TAMGACI: ARTIRIMSAL VE GERİ BESLEMELİ TÜRKÇE YAZAR ÇÖZÜMLEME

TURKISH AUTHORSHIP ANALYSIS WITH AN INCREMENTAL AND ADAPTIVE MODEL

OĞUZ ASLANTÜRK

Prof.Dr. HAYRİ SEVER
Tez Danışmanı

Hacettepe Üniversitesi
Lisansüstü Eğitim - Öğretim ve Sınav Yönetmeliğinin
Bilgisayar Mühendisliği Anabilim Dalı için Öngördüğü
DOKTORA TEZİ olarak hazırlanmıştır.

2014

OĞUZ ASLANTÜRK'ün hazırladığı **"TAMGACI: Artırımsal ve Geri Beslemeli Yazar Çözümleme"** adlı bu çalışma aşağıdaki jüri tarafından **BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI**'nda **DOKTORA TEZİ** olarak kabul edilmiştir.

Doç. Dr. Hasan OĞUL

Başkan .....

Prof.Dr. Hayri SEVER

Danışman .....

Prof. Dr. Haşmet GÜRÇAY

Üye .....

Doç. Dr. Ebru SEZER

Üye .....

Yrd. Doç. Dr. Mustafa EGE

Üye .....

Bu tez Hacettepe Üniversitesi Fen Bilimleri Enstitüsü tarafından **DOKTORA TEZİ** olarak onaylanmıştır.

Prof.Dr. Fatma SEVİN DÜZ
Fen Bilimleri Enstitüsü Müdürü

ETİK

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

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

beyan ederim.

31 / 01 / 2014

OĞUZ ASLANTÜRK

ÖZET

TAMGACI: ARTIRIMSAL VE GERİ BESLEMELİ TÜRKÇE YAZAR ÇÖZÜMLEME

Oğuz ASLANTÜRK

Doktora, Bilgisayar Mühendisliği Bölümü

Tez Danışmanı : Prof.Dr. Hayri SEVER

Ocak 2014, 105 sayfa

Yazar Çözümleme, bir metnin özelliklerini kullanarak o metnin yazarına ilişkin bilgi çıkartma eylemidir. Yaklaşık 130 yıllık geçmişe sahip bu araştırma konusunun kriminal, edebi, ticari ve akademik çerçevede önemli kullanım alanları bulunmaktadır. Yazar Tanıma ise bir metnin aday yazarlar arasından hangisi tarafından yazıldığını tespit etmeye çalışan ve Yazar Çözümleme disiplininin bir alt kolu olarak değerlendirilen bir araştırma alanıdır. Yazar Tanıma temelde bir sınıflandırma problemi olduğundan, makine öğrenmesi tekniklerinden bu alanda sıklıkla faydalanılmaktadır. Ancak yapılan çalışmalarda bugüne kadar 1000 kadar özellik incelenmiş olmasına rağmen, metne ait hangi özelliklerin en iyi, en başarılı özellikler olduğu konusunda halen ortak bir uzlaşı yoktur. Kullanılan metin özellikleri yazarların yüksek doğruluk oranlarıyla tanınmasında önemli olduğu kadar, oluşturulan sınıflandırıcıların eğitilmeleri için harcanan kaynaklar açısından da önemlidir. Çünkü özellik vektörleri büyüdükçe, kestirimsel modellerin eğitim süreleri de uzamakta, sınıflandırıcılar daha karmaşık hale gelmektedir. Öte yandan, zaman geçtikçe yazar üsluplarında meydana gelebilecek değişiklikler de

kullanılan metin özelliklerinin değiştirilmesi ya da sınıflandırıcıların yeniden eğitilmesini gerektirebilir.

Bu tez çalışmasında Türkçe için Yazar Tanıma amacıyla kullanılabilecek sözcüksel ve sözdizimsel metin özelliklerinden hangilerinin yazarları doğru olarak belirlemede en yüksek başarım oranını verdiği, buna ek olarak da başarımı yüksek bir en küçük metin özellikleri kümesinin tespiti araştırılmıştır. Bu amaçla Kaba Küme teorisinden faydalanılarak sınıflandırıcılar oluşturulmuş, belirlenen 37 metin özelliğinden hareketle tanımlanan 6 özellik grubunun bütün kombinasyonları ile, "Zamana Bağımlı" ve "Zamandan Bağımsız" modeller ile değişik zaman aralıkları için deneyler yapılmıştır. Deneyler gerçekleştirilirken, metin özelliklerinin yanısıra sınıflandırıcıların başarımlarının geçen zamanla birlikte değişip değişmediği, değişiyorsa nasıl değiştiği ve ne kadar süre boyunca yeniden eğitilmelerine gerek kalmadan kullanılabilecekleri de incelenmiştir.

12.000'den fazla köşe yazısı üzerinde yapılan 1134 deneyin sonuçları, Türkçe yazar tanımada en başarılı metin özelliklerinin bazı noktalama işaretleri (tire işareti, alt çizgi, taksim (*slash*) karakteri, ters taksim (*back slash*) karakteri, parantez, *ampersand* işareti), olduğunu, ayrıca hangi metin özelliklerinin kullanıldığından bağımsız olarak, sınıflandırıcıların en fazla 1 yıl süre ile yeniden eğitilmelerine gerek olmadan kullanılabileceklerini göstermiştir.

Anahtar Kelimeler: Yazar Çözümleme, Yazar Tanıma, Kaba Kümeler, Metin Özellikleri, Zamana Bağlı Değişim

ABSTRACT

TURKISH AUTHORSHIP ANALYSIS WITH AN INCREMENTAL AND ADAPTIVE MODEL

OĞUZ ASLANTÜRK

Doctor of Philosophy, Department of Computer Science

Supervison : Prof.Dr. Hayri SEVER

January 2014, 105 pages

Authorship Analysis is the analysis of a text to get information about the author of that text. It has a long history about 130 years with a wide range of studies, and is an important research topic for criminal, literary, commercial, and academic disciplines. Authorship Attribution is one of the distinct problems of Authorship Analysis and it deals with the identification of the author of a disputed text within a predefined set of candidate authors. Since it is basically a classification problem, machine learning techniques are widely employed for Authorship Attribution studies. However, although approximately 1000 stylistic features have been studied in different researches, there is still no consensus on which are the best and most distinctive. Stylistic features are very important for high prediction accuracies, as well as the resources needed to train the classifiers, because classification models become more complex when the size of input increased. On the other hand, changes of writing styles of authors in time may require to retrain the classifiers, or change the feature sets used.

In this thesis, lexical and syntactical stylistic features were analyzed for Authorship Attribution in Turkish. As well as finding the most distinctive

features for author detection, the smallest but distinctive sets of these features were investigated. Rough Set-based classifiers were constructed for this purpose, and all of the combinations of 6 feature groups defined from 37 features were analyzed with experiments which were performed using Time Dependent or Time Independent models for various periods of texts. By means of these models and periods, the effects of a possible temporal change on classifiers' performances were analyzed, as well as the distinctiveness of the features.

Results of 1134 experiments performed on more than 12.000 articles pointed that the most distinctive feature sets for Authorship Attribution in Turkish are some of the punctuation marks (hyphen, underscore, slash, back slash, paranthesis, ampersand). Additionally, independently of the features selected to train the them, classifiers should be used for at most 1 year before they are retrained.

Key Words: Authorship Analysis, Authorship Attribution, Author Detection, Rough Sets, Stylistic Features, Temporal Change

TEŞEKKÜR

Tez çalışmamda, değerli katkıları ve desteği için danışmanım Sayın Prof.Dr. Hayri SEVER'e, her aşamada bilgi ve birikimiyle yol gösteren Sayın Doç.Dr. Ebru SEZER'e, pozitif yorumlarıyla her fırsatta motive eden değerli arkadaşlarım Sayın Yrd. Doç. Dr. Erhan MENGÜŞOĞLU ve Sayın Yrd. Doç. Dr. Ahmet Burak CAN'a, kendi çalışmalarında edindikleri deneyim ve bilgiyi paylaşmaktan çekinmeyen ve tez çalışmasında kullanılan uygulamaların geliştirilmesinde katkıları olan arkadaşlarım Sayın Dr. Fatih Mehmet GÜLEÇ ve Yük. Müh. Tahir BIÇAKCI'ya, her zaman yanımda olan ve desteğini hiçbir zaman esirgemeyen sevgili aile fertlerime teşekkür ederim.

İÇİNDEKİLER

ÇİZELGELER

ŞEKİLLER

SİMGELER VE KISALTMALAR

BMR	Bayesian Multinomial Regression
BW	Balanced Winnow
CA	Cluster Analysis
CUSUM	Cumulative Sum
DA	Discriminant Analysis
DT	Decision Tree
kFCV	k-Fold Cross Validation
KLD	Kullback-Leibler Divergence
k-NN	k-Nearest Neighbour
LDA	Linear Discriminant Analysis
LR	Linear Regression
MLP	Multi Layer Perceptron
MRA	Multiple Regression Analysis
MANOVA	Multiple Analysis Of Variance
MVA	Multi Variate Analysis
MWF	Mean Word Frequency
NB	Naive Bayesian
NLP	Natural Language Processing
NN	Neural Networks
NSC	Nearest Shrunken Centroid
PCA	Principal Component Analysis
QSUM	Cumulative Sum
RF	Random Forest
RS	Rough Set
RBF	Radial Basis Function
SDA	Stepwise Discriminant Analysis
SMO	Sequential Minimal Optimization
SVM	Support Vector Machine
TTR	Type Token Ratio

1. GİRİŞ

Edebi üslubun çözümlenmesi ve bir metnin yazarı hakkında bilgi edinme eylemlerine genel olarak "yazar çözümleme (*authorship analysis*)" adı verilmektedir. İngiliz matematikçi ve mantıkçısı Augustus de Morgan'ın (1806 - 1871), "Aynı kişinin iki farklı konu hakkında yazdıklarının, aynı konu hakkında yazan iki farklı kişinin yazdıklarından daha çok birbirine benzemesini beklerim" [1] ifadesinin, Sophia de Morgan tarafından eşinin onuruna 1882 yılında yayınlanan anılarında [2] yer alışından beş yıl sonra T.C. Mendenhall tarafından yapılan inceleme, yazar çözümleme çalışmalarının başlangıcı olarak kabul edilir [3]. Mendenhall bu incelemesinde, aynı yazarın eserlerinde değişmeyen ama başka yazarlara ait eserler için makul biçimde değişen, böylece bir eserin yazarını biricik olarak tanımlayabilecek bir "sözcük spektrumu (*word-spectrum*)" ya da "karakteristik eğri (*characteristic curve*)" aramıştır. Bu çalışmadan günümüze kadar geçen yaklaşık 130 yıllık süre içerisinde Mendenhall'in açtığı yoldan birçok araştırmacı geçmiş, çok sayıda araştırma sonucu yayınlanmıştır. İlk zamanlar uzman kişilerin yoğun çalışmalarıyla yürütülen istatistiksel çalışmalar daha sonraları bilgisayar destekli hale gelmiş, istatistiksel işlemlerde bilgisayarların yüksek işlem güçlerinden faydalanmanın yanısıra makina öğrenmesi teknikleri de uygulanmaya başlamıştır. Bilgisayarlardan faydalanan çalışmalar "modern" olarak nitelendirilirken, uzman kişilerin emekleriyle yürütülmüş olan çalışmalara ise "geleneksel (*traditional/classic*)" denilmiştir. Modern yazar çözümleme çalışmalarının geleneksel çalışmalardan bir başka farkı da el yazısı metinler yerine elektronik ortamdaki metinler üzerinde işlem yapmasıdır.

Yazar çözümleme; eserlere ilişkin alıntı/intihal (*plagiarism*) iddialarının, tehdit, fidye, şantaj mektubu gibi suç oluşturan yazıların, yazarı kesin olarak bilinmeyen (ihtilaflı) yazıların ya da müşteri profillerinin çıkartılması gibi akademik, edebi, adli ve ticari alanlarda önemli kullanımları olan bir araştırma konusudur. Bu bağlamdaki araştırmalar dört alt grupta toplanabilir:

- Yazar Tanıma (*Authorship Attribution/Author Detection/Stylometry*): Metnin yazarının kim olduğunun belli olmadığı ama bir grup yazar arasından birisi tarafından yazıldığının bilindiği durumda, doğru yazarı bulma işidir. Birçok çalışmada "yazar tanıma" terimi ile "üslubun

ölçülmesi" anlamındaki "*stylometry*" terimi aynı anlamda kullanılmıştır. Bu tip çalışmalarda genellikle yazarı biricik olarak niteleyebilecek ve yazı özelliklerinden çıkarılmaya çalışılan niceliksel veriler incelendiğinden, "niceliksel yazar tanıma (*Quantitative Authorship Attribution*)" terimi de sık sık karşımıza çıkmaktadır.

- Yazar Doğrulama (*Author Verification*): Verilen bir metnin, belirli bir yazar tarafından yazıldığının ya da yazılmadığının doğrulanmasıdır.
- Yazar Profili Çıkarma (*Author Characterization*/*Author Profiling*): Verilen bir metnin yazarı hakkında yaşı, cinsiyeti, eğitim durumu, kültür yapısı.. vb bilgilerin çıkartılmasıdır.
- Benzerlik Bulma (*Similarity Detection*/*Plagiarism Detection*): İki metin arasındaki benzerliklerin bulunması işidir.

Yazar tanıma çalışmalarının önemli bir parçası, yazarın farklı yazılarında aynı kalan, yazarı biricik olarak niteleyebilecek olan, bir anlamda "yazarın yazılarındaki parmak izi" olarak adlandırılabilecek değişmezleri bulma çabasıdır. Bu değişmezlere, "yazarın değişmezi (*authorial or author's invariant*)" [4], "siber parmakizi (*cyper fingerprint*)" ya da "yazar-izi (*writerprint*)" [5] gibi isimler verilmektedir.

Yazar değişmezleri, metinlerden çıkartılan bir takım özellikler ya da özellik kümelerinden oluşur. Bu özellikler genellikle "biçemsel özellik (*stylistic feature*)" ya da "metin özelliği (*text feature*)" gibi isimlerle bilinmektedir. Yazar çözümleme çalışmaları sürecinde geçen zaman içerisinde 1000 kadar metin özelliği incelenmiştir [6]. Önceki araştırmalarda metindeki tek bir özelliğin yazarın değişmezi olduğu kabul edilip bu özellik belirlenmeye çalışılırken, modern yazar çözümleme bağlamında birden çok özellik bir arada kullanılmaktadır.

Metinler, şüphesiz yazıldıkları dile özgü özellikler taşır. Dolayısıyla yazarlara ait değişmezlerin kullanılan doğal dilden etkilenmesi beklenebilir. Bu bakış açısıyla, başta İngilizce olmak üzere, Yunanca, Rusça, Almanca, Arapça, Lehçe, Danca, Rumence, Portekizce, Hintçe ve Türkçe gibi birçok dilde yazar tanıma çalışmaları yapılmıştır. Bu çalışmaların bazıları hangi yöntemlerin daha iyi sonuç verdiğine odaklanırken, bazıları yeni özellikler belirlemeye çalışmakta, bazıları da çeşitli noktalardaki başarım iyileştirmelerine

odaklanmaktadır. Bu çalışmalar ile ilgili ayrıntılı bilgi Bölüm 2.1'de yer almaktadır.

Yazar tanıma probleminin çözümünde sıklıkla makina öğrenmesi tekniklerinden faydalanılmaktadır. Metinleri yazarlarına göre ayrıştıracak sınıflandırıcılar hazırlanmakta, bu sınıflandırıcıların eğitilmesinde ise bir ya da daha fazla gruptan seçilen özellikler ile oluşturulan özellik vektörleri kullanılmaktadır. Doğru özelliklerin seçilmesi (*feature selection*), makina öğrenmesinin kullanıldığı diğer problemlerde olduğu gibi yazar tanıma problemi için de önemli bir noktadır. Ayrıca, özellik vektörü büyüdükçe kestirim modeli karmaşıklaştığı, eğitim ve sınıflandırma işlemleri daha fazla işlemci gücü ve zamanı gerektirdiği için yazar değişmezlerini oluşturan doğru özellik kümelerinden en küçük olanlarının bulunması temel bir problem olarak karşımıza çıkmaktadır.

Yazar çözümleme çalışmalarında bir başka önemli nokta da gerek kullanılan dilin, gerekse yazarların dil kullanım özelliklerinin zamana bağlı değişimidir [7, 8]. Eğer böyle bir değişim varsa, bu değişim yazar tanıma amacıyla geliştirilen sınıflandırıcıların, eğitimlerinin devamında ne kadar süreyle etkin, geçerli ve güvenilir olarak kullanılabileceklerinin belirlenmesi açısından önem taşımaktadır. Ayrıca yazıldıkları tarihler 10, 20 ya da daha uzun yıllara yayılan metinlerdeki zamana bağlı değişim ile 1, 3 yıl gibi daha kısa aralıktaki metinlerin zamana bağlı değişimlerinin de birbirlerinden farklı oldukları düşünülebilir.

Yapılan literatür taramasında, Türkçe için yazar değişmezi olarak kullanılabilecek en küçük üslupsal özellik kümesinin bulunmasına ilişkin bir çalışmaya rastlanmadığı gibi, böyle bir kümenin geçerli ve güvenilir olarak kullanılabileceği sürenin belirlenmesine yönelik de herhangi bir çalışma bulunamamıştır. Tez çalışmasında bu problemlerin çözümüne odaklanılmıştır. Temel metin özelliklerinden oluşturulan birçok farklı özellik kümesi kombinasyonu üzerinde, zamana bağlı ve zamandan bağımsız modellerle, farklı gazetelerden, farklı alanlarda yazan yazarlar için gerçekleştirilen binden fazla deney sonucu birbirleriyle karşılaştırılarak en küçük üslupsal özellik kümesi ve bu kümenin ne kadar süreyle güvenilir olduğu araştırılmıştır.

Deneylerin gerçekleştirilmesinde Kaba Küme Teorisi (*Rough Set Theory*) ve araçlarından faydalanılmasının yanısıra, külliyatın hazırlanması, üslupsal özelliklerin çıkarılması, farklı kombinasyonlarla farklı modeller için girdi verilerinin hazırlanıp deneylerin gerçekleştirilmesi, sonuçların çözümlenmesi gibi birçok aşamada çeşitli uygulamalar geliştirilmiştir. Elde edilen sonuçlarda bazı özellik kümeleriyle %99'un üzerinde doğruluklarla metin yazarlarının belirlenebildiği görülmüştür. Dahası bu başarım değerleri, kullanılan özellik kümeleri içinde görece küçük kümeler tarafından yakalanabilmiştir. Öte yandan, aynı veri kümeleri üzerinde aynı özellikler kullanılarak yapılan zamana bağımlı ve zamandan bağımsız deneyler, özellik kümelerinin güvenli kullanılabilirlik sürelerinin bir yıl ile sınırlı olduğunu, ilk yıldan sonra başarımlarının düşmeye başladığını göstermektedir. Yazıların hangi alanda yazıldığına bağlı olarak, gündemin de etkisiyle bu başarım düşüşü hızlı ya da yavaş gerçekleşmekte, ancak her durumda bir düşüş yaşanmaktadır.

İkinci bölümde ilk olarak yazar çözümleme alanında daha önce yapılan çalışmalar aktarılmış, bu çalışmalarda kullanılan özellik, yöntem ve külliyatlar üzerinde durulmuştur. Problem tanımı ve çözüme ilişkin öneri açıklandıktan sonra, çözümde faydalanılan Kaba Küme Teorisine ilişkin temel bilgiler ile kaba kümeler bağlamındaki çeşitli algoritmaların gerçekleştirimlerini sunan ve tez kapsamında kullanılmış olan Rosetta uygulaması kısaca tanıtılmış, bu uygulama ile tarafımızdan gerçekleştirilen önceki bir çalışma özetle aktarılarak Kaba Kümelerin yazar tanıma bağlamında nasıl kullanılabileceği örneklenmiştir.

Üçüncü bölümde, çözümün doğrulanması amacıyla yapılan deneylerde kullanılan külliyatın, özelliklerin ve özellik gruplarının, zamana bağlı ve zamandan bağımsız kestirim modellerinin oluşturulması ve deneylerin gerçekleştirilme aşamaları ayrıntılarıyla anlatılmış, deney sonuçları irdelenmiştir.

Sonuç bölümünde ise varılan sonuçlar aktarılarak daha sonraki çalışmalarda incelenebileceği düşünülen fikirler üzerinde durulmuştur.

2. ARKA PLAN BİLGİSİ

Bu bölümde üzerinde çalışılan külliyat, özellik ve yöntemlere ilişkin literatür özeti verildikten sonra bu tez kapsamında kullanılan Kaba Küme teorisinin matematiksel temellerine kısaca değinilecek, ardından da kaba kümelerle sınıflandırma yapmak amacıyla faydalanılan bir uygulama tanıtılarak önceki bir çalışmamız üzerinde kullanımı aktarılacaktır.

2.1 Alanyazın Özeti

Yazar çözümleme, uzun zamana yayılmış ve çok sayıda uygulaması bulunan bir araştırma alanıdır. Bu alanda yapılan çalışmalara ilişkin ayrıntılı bilgi çeşitli kaynaklarda bulunabilir [9, 10, 11, 12, 13, 14, 15, 16]. Bu kaynaklarda çalışmalar genellikle kullanılan özellik, yöntem ya da külliyat açısından ele alınarak aktarılmaktadır. Burada ise kronolojik akış benimsenmiştir. Önemli çalışmalara tarihsel sırada değinilirken bazı noktalarda da özellik, yöntem ya da külliyata ilişkin kısa açıklamalar yapılmıştır.

Mendenhall'in 1887 yılında yaptığı ve yazar çözümleme bağlamındaki ilk çalışma olarak kabul edilen çalışmada sözcükler uzunluklarına göre gruplanarak bunların sıklık dağılımlarının eğrileri çizilmiş, böylece bir yazarı biricik olarak tanımlayabilecek bir "sözcük spektrumu"nun varlığı iddiası sınanmıştı. Mendenhall, üç farklı deney grubunun ilkinde Charles Dickens'ın "Oliver Twist" ve William Makepeace Thackeray'in "Vanity Fair" adlı romanlarını, ikinci deney grubunda John Stuart Mill'in "Political Economy" ve "Essay on Liberty" adlı denemelerini, üçüncü deney grubunda ise Edward Atkinson tarafından yapılmış iki konuşma metnini karşılaştırmıştı. Sonuçlar, sözcük spektrumunun çok başarılı bir ayrıştırıcı olmadığı gösteriyordu. Mendenhall'in kendi ifadesiyle: "eğer iki farklı derlem aynı eğrileri veriyorsa, aynı kaynaktan çıktıkları savı çok ikna edici değildir, çünkü olasılığı az olsa da iki yazarın aynı karakteristik eğriyi göstermeleri olanaklıdır" [3]. Bununla birlikte, sözcük uzunluklarını kullanarak çalışmaya devam etti ve Shakespeare ihtilafı üzerinde çalıştı. Shakespeare'in bir eser hırsızı olduğu, çeşitli sahneleri ya da dramaları başka kaynaklardan alıp derleyerek kendi oyunlarını ürettiğine dair süregelen iddiaları ilk ortaya atanlardan birisi Edmond Malone'dur. Shakespeare oyunlarının kronolojisi ve menşei hakkında bir uzman olan Malone, 1787 yılında yayınladığı çalışmasında üç bölümden

oluşan “Henry VI” adlı oyunun hiçbir bölümünün aslında Shakespeare tarafından yazılmadığını iddia etmişti [17]. Mendenhall, 1901 yılında yayınladığı çalışmasında William Shakespeare, Francis Bacon ve Christopher Marlowe’un sözcük spektrumlarını karşılaştırdı ve şu tespiti yaptı: “Christopher Marlow ile Skaespeare, Shakespear’in kendisi ile uyuştuğu kadar uyuşmaktadır” [18].

Mendenhall’in iki çalışmasının arasında, 1888 yılında Conrad Mascol’un iki kesimden oluşan çalışması yayınlandı [19, 20]. Bu yayınlarda, Paul adlı havari tarafından yazıldığı iddiasından dolayı “Paul Risaleleri (*Pauline Epistles*)” adıyla anılan ve erken dönem Hristiyanlığına ilişkin bilgiler içeren kutsal metinlerin her sayfasındaki ortalama cümle uzunluklarının yanısıra nokta, soru işareti ve virgül gibi noktalama işaretleri ile işlev sözcüklerinin sıklıkları incelenmişti.

1932 yılında Zipf [21], 1938 [22] ve 1944 [23] yıllarında da Yule tarafından yazar çözümleme çalışmalarına devam edildi. Zipf, bir yazara ait metinlerdeki farklı sözcüklerin sıklıklarını inceleyerek “Zipf Kanunu” olarak bilinen logaritmik ilişkiyi tanımladı. Yule ise önce cümle uzunluklarını kullansa da bunun çok güvenilir olmadığı sonucuna vardı ve sonrasında, Zipf kanununa dayanan ve büyük metinler için sözcük sıklığının ölçüsü olarak ifade edilen “Yule Karakteristiği” ya da “Yule’un Karakteristik K’sı” adıyla bilinen ölçüyü ortaya koydu. Fucks, 1952 ve 1954 yıllarındaki yayınlarında hece tabanlı sözcük uzunluklarının dağılımını inceledi [24, 25]. Sözcük uzunluklarını kullanan bir başka çalışma Brinegar tarafından on adet “*Quintus Curtius Snodgrass (QCS)*” mektubu üzerinde yapıldı [26]. Bunlar, New Orleans’ta yayınlanan “*The Daily Crescent*” adlı bir gazetede çıkan ve Mark Twain tarafından yazıldığı düşünülen bir dizi mektuptan oluşuyordu. Mektupların önemi, eğer bunlar gerçekten de Mark Twain tarafından yazılmışsa, Mark Twain’in Amerikan İç Savaşındaki rolüne ışık tutabilecek olmalarıydı. Ki-Kare Bağımsızlık Testi (*Chi-square test*) ve İki-Örneklem T Testi (*two-sample t-test*) uygulaması, Brinegar’ın Mendenhall’in yöntemi üzerine yaptığı iyileştirmelerdi.

Yazar çözümleme çalışmalarının ünlü problemi “*Federalist Papers*” üzerinde Mosteller ve Wallace tarafından yapılan çalışma bu alandaki kilometre taşlarından birisi oldu [27, 28]. *Federalist Papers*, New York halkını 1787

Amerikan anayasasını onaylamaya çağıran ve 1787-1788 yılları arasında yayınlanmış olan 85 adet denemeye verilen isimdir. Thomas Jefferson bu yazıları "devlet yönetimi ilkelerine ilişkin yazılmış en iyi yorum" olarak değerlendirmiştir [29]. "Publius" yazar adı ile yayınlanan bu yazıların Alexander Hamilton, James Madison ve John Jay elinden çıktığı bilinse de, yazılardan hangisinin kime ya da kimlere ait olduğu kesin olarak bilinmiyordu [29, 30]. Mosteller ve Wallace, metinlerin yazarlarını belirleyebilmek için işlev sözcüklerinin (*function words:* kendi başına pek bir anlam ifade etmeyen, diğer sözcükler arasında gramer ya da yapı ilişkilerini kurmaya yarayan yardımcı eylem, edat ya da bağlaç gibi sözcükler) sıklıklarını Bayes teoremi ile incelediler.

Morton'un Paul Risaleleri'ne ilişkin vardığı sonuçlar, 1963 yılında New York Times gazetesinde manşetten verildi. Morton, 1965 yılındaki bir dizi yayınında aynı çalışmaların üzerinden geçti [31, 32, 33]. Bu çalışmalarda cümle uzunlukları ve işlev sözcüğü sıklıkları ile cümlenin başına ya da sonuna göreli olarak belli pozisyonlardaki işlev sözcüklerinin sıklıklarını Ki-Kare testi ile karşılaştırdı ve bu Morton tarafından geliştirilen yeni bir özellikti.

1966 yılında O'Donnell cümle ve sözcük uzunluklarının yanısıra noktalı virgül ya da kesikli çizgiler gibi noktalama işaretlerinden de faydalanarak, yazarı öldüğü için başka bir romancı tarafından tamamlanan ve nereden itibaren ikinci yazar tarafından yazıldığı bilinmeyen bir romanı inceledi [34]. Bu çalışmanın ilginç bir yönü, kitabı tamamlayan yazar olan Robert Barr'ın, kitabın dörtte üçünü yazdığını iddia etmesi ve vefat eden yazar Crane'in dul eşinden, kitaptan elde edilen kazançtan da yazdığı oranda pay istemesiydi.

Levison, Morton ve Winspear, 1968'deki yayınlarında Plato'nun 7. Mektubunun aslında Plato tarafından yazılmadığını iddia ederken, Birikimli (Kümülatif) Toplam Grafiği (Cumulative Sum Plots) tekniği ile aldıkları sonuçları kanıt olarak gösteriyorlardı [35]. Plato'nun 7. Mektubu adıyla bilinen mektup Plato'nun Sicilya'daki etkinliklerinin bir otobiyografisi niteliğinde olup, Plato'ya atfedilen mektuplar arasında en uzunudur. Levison ve arkadaşlarının iddiasının önemi, bu mektubun Plato tarafından yazıldığına en çok güvenilen mektup olmasından kaynaklanır [36]. Morton, arkadaşları ile birlikte Antik Yunan eserleri üzerinde çalışmaya devam etti. 1972 yılında, Michaelson ile

birlikte bir dizi çalışmalarını yayınladılar. Bu çalışmalarda özel bir zamirin farklı çekim hallerinin birbirine oranı [37], özel bir bağlacın birbirini izleyen kullanımları arasındaki aralığın genişliği [38] ve cümle sonlarında kullanılan sözcüklerin sıklığı [39] gibi yeni özellikler kullandılar. Morton daha sonra 1978 yılında yayınladığı kitabında [40] en çok kullandığı tekniklerini anlattı. Bunlar işlev sözcüklerinin cümle içindeki belirli yerlerde bulunma sıklığı, arka arkaya kullanılan işlev sözcük dizilerinin ve eşanlamlı sözcük çiftlerinin sıklıklarıydı.

1978 yılında Kenny, "Aristo'nun Etikleri"ni incelemek üzere Yunanca işlev sözcüklerinin dağılımlarını kullanmış, dağılımların istatistiksel çözümlemesi için de Spearman'ın sıralama korelasyon katsayısı (Spearman'ın rho'su olarak da bilinir), Pearson korelasyon katsayısı ve Ki-Kare testi yöntemlerinden faydalanmıştır [41]. "Aristo'nun Etikleri", "Nichomachean Etiği" ve "Eudemian Etiği" adlarıyla bilinen ve Aristo tarafından yazılmış metinlere verilen isimdir [42]. Her ne kadar iki farklı isimle anılsa da bu isimler Aristo'nun kendisi tarafından verilmemiş, arkadaşı Eudemus ve oğlu Nicomachus tarafından metinlerde yapılan düzenlemelerden dolayı sonradan eklenmiştir. Metinlerin hangi sırayla yazıldığı bilinmemekte, ancak birçok kesimleri ortak olduğu için birinin diğerinin yeniden yazılmış hali olduğu düşünülmektedir. Kenny, incelemesinde "Eudamian Etiği"nin daha önce yazıldığı, "Nicomachean Etiği"nin de buna benzediği sonucuna varmıştır.

Thomas Merriam, Morton'un yöntemlerini destekleyen bir araştırmacı olarak, Morton'un 1978'deki kitabından sonra bu yöntemleri Shakespeare eserleri üzerinde uygulayan bir dizi çalışmasını yayınladı [43, 44, 45]. Bu çalışmalar, Morton'un yöntemlerinin güvenilmez sonuçlar ürettiğini savunan Smith'in 1985 yılında yayınladığı bir çift makaleden ikincisinde eleştirildi [46, 47]. Merriam'ın 1986'daki savunmasına [48], Smith, 1987 yılında yeniden atağa geçti [49]. Smith aynı yıl Morton'un *hapax legomena* (bir metinde yalnızca bir kez kullanılan sözcükler) sıklığına dayanan çalışmasını [50] da hedef alıyordu [51]. Merriam'ın, Smith'in eleştirilerine yanıtı fazla gecikmedi [52].

Bu karşılıklı eleştiri ve savunmalar sürerken, 1985 yılında Holmes'in yayınladığı incelemede [53] yazar çözümleme bağlamında kullanılan özellik ve yöntemler aktarılmıştı. Buna göre, kullanılan özellikler; sözcük uzunluğu sıklık dağılımları, sözcük başına ortalama hece sayısı ve hecelerin sözcüklere

göre dağılımı, ortalama cümle uzunluğu, konuşma parçalarının dağılımı, işlev sözcüklerinin sıklıkları, sözcük dağarcığı zenginliğine ilişkin çeşitli ölçüler (Simpson Indeksi (D), Yule Karakteristik K'sı, entropi, *Type-Token Ratio (TTR)* gibi), *hapax legomena* ya da *hapax dislegomena* (bir metinde yalnızca iki kez kullanılan sözcükler) gibi sözcük dağarcığı dağılımına ilişkin özellikler, sözcük sıklık dağılımları şeklinde gruplanabilir. Birçok çalışmada tek özellik üzerinde çalışılırken, daha sonraki çalışmalarda birden çok özellik bir arada ele alınmıştır. Uygulanan çözümleme yöntemleri arasında ise Ki-Kare testi, N-Örneklem T Testi, Bayes Analizi, Faktör Analizi (*Factor Analysis*), Diskriminant Analizi (*Discriminant Analysis*) ya da Kümeleme Analizi (*Cluster Analysis*) gibi istatistiksel yöntemler bulunmaktadır.

Yeni Ahit üzerindeki bir başka çalışma 1986 yılında yayınlandı [54]. Bu çalışmasında Kenny doksan dokuz farklı sözdizimsel özelliğin istatistiksel çözümlemesiyle Luke İncili (*Gospel of Luke*) [55] ile gene Luke tarafından yazıldığına inanılan Ameller Kitabının (*The Book of Acts, Acts of the Apostles* ya da kısaca *Acts* adlarıyla bilinir) [56] gerçekten de aynı elden mi çıktığını, John İncili (*Gospel of John*) [57] ile Vahiy Kitabının (*Book of Revelation, Revelation* ya da *the Apocalypse* adlarıyla bilinir) [58] yazarlarının aynı mı olduğunu ve Paul Risalaleri'nin gerçekten de Paul adlı havari tarafından mı yazıldığını araştırıyordu.

1987 yılında J. F. Burrows bir kitap yayınladı [59]. Bu kitapta, Jane Austen'in kurgusal karakterlerinin konuşmaları arasındaki farkları çözümlemek için, işlevsel sözcüklerin sıklıklarını PCA yöntemiyle incelemişti. Ertesi yıl Hassall ile birlikte yaptığı çalışmada Fielding problemini çözmek amacıyla artık *Burrows's Delta* adı verilen aynı yöntemden faydalandı [60]. Henry Fielding [61] ve kardeşi Sarah Fielding [62] İngiliz yazarlardı. Henry Fielding'in, eserlerindeki kadın karakterlerin geçmişlerini 3. tekil şahıs olan anlatıcı yerine, kadın karakterin kendi ağzından anlatmak gibi bir alışkanlığı vardı. Burada ilginç olansa, kadın karakterin ağzından anlatılan kesimlerin, gerçekte Sarah Fielding tarafından yazıldığından şüphelenilmesiydi. Burrows ve Hassall, bu problemi incelemek üzere her iki Fielding'in anlatılarından onar tanesinden çıkarttıkları işlev sözcükleri üzerinde PCA yöntemini uyguladı.

Yazar çözümleme araştırmacılarının en çok cezbeden problemlerden birisi olan *Federalist Papers*, 1989 yılında yeniden ele alınıyordu [63, 64]. Aynı yıl, Shakespeare'e ilişkin birçok çalışmadan bir başkası yayınlandı [65]. Bu çalışma, yazarı tam olarak bilinmeyen ama Shakespeare'e atfedilen "*A Funeral Elegy (Bir Cenaze Ağıtı)"* adlı ünlü şiiri inceliyordu ve sonraki yıllarda bu şiir tekrar tekrar gündeme gelecekti. Ertesi yıl ise yine tanıdık bir problem, Paul Risaleleri, bu sefer DA yöntemiyle mercek altına alınıyordu [66]. Aynı yıl Morton ve Michaelson, ortalama cümle uzunlukları ya da sesli harfle başlayan her cümledeki sözcük sayıları gibi değişik özellikleri Birikimli Toplam Grafiklerine (Cumulative Sum/CUSUM/QSum) yansıttıkları bir çalışma yayınladılar [67] ve Qsum tekniğinin yazar çözümleme literatüründe ilk kez bu çalışmada kullanıldığını iddia ettiler. Bu teknik daha sonra çeşitli kereler yeniden gündeme gelecekti [68, 69, 70, 71].

Burrows'un Delta yöntemi 1991 yılında Smith tarafından, Cyrill Tourneur ya da Thomas Middleton'a ait olduğuna inanılan "*The Revenger's Tragedy*" [72] adlı eser üzerinde sınandı [73]. İzleyen iki yıl içerisinde Smith aynı yöntemi Shakespeare'e atfedilen iki oyunu incelemek için kullandı [74, 75]. Bunlar, George Wilkins tarafından yazıldığı ileri sürülen "*Pericles"* [76] ile "*Edmund Ironside*" adlı oyunlardı [77]. 1992 yılında bu sefer Burrows'ın kendisi yine aynı yöntemi Bronte kardeşlerin [78] yazıları üzerinde sınadı [79]. İşlev sözcükleri ve PCA yöntemi, daha sonra da çeşitli çalışmalarda bir arada kullanıldı [80, 81, 82, 83]. David Holmes, 1992 yılında Mormon Kitabı'ndaki (*The Book of Mormon*) [84] yazar değişimlerini incelemek için PCA yöntemini *hapax legomena, hapax dislegomena*, Yule'un Karakteristik K'sı gibi sözcük haznesi zenginliğine ilişkin çeşitli kriterler üzerinde uyguladı [85].

1993 yılında Matthews ve Merriam, Yapay Sinir Ağlarını (*Neural Networks – NN*) kullanarak Shakespeare ve Fletcher üzerinde çalıştılar [86]. Bu çalışmada "sözcük oranı (*word ratio*)" adını verdikleri bir de kriter geliştirmişlerdi. Merriam aynı yıl bir başka Shakespeare ihtilafı üzerinde çalışırken [87], Ertesi yıl Matthews ve Merriam yine sözcük oranı kriteri ve NN kullanarak bu sefer Shakespeare ile Marlowe incelemesi yaptılar [88]. 1994 yılında yazar tanıma için NN kullanan bir başka araştırmacı, *Federalist Papers* üzerinde karakter n-gram'larından faydalanıyordu [89]. Aynı yıl bu sefer

Ledger ile birlikte çalışan Merriam, bir başka Shakespeare – Fletcher karşılaştırması için MVA tekniğinden faydalanıyordu [90]. 1994 yılına sığan birkaç diğer çalışmadan [91, 92] birisi de Holmes'un, o zamana kadar yazar tanıma alanında yapılan istatistiksel çözümlemeleri inceleyerek kullanılan özellikler/kriterleri özetlediği ünlü çalışmasıdır [9].

1995 yılı da yazar çözümleme çalışmaları açısından verimli bir yıldı. Holmes ve Forsyth, *Federalist Papers* üzerinde 3 özgün teknik denediler [93]. Bunlardan ilki kelime haznesi zenginliğine yönelik MVA, ikincisi ortak kullanılan yüksek sıklıklı sözcüklerin sıklık analizi ve son olarak da genetik algoritma kullanılmasıydı. Aynı yıl Ledger, MVA yöntemiyle Paul Risalelerini incelerken [94], Martindale ve McKenzie *Federalist Papers* üzerinde sözcük sıklıklarını kullanarak LDA ve NN ile çalışıyor [95], Lowe ve Matthews NN ve RBF ile yeni bir Fletcher – Shakespeare karşılaştırması yapıyor [96], Mealand ise Luke İncilini yeniden masaya yatırıyordu [97].

Ertesi yıl Baayen ve arkadaşları, sıklık ve hazne zenginliği kriterlerinin bulunmasında sözcükler yerine sözdizimsel yeniden yazım kurallarının (*syntactic rewrite rules*) kullanımını incelemek için önce [59]'daki, sonra da [85]'deki yöntemleri yeniden yazım kuralları ile uyguladılar [98]. Elliot ve Valenza ise Shakespeare'e ilişkin iddiaları sınamak için yaptıkları deneylerin sonuçlarını elli beş sayfalık bir makale ile yayınladılar [99]. Ward Elliot ve Robert J. Valenza, 1987 yılında Claremont Kolejinden 8 öğrencinin gerçek Shakespeare'i ve neleri gerçekten onun yazdığını bulmak amacıyla kurdukları *Claremont Shakespeare Clinic* ya da sadece *Shakespeare Clinic* [100] adlarıyla bilinen bir oluşumun üyeleriydi. Vardıkları sonuçlara göre "*A Funeral Elegy"* dışında bütün eserler gerçekten de Shakespeare'e aitti. Daha önce [65]'te "*A Funeral Elegy"*nin yazarının Shakespeare olduğunu söylemiş olan Foster bu yeni iddialara hemen cevap verdi ve Elliott ve Valenza'nın deneylerinin çoğunun hem tasarım hem de uygulama açısından ciddi hatalar içerdiğini yazdı [101]. Öte yanda Tweedie ve arkadaşları nöron ağları ile *Federalist Papers* üzerine eğilirken [102], Merriam [87]'de yaptığı Shakespeare-Marlowe karşılaştırmasına geri dönerek işlev sözcükleri üzerinde PCA uyguladı [82]. TTR ve MWF kriterleri ile Beatles şarkılarını inceleyen bir çalışma aynı yıl yayınlandı [103]. Forsyth ve Holmes ise, zaman

içerisinde yazar çözümleme alanında kullanılan birçok metin özelliğinin, çözümleyicinin içgüdülerine dayanılarak çıkartılan özellikler olduğunu, oysa hangi özelliklerin kullanılacağının seçiminin daha nesnel yapılması gerektiğini öneren bir araştırma yaptılar [104].

1998 yılında Holmes, istatistiksel yöntemlerin yazar çözümleme alanındaki kullanımlarının tarihsel gelişimini aktarırken [10], Rudman da o zamana kadar yapılan birçok modern yazar tanıma çalışmasının hala yöntem ya da özellikler bakımından bir uzlaşmaya varamayışını ve bir anlamda "her önüne gelenin yazar tanıma çalışması yapması"nı eleştiriyordu [6]. Merriam yeni bir Shakespeare incelemesi yaparken [105], Tweedie ve Baayen işlev sözcükleri ve sözcük haznesi zenginliği kriterleri üzerinde PCA ile çalışarak metin uzunluğuna göre değişmeyen, "sabit" kriterleri belirlemeye uğraşıyor [106], Elliot ve Valenza ise Foster'ın [101]'deki yanıtına yanıt veriyordu [107]. Ertesi yıl Hoorn ve arkadaşları harf dizilerinin güçlü bir araç olabileceği savlarını, NN, k-NN ve NB teknikleri ile sınadılar [108]. Binongo ve Smith, PCA yönteminin yazar tanıma alanında nasıl kullanılması gerektiğini açıklarken [109]. Foster ise *Shakespeare Clinic* eleştirisine devam ediyordu [110].

2000 yılında, Shakespeare'a atfedilen bir başka anonim oyun olan "Edward III", Merriam tarafından mercek altına alındı [111]. Merriam bu çalışmasında dişil sözcük ile biten satırlar, 10'dan daha az ya da daha çok hece içeren satırlar ve 5'ten daha az ya da daha çok vurgu içeren satırlar gibi bir takım değişkenkeri incelerken PCA ve QSUM çizgelerinden faydalandı. Waugh ve arkadaşları NN ile yeni bir *Federalist Papers* incelemesi yaparken [112], Stamatatos ve arkadaşları da Internet'ten indirdikleri günlük Yunan gazetelerinden derledikleri yazılar üzerinde "sözdizimsel öbek (*syntactic chunk)*" adını verdikleri özellikleri DA ve MRA ile incelediler [113]. Kullandıkları sözdizimsel parçalar, isim tamlamaları (*noun phrase*), ilgeç öbekleri (edat ve isimden oluşan sözcük öbekleri – *prepositional phrase*), eylem öbekleri (*verb phrase*) ve zarf öbekleri (*adverbial phrase*) ile iki öbeği birbirine bağlayan bağlaç dizileriydi.

Ertesi yıl Stamatatos ve arkadaşları benzer bir çalışma yayınladı [114]. Baayen'in sözcük kullanımlarınıı istatistiksel yöntemlerle çözümlenmesine ilişkin kitabı yayınlandı [115]. Craig ve Burrows, PCA tekniğini 17. Yüzyıldan

iki şiir üzerinde denediler [116]. de Vel ve arkadaşları, yazar tanıma çalışmaları külliyatları arasına e-postaları katarken SVM tekniğinden faydalandılar [117]. Khmelew ve Tweedie, karakter n-gramlarını Markov zincirleri (*Markov Chains*) ile çözümleyerek *Federalist Papers* problemini ele aldılar [118]. Chaski, bazı mahkeme kararlarında yazar tanımaya ilişkin deney sonuçlarının tanınması üzerine yaptığı çalışmayı yayınladı [119]. Bu çalışmasında aynı yaş, eğitim seviyesi ve entelektüel düzeyden iki Avrupa-Amerikalı, iki de Afrika-Amerikalı kadının yazdıkları üzerinde Ki-Kare testi ile sözdizim, noktalama, sözcük haznesi, içerik, noktalama hataları ya da yazım hataları gibi çeşitli özellikleri incelemişti. Grant ve Baker aynı yıl, hatta aynı derginin aynı sayısında bu yayına yanıt vererek dilbilimsel verinin incelenmesinde örneklem almanın karmaşıklığının genellikle hesaba katılmadığı iddia edip, güvenilir bir yöntem olarak da PCA tekniğini önerdiler [120]. Holmes, Gordon ve Wilson ile Pickett Mektuplarını incelemek için PCA tekniğinden faydalandı [121]. George Pickett Amerikan İç Savaşı sırasında görev almış bir generaldi ve dul eşi, generalin kendisine savaş alanında gönderdiğini iddia attiği mektupları yayınlamıştı. Oysa yazar ve İç Savaş tarihçileri bu mektupların en azının bir kısmının gerçekliğinden şüphe duyuyordu [122]. Holmes, bir yandan da Robertson ve Paez ile *New York Tribune* gazetesinde yayınlanan ve Stephen Crane'e ait olduğuna inanılan makaleler üzerinde PCA tekniğini uyguladıkları çalışmasını yayınladı [123]. Kukushkina ve arkadaşlarının 2000 yılında Rusça metinlerdeki karakter n-gramlarını Markov zincirleri ile incelediği çalışma İngilizce'ye çevrildi [124]. Hoover, CA tekniğinin etkinlik ve doğruluğunu sınamak üzere Amerikan ve İngiliz romanlarında oluşturduğu bir külliyattaki çok sık geçen sözcüklerin sıklıklarını kullandı [125]. Elliot ve Valenza ise yine bir Shakespeare incelemesi yaptılar [126].

Hoover ertesi yıl yayınladığı çalışmasında bu sefer sık geçen sözcük dizilerinin sıklıklarını ele alırken [127], 2003'deki çalışmasında ise sık geçen *collocation* (bir arada kullanılan farklı türden sözcükler) sıklıkları kullanıyordu [128]. Bu arada, Elliot ve Valenza ile Foster arasında geçen çekişmeye 2002 yılında son nokta koyulmuş, [129] editörün "... Bu yazı, Prof. Elliot ve Valenza'ya son bir yanıt verme olanağı sağlamak amacıyla dergide yer açmış

olduğumuz, serinin son yazısıdır...” notuyla yayınlanmıştı. Bu yıl, [12]’nin de literatüre kazandırıldığı yıl olmuştu.

2003 yılında Clement ve Sharp, beş eleştirmenin beş sinema filmine ilişkin yorumlarındaki karakter n-gramları üzerinde NB yöntemiyle çalışarak hem eleştirmenin hem de filmin kimliğini aradılar [130]. Diederich ve arkadaşları Almanca gazetelerden derledikleri külliyat ve SVM ile çalıştı [131]. Binongo, Oz Kitaplarını [132] işlev sözcükleri ve MVA ile incelerken [133], Hoover MVA ile yaptığı üç farklı çalışmayı aynı yıl yayınladı [128, 134, 135]. Koppel ve Schler, SVM ve DT yöntemi ile e-postalardaki kişiye özgü hatalar üzerinde çalıştı [136]. Peng, Keselj ve arkadaşları, iki ayrı çalışmada karakter n-gramları ile dilden bağımsız bir model geliştirilebileceğini göstermek için İngilizce, Çince ve Yunanca verileri kullandılar [137; 138]. Argamon ve arkadaşları ise British National Corpus külliyatının [139] geniş bir alt kümesi üzerinde yazar cinsiyetini çözümlemek amacıyla *winnow* sınıflandırmasndan faydalandılar [140]. 2004 yılında Hoover, yayınlanan iki çalışmasında *Burrows’s Delta* tekniğini kullanırken [141, 142], Peng ve arkadaşları Yunan gazetelerinden derledikleri külliyattaki karakter n-gram ve sözcük n-gramları üzerinde NB yöntemini uyguluyor [143], van Halteren ise Danca metinlerdeki sözcük n-gramları MVA ile çözümlüyordu [144]. Patton ve Can ise Yaşar Kemal’in İnce Memed dörtlemesini, sözcük tabanlı kriterler ve SDA ile MANOVA teknikleri yardımıyla incelerken [145], Collins ve arkadaşları *Federalist Papers* konusuna retorik bir bakış açısı katıyorlardı [146].

2005, Abbasi ve Chen’in Arapça Web forum iletilerini SVM ve DT ile inceledikleri [147], Chaski’nin dijital kanıt arayışında LDA tekniğine başvurduğu [148], Juola ve Baayen’in Danca metinlerdeki işlev sözcükleri üzerinde *cross-entropy* yöntemini kullandıkları [149], Burns’in Bayesian çıkarımı ile çalıştığı [150], Zhao ve Zobel’in işlev sözcükleri üzerinde NB, DT ve k-NN karşılaştırması yaptığı [151], Koppel ve arkadaşlarının SVM ile kişiye özgü hataları inceleyerek metnin dilini bulmaya çalıştıkları [152], Madigan ve arkadaşlarının yeni bir *Federalist Papers* incelemesini Bayes regresyonu ile gerçekleştirdikleri [153],yıldı.

Patrick Juola, yazar tanıma literatürüne ilişkin önemli incelemesini [14] yayınladığında yıl 2006 idi. Aynı yıl Koppel ve arkadaşları İbranice ve İbrani-

Aramice metinleri incelemek için BW yönteminden faydalanırken [154]. Zhao ve arkadaşları KLD ile ölçülen *relative entropy* yaklaşımını yazar tanıma alanına uyguladılar [155]. Zheng ve arkadaşları İngilizce ve Çince çevrimiçi haber gruplarından topladıkları iletileri NN, DT ve SVM ile incelediler [156]. Amasyalı ve Diri ise Türkçe çevrimiçi gazetelerden külliyat oluşturup karakter n-gramlarını NB, SVM, DT ve RF yöntemleri ile çözümleyerek yazar kimliği tespiti, belgeleri cinsiyete göre sınıflandırma ve yazarın cinsiyetini belirleme problemleri üzerinde çalıştılar [157].

2007 yılında Argamon ve arkadaşları SMO yöntemini kullanarak, yazar tanıma ya da cinsiyet belirleme gibi amaçlarla semantik işlevlere dayanan yeni bir özelliğin faydalı olabileceğini savundular [158], Burrows yazar tanıma için iki yeni test önerisinde bulundu [159], Hirst ve Feiguina, Bronte kardeşler problemine eğildi ve kısmi ayrıştırma ile elde ettikleri yeniden yazım kurallarının sıklıklarını SVM ile incelediler [160]. Pavelec ve arkadaşları ise SVM tekniğini Portekizce gazetelerden derledikleri külliyatta, bağlaç türleri üzerinde uyguladılar [161]. Zhao ve Zobel, bir önceki sene önerdikleri KLD yöntemini İngilizce literatürden çeşitli yazarlar üzerinde denerken, Shakespeare – Marlowe karşılaştırması yapmayı da ihmal etmediler [162]. Türkoğlu ve arkadaşları Türkçe çevrimiçi gazetelerden derlenmiş külliyat üzerinde sözcük tabanlı özelliklerden oluşturdukları on farklı özellik vektörünü NB, SVM, RF, k-NN ve MLP yöntemleri ile incelediler [163]. Stańczyk ve Cyran ise aynı sayıda yayınlanan çalışmalarından ilkinde iki Polonyalı yazarın romanlarını, işlev sözcükleri ve noktalama işaretleri üzerinde NN uygulayarak sınıflandırırken [164], ikincisinde ise dört Polonyalı yazarın romanlarından derledikleri külliyattaki noktalama işaretlerini incelemek üzere RS teorisinden faydalandılar [4]. Stańczyk, 2008'de en iyileştirilmiş (*optimised*) RS sınıflandırmasının [165], 2009'da da baskınlık tabanlı (*dominance-based*) RS sınıflandırmasının yazar tanıma bağlamında kullanımını sınayacaktı [166]. 2008 yılının bir başka önemli çalışması, Abbasi ve Chen'in yeni geliştirdikleri *Writeprint* adlı bir tekniği anlattıkları çalışmaydı [167]. Bu tekniği SVM ve PCA ile karşılaştırmak üzere çevrimiçi ortamdaki e-posta, yorum ve sohbet içeriklerini kullandılar. Aynı yıl Tearle ve arkadaşları, *Shakespeare – Marlowe* ve *Federalist Papers* incelemesinde NN'lere başvuruyor [168], Jockers ve

arkadaşları Mormon Kitabi'nı PCA ve NSC ile ele alıyor [169], Stamatatos ise İngilizce ve Arapça haberlerdeki n-gramları SVM ile inceliyordu [170]. 2009 yılında biri Koppel ve arkadaşları [15], diğeri de Stamatatos [16] tarafından olmak üzere, o zamana kadar kullanılan özellik ve yöntemleri anlatan iki güzel inceleme yayınlandı. Aynı yıl yine Koppel'in dahil olduğu bir çalışmada Argamon ve arkadaşları blog yazılarını BMR ile çözümleyerek yazarların cinsiyet, yaş, kişilik ve doğal dillerini bulmaya çalıştılar [171].

Elliot ve Valenza, Shakespeare incelemelerine 2010 yılında iki bölüm halinde yayınlanan bir araştırmayla geri döndüler [172, 173]. Holmes ve Crofts, "*The Diary of a Public Man*" [174] problemine PCA ve CA ile en çok kullanılan 50 işlev sözcüğünü inceleyerek yaklaşırken [175], Jockers ve Witten *Federalist Papers* problemi üzerinden Delta, k-NN, NSC, RDA ve SVM yöntemlerini karşılaştırdılar [176]. Aynı yıl biz de yazar tanıma bağlamında RS tabanlı sınıflandırıcıları kullanan bir model önerdik [177]. Ertesi yıl Koppel ve arkadaşları yazar tanımaya ilişkin üç önemli soruna birden yanıt arıyorlardı [178]. Bu sorunlar kabaca binlerce aday yazarın olduğu, isimsiz metnin aday yazarlardan hiçbirisine ait olmadığı ve yazar bilinen ya da bilinmeyen metinlerin çok sınırlı olduğu durumlardı. Rybicki ve Eder ise Burrows'un yöntemini sınamak üzere İngilizce, Almanca, Fransızca, Lehçe, Latince, Macarca ve İtalyanca metinler üzerinde deneyler yaptılar [179]. Eder, aynı külliyat ile 2013 yılında yaptığı çalışmada ise yazar tanımada kullanılabilecek en kısa metni arıyor, bunun için de yine Burrows'un Delta yönteminden faydalanıyordu [180]. 2012 yılnda Sayoud, Kur'an ve hadislerin aynı kalemden çıkıp çıkmadığını araştırmak üzere SVM, MLP ve LR yöntemleri ile karakter ve sözcük n-gramlarını kullandı [181]. Dahllöf ise SVM ile İsveçli politikacıların konuşma metinlerini cinsiyet, yaş ve politik görüşe göre sınıflandırdı [182]. Ertesi yıl Savoy çeşitli özellik seçme yöntemlerini karşılaştırmak üzere biri İtalyanca biri de İngilizce olmak üzere iki farklı gazetenin yazılarını KLD, Ki-Kare ve Delta yöntemlerini kullandı [183].

Uzunca bir sürede oldukça fazla çalışmanın olduğu yazar tanıma alanında, yukarıda bahsedilenlerin dışında daha birçok çalışma bulunmaktadır. Burada, yöntem, külliyat ve kullanılan özellikler açısından okuyucuya genel bilgi verilmeye çalışılmış, o nedenle de bir takım yenilikler öneren ya da önceki

çalışmaları ayrıntılı bir şekilde aktaran yayınlara yer verilmiştir. Böylece tezde yapılan çalışmanın özgün yönü de vurgulanabilecektir. Öncelikle, verilen birçok çalışmadan yalnızca iki tanesinde [4, 5] (her iki araştırma da Stanczyk ve Cyran tarafından yapılmıştır) yalın Kaba Kümeler ile oluşturulmuş sınıflandırıcılar kullanılmış, bunlarda da yalnızca 8 adet noktalama işaretinden oluşan çok küçük bir özellik kümesi incelenmiştir. Türkçe yazar tanıma çerçevesinde yapılan yayınlar incelendiğinde ise ne kaba kümelerin kullanıldığı, ne de çok sayıda üslupsal özellikten hangilerinin diğerlerinden daha ayrıştırıcı olduğunu belirlemeye yönelik bir çalışmaya rastlanmamıştır. Dolayısıyla bu tez çalışması hem kullanılan yöntem açısından, hem de araştırılan özellikler bakımından diğer çalışmalardan ayrılmaktadır. Bir başka özgün yönü ise yazar tanıma amacıyla oluşturulan sınıflandırıcıların kullanım sürelerine odaklanılmasıdır. Daha önce Türkçe metinlerdeki zaman etkisinin araştırıldığı bir çalışmada [8] Can ve Patton iki yazarın üslup özelliklerinde bir değişim meydana gelip gelmediğini incelemiş, kabaca 30 yıl arayla yazılmış metinleri “eski” ve “yeni” olarak sınıflandırmışlardır. Tez bağlamında ise 3-5 yıl aralığı gibi çok daha kısa vadedeki değişimler irdelenmiş, zamana bağlı değişimin varlığının yanısıra, sınıflandırıcıların bu değişimine karşı ne kadar süre direnip doğru bir şekilde yazarları ayrıştırabildikleri araştırılmıştır.

2.2 Problem Tanımı ve Çözüm Önerisi

Yazar tanıma, yazarı bilinmeyen bir metnin incelenerek, olası yazarlar kümesi içerisinden hangi yazar tarafından yazıldığı bilgisinin bulunması problemidir. Başka bir bakışla, metinlerin yazarlara göre sınıflandırılması problemi olarak da görülebilir. Sınıflandırmada kullanılan özellikler, metinler içerisinden çıkartılan sözcüksel, sözdizimsel, yapısal, içeriğe ya da yazara özgü çeşitli istatistiksel değerlerden oluşmaktadır. Zaman içerisinde yapılan birçok çalışmada toplamda bin kadar özellik kullanılmış [6] olsa da herkesçe kabul görmüş ve üzerinde uzlaşılmış bir özellik kümesi belirlenememiştir.

Birçok sınıflandırma probleminde olduğu gibi yazar tanımada da makina öğrenmesi teknikleri sıklıkla ve başarıyla kullanılmaktadır. Denetimli öğrenme (*supervised learning*) çalışmalarında oluşturulan kestirimci (*predictive*) modellerin eğitilmesinde kullanılan özellik vektörleri, metinlerden çıkartılan özellikler ile tanımlanmaktadır. Metinlerden çok sayıda özellik

çıkartılabilmesine rağmen, özellik vektörlerinin fazla sayıda eleman içermesi kestirimci modeli karmaşıklaştırmakta, buna bağlı olarak da eğitim ve sınıflandırma için gereken işlemci zamanını oldukça arttırmaktadır. Öte yandan daha çok metin özelliğinin kullanılması, yazarların daha yüksek yüzdelerle belirlenebileceği anlamına da gelmemektedir. Metinlerden basitçe çıkartılabilecek küçük özellik vektörleri, hızlı ve kolay adapte edilebilen, fazlaca işlem gücüne gereksinim duymayacak sistemlerin geliştirilmesi açısından önemlidir. Bir takım çalışmalarda dilden bağımsız özellikler tespit edilmeye çalışılsa da farklı dillerin farklı özellikleri olabileceği de gözardı edilmemelidir. Dolayısıyla, **dile özel ve en az sayıda özellikten oluşan, başarılı sınıflandırma sonuçları elde edilebilecek özellik kümelerinin bulunması** önemli bir problem olarak görülmektedir.

Bir başka soru ise yazarlar için belirlenen ayrıştırıcı özelliklerin zaman içinde değişip değişmediği, eğer böyle bir değişim varsa ne kadar sürede gerçekleştiği sorusudur. Bu, bir metnin yazarının doğru olarak belirlenebilmesinde oldukça önemli olabilir. Geçerliliğini yitirmiş özellik kümeleri ile yapılmaya çalışılan yazar tespiti, doğru yazarı bulamayabilir ya da daha kötüsü yanlış yazarları işaret edebilir. Örneğin tehdit içeren ve suç teşkil eden bir metnin yazarının hatalı olarak belirlenmesi kabul edilmesi mümkün olmayan bir durumdur. Bu bakış açısı ile, yazarlar için ayrıştırıcı özellik kümelerinin bulunmasının yanısıra, bu **özellik kümelerinin ne kadar süre ile geçerli olduğunun bulunması** da bir başka önemli problemdir.

Tez kapsamında bu iki problemin birden çözümüne ilişkin olarak önerilen yöntem;

- zamana bağımlı ve zamandan bağımsız olarak oluşturulacak kestirimci modellerle
- farklı alanlarda yazan farklı yazarların elinden çıkmış
- çok sayıda Türkçe metnin
- Türkçe metinler için ortak olan özelliklerinden faydalanan
- bu özelliklerin çok sayıda değişik kombinasyonları kullanılarak

birçok deney yapılması, elde edilen sonuçların irdelenmesidir. Böylece ilgisiz konularda, başka başka yazarlar tarafından, değişik zamanlarda yazılan

metinler, çeşitli özellik kümeleri için tekrar tekrar sınanmış olacağından hem özelliklerin birbirlerine göre başarımlarını, hem de zamana göre bu başarımların değişimlerini incelemek olanaklı hale gelecektir.

Yazar tanıma alanyazınında birçok farklı yöntem kullanılmıştır ve bu yöntemlerin hangileri olduğu önceki bölümde verilmiştir. Bu çalışmanın bir başka özgün yönü, yazar tanıma bağlamında özellik seçimi için Kaba Küme Teorisi'nin kullanılmasıdır. PCA, NN, k-NN ve SVM yöntemleri çok yaygın bir şekilde kullanılmış olmasına rağmen, başarılı bir sınıflandırıcı olan Kaba Kümelerden yalnızca birkaç çalışmada [4, 5, 165, 166] (hepsi de aynı kişi(ler) tarafından yapılmış olmak üzere) faydalanıldığı görülmüştür. Bu çalışmalarda ise birkaç özellikten oluşan küçük özellik vektörleri ile deneyler yapılmış, farklı özellikler dikkate alınmamıştır. Küçük özellik vektörlerinin kullanılma nedeni, çok sayıdaki özellik kümeleri ile eğitilen modellerin çok uzun çalışma sürelerine gereksinim duyması olabileceği gibi, yalnızca yönteme odaklanılarak hangi özelliklerin kullanılacağı konusunun fazla dikkate alınmaması da olabilir. Her iki senaryo da, tez kapsamında üzerinde durulan problemin ne kadar isabetli olduğuna işaret etmektedir.

2.3 Kaba Kümeler (*Rough Sets*)

Polonyalı matematikçi ve bilgisayar bilimcisi Zdzislaw Pawlak tarafından tanıtılan [184, 185] Kaba Küme Teorisi, tam olmayan bilginin belirsizliği ve kesin olmayışı sorunlarını ele alan bir araçtır. Verinin içinden çıkarılan bilgiyi kullanır ve başka herhangi bir parametreye gereksinim duymadan sınıflandırma yapmak amacıyla kullanılabilmektedir.

Klasik küme teorisinde kümeler sahip oldukları ya da olmadıkları elemanlarla tanımlanırken, Kaba Kümeler alttan yaklaşım (*lower approximation*) ve üstten yaklaşım (*upper approximation*) adı verilen bir çift küme ile tanımlanır. Alttan ve üstten yaklaşım kümelerini tanımlayabilmek için önce biçimsel tanımları verelim:

I, bir bilgi sistemi (*Information System*) olsun:

$$I=(U, A) \quad (1)$$

U, boş olmayan sonlu bir nesneler kümesi,

$$U = \{x_1, x_2, x_3, \ldots, x_n\} \quad (2)$$

A ise V_a, ***a*** niteliğinin alabileceği değerler kümesi olmak üzere, boş olmayan bir nitelikler kümesidir.

$$a: U \rightarrow V_a \ \forall a \in A. \qquad (3)$$

P, A kümesindeki niteliklerin bir alt kümesi olsun. Her ***P ⊆ A***, için bir ***IND(P)*** denklik ilişkisi *(equivalence)* vardır:

$$IND(P) = \{ (x_1, x_2) \in U^2 \mid \forall a \in P, a(x_1) = a(x_2) \} \qquad (4)$$

U/IND(P) (ya da ***U/P***), ***U***'nun ***IND(P)*** ile oluşturulan *bölümünü ifade etmek üzere kullanılır ve şöyle hesaplanır:*

$$U/IND(P) = \otimes \{ a \in P: U/IND(\{a\}) \} \qquad (5)$$

Burada

$$U/IND(\{a\}) = \{\{ x \mid a(x) = b, x \in U\} \mid b \in V_a\} \qquad (6)$$

ve

$$A \otimes B = \{ X \cap Y : \forall X \in A, \forall Y \in B, X \cap Y \neq \emptyset \} \qquad (7)$$

dir.

Eğer ***(x, y) ∈ IND(P)*** ise ***x*** ve ***y***, ***P*** kümesinden niteliklerle "ayırt edilemez (*indiscernible)"* denir. *P-**ayırt edilemezlik ilişkisi***nin *(P-indiscernibility relation)* denklik sınıfları $[x]_P$ ile gösterilir.

P kümesinin herhangi bir altkümesi R, için

$$[x]_R = [x]_P \qquad (8)$$

ise, R kümesine azaltılmış parça (***reduct)*** adı verilir. Bir bilgi kümesinde birden çok ***reduct*** bulunabilir.

X ⊆ U olacak şekilde bir ***X*** kümesini, ***P*** alt kümesindeki niteliklerle tanımlayalım. ***P***'deki nitelikler ile ayırt edilemeyen nesnelerden bazıları ***X*** kümesinin içindeyken bazıları dışında olabileceğinden, ***X*** kümesi ***P***'ye dayanarak kesin olarak ifade edilemez. Ancak ***X***'e, ***P***'nin ***X***'e alttan ve üstten yaklaşımları ile yaklaşılabilir.

P-alttan yaklaşım (*P-lower approximation*):

$$\underline{P}(X) = \{ x \mid [x]p \subseteq X \} \qquad (9)$$

P-üstten yaklaşım (*P-upper approximation*):

$$\bar{P}(X) = \{ x \mid [x]p \cap X \neq \emptyset \} \qquad (10)$$

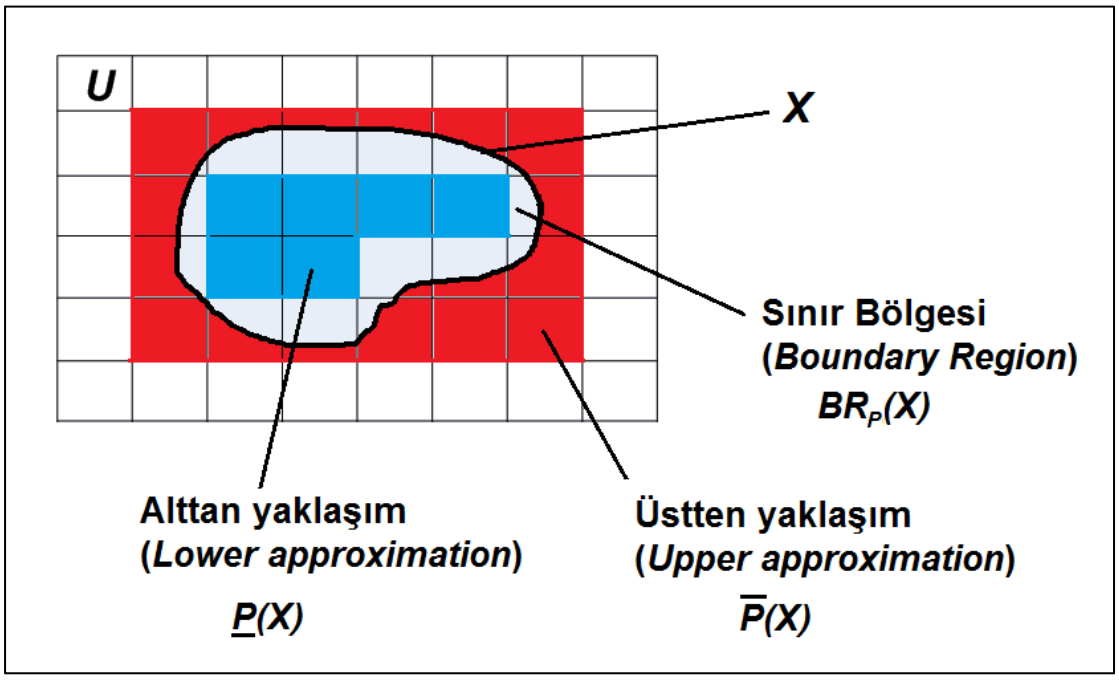
Alttan ve üstten yaklaşım kümelerinin farkı Sınır Bölgesi'ni (*Boundary Region*) verir:

$$BR_P(X) = \bar{P}(X) - \underline{P}(X) \qquad (11)$$

Kaba kümenin tanımı, sınır bölgesi ile verilir:

- Eğer X kümesinin sınır bölgesi boş küme değilse, X bir **kaba** (***rough***) kümedir.
- Eğer X kümesinin sınır bölgesi boş küme ise, X bir **keskin** (***crisp***) kümedir

Alttan ve üstten yaklaşımlar, sınır bölgesi ve kaba küme, Şekil 2.1'deki gibi gösterilebilir.



Şekil 2.1.Kaba Kümeler genel görünüm [185]

Bilgi sistemleri genellikle bir Karar Çizelgesi (*Decision Table*) ile gösterilir. Karar Çizelgesi, kolonları nitelik adlarıyla etiketlenmiş, hücrelerinde niteliklerin değerleri bulunan ve her bir satırı bir nesneyi ifade eden veri tablolarıdır. Çizelge 2.'de yazar tanıma problemi bağlamından örnek bir Karar Çizelgesi görülmektedir.

Çizelge 2.1. Kaba Küme için Örnek Karar Çizelgesi (f(x): x'in metinde geçme sıklığı)

Nesne	**Koşul Nitelikleri (*Condition Attributes*): Noktalama İşaretlerinin Sıklıkları***								**Karar Niteliği (*Decision Attribute*)**
Metin	**f(.)**	**f(,)**	**f(;)**	**f(:)**	**f(?)**	**f(!)**	**f(-)**	**f(“)**	**Yazar Adı**
M1	1	1	0	0	1	0	0	0	Y1
M2	1	1	1	0	0	1	0	0	Y1
M3	1	1	0	0	0	0	0	0	Y1
M4	1	1	1	0	0	1	0	0	Y2
M5	0	0	1	0	0	0	0	1	Y2
M6	1	1	0	0	1	0	0	0	Y2

Çizelge 2.1'in her bir satırı bir metne karşılık gelmekte ve her bir metin, o metinde kullanılan noktalama işaretlerinin sıklıklarıyla ifade edilmektedir. Bir noktalama işaretinin bir metindeki (örneğin '?'nin M1 metnindeki) sıklığı belli bir eşik değerden (örneğin o yazarın bütün metinlerindeki '?' sıklıklarının ortalamasından) daha yüksekse o niteliğin değeri 1, değilse 0 verilmiştir. Son kolonda ise o metnin hangi yazar tarafından yazıldığı görülmektedir.

Şimdi kaba kümeleri bu karar çizelgesi üzerinden açıklayalım:

f(x), *x*'in sıklığını göstermek üzere;

Nesne kümesi:

$$U = \{M1, M2, M3, M4, M5, M6\}$$

Koşul Kriterleri kümesi:

$$C = \{f(.), f(,), f(;), f(:), f(?), f(!), f(-), f(“)\}$$

Karar Kriterleri kümesi:

$$D = \{Y1, Y2\}$$

Nitelik kümesi:

$$A = C \cup D$$

Örnek nitelik kümesi olarak bir *P* kümesi seçelim; *P* = *C* olsun.

$$P = \{f(.), f(,), f(;), f(:), f(?), f(!), f(-), f(“)\}$$

Bu durumda, *M1* ile *M6*, *M2* ile *M4*, *P* kümesindeki niteliklerle ayırt edilemez olur.

Ayırt edilmezlik kümeleri:

$$IND_P = \{M1, M6\}, \{M2, M4\}, \{M3\}, \{M5\}$$

Rastgele bir *X* kümesi seçelim:

$$X = \{M1, M2, M3\}$$

olsun. *X* kümesinin, *P* nitelikleri cinsinden kesin tanımını veremediğimiz için alttan ve üstten yaklaşım yöntemini seçeriz:

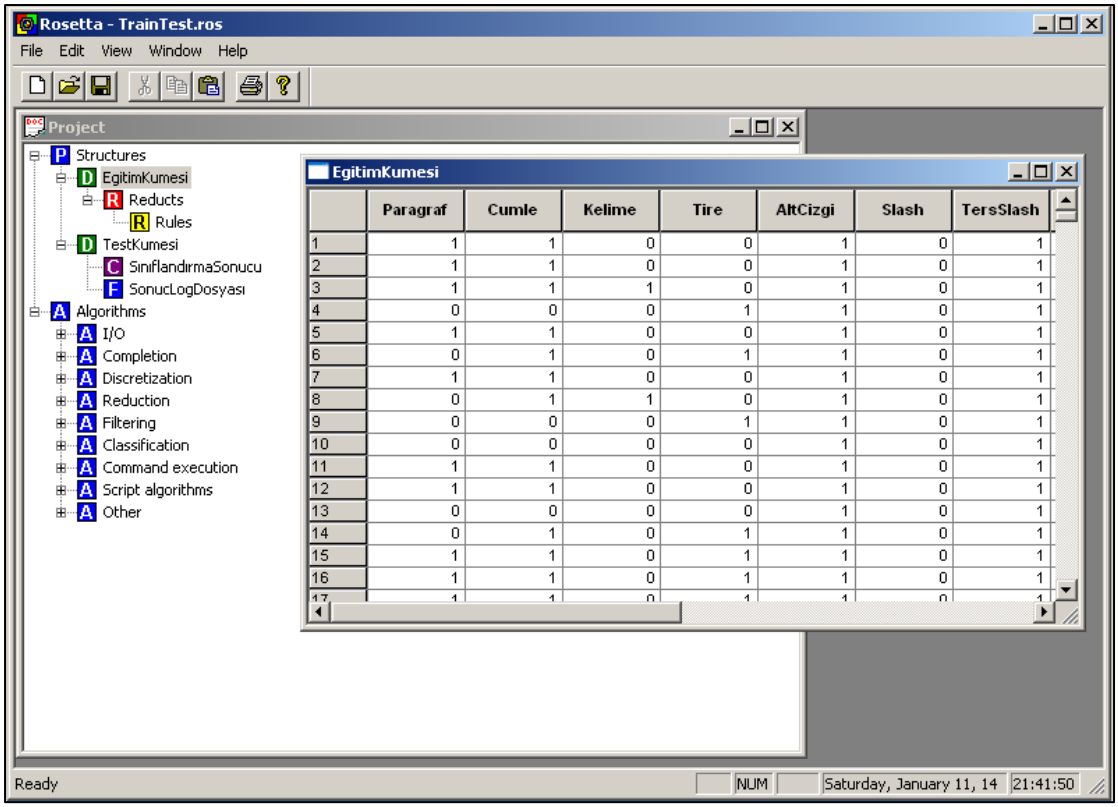
$$\underline{P}(X) = \{M1, M2, M3\}$$

$$\overline{P}(X) = \{M1, M6, M2, M4, M3\}$$

$$BRp(X) = \overline{P}(X) - \underline{P}(X) = \{M6, M4\}$$

2.4 ROSETTA ve CLROSETTA

ROSETTA, bir doktora çalışması kapsamında geliştirilmiş [186, 187], Kaba Küme Teorisi'ne ilişkin çeşitli algoritmaların gerçekleştirimleri ile karar çizelgesi biçimindeki verileri çözümlemek için kullanılabilecek genel amaçlı bir yazılımdır [188]. Veri madenciliği (*Data Mining*) ya da Bilgi Keşfi (*Knowledge Discovery*) gibi işlemler yapılabilmesini kolaylaştıran çeşitli araçları vardır. Bu araçların içinde; veri, kural, tablo, çizge ya da çeşitli formatlardaki nesneleri metin dosyası, Excel dosyası ya da veri tabanı gibi ortamlardan alabilme (*import*) ya da bu ortamlara yazabilme (*export*), veriyi gözden geçirme (*browse*), ayrıklaştırma (*discretization*), bölümleme (*partitioning*) gibi önişlemler (*preprocess*), denetimli (*supervised*) ya da denetimsiz (*unsupervised*) öğrenme, ayırt edilemezliğe ilişkin kuralların üretilmesi, betik (*script*) çalıştırma, çapraz doğrulama (*cross-validation*) sayılabilir. Windows ya da Linux ortamlarında çalışan program, Grafik Kullanıcı Arayüzü (*Graphical User Interface - GUI*) aracılığıyla kullanılmaktadır. Açık kaynak kodlu ve kullanımı ücretsiz olup Şekil 2.2'de uygulamadan bir ekran görüntüsü görülmektedir.



Şekil 2.2. ROSETTA uygulamasından bir ekran görünümü

CLROSETTA ise ROSETTA uygulamasının çekirdek işlevlerini içeren komut satırı sürümüdür. Her ne kadar CLROSETTA ile yapılabilen işlemlerin tamamı ROSETTA arayüzü üzerinden de gerçekleştirilebilir olsa da, özellikle çok sayıda deneyin söz konusu olduğu durumlarda CLROSETTA pratik olmanın ötesinde, kritik öneme sahiptir. Bunun nedeni, arayüz ile birkaç menüden geçilerek elle yapılabilecek bir işin, CLROSETTA sayesinde bir betik ile otomatikleştirilebilmesidir.

Yazar tanıma işlemi temelde, metinleri yazarlarına göre sınıflandırma işlemidir ve bu sınıflandırma işlemi ROSETTA ya da CLROSETTA ile yapılabilmektedir. Tez kapsamında iki farklı şekilde sınıflandırma işlemi yapılmıştır:

1. Yöntem: Seçilen bir eğitim kümesinden sınıflandırma kurallarının bulunması ve bu kuralların test kümesi üzerinde uygulanması. Yöntemin adımları Şekil 2.3'te verilmiştir.

i. Eğitim kümesi karar çizelgesinin yüklenmesi
ii. Karar çizelgesinden *reduct* kümesinin çıkarılması
iii. *reduct* kümesinden kural kümesinin çıkarılması
iv. Test kümesi karar çizelgesinin yüklenmesi
v. Kural kümesi ile test kümesinin sınıflandırılması

Şekil 2.3. Eğitim/Test yönteminin adımları

2. Yöntem: Yazılımın k-Katlı Çapraz Doğrulama (k-Fold Cross Validation, kFCV) yeteneğinden faydalanılması. Bu yöntemde bilgi sistemi eğitim ve test kümelerine ayrılmaz, yazılıma bir bütün olarak verilir. Kaç katlı doğrulama (örneğin 5) yapılacaksa, yazılım veriyi o kadar eşit parçaya bölerek bunlardan bir tanesini test, kalanını eğitim kümesi olarak kullanır. Eğitim kümesinden öğrenilen kurallarla test kümesi sınıflandırılır. Bu işlem, k kez tekrarlanır. Böylece bütün verilerin 1 kez test kümesinde, k – 1 kez de eğitim kümesinde yer alması sağlanmış olur.

i. Bilgi sistemi karar çizelgesinin yüklenmesi
ii. Kaçlı çapraz doğrulama yapılacağının belirtilerek sınıflandırmanın yapılması

Şekil 2.4. kFCV yönteminin adımları

Karar çizelgelerinin yazılıma yüklenmesi için veri tabanı, Excel ya da metin dosyaları kullanılabilirken, sınıflandırma sonuçları da adı kullanıcı tarafından belirtilen günlük dosyalarına kaydedilebilmektedir. Bu tez çalışmasında yapılan deneylerdeki karar çizelgelerinin ROSETTA'ya yüklenebileceği uygun formatlarda hazırlanabilmesi ve ROSETTA'nın oluşturduğu günlük dosyalarındaki sonuçların işlenebilmesi için uygulamalar geliştirilmiştir. Bu uygulamalar ilerleyen kesimlerde anlatılacaktır.

CLROSETTA, kFCV şeklindeki deneyleri gerçekleştirebildiği için 2. yöntem açısından uygundur. Ama bir eğitim kümesinin karar çizelgesinden öğrenilmiş

kuralları farklı bir karar çizelgesine uygulayamadığı için 1. yöntem ile gerçekleştirilen deneyler CLROSETTA ile gerçekleştirilememiş, zorunlu olarak ROSETTA kullanılmıştır.

2.5 ROSETTA ile yapılan örnek bir yazar tanıma çalışması

Bu kesimde, Bölüm 0'de değinilen ve detayları [177]'de bulunabilecek olan çalışma özetlenerek Rosetta yazılımının yazar tanıma amacıyla nasıl kullanılabileceği anlatılacaktır.

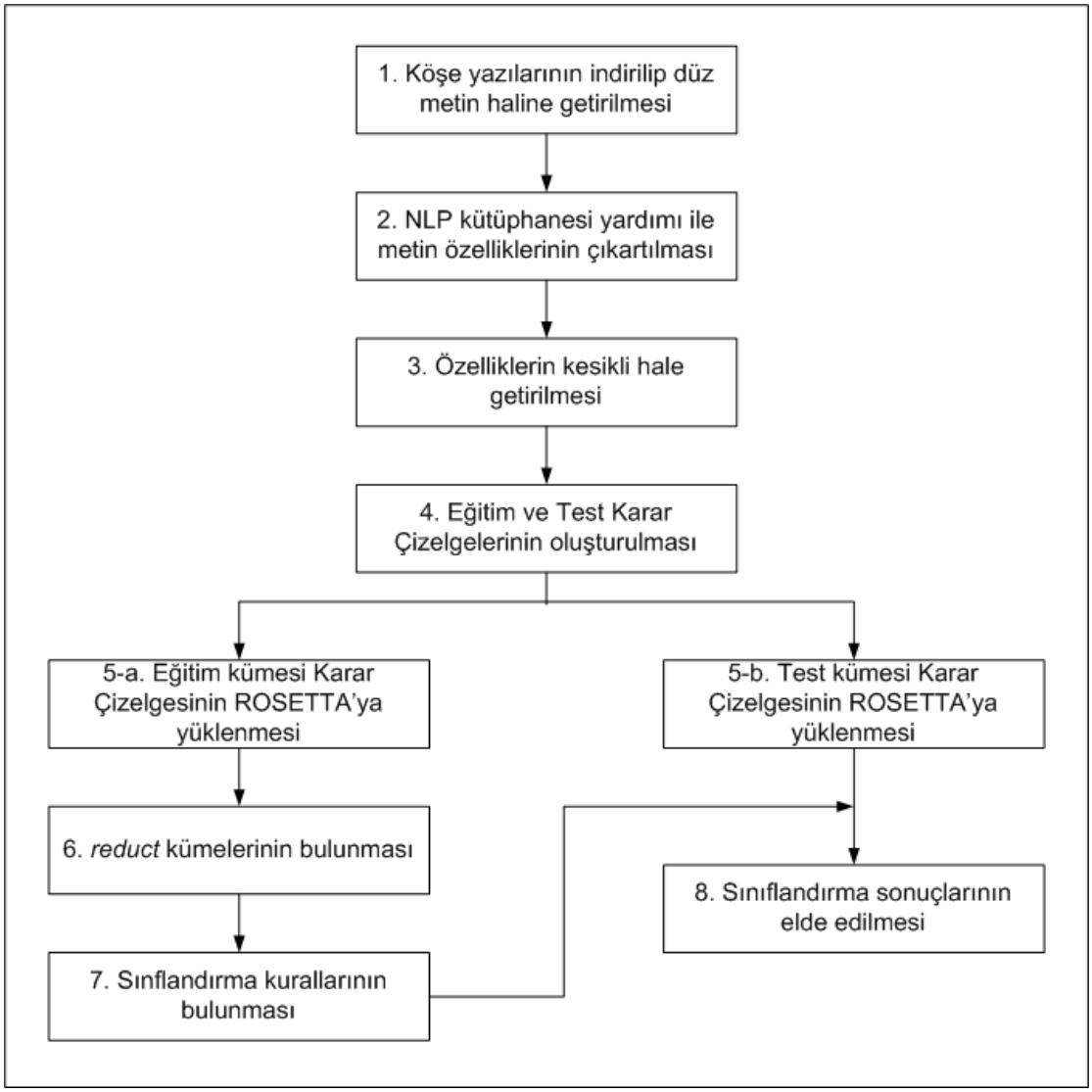
Kaba Küme tabanlı sınıflandırma yazar tanıma problemi bağlamında ilk olarak Stanczyk ve Cyran [4] tarafından kullanılmıştı. Sonrasında aynı ikili benzer çalışmayı dokuz adet noktalama işaretiyle gerçekleştirmiş [5], iki farklı çalışmada da Stanczyk en iyileştirilmiş [165] ya da ağırlıklandırılmış [166] kaba kümeler ile yazar tanıma problemine eğilmişti. Birçok sınıflandırma probleminde yaygın olarak başvurulan kaba kümelere, yazar çözümleme literatüründe bunların dışında rastlamayışımız hem ilginç gelimş, hem de bir motivasyon kaynağı olmuştu.

[177]'de önerdiğimiz model, homojen küçük kümelerin sınıflandırılmasının, heterojen ve daha büyük kümelerin sınıflandırılmasına oranla daha başarılı olabileceği fikrine dayanmaktadır. Belirlenen özelliklere göre özelleştirilmiş sınıflandırıcıları aşamalı bir şekilde kullanarak başarımın arttırılmasını hedefler. Bunun sınanabilmesi için oluşturulan deney setinde, günlük gazetelerin Internet sayfalarından indirilen metinler kullanılmış, 2008 yılının ikinci yarısında yazılmış, iki farklı alandan (yaşam ve siyaset), her yazar için 57 adet olmak üzere 9 yazara ait toplam 513 adet köşe yazısı indirilmiştir. Köşe yazıları HTML içeriklerinden ayrıştırılıp düz metin haline dönüştürüldükten sonra her bir metne ait sözcük türleri, noktalama işaretleri, cümle yapısı gibi toplam adet 34 özellik Türkçe NLP kütüphanesi olan Zemberek [189] yardımıyla çıkartılmıştır.

Köşe yazıları, hangi alanda yazılmış olduklarına göre eğitim ve test kümelerine ayrılmış, Çizelge 2.1'de olduğu gibi karar çizelgeleri şekline dönüştürülerek her bir metin kim tarafından yazıldığı da dahil olmak üzere toplam 35 nitelik ile tanımlanacak şekilde Excel dosyalarına kaydedilmiştir. Bu

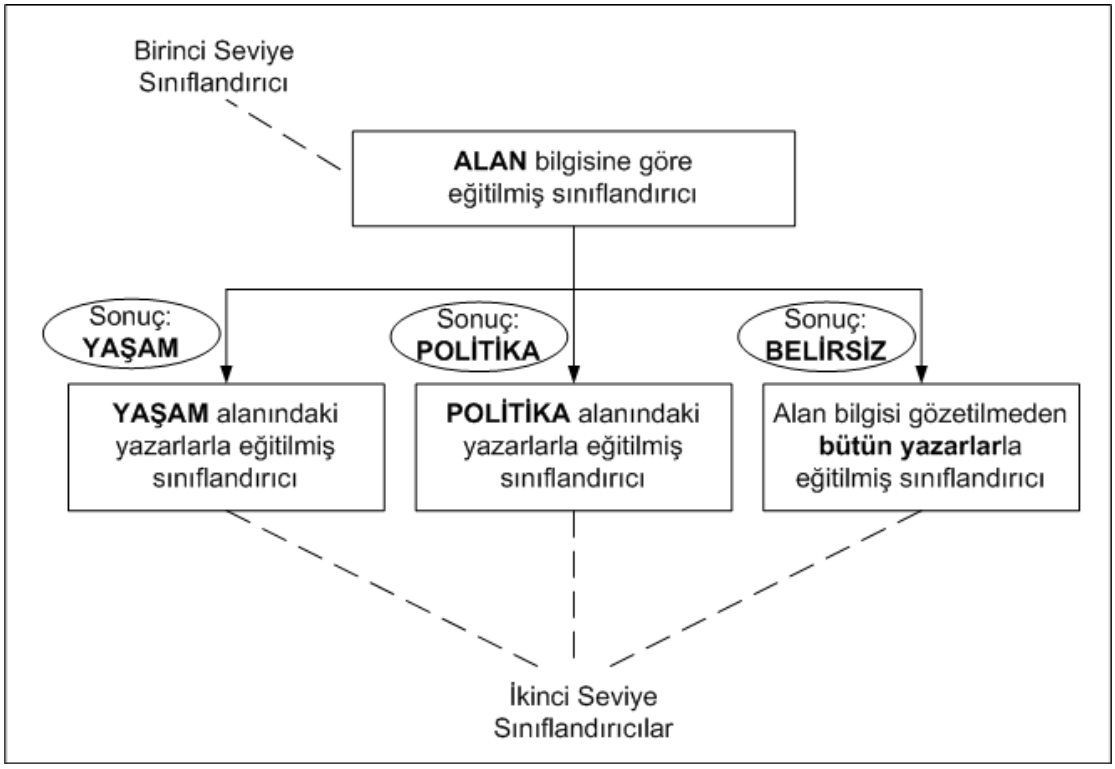
aşamadan sonra kaba küme tabanlı sınıflandırıcıların oluşturulmasında ROSETTA yazılımından faydalanılmıştır.

Daha önce de belirtildiği gibi ROSETTA, Excel dosyalarındaki veriyi alma olanağı sunmaktadır. Ancak kaba kümeler kesikli (*discrete*) değerlerle çalıştığından karar çizelgelerindeki verinin önce kesikli hale getirilmesi (*discretization*) gerekir. Bunun için çeşitli yöntemler bulunmakla birlikte, ortalamanın kullanılması basit bir seçenek olup, çalışmada da bu yöntem tercih edilmiştir. Buna göre her bir nitelik için karar çizelgesindeki değerler, o niteliğe ait ortalamadan yüksek olan değerler için 1, diğerleri için 0 ile güncellenerek veri kesikli hale getirilmiştir. Bundan sonra verinin ROSETTA'ya aktarılması, eğitim, kuralların bulunması ve sınıflandırmanın yapılması aşamaları Şekil 2.5'te görülmektedir.



Şekil 2.5.ROSETTA ile Yazar Tanıma

Çalışmada önerilen modeli sınamak üzere, birinci seviyede metinleri alan bilgisine göre sınıflandıran, ikinci seviyede ise alan bilgisi ile eğitilmiş sınıflandırıcıları kullanarak metinleri yazarlarına göre ayıran iki seviyeli bir yapı oluşturulmuştur (Şekil 2.6).

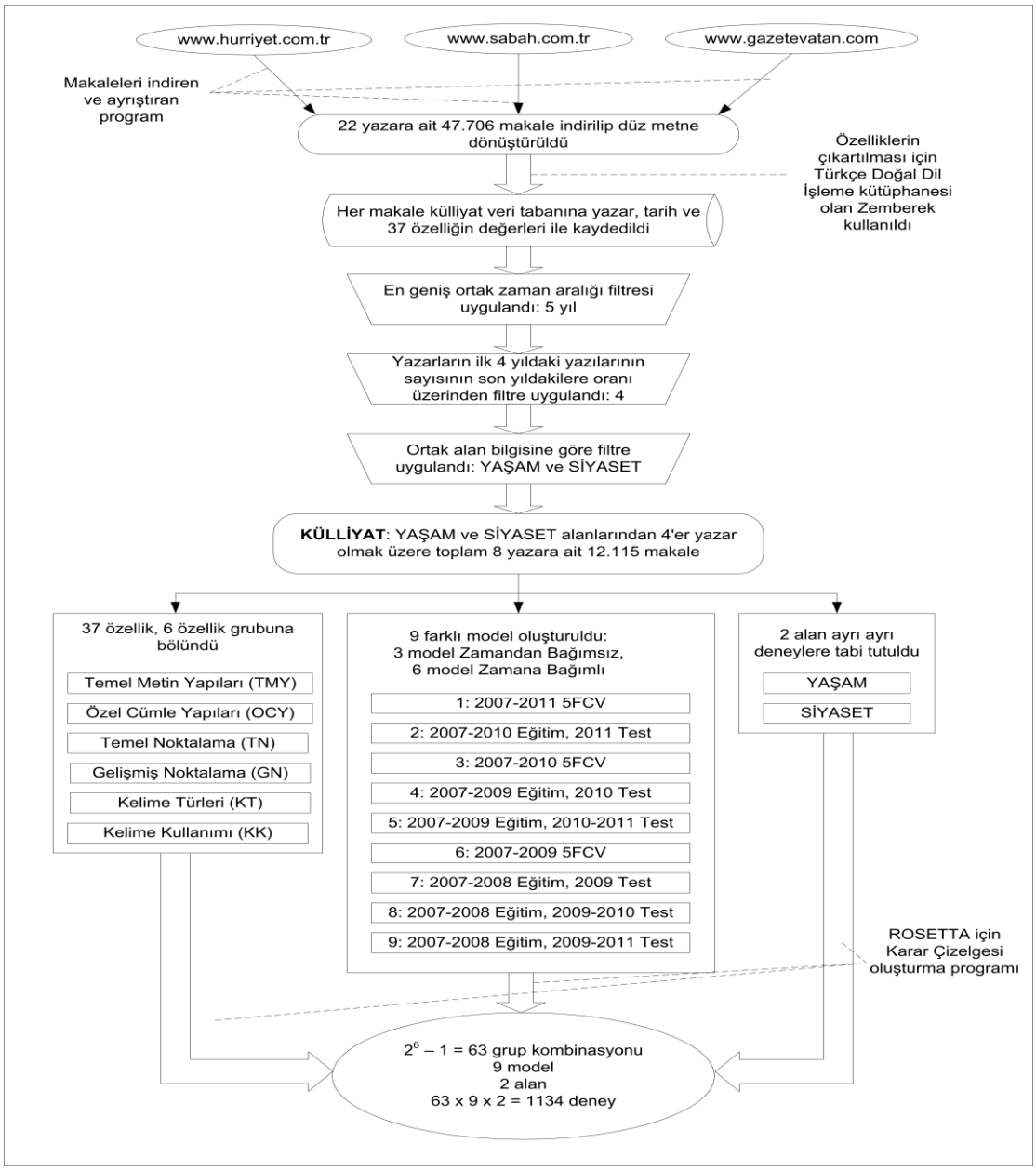


Şekil 2.6. Kaba küme tabanlı sınıflandırıcıların aşamalı olarak kullanılması

Deneylerin sonuçları incelendiğinde, sınıflandırıcıların homojen kümeler için eğitilerek aşamalı bir yapıda kullanılmasının sınıflandırma başarısını yaklaşık olarak %15 arttırdığı görülmüştür.

3. DENEYLER ve SONUÇLAR

Tez çalışması kapsamında binden fazla deney gerçekleştirilmiştir. Bu kesimde külliyatın oluşturulmasından başlayıp deney sonuçlarının alınmasına kadar geçen bütün aşamalarda yapılanlar, deneylerin neden ve nasıl kurgulandığı, kurulan modeller ve bütün bunların yapılabilmesi için kullanılan uygulamalar aktarılacaktır. Yapılanlara ilişkin özet bilgi Şekil 3.1'de verilmiştir.



Şekil 3.1. Külliyatın oluşturulup deneylerin gerçekleştirilmesine ilişkin genel görünüm

3.1 Külliyat (*Corpus*)

Internet'in yaygınlaşmasıyla birlikte birçok araştırmacı külliyat oluşturmak amacıyla çevrimiçi kaynaklara yönelmiş, e-posta, forum, sohbet odaları, kullanıcı yorumları ya da günlük gazeteler gibi Internet üzerinden kolayca erişilebilen veriler yazar tanıma çalışmalarında sıkça kullanılmıştır. Örneğin [113], [114], [138] ve [143]'de Yunan gazeteleri, [131]'de Alman gazeteleri, [157], [163] ve [177]'de Türk gazeteleri kullanılırken [147] ve [170]'de forum iletileri, [177] ve [136]'de e-posta içerikleri, [167]'de ise sohbet içerikleri incelenmiştir.

Bu çalışmaya konu olan külliyat, Hürriyet [190], Sabah [191], ve Vatan [192] gazetelerinin yazar arşivi sayfalarından indirilen köşe yazılarından oluşturulmuştur. 2000 – 2011 yılları arasında yayınlanmış, 22 yazara ait toplam 47706 adet yazı indirilmiş, düz metinlere dönüştürülüp yazar ve tarihlerine göre ayrılarak kaydedilmiştir. Bu noktada karşılaşılan çeşitli zorlukların aşılabilmesi için bir takım uygulamalar geliştirilmiştir. Bu zorluklar özetle:

- Gazetelerin yazar/yazı arşivi sayfaları insan okuyuculara hitap etmektedir. Belirli bir yazarın, belirli bir tarihteki yazısını arayan bir kimse, ekranlardaki yönlendirmeler aracılığıyla o yazıya ulaşıp okuyabilir ya da kaydedebilir. Ancak binlerce yazının bu şekilde indirilip kaydedilmesi söz konusu olduğunda bu işin elle yapılması çok zaman alıcı bir iş haline gelir.
- Yazıların indirilmesini otomatikleştirecek bir uygulamanın, her gazetenin kendi arşiv biçimine göre özelleştirilmesinin gerekli olacağı beklenen bir durumdur. Ancak gazetelerin arşiv sayfalarının kendi içinde yıllara ve hatta bazen aynı yıl içinde aylara göre değişen yapısı nedeniyle, yazıları indirecek olan uygulamanın yalnızca gazetelere göre değil, her gazetenin arşivindeki tarih aralıklarına göre de ele alınması gerekmektedir.
- İndirilen köşe yazıları web sayfası formatındadır. Dolayısıyla, saklanan dosyaların içinde gazetenin arşiv sayfasının düzenini belirleyen menü, reklam, başlık, gazetenin diğer kısımlarına bağlantılar gibi alanlar ile köşe yazısının başlığı, paragraf yapısı, ara başlıkları vb. gibi ögelerini

düzenleyen HTML imleri bulunmaktadır. Bu imlerin ayıklanması işlemini zorlaştıran, sayfa yapısının gazeteden gazeteye değişmesinin yanısıra, aynı gazete için bile gene yıllara ve hatta aylara göre farklılık gösterebilmesidir

- Türkçe metinler UTF-8, ISO-8859-9 ya da Win1254 gibi farklı kodlamalarla saklanabilmektedir. Bu konuda gazeteler ya da bir gazetenin arşivindeki bütün yazılar arasında bir standart olmadığı için, doğru kodlamanın her yazı için ayrı ayrı belirlenmesi gerekmektedir.

Köşe yazıları indirilen yazarlar, hangi alanda yazdıkları, tarih aralıkları ve toplam kaçar tane yazılarının indirilerek ayrıştırıldığı Çizelge 3.1'de görülmektedir.

Çizelge 3.1. Yazıları indirilen köşe yazarları

Yazar	Alan	Başlangıç		Son		Yazı Sayısı
		Yıl	Ay	Yıl	Ay	
Doğan Hızlan	Kültür/Sanat	2000	1	2009	12	2536
Selahattin Duman	Yaşam	2003	1	2011	12	2092
Zülfü Livaneli	Yaşam	2003	1	2011	12	1942
İclal Aydın	Yaşam	2003	1	2011	12	1615
Hıncal Uluç	Yaşam	2003	1	2011	12	2342
Yavuz Donat	Yaşam	2003	1	2011	12	2865
Reha Muhtar	Yaşam	2007	1	2011	12	1430
Oktay Ekşi	Siyaset	2000	1	2009	12	2667
Hadi Uluengin	Siyaset	2000	1	2009	12	2072
Güngör Mengi	Siyaset	2003	1	2011	12	3046
Ruhat Mengi	Siyaset	2003	1	2011	12	2801
Okay Gönensin	Siyaset	2003	1	2011	12	2939
Erdal Şafak	Siyaset	2003	1	2011	12	2806
Emre Aköz	Siyaset	2003	1	2011	12	2236
Mustafa Mutlu	Siyaset	2004	4	2011	12	2425
Can Ataklı	Siyaset	2007	1	2011	12	1481
Nazlı Ilıcak	Siyaset	2007	8	2011	12	1440
Ertuğrul Özkök	Siyaset/Yaşam	2000	1	2009	12	2769
Yalçın Bayer	Siyaset/Yaşam	2000	1	2009	12	2478
Mehmet Barlas	Siyaset/Yaşam	2006	4	2011	12	1685
Engin Ardıç	Siyaset/Yaşam	2008	2	2011	12	1114
Haşmet Babaoğlu	Siyaset/Yaşam	2008	9	2011	12	925

Çizelge 3.1‘e dikkatli bakılırsa yazıların farklı tarih aralıklarından olduğu gözlenecektir. Bu durum, seçilen yazarların diğer tarihlerde başka gazetelerde çalışması, o dönemde yazı yazmaması ya da yazı yazmayı tamamen bırakması gibi farklı nedenlerden kaynaklanmaktadır.

Çözülmeye çalışılan problemlerden ilki Türkçe için ayrıştırıcılığı en yüksek ve en az sayıda özellikten oluşan bir ya da daha fazla özellik kümesinin bulunmasıdır. Dolayısıyla, deneylerin birbirine tarih, alan ve yazılan yazı miktarı gibi kriterler açısından yakın (daha önceki çalışmamıza göndermeyle “homojen”) yazarlar üzerinde yapılması, alan ya da gündem farkı gibi parametrelerin etkisini azaltacaktır. Uğraşılan ikinci problem ise belirlenecek özelliklerin ne kadar süre ile kullanılabilir olacağıdır. Bu konu görüldüğü gibi doğrudan doğruya zaman bilgisiyle ilgilidir.

Dolayısıyla, 22 yazarın 50.000 civarında yazısı indirilmiş olmasına rağmen deneylerin birbirine yakın özelliklerdeki yazar/yazılar üzerinde yapılmasının

sağlanabilmesi için yazarın alanı, yazılarının zaman aralığı ve o aralıktaki yazı adetine bağlı olarak çeşitli eleme işlemleri yapılmıştır.

Tarih Aralığı: Aynı tarih aralığında yazılmış yazıların en geniş kümesi alınmış, geri kalanlar deneylere dahil edilmemiştir. Çizelge 3.1 incelendiğinde en geniş ortak tarih aralığının 2007/01 – 2011/12 olduğu görülür. Bu aralıkta yazıları bulunan 14 yazar, alanları ve yazı adetleri Çizelge 3.2'de verilmiştir.

Çizelge 3.2. 2007/01 – 2011/12 aralığında yazmış yazarlar

Yazar	**Alan**	**2007/01 – 2011/12 Aralığındaki Yazı Sayısı**
Selahattin Duman	Yaşam	1190
Zülfü Livaneli	Yaşam	953
İclal Aydın	Yaşam	834
Hıncal Uluç	Yaşam	1419
Yavuz Donat	Yaşam	1731
Reha Muhtar	Yaşam	1430
Güngör Mengi	Siyaset	1603
Ruhat Mengi	Siyaset	1566
Okay Gönensin	Siyaset	1558
Erdal Şafak	Siyaset	1674
Emre Aköz	Siyaset	1345
Mustafa Mutlu	Siyaset	1502
Can Ataklı	Siyaset	1481
Nazlı Ilıcak	Siyaset	1421
Mehmet Barlas	Siyaset/Yaşam	1190

Yazı Sayısı: Bazı yazarların hergün, bazılarınınsa haftada bir iki gibi sayılarda yazıları yayınlanmaktadır. Dolayısıyla yazarları yazı sayısı cinsinden dengelemek için, çizelgedeki yazı sayıları yerine kullanılabilecek başka bir kritere gereksinim duyulmuştur. Böyle bir kriter olarak beş yıllık zaman diliminin ilk dört yılındaki yazı sayısının son yıldaki yazı sayısına oranı kullanılmış, bu değeri aynı olan yazarlar seçilmiştir.

Aşağıdaki çizelgede yazarların ilk dört yıldaki ve son yıldaki yazılarının sayıları, bunların oranı ve o yazarın deneyler için seçilip seçilmediği görülmektedir.

Çizelge 3.3. Yazarların ilk dört yıldaki yazı sayılarının son yıldakilere oranı

Yazar	Alan	Yazı Sayısı 2007/01 - 2011/12	Yazı Sayısı 2007/01 – 2010/12	Yazı Sayısı 2011/01 – 2011/12	Oran	Seç? E(vet)/ H(ayır)
Selahattin Duman	Yaşam	1190	974	216	4	E
Zülfü Livaneli	Yaşam	953	816	137	5	H
İclal Aydın	Yaşam	834	701	133	5	H
Hıncal Uluç	Yaşam	1419	1161	258	4	E
Yavuz Donat	Yaşam	1731	1403	328	4	E
Reha Muhtar	Yaşam	1430	1156	274	4	E
Güngör Mengi	Siyaset	1603	1325	278	4	E
Ruhat Mengi	Siyaset	1566	1286	280	4	E
Okay Gönensin	Siyaset	1558	1304	254	5	H
Erdal Şafak	Siyaset	1674	1379	295	4	E
Emre Aköz	Siyaset	1345	1085	260	4	E
Mustafa Mutlu	Siyaset	1502	1217	285	4	E
Can Ataklı	Siyaset	1481	1201	280	4	E
Mehmet Barlas	Siyaset/Yaşam	1421	1095	326	3	H

Alan: Çizelge 3.3'te son kolondaki değeri 'E' olan yazarların iki farklı alanda yazdıkları görülmektedir: "Yaşam" ve "Siyaset". Aynı alanda yazan yazarlar gruplanarak deneyler bu iki alan için ayrı ayrı paralel olarak gerçekleştirilecektir. Ancak bu durumda Yaşam alanında yazan yazarlar grubunda 4 yazar varken, Siyaset grubuna 6 yazar düşmektedir. Sayıların eşitlenebilmesi amacıyla Siyaset grubundan, toplam yazı sayısı o grup için en düşük olan Emre Aköz ve Can Ataklı elenmiştir.

Bu elemelerin sonucunda oluşturulan külliyat, Yaşam ve Siyaset alanlarından 4'er, toplamda 8 yazarın 12.115 yazısından oluşmaktadır. Bu yazıların 5770 tanesi Yaşam, 6345 tanesi ise Siyaset alanında yazılmıştır ve külliyatın son hali Çizelge 3.4'te verilmiştir.

Çizelge 3.4. Külliyatın detayları

Alan	Yazar	Yazı Adeti	Toplam Yazı Adeti	Tarih Aralığı Başlangıcı	Tarih Aralığı Sonu
Siyaset	Erdal Şafak	1674	6345	2007/01	2011/12
	Güngör Mengi	1603			
	Ruhat Mengi	1566			
	Mustafa Mutlu	1502			
Yaşam	Yavuz Donat	1731	5770	2007/01	2011/12
	Reha Muhtar	1430			
	Hıncal Uluç	1419			
	Selahattin Duman	1190			

3.2 Özellikler (*Features*)

Yazar tanıma çalışmalarında yazarların birbirlerinden ayırt edilebilmesi için "biçemsel özellik (*stylistic feature*)" ya da "metin özelliği (*text feature*)" adı verilen ve metinlerden çıkartılan özellikler kullanılır. Dolayısıyla özellik çıkartma (*feature extraction*) ve özellik seçme (*feature selection*) yazar tanımanın önemli bir aşamasıdır. Bölüm 1'de aktarılan çalışmalarda da görüldüğü gibi, yazar tanıma çalışmaları kapsamında çok çeşitli özellikler kullanılagelmektedir. Bununla birlikte, [6]'da belirtildiği gibi bine yakın özelliğin farklı araştırmalarda kullanılmış olmasına rağmen, halen üzerinde uzlaşılmış ve genel kabul görmüş bir özellik kümesi oluşturulamamıştır.

İlk yapılan çalışmalarda genellikle tek bir özellik üzerinde durulmuştur. [3]'te sözcük uzunlukları ya da [22]'deki cümle uzunluklarının kullanılması buna örnek verilebilir. Ancak daha sonraki çalışmalarda birden çok özelliğin bir arada kullanılması yaygınlaşmış, paralelinde de birçok çalışmada çok değişkenli istatistiksel çözümleme yöntemlerine başvurulmuştur. Çok sayıda özellik arasından hangilerinin daha iyi sonuç verdiğinin araştırıldığı çalışmalarda, deneye girecek olan özellik kümelerinin kombinasyonlarının kolayca ve sistematik bir biçimde hazırlanabilmesi amacıyla özelliklerin gruplanması yaygın bir işlemdir. Ayrıca, geçmişte yapılmış olan çalışmalara ilişkin incelemelerde de, kullanılan özelliklerin daha anlaşılır bir şekilde aktarılabilmesi için benzer gruplamalara sıklıkla başvurulmaktadır. Yaygın olarak kullanılan 5 tür özellik grubu bulunmaktadır:

- Sözcüksel Özellikler (*Lexical Features*): Karakter ve sözcük tabanlı istatistiksel verilerdir. Toplam harf sayısı, sözcük sayısı, cümle sayısı, sözcük başına düşen ortalama harf sayısı, cümle başına düşen ortalama sözcük sayısı... vb.
- Sözdizimsel Özellikler (*Syntactic Features*): Tür tabanlı istatistiksel verilerdir. Noktalama işaretleri, sözcük dizileri (*n-gram*), sözcük türleri... vb.
- Yapısal Özellikler (*Structural Features*): Metnin genel yapısına ilişkin verilerdir. Yazıtipi özellikleri, başlık kullanımı, metnin içindeki resim ya da bağlantılar (*hyperlink*)... vb.
- İçeriğe Özgü Özellikler (*Content-Specific Features*): İçerik ya da alana özgü, diğer sözcük ya da cümleciklere göre daha önemli olan sözcük ya da cümlecik sayıları... vb.
- Kişiye Özgü Özellikler (*Idiosyncratic Features*): Yanlış sözcük kullanımları ya da gramer hataları... vb.

Bu tez çalışmasında aynı gruplamalar kullanılamamıştır. Deneylerde incelenen metinler gazetelerin web sayfalarından indirildiği için;

- yapısal özellikler kullanılamaz; çünkü bu özellikleri belirleyen yazarın kendisi değil, web sayfasının tasarımcılarıdır,
- kişiye özgü özellikler kullanılamaz; çünkü bu özelliklerin sorumluları yazarın kendisi değil, sayfa editörleri olabilir

Bir takım özelliklerin çıkartılması ise zaman alıcıdır. Örneğin uzun metinler için *hapax legomena* ya da *hapax dislegomena* gibi özellikler bu çerçevede düşünülebilir. Ayrıca Türkçe metinlere özel cümle ya da sözcük yapıları gibi özelliklerin de yazarları ayırt etmede faydası olabilir.

Tez çalışması kapsamında bütün bunlar dikkate alınarak, kolay ve hızlıca çıkarılabilecek, metinlerin Internet kanalından elde edilmiş olmasından etkilenmeyecek, Türkçe'ye has bazı dil kullanım şekillerini kapsayan özellikler üzerinde durulmuştur. Metinlerin ayrıştırılarak özelliklerin çıkartılması sürecinde, Zemberek [189] adlı Doğal Dil İşleme (*Natural Language Processing – NLP*) kütüphanesinden faydalanan bir uygulama geliştirilmiş ve çıkartılan 37 özellik, sonraki işlemlerde istenen kriterlere göre hızlıca

erişilebilmeleri amacıyla tasarlanan bir veri tabanına kaydedilmiştir. Herbir metin için çıkartılan özellikler ve bu özellikleri kolay yönetmek için tanımlanan gruplar Çizelge 3.5'te görülmektedir. Bütün özellikler, seçilen ögelerin metin içinde geçiş sıklıklarından oluşmaktadır.

Çizelge 3.5. Özellikler ve özellik grupları

Özellik	Özellik Grubu Adı	Grup Adı Kısaltması
paragraf sayısı cümle sayısı sözcük sayısı	Temel Metin Yapıları	TMY
edilgen cümle sayısı tek kelimelik cümle sayısı devrik cümle sayısı ünlem cümlesi sayısı	Özel Cümle Yapısı	OCY
nokta sayısı virgül sayısı soru işareti sayısı üç nokta sayısı tek tırnak saysı çift tırnak sayısı ünlem işareti sayısı iki nokta üst üste sayısı noktalı virgül sayısı	Temel Noktalama	TN
tire sayısı alt çizgi sayısı *slash* sayısı ters *slash* sayısı parantez sayısı *ampersand* sayısı	Gelişmiş Noktalama	GN
yansıma kelime sayısı zaman kelimesi sayısı özel isim sayısı kısaltma sayısı yabancı kelime sayısı Osmanlıca kelime sayısı argo kelime sayısı	Kelime Kullanımı	KK
isim sayısı sıfat sayısı fiil sayısı sayı kelimesi sayısı soru kelimesi sayısı edat sayısı bağlaç sayısı mastar sayısı	Kelime Türleri	KT

Yüksek başarımlı en küçük özellik kümesinin bulunabilmesi için bu özelliklerin farklı kombinasyonlarıyla sınıflandırmalar yapılmıştır. Bu noktada tek tek özellikleri kullanmak yerine özellik gruplarının kombinasyonlarının hazırlanması tercih edilmiştir. 6 farklı gruptan oluşturulan $2^6 - 1 = 63$ farklı kombinasyon Çizelge 3.6'da verilmiştir.

Çizelge 3.6. Özellik grubu kombinasyonları

Kombinasyondaki Grup Sayısı	Kombinasyondaki Gruplar	
1	TMY OCY GN TN KT KK	
2	TMY_TN TMY_GN TMY_OCY TMY_KT TMY_KK OCY_GN OCY_TN OCY_KT	OCY_KK GN_KK GN_KT GN_TN TN_KK TN_KT KK_KT
3	TMY_GN_OCY TMY_GN_KT TMY_GN_KK TMY_TN_GN TMY_TN_OCY TMY_TN_KT TMY_TN_KK TMY_KT_OCY TMY_KK_OCY TMY_KK_KT	OCY_KK_KT OCY_GN_KT OCY_GN_KK OCY_TN_GN OCY_TN_KT OCY_TN_KK GN_KK_KT GN_TN_KK GN_TN_KT TN_KK_KT
4	TMY_GN_KT_OCY TMY_GN_KK_OCY TMY_GN_KK_KT TMY_TN_GN_OCY TMY_TN_GN_KT TMY_TN_GN_KK TMY_TN_KT_OCY TMY_TN_KK_OCY	TMY_TN_KK_KT TMY_KK_KT_OCY OCY_GN_KK_KT OCY_TN_KK_KT OCY_TN_GN_KT OCY_TN_GN_KK GN_TN_KK_KT
5	TMY_GN_KK_KT_OCY TMY_TN_GN_KT_OCY TMY_TN_GN_KK_OCY TMY_TN_GN_KK_KT TMY_TN_KK_KT_OCY OCY_TN_GN_KK_KT	
6	TMY_TN_GN_KK_KT_OCY	

3.3 Model Türleri

Bu çalışmanın amaçlarından birisinin de Türkçe yazar tanıma bağlamında belirlenecek olan metin özelliklerinin ne kadar süreyle güvenli bir şekilde kullanılabileceğinin, başka bir ifadeyle zamana karşı ne kadar dirençli olduklarının bulunması olduğu daha önce belirtilmiştir. Seçilecek olan metin özellikleri ile eğitilmiş bir sınıflandırıcının doğru sonuçlar vermesi, özellikle suç teşkil eden durumların incelendiği kimi senaryolarda çok kritik olabilir. Dolayısıyla bir sınıflandırıcının ilk oluşturulduğunda çok başarılı olsa bile

geçen zamanla birlikte artık eskiyip geçerliliğini yitirip yitirmediğinin araştırılması gerekmektedir. Bu çerçevede daha önce yapılan çok az sayıdaki çalışmada [7, 8], uzun yıllara yayılan örneklemler üzerinde gerçekleştirilen incelemeler, gerçekten de zamanla birlikte metinlerde bir değişim olduğunu göstermekle birlikte bu değişimin kısa vadede ne kadar ve nasıl gerçekleştiğini ya da nasıl tespit edilebileceğini belirtmemiştir.

Bu tez kapsamında gerçekleştirilen deneylerde zamanın etkisini gözlemleyebilmek için iki farklı model üzerinde durulmuştur:

Zamandan Bağımsız Model: Örneklem kümeleri seçilirken metinlerin yazıldıkları tarihlerin etkisini ortadan kaldırabilmek için kullanılmış modeldir. Bu modelde eğitim ve test kümeleri, eldeki beş yıllık metinlerin her kesiminden alınan örneklerle oluşturulmakta ve sınıflandırıcının doğruluk değeri (*accuracy*), değişen test kümeleriyle beş kez tekrarlanan sınıflandırma sonuçlarının ortalama doğruluk değeri olarak hesaplanmaktadır. Bunun için beşli çapraz doğrulama (5FCV) tekniği kullanılmıştır. 5 yıllık karışık verinin 1/5'i test verisi olarak seçilip kalan 4/5'inin eğitim kümesini meydana getirdiği deneyler, herbir metin test kümesinde 1 kez bulunacak şekilde 5 kez tekrarlanmış, bu 5 tekrarda oluşturulan sınıflandırıcıların ortalama doğruluk değeri, o deneyin doğruluk değeri olarak alınmıştır. Böylece farklı tarihlerden metinlerin hem eğitim hem de test kümelerinde yer alması sayesinde modelin zamandan bağımsız olması sağlanmıştır.

Zamana Bağımlı Model: Bu modelde zaman bilgisinin aktif olarak kullanılabilmesi için eğitim ve test kümeleri belirlenen tarih aralıklarındaki metinlerden seçilmiştir. Örneğin 5 yıllık aralığın ilk 3 yılı eğitim amaçlı kullanılırken 4. ve 5. yıllara ait metinler test kümesini oluşturmuştur. Bu şekilde farklı tarih aralıkları ile sınıflandırıcılar oluşturularak, yıllar geçtikçe doğruluklarını kaybedip kaybetmedikleri ya da ne oranda kaybettikleri incelenmiştir.

Bu iki model, çeşitli tarih aralıklarından metinler üzerinde yapılacak deneyler için eğitim ve test örneklemlerinin hazırlanmasında kullanılmıştır. Çizelge 3.7'de bunların bir listesi görülmektedir.

Çizelge 3.7. Farklı modeller için oluşturulan örneklemler

Örneklem	Model – Tarih	Örneklem – Zaman İlişkisi
1	2007_2011_5FCV	Zamandan Bağımsız
2	2007_2010_Egitim_2011_Test	Zamana Bağımlı
3	2007_2010_5FCV	Zamandan Bağımsız
4	2007-2009_Egitim_2010_Test	Zamana Bağımlı
5	2007_2009_Egitim_2010_2011_Test	Zamana Bağımlı
6	2007_2009_5FCV	Zamandan Bağımsız
7	2007_2008_Egitim_2009_Test	Zamana Bağımlı
8	2007_2008_Egitim_2009_2010_Test	Zamana Bağımlı
9	2007_2008_Egitim_2009_2011_Test	Zamana Bağımlı

3.4 Deneylerin Gerçekleştirimi

Külliyat, çıkarılan özellikler ve modeller incelendiğinde; 2 farklı alan, 37 metin özelliğinden tanımlanmış 6 farklı grubun 63 kombinasyonu ve 2 farklı model için oluşturulmuş 9 örneklem olduğu görülmektedir. Bu parametrelerle yapılacak toplam deney sayısı (TDS):

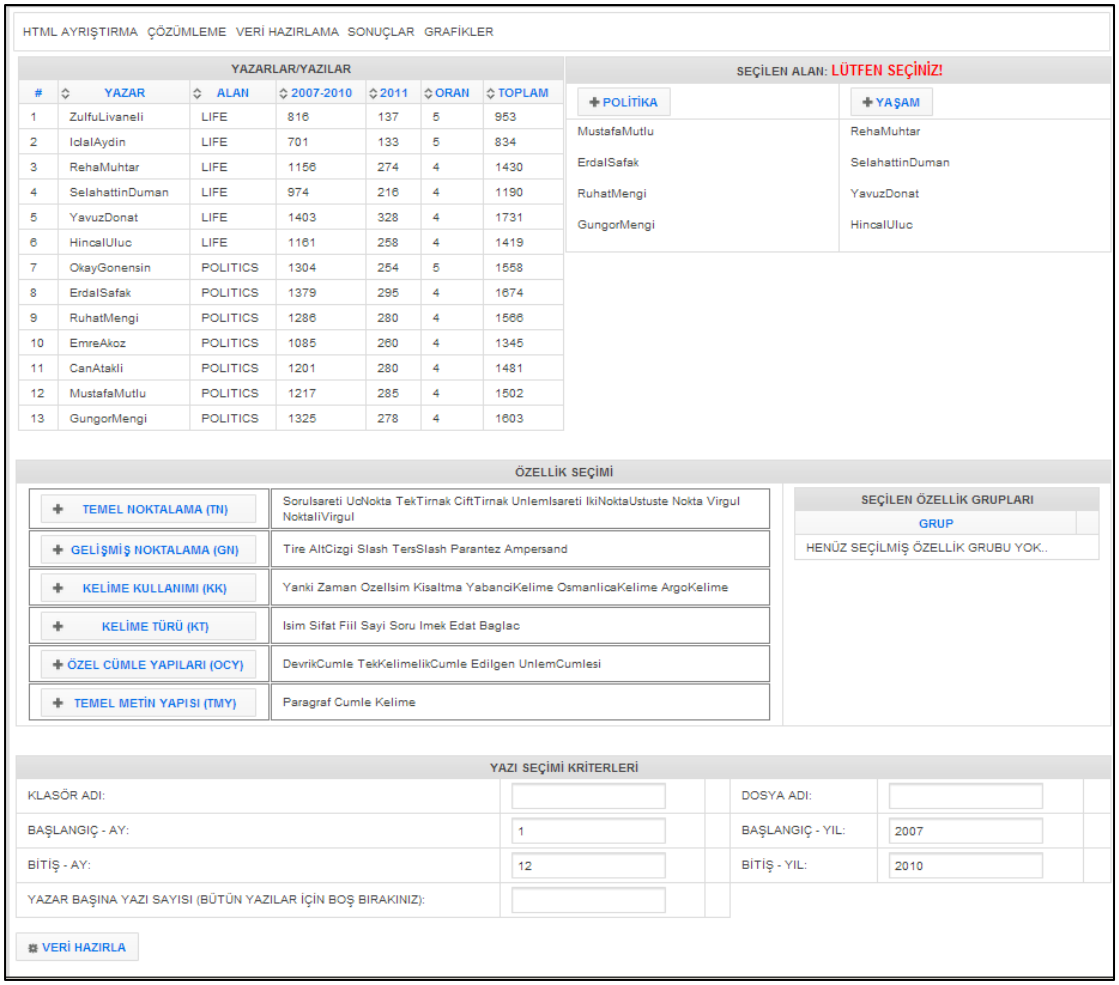
TDS = (Alan sayısı) x (Kombinasyon sayısı) x (Örneklem sayısı) (12)

formülü ile bulunabilir. Buna göre:

$$TDS = 2 \times 63 \times 9 = 1134$$

adet deney gerçekleştirilmiştir.

Deney sayısı fazla olduğu için deney verilerinin uygun alan, özellik ve modele göre hazırlanması elle yönetilemeyecek bir iş olarak değerlendirilmiş ve bu işlemin hatasız ve hızlı olması için bir uygulama geliştirilmiştir. Uygulamadan örnek bir ekran görünümü Şekil 3.2'de verilmiştir.



Şekil 3.2. Deney verisi hazırlama uygulaması ekran görünümü

Ekranın sol üst köşesinde yazarlar, onun hemen sağında ise bu yazarlardan seçilmiş olanlar alanlara göre ayrılmış olarak görülmektedir. Kullanıcı, verisini hazırlamak istediği yazar grubunu buradan seçebilir. Orta kısımda solda özellik grupları ve bu gruplarda hangi özelliklerin olduğu, sağda ise kullanıcının seçtiği özellik grupları bulunmaktadır. Ekranın alt kesiminde ise verilerin seçileceği tarih aralığı, ekranın sağ üstünde belirlenmiş olan yazarlara ait kaçar adet yazının istendiği (eğer kullanıcı bir adet belirtmezse, seçilen tarih aralığındaki bütün yazıları, adet belirtirse de metinlerden rastgele seçilen o kadar yazı) ve oluşturulan verinin saklanacağı klasör ve dosya adlarının girilebileceği alanlar bulunmaktadır.

Bu arayüz kullanıcıya çeşitli deney verileri hazırlama olanağı sunmak amacıyla hazırlanmış, böylece değişik kombinasyonların hızlıca

denenebilmesi sağlanmıştır. Tek tek deneyler için veri hazırlamayı hızlandırıyor olsa da, binden fazla deneyin verisini sistematik bir biçimde hazırlamak için bu uygulama yetersiz bulunmuş, toplu işlem (*batch process*) olarak çalışan ikinci bir modül geliştirilmiştir. Bu modül sayesinde deneylerin herbirisi için; belirlenen alandaki yazarların yazılarından, seçilen özellik grubu kombinasyonunda bulunan özellikler, belirlenen örneklemin tarih aralığına uygun şekilde seçilip, kesikli hale getirilmekte (*discretization*) ve ROSETTA ya da CLROSETTA uygulamalarının gereksinim duyduğu formatta bir karar çizelgesi olarak metin tabanlı dosyalarda saklanmaktadır.

Verinin kesikli hale getirilmesi için Şekil 3.3'teki algoritma kullanılmıştır.

1. Her ***y*** yazarı için
    1.1. Özellik grubu kombinasyonunda bulunan her ***ö*** özelliği için
        1.1.1. Örneklem modeli zamana bağımlı mı?
            1.1.1.1. Evet: yalnızca eğitim kümesindeki yazıları kullanarak ***ö*** özelliğinin ortalamasını bul: $\ddot{o}_{ort}$
            1.1.1.2. Hayır: örneklemdeki bütün yazıları kullanarak ***ö*** özelliğinin ortalamasını bul: $\ddot{o}_{ort}$
    1.2. ***y***'nin örneklemdeki her ***yz*** yazısı için
        1.2.1. Özellik grubu kombinasyonunda bulunan her ***ö*** özelliği için
        1.2.2. $\ddot{o} > \ddot{o}_{ort}$ mi?
            1.2.2.1. Evet: karar çizelgesinde ***yz***'nin ***ö*** özelliği için 1 ver
            1.2.2.2. Hayır: karar çizelgesinde ***yz***'nin ***ö*** özelliği için 0 ver

Şekil 3.3. Kesikli hale getirme algoritması

Bu tez çalışması kapsamında, zamandan bağımsız modelle yapılan deneyler için CLROSETTA tercih edilmiş, zamana bağımlı modelle yapılan deneylerde ROSETTA arayüzü manuel olarak kullanılmıştır. Bunun nedeni CLROSETTA'nın zamandan bağımsız modeller için kullanılan çapraz doğrulama tekniği için uygun olması, ancak bir karar çizelgesinden bulduğu kuralları başka bir karar çizelgesini sınıflandırmak amacıyla kullanamaması,

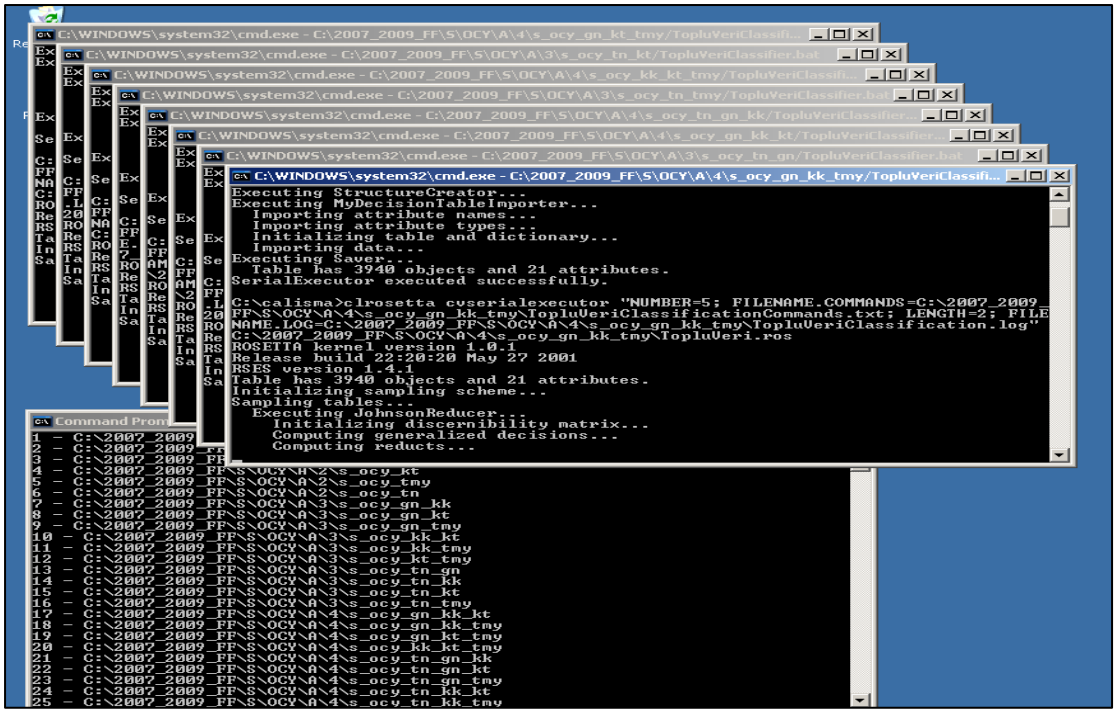
dolayısıyla da bu çalışmadaki zamana bağımlı modeldeki deneyler için uygun olmamasıdır.

CLROSETTA ile yapılan bir deney şu aşamalardan oluşur:

1. indirgeme, kural bulma ve sınıflandırma komutlarını içeren komut dosyanın hazırlanması.
2. komut dosyasının adını, dosyadaki komutların hangi karar çizelgesi üzerinde, kaçlı doğrulama yapılarak işletileceği ve çıktı dosyalarının hangi adla nerede saklanacağı bilgilerinin parametre olarak verilerek CLROSETTA programının çalıştırılması.

Şekil 3.4. CLROSETTA ile deney gerçekleştirim adımları

1134 deneyin herbiri için o deneyin karar çizelgesini içeren dosya adıyla bir komut dosyası hazırlanması ve sonra bu dosyanın parametre olarak verildiği bir komutun elle yazılıp işletilmesi doğru değildir. Dolayısıyla bu işlemleri otomatik hale getirecek bir program geliştirilmiştir. Program herbir deney için önce gerekli komut dosyasını hazırlamakta, sonra da bu dosyayı CLROSETTA'ya diğer parametrelerle birlikte vererek deneyi başlatan toplu işlem komutlarını paralel bir şekilde çalıştırmaktadır. Şekil 3.5'te örnek ekran görünümü verilmiştir.



Şekil 3.5. CLROSETTA ile deneylerin paralel olarak çalıştırılması

ROSETTA ile zamana bağımlı modeldeki bir deneyin yapılması aşamaları ise Şekil 3.6'da verilmiştir:

1. eğitim karar çizelgesinin yüklenmesi
2. indirgeme algoritması seçilerek eğitim karar çizelgesinin indirgenmesi (*reduct* bulma)
3. kuralların bulunması
4. test karar çizelgesinin yüklenmesi
5. test karar çizelgesi üzerinde işletilecek (3. adımda bulunmuş olan) kuralların seçilmesi
6. sonuçların kaydedileceği günlük dosyasının seçilerek sınıflandırmanın başlatılması

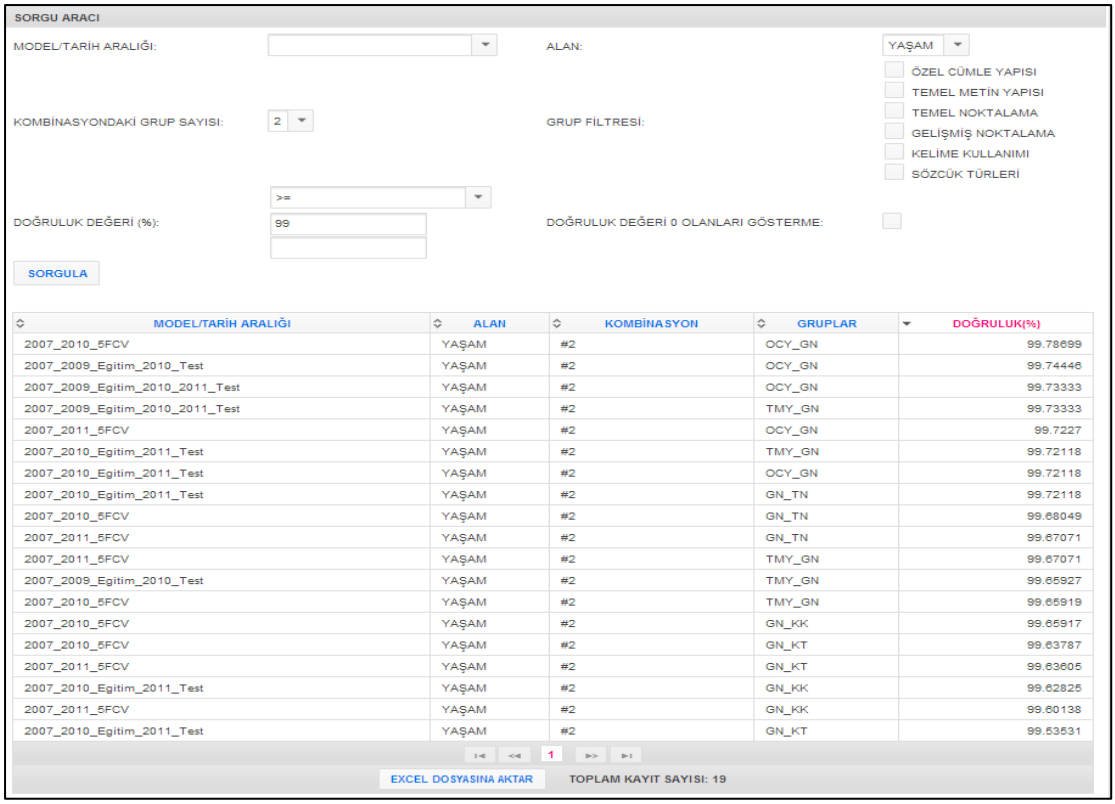
Şekil 3.6. ROSETTA ile deney yapma adımları

Zamana bağımlı model ile yapılan deneylerde bu adımlar manuel olarak gerçekleştirilmiştir.

3.5 Deney Sonuçları

ROSETTA ve CLROSETTA programları, yapılan bir sınıflandırmaya ilişkin sonuçları adı kullanıcı tarafından verilen metin tabanlı günlük dosyalarında saklayabilmektedir. Saklanan sonuçlar içerisinde, hangi deneyler için hangi sonuçların elde edildiği tek tek belirtilmekte, ayrıca doğru ya da yanlış olarak sınıflandırılan ya da hiç sınıflandırılamayan ögelerin yüzde cinsinden doğruluk değerleri de topluca verilmektedir. K-katlı Çapraz doğrulama ile yapılan deneylerde ise deneyin doğruluk değeri, k adet deneyin sonuçlarının ortalaması olarak kaydedilmektedir. İşte bu sonuçların dosyaların içerisinden alınıp detaylıca incelenebilmesi ve karşılaştırılabilmesi için, daha önce geliştirilen programlara yeni modüller eklenmiştir:

- Günlük Dosyası Ayrıştırıcısı (*Log File Parser*): Sonuçları içeren günlük dosyalarını tarayıp sonuçları toparlayarak daha sonra kolayca incelenebilmeleri için tasarlanan veri tabanına kaydeder.
- Sorgu Aracı (*Query Tool*): Deney sonuçları üzerinde çeşitli kriterler ile sorgular işletilerek sorgu sonuçlarının çizelge formatında alınmasını, istenirse Excel dosyalarına aktarılmasını sağlar. Sorgu Aracı'na ait bir ekran görünümü Şekil 3.7'de verilmiştir.
- Grafik Aracı (*Chart Tool*): Deneylerin birbirleriyle karşılaştırmalı grafiklerinin hazırlanmasında kullanılır. Örnek bir ekran görünümü Şekil 3.8'de görülmektedir.



Şekil 3.7. Sorgu Aracı ekran görünümü

Şekil 3.8. Grafik Aracı ekran görünümü

Bu kesimde incelenecek olan çizelge ve grafikler bu modüller yardımı ile oluşturulmuştur. Grafikler deney sonuçlarını birbirleriyle karşılaştırmak açısından faydalı olmakla birlikte, özellik grubu kombinasyonları ve doğruluk değerlerinin incelenmesinde çizelgeler daha başarılıdır. Dolayısıyla deney sonuçları verilirken çizelgelerden, sonuçlar karşılaştırılırken de grafiklerden faydalanılacaktır.

Deney sonuçları Çizelge 3.8 ve Çizelge 3.9'da toplu halde verilmiştir. Şekil 3.9 ve Şekil 3.10'da ise bu deney sonuçları grafikler üzerinde gösterilmiştir.

Çizelge 3.8. YAŞAM alanında yapılan deneylerin doğruluk değerleri

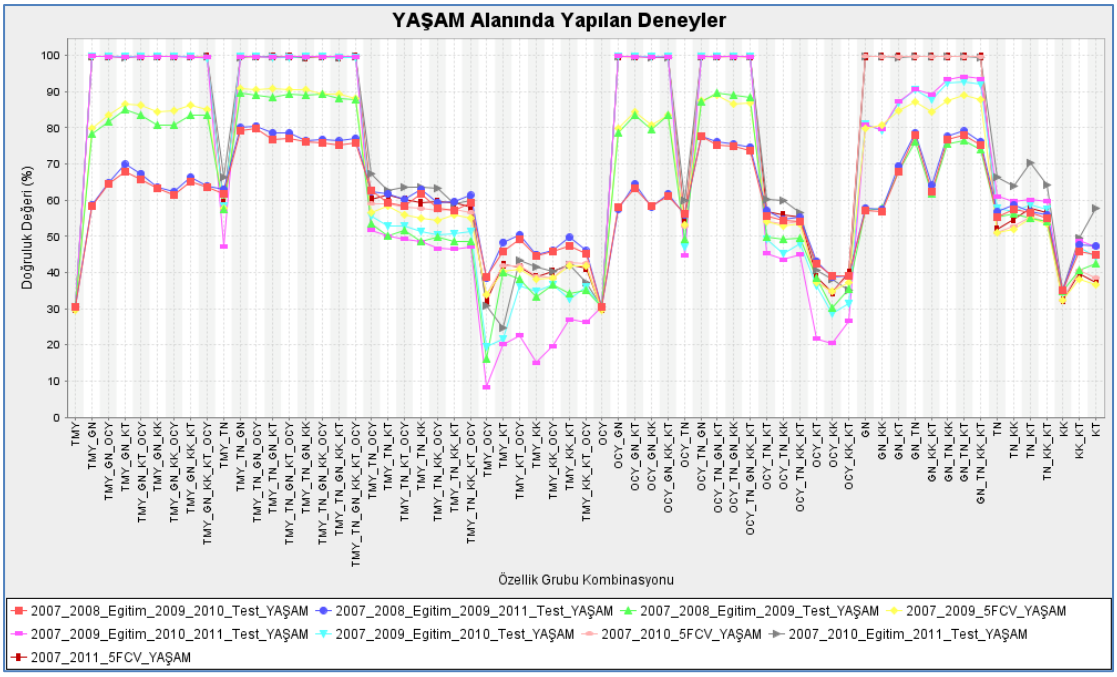
YAŞAM Alanında Yapılan Deneylerin Doğruluk Değerleri (%)											
Özellik Grubu Kombinasyonu	**Örneklem**									**Zamandan Bağımsız ve Zamana Bağımlı Modellerin Ortalamaları**	
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**ZB'sız**	**ZB'lı**
TMY	30,00	30,48	29,89	30,49	30,49	29,69	30,58	30,54	30,52	29,86	30,52
TMY_GN	99,67	99,72	99,66	99,66	99,73	79,91	78,28	58,26	58,73	93,08	82,40
TMY_KK	38,75	41,54	38,92	34,92	15,07	38,07	33,39	44,68	45,04	38,58	35,77
TMY_KT	41,99	24,72	41,90	21,55	20,13	40,17	40,12	45,91	48,39	41,36	33,47
TMY_OCY	31,87	30,86	33,79	19,51	8,36	33,95	16,18	38,67	38,38	33,20	25,32
TMY_TN	59,32	66,26	57,63	58,35	47,20	57,76	57,41	61,71	63,08	58,24	59,00
TMY_GN_KK	99,65	99,72	99,66	99,57	99,69	84,35	80,83	63,20	63,55	94,55	84,43
TMY_GN_KT	99,67	99,54	99,64	99,66	99,56	86,48	85,01	67,97	69,86	95,26	86,93
TMY_GN_OCY	99,71	99,72	99,70	99,74	99,60	83,44	81,52	64,61	64,72	94,28	84,98
TMY_KK_KT	42,01	42,10	42,59	32,71	27,07	41,82	34,07	47,36	49,65	42,14	38,83
TMY_KK_OCY	40,31	40,33	38,39	36,71	19,73	38,52	36,63	45,66	46,03	39,07	37,52
TMY_KT_OCY	41,39	43,22	41,88	36,37	22,62	40,99	38,25	49,02	50,38	41,42	39,98
TMY_TN_GN	99,60	99,54	99,60	99,74	99,64	90,91	89,61	79,22	80,02	96,70	91,30
TMY_TN_KK	59,29	63,57	57,67	51,28	48,44	54,89	48,47	61,58	62,91	57,28	56,04
TMY_TN_KT	61,37	62,64	58,97	52,81	50,04	58,27	50,17	59,24	61,65	59,53	56,09
TMY_TN_OCY	60,23	67,38	58,99	55,71	51,73	56,65	53,49	62,56	62,44	58,62	58,89
TMY_GN_KK_KT	99,64	99,63	99,57	99,66	99,51	86,25	83,56	65,03	66,36	95,15	85,62
TMY_GN_KK_OCY	99,67	99,72	99,68	99,57	99,60	84,86	80,75	61,54	62,27	94,74	83,91
TMY_GN_KT_OCY	99,71	99,72	99,64	99,57	99,69	86,16	83,56	65,80	67,38	95,17	85,95
TMY_KK_KT_OCY	41,09	37,08	42,52	36,03	26,31	41,70	35,09	45,32	45,97	41,77	37,63
TMY_TN_GN_KK	99,57	99,44	99,66	99,57	99,69	90,68	88,93	76,02	76,43	96,64	90,01
TMY_TN_GN_KT	99,64	99,63	99,55	99,32	99,51	90,85	88,50	76,87	78,59	96,68	90,40
TMY_TN_GN_OCY	99,64	99,63	99,55	99,57	99,64	90,45	89,01	79,81	80,52	96,55	91,37
TMY_TN_KK_KT	59,22	59,01	57,39	50,68	46,53	55,80	48,55	57,20	59,46	57,47	53,57
TMY_TN_KK_OCY	59,69	63,29	57,16	50,51	46,67	54,35	49,74	57,84	59,40	57,06	54,58
TMY_TN_KT_OCY	60,16	63,48	58,16	52,81	49,24	56,05	51,53	58,35	60,08	58,12	55,91
TMY_GN_KK_KT_OCY	99,64	99,63	99,64	99,23	99,56	85,11	83,56	63,63	63,99	94,80	84,93
TMY_TN_GN_KK_KT	99,57	99,54	99,55	99,49	99,60	89,18	88,16	75,21	76,29	96,10	89,71
TMY_TN_GN_KK_OCY	99,62	99,63	99,53	99,66	99,64	89,23	89,27	75,81	76,72	96,13	90,12
TMY_TN_GN_KT_OCY	99,60	99,63	99,57	99,49	99,64	90,60	89,27	77,17	78,48	96,59	90,61
TMY_TN_KK_KT_OCY	58,08	60,04	56,65	51,45	47,11	55,14	48,64	59,33	61,27	56,62	54,64
TMY_TN_GN_KK_KT_OCY	99,62	99,63	99,53	99,40	99,64	88,15	87,65	75,72	77,04	95,77	89,85
OCY	30,00	30,48	29,89	30,49	30,49	29,69	30,58	30,54	30,52	29,86	30,52
OCY_GN	99,72	99,72	99,79	99,74	99,73	79,77	78,71	58,01	57,59	93,09	82,25
OCY_KK	34,45	38,01	34,47	28,62	20,40	34,89	30,15	39,14	38,61	34,60	32,49
OCY_KT	38,86	40,71	37,62	36,37	21,69	37,22	38,59	42,29	43,14	37,90	37,13
OCY_TN	53,95	60,04	52,73	47,10	44,67	53,27	49,06	56,13	55,84	53,32	52,14
OCY_GN_KK	99,60	99,54	99,70	99,57	99,56	80,71	79,39	58,30	58,18	93,34	82,42
OCY_GN_KT	99,60	99,72	99,49	99,66	99,64	84,43	83,39	63,12	64,46	94,51	85,00
OCY_KK_KT	39,91	35,22	39,52	31,26	26,53	37,30	35,35	39,10	39,05	38,91	34,42
OCY_TN_GN	99,62	99,63	99,57	99,57	99,64	87,59	87,31	77,51	77,66	95,59	90,22

OCY_TN_KK	55,81	59,76	53,43	45,23	43,60	52,76	49,32	54,43	54,79	54,00	51,19
OCY_TN_KT	56,97	60,22	54,13	49,23	45,20	53,95	49,83	55,75	57,07	55,02	52,88
OCY_GN_KK_KT	99,65	99,54	99,64	99,66	99,60	83,61	83,56	61,24	61,74	94,30	84,22
OCY_TN_GN_KK	99,69	99,63	99,62	99,74	99,64	86,53	89,01	74,79	75,41	95,28	89,70
OCY_TN_GN_KT	99,62	99,63	99,55	99,66	99,60	89,06	89,69	75,34	76,11	96,08	90,01
OCY_TN_KK_KT	55,32	56,51	54,05	47,96	44,93	53,35	49,40	54,17	55,23	54,24	51,37
OCY_TN_GN_KK_KT	99,58	99,63	99,51	99,74	99,64	86,99	88,50	73,64	74,68	95,36	89,31
GN	99,72	99,72	99,72	81,09	80,84	79,72	56,98	57,16	57,83	93,05	72,27
GN_KK	99,60	99,63	99,66	79,30	79,56	80,77	57,67	56,90	57,56	93,34	71,77
GN_KT	99,64	99,54	99,64	86,71	87,29	84,63	67,80	67,89	69,54	94,63	79,79
GN_TN	99,67	99,72	99,68	90,55	90,62	87,13	76,06	77,94	78,56	95,49	85,58
GN_KK_KT	99,62	99,63	99,64	87,82	89,24	84,32	61,67	62,44	64,05	94,52	77,47
GN_TN_KK	99,69	99,63	99,68	92,50	93,33	87,41	75,55	76,62	77,57	95,59	85,87
GN_TN_KT	99,65	99,63	99,60	92,76	94,09	88,98	76,49	77,90	79,06	96,08	86,65
GN_TN_KK_KT	99,64	99,54	99,66	92,08	93,56	87,90	73,85	75,09	76,11	95,73	85,04
TN	51,92	66,36	50,96	57,75	60,98	50,94	55,28	55,37	56,72	51,27	58,74
TN_KK	54,64	63,94	52,71	56,90	59,69	51,82	56,22	57,54	58,67	53,06	58,83
TN_KT	57,80	70,35	55,26	58,52	60,18	55,03	55,11	56,43	56,78	56,03	59,56
TN_KK_KT	56,53	64,22	54,03	57,58	59,69	53,49	54,00	54,86	55,93	54,68	57,71
KK	32,20	35,22	32,19	34,33	34,22	32,36	34,75	35,14	35,08	32,25	34,79
KK_KT	39,60	49,63	40,56	47,02	48,80	38,07	40,63	45,70	47,66	39,41	46,57
KT	37,24	57,62	38,48	44,38	47,16	36,62	42,50	45,02	47,31	37,45	47,33

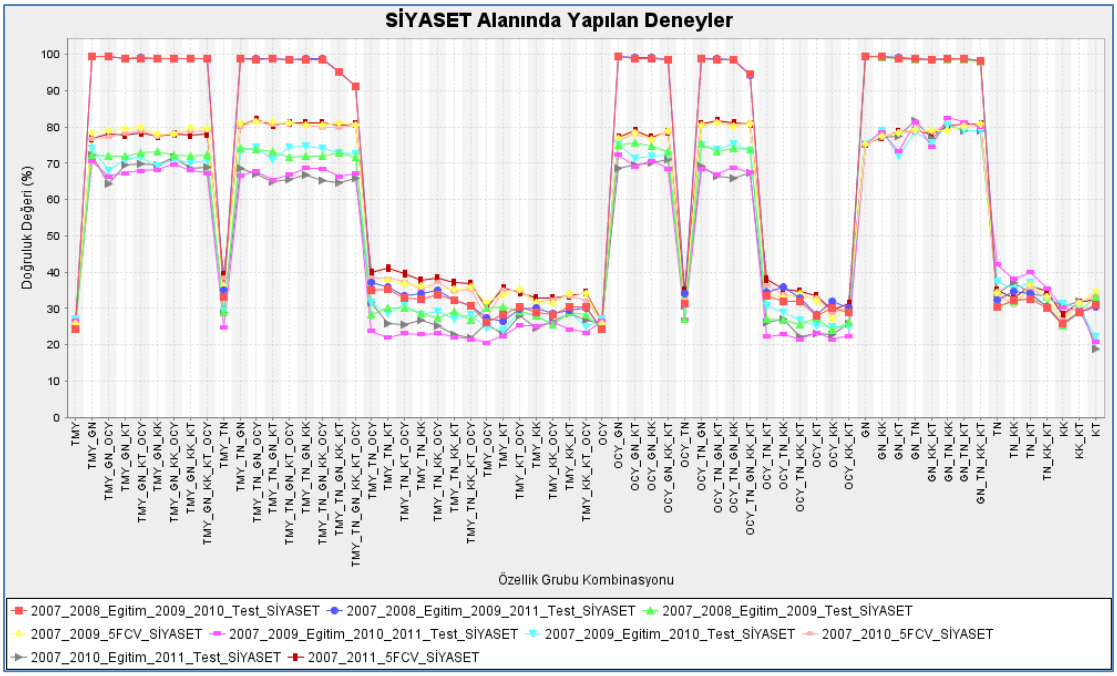
Çizelge 3.9. SİYASET alanında yapılan deneylerin doğruluk değerleri

SİYASET Alanında Yapılan Deneylerin Doğruluk Değerleri (%)											
Özellik Grubu Kombinasyonu	**Örneklem**									**Zamandan Bağımsız ve Zamana Bağımlı Modellerin Ortalamaları**	
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**ZB'sız**	**ZB'lı**
TMY	26,38	25,92	26,48	26,99	26,49	25,69	24,21	24,34	24,36	26,18	25,38
TMY_GN	76,75	72,32	76,72	74,11	70,52	78,27	72,31	99,45	99,49	77,25	81,37
TMY_KK	32,81	24,52	31,59	27,94	25,07	31,60	27,92	28,83	30,23	32,00	27,42
TMY_KT	35,49	22,67	35,11	23,99	22,41	33,91	30,39	28,20	26,39	34,84	25,68
TMY_OCY	30,12	25,92	27,27	24,55	20,50	31,55	30,16	26,13	27,39	29,65	25,78
TMY_TN	39,20	28,82	38,37	29,04	24,66	36,35	29,00	33,20	35,13	37,97	29,98
TMY_GN_KK	77,35	69,60	77,32	69,06	68,19	78,20	73,32	98,75	98,78	77,62	79,62
TMY_GN_KT	77,86	69,51	78,36	70,96	67,32	79,59	72,08	98,87	98,97	78,60	79,62
TMY_GN_OCY	77,98	64,24	77,22	68,03	66,36	78,86	72,00	99,34	99,38	78,02	78,22
TMY_KK_KT	33,35	28,47	33,36	28,97	24,28	33,93	28,77	30,27	29,45	33,55	28,37
TMY_KK_OCY	32,96	26,54	32,46	25,49	26,03	31,35	25,37	28,36	28,61	32,25	26,73
TMY_KT_OCY	34,33	28,03	34,90	29,12	25,28	35,30	29,31	30,39	29,94	34,84	28,68
TMY_TN_GN	80,60	68,72	80,12	73,56	66,69	81,04	74,01	98,79	98,92	80,59	80,12
TMY_TN_KK	37,71	26,80	35,26	28,26	22,79	36,19	28,62	32,62	34,13	36,39	28,87
TMY_TN_KT	41,06	25,75	38,31	28,65	21,91	37,79	30,24	35,23	35,86	39,05	29,61
TMY_TN_OCY	39,75	31,63	38,39	31,49	23,78	37,11	28,38	35,00	37,24	38,41	31,25
TMY_GN_KK_KT	77,73	68,63	78,72	70,09	67,98	79,80	71,85	98,91	98,97	78,75	79,40

TMY_GN_KK_OCY	78,12	71,79	78,32	71,03	69,73	78,20	72,16	98,79	98,78	78,21	80,38
TMY_GN_KT_OCY	78,44	69,86	78,97	71,74	67,86	79,85	72,78	98,98	99,05	79,09	80,05
TMY_KK_KT_OCY	34,47	26,71	32,15	24,94	23,33	34,19	28,15	30,31	30,15	33,60	27,27
TMY_TN_GN_KK	81,10	66,87	80,60	74,59	68,61	80,36	72,08	98,63	98,76	80,69	79,92
TMY_TN_GN_KT	80,61	65,03	80,51	71,03	65,49	81,57	73,09	98,79	98,86	80,90	78,71
TMY_TN_GN_OCY	82,16	67,14	81,53	74,27	67,73	81,42	73,70	98,55	98,73	81,70	80,02
TMY_TN_KK_KT	37,27	22,85	34,57	26,99	22,04	35,41	29,16	32,34	32,37	35,75	27,62
TMY_TN_KK_OCY	38,39	25,13	37,12	29,28	23,16	35,28	27,22	33,83	34,88	36,93	28,92
TMY_TN_KT_OCY	39,57	25,40	37,28	30,94	23,16	36,73	30,01	32,97	33,40	37,86	29,31
TMY_GN_KK_KT_OCY	78,19	68,98	79,09	70,72	67,32	79,49	72,31	98,87	98,95	78,92	79,52
TMY_TN_GN_KK_KT	80,47	64,76	79,84	72,77	66,36	81,02	72,78	95,12	95,19	80,44	77,83
TMY_TN_GN_KK_OCY	81,18	65,29	79,80	73,95	68,52	80,63	72,00	98,63	98,73	80,54	79,52
TMY_TN_GN_KT_OCY	81,23	65,64	80,97	74,35	66,78	81,02	71,77	98,48	98,54	81,07	79,26
TMY_TN_KK_KT_OCY	36,69	21,97	35,26	28,26	21,54	35,81	26,68	30,74	30,69	35,92	26,65
TMY_TN_GN_KK_KT_OCY	80,82	65,82	80,64	72,45	67,19	80,53	71,62	91,05	91,18	80,66	76,55
OCY	26,38	25,92	26,48	26,99	26,49	25,69	24,21	24,34	24,36	26,18	25,38
OCY_GN	77,01	68,63	76,28	75,45	72,31	76,98	74,86	99,45	99,51	76,76	81,70
OCY_KK	29,52	22,67	28,86	24,78	21,50	27,03	24,13	30,23	31,94	28,47	25,88
OCY_KT	33,40	22,93	32,86	25,10	23,37	32,13	27,46	28,05	28,15	32,80	25,84
OCY_TN	35,13	30,67	32,36	30,62	26,78	31,75	26,99	31,37	34,18	33,08	30,10
OCY_GN_KK	77,15	70,21	76,57	72,06	70,64	76,62	74,79	98,98	99,00	76,78	80,95
OCY_GN_KT	78,91	69,60	78,09	71,19	69,06	78,50	75,56	98,95	99,05	78,50	80,57
OCY_KK_KT	31,17	25,83	29,98	24,78	22,37	30,03	26,22	29,02	30,23	30,39	26,41
OCY_TN_GN	80,96	69,16	80,26	75,06	68,48	80,38	75,33	98,79	98,95	80,53	80,96
OCY_TN_KK	35,21	26,98	33,90	28,97	22,87	32,94	26,76	32,07	35,88	34,02	28,92
OCY_TN_KT	37,92	26,01	36,39	30,70	22,29	34,19	26,99	33,52	34,40	36,17	28,98
OCY_GN_KK_KT	78,64	71,18	78,53	72,22	68,57	78,83	73,32	98,59	98,59	78,67	80,41
OCY_TN_GN_KK	81,23	65,91	80,47	75,30	68,77	80,15	74,17	98,63	98,67	80,62	80,24
OCY_TN_GN_KT	81,58	66,43	81,31	73,88	67,03	81,07	73,09	98,63	98,81	81,32	79,64
OCY_TN_KK_KT	34,75	22,06	33,07	26,84	21,46	34,01	25,52	31,99	32,69	33,94	26,76
OCY_TN_GN_KK_KT	80,77	67,40	81,20	73,40	67,40	80,69	73,78	94,53	94,21	80,89	78,45
GN	75,37	75,40	75,21	75,30	75,43	75,30	99,30	99,45	99,51	75,29	87,40
GN_KK	77,27	77,07	77,53	78,93	78,54	77,59	99,07	99,30	99,30	77,46	88,70
GN_KT	78,68	77,50	78,45	72,06	73,22	78,48	98,76	98,98	99,03	78,54	86,59
GN_TN	79,01	81,63	78,61	78,77	81,21	79,21	98,61	98,87	98,95	78,94	89,67
GN_KK_KT	77,92	76,89	78,80	75,61	74,59	79,11	98,53	98,59	98,57	78,61	87,13
GN_TN_KK	80,30	80,40	80,87	80,51	82,54	79,06	98,45	98,75	98,81	80,08	89,91
GN_TN_KT	80,88	78,91	80,47	79,08	81,33	80,20	98,53	98,75	98,86	80,52	89,25
GN_TN_KK_KT	80,91	78,91	80,28	78,85	80,04	80,86	97,91	98,20	98,30	80,68	88,70
TN	34,96	33,92	33,82	37,41	42,12	34,29	30,55	30,35	32,37	34,36	34,45
TN_KK	32,92	36,99	30,86	35,20	38,09	31,73	32,02	32,38	34,67	31,84	34,89
TN_KT	36,25	34,01	35,47	37,02	40,00	36,62	32,71	32,50	34,15	36,11	35,07
TN_KK_KT	34,72	33,22	32,67	32,52	35,51	33,30	30,94	30,04	30,50	33,56	32,12
KK	28,12	30,49	26,87	31,18	30,15	25,84	25,21	25,66	25,72	26,94	28,07
KK_KT	31,73	31,99	31,82	30,15	29,52	31,32	29,16	28,79	28,93	31,62	29,76
KT	32,45	18,89	33,42	22,18	20,62	34,70	33,18	30,98	30,48	33,52	26,05



Şekil 3.9. YAŞAM alanında yapılan deneylerin doğruluk değerleri grafiği



Şekil 3.10. SİYASET alanında yapılan deneylerin doğruluk değerleri grafiği

3.5.1 Özelliklerin Ayırıcılığı

Şekil 3.9 ve Şekil 3.10 incelendiğinde şu önemli nokta kolayca görülmektedir: metinlerin ait olduğu alana, örneklem modeline ve tarih aralığına bağlı olarak deneylerin başarı oranları değişse de, en yüksek ya da en düşük sonuçlar hep aynı özellik grupları tarafından alınmıştır. Birçok noktada sonuçlar örtüşmekte ya da birbirlerine çok yakın bulunmaktadır. Bu durum, bazı özellik gruplarının diğerlerine göre daha yüksek ya da daha düşük oranda ayrıştırıcı olduğunu, ayrıca bu ayrıştırıcılığın alan, model ya da tarih aralığından bağımsız olduğunu göstermektedir. Bu noktada, hangi özellik grup ya da gruplarının daha ayrıştırıcı olduğunu belirlemek amacıyla, %99,5 üzerinde doğrulukla sonuçlanan deneylerde kullanılan özellik grubu kombinasyonları gösteren Çizelge 3.10 hazırlanmıştır. Çizelgedeki 32 farklı özellik grubu kombinasyonu toplam 139 deneyde %99,5 ve daha yüksek doğruluklara erişmiş, ancak çizelgeye yalnızca en yüksek değerleri ile alınmışlardır.

Çizelge 3.10. %99.5 üzerinde başarıyla sonuçlanan deneylerde kullanılan özellik grupları ve eriştikleri en yüksek doğruluk değerleri

Özellik Grubu Kombinasyonu	Doğruluk Değeri (%)
OCY_GN	99,79
TMY_TN_GN	99,74
TMY_GN_OCY	99,74
OCY_TN_GN_KK	99,74
OCY_TN_GN_KK_KT	99,74
TMY_GN	99,73
GN	99,72
GN_TN	99,72
TMY_GN_KT_OCY	99,72
OCY_GN_KT	99,72
TMY_GN_KK_OCY	99,72
TMY_GN_KK	99,72
OCY_GN_KK	99,70
TMY_TN_GN_KK	99,69
GN_TN_KK	99,69
TMY_GN_KT	99,67
TMY_GN_KK_KT	99,66
OCY_GN_KK_KT	99,66
TMY_TN_GN_KK_OCY	99,66
OCY_TN_GN_KT	99,66
GN_TN_KK_KT	99,66
GN_KK	99,66
GN_TN_KT	99,65
TMY_TN_GN_OCY	99,64
TMY_TN_GN_KT_OCY	99,64
TMY_TN_GN_KK_KT_OCY	99,64
OCY_TN_GN	99,64
GN_KT	99,64
TMY_GN_KK_KT_OCY	99,64
GN_KK_KT	99,64
TMY_TN_GN_KT	99,64
TMY_TN_GN_KK_KT	99,60

Çizelge 3.10 incelendiğinde, bütün kombinasyonlarda ortak olan grubun *Gelişmiş Noktalama (GN)* grubu olduğu görülür. Çizelge 3.8 ve Çizelge 3.9'da sadece bu grubun kullanıldığı deneylere bakıldığında, toplam 18 deneyden 6 tanesinin %99'un üzerinde doğrulukla sonuçlandığı görülmektedir. Başka hiçbir özellik grubu tek başına kullanıldığında bu değerlere ulaşamamıştır. *Gelişmiş Noktalama* dışındaki 5 özellik grubunun tek başlarına kullanıldığı toplam 90 deneyde en yüksek değer %66,36 ile *Temel Noktalama (TN)* grubuna ait olmuştur. Çizelge 3.11 ve Çizelge 3.12'de tek bir özellik grubu ile

gerçekleştirilen deneylerin sonuçları verilmiştir. Çizelgelerdeki her kolonda en yüksek değerin *Gelişmiş Noktalama* grubu ile elde edildiği görülmektedir.

Çizelge 3.11. YAŞAM, tek bir özellik grubuyla yapılan deneylerin sonuçları

Doğruluk Değerleri (%), Alan: YAŞAM									
Özellik Grubu	**Model**								
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
TMY	30,00	30,48	29,89	30,49	30,49	29,69	30,58	30,54	30,52
OCY	30,00	30,48	29,89	30,49	30,49	29,69	30,58	30,54	30,52
GN	**99,72**	**99,72**	**99,72**	**81,09**	**80,84**	**79,72**	**56,98**	**57,16**	**57,83**
TN	51,92	66,36	50,96	57,75	60,98	50,94	55,28	55,37	56,72
KK	32,20	35,22	32,19	34,33	34,22	32,36	34,75	35,14	35,08
KT	37,24	57,62	38,48	44,38	47,16	36,62	42,50	45,02	47,31

Çizelge 3.12. SİYASET,tek bir özellik grubuyla yapılan deneylerin sonuçları

Doğruluk Değerleri (%), Alan: SİYASET									
Özellik Grubu	**Özellik Grubu**								
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
TMY	26,38	25,92	26,48	26,99	26,49	25,69	24,21	24,34	24,36
OCY	26,38	25,92	26,48	26,99	26,49	25,69	24,21	24,34	24,36
GN	**75,37**	**75,40**	**75,21**	**75,30**	**75,43**	**75,30**	**99,30**	**99,45**	**99,51**
TN	34,96	33,92	33,82	37,41	42,12	34,29	30,55	30,35	32,37
KK	28,12	30,49	26,87	31,18	30,15	25,84	25,21	25,66	25,72
KT	32,45	18,89	33,42	22,18	20,62	34,70	33,18	30,98	30,48

Ayrıca, Çizelge 3.8 ve Çizelge 3.9'un detaylıca incelenmesiyle farkedilebilecek birkaç nokta ise şöyledir:

- *Gelişmiş Noktalama* grubunu içeren özellik grubu kombinasyonlarıyla yapılan deneylerin en yüksek ve en düşük doğruluk değerleri sırasıyla %99.79 ve %56.90 olmuştur.

- *Gelişmiş Noktalama* grubunu içermeyen özellik grubu kombinasyonlarıyla yapılan deneylerin en yüksek ve en düşük doğruluk değerleri ise sırasıyla %70.35 ve %8.36 olmuştur.
- Toplam 1134 deneyden 498 tanesi 70% üzerinde doğrulukla sonuçlanırken, bunlardan yalnızca 1 tanesinde *Gelişmiş Noktalama* grubu kullanılmamıştır.
- *Gelişmiş Noktalama* grubunu içermeyen özellik kümeleriyle yapılan toplam 558 deneyden yalnızca 27 tanesi %60'ın üzerinde doğrulukla sonuçlanmıştır. Bunların hepsinde ise *Temel Noktalama* grubu ortaktır (Çizelge 3.13: Özellik grupları birden çok deneyde %60 üzerinde başarı elde etse de çizelgeye sadece en yüksek değerleriyle alınmışlardır).
- *Gelişmiş Noktalama* ya da *Temel Noktalama* grubunu içermeyen toplam 270 deneyde erişilebilen en yüksek doğruluk değeri %57,62 olmuştur.

Çizelge 3.13. *Gelişmiş Noktalama* grubunu kullanmadan yapılan deneylerden %60 üzerinde doğrulukla sonuçlananlarda kullanılan özellik grupları

Özellik Grubu Kombinasyonu	**Doğruluk Değeri (%)**
TN_KT	70,35
TMY_TN_OCY	67,38
TN	66,36
TMY_TN	66,26
TN_KK_KT	64,22
TN_KK	63,94
TMY_TN_KK	63,57
TMY_TN_KT_OCY	63,48
TMY_TN_KK_OCY	63,29
TMY_TN_KT	62,64
TMY_TN_KK_KT_OCY	61,27
OCY_TN_KT	60,22
OCY_TN	60,04

Bu bilgiler gözönüne alındığında, Türkçe yazar tanıma için en ayrıştırıcı grubun *Gelişmiş Noktalama*, bundan sonraki en başarılı grubun ise Temel *Noktalama* olduğu sonucuna varılmıştır. Ancak her iki grup da bütün deneylerde tek başına yeterli görülmemektedir. Çizelge 3.14 ve Çizelge

3.15'te herbir deney için en yüksek doğruluk değerine erişen grup kombinasyonları ile, *Gelişmiş Noktalama* ve *Temel Noktalama* gruplarının bir karşılaştırması yapılmış, en son kolona ise bu iki grubun oluşturduğu kombinasyonun eriştiği sonuçlar yerleştirilmiştir. Çizelgeler incelendiğinde *Gelişmiş Noktalama* grubunun bazı deneyler için en yüksek doğruluk değerine ulaşmış olmasına rağmen, çoğu deneyde en yüksek değerin altında kaldığı, *Temel Noktalama* grubunun ise tek başına önemli bir başarı elde edemediği görülmektedir. Bu iki gruptan oluşturulan kombinasyonun eriştiği sonuçlar ise hem deneylerin çoğunda, hem de ortalamalarda ayrı ayrı *Gelişmiş Noktalama* ve *Temel Noktalama* gruplarınınkilerden yüksektir. Ayrıca bu kombinasyon, genellikle en yüksek doğruluk değerine oldukça yakın sonuçlara erişmiştir.

Çizelge 3.14. YAŞAM alanı için en yüksek doğruluk değerine erişen kombinasyonlar, en yüksek değer, GN, TN ve GN_TN karşılaştırması

Doğruluk Değerleri (%), Alan: YAŞAM					
Model	**En yüksek doğruluk değerine erişen özellik grubu kombinasyonları**	**En yüksek**	**GN**	**TN**	**GN_TN**
1	GN OCY_GN	99,72	99,72	51,92	99,67
2	GN GN_TN OCY_GN OCY_GN_KT TMY_GN TMY_GN_KK TMY_GN_OCY TMY_GN_KK_OCY TMY_GN_KT_OCY	99,72	99,72	66,36	99,72
3	OCY_GN	99,79	99,72	50,96	99,68
4	TMY_GN_OCY TMY_TN_GN OCY_GN OCY_TN_GN_KK OCY_TN_GN_KK_KT	99,74	81,09	57,75	90,55
5	TMY_GN OCY_GN	99,73	80,84	60,98	90,62
6	TMY_TN_GN	90,91	79,72	50,94	87,13
7	OCY_TN_GN_KT	89,69	56,98	55,28	76,06
8	TMY_TN_GN_OCY	80,52	57,83	55,37	78,56
9	TMY_TN_GN_OCY	79,81	57,16	56,72	77,94
Ortalama:		93,29	79,20	56,25	88,88

Çizelge 3.15. SİYASET alanı için en yüksek doğruluk değerine erişen kombinasyonlar, en yüksek değer, GN, TN ve GN_TN karşılaştırması

Doğruluk Değerleri (%), Alan: SİYASET					
Model	**En yüksek doğruluk değerine erişen özellik grubu kombinasyonları**	**En yüksek**	**GN**	**TN**	**GN_TN**
1	TMY_TN_GN_OCY	82,16	75,37	34,96	79,01
2	GN_TN	81,63	75,40	33,92	81,63
3	TMY_TN_GN_OCY	81,53	75,21	33,82	78,61
4	GN_TN_KK	80,51	75,30	37,41	78,77
5	GN_TN_KK	82,54	75,43	42,12	81,21
6	TMY_TN_GN_WT	81,57	75,30	34,29	79,21
7	GN	99,30	99,30	30,55	98,61
8	GN OCY_GN	99,51	99,51	30,35	98,95
9	GN OCY_GN TMY_GN	99,45	99,45	32,37	98,87
	Ortalama:	87,58	83,36	34,42	86,10

Gelişmiş Noktalama ve *Temel Noktalama* gruplarının oluşturduğu kombinasyonun başarılı olduğu görülmüştür. Bu durumda irdelenmesi gereken bir başka nokta, başka herhangi bir ikili kombinasyonunun daha yüksek doğruluk değerlerine ulaşıp ulaşamadığıdır.

Çizelge 3.16 ve Çizelge 3.17'de içinde *Gelişmiş Noktalama* grubunun bulunduğu ikili özellik grubu kombinasyonlarının karşılaştırmaları verilmiştir. Çizelgelerden de açıkça görüleceği üzere; hem daha yüksek değerlerin yakalandığı deneylerin sayısı bakımından, hem de ortalamada yakalanan doğruluk değerleri açısından *Gelişmiş Noktalama* ve *Temel Noktalama* gruplarının bir arada oluşturduğu ikili kombinasyon (*GN_TN*) diğerlerinden daha başarılıdır.

Çizelge 3.16. *Gelişmiş Noktalama* grubunu içeren ikili kombinasyonların doğruluk değerleri karşılaştırması, Alan: YAŞAM

Doğruluk Değerleri (%), Alan: YAŞAM					
Model	**OCY_GN**	**TMY_GN**	**GN_TN**	**GN_KK**	**GN_KT**
1	99,72	99,67	99,67	99,60	99,64
2	99,72	99,72	99,72	99,63	99,54
3	99,79	99,66	99,68	99,66	99,64
4	99,74	99,66	90,55	79,30	86,71
5	99,73	99,73	90,62	79,56	87,29
6	79,77	79,91	87,13	80,77	84,63
7	78,71	78,28	76,06	57,67	67,80
8	58,01	58,26	77,94	56,90	67,89
9	57,59	58,73	78,56	57,56	69,54
Ortalama:	85,87	85,96	**88,88**	78,96	84,74

Çizelge 3.17 *Gelişmiş Noktalama* grubunu içeren ikili kombinasyonların doğruluk değerleri karşılaştırması, Alan: SİYASET

Doğruluk Değerleri (%), Alan: SİYASET					
Model	**OCY_GN**	**TMY_GN**	**GN_TN**	**GN_KK**	**GN_KT**
1	77,01	76,75	79,01	77,27	78,68
2	68,63	72,32	81,63	77,07	77,50
3	76,28	76,72	78,61	77,53	78,45
4	75,45	74,11	78,77	78,93	72,06
5	72,31	70,52	81,21	78,54	73,22
6	76,98	78,27	79,21	77,59	78,48
7	74,86	72,31	98,61	99,07	98,76
8	99,45	99,45	98,87	99,30	98,98
9	99,51	99,49	98,95	99,30	99,03
Ortalama:	80,05	79,99	**86,10**	84,95	83,91

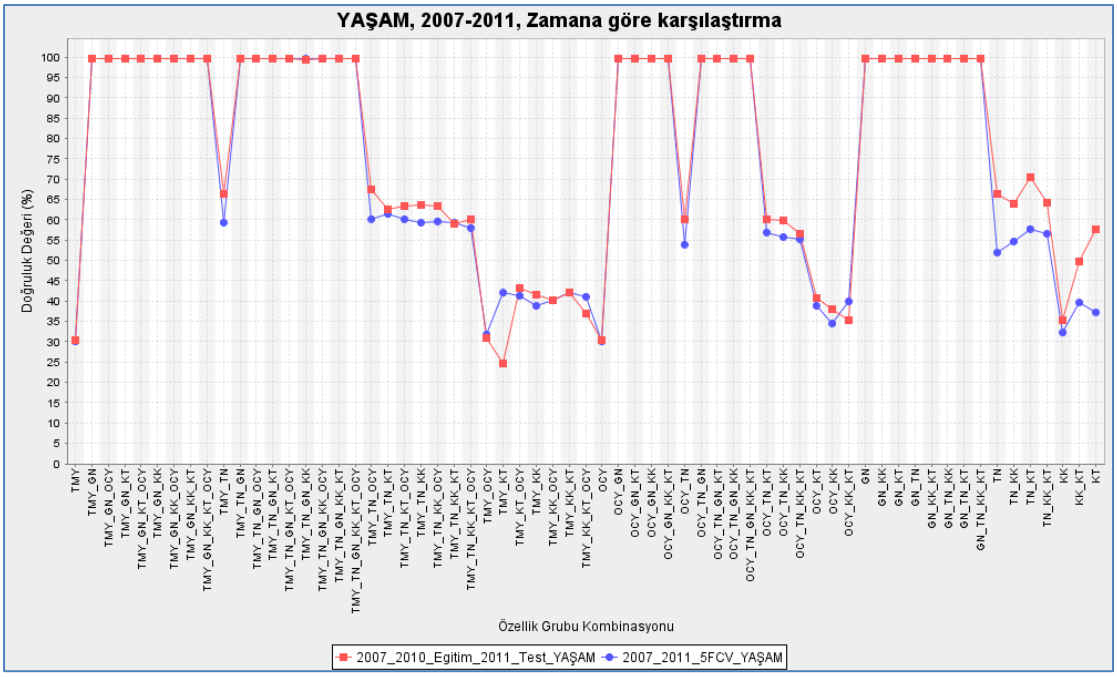
3.5.2 Zamanın Etkisi

Bu kesimde, zamanın, sınıflandırıcıların başarımı üzerindeki etkisini görebilmek amacıyla önce aynı tarih aralıkları için (Çizelge 3.18) zamandan bağımsız ve zamana bağımlı modellerle yapılan deney sonuçları karşılaştırılarak hangi modelin daha iyi sonuç verdiği gözlemlenecek, bunun ardından da sınıflandırıcının yeniden eğitilmeden ne kadar süreyle kullanılabileceği üzerinde durulacaktır.

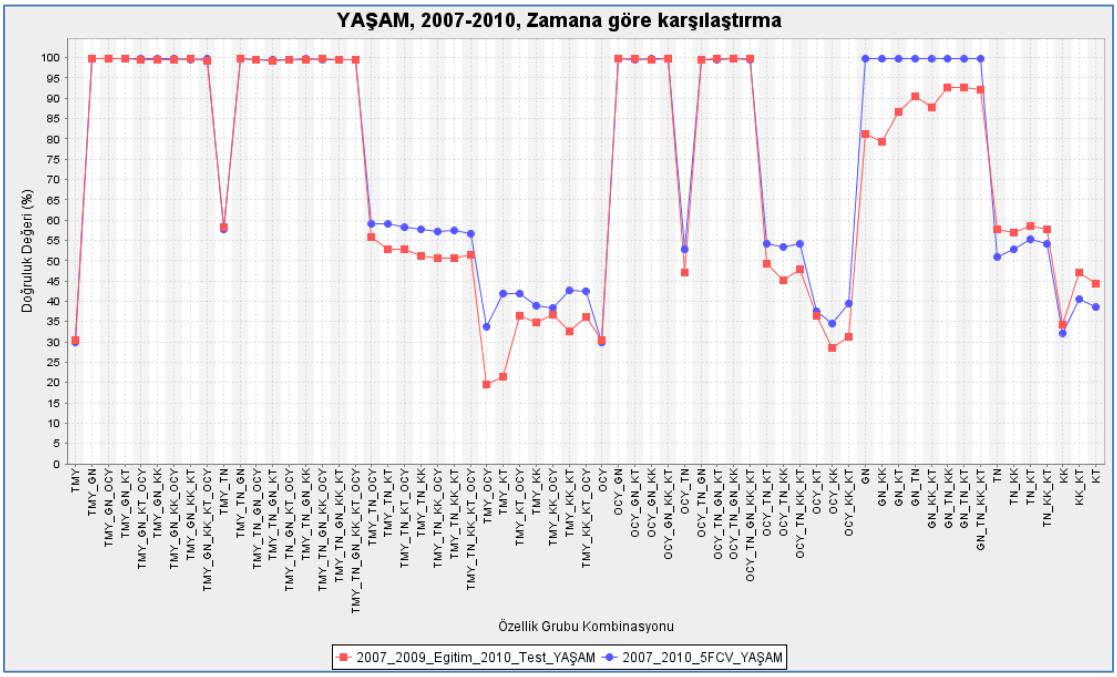
Çizelge 3.18. Aynı tarih aralıklarındaki metinler için yapılan deney grupları

Model	Tarih Aralığı	Yöntem	Zaman İlişkisi
1	2007 - 2011	5FCV	Zamandan Bağımsız
2		Egitim: 2007-2010 Test: 2011	Zamana Bağımlı
3	2007 – 2010	5FCV	Zamandan Bağımsız
4		Egitim: 2007-2009 Test: 2010	Zamana Bağımlı
6	2007- 2009	5FCV	Zamana Bağımlı
7		Egitim: 2007-2008 Test: 2009	Zamandan Bağımsız

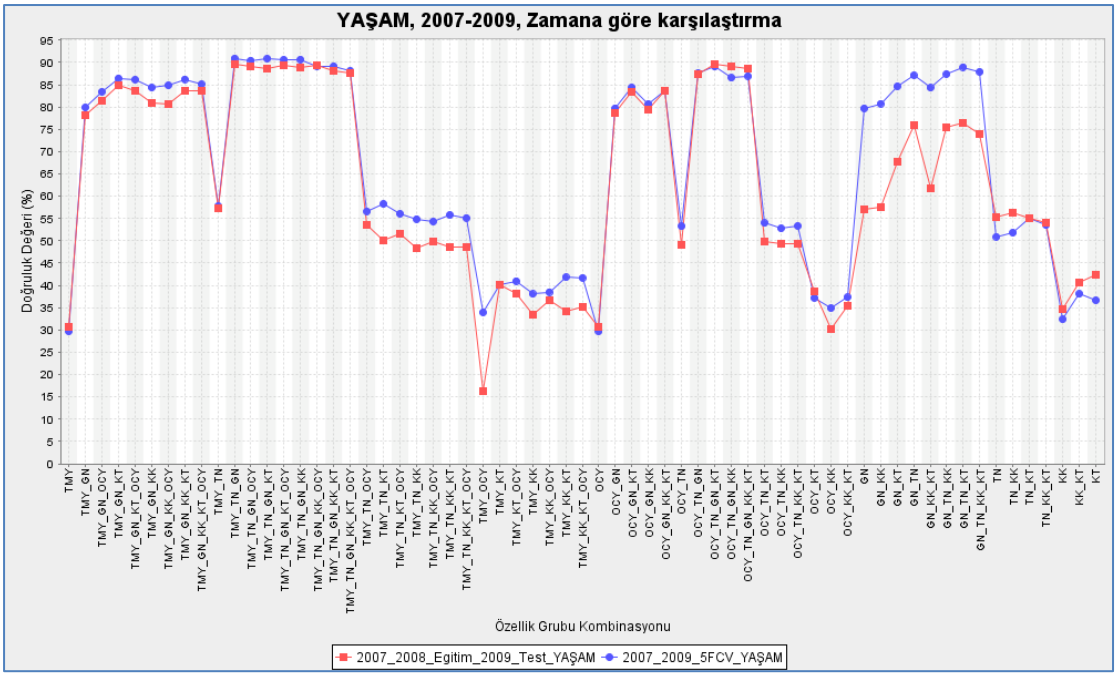
Çizelge 3.18'de verilen modeller ile YAŞAM ve SİYASET alanları için yapılan deneylerin karşılaştırmaları Şekil 3.11'den başlayarak Şekil 3.16'ya kadar grafiklerle gösterilmiştir. Grafiklerde mavi noktalar zamandan bağımsız, kırmızı noktalar ise zamana bağımlı modellerle yapılan deneylerin sonuçlarını göstermektedir. Ayrıca vurgulanması gereken bir durum ise şudur: her iki modelin de aynı ya da çok yakın değerleri yakaladığı yerlerde grafikte yalnızca kırmızı noktanın görünmesi, mavi noktayı örtmesinden kaynaklanmaktadır. Özellikle Şekil 3.11 ve Şekil 3.12'de bu duruma dikkat edilmelidir.



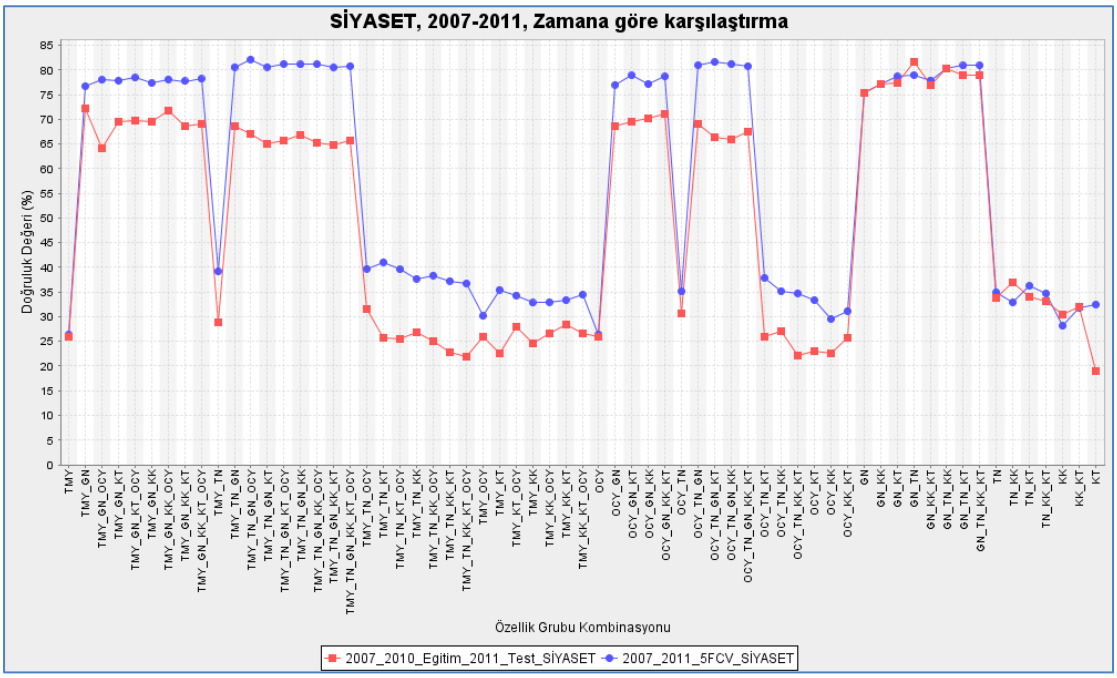
Şekil 3.11. YAŞAM, 2007-2011, Zamana göre karşılaştırma



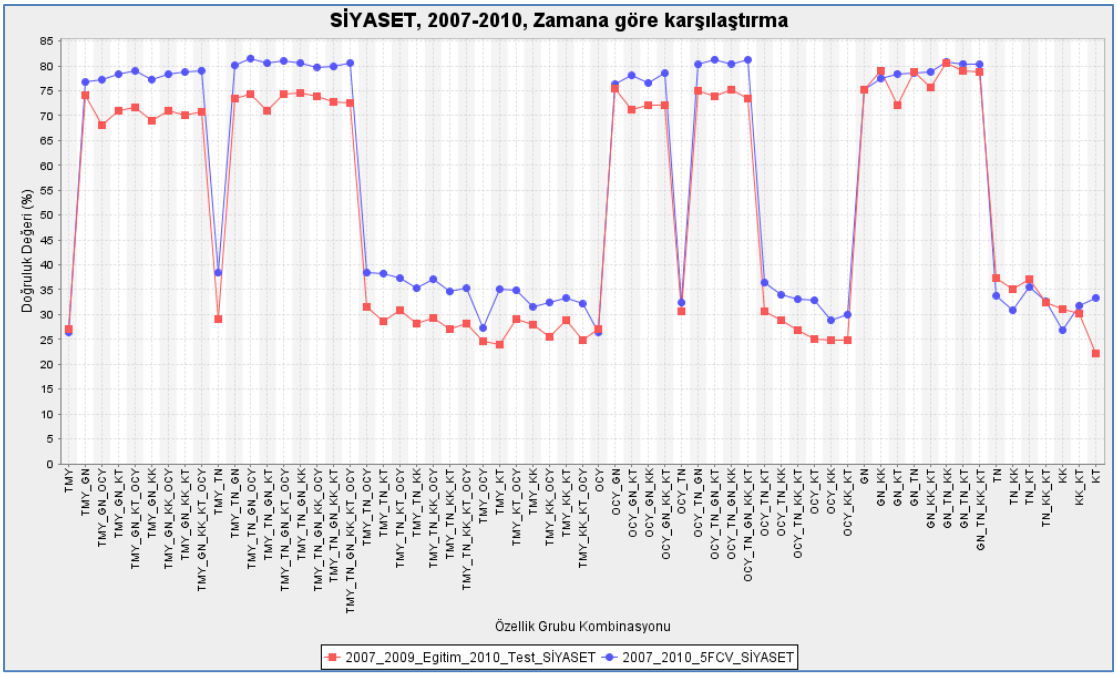
Şekil 3.12. YAŞAM, 2007-2010, Zamana göre karşılaştırma



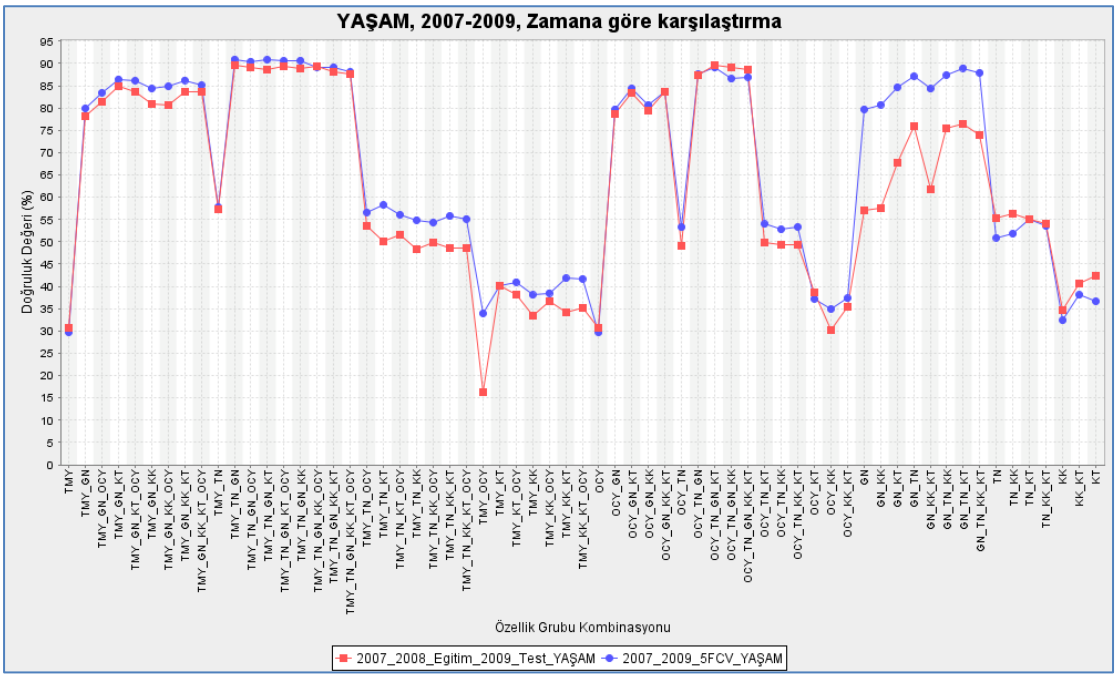
Şekil 3.13. YAŞAM, 2007-2009, Zamana göre karşılaştırma



Şekil 3.14. SİYASET, 2007-2011, Zamana göre karşılaştırma



Şekil 3.15. SİYASET, 2007-2010, Zamana göre karşılaştırma



Şekil 3.16. SİYASET, 2007-2009, Zamana göre karşılaştırma

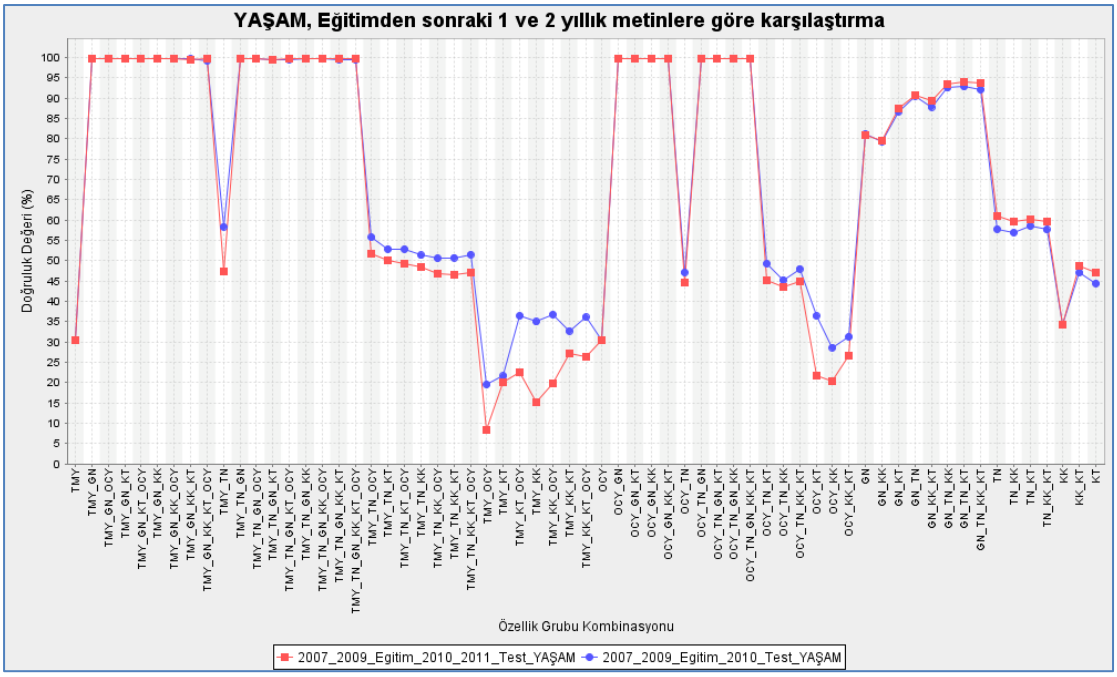
Grafikler incelendiğinde iki deneyde her iki modelin başarısının birbirine yakın olmasının dışında, genel olarak zamandan bağımsız modellerin daha yüksek sonuçlar aldığı görülmektedir. Benzer bir karşılaştırma Çizelge 3.19'da verilen ortalama doğruluk değerlerine göre yapıldığında da aynı yargıya varılmaktadır.

Çizelge 3.19. Deneylerin modellere göre ortalama doğruluk değerleri

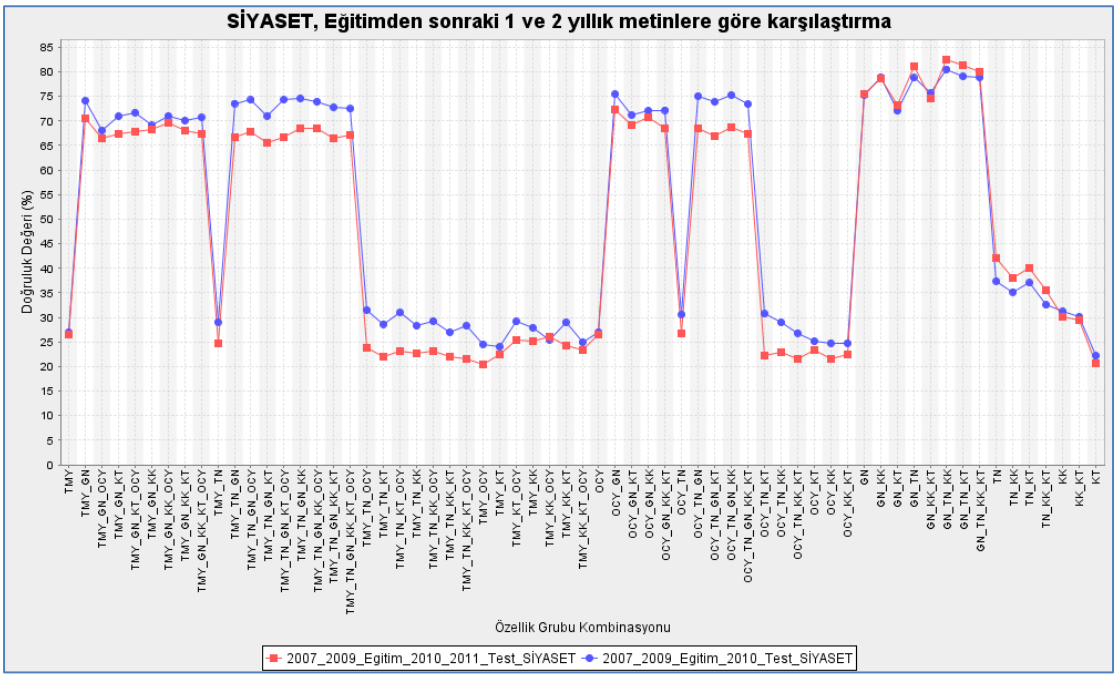
Ortalama Doğruluk Değerleri (%)						
Alan	**Model**					
	Zamandan Bağımsız	**Zamana Bağımlı**	**Zamandan Bağımsız**	**Zamana Bağımlı**	**Zamandan Bağımsız**	**Zamana Bağımlı**
	1	**2**	**3**	**4**	**6**	**7**
YAŞAM	74,10	75,75	73,66	70,44	66,36	62,48
SİYASET	57,16	49,19	56,48	51,58	56,61	54,28

Bu durum kabaca aynı sayıda metinle eğitilseler bile, zamandan bağımsız modellerin, zamana bağımlı modellere oranla daha başarılı olduğu, dolayısıyla metin özelliklerinde meydana gelebilecek değişimlerin kısa zaman aralıkları için bile dikkate alınması gerektiğini göstermektedir.

Zamana bağlı değişimler sınıflandırıcının başarımını etkilediğine göre, bir sınıflandırıcının ne kadar süre ile güvenle kullanılabileceği bilgisi önem kazanmaktadır. Dolayısıyla, deney sonuçları bu bakış açısıyla da irdelenmiştir. Şekil 3.17 ve Şekil 3.18'de, YAŞAM ve SİYASET alanlarında 4 ve 5. modellerle yapılan deneylerin karşılaştırması görülmektedir. Bu deneylerde sınıflandırıcının sınanmasında kullanılan metinler, o sınıflandırıcıların eğitilmesinde kullanılan metinlerin son tarihinden sonraki bir yıllık (mavi noktalar) ve iki yıllık (kızmızı noktalar) zaman aralığında yazılmış metinlerdir. YAŞAM alanındaki deneylerde en yüksek doğruluk değerleri için her iki modelin sonuçları birbirlerine çok yakınken, SİYASET alanındaki deneylerde ve her iki alanda da yüksek olmayan değerlerin genelinde ilk yıldan sonra başarımın düştüğü görülmektedir.



Şekil 3.17. YAŞAM, Eğitimden sonraki 1 ve 2 yıl için karşılaştırma



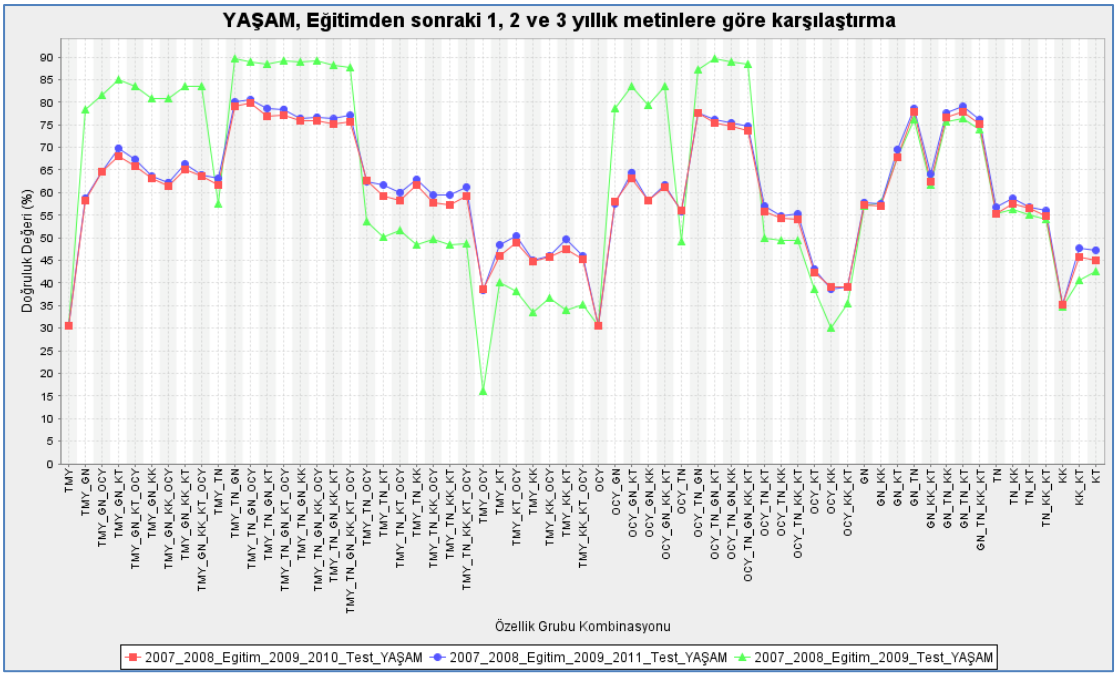
Şekil 3.18. SİYASET, Eğitimden sonraki 1 ve 2 yıl için karşılaştırma

Çizelge 3.20'de görülen ortalama doğruluk değerleri de zamana bağımlı olarak eğitilen sınıflandırıcıların başarısının, eğitimi izleyen ilk yıldan sonra düştüğünü göstermektedir.

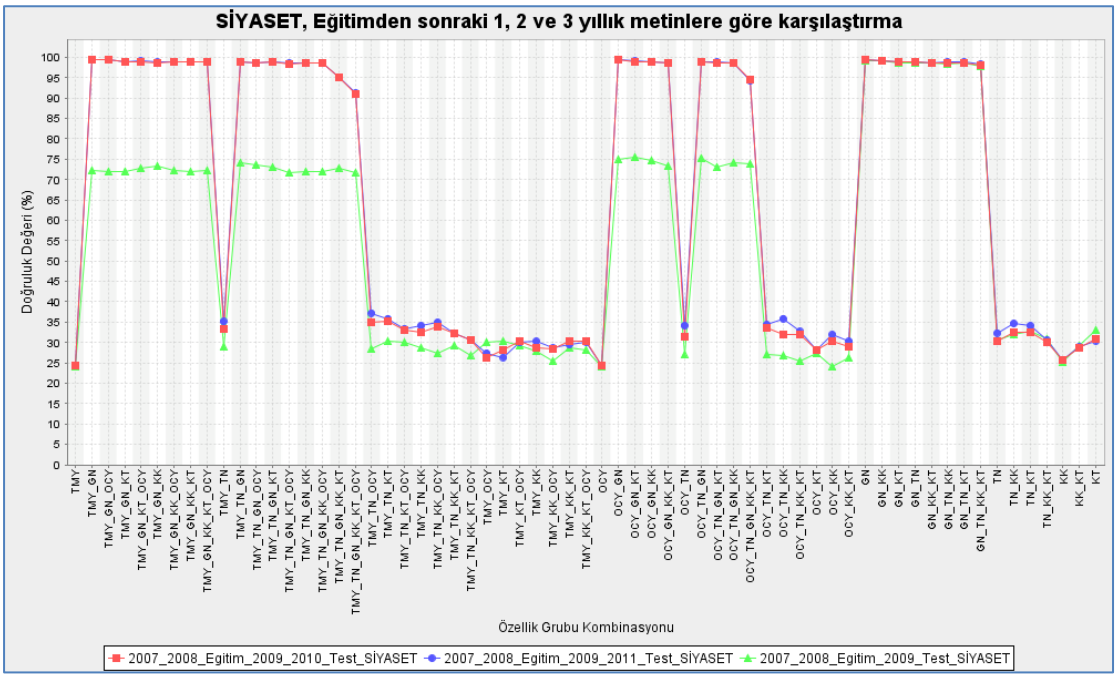
Çizelge 3.20. 4. ve 5. modellerle yapılan deneylerin ortalama doğruluk değerleri

Ortalama Doğruluk Değerleri (%)		
Alan	**Model**	
	4	**5**
YAŞAM	70,44	68,33
SİYASET	51,58	48,59

Şekil 3.19 ve Şekil 3.20'de ise YAŞAM ve SİYASET alanlarında 7, 8 ve 9. modellerle yapılan deneylerin karşılaştırması görülmektedir. Bu deneylerde sınıflandırıcının sınanmasında kullanılan metinler, o sınıflandırıcıların eğitilmesinde kullanılan metinlerin son tarihinden sonraki bir yıllık (yeşil noktalar), iki yıllık (kızmızı noktalar) ve üç yıllık (mavi noktalar) zaman aralığında yazılmış metinlerdir. YAŞAM alanındaki deney sonuçları incelendiğinde, beklendiği gibi ilk yıldan sonra başarımın düştüğü görülmektedir. SİYASET alanındaki deneylerde ise şaşırtıcı (ve yanıltıcı) bir şekilde, ilk yıldan sonra başarımın daha da yükselmesi söz konusudur. Bu durum sınıflandırıcıların ilk yıldan sonraki güvenilirliğinin daha da çok sarsılmasına neden olmaktadır. Böyle sonuçlar alınmasının nedeni, Türkiye'de çok hızlı değişen siyasi gündem olabileceği gibi sınıflandırıcının eğitilmesinde kullanılan metinlerin ait olduğu zaman aralığının (2 yıl), testlerde kullanılan metinlerin ait olduğu zaman aralığından (3 yıl) kısa olması da olanaklıdır.



Şekil 3.19. YAŞAM, Eğitimden sonraki 1, 2 ve 3 yıl için karşılaştırma



Şekil 3.20. SİYASET, Eğitimden sonraki 1, 2 ve 3 yıl için karşılaştırma

4. SONUÇ

Bu kesimde Bölüm 3.5'te aktarılan deney sonuçlarına dayanılarak varılan bulgular özetlenecek, sonraki çalışmalara yönelik önerilere değinilecektir.

Araştırılan sözcüksel ve sözdizimsel özellikler arasında ayırıcılığı en yüksek özelliklerin *Gelişmiş Noktalama* grubu altında toplanmış olan özellikler olduğu görülmüştür. Ancak bu özellik grubunun tek başına kullanılmasındansa, ikinci en belirleyici özellik grubu olarak belirlenen *Temel Noktalama* grubu ile birlikte kullanılması daha tatmin edici sonuçlar vermiştir. Bu kombinasyon bütün deneylerde en yüksek başarımı yakalayamasa da hem diğer kombinasyonlara göre daha yüksek doğruluk değerlerine ulaşması, hem özellik sayısının düşük olması, hem de kullanılan özelliklerin metinlerden çok kolay çıkartılabilmesi gibi nedenlerden dolayı Türkçe yazar tanıma çalışmaları için en uygun özellik kümesini oluşturmaktadır ve aşağıdaki 15 adet noktalama özelliği içermektedir.

Çizelge 4.1. Ayırıcılığı yüksek en küçük özellik kümesini oluşturan gruplar ve özellikler

Özellik Grubu	İçerdiği Özellikler (Metinde geçen)
Gelişmiş Noktalama	tire işaretlerinin ('-') sayısı alt çizgilerin ('_') sayısı *slash* ('/') karakterlerinin sayısı ters *slash* ('\') karakterlerinin sayısı parantezlerin ('( )') sayısı *ampersand* ('&') işaretlerinin sayısı
Temel Noktalama	noktaların ('.') sayısı virgüllerin (',') sayısı soru işaretlerinin ('?') sayısı üç noktaların ('...') sayısı tek tırnakların (''') saysı çift tırnakların ('"') sayısı ünlem işaretlerinin ('!') sayısı iki nokta üst üstelerin (':') sayısı noktalı virgüllerin (';') sayısı

Yazar tanımada kullanılmak üzere oluşturulan sınıflandırıcılar, eğitildikleri zaman aralığında yazılmış metinleri, bu zaman aralığının dışında yazılmış metinlere göre daha yüksek başarıyla tanımaktadır. Bu da zamana bağlı bir değişimin var olduğunu, sınıflandırıcıların zamanla birlikte etkinliklerini kaybettiklerini göstermektedir. Dolayısıyla, yazar tanıma bağlamında

kullanılacak sınıflandırıcıların yeni tarihlerde yazılmış metinlerle de sürekli eğitilmesi gerekmektedir. Buna ek olarak 3 – 4 yıllık bir zaman aralığındaki metinlerle eğitilen sınıflandırıcıların en çok 1 yıl süre ile yeniden eğitilmeden kullanılabileceği, aksi taktirde elde edilecek sınıflandırma sonuçlarının yeterince güvenilir olmadığı görülmüştür.

Bu tez çalışmasında Türkçe için etkin bir özellik kümesi araştırılırken, oluşturulan sınıflandırıcıların da zamandan nasıl etkilendikleri incelenmiştir. Bu amaçlarla oluşturulan külliyat 5 yıllık bir zaman aralığından 8 farklı yazara ait köşe yazılarını içermekte, üzerinde çalışılan özellikler ise sözcüksel ve sözdizimsel özelliklerden oluşmaktadır. Varılan sonuçların detaylıca sınanabilmesi için aşağıda belirtilen konuları kapsayan yeni çalışmalar yapılması planlanmaktadır:

- İşlev sözcüklerinin kullanılması: Farklı diller için yapılan çeşitli çalışmalarda işlev sözcükleri özellik kümesi olarak kullanılmıştır. Dolayısıyla Türkçe için işlev sözcüklerinin çıkartılarak bu tez kapsamındakine benzer bir incelemenin işlev sözcüklerini de kapsayacak şekilde gerçekleştirilmesi ve tez kapsamında en iyi olarak belirlenen noktalama işaretleri ile karşılaştırılması;
- Külliyatın genişletilmesi: SİYASET ve YAŞAM dışında kalan KÜLTÜR/SANAT, SPOR, SEYAHAT gibi çeşitli alanlardan metinlerin de dahil edilmesiyle alan etkisinin azaltılması ve özellik kümelerinin her alan için benzer başarımı gösterip göstermediğinin araştırılması;
- Çeşitli zaman aralıklarının ele alınması: Farklı uzunluklarda ve değişik yıllara yayılmış aralıklardan toparlanacak metinler üzerinde yapılacak incelemelerle zamanın etkisinin daha geniş çerçevede irdelenmesi

KAYNAKLAR

[1] Lord, R. D., Studies in the History of Probability and Statistics VIII: De Morgan and the Statistical Study of Literary Style. *Biometrika*, 45 (1-2) 282, **1958**.

[2] De Morgan, S., *Memoir of Augustus de Morgan by his Wife Sophia Elizabeth de Morgan With Selections From His Letters*, London: Longmans, Green, and Co., **1982**.

[3] Mendenhall, T. C., The characteristic curves of composition. *Science*, 9, 237-246, **1887**.

[4] Stańczyk, U., Cyran, K. A., On employing elements of rough set theory to stylometric analysis of literary texts. *Journal of Applied Mathematics and Informatics*, 1(4), 159–166, **2007**.

[5] Stańczyk, U., Cyran, K. A., Can punctuation marks be used as writer invariants? Rough set-based approach to authorship attribution. *In Proceedings of the 2nd Conference on European Computing*, 228–233, **2008**.

[6] Rudman, J., The State of Authorship Attribution Studies: Some Problems and Solutions. *Computers and the Humanities*, 31(4), 351-365, **1998**.

[7] Juola, P., The Time Course of Language Change, *Computers and the Humanities*, 37(1), 77-96, **2003**.

[8] Can, F., Patton, J.M., Change of Writing Style with Time, *Computers and the Humanities*, 38, 81-82, **2004**.

[9] Holmes, D. I., Authorship Attribution, *Computers and the Humanities*, 28(2), 87-106, **1994**.

[10] Holmes, D. I., The evolution of stylometry in humanities, *Literary and Linguistic Computing*, 13(3), 111-117, **1998**.

[11] McEnery, A. M., Oakes, M. P. , Authorship studies/textual statistics, *In Handbook of natural language processing*. Marcel Dekker Inc, Dallas, 234-248, **2000**.

[12] Love, H., *Attributing Authorship: An Introduction*. Cambridge University Press, **2002**.

[13] Grieve, J. W., Quantitative authorship attribution: A history and an evaluation of techniques, Yüksek Lisans Tezi, Simon Fraser University, **2005,** http://summit.sfu.ca/system/files/iritems1/8840/etd1721.pdf .

[14] Juola, P., Authorship Attribution, *Foundations and Trends in Information Retrieval*, 1(3), 233-334, **2006**.

[15] Koppel, M., Schler, J., Argamon, S., Computational methods in authorship attribution, *Journal of the American Society for Information Science and Technology*, 60(1), 9-26, **2009**.

[16] Stamatatos, E., A survey of modern authorship attribution methods, *Journal of the American Society for Information Science and Technology*, 60(3), 538-556, **2009**.

[17] Malone, E., A Dissertation on the Three Parts of King Henry VI. Tending to Show that Those Plays Were Not Written Originally by Shakspeare, London, **1787**

[18] Mendenhall, T. C., A mechanical solution of a literary problem. *Popular Science Monthly*, 60, 97-105, **1901**.

[19] Mascol, C., Curves of Pauline and Pseudo-Pauline Style i. *Unitarian Review*, 30 (November), 452–460, **1888**.

[20] Mascol, C., Curves of Pauline and Pseudo-Pauline Style ii. *Unitarian Review*, 30 (December), 539–546, **1888**.

[21] Zipf, G. K., *Selected studies of the principle of relative frequency in language*. Harvard University Press, Cambridge, MA., USA, **1932.**

[22] Yule, G.U., On sentence-length as a statistical characteristic of style in prose, with application to two cases of disputed authorship. *Biometrika*, 30(3/4), 363-390, **1938**.

[23] Yule, G. U., *The statistical study of literary vocabulary*. Cambridge University Press, **1944.**

[24] Fucks, W., On Mathematical Analysis of Style, *Biometrika*, 39, 122-129, **1952**.

[25] Fucks, W. On Nahordnung and Fernordnung in Samples of Literary Texts. *Biometrika*, 41, 116-132, **1958**.

[26] Brinegar, C. S., Mark Twain and the Quintus Curtius Snodgrass Letters: A Statistical Test of Authorship, *Journal of the American Statistical Association*, 58(301), 95-96, **1963**.

[27] Mosteller, F., Wallace, D. L., Inference in an Authorship Problem: A Comparative Study of Discrimination Methods Applied to the Authorship of the Disputed Federalist Papers, *Journal of the American Statistical Association*, 58(302), 275-309, **1963**.

[28] Mosteller, F., Wallace, D. L., *Inference in an disputed authorship: The Federalist*, Addison-Wesley, MA., USA, **1964.**

[29] The Federalist Papers, http://www.thefederalistpapers.org/federalist-papers (Ocak, 2014).

[30] Primary Documents In American History, The Federalist Papers, http://www.loc.gov/rr/program/bib/ourdocs/federalist.html (Ocak, 2014).

[31] Morton, A. Q., The authorship of Greek prose. *Journal of the Royal Statistical Society (A)*, 128, 169–233, **1965**.

[32] Morton, A. Q., The Integrity of the Pauline Epistles. *Manchester Statistical Society*, **1965**.

[33] Morton, A. Q., *The Authorship of the Pauline Epistles: A Scientific Solution*. University of Saskatoon, **1965**.

[34] O'Donnell, B., Stephen Crane's 'The O'Ruddy': A problem in authorship discrimination. *In Leed (Ed.), The computer and literary style,* 107–115. Kent State University Press, **1966**.

[35] Levison, M., Morton A. Q., Winspear, A. D., The Seventh Letter of Plato. *Mind*, 77, 309-325, **1968**

[36] Anonim, Seventh Letter (Plato), http://en.wikipedia.org/wiki/Seventh_Letter_(Plato), (Ocak, **2014**).

[37] Michaelson, S., Morton, A. Q., The New Stylometry: A One-Word Test of Authorship for Greek Writers, *The Classical Quarterly*, 22, 89-102, **1972**.

[38] Michaelson, S., Morton, A. Q., The spaces in between: A multiple test of authorship for Greek writers, *R.E.L.O. Review*, 1, 23-77, **1972**.

[39] Michaelson, S., Morton, A. Q., Last Words: A test of authorship for Greek writers, *New Testament Studies*, 18, pp. 192-208, **1972**.

[40] Morton, A. Q., *Literary Detection: How to Prove Authorship and Fraud in Literature and Documents*, Scribners, New York, **1978**.

[41] Kenny, A., *The Aristotelian Ethics: A Study of the Relationship between the Eudemian and Nicomachean Ethics of Aristotle*, Clarendon Press, Oxford, **1978**.

[42] Stanford Encyclopedia of Philosophy, http://plato.stanford.edu/entries/aristotle-ethics/, (Ocak, **2014)**.

[43] Merriam, T., What Shakespeare Wrote in Henry VIII (Part I), *The Bard*, 2, 81-94, **1979**.

[44] Merriam, T., What Shakespeare Wrote in Henry VIII (Part II), *The Bard*, 2, 111- 118, **1980**.

[45] Merriam, T., The Authorship of Sir Thomas More. *ALLC Bulletin*, 10, 1-7, **1982**.

[46] Smith, M.W.A., An Investigation of the Basis of Morton's Method for the Determination of Authorship, *Style*, 19, 341-368, **1985**.

[47] Smith, M.W.A., An Investigation of Morton's Method to Distinguish Elizabethan Playwrights, *Computers and the Humanities*, 19, 3-21, **1985**.

[48] Merriam, T., The Authorship Controversy of Sir Thomas More: Smith on Morton, *Literary and Linguistic Computing*, 1, 104-106, **1986**.

[49] Smith, M.W.A., Merriam's Application of Morton's Method, *Computers and the Humanities*, 2, 59-60, **1987**.

[50] Morton, A. Q., Once. A Test of Authorship Based on Words which are not Repeated in the Sample, *Journal of the Association for Literary and Linguistic Computing*, 1, 1-8, **1986**.

[51] Smith M.W.A., Hapax Legomena in Prescribed Positions: an Investigation of Recent Proposals to resolve problems of authorship, *Journal of the Association for Literary and Linguistic Computing*, 2, 145-152, **1987**.

[52] Merriam, T., A Reply to 'An Investigation of the Basis of Morton's Method for the Determination of Authorship', *Style*, 22(4), 646, **1988**.

[53] Holmes, D. I., The analysis of literary style: A review, *The Journal of the Royal Statistical Society (Series A)*, 148(4), 328–341, **1985**.

[54] Kenny, A., *A Stylometric Study of the New Testament*, Oxford University Press, **1986**.

[55] Anonim, Gospel of Luke, http://en.wikipedia.org/wiki/Gospel_of_Luke, (Ocak, **2014**)

[56] Anonim, Acts of the Apostles, http://en.wikipedia.org/wiki/Acts_of_the_Apostles, (**Ocak, 2014**)

[57] Anonim, Gospel of John, http://en.wikipedia.org/wiki/Gospel_of_John, (Ocak, **2014**)

[58] Anonim, Book of Revelation, http://en.wikipedia.org/wiki/Book_of_Revelation, (Ocak, **2014**)

[59] Burrows, J. F., *Computation into Criticism: a Study of Jane Austen's Novels and an Experiment in Method*, Clarendon Press, Oxford, **1987.**

[60] Burrows, J.F., Hassall, A. J., Anna Boleyn and the Authenticity of Fielding's Feminine Narratives, *Eighteenth Century Studies*, 21(4), 427-453, **1988**.

[61] Anonim, Henry Fielding, http://en.wikipedia.org/wiki/Henry_Fielding, (Ocak, **2014**).

[62] Anonim, Sarah Fielding, http://en.wikipedia.org/wiki/Sarah_Fielding, (Ocak, **2014**).

[63] Merriam, T., An Experiment with the Federalist Papers, *Computing and the Humanities*, 23, 251-254, **1989**.

[64] Farringdon, M. G., Morton, A. Q., *Fielding and the Federalist, Technical Report*, CSC 90/R6. University of Glasgow, **1989**.

[65] Foster, D, *'Elegy' by W. S.: A Study in Attribution*, Associated University Presses, Cranbury, New Jersey, **1989**.

[66] Neumann, K. J., *The Authenticity of the Pauline Epistles in the Light of Stylostatistical Analysis*, Scholar's Press, Atlanta, **1990**.

[67] Morton, A. Q., Michaelson, S., *The Qsum Plot, Technical Report,* CSR-3-90, University of Edinburgh, **1990**.

[68] Canter, D., An Evaluation of the 'Cusum' Stylistic Analysis of Confessions, *Expert Evidence*, 1, 93-99, **1992**.

[69] Hardcastle, R. A., Forensic Linguistics: An Assessment of the CUSUM Method for the Determination of Authorship, *Journal of the Forensic Science Society*, 33, 95-106, **1993**.

[70] Hilton, M, Holmes. D. I., An Assessment of Cumulative Sum Charts for Authorship Attribution, *Literary and Linguistic Computing*, 8, 73-80, **1993**.

[71] Holmes, D. I., Tweedie, F., Forensic Stylometry: A Review of the Cusum Controversy, *In Revue Informatique et Statistique dans les Science Humaine. University of Liege*, Belgium, 19-47, **1995**.

[72] Anonim, The Revenger's Tragedy, http://en.wikipedia.org/wiki/The_Revenger's_Tragedy, (Ocak, **2014**).

[73] Smith, M.W.A., The Authorship of *The Revenger's Tragedy*, *Notes and Queries*, 236, 508-511, **1991**.

[74] Smith, M.W.A., The Problem of Acts I-II of *Pericles*". *Notes and Queries*, 237, 346-355, **1992**.

[75] Smith, M.W.A., Edmund Ironside, *Notes and Queries*, 238, 202-205, **1993**.

[76] Anonim, Pericles, Prince of Tyre, http://en.wikipedia.org/wiki/Pericles,_Prince_of_Tyre, (Ocak, **2014**).

[77] Anonim, Edmund Ironside, http://en.wikipedia.org/wiki/Edmund_Ironside_(play), (Ocak, **2014**).

[78] Anonim, Bronte Family, http://en.wikipedia.org/wiki/Bronte_family, (Ocak, **2014**).

[79] Burrows, J. F., Not Unless You Ask Nicely: The Interpretative Nexus Between Analysis and Information, *Literary and Linguistic Computing*, 7, 91-109, **1992**.

[80] Burrows, J. F., *Computers and the Study of Literature. In C.S. Butler (ed.)*, *Computers and Written Texts*, Blackwell, Oxford, **1992**.

[81] Craig, H., Authorial Styles and the Frequencies of Very Common Words: Jonson, Shakespeare, and the Additions to 'The Spanish Tragedy', *Style*, 26, 199- 220, **1992**.

[82] Merriam, T., Marlowe's Hand in *Edward III* revisited, *Literary and Linguistic Computing*, 11(1), 19-22, **1996**.

[83] Craig, H., Authorial Attribution and Computational Linguistics: if you can tell authors apart, have you learned anything about them?, *Literary and Linguistic Computing*, 14, 103-113, **1999**.

[84] Anonim, Book of Mormon, http://en.wikipedia.org/wiki/Book_of_Mormon, (Ocak, **2014**).

[85] Holmes, D. I., A Stylometric Analysis of Mormon Scripture and Related Texts, *Journal of the Royal Statistical Society A*, 155, 91-120, **1992**.

[86] Matthews, R., Merriam, T., Neural Computation in Stylometry I: An Application to the Works of Shakespeare and Fletcher, *Literary and Linguistic Computing*, 8, 203-209, **1993**.

[87] Merriam, T. Marlowe's Hand in *Edward III*, *Literary and Linguistic Computing*, 8(2), 59-72, **1993**.

[88] Merriam, T., Matthews, R., Neural Computation in Stylometry II: An Application to the Works of Shakespeare and Marlowe, *Literary and Linguistic Computing*, 9, 1-6, **1994**.

[89] Kjell, B., Authorship determination using letter pair frequencies with neural network classifiers, *Literary and Linguistic Computing*, 9(2), 119–124, **1994**.

[90] Ledger, G., Merriam, T., Shakespeare, Fletcher, and the Two Noble Kinsmen, *Literary and Linguistic Computing*, 9, 119-124, **1994**.

[91] Smith, M.W.A., *Sir Thomas More, Pericles* and Stylometry, *Notes and Queries*, 239, 55-58, **1994**.

[92] Merriam, T., Letter Frequency as a Discriminator of Authorship, *Notes and Queries*, 239, 467-469, **1994**.

[93] Holmes, D. I., Forsyth, R., The *Federalist* Revisited: New Directions in Authorship Attribution, *Literary and Linguistic Computing*, 16, 403-420, **1995**.

[94] Ledger, G., An Exploration of Differences in the Pauline Epistles using Multivariate Statistical Analysis, *Literary and Linguistic Computing*, 10, 85-97, **1995**.

[95] Martindale, C., McKenzie, D. P., On the Utility of Content Analysis in Authorship Attribution: The Federalist, *Computers and the Humanities*, 29, 259- 270, **1995**.

[96] Lowe, D., Matthew, R., A Stylometric Analysis by Radial Basis Functions, *Computers and the Humanities*, 29, 449-461, **1995**.

[97] Mealand, D. L., Correspondence analysis of Luke, *Literacy and Linguistic Computing*, 10, 171–182, **1995**.

[98] Baayen, H., van Halteren, H., Tweedie, F., Outside The Cave of Shadows: Using Syntactic Annotation to Enhance Authorship Attribution, *Literary and Linguistic Computing*, 11, 110-120, **1996**.

[99] Elliott, W., Valenza, R. J., And Then There Were None: Winnowing The Shakespeare Claimants, *Computers and the Humanities*, 30, 191-245, **1996**.

[100] Elliot Online, The Shakespeare Clinic, http://www.claremontmckenna.edu/pages/faculty/welliott/shakes.htm, (Ocak, **2014**).

[101] Foster, D., Response to Elliot and Valenza, 'And Then There Were None', *Computers and the Humanities*, 30, 247-255, **1996**.

[102] Tweedie, F. J., Singh, S., Holmes, D. I., Neural Network Applications in Stylometry: the Federalist Papers, *Computers and the Humanities*, 30, 1-10, **1996**.

[103] Whissell, C., Traditional and Emotional Stylometric Analysis of the Songs of Beatles Paul McCartney and John Lemon, *Computers and the Humanities*, 30, 257-265, **1996**.

[104] Forsyth, R. S., Holmes, D. I., Feature-finding for Text Classification, *Literary and Linguistic Computing*, 11, 163-174, **1996**.

[105] Merriam, T., Heterogeneous authorship in Early Shakespeare and the problem of *Henry V*, *Literary and Linguistic Computing*, 13, 15-28, **1998**.

[106] Tweedie, F. J., Baayen, H., How Variable may a Constant be? Measures of Lexical Richness in Perspective, *Computers and the Humanities*, 32, 323-53, **1998**.

[107] Elliott, W., Valenza, R. J., The Professor Doth Protest Too Much, Methinks: Problems with the Foster 'Response', *Computers and the Humanities*, 32, 425-490, **1998**.

[108] Hoorn, J. F., Frank, S. L., Kowalczyk, W., van der Ham, F., Neural Network Identification of Poets using Letter Sequences, *Literary and Linguistic Computing*, 14(3), 311-338, **1999**.

[109] Binongo, J. N. G., Smith, M. W. A., The Application of Principal Component Analysis to Stylometry, *Literary and Linguistic Computing*, 14, 445-465, **1999**.

[110] Foster, D., The Claremont Shakespeare Authorship Clinic: How Severe Are the Problems, *Computers and the Humanities*, 32, 491-510, **1999**.

[111] Merriam, T., Edward III, *Literary and Linguistic Computing*, 15, 157-186, **2000**.

[112] Waugh, S., Adams, A., Tweedie, F. J., Computational stylistics Using Artificial Neural Networks, *Literary and Linguistic computing*, 15, 187-198, **2000**.

[113] Stamatatos, E., Fakotakis, N., Kokkinakis, G., Automatic Text Categorization in Terms of Genre and Author, *Computational Linguistics*, 26, 471-495, **2000**.

[114] Stamatatos, E., Fakotakis, N., Kokkinakis, G., Computer-Based Authorship Attribution Without Lexical Measures, *Computers and the Humanities*, 35, 193-214, **2001**.

[115] Baayen, R. H., *Word Frequency Distributions*, Kluwer Academic Publishers, Dordrecht, 2001.

[116] Burrows, J. F., Craig, H., Lucy Hutchinson and the Authorship of Two Seventeenth-Century Poems: A computational Approach, *The Seventeenth Century*, 16(2), 259-282, **2001**.

[117] de Vel, O., Anderson, A., Corney, M., Mohay, G. M., Mining E-mail Content for Author identification Forensics, *SIGMOD Record*, 30, 55-64, **2001**.

[118] Khmelev, D. V., Tweedie, F. J., Using Markov Chains for Identification of Writers, *Literary and Linguistic Computing*, 16, 299-308, **2001**.

[119] Chaski, C. E., Empirical Evaluation of Language-Based Author Identification Techniques, *Forensic Linguistics*, 8 (1), 1-65, **2001**.

[120] Grant, T. D., Baker, K. L., Identifying reliable, valid markers of authorship: a response to Chaski, *Forensic Linguistics*, 8(1), 67-79, **2001**.

[121] Holmes, D. I., Gordon, I., Wilson, C., A Widow and her Soldier: Stylometry and the American Civil War, *Literary and Linguistic Computing*, 16(4), 403-420, **2001**.

[122] Anonim, George Pickett, http://en.wikipedia.org/wiki/George_Pickett, (Ocak, **2014**).

[123] Holmes, D. I., Robertson, M., Paez, R., Stephen Crane and the New-York Tribune: A Case Study in Traditional and Non-Traditional Authorship Attribution, *Computers in the Humanities*, 35(3), 315-331, **2001**.

[124] Kukushkina, O. V., Polikarpov, A. A., Khmelev, D. V., Using Literal and Grammatical Statistics for Authorship Attribution, *Problems of Information Transmission*, 37(2), 172- 184, **2001**.

[125] Hoover, D. L., Statistical Stylistics and authorship Attribution: an Empirical Investigation, *Literary and Linguistic Computing*, 16(4), 421-443, **2001**.

[126] Elliott, W., Valenza, R. J., Smoking Guns and Silver Bullets: Could John Ford have written the Funeral Elegy?, *Literary and Linguistic Computing,* 16, pp. 205-232, **2001**.

[127] Hoover, D. L., Frequent Word Sequences and Statistical Stylistic, *Literary and Linguistic Computing,* 17(2), 157-180, **2002**.

[128] Hoover, D. L., Frequent Collocations and Authorial Style, *Literary and Linguistic Computing*, 18(3), 261-286, **2003**.

[129] Elliott, W., Valenza, R. J., So Many Hardballs, So Few Over the Plate, *Computers and the Humanities*, 36(4), 455-460, **2002**.

[130] Clement, R., Sharp, D., Ngram and Bayesian Classification of Documents, *Literary and Linguistic Computing*, 18(4), 423-447, **2003**.

[131] Diederich, J.,, Kindermann, J., Leopold, E., Paass, G., Authorship attribution with support vector machines, *Applied Intelligence*, 19(1), 109–123, **2003**.

[132] Anonim, List of Oz books, http://en.wikipedia.org/wiki/List_of_Oz_books, (Ocak, **2014**)

[133] Binongo, J. N. G., Who wrote the 15th Book of Oz? An application of multivariate analysis to authorship attribution, *Chance*, 16(2), 9–17, **2003**.

[134] Hoover, D. L., Multivariate analysis and the study of style variation, *Literary and Linguistic Computing*, 18, 341–360, **2003**.

[135] Hoover, D. L., Another perspective on vocabulary richness, *Computers and the Humanities*, 37, 151–178, **2003**.

[136] Koppel, M., Schler, J., Exploiting stylistic idiosyncrasies for authorship attribution, *In Proceedings of the IJCAI 2003 Workshop on Computational Approaches to Style Analysis and Synthesis*, 69–72., **2003**.

[137] Peng, F., Schuurmans, D., Keselj, V., Wang, S., Language independent authorship attribution using character level language models, In Proceedings of the *Tenth Conference of the European Chapter of the Association for Computational Linguistics*, 267-274, Morristown, New Jersey, **2003**.

[138] Keselj, V., Peng, F., Cercone, N., Thomas, C. N-gram-based author profiles for authorship attribution, *In Proceedings of Pacific Association for Computational Linguistics 2003*, 255–264, **2003**.

[139] Anonim, British National Corpus, http://www.natcorp.ox.ac.uk/corpus/index.xml, (Ocak, **2014**).

[140] Argamon, S., Koppel, M., Fine, J., Shimoni, A., Gender, genre, and writing style in formal written texts, *Text*, 23(3), 321–346, **2003**.

[141] Hoover, D. L., Testing Burrows's Delta, *Literary and Linguistic Computing*, 19(4), 453–475, **2004**.

[142] Hoover, D. L., Delta prime?, *Literary and Linguistic Computing*, 19(4), 477–495, **2004**.

[143] Peng, F., Schuurmans, D., Wang, S., Augmenting Naive Bayes text classifier with statistical language models, *Information Retrieval*, 7(3–4), 317–345, **2004**.

[144] van Halteren, H., Linguistic profiling for authorship recognition and verification, *In Proceedings of the 42nd Conference of the Association for Computational Linguistics*, Barcelona, Spain, 199–206, **2004**.

[145] Patton, J. M., Can, F., A Stylometric Analysis of Yaşar Kemal's İ*nce Memed* Tetralogy, *Computers and the Humanities*, 38(4), 457-467, **2004.**

[146] Collins, J., Kaufer, D., Vlachos, P., Butler, B., Ishizaki, S., Detecting Collaborations in Text: Comparing the Authors' Rhetorical Language Choices in The Federalist Papers*, Computers and the Humanities*, 38(1), 15-36, **2004**.

[147] Abbasi, A., Chen, H., Applying authorship analysis to extremist group Web forum messages, *IEEE Intelligent Systems*, 20(5), 67–75, **2005**.

[148] Chaski, C. E., Who's at the keyboard? Authorship attribution in digital evidence investigations, *International Journal of Digital Evidence*, 4(1), 1–13, **2005**.

[149] Juola, P., Baayen, H., A controlled-corpus experiment in authorship identification by cross-entropy, *Literary and Linguistic Computing*, 20(Suppl), 59–67, **2005**.

[150] Burns, K., Bayesian inference in disputed authorship: A case study of cognitive errors and a new system for decision support, *Information Sciences*, 176(11), 1570–1589, **2006**.

[151] Zhao, Y., Zobel, J., Effective authorship attribution using function words, *In Proceedings of the 2nd AIRS Asian Information Retrieval Symposium* 174–190, Jeju Island, Korea, 2005.

[152] Koppel, M., Schler, J., Zigdon, K., Determining an author's native language by mining a text for errors, In Proceedings of KDD 2005, 624–628, Chicago, IL, **2005**.

[153] Madigan, D., Genkin, A., Lewis, D. D., Argamon, S., Fradkin, D., Ye, L., Author identification on the large scale, *In Proceedings of the Meeting of the Classification Society of North America*, Piscataway, NJ, **2005**.

[154] Koppel, M., Mughaz, D., Akiva, N., New methods for attribution of Rabbinic literature. Hebrew Linguistics: A Journal for Hebrew Descriptive, *Computational and Applied Linguistics*, 57, 5–18, **2006**.

[155] Zhao, Y., Zobel, J., Vines, P., Using relative entropy for authorship attribution, *In Proceedings of the 3rd Asia Conference on Information Retrieval Technology*, 92–105, Singapore, **2006**.

[156] Zheng, R., Li, J., Chen, H., Huang, Z., 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, **2006**.

[157] Amasyalı, M. F., Diri, B., Automatic Turkish Text Categorization in Terms of Author, Genre and Gender, *In NLDB'06 Proceedings of the 11th international conference on Applications of Natural Language to Information Systems*, 221-226, Klagenfurt, Austria, **2006**.

[158] Argamon, S., Whitelaw, C., Chase, P., Hota, S., Garg, N., Levitan, S., Stylistic text classification using functional lexical features, *Journal of the American Society for Information Science and Technology*, 58(6), 802–821, **2007**.

[159] Burrows, J. F., All the way through: Testing for authorship in different frequency strata, *Literary and Linguistic Computing*, 21, 27–47, **2007**.

[160] Hirst, G., Feiguina, O., Bigrams of syntactic labels for authorship discrimination of short texts, *Literary and Linguistic Computing*, 22(4), 405–417, **2007**.

[161] Pavelec, D., Justino, E., Oliveira, L. S., Author identification using stylometric features, *Inteligencia Artificial*, 11(36), 59–65, **2007**.

[162] Zhao, Y., Zobel, J., Searching with style: Authorship attribution in classic literature, *In Proceedings of the 30th Australasian Conference on Computer Science,* 62, 59–68, Ballarat, Australia, **2007**.

[163] Türkoğlu, F., Diri, B., Amasyalı, M. F., Author attribution of Turkish texts by feature mining, *In Proceedings of the 3rd International Conference on Intelligent Computing: Advanced Intelligent Computing Theories and Applications. With Aspects of Artificial Intelligence,* 1086–1093, Qingdao, China, **2007**.

[164] Stańczyk, U., Cyran, K. A., Machine learning approach to authorship attribution of literary texts, *Journal of Applied Mathematics and Informatics*, 1(4), 151–158, **2007**.

[165] Stańczyk, U., On construction of optimised rough set-based classifier, *International Journal of Mathematical Models and Methods in Applied Sciences*, 1, 533-542, **2008**.

[166] Stańczyk, U., Dominance-based rough set approach employed in search of authorial invariants, *Computer Recognition Systems 3, Advances in Intelligent and Soft Computing*, 57, 293-301, **2009**.

[167] Abbasi, A., Chen, H., Writeprints: A stylometric approach to identity-level identification and similarity detection, *ACM Transactions on Information Systems*, 26(2), 1–29, **2008**.

[168] Tearle, M., Taylor, K., Demuth, H., An algorithm for automated authorship attribution using neural networks, *Literary and Linguist Computing*, 23(4), 425-442, **2008**.

[169] Jockers, M. L., Witten, D. M., Criddle, C. S., Reassessing authorship of the Book of Mormon using delta and nearest shrunken centroid classification, *Literary and Linguist Computing*, 23(4), 465-491, **2008**.

[170] Stamatatos, E., Author identification: Using text sampling to handle the class imbalance problem, *Information Processing and Management*, 44(2), 790–799, **2008**.

[171] Argamon, S., Koppel, M., Pennebaker, J., Schler, J., Automatically profiling the author of an anonymous text, *Communications of the ACM*, 52(2), 119-123, **2009**.

[172] Elliott, W., Valenza, R. J., Two tough nuts to crack: did Shakespeare write the 'Shakespeare' portions of *Sir Thomas More* and *Edward III*? Part I, *Literary and Linguist Computing*, 25(1), 67-83, **2010**.

[173] Elliott, W., Valenza, R. J., Two tough nuts to crack: did Shakespeare write the 'Shakespeare' portions of *Sir Thomas More* and *Edward III*? Part II: Conclusion, *Literary and Linguist Computing*, 25(2), 165-177, **2010**.

[174] Anonim, The Diary of a Public Man, http://en.wikipedia.org/wiki/The_Diary_of_a_Public_Man, (Ocak, **2014**).

[175] Holmes, D. I., Crofts, D. W., The diary of a public man: a case study in traditional and non-traditional authorship attribution, *Literary and Linguist Computing*, 25(2), 179-197, **2010**.

[176] Jockers, M. L., Witten, D. M., A comparative study of machine learning methods for authorship attribution, *Literary and Linguist Computing*, 25(2), 215-223, **2010**.

[177] Aslanturk, O., Sezer, E. A., Sever, H., Raghavan, V., Application of Cascading Rough Set-Based Classifiers on Authorship Attribution. *In Proceedings of the IEEE International Conference on Granular Computing*, 656-660, San Jose, California, USA **2010**.

[178] Koppel, M., Schler, J., Argamon, S., Authorship attribution in the wild, *Language Resources and Evaluation*, 45(1), 83-94, **2011**.

[179] Rybicki, J., Eder, M., Deeper Delta across genres and languages: do we really need the most frequent words?, *Literary and Linguistic Computing*, 26(3), 315-321, **2011**.

[180] Eder, M., Does size matter? Authorship attribution, small samples, big problem, *Literary and Linguist Computing*, Published online, DOI: 10.1093/llc/fqt066, **2013**.

[181] Sayoud, H., Author discrimination between the Holy Quran and Prophet's statements, *Literary and Linguist Computing*, 27(4), 427-444, **2012**.

[182] Dahllöf, M., Automatic prediction of gender, political affiliation, and age in Swedish politicians from the wording of their speeches - A comparative study of classifiability, *Literary and Linguist Computing*, 27(2), 139-153, **2012**.

[183] Savoy, J., Comparative evaluation of term selection functions for authorship attribution, *Literary and Linguistic Computing*, Published online, DOI: 10.1093/llc/fqt047, **2013**.

[184] Pawlak, Z., Rough Sets, *International Journal of Computer and Information Science,* 11(5), 341-356, **1982**.

[184] Pawlak, Z., *Rough Sets: Theoretical Aspects of Reasoning About Data*, Kluwer Academic Publishers, Norwell, MA, USA. ISBN:0792314727, **1991**.

[185] Suresh, G. V., Reddy, E. V., Reddy, E. S., Uncertain Data Classification Using Rough Set Theory, *In Proceedings of the International Conference on Information Systems Design and Intelligent Applications* 2012, 132, 869-877, Visakhapatnam, India, **2012**.

[186] Øhrn, A., *Discernibility and Rough Sets in Medicine: Tools and Applications*, PhD thesis, Department of Computer and Information Science, Norwegian University of Science and Technology (NTNU), Trondheim, Norway, ISBN 82-7984-014-1, **1999**.

[187] Øhrn, A., Komorowski, J., ROSETTA: A Rough Set Toolkit for Analysis of Data, *In Proceedings of Third International Joint Conference on Information Sciences, Fifth International Workshop on Rough Sets and Soft Computing (RSSC'97),* 3, 403-407, Durham, NC, USA, **1997**.

[188] ROSETTA, A Rough Set toolkit for Analysis of Data, http://www.lcb.uu.se/tools/rosetta/, (Ocak, **2014**)

[189] Zemberek, Open Source NLP Library for Turkic Languages, https://code.google.com/p/zemberek/, (Ocak, **2014**)

[190] Hürriyet Gazetesi Web Sitesi, www.hurriyet.com.tr, (Ocak, **2014**).

[191] Sabah Gazetesi Web Sitesi, www.sabah.com.tr, (Ocak, **2014**).

[192] Vatan Gazetesi Web Sitesi, www.gazetevatan.com, (Ocak, **2014**).

ÖZGEÇMİŞ

Kimlik Bilgileri

Adı Soyadı:	Oğuz Aslantürk
Doğum Yeri:	Ankara
Medeni Hali:	Bekar
E-posta:	oguz.aslanturk@gmail.com
Adresi:	Ümit Mah. 2483. Cad. Bahar Sitesi 12-C Blok Yasemin Apt. No: 10 Ümitköy Yenimahalle Ankara

Eğitim

Lise:	Adana Fen Lisesi, Adana, 1994
Lisans:	Hacettepe Ünv. Bilgisayar Müh. Böl., Ankara, 1998
Yüksek Lisans:	Hacettepe Ünv. Bilgisayar Müh. Böl., Ankara, 2002, Tez Konusu: SCORM Tabanlı Web Tabanlı Uzaktan Eğitim Yönetim Sistemi Geliştirilmesi
Doktora:	Hacettepe Ünv. Bilgisayar Müh. Böl., Ankara, 2014 Tez Konusu: Yazar Tanıma (*Authorship Recognition*)

Yabancı Dil ve Düzeyi

İngilizce:	Çok iyi

İş Deneyimi

Kurum:	Innova, Ankara
Eğitim Konuları:	Java ile Kurumsal Uygulama Geliştirme ve Web Servis eğitimleri/Ocak 2014
Kurum:	Hacettepe Üniversitesi Bilgisayar Müh. Böl.
Pozisyon/Süre:	Yarı zamanlı Öğretim görevlisi/Eylül 2013 – Şubat 2013
Verilen Dersler:	Yazılım Mimarileri
Kurum:	Türkiye Atom Enerjisi Kurumu, ANKARA
Pozisyon/Süre:	Yazılım Mimarı, Proje Yöneticisi – Kasım - Aralık 2013

Teknolojiler: Spring, Hibernate, JSF, PrimeFaces, Apache Tomcat Servlet Engine, IBM DB2

Sorumluluklar: Çeşitli projelerde mimarinin ve teknik altyapının hazırlanması, entegrasyon ve uygulama tasarım, ekip üyelerine teknik destek verilmesi.

NUMA Projesi, TAEK tarafından iki farklı yerleşkede gerçekleştirilen Numune Analizi işlemlerinin yönetildiği, var olan ve halen geliştirilen birkaç başka sistemle entegre çalışan Web tabanlı bir uygulamadır.

Kurum: Türkiye Atom Enerjisi Kurumu, ANKARA

Pozisyon/Süre: JEE Eğitimleri – Temmuz 2013

Eğitim Konuları: Java ile Nesneye Yönelik Programlama, JSF, Hibernate, Spring, Spring Security

Sorumluluklar: Yazılım geliştirme ekibine eğitim verilmesi

Kurum: Çevre ve Orman Bakanlığı, ANKARA

Pozisyon/Süre: Yazılım Mimarı/Ekip Lideri – Mart – Aralık 2013

Teknolojiler: Spring, Hibernate, JSF, PrimeFaces, JasperReports, Oracle DB

Sorumluluklar: Proje revizyonu çerçevesinde altyapı ve raporlamaya ilişkin değişikliklerin yapılması

Orman Köyleri Bilgi Sistemi Projesi, orman köylerini kalkındırma planları çerçevesinde yapılan çalışmalar, kredi tahsisleri ve tahsilatlarını yönetmek amacıyla bakanlık personeli tarafından kullanılan Web tabanlı bir uygulamadır.

Kurum: Hacettepe Üniversitesi, ANKARA

Pozisyon/Süre: Yazılım Mimarı/Ekip Lideri – Mart – Kasım 2013

Teknolojiler: Spring, Hibernate, JSF, PrimeFaces, JasperReports, Postgresql

Sorumluluklar:	Hacettepe Üniversitesi Sağlık Merkezleri Otomasyonu ile entegre çalışacak Öğrenci/Personel Sağlık Tarama Sistemi yazılımının tasarım ve gerçekleştirimi
Kurum:	Sürat Teknoloji, ANKARA
Pozisyon/Süre:	JEE Eğitimleri –Ocak – Mart - 2013
Eğitim Konuları:	Java ile Nesneye Yönelik Programlama, JPA ile Nesne İlişkisel Eşleştirme, EJB, SOAP ve REST Web Servisler
Sorumluluklar:	Yazılım geliştirme ekibine eğitim verilmesi
Kurum:	Mavinci Bilişim Endüstriyel Uygulamalar Ltd. Şti, ANKARA
Pozisyon/Süre:	Yazılım Mimarı – Ekim 2012 – Şubat 2013
Teknolojiler:	Spring, Hibernate, JSF, PrimeFaces, OpenLayers
Sorumluluklar:	Proje mimarisinin ve teknik altyapının hazırlanması, tasarım ve kodlama, ekip üyelerine teknik destek verilmesi AFKEN Projesi, Web ortamında çalışan, Coğrafi Bilgi Sistem Tabanlı bir uygulamadır. Herhangi bir afet durumunda oluşturulan Afet Geçici Çadır/Konteyner Kentlerde barınan sığınmacılar ile ilgili olarak her türlü bilginin yönetilip raporlanabildiği bir uygulamadır.
Kurum:	Başkent Üniversitesi Bilgisayar Müh. Böl.
Pozisyon/Süre:	Yarı zamanlı Öğretim görevlisi/Eylül 2012 – Şubat 2013
Verilen Dersler:	Bilgisayar Yazılımı I (C ile Yapısal Programlama)
Kurum:	Dış Ticaret Müsteşarlığı, ANKARA
Pozisyon/Süre:	Teknik Danışman – Temmuz 2012
Teknolojiler:	Spring, Hibernate, JSF, PrimeFaces, WebLogic Uygulama Sunucusu, Oracle DB, web servisler, e-imza
Sorumluluklar:	Denetime Dayalı Dış Ticaret Sistemi uygulamasının performans iyileştirmesi, bazı altdüzeyli problemlerin çözümü ve proje ekibine teknik destek verilmesi.

Kurum:	TRT, ANKARA
Pozisyon/Süre:	JEE Eğitimleri –Mayıs 2012
Eğitim Konuları:	Java Teknolojileri, JSP/Servlet, JSF, JPA, EJB, Spring, Kurumsal Uygulama Mimarileri ve Teknolojileri, uygulama sunucular, Apache Tomcat
Sorumluluklar:	Personele eğitim verilmesi

Kurum:	Altus Bilişim, ANKARA
Pozisyon/Süre:	Teknik Danışman – Nisan - Mayıs 2012
Eğitim Konuları:	JSF, Hibernate, Spring, Spring Security
Sorumluluklar:	Proje mimarisinin ve teknik altyapının hazırlanması, ekip üyelerine teknik destek verilmesi

Kurum:	Orman Genel Müdürlüğü, ANKARA
Pozisyon/Süre:	Java ve J2EE Eğitimleri – Eylül 2011 – Nisan 2012
Eğitim Konuları:	Java ile Nesneye Yönelik Programlama, J2EE ile Web Programlama, Spring, Spring Security
Sorumluluklar:	Personele eğitim verilmesi

Kurum:	OYAK Teknoloji, ANKARA
Pozisyon/Süre:	Teknik Danışman, Proje Mimarı, Temmuz – Ekim 2011
Teknolojiler:	Spring, Hibernate, JSF, PrimeFaces, Apache Tomcat Servlet Engine, XACML, Axiomatics, BPMN, Activiti, Jenkins, SOA, ESB, Governance Registry
Sorumluluklar:	Proje altyapısının çeşitli alanlarında çözümler üretmek, proje mimarisinin hazırlanmasına katkıda bulunmak, proje ekibine çeşitli konularda danışmanlık yapmak ODTÜ Bütünleşik Bilgi Sistemi projesi, ODTÜ bünyesinde gerçekleştirilen işlemlerin süreçlerinin tanımlanması, gerçekleştirilmesi, var olan sistemlerin bu sistem içerisine taşınması projesidir.

Kurum:	Çevre ve Orman Bakanlığı, ANKARA

Pozisyon/Süre: Proje Mimarı – Nisan 2011 – Haziran 2011

Teknolojiler: Spring, Hibernate, JSF, PrimeFaces, Apache Tomcat Servlet Engine, Oracle DB

Sorumluluklar: Proje mimarisinin ve altyapısının hazırlanması, tasarım ve kodlama, ekip üyelerine teknik destek verilmesi

Orman Köyleri Bilgi Sistemi Projesi, orman köylerini kalkındırma planları çerçevesinde yapılan çalışmalar, kredi tahsisleri ve tahsilatlarını yönetmek amacıyla bakanlık personeli tarafından kullanılan Web tabanlı bir uygulamadır.

Kurum: Dış Ticaret Müsteşarlığı, ANKARA

Pozisyon/Süre: Ekip Lideri – Eylül – Aralık 2010

Teknolojiler: Spring, Hibernate, JSF, PrimeFaces, WebLogic Uygulama Sunucusu, Oracle DB, web servisler, e-imza

Sorumluluklar: Proje mimarisinin ve altyapısının hazırlanması, tasarım ve kodlama, ekip üyelerine teknik destek verilmesi

Denetime Dayalı Dış Ticaret Sistemi, ithalat ve ihracat süreçlerinde firmaların, gümrük müdürlüklerinin, Dış Ticaret müsteşarlığının, müsteşarlığa bağlı bölge müdürlükleri ile laboratuvar çalışanlarının ve diğer bazı kamu kurumlarının kullanacağı Web tabanlı bir uygulamadır.

Kurum: Avrupa Birliği Eğitim ve Gençlik Programları Başkanlığı, ANKARA

Pozisyon/Süre: Java ve J2EE Eğitimleri – Ağustos 2010

Eğitim Konuları: J2EE ile Web Programlama, Spring, Hibernate, ExtJS

Sorumluluklar: Personele eğitim verilmesi

Kurum: Deniz Kuvvetleri Komutanlığı, ANKARA

Pozisyon/Süre: Java ve J2EE Eğitimleri – Aralık 2009

Eğitim Konuları: Java ile Nesneye Yönelik Programlama, J2EE ile Web Programlama, JSF

Sorumluluklar: Personele eğitim verilmesi

Kurum: Sahil Güvenlik Komutanlığı – MEBS Bşk, ANKARA
Pozisyon/Süre: Java ve J2EE Eğitimleri – Ocak 2009 – Mart 2009
Eğitim Konuları: Java ile Nesneye Yönelik Programlama, J2EE ile Web Programlama, JSF, Spring, Hibernate, JUnit, Maven
Sorumluluklar: Personele eğitim verilmesi, proje danışmanlığı

Kurum: Hacettepe Üniversitesi Bilgisayar Müh. Böl.
Pozisyon/Süre: Öğretim görevlisi/Ocak 2007 – Ağustos 2010
Verilen Dersler: Bilgisayar Programlama I (C ile Yapısal Programlama)
Bilgisayar Programlama II (Java ile Nesneye Yönelik Programlama)
Yazılım Mimarileri
Bilişim Projeleri Yönetimi (Bilişim Enstitüsü)

Kurum: Ankara Ticaret Odası, ANKARA
Pozisyon/Süre: Java ve JEE Eğitimleri – Şubat 2007 – Mayıs 2007
Sorumluluklar: Bilgi işlem personeline eğitim verilmesi

Kurum: Meteksan A.Ş.
Pozisyon/Süre: TÜİK AB Projesi: *Upgrading the Statistical System of Turkey (USST)* – Local Expert/Nisan 2006 – Kasım 2006
Sorumluluklar: TÜİK için geliştirilen Web tabanlı projelerde teknik liderlik/danışmanlık

Kurum: Meteksan A.Ş.
Pozisyon/Süre: Telekom Projesi – Yazılım Mühendisi, Mayıs–Temmuz, 2006
Sorumluluklar: Telekom ADSL ihalesi için test uygulamalarının geliştirilmesi.
ROTA (Web Tabanlı Abone/Santral Yönetimi) projesi kapsamında bakım çalışmaları.
Kurum: Hacettepe Üniversitesi Bilgisayar Müh. Böl.

Pozisyon/Süre: Ek öğretim görevlisi/Şubat 2006 – Haziran 2006

Kurum: Hacettepe Üniversitesi Sağlık Merkezleri
Pozisyon/Süre: Serbest programcı/Ocak 2006 – Eylül 2006
Sorumluluklar: Sağlık Merkezleri Web Tabanlı Hasta Takip Sistemi'nin J2EE teknolojileri ile geliştirilmesi

Kurum: TEDAŞ – Bilgi İşlem, ANKARA
Pozisyon/Süre: Java ve J2EE Eğitimleri – Aralık 2005
Sorumluluklar: Kurum personeline eğitim verilmesi

Kurum: Çevre Bakanlığı – Bilgi İşlem, ANKARA
Pozisyon/Süre: AB Projesi kapsamında Java ve J2EE Eğitimleri – Kasım, 2005
Sorumluluklar: Kurum personeline eğitim verilmesi

Kurum: AGEM Bilişim Yazılım Danışmanlık Eğitim Hiz. Ltd. Şti., ANKARA
Pozisyon/Süre: Kurucu ortak, Müdür – Şubat 2005 – Aralık 2005
Sorumluluklar: Yönetimsel faaliyetler, proje geliştirme, danışmanlık ve eğitim hizmetleri
Bazı projeler:
ROTA – Telekom ADSL Projesi kapsamında, abonelerin ve elektronik donanımların Web üzerinden yönetilmesini sağlayan bir uygulamadır. Avustralya-NEC şirketi ile birlikte geliştirilmiştir.
YAKAMOZ – Telekom ADSL Projesi kapsamında geliştirilen Web tabanlı keşif uygulamasıdır.

Kurum: Meteksan Sistem ve Bilgisayar Teknolojileri A.Ş., ANKARA
Pozisyon/Süre: J2EE Projeleri Danışmanlığı – Ağustos 2003 – Mart 2005
Sorumluluklar: Çeşitli projelerin tasarım ve gerçekleştirimlerinde teknik destek verilmesi, kurum içi eğitimler verilmesi.

Bazı projeler:

LUCA – Web Tabanlı Muhasebe Sistemi yazılımıdır (www.luca.com.tr). Türkiye Muhasebeciler Odası Birliği (TURMOB) tarafından projelendirilmiş olan uygulama halen belirtilen adreste kullanımdadır.

GÖKKUŞAĞI – Web tabanlı servis yönetimi uygulamasıdır.

Kurum: SYS Ltd Şti., ANKARA
Pozisyon/Süre: J2EE Proje Danışmanlığı – Ekim 2004 – Ocak 2005
Sorumluluklar: Kurum personeline eğitim verilmesi ve danışmanlık

Kurum: Emniyet Genel Md. İstihbarat Daire Başkanlığı Bilgi İşlem, ANKARA
Pozisyon/Süre: J2EE Proje Danışmanlığı – Kasım 2003 – Haziran 2004
Sorumluluklar: Kurum personeline eğitim verilmesi ve danışmanlık

Kurum: Çalışma ve Sosyal Güvenlik Bakanlığı – Bilgi İşlem, ANKARA
Pozisyon/Süre: Java ve J2EE Eğitimleri – Mayıs 2004 – Haziran 2004
Sorumluluklar: Kurum personeline eğitim verilmesi

Kurum: Yöntem Bilişim Teknolojileri Ltd. Şti., ANKARA
Pozisyon/Süre: Java ve J2EE Eğitimleri – Ocak 2004 – Şubat 2004
Sorumluluklar: Kurum personeline eğitim verilmesi

Kurum: Hacettepe Üniversitesi Bilgisayar Müh. Böl., ANKARA
Pozisyon/Süre: Araştırma Görevlisi/Öğretim Görevlisi – 1999 - 2004

Kurum: ASELSAN A.Ş., ANKARA
Pozisyon/Süre: Bilgisayar Mühendisi – 1998 – 1999
Sorumluluklar: Web-tabanlı ve/veya masaüstü kurumiçi uygulamaların geliştirilmesi

Kurum:	Halıcı Yazılım San. A.Ş., ANKARA
Pozisyon/Süre:	Yarı zamanlı programcı/rapor geliştirici – 1997 – 1998
Sorumluluklar:	İLSİS projesi için rapor ve ekranların geliştirilmesi

Kurum:	BİRKO Birleşik Koyunlu Halı Fabrikası A.Ş., NİĞDE
Pozisyon/Süre:	Yaz stajı – 1997
Sorumluluklar:	Kurum içi otomasyon projesinin geliştirilmesi/iyileştirilmesi

Kurum:	NİĞBAŞ. A.Ş., NİĞDE
Pozisyon/Süre:	Yaz stajı – 1996
Sorumluluklar:	Kurum içi otomasyon projesinin geliştirilmesi/iyileştirilmesi

Deneyim Alanları

Proje Yönetimi

Programlama:	C, C++, Java, JEE, çeşitli Web teknolojileri ve uygulama çatıları (JSP, Servlet, Struts, Tiles, JSF, PrimeFaces, RichFaces, Facelets, Hibernate, JPA, Spring, JUnit, ...)
RAD Araçları:	NetBeans, JDeveloper, IntelliJ IDEA, Eclipse, Microsoft Visual C++
J2EE Uyg. Sunucu:	WebLogic AS, WebSphere AS, Oracle AS 10g, Orion, Jboss, Apache Tomcat, Glassfish
Ağ:	LAN/WAN Mimarileri, TCP/IP Soket programlama
Veri tabanı:	Oracle, Postgres, MySQL, PL/SQL, MSSQL
Verdiği Eğitimler:	Bilgisayar Programlama, C, C++, Java, J2EE, Web Tabanlı Programlama, çeşitli uygulama çatıları (Struts, JSF, Tiles, EJB, JPA, Hibernate, Spring, JUnit)
Verdiği Dersler	C ile Yapısal Programlama, C++ ile Nesneye Yönelik Programlama, Java ile Nesneye Yönelik Programlama, J2EE Platformu, J2EE ile Web Tabanlı Uygulama Geliştirme, Yazılım Mimarileri, Bilişim Projeleri Yönetimi

Tezden Üretilmiş Projeler ve Bütçesi

-

Tezden Üretilmiş Yayınlar

Aslanturk, O., Sezer, E. A., Sever, H., Investigating the Effects of Stylistic Features on Authorship Attribution: Turkish Journalists (2007 – 2011), Information Processing & Management, İncelemede.

Tezden Üretilmiş Tebliğ ve/veya Poster Sunumu ile Katıldığı Toplantılar

Aslanturk, O., Sezer, E. A., Sever, H., Raghavan, V., Application of Cascading Rough Set-Based Classifiers on Authorship Attribution. In Proceedings of the IEEE International Conference on Granular Computing, 656-660, San Jose, California, USA 2010.

Aslanturk, O., Sezer, E. A., Sever, H., Investigating Stylistics Features of Turkish Texts. QQML 2014, Kabul edildi, Istanbul, Turkey 2014.