T.C.
ERCİYES ÜNİVERSİTESİ
BİLİMSEL ARAŞTIRMA PROJELERİ
KOORDİNASYON BİRİMİ

VERİ MADENCİLİĞİNDE METİN MADENCİLİĞİ (TEXT MINING) YAKLAŞIMI

Proje No: FBA-2014-4850

Normal Araştırma Projesi

SONUÇ RAPORU

Proje Yürütücüsü:
Feyza Gürbüz
Mühendislik Fakültesi/Endüstri Mühendisliği Bölümü

Esra Kahya Özyirmidokuz
Kayseri Meslek Yüksekokulu/ Bilgisayar Teknolojileri Bölümü

"Bu çalışma Erciyes Üniversitesi Bilimsel Araştırma Projeleri Koordinasyon Birimince Desteklenmiştir. Proje Numarası: FBA-2014-4850"

Şubat 2017
KAYSERİ

İÇİNDEKİLER

ÖZET

Günümüzde veri madenciliği firmalar açısından çok önemli hale gelmiştir. Firmalar sektörde rekabet avantajı sağlayabilmek için veri madenciliği tekniklerini kullanarak büyük veriden işlerine yarayacak, daha önceden keşfedilmemiş, kullanılabilir örüntüler elde eder. Gelişen haberleşme teknolojileri sonucu firmalarda biriken veri yığınları, firmalar için hayati önem taşıyan bilgileri içinde barındırır. Karar vericiler, klasik tekniklerle bu verilerden çıkarımlarda bulunurken, önemli bilgileri gözden kaçırırlar. Veriyi doğru yönetemeyen firmalar ise işlerine yaramayan veri yığınlarında kaybolur. Bir işletmeye ait sayısal platformdaki bu verilerin %80'i metin formundadır. Ancak yapısal olmayan verileri de içeren büyük veri klasik istatistiksel tekniklerle analiz edilen veriler kadar kolay işlenemez. Doğal dil işleme tekniklerinden faydalanılması gerekmektedir. Böylece, soyut ve yığın yapısal olmayan bilgiler, sayısal somut ifadelere dönüştürülebilmektedir. Bu araştırma, Kayseri'de bir imalat fabrikasında yapılan üst düzey toplantıların metin formatındaki tutanaklarını analiz ederek bilgi çıkarımı gerçekleştirmektedir. Yöneticilerin verdiği stratejik kararlarda önemli toplantı sonuçları çok etkilidir. Araştırmanın en genel amacı toplantıların kalitesini artırmaktır. Araştırmada, toplantı tutanaklarından kelime çıkarımı yapılarak, toplantılara ait genel konu başlıkları metin madenciliği ile elde edilecektir. Yöneticiler çeşitli madenleme teknikleriyle gruplanmış konu başlıklarına göre değerlendirme yaparak sonraki toplantıların kalitesini artırarak zaman kazanabilir.

Anahtar Kelimeler: Metin madenciliği, Doğal dil işleme, Yönetim, Üretim Toplantı Raporları

ABSTRACT

Recently, data mining has become crucial for firms. Using data mining, Firms, in order to have comparative advantage in industry / sector / market, obtain patterns that they can utilize and that have not been discovered before. The data accumulated as a result of the advanced communication channels within firms contain crucial information. Decision makers' undersees important information while they use classical techniques for data analysis. Firms that cannot manage data accurately get lost in piles of data that would not be useful for them. 80% of the data in the quantitative platform belonging to a firm is in text format. However, large data containing non-structural data cannot be analyzed as easily as the data analyzed by using classical statistical techniques. Natural language analysis techniques should be used. In this way, abstract and non-structural data can be converted into concrete and quantitative statements. In this analysis, information is inferred by the analysis of transcripts—in text format—of meetings among senior managers at a manufacturing company in Kayseri. Outcomes of the important meetings are very crucial in the decisions the directors take. One of the main goals of the study is to increase the quality of the meetings. In this research, the general themes of the meetings are found out by word inference from the meeting transcripts. Directors can have better time-management by increasing the quality of the future meetings by conducting evaluations according to the topics categorized by the data mining techniques.

Keywords: Text Mining, Natural language processing, Management, Manufacturing meeting reports

1. GİRİŞ

İşletmelerde güçlü bilgi sistemlerinde depolanan büyük veri yığınları, potansiyel yeni bilgiyi içinde barındırır. Bu verileri tüm süreçlerden eşzamanlı toplamak, otomatik analiz tekniklerini kullanarak bilgiyi seçmek ve analiz işlemleri çok karmaşıktır. Bu veriler, imalat firmalarında, makine, ürün, süreç, bakım, kalite kontrol, hata teşhisi, vs. veriler olabilir ve tipik olarak, veri tabanlarında depolanır. Bu veri yığınları, içlerinde de gizli ve değerli bilgiler barındırmaktadır. Bu bilgiler, günümüz koşullarında işletmelere rekabet avantajı sağlamak açısından önemlidir. Örneğin, bilgi çıkarımı sonucunda imalat sürecinde verimlilik artışı için yapılması gerekli ve önemli bilgiler elde edilir. Veri tabanlarında bilgi keşfi süreci (KDD)'nin bir adımı olan fakat zamanla KDD'nin kendisiyle aynı anlamda kullanılmaya başlayan veri madenciliği (DM), büyük miktardaki verilerden otomatik bir şekilde kullanışlı ve daha önceden keşfedilmemiş bilgiyi zeki tekniklerle çıkan bir bilgisayar bilimidir. Metin madenciliği (TM), yapısal olmayan veri türünün analizini yapmak için kullanılan DM yöntemlerini kapsar. Veri tabanlarında zaten mevcut örüntülerin otomatik olarak elde edilmesi sonucunda kullanılabilir modeller elde edilir. Bu modeller, karar vericiler ve mühendisler tarafından işletmenin performansını artırmak için doğrudan kullanılabilir.

Sayısal ortamda her tipte verinin sürekli ve artan şekilde depolanması sonucu yapısal olmayan verilerin analizi gerekli hale gelmiştir. Mühendis ve yöneticilerin bu tür yığın veriyi anlamaları için analiz etmeleri gerekir. Günümüzde, dünyada rekabet üstünlüğü elde etmek için kullanılan yapısal olmayan verinin analizi için geleneksel veri analiz yöntemleri yetersiz kalır. Veri tabanlarında bilgi keşfi süreci içerisinde verinin bütününün döküman analiz algoritmaları kullanılarak değerlendirilmesi gerekir.

Kalite, bir firmanın günümüz rekabet iş çevresinde ayakta durabilmesi için çok önemli bir faktördür. Teknolojik gelişme, imalat kalite kontrol performansını artırmak için kalite iyileştirmeye DM gibi yeni boyutlar katmıştır. Stratejik kararların verildiği üst düzey toplantılarda tutulan tutanaklar, toplantılarda görüşülen konuların genel bir çerçevesini içermektedir. Yöneticiler tek bir toplantı sonucunda birtakım kararlar vererek harekete geçerler. Biriken tutanaklar genel olarak içlerinde yöneticilerin de fark etmediği gizli ve değerli bilgileri elde etmek üzere değerlendirilmesiyle, daha sonraki stratejik toplantılara bir çerçeve geliştirilebilecektir. Böylece yöneticiler zaman kazanırken, uzun vadede toplantı kalitesi artmış olacaktır.

İmalat firmalarında son yıllarda artan metin dökümanlarının hacmi dolayısıyla, özellikle doğal dilde bu dökümanları işlemenin zorunluğunu beraberinde getirmiştir. Ancak, yapısal olmayan bu büyük miktarlardaki veri tiplerinin analizi çok karmaşık teknikleri kullanmayı gerektirmektedir. Doğal dil işleme gibi farklı bilgisayar disiplinleri ile ortak araştırma yapmayı gerektiren TMnin imalatta kullanımı son birkaç yıldır yaygınlaşmaya başlamıştır. Türkiye'de ise imalat sektöründe bu alanda çalışma neredeyse yoktur. DM uygulamalarında, yapısal olmayan veriden anahtar kelimelerin elde edilmesi güçlü tahmin edicilerdir. TM bu anahtar kelimelerin çıkarımını yapmada kullanıldığında, model performansı artar. Yapısal olmayan metin, tek tip anket formu sonuçlarından çok daha açıklayıcıdır (Thompson vd., 2012). Anahtar kelimeler, bir dökümanın içeriğini temsil eder. İdeal olarak, bir dokümanın gerekli içeriğinden yoğunlaştırılmış anahtar kelimeleri temsil eder. Anahtar kelimeler, tanımlaması, değiştirilmesi, hatırlanması ve paylaşılması kolay olduğu için, genelde, Bilgi Geri Kazanım (Information Retrieval, IR) içinde sorguları tanımlamada kullanılır. Matematiksel işaretlerin aksine, herhangi bir yapıdan bağımsızdır ve çoklu yapılara ve IR sistemlerine uygulanabilirler (Rose vd., 2010).

2. LİTERATÜR ÖZETİ

İmalat firmalarında DM uygulamaları uzun süredir çalışılmakta olan güncel bir konudur (Kumar vd., 2007; Li, Yeh, 2008; Çiflikli and Kahya-Özyirmidokuz, 2010; Gebus, Leiviska, 2009; Kusiak, Smith, 2007; Kang vd., 2009; Durán vd., 2010; Liao vd., 2012). Harding ve diğerleri (2006)'da ve Wang 2007'de DM'nin imalatta uygulamalarını araştırmıştır. 2010 ve 2012 yıllarında, Çiflikli ve Kahya Özyirmidokuz, yine Kayseri'de bir fabrikasından topladıkları binlerce veri içinde gizli bilgilerin çıkarımını yapmışlar, yöneticilerin kullanmaları için karar ağacı modellerini elde etmişlerdir. Çiflikli ve Kahya Özyirmidokuz (2012) çalışmalarında, kalite kontrol departmanındaki 73 adete indirgedikleri hata sebeplerini karar ağaçlarında kullanılan bilgi kazancı ile özellik uygunluk analizini kullanarak en önemli 23 adet hata sebebini elde etmişlerdir. Bu araştırmamızda ise, veri boyut indirgeme için bu teknikler yerine anahtar kelime çıkarımı yapılması yeterli olmuştur. Ayrıca, dökümanlar kesiklendirilmeden, doğrudan serbest metin halinde analize sokulmuştur.

Dünyada son birkaç yıldır iş dünyasında TM tekniklerinin kullanımı yaygınlaşmaya yeni yeni başlamıştır (Ittoo, Bouma, 2013; Thorleuchter, Van den Poel, 2014, Kahya-Özyirmidokuz, 2014, Kahya-Özyirmidokuz, Özyirmidokuz, 2014). TM ile ilgili üretim işletmelerinde de araştırmalar mevcuttur (Liu vd., 2006; Negahban, Smith, 2014).

Son zamanlarda TM işletmeler için önemli bir araştırma alanı olmuştur (Özyirmidokuz, 2014). Chang vd. (2009) internet ve e-ticaret müşterilerinin davranışlarını doğru şekillendirmek için veri ambarları ve veri madenciliği teknolojilerini kullanmışlardır. Gamon (2004), müşteri geri bildirim verilerinin otomatik sentimatik sınıflandırmasının yapılabilceğini belirterek, doğal dil işleme ve lineer destek vektör makinalarını sınıflandırma doğruluklarını yükseltmek için kullanmıştır. Gamon vd. (2005), çalışmalarında serbest müşteri geridönüşüm metinlerinin başlıklarını ve sentiment oryantasyonu yapan bir prototip sistem sunmuşlardır. Ittoo vd. (2006), online ürün özelleştirmedeki kararlarını vermede bir metin madenciliği temelli öneri sistemi sunmuşlardır. Coussement ve Van den Poel (2008), çalışmalarında otomatik bir e-mail sınıflandırma sistemi geliştirmişlerdir. Weng ve Liu (2004), çok yönlü e-mailleri başlıklarına göre düzenleyen bir şablon önermişlerdir. Özyurt ve Köse (2010), online görüşmelerin özelliklerini belirlemek üzere makine öğrenme ve veri madenciliği metodlarını kullanmışlardır. Thorleuchter vd. (2010), yeni ve işe yarar fikirleri çıkarmak için yapısal olmayan metinlerden fikir analizi yapmışlardır.

Tsai ve Kwee (2011), yeniliklerin analizinin uygunluğunu ve performansını veri tabanı optimizasyonu ile yapmıştır. Gopol vd. (2011), verinin ve metin madenciliğinin durumunu özetlemiştir. Sunikka ve Bragge (2012), araştırmaların kişiselleştirmek ve uyarlama için metin madenciliği yaklaşımını geleneksek literatür taraması ile kombine etmiştir. Onishi ve Manchanda (2012), Japon sinema kategorisinde yeni ürün ve reklam satış sonuçlarını analiz etmiştir. Armentano vd. (2013), metin analizinde farklı profil stratejilerinin etkilerini kullanıcıların rollerini de dikkate alarak belirlmiştir. Thorleuchter ve Van den Poel (2012), e-ticaret firmalarının websitelerinden alınan metinsel bilgilerin ticari başarıları üzerine etkilerini analiz etmişlerdir. Thorleuchter vd. (2012), web metin madenciliğini kullanarak bir Alman şirketinin müşterilerini analiz etmişlerdir. Ur-Rahman ve Harding (2012), metinsel verileri iki farklı sınıfa ayırmak için metin madenciliği ve metinsel veri madenciliğinin hibrid uygulamalarına yoğunlaşmışlardır. Hao (2012), doküman sınıflamada k-medoids ve k-medoids sosyal evrimsel programlama algoritmalarını karşılaştırmıştır. He vd. (2013), üç büyük pizza zincirinin facebook ve twitter sitelerindeki yapısal olmayan metin içeriklerine metin madenciliği uygulamışlardır. Kahya Özyirmidokuz (2014), Türkiye'de online alışveriş

sitelerini metin madenciliğininde doğal dil işleme kullanarak analiz etmiştir. Kahya Özyirmidokuz ve Özyirmidokuz (2014), web metin madenciliği ile Türkiye'deki en iy yedi ısıtma sistemi firmalarının müşteri şikayet dökümanlarını analiz etmişlerdir. Ordenes vd. (2014), dilsel tabanlı metin madenciliği modelini geliştirme ve iyileştirme prosesini modellemek için kullanmışlardır.

Peter vd. (2016) sanal toplantıların İsveç'te kamu ve özel kuruluşları üzerindeki olası etkilerini ve bu etkilerin varlığı ile gücünü incelemişlerdir. Kullanılan veriler, 3 farklı organizasyondan 23 farklı görüşme sonucunda ve literatür incelemesinden elde edilmiştir. Kurumlarda günlük toplantılar çok yaygın olarak yapılmasına rağmen çok az dikkate alınmaktadır. Stray vd. (2016) bu günlük toplantılarla ilgili çalışmalar yapmışlardır. Basu ve Murhy (2015) çalışmalarında yeni bir hiyerarşik ve geleneksel k-ortalama kümeleme tekniği ile combine edilmiş hybrid bir döküman sınıflandırma tekniği sunmuşlardır. Kim vd. (2016), çalışmalarında internette farklı platform ziyaretçilerinden toplanan büyük verileri analiz etmişlerdir. Hussain ve Suryani (2015) metinsel dökümanlardaki semantik benzerlikleri en yakın komşu algoritması ile belirleyerek intihalleri tespit eden bir çalışma önermişlerdir. Zhang ve Chow (2012), döküman analizinde hybrid döküman benzerliğini kullanan bir multi-seviyeli eşleştirme metodu sunmuşlardır. Çalışmada dökümanlar döküman ve paragraf seviyelerini içeren bir yapı sözkonusudur.

3. GENEL BİLGİLER

Bir veri madenciliği çalışması olarak kabul edilen Metin madenciliği çalışmalarında yazılı metinler veri kaynağı olarak kullanır. Amacı metin üzerinden yapısallaştırılmış verileri elde etmektir. Metin madenciliği, metinlerin sınıflandırılması (clustering), metinlerden konu çıkarılması (concept extraction), metinler için sınıf taneciklerinin üretilmesi (production of granular taxonomy), metinlerde görüş analizi yapılması (sentimental analysis), metin özetlerinin çıkarılması (document summarization) ve metinin özü ile ilgili ilişki modellemesi (entity relationship modelling) gibi çalışmaları kapsamaktadır. Metin madenciliği genellikle yapısal halde olmayan metinlerden ilgi çekici bilgi ve anlam çıkarma işlemi olarak tanımlanır (Hotho vd., 2005).

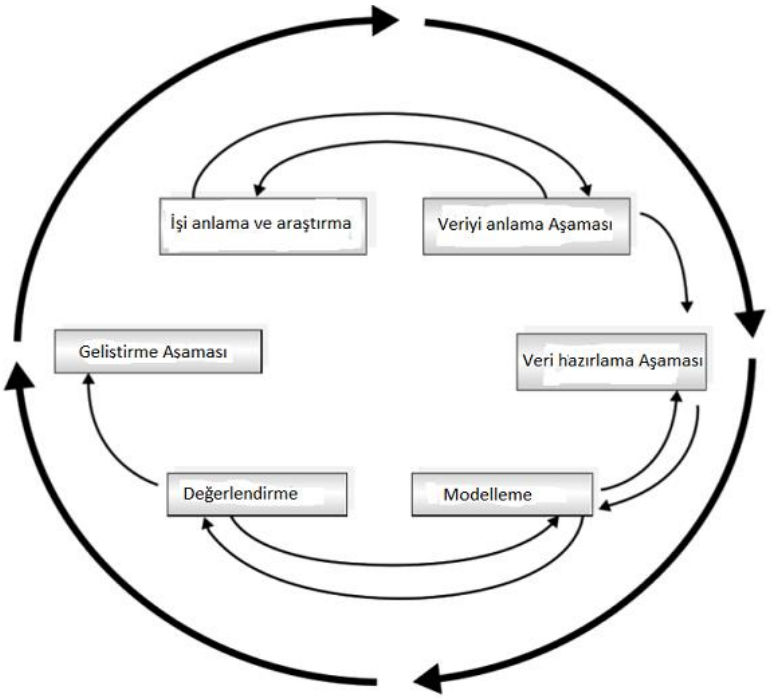
3.1. Metin madenciliği kavramı

DM tam olarak yapısal verileri analiz eden yöntemleri kapsar. Fakat bilgisayar ortamındaki büyük verinin %80'i metin tipindedir. Bu durum, yapısal olmayan verilerin analizini bir zorunluluk haline getirmektedir. Çok karmaşık bir süreç olan TM, yapısal olmayan verilerden anlamlı örüntüler bulmak için DM'den daha farklı veri analiz teknikleri kullanır.

TM, içeriklerinin ve konularının çıkarımını yapma ve yapısallaştırma, hızlı analiz yapma, gizli verinin keşfi, ve otomatik karar verme amaçlarıyla, bilgisayar ortamındaki büyük miktarlarda doğal dildeki metin verilerinin çeşitli tekniklerle otomatik işlenmesidir.

TM, çalışmanın tarihini veya yazarını belirlemede kullanılan metinlerin stillerini çalışan stylometriden farklıdır. Ancak, TM çok boyutlu istatistiğin gelişmiş yöntemlerini kullanarak, dilsel istatistik veya nicel dilsel yöntemler olarak adlandırılan lexicometri veya lexical istatistiğin bir uzantısıdır (Tuffery, 2011).Şekil 1'de sunulan ve1996 yılında DaimlerChrysler, SPSS, ve NCR firmalarını temsil eden analistler tarafından geliştirilmiş bir DM süreci olan (Larose, 2005: 6) CRISP-DM (Cross Industry Standard Process for Data Mining, DM için Çapraz Endüstri Standardı Süreci) metodolojisi, araştırmamızı sistematik bir çerçevede

yapmak için kullanılmıştır. Araştırmada CRISP-DM DM iş süreci, araştırma sorusuna çözüm bulmak üzere kullanılmıştır. Bu süreç içerisinde, kalitatif verilerin analizleri yapılmıştır. CRISP-DM, yapısal bir bilgi keşfi süreci yaklaşımıdır ve araştırmamızın analiz sıralarını adımlar. Bu yaklaşım, veri analiz sürecini daha hızlı, gerçekçi, daha yönetilebilir ve az maliyetli yapar (Sumathi, Sivanandam, 2006: 661).



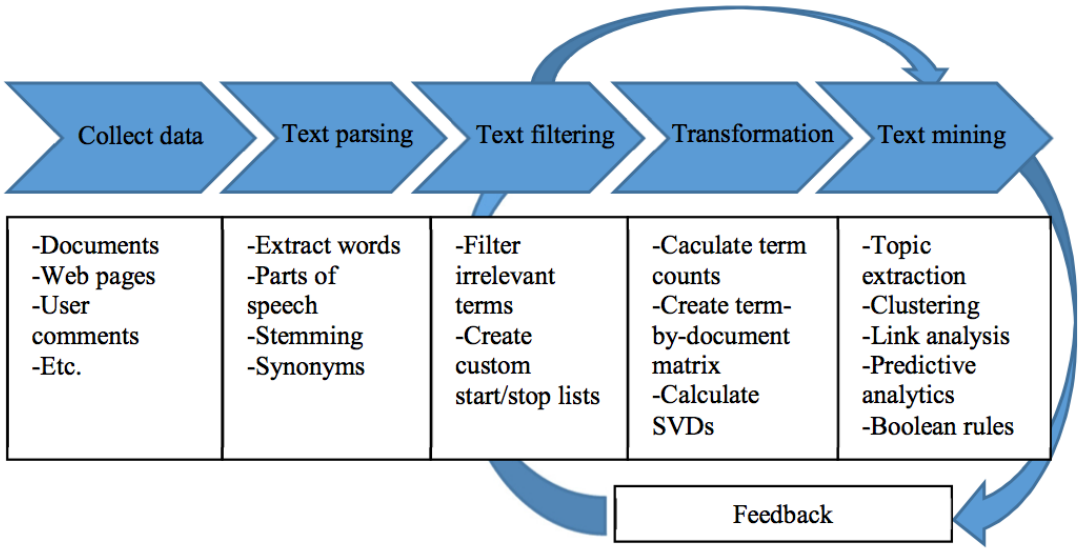
Şekil 1. CRISP-DM Prosesi

Metin dijital platformda yazılır. Mevcut metnin büyüklüğü artmaya devam etmektedir. Klasik veri madenciliği teknikleri yapısal olmayan verinin analizi için elverişli değildir. Bu nedenle bazı dilsel yaklaşımlardan faydalanmamız gerekir. Metin madenciliği teknikleri metni nümerik sayılara dönüştürerek istatistik ve makine öğrenimi dahil birçok veri madenciliği algoritmasının uygulanabilmesine hazır hale getirir. İşletmeler veri madenciliği ve metin madenciliğini rekabet üstünlüğü sağlayabilmek için müşterileri, rakipleri başta olmak üzere çevrelerini analiz etmek için kullanırlar.
Örneğin müşteri etkileşimlerinden elde edilen serbest formlu metin verisi, ürün geliştirme ve hizmet tahsislerinde girdi olan şikayet (ve ödül) alanları, garanti talepleri ve hata izlemeleri sırasında trendlerin anlaşılmasına izin verir (Feldman ve Sanger, 2007). Şekil 1 (Chakraborty, vd.,2013) metin madenciliği sürecini göstermektedir. Veri toplamadan sonraki adım doğal dil işleme algoritmalarını kullanarak veriyi nümerik indislere dönüştürmektir. Bu aşamada, veri parçalara ayrılır, cümleler belirlenir, kelimeler seçilir, gereksiz kelimeler temizlenir, kelime kökleri tespit edilir.

Bu adım varlıkları belirlemek için kelimelerin çıkarımını, dur-kelimelerin kaldırımlasını ve ve imla denetimi yapmayı içerir. Belgelerden sözcük çıkarmaya ek olarak, tarih, yazar, cinsiyet, kategori vb. gibi metinle ilişkili değişkenler elde edilir. Ayrıştırmanın ardından en önemli görev metin dönüşümüdür. Bu adım, latent semantik analiz (LSA), gizli semantik endeksleme

(LSI) ve vektör uzay modeli gibi doğrusal cebir tabanlı yöntemleri kullanarak metnin sayısal gösterimi ile ilgilidir. Bu alıştırma, bir belge terimi matrisinin (bir elektronik tablo veya metinsel verinin düz benzeri sayısal gösterimi) oluşturulmasıyla olur. Matrisin boyutları, belge sayısı ve koleksiyondaki terimlerin sayısına göre belirlenir. Bu adım, tekil değer ayrıştırması (SVD) kullanılarak belge başına matrisin boyut azaltılması içerebilir. Binlerce belgeden oluşan bir derleme sonucunda büyük olasılıkla belgeleri birbirinden ayırmakla ya da belgeleri özetlemekle alakasız birçok terim elde edilecektir. Alakasız terimleri ortadan kaldırmak için terimleri manuel olarak taramak gerekir. Bu genellikle tüm TM adımlarında en çok zaman alan ve öznel görevlerden biridir ve alan uzmanlığı gerektirir.

Terim filtrelemeye ek olarak, analizle ilgisiz dökümanlar anahtar kelimeler kullanılarak aranır. Tarih, kategori vb. gibi diğer doküman değişkenlerinden birine dayanan veya bazı terimleri içermeyen dökümanlar filtrelenir. Terim filtreleme veya doküman filtreleme, terim tabanlı doküman matrisini değiştirir. Terim tabanlı doküman matrisi terimin dökümanlarda bulunma sıklığını içerir. Belge-terimi matrisi her hücre için bir değer olarak bir doküman içindeki terimlerin varlığına göre bir dökümanda terimin ortaya çıkma sıklığını içerir. Bu sıklık matrisinden, çeşitli terim ağırlıklandırma teknikleri kullanılarak bir matris oluşturulur. TM adımı, kümeleme, sınıflandırma, ilişki analizi ve bağlantı (link) analizi gibi geleneksel veri madenciliği algoritmalarının uygulanmasını içerir. TM, farklı ayarları kullanarak analizin tekrar edilmesini ve daha iyi sonuçlar elde etmek için terimlerin dahil edilmesini veya hariç tutulmasını içeren, tekrar eden bir süreçtir. Bu adımın sonucu, doküman grupları, tek veya çok terimli başlıklar veya bir sınıflandırma probleminin cevabı olan kurallar olabilir (Chakraborty vd., 2013).



Şekil 2. Metin madenciliği prosesi (Chakraborty vd., 2013)

Bu çalışmada kullanılan döküman işleme prosesi teknikleri aşağıda verilmiştir (Ingersoll, 2013).

• Kelimelere Ayırma (Tokenization): Yapısal olmayan veriyi cümlelere, daha sonra da kelimelere parçalama işlemidir. Elde edilen belirteçleri bir dizin içine alarak tokenları elde etme işlemedir. Burada noktalama işaretlerini, sayı ve diğer sembolleri doğru, ve tutarlı işleme oldukça önemlidir.

• Küçük Harfe Dönüştürme (Downcasing): Bütün kelimeler araştırmayı kolaylaştırmak için küçük harfe dönüştürülür.

• Köklerine İndirme (Stemming): Kelimeleri eklerinden ayırarak kök kelime haline getirmek.

• Filtreleme (Stopword removal): Ve/veya gibi her türlü gereksiz kelimelerin çıkarılması işlemidir. Doğal dil işleme süreci içerisinde aslında elde edilen indekslerde filtreleme işlemleri yapılsa da son zamanlarda bazı algoritmalar probleme göre filtreleme işlemlerini ihmal etmektedir.

• Eş anlamlı Genişleme (Synonym expansion): Her bir simge için, eş anlamlılar bir eş anlamlılar listesinde aranır ve dizine eklenir. Eşanlamlılar listesindeki güncellemeler, sorguyu yeniden endekslemek zorunda kalmadan dinamik olarak hesaplanabildiğinden, genellikle dizin terimleri yerine sorgu terimlerinde yapılır. Bu çalışmada eş anlamlı genişleme kullanılmamıştır.

Metinlere yukarıda bahsedilen ön işleme algoritmaları uygulandıktan sonra elde edilen numerik veri kalıbına herhangi bir geleneksel istatistiksel ya da tahmin modeli ya da DM algoritması uygulanabilir (Chakraborty vd., 2013). Dolayısıyla modelleme aşamasına geçebilmek için yapısal olmayan veri numerik indislere dönüştürülmelidir.

3.2. Doğal dil işleme

Veri NLP teknikleri ile işlenirken TF-IDF (Term Frequency-Inverse Document Frequency; Terim Frekansı-Ters Döküman Frekansı) yöntemi kullanılmıştır. Veri doküman koleksiyonuna dönüştürülmesini sağlayan TF-IDF, bir kelimenin döküman içinde önemini ölçen nümerik bir istatistiktir. Formül 1, 2 ve 3'de gösterilmektedir.

$$tf \cdot idf(t,d) = tf(t,d) \cdot idf(t) \quad (1)$$

$$tf(t,d) = \sum_{i \in d}^{|d|} 1\{d_i = t\} \quad (2)$$

$$idf(t) = log\left(\frac{D}{\sum_{d \in D}^{D} |t \in d}\right) \quad (3)$$

TF-IDF yöntemi sonucunda çok miktarda kelime ve parametre ortaya çıkar. Bu istenmeyen durum, budama (prune) algoritması kullanılarak önlenir. Bu araştırmada, bu aşamada, budama yöntemi olarak yüzde miktar yöntemi tercih edilmiştir. Dökümanların %70'inden daha azında ortaya çıkan kelimeler budanmıştır.

NLP sürecinde veri analizine tokenization (kelimelere ayırma) ile kesiklendirerek başlanmıştır. Tokenization, metin verisini anlamlı parçalara bölme işlemidir. Harf olmayan ve harflerden oluşan veriler için ayrı ayrı tokenization işlemi uygulanmıştır. Daha sonra Porter stemming algoritması kullanılmıştır. Bu algoritma, kelimelerin köklerini elde etmeye yardımcı olur. Elde edilen dökümanlara filtreleme işlemleri uygulanmıştır. Öncelikle 2 karakterden küçük ve 25 karakterden büyük olan tüm kelimeler de dökümanlardan çıkarılması için

uzunluk filtreleme yapılmıştır. Ayrıca, dökümanlardaki tüm dur-kelimeleri (and, the, because, although, gibi) verilerden kaldırılmıştır. Önemsiz bu kelimelerin kaldırılması sonucunda, analizler daha hızlı ve kolay olacaktır. Ayrıca n-grams algoritması kelime gruplarını seçmek ve analizlere kelimeleri grup halinde dahil edebilmek için uygulanmıştır. Böylece, gruplardan oluşan bu kelimeler birlikte değerlendirilebilecektir.

3.3. Benzerlik temelli modelleme

Benzerlik temelli model elemanlara ait dilsel simgeler veya n-gram'lar gibi bir dağılım oluşturmaya ve sorgulanan örneğe en yakın dağılım ile dili belirlemek için bir benzerlik ölçüsü kullanmaya dayanır. Sınıflandırma yaklaşımı, dil için yüksek sayıda geniş örneklere ihtiyaç duyarken, benzerlik yaklaşımı tekbir dilde birleştirilmiş tüm metin gibi dil için tek bir geniş örneğe dayanır. Bu model, özünde k-en yakın komşuluk modelidir. Dil başına bir büyük metin olması bize, dilin doğru dağılımına en yakın simgelerin veya n-gram'ların dağılımının çıkarımlarını yapma imkanı verir (Hoffmann and Klinkenberg, 2014:220-221).
Benzerlik temelli yaklaşım bir dilin profili gibi n-gram karakterlerin dağılımını hesaplayarak ve sorgu metinlerinin profili ile bu profili karşılaştırmada kullanarak bir dile ait tüm cümleleri tek bir metinde birleştirir. Bu şekilde, bir dil profilinde bir modelin en özlü temsilinin o dilin modeli olduğu ve sınıflandırma sürecinin sadece benzerlik hesaplaması olduğu anlaşılır. Benzerlik temelli yaklaşım daha hızlı bir analiz, ve daha sonraki bir karşılaştırma için kalıcı bir profil sağlar. Buna ek olarak, n-gram ile profili oluşturmak yeni veri ile artan güncelleme olanağı sağlar (Hoffmann and Klinkenberg, 2014: 232).

Bu araştırmanın önişleme sürecinde veriyi seçmek ve indirgemek için benzerlik temelli modelleme kullanılmıştır. Dökümanlar arasındaki benzerlikler ortaya çıkmıştır. Araştırmada Eucliedian benzerliği uygulanmıştır. Eşitlik (4)'de verilen Cosine benzerlik ölçüsü en sık kullanılan benzerlik ölçülerindendir (Kahya-Özyirmidokuz, 2014: 9):

$$Sim(X_i, X_j) = (X_i' . X_j') = \sum_k X_{ik}' X_{jk}' \qquad (4)$$

Eşitlik (4)'de, x', x = xx 'in normalleştirilmiş vektörüdür.

Cosine ölçüsü, metin gruplandırmada, iki vektör arasındaki açının cosinesini alan popüler bir ölçüdür ve benzerliğin anlaşılmasında değişmez bir ölçü yakalar. Cosine benzerliği vektörlerin uzunluğuna dayanmaz, sadece yönüyle ilgilenir. Bu dökümanlara terimlerin aynı göreli dağılım ile işlem görmesini sağlar. Dökümanların boyutuna duyarsız olması da metinlerin analizinde çok popüler bir ölçü olmasının bir sebebidir. Ayrıca bu özelliği ile daha etkin bir önişleme için doküman vektörleri birim kürelere normalize edilebilir (Ghosh and Strehl, 2006).

4. GEREÇ VE YÖNTEM

Çalışmanın yapıldığı firma, ülkemizin önde gelen kablo firmalarından birisi olup, kablo ve tel sektöründe sağladığı başarı ve üstün performans sonucunda, bugün dünyaca tanınan uluslararası bir firma haline gelmiştir.

1974 yılında enerji kabloları üretmek üzere kurulan firma, çok hızlı bir gelişme kaydederek geçen zaman içerisinde bakır haberleşme kablosu, fiber optik kablo, enerji kablosu, yüksek

gerilim enerji kablosu, alüminyum iletken ve emaye bobin teli üretimini de kendi bünyesinde gerçekleştirerek kablo ve tel sektörünün tamamına hitap eden çok geniş bir ürün yelpazesine sahip olmuştur.

"Güven Veren Teknoloji" sloganı ile hareket eden firma, 40 yılı aşkın tecrübesi ve sunmuş olduğu yüksek kaliteli ürünler ile firma markasını uluslararası saygın bir marka haline getirmiştir.

Firma her türlü toplantılarını metin formatında kaydetmektedir. 2009-2015 her sene sonunda yapılmış 57 adet toplantıya ait tutanaktan tahminleme yapılmaktadır.Yöntem olarak, dökümanlara metin madenciliği süreci içerisinde, otomatik doğal dil işleme teknikleri kullanılmıştır. Otomatik analizlerin yanı sıra elde edilen yapısal veri tarafımızdan elle konrol edilerek yeniden gözden geçirilmiştir. K-ortalamalar ve benzerlik temelli modelleme araçları, sosyal ağ grafikleri verilere uygulanmıştır.

Firmadan temin edilen toplantı dökümanları doğal dil işleme algoritmalarıyla ön işlemesi tamamlanarak modelleme aşamasına hazır hale getirebilmek için vektör matrislere dönüştürülmüş yapısal olmayan verilerden nümerik sonuçlar otomatik olarak elde edilmiştir.

Ön işleme aşamasında öncelikle dökümanlar içerisindeki veriler ayrı ayrı tokenization işlemi ile parçalanmıştır. Sonra tüm kelimeler büyük harfe dönüştürülmüştür. Türkçe dökümanları işleyebilmek için Snowball'un Turkish stemming algoritması kullanılmıştır. Böylece, kelimelerin kökleri elde edilmiştir. 220 adet Türkçe dur kelimesi dökümanlardan çıkarılmıştır. Bu çalışmalar için PolyAnalist ve RapidMiner programı içerisindeki ilgili algoritmalar kullanılmıştır.

5. BULGULAR

Dökümanların TF-IDF kullanılarak doğal dil işleme algoritmaları uygulanması sonucunda dökümanlara ait ortak kelime matrisinin ilk 25 kelimesi Tablo 1'de gösterilmektedir.

Tablo 1. Ortak kelime matrisi

	word	attribute name	total occurences
1	üret	üret	297
2	hedef	hedef	218
3	kablo	kablo	203
4	bak	bak	175
5	iç	iç	169
6	yap	yap	167
7	yetersiz	yetersiz	166
8	gerçek	gerçek	143
9	enerji	enerji	139
10	çal	çal	132
11	yönet	yönet	129

12	saat	saat	122
13	tesis	tesis	112
14	çevre	çevre	109
15	güven	güven	108
16	içeri	içeri	102
17	genel	genel	100
18	oran	oran	99
19	süre	süre	96
20	hes	hes	95
21	miktar	miktar	90
22	ol	ol	88
23	sistem	sistem	86
24	şikayet	şikayet	84
25	haber	haber	83

Tablo 2. Kelime sıklığı tablosu

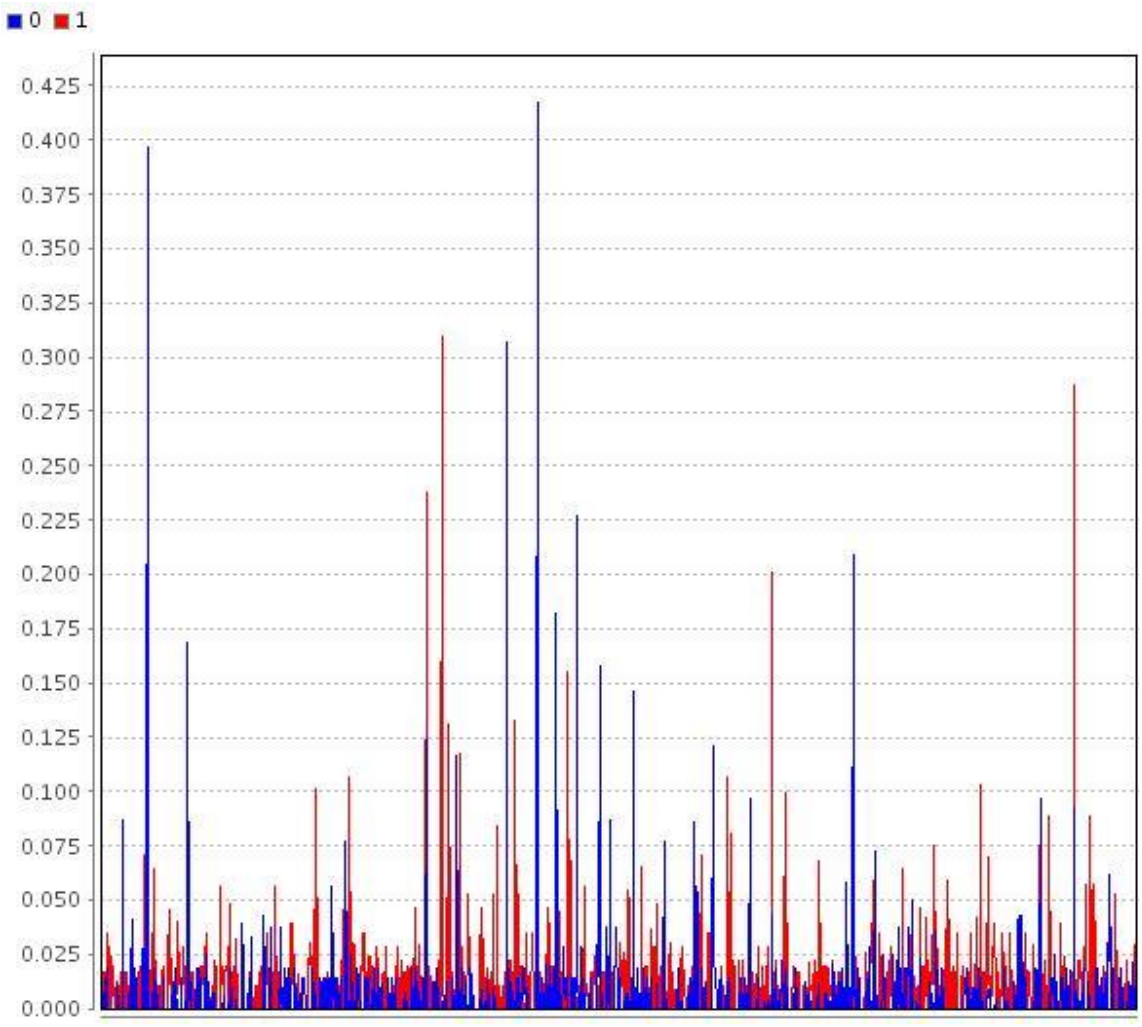
Rule name	Rec Count	%	Description		
kapasite	3	42.86	** unknown word **	404937	1
memnunum	1	14.29	** unknown word **	405145	1
Şirketimden	1	14.29	** unknown word **	405144	1
Şehir	1	14.29	** unknown word **	405134	1
raporla	1	14.29	** unknown word **	405133	1
akredite	1	14.29	** unknown word **	405132	1
mevcuttur	1	14.29	** unknown word **	405118	1
besleyen	1	14.29	** unknown word **	405117	1
hava	1	14.29	** unknown word **	405116	1

enerjide	1	14.29	** unknown word **	405115	1
kondensto	1	14.29	** unknown word **	405106	1
Ortalama	1	14.29	** unknown word **	405103	1
ilave	1	14.29	** unknown word **	405102	1
kondenstop	1	14.29	** unknown word **	405101	1
kwh	1	14.29	** unknown word **	405097	1
verimi	1	14.29	** unknown word **	405095	1
yanma	1	14.29	** unknown word **	405094	1
projesinde	1	14.29	** unknown word **	405093	1
Şikayetinden	1	14.29	** unknown word **	405052	1
sevk	1	14.29	** unknown word **	405046	1
barkoda	1	14.29	** unknown word **	405045	1
barkodu	1	14.29	** unknown word **	405043	1
Sistemde	1	14.29	** unknown word **	405042	1
Ambalajlama	1	14.29	** unknown word **	405040	1
Etiketlerdeki	1	14.29	** unknown word **	405038	1

Metrajlama	1	14.29	** unknown word **	405035	1
sorunlar	1	14.29	** unknown word **	405034	1
Malzemeden	1	14.29	** unknown word **	405033	1
8 TOP	1	14.29	** unknown word **	405032	1
bilgilendirmelerin	1	14.29	** unknown word **	405027	1
panolar	1	14.29	** unknown word **	405025	1
ilan	1	14.29	** unknown word **	405024	1
Politika	1	14.29	** unknown word **	405007	1
Ohsas	1	14.29	** unknown word **	405005	1
enetegre	1	14.29	** unknown word **	405003	1
sahip	1	14.29	** unknown word **	404999	1
belgelendirmesi	1	14.29	** unknown word **	404998	1
denetiminde	1	14.29	** unknown word **	404997	1
izoleli	1	14.29	** unknown word **	404984	1
izoleleli	1	14.29	** unknown word **	404983	1
Tambur	1	14.29	** unknown word **	404981	1

makinelerinin	1	14.29	** unknown word **	404980	1
kafes	1	14.29	** unknown word **	404979	1
telli	1	14.29	** unknown word **	404978	1
kapasiteye	1	14.29	** unknown word **	404914	1

Her kelimeye ait istatistiksel hesaplar da otomatik olarak yapılabilmektedir. Dökümanlara k-en yakın komşuluk gruplandırma uygulanmıştır. Dökümanlar 2 gruba ayrılmıştır. Bir grupta 25, diğerinde 32 doküman vardır. Dökümanlara ait gruplandırma grafiği Şekil'de verilmektedir. Gruplar incelendiğinde, aynı gruptaki dökümanlardan benzer kararlar çıktığı kolaylıkla anlaşılabilmektedir.



Şekil 3. Clustering sonucu elde edilen plot grafik

Dökümanlar, otomatik olarak içerik olarak 2 gruba ayrılsa da, dökümanlarda yazılı kelimeler bazında dökümanlar yeniden gruplandırılarak benzer kelimeler gruplara ayrılması sağlanmıştır. Doğal dil işlemesi algoritmaları uygulanan dökümanların önişlemesi tamamlanarak dökümanlara ait elde edilen ortak kelime vektörü elde edilmiştir. Daha sonra kelime vektörlerünün benzerlik modelleme ile gruplandırılarak 5 adet anahtar grup başka bir ifade ile tema elde edilmiştir.

Elde edilen bu temalar şu şekilde isimlendirilmiştir:
1. Metrajlama
2. Enerji
3. Kapasite
4. Lojistik
5. Süreç

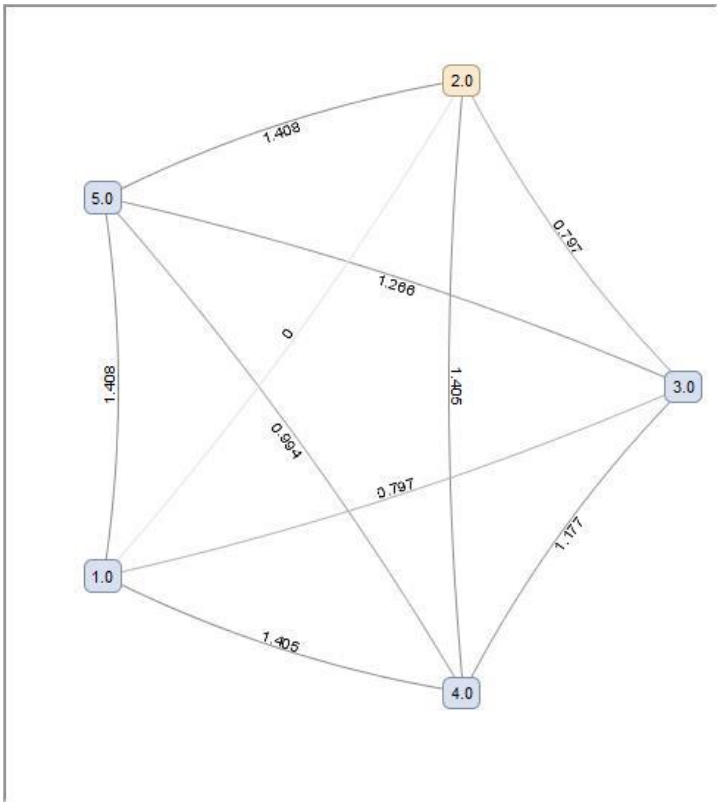
Bu gruplar/temalar arasındaki ilişkileri tespit etmek için ağ grafikleri kullanılmıştır. Elde edilen temalar arası mesafeler benzerlik analizleri ölçülmüştür. Cosine benzerlik analizlerinde Mixed Measures ölçümleri ve mixed Eucledian Distance ölçüm parametresi kullanılmıştır.

Aşağıdaki Tablo 3'de bu temaların birbirlerine benzerlik oranları verilmektedir.

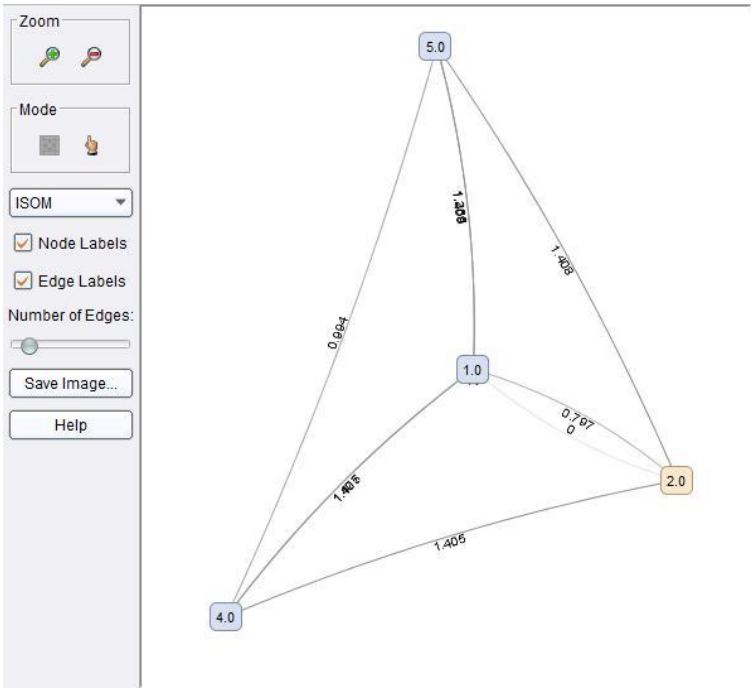
Tablo 3. Kelimelerin benzerliğini ifade eden uzaklık değerleri

first	second	similarity distance
1	2	1.47E+00
1	3	0,796815
1	4	1,405495
1	5	1,408278
2	3	0,796815
2	4	1,405495
2	5	1,408278
3	4	1,176727
3	5	1,266015
4	5	0,994268

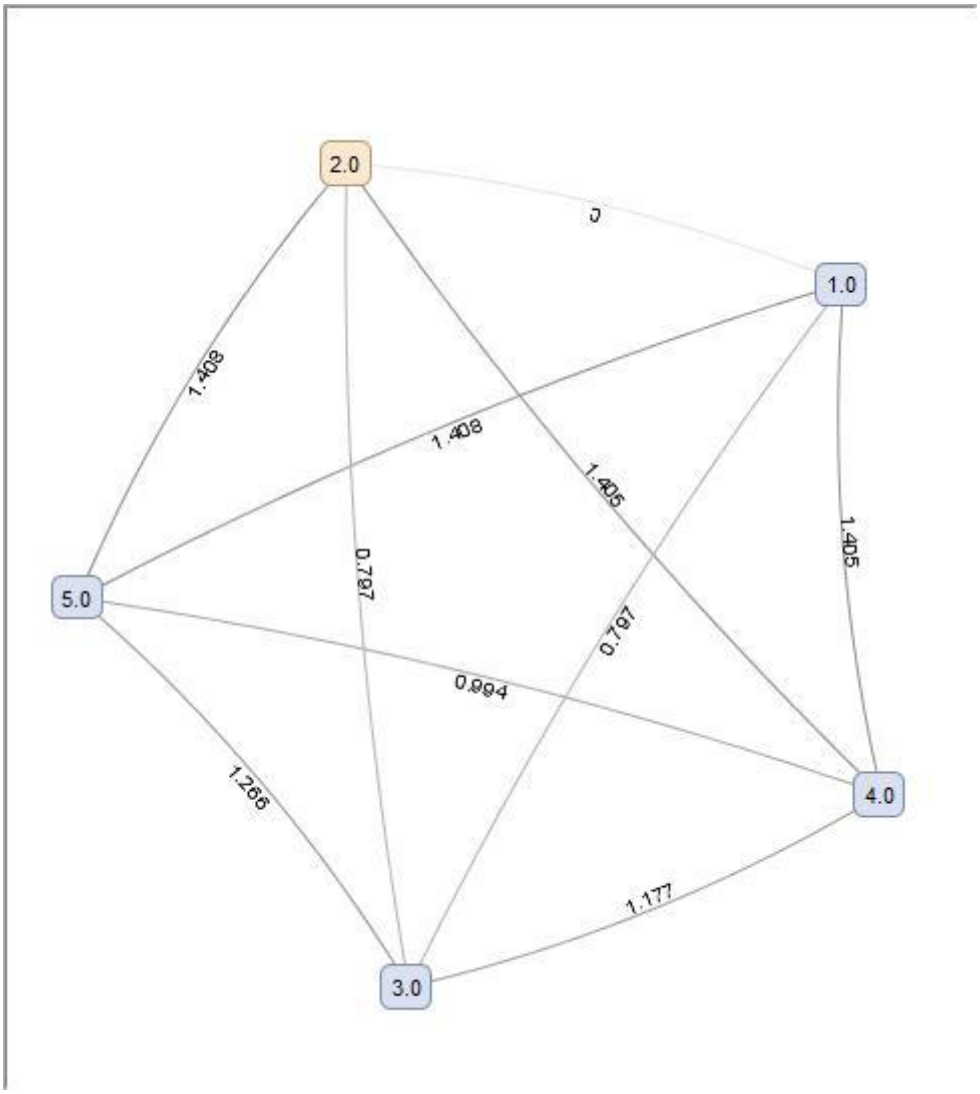
Doğal dil işleme teknikleri ile önişlemesi tamamlanan dökümanlara Cosine benzerlik ağ analizi uygulanması sonucu temalara ait sosyal ağ grafikleri elde edilmiştir. Şekillerde gösterilen ağ grafikleri birbirlerinin aynısı olup, farklı tarz gösterimleri sergilemektedir. Elde edilen ağ grafiğinden, toplantılara ait temaların birbirleriyle olan ilişkisi sayısal olarak ifade edilmiştir. Şekillerden de anlaşılacağı gibi, 1 nolu tema ve 2 nolu tema arasında en az benzerlik vardır. 2 nolu ve 4 nolu temalar ise yüksek benzerlik gösterirler. Bu temalar değerlendirildiğinde ve bu tema başlıkları altına düşen kelimeler neredeyse aynıdır.



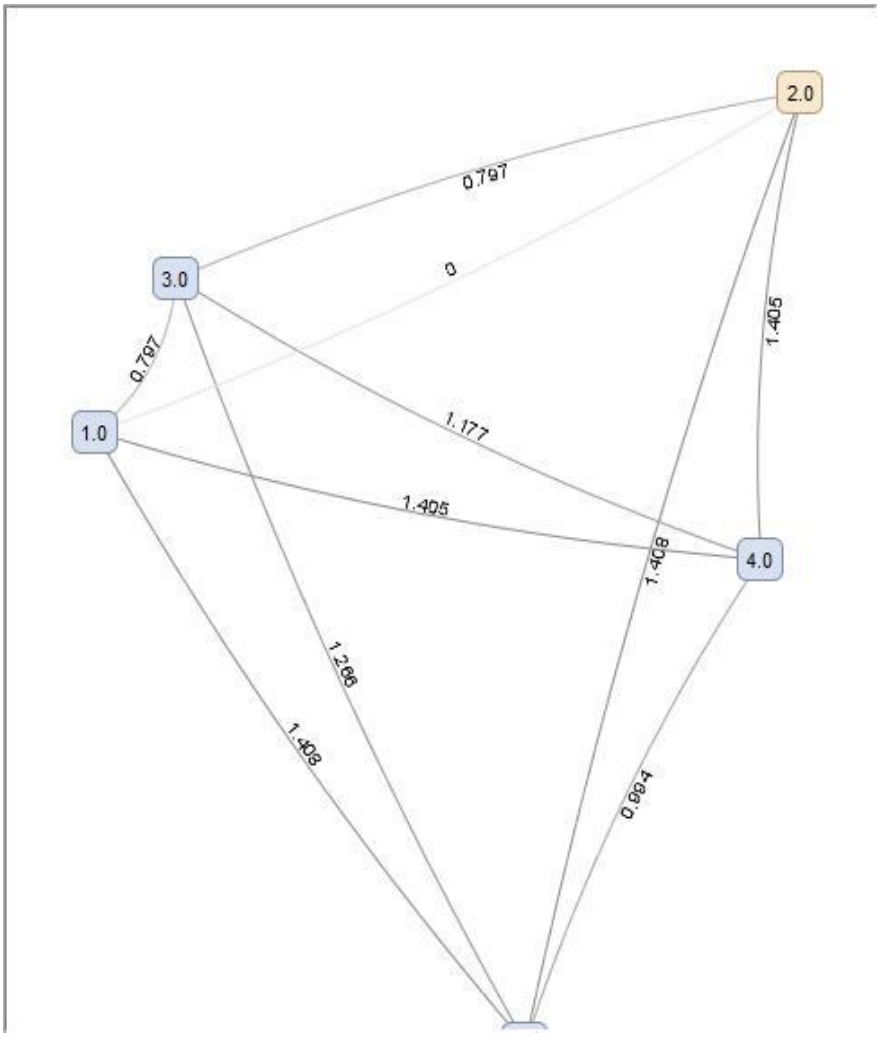
Şekil 4. Kelimelerin ilişkilerini gösteren uzaklık değerleri



Şekil 5. Kelimelerin ilişkilerini gösteren uzaklık değerleri

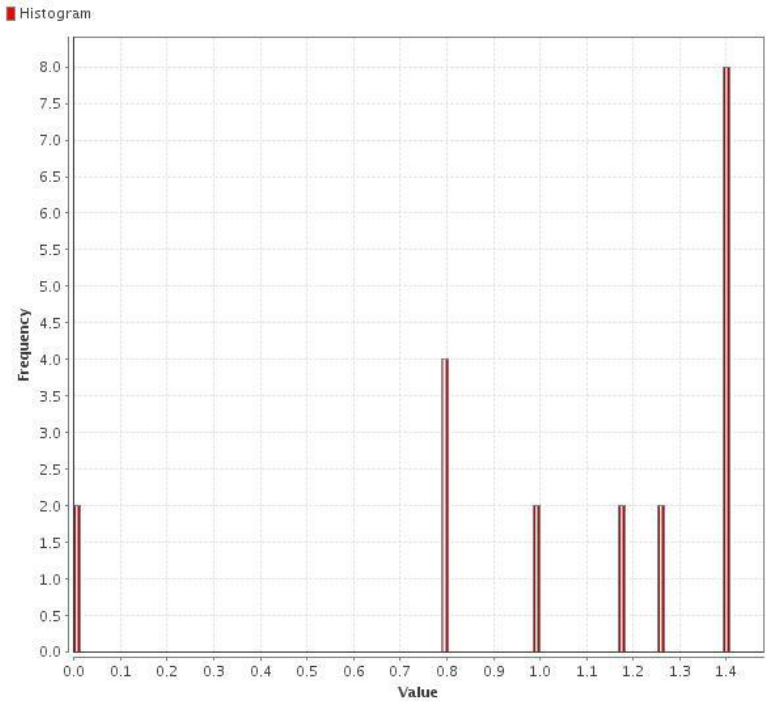


Şekil 6. Kelimelerin ilişkilerini gösteren uzaklık değerleri



Şekil 7. Kelimelerin ilişkilerini gösteren uzaklık değerleri

Aşağıda Şekil'de gruplara ait benzerlik histogramı verilmektedir. Temaların sıklık değişimi görülmektedir.



Şekil 8. Temalara ait sıklık değişim histogramı

6. TARTIŞMA VE SONUÇ

Bu araştırmada, Kayseri'de bir imalat fabrikasında kalite kontrol dökümanlarından bilgi çıkarımı yapılmıştır. Yapısal olmayan 57 döküman analiz edilmiştir. Kelime vektörü ve dökümanlardan toplantı temaları elde edilmiştir (Anahtar kelime matrisi elde edilmiştir). Böylece, dökümanlar gruplandırılmıştır.

Toplantılarda tutulan tutanak dökümanlarına kelimelere ayırma (tokenization), harf dönüştürme (transform cases), dur-kelimelerini filtreleme ve kök bulma (stemming) gibi doğal dil işleme teknikleri uygulanmıştır. TD-IDF (Terim Frekansı- Ters Belge Frekansı) önişleme analiz tekniği kullanılmıştır. Böylece dökümanlar sayısal matrislere dönüştürülerek, modelleme sürecine hazır hale getirilmiştir. Modelleme aşamasında gruplandırma ve sosyal ağ analizi yapılmıştır.

Gelecekte yapılacak araştırmalarda, daha kapsamlı ve büyük miktarda veriler kullanıbilir. Tematik analiz gibi kalitatif araştırma tekniklerinden faydalanılabilir. Alternatif TM yöntemleri çalışılabilir. Makine öğrenimi ve yapay zeka teknikleri ile güçlü modeller geliştirilebilir. Elde edilen temalar, ilerde tasarlanacak bilgi sistemlerinde girdi olarak kullanılabilir.

KAYNAKLAR

Armentano M. G., Godoy D., Amandi A. A., (2013). "Followee recommendation based on text analysis of micro-blogging activity," Information Systems, vol. 38, pp. 1116-1127.

Baxter, G., Sommerville, I. (2011). "STSs: From design methods to systems engineering", Interacting with Computers. 23 (2011) 4–17.

Chang C.W., Lin C.T., Wang L.Q., (2009). "Mining the text information to optimizing the customer relationship management," Expert Systems with Applications, vol. 36, pp. 1433–1443.

Coussement K., Den Poel D. V., (2008). "Improving customer complaint management by automatic email classification using linguistic style features as predictors," Decision Support Systems, vol. 44, pp. 870–882.

Çiflikli, C., Kahya-Özyirmidokuz, E., (2010), "Implementing A Data Mining Solution For Enhancing Carpet Manufacturing Productivity", Knowledge Based Systems, 23 (8) Pp.783-788.

Çiflikli, C., Kahya-Özyirmidokuz, E., (2012), "Enhancing Product Quality Of a Process", Industrial Management and Data Systems, 112, pp.1181-1200.

Durán, O., Rodriguez, N., Consalter, L.A., (2010), "Collaborative particle swarm optimization with a data mining technique for manufacturing cell design", Expert Systems with Applications, 37, pp.1563–1567.

Gamon M., (2004). "Sentiment classification on customer feedback data: Noisy data, large feature vectors, and the role of linguistic analysis," in Proc. the 20th international conference on Computational Linguistics, pp. 841-847, PA, USA: Association for Computational Linguistics Stroudsburg.

Gamon M., Aue A., Corston-Oliver S., Ringger E., (2005). "Pulse: Mining customer opinions from free text," LNCS, pp. 121-132, Heidelberg, Berlin: Springer-Verlag.

Gebus, S., Leiviska, K. (2009), "Knowledge acquisition for decision support systems on an electronic assembly line", Expert Systems with Applications, 36 (1), pp. 93-101.

Ghosh, J., Strehl, A., (2006). "Similarity-Based Text Clustering: A Comparative Study, in: Grouping Multidimensional Data: Recent Advances in Clustering", Jacob Kogan, Charles Nicholas, Marc Teboulle (Eds.), Springer-Verlag Berlin Heidelberg, pp. 73-98.

Gopal R. D., Marsden J. R., Vanthienen J., (2011). "Information mining - Reflections on recent advancements and the road ahead in data, text, and media mining," Decision Support Systems, vol. 51, pp. 727–731.

Hao Z.G., (2012). "A new text clustering method based on KSEP," Journal of Software, vol. 7, no. 6, pp. 1421-1425.

Harding, J.A., Shahbaz, M., Srinivas, Kusiak, A., (2006), "Data mining in manufacturing: A review", Journal of Manufacturing Science and Engineering, Manufacturing Engineering

Division of Asme 128, pp. 969- 976.

He W., Zha S., Li L., (2013). “Social media competitive analysis and TM: A case study in the pizza industry,” International Journal of Information Management, vol.33, no.3, pp. 464–472.

Hotho, A., Nurnberger, A., Paaß, G., (2005). “A Brief Survey of Text Mining. LDV Forum – GLDV”, Journal for Computational Linguistics and Language Technology 20(1), 19-62.

Hussain, S. F., Suryani, A., (2015). “On retrieving intelligently plagiarized documents using semantic similarity”, Engineering Applications of Artificial Intelligence 45, 246–258.

Ittoo, A., Bouma G., (2013), “Term extraction from sparse, ungrammatical domain-specific documents”, Expert Systems with Applications, 40, pp.2530–2540

Ittoo A. R., Zhang Y. R., Jiao J., (2006). “A TM based recommendation system for customer decision making in online product customization,” in Proc. International Conference on Management of innovation and technology, vol. 1, pp. 473-477, Singapore, China: IEEE.

Kahya-ÖzyiRmiDokuz, E., (2014), “Analyzing Social Network Unstructured Data”, Information Development, doi: 10.1177/0266666914528523.

Kahya Özyirmidokuz, E., Özyirmidokuz M. H., (2014) “Analyzing Customer Complaints : A Web Text Mining Application", in International Conference on Education and Social Sciences (INTCESS14), Ferit USLU (Ed.), İstanbul, 3-5 February 2014, pp.734-743.

Kahya Özyirmidokuz E., (2014). “Analyzing unstructured facebook social network data through web TM: A study of online shopping firms in Turkey,” Information Development, pp. 1–12, 2014.

Kahya Özyirmidokuz E., (2016). "Analyzing unstructured Facebook social network data through web text mining: A study of online shopping firms in Turkey", Information Development, vol.32, pp.70-80.

Kang, P., Lee H., Cho S., Kim, D., Park, J., Park, J. K., Doh, S., (2009), “A virtual metrology system for semiconductor manufacturing”, Expert Systems with Applications, 36, pp.12554–12561.

Kim, S. H., Park, S., Sun, M. R., Lee, J. H., (2016). “A Study of Smart Beacon-based Meeting, Incentive Trip, Convention, Exhibition and Event (MICE) Services Using Big Data”, Procedia Computer Science 91, 761 – 768.

Kumar, S., Nassehi, A., Newman S.T., Tiwari, M. K., (2007), “Process control in CNC manufacturing for discrete components: A STEP-NC compliant framework”, Robotics and Computer Integrated Manufacturing, 23, pp.667-676.

Kusiak, A., Smith M., (2007), “Data mining in design of products and production systems”, Annual Reviews in Control, 31, pp.147–156.

Larose, D. T., (2005), Discovering Knowledge in Data: An Introduction to Data Mining, USA: Wiley.

Li, D.C., Yeh C.W., (2008), "A non-parametric learning algorithm for small manufacturing data sets", Expert Systems with Applications, 34, pp.391– 398.

Liao, S. H., Chu, P. H., Hsiao, P. Y., (2012), "Data mining techniques and applications – A decade review from 2000 to 2011", Expert Systems with Applications, 39, pp.11303–11311.

Lindeblad, P. A., Voytenko, Y., Mont, O., Arnfalk, P., (2016). "Organizational effects of virtual meetings," Journal of Cleaner Production 123, 113-123.

Liu, Y., Lu, W. F., Loh, H. T., (2006), "A Framework of information and knowledge management for product design and development: A text mining approach", Information Control Problems in Manufacturing IFAC 12th, in INCOM 2006, Information control problems in manufacturing, pp. 635-640.

Mcafee, A., Brynjolfsson, E., (2012), "Büyük veri: Yönetim devrimi", Harvard Business Review Türkiye, 10, ss. 70-77.

Negahban, A., Smith, J.S., (2014), "Simulation for manufacturing system design and operation: Literature review and analysis", Journal of Manufacturing Systems, 33 (2), pp.241–261.

Onishi H., Manchanda P., (2012). "Marketing activity, blogging and sales," Intern. J. of Research in Marketing, vol. 29, pp. 221–234.

Ordenes F. V., Theodoulidis B., Burton J., Gruber T., Zaki M., (2014). "Analyzing customer experience feedback using TM: A linguistics-based approach," Journal of Service Research, pp. 1-18.

Özyurt Ö., Köse C., (2010). "Chatmining: Automatically determination of chat conversations' topic in Turkish text based chat mediums," Expert Systems with Applications, vol. 37, pp. 8705–8710.

Rose, S., Engel, D., Cramer, N., Cowley, W., (2010), "Automatic keyword extraction from individual documents", in: M.W. BERRY and J. KOGAN (Ed.), TM: Applications and Theory, Wiley, p.3-19.

Stray, V., Dag I.K., Sjøberg, T. D., (2016). "The daily stand-up meeting: A grounded theory study", The Journal of Systems and Software 114, 101–124.

Sumathi, S., Sivanandam, S.N. (2006). Introduction to DM and its Applications, Verlag Berlin Heidelberg: Springer.

Sunikka A., Bragge J., (2012). "Applying text-mining to personalization and customization research literature – Who, what and where?" Expert Systems with Applications, vol. 39, pp. 10049–10058.

Tanmay B., Murthy, C.A., (2015). "A similarity assessment technique for effective grouping of documents", Information Sciences 311, 149–162.

Thorleuchter D., Den Poel D. V., Prinzie A., (2010). "Mining ideas from textual information," Expert Systems with Applications, vol. 37, pp. 7182–7188.

Thorleuchter D., DenPoel D.V., (2012). “Predictinge-commercecompany success by mining the text of its publicly-accessible website,” Expert Systems with Applications, vol. 39, pp. 13026–13034.

Thorleuchter D., Den Poel D. V., Prinzie A., (2012). “Analyzing existing customers’ websites to improve the customer acquisition process as well as the profitability prediction in B-to-B marketing,” Expert Systems with Applications, vol. 39, pp. 2597–2605.

Thorleuchter, D., Van Den Poel, D., (2014), “Semantic compared cross impact analysis”, Expert Systems with Applications 41, pp. 3477– 3483.

Tsai S., Kwee A. T., (2011). “Database optimization for novelty mining of business blogs,” Expert Systems with Applications, vol. 38, pp. 11040–11047.

Tuffery, S. (2011), DM and Statistics for Decision Making. Wiley

Ur-Rahman N., Harding J. A., (2012). “Textual DM for industrial knowledge management and text classification: A business oriented approach,” Expert Systems with Applications, vol. 39, pp. 4729–4739.

Wang, K. (2007), “Applying data mining to manufacturing: the nature and implications”, Journal of Intelligent Manufacturing, 18 pp.487–495.

Weng S.S., Liu C.K., (2004). “Using text classification and multiple concepts to answer e-mails,” Expert Systems with Applications, vol. 26, pp. 529–543.

Zhang, H., Chow, T.W.S., (2012). “A multi-level matching method with hybrid similarity for document retrieval”, Expert Systems with Applications 39, 2710–2719.