T.C. İSTANBUL TİCARET ÜNİVERSİTESİ

LİSANSÜSTÜ EĞİTİM ENSTİTÜSÜ

DOĞAL DİL İŞLEME İLE OTOMATİK DOKÜMAN DOĞRULAMA

Ahmet TOPRAK

DOKTORA TEZİ
BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI
BİLGİSAYAR MÜHENDİSLİĞİ PROGRAMI
İSTANBUL – 2025

LİSANSÜSTÜ EĞİTİM / FEN BİLİMLERİ ENSTİTÜSÜ

DOĞAL DİL İŞLEME İLE OTOMATİK DOKÜMAN DOĞRULAMA

Ahmet TOPRAK
200008305

Danışman
Dr. Öğr. Üyesi Metin TURAN

DOKTORA TEZİ
BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI
BİLGİSAYAR MÜHENDİSLİĞİ PROGRAMI
İSTANBUL – 2025

AKADEMİK VE ETİK KURALLARA UYGUNLUK BEYANI

İstanbul Ticaret Üniversitesi, Fen Bilimleri Enstitüsü, tez yazım kurallarına uygun olarak hazırladığım bu tez çalışmasında,

- tez içindeki bütün bilgi ve belgeleri akademik kurallar çerçevesinde elde ettiğimi,
- görsel, işitsel ve yazılı tüm bilgi ve sonuçları bilimsel ahlak kurallarına uygun olarak sunduğumu,
- başkalarının eserlerinden yararlanılması durumunda ilgili eserlere bilimsel normlara uygun olarak atıfta bulunduğumu,
- atıfta bulunduğum eserlerin tümünü kaynak olarak gösterdiğimi,
- kullanılan verilerde herhangi bir tahrifat yapmadığımı,
- ve bu tezin herhangi bir bölümünü bu üniversitede veya başka bir üniversitede başka bir tez çalışması olarak sunmadığımı

beyan ederim.

21/01/2025

Ahmet TOPRAK

İÇİNDEKİLER

ÖZET

Doktora Tezi

DOĞAL DİL İŞLEME İLE OTOMATİK DOKÜMAN DOĞRULAMA

Ahmet TOPRAK

İstanbul Ticaret Üniversitesi
Lisansüstü Eğitim Enstitüsü
Bilgisayar Mühendisliği Anabilim Dalı
Bilgisayar Mühendisliği Programı

Danışman: Dr. Öğr. Üyesi Metin TURAN

2025, 127 sayfa

Doküman doğrulama, orijinal özet dokümanın orijinal tam metin doküman üzerinde doğrulanması işlemidir. Bu doğrulama süreçlerinde anlamsal kontrol oldukça kritiktir. Anlamsal doğrulamaya daha az odaklanan birçok güncel yaklaşımdan farklı olarak bu çalışmada, özellikle orijinal doküman veya dokümanlar için üretilen özetin tutarlılığını anlamsal olarak kontrol etmek amacıyla Doğal Dil İşleme tekniklerine dayalı bir otomatik doküman doğrulama sistemi tasarlanmıştır. Soyut özetlerin orijinal tam metin dokümanlar üzerinde doğrulanması Transformer tabanlı model aracılığıyla yapılmıştır. Çalışmada deneysel dokümanlar finansal türe ait seçilmiş olduğundan Reuters finansal veri seti ile eğitim yapılarak Transformer modeli oluşturulmuştur. Önerilen Transformer tabanlı anlamsal doküman doğrulama yaklaşımı, orijinal tam metin ve özet dokümanlar üzerinde test edilmiştir. Sistem, birçok Doğal Dil İşleme modelinde olduğu üzere, tam metin ve özet dokümanlar üzerinde veri ön işleme ve yazım denetimi işlemlerini uygulayarak başlamaktadır. Daha sonra özet doküman tam metin doküman üzerinde doğrulanacağı için Simhash ve Cross Encoder metin benzerliği algoritmaları kullanılarak tam metin doküman cümlelerinden özet doküman cümlelerine en çok benzeyen cümleler belirlenmiştir. Bu sezgisel bir yaklaşımdır ve özet içinde yer alan her cümlenin tam metin dokümanda hangi cümlelerle ilişkili olduğu tahmin edilmeye çalışılmaktadır. Özet doküman cümlesine en yakın (benzeyen) iki orijinal tam metin doküman cümlesi seçilmiştir. Daha sonra bu orijinal tam metin doküman

cümleleri eğitilmiş Transformer modeline girdi olarak verilmiştir, orijinal tam metin cümlelerinin soyut bir özeti üretilmiştir. Son aşamada orijinal özet ile Transformer modelinin ürettiği özet, benzerlikleri açısından hem Simhash hem de Cross Encoder metin benzerliği algoritmalarıyla karşılaştırılmış ve ortalama doküman doğrulama doğruluğu hesaplanmıştır. Önerilen Transformer tabanlı anlamsal doküman doğrulama sistemi, Reuters veri kümesindeki finansal dokümanlar üzerinde ortalama %84.1 anlamsal doküman doğrulama doğruluğu elde etmiştir. Bu çalışma, doküman doğrulamayı otomatikleştirmenin günümüz teknolojileri ile mümkün olduğunu göstermiştir. Hem Transformer hem de cümle gruplama tekniklerini ustalıkla bütünleştirerek otomatik soyut özetlemenin önünü açmış, insan özetlerinin yerine otomatik özetlerin başarı ile kullanılabileceğini göstermiştir.

Anahtar Kelimeler: *Adlandırılmış varlık tanıma, anlamsal analiz, derin öğrenme, doğal dil işleme, doküman doğrulama, finansal raporlar, soyut özetleme, Transformer, yazım denetleyicisi.*

ABSTRACT

Ph.D. Thesis

AUTOMATIC DOCUMENT VERIFICATION WITH NATURAL LANGUAGE PROCESSING

Ahmet TOPRAK

İstanbul Ticaret University
Institute of Graduate School
Department of Computer Engineering
Computer Engineering Program

Supervisor: Assist. Prof. Dr. Metin TURAN

2025, 127 pages

Document verification is the process of verifying an original summary document on the original full-text document. Semantic control is very critical in these verification processes. Unlike many current approaches that are less focused on semantic verification, in this study, an automatic document verification system based on Natural Language Processing techniques was designed to semantically check the consistency of the summary produced especially for the original document or documents. Verification of abstract summaries on original full-text documents was done through the Transformer-based model. Since the experimental documents in the study were selected to belong to the financial type, the Transformer model was created by training with Reuters financial dataset. The proposed Transformer-based semantic document verification approach was tested on the original full-text and summary documents. As in many Natural Language Processing models, the system starts by applying data pre-processing and Spell Checker operations on full-text and summary documents. Then, since the summary document will be verified on the full-text document, the sentences most similar to the summary document sentences from the full-text document sentences were determined by using Simhash and Cross Encoder text similarity algorithms. This is an intuitive approach, attempting to predict which sentences in the full-text document are associated with each sentence included in the summary. The two original full-text document sentences

closest (similar) to the summary document sentence were selected. Then, these original full text document sentences were given as input to the trained Transformer model, producing an abstract summary of the original full-text sentences. Finally, the transformer model produced an abstract summary of original full-text sentences. In the last stage, the original summary and the summary produced by the Transformer model were compared with both Simhash and Cross Encoder text similarity algorithms in terms of their similarities, and the average document verification accuracy was calculated. The proposed Transformer-based semantic document verification system achieved an average of 84.1% semantic financial document verification accuracy on the financial documents in the Reuters financial dataset. This study has demonstrated that document verification can be automated using modern-day technologies. By skillfully integrating both Transformer and sentence grouping techniques, the way has been paved for automated abstract summarization, showcasing the potential for the successful utilization of automated summaries in place of human-generated ones.

Keywords: *Document verification, natural language processing, deep learning, abstract summarization, semantic analysis, transformer, financial reports, named entity recognition, spell checker.*

TEŞEKKÜR

Bu çalışmanın yürütülmesi sırasında desteğini esirgemeyip, bilgi birikimi ve tecrübesi ile çalışmam boyunca ihtiyaç duyduğum her zaman yardım eden değerli danışman hocam Dr. Öğr. Üyesi Metin TURAN'a teşekkürlerimi sunarım.

Tez çalışmam boyunca yanımda olup, çalışmamın her aşamasında desteğini hiçbir zaman eksik etmeyen sevgili eşime ve eğitim hayatım boyunca her zaman desteklerini sunan ve yanımda olan sevgili aileme teşekkür ederim.

Ahmet TOPRAK
İSTANBUL, 2025

ŞEKİLLER

ÇİZELGELER

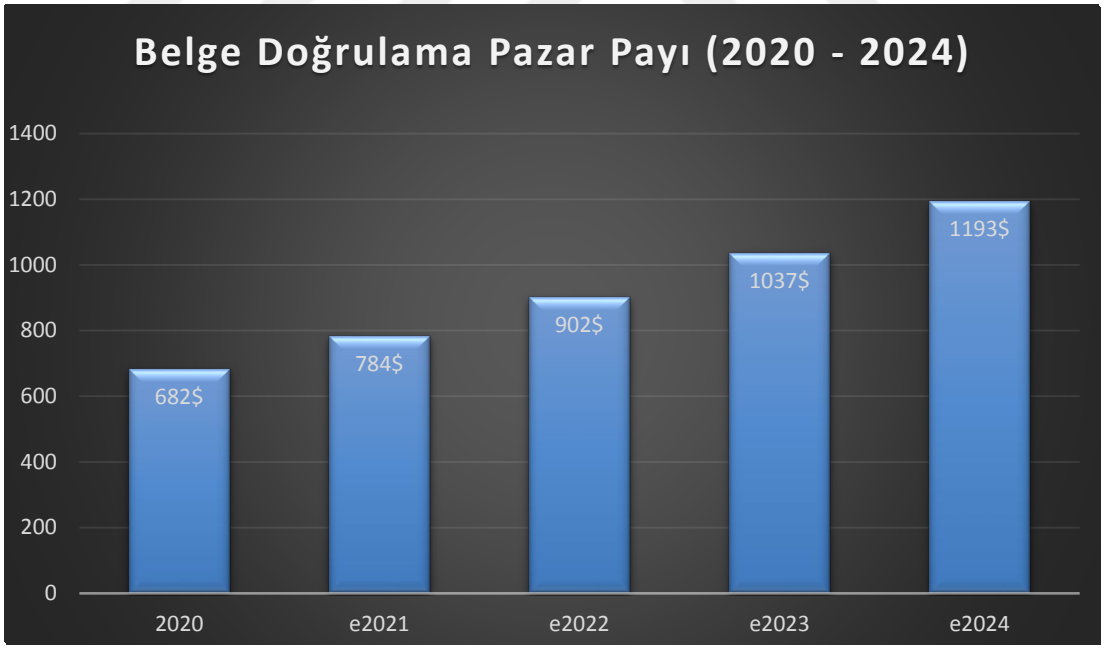
SİMGELER VE KISALTMALAR

BART	Bidirectional and Auto-Regressive Transformers
BERT	Bidirectional Encoder Representations from Transformers
BiLSTM	Bidirectional Long Short-Term Memory
BLEU	Bilingual Evaluation Understudy
CNN	Convolutional Neural Network
DDİ	Doğal Dil İşleme
GLUE	General Language Understanding Evaluation
GNN	Graph Neural Network
GSG	Gap Sentences Generation
GPT	Generative Pre-trained Transformer
HTML	Hypertext Markup Language
IDF	Inverse Document Frequency
IPFS	InterPlanetary File System
JSON	JavaScript Object Notation
LSA	Latent Semantic Analysis
LSTM	Long Short-Term Memory
MDS	Multi-Dimensional Scaling
MLM	Masked Language Model
MUC-6	Message Understanding Conference
MultiNLI	Multi-Genre Natural Language Inference
NLP	Natural Language Processing
OOV	Out-Of-Vocabulary
PDF	Portable Document Format
RNN	Recurrent Neural Network
ROUGE	Recall-Oriented Understudy for Gisting Evaluation
SVM	Support Vector Machine
SMIL	Synchronized Multimedia Integration Language
TB	Ticari Bilançolar
TF	Term Frequency
TF-IDF	Term Frequency-Inverse Document Frequency
TF-IDF-TR	Term Frequency-Inverse Document Frequency-TextRank
TFRSP	Text Frequency Ranking Sentence Prediction
TKP	Ticari Kredi Puanı
TKR	Ticari Kredi Raporu
T2SAM	Time-Tree Structured Self-Attention Model
WCAG	Web Content Accessibility Guidelines
WWW	World Wide Web

1. GİRİŞ

Dünyadaki veri hacmi ve veri çeşitliliği, insanlık tarihinde daha önce hiç görülmediği hızda artmaktadır. İnternet teknolojilerinin ve sosyal medyanın hayatımızın her evresine ve hatta cep telefonlarımıza girmesiyle, insanlar günlük faaliyetlerinde bile veri üretir duruma gelmiştir. Dünün manuel olarak çalışan araç gereçleri, bugün akıllı cihazlar olarak anılmakta ve hemen hepsi sensörleri vasıtasıyla veri üretmektedir (Aktan, 2018).

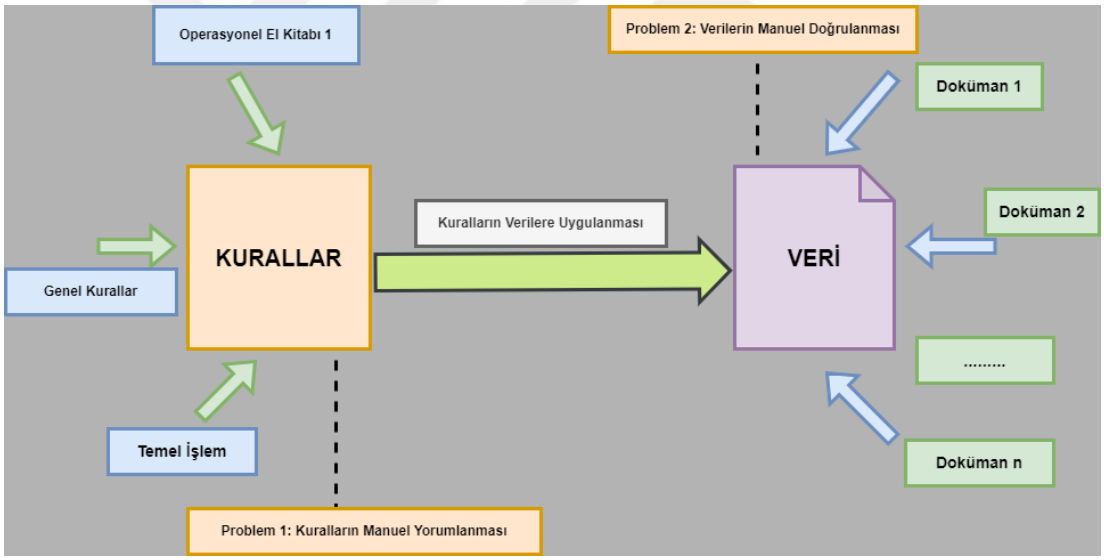
Verinin bu kadar çok olduğu bir ortamda istenilen bilgiye ulaşmak ciddi bir problemdir. Günümüz bilgi çağı, aranan bilgiye daha çabuk ve hızlı erişmek için otomatik doküman doğrulama sistemlerinin bilgi çıkarımı ile ilgili birçok alanda kullanımını zorunlu hale getirmektedir. Şekil 1.1, 2020-2024 yılları arasındaki tahmini doküman doğrulama pazar payını göstermektedir. Görüldüğü üzere, bu alandaki pazar payının sürekli olarak artması beklenmektedir.



Şekil 1.1 2020-2024 yılları arasındaki doküman doğrulama pazar payı (Novarica, 2023)

Doküman doğrulama, belirli bir konuya ait orijinal doküman/dokümanlar üzerinde, özet dokümanın doğrulanması işlemidir. Kısacası özet içeriğinin orijinal dokümanların içeriği ile uyumluluğunun geçerli kılınması işlemidir.

Doküman doğrulama süreci elle (Morocho vd., 2017; Ghazali vd., 2019), yarı otomatik (Stone vd., 2017; Liu vd., 2021; Zhi vd., 2021) ve otomatik (Hsu, 2020; Attivissimo vd., 2019; Gupta vd., 2018) olmak üzere üç farklı şekilde gerçekleştirilmektedir. Elle ve yarı otomatik doğrulamalar insan müdahalesi gerektirir ve doğrulama sonucunda elde edilen doğruluk, doğrulamayı yapan kişinin performansına doğrudan bağlıdır. Otomatik doküman doğrulama süreçlerinde ise durum böyle değildir. Elle ve yarı otomatik doküman doğrulama işlemi Şekil 1.2'de verilmektedir. Görüldüğü üzere doğrulama işlemleri kural bazlı olarak yürütülmektedir.



Şekil 1.2 Elle ve yarı otomatik doküman doğrulama süreci (Roychoudhury vd., 2016)

Otomatik doküman doğrulama ise, orijinal tam metin dokümanlar veya belirli bir konuya ait dokümanlardan üretilen özetlerin içerik ve anlam açısından otomatik olarak doğrulanmasını sağlayan bir süreçtir. Bu süreç, algoritmalar ve belirli kurallar kullanılarak gerçekleştirilmektedir. Amacı, özetin orijinal doküman(lar) ile uyumlu olup olmadığını hızla ve yüksek doğrulama başarım oranı ile belirlemektir. Bu, özellikle büyük veri kümelerinin özetlerini kontrol etmek veya

belirli bir konu hakkında çok sayıda dokümanı hızla özetlemek için oldukça kullanışlıdır. Otomatik doküman doğrulama, manuel incelemelere kıyasla çok daha hızlı ve tutarlı sonuçlar sağlamaktadır.

1.1 Problem Tanımı

Günümüzde hem bireysel hem de kurumsal düzeyde büyük miktarda dokümanın üretimi ve kullanımı söz konusudur. Bankacılık işlemlerinden yasal dokümanlara, eğitim materyallerinden ticari faturalara kadar geniş bir yelpazede dokümanlarla karşı karşıyayız. Bu dokümanların doğruluğunun ve bütünlüğünün teyit edilmesi, özellikle hukuki ve mali sorumluluklar söz konusu olduğunda hayati öneme sahiptir.

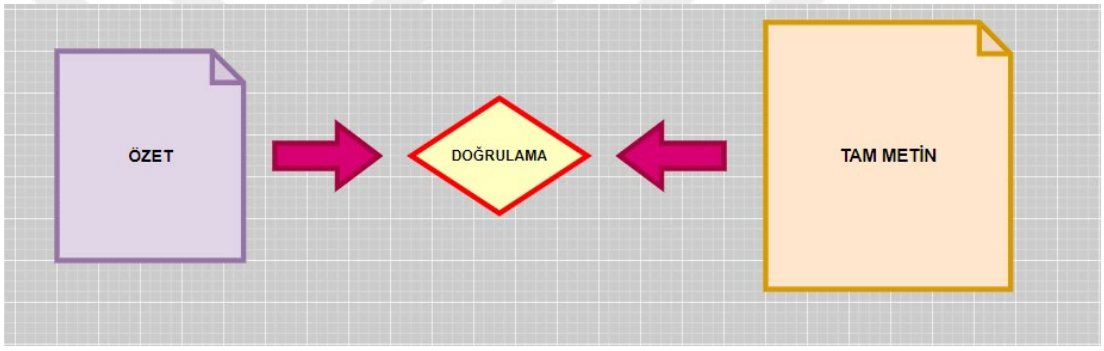
İnsan gücü ile doküman doğrulama yöntemleri, bünyesinde bir dizi zorluk barındırmaktadır. Öncelikle, bu süreçler genellikle zaman alıcı ve işgücü yoğundur. Belirli bir dokümanın doğru olup olmadığını belirlemek için gereken süre, bu dokümanın karmaşıklığına ve içeriğine bağlı olarak önemli ölçüde değişiklik gösterebilmektedir. Ayrıca, insan kaynaklı hataların riski her zaman mevcuttur. Bir kişi, dokümanın içeriğini gözden geçirirken gözden kaçırdığı veya yanlış yorumladığı detaylar olması muhtemeldir. Bu durum, yanıltıcı veya hatalı bilgilere dayalı kararların alınmasına neden olabilmektedir.

Ayrıca, insan gücü ile doküman doğrulama yöntemleri büyük veri setleriyle kolayca ölçeklenemezler. Modern iş dünyası ve akademik araştırmalar, her geçen gün artan miktarda veri üretmekte ve kullanmaktadır. Bu yöntemler, hızlı veri akışını yönetmek ve doğrulamak için gereken hızda ve verimlilikte icra edilemezler.

İnsan gücü ile doküman doğrulama, genellikle yüksek operasyonel maliyetlerle ilişkilidir. Nitelikli profesyonellerin saatlerini bu tür incelemelere ayırmaları, işgücü maliyetlerinin önemli ölçüde artmasına neden olmaktadır. Ayrıca, hatalı doğrulamadan kaynaklanabilecek olası hukuki sorunlar ve itibar kaybı, maliyetleri daha da artırabilmektedir.

İnsan gücü ile doküman doğrulama yöntemleri ayrıca gelişmiş teknolojilerin sunduğu avantajlardan yararlanamazlar. Otomatik doküman doğrulama sistemleri, yapay zeka ve DDİ gibi teknolojileri kullanarak, anlam analizi, içerik karşılaştırması ve hatta sahteciliğin tespiti gibi daha sofistike görevleri yerine getirebilmektedir.

Bu zorluklar, birçok kuruluşun ve araştırma kurumunun, doküman doğrulama süreçlerini otomatikleştirmeye yönelmesinin temel nedenlerindendir. Otomatik doküman doğrulama, hız, tutarlılık ve maliyet etkinliği açısından önemli avantajlar sunmaktadır. Bu çalışma Şekil 1.3'te verilen probleme çözüm olarak gerçekleştirilmiştir.



Şekil 1.3 Problem tanımı

Bu teze konu olan otomatik doküman doğrulama çalışmasında, orijinal dokümana ait üretilen özet dokümanın tutarlılığını okumak, yorumlamak ve kontrol etmek için Doğal Dil İşleme (DDİ) tekniklerine dayanan otomatik doküman doğrulama sistemi tasarlanmıştır. Amacımız, özelleşmiş problemler üzerinde yüksek başarım oranları elde edecek özellikle finansal dokümanların doğrulanması için genelleştirilebilir otomatik doküman doğrulama altyapısı oluşturmaktır.

1.2 Çalışma Konusu ve Amacı

İnsan gücü ile doküman doğrulama, büyük bir doküman havuzunda işlem yapıldığında veya kuralların yorumlanmasına maruz kalındığında önemli ölçüde

zaman ve çaba gerektirir. Doküman doğrulama, dokümanın içerik olarak doğrulanması veya dokümanın sahipliğini kanıtlamaya yönelik bir süreçtir. Mali tabloların, pasaportların, lisansların, elektrik faturalarının ve sertifikaların gerçek olduğunu ve bunları sunan başvuru sahibine ait olduğunu doğrulamak sahiplik kanıtlanmasına dair örneklerdir.

Bu teze konu olan doküman doğrulama, belirli bir konuya ait dokümanlar için üretilen özetin orijinal dokümanların içeriği ile doğrulanması işlemidir. Kısaca orijinal doküman veya dokümanlardaki özetin içerik ve anlam açısından doğrulanması işlemidir. Özet, orijinal dokümandaki tam olarak aynı cümlelerle (çıkarımsal özet) veya tamamen farklı cümlelerle (soyut özet, insan özetleri genellikle bu yöntemde ele alınır) yazılabilir. Çıkarımsal özetlerin orijinal dokümanlardaki cümlelerini eşleştirmek yeterliyken, soyut özetlerin doğrulanması dilbilimde daha karmaşık ve anlamsal analiz gerektirir. Hangi özetleme tekniğinin kullanıldığı dokümanların analizi açısından doğrulamada önemli bir noktadır. Soyutlayıcı bir özetleme algoritmasıyla elde edilen bir özet dokümanın doğal olarak doğrulanması daha zor bir iştir.

Otomatik doküman doğrulama, finans gibi çok sayfalı raporlardan (şirket bilanço özet raporu, yatırımcı fon raporları vb.) özetlerin üretildiği sektörlerde periyodik olarak (genellikle aylık) yapılan görevlerden biridir. Özet üretmek için personel tahsis edilmektedir. Amaç, detaylı iç raporlardan dış paydaşlara (müşteri veya kurumlara) birkaç sayfayı geçmeyecek özetler oluşturmaktır. Genel olarak zahmetli, titiz ve hataya açık bir iştir. Daha sonra insanlar tarafından oluşturulan bu dokümanların tutarlılığının yorumlanması ve doğrulanması gerekmektedir. Ancak bu süreç için kaynak kısıtlılığı nedeniyle genellikle personel tahsisi yapılmamaktadır. Kaynak ayrılsa bile tüm sorumluluk özeti hazırlayan personele aittir. Doğrulama için farklı personelin tahsis edilse bile, yine iş tamamen insan tarafından yapıldığı için tekrar hataya açık ve özet çıkaranın işine göre daha sıkıcı ve uzun bir süreç olacaktır. Bu nedenlerden dolayı doğrulama işleminin bilgisayar yazılımları ile otomatik ve yüksek doğrulukla yapılması önemlidir.

Otomatik doküman doğrulamanın önemli fayda sağlayacağı bir diğer uygulama alanı da finans sektöründe fon yönetimidir. Fonlar Avrupa'da ve dünyanın birçok yerinde iki dokümanda anlatılmaktadır. Biri özet olarak adlandırılan en önemli bilgi setini sunar, diğeri ise her bir bilgiyi detaylandıran ve fon bileşenlerini doğru bir şekilde açıklayan ana dokümandır (tam metin). Bu fonlarla işlem yapan kişi ve kurumlar için bu özet dokümanların ana dokümanları iyi yansıtması özellikle önemlidir. Örneğin ana dokümanda 3.000.000$, özet dokümanda ise 300.000$ olarak yazılan bir beyan ciddi bir sorundur. Bu ve benzeri sorunların tespitinde otomatik doküman doğrulama altyapısı kullanılarak ciddi mali kayıpların önüne geçilebilmektedir.

Bu çalışmada, orijinal dokümanlar (özellikle finansal türde) için üretilen soyut özetin tutarlılığının anlamsal olarak kontrol edilmesi amacıyla DDİ tekniklerini temel alan bir otomatik doküman doğrulama sistemi tasarlanmıştır. Orijinal tam metin dokümanlardaki soyut özetlerin doğrulanması Transformer tabanlı model aracılığıyla yapılmıştır. Çalışmada doğrulanacak referans dokümanlar finansal türe ait olduğundan Reuters finansal veri seti ile eğitim yapılarak Transformer modeli oluşturulmuştur. Önerilen Transformer tabanlı anlamsal doküman doğrulama yaklaşımı, Reuters finansal veri kümesindeki orijinal tam metin ve özet dokümanlar üzerinde test edilmiştir. Veri seti (finans) Kaggle sisteminden (Archive 2023) elde edilmiştir. Bu veri setleri ilk olarak veri ön işleme tabi tutulmuştur. Veri ön işleme adımında veri setindeki hem orijinal tam metin hem de özet dokümanlar uygun forma dönüştürülmüştür. Bu amaçla, dokümanlarda (doğrulanacak orijinal ve özet dokümanlar) öncelikle kelimelerin dil kurallarına uygun yazılıp yazılmadığı kontrol edilmiştir. Özet oluşturma sırasında yazım hataları (harflerin değişmesi, kelimelerin farklı anlamlarda kullanılması vb.) veya anlam hataları (olumlu bir ifadenin olumsuz yazılması, sayısal büyüklüğün ters ifade edilmesi vb.) meydana gelebilir. Bu hatalar otomatik anlamsal doküman doğrulama sisteminin başarısı için önemlidir. Bu hataların varlığı kontrol edilmiş olup hatalı ifadeler varsa WordNet Yazım Denetleyicisi algoritması ile düzeltilmiştir. Daha sonra özet doküman tam metin dokümanlar üzerinde doğrulanacağı için Simhash ve Cross Encoder metin benzerliği algoritmaları kullanılarak tam metin dokümanların cümlelerinden özet doküman cümlelerine

en çok benzeyen cümleler belirlenmiştir. Özet doküman cümlesine en çok benzeyen orijinal tam metin doküman cümleleri (deneysel olarak) seçilmiştir. Daha sonra bu orijinal tam metin doküman cümleleri Transformer modeline eğitim verisi olarak verilmiştir. Son olarak Transformer modeli orijinal tam metin cümlelerin soyut özetini üretmiştir. Son aşamada orijinal özet ile Transformer modelinin ürettiği özet, benzerlikleri açısından hem Simhash hem de Cross Encoder metin benzerliği algoritmalarıyla karşılaştırılmış ve ortalama doküman doğrulama doğruluğu hesaplanmıştır. Doğrulama başarım oranının yüksek olması, özet dokümanın tam metin doküman temsil gücünün yüksek ve önemli bilgiler içerdiği anlamına gelirken, düşük olması, özet dokümanın tam metin doküman temsil gücünün düşük olduğu ve tam metin doküman temsil gücünün bazı önemli bilgileri içermediği anlamına gelmektedir. Önerilen Transformer tabanlı anlamsal doküman doğrulama yaklaşımı, Reuters veri setinde ortalama %84,1 anlamsal doküman doğrulama başarımı elde etmiştir. Amacımız, kurumların elle, zaman alıcı ve kontrolsüz bir şekilde yaptıkları doğrulama maliyetlerini ortadan kaldırmak, kaynaklarını daha katma değerli alanlarda kullanmalarını sağlamak ve insan kaynaklı hataları önlemek amacıyla genelleştirilebilir Transformer tabanlı bir otomatik anlamsal doküman doğrulama yaklaşımı oluşturmaktır.

Bu çalışma, soyut doküman özetlerinin anlamsal doğrulanması için tasarlanmış Transformer tabanlı modellerin yenilikçi uygulamasıyla öne çıkmaktadır. Yalnızca sözdizimsel özelliklere dayanan mevcut birçok yöntemin aksine, yaklaşımımız, doküman doğrulamanın başarımını ve etkinliğini artırmak için DDİ teknikleri kullanılarak elde edilen anlamsal bilgilerden yararlanmıştır. Transformer modelini özel bir Reuters veri seti üzerinde titizlikle eğiten yöntemimiz, yalnızca finansal terimlerin karmaşıklığını yakalamakla kalmamış, aynı zamanda yüksek doğrulama başarımı da sağlamıştır.

Bu yaklaşımın teknik başarısının ötesinde, veri ön işlemeden Simhash ve Cross Encoder algoritmalarını kullanan cümle benzerliği kontrollerine kadar uzanan doğrulama yaklaşımı, sürece derinlik sağlamıştır. Tam metin dokümandaki özet dokümana en çok benzeyen cümleleri seçmek sistemimizin performansını daha

da arttırmıştır. Elde edilen sonuçlar finans sektörüne somut faydalar vaat etmektedir. %84,1 olarak kaydedilen doğrulama doğruluğu, insan kaynağında önemli azalmalar sağlayarak fon yönetimi ve diğer finansal alanlarda uygun maliyetli ve verimli operasyonların önünü açmaktadır.

Ayrıca, bu çalışmanın önemli bir katkısı Transformer tabanlı soyutlayıcı özetleme yaklaşımının deneysel uygulamasıdır. Transformer ve cümle gruplamayı kullanarak otomatik özetleme için hibrit bir yöntem uygulanmıştır. Süreç, orijinal tam metnin cümlelere ayrıştırılması, aralarındaki benzerliklerin hesaplanması ve ardından bu cümlelerin hash mesafelerine göre sıralanması şeklindedir. En benzer orijinal tam metnin cümleler gruplandırılmış, bu gruplandırılmış cümleler daha sonra Transformer modeline beslenerek her grup için soyutlayıcı bir cümle üretilmiş ve tutarlı ve özlü soyutlayıcı özetler elde edilerek süreç tamamen otomatik hale getirilmiştir.

2. LİTERATÜR ÖZETİ

Doküman doğrulama çalışmaları özellikle verinin çok yoğun olduğu birçok sektör için önemli bir konu haline gelmiştir. Belgelerinin tutarlılığının yorumlanması ve doğrulanması zaman alıcı bir görevdir ve maliyetlidir. Bu sorunla baş etmek ve insan işgücünü desteklemek üzere akıllı araçlar gereklidir (Hassanpour vd., 2011).

Doküman özetlerini doğrulama üzerine çeşitli çalışmalar (Charlow vd., 2008; Alpuente vd., 2008; Thongkor vd., 2012; Li vd., 2016) yapılmıştır. Son zamanlarda araştırmacılar özellikle yazılım mühendisliği problemleri üzerinde yoğunlaşmışlardır. Gereksinimlerin otomatik olarak çıkarılması (Mu vd., 2009), dokümanların kalitesi (Thitisathienkul vd., 2015) ve doğrulama (Ando vd., 2015) çalışmalarında, doküman doğrulama için DDİ uygulanmıştır.

Doğrulama çalışmalarındaki kritik konulardan biri anlamsal (semantik) doğrulamadır. Doğrulanacak özet doküman soyut özetleme teknikleri ile elde edilmiş ise, doğrulamanın da anlamsal olarak yapılması gerekmektedir. Dokümanların anlamsal olarak karşılaştırılması için farklı ölçüm yaklaşımları bulunmaktadır. Literatürde bu ölçümü gerçekleştirmek amacıyla metinlerin yapısal ya da anlamsal özelliklerini inceleyen çeşitli çalışmalar yapılmıştır. Bu çalışmalardan ilki Philip Resnik'in 1995 yılı (Resnik, 1995) çalışmadır. Resnik bu çalışmada, bilgi içeriği kavramına dayanan bir taksonomide anlamsal benzerliğin belirlenmesine yönelik bir yöntem önermiştir. İki kelime/kavram arasındaki benzerlik, aralarındaki ortak bilgiler dikkate alınarak değerlendirilmiş ve bu görev için WordNet isim sınıfı taksonomisinden elli bin düğümlü bir küme dikkate alınmıştır. Taksonomideki kavramların sıklıkları, haber makalelerinden bilim kurguya kadar çeşitli türlerde geniş (1.000.000 kelime) bir metin koleksiyonu olan Brown derleminden (Brown Üniversitesi Standart Günümüz Amerikan İngilizcesi Derlemesi) alınan isim sıklıkları kullanılarak tahmin edilmiştir. Sonuçlar, ölçümün, insan benzerliği yargılarından oluşan bir kıyaslama seti ile %79'luk bir korelasyonla cesaret verici derecede iyi bir performans sergilediğini göstermiştir.

1997 yılında Jiang ve arkadaşlarının (Jiang vd., 1997) çalışmasında kelimeler ve kavramlar arasındaki anlamsal benzerliği/uzaklığı ölçmek için yeni bir yaklaşım ortaya konmuştur. Sözcüksel bir taksonomi yapısını derlem istatistiksel bilgisiyle birleştirmiş, böylece taksonomi tarafından oluşturulan semantik uzaydaki düğümler arasındaki semantik mesafe, derlem verilerinin dağılım analizinden elde edilen hesaplamalı kanıtlarla daha iyi ölçülebilmiştir. Önerilen ölçüm yöntemi, kenar sayma şemasının kenar tabanlı yaklaşımını devralan ve daha sonra bilgi içeriği hesaplamasının düğüm tabanlı yaklaşımıyla geliştirilen birleşik bir yaklaşımdır. Önerilen yaklaşım, ortak bir kelime çifti benzerlik derecelendirme veri seti üzerinde test edildiğinde diğer hesaplama modellerinden daha iyi performans göstermiştir. İnsan benzerliği yargılarına dayalı bir kıyaslamayla %82'lik en yüksek korelasyon değerini vermiştir.

World Wide Web (WWW) şu an tüm dünyada kullanılan temel iletişim altyapısıdır. Bu nedenle web sitesi tasarlayıcılarının WWW standartlarına uygun, doğrulanabilir, evrensel olarak erişilebilir siteler oluşturmaları gerekmektedir. WWW tabanlı sayfaların doğrulanmasına yönelik önemli çalışmalardan biri Takata ve ekibinin (Takata vd., 2004) çalışmasıdır. Web Content Accessibility Guidelines (WCAG) gibi erişilebilirlik yönergeleri, içerik oluşturuculara Web sayfalarını oluşturma noktasında erişilebilirlik ve doğrulama süreçlerinde referans kaynak olmaktadır. Belirli bir web sayfasının yönergelere uygun olup olmadığını otomatik olarak doğrulayan araçlar sağlanmaktadır. Bu çalışmada, WCAG doğrulamasının, HTML (Hypertext Markup Language) dışındaki herhangi bir belge biçimine veya WCAG dışındaki herhangi bir yönergeye uygulanmasının zor olduğundan ötürü yeni bir yaklaşım önerilmiştir. Önerilen yaklaşımda, yönerge ve doğrulama kriterlerini belirlemek için basit akış süreci oluşturulmuştur. Belge türü HTML'den soyutlanmış ve her bir belge türüne (Extensible Markup Language (XML), HTML, JavaScript Object Notation (JSON)) göre, farklı bir doğrulama süreçleri işletilmiştir. Önerilen doğrulama aracı kullanılarak ABD ve Japonya'daki kırk büyük kuruluşun yaklaşık 3.000 Web sayfasında doğrulama gerçekleştirilmiş ve yaklaşık %60 doğrulama oranı elde edilmiştir.

Ameur Bensefia ve ekibi (Bensefia vd., 2005) çalışmalarında, yazar doğrulama alanında umut verici sonuçlar elde etmişlerdir. Bu çalışmada, bitişik el yazısının segmentasyonundan çıkarılan grafikler gibi yerel özellikler kullanılarak hem yazar tanımlama hem de yazar doğrulama görevleri gerçekleştirilmiştir. Yazar tanımlama sürecinde metin tabanlı bilgi erişim modeli kullanılmıştır. Yazar-eser eşleştirilmesinde özellik frekanslarına dayalı olarak ilişki kurularak doğrulama yapılmıştır. Değerlendirme için PSIDataBase veri seti ve IAMDataBase veri seti olmak üzere iki veri seti kullanılmıştır. PSIDataBase veri kümesi 55 yazardan 550 sayfa içerirken, IAMDataBase veri kümesi 657 yazardan 1.539 sayfa içermektedir. Önerilen yaklaşım, PSIDataBase veri kümesinde %95, IAMDataBase veri kümesinde ise %86 doğruluk elde etmiştir. Bu yaklaşımın genel ve el yazısı belge sorgulama ve yazar doğrulama alanındaki büyük ölçekli uygulamalar için oldukça faydalı olacağı belirtilmiştir.

2006 yılı Rada Mihalcea ve ekibinin (Mihalcea vd., 2006) çalışmasında, derlem tabanlı ve bilgiye dayalı benzerlik ölçütlerini kullanarak metinlerin anlamsal benzerliğini ölçmek amacıyla bir yöntem sunulmuştur. Bu konuyla ilgili geçmişte yapılan çalışmaların, ya büyük boyutlu belgelere (Metin sınıflandırması, bilgi erişimi) ya da tek tek sözcüklere (Eş anlamlılık testleri) odaklanıldığı belirtilmiştir. Ancak, günümüzde, Web'de ve başka yerlerde bulunan bilgilerin büyük bir bölümünün kısa metin parçalarından (Bilimsel belgelerin özetleri, hayali alt yazılar, ürün açıklamaları) oluştuğu göz önüne alındığında, bu çalışmada kısa metinlerin anlamsal benzerliğinin ölçümlenmesine odaklanılmıştır. 100 kayıtlık elle hazırlanmış bir açıklama veri setiyle (Shinyama vd., 2002) gerçekleştirilen deneylerde, anlamsal benzerlik yönteminin basit sözcük eşleştirmeye dayalı benzerlik yöntemlerinden daha iyi performans gösterdiğini ve bunun sonucunda geleneksel vektör tabanlı benzerlik metriğine göre hata oranında %13'e varan azalma sağladığı gösterilmiştir.

Doğrulama işlemi konuşma verileri üzerinde de yapılmaktadır. Bu çalışmalardan biri de Chung-Hsien Wu ve ekibinin (Chung vd., 2007) çalışmasıdır. Bu çalışmada, konuşulan anahtar kelime çıkarmayı ve konuşma anlamaya yönelik anlamsal

doğrulamayı kullanan yeni bir yaklaşım sunmaktadır. Önemli bilgi/anahtar kelime çıkarımı için akustik, aruz ve dilsel bilgilere dayalı bir bilgi çıkarma yaklaşımı önerilmiştir. Alınan belgelerin kesinliğini artırmak için ileri-geri yayılımlı anlamsal doğrulama uygulanmıştır. İleri prosedürde, kelime anlamlılık ölçüsüne göre bir torba anahtar kelimeler seçilmiştir. Geriye dönük prosedürde, anlamsal doğrulama için anahtar kelimeler arasındaki anlamsal ilişki tahmin edilmiştir. Doğrulama puanı daha sonra, alma performansını iyileştirmek için alınan belgeleri ağırlıklandırmak ve yeniden sıralamak için kullanılmıştır. Deneyler, 198 saatlik yayın haber veri tabanında gerçekleştirilmiştir. Deneysel sonuçların, önerilen yöntemlerin diğer geleneksel indeksleme yöntemlerine kıyasla daha iyi erişim sonuçları elde ettiği belirtilmiştir. 3603 adet MATBN Çince veri kümesinde %80 oranında doğrulama başarımı elde edilmiştir.

Doküman doğrulama sahtecilik gibi yasal olmayan süreçlerde de oldukça önemli bir yere sahiptir. Özellikle finansal belgelerde yapılan bu işlemler maddi büyük kayıplara neden olmaktadır. Utpal Garain ve ekibi (Garain vd., 2008) çalışmasında, banka çekleri, piyango biletleri, uçak biletleri gibi çeşitli belgelerde yapılan sahteciliklerı tespit etmek ve otomatik orijinallik doğrulaması yapmak amacıyla genel bir çerçeve geliştirmeye çalışmışlardır. Önerilen yöntem, önce belge görüntülerinden güvenlik özelliklerini hesaplamalı olarak çıkarır ve daha sonra özellik alanında orijinallik ve özgünlük kavramı tanımlanır. Deneylerin yapılmasında banka çekleri referans alınır. Bu çeklerin gerçekliğini doğrulamak için SVM sınıflandırıcıları ve doğrusal olmayan çekirdek işlevleri kullanılır. Sonuçlar, polinom çekirdek tabanlı bir SVM sınıflandırıcısının, mükerrer kontrolleri orijinal olanlardan ayırt etmede yaklaşık %99,5 doğruluk sağladığını göstermektedir. Bu, basılı güvenlik belgelerinin makine doğrulaması için önerilen yaklaşımın uygulanabilirliğini güçlü bir şekilde kanıtlamaktadır.

Doküman sahteciliği tespiti amacıyla yapılan çalışmalardan bir diğeri de Joost van Beusekom ve ekibinin (Beusekom vd., 2011) çalışmasıdır. Bu çalışmada, taranan veya yeniden tasarlanmış belgelerde yapılan kanunsuz düzenlemeleri tespit edecek otomatik sahte belge doğrulama için bir yaklaşım sunulmuştur. Bozulmaların tespiti, genellikle faturalarda görünen üstbilgiler ve altbilgiler gibi

sabit belge kısımlarında yapılmıştır. Sorgulanan faturalar aynı kaynaktan gelen diğer faturalar ile eşleştirerek toplam eşleştirme kalitesi hesaplanmış ve sonraki aykırı değerler tespiti için özellik olarak kullanılmıştır. Bir aykırı değer tespit edilirse, o belge şüpheli olarak kabul edilmiştir. Oluşturulan modeli değerlendirmek amacıyla iki farklı türde fatura veri seti oluşturulmuştur. Bu veri setleri üzerinde yapılan deneylerde %82'lik bir belge doğrulama oranı ile sahte belge tespiti yapılmıştır. Oluşturulan bu modelin, insan müdahalesi gerektirmeden denetimsiz bir şekilde sahte belge tespiti ve doğrulamasında kullanılabileceği belirtilmiştir.

2011 yılında Jenq-Haur Wang ve ekibi (Wang, 2011) çalışmalarında, istatistiksel terim çıkarma ve Web tabanlı terim doğrulama yöntemlerini kullanarak metin akışlarından anahtar kavramları çıkarmanın ve doğrulamanın otomatik bir yolunu önermişlerdir. Çalışmadaki ana hedef amaçlanan web aramasını en efektif şekilde yapmak ve web arama sonucunda gelen verileri doğrulamaktır. Bu amaçla, öncelikle başlangıç veri seti için anahtar kelime tespiti yapılmış ve bu işlem için Term Frequency-Inverse Document Frequency (TF-IDF) algoritması kullanılmıştır. Belirlenen parametre değeri kadar terim kullanılarak web araması yapılmıştır. Web araması sonucunda gelen verilerin terim doğrulaması çeşitli Jaccard benzerlik ölçütleriyle karşılaştırılmıştır. En yüksek Jaccard benzerlik oranına sahip web sayfasının başlangıç veri setini daha iyi temsil edeceği belirtilmiştir. Çince ve İngilizce haber alıntıları için yapılan deneysel sonuçlarda, İngilizce veri kümesinde %81, Çince veri kümesinde ise %63 oranında doğrulama başarımı elde edilmiştir. Önerilen yaklaşımın verimlilik ve etkililik potansiyeli gösterdiği belirtilmiştir.

Abdulwahed Almarimi ve ekibinin (Almarimi vd., 2015) çalışmasında, n-gram ve kelimelerin histogramlarını kullanarak belge doğrulama yapılmıştır. Çalışmada, n-gramlara ve İngilizce ve Arapça sözcükleri için oluşturulan yerel histogramlara dayalı belge doğrulama için kullanılabilir yöntemler analiz edilmiş ve karşılaştırmalı sonuçları listelenmiştir. İngilizce ve Arapça metinler birçok istatistiksel özellik açısından analiz edilmiştir. Her iki dil arasında bazı istatistiksel farklılıklar keşfedilmiş ve metin parçalarının farklılıklarını tespit

etmek için uygulamalı n-gram analizi ve yerel histogramlar bulunmuştur. Daha sonra bu histogramlar kullanılarak belgeler doğrulanmaya çalışılmıştır. İngilizce veri kümesinde %70, Arapça veri kümesinde ise %80 oranında doğrulama başarımı elde edilmiştir. Elde edilen sonuçların umut verici olduğu belirtilmiştir.

Doğrulama işlemi sadece doküman ya da metin dosyaları ile ilgilenen kişi ve kurumlar için değil, yazılım geliştiricilerin hayatında da oldukça önemli bir yere sahiptir. Yazılan kodların insan gereksinimi olmadan otomatik olarak doğrulanması ve standartlara uygun olarak yazılmayan kodların üretim ortamına alınmaması gerekmektedir. Günümüzde, cep telefonları, otomobiller, uzay ekipmanları gibi çeşitli alanlarda kullanılan yazılımlar, hayatımızın birçok farklı yönüne entegre edilmiştir. Yüksek güvenilirlik gerektiren gömülü yazılımların geliştirilmesinde, geliştirmenin her aşamasında titiz doğrulamalar gereklidir. Bu alanda yapılan çalışmalardan biri de Takahiro Ando ve ekibi (Ando vd., 2015) tarafından gerçekleştirilmiştir. Çalışmanın amacı, yazılan kodların doğrulamasını geliştiriciden farklı bir bakış açısıyla incelemek ve geliştiricinin gözden kaçırdığı hususları daha sıkı bir kontrole tabi tutmaktır. Bu çalışmada, her bir farklı türden kaynak kod deposu için şablon denetleme mekanizması kurulmuştur. Bu şablon mekanizmasına "Referans Modeller" adını vermişlerdir. Referans Modellerde, gereksinim ve tasarım belirtimi doğrulamasının verimli bir şekilde uygulanmasını desteklemek için kontrol noktaları bulunmaktadır. Yazılım geliştiricin yazdığı kod üretime alınmadan önce bu sistemden geçirilmekte ve ilişkili referans model ile karşılaştırılmaktadır. Doğrulama oranı belli seviyenin altında olan kodlar tekrar düzenlenmek üzere yazılım geliştiriciye tekrar iletilmektedir. Kaynak kodların olduğu elle hazırlanmış bir veri setinde %81,6 doğrulama başarımı elde edilmiştir. Elde edilen sonuçların ilgili veri setini doğrulama açısından iyi seviyede olduğu belirtilmiştir.

2016 yılında Zhao ve arkadaşlarının (Zhao vd., 2016) çalışmasında, anlamsal rollere dayalı yeni bir anlamsal benzerlik ölçüsü önerilmiştir. Yazarlar, geleneksel anlamsal benzerlik ölçümlerinin genellikle sözcüksel bilgilere dayandığını ve sözcükler arasındaki anlamsal ilişkilerin tamamını yakalayamayabileceğini belirtmiştir. Önerilen yaklaşım, cümlelerden anlamsal rolleri çıkarmak için bir

bağımlılık ayrıştırıcısı kullanmakta ve bu rollerin örtüşmesine dayalı olarak benzerliği hesaplamaktadır. SICK veri seti ve SemEval 2015 Task 2 veri seti üzerinde deneyler gerçekleştirilmiştir. SICK veri setinde 0,805'lik bir Pearson korelasyon katsayısına ulaşırken, SemEval 2015 Task 2 veri setinde ise %87 doğruluk elde etmiştir.

Suman Roychoudhury ve ekibinin (Roychoudhury vd., 2016) çalışmasında, özellikle yabancı hisse senedi, e- akreditif mektubu, konşimento, ticari fatura gibi veri miktarının yoğun olduğu sistemlerin insan tarafından doğrulamasından kaynaklanan zorlukları gidermek amacıyla kural tabanlı otomatik web sayfası doğrulamasını sağlayan bir yöntem önermişlerdir. Çalışmada referans alınan doküman seti 302.000 düğüm ve 10 MB büyüklüğünde, Web'ten elde edilen XML ve HTML uzantılı dosyalardır. Özet dokümanların doğrulaması cümle benzerliği ile yapılmıştır. Cümle benzerliğinin hesaplanması Verdi Performance Analyzer Tool modelinin yeni bir versiyonu olarak tasarlanan ve alphaVerdi olarak isimlendirilen model ile yapılmıştır. XML ve HTML veri setinde %87,3 doğrulama başarımı elde edilmiştir. Elde edilen sonuçların ilgili veri setini doğrulama açısından iyi seviyede olduğu belirtilmiştir.

2017 yılı Haoyu Pu ve ekibi (Pu vd., 2017) çalışmasında, kısa metinler üzerinde anlamsal bir benzerlik yöntemi önermişlerdir. Kısa metinlerin sosyal medyada yaygın olarak kullanıldığı belirtilmiştir. Kısa metnin benzerlik hesaplaması, kelimelerinin kısa olması özelliğinden etkilenir ve doğruluk oranı oldukça düşüktür. Bu nedenle kelime anlam benzerliği ile kısa metinlerin benzerliğini hesaplamak yaygın bir iyileştirme yöntemi olduğu vurgulanmıştır. Bu çalışma, bilgi tabanlı (Knowledge-based) yöntem ile derlem tabanlı (Corpus-based) yöntemi birleştiren kısa metinlerde anlamsal benzerlik hesaplama yöntemi ortaya koymaktadır. Bu yöntem, geliştirilmiş sözcük anlamsal benzerlik hesaplama ve genel kısa metin anlamsal benzerlik hesaplama yöntemlerine dayanmaktadır. Kelime benzerliği hesaplama yöntemi, iki kelimenin anlamsal benzerliğini bazı stratejilerle birleştirir. Tek yöntemin dezavantajlarının üstesinden gelmek için iki yöntemin avantajlarını kullanır, metinlerdeki kelimeler arasında daha fazla anlamsal ilişki bulur ve kelime benzerliği

hesaplamasının doğruluğunu arttırmaktadır. Geliştirilen yöntemin hem kelime hem de metin benzerliği açısından diğer yöntemlere göre insan derecelendirmelerine daha yakın bir sonuca sahip olduğu belirtilmiştir.

2017 yılında Rehman ve ekibi (Rehman vd., 2017) çalışmasında, grafik tabanlı bir yaklaşım önermiştir. Yazarlar belgeyi, düğümlerin kelimeleri ve kenarların da aralarındaki ilişkileri temsil ettiği bir grafik olarak temsil etmişlerdir. Daha sonra, belge ile iddia edilen yazar arasındaki benzerliği karşılaştırmak için grafik düzenleme mesafesi gibi grafik benzerlik ölçümleri kullanılmıştır. Yazarlar değerlendirme için 5 yazardan 50 belge içeren C50 veri setini kullanmıştır. Önerilen yaklaşım C50 veri kümesinde %96 doğruluk elde etmiştir.

2018 yılında Zeng ve arkadaşlarının (Zeng vd., 2018) çalışmasında, derin öğrenme temelli bir anlamsal doküman doğrulama tekniği tanıtılmıştır. Bu teknik, belgeyi anlamsal bir temsil modeli ile ifade etmektedir. Daha sonra, model belge ile iddia edilen yazar arasındaki benzerliği karşılaştırmak için Siamese sinir ağları kullanılmıştır. Yazarlar değerlendirme için iki veri kümesi kullanmıştır: Enron e-posta veri kümesi ve PAN 2013 veri kümesi. Enron e-posta veri kümesi 500.000'den fazla e-posta içerirken, PAN 2013 veri kümesi 40 yazardan gelen 40.000 belge içermektedir. Önerilen yaklaşım Enron e-posta veri kümesinde %99,2 ve PAN 2013 veri kümesinde %90,1 doğruluk oranı elde etmiştir.

2019 yılında Alzahrani ve ekibinin (Alzahrani vd., 2019) çalışmasında anlamsal doküman doğrulama için metinsel ve görsel özellikleri birleştiren hibrit bir yaklaşım kullanılmıştır. Öncelikle dokümandan hem n-gram ve bag-of-words gibi metinsel özellikler, hem de renk ve doku gibi görsel özellikler çıkarılmıştır. Daha sonra, doküman ile iddia edilen yazar arasındaki benzerliği karşılaştırmak için bir Support Vector Machine (SVM) sınıflandırıcısı kullanılmıştır. Yaklaşımın değerlendirilmesi için 657 yazarın 1.539 el yazısı sayfasını içeren IAM veri setinden faydalanılmıştır. Önerilen yaklaşım, diğer mevcut yöntemlere göre daha iyi performansla %97,96'lık bir doğruluk elde etmiştir.

2019 yılı Yang ve ekibinin (Yang vd., 2019) çalışmasında, terimler arasındaki anlamsal ilişkiyi ölçmek amacıyla hibrit bir model sunulmaktadır. Anlamsal benzerlik yöntemleriyle ilgili önceki çalışmaların, ya terimler arasındaki anlamsal ağın yapısına ya da yalnızca terimlerin bilgi içeriğine odaklanıldığı ve bu nedenle mevcut yaklaşımların bilgi tabanının ve derlemin boyutuyla sınırlı olacağı belirtilmiştir. Bu çalışmada, büyük ölçekli bir anlamsal ağ kullanarak anlamsal benzerliği hesaplamak için verimli ve etkili bir yaklaşım önerilmektedir. Bu yaklaşım, büyük bir kavram grafiği olan Probase yapısına dayanmaktadır. Probase içerisindeki bilgiler, milyarlarca web sayfasından ve yılların arama günlüğünden yararlanmaktadır. MC, WS353-Sim ve R&G kelime benzerliği veri kümeleri üzerinde gerçekleştirilen deneylerde, önerilen modelin MC veri setinde %88, WS353-Sim veri setinde %70 ve R&G veri setinde %91 oranında doğruluk elde ettiği belirtilmiştir.

2019 yılında Wang ve arkadaşlarının (Wang vd., 2019) çalışmasında, doküman doğrulama için derin öğrenmeye dayalı bir yaklaşım uygulanmıştır. Ekip, dokümandan özellikleri çıkarmak için Convolutional Neural Network (CNN) ve ardından çıkarılan özellikler arasındaki zamansal bağımlılıkları yakalamak için Long Short-Term Memory (LSTM) makine öğrenmesi yöntemlerini kullanmıştır. Son olarak, belge ile iddia edilen yazar arasındaki benzerliği karşılaştırmak için bir SoftMax sınıflandırıcısı uygulanmıştır. Yaklaşımın değerlendirilmesi için 170 yazardan 1.700 doküman içeren PAN 2014 veri seti kullanılmıştır. Önerilen yaklaşımın, PAN 2014 veri kümesinde %89,2 doğrulukla en son teknolojiye sahip sonuçlara ulaştığı belirtilmiştir.

DDİ'de derin öğrenmenin ortaya çıkışı bir paradigma değişimine işaret ediyordu. RNN mimarisini ve LSTM kullanan diziden diziye modeller, değişken uzunluklu dizileri işleme yetenekleri göz önüne alındığında önem kazanmıştır (Ladwani vd., 2022, Zhai vd., 2023). Özellikle Rush (Rush vd., 2015) ve Cho (Cho vd., 2014), bağlama duyarlı özetlemenin yolunu açan dikkat mekanizmalarını tanıttı (Zhang vd., 2016, Sunderrajan vd., 2016). Vaswani ve ekibi (Vaswani vd., 2017), Transformer mimarisinin tanıtımı önemli bir kilometre taşıydı. Transformer mimarisi, kendi dikkat mekanizmalarıyla metin bölümlerinin paralel işlenmesini

mümkün kılmıştır. Transformer mimarisi bu yönleriyle metin üretme, özetleme ve anlamsal analiz noktasında kullanılmaya başlanmıştır.

2019 yılında Jacob Devlin ve ekibi (Devlin vd., 2019), BERT adlı bir dil temsil modelini tanıtmışlardır. BERT modelinin amacı, tüm katmanlarda hem sol hem de sağ bağlamı kullanarak, etiketlenmemiş metinden derinlemesine çift yönlü gösterimleri önceden eğitmektir. BERT, önceden eğitildikten sonra, soru yanıtlama ve dil çıkarımı gibi çeşitli görevler için yüksek performanslı modeller geliştirmek üzere yalnızca ek bir çıktı katmanıyla iyileştirilebilir. Daha da önemlisi, bu, göreve özgü mimaride önemli değişiklikler yapılmasını gerektirmez. BERT kavram olarak nispeten basit olmasına rağmen ampirik olarak dikkate değer sonuçlar göstermiştir. On bir DDİ görevinde en gelişmiş sonuçları elde etmiştir. Bunların arasında General Language Understanding Evaluation (GLUE) puanının 80,5'e yükseltilmesi, Multi-Genre Natural Language Inference (MultiNLI) doğruluğunun %86,7'ye yükseltilmesi, v1.1 soru yanıtlama Testi F1'in 93,2'ye yükseltildiği Stanford Question Answering Dataset (SQuAD) iyileştirilmesi ve SQuAD v2.0 Testi F1 puanının 83.1'e yükseltilmesi de yer almaktadır.

2019 yılında Xin Liu'nun (Liu vd., 2019) çalışmasında, öz dikkat ve Transformer mimarilerinin son zamanlarda metin oluşturma görevlerinde sıklıkla kullanıldığı ve çok iyi sonuçlar elde edildiği belirtilmiştir. Bu çalışmada, metin özeti oluşturmak için LSTM ve Transformer dayanan LTABS adı verilen yeni bir sıralı model üzerinde soyutlayıcı özetleme tekniği uygulanmıştır. Bu model, Transformer modelinin özet oluşturma görevinde kodlayıcı kısmı için konum bilgisi olarak LSTM modelinin gizli katman sonucunu kullanmıştır. Aynı zamanda Transformer, LSTM modelinin dikkat matrisi olarak kullanılarak Transformer modelinin konum bilgisi almasına ve yerel dikkatin artmasına olanak sağlamıştır. Özet görevlerin Out of Vocabulary sorununu çözmek için yeni bir ağ kopyalama mekanizması da eklenmiştir. Önerilen model CNN/Daily Mail ve XSum veri setleri üzerinde değerlendirilmiştir. CNN/Daily veri setinde Rouge-1 ve Rouge-2 skorlarını sırasıyla 41,68 ve 19,33 elde ederken, XSum veri setinde 33,21 ve 12,12 elde etmiştir. LTABS modelinin anlamsal ve sözdizimsel yapı açısından en

gelişmiş modelden üstün olduğunu ve dil kalitesi değerlendirmesinde rekabetçi sonuçlar elde ettiğini vurgulanmıştır.

2020'de Zhang ve ekibi (Zhang vd., 2019) Transformer modeli olan "Pegasus" modelini tanıtmıştır. Pegasus, özetin bazı bölümlerinin çıkarıldığı ve modelin boşlukları doldurmak üzere eğitildiği Gap Sentences Generation (GSG) eğitim hedefi nedeniyle diğerlerinden farklıdır. Denetimsiz eğitim yaklaşımı, özellikle CNN/Daily Mail ve GigaWord gibi veri kümelerinde yeni ölçütler belirleyerek Bidirectional and Auto-Regressive Transformers (BART) ve T5 gibi modellerden daha iyi performans göstermiştir. Üstelik insan değerlendiriciler Pegasus modelinin özetlerini önceki modellerin özetlerine tercih etmiştir.

2020 yılında, Ruyun Wang ve ekibinin (Wang vd., 2020) çalışmasında, kaynak kodunu yorumlama ve doğal dil açıklamaları üretmenin karmaşıklıklarına atfedilen, yazılım mühendisliğinde uzun süredir devam eden bir zorluk olan kod özetlemenin karmaşıklığını araştırmıştır. Yaygın yaklaşımlar, yapısal kaynak kodu ayrıntılarını çıkarmak ve iyi yorumlar sağlamak için dil modelleriyle eşleştirilmiş soyut sözdizimi ağaçlarından yararlanırken, derin kodu anlamada genellikle yetersiz kalıyor ve uzun bağımlılık sorunuyla boğuşuyordu. Bu zorlukların üstesinden gelmek için ekip, “Functional Reinforced Transformer with BERT” kısaltması olan "Fret" modelini tanıtmıştır. Bu öncü model, kod işlevselliği anlayışını geliştirerek kod yorumu oluşturmayı geliştirmekte ve böylece uzun bağımlılık sorununu azaltmaktadır. Kod işlevselliklerini gösteren daha kesin özetlerin oluşturulmasını sağlayan, işlevsel kod yönlerini kavramak için benzersiz bir pekiştirici entegre edilmiştir. Ayrıca kaynak kod yapısının daha iyi anlaşılması için yenilikçi bir algoritma geliştirilmiştir. Deneysel sonuçlar Fret modelinin olağanüstü etkinliğini doğrulamıştır. Mevcut üst düzey yöntemleri geride bırakarak Java kod doğrulama ve özetleme için BLEU-4 puanını 24,32'ye (%14,23 mutlak artış) ve Python için 40,12'lik Recall-Oriented Understudy for Gisting Evaluation (Rouge)-L puanı elde etmiştir.

2021 yılında Imam ve ekibi (Imam vd., 2021) çalışmasında, Ethereum blok zincirini kullanarak dijital doküman doğrulaması için tasarlanan DocBlock adlı merkezi olmayan bir web uygulamasını tanıtmışlardır. Bu uygulama,

kullanıcıların dijital belgeleri yüklemesine, doğrulamasına ve indirmesine olanak tanımaktadır. Dokümanın hash değeri, kuruluşun genel anahtarına ve yükleme tarihine bağlanarak ek orijinallik doğrulaması sağlamaktadır. Dokümanda herhangi bir değişiklik yapılması durumunda hash değeri değiştirilecek ve böylece otoritenin yolsuzlukları tespit etmesi sağlanacaktır. Doküman eklendikten sonra kullanıcılar, gelecekte kullanmak üzere depolamaktan sorumlu oldukları tek seferlik InterPlanetary File System (IPFS) hash değeri almaktadır. Kullanıcılar sisteme IPFS hash değerini girdiğinde, sistem ilgili dosyayı aramakta ve orijinal dosyayı ikili kod olarak döndürmektedir. Sistem, doküman doğrulamada %80 başarı oranına ulaşmıştır. Bir belge blok zincirine yüklendikten sonra kullanıcılar sistem içinde belge üzerinde sınırlı kontrole sahip olabilir. Belge, şeffaf ve denetlenebilir bir sistem sağlayarak geleneksel teknolojilere göre üstün bir çözüm sunarken, kullanıcı kontrolü sağlamamaktadır.

2021 yılında Zhi ve ekibi (Zhi vd., 2021) çalışmasında, emtia vadeli işlem piyasasını denetleyen düzenleyici kurumlar için belge doğrulamanın önemi ele alınmıştır. Çalışmada, emtia vadeli işlem belgelerinin doğrulanmasında metin tanıma için yeni bir hibrit çözüm önerilmiştir. Metin algılama aşamasında, bilinen şablonlara göre geçerli metnin konumunu belirlemek için elle etiketlenmiş referans noktaları kullanılmıştır. Metin tanıma aşamasında, rastgele uzunluktaki metni tanımak için konuşma tanıma alanından ödünç alınan Bağlantılı Zamansal Sınıflandırma kaybını (Connectionist Temporal Classification Loss) tanıtılmıştır. Hibrit çerçeve daha sonra kalite kontrol belgeleri ve iş sözleşmelerinin metin tanıma işlemine uygulanmıştır. RSS3 belge veri seti üzerinde yapılan deneysel çalışmalar, ortalama bir doğruluk oranı olan %90'ı ortaya koymuştur. Bu yaklaşım güvenilir metin tanıma sonuçları sağladı ve iş gücü maliyetlerini önemli ölçüde azaltarak belge doğrulama alanına önemli bir katkıda bulunmuştur.

2021 yılında Dan Jiang ve arkadaşlarının (Jiang vd., 2021) çalışmasında, BSSA adı verilen BERT Transformer ve Seq2Seq (Sıradan Sıraya) modellerine dayanan otomatik bir doküman özetleme ve anlamsal doğrulama sistemi önerilmiştir. Çalışma üç aşamadan oluşmaktadır. BERT modeli başlangıçta metin özelliğini

kodlayıcının bir parçası olarak çıkarmıştır. Daha sonra LSTM modeli, kodlayıcının başka bir ana parçası olarak metnin anlamsal özelliğini üretmiştir. Son olarak dikkat mekanizmasına sahip kod çözücü soyut özetler üretmiştir. Model, hem Çin Bilimsel Literatürü LCS hem de İngilizce GigaWord veri kümeleri üzerinde deneysel olarak test edilmiştir. Modeli değerlendirmek için Rouge metriği kullanılmıştır. Rouge-1, Rouge-2 ve Rouge-L puanları LCS veri setinde sırasıyla 38,6, 25,0 ve 35,1 iken GigaWord veri setinde 40,3, 20,8 ve 35,7'dir. Sonuçlara göre BSSA modelinin temel modellere göre önemli iyileştirmeler sağladığı belirtilmiştir.

2021 yılında Aghajanyan ve ekibi (Aghajanyan vd., 2021) "PreSumm" adlı Transformer tabanlı bir model önermiştir. PreSumm modeli, DDİ görevlerinde oldukça etkili olduğu gösterilen bir tür sinir ağı olan Transformer mimarisini temel almaktadır. PreSumm modeli iki aşamalı bir yaklaşım kullanmaktadır. İlk aşamada giriş belgesindeki en önemli cümleleri belirleyerek çıkarımsal özetleme gerçekleştirir. Bunu, giriş belgesini bir dizi yerleştirmeye kodlamak için önceden eğitilmiş bir Transformer modeli kullanarak yapar; bunlar daha sonra en belirgin cümleleri seçen grafik tabanlı bir sinir ağına beslenir. İkinci aşamada model, seçilen cümlelerden yola çıkarak bir özet üreterek soyutlama gerçekleştirir. Bu, çıkarılan bilgilere dayanarak tutarlı özetler üretmek üzere eğitilmiş dönüştürücü tabanlı bir dil modeli kullanılarak yapılmıştır. Yazarlar PreSumm modelini üç veri kümesinde değerlendirdiler: CNN/Daily Mail, New York Times ve WikiHow. CNN/Daily Mail veri setinde %45,17, New York Times veri setinde %39,31 ve son olarak WikiHow veri setinde %57,04'lük ROUGE-1 skoruna ulaştığı belirtilmiştir.

2022 yılında Amanuel Alambo ve ekibinin (Alambo vd., 2022) çalışması, klinik alanda yapılan önemli anlamsal metin özetleme çalışmalarından biridir. Bu çalışmada, iki aşamalı bir yaklaşımla bilgi güdümlü çok amaçlı optimizasyon kullanılarak klinik belgelerin soyutlayıcı özetlenmesi için bir model önerilmiştir. İlk aşamada, adlandırılmış varlıklar (Named entities) kullanılarak alana özgü bilgi kaynaklarından klinik bilgi alımı yapılmıştır. İkinci aşamada, çok amaçlı optimizasyon kullanılarak uçtan uca Transformer kodlayıcı-kod çözücü modellerinin ortak eğitimi gerçekleştirilmiştir. Önerilen Transformer modelinin

eğitimi sırasında üretim kaybı, varlık kaybı ve bilgi kaybı maliyet fonksiyonları birlikte optimize edilmiştir. Önerilen model üç farklı veri setinde değerlendirilmiştir: kalp yetmezliği hastalarının klinik notları (HF), Indiana Üniversitesi Göğüs Röntgeni Koleksiyonu (IU X-Ray) ve MIMIC-CXR. Model, HF, IU X-Ray ve MIMIC-CXR veri setlerinde sırasıyla 31,62, 36,25 ve 36,45 F-1 puanları elde etmiştir.

2023 yılında Yashoda Saisree Bareedu ve ekibi (Bareedu vd., 2023) çalışmasında, yapılandırılmamış belgelerde bulunan modelleme kısıtlamalarını anlamsal doğrulama için açık kurallar olarak resmi olarak temsil edecek bir yaklaşım geliştirmiştir. Çalışmada bu yaklaşım OPC UA endüstriyel standardı bağlamında kullanılmış ve önemli bulgular elde edilmiştir. Modelleme kısıtlarının OWL ve SPARQL gibi anlamsal Web teknolojileri kullanılarak sistematik olarak kurallara dönüştürülebileceği belirlenmiş ve bunlar daha sonra bir taksonomi halinde düzenlenmiştir. Araştırmacılar ayrıca tabloları ve belge metnini analiz ederek spesifikasyon belgelerindeki kısıtlamaları (%87 hassasiyet ve %94'e varan F1 puanıyla) otomatik olarak belirleme yeteneğini de göstermiştir. Ayrıca, modelleme kısıtlamalarının kurallara dönüştürülmesi, kısıtlamalar tablolardan çıkarıldığında tamamen otomatik hale getirilebiliyor, metinden çıkarım ise süreçte İnsan gerektiriyordu. Dolayısıyla bu çalışma, endüstriyel standartlara dayalı anlamsal doğrulama süreçlerinin otomasyonunda önemli bir ilerlemeye işaret etmiştir.

2023 yılında Liuchang Xu ve ekibi (Xu vd., 2023) çalışmasında, akıllı şehirlerin gerçekleştirilmesine yönelik kritik bir adım olan adres eşleştirme sorununun üstesinden gelmek için yeni bir yaklaşım geliştirmiştir. Kural tabanlı veya metin benzerliği teknikleri gibi geleneksel yöntemlerin sınırlamalarının farkında olarak, özellikle standart dışı adres verileriyle uğraşırken, çeşitli adresler arasındaki anlamsal benzerlikleri belirlemek için Transformer mimarisi kullanılmıştır. Kapsamlı bir adres veri seti kullanarak adres semantik modelini önceden eğitmişlerdir. Bu model daha sonra, adrese özgü coğrafi özellikler açısından zengin, özel olarak derlenmiş bir veri kümesiyle ince ayarlanmıştır. Bu süreç, karmaşık eşleştirme problemini etkili bir şekilde ikili sınıflandırma tahmin

görevine dönüştürülmüştür. Bu yaklaşım, çeşitli yerleşik adres eşleştirme tekniklerini %90'lık bir F1 puanı ile geride bırakmıştır.

3. MATERYAL VE YÖNTEM

3.1 Doğal Dil İşleme

DDİ, insan dilini bilgisayarlar tarafından anlaşılabilir ve işlenebilir bir formata çevirmek için kullanılan bir yapay zeka alt alanıdır. Bu, bilgisayarların metin işleme, analiz etme, metin üretme ve insanlarla doğal bir dilde etkileşimde bulunma yeteneğini geliştirme amacını taşır.

DDİ, metinlerin sözcük düzeyinde analiz edilmesi, dilbilgisi kurallarının ve anlamın çıkarılması gibi detaylı işlemleri içermektedir. Bu sayede, metinleri anlayabilen ve içeriğini yorumlayabilen bilgisayar programları ve sistemlerinin geliştirilmesi sağlanabilmektedir. DDİ'nin temel bileşenleri şunlardır:

- **Metin Anlama:** DDİ, metinleri anlama yeteneği sunar. Bu, metinlerin içeriğini analiz ederek anlamak anlamına gelir. Metinlerin dil bilgisel ve anlamsal yapıları çözümlenir (Li vd., 2019).

- **Metin Üretme:** DDİ, bilgisayarların metinler üretmesini sağlar. Bu, otomatik metin oluşturma, makine çevirisi ve otomatik özetleme gibi işlevleri içerir (Seo vd., 2023).

- **Duygu Analizi:** DDİ, metinlerdeki duygusal ifadeleri ve tonlarını tespit edebilir. Bu, metinlerin pozitif, negatif veya nötr gibi basit veya daha detaylı duygu analizi için kullanılır (Durga vd., 2023).

- **Metin Sınıflandırma ve Kategorizasyon:** DDİ, metinleri belirli kategorilere veya sınıflara ayırabilir. Örneğin, e-posta spam filtreleri ve haber makalelerini konularına göre sınıflandırma problemleri verilebilir (Fiok, 2021).

- **Konuşma Tanıma ve Sentezi:** DDİ, konuşmayı metne çevirme (konuşma tanıma) veya metni konuşmaya dönüştürme (konuşma sentezi) amaçlı

kullanılabilir. Bu, sesli komutları anlama, sesli yanıtlar üretme ve sanal asistanlar gibi uygulamalarda kullanılır (Jin, 2024).

DDİ'nin uygulama alanları çok geniştir. Metin üreten ve bunları işlemek isteyen her sistemde, örneğin metin madenciliği, dil çevirisi, sağlık bilgi sistemleri, eğitim, hukuk, sosyal medya analizi ve daha birçok alanda önemli uygulamaları bulunur. Bu teknoloji ile, büyük miktarda metin verisi ile çalışan, otomatikleştirilmiş ve insan-makine etkileşimini gerçekleştiren sistemler (Siri (Tenhundfeld vd., 2022), IBM Watson (Conde, 2022)) tasarlamak mümkün olmuştur.

3.2 Doküman Doğrulama

Doküman doğrulama, bilginin güvenilirliğini, kaynağından doğrulamayı amaçlayan bir süreçtir. Özellikle büyük miktarda verinin dolaştığı ve paylaşıldığı çağımızda, yanlış veya sahte bilgilerin yayılması potansiyel olarak ciddi sonuçlara yol açabilir. Bu nedenle, doküman doğrulama modern çağda hayati bir rol oynamaktadır. Doküman doğrulama aşağıdaki nedenlerle önemlidir:

- **Bilgi Güvenliği:** Doğru bilgiye sahip olmak, bireylerin ve kuruluşların güvenliğini sağlamada kritik bir rol oynar. Yanlış veya sahte bilgiler, sahtekarlık veya siber saldırılar gibi tehlikelere yol açabilir.

- **Karar Alma Süreçleri:** Özellikle iş dünyasında ve yönetimde, doğru bilgilere dayalı kararlar almak önemlidir. Yanlış veya yanıltıcı bilgilere dayalı kararlar, işletmelerin başarısını olumsuz etkileyebilir.

- **Kamu Politikaları ve Kamuoyu:** Kamu politikalarının oluşturulması ve kamuoyunun bilgilendirilmesi, doğru ve güvenilir verilerle olmalıdır.

- **Bilimsel Araştırmalar:** Bilimsel araştırmalarda doğru ve güvenilir verilerin kullanılması önemlidir. Yanlış veya sahte bilgiler, bilimsel çalışmaların itibarını zedeler.

Doküman doğrulamanın temel amacı, bir özet doküman içeriğinin, kaynak olarak kullanılan orijinal doküman veya dokümanlarla doğruluğunu değerlendirmektir. Özet metin, orijinal dokümanların önemli bilgilerini özümseyerek daha kısa bir versiyonunu sunar. Doğrulama işlemi de bu özet metnin ne derecede başarılı bir şekilde orijinal dokümanları yansıttığını ölçmeyi hedeflemektedir.

Doküman doğrulama sürecinde kullanılan özetleme teknikleri oldukça önemlidir. Özetler, çıkarımsal (Naik vd., 2017) veya soyut özetleme (Dilawari vd., 2023) teknikleri kullanılarak oluşturulabilir. Çıkarımsal özetler, orijinal dokümandaki cümleleri özet için seçerken, soyut özetler cümleleri yeniden üretmektedir.

Doküman doğrulama, genellikle üç farklı yöntemle gerçekleştirilir.

3.2.1 Elle Doküman Doğrulama

El ile doküman doğrulama, uzmanların bilgi ve deneyimlerini kullanarak gerçekleştirilmektedir. Bu yöntemde, belirli bir konuda uzman olan kişilerin özet doküman ile orijinallerini karşılaştırarak doğrulama yapması amaçlanmaktadır. El ile doğrulama süreci, özellikle karmaşık ve uzmanlık gerektiren konular için tercih edilir. Uzmanların dikkati ve deneyimi, doğrulama sürecinin başarısını doğrudan etkileyebilir.

3.2.2 Yarı Otomatik Doküman Doğrulama

Yarı otomatik doküman doğrulama hem otomasyonun hem de insan müdahalesinin kullanıldığı bir yöntemdir. Özet dokümanları otomatik özetleme algoritmaları oluşturur ve daha sonra insanlar bu özetleri inceleyerek orijinal dokümanla karşılaştırır. Bu yöntem, büyük veri havuzlarını hızlı bir şekilde değerlendirmek ve doğrulamak için kullanışlıdır. Otomatik özetleme algoritmalarının kullanılması, büyük veri akışlarını daha etkili bir şekilde işlememizi sağlar.

İnsan faktörü, özellikle aşağıdaki durumlarda kullanılmaktadır.

1. **Uzmanlık Gerektiren Konular:** Karmaşık ve uzmanlık gerektiren konularda, insanlar daha derinlemesine anlayış ve uzmanlık sağlayabilir.

2. **Duygusal ve Yaratıcı Değerlendirme:** İnsanlar, duygusal tonlamaları, ironiyi ve yaratıcılığı daha iyi değerlendirebilir. Bu özellikle sanatsal, edebi veya duygusal içeriklerde önemlidir.

3. **Özel Bağlamlar ve İnce Nüanslar:** Dilin özel bağlamlarını ve ince nüanslarını anlama yeteneği, insanların, özellikle belirli terminoloji veya kültürel referanslar içeren metinleri daha iyi değerlendirebilmelerini sağlar.

Yarı otomatik doküman doğrulamanın elle doğrulamadan farkı, otomatik algoritmaların hızlı bir şekilde özet oluşturmasına ve büyük veri havuzlarını işlemesine odaklanırken, insan faktörünün daha derin anlayış, uzmanlık ve duygusal değerlendirmeler sağlamada etkin olduğu noktada yatar. Bu birleşim, doğrulama sürecini hızlandırırken aynı zamanda kalite ve doğruluk düzeyini artırabilir.

3.2.3 Otomatik Doküman Doğrulama

Otomatik doküman doğrulama, tamamen otomatik bir süreçtir. Bu yöntemde, özet dokümanlar otomatik özetleme algoritmaları kullanılarak oluşturulur ve bilgisayar programları veya yapay zeka sistemleri tarafından orijinal dokümanlarla karşılaştırılır. Otomatik doğrulama, büyük miktarda verinin hızlı bir şekilde işlenmesini sağlar ve insan müdahalesini en aza indirir.

3.3 Doğrulama Kriterlerinin Belirlenmesi

Doğrulama yapılacak veri setinin türü, başarım oranı tespiti açısından oldukça önemlidir. Çalışılan doküman türü için hangi verilerin önemli olduğunun

bilinmesi çalışmayı kolaylaştıracaktır. Bu tez kapsamında özellikle finansal dokümanlar üzerinde çalışılacağından, doğrulama aşamasında finansal doküman için önemli özellikler (Örneğin, finansal terimler ve semboller, finansal rakam ve sayılar, finansal kurum ve kuruluş isimleri vb.) göz önüne alınmıştır.

3.4 Doküman Doğrulama Başarım Oranının Belirlenmesi

Bu aşamada, her bir dokümanın bilgi çıkarımları yapıldıktan sonra, ilgili dokümanın konusuna göre belirlenen doğrulama özellikleri göz önüne alınarak, orijinal doküman ile özet dokümanın karşılaştırılması yapılmaktadır. Bu şekilde özet doküman, orijinal doküman ile karşılaştırılarak doğrulama gerçekleştirilir. Doküman doğrulama başarım oranı tespiti için farklı teknikler bulunabilir. Burada en basit şekilde özel kimlik (named entities) tabanlı (Toprak vd., 2023) karşılaştırmanın yanı sıra, daha karmaşık olan semantik analiz (Zeng vd., 2018) yöntemleri kullanılmıştır.

3.5 Doküman Özetleme

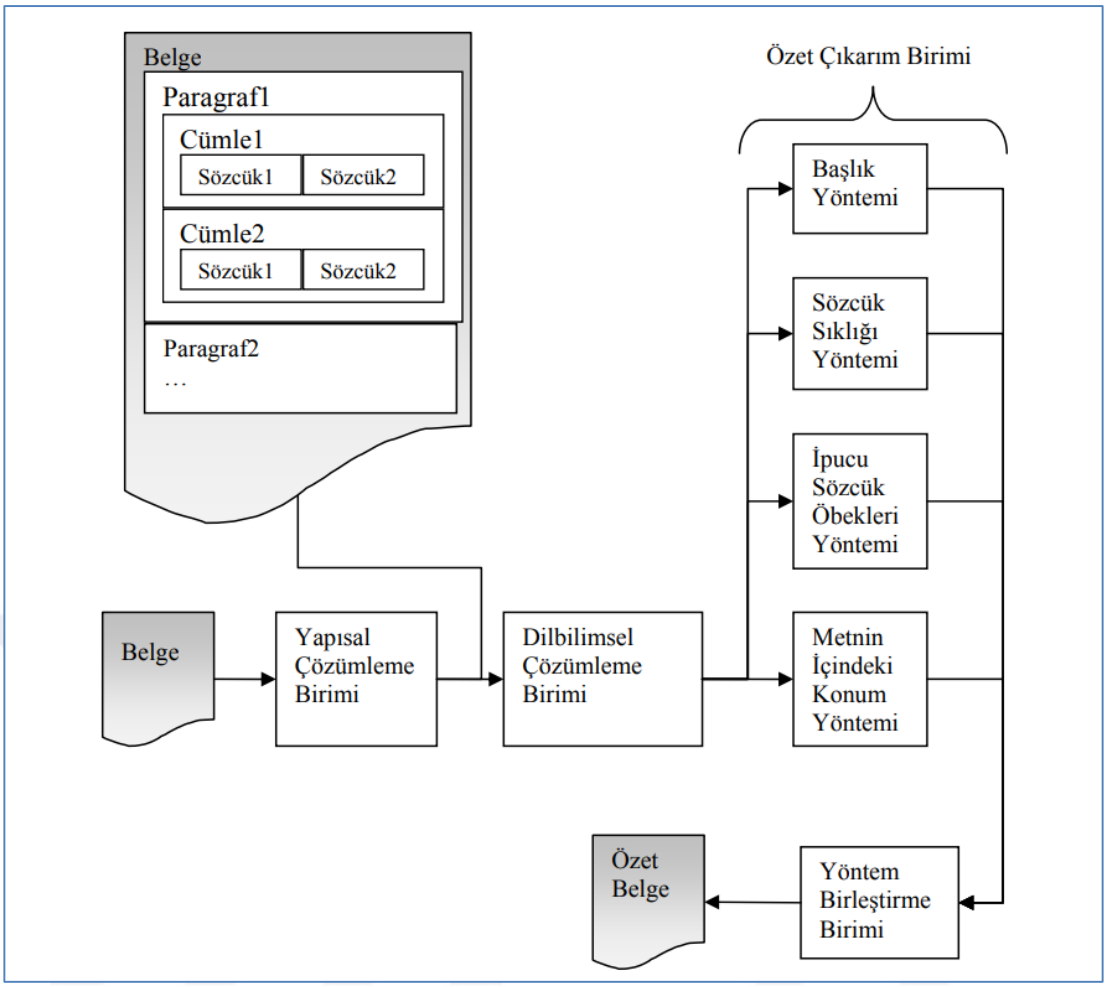
Dijital içeriğin katlanarak büyümesiyle birlikte, büyük hacimli bilgileri kısa ve bilgilendirici özetlere dönüştürmek için etkin yöntemlere acil bir ihtiyaç oluşmuştur. Bu ihtiyaç gazetecilik, hukuk, bilimsel araştırma ve iş zekası dahil olmak üzere çeşitli alanlarda ortaya çıkmaktadır. İnsanlar tarafından oluşturulan özetler zaman alıcı, öznel ve önyargılara açık olabilir; bu da otomatik özetleme tekniklerinin geliştirilmesini önemli hale getirmektedir. Doküman özetlemenin amacı, temel bilgileri ve ana fikirleri korurken metinsel belgelerin yoğunlaştırılmış temsillerini oluşturmaktır. İyi yazılmış bir özet, tam metin belgelerin önemli ayrıntılarını ve genel içeriğini kapsamalıdır. Okuyuculara belgenin içeriği kapsamlı bir şekilde sunulmalı, metnin tamamını okumak zorunda kalmadan ana mesajı yakalamalarına olanak sağlanmalıdır.

Doküman özetlemeye yönelik ilk yaklaşımlar, kaynak dokümanlardan cümleleri veya paragrafları seçip birleştirmeyi ve özeti oluşturmayı içeren çıkarımsal yöntemlere (Suleiman vd., 2019; Shirwandkar vd., 2018) dayanıyordu. Bu

yöntemlerde, her cümlenin dikkat çekiciliğini ve alaka düzeyini belirlemek için sıklıkla istatistiksel ve dilsel özelliklerden yararlanır. Çıkarımsal özetleme teknikleri umut verici sonuçlar vermiş olsa da sınırları vardır. Alıntılanan cümleler her zaman doğal veya tutarlı bir şekilde sıralanamayabilir, bu da özetlerin okunabilirlik düzeyinin düşük olmasına neden olabilir.

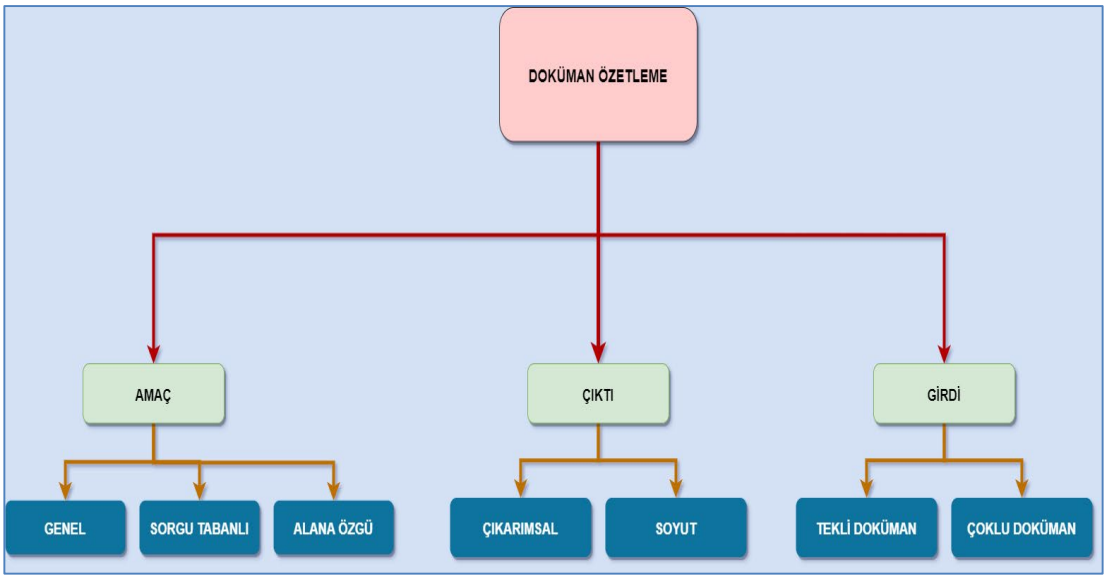
Derin öğrenmede ile çıkarımsal özetlemenin bu sınırlamasını ortadan kaldırmak için araştırmacılar dikkatlerini soyutlayıcı özetleme yaklaşımlarına (Moawad vd., 2012; Bhargava vd., 2016; Liaqat vd., 2022; Etemad vd., 2021) çevirmişlerdir. Soyut özetlemede, dokümanın içeriğini tam olarak anlayan, orijinal metindeki anlamı koruyan, bilgi kaybı olmadan yeniden oluşturulmuş anlamlı cümlelerle bir özet oluşturulmaktadır.

Soyut özetleme, kaynak dokümandaki bilgileri daha kısa ve tutarlı bir şekilde anlamak ve yeniden ifade etmek için gelişmiş DDİ ve derin öğrenme modellerine dayanır. Bu yöntemler, yalnızca cümle çıkarmanın ötesine geçen insan benzeri özetler üretme konusunda büyük potansiyel göstermiştir (Erkan vd., 2004; Gupta vd., 2019). Şekil 3.1, doküman özetlemenin sistem mimarisin görülmektedir.



Şekil 3.1 Doküman özetleme sistemi mimarisi

Doküman özetleme, giriş dokümanlarının sayısal değerine göre tekli (Mao vd., 2021; Hariharan vd., 2008) ve çoklu doküman (Lee vd., 2013; Jin vd., 2020) özetleme olmak üzere ikiye ayrılır. Doküman özetleme türlerini içeren örnek bir şema Şekil 3.2'de verilmiştir.



Şekil 3.2 Doküman özetleme türleri

Soyutlayıcı doküman özetleme çeşitli nedenlerden dolayı önemlidir. İlk olarak, uzun dokümanları kısa özetlere sıkıştırarak bilgi çıkarımını kolaylaştırır. Çok fazla miktarda veriye maruz kaldığımız günümüz bilgi çağında soyutlama, dokümanlardan en önemli bilgilerin çıkarılmasına yardımcı olur. Örneğin, haber makalelerinde bir özet, tartışılan ana olaylara ve önemli noktalara kısa bir genel bakış sağlayabilir. İkinci olarak, soyutlayıcı özetleme zamandan tasarruf sağlar ve verimliliği artırır. Tam metin dokümanların özetlenmiş bir versiyonunu sunarak, kullanıcıların dokümanların tamamını okumakla zaman kaybetmeden içeriği ve alaka düzeyini hızlı bir şekilde değerlendirmesine olanak tanır. Bu, özellikle zamanın sınırlı olduğu veya büyük miktarda doküman ile uğraşıldığı senaryolarda faydalıdır. Hukuk veya araştırma gibi alanlardaki profesyoneller, soyut özetler aracılığıyla önemli bilgileri verimli bir şekilde tanımlayabilir. Dahası, soyutlayıcı özetler, altta yatan bağlamı ve anlamı yakalayarak içeriğin anlaşılmasına yardımcı olur. Tutarlı ve bağlamsal olarak uygun özetler oluşturarak soyutlayıcı yaklaşım, tam metin içeriğinin özünü yakalar ve onu daha sindirilebilir bir formatta sunar. Bu özellikle karmaşık kavramların ve bulguların daha anlaşılır bir şekilde açıklanabildiği eğitim materyalleri, bilimsel makaleler ve teknik belgeler için kullanışlıdır. Örneğin, yıllık değerlendirme raporları, kazanç açıklamaları ve mali tablolar gibi finansal belgeler kapsamlı bilgi ve veriler içerir. Soyutlayıcı doküman özetleme, bu belgelerden en önemli iç

görülerin ve temel bulguların çıkarılmasına yardımcı olur. Yatırımcılar, analistler ve karar vericiler de dahil olmak üzere paydaşların, belgenin tamamını incelemeye gerek kalmadan finansal performansı, temel ölçümleri ve önemli olayları hızlı bir şekilde anlamalarını sağlar. Soyutlayıcı özetler arama motoru, öneri sistemi veya çeşitli alanlardaki bilgi erişim amacına yönelik, bilginin erişilebilirliğini ve kullanılabilirliğini artırır.

3.6 Veri Seti

Bu çalışmada doğrulama yapılacak hedef doküman bir finans kuruluşunun finansal raporu olarak seçilmiştir. Verilerin gerçek hayat verisi olması, çalışmanın başarımını test etmek açısından oldukça önemlidir. Çalışmada Transformer modelini eğitmek için kullanılan finansal veri seti Kaggle sisteminden (Archive 2023) alınmıştır. Örnek bir orijinal tam metin doküman ve özeti Çizelge 3.1'de verilmiştir.

Çizelge 3.1 Transformer modelini eğitmek için kullanılan veri seti

Orijinal Tam Metin Doküman	**Orijinal Özet Doküman**
Ask Jeeves has become the third leading online search firm this week to thank a revival in internet advertising for improving fortunes. The firm's revenue nearly tripled in the fourth quarter of 2004, exceeding $86m (£46m). Ask Jeeves, once among the best-known names on the web, is now a relatively modest player. Its $17m profit for the quarter was dwarfed by the $204m announced by rival Google earlier in the week. During the same quarter, Yahoo earned $187m, again tipping a resurgence in online advertising. The trend has taken hold relatively quickly. Late last year, marketing company Doubleclick, one of the leading providers of	Ask Jeeves has become the third leading online search firm this week to thank a revival in internet advertising for improving fortunes. Its $17m profit for the quarter was dwarfed by the $204m announced by rival Google earlier in the week. During the same quarter, Yahoo earned $187m, again tipping a resurgence in online advertising. Neither Ask

online advertising, warned that some or all of its business would have to be put up for sale. But on Thursday, it announced that a sharp turnaround had brought about an unexpected increase in profits. Neither Ask Jeeves nor Doubleclick thrilled investors with their profit news, however. In both cases, their shares fell by some 4%. Analysts attributed the falls to excessive expectations in some quarters, fuelled by the dramatic outperformance of Google on Tuesday.	Jeeves nor Doubleclick thrilled investors with their profit news, however. Ask Jeeves, once among the best-known names on the web, is now a relatively modest player.

Reuters veri seti, birçok DDİ ve metin madenciliği araştırması için kullanılan ünlü bir metin veri setidir. Bu veri seti, haber makalelerini içerir ve genellikle metin sınıflandırma, metin sıralama, metin kümeleme, bilgi çıkarma ve duygu analizi gibi metin madenciliği ve DDİ görevlerinin değerlendirilmesi için kullanılır. Reuters veri seti şunları içerir:

- **Metin Belgeleri:** Veri setinde, Reuters haber ajansına ait çok sayıda haber makalesi bulunmaktadır. Bu metin belgeleri, genellikle finans, iş, teknoloji, spor ve diğer haber kategorilerindeki haberleri kapsar.

- **Etiketler:** Her haber makalesi, ilgili kategoriyi veya konuyu belirten etiketlerle (etiketler) ilişkilendirilmiştir. Örneğin, bir makale "finans" veya "spor" gibi bir etiketle işaretlenmiştir.

Reuters finansal veri seti, Reuters web sitesinden alınan “Financial News Articles”ın geniş bir koleksiyonudur. Başlangıçta “Using Structured Events to Predict Stock Price Movement: An Empirical Investigation” başlıklı makale için kullanılan bu yapılandırılmamış veri seti, Finansal Verilerin güçlü bir tarihi deposudur. Reuters-21578 veri seti, metin sınıflandırma araştırması için en yaygın kullanılan veri koleksiyonlarından biridir (Wei vd., 2014). Reuters veri setinin genel içerik bilgisi Çizelge 3.2’de görülmektedir.

Çizelge 3.2 Reuters veri setinin genel içerik bilgisi

Parametre Adı	**Parametre Değeri**
Veri tipi	Metin
Farklı Belge Konusu Sayısı	20
Her Belge Konusundaki Haber Sayısı	15
Özetin Uzunluk Sınırı	1050 Bayt
Veri Setindeki Toplam Cümle Sayısı	27670

Reuters veri seti, metin sınıflandırma modellerini eğitmek ve test etmek için kullanılan bir öğrenme veri seti olarak oldukça değerlidir. Araştırmacılar, bu veri setini kullanarak metin sınıflandırma algoritmalarını, metin özetleme sistemlerini ve metin madenciliği tekniklerini geliştirmek için çalışmaktadırlar. Reuters veri seti, genellikle DDİ ve metin madenciliği alanındaki çalışmaların sonuçlarını karşılaştırmak ve değerlendirmek için standart bir referans olarak kullanılır. Bu veri seti, farklı dil modelleri ve algoritmalarının performansını test etmek ve karşılaştırmak için de kullanışlıdır.

3.7 Ön İşleme

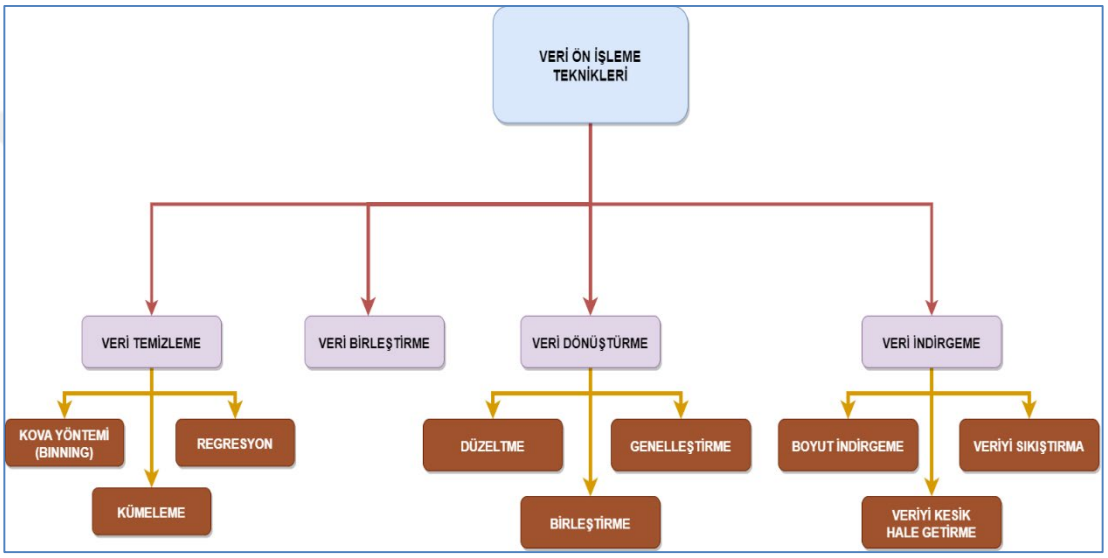
Bilişim dünyasının gelişimi insan yaşamına birçok açıdan büyük kolaylıklar getirmiştir. Bu kolaylıkların yanında başka açılardan birtakım zorluklar da getirdiği bilinmektedir. Kullanılan teknolojik araçların çokluğu, mobil uygulamaların yaygınlaşması, internetin kolay erişilebilirliği gibi olanaklar sonucu çok büyük miktarda veri birikimi oluşmuştur. Bu büyük verileri işlemek artık klasik yöntemler ile güçleşmiş hatta imkânsız hale gelmiştir (Oğuzlar, 2003; Zhou, 2002).

DDİ çalışmalarında başarılı sonuçlar elde etmek için kullanılacak verilerin kaliteli olması gerekir. Kaliteli veriler elde etmek için kullanılacak verilerin ön işleme süreçlerinden geçmesi gerekir. Böylece DDİ çalışmalarında başarılı sonuçlar elde edilebilir. Çok sayıda uygulamada veri ön işlemenin türlerinin bir tanesinden daha fazlasına ihtiyaç duyulmaktadır. Bu sebeple veri ön işlemede tür

belirlenmesi önemlidir (Famili vd., 1997). Bu yaklaşımdan yola çıkarak çok sayıda veri ön işleme tekniği geliştirilmiştir. Bunlara; veri temizleme, veri birleştirme, veri dönüştürme ve ayrıklaştırma, veri indirgeme gibi teknikler örnek olarak verilebilir.

3.7.1 Veri Ön İşleme Teknikleri

Ön işleme yöntemlerinin detaylı hali Şekil 3.3.'de gösterilmiştir.



Şekil 3.3 Veri ön işleme teknikleri

Veri temizleme: Toplanan verilerin ilk ön işleme aşamasıdır. Bu aşamada, eksik verilerin tamamlanması, gürültünün ve verilerdeki tutarsızlıkların giderilmesi işlemleri yapılmaktadır. Herhangi bir değişkene ilişkin eksik değerlerin doldurulması için farklı yollar vardır. Bunlardan bazıları aşağıda kısaca açıklanmaktadır.

- Eksik değer içeren kayıt veya kayıtlar atılabilir.
- Değişkenin ortalaması eksik değerlerin yerine kullanılabilir.
- Aynı sınıfa ait tüm örneklemler için değişkenin ortalaması kullanılabilir.

- Var olan verilere dayalı olarak en uygun değer kullanılabilir. Burada sözü edilen en uygun değerin belirlenmesi için regresyon veya karar ağacı gibi teknikler kullanılabilir. Örneğin yaşı x, eğitim düzeyi y olan bir kişi için ücret durumu, mevcut verilerden yukarıdaki tekniklerden birinin kullanılmasıyla tahmin edilebilir.

Veri temizleme tekniğinin kullanıldığı bir diğer problem ise gürültülü verilerdir. Gürültülü verilerin oluşmasının nedeni, bir değerin hatalı olarak ölçülmesi, hatalı veri toplama aracı, ya da verinin hatalı niteliklerinden kaynaklı olabilir. Gürültü verilerinin giderilmesi içinde çeşitli yöntemler mevcuttur. Aşağıda veri düzeltme tekniklerinden bazıları açıklanmaya çalışılmıştır:

- **Kova yöntemi (Binning):** Kova yöntemi, küçükten büyüğe veya büyükten küçüğe sıralanmış verileri düzeltmek için kullanılır. Binning yönteminde öncelikle sıralanmış veriler eşit büyüklükteki bin'lere ayrılır. Daha sonra bin'ler, bin ortalamaları, bin medyanları veya bin sınırları yardımıyla düzeltilir.

- **Kümeleme (Clustering):** Veriler içerisindeki aykırı değerler kümeler ile belirlenebilir. Benzer değerler aynı grup veya aynı küme içinde yer alırken, aykırı değerler kümelerin dışında yer alacak şekilde sınıflandırılır.

- **Regresyon (Regression):** Veriler, regresyon ile verilere bir fonksiyon uydurularak düzeltilebilir. Uydurulan fonksiyona uymayan noktalar aykırı değerlerdir.

Veri Birleştirme: Veri madenciliğinde genellikle farklı veri tabanlarındaki veriler veri ambarı denilen alanda birleştirilirler. Farklı veri tabanlarındaki verilerin tek bir veri tabanında birleştirilmesiyle şema birleştirme hataları (Schema Integration Errors) oluşur. Bu hatalar çoğunlukla tablo kolonlarının farklı isimlerde verilmesinden kaynaklı ortaya çıkmaktadır. Şema birleştirme hatalarından kaçınmak için meta veriler kullanılır. Meta veri, veriye ilişkin

veridir. Raporlama işlemlerinin sık olarak yapıldığı kurumlarda veri ambarı alanı oluşturulur ve kayıtlar meta veri olarak tutulur.

Veri Dönüştürme: Veri dönüştürme işleminin amacı, verileri veri madenciliği için uygun formlara dönüştürmektir. Veri dönüştürme; düzeltme, birleştirme, genelleştirme ve normalleştirme işlemlerden bazılarını içerebilir.

- **Veri İndirgeme:** Veri indirgeme, işlem yapılacak veri boyutunun (özellik sayısının) çok büyük olması durumunda, veriyi daha küçük boyutlu hale getirme işlemidir. Bu işlem sonucunda elde edilen küçük boyutlu veriler ile daha etkin sonuçlar elde edilebilir.

Ön işleme tekniklerine, verilerle çalışılan her alanda ihtiyaç duyulmaktadır. Yukarıda bahsedilen teknikler veri madenciliği çalışmalarında, çalışma bazında farklılık gösterse de çoğu zaman bu sırada ilerler. DDİ çalışmalarında ise, ön işleme teknikleri sisteme verilen dokümanlara uygulanarak, doküman içindeki kelimelerin standart bir forma dönüştürülmesi sağlanır. D. Tanasa (Tanasa ve Trousse, 2004) ve V. Chitraa, (Chitraa ve Davamani, 2010) çalışmalarında ön işleme tekniklerine yer vermişlerdir.

3.7.2 Çalışmada Uygulanan Ön İşleme Teknikleri

Doküman doğrulama sürecine tabi tutulacak veri setinin belli bir standartta olması gerekmektedir. Bu çalışmada kullanılacak veri seti İngilizce diline aittir. İngilizce diline ait dokümanların (özellikle finansal türde) veri ön işleme teknikleri kullanılarak standart hale getirilmiştir. Veri ön işlemi adımına tabi tutulmuş dokümanlar, doğrudan doğrulama sürecine alınmamıştır. Çünkü bu dokümanlarda yazım veya anlam hataları oluşması olasıdır. Bu çalışmada, bu tür sorunların üstesinden gelmek için veri ön işleme sonrasında hem özet hem de tam metin dokümanlar Yazım Denetleyici (Spell Checker) sürecine tabi tutulmuştur.

Çalışmamızda çeşitli kaynaklardan toparlanan dokümanlar üzerinde işlemler yapılmaktadır. Doküman türleri ve formatları birbirinden farklı olarak verilebilmektedir. Bu da beraberinde verilerin ön işlemeden geçirilerek veri temizleme yöntemi ile dokümanların uygun formata getirilmesini gerektirmektedir. İlk aşama olarak sisteme doğrulanmak üzere verilen dokümanların işleme alınabilecekleri birimlere ayrıştırılması gerekmektedir. Çalışmamızda işlemler cümle bazlı ele alınmaktadır. Bu nedenle doğrulama yapılacak orijinal özet doküman ve bu özete ait orijinal tam metin dokümanda bulunan cümleler tespit edilir. Bu amaçla TextBlob algoritması kullanılmaktadır. Daha sonra, cümlelerine ayrılmış metin parçacıkları kelimelerine ayrılır. Dokümanlar metin içinde anlamı olmayan birçok kelime içermektedir ve bu kelimeler (Stop words) cümleden çıkarıldıklarında da anlamsal bir kayba yol açmazlar. Ancak bu kelimeler ön işleme adımlarına tabi tutulmazsa sonuçlara olumsuz yönde etki edeceklerdir. Çalışmada İngilizce doküman doğrulama yapılacağından İngilizce diline ait sonlama kelimelerinin dokümanlardan atılması gerekmektedir.

3.8 Anlamsal (Semantik) Analiz

Semantik anlam, kelimelerin konu bütünlüğü (context) içinde ne anlama geldiğini ifade eden kavramdır. Dil, işaretlerin bir dizisidir ve bu işaretlerin analizi semantik anlamı meydana getirir. Bir dildeki küçük ses birimleri kelimeleri, kelimeler de cümleleri oluşturur. Bu bağlamda, dil ile dünya arasındaki bağlantı, semantik anlamı şekillendirir. Bir cümle, dünyadaki karşılıklarının bir işareti olarak ortaya çıkar. Dilde kullanılan kelimelerin anlamlarının incelenmesi ve bu anlamların ilişkilendirildiği şeyler arasındaki bağlantı, semantik anlamı oluşturur (Dipsizkuyu, 2022).

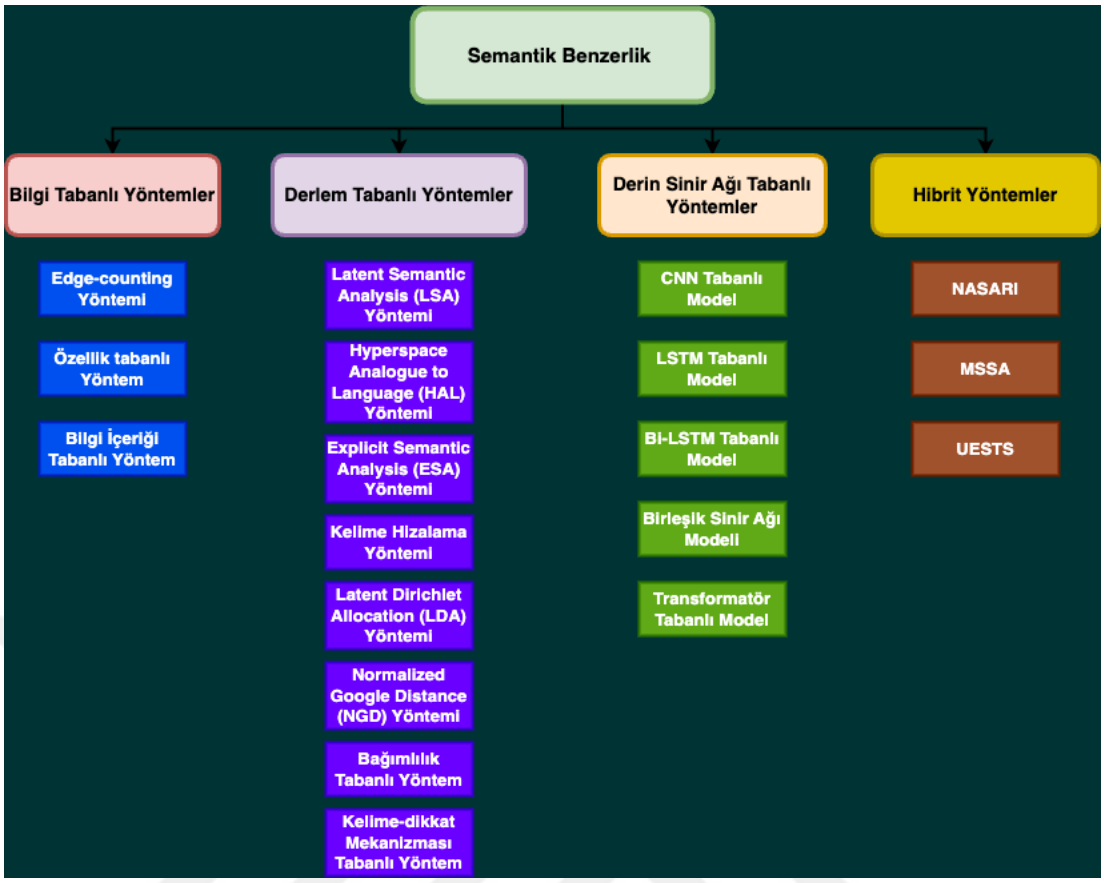
Semantik analiz ise, temel anlamı oluşturan kelimelerin ilgili bağlamlara yerleştirilmesini ifade etmektedir. Yani, kelimelerdeki ifadenin temel anlamı, semantik analizi belirlemektedir. Bu kelimeler ve anlamları, bütünsel bir yapı oluştururken, alanlar arasında ilişkilendirilmiş kelimelerin ortak bir anlam meydana getirmesi de analizin bir parçasıdır. Örneğin, “Yemek yapacağım”

ifadesi genel olarak yemeği tanımlar, ancak ne tür bir yemek olduğuna dair bilgi vermez.

Doküman doğrulama sürecinde özellikle finansal dokümanların doğrulanması sırasında tutarlılık çok önemlidir. Doğrulanacak özet doküman soyutlayıcı özetleme yöntemiyle oluşturulduğunda doğrulama süreci daha da zorlaşmaktadır. Çünkü özetteki cümleler ile orijinal tam metin dokümandaki cümleler tamamen farklı yapılarda olabilmektedir. Bu durumda iki cümlenin anlamsal olarak ölçülmesi gerekmektedir. Bu noktada anlamsal analiz yöntemlerine ihtiyaç duyulmaktadır.

Anlamsal ölçümler, anlamsal mesafeleri temsil etmek için kelimelere sayısal benzerlik puanları atamaktadır. Hesaplamalı dilbilimdeki anlamsal ilişki, iki nesnenin herhangi bir anlamsal ilişkisi varsa anlamsal olarak ilişkili olduğunu varsayar (Budanitsky vd., 2001). Anlamsal ölçü, iki kavramın hiyerarşik ilişkilerine dayalı ortaklığını temsil eden özel bir ölçüdür (Zhu vd., 2017). Genel olarak, daha genel bir kavram olan ve mutlaka hiyerarşik ilişkilere dayanması gerekmeyen anlamsal ilişkilerin özel bir durumudur. Bilgi erişimi (Kim vd., 2017), metin kümeleme (Wei vd., 2014), metin sınıflandırma (Kim vd., 2014), metin özetleme (Mohamed vd., 2019) ve kelimelerin belirsizliğini giderme (Miller vd., 2012) birçok DDİ görevinde önemli bir role sahiptir. Ayrıca öneri sistemlerinde (Wang vd., 2016), jeo-informatik (Janowicz vd., 2011) ve biyomedikal bilişim (Guzzi vd., 2011) alanlarında da uygulanmaktadır.

Şekil 3.4'te, çeşitli semantik varlık türlerini (kelimeler, kavramlar, örnekler) karşılaştırmak için kullanılabilecek semantik benzerlik yöntemlerinin genel bir sınıflandırılması verilmiştir.

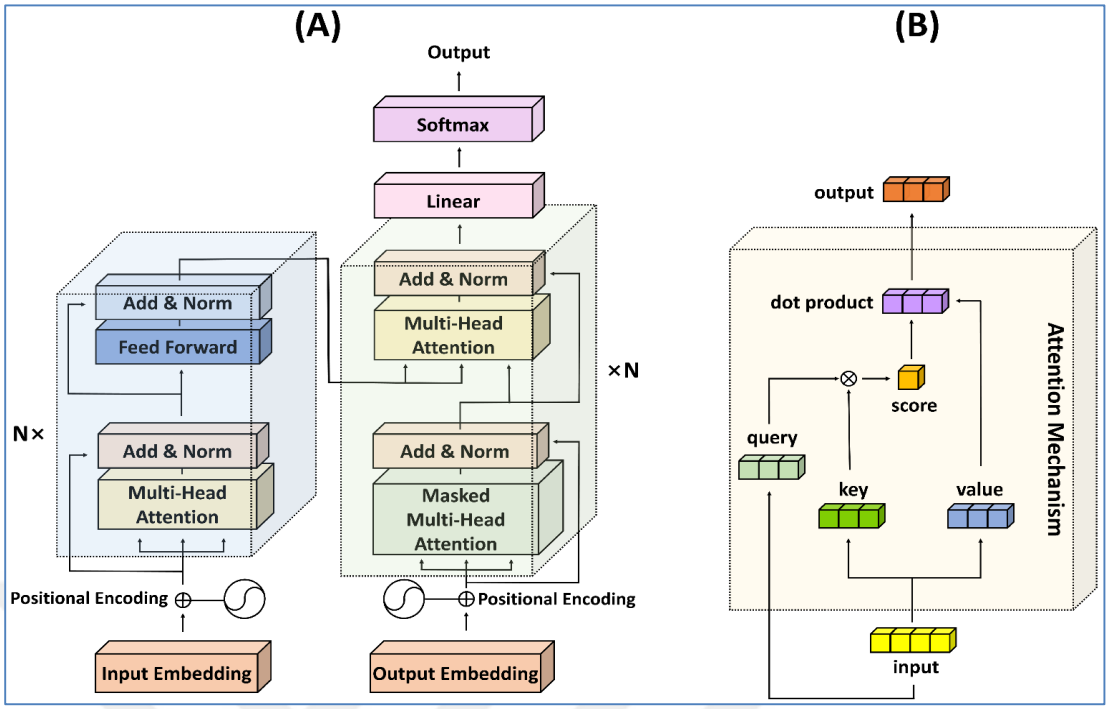


Şekil 3.4 Semantik ölçümlere genel bakış

Bu çalışmada semantik analiz görevi için Transformer mimarisinden yararlanılmıştır.

3.9 Transformer Mimarisi

Transformer, derin öğrenmenin 2012 yılında ImageNET'te CNN mimarisini kullanarak büyük bir atılım yapmasının ardından “Attention is All You Need” (Vaswani vd., 2017) başlıklı makale ile ortaya çıkan bir derin öğrenme mimarisidir. Transformer, dikkat açısından sıralı veriler üzerinde daha başarılı sonuçlar elde edebildiği için DDİ problemlerinde en çok tercih edilen mimaridir. Şekil 3.5'te Transformer mimarisinin çalışma yapısı verilmiştir.



Şekil 3.5 Transformer mimarisinin çalışma yapısı

Transformer, Şekil 3.5'te verilen kelime gibi sıralı verilerdeki ilişkileri izleyerek bağlamı ve dolayısıyla anlamı öğrenen bir sinir ağıdır. Transformer mimarisinden önce kullanıcılar, oluşturulması maliyetli ve zaman alıcı olan büyük, etiketli veri kümeleriyle sinir ağlarını eğitmek zorundaydı. Transformer, elementler arasındaki kalıpları matematiksel bir yaklaşımla tespit ederek trilyonlarca görsel ve petabaytlarca metin verisini web ve kurumsal veri tabanlarında kullanılabilir hale getirmiş ve bu ihtiyacı ortadan kaldırmıştır. Ayrıca Transformer mimarisinin kullandığı matematiksel hesaplamalar paralel işleme uygun olduğundan bu modeller hızlı çalışabilmektedir.

Transformer ile diğer yinelemeli sinir ağları arasındaki temel farklardan biri, yinelemeli sinir ağlarının t-1 zamanındaki değeri hesaplamadığı sürece t zamanındaki çıktının değerini elde edememesidir. Yani veriyi işlerken paralelleştirme sağlayamamaktadır.

Transformer mimarisinin bir farkı da artık toplama (Resudial) yapıda olmasıdır. Yani her bir işlem adımından sonra, işlem adımının sonucu ile işleme giren kısma

belirli bir toplama ve normalleştirme fonksiyonu uygulanarak iletimin sağlanmasıdır. Bu artık bağlantılar gradyanların doğrudan ağlar üzerinde akmasına izin vererek ağın eğitimine yardımcı olur. Aynı zamanda Transformer bloklarının alt katmanlarında kontrollü olarak t adıma geçilirken, t-1 dahil olmak üzere bu adıma kadar olan bilginin kaybolmadan aktarılmış olması sağlanır.

DDİ problemlerinin çözümlerinde Yapay Sinir Ağları (YSA) mimarisi kullanılırsa, kodlayıcı ve kod çözücü olmak üzere iki bileşenden oluşan bir model uygulanmaktadır. Bu yöntemde sekanslar, kodlayıcı adımları boyunca YSA ile eğitildikten sonra kod çözücü yine YSA ile çıktıyı tahmin etmeye çalışmaktadır. Buradaki temel problem aslında öğrenilmeye çalışılan girdinin tek bir bağlam özetine sıkıştırılmaya çalışılmasıdır. Bu durumda da tek seferde eğitim için modele giren sekansların uzunluğu arttıkça, kod çözücüye tahmin etmeye başlaması için girdi olarak verilen bağlam özeti de eğitim girdisini tam olarak ifade edemeyip, uygun çıktıların hesaplanamamasına neden olacaktır. Transformer kodlayıcı ve kod çözücü mimarisinde ise her kodlayıcının çıktısı kod çözücüyü besleyerek bağlamdan uzaklaşılmamasını sağlamaktadır.

Transformer mimarisinin bir dizideki kelimeler veya karakterler arasındaki uzun vadeli bağımlılıkları ele alma yeteneği, bu mimarinin kritik gücüdür. Geleneksel Tekrarlayan Sinir Ağlarında (RNN) veya LSTM ağlarında, sıralı işleme doğası ve yok olan gradyanlar gibi sorunlar, bu bağımlılıkların sürdürülmesini ve öğrenilmesini zorlaştırır. Bununla birlikte Transformer mimarisinin içindeki dikkat mekanizması, bilgi akışı için doğrudan bir yol sunarak, her bir kelimenin veya karakterin dizideki diğer her kelime veya karakterle doğrudan 'ilgilenmesini' sağlayarak bu soruna bir çözüm sağlar (Vaswani vd., 2017).

Dikkat mekanizmalarının bu yenilikçi kullanımı, Transformer modelinin önceki diziden diziye modellerden daha iyi performans göstermesinin temel nedenidir. Transformer mimarisine, yalnızca önceki öğelere dayalı olarak değil, aynı zamanda dizi içindeki tüm öğeleri dikkate alarak belirli bir giriş dizisinin bağlamını anlama yeteneği vermektedir. Bu, verilerin paralel işlenmesini

kolaylaştırarak modelin hesaplama verimliliğini artırır (Jamil vd., 2023; Han vd., 2023).

Ayrıca, Transformer mimarisinin önemli bir değişiklik yapmadan çeşitli görevlere uyum sağlama yeteneği, ona DDİ problemlerinde avantaj sağlar. Bu esneklik mümkündür çünkü Transformer doğası gereği sıralı değildir. Yani, geleneksel dil işleme modelleri, metinleri cümleler veya sözcükler şeklinde sıralı bir şekilde işlerler. Ancak Transformer mimarisi, girdi verilerini sıralı bir şekilde işlemek yerine, metindeki ilişkileri ve bağlantıları daha serbest bir şekilde öğrenir. Karmaşık yapılara sahip diller ve metinlerle uğraşırken temel bir avantaj olan, sözcük grubu (phrase) sıralanmasına gerek kalmadan veri içindeki grupları öğrenirler (Tay vd., 2022).

İleriye dönük olarak, ölçeklenebilirliği ve verilerdeki karmaşık sözcük grupları öğrenme yeteneği ile Transformer mimarisinin DDİ araştırmalarının ön saflarında yer almaya devam edeceği açıktır. İnsan dilini artan doğrulukla anlama ve üretme kapasitesine sahip daha gelişmiş dil modellerinin geliştirilmesinde etkili olacaktır. Sonuç olarak, bu modellerin uygulama yelpazesi büyümeye devam edecek ve bizi gerçek doğal dil anlayışına daha da yaklaştıracaktır (Patwardhan vd., 2023).

Transformer modeli yedi bölümden oluşmaktadır. Bu bölümlerin detayları aşağıda verilmiştir.

3.9.1 Giriş Katmanı (Input Layer)

Transformer modeline giriş olarak metin verilir. Her bir kelime veya altı kelime birimi, önce sayısal verilere dönüştürülür. Bu dönüşüm, her kelimenin belirli bir sayısal vektörle temsil edilmesini sağlar. Özellikle, kelime gömme (Word Embedding) yöntemi kullanılır.

3.9.2 Kelime Gömme Katmanı (Word Embedding Layer)

Kelime gömme katmanı, her kelimenin sayısal temsilini oluşturur. Her kelime bir vektörle temsil edilir ve bu vektörler, dil içindeki kelimelerin benzerliklerini ve ilişkilerini yansıtır. Öğrenilebilir ağırlıklar kullanılarak kelime gömme vektörleri oluşturulur.

3.9.3 Pozisyon Gömme Katmanı (Positional Embedding Layer)

Transformer, metindeki kelimenin sırasının önemli olduğunu anlar, ancak kelime gömme katmanı kelimenin konum bilgisini içermez. Pozisyon gömme katmanı, her kelimenin metindeki pozisyonunu temsil eden sayısal vektörleri ekler. Bu, modelin metindeki kelimenin konumunu dikkate almasını sağlar.

3.9.4 İşlem Katmanları (Transformer Encoder ve Decoder)

Transformer, kodlayıcı (Encoder) ve çözücü (Decoder) olarak adlandırılan iki temel bölümden oluşur. Kodlayıcı (Encoder), metni bir gösterimde (Representation) kodlar. Kodlayıcı, girdi metnini işleyerek anlamsal gösterimler oluşturur. Kodlayıcı, birden fazla kodlayıcı katmanından oluşur. Her bir katman, dikkat mekanizması (Attention Mechanism) ve çok başlıklı dikkat (Multi-Head Attention) gibi bileşenleri içerir. Bu katmanlar, metindeki farklı kavramların ilişkilerini ve bağlantılarını anlamak için kullanılır. Çözücü (Decoder) ise, kodlayıcının çıktısını kullanarak sonuç metnini oluşturur. Çözücü, dil modellemesi yaparak ve sonuç metni adım adım oluşturarak çeviri, özetleme veya diğer görevleri gerçekleştirir. Çözücü de birden fazla çözücü katmanından oluşur ve her bir katman, dikkat mekanizması gibi bileşenleri içerir.

3.9.5 Dikkat Mekanizması (Attention Mechanism)

Transformer mimarisinin temel özelliği olan dikkat mekanizması, metindeki kelimeler arasındaki ilişkileri öğrenme yeteneği sağlar. Dikkat mekanizması, özellikle kelime hizalaması ve anlam belirleme açısından önemlidir. Her bir

kodlayıcı ve çözücü katmanı içinde kullanılan bu mekanizma, belirli kelimelerin diğerlerine göre daha fazla dikkate alınmasını sağlar. Dikkat mekanizması, özellikle çeviri, metin sınıflandırma ve metin üretimi gibi görevlerde kullanılır. Dikkat mekanizmasının çalışma mantığı aşağıda detaylandırılmıştır.

Dikkat mekanizması, kodlayıcı-kod çözücü mimarisi içinde, kod çözücünün her bir çeviri adımında kodlayıcı tarafından üretilen bağlam vektörünün hangi kısımlarına odaklanacağını belirlemek için kullanılmaktadır. Önce, giriş metni üzerinde kodlayıcının çalışmasını gözden geçirelim:

Giriş metin: "Bir tarih dersi için önemli kişiler hakkında bilgi edinmek isterseniz, Albert Einstein ve Leonardo da Vinci gibi büyük figürleri inceleyebilirsiniz. Einstein, modern fizikteki devrimci çalışmalarıyla tanınırken Da Vinci, Rönesans döneminde sanat ve bilimde büyük bir etki yaratmıştır."

Encoder İşlemi:

- Giriş metin, kelime seviyesinde işlenir ve her kelimenin gömülü (embedding) temsilini alır.

- Encoder, bu temsilleri bir sıralı şekilde işler ve bağlam vektörünü üretir. Bağlam vektörü, giriş metindeki tüm bilgiyi kodlayan bir vektördür.

Dikkat mekanizması, kod çözücünün her çeviri adımında kullanılır:

Örnek kelime: "Albert":

- Kod çözücü, çeviriye başlar ve ilk kelime olarak "Albert" kelimesini oluşturur.

- Dikkat mekanizması devreye girer. Dikkat mekanizması, kodlayıcı tarafından üretilen bağlam vektörünü ve çeviri adımında üretilen kelimenin temsilini kullanır.

- Dikkat mekanizması, bu iki temsil arasındaki benzerliği hesaplar ve her kelimenin giriş metnindeki hangi kısımla daha fazla ilişkili olduğunu belirler.

- Örneğin, "Albert" kelimesi için dikkat mekanizması, "Albert Einstein" ifadesine daha fazla dikkat edilmesi gerektiğini belirleyebilir ve bağlam vektörünün bu kısmına odaklanır.

- Bu odaklanma, "Albert" kelimesinin daha iyi bir çeviri yapmasına yardımcı olur.

Örnek kelime: "tanınırken":

- Kod çözücü, ikinci kelimeyi oluşturur ve "tanınırken" kelimesini seçer.

- Dikkat mekanizması tekrar devreye girer ve yeni kelimenin temsili ile bağlam vektörü arasındaki ilişkiyi değerlendirir.

- Bu adımda, "tanınırken" kelimesi için dikkat mekanizması, önceki cümledeki "Einstein, modern fizikteki devrimci çalışmalarıyla" bölümüne daha fazla dikkat edilmesi gerektiğini belirleyebilir ve bağlam vektörünün bu kısmına odaklanır.

Bu şekilde, dikkat mekanizması her çeviri adımında hangi kısımlara odaklanılması gerektiğini belirler ve çeviri sürecinde daha iyi sonuçlar elde edilmesine yardımcı olur. Bu sayede, çeviri modeli giriş metnindeki önemli bağlantıları ve bilgileri doğru bir şekilde kullanabilir ve daha iyi çeviriler üretebilir.

3.9.6 Çok Başlıklı Dikkat (Multi-Head Attention)

Çok başlıklı dikkat, modelin farklı özellikleri ve ilişkileri daha iyi anlamasına yardımcı olur. Birden fazla dikkat başlığı kullanarak, metindeki farklı özelliklerin

ve ilişkilerin yakalanması hedeflenir. Her bir başlık, farklı bir öğrenme görevini üstlenir ve sonuçları birleştirilir.

3.9.7 Toplama ve Besleme İleri Ağları (Feed-Forward Neural Networks)

Her kodlayıcı ve çözücü katmanı içinde bulunan bu katmanlar, metindeki bilgileri daha da işler ve öğrenilmiş özellikleri geliştirir. Besleme ileri ağları, metindeki özelliklerin daha karmaşık bir şekilde temsil edilmesine yardımcı olur.

3.10 Özel Kimliklerin Belirlenmesi (Named Entity Recognition)

Özet dokümanın, orijinal doküman üzerinde doğrulanması için öncelikle her bir doküman üzerinde bilgi çıkarımının yapılması gerekmektedir. Bu amaçla kullanılan farklı bilgi çıkarımı teknikleri bulunmaktadır. Dokümanlar üzerinde gerçekleştirilen en etkin bilgi çıkarım yöntemlerinden biri, özel kimliklerin belirlenmesi işlemidir. Özel kimliklerin belirlenmesi, yer, organizasyon gibi önceden tanımlanmış kategorilerin metin dokümanları üzerinden çıkarılma işlemidir. Bilgi çıkarımının (Information Extraction) bir alt dalı olup makine çevirilerinden duygu analizine kadar birçok DDİ probleminde kullanılmaktadır. Tanımı ilk olarak 1995 yılında MUC-6 (Message Understanding Conference) konferansında yapılmıştır (Medium, 2023). Enamex, Numex ve Timex olmak üzere üç temel kategoride tanımlamalar yapılmaktadır. Bu kategoriler, metin içindeki belirli türdeki varlıkları (named entities) tanımak ve sınıflandırmak için kullanılmaktadır.

3.10.1 Enamex (Entity Names and Expressions)

Enamex, metin içindeki isimlendirilmiş varlıkları veya ifadeleri belirlemek ve bu varlıkları belirli kategorilere sınıflandırmak için kullanılan bir etiketleme sistemidir. Bu tür varlıklar genellikle kişiler, yerler, organizasyonlar, ürünler, kitaplar, film karakterleri gibi isimlendirilmiş varlıkları içerir. Enamex, metindeki varlıkları aşağıdaki kategori etiketlerine sınıflandırır:

- PER: Kişiler (Person)
- LOC: Yerler (Location)
- ORG: Organizasyonlar (Organization)
- MISC: Diğer çeşitli varlık türleri (Miscellaneous)

3.10.2 Numex (Numeric Expressions)

Numex, metindeki parasal ve yüzdesel ifadeleri tanımlamak ve sınıflandırmak için kullanılan bir etiketleme sistemidir. Bu tür ifadeler genellikle finansal raporlarda, fiyat listelerinde ve ekonomi metinlerinde bulunur. Numex, metindeki ifadeleri aşağıdaki kategori etiketlerine sınıflandırır:

- MONEY: Parasal ifadeler (örneğin, "$100" veya "2,5 milyon dolar")
- PERCENT: Yüzdesel ifadeler (örneğin, "%10" veya "yüzde 50")

3.10.3 Timex (Time Expressions)

Timex, metindeki tarih ve zaman ifadelerini tanımak ve sınıflandırmak için kullanılan bir etiketleme sistemidir. Bu tür ifadeler genellikle takvim tarihleri, saat dilimleri, yıl ve ay ifadeleri gibi zamansal bilgileri içerir. Timex, metindeki ifadeleri aşağıdaki kategori etiketlerine sınıflandırır:

- DATE: Tarih ifadeleri (örneğin, "01 Ocak 2023" veya "15 Temmuz")
- TIME: Saat ifadeleri (örneğin, "14:30" veya "öğleden sonra 3")
- DURATION: Süre ifadeleri (örneğin, "2 saat" veya "3 gün")

Bu etiketleme sistemleri, metin madenciliği, otomatik bilgi çıkarma ve duygu analizi gibi birçok uygulamada kullanılır. Metin içindeki bu özel varlık türlerini tanımlamak ve sınıflandırmak, metinlerden anlamlı bilgi çıkarmak ve analiz etmek için önemlidir. Özel kimliklerin belirlenmesi işlemi uygulanan dokümanlarda, Çizelge 3.3'te belirtilen özel kimlik türleri tespit edilmektedir.

Çizelge 3.3 Özel kimlik kategorileri (Spacy, 2023)

Tür	**Açıklama**
PERSON	İnsanların adları, soyadları veya özel isimleri. Örneğin: John Smith, Mary Johnson.
ORGANIZATION	Şirketler, kurumlar veya organizasyonların adları. Örneğin: Microsoft, NASA, United Nations.
LOCATION	Coğrafi yerlerin adları. Şehirler, ülkeler, dağlar, nehirler vb. Örneğin: Paris, Türkiye, Himalayalar, Nil Nehri.
DATE	Tarihleri ifade eden ifadeler. Örneğin: 1 Ocak 2023, 20. yüzyıl, Pazartesi.
TIME	Zamanı ifade eden ifadeler. Örneğin: 10:30, öğleden sonra, sabah.
MONEY	Para birimlerini ve miktarlarını ifade eden ifadeler. Örneğin: 1000 dolar, 50 euro.
PERCENT	Yüzdelik dilimleri ifade eden ifadeler. Örneğin: %10, %50, %100.
PRODUCT	Ürünlerin adları veya markaları. Örneğin: iPhone, Coca-Cola, Toyota.
EVENT	Önemli etkinliklerin veya olayların adları. Örneğin: Olimpiyatlar, Dünya Kupası, Festival.
WORK_OF_ART	Sanatsal eserlerin veya edebi eserlerin adları. Örneğin: Mona Lisa, Hamlet, Bohemian Rhapsody.

LANGUAGE	Dillerin adları. Örneğin: İngilizce, Fransızca, Mandarin.
LAW	Hukuki metinlerdeki yasaların veya düzenlemelerin adları. Örneğin: Anayasa, Ceza Kanunu.
QUANTITY	Miktarları ifade eden ifadeler. Örneğin: 5 litre, 10 kilogram, 3 metre.
CARDINAL	Sayıları ifade eden ifadeler. Örneğin: 1, 1000, beş.
ORDINAL	Sıralamayı ifade eden ifadeler. Örneğin: ilk, ikinci, üçüncü.
GPE (Geopolitical Entity)	Coğrafi bölge veya ülkelerin adları. Örneğin: ABD, Çin, Avrupa Birliği.
FACILITY	Tesislerin veya yapıların adları. Örneğin: hastane, okul, havaalanı.

Orijinal ve özet dokümanlara ait özel kimliklerin tespiti amacıyla kullanılan farklı özel kimlik belirleme algoritmaları (Azure Cognitive Service, Spacy, Stanford CoreNLP, NLTK, NLP-NER, MonkeyLearn) bulunmaktadır. Bu algoritmaların her biri Çizelge 3.3'te belirtilen özel kimlik türlerinden bazılarını tespit edebilmektedir.

Bu algoritmaların yanı sıra, bu çalışmada İngilizce dili üzerinde doküman doğrulama işlemi gerçekleştirileceğinden, İngilizce dilinde özel kimlik tespiti amacıyla yeni bir yaklaşım önerilmiştir. Önerilen özel kimlik tabanlı doküman doğrulama süreci daha sonraki bölümlerde detaylı olarak açıklanacaktır.

3.11 Yazım Denetleyicisi (Spell Checker) Süreci

DDİ alanında girdi verilerinin kalitesi, modelin performansının belirlenmesinde çok önemli bir rol oynamaktadır. Yazım hataları yinelenen bir sorun olarak öne çıktığından, metin verileri genellikle zorluklarla doludur. Bu tür hatalar özetleme modelinin etkinliğini olumsuz yönde etkileyebilir. Bu etkiyi azaltmak için

uzmanlar genellikle eğitim aşamasını başlatmadan önce veri seti üzerinde bir yazım denetiminin uygulanmasını savunmaktadır. Bu adımın gerçekleştirilmesi yalnızca veri setindeki yazım hatalarını temizlemekle kalmaz, aynı zamanda modelin performansını ve üretilen özetlerin okunabilirliğini artırır.

Veri ön işleme adımına tabi tutulmuş dokümanlar, doğrudan doğrulama sürecine alınmamalıdır. Çünkü özet oluşturma esnasında yazım veya anlam hataları oluşması olasıdır. Bu hatalar otomatik doğrulama sisteminin başarısı için önemlidir. Bu hataların varlığını kontrol etmek ve eğer hatalı ifadeler var ise bunları düzeltmek gerekmektedir. Aksi takdirde doküman doğrulama başarım oranını, özellikle finansal dokümanlarda olumsuz yönde etkiler. Sıklıkla yapılan yazım hataları aşağıdaki gibidir.

- Parasal/finansal kelimelerde bulunan yazım hatalarının özellikle finansal raporlarda yapılacak işlemlerde ciddi sorunlara yol açmaktadır.

- Bazı ülke isimlerinde (Örneğin, “The United States” yerine “The Unted States” gibi) yapılan hatalı yazımlar doküman özetleme ve doğrulama süreçlerinde hatalı model oluşturmaya neden olmaktadır.

Örneğin, aşağıda verilen örnek doküman içerisinde işaretlenmiş olan yazım hataları bulunmaktadır.

“Football and baseball have been lockd in a perpetual battle for the affection of sports in te Unted Sttes. Football has been the most popular in the United States with its bone-crushing hits and high-flying action. However, baseball has long been considered America's pastime, with its rich history of almost mythical characters and dramatic moments. Which sport is America's favorite? There is no other sport in the United States that garners so much attention during reular seaon games, and so much sadnes when it is gone. Every week, from Septemper to February, America eats, sleeps and breathes football. From NFL Sunday to Colege Football Saturday, the sport is being played somewhere at a high level on the wekend.

According to the latest Harris Poll, 33 percent of the people surveyed say football is their favorite sport. It's no surprise why. Some of the world's best athletes compete in football. No one is supposed to move with the speed and power that those monsrous human beings do. 350-poud men thunder down the field and crash into one another. As brutal as it is, fans revel in the violence, cheering on their warriors as they clash with other teams on the field of batle. Footbll is fast-paced, hard-hiting and exciting. Baseball has become bogged down with slow moving action, and archaic unwriten rules that deter from the on-field product."

Yazım hatalı ifadelerin düzeltilmesi amacıyla farklı algoritmalar (AutoCorrect, Azure Cognitive Service, NLTK Jaccard Distance, Pattern, RapidApi, Spacy Contextual, SpellChecker, Spello, TextBlob, WordNet) bulunmaktadır. Bu algoritmalar ile dokümanlarda hatalı yazım tespiti yapılmaktadır. 35 adet yazım hatası içeren bir veri setinde yapılan çalışmada bu algoritmaların elde ettiği başarım oranları Çizelge 3.4'te verilmiştir.

Çizelge 3.4 Yazım Denetleyici algoritmalarının başarım oranları

Yazım Denetleyicisi	**Hatalı Yazım Düzeltme Sayısı**	**Başarım Oranı (%)**	**Çalışma Süresi (sn)**
WordNet	30/35	0.86	12
Auto Correct	19/35	0.54	17
TextBlob	21/35	0.60	8
Spello	28/35	0.80	7
Speller	17/35	0.49	19
NLTK Jaccard Distance	26/35	0.74	9

Bu çalışmada, yukarıda belirtilen sorunların üstesinden gelmek için Tablo 1'deki sonuçlar dikkate alınarak WordNet Yazım Denetleyici algoritması kullanılmıştır. Bu algoritma ile belgelere ait cümleler yapısal olarak kontrol edilmiş, hatalı kelimeler düzeltilmiş ve özel isimler büyük harfle başlatılmıştır.

3.12 Anlamsal Benzerlik (Semantic Similarity)

Anlamsal benzerlik, metinlerin veya dil içeriğinin anlamsal olarak ne kadar benzer veya ilişkili olduğunu ölçen bir kavramdır. Bu ölçüm, metin madenciliği, metin sınıflandırma, özetleme, çeviri, öneri sistemleri ve daha birçok DDİ uygulamasında önemlidir. Anlamsal benzerlik, metinler arasındaki anlam ilişkilerini değerlendirir ve metinlerin veya metin içeriğinin anlamını analiz eder. Anlamsal benzerlik ölçümü için kullanılan temel bileşenler ve yöntemler aşağıdaki gibidir:

3.12.1 Kelime Gömme (Word Embeddings)

Kelime gömme, kelimeleri sayısal vektörlerle temsil etmek için kullanılan bir tekniktir. Bu vektörler, kelimenin anlamını yansıtır. Örneğin, "kral" ve "kraliçe" kelimeleri benzer anlamlara sahipse, bu kelimelerin gömme vektörleri de benzer olur.

3.12.2 Dil Modelleri

Dil Modelleri: Büyük metin koleksiyonlarında eğitilen dil modelleri (örneğin, BERT, Transformer, GPT-3), metinlerin anlamını daha derinlemesine analiz edebilir. Bu modeller, kelime düzeyinden daha fazla bağlamı ve anlamı anlarlar.

3.12.3 Ölçü (Metric) Tabanlı Yaklaşımlar

Ölçü tabanlı yöntemler, metinlerin veya metin içeriğinin benzerliğini hesaplamak için metrikler kullanır. Örneğin, kosinüs benzerliği, Jaccard benzerliği, Simhash veya Cross-Encoder gibi ölçüler, metinlerin vektör temsilleri arasındaki benzerliği ölçmek için kullanılabilir.

3.12.4 Derin Öğrenme Modelleri

Derin öğrenme modelleri, metinler arasındaki anlamsal benzerliği hesaplamak için kullanılabilir. Özellikle, çift metin sınıflandırma görevlerinde (örneğin, iki metnin aynı anlam taşıyıp taşımadığını belirleme) başarılıdır. Anlamsal benzerlik ölçümü, metinler arasındaki anlamsal ilişkileri yakalamak ve metinleri sınıflandırmak, karşılaştırmak veya özetlemek gibi birçok uygulamada kullanılır. Örneğin, bir metin özetleme sistemi, metinler arasındaki cümlelerin anlamsal benzerliğini değerlendirerek en önemli cümleleri seçebilir. Ayrıca, bir çeviri sistemi, iki dil arasındaki anlamsal benzerliği kullanarak daha iyi çeviriler üretebilir. Bu nedenle, anlamsal benzerlik ölçümü, metin analizi ve DDİ alanında temel bir bileşendir ve metinlerin anlamını daha iyi anlamak ve analiz etmek için kullanılır.

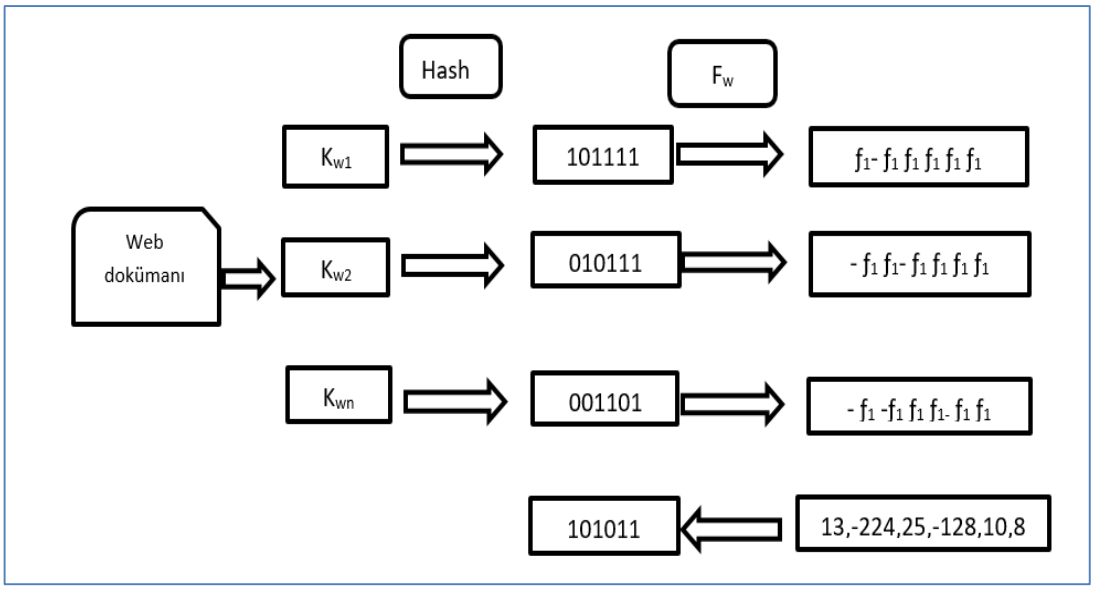
Özet dokümana ait cümlelerin, orijinal tam metin doküman içerisinde hangi cümlelere karşılık geldiğinin tespit edilmesi doküman doğrulama süreci için oldukça önemlidir. Çünkü orijinal özet doküman içerisinde bulunan cümlelerin içerdiği özel kimlikler ile orijinal tam metin dokümanda bu özet cümleleriyle benzer cümlelerin içerdiği özel kimlikler kıyaslanarak belgenin doğrulanması sağlanmaktadır. Eğer orijinal özet dokümana ait cümlede geçen bir özel kimlik verisi, orijinal tam metin dokümandaki benzer cümlede geçmiyorsa, ilgili özet dokümanın doğrulama başarım oranı düşük anlamına gelmektedir. Tam metin dokümanda benzer cümlelerin tespiti amacıyla Simhash ve Cross-Encoder metin benzerlik algoritmaları ayrı ayrı olarak kullanılmış ve deneysel sonuçları değerlendirilmiştir.

Simhash algoritması, özellikle metin işlemenin yoğun olduğu, arama motoru gibi uygulamalarda dosyaların veya web sitelerinin birbirine olan benzerliğini bulmak için kullanılan bir algoritmadır. Simhash algoritması, iki dosyayı birer vektör olarak görür ve bu vektörler (Yöney, Vektör) arasındaki kosinüs (Cosine) bağlantısını bulmaya çalışır. Simhash algoritması, benzerlik tespiti amacıyla birçok farklı çalışmada kullanılmıştır. Jiang (Jiang vd., 2011) ve Pi (Pi vd., 2019) çalışmalarında, doküman benzerliğini elde etmek amacıyla Simhash algoritması

kullanılmıştır. Şekil 3.6'da Simhash algoritmasının sözde kodu verilmiş, Şekil 3.7'de ise Simhash algoritmasının uygulanışı görülmektedir.

```
(1) function Simhash (document D)
(2) {
(3)    Init vector Sim[0, . . . , (f − 1)] = 0;
(4)    for each Keyword Kw in D
(5)    {
(6)       Kw is hahsed into f bit Hash value X;
(7)          for (i = 0; i < f; i ++)
(8)          {
(9)             if (X == 1)
(10)            {
(11)               Sim[i] = Sim[i] + weight(Kw);
(12)            } else {
(13)               Sim[i] = Sim[i] − weight(Kw);
(14)            } //end if
(15)         } //end for
(16)   }//end for
(17)   for (i = 0; i < f; i ++)
(18)   {
(19)      if (Sim[i] > 0) Sim[i] 1;
(20)      else Sim[i] = 0;
(21)   } //end for
(22) }
```

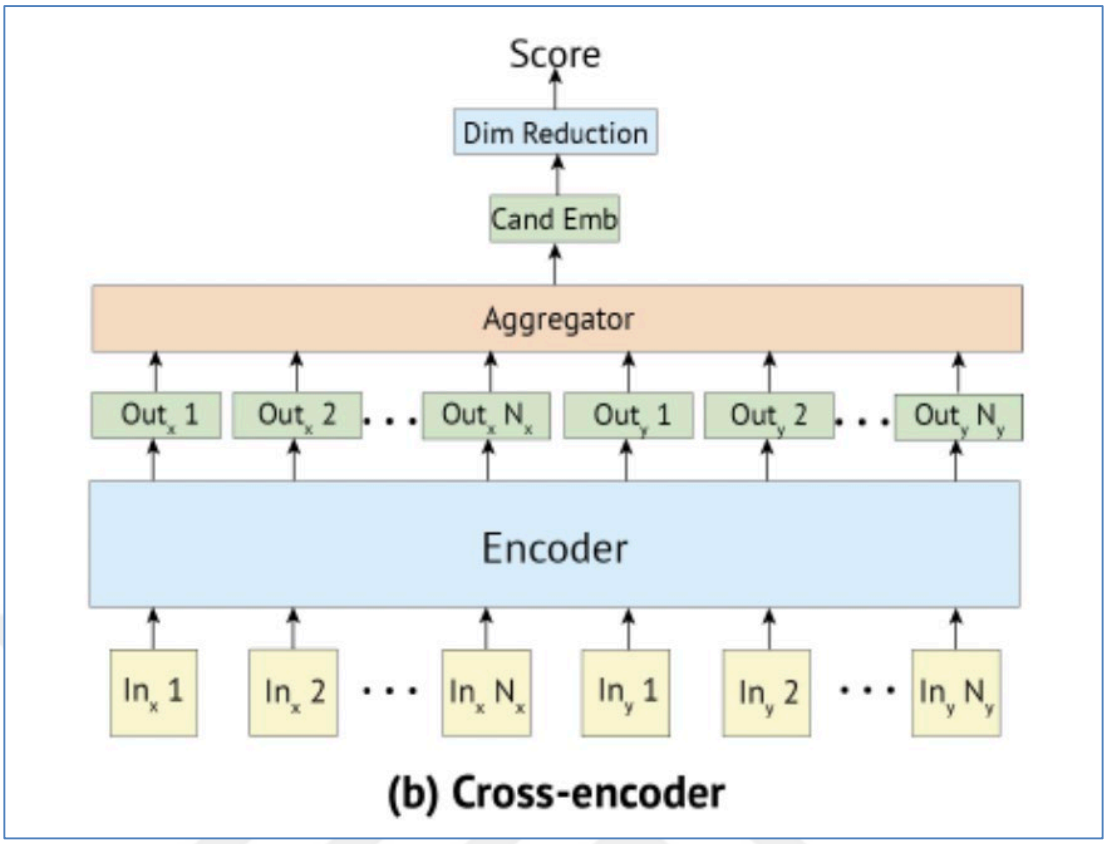
Şekil 3.6 Simhash algoritmasının sözde kodu (Ho vd., 2014)



Şekil 3.7 Simhash algoritmasının uygulanması

Simhash algoritmasının uygulanması şu şekilde özetlenebilir: Eğer doküman içinde yer alan kelime Kw'nin hash değeri X'in (f uzunluğunda) ilgili bit değeri 1 ise Sim vektöründeki ilgili benzerlik değeri o kelimenin (Kw) ağırlığı kadar artırılır, 0 ise o kelimenin (Kw) ağırlığı kadar azaltılır. Nihai olarak, Sim vektöründeki değerler sıfırdan büyükse 1'e, küçükse 0'a dönüştürülür.

Cross Encoder anlamsal benzerlik algoritması, iki metin parçası arasındaki anlamsal benzerliği ölçmek için kullanılan Cross Encoder mimarisinin özel bir uygulamasıdır. Metin benzerliği, açıklama tespiti ve kopya tespiti gibi DDİ görevlerinde yaygın olarak kullanılır (Weaviate, 2023). Algoritma, iki giriş metnini alıp bunları ayrı kodlama ağları kullanarak sabit uzunluktaki vektörlere kodlayarak çalışır. Bu kodlanmış vektörler daha sonra birleştirilir ve anlamsal içeriklerine göre iki girdi arasındaki benzerlik puanını hesaplayan bir çapraz dikkat katmanına beslenir. Bu benzerlik puanı daha sonra iki metnin benzer mi yoksa farklı mı olduğunu tahmin etmek için kullanılmaktadır. Şekil 3.8'de Cross-Encoder algoritmasının çalışma prosedürü yer almaktadır.



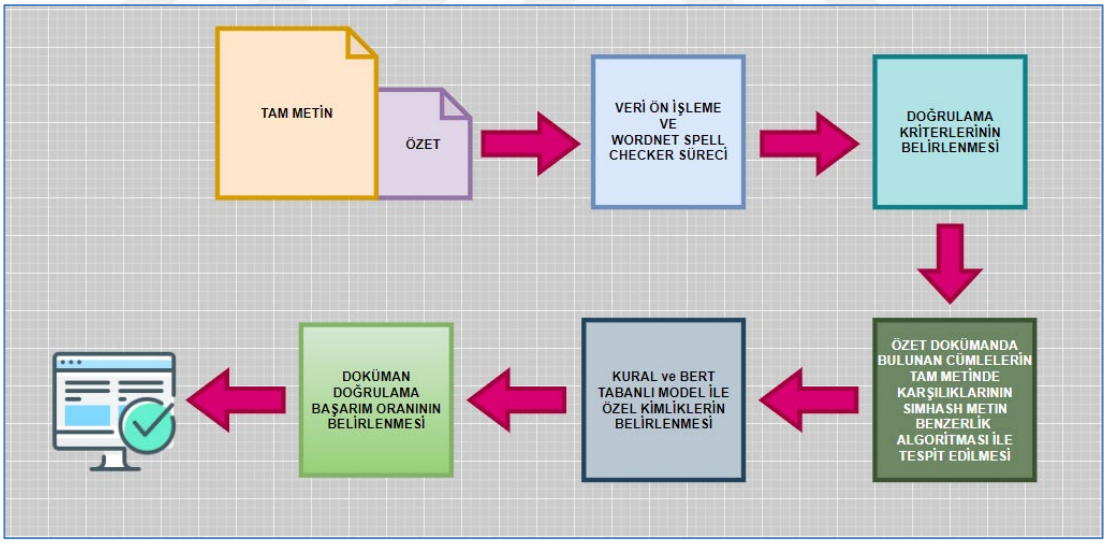
Şekil 3.8 Cross-Encoder algoritmasının çalışma prosedürü

4. OTOMATİK DOKÜMAN DOĞRULAMA SÜRECİ

Tasarlanacak DDİ tabanlı otomatik doküman doğrulama altyapısı üç bölümden oluşmaktadır. İlk olarak anlamsal analiz olmadan özel kimlik tespiti vasıtasıyla doküman doğrulama, ikinci olarak Transformer mimarisi kullanılarak anlamsal doküman doğrulama ve son olarak tüm bu sürecin otomatik hale getirilmesi amacıyla sisteme verilen orijinal tam metnin soyut özetinin sistem tarafından üretilmesi sağlayan özetleme aşamasıdır. Bu aşamalar aşağıdaki bölümlerde detaylı olarak ele alınmıştır.

4.1 Özel Kimlik Tespiti Yöntemi Vasıtasıyla Otomatik Doküman Doğrulama

Bu çalışmada önerilen özel kimlik tespiti yönteminin aşamaları Şekil 4.1'de verilmiştir.



Şekil 4.1 Özel kimlik tespiti yöntemi vasıtasıyla doküman doğrulama sürecinin aşamaları

Bu aşamaların detayı aşağıdaki gibidir.

1. Doğrulanması istenen doküman ve belge setinin temini.

2. Dokümanların standart hale getirilmesi.

3. Orijinal ve özet dokümanların Spell Checker sürecine tabi tutulması.

4. Üçüncü aşamanın her bir sonucu için;

    a. Dokümanların doğrulama kriterlerinin alınması

    b. Hem orijinal belge seti hem de özet dokümandan bilgi çıkarımının yapılması (Özel Kimliklerin Belirlenmesi).

    c. Belge seti ile özet dokümanın karşılaştırılması.

    d. Doğrulama kriterleri baz alınarak elde edilen doküman doğrulama başarım oranı üzerinden bir çıktı üretilmesi.

4.1.1 Özel Kimlik Tespiti

Öncelikle, Kaggle (Kaggle, 2023) ortamından alınan ve Çizelge 4.1 ve Çizelge 4.2'de detayları verilen veri seti kullanılarak eğitilen sınıflandırma tabanlı bir özel kimlik tespit modeli oluşturulmuştur.

Çizelge 4.1 Özel kimlik tespiti amacıyla kullanılan eğitim veri seti metrikleri

Cümle Sayısı	**Kelime Sayısı**	**Özel Kimlik (Named Entity) Sayısı**
47959	1048575	8

Çizelge 4.2 Özel kimlik tespiti amacıyla kullanılan eğitim veri setinin bir kesiti

Cümle No	**Kelime**	**Etiket**
1	Thousands	O
1	of	O
1	demonstrators	O

1	have	O
1	marched	O
1	through	O
1	London	B-geo
1	to	O
1	protest	O
1	the	O
1	war	O
1	in	O
1	Iraq	B-geo
1	and	O
1	demand	O
1	the	O
1	withdrawal	O
1	of	O
1	British	B-gpe
1	troops	O
1	from	O
1	that	O
1	country	O

Sınıflandırma modelinin eğitimi için kullanılan parametre değerleri Çizelge 4.3'te verilmiştir.

Çizelge 4.3 Sınıflandırma modelinin eğitimi için kullanılan parametre değerleri

Parametre Adı	**Parametre Değeri**
num_train_epochs	16
learning_rate	1e-4
overwrite_output_dir	True
train_batch_size	32
eval_batch_size	32

model_type	Bert
model_name	Bert-base-cased
use_cuda	False
validation_split	0.2
drop_out	0.1

İlgili veri seti, bazı özel kimlik türlerini (parasal, e-posta, yüzde) içermemektedir. Çalışmamızda finansal dokümanları doğrulamamız gerektiğinden, bu türlerin tespit edilememesi, doküman doğrulama başarım oranını etkileyecektir. Bu nedenle, belirtilen özel kimlik türleri için Regex kural tabanlı özel kimlik tespit algoritması yazılmıştır. Yazılan Regex kurallarından biri aşağıdaki gibidir.

["TEXT": "REGEX": "[Uu](.?|nited)" or ["ORTH": "Google", "ORTH": "I", "ORTH": "/", "ORTH": "O"])

Hem sınıflandırma modeli hem de Regex kuralları ile tespit edilen özel kimlikler ile tam metin ve özet doküman doğrulaması yapılmaktadır. Hem tam metin hem de özet doküman için cümle bazında özel kimlikler belirlendiğinde doğrulama başarısı hesaplanabilmektedir. Basit doğrulama için, özet dokümanda tanımlanan her özel kimlik, orijinal belgede aranır. Daha sonra doğrulama başarısı, orijinal dokümanda eşleşen özel kimliklerin sayısının özette bulunan özel kimliklerin sayısına bölünmesiyle hesaplanır. Orijinal doküman/dokümanlarda aynı isim ve türde iki özel kimliğin mevcut olması durumunda doğru cümleyi bulmak için şu anda herhangi bir anlamsal kontrol uygulanmamaktadır; bu elbette özellikle uzun belgeler (farklı cümlelerde birden fazla kez tekrarlanan varlık) için olası bir durumdur.

Doğrulama başarısı formülü Formül 4.1'deki şekilde tanımlanır:

P = Doğruluk (%)

Sm = Orijinal dokümanla eşleşen özel kimliklerin sayısı.

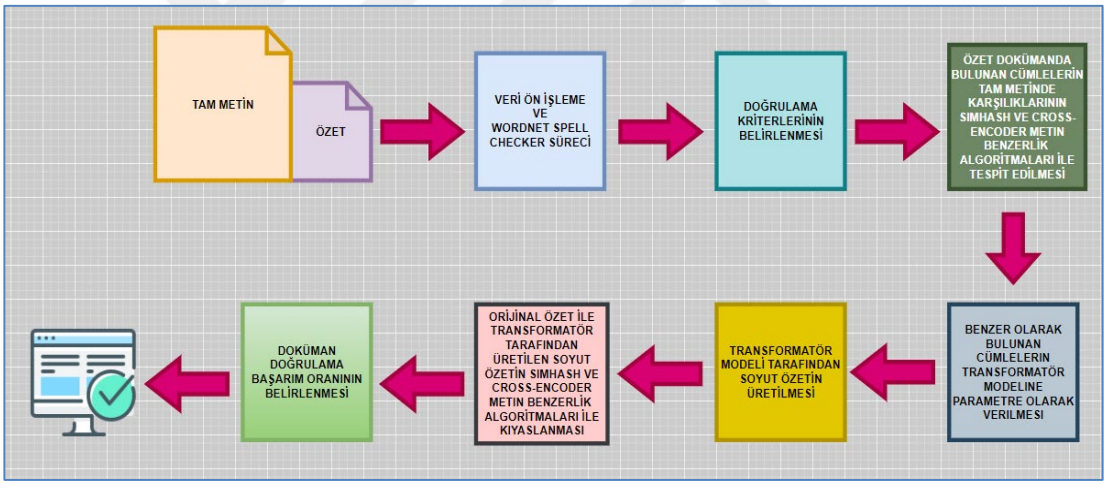
St = Özetteki özel kimliklerin sayısı.

$$\boldsymbol{P}\ (\%) = \frac{\mathbf{Sm}}{\mathbf{St}} \tag{4.1}$$

Özel kimlik tespiti vasıtasıyla doküman doğrulama sürecine ilişkin süreç bir sonraki bölümde uçtan-uca ele alınmıştır.

4.2 Transformer Mimarisi Vasıtasıyla Otomatik Anlamsal (Semantik) Doküman Doğrulama

Bu çalışmada semantik analiz görevi için Transformer mimarisinden yararlanılmıştır. Önerilen semantik analiz yaklaşımı Şekil 4.2'de verilen aşamalardan oluşmaktadır.



Şekil 4.2 Transformer mimarisi vasıtasıyla doküman doğrulama sürecinin aşamaları

Bu aşamaların detayı aşağıdaki gibidir.

1. Doğrulanması istenen doküman ve belge setinin temini
2. Dokümanların standart hale getirilmesi
3. Orijinal ve özet dokümanların Spell Checker sürecine tabi tutulması
4. Üçüncü aşamanın her bir sonucu için;

a. Orijinal özet dokümanda bulunan her bir cümlenin, orijinal tam metin dokümana ait tüm cümleler ile benzerlikleri bakımından kıyaslanması ve en benzer tam metin cümlelerinin tespit edilmesi.

b. Tespit edilen orijinal tam metin doküman cümlelerinin oluşturulan Transformer modeline verilmesi ve bu cümlelerin soyut özetlerinin üretilmesi.

c. Üretilen soyut cümle özetlerinin birleştirilmesi ve Transformer özet dokümanın elde edilmesi.

d. Son aşamada, üretilen Transformer özeti ile orijinal özet dokümanın özel kimlik ve metin benzerlikleri bakımından anlamsal doğrulamasının yapılması.

Bu bölümde önerilen otomatik belge doğrulama modeli aşağıda ayrıntılı olarak anlatılmaktadır. Esasen Transformer tabanlı bir anlamsal belge doğrulama yaklaşımıdır. Bu yaklaşımın ayrıntılı adımları aşağıdaki algoritma akışında sunulmaktadır.

Algoritma 1 Önerilen Model

procedure TextSummarization (originalText, originalSummary)
standardizedText ← STANDARDIZE (originalText)
standardizedSummary ← STANDARDIZE (originalSummary)
standardizedText ← SPELLCHECK (standardizedText)
standardizedSummary ← SPELLCHECK (standardizedSummary)
groupsOf SimilarSentences ← [ ]
for each sentence in standardizedSummary **do**
similarSentences ← COMPAREWITHFULLTEXT (sentence, standardizedText)
groupsOfSimilarSentences.append (similarSentences)
end for
abstractSummaries ← [ ]

```
        for each group in groupsOfSimilarSentences do
            abstractSummary ← TRANSFORMERMODEL (group)
            abstractSummaries.append (abstractSummary)
        end for
        transformerSummary ← COMBINESUMMARIES
        (abstractSummaries)
        similarityScore ← EVALUATESIMILARITY
        (transformerSummary, originalSummary)
        return similarityScore
end procedure
```

```
function STANDARDIZE (document)
        Metni küçük harfe dönüştürün, özel karakterleri kaldırın vb.
    return standardizedDocument
end function
function SPELLCHECK (document)
        Yazım hatalarını kontrol edin ve düzeltin.
    return correctedDocument
end function
function COMPAREWITHFULLTEXT (sentence, fullText)
        Cümleyi tam metindeki tüm cümlelerle karşılaştırın.
    return similarSentences
end function
function TRANSFORMERMODEL (groupOfSentences)
        Cümle grubunu Transformer modeline besleyin.
    return summary
end function
function COMBINESUMMARIES (summaries)
        Tüm özetleri tek bir belgede birleştirin.
    return combinedSummary
end function
function EVALUATESIMILARITY (summary1, summary2)
        İki özet arasındaki benzerlik puanını hesaplayın.
    return similarityScore
```

end function

4.2.1 Transformer Modelinin Oluşturulması

Bu çalışmada hem doküman doğrulama hem de otomatik soyut özetleme aşamasında Transformer mimarisinden faydalanılmıştır. Çalışmada hedef doküman türü finansal türe ait olduğundan, Transformer modelinin bu amaca yönelik oluşturulması gerekmektedir. Bu nedenle modelin finansal türe ait dokümanlar ile eğitilmesi gerekmektedir. Transformer modelini eğitmek için Bölüm 3.4'te detayları verilen Reuters veri seti kullanılmıştır. Reuters veri setine ait örnek tam metin ve bu tam metin dokümana ait özet doküman Çizelge 4.4'te verilmiştir.

Çizelge 4.4 Reuters veri setine ait örnek tam metin ve bu tam metin dokümana ait özet doküman

Orijinal Tam Metin Doküman	**Orijinal Özet Doküman**

Ask Jeeves has become the third leading online search firm this week to thank a revival in internet advertising for improving fortunes. The firm's revenue nearly tripled in the fourth quarter of 2004, exceeding $86m (£46m). Ask Jeeves, once among the best-known names on the web, is now a relatively modest player. Its $17m profit for the quarter was dwarfed by the $204m announced by rival Google earlier in the week. During the same quarter, Yahoo earned $187m, again tipping a resurgence in online advertising. The trend has taken hold relatively quickly. Late last year, marketing company Doubleclick, one of the leading providers of online advertising, warned that some or all of its business would have to be put up for sale. But on Thursday, it announced that a sharp turnaround had brought about an unexpected increase in profits. Neither Ask Jeeves nor Doubleclick thrilled investors with their profit news, however. In both cases, their shares fell by some 4%. Analysts attributed the falls to excessive expectations in some quarters, fuelled by the dramatic outperformance of Google on Tuesday.	Ask Jeeves has become the third leading online search firm this week to thank a revival in internet advertising for improving fortunes. Its $17m profit for the quarter was dwarfed by the $204m announced by rival Google earlier in the week. During the same quarter, Yahoo earned $187m, again tipping a resurgence in online advertising. Neither Ask Jeeves nor Doubleclick thrilled investors with their profit news, however.Ask Jeeves, once among the best-known names on the web, is now a relatively modest player.

Transformer modelinin eğitimi için kullanılan parametre değerleri Çizelge 4.5'te verilmiştir.

Çizelge 4.5 Transformer modelinin eğitimi için kullanılan parametre değerleri

Parametre Adı	**Parametre Açıklaması**	**Parametre Değeri**
Öz-dikkat kafalarının sayısı	Modelin çok kafalı dikkat mekanizmasında kullanılacak dikkat kafası sayısını temsil etmektedir.	8

Doğrusal olarak yansıtılan sorguların ve anahtarların boyutluluğu	Doğrusal projeksiyon sonrasında sorgu ve anahtar vektörlerin boyutluluğu temsil etmektedir. Bu genellikle modelin gömme boyutu veya gizli boyutuyla aynı olacak şekilde ayarlanır.	64
Doğrusal olarak yansıtılan değerlerin boyutluluğu	Doğrusal izdüşümü sonrasında değer vektörlerinin boyutluluğunu temsil etmektedir. Bu genellikle modelin gömme boyutu veya gizli boyutuyla aynı olacak şekilde ayarlanır.	64
Model katmanlarının çıktılarının boyutluluğu	Model katmanlarının çıktı vektörlerinin boyutluluğunu temsil etmektedir. Bu, modelin özel mimarisine bağlı olarak değişebilir.	512
Tamamen bağlı iç katmanın boyutluluğu	Transformer kodlayıcı ve kod çözücü bloklarında kullanılan ileri beslemeli ağdaki ara katmanın boyutluluğunu temsil etmektedir.	2048
Kodlayıcı ve kod çözücü yığınındaki katman sayısı	Modeldeki kodlayıcı ve kod çözücü katmanlarının sayısını temsil etmektedir. Bu, modelin özel mimarisine bağlı olarak değişebilir.	6
Eğitim tur (Epoch) sayısı	Eğitim süreci sırasında tüm eğitim veri kümesinin modelden geçme sayısıdır.	2
Parti boyutu (Batch size)	Eğitim sırasında bir ileri/geri geçişte işlenen eğitim örneklerinin sayısını temsil etmektedir.	64

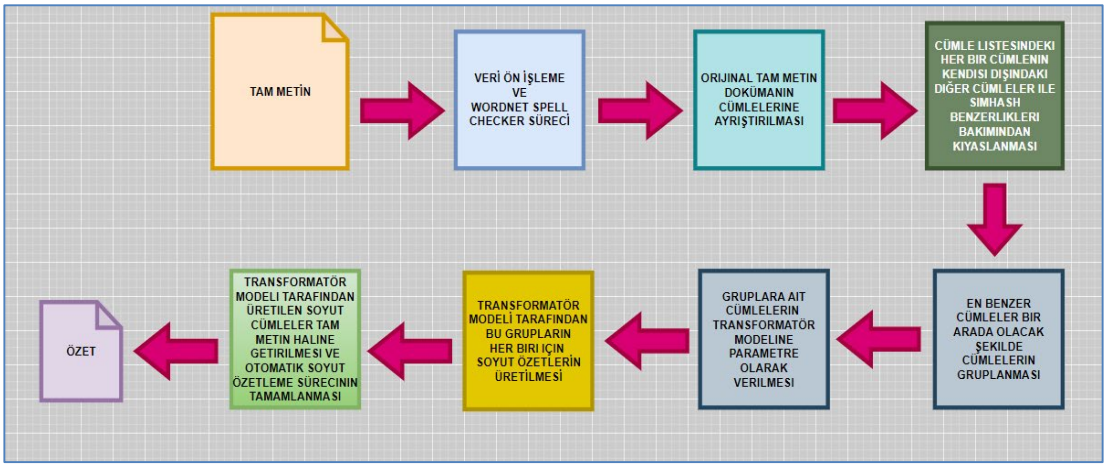
Sorgu sayısı (Number of queries)	Modelin çok başlıklı dikkat mekanizmasında kullanılan sorgu vektörlerinin sayısını temsil etmektedir.	128
Öz dikkat sayısı (Number of self-attentions)	Modelde öz-dikkat mekanizmasının uygulanma sayısını temsil etmektedir.	8
Öğrenme oranı (Learning rate)	Eğitim süreci sırasında model ağırlıklarının güncellenme hızını temsil etmektedir.	1e-9
Sönümleme (Dropout)	Aşırı uyumun önlenmesine yardımcı olan, eğitim sırasında modeldeki bazı nöronları rastgele bırakan (sıfıra ayarlayan) bir düzenleme tekniğidir. Sönümleme oranı, bırakılacak nöronların oranını belirlemektedir.	0,1

Eğitim sırasında, giriş metninin özünü yakalamayan özetler üretmesi nedeniyle modeli cezalandırmak için ikili çapraz entropi kaybı fonksiyonu kullanılmıştır. Geri yayılım sırasında hesaplanan gradyanlara dayalı olarak modelin parametrelerini güncellemek için Adam optimizasyon fonksiyonu uygulanmıştır. Veri seti sırasıyla %80 ve %20 oranında eğitim ve doğrulama olarak ayrılmıştır. Veri kümesindeki her doküman, durdurma sözcüklerini (stop words) ve alfa sayısal olmayan karakterleri kaldırmak için veri ön işleme sürecine tabi tutulmuş ve elde edilen nihai doküman, önceden eğitilmiş bir sözcüklere ayırma modeli (tokenizer) kullanılarak bir tamsayı dizisine dönüştürülmüştür. Eğitimin her aşamasında (epoch), eğitim veri kümesinden orijinal tam metin ve bunların özet grupları rastgele örneklenmiştir. Daha sonra dokümanlar kodlayıcı (encoder) aracılığıyla modele beslenmiş ve kod çözücü (decoder) soyut özet oluşturulması için kullanılmıştır. Oluşturulan özet, ikili çapraz entropi kaybı fonksiyonu kullanılarak orijinal özetiyle karşılaştırılmış ve gradyanlar, parametrelerini güncellemek için model boyunca geriye doğru yayılmıştır. Öğrenme oranı 1e-9

olarak başlatılmış ve doğrulama kaybının art arda 3 dönem boyunca iyileşmemesi durumunda 10 kat azaltılmıştır. Ayrıca aşırı öğrenmeyi (overfitting) önlemek için 0,1'lik bir sönümleme oranı (dropout) kullanılmıştır. Model katmanlarının çıktı boyutu, modelin yapısına ve belirli kullanım durumlarına bağlı olarak değişebilir ancak modelin daha karmaşık ilişkileri öğrenmesine yardımcı olabileceğinden genellikle daha büyük boyutlar tercih edilmektedir. Bu çalışmada model katmanlarının çıktı boyutu 512 olarak ayarlanmıştır. Ayrıca grup büyüklüğü de (Batch size) özellikle büyük veri kümeleri üzerinde eğitimde önemli bir faktördür. İyi performans elde etmek için doğru büyüklüğü (örneklem sayısı) seçmek gereklidir. Burada 64'lük bir grup büyüklüğü seçilmiştir. Çünkü daha büyük grup büyüklüğü eğitim süresini ve bellek kullanımını artırabilirken, daha küçük grup büyüklüğü eğitim süresini artırabilmektedir.

4.3 Transformer Mimarisi Vasıtasıyla Otomatik Soyut Özetleme

Bu çalışmada, Bölüm 4.2.1'de önerilen Transformer modeli kullanılarak, parametre olarak alınan orijinal bir tam metin dokümanın soyut özetinin üretilmesi sağlanmıştır. Önerilen Transformer mimarisi vasıtasıyla otomatik soyut özetleme yaklaşımı Şekil 4.3'te verilen aşamalardan oluşmaktadır.



Şekil 4.3 Transformer mimarisi vasıtasıyla otomatik soyut özetleme sürecinin aşamaları

Bu aşamaların detayı aşağıdaki gibidir.

1. Soyut özetinin üretilmesi istenen orijinal tam metin doküman setinin temini.
2. Dokümanların standart hale getirilmesi.

3. Orijinal tam metin dokümanların Spell Checker sürecine tabi tutulması

4. Üçüncü aşamanın her bir sonucu için;

    a. Orijinal tam metin dokümanın cümlelerine ayrıştırılması.

    b. Cümle listesindeki her bir cümlenin kendisi dışındaki diğer tüm cümleler ile benzerlikleri bakımından kıyaslanması ve en benzer tam metin cümlelerinin tespit edilmesi.

    c. Deneysel olarak belirlenen grup sayısı parametresine göre, en benzer tam metin cümlelerin gruplandırılması.

    d. Her bir grup cümle listesinin Transformer modeline parametre olarak verilmesi ve Transformer modeli tarafından her bir grup için soyut cümle özetinin üretilmesi.

    e. Üretilen soyut cümle özetlerinin birleştirilmesi ve Transformer soyut özet dokümanının elde edilerek sürecin sonlandırılması.

Bu aşamaların her biri bir sonraki uçtan uca entegrasyon sürecinde detaylandırılacaktır.

5. UÇTAN UCA DOKÜMAN DOĞRULAMA ENTEGRASYONU

Bu bölümde 4. Bölümde önerilen doküman doğrulama ve soyut doküman özetleme süreçleri bir örnek doküman seti üzerinde ele alınmıştır.

5.1.1 Özel Kimlik Tespiti Yöntemi Vasıtasıyla Otomatik Doküman Doğrulama

Bu çalışmada, Reuters veri setindeki tam metin ve bu metinlere ait özet dokümanların doğrulanması sağlanmaktadır. Öncelikle doğrulacak özet dokümandaki kelimelerin orijinal tam metin dokümandaki hangi cümleler ile doğrulanacağının belirlenmesi gerekmektedir. Çizelge 5.1'de doğrulama yapılacak özet ve özete ait tam metin doküman örneği yer almaktadır.

Çizelge 5.1 Doğrulama yapılacak özet ve özete ait tam metin doküman

Orijinal Tam Metin Doküman	**Orijinal Özet Doküman**

For the second time during his papacy, Pope Francis has announced a new group of bishops and archbishops set to become cardinals -- and they come from all over the world. Pope Francis said Sunday that he would hold a meeting of cardinals on February 14 "during which I will name 15 new Cardinals who, coming from 13 countries from every continent, manifest the indissoluble links between the Church of Rome and the particular Churches present in the world," according to Vatican Radio. New cardinals are always important because they set the tone in the church and also elect the next pope, CNN Senior Vatican Analyst John L. Allen said. They are sometimes referred to as the princes of the Catholic Church. The new cardinals come from countries such as Ethiopia, New Zealand and Myanmar. "This is a pope who very much wants to reach out to people on the margins, and you clearly see that in this set," Allen said. "You're talking about cardinals from typically overlooked places, like Cape Verde, the Pacific island of Tonga, Panama, Thailand, Uruguay." But for the second time since Francis' election, no Americans made the list. "Francis' pattern is very clear: He wants to go to the geographical peripheries rather than places that are already top-heavy with cardinals," Allen said. Christopher Bellitto, a professor of church history at Kean University in New Jersey, noted that Francis announced his new slate of cardinals on the Catholic Feast of the Epiphany, which commemorates the visit of the Magi to Jesus' birthplace in Bethlehem. "On feast of three wise men from far away, the	The 15 new cardinals will be installed on February 14. They come from countries such as Myanmar and Tonga. No Americans made the list this time or the previous time in Francis' papacy

Pope's choices for cardinal say that every local church deserves a place at the big table." In other words, Francis wants a more decentralized church and wants to hear reform ideas from small communities that sit far from Catholicism's power centers, Bellitto said. That doesn't mean Francis is the first pontiff to appoint cardinals from the developing world, though. Beginning in the 1920s, an increasing number of Latin American churchmen were named cardinals, and in the 1960s, St. John XXIII, whom Francis canonized last year, appointed the first cardinals from Japan, the Philippines and Africa. In addition to the 15 new cardinals Francis named on Sunday, five retired archbishops and bishops will also be honored as cardinals. Last year, Pope Francis appointed 19 new cardinals, including bishops from Haiti and Burkina Faso.	

Orijinal özet dokümanda bulunan her bir cümle, orijinal tam metin dokümana ait tüm cümleler ile Simhash metin benzerlik algoritması kullanılarak benzerlikleri bakımından kıyaslanmış ve en benzer olan iki cümle seçilmiştir. İki cümle seçilmesinin nedeni, özetin maksimum %50 olacağı varsayımına dayalı olarak, özetteki her cümlenin orijinal tam metinde 2 cümle ile eşleşmesi öngörülmüştür. Daha kısa özetlerde bu cümle seçimlerinin daha seçici olacağı, uzun özetlerde sapmalara sebep vereceği öngörülmektedir. Çizelge 5.1'de verilen özet ve orijinal tam metin doküman için benzer olarak seçilen cümleler Çizelge 5.2'de verilmiştir.

Çizelge 5.2 Özet dokümanda bulunan cümlelere en benzer orijinal tam metin doküman cümleleri

Orijinal Özet Doküman Cümle	**En Benzer 2 Cümle**

The 15 new cardinals will be installed on February 14.	**1-** Pope Francis said Sunday that he would hold a meeting of cardinals on February 14 "during which I will name 15 new Cardinals who, coming from 13 countries from every continent, manifest the indissoluble links between the Church of Rome and the particular Churches present in the world," according to Vatican Radio. **2-** In addition to the 15 new cardinals Francis named on Sunday, five retired archbishops and bishops will also be honored as cardinals.
They come from countries such as Myanmar and Tonga.	**1-** The new cardinals come from countries such as Ethiopia, New Zealand and Myanmar. **2-** Pope Francis said Sunday that he would hold a meeting of cardinals on February 14 "during which I will name 15 new Cardinals who, coming from 13 countries from every continent, manifest the indissoluble links between the Church of Rome and the particular Churches present in the world," according to Vatican Radio.
No Americans made the list this time or the previous time in Francis' papacy	**1-** But for the second time since Francis' election, no Americans made the list. **2-** For the second time during his papacy, Pope Francis has announced a new group of bishops and archbishops set to become cardinals -- and they come from all over the world.

Özet doküman cümlelerine ait tespit edilen özel kimlikler Çizelge 5.3'teki gibidir.

Çizelge 5.3 Özet doküman cümlelerine ait tespit edilen özel kimlikler

Özet Doküman Cümle	**Özel Kimlik Değeri**
The 15 new cardinals will be installed on February 14.	15 February 14
They come from countries such as Myanmar and Tonga.	Myanmar Tonga
No Americans made the list this time or the previous time in Francis' papacy.	Americans Francis

Özet doküman cümlelerine en benzer orijinal tam metin doküman cümleleri için tespit edilen özel kimlikler Çizelge 5.4'teki gibidir.

Çizelge 5.4 Benzer tam metin doküman cümlelerine ait özel kimlikler

Tam Metin Cümle	**Özel Kimlik Değeri**
Pope Francis said Sunday that he would hold a meeting of cardinals on February 14 "during which I will name 15 new Cardinals who, coming from 13 countries from every continent, manifest the indissoluble links between the Church of Rome and the particular Churches present in the world," according to Vatican Radio.	Pope Francis Sunday February 14 15 13 Churches Vatican Radio
In addition to the 15 new cardinals Francis named on Sunday, five retired archbishops and bishops will also be honored as cardinals.	15 Francis Sunday five
For the second time during his papacy, Pope Francis has announced a new group of bishops and archbishops set to become cardinals -- and they come from all over the world.	Second Pope Francis

But for the second time since Francis' election, no Americans made the list.	Second Francis Americans
The new cardinals come from countries such as Ethiopia, New Zealand and Myanmar.	Ethiopia New Zealand Myanmar

Yapılan çalışmanın başarısı için doğrulama oranı aşağıdaki gibi hesaplanır:

Özet dokümanda yer alan her bir cümle için orijinal tam metin dokümanda Simhash algoritması kullanılarak en benzer olarak belirlenen cümlenin özel kimliklerinin eşleşmesi bakımından doğrulama oranı hesaplanır. Tüm özet cümleleri için bu süreç işletildikten sonra tüm doğrulama oranları toplanıp, toplam özet cümle sayısına bölünerek ortalama doküman doğrulama oranı tespit edilmektedir.

Örnek olarak Çizelge 5.2'deki ilk özet cümlesi ve bu özet cümlesine orijinal tam metin dokümanda benzer cümle için doğrulama oranı aşağıdaki şekilde hesaplanmıştır.

"The 15 new cardinals will be installed on February 14." özet cümlesine ait özel kimlikler;

- 15, February 14

En benzer orijinal tam metin cümlesi ***"Pope Francis said Sunday that he would hold a meeting of cardinals on February 14 "during which I will name 15 new Cardinals who, coming from 13 countries from every continent, manifest the indissoluble links between the Church of Rome and the particular Churches present in the world," according to Vatican Radio."*** ait özel kimlikler;

- Pope Francis, Sunday, February 14, 15, 13, Churches, Vatican Radio

Özet cümlesi için doğrulama oranı yukarıda detayları verilen Formül 4.1'e göre hesaplandığında;

V = 2/2 = %100 olarak tespit edilir.

Bu sonuç, özet cümlesinde özel kimlik bakımından herhangi bir bilgi kaybının olmadığı anlamına gelmektedir.

5.1.2 Türkçe Haber Sitesi Veri Seti Kullanılarak Özel Kimlik Tespiti Yöntemi Vasıtasıyla Otomatik Doküman Doğrulama

Veri setindeki tam metin ve özetleri, Web'ten arama yöntemi ile üretilmiştir. Öncelikle başlangıç bir tohum kelime ile Web'ten Türkçe haber sitelerinin içerikleri tam metin, manşet kısımları ise özet doküman olarak üretilmiştir. Üretilen tam metin ve özet dokümanların bir örneği Çizelge 5.5'teki gibidir.

Çizelge 5.5 Doğrulama yapılacak özet ve özete ait tam metin doküman

Orijinal Tam Metin Doküman	**Orijinal Özet Doküman**
Ziraat Bankası'nın açıklamasında: "İmar Barışı'ndan faydalanmak için 15 Haziran 2019 tarihine kadar başvuruda bulunan vatandaşlarımız için yapı kayıt belgesi bedellerinin en son yatırılabileceği tarih 30 Haziran 2019'dur.Vatandaşlarımızın bu tarihe kadar gerçekleştireceği ödemelerle ile ilgili yoğunluk yaşamamaları için Ziraat Bankası, Halkbank ve Vakıfbank 29 Haziran Cumartesi ve 30 Haziran Pazar günleri 10.00-16.30 saatleri arasında 81 ilde yeterli sayıda şube ile İmar Barışı ödemelerini almaya devam edeceklerdir. Ziraat Bankası, Halkbank ve Vakıfbank'ın 29 Haziran Cumartesi ve 30 Haziran Pazar günleri 10.00-16.30 saatleri	İmar Barışı ödemelerinde yoğunluk yaşanmaması için 29 ve 30 Haziran tarihlerinde Ziraat Bankası, Halkbank ve Vakıfbank veznelerinin saat 10:00 – 16:00 saatleri arasında ödeme tahsil edeceği duyuruldu.

arasında açık olacak şubeleri www.ziraatbank.com.tr , www.halkbank.com.tr ve www.vakifbank.com.tr internet sitelerinden öğrenilebilecektir" denildi. İmar Barışı ödemelerinin, 30 Haziran 2019 Pazar günü saat 23.59'a kadar Ziraat Bankası ATM'lerinden, Ziraat Bankası, Halkbank ve Vakıfbank internet şubelerinden, Ziraat Bankası ve Vakıfbank mobil uygulamalarından da yapılabileceği belirtildi.	

Orijinal özet dokümanda bulunan her bir cümle, orijinal tam metin dokümana ait tüm cümleler ile Simhash metin benzerlik algoritması kullanılarak benzerlikleri bakımından kıyaslanmış ve en benzer olan 2 cümle seçilmiştir. Çizelge 5.5'te verilen özet ve orijinal tam metin doküman için benzer olarak seçilen cümleler Çizelge 5.6'da verilmiştir.

Çizelge 5.6 Özet dokümanda bulunan cümlelere en benzer orijinal tam metin doküman cümleleri

Orijinal Özet Doküman Cümle	**En Benzer 2 Cümle**
İmar Barışı ödemelerinde yoğunluk yaşanmaması için 29 ve 30 Haziran tarihlerinde Ziraat Bankası, Halkbank ve Vakıfbank veznelerinin saat 10:00 – 16:00 saatleri arasında ödeme tahsil edeceği duyuruldu.	1- Ziraat Bankasi'nin açiklamasinda: "Imar Barisi'ndan faydalanmak için 15 Haziran 2019 tarihine kadar basvuruda bulunan vatandaslarimiz için yapi kayit belgesi bedellerinin en son yatirilabilecegi tarih 30 Haziran 2019'dur. Vatandaslarimizin bu tarihe kadar gerçeklestirecegi ödemelerle ile ilgili yogunluk yasamamalari için Ziraat Bankasi, Halkbank ve Vakifbank 29 Haziran Cumartesi ve 30 Haziran Pazar günleri

	10.00-16.30 saatleri arasinda 81 ilde yeterli sayida sube ile Imar Barisi ödemelerini almaya devam edeceklerdir. 2- Ziraat Bankasi, Halkbank ve Vakifbank'in 29 Haziran Cumartesi ve 30 Haziran Pazar günleri 10.00-16.30 saatleri arasinda açik olacak subeleri www.ziraatbank.com.tr, www.halkbank.com.tr ve www.vakifbank.com.tr internet sitelerinden ögrenilebilecektir" denildi.

Özet doküman cümlelerine ait tespit edilen özel kimlikler Çizelge 5.7'deki gibidir.

Çizelge 5.7 Özet doküman cümlelerine ait tespit edilen özel kimlikler

Özet Doküman Cümle	**Özel Kimlik Değeri**
Imar Barisi ödemelerinde yogunluk yasanmamasi için 29 ve 30 Haziran tarihlerinde Ziraat Bankasi, Halkbank ve Vakifbank veznelerinin saat 10:00 – 16:00 saatleri arasinda ödeme tahsil edecegi duyuruldu.	Ziraat Bankasi Halkbank Vakifbank 10:00 16:00 29 30

Özet doküman cümlelerine en benzer orijinal tam metin doküman cümleleri için tespit edilen özel kimlikler Çizelge 5.8'deki gibidir.

Çizelge 5.8 Benzer tam metin doküman cümlelerine ait özel kimlikler

Tam Metin Cümle	**Özel Kimlik Değeri**
Ziraat Bankasi, Halkbank ve Vakifbank'in 29 Haziran Cumartesi ve 30 Haziran Pazar günleri 10.00-16.30 saatleri arasinda açik olacak subeleri	Halkbank Vakifbank 29

www.ziraatbank.com.tr, www.halkbank.com.tr ve www.vakifbank.com.tr internet sitelerinden ögrenilebilecektir" denildi.	30 10.00 16.30 www.ziraatbank.com.tr www.halkbank.com.tr www.vakifbank.com.tr
Ziraat Bankasi'nin açiklamasinda: "Imar Barisi'ndan faydalanmak için 15 Haziran 2019 tarihine kadar basvuruda bulunan vatandaslarimiz için yapi kayit belgesi bedellerinin en son yatirilabilecegi tarih 30 Haziran 2019'dur.Vatandaslarimizin bu tarihe kadar gerçeklestirecegi ödemelerle ile ilgili yogunluk yasamamalari için Ziraat Bankasi, Halkbank ve Vakifbank 29 Haziran Cumartesi ve 30 Haziran Pazar günleri 10.00-16.30 saatleri arasinda 81 ilde yeterli sayida sube ile Imar Barisi ödemelerini almaya devam edeceklerdir.	Ziraat Bankasi 15 2019 30 Halkbank Vakifbank 2019 29 Cumartesi 30 10.00 16.30 81

Tüm bu aşamalardan sonra, özet dokümanda yer alan her bir cümle için orijinal tam metin dokümanda Simhash algoritması kullanılarak en benzer olarak belirlenen cümlenin özel kimliklerinin eşleşmesi bakımından doğrulama oranı hesaplanır. Tüm özet cümleleri için bu süreç işletildikten sonra tüm doğrulama oranları toplanıp, toplam özet cümle sayısına bölünerek ortalama doküman doğrulama oranı tespit edilmektedir. Çizelge 5.6'daki ilk özet cümlesi ve bu özet cümlesine orijinal tam metin dokümanda benzer cümle için doğrulama oranı aşağıdaki şekilde hesaplanmıştır.

"İmar Barışı ödemelerinde yoğunluk yaşanmaması için 29 ve 30 Haziran tarihlerinde Ziraat Bankası, Halkbank ve Vakıfbank veznelerinin saat 10:00

– 16:00 saatleri arasında ödeme tahsil edeceği duyuruldu." özet cümlesine ait özel kimlikler;

- Ziraat Bankasi, Halkbank, Vakifbank, 10:00, 16:00, 29, 30

En benzer orijinal tam metin cümlesi ***"Ziraat Bankasi'nin açiklamasinda: "Imar Barisi'ndan faydalanmak için 15 Haziran 2019 tarihine kadar basvuruda bulunan vatandaslarimiz için yapi kayit belgesi bedellerinin en son yatirilabilecegi tarih 30 Haziran 2019'dur.Vatandaslarimizin bu tarihe kadar gerçeklestirecegi ödemelerle ile ilgili yogunluk yasamamalari için Ziraat Bankasi, Halkbank ve Vakifbank 29 Haziran Cumartesi ve 30 Haziran Pazar günleri 10.00-16.30 saatleri arasinda 81 ilde yeterli sayida sube ile Imar Barisi ödemelerini almaya devam edeceklerdir."*** ait özel kimlikler;

- Ziraat Bankasi, 15, 2019, 30, 2019, Halkbank, Vakifbank, 29, Cumartesi, 30, 10.00, 16.30, 81

Özet cümlesi için doğrulama oranı yukarıda detayları verilen Formül 4.1'e göre hesaplandığında;

V = 7/7 = %100 olarak tespit edilir.

Bu sonuç, özet cümlesinde özel kimlik bakımından herhangi bir bilgi kaybının olmadığı anlamına gelmektedir.

5.1.3 Transformer Mimarisi Vasıtasıyla Otomatik Semantik Doküman Doğrulama

Bu bölümde, Reuters veri setindeki tam metin ve bu metinlere ait özet dokümanların Transformer mimarisi ile anlamsal olarak doğrulanması sağlanmaktadır. Öncelikle doğrulacak özet dokümandaki kelimelerin orijinal tam metin dokümandaki hangi cümleler ile doğrulanacağının belirlenmesi

gerekmektedir. Çizelge 5.9'da doğrulama yapılacak özet ve özete ait tam metin doküman örneği yer almaktadır.

Çizelge 5.9 Semantik doğrulama yapılacak özet ve özete ait tam metin doküman

Orijinal Tam Metin Doküman	**Orijinal Özet Doküman**
For the second time during his papacy, Pope Francis has announced a new group of bishops and archbishops set to become cardinals -- and they come from all over the world. Pope Francis said Sunday that he would hold a meeting of cardinals on February 14 "during which I will name 15 new Cardinals who, coming from 13 countries from every continent, manifest the indissoluble links between the Church of Rome and the particular Churches present in the world," according to Vatican Radio. New cardinals are always important because they set the tone in the church and also elect the next pope, CNN Senior Vatican Analyst John L. Allen said. They are sometimes referred to as the princes of the Catholic Church. The new cardinals come from countries such as Ethiopia, New Zealand and Myanmar. "This is a pope who very much wants to reach out to people on the margins, and you clearly see that in this set," Allen said. "You're talking about cardinals from typically overlooked places, like Cape Verde, the Pacific island of Tonga, Panama, Thailand, Uruguay." But for the second time since Francis' election, no Americans made the list. "Francis' pattern is very clear: He wants to go to the geographical peripheries rather than places that are already top-heavy with cardinals," Allen said. Christopher	The 15 new cardinals will be installed on February 14. They come from countries such as Myanmar and Tonga. No Americans made the list this time or the previous time in Francis' papacy

Bellitto, a professor of church history at Kean University in New Jersey, noted that Francis announced his new slate of cardinals on the Catholic Feast of the Epiphany, which commemorates the visit of the Magi to Jesus' birthplace in Bethlehem. "On feast of three wise men from far away, the Pope's choices for cardinal say that every local church deserves a place at the big table." In other words, Francis wants a more decentralized church and wants to hear reform ideas from small communities that sit far from Catholicism's power centers, Bellitto said. That doesn't mean Francis is the first pontiff to appoint cardinals from the developing world, though. Beginning in the 1920s, an increasing number of Latin American churchmen were named cardinals, and in the 1960s, St. John XXIII, whom Francis canonized last year, appointed the first cardinals from Japan, the Philippines and Africa. In addition to the 15 new cardinals Francis named on Sunday, five retired archbishops and bishops will also be honored as cardinals. Last year, Pope Francis appointed 19 new cardinals, including bishops from Haiti and Burkina Faso.	

Öncelikle orijinal tam metin ve özet doküman cümlelere ayrıştırılmıştır (Çizelge 5.10 ve Çizelge 5.11).

Çizelge 5.10 Orijinal tam metin doküman cümleleri

Cümle No	**Cümle Metni**
1	For the second time during his papacy, Pope Francis has announced a new group of bishops and archbishops set to become cardinals -- and they come from all over the world.

2	Pope Francis said Sunday that he would hold a meeting of cardinals on February 14 "during which I will name 15 new Cardinals who, coming from 13 countries from every continent, manifest the indissoluble links between the Church of Rome and the particular Churches present in the world," according to Vatican Radio.
3	New cardinals are always important because they set the tone in the church and also elect the next pope, CNN Senior Vatican Analyst John L. Allen said.
4	They are sometimes referred to as the princes of the Catholic Church.
5	The new cardinals come from countries such as Ethiopia, New Zealand and Myanmar.
6	"This is a pope who very much wants to reach out to people on the margins, and you clearly see that in this set," Allen said.
7	"You're talking about cardinals from typically overlooked places, like Cape Verde, the Pacific island of Tonga, Panama, Thailand, Uruguay."
8	But for the second time since Francis' election, no Americans made the list.
9	"Francis' pattern is very clear: He wants to go to the geographical peripheries rather than places that are already top-heavy with cardinals," Allen said.
10	Christopher Bellitto, a professor of church history at Kean University in New Jersey, noted that Francis announced his new slate of cardinals on the Catholic Feast of the Epiphany, which commemorates the visit of the Magi to Jesus' birthplace in Bethlehem.
11	"On feast of three wise men from far away, the Pope's choices for cardinal say that every local church deserves a place at the big table."

12	In other words, Francis wants a more decentralized church and wants to hear reform ideas from small communities that sit far from Catholicism's power centers, Bellitto said.
13	That doesn't mean Francis is the first pontiff to appoint cardinals from the developing world, though.
14	Beginning in the 1920s, an increasing number of Latin American churchmen were named cardinals, and in the 1960s, St. John XXIII, whom Francis canonized last year, appointed the first cardinals from Japan, the Philippines and Africa.
15	In addition to the 15 new cardinals Francis named on Sunday, five retired archbishops and bishops will also be honored as cardinals.
16	Last year, Pope Francis appointed 19 new cardinals, including bishops from Haiti and Burkina Faso.

Çizelge 5.11 Orijinal özet doküman cümleleri

Cümle No	**Cümle Metni**
1	The 15 new cardinals will be installed on February 14.
2	They come from countries such as Myanmar and Tonga.
3	No Americans made the list this time or the previous time in Francis' papacy

Orijinal özet dokümanda bulunan her bir cümle, orijinal tam metin dokümana ait tüm cümleler ile Simhash metin benzerlik algoritması kullanılarak benzerlikleri bakımından kıyaslanmış ve en benzer olan 2 cümle seçilmiştir. Çizelge 5.10'da verilen özet ve orijinal tam metin doküman için benzer olarak seçilen cümleler Çizelge 5.12'de verilmiştir.

Çizelge 5.12 Özet dokümanda bulunan cümlelere en benzer orijinal tam metin doküman cümleleri

Orijinal Özet Doküman Cümle	**En Benzer 2 Cümle**

The 15 new cardinals will be installed on February 14.	1- Pope Francis said Sunday that he would hold a meeting of cardinals on February 14 "during which I will name 15 new Cardinals who, coming from 13 countries from every continent, manifest the indissoluble links between the Church of Rome and the particular Churches present in the world," according to Vatican Radio. 2- In addition to the 15 new cardinals Francis named on Sunday, five retired archbishops and bishops will also be honored as cardinals.
They come from countries such as Myanmar and Tonga.	1- The new cardinals come from countries such as Ethiopia, New Zealand and Myanmar. 2- Pope Francis said Sunday that he would hold a meeting of cardinals on February 14 "during which I will name 15 new Cardinals who, coming from 13 countries from every continent, manifest the indissoluble links between the Church of Rome and the particular Churches present in the world," according to Vatican Radio.
No Americans made the list this time or the previous time in Francis' papacy	1- But for the second time since Francis' election, no Americans made the list. 2- For the second time during his papacy, Pope Francis has announced a new group of bishops and archbishops set to become cardinals -- and they come from all over the world.

Daha sonra bu benzer olarak bulunan cümleler önce birleştirilmiş sonra Transformer modeline parametre olarak verilmiştir. Transformer modeli bu

cümle gruplarının her biri için soyut özetler üretmiştir. Her bir cümle grubu için üretilen Transformer çıktısı Çizelge 5.13'te yer almaktadır.

Çizelge 5.13 En benzer tam metin doküman cümleleri için Transformer modelinin ürettiği soyut özetler

En Benzer 2 Cümle	**Transformer Özeti**
1- Pope Francis said Sunday that he would hold a meeting of cardinals on February 14 "during which I will name 15 new Cardinals who, coming from 13 countries from every continent, manifest the indissoluble links between the Church of Rome and the particular Churches present in the world," according to Vatican Radio. 2- In addition to the 15 new cardinals Francis named on Sunday, five retired archbishops and bishops will also be honored as cardinals.	Pope Francis said he would hold a meeting of cardinals on February 14. He will name 15 new cardinals who" manifest the indissoluble links between the Church of Rome and the particular Churches present in the world," he said. The Vatican Radio reports that the cardinals will come from 13 countries from every continent. In addition to the 15 new cardinals that Francis named on Sunday, five retired archbishops and bishops will be honored as cardinals.
1- The new cardinals come from countries such as Ethiopia, New Zealand and Myanmar. 2- Pope Francis said Sunday that he would hold a meeting of cardinals on February 14 "during which I will name 15 new Cardinals who, coming from 13 countries from every continent, manifest the indissoluble links between the Church of Rome and the	New cardinals are being brought in from countries like New Zealand, Myanmar, and Ethiopia. Pope Francis said he would hold a meeting of cardinals on February 14. He will name 15 new cardinals who " manifest the indissoluble links between the Church of Rome and the particular Churches present in the world," he said. The Vatican Radio reports that the cardinals will come from 13 countries from every continent.

particular Churches present in the world," according to Vatican Radio.	
1- But for the second time since Francis' election, no Americans made the list. 2- For the second time during his papacy, Pope Francis has announced a new group of bishops and archbishops set to become cardinals -- and they come from all over the world.	No Americans made it to Francis' list for the first time since the pope's election. Pope Francis has announced a new group of bishops and archbishops set to become cardinals. The new cardinals come from all over the world. They are the second time in his papacy that has been announced by Pope Francis. They come from around the world.

Son aşamada ise orijinal özet doküman cümleleri ile Transformer tarafından üretilen özet doküman cümleleri arasındaki benzerliği Simhash ve Cross Encoder metin benzerliği algoritmaları kullanılarak değerlendirilmiştir. Karşılaştırmalı olarak doküman doğrulama sürecinin ortalama doğruluğu hesaplanmıştır. Çizelge 5.14'te Transformer modeli tarafından üretilen özet doküman verilmiştir.

Çizelge 5.14 Cümle gruplarının bir arada kullanılmasıyla Transformer modeli tarafından üretilen özet doküman

Transformer Özet Doküman
Pope Francis said he would hold a meeting of cardinals on February 14. He will name 15 new cardinals who " manifest the indissoluble links between the Church of Rome and the particular Churches present in the world," he said. The Vatican Radio reports that the cardinals will come from 13 countries from every continent. In addition to the 15 new cardinals Francis named on Sunday,

five retired archbishops and bishops will also be honored as cardinals. New cardinals are being brought in from countries like New Zealand, Myanmar, and Ethiopia. No Americans made it to Francis' list for the first time since the pope's election. Pope Francis has announced a new group of bishops and archbishops set to become cardinals. The new cardinals come from all over the world. They are the second time in his papacy that has been announced by Pope Francis. They come from around the world.

Ayrıca, Çizelge 5.12'deki gruplarda yer alan cümleleri Transformer modeline tek tek uygulayarak bağlam genişletmenin etkisi değerlendirilmeye çalışılmıştır. Bu kapsamda üretilen Transformer modeli özeti Çizelge 5.15'te görülmektedir.

Çizelge 5.15 Cümlelerin ayrı ayrı kullanılmasıyla Transformer modeli tarafından üretilen özet doküman

Transformer Özet Doküman
Pope Francis announced on Sunday his intention to convene a meeting of cardinals on February 14th. Besides the 15 new cardinals appointed by Pope Francis on Sunday, there will also be five retired archbishops and bishops. The newly appointed cardinals hail from countries including Ethiopia, New Zealand, and Myanmar. No Americans were included in the list of appointed cardinals. Pope Francis has announced a diverse group of bishops and archbishops from around the world to become cardinals, marking the second time during his papacy.

Orijinal tam metin doküman ile orijinal özet doküman arasında %82,1 Simhash ve %79,8 Cross Encoder metin benzerliği elde edilirken, orijinal tam metin doküman ile Transformer özeti arasında %84,8 Simhash ve %81,9 Cross Encoder metin benzerliği elde edilmiştir (Orijinal doküman cümlelerini gruplama yaklaşımını kullanarak). Transformer modeliyle üretilen özet dokümanda özel kimliklerin (Named entities) korunduğu görülmektedir. Ayrıca bu sonuç, Transformer modelinin ürettiği özet dokümanda bulunan bilgilerin orijinal özet belgede de yüksek oranda yer aldığını göstermektedir.

Ancak soyut özetler oluşturmak için Transformer modeline en benzer orijinal tam metin doküman cümleleri ayrı ayrı verildiğinde %81,3 Simhash ve %79,45 Cross Encoder ile daha düşük metin benzerliği elde edilmiştir. Bu, cümleleri Transformer modeline ayrı ayrı sağlamanın bağlamsal bilgi kaybına yol açarak benzerliğin azalmasına yol açtığını açıkça göstermektedir. Öte yandan, Transformer modeline verilmeden önce cümleler birleştirildiğinde cümleler arasındaki bağlamsal bilgiler korunarak anlamsal sonuçların iyileştirilmesi sağlanmıştır.

5.1.4 Transformer Mimarisi Vasıtasıyla Otomatik Soyut Özetleme

Bu bölümde, Reuters veri setindeki tam metin ve bu metinlere ait özet dokümanların Transformer mimarisi ile soyut özetleme yapılması sağlanmaktadır. Çizelge 5.16'da soyut özetleme yapılacak orijinal tam metin doküman örneği yer almaktadır.

Çizelge 5.16 Soyut özetleme yapılacak orijinal tam metin doküman

Orijinal Tam Metin Doküman
For the second time during his papacy, Pope Francis has announced a new group of bishops and archbishops set to become cardinals -- and they come from all over the world. Pope Francis said Sunday that he would hold a meeting of cardinals on February 14 "during which I will name 15 new Cardinals who, coming from 13 countries from every continent, manifest the indissoluble links between the Church of Rome and the particular Churches present in the world," according to Vatican Radio. New cardinals are always important because they set the tone in the church and also elect the next pope, CNN Senior Vatican Analyst John L. Allen said. They are sometimes referred to as the princes of the Catholic Church. The new cardinals come from countries such as Ethiopia, New Zealand and Myanmar. "This is a pope who very much wants to reach out to people on the margins, and you clearly see that in this set," Allen said. "You're talking about cardinals from typically overlooked

places, like Cape Verde, the Pacific island of Tonga, Panama, Thailand, Uruguay." But for the second time since Francis' election, no Americans made the list. "Francis' pattern is very clear: He wants to go to the geographical peripheries rather than places that are already top-heavy with cardinals," Allen said. Christopher Bellitto, a professor of church history at Kean University in New Jersey, noted that Francis announced his new slate of cardinals on the Catholic Feast of the Epiphany, which commemorates the visit of the Magi to Jesus' birthplace in Bethlehem. "On feast of three wise men from far away, the Pope's choices for cardinal say that every local church deserves a place at the big table." In other words, Francis wants a more decentralized church and wants to hear reform ideas from small communities that sit far from Catholicism's power centers, Bellitto said. That doesn't mean Francis is the first pontiff to appoint cardinals from the developing world, though. Beginning in the 1920s, an increasing number of Latin American churchmen were named cardinals, and in the 1960s, St. John XXIII, whom Francis canonized last year, appointed the first cardinals from Japan, the Philippines and Africa. In addition to the 15 new cardinals Francis named on Sunday, five retired archbishops and bishops will also be honored as cardinals. Last year, Pope Francis appointed 19 new cardinals, including bishops from Haiti and Burkina Faso.

İlk olarak orijinal tam metin doküman cümlelerine ayrıştırılır. Çizelge 5.17'de soyut özetleme yapılacak orijinal tam metin dokümanın cümleleri verilmiştir.

Çizelge 5.17 Soyut özetleme yapılacak orijinal tam metin doküman cümleleri

Cümle No	**Cümle Metni**
1	For the second time during his papacy, Pope Francis has announced a new group of bishops and archbishops set to become cardinals -- and they come from all over the world.
2	Pope Francis said Sunday that he would hold a meeting of cardinals on February 14 "during which I will name 15 new Cardinals who, coming from 13 countries from every

	continent, manifest the indissoluble links between the Church of Rome and the particular Churches present in the world," according to Vatican Radio.
3	New cardinals are always important because they set the tone in the church and also elect the next pope, CNN Senior Vatican Analyst John L. Allen said.
4	They are sometimes referred to as the princes of the Catholic Church.
5	The new cardinals come from countries such as Ethiopia, New Zealand and Myanmar.
6	"This is a pope who very much wants to reach out to people on the margins, and you clearly see that in this set," Allen said.
7	"You're talking about cardinals from typically overlooked places, like Cape Verde, the Pacific island of Tonga, Panama, Thailand, Uruguay."
8	But for the second time since Francis' election, no Americans made the list.
9	"Francis' pattern is very clear: He wants to go to the geographical peripheries rather than places that are already top-heavy with cardinals," Allen said.
10	Christopher Bellitto, a professor of church history at Kean University in New Jersey, noted that Francis announced his new slate of cardinals on the Catholic Feast of the Epiphany, which commemorates the visit of the Magi to Jesus' birthplace in Bethlehem.
11	"On feast of three wise men from far away, the Pope's choices for cardinal say that every local church deserves a place at the big table."
12	In other words, Francis wants a more decentralized church and wants to hear reform ideas from small communities that sit far from Catholicism's power centers, Bellitto said.

13	That doesn't mean Francis is the first pontiff to appoint cardinals from the developing world, though.
14	Beginning in the 1920s, an increasing number of Latin American churchmen were named cardinals, and in the 1960s, St. John XXIII, whom Francis canonized last year, appointed the first cardinals from Japan, the Philippines and Africa.
15	In addition to the 15 new cardinals Francis named on Sunday, five retired archbishops and bishops will also be honored as cardinals.
16	Last year, Pope Francis appointed 19 new cardinals, including bishops from Haiti and Burkina Faso.

Cümle listesindeki her bir cümle kendi dışındaki diğer cümleler ile Simhash metin benzerlikleri bakımından kıyaslanmış ve en benzer cümleler bir arada olacak şekilde cümleler gruplanmıştır. Örneğin Çizelge 5.18'de ***Cümle 1***'in diğer cümlelerle olan Simhash metin benzerlik oranı hesaplanmıştır.

Çizelge 5.18 Cümle 1 için hesaplanan Simhash metin benzerlik oranları

Kaynak Cümle No	**Hedef Cümle No**	**Benzerlik Oranı (%)**
1	2	92.1
1	3	88.4
1	4	89.2
1	5	85.3
1	6	87.2
1	7	86.1
1	8	88.65
1	9	87.45
1	10	85.85
1	11	84.65
1	12	93.2
1	13	91.8
1	14	89.07

1	15	86.3
1	16	83.98

Grup sayısı parametrik olarak yönetilmiştir. Bu gruplama adımı, ortak anlamsal özellikleri paylaşan cümle gruplarının oluşturulmasını kolaylaştırmış ve böylece doküman içinde tutarlı bilgi birimleri oluşturulmuştur. Dengeli bir temsili sağlamak için grup sayısı, dokümandaki toplam cümle sayısının yüzdesi olarak belirlenmiştir. Yapılan deneysel araştırmalarda grup sayısı doküman toplam cümle sayısının %25'i olacak şekilde ayarlandığında, yüksek başarım oranları elde edilmiştir. Literatürde bilinen özetleme algoritmaları da (Pedrycz vd., 2022; Huan vd., 2022) %25-30 bandında özet üretmektedir. Bu örnekte grup sayısı 4 olarak belirlenmiştir. Simhash benzerliklerine göre en yüksek benzerliği gösteren cümleler gruplandırılmıştır. Çizelge 5.19'da grup bazında yer alan cümle listeleri verilmiştir.

Çizelge 5.19 Grup bazında yer alan cümle listeleri

Grup	**Cümle Listesi (Cümle Numaraları)**
Grup 1	1 2 12 13
Grup 2	3 10 14 15
Grup 3	4 5 7 16
Grup 4	6 8 9

	11

Daha sonra bu 4 gruba ait cümleler Transformer modeline parametre olarak verilmiştir. Transformer modeli bu grupların her biri için soyut cümleler üretmiştir. Çizelge 5.20'de gruplanmış cümleler ve Transformer modeli tarafından üretilen cümleler verilmiştir.

Çizelge 5.20 Gruplanmış cümleler ve Transformer modeli tarafından üretilen cümleler

Grup No	**Transformer Soyut Cümle**
Grup 1 For the second time during his papacy, Pope Francis has announced a new group of bishops and archbishops set to become cardinals -- and they come from all over the world. Pope Francis said Sunday that he would hold a meeting of cardinals on February 14 "during which I will name 15 new Cardinals who, coming from 13 countries from every continent, manifest the indissoluble links between the Church of Rome and the particular Churches present in the world," according to Vatican Radio. In other words, Francis wants a more decentralized church and wants to hear reform ideas from small communities that sit far from Catholicism's power centers, Bellitto said. That doesn't mean Francis is the first pontiff to appoint cardinals from the developing world, though. Last year, Pope Francis appointed	Pope Francis appoints 15 new cardinals from 13 countries from all over the world. Last year, he appointed 19 cardinals, including bishops from Haiti and Burkina Faso. He wants to hear reform ideas from small communities that sit far from Catholicism's power centers.

19 new cardinals, including bishops from Haiti and Burkina Faso.	
Grup 2 New cardinals are always important because they set the tone in the church and also elect the next pope, CNN Senior Vatican Analyst John L. Allen said. Christopher Bellitto, a professor of church history at Kean University in New Jersey, noted that Francis announced his new slate of cardinals on the Catholic Feast of the Epiphany, which commemorates the visit of the Magi to Jesus' birthplace in Bethlehem. Beginning in the 1920s, an increasing number of Latin American churchmen were named cardinals, and in the 1960s, St. John XXIII, whom Francis canonized last year, appointed the first cardinals from Japan, the Philippines and Africa. In addition to the 15 new cardinals Francis named on Sunday, five retired archbishops and bishops will also be honored as cardinals.	New cardinals are important because they set the tone in the church and elect the next pope. Five retired archbishops and bishops will also be honored as cardinals. St. John XXIII, whom Francis canonized last year, appointed the first cardinals from Japan, the Philippines and Africa in 1960s.
Grup 3 They are sometimes referred to as the princes of the Catholic Church. The new cardinals come from countries such as Ethiopia, New Zealand and Myanmar. "You're talking about cardinals from typically overlooked places, like Cape Verde, the Pacific island of Tonga, Panama, Thailand, Uruguay." CNN's Daniel Burke and Christabelle Fombu contributed to this report.	The new cardinals come from countries such as Ethiopia, New Zealand and Myanmar. They are sometimes referred to as the princes of the Catholic Church. "You're talking about cardinals from typically overlooked places, like Cape Verde, the Pacific island of Tonga, Panama, Thailand, Uruguay"

Grup 4 "This is a pope who very much wants to reach out to people on the margins, and you clearly see that in this set," Allen said. But for the second time since Francis' election, no Americans made the list. "Francis' pattern is very clear: He wants to go to the geographical peripheries rather than places that are already top-heavy with cardinals," Allen said. "On feast of three wise men from far away, the Pope's choices for cardinal say that every local church deserves a place at the big table."	For the second time since Francis' election, no Americans made the list. "Francis' pattern is very clear: He wants to go to the geographical peripheries," Allen says. "This is a pope who very much wants to reach out to people on the margins," he says.

Son olarak Transformer modelinin her grup için ürettiği soyut cümleler bir araya getirilerek soyut özet doküman oluşturulmuş ve otomatik soyut özetleme süreci tamamlanmıştır. Bu doküman, orijinal içeriğin damıtılmış özünü temsil etmekte, kilit noktaları ve ana fikirleri kısa ve öz bir şekilde yakalamaktadır. Yeni soyut özet doküman ile tam metin doküman arasındaki anlamsal benzerlik değerlendirilmiştir. Transformer modelinin benzerlik ölçümleri, orijinal özeti aslına sadık kalarak temsil etme konusundaki ilgili performanslarını belirlemek için ayrı ayrı değerlendirilmiştir. Sonuçlar deneyler bölümünde detaylı olarak paylaşılmıştır. Çizelge 5.21'de Transformer modeli tarafından üretilen soyut özet dokümanın son hali verilmiştir.

Çizelge 5.21 Transformer modeli tarafından üretilen soyut özet doküman

Transformer Modeli Tarafından Üretilen Soyut Özet Doküman
Pope Francis appoints 15 new cardinals from 13 countries from all over the world. Last year, he appointed 19 cardinals, including bishops from Haiti and Burkina Faso. He wants to hear reform ideas from small communities that sit far from Catholicism's power centers. New cardinals are important because they set the tone in the church and elect the next pope. Five retired

archbishops and bishops will also be honored as cardinals. St. John XXIII, whom Francis canonized last year, appointed the first cardinals from Japan, the Philippines and Africa in 1960s. The new cardinals come from countries such as Ethiopia, New Zealand and Myanmar. They are sometimes referred to as the princes of the Catholic Church. "You're talking about cardinals from typically overlooked places, like Cape Verde, the Pacific island of Tonga, Panama, Thailand, Uruguay". For the second time since Francis' election, no Americans made the list. "Francis' pattern is very clear: He wants to go to the geographical peripheries," Allen says. "This is a pope who very much wants to reach out to people on the margins," he says.

6. DENEYSEL ÇALIŞMALAR

Bu bölümde, önerilen sistemin performansını değerlendirmek amacıyla gerçekleştirilen deneysel çalışmalar sunulmaktadır. Çalışma kapsamında, doküman doğrulama sürecinin farklı aşamaları için kapsamlı deneyler gerçekleştirilmiş ve elde edilen sonuçlar analiz edilmiştir. Deneysel çalışmalar, özel kimlik tespiti yönteminin yanı sıra, Transformer tabanlı anlamsal doküman doğrulama ve soyut özetleme süreçlerinin başarı oranlarını kapsamaktadır. Bu deneyler, sistemin hem anlamsal doğruluk hem de sözdizimsel uyumluluk açısından genel performansını ortaya koymayı amaçlamaktadır.

İlk olarak, özel kimlik tespiti yöntemiyle doküman doğrulama süreçleri analiz edilmiştir. Özel kimlik tespiti (NER), özellikle kişi, yer ve organizasyon gibi kritik varlıkların tespit edilmesine odaklanmıştır. Bu yöntem, özetleme ve doğrulama süreçlerinin bilgi bütünlüğünü sağlaması açısından önemli bir bileşen olarak değerlendirilmiştir. NER doğrulaması, özetleme işlemleri sırasında metindeki kritik bilgilerin doğru ve eksiksiz bir şekilde aktarılıp aktarılmadığını tespit etmek için kullanılmıştır. Bu bağlamda hem orijinal tam metin dokümanları hem de özet dokümanlar üzerinde NER işlemleri gerçekleştirilmiş ve bu işlemler sonucunda sistemin doğruluk oranları hesaplanmıştır.

İlerleyen aşamalarda, Transformer tabanlı modeller ile anlamsal doküman doğrulama süreçleri incelenmiştir. Bu deneyler sırasında, orijinal tam metin ve oluşturulan özetler arasındaki semantik uyum, Simhash ve Cross-Encoder algoritmaları ile değerlendirilmiştir. Bununla birlikte, yalnızca anlamsal uyumu ölçmekle yetinilmemiş, özetlerin içerdiği özel varlıkların doğruluğu da analiz edilmiştir. Bu değerlendirme, özetlerin hem anlam hem de içerik açısından ne kadar başarılı olduğunu ortaya koymak amacıyla gerçekleştirilmiştir.

Son olarak, soyut özetleme süreçleri kapsamında elde edilen sonuçlar detaylı bir şekilde analiz edilmiştir. Transformer tabanlı modelin soyut özetleme başarımı, orijinal dokümanlarla kıyaslanarak ölçülmüş ve bu süreçte hem anlamsal hem de NER doğrulama başarım oranları sunulmuştur. Bu deneyler, sistemin farklı veri

türlerinde gösterdiği performansı anlamak ve genel başarısını değerlendirmek için yapılmıştır.

Bu deneysel çalışmalar, her aşamanın yalnızca anlamsal doğruluk açısından değil, aynı zamanda içerdiği özel varlıklar bakımından da güçlü bir doğruluk mekanizması sunduğunu ortaya koymaktadır. Özellikle finansal dokümanlar gibi hassasiyet gerektiren alanlarda, sistemin bilgi kaybını en aza indirerek güvenilir sonuçlar ürettiği gözlemlenmiştir. Sonraki alt bölümlerde, bu deneylere ait detaylı sonuçlar ve analizler sunulmaktadır.

6.1 Özel Kimlik Tespiti Yöntemi Vasıtasıyla Otomatik Doküman Doğrulama

Bu bölümde, önerilen özel kimlik (Named entities) tabanlı doküman doğrulama yaklaşımı, Reuters veri seti üzerinde test edilmiştir. Ayrıca, yazım denetleyicisi sürecinin ve adlandırılmış varlık tespiti amacıyla eklenen normal ifade kurallarının doküman doğrulama doğruluğu üzerindeki etkisini gözlemlemek amacıyla, yazım hatalarına sahip aynı sayıda veri seti üzerinde de deneyler yapılmış ve sonuçlar karşılaştırmalı olarak değerlendirilmiştir.

Makine öğrenmesi modeli aracılığıyla konuya özel Regex kuralları ve yazım denetleyicisi de dikkate alınarak bunların özel kimlik tespiti modeline etkileri değerlendirilmiştir. Özel kimlik tespiti vasıtasıyla doküman doğrulamasının arkasındaki fikir, anlamsal doğrulamanın çözümün bir parçası olmadığı durumlarda dokümanda eşleştirilecek doğal özellikleri sunmasıdır. Anlamsal doğrulama karmaşık ve zaman alıcı bir işlemdir. Öte yandan, önerilen model herhangi bir konuya genellenebilir, uyarlanması ve geliştirilmesi basit, hızlıdır ve literatürdeki doğrulama araştırmalarıyla karşılaştırıldığında kabul edilebilir bir doğruluğa sahiptir.

Bu yaklaşım, onayın kullanıcı tarafından alındığı yarı denetimli (semi-supervised) bir sistem olarak da düşşünülebilir. Kullanıcı, yaklaşımın önerdiği cümlelerden birini ana dokümanlardan özel kimlikler mevcutsa seçer, ilgili

değilse reddeder. Bu biraz daha zaman alır, ancak anlamsal kontrol kullanıcı tarafından yapılabilir ve daha güvenli görünebilir (Dokümanların okunması otomatikleştirilmiştir).

Önerilen özel kimlik tespiti vasıtasıyla doküman doğrulama yaklaşımını test etmek için iki farklı deney gerçekleştirilmiştir. Deneylerde farklı sayıda doküman içeren veri setleri kullanılmıştır. Dokümanlar finans ve spor konularına aittir. Önerilen özel kimlik tabanlı model özellikle finansal özel kimlikleri kapsamaktadır. Ancak farklı doküman konularındaki performansını gözlemlemek amacıyla spor konusuna ait veri seti üzerinde de bir deney yapılmıştır.

6.1.1 Deney I: Yazım Denetleyicisi (Spell Checker) Vasıtasıyla

Bu deneyde, yazım denetleyicisi sürecine dahil edilen 10, 100 ve 1000 Reuters veri seti üzerinde doğrulama gerçekleştirilmiştir. Farklı konu ve sayılara sahip her veri setinde, Regex kuralları eklenerek ve eklenmeden elde edilen doğrulama doğruluğu ve ortalama işlem süresi hesaplanmıştır. Doğrulama sürecinde elde edilen doğruluk Çizelge 6.1'de verilmiştir.

Çizelge 6.1 Özel kimlik tespiti yöntemi (Yazım denetleyicisi ile) vasıtasıyla otomatik doküman doğrulama doğruluğu sonuçları

DS	DK	CME	CMER	CMEB	DNEB	DNER	AAWR	AAOR	AOT
10	Finans	25	7	7	35	16	94.5	91.6	2.4
100	Finans	183	28	32	218	102	91.7	89.7	14.8
1000	Finans	297	75	109	1292	1113	88.6	86.5	193.6

10	Spor	18	5	4	39	39	82.1	82.1	2.6
100	Spor	167	19	24	183	182	79.2	78.9	16.3
1000	Spor	201	45	64	978	975	75.9	75.3	199.8

Çizelge 6.2 Çizelge 6.1'de kullanılan kısaltmalar ve açıklamaları

Kısaltma	**Açıklama**
Doküman Sayısı (DS)	Bir veri setinde bulunan toplam doküman sayısını ifade eder. Örneğin, bu deneyde 10, 100 ve 1000 Reuters dokümanı veri seti üzerinde doğrulama gerçekleştirilmiştir.
Doküman Konusu (DK)	Dokümanların hangi konuları kapsadığını gösterir. Örneğin, bir veri setindeki dokümanlar yapay zeka, makine öğrenmesi ve veri bilimi gibi konuları içerebilir. Bu deneyde, finans ve spor doküman türlerinde doğrulama gerçekleştirilmiştir.
Düzeltilmiş Yanlış Yazılan Girişlerden Regex Kurallarına Göre Tespit Edilen Özel Kimliklerin Sayısı (CMER)	Yanlış yazılan veri girişlerinin düzeltilmesi ve ardından regex (regular expression) kuralları kullanılarak bu girişler içinden özel kimliklerin tespit edilme sayısını ifade eder. Örneğin, bir metin içerisinde yanlışlıkla "Aple" olarak yazılmış "Apple" kelimesi düzeltilip, ardından regex kurallarıyla belirli bir örüntüye göre özel kimlik olarak tespit edilebilir.
Düzeltilmiş Yanlış Yazılan Girişlerden Bert Transformer Tabanlı Sınıflandırma	Yanlış yazılan veri girişlerinin düzeltilmesi sonrası BERT Transformer tabanlı bir sınıflandırma modeli kullanılarak özel kimliklerin tespit edilme sayısını gösterir. Örneğin, "Googel" olarak yanlış yazılan

Modeli Tarafından Tespit Edilen Özel Kimliklerin Sayısı (CMEB)	"Google" kelimesi düzeltilip, BERT modeli kullanılarak bir özel kimlik olarak tespit edilebilir.
Bert Transformer Tabanlı Sınıflandırma Modeline Göre Tespit Edilen Özel Kimliklerin Sayısı (DNEB)	BERT Transformer tabanlı bir sınıflandırma modeli kullanılarak doğrudan tespit edilen özel kimliklerin sayısını ifade eder. Bu model, metin içerisinden belirli özelliklere sahip özel kimlikleri doğrudan tespit edebilir.
Regex Kurallarına Göre Tespit Edilen Özel Kimliklerin Sayısı (DNER)	Regex (regular expression) kuralları kullanılarak metin içerisinden tespit edilen özel kimliklerin sayısını gösterir. Regex, metin içinde belirli desenleri bulmak için kullanılan bir yöntemdir.
Regex Kurallarıyla Doğrulama Doğruluğu (%) (AAWR)	Regex kuralları kullanılarak yapılan tespitlerin doğruluğunu yüzdesel olarak ifade eder. Örneğin, tüm tespitlerin %95'i doğru ise doğrulama doğruluğu %95 olarak belirtilir.
BERT ile Doğrulama Doğruluğu (%) (AAOR)	BERT Transformer tabanlı sınıflandırma modeli kullanılarak yapılan tespitlerin doğruluğunu yüzdesel olarak ifade eder. BERT modeli, metin üzerinde daha karmaşık analizler yapabilme yeteneğine sahiptir, bu yüzden doğruluk oranı önemli bir metriktir.
Ortalama Çalışma Süresi (sn) (AOT)	İşlemin başlangıcından sonuna kadar geçen sürenin ortalamasını saniye cinsinden gösterir. Bu, bir algoritmanın veya işlemin ne kadar hızlı çalıştığını gösterir.
Düzeltilmiş Yanlış Yazılan Girişler (CME)	Başlangıçta yanlış yazılmış ve sonradan düzeltilmiş veri girişlerini ifade eder. Bu, veri temizleme sürecinin bir parçası olarak görülebilir.

6.1.2 Deney II: Yazım Denetleyicisi (Spell Checker) Olmadan

Bu deneyde, yazım denetleyicisi sürecine dahil edilmeyen 10, 100 ve 1000 Reuters veri seti üzerinde doğrulama gerçekleştirilmiştir. Farklı konu ve sayılara sahip her veri setinde, Regex kuralları eklenerek ve eklenmeden elde edilen doğrulama doğruluğu ve ortalama işlem süresi hesaplanmıştır. Doğrulama sürecinde elde edilen doğruluk Çizelge 6.3'te verilmiştir.

Çizelge 6.3 Özel kimlik tespiti yöntemi (Yazım denetleyicisi olmadan) vasıtasıyla otomatik doküman doğrulama doğruluğu sonuçları

DS	DK	DNEB	DNER	AAWR	AAOR	AOT
10	Finans	32	16	90.1	90	2.0
100	Finans	230	71	85.6	83.2	12.4
1000	Finans	1001	209	82.4	81.4	178.9
10	Spor	33	20	82	82	2.2
100	Spor	175	173	76.4	76.1	12.8
1000	Spor	978	975	75.9	75.3	199.8

Deneylerde elde edilen sonuçlar incelendiğinde hem deney I hem de Deney II'de, finansal veri setlerinde elde edilen ortalama doküman doğrulama doğruluğu, farklı sayıda veri setleri için spor veri setlerinden daha yüksektir. Deney I'de 10,

100 ve 1000 finans doküman veri setindeki ortalama doküman doğrulama doğruluğu sırasıyla %94,5, %91,7 ve %88,6 iken, Deney II'de aynı sayıdaki spor veri kümeleri için bu oran %82, %76,4 ve %74,4 olmuştur. Bu, finansal tür için BERT Transformer modelinin tanımlamakta eksik kaldığı özel kimlik türleri için yazılan Regex kurallarının olumlu etkisinin sonucudur.

BERT Transformer modelinin kapsamadığı bazı özel kimlik türlerini tespit etmek amacıyla yazılan Regex kurallarının etkileri de yapılan deneylerde görülmüştür. Hem Deney I hem de Deney II'de, Regex kuralları entegre edilen senaryolarda daha yüksek doğruluk elde etmiştir. Ayrıca finansal türlerin (Euro, Dolar, Türk Lirası gibi para birimleri) belirlenmesine yönelik Regex kuralları yazıldığı için finansal konuda spor konusuna göre daha yüksek doğruluk elde edilmiştir.

Yapılan deneylerde yazım denetleyicisi işleminin doküman doğrulama doğruluğuna olumlu katkı sağladığı görülmüştür. Hem finans hem de spor dokümanları ile yapılan tüm deneylerde, doğrulanacak dokümanlar ilk kez yazım denetleyicisi sürecine dahil edildiğinde, ortalama doğrulama doğruluğu, yazım denetleyicisi kullanılmadığı duruma göre yaklaşık %5 daha yüksek olmuştur.

Deney II'de, Deney I'de olduğu gibi, farklı sayıdaki finansal ve spor veri setlerine ait dokümanlar üzerinde deneyler yapılmıştır. Bu deneylerde kullanılan dokümanlar, yazım denetleyicisi sürecine dahil edilmeden doğrudan doğrulama sürecine tabi tutulmuştur. Sonuçlar incelendiğinde veri seti büyüklüğü ve konusu ne olursa olsun tüm deneylerde Deney I'e göre daha düşük doğrulama doğruluğu elde edilmiştir. Dokümanlar yazım denetleyicisi sürecine alınmadan önce doğrulandığından, hem Bert Transformer tabanlı sınıflandırma modeli hem de Regex kuralları tarafından tespit edilen adlandırılmış varlıkların sayısı Deney I'den daha düşüktür. Sonuç olarak, bu durum orijinal tam metin ve özet dokümanlardaki özel kimliklerin eşleşmemesine neden olmuştur.

Son olarak yapılan deneylerde doküman sayısının ve yazım denetleyicisi işleminin doküman doğrulama sürecine etkisi de incelenmiştir. Doküman sayısı arttıkça ortalama doküman doğrulama süresinin arttığı gözlemlenmiştir. Ayrıca,

yazım denetleyicisi sürecine dahil olmayan dokümanların doğrulanması, dahil edilenlere göre çok daha az zaman almıştır. Ancak ortalama doğrulama doğruluğu daha düşük kalmıştır.

6.2 Transformer Mimarisi Vasıtasıyla Otomatik Semantik Doküman Doğrulama

Bu bölümde Transformer tabanlı anlamsal doküman doğrulama yaklaşımı Reuters veri seti üzerinde test edilmiştir. Veri setinde 3 farklı kategori (İş, Spor ve Finans) bulunmaktadır. Yazım denetleyicisi (Spell Checker) işleminin Transformer tabanlı anlamsal doküman doğrulama doğruluğu üzerindeki etkisini gözlemlemek için aynı sayıda yanlış yazılan veri seti üzerinde de deneyler yapılmış ve sonuçlar karşılaştırmalı olarak değerlendirilmiştir. Ayrıca Transformer modeliyle üretilen özet doküman ile orijinal özet dokümanın metin benzerliği karşılaştırması hem cümle-cümle hem de belge bazında hesaplanmış ve sonuçlar karşılaştırmalı olarak paylaşılmıştır.

6.2.1 Deney I: Yazım Denetleyicisi (Spell Checker) Vasıtasıyla

Bu deneyde, Yazım denetleyicisi sürecine dahil edilen 10, 100 ve 1000 Reuters finans, iş ve spor belgesi veri kümeleri üzerinde Transformer tabanlı anlamsal doküman doğrulaması gerçekleştirilmiştir. Farklı konu ve sayılara sahip her veri setinde doğrulama doğruluğu ve ortalama işlem süresi hesaplanmıştır. Doğrulama sürecinde elde edilen doğruluk oranları Çizelge 6.4'te verilmiştir.

Çizelge 6.4 Transformer mimarisi (Yazım denetleyicisi ile) vasıtasıyla otomatik doküman doğrulama doğruluğu sonuçları

DS	**DK**	**CME**	**ASVA Simhash (%)**	**ASVA Cross Encoder (%)**	**ANVA**	**AOT Simhash (sn)**	**AOT Cross Encoder (sn)**
10	Spor	18	82.30	79.21	86.0	2.6	3.0

100	Spor	167	80.43	79.02	84.15	16.9	23.6
1000	Spor	201	80.10	76.58	82.20	200.1	238.7
10	İş	14	83.09	81.36	89.50	3.2	3.3
100	İş	143	81.23	75.89	84.85	17.8	24.1
1000	İş	198	77.01	78.65	83.05	199.8	224.9
10	Finans	25	84.10	82.45	91.50	3.1	3.3
100	Finans	183	83.35	80.80	89.75	17.7	23.1
1000	Finans	297	80.56	77.71	87.50	205.6	216.8

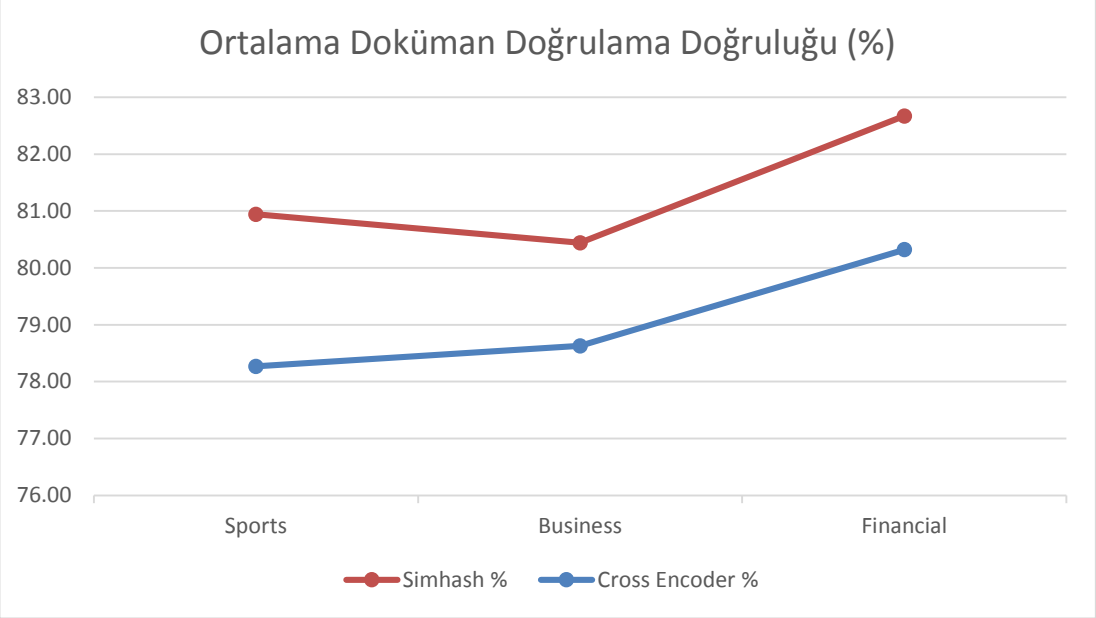
Not: Çizelge 6.4'te kullanılan kısaltmalar: DS= Doküman Sayısı, DK = Doküman Konusu, ASVA = Ortalama Anlamsal Doğrulama Başarımı (Average Semantic Verification Accuracy), AOT = Ortalama İşlem Süresi (Average Operation Time), CME = Düzeltilmiş Yazım Hatalı Girdiler (Corrected Misspelled Entries). ANVA (%) = Ortalama Özel Kimlik Doğrulama Başarımı. AOT, bilgisayar donanımına bağlıdır. Tüm deneyler, aşağıdaki özelliklere sahip bir bilgisayarda gerçekleştirilmiştir: Microsoft Windows 11 Pro, 10.0.22631 Build 22631, HP EliteBook 640 14 inç G9 Notebook PC, 12. Nesil Intel(R) Core (TM) i7-1255U, 1700 Mhz, 10 Çekirdek, 12 Mantıksal İşlemci.

Çizelge 6.4, önerilen Transformer tabanlı semantik doküman doğrulama yaklaşımını farklı sayıda belge üzerinde, farklı belge konularıyla ve anlamsal doğrulama doğruluğunu ölçmek için farklı yöntemler kullanan (Simhash veya Cross Encoder metin benzerliği algoritmalarını kullanarak) test eden deneylerin sonuçlarını sunmaktadır.

Sonuçlar, her doküman konusu için (Spor, İş ve Finans), doküman sayısı 10'dan 1000'e çıktıkça, düzeltilen yazım hatalarının sayısının da arttığını göstermiştir. Bu da sistemin yanlış yazılan kelimeleri tespit etme ve düzeltme konusunda ne kadar etkili olduğunun bir göstergesidir.

Çizelge 6.4 ayrıca Simhash ve Cross Encoder metin benzerliği algoritmaları kullanılarak ortalama semantik doğrulama doğruluğunun, doküman konusuna ve sayısına bağlı olarak değiştiğini, her ikisinin de ilişkili olduğunu, ancak ikincisinin daha düşük metrik değerleri verdiğini göstermiştir. Örneğin, Spor belgeleri için Simhash kullanılarak ortalama anlamsal doğrulama doğruluğu %82,30 (10 belge için) ile %80,10 (1000 belge için) arasında değişirken Cross Encoder kullanılarak ortalama anlamsal doğrulama doğruluğu %79,21 (10 belge için) %76,58'e (1000 belge için) arasında değişmektedir. Önerilen yaklaşımın daha küçük veri kümelerinde daha büyük veri kümelerine göre daha iyi performans gösterdiği sonucuna varılmıştır.

Şekil 6.1'deki grafikte Deney I'de doküman türü ve anlamsal benzerlik algoritması bazında elde edilen ortalama doküman doğrulama doğruluğu verilmiştir.



Şekil 6.1 Deney I'de doküman türü ve anlamsal benzerlik algoritması bazında elde edilen ortalama doküman doğrulama doğruluğu

6.2.2 Deney II: Yazım Denetleyicisi (Spell Checker) Olmadan

Bu deneyde, Yazım denetleyicisi sürecine dahil olmayan 10, 100 ve 1000 Reuters finans, iş ve spor belgesi veri kümeleri üzerinde Transformer tabanlı anlamsal

doküman doğrulaması gerçekleştirilmiştir. Farklı konu ve sayılara sahip her veri setinde doğrulama doğruluğu ve ortalama işlem süresi hesaplanmıştır. Doğrulama sürecinde elde edilen doğruluk oranları Çizelge 6.5'te verilmiştir.

Çizelge 6.5 Transformer mimarisi (Yazım denetleyicisi olmadan) vasıtasıyla otomatik doküman doğrulama doğruluğu sonuçları

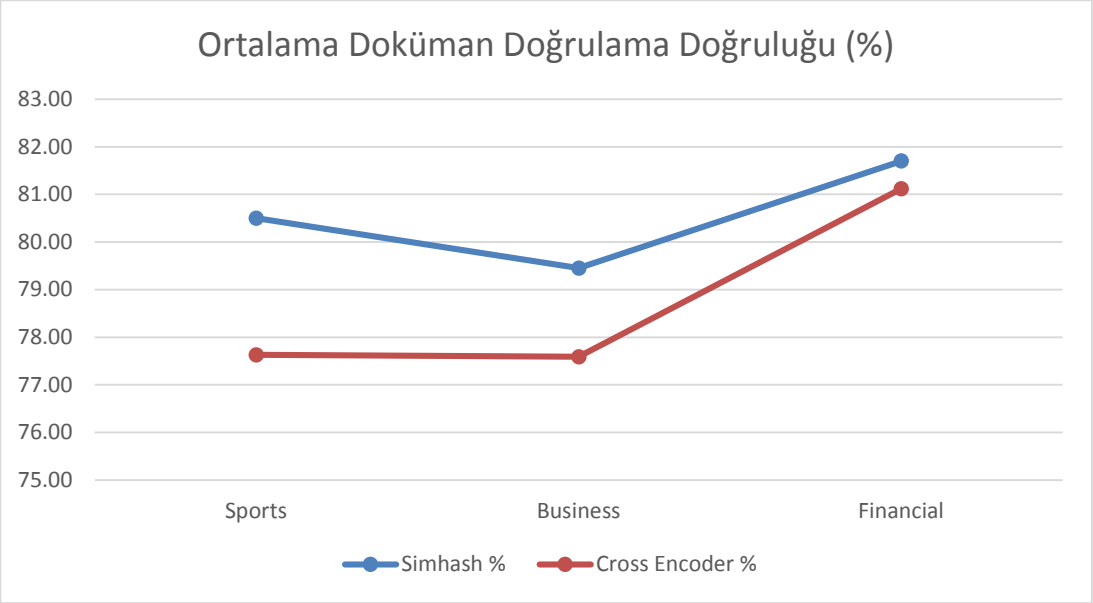
DS	**DK**	**ASVA Simhash (%)**	**ASVA Cross Encoder (%)**	**ANVA**	**AOT Simhash (sn)**	**AOT Cross Encoder (sn)**
10	Spor	82.10	80.14	85.0	2.4	2.9
100	Spor	80.02	77.79	82.15	14.8	20.1
1000	Spor	79.40	74.96	80.50	193.6	236.8
10	İş	81.67	79.70	86.50	2.6	3.0
100	İş	80.09	78.08	84.25	16.3	22.5
1000	İş	76.60	75.01	82.0	199.8	220.0
10	Finans	82.89	81.11	89.50	2.1	2.9
100	Finans	82.01	80.09	86.25	15.3	21.2
1000	Finans	80.21	82.16	85.0	195.2	228.9

Not: Çizelge 6.5'te kullanılan kısaltmalar: DS= Doküman Sayısı, DK = Doküman Konusu, ASVA = Ortalama Anlamsal Doğrulama Başarımı (Average Semantic Verification Accuracy), AOT = Ortalama İşlem Süresi (Average Operation Time). ANVA (%) = Ortalama Özel Kimlik Doğrulama Başarımı. AOT, bilgisayar donanımına bağlıdır. Tüm deneyler, aşağıdaki özelliklere sahip bir bilgisayarda gerçekleştirilmiştir: Microsoft Windows 11 Pro, 10.0.22631 Build 22631, HP EliteBook 640 14 inç G9 Notebook PC, 12. Nesil Intel(R) Core (TM) i7-1255U, 1700 Mhz, 10 Çekirdek, 12 Mantıksal İşlemci.

Çizelge 6.5'teki sonuçlar anlamsal doğrulama doğruluğunun doküman sayısına ve konusuna bağlı olarak değiştiğini göstermiştir. Örneğin, Spor belgeleri için Simhash metin benzerliği algoritması kullanıldığında doğruluk %82,10 (10 belge için) ile %79,40 (1000 belge için) arasında değişirken Cross Encoder

kullanıldığında doğruluk %80,14 (10 belge için) %74,96'ya (1000 belge için) arasında değişmektedir.

Şekil 6.2'deki grafikte Deney II'de doküman türü ve anlamsal benzerlik algoritması bazında elde edilen ortalama doküman doğrulama doğruluğu verilmiştir.



Şekil 6.2 Deney II'de doküman türü ve anlamsal benzerlik algoritması bazında elde edilen ortalama doküman doğrulama doğruluğu

6.3 Transformer ve Cümle Gruplaması Yöntemi Vasıtasıyla Otomatik Soyut Özetleme

Bu bölümde Transformer ve cümle gruplama tabanlı otomatik doküman özetleme hibrit yaklaşımı Reuters veri seti üzerinde test edilmiştir. Dokümanlar veri kümesinde üç farklı kategoriye (İş, Spor ve Finans) aittir. Deneyin amacı, hibrit yaklaşımın farklı kategorilerdeki dokümanları özetlemedeki etkinliğini değerlendirmek ve doğru ve bilgilendirici özetler üretip üretemediğini belirlemektir.

Önerilen yaklaşım, finansal dokümanların özetlenmesinde doğruluk ve verimlilik açısından umut verici sonuçlar vermiştir. Önerilen yaklaşım, Transformer modelini kullanarak, her cümle grubu için soyutlayıcı yeni cümleler üretebildi; bu da daha bilgilendirici ve kısa özetlerin oluşumunu sağlamıştır. Ayrıca Simhash metin benzerliği algoritmasının kullanılması, benzer cümlelerin etkili bir şekilde gruplandırılmasına olanak tanıyarak özetlerin kalitesini daha da arttırmıştır. Deneyler özetlemenin doğruluğunun veri kümesi, doküman konusu, doküman sayısı ve Transformer modelinin eğitimi gibi çeşitli faktörlerden etkilendiğini göstermiştir. Ayrıca veri kümesinin boyutunu artırmanın ve modelin eğitiminin iyileştirilmesinin doğrulukta potansiyel olarak daha fazla iyileşmeye yol açabileceği de gözlemlenmiştir.

Bu tez çalışması kapsamında, Transformer ve cümle gruplandırmayı kullanarak otomatik doküman özetleme için önerilen yaklaşımın performansı üzerinde çeşitli faktörlerin etkisi değerlendirilmiştir. Öncelikle modeli eğitmek ve test etmek için kullanılan veri kümesinin performans üzerinde önemli bir etkiye sahip olduğu görülmüştür. Spesifik olarak, öncelikle finansal belgeleri içeren Reuters veri seti kullanılmış ve bu yaklaşımın, %93,2'lik ortalama soyutlama özetleme doğruluğu ile bu veri seti üzerinde en yüksek doğruluğa ulaştığı görülmüştür. Bu durum, bu yaklaşımın finansal dokümanlar için etkili olabileceğini göstermiştir.

Daha sonra doküman konusunun soyutlama doğruluğu üzerindeki etkisi değerlendirilmiştir. Önerilen yaklaşım üç farklı doküman konusu üzerinde test edilmiştir: iş, spor ve finans. Önerilen yaklaşımın finansal dokümanlarda en yüksek doğruluğu sağladığını, ardından iş ve sporun geldiği gözlenmiştir.

Ayrıca doküman sayısının özetleme doğruluğuna etkisi de araştırılmıştır. Önerilen yaklaşım üç doküman konusunun her biri için 10, 100 ve 1000 dokümanla test edilmiştir. Doküman sayısı arttıkça doğruluğun azaldığı tespit edilmiş, bu da yaklaşımın daha küçük doküman kümeleri için daha etkili olabileceğini, konu bütünlüğü bozulduğunda başarının azaldığı gözlemlenmiştir.

Önerilen Transformer yaklaşımıyla oluşturulan özetlerin doğruluğunu doğrulamak için Transformer modeliyle üretilen özet belgeler ile orijinal özet belgeler arasındaki ortalama özetleme doğruluğu da hesaplanmıştır. Tüm deneylerin sonuçları değerlendirildiğinde, önerilen Transformer yaklaşımıyla üretilen özetlerin doğruluğu daha yüksek olduğu gibi, orijinal özetlere oldukça benzer özetler de ortaya çıkmıştır.

Son olarak Transformer model eğitiminin özetleme doğruluğuna etkisi değerlendirilmiştir. Transformer modelinin hiper parametreleri deneylerle belirlenmiş ve modelin bu parametrelerle en yüksek doğruluğa ulaştığı görülmüştür.

6.3.1 Deney I: İş Doküman Türü

Bu deneyde Transformer ve cümle gruplandırma tabanlı otomatik doküman özetleme hibrit yaklaşımı 10, 100 ve 1000 adet iş konusu dokümanları üzerinde gerçekleştirilmiştir. Ortalama soyut özet üretme doğruluğu ve ortalama işlem süresi ayrıca hesaplanmıştır. Özetleme sürecinde elde edilen sonuçlar Çizelge 6.6'da verilmiştir.

Çizelge 6.6 Transformer ve cümle gruplandırma tabanlı otomatik doküman özetleme hibrit yaklaşımının iş doküman konusundaki ortalama özetleme doğruluğu

DS	**DK**	**TTASA**	**TOASA**	**OOASA**	**ANVA**	**AOT**
10	İş	86.97	89.09	83.56	89.65	3.9
100	İş	83.09	87.34	81.87	87.50	23.2
1000	İş	81.45	85.78	80.01	86.30	241.82

Not: Çizelge 6.6'da kullanılan kısaltmalar: DS= Doküman Sayısı, DK = Doküman Konusu, TTASA = Transformer ile Orijinal Tam Metin Dokümanlar Arasındaki Ortalama Özetleme Doğruluğu (%) TOASA = Transformer ile Orijinal Özet Dokümanlar Arasındaki Ortalama Özetleme Doğruluğu (%) OOASA = Orijinal

Tam Metin ile Orijinal Özet Dokümanlar Arasındaki Ortalama Özetleme Doğruluğu (%), ANVA (%) = Ortalama Özel Kimlik Doğrulama Başarımı.

6.3.2 Deney II: Spor Doküman Türü

Bu deneyde Transformer ve cümle gruplandırma tabanlı otomatik doküman özetleme hibrit yaklaşımı 10, 100 ve 1000 adet spor konusu dokümanları üzerinde gerçekleştirilmiştir. Ortalama soyut özet üretme doğruluğu ve ortalama işlem süresi ayrıca hesaplanmıştır. Özetleme sürecinde elde edilen sonuçlar Çizelge 6.7'de verilmiştir.

Çizelge 6.7 Transformer ve cümle gruplandırma tabanlı otomatik doküman özetleme hibrit yaklaşımının spor doküman konusundaki ortalama özetleme doğruluğu

DS	**DK**	**TTASA**	**TOASA**	**OOASA**	**ANVA**	**AOT**
10	Spor	81.02	84.32	79.65	85.75	3.8
100	Spor	79.98	82.81	77.07	84.80	22.4
1000	Spor	79.13	81.90	75.21	83.0	243.76

Not: Çizelge 6.7'de kullanılan kısaltmalar: DS= Doküman Sayısı, DK = Doküman Konusu, TTASA = Transformer ile Orijinal Tam Metin Dokümanlar Arasındaki Ortalama Özetleme Doğruluğu (%) TOASA = Transformer ile Orijinal Özet Dokümanlar Arasındaki Ortalama Özetleme Doğruluğu (%) OOASA = Orijinal Tam Metin ile Orijinal Özet Dokümanlar Arasındaki Ortalama Özetleme Doğruluğu (%), ANVA (%) = Ortalama Özel Kimlik Doğrulama Başarımı.

6.3.3 Deney III: Finans Doküman Türü

Bu deneyde Transformer ve cümle gruplandırma tabanlı otomatik doküman özetleme hibrit yaklaşımı 10, 100 ve 1000 adet finans konusu dokümanları üzerinde gerçekleştirilmiştir. Ortalama soyut özet üretme doğruluğu ve ortalama

işlem süresi ayrıca hesaplanmıştır. Özetleme sürecinde elde edilen sonuçlar Çizelge 6.8'de verilmiştir.

Çizelge 6.8 Transformer ve cümle gruplandırma tabanlı otomatik doküman özetleme hibrit yaklaşımının finans doküman konusundaki ortalama özetleme doğruluğu

DS	**DK**	**TTASA**	**TOASA**	**OOASA**	**ANVA**	**AOT**
10	Finans	93.20	95.73	88.10	93.0	3.8
100	Finans	91.12	93.56	87.01	91.75	22.4
1000	Finans	87.78	91.18	85.63	90.5	243.76

Not: Çizelge 6.8'de kullanılan kısaltmalar: DS= Doküman Sayısı, DK = Doküman Konusu, TTASA = Transformer ile Orijinal Tam Metin Dokümanlar Arasındaki Ortalama Özetleme Doğruluğu (%) TOASA = Transformer ile Orijinal Özet Dokümanlar Arasındaki Ortalama Özetleme Doğruluğu (%) OOASA = Orijinal Tam Metin ile Orijinal Özet Dokümanlar Arasındaki Ortalama Özetleme Doğruluğu (%), ANVA (%) = Ortalama Özel Kimlik Doğrulama Başarımı.

7. SONUÇ VE ÖNERİLER

Bu tez kapsamında, üç bölümden oluşan DDİ tabanlı otomatik doküman doğrulama altyapısı oluşturulmuştur. İlk olarak anlamsal analiz olmadan özel kimlik tespiti vasıtasıyla doküman doğrulama, ikinci olarak Transformer mimarisi kullanılarak anlamsal doküman doğrulama ve son olarak tüm bu sürecin otomatik hale getirilmesi amacıyla sisteme verilen orijinal tam metnin soyut özetinin sistem tarafından üretilmesi sağlayan özetleme aşamasıdır. Bu aşamaların her biri için deneysel çalışmalar yapılmış ve sonuçları önceki bölümlerde paylaşılmıştır. Ancak asıl odak noktamız anlamsal analiz olduğu için, özellikle finansal türdeki dokümanlar için üretilen özetin tutarlılığını NLP tekniklerini temel alarak anlamsal olarak kontrol eden Transformer tabanlı anlamsal doküman doğrulama çözümü ön plandadır.

Transformer tabanlı anlamsal doğrulamanın ardındaki fikir, orijinal tam metinli dokümandan elde edilen özet dokümanın, o orijinal tam metinli dokümanın anlamsal bütünlüğünü içerdiğini hesaplamaktır. Çalışmada doğrulanacak referans dokümanlar finansal türe ait olduğundan Reuters finansal veri seti ile eğitim yapılarak Transformer modeli oluşturulmuştur. Önerilen Transformer tabanlı anlamsal doğrulama yaklaşımının anlamsal doğrulama doğruluğunu test etmek için Reuters veri seti kullanılmıştır.

Deneysel sonuçlar bu yaklaşımın finansal alandaki etkinliğine işaret etmektedir. Beklendiği gibi, finansal veri kümeleriyle eğitilmiş Transformer modeli, iş veya spor dokümanlarına kıyasla finansal dokümanlar için daha yüksek anlamsal doğrulama doğruluğu sergilemiştir. Finansal veri kümeleri üzerinde göz ardı edilemeyecek ortalama %84,1 anlamsal doküman doğrulama doğruluğuna sahiptir ve bu doğruluk oranı, finans sektöründeki pratik uygulamalar için kullanışlı olduğunu göstermektedir.

Önerilen yaklaşım öncelikle finansal dokümanlara odaklanarak tasarlanmış olsa da genelleştirilmiş anlamsal doküman doğrulama modeli için sağlam bir temel sunmaktadır. Bu model, çeşitli konu ve alanlara göre ayarlanabilir veya

genişletilebilir durumdadır. Diğer DDİ teknikleri ve algoritmalarıyla bir araya getirilerek hem doğruluk hem de verimlilik açısından önemli iyileştirmeler yapma potansiyeli taşımaktadır.

Sunulan bulgular ve yeniliklerin ışığında, bu tez çalışması yalnızca anlamsal doküman doğrulama alanında önemli araştırma bulguları sunmakla kalmayıp, aynı zamanda özellikle finans sektöründe değerli bir emsal oluşturmaktadır. Önerilen DDİ tekniklerini Transformer tabanlı bir modelle bir araya getiren çalışma, metnin semantik katmanlarının derinliklerine inmekte ve doküman özetlerinin tutarlılığını sağlamak için iyileştirilmiş bir yaklaşım sunmaktadır. Bu arada, modeli Reuters finansal veri seti üzerinde eğitmek, modelin kesinliğini daha da arttırmakta ve finansal jargona hâkim olmasına olanak tanımaktadır.

Önerilen Transformer tabanlı yaklaşım, yalnızca doküman özetlerinin anlamsal bütünlüğünü geliştirme olanağı sağlamakla kalmayıp, aynı zamanda doküman doğrulamanın geleneksel sınırlarını aşan otomasyonlaşma ve kesinliği bünyesinde barındıran yeni araştırmalara ışık tutmaktadır. Bu durum, bu alanı bekleyen gelecekteki ilerlemelerin ve bu karmaşık süreci daha da iyileştirme ve geliştirme potansiyelinin de bir kanıtıdır.

Ayrıca bu çalışma, Transformer tabanlı bir doküman doğrulama yaklaşımını Transformer tabanlı soyutlayıcı özetleme yaklaşımı ile bütünleştirmiştir. Hibrit yaklaşım, doğru ve bilgilendirici soyut özetler oluşturmak için Transformer ve cümle gruplama tekniklerini kullanmıştır. Amaç, doğrulama ve soyut özet oluşturma süreçlerini uçtan uca bir sistemde buluşturmaktır.

Transformer tabanlı anlamsal doğrulama ve soyut özetleme yaklaşımımız, oluşturulan özetlerin yalnızca anlamsal doğruluğunu değil, aynı zamanda yapısal bütünlüğünü de koruduğunu göstermektedir. Elde edilen yüksek ANVA (Ortalama Özel Kimlik Doğrulama Başarımı) değerleri, özetleme sürecinde özel kimliklerin doğru bir şekilde tespit edildiğini ve orijinal metindeki bilgilerle tutarlı bir şekilde aktarıldığını kanıtlamaktadır. Bu durum, sistemin hem anlamsal hem de yapısal bilgi kaybını önlediğini, kritik bilgilerin güvenilir bir

şekilde özetlerde yer aldığını açıkça ortaya koymaktadır. Özellikle finansal dokümanlarda gözlemlenen üstün ANVA başarımı, modelin hassas ve öngörülebilir veri yapılarıyla etkili bir şekilde çalıştığını göstermekte, önerilen yöntemin bilgi kaybını önleme ve doğruluk sağlamadaki yetkinliğini desteklemektedir.

Sonuç olarak, ANVA analizleri, Transformer tabanlı yaklaşımın yalnızca anlamsal doğrulukta değil, aynı zamanda yapısal bütünlüğü koruma açısından da güçlü bir doğruluk mekanizması sunduğunu ortaya koymaktadır. Bu bulgular, sistemin geniş bir uygulama yelpazesinde kullanılabilirliğini desteklemekte ve bilgi ile özel kimlik kaybını önleme konusunda gelecek çalışmalara sağlam bir temel sunmaktadır.

Bu tez çalışmasının bulgularına dayanarak, otomatik doküman doğrulama konusunda gelecekte yapılacak araştırmalar için bazı potansiyel öneriler şunlardır:

- Her özel kullanım durumu için en uygun hiper parametreleri bulmak amacıyla farklı veri kümeleri ile deneyler yapılabilir.
- Anlamsal doküman doğrulamanın doğruluğunu daha da artırmak için hızla gelişen farklı Transformer tabanlı modellerin kullanımı araştırılabilir.

- Özellikle finansal alanda manipüle edilmiş veya hileli dokümanları tespit etmek için daha karmaşık algoritmalar geliştirilebilir.

KAYNAKLAR

Aghajanyan, A., Yousefi-Azar, A., Cho, Y., 2021. PreSumm: A Transformer-based Model for Abstractive and Extractive Summarization, in IEEE Access, 9, pp. 6267-6279.

Aktan, E., 2018. Büyük Veri: Uygulama Alanları, Analitiği ve Güvenlik Boyutu Big Data: Application Areas, Analytics and Security Dimension. Bilgi Yönetimi. 1. 10.33721/by.403010.

Alambo, A., Banerjee, T., Thirunarayan, K., Cajita, M., 2022. Improving the Factual Accuracy of Abstractive Clinical Text Summarization using Multi-Objective Optimization, 2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), 1615-1618.

Almarimi, A., Andrejková, G., Sedmák, P., 2015. Document verification using n-grams and histograms of words, 2015 IEEE 13th International Scientific Conference on Informatics, 21-26.

Alpuente, M., Ojeda, P., Romero, D., Ballis, D., ve Falaschi, M., 2008. An Abstract Generic Framework for Web Site Verification, 2008 International Symposium on Applications and the Internet, 2008,104 – 110.

Alzahrani, S., & Mahmood, A. (2019). Hybrid textual and visual feature-based approach for authorship verification. IEEE Access, 7, 150515-150525.

Ando, T., Yatsu, H., Hisazumi, K., Fukuda, A., Matsumoto, M., Michiura, Y., 2015. Reference Model of Specifications Toward Independent Verification and Validation, TENCON 2015- 2015 IEEE Region 10 Conference, 2015,1 – 3.

Archive, 2023. Reuters 21578 Text Categorization Collection. Erişim Tarihi: 30.09.2023. https://archive.ics.uci.edu/ml/datasets/reuters-21578+text+categorization+collection

Attivissimo, F., Giaquinto, N., Scarpetta, M., Spadavecchia, M., 2019. An Automatic Reader of Identity Documents" 2019 IEEE International Conference on Systems, Man and Cybernetics (SMC), 3525-3530.

Bareedu, Y. S., Frühwirth, T., Niedermeier, C., Sabou, M., Steindl, G., Thuluva, A. S., Tsaneva, S., & Ozkaya, N. T., 2023. Deriving Semantic Validation Rules from Industrial Standards: an OPC UA Study, Semantic Web, 15(2), 517-554.

Bensefia, A., Paquet, T., Heutte, L., 2005. A writer identification and verification system, Pattern Recognition Letters, 26(1), 2080-2092.

Beusekom, J.V., and Shafait, F., 2011. Distortion Measurement for Automatic Document Verification, 2011 International Conference on Document Analysis and Recognition, 2011, pp. 289-293, doi: 10.1109/ICDAR.2011.66.

Bhargava, R., Sharma, Y., Sharma, G., 2016. ATSSI: Abstractive Text Summarization Using Sentiment Infusion, Procedia Computer Science, 89(1), 404-41.

Budanitsky, A., Hirst, G., 2001. Semantic distance in wordnet: An experimental, application-oriented evaluation of five measures, in: Workshop on WordNet and other Lexical Resources, 2(1), 2–2.

Charlow, K., Broyles, B., Stutzman, M., 2008. Standardized documentation for verification, validation, and accreditation (VV&A) — helping assure mission success, MILCOM 2008 IEEE Military Communications Conference, 2008, 1-7.

Christiane, F., George, M., 1998. Combining Local Context and Wordnet Similarity for Word Sense Identification, in WordNet: An Electronic Lexical Database, MIT Press, 265-283.

Chung, H.W., Chien-Lin H., Chin-Shun H., and Kuei-Ming L., 2007. Speech retrieval using spoken keyword extraction and semantic verification, TENCON 2007- 2007 IEEE Region 10 Conference, 2007, pp. 1-4, doi: 10.1109/TENCON.2007.4429138.

Devlin, J., Chang, M.W., Lee, K. and Toutanova, K., 2019. BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding, Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 1, 4171-4186.

Dilawari, A., Khan, M. U. G., Saleem, S., Zahoor, R., Shaikh, F. S., 2023. Neural Attention Model for Abstractive Text Summarization Using Linguistic Feature Space, in IEEE Access, 11(1), 23557-23564.

Dipsizkuyu, 2022. Erişim Tarihi: 15.12.2022. https://dipsizkuyu.net/semantik-analiz-nedir/

Erkan, G., Radev, D. R., 2004. LexRank: Graph-based Lexical Centrality as Salience in Text Summarization. Journal of Artificial Intelligence Research, 22, 457-479.

Etemad, A. G., Abidi, A. I., Chhabra, M., 2021. A Review on Abstractive Text Summarization Using Deep Learning, 2021 9th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), 1-6.

Garain, U. and Halder, B., 2008. On Automatic Authenticity Verification of Printed Security Documents, 2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing, 2008, pp. 706-713, doi: 10.1109/ICVGIP.2008.67.

Ghazali, O., Al-Maatouk, Q., Saleh, O., 2019. Graduation Certificate Verification Model: A Preliminary Study, International Journal of Advanced Computer Science and Applications, 10(7), 575-582.

Github, 2022. Erişim Tarihi: 09.06.2022. https://github.com/stefan-it/turkish-bert/files/4558187/nerdata.txt

Gupta, S., Meena, A. K., 2018. A Practical Implementation of Automatic Document Analysis and Verification using Tesseract [1] [2], 2018 International Conference on Computational Techniques, Electronics and Mechanical Systems (CTEMS), 401-406.

Gupta, P., Khatter, M. 2019. Text summarization: a survey of current approaches. International Journal of Computer Applications, 180(37), 9-15.

Guzzi, P., Mina, M., Guerra, C., Cannataro, M., 2011. Semantic similarity analysis of protein data: Assessment with biological features and issues, Briefings in Bioinformatics, 13(1), 569-85.

Han, K., 2023. A Survey on Vision Transformer" in IEEE Transactions on Pattern Analysis and Machine Intelligence, 45(1), 87-110.

Hariharan, S., Srinivasan, R., 2008. Investigations in single document summarization by extraction method, 2008 International Conference on Computing, Communication and Networking, 1-5.

Hassanpour, S., O'Connor, M.J., & Das, A.K. 2011. A Framework for the Automatic Extraction of Rules from Online Text. RuleML Europe.

Ho, T., Sung, K., 2014. Fingerprint-Based Near-Duplicate Document Detection with Applications to SNS Spam Detection. International Journal of Distributed Sensor Networks, 10, 1-8.

Hsu, W.P., 2020. Intelligent Document Recognition on Financial Process Automation, 2020 International Symposium on VLSI Design, Automation and Test (VLSI-DAT), 1-1.

Imam, I. T., Arafat, Y., Alam, K. S., Shahriyar, S. A., 2021. DOC-BLOCK: A Blockchain Based Authentication System for Digital Documents, 2021 Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV), Tirunelveli, 1262-1267.

Jamil, S., Jalil, M., Kwon, O., 2023. A Comprehensive Survey of Transformers for Computer Vision" Drones, 7(5), 287-297.

Janowicz, K., Martin, R., Kuhn, W., 2011. The semantics of similarity in geographic information retrieval, J. Spatial Information Science, 2(1), 29-57.

Jiang, J.J., Conrath, D.W., 1997. Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy, ROCLING/IJCLCLP, 19-33.

Jiang, Q., Sun, M., 2011. Semi-Supervised SimHash for Efficient Document Similarity Search. The 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, 19-24 June, Portland, Oregon, USA, 93-101.

Jiang, D., Cao, S., Yang, S., 2021. Abstractive summarization of long texts based on BERT and sequence-to-sequence model" 2021 2nd International Conference on Information Science and Education (ICISE-IE), 460-466.

Jin, H., Xiaojun, W., 2020. Abstractive Multi-Document Summarization via Joint Learning with Single-Document Summarization, Findings of the Association for Computational Linguistics: EMNLP 2020, 2545–2554.

Kaggle, 2022. Erişim Tarihi: 09.06.2022. https://www.kaggle.com/pariza/bbc-news-summary

Kim, S., Fiorini, N., Wilbur, W., Zhiyong, L., 2017. Bridging the gap: Incorporating a semantic similarity measure for effectively mapping PubMed queries to documents, Journal of Biomedical Informatics, 75(5), 122-127.

Kim, Y., 2014. Convolutional Neural Networks for Sentence Classification, Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, 1746–1751.

Ladwani, V. M., Ramasubramanian, V., 2022. M-ary Hopfield neural network for storage and retrieval of variable length sequences: Multi-limit cycle approach, 2022 IEEE Symposium Series on Computational Intelligence (SSCI), 436-441.

Lee, S., Belkasim, S., Zhang, Y., 2013. Multi-document Text Summarization Using Topic Model and Fuzzy Logic, In: Perner, P. (eds) Machine Learning and Data Mining in Pattern Recognition. MLDM 2013. Lecture Notes in Computer Science, 7988, 159–168.

Li, X., Xu, X., ve Malik, T., 2016. Interactive provenance summaries for reproducible science, 2016 IEEE 12th International Conference on e-Science (e-Science), 2016, 355 – 360.

Li, J., Huang, G., Chen J., Wang, Y., 2019. Short Text Understanding Combining Text Conceptualization and Transformer Embedding, in IEEE Access, 7(1), 122183-122191.

Liaqat, M. I., Hamid, I., Nawaz, Q., Shafique, N., 2022. Abstractive Text Summarization Using Hybrid Technique of Summarization, 2022 14th International Conference on Communication Software and Networks (ICCSN), 141-144.

Liu, X., Xv, L., 2019. Abstract Summarization Based on the Combination of Transformer and LSTM, 2019 International Conference on Intelligent Computing, Automation and Systems (ICICAS), Chongqing, 923-927.

Liu, S., Li, D., Chen, Y., Yang, G., 2021. (Semi) automatic Assertion Generation from Con-trolled Chinese Natural Language: A Practice in Aerospace Industry, 2021 IEEE 21st Inter-national Conference on Software Quality, Reliability and Security Companion (QRS-C), 778-782.

Mao, X., Shaobin, H., Shen, L., Li, R., Yang, H., 2021. Single document summarization using the information from documents with the same topic, Knowledge-Based Systems, 228(1), 107265-107285.

Medium, 2022. Erişim Tarihi: 09.06.2022. https://medium.com/codable/named-entity-recognition-varlık-ismi-tanıma-b21315a30029

Mihalcea, R., Corley, C., Strapparava, C., 2006. Corpus-based and Knowledge-based Measures of Text Semantic Similarity, Proceedings of the National Conference on Artificial Intelligence, 775–780.

Miller, T., Biemann, C., Zesch, T., Gurevych, I., 2012. Using Distributional Similarity for Lexical Expansion in Knowledge-based Word Sense Disambiguation, 24th International Conference on Computational Linguistics - Proceedings of COLING 2012: Technical Papers, 1781-1796.

Moawad, I. F., Aref, M., 2012. Semantic graph reduction approach for abstractive Text Summarization, 2012 Seventh International Conference on Computer Engineering & Systems (ICCES),132-138.

Mohamed, M., Oussalah, M., 2019. SRL-ESA-TextSum: A text summarization approach based on semantic role labeling and explicit semantic analysis, Information Processing & Management, 56(1), 1356-1372.

Morocho, D., Morales, A., Fierrez J., Vera-Rodriguez, R., 2017. Human-Assisted Signature Recognition Based on Comparative Attributes, 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), 5-9.

Mu, Y., Wang, Y., ve Guo, J., 2009. Extracting Software Functional Requirements from Free Text Documents, 2009 International Conference on Information and Multimedia Technology, 2009, 194 – 198.

Naik, S. S., Gaonkar, M. N., 2017. Extractive text summarization by feature-based sentence ex-traction using rule-based concept, 2017 2nd IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT), 1364-136.

Novarica, 2022. Erişim Tarihi: 09.06.2022. https://aite-novarica.com/report/global-document-identification-and-verification-market-overview

Patwardhan, N., Marrone, S., Sansone, C., 2023. Transformers in the Real World: A Survey on NLP Applications, Information, 14(4), 242-259.

Pedrycz, W., Gacek, A., Wang, X., 2022. A Hierarchical Approach to Interpretability of TS Rule-Based Models, in IEEE Transactions on Fuzzy Systems, 30(8), 2861-2869.

Pi, B., Fu, S., Wang, W., Han, S., Inc, R., China, P., 2019. SimHash-based Effective and Efficient Detecting of Near-Duplicate Short Messages. Proceedings of the Second Symposium International Computer Science and Computational Technology (ISCSCT '09), 26-28 December, Huangshan, P. R. China, 20-25.

Pss, S., Bhaskar, M., 2016. RFID and pose invariant face verification based automated classroom attendance system, 2016 International Conference on Microelectronics, Computing and Communications (MicroCom), 1-6.

Pu, H., Fei, G., Zhao, H., Hu, G., Jiao, C., Xu, Z., 2017. Short Text Similarity Calculation Using Semantic Information, 2017 3rd International Conference on Big Data Computing and Communications (BIGCOM), 144-150.

Reddy, J. M., Prasad, S. V. A. V., 2016. The role of verification and validation in software testing, 2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom), 1298-1301.

Rehman, S., Khan, S., & Faye, I. (2017). Graph-based authorship verification. In Proceedings of the 15th International Conference on Frontiers in Handwriting Recognition (pp. 413-418).

Resnik, P., 1995. Using information content to evaluate semantic similarity in a taxonomy, In Proceedings of the 14th international joint conference on Artificial intelligence, 1, 448–453.

Roychoudhury, S., Bellarykar, N., and Kulkarni, V., 2016. A NLP Based Framework to Support Document Verification-as-a-Service, 2016 IEEE 20th International Enterprise Distributed Object Computing Conference (EDOC), 2016, pp. 1-10, doi: 10.1109/EDOC.2016.7579376.

Rush, A., Chopra, S., Weston, J., 2015. A Neural Attention Model for Abstractive Sentence Summarization, 10(1), 379–389.

Sampaio, P. N. M., Santos, C. A. S., and Courtias, J. P., 2000. About the semantic verification of SMIL documents, 2000 IEEE International Conference on Multimedia and Expo. ICME2000. Proceedings. Latest Advances in the Fast

Changing World of Multimedia (Cat. No.00TH8532), 2000, pp. 1675-1678 vol.3, doi: 10.1109/ICME.2000.871093.

Shinyama, Y., Sekine, S., Sudo, K., 2002. Automatic Paraphrase Acquisition from News Articles, In Proceedings of NAACL-HLT, 313–318.

Shirwandkar, N. S., Kulkarni, S., 2018. Extractive Text Summarization Using Deep Learning, 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA), 1-5.

Spacy, 2022. Erişim Tarihi: 09.06.2022. https://spacy.io/api/annotation#named-entities

Steingrímsson, S., Guðnason, J., Helgadóttir, S., Rögnvaldsson, E., 2017. Málrómur: A Manually Verified Corpus of Recorded Icelandic Speech, 21st Nordic Conference on Computational Linguistics, Göteborg, Sweden, 237-240.

Stone, C., Chothia, T., Garcia, F., 2017. Spinner: Semi-Automatic Detection of Pinning without Hostname Verification, ACSAC 17: Proceedings of the 33rd Annual Computer Security Applications Conference, 176-188.

Suleiman, D., Awajan, A. A., 2019. Deep Learning Based Extractive Text Summarization: Approaches, Datasets and Evaluation Measures, 2019 Sixth International Conference on Social Networks Analysis, Management and Security (SNAMS), 204-210.

Sunderrajan, S., Manjunath, B. S., "Context-Aware Hypergraph Modeling for Re-identification and Summarization," in IEEE Transactions on Multimedia, vol. 18, no. 1, pp. 51-63, Jan. 2016, doi: 10.1109/TMM.2015.2496139.

Takata, Y., Nakamura, T., Seki, H., 2004. Accessibility verification of WWW documents by an automatic guideline verification tool, 37th Annual Hawaii International Conference on System Sciences, 10-20.

Tanasa, D., Trousse, B., 2004. Advanced data preprocessing for intersites web usage mining, IEEE Intelligent Systems, 19(2), 59–65.

Tay, Y., Mostafa, D., Bahri, D., Donald M., 2022. Efficient Transformers: A Survey, ACM Comput. Surv, 55(6), 28-40.

Thitisathienkul, P., ve Prompoon, N., 2015. Quality Assessment Method for Software Requirements Specifications Based on Document Characteristics and Its Structure, 2015 Second International Conference on Trustworthy Systems and Their Applications, 2015, 51 – 60.

Thongkor, K., Pramoun, T., Chaisri, C., ve Amornraksa T., 2012. Integrity verification method of Thai content for faxed document, 9th International

Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology, 2012, 1-4.

Turney, P., 2001. Mining the web for synonyms: PMI-IR versus LSA on TOEFL. In Proceedings of the Twelfth European Conference on Machine Learning (ECML-2001), 491–502.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I., 2017. Attention Is All You Need, 1-15.

Wang, J., 2011. Web-Based Verification on the Representativeness of Terms Extracted from Single Short Documents, 2011 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology, 2011, pp. 114-117, doi: 10.1109/WI-IAT.2011.258.

Wang, X., Liu, Y., Xiong, F., 2016. Improved personalized recommendation based on a similarity network, Physica A: Statistical Mechanics and its Applications, 456(1), 271-280.

Wang, J., Li, J., Huang, Q., & Zhang, C., 2019. Deep learning-based semantic document verification. IEEE Transactions on Information Forensics and Security, 14(1), 68-80.

Wang, R., Zhang, H., Lu, G., Lyu, L., Lyu, C., 2020. Fret: Functional Reinforced Transformer With BERT for Code Summarization, in IEEE Access, 8(1), 135591-135604.

Weaviate, 2023. Using Cross-Encoders as reranker in multistage vector search. Erişim Tarihi: 30.09.2023. https://weaviate.io/blog/cross-encoders-as-reranker.

Wei, T., Zhou, Q., Chang, H., Lu, Y., Bao, X., 2014. A Semantic Approach for Text Clustering using WordNet and Lexical Chains. Expert Systems with Applications, 42(4), 264-2275.

Wessels, T., Omlin, C.W., 2000. A hybrid system for signature verification, Proceedings of the IEEE-INNS-ENNS International Joint Conference on Neural Networks. IJCNN 2000. Neural Computing: New Challenges and Perspectives for the New Millennium, 2000, 5(1), 509-514.

Wu, Z., Palmer, M., 1994. Verb semantics and lexical selection. In Proceedings of the Annual Meeting of the Association for Computational Linguistics, 133-138.

Xu, L., 2023. Deep Transfer Learning Model for Semantic Address Matching" Applied Sciences, 12(19), 10110-10130.

Yang, T., Wu, S., Feng, J., Fu, N., Tian, M., 2019. Semantic Network Based Approach to Compute Term Semantic Similarity, 2019 3rd International Conference

on Electronic Information Technology and Computer Engineering (EITCE), 654-658.

Zhai, Y., 2023. ByteTransformer: A High-Performance Transformer Boosted for Variable-Length Inputs, 2023 IEEE International Parallel and Distributed Processing Symposium (IPDPS), 344-355.

Zhang, S., Zhu, Y., Roy-Chowdhury, A. K., 2016. Context-Aware Surveillance Video Summarization, in IEEE Transactions on Image Processing, 25(11), 5469-5478.

Zhang, J., Zhao, Y., Saleh, M., Liu, P., 2019. PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization, 1051, 11328-11339.

Zhi, X., Zhao, B., Wang, Y., 2021. A Hybrid Framework for Text Recognition Used in Commodity Futures Document Verification" 2021 6th International Conference on Computational Intelligence and Applications (ICCIA), 143-148.

Zhu, G., Iglesias, C.A., 2017. Computing semantic similarity of concepts in knowledge graphs, IEEE Trans. Knowl. Data Eng, 29(1), 72–85.

Zeng, X., Lu, Y., & Zhu, J. (2018). A deep learning-based approach for semantic document verification. In Proceedings of the 27th International Conference on Computational Linguistics (pp. 2837-2847).

TEZ SÜRECİNDE ÜRETİLMİŞ YAYINLAR

Toprak, A., Turan, M. 2023. Enhanced Named Entity Recognition algorithm for financial document verification. J Supercomput, 79, 19431–19451.

Toprak, A., Turan, M. 2024. Transformer-Based Approach for Automatic Semantic Financial Document Verification. IEEE Access, 12, 184327-184349.

Toprak, A., Turan, M. 2024. Automated thematic dictionary creation using the web based on WordNet, Spacy, and Simhash. Data and Information Management, 100088, 1–14.