

NoSQL VERİ TABANLARI ÜZERİNDE BİR METİN MADENCİLİĞİ UYGULAMASI

Metin Zontul^[*], Gürkan Aydın^[**]

Özet

Metin madenciliği metni veri kaynağı olarak dikkate alan veri madenciliği çalışmasıdır. Metin madenciliği düzensiz metinlerden oluştuğundan veri madenciliği algoritmaları ile yaklaşmak mümkündür. Asıl fark düzensiz metinleri anlamlı bir düzene oturtmaktır. Bu sebeple belirli bir düzene yaklaştırılan metinler üzerinde veri madenciliği yapmak kolaylaşmaktadır. Zor olan ise kaynaklar toplayıp derlemek ve anlamlı sayısal ifadelere çevirebilmektir. Projedeki amaç müşteri şikâyetlerini anlamlandırıp ölçümlemektir. Projede GSM firmaları dikkate alınmıştır. Sonraki süreçte ise hedef öğrenci yazılılarını okumaktır. Öğrencilere sınavlarına uygulanacak metin madenciliği çalışması ise farklı bir uzmanlık gerektirmektedir. Sözlük yaklaşımında olduğu gibi her soru için kategori ve kategori altındaki ağırlıklı kelimeler bu uzmanlar tarafından belirlenmelidir. Doğru sözlük tablosu oluşturulduğunda yazılım hızla değerlendirmesini yapacaktır. Milyonlarca öğrencinin cevap kâğıdının okunduğunu düşünürsek yazılımın milyonlarca işlem karşısında yorgunluk ve dikkatsizlik gibi davranışlarda bulunması beklenemez. Sonuç olarak milyonlarca iş yükünün altında bile metin madenciliği yaklaşım projesinin eşit, adil ve hızlı bir değerlendirme yapacağı aşikârdır.

Anahtar Kelimeler: Metin Madenciliği, Joker Yöntemi, Sözlük Yöntemi, Gövdeleme, MongoDB, NoSQL

A Text Mining Application on NoSQL Databases

Abstract

Text mining is a data mining operation that considers text as a data source. Since text mining is composed of irregular texts, it is possible to approach it with data mining algorithms. The main difference is to place irregular text on a meaningful level. For this reason, it is easy to make data mining on a certain level of approximated text. The hard one is to collect resources and translate meaningful digital expressions. The purpose of the project is to understand and measure customer complaints. GSM companies are considered in the project. The next step is to read the target student's writings. The study of text mining that will be applied to the students' exams requires different expertise. As with the dictionary approach, the weighted words under the category and category for each question must be determined by these experts. When the correct dictionary table is created, the software will make a quick evaluation. If we think that millions of learners read the answer paper, it can not be expected that the software will be found in behaviors like fatigue and carelessness in the face of millions of transactions. As a result, even under millions of workloads, the text mining approach project is likely to make an equitable, fair and rapid assessment.

Keywords: Text Mining, Joker Method, Dictionary Method, Mocking, MongoDB, NoSQL.

[*] İstanbul Aydın Üniversitesi, Yazılım Mühendisliği, İstanbul, metinzontul@aydin.edu.tr

[**] İstanbul Aydın Üniversitesi, Bilgisayar Mühendisliği (YL), İstanbul, gurkan@gurkanaydin.com

1. GİRİŞ

Mobil; Fransızcadan dilimize katılan bir terimdir. Fransızcada “hareketli, hareket eden” anlamına gelmektedir. Sıfat olarak hareketli, taşınabilir anlamına gelir (TDK, 2016). Zamanımızda ise Mobil terimi “Hareketli iletişimi” temsil etmektedir.

Bilgi çağında teknolojinin de ilerlemesiyle giyilebilir giysi ya da cihazlarda üretilmeye başlamıştır. Bununla beraber televizyonlar, telefonlar, saatler, dijital bileklikler, ayakkabılar dahi sanal dünyaya anlık olarak veri gönderebilmektedirler. Bu kadar teknolojik hareketlerde farklı cihazlarla farklı uygulamaların doğmasına sebep olmaktadır. Bu cihazların ortak değeri ise veridir. Tüm bu cihazlar sosyal paylaşım sitelerine otomatik olarak veri yüklemek suretiyle büyük verinin oluşmasına katkıda bulunmaktadırlar. Dolayısıyla, sosyal medyada yer alan herhangi bir konu üzerinde metin madenciliğinin uygulanması mümkün olabilmektedir.

Dünyadaki insanların yaklaşık yarısı internet kullanmaktadır (Digital in 2016, 2016). Bu insanlar ya mobil cihazlar üzerinden ya da online web siteleri üzerinde anlık veri girmekte ve internet üzerinde bir saniyelik zaman diliminde terabyte’lar düzeyinde büyük veriyi oluşturmaktadırlar. Oluşan bu veri üzerinde gizli bilginin mantıksal bir zaman diliminde açığa çıkartılabilmesi için akıllı analiz algoritmalarının kullanılması gerekmektedir. Bu algoritmalar veri madenciliği ya da metin madenciliği başlığı altında incelenebilirler. Veri madenciliği, bilginin gizli olduğu veri kaynaklarından faydalı olan bilginin elde edilmesidir. Bu da; kümeleme, veri özetleme, analiz gibi yöntemleri kapsamaktadır.

Büyük veri dediği zaman Veri Madenciliği akla gelmektedir. William Frawley ve Gregory Piatetsky-Shapiro tarafından “önceden bilinmeyen, fakat yarar sağlama potansiyeli yüksek olan verinin keşfedilmesi” dir. M. Berry ve G. Linoff ise “Anlamlı kuralların ve örüntülerin bulunması için geniş veri yığınları üzerine yapılan keşif ve analiz işlemleridir” şeklinde tanımlamışlardır. Veri madenciliği bilişim literatüründe; veri tabanlarında bilgi madenciliği (knowledge mining from databases), bilgi çıkarımı (knowledge extraction), veri ve örüntü analizi (data/pattern analysis), veri arkeolojisi gibi kavramlarla da açıklanmaya çalışılmıştır (Özcan, 2014).

Karar alma mekânizmalarının doğru tercihler yapması adına bu verilerin basitleştirilmesi, anlamlandırılması ya da görselleştirilmesi veri madenciliğinin kapsamına girmektedir. Veri madeni raporları doğrultusunda stratejik karar mekânizmaları bilimsel çıkarım yapabilir ya da bilimsel tercihlerde bulunabilirler. Büyük veri, yapılandırılmış veya yapılandırılmamış veri şeklinde olabilir. Yapılandırılmış verilerde her satırdaki sütun sayısı sabitken yapılandırılmamış verilerde her bir satırda farklı sayıda sütun olabilir. Veri madenciliği genellikle yapılandırılmış verilerle çalışırken metin madenciliği ise yapılandırılmamış verilerle uğraşır. Büyük veri üzerinde metin madenciliği uygulaması geliştirebilmek; Seçim, Aktarım, Barınma, İşlem ve Çıktı aşamalarından oluşmaktadır. Bu adımların gerçekleştirilmesi adına birçok uzmanlık alanına ihtiyaç duyulmaktadır:

Şirketlerin vizyon analizi yapılan metin madenciliği çalışmasında kaynak olarak 2010 Yılı Capital dergisinde bulunan Türkiye’nin En Beğenilen Şirketleri araştırması kaynak olarak kullanılmıştır. Ayrıca kaynak olarak yöneticilerle uygulanan online anket de kullanılmıştır. Çalışmada geliştirme teknolojileri olarak

Visual Basic ve Statistica programı kullanılmıştır. MELEK çalışmasında "Tekil değer ayrışımı özellik çıkarımı" ve anlam endeksini oluşturabilmek için Statistica programının analizini kullanmıştır (Melek, 2012). Sosyal medya üzerinde duygu analizi ve pazar değerlendirme çalışmasında yazılımsal teknoloji için Botego firmasının destek aldıklarını beyan ederler. Sosyal medya da kaynak olarak Twitter seçilmiştir. Kendi ifadesi ile "Twitter verileri, yaratmış oldukları yazılımın alt yapısını oluşturan makine öğrenmesi temeline dayalı duygu analizi ile pozitif, negatif ve nötr olarak sınıflandırılmıştır." Şeklinde açıklamıştır. Metin madenciliği yaklaşımı olarak Kümeleme Yöntemi kullanılmıştır. İnternet, spor alış veriş, romantik, reklam ve diğerleri başlıkları altında kümeleri toplanmıştır (Beyhan, 2014).

Sosyal medya üzerinde yapılan bir metin madenciliği çalışmasında, fast-food hizmeti veren ünlü zincir markaları müşteri şikâyetlerine göre hijyen, hız, lezzet ve memnuniyet kalitesi gibi kategorilere ayrılmıştır. İlgili çalışmada sözlük algoritması ve joker yaklaşımı kullanılarak firmalar hakkındaki şikâyetler kategorilere ayrılarak derecelendirilmiştir. Sözlükte yer alan kelimeler negatif (-1) veya pozitif (+1) olarak değerlendirilmiştir. Çalışmada geliştirilen yazılım, Entity Framework, WPF ve MSSQL teknolojileri kullanılarak C# programlama dili ile yazılmıştır (Aravi, 2014). Müşteri memnuniyeti çalışmasında ise oluşturulan veri tabanına verilerin yazılan bir program aracılığıyla girilmesi sağlanmıştır. Metin madenciliği kaynağı olan veriler düzensiz yapıda olduğundan elle giriş sırasında kullanıcı tarafından mutlu ya da mutsuz diye ayrıştırılmaktadır. Oluşturulan girdi verisinin toplam kelimesi ile yargıya varılmıştır. Ayrıca bir ikinci yöntem olarak Karşılık Matrisini kullanarak sonuca ulaşılmıştır (Kuzucu, 2015).

Sosyal medyada bulunan metin ve fotoğraf eşleşmesinin incelenmesi üzerine metin madenciliği yapılmıştır. Sosyal medyalarda bulunan fotoğraf ve fotoğrafla ilgili metnin bağının incelenmesi üzerine öncelikle fotoğrafların Pixel bazında her hücresi incelenerek renk ağırlığı tespit edilmiştir. Tespit sonrası veri analiz programı olarak Knime kullanılmıştır. Metin madenciliğinde ise Kümeleme Yöntemi kullanılmıştır. Program üzerinden elde edilen analizlerle renk uzayı ve metin madenciliğinin sonuçları birlikte yorumlanmıştır (Demirel, 2015).

Talep tanıma ve yönlendirme çalışmasında ise dilekçe tanıma ve anlamlandırma konusu üzerinde çalışılmıştır. Dilekçeleri önce optik karakter tanıma yöntemi ile karakter taraması sırasında eksik ya da yanlış kelime var ise burada da Levenshtein algoritmasından faydalanılarak oluşan veriler veri tabanına gönderilmektedir. Naive Bayes Multinomial kullanılarak dilekçe sınıflandırılması yapılmıştır. Sınıflandırılan dilekçeler veri tabanı üzerinde kayıt kişilere belge talep, ders kayıt, harç, sınav ve staj gibi kategorilere ayrıştırılarak ilgililere aktarılması sağlanmıştır. (Sancar, 2016).

Bu çalışmasında, sikayetvar.com üzerinden GSM firmaları hakkında alınan şikâyetlerin kategorilere ayrılarak derecelendirilmesi metin madenciliği ile yapılmıştır. Bu çalışmada da sözlük algoritması ve joker yaklaşımı kullanılmakla beraber sözlükte kullanılan kelimelere (0,1) aralığında ağırlıklar verilmiştir. Ayrıca, geliştirilen yazılımın performansını artırmak için class yapısı C# dilinde elle oluşturulmuş ve ilişkisel veri tabanı yerine NoSQL veri tabanı tercih edilmiştir. Oluşturulan yazılım mimarisi bulut mimarisi ile uyumlu olup ileride web servis şeklinde hizmet vermesi düşünülmektedir.

2. METİN MADENCİLİĞİ NEDİR?

Metin madenciliği, metin kaynaklarını girdi olarak kullanır ve bu metinler üzerinden anlamlı çıktılar elde etmeyi amaçlar. Metin madenciliğine kaynak olarak internet çevrimiçi ulaşılması kolay bir veri kaynağıdır. Veri tabanına aktarımı zor olan lokal veri ya da kağıt üzerinde bulunabilirler. Örneğin fatura, hesap belgesi, mektup, mail, epikriz gibi. Bu dokümanları toplamak, derlemek ve veri tabanına doğru bir şekilde aktarmak çevrimiçi sistemlere göre daha fazla zaman alabilirler.

Metin madenciliğinde amaca hizmet eden algoritmalar seçilerek veri tabanı incelemesi yapılır. Günümüzde internet kullanımının yaygınlaşmasıyla beraber bireylerinde sosyal mecralar üzerinden kişisel paylaşımları oldukça artmıştır. Buda internet dünyasını metin madenciliğinin merkezine oturtmaktadır.

Metin madenciliği, biçimi olmayan ifadelerin içinden bilgiler çıkarılması ve yapılandırılmamış verilerin biçimlendirilmesi sürecidir. Metin madenciliği dört adımdan oluşmaktadır (Oğuzlar, 2011);

1. Metin madeni oluşturma
2. Metin işleme
3. Veri Analizi
4. Değerlendirme ve Raporlama

Metin madenciliğinden birçok alanda yararlanılmaktadır. Bunlardan en önemlilerinden bir tanesi de Tıp'tır. Tıp'ta ki verilerin genel olarak serbest metin biçiminde olması hasta ile ilgili önemli bilgilerin gözden kaçırılmasına, bilgiye erişimin zorlaşmasına sebep olmaktadır. Sağlık sistemin başarısının, klinik dokümantasyonun serbest metin biçiminde oluşmasından dolayı bu tür sistemlere olan ihtiyacı ortaya çıkarmıştır. Yapılan klinik çalışmalar, araştırma raporları, hastane kayıtları, doktor notları ve ya faturalar tıptaki en önemli veri kaynaklarıdır (Oğuz, 2009).

Metin madenciliği, son zamanlarda daha çok sanal dünyanın sosyal etkileşiminin popülerliğinden dolayı sanal ticaret konularında daha çok rastlanmaktadır. Ürün pazarlama ve tanımlama sosyal medyalar üzerinde daha rahat geliştirilip daha rahat incelenebilmektedir. Sosyal medyalarda sayısal ya da metin verilerini toplamak, derlemek ve analiz etme süreçleri daha kısa ve düşük maliyetli olduğundan pazarlama beceri analizleri bu mecralar üzerinde daha çok yapılmaktadır. Reklam ve tüketim dünyasının ilişkisini düşündüğünde ihtiyacın karşılığı daha net görülecektir.

2.1 Hazırlık

Metin madenciliği öncesinde verilerin işlemlere uygun hale getirilmesi ve hazırlanması gerekmektedir. Ham veri üzerinden çalışılması sonuçlar üzerinde farklı etkiler oluşturacağı gibi sürecinde uzamasına sebep olacaktır. Ön hazırlık için ham veriyi temizlemekle işe başlanmalıdır. Veri tabanımızda bulunan ham verilerin madencilik işlemlerine uygun hale gelmesi ve optimum sonucun elde edilmesi adına madencilikten önce hazırlık işlemleri uygulanmalıdır. Metin ile ilgili çalışmada atılacak ilk adım işaretleme işlemidir. Ham metin verilerinde bulunan bütün tümcelerin sağlıklı bir şekilde edilmesi işlemine işaretleme denir.

Elimizdeki ham verinin daha kaliteli hale getirilmesi, veri boyutunun da küçültülerek işlem kabiliyetimizin artırılması adına metnin sadeleştirilmesi gerekmektedir. Bu sebeple her kelimeyi ayrıştırabilmemiz için toplam metni sadeleştirmek ve işaretlememiz gerekmektedir. Bu esnada metin içerisinde bulunan noktalama işaretleri, tek boşluk karakterinden fazla olan boşluklar “white space” ve diğer metine konu olmayan karakterlerin temizlenmesi şekliyle işaretleme gerçekleşir. Böylelikle metin olarak geriye kelimeler ve kelimeler arasındaki birer boşluklar halinde ki sade metin kalır.

2.2 Gövdeleme

Veri kümesinin işaretlenmesidir. İşaretleme belirlendikten sonra bu işaretleme tüm veri için uygun olacak halde standardize edilmesi gerekir. Gövdeleme aşaması her uygulamada farklı olabilir. Bu farklılıklar fayda sağlayacağı gibi işleme zarar da verebilir. Gövdeleme iki ana başlıktan oluşmaktadır (Çelikyay, 2010).

Joker Yöntemi: Türk dili sondan eklemeli yapıya sahip bir dildir. “okul” kelimesi ile “okuldan”, “okulu”, “okula”, ve “okulun” kelimeleri aynı kelime köküne farklı anlamlar yüklemektedir. Konunun temeli olan “okul” kelimesi bütünlüğünü korumaktadır (Öner, 1999), Eğer her ek almış kelimenin hallerini farklı kelimeler olarak kabul edilebilir. Kelimenin ek almış halleri işlem sürecinin uzatacağı gibi sözlük de barınacak kelime sayısının da artması ile başarı oranı düşecektir. Joker kelime, aynı köke sahip farklı ekler alan kelime öbeğidir. Bu farklı öbekleri bir çatı altında toplayan kelimeler joker kelimelerdir. Belirtilen örnek kelime “ okul* ” (kelimenin joker olduğunu * işaretinden anlıyoruz), “okul” kelimesinden sonra gelecek tüm ekler “okul” terimi olarak kabul görmüş olacaktır (Oğuzlar, 2011).

Gövdeleme Yöntemi: Gövdeleme yönteminin kuralları daha katı kurallar içermektedir. Bu katı kurallar ile joker yönteminden ayrılır. Joker yöntemine göre daha katı kurallardan oluşan bir yöntemdir. Gövdeleme kelimenin kökünü hedef almaktadır. Buda çakışmalar ile karşılaşmamıza sebep olacaktır. (Şentürk & Gündüz, 2014). Gövdeleme yöntemi sondan eklemeli olan Türkçe ‘de çekim eklerinin filtrelenerek yapımlarının korunması ile sağlanmaktadır.

2.3 Sözlük oluşturma

Veri tabanımızda bulunan tüm kelimelerin dağılım frekansını tespit eden bir program yazıldı. Bu programın analizi sonucunda en çok kullanılan kelimeler tespit edildi. Bu kelimelerden kategorilere uygun olanlar belirlenerek aşağıdaki sözlük tablosu oluşturulmuştur. Sözlük oluşturulurken kalabalık edecek gereksiz kelimelere yer verilmemesi önem arz eder. Dikkat edilmemesi durumunda çıktının sağlayacağı faydanın değeri de düşecektir.

Tablo I. Oluşturulan Sözlük Tablosu ve Ağırlık Puanları

Kapsama	Ağırlık	İnternet	Ağırlık	Fatura	Ağırlık	Hizmet	Ağırlık
Sokak	0.1	Yavaş	0.1	Tarife	0.1	Hizmet	0.15
Sinyal	0.1	Hız	0.1	Ücret	0.1	Alay	0.1
İletişim	0.1	Kısıtlama	0.05	Fiyat	0.1	Yapılmadı	0.15
Göbeği	0.3	Kopma	0.05	Abonelik	0.1	Ukala	0.1
Ev	0.1	İnternet	0.1	Fatura	0.05	Ulaşamıyorum 0.1	0.1
Çek*	0.3	GB	0.2	Faiş	0.05	Saygısızlık	0.05
Kapsam*	0.1	Kota	0.1	Kapama	0.1	Çağrı	0.1
Cayma	0.1	3G	0.1	Kredi	0.1	Umursamaz 0.1	0.05
Bölge	0.1	4.5G	0.2	TL	0.1	Kapat	0.05
				Ödeme	0.1	Davranış	0.05
				Paha	1.5	Sorun	0.05

Tablo I'de Kategori altındaki kelimeler ve ağırlık puanları görülmektedir. Örnek olarak Kapsama kategorisi altındaki "çek*" kelimesini incelersek "çekim, çekmiyor, çekmez, çekmeyen" gibi veri tabanımızda bulunan kelimelerin Joker ifadesini oluşturmaktadır. Kapsam* Joker ifadesi ise "kapsam, kapsamıyor, kapsamamakta" gibi kelimelerin joker ifadesidir.

Bir dilin veya dillerin kelime haznesini (sözvarlığını), söyleyiş ve yazılış şekilleriyle veren, kelimenin kökünü esas alarak, bunların başka unsurlarla kurdukları sözleri ve anlamlarını, değişik kullanışlarını gösteren eserlere sözlük denir. Bir başka kaynağa göre sözlük, bir dilin veya dillerin kelime haznesini (sözvarlığını), söyleyiş ve yazılış şekilleriyle veren, sözcüğün kökünü esas alarak, bunların başka unsurlarla kurdukları sözleri ve anlamlarını, değişik kullanışlarını gösteren yazılı eserdir (Boz, 2015).

Kelime sayısının niceliğinden çok niteliği önemlidir. Bu sebeple çok sayıda kelime tespit etmektense faydaya yakın kelimeleri belirlemek daha da önemlidir. Bu sebeple kelime sayısı az olan sözlüklerin işlem sürelerinin kısa olduğu gibi başarı yüzdesinin de yüksek olması gözlemlenebilir. Veri tabanından çağrılan veriler ilk önce oldukları gibi RAM'e aktarılırlar. RAM'e verileri aktararak işlem süreleri kısaltılmış olur. RAM'e yazılmış kategoriler sırası ile dönülürken bu kategoriye ait sözlük kelimeleri dikkate alınır. GSM şirketine ait her bir yorum için kategoriye ait sözlük elemanları dikkate alınarak kategori bazında puanlama yapılması sağlanır. Süreçte yoruma ait her kelime için kategoriye ait her bir sözlük kelimeyle eşleşme durumuna bakılır ve eğer eşleşme bulunursa; sözlük kelimesinin ilgili ağırlık puanı dikkate alınarak kategoride puan artışı sağlanır. Aynı yorum içerisindeki kelimeler dikkate alındığından ilgili yorum birden çok kategorinin puanlamasına etkide bulunabilir.

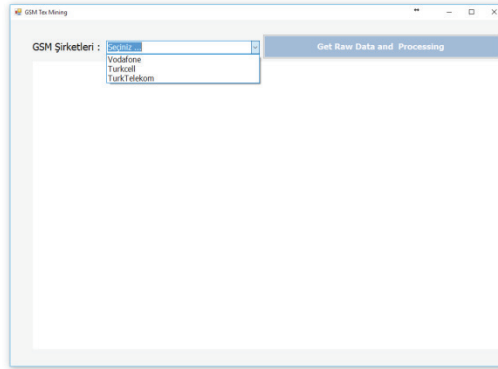
Örnek : "Vodafone'dan 24 ay taksitli Discovery Mobile GM5 Plus telefon aldım. Piyasada 1000 TL olan telefonu tarifesiyle birlikte 2400 TL'ye aldım farka bakın 1400 TL 400 TL telefon farkı ödüyorum zaten ama bir de Vodafone beni şu an piyasada diğer operatörlerinde sunduğu 20 TL'ye verilen tarifeyi 40 TL'ye verdi. 4.5G gibi bir teknolojinin olduğu dönemde 2 GB gibi küçük bir şey veriyor bana. Telefondan da memnun değilim kamerası çok kötü telefonun ya parasını verip tarifeden çıkmak istiyorum ya da tarifeyi 20 TL'ye düşürmek istiyorum Vodafone'u aradım ama 3. sıradasınız deyip 5 dakika sırada bekletip aramayı sonlandırdılar amaç arayanları yıldırmak yani hizmet değil. Koskoca şirketin büyük bir çağrı merkezinde yok bu şirket sizin onu aramanızı istemiyor zaten." Yorumunu ele alalım. Fatura, Hizmet, İnternet kategorilerimize ait sözlük kelimeleri barındırmaktadır. İlgili yorum hem Fatura hem İnternet hemde Hizmet kategorisinin puanını etkilemektedir. "çağrı", "tarife", "4.5G", "TL", "hizmet" kelimeleri farklı kategorilerinden oluşmasına karşın aynı yorum içerisinde yer almaktadır.

Fatura : Tarife, TL

Hizmet : Çağrı

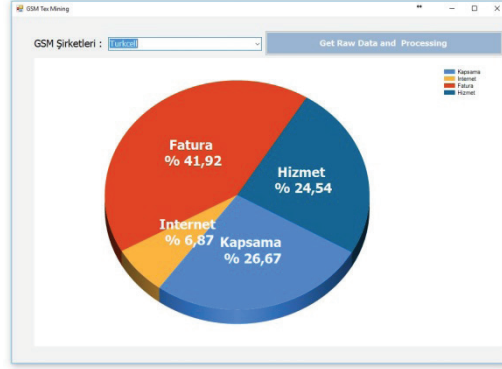
İnternet : 4.5G, GB

Yorum içerisinde geçen kelimeleri incelediğimizde yukarıda ifade edilen üç kategori altındaki sözlük kelimelerimiz görünmektedir. Örnek yorumumuz dört kategorimizden Fatura, Hizmet ve İnternet kategorilerini kapsamaktadır. Algoritmaya tabi tutulan GSM şirketine ait her yorumun analiz sonrasında RAM'e aktarılan rapor verilerini programımızda seçilen şirket adı altında sunumu yapılmaktadır.



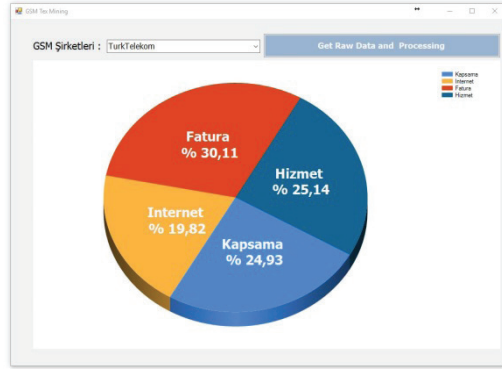
Şekil 1. Program Ara Yüzü

Şekil 1'de programın ara yüzü görülmektedir. Öncelikle tüm datanın çağırılması ve işlenmesi gerekmektedir. Bu sebeple ilk önce "Get Raw Data and Processing" butonuna basmalıyız. Bu süreçte buton adından da anlaşılacağı gibi tüm veri, veri tabanından çekilerek algoritmaya tabi tutulmaktadır. Sonuçlar RAM alınarak kullanıcı için hazır bekletilmektedir. Sonrasında GSM şirketlerinin (GSM 1,2,3) adının bulunduğu dropdownlist 'den sonucunu görmek istediğimiz şirketi seçerek programın gövdesinde PIE CHART olarak raporun ekrana yansıtılması sağlanır.

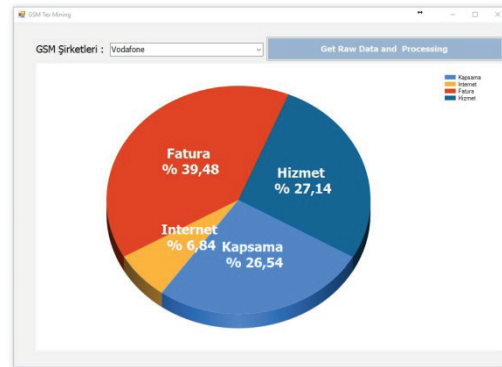


Şekil 2. Program GSM 1 Sonuçları

Şekil 2’de görüldüğü **kapsama, internet, fatura ve hizmet** kategorileri dikkate alınarak hazırlanan rapor PIE CHAR üzerinde yüzdeleri dilimlerle ekrana yansıtılmıştır. Şekil 3’de GSM 1’in aldığı yorumlar üzerinden yüzdeleri dilimleri görünmektedir.



Şekil 3. Program GSM 2 Sonuçları



Şekil 4. Program GSM 3 Sonuçları

Şekil 2, Şekil 3 ve Şekil 4 de tüm çalışmanın sonuçları görünmektedir. Bu raporlar üzerinden anlaşılacağı gibi toplam değerlendirmelerinde birbirlerinden farklılıkları görünmektedir.

3. SONUÇ

Sonuç olarak GSM şirketlerinin sunulan hizmet ve tüketici arasındaki ilişkileri üzerine yorum yapabiliriz. Firmalar bu doğrultuda pazarlama ya da iyileştirme adına stratejilerini belirleyebilirler. Turkcell ve Vodafone 'nun birbirlerine işlem kalitesi olarak yakın oldukları görünmektedir. Fakat Türk Telekom 'un müşterilerine **“internet”** kategorisinde daha çok sorun yaşattığı görülürken **“fatura”** kategorisinde ise diğer iki firmaya karşın çok daha avantajlı görünmektedir. Bu ekranlar sonucunda tüketicide hangi firmayı seçeceğine ya da seçeneklerinin neler olduğuna daha kolay verebilecektir.

Pazarlama stratejilerini düşündüğümüzde karar vericiler için en önemli unsur doğru bilgidir. Bu bilgiye ulaşmak ya da elde etmek çok maliyetli olabilmektedir. Bununla beraber bilgiye ulaşma anlamında yeterli kadar insan ile çözümlene yapılamadıysa elde edilen bilginin de doğruluğu azalacaktır. Buda stratejilerin yanlış temeller üzerinde kurgulanmasına sebep olacaktır.

Zamanımızın aracı olan sanal dünya ölçülebilir ve takip edilebilir bir yapıya sahiptir. Buna dikkat eden günümüz firmaları her sosyal mecrada yer alamaya çalışmaktadır. Bilinen hemen hemen her firma Twitter, Facebook, Instagram benzeri sosyal mecralarda kurumsal hesaplar açmaktalar. Firmalar sosyal mecraları aynı zamanda aktif olarak kullanmaktadırlar. Buda hem firmalar hem de insanların sosyal medyalar üzerinde bıraktıkları izlerin tarafımızdan kolayca izlenmesini sağlamaktadır. Bu büyük verinin farkında olan firmalar daha az enerji ile daha doğru sonuçlara bu sayede ulaşabilirler. Bu büyük veri üzerinde metin madenciliği işlemini yerine getirerek firmaların hizmetleri hakkında başarılı ya da başarısız oldukları ölçülebilmektedir. Bu sayede belirleyecekleri stratejileri hakkında karar verme süreçlerinde daha doğru bilgilere ulaşılmasını sağlanabilmektedir.

Projemizin sonuç ekranlarına bakıldığında GSM firmalarından müşterilerin beklentilerini veya firmaların birbirleri arasında ki farkı ya da benzerlikleri rahatlıkla görebiliyoruz. Böylelikle her GSM şirketi kendi eksiklerini belirleyerek müşteri memnuniyet kalitesinin artmasını sağlayabilirler. Alınan kararlar sonrasında ki iyileştirme süreçlerinin de ölçümleri yapılarak gelişmeleri de analiz edebilirler. Bu sayede metin madenciliğinin hızından faydalanarak her iyileştirme sonuçlarını ayrı ayrı analiz edebilirler. Örneğin GSM 1 internet kategorisi sorunu % 19,82 iken GSM 2 ve GSM 3 'de bu sorun % 6 seviyelerinde. Tüm GSM şirketlerinin kapsama ve hizmet yaklaşım yüzdeleri birbirine yakın. Fatura kategorisinde farklılık göstermektedirler. Fatura kategorisinde yaşanan sorunlarda GSM 1 %41,92, GSM 2 %39,48 ve GSM 3 %30,11 şeklinde sıralanmaktalar. Turkcell ve Vodafone 'nun internet kategorisinde daha az sorun yaşadıklarını da açıkça görebiliyoruz. Bu rakamlara bakıldığında hangi firmanın hangi eksikliğini öncelikle gidermesi gerektiği açıkça görünmektedir.

Metin madenciliği burada kullanıldığı gibi birçok alanda kullanılmaktadır. Örneğin sağlık problemleri olarak alınan hasta şikâyetleri üzerinden yeni gelen bir hastanın ön tespiti metin madenciliği uygulaması ile çok daha hızlı ve başarılı yapılabilir Geliştirilecek uzman sistem ile insan kullanımı hata faktörü de ortadan kalkar. Hastayı karşılayan ilk birimi düşünürsek mesai bitimine doğru yorgunluk sebebi ile hata yapma

oranı riski yüksektir. Metin madenciliği ile böyle bir uygulama geliştirilebilir. Böylelikle kişiye özel metin madenciliği uygulamasını da geliştirmiş olduk. Yakın gelecekte metin madenciliği konusunun hayatımıza sağlayacağı hizmeti ve faydası büyük olacaktır.

REFERANSLAR

Özcan C. (2014) Veri Madenciliğinin Güvenlik Uygulama Alanları ve Veri Madenciliği ile Sahtekârlık Analizi, İstanbul Bilgi Üniversitesi Sosyal Bilimler Enstitüsü Bilişim Ve Teknoloji Hukuku Yüksek Lisans Programı

Melek C. (2012) Metin Madenciliği Teknikleri Şirketlerin Vizyon İfadelerinin Analizi, Dokuz Eylül Üniversitesi Sosyal Bilimler Enstitüsü Ekonometri Anabilim Dalı Ekonometri Programı

Beyhan H. D. (2014) Sosyal Medya Üzerinden Metin Madenciliği ve Duygu Analizi ile Pazar Değerlendirme, İstanbul Teknik Üniversitesi Fen Bilimleri Enstitüsü Endüstri Mühendisliği Anabilim Dalı Mühendislik Yönetimi Programı

Aravi G. (2014) Metin Madenciliği İle Sosyal Medya Analizi Yüksek Lisans Tezi, İstanbul Aydın Üniversitesi Fen Bilimleri Enstitüsü Bilgisayar Mühendisliği Anabilim Dalı Bilgisayar Mühendisliği

Kuzucu K. (2015) Müşteri Memnuniyeti Belirlemek İçin Metin Madenciliği Tabanlı Bir Yazılım Aracı, Fen Bilimleri Enstitüsü Bilgisayar Mühendisliği Anabilim Dalı

Demirel A. (2015) Metin Madenciliği Yöntemleri İle Sosyal Medyadan Toplanan Fotoğraflı Paylaşımların, Metin – Fotoğraf Eşleşmesinin İncelenmesi, Beykent Üniversitesi Fen Bilimleri Enstitüsü Matematik Bilgisayar Anabilim Dalı Bilgi Teknolojileri Bilim Dalı

Sancar Y. (2016) Metin Madenciliği Kullanılarak Talep Tanıma Ve Yönlendirme Sistemi, Atatürk Üniversitesi Fen Bilimleri Enstitüsü Bilgisayar Mühendisliği Anabilim Dalı

Döven S. (2013) Metin Madenciliği İle Dokümanlar Arasındaki Benzerliklerin Bulunması, Bahçeşehir Üniversitesi Fen Bilimleri Enstitüsü Bilgi Teknolojileri Anabilim Dalı

Durmuş M.S. (2005) Veri Kümeleme Algoritmalarının Performansları Üzerine

Karşılaştırmalı Bir Çalışma, Pamukkale Üniversitesi Fen Bilimleri Enstitüsü

Chung-Hoon ParkYoung-Gul Kim (2003) Identifying key factors affecting Consumer Purchase Behavior in An Online Shopping Context International Journal of Retail & Distribution Management, Vol. 31 Iss: 1, Pp.16 - 29

Silahtaroğlu G. (2008), Veri Madenciliği, İstanbul: Papatya Yayınları

Oğuzlar A. (2011), Temel Metin Madenciliği, Bursa : Dora Yayınları

Melek C. (2012), Metin Madenciliği Teknikleri İle Şirketlerin Vizyon İfadelerinin Analizi, Dokuz Eylül Üniversitesi Sosyal Bilimler Enstitüsü Ekonometri Anabilim Dalı Ekonometri Programı

- Oğuz B. (2009), Kulak Burun Boğaz Epikriz Notlarından Birliktelik Kurallarının Çıkarılması, Akdeniz Üniversitesi
- Boz E. (2015) Kullanıcı Ve Sözlük İlişkisi, İç Anadolu Dilcileri Sürekli Çalıştayı-II
- Altan Z. (2016) Dil Modellemede Belirsizlik Probleminin Etmenlenmiş Dilbilgisi İle Giderilmesi, İstanbul Üniversitesi Mühendislik Fakültesi Bilgisayar Mühendisliği Bölümü, Makale
- Şentürk F. & Gündüz G. (2014), Arama Motoru Bing 'in Tekil-Çoğul Kelimeler İçin Gövdeleme Mekânizması, Pamukkale Üniversitesi Mühendislik Bilimleri Dergisi, Cilt 20, Sayı 2, 2014, Sayfalar 42-45
- Öner M. (1999), Türkçede Edatlı (Sentaktik) İsim Çekimi, Türk Dili, sayı:565, Ocak-1999, 1018. s.
- Çelikyay E. K. (2010), Metin Madenciliği Yöntemiyle Türkçede En Sık Kullanılan Ve Birbirini Takip Eden Harflerin Analizi Ve Birliktelik Kuralları, Beykent Üniversitesi Fen Bilimleri Enstitüsü Matematik Bilgisayar Anabilim Dalı Bilgisayar Ağları Ve İnternet Teknolojileri Bilim Dalı
- Gürsoy U. T. Ş. (2010), Uygulamalı Veri Madenciliği Sektörel Analizler, 2012 3. Baskı, ISBN: 978-605-364-040-0
- ŞENTÜRK A. (2006), Veri Madenciliği Kavram ve Teknikleri, ISBN: 975-8768-24-7
- Pektaş A. O. (2013), SPSS ile Veri Madenciliği, ISBN: 978-605-86660-4-7