Haziran 2024

BÜYÜK VERİ ANALİTİĞİ FİNAL ÖDEVİ

Doç. Dr. Duygu İÇEN

Ece KANLI
2200329048

1 Büyük Veri ve İstatistik

1.1 İstatistik

İstatistik, veri toplama, analiz etme, anlama ve ilgili belirsizlikleri hesaba katma bilimidir. Bu nedenle, fiziksel, doğal ve sosyal bilimlere; halk sağlığına; tıbba; iş dünyasına ve politikaya nüfuz eder[1].

1.2 Büyük Veri

Büyük veri, hacim, çeşitlilik ve bazı durumlarda toplanma hızı açısından karmaşık olan veri setlerinin toplanması ve analiz edilmesidir. Bazı büyük veriler, belirli bir bilimsel soruyu ele almak için toplanmadığından zorlu olabilmektedir.[1].

Büyük veri problemleri, genellikle çok disiplinli takımları gerektirir. Tipik olarak, konu alanı uzmanlarını, hesaplama uzmanlarını, makine öğrenimi uzmanlarını ve istatistikçileri gerektirir[1].

1.3 İstatistiğin, büyük veri için kilit disiplinlerden biri olması neden önemlidir?

İstatistik, büyük veriden anlamlı ve doğru bilgilerin çıkarılmasını sağlamak için temel öneme sahiptir.

Büyük veride istatistik bilimleri ve alan bilimleri her zamankinden daha iç içe geçmiştir ve istatistiksel metodoloji çıkarım yapmak için kesinlikle önemlidir.

Aşağıda belirtilen durumlar, istatistik için önemlidir ancak büyük veriyle birlikte problemlere neden olabilmektedir:

- Verinin kalitesi ve eksik gözlemler
- Tahminlerin, öngörülerin ve modellerin belirsizliğinin ölçülmesi

Bilimsel istatistik disiplini, bu konulara gelişmiş teknikler ve modeller getirmektedir.

İstatistikçiler, bilimsel sorunun istatistiksel bir soruya dönüştürülmesine yardımcı olurlar, bu da veri yapısının, veriyi oluşturan temel sistemin(modelin) ve değerlendirilmek istenen tahminin (parametre veya parametrelerin) özenli bir şekilde tanımlamasını içerir[1].

1.4 İstatistik ve Büyük Verinin Farklılıkları

Büyük veri ve istatistikler ilgili alanlar olmasına rağmen bazı önemli farklılıklara sahiptir[2]:

- **Veri Boyutu:** Büyük veriler özel olarak, geleneksel veri işleme uygulamaları kullanılarak işlenemeyecek kadar karmaşık ve büyük veri kümelerini içerir. Öte yandan, istatistikler hem küçük hem de büyük veri kümelerinin analizini içerebilir, ancak istatistiksel yöntemler genellikle daha yapılandırılmış ve daha küçük veri kümeleri için tasarlanmıştır.

- **Veri Çeşitliliği:** Büyük veriler genellikle metin, resim, video, sensör verileri, sosyal medya verileri vb. gibi çeşitli veri kaynaklarını içerir. İstatistikler geleneksel olarak sayısal veriler, kategorik veriler vb. gibi daha yapılandırılmış veri türleriyle ilgilenir.

- **Veri Hızı:** Büyük veriler genellikle yüksek hızda üretilir ve gerçek zamanlı işleme gerektirir. İstatistikler genellikle daha kontrollü veri toplama süreçlerini içerir ve her zaman gerçek zamanlı analiz gerektirmez.

- **Amaç:** İstatistikte, odak genellikle çıkarımsal analiz, hipotez testi ve bir örnekleme dayalı tahminler yapmaktır. Büyük veri analizi, genellikle önceden tanımlanmış bir hipotez olmadan, büyük hacimli verilerden içgörü, model ve eğilimleri çıkarmaya daha fazla odaklanmıştır.

- **Araçlar ve Teknikler:** Büyük veri analizi Hadoop, Spark, NoSQL veritabanları ve büyük miktarda veriyi işlemek için özel olarak tasarlanmış makine öğrenme algoritmaları gibi araçların kullanımını içerir. İstatistik, R, Python dillerinde regresyon analizi, Anova vb. gibi geleneksel istatistiksel yöntemler kullanır.

- **Yanlılık:** İstatistikte yanlılığın ele alınması köklü bir uygulamadır, oysa büyük verilerde, yan konusu, veri kaynaklarının çeşitliliği ve hacmi nedeniyle daha karmaşık olabilmektedir.

- **Yorumlama:** İstatistik genellikle değişkenler arasındaki ilişkiyi anlamaya ve belirli bir bağlamda tahminler yapmaya odaklanır. Öte yandan, büyük veri analizi, temel mekanizmaları anlamaksızın veri içindeki kalıpları ve korelasyonları tanımlamaya öncelik verebilir.

Uygulamada, büyük veri ve istatistik arasında önemli bir örtüşme vardır ve birçok istatistiksel yöntem büyük veri analitiğine uyarlanmakta ve uygulanmaktadır. Büyük veri analitiği geliştikçe ve daha fazla istatistiksel teknik içerdikçe, iki alan arasındaki ayrım daha az belirgin hale geliyor.

2 Veri

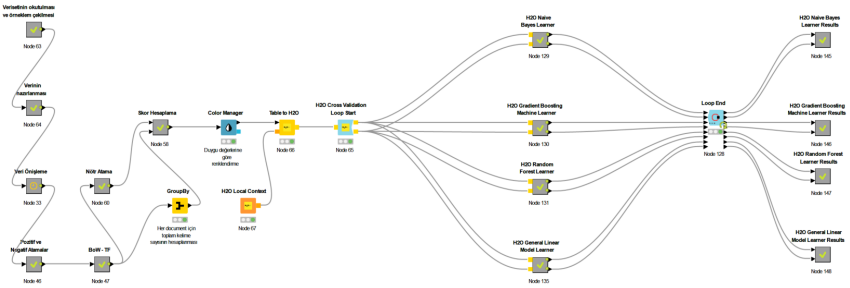
Amazon Kitap Görüşleri

- 1996 Mayıs - 2014 Temmuz tarihleri arasında Amazon'daki 3 milyon kullanıcının 212.404 kitap için yaptıkları 142.8 milyon görüşü içeren bir veri seti kullanılmıştır.
- Veri setine ilişkin değişkenler Tablo 1'de verilmiştir.

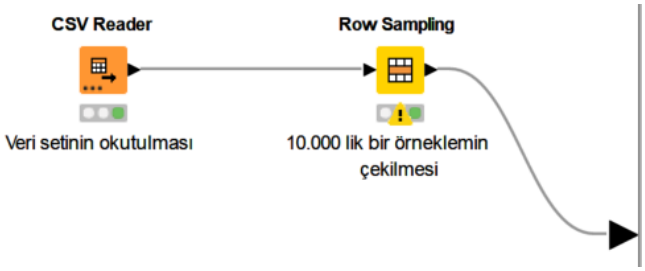
Books rating	
Id (Kitap ID)	review/helpfulness
Title (Kitabın başlığı)	review/score
Price (Kitabın ücreti)	review/time
User_id (Kullanıcı ID)	review/summary (Görüş özeti)
ProfileName (Kullanıcının ismi)	review/text (Görüş)

Tablo 1: Veri Seti Değişkenleri

3 Uygulama

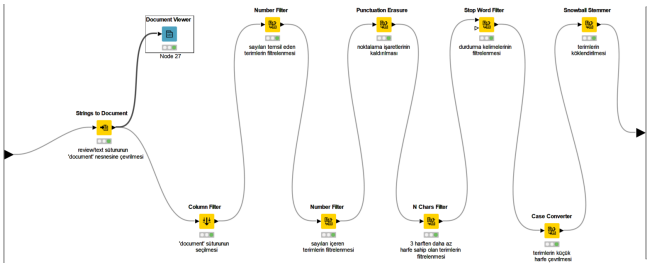


- Kaggle üzerinden [3] alınan veri seti ile sözlük tabanlı duygu ataması gerçekleştirilmiş ve elde edilen duygu etiketlerinin sınıflandırması, H2O AI 5-katlı çapraz geçerlilik kullanılarak makine öğrenimi algoritmaları ile uygulanmıştır.

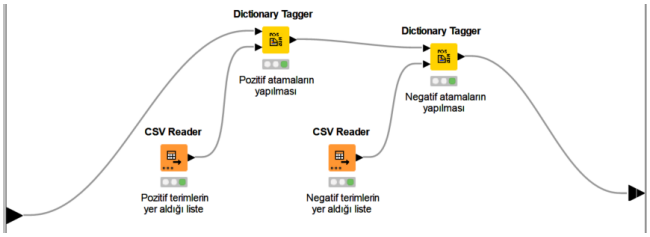


- 3 milyon gözlem ve 10 sütundan oluşan veri setinden sadece **review/text** sütunu ve 10.000 gözlemden oluşan örneklem çekilmiştir.

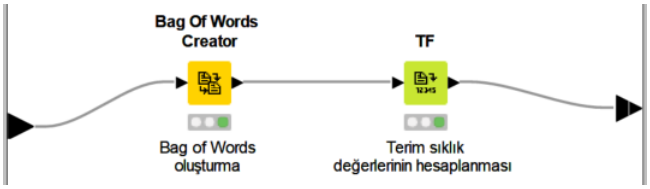
- Veri setinden yinelenen satırlara ve eksik gözleme sahip olan satırlar çıkarılmıştır.



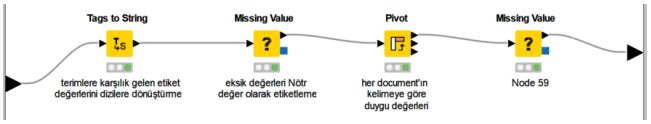
- **review/text** sütunu, **document** nesnesine çevrilerek her bir satırdaki döküman için ön işlemeler yapılmıştır.



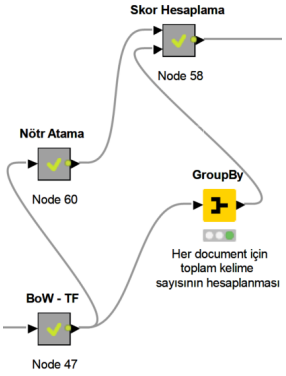
- Her bir dökümana duygu atamasının yapılabilmesi için pozitif ve negatif sözlükler okutularak pozitif ve negatif atamalar yapılmıştır.



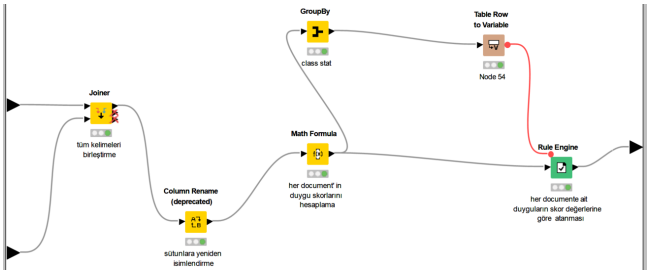
- Her bir döküman için **bag of words** oluşturulmuş ve terim sıklıkları hesaplanmıştır.



- Her bir dökümandaki terimlere karşılık gelen duygu etiketleri dizelere dönüştürülmüş, ardından eksik gözlem değerleri **nötr** duygu etiketi ile doldurulmuştur. Daha sonra her bir dökümandaki toplam pozitif, negatif ve nötr kelime sayıları hesaplanmış, eksik gözlemler **0** değeri ile doldurulmuştur.



- **GroupBy** düğümü ile her dökümandaki toplam kelime sayısı hesaplanmıştır. Ardından elde edilen tablo, duygu skorlarını hesaplamak için **Skor hesaplama** düğümüne bağlanmıştır.



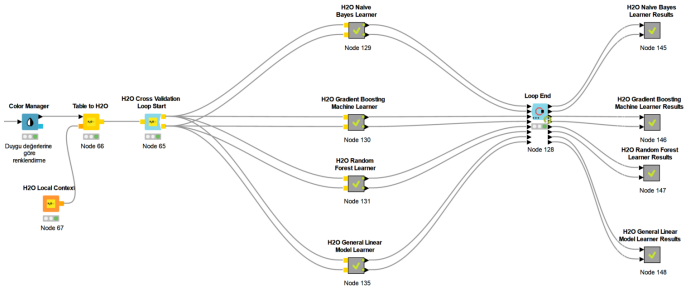
- Pozitif, negatif, nötr ve toplam kelime sayıları birleştirilmiş, her bir duygu etiketi için duygu skorları hesaplanmış ve skorlara göre duygu ataması yapılmıştır.

Table "default" - Rows: 10025 | Spec - Columns: 7 | Properties | Flow Variables

Row ID	Document	Negative Words	Neutral Words	Positive Words	All Words	Sentiment Score	Sentiment
Row0_Row4100	"--dr frankensteinpursuit creation artifici race humancall ra...	60	256	52	368	0.674	NEG
Row1_Row3939	"aaron sisson coal miner amateur flutist midland abandon ...	36	228	60	324	0.778	POS
Row2_Row6983	"abl buy textbook ten dollar includ ship classmat skeptic lo...	8	36	2	46	0.652	NEG
Row3_Row2487	"abl inform book overal found frustratingth author make s...	18	154	30	202	0.822	POS
Row4_Row8522	"abl read bit held attentmayb tri sometim suppos classic tri"	0	20	2	22	1	POS
Row5_Row7579	"abl understand vivid descript symbol book hawthorn write...	2	24	8	34	0.882	POS
Row6_Row3204	"abli translat annot jonathan galassi revis bilingu edit euge...	32	138	16	186	0.656	NEG
Row7_Row6183	"abort clinic twenti ann bakerrequir read counselor staff ad...	6	56	16	78	0.846	POS
Row8_Row8787	"abridg version book read sin inevit bad situat women stro...	24	52	8	84	0.429	NEG
Row9_Row1615	"absolut amaz book extrem written author creat round cha...	0	34	4	38	1	POS
Row10_Row4...	"absolut amaz fun excit emot drama love erot romant passi...	6	18	6	30	0.6	NEG
Row11_Row9...	"absolut captiv novel throughout novel major event occur ...	18	80	12	110	0.673	NEG
Row12_Row8...	"absolut found kerouacknowmix bit buddha wild prose andi...	22	84	20	126	0.651	NEG
Row13_Row9...	"absolut hook chapter devour book fastest finish book ten...	12	56	14	82	0.707	POS
Row14_Row625	"absolut html book book time job web develop main job lay...	6	80	6	92	0.87	POS
Row15_Row2...	"absolut love book ago book talk person growth improv yo...	16	136	30	182	0.824	POS

Şekil 1: Duygu Etiketleri

- Duygu etiketlerinin elde edildiği son tablodan bir kısım Şekil 1'de verilmiştir.



- Elde edilen duygu etiketlerine göre **Color Manager** düğümü ile duygu etiketlerine renklendirme yapılmıştır.
- Sınıflandırma algoritmalarının uygulanması için önce elde edilen son tablo H2O çerçevesine dönüştürülmüştür. Ardından 5 katlı çapraz geçerlilik döngüsü başlatılmıştır.
- H2O Naive Bayes, H2O Gradient Boosting MAchine, H2O Random Forest, H2O General Linear Model olmak üzere 4 sınıflandırma algoritması uygulanmıştır.
- Son olarak uygulanan her bir algoritmanın eğitim ve test kümesi için doğruluk skorları hesaplanarak değerlendirilmiştir.

3.1 Makine Öğrenmesi Algoritmaları

3.1.1 H2O Naive Bayes Learner

	Doğruluk oranı (%)	
İterasyon	**Eğitim Kümesi**	**Test Kümesi**
0	0.830	0.833
1	0.828	0.820
2	0.835	0.843
3	0.827	0.823
4	0.831	0.834

Tablo 2: H2O Naive Bayes Doğruluk Oranları

- Tablo 2'de yer alan sonuca göre eğitim ve test kümesinin doğruluk oranlarının dengede olduğu ve %83 civarında olduğu söylenebilir. İki küme birlikte değerlendirildiğinde algoritma, en iyi sınıflandırmayı 4. iterasyonda yapmıştır.

		Tahmin		
		Negatif	**Pozitif**	**Nötr**
Gerçek	**Negatif**	2793	939	82
	Pozitif	305	3818	20
	Nötr	0	0	7

Tablo 3: NB Eğitim Kümesi

		Tahmin		
		Negatif	**Pozitif**	**Nötr**
Gerçek	**Negatif**	749	241	18
	Pozitif	76	968	8
	Nötr	0	0	1

Tablo 4: NB Test Kümesi

- Tablo 3 ve Tablo 4'te 4. iterasyona ait confusion matrisleri verilmiştir. Hem eğitim hem de test kümesinde en fazla yanlış sınıflandırılan duygu etiketi negatiftir. Nötr duygu etiketlerinde ise yanlış sınıflandırma olmamıştır.

3.1.2 H2O Gradient Boosting Machine Learner

	Doğruluk oranı (%)	
İterasyon	**Eğitim Kümesi**	**Test Kümesi**
0	1.00	1.00
1	1.00	1.00
2	1.00	1.00
3	1.00	1.00
4	1.00	1.00

Tablo 5: H2O Gradient Boosting Machine Doğruluk Oranları

- Tablo 5'te yer alan sonuca göre yapılan sınıflandırmada, eğitim ve test kümesinin her iterasyonda %100 doğru sınıflandırma yaptığı söylenebilir.

		Tahmin		
		Negatif	**Pozitif**	**Nötr**
Gerçek	**Negatif**	3814	0	0
	Pozitif	0	4143	0
	Nötr	0	0	7

Tablo 6: GBM Eğitim K.

		Tahmin		
		Negatif	**Pozitif**	**Nötr**
Gerçek	**Negatif**	969	0	0
	Pozitif	0	1062	0
	Nötr	0	0	2

Tablo 7: GBM Test Kümesi

- Tablo 6 ve Tablo 7'de rasgele seçilen bir iterasyonun confusion matrisleri verilmiştir. Tüm duygu etiketleri eğitim ve test kümesinde doğru sınıflandırılmıştır.

3.1.3 H2O Random Forest Learner

	Doğruluk oranı (%)	
İterasyon	**Eğitim Kümesi**	**Test Kümesi**
0	1.00	1.00
1	1.00	1.00
2	1.00	1.00
3	1.00	1.00
4	1.00	1.00

Tablo 8: H2O Random Forest Doğruluk Oranları

- Tablo 8'de yer alan sonuca göre yapılan sınıflandırmada, eğitim ve test kümesinin her iterasyonda %100 doğru sınıflandırma yaptığı söylenebilir.

		Tahmin		
		Negatif	**Pozitif**	**Nötr**
Gerçek	**Negatif**	3814	0	0
	Pozitif	0	4143	0
	Nötr	0	0	7

Tablo 9: RF Eğitim Kümesi

		Tahmin		
		Negatif	**Pozitif**	**Nötr**
Gerçek	**Negatif**	1008	0	0
	Pozitif	0	1052	0
	Nötr	0	0	1

Tablo 10: RF Test Kümesi

- Tablo 9 ve Tablo 10'da rasgele seçilen bir iterasyonun confusion matrisleri verilmiştir. Tüm duygu etiketleri eğitim ve test kümesinde doğru sınıflandırılmıştır.

3.1.4 H2O General Linear Model Learner

	Doğruluk oranı (%)	
İterasyon	**Eğitim Kümesi**	**Test Kümesi**
0	0.994	0.996
1	0.992	0.989
2	0.996	0.996
3	0.994	0.994
4	0.997	0.998

Tablo 11: H2O General Linear Model Doğruluk Oranları

- Tablo 11'de yer alan sonuca göre eğitim ve test kümesinin doğruluk oranlarının dengede olduğu ve %99 civarında olduğu söylenebilir. İki küme birlikte değerlendirildiğinde algoritma en iyi sınıflandırmayı 4. iterasyonda yapmıştır.

		Tahmin		
		Negatif	**Pozitif**	**Nötr**
Gerçek	**Negatif**	3805	9	0
	Pozitif	5	4138	0
	Nötr	7	0	0

Tablo 12: GLM Eğitim K.

		Tahmin		
		Negatif	**Pozitif**	**Nötr**
Gerçek	**Negatif**	1005	3	0
	Pozitif	0	1052	0
	Nötr	1	0	0

Tablo 13: GLM Test Kümesi

- Tablo 12 ve Tablo 13'te 4. iterasyona ait confusion matrisleri verilmiştir. Test kümesinde, pozitif duygu etiketinde yanlış sınıflandırma olmadığını gözlemlerken; nötr duygu etiketinin hem eğitim hem de test kümesinde hiç doğru sınıflandırılamadığı ve negatif duygu etiketi olarak sınıflandırıldığı söylenebilir.

4 Sonuç ve Tartışma

- Bu çalışmada, Amazon kitap görüşleri veri seti üzerinde sözlük tabanlı duygu analizi gerçekleştirilmiş ve ardından çeşitli makine öğrenimi algoritmaları ile sınıflandırma işlemleri uygulanmıştır.

- Sözlük tabanlı yaklaşım ile her bir döküman için pozitif, negatif ve nötr kelime sayıları hesaplanmış ve duygu etiketleri belirlenmiştir.

- Elde edilen bu etiketler, H2O AI ile 5 katlı çapraz geçerlilik kullanılarak dört farklı makine öğrenimi algoritması **Naive Bayes, Gradient Boosting Machine, Random Forest, General Linear Model** ile sınıflandırılmıştır.

- Naive Bayes algoritması, eğitim ve test kümelerinde yaklaşık %83 doğruluk oranları ile bir sınıflandırma yapmıştır. Negatif duygu etiketlerinde yanlış sınıflandırma fazla iken nötr duygu etiketlerinde yüksek doğruluk oranı dikkat çekmektedir.

- Gradient Boosting Machine algoritması, hem eğitim hem de test kümelerinde %100 doğruluk oranı ile en yüksek performansı gösteren algoritmadan biridir. Tüm iterasyonlarda duygu etiketlerini doğru sınıflandırmıştır.

- Random Forest algoritması da GBM gibi, %100 doğruluk oranı ile mükemmel bir performans sergileyen diğer algoritmadır. Tüm duygu etiketleri, eğitim ve test kümelerinde doğru sınıflandırılmıştır. Bu durum, hem GBM'in hem de Random Forest algoritmasının iyi bir şekilde öğrendiğini ve duygu etiketlerini etkili bir şekilde ayırt edebildiğini göstermektedir.

- General Linear Model (GLM) algoritması, %99 doğruluk oranları ile yine yüksek bir performans göstermiştir. Ancak, nötr duygu etiketlerinin doğru sınıflandırılamadığı ve negatif duygu etiketi olarak etiketlendiği gözlemlenmiştir. Bu durum, GLM algoritmasının nötr etiketlerini diğer duygu etiketlerinden ayırmada zorlandığını göstermektedir.

- Sonuç olarak, bu çalışmada kullanılan makine öğrenimi algoritmaları genel olarak yüksek doğruluk oranları ile başarılı bir performans sergilemiştir. Özellikle Gradient Boosting Machine ve Random Forest algoritmaları, %100 doğruluk oranı ile mükemmel performans göstermiştir. Naive Bayes algoritması ise %83 doğruluk oranı ile diğer algoritmaların gerisinde kalmıştır. General Linear Model algoritması, %99 doğruluk oranı ile yüksek performans göstermiş ancak nötr duygu etiketlerinde zayıf kalmıştır.

- Sözlük tabanlı duygu analizi ve H2O AI kullanılarak gerçekleştirilen bu çalışma, büyük veri setleri üzerinde makine öğrenimi algoritmalarının etkin bir şekilde kullanılabileceğini göstermiştir. Farklı algoritmaların performans karşılaştırması, duygu analizi ve sınıflandırma problemlerinde hangi algoritmaların daha etkili olabileceğini göstermesi açısından önemlidir.

Kaynaklar

[1] Amazon Web Services, "Statistics and big data." https://higherlogicdownload.s3.amazonaws.com/AMSTAT/UploadedImages/49ecf7cf-cb26-4c1b-8380-3dea3b7d8a9d/BigDataOnePager.pdf [Accessed: 25.05.2024].

[2] Quora, "How does big data differ from statistics?." https://www.quora.com/How-does-big-data-differ-from-statistics [Accessed: 25.05.2024].

[3] Kaggle, "Amazon books reviews." https://www.kaggle.com/datasets/mohamedbakhet/amazon-books-reviews?resource=download&select=Books_rating.csv [Accessed: 09.05.2024].

[4] Knime Community Forum, "Sentiment score formula for sentimental analysis with neutral sentiments." https://forum.knime.com/t/sentiment-score-formula-for-sentimental-analysis-with-neutral-sentiments/15575 [Accessed: 12.05.2024].

[5] Knime Blog, "Lexicon based sentiment analysis: What it is how to conduct one." https://www.knime.com/blog/lexicon-based-sentiment-analysis [Accessed: 12.05.2024].

[6] Knime Community Hub, "Lexicon based approach for sentiment analysis." https://hub.knime.com/knime/spaces/Examples/08_Other_Analytics_Types/01_Text_Processing/26_Sentiment_Analysis_Lexicon_Based_Approach~zp_hhUROHNXToZHX/current-state [Accessed: 12.05.2024].

[7] Knime Community Hub, "Sentiment classification." https://hub.knime.com/knime/spaces/Examples/08_Other_Analytics_Types/01_Text_Processing/03_Sentiment_Classification~ZHAExldZ5M7q6hdG/current-state [Accessed: 12.05.2024].

[8] Knime Community Hub, "H2o cross validation." https://hub.knime.com/knime/spaces/Examples/04_Analytics/15_H2O_Machine_Learning/04_H2O_Crossvalidation~zzTdxkw7hfziD4Ik/current-state [Accessed: 25.05.2024].