YBS Ansiklopedi

www.YBSAnsiklopedi.com

Cilt 4, Sayı 4, Aralık 2017

Veri Madenciliği Yöntemleri ile Twitter Üzerinden Girişimcilik Analizi

Kubilay GAZİOĞLU[1], Şadi Evren ŞEKER[2]

1. Istanbul Üniversitesi, İktisat Fakültesi, Ekonometri Bölümü
2. İstanbul Şehir Üniversitesi, Yönetim Bilişim Sistemleri Bölümü

Özet

Bu çalışmanın amacı, sosyal medya kullanan kişilerin, kendi beyanarını esas alarak, başarılı girişimciler ile görece daha az başarılı sayılabilecek insanların tweet'lerini incelemek ve bu kişilerin atmış olduğu tweetler üzerinde yapılan analizler sonucunda başarılı kişileri ayırt edici belirgin belirgin farkları çıkarıp bunları öğrenen bir yapay zekâ tasarlamaktır. Bu yapay zekânın amacı, veri madenciliği, makine öğrenmesi ve veri bilimi tekniklerini kullanarak, herhangi bir kişinin tweet'lerini incelemek ve kişinin girişimcilik potansiyeli hakkında bilgi vermektir.

Anahtar Kavramlar: *Girişimcilik Analizi, Yapay Zeka,Twitter, Veri Madenciliği, Veri Bilimi*

Abstract

The purpose of this study is to develop an artificial intelligence system, which detects if there is any explicit difference between the tweets of successful entrepreneurs and the ones of relatively less successful people. This artificial intelligence, using data mining, machine learning and data science techniques, aims to scan any person's tweets and to analyze whether this person could be an entrepreneur or not.

Keywords: *Analysis of Entrepreneurship,Artifical Intelligence, Twitter, Data Mining, Data Science*

1. Giriş

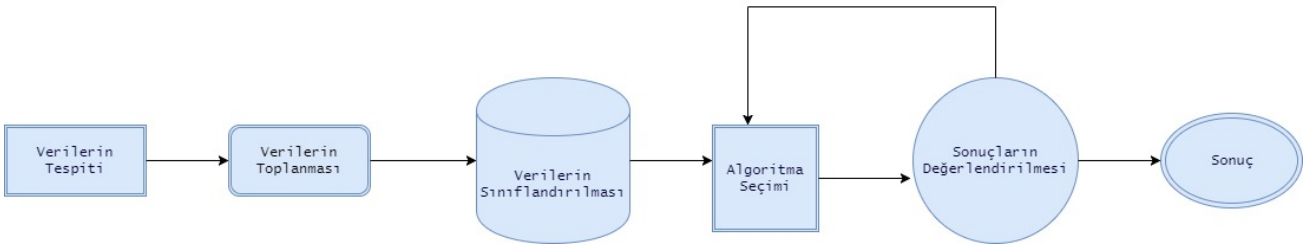
Girişimcilik, kelime anlamıyla, risk alma, yenilikleri yakalama, fırsatları değerlendirebilme ve tüm bunları hayata geçirebilme sürecidir[1]. Dolayısıyla, girişimci olarak adlandırdığımız insanların yukarıda belirtilen özelliklere sahip olması beklenir. Bu özellikler, Twitter gibi oldukça öznel fakat kamuya açık paylaşım platformlarında da kendi yansımasını kazanır. Genel anlamda, Twitter üzerinden kişilik analizi yapılabildiği gibi, bu analizlerden elde edilen sonuca göre kişinin sosyoekonomik statüsü de öngörülebilir olacaktır. Twitter'ın dünyanın en popüler 3. sosyal medya platformu olması [2], yapılan bu öngörülerin istatistiksel olarak daha güvenilir olmasını sağlamaktadır. Bu çalışmada bu platformun veri ambarı olarak seçilmesinin temel nedeni de budur.

Yapay zeka tasarımı, bir zeka oluşturarak istenen verileri toplayıp tümevarım yöntemi ile verilerin analizinin yapılabilmesi için başvurulan yöntemlerden biridir.[3] Yapay zeka çalışmaları genellikle insan zekası baz alınarak

bilgisayarlara algoritmalar geliştirilmesi şeklinde gerçekleşir.[4] Bu çalışmada veri madenciliği, makine öğrenmesi ve veri bilimi kullanılarak bir yapay zeka yaratılmış ve tweetler seçilen algoritmalara göre analiz edilmiştir.

2.Metodoloji

Uygulanan metodolojik sıralama aşağıdaki gibidir ve bu adımların akışı, Şekil 1'de görselleştirilmiştir.



Şekil-1 Metodoloji Adımları Akış Şeması

- Verilerin Tespiti: bu aşamada probleme özel olarak veri çekilmesi için uygun twitter hesapları belirlenmiş ve problem çözümüne yol gösterecek kutupsal sınıflar (görece olarak başarılı ve başarısız girişimciler) için twitter hesapları belirlenmiştir.
- Verilerin Toplanması: Bu aşamada twitter'a, Twitter API'ı üzerinden bağlantı sağlanarak bir önceki aşamada belirlenen hesaplardan veriler toplanmıştır.
- Verilerin Sınıflandırılması: Bu aşamada veri madenciliği teknikleri kullanılmış, veriler üzerinde özellik çıkarımı yapılmış ve daha sonra veriler için görece olarak başarılı ve başarısız girişimci etiketi eklenmiştir.
- Algoritma Seçimi : Bu aşamada veri madenciliği algoritmaları üzerinde testler yapılarak uygun algoritmanın seçilmesi ve seçilen algoritmanın, problem özelinde daha başarılı sonuçlar çıkarması için ayarlamalar yapılmıştır.
- Sonuçların Değerlendirilmesi: Bu aşamada, elde edilen algoritma sonuçlarının başarısı incelenmiş ve sonuçlardaki değerlere göre algoritmalar üzerinde çalışılarak, tatmin edici sonuçlar çıkarana kadar bir döngü içerisinde yeniden değerlendirme süreci sürdürülmüştür.
- Sonuç: Bu aşamada çalışan bir sistem kurgulanmış, başarılı olan algoritmalar ve sonuçları bu makale ile sunulmuş ve ileride geliştirilecek olan otomasyon sistemleri için yol gösterici sonuçlar elde edilmiştir.

3. Veriler

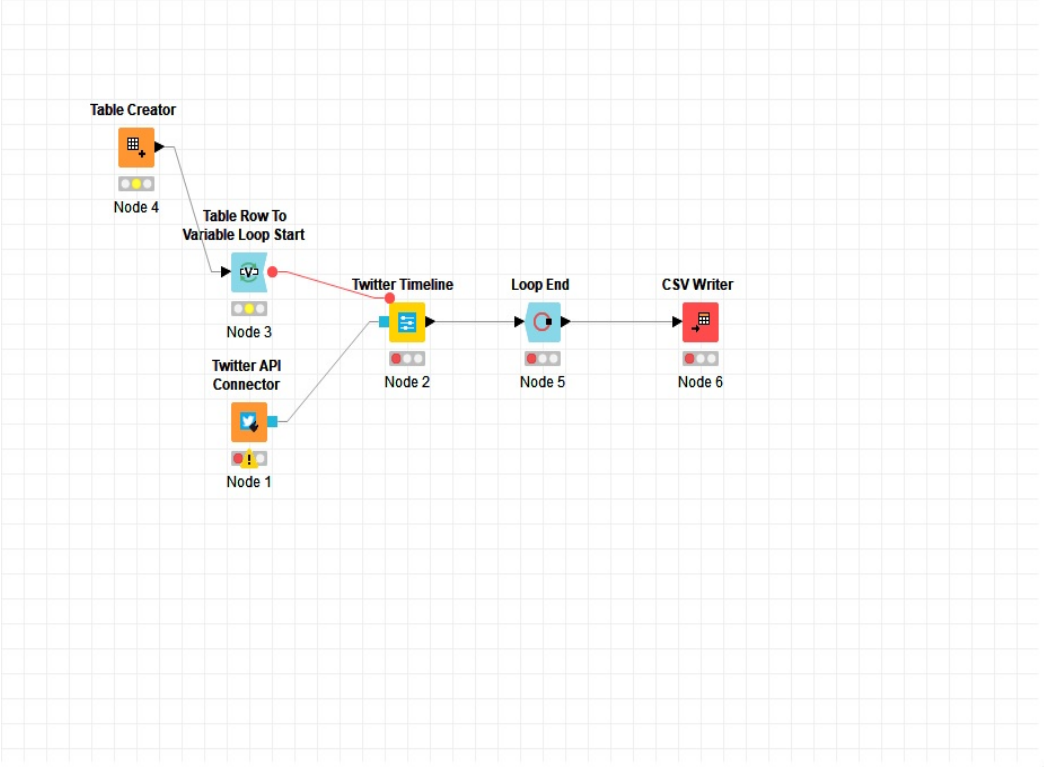
Bu çalışmada sektörlerinde bir ilki gerçekleştirmiş, kendi alanlarında başarılı olmuş girişimciler ile daha az başarılı olarak kabul edebileceğimiz veya bir girişimde bulunup başarısızlıkla sonuçlanmış insanların tweet'leri çekilmiş ve bu tweet'ler üzerinde veri analizi yapılmıştır. Toplam çekilen tweet sayısı 142.656'dır.

3.1 Verilerin Tespiti Aşaması

İki kutuplu bir (bipolar) bir yaklaşımla, iki uç kutupta bulunabilecek, başarısız ve başarılı insanlar internetten yararlanılarak liste haline getirildi ve bu kişilerin twitter hesapları tespit edildi. Bu hesapların belirlenmesinde, kişilerin kendi beyanları esas kabul edildi ve bu aşamada bir araştırmacının twitter hesaplarını takip etmesi ve kişilerin beyanlarını okuması sayesinde elle yapıldı. İki kutuplu yaklaşımda kişilerin bu kutupların en uç noktasında olması gerekmemekle birlikte bu kutuplardan birisine daha yakın olması, kişinin bu şekilde etiketlenmesi için yeterli görüldü.

3.2 Verilerin Toplanması

Twitter hesapları tespit edilen kişilerin Knime programı yardımı ile döngü oluşturularak tweetleri çekildi. Knime içinde bulundurduğu düğümler ile birlikte(node) farklı senaryolarda kullanabilecek Java tabanlı bir veri madenciliği programıdır[5]. Knime da twitter üzerinden veri çekebilmek için Twitter API Connector oluşturulması gerekmektedir Liste oluşturucu(table creator) ile liste oluşturulup tweet'leri çekilecek kişilerin isimleri listeye eklendi. Daha sonra döngü(table row to variable loop start) başlatılıp twitter zaman tüneli üzerinden(twitter timeline) kişiler listedeki sıralarına göre aratılmaya başlandı. Listedeki son kişi de aratıldıktan sonra döngü sonlandırılıp (loop end), CSV dosyasına tüm tweet'ler yazdırıldı(CSV Writer).Bu işlem başarılı girişimciler ile başarısız insanlar için iki kez tekrarlanıp, iki ayrı CSV'ye kaydedilmiştir. Knime iş akışı(workflow) görseli Şekil-2'de gösterilmiştir.



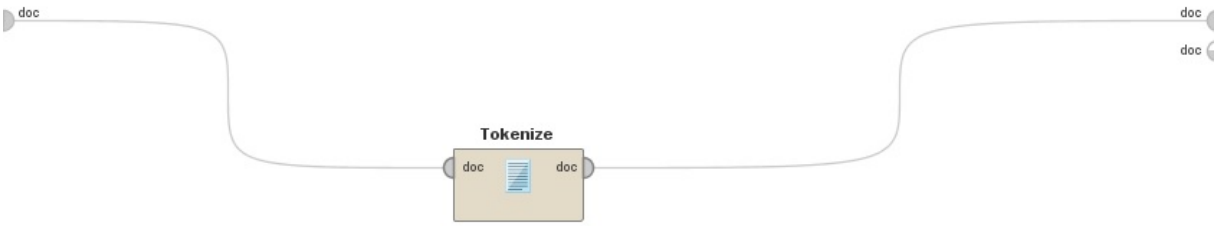
Şekil-2 Knime İş Akışı

3.3 Verilerin Sınıflandırılması

İki ayrı kutup için iki ayrı CSV formatındaki veriler etiketlenerek, birleştirilip tek bir CSV dosyasına dönüştürüldü. Bu sırada, başarılı insanlara 1, diğerlerine 2 etiketi verilerek iki ayrı gruba ayrıldı. Çalışmanın bundan sonraki aşamasında Rapid Miner programından yararlanıldı.

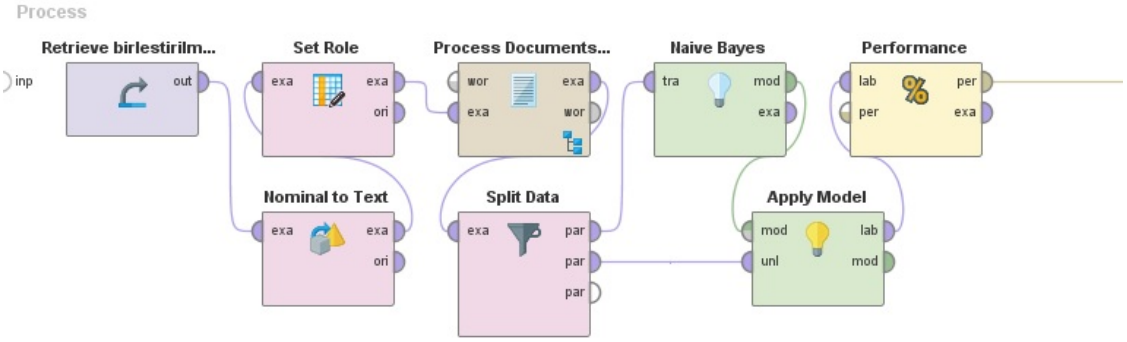
Rapid Miner 22 adet dosya formatındaki veriyi işleyebilen veri madenciliği, metin madenciliği [6] ve makine öğrenmesi algoritmalarını da kapsayan Java tabanlı bir yazılımdır.[7] Retrieve operatörü ile CSV formatındaki veri-

ler belirtilen adresten Rapid Miner'a çekilmiştir. Nominal to text operatörü ile nominal özelliklere sahip olan veriler, metin formatına dönüştürülmüştür. Set Role operatörü ile birleştirilen verilerdeki "id" sütunu hedef sütun olarak belirlendi. Veri girdilerinin sınıflandırılması için(tokenize), Process Document from Data operatöründen yararlanıldı. Aynı zamanda bu operatör ile kelime vektör dönüşümü(Word2Vec) yapıldı. [8] Process Document from Data operatörünün içindeki Tokenize işlemi Şekil-3'de gösterilmiştir.



Şekil-3 Tokenize İşlemi

Split Data operatörü elde edilen verileri 0.7 ve 0.3 olmak üzere rassal olarak iki gruba bölmüştür [9]. Bölünen bu iki grup içerisinde 0.7 'lik kısım eğitim (traning), 0.3'lük kısım test için kullanılmıştır. Apply Model operatörü ile veri setine istenilen model uygulandı. Şekil-4'de görüleceği üzere veri setine Naive Bayes algoritması uygulanmıştır.



Şekil-4 Rapid Miner Akış Şeması

4.Algoritma Seçimi

Bu çalışmada i5-6300 HQ Intel işlemci, 8 GB DDR4 2333 Mhz Ram ve x64 mimarisi özelliklerine sahip bir bilgisayar kullanılmıştır.

Algoritma seçiminde çalışmayı sınırlandıran en büyük unsur donanım yetersizliğidir. Bazı algoritmalar hiç çalıştırılamamıştır. Çalıştırılamayan algoritmalar aşağıdaki gibidir.

- Neural Net
- Auto MLP
- Linear Regression
- SVM
- Logistic Regression
- Gaussian Process

Rapid Miner'ın algoritma deposunda bulunan bütün algoritmalar denenerek çalışmayan algoritmalar elenmiş ve çalışan algoritmaların sonuçları, bu çalışmanın metodolojisindeki bir sonraki adıma taşınmıştır. Bu anlamda çalışan ve bir sonraki aşamaya taşınan algoritmaların detayları, 5. bölümde sunulmuştur.

5.İstatiksel Analiz ve Sonuçların Karşılaştırılması

CHAID, Decision Tree, Deep Learning, Gradient Boosted Tree, KNN, Naive Bayes ve Random Tree sonuç alınabilen algoritmalardır. Bu algoritmalar arasındaki başarı oranı en yüksek algoritma %99.89 ile Gradient Boosted Tree olmuştur.

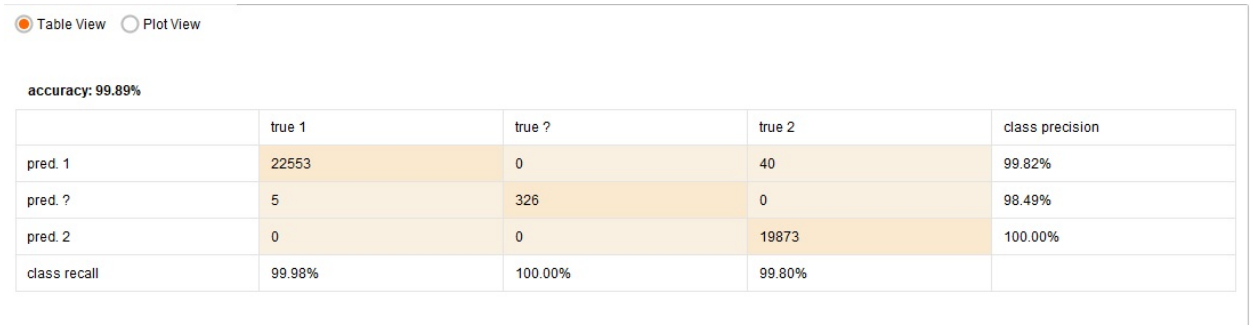


Table View Plot View

accuracy: 99.89%

	true 1	true ?	true 2	class precision
pred. 1	22553	0	40	99.82%
pred. ?	5	326	0	98.49%
pred. 2	0	0	19873	100.00%
class recall	99.98%	100.00%	99.80%	

Şekil-5 Gradient Boosted Tree Confusion Matrisi

Diğer algoritmaların başarılarını gösteren karışıklık matrisleri (confusion) de Şekil 7-11 arasında görselleştirilmiştir:

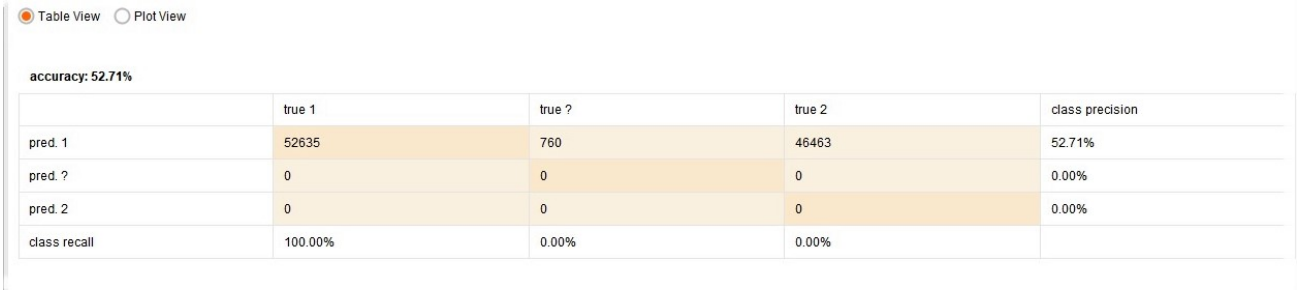


Table View Plot View

accuracy: 52.71%

	true 1	true ?	true 2	class precision
pred. 1	52635	760	46463	52.71%
pred. ?	0	0	0	0.00%
pred. 2	0	0	0	0.00%
class recall	100.00%	0.00%	0.00%	

Şekil-6 CHAID Confusion Matrisi

Table View Plot View

accuracy: 52.71%

	true 1	true ?	true 2	class precision
pred. 1	52635	760	46463	52.71%
pred. ?	0	0	0	0.00%
pred. 2	0	0	0	0.00%
class recall	100.00%	0.00%	0.00%	

Şekil-7 Decision Tree Confusion Matrisi

Table View Plot View

accuracy: 52.71%

	true 1	true ?	true 2	class precision
pred. 1	22558	326	19913	52.71%
pred. ?	0	0	0	0.00%
pred. 2	0	0	0	0.00%
class recall	100.00%	0.00%	0.00%	

Şekil-8 Deep Learning Confusion Matrisi

Table View Plot View

accuracy: 53.06%

	true 1	true ?	true 2	class precision
pred. 1	52614	720	46135	52.89%
pred. ?	0	40	0	100.00%
pred. 2	21	0	328	93.98%
class recall	99.96%	5.26%	0.71%	

Şekil-9 K-NN Confusion Matrisi

Table View Plot View

accuracy: 53.46%

	true 1	true ?	true 2	class precision
pred. 1	22546	286	19620	53.11%
pred. ?	0	40	0	100.00%
pred. 2	12	0	293	96.07%
class recall	99.95%	12.27%	1.47%	

Şekil-10 Naive Bayes Confusion Matrisi

Table View Plot View

accuracy: 52.71%

	true 1	true ?	true 2	class precision
pred. 1	52635	760	46463	52.71%
pred. ?	0	0	0	0.00%
pred. 2	0	0	0	0.00%
class recall	100.00%	0.00%	0.00%	

Şekil-11 Random Tree Confusion Matrisi

Şekillerde gösterilen 1. Sınıf, görece başarılı, 2. Sınıf, görece başarısız ve ? sınıfı ise eksik verileri ifade etmektedir. Bütün algoritmalar Rapid Miner üzerindeki varsayılan parametreleri ile test edilmiş ve herhangi bir değişiklik yapmadan sonuçları bu makaleye alınmıştır. Şekillerde görüleceği üzere GBT dışındaki algoritmalarda tek sınıfa yığılma olmuş ve sınıflar arasındaki farklılığın öğrenilmediği gözlemlenmiştir. Bu açıdan, sadece KNN algoritması ufak bir farklılık göstermektedir.

6.Sonuç

Bu çalışmada veri madenciliği, makine öğrenmesi ve veri bilimi yöntemleri Twitter üzerinden çekilen verilere uygulanmıştır. Toplamda 142.656 tweet çekilmiş ve bu tweet'ler üzerinde çalışılmıştır. Uygulanan algoritmalar arasında en yüksek başarı oranı Grandient Boosted Tree'ye aittir ve neredeyse dengeli ve iki kutuplu veri kümesi üzerinde, %99 üzerinde bir başarı ile doğru sınıflandırma yapabilmektedir.

Uygulanan makine öğrenmesi algoritmalarının sonucunda, başarılı girişimciler ile başarısız insanların cümle yapıları, kullandığı kelimeler ve yaptığı paylaşımlar arasında belirgin bir fark olduğu görülmüştür. Makine öğrenmesi sayesinde kişilerin twitter hesaplarını analiz ederek başarılı birer girişimci olup olamayacağı öğrenilebilmektedir. Yeterli donanım kapasitesine sahip olunduğunda, diğer makine öğrenme algoritmaları da denenebilir.

Kaynakça

[1]Aysun BAYHAN, "Girişim nedir? Girişimci Kime Denir?", https://paratic.com/girisimcilik-nedir-girisimci-kime-denir/ , Tarama : Eylül 2017

[2] Sosyal Ağ Siteleri: http://www.ebizmba.com/articles/social-networking-websites, Tarama: Eylül 2017

[3]Stuart J RUSSEL and Peter NORVIG, "Artificial Intelligence, A Modern Approach" syf. 3-5

[4] Yapay Zeka Nedir? Internet: http://uzaycagi.com/yapay-zeka-nedir/ Tarama: Eylül 2017

[5]Bernd JAGLA, "Extending KNIME for next-generation sequencing data analysis", https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btr478

[6] Seker, Sadi Evren. "Metin Madenciliği (Text Mining)." YBS Ansiklopedi, v. 2, is 3, pp. 30-32, Eylül 2015

[7] YILDIZ, Merve, and Şadi Evren ŞEKER. "Veri Madenciliği Araçları (Data Mining Tools).", YBS Ansiklopedi v. 3, is. 4, pp. 10 – 19, Aralık 2016

[8] Amine YEŞİLYURT, Şadi Evren ŞEKER, "Metin Madenciliği Yöntemleri ile Twitter Duygu Analizi", YBS Ansiklopedi, v. 4, is. 2, pp. 26- 36, Haziran 2017

[9] Seker, Sadi Evren, and Khaled Al-Naami. "Sentimental analysis on Turkish blogs via ensemble classifier." Proceedings of the International Conference on Data Mining (DMIN). The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp), 2013.