

T.C.
KASTAMONU ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ
MALZEME BİLİMİ VE MÜHENDİSLİĞİ ANA BİLİM DALI



DOĞAL DİL İŞLEME ALGORİTMASI İLE KALIPLARA
DAYANARAK ARAPÇA MORFOLOJİK ANALİZİN
OLUŞTURULMASI

ABDULMONEM ALI ABDULSALAM AHMED

DOKTORA TEZİ

PROF. DR. AYBABA HAÇERLİOĞULLARI

HAZİRAN - 2024

KASTAMONU

TEZ ONAYI

ABDULMONEM ALI ABDULSALAM AHMED tarafından hazırlanan “**DOĞAL DİL İŞLEME ALGORİTMASI İLE KALIPLARA DAYANARAK ARAPÇA MORFOLOJİK ANALİZİN OLUŞTURULMASI**“ adlı tez çalışmasının savunma sınavı **13.06.2024** tarihinde yapılmış olup aşağıda verilen jüri tarafından oy birliği ile Kastamonu Üniversitesi Fen Bilimleri Enstitüsü **Malzeme Bilimi ve Mühendisliği Ana Bilim Dalı Doktora Tezi** olarak kabul edilmiştir.

Danışman	Prof. Dr. Aybaba HANÇERLİOĞULLARI Kastamonu Üniversitesi	
Jüri Üyesi	Prof. Dr. Fatih Vehbi ÇELEBİ Yıldırım Beyazıt Üniversitesi	
Jüri Üyesi	Doç. Dr. M. Serdar ÇAVUŞ Kastamonu Üniversitesi	
Jüri Üyesi	Dr. Öğrt. Üyesi Ümit TOKEŞER Kastamonu Üniversitesi	
Jüri Üyesi	Dr. Öğrt. Üyesi Melisa RAHEBİ İstanbul Topkapı Üniversitesi	

Jüri üyeleri tarafından kabul edilmiş olan bu tez Kastamonu Üniversitesi Fen Bilimleri Enstitüsü Yönetim Kurulunca onanmıştır.

Enstitü Müdürü	Doç. Dr. Selçuk MEMİŞ	
----------------	-----------------------	-------

TAAHHÜTNAME

Bu tezin tasarımı, hazırlanması, yürütülmesi, araştırmalarının yapılması ve bulgularının analizlerinde bütün bilgilerin etik davranış ve akademik kurallar çerçevesinde elde edilerek sunulduğunu; ayrıca tez yazım kurallarına uygun olarak hazırlanan bu çalışmada bana ait olmayan her türlü ifade ve bilginin kaynağına eksiksiz atıf yapıldığını, bilimsel etiğe uygun olarak kaynak gösterildiğini bildirir ve taahhüt ederim.

Abdulmonem Ali ABDULSALAM AHMED

ÖZET

DOKTORA TEZİ

DOĞAL DİL İŞLEME ALGORİTMASI İLE KALIPLARA DAYANARAK ARAPÇA MORFOLOJİK ANALİZİN OLUŞTURULMASI

ABDULMONEM ALI ABDULSALAM AHMED

KASTAMONU ÜNİVERSİTESİ FEN BİLİMLERİ ENSTİTÜSÜ
MALZEME BİLİMİ VE MÜHENDİSLİĞİ ANA BİLİM DALI
DANIŞMAN: PROF. DR. AYBABA HANÇERLİOĞULLARI

Morfolojik analiz, metin kategorizasyonu, makine çevirisi, indeksleme, bilgi erişimi ve çok daha fazlasını içeren doğal dil İşleme uygulamalarının büyük çoğunluğunda gerekli bir adımdır. Arapça, karmaşık kelime yapısına sahip, türetme açısından zengin, derin anlamlara sahip bir dildir. Bu yüzden kelimeleri bileşenlerine ayırmak karmaşık ve yoğun bir çalışma gerektirir. Arapça kelimeler köklerden ve kalıplardan oluşur. Arapça, çekimleri çok olan ve karmaşık bir morfolojik yapıya sahip bir dildir, hemen hemen tüm kelimelerin köklerden kalıplarla türediği, çok zengin bir türetme morfolojisine sahip bir türetme dilidir. Dilin son derece spesifik kalıpları olmasına rağmen, her türlü Arapça kelimeyi barındıracak çok sayıda kalıp oluşturulmuş ve genişletilmiştir. Kökleri belirlemek için kalıplar kullanıldıktan sonra kelimelerin morfolojik yapıları belirlenebilir veya bu kalıplardan kelime veya kök oluşturulabilir. Bu morfolojik analiz yöntemi kökleri, isimleri, fiilleri ve her türlü çoğul ve tekili tanıyabilir ve ayrıca herhangi bir kelime için tüm çekimleri oluşturabilir. Bu tez çalışmasında iki algoritma yöntemi ilk kez uygulanmıştır. Birincisi, herhangi bir dil kuralına dayanmamasıdır, bu da orijinal harfler ile kelimeye eklenen harfler arasında herhangi bir karışıklık olasılığını ortadan kaldırır; ikincisi, tüm kelime ön eklerini, son eklerini ve ara eklerini korur, bu da işlem süresinin kısılmasını sağlar. Son olarak önerilen algoritmalar, performans metrikleri kullanılarak değerlendirilmiş ve doğruluk %95, geri çağırma %98 ve F1 puanı %97 olarak elde edilmiştir.

ANAHTAR KELİMELE: Arapça, morfoloji, kök, kalıp, kök çıkarma, Algoritma

Haziran 2024, 73 Sayfa

ABSTRACT

PH.D THESIS

BUILDING ARABIC MORPHOLOGICAL ANALYSIS WITH NATURAL LANGUAGE PROCESSING ALGORITHM BASED ON PATTERNS

ABDULMONEM ALO ABDULSALAM AHMED

KASTAMONU UNIVERSITY INSTITUTE OF SCIENCE

DEPARTMENT OF MECHANICAL ENGINEERING

SUPERVISOR: PROF. DR. AYBABA HANÇERLİOĞULLARI

Morphological analysis is a necessary step in the great majority of Natural Language Processing (NLP) applications, including text categorization, machine translation, indexing, information retrieval, and many more. Arabic is a deep-meaning language with a complex word structure that is rich in derivatives. That's why it requires a lot of work to cope with. Arabic words are constructed from roots and patterns. Arabic is a language with a lot of inflections and a complicated morphological structure, it is a derivational language with a very rich derivational morphology, with nearly all words deriving from roots through patterns. We have created and extended numerous patterns to accommodate all kinds of Arabic words, even though the language has highly specific patterns. After patterns are used to identify roots, words' morphological structures can be ascertained, or words or stems can be formed from these patterns. This morphological analyzer can recognize roots, nouns, verbs, and all types of plurals and singulars, and it can also generate all inflections for any word. In this thesis proposed two methods; first, it does not rely on any linguistic rules, which eliminates any possibility of confusion between the original letters and the letters added to the word; second, it retains all word affixes, suffixes, and infixes, resulting in a shorter processing time. Finally, the algorithm was evaluated using performance metrics and 0.94 for Precision, 0.97 for Recall, 0.95 for F1-score, and 0.92 for Accuracy were obtained. This shows that the algorithm has a high level of performance.

KEYWORDS:Arabic, Morphology, Root, Pattern, Stemming, Algorithm

June 2024, 73 Page

TEŞEKKÜR

Öncelikle çalışmam boyunca bana destek olan ve yardımlarını esirgemeyen danışmanım Prof. Dr. Aybaba Hañçerlioğulları'na teşekkür ederim. Ayrıca eş danışmanım Dr. Öğrt. Üyesi Ali Rıza Tosun hocama simülasyon ve programlama çalışmalarında sürekli yanımda olduğu için teşekkür ediyorum. Tezimin ilerlemesinde bana her zaman destek olan, değerli bilgiler veren tez izleme komitemdeki hocalarıma şükran ve takdirlerimi sunarım. Ayrıca öğrenimim boyunca bana destek olan tüm arkadaşlarıma ve maddi desteklerinden dolayı Libya hükümetine özel teşekkürlerimi sunmak isterim. Çalışmalarım boyunca desteklerini esirgemeyen aileme teşekkür ederim.

ABDULMONEM ALİ ABDULSALAM AHMED

Kastamonu, 2024

İÇİNDEKİLER

	<u>Sayfa</u>
TEZ ONAYI	ii
TAAHHÜTNAME	iii
ÖZET.....	iv
ABSTRACT	v
TEŞEKKÜR	vi
İÇİNDEKİLER	vii
ŞEKİLLER DİZİNİ	ix
TABLolar DİZİNİ	x
SİMGELER VE KISALTMALAR DİZİNİ	xi
1. GİRİŞ.....	1
1.1 Tezin Amacı.....	3
1.2 Hedefler.....	4
1.3 Katkılar	4
2. LİTERATÜR TARAMASI	5
2.1 Arapça Dilinin Arka Planı	5
2.2 Arapçaya Genel Bakış.....	7
2.3 Arapça Morfolojisi.....	8
2.3.1 Türetme	10
2.3.2 Çekim.....	11
2.4 Arap Sisteminin Zorlukları	11
2.4.1 Aksan İşaretleri	12
2.4.2 Büyük Harf Eksikliği	12
2.4.3 Çekimli Dil	13
2.5 Kök Çıkarma.....	13
2.6 Arapça Kökler.....	14
2.7 Arapça Kalıplar	15
2.8 Kısa Ünlü veya Aksan	16
2.9 Arapçada Çoğulluk	16
2.9.1 Arapça Düzenli Çoğul	17
2.9.1.1 Arapça eril düzenli çoğul	17
2.9.1.2 Arapça dişil düzenli çoğul.....	17
2.9.2 Arapça Düzensiz Çoğul	17
2.10 İsimlendirilmiş Varlık Tanıma.....	18
2.10.1 İsimlendirilmiş Varlık Tanıma Uygulamaları.....	18
2.10.2 Arapçada İsimleştirilmiş Varlık Tanıma Zorlukları	20
2.11 Arapça Fiiller	21
2.12 Arapça Doğal Dil İşleme	22
2.12.1 Düzgünleştirme	23
2.12.2 Bölütleme.....	24
2.12.3 İsimlendirilmiş Varlık Tanıma Çıkarımı	24
2.12.4 Kelime Etiketleme Tanımlaması	25
2.12.5 Arapça Gereksiz Kelimelerin Kaldırılması.....	25
2.13 İlgili Çalışmalar	25
2.13.1 Kök Tespiti ile İlgili Çalışmalar	26

2.13.2	İsimlendirilmiş Varlık Tanımayla İlgili Çalışmalar.....	27
2.13.3	Fiillerle İlgili Çalışmalar.....	29
2.13.4	Çoğullarla İlgili Çalışmalar	31
2.13.5	Kelime Üretme ile İlgili Çalışmalar.....	33
2.13.6	Özet.....	35
3.	METODOLOJİ VE YÖNTEMLER.....	36
3.1	Ön İşleme Aşaması	37
3.1.1	Bölütleme Süreci.....	38
3.1.2	Düzgünleştirme Süreci.....	38
3.1.3	Gereksiz Kelimeleri Kaldırma Süreci.....	39
3.2	Tasarım Aşaması.....	40
3.2.1	Kök Tespiti Algoritması	41
3.2.2	Check Pattern Fonksiyonu	42
3.3	Kalıp Modeli	43
3.3.1	Arapça Gereksiz Kelimelerin Kaldırılması.....	46
3.3.2	Fiil Kalıpları.....	46
3.3.3	Varlık İsmi Kalıpları	47
3.4	Özet.....	47
4.	DEĞERLENDİRME VE DENEYSEL SONUÇLAR	49
4.1	Doğal Dil İşleme Algoritması Metrikleri.....	49
4.1.1	Veri Kümesi.....	52
4.3	Örnek Üçüncü Düzey Alt Başlık	53
4.4	Deney Sonuçları.....	54
4.4.1	Birinci Örnek	55
4.4.2	İkinci Örnek	55
4.4.3	Üçüncü Örnek	56
4.4.4	Dördüncü Örnek	57
5.	SONUÇ VE ÖNERİLER	59
5.1	Öneriler	60
KAYNAKLAR		61
EKLER.....		71
EK A. Terminoloji.....		72
ÖZGEÇMİŞ.....		73

ŞEKİLLER DİZİNİ

Sayfa

Şekil 3.1 Önerilen modelin şeması.....	37
Şekil 3.2 Bölütleme fonksiyonu.....	38
Şekil 3.3 Düzgünleştirme fonksiyonu.....	39
Şekil 3.4 Gereksiz kelimeleri kaldırma fonksiyonu.....	40
Şekil 3.5 Kök tespit algoritması.....	41
Şekil 3.6 Check pattern fonksiyonu.....	43
Şekil 4.1 Karışıklık Matrisi.....	50
Şekil 4.2 Dokümanların kategorisel dağılımı.....	53
Şekil 4.3 Birinci örnek.....	55
Şekil 4.4 Örnek iki.....	56
Şekil 4.5 Üçüncü örnek.....	57
Şekil 4.6 Dördüncü örnek.....	58

TABLÖLAR DİZİNİ

	<u>Sayfa</u>
Tablo 3.1 Kökleri tespit etmek için kullanılan kalıplar.....	44
Tablo 3.2 Eril düzenli çoğul kalıpları	46
Tablo 3.3 Dişil düzenli çoğul kalıpları.....	46
Tablo 3.4 Düzensiz çoğul kalıpları	46
Tablo 3.5 Fiiller için kalıplar	47
Tablo 3.6 Varlık isimlerine ilişkin kalıplar	47
Tablo 4.1 Karışıklık matrisi parametreleri	50
Tablo 4.2 Veri kümesinin istatistiği.....	52
Tablo 4.3 Parametre değerleri	53
Tablo 4.4 Metrik hesaplamaları	54

SİMGELER VE KISALTMALAR DİZİNİ

Kısaltmalar

AI	: Yapay Zekâ (Artificial Intelligence)
AMIR	: Arapça Morfoloji Bilgi Erişimi
FN	: Yanlış Negatif (False negative)
FP	: Yanlış Pozitif (False positive)
GUI	: Grafik Kullanıcı Arayüzünde
LSTM	: Uzun-Kısa Süreli Bellek (Long / Short Term Memory)
MSA	: Modern Standart Arapça
RNN	: Tekrarlayan Sinir Ağı (Recurrent Neural Network)
TN	: Gerçek Negatif (True Negative)
TP	: Gerçek Pozitif (True Positive)
vb.	: ve benzeri
vd.	: ve diğerleri

1. GİRİŞ

Arapça, milyonlarca anadil olarak konuşanı ile dünyada en çok konuşulan beş dilden biri ve Birleşmiş Milletler'in altı resmi dilinden biridir (Ameur vd., 2020). Ayrıca internette en yaygın kullanılan dördüncü dil olarak kabul edilmektedir (Guellil vd., 2021). Arapça'yı diğer dillerden ayıran, zengin ve derin bir morfolojiye sahip olmasıdır. Arapça'dan ve Arapçaya bilgisayarlı çeviri, diğer dil bileşenlerine Arapça morfolojisinin eklenmesi nedeniyle çok daha zor hale gelmiştir. Arapça, karmaşık morfolojisiyle bilinmektedir (Tahat, 2020). Tek bir kelimenin çeşitli anlamlara gelmesi sebebiyle, belirsizlik karakteri ile ünlüdür (Abdallah vd., 2012; Zayed ve El-Beltagy, 2012), ayrıca Arapça yazarken ses simgeleri (aksan işaretleri) kullanılabilmektedir. Sesli harflerde herhangi bir harf işareti yoktur ve Arapça kelimelerdeki her harfte "hareketler" anlamına gelen bir işaret (haraka) bulunur (Mansouri vd., 2008; Chiu ve Nichols, 2016). Bunlar "Fathah", "Kasrah", "Dammah", "Sukun", "Şeddah" ve "Tenvin" olarak adlandırılır. Arap dili güçlü bir yapıya sahip olup çekim ve türetme derecesi en yüksek dil olarak kabul edilmektedir. Arapça, hata teşhisini ve geri bildirimi zorlaştıran bazı morfolojik özelliklere sahiptir (Alasmari vd., 2017; Shaalan vd., 2015; Ray ve Shaalan, 2016). Arapça morfolojisinin değerlendirilmesi ihtiyacı ve morfolojinin makine analiz sistemlerinde kullanılması, Arap dilinin şu anda ilgilendiği iki konudur. Sonuç olarak, kelime analizi ve bilgi erişimi, birçok Doğal Dil İşleme uygulamasında kullanılan önemli adımlardır (Muhammed vd., 2018). Kelimelerin iç yapısıyla ilgilenmeye morfoloji denir; bu analiz, ekleştirme davranışı, kökler ve kalıp özellikleri dahil olmak üzere kelime yapısının incelenmesine yardımcı olur (Elmadany vd., 2015; Attia vd., 2011). Çekimsel ve türetme morfolojisi, morfolojiyi kategorize etmek için kullanılan iki temel fikirdir. Çekimsel morfolojide, belirli bir formasyona sahip bir kelime gövdesine çekim uygulanır. Kelimenin isim, fiil vb. gramer sınıfı etkilenmez. Çekimden etkilenebilecek özellikler arasında büyük/küçük harf, cinsiyet, sayı, zaman, kişi, durum ve ses yer alır. Öte yandan, türetme morfolojisi belirli bir kelimeye sözdizimsel kategorisini değiştirebilecek bir grup morfem ekler (Farghaly ve Shaalan, 2009). İnternette bulunabilecek bilgi miktarı son zamanlarda önemli ölçüde artmıştır (Alshalabi vd., 2013). Sıradan arama motorlarının yanıtlaması zor olan soruları

cevaplayabilmesi için mevcut anlamsal web araştırması bu bilgiye erişmeyi amaçlamaktadır. Bilgi çıkarımı, metindeki öğeler arasındaki ilişkileri otomatik olarak belirleyerek, yapılandırılmamış bir metinden yapılandırılmış veriler çıkarmaya yönelik bir tekniktir (Alshalabi vd., 2013). İlişki çıkarımı, bir metindeki iki element arasındaki bağlantıyı belirleme tekniğidir. Bu elementler, ana ve yan elementler olarak ikiye ayrılır (Zakria vd., 2019). Metinden bilgi çıkarımının ana amacı ilişki çıkarımıdır. Arapça cümleler, eylemsel veya isimsel olmak üzere iki kategoriden birine dahildir (Zakria vd., 2019). Arapça üç ana çeşide ayrılır: Edebi materyallerde ve Kuran'da kullanılan Klasik Arapça; yazılı ve resmi konuşmalarda kullanılan Modern Standart Arapça ve gündelik konuşmalarda ve günlük etkileşimlerde kullanılan Arapça Lehçeler (Guellil vd., 2021). Cümlelerin ve kelimelerin morfolojik özelliklerinden dolayı metnin sözdizimi ve anlambilimi zorlayıcıdır (Ismail vd., 2017). Arap dili zengindir ve diğer diller arasında en karmaşık dil olarak kabul edilir. Cümle karmaşıklığı hem sözdizimi hem de anlambilimi olarak yüksektir. Aynı zamanda eş anlamlı kelimeler, zıt anlamlılar ve isim veya fiil olarak kelime kökleri de dahil olmak üzere birçok kelime hazinesi içerir (Othman vd., 2020). Halen var olan Arapça metinlerin çoğunluğu resmi Modern Standart Arapça ile yazılmıştır, ancak daha fazla insan Facebook, Google ve Twitter gibi sosyal medya sitelerini kullandıkça, resmi olmayan, günlük konuşma dilindeki metinlerin miktarı da artmaktadır (Hegazi vd., 2021). Önceki sistemler, varlık tanıma sorununu çözmek için ayrıştırıcılar ve etiketleyicilerin yanı sıra çok sayıda dizin ve bazı durumlarda devasa eğitim setleri kullanmaktaydı. Öte yandan, günlük Arapça yazılmış metinlerden isimlendirilmiş öğeleri çıkarmanın zorluğu, günlük Arapça'nın ayrıştırıcılarının yeterince doğru olmaması nedeniyle halihazırda kullanımda olan yöntemleri geçersiz kılmaktadır. Bunun nedeni, günlük Arapçanın kuralsızlığı, eksikliği ve sözcük dizilimindeki düzensizliktir. Konuşma dili Arapçası halk arasında konuşulan bir dil olduğu için yerleşik kurallara veya gramer yapısına sahip değildir (Zayed ve El-Beltagy, 2012). Bu konuyla ilgili daha önceki çalışmaların çoğunluğu İngilizce, Fransızca veya Çince olarak üretilmemiştir. Bu konuda Arapça ile ilgili çok fazla kitap da bulunmamaktadır. Ayrıca Arapçanın pek çok ayırt edici özelliği olması nedeniyle metinden ilişkileri tespit etmek zor olabilir. Örneğin aksan işaretleri bir ifadenin anlamını değiştirir. Arapçanın işlenmesi başka zorlukları da beraberinde getirir (Zakria vd., 2019). Arapça kelimelerin çoğunluğunun üç harfli (üçlü) kökü vardır. Ayrıca teorik olarak her üçlü

kökten farklı fiil zamanları oluşturulabilmektedir. Arapça günlük konuşma edatlardan, artikellerden, bağlaçlardan, zamirlerden, sıfatlardan ve zarflardan oluşur. Bu araştırma Arapça metinlerde fiilin tespitine yönelik olduğundan fiilin tanımı, çeşitleri ve türevleri üzerinde durulacaktır. Fiil, belirli bir anda bir durumu veya bir eylemin meydana geldiğini ifade eden bir kelimedir; zamana bağlı olan ve şu anda, geçmişte veya gelecekte gerçekleşebilecek bir eylemi gösterir (Siswoyo, 2016). Bu çalışmada, kökler sözlüğünden bağımsız olan ve sorunları çözmek için morfolojik analiz yaparken kuralları kullanmayan bir morfolojik analizör oluşturulmuştur. Önerilen model, karakterleri diğer uyarlanmış elementlerden ayırırken, karakterlerden özgün fiil kökleri çıkarma konusunda yeteneklidir. Fiilleri, isimleri, çoğulları ve diğer dil yapılarını tanımlamak için bu çalışma, köklerin çıkarılmasına yönelik yöntemlerden yararlanan kök kalıbına dayalı bir strateji sunmaktadır. Analiz edilen kelimenin yalnızca yüzeysel kalıpları kullanılır. Arapça dil işleme araştırmalarının çoğunluğu kelimelerden ekleri ayırtmak için bir teknik kullanmaktadır. Yöntem, uygulama sırasında herhangi bir eki kaldırmaz. Bu tezdeki yaklaşım, gramer veya makine öğrenimi uygulamaları gerektirmemektedir. Bunun yerine yaklaşım, kelimeyi bulunana kadar tekrar tekrar eşit uzunluktaki kalıplarla karşılaştırır. Kalıpları kullanarak kelimenin bir kökü olup olmadığını belirlemekle başlar. İlerleyen bölümlerde süreç daha ayrıntılı olarak ele alınacaktır.

1.1 Tezin Amacı

Tezin amacı, Arap morfolojisinin karmaşıklıklarına takılıp kalmayan bir Morfolojik Analizör oluşturmaktır. Hedef, Arapça dil işleme üzerinde çalışan geliştiricilerin ve araştırmacıların, dilin karmaşık özelliklerini öğrenmek zorunda kalmadan uygulamalarını oluşturmalarını kolaylaştıracak kütüphaneler ve bir sistem geliştirmeye başlamaktır. Temel ve önemli araçlar sisteme dahil edilecektir. Önerilen algoritma kökleri tanımlamak için kalıpları kullanır; Arapça kelimelerin morfolojik yapısı, kökü belirlemek için kullanılan kalıplar sayesinde tanınır. Kullanılan hiçbir kelimeden hiçbir ek çıkarılmamıştır, bu nedenle kelimenin temel harflerinin ve eklerinin ne olduğu her zaman belirlenebilecektir. Oluşturulan sistem arama motorlarına, veritabanı motorlarına, bilgi çıkarma uygulamalarına ve Arapça dil işlemeyi kullanan diğer yapay zekâ uygulamalarına hitap edecektir.

1.2 Hedefler

Arapçaya özel morfoloji çalışmalarını içeren yepyeni bir strateji oluşturulmuştur; önerilen yöntem, öncelikle kalıp tanıma yoluyla kökü tanımlayarak çalışmaktadır. Kullanılan kalıp çok önemlidir çünkü kelimenin tüm morfolojik yapılarını oluşturmak ve anlamak için kullanılır. Önerilen algoritma hiçbir son eki, ön eki veya iç eki hariç tutmaz; bunun yerine, kelimenin ve kalıbın harflerini doğrudan karşılaştırır; bu, kalıbın kökü tanımlayıp tanımlayamayacağını belirlemeye yardımcı olur. Ayrıca algoritmada kural tabanlı yaklaşımlar kullanılamaz. İşlem süresinin kısaltılması bunların arkasındaki itici güçtür. Hem kelimenin hem de kalıbın ana ve ek harfleri arasında oluşabilecek karışıklığı ortadan kaldırmak için, modelin orijinal harfleri Latin harfleriyle değiştirilmiştir.

1.3 Katkılar

Birinci bölümde tezin konusu hakkında giriş niteliğinde bilgiler verilmektedir. Arapçanın açıklanmasıyla başlar ve tezin amacı ve motivasyonu ile biter. İkinci bölüm, Arap dilinin ve doğal dil işlemenin temel terminolojilerini ve bu alanlardaki ilgili çalışmalar hakkındaki bilgileri içermektedir. Üçüncü bölümde bu çalışmada kullanılan algoritmaların tasarım metodolojileri ve yöntemleri gösterilmiştir. Dördüncü bölüm sonuçları tartışır ve değerlendirme sürecine genel bir bakış sunar ve beşinci bölümde bu araştırmanın sonucu ve öneriler yer almaktadır.

2. LİTERATÜR TARAMASI

Bu bölümde Arapça ile doğrudan ilişkisi olan temel kavramlara değinileceği gibi, bu tezde konuyla doğrudan ilişkili olan önceki çalışmalardan bazıları da sunulacaktır.

2.1 Arapça Dilinin Arka Planı

Arapça dünya çapında ve internette en çok konuşulan ve kullanılan dördüncü dildir. Aynı zamanda dünyada bilinen en eski konuşulan dillerden biridir (Guellil vd., 2021). Yirmi iki ülkenin resmi dilidir ve dünya çapında yaklaşık dört yüz milyon insan tarafından kullanılmaktadır. Üç ana lehçeye ayrılmıştır: Arapça diyalektiği, Modern Standart Arapça ve Klasik Arapça. Günlük iletişimde, gündelik alışverişlerde vb. diyalektik kullanılırken, yazma ve resmi etkileşimler için modern standart Arapça, edebi metinlerde ve Kuran’da ise klasik Arapça kullanılır (Guellil vd., 2021). Hem türetme hem de çekim açısından zengin morfolojisi ve karmaşık dil yapısı nedeniyle Arapça, son derece zor bir dil olarak kabul edilmektedir (Azmi ve Alshenaifi, 2017; El Ouardi, 2022; Hanani ve Naser, 2020). Arapçada türetilmiş bir kelime, bir grup kök kelimedenden kaynak alarak oluşur. Arapça, kök kalıp şemalarına dayanan türetilmiş bir dildir (El-Defrawy vd., 2017). Bu, kelimelerin üç veya daha fazla ünsüz kullanarak, geniş bir anlam veya kavramı ifade eden köklerden geldiği anlamına gelir. Birkaç morfolojik kalıp modeline bağlı kalırlar (Azmi ve Alshenaifi, 2017; El-Defrawy vd., 2017; Ismail vd., 2017). İsimler, sıfatlar ve fiiller, kalıplar kullanılarak köklerden türetilir. Örneğin (فعل) kalıbı, (درس) kökünü içerir. Bu kalıbın (مفعول) kullanılması, “öğretmen” anlamına gelen (مدرس) kelimesini verir; (فعلول) kalıbı kullanıldığında “dersler” anlamına gelen (دروس) kelimesi elde edilir; (مفعلة) kalıbının kullanılması “okul” anlamına gelen (مدرسة) kelimesini verir; (فعالة) kalıbının uygulanması “çalışma” anlamına gelen (دراسة) kelimesini verir ve bu şekilde devam eder. Burada aynı kökten farklı kelimelerin türetilmesinin yalnızca dört örneği sunulmuştur. Aynı kökten türeyen bu kadar çok terimin anlamları her zaman aynı olmamaktadır. Geleneksel Arapça sözlüklerin köke göre düzenlenmesi nedeniyle kök, herhangi bir morfolojik çalışmanın çok önemli bir bileşenidir (Alazzawie, 2020). Bir kök, birkaç kalıp ve birkaç ekten onlarca, yüzlerce kelime üretilebildiği göz önüne alındığında Arapça,

morfolojik açıdan en karmaşık ve zengin dildir. Ünlülerin yokluğu ve eklenen karakterlerin kök veya kök harflere benzerliği gibi çeşitli faktörler nedeniyle Arapça önemli derecede belirsizlik sergiler (Ameur vd., 2020). Kelimeler geleneksel gramerciler tarafından üç kategoriye ayrılmıştır: fiiller, isimler ve edatlar (Alnaied vd., 2020). Morfolojinin analizi zorlu bir iştir. Bu durum kelimelerde bulunan eklerin çokluğuyla olduğu kadar doğrudan kökten türetilen kelimelerin çokluğuyla da açıklanabilir. Farklı aksan işaretlerinin aynı kelime için farklı anlamlara sahip olması morfoloji açısından başka bir zorluktur (Alnaied vd., 2020). Aksan işareti, bir harfe karşılık gelen sesi değiştirmek veya onu aynı harflerden oluşan diğer kelimelerden ayırmak için kullanılan bir işarettir. Arapça'da aksan işaretleri, morfolojik yapı, gramer işlevi, kelimelerin anlamsal içeriği ve ses ve dil özellikleri dahil olmak üzere dilin çeşitli elementlerini belirlemek için kullanılır (Chennoufi ve Mazroui, 2017). Fonetikte kök; son ek veya ön eklerin kökte birleştirilmesiyle oluşturulan fiilleri, sıfatları ve isimleri içeren ekli türevlerin kaynağını oluşturan en temel kelimedir. Arapça isim ve fiiller genellikle bir kök ailesinden gelir. Arapça isimler cinsiyet (dişil veya eril), çoğulluk (çoğul, tekil veya ikili) ve gramatik vurgu (şimdiki, geçmiş, gelecek ve emir) dahil olmak üzere bazı özelliklere bağlı olarak herhangi bir biçime sahip olabilir. Öte yandan kök çıkarma, kelimeleri kök karşılıklarına indirgeyen bir bilgisayar yöntemini tanımlamaktadır (Salloum ve Habash, 2022). Arapçada kelimeler genellikle bir kalıp kullanılarak kökten türetilir. Daha fazla kelime oluşturmak için bir köke ekler eklenebilir ve kökün kendisi de çeşitli kalıplarda birçok kelime üretir. Doğal diller belirsizliğe neden olabilecek çekim yapılarına sahip olduğundan, bunları incelemek yazarlar, çevirmenler ve hatta konuşanlar için bile zorlayıcı olabilir. İnsanlar sezgilidir ve bağlamdan anlam çıkarabilir; algoritmalar ve bilgisayar programları ise bu başaramaz ve birçok hata yapabilir. Tüm bu sebepler ve daha fazlası nedeniyle araştırmacılar bu alanda araç ve uygulama üretmek için sürekli olarak çalışmalarını geliştirmek ve ilerletmek için çalışmaktadırlar. İnternette mevcut olan veri miktarı gün geçtikçe artmakta ve bu da otomatik analizi önemli hale getirmektedir. Burada kök çıkarma adı verilen ve kelimelerin kök veya morfolojik köklerine ayrıştırılmasını içeren çok önemli bir prosedür bulunmaktadır (Al-Serhan ve Ayes, 2020; Salloum ve Habash, 2022). Birçok Doğal Dil İşleme uygulamasında önemli ön işleme aşamalarından biri, türetilmiş sözcüklerin temel biçimlerine döndürülmesi süreci olan kök çıkarmadır (Alhaj vd., 2019; Mustafa vd., 2019). Uygun belgelerin

toplanması sürecini kolaylaştırmak için kullanılsa da kök çıkarma basit bir işlem değildir ve önemi artmaktadır (Alkhatib vd., 2020; Jaafar vd., 2017; Naili vd., 2019; Salloum ve Habash, 2022). Kök çıkarma süreci bir gövde veya kök verir. Bilgi erişimi de dahil olmak üzere diğer alanlarda da sorguları genişletmek ve artan erişim ve doğrulukla en uygun verileri elde etmek için gövde ve kök kullanılır. Az kök çıkarma, kelimedenden eklerin (ön ekler ve son ekler) çıkartılarak gövde elde edilmesi yöntemidir; çok kök çıkarma ise asıl kökü elde etme işlemidir (Azmi ve Alshenaifi, 2017; El Bazzi vd., 2016). Ekleri çıkararak veya çıkarmadan kök çıkarma işlemini tamamlamak mümkündür. Arapça kelimeler miktarsal olarak üçe ayrılır; tekil, ikili ve çoğul, ayrıca iki cinsiyete ayrılır; eril ve dişil (Othman vd., 2020). Ön ekler, son ekler ve iç ekler Arapçadaki eklerin örnekleridir. Bir kelimenin başı ön eklerin eklendiği, sonu son eklerin eklendiği ve merkezi ise iç eklerin bulunduğu yerdir. Kök çıkarma, Arapça’da bir kelimenin orijinal kökünü eklerinden ayırma işlemidir. Arapça bir cümle oluştururken kullanılan üç öge isim, fiil ve edattır. İsim ve fiiller, yaklaşık on bin adet kök içeren bir kapalı kümenin elemanlarından oluşur. Kökler nadiren beş harften oluşur, genellikle üç veya dört harften oluşur. Uygun kökü belirlemek için çoğu araştırmacı, ekleri silerek kökleri çıkarmaya yönelik algoritmalar ve teknikler kullanır (Salloum ve Habash, 2022).

2.2 Arapçaya Genel Bakış

Arapçada sağdan sola yazılan yirmi sekiz harf vardır. Daha fazla şekil, işaret ve sesli harf eklenerek olası harf sayısı doksana çıkarılabilir (Oueslati vd., 2020). Arapçada çok fazla çekim bulunur. Daha fazla zamir ve edat (bağlaçlar, edatlar vb.) ekleme sürecine ek olarak Arapça, cinsiyet, sayı, kişi, görünüş, ruh hali, durum, konum ve sesi içeren kapsamlı bir çekim morfolojisi sistemine sahiptir (Obeid vd., 2020). Arapçanın zengin morfolojisi ve karmaşık dil yapısı onu öğrenmeyi zor bir dil haline getirmektedir (Hanani ve Naser, 2020). Üç harfli (üç ünsüz harfli) fiiller, çoğu yerel Arapça ismin kaynağıdır. Bunun yanı sıra, yerel Arapça kelimelerin küçük bir kısmı dört, beş veya altı harfli köklere sahiptir (Al-Kabi, 2013). Arapça kelimelerin çoğunluğu önceden belirlenmiş kurallar veya yönergeler kullanılarak köklerden türetilmiştir. Köklere ekler getirilerek ve gramer kurallarına bağlı kalarak, şablon olarak kabul edilen bu kalıplar uygulanır. İyelik ekleri, çoğullar, biçimler, cinsiyet vb.

gibi diğer gramer kullanımlarını elde etmek için ekler (ön ekler, iç ekler ve son ekler) eklenebilir (Boubas vd., 2011). Arapça'da üç tür kelime vardır: fiil, isim ve edat (Ismail vd., 2017). Arapçanın çok sayıda güçlüğü vardır ve bu tezde bunlardan üçü tartışılacaktır. Öncelikle bir kelimenin metinde geçtiği yere göre değişen birden fazla anlamı varsa kelime çok anlamlı olabilir. Bunun bir örneği (عقد) kelimesidir, birbirinden farklı altı anlamı bulunur. İkinci olarak, aynı kökü paylaşan ancak farklı anlamlara sahip olan kelimeler vardır; (حسب) hesaplamak, (حاسوب) bilgisayar ve (محاسبة) muhasebe anlamına gelir. Son olarak, aynı kökü paylaşan ancak farklı anlamlara sahip olan kelimeler de bulunur. Bazı kelimeler birden fazla köke sahip olabilir. Arapçada (ي), (ا) ve (و) harfleri uzun sesli harfler, geri kalan yirmi beş harf ise sessiz harflerdir. Ayrıca Arapçada harf olmayan kısa ünlüler de vardır; bunlar aksan işaretleri olarak bilinir. Bir kelimedeki harfler bu işaretlere sahip olduğunda anlam açısından daha az belirsiz hale gelir ve dil bilgisinin kontrolü ve doğru kökün belirlenmesi kolaylaşır (Goweder vd., 2004; Hassan, 2015). Arapça kök takımlarına dayandığı için dildeki bütün isim ve fiillerin kökeni bu köklerden gelir. Kökler dört kategoriye ayrılabilir: beş karakterli kökler, üç karakterli kökler, dört karakterli kökler ve türevi olmayan iki karakterli kökler (Hassan, 2015). Herhangi bir kelimenin gövdesini veya kökünü belirleme yöntemi, kök çıkarma olarak bilinir ve bilgi erişimi de dahil olmak üzere çeşitli alanlarda kullanılır (Taylor ve Brychcín, 2018). Arapça benzersiz nitelikler içerdiğinden, çıkarma işleminin başlayabilmesi için ön işleminin tamamlanması gerekir. Ön işleme, verileri bir sonraki metin işleme adımına hazır olacak şekilde dönüştürme işlemidir. Hesaplamaları hızlandırmak, doğruluğu ve verimliliği artırmak için kullanılır. Bölütleme, düzgünleştirme, gereksiz kelimelerin çıkarılması ve Arapça olmayan kelimelerin, sayıların, rakamların ve noktalama işaretlerinin kaldırılması, ön işleme adımının aşamalarıdır (Al-Nashashibi vd., 2010).

2.3 Arapça Morfolojisi

Morfoloji bilimi, bir kelimenin farklı biçimlerde farklı anlamlara sahip olacak şekilde nasıl değiştirilebileceğini açıklar. Dilbilimsel morfolojiye göre kurallar, yeni kelimelerin ve kelime türlerinin oluşturulmasını yönetir ve aynı zamanda sözlükte zaten var olan karmaşık kelimeler için yedek kurallar olarak da hizmet eder (Sawalha vd., 2013). Durum analizi, bir kelimenin iç yapısıyla ilgilenme eylemidir. Kalıp

nitelikleri ve ekleme davranışının nedenleri gibi unsurları ele alarak kelime formunun yönetilmesine yardımcı olur (Alazzawie, 2020). Bu en temel katmanın üzerine daha yüksek sözdizimsel ve anlamsal katmanlar inşa edilir. Bir terimin doğal dildeki asıl anlamını belirleme sürecine morfolojik analiz denir. Doğal dil işleme uygulamalarının çoğunluğu, önemli bir ilk adım olan morfolojik analizle başlar (Abouenour vd., 2008). Kelimeler geleneksel gramerciler tarafından üç kategoriye ayrılır: fiiller, isimler ve edatlar (Alnaied vd., 2020). Arapçanın morfolojik analizinin karmaşıklığı, kelimelerdeki eklerin çokluğuna ve aynı zamanda doğrudan kökten türetilen sözcüklerin çokluğuna bağlanabilir. Farklı aksan işaretlerinin aynı kelime için farklı anlamlara sahip olması morfoloji açısından başka bir zorluktur (Alnaied vd., 2020). Aksan işareti, bir harfe karşılık gelen sesi değiştirmek veya onu aynı harflerden oluşan diğer kelimelerden ayırmak için kullanılan bir işarettir. Arapça'da aksan işaretleri, morfolojik yapı, gramer işlevi, kelimelerin anlamsal yapısı ve ses ve dil özellikleri dahil olmak üzere dilin çeşitli yönlerini açıklamak için kullanılır (Chennoufi ve Mazroui, 2017). Fonetikte kök, ön eklerin veya son eklerin kökte birleştirilmesiyle oluşturulan fiiller, sıfatlar ve isimleri içeren ek türevlerin temelini oluşturan en temel kelimedir. Bu tür yöntemlerle olası bir kök bulabilmek için ön işleme uygulamak gereklidir (Thalji vd., 2018). Kök çıkarma, yalnızca saf dil analizi için değil, aynı zamanda konuşmadaki çeşitli cümle, amaç ve etki argümanlarının belirlenmesi ve genel olarak daha gelişmiş doğal dil kavrama sistemlerinin geliştirilmesi için de hayati öneme sahiptir (Azman, 2019). Arapça isim ve fiiller genellikle bir kök ailesinden gelir. İsimler, cinsiyet (dişil veya eril), çoğulluk (çoğul, tekil veya ikili) ve gramatik vurgu (şimdiki, geçmiş ve emir) dahil olmak üzere birçok özelliğe bağlı olarak herhangi bir biçime sahip olabilir (Abdelkader vd., 2019). Öte yandan kök çıkarma, kelimeleri kök karşılıklarına indirgemek için kullanılan bir bilgisayar yöntemini tanımlamaktadır (Abouenour vd., 2008). Arapçada kelimeler genellikle bir kalıp kullanılarak kökten türetilir. Daha fazla kelime oluşturmak için bir köke ekler eklenebilir ve kökün kendisi de çeşitli kalıplarda birçok kelime üretir. Kök ve gövde çıkarımı için veri tabanı aramaları, istatistiksel analizler ve kurallara dayalı teknikler olarak üç yöntem vardır. En basit yaklaşım, yalnızca arama tablosunda kelimenin kökünü arayan veri tabanı arama yöntemidir. Çeşitli kelimelere karşılık gelen ve kökleri ayırt etmek için kullanılabilecek bir kalıp koleksiyonu da veri tabanına dahil edilmiştir. Bilinen çalışmaların çoğu bu şekilde yapılmıştır. Kelimenin kökünü ve

kalıbı oluşturmak için algoritma kullanılması önerilmiştir. Sorunlar genellikle veri tabanında kalıp ve kök eşleşemediğinde ortaya çıkar. Bu yöntemin dezavantajı veri tabanının düzenli olarak güncellenmesinin gerekmesidir. Ayrıca algoritma belirli bir kelime için birden fazla kalıp tanımlayabilir. İkinci yöntem Ağırılık Esaslı Yaklaşım denir; bu yöntemde algoritma, kökü belirlemek için matematiksel hesaplamalar yapmadan önce sözcükteki her harfe farklı bir ağırlık verir. Bu algoritmanın temel kusuru, fazlalık harflere orijinal harflerle aynı önceliği atamasıdır. Son strateji ise, günümüz literatüründe büyük ölçüde kullanılan Kural Tabanlı Yaklaşımdır. Bu algoritmada kelimenin Arapça kökünü belirlemek için bir kurallar sistemi oluşturulmuştur. Bu yöntem eklerden ve kalıplardan oluşan bir veri tabanından da yararlanır. Kuralların hem miktarı hem de sırası bu algoritmada önemli etkiye sahiptir. Bu algoritma ayrıca potansiyel bir kökü tanımlamak için bir ön işleme adımını da içerecektir (Thalji vd., 2018). Genel olarak Arapça kelimeler bir kalıp kullanılarak köklerinden türetilmiştir. Aynı kökten farklı gramer yapılarına sahip çok sayıda kelime meydana gelir ve ekler eklenerek ilave kelimeler oluşturulabilir. Arap morfolojisi çok sistematiktir ve açıklanan kavramlarla bağlantılı yöntemleri türetme ve çekim olmak üzere iki modele dönüştürür.

2.3.1 Türetme

Arapça morfolojisinde, ünsüz köklerden oluşan bir sisteme dayalı olarak morfemlerin birbirine bağlanması yoluyla yeni kelimeler oluşur (Thalji vd., 2018). Köklerde ünlüler de mevcut olabilir ve bunlar da kökün temel bileşenleri olarak kabul edilecektir. Genel olarak Arapçada bir kök veya kalıbın tek başına kullanılması mümkün değildir; kelimeler oluşturmak için birbirleriyle etkileşime girmeleri gerekir (Thalji vd., 2018). Bir kökten veya daha temel bir sözcük biriminden ek sözcük birimleri oluşturma işlemine türetme denir. Türetilmiş sözcük birimi ile temel sözcük birimi veya kök arasındaki bağlantıyı bunlardan biri ya da her ikisi ile açıklanabilir: Morfolojik kategoride değişim; fiilden türeyen isim, sıfattan türeyen fiil, isimden türeyen zarf vb.: Ve anlam kaymaları; geçişsizden geçişli hale gelmesi, nedenselin geçişli hale gelmesi ve olumsuzun pozitif hale gelmesi vb. Bazı durumlarda, anlam veya kategoride (önemli) bir değişiklik olmasa bile, türetilmiş sözcük biriminin temelden farklı olduğu bilinmektedir. İngilizce bir örnek olarak, her ikisi de “daireler halinde hareket etmek”

veya "düzenli aralıklarla meydana gelmek" anlamına gelen "cyclic" ve "cyclical" sıfatları verilebilir. Öte yandan "cyclical", düzenli olarak tekrarlamayı ifade eder. Bu örnekte "-al" bir türetme morfudur çünkü bunlar farklı sözcük birimlerdir ve her ikisi de sözlüğe dahil edilmiştir. Yeni bir kelime üreten herhangi bir prosedüre türetme denir. Çekim morfolojisi genellikle bir dilin tarihsel evrelerini temsil edebilen türetimsel morfolojiden daha verimlidir. Ek olarak, çekimsel morfolojiyle karşılaştırıldığında türetimsel morfoloji genellikle daha fazla istisna gösterir.

2.3.2 Çekim

Bir kelime bir cümlede kullanılırken fonolojik modifikasyona uğradığında, kelimenin temel anlamı değişmeden bırakılsa bile oluşan değişiklikler "çekim" olarak ifade edilir (Habash, 2007). Arapça kelimelerdeki çekim sürecini göstermek için, genellikle bu kelimelere gramerik kategoriler atanır. Arapçanın sekiz ana gramer kategorisi vardır: zaman/görünüş, kişi, ses, mod, cinsiyet, sayı, durum ve kesinlik (Thalji vd., 2018). Dilbilgisel anlamları işaretlenmiş bir bütün sözcük, bu kategorilerin bir köke uygulanmasıyla ortaya çıkacaktır. Çekim olarak bilinen kelime yapısı, kelimenin kategorisini değiştirmeden veya yeni sözcükbirimler yaratmadan, belirli bir anlam içinde bir sözlüğün uygun gramer formunu sunmaya hizmet eder. Çekim morfolojisinin bir başka özelliği de genellikle daha az istisnaya sahip olması ve türetme morfolojisinden daha üretken olmasıdır. Arapçadaki türetme morfolojisinin birleşik olmayan bir temel kullanması daha muhtemelken, çekimsel morfolojinin birleşik bir temel kullanması daha olasıdır. Arapça hem birleşik hem de birleşik olmayan morfoloji sergiler.

2.4 Arap Sisteminin Zorlukları

Arapça, iki temel özelliği olan kendine özgü bir yazı dilidir: sağdan sola doğru yazılır ve etrafındaki harfe bağlı olarak karakter biçimi değişir (Alsayat ve Elmitwally, 2020; Cheikh ve Laffet, 2017). Bunlar dilin karşılaştığı sayısız zorluktan sadece ikisidir. Yüksek kaliteli Arapça morfolojisi, bilgi kazanımı ve makine çevirisi de dahil olmak üzere birçok doğal dil işleme uygulamasında kritik bir seviye olarak kabul edilirken, bunu analiz etmek ilk önemli engeli oluşturur. Morfolojinin makine çeviri

sistemlerinde uygulanması ikinci problemin konusunu oluşturmaktadır. Üçüncü konu ise morfolojik analizin genellikle sözdizimsel analizden önce yapılmasıdır (Muhammed vd., 2018). Arapça zengin bir dil mirasına sahiptir. Diğer diller için oluşturulan düzenli doğal dil işleme sistemleri bu zenginliği yönetememektedir; bunun yerine özel muamele gereklidir. Arapça için sistemler oluşturmak zordur. Dilin kendine has özellikleri, uygun araçların bulunmaması ve alan araştırmacılarının erişebileceği kaynakların yetersizliği gibi engellerin tamamı, Arapça sistemi oluşturmayı daha da zorlaştırmaktadır (Alwaneen vd., 2022). Bu zorluklardan birkaçı bu bölümde ele alınacaktır.

2.4.1 Aksan İşaretleri

Yazılı Arapçada, aksan işaretleri ve uzun ünlüler olmak üzere iki tür ses bulunmaktadır. Kelime telaffuzlarını belirtmek için aksan işaretleri adı verilen bir dizi ortografik sembol kullanılır. Arapça, yazılı formda kısa sesli harfleri belirtmek için aksan işaretlerini kullanır (Azmi ve Aljafari, 2018). Bir sesi belirtmek için bir harfin üstüne veya altına yerleştirilen işarete aksan işareti denir. Modern Arapça yazı, tipik olarak, aksan işaretleri olmadan yazılır; bu da kelimenin aynı biçimine birden fazla yorum verilebilmesi nedeniyle önemli derecede belirsizliğe yol açar. Arapça Doğal Dil İşleme araştırmacılarının karşılaştığı zorluklardan biri de budur (Alwaneen vd., 2022; Shaalan vd., 2019).

2.4.2 Büyük Harf Eksikliği

İsimlendirilmiş varlık tanıma, gerçek soruları yanıtlamada birincil adımdır. Bu adımda tanıyıcı, metinden kişi veya yer adlarını çıkarır. İngilizce gibi diğer dillerde bu isimlerin büyük harflerle başlaması bu süreci kolaylaştırır. Ancak Arap dilinde büyük harf kullanımı yoktur, bu durum isimlendirilmiş varlık tanıma görevlerinde yüksek performans elde edilmesine engel teşkil etmekte ve soru analizi ve cevap çıkarma işlemlerinde çok fazla belirsizlik yaratmaktadır. Sonuç olarak sistem, özel isimleri bulmak ve sıfatlar veya fiiller gibi diğer formlardan ayırmak için daha fazla araca ihtiyaç duyabilir (Alwaneen vd., 2022; Alothman ve Alsalman, 2020).

2.4.3 Çekimli Dil

Arapça son derece çekimli bir dildir. Hem ön ek hem de son ek birleştirilebilir ve bu nedenle bir kelimenin sıfır veya daha fazla eki olabilir. Bilgi erişimi Soru-Cevap sistemlerinde önemli bir adımdır. Kullanıcının verdiği soruya en uygun dokümanı bulur. En alakalı dokümanı bulmak esas olarak dokümandaki soru anahtar sözcüklerini bulmaya bağlıdır. Ancak belgede bu kelimeler eklerle birlikte bulunabilir ve bu da dokümanı geçersiz kılar. Bu nedenle Arapça morfolojisinin karmaşıklığı, Arapça Doğal Dil İşleme sisteminin geliştirilmesini karmaşık bir görev haline getirmektedir (Alwaneen vd., 2022; Alothman ve Alsalman, 2020).

2.5 Kök Çıkarma

Bir kelimeyi köküne indirgemek genellikle ön eklerin ve son eklerin silinmesini içerir. Bu prosedür kök çıkarma olarak bilinir (Thalji, 2024). Kök veya gövde çıkarma işlemi, herhangi bir Arapça kelimenin kök veya gövdesinin nasıl tespit edileceği veya tanınacağıdır, bu kök çıkarma işlemi Bilgi Erişimi gibi farklı alanlarda kullanılmaktadır (Al-Kabi vd., 2015; Avvad, 2019; Zakria vd., 2019). Modern Standart Arapça’da kullanılan Arapça kelimelerin tümü orijinal Arapça kelimeler değildir ve Arapça’nın kökenlerinden gelmemektedir. Bunlar, “televizyon”, “internet” gibi yerel Arapça kökenlerden türetilmemiş olan ve dolayısıyla otantik Arapça kökenlerinden yoksun olan kelimeleri içerir. Doğal dillerde, aynı kökten veya gövdeden türemiş birçok kelime bulmak mümkündür. Bu nedenle, aynı kökü paylaşan kelimelerin aynı anlama sahip ve eşanlamlı olarak düşünmek yaygındır. Ancak aynı kaynaktan gelen her kelime aynı anlama sahip değildir (Al-Kabi vd., 2015). Kök çıkarma işleminden önce ön işleme yapılmalıdır. Ön işleme, belgelerden Arapça olmayan tüm ifadeleri, gereksiz sözcükleri, rakamları, sayıları ve noktalama işaretlerini kaldırır ve bunları sonraki metin işleme süreci için en uygun hale getirir (Alhaj vd., 2019). Ayrıca hesaplamayı hızlandırır ve doğruluğu ve verimliliği artırır. Ön işleme, gereksiz sözcükleri elimine etmeyi, bölütlemeyi, düzgünleştirmeyi ve diğer ilgili prosedürleri içeren bir süreçtir. Kök çıkarma, doğal dil işleyen her sistemde standart bir tekniktir. Kök çıkarma, kullanıldığı herhangi bir programın performansını doğrudan etkileyebileceğinden, yetkin bir kök çıkarıcı geliştirmek çok önemlidir.

Arapça çok çekimli ve türetilmiş bir yapıya sahip olduğundan Arapça terimlerde kök çıkarma oldukça zordur. Genel olarak kök çıkarmanın amacı, etkili bilgi erişimi için bir kelimenin temsili dizinleme formunu keşfetmektir (Alshalabi vd., 2022). Yaygın olarak yazılan bir sözcük biçiminden çekimli ve türetilmiş sözcükleri temellerine (kök veya gövde) indirgemek kök çıkarmanın amacıdır. Başka bir deyişle kök çıkarma, bir kelimenin tüm son ek ve öneklerini ortadan kaldırarak kökünü veya gövdesini üretmeye yönelik hesaplamalı işlemdir. Ayrıca, bir ek alabilen bir morfem veya bir dizi birleştirilmiş morfem, bir gövde olarak tanımlanabilir. Ancak Arapçada isimler genellikle ön ekler kaldırıldığında anlamını kaybeder (Alshalabi vd., 2022). Gerekli analiz düzeyine bağlı olarak Arapça kök çıkarma algoritmaları iki kategoriden birine girer: az veya kural tabanlı. Başka bir deyişle, kural tabanlı algoritmalar gövdeleri köklere indirgerken, az kök çıkarma algoritmaları Arapça kelimelerdeki ön ek ve son ekleri ortadan kaldırır. Mevcut çalışmanın ana amacı, etkili bir türev oluşturarak bir Arapça az kök çıkarıcı geliştirmektir. Bu, daha fazla ön ek ve sonek ekleyerek ve kelime uzunluğuna dayalı belirli kısıtlamaları uygulayarak yapılacaktır (Alshalabi vd., 2022). Metin madenciliği, bilgi erişimi ve doğal dil işleme, ilgili kelimeleri temsil etmek için bir temel dize üreten ve otomatik bir yöntem olan kök çıkarıcıları kullanan sayısız uygulamadan sadece birkaçıdır. Kelimeler kök çıkarma yöntemiyle en temel hallerine indirgenir. Metinlerin söz dizimi ve anlambilimini değerlendirip kavramak için kullanılan sistem sayısı arttıkça sonuçların doğruluğu da artar. Bazı Arapça kelimelerin anlamsal olarak bağlantılı olması gerekmeyen morfolojik formları olduğundan, Arapça kelimelerin kök ayrımı basit bir işlem değildir (Al-Abweeny ve Zaid, 2018). Gövde, birleştirilmiş morfolojinin temeli olan yüzeysel bir kelimedir ve bir kalıba üst üste koyulan kökler ve kısa sesli harflerle oluşturulur (Alothman ve Alsalman, 2020).

2.6 Arapça Kökler

Bir dildeki bir kelimenin, kendisinin olduğu bir temeli vardır; bu temel, dilsel kök olarak adlandırılır ve çok sayıda kelime tarafından paylaşılan harfler olarak tanımlanır. Genellikle üç ünsüzden oluşur, ancak bazen dört, beş veya altı da olabilir. Farklı sistemler kullanılarak elde edilebilir. Bir kelimenin genel anlamını gösterir (Alothman ve Alsalman, 2020). Arapça kelimeler bir dizi kökten türetildiği için hem isimlerin

hem de fiillerin birden fazla kökü vardır. Arapça kökler sıklıkla üç harfli, dört harfli veya beş ünsüz içeren beş harflidir (Mustafa vd., 2019). Arapça kelimeler oluşturmak için bu köklere ekler (iç ek, ön ek ve son ek gibi) eklenir, bu da Arapçada her kök için çok sayıda olası kelime olduğu anlamına gelir. Bağlaçlar ve edatlar ön ek örnekleridir; çift dişil ve çoğul eril ekler ise son ek örnekleridir. Arapça kelimeler eril ve dişil olarak sınıflandırılır. Buna ayrıca zamirleri (yani, ikinci ve üçüncü şahıs) belirten ekler de dahil edilebilir, bu da onun Arapça bir son ek olduğunu gösterir (Mustafa vd., 2019). Kelimeleri ve kökleri listelemek ve düzenlemek için ilk Arap dilbilimcileri tarafından bir sözlük koleksiyonu üretilmiştir. Bu sözlüklerin düzenlenme şekli onları çeşitli gruplara ayırmaktadır, bu da farklı arama stratejilerinin uygulandığı anlamına gelmektedir. Sözlük, belirli bir sıraya göre düzenlenmiş dilsel terminolojinin yanı sıra sözcük dağılımının açıklamasını ve başka bir dile çevirisini içeren bir kitaptır. Arapça sözlük, Arapça dilini korumak için değerli bir araçtır ve Kur'an-ı Kerim'in anlamlarını açıklamada önemli bir rol oynar. Arap dil biliminde bir kelimenin temel anlam birimine kök denir. Sesli harf ve ek modifikasyonları, genellikle ünsüz harfli kökler ve kalıpların bir kombinasyonundan oluşturulan Arapça kelimelerdeki farklı gramer rollerini ve nüansları ifade etmek için kullanılır. Örneğin, yazmanın genel kavramı (كتب) köküyle belirtilir. Bu köke çeşitli sesli harfler ve ekler eklenerek, “kitap” anlamına gelen (كتاب), “yazar” anlamına gelen (كاتب) ve “kütüphane” anlamına gelen (مكتبة) gibi çok sayıda kelime oluşturulabilir. Arapça morfoloji ve sözlük biliminde köklerin anlaşılması önemlidir çünkü bu, konuşmacıların kalıpları belirlemesine ve yabancı kelimelerin anlamlarını onları oluşturan parçalardan çıkarmasına olanak sağlar.

2.7 Arapça Kalıplar

Morfolojik kalıp, Arapça kelimelerin iç yapısını incelemeye yönelik bir modeldir. Kökü, kısa ünlüleri ve belirli ekleri düzenleyerek soyeksenli bir yapıda gövde oluşturmayı gösterir. Zaman gibi dilbilgisel anlamları ve kelime türlerini ifade eder (Alothman ve Alsalman, 2020). Üç Arapça harf kalıbıyla temsil edilen üç harfli kökün sırasıyla birinci, ikinci ve üçüncü harflerinden oluşur (Al-Kharashi ve Al-Sughaiyer 2004). Kök harf boşluklarını çevreleyen karakterlerle oluşturulan bir şablondur. Bu boşlukları görüntülemek için kalıpta indeksler kullanılır. Boşlukların arasında ve

çevresinde kısa ünlülerin yanı sıra bazen ünsüz ya da uzun ünlü harflerden oluşan kalıplar görülmektedir (Hassan, 2015). Dörtlü kökler durumunda kökün dördüncü harfi, üçüncü harfin kopyası ile temsil edilir. Ayrıca bir veya daha fazla kısa sesli harf ya da sıfır veya daha fazla çoğaltılmış harf de yerleştirilir (Al-Kharashi ve Al-Sughaiyer, 2004). Arap dilbilimciler, sunumunun biçimi, gecikmesi, hareketi ve durgunluğuyla ağırlıklı olanın kabulüne kendilerini adadılar. Konuşma yapılarını ve bu yapıların koşullarını anlayacak bir ölçüğe sahip olmak için “kalıp” adını verdikleri sözel standart ve morfolojik dengeyi kullandılar. Kelimenin kökünü belirlemek ve kelimeyi ağırlık açısından yönetmek için kelimenin morfolojik ağırlığını belirlemek amacıyla bir yöntem oluşturdular. Orijinal fiil zamanlarının ve Arapçadaki soyut isimlerin ve fiillerin çoğunun üçlü dil köküne sahip olduğu gerçeğini dikkate aldığına, dilbilimciler oluşumun başlangıcını üç harf etrafında inşa ettiler; “fiil” anlamına gelen ve birlikte fiili oluşturan "ع", "ف", ve "ل".

2.8 Kısa Ünlü veya Aksan

Genellikle bir harfin üstüne veya altına yerleştirilen sembollere aksan işaretleri denir. Shadda, sadece isimlerin sonlarında görülen üç nunlama, ve üç kısa ünlü (“Fetha”, “Damma” ve “Kesra”) aksan işaretlerin örnekleridir (Alothman ve Alsalman, 2020).

2.9 Arapçada Çoğulluk

Arapça sayılar üç çeşittir: tek, ikili ve çoğul. Arapça çoğul üç ile başlar. Arapça’da iki tür çoğul vardır; düzenli ve düzensiz (Goweder vd., 2004). Düzensiz çoğul, çözülmesi çok fazla çaba ve zaman gerektirdiği için zorlayıcı bir durumdur. Düzensiz çoğul oluşturma ve tanımlama konusunda kesin kurallar yoktur (örneğin, tek bir forma bir veya daha fazla harf eklemek veya bir veya daha fazla harfi silmek). Bu işlemleri yapmanın birkaç yolu vardır. Her stratejinin avantajları ve dezavantajları bulunur (Ellouze vd., 2010). Bu yaklaşımlardan bazıları eklerin kaldırılması, dil bilgisi kurallarının kullanılması ve sözlüklerin kullanılmasıdır.

2.9.1 Arapça Düzenli Çoğul

Arapça düzenli çoğulun oluşabilmesi için kökteki fazladan harflerden arındırılması gerekir. Arapça bir kelimenin kalıbının genellikle (فعل) harfleriyle temsil edilen üç harften oluştuğunu hatırlamak da önemlidir. Arapça isim ve sıfatların iki cinsiyeti vardır: eril ve dişil. Bu nedenle çoğullar çeşitli çekim ekleri kullanılarak oluşturulur ve değişiklikler genellikle görülmez (Goweder vd., 2004). Çoğulun bu bölümünde iki tür vardır:

2.9.1.1 Arapça eril düzenli çoğul

Standart eril çoğullarda kullanılabilen isimler, kökleri sesli harfle bitmeyen eril özel isimler ve eril sıfatlardır. Eril düzenli biçime "ون" veya "ين" son ekleri eklenir (Al-Sughaiyer ve Al-Kharashi, 2004; Goweder vd., 2004).

2.9.1.2 Arapça dişil düzenli çoğul

Bu çoğul Arapça'da sıfatların yanı sıra hem insan hem de insan dışı geniş bir isim sınıfını tanımlamak için yaygın olarak kullanılır. Tipik dişil çoğullar "ات" son ekiyle şekillendirilir; dikkat edilmesi gereken nokta, kökün o harfle bittiği durumlarda ek olan "ة" harfini değiştirmesidir (Goweder vd., 2004; Shafah vd., 2016).

2.9.2 Arapça Düzensiz Çoğul

Arapçada belirli bir kalıbı olmayan düzensiz çoğulların oluşumunu denetleyen çeşitli katı kurallar vardır (Goweder vd., 2004; Shaalan, 2014). Arapça kavramlar kök adı verilen soyut yapı taşlarından oluşur. Kök, fonetikteki en temel terimdir ve son ekler veya ön eklerle birleştirilmesiyle fiiller, sıfatlar ve isimler de dahil olmak üzere diğer türevleri oluşturur. Arapça kelimelerin neredeyse %85'inin üç harfli köklere sahip olması dikkat çekicidir, bu da dilin önemli çekim içeriğini göstermektedir (Etaiwi ve Awajan, 2018). Arapçada fiiller ve isimler genellikle 10.000 üye içeren bir kök grubundan oluşturulur (Issa, 2018). Düzensiz çoğullar büyük Arapça derlemlerdeki belgelerin %10'unu oluştururken, düzenli çoğullar %41'ini oluşturur (Shafah vd., 2016). Bu nedenle, düzensiz çoğul tanımlama, bilgi erişim uygulamalarına yönelik az

kök çıkarma algoritmaları için ciddi bir zorluk teşkil etmektedir. Ancak düzensiz çoğul tespitin etkileri çok az anlaşılmıştır. İnternet veya sosyal medyada bulunan büyük miktardaki veriyi kullanarak, kullanıcının isteklerine özgü sorgu yanıtları sunmak, arama motorlarının temel hedeflerinden biridir. Bu dönüş, kullanıcının ihtiyaçlarını karşılamalıdır. Arapça diğer dillerle karşılaştırıldığında kendine özgü fonetik ve anlamsal kalıplara sahiptir. Bu fark, Arap kullanıcıların arama motoru optimizasyonundan nasıl yararlandıkları konusunda önemli bir konuyu da gündeme getirmiştir. Son zamanlarda akademisyenler tüketicilerin arama sonuçlarını geliştirmek için Arapça kullanmaya ilgi duymaya başlamıştır. Ana sorunlardan biri Arapça web sitelerinin eksikliğidir. Bir diğer zorluk ise Arapça bilgi erişim sistemlerinin kelime belirsizliği (köklerin çoğunluğu üç harften oluşur), imla farklılıkları ve karmaşık ve son derece zengin morfoloji gibi sorunları çözmemesi olmuştur (Shafah vd., 2016).

2.10 İsimlendirilmiş Varlık Tanıma

Bir isimlendirilmiş varlık, benzer özelliklere sahip bir kümenin içindeki bir öğeyi belirgin bir şekilde belirten bir kelime veya ifadedir. Biyomedikal alanında genlerin, proteinlerin, ilaçların ve hastalıkların adları, küresel alanda ise organizasyonların, insanların ve yerlerin adları buna örnektir (Shrestha ve Bal, 2020; Sghaier ve Zrigui, 2020). Metinde bahsedilen varlıkları önceden tanımlanmış varlık kategorilerine yerleştirme ve sınıflandırma tekniği, isimlendirilmiş varlık tanıma olarak bilinir (Ali vd., 2020).

2.10.1 İsimlendirilmiş Varlık Tanıma Uygulamaları

İsimlendirilmiş Varlık Tanıma, Doğal Dil İşlemenin çeşitli disiplinlerinde kullanılmaktadır. Aşağıda bu alanlara ait bazı örnekler verilmiştir:

Bilgi Erişimi: Bunun için gelen bir isteğe yanıt olarak bir veri kaynağından ilgili bilgilerin bulunması ve elde edilmesi gerekir. İsimlendirilmiş Varlık Tanıma, talepteki isimlendirilmiş varlığı tanımlayarak ve aranan bilgilerde isimlendirilmiş varlığı konumlandırarak bilgi çıkarmaya yardımcı olabilir. Son olarak, sınıflandırılmış

isimlendirilmiş varlıkları ve bunların taleple ilişkilerini hesaba katarak ilgili bilgileri elde edebilir (Ali vd., 2020).

Soru Cevaplama: Soru-cevap teknikleri, soruları girdi olarak alır ve kısa ve öz, nokta odaklı yanıtlar sağlar (Alwaneen vd., 2022). Ayrıca, İsimlendirilmiş Varlık Tanıma görevi, sorgunun işleme aşaması boyunca isimlendirilmiş varlığı belirlemede yardımcı olmak için kullanılabilir. Sonuç olarak bu, ilgili bilgileri bulmamızı ve uygun bir yanıt sunmamızı sağlayacaktır. İsimlendirilmiş Varlık Tanıma, birçok gerçek sorgunun ihtiyaç duyduğu gibi, soru yanıtlama sistemleri için yararlı bir özelliktir (Alwaneen vd., 2022; Ali vd., 2020).

Makine Çevirisi: Bu, bir metnin diller arasında otomatik olarak çevrilmesi işlemidir (Zirikly ve Diab, 2015). İsimlendirilmiş varlığın hangi bölümlerinin karakter çevirisi yapılması ve hangilerinin kelime çevirisi yapılması gerektiğinin belirlenmesine özellikle dikkat edilmelidir (Ali vd., 2020; Elmenshawy ve El-Nady, 2019). İsimlendirilmiş Varlık Tanıma sisteminin kalitesi, makine çevirisi sisteminin genel kalite düzeyinin değerlendirilmesinde çok önemlidir. Bu nedenle İsimlendirilmiş Varlık çevirisi, çok dilli uygulama sistemlerinin çoğunluğuna dahil edilmiştir (Elmenshawy ve El-Nady, 2019).

Metin Kümeleme: Oluşturulan kümeleri her kümedeki varlıkların oranına göre derecelendirerek, İsimlendirilmiş Varlık Tanıma, çalışma sonuçlarını kümeler halinde düzenlemek için kullanılabilir (Zirikly ve Diab, 2015). Bu, her kümenin özelliklerini analiz etme sürecini iyileştirir ve önceden belirlenmiş kriterler açısından kümeleme tekniğini güçlendirir. Bir belgedeki olayların tam zamanını ve yerini belirlemek için zaman ifadeleri ve İsimlendirilmiş Varlıklar gibi değişkenler kullanılabilir (Zirikly ve Diab, 2015).

Navigasyon Sistemleri: Dijital haritalar ile navigasyonu kolaylaştıran bu teknolojiler, günlük hayatımızın vazgeçilmez bir parçası haline gelmiştir. Bunlar, yol tariflerini, (diğer internet hizmetleriyle bağlantılı olabilen) yerel varış noktalarına ilişkin bilgileri ve dolaşım bilgilerini içerir. Bu tür sistemlerde ilgili noktalar, coğrafi koordinatlara sahip isimlendirilmiş varlıklar olarak bir veritabanında saklanır. Otoparklar,

mağazalar, hastaneler, restoranlar, üniversiteler, okullar ve simge yapılar gibi turistler, ziyaretçiler ve arayıcılar için önemli olan alanları kapsarlar (Ali vd., 2020).

2.10.2 Arapçada İsimleştirilmiş Varlık Tanıma Zorlukları

Arapça yaygın olarak konuşulan bir dildir ve BM'nin resmi dillerinden biridir. Araştırmacılar, Arapça bir uygulama geliştirirken dilin kendine has özelliklerinden dolayı önemli zorluklarla karşılaşmaktadırlar. Bu zorluklar arasında zengin bir morfoloji, karmaşık yazım ve farklı derecelerde belirsizlikler yer alır (Abdallah vd., 2012). İsimlendirilmiş Varlık Tanıma, Arapça doğal dil işleme teknolojileri ve uygulamalarının önemli bir bileşenidir. Arapça İsimlendirilmiş Varlık Tanıma araştırmalarında son yirmi yılda kaydedilen muazzam ilerlemeye rağmen, Arapça dilinin, isimlerin diğer kelimelerden ayrılmasına yardımcı olan büyük harfin kullanıldığı İngilizce'den farklı olarak büyük harf eksikliği gibi çeşitli özellikleri nedeniyle hala zorluklar devam etmektedir (Hassan vd., 2018). Bileşikleştirme, Arapça'da kelimelerin bir veya daha fazla ön ek, kök ve son eklerden oluşan birden çok grupta içinde oluşturulduğu karmaşık bir morfolojik özelliktir. Diğer bir konu ise yazı stilindeki tutarlılık eksikliğidir (Ali vd., 2019). Arapça diğer dillere göre daha fazla konuşma sesine sahip olduğundan, Arapçaya çevrilen adların aynı anlamı korurken birden fazla NE varyantı olacaktır (Ali vd., 2019). İsteğe bağlı kısa ünlüler başka bir konudur. Arapçadaki ünlü harflerin çoğu, fonetiği değiştiren ve aynı kelimeye farklı anlamlar yükleyen aksan işaretleri ile temsil edilir. Bugün kullanılan Arapçanın modern versiyonunda aksan işaretleri yoktur ve bu durum sözcüksel karışıklığa yol açmaktadır (Hassan vd., 2018). Kaynak eksikliği de bir engeldir, çünkü internette büyük miktarda veri vardır ancak Arapça İsimlendirilmiş Varlık Tanıma için kapasite veya kapsam bakımından sınırlı olan veya pahalı bir lisans gerektiren yalnızca birkaç kaynak mevcuttur (Shaalan, 2014; Ali vd., 2019). Ayrıca, bazı içerikler genellikle Arap dünyasının farklı bölgelerinden lehçelerin temsil edildiği konuşma dilinde yazılmıştır, bu da Arapça İsimlendirilmiş Varlık Tanıma zorluğunu artırır. Resmi ve günlük olmak üzere dilin iki farklı türünün bulunması daha teknik bir sorundur. Modern Standart Arapça (MSA), çoğu yazılı kaynakta ve resmi konuşmalarda kullanılır, ancak günlük iletişimde kullanılmaz; günlük konuşma dilindeki Arapça ise günlük iletişim için kullanılır ve coğrafi olarak farklılık

gösterebilir. Pek çok bilim adamı Arapçanın benzersiz özelliklerine ilişkin kapsamlı araştırmalar yürütmüştür (Zayed ve El-Beltagy, 2015).

2.11 Arapça Fiiller

Arapça dili zengindir ve diğer diller arasında en karmaşık dil olarak kabul edilir. Bir ifadenin karmaşıklığı, sözdizimi ve anlambilimi içinde doğal olarak bulunur. Ayrıca hem isim hem de fiil olarak, kelime eşanlamlıları, zıt anlamlılar ve kelime kökleri gibi geniş bir kelime hazinesi de içerir (Othman vd., 2020). Biçim ve anlam bakımından karmaşıklığı nedeniyle Arapça fiil morfolojisi, Arapça bir cümlenin oluşumunda kritik öneme sahiptir. Bununla birlikte, yaygın olarak kabul edilen bir modele göre, Arapça fiiller, üç veya dört harfli kökler kullanılarak çekilebilir. Arapça fiiller dört kip (gösterme, şartlı, emir ve kuvvetli), üç zaman (mükemmel, eksik ve emir), ve iki ses (etken ve edilgen) olarak bölünür. Fiiller, morfolojik kurallara uyularak bazı zaman, ses ve kiplerde çekilebilir. İstisnalara bazı kök ünsüzlerin fonolojik kısıtlamaları neden olur. Arapça esas olarak türetilmiş bir dildir; yani kelimelerinin çoğu, belirli bir kök kümesinden belirli bir morfolojik ‘biçim’ veya ‘kalıp’ kümesi aracılığıyla türetilmiştir. Arapça bir kelime fiil, isim veya edat olabilir. Arapçada edat sayısı yaklaşık 80’dir. Arapça fiiller emir, şimdiki zaman ve geçmiş olmak üzere üç kategoriye ayrılır. Fiiller ayrıca yalın, uzatılmış, düzenli veya eksikli olarak da sınıflandırılabilir. Önceki araştırmacılar ve profesyoneller tarafından geliştirilen az sayıda morfolojik form, her kategorinin morfolojik özelliklerini etkili bir şekilde özetlemektedir. Bu formlar, Arap nesirinde yaygın olan zamirlerle birleştirildiğinde fiil eylemini gösterir. Dolayısıyla fiiller hakkındaki bilgiyi kodlamanın en iyi yollarından biri, ilgili tüm morfolojik formları ve bunların niteliklerini depolamaktır (Kanan vd., 2019). Arapça bir fiilin ünsüz kökü, genellikle fiilin tüm çekimli biçimlerinde bulunan üç veya dört ünsüzden oluşur; çekimli biçimin kök dışındaki kısmına kalıp denir. Kök bu ünsüzlerden oluşur ve kalıplar kökün anlamını değiştirerek çok sayıda ilgili kelime oluşturur (Azman, 2019). Dolayısıyla kök fiiller, olay fiillerini ve onların değişken yüzey formlarını dikkate alan neredeyse tüm morfoloji sistemlerine ve dil bilgisel analiz sistemlerine dahildir. Buna ek olarak, bazı sistemler bir fiil çekimini çıkarmayı amaçlarken, ayrıştırıcılar gibi diğerleri, anlamı kaybetmeden hatalı bir metni yeniden ifade etmek için sadece fiilin yüzey formuna ihtiyaç duyar. Anlamsal analiz aynı zamanda cümle

argümanlarının tematik işlevlerine göre derlenmesini de gerektirir ki bu sistem içerisinde daha önemlidir (Kambhatla ve Zitouni, 2013). Web Araması, Bilgi Erişimi, Makine Çevirisi, Soru Cevaplama ve benzeri popüler çalışma alanlarındaki neredeyse tüm problemler, doğru fiil tanımlama ve çıkarma gerektirir (Ahmed vd., 2023).

2.12 Arapça Doğal Dil İşleme

Yazılı metinler, insan müdahalesine gerek kalmadan doğal dil işleme yazılımları ile analiz edilebilmektedir (Kanan vd., 2019). İnsan dilinin makine tarafından anlaşılması, doğal dil işlemenin nihai hedefidir. Ön işleme, belgelerden Arapça olmayan tüm ifadeleri, gereksiz kelimeleri, rakamları, sayıları ve noktalama işaretlerini kaldırır ve bunları sonraki metin işleme süreci için en uygun hale getirir. Ön işleme, hesaplamayı hızlandırmak ve doğruluğu artırmak için de kullanılabilir. Bölütleme, düzgünleştirme ve gereksiz kelimelerin kaldırılması adımları olarak tanımlanır (Guellil vd., 2021). Arapça Doğal Dil İşleme, Arapça yazılmış metinlerin hesaplamalı araçlar kullanılarak analiz edilmesi, anlaşılması ve üretilmesi sürecidir. Doğal dil işleme teknikleri ve ilkeleri dile özgü olmasa da Arapça doğal dil işleme hâlâ ileri teknolojiden uzaktır. Arap alfabesiyle yazılan Arapça yazılar, Arapça doğal dil işlemenin odak noktasıdır. Arapçanın az kaynaklı bir dil olması, doğal dil işlemenin bu dalını özellikle benzersiz kılmaktadır. Son yıllarda bu teknolojiye hem ticari olarak hem de araştırma ilgisinde bir patlama yaşanmış ve bu da devasa transformatör tabanlı ağlar kullanan belirli derin öğrenme modellerinin oluşturulmasını sağlamıştır. Doğal dil işleme, bu modeller sayesinde büyük ölçüde ilerleme kaydetmiş, ancak aynı zamanda, özellikle denetimsiz dil öğreniminde, derin öğrenmenin ve Doğal Dil İşlemenin bulunduğu alanlar da öne çıkmıştır. Arapça diğer dillere göre daha fazla sayıda konuşma sesine sahiptir ve isimlerin aynı anlama gelen çeşitli biçimleri vardır. İsteğe bağlı kısa sesli harflerin kullanılması da başka bir sorundur (Ahmed vd., 2022). Sesli harflerin çoğu, fonetik temsili değiştiren ve aynı kelimeye farklı bir anlam kazandıran aksan işaretleriyle temsil edilir. Günlük Arapça dil kullanımında aksan işaretleri eksiktir, bu da sözcüksel zorluklara neden olur (Hassan vd., 2018). Bu bölümün geri kalanında Arapça dilinin işlenmesinde yer alan temel süreçler genel olarak tartışılacaktır.

2.12.1 Düzgünleştirme

Arapça düzgünleştirme, genellikle metin analizini ve işlenmesini kolaylaştırmak için Arapça metni standartlaştırma veya normalleştirme uygulamasıdır. Arapça metin, düzgünleştirme yoluyla ele alınması amaçlanan imla, morfolojik ve diyalektik değişkenler gibi bazı dilsel farklılıklar içerir. Arapça işlemede birkaç tipik düzgünleştirme yöntemi aşağıdaki gibidir:

Ortografik Düzgünleştirme: Arapça kelime yazımı, birçok biçime sahip harfler ve bitişik harfler kanonik biçimlerine dönüştürülerek standartlaştırılmaktadır. Bu, aksan işaretleri veya aynı harfin farklı biçimleri gibi farklı karakterlerin yerine temel formların kullanılmasını gerektirir.

Aksan Kaldırma: Arapça yazı, belirli gramer kurallarını veya sesli harfleri belirtmek için genellikle kısa sesli harf işaretleri olarak bilinen aksan karakterlerini kullanabilir. Aksansız bir metin oluşturmak için Arapça kelimelerin aksan işaretleri kaldırılmıştır. Bu, bölütleme ve kök çıkarma gibi metin işleme işlemlerini kolaylaştırabilir.

Kök Çıkarma ve Çözümleme: Çekimli ve türetilmiş kelimeleri temel veya kök biçimlerine döndürme teknikleri, kök çıkarma ve çözümleme içerir. Arapça’da çözümleme, kelimeleri sözlük biçimlerine veya başsözcüğe eşleme işlemidir; oysa kök çıkarma, kök harflerini bulmak için kelimelerden önek ve son eklerin kaldırılmasıdır. Kök çıkarma, kelimedeki ek ve önekleri kaldırarak kelimeyi köküne döndürmeyi amaçlamaktadır. Eklerin ön ekler, son ekler ve iç ekler olarak türleri vardır.

Sayısal Düzgünleştirme: Arap literatüründe sayıların temsil edilme şeklinin standartlaştırılması, Arap rakamlarının (٣, ٢, ١ , vb.) karşılık gelen Batı rakamlarına (1, 2, 3 vb.) veya tam tersine çevrilmesiyle sağlanabilir. Metin işleme proseslerinde bu düzgünleştirme aşaması, sayısal verilerin tutarlı şekilde işlenmesini garanti etmeye yardımcı olur.

2.12.2 Bölütleme

Arapçada bölütleme, Arapça bir metni sözcükleri, ifadeleri, noktalama işaretlerini ve diğer dilsel bileşenleri kapsayan kurucu belirteçlere veya birimlere ayırma eylemini ifade eder. Metnin bilgisayar tarafından anlaşılmasını ve analizini geliştirmek için bölütleme, Doğal Dil İşleme etkinliklerinde önemli bir ön işleme adımıdır. Ancak Arapçanın karmaşık bir imla sistemi ve zengin bir morfolojiye sahip olması nedeniyle Arapça metni bölütlemenin kendine has zorlukları vardır. Metni, her kelimeyi tek başına bırakacak ve ilk boşluktan sonraki kelimeyi ayıracak şekilde bölme işlemidir; her bölüme bölüt adı verilir. Arapça metnin etkili bir şekilde bölütleştirilmesi, makine çevirisi, isimlendirilmiş varlık tanıma, sözdizimsel ayrıştırma ve konuşmanın bir kısmını etiketleme gibi daha sonraki görevler için gereklidir. Doğal Dil İşleme algoritmaları, Arapça metindeki anlamlı belirteçleri doğru bir şekilde bölümlere ayırarak, çok sayıda Arapça veri kümesinden önemli dilsel bilgiler ve içgörüler çıkarılabilir. Arapça bölütlemeye yönelik yaklaşımlar Kural Tabanlı Bölütleme, İstatistiksel Bölütleme ve Makine Öğrenimi Tabanlı Bölütleme'dir

2.12.3 İsimlendirilmiş Varlık Tanıma Çıkarımı

Arapça İsimlendirilmiş Varlık Tanıma görevi, Arapça metindeki isimlendirilmiş varlıkların kişiler, yerler, tarihler ve diğer varlık türleri dahil olmak üzere belirli kategoriler halinde tanımlanmasını ve sınıflandırılmasını içerir. Belirli kişilerle, gruplarla, konumlarla, tarihlerle ve diğer varlıklarla ilgili ifadeler, isimlendirilmiş varlıklar olarak bilinir. Bilgi erişimi, soru cevaplama ve belge özetleme, Doğal Dil İşleme'ye ve onun temel işi olan İsimlendirilmiş Varlık Tanıma'ya bağlı uygulamalardan sadece birkaçıdır (Hassan vd., 2018). Arapça İsimlendirilmiş Varlık Tanıma, dilin farklı kelime düzeni, zengin morfolojisi ve lehçe farklılıkları nedeniyle birçok zorlukla karşı karşıyadır. Morfolojik karmaşıklık, imla farklılıkları, lehçe farklılıkları ve sınırlı dilsel kaynaklar bu zorluklardan birkaçıdır. İsimlendirilmiş Varlık Tanıma, Arapça metinden önemli bilgilerin çıkarılmasını kolaylaştırarak bilgi erişim sistemleri, makine çevirisi ve soru cevaplama gibi faaliyetleri destekler (Ahmed vd., 2022). İsimlendirilmiş Varlık Tanıma Duyarlılık analizi, tanımlanmış varlıklara yönelik görüş veya duyguların ifade edilmesini içerebileceğinden, bunların

tanımlanması önemlidir. İsimlendirilmiş Varlık Tanıma, Arapçadaki verilerin düzenlenmesini ve sınıflandırılmasını, analizini ve kullanımını kolaylaştırır. Kelime sınırlayıcıların eksikliği, değişken harf şekilleri ve aksan işaretlerinin (Harakat) varlığı, Arapça İsimlendirilmiş Varlık Tanıma ile ilgili zorluklardır (Hassan vd., 2018).

2.12.4 Kelime Etiketleme Tanımlaması

Metindeki her kelimenin isim, fiil, zarf, sıfat gibi görünüm ve anlatımdaki konumuna göre tanımlanması işlemidir. Her kelime, Doğal Dil İşleme'deki "etiketleme" olarak bilinen dilbilimsel işlem tarafından belirli bir kelime türüne (fiil, sıfat, zarf vb.) veya dilbilgisel kategoriye atanır (Mohamed ve Kübler, 2010). BET, bir cümledeki her kelimeye, cümledeki amacını belirten benzersiz bir etiketin verildiği yöntemdir. Arapça Aksan işaretlerinin eksikliği ve Arapça morfolojisinin karmaşıklığından kaynaklanan yüksek derecede belirsizlik nedeniyle ABET kolay bir işlem değildir. Her kelime için en uygun morfolojik etiket ABET tarafından atanır (El Hadj vd., 2009). Kurala dayalı yaklaşım, İstatistiksel yaklaşım ve Hibrit yaklaşım, etiketlemede kullanılan üç genel stratejidir (Alharbi vd., 2018).

2.12.5 Arapça Gereksiz Kelimelerin Kaldırılması

Gereksiz sözcükler, metni işlemeyen önce filtrelememiz gereken sözcüklerdir. Gereksiz kelimeler sonuçları yanıltıcı olabileceğinden kaldırılmalıdır, bu nedenle araştırma sürecini iyileştirmek için bunları görmezden gelmemiz gerekir (Kanan vd., 2019).

2.13 İlgili Çalışmalar

Arap dili ve Arapça metin üzerine yoğunlaşan pek çok araştırma bulunmaktadır. Bu çalışmada Arap dili ile ilgili belirli konular incelenmiştir ve bu nedenle bu belirli konularla ilgili daha önce yapılmış bazı çalışmalar sunulacaktır. Bu konular; köklerin çıkarılması ve tespiti, isimlerin tanınması, fiillerin tanınması, çoğulların belirlenmesi, kelime ve kök üretilmesidir.

2.13.1 Kök Tespiti ile İlgili Çalışmalar

Arapça kelimelerin köklerinin ve gövdelerinin çıkarılmasının ve bunların çeşitli alanlarda kullanılmasının önemi nedeniyle Arapça dilinin morfolojik analizi, farklı dilbilimsel derinlik ve stil çeşitliliğiyle geniş kapsamlı bir şekilde incelenmiştir. Çok sayıda çalışma, bu prosedürleri, az veya çok kök çıkarma ve eklerin ayrılması dahil olmak üzere çeşitli yöntemlerle gerçekleştirmiştir. Bu eserler kelime sözlüklerinden, kök kelime sözlüklerinden, bu dilin morfolojik analizinin genel kurallarından veya Arapçadaki kelime kalıplarından türetilmiştir.

Zerrouki (2024), Arapça için kapsamlı bir kök bilgisi sunan Tashaphyne adında bir Python kütüphanesi oluşturmuştur. Tashaphyne, algoritmayı değiştirmeden kök ayırıcıları özelleştirmek için kullanılabilen son derece esnek bir araçtır, çünkü tüm olası segmentasyonları oluşturan değiştirilmiş sonlu durum otomatlarını kullanır. Ayrıca Tashaphyne varsayılan olarak ön ekler ve son ekler sunar ve daha karmaşık kök çıkarma görevlerinin yerine getirilmesi için özelleştirilmiş listelerin kullanımını destekler. Tashaphyne, açık kaynak olarak Arapça dil işleme topluluğuna verilen önemli bir katkıdır.

Al-Khateeb vd. (2021), özellikle tıbbi makalelerle ilgili bilgi erişimi sorununu ele almak için geliştirilmiş bir Arapça kök belirleme algoritması sunmuşlardır. Giriş verilerinden kökleri ve gövdeleri çıkarmak için kullanılan algoritma, az kök çıkarma algoritması olarak adlandırılmıştır. Önerilen yükseltilmiş kök ayırıcıyı doğrulamak için dört veri seti kullanılmıştır. Diğer beş iyi bilinen Arapça az kök ayırıcı ile aynı veri setlerini kullanarak, bu tezde önerilen az kök ayırıcının ürettiği gövdeler ve kökler, sonuçlarla karşılaştırılmıştır.

Alnaied vd. (2020), ilgili kelimelerin kökünü ve gövdesini bulmak için Arap harfleri arasındaki ilişkiye dair bir dizi kural uygulayarak kökler oluşturmak veya çıkarmak için Arapça Morfoloji Bilgi Erişimi (AMIR) adı verilen Arapça kök yaklaşımını ele almıştır. Arapça erişim sistemlerinde metin arama için indeksleme terimleri olarak kullanılır. Arapça içeriğin çıkarılmasına yönelik farklı sistemler için önerilen kuralların avantajlarını değerlendirmişlerdir. Son olarak, performansını diğer

yöntemlerle karşılaştırarak ve ortalama doğruluk açısından kök benzeri bir sistem bulunmayan AMIR sistemini değerlendirmişlerdir. Elde edilen sonuçlar, AMIR'in ortalama yüzde 0,34'lük bir doğruluk elde ettiğini, diğerlerinin ise daha düşük yüzdeli doğruluk sunduğunu göstermektedir; bu, AMIR'in Arapça kök ayırıcısı geliştirebileceği ve iyileşmeyi artırabileceği ve aynı zamanda herhangi bir kök biçimine karşı kullanışlı olabileceği anlamına gelmektedir.

Alsaad ve Abbod (2014), iki ana aşamadan oluşan bir dilsel kök çıkarma stratejisi sunmuştur. İlk adımda ön ekler, son ekler ve ara ekler dahil olmak üzere eklerin kaldırılmasıyla ilgilenmişlerdir. Her kesintiden sonra ekleri ortadan kaldırmak için morfolojik model test edilirken, ifadenin uzunluğuna bağlı olarak ön ekler ve son ekler atlanır. Metinlerde rasyonel olarak çok sayıda düzensiz Arapça kelime bulunması nedeniyle, ikinci aşamada ince, ayrık, uzun ünlü ve iki harfli çiftleyen kelimeleri işlemek için kök çıkarma algoritması daha da geliştirilmiştir. Kökler çıkarılmadan önce, önceden tanımlanmış 3 800 üç harfli ve 900 dört harfli kök listesine göre test edilirler. Önerilen algoritmanın verimliliğini artırmak ve doğrulamak için birçok deney yapılmıştır.

Ali vd. (2019), sesli hadis külliyatındaki kelimelerle etiketleri ilişkilendirmeyi amaçlayan Arapça açıklamalar için bir web uygulaması sunmaktadır. Birkaç ek açıklayıcı arasındaki ortak çalışma alanı uygulama tarafından sağlanır. Oturumundaki her yorumcu bir veya daha fazla hadisi ele alır. Her hadisi değerlendirmek için morfolojik analiz cihazı AlKhalil kullanılmıştır. Açıklayıcı, analizi düzeltir ve doğru açıklama için değerleri atar. Bir cümlelerin bölümleri, POS etiketi, kök ve kalıp, ek açıklamanın parçalarıdır. Nihai derlem ek açıklaması sonuçları XML dosyalarında yer alır. Ek açıklama sonuçlarını bir Ağaç bankasında görüntülerler. Araç, açıklamalı bir Arapça külliyat elde etmek için farklı açıklayıcılardan gelen girdileri kullanan kitle kaynak kullanımına yönelik bir çözüm olarak kabul edilmektedir.

2.13.2 İsimlendirilmiş Varlık Tanımayla İlgili Çalışmalar

Doğal dil işlemede önemli bir görev olan varlık tanımda iyi sonuçlara ulaşmak, tarihsel sözlüklerden ve özellik mühendisliğinden önemli miktarda veri gerektirmiştir

(Chiu ve Nichols, 2016). Bu bölümün geri kalanında Arapça İsimlendirilmiş Varlık Tanıma konusunda yapılan önceki çalışmalardan bazıları sunulmaktadır.

Ahmed vd. (2022), isimleri köklerden belirlemek için kalıplardan yararlanan bir algoritma ortaya koymuşlardır. Yöntem Python'da yazılmıştır ve kalıpları kolayca ekleyip değiştirmek ve sonuçları hemen görüntülemek için “pyqt5” görsel modülü kullanılmıştır. Algoritmanın tanımlayamadığı isimlerin, bilinen kökler listesinde Arap kökenlerine sahip olmadığını göstermektedirler. Yöntemin köklü isimleri ne kadar hızlı ve doğru bir şekilde tanımlayabildiğini, ancak bilinen Arapça kökler listesinde köksüz olan isimleri tespit edemediğini göstermişlerdir. Son olarak birçok fikri birleştiren hibrit bir yaklaşımın kullanılmasını tavsiye etmişlerdir.

Ali vd. (2020) çalışmalarında, 2014 yılından 2018 yılına kadar sosyal medyada İsimlendirilmiş Varlık Tanıma konusunda yayınlanan birçok makalenin literatür değerlendirmesini sunarken, Arapça ile İngilizce arasındaki ayrımları da göstermeyi amaçlamışlardır. Arapça İsimlendirilmiş Varlık Tanıma sistemlerinin mevcut durumu hakkında kapsamlı araştırmalar yapmışlardır. Bu fikirleri kullanan önemli makaleleri tartışmış ve Arapça İsimlendirilmiş Varlık Tanıma'nın temel unsurlarını gözden geçirmeyi amaçlamışlardır.

Almarimi ve Enbiah 2020 yılında Libya'daki kişilerin, konumların ve eğitim kurumlarının adlarını tanımlayabilen bir sistem geliştirdiler. Başlangıç olarak, varlık adları için sözlükler ve bir eğitim derlemleri oluşturdular. Daha sonra varlık adlarının niteliklerini inceleyerek kalıplar ve standartlar oluşturdular. Uygulama daha sonra Nooj sözel dili kullanılarak tamamlandı. Eğitim derleminin istatistikleri, belirli isimlerin, üniversitelerin ve şehirlerin diğerlerinden daha yaygın olduğunu gösterdi. Sonuçların cesaret verici olduğunu ve Arapça isimli işletmelerin belirlenmesindeki zorluğun çözümüne yönelik çalışmanın hedeflerini karşıladığını ortaya koydular.

Ali vd. (2019), Arapça isimlendirilmiş varlık tanıma görevlerini çıkarmak için çoklu dikkat katmanına sahip bir sinir ağı modeli önerdi ve araştırdı. Kullanılan iki adet dikkat edilmesi gereken birim ortaya kondu. Bunlardan ilki, en iyi sözcük temsilini seçmek için karakter ve sözcük yerleştirmeyi kullanan yerleştirme dikkat katmanıdır.

Diğer dikkat birimi, model performansını artırmak için daha fazla anlama sahip kelimelere odaklanarak etiketleme tahmini görevine yardımcı olan kelime üzerinde kullanılan kodlayıcı katmanın üstünde kullanılan öz-dikkat birimidir. Bu birimler derin sinir ağlarının, Arapça gibi zengin morfolojiye sahip, özellik mühendisliği üzerinde daha az çalışma yapılan ve dış kaynaklara daha az bağımlı olan bir dil için rekabetçi performans elde edebileceğini göstermektedir. Ayrıca dikkat mekanizması yalnızca İsimlendirilmiş Varlık Tanıma ve dizi etiketleme görevi modelleri için değil aynı zamanda diğer görevler için de önemlidir.

Zirikly ve Diab (2015), tweetlerden Arap kişilerin adlarını çıkarmak için morfolojik analize veya dile özgü faktörlere dayanmayan bir model tanıtmıştır. Önerilen yaklaşımda, kural tabanlı bir model istatistiksel bir modelle entegre edilmiştir. Bir kişinin adını belirlemek için bu yöntem denetimsiz desen öğrenimi ve kümelendirilmiş sözlükler kullanır. Yöntem, ad sözlüklerini ve kümelerini denetimsiz bir istatistiksel modelle birleştirerek, bağlam unigram desenlerini çıkarmak için kullanılır. Bu desenler, insan adlarının varlığını belirtmek için kullanılır. Yaklaşımın temel dayanağı, ad kümesi ve seçilmiş kalıpları kullanarak birleşimsel kısıtlamaları öğrenmektir. Yöntem, kamuya açık kaynaklardan derlenen tam adların bir listesini kullanmıştır. Sonuçlara göre, önerilen yöntem, Arapça isimli öğeleri tweet’lerden çıkarmaya yönelik son girişimlerin tümünü geride bırakmaktadır.

2.13.3 Fiillerle İlgili Çalışmalar

Arapça fiil morfolojisi bu kadar zengin bir şekil ve manaya sahip olmasının yanı sıra Arapça bir cümlelerin inşası için esastır. Öte yandan Arapça fiiller, üç veya dört harfli köklerden, yerleşik kalıplardan biri kullanılarak çekilebilir (Shaalan vd., 2015). Kök çıkarma, doğal dil işleyen her sistemde standart bir tekniktir. Arapçada fiil tipik olarak birincil fiilin, belirli veya genel bir durumun aktardığı önceki bir durumu veya oluşumu belirtmek için tekrarlanır (Guellil vd., 2021). Bu çalışmanın geri kalan kısmında fiil tanımlamasını farklı bağlamlarda ve farklı yöntemler kullanarak inceleyen birkaç çalışma ele alınacaktır.

Ahmed vd. (2023), Arapça metinlerdeki Arapça fiilleri tespit edebilecek bir algoritma geliştirmiş ve uygulamaya koymuşlardır. Tüm fiil zamanları (geçmiş, şimdiki zaman ve emir) ile tek, çoğul, eril ve dişil zamirler on yedi farklı kalıpla temsil edilmiştir. Bir fiilin kökü olduğunda bu kalıplara uyan tüm fiiller kullanılmıştır ve sonuçlar tutarlı olmuştur. Aynı yapıya sahip tüm fiilleri tanımlamak için yöntemin kodu veya tasarımı değiştirmesine gerek olmadığını göstermişlerdir. Arapça kökler, tekniğin ayırt edemediği fiillerin öncüllerini içermez.

Alhaj vd. (2019), üç harfli edilgen geçmiş zaman fiillerin modeli Raghad Alkhudair ve Mohammad Aljutaily tarafından kapsamlı bir şekilde analiz edilmiştir. Verileri üç farklı nesilden anadili İngilizce olan 31 kişi sağlamıştır. Araştırmacılar nitel veri toplamak amacıyla gözlemlerinin yanı sıra yapılandırılmış ve yarı yapılandırılmış görüşmeler de gerçekleştirmişlerdir. Bunlar bulgulardan çıkarılacak iki ana sonucu göstermektedir. İlkinde, hibrit morf fonolojik sistemi nedeniyle, kalıpları ve prozodik yapıları bakımından Modern Standart Arapça'dan farklı olan birçok üç harfli edilgen geçmiş zaman dilbilgisi kalıpları ve bunları oluşturmak için kullanılan morf fonolojik teknikler benimsenmiştir. İkincisi, bu model, konuşmacı kuşakları arasında çeşitli artzamanlı geçişleri gösterir ve değişebilirliğin bir dilin bir özelliği olduğu hipotezini destekleyen kanıtlar sağlar.

Othman vd. (2020), Arapça bir cümleyi anlamaya yönelik bir model sunmuştur. Arapça fiilleri tanımak için düzenli ifadelerle dayalı yeni bir morfolojik model geliştirilmiştir. Üç harfli tüm Arapça fiil köklerini düzenlemek için bir komut tablosu kullanılmıştır. Arapça üç harfli köke sahip 23 090 fiil içerir. Toplamda 6 104 kök vardır. Arapça dilbilgisi, Arapça cümlelerin sözdizimini türetmek ve analiz etmek için kullanılan bir dizi kuraldır. Gösterilen fiillerin yaklaşık %87'si düzenli ifade motoru tarafından tanınabilmektedir.

Mustafa vd. (2019), mevcut morfoloji sistemlerinde kullanılan geleneksel yöntemleri kullanmadan fiil kökü tanıma ve çıkarma için RootIT adlı bir araçta geliştirilen yeni bir model sundular. Bu sistem tasarımının önemli bir yönü hem bağımsız olarak hem de diğer dilbilgisel analizörlerle verimli bir arayüz olarak çalışabilme kapasitesidir. Yüzey fiilleri, seçilen teknik kullanılarak ilişkisel veritabanında halihazırda

kullanımda olan tam ölçekli türev fiillerle eşlenir. Ayrıca önerilen sistemin değerlendirilmesinde CoreNLP sisteminden çıkarılan PATB fiillerinden oluşturulan veri seti kullanılmıştır.

Shaalan vd. (2015), Arapça fiillerin hata tespiti ve geri bildirimle ilgili sorunları ve bu sorunların, ikinci dil öğrenenler için web tabanlı akıllı bir dil öğretim sisteminin oluşturulması üzerindeki etkisini tartışmaktadır. Ana amaç, Arapça için bu sorunları ele alan ve ikinci bir dil öğrenen dil öğrencilerinin dil yeterliliklerini geliştirmelerine yardımcı olan akıllı bir dil öğretim sistemi oluşturmaktır. Öğrenciler, kötüye kullandıkları dilin yanlış amaçlarını bağımsız olarak belirleme konusunda eğitilir ve çeşitli ortam ve bağlamlarda özgürce katkıda bulunmaya teşvik edilir. Ek olarak, öğrenme deneyiminin kişiselleştirilmesine izin veren ve her hata kategorisi için öğrencinin yeterlilik düzeyine göre bilgi geri bildirimi sunan bir sistem ortaya koyarlar. Mevcut Arap morfolojik analizörleri hata teşhisine izin vermemektedir. Böylece, kısıtlamayı gevşetme ve mesafeyi düzenleme teknikleri kullanılarak öğrencilere hataya özel teşhis ve uyarlanabilir geri bildirim başarıyla verilir. Arapçadaki zayıf fiillerin karmaşık morfolojik kurallar uygulanarak oluşturulan hatalarını tespit etmek için bu yöntemlerin nasıl kullanılabileceğine dair örnekler verdiler. Öğrencinin hatalarını tespit eden ve geri bildirim veren kısımları uygulamaya koydular ve bu kısımları gerçek sınıf ortamlarından elde edilen test sonuçlarıyla başarılı bir şekilde karşılaştırdılar.

2.13.4 Çoğullarla İlgili Çalışmalar

Arapça morfolojik analiz ve türetme farklı dil düzeylerinde ve çeşitli metodolojiler kullanılarak kapsamlı bir şekilde çalışılmış olmasına rağmen, Arapça çoğul üzerine yalnızca az sayıda araştırma yoğunlaşmıştır. Bu az sayıdaki çalışma, tekil isimlerin çoğul hallerinden çıkarılmasını ve çoğulluğun tekil veya kökten türetilmesini konu alan çalışmaları ele almaktadır. Genel olarak Arap morfolojik analizindeki teknikleri ve çalışma yöntemlerini daha iyi anlamak için bu bölümde birkaç çalışma ele alınacaktır.

Mevcut kök tabanlı yaklaşımın önerilen kuralları kullanarak doğru kökleri çıkaramadığı durumu ele almak için Alshalabi vd. (2022) yeni bir düzensiz çoğul kural algoritması çözümü sunmuşlardır. Algoritma, düzensiz olarak çoğul yapılmış Arapça kelimelerin doğru köklerini çıkarmak için ana ve alt kurallar da dahil olmak üzere birden fazla kural sunmuştur. Arapça standart bir veri kümesinin kökleri alınmış ve daha sonra BPR algoritması, algoritmanın etkinliğini değerlendirmek amacıyla kök tabanlı bir Arapça kök belirleyiciyi geliştirmek için kullanılmıştır. Değerlendirme sonuçlarına göre, hatalı köklerin yalnızca küçük bir yüzdesi ayrılmış ve geliştirilmiş kök cesaret verici sonuçlar vermiştir. Kök tabanlı Arapça kök ayırıcının performansı, önerilen algoritmanın bazı düzensiz çoğul konularını çözmesiyle iyileştirilmiştir.

Arapçadaki düzensiz çoğulların kök ayrımı yapılmadan tanınmasına yönelik bir yöntem Shafah vd. (2016) tarafından önerilmiştir. Eğitim verileri olarak oldukça büyük bir Arapça külliyat üzerinde popüler karar ağacı çerçevesini (WEKA J48) kullanarak bir sınıflandırıcı (model) oluşturmuşlardır. Oluşturulan sınıflandırıcı, tanımlanamayan bir test seti kullanılarak değerlendirilmiştir. Sonuçlar, doğal dil işleme uygulamaları için umut verici bir düzensiz çoğul tanıyıcının geliştirilmesinin ve uygulanmasının mümkün olduğunu göstermektedir.

Çağdaş Arapça Derlemi'nde Dawdy-Hesterberg ve Pierrehumbert (2014), bu son derece karmaşık sistem için ne tür dilsel verilerin morfolojik genelleme açısından istatistiksel olarak önemli olduğunu araştıran tekil-çoğul çiftlerinin bir hesaplama analizi sunmuştur. Genelleştirilmiş bağlam modelini kullanan benzer bir yöntemin, herhangi bir tekil biçimin çoğul biçimini ne kadar iyi tahmin edebildiğini göstermişlerdir. Bu modelin on çapraz doğrulama döngüsü boyunca kararlılığı, dalgalanmalara karşı dayanıklılığını gösterir. Sadece bir segmentin ünsüz veya ünlü olduğunu belirten bir temsil olan CV şablonu, neredeyse tüm öngörü gücünü taşır. Bu ayrıca McCarthy ve Prince'in aruz kalıplarını Arap morfolojik sisteminden ayırma önerisini de desteklemektedir.

Ben Ismail vd. (2018), düzensiz çoğulları ele almak için baş yönlendirmeli tümce yapısı dilbilgisini kullanan bir Arapça morfolojik çözümleyici oluşturdu. Böylece, dilsel bir analizin ardından, çeşitli düzensiz çoğul kalıpları tanıyarak ve bunları ifade

etmek için baş yönlendirmeli cümle yapısı dilbilgisini kullanarak işe başladılar. İstenilen dilbilgisi bir tür tanımlama dilinde tanımlandıktan sonra, deney için bir dilsel bilgi oluşturma sistemi kullanıldı. Yöntemimiz Arapça tekil isme sahip olabilecek tüm düzensiz çoğul halleri üretebildiğinden, elde edilen sonuçlar tatmin ediciydi.

Goweder vd. (2004), Arapçadaki düzensiz çoğulların tespit edilmesine yönelik çeşitli yöntemleri araştırdı ve modern standart Arapçadaki düzensiz çoğulların tanınmasının dil mühendisliği ve bilgi erişim uygulamaları için zorlu bir görev olduğunu gösterdi. Düzensiz çoğulların tanımlanması için birçok teknik geliştirdiler ve test ettiler. Düzensiz çoğul algılama bileşeninin dahil edilmesiyle, düzensiz ve düzenli çoğulu tekil biçimleriyle karıştıran yeni bir az kök çıkarma yöntemi oluşturuldu. Son olarak en yeni az kök çıkarma tekniğinin bilgi erişim performansını değerlendirdiler

2.13.5 Kelime Üretme ile İlgili Çalışmalar

Yazılım araştırma ve geliştirmesi ağırlıklı olarak doğal dil işlemeye odaklanmaktadır. Şablon ve ek morfeplerinin karışımı, karmaşık morfolojik kurallar ve sağlam bir özellik sistemi, Arapçanın son derece zengin morfolojisini tanımlar (Habash, 2007). Arap dilinin morfolojik yapısı İngilizceye göre daha karmaşık olduğundan Arapça kök oluşturunucuların daha iyi kök çıkarma algoritmaları kullanması gerekmektedir (Alshalabi vd., 2022). Biçim ve anlam bakımından karmaşıklığından dolayı Arapça fiil morfolojisi, Arapça bir cümlenin oluşturulmasında esastır. Ancak Arapça fiiller, yaygın olarak kabul edilen formlardan biri olan üç veya dört harfli köklerden elde edilir (Shaalan vd., 2015). Kök çıkarma, doğal dil işlemeye yönelik tüm sistemlerde kullanılan rutin bir işlemdir. Kök çıkarma, kullanıldığı herhangi bir uygulamanın etkinliği üzerinde önemli bir etkiye sahip olabileceğinden, yetenekli bir kök düzeltme uygulaması çok önemlidir (Alshalabi vd., 2022). Arapça dili için terim seçiminde çok kök çıkarma, istatistiksel ve hibrit kök çıkarma, az kök çıkarma ve yapay zekâ kök çıkarma olmak üzere dört farklı kök çıkarma stratejisi kullanılmaktadır. Çok kök ayırıcılar, Arapça kelimenin kökünü izole etmek için morfolojik analiz kullanırlar (Alshalabi vd., 2022).

Hegazi (2016), Arapça kelime dağarcığının evrimi ve köklerin üretimi için yenilikçi bir yöntem önermektedir. Yöntem üç algoritma sunmaktadır; İlk algoritma permütasyon ve kombinasyon ilkesini kullanarak Arapça kelime kökleri oluşturur. Algoritma Arap alfabetik harflerine permütasyon uygulayarak kökler oluşturur. Daha sonra ikinci yöntem, Arapça morfoloji şablonunu kullanarak bu kökten farklı kelimeler oluşturmak için kullanılır; bu algoritma Arapça şablona göre kökleri oluşturarak Arapça kelimeler oluşturur.

Aqel vd. (2015), verilen köklerin neredeyse tamamının kelime olarak oluşturulmasını sağlayacak bir teknik önermiştir. Aynı teknikler kullanılarak çekimli kelimeler ve türevli kelimeler üretilebilirse, yöntem soruyu yanıtladı anlamına gelir. Önerilen metodolojinin nasıl geliştirileceğine dair çok sayıda örnek sunmuşlardır.

Alnaied vd. (2020), Arapça Morfoloji Bilgi Erişimi olarak bilinen Arapça kök bulma yöntemini tartışmışlardır. Bu yöntem, Arap harfleri arasındaki karşılıklar üzerine bir dizi kural uygulayarak Arapça bilgi erişim sistemlerinde kullanılan karşılık gelen ifadelerin kök veya gövdesini belirlemek için kullanılan dizinleme terimleri üretir. Çeşitli Arapça içerik erişim sistemleri için önerilen yönergelerin faydalarını vurgulamışlardır. Son olarak, ortalama doğruluk açısından performansını diğer yaklaşımlarla ve kök ayırıcı eşdeğeri olmayan sistemle karşılaştırarak sistemi değerlendirmişlerdir. Verilere göre sistem, Arapça kök ayırıcıyı iyileştirebilecek, iyileşmeyi hızlandırabilecek ve her türlü kök ayırıcıya karşı etkili olabilecektir.

Azman (2019), mevcut morfoloji sistemlerinde kullanılan geleneksel yöntemleri kullanmadan fiil kökü çıkarmak için RootIT adlı bir araçta geliştirilen fiil kökü tanımda devrim niteliğinde bir model önermektedir. Bu sistem için çok önemli bir tasarım faktörü hem bağımsız olarak hem de diğer dil analizcileriyle birlikte çalışabilme kapasitesidir. Seçilen yöntem, yüzeydeki fiillerin ilişkisel veritabanında uygulanan tam gelişmiş türetilmiş fiillere eşlenmesini içerir. Önerilen sistemi değerlendirmek için CoreNLP sisteminden elde edilen PATB fiillerinden oluşturulan veri seti de kullanılmıştır.

Souri vd. (2018), Arapça dilinde tekrarlayan bir sinir ağı (RNN) dil modelini eğitmek ve test etmek için “Arap Dünya Kitapları” ve “Hindawi”den ücretsiz Arapça metin veri kümelerini kullanmıştır. Arapça özellikleri ele almak için bir RNN modelini modifiye etmişlerdir. Önerileri Arapça diline duyarlı Uzun-Kısa Süreli Bellek (LSTM) modelidir. Geleneksel LSTM modeline kıyasla önerilen LSTM modelinin Arapça metin üretmede ne kadar etkili olduğunu göstermişlerdir.

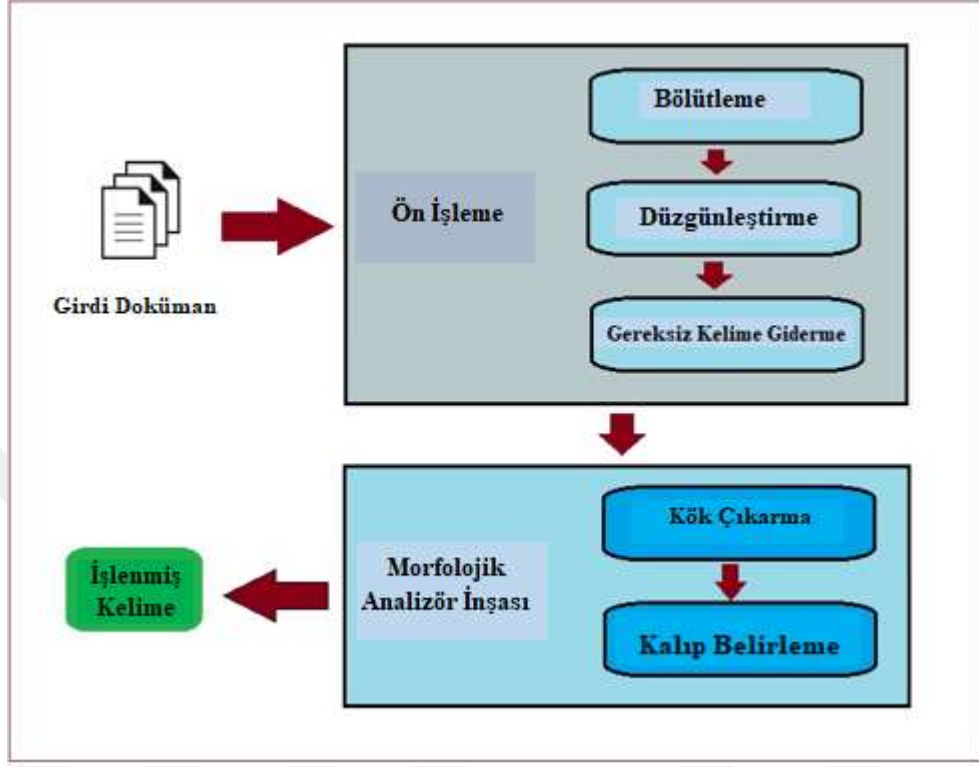
2.13.6 Özet

Bu alanda daha önce yapılan bazı çalışmalar sergilenmiştir. Arapçanın işlenişi geçmişte kapsamlı çalışma ve incelemelerin konusu olmuştur. Bilimdeki son gelişmeler, bu konularda çok sayıda çalışma ve araştırmanın yanı sıra, özellikle yapay zekâ ve doğal dil işleme alanlarında çeşitli teknikler ve algoritmalar ortaya çıkarmıştır. Arapçanın hem köke hem de kalıba ne kadar bağlı olduğu göz önüne alındığında kökün belirlenmesi, dil öğretiminin en önemli yönlerinden biri olarak kabul edilmektedir. Bu alanda yapılan pek çok çalışma bulunmaktadır. Ayrıca, özellikle antik çağ araştırmalarında çok sayıda eski yayının yalnızca dil kurallarını veya eklerin kaldırılmasını kullandığı da gözlemlenmiştir. Kelimenin baş harflerinin, kök ve eklerinin kelimedede ve kalıpta örtüşmesi birçok hataya ve bazen de önemli sorunlara yol açabilmektedir. Bu çalışmada ekleri ortadan kaldıran, gramer kurallarına uymayan ve bu hususları içeren metinlerde ayırıcı işaretleri de ortadan kaldıran alternatif bir yaklaşım önerilmiştir.

3. METODOLOJİ VE YÖNTEMLER

Desenlere Dayalı bir Arapça Morfolojik Analizi tasarlamak ve uygulamak için öncelikle birkaç temel kavramın üzerinden geçilecektir. Daha sonra tasarlanan yöntem, kullanılan algoritmalar, işin fikri ve ayrıca her görev için oluşturulan kalıplar ayrıntılı olarak açıklanacaktır. Arap dilinin temeli kökleridir. Bir kelimenin en temel şekli köküdür; kelimenin tüm ekler çıkarıldıktan sonra kalan kısmıdır ve anlamını kaybetmeden daha fazla bölünemez. Kökün birçok farklı anlamın temelini oluşturan geniş ve ana bir anlamı vardır. Kelimelerin çoğunluğu bir kök koleksiyonundan gelmektedir (Qussai ve Ismail, 2014). Bu kökler, daha karmaşık türetmeler için eklerin eklenebileceği basit isimler ve fiiller oluşturmak için çeşitli sesli harf kalıplarıyla birleştirilir. Arapça kökten türetilen kalıplar, Arapça kelimeler üretmek için kullanılır (Boudchiche ve Mazroui, 2018). Bu araştırmanın temel amacı, kalıplardan yararlanarak kelimenin köklerini tespit ederek bir morfolojik analiz cihazı oluşturmak ve kullanılan kalıptan morfolojik yapının tespitini sağlamaktır. Harf sayısı bakımından belirli kalıplar bulunsa ve aksan ve sesli harflerin eklenmesi ile ses bakımından farklılık gösterse de günümüzde Arapçada kullanılan kelimelere uygun olacak şekilde pek çok kalıp eklenmiştir. Sonraki bölümlerde tüm bu kalıplar ayrıntılı olarak ele alınacaktır. Harflerin varlığı, dizilişleri ve tipik harflerle karşılaştırılması açısından kelimenin yapısına değinilecektir. Eklerin çıkarılması işlemi için gereken zamanın yanı sıra, eklerin çıkarılması bazen bazı orijinal kelimenin temel harflerini atlayarak hatalara neden olabilir. Ekler, eklenmek veya çıkarılmak yerine kalıptan eklenir veya çıkarılır ve aynı uzunluktaki kelimeler ve kalıplar karşılaştırıldıktan sonra sonuç, yani kök gözden geçirilir. Algoritmanın konsepti basit olsa da diğer algoritmalarla çalışırken son derece etkilidir ki bu da modüler programlamanın faydalarından biri olarak görülmektedir. Eklerin kaldırılması zaman alan ve ek kuralların uygulanmasını gerektiren zor bir işlem olduğundan, nihayet kökleri hızlı ve kolay bir şekilde elde edilebilir hale gelmiştir. Model iki aşamadan oluşmaktadır: İlk aşama ön işleme aşamasıdır; bölütleme uygular, düzgünleştirmeyi gerçekleştirir ve belirteçleri gereksiz sözcüklerden ve kökten filtreler. İkinci aşama, kökün belirlenmesi, ardından kullanılan kalıbın belirlenmesi ve bu kalıbı kullanarak kelimenin kompozisyonunun ve yapısının

belirlenmesi olan önemli bir aşamadır. Şekil 3.1 önerilen modelin şemasını göstermektedir.



Şekil 3.1 Önerilen modelin şeması

3.1 Ön İşleme Aşaması

Ön işleme, çoğu doğal dil işleme uygulamasındaki bir dizi temel prosedürdür. Metinden dile bağımlı unsurları çıkarmak için kullanılır ve gereksiz kelimelerin çıkarılması, bölütleme ve kök alma işlemlerini içerir (Al-Nashashibi vd., 2010). Belgelerin ön işleme tabi tutulması veya boyutların azaltılması, diğer uygulamalarda vasıflı veri manipülasyonuna olanak tanır. Boyutsallığın azaltılması, doğal dil işleme sürecinde önemli bir adım olarak kabul edilir. Bir metnin önemsiz özelliklerinin silinmesine olanak tanır, bu da bazen sınıflandırmanın verimliliğinin en aza indirilmesine yol açar ve aynı zamanda bunların doğruluğunu ve hızını da en aza indirir (Alhaj vd., 2019). Bu ön işleme fonksiyonları tüm ön işleme işlemleri için tasarlanmıştır. İlerleyen paragraflarda bu yöntemler, algoritmaların ve fonksiyonların bu sorunları çözmek için nasıl tasarlandığı açıklanacaktır.

3.1.1 Bölütleme Süreci

Arapça bölütleme, Arapça yazılmış bir metin parçasını birimlerine veya bölümlerine ayırma işlemidir. Bu bölümler genellikle kelimelerdir ancak uygulanan belirli bölütleme kurallarına bağlı olarak noktalama işaretlerini, sayıları veya diğer dil birimlerini de içerebilirler. Metni, her kelimeyi tek başına bırakacak ve ilk boşluktan sonraki kelimeyi ayıracak şekilde bölme işlemidir; her bölüme bölüt adı verilir (Guellil vd., 2021). Şekil 3.2 bölütleme fonksiyonunu göstermektedir.

```
1. Function Tokenization(text):  
2.   // input Arabic text  
3.   // output words list  
4.   newList = [] // new list  
5.   wordList = [] // new list  
6.   wordList = text.split()  
7.   For Each word in wordList:  
8.     IF word NOT in newList:  
9.       newList.append(word)  
10.    End IF  
11.  End For  
12.  Return newList  
13. End Function
```

Şekil 3.2 Bölütleme fonksiyonu

Bu fonksiyonun girdisi bir belge veya metindir. Adım 3 ve 4, yeni iki liste ve başlatma değerleri oluştururken, adım 5 listeyi girdi metnindeki tüm bölütlerle doldurur. 7 ve 11 arasındaki adımlarda, tüm bölütler için bir döngü vardır ve tekrarlanan sözcükleri kaldırır. Çıktı, tekrarlanan sözcükleri içermeyen bir listedir.

3.1.2 Düzgünleştirme Süreci

Düzgünleştirme süreci, bir metinden istenmeyen harflerin, noktalama işaretlerinin ve harf olmayan karakterlerin kaldırılmasıdır (Alshalabi vd., 2022; Thalji vd., 2018). Genel olarak düzgünleştirme sürecinin adımları, harf olmayan karakterlerin kaldırılması, noktalama işaretlerinin kaldırılması, aksan işaretlerinin kaldırılması, bazı harflerin diğerleriyle değiştirilmesi ve bazı harflerin orijinal konumlarına döndürülmesidir. Arapçada, kelime içindeki konumlarına göre birincil karakterlerin

diğer formları olan 8 karakter ekstra karakter olarak kullanılabilir. Şekil 3.3 düzgünleştirme fonksiyonunu göstermektedir.

```
1. Function Normalization(text):
2. Import regex // Regular Expression
3. Noise = re.compile(""" Fatha|Kasra|Dama|
    Sukun|Shadda|Tanwin""",re.VERBOSE)
4. text = re.sub(noise,"",text)
5. text = re.sub("|"," ",text)
6. text = re.sub("ي"," ",text)
7. text = re.sub("و"," ",text)
8. text = regex.sub(r'^\p{Arabic}','','', text)
9. Return text
10. End Function
```

Şekil 3.3 Düzgünleştirme fonksiyonu

Girdi metindir, çıktı ise Arap harfleri dışındaki harfleri ve tüm aksan işaretleri kaldırıldıktan sonra aynı metindir. Adım 2’de fonksiyon, bir arama modelini oluşturan bir karakter dizisi olan kurallı ifadeyi kullanır ve içe aktarır. Adım 3, tüm Arapça aksan işaretlerini metinden çıkarılacak bir değişkene yerleştirir. Geri kalan adımlar, gerekli olmayan tüm harfleri ve aksan işaretlerini metinden kaldırır ve son olarak temiz metni döndürür.

3.1.3 Gereksiz Kelimeleri Kaldırma Süreci

Gereksiz sözcükler, bir gereksiz sözcük listesine dahil edilen, minimum anlam taşıyan sözcüklerdir. Ayrıca bu kelimelerin konuya ilişkin bir anlamı yoktur; yalnızca sözdizimsel kuralları tanımlamak için kullanılırlar. Bu gereksiz sözcüklerin Doğal Dil İşleme üzerinde iki farklı sonucu vardır (Alshalabi vd., 2013). Göreceli olarak yüksek frekansları nedeniyle, bilgi erişim sürecini etkileme potansiyeline sahiptirler. Ayrıca daha az yaygın olan kelimeler arasındaki frekans dalgalanmalarının etkisini de azaltma eğilimindedirler ve bu da sonuçta tartım sürecini etkiler (Abd Alameer ve Kadhem, 2017). Bu çalışmada tasarladığımız algoritmada işlem yapmaya gerek kalmayacak tüm kelimeleri içeren bir liste hazırlanmıştır. Şekil 3.4’te gereksiz sözcükleri kaldırma işlevi gösterilmektedir.

```
1. Function removeStopWords(textList)
2.   // input text as list
3.   // output text without stop words
4.   newList = [] // create new list
5.   stopWordList = ReadTextFile('StopWords.txt')
6.   For Each word in textList:
7.     IF not word not in stopWordList:
8.       newList.append(word)
9.     End IF
10.  End For
11.  Return (newList)
12. End Function
```

Şekil 3.4 Gereksiz kelimeleri kaldırma fonksiyonu

Bu fonksiyonun açıklaması net ve basittir, çünkü tüm gereksiz sözcükler adım 4'teki gibi metin dosyasından okunur ve bir listeye yerleştirilir ve ardından adım 5'te yeni bir liste oluşturulur. Bundan sonra, 5 ve 10 arasındaki adımlar, bu fonksiyona girilen listedeki tüm kelimeleri okuyan ve ardından her bir kelimeyi, gereksiz kelimeler listesi (Adımlar (7-9)) arasında olup olmadığını görmek için test eden bir döngü içerir. Listede yoksa Adım 8'deki yeni listeye eklenir. Ön işlemenin tüm fonksiyonlarının çıktısı, diğer doğal dil işleme uygulamaları veya sınıflandırması için kullanılabilecek temiz Arapça kelimelerin bir listesidir.

3.2 Tasarım Aşaması

Bu aşama, çalışmamızın en önemli konusu olan algoritmanın yapısının ve adım adım nasıl çalışabileceğinin daha detaylı olarak ele alınacağı aşamadır. Algoritmanın tasarımı Arapça morfolojiden, özellikle kelime, kökü ve kalıbı arasındaki ilişkiden yararlanmaktadır. Yöntem, algoritmanın kendisi ve diğer üç fonksiyonun bulunduğu ana program olarak tasarlanmıştır. Algoritma fonksiyonları çağırır ve onlar da birbirlerini çağırabilirler. Prosedürel programlama, her iş için görevleri ayırarak ve fonksiyonları açıklayıcı, basit ve öz olacak şekilde, etkinlik ve anlaşılabilirliği artırmak amacıyla kullanılmıştır. Ağaç işlevleri “checkPattern”, “returnRoot” ve “returnWord” olup, her birinin belirli bir görevi vardır ve tümü, kökü tespit etmek veya çıkarmak ve Arapça kelimelerin morfolojik yapısını bilmek için tüm işi yapmak üzere

algoritma ile birlikte çalışabilir. Sonraki birkaç bölümde algoritmalar ve fonksiyonlar daha ayrıntılı olarak tartışılacaktır.

3.2.1 Kök Tespiti Algoritması

Bir kelimenin en temel şekli köküdür; kelimenin tüm ekler çıkarıldıktan sonra kalan kısmıdır ve anlamını kaybetmeden daha fazla bölünemez. Arapça kök oluşturmak için üç harfli veya daha uzun bir alfabetik dizi düzenlenir. Kökün birçok farklı anlamın temelini oluşturan geniş ve ana bir anlamı vardır (Boudchiche ve Mazroui, 2018). Bu kökler, daha karmaşık türetmelerde eklerin eklenebileceği basit isimler ve fiiller oluşturmak için çeşitli sesli harf kalıplarıyla birleştirilir. Arapça kökten türetilen kalıplar, Arapça kelimeler üretmek için kullanılır (Boudchiche ve Mazroui, 2018). Şekil 3.5'te algoritmanın sorgusu bulunmaktadır.

```
1. Algorithm detect_roots(text):
2.   // input : Arabic text
3.   // output: List of detected roots.
4.   patterList = []
5.   rootList = []
6.   text = Preprocessing(text)
7.   patternList = ReadTextFile('Patterns.txt')
8.   For word in List_of_words:
9.     For currentPattern in patternList:
10.      IF (len(currentPattern) == len(word)):
11.        IF (checkPattern(currentPattern, word) == TRUE:
12.          Root = returnRoot(currentPattern, word)
13.          currentWord = returnWord(currentPattern, root)
14.          IF (word == currentWord):
15.            rootList.append(root, currentPattern)
16.          End IF
17.        End IF
18.      End IF
19.    End For
20.  End For
21. End Algorithm
```

Şekil 3.5 Kök tespit algoritması

Şekil 3.5’te görebileceğimiz gibi, bu algoritmanın girdisi Arapça bir metin, çıktısı ise tespit edilen köklerin bir listesidir. Adım 4’ün amacı, adım 7’deki gibi metin dosyasından okuma kalıpları için yeni bir liste oluşturmaktır. Adım 5’te, giriş metninden algılanan tüm köklerin içine yerleştirilecek boş liste oluşturulur. 8 ve 20 arasındaki adımlardan itibaren tüm kelimelerin köklerini tespit edecek bir döngü vardır, 9 ve 19 arasındaki adımlar bu algorithmada kullanılan tüm kalıpları getirmek için bir iç döngüdür ve 10 ve 18 arasındaki adımlar sadece kalıpların girilen listedeki mevcut kelimeyle aynı uzunlukta olup olmadığını kontrol eder. Adım 11’de, Şekil 3.6’da açıklanan “checkPattern” fonksiyonuna çağrı yapılır, eğer fonksiyondan dönen değer doğruysa, daha sonra 12 ve 13 adımlarındaki gibi “returnRoot” ve “returnWord” fonksiyonlarına çağrı yapılır. Adım 14’te koşul, orijinal kelimeyi adım 13 sonucunda gelen başka bir önerilen kelimeyle karşılaştırarak kökün doğru olup olmadığına karar verecektir. Son olarak, eğer kök algoritma koşullarını karşılıyorsa, adım 15’teki gibi kökler listesine eklenecektir.

3.2.2 Check Pattern Fonksiyonu

Bu fonksiyon, gönderilen kalıbın, temel harfler ve ek harflerin uygun yerlerine düşme olasılığı açısından kelimeye uygun olup olmadığını test etmek için tasarlanmıştır ve mantıksal olarak “true” değerini döndürür, aksi halde herhangi biri farklıysa “false” mantıksal değerini döndürür. Bu fonksiyonun temel avantajları hız ve doğruluğun yanı sıra Şekil 3.6’daki gibi basitlik ve tasarım kolaylığıdır.

```

Function returnWord(pattern, root):
  word = ""
  For i in range(len(pattern)):
    IF ( pattern [i] == "F" ):
      word += root[0]
    Elseif ( pattern [i]=="A" ):
      word += root[1]
    Elseif ( pattern [i]=="L" ):
      word += root[2]
    Else:
      word += pattern [i]
    End IF
  End For
  Return word
End Function

```

Şekil 3.6 Check pattern fonksiyonu

Fonksiyon, girdi olarak hem deseni hem de kökü alır ve kelimeyi çıktı olarak döndürür. Buradaki fikir, bir kalıptaki tüm harfleri kontrol etmektir ve eğer bu harf, kalıbın ana harflerine aitse o zaman kelimenin harfi, kökteki karşılık gelen bir harften gelecektir, aksi takdirde harf, kalıptaki karşılık gelen harften gelecektir. Bu işlev çok hızlıdır çünkü “FOR” ifadesinin içinde normalde dört ila dokuz adımdan oluşan yalnızca bir “IF” ifadesi içerir.

3.3 Kalıp Modeli

Önerilen programın en önemli özelliklerinden biri kökü bulunacak tüm kelimelere uygun kalıpların kullanılmasıdır. Arap edebiyatını incelerken tespit edilebilecek üslupların sayısı oldukça azdır, toplamda 15 civarındadır. Bu çalışmada Arapçadaki temel kelimelere karşılık gelecek yeni kalıplar oluşturulmuş ve ayrıca bu kalıplara kelime yapılarına daha uygun hale getirmek için ekler de eklenmiştir. Kökleri tespit etmek için Tablo 3.1’de gösterildiği gibi 504 kalıp kullanılmıştır.

Tablo 3.1 Kökleri tespit etmek için kullanılan kalıplar

فعلنا	فعلن	فعللي	فعلل	فعلكم	فعلك	فعل
فعلون	فعلوا	فعلو	فعلهم	فعلهما	فعلها	فعله
فعلين	فعليكما	فعلينتا	فعلية	فعللي	فعالا	فعللي
فعاليات	فعالي	فعاله	فعالك	فعالة	فعول	فعالات
افعلت	افعله	افعلاء	للافتعال	فيعل	افعلي	فوا عليهم
افعالهم	افعالهن	افاعل	افتعل	افتعلت	افتعلهم	افعللي
استفعل	استفعلوا	استفعلوا	استفعل	استفعلها	افتعالهم	الفعل
الفعله	الفعلاني	الفعللة	الفعلن	الفعللية	الفعللي	الفعللي
الفعالة	الفعالات	الفعال	الفعليين	الفعليون	الفعلين	الفعلية
الفاعلين	الفواعل	الفواعل	الفيعال	الافعل	الافعله	الافعلين
تستفعل	الافتعالي	الافتعالية	تفعيلها	الانفعال	التفعلة	التفعل
تفاع	المفعلة	المفعولون	المفعلين	المفعولون	المفعول	تفعيلية
انفعل	بفعول	بفعلهم	المفعول	المستفعله	بفعولة	المفاعلة
بافعال	بالفعل	بالفعالات	بالافعال	بفعلاتهم	بالتفعل	بفاعلة
تفعلاها	تفعلتي	بتفعيله	بتفعيل	المفعولات	تفعلا	تفعيله
فنفعل	فأفعلهم	فأفعلنا	ففعلوهم	الاستفعل	ففعلوا	ففعلنا
للمفعّل	للمفعول	مفعلة	لافعال	المفعلات	كفعل	كفعلة
مفعلا	مفعّل	لمفعّل	للمفعلة	الافتعالات	للفعل	للفاعلة
مفعّلين	مفعولّين	مفعولون	مفعّلها	التفعيلات	مفعّله	مفعولون
يفعلها	مفتعلة	مفاعيل	مفاعل	الفاعلات	مفعّاله	مفعلي
يفعله	يتفعل	مفتعل	مفاعله	باستفعل	مفاعلة	مفاعلات
يستفعل	افعل	يفعلان	يفعلا	بالفاعلين	يفاعلون	يفعل
فعلونه	المفتعلة	يستفعلون	فعلاته	بالمفعول	يتفعلون	الفعولية
سيفعل	مفاعنا	فسيفعلكم	ففعلهن	فأفعلاتكموه	لتفعلوا	افعالكم
بالمفعلة	ليفعولون	فأفعلوا	فلنفعلناك	مفاعيعها	بفعلات	فاعلك
لفعلت	وليفعل	فاعلي	فأفعلون	مستفعله	فعلتم	افعلين
فأفعلها	تفاعلوا	يفاعلوكم	بالفعلاء	تفاعلوهم	فعلكم	فعلان
فأفعلت	بالفعول	ففعلوكم	فعلنا	ستفعلونهم	يفعلنا	فعلنا
بفعلهم	لمفتعلن	للفعلاء	فأفعلوه	فأفعلكم	فعلكم	فعلين
بالفعال	افعلتكم	تتفعلون	فعلهن	مستفعلون	بفعلين	افعلوهم
استفعلتم	تفاعلت	يتفاعلا	ففعله	تستفعلوا	بأفعلهن	مفعولا
ففعلناها	بفعلته	كفعلكم	يفعلوها	تفاعلوهم	بمفعول	فأفعال
فأفعلوا	فعلوه	ففتعلنا	ففعلوها	اتستفعلون	فعالها	بالفعلة
بالفاعل	يفعلونكم	فعلتكم	فعلياكم	ففعلتكم	بفعله	فأفعلت
افاعلهم	افعالا	انفعل	يفعلكم	افعلوكم	بفتعالكم	تفعلا
وفعلها	وفعيلة	فعلتموهم	فيعلهم	يفاعلونكم	افوعلها	يفعلونك
فعلولا	فوعل	فعولة	المفعولة	فعلوها	فعلناهم	فعلناهم
فعلناهم	فاعلين	فاعلية	المفاعل	فعلهم	فعلولون	فعل
فعال	فاعلين	فاعلات	المفاعلات	فعيلة	فعيلا	فعيلي
فعالية	فاعلون	فاعلا	المفاعلة	فاعلة	فاعلون	فعيلها
افعلوا	افعلهم	افعلها	المتفاعل	افعلكم	افعلت	افعلت

Tablo 3.1'in devamı.

افتعال	أفعالية	أفعال	المتفعلة	أفعال	أفعالية	افتعال
الفعلاء	أفعال	الفعال	المفتعل	افتعاله	أفعال	الفعلاء
الفعولون	الفعلات	الفعلان	المفتعلين	الفعول	الفعولة	الفعولون
الفعيل	الفعليات	الفعائل	المفتعل	الفاعل	الفاعلة	الفعيل
الفعيلة	الفعلة	الفاعلون	المفاعلون	الفاعل	الفاعلة	الفعيلة
الفعليات	التفعيل	الافعال	المفاعلين	الافعال	التفعيل	الفعليات
تفعله	التفاعل	الفاعل	المفتعلات	التفعيلية	التفاعل	تفعله
تفعليها	تفعّل	بفعالنا	بفعالنا	تفعّلنا	تفعّل	تفعليها
بالفعالو	تفعّل	بأفعالهم	بأفعالهم	بأفعاله	تفعّل	بالفعالو
ففعلت	تفعّل	تفعّل	بمفاعلة	تفاعلا	تفعّل	ففعلت
ففعّلهم	فيفعل	كفاعل	تفاعلا	كفاعل	فيفعل	ففعّلهم
لفعلنا	لفعلوها	للمفعلين	فيفعلون	للمفعلين	لفعلوها	لفعلنا
للمفعّل	ليفعل	ليفعل	كفعلات	ليفعل	ليفعل	للمفعّل
مفاعلا	مفعول	مفعولي	لاستفعال	مفعولي	مفعول	مفاعلا
نفعّل	متفعلا	نفعله	متفاعلة	متفاعلا	متفعلا	نفعّل
بفاعلكم	بفعولون	بفعّل	بفعولونهم	بفعّل	بفعولون	بفاعلكم
فبعلية	بالفاعلة	الفاعلي	يفاعلان	يفاعلي	بالفاعلة	فبعلية
يتفعّل	المتفعّل	تفعيلات	يتفعّلان	يتفعّلان	المتفعّل	يتفعّل
فعليه	لنفعّل	مفتعل	الفعلا	مفتعل	لنفعّل	فعليه
بفاعل	افعلنا	ويفعّلكم	يفعلونه	بفاعلين	يفعلهم	بفاعل
افعلهن	الفعالا	ففعّل	فاعلو هن	ففعّلنا	مفعولات	تفعّلوها
يفعلّك	انفعّل	نفاعّل	ففعّلوهن	نفاعّل	ففعّلها	ففعّلها
توفعلنا	تفعّلوه	لأفعالهم	ففعّلناكم	ففعّلناكم	بفعليه	ففعّلها
ففعليته	انفعّلون	ففعّلنا	تفاعلتهم	ففعّلنا	مفعّلكم	ففعّلكم
يفعلّن	ففعّلون	بفعّلها	انفعّلونهم	بفعّلها	ففعّلناكم	لأفعّلكم
تفعّلن	يفعلّك	تفعّلون هن	ففاعلو هن	مفاعلكم	قفعّلونا	فاعلو
يفعلّوا	مفعّلهم	افعلّوه	المتفعّلين	بفعّلك	ففاعلتهم	افعلّتم
تفعّلتم	افتفعّلون	ففعّلناكم	ففعّلناكم	بفعّلكم	تفاعلو	افعلّوني
متفاعلا	لفعيلة	بفعلي	فيتفعّلون	ففاعّل	ففعّلكم	ففعّلكم
فتعّلتم	ففاعّلناكم	ففعّلناكم	ففاعّلها	للفعالة	ففاعّلهم	افعلّتهم
وفعّل	وفعّلها	وفعّلها	وفعّلها	وفعّلها	وفعّلها	وفعّلها

Yeni kalıplar oluşturma temel ve önemli nedeni, farklı kelimeleri tüm ekleriyle birlikte köke sığdırmaktır. Bunun faydası, sonuçların doğruluğu ve algoritmanın mekanizmasının anlaşılmasının kolaylığının ve kök bulmadaki belirsizlik ve karışıklığın ortadan kaldırılmasının yanı sıra işlem hızıdır. Aynı kalıp, aynı harf sayısına ve aynı morfolojik yapıya sahip tüm kelimeler için işe yarayabilir. Her kelimenin birkaç harfi vardır, bu harflerin bir kısmı kökü oluşturan temel harfler, geri kalan harfler ise genel formun bir bileşimidir.

3.3.1 Arapça Gereksiz Kelimelerin Kaldırılması

Üç tür Arapça sayı vardır: tekil, ikili ve çoğul. Arapça çoğullar da üçe ayrılır; Eril Düzenli Çoğul, Dişil Düzenli Çoğul ve Düzensiz Çoğullar (Ameur vd., 2020). Tablo 3.2, Tablo 3.3 ve Tablo 3.4'te sırasıyla her çoğul türü için kalıplar yer almaktadır.

Tablo 3.2 Eril düzenli çoğul kalıpları

ويفعلون	ومفعلين	مفعلين	مفعولون	فيفعلو	يفعلون
يفاعلون	بمفعلين	للمفعلين	والمفعلين	والمفعولون	المفعلين
الفاعلين	الفاعلون	وللفاعلين	للفاعلين	فاعلين	فاعلون
تفعلون	وبالفاعلين	بالفاعلين	مستفعلون	والفاعلين	والفاعلون
يفتعلون	اتفعلون	وفعلون	فعلين	فعلون	وتفعلون
ويتفعلون	يتفعلون	يستفعلون	افتعملون	اتستفعلون	ويستفعلون
فعالين	للفعلين	والفعلين	والفعلون	الفعلين	الفعلون
تتفعلون	فللفاعلين	الفعالين	تفاعلون	والافعالين	افعالين
افعلون	فيعلين	بفاعلين	ليفعلون	المتفعلين	تفاعلون

Tablo 3.3 Dişil düzenli çoğul kalıpları

الفعلات	فعالات	مفعولات	وفعلات	فعلات
والفاعلات	الفاعلات	والفعالات	بالفعلات	بفعلات
الفعالات	والفعلات	المفعلات	وللمفعلات	والمفعلات

Tablo 3.4 Düzensiz çoğul kalıpları

فعليهم	فعلياكم	افو عليكم	افعال	فعال
فعلوا	فأفعلنا	بفعليهم	فعلهم	افعال
فعلناهم	مفعلهم	افعلناكم	افعالهم	فعلة
الفواعل	يفعلونهم	افعلوا	افو علها	افعلنا
بفعلهم	يفعلهم	فعلتكم	فعلولها	فعلول
الافعال	تفعلتم	يفعلونكم	مفاعيل	تفاعيل
الفعالية	افعلواهم	فعلتم	فعلولهم	مفاعيل
فأفعلوا	فعلة	بأفعلناكم	افعلنا	فعلاء
فعلتنا	فعلولكم	فأفعلوا	افعالكم	فعلاءكم
بفعليهم	ليفعلنا	فعلونا	الافعل	فعالككم

3.3.2 Fiil Kalıpları

Arapçada harf oluşumu bakımından ve aksansız 9.300 civarında fiil bulunmaktadır. Bugün kullanılan dilde, özellikle basında ve sosyal medyada kullanılmayan çok sayıda fiil vardır. Tablo 3.5 genel olarak fiilleri tanımlamak için kullanılan kalıpları içermektedir.

Tablo 3.5 Fiiller için kalıplar

فاعِل	فَعِلُوا	فَعِلن	يَفْعَلون	وَيَفْعَلون
فَعِلت	فَعِلَا	وَيَفْعَل	تَفْعَل	أَفْعَل
يَفْعَلون	يَفْعَلن	تَفْعَلن	يَفْعَلن	أَفْعَلِي
أَفْعَلَا	أَفْعَلُوا	أَفْعَلن	وَفْعَلت	يَفْعَل

3.3.3 Varlık İsmi Kalıpları

Önceki yöntemlerle, desenler isimlere, özellikle de kişi isimlerine uyacak şekilde tasarlanmıştır. Şehir isimlerinin ise herhangi bir kuralı yoktur ve bu isimler genellikle başka dillerden gelmektedir, bu nedenle kullanılan yöntem uygun değildir. Tablo 3.6, adları tanımlamak için kullanılan kalıpları göstermektedir.

Tablo 3.6 Varlık isimlerine ilişkin kalıplar

فَعِيلَة	فَعُول	فَعِيلَة	فَعْلَاء	فَاعِلَة
فَعْلِي	أَفْتَعَل	أَفْعَال	أَفْعَل	مَفْعَل
فَاعِل	فَعِيل	فَعْلُو	عَبْد المَفْعَل	عَبْد المَفْعَلن
عَبْد الفَعْل	عَبْد الفَع	عَبْد المَفْعَلِي	عَبْد المَفْعَل	عَبْد الفَعُول
عَبْد الفَعِيل	عَبْد الفَاعِل	عَبْد الفَعَال	عَبْد المَفْعَل	عَبْد الفَعْلن
فَعَال	فَعْلَان	مَفْعُول	فَاعِل	مَفْعَل
مَفَاعِل	مَفْعَال	فَعْلَة	فَعْلون	فَعْلت
مَفْعَل	فَعْلَان	فَعِيل	فَاعُول	فَعْلِي
فَعْلَات	الفَاعِل	الفَعِيل	المَفْعَل	مَفْعُولَة

3.4 Özet

Bu bölüm, ana programın verilen fonksiyon koleksiyonunu nasıl çağırdığını açıklar, bunların bazıları görevin gerektirdiği işleme dayanarak diğer fonksiyonları çağırır. Prosedürel programlama kullanılmıştır ve kullanılan fonksiyonların tümü farklı ve anlaşılır işler gerçekleştirmektedir. Bir uygulamanın etkili olabilmesi için aşamalar arasındaki birçok görevin tamamlanması gerekir. En iyi sonuçları elde etmek için deneyimlenmiş yöntemlerin ve uzman taktiklerin kullanılmasına büyük özen gösterilmiştir.

Python şu anda makine öğrenimi ve doğal dil işleme de dahil olmak üzere birçok sektör ve alanda en yaygın kullanılan programlama dillerinden biri olduğundan, önerilen algoritma ve tüm fonksiyonlar bu programlama dili tarafından uygulanmaktadır. Son olarak, algoritmayı Grafik Kullanıcı Arayüzünde (GUI) uygulamak için (pyqt5) paketi

kullanılmıştır; bu, sonuçların doğrudan görülmesine ve hem kalıpları hem de gereksiz sözcükleri kolayca değiştirebilmesine veya kaldırabilmesine yardımcı olmuştur.



4. DEĞERLENDİRME VE DENEYSEL SONUÇLAR

Doğal Dil İşleme algoritmalarının bilgi sistemleri performansındaki etkinliğinin ve doğruluğunun değerlendirilmesi oldukça önemli bir konudur. Farklı kriter ve metrikler kullanarak algoritmaları ölçmenin birçok yolu vardır. Doğal Dil İşleme algoritmaları için iki ana değerlendirme yöntemi türü vardır: içsel ve dışsal. İçsel değerlendirme, algoritmanın dilbilgisi, sözcük dağarcığı, sözdizimi ve anlambilimi gibi iç kalitesine ve özelliklerine odaklanır. Dışsal değerlendirme, Doğal Dil İşleme algoritmasının işlevselliği, kullanılabilirliği ve kullanıcı memnuniyeti gibi dış etkisine ve kullanışlılığına odaklanır. Hangi yöntemin seçileceği amaca ve kapsama bağlıdır; amaç metin verilerini Doğal Dil İşleme ile analiz etmekse, algoritmanın doğruluğunu ve eksiksizliğini ölçmek için içsel değerlendirmeyi kullanmak gerekecektir. Göreve ve etki alanına bağlı olarak algoritmalar için birçok farklı ölçüm vardır. Bu bölümde, bu çalışma için tasarlanan algoritma, böyle bir duruma uygun bazı metrikler kullanılarak değerlendirilecek ve ardından algoritmanın karşılaşılabileceği tüm olası durumları açıklamak için örneklerle resimlendirilen pratik deneyler yapılacaktır.

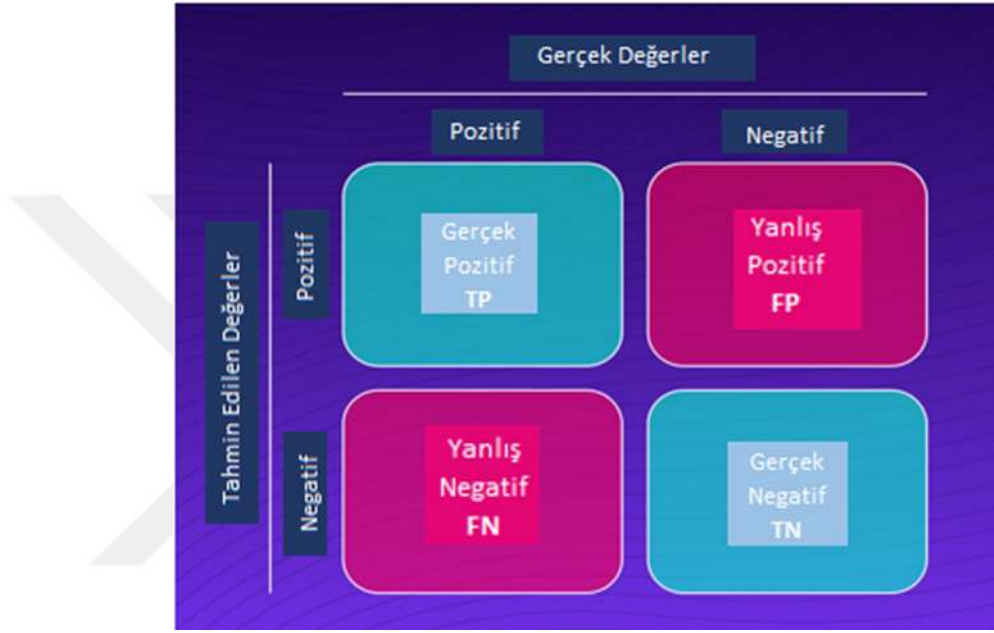
4.1 Doğal Dil İşleme Algoritması Metrikleri

Değerlendirme ölçümleri, Doğal Dil İşleme sistemlerinin performansını ve etkililiğini değerlendirmek için kullanılan niceliksel ölçümlerdir. Bu ölçümler, belirli bir sistemin makine çevirisi, duyarlılık analizi veya isimlendirilmiş varlık tanıma gibi amaçlanan görevi ne kadar iyi yerine getirdiğini değerlendirmeye yardımcı olur. Doğal Dil İşleme algoritmalarını değerlendirmek için kullanılan ortak metriklerden bazıları anlatılacak ve tanımlanacaktır, çünkü bunlar tasarladığımız ve uyguladığımız algoritmayı değerlendirmek için kullanılacaktır, ancak öncesinde karışık matris tanımlanacak, daha sonra bu çalışmada kullanılan bir veri kümesi tartışılacaktır.

Karışıklık Matrisi:

Bir ekseninde gerçek değerler ve diğerinde tahmin edilen değerler ile ikili sınıflandırma için kullanılan, 2x2 boyutunda bir matristir. Bu bir değerlendirme ölçütü değil, değerlendirmeye yardımcı olabilecek bir kayıttır. Tahmin edilen ve gerçek sonuçları

karşılaştırır. Algoritma sonuçlarını görselleştirme yöntemidir. Daha spesifik olarak bu tablo, gerçek örneklerin sayısını ayrıntılı bir şekilde gösteren bir tablodur. Diğerlerinin yanı sıra hassasiyet ve geri çağırma gibi diğer birçok model performans ölçümünü hesaplamak için kullanılabilir. Karışıklık matrisi, sınıflandırma problemlerinin çözümünde kullanılan çok popüler bir ölçümdür. Şekil 4.1 karışıklık matrisinin yapısını göstermektedir.



Şekil 4.1 Karışıklık Matrisi

Tablo 4.1, Kesinlik, Geri Çağırma, F1 Puanı ve Doğruluğu hesaplamak için kullanılacak bir karışıklık matrisindeki dört değeri veya parametreyi açıklamaktadır.

Tablo 4.1 Karışıklık matrisi parametreleri

İsim	Tanım
TP (Gerçek Pozitif)	Doğru şekilde tespit edilen pozitif vakaların sayısı,
TN (Gerçek Negatif)	Doğru şekilde tespit edilen negatif vakaların sayısı
FP (Yanlış Pozitif)	Yanlış tanımlanan pozitif vakaların sayısı
FN (Yanlış Negatif)	Yanlış tanımlanan negatif vakaların sayısı

Kesinlik:

Gerçek pozitiflerin (doğru tespit edilmiş vakalar) pozitif olarak tanımlanan toplam vaka sayısına oranıdır. Kesinlik, pozitif olarak tanımlanan vakalardan kaçının pozitif olduğunu ölçmek için kullanılır.

$$\text{Kesinlik} = \text{TP} / (\text{TP} + \text{FP})$$

Geri Çağırma:

Gerçek pozitiflerin toplam pozitif vaka sayısına oranıdır. Geri çağırma, pozitif vakalardan kaçının model tarafından doğru şekilde tanımlandığını ölçmek için kullanılır.

$$\text{Geri çağırma} = \text{TP} / (\text{TP} + \text{FN})$$

F1 puanı:

F1 puanı, kesinlik ve geri çağırmanın harmonik ortalamasıdır ve kesinlik ile geri çağırma arasındaki dengeyi sağlamak için kullanılır. 0 ile 1 arasında değişir ve 1 mümkün olan en iyi puandır. F1 puanı genellikle bir modelin genel performansını değerlendirmek için kullanılır.

$$\text{F1 puanı} = 2 * (\text{Kesinlik} * \text{Geri Çağırma}) / (\text{Kesinlik} + \text{Geri Çağırma})$$

Doğruluk:

Doğru sınıflandırılmış vakaların toplam örnek sayısına oranıdır. Doğruluk genellikle bir modelin genel performansını ölçmek için kullanılır.

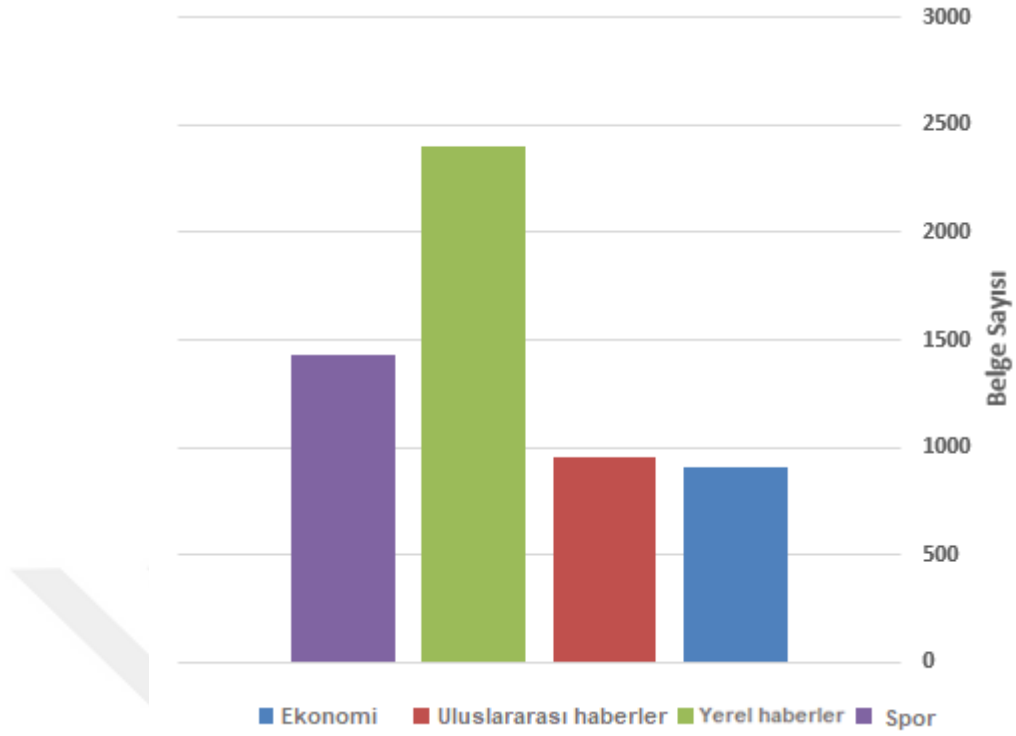
$$\text{Doğruluk} = (\text{TP} + \text{TN}) / (\text{TP} + \text{TN} + \text{FP} + \text{FN})$$

4.1.1 Veri Kümesi

Veri kümesi, genellikle standart bir formatta düzenlenen ilgili verilerin bir koleksiyonudur. Veri kümeleri analitik, iş zekâsı, yapay zekâ (AI) model eğitimi ve diğer çeşitli kullanım durumları için kullanılır. Veri kümeleri, verinin hem boyutu hem de türü açısından önemli ölçüde farklılık gösterebilir. Örneğin bir veri seti, ağaç türleri, okyanus sıcaklıkları, bölgesel satış toplamaları, meyve fiyatları, piyango talihlileri, hastalıklar veya diğer herhangi bir veri türü hakkında bilgi içerebilir. Yüksek kaliteli veri kümeleri, Doğal Dil İşleme projelerinde iyi performansın anahtarıdır. Arapça gibi bulunması zor, düşük kaynaklı dil veri kümeleri olmasına rağmen internette ücretsiz bir derlem bulunabilmektedir. Arapça bir derlem olan “Khaleej-2004”, ücretsiz bir şekilde <http://sites.google.com/site/mouradabbas9/corpora> (URL-1, 2020) sitesinden indirilebilir. Derlem 5.690 belge içerir. Ekonomi, Uluslararası haberler, Yerel Haberler ve spor olmak üzere dört konu kategorisine ayrılmıştır. Derlemdeki belgeler “HTML” biçiminde olduğundan metin dosyalarına dönüştürülmeleri gerekmektedir. Tüm dokümanlardaki kelime sayısı 2.472.763 olup, bunların 115.801’inin kopyası yoktur, bunlar kontrol ve test edilecektir. İstatistikler ve veri sınıfı dağılımı Tablo 4.2 ve Şekil 4.2’de açıklanmıştır.

Tablo 4.2 Veri kümesinin istatistiği

Kategori adı	Belge Sayısı
Ekonomi	909
Uluslararası Haberler	953
Yerel Haberler	2.398
Spor	1.430
Toplam	5.690



Şekil 4.2 Dokümanların kategorisel dağılımı

4.3 Örnek Üçüncü Düzey Alt Başlık

Algoritmayı daha önce bahsedilen ölçümlerle değerlendirmek için daha önce belirtilen değerler hesaplanır: TP (Doğru Pozitif), TN (Gerçek Negatif), FP (Yanlış Pozitif) ve FN (Yanlış Negatif). Daha sonra, her ölçüme ilişkin denklemleri hesaplamak ve sonuçları karşılaştırmak için bu hesaplamalar kullanılır. Tablo 4.3 bu metriklerin değerlerini içermektedir.

Tablo 4.3 Parametre değerleri

Ölçek	Değer
TP (Gerçek Pozitif)	97.852
TN (Gerçek Negatif)	8.453
FP (Yanlış Pozitif)	6.716
FN (Yanlış Negatif)	2.780
Toplam	115.801

Tablo 4.4, önerilen yöntemi değerlendirmek için yapılan tüm metrik hesaplamaları içermektedir.

Tablo 4.4 Metrik hesaplamaları

Metrik	Denklem	Hesaplama	Değer
Kesinlik	$TP / (TP + FP)$	$97.852 / (97.852 + 6.716)$	%0,94
Geri Çağırma	$TP / (TP + FN)$	$97.852 / (97.852 + 2.780)$	%0,97
F1 puanı	$2 * (Kesinlik * Geri Çağırma) / (Kesinlik + Geri Çağırma)$	$2 * (0,94 * 0,97) / (0,94 + 0,97)$	%0,95
Doğruluk	$(TP + TN) / (TP + TN + FP + FN)$	$(97.852 + 8.453) / (97.852 + 8.453 + 6.716 + 2.780)$	%0,92

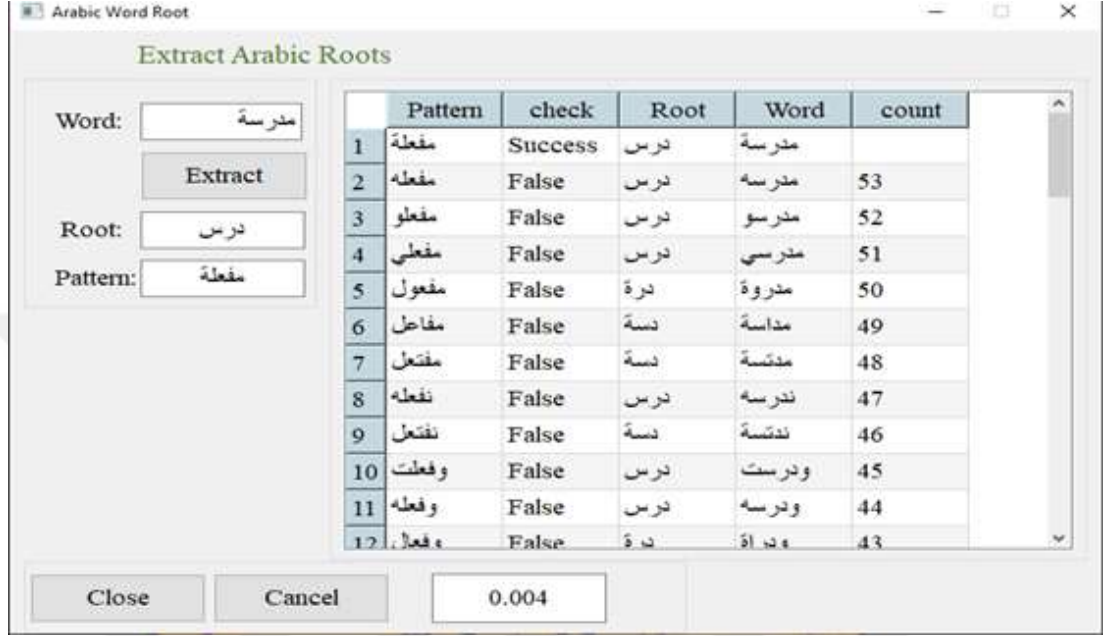
Tablo 4.3'te gösterildiği gibi, algoritmanın pozitif olarak tespit ettiği tüm durumlar arasında doğru pozitif vakaların yüzdesi, kesinliği, %94'tür. Genellikle hassasiyet olarak adlandırılan geri çağırma, veri kümesindeki gerçek olan tüm olumlu örnekler arasında doğru şekilde tahmin edilen pozitif vakaların yüzdesini ölçer. Algoritma, geri çağırma oranıyla gerçek pozitif vakaların yüzdesini %97 olarak hesaplamıştır. F1 puanı kesinlik ve geri çağırmanın harmonik ortalamaları olarak hesaplanır. Yüksek F1 puanı, modelin geri çağırma ve kesinlik açısından iyi performans gösterdiği anlamına gelir. %95'lik F1 puanı, iyi dengelenmiş bir performansı göstermektedir. Doğru şekilde tahmin edilen örneklerin veri kümesindeki tüm örneklerle oranı, doğruluğu hesaplamak için kullanılır; algoritma, veri kümesindeki durumların %92'sini doğru bir şekilde algılamıştır.

4.4 Deney Sonuçları

Uygulama Python 3.9.0 programlama dili kullanılarak hayata geçirilmiştir. “pyqt5” paketi, uygulamanın çalışmasını doğrudan simüle etmek amacıyla GUI'yi temsil etmek için kullanılmıştır. Bu fikir, varsa hatalardan faydalanmak ve bunların yerini kolay ve doğru bir şekilde belirlemek için kullanılabilir. Türetilmiş kökün doğru olup olmadığını bilmenin dört spesifik olasılığı vardır; birincisi, kök doğrudur ve algoritma bu kökü tespit etmektedir; ikincisi, kelimenin kökeni Arapça değildir ve bu nedenle kökü yoktur ve algoritma onu tespit etmemektedir; üçüncüsü, kelimenin bir kökü vardır ama algoritma onu bulamaz; sonuncusu da kelimenin doğru bir kökü vardır ama algoritma keşfetme işlemini atlar ve yanlış kökü bulur. Bütün bunları dört farklı örnekle anlatacağız.

4.4.1 Birinci Örnek

Bu örnekte, Şekil 4.3'te gösterildiği gibi, test edilen kelimenin doğru bir kök içermesi ve algoritmanın onu doğru algılaması durumu olan ilk olasılık sunulmaktadır.



	Pattern	check	Root	Word	count
1	مفعلة	Success	درس	مدرسة	
2	مفعلة	False	درس	مدرسه	53
3	مفعلو	False	درس	مدرسو	52
4	مفعلي	False	درس	مترسي	51
5	مفعول	False	دره	مدروة	50
6	مفاعل	False	دسه	مداسه	49
7	مفتعل	False	دسه	مدتسه	48
8	نفعله	False	درس	ندرسه	47
9	نفتعل	False	دسه	ندتسه	46
10	وفعلت	False	درس	ودرست	45
11	وفعله	False	درس	ودرسه	44
12	وفعال	False	دسه	وداسه	43

Şekil 4.3 Birinci örnek

Bu örnekte “okul” anlamına gelen "مدرسة" kelimesinin kökünün nasıl bulunacağı anlatılmaktadır. Kelime beş harf içerir, bu nedenle kökü bulunana kadar beş harf uzunluğundaki tüm kalıplar getirilerek kelimeyle karşılaştırılacaktır. Bu durumda, gördüğümüz gibi, 54. adımda " درس " kalıbını kullanana ve tespit edilen kök "مفعلة" olana kadar 53 kez başarılı olunamamıştır. Örneğin 49. adımda "مفاعل" kalıbı kullanılmış ancak algoritma kalıbı kontrol ettikten sonra kalıptaki üçüncü harfle kelimenin eşleşmediğini bulmuş ve bu nedenle bu kalıbı göz ardı etmiştir çünkü devam ederse yanlış kök olan "دسه" üretilcektir.

4.4.2 İkinci Örnek

Bu örnek, kelimenin herhangi bir nedenle kökü olup olmadığını göstermek içindir. Kullanılan kelime “Bilgisayar” anlamına gelen "كومبيوتر"dur. Bu kelime İngilizce kökenlidir ve günümüzde Arapçada da kullanılmaktadır. Aynı kelime uzunluğuna

sahip tüm kalıplar kontrol için getirilir, ancak bu kalıplardan hiçbiri Şekil 4.4'te görüldüğü gibi başarılı bir şekilde kök alamaz.

Arabic Word Root

Extract Arabic Roots

Word: كومبيوتر

Extract

Root: ---

Pattern: ---

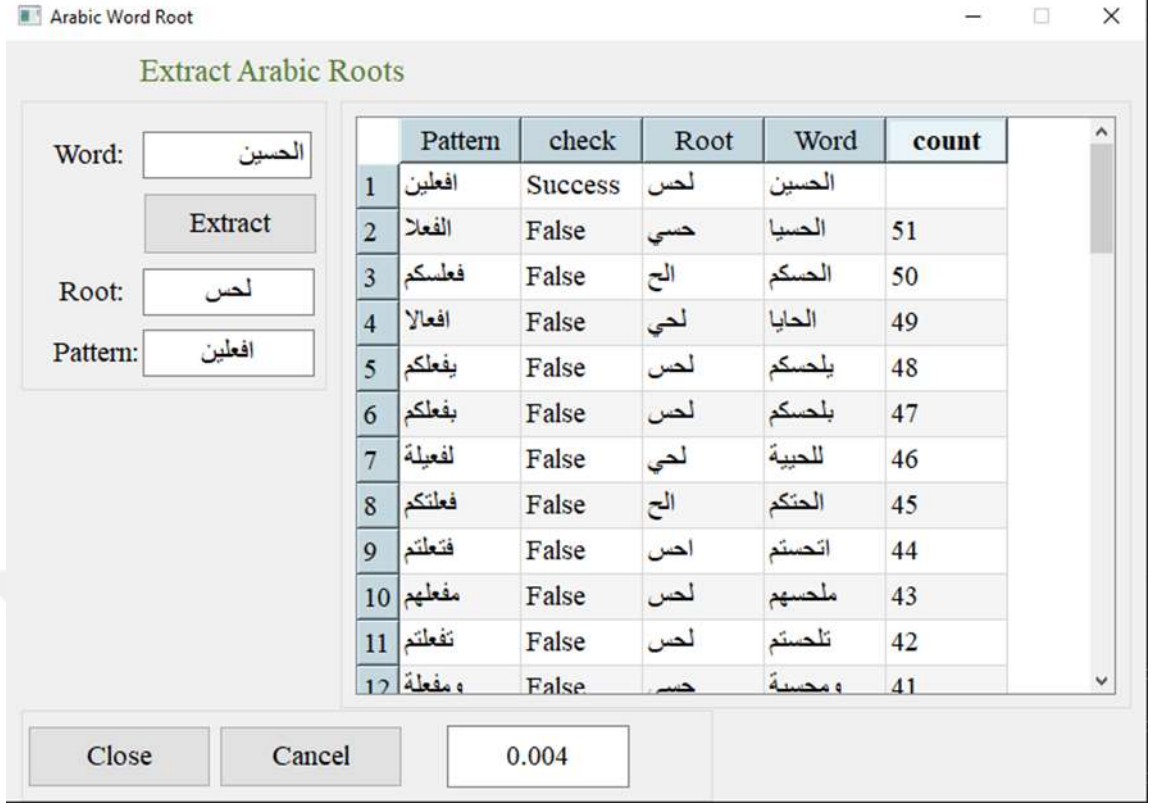
	Pattern	check	Root	Word	count
1	لافتعال	False	كبو	ولاكتبال	108
2	افتعالات	False	وبو	اوتباوات	107
3	استفعلوا	False	بيو	استبيروا	106
4	افتعالهم	False	وبو	اوتباوهم	105
5	الفعالة	False	مبي	المبيالة	104
6	الفعالي	False	مبي	المبياني	103
7	الفعالية	False	مبي	المبيلية	102
8	الفعليات	False	مبي	المبييات	101
9	الفعليون	False	مبي	المبييون	100
10	الفعليين	False	مبي	المبييين	99
11	الفعالات	False	مبو	المباوات	98
12	الفاعلات	False	مبه	المباهات	97

Close Cancel 0.005

Şekil 4.4 Örnek iki

4.4.3 Üçüncü Örnek

Bu durumda kelimenin kökü doğrudur ve kök belirlenmiştir ancak Şekil 4.5'te gösterildiği gibi hatalıdır.

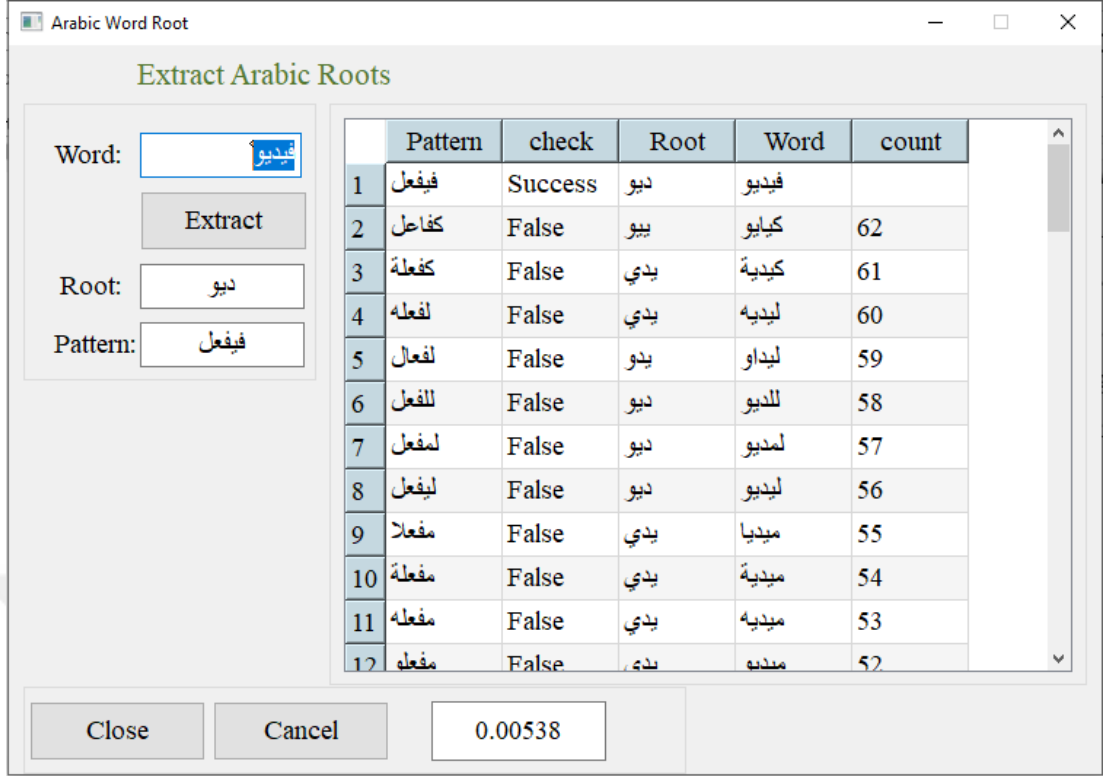


Şekil 4.5 Üçüncü örnek

"الحسين" kelimesinin bir "حسن" kökü vardır, ancak algoritma "افعلين" kalıbını kullanarak yanlış "لحس" kökü üretir; burada aslında doğru kalıp "الفعيل"dır. Bunun nedeni, sıralamadaki bu "افعلين" kalıbının doğru kalıptan önce gelerek kelimedeki karşılık gelen harfin aynı olması şartını yerine getirmesi ve üretilen kelimeyi orijinal kelimeyle karşılaştıran ikinci testte aynı kelimeyi üretmesidir. Bu sorunu çözmek için harf simetrisi koşulunu sağlayan tüm desenler seçilebilir ve amaca en yakın olanı hesaplanabilir.

4.4.4 Dördüncü Örnek

Algoritmayı kullanırken karşılaşılan son durum, kelimenin doğru köke sahip olmaması ancak programın yanlış kök üretebilmesi veya tespit edebilmesidir, Şekil 4.6 bu durumu göstermektedir.



Şekil 4.6 Dördüncü örnek

Şekil 4.6’da olduğu gibi, “video” anlamına gelen "فيديو" kelimesi orijinal Arapça bir kelimedir, ancak koşullar karşılandığı için uygulama yanlış kökü tespit etmiştir. Bu sorunu çözmek için sözlük kullanılabilir veya kurallar kullanılmalıdır.

5. SONUÇ VE ÖNERİLER

Doğal dil işleme, bilgisayarların insan dilini anlamasını, yaratmasını ve değiştirmesini sağlayan yapay zekanın bir alt kümesidir. Doğal dil işleme, verileri yerel dildeki metin veya konuşmaları kullanarak inceleyebilir. Doğal Dil İşleme destekli diğer çözümler arasında çevrimiçi arama, e-posta spam filtrelemesi, otomatik metin veya konuşma çevirisi, belge özetleme, duygu analizi ve dil bilgisi/yazım denetimi yer alır. Doğal dil işleme araştırmaları, dijital bilgisayarların piyasaya sürülmesinden hemen sonra başlamıştır ve yapay zekanın yanı sıra dil biliminden de yararlanmaktadır. Ancak son zamanlardaki başarılar, verilerden öğrenen ve genelleme yapan sistemler yaratan, bir tür yapay zekâ olan makine öğrenimi tarafından kazanılmıştır. Derin öğrenme, büyük veri kümelerinden son derece karmaşık kalıpları öğrenebilen ve web kaynaklı veri kümelerinden doğal dilin karmaşıklığını anlamak için onu uygun hale getiren bir tür makine öğrenimidir. Arapça, dünyanın en güçlü ve en yaygın konuşulan dillerinden biri olmasına rağmen çok sayıda örtüşen ve yabancı terime sahiptir. Arapçada kelime bir kök-kalıp bağlantısı olarak görülmektedir. Bir kök, kalıpla etkileşime girdiğinde çok sayıda kelime öbeği oluşabilir. Aynı kök, başka bir kalıpla etkileşime girdiğinde birden fazla kelime üretebilir. Kalıp ve kök arasındaki ilişki Arapça kelimelerin ana odağıdır. Bu çalışmada Arapça kelimelerden kök çıkarmak için morfolojik özelliklere ve kalıplara dayalı bir algoritma tanıtılmıştır. Üç unsur olan “kelime”, “kalıp” ve “kök” arasındaki ilişkiler araştırılmıştır. Kelimenin kökünü ayırmak zor bir problem olduğu için uygun kalıbın seçilmesi önemlidir. Öncelikle girdi kelimenin varsa kökü belirlenir. Daha sonra doğru kökü belirlemek için kullanılacak doğru kalıp seçilir. Bu, işlemi hızlandırır ve kelimenin başlangıç harflerinin kaldırılmasını önler çünkü ekleri kaldırmadan doğrudan köklere ulaşabiliriz. Kelimedeki harfler ve kalıptaki harfler, varlıkları, dizilişleri ve karşılaştırılmaları ile ilgili olarak, eklerin kaldırılması işlemi için gereken zamanın yanı sıra, bazen orijinal kelimenin temel harflerinden bazılarının silinmesiyle hatalara neden olabilir. Aynı uzunluktaki sözcükleri ve kalıpları karşılaştırdıktan sonra, ekleri eklemek veya çıkarmak yerine kalıptan ekler eklenir veya çıkarılır ve sonuç, yani kök değerlendirilir. Kök gerekli olduğundan ve orijinal kelime bulunursa doğru kabul edildiğinden, elde edilen kök daha sonra orijinal kelimeyi oluşturmak için ters yönde kullanılır. Uygulamanın Python’da inşa

edilmesinin birçok nedeni vardır; bunların en önemlisi, kolaylığı ve kaynakların kullanılabilirliğinin yanı sıra açık kaynaklı bir dil olması ve yapay zekâ uygulamalarında birinci dil olarak kabul edilmesidir. Python dilinin yanı sıra görsel programlama araçları sağlamak üzere tasarlanmış pyqt5 paketi kullanılmıştır. Uygulamayı test etmek için 115.801 benzersiz kelime içeren 2.472.763 farklı belgeden oluşan Arapça derlem kullanılmış ve doğal dil işleme algoritmalarını değerlendirmek için kullanılan kriterler uygulanmış olup doğruluk %92, geri çağırma %97, kesinlik ise %94 olarak hesaplanmıştır.

5.1 Öneriler

Doğal diller genellikle karmaşık olarak kabul edildiğinden ve çekimler içerdiğinden, bunları hesaplamalı olarak işlemek basit değildir. Arapça öğrenilmesi en zor dillerden biri olarak kabul edilir. Bu dil üzerine yapılan araştırmaların çoğunluğu dil kurallarının uygulanması, dil edinimi ve diğer yaklaşımlar dahil olmak üzere bir dizi tekniğe dayanmaktadır. Bu teknikler genellikle sözlüklerden, dilbilgisi kurallarından ve son eklerin kaldırılmasından yararlanır. Daha önce bahsedilenlerden bağımsız, farklı bir yaklaşım kullandığımızı göz önüne alarak şunları önermekteyiz; Arapça ile çalışırken karma yaklaşımlardan yararlanmak, bazı popüler sözlüklerden yararlanmak ve daha fazla kelime yapısını karşılamak için daha fazla kalıp eklemek.

KAYNAKLAR

- Abd Alameer, A. Q. & Kadhem, S. M. (2017). Finding the similarity between two Arabic texts. *Iraqi Journal of Science*, 58(1A), 152-162.
<https://ijs.uobaghdad.edu.iq/index.php/eijs/article/view/6236>
- Abdallah, S., Shaalan, K., & Shoaib, M. (2012, March). Integrating rule-based system with classification for arabic named entity recognition. In *International Conference on Intelligent Text Processing and Computational Linguistics* (pp. 311-322). Berlin, Heidelberg: Springer Berlin Heidelberg.
https://www.academia.edu/download/70522079/Integrating_Rule-Based_System_with_Class20210929-5742-28x9td.pdf
- Abdelkader, A., Najeeb, M., Alnamari, M., & Malik, H. (2019). How Existing NLP Tools of Arabic Language Can Serve Hadith Processing. *International Journal of Computer Engineering and Technology (IJCET)*, 10(6), 22-31.
<http://www.iaeme.com/IJCET/issues.asp?JType=IJCET&VType=10&IType=6>
- Abouenour, L., El Hassani, S., Yazidy, T., Bouzouba, K., & Hamdani, A. (2008). Building an Arabic morphological analyzer as part of an open Arabic NLP platform. In *Workshop on HLT and NLP within the Arabic world: Arabic Language and local languages processing Status Updates and Prospects At the 6th Language Resources and Evaluation Conference (LREC'08)*.
https://www.researchgate.net/publication/228730940_Building_an_Arabic_morphological_analyzer_as_part_of_an_open_Arabic_NLP_platform
- Ahmed, A., Hançerlioğulları, A., & Tosun, A. R. (2022). Utilizing roots and patterns to identify Arabic named entities. *Asian Journal of Mathematics and Computer Research*, 29(2), 33-42.
<https://doi.org/10.56557/ajomcor/2022/v29i27922>
- Ahmed, A., Hançerlioğulları, A., & Tosun, A. R. (2023). Using Roots and Patterns to Detect Arabic Verbs without Affixes Removal. *International journal of computer science and network security: IJCSNS*, 23(4), 1-6.
<https://doi.org/10.22937/ijcsns.2023.23.4.1>
- Al-Abweeny, W., & Zaid, N. A. (2018). Arabic stemmer system based on rules of roots. *International Journal of Information Technology*, 2(1), 19-26.
<https://journals.sfu.ca/ijitls/index.php/ijitls/article/download/17/pdf/0>
- Alasmari, J., Watson, J. C., & Atwell, E. (2017). A comparative analysis of verb tense and aspect in Arabic and English using Google Translate. *International Journal on Islamic Applications in Computer Science and Technology*,

- 5(3), 9-13.
https://www.researchgate.net/publication/325333623_A_comparative_analysis_of_verb_tense_and_aspect_in_Arabic_and_English_using_Google_Translate
- Alazzawie, A. (2020). The focus function and derivation of object displacement in Standard Arabic. *Heliyon*, 6(4).
<https://doi.org/10.1016/j.heliyon.2020.e03792>
- Alhaj, Y. A., Xiang, J., Zhao, D., Al-Qaness, M. A., Abd Elaziz, M., & Dahou, A. (2019). A study of the effects of stemming strategies on arabic document classification. *IEEE access*, 7, 32664-32671.
[10.1109/ACCESS.2019.2903331](https://doi.org/10.1109/ACCESS.2019.2903331)
- Alharbi, R., Magdy, W., Darwish, K., AbdelAli, A., & Mubarak, H. (2018, May). Part-of-speech tagging for Arabic Gulf dialect using Bi-LSTM. In *Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018)*. <https://aclanthology.org/L18-1620.pdf>
- Ali, B. A. B., Mihi, S., El Bazi, I., & Laachfoubi, N. (2020). A Recent Survey of Arabic Named Entity Recognition on Social Media. *Rev. d'Intelligence Artif.*, 34(2), 125-135. <https://doi.org/10.18280/ria.340202>
- Ali, M. N. A., Tan, G., & Hussain, A. (2019). Boosting Arabic named-entity recognition with multi-attention layer. *IEEE Access*, 7, 46575-46582.
[10.1109/ACCESS.2019.2909641](https://doi.org/10.1109/ACCESS.2019.2909641)
- Al-Kabi, M. N. (2013, December). Towards improving Khoja rule-based Arabic stemmer. In *2013 IEEE Jordan Conference on Applied Electrical Engineering and Computing Technologies (AEECT)* (pp. 1-6). IEEE.
[10.1109/AEECT.2013.6716437](https://doi.org/10.1109/AEECT.2013.6716437)
- Al-Kabi, M. N., Kazakzeh, S. A., Ata, B. M. A., Al-Rababah, S. A., & Alsmadi, I. M. (2015). A novel root based Arabic stemmer. *Journal of King Saud University-Computer and Information Sciences*, 27(2), 94-103.
<https://doi.org/10.1016/j.jksuci.2014.04.001>
- Al-Khateeb, A., Zerrouki, T., Shquier, M. M. A., Balla, A., & Al-Khateeb, A. (2021). A New Enhanced Arabic Light Stemmer for IR in Medical Documents. *Computers, Materials & Continua*, 68(1). [10.32604/cmc.2021.016155](https://doi.org/10.32604/cmc.2021.016155)
- Alkhatib, M., Monem, A. A., & Shaalan, K. (2020). Deep learning for Arabic error detection and correction. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, 19(5), 1-13.
<https://doi.org/10.1145/3373266>
- Almarimi, A. A., & Enbiah, E. M. (2020). Recognition System for Libyan Entity Names. *European Journal of Electrical Engineering and Computer Science*, 4(6). [10.24018/ejece.2020.4.6.263](https://doi.org/10.24018/ejece.2020.4.6.263)

- Alnaied, A., Elbendak, M., & Bulbul, A. (2020). An intelligent use of stemmer and morphology analysis for Arabic information retrieval. *Egyptian Informatics Journal*, 21(4), 209-217. <https://doi.org/10.1016/j.eij.2020.02.004>
- Al-Nashashibi, M. Y., Neagu, D., & Yaghi, A. A. (2010, November). An improved root extraction technique for Arabic words. In *2010 2nd International Conference on Computer Technology and Development* (pp. 264-269). IEEE. [10.1109/ICCTD.2010.5645872](https://doi.org/10.1109/ICCTD.2010.5645872)
- Allothman, A., & Als Salman, A. (2020). Arabic Morphological Analysis Techniques. *International Journal of Advanced Computer Science and Applications*, 11(2). <https://api.semanticscholar.org/CorpusID:212656007>
- Alsaad, A., & Abbod, M. (2014, March). Arabic text root extraction via morphological analysis and linguistic constraints. In *2014 UKSim-AMSS 16th International Conference on Computer Modelling and Simulation* (pp. 125-130). IEEE. [10.1109/UKSim.2014.43](https://doi.org/10.1109/UKSim.2014.43)
- Alsayat, A., & Elmitwally, N. (2020). A comprehensive study for Arabic sentiment analysis (challenges and applications). *Egyptian Informatics Journal*, 21(1), 7-12. <https://doi.org/10.1016/j.eij.2019.06.001>
- Al-Serhan, H., & Ayesh, A. (2006, November). A trilateral word roots extraction using neural network for Arabic. In *2006 International Conference on Computer Engineering and Systems* (pp. 436-440). IEEE. [10.1109/ICCES.2006.320487](https://doi.org/10.1109/ICCES.2006.320487)
- Alshalabi, H., Tiun, S., Omar, N., & Albared, M. (2013). Experiments on the use of feature selection and machine learning methods in automatic malay text categorization. *Procedia Technology*, 11, 748-754. <https://doi.org/10.1016/j.protcy.2013.12.254>
- Alshalabi, H., Tiun, S., Omar, N., Abdulwahab Anaam, E., & Saif, Y. (2022). BPR algorithm: New broken plural rules for an Arabic stemmer. *Egyptian Informatics Journal*, 23(3), 363-371. <https://doi.org/10.1016/j.eij.2022.02.006>
- Alshalabi, H., Tiun, S., Omar, N., Al-Aswadi, F. N., & Alezabi, K. A. (2022). Arabic light-based stemmer using new rules. *Journal of King Saud University-Computer and Information Sciences*, 34(9), 6635-6642. <https://doi.org/10.1016/j.jksuci.2021.08.017>
- Al-Sughaiyer, I. A., & Al-Kharashi, I. A.. (2004). Arabic morphological analysis Techniques: A comprehensive survey. *Journal of the American society for information science and technology*, 55(3), 189-213. <https://doi.org/10.1002/asi.10368>
- Alwaneen, T. H., Azmi, A. M., Aboalsamh, H. A., Cambria, E., & Hussain, A. (2022). Arabic question answering system: a survey. *Artificial Intelligence*

- Review, 55(1), 207-253. [https://doi.org/10.1007/s10462-021-10031-1\(0123456789\(\),-volV\)\(0123456789\(\),-volV\)](https://doi.org/10.1007/s10462-021-10031-1(0123456789(),-volV)(0123456789(),-volV))
- Ameur, M. S. H., Meziane, F., & Guessoum, A. (2020). Arabic machine translation: A survey of the latest trends and challenges. *Computer Science Review*, 38, 100305. <https://doi.org/10.1016/j.cosrev.2020.100305>
- Aqel, A., Alwadei, S., & Dahab, M. (2015). Building an Arabic words generator. *International Journal of Computer Applications*, 112(14). [10.5120/19738-1540](https://doi.org/10.5120/19738-1540)
- Attia, M., Pecina, P., Toral, A., Tounsi, L., & Van Genabith, J. (2011, July). An open-source finite state morphological transducer for modern standard Arabic. In *Proceedings of the 9th International Workshop on Finite State Methods and Natural Language Processing* (pp. 125-133). <https://aclanthology.org/W11-4417.pdf>
- Azman, B. (2019). Root identification tool for Arabic verbs. *IEEE Access*, 7, 45866-45871. [10.1109/ACCESS.2019.2908177](https://doi.org/10.1109/ACCESS.2019.2908177)
- Azmi, A. M., & Aljafari, E. A. (2018). Universal web accessibility and the challenge to integrate informal Arabic users: a case study. *Universal Access in the Information Society*, 17, 131-145. <https://link.springer.com/article/10.1007/s10209-017-0522-3>
- Azmi, A. M., & Alshenaifi, N. A. (2017). Lemaza: An Arabic why-question answering system. *Natural Language Engineering*, 23(6), 877-903. <https://doi.org/10.1017/S1351324917000304>
- Ben Ismail, S., Boukedi, S., & Haddar, K. (2018). Generation of Arabic broken plural within LKB. In *Text, Speech, and Dialogue: 21st International Conference, TSD 2018, Brno, Czech Republic, September 11-14, 2018, Proceedings 21* (pp. 205-212). Springer International Publishing. https://link.springer.com/chapter/10.1007/978-3-030-00794-2_22
- Boubas, A., Lulu, L., Belkhouche, B., & Harous, S. (2011, April). GENESTEM: A novel approach for an Arabic stemmer using genetic algorithms. In *2011 International Conference on Innovations in Information Technology* (pp. 77-82). IEEE. [10.1109/INNOVATIONS.2011.5893872](https://doi.org/10.1109/INNOVATIONS.2011.5893872)
- Boudchiche, M., & Mazroui, A. (2018, April). Improving the Arabic root extraction by using the quadratic splines. In *2018 International Conference on Intelligent Systems and Computer Vision (ISCV)* (pp. 1-5). IEEE. [10.1109/ISACV.2018.8354062](https://doi.org/10.1109/ISACV.2018.8354062)
- Cheikh, I. B., & Laffet, A. (2017, November). New Morphological Markovian Approach for Analysis and Recognition of Open Arabic Canonical Vocabulary. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)* (Vol. 1, pp. 183-188). IEEE. [10.1109/ICDAR.2017.38](https://doi.org/10.1109/ICDAR.2017.38)

- Chennoufi, A., & Mazroui, A. (2017). Morphological, syntactic and diacritics rules for automatic diacritization of Arabic sentences. *Journal of King Saud University-Computer and Information Sciences*, 29(2), 156-163. <https://doi.org/10.1016/j.jksuci.2016.06.004>
- Chiu, J. P., & Nichols, E. (2016). Named entity recognition with bidirectional LSTM-CNNs. *Transactions of the association for computational linguistics*, 4, 357-370. https://doi.org/10.1162/tacl_a_00104
- Dawdy-Hesterberg, L. G., & Pierrehumbert, J. B. (2014). Learnability and generalisation of Arabic broken plural nouns. *Language, cognition and neuroscience*, 29(10), 1268-1282. <https://doi.org/10.1080/23273798.2014.899377>
- El Bazzi, M. S., Zaki, T., Mammass, D., & Ennaji, A. (2016, October). Stemming versus multi-words indexing for Arabic documents classification. In *2016 11th International Conference on Intelligent Systems: Theories and Applications (SITA)* (pp. 1-5). IEEE. [10.1109/SITA.2016.7772288](https://doi.org/10.1109/SITA.2016.7772288)
- El Hadj, Y., Al-Sughayeir, I., & Al-Ansari, A. (2009, April). Arabic part-of-speech tagging using the sentence structure. In *Proceedings of the Second International Conference on Arabic Language Resources and Tools, Cairo, Egypt* (pp. 241-45). <https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=ea2246d6775145faf60ad950503c33c878f961e8>
- El Ouardi, L. (2022). The Moroccan Arabic verb and noun test for language mapping (MAVNT-LP) under nTMS and DES. *Applied Neuropsychology: Adult*, 29(6), 1413-1424. <https://doi.org/10.1080/23279095.2021.1883020>
- El-Defrawy, M., Belal, N. A., & El-Sonbaty, Y. (2017, September). An efficient rank based Arabic root extractor. In *2017 Intelligent Systems Conference (IntelliSys)* (pp. 870-878). IEEE. [10.1109/IntelliSys.2017.8324232](https://doi.org/10.1109/IntelliSys.2017.8324232)
- Ellouze, S., Haddar, K., & Abdelwahed, A. (2010). NooJ disambiguation local grammars for Arabic broken plurals. In *Proceedings of the NooJ 2010 International Conference* (pp. 62-72). https://gavriilidou.gr/wp-content/uploads/2021/04/Proceedings_of_the_Nooj_2010.pdf#page=64
- Elmadany, A. A., Abdou, S. M., & Gheith, M. (2015). A survey of Arabic dialogues understanding for spontaneous dialogues and instant message. *arXiv preprint arXiv:1505.03084*. <https://doi.org/10.48550/arXiv.1505.03084>
- Elmenshawy, K., & El-Nady, A. (2019). Practical English to Arabic Machine Translation System by Applying Rules-Based Method. *Int. J. Eng. Sci. Comput*, 9(1), 19556-19560.

- Etaiwi, W., & Awajan, A. (2018). Graph-based Arabic NLP techniques: a survey. *Procedia computer science*, 142, 328-333. <https://doi.org/10.1016/j.procs.2018.10.488>
- Farghaly, A., & Shaalan, K. (2009). Arabic natural language processing: Challenges and solutions. *ACM Transactions on Asian Language Information Processing (TALIP)*, 8(4), 1-22. <https://doi.org/10.1145/1644879.1644881>
- Goweder, A., Poesio, M., De Roeck, A., & Reynolds, J. (2004, July). Identifying broken plurals in unvowelised arabic tex. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing* (pp. 246-253). <https://aclanthology.org/W04-3232.pdf>
- Guellil, I., Saâdane, H., Azouaou, F., Gueni, B., & Nouvel, D. (2021). Arabic natural language processing: An overview. *Journal of King Saud University-Computer and Information Sciences*, 33(5), 497-507. <https://doi.org/10.1016/j.jksuci.2019.02.006>
- Habash, N. (2007). Arabic morphological representations for machine translation. In *Arabic computational morphology: Knowledge-based and empirical methods* (pp. 263-285). Dordrecht: Springer Netherlands. https://link.springer.com/chapter/10.1007/978-1-4020-6046-5_14
- Hanani, A., & Naser, R. (2020). Spoken Arabic dialect recognition using X-vectors. *Natural Language Engineering*, 26(6), 691-700. <https://doi.org/10.1017/S1351324920000091>
- Hassan, A. J. (2015). Translating Arabic verb repetition into English. *Arab World English Journal (AWEJ)*, 6(2), 144-153. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2834344
- Hassan, M. M., Al-Nasrawi, D. A., Hassan, R. J., & Mahdi, N. T. (2018). Rule based method of name entity recognition for matching Allah's finest names in Holy Quran. *Journal of Engineering and Applied Sciences*, 13(10), 3618-3623. https://www.researchgate.net/publication/326544229_Rule_based_method_of_name_entity_recognition_for_matching_Allah's_fine_names_in_Holy_Quran
- Hegazi, M. O. (2016). An approach for Arabic root generating and lexicon development. *IJCSNS Int. J. Comp. Sci. Network Security*, 16(1).
- Hegazi, M. O., Al-Dossari, Y., Al-Yahy, A., Al-Sumari, A., & Hilal, A. (2021). Preprocessing Arabic text on social media. *Heliyon*, 7(2). <https://doi.org/10.1016/j.heliyon.2021.e06191>
- Ismail, S. B., Boukédi, S., & Haddar, K. (2017, May). Transformation system to generate derivational forms of an Arabic verb with HPSG. In *2017 International Conference on Engineering & MIS (ICEMIS)* (pp. 1-5). IEEE. [10.1109/ICEMIS.2017.8272994](https://doi.org/10.1109/ICEMIS.2017.8272994)

- Issa, E. (2018, August). An OpenNMT Model to Arabic Broken Plurals. In *Proceedings of the First International Workshop on Language Cognition and Computational Models* (pp. 22-30). <https://aclanthology.org/W18-4103>
- Jaafar, Y., Namly, D., Bouzoubaa, K., & Yousfi, A. (2017). Enhancing Arabic stemming process using resources and benchmarking tools. *Journal of King Saud University-Computer and Information Sciences*, 29(2), 164-170. <https://doi.org/10.1016/j.jksuci.2016.11.010>
- Kambhatla, N., & Zitouni, I. (2013). U.S. Patent No. 8,527,262. Washington, DC: U.S. Patent and Trademark Office.
- Kanan, T., Sadaqa, O., Aldajeh, A., Alshwabka, H., AlZu'bi, S., Elbes, M., ... & Alia, M. A. (2019, April). A review of natural language processing and machine learning tools used to analyze arabic social media. In *2019 IEEE Jordan International Joint Conference on Electrical Engineering and Information Technology (JEEIT)* (pp. 622-628). IEEE. [10.1109/JEEIT.2019.8717369](https://doi.org/10.1109/JEEIT.2019.8717369)
- Mansouri, A., Affendey, L. S., & Mamat, A. (2008). Named entity recognition using a new fuzzy support vector machine. *IJCSNS*, 8(2), 320. <https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=73e3a8af095db72f164cb6a3b2cb37733e501a4c>
- Mohamed, E., & Kübler, S. (2010, May). Arabic Part of Speech Tagging. In *LREC*. http://lrec.elra.info/proceedings/lrec2010/pdf/384_Paper.pdf
- Muhammed, M. H., Salih, B. M., & Jasim, O. K. (2018, November). An emerging standard miniaturization in Arabic morphological analysis. In *2018 1st Annual International Conference on Information and Sciences (AiCIS)* (pp. 112-116). IEEE. [10.1109/AiCIS.2018.00032](https://doi.org/10.1109/AiCIS.2018.00032)
- Mustafa, M., Aldeen, A. S., Zidan, M. E., Ahmed, R. E., & Eltigani, Y. (2019). Developing two different novel techniques for Arabic text stemming. *Intelligent Information Management*, 11(01), 1. [10.4236/iim.2019.111001](https://doi.org/10.4236/iim.2019.111001)
- Naili, M., Chaibi, A. H., & Ghezala, H. H. B. (2019). Comparative study of Arabic stemming algorithms for topic identification. *Procedia Computer Science*, 159, 794-802. <https://doi.org/10.1016/j.procs.2019.09.238>
- Obeid, O., Zalmout, N., Khalifa, S., Taji, D., Oudah, M., Alhafni, B., ... & Habash, N. (2020, May). CAMEL tools: An open source python toolkit for Arabic natural language processing. In *Proceedings of the Twelfth language resources and evaluation conference* (pp. 7022-7032). <https://aclanthology.org/2020.lrec-1.868>
- Othman, M. T. B., Al-Hagery, M. A., & El Hashemi, Y. M. (2020). Arabic text processing model: Verbs roots and conjugation automation. *IEEE Access*, 8, 103913-103923. [10.1109/ACCESS.2020.2999259](https://doi.org/10.1109/ACCESS.2020.2999259)

- Oueslati, O., Cambria, E., HajHmida, M. B., & Ounelli, H. (2020). A review of sentiment analysis research in Arabic language. *Future Generation Computer Systems*, 112, 408-430. <https://doi.org/10.1016/j.future.2020.05.034>
- Ray, S. K., & Shaalan, K. (2016). A review and future perspectives of arabic question answering systems. *IEEE Transactions on Knowledge and Data Engineering*, 28(12), 3169-3190. [10.1109/TKDE.2016.2607201](https://doi.org/10.1109/TKDE.2016.2607201)
- Salloum, W., & Habash, N. (2022). Unsupervised Arabic dialect segmentation for machine translation. *Natural Language Engineering*, 28(2), 223-248. <https://doi.org/10.1017/S1351324920000455>
- Sawalha, M., Atwell, E., & Abushariah, M. A. (2013, February). SALMA: standard Arabic language morphological analysis. In *2013 1st International Conference on Communications, Signal Processing, and their Applications (ICCSPA)* (pp. 1-6). IEEE. [10.1109/ICCSPA.2013.6487311](https://doi.org/10.1109/ICCSPA.2013.6487311)
- Sghaier, M. A., & Zrigui, M. (2020). Rule-based machine translation from Tunisian dialect to modern standard Arabic. *Procedia Computer Science*, 176, 310-319. <https://doi.org/10.1016/j.procs.2020.08.033>
- Shaalan, K. (2014). A survey of arabic named entity recognition and classification. *Computational Linguistics*, 40(2), 469-510. https://doi.org/10.1162/COLI_a_00178
- Shaalan, K., Magdy, M., & Fahmy, A. (2015). Analysis and feedback of erroneous Arabic verbs. *Natural Language Engineering*, 21(2), 271-323. <https://doi.org/10.1017/S1351324913000223>
- Shaalan, K., Siddiqui, S., Alkhatib, M., & Abdel Monem, A. (2019). Challenges in Arabic natural language processing. In *Computational linguistics, speech and image processing for arabic language* (pp. 59-83). https://doi.org/10.1142/9789813229396_0003
- Shafah, A., Goweder, A., Eshafah, S., & Rgibi, A. (2016, December). Irregular Arabic plurals recognition without stemming. In *2016 4th International Conference on Control Engineering & Information Technology (CEIT)* (pp. 1-6). IEEE. [10.1109/CEIT.2016.7929052](https://doi.org/10.1109/CEIT.2016.7929052)
- Shrestha, B. B., & Bal, B. K. (2020, December). Named-entity based sentiment analysis of Nepali news media texts. In *Proceedings of the 6th workshop on natural language processing techniques for educational applications* (pp. 114-120). <https://aclanthology.org/2020.nlptea-1.16>
- Siswoyo, S. (2016). Similarities and differences between English and Arabic verb. *Jurnal Smart*, 2(2). <http://ejournal.stkipmpringsewu-lpg.ac.id/index.php/smart>

- Souri, A., El Maazouzi, Z., Al Achhab, M., & El Mohajir, B. E. (2018). Arabic text generation using recurrent neural networks. In *Big Data, Cloud and Applications: Third International Conference, BDCA 2018, Kenitra, Morocco, April 4–5, 2018, Revised Selected Papers 3* (pp. 523-533). Springer International Publishing. https://link.springer.com/chapter/10.1007/978-3-319-96292-4_41
- Tahat, A. (2020). Morphological patterns of personal naming practice: a case study of non-concatenative Arabic language. *International Journal of Academic Research in Business and Social Sciences*, 10(6), 572-585. <http://dx.doi.org/10.6007/IJARBS/v10-i6/7333>
- Taylor, S., & Brychcín, T. (2018, April). Arabic word analogies and semantics of simple phrases. In *2018 2nd International Conference on Natural Language and Speech Processing (ICNLSP)* (pp. 1-6). IEEE. [10.1109/ICNLSP.2018.8374386](https://doi.org/10.1109/ICNLSP.2018.8374386)
- Thalji, N. J. (2024). A Survey of Arabic Root Extraction Algorithms. *Iraqi Journal of Science*. <https://doi.org/10.24996/ijs.2024.65.4.43>
- Thalji, N., Hanin, N. A., Hani, W. B., Al-Hakeem, S., & Thalji, Z. (2018). A novel rule-based root extraction algorithm for Arabic language. *International Journal of Advanced Computer Science and Applications*, 9(10). <https://pdfs.semanticscholar.org/2d28/d7139b6f45ac793e8b5436b24d61646a61ae.pdf>
- URL-1 (2020). <http://sites.google.com/site/mouradabbas9/corpora> Erişim tarihi: 05/05/2020
- Zakria, G., Farouk, M., Fathy, K., & Makar, M. N. (2019). Relation extraction from arabic wikipedia. *Indian Journal of Science and Technology*, 12(46), 01-06. <https://pdfs.semanticscholar.org/f3f1/906565b394d9493c1bb08420d773429e8c52.pdf>
- Zayed, O. H., & El-Beltagy, S. R. (2012, May). Person name extraction from modern standard Arabic or colloquial text. In *2012 8th International Conference on Informatics and Systems (INFOS)* (pp. NLP-44). IEEE. <https://www.infona.pl/resource/bwmeta1.element.ieee-art-000006236607>
- Zayed, O., & El-Beltagy, S. R. (2015, September). Named entity recognition of persons' names in Arabic tweets. In *Proceedings of the international conference recent advances in natural language processing* (pp. 731-738). <https://aclanthology.org/R15-1093.pdf>
- Zerrouki, T. (2024). Tashaphyne: A Python package for Arabic Light Stemming. *Journal of Open Source Software*, 9(93), 6063. <https://doi.org/10.21105/joss.06063>

Zirikly, A., & Diab, M. (2015, June). Named entity recognition for arabic social media. In *Proceedings of the 1st workshop on vector space modeling for natural language processing* (pp. 176-185).
<https://aclanthology.org/W15-1524.pdf>





EKLER

EK A. Terminoloji

Derlem: Dilbilimsel çalışmalarda kullanılan dil kullanımının yazılı ve ara sıra konuşulan örnekleri.

Sözlük: Bir dildeki kelimeleri sıralayan ve bunların tanımlarını ya da başka bir dile tercümelerini sağlayan bir kitap.

Lügat: Bir kişinin, bir dilin veya bir bilgi alanının sözcük dağarcığı.

Yüzey Formu: Yazılı bir kelimenin metinde görüldüğü şeklidir.

Başsözcük: Sözlük tanımı olan ifade. Bir kelimenin kurallı biçimi olarak tanımlanabilir.

Gövde: Bir kelimenin tüm ekler çıkarılmış hali.

Kök: Kelimenin sözsel kaynağıdır. Yalnızca orijinal harflerden oluşur ve bunlar daha sonra türetme ve çekim yoluyla çok sayıda başka kelime oluşturmak için kullanılır.

Kök Çıkarma: Kökün elde edilmesi. Yani tüm ekler çıkarıldıktan sonra elde edilen kelime.

Az Kök Çıkarma: Bir diğer adı da gövde tabanlı kök çıkarmadır. Kelime kökünün en temel çekim şekliyle elde edilmesi işlemidir.

Çok Kök Çıkarma: Kök tabanlı kök çıkarmadır. Kelimenin asıl kökünü tespit etme işlemidir. Bu kök, kelimeye yüzey formunu vermek için türetilen ve çekimlenen fiildir.

Morfolojik Analiz: Girdi metninin yüzey formları alt kategoriye (sayı, cinsiyet vb.) ve cümle öğelerine (isim, fiil vb.) göre kategorize edilir. Bu noktada başsözcük, kök ve eğer kelime türetilmişse gövde, her yüzey formu için olası tüm “analizler” ile birlikte çıktı olarak sunulur.