RapidMiner ile Twitter Verilerinin Konu Modellemesi*

Topic Modeling of Twitter Data via RapidMiner

Bilgi Yönetimi Dergisi
Cilt: 3 Sayı: 1 Yıl: 2020
https://dergipark.org.tr/tr/pub/by

Hakemli Makaleler
Araştırma Makalesi

Makale Bilgisi

Gönderildiği tarih: 02.11. 2019
Kabul tarihi: 02.03. 2020
Yayınlanma tarihi: 30.06. 2020

Article Info

Date submitted: 02.11. 2019
Date accepted: 02.03. 2020
Date published: 30.06. 2020

Anahtar sözcükler

Veri Madenciliği, Veri Analizi, Konu Modelleme, Twitter, RapidMiner

Keywords

Data Mining, Data Analysis, Topic Modeling, Twitter, RapidMiner

DOI numarası

10.33721/by.641878

ORCID

0000-0002-7968-485X (1)
0000-0002-2204-3170 (2)

Ela ANKARALI
Hacettepe Üniversitesi Bilgi ve Belge Yönetimi Bölümü Doktora Öğrencisi,
ela.ankarali@hacettepe.edu.tr

Özgür KÜLCÜ
Hacettepe Üniversitesi Bilgi ve Belge Yönetimi Bölümü Öğretim Üyesi,
kulcu@hacettepe.edu.tr

Öz

Bu çalışmada öncelikle RapidMiner kullanılarak Twitter'da belirli kelimeleri içeren tweet verileri elde edildi, bu veriler ön işlemden geçirildi ve sonrasında tweetlerin konu modellemesi yapıldı. Ön işleme için "Search Twitter", "Select Attributes", "Nominal to Text" blokları kullanıldı. Ön işlemden geçen Twitter verileri "Tokenize", "Aggregate" ve "Discretize" operatörleri kullanılarak analiz edildi. Tweetlerde en çok kullanılan kelimeler belirlendi ve kullanım sıklığına göre kelime grupları oluşturuldu. Daha sonra Twitter verilerine nasıl konu bazlı kümeleme yapılacağı anlatıldı. Bu işlem için Latent Dirichlet Allocation modelini kullanan "Extract Topics From Documents (LDA)" operatörü kullanıldı. Tweetlerde en fazla kullanılan kelimeler ve kullanıcı başına atılan tweet sayıları, grafik ve tablolarla incelendi, ayrıca konu modellemesi sonucunda elde edilen konuların kelime bulutu oluşturuldu.

Abstract

In this study, firstly, tweets containing specific words on the Twitter platform were obtained and pre-processed using the RapidMiner software. After that, the tweets are clustered based on the topic modeling approach. "Search Twitter", "Select Attributes", and "Nominal to Text" blocks were used for preprocessing. This preprocessed data is then analyzed using "Tokenize", "Aggregate", and "Discretize" operators. The most used words were determined, and tweets are grouped according to their frequencies. Then, it is explained how to perform topic-based modeling and clustering on Twitter data. "Extract Topics From Documents (LDA)" operator, which uses the Latent Dirichlet Allocation model, was used for this process. The most commonly used words in tweets, and the number of tweets per user were extracted and investigated via tables and graphical illustrations. In addition, the word cloud of each topic, obtained as a result of the topic modeling process, was created.

1. Giriş

Sosyal medya hızlı bir şekilde büyüyerek, sosyal yaşamın önemli bir parçası haline geldi. Bu gelişimin paralelinde ise, makine öğrenmesi ve veri madenciliği araçları hemen hemen bütün bilim dallarında aktif olarak kullanılan yöntemler haline geldi (Mitchell, 1999; LeCun, Bengio ve Hinton, 2015). Bu bağlamda sosyal medya veri madenciliği ve analizi üzerine birçok çalışma yayımlandı (Corley, Cook, Mikler ve Singh, 2010; Majid, Chen, Mirza, Hussain ve Woodward, 2013). Bu çalışmalarda kullanılan veri bilimi

**Bu çalışma e-BEYAS 2019 Sempozyumunda sunulan Poster Sunumun genişletilerek hazırlanmış tam metnidir.*

yazılım platformlarından bir tanesi de RapidMiner programıdır. Bu platform veri analizi, makine öğrenmesi, metin madenciliği gibi işlemleri gerçekleştirmek için grafiksel bir kullanıcı arayüzü sunmaktadır.

Twitter bireylerin belirli bir konudaki görüşlerini halka açık bir şekilde ifade etmek için yaygın olarak kullandığı sosyal medya platformlardan biridir (Jain ve Katkar, 2015). Bu nedenle, Twitter verileri üzerinde veri madenciliği uygulamalarıyla bilimsel veya pratik sonuçlar elde etmeyi amaçlayan birçok araştırma mevcuttur (Conover, Gonçalves, Ratkiewicz, Flammini, & Menczer, 2011; Culotta, 2010; Jiang ve Zheng, 2013; Earle, Bowden ve Guy, 2011). Conover ve diğerleri (2011), 2010 yılında Amerika Birleşik Devletleri'nde yapılan ara seçimlerle ilgili olarak Twitter üzerinden paylaşılan siyasi mesajların içeriği ve yapısına dayanarak, bireylerin siyasi yönelimlerini tahmin etme amaçlı çeşitli yöntemler geliştirmişlerdir. Culotta (2010), Twitter mesajlarından influenza ile ilgili olanları elde etmiş ve ortaya çıkan verileri regresyon yöntemi ile CDC (Hastalık Kontrol ve Önleme Merkezleri) istatistikleri ile ilişkilendirerek, tweetlerden influenza salgınını tahmin eden bir araç geliştirmiştir. Jiang ve Zheng (2013) ise, ilaç test deneylerinde gönüllü olan bireylerin Twitter mesajlarını inceleyerek ilaçların yan etkilerinin takip ve tespit edilmesine yönelik bir çalışma yürütmüştür. Earle ve diğerleri (2011), sadece Twitter verilerine dayanan bir deprem tespit aracı geliştirmişler ve bu yöntemi sismografik deprem tahmin yöntemleri ile karşılaştırmışlardır. Geliştirdikleri aracın, teknolojik olarak geri kalmış bölgelerde, sismografik yöntemlere göre çok daha hızlı bir şekilde depremi tahmin edebildiğini belirtmişlerdir.

Bu çalışmada öncelikle RapidMiner kullanılarak Twitter'da belirli kelimeleri içeren tweetler analiz edildi, sonrasında ise konu modelleme yapıldı, sonuçlar kelime bulutu ve grafikler oluşturularak incelendi. Bu bağlamda, Twitter gündeminin anlık olarak takip edilmesini sağlayan bir yaklaşım geliştirilmiş oldu. Çalışmanın amacı, geleneksel yöntemlerle analiz edilmesi mümkün olmayan büyük verinin analiz edilmesi için bir yöntem sunmaktır.

Ön işleme için "Search Twitter", "Write Excel", "Select Attributes", "Nominal to Text" blokları kullanıldı. Ön işlemden geçen Twitter verileri "Tokenize", "Aggregate" ve "Discretize" operatörleri kullanılarak analiz edildi. Tweetlerde en çok kullanılan kelimeler belirlendi ve kullanım sıklığına göre kelime grupları oluşturuldu. Daha sonra Twitter verilerine nasıl "konu" bazlı kümeleme yapılacağı anlatıldı. Bu işlem için "Extract Topics From Documents (LDA)" operatörü kullanıldı.

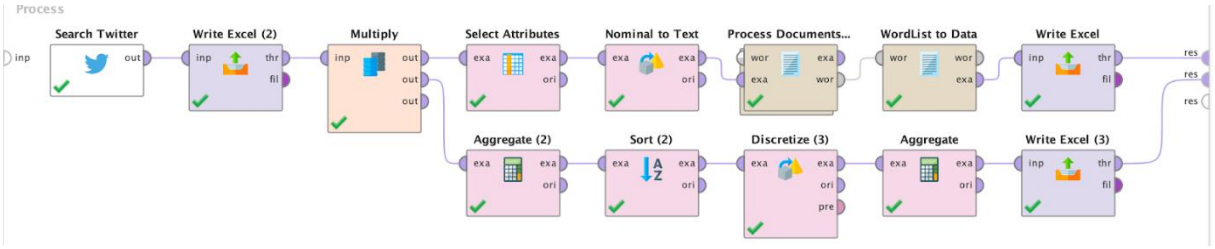
Bu çalışma kapsamında, örnek olarak "Hacettepe" kelimesinin geçtiği tweetlerin konu modellemesi yapılarak her konunun kelime bulutu oluşturulmuş, konuların benzerlik ve farklılıkları incelenmiştir.

2. Yöntem

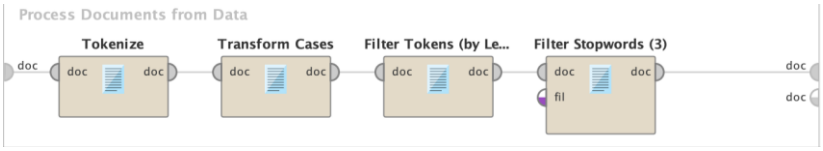
Birinci kısımda RapidMiner ile Twitter verilerinin nasıl elde edildiği açıklanmış, elde edilen veriler tablo, grafik ve kelime bulutu oluşturularak analiz edilmiştir. İkinci kısımda ise Twitter verilerinden konu modellemesi yapılarak, her konunun kelime bulutu oluşturulmuştur. Bu yöntem kullanılarak elde edilebilecek sonuçlara örnek teşkil etmesi açısından çalışmanın ilk kısmında "Ordu" kelimesini içeren tweetler, ikinci kısımda ise "Hacettepe" kelimesini içeren tweetler analiz edilmiştir. Seçilen kelimelerin çalışma konusuyla doğrudan ilişkisi bulunmamaktadır.

2.1. Twitter Veri Analizi ve Kelime Bulutu Oluşturma

Bu bölümde RapidMiner kullanılarak Twitter'da Ordu kelimesini içeren tweetler analiz edildi ve sonuçlar kelime bulutu ve grafikler oluşturularak incelendi. Şekil 1'de bu işlem için RapidMiner programında oluşturulan modelin ekran görüntüsü verilmiştir.



a)



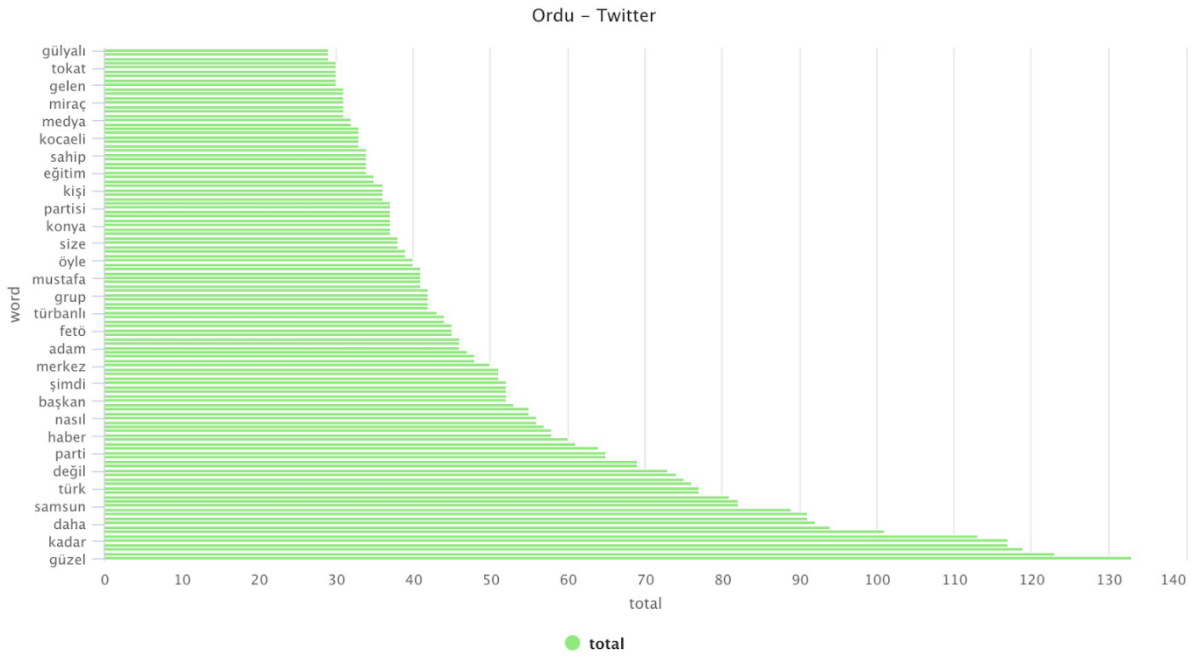
b)

Şekil 1. a) RapidMiner blok şeması – veri analizi b) RapidMiner blok şemasındaki process documents modülü içinde kullanılan alt operatörler

Öncelikle "Search Twitter" operatörü kullanılarak 1 – 9 Nisan 2019 tarihli Tweetler elde edildi. "Search Twitter" modülüne "query" olarak "Ordu -RT" girildi, böylelikle içinde Ordu sözcüğü geçen tweetler seçilirken Retweet edilen tweetler elenmiş oldu. Daha sonra "Write Excel" operatörü kullanılarak elde edilen veri excel dokümanı olarak kaydedildi. "Multiply" operatörü ile Twitter verisi iki ayrı proseste kullanılmak üzere çoğaltıldı.

İlk proseste önce "Select Attributes" bloğu kullanılarak Twitter verilerinden tweetleri ifade eden "Text" verileri filtrelendi. Daha sonra, "Nominal to Text" operatörü kullanılarak tweetler metin formatına çevirildi. "Attribute filter type" olarak "single", "attribute" olarak ise "Text" seçildi. Daha sonra "Process Documents from Data" operatörü içinde sırasıyla "Tokenize" (mode: non-leters), "Transform Cases" (transform to: lover case) "Filter Tokens (by Length)" (min chars: 4, max chars: 15) operatörleri kullanılarak tweetler kendilerini oluşturan kelimelere ayrıldı. Ayrıca "Filter Stopwords" operatörü ile bağlaç ve edat gibi tek başına anlamı olmayan sözcükler filtrelendi. Son olarak işlenmiş veri tekrar "Write Excel" operatörü kullanılarak excel dokümanı olarak kaydedildi. Daha sonra "Visualization" seçeneklerinden wordcloud kullanılarak elde edilen verinin kelime bulutu oluşturuldu. Şekil 2'de "Ordu" kelimesinin geçtiği tweetlerin kelime bulutu (argo sözcükler kaldırılarak) verilmiştir. Şekil 3'te ise Ordu kelimesinin geçtiği tweetlerde bulunan diğer kelimelerin kullanım sayısını gösteren grafik verilmiştir; grafikte kullanım sayısına göre oluşturulan aralıkların her biri için örnek olarak tek kelime grafik üzerine yazılmıştır.

Şekil 2. Ordu kelimesinin geçtiği tweetlerin kelime bulutu



Şekil 3. Ordu kelimesinin geçtiği tweetlerde bulunan diğer kelimelerin kullanım sayısı

İkinci proseste önce "Aggregate" operatörü ile her kullanıcının kaç tweet attığı belirlendi. "Sort" operatörü kullanılarak veriler tweet sayısına göre sıralandı. (attribute name: count(From-User), sorting direction: decreasing). Daha sonra "Discretize by user Specification" operatörü kullanılarak kullanıcı başına atılan tweet sayısı 0-2, 3-4, 5-8, 9-sonsuz aralıklarına ayrılarak kullanıcı başı atılan tweet sayısının dağılımı elde edildi. "Aggregate" operatörü ile her aralıkta toplam kaç kullanıcı olduğu

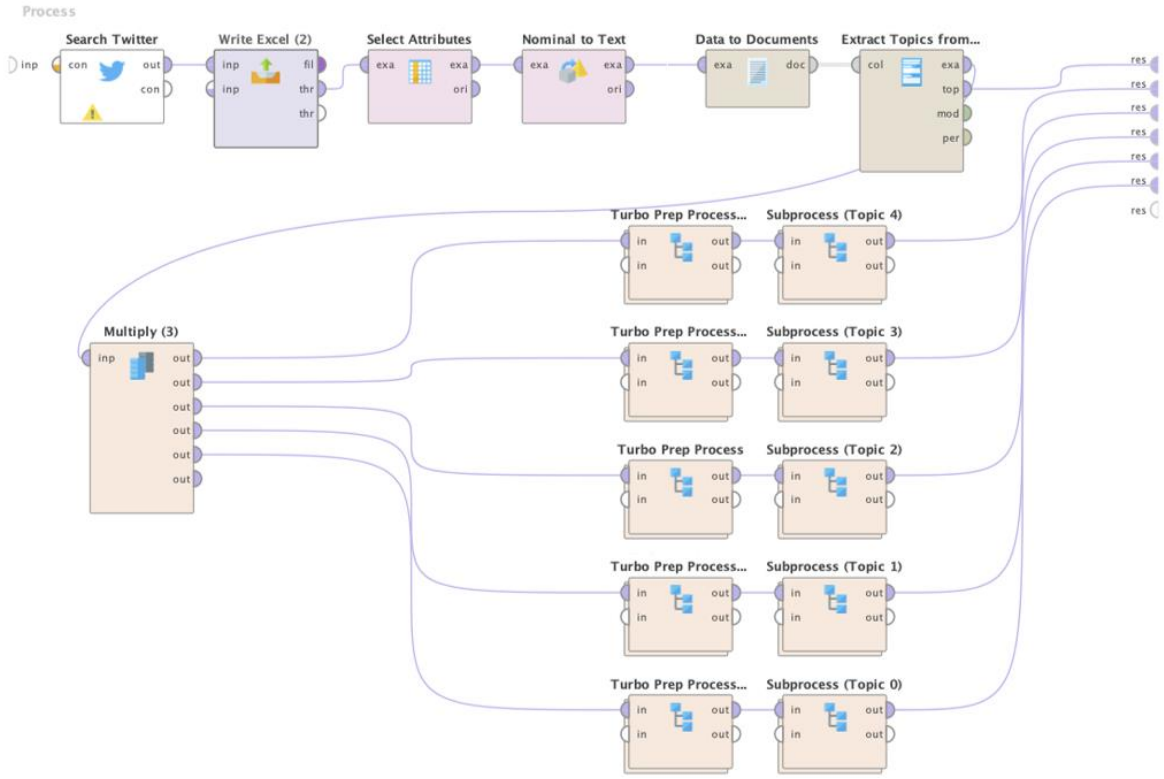
hesaplandı ve son olarak "Write Excel" ile elde edilen veri excel formatında yazdırıldı. "Visualization" ile çeşitli görseller elde edildi. Tablo 1'de Ordu kelimesinin geçtiği tweetlerdeki kullanıcı başında düşen tweet sayısının kullanıcılara göre dağılımı verilmiştir. Tabloda verilen sonuçlara göre, 2010 kişi 1-2 tweet, 94 kişi 3-4 tweet, 35 kişi 5-8 tweet, 18 kişi ise 9 veya daha fazla sayıda tweet atmıştır.

Row No.	count(From-User)	count(count(From-User))
1	0-2	2010
2	3-4	94
3	5-8	35
4	9-sonsuz	18

Tablo 1. Ordu kelimesinin geçtiği tweetlerdeki kullanıcı başında düşen tweet sayısının kullanıcılara göre dağılımı

2.2. Twitter Verilerinden Konu Modelleme

Bu kısımda RapidMiner kullanılarak taranan Twitter verilerine nasıl konu bazlı kümeleme yapılacağı anlatılmıştır. Şekil 4'te bu işlem için RapidMiner programında oluşturulan modelin ekran görüntüsü verilmiştir.



Şekil 4. RapidMiner blok Şeması - konu modelleme

Bu kısımda örnek olarak Twitter platformundaki, 9 – 18 Mayıs 2019 tarihli, Hacettepe kelimesini içeren tweetler incelendi. Twitter platformu, sadece güncel tarihlere ait tweet verilerini paylaştığı için, verinin elde edildiği tarih olan 19 Mayıs 2019 tarihi öncesinde platformun verilerini açtığı tarih aralığına ait veriler analiz edildi. Verileri elde etmek için "Search Twitter" operatörü kullanıldı ve "query" olarak "Hacettepe -RT" girildi. "-RT" ifadesi ile konu modellemesi için kullanılacak olan veriden retweet edilen tweet datası çıkarılmış oldu. Kümeleme yapabilmek için bu çalışmanın bir önceki kısmındaki ilk

4 blok/proses, "Search Twitter", "Write Excel", "Select Attributes", "Nominal to Text" doğrudan kullanıldı. "Nominal to Text" operatörünün çıktısında her kullanıcı için bir yazı verisi yani "text" mevcuttur. Bu operatörün çıktısını "Extract Topics From Documents (LDA)" girdisine uyumlu hale getirmek için "Data to Documents" operatörü kullanılmıştır. "Extract Topics From Documents (LDA)" operatörü kümeleme yapmak için Latent Dirichlet Allocation yöntemini kullanır.

Konu modellemesinin çeşitli yöntemleri bulunmakla birlikte, RapidMiner yazılımı içinde de farklı konu modelleme operatörleri bulunmaktadır. Latent Dirichlet Allocation (LDA) algoritması yaygın olarak kullanılan konu modelleme tekniklerinden biridir (Blei, Ng ve Jordan, 2003). Doğal dil işlemede (Natural Language Processing) LDA yöntemi istatistiksel bir konu çıkarma ve modelleme yöntemidir. LDA her verideki kelimelerin birbiriyle ilgili olduğunu varsayar ve verinin nasıl oluşturulduğunu ve kelime dağılımını çözmeye çalışır. Böylece kullanıcı tarafından belirlenen sayı kadar konu altında veriyi gruplar, böylece benzer konudaki veriler belirlenmiş olur (Lamba ve Madhusudhan, 2018, s. 2).

Bu çalışmada elde edilen veriler 5 konuda (topic) gruplanmıştır. Bu çalışmada örneklem, tweetler içinde kullanılan kelimelerden oluşurmuştur. LDA her tweetin az sayıda konunun bir karışımı olduğunu ve her kelimenin atanan konulardan birine atfedildiğini belirtir. Tablo 2'de "Search Twitter" operatörünün verdiği ilk 11 tweete ait LDA tabanlı konu modelleme sonucu gösterilmiştir. LDA analizi sonucunda, her bir tweetin modellenen konuları içerme olasılığı ortaya çıkarılır. Örneğin, 1. tweetin içinde Topic 0 olma olasılığı yaklaşık olarak %0, Topic 1 olma olasılığı ise %95'tir. Diğer konuların olma olasılığı da Topic 1'e göre oldukça düşüktür. İstatiksel açıdan, 1. tweet Topic 1 grubuna aittir. Bu bağlamda "Extract Topics From Documents (LDA)" operatörü tweetleri olasılığı en yüksek olan konu altında gruplar.

Tweet No	**Konu Tahmini**	**Olasılık Topic 0**	**Olasılık Topic 1**	**Olasılık Topic 2**	**Olasılık Topic 3**	**Olasılık Topic 4**
1	Topic_1	0.00	0.95	0.00	0.00	0.03
2	Topic_0	0.77	0.15	0.01	0.01	0.06
3	Topic_1	0.01	0.55	0.40	0.04	0.01
4	Topic_1	0.01	0.97	0.01	0.01	0.01
5	Topic_4	0.02	0.46	0.02	0.02	0.49
6	Topic_1	0.01	0.64	0.33	0.01	0.01
7	Topic_1	0.01	0.98	0.01	0.01	0.01
8	Topic_1	0.02	0.84	0.02	0.02	0.11
9	Topic_0	0.49	0.47	0.01	0.01	0.01
10	Topic_0	0.87	0.05	0.07	0.01	0.01
11	Topic_4	0.01	0.33	0.01	0.08	0.56

Tablo 2. İlk 11 tweet için LDA tabanlı konu modellemesi sonuçları

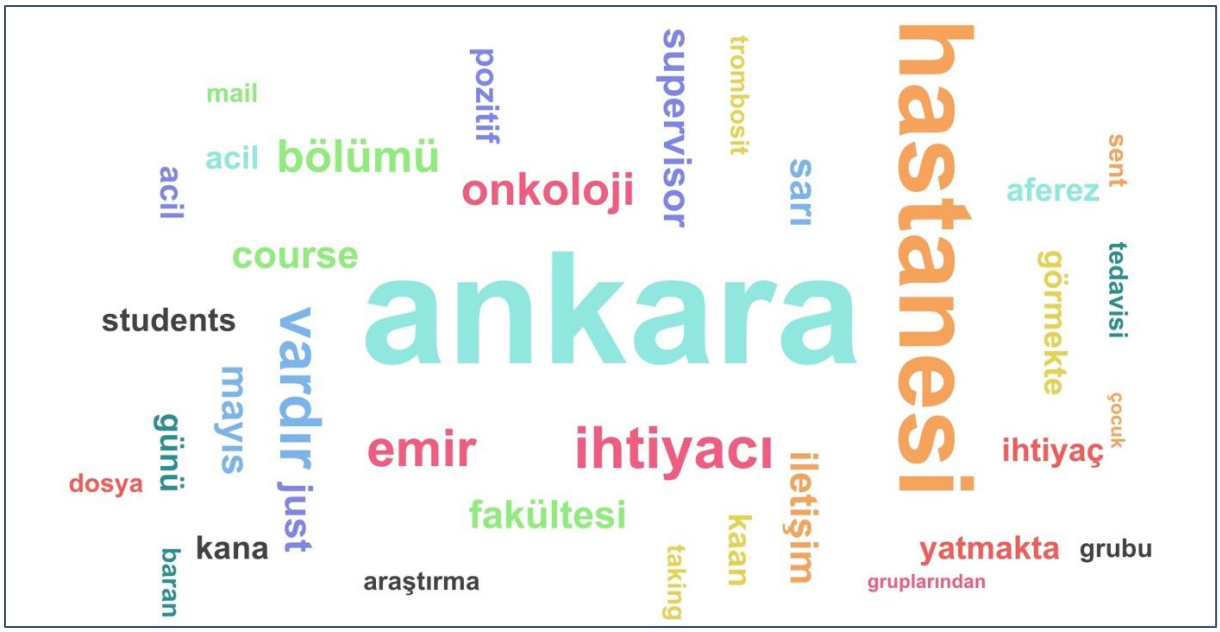
Daha sonra modellenen konuların ayrı ayrı kelime bulutu oluşturulmuştur. Bu işlem için, RapidMiner'ın "Turbo Prep Process" özelliği kullanılmıştır. Bu özellik sayesinde elde edilen analiz sonuçlarını filtreleyerek çalışmak mümkün olmuştur ve her konuya ait tweetler için ayrı operatör oluşturulması sağlanmıştır. Şekil 4'te ilgili bloklar görülebilir.

Her konuya ait kelime bulutu Şekil 5-9'da verilmiştir. Görüldüğü üzere, kelime bulutları arasında benzerlikler bulunmakla birlikte belirgin farklılıklar mevcuttur. Örneğin "hastane" kelimesi ve hastaneyle ilgili diğer kelimelerin sadece Topic_3 konu grubunda ortaya çıktığı görülmektedir. Topic_4 konu grubunda ise öne çıkan kelimenin ODTU olduğu gözlemlenmiştir.

Şekil 5. Topic 0 kelime bulutu

Şekil 6. Topic 1 kelime bulutu

Şekil 7. Topic 2 kelime bulutu



Şekil 8. Topic 3 kelime bulutu

Şekil 9. Topic 4 kelime bulutu

3. Sonuç ve Öneriler

Bu çalışmada, RapidMiner yazılımı kullanarak, Twitter'da kullanıcılar tarafından belirlenen kelimeleri içeren tweetlerin nasıl elde edileceği, hangi yöntemler kullanılarak işlenip analiz edileceği ve Latent Dirichlet Allocation algoritması kullanılarak konu modellemesinin nasıl yapılabileceği örnekler üzerinden anlatıldı. İlk olarak Twitter'dan "Search Twitter" operatörü ile elde edilen tweetlerden nasıl kelime bulutu oluşturulabileceği anlatıldı.. Bu işlemi gerçekleştirmek için RapidMiner içindeki "Select Attributes", "Nominal to Text", "Process Documents from Data" ve "Word List to Data" blokları seri bir şekilde bağlanarak çalıştırılmıştır. Şekil 1.a'da görülebileceği üzere bu blokların aralarında iki adet "Write Excel" operatörü kullanılmıştır ve bu blokların sürece doğrudan bir etkisi yoktur. Sadece ara fazlardaki işlenmiş verileri Excel dosyasına kaydederler. Örnek bir görsel çıkartma amacı ile, bu kapsamda "Ordu" sözcüğü geçen tweetler incelenmiş ve Şekil 2'de verilen kelime bulutu oluşturulmuştur.

İkinci analizde ise, "Aggregate", "Sort" ve "Discretize by user Specification" blokları kullanılarak elde edilen tweetler içinde 0-2, 3-4, 5-8, 9-sonsuz aralıklarında tweet atan kaç kullanıcı olduğu ortaya çıkarılarak sonuçları Tablo 1'de gösterildi. Elde edilen bu tabloda, kullanıcıların yaklaşık %95'inin Ordu ile ilgili sadece 0-2 tweet attığı gözlemlenmektedir.

Son analiz yönteminde ise, RapidMiner kullanarak belirli kelemlerin geçtiği tweetlerden nasıl istatiksel konu modellemesi yapılacağı detaylı olarak anlatılmıştır. Bu çalışmada doğal dil işlemede yoğun bir şekilde kullanılan Latent Drichlet Allocation algoritması tercih edilmiştir. RapidMiner yazılımı içindeki

"Extract Topics From Documents (LDA)" operatörü bu algoritmayı kullanarak girdi olarak verilen metin verileri üzerinden konu modellemesi yapar. Bu yöntemi test etmek amacıyla, örnek olarak "Hacettepe" kelimesini içeren tweetler incelenmiş ve elde edilen veriler 5 konuda (topic) gruplandırılmıştır. "Extract Topics From Documents (LDA)" operatörü içinde konu sayısı kullanıcı tarafından seçilebilmektedir ve konu sayısına bağlı olarak sonuçların nasıl etkileneceği gelecekteki çalışmalarımızda incelenecektir.

Son olarak konu modellemesi işleminin nasıl bir sonuç ortaya çıkardığını gözlemleyebilmek için her konu grubuna ait tweetlerin ayrı ayrı kelime grupları elde edilmiştir ve bu kelime gruplarından kelime bulutları oluşturulmuştur (Şekil 5, 6, 7, 8 ve 9).

9 – 18 Mayıs 2019 tarihli Hacettepe Üniversitesiyle ilgili haberler incelendiğinde, Hacettepe Üniversitesinin iş makinaları üreten bir firma ile imzaladığı iş birliği protokolü, ODTÜ öğrencilerine bir organizasyonda Hacettepe Üniversitesi öğrencilerinin destek olması ve Hacettepe Uluslararası Dostluk Gününün ilgili tarihlerdeki tweetleri, dolayısıyla elde edilen konuları etkilediği düşünülmektedir. Şekil 8'de verilen konunun ise Hacettepe Üniversitesi Hastaneleri ile ilgili olduğu görülmektedir. Bu çalışmada sunulan yöntem sadece Twitter verisine yönelik olmayıp, her türlü büyük metin verisi analizi için kullanılabilir.

Tong ve Zhang (2016, s. 209), LDA algoritmasını Wikipedia makalelerinden konu modellemesi yapmak ve ortaya çıkan verileri analiz etmek için kullanmıştır. Yaptıkları çalışma ile makalelerin organize edilmesine ve okuyuculara ilgilerini çekebilecek yeni makaleleri önerme yöntemleri geliştirilmesine katkı sağlamayı amaçlamışlardır. Lamba ve Madhusudhan (2018), RapidMiner programını kullanarak "DESIDOC Journal of Library and Information Technology" dergisinde 2017 yılı içinde yayınlanan 50 makale verisine LDA ile konu modellemesi yapmış ve makaleleri 5 alt konu grubuna ayırmıştır. Elde edilen sonuçların ve yöntemlerin kütüphanelerde makalelerin araştırılmasında ve önerilmesinde faydalı olabileceğini vurgulamıştır.

Bu çalışmada kullanılan yöntem, kütüphanelerin kitap kataloglarını gruplama, dergilerin detaylı içerik analizi, belirli bir alandaki yayınların yıllara göre konu dağılımı değişimi vb. çalışmalarda kullanılabilir.

Teşekkür

Görüşleri ve yorumları için Mustafa Mert Ankaralı'ya teşekkür ederiz.

Kaynakça

Blei, D. M., Ng, A. Y. and Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3(Jan), 993-1022.

Conover, M. D., Gonçalves, B., Ratkiewicz, J., Flammini, A. and Menczer, F. (2011, October). Predicting the Political Alignment of Twitter Users. *In 2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing* (pp. 192-199). IEEE.

Corley, C., Cook, D., Mikler, A. and Singh, K. (2010). Text and Structural Data Mining of Influenza Mentions in Web and Social Media. *International Journal of Environmental Research and Public Health*, 7(2), 596-615.

Culotta, A. (2010, July). Towards Detecting Influenza Epidemics by Analyzing Twitter Messages. *In Proceedings of the First Workshop on Social Media Analytics* (pp. 115-122). Acm.

Earle, P. S., Bowden, D. C. and Guy, M. (2012). Twitter Earthquake Detection: Earthquake Monitoring in a Social World. *Annals of Geophysics*, 54(6).

Jain, A. P. and Katkar, V. D. (2015). Sentiments Analysis of Twitter Data Using Data Mining. *In 2015 International Conference on Information Processing* (ICIP) (pp. 807-810). IEEE.

Jiang, K. and Zheng, Y. (2013, December). Mining Twitter Data for Potential Drug Effects. *In International Conference on Advanced Data Mining And Applications* (pp. 434-443). Springer, Berlin, Heidelberg.

Lamba, M. and Madhusudhan, M. (2018). Application of Topic Mining and Prediction Modeling Tools for Library and Information Science Journals. *Library Practices in Digital Era.* Eds. MR Murali Prasad et al. Hyderabad: BS Publications, 395-401.

LeCun, Y., Bengio, Y. and Hinton, G. (2015). Deep Learning. *Nature*, 521(7553), 436-444.

Majid, A., Chen, L., Chen, G., Mirza, H. T., Hussain, I. and Woodward, J. (2013). A Context-Aware Personalized Travel Recommendation System Based on Geotagged Social Media Data Mining. *International Journal of Geographical Information Science*, 27(4), 662-684.

Mitchell, T. M. (1999). Machine Learning and Data Mining. Communications of the ACM, 42(11).

Tong, Z. and Zhang, H. (2016). A Text Mining Research Based on LDA Topic Modelling. *International Conference on Computer Science, Engineering and Information Technology* (pp. 201-210).