OKU Fen Bilimleri Enstitüsü Dergisi
Cilt 4, Sayı 3, 242-249, 2021

OKU Journal of the Institute of Science and Technology
Volume 4, Issue 3, 242-249, 2021

Osmaniye Korkut Ata Üniversitesi Fen Bilimleri Enstitüsü Dergisi

Osmaniye Korkut Ata University Journal of the Institute of Science and Technology

Apache Spark Tabanlı Duygu Analizi

Emre YILDIRIM[1*], Ali ÇALHAN[2]

[1]Osmaniye Korkut Ata Üniversitesi Osmaniye Meslek Yüksekokulu Bilgisayar Teknolojileri Bölümü, 80000 Osmaniye
[2]Düzce Üniversitesi Mühendislik Fakültesi Bilgisayar Mühendisliği Bölümü, 81000 Düzce

[1]https://orcid.org/0000-0002-9072-9780
[2]https://orcid.org/0000-0002-5798-3103
*Sorumlu yazar: emreyildirim@osmaniye.edu.tr

Araştırma Makalesi

Makale Tarihçesi:
Geliş tarihi: 27.04.2021
Kabul tarihi:02.08.2021
Online Yayınlanma: 15.12.2021

Anahtar Kelimeler:
Apache spark
Duygu analizi
Makine öğrenmesi

ÖZET

Bu çalışmada, büyük verileri bellek içi hesaplama yöntemi ile hızlı bir şekilde işleyebilen Apache Spark açık kaynak kodlu çerçeve kullanılarak duygu analizi gerçekleştirilmiştir. Duygu analizi işleminde Spark içerisinde bulunan MLlib makine öğrenimi kütüphanesi kullanılmıştır. Lojistik regresyon (LR), destek vektör makinesi (SVM) ve Naive Bayes (NB) makine öğrenmesi sınıflandırma algoritmaları kullanılmıştır. Çalışmada, duygu analizinde kullanılan makine öğrenmesi algoritmaları doğruluk, kesinlik ve duyarlılık performanslarına göre değerlendirilmektedir. Sonuçlar, SVM algoritmasının çalışmada kullanılan iki farklı veri setinde de sırasıyla %91, %88 doğruluk, %91, %90 kesinlik ve %91, %87 duyarlılık değerleri ile en iyi performansa sahip olduğunu göstermektedir.

Apache Spark Based Sentiment Analysis

Research Article

Article History:
Received: 27.04.2021
Accepted: 02.08.2021
Published online: 15.12.2021

Keywords:
Apache spark
Sentiment analysis
Machine learning

ABSTRACT

In this study, sentiment analysis is carried out using the Apache Spark open source framework, which is capable of processing big data quickly with the method of computing in memory. MLlib machine learning library in Spark is used in the sentiment analysis process. Logistic regression (LR), support vector machine (SVM) and Naive Bayes (NB) machine learning classification algorithms were used. In the study, machine learning algorithms used in sentiment analysis are evaluated according to their accuracy, precision and sensitivity performances. The results show that the SVM algorithm has the best performance in the two different data sets used in the study, with 91%, 88% accuracy, 91%, 90% precision and 91%, 87% sensitivity, respectively.

To Cite: Yıldırım E., Çalhan A. Apache Spark Tabanlı Duygu Analizi. Osmaniye Korkut Ata Üniversitesi Fen Bilimleri Enstitüsü Dergisi 2021; 4(3): 242-249.

Giriş

Çoğu insan günümüzde farklı ortamları kullanarak anılarını, deneyimlerini, fikirlerini ve duygularını paylaşmaktadır. Bu ortamlarda, kullanıcılar çeşitli cihazlar kullanarak mesajlar gönderir ve bu mesajlar sağlık, ekonomi, spor gibi içeriklere sahiptir. Bu mesajlar, kısa bir süre içerisinde çok büyük verilere sebep olmaktadır. Bu nedenle, farklı ortamlardan atılan mesajların içeriğindeki duyguları ve fikirleri analiz edebilmek amacıyla günümüzde araştırmacılar popüler bir alan olan duygu analizi yaklaşımını kullanmaktadır. Duygu analizi, doğal dil işlemenin (NLP) önemli bir alt dalı olarak tanımlanabilir. Amacı, insanların belirli varlıklar hakkındaki görüşlerini ortaya çıkarmaktır. Duygu analizi birçok uygulamada kullanılmaktadır. Yapılan yorumlar sonucunda bir otelle ilgili olumlu ya da olumsuz durumları analiz

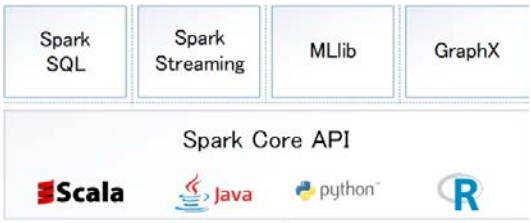
edilebilmekte, bir filmin içeriği hakkında analizler yapılabilmek, sosyal medya üzerinden tanıtılan ürünlerin kullanıcı yorumları ile müşteri memnuniyetinin belirlenmesi ve haber sitesinde yayınlanan olaylarla ilgili toplumun olaylara ait tutumlarının belirlenmesi bunlardan bazılarıdır. Bununla birlikte, verinin gün geçtikçe artması büyük veri işleme teknolojilerini de geliştirmektedir. Hadoop, Apache Spark, Apache Storm gibi büyük veri çerçeveleri ve HDFS, HBase, MongoDB gibi dağıtık veri tabanları, çok büyük miktarda verinin işlenmesini neredeyse zahmetsiz hale getirecek şekilde tasarlandıkları için popüler hale gelmektedir. Bu tür sistemler gün geçtikçe gelişmekte ve makine öğrenimi tekniklerinin kullanılmasını mümkün kılmaktadır.
Bu çalışmada, Apache Spark tabanlı duygu analizi yapılmıştır. Ayrıca, analiz için Apache Spark MLlib kütüphanesinden LR, SVM ve NB sınıflandırma algoritmaları kullanılmaktadır.

İlgili Araştırmalar

Geçmiş yıllarda duygu analizi ve duygusal modeller üzerine yapılan çalışmalar yoğun ilgi görmüştür. Bunun nedeni, son zamanlarda insanlar, bakış açılarını, düşüncelerini ve yorumlarını farklı ortamları kullanarak paylaşmasıdır. Mohapatra ve ark. (2018), Twitter verilerinden çıkarılan duygulara göre, gerçek zamanlı yeni bir kripto para birimi fiyat tahmin platformu olan KryptoOracle'ı tanıtmışlardır. Kouloumpis ve ark. (2011) Twitter mesajlarının duyarlılığını analiz etmek için dil özelliklerinin faydasını araştırmışlardır. Denetimli öğrenme algoritmalarının kullanıldığı çalışmada, Twitter hashtag'lerine göre alınan tweetler pozitif, negatif ve nötr olarak sınıflandırmışlardır. Pang ve ark. (2002), film yorumlarından alınan verilerin duyarlılığını analiz etmek için naive bayes, maksimum entropi ve destek vektör makinesi sınıflandırıcılarını kullanmışlardır. Wang ve ark. (2012), ABD başkanlık seçimi oylamasına ilişkin gerçek zamanlı bir Twitter duygu analizi sistemi oluşturmuş ve 17000 tweeti eğitim veri setinde kullanmışlardır. Neethu ve Rajasree (2013) cep telefonları, dizüstü bilgisayarlar vb. elektronik ürünler hakkındaki twitter mesajlarına göre duygu analizi gerçekleştirmişlerdir. Makine öğrenimi algoritmalarının kullanıldığı çalışmada, insanların ürünler hakkındaki fikirlerini olumlu ve olumsuz olarak sınıflandırmışlardır.
Yapılan çalışmalar incelendiğinde Apache Spark veri işleme teknolojisinin duygu analizi çalışmalarında kullanılmadığı görülmektedir. Bu bağlamda çalışmamızda Apache Spark ve ona ait makine öğrenmesi algoritmaları kullanılarak bir duygu analizi yapılmıştır. Böylece, literatüre katkı sağlanması amaçlanmaktadır.

Apache Spark

Apache Spark, hem toplu hem de gerçek zamanlı olarak veri işlemeyi desteklediği için hibrit bir analiz motorudur (Zaharia ve ark., 2012). Spark, Hadoop yazılımı Map-Reduce yapısındaki birçok ilkeyi kullansa da, Spark bellek içi hesaplama özelliği ile performans açısından Hadoop'a göre daha iyi performans göstermektedir. Spark, bağımsız olarak da çalışabilirken, ayrıca Hadoop ile de entegre edilerek kullanılabilmektedir.
Spark, HDFS, NoSQL veritabanları ve SQL benzeri veri depoları dahil olmak üzere farklı veri kaynaklarından gelen verileri işleyebilmektedir. Spark, Scala dilinde yazılmıştır, ancak Java, Python ve R dillerini de desteklemektedir.

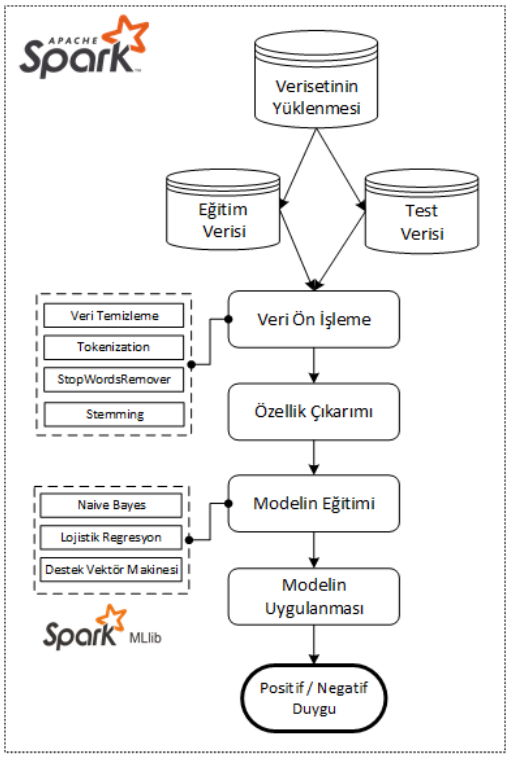


Şekil 1. Apache spark ekosistemi

Apache Spark ekosisteminde analiz işlemlerine yardımcı olacak farklı özellikte kütüphaneler bulunmaktadır. Bu ekosistem Şekil 1'de de gösterildiği gibi Spark SQL, Spark Streaming, MLlib ve Spark GraphX kütüphanelerinden oluşmaktadır. Bu çalışmada, Spark'ın dağıtık olarak çalışabilen makine öğrenimi kütüphanesi olan MLlib kütüphanesi kullanılmıştır. MLlib kütüphanesinde birçok sınıflandırma, kümeleme ve regresyon algoritmaları bulunmaktadır (Meng ve ark., 2016). Çalışmamızda ise Logistik Regresyon, Destek Vektör Makinesi ve Naive Bayes sınıflandırma algoritmaları kullanılmıştır.

Apache Spark Tabanlı Duygu Analizi

Duygu analizi için Apache Spark tabanlı bir analiz sistemi geliştirilmiştir. Sistemin uygulaması Apache Spark 2.4.6 versiyonu temel alınarak geliştirilmiştir. Python için Spark programlama modelini destekleyen PySpark kullanılmıştır. Sistem, Şekil 2'de gösterildiği gibi dört ana bileşene sahiptir. Her bileşenin açıklaması alt başlıklarda detaylı bir şekilde açıklanmıştır.



Şekil 2. Apache Spark Tabanlı Duygu Analizi Sistemi.

Veri Seti

Duygu analizinde iki farklı veri seti kullanılmıştır. Bu veri setleri, Kaggle veri seti merkezinde bulunan Yelp mekan keşfetme uygulama yorumları (Kaggle, 2021a) ve IMDB film incelemeleridir (Kaggle, 2021b)

Çalışmamızda kullanılan veri setlerinde pozitif ve negatif sınıflara ait metinler bulunmaktadır. Yelp ve IMDB veri setlerinin detaylı sayısal bilgileri Tablo1'de sunulmuştur.

Tablo 1. Yelp ve IMDB veri seti detayları

Veri Seti	Eğitim Verisi	Test Verisi	Toplam
Yelp	560000	38000	598000
IMDB	45000	5000	50000

Veri Önişleme

Veri seti mesaj metinlerinden oluşmaktadır. Bu nedenle analiz işleminin en doğru sonucu verebilmesi için bir takım ön işleme aşamalarından geçmesi gerekmektedir. Çalışmamızda yapılan veri ön işleme aşamaları aşağıdaki gibidir:

- *Veri Temizleme*: Tüm metindeki internet adreslerini, hashtag'leri, noktalama işaretleri gibi gereksiz karakterleri metinden temizleme işlemi bu aşamada gerçekleşmektedir. Ayrıca, metindeki büyük harfler küçük harfe dönüştürülmektedir.
- *Jetonlaştırma (Tokenization)*: Metnin çeşitli karakterler (boşluk, virgül vb.) temel alınarak kelimelere (jetonlara) ayrılması anlamına gelir. Metin boşluklara göre ayrılarak kelime dizileri haline getirilmektedir.
- *Durak Kelimelerinin Çıkarılması (StopWordsRemover)*: Metin içerisinde yaygın olarak kullanılan a, an, the, has, have vb. gibi anlam taşımayan, yani analiz sırasında metnin duyarlılığını belirlemede yardımcı olmayan kelimeler çıkartılmaktadır.
- *Kök Bulma (Stemming):* Türetilmiş sözcükleri dile bağlı bir türetme algoritması kullanarak temel biçimlerine indirger. Dolayısıyla, farklı biçimsel durumlarda bulunan ancak kök hali aynı olan kelimeler köke indirgenmektedir.

Özellik Çıkarımı

Duygu analizi gibi metin analizlerinin zorluklarından biri de büyük boyutlu verilerden makine öğrenimi için özellik çıkarımıdır. Metni bir özellik matrisine dönüştürmek için bazı özellik çıkarma yöntemlerini kullanmak en iyisidir. Bu nedenle hem çalışmamıza uygunluğu hem de geçerliliği yüksek bir özellik çıkarma yöntemlerinden biri olan Terim Frekansı – Ters Doküman Frekansı (TF-IDF) uygulanmıştır.

TF-IDF, işlenen bir metnin içindeki kelimelerin önem düzeyini değerlendirmede kullanılan popüler bir yöntemdir. TF-IDF'in amacı, metin içindeki kelime sıklığını hesaplamaktır.

Makine Öğrenmesi Sınıflandırma Algoritmaları

Bu çalışmada, duygu analizi için denetimli öğrenme makine öğrenimi sınıflandırma algoritmaları kullanılmıştır. Bunun için Apache Spark'ın MLlib kütüphanesinden LR, SVM ve NB sınıflandırma algoritmaları kullanılmıştır. Uygulanan sınıflandırıcılar bu bölümde detaylı bir şekilde açıklanmaktadır.

Naive Bayes (NB): NB, problemde her örneğin bir özellik vektörü olarak sunulduğu ve her bir özellik değerinin diğer herhangi bir özelliğin değerinden ayrı ve bağımsız olarak varsayıldığı, Bayes teoremine dayanan bir sınıflandırma tekniğidir (Goel ve ark., 2016). NB sınıflandırıcı, belgedeki kelimelerin dağılımına bağlı olarak bir sınıfın son olasılığını hesapladığı için metin sınıflandırmasında tercih edilmektedir (Jain ve Dandannavar, 2016).

Lojistik Regresyon (LR): LR, bir regresyon analiz modelidir. Çoğunlukla bağımlı değişkenin tutarlı sayıdaki değerlerden birini alabilen Binary olduğu durumlarda kullanılır. Hedef sınıf ile girdiden çıkarılan özellikler arasındaki ilişkiyi değerlendirir ve açıklar. Sadece ikili sınıflandırma için kullanılmaz, aynı zamanda çok sınıflı sınıflandırma problemleri için de kullanılabilir (Hosmer ve ark., 2013).

Destek Vektör Makinesi (SVM): SVM'nin amacı, arama alanında farklı sınıfları en iyi şekilde ayırabilen doğrusal ayırıcıları veya hiper düzlemi bulmaktır. Sınıfları ayıran birkaç hiper düzlem olabilir, ancak bu düzlemlerden en uygun olanı seçilen veri noktalarından herhangi birinin normal mesafesinin en büyük olduğu, yani maksimum ayırma mesafesini gösterdiği hiper düzlemdir (Zhu ve Blumberg, 2002).

Performans Değerlendirmesi

Duygu analizi amacıyla geliştirilen tahmin modelinde üç farklı sınıflandırma algoritması kullanılmıştır. Bu algoritmaların aşırı uyum sorunundan kaçınarak performans değerlendirmesinin yapılabilmesi için öncelikle k-kat çapraz doğrulama kullanılmalıdır. Çalışmada, model 5 kat çapraz doğrulama ile model oluşturulmaktadır. Sınıflandırma algoritması önce veri setinin bir bölümü ile eğitilmekte ve tahmine dayalı bir model oluşturulmaktadır. Eğitilen model daha sonra doğruluk ölçütü değerini belirlemek için kalan örneklerle test edilmektedir. Bir modelin tahmin doğruluğu aşağıdaki denklem ile hesaplanır:

$$\text{Doğruluk} = \frac{TP + TN}{TP + TN + FP + FN} \qquad (1)$$

Doğruluk hesaplamasında bahsedilen ilişkide TP, TN, FP ve FN, sırasıyla True Positive, True Negative, False Positive ve False Negative'dir. Çalışmada, TP doğru tahmin edilen pozitif duyguları, TN doğru tahmin edilen negatif duyguları, FP yanlış tahmin edilen pozitif duyguları, FN ise yanlış tahmin edilen negatif duyguları belirtmektedir.

Çalışmamızda, modelin tahmin doğruluğuna ek olarak, kesinlik ve duyarlılık değerleri de hesaplanmaktadır. Kesinlik ve duyarlılık değerleri aşağıdaki denklem ile hesaplanmaktadır.

$$\text{Kesinlik} = \frac{TP}{TP + FP} \qquad (2)$$

$$\text{Duyarlılık} = \frac{TP}{TP + FN} \qquad (3)$$

Bulgular ve Tartışma

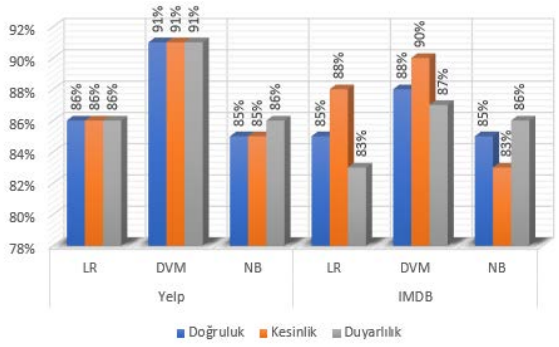
Bu çalışmada, Apache Spark tabanlı bir duygu analizi sistemi oluşturulmuştur. Sistemde analiz için Spark MLlib kütüphanesinden LR, SVM ve NB sınıflandırma algoritmaları kullanılmıştır. Duygu analizinde Yelp

ve IMBD film yorumları veri setleri ayrı ayrı kullanılarak sınıflandırma algoritmalarının performansları incelenmiştir. Performans analizinde kullanılan iki veri seti veri ön işleme aşamasından geçirilerek analiz edilmiştir. Analiz sonucunda her iki veri seti için farklı performans değerleri ortaya çıkmıştır. Algoritmaların tahminlerine göre oluşan karmaşıklık matrisi Tablo 2'de verilmiştir.

Tablo 2. Duygu analizi sonucundaki tahmin değerleri

		TP	FP	FN	TN
Yelp	LR	16360	2640	2573	16427
	DVM	17298	1702	1688	17312
	NB	16111	2889	2626	16374
IMDB	LR	2214	309	446	2091
	DVM	2264	259	332	2205
	NB	2094	429	350	2187

Tablo 2'de verilen tahmin değerlerine göre sınıflandırma algoritmalarının doğruluk ve f-ölçüm değerleri hesaplanmış ve bu değerlere göre performansları karşılaştırılmıştır. Hesaplamalar sonucunda, sınıflandırma algoritmalarının performans değerleri Şekil 3'te yüzdelik olarak gösterilmektedir.



Şekil 3. Sınıflandırma Algoritmalarının Performansları.

Şekil 3 incelendiğinde, duygu analizinde kullanılan sınıflandırma algoritmalarının veri setlerine göre farklı performans değerleri ortaya çıkmaktadır. Bu bağlamda sonuçlar, SVM'nin her iki veri setinde üç farklı kritere göre en iyi performansa sahip olduğu görülmektedir. NB ise diğer algoritmalara göre daha düşük bir performansa sahiptir.

Sonuçlar

Duygu analizi, ağırlıklı olarak makine öğrenimi algoritmalarının kullanımıyla çözülebilen bir sınıflandırma problemidir. Farklı ortamlarda paylaşılan mesajların içerisindeki duyguları çıkarmak için kullanılan çok sayıda yöntem vardır. Bu çalışmada, pozitif ve negatif duyguların oluşturduğu iki farklı veri setinin kullanıldığı Apache Spark tabanlı bir duygu analizi sistemi sunulmuştur. Bu sistem doğal dil işleme tekniklerinin yanı sıra, LR, SVM ve NB sınıflandırma algoritmalarını kullanarak veri setleri içerisindeki duyguları analiz etmektedir. Tüm sınıflandırma algoritmaları Apache Spark'ın MLlib kütüphanesi sisteme

entegre edilerek sistem içerisinde kullanılmaktadır. Bu algoritmalar, farklı sınıfların olduğu veri setleri kullanılarak eğitilir ve test edilir. Bu çalışmada, MLlib kütüphanesine ait sınıflandırma algoritmaları doğruluk, kesinlik ve duyarlılık ölçüm kriterleri üzerinden performansları açısından karşılaştırılmaktadır. Sonuçlar, SVM algoritmasının iki farklı veri setinde de sırasıyla %91, %88 doğruluk, %91, %90 kesinlik ve %91, %87 duyarlılık değerleri ile en iyi performansa sahip olduğunu göstermektedir.

Gelecek çalışmalarda, popüler bir veri kaynağı olan Twitter sosyal medya ortamından alınan gerçek zamanlı verilerin kullanıldığı duygu analizi sistemi oluşturulması planlanmaktadır. Bunun için Apache Spark'ın gerçek zamanlı veri akışlarını işlemek için kullanıldığı Spark Streaming kütüphanesinden yararlanılması planlanmaktadır.

Çıkar Çatışması Beyanı

Makale yazarları aralarında herhangi bir çıkar çatışması olmadığını beyan ederler.

Araştırmacıların Katkı Oranı Beyan Özeti

Yazarlar makaleye eşit oranda katkı sağlamış olduklarını beyan ederler.

Kaynakça

Goel A., Gautam J., Kumar S. Real time sentiment analysis of tweets using naive bayes. Proceedings on 2016 2nd International Conference on Next Generation Computing Technologies (Ngct) 2016; 257-261.

Hosmer DW., Lemeshow S., Sturdivant RX. Applied logistic regression third edition preface. Applied Logistic Regression, 3rd Edition, p. Xiii-+2013.

Jain AP., Dandannavar P. Application of machine learning techniques to sentiment analysis. Proceedings of the 2016 2nd International Conference on Applied and Theoretical Computing and Communication Technology (Icatcct), p. 628-632.

Kaggle. Yelp review sentiment dataset. 2021a. https://www.kaggle.com/ilhamfp31/yelp-review-dataset. (Erişim tarihi: 19.02.2021)

Kaggle. IMDB Dataset of 50K Movie Reviews, 2021b. https://www.kaggle.com/lakshmi25npathi/imdb-dataset-of-50k-movie-reviews. (Erişim tarihi: 19.02.2021).

Kouloumpis E., Wilson T., Moore J. Twitter sentiment analysis: The good the bad and the omg!, In Proceedings of the International AAAI Conference on Web and Social Media 2011; 5(1): 538-541.

Meng XR., Bradley J., Yavuz B., Sparks E., Venkataraman S., Liu D., Freeman J., Tsai DB., Amde M., Owen S., Xin D., Xin R., Franklin MJ., Zadeh R., Zaharia M., Talwalkar A. MLlib: Machine learning in apache spark. Journal of Machine Learning Research 2016; 17(1): 1235-1241.

Mohapatra S., Ahmed N., Alencar P. KryptoOracle: A real-time cryptocurrency price prediction platform using twitter sentiments. 2019 IEEE International Conference on Big Data (Big Data), p. 5544-5551.

Neethu MS., Rajasree R. Sentiment analysis in twitter using machine learning techniques. 2013 Fourth International Conference on Computing, Communications and Networking Technologies (Icccnt), pp: 1-5.

Pang B., Lee L., Vaithyanathan S. Thumbs up? Sentiment classification using machine learning techniques. Appears in Proc. 2002 Conf. on Empirical Methods in Natural Language Processing (EMNLP).arXiv preprint cs/0205070.

Wang H., Can D., Kazemzadeh A., Bar F., Narayanan S. A system for real-time twitter sentiment analysis of 2012 US presidential election cycle. In Proceedings of the ACL 2012 system demonstrations. 2012, 115-120.

Zaharia M., Chowdhury M., Das T., Dave A., Ma J., McCauley M., Franklin MJ., Shenker S., Stoica I. Resilient distributed datasets: A fault-tolerant abstraction for in-memory cluster computing. In 9th {USENIX} Symposium on Networked Systems Design and Implementation 2012; 12, 15-28.

Zhu GB., Blumberg DG. Classification using ASTER data and SVM algorithms: The case study of Beer Sheva, Israel. Remote Sensing of Environment 2002; 80(2): 233-240.