**TÜRKİYE AÇIK KAYNAK PLATFORMU**

**ONLİNE YARIŞMA PROGRAMI**

**TÜRKÇE DOĞAL DİL İŞLEME**

**METİN MADENCİLİĞİ LE MEVLANA’NIN MESNEVİ’SİNİN İNCELENMESİ**

**Takım Adı: Teknodil**

**TAKIM ÜYELERİ:**

**LEYLA AKÜZÜM**

**BURAK DEDEOĞLU**

**İÇİNDEKİLER**

**Proje Özeti**

**Projenin Odak aldığı Sorun ve Sağladığı Çözüm ……………………………………………………………………………….**

**Giriş**

**1.Veri Madenciliği ……………………………………………………………………………………………………………………………**

**2.Metin Madenciliği …………………………………………………………………………………………………………………………**

**3.Söz Varlığı …………………………………………………………………………………………………………………………………….**

**Yöntem**

**Kaynakça**

**Proje Özeti**

Bu çalışmanın amacı; Türk-İslam dünyasının yetiştirdiği önemli bir şahsiyet olan Mevlana’nın çok derin konuları bile rahat ve anlaşılır bir dille anlattığı Mesnevi’sini veri madenciliği ve metin madenciliği yöntemlerinin farklı yönleri ile ele almaktır. Aynı zamanda dijital çağda dil, edebiyat, tarih ve felsefe alanındaki çalışmalarda da teknolojinin kullanılmasını yaygınlaştırmak ve araştırmacılara bazı öngörüler sağlamak amaçlanmıştır. Proje kapsamında metin madenciliğinin çeşitli yöntemleri kullanılarak metnin söz varlığına ait sınıflamalar, kavramsal analizler ve örüntüler bulunacaktır. Ayrıca edebi metinler sayısallaştırılacak ve Mevlana’nın Mesnevi’si metin verisinden anlamlı bilgiler çıkarmak ihtiyacıyla dil bilim çalışmalarında yazılım yöntemlerinin kullanılmasının önemine dikkat çekilecektir. Bu çalışmada kullanılacak veri setinin depolanması için Google firmasına ait bulut tabanlı bir depolama sistemi olan Google Drive programı seçilmiştir. Depolanan veri üzerinde veri analizlerinin yapılması ve gerekli kodların yazılması için Python programlama dili ve bu dile ait NumPy, Pandas, Seaborn gibi veri madenciliğinde sıkça kullanılan kütüphaneler kullanılacaktır. Bu analiz çalışmaları ile elde edilen çıktıların ileri seviyede görselleştirilmesi için Microsoft firmasına ait Power BI programı kullanılacaktır.

**Projenin Odak aldığı Sorun ve Sağladığı Çözüm**

Veri madenciliği kurumlardaki milyonlarca verinin ihtiyacı karşılayacak şekilde bilişim sektörleri tarafından analiz edilmesidir. Metin madenciliğinde ise metin; veri olarak kabul edilir ve örüntüler doğal dil metinlerinden çıkarılır. Metin madenciliğinin dilbilim araştırmalarında kullanılmasıyla yapılan işlemler, yüksek kalitede yeni bilgi üretimi sağlamaktadır (Anandarajan vd. 2019, s. 2). Dilbilimde söz varlığı incelemeleri çeşitli yöntemlerle yapılabilmektedir ancak geleneksel araçlarla büyük miktardaki metin verilerinin depolanması, işlenmesi ve analiz edilmesi zordur. Bu bağlamda proje kapsamında metin madenciliğinin çeşitli yöntemleri kullanılarak metnin söz varlığına ait sınıflamalar, kavramsal analizler ve örüntüler bulunacaktır. Ayrıca edebi metinler sayısallaştırılacak ve Mevlana’nın Mesnevi’si metin verisinden anlamlı bilgiler çıkarmak ihtiyacıyla dil bilim çalışmalarında yazılım yöntemlerinin kullanılmasının önemine dikkat çekilecektir.

**GİRİŞ**

Hoşgörü ve kardeşliğin evrensel temsili olan Mevlana Celaleddin Rumi 13. Yüzyılda yaşamış ve tüm dünyaya mal olmuş önemli bir düşünür, şair, İslam âlimi ve sufidir. Mevlana gibi sesini, sözünü dünyanın her kesimine duyurabilmek, her kesimden insanı etkileyebilecek bir yazar olmak pek de kolay değildir. Öyle ki eserleri, onlarca dile çevrilmiş ve ardında bıraktığı sema geleneği UNESCO İnsanlığın Somut Olmayan Kültürel Mirası Temsili Listesi’ne girmiş, her yıl yapılan Şeb-i Arus törenleri ile de evrensel bir boyut kazanmıştır. Dünya edebiyatının önemli eserleri arasında görülen Mesnevi’den Alman yazar Goethe de etkilenmiş ve eseri daha iyi anlayabilmek için farsça öğrenmeye çalışmıştır (Mitler, 1988: 86). Kuşkusuz, bir eserin asırlar geçmesine rağmen hala okunmasının, ölümsüz olarak kabul edilmesinin birçok sebebi vardır. Edebiyat eserlerini anlamak, yorumlamak ve değerlendirmek için eserin bir varlık/veri olarak ele alındığı birçok geleneksel ve modern yaklaşım bulunmaktadır. Bu bağlamda çalışmanın amacı; Türk-İslam dünyasının yetiştirdiği önemli bir şahsiyet olan Mevlana’nın çok derin konuları bile rahat ve anlaşılır bir dille anlattığı Mesnevi’sini veri madenciliği ve metin madenciliği yöntemlerinin farklı yönleri ile ele almaktır. Aynı zamanda dijital çağda dil, edebiyat, tarih ve felsefe alanındaki çalışmalarda da teknolojinin kullanılmasını yaygınlaştırmak ve araştırmacılara bazı öngörüler sağlamak amaçlanmıştır.

**1. Veri Madenciliği**

Veri madenciliği, kurumlardaki milyonlarca verinin ihtiyacı karşılayacak şekilde bilişim sektörleri tarafından analiz edilmesidir. Özellikle işletmelerin milyonlarca veriyi yazılım sistemleri yoluyla işleyip anlamlı hale getirmesi gerekir. Veri madenciliğinde esas amaç; şahıslar ve işletmeler için işlevsel değeri olan, faydalanabilecekleri verileri anlamsız ve işlevsiz veriden ayırarak belirli tekniklerle işlemek ve kullanışlı hale getirebilmektir. Veri sayısı özellikle İnternet ortamında geçirilen süre nedeniyle her geçen gün hızla arttığı için veri madenciliği alanı önem kazanmıştır. Veri madenciliği uygulama alanları; pazar sepeti analizi, risk yönetimi, dolandırıcılık tespiti, müşteri değerlendirme, satın alma davranışlarının belirlenmesi gibi bölümlere ayrılabilir (Silahtaroğlu, 2008).

**2. Metin Madenciliği**

Metin madenciliği, doğal dil metninde yer alan yapısal olmayan verilerden çeşitli yöntemler kullanılarak yapısal, düzenli ve anlamlı bilgiler elde etme sürecidir. Metin veri madenciliği olarak da bilinen metin madenciliği, doğal dil işleme algoritmaları ile ilgilenmektedir ((Weiss vd., 2010). Bu algoritmalar yoluyla yapılandırılmamış verilerden yüksek kalitede, işlevsel, düzenli, yapısal bilgiler elde edilmeye çalışılmaktadır (Bose, 2018). Metin madenciliği; Doğal dildeki metin verisinde yer alan anahtar kelimeleri, kavramları, konuları ve tanımlanabilecek diğer belirleyici özelliklerin ilgili yazılımlar tarafından incelenip analiz edilmesi süreci olarak tanımlanabilmektedir. Metnin analiz sürecinde veri kümeleri arasında sınıflama, sıralama yapmak ve ilişki kurmak için metin madenciliği teknikleri kullanılır (Rouse, 2018). Günümüzde bilişim alanındaki ilerlemelerle birlikte veri bilimciler ve diğer kullanıcılar için metin madenciliği, yapılandırılmamış büyük veri yığınlarını derin öğrenme algoritmaları yoluyla işleyip kullanışlı hale getirebilmektedir (Linguamatics, 2018).

**3. Söz Varlığı**

Söz varlığı, bir dilde yer alan sözcüklerin biçimsel ve anlamsal niteliklerinin yanı sıra sayısal durumlarını da yansıtan anlamlı ögelerdir. (ç13) Aksan, söz varlığı kavramı içinde şu ögelerin bulunduğunu ifade eder: “Bir dilin söz varlığı denince, yalnızca, o dilin sözcüklerini değil, deyimlerin, kalıp sözlerin, kalıplaşmış sözlerin, atasözlerinin, terimlerin ve çeşitli anlatım kalıplarının oluşturduğu bütünü anlıyoruz.” (Aksan, 2018 ). Başka bir deyişle kapsam olarak sözcükten daha geniş, kalıplaşmış sözler, deyim, atasözü gibi ifadelerin söz varlığı içinde yer aldığı söylenebilir. Bir dilin söz varlığı; o dili konuşan insanların maddi ve manevi kültürünü, dünyayı algılayış biçimini, ifadesini ve bu ifadelere dayalı üretilen sözcükler ve sözcük gruplarından oluşur (Yalçın, 2005: 25). Günümüzde metne dayalı söz varlığı incelemelerinde bilişim destekli araştırmalar artarak devam etmekte ve metin madenciliğine dayalı yöntemlerde de söz varlığı araştırmalarında kullanılabilmektedir. Metihlerdeki sözcük frekanslarının gösterilmesi ve bu sayısal göstergelerin metinlerde konu, kavram, sözcük vb. bulmada yararlanılmaktadır. Yine benzer şekilde eşdizim dağılım göstergeleri ve ilgili görselleştirmeler metin madenciliği ve dil bilim alanlarındaki söz varlığı araştırmalarında kullanılabilmektedir.

**Yöntem**

Çalışmada emekli Albay Şefik Can tarafından çevirisi yapılan Mevlana’nın Mesnevi’si esas alınmış ve metin sayısallaştırılarak metin işleme yazılımları için hazır duruma getirilecektir. Eserin tam metni üzerine birçok çalışma bulunmaktadır. Şefik Can tarafından yapılan çeviri metninin bu çalışmada tercih edilmesinin sebebi akademik alanda yaygın biçimde kullanılmasıdır. Metin işleme için farklı yazılımlar kullanılacaktır. Kullanılacak veri setinin depolanması için Google firmasına ait bulut tabanlı bir depolama sistemi olan Google Drive programı seçilmiştir. Depolanan veri üzerinde veri analizlerinin yapılması ve gerekli kodların yazılması için Python programlama dili ve bu dile ait NumPy, Pandas, Seaborn gibi veri madenciliğinde sıkça kullanılan kütüphaneler kullanılacaktır. Python programlama dili kullanılarak incelenecek metinde bulunan kelimeler, kullanım sıklıkları gibi analizler yapılacaktır. Bu analiz çalışmaları ile elde edilen çıktıların ileri seviyede görselleştirilmesi için Microsoft firmasına ait Power BI programı kullanılacaktır.

**Kaynakça**

Aksan, D. (2018). *Türkçenin Sözvarlığı* (2nd ed.). Ankara: Bilgi.

Anandarajan, M., Hill, C., ve Nolan, T. (2019). Practical Text Analytics: Maximizing the Value of Text Data. Advances in Analytics and Data Science: Vol. 2. Cham: Springer International.

Bose, B. (2018). Techniques and Aplications of Text Mining. https://www.digitalvidya.com/blog/techniquesapplications-text-mining/, (Erişim Tarihi: 10.06.2018).

Linguamatics (2018). What is NLP Text Mining?. https://www.linguamatics.com/what-is-text-mining-nlp-machine-learning, (Erişim Tarihi: 20.12.2018).

MİTLER, Louis. (1988). *Ottoman Turkish Writers: a Bibliographical Dictionary of Significant Figures in pre-Republican Turkish Literature*. New York; Bern; Frankfurt am Main; Paris: Lang.

Rouse, M. (2018). Text mining (text Analytics). https://searchbusinessanalytics.techtarget.com/definition/text-mining, (Erişim Tarihi: 20.12.2018).

Silahtaroğlu, G. (2008). *Kavram ve Algoritmalarıyla Temel Veri Madenciliği*. İstanbul: Papatya.

Weiss, S. M., Indurkhya, N., Zhang, T. ve Damerau, F. (2010). Text mining: predictive methods for analyzing unstructured information. Springer Science & Business Media.