Gelişmiş Metin Madenciliği Teknikleri ile Optimum Konu Sayısının Araştırılması

Müşerref Özkan, Yıldız Teknik Üniversitesi, muserref.ozkan@std.yildiz.edu.tr

1. Özet

Gelişmiş metin madenciliği teknikleri ile optimum konu sayısının araştırılması: Sürdürülebilir Enerji Araştırmaları (Amer Farea, Shailesh Tripathi, Galina Glazko, Frank Emmert-Streib) makalesinde uygulanan yöntemler bu proje kapsamında uygulanmıştır. Veri kümesini tutarlı konular halinde düzenlemek ve sınıflandırmak için, Gizli Dirichlet Tahsisi (LDA) ve BERTopic algoritmaları kullanılarak karşılaştırma yapılmıştır.

2. Giriş

Konu modelleme, metinsel veri setlerindeki gizli temaları ortaya çıkararak etkili bilgi erişimi ve tematik analiz sağlamada kritik bir rol oynamaktadır. Metin verilerinin hacminin artmasıyla, Latent Dirichlet Allocation (LDA) ve BERTopic gibi ileri düzey teknikler, büyük ölçekli veri analizi için öne çıkmıştır. LDA, temaları belirlemek için olasılıklı bir çerçeve sağlarken, BERTopic, bağlamsal olarak zengin temsiller için dönüştürücülere dayalı modeller kullanır. Bu çalışma, bilimsel makalelerin başlık ve özetlerinden oluşan bir veri setini analiz etmek için bu yöntemleri inceleyerek optimal konu sayısını belirlemeyi amaçlamaktadır.

Bu çalışmanın hedefleri:

- Etkili modelleme için metin verilerinin önişlenmesi ve temizlenmesi.
- Anlamlı konular çıkarmak için LDA ve BERTopic'in uygulanması.
- Sonuçları karmaşıklık ve tutarlılık puanlarıyla değerlendirmek.

3.Yöntem

3.1. Veri Hazırlama

Doğal dil işleme ve metin analizi alanında, ön işleme, sonraki analiz ve modelleme için ham metin verilerinin rafine edilmesinde çok önemli bir rol oynar.

Bu bağlamda örnek makale incelendiğinde yazar adları, dil, belge türü, atıf yapılan zamanlar, yayıncı ile ilgili bilgiler, dergi kısaltması, cilt, sayı, kategoriler vb. gibi ilgisiz veriler çıkarılırken içgörü ve çıkarımlar içeren makale başlığı, anahtar kelimeler, özet, yayın tarihi ve kaynak başlığı verilerinin birleştirildiği görülmüştür.

Çalışma için kullanılan veri setinde makaleler hakkında bilgiler bulunmaktadır. Veri seti, 8989 satır ve id, başlık ve özet olmak üzere üç sütundan oluşmaktadır. Id bilgisinin çalışmaya bir katkısı olmayacağı için kaldırılmış, başlık ve özet sütunları örnek alınan makalede olduğu gibi birleştirilmiştir.

Ön işleme için hazır hale gelen veri setine örnek makalede olduğu gibi noktalama işaretlerinin kaldırılması, metnin küçük harfe dönüştürülmesi ve kelimelerin köklerine indirgenmesi, durdurma kelimelerinin ve üç harften az olan tokenlerin çıkarılması işelmleri uygulanmaktadır. Bu ön işleme adımları metin verilerinin kalitesini ve tutarlılığını yükselterek ham metni çeşitli doğal dil işleme görevleri için yönetilebilir ve bilgilendirici bir biçime dönüştürmeye daha uygun hale getirir.

3.2. Latent Dirichlet Allocation (LDA)

LDA, her belgenin bir konu karışımı ve her konunun kelimelerden oluşan bir karışım olduğunu varsayan olasılıklı bir modeldir. Model, bir konuya verilen her kelimenin olasılığını ve bir belgeye verilen her konunun olasılığını belirler.

3.3. BERTopic

BERTopic, metin belgeleri için bağlama duyarlı katıştırmalar oluşturmak üzere dönüştürücü tabanlı modellerden yararlanmaktadır. Konuları tanımlamak için bu katıştırmalar üzerinde kümeleme algoritmaları kullanır ve bağlamsal olarak daha zengin temsiller sağlar.

3.4. Optimal Konu Sayısının Belirlenmesi

Optimal konu sayısını belirlemek için karışıklık (perplexity) ve tutarlılık (coherence) skorları kullanılmıştır.

Karışıklık skoru, modelin metin verisini ne kadar iyi tahmin ettiğini değerlendirir. İdeal olarak düşük karışıklık skorları, modelin daha başarılı olduğunu gösterir.

Tutarlılık skorları, bir konu içerisindeki kelimelerin anlamsal bağlılığını ölçer:

UCI Coherence**:** Kelime çiftleri arasındaki anlamsal benzerliği değerlendirir. Daha yüksek skorlar, daha tutarlı konuları ifade eder.

UMass Coherence**:** Kelime çiftlerinin doküman içindeki frekanslarına dayanır. Daha negatif skorlar daha iyi tutarlılığı ifade eder.

CV Coherence: Noktasal karşılıklı bilgi (NPMI) ve kosinüs benzerliği gibi metrikler ile hesaplanır. Daha yüksek skorlar daha iyi bağlılığı temsil eder

Optimum konu sayısının belirlenmesi için ölçütler birleştirilerek değerlendirilmiştir. Bu amaçla, uygulana optimizasyonun ilk adımında, skorlara normalizasyon işlemi uygulanmıştır.

Min-Max Normalizasyonu: Skorları 0 ile 1 arasına çeker.

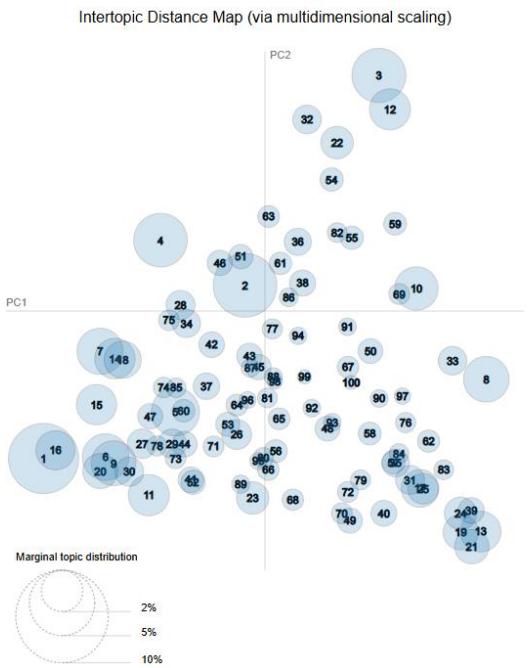
Ortalama Normalizasyon: Skorları, ortalamadan uzaklığına göre normalize eder.

T-Skor Normalizasyonu: Skorların standart sapma ve ortalamaya göre normalizasyonunu yapar.

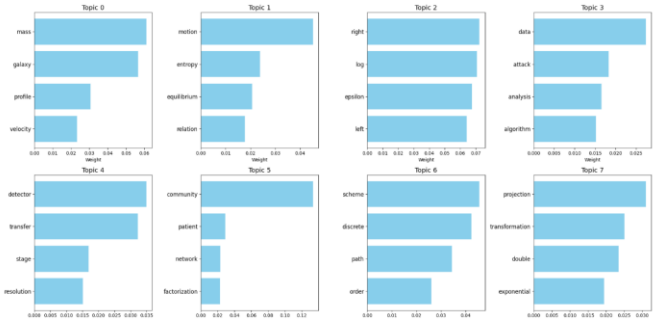
Skorlar üzerine uygulanan normalizasyon sonuçları lamda değerinin etkisiyle birleştirilerek birleşik skor değeri elde edilmiştir. Lambda değerleri, farklı normalizasyon yöntemlerinden (average normalization, min-max normalization, ve t-score normalization) gelen skorların birleşik skora katkısını ağırlıklı olarak belirlemek için kullanılır. Farklı metriklerin önem derecesini ayarlamak ve analiz edilen skorların etkisini dengelemek amacıyla kullanılır. Sonuçlar değerlendirilerek optimal konu sayısı belirlenmiştir.

3.DENEYSEL ANALİZ

Bu bölümde, proje kapsamında gerçekleştirilen deneysel çalışmaların sonuçları detaylı bir şekilde incelenmiştir. Öncelikle, Latent Dirichlet Allocation (LDA) ve BERTopic yöntemleri kullanılarak veri kümesindeki temaların modellenmesi ve görselleştirilmesi sağlanmıştır. LDA ile optimal konu sayısının belirlenmesi ve farklı konu dağılımlarının analiz edilmesi hedeflenirken, BERTopic modeli bağlamsal güçlü temsil yeteneği sayesinde tematik yapıları daha ayrıntılı bir şekilde ortaya koymuştur. Görseller ve metrikler yardımıyla her iki yöntemin güçlü ve zayıf yönleri karşılaştırılmış, elde edilen sonuçlar detaylandırılmıştır. Bu analizler, metin verisi üzerindeki tematik ayrıştırmanın etkinliğini değerlendirmenin yanı sıra, kullanılan yöntemlerin performansını kapsamlı bir şekilde ortaya koymaktadır.

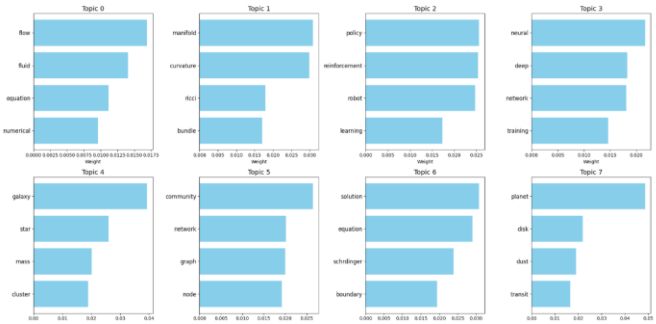


Şekil1. LDA modelinin tüm konuları için konular arası mesafe haritası



Şekil2. LDA modelinin tüm konuları için 8 başlıkta bulunan ilk dört terim.

Şekil3. BERTopic modelinin tüm konuları için konular arası mesafe haritası



Şekil4. BERTopic modelinin tüm konuları için 8 başlıkta bulunan ilk dört terim.

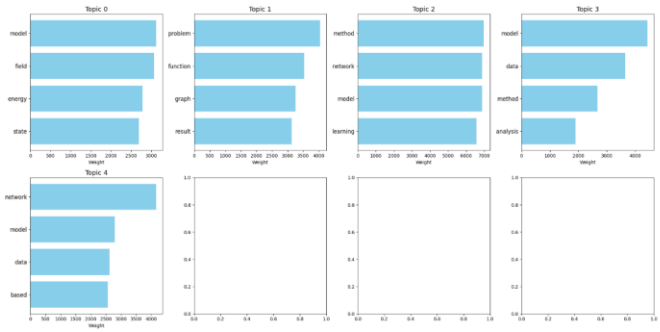
Şekil1'de dairelerin boyutları, o konuya ait dokümanların toplam ağırlığını ifade ederken, daireler arasındaki mesafeler ise konular arasındaki anlamsal yakınlığı yansıtır. Bu haritalar, modeldeki konuların birbirine ne kadar bağımsız veya benzer olduğunu anlamak için kullanılır. Kısıtlanmamış bir konu sayısında, her iki harita (Şekil1, Şekil3) içinde yoğun bir şekilde dağılmış, birden çok küçük ve birbirine yakın grup bulunmaktadır. Bu durum, modelin belirli bir sayıda anlamlı ve ayrıştırılmış konuya ulaşmakta zorlandığını işaret eder. Optimal bir konu sayısı belirlenerek bu harita, daha düzenli ve anlamlı bir yapıya dönüşebilir.

Şekil2 ve Şekil 4 için terimlerin ağırlıkları, ilgili konunun bu terimlerle ne kadar güçlü bir şekilde temsil edildiğini gösterir. Daha yüksek ağırlığa sahip kelimeler, o konunun belirgin anahtar kelimeleri olarak değerlendirilebilir. Her bir konu, diğerlerinden belirgin bir şekilde ayrılan terimler içeriyor. Ancak bazı konular arasında örtüşme potansiyeli bulunabilir; bu da optimal bir konu sayısı belirlenerek daha iyi ayrıştırılabilir.

Optimal konu sayısı, örnek makalede de olduğu gibi LDA modeli üzerinden belirlenmiştir. Karışıklık ve tutarlılık (UCI, UMass, CV) skorları üzerinde normalizasyon işlemi uygulanarak lamda değerleri de kullanılarak birleşik skor değerleri elde edilmiştir. Hesaplama sırasında lambda değerleri için 0.25, 0.5, 0.75, 1.0 değerlerinin kombinasyonları denenmiştir. Birleşik skoru en yük değerler lambda için [1.0, 1.0, 1.0, 1.0] olacak şekilde eşit ağırlıklarken konu sayısı 20'dir. Bu konu sayısı üzerinden LDA ve BERTopic için uygulama gerçekleştirilmiştir.

Şekil5. Konu sayısı 16 için LDA modelinin konular arası mesafe haritası

Şekil 5'e göre konuların birçoğu birbirinden belirgin bir şekilde ayrılmıştır. Bu durum, optimal konu sayısının daha iyi tematik ayrışma sağladığını gösterir. Örneğin, "Topic 1", diğer konulardan net bir şekilde ayrışmış, belirgin bir grup oluşturmuştur. Bazı konular ise (örneğin "Topics 2, 4, 14"), birbiriyle daha yakın mesafede yer almakta, bu da ilgili konular arasında potansiyel bir örtüşmeyi ifade edebilir.



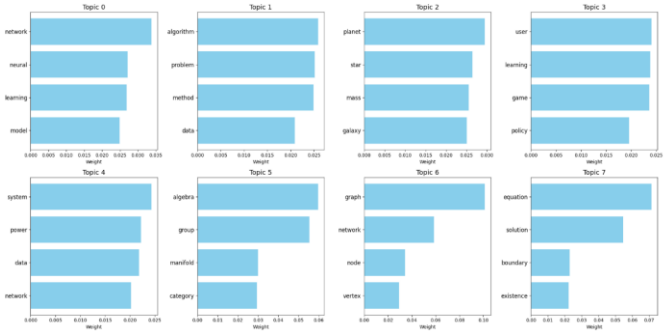
Şekil6. LDA modelinin 16 konu için 8 başlıkta bulunan ilk dört terim.

Şekil6'ya göre optimal konu sayısı (16) belirlendiğinde, LDA modeli daha anlamlı tematik ayrışma sağlamış ve konular arasındaki anlamsal ilişkileri daha net bir şekilde ortaya koymuştur. Konuların bazıları çok güçlü temalar içerirken, birkaç konu arasındaki potansiyel örtüşmeler, daha spesifik ayrıştırma gerekliliğine işaret edebilir.

Topic 2: "method", "network", "model", ve "learning" gibi kelimeler, makine öğrenmesi ve veri bilimi ile ilgili belirgin temaları ifade etmektedir.
Topic 0: "model", "field", "energy", ve "state" terimleri, fiziksel bilimler veya mühendislik konularını çağrıştırmaktadır.
Topic 4: "network", "model", ve "data" gibi kelimeler, ağ analizi veya veri analitiğiyle ilişkili olabilir.

Şekil7. Konu sayısı 16 için BERTopic modelinin konular arası mesafe haritası

Şekil7'ye göre bazı konular birbirine daha yakın bir şekilde gruplanmıştır. Bu, yakın konuların içerik açısından örtüşme potansiyeline işaret edebilir. Ancak yakın konuların diğerleri ile ayrıştığı da açık şekilde görülmektedir.



Şekil8. BERTopic modelinin 16 konu için 8 başlıkta bulunan ilk dört terim.

Şekil8' bakıldığında aşağıdaki çıkarımlar yapılabilir.
Topic 0: "network", "neural", "learning", "model" gibi kelimeler, makine öğrenimi ve sinir ağlarıyla doğrudan ilişkilidir. Bu, bu konunun veri bilimiyle bağlantılı bir tema olduğunu gösterir.
Topic 1: "algorithm", "problem", "method", "data" gibi kelimeler, algoritma geliştirme ve problem çözme odaklı bir konuyu ifade eder.
Topic 2: "planet", "star", "mass", "galaxy" gibi kelimeler, bu konunun astrofizik veya uzay bilimi ile ilgili olduğunu açıkça ortaya koymaktadır.
Topic 3: "user", "learning", "game", "policy" gibi kelimeler, kullanıcı deneyimi, oyun ve politika odaklı bir temayı işaret etmektedir.
Topic 7: "equation", "solution", "boundary", "existence" gibi kelimeler, matematiksel modelleme veya teorik bir konuyla ilişkilendirilebilir.
Ağırlıkların Yoğunluğu: Her konu için en önemli dört kelime, o temanın anahtar noktalarını iyi bir şekilde temsil ediyor. Özellikle, "network" ve "learning" gibi yaygın terimler, birden fazla konuda benzer temalara işaret edebilir.

BERTopic'in daha geniş dağılımlar ve daha net ayrışmalar sağladığı gözlemlenebilir. Bazı konuların gruplar halinde yakın bir alanda bulunması, bu konular arasında potansiyel bir örtüşmeyi gösterebilir. Daha spesifik temalar ortaya koymak için bu konular yeniden analiz edilebilir. Konular için anahtar kelime seçimleri oldukça anlamlıdır, ancak daha fazla terim eklenerek konuların temsili genişletilebilir.

3.TARTIŞMA

LDA modeli ile konu sayısı başlangıçta sınırlandırılmadığında, konular arasındaki örtüşmeler ve anlamsal yakınlıkların yoğun olduğu bir yapı gözlemlenmiştir. Bu durum, görsellerdeki konu dağılımlarında çoklu grupların bir arada yer almasıyla desteklenmiştir.

Optimal konu sayısının 16 olarak belirlenmesi, daha anlamlı ve ayrıştırılmış temaların elde edilmesini sağlamış olsa da, bazı konuların hâlâ örtüşme gösterdiği ve bu nedenle anlamlı ayrıştırma yapılamadığı anlaşılmıştır.
Konu ağırlıklarında öne çıkan terimler, özellikle teknik ve bilimsel alanlarla ilişkili temaların ayrıştırılmasında güçlü bir katkı sunmuş olsa da, bazı genel terimlerin (örneğin "model", "data") birçok konuda tekrar ettiği gözlemlenmiştir. Bu durum, modelin tematik ayrıştırma kabiliyetini sınırlayan bir neden olabilir.

Uzaklık haritasında görüldüğü üzere, BERTopic konuları daha bağımsız ve net bir şekilde gruplandırmıştır. Bu durum, modelin veri kümesindeki tematik farklılıkları daha iyi yansıttığını ortaya koymaktadır.
Bununla birlikte, bazı konuların bir araya gruplaşması ve örtüşmesi, BERTopic için de kısmen bir sorun olarak devam etmektedir.

Optimal konu sayısının seçimi hem LDA hem de BERTopic modelleri için kritik bir parametredir. Daha iyi bir optimizasyon yöntemi uygulanabilir.

Proje kapsamında elde edilen bulgular, özellikle bilimsel makaleler gibi teknik ve karmaşık veri setlerinde, LDA ve BERTopic modellerinin güçlü yanlarını ve sınırlamalarını açıkça ortaya koymuştur. Bununla birlikte, kullanılan veri setinin sınırlılığı, sonuçların genellenebilirliğini kısmen etkilemektedir. Daha geniş ve farklı türde veri setleriyle modellerin performansının tekrar değerlendirilmesi, bu bulguların daha fazla genellenebilirliğini sağlayabilir.