılırken, ikinci yöntemde ise belirli sayıda tweet kümesi sınıflandırılarak tweetlerin paylaşıldığı Twitter hesabının hangi yazara ait olduğu tespit edilmeye çalışılmıştır.

Türkçe metinler üzerinde yapılan bu çalışmada, tweetlerde ve makalelerde geçen kelimelerin kendileri, kökleri ve 3-gramları üzerinde çalışılmıştır. Kelimelerin kökleri bulunurken Zemberek aracından faydalanılmıştır [5]. Sınıflandırıcı olarak ise WEKA [6] kütüphanesinde yer alan Naive Bayes (NB), Karar Destek Makinesi (SMO), Karar Ağacı (J48) ve Bagging sınıflandırıcıları kullanılmıştır.

2. Veri Toplama

Çalışmada Türkiye'de en çok takip edilen Twitter kullanıcıları arasında yer alan 10 köşe yazarının 200'er tweeti ve 50'şer makalesi kullanılmıştır. Bu yazarların Twitter'daki takipçi sayıları Tablo 1'de listelenmiştir.

Tablo 1: Çalışmada kullanılan köşe yazarlarının takipçi sayıları

Yazar No	Takipçi Sayısı
Yazar1	4,47 Mn
Yazar2	2,68 Mn
Yazar3	1,77 Mn
Yazar4	1,17 Mn
Yazar5	1,04 Mn
Yazar6	850 B
Yazar7	848 B
Yazar8	786 B
Yazar9	721 B
Yazar10	711 B

Çalışmada kullanılan makaleler yazarların çalıştıkları günlük gazetelerin web sitelerinin arşivinden toplanmıştır. Yazarların makaleleri toplanırken, en güncel yazılan makaleden geriye doğru ardışık tarihlerde yazdığı makalelerin herhangi biri elenmeden 50 tanesi alınmıştır.

Her bir yazarın 200'er tweeti toplanırken ise yine en güncel paylaşılan tweetinden başlanarak geriye doğru gidilmiştir ve yazarın kendi görüşünü açık bir şekilde ifade etmeyen bazı "değersiz" tweetleri elenmiştir. Bu çalışma için "değersiz" olarak nitelendirilen ve elenen tweetlerin özellikleri şu şekildedir:

1. Retweetler,

2. Üç kelimeden az kelime sayısına sahip tweetler,

3. Facebook, instagram gibi diğer sosyal medya hesaplarında yaptıkları paylaşımlarından otomatik olarak atılan tweetler,

4. Haber siteleri başta olmak üzere bir web sitesi sayfasında bulunan Twitter'da paylaşma seçeneği ile paylaşılan tweetler,

5. Yaptıkları programlar başta olmak üzere kendi programları hakkındaki duyurular için attıkları tweetler,

6. Tek başına anlamlı olmayan, başka bir kullanıcıya yanıt olarak yazılan "mention" olarak adlandırılan sohbet tweetleri,

7. Bir özdeyişin paylaşıldığı, alıntının yapıldığı tweetler.

Bu çalışmada metinleri analiz edilen bazı yazarların kullanılan tweetlerinin paylaşıldığı tarih aralığı ile kullanılan makalelerinin yazıldığı tarih aralığı büyük ölçüde örtüşürken, uzun süre önce günlük gazete yazarlığını bırakan bazı yazarların ise kullanılan tweetlerinin paylaşıldığı tarih aralığı ile kullanılan makalelerinin yazıldığı tarih aralığı hiç örtüşmemektedir.

3. Metodoloji

Çalışmada uygulanan ilk yöntemde yazarların 200'er tweetinin her biri birer metin olarak değerlendirilmiştir. Bu bölümde 10 yazar için 50'şer makale ve 200'er tweet olmak üzere 2500 adetten oluşan bir veri kümesi kullanılmıştır. Tweetlerde ve makalelerde geçen kelimelerin kendileri, kökleri ve 3-gramlarının metinde geçme durumuna göre Binary, TF, TF-IDF ve Normalized-TF matrisleri ile 4 farklı arff dosyası üretilmiştir. Arff (*Attribute-Relation File Format*), WEKA aracına özgü bir dosya yapısıdır [7]. Üretilen arrf dosyaları şunlardır:

1. Binary : Terimin metinde geçme duruma göre 1 veya 0.

2. TF *(Term Frequency)* : Terimin metinde toplam geçme sayısı.

3. TF-IDF : TF*IDF.

IDF *(Inverse Document Frequency)* : $\log_e$(Toplam metin sayısı/ Terimin geçtiği metin sayısı).

4. Normalized-TF : Terimin metinde toplam geçme sayısı/Metindeki toplam terimin sayısı.

Kelimelerin kendilerinin kullanılarak üretilen arff 62742 özellik sayısına sahipken, bu özellik sayısı kelimelerin kökleri ile üretilen arff için 16255, 3-gramlarla üretilen arff için 18793'tür.

Çalışmada uygulanan ikinci yöntemde yazarların 200'er tweeti 40'ar adet olarak 5 eşit gruba bölünmüştür. Bu gruplardaki 40'ar tweet birleştirilerek bütün halinde tek bir metin olarak değerlendirilmiş, böylece her yazar için 5 adet birleştirilmiş tweet metni oluşturulmuştur. 10 yazar için 50'şer makale ve 5'er birleştirilmiş tweet metni olmak üzere bu bölümde toplam 550 adetten oluşan bir veri kümesi kullanılmıştır. Birinci yöntemde olduğu gibi yine Tweet metinlerinde ve makalelerde geçen kelimelerin kendileri, kökleri ve 3-gramları için benzer şekilde Binary, TF, TF-IDF ve Normalized-TF matrisleri ile 4 farklı arff dosyası üretilmiştir.

Üretilen arff dosyaları ile dört farklı test yapılmıştır:

1. Makaleler ile eğitim, tweetler ile test,
2. Tweetler ile eğitim, makaleler ile test,
3. Sadece makaleler ile 10 kez çapraz doğrulama
4. Sadece tweetler ile 10 kez çapraz doğrulama

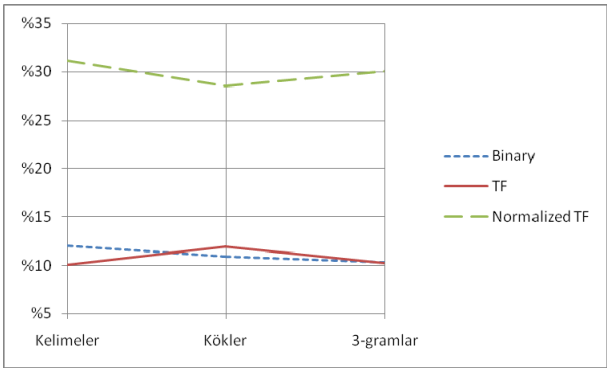
Daha sonra WEKA aracı içerisindeki, ki-kare istatistik değerini hesaplayarak özelliklerin değerini belirleyen ChiSquaredAttributeEval fonksiyonu kullanılarak tüm arff dosyaları için özellik sıralaması yapılmış ve ilk 1000 özellik seçilerek yeni arff dosyaları oluşturulmuş ve bu şekilde aynı testler tekrar yapılmıştır.

4. Deneysel Sonuçlar

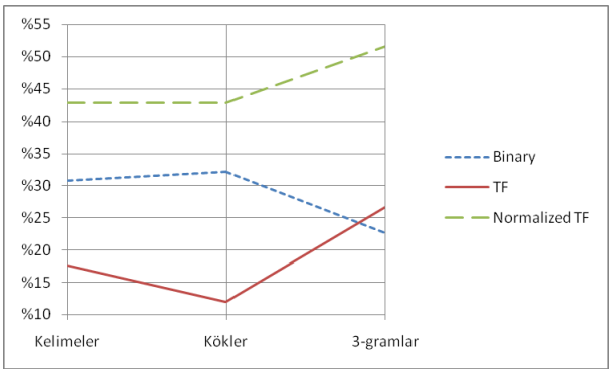
Yapılan tüm deneylerin büyük bir çoğunluğunda SMO sınıflandırıcı NB, J48 ve Bagging sınıflandırıcılarına göre daha başarılı sonuçlar verdiği için sadece SMO ile elde edilen değerler sunulmaktadır. Yapılan tüm deneylerin genelinde TF ve TF-IDF ile yapılan deney sonuçlarının birbirine çok yakın olması nedeniyle şekillerde TF-IDF sonucu gösterilmemiştir.

Özellik seçimi ile yapılan testlerde başarı oranları hemen hemen hepsinde düşmüş, sadece birleştirilmiş tweetler ile 10 kez çapraz doğrulama yapıldığında daha yüksek başarı elde edilmiştir. Bu yüzden özellik seçimi ile yapılan testlerden sadece bu testin sonucu sunulmaktadır. Tüm özellikler ile elde edilen sonuçlar Kelimeler, Kökler ve 3-gramlar şeklinde, özellik seçimi ile gerçekleştirilen testlerin sonuçları ise Kelimeler', Kökler' ve 3-gramlar' şeklinde ifade edilmiştir.

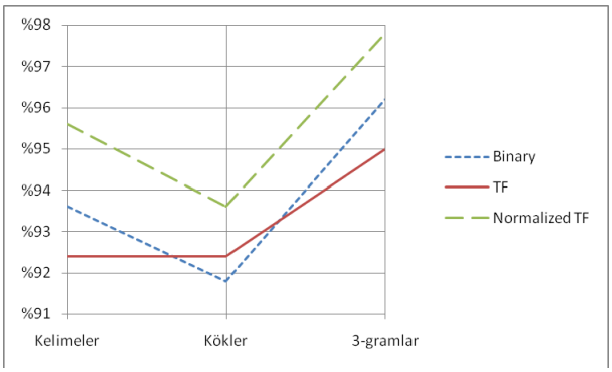
Her bir tweet birer metin olarak değerlendirildiğinde tüm öznitelikler kullanılarak ile elde edilen sonuçlar Şekil 1, 2, 3 ve 4'te görülmektedir.



Şekil 1. Makaleler eğitim, tekil tweetler test verisi olarak kullanıldığında elde edilen başarı oranları



Şekil 2. Tekil tweetler eğitim, Makaleler test verisi olarak kullanıldığında elde edilen başarı oranları

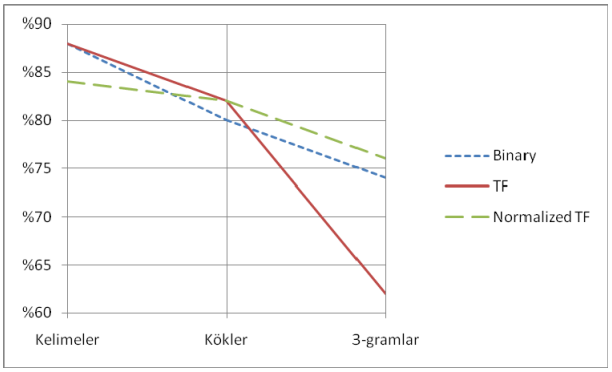


Şekil 3. Sadece makaleler ile 10 kez çapraz doğrulama yapıldığında elde edilen başarı oranları

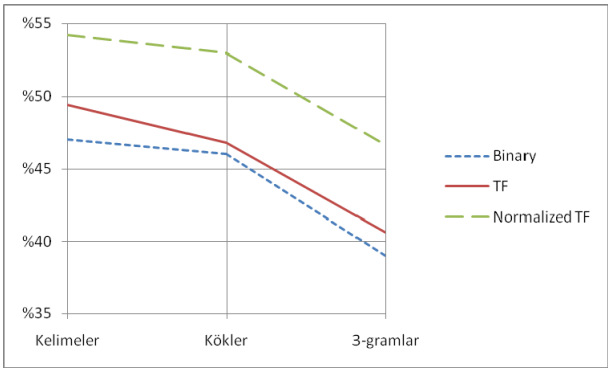
Şekil 4. Sadece tekil tweetler ile 10 kez çapraz doğrulama yapıldığında elde edilen başarı oranları

Şekil 1-4'te en başarılı sonuçların Normalized-TF ile alındığı açıkça görülmektedir. Şekil 3'te görüldüğü üzere makaleler kendi arasında 10 kez çapraz doğrulama yapıldığında 3-gramlar ile %97.8 gibi çok yüksek bir başarı oranı elde edilirken, Şekil 4'te görüldüğü gibi tekil tweetler kendi arasında 10 kez çapraz doğrulama yapıldığında yine 3-gramlar ile %57.85'lik bir başarı oranı ortaya çıkmıştır. Şekil 2'de görüldüğü gibi, tekil tweetler eğitim, makaleler test verisi olarak kullanıldığında 3-gramlar ile %51.6'lık bir başarı oranı yakalanırken, tam tersi şekilde makaleler eğitim, tekil tweetler test verisi olarak kullanıldığında ise en yüksek başarı oranının kelimelerin kendisi ile %31.15 olarak ortaya çıktığı Şekil 1'de görülmektedir.

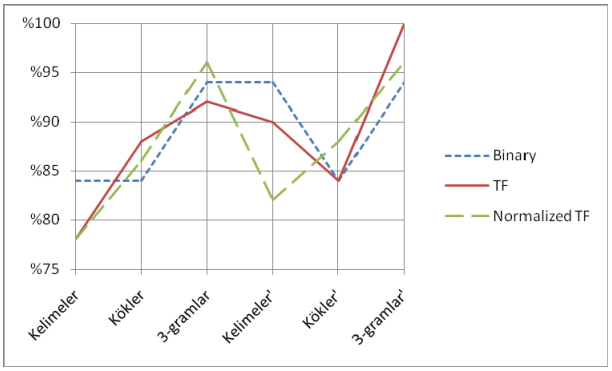
Çalışmanın ikinci bölümünde her bir tweet ayrı birer metin olarak değerlendirilmemiş, her bir yazar için toplanan 200'er tweet 40'ar adetlik 5 gruba ayrılmış ve her gruptaki tweetler birer bütün olarak tek bir tweet metni olarak değerlendirilmiştir. Bu şekilde kelimelerin kendileri, kökleri ve 3-gramlar için özellik seçimi yapmadan elde edilen sonuçlar Şekil 5, 6 ve 7 'de görülmektedir.



Şekil 5. Makaleler eğitim, birleştirilmiş tweetler test verisi olarak kullanıldığında elde edilen başarı oranları



Şekil 6. Birleştirilmiş tweetler eğitim, Makaleler test verisi olarak kullanıldığında elde edilen başarı oranları



Şekil 7. Sadece birleştirilmiş tweetler ile 10 kez çapraz doğrulama yapıldığında elde edilen başarı oranları

Çalışmanın bu bölümünde, yazarların makalelerini eğitim, birleştirilmiş tweetleri test verisi olarak kullanarak yapılan testte bu tweetlerin paylaşıldığı Twitter hesabının yazarını %88 gibi yüksek bir oranda başarılı tahmin edebildiğimiz Şekil 5'te görülmektedir. Şekil 7'deki sonuçlardan anlaşıldığı gibi birleştirilmiş tweetler kendi arasında 10 kez çapraz doğrulamayla test edildiğinde gayet yüksek başarı oranları yakalanmıştır. Bu testte, özellik seçimi yapılmış 3-gramlar ile %100 gibi bir başarı oranı elde edilmesinin sebebi kullanılan verinin örnek sayısının düşük olmasıdır. Şekil 6'da başarı oranlarının çok yüksek çıkmamasının sebebi de her yazar için 5'er birleştirmiş tweet metninin eğitim, 50'şer makalenin test olarak kullanılması dolayısıyla eğitim kümesinin çok küçük olmasıdır.

Ayrıca, sınıflandırma testleri sonrasında ortaya çıkan karışıklık matrisleri incelendiğinde, yazarların kullanılan makaleleri ve tweetlerinin yazıldığı tarih aralıklarının örtüşmesi veya örtüşmemesinin sınıflandırma başarısında bir etkisi olmadığı görülmüştür.

5. Tartışma

Bu çalışmada elde edilen sonuçlara göre birleştirilmiş bir tweet grubunun kim tarafından yazıldığını tespit etmenin bir tweetin kim tarafından yazıldığını tespit etmekten çok daha kolay olduğu ortaya çıkarılmıştır. Kimliği belirli olmayan bir Twitter kullanıcısının paylaştığı tweetlerin her birinin bir metin olarak değerlendirmesinden ise belirli sayıda tweeti birleştirilerek bir bütün olarak ele alınmasıyla çok daha uzun tweet metinleri oluşturulabilir ve aday köşe yazarlarının makaleleri ile karşılaştırılarak yüksek bir başarı oranıyla yazarın kimliğine dair tahmin yapılabilir. Ayrıca, bu şekilde yapılan testlerde birleştirilmiş tweetlerin kendi aralarında yapılan çapraz doğrulamalarda da çok başarılı sonuç vermesi, bir Twitter hesabının el değiştirip değiştirmediğinin anlaşılabileceğini de göstermektedir.

Kaynakça

[1] Statista, "Number of monthly active Twitter users worldwide from 1st quarter 2010 to 4th quarter 2015 (in millions)" www.statista.com. [Online]. Available: http://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/, 2016.

[2] M. F. Amaysalı, B. Diri ve F. Türkoğlu, "Farklı Özellik Vektörleri ile Türkçe Dokümanların Yazarlarının Belirlenmesi," *The Fifteenth Turkish Symposium on Artificial Intelligence and Neural Networks (TAINN'2006),* Muğla, Turkey, 21-24 June, 2006.

[3] R. Schwartz, O. Tsur, A. Rappoport ve M. Koppel, "Authorship Attribution of Micro-Messages," *Conference on Empirical Methods in Natural Language Processing*, pp. 1880–1891, Washington, USA, 18-21 October 2013.

[4] R. M. Green ve J. W. Sheppard, "Comparing Frequency- and Style-Based Features for Twitter Author Identification," *The Twenty-Sixth International Florida Artificial Intelligence Research Society Conference*, pp. 64-69, 2013.

[5] M. D. Akın ve A. A. Akın, "Türk Dilleri İçin Açık Kaynaklı Doğal Dil İşleme Kütüphanesi," *Electricity Engineering – Elektrik Mühendisliği*, vol. 431, pp. 38-44, 2007.

[6] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann ve I. Witten, "The weka data mining software: an update". *SIGKDD Explor*. Newsl. 11(1): 10–18, 2009.

[7] The University of Waikato, "Attribute-Relation File Format (ARFF)" cs.waikato.ac.nz. [Online]. Available: http://www.cs.waikato.ac.nz/ml/weka/arff.html, 2016.