YBS Ansiklopedi

www.YBSAnsiklopedi.com

Cilt 4 Sayı 2, Haziran 2017

Veri Madenciliği Yöntemleri ile Twitter Üzerinden MBTI Kişilik Tipi Analizi

Kader BASTEM[1], Şadi Evren ŞEKER[2]

1. Fırat Üniversitesi, Teknoloji Fakültesi, Yazılım Mühendisliği Bölümü

2. İstanbul Şehir Üniversitesi, Yönetim Bilişim Sistemleri Bölümü

Özet

Bu çalışmanın amacı sosyal medya üzerinde insanların yapmış olduğu paylaşımları kullanarak insanların kişilik analizini yapan bir yapay zeka sistemi geliştirmektir. Makine öğrenmesi, veri madenciliği ve veri bilimi tekniklerini kullanarak, insanların twitter üzerinden atmış oldukları tweet'lere bakılarak MBTI kişilik tiplerinden kişiye en yakın tipe sınıflandırılması ve kişilik tahmini gerektiren, insan kaynakları veya kariyer planlama gibi alanlarda kullanılmasını sağlamaktır. MBTI kişilik göstergesi, gerek özel hayatta, gerek toplum hayatında, gerekse iş hayatında kullanılabilen ve verimliliğin oldukça artmasına olanak sağlayan sınıflandırma yöntemlerinden birisidir ve bu proje kapsamında kullanılmıştır.

Anahtar Kavramlar: Kişilik analizi, twitter ,mbti, veri madenciliği.

Abstract

The aim of this study is to develop an artificial intelligence system that uses people's social media sharing to analyze people's personality. Machine learning is the use of data mining and data mining techniques to look at the tweets people have twittered and to use them in areas such as human resources or career planning where the personality classification is needed and personality prediction is required. MBTI personality demonstration is one of the classification methods that can be used in private life, in public life, in business life and in the productivity, and it is used within the scope of this project.

Keywords: *Personality analysis ,twitter, mbti ,data mining.*

1. Giriş

Bir tanıma göre, kişilik, insanın, sosyal yaşamında sergilediği duygu ve düşüncelerinin sonucudur.[1]. Myers-Briggs kişilik tipi Göstergesi (MBTI) [2], insanların dünyayı nasıl algıladığını ortaya koyan bir kişilik analizi yöntemidir. MBTI'ın temelleri psikiyatr Carl G. Jung tarafından atılmıştır. Isabel Myers ve Katherine Briggs Jung'ın yaptığı çalışmalardan yararlanarak, kişilikler dört ana gruba bölmüş ve içe dönük olma, dışa dönük olma gibi tercihlere göre dört harfli ve toplamda 16 adet farklı kişilik tipi tanımlamıştır[3]. İlk önerildiği yıldan bu yana

kişilik analizi, kariyer planlama ve insan davranışlarını anlama gibi alanlarda en çok kabul gören yöntemlerden birisidir[4].

İnsanlar, sosyal medya üzerinde daha rahat hareket edip, düşüncelerini daha özgürce paylaşabildikleri için, kişilik özelliklerini yapmış oldukları paylaşımlardan analiz edip bir sonuca varmak daha kolaydır. Sosyal medya üzerinde yapılan çalışmalar, kişilik analizi için bu kapsamda önem gösterir.

Bu alanda yapılan bir diğer çalışma Barbara Plank tarafından, 6 batı dilinde(Hollanda, Almanca, Fransızca, İtalyanca, Portekizce ve İspanyolca) paylaşılmış olan tweet bilgileri ve cinsiyet bilgilerini veri olarak işlemektedir. Seçilen 6 dil twitter'da en çok rastlanan 15 dil arasından seçilmiştir. Diller göre atılan tweet'ler gruplanmış ve her gruptaki MBTI tipleri arasında en fazla geçen tipler ortaya çıkarılmıştır. Ayrıca çalışmada kadın ve erkek gruplar arasında MBTI tiplerinin dağılımı da ayrıca incelenmiştir. Çalışmada istatistiksel model olarak Logistic Regrezizasyon kullanılmıştır[5].

2. MBTI kişilik tipleri

Bu çalışmada, Myers ve Briggs tarafından geliştirilen her bir psikolojik tip "MBTI tipi" olarak adlandırılmıştır. Söz konusu kişilik profillerinin açılımları şunlardır

E :Dışa dönük / I :İçe dönük

S: Sağduyulu / N:Sezgilerini kullanan

T:Düşünen / F: Hisseden

J:Yargılayan / P: Kavrayan

Yukarıdaki her satırda iki alternatiften birisini içeren 4 harfli kişilikler ile toplam 16 farklı kişilik tipi olasılığı bulunmaktadır. Bu olasılıklar aşağıdaki şekilde sıralanabilir:

INFJ: Hedeflerini gerçekleştirmek adına adım atan idealist kişilerdir.[6].

INTJ: Kendine güvenen, lider kapasiteli ve net kişilerdir.[6].

INTP: Düşüncelere önem veren konsantrasyonu yüksek, düşüncelerini kolay ifade eden[6].

INFP: Kişisel değerlerini önemseyen kişilerdir ve kararlarında bu değerlerin etkileri görülür.[6].

ENTP: Çevresindeki şeylere karşı meraklı, her konuda en iyi ve tek olmak isteyen kişilerdir[6].

ENFJ: Yardım etme konusunda istekli, iletişim konusunda doğuştan yetenekli ve lider kişileridir.[6].

ENTJ: Sorumluluk almayı ve yönetmeyi sever. Grup çalışmalarında hedefe ulaşmak için insanları kolayca organize eden kişilerdir[6].

ISTP: Pratik, iyimser, az ve öz konuşan kişilerdir [6].

ENFP: Güzel sanatlara eğilimli, plan yapmayı sevmeyen, kendini ifade ederken kelimelerden değil eylemlerden yararlanmaktan hoşlanan kişilerdir.[6].

ISFP: Güzel sanatlara ilgilidir. Hisleri kuvvetlidir ve estetik duygusu güçlüdür[6].

ISFJ: Seçilmiş ve saygın olmayı sever. Başkalarına yardım etmekten hoşlanan ve bunu görev edinen kişilerdir[6].

ISTJ: Karar verme mekanizması güçlüdür. Her şeyin bir düzende olmasını tasdik etmek isterler, sessiz ve ciddi kişilerdir[6].

ESTP: Başka insanları yönlendirme yeteneğine sahip, girişimci kişilerdir[6].

ESFP: Daima pozitif, yalnız kalmaktan çekinen ve yumuşak huylu kişilerdir[6].

ESTJ: Çevresindekilere sürekli kurallar koyan, iş akışı için yönergelerden ve belirlenen talimatların dışına çıkmayan kişilerdir[6].

ESFJ: Tüm tipler içinde en sosyal olan kişilerdir. Başkalarının fikirlerine saygılı ve insanlarla uyumlu kişilerdir[6].

3.Veriler

Bu çalışmada, toplumda etkisi oldukça fazla olan, dünyaca tanınmış ve milyonlarca insanı etkisinde bırakmış meşhur kişilerin tweet'leri çekilmiş ve bu tweetler üzerinde veri analizi yapılmıştır. MBTI tipi bilinen 64 kişinin 800 ila 1000 arasında, toplam 63384 tweet'i toplanmıştır. Veriler 16 MBTI türünü de içermekte olup her biri için ayrı tweet örneği bulundurmaktadır. 16 MBTI tipi için alınan örnek kişilerin isimleri ve hangi tipte olduklarını gösteren referanslar tablo 1'de sunulmaktadır.

Tablo 1. Farklı MBTI Kişilik tipleri için çalışmada kullanılan Ünlüler.

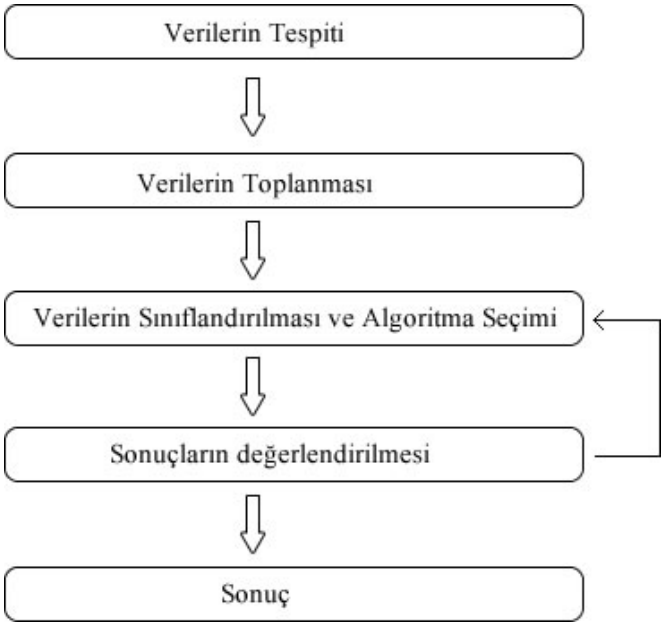
ENFP Will Smith[7] Alicia Silverstone[8] Robin Williams[7] Jerry Seinfeld[9]	**INFP** Fred Savage[8] Lisa Kudrow[8] Amy Tan[8] James Taylor[8]	**ENFJ** Barack Obama[8] Ronald Reagan[8] Ben Stiller[8] Bob Saget[8]	**INFJ** Nicole Kidman[8] Evangeline Lilly[8] Mel Gibson[8] George Harrison[9]
ESTJ Bill Frist, M.D.[8] Billy Graham[8] Emma Watson[9] Alec Baldwin[9]	**ISTJ** Jackie Joyner-Kersee[8] Evander Holyfield[8] Natalie Portman[9] Sting[7]	**ESFJ** Danny Glover[7] Martha Stewart[9] Ariana Grande[9] Selena Gomez[9]	**ISFJ** Robin Roberts[8] Kristi Yamaguchi[8] Johnny Carson[8] 50 Cent[9]
ENTP Rachael Ray[8] David Spade[8] Al Yankovic[7] Celine Dion[7]	**INTP** Richard Dawkins[9] Kristen Stewart[8] Tiger Woods[8] Henry Mancini[10]	**ENTJ** Jim Carrey[8] Bill Gates[9] Steve Martin[8] Patrick Stewart[8]	**INTJ** Hillary Clinton[8] Arnold Schwarzenegger[7] Charles Rangel[8] Russell Crowe[9]
ESTP Donald Trump[8] Nicolas Sarkozy[7] Madonna[7] Taylor Swift[9]	**ISTP** Tom Cruise[9] Keith Richards[8] Michael Jordan[7] Milla Jovovich[7]	**ESFP** Justin Bieber[9] Katy Perry[9] Chloe Grace Moretz[9] Kyle petty[8]	**ISFP** Rihanna[9] Kevin Kostner[7] Britney Spears[7] Justin Timberlake[9]

3.Metodoloji

Bu çalışmadaki metodolojide 4 ana aşama bulunmaktadır:

1. Verilerin Tespiti
2. Verilerin Toplanması
3. Verilerin Sınıflandırılması ve Algoritma Seçimi
4. Sonuçların değerlendirilmesi

Bu aşamaların görsel olarak birbiri ile ilişkisi ve çalışmanın ana hatları ile akışı Şekil 1'de görselleştirilmiştir.



Şekil 1 Metodoloji Akış Şeması

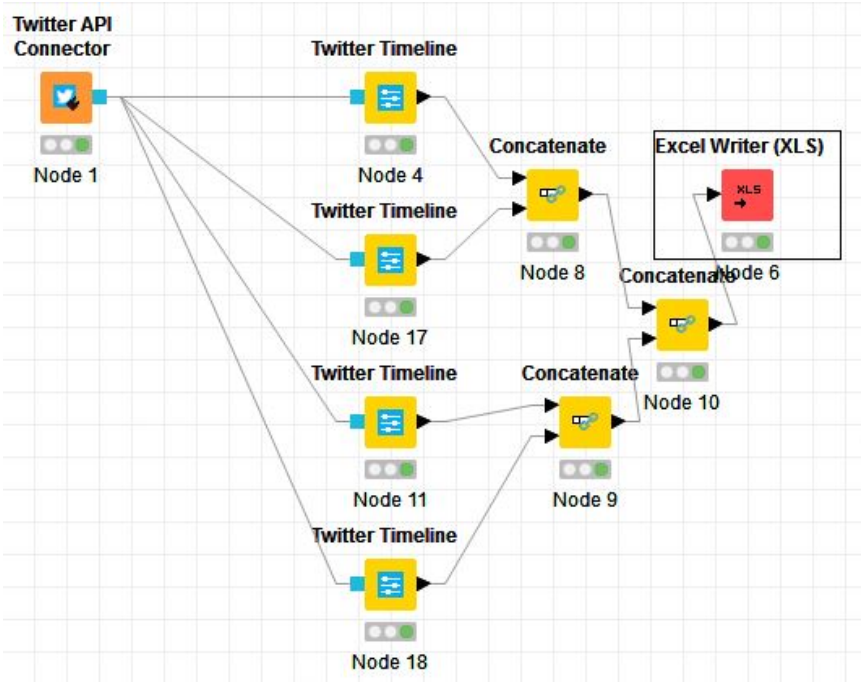
Şekil 1'de sunulan metodoloji akış şemasındaki her aşamanın açıklaması aşağıda açıklanmıştır:

1. **Verilen tespiti aşaması: İnternet'ten** yararlanılarak, seçilen ünlü kişilerin MBTI kişilik grubu araştırılarak bir liste hazırlandı ve bu kişilerin twitter hesapları tespit edildi.
2. **Verilerin Toplanması:** Twitter hesapları tespit edilen kişilerin kişilik tipleri 16 gruba ayrıldı ve Knime programı ile her kişi için 800 ila 1000 arasında değişen sayıda tweet otomatik olarak indirildi. Daha sonra ayrılan 16 grup sınıflandırılma ve algoritma seçimi için tek dosya haline getirildi.
3. **Verilerin Sınıflandırılması ve Algoritma Seçimi:** Rapid Miner programı ile sınıflandırma işlemleri gerçekleştirilen verilerin algoritma seçimi programın çalışması sonucu alınan başarı oranına göre seçilmiştir.
4. **Sonuçların değerlendirilmesi: Algoritma** sonuçları dikkate alındığında %54.98 oranında başarı gösteren çalışma özellikle insan kaynakları ve çalışan davranışlarını anlama gibi kişilik tahmini gerektiren alanlarda kullanımının başarı göstereceği belirlenmiştir.

3.1 Kullanılan programlar

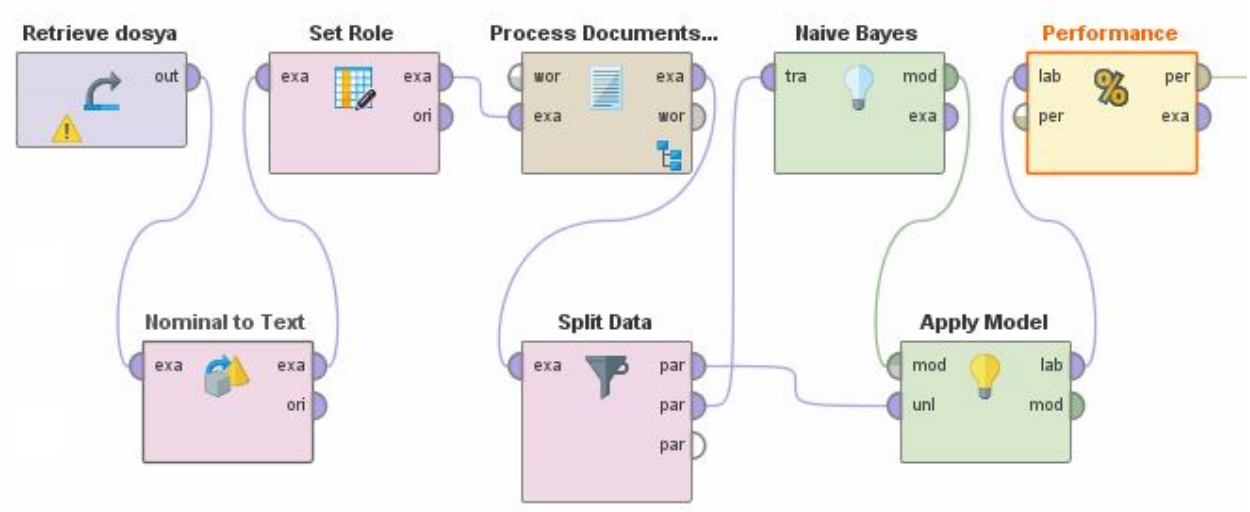
Çalışmada, twitter API erişimi ve otomatik olarak ilgili kişinin attığı tweet'lerin aranması ve indirilmesi için Knime ve veri bilimi aşamaları için de Rapid Miner programları kullanılmıştır. Veriler CSV formatında toplanarak iki yazılım arasında veri akışı sağlanmıştır.

Knime: Knime açık kaynak kodlu bir veri analiz platformudur ve üzerinde görsel işlemlerin gerçekleştirilebileceği görsel bir ara-yüz bulunmaktadır. Bu ara-yüzde node'lar birbirine bağlanır ve yapılacak işleme uygun bir akış şeması elde edilir ve node'lar çalışarak bir sonuç üretir. Bu çalışmanın veri toplama aşamasındaki Knime görseli şekildeki gibidir.



Şekil 2 Çalışmanın Knime Görseli

Rapid Miner: Kullanılan diğer program ise Rapid Miner'dır. Veri madenciliği ve makine öğrenme algoritmalarını da kapsayan Rapid Miner, literatürde bilinen çoğu veri bilimi algoritmasını kullanma imkanı sağlamaktadır.. Veri analizi, önişleme, sınıflama, kümeleme, birliktelik kuralları çıkarımı, nitelik seçimi işlemlerini içermektedir [11]. Bu çalışmanın Rapid Miner görseli aşağıdaki gibidir



Şekil 3 Çalışmanın Rapid Miner Görseli

Çalışmada Knime ile veriler toplanmış ve Rapid Miner programı ile algoritma uygulaması ve sınıflandırma işlemi gerçekleştirilmiştir. Verilerin işlenmesi sırasında metin kaynağı olarak gelen veriler üzerinden Word-gram (kelime sayısı) yöntemi ile özellik çıkarımı yapılmış ve bu sayılar üzerinde makine öğrenmesi algoritmaları denenmiştir [12]. Ayrıca başarı ölçümü sırasında eğitim ve test kümeleri, orijinal veri üzerinden rasgele olarak %70 eğitim , %30 test kümesi seçilerek oluşturulmuştur.

4.İstatistiksel Analiz ve Karşılaştırma

Rapid Miner kütüphanesinde yer alan farklı makine öğrenmesi ve istatistiksel metotlar bu çalışma kapsamında kullanılmıştır. Çalışma kapsamında, algoritmaların çalıştırıldığı bilgisayar i7-4702 MQ İntel işlemci 8 GB ram x64 taban özelliklerine sahip bir adet bilgisayar kullanılmıştır.

Toplam 63bin civarı tweet üzerinde, bütün algoritmaların çalışması, sınırlı hafıza ve donanım kapasitesi ile, uzun süren denemeler gerektirebilmektedir. Bazı algoritmalar ise donanım sınırları yüzünden ya hiç çalıştırılamamış ya da günler süren uzunlukta işlem gerektirdiği için çalıştırılmayarak durdurulmuştur. Bu denemelerde uzun süre bilgisayarın çalışmasına rağmen sonuç alınamayan algoritmalar: K-NN, ID3, Decision Tree, Decision Stump, Neural Net, Deep Learning, Auto MLP, Linear Regresyon, Support Vector Machine yöntemlerdir. Bununla birlikte, Naive Bayes (NB), Random Tree (RT) ve Gradiend Boosted Tree (GBT) algoritmalarında sonuç alınabilmiştir ve bu başarı değerleri tablo 2'de sunulmuştur:

Tablo 2 Kullanılan Üç Metodun Accuracy (Kesinlik) Değerlerinin Karşılaştırılması

NB	GBT	RT
%54.98	%48.86	%6.16

Naive Bayes (NB), Random Tree (RT) ve Gradiend Boosted Tree (GBT) sınıflandırma algoritmalarının gösterdikleri başarım sonuçlarına etkisi incelenmiştir. Elde edilen deneysel sonuçlarda en iyi performans gösteren sınıflandırma algoritması %54.98 ortalama doğruluk başarı oranıyla NB olmuştur. Kullanılan algoritmaların performans görselleri aşağıdaki gibidir.

accuracy: 6.16%

	true ESFP	true ISFP	true ESTP	true ISTP	true INTJ	true ENTJ	true ENFP	true ENFJ	true ENTP	true INFP	true INFJ	true ISFJ	true ESFJ	true ISTJ	true ESTJ	true INTP
pred. ES...	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
pred. ISFP	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
pred. ES...	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
pred. ISTP	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
pred. INTJ	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
pred. EN...	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
pred. EN...	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
pred. EN...	2818	2824	2762	2822	2821	2783	2814	2731	2749	2809	2841	2776	2787	2793	2840	2398
pred. EN...	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
pred. INFP	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
pred. INFJ	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
pred. ISFJ	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
pred. ES...	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
pred. ISTJ	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
pred. ES...	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
pred. INTP	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
class rec...	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	100.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%

Şekil 4 Random Tree Confusion Matrisi

accuracy: 48.86%

	true ESFP	true ISFP	true ESTP	true ISTP	true INTJ	true ENTJ	true ENFP	true ENFJ	true ENTP	true INFP	true INFJ	true ISFJ	true ESFJ	true ISTJ	true ESTJ	true INTP
pred. ES...	1709	655	472	380	604	505	385	450	450	298	310	451	631	278	319	150
pred. ISFP	4	1227	147	314	11	344	55	36	218	204	29	124	255	179	55	179
pred. ES...	3	7	1228	1	19	2	1	6	59	6	3	1	15	0	7	6
pred. ISTP	3	2	7	1443	9	1	1	3	2	0	2	2	3	1	1	4
pred. INTJ	36	4	61	3	1243	9	11	47	4	64	4	10	24	0	37	39
pred. EN...	350	225	292	230	353	1214	249	259	303	349	213	332	367	134	165	283
pred. EN...	519	527	325	296	319	536	1809	424	328	596	320	271	397	783	401	517
pred. EN...	3	2	71	4	39	11	11	1225	3	15	2	1	5	0	11	5
pred. EN...	7	4	5	6	4	15	10	4	1247	9	3	10	9	2	2	3
pred. INFP	0	21	14	7	6	7	2	6	3	1104	6	0	2	4	5	5
pred. INFJ	2	6	1	0	3	2	3	8	4	0	1817	1	2	2	2	1
pred. ISFJ	8	7	2	1	5	5	12	9	5	5	5	1474	11	8	5	8
pred. ES...	70	49	38	41	36	65	51	40	30	46	50	27	864	45	37	36
pred. ISTJ	0	1	4	0	3	7	0	1	0	1	1	0	0	1224	0	0
pred. ES...	101	84	90	96	164	55	210	208	93	104	72	69	200	130	1789	102
pred. INTP	3	3	5	0	3	5	4	5	0	8	4	3	2	3	4	1060
class rec...	60.65%	43.45%	44.46%	51.13%	44.06%	43.62%	64.29%	44.86%	45.36%	39.30%	63.96%	53.10%	31.00%	43.82%	62.99%	44.20%

Şekil 5 Gradiend Boosted Tree Confusion Matrisi

accuracy: 54.98%

true ESFP	true ISFP	true ESTP	true ISTP	true INTJ	true ENTJ	true ENFP	true ENFJ	true ENTP	true INFP	true INFJ	true ISFJ	true ESFJ	true ISTJ	true ESTJ	true INTP	class pr...
1571	339	132	146	165	196	207	122	129	153	114	107	408	234	122	119	36.84%
185	1464	126	127	90	135	90	81	129	152	70	123	224	155	72	94	44.14%
48	47	1531	21	61	31	35	58	70	37	18	27	45	18	41	33	72.18%
76	89	66	1831	49	69	48	47	74	38	36	40	59	45	39	37	69.28%
57	45	108	45	1467	136	92	197	51	72	40	61	89	43	85	66	55.28%
51	39	46	54	61	1013	97	61	67	79	32	47	39	38	106	102	52.43%
172	140	81	107	144	240	1567	144	145	181	91	118	166	124	175	130	42.07%
58	58	112	31	149	109	63	1409	68	57	40	51	78	52	75	46	57.37%
41	57	32	31	48	51	71	29	1462	71	25	56	54	19	33	38	69.03%
67	68	88	49	78	126	79	71	115	1374	55	71	78	63	74	50	54.83%
66	58	40	71	39	90	50	36	52	69	2044	40	49	78	49	44	71.10%
86	71	61	47	59	65	64	80	99	95	30	1769	67	76	42	49	64.09%
90	60	88	44	89	88	71	94	86	89	34	59	1178	30	80	72	52.31%
117	123	60	88	90	107	93	72	63	103	76	99	91	1704	83	91	55.69%
42	59	79	25	106	128	69	122	50	70	45	33	73	36	1642	59	62.24%
91	107	112	105	126	199	118	108	89	169	91	75	89	78	122	1368	44.90%
55.75%	51.84%	55.43%	64.88%	52.00%	36.40%	55.69%	51.59%	53.18%	48.91%	71.95%	63.72%	42.27%	61.01%	57.82%	57.05%	

Şekil 6 Naive Bayes Confusion Matrisi

5.Sonuç

İnsanların iletişimde bulunurken kullandığı kelimeler , cümleler kişiliklerini de yansıtır ve insanların düşüncelerini paylaşmak adına sosyal medya üzerinde kendilerini daha rahat ifade ettikleri görülmüştür. Yapılan çalışmamız sonucunda iş görüşmesi gibi karşımızdaki insanı tanımadığımız veya kişi hakkında sınırlı bilgiye sahip olunan durumlarda sosyal medya hesaplarının kişilik analizi için kullanılabileceğine dair %54'e varan olumlu sonuçlar elde edilmiştir. İnsan kaynakları danışmanlığı yapan firmalar ve özel sektör kuruluşlarının insan kaynakları bölümlerinde aday seçiminde kişilik faktörlerini tespit etmeye yönelik kullanıma gibi farklı gerçek hayat uygulamalarına imkan verecektir.

Bu çalışma tek bir bilgisayar üzerinden sınırlı donanım kaynakları ile yürütüldüğü için çok sayıdaki makine öğrenmesi algoritması çalıştırılamamıştır. Gelecekte yapılabilecek çalışmalardan birisi de, projenin çok daha fazla tweet içerecek ve çok daha geniş donanım kaynakları ile farklı algoritmalar deneyebileceği büyük veri platformlarında yeniden tasarlanarak çalışmasıdır.

Kaynakça

[1] Ozsoy Emrah; Yıldız Gultekin (2013),Kişilik Kavramının Örgütler Açısından Önemi: Bir Literatür Taraması, v.1, is.2, pp. 2

[2] Briggs-Myers, I., & Briggs, K.C. (1985). Myers-Briggs Type Indicator (MBTI). Palo Alto, CA: Consulting Psychologists Press.

[3] Şadi Evren ŞEKER (2014), *" Myers Briggs Tip Göstergesi (Myers-Brigs Type Indicator) ",*YBS Ansiklopedi, v .1,is .1, pp. 30 - 34

[4] Myers-Briggs Kişilik Tip Belirleme Envanteri, http://www.mentorink.com/kisilik-testi, Tarama: 2017

[5] Verhoven, Ben; Daelemans, Walter; Plank, Barbara(2016),TWISTY: a Multilingual Twitter Stylometry Corpusfor Genderand Personality Profiling, syf 1-3

[6] 2013, MBTI yöntemi ve genel açıklama, http://karaktertestleri.blogspot.com.tr/2013/01/mbti-kisilik-test-yontemi-nedir.html,2017

[7] Kişilik Tipleri, https://www.16personalities.com/tr/ki%C5%9Filik-tipleri,2017

[8] Psychological Type Profiles, http://www.typelogic.com/2017

[9] Kişilik Karakterleri, https://mbtiturkiye.wordpress.com/,2017

[10] Famous INTPS, http://www.intp.org/famous.html,2017

[11] Kaya Mumine; Ozel S.A(2014), Açık Kaynak Kodlu Veri Madenciliği Yazılımlarının Karşılaştırılması,

[12] Şadi Evren ŞEKER, (2015), *"Metin Madenciliği (Text Mining)",* YBS Ansiklopedi, v. 2, is. 3, pp. 30-32