T.C.
DÜZCE ÜNİVERSİTESİ
SAĞLIK BİLİMLERİ ENSTİTÜSÜ

METİN MADENCİLİĞİ VE SAĞLIK ALANINDA BİR UYGULAMA

Selçuk Göksel TOPLU

YÜKSEK LİSANS TEZİ

BİYOİSTATİSTİK VE TIBBİ BİLİŞİM ANABİLİM DALI

DANIŞMAN

Doç. Dr. Şengül CANGÜR

DÜZCE, 2019

BEYAN

Bu tez çalışmasının kendi çalışmam olduğunu, tezin planlanma aşamasından yazım aşamasına kadar bütün aşamalarda etik dışı davranışımın olmadığını, bu tezdeki bütün bilgileri akademik ve etik kurallar içinde elde ettiğimi, elde edilen bütün bilgi ve yorumlara kaynak gösterdiğimi ve yazımı sırasında patent ve telif haklarını ihlal edici bir davranışımın olmadığını beyan ederim.

05/ 08 /2019

Selçuk Göksel TOPLU

ÖNSÖZ

Yüksek Lisans eğitimi boyunca ilminden faydalandığım, her zaman bana destek olan, insani ve ahlaki değerleri ile de örnek edindiğim, yanında çalışmaktan onur duyduğum ve ayrıca tecrübelerinden yararlanırken göstermiş olduğu hoşgörü ve sabırdan dolayı değerli hocam Sayın Doç. Dr. Şengül Cangür'e, öğrenimim süresince daima yanımda olduğunu hissettiğim, çalışma süresince değerli bilgi ve becerilerini esirgemeyen değerli hocam Sayın Prof. Dr. Handan Ankaralı'ya ve tez zamanı boyunca hep yanımda olan, en zor zamanlarımda bana desteğini hiç bir zaman esirgemeyen sevgili ablam Op. Dr. Gaye Toplu'ya sonsuz teşekkür eder,

Saygılarımı Sunarım.

Selçuk Göksel TOPLU

İÇİNDEKİLER

KISALTMALAR VE SİMGELER

ARFF : Attribute Relation File Mormat

Cos : Cosinus

CRISP-DM : Cross Industry Standard Process Model for Data Mining

DDİ : Doğal Dil İşleme

Dist : Distance

DNA : Deoxyribonucleic Acid

FN : False Negative

FP : False Positive

HTML : Hypertext Markup Language

IDF : Inverse Document Frequency

K-NN : K - Nearest Neighbor

NLPBA : Natural Language Processing in Biomedical Applications

NLP : Natural Language Processing

POS : Part of Speech

RNA : Ribonucleic Acid

Sim : Similarity

TN : True Negative

TP : True Positive

TF : Term Frequency

XML : Extensible Markup Language

ŞEKİLLER LİSTESİ

ÖZET

METİN MADENCİLİĞİ VE SAĞLIK ALANINDA BİR UYGULAMA

Selçuk Göksel TOPLU

Yüksek Lisans Tezi, Biyoistatistik ve Tıbbi Bilişim Anabilim Dalı

Tez danışmanı Doç. Dr. Şengül CANGÜR

Ağustos 2019, 72 Sayfa

Teknolojinin hızla gelişmesi, bilgisayarların ve internetin gündelik yaşama daha fazla entegre olmasıyla birlikte veri tabanlarındaki verilerin de hızla artış göstermesine sebebiyet vermiştir ve birçok işlemin elektronik ortamda kayıt altına alınması, bu kayıtların saklanabilmesini, istendiğinde erişilebilmesini hem kolaylaştırmış hem de daha ucuza sahip olunmasını sağlamıştır. Bu durumda ham verilerin veritabanlarıyla birlikte günden güne artış göstermesiyle beraber, elde edilmek istenen bu verilerin doğru ve güvenilir olma ihtiyacı da ortaya çıkmış ve gereklilik haline gelmiştir. Bundan dolayı veri madenciliği oldukça önemli bir çalışma alanı bulmuştur. Veri madenciliğinde sayısal haldeki verilerin analizi yapılabilmekteyken, metinsel durumda bulunan yani sayısal olmayan verilerin analiz edilmesi de önemli bir ihtiyaç haline gelmiştir. Bu ihtiyaçtan dolayı metin madenciliğine yönelik çalışmalar da hız kazanmıştır. Metinsel verileri sayısal hale getirerek veri madenciliği algoritmalarına uygulanabilir hale getiren metin madenciliği, günümüz dünyasında büyük önem teşkil etmektedir. Bu tez çalışmasının amacı, metin madenciliği yöntemini ve uygulama adımlarını tanıtmak, ve sağlık alanında belirlenen bir konuda uygulamasını göstermektir. Çalışmanın uygulama aşamasında; "insanlarda görülen kanser vakaları (human and cancer)" ve "farelerde kanser araştırmaları (mouse and cancer)" şeklinde belirlenen iki farklı konu başlığı altında en sık kullanılan Pubmed veritabanından ayrı ayrı elde edilen dokümanlar birleştirilerek, bu dokümanlara sırasıyla metin madenciliği tekniklerinin uygulanmasına, Knime programının metin madenciliğinde nasıl kullanıldığına ve elde edilen dokümanlara uygulanan adımların neler olduğuna ayrıntılı olarak yer verilecektir.

Anahtar sözcükler: Biyoistatistik, Kanser, Knime, Metin Madenciliği, Veri Madenciliği

ABSTRACT

TEXT MINING AND AN APPLICATION IN HEALTH

Selçuk Göksel TOPLU

Master of Science Thesis, Department of Biostatistics and Medical Informatics

Supervisor Assoc. Prof. Dr. Şengül CANGÜR

August 2019, 72 Pages

The development of technology at speed has led to a rapid increase in the data in the databases as computers and the Internet are more integrated into daily life and recording of many transactions in the electronic environment has made it possible to store these records and make them easier to access and provide cheaper. In this case, as the raw data increased day by day along with the databases, the need to be accurate and reliable emerged. Therefore, data mining has become an important field of study. While it is possible to analyze numerical data in data mining, analyzing non-numerical data in text mining has become an important need. Because of this need, studies on text mining have gained momentum. Text mining digitizes textual data and makes it applicable to data mining algorithms and it is of great importance in today's world. In the application phase of the study, there are two different topics that are identified as "cancer cases in humans ”and "cancer research in mice" and these documents obtained from the most commonly used Pubmed database have been combined and text mining techniques have been applied to these documents respectively. It will be given detail about the Knime program is used in text mining and the steps taken in the documents obtained will be given in detail.

Key words: Biostatistics, Cancer, Data Mining, Knime, Text Mining

1. GİRİŞ ve AMAÇ

Son yıllarda gelişen bilgi teknolojileri sayesinde üretilen veri miktarı da hızla büyümektedir. Bilgisayar teknolojilerindeki hızlı gelişmeler hayatın her alanında kolaylıklar sağlamakta ve beklentileri her geçen gün arttırmaktadır. Bu beklentilerin karşılanması amacıyla ihtiyaca göre birçok sistem geliştirilmekte ve bu sistemler günümüzde giderek daha fazla özelliğe sahip olmakta ve fonksiyonel hale gelmektedir[1].

Veritabanlarının gün geçtikçe artması, teknolojinin ilerlemesi ve internet kaynaklı veritabanlarının devamlı olarak kendisini güncellemesi neticesinde çok daha fazla ebatlı verilerden gereksinim duyulan verileri elde etmek daha da zorlaşmaktadır. Daha önceleri kiloByte olarak adlandırılan daha düşük ölçekli veriler, teknolojinin hızla gelişmesi neticesinde yetersiz kalmış, günümüze kadar yottaByte boyutuna ulaşarak ifade edilmeye başlanmıştır. Böylelikle büyük veri kavramı da ortaya çıkmıştır. Buna bağlı olarak verinin depolanma sistemleri de gelişme göstermiştir. Bu durum, veriye geksinim duyan kişi ve kurumların işini zorlaştırmakta, zaman ve ekonomik kayıplara neden olmaktadır[2].

Üretilen ve depolanan verinin giderek artması, verinin modellenmesi ve veri içindeki "saklı" bilgiye ulaşılmasına aracı olan tekniklerin de değişmesine neden olmuştur. Farklı kaynaklarda depolanmış, büyük boyutlu veri, günümüzde "Veri Madenciliği", "Makine Öğrenmesi" ve "Büyük Veri" alanlarını ortaya çıkarmıştır. Veri madenciliği ile kişi ve ilgili kurumlar kendi veritabanlarından veya başka bir yüksek kapasiteli veritabanlarından çok hızlı bir şekilde gereksinim duydukları anlamlı bilgiyi elde edebilecek aşamaya erişmişlerdir.

Veri Madenciliği uygulamaları çoğunlukla yapısal veriler üzerinde gerçekleştirildiğinden, sadece metinden oluşan ve yapısal olmayan verilerin yapısal verilere dönüştürülmesi gerekmektedir. Bu durumda metin madenciliği devreye girmektedir. Metin madenciliği metin formatındaki verileri kullanarak yapısal

olmayan verileri yapısallaştırır ve metinlerden nümerik değerler elde ederek bilgiye ulaşılmasını sağlar.

Metin madenciliği teknikleri, çeşitli anlamsal bilgileri otomatik olarak tanıma kabiliyetleri sayesinde eş anlamlı kavram biçimleri ve kavramlar arasındaki ilişkileri kullanma konusunda yardımcı olabilirler[3].

Bu çalışma; sağlık alanında elde edilmiş dokümanları analiz ederek bilgi çıkarmayı amaçlayan araştırıcı ve analistlerin kullandığı "Metin Madenciliği" ni ele almaktadır. Metin Madenciliği, araştırma konusu olan verinin sadece doküman ya da dokümanlardan oluştuğu ve "metin" üzerinde analizlerin gerçekleştirildiği bir alandır.

Çalışmanın birinci bölümünde veri madenciliği, veri madenciliği uygulama alanları, veri madenciliği modelleri ve veri işleme sürecinden bahsedilmiştir.

İkinci bölümde metin madenciliği ve metin madenciliği uygulama alanlarına yer verilmiş ve metin madenciliğinde en sık tercih edilen sınıflama analizi ayrıntılı olarak ele alınmıştır. Bu çalışmada metin madenciliğinde veriyi hazırlama, ön işleme gibi aşamalarından bahsedilmiş, ön işlemlerden geçirilen kelime kökleri elde edilerek, bu köklerin metin içerisinde tekrarlanma sıklığı olan frekansları üzerinde uygun olan veri madenciliği analizleri yapılmıştır.

Çalışmanın üçüncü bölümünde ise sağlık alanıyla ilgili bir uygulamaya yer verilmiştir. İnsanlardaki kanser vakalarıyla alakalı yapılan araştırmaları içeren, Pubmed veri tabanından elde edilmiş dokümanlar ile fareler üzerinde yapılan kanser araştırmalarını içeren yine bu veri tabanından elde edilen dokümanlara, açık kaynak kodlu bir veri analiz uygulaması olan Knime aracılığı ile metin madenciliği algoritmaları uygulanmıştır.

Son bölüm olan dördüncü bölümünde ise elde edilen sonuçlar paylaşılarak tartışılmıştır.

2. GENEL BİLGİLER

2.1. Veri Madenciliği

Veri madenciliği, çok sayıdaki incelenmiş verilerden kurallar, örüntü ve modellerin elde edilmesidir. Bir diğer anlatım ile veri madenciliği, veri tabanları veya veri depolarında yer alan kütle veri içindeki saklı örüntüleri ve bağlantıları keşfetmek için istatistiksel algoritmaları ve yapay zekâ metotlarını işleten komplike bir veri arama kabiliyeti olarak tarif edilebilir. Veri madenciliği; bununla birlikte bilgisayar bilimini, makine öğrenmesini, veri tabanı yönetimini, matematiksel algoritmaları ve istatistiği bir araya getiren disiplinler arası bir alandır. Aynı zamanda veri madenciliğini değişik araştırmacılar aracılığıyla,

- Veri madenciliği makro veri kümeleri içinde gizli olan, yararlı bilgilerle umumi olarak kestirilemeyen eğilim ve bağlantıların ortaya çıkarılması için bir eleme çalışmasıdır[4].
- Veri madenciliği veritabanı sahibi için önemli miktardaki veriden bilinmeyen bağlantı ve uyumların ortaya çıkarılması ile yararlı ve tartışmaya açık olmayacak neticelere ulaşmayı amaçlayan seçme, araştırma ve modelleme sürecidir[5].
- Veri madenciliği, bilinmeyen ilişkilerin bulunması ve verinin değişik şekillerde özetlenmesi için gözlemsel verilerin, veri sahibi için anlaşılır ve yararlı olacak şekilde analiz edilmesidir[6]

olarak ifade edilmektedir.

Veri madenciliği, veri tabanındaki bilginin ortaya çıkarılması aşamasının bir etabıdır. Bilginin ortaya çıkarılması aşamasındaki adımları şu şekilde belirtilebilir. Bu aşamalar interaktif olup gerektiği durumlarda sıralama farklı şekilde gerçekleşebilir.

Veri temizleme: Gürültülü ve tutarlı olmayan verilerin işlem dışı tutulması.

Veri bütünleştirme: Farklı veri kaynaklarını bir araya getirmek.

Veri seçme: Uygulanacak analizle alakalı verileri belirlemek.

Veri dönüşümü: Verinin veri madenciliği tekniğiyle birlikte değerlendirilebilecek duruma getirilmesini sağlamak.

Veri madenciliği: Verideki örüntüleri ele geçirebilmek için teknikleri uygulamak.

Bilgi sunumu: Madenciliği gerçekleştirilmiş olan ele geçirilmiş bilginin kullanıcıya sunumunu ortaya koymak.

Veri madenciliğinde örüntü belirleme etkinlikleri üç ana aşamada bir araya getirilebilir. Bunlar; keşif (discovery), tahmin edici modelleme (predictive modelling) ve adli analizdir (forensic analysis). Keşif, bir veri kütlesindeki saklı olan örüntüleri daha evvelden saptanmış bir düşünce veya hipotez olmaksızın oluşturma sürecidir. Tahmin edici modelleme, bulunan örüntüler ile geleceği kestirmek için değerlendirilmektedir. Adli analiz ise gerçekleşmiş örüntülerin, kural dışı veya anormal veri elemanlarını keşfetmek için değerlendirilmesi süreci olarak tarif edilebilir[7].

Veri madenciliğinin ortaya çıkması, kavramsal şekilde 1960'lı senelerde, bilgisayarların veri analiz problemlerine çözüm üretmek için değerlendirmeye başlamasıyla gerçekleşmiştir. Veri madenciliği kavramı ortaya çıkmadan evvel, veri taraması (data dredging) ve veri yakalanması (data fishing) gibi türlü isimler ile anılmaktaydı. 1960'lı senelerde veri toplama ile ortaya çıkan bu süreç, 1970'lerde veritabanlarının meydana gelmesi ile sürmüştür. 1990'lı senelerde ise veri madenciliği ismi, Rakesh Aggrawal liderliğinde birtakım bilgisayar mühendisleri tarafından öne sürülmüştür. Daha sonra ise veri madenciliğine türlü yaklaşımlar getirilmeye başlanmıştır. İstatistik, makine öğrenimi (machine learning), veritabanları, otomasyon, pazarlama, araştırma gibi disiplinler ve kavramlar bu yaklaşımların temelinde bulunmaktadır[8].

2.1.1. Veri Madenciliği Uygulama Alanları

Günümüzde bilgisayar sistemlerinin gelişmesiyle birlikte veri madenciliğinin uygulama alanları da oldukça gelişmiştir. Analiz edilen verinin yapısı ve boyutları, farklı bilim dallarında ve sektördeki uygulama alanlarına göre değişiklik göstermektedir. Veri madenciliği uygulama alanları şu şekilde incelenebilir:

Tıp; DNA içerisinde bulunan genlerin sıralarının ortaya konulması, protein analizlerinin yapılması, hastalık anahatlarının oluşturulması.

Perakendecilik; genel piyasa analizleri, optimal müşteriler veya müşteri departmanlarının ortaya çıkarılmasında, piyasa analizleri, hisse senedi tahminleri, satış sonrası analizler, alışveriş analizleri.

Pazarlama; müşterilerin satın alma alışkanlıklarının keşfedilmesi, yeni müşterilerin kazanılıp, eski müşterilerin elde tutulması, satış tahminleri, Pazar sepeti analizi, müşterilerin demografik özellikleri arasındaki bağlantının keşfedilmesi, çapraz satış analizleri.

Bankacılık; kredi isteklerinin değerlendirilmesi, farklı mali indeksler arasındaki saklı korelasyonların ortaya çıkarılması, müşterilerin kredi kartı harcamalarına göre gruplandırılması, usulsüzlük saptanması ve risk analizi.

Sigortacılık; müşterilerin yeni poliçe talepleri doğrultusunda tahmin edilmesi, riskli müşteri örüntülerinin tespit edilmesi, sigorta dolandırıcılıklarının saptanması.

Endüstri; lojistik, kalite kontrol analizleri, üretim süreçlerinin optimize edilmesi.

Telekomünikasyon; iletişim ağlarında problemli yerlerin saptanması, kaçak hat kullanımlarının tespit edilmesi, kullanıcı tutumlarının ortaya çıkarılması, müşteri tutumlarına göre daha yeni hizmetlerin sunulması olarak belirtilebilir[9].

2.1.2. Veri Madenciliği Süreci

Veri madenciliği, bununla birlikte ortaya konan sürecin işlenmesidir. Nitelikli ve kullanışlı veri madenciliği neticeleri elde edebilmek için kayıp verilerin, yanlış kodlanmış veya yanlış işlenmiş verilerin olması, gerçek hayattaki verilerin çok büyük miktarlarda olması, hatalı değerler içeren gürültülü verilerin olması gibi sebeplerden dolayı veri madenciliği süreçleri uygulanmadan evvel veri işleme tekniklerinin uygulanması gereklidir. Bilgi keşfi aşamasında örüntüleri süzmek ve bir sonraki aşamaya hazır hale getirmek de bu sürecin bir parçasıdır. Üzerinde inceleme yapılan verinin özelliklerinin bilinmesi son derece önemlidir. Aksi takdirde bu veri yığını ne kadar etkin olursa olsun hiçbir veri madenciliği algoritmasından fayda sağlamak mümkün olmayacaktır. Bu nedenle, veri madenciliği sürecine girilmeden evvel, analizlerin ilk koşulu, iş ve veri özelliklerinin detaylı analiz edilmesidir[8,9,10].

Problemin tanımlanması: Veri madenciliği çalışmalarının olmazsa olmazı, problemin tanımlanma aşamasıdır. Problemin amacının net bir şekilde ifade edilmesi gerekmektedir. Hangi işletme amacı için yapılacağı ve elde edilecek sonuçların başarısının nasıl ve neye göre ölçüleceğinin belirlenmesi ve tanımlamasının yapılması en önemli aşamadır.

Verilerin hazırlanması: Bu basamak, çalışmaya temel olacak son verilere dönüştürme aşaması olarak da tanımlanabilir. Aynı zamanda problemin hazırlanmasından sonraki basamaktır. Modelin kurulması basamağında ortaya çıkacak beklenmedik sorunlar, bu basamağa sürekli geri dönülmesine ve verilerin yeniden düzenlenmesine neden olabilecektir. Verilerin hazırlanması, “toplama”, “değer biçme”, “birleştirme ve temizleme”, “örneklem seçimi” ve “dönüştürme” basamaklarından meydana gelmektedir.

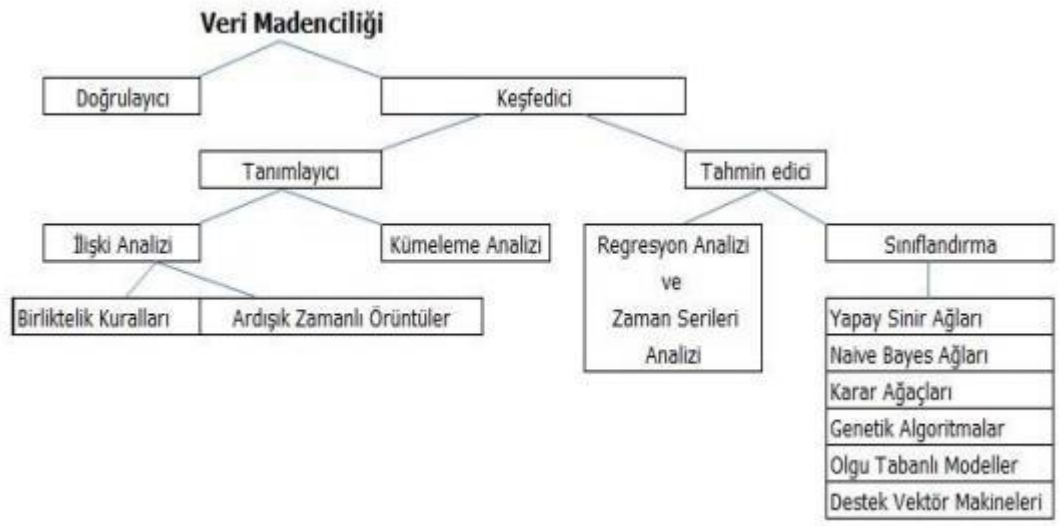
Modelin kurulması ve değerlendirilmesi: Oldukça fazla miktarda modelin kurularak denenmesi, tanımlanan model için en uygun modelin saptanması için gereklidir.

Bundan dolayı, veri hazırlama ve model oluşturma basamakları en iyi olduğu varsayılan modeli keşfedinceye kadar tekrar edilen bir süreçtir.
Modelin kullanılması: Hazırlanan ve onaylanan model, direkt olarak bir uygulama olabileceği gibi, başka bir uygulamanın alt uygulaması olarak da işlevselliğini sürdürebilir.

Modelin izlenmesi: Zamanla tüm sistemlerin özellikleri ve dolayısıyla meydana getirdikleri verilerde değişim ve farklılaşma gözlenebileceğinden, bu durum, oluşturulan modellerin sürekli olarak izlenmesi ve gerekirse yeniden bir daha düzenlenmesi ihtiyacını doğuracaktır.

2.1.3. Veri Madenciliği Modelleri

Veri madenciliğinde kullanılan modeller farklı şekillerde sınıflandırılabilmektedir. Bu modeller doğrulayıcı ve keşifleyici olarak iki temel sınıfa ayrılabilir. Keşfedici modeller ise yine kendi içinde tahmin edici ve tanımlayıcı olarak iki grupta ele alınabilir. Şekil 2.1 bu bağlamda ele alınan veri madenciliği modellerinin sınıflandırılmasını göstermektedir[10].



Şekil 2.1. Veri madenciliği modelleri[10].

2.1.3.1. Doğrulayıcı ve Keşfedici Modeller

Doğrulayıcı modeller; araştırmacının bilgi birikimi ve geçmiş deneyimlerinin de yardımıyla araştırmak istediği konuyla alakalı veya belirlenmiş bir hipotezin değerlendirilmesi ile ilgilenir. Bu yöntemler, keşfedici veri madenciliği ile pek de ilgili sayılmamakla birlikte keşfedici analiz sürecinde, gerekli görüldüğünde kullanılmaktadır. Doğrulayıcı modeller; uyumluluk testi, varyans analizi, t-testi gibi süregelmiş istatistiksel yöntemleri içerir. Keşfedici modeller ise tahmin edici ve tanımlayıcı olmak üzere iki gruptan oluşan modellerdir ve genellikle veri kümesi içindeki örüntüleri yakalamak için kurulur[11].

2.1.3.1.1. Tanımlayıcı Modeller

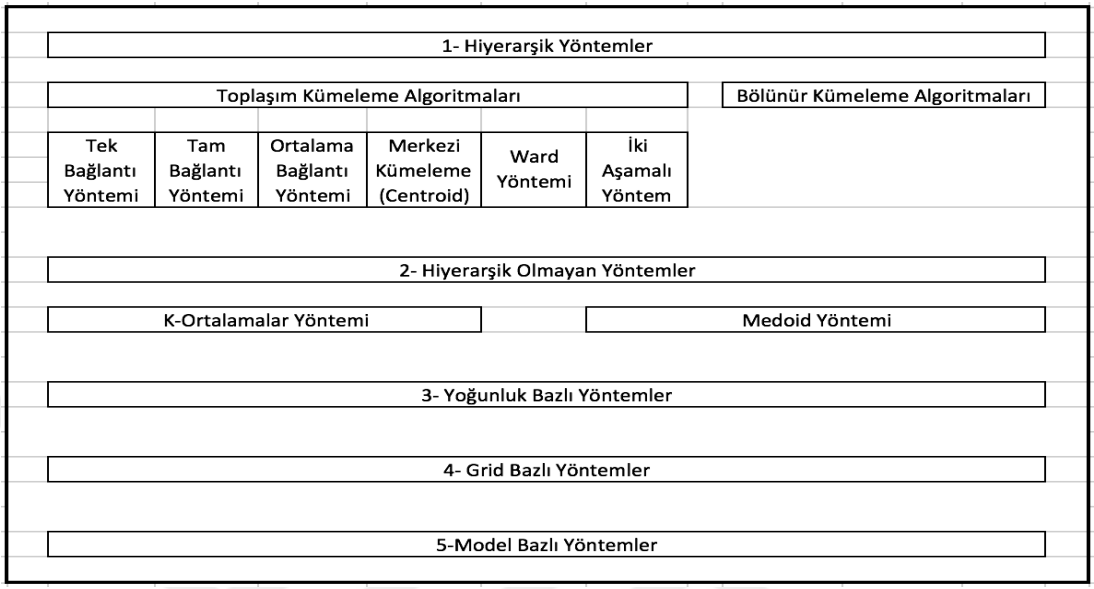
Analiz edilen veri kümesinin altındaki bilgilerin ortaya çıkmasını sağlayan, yani veri kümesindeki var olan örüntüleri anlamlandıran ve tanımlayan, karar verme aşamasına rehberlik sağlamak amacıyla kullanılabilen modellerdir. Kümeleme, birliktelik kuralları ve ardışık zamanlı örüntüler, tanımlayıcı modellerdir.

2.1.3.1.1.1. Kümeleme Analizi

Bu analiz yönteminin amacı, veri içerisinde yer alan ve birbirine benzer üyeleri olan farklı grupları keşfedip ortaya çıkarmaktır. Kümeleme analizinde, aynı grup üyelerinin homojen yani birbirlerine benzer, farklı grup üyelerinin ise heterojen yani birbirlerinden farklı özelliklerde olması beklenmektedir.

Kümeleme analizi; oluşturulan grafikler ile veri seti içerisindeki grupların benzerliklerinin basit bir şekilde görselleştirilmesi, veri içerisindeki farklı ve aykırı gözlemlerin basit bir şekilde belirlenmesi, büyük miktarda veriler nedeniyle çalışamayan algoritmalar için örneklemler yaratması gibi veri madenciliğinde farklı

amaçlar için de kullanılabilmektedir. Şekil 2.2 bu bakış açısıyla kümeleme algoritmalarının sınıflandırılmasını gösterilmektedir[10,12].



Şekil 2.2. Kümeleme algoritmaları sınıflandırması.

Kümeleme analizinde gruplara ayırma işlemlerinde, elde edilen gözlemlerin birbirlerine ne derece benzerlik gösterdiklerini belirlemek için gözlemler arasındaki mesafeler kullanılır. Mevcut veri setindeki değişkenlerin ölçeğine göre Öklit (Euclidean), Karesel Euclidean, Pearson, Manhattan, Minkowski, Mahalanobis mesafe ölçüleri veya Jaccard, Ochiai, Rao benzerlik katsayılarından faydalanılmaktadır[12].

2.1.3.1.1.2. Birliktelik Kuralları

Veri setinde bir arada sıklıkla görülen, eş zamanlı gerçekleşen olayları ortaya çıkarmak için birliktelik kuralları kullanılır. Birliktelik kuralları analiz süreci, market sepeti analizi olarak da tanımlanır. Bu kurallar pek çok verinin depolandığı büyük bir veri tabanı içerisinde, farklı özellikler içerisinden hemen gözlemlenemeyen birçok bağlantının keşfedilmesi, önemli ve stratejik kararların alınmasına imkân sağlayabilmektedir. Fakat çok sayıda verinin içerisinden bu ilişkilerin ortaya

çıkarılması kolay bir süreç değildir. Meydana gelen bu süreç birliktelik kuralı madenciliği (association rule mining) olarak bilinmektedir[12].

2.1.3.1.2. Ardışık Zamanlı Örüntüler

Birbirleri ile bağlantılı durumda bulunan ve art arda dönemlerde meydana gelen olaylar arasındaki bağlantıyı tanımlamada kullanılır. Örneğin, X ameliyatı yapıldığında, Y enfeksiyonu oluşma ihtimali olduğunun belirlenmesi birliktelik kuralları ile bulunurken, birbirini izleyen ameliyatlarda bu eğilimin tahmini miktarının elde edilmesi ardışık zamanlı örüntüler ile ortaya konulur[10].

2.1.3.2. Tahmin Edici Modeller

Bu modeller; mevcut verileri kullanarak, bilinmeyen herhangi bir değeri tahmin etmeye çalışırlar. Oluşturulan veri kümesinden hareketle yeni bir model geliştirilmesi ve bu geliştirilen bu modelden faydalanılarak yeni ve sonuçları bilinmeyen veri kümeleri için sonuç tahminleri yapılması hedeflenmektedir. Hangi verinin en anlamlı olduğu, her bir değişkenin önemi, bu modellemelerde ortaya koyulur. Tahmin edici modeller olarak sınıflandırma, regresyon analizi, zaman serileri analizi modelleri kullanılmaktadır.

2.1.3.2.1. Sınıflandırma

Denetleyici (supervised) öğrenme gerçekleştiren bu modeller, kümeleme ile birlikte veri madenciliği teknikleri arasında en çok kullanılan modellerden birisidir. Sınıflandırma tekniklerinin kullanıldığı alanlar arasında; hastalık tanıları, kalite kontrol ve pazarlama, resim ve örüntü tanıma, dolandırıcılık tespiti, kredi kartı başvurusu değerlendirme gibi konular yer almaktadır. Amaç, verilen birden fazla kategoriye ait verileri birbirinden ayırarak önceden bilinen farklı gruplara atamak suretiyle bir model oluşturmaktır. Yapılan bu atama yardımıyla yeni karşılaşılan

verilerin hangi sınıfa ait olduğu tahmin edilmektedir. Birçok sınıflandırma modeli mevcuttur. En çok kullanılan sınıflama modelleri şu şekilde sıralanabilir[10,12,13]:

Karar Ağaçları (Decision Trees); Karar ağaçlarında temel düşünce, veri kümesine ait ögelerin gruplara ayrılmasına dayanır. Burada gaye grubun bütün ögeleri benzer sınıf etiketine (label) sahip olana kadar işlemi devam ettirmektir. Veride en iyi seçimin yapılmış olması demek, veri bir özelliğe göre parçalandığında ortaya çıkan her bir veri kümesinin belirsizliği minimum ve dolayısıyla bilgi kazancı maksimum olması anlamına gelmektedir. Bunun için özellik vektörleri incelenir ve en fazla ölçüde bilgi kazancına (information gain) sahip olan özellik, bir çok kombinasyon çözülerek, ağaçta dallanma yapmak amacıyla tercih edilir. Ağacın şekli kullanılan algoritmaya göre değişiklik gösterebilir.

Yapay Sinir Ağları (Neural Networks): En basit tanımla, insan beyninin çalışma yapısını taklit ederek modelleme yapan algoritmalardır. Yapay sinir ağları; öğrenme sürecinin insan beyninden esinlenilmesi suretiyle, matematiksel olarak modellenmeye çalışılması neticesinde oluşmuş bir algoritma çeşitidir. Bundan dolayıdır ki, ilk olarak bu konudaki araştırmalar, insan beynini bir araya getiren, biyolojik birer birim olan nöronların incelenip modellenmesi ve bu modellemelerin bilgisayar sistemlerinde uygulanması ile başlamış, daha sonraları bilgisayar sistemlerinin gelişip ileri bir seviyeye ulaşmasıyla birlikte pek çok alanda kullanılır duruma gelmiştir. Yapay sinir ağlarının, eksik, mutlak olmayan, kompleks, gürültülü, hatalı, hata ihtimali fazla olan sensör verilerinin mevcut olması ve problemi çözümlemek amacıyla matematiksel modelin ve algoritmaların mevcut olmadığı, yalnızca örneklerin bulunduğu durumlarda yaygın bir şekilde kullanıldıklarına rastlanmaktadır. Ayrıca analizcinin bilgi ve tecrübesine dayanır, parametre ayarları ve iteratif yapısından dolayı her denemede farklı sonuçlar doğurur. Konstrüktif olmaları, duyarlı eşleştirmeleri başarı ile ortaya koymalarıyla ve yapısallıkları ile giderek daha çok uygulama alanları bulabilmektedirler.

Bayes Ağları (Bayesian Networks): Bayes ağları yönlü dönüşsüz olasılıksal ağlardır (directed acyclic network). Her düğüm ayrı bir değişkeni ifade eder. Ayrıca bu

rastgele değişkenler arasındaki sıralama da bayes ağları ile yönlü oklar aracılığıyla basitçe bir düğümden diğer düğüme geçiş sırası şeklinde gösterilebilir. Genel olarak bir bayes ağı iki ana parçadan oluşur. Bunlar, düğümler ve oklar yardımıyla değişkenler ve değişkenler arası olasılıksal bağlantıların gösteriminin gerçekleştiği grafiksel bölüm ve değişkenlere ait şartlı olasılık tablolarıdır.
Genetik Algoritmalar: Yapay zekânın bir araştırma alanıdır ve birçok alanda kullanılmaktadır. Doğal seçim prensiplerine dayanan bir arama ve optimizasyon yöntemidir. Bilinen optimizasyon yöntemlerinden farklı olan bu algoritmalar, parametre kümesini kullanmayıp, bunun kodlanmış şeklini kullanırlar. Sadece amaç fonksiyonuna gereksinim duyan bu algoritma çeşidi, olasılık kurallarına göre çalışırlar. Mekanik öğrenme, fonksiyon optimizasyonu, çizelgeleme, hücresel üretim, tasarım gibi alanlarda başarılı uygulamaları söz konusudur.

Olgu Tabanlı (Instance Based) Modeller: Tahmin işleminde önceden düzenlenmiş tahmini ve soyut çıkarımlardan farklı olarak, özel ve farklı örnekler kullanır. Olasılık kavramları tanımlayan ifadeler kullanabilen bu algoritmalar, örnekleri kategorize ederken doğru eşleşmeyi elde etmek için benzer fonksiyonları kullanırlar.

Destek Vektör Makineleri (Support Vector Machines): Sınıflandırmayı doğrusal ya da doğrusal olmayan bir fonksiyon yardımıyla yapan algoritmalardır. En elverişli fonksiyonun, veriyi birbirinden ayırmak amacıyla tahmin edilmesi esasına dayanmaktadır.

2.1.3.2.2. Regresyon ve Zaman Serileri Analizi

Regresyon analizi ve zaman serileri analizi şahsi hükümlerden etkilenmeyen, objektif tahminler geliştirilebilmesi ve işletmelere yerinde ve isabetli kararlar alabilmelerinde önemli avantajlar sağlamaktadır[10,12,13].

Bir zaman serisi, belirli aralıklarda gözlemlenen ve değerleri kaydedilen bir büyüklüğün zaman içerisinde sıralanmış ölçümlerinin bir kümesidir. Zaman serisi ile ilgili bu analizin yapılma amacı ise, gözlem kümesince temsil edilen gerçeğin

anlaşılması ve zaman serisindeki değişkenlerin gelecekteki değerlerinin doğru bir şekilde tahmin (forecast) edilmesidir. Bu özelliği nedeniyle zaman serileri analizi, sabit şartlar altında daha fazla etkin olmaktadır. Regresyon analizinin kullanılması ise, iki ya da daha fazla değişken arasındaki ilişkinin ölçülmesi ile ilgilidir. Değerleri tahmin edilecek değişkenle ilişkili olan diğer değişkenlerin saptanmasını kapsamaktadır. Belirlenen bu değişkenlerden sonra meydana getirilen istatistiksel model, tahmin değişkeni ile diğer değişkenler arasındaki ilişkiyi tanımlayıp ilgilenilen değişken ile alakalı tahminler yürütülmesinde kullanılmaktadır.

2.2. Metin Madenciliği

Metin madenciliği, kişisel veya özel amaçlar doğrultusunda metinsel ve yapısal olmayan dokümanlardan bir takım bilgiler çıkarmak için, metnin analiz edilmesi işlemidir[14]. En kısa tanımla metin madenciliği, veri madenciliğinin dokümanlar üzerinde bulunan metinlere uygulanması işlemidir. Metin madenciliği, belirli bir biçimde olmayan, metin türündeki veriler içerisinde gizli kalmış vasıflı bilginin ortaya çıkarılması, düzenli bir durumda olmayan verinin biçimlendirilmesi sürecidir. Metin sınıflandırması ise önceden saptanmış gruplara göre, doğal dil metinlerinin sınıflandırılmasıdır[15]. Metin tabanlı bilgileri işleyen hesaplama düzeneklerinin temel stratejisi, çok fazla sayıda olan doğal dil girdilerini, küçük kategoriler kümesine indirgemektir. Günümüzde dijital ve basılı dokümanların sayısı oldukça fazladır ve gün geçtikçe artmaktadır. Büyük ölçekte yapısal olmayan veri barındıran bu dijital dokümanlar, web sayfaları, e-postalar ve yazılı ortamdaki dokümanların dijitale dönüştürülmesiyle elde edilen dijital kaynaklar olarak örneklendirilebilir. Bu yapısal olmayan verilerin işlenmesi ve analiz edilmesi, sayısal verilere göre farklılıklar gösterebilmektedir. Metin madenciliğinde araştırmacılar düzenli durumdaki verileri analiz ettikleri gibi, makalelerden, internet sayfalarındaki metinlerden, tıbbi raporlardan, fatura bilgilerinden, kısaca metinsel halde olan verilerin de analizini gerçekleştirebilmektedirler[16].

Metin Madenciliği, Metin Veri Madenciliği (Text Data Mining) ve Metin Veritabanlarından Bilgi Keşfi (Knowledge Discovery from Textual Databases)

olarak da adlandırılır[17]. Bazı metin madenciliği çalışmaları, literatürde farklı isimlerle yer alabilmektedir. Örneğin, sadece internetin incelenmesi, internet analizi (web mining); fikir anlatan terimler, duygu analizi (sentiment analysis); sosyal medyada yer bulan kısa metinler, sosyal medya analizi olarak ifade edilebilir. Bu uygulamalarda genellikle sınıflama, kümeleme ve birliktelik analizleri kullanılmaktadır.

Metin madenciliği yeni bir terim olmasıyla beraber, bilgi erişim sistemleri ve Doğal Dil İşleme (DDİ) ile alakalı gerçekleştirilen çalışmalara bağlı olarak meydana gelmiştir. Bilim ve teknoloji metin madenciliğini "bilginin teknik literatürden çıkartılması" şeklinde tanımlamış olan Kostoff ve DeMarco[18], metin madenciliğini bilgi erişim, bilgi işleme ve bilgi entegrasyonu olacak şekilde üç bileşenden meydana geldiğini ifade etmişlerdir. Bilgi işlemeyi, ulaşılan dokümanlardaki örüntülerin ortaya konulması işlemi, bilgi entegrasyonunu ise ulaşılan ilgili dokümanların okunarak bilgi işleme safhasından sonra çıkan sonuçlarla birleşiminin gerçekleştirilmesi aşaması şeklinde tanımlamışlardır. Ayrıca Losiewicz[19], metin veri madenciliğini, metin derlemlerinden bilgiye ulaşmayı, bireysel metinlerden bilgi çıkarmayı, veritabanlarından bilgi keşfini, kurumlarda bilgi yönetimini ve veriyle bilginin görselleştirilmesi basamaklarını bir araya getiren bir mimari olarak tanımlamıştır. Çok sayıda dijital metnin kısa sürede analiz edilmesi ve nitelikli bilgilere kısa sürede erişebilmek için metin madenciliği, sıklıkla kullanılan yöntem haline gelmiştir.

2.2.1. Metin Madenciliği Uygulama Alanları

Metin madenciliğinde uygulama alanları şu şekilde sıralanabilir[20,21];

Enformasyon Getirimi (Information Retrieval): Bu evre ilgilenilen korpus (derlem) ile alakalı ön bilginin elde edildiği evredir. Örneğin metin madenciliği herhangi bir dosya düzeni üzerinde yapılacaksa dosyaların tarihleri, kullanıcı bilgileri, dosya isimleri, dizin hakkındaki bilgiler veya web tabanlı veri kaynakları kullanılarak gerçekleştirilecekse web sayfaları, web adresleri gibi bilgilerin derlendiği evredir.

Doğal Dil İşleme Aşaması (Natural Language Processing): Yapay zekânın gelişimiyle birlikte dil bilimle beraber geliştirilen çalışmalar neticesinde ortaya çıkmış bir terimdir. Genel tanımıyla doğal dil işleme, Türkçe, İngilizce vb. doğal dillerdeki metinlerin, bilgisayar algoritmaları kullanılarak yazılım programlarında analiz edilmesi ve bilgisayar ortamına iletilmesidir. Özellik çıkarımı ve metinden bir takım anlamsal bilgilerin elde edilmesi, bütün metin madenciliği adımlarında kullanılmasa bile bu adımda sıklıkla başvurulur. Örneğin, konuşma parçalarının etiketlenmesi (part of speech tagging) veya cümlebilimsel parçalama (syntactic parsing) veya diğer dilbilimsel işlemler doğal dil işleme adımında yapılır. Doğal dil işleme, Türkçe, İngilizce gibi doğal dillerin kurallı yapısının ayrıştırılarak ortaya çıkarılmasını veya tekrar üretilmesini amaçlar. Bu analiz aşamasının, yazılı metinlerin otomatik olarak çevrilmesi, komut algılama ve otomatik konuşma, soru-cevap makineleri, konuşma üretimi, bilgi elde etme, kendiliğinden metin özetleme ve komut anlama, konuşma sentezi, otomatik metin özetleme, bilgi tedariği gibi pek çok konuda kolaylıklar sunacağı söylenebilir.

Varlık İsmi Tanımlama (Named Entity Recognition): Çoğunlukla metin işleme safhasında birtakım istatistiksel özelliklerin ortaya çıkarılması amacıyla kullanılır. Örneğin, metindeki şahıs isimleri, kısaltmalar, yer isimleri, semboller vb. bu metotla bulunur. Metin madenciliği araştırmaları her zaman temiz metinlerde yapılamamaktadır. Örneğin, sosyal medya üzerinde yapılan yazışmalar, facebook, twitter mesajları, telefonlardan gönderilen kısa mesajlar gibi iletilerin çoğunda kısaltmalar ve yazımdan kaynaklı hatalar bulunmaktadır. Metin madenciliği bu ihtimallerin de meydana gelebileceğinin unutulmaması gereken çalışmalardır. Adlandırılmış varlık tanıma çalışmalarında, hedeflenen kelime gruplarının metin içerisinden ayıklanıp çıkarılması, miktarının belirlenmesi için sayılması, yoğunluğunun bulunması, etiketlenmesi gibi işlemler yapılabilmektedir.

Örüntüsü Tanımlı Varlıkların Bulunması (Pattern Identified Entities): Metin içerisindeki özel bazı bilgilerin metin madenciliğine konu olması neticesinde

kullanılır. Örneğin, telefon numaraları, adresler, tarihler gibi bir takım bilgiler özel olarak elde edilmek istendiğinde kullanılır.

Eş Atıf (Coreference): Metinde bir varlığı belirten, o varlığa işaret eden (atıfta bulunan) isim, kelime grupları ve terimlerin bulunmasını ve ayrılmasını amaçlar.

İlişki, Kural, Olay Çıkarımları: Türlü nedenlerle metnin içerisinden birtakım bilgilerin çıkarılması istenebilir. Örneğin, bir çalışma sırasında, verilen bir metnin içerisindeki olayların çıkarılarak sıralanması (event ordering), Türkçedeki fiil yapılarını, olay belirten kelime gruplarını, zaman kalıplarını ve bütün bu kelime grupları arasındaki olası ilişkileri gösteren özel bir algoritma tasarlanmak istenebilir.

Duygu Analizi (Sentimental Analysis): Metinlerde geçen duygusal ifadelerin elde edilmesini amaçlar. En sık kullanılanı duygusal kutupsallıktır (sentimental polarity). Yani metinde herhangi bir konu üzerinde bahsedilen mesajların veya yazıların olumlu veya olumsuz yer almasına göre iki kategoriye ayrılması amaçlanır. Bununla birlikte duygu analizi bundan farklı olarak, metinlerdeki, kanaat, düşünce, ruh hali ve daha kompleks duyguların ortaya konulması üzerinde de çalışmaktadır.

Metin madenciliğinin uygulama alanları ise aşağıdaki şekilde sıralanabilir:

- Sağlık alanı,
- Web içerikleri sınıflandırma,
- Yazar tanıma sistemleri,
- Soru ve cevap sistemleri,
- Benzer içeriklerin tayin edilmesi,
- Müşteri ilişkileri yönetimi,
- Sahtekârlık tespiti,
- Pazar araştırmaları,
- Doküman özetleme ve sınıflamadır.

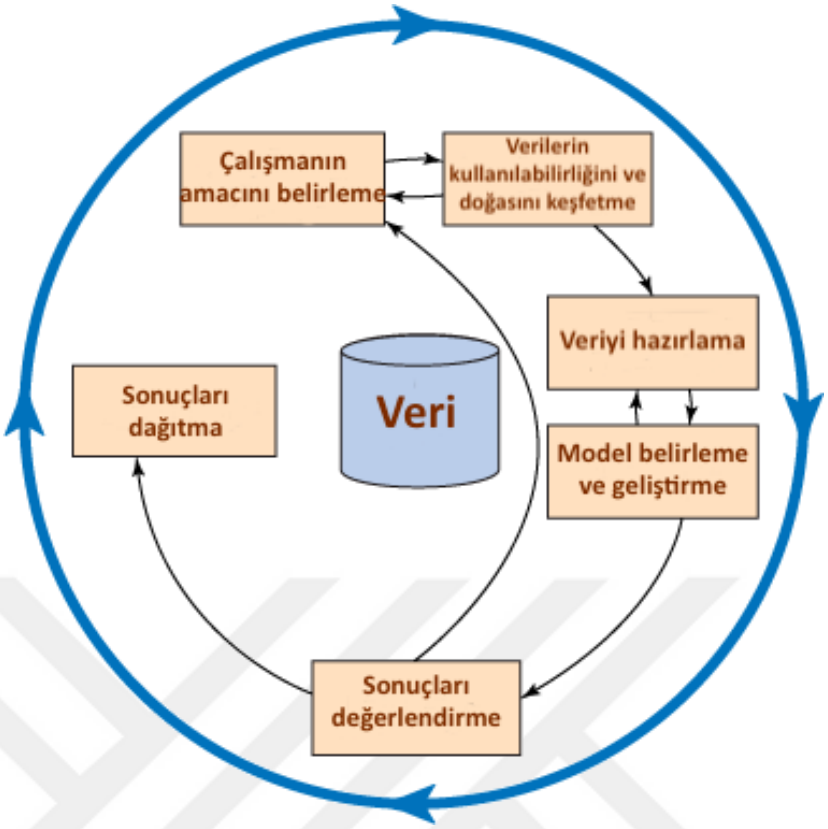
2.2.2. Metin Madenciliği Yöntemi

Metin madenciliği genel olarak beş adımdan meydana gelmektedir, ancak metin madenciliği için tam olarak kabul edilen bir süreç modeli mevcut değildir. Çalışmanın bu kısmında metin madenciliği için önerilen bir süreci anlatmaya ayrılmıştır.

Metin madenciliği farklı insanlar için farklı şeyler demektir. Hatta bunun tanımı ve kapsadığı şey çok kararsız ve tartışılabilir konulardır. Verilerin yapılandırılmamış yapısı çok farklı yelpazelerde keşfedici yollar açar. Bazıları yarı-yapılandırılmış (HTML ve XML dosyaları gibi) olmak üzere pek çok yapılandırılmamış veri türü vardır. Eldeki verilerin büyüklüğü erken örnekleme ve basitleştirme faaliyetlerini teşvik eder. Tüm bu sebepler, metin madenciliği uygulamalarında bir yöntem boşluğu olmasından kaynaklanır. Veri madenciliği yöntemleri nispeten olgunlaşırken herhangi bir alanda uygulamaların özünü yansıtan ve kabul edilen bir metin madenciliği yöntemi yoktur. Bu bağlamda, veri madenciliğinin yaygın olarak kullanılan işleyiş yöntemi olan Veri Madenciliği İçin Çapraz Endüstri Standart Süreci (CRISP-DM) tercih edilebilir[20].

Veri madenciliğinin çok kullanılan ve işleyiş süreci olan CRISP-DM, metin madenciliği için de tercih edilebilmektedir ve altı aşamalı bir döngüden meydana gelmektedir (Şekil 2.3):

1. İşi anlama
2. Veriyi anlama
3. Veriyi hazırlama (önişleme, öznitelik seçimi)
4. Modelleme
5. Değerlendirme
6. Dağıtım



Şekil 2.3. Veri madenciliği için çapraz endüstri standart süreci (CRISP-DM) ile metin madenciliği işleme süreci[25].

2.2.2.1. Çalışmanın Amacını Belirleme

Her çalışma gibi metin madenciliğinde de önce çalışmanın amacı belirlenir. Sistemin yapısını, kısıtlamalarını ve var olan kaynakları iyi bir şekilde değerlendirmek için alanında uzman kişiler ile etkileşim içinde olunması gerekir. Çalışmanın yönünü yönetmek için gerçekçi hedef ve amaçlara ancak bu şekilde ulaşılabilir.

2.2.2.2. Verilerin Kullanılabilirliğini ve Doğasını Keşfetme

Bu aşamada, metinsel veri kaynağının belirlenmesi, verilerin erişebilirliğinin ve kullanılabilirliğinin değerlendirilmesi, ilk veri kümesinin toplanması, verilerin zenginliğinin araştırılması, verilerin netlik ve kalitesinin değerlendirilmesi gibi bazı görevler uygulanmaktadır. Böylelikle çalışmanın amacı belirlendikten sonra mevcut

yapılacak çalışma için verilerin kullanılabilirliği, elde edilebilirliği ve uygulanabilirliği değerlendirilir.

2.2.2.3. Veriyi Hazırlama

Veri madenciliği ile metin madenciliği arasındaki en önemli farklar, bu aşamada ortaya çıkmaktadır. Projede kullanılacak olan veri setinin modelleme amacıyla hazırlanması, modelleme sonrasında bir kez daha veri üzerinde türlü kodifikasyonların gerçekleştirilmesini ihtiva eder ve veri hazırlama etabı birden çok kere yinelenebilir. Şekil 2.2'de yer alan veri hazırlama ve model geliştirme aşamaları, metin madenciliğinin veri madenciliğine göre içerik olarak farklılaşan aşamalarını belirtmektedir[20].

Bir metin içerisindeki sözcükleri elde etmek için genellikle dizgeciklere (token) ayırma (tokenization) işlemi gereklidir. Bu işlem ile metinde yer alan bütün noktalama işaretleri ve satır sonu karakterleriyle birlikte diğer tüm okunabilir olmayan (non-text ve non-readable) karakterler boşlukla (white space) değiştirilir. Bu da metnin bir sonraki aşama için daha elverişli ve temiz bir duruma getirilmesini sağlar. Koleksiyondaki tüm dokümanlarda dizgeciklere ayırma işlemi uygulandıktan sonra tüm dokümanlarda yer alan sözcüklerin tümü, ilgili koleksiyonun "sözlüğü"nü (dictionary) oluşturur. Korpusu yapısal hale getirebilmek için metinlerde yer alan rakamların ve noktalama işaretlerinin metinden kaldırılması gerekir. Tekrarlı boşluklar ve beyaz boşluklar da korpustan kaldırılmalıdır. Ayrıca korpus yapısı web sayfalarından ya da HTML, XML gibi formatlardan derlenmişse tablo, şekil ve resimlerden de arındırılması gerekmektedir[22].

Sözlük boyutunun, dolayısıyla da koleksiyonlardaki dokümanları temsil eden veri yapılarının (örneğin vektör uzayı modelindeki doküman vektörlerinin) boyutunun küçültülmesi için çeşitli ön işleme yöntemleri kullanılabilir.

Filtreleme (Filtering) yöntemi ile sözlükteki ve dolayısıyla dokümanlardaki sözcükler filtrelenebilir. En yaygın filtreleme yöntemi durak sözcükleri (stop words)

filtreleme yöntemidir. Buradaki amaç, tek başına bir anlam veya duygu durumu belirtmeyen ve içeriğe bir etkisi olmayan edat, bağlaç, zamir gibi kelimelerin sistemden çıkarılmasıdır. Bununla beraber, dokümanlarda sıklıkla geçen ve istatistiksel olarak bir anlam ifade etmeyen sözcükler de filtrelenebilir[23].

Temel hale döndürme (Lemmatization) yöntemleri çoğul haldeki isimleri tekil hale dönüştürmek amacıyla ya da çoğunlukla fiil çekimlerini mastar duruma dönüştürmek amacıyla kullanılır. Bu işlemin pahalı, zor ve hataya açık bir işlem olmasının nedeni, sözcüklerin cümlede yer alan konumlarını ve sözcüklerin sahip oldukları görevlerini de bilmeyi gerektirdiğinden kaynaklanmaktadır. Bu nedenle pratikte "kökenine döndürme" (stemming) yöntemleri daha fazla tercih edilebilir.

Kökenine Döndürme (Stemming) yöntemi kelimeleri basit hallerine çevirmek için kullanılır. Örneğin fiil çekim eklerinin fiilden ayrılarak fiil kökünün yalın hale getirilmesi, çoğul ekinin isimlerden atılması gibi işlemler stemming olarak adlandırılır. Türkçe için bu amaçla açık kaynak platform, bağımsız ve genel amaçlı bir Doğal Dil İşleme Kütüphanesi olan Zemberek geliştirilmiştir ve Java ile çalışmaktadır. Zemberek kullanılarak Türkçe kelimelerde stemming yapılabilir. Ayrıca birtakım üniversiteler de kendi stemming algoritmalarını geliştirmektedir[24].

Örneğin İstanbul Teknik Üniversitesi'nin İTÜ NLP, Yıldız Teknik Üniversitesi'nin Kemik adında stemming algoritmaları mevcuttur. Temizlenen korpusdan sonra modellemenin ilk kısmı için uygun olan bir temsil metodu seçilir. Metinlerin anlamsal içeriklerinden daha fazla yararlanabilmek için çeşitli teknikler geliştirilmiştir. Çoğu metin madenciliği uygulaması bir metin dokümanını metinde yer alan sözcüklerin kümesi olarak temsil etme fikri üzerinde geliştirilmiştir (bag-of-words, temsil yöntemi). Sözcüklerin doküman içindeki önemlerinin de temsil edilmesine olanak sağlayan vektörel bir temsil şekli vardır (vector representation). Bu modelin adı vektör uzayı modelidir (vector space model)[25].

2.2.2.4. Ön İşleme (Pre-Processing) Aşaması

Tüm metin sınıflama algoritmaları için ilk adım olan veri ön işleme aşaması aşağıda belirtilen sebeplerden dolayı gerçekleştirilmektedir.

- Veri üzerinde meydana gelen problemleri gidermek
- Verinin doğal yapısını keşfederek daha hassas ve nitelikli analiz yapabilmek
- Verilerden daha işe yarar ve anlamlı bilgiler üretebilmektir[26].

Türkçe dili ele alınırsa, eklemeli diller grubunda olan Türkçeye eklenen her ek o kelimenin anlamını farklılaştırdığından ve/veya değiştirdiğinden dolayı, bu dilin ön işleme aşaması zor olmaktadır. Diğer yabancı dillerden farklı olarak, Türkçe kelimelerden çok sayıda değişik anlamlı kelimeler oluşturulabilir. Bu karmaşık yapı nedeniyle, Türkçe için diğer dillerden daha farklı metin işleme teknikleri gerekebilir. Bundan dolayı, özellikle Türkçe metinlerde noktalama işaretlerinin kaldırılması ve bütün kelimelerin küçük harflere çevrilmesi dışında; joker kelimeler ve anahtar kelimelerin meydana getirilip düzenlenmesi gibi ön hazırlıklar yapılması gerekmektedir.

2.2.2.4.1 Ön İşleme Genel Adımları

Metinler, doğal yazılışları ile birlikte bir kelime vektörü olarak ifade edilemediğinden dolayı birden fazla zorluk bulunmaktadır. Mesela dokümanlarda çok sayıda kelime bulunmakta ve bu dokümanlardan da fazla miktarda bulunmaktadır. Ayrıca bu dokümanlarda çok çeşitli bilgilere yer verilmekte ve bu bilgilere yer verilirken insanlar tarafından yazılan birçok hatayı da barındırmış olmakta; noktalama işaretleri ve kısaltmalar içermektedir. Çoğunlukla metin tabanlı ön işleme teknikleri metin madenciliği operasyonları için yeterli olmaktadır. Ancak, çeşitli durumlarda dilbilimsel ön işleme yöntemleri ile terimler hakkında daha fazla bilgi sahibi olmak ve bundan yararlanmak mümkün olabilmektedir[24]. Bu nedenle ön

işleme evresi etkili bir sınıflandırma için gerekli bir adımdır. Ön işleme genel adımları aşağıdaki gibi sıralanabilir:

1- Kategoriler belirlenir ve bu kategoriler ile bağlantılı olabilecek kelimeler sözlüğe eklenir (Bu tezde … kategorilerini kullanmış olacağız)
2- Oluşturulan sözlükte her kelime teker teker incelenir. Joker (Wild Card) olarak değerlendirilebilecek kelimeler keşfedilip sözlük güncellenir.
3- Her bir doküman, joker kelimeler de dahil olmak üzere, sözlükte oluşan tüm kelimelerin boyutundaki vektörün ağırlıklandırılması ile gösterilir.

2.2.2.4.1.1. Joker (Wild Card) Yöntemi

Bu tez çalışmasında, daha çok yabancı veri kaynaklarından yararlanılacağı için, bu sorunla karşılaşılmamıştır. Ancak Türkçe veri kaynaklarından elde edilen verilerde sıklıkla karşılaşılan bir yöntem olduğundan dolayı kısaca değinecek olunursa joker yöntemi, genellikle sondan eklemeli dillerde uygulanan bir yöntemdir. Türkçe gibi sondan eklemeli dillerde, bir gövdenin sonuna farklı ekler getirilerek farklı kelimeler karşımıza çıkabilmektedir. Dolayısıyla, sistemde yer alan metinlerin içinde bulunan kelimelerin gövdeleri, kelimelerin kendilerinin yerine tercih edilmektedir. Örneğin "ev" kelimesi ile "evden", "evi", "evde" ve "evin" kelimeleri ayrı birer kelime olarak görülecekti. Bundan dolayı hem oluşturulan sözlük boyutu artacak hem de sınıflandırma başarısı düşecektir[27].

Bir takım ekler almış olmasına karşın yakın anlamda olan ve aynı söz dizimi ile başlayan sözcükleri tek bir gösterimle bir grup altında bir araya getiren kelimeler, joker kelime olarak adlandırılabilir. Gövdeleme yönteminden farklı olarak burada köke indirgeme koşulu yoktur. Kökle beraber ek de alabilmektedir[26]. Joker kelimeler, kategori belirlenmesinde yardımcı olan anahtar kelimeler ile birlikte, sık kullanılan kelimelerden de seçilebilir.

2.2.2.4.1.2 Veri Filtreleme ve Vektörün Ağırlıklandırılması

Elde edilen dokümanların üzerinde etkili bir ön işleme yapabilmek için öncelikle metinlerden noktalama işaretlerinin atılması ve tüm büyük harflerin küçük harflere çevrilmesi gerekmektedir. Daha sonra noktalama işaretleri atılmış dokümanda var olan bütün kelimeler bir diziye aktarılır. Dizideki elemanlar sözlükteki elemanlar ile karşılaştırılarak vektörün elemanlarının ağırlıkları belirlenir. Örneğin kelime dizimiz şu şekilde belirlenmiş olsun; (Children, laboratory, period). Sözlüğümüzün de aşağıdaki kelimelerden oluştuğu varsayılırsa;

child*
laborant*
active*
group*

Vektörü (1,1,0,0) olarak oluşmuş olur. Hem sınıflandırılacak dokümanların ağırlıklandırılmasında hem de eğitim dokümanlarının ağırlıklandırılmasında ve vektörlerin oluşturulmasında bu yöntem kullanılır. Fakat ağırlıklandırma yöntemi değişebilir. Bu tez içerisinde terim frekansı (TF) ağırlıklandırma yöntemi olarak kullanılacaktır.

2.2.2.4.1.3 Kelime Değerleri

Kelime değerleri, başta dokümanlar olmak üzere kelimeler ve onların ağırlıkları ile ifade edilir. Kategorizasyonun başarılı olması, ağırlıklandırmanın ne kadar iyi yapılmasına bağlıdır. Ağırlıklandırma konusu önemli bir konu olduğundan dolayı bu konu hakkında birçok teknik geliştirilmiştir. Bu teknikler/algoritmalar şu şekildedir[37]:

-Terim Frekansı (TF),
-Ters Doküman Frekansı (IDF),
-Terim Frekansı-Ters Doküman Frekansı (TF-IDF),
-Terim Ayrıştırma Değeri,
-Olasılıksal Terim Ağırlıklandırma,

-Tek Terim Doğruluğu,

-Genetik Algoritmalardır.

Bu algoritmaların basitçe kavranabilmesi ve daha iyi anlatımı için aşağıdaki gibi bir Türkçe örnek eğitim dokümanı belirlenmiştir.

1- İlaç firmalarının yaptıkları araştırmalarda, sonbaharda grip salgını ve gribe bağlı hastalıklarda artış gözlenmektedir (Sağlık).
2- Mevsim geçişlerinde ilaç satışları artmaktadır (Sağlık).
3- Yıllık enflasyon oranı bu sene de yükselişte (Ekonomi).
4- Tarımda makineleşme, tarımla uğraşanlar için çok büyük kolaylık yaratmıştır (Ekonomi).
5- Hakemin verdiği penaltıyı hatalı kullandı, hakem penaltıyı tekrarlattı (Spor).
6- İlaç gibi gelen erken golden sonra taraftarlar çok sevindi ve taraftarlar bol tezahürat yaptı (Spor).

Vektörler oluşturulurken joker kelimeler de dahil olmak üzere kelimeler sırası ile dokümanlarda aranıp bulunan kelime sayısı boyuta eklenmelidir.

Kategorisinin bulunması istenen metin; “Taraftarlar hakeme tepki gösterdiler. Hakem sahayı terk etti.” olsun.

Sözlük grubu: {enflasyon*, grip*, hakem*, ilaç*, taraftar*, tarım*}

Eğer vektörler kelime frekanslarına göre şu şeklinde ifade edilirse;

D1=(0,2,0,1,0,0)

D2=(0,0,0,1,0,0)

D3=(1,0,0,0,0,0)

D4=(0,0,0,0,0,2)

D5=(0,0,2,0,0,0)

D6=(0,0,0,1,2,0)

DQ=(0,0,2,0,1,0)

Eğer vektörler bitsel şekilde ifade edilirse;

D1=(0,1,0,1,0,0)

D2=(0,0,0,1,0,0)

D3=(1,0,0,0,0,0)

D4=(0,0,0,0,0,1)

D5=(0,0,1,0,0,0)

D6=(0,0,0,1,1,0)

DQ=(0,0,1,0,1,0)

2.2.2.5. K En Yakın Komşuluk (K Nearest Neighbor, K-NN) Algoritması ve Vektör Uzay Modeli

K en yakın komşuluk algoritması sorgu vektörünün, en yakın K komşuluğundaki vektör ile birlikte sınıflandırılmasının bir neticesi olan denetimli öğrenme algoritmasıdır. Doküman vektörü ve eğitim denetlemeleri vektörleri, yeni bir vektörü sınıflandırabilmek için kullanılır. Herhangi bir sorgu örneğindeki sorgu

noktasına en yakın K tane eğitim noktası bulunur. Sınıflandırma da bu K tane objenin en çok bulunanı ile gerçekleştirilir[28]. K en yakın komşuluk uygulaması yeni sorgu örneğini sınıflandırmak amacıyla kullanılan bir komşuluk sınıflandırma algoritmasıdır.

K en yakın komşulukları bulup ortaya çıkarmak için mevcut olan sorgu örneği ile birlikte eğitim dokümanları arasındaki en az uzaklıklar dikkate alınır. En yakın komşuluklar keşfedildikten sonra bunlardan kategorisi fazla olan, dokümanın kategorisini kestirmek için kullanılır.

Avantajları: Uygulanabilirliği oldukça kolay bir algoritmadır. Gürültülü eğitim dokümanlarına karşı oldukça dayanıklıdır. Eğitim dokümanlarının sayısı fazla ise etkili olmaktadır.

Dezavantajları: K parametreye ihtiyaç vardır. Uzaklık bazlı öğrenme algoritması kullanıldığında, bu algoritmanın en iyi sonuca ulaşmak için, hangi uzaklık çeşidinin ve hangi özelliğin kullanılacağı konusunda herhangi bir netlik olmamasıdır. Her bir sorgu örneğinin tüm eğitim örneklerine olan uzaklığı hesaplandığı için hesaplama maliyeti oldukça fazladır.

Bu algoritma, en yakın komşu ilkesine dayanır ve tüm dokümanlar vektörel olarak temsil edilir. Diğer dokümanlar ile sorgu dokümanı arasındaki kosinüs benzerliği hesaplanır. Similatry oranı 1'e en fazla yaklaşan n adet vektörün fazla olanı dokümana atanır.

$$d_i = \left(wd_{i1}, wd_{i2}, \ldots\ldots\ldots\ldots, wd_{ij}\right) \qquad (2.1)$$

w_{ij} terimin doküman içerisindeki ağırlığı, d_i eğitim dokümanı vektörüdür. q ise bulunduğu sınıfın belirlenmesi istenen vektördür.

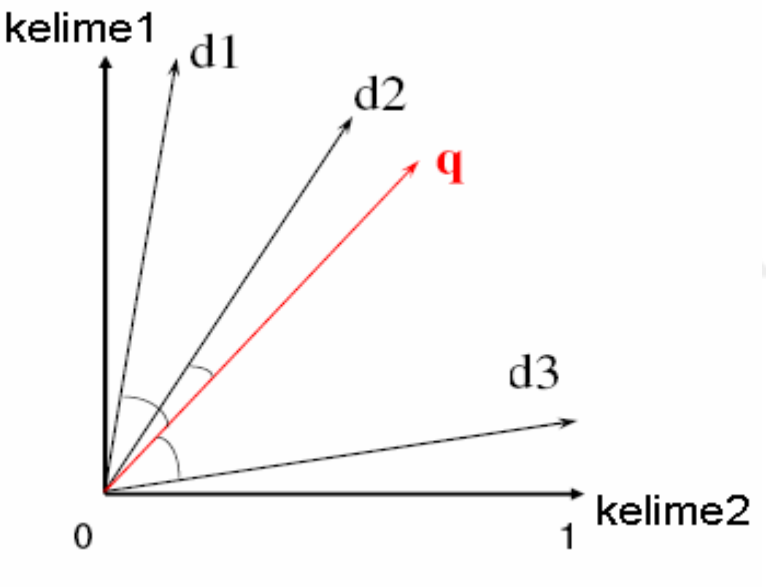
$$sim(d_i,q) = \cos\theta \qquad (2.2)$$

$$\text{sim}(d_i,q) = \frac{d_i \cdot q}{|d_i||q|} = \frac{\sum_j W_{i,j} * W_{q,j}}{\sqrt{\sum_j W_{i,j}^2}\sqrt{\sum_j W_{q,j}^2}} \quad \frac{d_i \cdot q}{|d_i||q|} = \frac{\sum_j W_{i,j} * W_{q,j}}{\sqrt{\sum_j W_{i,j}^2}\sqrt{\sum_j W_{q,j}^2}} \qquad (2.3)$$

$\text{sim}(d_i,q)=1 \Rightarrow d=q$

$\text{sim}(d_i,q)=0$ ise terim paylaşımı yoktur.

Hangi sınıfa ait olduğunun bilinmesi istenen doküman ve tüm dokümanlar bu kurallar doğrultusunda vektörel olarak gösterilirler (Şekil 2.4). Burada aslında her bir boyut, kelimeleri ifade etmektedir.



Şekil 2.4. Vektör uzay modelinde dokümanlar[37].

Burada d1, d2 ve d3 eğitim dokümanlarından oluşan vektörler, q ise sınıfı bulunmak istenen vektördür.

Bu model kelimelerin doküman içerisindeki önemlerinin temsil edilmesine yarayan vektörel bir temsil şeklidir. Büyük boyutlardaki veri dokümanlarının önemli ölçüde dizinlenmesini ve veri analizinin etkin bir şekilde yapılması için kullanılır. Burada her bir obje/nesne, vektör olarak tanımlanmaktadır. Vektör uzayının eksenlerini,

tanımlanan bu objelerin sahip oldukları farklı nitelikler oluşturmakta ve her bir obje sahip olduğu niteliklere göre vektör uzayında belirli bir konuma sahip olmaktadır[29]. Özellik vektör uzayı, doküman sınıflandırma çalışmalarında kullanılmakla beraber, sözcüklerin dokümanlardaki ortaya çıkma sıklıklarına dayanmaktadır. Yani her bir doküman içinde yer alan sözcüklerin dokümanlardaki frekansları hesap edilerek sözcük vektör uzayı meydana getirilir[30]. Bir metnin, vektör uzay modelinde gösterimi amacıyla farklı üç metot kullanılmaktadır:

Binary Vektör: Bu yöntem ile metinsel veriler 1 ve 0 olarak belirtilip ifade edilmektedir. Veri içerisinde yer alan kelimeler sözlükteki mevcudiyetlerine göre bu değerleri almaktadırlar[31]. Veri setindeki kelimelerin sahip olacağı değerler binary vektör temsilinde {1,0,0,1...} şeklinde olmaktadır.

Frekans Vektör: Binary tanımlamasından farklı olacak şekilde veri içerisinde mevcut olan kelime köklerinin kaç defa kullanıldığı bilgisinin de ele alınarak yapıldığı bir tanımlama biçimidir[31]. Veri setindeki kelimelerin sahip olacağı değerler frekans vektör gösteriminde {2,0,3,1...} şeklinde olmaktadır.

Terim Frekansı (TF-term frequency) - Ters Doküman Frekansı (IDF-inverse document frequency) Vektör: Terim Frekansı - Ters Doküman Frekansı (TF-IDF) ağırlıklandırmasında her bir dokümandaki kelimelerin frekansı etkili olmaktadır. Terim Frekansı (TF) değeri frekans bilgisini yani terimin veri setinde kaç defa geçtiğini hesaplar. Ters Doküman Frekansı (IDF) ise tüm dokümanlarda nadir olarak geçen kelimelerle alakalı bir ölçü verir. Kelimenin bir doküman için belirleyici bir özelliğinin olması, kelimenin tüm eğitim dokümanları incelendiğinde yalnızca o dokümanda geçmesine bağlıdır. Kısacası kelimenin o doküman için belirleyici özelliği var olmuş olur[32].

Eşitlik (2.4) ve (2.5)'te sırasıyla TF ve IDF hesaplamaları verilirken, eşitlik (2.6)'da ağırlık hesaplaması verilmiştir.

$$TF_{ij} = \frac{n_{ij}}{|d_i|} \qquad (2.4)$$

$$IDF_{ij} = \log\left(\frac{n}{n_j}\right) \quad (2.5)$$

$$W_d = TF_{ij} \times IDF_{ij} \quad (2.6)$$

TF değerinin hesaplanmasında kullanılan *n* değeri, *j* nci kelime kökünün toplanan *i* nci veri seti içinde kaç kez geçtiği sayıyı ifade eder. *d* değeri ise veri seti içerisinde yer alan bütün kelime köklerinin sayısını belirtir. Formül içinde yer alan *i* değeri ise e-posta içerisinde bulunan kelimelerin sayısıdır. IDF değerinin hesaplanması için kullanılan *n* değeri toplam belge miktarının n_j ise *j*. terimin göründüğü belgelerin miktarını ifade eder. Ağırlıklandırma ise bu iki değerin çarpımı suretiyle ortaya çıkar[30].

Vektör uzayı modelinde doküman ve sorgular m-boyutlu vektörlerle temsil edilirler. Burada m sözlükteki terim sayısıdır. Vektör uzayı modelinde her bir doküman sayısal bir öznitelik vektörüyle temsil edilir: w(d) = (w(d,),...,w(d,)). Vektörün her bir boyutunda ilgili terimin dokümanlardaki ağırlığı da yer almaktadır[20].

2.2.2.6. Model Belirleme ve Geliştirme

W(d) içeriği eşitlik (2.6) aracılığıyla elde edildikten sonra benzerlik ölçüleri hesaplanır. Eğer metin madenciliğinde kümeleme algoritmalarının kullanılması amaçlanmışsa, iki doküman arasındaki benzerliğin ölçülmesi gerekmektedir. Kümeleme analizinde benzerlik hesaplamak için çeşitli ölçüm yöntemleri mevcuttur fakat metin madenciliğinde doküman kümelemesi için Cosine ölçüsü kullanılmaktadır[23].

$$Cosine\,(d_1, d_2) = (d_1 \cdot d_2) \,/\, \|d_1\| \cdot \|d_2\| \quad (2.7)$$

Eşitlik (2.7)'de $(d_1 \cdot d_2)$, d_i vektörlerinin çarpımını, $\|d\|$ ise d_i vektörünün uzunluğunu ifade etmektedir.

Benzerlik ölçüleri haricinde, uzaklık ölçüleriyle de kümeleme modeli kurulabilir. En çok kullanılan uzaklık ölçüsü Öklit (Euclidean) ölçüsüdür. Bu ölçüt, (2.8) eşitliği ile hesaplanmaktadır.

$$dist\ (d_1, d_2) = \sqrt{\sum_{k=1}^{n} |\mathrm{w}(\mathrm{d}_1\ ,\ \mathrm{t}_k)| - |\mathrm{w}(\mathrm{d}_2, \mathrm{t}_k)|^2} \qquad (2.8)$$

Uzaklık ölçüsü belirlendikten sonra modellemeye geçilir. Metin madenciliğinde elde edilen terim frekans matrisinin oldukça büyük olması hiyerarşik bir kümeleme algoritmasının kullanılmasına engel oluşturur. Bu nedenle Mac Queen tarafından geliştirilen K-ortalamalar (K-means) algoritması tercih edilmektedir.

K-ortalamalar, özellik çıkarımı yapılmış bir grup verinin hangi kümeye ait olduğunun birden fazla küme özelliği kullanılarak bulunmasıdır. Her verinin sadece bir kümede yer almasına imkan tanınır. Bu nedenle keskin ve değişmez bir algoritmadır. Kullanılan matematiksel yöntem, yeni kümelerin her bir sınıf için merkez olarak belirlenen noktaya uzaklığa göre bu kümelerin yerleştirilmesidir.

Algoritma temelde şu dört basamaktan meydana gelir:

1. Küme merkezlerinin tayin edilmesi
2. Merkez dışındaki örneklerin uzaklıklarına göre gruplandırılması
3. Yapılan gruplandırmaya göre yeni merkezlerin tayin edilmesi (veya eski merkezlerin yeni merkeze kaydırılması)
4. İstikrarlı duruma (stable state) gelinceye kadar 2. ve 3. adımların yinelenmesi olarak ifade edilebilir[12,13].

2.2.2.7. Sonuçları Değerlendirme

Sonuçların paylaşılma aşamasından önce modellerin oluşturulup, tüm işlemlerin doğru bir şekilde yürütüldüğünün sağlamasını yapmak gerekmektedir. Bu

doğrulamayı yaptıktan sonra sonuçların paylaşılması aşaması gerçekleştirilebilir. Sürecin bu şekilde geniş ve kapsamlı olarak değerlendirilmesi, karar verme sürecinde hata yapılma ihtimalini en aza indirip, geri dönüşü olmayan zararlara yol açmasının önüne geçilebilir.

2.2.2.8. Sonuçların Sunulması

Sunum aşaması, modeller ve modelleme sürecinin başarı ile gerçekleşmesinden sonra geçilen son adımdır. Bu aşamada sunulan modeller karar vericilere hitap edecek kadar basit ya da kompleks olabilir. Daha iyi bir karar süreci ortaya koyabilmek için, model sonuçları defalarca kullanılabilir. Model sonuçlarının yeni verilerle periyodik olarak güncellenmesi, bu oluşturulan modellerin zaman içerisinde doğruluk ve uygunluğunun kaybolmasından ileri gelmektedir. Dolayısıyla bu yenileme işlemi; sürekli olarak yeni bir model yaratıp, yeni bir analiz süreci başlatmaktan çok daha kazançlı olacaktır[20].

3. GEREÇ ve YÖNTEM

Bu tez çalışmasının amacı, metin madenciliği yöntemini ve uygulama adımlarını tanıtmak, ve sağlık alanında belirlenen bir konuda uygulamasını göstermektir. Çalışmanın uygulama aşamasında; "insanlarda görülen kanser vakaları (cancer cases in humans)" ve "farelerde kanser araştırmaları (cancer research in mouse)" şeklinde belirlenen iki farklı konu başlığı altında en sık kullanılan Pubmed veritabanından ayrı ayrı elde edilen dokümanlar birleştirilerek, bu dokümanlara metin madenciliği yönteminin uygulanmasına, Knime programının metin madenciliğinde nasıl kullanıldığına ve elde edilen dokümanlara uygulanan adımların neler olduğuna ayrıntılı olarak yer verilecektir.

Bu bölümde çalışmanın uygulama aşaması açıklanmaktadır. Uygulama esnasında, analiz edilen dokümanlar Knime programı aracılığıyla analiz edilmiştir.

3.1. Knime Yazılımı

Knime, Düğüm Havuzu (Node Repository) altında yer alan Düğümler (Node'lar) arasında ilişkilendirmeler yapılarak verinin işlenmesi, yorumlanması, görselleştirmesi ve raporlanmasını sağlayan, workflow mantığıyla çalışan açık kaynak kodlu bir veri analiz platformudur.

Knime yazılımında düğüm (node) adı verilen kutucuklar vardır ve bunlar birbirine bağlanarak bir akış diyagramı oluşturulur. Analiz etmek istenilen veri sisteme okutulduktan sonra uygulamak istenilen düğümler sırasıyla seçilir ve akış gerçekleştirilerek veri kolayca analiz edilir.

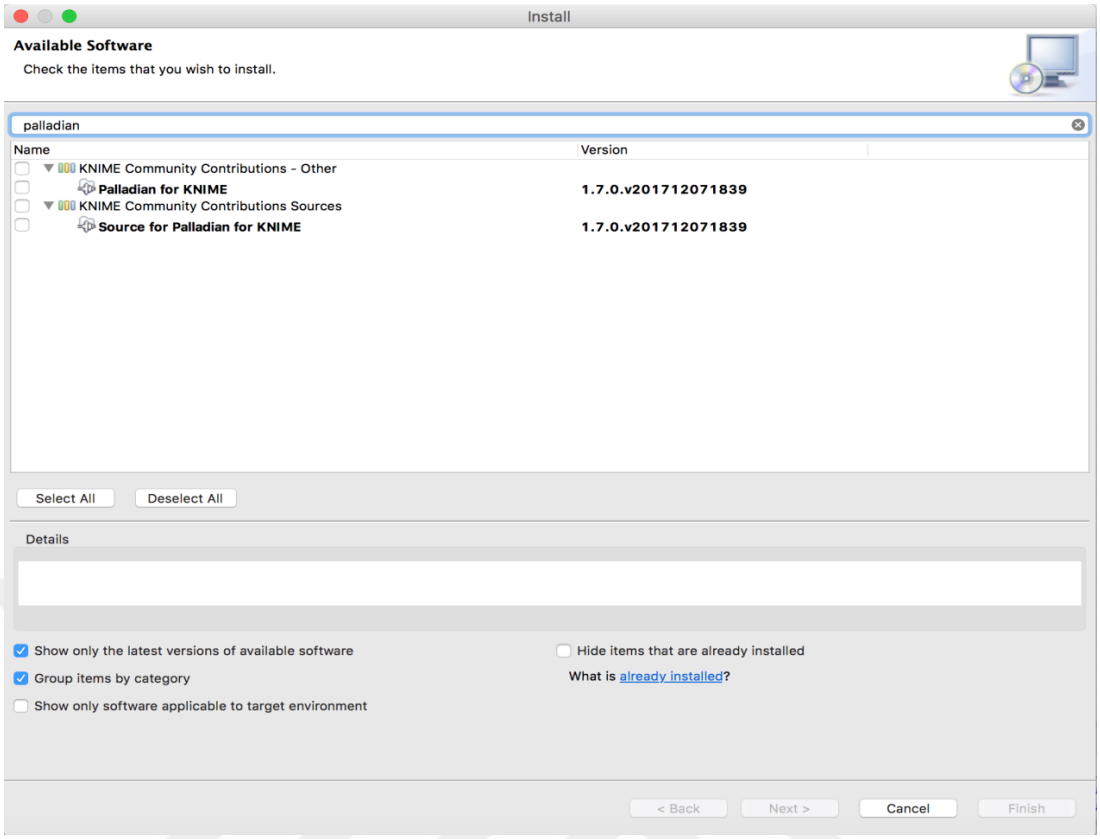
Java ile yazılmış ve Eclipse tabanlı kurulmuş olan Knime, mevcut sabit disk alanıyla sınırlı olan Büyük Veri (Big Data) süreçlerinde de kullanıma uygun olarak tasarlanmıştır. Ağırlıklı olarak; müşteri ilişkileri yönetimi, iş zekası süreçlerindeki

veri analizi, sosyal medya, ilaç/sağlık, üretim, finans, sosyal medya, e-ticaret vb. alanlarda çeşitli analitik çözümler için kullanılabilir.

Modül olarak ifade edilebilen işlem hattı yapısıyla Knime, makine öğrenimi ve veri madenciliği ihtiyaçlarına yönelik birçok bileşene sahiptir ve bu bileşenler uygulama içerisinde "düğüm (node)" olarak ifade edilir. Açık kaynak kodlu yapısı sayesinde Knime node yapısı geliştirilebilmekte ve ihtiyaçlara uygun şekilde özelleştirilebilmektedir. Knime, var olan node yapıları vasıtasıyla Weka, Tableau ve Rapid Miner gibi diğer veri analizi ve makine öğrenimi uygulamaları ile kolaylıkla entegre olabilmektedir. Bu anlamda ARFF formatını tanımakta, C, C++, R, Python, Java ve JavaScript kodlarının kullanımına izin vermektedir. İhtiyaçlara uygun olarak Windows, Linux, macOS, gibi işletim sistemleri için sunulan bu uygulama http://www.knime.com internet sitesi üzerinden tamamen ücretsiz bir şekilde elde edilebilmektedir[36].

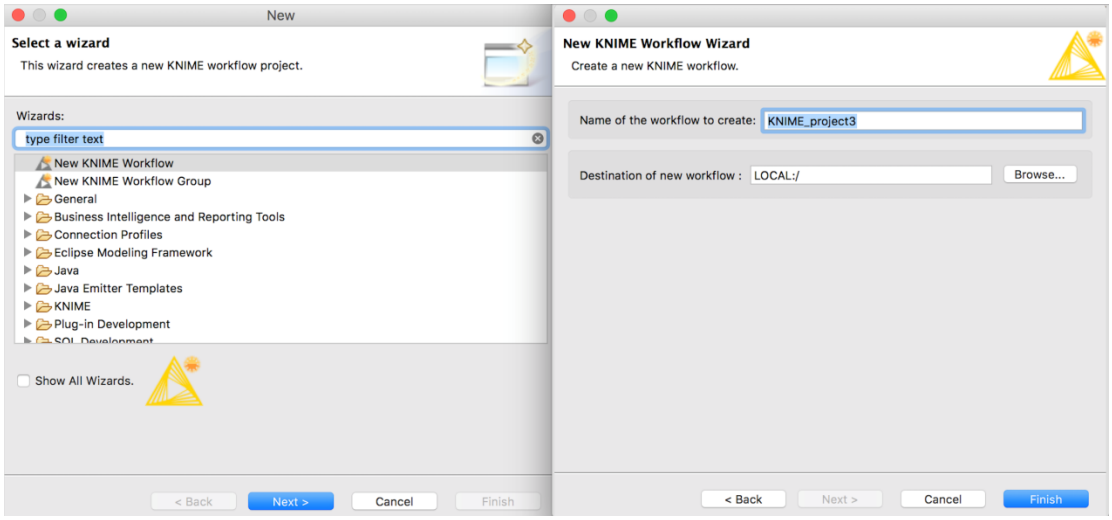
3.1.1. Knime ile Workflow Oluşturma

Knime'da metin madenciliği uygulamalarını kullanabilmek için "palladian" isimli mini yazılımın Knime üzerinde kurulması gereklidir. Palladian, java üzerinde geliştirilmiş ve internet üzerinden download yapmayı sağlayan bir ek yazılımdır. Bunun için "File" menüsünden "Install Knime Extensions" seçilerek arama kısmına "palladian" yazılarak bu ek program kurulabilir (Şekil 3.1).

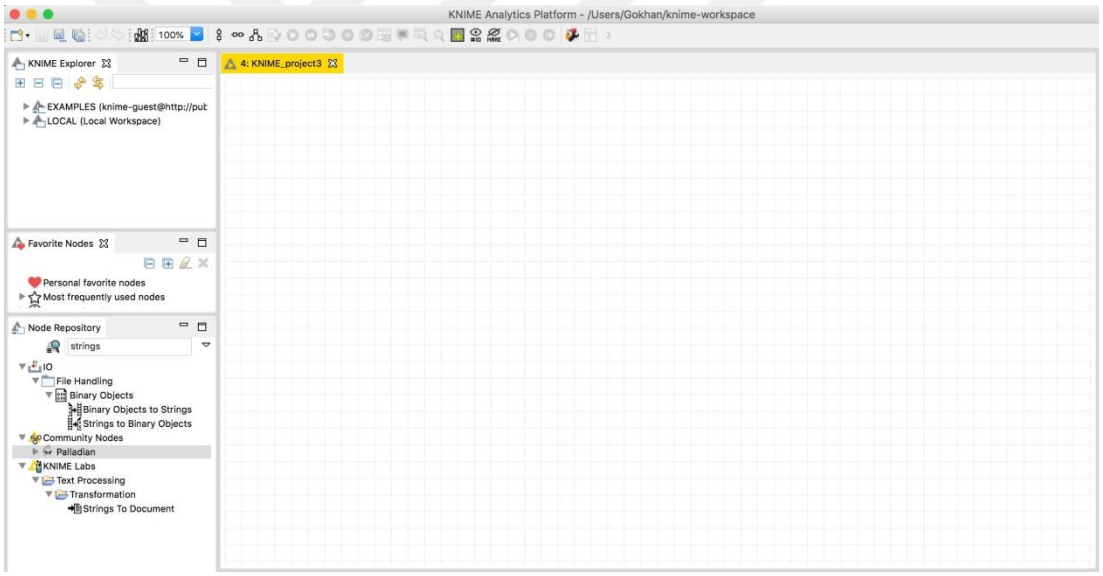


Şekil 3.1. Palladian yazılımı arama penceresi.

Daha sonra, Knime üzerinde gerçekleştirmek istenen proje için yeni bir workflow, yani çalışma penceresi Şekil 3.2'deki adımlar takip edilip açılarak burada yapılmak istenen proje, node adı verilen ve her birinin farklı bir görevi olan kutucuklar sayesinde gerçekleştirilir. Bu yeni workflow, File – New – New Knime Workflow sekmeleri takip edilerek, yeni bir proje adı verilerek oluşturulur. Ardından Şekil 3.3'te görüldüğü gibi ekrana gelen yeni workflow sayfasında proje adımları gerçekleştirilir.



Şekil 3.2. Knime'da yeni workflow oluşturma pencereleri.



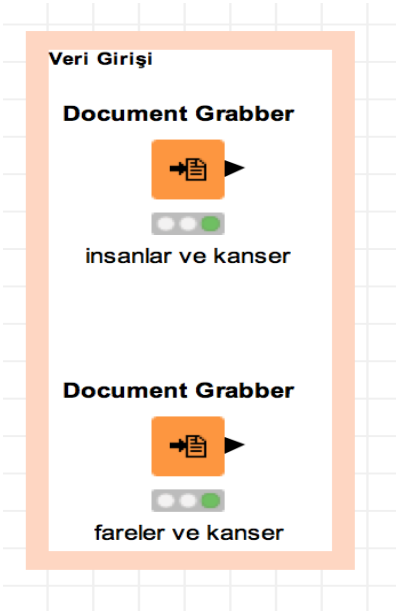
Şekil 3.3. Yeni oluşturulmuş workflow sayfası.

Burada yapılan çalışmada insanlarda görülen kanser vakaları ile alakalı Pubmed'ten indirilmiş dokümanlar ile farelerde kanser araştırmaları ile alakalı yine Pubmed'ten elde edilen dokümanların birleştirilmesiyle bu dokümanlara sırasıyla metin madenciliği teknikleri uygulanarak, Knime programının metin madenciliğinde nasıl kullanıldığını ve elde edilen dokümanlara uygulanan adımların neler olduğuna değinilecektir. Sonuç itibariyle Pubmed'ten indirilen bu iki dokümanda da en fazla

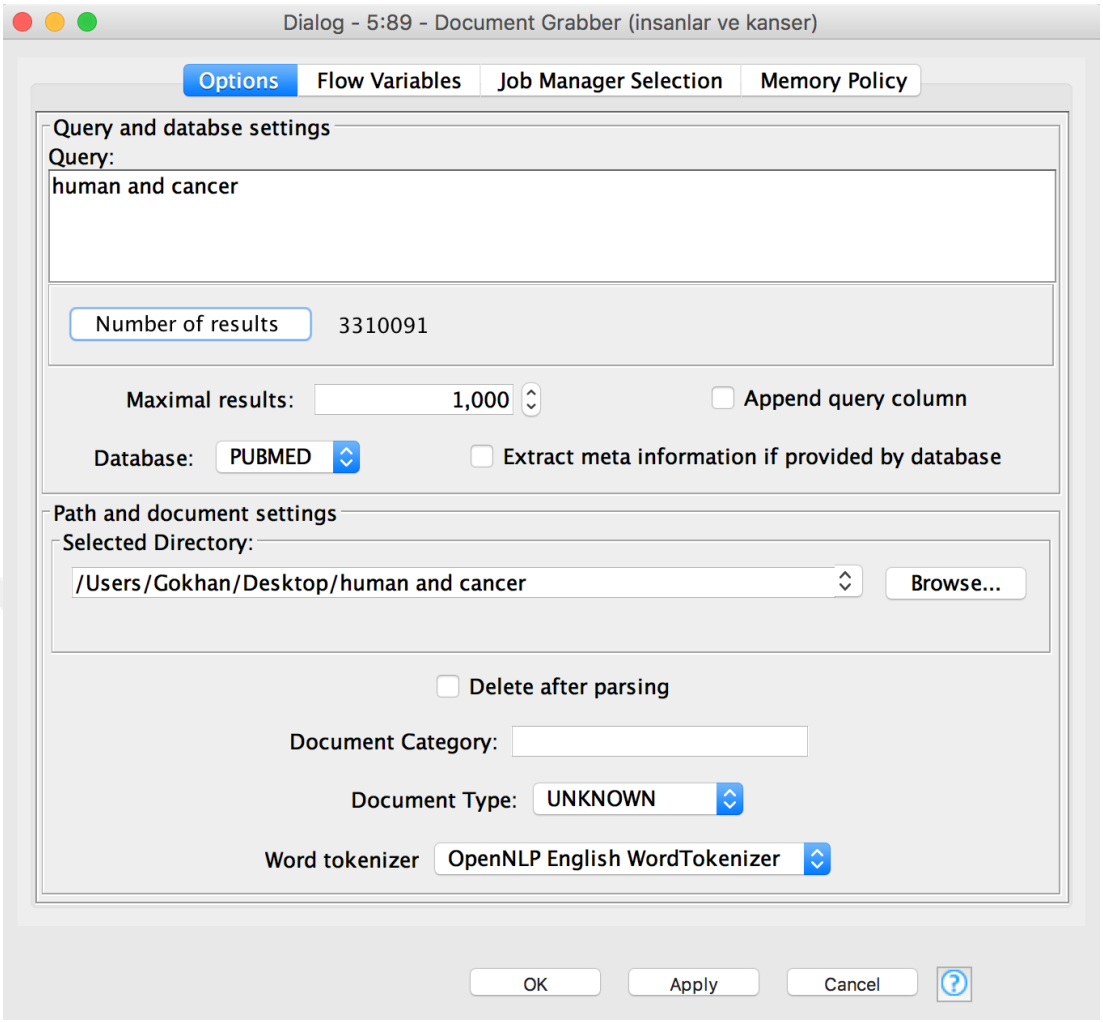
kullanılan kelimelerin neler olduğu araştırılıp farklı filtreleme yöntemleri kullanıldığında bu en çok kullanılan kelimelerin değişip değişmediği incelenecektir.

3.2. Veri ve Veri Ön İşlemleri

Öncelikle Document Grabber düğümü kullanılarak anahtar kelimeler vasıtasıyla Pubmed veri tabanından metinsel doküman verileri elde edilir ve önceden bilgisayarda açılmış olan boş klasöre bu veriler kaydedilir ve kullanılmaya hazır hale getirilir. Şekil 3.4'deki Document Grabber düğümünün şeçenekler penceresinde, "query" kısmına belirtilen veri tabanında aranacak anahtar sorgu kelimesi yani keywords'ler yazılır. "Number of results" butonuna tıklandığında elde edilen sonuçların miktarı görünmektedir. İnsanlar ve kanser dokümanları için Pubmed veri tabanında üç milyona yakın sonuç çıkmış, "maximal results" sekmesiyle bu sonuçları bin adetle sınırlandırılması gerçekleştirilmiştir. Daha sonra kelime tokenizasyon yöntemi, doküman tipi ve kategorizasyonu ile elde edilecek dokümanların saklanacağı klasör oluşturulup veriler işlenmeye hazır hale getirilmiştir. Fareler ve kanser verilerinin elde edilmesi için de aynı işlemler tekrarlanmıştır.

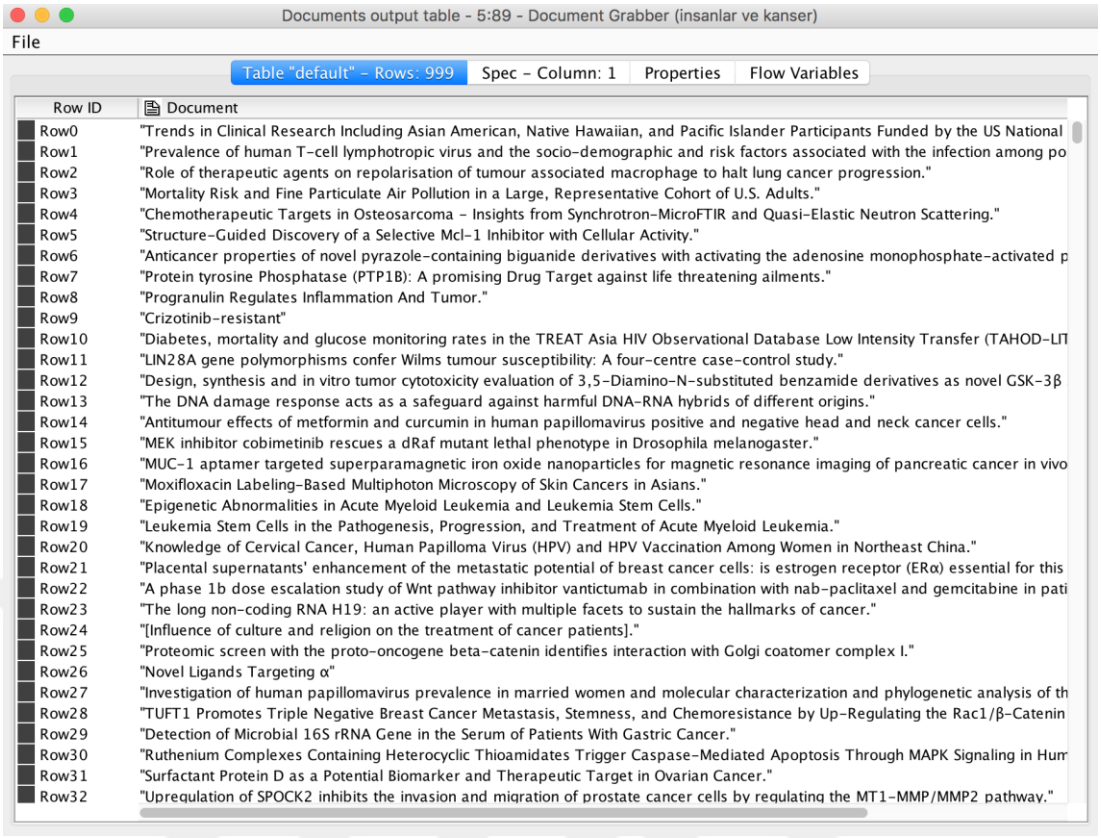


Şekil 3.4. Document Grabber düğümlerinin workflow'da gösterimi.



Şekil 3.5. Documant Grabber seçenekler penceresi.

Sonra elde edilen bu dokümanlar yine Document Grabber düğümü kullanılarak Knime workflow sayfasına aktarılır ve ön işleme hazır halde bulunur. "Output table" butonuna tıklandığında Şekil 3.6'da bin adet dokümanın satırlar halinde işleme hazır halde olduğu görülmektedir.



Şekil 3.6. İnsanlar ve kanser ile alakalı elde edilen doküman verisi.

Verileri içerisinde barındıran Document Grabber düğümü verileri işleme sokmak için veri akışını temsil eden oklar ile adlandırılmış varlık tanıma aşaması için POS tagger ve Abner tagger düğümlerine aktarılır. Bu düğümler ile birlikte veriye zenginleştirme (enrichment) yapılır. Zenginleştirme kategorisi konuşma etiketlerinin bir kısmını atayan ve standart adlandırılmış varlıkları tanıyan düğümleri içerir. Bunlara örnek olarak kişilerin, kuruluşların veya konumların adları, genlerin veya proteinlerin adları ve kimyasal yapılar gibi biyomedikal olarak adlandırılan varlıklar verilebilir.

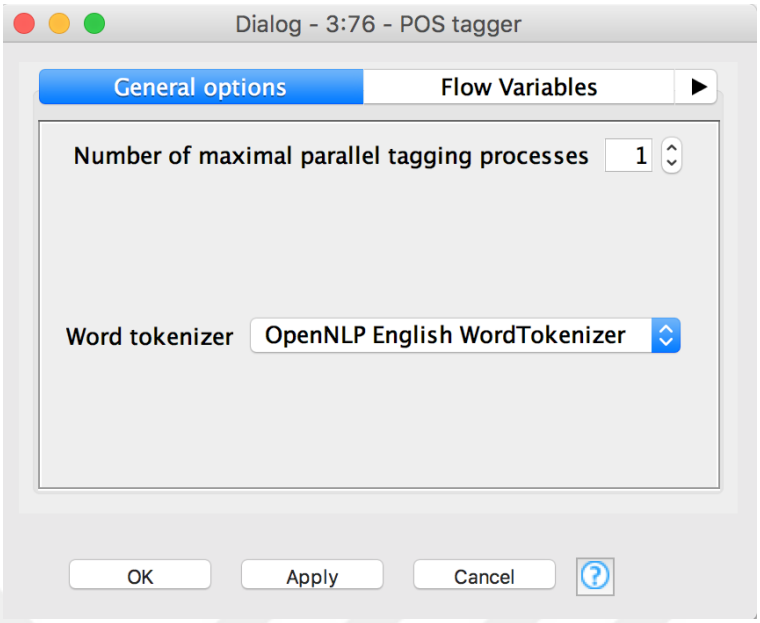
Her tanınan adlandırılmış varlığa bir etiket değeri atanır. Etiket türü, etki alanını veya bir etiketleyicinin türünü temsil eder (biyomedikal adlandırılmış varlıklar veya kimyasal adlandırılmış varlıklar gibi). Değer, o alandaki belirli bir özelliği temsil eder (örneğin, biyomedikal alanındaki gen gibi).

POS Tagger (Kelime Türü); bir kelimenin kökü bulunduktan sonraki aşama kelimenin türünün bulunmasıdır. Bu işleme Pos Tagging denir. Pos Tagging iki aşamadan meydana gelir.

Birincisi eğitim (training) aşamasıdır. Bu aşamada kelimelerin kökleri manuel olarak tanımlanmış algoritmalar kullanılarak machine learning sistemi vasıtasıyla işlenir. İkinci aşama ise tagging aşamasıdır. Bu aşamada, ilk adımda kullanılan algoritma, öğrenilen parametrelere göre yeniden işlenir ve kelimeler türlerine ayrılır.

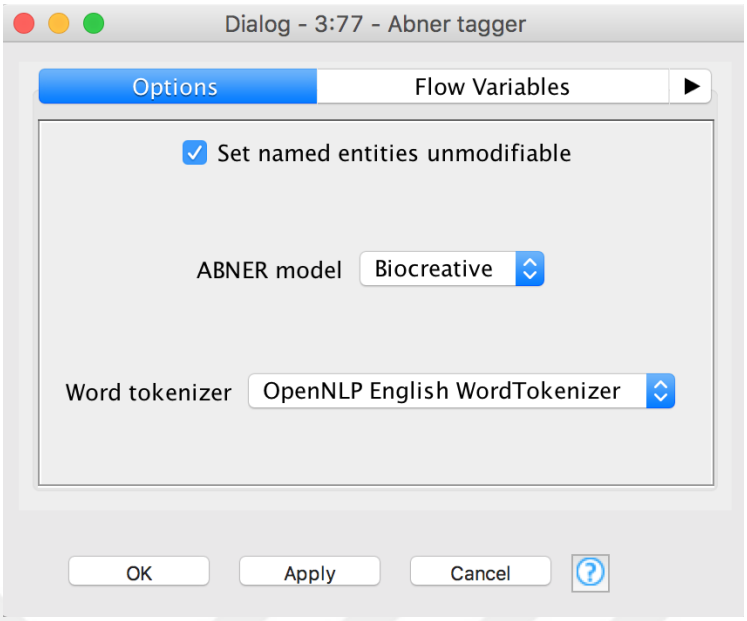
Kısaca POS tagger düğümü konuşma etiketlerinin bir bölümünü atar. Her bir etiketleyici genellikle kendi etki alanının etiketlerini atar ve bu nedenle kendi etiket türünü ve değer kümesini kullanır. Bu etiket türlerine ve değerlerine bağlı olarak, filtreleme daha sonra uygulanabilir, böylece adlandırılmış varlıklar çıkarılabilir ve görselleştirilebilir.

POS tagger düğümüne iki kere tıklandığında açılan pencereden hangi terime hangi etiketin atanacağına karar veren Tokenizer seçenekleri görüntülenir. Tokenizer; metni kelime bazında parçalara bölme, dizgeciklere ayırma demektir. Bu çalışmada Pubmed'ten elde edilen İngilizce dokümanlar kullanıldığı için Açık Doğal Dil İşleme (OpenNLP) English Word Tokenizer segmesi seçilmiştir.

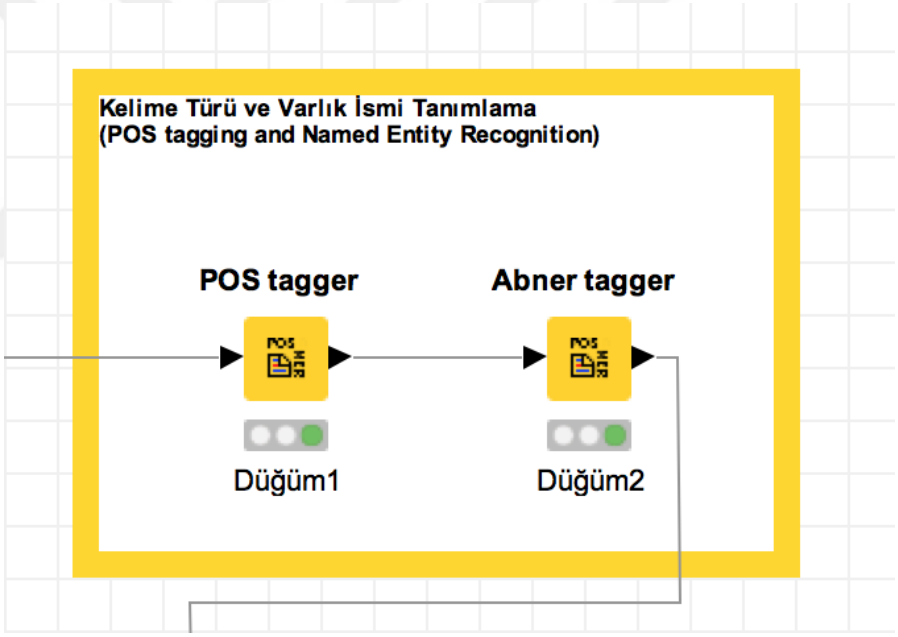


Şekil 3.7. POS Tagger genel seçenekler penceresi.

Abner tagger ise genler, proteinler veya hücreler gibi biyomedikal olarak isimlendirilen varlıkları tanır ve "Protein", "RNA", "DNA", "Hücre Hattı" veya "Hücre Tipi " gibi karşılık gelen terimlere etiketler atar. Ayrıca, adlandırılmış varlıkları (named entities) değiştirilemez olarak işaretlenebilir, yani daha sonra herhangi bir düğüm tarafından değiştirilmeyecekleri belirtilebilir. Abner Tagger düğümüne iki kere tıklandığında açılan pencereden kullanılacak ABNER model seçilebilir. Örneğin; Biocreative modeli sadece proteinleri tanır, NLPBA modeli de hücreleri, DNA ve RNA hatlarını tanıyıp etiketler.



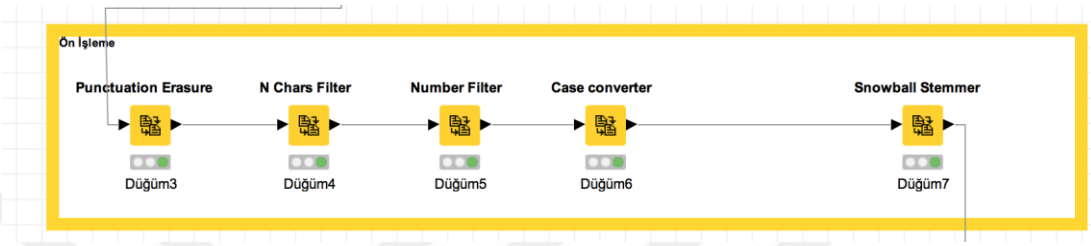
Şekil 3.8. Abner tagger seçenekler penceresi.



Şekil 3.9. Pos tagger ile Abner tagger düğümlerinin birbirine bağlanması.

Bu iki düğümün birbirine bağlanmasından sonra Adlandırılmış Varlık Tanımlama (Named Entity Recognition) süreci de gerçekleşmiş olur. Adlandırılmış Varlık Tanımlama; metinlerde daha önceden çıkarılmış veya elde var olan bilgileri kullanarak istenilen varlıkları tanımlama işlemidir. Aynı zamanda elde edilen varlık isminin, elde daha önceden var olan bir bilgiye göre neye ait olduğunun bulunmasını da kapsar.

Ön İşleme (Pre-Prossesing): Metni kelimelere ayırma, kelimelerin anlamsal değerlerini bulmak, kelimeleri köklerine ayırma ve gereksiz kelimeleri ayıklama, yazım kurallarına uygunluğunu tespit etme ve var olan hataları düzeltme gibi metin belgelerin yapıtaşı olan kelimelerle ilgili işlemleri içeren süreçtir. Knime'da uygulanan metin ön işleme şeması Şekil 3.10'daki gibidir.



Şekil 3.10. Knime'da metin ön işleme ve uygulanan düğümler.

Punctuation Erasure; dokümanda bulunan terimlerin tüm noktalama işaretlerini kaldırır.

N Chars Filter; dokümanda yer alan tüm terimleri belirtilen N sayıda karakterden daha az olarak filtreler.

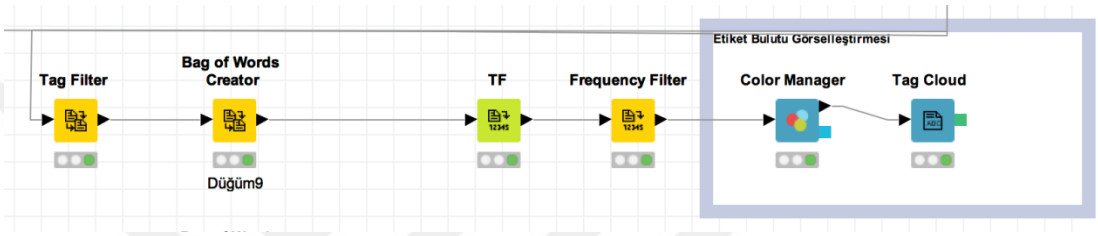
Number Filter; dokümanda yer alan ve rakamlardan oluşan ve ondalık ayırıcılar "," ve "." dahil olmak üzere tüm terimleri filtreler.

Case Converter; dokümanda yer alan tüm terimleri küçük veya büyük harfe dönüştürür.

Snowball Stemmer; bilgi edinme alanında kullanılmak üzere gövde algoritmaları oluşturmak için tasarlanmış küçük bir dize işleme dilidir. Yani Snowball Stemmer bir sözcük kökü ayırıcıdır. Sözcükleri anlamlı gövdelere yani köklere ayırır. Kök bulmak için tasarlanmış küçük bir karakter işleme dilidir. Snowball kullanılarak birçok dil için kök bulma algoritmaları geliştirilmiştir. Bu uygulamada kullandığımız dokümanlar İngilizce dilinde olduğundan dolayı gövdeleme işlemi (stemming) için sunulan algoritmalardan "English" olanı seçilmiştir.

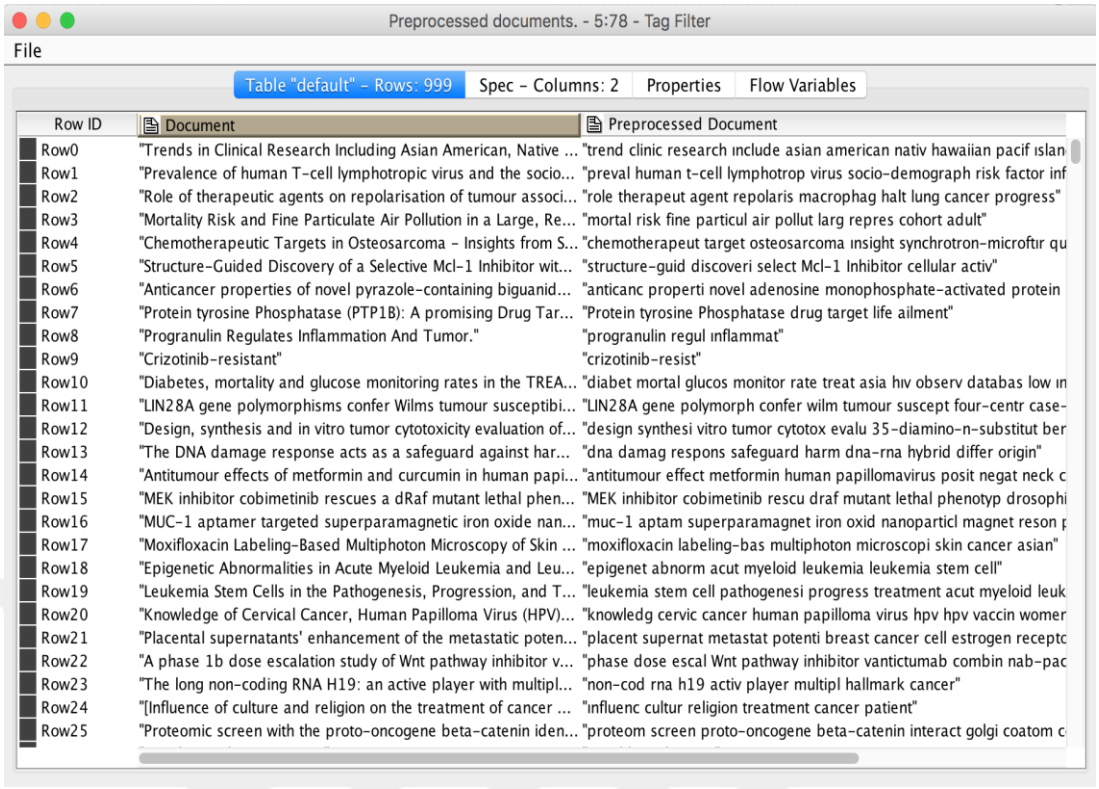
3.3. Frekans Belirleme ve Analiz

Ön işleme tamamlandıktan sonra, belgelerdeki terimlerin frekansları (sıklıkları) ve tüm korpus hesaplanabilir. Filtreleme yoluyla ön işleme ve kök bulma işlemleri yapıldıktan sonra Bag of Words oluşturulur ve frekanslar hesaplanır. Bu uygulamada TF Terim Frekansı düğümü kullanılmıştır. Yeniden filtreleme bu frekans değerlerine göre yapılır. Son olarak kalan terimler bir etiket bulutu yardımıyla görselleştirilir.



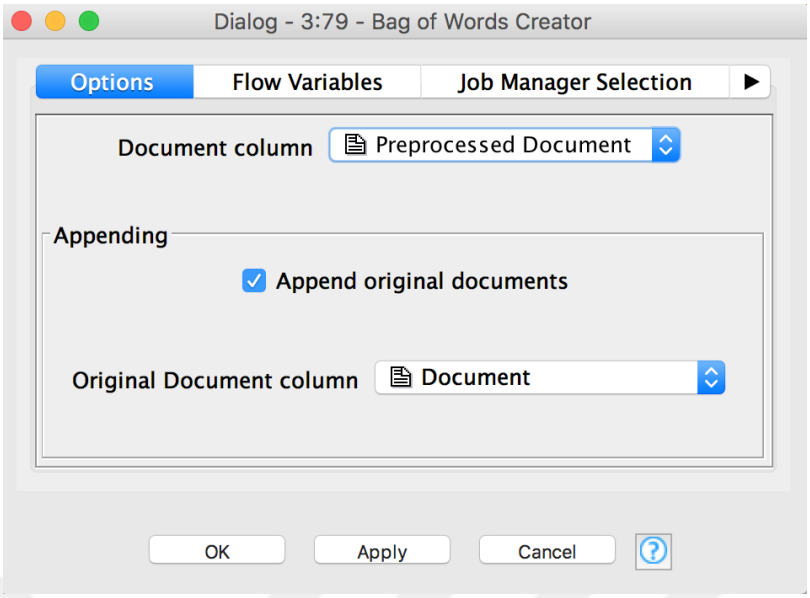
Şekil 3.11. Tag Filter ile oluşturulan döngü.

Tag Filter; dokümanda belirli etiketlerin atandığı terimleri filtreler. Bir terim en az bir tanesinde filtrelenmemişse, belirtilen etiketlerden birtanesi o terime atanır. Eğer sıkı filtreleme (strict filtering) ayarlanmışsa bir terimin atanmış tüm etiketleri belirtilen etiketlerden oluşmalıdır. Bu uygulamada isim ve sıfatlara yönelik bir fitreleme uygulanmış olup, gereksiz gürültülü kelimelerin Tag Filter yardımıyla filtrelenmesi amaçlanmıştır.



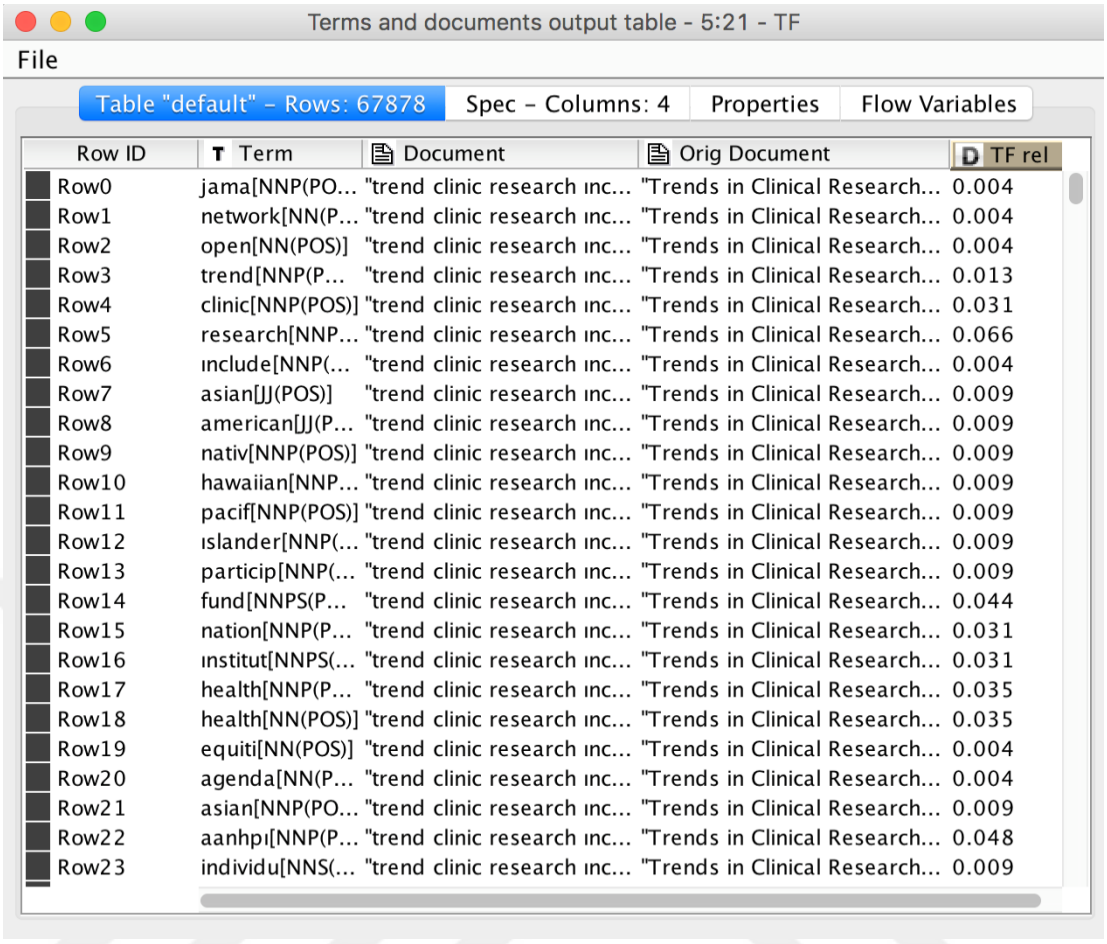
Şekil 3.12. Ön işleme dokümanına uygulanan Tag Filter penceresi.

Bag of Words Creator; bu aşamada gruplanan tüm dokümanlardaki tüm kelimelerin kullanım sıklıkları hesaplanır ve bir havuzda toplanır. Daha sonrasında ise bu kelimelerin değerleri (word weighting) hesaplanır. Kelime değeri, bir kelimenin belirli bir alan ile ilgili bir metnin içinde bulunma sıklığı olarak açıklanabilir. Yani dokümandan elde edilen her etiketi sayar ve gruplandırır. Bir Bag of Words, biri dokümanı içeren ve biri de ilgili dokümanda ortaya çıkan terimleri içeren en az iki sütundan oluşur.



Şekil 3.13. Bag of Words seçenekler penceresi.

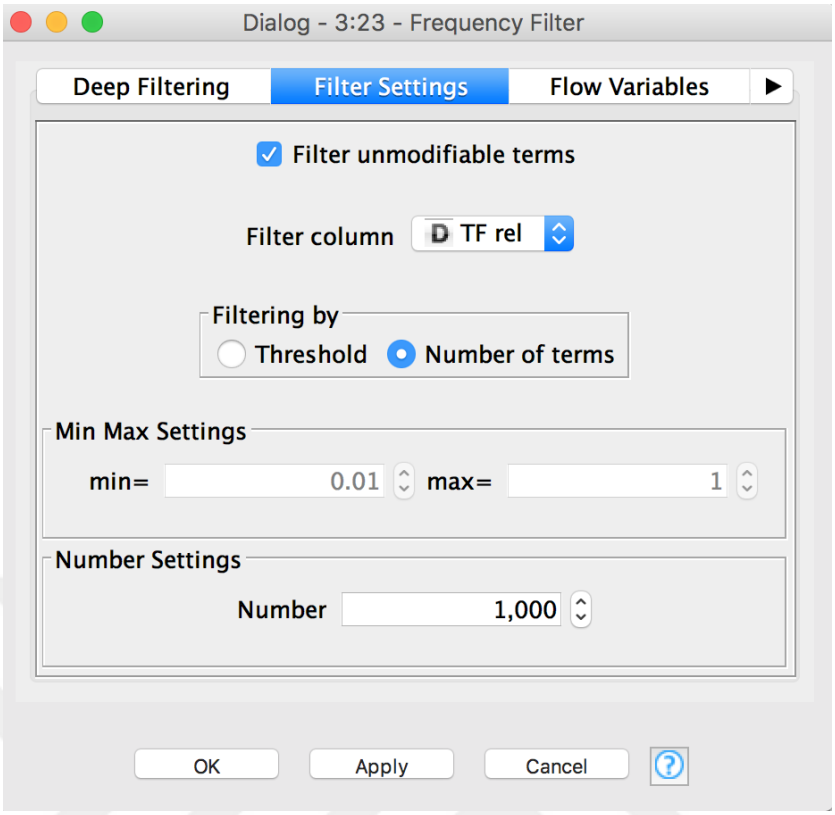
Term Frequency (TF) değeri frekans bilgisini yani terimin veri setinde kaç kez geçtiğini tutar. Ağırlıklandırmasında her bir dokümandaki kelimelerin frekansı rol oynamaktadır. Başka bir deyişle her bir dokümana göre her bir terimin bağlantılı terim frekansını (TF) hesaplar ve TF değerini içeren bir sütun ekler. Değer, bir belgeye göre bir terimin mutlak sıklığının, o belgenin tüm terimlerinin sayısına bölünmesiyle hesaplanır.



Row ID	Term	Document	Orig Document	TF rel
Row0	jama[NNP(PO...	"trend clinic research inc...	"Trends in Clinical Research...	0.004
Row1	network[NN(P...	"trend clinic research inc...	"Trends in Clinical Research...	0.004
Row2	open[NN(POS)]	"trend clinic research inc...	"Trends in Clinical Research...	0.004
Row3	trend[NNP(P...	"trend clinic research inc...	"Trends in Clinical Research...	0.013
Row4	clinic[NNP(POS)]	"trend clinic research inc...	"Trends in Clinical Research...	0.031
Row5	research[NNP...	"trend clinic research inc...	"Trends in Clinical Research...	0.066
Row6	include[NNP(...	"trend clinic research inc...	"Trends in Clinical Research...	0.004
Row7	asian[JJ(POS)]	"trend clinic research inc...	"Trends in Clinical Research...	0.009
Row8	american[JJ(P...	"trend clinic research inc...	"Trends in Clinical Research...	0.009
Row9	nativ[NNP(POS)]	"trend clinic research inc...	"Trends in Clinical Research...	0.009
Row10	hawaiian[NNP...	"trend clinic research inc...	"Trends in Clinical Research...	0.009
Row11	pacif[NNP(POS)]	"trend clinic research inc...	"Trends in Clinical Research...	0.009
Row12	islander[NNP(...	"trend clinic research inc...	"Trends in Clinical Research...	0.009
Row13	particip[NNP(...	"trend clinic research inc...	"Trends in Clinical Research...	0.009
Row14	fund[NNPS(P...	"trend clinic research inc...	"Trends in Clinical Research...	0.044
Row15	nation[NNP(P...	"trend clinic research inc...	"Trends in Clinical Research...	0.031
Row16	institut[NNPS(...	"trend clinic research inc...	"Trends in Clinical Research...	0.031
Row17	health[NNP(P...	"trend clinic research inc...	"Trends in Clinical Research...	0.035
Row18	health[NN(POS)]	"trend clinic research inc...	"Trends in Clinical Research...	0.035
Row19	equiti[NN(POS)]	"trend clinic research inc...	"Trends in Clinical Research...	0.004
Row20	agenda[NN(P...	"trend clinic research inc...	"Trends in Clinical Research...	0.004
Row21	asian[NNP(PO...	"trend clinic research inc...	"Trends in Clinical Research...	0.009
Row22	aanhpi[NNP(P...	"trend clinic research inc...	"Trends in Clinical Research...	0.048
Row23	individu[NNS(...	"trend clinic research inc...	"Trends in Clinical Research...	0.009

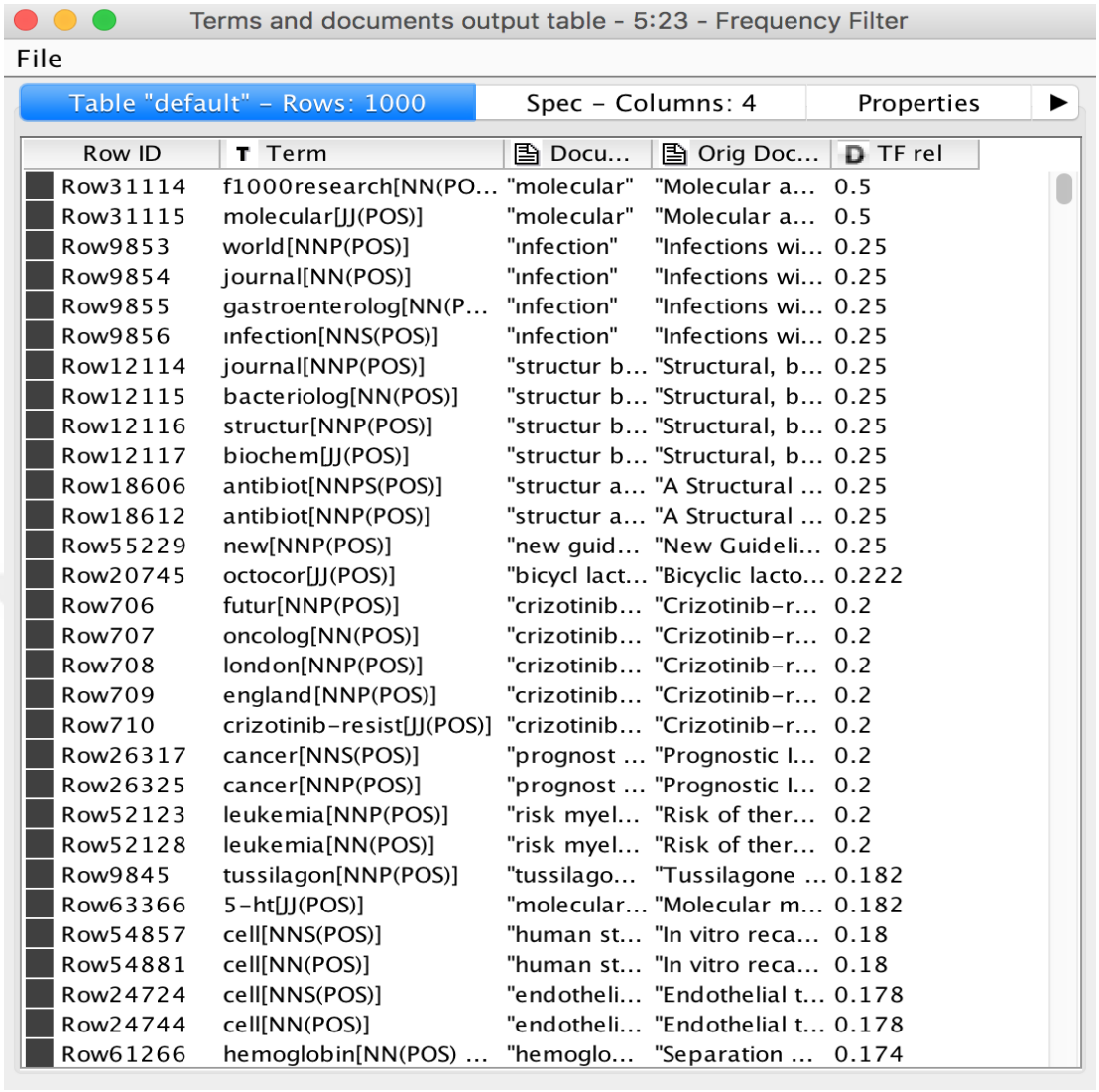
Şekil 3.14. TF ile dokümanda yer alan ağırlıklandırılmış terimler ve TF değerleri.

Frequency Filter; Hesaplanan frekanslara bağlı olarak, yüksek frekanslı terimleri korumak için "Frekans Filtresi" düğümü tarafından filtreleme uygulanabilir. Bir yandan da filtreleme için kullanılacak minimum ve maksimum değerler tanımlanabilir. Belirtilen bir frekans sütununun değeri minimumdan küçük veya maksimum değerden büyükse, terim filtrelenir. Öte yandan, k adet terim sayısı tanımlanabilir. Sadece en yüksek frekans değerine sahip olan k adet terim tutulur, gerisi filtre edilir.



Şekil 3.15. Frekans filtre ayarları penceresi.

Burada filtrelenecek değerler min=0.01 max=1 olarak seçilmiştir. Yani TF ağırlık değeri bu değerler dışında kalan terimler filtrelenecektir.

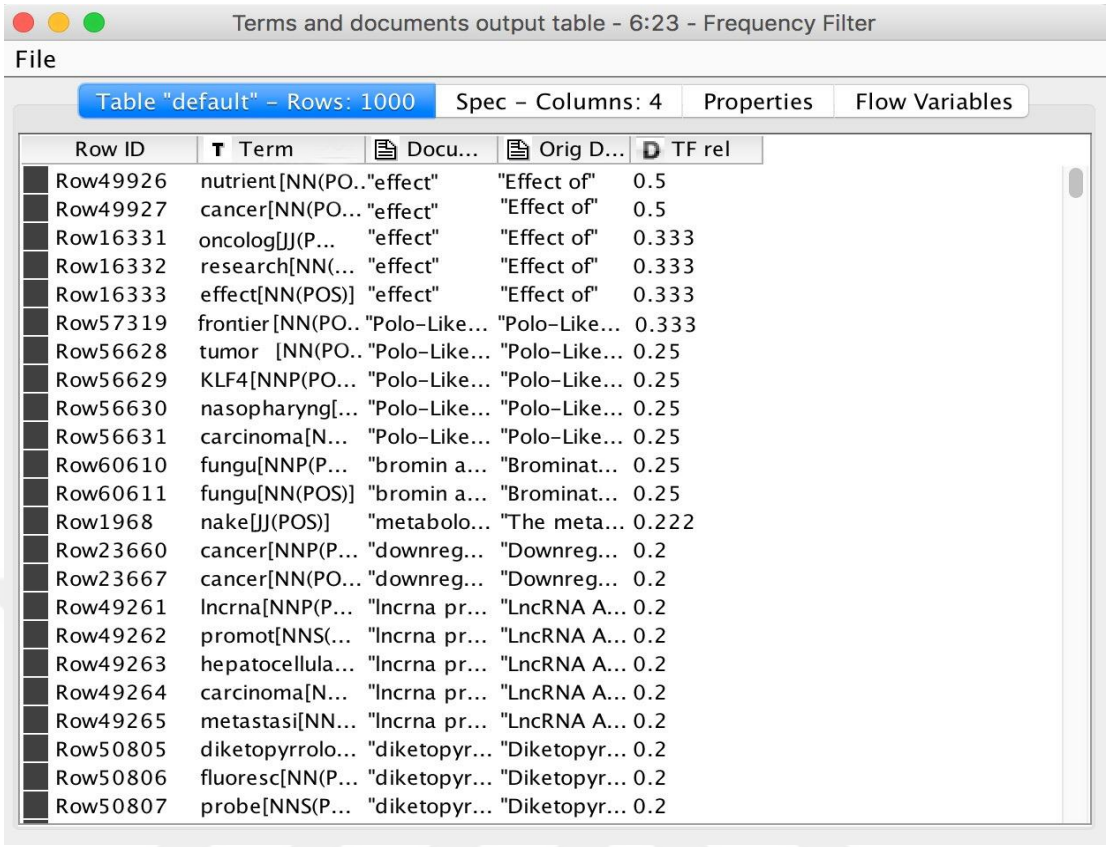


Row ID	Term	Docu...	Orig Doc...	TF rel
Row31114	f1000research[NN(PO...	"molecular"	"Molecular a...	0.5
Row31115	molecular[JJ(POS)]	"molecular"	"Molecular a...	0.5
Row9853	world[NNP(POS)]	"infection"	"Infections wi...	0.25
Row9854	journal[NN(POS)]	"infection"	"Infections wi...	0.25
Row9855	gastroenterolog[NN(P...	"infection"	"Infections wi...	0.25
Row9856	infection[NNS(POS)]	"infection"	"Infections wi...	0.25
Row12114	journal[NNP(POS)]	"structur b...	"Structural, b...	0.25
Row12115	bacteriolog[NN(POS)]	"structur b...	"Structural, b...	0.25
Row12116	structur[NNP(POS)]	"structur b...	"Structural, b...	0.25
Row12117	biochem[JJ(POS)]	"structur b...	"Structural, b...	0.25
Row18606	antibiot[NNPS(POS)]	"structur a...	"A Structural ...	0.25
Row18612	antibiot[NNP(POS)]	"structur a...	"A Structural ...	0.25
Row55229	new[NNP(POS)]	"new guid...	"New Guideli...	0.25
Row20745	octocor[JJ(POS)]	"bicycl lact...	"Bicyclic lacto...	0.222
Row706	futur[NNP(POS)]	"crizotinib...	"Crizotinib-r...	0.2
Row707	oncolog[NN(POS)]	"crizotinib...	"Crizotinib-r...	0.2
Row708	london[NNP(POS)]	"crizotinib...	"Crizotinib-r...	0.2
Row709	england[NNP(POS)]	"crizotinib...	"Crizotinib-r...	0.2
Row710	crizotinib-resist[JJ(POS)]	"crizotinib...	"Crizotinib-r...	0.2
Row26317	cancer[NNS(POS)]	"prognost ...	"Prognostic l...	0.2
Row26325	cancer[NNP(POS)]	"prognost ...	"Prognostic l...	0.2
Row52123	leukemia[NNP(POS)]	"risk myel...	"Risk of ther...	0.2
Row52128	leukemia[NN(POS)]	"risk myel...	"Risk of ther...	0.2
Row9845	tussilagon[NNP(POS)]	"tussilago...	"Tussilagone ...	0.182
Row63366	5-ht[JJ(POS)]	"molecular...	"Molecular m...	0.182
Row54857	cell[NNS(POS)]	"human st...	"In vitro reca...	0.18
Row54881	cell[NN(POS)]	"human st...	"In vitro reca...	0.18
Row24724	cell[NNS(POS)]	"endotheli...	"Endothelial t...	0.178
Row24744	cell[NN(POS)]	"endotheli...	"Endothelial t...	0.178
Row61266	hemoglobin[NN(POS) ...	"hemoglo...	"Separation ...	0.174

Şekil 3.16. İnsan ve kanser verileri için Frekans Filtreleme ile elde edilen TF değerleri.

Tag Cloud (Etiket Bulutu); oluşturulan etiketleri görsel olarak gösterilmesini sağlar. Sık kullanılan etiketleri daha büyük diğerlerini daha küçük yazar.

Document Grabber düğümü ile elde edilen insan ve kanser metinsel verilerine uygulanan bu aşamaların aynısı fareler ve kanser metinsel verilerine de uygulanmış ve sonuçları değerlendirilmiştir.



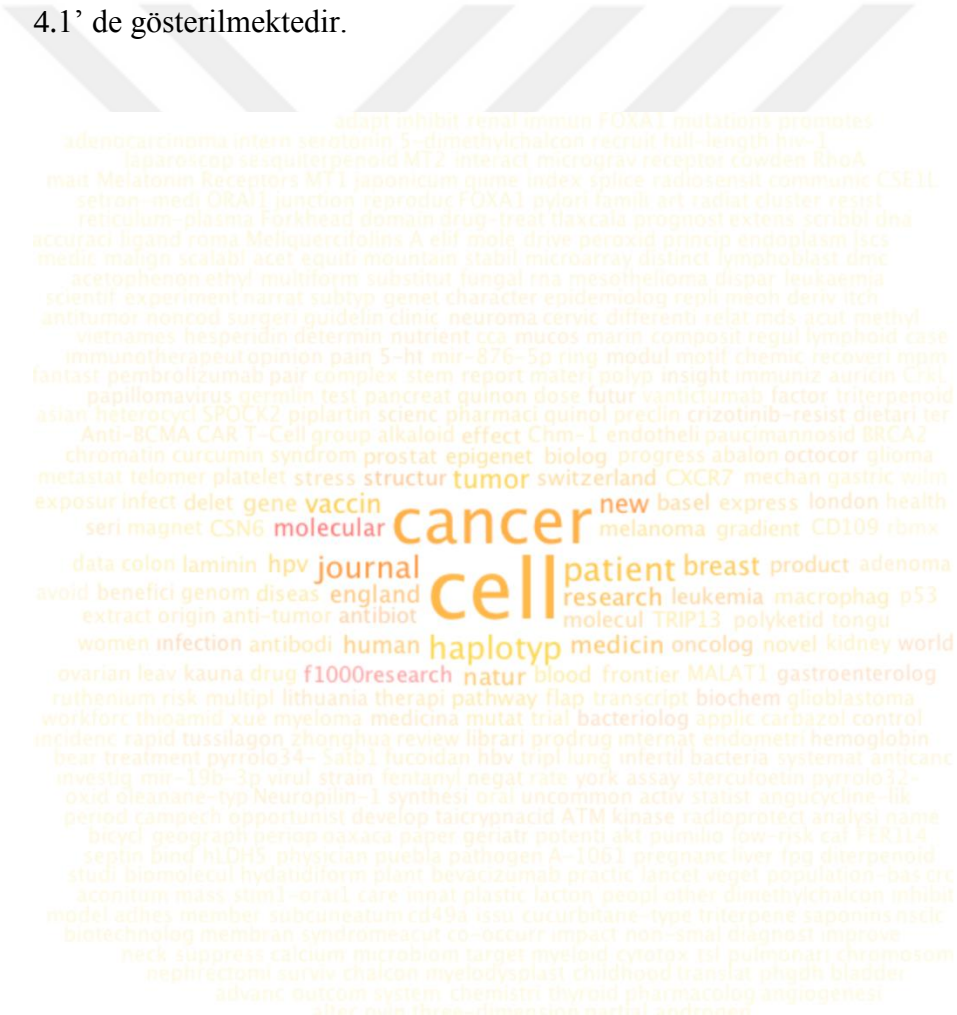
Row ID	Term	Docu...	Orig D...	TF rel
Row49926	nutrient[NN(PO..	"effect"	"Effect of"	0.5
Row49927	cancer[NN(PO...	"effect"	"Effect of"	0.5
Row16331	oncolog[JJ(P...	"effect"	"Effect of"	0.333
Row16332	research[NN(...	"effect"	"Effect of"	0.333
Row16333	effect[NN(POS)]	"effect"	"Effect of"	0.333
Row57319	frontier[NN(PO..	"Polo-Like...	"Polo-Like...	0.333
Row56628	tumor [NN(PO..	"Polo-Like...	"Polo-Like...	0.25
Row56629	KLF4[NNP(PO...	"Polo-Like...	"Polo-Like...	0.25
Row56630	nasopharyng[...	"Polo-Like...	"Polo-Like...	0.25
Row56631	carcinoma[N...	"Polo-Like...	"Polo-Like...	0.25
Row60610	fungu[NNP(P...	"bromin a...	"Brominat...	0.25
Row60611	fungu[NN(POS)]	"bromin a...	"Brominat...	0.25
Row1968	nake[JJ(POS)]	"metabolo...	"The meta...	0.222
Row23660	cancer[NNP(P...	"downreg...	"Downreg...	0.2
Row23667	cancer[NN(PO...	"downreg...	"Downreg...	0.2
Row49261	lncrna[NNP(P...	"lncrna pr...	"LncRNA A...	0.2
Row49262	promot[NNS(...	"lncrna pr...	"LncRNA A...	0.2
Row49263	hepatocellula...	"lncrna pr...	"LncRNA A...	0.2
Row49264	carcinoma[N...	"lncrna pr...	"LncRNA A...	0.2
Row49265	metastasi[NN...	"lncrna pr...	"LncRNA A...	0.2
Row50805	diketopyrrolo...	"diketopyr...	"Diketopyr...	0.2
Row50806	fluoresc[NN(P...	"diketopyr...	"Diketopyr...	0.2
Row50807	probe[NNS(P...	"diketopyr...	"Diketopyr...	0.2

Şekil 3.17. Fareler ve kanser verileri için Frekans Filtreleme ile elde edilen TF değerleri.

4. BULGULAR

Bu çalışmada sağlık alanında belirlenen iki konu başlığı altında elde edilen dokümanlar iki kategoriye ayrılmıştır ve bunlar kategori görevlerine göre temelde iki küme halinde bulunmaktadır. İlk küme insanlarda meydana gelen kanser vakalarıyla alakalı metinsel dokümanlardan, ikinci küme ise fareler üzerinde yapılan kanser çalışmalarıyla ilgili metinsel dokümanlardan meydana gelmektedir.

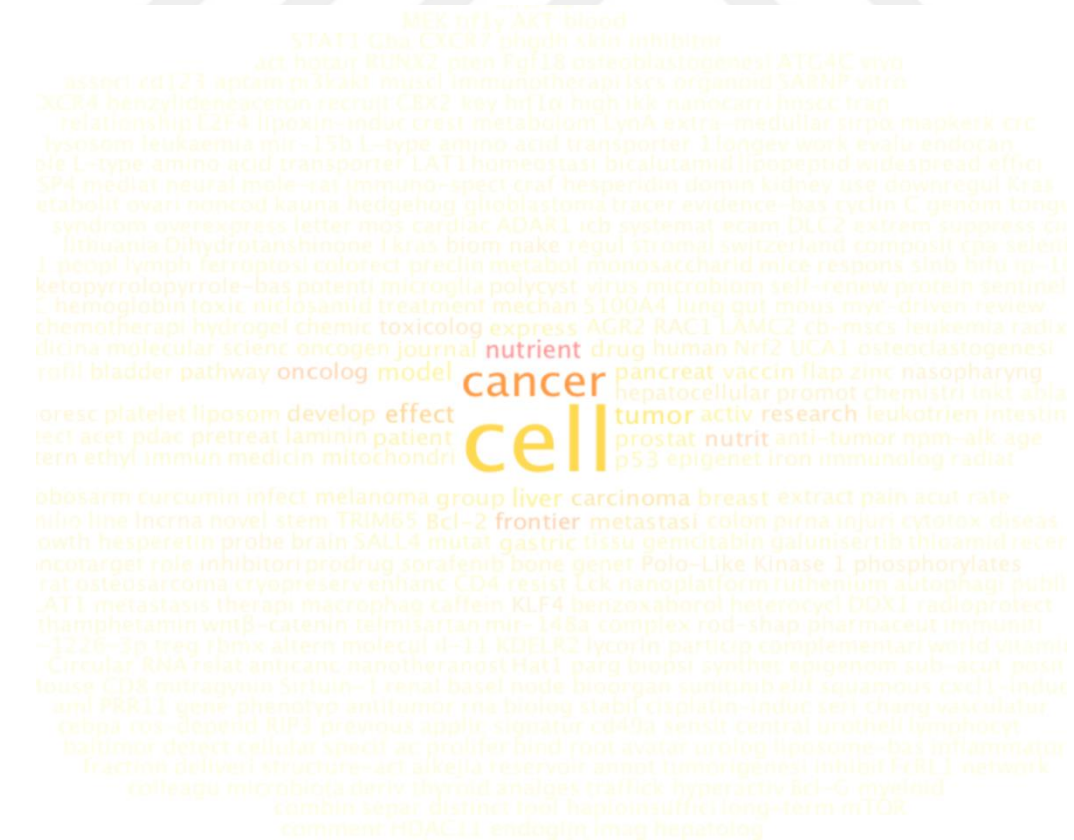
İnsanlarda meydana gelen kanser vakalarıyla alakalı metinsel dokümanlardan elde edilen veriler kullanılarak yapılan analiz sonucu elde edilen Tag Cloud grafiği Şekil 4.1' de gösterilmektedir.



Şekil 4.1. Tag Filter kullanılarak elde edilen Tag Cloud grafiği.

Grafikten görüleceği üzere insan ve kanser dokümanlarında yer alan etiketlerden en sık kullanılanları daha büyük harflerle yazılmış olup TF değerleri en düşükten en yükseğe doğru olacak şekilde Color Manager düğümü kullanılarak, sarıdan kırmızıya doğru renklendirilmiştir. Sık kullanılan etiketlerin terim frekans değerleri farklı olabilir. Çok sık kullanılmayan ancak terim frekans değeri fazla olan etiketler küçük ancak koyu renkli, sık kullanılan ama terim frekans değeri düşük olan etiketler büyük ancak açık renkli olabilir. Dolayısıyla dokümanda TF değeri en yüksek olan ve en sık geçen kelime "cell" yani "hücre" kelimesidir. Hemen arkasından "cancer" kelimesi gelmektedir. Sonrasında önem sırasına göre tıpta genetik bir terim olan "haplotyp" (gen terimi) , "journal" (dergi), "vaccin" (aşı), "patient" (hasta), "tumor" (tümör), "molecular" (moleküler), "breast" (meme), "medicin" (tıp) gibi kelimeler frekanslar tarafından ağırlıklandırılarak bu dokümanlarda en çok söz edilen kelime grupları olmuşlardır.

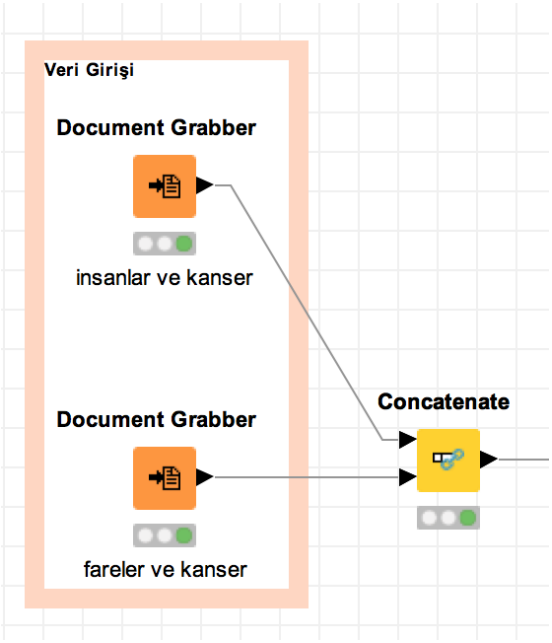
Benzer şekilde farelerde yapılan kanser araştırmalarıyla alakalı metinsel dokümanlardan elde edilen Tag Cloud grafiği de Şekil 4.2' de gösterilmektedir.



Şekil 4.2. Tag Filter kullanılarak elde edilen Tag Cloud grafiği.

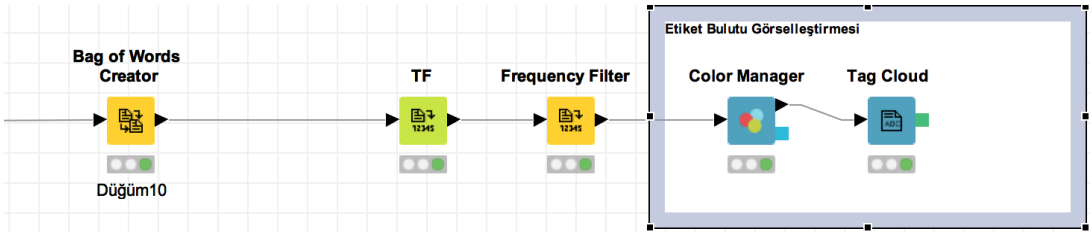
Burada yine grafikten görüleceği üzere farelerde kanser araştırmalarıyla alakalı elde edilen metinsel verilere metin madenciliği uygulanmış ve bu dokümanlarda en sık rastlanan kelime grupları etiketlenerek Tag Cloud'da (etiket bulutu) grafiksel olarak ifade edilmiştir. Burada öne çıkan kelimeler yine "cell" (hücre) ve "cancer" (kanser) kelimeleridir. Daha sonra sırasıyla "nutrient" (besin), "effect" (tesir), "tumor" (tümör), "carcinoma" (bir çeşit ur), "liver" (karaciğer), "model" (model), "drug" (ilaç), "patient" (hasta), "pancreat" (pankreas), "oncolog" (onkolog) gibi kelimeler ön plana çıkmaktadır. Bir önceki etiket bulutuyla bu analiz sonucu ortaya çıkan etiket bulutunda benzer ve farklı kelimelerin ön plana çıktığı görülmektedir.

Bu kez bu iki veriyi birleştirerek Tag Filter kullanmadan analiz sürecini tekrarlayalım. İnsanlarda kanser vakalarıyla, farelerde kanser araştırmalarına ait olan metinsel verilerin bulunduğu Document Grabber düğümleri, veri akışını temsil eden oklar ile Concatenate (birleştirme) düğümüne bağlanır. Concatenate düğümü kendinden önceki iki tablodaki dokümanları birleştirmeye yarar. Bir giriş tablosunda diğer tablonun bulunmadığı sütun adları varsa, sütunlar çıktı tablosunda olmayacak şekilde eksik değerlerle veya filtrelenmiş değerlerle doldurulabilir.



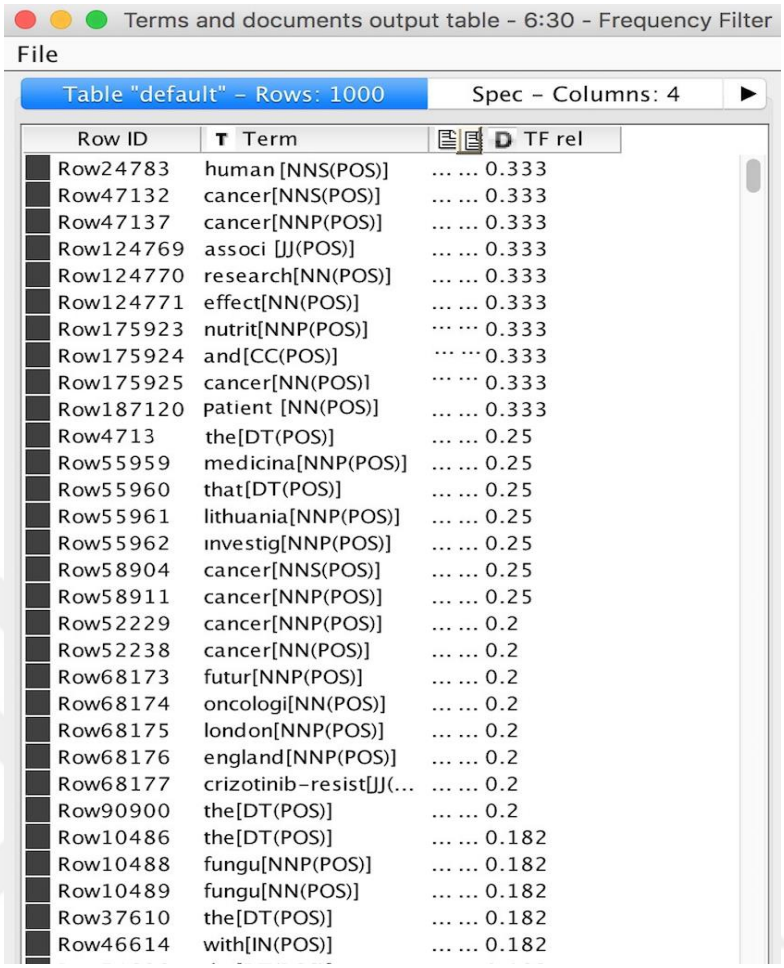
Şekil 4.3. Document Grabber ve Concatenate düğümlerinin birbirine bağlanması.

Tag Filter Filtresini kullanmadan oluşturduğumuz Knime döngüsü Şekil 4.4' te olduğu gibidir.

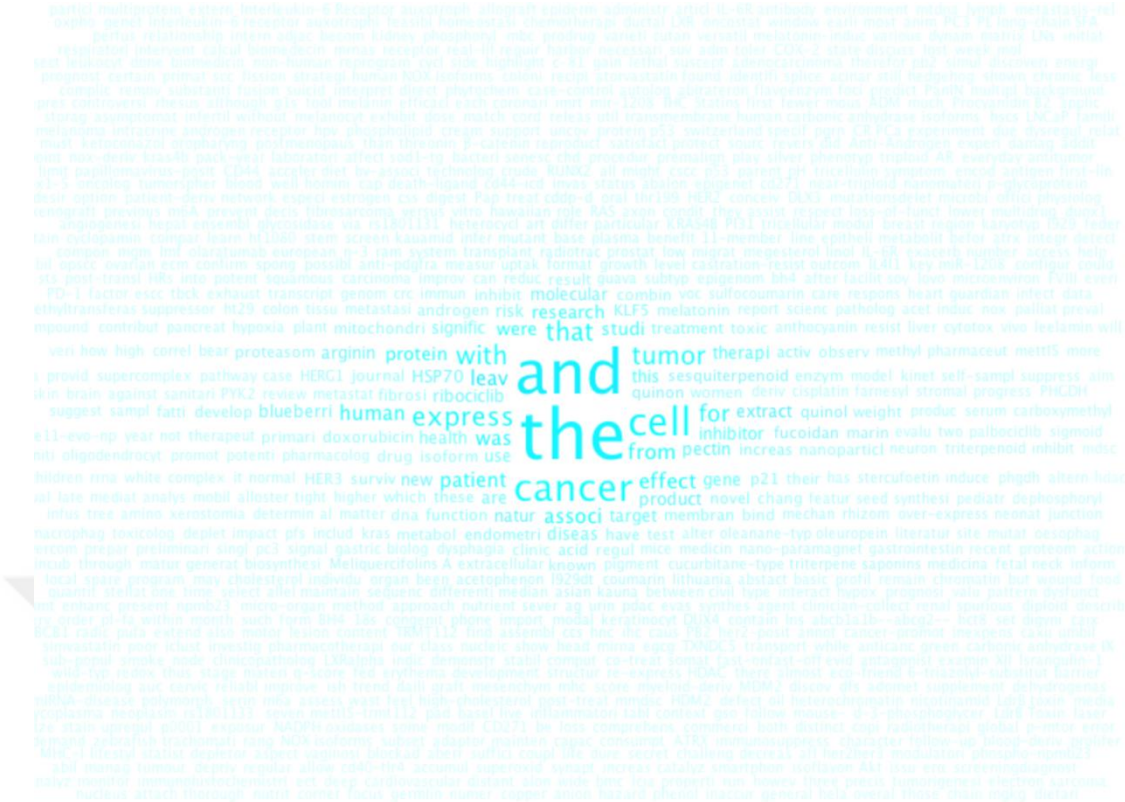


Şekil 4.4. Tag filter kullanmadan oluşturulan döngü

İki metinsel veri birleştirildikten sonra Tag Filter olmadan oluşturulan Knime döngüsünde, frekans filtrelemeyle elde edilen TF değerleri ve etiket bulutu görseli Şekil 4.5 ve Şekil 4.6' da gösterilmiştir.



Şekil 4.5. Her iki veri için Tag Filter kullanılmadan Frekans Filtreleme ile elde edilen TF değerleri.

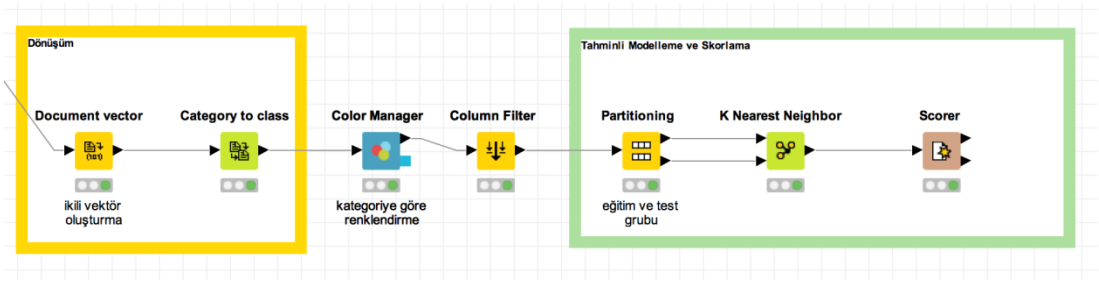


Şekil 4.6. Tag Filter kullanılmadan elde edilen Tag Cloud grafiği.

Grafikten de anlaşılacağı üzere İngilizce dilinde kullanılan “and”, “the”, “with”, “were”, “that” gibi bağlaçlar filtrelenmemiştir. Bu gibi bağlaçlar metin içerisinde çok sık kullanıldığından dolayı çalışma yapılan dokümanlarda bunun gibi anlamsız kelimeler, anlamlı gibi değerlendirilip terim frekansları ile ağırlıklandırılmış ve etiket bulutunda büyük harflerle yazılıp koyu renk ile gösterilmişlerdir. Asıl anlamlı olması gereken diğer terimler ise daha geri planda kalmıştır. Bununla beraber her iki dokümanlarda da frekans değeri yüksek çıkan “cell”, “cancer”, “tumor”, “patient” gibi kelimelerin veriler birleştirildikten sonra yapılan analizde de yüksek oranla çıktığı gözlenmiştir.

Son olarak birleştirilmiş veri setine Knime aracılığıyla K-NN algoritmasını uygulayalım. İnsanlarda kanser vakalarıyla, farelerde kanser araştırmalarına ait olan metinsel verilerin bulunduğu Document Grabber düğümleri, Concatenate düğümüyle birbirine bağlandıktan sonra yukarıdaki metin ön işleme bölümlerinin hepsi

tekrarlanarak terim frekansları hesaplanır. Bu ön işleme aşamasından sonra K-NN hesaplaması için oluşturulan Knime döngüsü Şekil 4.7' de görüldüğü gibidir.



Şekil 4.7. K-NN algoritması için oluşturulan Knime döngüsü

Document vector düğümü, terim alanında onu temsil eden her doküman için bir doküman vektörü oluşturur. Özellik vektörlerinin değerleri belirtilen bir sütunun değerleri olabilir. TF değeri, vektör değeri olarak atanır. Her bir TF değeri için oluşturulan doküman vektörleri Şekil 4.8'de görülmektedir.

Documents output table - 10:52 - Document vector (binary vector)

File

Table "default" - Rows: 1999 | Spec - Columns: 1197 | Properties | Flow Variables

Row ID	Docu...	compl...	altern	medi...	cervic	cancer	common	type	women	world...	remair
Row60	"ajoen ma...	0	0	0	0	1	0	0	0	0	0
Row61	"alc1 knoc...	0	0	0	0	1	0	0	0	0	0
Row62	"aldh2 ca...	0	0	1	0	1	0	0	0	0	0
Row63	"aldh2 def...	0	0	0	0	1	0	0	0	0	1
Row64	"alkaloid i...	0	0	1	0	1	0	0	0	0	0
Row65	"all-tran r...	0	0	0	0	0	0	0	0	0	1
Row66	"all-tran r...	0	0	0	0	0	0	0	0	0	1
Row67	"allicin pro...	0	0	0	0	0	0	0	0	0	0
Row68	"allogen b...	0	0	0	0	1	0	0	0	0	0
Row69	"allostat a...	0	0	0	0	1	0	0	0	0	0
Row70	"alpha-lin...	0	0	0	0	0	0	1	0	0	0
Row71	"alter intra...	0	0	0	0	1	0	0	0	0	0
Row72	"altern me...	0	1	1	0	1	0	1	0	0	0
Row73	"altern spl...	0	1	0	0	0	0	0	0	0	0
Row74	"ambient ...	0	0	0	0	1	0	0	0	0	1
Row75	"amelior e...	0	0	0	0	1	0	0	0	0	0
Row76	"amh amh...	0	0	0	0	0	1	1	0	0	0
Row77	"amid pro...	0	0	0	0	1	0	0	0	0	0
Row78	"amido-p...	0	0	0	0	1	0	0	0	0	0
Row79	"ampk-rel...	0	0	0	0	1	1	0	0	0	1
Row80	"analys m...	0	0	0	0	0	0	0	0	0	0
Row81	"anaplast ...	0	0	0	0	0	0	0	0	0	0
Row82	"andrea v...	0	0	0	0	0	1	0	0	0	0
Row83	"androgen...	0	0	1	0	1	0	0	0	0	1

Şekil 4.8. Doküman vektörleri tablo görüntüsü.

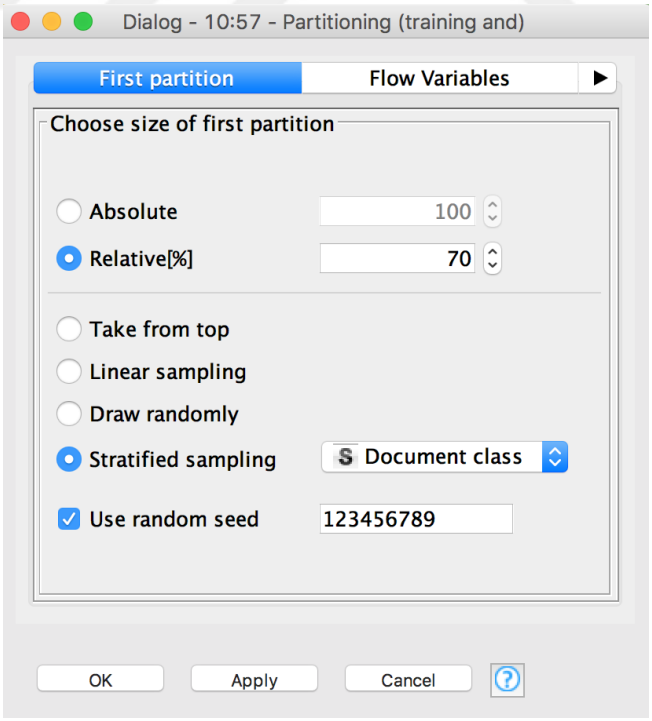
Category to class düğümü, bir doküman hücresi içeren her satıra bir sınıf (string=dize) sütunu eklemesini sağlar. Sınıfın değeri dokümanın dize (string) olarak kategorisidir. Hiçbir kategori tanımlanmadıysa, sınıf değeri "tanımsız" olarak

ayarlanır. Eğer birden fazla kategori tanımlanmışsa, ilk kategori (alfabetik olarak sıralanan kategoriler listesinin sınıfı) sınıf değeri olarak kullanılır.

Color manager düğümünde, kategoriye göre renklendirme işlemi yapılmıştır. İnsanlar ve kanser kategorisi kırmızı, fareler ve kanser kategorisi ise mavi renk ile temsil edilmiştir.

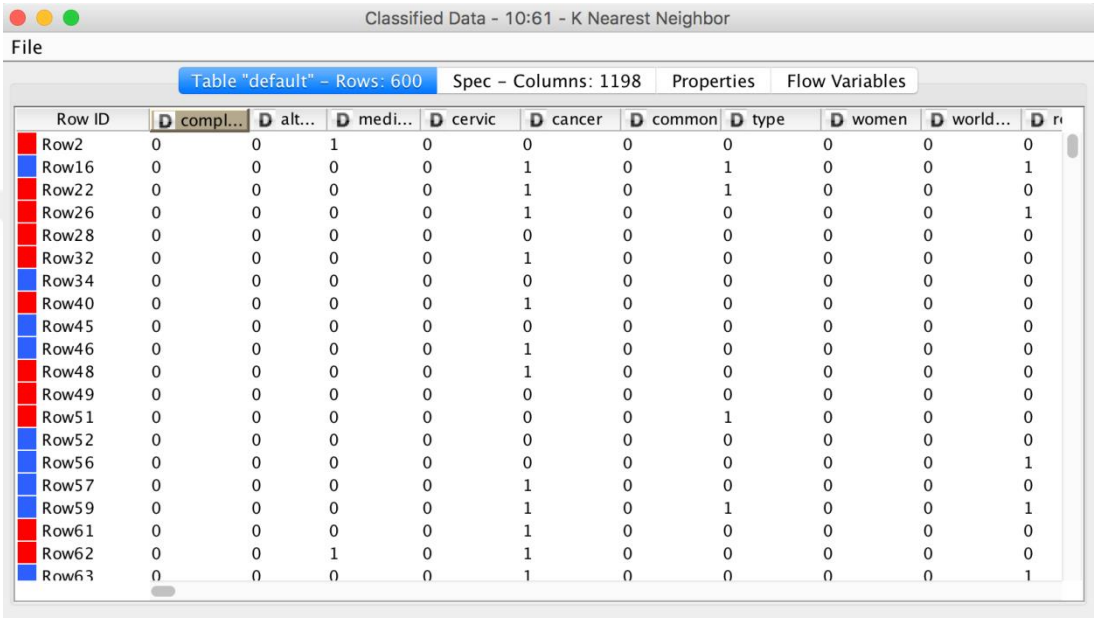
Column filter düğümüyle sütunların giriş tablosundan filtrelenmesine izin verirken yalnızca kalan sütunlar çıkış tablosuna iletilir.

Partitioning düğümünde giriş tablosu eğitim ve test verisi olarak iki bölüme ayrılır. Şekil 4.9'da görüldüğü gibi eğitim verisini oluşturacak satır yüzdesi buradan oluşturulabilmektedir. Kalan satırlar ise test verisi olarak ikinci bölümü oluşturarak K Nearest Neighbor düğümüne girecektir. Ayrıca diyalog penceresinde tabakalı örnekleme seçeneği seçilmiş, yeniden yürütme sonrasında tekrarlanabilir sonuçlar (reproducible results) elde etmek için alttaki kutucuğa da sabit bir değer girilmiştir.



Şekil 4.9. Partitioning seçenekler penceresi

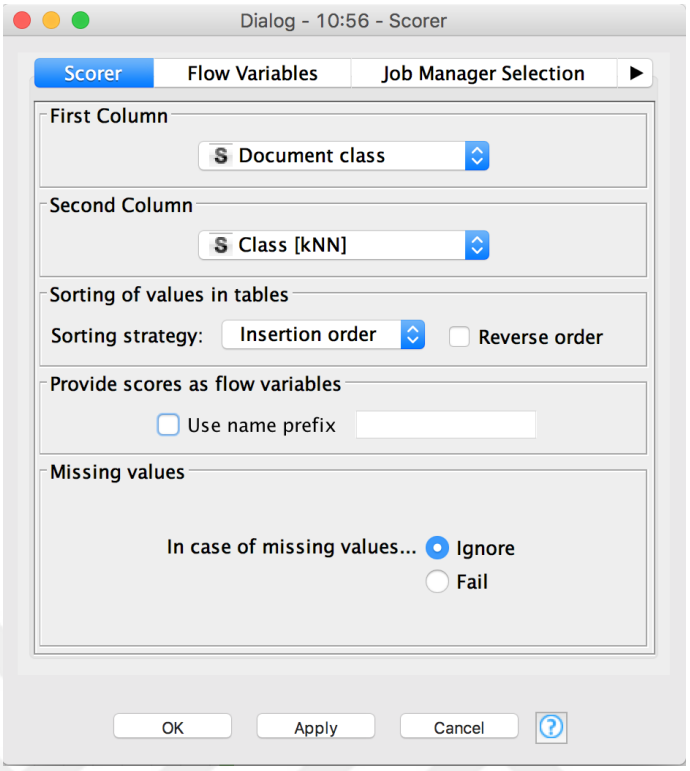
K Nearest Neighbor düğümü, eğitim verilerini kullanarak K en yakın komşu algoritmasına göre bir test verisi kümesini sınıflandırır. Sınıflandırma için en yakın komşu değeri K=3 olarak belirlenmiştir. Şekil 4.10'de sınıflandırılmış olan kelime vektörleri ile doküman satırları görülmekte, doküman satırındaki kırmızı ve mavi renkler ise o dokümanın hangi kategoriye ait olduğunu ifade etmektedir. Kırmızı renk insanlar ve kanser, mavi renk ise fareler ve kanser kategorilerini temsil etmektedir.



Row ID	compl...	alt...	medi...	cervic	cancer	common	type	women	world...
Row2	0	0	1	0	0	0	0	0	0
Row16	0	0	0	0	1	0	1	0	0
Row22	0	0	0	0	1	0	1	0	0
Row26	0	0	0	0	1	0	0	0	0
Row28	0	0	0	0	0	0	0	0	0
Row32	0	0	0	0	1	0	0	0	0
Row34	0	0	0	0	0	0	0	0	0
Row40	0	0	0	0	1	0	0	0	0
Row45	0	0	0	0	0	0	0	0	0
Row46	0	0	0	0	1	0	0	0	0
Row48	0	0	0	0	1	0	0	0	0
Row49	0	0	0	0	0	0	0	0	0
Row51	0	0	0	0	0	0	1	0	0
Row52	0	0	0	0	0	0	0	0	0
Row56	0	0	0	0	0	0	0	0	1
Row57	0	0	0	0	1	0	0	0	0
Row59	0	0	0	0	1	0	1	0	1
Row61	0	0	0	0	1	0	0	0	0
Row62	0	0	1	0	1	0	0	0	0
Row63	0	0	0	0	1	0	0	0	1

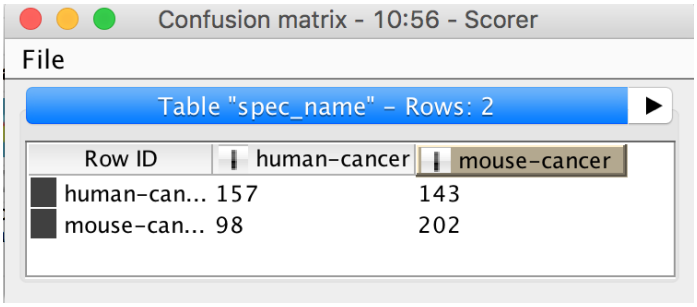
Şekil 4.10. K-NN algoritması ile sınıflandırılmış veri penceresi.

Scorer düğümü ise iki sütunu öznitelik değer çiftleriyle karşılaştırır ve hata matrisini gösterir. Şekil 4.11'daki scorer seçenekler penceresi, karşılaştırma için iki sütun seçilmesine izin verir; seçilen ilk sütundan gelen değerler hata matrisinin (confusion matrix) satırlarında ve ikinci sütundaki değerlerin hata matrisinin sütunlarıyla temsil edilir. İlk sütun, verinin gerçek sınıflarını temsil eder. İkinci sütun, K-NN algoritması kullanılarak oluşturulmuş öngörülen veri sınıflarını temsil eder.

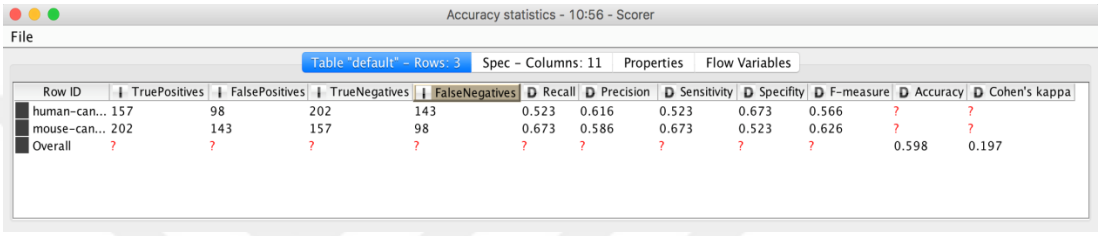


Şekil 4.11. Scorer seçenekler penceresi.

Scorer düğümünün çıktısı, her bir hücrede eşleşme sayısının yer aldığı hata matrisi ile doğruluk istatistikleri (accuracy statistics) tablosudur. Kullanılan sınıflandırma modellerinin performansını değerlendirmek için hedef niteliğe ait tahminlerin ve gerçek değerlerin karşılaştırıldığı hata matrisi, sıklıkla kullanılmaktadır. Doğruluk istatistikleri tablosunda ise True Positive (TP, Doğru Pozitif), True Negative (TN, Doğru Negatif), False Positive (FP, Yanlışa Pozitif), False Negative (FN, Yanlış Negatif), Precision (Kesinlik, Pozitif Tahmini Değer), Recall / Sensitivity (Duyarlılık), Specificity (Özgüllük), F-measure (F-ölçütü), Accuracy (Doğruluk) ve Cohen's kappa gibi bazı istatistiksel ölçüler yer alır.



Şekil 4.12. Hata (Confusion) Matrisi.



Şekil 4.13. Doğruluk (Accuracy) istatistikleri tablo görüntüsü.

Şekil 4.12'de K-NN algoritması kullanılarak oluşturulan test verileri yardımı ile gerçekleşen tahminlerin matrisi görülmektedir. Burada K-NN algoritması kullanılarak toplamda 600 adet olan test dokümanlarının 255 tanesi insan ve kanser sınıfına ait olan dokümanlar olarak sınıflanmış olup bunların 157 tanesi insan ve kansere sınıfına ait olduğu yani doğru tahmin edildiği gözlenmiştir. Geriye kalan 98 tanesinin hatalı tahmin olduğu, aslında fareler ve kanser sınıfına ait olan bu dokümanları da insanlar ve kanser sınıfına ait doküman gibi algılayıp hatalı sınıflandırıldığı görülmüştür. Buradaki başarı yüzdesi %61.6 olarak hesaplanmış ve Şekil 4.13'de Precision sütununda gösterilmiştir. Aynı şekilde toplamda 600 adet olan test dokümanlarından fareler ve kanser sınıfına ait olarak 345 doküman tespit edilmiş olup bunun 202 tanesinin gerçekten fareler ve kanser sınıfına ait dokümanlar olduğu, geri kalan 143 tanesinin de insanlar ve kanser sınıfına ait dokümanlar olduğu hata matrisinde yer almaktadır. Buradaki başarı yüzdesi de %58.6 olarak hesaplanmış, Şekil 4.13'de Precision sütunu olarak gösterilmiştir.

Aynı zamanda algoritma tarafından doğru tahmin edilen insan ve kanser sınıfına ait doküman miktarının gerçekte test verisi olarak kullanılan 300 adet insan ve kanser

dokümanlarına oranı %52.3 olarak hesaplanmıştır ve bu Şekil 4.13'de Recall / Sensitivity (Duyarlılık) sütunu olarak gösterilmiştir. Aynı şekilde bu oran fareler ve kanser dokümanlarında %67.3 olarak hesaplanmıştır. Bu sonuçlara göre; fareler ve kanser dokümanlarındaki doğru sınıflandırma yüzdesi oldukça yüksek çıkmıştır. Precision, tahmin edilen sınıflar içerisinde gerçekten kaç tanesi doğru, sorusuna cevap olarak verilen bir ölçüttür. Recall / Sensitivity ise pozitif kategoriye ait sınıflardan kaç tanesi doğru tahmin edildi, sorusuna cevap veren bir ölçüttür.

Precision (Kesinlik) ve Recall / Sensitivity (Duyarlılık) ölçütlerinin harmonik ortalaması ise F ölçütünü verir. Bu ölçütün avantajı iki farklı ölçüt yerine tek bir ölçütün değerlendirilmesine olanak sağlamasıdır. Sınıflandırma doğruluğu olarak bilinen F ölçütü, Precision ve Recall / Sensitivity değerlerini aynı anda ölçmeye yardım eder. Şekil 4.13'de F ölçütüne göre insanlar ve kanser dokümanlarının %56.6'lık, fareler ve kanser dokümanlarının ise %62.6'lık bir doğru sınıflandırılma yüzdesi mevcuttur.

Accuracy (Doğruluk), kullanılan algoritmanın ne kadar doğru sonuç verdiğini ortaya koyar. Sistemde doğru olarak yapılan tahminlerin tüm tahminlere oranıdır. K-NN algoritması kullanılarak yapılan doğru doküman sınıflama oranı %59.8 olarak elde edilmiştir.

Cohen's kappa değeri ise her iki sınıflama, gerçek sonuç ile sınıflandırma sonucu arasındaki toplam rastgele uyum olasılığının %19.7 olduğu ve Cohen Kappa değeri genel sınıflamasına göre uyumun zayıf düzeyde (≤ 20) olduğu söylenebilir[21].

5. TARTIŞMA ve SONUÇ

Bu tez çalışmasında; metin madenciliği yöntemi hakkında bilgiler verilmiş ve sağlık alanında belirlenen bir konuda uygulaması gösterilmiştir. Çalışmanın uygulama aşamasında; "insanlarda görülen kanser vakaları (human and cancer)" ve "farelerde kanser araştırmaları (mouse and cancer)" şeklinde belirlenen iki farklı konu başlığı altında en sık kullanılan Pubmed veritabanından ayrı ayrı elde edilen dokümanlara ve daha sonra bu dokümanlar birleştirilerek, birleştirilmiş olan dokümanlara metin madenciliği yöntemi uygulanmış, Knime programıyla K-NN algoritması kullanılarak doküman sınıflarının nasıl oluşturulduğuna, Knime programının metin madenciliğinde nasıl kullanıldığına ve elde edilen dokümanlara uygulanan adımların neler olduğuna ayrıntılı olarak yer verilmiştir.

Metin madenciliği yöntemleri, günümüzde pek çok alanda uygulanabildiği gibi, tıp ve biyoloji alanında da sıkça uygulanan bir yöntemdir. Thampson, "Tıp tarihinde metin madenciliği" başlıklı araştırmasında metin madenciliği tekniklerinin çeşitli anlamsal bilgileri otomatik olarak tanıma yetenekleri (yerler, tıbbi durumlar, ilaçlar, vb.) ve eş anlamlı / değişken kavram biçimleri ve kavramlar arasındaki ilişkileri (hangi tıbbi durumların tedavisinde hangi ilaçların kullanılması gibi) anlama yeteneğinden faydalanmıştır. Thampson P, araştırmasında 19. yüzyılın ortalarına dayanan yayımlanmış tarihi tıbbi metinlerin anlamsal analizini yapmıştır[3]. Lam C. ise 2000-2013 yılları arasında yayımlanan uyku bozuklukları ile ilgili dergi makalelerindeki yayın eğilimlerini belirlemek ve yine uyku bozuklukları ile metodolojik terimler arasındaki ilişkiyi keşfetmek için metin madenciliği yöntemini kullanmıştır[33]. Hoa, araştırmasında sağlık hizmetlerinin daha efektif olarak sağlanabilmesi amacıyla, Çinli hastaların web üzerinden tedavi oldukları doktorlar hakkında yaptıkları yazılı değerlendirmeler ile olumlu ve olumsuz yorumların, doktorların uzmanlık alanına göre farklılık gösterip göstermediğini metin madenciliği yöntemini kullanarak incelemiştir. Bu çalışmasında 2006-2014 dönemindeki tüm yorumları tanımlayıcı istatistikleri kullanarak inceleyip, konu modelleme algoritması olarak bilinen Latent Dirichlet Allocation algoritmasını Çinli hastaların doktor ziyaretleri hakkında hangi yorumları yaptıklarını anlayabilmek için

500.000'den fazla metin üzerinde uygulamıştır[34]. Mahgoub H, araştırmasında kuş gribi ile alakalı birçok kaynaktan (BBC, Reuters, Yahoo, Medical News Today, vb.) elde edilen örnek 100 adet internet sayfasının XML formatına dönüştürüldükten sonra anahtar kelimeler arasındaki bağlantılara bakılmış, hastalıkla alakalı özellikler (lokasyonu, hastanın durumu, vb.) EART isimli metin madenciliği sistemi ile ortaya konulmaya çalışılmıştır[35].

Bu çalışmanın analiz sürecinde, yapısal olmayan dokümanlar metin madenciliği yöntemleriyle yapısal hale getirilmiştir. Bu işlemler için açık kaynak kodlu bir veri madenciliği uygulaması olan Knime yazılımı kullanılmıştır. Knime uygulamasında bulunan ve Pubmed'ten elde edilmiş sağlık alanındaki dokümanlar iki kategori halinde uygulamaya dahil edilmiştir. İlk küme insanlarda kanser vakaları ile ilgili dokümanlardan, ikinci küme ise fareler de kanser araştırmaları ile ilgili dokümanlardan oluşmaktadır.

Bu dokümanlara Knime üzerine çeşitli işlemler yapılmadan önce, Knime uygulamasında metin madenciliği araçlarını kullanmaya olanak sağlayan "palladian" isimli mini ek yazılım yüklemiştir.

Pubmed'ten elde edilmiş olan, iki kategoriden oluşan bu dokümanlar ayrı ayrı Document Grabber düğümü yardımıyla Knime workflow sayfasına aktarılmıştır. Daha sonra bu dokümanlar POS tagger ve Abner tagger düğümlerine aktarılarak kelimelerin köklerine ve türlerine ayrılması sağlanmıştır. Bu aşamada Tokenizer ile metinler kelime bazında parçalara bölünmüş, teknik anlamda dizgiciklere ayrılmıştır. Abner tagger sayesinde ise metinlerde bulunan ve biyomedikal olarak adlandırılan varlıklar tanımlanmış ve bu terimler etiketlenmiştir.

Metin ön işleme safhasında ise Punctuation Erasure düğümü sayesinde dokümanlarda bulunan terimlerin noktalama işaretleri kaldırılmış, Number Filter düğümüyle de dokümanlarda yer alan ondalık ayıraçlar ve diğer işaretler filtrelenmiştir. Case Conventer düğümü ile de dokümanlarda yer alan tüm büyük harfler, küçük harflere dönüştürülmüştür.

Noktalama işaretlerinden arınan, köklerine ve türlerine göre ayrılmış olan metinsel veri, Snowball Stemmer düğümüne aktarılarak bu sözcüklerin anlamlı gövdelere yani anlamlı köklere ayırma işlemi yani kök bulma işlemi gerçekleşmiştir.

Metin ön işleme aşamalarından sonra Tag Filter yani etiketleme filtreleri kullanılarak ve veriler Concatenate düğümü sayesinde birleştirildikten sonra Tag Filter kullanılmayarak iki aşamalı olarak analize devam edilmiştir. Burada amaç iki verinin ayrı ayrı analizi sonucunda oluşan etiket bulutundaki sık kullanılan kelimeler ile birleştirildikten sonra etiket bulutundaki sık kullanılan kelimelerde ne gibi bir değişiklik olduğunu incelemektir. Tag Filter yani etiketleme filtresi kullanılarak yapılan analizde gereksiz ve tek başına bir anlam ifade etmeyen sözcüklerin ortadan kaldırılması hedeflenmiştir. Daha sonra Bag of Words düğümü kullanılarak gürültülü terimlerden arındırılan ve etiketlenen metinsel verilerin her bir etiketi sayılıp metnin içinde bulunma sıklığına göre gruplandırılmıştır. Terim Frekansı (TF) düğümü kullanılarak her bir terimin veri setinde kaç kez geçtiği hesaplanmıştır ve bunlar kullanım sıklığına göre "TF değeri" adı altında hesaplanarak bir sütun olarak gösterilmiştir.

Son olarak Frekans Filtresi düğümü kullanılarak elde edilen terimler belirli bir frekans değeri ile filtrelenmiştir. Burada filtrelenecek değerler min=0.01 ve max=1 olarak seçilmiştir. Yani TF değeri bu değerler dışında kalan terimler filtrelenmiştir. Filtrelenen bu etiketler Color Manager düğümü kullanılarak renklendirilmiştir. Burada terim frekans değerleri yüksek olanlar koyu renk ile terim frekans değerleri düşük olanlar daha açık renk ile görselleştirilerek terim frekans değerlerinin yüksek olup olmadığı gözlemlenmiştir.

Etiket bulutu görselleştirmesi ile insanlar ve kanser dokümanları ile fareler ve kanser dokümanlarında var olan etiketlerden en sık kullanılanı "cell" (hücre) ve "cancer" (kanser) kelimeleri olduğu görülmüştür. Fareler ve kanser dokümanlarında hücrenin daha ön planda olduğu, insanlar ve kanser dokümanlarında ise kanser kelimesinin daha sık geçtiği görülmektedir. İnsanlarla ilgili "haplotyp" (gen terimi) , "journal" (dergi), "vaccin" (aşı), "patient" (hasta), "tumor" (tümör), "molecular" (moleküler),

“breast” (meme), “medicin” (tıp) gibi kelimeler frekanslar tarafından ağırlıklandırılarak bu dokümanlarda en fazla söz edilen kelime grupları olmuşlardır. Benzer şekilde fareler için yapılan araştırmalarda ise hücre ve kanser kelimeleriyle birlikte sırasıyla sırasıyla “nutrient” (besin) , “effect” (tesir), tumor, “carcinoma” (bir çeşit ur), “liver” (karaciğer), “model”, “drug” (ilaç), “patient” (hasta) gibi kelimelerin ön plana çıktığı görülmektedir.

Tag Cloud yani etiket bulutu görselleştirmesi ile insanlar ve kanser dokümanları ile fareler ve kanser dokümanlarında birlikte yer alan etiketlerden en sık kullanılanları daha büyük harflerle yazılarak görselleştirilmiştir. Burada iki dokümanda da TF değeri en yüksek olan ve en sık geçen kelime “cell” yani “hücre” kelimesiyle “cancer” kelimesi olduğu görülmüştür. Fareler ile ilgili kanser araştırmalarında hücre ve kanser kelimeleriyle birlikte birlikte sırasıyla “nutrient” (besin) , “effect” (tesir), “tumor” (tümör), “carcinoma” (bir çeşit ur), “liver” (karaciğer), “model” (model), “drug” (ilaç), “patient” (hasta), “pancreat” (pankreas), “oncolog” (onkolog) gibi kelimeler ön plana çıktığı görülmektedir. Aynı şekilde insanlarda görülen kanser vakalarıyla ilgili verilerin analizi sonucu elde edilen etiket bulutunda hücre ve kanser etiketleriyle birlikte bir gen terimi olan “haplotyp” kelimesinin yanı sıra “journal” (dergi), “vaccin” (aşı), patient (hasta), “tumor” (tümör), “molecular” (moleküler), “breast” (meme), gibi kelimeler frekanslar tarafından ağırlıklandırılarak bu dokümanlarda en çok söz edilen kelime grupları olmuşlardır. Pubmed’ten elde edilen bu dokümanlarda en sık geçen kelime “hücre” kelimesidir. “İnsanlar ve kanser” ile “fareler ve kanser” vakalarıyla ilgili elde edilen metinlerde geçen kelimelerin sıklığı, bu iki vakanın insan ve hayvan hücreleriyle çok güçlü bir ilişkisi olduğunu öne çıkarmaktadır ve terim frekans değeri de bu bağlamda yüksektir. Bu iki kategoride de “tumor” kelimesinin etkili olduğu ön plana çıkmaktadır. Terim frekans değeri “cell” yani “hücre” kelimesine göre nispeten daha düşüktür.

Aynı çalışma Tag Filter düğümü kullanılmadan yapıldığında görülmektedir ki “and”, “the”, “with”, “were”, “that” gibi bağlaçlar metin içerisinde çok sık kullanıldığından filtrelenmemiştir. Asıl anlamlı olması gereken diğer terimler ise daha geri planda kaldığından dolayı bu tür gürültülü verilerin metin madenciliği çalışmalarında Tag

Filter ile filtrelenmesi önem arz etmektedir. Aksi takdirde yapılan çalışma hem süre olarak daha uzun bir süreye yayılacak, çalışmanın verimliliği azalacak, gereksiz veri kullanılması nedeniyle maliyet artacak ve istenilen sonuca ulaşılmada güçlük yaşanacaktır.

Bu adımlardan sonra K-NN algoritmasıyla doküman sınıflandırması yapmak amacıyla doküman vektörleri oluşturulmuş, sonra bu veriler eğitim ve test grubu olarak ikiye ayrılmıştır. Burada K-NN algoritması, eğitim verilerini kullanarak bir test verisi kümesi sınıflandırır. 600 adet olan test dokümanlarının 255 tanesi insan ve kanser sınıfına ait olan dokümanlar olarak sınıflanmış olup bunların 157 tanesi insan ve kanser sınıfına ait olduğu yani doğru tahmin edildiği görülmektedir ve kesinlik değeri %61.6 olarak hesaplanmıştır. Bu precision yani pozitif tahmini değer hesaplaması, fareler ve kanser verilerinin sınıflandırılması için de yapılmış olup, tespit edilen 345 dokümanın 202 tanesinin gerçekten fareler ve kanser sınıfına ait dokümanlar olduğu %58.6 oranıyla ortaya konulmuştur. Benzer şekilde algoritma tarafından doğru tahmin edilen her iki sınıfa ait doküman sayısının, gerçek test verisi miktarına oranları hesaplanarak recall / sensitivity yani duyarlılık değeri hesaplanmış ve bunlar insanlar ve kanser sınıfına ait dokümanlar için %52.3 fareler ve kanser dokümanlarında %67.3 olarak elde edilmiştir. Burada fareler ve kanser dokümanlarına ait sınıflamada yüksek oran göze çarpmaktadır. Yine bu iki ölçütten yola çıkılarak hesaplanan F ölçütüne göre; insanlar ve kanser dokümanlarının %56.6'lık, fareler ve kanser dokümanlarının ise %62.6'lık bir doğru sınıflandırılma yüzdesi elde edilmiştir. Sonuç olarak kullanılan K-NN algoritması ile %59.8 oranında kısmen başarılı bir doküman sınıflama tahmini yapılmıştır.

6. KAYNAKLAR

1. Pelin Y, Derya B. Bulut bilişimde veri madenciliği tekniklerinin uygulanması: Bir literatür taraması. Pamukkale Ün. Müh. Bilim Derg. 2018; 24(2): 336-343. DOI: 10.5505/pajes.2017.65642.

2. Alabay N. Büyük veri (Big data), https://dralabay.wordpress.com/2014/01/20/buyuk-veri-big-data/, Erişim Tarihi: 04.03.2019.

3. Thompson P, Batista-Navarro RT, Kontonatsios G, Carter J, Toon E., McNaught J, Timmermann C, Worboys M, Ananiadou S. Text Mining the History of Medicine. PlosOne. 2016; Jan 6;11(1):e0144717. doi: 10.1371/journal.pone.0144717, eCollection 2016.

4. Teorey TJ. Database Modeling & Design. USA, San Francisco: Morgan Kaufmann Publishers, 1998: 5-11.

5. Giudici P. Applied Data Mining: Statistical Methods for Business and Industry. England: Wiley, 2003: 10-17.

6. Hand D, Mannila H, Smyth P. Principles of Data Mining. Cambridge, MA, MIT Press, 2001: 8-12.

7. Gökay Emel G, Taşkın Ç. Veri Madenciliğinde Karar Ağaçları ve Bir Satış Analizi Uygulaması. Eskişehir Osmangazi Üniversitesi Sosyal Bilimler Dergisi. 2005; 6(2): 224.

8. Fayyad U, Piatetsky-Shapiro G, Smyth P. Knowledge Discovery and Data Mining: Towards a Unifying Framework. Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96), Portland, 1996: 12-21.

9. Şen F. Veri Madenciliği ile Birliktelik Kurallarının Bulunması. 2008, Sakarya Üniversitesi, Fen Bilimleri Enstitüsü, Yüksek Lisans Tezi, 8-11, Sakarya, (Yrd. Doç. Dr. Nilüfer YURTAY).

10. Bölükbaş MA. Veri Madenciliği Teknikleri Kullanılarak Çalışan Memnuniyetinin İncelenmesi. 2013, Mimar Sinan Güzel Sanatlar Üniversitesi, Fen Bilimleri Enstitüsü, Yüksek Lisans Tezi, 9-11, İstanbul, (Yrd. Doç. Dr. Elif Özge ÖZDAMAR).

11. Nisbet R, Elder J, Miner G. Handbook of Statistical Analysis and Data Mining Applications. Canada: Elsevier Inc., 2009: 221-231.

12. Kantardzic M. Data Mining: Concepts, Models, Methods and Algorithms, New Jersey, Wiley-IEEE, 2011: 56-75.

13. Roiger RJ. Data Mining: A Tutorial-Based Primer. North West, CRC Press, 2017: 25-34.

14. Visa A. Technology of Text Mining. Tampare University of Technology, 2001: 1-11.

15. Soucy P, Mineau W. Beyond TFIDF Weighting for Text Categorization in the Vector Space Model. IEEE International Conference, Edinburgh-Scotland, July 30- August 2005: 1130-1135.

16. Cerrito P. Inside text mining: Text mining provides a powerful diagnosis of hospital quality rankings - Data Warehousing/Mining, http://findarticles.com/p/articles/mi_m0DUD/is_3_25/ai_114167705/pg_2, Erişim Tarihi: 20.12.2017.

17. Sehgal AK. Text Mining: The Search for Novelty in Text, http://www.cs.uiowa.edu/~sehgal/Papers/comp04.pdf, Erişim Tarihi: 12.02.2017.

18. Kostoff RN, DeMarco RA. Information extraction from scientific literature with text mining, http://www.onr.navy.mil/sci_tech/special/technowatch/kdocs/anchem2/txt, Erişim Tarihi: 22.02.2017.

19. Losiewicz P, Oard DW, Kostoff RN. Textual Data Mining to Support Science and Technology Management. Journal of Intelligent Information Systems. 2000; 15(2): 99-119.

20. Miner G, Elder IV John, Hill T, Nisbet R, Delen D, Fast, A. Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications. USA, San Francisco: Academic Press, 2017: 5-91.

21. Veri Madenciliği ders notları, http://kergun.baun.edu.tr/veri_madenciligi_hafta11.pdf, Erişim Tarihi: 10.08.2017.

22. Veri Analitiği, http://www.verianalitigi.org/metin-madenciligi/metin-madenciliginin-asamalari/, Erişim Tarihi: 11.10.2017.

23. Berry MW, Kogan J. Text Mining Applications and Theory. UK, West Sussex: Wiley, 2018: 881-893.

24. Akın AA, Akın MD, Zemberek an Open Source NLP Framework for Turkic Languages. Structure. 2007; 10: 1-5.

25. Hotho A, Nürnberger A, Paab G, A Brief Survey of Text Mining, GLDV Journal for Computational Linguistics and Language Technology. 2005: 3-10.

26. İlhan U. Application of KNN and FPTC Based Text Categorization Algorithms to Turkish News Reports. 2001, Bilkent University, Institute of Engineering and Science, Thesis, 3-17, Ankara, (Assoc. Prof. Dr. Halil Altay GÜVENİR).

27. Amasyalı F, Yıldırım T. Otomatik Haber Metinleri Sınıflandırma. Yıldız Teknik Üniversitesi, İstanbul, 2004, 224-226.

28. Kutlu F. Categorization In A Hierarchically Structured Text Database. Bilkent University, İstanbul, 2001.

29. İlhan S, Duru N, Karagöz Ş, Sağır M. Metin Madenciliği ile Soru Cevaplama Sistemi, Elektronik ve Bilgisayar Mühendisliği Sempozyumu (ELECO), Bursa, 26-30 Kasım 2008: 356-359.

30. Aşlıyan R, Günal K. Metin İçerikli Türkçe Doküman Sınıflandırılması. Akademik Bilişim Konferansı Bildirileri, Muğla Üniversitesi, Muğla, 10-12 Şubat 2010: 659-665.

31. Çalış K, Gazdağı O, Yıldız O. Reklam İçerikli Epostaların Metin Madenciliği Yöntemleri ile Otomatik Tespiti, Bilişim Teknolojileri Dergisi. 2013; 6(1): 1-7.

32. Soucy P, Mineau W. Beyond TFIDF Weighting for Text Categorization in the Vector Space Model. IEEE International Conference, Edinburgh-Scotland, July 30- August 2005: 1130-1135.

33. Lam C, Lai FC, Wang CH, Lai MH, Hsu N, Chung MH. Text Mining of Journal Articles for Sleep Disorder Terminologies. Plos One. 2016; 11(5):e0156031. doi: 10.1371/journal.pone.0156031.

34. Hao H, Zhang K. The Voice of Chinese Health Consumers: A Text Mining Approach to Web-Based Physician Reviews. J Med Internet Res. 2016; 18(5):e108. doi: 10.2196/jmir.4430.

35. Mahgoub H, Rösner D, Ismail N, Torkey F. A Text Mining Technique Using Association Rules Extraction. International Journal of Computational Intelligence. 2008; 4:1.

36. About Knime Home, https://www.knime.com/about, Erişim Tarihi: 22.03.2019

37. Pilavcılar İF. Metin Madenciliği ile Metin Sınıflandırma. 2007, Yıldız Teknik Üniversitesi, Fen Bilimleri Enstitüsü, Yüksek Lisans Tezi, 10-17, İstanbul, (Yrd. Doç. Dr. Nilgün Güler BEYAZIT).

ÖZGEÇMİŞ

30.10.1983 yılında Gebze'de doğdum. İlkokul, ortaokul ve lise eğitimimi Ankara'da tamamladıktan sonra 2002 yılında Ondokuz Mayıs Üniversitesi Fen-Edebiyat Fakültesi İstatistik Bölümünü kazanıp, 2008 yılında üniversiteden mezun oldum. 2010 yılında yabancı dil eğitimim için Long Island University, ingilizce dil programına bir yıl katıldım. 2013 Eylül ayında Düzce Üniversitesi Sağlık Bilimleri Enstitüsü Biyoistatistik ve Tıbbi Bilişim Anabilim Dalında yüksek lisans eğitimime başladım. 2019 yılının Ağustos ayında "Metin Madenciliği ve Sağlık Alanında Bir Uygulama" isimli tezimle yüksek lisans eğitimimin sonuna gelmiş bulunmaktayım.