Nur Alabuğa

YouTube Yorumları İçin Spam Algılama

1.Amaç

Spam araştırması için toplanan herkese açık bir yorumlar kümesi veriseti üzerinden makine öğrenmesi algoritmaları kullanılarak yöntem karşılaştırması yapılması amaçlanmıştır.

2.Giriş

Proje implementasyonu Python dilinde gerçekleştirilmiştir. Veriseti çok yaygın olarak kullanılmakta olan UCI Makine Öğrenimi Havuzundan toplanmıştır. Beş adet özniteliğe sahip beş adet farklı video bulunmaktadır. Bu öznitelikler "COMMENT_ID, AUTHOR, DATE, CONTENT, CLASS" olarak adlandırılmıştır (Şekil 2.1). Proje kapsamında ilgilenilen öznitelik "CONTENT", yani video yorumlarının bulunduğu sütun ile yorumun spam olup olmadığını gösteren "CLASS" sütunudur.

COMMENT_ID	AUTHOR	DATE	CONTENT	CLASS
LZQPQhLyRh80UYxNu	Julius NM	2013-11-07T06:20:48	Huh, anyway check out this you[tube] channel: kobyoshi02	1
LZQPQhLyRh_C2cTtd9	adam riyati	2013-11-07T12:37:15	Hey guys check out my new channel and our first vid THIS IS US THE MONKEYS	1
LZQPQhLyRh9MSZYnf	Evgeny Murashkin	2013-11-08T17:34:21	just for test I have to say murdev.com	1
z13jhp0bxqncu512g22	ElNino Melendez	2013-11-09T08:28:43	me shaking my sexy ass on my channel enjoy ^_^ï»¿	1
z13fwbwp1oujthgqj04	GsMega	2013-11-10T16:05:38	watch?v=vtaRGgvGtWQ Check this out .ï»¿	1
LZQPQhLyRh9-wNRtlZ	Jason Haddad	2013-11-26T02:55:11	Hey, check out my new website!! This site is about kids stuff. kidsmediausa .	1
z13lfzdo5vmdi1cm123	ferleck ferles	2013-11-27T21:39:24	Subscribe to my channel ï»¿	1
z122wfnzgt30fhubn04	Bob Kanowski	2013-11-28T12:33:27	i turned it on mute as soon is i came on i just wanted to check the views...ï»¿	0
z13ttt1jcraqexk2o234g	Cony	2013-11-28T16:01:47	You should check my channel for Funny VIDEOS!!ï»¿	1
z12avveb4xqiirsix04ch	BeBe Burkey	2013-11-28T16:30:13	and u should.d check my channel and tell me what I should do next!ï»¿	1
z13auhww3oufjn1qo0	Huckyduck	2013-11-28T17:06:17	Hey subscribe to meï»¿	1
z13xit5agm2zyh4f523r	Lone Twistt	2013-11-28T17:34:55	Once you have started reading do not stop. If you do not subscribe to me wit	1
z13pejoiuozwxtdu323	Archie Lewis	2013-11-28T17:54:39	https://twitter.com/GBphotographyGBï»¿	1
z121zxaxsq25z5k5o04	TheUploadaddict	2013-11-28T18:12:12	subscribe like commentï»¿	1
z12oglnpoq3gjh4om04	Francisco Nora	2013-11-28T19:52:35	please like :D https://premium.easypromosapp.com/voteme/19924/61637535	1
z13phrmwrkfisn5er22	Gaming and Stuff PRO	2013-11-28T21:14:13	Hello! Do you like gaming, art videos, scientific experiments, tutorials, lyrics	1
z13bgdvyluihfv11i22rg	Zielimeek21	2013-11-28T21:49:00	I'm only checking the viewsï»¿	0
z13vxpnoxsyeuv2jr04	OutrightIgnite	2013-11-28T21:55:02	http://www.ebay.com/itm/171183229277?ssPageName=STRK:MESELX:IT&amp	1
z12qth5j0ob1fx3q404	Tony K Frazier	2013-11-28T23:57:13	http://ubuntuone.com/40beUutVu2ZKxK4uTgPZ8Kï»¿	1
z13etj0bclzfztuwc04cg	Jose Renteria	2013-11-29T00:22:01	We are an EDM apparel company dedicated to bringing you music inspired de	1

Şekil 2.1. Verisetinden alınan öznitelik örnekleri

"CONTENT" ve "CLASS" sütunundaki bilgiler ilk önce önişlemden geçirilecek, sonra eğitim için ayrılan bilgiler ile model eğitilecek, en son olarak ise test için ayrılan yorum bilgilerinin spam olup olmadığı bilgisi üretilerek üretilme doğruluğu control edilecektir.

3.Ön İşlem Adımları

İlk olarak beş farklı video için CSV dosyalarındaki veriler birleştirirlir ve “COMMENT_ID, AUTHOR, DATE” sütunlarındaki bilgiler silinerek sadece “CONTENT, CLASS” sütunlarındaki bilgiler tutulur (Şekil 3.1).

	CONTENT	CLASS
0	Huh, anyway check out this you[tube] channel: ...	1
1	Hey guys check out my new channel and our firs...	1
2	just for test I have to say murdev.com	1
3	me shaking my sexy ass on my channel enjoy ^_^...	1
4	watch?v=vtaRGgvGtWQ Check this out .ï»¿	1

Şekil 3.1. “CONTENT” ve “TAG” sütunlarından alınan örnekler

Şekil 2’de görüldüğü gibi “CONTENT” sütunundaki yorum içeriğinde noktalama işaretleri ve büyük-küçük harf karmaşası bulunmaktadır. Bu sorun çözülerek “processed_content” sütununa düzeltilmiş olan yorumlar eklenmiştir (Şekil 3.2).

	CONTENT	CLASS	processed_content
0	Huh, anyway check out this you[tube] channel: ...	1	huh anyway check out this you tube channel kob...
1	Hey guys check out my new channel and our firs...	1	hey guys check out my new channel and our firs...
2	just for test I have to say murdev.com	1	just for test i have to say murdev com
3	me shaking my sexy ass on my channel enjoy ^_^...	1	me shaking my sexy ass on my channel enjoy
4	watch?v=vtaRGgvGtWQ Check this out .ï»¿	1	watch v vtarggvgtwq check this out

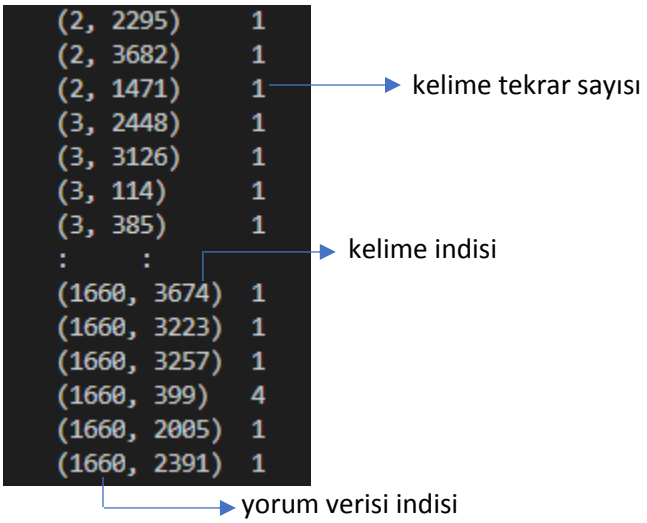
Şekil 3.2. “CONTENT” sütunundaki yorumların sadece alfabe içerecek şekilde ve tüm harflerin küçük harfe dönüştürülmek üzere işlenip “processed_content” sütununa eklenmesi

Sonraki aşama ise “CONTENT” sütununun silinmesi ve düzeldilmiş olan yorumlar sütununun tutulmasıdır (Şekil 3.3).

	CLASS	processed_content
0	1	huh anyway check out this you tube channel kob...
1	1	hey guys check out my new channel and our firs...
2	1	just for test i have to say murdev com
3	1	me shaking my sexy ass on my channel enjoy
4	1	watch v vtarggvgtwq check this out

Şekil 3.3.

Önişlemden geçirilmiş olan veri setinin %85’i eğitim ve %15’i test olarak rastgele ayrılmıştır. Ayrılan eğitim ve test verilerindeki yorumlar kümesine ‘bag of words’ adı verilen kelime torbası modelini uygulanmıştır. Bu model metin işleme konularında sıkça kullanılmaktadır. Çıktı olarak ise metinde hangi kelimeden kaç tane kullanıldığını hesaplayıp bir matrise döker. Şekil 3.4’te oluşturulan matrisin bir kısmı gösterilmiştir.



Şekil 3.4.

Şekil 5'teki çıktıyı daha iyi anlamak için aşağıdaki tablo örnek olarak verilebilir.

	and	back	channel	grow	guys	help	i	me	my	please	plz	subscribe	to	xx
Birinci yorum	1	1	1	0	0	0	1	0	1	0	1	2	1	1
İkinci yorum	0	0	1	1	2	1	0	1	1	1	0	1	1	0

Tablo 3.1.

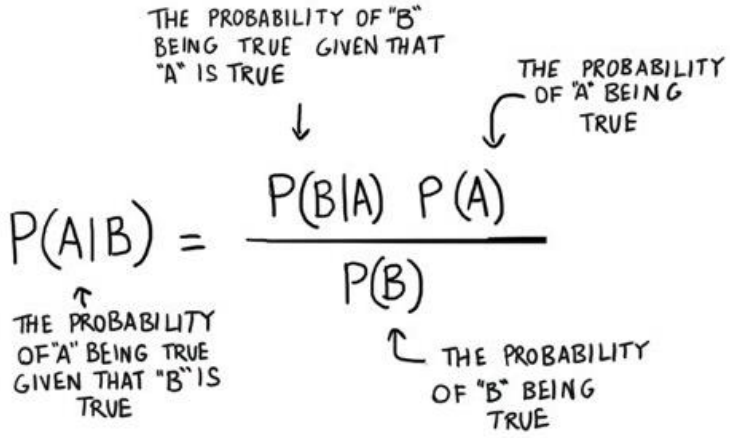
Çok yaygın kelimeleri azaltmak ve nadir olanları vurgulamak için yapılması gereken şey ise her kelimenin ham sayısından ziyade göreceli önemini kaydetmektir. Bu da TF-IDF olarak bilinen metindeki bir kelime veya terimin ne kadar yaygın olduğunu ölçen terim sıklığı (tf) ve ters belge sıklığı (idf); TF x IDF'dır. Sonuç olarak, tüm metinlerde geçen (the, is, an vb.) gibi daha yaygın kelimelerin ağırlığı azaltılmış oldu.

4.Model Oluşturma

Model olarak Multinomial Naive Bayessian, Support Vector Machine(SVM), Logistic Regression ve Random Forest Classifier modelleri kullanılmıştır.

4.1. Multinomial Naive Bayessian

Bayes teoremi bir rassal değişken için olasılık dağılımı içinde koşullu olasılıklar ile marjinal olasılıklar arasındaki ilişkiyi gösterir (Şekil 4.1.1).



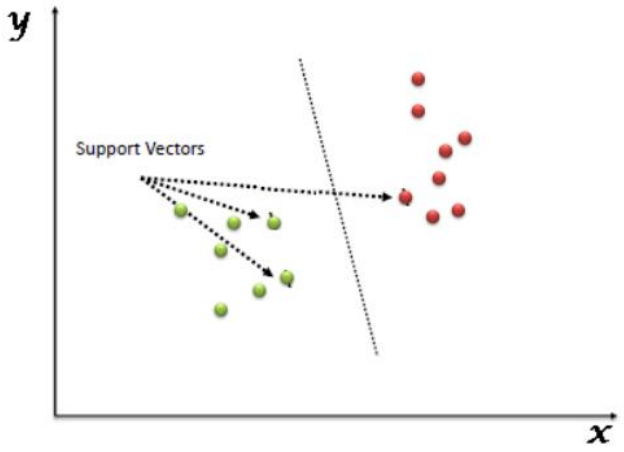
Şekil 4.1.1.

Naive Bayes sınıflandırıcısının temeli Bayes teoremine dayanır. Algoritmanın çalışma şekli bir eleman için her durumun olasılığını hesaplar ve olasılık değeri en yüksek olana göre sınıflandırır. Multinomial Naive Bayes algoritması ise özelliklerin çok terimli bir dağılımdan alındığını varsayar.

4.2. SVM

SVM olarak bilinen Destek Vektör Makinesi, sınıflandırma için kullanılan makine öğrenmesi yöntemlerden birisidir. Temel olarak iki sınıfı bir doğru veya düzlem ile birbirinden ayırmaya çalışır. Bu ayırmayı da sınırdaki elemanlara göre yapar.

Eldeki verilerden yeni kalıpları tanımlamak için büyük miktarda veriyi analiz ederler. SVM'ler, aşağıdaki şekilde gösterildiği gibi bir veri kümesini en iyi iki sınıfa ayıran bir hiper düzlem bulma amacı ile oluşturulur.



Şekil 4.2.1

4.3. Logistic Regression

Logistic Regression yöntemi; sınıflandırma teknikleri, makine öğrenimi ve veri madenciliği uygulamalarının önemli bir parçasıdır ve ikili sınıflandırma problemini çözmek için kullanışlı bir regresyon yöntemidir.

Lojistik regresyondaki bağımlı değişken Bernoulli Dağılımını takip eder. Tahmin ise maksimum olasılıkla yapılır.

4.4. Random Forest Sınıflandırıcısı

Random Forest yöntemi, hem regresyon hem de sınıflandırma problemlerine uygulanabilir olmasından dolayı popüler makine öğrenmesi modellerinden biridir.

Geleneksel yöntemlerden biri olan karar ağaçlarının en büyük problemlerinden biri aşırı öğrenme-veriyi ezberlemedir (overfitting). Random Forest modeli bu problemi çözmek için hem veri setinden hem de öznitelik setinden random olarak 10'larca 100'lerce farklı alt-setler seçiyor ve bunları eğitiyor. Bu yöntemle 100'lerce karar ağacı oluşturuluyor ve her bir karar ağacı bireysel olarak tahminde bulunuyor. Sonuç olarak ise tahminler arasında en çok oy alan seçiliyor.

5.Sonuçlar

Sınıflandırma algoritmalarını kullanarak yapılan çalışmalarda en büyük yanılgılardan biri başarı kriteri olarak sadece doğruluk oranına bakmaktır. Doğruluğa ek olarak iki metriğe daha bakmakta fayda var: recall (duyarlılık) ve precision (kesinlik).

Burada confusion matrixten bahsetmek gerekiyor (Şekil 5.1):

TP (True positive — Doğru Pozitif): Spama spam demek.

FP (False positive — Yanlış Pozitif): Spam olmayana spam demek.

TN (True negative — Doğru Negatif): Spam olmayana spam değil demek.

FN (False negative — Yanlış Negatif): Spam olana spam değil demek.

ACTUAL			
Positive	*Negative*		
True Positive	False Positive	*Positive*	**PREDICTION**
False Negative	True Negative	*Negative*	

Şekil 5.1. Confusion Matrix

Recall (Duyarlılık): Spam olanları doğru tespit etme oranı:

$$recall = \frac{TP}{TP+FN}$$ Denklem (5.1)

Precision (Kesinlik): Spam tespit edilenlerin gerçekten kaçı spam:

$$precision = \frac{TP}{TP+FP}$$ Denklem (5.2)

Recall metriği bazı anomali vakalarını doğru tespit etmek için olumlu bir metriktir. Yani false negative false positiveden daha kritik olduğu durumlar için kullanılır. Doğru olanları yüksek doğrulukla tespit eder fakat yanlış olanların çoğunu yanlış olarak yespit edemez (FP yüksek). Precision metriği ise doğru olanı seçerken iyice düşünüp öyle karar vermeyi sağlıyor fakat burada da FN yüksektir. Görüldüğü üzere bu iki metrik arasında bir trade-off vardır. Bu sorunu çözmek için f1-skoru üretilmiştir. F1-skoru formülü Denklem (5.3)'te verilmiştir.

$$f1 = 2 * \frac{precision * recall}{precision + recall}$$ Denklem (5.2)

Proje kapsamında model başarı kriteri f1-skoru hesabı ile belirlenmiştir. Yöntem sonuçları için oluşturulmuş olan confusion matrisler Şekil 5.2'de verilmiştir.

	precision	recall	f1-score	support
0	0.94	0.88	0.91	138
1	0.90	0.95	0.92	156
accuracy			0.91	294
macro avg	0.92	0.91	0.91	294
weighted avg	0.92	0.91	0.91	294

a

	precision	recall	f1-score	support
0	0.90	0.96	0.93	138
1	0.96	0.90	0.93	156
accuracy			0.93	294
macro avg	0.93	0.93	0.93	294
weighted avg	0.93	0.93	0.93	294

b

	precision	recall	f1-score	support
0	0.87	0.96	0.91	138
1	0.96	0.88	0.92	156
accuracy			0.91	294
macro avg	0.92	0.92	0.91	294
weighted avg	0.92	0.91	0.92	294

c

	precision	recall	f1-score	support
0	0.90	0.99	0.94	138
1	0.99	0.90	0.94	156
accuracy			0.94	294
macro avg	0.94	0.94	0.94	294
weighted avg	0.95	0.94	0.94	294

d

Şekil 5.2 Confusion matrisler: a)Naive Bayessian b)SVM c)Logistic Regression d)Random Forest

Yöntemlerin f1-skor tablosuna bakıldığında Random Forest sınıflandırma modelinin başarımı daha yüksektir (Tablo 5.1).

	Naive Bayessian	SVM	Logistic Regression	Random Forest
F1-skor	0.91	0.93	0.92	0.94

Tablo 5.1