AZƏRBAYCAN RESPUBLİKASI ELM VƏ TƏHSİL NAZİRLİYİ

AZƏRBAYCAN TEXNİKİ UNİVERSİTETİ

YÜKSƏK TƏHSİL İNSTİTUTU

Nizamova Aytac Yusif qızı

Abbasova Kəmalə Bəylər qızı

Həsənov Fariz Rəfail

İnformasiya təhlükəsizliyində maşın öyrənmə alqoritimlərindən istifadə mövzusunda

MAGİSTRLİK DİSSERTASİYASI

060509 Kompüter elmləri

Kompüterli modelləşdirmə

Elmi rəhbər: T.ü.f.d Qurbanova Gülnar Həsən

BAKI – 2024

AZƏRBAYCAN TEXNİKİ UNİVERSİTETİ

YÜKSƏK TƏHSİL İNSTİTUTU

MAGİSTRANTIN ANDI

İnformasiya təhlükəsizliyində maşın öyrənmə alqoritimlərindən istifadə mövzusunda təqdim etdiyimiz (Magistrlik dissertasiyasının mövzusu) magistrlik dissertasiyasını elmi əxlaq normalarına və istinad qaydalarına tam riayət etməklə və istifadə etdiyim bütün mənbələri ədəbiyyat siyahısında əks etdirməklə yazdığıma and içirik və magistrlik dissertasiyasının AzTU Kitabxana İnformasiya Mərkəzində saxlanılması, həmin mərkəz tərəfindən AzTU Rəqəmsal Repozitoriyasına daxil edilərək repozitoriyanın veb saytında yerləşdirilməsinə icazə veririk.

Nizamova Aytac Yusif qızı

Abbasova Kəmalə Bəylər qızı

Həsənov Fariz Rəfail

Tarix

XÜLASƏ

Elektron rabitənin mühüm xidmətlərindən biri də elektron poçtdan istifadədir. İnternetin yaranması və geniş yayılması ilə ən çox istifadə edilən ünsiyyət vasitələrindən birinə çevrilən e-poçt xidmətində baş verən inkişaflar bəzi problemləri də ortaya çıxarmışdır. Bu problemlərdən biri də elektron məktublar vasitəsilə "spam" mesajların yayılmasıdır. Gündən-günə kritik problemə çevrilən spam mesajlar e-poçt xidmətlərinin təhlükəsizliyini və etibarlılığını təhdid edərək istifadəçilərin narahat olmasına səbəb olur. Ortaya çıxan bu təhlükə qarşısında bütün dünyada texniki və hüquqi tədbirlər görülür və problemlə mübarizə üçün səylər göstərilir. Bunun aradan qaldırılması üçün spam filtrasiyası və spam aşkarlama metodunun inkişafı kimi proseslər həyata keçirilir. E-poçtlarında spam e-poçtları aşkar etmək və onların aradan qaldırılması maşın öyrənmə alqoritmlərindən istifadə edilir. Maşın öyrənmə alqoritmlərindən istifadə edərək, e-poçtlarda arzuolunmaz e-poçtların süzgəcdən keçirilməsi məqsədəuyğundur. İndiki vaxtda elektron məktubların göndərilməsi asan olduğundan, onlar təkcə dostlarla ünsiyyət vasitəsi kimi deyil, həm də tez-tez reklam məqsədləri üçün spam e-poçt qutularını məhv etmək üçün bir vasitə kimi böyük populyarlıq qazanmışdır. Bunun üçün Random Forest, Logistic Regression, Naive Bayes, Artificial Neural Networks maşın öyrənmə üsulları təhlil edilmişdir. E-poçtların spam və ya normal kimi təsnif etmək üçün əvvəlcə mətn əsaslı spam və normal e-poçt nümunələrindən ibarət məlumat dəsti istifadə edilmişdir. Sonra hər bir e-poçtun məzmunu təhlil edildi və hər bir e-poçtda görünən müxtəlif sözlər və ya terminlər tapıldı. Bu araşdırmada iki fərqli e-poçt verilənlər bazasında fərqli maşın öyrənmə alqoritmindən istifadə edilərək təhlükəli e-poçtlar aşkar edilməyə çalışıldı. Bu alqoritmlərdən istifadə etməzdən əvvəl verilənlər bazasında əvvəlcədən emal addımları yerinə yetirildi. Daha sonra xüsusiyyət çıxarılması və xüsusiyyət seçimi həyata keçirildi. Xüsusiyyət seçildikdən sonra xüsusiyyət vektoru yaradıldı və maşın başa düşə biləcəyi formatda dəyərlər əldə edildi. Xüsusiyyət vektoru maşın öyrənmə alqoritmləri ilə sınaqdan keçirildi və yaramaz e-poçt filtrləmə prosesi ilə əldə edilən performans nəticələri qiymətləndirildi.

Açar sözlər: E-poçt təsnifatı, spam,spam filtrasiyası, maşın öyrənməsi, kibertəhlükəsizlik, təbii dil emalı, Word Cluster Technique, mətn təsnifatı.

MÜNDƏRİCAT

QRAFİKLƏR

ŞƏKİLLƏR

QISALTMALAR

SPAM - İstənməyən E-poçt

SVM -Vektor maşınlarına dəstək

ARPANET- Qabaqcıl Araşdırma Layihələri Agentliyi Şəbəkəsi

MALWARE - Zərərli proqram təminatı

FTC - Federal Ticarət Komissiyası

BOW - Word Cluster Technique (Söz Çantası)

KNN -K Ən yaxın qonşuluq

ANN - Süni Neyron Şəbəkələri

İst. EP - Spam

USD - ABŞ dolları

TM-Turing Machine

GİRİŞ

Elektron poçt (e-poçt) bu gün dünyada milyonlarla insanın ünsiyyət qurmasına imkan verən ən mühüm və geniş yayılmış kommunikasiya vasitələrindən biridir. İnternetin yaranması və geniş yayılması, informasiya-kommunikasiya texnologiyalarının inkişafı, həyatımızın mühüm tərkib hissəsinə çevrilməsi və demək olar ki, hər bir sahədə istifadə olunması elektron poçt xidmətlərinin inkişafına və yayılmasına müsbət təsir göstərmişdir.İnformasiya cəmiyyəti anlayışının ön plana çıxdığı və informasiyanın ən sürətli şəkildə paylaşılmasının son dərəcə vacib olduğu bir dövrdə e-poçt xidməti təsirli və mühüm elementə çevrilmişdir. İnsanlar arasında anlıq səviyyəyə çatan ünsiyyət, asan istifadə imkanı təqdim edən və poçt xidmətlərini əvəz edən ucuz bir xidmətdir.Elektron poçt xidməti gündəlik həyatın ayrılmaz hissəsinə çevrilmiş və texnoloji inkişaflar bəzi problemləri də ortaya çıxarmışdır. İnternetin ilk xidmətlərindən biri olduğundan, hazırda internet xidmətlərinə gəldikdə son dərəcə ehtiyac duyulan təhlükəsizlik və autentifikasiya kimi tələblər nəzərə alınmamışdır. Buna görə də elektron poçt xidməti bu gün internetin ən böyük problemlərini özündə saxlayır .Əsas problem spam mesajlardır.Elektron poçt xidmətlərində əsas problem olan spam mesajların sayı günü-gündən artır və təhlükəli hala gəlir. Spam e-poçtların tarixinə nəzər saldıqda görürük ki, ilk spam e-poçt 1978-ci ildə ARPANET üzərindən göndərilib [5]. Spam, alıcı ilə cari əlaqəsi olmayan göndərici tərəfindən elektron şəkildə göndərilən istənməyən və arzuolunmaz mesaj kimi müəyyən edilə bilər . Elektron spamın bir neçə alt dəsti mövcuddur. Spam mesajları e-poçt, SMS, sosial şəbəkələr və ya onlayn alış-veriş platformaları kimi bir çox rabitə kanalları vasitəsilə göndərilə bilər. Spam istifadəçilərin vaxtını sərf edir, çünki istifadəçilər arzuolunmaz mesajları müəyyən etməli və silməlidirlər, o, həmçinin məhdud poçt qutusunda yer tutur və mühüm şəxsi e-poçtları basdırır .Spam mesajları əl ilə və ya avtomatik olaraq süzülə bilər. Aydındır ki, spam mesajını müəyyən etməklə və onu silməklə manual spam filtrasiyası çox vaxt aparan bir işdir. Üstəlik spam mesajlarında fişinq veb-saytlarına və ya zərərli proqram təminatı olan serverlərə keçidlər kimi təhlükələr ola bilər. Buna görə də, bir neçə onilliklər ərzində tədqiqatlar və praktiklər avtomatik

spam filtrləmə alqoritmlərinin təkmilləşdirilməsi üzərində işləmişlər. Maşın öyrənmə üsulları spam mesajlarının aşkarlanmasında yüksək dəqiqliyə malikdir. Hər bir sözə uyğun olaraq, spam göndərənlər aşkarlanma ehtimalını azaltmaq üçün ümumi qanuni mesajları spam mesajına daxil etməyə meyllidirlər. Neyron şəbəkələri (Ni) Support vektor maşınları (SVM) kimi spam filtrinə tətbiq olunan bir sıra mövcud maşın öyrənmə alqoritmləri mövcuddur. Naive Bayes (NB) və təsadüfi meşə (RF) Kaur və digərlərinin sorğusuna əsasən (2018), təsadüfi meşə kimi öyrənmə üsulları ənənəvi tək təsnifatçılardan üstündür. Son sübutlar göstərdi ki, nizamlama üsulları ilə təchiz olunmuş NN-lər e-poçt və SMS spamını aşkar etməkdə yüksək dəqiqliyə malik ola bilər .

Ümumiyyətlə, spam filtrləmə tapşırığı binar təsnifat probleminə aiddir, hər bir mesaj ya spam ya da vetçina kimi müəyyən edilməlidir. Yüksək dəqiqliklə yanaşı, spam filtrləmə alqoritmləri də yalan müsbət nisbətə gəldikdə yaxşı fəaliyyət göstərməlidir. Bundan əlavə, dəqiqlikdən istifadə edərək ənənəvi təsnifat performans ölçüsü I və II növ səhvlərlə bağlı müxtəlif xərcləri nəzərə almır. Çox vaxt yüksək balanssız spam məlumat dəstləri üçün dəqiqlikdən istifadə də səhv nəticələrə səbəb ola bilər, çünki azlıq sinfi (adətən spam mesajlar sinfi) qanuni mesajların əksəriyyəti sinfi ilə müqayisədə dəqiqliyə az təsir göstərir. Buna görə də, spam filtrləmə alqoritmlərini qiymətləndirərkən çoxlu performans ölçüləri nəzərə alınmalıdır.

Yuxarıda qeyd edildiyi kimi, məzmuna əsaslanan maşın öyrənmə modellərinin əsas ideyası söz (ifadə) siyahısı yaratmaq və hər bir söz və ya ifadəyə (sözlər çantası) və ya söz kateqoriyasına (nitqin hissələrinin etiketlənməsi və ya) çəki təyin etməkdir..Bu dissertasiya işi e-poçt, SMS, sosial şəbəkə mesajları və onlayn rəylərin semantik təsvirini əldə etmək üçün söz əlavələrindən istifadə edir. 2016-cı ilin əvvəlinə nəzər saldıqda, spam şirkətlərin daha çox zərərli makroları ehtiva edən Office sənədlərindən istifadə etdiyi görüldü. Bu baxımdan zərərli proqram təminatının əsas distribyutorlarından biri kimi tanınan Necurs JavaScript və Office makro əlavələri vasitəsilə zərərli proqramların yayılmasına böyük təsir göstərir (Internet Security Threat Report. https://www.symantec.com/content/dam/symantec/docs/reports/istr-

22-2017- en.pdf, (Şubat 2018)). Hücumların şirkətlərə təsirlərinə baxdığımızda kiçik şirkətlərə göndərilən spam e-poçtların sayının böyük şirkətlərə göndərilən spam e-poçtlardan çox az olduğunu görmək olar. Bu onu göstərir ki, hücumlar böyük miqyasda mənfəət əldə edəcəkləri şirkətlərə yönələcək.Spam e-poçtları müəyyən etmək istədikdə deyə bilərik ki, bu, aşağıdakı üç meyardan birinə cavab verən istənilən e-poçtdur. Bu meyarlar aşağıda verilmişdir (GAO, (Eylül, 2016)):

* Anonimlik: Göndərənin ünvanı və şəxsiyyəti məxfidir.

* Kütləvi poçt: E-poçtlar reklam və marketinq kimi məqsədlər üçün kütləvi qruplara göndərilir.

* İstenmeyen: E-poçtlar alıcı tərəfindən istənmir.

Bir çox insanlar gələnlər qutusundakı arzuolunmaz (spam) e-poçtu asanlıqla tanıya və görməməzliyə vura bilərlər. Çox vaxt bu e-poçtlar narahat olmaya bilər, çünki insanlar spam e-poçtlarla bağlı təhlükələrdən xəbərsizdirlər və aldıqları spam e-poçtların sayı azdır. Bununla belə, yalnız bir gələn e-poçt olsa belə, bu, spam e-poçt təhlükəsini aradan qaldırmır. Tək bir spam mesajına cavab verməklə spam göndərənlərin tələb etdiyi maddi və ya şəxsi məlumatları təqdim etmək mümkündür. Nəticə spam göndərənlər üçün qazanc və alıcılar üçün maliyyə itkisi ola bilər.Xərc nöqteyi-nəzərindən ağla gəlir ki, istənməyən e-poçtlara reaksiya verəcək kiçik bir auditoriya üçün bu qədər insana e-poçt göndərməyin mənası varmı?.Ümumilikdə spamer nə qədər çox insana çatırsa, alıcıların e-poçta cavab vermə ehtimalı bir o qədər yüksəkdir. Digər tərəfdən, bu e-poçtların göndərilməsinin dəyəri kifayət qədər aşağıdır, çünki spam göndərənlər öz serverlərindən istifadə edirlər və ya ucuz Proksi serverləri icarəyə götürürlər (GAO, (Eylül, 2016)).Hüquqi nöqteyi-nəzərdən ABŞ Federal Hökuməti, xüsusən də Federal Ticarət Komissiyası (FTC) spam e-poçt məsələsini gündəmə gətirdi və 2003-cü ildə CAN-SPAM federal qanunvericiliyini qəbul etdi. FIC-nin vəzifəsi istehlakçı hüquqlarını qorumaqdır. Spam e-poçtlar istehlakçıları iki fərqli şəkildə riskə ata bilər (GAO, (Eylül, 2016)):

1. Maliyyə və Məxfilik Riskləri: Məqsəd alıcının maliyyə məlumatlarını əldə etmək olduğundan, əksər istənməyən e-poçtlar alıcıdan kredit kartı nömrələri və ya sosial təminat nömrələri kimi şəxsi məlumatları tələb edir. Bu məlumatlar daha sonra şəxsiyyət oğurluğu, kredit kartı fırıldaqçılığı və s səbəb olur.
2. Uşaqları qorumaq: E-poçt göndərən şəxsin e-poçt göndərilən istifadəçinin yaşını bilməməsi üçün heç bir yol yoxdur. CAN-SPAM-uşaqlar üçün uyğun olan spamı aradan qaldırmaq üçün hazırlanmışdır (GAO, (Eylül, 2016)).

Qısacası, CAN-SPAM e-poçt hesabınıza spam axınını dayandırmaq üçün filtr rolunu oynamasa da, əksər e-poçt xidməti təminatçılarında arzuolunmaz e-poçtları gələnlər qutusundan uzaqlaşdırmaq üçün qorxu yaradır. Əks halda cəza tətbiq edilir. Marketinq məqsədləri üçün göndərilən elektron məktubları və tələb olunmayan e-poçtları ayırd etmək üçün CAN-SPAM bildirir ki, aşağıdakı qaydalara əməl edilməlidir (GAO);

- Özünüzü dediyiniz şəxs olun: İstifadəçiyə e-poçtu açmağa və ya e-poçtun lazımsız hala gəlməsinin qarşısını almaq üçün başqa veb-sayt və ya şirkət kimi davrana bilməzsiniz. Bu, onların spam e-poçtlarının filtrlərə tutulmasının qarşısını almaq üçün istifadə etdikləri məşhur hiylədir, lakin bu, qanunsuzdur.
- Mövzu hissəsində dürüst olun.
- Alıcılara reklam olduğunuzu aydınlaşdırın.
- Həqiqi fiziki ünvan təqdim edin: Bu, fırıldaqçı olmadığınızı təmin edir və müştərilərə təsdiqlənmiş əlaqə yolu göndərir.
- Qəbul edənləri sorğu əsasında e-poçt siyahısından çıxmağa istiqamətləndirin.
- Qəbul edənlər tələb edərsə, onları e-poçt siyahınızdan çıxarın.

E-poçt xidməti təminatçılarının əsas müştəriləri e-poçt hesabı olan insanlardır. Əksər e-poçt xidməti təminatçıları gəlirlərini istifadəçinin gələnlər qutusunda keçirdiyi vaxta əsasən artırır. Məsələn, əksər onlayn e-poçt xidməti təminatçıları e-poçt gələnlər qutusunun onlayn versiyası daxilində veb əsaslı reklamlara xidmət göstərir. İstifadəçilərin e-poçt qutusunda keçirdikləri vaxt artdıqca, xidmət provayderinin

təklif etdiyi reklamın üzərinə kliklənmə ehtimalı artır və daha çox reklam göstərmək şansı yaranır. Spam e-poçtlarla dolu gələnlər qutusunda vaxt itirmək əvəzinə, istifadəçi həqiqətən vacib e-poçtlarına daxil olmaq üçün daha az vaxt sərf edəcək və daha yaxşı təcrübə əldə edəcək başqa bir e-poçt xidməti təminatçısına keçə bilər. Buna görə e-poçt xidməti təminatçıları yaxşı spam filtrlərinə sahib olmağa çalışırlar.

1.1 Dissertasiyanın məqsəd və vəzifələri

Arzuolunmaz e-poçtların təsnifləşdirilməsi üçün aparılan araşdırmalar araşdırıldığında, e-poçtların başlıq və əsas hissələrində müxtəlif texnikalar və müxtəlif məlumatlardan istifadə edilməklə təsnifat aparıldığı bildirildi. Bu araşdırmalardan fərqli olaraq tədqiqatın əsas məqsədi e-poçtların məzmununda olan keçidlərin mətinlərinə diqqət yetirməkdir.Tədqiqatda verilənlər toplularının yaradılmasında Word Group Technique sinfinin icazə və maşın öyrənmə alqoritmləri istifadə olunur. BOW nəticəsində yaranan müxtəlif uzunluqlu N qramın təsnifatı spam təsnifatı üçün müxtəlif maşın öyrənmə üsullarının performansa təsiri və müvəffəqiyyəti təhlil edilmişdir.Bununla əlaqədar olaraq dissertasiyanın məqsəd və vəzifələri aşağıdakılardır:

- İstənməyən e-poçtlar, istənməyən e-poçtların təhlükələri, xərclər və spamların göndərilməsi zamanı hədəflənən əsas məqsədlər haqqında məlumatların verilməsi.
- İstənməyən e-poçtları ayırd etmək üçün aparılan işlər haqqında məlumatların verilməsi və bu işlərdə istifadə edilən üsul və modellərin araşdırılması.
- Ədəbiyyatdakı tədqiqatlardan fərqli olaraq, link mətnlərinə görə arzuolunmaz e-poçtların təsnifləşdirilməsi metodunun müəyyən edilməsi.
- İstənməyən e-poçtların xüsusiyyətlərinin çıxarılmasında istifadə olunan Word Cluster Texnikası və bu texnikanın tətbiqi üçün nəzərdə tutulmuş alqoritm haqqında məlumat vermək.
- Spam e-poçtların təsnifatında müxtəlif maşın öyrənmə üsullarının performansını müşahidə etmək.
- Təsnifat uğurunu ölçən metriklərin araşdırılması.

- Bağlantı mətnlərinə görə təsnifat metodunun tətbiqi və sınaqdan keçirilməsi.
- Təklif olunan modelin mövcud həllərlə müqayisəsi.

Dissertasiyanın əhatə dairəsi

Tədqiqat çərçivəsində, Word Cluster Technique və müxtəlif maşın öyrənmə üsullarından istifadə etməklə istənməyən e-poçtların təsnifat metodu araşdırıldı. Seçilmiş maşın öyrənmə üsulları tətbiq edilərkən, e-poçtlardakı keçidlərin mətnlərindən istifadə edilmişdir ki, bu da ədəbiyyatdakı araşdırmalardan fərqlidir. Təsnifat müvəffəqiyyətini ölçən metrikləri araşdıraraq, yaradılan N-qramların təsnifat uğuruna təsiri və müxtəlif maşın öyrənmə üsullarının spam e-poçtların təsnifatındakı uğuru müqayisə edilmişdir.Tədqiqat spam e-poçtların aşkarlanması üçün bir həll təqdim edir.

1.3 Dissertasiya işinin strukturu

Tədqiqatın strukturu aşağıda verilmişdir.

- Giriş bölməsində; istənməyən elektron məktublar ümumiləşdirilir və işin məqsədi, vəzifələri, əhatə dairəsi və strukturu haqqında məlumat verilir.
- İkinci hissədə; Bu dissertasiyanın mövzusu olan maşın öyrənmə üsulları haqqında məlumat verilir.
- Üçüncü hissədə; İstənməyən elektron məktubları təsnif etmək üçün ədəbiyyatda istifadə olunan anlayışlar, modellər, alətlər və ölçülər qeyd olunur.
- Dördüncü hissədə; Tədqiqatın metodu təqdim olunur.
- Beşinci fəsildə; Həll üçün təklif olunan model həyata keçirilir və əldə edilən nəticələr təqdim edilib müzakirə edilir.
- Altıncı bölmədə tədqiqatın nəticələri təqdim olunur.

I FƏSİL

SAHƏ BİLİKLƏRİ

Maşın Öyrənməsi verilənlər əsasında problemi modelləşdirən alqoritmlərin ümumi adıdır. Ümumiyyətlə Nəzarətli, Nəzarətsiz və Yarı Nəzarətli öyrənməyə bölünür (Brownlee).

Nəzarət olunan Öyrənmə: Bunlar X giriş dəyəri və Y çıxış dəyəri əsasında girişdən çıxışa xəritəçəkmə funksiyasını öyrənən alqoritmlərdir. Başqa sözlə, hər bir verilənlərin hansı sinfə aid olduğu məlum olduqda istifadə olunan alqoritmlərdir (Brownlee, (Nisan, 2016)). Aşağıdakı tənlik f xəritələşdirmə funksiyasını göstərir, burada X giriş və Y çıxışdır.

$$Y=f(X) \qquad (1)$$

(1)tənliyində f uyğunluq funksiyasının əsas məqsədi yeni X girişi gəldikdə Y çıxışını proqnozlaşdırmaqdır. Təlim verilənlər toplusunda bütün verilənlərin hansı sinfə aid olduğu və bu verilənlərin təlim prosesini istiqamətləndirdiyi məlum olduğu üçün ona "Nəzarət olunan təlim" deyilir. Məqbul performans səviyyəsinə çatdıqda öyrənmə prosesi dayanır.

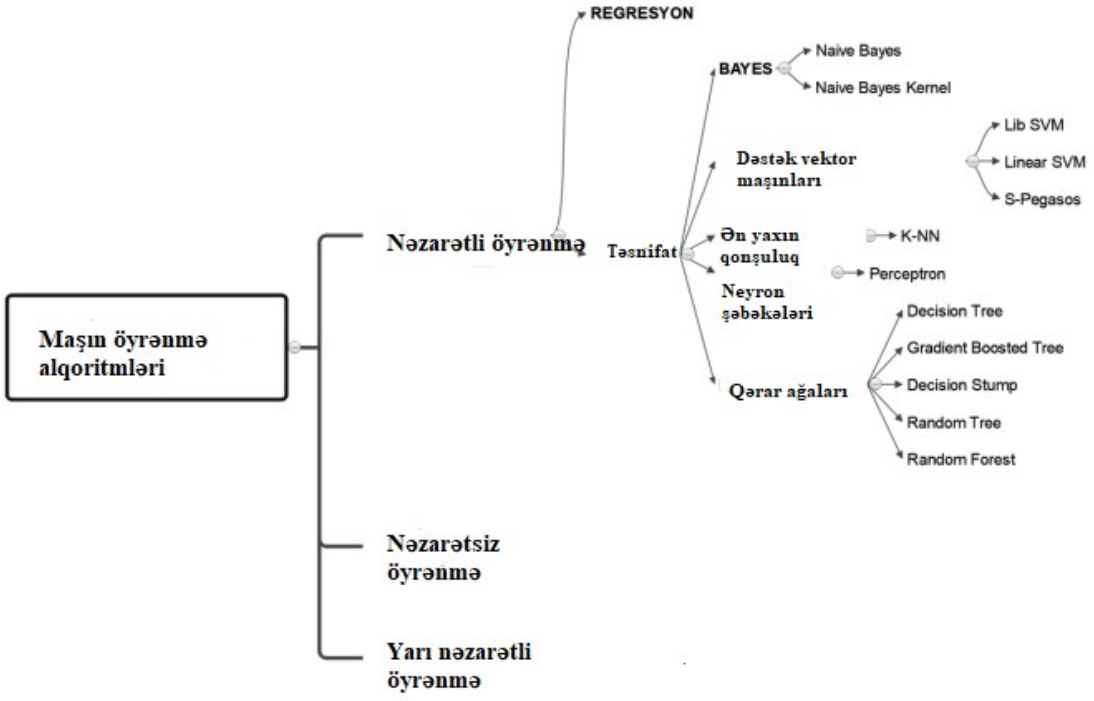
Nəzarətli Öyrənmə: Onları iki qrupa bölmək olar: təsnifat və reqressiya (Brownlee, (Nisan, 2016)). Təsnifat çıxış dəyərinin kateqoriyalı olmasıdır. Məsələn, spam/e-poçt təsnifatı. Reqressiya, çıxış dəyərinin real ədəd olmasıdır.

Nəzarətsiz Öyrənmə: Yalnız X girişinin olduğu və bu X dəyərinə uyğun gələn Y çıxışının naməlum olduğu vəziyyətlərdə öyrənmə üsuludur. Nəzarətsiz öyrənmənin məqsədi bu barədə daha çox öyrənmək üçün məlumatların əsas strukturunu və ya paylanmasını modelləşdirməkdir. Nəzarət olunan Öyrənmədən fərqli olaraq, öyrənmə prosesini istiqamətləndirəcək heç bir məlumat yoxdur. Nəzarətçi olmadan öyrənmə; Klasterləşmə və Assosiasiya olaraq iki yerə bölünür: (Brownlee, (Nisan, 2016)).

Klasterləşmə problemi diskret qrupların aşkar edildiyi yerdir.Assosiasiya məlumatların böyük bir hissəsini təsvir edən qaydaların aşkarlanması problemidir (Brownlee, (Nisan, 2016)).

Yarı Nəzarətli Öyrənmə: X və Y çıxışlarının yalnız bəzilərinin etiketlərinin məlum olduğu öyrənmə üsuludur. O, həm Nəzarət olunan, həm də Nəzarətsiz Öyrənmə üsullarını əhatə edir. Məlumata etiket verilməsi ayrıca xərcə malikdir. Bu mənada, etiketsiz məlumatlar daha ucuzdur və toplamaq və saxlamaq daha asandır. Maşın öyrənmə üsullarının əksəriyyəti bu aspektlərdə praktikliyinə görə bu sahəyə düşür.

Nəzarətsiz Öyrənmə metodlarından giriş strukturunu kəşf etmək və öyrənmək istədiyimiz hallarda, nəzarət edilən öyrənmə metodlarından isə etiketlənməmiş məlumatlar haqqında ən yaxşı proqnozlar vermək istədiyimiz hallarda istifadə edilməlidir (Brownlee, (Nisan, 2016)).Tədqiqat çərçivəsində istifadə edilən bütün məlumatların hansı sinfə aid olduğu məlumdur. Bu səbəbdən yuxarıda izah edilən maşın öyrənmə üsullarından biri olan nəzarət edilən öyrənmə alqoritmlərindən "təsnifat alqoritmləri" tezis çərçivəsində müzakirə edilir.



Şəkil 1.1 Maşın öyrənmə üsulları

Tədqiqat işi çərçivəsində öyrənilən nəzarətli öyrənmə alqoritmləri seçilərkən əsasən təsnifat uğuru nəzərə alınıb. Buna görə, ilk növbədə, təlim aşağıdakı hissələrdə izah edilən 50 limitlə məhdudlaşan 3 qramlıq məlumat dəsti ilə həyata keçirildi. Bu məlumat dəstinin seçilməsinin səbəbi xüsusiyyətlərin sayının digər məlumat

dəstlərindən az olması və qram olaraq orta qiymətdədir. Bu təlimin sonunda müvəffəqiyyət nisbəti 95%-dən yuxarı olan alqoritmlər üçün N qrama görə performans dəyişikliyi yoxlanılıb. Uğur 95%-dən aşağı olan alqoritmlərlə yalnız bir təcrübə aparıldı. N qramın performansa təsirini araşdıran, başqa sözlə 95%-dən çox dəqiqlik göstərən alqoritmlər aşağıda ətraflı şəkildə açıqlanır.

1.1 Bayes Təsnifatçısı

1.1.1 Naive Bayes Təsnifatçısı

Naive Bayes klassifikatoru öz adını 17-ci əsrdə yaşamış İngilis riyaziyyatçısı Tomas Bayesdən götürür [9]. Onun əsası müstəqil dəyişən fərziyyələri ilə məşğul olan Bayes teoremində yatır. Naive Bayes, xüsusilə girişlərin ölçüsü yüksək olduqda, uyğun bir təsnifatçı kimi təyin edilə bilər. Sadə məntiq üzərində qurulsa da, Naive Bayes təsnifatı ümumiyyətlə təəccüblü dərəcədə yaxşı işləyir, buna görə də geniş istifadə olunur. Tipik istifadələr real vaxt proqnozu, spam aşkarlanması, hisslərin təhlili, mətn təsnifatı və s. kimi məsələlərdə istifadə olunur (Miner, (Ocak, 2016)).

Bayes teoremi (2) [8] tənliyindəki kimidir;

$$P(A\backslash B) = \qquad (2)$$

P(A\B): B hadisəsinin baş verdiyi anda A hadisəsinin baş vermə ehtimalı

P(B\A): A hadisəsi baş verdikdə B hadisəsinin baş vermə ehtimalı

P(A) və P(B): A və B hadisələrinin aprior ehtimalları.

Naive Bayes [8] ilə təsnifat apararkən;

1. Məlumat dəsti tezlik cədvəlinə çevrilir.

2. Ehtimal təsnifatı üçün hər bir dəyişənin ehtimalı hesablanır.

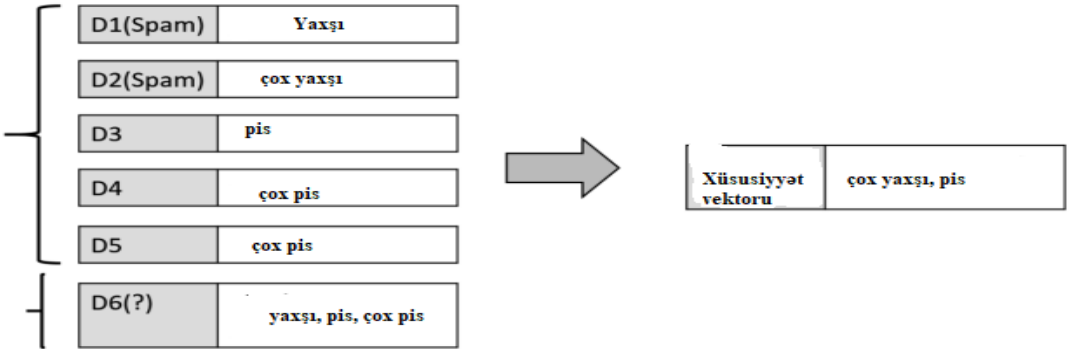
3. Hər bir sinfin ehtimalını hesablamaq üçün sadələvh Bayes tənliyindən istifadə edilir. Ən yüksək ehtimala malik sinif proqnoz nəticəsidir.

Naive Bayes Təsnifatı spam təsnifatında tez-tez üstünlük verilən üsullardan biridir. Spam təsnifatındakı bu teoremi aşağıdakı misalla izah etmək olar [9];

- Hər bir e-poçt sözlərdən ibarət xüsusiyyət vektorları ilə təmsil olunur.
- X xüsusiyyətlərini təmsil etmək üçün əslində 30000 sözdən ibarət xüsusiyyət vektorunda P(xy) olması arzu edilir. P(xy) tənlik (3) kimi hesablanır.

$$P(x_1, \dots \dots \dots . x_{3000}|y) \quad (3)$$
$$=P(x_1|y)P(x_2|y,x_1)P(x_3|y,x_1,x_2) \dots \dots . . P(x_{3000}|y,x_1,x_2 \dots \dots . x_{2999})$$
$$=\prod_{i=1}^{n} P(x_i|y)$$

Nümunə verilənlər toplusu və verilənlər toplusundan yaranan xüsusiyyət vektoru Şəkil 2.2-də göstərildiyi kimidir.



Şəkil 2.2 Naive Bayes Təsnifat nümunəsi verilənlər toplusu

Xüsusiyyət vektoru yaradıldıqdan sonra ehtimallar Cədvəl 1.1-dəki kimi hesablanır.

P(X)=X/(X+Y) (X: Spam, Y: E-poçt)	2/(2+3)=0.4
P(Y) = Y/ (X+Y)=	3/(2+3)=0.6
P("Çox"/X)=	$N_{çox\|X}$ + 1)/{ ($N_{çox\|X}$ +1)+ ($N_{yaxşı\|X}$ + 1)+ (pis=(1+1)/{(1+1)+(2+1)+(0+1)=0.33
P("Yaxşı"/X)=	($N_{yaxşı\ l\|X}$ + 1)/{ ($N_{çox}\|X$ + 1)+ ($N_{yaxşı}$ ($+l\|N$) X + 1)}=(2+1)/{(1+1)+(2+1)+(0+1)=0.5
P("Pis"/X)=	($N_{pis}\|X$ + 1)/{ ($N_{çox}\|X$ + 1)+ ($N_{yaxşı}\|X$ +1)+ (N_{pis} 1)(N_{pis} 1)(1)+ +(2+ 1)+(0+1)=0,17
P("Çox"/Y)=	$N_{çox\|Y}$ + 1)/{ ($N_{çox\|Y}$ + 1)+ ($N_{yaxşı\|Y}$ + 1)+ ($N_{pis\|Y}$ + 1)}x=(3+1)/{(3+1)+0 1)+(4+1)=0,4
P("Yaxşı"/Y)=	($N_{yaxşı\|Y}$ + 1)/{ ($N_{çox\|Y}$ +1)+ ($N_{yaxşı\|Y}$ + 1)+ ($N_{pis\|Y}$ +1)}=(0+1)/{(3+1)+(0+1)+(4+1)=0.1
P("Pis"/Y)=	**($N_{pis\|Y}$ +1)/{ ($N_{çox\|Y}$ + 1)+ ($N_{yaxşı\|Y}$ +1)+ ($N_{pis\|Y}$ + 1)} =(4+1)/{(3+1)+(0+1)+(4+1)=0.5**

Test məlumatları üçün Bayes qaydası D6="yaxşı, pis, çox pis" Cədvəl 1.2-də olduğu kimi hesablanır. Hesablama nəticəsində D6 elektron poçtunun spam olduğu qənaətinə gəlinir.

Cədvəl 1.2 Naive Bayes təsnifatı ilə etiket proqnozu

P(D6|X)=	P(yaxşı|X)P(pis|X)P(çox|X)P(pis|X)=0.5*0.17*0.33*0.17=0.004
P(D6|Y)=	P(yaxşı|Y)P(pis|Y)P(çox|Y)P(pis)=0.1*0.5*0.4*0.5=0.010
P(X|D6)	=P(D6|X)*P(X)/P(D6)=0.0048*0.4/P(D6)
P(Y|D6)	**=P(D6|Y)*P(Y)/P(D6)=0.010*0.6/P(D6)**

Yuxarıdakı nümunədə göründüyü kimi, oxşar üsul müxtəlif xüsusiyyətlərə əsaslanan müxtəlif siniflərin ehtimalını qiymətləndirmək üçün istifadə olunur.

Naive Bayes alqoritminin üstünlükləri və çatışmazlıqlarını aşağıdakı kimi sıralamaq olar [11];

- Məlumatların təsnifatında tez və asanlıqla tətbiq oluna bilər. Çox səviyyəli proqnozlarda yaxşı çıxış edir.
- Müstəqillik fərziyyəsi qüvvədə olduqda, Naive Bayes klassifikatoru logistik reqressiya kimi digər modellərlə daha yaxşı müqayisə edir və daha az təlim məlumatı tələb edir.
- Kateqorik dəyişənlər üçün ədədi dəyişənlərə nisbətən daha yaxşı nəticələr verir.
- Təlim məlumat dəstində olmayan kateqoriyalı dəyişən sınaq zamanı baş verərsə, Naive Bayes modeli sıfır ehtimalını hesablayır və proqnoz verə bilmir. Bu vəziyyət "Sıfır Tezlik" adlanır. Bu vəziyyəti həll etmək üçün "Smoothing Technique" istifadə olunur. Ən sadə düzəliş texnikası "Laplas qiymətləndirməsidir". Laplace: Nate Windows hesablamasında adətən ehtimala +1 əlavə edilir.
- Ehtimalların ilkin dəyərlərinə ehtiyac var. Bu ehtimallar məlum deyilsə, ehtimal adətən mövcud məlumatlar əsasında hesablanır.

- Real həyatda müstəqil dəyişənləri tapmaq həmişə asan olmaya bilər.

1.1.2 Sadə Bayes nüvəsi

Kernel qeyri-parametrik qiymətləndirmə üsullarında istifadə edilən çəki funksiyasıdır. Sıxlığın qiymətləndirilməsində təsadüfi dəyişənlərin sıxlıq funksiyalarını qiymətləndirmək üçün nüvədən istifadə olunur (Miner, (Ocak, 2016)). Bundan əlavə, təsnifat üçün Kernel Density metodundan da istifadə etmək olar (Breheny, (Ocak,2016)). Fərz edək ki, x davamlı dəyərdir və y k müxtəlif kateqoriyalarda qiymət ala bilən diskret dəyərdir. n {xi,, yi} müşahidələri üçün gələcək müşahidələrdə P ($y_1 = j \mid x_1$) proqnozunu vermək üçün x müşahidə edilir, lakin y verilənlərin müşahidə olunmadığı bir üsul əldə etmək istənildikdə, (4) (Breheny, (Ocak,2016)) tənliyində olduğu kimi Naive Bayes Kernel ilə asanlıqla hesablana bilər.

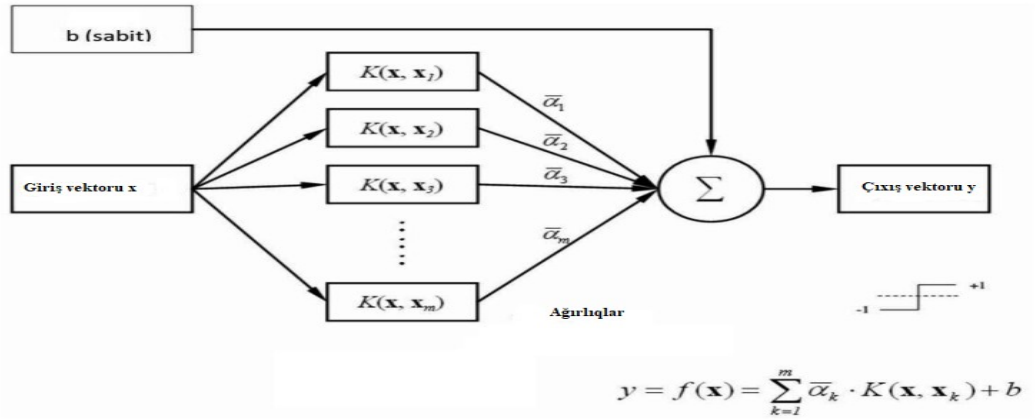
$$P(y=j|x_0)= \quad (4)$$

π,: j sinfinin əvvəlcədən qiymətləndirilməsi (Adətən π; j sinfinə daxil olan nümunələrin nisbətidir) $f(x_o)$: yalnız j sinfindən müşahidələri ehtiva edən nüvə sıxlığına əsaslanaraq xo-da təxmin edilən sıxlığı ifadə edir.Naive Bayes klassifikatorunun üstünlüyü ondan ibarətdir ki, təsnifat üçün tələb olunan dəyişənlərin vasitələrini və fərqlərini qiymətləndirmək üçün yalnız az miqdarda təlim məlumatı tələb olunur. Müstəqil dəyişənlər nəzərdə tutulduğundan, bütün kovariasiya matrisini deyil, yalnız hər bir etiket üçün dəyişənlərin dispersiyalarını müəyyən etmək lazımdır. Naive Bayes operatorundan fərqli olaraq, Naive Bayes (Kernel) operatoru ədədi atributlara tətbiq oluna bilər (Miner, (Ocak, 2016)).

1.2 Dəstək Vektor Maşınları (SVM)

Vapnik tərəfindən hazırlanmış SVM statistik öyrənmə nəzəriyyəsinə, başqa sözlə Vapnik-Chervonenkis (VC) nəzəriyyəsi və struktur riskinə əsaslanır və dəyişənlər arasındakı nümunələrin məlum olmadığı məlumat dəstlərində təsnifat, reqressiya və nümunənin tanınması problemləri üçün istifadə olunur.Bu minimuma endirməyə

əsaslanan nəzarət edilən öyrənmə üsuludur. SVM verilənlərlə bağlı hər hansı birgə paylama funksiyası məlumatını tələb etmədiyi üçün paylanmadan müstəqildir (Ayhan, 2014). SVM-in şəbəkə strukturu Şəkil 1.3-də verilmişdir (Ayhan, 2014). "Burada K(x, x_i) nüvə funksiyalarını, α isə Laqranj çarpanlarını ifadə edir. Girişlərin daxili məhsulları nüvə funksiyalarının köməyi ilə hesablanır. Laqranj çarpanları çəkiləri göstərir.SVM-də nümunənin çıxış dəyəri girişlərin daxili məhsullarının və Laqranj çarpanlarının müstəqil birləşmələrinin cəminə bərabərdir (Ayhan, 2014).

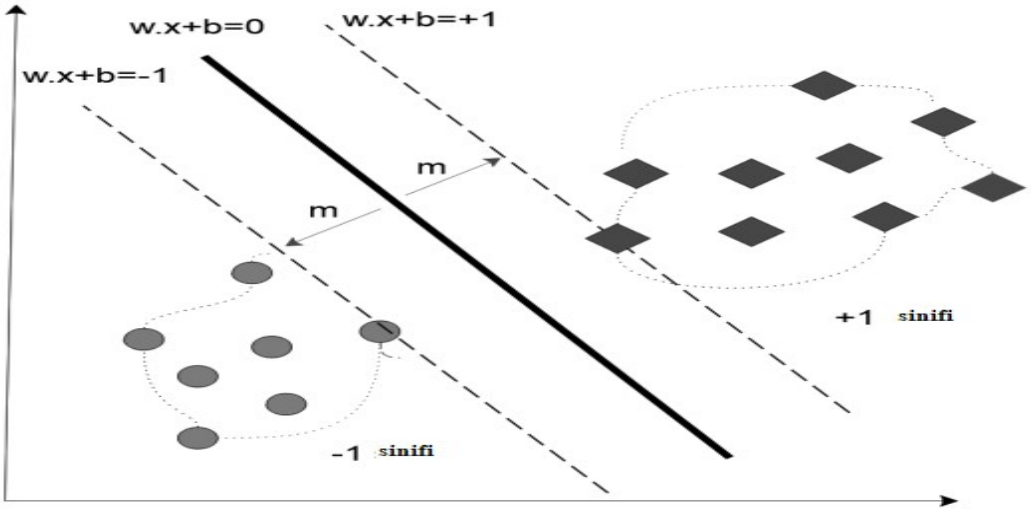


Şəkil 1.3 Dəstək Vektor Maşın şəbəkə strukturu

SVM-nin əsas məqsədi müxtəlif siniflərə aid vektorları bir-birindən uzaqlaşdırmaq və optimal ayırma hipertəpəsini əldə etməkdir.

1.2.1 Xətti SVM

Ən əsas SVM tətbiqi olan xətti SVM-lər maksimum marja çox müstəvili yanaşma ilə həyata keçirilir. Spam e-poçtların təsnifatı kimi 2-sinif və ikiölçülü təsnifat problemi üçün xətti SVM dəstək vektorları Şəkil 1.4-də göstərilmişdir. Dəstək vektorları adlanan vektorlar ayırma müstəvisinə ən yaxın olan hər iki sinfə misal olaraq göstərilmişdir (Ayhan, 2014).



Şəkil 1-4 2-sinif problemlər üçün SVM

Şəkil 4.2-də kəsik-kəsik xətlər kimi göstərilən ayıran çox təyyarəyə paralel çəkilmiş iki bərabər məsafəli çox təyyarələr arasındakı məsafə haşiyə adlanır. Tənlik (5) m və kənar məsafəsi arasındakı bərabərliyi göstərir. X∈p yüksək ölçülü giriş vektoru olduğu xətti SVM-dən istifadə; (x_i, y_i) cütlərindən ibarət məşq dəstini ən yaxşı şəkildə ayıracaq təyyarə tənlik (5)-də olduğu kimi hesablanır:

$$y_i=+1 \text{ üçün } wx_i+b \geq +1 \qquad (5)$$

$$y_i=-1 \text{ üçün } wx_i+b \leq -1$$

$$m=\frac{2}{\sqrt{w*w}} \rightarrow f_{min}(w) = \frac{w*w}{2}$$

x_i: çoxlu müstəvidə istənilən nöqtə w*x+b=0

y_i: sinif etiketləri, yi€{+1,-1}

w: hiperplan normal, çəki vektoru

m: sərhəd müstəvisi arasındakı məsafə və b: sabit

1.2.2 Lib SVM

Hal-hazırda Dəstək Vektor Maşını (SVM) müxtəlif təsnifat problemlərində verdiyi yaxşı nəticələrə görə təsnifat problemlərində geniş istifadə olunur.Standart SVM qeyri-ehtimal, binar, xətti təsnifatçısıdır. Qeyri-xətti təsnifat imkanları əlavə etmək və sinifləri xətti olaraq ayrıla bilən etmək üçün problemi daha yüksək ölçülü xüsusiyyət məkanına çevirən Kernel funksiyaları lazımdır. LibSVM müxtəlif Kernel funksiyaları, çox qatlı təsnifat və fərqli çarpaz doğrulama kimi müxtəlif funksiyalar təklif edir (Athanasopoulos, 2011). Hsu və Çanq (Chang, 2011, p. 27.) tərəfindən hazırlanmış LibSVM, Support Vector Machines-in incəliklərini bilməyən insanlara SVM-i asanlıqla həyata keçirməyə kömək etmək məqsədi daşıyan hazır kitabxanadır. İlk versiyası çıxanda yalnız 2-sinif problemləri dəstəklədi. Daha sonra o, işlənib hazırlanmış və çoxsinifli problemlər və ehtimalların qiymətləndirilməsi imkanları əlavə edilmişdir (Ayhan, 2014).

SVM ilə tanış olmayan istifadəçilər adətən funksiyalar və sinif etiketlərindən ibarət məlumat dəstini vektor maşınları üçün uyğun formata çevirir. Daha sonra təsadüfi olaraq “xətti, çoxhədli, radial əsas funksiya, sigmoid” kimi parametrlərdən birini seçib sınaqdan keçirirlər. Hsu və Chang LibSVM ilə bu vəziyyəti asanlaşdırdı. LibSVM ilə hər bir parametrin optimal dəyəri çarpaz doğrulama ilə hesablanır.SVM-də təsnifat iki hissədə aparıla bilər: nüvə matrisinin hesablanması və təsnifat modelinin yaradılması. Giriş məlumatları böyük olduqda, Kernel matrisini hesablamaq və yaddaşda saxlamaq mümkün deyil. Buna görə həlledici dərhal hesablama prosessor və yaddaş bant genişliyi tələb edir. Çarpaz doğrulama istifadə edildikdə, Kernel matrisinin dəyərlərinin hesablanması bir dəfədən çox təkrarlanmalıdır. Bu nöqtədə, LibSVM kitabxanasındakı “easy.py” skripti bu prosesi standart parametrlərlə asanlaşdırır. O, təsnifat prosesinin sürətini təmin edir (Ayhan, 2014).

1.2.3 Pegasos SVM

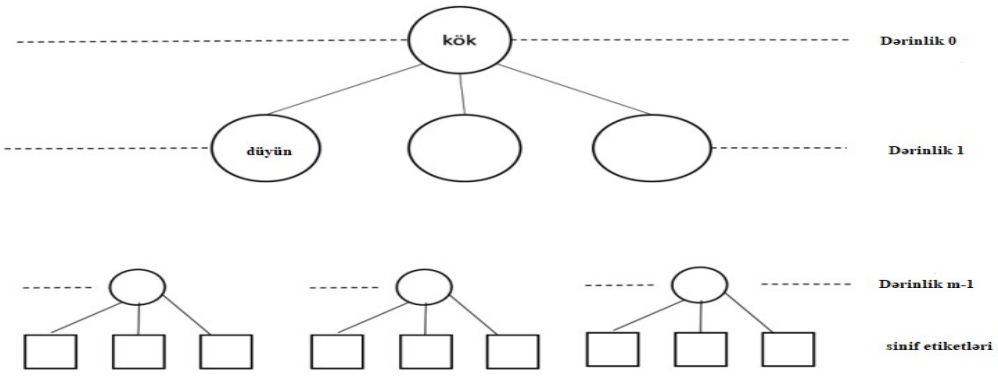
Shai və Yoram (Shalev-Shwartz, 2011) tərəfindən görülən işlərlə ortaya çıxan Pegasos SVM, "İlkin Təxmini Alt Qradient Həlledici" deməkdir. Pegasos Dəstək Vektor Maşınlarında optimallaşdırma problemini həll etmək üçün sadə və effektiv stoxastik “sub-gradient eniş” həllidir.Psevdokodları Şəkil 2.5-də verilmiş Pegasos alqoritmini tətbiq edərkən hər bir təlim addımında təsadüfi təlim məlumatları seçilir. Seçilmiş məlumatlar üçün 15 proqnoz sub-qradientlə aparılır və əks istiqamət əvvəlcədən müəyyən edilmiş addımların sayı ilə aparılır. Burada məqsəd təsadüfi seçilmiş nümunələrdə ən yüksək ehtimalı göstərməkdir.Pegasos böyük mətn təsnifatı problemləri üçün SVM metodları ilə müqayisədə daha sürətli nəticələr verir. Pegasos SVM alqoritmi Şəkil 1.5-də göstərildiyi kimidir:

INPUT: S, λ, T
INITIALIZE: Set $\mathbf{w}_1 = 0$
FOR $t = 1, 2, \ldots, T$
  Choose $i_t \in \{1, \ldots, |S|\}$ uniformly at random.
  Set $\eta_t = \frac{1}{\lambda t}$
  If $y_{i_t} \langle \mathbf{w}_t, \mathbf{x}_{i_t} \rangle < 1$, then:
    Set $\mathbf{w}_{t+1} \leftarrow (1 - \eta_t \lambda)\mathbf{w}_t + \eta_t y_{i_t} \mathbf{x}_{i_t}$
  Else (if $y_{i_t} \langle \mathbf{w}_t, \mathbf{x}_{i_t} \rangle \geq 1$):
    Set $\mathbf{w}_{t+1} \leftarrow (1 - \eta_t \lambda)\mathbf{w}_t$
  [Optional: $\mathbf{w}_{t+1} \leftarrow \min\left\{1, \frac{1/\sqrt{\lambda}}{\|\mathbf{w}_{t+1}\|}\right\} \mathbf{w}_{t+1}$]
OUTPUT: $\mathbf{w}_{T+1}$

Şəkil 1.5 Pegasos Alqoritmi

1.3 Qərar Ağacları

Morgan və Songuist tərəfindən yaradılmış alqoritmdə məlum siniflərə malik nümunə verilənlər sadə qərar qəbul etmə addımları ilə kiçik qruplara bölünür. Hər bir bölmə prosesi ilə oxşar məlumatlar qruplaşdırılır və induktiv üsulla təsnif edilir (Albayrak).Qərar ağaclarının ümumi görünüşü Şəkil 1.6 (Albayrak)-də göstərildiyi kimidir.



Şəkil 1-6 Qərar ağacının nümunəsi

Qərar ağaclarında əsas addım qərar qovşaqlarının yaradılmasıdır. Qərar qovşaqları yaradılarkən ağacın balanslı şəkildə budaqlanması və təsnifat prosesinin düzgün aparılması üçün düyün kimi ən yaxşı atribut seçilməlidir. Bunun üçün bütün sistemdə gözlənilən dəyər Şennon və Uiver [19] tərəfindən təqdim edilən "İnformasiya Qazanma Nəzəriyyəsi" ilə hesablanır. İnformasiya qazancı 6-cı tənlikdəki kimi hesablanır.

$$H=-\sum_{i=1}(p_i \log p_i) \qquad (6)$$

$$\text{Gain(S, A)=H(S) - } \sum_{v \in Values(A)} \frac{|S_v|}{S} H(S_v)$$

p_i: i sinifinin əvvəlki ehtimalı

H: Entropiya

S: Nümunə sahəsi

S_v : Nümunə boşluq alt çoxluğu

Nəzarət olunan öyrənmə alqoritmləri arasında çox yayılmış qərar ağaclarının alqoritm məntiqi Cədvəl 1.3-də göstərildiyi kimidir:

Cədvəl 1-3 Qərar Ağacları Alqoritmi

Məlumat toplusunun yaradılması alqoritmi

Daxiletmə: T (Öyrənmə Seti)
1. Hər bir atribut üçün məlumat qazancını hesablayın
2. "Fərqləndirici atribut" kimi ən yüksək məlumat qazandıran atributu seçin
3. T-ni fərqləndirici xüsusiyyətlə bölün və düyün yaradın.
Əgər(nümunələr eyni sinifə aiddir || nümunələri bölmək üçün heç bir atribut yoxdur || qalan atribut üçün nümunə yoxdur)
{
Prosesi bitir
}
başqa
{
1-ci addıma qayıdın
}

Qərar ağaclarının davamlı və diskret dəyərlər üçün istifadə oluna bilməsi, ağacların yaradılması asanlığı və asan başa düşülən qaydalar kimi üstünlükləri var. Digər tərəfdən, atributları daim proqnozlaşdırmaq qərar ağacları üçün çətindir. Çox sinifli verilənlərdə üstünlük təşkil etmir (Shalev-Shwartz, 2011). "Qərar ağacları yaratmaq üçün hazırlanmış alqoritmlərə CHAID (Chi-Squared Avtomatik Qarşılıqlı Əlaqə Detektoru), Tam CHAID, CRT (Təsnifat və Reqressiya Ağacları), ID3, C4.5, MARS (Çoxvariantlı Adaptiv Reqressiya Splines), QUEST (Quick, 17 Biased, Efficient Statistical Tree), C5.0, SLIQ (Questdə Nəzarət Edilən Öyrənmə), SPRINT (Scalable Paralleizable Induction of Decision Trees) kimi alqoritmlər (Shalev-Shwartz, 2011)".

1.4 Süni neyron şəbəkələri

İlk dəfə 1943-cü illərin əvvəllərində Warren McCulloch və W.A Pitts tərəfindən hazırlanmış və insan beynindən ilhamlanaraq yaradılan Süni Neyron Şəbəkələri (ANN) mövcud məlumatlar əsasında ümumiləşdirmələr aparır və heç vaxt görmədikləri məlumatlar haqqında bu ümumiləşdirmələrə əsaslanaraq qərarlar verə bilirlər. ANN-də hər bir neyron öz məlumat emal strukturuna malikdir və digər neyronlarla çəkili bağlantılar vasitəsilə əlaqələndirilir (Şengöz, (Ocak, 2017)).

ANN-in digər öyrənmə alqoritmlərindən fərqlərini aşağıdakı kimi sıralamaq olar:

- Digər alqoritmlər “girişlər müəyyən edilmiş qaydalara tətbiq edildikdə çıxışlar baş verir”.Məntiqə əsaslanarkən, ANN giriş-çıxış məlumatlarını təmin etməklə qaydaları müəyyən edir.
- ANN təcrübədən faydalanır. Digər alqoritmlərdə məlumat dəqiqdir.
- Hesablama; toplu, asinxron və öyrəndikdən sonra paralel.
- Digər alqoritmlərdən daha yavaşdır.
- Başqa alqoritmlərdə yaddaş paketlənir və hərfi informasiya saxlanılırsa, ANN-də yaddaş ayrılır və bütün şəbəkəyə yayılır.
- Əgər başqa alqoritmlərdə səhvlərə dözümlülük yoxdursa, ANN-də mövcuddur.

Şəkil 2.7 (Şengöz, (Ocak, 2017)) süni sinirin hissələrini göstərir

Girişlər: ($X1$, $X2$,….. Xn) Ətrafdan aldığı məlumatı sinirə gətirir. Girişlər neyron şəbəkəyə əvvəlki neyronlardan və ya xarici dünyadan əlavə edilə bilər.

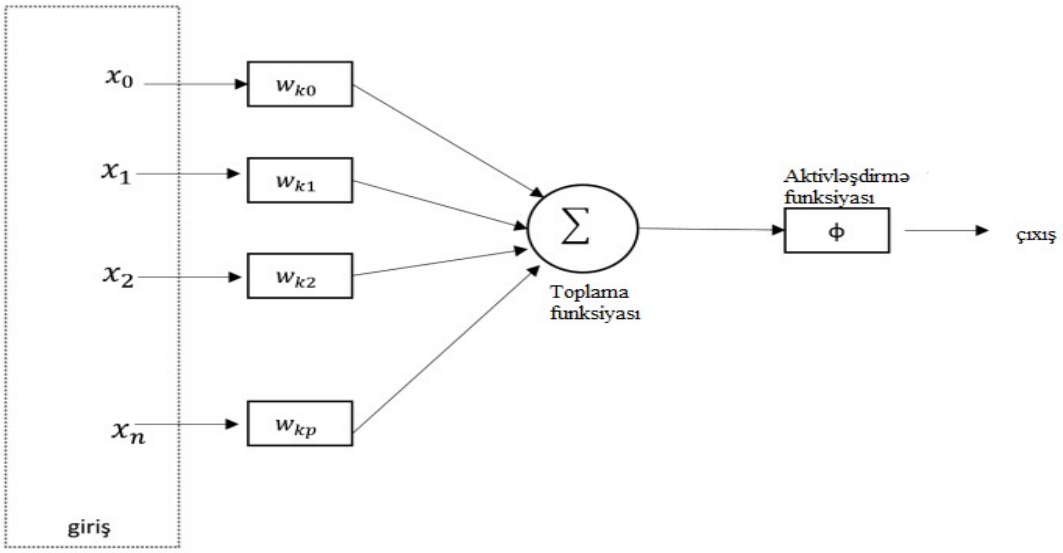
Çəkilər: ($w1$, $w2$,….. wn) süni sinirin qəbul etdiyi girişlərin sinirə təsirini təyin edən uyğun əmsallardır. Hər girişin öz çəkisi var.

Toplama funksiyası: Hər bir çəkinin neyronda aid olduğu girişlərlə çarpmasının cəmini eşik dəyəri ilə toplayır və aktivləşdirmə funksiyasına (fəaliyyət funksiyası) göndərir. Bəzi hallarda toplama funksiyası minimum (min), maksimum (max), çoxluq və ya az ola bilər.Normallaşdırma alqoritmi kimi daha mürəkkəb ola bilər.

Aktivləşdirmə funksiyası (fəaliyyət funksiyası): Əlavə prosesinin nəticəsi fəaliyyət funksiyasından keçir və çıxışa çatdırılır. Səmərəlilik funksiyasının çıxışı y_i giriş vektorları ilə uyğunlaşdırıldıqda (7) tənliyində olduğu kimi müəyyən edilir.

$$y_i = \begin{cases} 1 & w_1 x_1 + w_2 x_2 + w_3 x_3 + \ldots + w_n x_n \geq T \\ 0 & w_1 x_1 + w_2 x_2 + w_3 x_3 + \ldots + w_n x_n \leq T \end{cases} \quad (7)$$

İkili girişlərin nümunəsini nəzərə alsaq, səmərəlilik funksiyası ya sıfır, ya da bir çıxış edəcək.

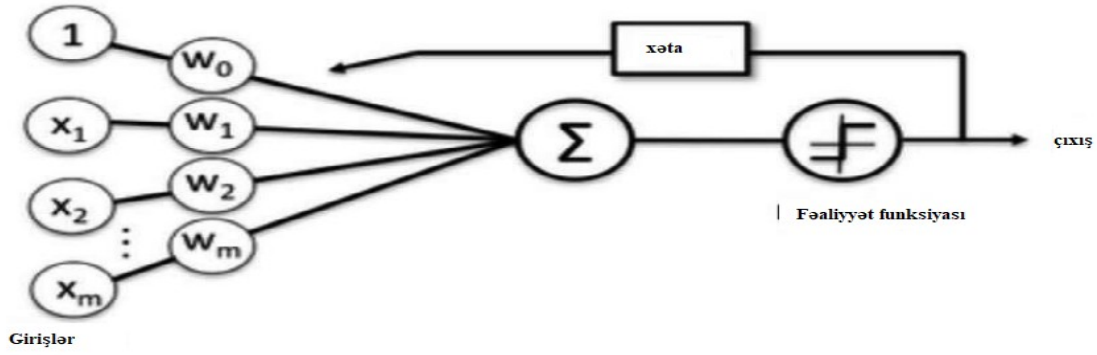


Şəkil 1-7 Süni neyron şəbəkəsi

1.4.1 Perseptron

"F. Neyron şəbəkələri 1960-cı illərdə Rosentblatt tərəfindən təklif edilmişdir. Rosentblatt Perceptronları adi maşınların həll edə bilmədiyi problemləri həll edə bilən "yeni kompüter növü" kimi təsvir etdi (Şengöz, (Ocak, 2017)) "Perceptron beyin funksiyalarını modelləşdirmək üçün istifadə olunur".Bu məqsədlə aparılan tədqiqatlar nəticəsində ortaya çıxan, tək çıxışlı, tək qatlı, öyrədilə bilən süni neyron şəbəkəsidir (Elmas, 2011). Birdən çox girişin alınmasına və tək bir çıxışın istehsalına cavabdehdir. Şəkil 1.8-də göstərildiyi kimi göstərilə bilər. "Perseptron xətti funksiya ilə iki hissəyə bölünə bilən məsələlərdə istifadə edilə bilər. Və ya ,deyil vəziyyətləri bu problemlərə misal olaraq verilə bilər (Kabalcı, (Ocak, 2017))."

Şəkil 1-8 Perseptron nümunəsi



Perseptron daxilolmaların çəkili cəmini (7) tənliyində müəyyən edilmiş həd dəyəri T ilə müqayisə edir. Çəkili cəmi həd dəyərdən böyükdürsə, nəticə 1, əks halda 0-dır.

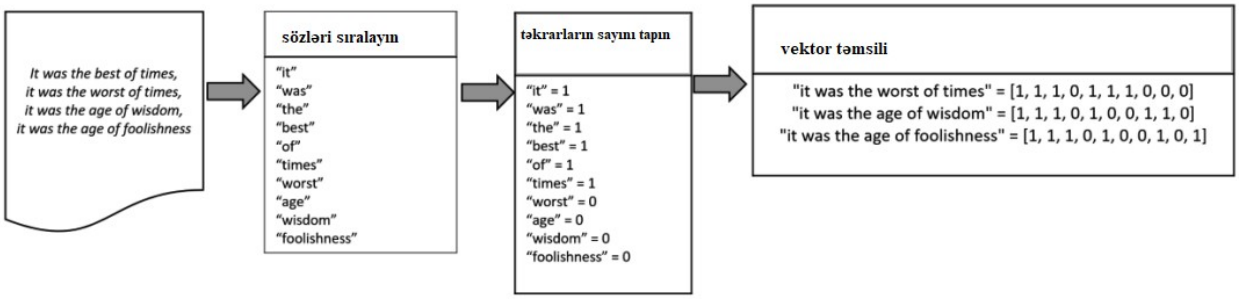
1.5 Ən yaxın qonşu K-NN

T. M. Cover və P. E. Hart tərəfindən təklif edilən K Nearest Neighbor (K-NN) alqoritmi təsnif ediləcək verilənlərin k əvvəlki verilənlərə yaxınlığına baxmaq məntiqinə əsaslanan təsnifat alqoritmidir (Taşcı, 2016). K təsadüfi qiymətdir. Təsnifatlaşdırılacaq məlumatlar əvvəllər təsnif edilmiş və ya başqa sözlə etiketləri artıq məlum olan k ən yaxın verilənləri götürür. Bu elementlər hansı sinfə aiddirsə, təsnif ediləcək element də həmin sinfə aiddir. Elementlər arasındakı məsafə adətən "Evklid məsafəsi" və ya "Manhetten məsafəsi" ilə hesablanır. Nearest Neighbor Alqoritmində təsnifat performansına birbaşa təsir edən parametrlərdən biri də "K" qiymətidir (Bulut, (ASYU2014), 2014). Bulut və Amasyalı tərəfindən aparılan araşdırmada k dəyərinin dəyişdirilməsi ilə k dəyərinin təsnifat performansına təsiri müşahidə edilmişdir. K dəyəri həssas dəyərdir. Bu səbəbdən K-nın artırılması bəzi təsnifat işlərində performansa müsbət təsir göstərdiyi halda, digərlərində əks təsir göstərmişdir. Buna görə də k parametri dəqiq seçilməlidir.K-NN alqoritmi bir çox yerdə analitik olaraq izlənilə bilən və asan olması və kənar məlumatlardan təsirlənməməsi kimi səbəblərə görə üstünlük verilir.

1.6 Söz Çantası Texnikası

Mətn məlumatlarından öyrənərkən ən vacib problem hər bir sənədin müxtəlif uzunluqda olmasıdır. Belə hallarda hər bir sözə bir xüsusiyyət kimi yanaşmaq

lazımdır. Lakin bu vəziyyəti geniş miqyaslı məlumatlarda tətbiq etmək çox çətindir. Bu vəziyyətdən çıxmaq üçün ən vacib üsullardan biri Söz Çantası Texnikasıdır (BOW). Bu üsulda hər bir söz unikal hesab olunur və hər bir unikal sözün 20-si varSənəddə bir sıra təkrarlar var. Beləliklə, maşın öyrənmə alqoritmlərinin modelləşdirilməsində istifadə olunacaq xüsusiyyətlər mətndən çıxarılır (Bulut, (ASYU2014), 2014) (Brownlee, (Ocak,2016)). Onun istifadəsi aşağıdakı nümunədəki kimidir (Soumya):



Şəkil 1-9 BOW metodunun nümunəsi

Bu nümunədə əvvəlcə sənəddəki unikal sözlərin siyahısı çıxarılır. Sonra bu sözlərin sənəddə rast gəlinmə sayı hesablanır. Ümumilikdə 10 unikal söz var. Bu o deməkdir ki, bütün vektorların ölçüsü 10 olacaq. Hər bir cümlə üçün vektor təsviri Şəkil 1.9-dakı kimidir. Verilmiş sözlər toplusunda təkrarları tapmaq və ehtimalları sözlərə hesablamaq üçün N qramdan istifadə olunur.

Təyinat üçün faydalıdır. N-qram simvolu ardıcıl n simvol və ya sözlərin ardıcıllığıdır (Cianflone, 2016). 8-ci tənlikdəki kimi hesablanır.

$$P(w_{1)}^{n} = \prod_{k=1}^{n} P(w_k | w_{k-N-1}^{k-1}) \qquad (8)$$

N bir olduqda uniqram, iki olduqda biqram, üç olduqda isə triqram adlanır. Cümlə üçün triqramın yaradılması aşağıda təsvir edilmişdir:

II FƏSİL

ƏDƏBİYYATIN XÜLASƏSİ

Spam/e-poçtun təsnifatı üzrə bir çox tədqiqatlar təqdim edilmişdir.Tədqiqatların əksəriyyətində tədqiqatçılar spam e-poçtları ayırd etmək üçün maşın alətlərindən istifadə etmişlər.Onlar təlim alqoritmlərinin hazırlanmasında istifadə olunacaq məzmuna əsaslanan yanaşmalara müraciət ediblər. Bu məzmun əsaslı yanaşmaları dörd əsas kateqoriyaya bölmək olar.Bunlar:başlıq xüsusiyyətləri, mövzu xüsusiyyətləri, e-poçtun əsas xüsusiyyətləri və e-poçt əlavəsi xüsusiyyətləridir (Sah, 2017). Tədqiqat nümunələrini ardıcıllıqla araşdırsaq, Sah və başqalarının (Sah, 2017) apardıqları araşdırmada məlumatların toplanması, verilənlərin əvvəlcədən işlənməsi, verilənlərdən xüsusiyyətlər çıxarılması, təsnifatlaşdırılması yolu ilə mətn əsaslı təsnifat araşdırması aparılmışdır.Müvafiq olaraq məlumatların və nəticələrin təhlili üçün tədqiqatda, Support Vector Machine (SVM) və Naive Bayes alqoritmləri ilə təsnif etmək üçün təlim üçün 702 e-poçt və sınaq üçün 260 e-poçt istifadə edilmişdir.Tədqiqatda həyata keçirilən xüsusiyyət çıxarma mərhələsi bu dissertasiya işində xüsusiyyət çıxarılması üçün həyata keçirilən prosesə bənzəyir. Araşdırma nəticəsində, Naive Bayes və SVM alqoritmlərin oxşar müvəffəqiyyət göstərdiyi qənaətinə gəlindi. Digər oxşar tədqiqat Hadoop mühitində aparılan təsnifat işində Renuka və başqaları (Renuka, 2017) tərəfindən aparılmışdır.

Spam e-poçtları ayırd etmək üçün Gradient Boost və Naive Bayes təsnifat üsullarından istifadə edilmişdir. Tədqiqat əsasən iki mərhələdə aparıldı: təlim və sınaq. Tədqiqatda təsnifat performansını yaxşılaşdırmaq üçün tək qovşaqlı Hadoop mühitindən istifadə edərək müxtəlif təsnifat üsullarını birlikdə istifadə edən hibrid model istifadə edilmişdir. Tədqiqatın performansı dəqiqlik, dəqiqlik və geri çağırma metrikləri ilə ölçüldü.

Spam/e-poçt ayrı-seçkiliyində funksiyaların çıxarılması təsnifat uğuruna birbaşa təsir edən bir prosesdir. İstənməyən e-poçtlarda adətən təsnif etməyə çalışdığımız obyektlər mətnlərdir. Spam e-poçtların mətn hissələrindən funksiyaların çıxarılması

tez-tez dəyişir, çünki spam göndərənlər öz texnikalarını daim dəyişirlər.Digər tərəfdən, funksiyaların çıxarılması mərhələsində bir növ etiketli məlumat yaradılır. Etiketli məlumatlardan istifadə etməklə təlim performansı qorumağa kömək edə bilsə də, daim etiketlənmiş məlumatların hazırlanması olduqca bahalıdır. Kumagai və digərləri (Kumagai, 2017) tərəfindən aparılan tədqiqatda təsnifat daha az xərc gətirəcək öyrənmə prosesi üçün etiketlənməmiş verilənlərlə yanaşı, etiketli məlumatlardan istifadə etməklə aparılmışdır. Tədqiqatın əsas məqsədi təlim mərhələsində heç vaxt görülməmiş bir xüsusiyyətin performansına mənfi təsirləri aradan qaldırmaqdır. Digər mühüm qazanc, təlim və sınaq mərhələləri arasında paylanmanın dəyişməsi nəticəsində yaranan performansın pisləşməsini yaxşılaşdırmaqdır. Təklif olunan metod hər iki problemin öhdəsindən gələn ən müasir təsnifatçıları öyrənir. Təklif olunan metodla mövcud xüsusiyyətləri nəzərə alaraq yeni xüsusiyyətlərin şərti paylanması etiketsiz verilənlərdən istifadə etməklə öyrənilir. Təcrübələrdə təklif olunan metodun mövcud metodlardan daha yaxşı performans göstərdiyi, yeni xüsusiyyətlərə malik olduğu, təlim və sınaq paylamalarının fərqli olduğu izah edilmişdir.Xüsusiyyətlərin çıxarılmasına maşın öyrənmə üsulları nöqteyi-nəzərindən baxsaq, arifmetik obyektləri (rəqəmlər və ya vektorlar) mətnlərdən daha yaxşı təsnif etmək olar. Ajaz və digərləri (Ajaz, 2017) tərəfindən aparılan spam/e-poçt təsnifatı tədqiqatında bütün mesajlar əvvəlcə ədədlərdən ibarət xüsusiyyət vektorlarına çevrilmişdir. Tədqiqatda xüsusiyyət vektorlarını çıxararkən məlumatın tez-tez itirildiyi və xüsusiyyət çıxarıcının təyin edilməsinin təsnifat uğuru üçün vacib olduğu vurğulanmışdır. Bundan əlavə, xüsusiyyəti seçərkən mesaj mətnindən bütün xüsusiyyətlərin alınmaması lazım olduğu və xüsusiyyət çıxarma prosesinə öz məlumatlarımızın əlavə oluna biləcəyi ifadə edildi.

Xüsusilə e-poçtların mətn hissələrinə diqqət yetirən təsnifat işlərinə əlavə olaraq, mövzu məlumatlarına əsaslanaraq təsnif edən araşdırmalar da var. Li və başqaları (Lee, 2017) tərəfindən aparılan araşdırmada çəkili Naive Bayes təsnifat metodu əsasında spam filtri hazırlanmışdır. Hazırlanmış filtr yalnız arzuolunmaz e-poçtların mövzusunu yoxlayır. İş zamanı tez-tez istifadə olunan sözləri silmək üçün söz çantası

(BOW) texnikasına əlavə olaraq, e-poçt mövzularında yeni üsullardan da istifadə olunur.Xüsusiyyətləri aşkar etmək üçün təbii dil emal üsullarından da istifadə edilmişdir. Tədqiqatda digər tədqiqatlarda tez-tez istifadə edilən spam/e-poçt məlumat dəstlərindən (TREC, EnronSpam) istifadə edilmişdir. (Lee, 2017) Qeyd olunan metod 94,85% dəqiqliklə təsnif edilmişdir. Tədqiqat eyni tipli digər tədqiqatlarla müqayisədə 2,43% daha yaxşı uğur qazanıb. Spam e-poçt təsnifatı tədqiqatlarında SVM və Naive Bayes alqoritmlərinin məşhur olduğunu görürük. Carreras və Markes (Carreras, 2001) tərəfindən aparılan araşdırmada bunlardan fərqli olaraq AdaBoost alqoritmi ilə spam e-poçtların filtrasiyası həyata keçirilmişdir. Tədqiqatda açıq məlumat dəstlərindən biri olan PU1 (Carreras, 2001) istifadə edilmişdir. Tədqiqat nəticəsində, Boosting əsaslı metodların PU1 məlumat dəstləri üçün Naive Bayes və Qərar Ağaclarını istifadə edən metodlardan üstün müvəffəqiyyət göstərdiyi görüldü. Digər tərəfdən, AdaBoost təsnifatının yüksək dəqiqlikli təsnifatçıların tələb olunduğu ssenarilərdə yaxşı işlədiyi sübut edilmişdir. Mürəkkəbliyin artırılması daha yüksək dəqiqliklə təsnifatlaşdırıcıları əldə etməyə imkan verir. Mürəkkəblik maya dəyərini artırsa belə, bu xərc səhv təsnifatın qiymətindən aşağı olacaq. Təsnifat tədqiqatlarında üstünlük verilən başqa bir alqoritm Süni Sinir Şəbəkələridir (ANN). Chuan və digərləri (Chuan, 2005) tərəfindən aparılan araşdırmada spam e-poçtlar açıq mənbəli e-poçt verilənlər bazasından istifadə edərək ANN-LVQ və ANN-BP yanaşması ilə məzmunlarına görə təsnif edilmişdir.

Tədqiqatda vurğulanıb ki, neyroşəbəkə təlimlərinin sayı təsnifat uğuruna birbaşa təsir edir və bu rəqəm 1500-ə çatdıqda performans yaxşılaşır. Tədqiqatda, süni neyron şəbəkəsinə əsaslanan alqoritmlərin spam/e-poçtların hər bir xüsusiyyətinin bütövlüklə əlaqəsini nəzərə aldıqları üçün hər bir xüsusiyyəti bir-birindən müstəqil hesab edən Bayes alqoritmlərindən daha yaxşı olduğu irəli sürülüb. Eksperimental nəticələrə nəzər saldıqda ANN BP alqoritminin 98,42%, ANN-LVQ alqoritminin 98,97%, Naive Bayes alqoritminin isə 97,63% müvəffəqiyyət əldə etdiyi görüldü. Sharma və Sahni [35] tərəfindən aparılan araşdırmada spam e-poçt məlumatları Weka mühitində

ID3, J48, Simple CART, Alternating Decision Tree (ADT) alqoritmlərindən istifadə etməklə təsnif edilmişdir. Tədqiqatda UCI Machine Learning repozitoriyasından məlumat dəstlərindən istifadə edilmişdir. Tədqiqat nəticəsində uğur göstəricisi ID3 alqoritmi üçün 89%, J48 alqoritmi üçün 92%, ADT alqoritmi üçün 90,91%, SimpleCART alqoritmi üçün 92,63% təşkil edib. J48 klassifikatorunun təsnifat dəqiqliyi baxımından ID3, CART və ADTree-dən üstün olduğu göstərilmişdir.Ədəbiyyatdakı araşdırmalara və bu araşdırmalardakı məlumatlara görə, ümumiyyətlə spam/e-poçt təsnifatı araşdırmaları ilə bağlı aşağıdakı nəticələr çıxarılır;

- Zərərli spam araşdırmalarının əksəriyyətində tədqiqatçılar spamı ayırd etmək üçün maşın öyrənmə alqoritmlərini öyrətmək üçün məzmuna əsaslanan yanaşmalara müraciət ediblər. Bu yanaşmaları dörd əsas kateqoriyaya bölmək olar. Bunlar;başlıq xüsusiyyətləri, mövzu xüsusiyyətləri, e-poçtun əsas xüsusiyyətləri və e-poçt əlavəsi xüsusiyyətləridir (Sah, 2017).
- Tədqiqatlarda ən çox istifadə olunan maşın öyrənmə üsulları; Bunlar SVM, ANN, RF, Naive Bayes və AdaBoost idi (Sah, 2017).
- Bəzi tədqiqatlarda Kobud dəstlər kimi qayda-əsaslı yanaşmaları, eləcə də məlum maşın öyrənmə üsullarını ehtiva edən qarışıq yanaşmalar nümayiş etdirilmişdir (Sah, 2017).
- Tədqiqatlarda ümumiyyətlə istifadə edilən e-poçt məlumatları Spambase (1999), Spam Assassin (2006), TREC (2007)-dir (Sah, 2017).

III FƏSİL

MATERİALLAR VƏ METODLAR

Bu bölmədə spam/e-mail təsnifat metodu ətraflı müzakirə olunur. Bu fəsildə müvafiq olaraq məlumatların işlənməsi, alət seçimi və təcrübələrin yerinə yetirilməsi izah ediləcəkdir.

3.1 Məlumatların əvvəlcədən emalı

Araşdırması çərçivəsində istifadə edilən bütün məlumatlar istənməyən e-poçtlardakı keçid məlumatlarını, bu bağlantılardakı mətn məlumatlarını və unikal e-poçt nömrəsini ehtiva edir.

Xam məlumatlar Cədvəl 3.1-də formatda təqdim olunur.

Cədvəl 3-1 İşlənməmiş məlumat formatı

FORMAT
E-poçt
Unikal E-poçt Nömrəsini göndərin ||| E-poçt Linki (URL) ||| Link Mətni
Misal:
00000460-6b36-41c3-aeaf
35bb32c02766|||http://www.bigbv.top/5D899TU358EM391XL1721W2346IN18T678649B3249141020.php|||Bura klikləyin
İstənməyən E-poçt
Unikal E-poçt Nömrəsi ||| Spam Link(URL) ||| Link Mətni
Misal:
0002d83c-90c4-48ab-80ea
594c411463db|||http://www.totalspas.top/911/362/395/1734/2360.19tt1319463AAF3.php|||Start 12.000 Tökmə Planı Kolleksiyası ilə Heyrətamiz Tövbələr Tikintisi Daha Asan Yol!

Xam məlumat dəsti aşağıdakı xüsusiyyətlərə malikdir;

- E-poçt nömrəsi, keçid və keçid mətnləri "|||" xarakterlərinə görə bir-birindən seçilirlər.
- Elektron poçt nömrələri hərf və rəqəmlərdən ibarət unikal nömrələrdir.
- Elektron məktubda birdən çox keçid ola bilər. Bunu unikal e-poçt nömrələri ilə fərqləndirmək olar.
- Linklər xüsusiyyət çıxarmaq üçün uyğun deyil. Hər bir keçid müxtəlif uzunluqda və fərqli formatdadır.
- Anker mətnləri dildən müstəqildir. Bundan əlavə, ingilis və türkcə mətnlərin çoxluq təşkil etdiyi müşahidə edilib.

- Bağlantı mətnlərində durğu işarələrindən tez-tez istifadə olunduğu müşahidə edilmişdir. Bu, maşın öyrənmə texnikalarının performansını artırmaq üçün arzuolunmaz bir vəziyyətdir.

Xam məlumatların xüsusiyyətlərinə görə aşağıdakı tələblər alınır.

- Hər bir məlumat sırası unikal e-poçt nömrəsinə, keçidə və lövbər mətninə bölünməlidir.
- Link mətnlərindəki rəqəmlər və naməlum simvollar silinməlidir. Silinmə prosesi zamanı simvolların silindiyi yerdə boşluq əlavə olunacaq. Bu boşluqlar sözləri ayırmağa kömək edəcək. Bununla belə, apostrof üçün xüsusi bir hal var. Apostrof çıxarılarkən apostrofun yerinə boşluq qoyulmayacaq ki, sözün şəkilçisi və kökü iki ayrı söz kimi qəbul edilməsin.
- İki simvollu sözlərin heç bir mənası olmadığı üçün silinməlidir.
- Xüsusi məna daşıyan bəzi sözlərin fərqləndirici xüsusiyyəti olmadığı üçün onları çıxarmaq lazımdır. Bunlar; Türk və İngilis dillərində həftənin günləri, türk və ingilis dillərində günlərin, ilin aylarının abbreviaturaları, ayların abbreviaturaları, valyuta abbreviaturaları, “WWW, http, HTTPS, COM, AND, WITH, THE və s.” kimi sözlər. Bu sözlər verilənlər toplusunu araşdıraraq qərara alınıb.
- Türk və ingilis əlifbaları arasındakı fərqə görə yaranan səhvləri minimuma endirmək üçün bütün hərflər böyük hərflərə çevriləcək. Türk dilində nöqtəli hərflər nöqtəsiz hərflərə çevriləcək. Bu, simvolların eyni söz olması və birində nöqtə ilə, digərində isə nöqtəsiz yazılması səbəbindən sözlərin fərqli qəbul edilməsinin qarşısını alacaq.Məlumatların ilkin emalı mərhələsini həyata keçirmək üçün 16 GB RAM və prosessorlu kompüter Windows 10 əməliyyat sistemi mühitində 2,60 GHz sürətindən istifadə edilmişdir. Yuxarıda sadalanan tələbləri həyata keçirmək üçün proqram Visual Studio 2015 mühitində C# proqramlaşdırma dilindən istifadə etməklə hazırlanmışdır.

Tətbiqin psevdokodları Cədvəl 3.2-dəki kimidir;

Məlumatların əvvəlcədən emalı alqoritmi
Daxiletmə: D:Spam/e-poçt verilənlər toplusu (“e-poçt nömrəsi|||link|||link mətni” formatında) **Çıxış O: İşlənmiş məlumat dəsti** **foreach (D-də məlumat)** **{** **“|||” işarə ilə ayırmaq;** **dataline=dataline- e-poçt nömrəsi;** **link mətn=dataline- keçid;** **Əgər(link mətni!=null)** **{** **Link mətnini böyük hərflərə çevirmək;** **Türk simvollarını nöqtəsiz simvollara çevirin;** **Xüsusi simvolları çıxarın, onları boşluqlarla əvəz edin;** **Əgər (Xüsusi simvol “ ‘ ” olarsa)** **{** **Apostrofu çıxarın, boşluq qoymayın;** **}** **Wordlist=link mətnini sözlərə bölmək;** **}** **Foreach (söz siyahısında söz)** **{** **Əgər (söz uzunluğu>2)** **{** **Əgər(söz!=xüsusi söz)** **tək söz siyahısı əlavə edin** **}** **}** **Foreach (bir söz siyahısında bir söz)** **{** **Təmizlənmiş lövbər mətni+=bir söz** **}** **Təmizlənmiş lövbər mətni siyahısı=təmizlənmiş lövbər mətni** **}** **O=təmizlənmiş lövbər mətn siyahısı**

Məlumatların əvvəlcədən işlənməsi spam və e-poçtlar üçün ayrı-ayrı fayllarda həyata keçirilib.E-poçtda birdən çox keçid ola bilər. Eyni e-poçtdakı keçidlər və lövbər mətnləri eyni ola bilər. Fərqli nümunələri görmək üçün eyni keçid və keçid mətni

olan elektron məktublardan yalnız biri işlənmişdir.Bu fərqi etmək üçün Cədvəl 3.3-də psevdokodları verilmiş alqoritm kodlaşdırılmışdır.

Cədvəl 3-3 Məlumatların tanınması alqoritminin psevdokodlarının təkrarlanması

Dublikat verilənləri ayırd etmək üçün alqoritm

Daxiletmə: D:Emal edilmiş e-poçt/e-poçt verilənlər toplusu ("e-poçt nömrəsi|||link|||link mətni" formatında)
Çıxış O: Məlumat dəsti dublikat qeydlərdən təmizləndi
M: Dublikat siyahı=boş; //e-poçt nömrəsini və keçid mətnini saxlayır.
Temperatur=doğru;
Obyekt=null; //e-poçt nömrəsi və keçid mətni xüsusiyyətlərinə malikdir.
foreach (D-də verilənlər bazası)
{
"|||" nişanla ayırmaq;
e-poçt nömrəsi=ilk "|||" ilə ayrılmış məlumatlar.
data line=data line-email number;
link mətni=server bağlantısı;
N.emailnumber=e-poçt nömrəsi;
N.linktext=
keçid mətni
If(temp)
{
M[0]+=N;
N=null;
temp=false;
continue;
}
If(bağlantı metni!=null)
{
If(M contains N==false)
M.add(N);
N=null;
}
}
O=M;

Bu proseslər nəticəsində spam e-poçtlar və e-poçt keçidlərindəki keçid mətnləri fərqləndirilib və əvvəlcədən işlənib. Əvvəlcədən emal mərhələsi sayəsində mətnləri əlaqələndirin.Onların hamısı böyük hərflərə çevrilib, əcnəbi simvol problemi həll olunub, durğu işarələri çıxarılıb, bəzi xüsusi mənalı sözlər və 3 hərfdən qısa sözlər fərqləndirilib. İlkin emal nəticəsində tələb olunmayan e-poçtlara və ".txt" uzantılı e-

poçtlara aid iki məlumat faylı yaradılıb. İlkin emaldan əvvəl və sonrakı məlumatların nömrələri Cədvəl 3.4-dəki kimidir;

Cədvəl 3-4 Məlumatların əvvəlcədən işlənməsi mərhələsindən sonra məlumatların hesablanması vəziyyəti

	emaldan əvvəl sətirlərin sayı	emaldan sonra verilənlər sətirlərinin sayı	emalla silinmiş məlumat sətirlərinin sayı (fərq)
E-poçt məlumat sətirlərinin sayı	141414	107163	34251
Spam məlumat sətirlərinin sayı	150000	132254	17746

E-poçtda birdən çox keçid ola bilər. İlkin emaldan əvvəl və sonrakı e-poçtların sayı Cədvəl 3.5-dəki kimidir;

Cədvəl 3-5 Məlumatların əvvəlcədən işlənməsindən sonra e-poçtların sayı

	Emaldan əvvəl e-poçtların sayı	Emaldan sonra e-poçtların sayı
E-post	8506	8500
İstənməyən E-poçt	47382	47213
Ümumi	55888	55713

3.2 N Gram Xüsusiyyətlərinin çıxarılması

Maşın öyrənmə üsullarında əvvəlcədən işlənmiş məlumatlardan istifadə etmək üçün onlar təlim və sınaq proseslərində istifadə ediləcək məlumat dəstinə çevrilməlidir. Bu məqsədlə Word Cluster Texnikasından istifadə edilməklə verilənlərin 1 qram, 2 qram, 3 qram, 4 qram və 5 qram kimi xüsusiyyətləri çıxarılmışdır.Xüsusiyyətlərin çıxarılması üçün proqram Windows 10 əməliyyat sistemi, 16 GB RAM və 2,60 GHz prosessor sürətinə malik kompüterdə Visual Studio 2015 inkişaf aləti və C# proqramlaşdırma dili ilə hazırlanmışdır. N Gram alqoritminin psevdo kodları Cədvəl 3.6-dakı kimidir,

Cədvəl 3-6 N Qram Alqoritminin psevdokodları

N Qram alqoritmi

```
Giriş: D: Əvvəlcədən işlənmiş məlumat dəsti (bir və ya daha çox sözdən ibarət cümlə formatında)
  N:N qram ölçüsü
Çıxış O: N qram
foreach (D-də məlumat)
{
  Tək söz siyahısı=məlumat xəttini sözlərə bölmək;
}
for (i=0; i<=tək sözlü siyahı elementlərinin sayı -N;i++)
 { switch(N)
 {
 case 1:
 NGram+= tək sözlü siyahı [i];
 break;
 case 2:
 NGram+= tək sözlü siyahı [i]+" "+tək sözlü siyahı [i+1];
 break;
 case 3:
 NGram+=tək sözlü siyahı [i]+" "+ tək sözlü siyahı [i+1] +" "+ tək sözlü siyahı [i+2];
 break;
 case 4:
 NGram+=tək sözlü siyahı [i] +" "+ tək sözlü siyahı [i+1] +" "+tək sözlü siyahı [i+2]+" "+
tək sözlü siyahı [i+3];
 case 5:
 NGram+=tək sözlü siyahı [i] +" "+ tək sözlü siyahı [i+1] +" "+ tək sözlü siyahı [i+2]+" "+
tək sözlü siyahı [i+3] +" "+ tək sözlü siyahı [i+4];
 break;
 }
add NGram Listesi (Ngram)
 }
O=NGram Listesi
```

Xüsusiyyətlərin çıxarılması spam və e-poçtlar üçün ayrıca həyata keçirilirdi. Ayrı-ayrılıqda yaradılan bu xüsusiyyət dəstləri birləşdirildi və ümumi qramlardan tək biri xüsusiyyət kimi qəbul edildi. Yaradılan hər xüsusiyyət, başqa sözlə, hər qram maşın öyrənməsidir. Texnikalarda istifadə ediləcək matris strukturunda məlumat dəstlərinin sütunlarını təşkil edirlər. Qramlar yaradıldıqdan sonra əldə edilən nəticələr, xüsusiyyətlərin sayının kifayət qədər çox olduğu görüldü. İşlənmiş məlumat dəstində hər qramın görünmə sayı hesablanmışdır. Nəticələrə əsasən, xüsusiyyətlərin sayı müəyyən limitlərə uyğun olaraq azaldılıb. Xüsusiyyətlərin sayının azaldılması: 30 təkrar, 40 təkrar və 50 təkrar limit kimi müəyyən edilib və bu ədədlərdən daha az

təkrarlanan qramlar xüsusiyyət dəstinə daxil edilməyib. Xüsusiyyətlər dəstinin bu şəkildə dizayn edilməsi, həmçinin funksiyaların sayının maşın öyrənmə texnikalarının uğuruna təsirini göstərəcəkdir. Limitlər əlavə edilmədən və 30, 40, 50 təkrar limitlərinə uyğun olaraq yaradılan funksiyaların sayı Cədvəl 3.7-dəki kimidir. burada arzuolunmaz elektron məktubların xüsusiyyətlərinin sayı və e-poçtların xüsusiyyətlərinin sayı verilmişdir. Toplam verilən xüsusiyyətlərin sayı məlumat dəstlərində istifadə ediləcək funksiyaların sayını göstərir. Ümumi sayı hesablanarkən, spam xüsusiyyətləri və e-poçt xüsusiyyətləri birləşdirildi və hər iki dəstdəki ümumi xüsusiyyətlər bir dəfə hesablandı.

Cədvəl 3-7 Limitlərə uyğun xüsusiyyət nömrələri

		1 Qram	2 Qram	3 Qram	4 Qram	5Qram
LİMİT YOX	İst. EP	19450	82115	127040	171669	214554
	EP	35741	114861	147189	168251	183528
	Toplam	48580	193049	272615	339069	397587
LİMİT=30	İst. EP	1940	3041	3044	2714	2332
	EP	1869	1377	1270	1198	1161
	Toplam	3335	4332	4286	3902	3487
LİMİT=40	İst. EP	1637	2352	2278	1995	1690
	EP	1362	959	868	836	838
	Toplam	2658	3239	3131	2825	2525
LİMİT=50	İst. EP	1373	1844	1756	1521	1280
	EP	1083	757	701	675	663
	Toplam	2188	2550	2448	2192	1941

3.3 Məlumat dəstlərinin yaradılması

Maşın öyrənmə üsullarını tətbiq edərkən istifadə ediləcək məlumat dəstləri matris formatında hazırlanır. Əvvəlki addımda yaradılmış 1, 2, 3, 4, 5 qramlıq xüsusiyyətlər məlumat dəstlərinin sütunlarını təşkil edir. Məlumat dəstlərinin sətirləri xam verilənlərdəki keçid mətnlərindən ibarətdir. Sətirlər yaradarkən;

- Xam formatda olan məlumatlar idarə olunur. E-poçt nömrəsi, keçid və keçid mətni təhlil edilir.
- Çapa mətni əvvəlcədən işlənir.
- İstifadə olunan xüsusiyyət vektoruna uyğun olaraq keçid mətninin qramları çıxarılır.

- Link mətninin qramları xüsusiyyət vektoru ilə müqayisə edilir. “1” üst-üstə düşən xüsusiyyətlər üçün, “0” isə üst-üstə düşməyən xüsusiyyətlər üçün yazılır.
- Link mətnində eyni qramlar varsa, rəqəm tapılan qramların sayı kimi hesablanır. Verilənlər toplusunda "1" əvəzinə ümumi rəqəm yazılır.
- Əgər anker mətninin qramları xüsusiyyət vektorunda heç bir xüsusiyyətlə üst-üstə düşmürsə, bu çapa mətni verilənlər bazasına daxil edilmir.
- Çapa mətnindən yaradılmış hər bir xüsusiyyəti xüsusiyyət vektoru ilə müqayisə edərkən etiket məlumatı əlavə edilməlidir.

Cədvəl 3-8 Verilənlər toplusunun yaradılması alqoritmi psevdokodları

Məlumat toplusunun yaradılması alqoritmi

```
Daxiletmə: F[n]=x qram xüsusiyyət vektoru, n: funksiyaların sayı
  D: Əvvəlcədən işlənmiş məlumat dəsti (yalnız anker mətnləri ehtiva edir)
  N:N qram ölçüsü
Nəticə: X[n]= Məlumat dəsti vektoru, n: xüsusiyyətlərin sayı
SET X[n] “Sıfır”
foreach (D-də məlumat)
{
  Ngram Siyahısı=ZƏNG ET NGram yarat(DataLine,N)
  Nqram sayı=NgramSiyahısının uzunluğu

  for(i=0 "xüsusiyyətlərin sayı" addım 1)
  {
  for(a=0 - "Nqramların sayı" addım 1)
  {
  Əgər(NgramList[a]=F[i])
  X[i]+=1
  }
  }
  X[n] fayla yazın
  SET X[n] “Sıfır”
}
```

Məlumat dəsti matrisinin sütunları üçün əvvəllər təsvir edilmiş limit dəyərləri (30, 40, 50) nəzərə alınmaqla müxtəlif məlumat dəstləri yaradılmışdır. Maşın öyrənmə üsullarını tətbiq etmək üçün istifadə ediləcək alətin tutumuna görə maksimum 50.000

sətir məlumat emal edilə bilər. Bu səbəbdən maksimum 50000 sətirdən ibarət bütün məlumat dəstləri yuxarıda izah edilən alqoritm məntiqi ilə hazırlanmışdır. Ümumilikdə 15 müxtəlif məlumat toplusu hazırlanmışdır. Məlumat dəstlərinin ölçüləri aşağıdakılardır.

Cədvəl 3-9 Limitlərə uyğun olaraq verilənlər dəsti matrisinin ölçüləri

		1 Qram	2 Qram	3 Qram	4 Qram	5Qram
30 LİMİTİ	İst. EP	25000	25000	25000	25000	23677
	EP	25000	25000	19864	8818	5125
	Ümumi	50000	50000	44864	33818	28802
40 LİMİTİ	İst. EP	25000	25000	25000	25000	21592
	EP	25000	25000	18045	7613	4081
	Ümumi	50000	50000	43045	32613	25673
50 LİMİTİ	İst. EP	25000	25000	25	25000	3882
	EP	25000	25000	17299	7368	19350
	Ümumi	50000	50000	17324	32368	23232

	1 Qram	2 Qram	3 Qram	4 Qram	5Qram
30 LİMİTİ	50000 X 3335	50000 X 4332	44864 X 4286	33818 X 3902	28802 X 3487
40 LİMİTİ	50000 X 2658	50000 X 3239	43045 X 3131	32613 X 2825	25673 X 2525
50 LİMİTİ	50000 X 2188	50000 X 2550	17324 X 2448	32368 X 2192	23232 X 1941

3.4 Nəqliyyat vasitəsinin seçimi

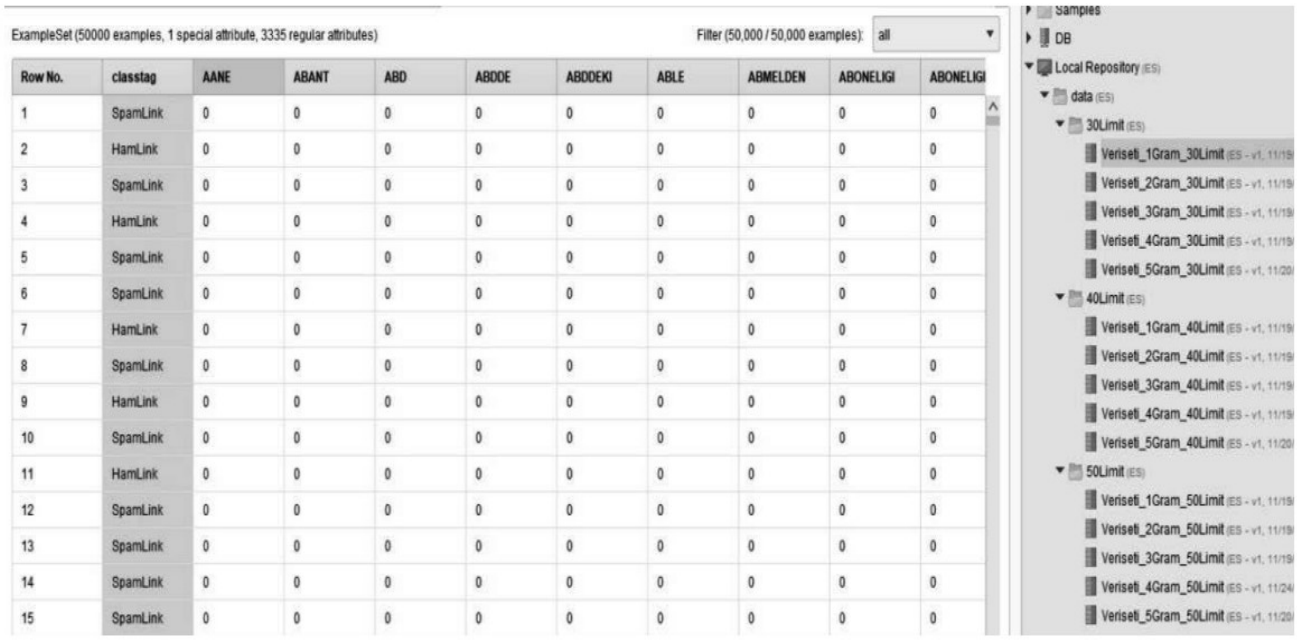
Məlumat dəstləri yaradıldıqdan sonra maşın öyrənmə üsullarının tətbiqi üçün müvafiq alət seçimi tədqiqatı aparıldı. İlk növbədə, biz maşın öyrənmə tədqiqatlarında ən çox istifadə edilən və populyar vasitələrdən biri olan Weka (Weka-Out of Memory Hatası, (Eylül, 2016)) aləti ilə işlədik. Weka çoxlu müxtəlif maşın öyrənmə alqoritmlərini ehtiva edən bir vasitədir. Weka sizə alqoritmləri birbaşa verilənlər bazasına tətbiq etməyə və ya öz Java kodunuzdan çağırmağa imkan verir. Weka məlumatların ilkin emalı, təsnifatı, qruplaşdırılması, əlaqə qaydaları və vizuallaşdırılması üçün alətləri əhatə edir. O, həmçinin yeni maşın öyrənmə sxemlərinin hazırlanması üçün uyğundur. Bu baxımdan Weka alətinin xüsusiyyətləri bu tezis işinin ehtiyaclarını ödəmək üçün uyğun tapıldı. Bununla belə, verilənlər

toplusu matrislərinin böyük ölçüsünə görə, Weka (Weka-Out of Memory Hatası, (Eylül, 2016)) tərəfindən də bildirilmiş "yaddaşdan kənar istisna" ilə qarşılaşdı. Daha çox RAM tutumu olan kompüter mühitində və WEKA tərəfindən müəyyən edilmiş həllər ilə yaddaşın daşması xətasını həll etməyə cəhdlər edilmişdir. Bununla belə, bütün hallarda böyük ölçülü məlumat dəstləri üçün eyni xətaya rast gəlindiyi halda, kiçik ölçülü məlumat dəstlərində heç bir problem yox idi. Bu vəziyyətdə başqa bir vasitə seçmək üçün araşdırma aparıldı. RapidMiner aləti ilə işləməyə qərar verildi (RapidMiner Studio 7.6 Sürümü, (Ekim, 2016)). RapidMiner Studio məlumat alimləri üçün pulsuz iş axını dizayneridir. Vizual dizayn imkanları çox faydalı olduğundan, ideyalar tez yaradıla bilər.Bu, prototipləşdirməyə və modellərin etibarlılığını yoxlamağa imkan verir. İstənilən formatda saxlanılan məlumatların avtomobilə asanlıqla ötürülməsinə imkan verir. Məsələn, bu tədqiqatdakı məlumatlar txt uzantılı mətn fayllarında saxlanılırdı. RapidMiner aləti ilə bu məlumatlar avtomobilin məlumat anbarına asanlıqla ötürülür və tip, etiket və s. kimi məlumatlar müəyyən edilirdi. RapidMiner verilənlərdəki nümunələri asanlıqla aşkar edir. Məsələn, RapidMiner mühitinə köçürdüyümüz məlumatların spam/e-poçt paylamaları alət tərəfindən çıxarılıb və çatışmayan dəyər yoxlanışı aparılıb. RapidMiner,məlumatları qarışdırmaq, modellər yaratmaq və təsdiqləmək və performansa nəzarət etmək imkanı verir. Böyük məlumatlarla işləmək üçün Weka ilə müqayisədə üstün bir vasitə olduğu göstərilmişdir. O, həmçinin bütün Weka alqoritmlərini ehtiva edir. Bu kimi səbəblərdən bu tədqiqatda istifadə etmək üçün RapidMiner aləti seçilmişdir. İş zamanı avtomobildən istifadə ilə bağlı səhvlər və s. Belə problemlərə rast gəlinmədiyi müşahidə olunub. RapidMiner 7.6 versiyası işləyir.

3.5 Təcrübələrin layihələndirilməsi

Məlumat dəstləri hazırlandıqdan və alət seçiminə qərar verildikdən sonra son mərhələdə maşın öyrənmə təcrübələri tərtib edildi. Hər bir maşın öyrənmə metodu üçün 15 məlumat dəstinə təlim və sınaq prosesləri tətbiq edilmişdir. Təcrübələrə başlamazdan əvvəl bütün məlumat dəstləri RapidMiner mühitinə yükləndi. Bunu

edərkən, "sinif etiketi" kimi təyin olunan sütun qeyd olunur. Şəkil 3.1-də RapidMiner mühitinə yüklənmiş məlumat dəstləri və bu məlumat dəstlərinin görünüşü göstərilir.

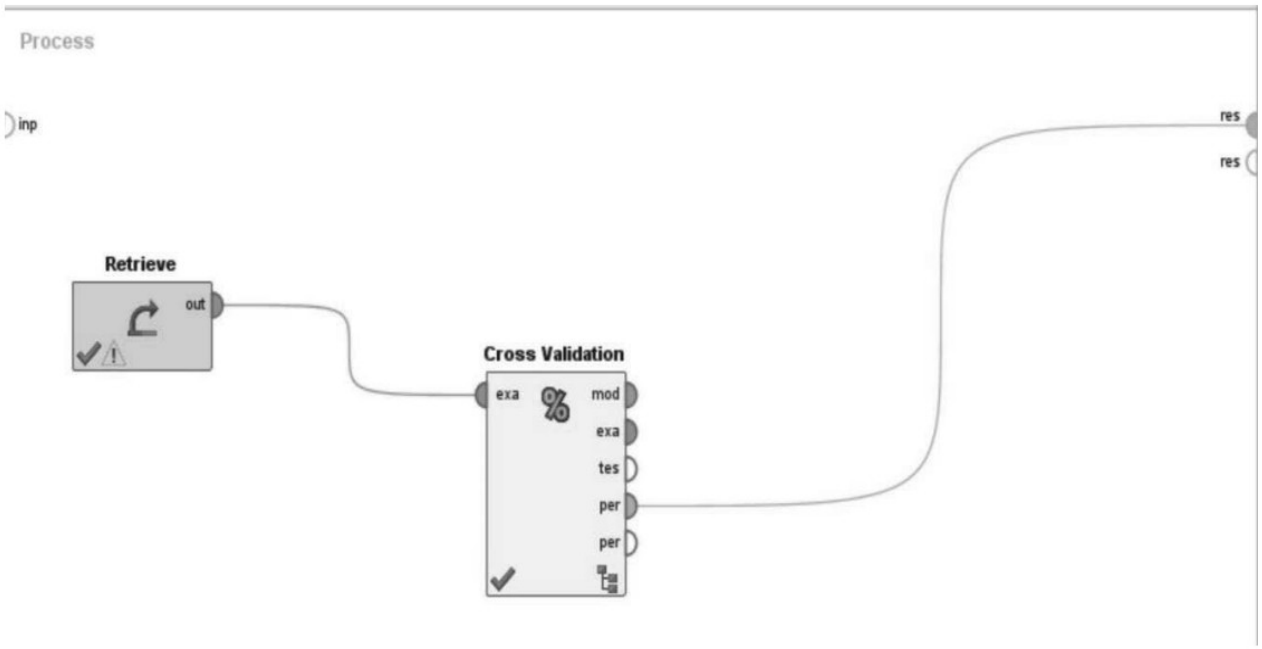


ExampleSet (50000 examples, 1 special attribute, 3335 regular attributes) Filter (50,000 / 50,000 examples): all

Row No.	classtag	AANE	ABANT	ABD	ABDDE	ABDDEKI	ABLE	ABMELDEN	ABONELIGI	ABONELIGI
1	SpamLink	0	0	0	0	0	0	0	0	0
2	HamLink	0	0	0	0	0	0	0	0	0
3	SpamLink	0	0	0	0	0	0	0	0	0
4	HamLink	0	0	0	0	0	0	0	0	0
5	SpamLink	0	0	0	0	0	0	0	0	0
6	SpamLink	0	0	0	0	0	0	0	0	0
7	HamLink	0	0	0	0	0	0	0	0	0
8	SpamLink	0	0	0	0	0	0	0	0	0
9	HamLink	0	0	0	0	0	0	0	0	0
10	SpamLink	0	0	0	0	0	0	0	0	0
11	HamLink	0	0	0	0	0	0	0	0	0
12	SpamLink	0	0	0	0	0	0	0	0	0
13	SpamLink	0	0	0	0	0	0	0	0	0
14	SpamLink	0	0	0	0	0	0	0	0	0
15	SpamLink	0	0	0	0	0	0	0	0	0

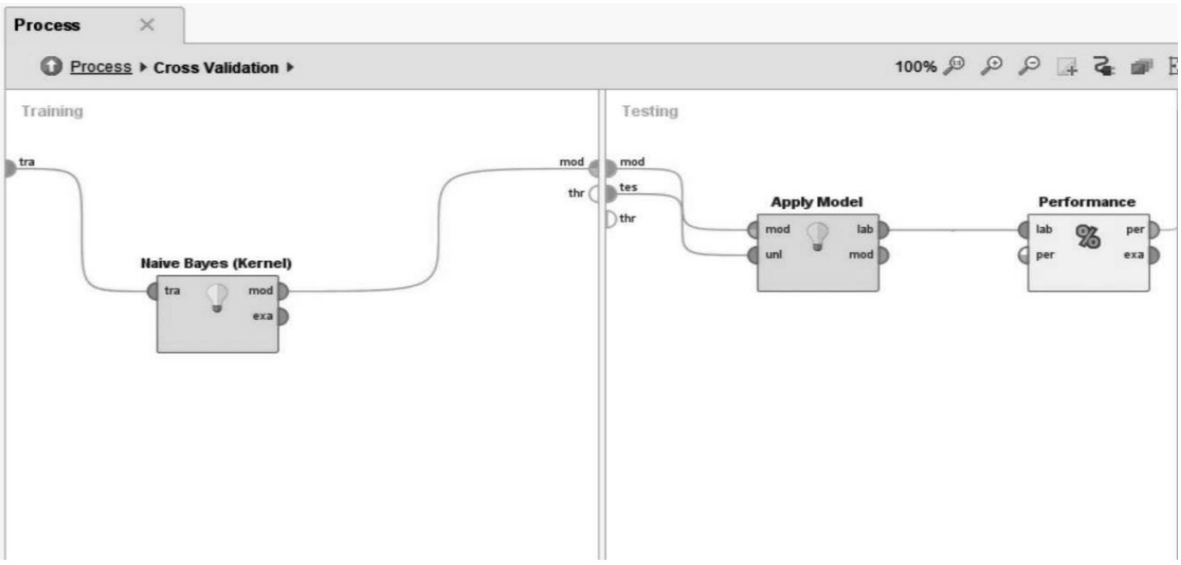
Şəkil 3-1 RapidMiner mühitində verilənlər toplusunun görünüşü

Məlumat dəstləri RapidMiner məlumat anbarına yükləndikdən sonra ikinci mərhələdə eksperimental dizayn həyata keçirildi. Bütün təcrübələr 10-qat çarpaz doğrulama ilə aparılmışdır. Məlumat RapidMiner mühitində Retrive bloku ilə eksperimental mühitə ötürülür. Bu blokun çıxışı digər bloka, Çarpaz Validasiya blokuna girişi təmin edir. Təlim və sınaq prosesləri Cross Validation blokunda həyata keçirilir. Cross Validation blokunun nəticəsi olaraq, performans nəticələri "per" abbreviaturası ilə verilir. Şəkil 3.2-də RapidMiner Cross Validation göstərilir.



Şəkil 3-2 RapidMiner Cross Validation

Bu blok çarpaz doğrulama blokunda təlim və sınaq proseslərini dizayn etmək üçün daxil edilir. Şəkil 3.3-də Naive Bayes Kernel metodu üçün nəzərdə tutulmuş model göstərilir. Model iki hissədən ibarətdir: "təlim" və "sınaq". Təlim bölməsində Naive Bayes Kernel bloku əlavə edilmişdir çünki Naive Bayes Kernel metodu ilə məşq etmək istənilir. Blokun girişi təlim məlumatları, çıxışı isə modeldir. Test bölməsində iki blok var: "model tətbiq et" və "performans". Model tətbiqi blokunda təlim nəticəsində əldə edilən model və test məlumatları daxil edilir. Performans bloku performansı müşahidə etmək üçün daxil edilir və bu modelin tətbiqinin nəticəsidir.İstədiyiniz parametrlərə görə model performansını göstərir.



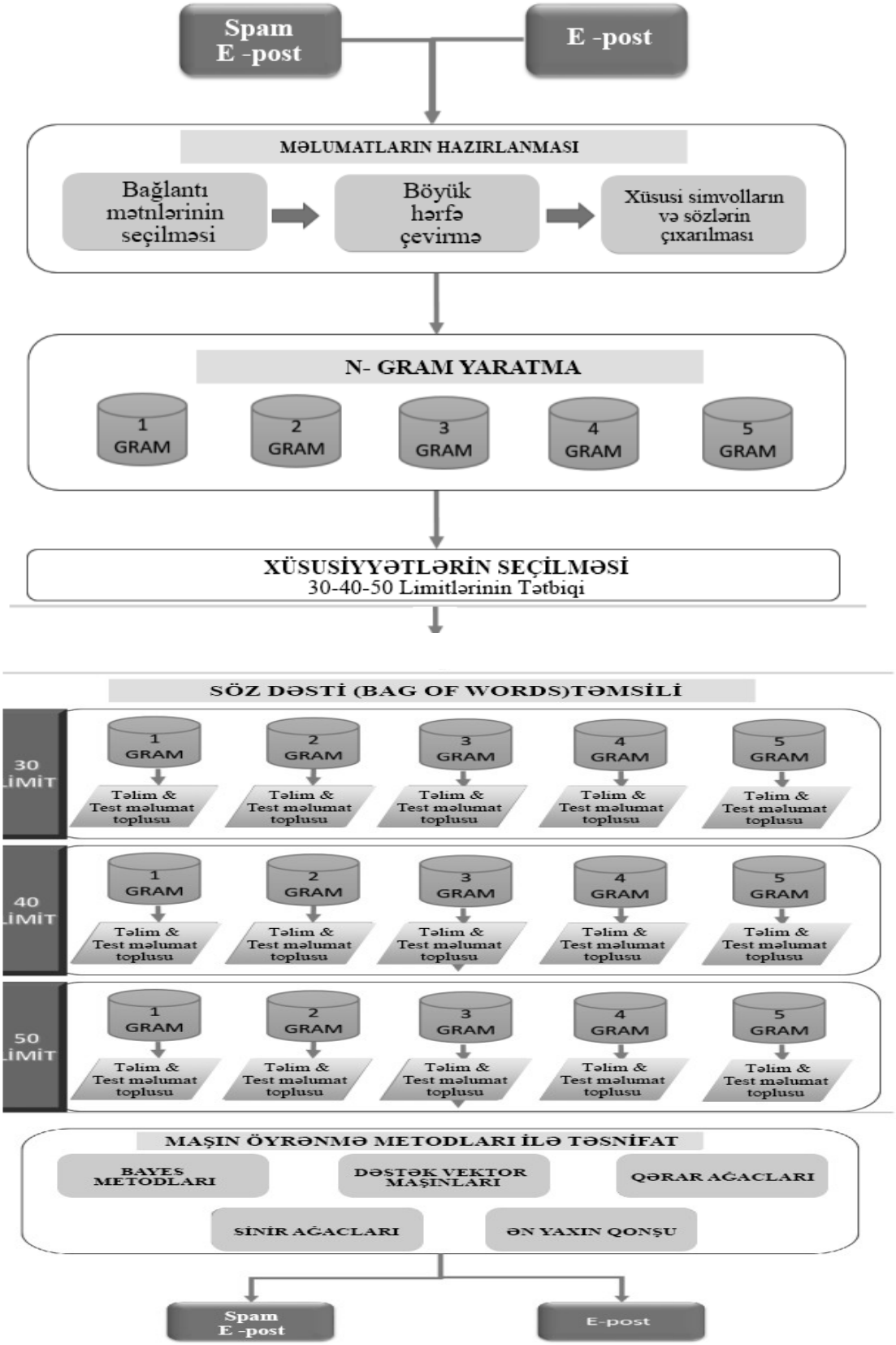
Şəkil 3-3 RapidMiner mühitində nümunə sınaq dizaynı

Şəkil 3.3 bütün təcrübələr üçün ümumidir. Təcrübələr hər dəfə Retrive blokuna 15 məlumat dəstini yenidən təyin etməklə həyata keçirilir. Eksperimental dizayn hər bir öyrənmə modeli üçün fərqlidir.

Tədqiqat çərçivəsində aşağıdakı maşın öyrənmə üsulları müzakirə olunur.

- Naive Bayes
- Naif Bayes (Kernel)
- Dəstək Vektor Maşını (Pegasos)
- Dəstək Vektor Maşını (Lib SVM)
- Dəstək vektor maşını (xətti)
- Perseptron
- KNN
- Qərar ağacı
- Gradient gücləndirilmiş ağac
- Qərar kökü
- Təsadüfi ağac
- Təsadüfi meşə

İndiyə qədərki addımlar Şəkil 3.4-də olduğu kimi ümumiləşdirilmişdir.



Şəkil 3-4 Spam e-poçtların təsnifatı üsulu

IV FƏSİL

EKSPERİMENTAL TƏDQİQATLAR VƏ NƏTİCƏLƏR

Bu bölmədə tədqiqatda nəzərdə tutulan prosesin eksperimental qiymətləndirmələri və tapıntıları yer alır. Eksperimental tədqiqatlar materiallar və metodlar bölməsində təsvir olunduğu kimi aparılmışdır. Eksperimental tədqiqatın nəticələrini və tapıntılarını başa düşmək üçün əvvəlcə performans göstəriciləri müzakirə olunacaq.

4.1 Performans Metrikləri

Təsnifat problemi üçün model yaradıldıqda və ya mövcud modellərdən istifadə edildikdə, həmin modelin uğuru edilən bütün proqnozlardan düzgün proqnozların sayı kimi qəbul edilir. Lakin bu məlumat yalnız təsnifatın dəqiqliyini təmin edir. Təkcə təsnifat dəqiqliyi çox vaxt modelin kifayət qədər yaxşı olub-olmadığını müəyyən etmək üçün kifayət qədər məlumat deyil.

Mürəkkəblik Matrisi: Təsnifatlayıcının proqnoz nəticələrini təqdim etməyin aydın yolu mürəkkəblik matrisindən istifadə etməkdir. Mürəkkəblik matrisi, məlum həqiqi dəyərləri olan bir sıra test məlumatları ilə təsnifat modelinin performansını təsvir etmək üçün tez-tez istifadə olunan bir cədvəldir. Onun 4 parametri var (Joshi, (Eylül, 2017)).

Cədvəl 4-1 Mürəkkəblik Matrisi

		TƏXMİN EDİLƏN SİNİF	
HƏQİQİ SİNİF		Spam E-post	E-Post
	Spam E-post	Doğru Pozitif(DP)	Yanlış Negatif(YP)
	E-Post	Yanlış Pozitif(YP)	Doğru Negatif(DN)

True Positive (DP) - Düzgün proqnozlaşdırılan müsbət dəyərlər. Bu, faktiki sinif və proqnozlaşdırılan sinfin dəyərinin eyni olduğunu göstərir. Burada DP dəyəri istənməyən e-poçtu spam kimi təsnif etdikdə tapılır.

True Negative (DN)- Bunlar düzgün proqnozlaşdırılan mənfi dəyərlərdir. Bu əsl sinifdir.Bu, dəyərin və proqnozlaşdırılan sinfin eyni olduğunu göstərir. Burada e-poçtu e-poçt kimi təsnif etdikdə DN dəyəri tapılır.

False Positive (FP) - Bu dəyər faktiki sinifiniz və proqnozlaşdırılan siniflə ziddiyyət təşkil etdikdə baş verir. Burada FP dəyəri e-poçtu spam kimi təsnif etdikdə tapılır.

False Negative (FN) - Bu dəyər faktiki sinifiniz və proqnozlaşdırılan siniflə ziddiyyət təşkil etdikdə baş verir. Burada istənməyən e-poçtu e-poçt kimi təsnif etdikdə FN dəyəri tapılır.

Burada həqiqi müsbət və həqiqi mənfi sahələrin artırılması istənilərkən, yalançı müsbət və yalançı mənfi sahələrin azaldılması təsnifat performansının yaxşı olduğunu göstərir.Mürəkkəblik matrisi sayəsində aşağıdakı ölçüləri hesablamaq olar.

Doğruluq: Doğruluq ən intuitiv performans ölçüsüdür və düzgün proqnozlaşdırılan müşahidələrin ümumi müşahidələrə nisbətidir. İstifadə olunan model yüksək dəqiqliyə malikdirsə, model ən yaxşı hesab edilə bilər. Bununla belə, yanlış müsbət və yanlış mənfi dəyərlərin sayı bir-birindən tamamilə fərqli və çox olduğu hallarda, modelin performansını qiymətləndirmək üçün digər parametrlər araşdırılmalıdır (Joshi, (Eylül, 2017); Brownlee, (Şubat, 2018)). Doğruluq(9) tənliyində olduğu kimi hesablanır.

Doğruluq =(9)

Precision -Həssaslıq düzgün proqnozlaşdırılan müsbət müşahidələrin ümumi proqnozlaşdırılan müsbət müşahidələrə nisbətidir. Buna müsbət proqnozlaşdırıcı dəyər də deyilir. Həssaslıq təsnifatçının dəqiqliyinin ölçüsü kimi düşünülə bilər. Aşağı həssaslıq çoxlu sayda yanlış pozitivləri də göstərə bilər (Joshi, (Eylül, 2017); Brownlee, (Şubat, 2018)). Həssaslıq (10) tənliyində olduğu kimi hesablanır.

Precision = (10)

Dəqiqlik - Düzgün proqnozlaşdırılan nəticələrin müsbətlərin ümumi sayına nisbəti. Təsnifatdakı bütün müşahidələr üçün düzgün proqnozlaşdırılan müsbət müşahidələrin nisbətidir. Dəqiqlik təsnifatçıların tamlığının ölçüsü kimi düşünülə bilər. Aşağı həssaslıq yalan neqativlərin yüksək olduğunu göstərir (Joshi, (Eylül, 2017); Brownlee, (Şubat, 2018)). Həssaslıq (11) tənliyindəki kimi hesablanır.

Recall = (11)

Fl hesabı- həssaslıq və həssaslığın harmonik ortasıdır. Buna görə də həm yalançı müsbətləri, həm də yanlış neqativləri nəzərə alır.Xüsusən qeyri-bərabər sinif paylanmasının olduğu hallarda F1 balına baxmaq dəqiqlik dəyərinə baxmaqdan daha faydalıdır. Oxşar sayda yalançı pozitivlər və yalan neqativlər baş verərsə, dəqiqlik dəyərinə baxmaq təsnifat uğuru üçün ən yaxşı nəticəni verir. Yanlış müsbət və yalançı neqativlərin sayları çox fərqlidirsə, həm dəqiqliyə, həm də həssaslığa baxmaq lazımdır (Joshi, (Eylül, 2017); Brownlee, (Şubat, 2018)). FI balı (12) tənlikdəki kimi hesablanır.

F1 hesabı = 2 (12)

4.2 Bayes alqoritmləri ilə təcrübələr

4.2.1 Sadə Bayes Təcrübəsinin Nəticələri

Ən yüksək dəqiqlik dərəcəsi 98,80% və F1 balı 98,04% Naive Bayes Alqoritmi ilə 1,2,3,4,5 qramlıq məlumat dəstləri ilə aparılan təcrübələrdə əldə edilmişdir. Cədvəl 5.3-də verilmiş dəqiqliyə və F1 Skoru nəticələrinə nəzər saldıqda aşağıdakı nəticələr əldə edilmişdir:

- Xüsusiyyət vektorunda sözlərin sayı, başqa sözlə, qramların sayı artdıqca, Naive Bayes alqoritminin dəqiqlik dərəcəsi müntəzəm olaraq artır. Bu, Naive Bayes alqoritminin xüsusilə mətn təsnifatında və uzun mətnlərin təsnifatında kifayət qədər uğurlu olduğunu göstərir.
- Xüsusiyyətlərin sayının performansa təsirini araşdırmaq istənildikdə, 30, 40, 50 limitləri ilə hazırlanmış eksperimental dəstlər araşdırıldı. Aparılan

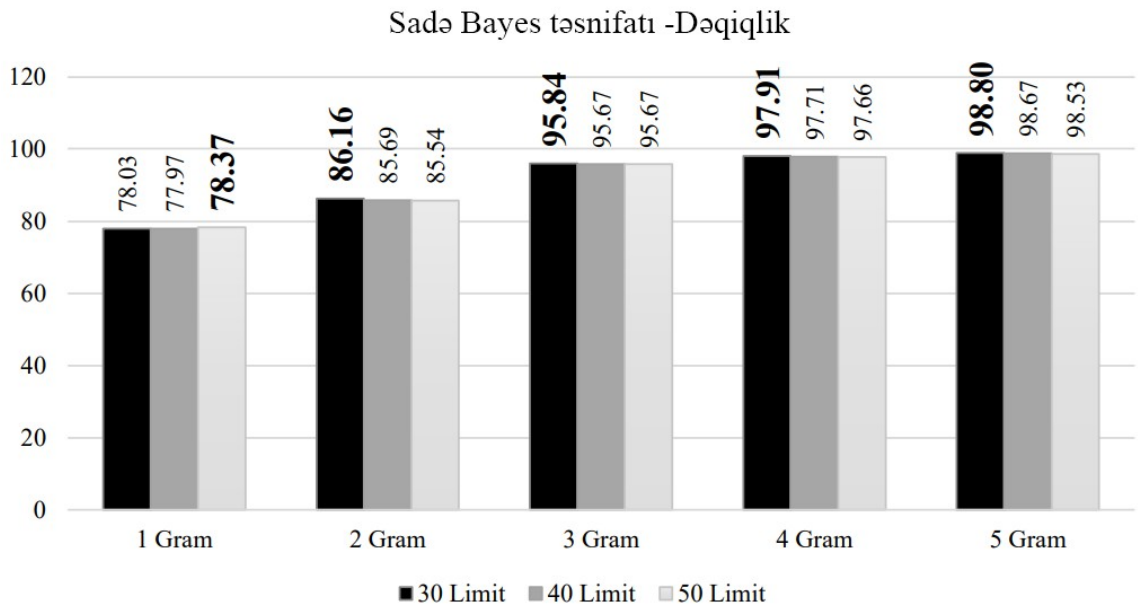
təcrübələrdə 1 qram üçün 50 limitlə aparılan bir təcrübə istisna olmaqla, bütün təcrübələrdə ən yüksək dəqiqliyin 30 limitlə əldə edildiyi müşahidə edilmişdir. Limitlərə görə xüsusiyyət nömrələrinin dəyişməsi üçün cədvəl 4.2-dəki məlumatlara baxdığımızda ən yüksək dəqiqliyin əldə edildiyi 30 həddi ilə digər hədlər arasında təxminən 1500 fərq olduğu görülür. Bu göstəricilər açıq şəkildə göstərir ki, xüsusiyyət vektorunun ölçüsünü artırmaq Naive Bayes alqoritminə mənfi təsir göstərmir, əksinə, performansa daha yaxşı təsir göstərir.

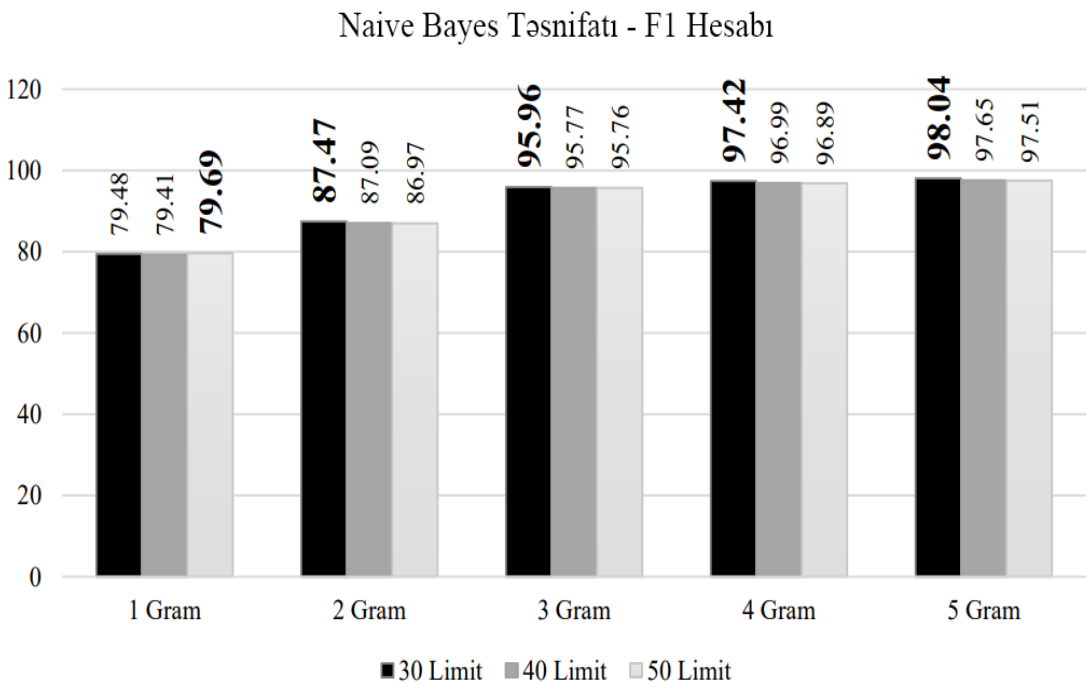
- Dəqiqlik və F1 skorunu müqayisə etsək, mənim nəticələrim böyük ölçüdə uyğun gəlir. Bu, məlumatların paylanmasının müntəzəm olduğunu göstərir.
- Naive Bayes 98,8% dəqiqliklə spam/e-poçt təsnifatında çox uğurlu olmuşdur.

Cədvəl 4-2 Xüsusiyyət nömrələri arasındakı fərqlər

	1 Gram	2 Gram	3 Gram	4 Gram	5Gram
LİMİT=30	**3335**	**4332**	**4286**	**3902**	**3487**
LİMİT=40	2658	3239	3131	2825	2525
LİMİT=50	2188	2550	2448	2192	1941

Cədvəl 4-3 Naive Bayes təsnifat təcrübəsinin nəticələri





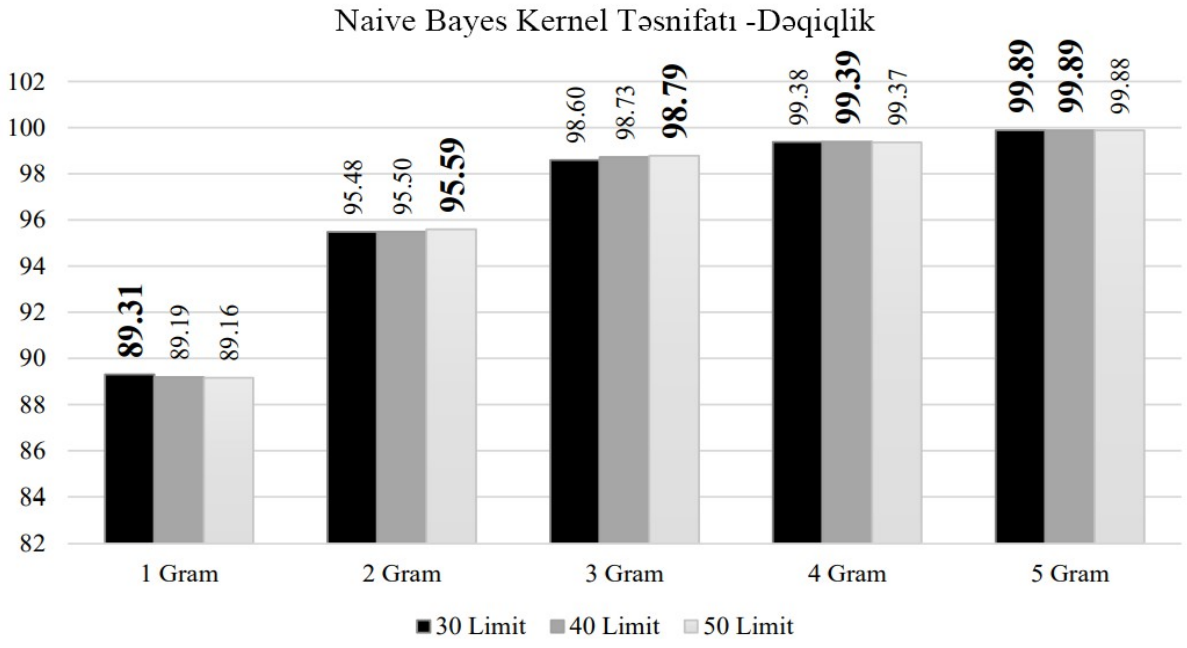
4.2.2 Sadə Bayes Kernel Təcrübə Nəticələri

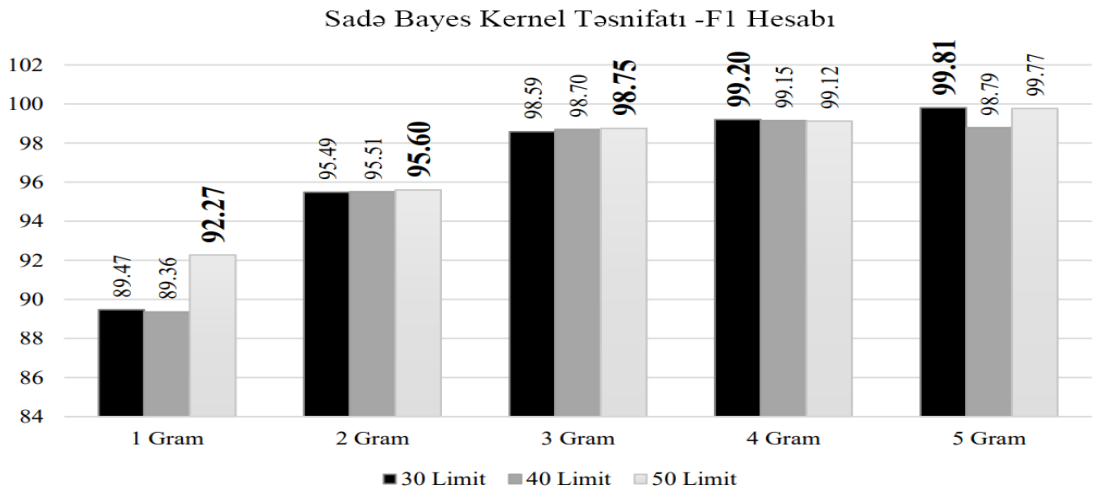
Ən yüksək dəqiqlik dərəcəsi 99,89% və 99,81% F1 Hesabı, Naive Bayes Kernel alqoritmi ilə 1,2,3,4,5 qramlıq məlumat dəstləri ilə aparılan təcrübələrdə əldə edilmişdir. Cədvəl 4.4-də verilmiş dəqiqlik və F1 skoru nəticələrinə baxdıqda aşağıdakı nəticələr əldə edilmişdir.

- Xüsusiyyət vektorunda sözlərin sayı, başqa sözlə, qramların sayı artdıqca, Naive Bayes Kernel alqoritminin dəqiqlik dərəcəsi ümumiyyətlə artır. 1 qram və 2 qramlıq təcrübələr arasında dəqiqlikdə böyük fərq var. Ancaq 1 qramlıq təcrübələrin F1 ballarına baxdığımızda məlumatların paylanmasında bir balanssızlığın olduğunu görürük. Dəqiqlik dərəcələrindəki anormal fərqin səbəbi budur. 4 qramlıq təcrübələrdə dəqiqlik dərəcəsi 3 qramdan aşağıdır. Yenə buradakı azalma səbəbi ilə F1 bal dəyərlərinə baxdığımızda görürük ki, 3 və 5 qramlıq təcrübələrin dəqiqlik və F1 bal dəyərləri oxşar olsa da, 4 qramda balanssızlıq var.Buradakı nəticələrdən göründüyü kimi, nəticələr arasındakı anormal fərqləri anlamaq üçün F1 balını və dəqiqlik dəyərini birlikdə şərh etmək lazımdır. Bu anormal fərqlər nəzərə alınmadıqda, ümumiyyətlə, bu alqoritm, Naive Bayes alqoritmi kimi artan qram sayına müsbət reaksiya verdi.

- Xüsusiyyətlərin sayının performansa təsirini araşdırmaq istənildikdə, 30, 40, 50 limitləri ilə hazırlanmış məlumat dəstləri ilə aparılan təcrübələr araşdırıldı. Təcrübə nəticələrinin nizamlı bir paylanma göstərmədiyi və kiçik fərqlər nəzərə alınmadıqda, 50 ilə məhdudlaşan xüsusiyyətlərin təsnifat üçün daha yaxşı olduğu müşahidə edildi. Naive Bayes-dən fərqli olaraq, Naive Bayes Kemel alqoritminin xüsusiyyətlərin sayının artmasına mənfi təsir göstərdiyi müşahidə edilmişdir.
- Dəqiqlik və F1 skoru müqayisə edildikdə, 1 qram və 4 qram ilə hazırlanmış məlumat dəstləri ilə aparılan təcrübələrdə limitlərə görə balanssızlığın olduğu və bu vəziyyətin performansa mənfi təsir göstərdiyi görüldü.
- Naive Bayes Kernel 99,89% dəqiqlik dərəcəsi ilə spam/e-poçt təsnifatında uğur qazandı.

Cədvəl 4-4 Naive Bayes Kernel təsnifatı təcrübəsinin nəticələri





4.3 Dəstək vektor maşın alqoritmləri ilə edilən təcrübələr

4.3.1Xətti SVM (Linear SVM) Təcrübə Nəticələri

Xətti SVM alqoritmi ilə 1,2,3,4,5 qramlıq məlumat dəstləri ilə aparılan təcrübələrdə ən yüksək 99,89% dəqiqlik dərəcəsi və 98,81% F1 balı əldə edilib. Cədvəl 4.5-də verilmiş dəqiqlik və F1 Score nəticələrinə baxıldıqda aşağıdakı nəticələr əldə edildi:

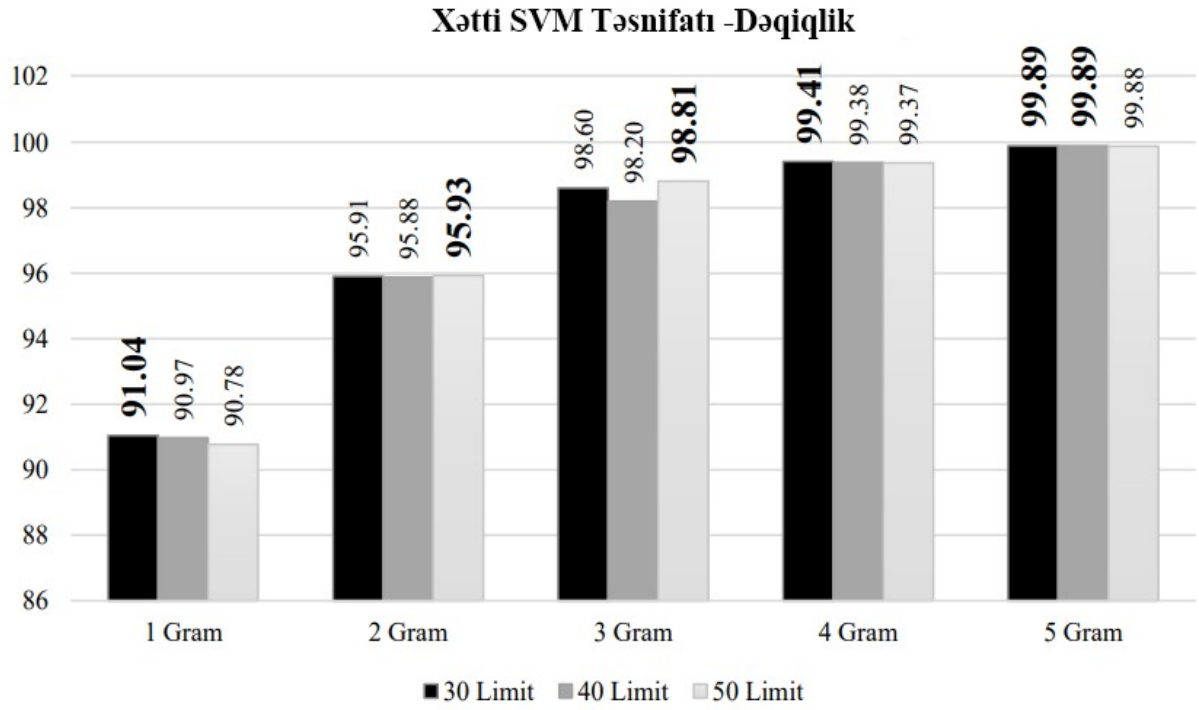
- Xüsusiyyət vektorunda qramların sayı artdıqca, Xətti SVM Alqoritminin dəqiqlik dərəcəsi müntəzəm olaraq artır və 99,89% çox yüksək müvəffəqiyyət dərəcəsinə çatır. Bu onu göstərir ki, Linear SVM alqoritmi, Sadə Bayes alqoritmləri kimi, uzun mətnlərin təsnifatında çox uğurludur.

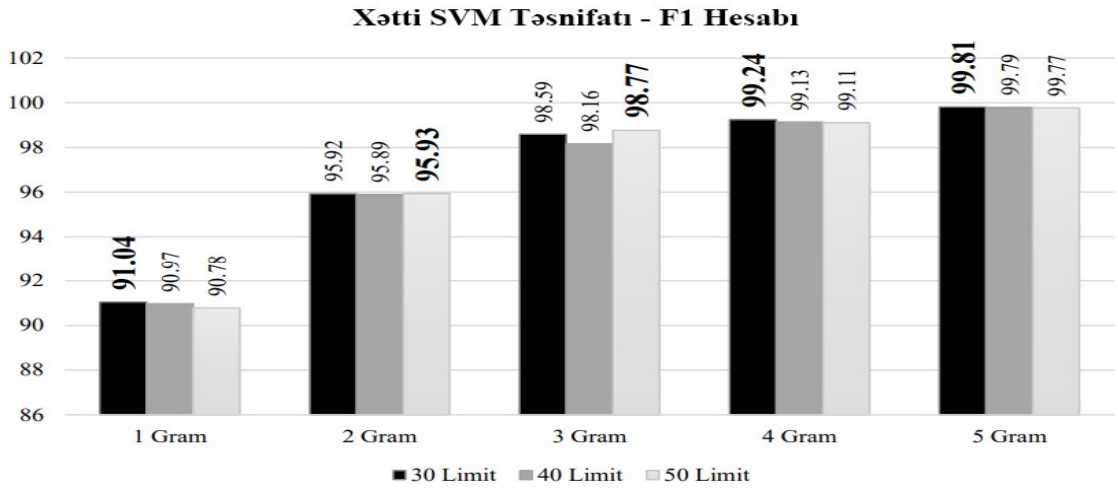
- Xüsusiyyətlərin sayının performansa təsirini araşdırmaq istənildikdə, 30, 40, 50 limitləri ilə hazırlanmış eksperimental dəstlər araşdırıldı. Təcrübələrdə müntəzəm paylama yoxdur. Feature Vector ölçüsünün performansa təsiri dəyişkəndir.

- Dəqiqlik və F1 Skoru müqayisə edildikdə, nəticələrin böyük ölçüdə üst-üstə düşdüyünü görmək olar. Bu, məlumatların paylanmasının müntəzəm olduğunu göstərir.

- Xətti SVM 99,89% dəqiqlik dərəcəsi ilə spam/e-poçt təsnifatında çox uğurlu olmuşdur.

Cədvəl 4-5 Xətti SVM təsnifatı təcrübəsinin nəticələri





4.3.2LibSVM Təcrübə Nəticələri

LibSVM alqoritmi ilə 1,2,3,4,5 qramlıq məlumat dəstləri ilə aparılan təcrübələrdə ən yüksək dəqiqlik dərəcəsi 99,76% və 99,6% F1 hesabı əldə edilmişdir. Cədvəl 4.6-da verilmiş dəqiqlik və F1 hesabı nəticələrinə baxdıqda aşağıdakı nəticələr əldə edilmişdir.

• Xüsusiyyət vektorunda qramların sayının artırılması performans artımında davamlı yaxşılaşma təmin etmədi. 1 qramlıq təcrübələrlə müqayisədə 2 qram ilə edilən təcrübələrdə performans artarkən, 3 qram ilə edilən təcrübələrdə 2 qramlıq təcrübələrə nisbətən dəqiqlik nisbəti azalıb.

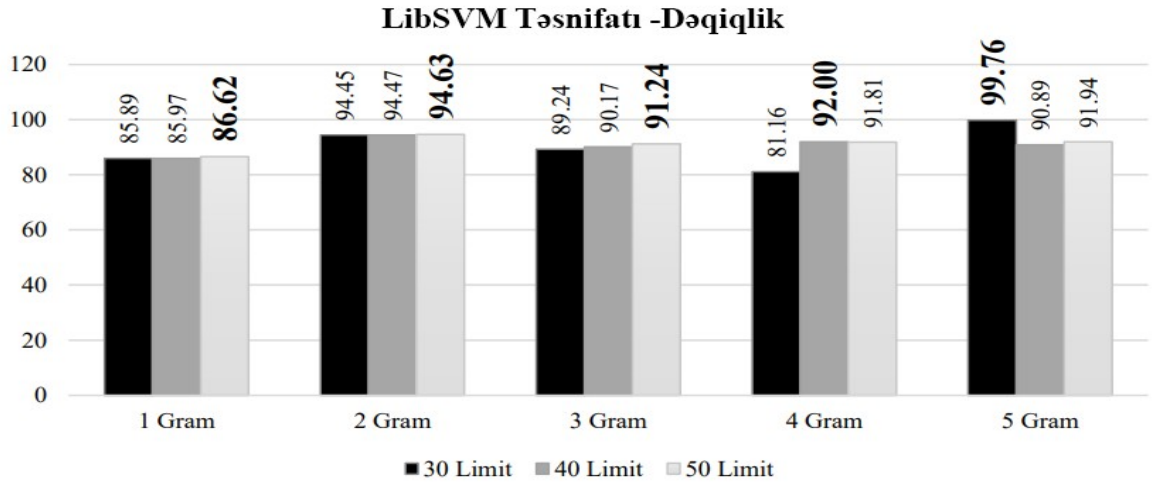
• Xüsusiyyətlərin sayının performansa təsirini araşdırmaq istənildikdə, 30, 40, 50 limitləri ilə hazırlanmış eksperimental dəstlər araşdırıldı. Təcrübələrdə 50 limitinin yüksək dəqiqlik təmin etdiyi görünsə də, ən yüksək dəqiqlik 5 qramlıq 30 limiti ilə edilən təcrübə ilə əldə edilmişdir.

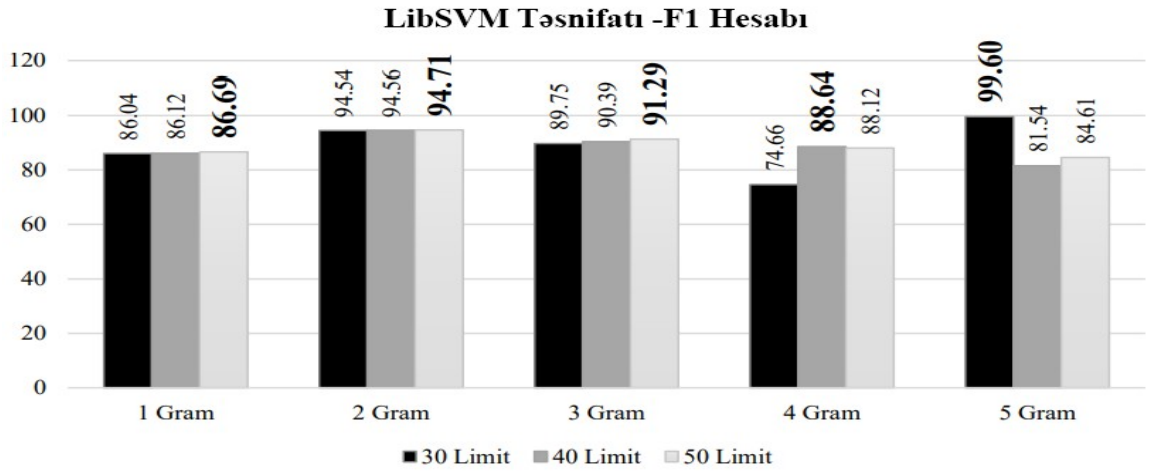
• LibSVM 1 qram ilə aparılan təcrübələrdə müvəffəqiyyət nisbəti olduqca aşağı olmuşdur.

• Dəqiqlik və F1 Skoru müqayisə edildikdə, nəticələrin böyük ölçüdə üst-üstə düşdüyünü görmək olar. Bu, məlumatların paylanmasının müntəzəm olduğunu göstərir.

• Lib SVM ən yüksək dəqiqlik dərəcəsi 99,76% ilə spam/e-poçt təsnifatında çox uğurlu olduğu aşkar edilmişdir.

Cədvəl 4-6 LibSVM təsnifatının sınaq nəticələri





4.3.3 Pegasos SVM Təcrübə Nəticələri

Pegasos SVM Alqoritmi ilə 1,2,3,4,5 qramlıq məlumat dəstləri ilə aparılan təcrübələrdə 99,89% və 99,8% F1 balı ilə ən yüksək dəqiqlik dərəcəsi əldə edilmişdir. Cədvəl 4.7-də verilmiş dəqiqlik və F1 Skoru nəticələrinə nəzər saldıqda aşağıdakı nəticələr əldə edilmişdir.

- Xüsusiyyət vektorunda sözlərin sayı, başqa sözlə, qramların sayı artdıqca dəqiqlik dərəcəsi müntəzəm olaraq artır. Bu, Pegasos SVM alqoritminin xüsusilə mətn təsnifatında və uzun mətnlərin təsnifatında çox uğurlu olduğunu göstərir. Bundan əlavə, qram sayı effekti LibSVM alqoritmindən daha aydın görünür.

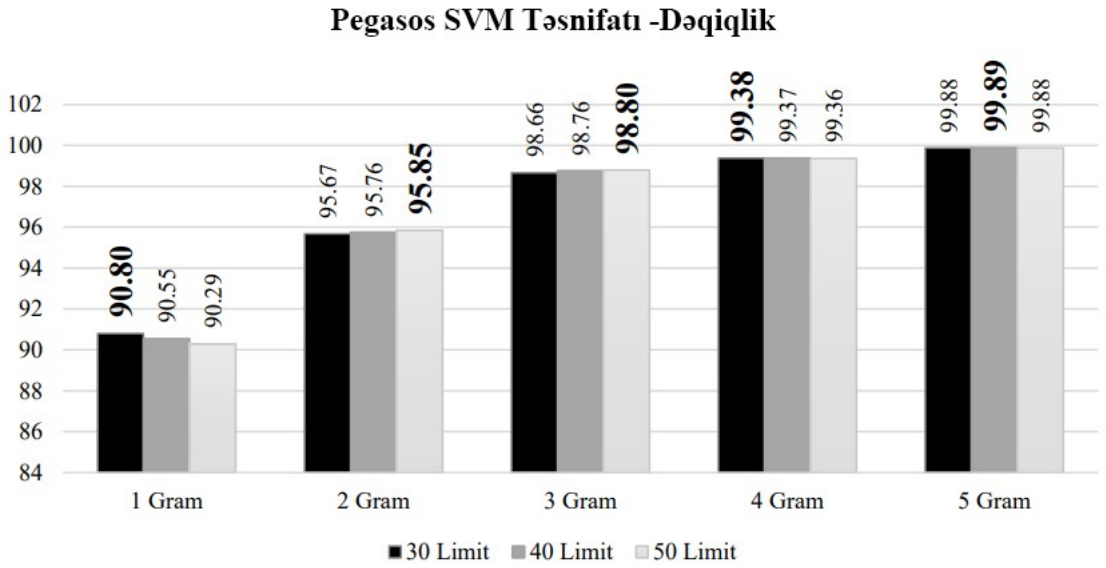
- Xüsusiyyətlərin sayının performansa təsirini araşdırmaq istənildikdə, 30, 40, 50 limitləri ilə hazırlanmış eksperimental dəstlər araşdırıldı. Müntəzəm artım olmasa da, limitlərin dəqiqlik dərəcələri bir-birinə kifayət qədər yaxındır. Xüsusilə 40 limiti ilə aparılan təcrübələr ümumiyyətlə uğurlu görünür.

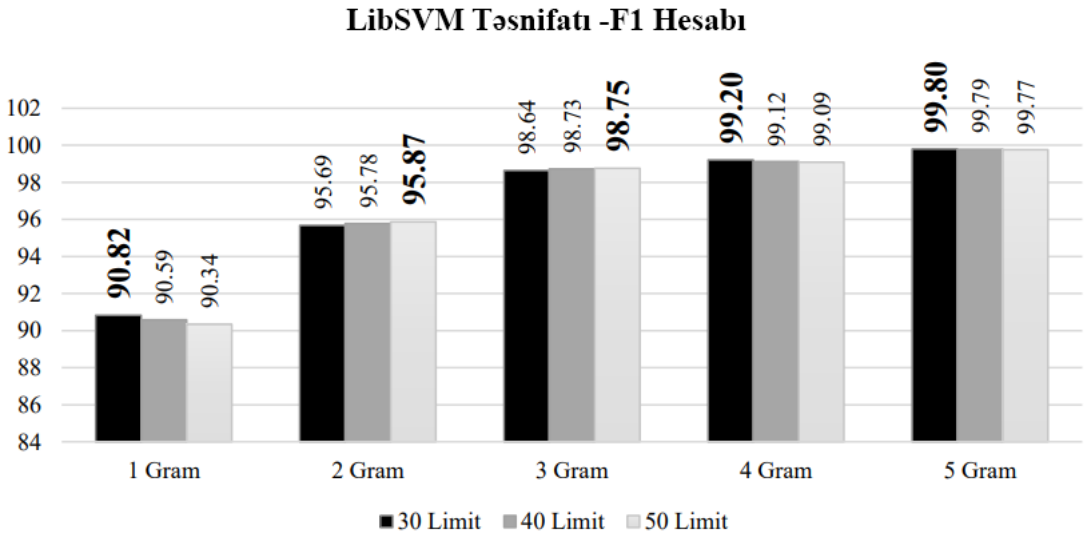
- Dəqiqlik və F1 Skoru müqayisə edildikdə, nəticələrin böyük ölçüdə üst-üstə düşdüyünü görmək olar. Bu, məlumatların paylanmasının müntəzəm olduğunu göstərir.

- 1 qram ilə aparılan təcrübələr istisna olmaqla, Pegasos SVM-nin ümumiyyətlə Lib SVM-dən daha yaxşı olduğu və dəqiqlik baxımından Linear SVM-ə yaxın olduğu

müşahidə edilmişdir. Fərqli qramlara və məhdudiyyətlərə görə, ümumi müvəffəqiyyət nisbəti 96% -dən yuxarı qaldı. Alqoritmin ən yüksək dəqiqlik dərəcəsi 99,89% olmaqla, spam/e-poçt təsnifatında çox uğurlu olduğu aşkar edilmişdir.

Cədvəl 4-7 Pegasos SVM təsnifat testinin nəticələri





4.4 Qərar ağacı alqoritmləri ilə edilən təcrübələr

Qərar Ağacı, Qradient Gücləndirilmiş Ağaclar, Təsadüfi Ağac və Təsadüfi Meşə kimi alqoritmlərlə aparılan təcrübələr əvvəlcə 50 limitlə məhdudlaşan 3 qramlıq məlumat dəsti ilə aparıldı. Ədəbiyyat araşdırmalarından əldə edilən nəticələr və eksperimental mühitdə müvəffəqiyyət nisbətinin kifayət qədər aşağı olması səbəbindən qərar verilmədi.

Ağac alqoritmlərinin spam təsnifatı üçün uyğun olmadığı qənaətinə gəlindi. Bu qənaətə gəlməyin əsas səbəblərini aşağıdakı kimi sadalamaq olar:

- Təcrübələrdə istifadə olunan verilənlər toplusunda xüsusiyyət vektorunun ölçüsü 2448-dir. Xüsusiyyətlərin çoxluğu həm ağacın yaradılmasının, həm də ağacın budamasının mürəkkəbliyini artırır.
- Verilənlərin çoxluğuna və Feature vektorunun böyük ölçüsünə görə bütün verilənlərin xüsusiyyətlərini modelləşdirə bilməyən ağac yaradılmışdır.

Qərar ağacı öyrənənlər bəzi dərsləri artıq öyrəniblərsə, qərəzli ağaclar yaradırlar.

Cədvəl 4.8-dəki nəticələr tədqiq edildikdə, qərar ağacı alqoritmlərinin spam təsnifatında çox aşağı müvəffəqiyyət qazandığı görülür.

Cədvəl 4-8 Qərar ağacı alqoritmləri ilə təsnifat təcrübəsinin nəticələri

	Qərar ağacları	Gradient Boosted Trees	Decision Stump	Random Tree	Random Forest
Doğruluq	86.49	79.19	63.26	59.10	59.41
F1- Hesabı	91.00	78.78	65.52	37.15	61.71
Təsnifat səhvi	13.51% +/- 0.38%	20.81% +/- 19.15%	36.74% +/- 0.22%	40.90% +/- 0.01%	40.59% +/- 0.36%

4.5 Süni neyron şəbəkələri ilə edilən təcrübələr

4.5.1 Perceptron Təcrübə Nəticələri

Ən yüksək dəqiqlik dərəcəsi 99,70% və 99,43% F1 hesabı perceptron alqoritmi ilə RapidMiner mühitində 1,2,3,4,5 qramlıq məlumat dəstləri ilə aparılan təcrübələrdə əldə edilmişdir. Cədvəl 5.9-da verilmiş dəqiqlik və F1 Skoru nəticələrinə baxdıqda, aşağıdakı nəticələr əldə edilmişdir.

- Xüsusiyyət vektorunda qramların sayı artdıqca Perceptron alqoritminin dəqiqlik dərəcəsi müntəzəm olaraq artır.

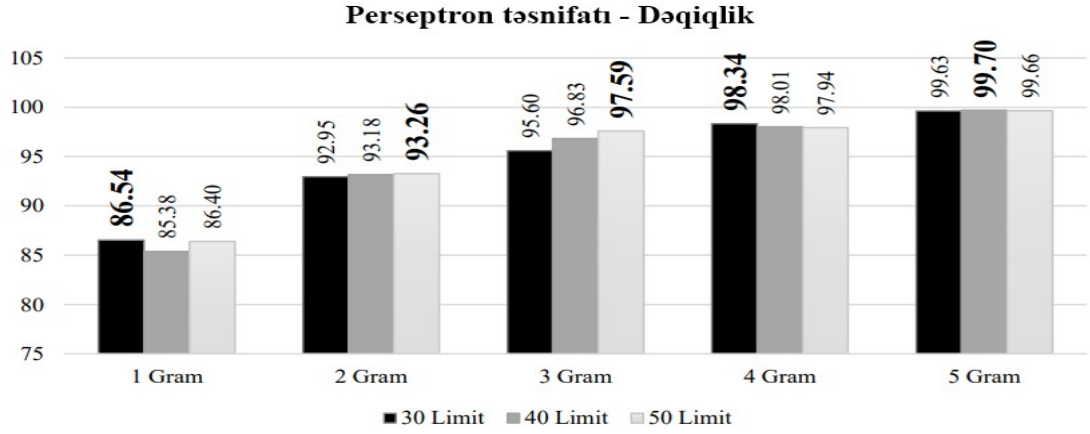
- Ən yaxşı performans 5 qram xüsusiyyət vektorları ilə hazırlanmış məlumat dəstlərində müşahidə edilmişdir. Ən aşağı performansın 1 qram xüsusiyyət vektoru ilə hazırlanmış məlumat dəstlərində olduğu müşahidə edilmişdir.

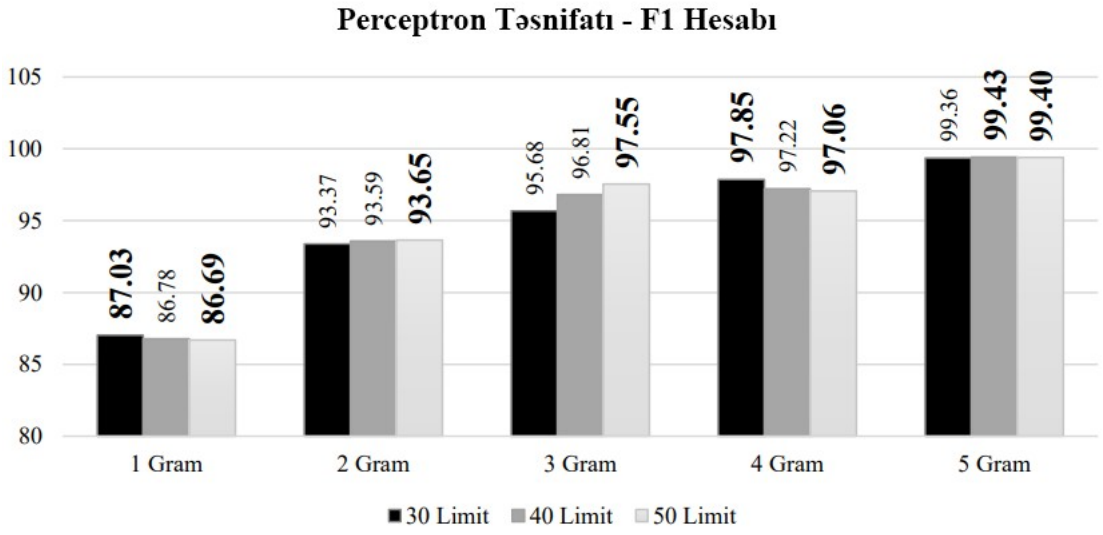
- Xüsusiyyətlərin sayının performansa təsirini araşdırmaq istənildikdə, 30, 40, 50 limitləri ilə hazırlanmış eksperimental dəstlər araşdırıldı. Təcrübələrdə ən yüksək dəqiqliyin ümumiyyətlə 50 həddi ilə əldə edildiyi müşahidə edildi.

- Dəqiqlik və F1 Skoru müqayisə edildikdə, nəticələrin böyük ölçüdə üst-üstə düşdüyünü görmək olar. Bu, məlumatların paylanmasının müntəzəm olduğunu göstərir.

- Perceptron 98,8% dəqiqliklə spam/e-poçt təsnifatında çox uğurlu olmuşdur.

Cədvəl 4-9 Perseptron təsnifatı üzrə təcrübənin nəticələri





4.6 Ən yaxın qonşu (K-NN) alqoritmi ilə edilən təcrübələr

KNN alqoritmi ilə RapidMiner mühitində 1,2,3,4,5 qramlıq məlumat dəstləri ilə aparılan təcrübələrdə 99,72% və 99,52% F1 hesabı ilə ən yüksək dəqiqlik dərəcəsi əldə edilmişdir. Cədvəl 5.10-da verilmiş dəqiqliyə və F1 Skoru nəticələrinə nəzər saldıqda, aşağıdakı nəticələr əldə edilmişdir:

- Xüsusiyyət vektorunda sözlərin sayı, başqa sözlə, qramların sayı artdıqca KNN alqoritminin dəqiqlik dərəcəsi müntəzəm olaraq artır. Bu vəziyyət KNN alqoritmini xüsusilə mətn təsnifatında və uzun mətnlərin təsnifatında çox çətinləşdirir və uğurlu olduğunu göstərir.

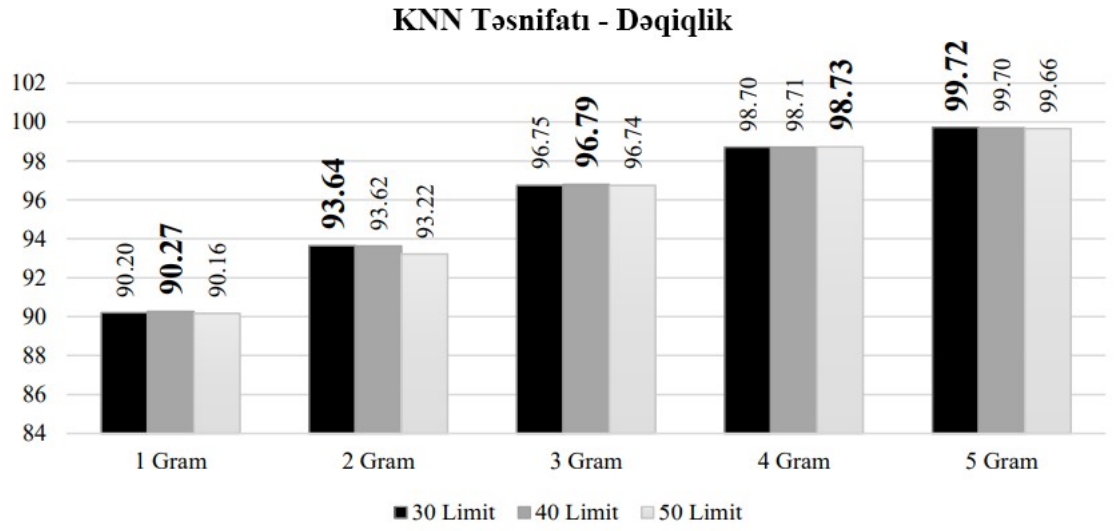
- Xüsusiyyətlərin sayının performansa təsirini araşdırmaq istənildikdə, 30, 40, 50 limitləri ilə hazırlanmış eksperimental dəstlər araşdırıldı. Ümumiyyətlə, 30 limiti ilə aparılan təcrübələrin yüksək dəqiqlik dərəcəsinə malik olduğu görünür.

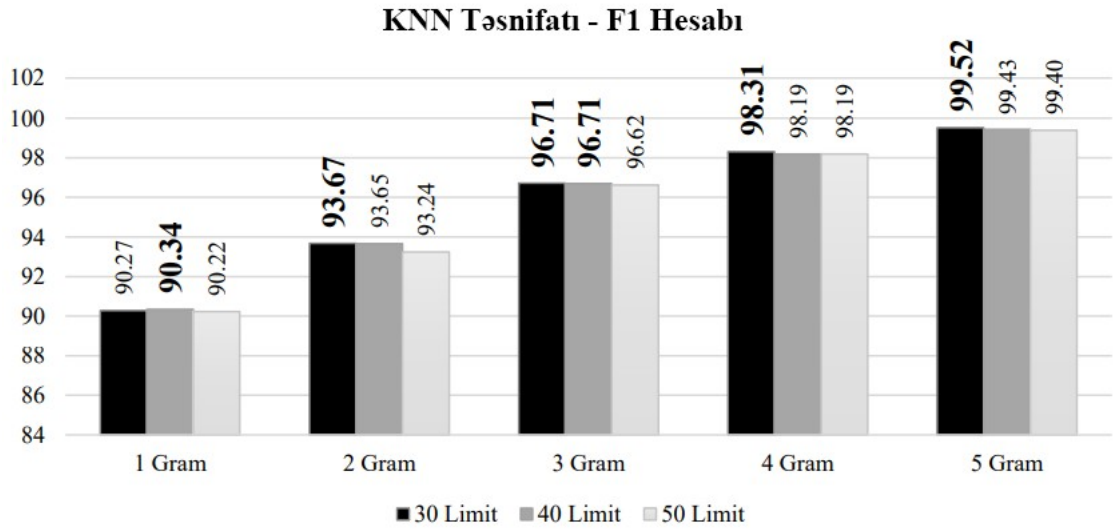
- 5 qramla aparılan təcrübələrin ən uğurlu olduğu görünür.

- Dəqiqlik və F1 Skoru müqayisə edildikdə, nəticələrin böyük ölçüdə üst-üstə düşdüyünü görmək olar. Bu, məlumatların paylanmasının müntəzəm olduğunu göstərir.

• Naive Bayes 99,72% dəqiqlik dərəcəsi ilə spam/e-poçt təsnifatında çox uğurlu olduğu aşkar edilmişdir.

Cədvəl 4-10 KNN təsnifat təcrübəsinin nəticələri





NƏTİCƏ

Bu tədqiqatın əsas məqsədi elektron məktubların məzmununda olan keçid mətnləri ilə spam e-poçtları fərqləndirməkdir. Tədqiqatda istifadə ediləcək məlumat

dəstləri müxtəlif ön proseslərdən keçərək maşın öyrənmə texnikalarında istifadəyə hazır vəziyyətə gətirilir. Məlumat dəstləri hazırlanarkən Word Set Technique (BOW) ilə 1,2,3,4,5 qram olaraq hazırlanmış 5 xüsusiyyət vektorundan istifadə edilmişdir. Müxtəlif qramlarda xüsusiyyət vektorlarının yaradılmasında əsas məqsəd qramların sayının təsnifat göstəricilərinə təsirini müşahidə etməkdir. Yaradılan N qramların sayı çox yüksəkdir, başqa sözlə, xüsusiyyət vektorunun ölçüsü çox uzundur və buna maşın öyrənməsi səbəb olur.

Təlim və sınaq zamanı çətinliklərə görə xüsusiyyət vektorlarının ölçüsü müxtəlif məhdudiyyətlərlə azaldılıb. Bu limitlər 30,40,50-dir. Limitlər eyni N qramın təkrar sayı nəzərə alınmaqla yaradılmışdır.Maşın öyrənmə üsullarını tətbiq edərkən əvvəlcə 50 limit 3 Gram xüsusiyyət vektoru ilə hazırlanmış məlumat dəsti öyrədildi və sınaqdan keçirildi. Nəticə olaraq, təsnifat dəqiqliyi dərəcəsi orta hesabla 95%-dən yuxarı olan təsnifat alqoritmləri üçün N qram sayından istifadə edilir.Onun performansa təsiri araşdırıldı. Dəqiqlik dərəcəsi 95%-dən aşağı olan təsnifat alqoritmləri üçün N qram sayının performansa təsirini müşahidə etmək mənasız idi, çünki dəqiqlik dərəcəsi istənilən səviyyədə deyildi.Tədqiqatın başqa bir məqsədi spam e-poçt təsnifatı üçün müxtəlif maşın öyrənmə üsullarının müvəffəqiyyət səviyyəsini araşdırmaqdır. Təhlil aparılarkən ilk növbədə verilənlərin növünə uyğun təsnifat alqoritmləri nəzərdən keçirilmiş və müvafiq alqoritmlər seçilmişdir.Maşın öyrənmə təcrübələri aparılıb. Tədqiqat çərçivəsində müzakirə edilən maşın öyrənmə üsulları "Bayes, Support Vector Machines, Decision Trees, Süni Sinir Şəbəkələri və Perceptron Alqoritmləri" kimi qruplaşdırılıb. Bu qruplardan qərar ağacı N qramın performansa təsiri istisna olmaqla bütün alqoritmlər üçün yoxlanılmışdır. Qərar Ağacı Alqoritmləri üçün orta dəqiqlik dərəcəsi 65% təşkil etmişdir. Bütün eksperimentlər RapidMiner alətindən istifadə etməklə 10-qat çarpaz doğrulama ilə aparılmışdır.Bütün eksperimentlər tamamlandıqdan sonra analiz prosesləri üçün dəqiqlik, F1 balı və təsnifat xəta dərəcəsi metrikləri və spam təsnifatı performansı yoxlanıldı.Cədvəl 6.1-dəki təsnifat alqoritmlərinin dəqiqlik dərəcələri tədqiq edildikdə aşağıdakı nəticələr əldə edilmişdir:

- Qramların sayını nəzərə alsaq, bütün təcrübələrdə ən aşağı dəqiqlik dərəcəsinin 1 qram xüsusiyyət vektoru ilə hazırlanmış verilənlər toplusundan alındığını görmək olar. Bu vəziyyətdən 1 qram xüsusiyyət vektorlu spam e-poçtların təsnifatında istifadə edilə bilər.Hazırlanmış məlumat dəstlərinin spam e-poçtları ayırd etmək üçün kifayət etmədiyi qənaətinə gəlindi.

- Yenə qram sayına görə baxdıqda, N qram sayı artdıqca maşın öyrənmə alqoritmlərinin dəqiqlik dərəcələrinin getdikcə artdığı görünür. Bu, N qramla aparılan spam e-poçt təsnifat işlərində N qramda N sayının artırılmasının təsnifat dəqiqliyi üçün çox yaxşı bir həll olacağını göstərir.

- Müşahidə edilmişdir ki, alqoritmlərin əksəriyyətində ən yüksək dəqiqlik dərəcəsi 5 qram 30 limit məlumat dəsti ilə aparılan təcrübələrdə əldə edilir. Digər məlumat dəstləri ilə aparılan təcrübələrdə ən yüksək dəqiqliyə 5 qram 40 limit məlumat dəsti ilə aparılan təcrübələrdə nail olunub. Bu onu göstərir ki, 5 qramlıq xüsusiyyət vektoru ilə hazırlanan məlumat dəstləri spam e-poçtların təsnifatında əhəmiyyətli fərq yaradır.

- Çox aşağı dəqiqlik dərəcəsi ilə təsnif edilən Qərar Ağacı Alqoritmləri arasında Qərar Ağacı, Qradient Gücləndirilmiş Ağaclar, Qərar Ağacı, Təsadüfi Ağac və Təsadüfi Meşə alqoritmlərinin təsnifatında alqoritmlərin ümumi uğurunu nəzərə alaraq. Əksinə, Cədvəl 6.1-də verilmişdir

Nəticələrə görə, Naive Bayes (98,8%), Naive Bayes Kernel (99,89%), LibSVM (99,76%), LinearSVM (99,89%), Pegasos SVM (99,89%), Perceptron (99,7%), KNN (99,72%) alqoritmlər %. 98,8-dən yuxarı dəqiqlik dərəcəsi spam e-poçtların təsnifatında çox uğurlu olmuşdur.

Cədvəl 5-1 Təsnifat alqoritmləri ilə aparılan təcrübələrin dəqiqlik nəticələri

Məlumat dəsti		Metodların dəqiqliyi						
	N Gram	NB	NBK	Lib SVM	Linear SVM	Pegasos	Percep.	K-NN
Limit=30	1 Gram	78.03	89.31	85.89	91.04	90.80	86.54	90.20
	2 Gram	86.16	95.48	94.45	95.91	95.67	92.95	93.64
	3 Gram	95.84	98.60	89.24	98.60	98.66	95.60	96.75
	4 Gram	97.91	99.38	81.16	99.41	99.38	98.34	98.70
	5 Gram	**98.80**	**99.89**	**99.76**	**99.89**	99.88	99.63	**99.72**
Limit=40	1 Gram	77.97	89.19	85.97	90.97	90.55	85.38	90.27
	2 Gram	85.69	95.50	94.47	95.88	95.76	93.18	93.62
	3 Gram	95.67	98.73	90.17	98.20	98.76	96.83	96.79
	4 Gram	97.71	99.39	92.00	99.38	99.37	98.01	98.71
	5 Gram	98.67	**99.89**	90.89	**99.89**	**99.89**	**99.70**	99.70
Limit=50	1 Gram	78.37	89.16	86.62	90.78	90.29	86.40	90.16
	2 Gram	85.54	95.59	94.63	95.93	95.85	93.26	93.22
	3 Gram	95.67	98.79	91.24	98.81	98.80	97.59	96.74
	4 Gram	97.66	99.37	91.81	99.37	99.36	97.94	98.73
	5 Gram	98.53	99.88	91.94	99.88	99.88	99.66	99.66

Cədvəl 5-2 Xüsusiyyət vektoru limitlərinə görə dəqiqlik dərəcələri arasındakı fərqlər

	Limitlərə Görə Dəqiqlik Dərəcələri Arasındakı Fərqlər							
	N Gram	NB	NBK	Lib SVM	Linear SVM	Pegasos	Perceptron	K-NN
30 Limit-40 Limit fərqi	1 Gram	0.06	0.12	0.08	0.07	0.25	1.16	0.07
	2 Gram	0.47	0.02	0.02	0.03	0.09	0.23	0.02
	3 Gram	0.17	0.13	0.93	**0.40**	0.10	1.23	0.04
	4 Gram	0.20	0.01	**10.84**	0.03	0.01	0.33	0.01
	5 Gram	0.13	0.00	8.87	0.00	0.01	0.07	0.02
40 Limit-50 Limit fərqi	1 Gram	0.40	0.03	0.65	0.19	0.26	1.02	0.11
	2 Gram	0.15	0.09	0.16	0.05	0.09	0.08	0.40
	3 Gram	0.00	0.06	1.07	0.61	0.04	0.76	0.05
	4 Gram	0.05	0.02	0.19	0.01	0.01	0.07	0.02
	5 Gram	0.14	0.01	1.05	0.01	0.01	0.04	0.04
30 Limit-50 Limit fərqi	1 Gram	0.34	0.15	0.73	0.26	**0.51**	0.14	0.04
	2 Gram	**0.62**	0.11	0.18	0.02	0.18	0.31	**0.42**
	3 Gram	0.17	**0.19**	2.00	0.21	0.14	**1.99**	0.01
	4 Gram	0.25	0.01	10.65	0.04	0.02	0.40	0.03
	5 Gram	0.27	0.01	7.82	0.01	0.00	0.03	0.06

• Cədvəl 6.3-də təsnifat xətalarının dərəcələrinə nəzər saldıqda onların Cədvəl 6.1-dəki nəticələri dəstəklədiyi görünür. Bütün alqoritmlərdə ən yüksək təsnifat xətasının 1 qram xüsusiyyət vektoru ilə hazırlanmış məlumat dəstlərində olduğu görülür.

Cədvəl 5-3 Təsnifat xəta dərəcələri

Məlumat dəsti		Metodların Təsnifat Səhvləri Dərəcəsi						
	N Gram	NB	NBK	Lib SVM	Linear SVM	Pegasos	Percep.	K-NN
Limit=30	1 Gram	21.97% +/- 0.41%	**10.69% +/- 0.47%**	**14.11% +/- 0.44%**	8.96% +/- 0.29%	4.33% +/- 0.22%	13.46% +/- 0.42%	9.84% +/- 0.43%
	2 Gram	13.84% +/- 0.56%	4.52% +/- 0.32%	5.55% +/- 0.28%	4.09% +/- 0.21%	1.34% +/- 0.18%	7.05% +/- 0.32%	6.78% +/- 0.42%
	3 Gram	4.16% +/- 0.31%	1.40% +/- 0.17%	10.76% +/- 0.86%	1.40% +/- 0.16%	1.34% +/- 0.18%	4.40% +/- 1.51%	3.26% +/- 0.29%
	4 Gram	2.09% +/- 0.27%	0.62% +/- 0.14%	18.84% +/- 0.48%	0.59% +/- 0.12%	0.62% +/- 0.11%	1.66% +/- 0.52%	1.27% +/- 0.34%
	5 Gram	1.20% +/- 0.20%	0.11% +/- 0.07%	0.24% +/- 0.08%	0.11% +/- 0.07%	0.12% +/- 0.07%	0.37% +/- 0.07%	0.34% +/- 0.13%
Limit=40	1 Gram	**22.03% +/- 0.50%**	10.81% +/- 0.28%	14.03% +/- 0.55%	**9.03% +/- 0.30%**	9.45% +/- 0.26%	**14.62% +/- 0.48%**	9.80% +/- 0.67%
	2 Gram	14.31% +/- 0.32%	4.50% +/- 0.23%	5.53% +/- 0.34%	4.12% +/- 0.25%	4.24% +/- 0.29%	6.82% +/- 0.29%	6.36% +/- 0.49%
	3 Gram	4.33% +/- 0.27%	1.27% +/- 0.12%	9.83% +/- 0.15%	1.80% +/- 1.72%	1.24% +/- 0.09%	3.17% +/- 0.35%	3.25% +/- 0.32%
	4 Gram	2.29% +/- 0.23%	0.61% +/- 0.14%	8.00% +/- 0.21%	0.62% +/- 0.15%	0.63% +/- 0.14%	1.99% +/- 0.30%	1.30% +/- 0.17%
	5 Gram	1.33% +/- 0.21%	0.11% +/- 0.07%	9.11% +/- 0.46%	0.11% +/- 0.07%	0.11% +/- 0.07%	0.30% +/- 0.12%	0.28% +/- 0.09%
Limit=50	1 Gram	21.63% +/- 0.31%	10.84% +/- 0.25%	13.38% +/- 0.52%	9.22% +/- 0.47%	**9.71% +/- 0.43%**	13.60% +/- 0.67%	**9.73% +/- 0.41%**
	2 Gram	14.46% +/- 0.34%	4.41% +/- 0.26%	5.37% +/- 0.18%	4.07% +/- 0.28%	4.15% +/- 0.24%	6.74% +/- 0.27%	6.38% +/- 0.40%
	3 Gram	14.46% +/- 0.34%	1.21% +/- 0.11%	8.76% +/- 0.41%	1.19% +/- 0.12%	1.20% +/- 0.13%	2.41% +/- 0.20%	3.21% +/- 0.26%
	4 Gram	2.34% +/- 0.27%	0.63% +/- 0.10%	8.19% +/- 0.76%	0.63% +/- 0.10%	0.64% +/- 0.10%	2.06% +/- 0.28%	1.29% +/- 0.18%
	5 Gram	1.47% +/- 0.22%	0.12% +/- 0.08%	8.06% +/- 0.27%	0.12% +/- 0.08%	0.12% +/- 0.08%	0.34% +/- 0.13%	0.30% +/- 0.12%

Nəticə olaraq, bu işdə, maşın öyrənmə üsulları, bu bağlantılarla əlaqəli bağlantılar və mətnlər olan e-poçtlara qarşı araşdırıldı və dəqiqlik dərəcəsi olduqca yüksək idi. Bu tədqiqat həmçinin spam mesajlarının aşkar edilməsində mətnin əvvəlcədən işlənməsi strategiyalarının mərkəzi əhəmiyyətini nümayiş etdirir. Nəticələr göstərir ki, ümumi nümunələri müşahidə etmək olar. Çıxarılan söz seqmentlərinin sayı və uzunluğu klassifikatorların işinə böyük təsir göstərir.

Ümumilikdə tədqiqata keçidləri olmayan e-poçtlar daxil edilmir. Bu səbəblə gələcək tədqiqatlar burada araşdırmaya əlavə olaraq gövdə və başlıq kimi sahələri də daxil etməklə arzuolunmaz e-poçtları aşkar etməyi hədəfləyir.

ƏDƏBİYYAT

1. Ajaz, S., Nafis, M. T., & Sharma, V. (2017). Spam Mail Detection Using Hybrid Secure Hash Based Naive Classifier. *International Journal of Advanced Research in Computer Science*, *8*(5).
2. Albayrak, A. S., & Yilmaz, S. K. (2009). VERİ MADENCİLİĞİ: KARAR AĞACI ALGORİTMALARI VE İMKB VERİLERİ ÜZERİNE BİR UYGULAMA. *Suleyman Demirel University Journal of Faculty of Economics & Administrative Sciences*, *14*(1).
3. Aliusta, C., & Benzer, R. (2018). Avrupa Siber Suçlar Sözleşmesi ve Türkiye'nin Dahil Olma Süreci. *Uluslararası Bilgi Güvenliği Mühendisliği Dergisi*, *4*(2), 35-42.
4. Aliusta, C., & Benzer, R. (2018). Avrupa Siber Suçlar Sözleşmesi ve Türkiye'nin Dahil Olma Süreci. *Uluslararası Bilgi Güvenliği Mühendisliği Dergisi*, *4*(2), 35-42.
5. Andress, J., & Winterfeld, S. (2013). Cyber warfare: techniques, tactics and tools for security practitioners. Elsevier.
6. Athanasopoulos, A., Dimou, A., Mezaris, V., & Kompatsiaris, I. (2011, April). GPU acceleration for support vector machines. In *Procs. 12th Inter. Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS 2011), Delft, Netherlands* (Vol. 164).
7. Ayhan, S., & Erdoğmuş, Ş. (2014). Destek vektör makineleriyle sınıflandırma problemlerinin çözümü için çekirdek fonksiyonu seçimi. *Eskişehir Osmangazi Üniversitesi İktisadi ve İdari Bilimler Dergisi*, *9*(1), 175-201.
8. Breheny, P. Kernel density classification. *STA*, *621*.
9. Brownlee, J. (2014). Classification accuracy is not enough: More performance measures you can use. *Machine Learning Mastery*, *21*.
10. Brownlee, J. (2014). Discover feature engineering, how to engineer features and how to get good at it. Machine Learning Process.

11. Brownlee, J. (2017). A gentle introduction to the bag-of-words model. *Machine Learning Mastery*, *21*.
12. Bulut, F., & Amasyalı, M. (2014). En Yakın k Komşuluk Algoritmasında Örneklere Bağlı Dinamik k Seçimi. *ASYU'2014: Akıllı Sistemlerde Yenilikler ve Uygulamaları*, 62-66.
13. Carreras, X., & Marquez, L. (2001). Boosting trees for anti-spam email filtering. *arXiv preprint cs/0109015*.
14. Chang, C. C., & Lin, C. J. (2011). LIBSVM: a library for support vector machines. *ACM transactions on intelligent systems and technology (TIST)*, *2*(3), 1-27.
15. Chen, S., Han, Z., Elahi, M. M., Habib, K. M., Wang, L., Wen, B., ... & Dean, C. R. (2016). Electron optics with pn junctions in ballistic graphene. *Science*, *353*(6307), 1522-1525.
16. Chuan, Z., Xianliang, L., Mengshu, H., & Xu, Z. (2005). A LVQ-based neural network anti-spam email approach. *ACM SIGOPS Operating Systems Review*, *39*(1), 34-39.
17. Cianflone, A., & Kosseim, L. (2017). N-gram and neural language models for discriminating similar languages. *arXiv preprint arXiv:1708.03421*.
18. Elmas, Ç. (2018). Yapay zeka uygulamaları.
19. Halawa, H., Ripeanu, M., Beznosov, K., Coskun, B., & Liu, M. (2017, November). An early warning system for suspicious accounts. In *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security* (pp. 51-52).
20. Kumagai, A., & Iwata, T. (2017, August). Learning Latest Classifiers without Additional Labeled Data. In *IJCAI* (pp. 2039-2045).
21. Lee, C. N., Chen, Y. R., & Tzeng, W. G. (2017, August). An online subject-based spam filter using natural language features. In *2017 IEEE Conference on Dependable and Secure Computing* (pp. 479-487). IEEE.
22. Ray, S. (6). Easy Steps to Learn Naive Bayes Algorithm with codes in Python and R. *Analytics Vidhya*.

23. Renuka, D. K., & Rajamohana, P. V. S. (2017). An ensembled classifier for email spam classification in hadoop environment. *Appl. Math*, *11*(4), 1123-1128.
24. Safavian, S. R., & Landgrebe, D. (1991). A survey of decision tree classifier methodology. *IEEE transactions on systems, man, and cybernetics*, *21*(3), 660-674.
25. Sah, U. K., & Parmar, N. (2017). An approach for malicious spam detection in email with comparison of different classifiers. *International Research Journal of Engineering and Technology (IRJET)*, *4*(8), 2238-2242.
26. Shalev-Shwartz, S., Singer, Y., & Srebro, N. (2007, June). Pegasos: Primal estimated sub-gradient solver for svm. In *Proceedings of the 24th international conference on Machine learning* (pp. 807-814).
27. Sharma, A. K., & Sahni, S. (2011). A comparative study of classification algorithms for spam email data analysis. *International Journal on Computer Science and Engineering*, *3*(5), 1890-1895.
28. Sharmila, A., & Geethanjali, P. J. I. A. (2016). DWT based detection of epileptic seizure from EEG signals using naive Bayes and k-NN classifiers. *Ieee Access*, *4*, 7716-7727.
29. Soumya George, K., & Joseph, S. (2014). Text classification by augmenting bag of words (BOW) representation with co-occurrence feature. *IOSR J. Comput. Eng*, *16*(1), 34-38.
30. Şahin, E. (2018). *Makine öğrenme yöntemleri ve kelime kümesi tekniği ile istenmeyen e-posta/e-posta sınıflaması* (Master's thesis, Fen Bilimleri Enstitüsü).
31. Şengöz, M. (2024). YAPAY ZEKÂNIN KAMUOYU ALGISININ YÖNETİLMESİ NOKTASINDA KULLANILABİLMESİNE DAİR BİR DEĞERLENDİRME. *Habitus Toplumbilim Dergisi*, *5*(5), 95-114.
32. Taşcı, E., & Onan, A. (2016). K-en yakın komşu algoritması parametrelerinin sınıflandırma performansı üzerine etkisinin incelenmesi. *Akademik Bilişim*, *1*(1), 4-18.

33. Wang, J., Gao, K., Jiao, Y., & Li, G. (2009). Study on ensemble classification methods towards spam filtering. In *Advanced Data Mining and Applications: 5th International Conference, ADMA 2009, Beijing, China, August 17-19, 2009. Proceedings 5* (pp. 314-325). Springer Berlin Heidelberg.