T.C.

İSTANBUL AYDIN ÜNİVERSİTESİ

FEN BİLİMLERİ ENSTİTÜSÜ

DOĞAL DİL İŞLEME TEKNİKLERİYLE YAZAR-KİTAP TANIMA

YÜKSEK LİSANS TEZİ

SAMET KAYA

BİLGİSAYAR MÜHENDİSLİĞİ ANA BİLİM DALI

BİLGİSAYAR MÜHENDİSLİĞİ PROGRAMI

HAZİRAN 2018

T.C.

İSTANBUL AYDIN ÜNİVERSİTESİ

FEN BİLİMLERİ ENSTİTÜSÜ

DOĞAL DİL İŞLEME TEKNİKLERİYLE YAZAR-KİTAP TANIMA

YÜKSEK LİSANS TEZİ

Samet KAYA

(Y1413010034)

Bilgisayar Mühendisliği Ana Bilim Dalı

Bilgisayar Mühendisliği Programı

Tez Danışmanı: Prof. Dr. Ali GÜNEŞ

HAZİRAN 2018

T.C.
İSTANBUL AYDIN ÜNİVERSİTESİ
FEN BİLİMLER ENSTİTÜSÜ MÜDÜRLÜĞÜ

Yüksek Lisans Tez Onay Belgesi

Enstitümüz Bilgisayar Mühendisliği Ana Bilim Dalı Bilgisayar Mühendisliği Tezli Yüksek Lisans Programı **Y1413.010034** numaralı öğrencisi **Samet KAYA** ' nın **"DOĞAL DİL İŞLEME TEKNİKLERİYLE YAZAR-KİTAP TANIMA"** adlı tez çalışması Enstitümüz Yönetim Kurulunun 23.05.2018 tarih ve 2018/09 sayılı kararıyla oluşturulan jüri tarafından ...ejbirliği.. ile Tezli Yüksek Lisans tezi olarak ...kabul.....edilmiştir.

Öğretim Üyesi Adı Soyadı	İmzası
Tez Savunma Tarihi : 19/06/2018	
1)Tez Danışmanı: Prof. Dr. Ali GÜNEŞ	.....
2) Jüri Üyesi : Doç. Dr. Metin ZONTUL	.....
3) Jüri Üyesi : Dr. Öğr. Üyesi Ferdi SÖNMEZ	.....

Not: Öğrencinin Tez savunmasında **Başarılı** olması halinde bu form **imzalanacaktır**. Aksi halde geçersizdir.

YEMİN METNİ

Yüksek Lisans / Doktora tezi olarak sunduğum "Doğal Dil İşleme Teknikleriyle Yazar-Kitap Tanıma" adlı çalışmanın, tezin proje safhasından sonuçlanmasına kadarki bütün süreçlerde bilimsel ahlak ve geleneklere aykırı düşecek bir yardıma başvurulmaksızın yazıldığını ve yararlandığım eserlerin Bibliyografya'da gösterilenlerden oluştuğunu, bunlara atıf yapılarak yararlanılmış olduğunu belirtir ve onurumla beyan ederim. (08/05/2018)

Samet KAYA

Aileme,

ÖNSÖZ

Bu tezde Türkçe kitaplarda yazar tanıma işlemi gerçekleştirilmiştir. Bu çalışmayı yaparken FATİH SULTAN MEHMET VAKIF ÜNİVERSİTESİ ailesine sağladığı destekten dolayı teşekkür ederim. Aynı camiada çalışmakta olduğum ve benden bilgi ve tecrübelerini esirgemeyen Prof. Dr. A. Yılmaz ÇAMURCU, Yrd. Doç. Dr. Ayla GÜLCÜ, Yrd. Doç. Dr. Ali NİZAM, Doç. Dr Sadullah ÖZTÜRK, Yrd. Doç. Dr. Süha TUNA hocalarıma saygılarımı sunarım. Ailem ve hayatımın birçok kademesinde yanımda olan arkadaşlarım Öğr. Gör. Musa AYDIN'a, Öğr. Gör. Enes ÇELİK'e ve Ümit DEMİRBAĞA'ya da maddi ve manevi desteklerinden ötürü teşekkürlerimi borç bilirim. Tüm bunlara ek olarak tez çalışmamda bana yol gösteren değerli danışman hocam Prof. Dr. Ali GÜNEŞ'e minnetlerimi iletirim.

Haziran 2018 Samet KAYA

İÇİNDEKİLER

Sayfa

KISALTMALAR

DDİ :	Doğal Dil İşleme
D :	Doküman Seti
d :	Doküman
T :	Doküman Koleksiyon Sözlüğü
T :	Doküman Koleksiyon Sözlüğünde bir sözlük
PR :	Profil Fonksiyonu
Me :	Eğitim Metni
Mt :	Test Metni
f :	frekans

ÇİZELGE LİSTESİ

ŞEKİL LİSTESİ

DOĞAL DİL İŞLEME TEKNİKLERİYLE YAZAR-KİTAP TANIMA

ÖZET

İnsanlık yazının bulunmasından bu yana farklı yollarla birçok yazılı doküman üretmiştir. Yazılmış olan her yazı onu üreten yazarının izlerini taşımaktadır. Yazarın kelime hazinesi, düşünüş biçimi, mantık çıkarımları hatalı ya da eksil bilgileri, yazım alışkanlıkları metne yansımaktadır.

Bu bakış açısıyla, yazılan her dokümanın yazarın metinsel parmak izi olduğunu söyleyebiliriz. Ancak gerçek parmak izinde olduğu gibi izde bulunan yazara ait olan özellikleri çıkarmak insan yeteneğini aşmaktadır. Metin üzerimdeki kişisel karakteristiği çıkarmak bilgisayar devriminden önce oldukça zor bir görevdi bunun yanında bilgisayarlar bu işlemi yapabilmektedir. Yazar tanıma işlemi için, çeşitli yazar özellikleri yazara ait eğitim metinlerinden tespit edilmekte ve daha sonra sisteme sokulan başka bir metinin öndeki eğitimden çıkarılmış karakteristik vektörüyle ile benzerliği hesaplanmaktadır. Metin üzerindeki yazar özelliklerinden bazıları: kelime hazinesi, yazım hataları, karakter ve kelime n-gram izleri vs. Bilgisayarlar sayesinde bu tip özellikleri metinin içerisinden çıkarabiliyor ve bir dokümanın yazara aitliğini tespit edebiliyoruz.

Bu tezde, yazar tanıma işlemi yapılmıştır. 20 Türk yazarın farklı dağılımlarda yazmış olduğu 120 farklı Türkçe kitap üzerinde çalışılmıştır. Karakter n-gram yazarın stilometri özelliği olarak kullanışmış ve Naive Bayes sınıflayıcı metodu ile de sınıflama işlemi yapılmıştır. Tez kapsamında ilk önce, 120 Türkçe kitap bulunmuş ve txt formatına dönüştürülmüştür. Ardından, tüm kitaplar bir ön işleme sokularak boşluklar, karakter hataları, sayısal ve alfabetik olmayan ifadeler, noktalamalar, Türkçe olmayan karakterler yazıdan çıkarılmıştır. Ön işlemeden sonra, 120 kitap rasgele 20 yazar için 20 eğitim kitabı ve 100 test kitabı olarak iki farklı gruba bölünmüştür. Eğitim kitaplarında yazar etiketi bulunmaktadır. Yazar özelliği olarak

bi-gram, tri-gram, quadri-gram özellikleri eğitim kitaplarından frekansı hesaplanarak çıkarılmış ve en sık 200 tanesi yazarın stilometrik vektör uzayı oluşturulmuştur. Bu noktada sistemimiz yazar tanıma işlemi için hazır durumdadır. Sistemimizi test etmek için, her bir test kitabını yazar etiketsiz olarak tek tek sisteme soktuk. Her bir test kitabı da tıpkı eğitim kitabı gibi bi-gram, tri-gram, quadri-gram özellikleri çıkarılarak en sık 200 tanesi yazar özelliği olarak aldık. Sonunda sistemde bulunan yazar özellikleriyle her hangi bir test kitabından çıkardığımız vektörü naive bayes sınıflandırıcı ile sınıflandırma sonuçlarını aldık. Test kitabının gerçekte olan yazarı ile sistemin tahmin ettiği yazar ismini karşılaştırarak sistemimizin başarısını ölçtük ve kaydettik.

Tez çalışmasında farklı n-gram performansları Naive Bayes sınıflayıcı üzerinde performansları karşılaştırılmıştır. N-gram vektör uzaylarının yazar tanıma başarımları ölçülmüştür. Gözlemlerin sonucu olarak bi-gram vektör uzayı başarısız olmuştur. Bunun yanında tri-gram ve quadri-gram iyi sonuçlar vermiştir. En iyi performansı %82 başarım ile quadri-gram vermiştir. Tez sonunda tüm sonuçlar, karmaşıklık matrisi verilmiştir.

İnternet çağıyla birlikte explonansiyel artmış olan elektronik dokümanların plagarizim, adli araştırma gibi yönlerden incelenebilmesi için tez konusu önemlidir. Alanda birçok İngilizce çalışma bulunmasına rağmen Türkçe çalışma oldukça azdır. Bilgisayar çağında, bilgisayarların insan dilini anlaması ve üretmesi üzerine çalışmalar yürütülmektedir. Türkçe'nin de diğer dillerin gerisinde kalmaması için bu tip çalışma önem arz etmektedir. Bu bakımdan tez Türkçe doğal dil işlemeye katkıda bulunmuştur.

Anahtar Kelimeler : Metin sınıflama, Yazar tanıma, Naive bayes sınıflama, N-gram

AUTHOR-BOOK RECOGNITION WITH NATURAL LANGUAGE PROCESSING TECHNIQUES

ABSTRACT

Since the discovery of the manuscript, humanity has produced many written document in various ways. Every text carries traces of its author. There are author's thesaurus, thinking and logic execution, wrong or incomplete informations, spelling habits on the text.

From this point of view, we can say that a written document is the textual fingerprints of the people who write it. However, just like fingerprints, these features are difficult to detect from the text with human abilities. It is difficult to determine the personal characteristics of texts before the computer revolution, but these processes can be achived with computers today. For the author recognition process, various author features are determined by training text and then compared with the features of other texts to look for similarities. Some of the authors features on the text: thesaurus, typographical errors, character n-gram traces, word n-gram traces. Thanks to the computers, these types of properties are extracted from the text and analysis of the status of the writer's ownership of a document.

In this thesis, author detection process was studied. 120 different Turkish books written by 20 Turkish authors in different distributions were studied. Character n-gram for stylometry and Naïve Bayes classifier is used to recognize the author of the text. First of all, we gather 120 Turkish novels and convert them to txt format. Then, all book were pre-processed and gaps deleted, erroneous characters corrected, alphanumeric characters, punctuation, and non-Turkish characters removed. After preliminary operations, the books were divided into two group as 20 training and 100 test book. The training books have the author name label and author stylometric features

extracted from them. These stylometric features are separately bi-gram, tri-gram, quadri-gram. We calculate the frequency of these features and get 200 for author's stylometry vector space. After all, our system is ready for the author recognition process. For the text, We take the test books one by one and extract bi-gram, tri-gram, quadri-gram for authorship properties as we do for educational books. Following, Extracted most frequent 200 n-gram pass to naive bayes classifier. Naïve Bayes classifier decides who write the book among the authors previously introduced to the system. The system's estimation label and the real author of the book is compared and the results are noted at the end of work.

The comparison of the n-grams performances attained by Naïve Bayesian method is examined through this thesis. The achievements of the n-gram vector spaces about author detection were observed in this study. As a result of our observations, bigram vector space is failed. Besides, trigram and quadri-gram gave good accuracy result. It should be noted the best performance with 82% accuracy belongs to the quadri-gram. The other results and confusion matrix is located in the thesis.

With the recent developments of computer architectures, the amount of proceedings and articles about the understanding of human languages by the computers has enormously increased. Unfortunately, a large amount of these papers are related with English language. From this perspective, the thesis contributes the Natural Language Processing in Turkish language and provides motivation for the further studies on the field.

Keywords : Text classification, Author detection, Naïve Bayesian approach, N-gram.

1 GİRİŞ

1.1 Temel Dil Bilimi Kavramları

1.1.1 İletişim

İletişim: gönderici ve alıcı arasında gerçekleşen duygu, düşünce, davranış, fikir, his, bilgi v.s. alışverişidir ("İletişim Nedir," n.d.). İnsanlar tarihin başından beri doğada bulunan canlı cansız neredeyse her şeyle iletişim kurmaya çalışmıştır. Bu iletişimin konusu çoğu zaman insandan insana olsa da insanın hayvanlarla ya da belirli bir algısı var olan daha primitif varlıklarla iletişim kurmuştur.

İletişim kurulabilmesi için bir kaç unsura ihtiyaç vardır. Bunlar: gönderici(kaynak), ileti(mesaj), kanal, alıcı(hedef), dönüt(geri bildirim).

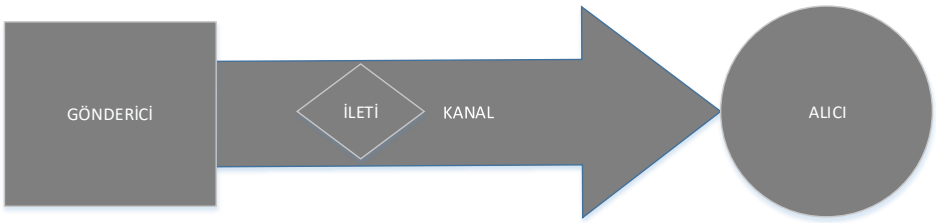
Gönderici (Kaynak): iletişimi başlatan unsurdur. Gönderici vermek istediği iletiyi anlamlı bir bütün haline getiren unsurdur.

İleti (Mesaj): gönderici ve alıcı arasında iletilmek istenen duygu, düşünce, bilgi veya verilerdir. İleti içeriği herhangi bir şey olabilir.

Kanal: gönderici ve alıcı arasında iletilecek olan mesajın gönderilme ortamıdır. Gönderici ve alıcının iletişim bağlantısıdır denilebilir.

Alıcı (Hedef): Gönderici tarafından iletilen mesajın ulaştığı öğedir.

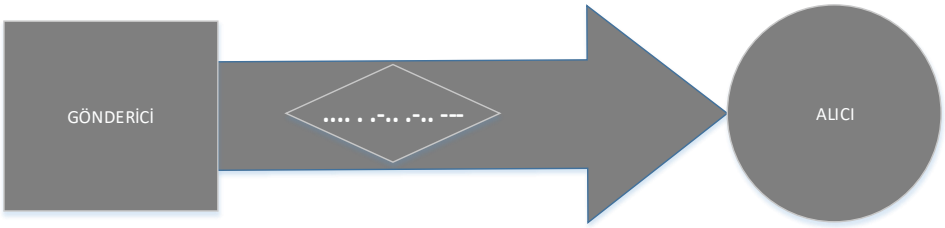
Dönüt (Geri Bildirim): Alıcının gelen iletiye karşı göndericiye verdiği cevaptır.



Şekil 1.1: İletişim

İletişimin kurulabilmesi için **Şekil 1.1**'de gösterildiği gibi gönderici, ileti, kanal ve alıcı olmak üzere dört temel unsura ihtiyaç vardır. Gönderici iletişim konusu olan iletiyi hazırlar ve kanal yardımıyla alıcıya gönderir ve böylece iletişim kurulmuş olur.

Bu dört temel unsur iletişimin sadece kurulmasını sağlar. Bunun yanında iletişimin anlamlandırılabilmesi için göndericinin gönderdiği iletiyi alıcının algılayabiliyor ve anlamlandırabiliyor olması gerekir. Bu duruma bir örnek olarak aşağıda görülmekte olan **Şekil 1.2**'ye bakınız.



Şekil 1.2: İletişim Anlamlandırma

Şekil 1.2'de gönderici alıcıya mors alfabesiyle "hello"(merhaba) mesajını göndermektedir. Bu iletişim de dört temel unsurun hepsi bulunmaktadır ve mesaj iletilmiştir. Fakat iletişimin anlamlandırılabilmesi için alıcının da mors alfabesini bilmesi gerekmektedir. Aksi takdirde gönderilmiş iletinin bir anlamı olmayacaktır. Bu noktada dikkat edilmesi gereken göndericinin iletiyi oluşturması ve alıcının iletiyi anlamlandırması kısımlarıdır.

İnsanlar tarih boyunca kendilerini ifade edebilmek için -ileti oluşturabilmek için- çeşitli yöntemler kullanmıştır. Bunlardan bazıları;

Basit seviyede:

- El işaretleri
- Görsel ifadeler
- Dokunma
- Basit sesler

Orta seviyede:

- Resimler
- Konuşma
- Yazışma

İleri seviyede:

- Telepati

Olarak verilebilir.

*Yukarıda ki örneklere bir ok örnek daha eklenebilir, tezin konusu iletişim olmadığı için kısa tutulmuştur.

Örnekler ne kadar çoğalsa da insanların ileti oluşturmaktaki en etkili aracı dildir. Dil bir başka ismi ile lisan (language) insanların ulaştırmak istedikleri iletileri belirli bir forma dönüştürebilecekleri ve bu formda gelen iletileri anlamlandırabilecekleri bir ifadeler bütünüdür.

1.1.2 Dil

İnsanoğlu birbiriyle bilgi, inanış, fikir, dilek, emir, teşekkür, sözler, hisler vesaire birçok şeyi paylaşabilir; bunun limiti sadece hayal gücüdür. Biz, gülerek: heyecanımızı, mutluluğumuzu: bağırarak: kızgınlığımızı, korkumuzu; yumruklarımız sıkarak: tehditlerimizi, şiddetimizi; gözlerimizi büyük açarak: şaşkınlığımızı, kınamalarımızı ifade edebiliriz (Eifring & Theil, n.d.) Bu iletişimlerden hiç biri dil kadar etkili değildir. Dil hakkında söylenebilecek onlarca şey vardır fakat hepsinden önce genel bir dil tanımı yapmak gerekir. Dil basit seslerin birleşimiyle kelimeler, kelimelerin birleşimiyle cümleleri oluşturan sözlü ya da yazılı bir iletişim aracıdır (Eifring & Theil, n.d.). İnsan dışındaki türler de iletişim kurmaktadır fakat bunların hiçbirinin kullandıkları iletişim insanların kullandıkları dil sistemleri gibi karmaşık değildir.

Dil binlerce farklı formlarda ve anlamlarda kelimeyi bünyesinde barındırır.

İnsanlar el hareketleri ve temas ile başladıkları iletişim kurma maceralarına basit sesler çıkararak, basit şekiller çizerek devam etmişlerdir. Basit sesleri yan yana getirerek daha karmaşık kelimeler ve kelimelerin bir araya gelmesi ile de cümleleri oluşturmuştur. Dilin birikimli olması özelliği olarak insanlar dili kullandıkça dil zenginleşmiş ve insanların çoğu şeyi ifade edebilmesine olanak sağlamıştır.

Dil zamandan, mekândan, insandan bağımsız olmayan yaşayan ve gelişen, değişen bir olgudur. Dil insanların kullanımı ile zamanla gelişmiş ve farklılaşmıştır. Dil; İnsanlar bir sürekli iletişim kurdukları insanların konuşma biçimi, kullandıkları kelimelerden etkilenerek; insanların çevrelerinde ifadede etmeleri gereken olaylara, olgulara kelimeler, cümleler üretmelerinden, yer yer unutulan ve farklılaşan yapılarıyla değişimlere uğramış ve günümüzdeki dünya dilleri ortaya çıkmıştır.

Dillerde var olan sesleri işaret etmesi için işaretlerle her ses sembolize edilmiştir. Bu sembollere harf ve harflerin bir araya geldiği kümeye de alfabe denir. Alfabedeki harfler bir araya gelerek kelimeleri oluşturur.

Kelimenin üç temel unsuru vardır: telaffuz, yazım, anlam. Anlam, telaffuz ve yazımdan tamamen bağımsızdır. Telaffuz ve yazım, anlama göre daha birbiri ile ilişkilidir fakat bazı dillerde bu ilişki daha zayıf olabilir.

1.1.2.1 Anlam

Kelimeler işaretlerin bir araya ya da seslerin yan yana gelmesiyle oluşmuş öbekler gibi görünse de; anlamı olamayan bir kelimenin dilde yeri yoktur. Dil içindeki anlamlı ses öbeklerine kelime denir. Kelimelerin anlamları dili kullanan kişilerce değişebilir. Bir kişi için anlamlı olan bir kelime diğeri için anlamsız ya da başka bir anlama gelebilir.

1.1.2.2 Telaffuz

Dilde var olan bir kelimenin okunuş biçimine telaffuz(pronunciation) denir. Dil; ilk olarak sesler ve seslerin birleşmesiyle ortaya çıkmıştır. Bu sesler çok daha sonrasında yazıya dökülmüştür. Dilde bulunan onlarca ses ve bunların bir araya gelmesiyle meydana çıkan onlarca ses birleşimi ile kelimeler oluşur.

1.1.2.3 Yazım

Dillerin alfabedeki kelimeleri bir araya getirip anlamlı kelimler üretmek için belirli kuralları vardır. Bunlara yazım kuralları denir. Eğer kurallara uyulmazsa dilin standardından çıkılmış olur. Dil standardından çıkılması anlamsız bir kelime üretileceği anlamına gelmez. Nitekim belirli dil içinde kendi dil standartlarına uymayan birçok anlamlı kelime var olabilir.

2 DOĞAL DİL İŞLEME

Doğal dil işleme insan dilini bilgi sayımsal (computational) olarak eğitme çalışmasıdır. Bir başka deyişle bilgisayarlara insan dilini nasıl anlayacaklarını ve üreteceklerini öğretme bilimidir (P. Dragomir Radev, n.d.) .

Bilgisayarlar bir dili anlayabilir ve insanlarla iletişime devam edebilmesi yirminci yüzyılın ilk yarısında bir bilimkurgu film sahnesi gibi gelmişti. Alan Turing'in makine ve zekâyı işleme (Computing Machinery and Intelligence) isimli klasik makalesinde "Makineler düşünebilir mi?" diye sormuş ve Turing deneyiyle makinenin insan gibi konuşup konuşamayacağını sorgulamıştır. Yapay zekâ çalışmalarıyla alanda birçok gelişme yaşanmış ve bugünkü seviyeye gelinebilmiştir. Bugün birçok web sitesi otomatik dil çevrimi yapabilmekte, cep telefonları insanların ne demek istediklerini anlayabilmekte, arama motorları yazdıklarımızı düzeltip, tamamlayabilmektedir. Ancak doğal dili tamamen anlayabilen makinelerden hala uzağız. Otomatik çevirimler hala bir insan danışmanın gözlemimden geçmek durumunda kalıyor, Turing testini hala tam geçebilmiş bir makine bulunmamaktadır (Kibble, 2013).

İnternetin gelişiminden önce birçok bilgi elektronik ortamlarda yapılandırılmış veritabanı(structured database) sistemlerinde tutuluyordu. Bu yapısallaştırılmış veriler yapılandırılmış sorgulama dili (Structured Query Language - SQL) ile sorgulama gereken bilgi istenildiği gibi veritabanı sisteminden getirilebiliyordu. İnternetin gelişmesiyle günümüzde birçok veri elektronik ortamda yapılandırılmamış şekilde durmaktadır. Bu durum var olan verileri işleyebilmek, anlam ve bilgiden bilgi çıkarmak gibi sorunları getirmektedir. Verilerin şifresini çözecek sistemlere ihtiyaç duyulmaktadır. Doğal dil işlemenin metin analitik, başlık tespiti, bilgi çıkarımı gibi alanları bu sorunlara çözüm aranmaktadır (Kibble, 2013).

Bazı modern uygulamalar:

- Arama motorları (Google, Yahoo, Bing, Baidu)
- Soru-Cevap (IBM's Wahtson)
- Doğal Dil Asistanı (Apple's Siri, Google Assistant)

- Translatioın Systems (Google Translate)
- Haber Özetleri (Yahoo)
- Otomatik Deprem Raporlayıcı (LA Times)

2.1 Tarihçe

Natural Language Processing (NPL) dilimizde Doğal Dil İşleme (DDİ) olarak geçmiş yapay zekânın bir dalıdır. Doğal dil işlemedeki amaç: insanların günlük hayatta yazarak, konuşarak kullandıkları dili bilgisayarlarda işleyerek bilgisayarın anlama ve anlamlandırma yeteneklerini artırmaktır. Kısacası insan dili ile bilgisayarlar arasında anlamsal bir bağ kurmayı amaçlar.

Doğal dil işlemenin tarihi 1950'lili yıllarda Alan Turing'in yazdığı "Computing Machinery And Intelligence" makalesiyle başladığı farz edilir, daha sonra bu makale içeriği "Turing Test" ismi ile anılmaya başlanmıştır. Bu makale makinelerin insan zekâsına yakın işler yapabilmesinin de başlangıcıdır.

1954 yılında "The Georgetown Experiment" altı Rusca cümleyi İngilizce'ye çevirebildi. Deney sahibi 3-5 yıl içerisinde dil çevrimi probleminin çözüleceğini öngörmüştü yalnız bu öngörü gerçekleşemedi. 1966 ALPAC (Automatic Language Processing Advisory Committee) raporuna göre, yıllar süren çalışmalara rağmen iyi sonuçlar alınamadı ve makine çevrimi fonlamaları düştü. 1980'de İstatiksel makine çevirim sistemleri geliştirilene kadar oldukça az araştırmalar gerçekleşti.

1960'larda oldukça başarılı sayılabilecek DDİ sistemi "SHRDLU" Terry Winograd tarafından geliştirildi. Bu yazılım sınırlı kelime ile sınırlı durumları işleyebiliyordu. Buna benzer bir program olan "ELIZA" kullanıcı ile insana benzer iletişime geçebiliyordu. Örneğin hasta "Benim kalbim acıyor" dese ELIZA "Neden kalbim acıyor dedin" diye cevap verebiliyordu.

1970'lerde varlığı kavrayan birçok program yazıldı; MARGIE (Schank, 1975), SAM (Cullingford, 1978), PAM (Wilensky, 1978), TaleSpin (Meehan, 1976), QUALM (Lehnert, 1977), Politics (Carbonell, 1979), Plot Units (Lehnert 1981). Bu süreçte insan konuşmasını simüle eden programlarda yazıldı: PARRY, Racter, Jabberwacky.

1980 yıllarının sonlarına doğru makine öğrenmesi algoritmalarıyla doğal dil işleme araştırmaları yeni bir dönemece girdi. Bu dönemecin sebebi; hesaplama gücünde istikrarlı artış ve Chomskyan dilbilimi teorileri egemenliğinin yerini makine öğrenmesi yaklaşımına uygun türden dil bilimine bırakmasıydı. Karar ağaçları

(Decision Trees) gibi algoritmalar zor koşul yapılarını (if-then) çözmeye yardımcı oldu. Daha sonraları "Hidden Markov Model", "İstatiksel model", "Olasılıkçı kararlar (probalistic decisions)" gibi yapılar çeşitli problemlerin çözümü için kullanıldı.

Dikkate değer en başarılı makine çevirimi IBM tarafından oldukça karmaşık bir istatiksel model kullanılarak yapıldı. IBM araştırmacıları çok dilli metinler üzerinde öğrenme algoritmalarını denediler.

Doğal dil işleme hala devam eden araştırma konusudur. Dünya üzerinde birçok bilim insanı doğal dil işleme üzerine çalışmalarını devam ettirmektedir.

2.2 Çalışma Alanları

DDİ çalışmaları oldukça farklı alanlarda kullanılmaktadır. Bazı kullanımları direkt olarak gerçek dünyayla ilgiliyken, bazen büyük problemleri çözmek için alt görevler olarak çalışmaktadır. Aşağıda DDİ üzerine yapılan araştırmalar özetlenmeye çalışılmıştır.

2.2.1 Otomatik Özetleme

Otomatik özetleme DDİ toplumu tarafından uzun yıllardır araştırılmakta olan bir alandır. Dragoimir R. Radev 2002'deki bir makalesinde otomatik özetlemeyi "bir ya da birden fazla metinden üretilen ve orijinal metinlerin önemli bilgilerini içeren, genellikle diğer metinlerden daha kısa olan metin" olarak tanımlamıştır (Radev, Hovy, & McKeown, 2002). Bu tanım bize otomatik özetleme hakkında üç önemli bakış açısı kazandırıyor (Das, 2007).

- Özet bir ya da birden fazla dokümandan üretilmeli.
- Kaynak dokümanların önemli kısımları yer almalı.
- Kısa olmalı.

Bu üç madde söylemde kolay olsa da bu işlemi yapmak biraz uğraş gerektiriyor. Özetleme işlemi için bir kaç adımı tanımlamak gerekir (Das, 2007).

1. Bilgi çıkarma (Extraction): Metinin önemli kısımlarını çıkarmak.
2. Soyutlama (Abstraction): Metinden çıkarılan kısımları daha genel ifade etmektir.
3. Birleştirme (Fusion): Çıkarılıp genellenen kısımları tutarlı bir şekilde bir araya getirmektir.
4. Kısaltma (Compression): Gereksiz kısımları atmaktır.

Google aramalarında arama sayfasına gelen internet sayfalarının özetleri ve Yahoo haber özetleri bu alana örnek sayılabilir.

2.2.2 Söylev Analizi

Bir söylemin ne istediğini, ince ayrımlarını açıklamaya çalışan DDİ alt bilimidir. Bu başlık birkaç ilişkili görevi kapsar. Bir görevler yazının yapısını tanımlar, bir başkası cümleler arası ilişki doğasını çözer, bir başka görev ise tanımlama ve kümeleme işlemini yapar.

2.2.3 Makine Çevirimi

Makine çevrimi; makine tarafından belirli bir dille yazılmış bir metni insan yardımı olamadan analiz ederek hedeflenen dile ait bir metin üretilmesidir. Oldukça zor bir DDİ alt alanıdır. Makine çevriminde önemli olan iki zorluk; yeterlilik (adequacy) ve akıcılıktır (fluency) (Antony, 2013).
Makine çeviriminde farklı yaklaşımlara dayalı çözümler bulunabilmiştir. Bunlardan bazıları (D.V, B.M.SAGAR, & S, 2014):

1. Direkt çevrim yaklaşımı: Bir dilden diğerine direkt olarak çevrim yapar. İki dilinde sözlük kütüphanesinin bulunması gerekir. Herhangi bir gramere ait yapıyı göz önünde bulundurmaz. Cümlenin anlamı korunmayabilir.
2. Transfer tabanlı yaklaşım: Bu yaklaşım cümlenin anlamını korumaya çalışır. Her iki dilinde gramerine ait yapısını göz önünde bulundurur.
3. Bileşik yaklaşım: Birden fazla yaklaşımın bir arada kullanılmasıdır.
4. Örnek temelli yaklaşım: Daha önden kaynak ve hedef dil arasındaki var olan çevrimlere bakarak çevrim işlemini gerçekleştiren yaklaşımdır.
5. İstatistiksel yaklaşım: Bu modelde danışmanlı-danışmansız makine öğrenmesi teknikleri kullanılır. Kaynak dil hedef dilde olası tüm olasılıklara çevrilir ve bu olasılıklardan en uygunu seçilir.
6. İnterlingua yaklaşımı: kaynak dil ilk önce aracı bir dile çevrilir. Bu aracı dil Aradil (İnterlingua) olarak geçer. Ardından aradil hedef dile çevrilir.

2.2.4 Biçimsel Bölütleme

Bir dildeki biçimsel kulları inceleyen DDİ alanıdır. Dünya üzerinde diller birbirlerine göre oldukça farklıdır. Örnek vermek gerekirse kimi diller ön ek kimileri son ek

almaktadır. Bu dillerin bu tip yapılarını çözümlemek biçimsel bölütlemedir. Biçimsel bölütleme ayrıca bir kelimeyi köküne kadar analiz edebilmemizi sağlar. Bir dilin biçimsel analizi için o dilde bulunan kelimeler, ekler, bağlaçlar gibi birimlerin bulunduğu bir sözlük kütüphanesine ve o dilin yapısına göre çalışan bir otomat programa ihtiyaç vardır (Nouri & Yangarber, 2011).

2.2.5 Doğal Dil Üretimi

Doğal dil üretimi makinelerin insan dili ile konuşabilir hale gelmesini hedefler. Makineler kendi dilleriyle işlemler yapıp kendi dillerince sonuçlar üretir. Bu sonuçları renkler, yazılar, grafikler olarak insanların anlayabileceği halde görürüz. Fakat zamanla sadece grafik değil makinelerin insanlarla konuşarak iletişime geçmesi, sonuçları insan dili ile söylemesi öngörülmektedir. Şimdilik bunu başaran bir kaç uygulama görmekteyiz. Örneğin Apple-SIRI.

Doğal dil üretimi, doğal dili anlamanın tam karşısında gibi dursa da doğal dili anlamayan makine doğal dil üretemeyecektir. Yani makine ilk önce insan dilini analiz etmeli ardından hedeflenen dile uygun gramer, sözdizimi ile anlamlı cümleler kurmalıdır (“Natural Language Generation,” n.d.).

2.2.6 Soru Cevaplama

2000’li yıllara girdiğimizde internet kullanıcı sayısı üstsel biçimde artmaya devam ederek milyonlara hatta milyarlara ulaştı. İnternet kullanıcıları ulaşmak istedikleri bilgileri bulmak için arama motorlarını kullanmaktadır. İnternet arama motorlarında insanların aradıkları şeyleri en doğru şekilde getirebilmek için yazdıkları sorguyu doğru anlamak durumunda ve ardından bu sorguya en doğru cevabın hangi sitede yer aldığının cevabını vermelidir. Burada devreye doğal dil işleme ve soru cevaplama teknikleri giriyor (Pundge, 2016).

Soru cevaplama DDİ’nin alt bilimdir. Bu bilim insan tarafından sorulan soruyu algılayacak ve ya kendi yapay zekâsıyla ya da sahip olduğu veritabanından en doğru cevabı getirmeye çalışmasıdır. İki tip soru-cevap sistemi vardır; açık alan (open domain), kapalı alan (closed domain). Açık alan soruyu her hangi bir alan kısıtlaması olmadan cevaplarken, kapalı alan kısıtlaması getirir(fizik, kimya) (Pundge, 2016).

2.2.7 Bilgi Çıkarımı

Bilgi çıkarımı bilgisayarsal dilde kritik rol oynar. Martin ve Jurafsky bilgi çıkarımını; "yapılandırılmamış bilgiden yapılandırılmış bilgi elde etme olarak tanımlamıştır" (Jurafsky & Martin, 1999). Aynı kavrama Grisman'ın tanımı ise: "doğal dil metninde bulunan ilişki veya olay kümelerinden anlamlı argumanlar çıkarma" (Grishman, 1997). Bilgi çıkarımı soru-cevap, bilgi getirme gibi alanlarda kullanılmaktadır.

Genellikle metinde bulunan nesneleri tespit ve sınıflama ile başlar bu da isimlendirilmiş varlık tanıma (Named Entity Recognation) işlemi demektir. Sonraki adım; referans çözümlemesi (Coreference Resolution) işlemidir. Bu işlemde nesnelerin birbiri yerine kullanıldığı kısımlar çözülür. Ardından metin içerisinde problemi çözüme götürecek bilgi aranır. Son adım ise metin içerisinden bulunan bilginin anlamlı bir yapıda sunulmasıdır (Porshnev & Redkin, 2014).

2.3 Doğal Dil İşleme İşlem Basamakları?

Geleneksel olarak doğal dil işleme sözdizimi(Syntax), anlambilim(Semantic) ve faydabilim (Pragmatic) ana başlıklarına ayrılarak analiz edilir. İlk olarak bir cümle sözdizimi olarak analiz edilir. Bu anlambilim açısından ya da sözlüksel açıdan bir sıra ve yapı sunar. Bunu kelimenin metin içerisinde söyleniş veya konumunun cümleye kattığı anlamı irdeleyen faydabilim analizi takip eder. Son kısım genellikle kelimenin cümle içerisindeki ilişkisel anlamını inceleyen söylem çözümlemesi(Discourse Analysis) ile ilgilidir. Kelimenin cümle içinde ne anlama geldiği, cümlenin metin içerisinde ne anlama geldiği gibi kısımlar bazı karışıklıklara neden olur. Fakat doğal dil işlemeyi kısımlara ayırmak yazılımsal açıdan anlaşılırlığını ve çözümlemesini kolaylaştırır (Indurkhya & Damerau, 2010).

Bununla beraber, üçlü ayrım bir metni doğal dil işleme işlemleri için sadece iyi bir başlangıç noktası olabilir. Eğer bir veri üzerinde doğal dil işleme yapılacaksa bir başka farkı adımlara ihtiyaç duyulur. **Şekil 2.1** doğal dil işleme adımları gösterilmiştir. Verilen şekilde DDİ adımları metinden anlam çözümleme adımlarına kadar verilmiştir. Bu işlem basamakları arasında ihtiyaca göre yeni adımlar girebilir veya bir basamak aradan çıkabilir. Ancak genel olarak bir DDİ çalışmasında bu işlem adımları bulunur.

Şekil 2.1 : Doğal Dil İşleme Adımları

Doğal dil işlemede ilk adım jetonlama ve cümle bölütlemesidir. Bu adım DDİ'de hayati öneme sahiptir. Elektronik ortamda bulunan metinler genellikle kısa, düzenli, iyi ayrılmış, yazım kurallarına uygun değillerdir. Bu metinlerin temizlenmesi ve işlenebilir parçalar haline gelmesi için bir adım gerekir.

2.3.1 Metin Ön İşleme

Bir ses ya da yazı olarak bulunan işlenecek veriyi doğal dil işleme ile işlenebilecek hale getirme basamağıdır. Bu aşamada veri çeşitli gürültü, hata gibi vesaire akışı bozan kısımlarından arındırılır ve jetonlamaya ve bölütlemeye uygun hale getirilir. Bu işlem verinin durumuna, işlenecek olan dilin özelliklerine göre değişiklik gösterir ve bundan dolayı kolay bir cevabı yoktur. Olabilecek durumlara ilişkin bir kaç paragraflık örnekler verebiliriz.

Her metin dil kurallarına uygun olmaz hatta içerisinde hedef dile uygun olamayan, kodlama farklılığından (UTF8, ISO-8859-1) dolayı eksik ya da yanlış karakterler bulunabilir. Bunların temizlenmesi ya da düzeltilmesi gerekir.

Tüm diller boşluklarla sınırlandırılmış kelimeler sunmaz. Örneğin İngilizce kelimeler boşluklarla ayrılabilirken, Çince, Japonca gibi diller boşluklarla ayrılmaz ve kelime bölütleme işleminde faklı bir yol izlenmesine ihtiyacı duyarlar. Farklı diller faklı işlemlere tabi tutulabilir.

2.3.2 Jetonlama

Bu işlem; işlenmeye uygun hale getirilmiş bütün veriyi işlenmek için kullanılacak küçük parçalar haline getirmektir. Bu parçalar kelimeler, cümleler, noktalamalar, harf kümeleri vesaire olabilir. Her bir jeton sonraki işlemlerde anlamlandırılabilecek şekle sokulur.

Jetonlama programlama dilleri gibi yapay dillerde oldukça kolaydır. Yapay dillerde kelimeler ve yapılarda belirsizlik yoktur. Doğal dillerde böyle bir lüks bulunmamaktadır. Doğal dillerin sınırlarını jetonlama şekli dilden dile, içerikten içeriğe, uygulamadan uygulamaya değişmektedir.

2.3.3 Sözcüksel analiz

Önceki işlemler bütünsel bir veri kümesini temizleyip jetonlarına ayırdık. Bu jetonların dil için ne anlama geldiği ya da yapacağımız işlem için ne anlama geleceğini belirtme işlemi, sözcüksel analiz kısmında yapılır. Bu analiz daha sonraki kısımlar için bir bilgisayarsal biçimbilim(Morphology) oluşturur. Ayrıca gelen her jeton ya da bölütlenmiş veri atomik olmayabilir, burada veriyi ihtiyaca göre ekleri ayrılabilir daha fazla bilgi çıkarımı yapılabilir. Bu işlemi anlamak için en basit örnek; kelimelerin o dilde isim, sıfat ya da fiil olarak sınıflandırılması verilebilir.

2.3.4 Sözdizimsel Analiz

Dil anlamlı kurallar bütünüdür ve her dilin anlamlı bir söz dizimi bulunmaktadır. Veri kümeleri her zaman dil kullarına uygun gelmese de bir cümlenin anlamını çözmek için dilin sözdizimsel kuralları kullanılır. Sözdizimsel analiz dilin gramerine ait ya da sözdizimsel özelliklerinin anlaşıldığı veya uygulandığı kısımdır.

2.3.5 Anlambilimsel Analiz

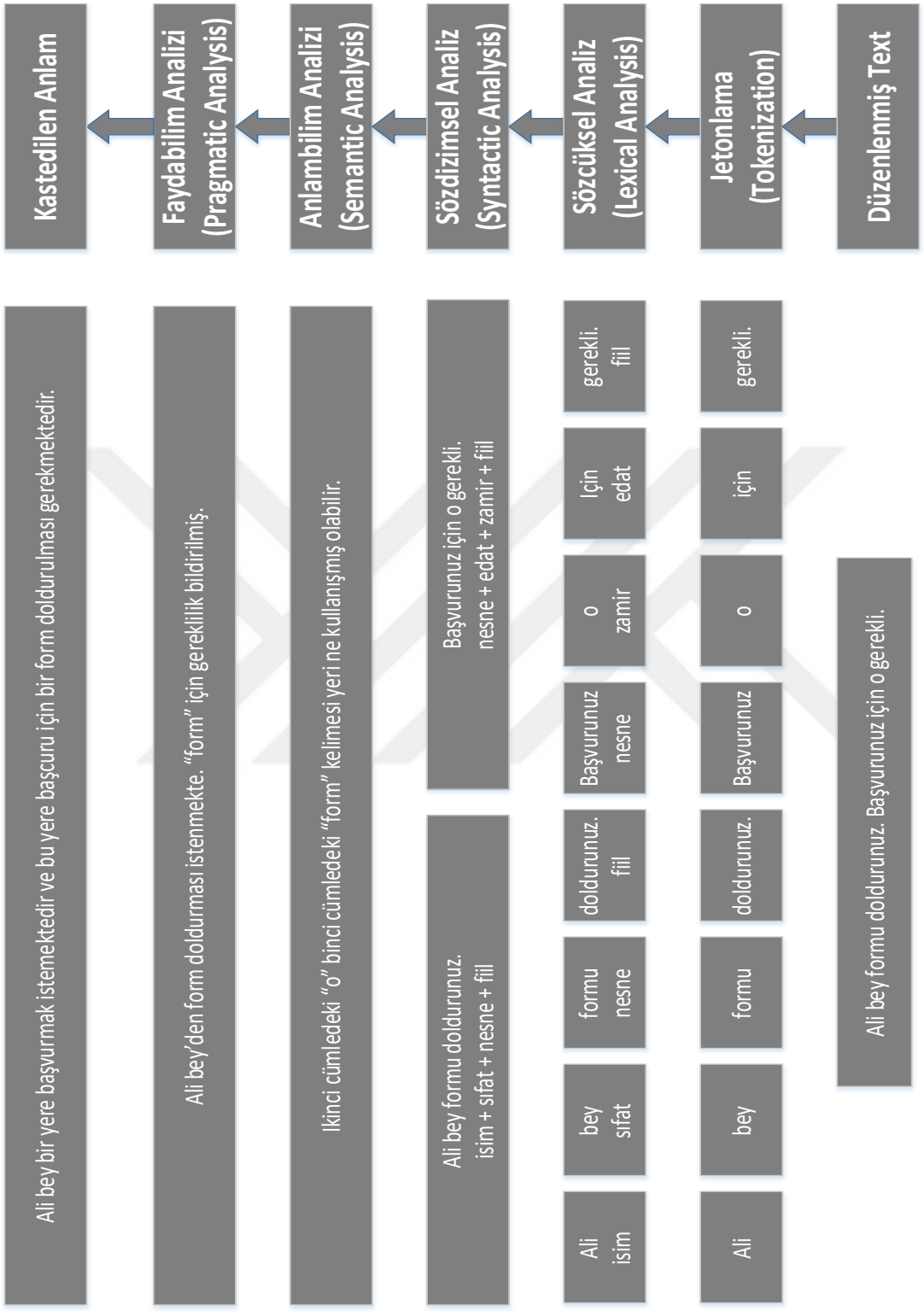
Sözdizimsel analiz edilmiş olan kelimelerin ya da jetonların anlamının anlaşılmaya çalışıldığı kısımdır. Bir kelimenin bulunduğu cümlede ne anlama geldiği gibi çıkarımlar yapılır. Burada unutulmaması gereken bir kelime bir cümle içinde birden fazla anlama gelebilir. Kelimenin olası anlamlarını çıkarmak bu bölümde olsa da ileri seviye neyin kastedildiği bir sonraki bölümün işidir.

2.3.6 Faydabilim Analizi

Bu bölümde anlambilimsel analizi yapılmış olan kelimelerin ve cümlelerin aslında neyi kastettiği bulunmaya çalışılır. Tüm metin içerisinde ne anlama geldiği, neye atıf yaptığı gibi ilişkiler faydabilim analizinde yer alır. Bu işlem oldukça karmaşıktır ve hala tam anlamıyla çözülebilmiş değildir.

Şekil 2.2'de bir örnek üzerinde DDİ adımları verilmiştir. Bir örnek cümle üzerinden DDİ'in nasıl yapılacağı, hangi basamakta ne olduğu ifade edilmeye çalışılmıştır. İlk basamakta bir cümle verilmiş ve DDİ uygulanması istenmiştir. İkinci basamakta bu cümle jetonlarına ayrılmıştır. Duruma göre jetonlama kelimelere bölme dışında başka da olabilmektedir. Bir üst basamakta kelimelerin gramersel olarak öğreleri bulunmuştur. Söz dizimsel analizi sonraki basamakta, söz dizimsel analiz kısmında gelmiştir. Semantik, cümleden anlam çıkarma kısmında eş atıf çözümlemesi yapılmış ve kelimelerin birbirini ifade edtmesi çözülmüştür. Diğer adım fayda bilim adımıdır. Fayda bilim adımında örneğin telefon formu unutmaması için Ali Bey'i uyarabilir, yani kurulan cümlenin faydalı bir şekilde kullanılması anlamına gelir. Kastedilen anlam kısmı ise artık cümlenin tamamen çözümlenip makinenin tıpkı bir insan gibi cümleyi yorumlamasıdır.

Örnekler çoğaltılabilir, farklılaştırılabilir ancak burada temel olarak DDİ deyince ne demek istenildiğini göstermek için basit bir örnek üzerinde gösterilmeye çalışılmıştır.



Şekil 2.2 : Doğal Dil İşleme Örnek Adımları

3 METİN İŞLEME

Bilgisayarların yaygınlaşmasıyla metin dokümanların elektronik ortamlarda bulundurulması artmıştır. Günümüzde elektronik ortamlarda bulunmayan dokümanlar giderek azalmaktadır. Bilim, ekonomi, edebiyat, devlet işleri vesaire başlıklardaki neredeyse tüm dokümanlar elektronik ortamlarda bulundurulmaya başlanmıştır. Ne yazık ki elektronik ortamlara taşınan bu dokümanları işlemek yapılandırılmış veriler gibi kolay olamamaktadır. Yapısal olmayan metin dokümanları bulanık mantık ve belirsiz ilişkiler içerdiğinden belirli zorluk getirmektedir. Metin işlemenin amacı metin içerisinde bulunan birçok kelime arasında bulunan gizli bilgileri çıkarmak, belirsizlikleri, bulanık mantıkları, muğlaklıkları çözümlemektir (Hotho, Nürnberger, & Paaß, 2005).

Literatürde bilgi keşfi (Knowledge Discovery) ya da veritabanı üzerinde bilgi keşfi üzerine çeşitli tanımlar bulunmaktadır. Bu tanımlardan biri; "Veritabanı üzerinde bilgi keşfi veri içerisindeki değerli, yeni, yararlı ve anlamlı bağıntıları tanımlama işlemidir." (Fayyad, Piatetsky-Shapiro, & Smyth, 1996). Veritabanının analizi içerisinde bulunan veriler arası gizli bağlantıları ve bağlamları bulmayı amaçlar (Hotho et al., 2005). Veritabanı yapılandırılmış veri olduğundan bu verilerin işlenmesi bakış açısına göre kolay olabilir.

Veritabanlarının aksine metin dosyaları yapılandırılmamış verilerdir. Her şeyden önce bir metin dokümanı içerisinde oldukça fazla gürültü ve anlamlandırılmamış(bilgisayar tarafından) yapılar barındırır. Gürültü olarak karakter hataları, gramer hataları vesaire sayılabilir. Anlamlandırılmamış yapılardan kasıt ise yazılan yazılar insanlar için anlamlı olsa da bilgisayar için anlamsızdır. Metin işleme ile ilk önce metin dokümanındaki gürültü giderilerek ardından bilgisayara anlamsız gelen yapıların çözümlenmesi gelir.

Büyük doküman dizinlerinde işlem yaparken metinleri ön işleme sokarak bilgileri veri yapıları biçimine sokmak gerekir. Metin içerisindeki anlambilimsel, sözdizimsel yapıları çıkarabilecek bir kaç farklı metot bulunsa dahi birçok metin işleme yaklaşımı

metin dokümanlarının kelime yığıtlarından oluştuğu varsayımına dayanır. Ancak bir kelimenin metin içerisindeki önemini göstermek için sayısal önemini ile vektör gösterimi kullanılır. Baskın olan yaklaşımlar vektör uzayı modeli (vektör space model), olasılıkçı model (probabilistic model), mantıksal model (logical model) (Hotho et al., 2005). Metin Çözümleme

Büyük doküman koleksiyonlarında metin işleme yapmak için dokümanları ön işlemlerden geçirip onların bilgilerini veri yapıları içerisinde tutmamız gerekir. Metin içerisindeki bilgileri yapısallaştırmak için sözdizimsel ve anlamsal yapısından faydalanmak gibi bir kaç metot bulunsa da temel fikir her metinin bir kelimeler seti ile temsil edilebileceğidir. Buna İngilizce olarak "bag of words" denilmektedir. Metin çözümleme işlemi metni temsil edecek kelime kümelerinin bulunmasıdır.

3.1 Metin Önişleme

Bir dokümandaki tüm kelimeleri analiz edebilmek için ilk olarak jetonlama işlemi yapılmalıdır (tokenization). Jetonlama işleminde genellikle –özel durumlar hariç- tüm noktalamalar, boşluklar, alfabetik olmayan karakterler yok sayılır. Jetonlama işlemi sonucunda metinin bir sözlüğü oluşturulmuş olur.

Metin ön işleme tam bir standarttı olmasa da genel bir algoritma tanımlayabiliriz.

D → Döküman seti

d → Döküman seti içerisindeki bir döküman

T → Döküman kolektsiyonunun sözlüğü

t → Döküman seti içerisindeki bir dökümanın sözlüğü

$$d \in D \qquad t \in T$$

$$T = \{t_1, \dots, t_m\}$$

$$Terim\ Sıklığı(Term\ Frequency) = tf(d, t)$$

$$\overrightarrow{t_d} = (\, tf(d_1, t_1), \dots, tf(d_m, t_m)\,) A = \pi r^2$$

$$\overrightarrow{t_X} := \frac{1}{|X|} \sum_{t_d \in X} t_d$$

Denklem 3.1 : Terim Sıklığı

3.1.1 Filtreleme, Kelime Kökeni, Köke İnme

Dokümanın ön işlemede çıkartılmış olan sözlüğünün boyutu düşürmek için filtreleme, kelime kökeni bulma ve bölütleme yapılabilir.

Filtreleme: genellikle metin hakkında pek bir bilgi içermeyen kelimeleri sözlükten çıkarma işlemidir. Örneğin bağlaçlar, edatlar sözlükte yer almasına gerek yoktur bu yüzden filtrelenir. Çok sık tekrar eden veya çok az tekrar eden kelimeler de uçbirim (outlier) olarak görülüp filtreleme işlemine tabi tutulabilir (Hotho et al., 2005).

Kelime Kökeni(Lemmatization): Bir kelimeyi parçalarına ayırarak içerisinde bulunan en temel birimleri bulmaya çalışır. Kelimelerin temel birimlerine ulaştıkça sözlükteki kelime sayısı düşecektir. Bu işlem oldukça zor olabilir ve zaman alacaktır (Hotho et al., 2005).

Köke inme (Stemming): Bu işlem kelime kökeni bulmaya benzese de küçük bir ince ayrım farkı vardır. Bu işlem kelimedeki sadece sık kullanılan ekleri siler. Bir örnek vermek gerekirse çoğul, aitlik ekleri kelime gövdesinden ayrılır (Hotho et al., 2005).

Terim Seçimi

Dokümanların sözlüklerini daraltmak için kullanılan başka bir teknikte terim seçimidir. Burada amaç dokümanı en iyi temsil edecek sözcükleri seçmektir. Bunun için anahtar seçme algoritmaları kullanılabilir.

Kelime seçiminde doküman içerisinde ne kadar değerli olduğu bulunur. Elime seçimine ilişkin bir yöntem (Lochbaum & Streeter, 1989) numaralı makalede matematiksel olarak açıklanmıştır. Bu makalede bir kelime birden fazla dokümanda geçiyorsa düşük değerli, buna karşın bir dokümana özelse yüksek değerli olarak hesaplanmaktadır. Bu işlemin ardından doküman sözlüğünde o dokümana özel kelimelerin dokümana ait değerleri tutulmuş olur.

3.1.2 Vektör Uzayı Modeli

Basit veri yapıları kullanılarak herhangi bir anlamsal bilgi kullanmaksızın oluşturulan vektör uzayı modeli büyük dokümanlarda oldukça etkili analizler yapabilmemizi sağlar. İlk olarak Salton tarafından (Salton, Wong, & Yang, 1975) numaralı makalede önerilmiştir.

Vektör uzayı modeli dokümanları m-boyutlu uzayla temsil eder ve her bir doküman (d) sayısal özellik vektörüyle tanımlanır. Böylece dokümanları karşılaştırmak, içerisine sorgu yazıp bilgi çekmek kolaylaşır.

Vektörün her elemanı genellikle bir kelimeye karşılık gelir. Vektörün boyutu dokümandan çıkacak olan kelime sayısıyla orantılıdır. Vektör uzayı uygulamalarından en basiti ikilik sistemli olanıdır. Burada bir kelimenin dokümanda varlığı ve yokluğu gösterilir. Vektör uzayının performansını artırmak için her elemanın ağırlıkları ile oynanabilir. Ağırlık değeri olarak daha önce bahsettiğimiz terim seçimi algoritmalarından faydalanılabilir.

3.1.3 Dilsel Ön İşleme

Çoğu zaman metin işleme ön işlemeden sonar daha fazla işleme ihtiyaç duymaz. Bununla beraber bazen dokümanlar dile ait işlemlerden geçirilmesi daha fazla bilgi çıkarımına yarayabilir. Bunu yapabilmek için aşağıdaki yaklaşımlar kullanılabilir (Hotho et al., 2005).

Metin parçası isimlendirme: Dokümanda bulunan kısımlara bir başlık altında gruplamadır. Bir örnek verirsek isim, fiil, sıfat olarak yapılan ayrım olabilir.

Metin Bölütleme: birbiriyle ilişkili kelimeleri gruplama işlemidir.

Kelime Duyarlı Belirsizlik Giderme: Kelimeler cümle içerisinde faklı anlamlarda kullanılabilirler. Birden fazla anlam belirsizlik demektir. Kelimelerin cümle içesindeki muhtelif giderme işlemi yapıldığı takdirde daha anlamlı analizler olacaktır.

3.2 Metin İşleme İçin Veri Madenciliği Teknikleri

Metinler ön işlemlerden geçirildikten sonraki işlem bilgi çıkarımıdır. Yapılandırılmış metinin veri madenciliği ile işlenmesi bir nebze daha kolaylaşır. Metinleri işlemede kullanılan birçok veri madenciliği tekniği bulunmaktadır. Bu metotları bilmek ve doğru yerde, doğru tekniği kullanmak veri madenciliği etkili kılacaktır. Bu bölümde genel veri madenciliğinde kullanılan teknikleri ele alacağız.

3.2.1 Sınıflama

Sınıflama, belirli nesneleri önceden tanımlanmış sınıflara göre kategorize etmektir. Bu işlemi metinler üzerinde de yapabiliriz. Bir örnek vermek gerekirse internete bulunan haberleri spor, politika, sanat olarak sınıflayabiliriz.

Veri madenciliği açısından sınıflama ilk olarak bir eğitim seti ile başlar. Eğitim seti önceden sınıflanmış dokümanlardır. Eğitim seti ile birlikte sınıflama kuralları belirlenmeye çalışılır. Sistemin eğitimi için dokümanlar belirli bir eğitim fonksiyonuna sokulur. Bu fonksiyon kullanılan eğitim modeline göre değişecektir. Bu konuya daha sonra değineceğiz.

$D_t \rightarrow$ Döküman Eğitim seti (Documents Traning set)

d → Döküman Etiğim seti içerisindeki bir döküman (Document)

L → Etiket Seti (Label Set)

l → Etiket (Label)

$$D_t = \{d_{t1}, \dots, d_{tn}\}$$

$$L = \{l_1, \dots, l_m\}$$

$$f: D \rightarrow L \qquad f(d) = l$$

Denklem 3.2 : Etiket ve Doküman Seti

Sınıflama fonksiyonu eğitildikten sonra bu fonksiyona eğitilen sisteme uygun herhangi bir metin sokulabilir. Sisteme girilen dokümanlara test dokümanları seti olarak adlandırabiliriz (D_{ts}). Test dökümanlarının doğru sınıflanmış olması sınıflama fonksiyonumuzun performansını gösterir. Bu doğruluğu değerlendirmek için bilimsel bazı hesaplamalar gerekir. Tezde kullanılmış olan değerlendirme merikleri bir sonraki başlıkta bulunmaktadır.

3.2.2 Değerlendirme

Sistemin çalışması sonucunda değerlendirme işlemi yapılmalıdır. Değerlendirme işlemleri çeşitli formül hesaplamalarıyla yapılabilir. Bu tez kapsamında Doğruluk, Hata oranı, Precision, Recall ve f-score değerleri hesaplanmıştır.

Toplam dökümanın doğru olarak sınıflanmış dökümana oranı sistemin doğruluğunu (accuracy) verir.

$D_{ts} \rightarrow$ Döküman test seti (Documents Test Set)

$d_{ts} \rightarrow$ Döküman Etiğim seti içerisindeki bir döküman (Document)

$C_{td} \rightarrow$ Doğru Sınıflanmış dökümanlar

$C_{fd} \rightarrow$ Yanlış Sınıflanmış dökümanlar

$$Doğruluk(Accuracy) = \frac{\sum C_{td}}{\sum d_{ts}}$$

Denklem 3.3 : Doğruluk Hesabı

Sınıflamada çoğu zaman girilen dokümanlar hiçbir sınıfta bulunmayabilir. Bu da doğruluğu etkileyen bir durumdur. Burada sınıfı bulunamayan dokümanları alternatif

$$F - Score = \frac{2}{{}^{1}/_{Hassasiyet} + {}^{1}/_{Kesinlik}}$$

Denklem 3.5 : F-Score Hesabı

bir sınıfa atarsak yüksek bir doğruluk yakalarız. Bu etkilerden kaçınmak için kesinlik (Precision) ve hassasiyet (Recall) hesaplamaları kullanılabilir. Hassasiyet; işleme sokulan dokümanlar arasında hedef sınıflara ait olan doküman sayısının oranıdır. Kesinlik; işleme sokulan belgelerin hangi kesimin alındığıdır. Görüldüğü gibi iki formül arasında sadece bir yer değiştirme bulunmaktadır.

$$Kesinlik(Precision) = \frac{\{İlgili\ dökümanlar \cap Tüm\ dökümanlar\}}{Tüm\ dökümanlar}$$

$$Hassasiyet\ (Recall) = \frac{\{İlgili\ dökümanlar \cap Tüm\ dökümanlar\}}{İlgili\ dökümanlar}$$

Denklem 3.4 : Hassasiyet ve Kesinlik Hesabı

Birçok sınıflayıcı hedef sınıflara üyelik derecesini hesaplar. Eğer dokümanlar hedef sınıflarda varsa hassasiyet oldukça yüksek çıkacak. Bununla beraber yanlış sınıflandırılmış dokümanlar düşük hassasiyet değerine neden olacaktır. Sonuç olarak kesinlik yükselirken hassasiyet düşmektedir. Burada sınıflayıcının performansını hesaplamak için F-Score değeri hesaplanır.

4 YAZAR TANIMA

İstatistiksel ve bilgisayarlara dayalı yazar tanıma Mosteller ve Wallace (1964) tarafından "the authorship of the disputed Federalist Papers" isimli çalışmalarında ismi anılmıştır (Mosteller & Wallace, 1964). Elektronik dokümanların artmasıyla yazar tanıma artık bir ihtiyaç haline gelmiştir. Ancak sonraki yıllarda çalışmalar olsa da makine öğrenmesi, doğal dil işleme alanları belirli bir bilimsel seviyeye gelene kadar bir sonuç alınamamıştır. İstatistik ve bilgisayar algoritmaları kullanılarak günümüzde bu mümkün hale gelmiştir.

Yazar tanımadaki ana fikir elektronik ortama aktarılmış metin belgelerinin daha önce belirtilmiş özelliklere göre hangi yazara ait olduğunu tespit eden otomatik sistemler oluşturmaktır.

4.1 Literatür taraması

Yazar tanıma üzerine ilk çalışmalar Mosteller ve Wallace tarafından 1964 yılında yapıldı (Mosteller & Wallace, 1964). Mosteller ve Wallace faklı yazarların yazmış olduğu 146 politik makaleyi kimin yazdığını buldurmaya çalıştılar. Bu işlemi yaparken yaygın kelimelerin sıklık bilgilerini bayes istatistik hesaplamalarına soktular. Aldıkları sonuç oldukça pozitifti.

Yapılan bu çalışma sadece kelime sıklığına bakıyor aslında yazarın yazı stili kimliği hakkında bize bilgi vermiyordu. 1990'larda yapılan bir çalışmada cümle uzunluğu, kelime uzunluğu, kelime sıklığı, karakter sıklığı, kelime zenginliği gibi özellikleri içeren bir çalışma yapıldı (Bennett & Mangasarian, 1992). Daha sonra 1998 yılında Joseph Rudman'ın kendi belirlemiş olduğu bine yakın özellikle bir yazar tanıma çalışması yayınlamıştır (Rudman, 1998). Tüm bu çalışmalarda bilgisayar bir asistan gibi görülmüş yani yardımcı olarak kullanılmıştır. Bir başka değişle tamamen otomatik bir yazar tanıma için çalışılmamıştır. Buna en güzel örnek Morton ve Michealson'nun geliştirmiş olduğu CUSUM tekniğidir. Bu teknik mahkemelere dahi kabul edilmiştir. Bu ön çalışmalarda en büyük eksiklik objektif bir metin işleminin olmayışıdır.

Tahminin doğruluğunu hesaplayabilecek bir sistem geliştirilememiştir. Bunun bir kaç nedeni bulunmaktadır:

- Metin dosyalarının oldukça büyük olması.
- Kelime, cümle yayılımının homojen olmaması.
- Yazar sayısının az olması.
- Yazı gelişiminin başlıktan bağımsızlaşması.
- Faklı metotların karşılaştırılma zorluğu.

90'lı yılların sonlarına doğru yazar tanıma çalışmalarında bazı şeyler değişti. Özellikle internetin gelişimiyle dokümanların elektronik ortamlara aktarılması yazar tanıma işleminde oldukça değişime neden olmuştur. Bununla birlikte bilgi çıkarımı, makine öğrenmesi ve doğal dil işleme alanlarındaki gelişmeler de yazar tanıma probleminin çözümüne ilişkin faklı bakış açıkları geliştirilmesini sağlayan bazı gelişmeler şunlardır:

- Bilgi çıkarımı alanında yaşanan metin gösterimi ve sınıflamasındaki etkili teknikler.
- Çok boyutlu ve ayrık veri üzerindeki makine öğrenmesi algoritmalarındaki gelişmeler.
- Doğal dil işlemenin metin analizi ve stil işleme işlemlerindeki etkili çözümler.

Metin dosyalarının elektronik ortamlarda erişilebilirliği artınca birçok farklı uygulama yapılmaya başlanmıştır; suçlu tespiti, kopya suçları, virüs tespiti vs.

Tipik yazar tanıma probleminde yazarı bilinmeyen bir metin verilen aday yazarların yazarlık özellikleri arasında hangisine uygunsa onun olduğu belirtilir. Makine öğrenmesi bakış açısına göre bu işlem çok sınıflı tek etiketli kategorize etme işlemidir. Bu işlem genellikle bilgisayar bilim adamlarınca yazar tanıma olarak adlandırılır. Bu problemin ötesinde yazar analizi işlemi aşağıdaki gibi tanımlanabilir:

- Yazar doğrulama (Verilen metinin verilen yazara aitliğini belirler) (Moshe Koppel & Schler, 2004).
- Eser hırsızlığı tespiti (iki metin arasındaki benzerlik derecesini bulma) (Stein, Koppel, & Stamatatos, 2007) .
- Yazar profili ya da kategorizesi (yaşı, cinsiyeti, eğitimi)(M. Koppel, 2002) .
- Stilistik uyumsuzluklar (çok yazarlı metinlerde) (Collins, Kaufer, Vlachos, Butler, & Ishizaki, 2004).

4.2 Yazar Tanıma İşleminde Kullanılan Metodolojiler

4.2.1 Yazarın Metin Üzerindeki Stilsel Özellikleri

Yazar tanımada önerilen stil işaretçileri olarak isimlendirilen yazarın yazı özelliklerini faklı kıstas ve etiketlerle belirterek kullanan model önerilmektedir (Holmes, 1994) (Zheng, Li, Chen, & Huang, 2006). Şimdiye kadar yazarın biçembilim özellikleri olarak algoritmalara dâhil edilen kısmı sadece bilgisayarsal hesaplanabilecek ölçümlerdir. İlk olarak sözcüksel ve karakter özellikleri ardından kelime karakter sıraları özellik olarak alınabilir. Sözcüksel özellikler karakter özelliklerine göre daha karmaşık olsa da geleneksel teknikler ilk onları temele aldığı için başlangıçta bununla başlanabilir. Ardından sözdizimsel (syntactic) ve anlambilimsel(semantic) özellikler daha derin dilsel analiz istemektedirler. Bunlar dışında başka özelliklerde yazarın sitilsel özellikleri dâhil edilebilir.

4.2.1.1 Sözcüksel Özellikler

Bir metini göstermek için en kolay yollardan biri onu jetonlarına ayırmaktır. Jeton bir kelime, sayı ya da noktalama işareti olabilir. Metinde bulunan cümle ve kelime uzunluklarının bir özellik olarak kullanıldığı çalışmalar bulunmaktadır (Mendenhall, 1887) . Bu yaklaşımın en basit avantajı dilden bağımsız olmasıydı yani her hangi bir dil üzerinde çalışabilmekteydi. Ancak bazı diller için bu işlem geçersiz olabilmekteydi: Cince. Bu dillerde kelimelerin jetonlaşrılma işlemi oldukça zor olabilmektedir. Bu gibi dillerde ancak cümle boyutları belirlidir ve cümleler jetonlaştırma işleminin söz konusu olurlar.

Bir metinde bulunan kelime zenginliği de bir özellik olarak yazar özelliklerine katılabilmektedir. Kullanılan kelimelerin oranı (Fiil/İsim) dahi bir stil özelliği olabilmektedir (de Vel, Anderson, Corney, & Mohay, 2001). Ne yazık ki kelime zenginliği metin uzunluğuna da bağlı olan bir özelliktir. Bu özelliği dengelemek için bazı fonksiyonlar önerilmiştir (Tweedie & Baayen, 1998). Buna rağmen tek başına kullanımı önerilmez bir durumdur.

En basit yaklaşım olarak kelime sıklığı vektörü oldukça yaygın kullanılmaktadır. Yazar tanıma çalışmalarının birçoğu kelime özelliklerine dayanmaktadır. Bu başlık temelli sınıflama yapan araştırmacıların da kelime kesesi (bag-of-word) yaklaşımıdır (Sebastiani, 2002). Stil temelli yaklaşım ile kelime sıklığı yaklaşımı arasındaki temel fark stil temellide bir kelimenin diğer yazarlarda kullanılmama oranı değer olarak

alınabilirken kelime sıklığı sadece doküman bazında kalmaktadır. Bir başka değişle stil temelli yaklaşım vektör uzaklığını hedef alır (Burrows, 1987) (Argamon & Levitian, 2005).

Kelimeler, cümleler arasındaki bağlantıyı veren kelimelere fonksiyon kelimeleri (function word or functor) denir. Bu kelimeler bağlaç, fiil, edat vesaire olabilirler. Özellik için seçilecek fonksiyon kelimeleri rasgele ve dil bağımsız olarak seçilmeye özen gösterilir. Buna rağmen birçok fonksiyon kelimesi İngilizce olarak seçilmiştir. Abbasi ve Chen (Abbasi & Chen, 2005) nolu makalesinde 150 fonksiyon kelimesi, Argamon, Saric ve Stein (Argamon & Levitian, 2005) nolu konferans bildirisinde 303 kelime önermiştir. Bunun dışında başka çalışmalarda 365, 480, 675 kelime önerildiği olmuştur.

Oldukça basit ve başarılı metotlardan biri: metin içerisinde bulunan kelime sıklıkları çıkarmaktır. Eğer bir yazarın bilinen birden fazla metini varsa bu diğer metinler üzerinde de uygulanabilir. Bundan sonrası kaç kelimenin yazar özelliği olarak kullanılacağıdır. Daha önceki çalışmalarda en fazla 100 en sık kelimenin bir yazarı temsil için yeterli olduğunu söylemektedir (Burrows, 1987) (Burrows, 1992). Bununla birlikte yazar temsil uzayının çok genişlediği çalışmalar vardır. Kimi araştırmacılar metinde en az iki kere geçen her kelimeyi alırken bazıları belirli bir limit değeri koyarak yazar özellikleri çıkarmışlardır (Moshe Koppel, Schler, & Bonchek-Dokow, 2007) (Efstathios Stamatatos, 2006) (Madigan et al., 2005).

Kelimeleri jetonlarına ayırmak kolay olsa da kelime temelli yazar tanımada büyük-küçük harf çevrimi gibi kolay işlemlerden eklerin atımı(stemmering) (Sanderson & Guenter, 2006), kök bulma (lemmatizing) (Tambouratzis et al., 2004) gibi zor işlemlere kadar farklı işlemlerden geçirildiği çalışmalar olmuştur. Bir başka çalışmada kelimelerin soyut kelimelere dönüştürüldüğünü görmekteyiz (Halteren, 2007).

Kelime kesesi yaklaşımı basit ve etkili bir yaklaşım fakat kelime sırasını göz ardı etmektedir. Kelimelerin ardaşıllığı bir özellik olarak kullanılabilirler. Kelime kesesinde ise kelimeler ayrı ayrı jetonlaştırıldıklarından kelimelerin ardışıklık özellikleri kaybolmaktadır. Bunun önüne geçebilmek için kelime n-gram yaklaşımı önerilmiştir (Sanderson & Guenter, 2006) (Coyotl-Morales, Villaseñor-Pineda, Montes-y-Gómez, & Rosso, 2006) . Kelime n-gram yaklaşımında kelimelerin n sayısınca jetonlara ayrılır ve bu jetonlarla işlem yapılır. Bu yaklaşımda sorun kısa metinlerde yazar tanıma için önemli olabilecek kelime ardışıklarının yer almamasıdır.

Bir başka n-gram yaklaşımı da yazarın stil özelliklerinden çok içerik bağımlı frekanslar vermesidir (Gamon & Grey, 2004).

Bir başka yaklaşım yazar hatalarından yola çıkmaktadır (Moshe Koppel & Schler, 2003). Heceleme, boşluk, format hataları yazar için özellik olarak kullanılarak yazar özellikleri çıkarılır. Bu yöntemin zayıf noktası günümüz teknolojisiyle birlikte yazım hatalarını en aza indiren programlar olmasıdır.

4.2.1.2 Karaktersel Özellikler

Teoride bir metin karakter sıralamalarından oluşmaktadır. Durum böyle olunca karakterlerden yazar özellikleri çıkarılabilir olarak görülmektedir. Büyük-küçük harfler, sayı sıklıkları, karakter sıklıkları, harf sıklıkları gibi ölçümler yazar vektörünü oluşturabilir (Zheng et al., 2006). Bu tip bilgiler her metin üzerinde kolayca erişilebilirler.

Karakterlerden yazar özellikleri çıkarmak için en basit yaklaşım n-gram tekniğinden faydalanmak olabilir. N-gram ikili, üçlü, dörtlü vesaire olabilir. Bu teknikle hem sözcüksel özelliklere hem de içeriksel özelliklere erişilebilir. N-gram tekniği yazım, noktalama, boşluk hatalarından pek etkilenmez (Moshe Koppel & Schler, 2003). N-gram tekniği jetonlamada zorluk çekilen doğu dillerinde de özellik çıkarmada kolaylık sağlamaktadır (MATSUURA & KANADA, 2000).

N-gram tekniğinde yazarın en önemli özelliği en çok tekrar eden gramdır. Fakat bir problemde dilde bulunan bağlaç, edat gibi kelime ve cümle birleştiricilerdir. Bu kelimeler bir metinde en çok tekrar eden gramlar olabilir ya da özellik listesine girerler. Bununla beraber bir yazarın bu kelimeleri kullanma sıklığı da bir özellik olabilir (Efstathios Stamatatos, 2006).

4.2.1.3 Söz Dizimsel Özellikler

Daha ayrıntılı özellik uzayı çıkarmak için kullanılabilecek bir yöntem de söz dizimsel özellikleri kullanmaktır. Burada ana fikir bir yazarın benzer söz dizimsel özellikleri yazılarına yansıtacağıdır. Bundan dolayı yazarın kullandığı söz dizimlerinin yazarın parmak izi gibi olacağı düşünülmektedir. Genellikle fonksiyon kelimeleri (bağlaç, edat) dizimsel özelliklerde kullanılabilir. Fakat bir kelimenin fonksiyon kelimesi olup olmadığını anlama işlemi DDİ araçlarıyla ancak çıkarılabilir.

Dil bağımsız söz dizimsel özellikler çıkarmak ancak jetonlaştırılmış kelimelerin ardışıklığına bakılarak olabilir. Bu durumu ön ve son ekler düşünüldüğünde oldukça

zor olduğu anlaşılabilir. Dil bağımlıda bir kelimenin köküne inmek gerekir. Bu işlem ise metni ön bir DDİ işlemine sokmak demektir.

Söz dizimsel özeliklerin yazar tanımada kullanıldığını ilk olarak 1996 yılında görüyoruz (Baayen, Halteren, & Tweedie, 1996). Bu çalışmada her bir cümle için ağaçlar oluşturulmuş ve bu dizilimlerin sıklıkları ölçüm olarak kaydedilmiştir. Var olan çalışmalarda, kelime köklerine göre, kelime tiplerine göre ayrımların kullanıldığını görüyoruz. Çalışmaların sonucunda dizilimsel özelliğin kullanılmasının yazar tanıma işlemini iyileştirildiği görüşmüştür (Gamon & Grey, 2004).

Bir başka çalışmada kelimelerin tiplerinin ardışıklık özellikleri çıkarılmıştır (Baayen et al., 1996). Bu yaklaşım kelimelere daha soyut bir anlayışla bakmaktadır. Algoritma açısından kolaylık sağlasa da metinin ön işlemesi biraz daha karmaşıktır. Fiil, isim, sıfat öbeklerinin sıklık sayıları tutularak çıkarılmaya çalışılan yaklaşımda oldukça etkili sonuçlar alınmıştır. Bir başka çalışma olan Stamatatos'un 2000 yılındaki çalışmasında ise analiz-seviyesi ölçümler kullanılmıştır (E Stamatatos, Fakotakis, & Kokkinakis, 2001). Çalışmada metinin özellikleri birkaç adımda çıkarılmaktadır. İlk adım basit durumları analiz ederken son adım önceki adımların çıktılarından daha karmaşık ölçüler çıkarmaktadır. Burada kullanılan yöntemde daha çok dile özel anlamsal yaklaşım vardır.

Konuşmanın bölümleri (Part-of-speech) yaklaşımıyla ile yapılan bir çalışmada, metinin kısımları etiketlenerek basit bir yaklaşım denenmiştir. Biçim bilimsel yapıdan kelime jetonlarında yüklemiş içeriksel bilgiler kullanılarak kısımlara belirli etiketler verilmeye çalışılmıştır (Argamon-Engelson, Koppel, & Avneri, 1998) (Diederich et al., 2003). Ancak bu çalışmalar kelime dizilimleri değil, anlamsal bir yaklaşımdır.

4.2.1.4 Anlamsal Özellikler

Şimdiye kadar görülen çalışmalarda daha karmaşık metin analizi daha gürültülü ölçümler meydana getirdiğinin anlaşılmış olması gerekir. Doğal dil işleme araçları düşük seviye doğal dil işlerini – konuşmanın kısımlarının başlıklandırılması, metin bölütleme, kısmi bölütleme, cümle bölütleme- oldukça kolay bir şekilde gerçekleştirebilirler. Bu özellikler oldukça rahat bir şekilde ölçülebilir ve metinden çıkartılan özelliklerin gürültüleri düşük seviyede kalır. Bir diğer taraftan daha karmaşık işlemler – tam söz dizimsel bölütleme, anlamsal analiz, fayda analizi- sınırlamasız metinler günümüz DDİ araçlarıyla çok iyi başarılara ulaşamamaktadır. Bundan dolayı konuda başarılabilmiş çalışmalar sınırlı sayıda kalmaktadır.

Gamon'nun 2004 yılında anlamsal bağımlılık grafik yapısını çıkardığı bir çalışması bulunmaktadır fakat kullandığı araçların doğruluk analizini yapamamıştır (Gamon & Grey, 2004). Metinlerden iki tip bilgi çıkarılmıştır: binary anlamsal özellikler ve anlamsal niteleme ilişkileri. Bu konuda ilk adım sayı, insan isimleri, zamanlar ve fiil durumları özellikleri çıkarılıyor. Ardından düğümler arası dizilimsel ve anlamsal ilişkiler bulunmaya çalışıyor. Yapılan çalışmalar gösteriyor ki anlamsal bilgiler kelimesel ve dizilimsel bilgilerle birleşince sınıflama doğruluğunu artırıcı etki gösteriyor.

McCarty ve arkadaşlarının 2006'da yaptığı bir çalışmada anlamsal özellik çıkarımına faklı bir yaklaşım tanımlamıştır (McCarthy, Lewis, Dufty, & McNamara, 2006). WordNet aracına dayanan yaklaşım kelimelerin benzer ve daha soyut üst kelimelerine göre anlamlandırmayı denemişlerdir. Ek olarak kelimelerin benzerliklerinden gizli anlamsal analizini de yapmaya çalışmışlardır. Ancak tam bir model öneremişlerdir. Bu konuda belki de en önemli metodu Argamon 2007 yılında önermiştir (Argamon et al., 2007). Argamon "Systematic Functional Grammar" tekniği kullanarak anlamsal özellikler çıkarmayı denemiştir. Bu çalışmada kelimelerin fonksiyon özellikleri ve faklı anlamları bir araya getirilmiştir. Çalışmanın doğruluğu tam olarak ıspatlanamasa da sınıflandırma doğruluğunu artırdığı görülmüştür.

4.2.2 Özellik Seçimi ve İndirgenmesi

Yazar tanıma işlemlerinde özellik uzayı sık sık birçok faklı özelliği içerisinde barındırır. Faklı özellikler, özellik uzayının boyutunu artırsa da bununla doğru orantılı olarak yazar tanıma doğruluğunu da artırmaktadır. Ancak çoğu durumda özellik uzayındaki her özellik ayırıcılık oranına aynı katkıyı yapmamakta, hatta tarafsız ve negatif ayırıcılık yapan özellikler de bulunabilmektedir. Bunun için özellik seçim algoritmaları ile özellik uzayının boyutu küçültülüp, özellik uzayının etkililiği artırılmaya çalışılır (Forman, 2003).

Genellikle yazarların metin üzerindeki ayırt edici özellikleri seçilmeye çalışılır. Fakat testler yapıldığında seçilen özellikler yazarın stilini tam yansıtmadığı görülmektedir. Destek vektör makinesi algoritması gibi algoritmalar ile seçilen özellikler yazarın stilini daha iyi yansıtmaktadır (Brank, Grobelnik, Milic-Frayling, & Mladenic, 2002). Bir başka çalışmada genetik algoritmalar ile yazarın özellik uzayı ayırt edicilik bakımından düşürülmeye çalışılmıştır (Li, Zheng, & Chen, 2006). Bu gibi çalışmaların

sonunda 270 özellikli bir yazar stili uzayı 134'e indirgenmiş ve sınıflama doğruluğu artmıştır.

Tüm bunların yanında çıkartılan özellikler genellikle içerik bağımlı olduğundan farklı içerikteki metinlerden çıkarılan yazar stilleri farklılaşabilmektedir. Bu bağlamda özellik çıkarma algoritmalarının içerik bağımsız olması daha fazla önerilmektedir. Bu konuda 1964 Mosteller ve Wallace'in çalışmaları incelenebilir (Mosteller & Wallace, 1964). Yapılan çalışmada bağımsız evrensel özelikler belirlenmiştir.

Bir metinden özellik seçilirken en belirleyici ölçüm özelliğin tekrarıdır. Genellikle en çok tekrar eden özellik yazarın en belirleyici stil özelliği olmaktadır. Forsyth 1996 yılında yaptığı bir çalışmada karakter n-gram sıklıkları en fazla olan ile en az olanları alarak bir çalışma yapmıştır (Forsyth & Holmes, 1996). Bu çalışmada sıklığı fazla olan özellikleri aldığı durum diğerine göre daha ayırt edici olmuş ve daha doğru çıkmıştır.

Özellik seçimi üzerine yapılan bir başka çalışmada "istikrarsızlık" durumu kıstas olarak alınmıştır (Moshe Koppel, Akiva, & Dagan, 2006). Bu çalışmada bir metindeki değişmeyen tüm kelimeler –and, or, the v.s- stabil olarak kabul edilmiştir. Buna zıt olarak değişebilir yani benzer anlamı bulunan kelimeler istikrarsız olarak kabul edilmiş ve özellik seti için kullanılmıştır. Bir yazarın metinde benzer anlamlı kelimler arasından hangisini seçtiği bir stil özelliği olarak alınmıştır.

4.2.3 Özellik Metotları

Her yazar tanıma probleminde bir aday yazar seti, bu yazarlara ait yazarların stil özelliklerinin çıkarılacak metin örnekleri (eğitim seti) ve uygulamanın tutarlılığını ve doğruluğunu test edebileceğimiz metin örnekleri (test seti) bulunur. Bu bölümde yazarların sınıflanması yapılırken kullanılan yaklaşımları inceleyeceğiz.

4.2.3.1 Profil Temelli Yaklaşım

Yazar başına düşen erişilebilir eğitim belgelerini bir belgede toplamak bir yöntem olabilir. Bu belge her yazar için bir tane olacaktır ve yazar özellikleri bu belgeden çıkarılacaktır. Daha sonra verilen yazarı bilinmeyen metin belgesi uzaklık ölçümüne göre tahmin edilebilir. Yazarların eğitim belgelerinden oluşturulmuş bir belge oldukça büyük ve yeniden işlenmesi oldukça zor ve gereksiz bir işlem olabilir. Bunun yerine bir kere yazar profilini çıkarıp bunu bir dosyaya veya veritabanına kaydetmek daha uygun bir yöntemdir.

Profil temelli yaklaşım oldukça basit bir eğitim aşaması içerir. Aslında eğitim aşaması sadece yazarların profillerinin çıkarılmasıdır. Ardından verilen metin dosyasının profili çıkarılarak uzaklık durumuna göre sınıflanacaktır.

$$PR\ (x) \rightarrow Profil\ fonksiyonu$$
$$dfun(x, y) \rightarrow Uzaklık\ fonksiyonu$$
$$m_e\ \rightarrow Eğitim\ Metni$$
$$m_t\ \rightarrow Test\ Metni\ Yazar(x) = \min dfun(PR(m_e), PR(m_t))$$

Denklem 4.1: Profil Temmelli Yaklaşım

Burada uzaklıkların nasıl hesaplanacağı ile ilgili birkaç temel yaklaşım bulunmaktadır. Bunlar: olasılıkçı (Probabilistic), sıkıştırma (compression) ve CNG-Varyans metotlarıdır.

4.2.3.1.1 Olasılıkçı Metot

Yazar tanımada hala kullanılan ve en eski yaklaşımlardan biridir. Oldukça fazla modern yazar tanıma uygulaması bu yaklaşım ile problemlerini çözmüştür (Mosteller & Wallace, 1964) (Madigan et al., 2005). Bu yaklaşımda test metininin yazar adayları arasında olasılığı en yüksek olanı metinin yazarı olarak tanır.

$$Yazar(x) = \max log_2 \frac{P(x|a)}{P(x|\bar{a})}$$

Denklem 4.2: Olasılıkçı Yaklaşım

Yazar adayları arasında olasılık hesabı en yüksek olarak hesaplanan yazar sisteme girilmiş olan metin dosyasının yazarı olarak tespit edilmiş olur. Buradaki olasılık hesabı için faklı matematiksel olasılık işlemleri kullanılabilir. Naive bayes olasılık hesabı Peng'in 2004 yılındaki çalışmasında önerilmiş ve oldukça başarılı olmuştur (Peng, Schuurmans, & Wang, 2004). Olasılık hesabı faklı yöntemlerle optimize edilebilmektedir. Bu modelde önemli olan olasılık hesabına sokulacak özelliklerin seçilmesidir.

4.2.3.1.2 Sıkıştırma Metodu

Profil temelli yaklaşımda kullanılan oldukça başarılı olmuş olan sıkıştırma yaklaşımı birçok çalışmada kullanılmıştır (Kukushkina, Polikarpov, & Khmelev, 2001)(Khmelev & Teahan, 2003). Bu metotta yazarlara ait olan eğitim metinleri ayrı ayrı bir büyük dosyada toplanır. Ardından bu büyük dosya sıkıştırma algoritması

kullanılarak benzerlikleri bir yere toplanarak sıkıştırılır. Bu işlem yazarın özellik uzayını küçültür ve daha yoğun bir hale getirir. Aynı işlem test metni için de yapılıp karşılaştırma yapılarak yazar sınıflaması yapılır.

4.2.3.1.3 CNG ve Varyans Metodu

CNG (Common n-Grams) yaklaşımı Keselj tarafından tanımlanmıştır (Kešelj, Peng, Cercone, & Thomas, 2003). Bu yaklaşım yazar profilini oluştururken n-gramlarını çıkararak bunlar arasında en sık olanları alır ve bunları yazarın profili yapar. Yazarı tanıma işlemini oluşturulan profiller ve metin metininden çıkarılan profilin uzaklık hesabını çıkararak yapar.

Yukarıda görülen denklem iki metin arasındaki benzersizliği verir. Benzersizlik hesabı sonucu en düşük çıkan yazar metnin sahibidir.

$$dfun\big(PR(m_e), PR(m_t)\big) = \sum_{g \epsilon PR(m_e) \cup PR(m_t)} \left(\frac{2\left(f_x(g) - f_y(g)\right)}{f_x(g) + f_y(g)} \right)^2$$

$$f \ \rightarrow Frekans$$

Denklem 4.3: Profil Uzaklık Hesabı

Bu yöntemde önemli olan parametreler n-gram uzunluğu, profil oluşturmada alınacak sıklık limitidir. Çeşitli çalışmalarda n-gram sayısının 3 ve alınacak sıklık limitinin 1000 olduğu gösterilmiştir. Ancak kısa metinlerde bu sayılar sıkıntılı olabilmektedir.

4.2.4 Örnek Temelli Yaklaşım

Modern yazar tanıma yaklaşımlarında her bir eğitim metni problemin kendisi olduğu düşünülmektedir. Bir başka deyişle her bir metin çözülmesi gereken bir problemdir. Bu yaklaşımda metinler verilerek doğru yazar sınıflaması için sınıflama algoritması eğitilir.

Sisteme verilen örnek metinlerin çokluğu sınıflama algoritmasının daha iyi eğitilmesi anlamına gelir. Bunun için eğer bir yazarın büyük bir metni varsa bölünerek –yakın boyutlarda- birçok örnek haline getirilir ya da bir yazarın birden fazla metni sisteme eğitim verisi olarak verilir (Sanderson & Guenter, 2006). Genellikle verilen örneklerin boyutları normalize edilir.

Örnek temelli yaklaşımda yazarların nasıl sınıflanacağı ile ilgili olan bazı temel metotlardan biri tercih edilebilir. Bu metotlar; vektör uzayı, benzerlik temelli;

4.2.4.1.1 Vektör Uzayı Metodu

Verilen eğitim metinleri çok değişkenli bir vektör uzayı olarak düşünülür. Sınıflama modelini yapmak için güçlü istatistiksel algoritmalar ve makine öğrenme algoritmaları kullanılır. Bu algoritmalar; ayrıştırma analizi, destek vektör makineleri, kakar ağaçları, yapay sinir ağları.

Vektör uzayının etkinliği sistem kullanıldıkça artmaktadır. Uzaydaki verilerin ağırlıkları değişmekte, duruma özellik değerleri artmakta ya da azalmaktadır (Efstathios Stamatatos, 2008). Bu işlemlerin sonucunda metin örnekleri az olan yazarları dahi tanıyan bir sistem oluşturulabilmektedir.

4.2.4.1.2 Benzerlik Temelli Metot

Benzerlik temelli metodun temel yaklaşımı verilen metin metini eğitim metinlerinden çıkarılan vektör uzayı ile karşılaştırıp benzerlik ölçümlerini almaktır. Buradan çıkan sonuç ile metin metnini en çok benzeyen yazarın sınıfına atacaktır. Bu işlemi yaparken en yaygın kullanılan algoritma en yakın komşu algoritmasıdır. Ek olarak başka algoritmalar da kullanılabilir.

5 KİTAPLARDA YAZAR TANIMA

Bu tez kapsamında Türkçe kitapların yazarlarını tanıma işlemini bilgisayar üzerinde çalışan bir program yardımıyla gerçekleştirmeye çalıştık. Türkçe Metinlerde yazar tanıma işlemini yaparken dil ve içerik bağımsız bir çalışma yürütülmüştür. Çeşitli deneyler sonucunda en avantajlı ve çözülebilir bir yöntem olan n-gram ve naive bayes sınıflama algoritması kullanılmıştır. Bu bölümde tez çalışmasını derinlemesine incelemesini okuyacaksınız.

5.1 Materyal

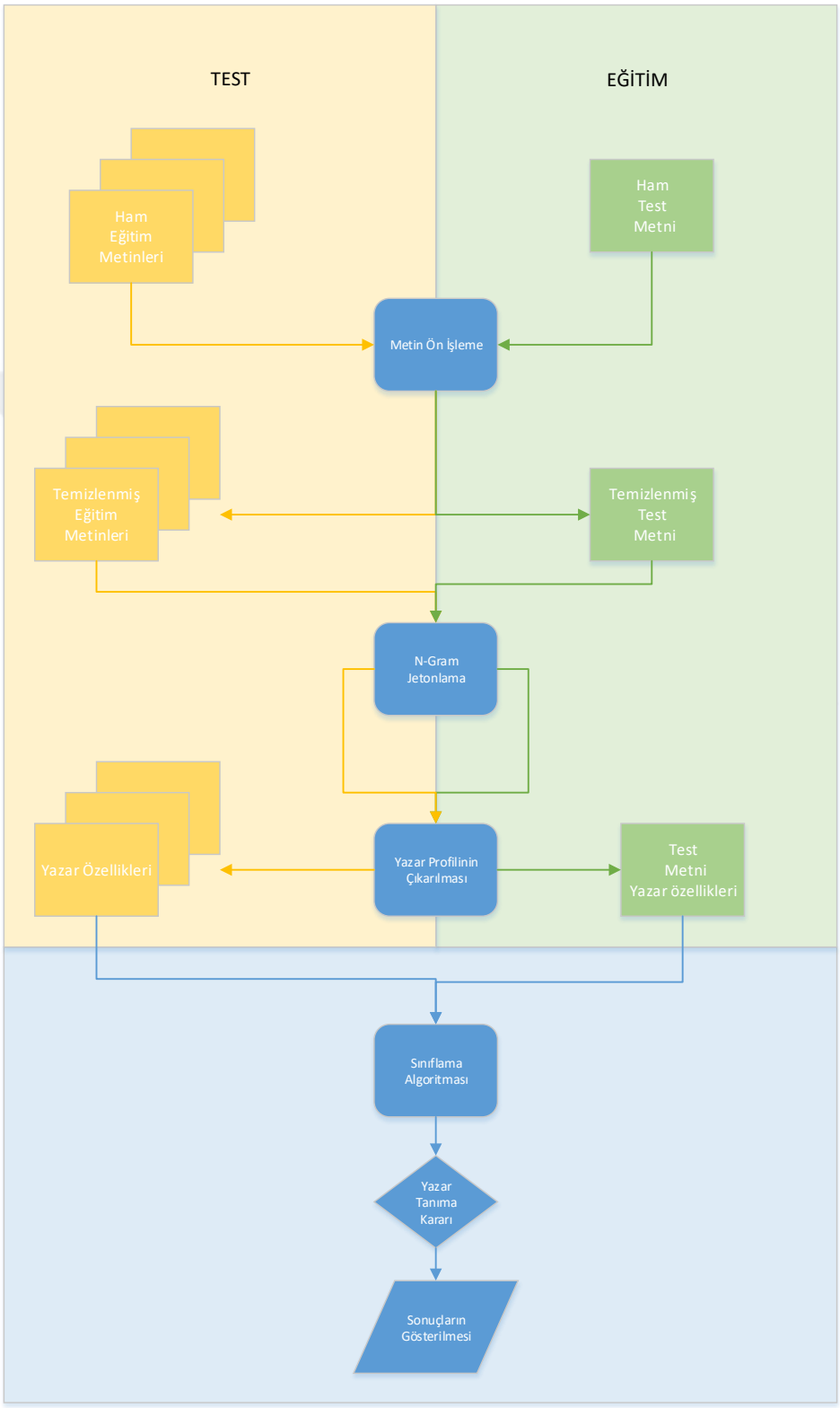
Tez kapsamında gerçekleştirilecek olan program için bir yazılım dili araştırması sonucunda birkaç seçenek üzerinde çalışılmıştır. Bu seçenekler Matlab, Python, Java, C# olarak belirlenmiştir. Bunlar arasında yazılımı gerçekleştirmek üzere sahip olduğu DDİ, matematik kütüphaneleri ve çalışabilir program haline getirilebilmesi ile ön plana çıkan Python 3.0 programlama dili kullanılmıştır. Python programlama dili ile birlikte jetonlama ve n-gram çıkarımında NLTK (Natural Language Toolkit) kütüphanesinden faydalanılmıştır.

Program Intellij Idea firmasının Python editörü olan Pycharm idesi üzerinde geliştirilmiştir ve Python 3.0 kurulu olan her bilgisayarda çalışabilmektedir.

5.2 Metodoloji

Tez çalışmasında yapılmış olan uygulama da ilk önce 120 adet Türkçe kitap bulunmuştur. Bulunan kitaplar pdf formatındadır. Pdf formatı işlenebilir durumda olmadığından, pdf içerisindeki bilgiler txt gormatına çekilmiştir. Çevrim işleminde bazı veriler bozulmuştur. Hem bozulan verileri düzeltmek hem de gereksiz ve işe yaramaz bilgileri metin içerisinden atmak için bu kitapları bir metin ön işleme fonksiyonundan geçirilip temizlenmiştir. Ardından rasgele 20 yazarın 20 adet kitabı seçilmiş ve eğitim verisi olarak kullanılmıştır ve geri kalan 100 kitap test verisi olarak

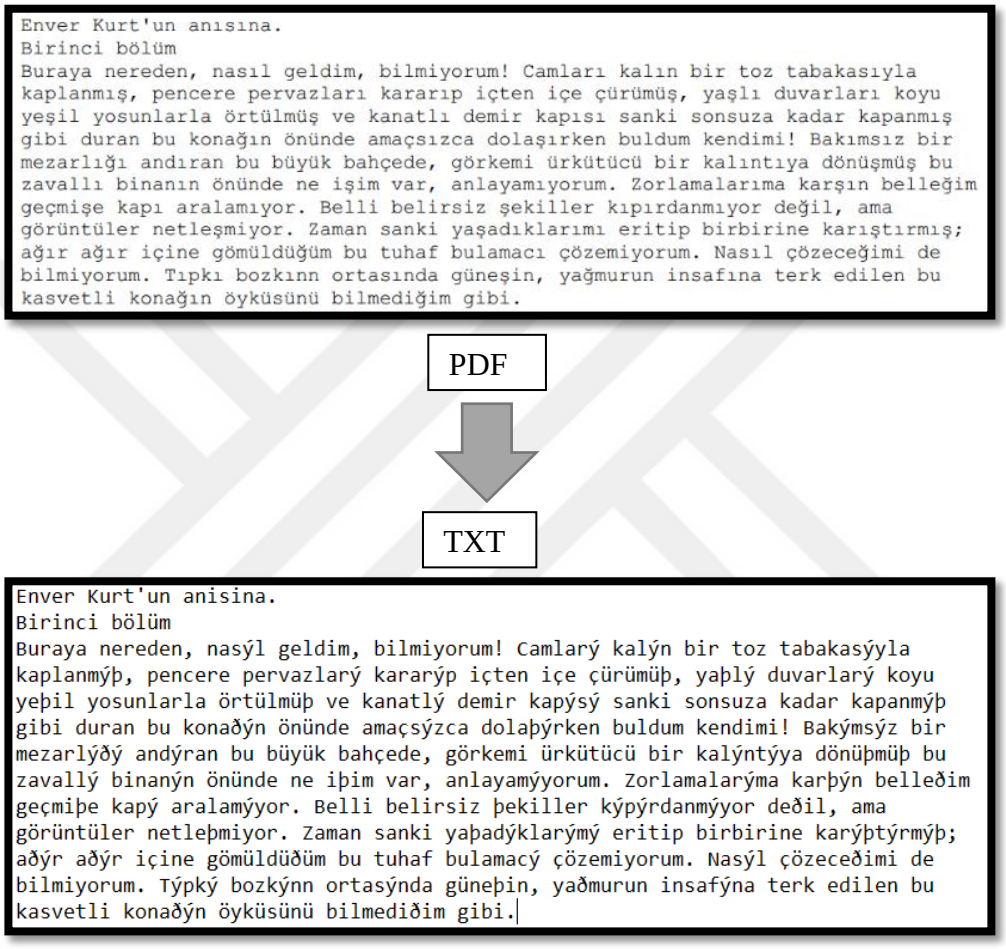
alınmıştır. Test kitapları eğitilmiş sisteme sokulup naive bayes sınıflayıcı fonksiyonunda çıktılarının sonuçları gözlemlenmiştir. Tez boyunca izlenmiş olan işlem adımları ***Şekil 4.1*** üzerinde gösterilmiştir.



Şekil 4.1: İşlem Adımları Diyagramı

5.2.1 Metin Ön İşleme

Elektronik ortamda yazarların birçok metini bulunmaktadır. Bu metinler çeşitli formatlarda olabilmektedir. Bu tez kapsamında bulunmuş olan metinler pdf formatındadır. Pdf dosyalarının yazılımla okunması zor olduğundan pdf formatında olan metinler manuel olarak txt formatına çevrilmiştir. Txt formatına çevrilme işleminden sonra gelen işlemlerin hepsi otomatik olarak gerçekleştirilmektedir.



Şekil 5.1 : Pdf Metin Dönüşümü

Şekil 5.1 Pdf dosyasının txt dosyasına çevrim işlemini görtermektedir. Manuel çevrim işleminde içindekiler gibi kısımlar, son yazılar vesaire oluşturulan txt metni içerisinde alınmamıştır. **Şekil 5.1**'de de görüldüğü gibi bu çevrim işleminden sonra yazıda karakter bozulmaları oluşmaktadır. Bu karakterler DDİ yaparken bizim sistemimizin hata yapmasına neden olmaktadır, hatalı çevrilen karakterler tekrar düzeltilmelidir. Buna ek olarak sistemimize sokacağımız metin dosyasında boşluk, noktalama işaretleri, özel karakterleri, sık geçen bağlaç, edatları, sayıları da metnin içerisinden silmemiz gerekmektedir.

Anlamsız karakterler metnin içerisinden çıkarılmaktadır. Anlamsız karakterler yazar özelliği olmadığı gibi çeşitli nedenlerden –karakter setleri, programsal özellikleri- oluşan durumlardır.

Sayılar genellikle sıklık açısından düşük sıralamalarda kalmakta ve yazar özellikleri açısından da bize pek bir bilgi vermemektedir. Bu yüzden sayıları metin içerisinden siliyoruz.

Bir başka yazar stili olabilecek özellik noktalama işaretleri. Noktalama işaretlerinin hepsi tek karakterden oluşmaz. Burada yapılabilecek işlem noktalama işaretini tekil bir karakterle değiştirmek ya da silmektir. Silme mantığında noktalama işaretlerinin n-gram sistemine göre güçsüz kalmasıdır.

Bir başka sorun dilde çok geçen yani herkes tarafından çok kullanılan terimlerdir.

Şekil 5.2: Metin Ön İşleme

Bağlaçlar bunlara örnek gösterilebilir. Bundan dolayı sıklık açısından her yazarda üst sıralarda yer almakta ayırıcılığı düşürmektedir. Bu durumu aşmak için iki yol

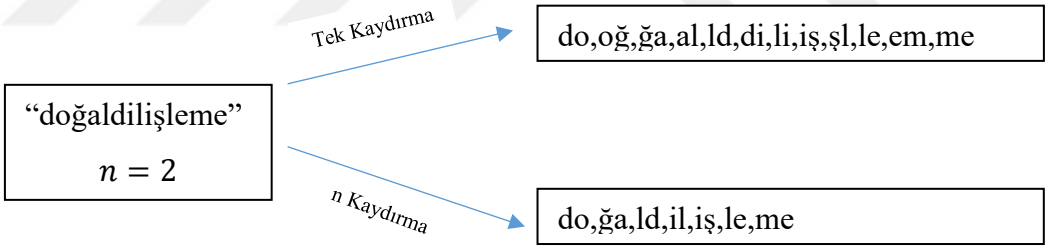
bulunmaktadır. Sık geçen terimleri tekbir tekil karakterle değiştirmek ya da yazıdan silmek.

5.2.2 Jetonlama

Doğal dil işleme için verilen metin belgesi ön işlemeden geçip temizlendikten sonra dilsel olarak işlenebilecek şekilde jetonlama işlemine geçilir. Jetonlama işlemi yapılacak DDİ işlemine göre değişiklik gösterebilir. Çoğu DDİ çalışmasında kelimeler jetonlama için kullanılmaktadır. Bununla birlikte cümleler, kelime ekleri, sadece harfler-noktalamalar jeton olarak kabul edilebilir. Bu çalışmada yapılacak olan jetonlama işlemi n-gram çıkarımıdır.

5.2.2.1 N-Gram Jetonlama

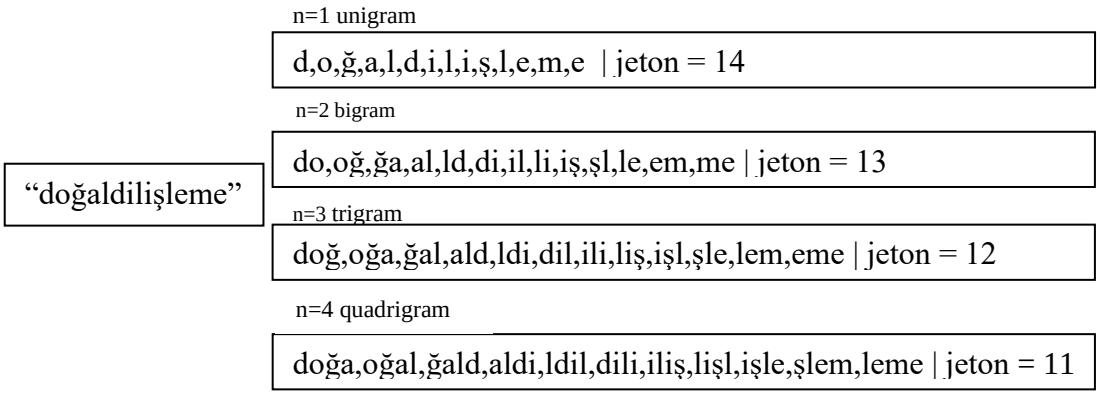
N-Gram yaklaşımı tüm metnin belirli büyüklükte bölütlerine ayrılmasıdır. Bu büyüklük değeri "n" ile ifade edilir. Tüm metin "n" büyüklüğünde parçalara ayrılır. Burada iki türlü n-gram jetonlama vardır. Birincisi teker teker kaydırma ile yapılan n-gram, ikincisi n kadar kaydırmalı n-gram yaklaşımıdır. **Şekil 5.3**' de tekli kaydırma ve n kadar kaydırma işlemi n=2'ye göre örneklenmiştir.



Şekil 5.3 : N-Gram Jetonlama

Şekil 5.3'de görülebileceği gibi tek kaydırma işleminde çok jeton çıkmakta n kadar kaydırma işleminde ise daha az jeton çıkmaktadır. N sayısı büyüdükçe aradaki fark açılmaktadır. Uzun metin belgelerinde aradaki fark oldukça açılmaktadır. DDİ açısından iki metodun da kullanılabildikleri yerler vardır. Ancak yazar özellikleri çıkarmak açısından tek kaydırma metodu daha uygun olmaktadır. N kaydırma metodunda bir özellik karakter kayması nedeniyle kaçırılmaktadır. Tekli kaydırma çok fazla jeton çıkarsa da bilgisayarlar açısından bu bir sorun oluşturmamaktadır.

N-gram yönteminde n yerine gelecek sayı ile birlikte gramlara verilen isimler değişmektedir; uni-gram, bi-gram, tri-gram, quadri-gram. **Şekil 5.4** n-gram jetonlama örnekleri verilmiştir.



Şekil 5.4 : N-Gram Jetonlama Örneği

Şekil 5.4'te örneklendiği üzere n-gram jetonlamada n uzunluğuna göre avantajlı ve dezavantajlı durumlar bulunmaktadır. Her şeyden önce n sayısı ile jeton sayısı ters orantılıdır. Bir başka deyişle n sayısı artıkça jeton sayısı düşmektedir. Kısa bir metin için çok farkı yokmuş gibi görünse de uzun metinlerde jeton sayısının azlığı sıklık hesabı ve karşılaştırma hızında fayda saylayacaktır. N sayısının büyük olup jeton sayısının azlığı bilgisayar hafızasında(RAM) daha az yer kaplayacağı anlamına gelmektedir, bunu bir avantaj olarak düşünebiliriz. Bunun yanında jeton sayısının fazlalığı seçim için yazar özelliği fazlalığı demektir. Tüm bu söylenenlerin yanında tavsiye edilen n sayısı 2, 3, 4 olmaktadır. Bi-gram sadece harf özelliği olacağından pek bir ayırıcılık özelliği olmaz. Bu tez kapsamımda 2, 3, 4 n-gramlar ile test işlemleri yapılmıştır.

5.2.3 Yazar Stil Özelliği Vektör Uzaylarının Çıkarımı

Yazar sitilleri araştırmaları kısmında anlamsal, biçimsel özellikler olarak metin üzerinde birçok şeyin yazar özelliği olabileceği anlatılmıştır. Buradaki iki kısım – anlamsal, biçimsel- özelliklerin çıkarım zorlukları birbirinden farklıdır.

Anlamsal yazar özellikleri yazarın kastettiği anlamlar, verdiği örnekler, söylem biçimi hatta fikrinin çıkarımına kadar gidebilmektedir. Sezileceği gibi bunu yapmak oldukça zordur. Metinin kelimesel, cümlesel, kısımsal analizlerinin yapılıp anlamsal çıkarımlarda bulunabilecek sistemler üzerinde hala çalışılmaktadır fakat tam bir başarı ortaya konulabilmiş değildir. Her şeyden önce anlamsal analiz bir sözlük

gerektirmektedir. Bu sözlükle kelimelerin, cümlelerin faklı anlamları çıkarılacak ve hatta kelimelerin köklerine inilip söylem analizi yapılmasına imkân sağlayacaktır. Bunun için birkaç Türkçe sözlük çalışması bulunsa da bizim burada kullandığımız yöntem anlamsal çözümleme olmayacaktır.

Biçimsel özellikler dilden bağımsız olarak yazarın yazıda kullandığı kelimeler, kelime sıraları, noktalama işaretleri, n-gram dizileri vesaire olabilir. Bunların hepsi ya da bir kaçı ile yazar tanıma işlemi test edilebilir. Ne kadar farklı özellik bir arada kullanılırsa ayırt edicilik artacaktır. Bu tezde n-gram sıklıklarını yazar özelliği olarak metinden çıkardık ve tanıma işlemi için kullandık.

N-gram olarak jetonlanmış metnimizde, bu n-gramların sıklık hesabını yapıyoruz. Bir yazar için n-gram vektör uzayında terim sıklığı ve metin uzunluğuna göre terim sıklığı bölütünü hesaplatıyoruz. Bu hesaptan sonra terimlerin sıklıklarına göre bir sıralamaya tabi tutulmaktadır. Burada en az tekrar eden ya da en fazla tekrar eden terimleri ayırt edicilik için kullanabiliriz. Bu tezde en fazla tekrar eden terim ve sıklık benzerliklerinden yararlanarak yazar tanıma işlemini yapıldı. **Şekil 5.5** ön işlemeden geçirilmiş bir metinden özellik çıkarılmış hali gösterilmiştir. Şekilde çıkarılmış özelliklerin kısa bir özeti gösterilmiştir.

Şekil 5.5'te gösterilen şekilde Türkçe karakterlerin İngilizce karakterlere çevrilmiş halini görmemizin nedeni: şekil Pycharm editöründen alınmıştır ve pycharm editöründen kaynaklı öyle görünmetedir. Yani veride herhangi bir hata yoktur.

Şekil 5.5 : Yazar N-gram Sıklık Profil Uzayı

5.2.3.1.1 Yazar Vektör Uzaylarının Karşılaştırması ve Sınıflama

Metinden çıkarmış olduğumuz yazar özelliği ve eğitim metinlerinden çıkarmış olduğumuz yazar özellikleri vektör uzaylarından en yakın olduğu ile eşleştirip test metninin yazarını tespit etme işleminin yapması gerekmektedir. Bu işlemi yapabilmek için birçok sınıflama kümeleme algoritması bulunmaktadır. Bu algoritmaların yerine göre kullanım alanları ve amaçları vardır. Bu tez kapsamında naive bayes ile sınıflama işlemi yapılmıştır.

5.2.3.2 Naive Bayes Sınıflama Algoritması

Naive Bayes: Makine öğrenmesinde özellikler arasındaki bağımsızlık varsayımlarını kullanarak bayes teoremini uygulayan basit olasılıkçı bir sınıflandırma algoritmasıdır.

Naive bayes 1950'li yıllarda çalışılmaya başlanmıştır. Sonraki yıllarda metin kategorizasyonu işlemlerinde kullanılmaya başlanmıştır (Raschka, 2014).

Naive bayes teoremi bir özelliğin diğeriyle bağımsız olduğunu varsayımından yola çıkmaktadır. Özelliklerin bağımsız olasılıkları hesaplanıp çarpılarak bu olasılıkların bir araya gelme durumundaki sonuçları bize vermektedir. Naive bayes bütün veri setleri üzerine kolaylıkla uygulanabilmektedir (Raschka, 2014). **Denklem 5.1** Naive bayes formülünü göstermektedir. **Denklem 5.2** Naive bayes teoremini örneklemektedir.

$P(X|C_i) \rightarrow Özelliğin\ Bağımsız\ Olasılığı$

$P(C_i) \rightarrow Tüm\ olasılık$

$P(C_i|X) \rightarrow Olsılık\ Sonuç$

$$P(C_i|X) = \frac{P(X|C_i)\,P(C_i)}{P(X)}$$

$İndirgenmiş\ Formül:\ P(C_i|X) = P(X|C_i)\,P(C_i)$

Denklem 5.1: Naive Bayes Sınıflandırma

Bir sonraki sayfadaki **Denklem 5.2** naive bayes hesaplamasını bir satış örneği üzerinde göstermektedir. Bu tez kapsamında metinin sınıflandırması n-gram sıklıklarının alınması ve sonra test metninin alınmış n-gram sıklıklarıyla naive bayes hesabına sokularak olasılığı yüksek olan yazara test metninin atanması olarak gerçekleşir.

Araba	Renk	Marka	Satış
1	Mavi	Renault	Satıldı
2	Kırmızı	Mercedes	Bekliyor
3	Mavi	Renault	Satıldı
4	Mavi	Mercedes	Bekliyor
5	Kırmızı	Mercedes	Satıldı
6	Mavi	Renault	Bekliyor
7	Kırmızı	Mercedes	Satıldı
8	Kırmızı	Renault	Satıldı
9	Kırmızı	Renault	Satıldı
10	Mavi	Mercedes	Satıldı

P(Satıldı)	$^{7}/_{10}$
P(Bekliyor)	$^{3}/_{10}$

P(Kırmızı|Satıldı)	$^{4}/_{7}$	P(Kırmızı|Bekliyor)	$^{1}/_{3}$
P(Mavi|Satıldı)	$^{3}/_{7}$	P(Mavi| Bekliyor)	$^{2}/_{3}$

P(Mercedes|Satıldı)	$^{3}/_{7}$	P(Mercedes |Bekliyor)	$^{2}/_{3}$
P(Renault |Satıldı)	$^{4}/_{7}$	P(Renault | Bekliyor)	$^{1}/_{3}$

<Kırmızı, Mercedes> olasılıkları

$$P(Satıldı|X) = P(Kırmızı|Satıldı)xP(Mercedes|Satıldı)x\ P(Satıldı)$$

$$P(C_i|X) = {}^{4}/_{7}\ x\ {}^{3}/_{7}\ x\ {}^{7}/_{10} = 0{,}1714$$

$$P(Bekliyor|X) = P(Kırmızı|Bekliyor)xP(Mercedes|Bekliyor)x\ P(Bekliyor)$$

$$P(C_i|X) = {}^{1}/_{3} \times {}^{2}/_{3} \times {}^{3}/_{10} = 0{,}0666$$

<Kırmızı, Mercedes>

Satıldı= 0,1714

Bekliyor= 0,066

Satıldı > Bekliyor

Satılma olasılığı fazla.

Denklem 5.2: Naive Bayes Sınıflandırıcı Örnek

5.3 Deneysel Çalışma

Yapılan deneysel çalışmada yirmi adet yazar ve bu yazarların faklı oranlarda yüz adet kitabı kullanılmıştır. Yazarın özellik çıkarıldığı eğitim kitabı ve test kitapları tamamen rasgele seçilmiştir. Yazarın özellik uzayı tek bir kitabından çıkarılmıştır. Rasgele verilmiş olan yazar kitabının n-gram özellikleri çıkarılıp sıralanmış ve en yüksek olan bin adet veri yazar özelliği olarak alınmıştır.

Sisteme verilen bir kitap yazar eğitimindeki gibi işlemlerden geçerek n-gramları ile yazar özellik vektörü çıkarıldıktan sonra naive bayes algoritmasına sokularak tanıma işlemi yapılmaktadır.

Bu çalışmada iki defa deney yapılacaktır. İlkinde yirmi yazarın yirmi kitabı rasgele seçilecek sisteme yazar özelliği olarak verilecek ve sonuçlar incelenecektir. Ardından aynı işlem yirmi yazarın önceki deneye göre farklı yirmi kitabı ile yazar özellik uzayı çıkarılacak ve sonuçlar incelenip önceki işlem ile karşılaştırılacaktır.

Şekil 5.6 yazar özellik vektörlerinin çıkarıldığı kitaplar ve karakter uzunlukları verilmiştir.

Yazar Adı - Kitap	Uzunluk
ahmet hamdi tanpınar - huzur.txt	731420
ahmet ümit - babı esrar.txt	686573
aziz nesin - bay duduk.txt	120618
cemil meriç - umrandan uygarliga.txt	528054
doğan cüceloğlu - savaşcı.txt	747830
elif şafak - aşk.txt	620628
fakir baykurt - yılanların öcü.txt	456147
hüseyin nihal adsız - ruhadam.txt	406820
hüseyin rahmi gurpınar - aşk batağı.txt	318109
ilber ortaylı - ikinci aldulhamit döneminde alman nüfuzu.txt	314087
kemal tahir - bozkırdaki çiçekler.txt	706356
necip fazıl kısakürek - aynadaki yalan.txt	274971
ömer seyfettin - terakki.txt	143896
orhan kemal - bereketli topraklar üzerine.txt	504844
orhan pamuk - kar.txt	808656
peyami safa - yalnızız.txt	571449
sebahattin ali - kurk mantolu madonna.txt	296786
soner yalcın - bu dinciler o müslümanlara benzemiyor.txt	905048
ugur mumcu - cagın sucu.txt	454419

Şekil 5.6 : Eğitim Kitapları

ahmet ümit - beyoğlu rapsodisi.txt	723706
ahmet ümit - bir ses böler geceyi.txt	276699
ahmet ümit - kar kokusu.txt	475323
ahmet ümit - kavim.txt	685874
ahmet ümit - kukla.txt	933559
ahmet ümit - ninattanin bileziği.txt	82261
ahmet ümit - patasana.txt	698271
ahmet ümit - seytan ayrintida gizlidir.txt	308894
ahmet ümit - sis ve gece.txt	442780
attilla ilhan - bıçağın ucu.txt	783962
aziz nesin - bay duduk.txt	135547
aziz nesin - borçlu olduklarimiz.txt	80817
aziz nesin - damda deli var.txt	121592
aziz nesin - gerçeğin masalı.txt	159470
aziz nesin - kazan toreni.txt	207599
aziz nesin - memleketin birinde.txt	169482
aziz nesin - nihat bey neden kasınıyor.txt	210207
aziz nesin - simdiki cocuklar harika.txt	221048
aziz nesin - tatli betus.txt	594227
aziz nesin - zübüklüğün sonu yok.txt	167157
cemil meriç - bir facianın hikayesi.txt	179184
cemil meriç - jurnal 1.txt	595483
cemil meriç - jurnal 2.txt	474074
doğan cüceloğlu - insan insana.txt	469947
doğan cüceloğlu - yetişkin çocuklar.txt	409897

elif şafak - araf.txt	585198
elif şafak - baba ve piç.txt	641156
elif şafak - bit palas.txt	172153
elif şafak - mahrem.txt	452703
elif şafak - pinhan.txt	396355
elif şafak - siyah süt.txt	331110
fakir baykurt - keklik.txt	572069
fakir baykurt - köygöçüren.txt	1058937
hüseyin nihal adsız - bozkurtlar diriliyor.txt	270462
hüseyin nihal adsız - bozkurtların ölümü.txt	552106
hüseyin nihal adsız - delikurt.txt	295639
hüseyin nihal adsız - makaleler.txt	510300
hüseyin rahmi gurpınar - dirilen iskelet.txt	458619
hüseyin rahmi gurpınar - gulyabani.txt	220337
hüseyin rahmi gurpınar - muhabbet tılsımı.txt	348326
hüseyin rahmi gurpınar - nimetşinas.txt	247309
ilber ortaylı - gelenekten geleceğe.txt	284099
ilber ortaylı - osmanlı barışı.txt	382840
ilber ortaylı - osmanlı devletinde kadı.txt	131625
ilber ortaylı - tarihin izinde.txt	188073
ilber ortaylı - teşkilat ve idare tarihi.txt	977640
kemal tahir - büyük mal.txt	761905
kemal tahir - devlet ana.txt	1013786
kemal tahir - namuscular.txt	646429
kemal tahir - sağırdere.txt	427444

necip fazıl kısakürek - babıali.txt	466619
necip fazıl kısakürek - çöle inen nur.txt	562360
necip fazıl kısakürek - dünya bir inkilap bekliyor.txt	168506
necip fazıl kısakürek - moskof.txt	504927
necip fazıl kısakürek - peygamber halkası.txt	277746
necip fazıl kısakürek - rapor 1.txt	142340
necip fazıl kısakürek - rapor 2.txt	123998
necip fazıl kısakürek - tarih boyunca büyük mazlumlar.txt	855621
necip fazıl kısakürek - tasavvuf bahçeleri.txt	197297
necip fazıl kısakürek - vahudiddin.txt	368143
necip fazıl kısakürek - yeniçeri.txt	481893
ömer seyfettin - ant.txt	78012
ömer seyfettin - asilzadeler.txt	341528
ömer seyfettin - düşünme zamanı.txt	16813
ömer seyfettin - harem.txt	194139
ömer seyfettin - inat.txt	26879
ömer seyfettin - külah.txt	43087
ömer seyfettin - üç öğüt.txt	56628
orhan kemal - baba evi 2.txt	146187
orhan kemal - baba evi.txt	147644
orhan kemal - çamaşırcının kızı.txt	163456
orhan kemal - cemile.txt	203798
orhan kemal - dünya evi.txt	364937
orhan kemal - ekmek kavgası.txt	189512
orhan kemal - eskici dükkanı.txt	529654

orhan kemal - gurbet kuşları.txt	510973
orhan kemal - hanımın çiftliği.txt	510525
orhan kemal - murtaza.txt	487839
orhan kemal - vukuat var.txt	590978
orhan pamuk - beyaz kale.txt	278049
orhan pamuk - kara kitap.txt	886785
orhan pamuk - kmasumiyet müzesi.txt	1065806
orhan pamuk - sessiz ev.txt	532217
orhan pamuk - yeni hayat.txt	438438
peyami safa - dokuzuncu hariciye koğuşu.txt	132037
peyami safa - selma ve gölgesi.txt	261414
soner yalcın - beco.txt	421975
soner yalcin - beyaz muslumanlarin buyuk sırrı.txt	463656
soner yalcın - bu dinciler o müslümanlara benzemiyor.txt	905169
soner yalcın - erseverin itirafları.txt	372358
soner yalcın - pipo.txt	947270
soner yalcın - reis.txt	647299
soner yalcın - siz kimi kandırıyorsunuz.txt	621476
soner yalcin - teskilatin iki silahsoru.txt	499920
ugur mumcu - cagın sucu.txt	455563
ugur mumcu - devlet silah adalet.txt	372598
ugur mumcu - sakıncalı piyade.txt	167226
uğur mumcu - yolsuzluk şiddet bağımlılık.txt	407516
ahmet hamdi tanpınar - beş şehir.txt	371233
ahmet ümit - ask köpekliktir.txt	350392

Şekil 5.7 : Tets Kitapları

Yazarların özelliklerinin çıkarıldığı eğitim metinleri gibi toplam yüz adet olan test metinleri de aynı uzunluk ve özelliklere sahiptir. **Şekil 5.7** test kitapları isimleri ve boşluksuz karakter uzunlukları verilmiştir.

Deney 1'de işleme sokulan yüz kitabın sonuçları **Çizelge 5.1:** Özet Sonuçlar'da verilmiştir(arada iki örnek kitap gösterim için koyulmuştur). Sistemin çalışmış hali ekler kısmında ve Deney çıktılarının tamamı **EK A**'da verilmiştir. Bu bölümde yapılan değerlendirme formülleri 3.2.2 Değerlendirme bölümünde ***Denklem* 3.4 :** Hassasiyet ve Kesinlik Hesabı, **Denklem *3.3* :** Doğruluk Hesabı, **Denklem *3.5* :** F-Score Hesabı formüllerinde verilmiştir.

Çizelge 5.1: Özet Sonuçlar

KİTAP	YAZAR	Bi-Gram		Tri-Gram		Quad-Gram	
		TAHMİN	S	TAHMİN	S	TAHMİN	S
ask köpekliktir	a. ümit	ö. seyfettin	F	a. ümit	T	a. ümit	T
üç öğüt	ö. seyfettin	ö. seyfettin	T	ö. seyfettin	T	ö. seyfettin	T
Toplam Doğru Sayısı			12		71		82
Toplam Yanlış Sayısı			88		29		18

$$Bi - Gram$$

$$Doğruluk\ (Accuracy) = \frac{12}{100} = 0.12$$

$$Hata\ Oranı\ (Error\ Rate) = 1 - 0.12 = 0.88$$

$$Tri - Gram$$

$$Doğruluk\ (Accuracy) = \frac{71}{100} = 0.71$$

$$Hata\ Oranı\ \left(Error\ Rate\right) = 1 - 0.71 = 0.29$$

$$Quad - Gram$$

$$Doğruluk\ (Accuracy) = \frac{82}{100} = 0.82$$

$$Hata\ Oranı\ \left(Error\ Rate\right) = 1 - 0.82 = 0.18$$

Denklem 5.3: Doğruluk ve Hata Oranı Hesapları

Çizelge 5.2: Bi-Gram Karmaşıklık Matrisi

Bi-Gram		DOĞRU ETİKET																				Toplam	Precision
		a. kulin	a. ümit	r. n. güntekin	a. nesin	c. meriç	d. cüceloğlu	e. şafak	f. baykurt	h. n. adsız	h. r. gurpınar	i. ortaylı	k. tahir	n. f. kısakürek	ö. seyfettin	o. kemal	o. pamuk	p. safa	s. ali	s. yalcın	u. mumcu	tanımsız		
TAHMİN ETİKET	a. kulin	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	a. ümit	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1
	r. n. güntekin	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	a. nesin	0	0	0	4	0	0	0	2	0	2	0	2	3	1	6	0	0	0	0	1	0	21	0.19
	c. meriç	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	d. cüceloğlu	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	e. şafak	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	f. baykurt	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	h. n. adsız	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	h. r. gurpınar	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	i. ortaylı	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	k. tahir	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	n. f. kısakürek	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	ö. seyfettin	4	9	5	4	3	2	6	0	4	2	3	2	6	6	3	5	4	0	6	2	0	76	0.079
	o. kemal	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	o. pamuk	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	p. safa	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	s. ali	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0
	s. yalcın	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	u. mumcu	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	1
	tanımsız	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Toplam		4	10	5	8	3	2	6	2	4	4	3	4	9	7	9	5	4	0	7	4	0	100	0.108
Recall		0	0.1	0	0.5	0	0	0	0	0	0	0	0	0	0.857	0	0	0	0	0	0.25	0	0.081	

Çizelge 5.3: Tri-Gram Karmaşıklık Matrisi

Tri-Gram		DOĞRU ETİKET																				Toplam	Precision
		a. kulin	a. ümit	r. n. güntekin	a. nesin	c. meriç	d. cüceloğlu	e. şafak	f. baykurt	h. n. adsız	h. r. gurpınar	i. ortaylı	k. tahir	n. f. kısakürek	ö. seyfettin	o. kemal	o. pamuk	p. safa	s. ali	s. yalcın	u. mumcu	tanımsız		
TAHMİN ETİKET	a. kulin	4	0	0	1	0	0	0	0	0	0	0	0	0	1	4	1	0	0	0	0	0	11	0.364
	a. ümit	0	9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	9	1
	r. n. güntekin	0	0	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5	1
	a. nesin	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	1
	c. meriç	0	0	0	0	1	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	3	0.333
	d. cüceloğlu	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	1
	e. şafak	0	0	0	2	0	0	6	0	0	0	0	1	0	0	2	0	0	0	0	0	0	11	0.545
	f. baykurt	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1
	h. n. adsız	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	2	1
	h. r. gurpınar	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	3	1
	i. ortaylı	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	k. tahir	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	3	1
	n. f. kısakürek	0	0	0	1	2	0	0	0	0	1	0	0	9	0	0	0	0	0	0	0	0	13	0.692
	ö. seyfettin	0	0	0	0	0	0	0	0	0	0	0	0	0	6	0	0	0	0	0	0	0	6	1
	o. kemal	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	3	1
	o. pamuk	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	3	1
	p. safa	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	2	1
	s. ali	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	2	0
	s. yalcın	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	7	0	0	9	0.778
	u. mumcu	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	0	4	1
	tanımsız	0	1	0	2	0	0	0	1	1	0	0	0	0	0	0	1	0	0	0	0	0	6	0
Toplam		4	10	5	8	3	2	6	2	4	4	3	4	9	7	9	5	4	0	7	4	0	100	0.748
Recall		1	0.9	1	0.25	0.333	1	1	0.5	0.5	0.75	0	0.75	1	0.857	0.333	0.6	0.5	0	1	1	0	0.632	

Çizelge 5.4: Quadri-Gram Karmaşıklık Matrisi

Quadri-Gram		DOĞRU ETİKET																				Toplam	Precision
		a. kulin	a. ümit	r. n. güntekin	a. nesin	c. meriç	d. cüceloğlu	e. şafak	f. baykurt	h. n. adsız	h. r. gurpınar	i. ortaylı	k. tahir	n. f. kısakürek	ö. seyfettin	o. kemal	o. pamuk	p. safa	s. ali	s. yalcın	u. mumcu	tanımsız		
TAHMİN ETİKET	a. kulin	4	0	0	1	0	0	0	0	0	0	0	1	0	1	0	1	0	0	0	0	0	8	0.5
	a. ümit	0	10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	10	1
	r. n. güntekin	0	0	5	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	6	0.833
	a. nesin	0	0	0	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	1
	c. meriç	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	2	0.5
	d. cüceloğlu	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	1
	e. şafak	0	0	0	1	1	0	6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	8	0.75
	f. baykurt	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1
	h. n. adsız	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	3	1
	h. r. gurpınar	0	0	0	0	0	0	0	0	0	4	0	0	0	0	0	0	0	0	0	0	0	4	1
	i. ortaylı	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	k. tahir	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	3	1
	n. f. kısakürek	0	0	0	0	1	0	0	0	0	0	0	0	9	0	0	0	0	0	0	0	0	10	0.9
	ö. seyfettin	0	0	0	0	0	0	0	0	0	0	0	0	0	6	0	0	0	0	0	0	0	6	1
	o. kemal	0	0	0	0	0	0	0	0	0	0	0	0	0	0	7	0	0	0	0	0	0	7	1
	o. pamuk	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	3	1
	p. safa	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	3	1
	s. ali	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	s. yalcın	0	0	0	1	0	0	0	0	1	0	2	0	0	0	0	0	0	0	7	0	0	11	0.636
	u. mumcu	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	0	4	1
	tanımsız	0	0	0	1	0	0	0	1	0	0	0	0	0	0	2	1	0	0	0	0	0	5	0
Toplam		4	10	5	8	3	2	6	2	4	4	3	4	9	7	9	5	4	0	7	4	0	100	0.768
Recall		1	1	1	0.5	0.333	1	1	0.5	0.75	1	0	0.75	1	0.857	0.778	0.6	0.75	0	1	1	0	0.706	

$$Bi - Gram$$

$$f - score = \frac{2}{\frac{1}{0.0813} + \frac{1}{0.108068}} = 0.09279211271$$

$$Tri - Gram$$

$$f - score = \frac{2}{\frac{1}{0.6321} + \frac{1}{0.748215}} = 0.68527358103$$

$$Quad - Gram$$

$$f - score = \frac{2}{\frac{1}{0.7056} + \frac{1}{0.767605}} = 0.73529765104$$

Denklem 5-4: F-Score Değerleri

Çizelge 5.1'de yansıtılan sonuçlara göre sisteme sokulan yüz kitaptan bi-gram sonucuna göre 88 yanlış 12 doğru, tri-gram sonucuna göre 71 doğru 29 yanlış, quadri-gram sonucuna göre 82 doğru 18 yanlış olarak çıkmıştır.

Denklem 5.3'de çıkan sonuçların doğruluk ve hata oranı hesaplamaları yapılmıştır. Bilindiği gibi doğruluk bire yakınsaması, hata oranının sıfıra yakınsaması sonuçların doğruluğunu belirtmektedir. **Denklem 5.3**'de yapılan hesaplamalara göre bi-gram doğruluk değeri 0.12 ve hata oranı 0.88 çıkmıştır. Bu bi-gram ile yapılan yazar tanımanın başarısız olduğu anlamına gelmektedir. Tri-gram doğruluk değeri 0.71 ve hata oranı 0.29 olarak çıkmıştır. Tri-gram ile yazar tanıma işleminin oldukça başarılı olduğunu göstermektedir. Bir diğer özellik çıkarımı için kullandığımız quadri-gram hesabında doğruluk değeri 0.82 ve hata oranı 0.18 olarak hesaplanmıştır. Deney 1 sonucuna göre 100 kitap ve 20 yazar arasında yapılmış olan en başarılı yazar tanıma işlemini quadri-gram özellik uzayı göstermiştir. Tri-gram sonucu kabul edilebilir olsa da bi-gram sonucu başarısızdır.

Çizelge 5.2, **Çizelge 5.3**, **Çizelge 5.4** deney çıktılarının karmaşıklık matrisi çıkarılmıştır. Karmaşıklık matrisinde sisteme girilen kitapların doğru yazar etiketleri

ve sistemin tahmin ettiği yazar etiket değeri eşleşme durumları yazarlara göre ayrılarak gösterilmiştir. Tahmin durumlarının recall ve precision değerleri hesaplanmıştır. Hassasiyet (recall) bir yazara ait kitabın doğru etiketleme durum oranını göstermekte ve kesinlik (precision) doğru etiketlenmenin ilgili etiketle oranını gösteriyor. Burada belirtmek gerekir ki bazı yazarların sisteme sokulan test kitap sayısı veri bulunması bakımından azdır. Fakat kimi yazarların sisteme sokulan sayıları fazla olanlarda benzer sonuçları vermektedir. Sonuçlara bakılacak olursa bi-gram hassasiyet ve kesinlik değerleri düşük çıkmaktadır. Bu sonuçların çıkacağı doğruluk ve hata oranından da anlaşılabilir. Bunun yanında yine doğruluk ve hata oranı sonuçlarına bakarak tri-gram ve quadri-gram hassasiyet ve kesinlik değerlerinin başarılı çıkacağını tahmin edebiliriz. Sonuçlara baktığımızda ise bu değerlerin tri-gram ve quadri-gramda yüksek çıktığını görüyoruz.

Hassasiyet ve kesinlik değerleri tek başlarına yeterli bilgiyi sağlamaz. Bunun için bu iki değerin f-değeri (f-score) hesaplaması yapılır. F-score hesabı **Denklem 5-4**'te gösterilmiştir. Burada yapılan hesaplamaya göre en yüksek değeri quadri-gram vermiştir.

6 SONUÇ VE ÖNERİLER

Bu çalışmada çeşitli yazarların kitaplarıyla yaptığımız n-gram özelliklerini kullanarak otomatik yazar tanıma işlemini yaptık. Sisteme sokulan 100 adet test kitabıyla aldığımız sonuçlara göre n-gram başarımları arasında en başarılı olan quadri-gram özellik uzayı oldu. Tri-gram başarılı bir sonuç verse de %70 başarı tam anlamıyla yeterli gelmemektedir. Bi-gram özellik uzayı çıkarımı ise başarısız yani geçersiz bir özellik uzayıdır.

Burada uygulanan sistemde en başarılı sonucu quadri-gram vermiştir. Gelecekte yapılacak çalışmalar için n-gram özellik çıkarımıyla yazar tanıma işleminde "n sayısı artıkça daha başarılı sonuçlar alınabilir mi" sorusunun cevabı aranabilir. N değeri belirli bir sayıya kadar denenebilir belirli sayıdan sonra ayırıcılık özelliğini yitirecektir.

Yazar tanımak için n-gram özellik uzayına başka özellikler de eklenebilir ve ağırlıklandırılabilir. Örneğin ardışık kelimeler, cümle başlangıç ve bitiş harfleri, kullanılan noktalama durumları vesaire. Bu özellikler yazarın özellik uzayına katılırsa sistem ayırıcılığı artabilir. Yazarın yazım hataları da buna eklenebilir ancak genellikle kitaplar yayınlanmadan bir editör tarafından hataları düzeltilmektedir, bu yüzden eklenmesi uygun olmayacaktır.

Bu tezde sistemin sahip olduğu yazar özellik uzayı ile test kitabı yazar özellik uzayı karşılaştırılırken yakınlık durumuna bakılmıştır. Bir başka çalışmada uzaklık durumları karşılaştırılarak bir çıkarım yapılabilir.

Tezde n-gram sıklık durumları alındı ve en sık kullanılan 200 n-gramlar yazarın özellik uzayı olarak alındı. Bunun tam tersi en az kullanılan n-gramlar alınarak bir test yapılabilir. Bu işlemin de sonucu bilimsel açıdan cevap beklemektedir.

Sonuç olarak n-gram sıklığı yazar tanıma işlemi için başarılı olmuştur. Başka özelliklerle desteklenirse başarı oranı artması öngörülmektedir.

Çalışma tez sahibinin ilk DDİ çalışmamızdır. Daha sonraki çalışmalarda çok daha gelişmiş ve değişik doğal dil işleme Türkçe ve diğer dünya dilleri üzerinde

yapılacaktır. Çalışma boyunca birçok fikir geliştirilmiş ancak bir başka çalışma için ayrılmıştır.

Bilgisayar ve İnternet çağında Türkçe'nin diğer dillerden geri kalmaması için bir an önce makine dili ve Türkçe arasında daha sıkı bağlar kurulmalıdır. Mevcut durumda makineler ile en iyi anlaşabilen dilin İngilizce olduğu kabul gören bir gerçektir. Gelecek yüzyıllarda makineler ile anlaşamayan dillerin yok olma ile karşı karşıya kalabileceği kişisel ön görümdür. Türkçe'yi bu tehlikeden uzaklaştırmamız için makine dili ve Türkçe arasında iletişim sağlamamız gerekir. Şurası bir gerçektir ki makineler artan zekâsı ve nüfusuyla dünyada en kalabalık milleti sayılabilir. "Millet" kelimesi burada mecazi manada kullanılmış olsa da gerçekten bir millet olma olasılıkları çeşitli ütopya ve distopya kurgularında işlenmiştir. Dünyanın şu an henüz "agulama" devresinde olan makine milletinin bebekleri ilerde büyüdüklerinde hangi dille yetişmişlerse o dille konuşacaklar ve o dili bilenlerle daha iyi iletişim kuracaklardır. Bu yeni nüfusla iletişim kuramamak ya da daha az iletişim kurabilmek Türkçe'nin dezavantajı olacaktır. Bu durum birkaç yüz yıl sonra matbaanın topraklarımıza geç gelişi kadar milletimizin kaderini etkileyebileceği bir kıyas olarak verilebilir. Makine nüfusunu kaybetmemek için Türkçe doğal dil işleme çalışmalarına ağırlık verilmelidir.

KAYNAKLAR

Abbasi, A., & Chen, H. (2005). Applying authorship analysis to extremist-group Web forum messages. *IEEE Intelligent Systems*. https://doi.org/10.1109/MIS.2005.81

Antony, P. (2013). Machine Translation Approaches and Survey for Indian Languages. *Computational Linguistics and Chinese Language …*, *18*(1), 47–78. Retrieved from http://www.aclclp.org.tw/clclp/v18n1/v18n1a3.pdf

Argamon-Engelson, S., Koppel, M., & Avneri, G. (1998). Style-based text categorization: What newspaper am I reading? *Proceedings of AAAI Workshop on Learning for Text Categorization*, 1–4. Retrieved from http://www.aaai.org/Papers/Workshops/1998/WS-98-05/WS98-05-001.pdf

Argamon, S., & Levitian, S. (2005). Measuring the usefulness of function words for authorship attribution. *Proc. of the ACH/ALLC*, 1–3.

Argamon, S., Whitelaw, C., Chase, P., Hota, S. R., Garg, N., & Shlomo Levitan, L. (2007). Stylistic text classification using functional lexical features. *Journal of the American Society for Information Science and Technology*, *58*(6), 802–822. https://doi.org/10.1002/asi.20553

Baayen, H., Halteren, H. Van, & Tweedie, F. (1996). Outside the cave of shadows: using syntactic annotation to enhance authorship attribution. *Literary and Linguistic Computing*, *11*(3), 121–132. https://doi.org/10.1093/llc/11.3.121

Bennett, K., & Mangasarian, O. L. (1992). Robust linear programming discrimination of two linearly inseparable sets. *Optimization Methods and Software*, *1*(1), 23–34. https://doi.org/10.1080/10556789208805504

Brank, J., Grobelnik, M., Milic-Frayling, N., & Mladenic, D. (2002). Interaction of feature selection methods and linear classification models. *… on Text Learning Held at ICML*. Retrieved from http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:Interaction+of+Feature+Selection+Methods+and+Linear+Classification+Models#0

Burrows, J. F. (1987). Word-patterns and story-shapes: The statistical analysis of narrative style. *Literary and Linguistic Computing*, *2*(2), 61–70. https://doi.org/10.1093/llc/2.2.61

Burrows, J. F. (1992). Not unles you ask nicely: The interpretative nexus between analysis and information. *Literary and Linguistic Computing*, *7*(2), 91–109. https://doi.org/10.1093/llc/7.2.91

Collins, J., Kaufer, D., Vlachos, P., Butler, B., & Ishizaki, S. (2004). Detecting collaborations in text. *Computers and the Humanities*, *38*(1), 15–36. https://doi.org/10.1023/B:CHUM.0000009291.06947.52

Coyotl-Morales, R., Villaseñor-Pineda, L., Montes-y-Gómez, M., & Rosso, P. (2006). Authorship attribution using word sequences. *Progress in Pattern Recognition Image Analysis and Applications*, *4225*, 844–853. https://doi.org/10.1007/11892755_87

B.M.Sagar, (2014). Survey on Machine Translation and Its Approaches. *International Journal of Advanced Research in Computer and Communication Engineering*, *3*(6), 7317–7320.

Das, D. (2007). A Survey on Automatic Text Summarization Single-Document Summarization. *Language*, *4*, 1–31. https://doi.org/10.1016/B0-08-044854-2/00957-3

de Vel, O., Anderson, A., Corney, M., & Mohay, G. (2001). Mining e-mail content for author identification forensics. *ACM SIGMOD Record*, *30*(4), 55. https://doi.org/10.1145/604264.604272

Diederich, J., Diederich, J., Kindermann, J., Kindermann, J., Leopold, E., Leopold, E., Paass, G. (2003). Authorship attribution with support vector machines. *Applied Intelligence*, *19*, 109–123. https://doi.org/10.1.1.33.7558

Eifring, H., & Theil, R. (n.d.). The 2005 manuscript version of Halvor Eifring & Rolf Theil: Linguistics for Students of Asian and African. Retrieved from https://www.uio.no/studier/emner/hf/ikos/EXFAC03-AAS/h05/larestoff/linguistics/

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). Knowledge Discovery and Data Mining: Towards a Unifying Framework. *Int Conf on Knowledge Discovery and Data Mining*, 82–88. https://doi.org/10.1.1.27.363

Forman, G. (2003). An Extensive Empirical Study of Feature Selection Metrics for Text Classification. *Journal of Machine Learning Research*, *3*, 1289–1305. https://doi.org/10.1162/153244303322753670

Forsyth, R. S., & Holmes, D. I. (1996). Feature-Finding for Text Classification. *Literary and Linguistic Computing*, *11*(4), 164–174. https://doi.org/10.1093/llc/11.4.163

Gamon, M., & Grey, A. (2004). Linguistic correlates of style : authorship classification with deep linguistic analysis features. *Proceedings of the 20th International Conference on Computational Linguistics*, *4*, 611. https://doi.org/10.3115/1220355.1220443

Grishman, R. (1997). Information extraction: Techniques and challenges. In M. T. Pazienza (Ed.), *Information Extraction A Multidisciplinary Approach to an Emerging Information Technology: International Summer School, SCIE-97 Frascati, Italy, July 14--18, 1997* (pp. 10–27). Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/3-540-63438-X_2

Halteren, H. Van. (2007). Author verification by linguistic profiling. *ACM Transactions on Speech and Language Processing*, *4*(1), 1–17. https://doi.org/10.1145/1187415.1187416

Holmes, D. I. (1994). Authorship attribution. *Computers and the Humanities*, *28*(2), 87–106. https://doi.org/10.1007/BF01830689

Hotho, A., Nürnberger, A., & Paaß, G. (2005). A Brief Survey of Text Mining. *LDV Forum - GLDV Journal for Computational Linguistics and Language Technology*, *20*, 19–62. https://doi.org/10.1111/j.1365-2621.1978.tb09773.x

İletişim Nedir. (n.d.). Retrieved from http://www.turkedebiyati.org/iletisim-nedir.html

Indurkhya, N., & Damerau, F. J. (2010). *Handbook of Natural Language Processing*. *Processing* (Vol. 2). https://doi.org/10.1038/nbt1267

Jurafsky, D., & Martin, J. H. (1999). Speech and Language Processing. *Speech and Language Processing*. Retrieved from http://www.cs.colorado.edu/~martin/slp.html

Kešelj, V., Peng, F., Cercone, N., & Thomas, C. (2003). N-gram-based author profiles for authorship attribution. *Pacific Association for Computational Linguistics*, 255–264. https://doi.org/10.1.1.9.7388

Khmelev, D. V, & Teahan, W. J. (2003). A repetition based measure for verification of text collections and for text categorization. *Proceedings of SIGIR-03, 26th ACM International Conference on Research and Development in Information Retrieval*, 104–110. https://doi.org/10.1145/860454.860456

Kibble, R. (2013). Introduction to natural language processing. *University of London*.

Koppel, M. (2002). Automatically Categorizing Written Texts by Author Gender. *Literary and Linguistic Computing*, *17*(4), 401–412. https://doi.org/10.1093/llc/17.4.401

Koppel, M., Akiva, N., & Dagan, I. (2006). Feature instability as a criterion for selecting potential style markers. *Journal of the American Society for Information Science and Technology*, *57*(11), 1519–1525. https://doi.org/10.1002/asi.20428

Koppel, M., & Schler, J. (2003). Exploiting Stylistic Idiosyncrasies for Authorship Attribution. *IJCAI'03 Workshop on Computational Approaches to Style Analysis and* ***Synthesis*****, (2000),** 69–72. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.2.3019

Koppel, M., & Schler, J. (2004). Authorship verification as a one-class classification problem. In *Twenty-first international conference on Machine learning - ICML '04* (p. 62). https://doi.org/10.1145/1015330.1015448

Koppel, M., Schler, J., & Bonchek-Dokow, E. (2007). Measuring differentiability: unmasking pseudonymous authors. *Journal of Machine Learning Research*, *8*, 1261–1276. Retrieved from http://eprints.pascal-network.org/archive/00003399/

Kukushkina, O. V, Polikarpov, A. A., & Khmelev, D. V. (2001). Using Literal and Grammatical Statistics for Authorship Attribution. *Problems of Information Transmission*, *37*(2), 172–184. https://doi.org/10.1023/A:1010478226705

Li, J., Zheng, R., & Chen, H. (2006). From fingerprint to writeprint. *Communications of the ACM*, *49*(4), 76–82. https://doi.org/10.1145/1121949.1121951

Lochbaum, K. E., & Streeter, L. A. (1989). Comparing and combining the effectiveness of latent semantic indexing and the ordinary vector space model for information retrieval. *Information Processing and Management*, *25*(6), 665–676. https://doi.org/10.1016/0306-4573(89)90100-3

Madigan, D., Genkin, A., Lewis, D. D., Argamon, S., Fradkin, D., & Ye, L. (2005). Author identification on the large scale. In *Meeting of the Classification Society of North America*. https://doi.org/10.1.1.60.5324

MATSUURA, T., & KANADA, Y. (2000). Extraction of Authors' Characteristics from Japanese Modern Sentences via N-gram Distribution. In S. Arikawa & S. Morishita (Eds.), *Discovery Science: Third International Conference, DS 2000*

Kyoto, Japan, December 4--6, 2000 Proceedings (pp. 315–319). Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/3-540-44418-1_38

McCarthy, P., Lewis, G., Dufty, D., & McNamara, D. (2006). Analyzing Writing Styles with Coh-Metrix. *Proceedings of the Florida Artificial Intelligence Research Society International Conference*, (1995), 764–769.

Mendenhall, T. C. (1887). The characteristic curves of composition. *Science (New York, N.Y.)*, *9*(216), 297. https://doi.org/10.1126/science.ns-9.216.297

Mosteller, F., & Wallace, D. L. (1964). *Inference and disputed authorship: The Federalist*. *Addison-Wesley series in behavioral science. Quantitative methods.*

Natural Language Generation. (n.d.). Retrieved from https://en.wikipedia.org/wiki/Natural_language_generation

Nouri, J., & Yangarber, R. (2011). A Novel Evaluation Method for Morphological Segmentation. *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, 3102–3109.

P. Dragomir Radev. (n.d.). Introduction Natural Language Processing.

Peng, F., Schuurmans, D., & Wang, S. (2004). Augmenting Naive Bayes Classifiers with Statistical Language Models. *Information Retrieval*, *7*(3/4), 317–345. https://doi.org/10.1023/B:INRT.0000011209.19643.e2

Porshnev, A., & Redkin, I. (2014). Analysis of Images, Social Networks and Texts. *Communications in Computer and Information Science*, *436*(April), 190–197. https://doi.org/10.1007/978-3-319-12580-0

Pundge, A. M. (2016). Question Answering System , Approaches and Techniques : A Review. *International Journal of Computer Applications*, *141*(3), 34–39. https://doi.org/10.5120/ijca2016909587

Radev, D., Hovy, E., & McKeown, K. (2002). Introduction to the special issue on summarization. *Computational Linguistics*, *28*(4), 399–408. https://doi.org/10.1016/j.jbi.2011.03.008

Raschka, S. (2014). Naive Bayes and Text Classification I - Introduction and Theory. *arXiv Preprint arXiv:1410.5329*, 20. Retrieved from http://arxiv.org/abs/1410.5329

Rudman, J. (1998). The State of Authorship Attribution Studies: Some Problems and Solutions. *Computers and the Humanities*, *31*, 351–365. https://doi.org/10.1080/0013838X.2012.668785

Salton, G., Wong, a., & Yang, C. S. (1975). A vector space model for automatic indexing. *Communications of the ACM*, *18*(11), 613–620. https://doi.org/10.1145/361219.361220

Sanderson, C., & Guenter, S. (2006). Short Text Authorship Attribution via Sequence Kernels, Markov Chains and Author Unmasking: An Investigation. *Computational Linguistics*, (July), 482–491. https://doi.org/10.3115/1610075.1610142

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Computing Surveys*, *34*(1), 1–47. https://doi.org/10.1145/505282.505283

Stamatatos, E. (2006). Authorship attribution based on feature set subspacing ensembles. *International Journal on Artificial Intelligence Tools*, *15*(5), 823–838. https://doi.org/10.1142/s0218213006002965

Stamatatos, E. (2008). Author identification: Using text sampling to handle the class imbalance problem. *Information Processing and Management*, *44*(2), 790–799. https://doi.org/10.1016/j.ipm.2007.05.012

Stamatatos, E., Fakotakis, N., & Kokkinakis, G. (2001). Computer-Based Authorship Attribution Without Lexical Measures. *Computers and the Humanities*, *35*, 193–214. https://doi.org/10.1.1.102.9514

Stein, B., Koppel, M., & Stamatatos, E. (2007). Plagiarism analysis, authorship identification, and near-duplicate detection PAN'07. *ACM SIGIR Forum*, *41*(2), 68. https://doi.org/10.1145/1328964.1328976

Tambouratzis, G., Markantonatou, S., Hairetakis, N., Vassiliou, M., Carayannis, G., & Tambouratzis, D. (2004). Discriminating the Registers and Styles in the Modern Greek Language-Part 1: Diglossia in Stylistic Analysis. *Literary and Linguistic Computing*, *19*(2), 197–220. https://doi.org/10.1093/llc/19.2.197

Tweedie, F. J., & Baayen, R. H. (1998). How Variable May a Constant be? Measures of Lexical Richness in Perspective. *Computers and the Humanities*, *32*(5), 323–352. https://doi.org/10.1023/A:1001749303137

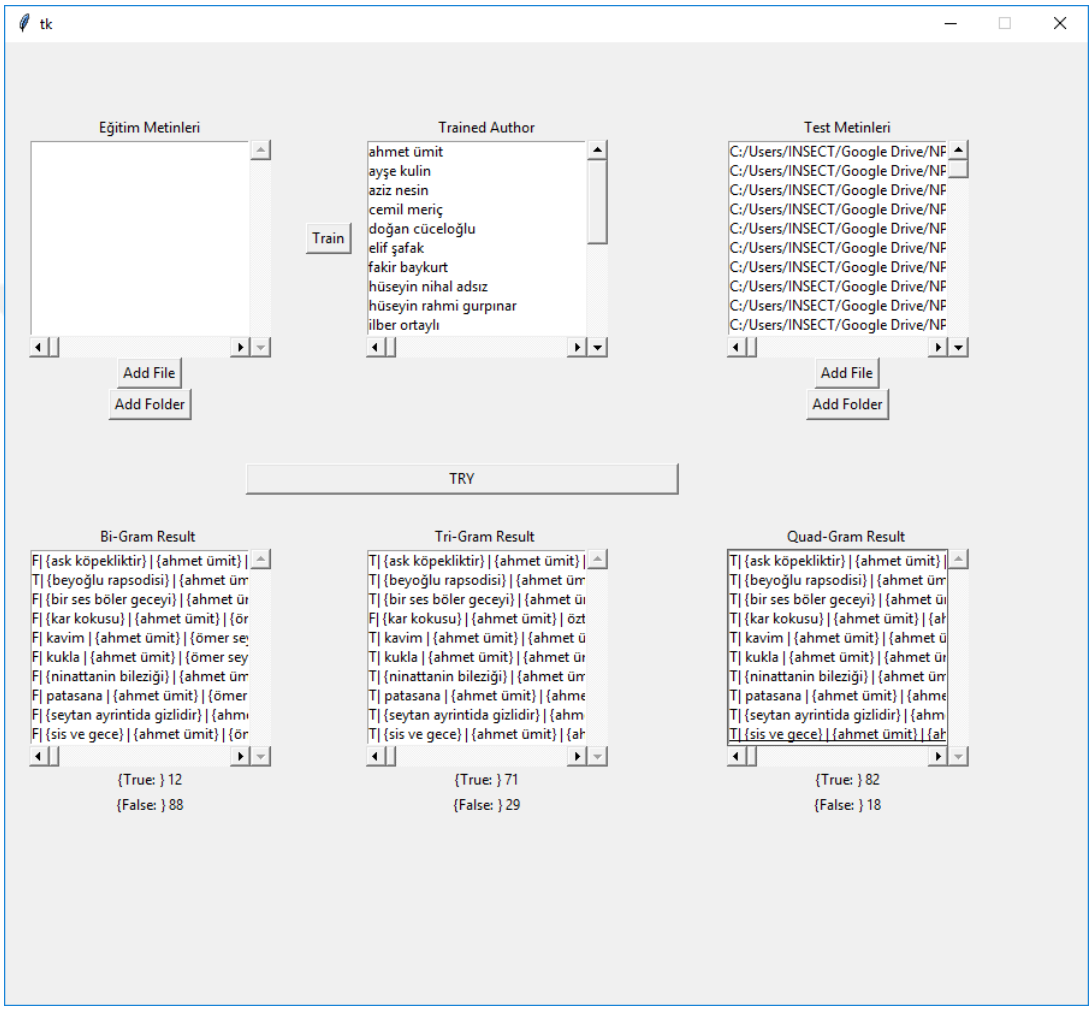
Zheng, R., Li, J., Chen, H., & Huang, Z. (2006). A framework for authorship identification of online messages: Writing-style features and classification techniques. *Journal of the American Society for Information Science and Technology*, *57*(3), 378–393. https://doi.org/10.1002/asi.20316

EKLER

EK A: Program Arayüzü

EK B: Text Kitapları ve Tahmin Sonuçları

EK A


tk
Eğitim Metinleri
Add File
Add Folder
Train
Trained Author
ahmet ümit
ayşe kulin
aziz nesin
cemil meriç
doğan cüceloğlu
elif şafak
fakir baykurt
hüseyin nihal adsız
hüseyin rahmi gurpınar
ilber ortaylı
Test Metinleri
C:/Users/INSECT/Google Drive/NF
Add File
Add Folder
TRY
Bi-Gram Result
F| {ask köpekliktir} | {ahmet ümit} |
T| {beyoğlu rapsodisi} | {ahmet üm
F| {bir ses böler geceyi} | {ahmet üı
F| {kar kokusu} | {ahmet ümit} | {ör
F| kavim | {ahmet ümit} | {ömer se
F| kukla | {ahmet ümit} | {ömer sey
F| {ninattanin bileziği} | {ahmet üm
F| patasana | {ahmet ümit} | {ömer
F| {seytan ayrintida gizlidir} | {ahm
F| {sis ve gece} | {ahmet ümit} | {ör
{True: } 12
{False: } 88
Tri-Gram Result
T| {ask köpekliktir} | {ahmet ümit} |
T| {beyoğlu rapsodisi} | {ahmet üm
T| {bir ses böler geceyi} | {ahmet üı
F| {kar kokusu} | {ahmet ümit} | özt
T| kavim | {ahmet ümit} | {ahmet ü
T| kukla | {ahmet ümit} | {ahmet üı
T| {ninattanin bileziği} | {ahmet üm
T| patasana | {ahmet ümit} | {ahme
T| {seytan ayrintida gizlidir} | {ahm
T| {sis ve gece} | {ahmet ümit} | {ah
{True: } 71
{False: } 29
Quad-Gram Result
T| {ask köpekliktir} | {ahmet ümit} |
T| {beyoğlu rapsodisi} | {ahmet üm
T| {bir ses böler geceyi} | {ahmet üı
T| {kar kokusu} | {ahmet ümit} | {ah
T| kavim | {ahmet ümit} | {ahmet ü
T| kukla | {ahmet ümit} | {ahmet üı
T| {ninattanin bileziği} | {ahmet üm
T| patasana | {ahmet ümit} | {ahme
T| {seytan ayrintida gizlidir} | {ahm
T| {sis ve gece} | {ahmet ümit} | {ah
{True: } 82
{False: } 18

EK B

KİTAP	YAZAR	Bi-Gram		Tri-Gram		Quad-Gram	
		TAHMİN	S	TAHMİN	S	TAHMİN	S
ask köpekliktir	a. ümit	ö. seyfettin	F	a. ümit	T	a. ümit	T
beyoğlu rapsodisi	a. ümit	a. ümit	T	a. ümit	T	a. ümit	T
bir ses böler geceyi	a. ümit	ö. seyfettin	F	a. ümit	T	a. ümit	T
kar kokusu	a. ümit	ö. seyfettin	F	tanımsız	F	a. ümit	T
kavim	a. ümit	ö. seyfettin	F	a. ümit	T	a. ümit	T
kukla	a. ümit	ö. seyfettin	F	a. ümit	T	a. ümit	T
ninattanin bileziği	a. ümit	ö. seyfettin	F	a. ümit	T	a. ümit	T
patasana	a. ümit	ö. seyfettin	F	a. ümit	T	a. ümit	T
seytan ayrintida gizlidir	a. ümit	ö. seyfettin	F	a. ümit	T	a. ümit	T
sis ve gece	a. ümit	ö. seyfettin	F	a. ümit	T	a. ümit	T
bir gün	a. kulin	ö. seyfettin	F	a. kulin	T	a. kulin	T
köprü	a. kulin	ö. seyfettin	F	a. kulin	T	a. kulin	T
nefes nefese	a. kulin	ö. seyfettin	F	a. kulin	T	a. kulin	T
sevdalinka	a. kulin	ö. seyfettin	F	a. kulin	T	a. kulin	T
bay duduk	a. nesin	a. nesin	T	a. nesin	T	a. nesin	T
borçlu olduklarimiz	a. nesin	ö. seyfettin	F	tanımsız	F	s. yalcın	F
damda deli var	a. nesin	a. nesin	T	e. şafak	F	a. nesin	T
gerçeğin masalı	a. nesin	ö. seyfettin	F	e. şafak	F	tanımsız	F
kazan toreni	a. nesin	a. nesin	T	tanımsız	F	a. nesin	T
memleketin birinde	a. nesin	ö. seyfettin	F	n. f. kısakürek	F	e. şafak	F
tatli betus	a. nesin	a. nesin	T	a. kulin	F	a. kulin	F
zübüklüğün sonu yok	a. nesin	ö. seyfettin	F	a. nesin	T	a. nesin	T
bir facianın hikayesi	c. meriç	ö. seyfettin	F	c. meriç	T	c. meriç	T
jurnal 1	c. meriç	ö. seyfettin	F	n. f. kısakürek	F	n. f. kısakürek	F
jurnal 2	c. meriç	ö. seyfettin	F	n. f. kısakürek	F	e. şafak	F
insan insana	d. cüceloğlu	ö. seyfettin	F	d. cüceloğlu	T	d. cüceloğlu	T
yetişkin çocuklar	d. cüceloğlu	ö. seyfettin	F	d. cüceloğlu	T	d. cüceloğlu	T
araf	e. şafak	ö. seyfettin	F	e. şafak	T	e. şafak	T
baba ve piç	e. şafak	ö. seyfettin	F	e. şafak	T	e. şafak	T
bit palas	e. şafak	ö. seyfettin	F	e. şafak	T	e. şafak	T
mahrem	e. şafak	ö. seyfettin	F	e. şafak	T	e. şafak	T
pinhan	e. şafak	ö. seyfettin	F	e. şafak	T	e. şafak	T
siyah süt	e. şafak	ö. seyfettin	F	e. şafak	T	e. şafak	T
keklik	f. baykurt	a. nesin	F	f. baykurt	T	f. baykurt	T
köygöçüren	f. baykurt	a. nesin	F	tanımsız	F	f. baykurt	T
bozkurtlar diriliyor	h. n. adsız	ö. seyfettin	F	h. n. adsız	T	h. n. adsız	T
bozkurtların ölümü	h. n. adsız	ö. seyfettin	F	tanımsız	F	h. n. adsız	T
delikurt	h. n. adsız	ö. seyfettin	F	h. n. adsız	T	h. n. adsız	T
makaleler	h. n. adsız	ö. seyfettin	F	s. yalcın	F	s. yalcın	F
dirilen iskelet	h. r. gurpınar	ö. seyfettin	F	n. f. kısakürek	F	h. r. gurpınar	T
gulyabani	h. r. gurpınar	a. nesin	F	h. r. gurpınar	T	h. r. gurpınar	T
muhabbet tılsımı	h. r. gurpınar	ö. seyfettin	F	h. r. gurpınar	T	h. r. gurpınar	T
nimetşinas	h. r. gurpınar	a. nesin	F	h. r. gurpınar	T	h. r. gurpınar	T
gelenekten geleceğe	i. ortaylı	ö. seyfettin	F	s. yalcın	F	c. meriç	F
osmanlı barışı	i. ortaylı	ö. seyfettin	F	c. meriç	F	s. yalcın	F

osmanlı devletinde kadı	i. ortaylı	ö. seyfettin	F	c. meriç	F	s. yalcın	F
büyük mal	k. tahir	a. nesin	F	k. tahir	T	k. tahir	T
devlet ana	k. tahir	ö. seyfettin	F	k. tahir	T	k. tahir	T
namuscular	k. tahir	a. nesin	F	e. şafak	F	a. kulin	F
sağırdere	k. tahir	ö. seyfettin	F	k. tahir	T	k. tahir	T
babıali	n. f. kısakürek	ö. seyfettin	F	n. f. kısakürek	T	n. f. kısakürek	T
dünya bir inkilap bekliyor	n. f. kısakürek	ö. seyfettin	F	n. f. kısakürek	T	n. f. kısakürek	T
moskof	n. f. kısakürek	ö. seyfettin	F	n. f. kısakürek	T	n. f. kısakürek	T
peygamber halkası	n. f. kısakürek	ö. seyfettin	F	n. f. kısakürek	T	n. f. kısakürek	T
rapor 1	n. f. kısakürek	a. nesin	F	n. f. kısakürek	T	n. f. kısakürek	T
rapor 2	n. f. kısakürek	a. nesin	F	n. f. kısakürek	T	n. f. kısakürek	T
tarih boyunca büyük mazlumlar	n. f. kısakürek	ö. seyfettin	F	n. f. kısakürek	T	n. f. kısakürek	T
tasavvuf bahçeleri	n. f. kısakürek	a. nesin	F	n. f. kısakürek	T	n. f. kısakürek	T
çöle inen nur	n. f. kısakürek	ö. seyfettin	F	n. f. kısakürek	T	n. f. kısakürek	T
baba evi 2	o. kemal	ö. seyfettin	F	a. kulin	F	tanımsız	F
baba evi	o. kemal	ö. seyfettin	F	a. kulin	F	tanımsız	F
cemile	o. kemal	ö. seyfettin	F	o. kemal	T	o. kemal	T
dünya evi	o. kemal	a. nesin	F	a. kulin	F	o. kemal	T
ekmek kavgası	o. kemal	a. nesin	F	o. kemal	T	o. kemal	T
eskici dükkanı	o. kemal	a. nesin	F	o. kemal	T	o. kemal	T
gurbet kuşları	o. kemal	a. nesin	F	e. şafak	F	o. kemal	T
hanımın çiftliği	o. kemal	a. nesin	F	e. şafak	F	o. kemal	T
çamaşırcının kızı	o. kemal	a. nesin	F	a. kulin	F	o. kemal	T
beyaz kale	o. pamuk	ö. seyfettin	F	tanımsız	F	tanımsız	F
kara kitap	o. pamuk	ö. seyfettin	F	o. pamuk	T	o. pamuk	T
kmasumiyet müzesi	o. pamuk	ö. seyfettin	F	o. pamuk	T	o. pamuk	T
sessiz ev	o. pamuk	ö. seyfettin	F	a. kulin	F	a. kulin	F
yeni hayat	o. pamuk	ö. seyfettin	F	o. pamuk	T	o. pamuk	T
canan	p. safa	ö. seyfettin	F	p. safa	T	p. safa	T
dokuzuncu hariciye koğuşu	p. safa	ö. seyfettin	F	s. ali	F	r. n. güntekin	F
fatih harbiye	p. safa	ö. seyfettin	F	s. ali	F	p. safa	T
selma ve gölgesi	p. safa	ö. seyfettin	F	p. safa	T	p. safa	T
aateş gecesi	r. n. güntekin	ö. seyfettin	F	r. n. güntekin	T	r. n. güntekin	T
acımak	r. n. güntekin	ö. seyfettin	F	r. n. güntekin	T	r. n. güntekin	T
akşam güneşi	r. n. güntekin	ö. seyfettin	F	r. n. güntekin	T	r. n. güntekin	T
dudaktan kalbe	r. n. güntekin	ö. seyfettin	F	r. n. güntekin	T	r. n. güntekin	T
yaprak dökümü	r. n. güntekin	ö. seyfettin	F	r. n. güntekin	T	r. n. güntekin	T
beco	s. yalcın	s. ali	F	s. yalcın	T	s. yalcın	T
beyaz muslumanlarin buyuk sırrı	s. yalcın	ö. seyfettin	F	s. yalcın	T	s. yalcın	T
erseverin itirafları	s. yalcın	ö. seyfettin	F	s. yalcın	T	s. yalcın	T
pipo	s. yalcın	ö. seyfettin	F	s. yalcın	T	s. yalcın	T
reis	s. yalcın	ö. seyfettin	F	s. yalcın	T	s. yalcın	T
siz kimi kandırıyorsunuz	s. yalcın	ö. seyfettin	F	s. yalcın	T	s. yalcın	T
teskilatin iki silahsoru	s. yalcın	ö. seyfettin	F	s. yalcın	T	s. yalcın	T
cagın sucu	u. mumcu	u. mumcu	T	u. mumcu	T	u. mumcu	T
devlet silah adalet	u. mumcu	ö. seyfettin	F	u. mumcu	T	u. mumcu	T
sakıncalı piyade	u. mumcu	a. nesin	F	u. mumcu	T	u. mumcu	T
yolsuzluk şiddet bağımlılık	u. mumcu	ö. seyfettin	F	u. mumcu	F	u. mumcu	F
ant	ö. seyfettin	ö. seyfettin	T	a. kulin	F	a. kulin	F
asilzadeler	ö. seyfettin	ö. seyfettin	T	ö. seyfettin	T	ö. seyfettin	T
düşünme zamanı	ö. seyfettin	ö. seyfettin	T	ö. seyfettin	T	ö. seyfettin	T
harem	ö. seyfettin	a. nesin	F	ö. seyfettin	T	ö. seyfettin	T
inat	ö. seyfettin	ö. seyfettin	T	ö. seyfettin	T	ö. seyfettin	T
külah	ö. seyfettin	ö. seyfettin	T	ö. seyfettin	T	ö. seyfettin	T
üç öğüt	ö. seyfettin	ö. seyfettin	T	ö. seyfettin	T	ö. seyfettin	T
Toplam Doğru Sayısı			12		71		82
Toplam Yanlış Sayısı			88		29		18

ÖZGEÇMİŞ

Ad-Soyad: Samet KAYA

Doğum Yeri: Bakırköy

Doğum Tarihi: 1987

E-mail: kysamet@gmail.com

ÖĞRENİM DURUMU:

Lisans:

- 2012, Marmara Üniversitesi, Teknik Eğitim Fakültesi, Bilgisayar Ve Kontrol Öğrentmenliği(İnglizce)
- 2017, Sakarya Üniversitesi, Bilgisayar Ve Bilişim Fakültesi, Bilgisayar Mühendisliği

TEZDEN TÜRETİLEN YAYINLAR, SUNUMLAR VE PATENTLER:

S. KAYA and A. GUNES, "Automatic Author Detection in Turkish books Using N-Gram and Naïve Bayesian Approach" International Conference on Advanced Technologies, Computer Engineering and Science (ICATCES), 2018.