AZƏRBAYCAN RESPUBLİKASI ELM və TƏHSİL NAZİRLİYİ

AZƏRBAYCAN TEXNİKİ UNİVERSİTETİ

Əlyazması hüququnda

Abdullayev Elnur Nasir oğlu

"MƏTİNLƏRİN KONTEKSƏ GÖRƏ SİNİFLƏNDİRİLMƏSİ ÜÇÜN İNTELLEKTUAL METODUN YARADILMASI"

mövzusunda

MAGİSTRLİK DİSSERTASİYASI

İxtisasın: 060509 - Kompüter elmləri

İxtisaslaşma: İntellektual Sistemlər

Elmi rəhbər: r.ü.f.d. Hüseynova N.Ş.

BAKI-2023

MÜNDƏRİCAT

GİRİŞ

Mövzunun aktuallığı. Təbii Dil Emalı (NLP) tədqiqatı irəlilədikcə və yeni tətbiqlər ortaya çıxdıqca, mətn məlumatlarının effektiv şəkildə idarə edilməsi və anlaşılması problemlərini həll edərək mətn təsnifatının əhəmiyyətinin artmağa davam edəcəyi gözlənilir. Sənədlərin sinifləndirilməsi üçün mətnlərin təsnifatı üsullarından istifadənin aktuallığı və təsiri hazırda əhəmiyyətlidir. Bir neçə mühüm element onun cari aktuallığını vurğulayır:

Mətn məlumatlarının səmərəli təşkili və strukturlaşdırılması rəqəmsal məzmunun eksponensial artımı səbəbindən getdikcə daha çox əhəmiyyət kəsb edən sənəd təsnifatı ilə mümkün olur. Sənədləri müvafiq siniflər və ya mövzular üzrə təsnif etməklə, məlumatı axtarmaq, əldə etmək və idarə etmək asanlaşır, məhsuldarlığın artmasına və iş axınının sadələşdirilməsinə səbəb olur.

Avtomatlaşdırılmış mətn təsnifatı sənədlərin kateqoriyalara bölünməsi prosesini asanlaşdırır və bununla da insanların iştirakı zərurətini minimuma endirir. Bu təcrübə, həm vaxta, həm də resurslara qənaətlə nəticələndiyi üçün əhəmiyyətli miqdarda sənədlərin tez-tez işlənməsinə ehtiyac olduğu hallarda xüsusilə faydalıdır. Avtomatlaşdırılmış sənədlərin təsnifatı hüquq, maliyyə və səhiyyə daxil olmaqla müxtəlif sənaye sahələri üçün dəyərli vasitədir. Bu, sənədlərin nəzərdən keçirilməsini, uyğunluq yoxlamalarını və məlumat axtarışını sürətləndirməyə kömək edir və bununla da bu sektorların səmərəliliyini artırır.

Sənədlərin təsnifatı istifadəçilərə təsnifat etiketlərinə əsasən xüsusi növ sənədləri süzgəcdən keçirməyə və əldə etməyə imkan verməklə axtarış imkanlarını artırır. Bu, axtarış nəticələrinin səmərəliliyini və dəqiqliyini əhəmiyyətli dərəcədə artırır, istifadəçilərə müvafiq məlumatları daha tez tapmağa imkan verir.

Sənədlərin təsnifatı üçün mətn təsnifat metodlarından istifadənin aktuallığı rəqəmsal məlumatların artan kəmiyyəti, effektiv məlumat idarəçiliyinə tələbat və mürəkkəb maşın öyrənmə modelləri və texnikalarının əlçatanlığı ilə izah olunur. Sənədlərin təsnifatı prosesi geniş mətn məlumatlarının artan istehsalı və toplanması səbəbindən müəssisələr və təşkilatlar üçün həlledici bir vasitə kimi ortaya çıxdı.

İşin əsas məqsədi. Dissertasiya işinin əsas məqsədi mətn sənədlərinin məzmununa görə əvvəlcədən müəyyən edilmiş kateqoriyalara və ya siniflərə avtomatlaşdırılmış təhlil və təsnifat aparmaqdır. Proses kateqoriyalara ayrılmış verilənlər toplusundan istifadə edərək maşın öyrənməsi alqoritminə təlimat verilməsini tələb edir, hər bir sənəd ayrı bir təsnifata təyin edilir. Təlim prosesi başa çatdıqdan sonra model yeni və müşahidə olunmamış sənədləri öz siniflərinə dəqiq şəkildə təsnif etmək imkanına malikdir.

İşin praktiki əhəmiyyəti. Mətn təsnifatının praktiki əhəmiyyəti onun çoxsaylı real tətbiqetmələrində və təklif etdiyi maddi faydalardadır. Mətnin təsnifatı nəhəng mətn məlumatlarını səmərəli şəkildə təşkil edir və tənzimləyir, axtarışı, əldə etməyi və idarə etməyi asanlaşdırır. Sənədləri mövzulara avtomatik təsnif etməklə məhsuldarlığı artırır. Avtomatlaşdırma, xüsusən bir çox sənədlərin işlənməsi zamanı vaxt və resurslara qənaət edir. Mətn təsnifatı istifadəçilərə sənədləri öz təsnifat etiketlərinə əsasən süzgəcdən keçirməyə və əldə etməyə imkan verir, məlumat axtarışını daha sürətli və daha dəqiq edir.

Tədqiqat işinin metodu və nəzəri əsasları. Python proqram dilindən istifadə edərək, mətnin təsnifatı işlənməmiş mətni təhlil üçün hazırlamaq üçün əvvəlcədən emaldan başlayır. Xüsusiyyətlərin çıxarılması mətni maşın öyrənmə alqoritmlərinin emal edə biləcəyi ədədi formata çevirir. Mətn təhlili metodlarına hər bir sənəddəki sözlərin tezliyini təmsil etmək üçün vektordan istifadə edən bagofwords strategiyası və sözlərin sənəddə və korpusda əhəmiyyətindən asılı olaraq çəki verən TF-IDF metodologiyası termini daxildir. Mətn daxiletməsini təsnif etmək üçün bir neçə maşın öyrənmə alqoritmi nümunələri və əlaqələri öyrənir. Bu üsullara Naive Bayes, dəstək vektor maşınları (SVM), təsadüfi meşələr və konvolyusiya neyron şəbəkələri (CNN) kimi dərin öyrənmə modelləri daxildir. Sənədlərə siniflərin təyin olunduğu etiketli məlumatlar mətn təsnifat modelini öyrədir. Model daha sonra parametrləri dəyişdirmək üçün accuracy, precision, recall, and F1-score kimi ölçülərdən istifadə etməklə qiymətləndirilir. Mətnin təsnifatı statistik nəticədən, ehtimaldan, məlumatdan və dilçilikdən istifadə edir. Ehtimal nəzəriyyəsi sənədin müşahidə olunan xassələri əsasında təsnifatını müəyyən etmək üçün istifadə olunur. Linqvistik nəzəriyyələr dilin

strukturunu, sintaksisini və semantik əlaqələrini anlamağa kömək edir, xüsusiyyət çıxarmağa və model yaratmağa imkan verir.

Tədqiqat işinin aprobasiyası və əməli reallaşdırılması. İşdə alınmış nəticələr "İnformasiya və telekommunikasiya texnologiyaları" kafedrasının seminarlarında məruzə və müzakirə olunmuş və "(Azərbaycanın Ümummilli Lideri Heydər Əliyevin anadan olmasının 100-cü il dönümünə həsr olunmuş <<Tətbiqi Riyaziyyatın müasir problemləri >> Respublika elmi konfransının materiallari XXI (23 May 2023-cü ildə) " çap olunmuşdur.

Dissertasiyanın həcmi və quruluşu. Dissertasiya işi girişdən, üç fəsildən, nəticə və ədəbiyyat siyahısından ibarətdir.

Dissertasiya işinin birinci fəsli üç yarımfəsildən ibarətdir.

Birinci yarımfəsildə Süni İntelekt və onun tarixi haqqında məlumat verilmişdir.

İkinci yarımfəsildə Süni İntellektin önəmi və istifadə sahələri haqqında məlumat verilmişdir.

Üçüncü yarımfəsildə Süni intellektin həyata keçirilməsi üçün strategiyalar verilmişdir.

Dissertasiya işinin ikinci fəsli üç yarımfəsildən ibarətdir.

Birinci yarımfəsildə təbii dil emalı (NLP) giriş haqqında ətraflı məlumatla təmin olunmuşdur.

İkinci yarımfəsildə mövzu təhlili, onun əhatə dairəsi, önəmi və işləmə prinsipi haqqında məlumatlar verilmişdir.

Üçüncü yarımfəsildə mövzu təhlilinin metodlarına toxunulmuşdur.

Dissertasiya işinin üçüncü fəsli üç yarımfəsildən ibarətdir.

Birinci yarımfəsildə mətinlərin kontekstə görə sinifləndirilməsi problemləri və onların tətbiqi istiqamətləri haqqında danışılmışdır.

İkinci yarımfəsildə mətinlərin kontekstə görə sinifləndirilməsinin işlənilmə instrumental vasitələri verilmişdir.

Üçüncü yarımfəsildə mətinlərin sinifləndirilməsinin prosesi təhlil olunmuşdur.

FƏSİL 1. SÜNİ İNTELLEKTƏ GİRİŞ

1.1 Süni İntelekt və onun tarixi

Milyarder Microsoft-un həmtəsisçisi Bill Qeyts süni intellektin gücü "o qədər inanılmazdır ki, cəmiyyəti çox dərin mənalarda dəyişdirəcək" dedi. Süni intellekt (AI) Con Makkarti tərəfindən tapılıb və o süni intellektin atası kimi tanınır. 1950-ci illərin ortalarında McCarthy "Süni intellekt" ifadəsini təqdim etdi və onun üçün "ağıllı maşınların yaradılması ilə əlaqəli tədqiqat sahəsi" kimi bir tərif verdi. Süni intellekt adətən intizamın həm elmi, həm də mühəndislik aspektlərini əhatə edən, intellektual maşınların inkişafı ilə əlaqəli təhsil və təcrübə sahəsi kimi istinad edilir. Süni intellekt insanın idrak və davranışını simulyasiya etmək üçün kompüter elmləri proqramlaşdırmasının tətbiqinə aiddir. Bu, məlumatların və ətraf mühit amillərinin təhlili, problemlərin həlli və müxtəlif tapşırıqlara effektiv şəkildə uyğunlaşmaq üçün biliklərin əldə edilməsi və ya öz-özünə öyrənmə yolu ilə əldə edilir. Süni intellekt müxtəlif gündəlik fəaliyyətləri asanlaşdırmaq üçün müntəzəm olaraq istifadə edildiyi üçün müasir cəmiyyətdə geniş yayılmışdır. Telefonunuzdakı Siri və Amazon-un Alexa-sı tapşırıqları yerinə yetirmək və sorğulara cavab vermək üçün, Facebook Feed fərdi istifadəçilərə görə ən uyğun məzmunu təxmin etmək və müəyyən etmək üçün, Netflix fərdi baxış vərdişlərinə əsaslanan fərdi tövsiyələr yaratmaq üçün süni intellekt texnologiyasından istifadə edir. Ümumiyyətlə, süni intellekt gündəlik təcrübələrimizi təkmilləşdirdi və müxtəlif tapşırıqların yerinə yetirilməsini və çoxlu sayda insan üçün məlumat istehlakını asanlaşdırdı.

Süni intellektin mənşəyi filosofların süni varlıqların, mexaniki insanların və avtomatların digər formalarının mövcudluğunun mümkünlüyünü düşündüyü qədim dövrlərə gedib çıxır. Süni intellektin yaranması öz mənşəyini erkən mütəfəkkirlərin qabaqcıl işlərinə borcludur, onların töhfələri 18-ci əsrdə və sonrakı dövrlərdə onun tədricən maddiləşməsinə kömək etmişdir. Ağıllı qeyri-insani maşınlar vasitəsilə insan təfəkküründün süni şəkildə mexanikləşdirilməsi və manipulyasiyası konsepsiyası filosoflar arasında fikir mövzusu olmuşdur. Süni intellektə marağı artıran düşüncə prosesləri klassik filosoflar,

riyaziyyatçılar və məntiqçilər simvolların manipulyasiyasını (mexaniki olaraq) nəzərdən keçirdikdə yaranıb və nəticədə 1940-cı illərdə proqramlaşdırıla bilən rəqəmsal kompüter Atanasoff Berry Kompüterinin (ABC) ixtirasına gətirib çıxarıb. Bu xüsusi yenilik tədqiqatçılar üçün "elektron beyin" və ya sintetik intellektual varlıq yaratmaq konsepsiyasını həyata keçirmək üçün katalizator rolunu oynadı. Süni intellekt nişanlarının sahənin hazırkı vəziyyətinin dərk edilməsinə töhfə verməsi təxminən on il çəkdi. Riyaziyyat və digər sahələrdə təcrübəsi olan bir polimath Alan Turinq, maşının insan hərəkətlərini insan davranışından hiss olunmayan səviyyəyə qədər təqlid etmək qabiliyyətini qiymətləndirən bir test təqdim etdi. Hər bir ardıcıl onilliyin gəlişi süni intellekt anlayışımızı əsaslı şəkildə dəyişdirən yeni kəşflər və texnoloji irəliləyişlərlə yadda qaldı. Bu tarixi irəliləyişlər süni intellektə sadəcə təxəyyül məhsulundan indiki və gələcək nəsillər üçün əlçatan olan konkret reallığa keçməyə imkan verib.

1.2 Süni İntellektin əhəmiyyəti və istifadə sahələri

Süni intellekt insan intellektini təqlid edən, kompüter əsaslı sistemlərə iterativ emal və alqoritmik təlim vasitəsilə bilik əldə etməyə imkan verən texnoloji irəliləyişdir. Süni intellekt sistemləri məlumat emalının hər bir uğurlu iterasiyası ilə təkmilləşdirilmiş intellekt səviyyəsini nümayiş etdirir. Bunun səbəbi, hər bir qarşılıqlı əlaqənin sistemə həlləri qiymətləndirməyə və kəmiyyətləndirməyə və verilən tapşırıqda bacarıq əldə etməyə imkan verir. Süni intellekt sistemləri insanlardan əhəmiyyətli dərəcədə daha sürətli təcrübə əldə etmək potensialına malikdir və onları ağıllı qərar qəbul etməyi tələb edən vəzifələr üçün yüksək səmərəli edir. Bu, bir insanın oxşar işi yerinə yetirə biləcəyi sürəti üstələyərək, bu cür tapşırıqları yüksək sürətlə yerinə yetirmək qabiliyyəti ilə bağlıdır. Süni intellekt insan beyninin əldə edə biləcəyindən əhəmiyyətli dərəcədə daha sürətli templərlə və daha böyük emal imkanları ilə olsa da, kompüterlərə insan düşüncə və davranışını təqlid etmək qabiliyyətinə görə çox güclü və dəyərli bir texnologiyadır. Süni intellekt sistemləri çoxsaylı

tətbiqlərdə insanlarla müqayisədə üstün performans nümayiş etdirərək, onları müasir iqtisadiyyatın mühüm komponentinə çevirib.

Süni intellekt texnologiyası onu demək olar ki, hər hansı bir müasir qurum üçün yüksək effektiv alət edən çoxsaylı mühüm üstünlüklər təqdim edir. Bu üstünlüklərə aşağıdakılar daxildir:

- Süni intellekt bir insandan fərqli olaraq, heç bir tükənmə və ya fasilə tələb etmədən əvvəllər əl ilə yerinə yetirilən monoton tapşırıqları avtomatlaşdırmaq qabiliyyətinə malikdir.
- Süni intellekt insanlarla müqayisədə məlumatların təhlilini əhəmiyyətli dərəcədə sürətləndirmək qabiliyyətinə malikdir. Bu, süni intellektə nümunələri daha yüksək effektivliklə aşkarlamağa imkan verir.
- Süni intellekt insanlardan daha böyük ölçülərə malik məlumat dəstlərini təhlil edə bilir və bununla da insan qavrayışından yayına bilən nümunələri müəyyən etməyə imkan verir.
- Süni intellekt məlumat toplamaq və təhlil etmək qabiliyyətindən istifadə edərək insan dəqiqliyini üstələmək potensialına malikdir.
- Süni intellektin istifadəsi mürəkkəb və çoxşaxəli əlaqələri effektiv şəkildə təhlil edərək məlumatların investisiya gəlirini artıra bilər. Bu texnologiya fasilələrə ehtiyac olmadan və minimal səhvlərlə davamlı olaraq fəaliyyət göstərir, bu da onu məlumatlardan asılı olan və geniş miqyasda fəaliyyət göstərən müəssisələr üçün mühüm alətə çevirir.
- Süni intellekt təşkilatlara strateji qərar qəbul etmə proseslərinin sürətini və dəqiqliyini artırmaqla öz əsas biznes proseslərini təkmilləşdirməyə imkan verir və bununla da qərarların qəbulu nəticələrinin yaxşılaşdırılmasına gətirib çıxarır.
- Süni intellektin gələcəyə potensial inqilabi təsirini əhatə edən diskurs, əsasən, süni intellekt həllərinin sənayenin geniş spektrində nəzərəçarpacaq müvəffəqiyyətlə

tətbiq olunması ilə bağlıdır. Süni intellektin effektivliyi bir neçə xüsusi tətbiqdə nümayiş etdirilmişdir, onlardan bir neçəsi aşağıda qeyd edilmişdir.

- Səhiyyə sənayesi, xəstələrə uyğunlaşdırılmış tibbi müalicəni idarə etmək üçün Süni intellekt tətbiqlərini tətbiq etdi. Bu AI sistemləri dərman xatırlatmaları təklif edir və hansı xüsusi məşqləri yerinə yetirməli olduqları barədə təkliflər verir.
- İnventarlaşdırma, optimallaşdırılmış mağaza planlarını idarə etmək və Amazon-un "Siz də bəyənə bilərsiniz" funksiyası və Netflix-in maşın öyrənməsinə əsaslanan alqoritmi vasitəsilə fərdiləşdirilmiş baxış tövsiyələri vasitəsilə xüsusi alış-veriş tövsiyələri təklif etmək üçün AI texnologiyası pərakəndə satış parametrlərində istifadə olunur.
- İstehsal sahəsində, fabriklər üçün yük və tələbatı proqnozlaşdırmaq üçün Süni intellekt həlləri istifadə olunur. Bu, logistika, materialların sifarişi və layihənin tamamlanma müddətləri ilə bağlı əsaslandırılmış qərarlar qəbul etməyə imkan verməklə fabriklərin səmərəli idarə edilməsini asanlaşdırır.
- Bank sənayesində süni intellektin tətbiqi saxta maliyyə əməliyyatlarının aşkarlanmasını asanlaşdırıb, kredit xallarının qiymətləndirilməsinin dəqiqliyini artırıb və məlumatların əl ilə daxil edilməsini və idarə olunmasını tələb edən avtomatlaşdırılmış tapşırıqları yerinə yetirib.
- Həyat Elmləri - Süni intellekt texnologiyası yeni dərmanları sınaqdan keçirmək üçün tətbiq edilir ki, bu da təşkilatlara onları daha tez bazara çıxarmağa və yeni, daha effektiv müalicə üsullarını və əczaçılıq dərmanlarını kəşf etməyə kömək edən böyük və mürəkkəb məlumat dəstlərini təhlil etməyə imkan verir.
- Şübhəsiz ki, süni intellekt saysız-hesabsız əhəmiyyətli prosedurlarla həyata keçirilib. Bununla belə, müasir iqtisadiyyatın demək olar ki, hər bir sahəsində AI inteqrasiyası üçün çoxlu istifadə edilməmiş imkanlar mövcuddur.

1.3 Süni intellektin həyata keçirilməsi üçün strategiyalar

Süni intellekt sahəsi hazırda kompüter elmləri sahəsində əhəmiyyətli dərəcədə populyarlıq qazanır. Buna baxmayaraq, yeni texnologiyaların və tədqiqatların yayılması səbəbindən sahə sürətlə genişlənir və bununla da müxtəlif anlayışlar arasındakı fərqləri ayırd etmək çətinləşir. Üstəlik, Süni İntellekt hər biri özünəməxsus alqoritmlər dəsti ilə müxtəlif sahələri əhatə edir. Beləliklə, AI-nin tək bir intizam deyil, müxtəlif sahələrin birləşməsi olduğunu qəbul etmək vacibdir. Süni intellekt kompüterlərin insanlar tərəfindən edildiyi təqdirdə intellekt tələb edən işləri yerinə yetirə bilməsi üçün ümumi termindir. Süni intellekt iki əsas sahəyə bölünə bilər: Maşın Öyrənməsi (ML) və Neyron Şəbəkələr (NN). Bu domenlərin hər ikisi Süni İntellektin çətiri altına düşür və hər birinin problemlərin həllinə kömək edən öz fərqli texnika və alqoritmləri var. Dərin Öyrənmə ən daxili dairədir. Maşın Öyrənmə sahəsi yuxarıda qeyd olunan sahədən kənarda yerləşir, qalanları əhatə edən ən geniş sahə isə Süni İntellektə aiddir.

Maşın Öyrənmə

Maşın öyrənməsi süni intellektin alt sahəsidir. Maşın öyrənmənin məqsədi ümumiyyətlə məlumatların strukturunu anlamaq və bu məlumatları insanlar tərəfindən başa düşülə və istifadə edilə bilən modellərə uyğunlaşdırmaqdır. Maşın öyrənməsi, kompüter elminin alt sahəsi olmasına baxmayaraq, ənənəvi hesablama metodologiyalarından kənara çıxır. Ənənəvi hesablamada alqoritmlər hesablamaq və ya problemi həll etmək üçün kompüterlər tərəfindən istifadə olunan açıq şəkildə proqramlaşdırılmış təlimatlar toplusudur. Maşın öyrənmə alqoritmləri kompüterlərə verilənlərin daxil edilməsi üzrə təlim keçməyə və əvvəlcədən müəyyən edilmiş diapazonda məhdud olan çıxışları yaratmaq üçün statistik analizdən istifadə etməyə imkan verir. Maşın öyrənməsi kompüterlərə nümunə verilənlərdən modellər qurmağa imkan verir və bununla da məlumat daxilolmalarına əsaslanan qərar qəbuletmə prosedurlarını avtomatlaşdırır.

Hal-hazırda, hesablama prosesləri üçün maşın öyrənmə metodologiyalarından istifadə edən avtonom nəqliyyat vasitələri, dilin tərcüməsi alqoritmləri və üz

identifikasiyası sistemləri kimi diqqətəlayiq tətbiqlər mövcuddur. Bu fenomenin mənşəyini 1943-cü ildə, neyrofizioloq Uorren Makkulloch və riyaziyyatçı Uolter Pitsin neyronların funksiyasını işıqlandıran bir kağız üzərində əməkdaşlıq etdiyi vaxta qədər görmək olar. Neyron şəbəkənin başlanğıcını elektrik sxemlərindən istifadə edən bir modelin inkişafı ilə izləmək olar. 1950-ci ildə Alan Turinq tərəfindən kompüterlərdə həqiqi zəkanın mövcudluğunu müəyyən etmək üçün məşhur qiymətləndirmə metodu olan Turing Testi sınaqdan keçirildi. Testdən uğurla keçmək üçün sistem insana bənzər davranışı ardıcıl surətdə simulyasiya etməli və bununla da insanı kompüter deyil, insan olduğuna inandırmalıdır. 1952-ci ildə Artur Samuel dama oyunu ilə məşğul olarkən bilik əldə edə bilən ilkin kompüter proqramını yaratdı. 1957-ci ildə Frank Rosenblatt ilkin neyron şəbəkəsini inkişaf etdirdi və onu perseptron adlandırdı. 1990-cı illərdə biliyə əsaslanan yanaşmadan verilənlərə əsaslanan metodologiyaya keçid zamanı maşın öyrənməsi sahəsində əhəmiyyətli transformasiya baş verdi. Bu dəyişiklik, əsasən , bu müddət ərzində asanlıqla əldə edilə bilən böyük miqdarda məlumatların bolluğu ilə əlaqələndirildi. 1997-ci ildə IBM-in Deep Blue şirkəti şahmat oyununda dünya çempionunu məğlub edərək mühüm bir mərhələyə çatdı və ilk dəfə bir maşın belə bir uğura imza atdı. Müəssisələr tərəfindən etiraf edilmişdir ki, mürəkkəb hesablamalar üçün imkanlar maşın öyrənməsi vasitəsilə artırıla bilər. Bu yaxınlarda həyata keçirilən tədbirlərdən biri 2012-ci ildə yaradılmış Google Brain-ə aiddir. O, şəkillər və videolardakı nümunələrin identifikasiyası üzərində cəmləşən dərin neyron şəbəkəsini təşkil edir. Daha sonra, YouTube platformasına yüklənmiş videolardakı obyektləri müəyyən etmək üçün istifadə edilmişdir. 2014-cü ildə Facebook insan kimi tanıma imkanlarını təqlid edən qabaqcıl üz tanıma texnologiyası olan Deep Face-i inkişaf etdirdi. 2014-cü ildə Deep Mind Alpha Go adlı kompüter alqoritmini işləyib hazırlayıb və bu alqoritm stolüstü oyunda bacarıqlı Go oyunçusuna qalib gəlib. Oyun mürəkkəb təbiətinə görə süni intellekt üçün klassik oyun kimi qəbul edilir və bu, onu çox çətinləşdirir. Tanınmış alimlər Stiven Hokinq və Stüart Rasselin fikirlərinə görə, süni intellektin sürətlənən sürətlə özünü çoxalma qabiliyyətinin yaratdığı “kəşfiyyat partlayışı”

ehtimalı bəşər övladının yox olması ilə nəticələnə bilər. Elon Musk süni intellektin bəşəriyyət üçün mühüm ekzistensial təhlükə olduğunu müəyyən edib. Buna cavab olaraq o, 2015-ci ildə həm təhlükəsiz, həm də bəşəriyyət üçün faydalı süni intellektin inkişafına həsr olunmuş Open AI təşkilatını qurdu. Müasir dövrdə süni intellekt sahəsində əhəmiyyətli irəliləyişlərin şahidi olan bir neçə sahəyə Kompüter Görməsi (CV), Təbii Dil Emalı (NLP) və Gücləndirici Öyrənmə (RL) daxildir.

Maşın öyrənmə sahəsi davamlı irəliləyişlər və tərəqqi ilə xarakterizə olunur. Maşın öyrənmə metodologiyaları ilə məşğul olarkən və ya maşın öyrənmə prosedurlarının nəticələrini araşdırarkən müəyyən mülahizələri nəzərə almaq vacibdir. Tapşırıqların təsnifatı maşın öyrənməsi sahəsində ümumi bir təcrübədir. Yuxarıda qeyd olunan kateqoriyalar öyrənmənin necə əldə olunduğuna və ya işlənmiş sistemə əks əlaqənin verildiyi metodologiyaya əsaslanır. Nəzarət olunan öyrənmə və nəzarətsiz öyrənmə iki görkəmli maşın öyrənmə texnikasıdır.

Nəzarət olunan öyrənmə kompüterə etiketlənmiş nümunə daxiletmələrin və onların müvafiq arzu olunan nəticələrinin təqdim edilməsini nəzərdə tutur. Bu yanaşmanın məqsədi uyğunsuzluqları aşkar etmək və sonradan modeli tənzimləmək üçün onun faktiki çıxışını göstəriş verilmiş nəticələrlə müqayisə edərək alqoritmin bilik əldə etməsini asanlaşdırmaqdır. Nəzarət edilən öyrənmə, etiketlənməmiş məlumatlar üzərində etiket dəyərləri haqqında proqnozlar vermək üçün nümunələrdən istifadə edir. Nəzarət olunan öyrənmənin geniş yayılmış tətbiqlərindən biri statistik vasitələrlə ehtimal olunan gələcək hadisələri proqnozlaşdırmaq üçün keçmiş məlumatlardan istifadə etməyi əhatə edir. Tarixi fond bazarı məlumatları gələcək bazar dalğalanmalarını proqnozlaşdırmaq üçün istifadə oluna bilər, spam e-poçtları isə bu texnologiyadan istifadə etməklə effektiv şəkildə yoxlanıla bilər.

Nəzarətsiz öyrənmə, etiketlənməmiş məlumatların istifadə edilməsi, bununla da öyrənmə alqoritminə giriş verilənləri daxilində nümunələri və oxşarlıqları müəyyən etmək tapşırığı verilir. Nəzarətsiz öyrənmə maşın öyrənməsində dəyərli bir yanaşmadır, çünki o,

etiketli məlumatlardan daha çox yayılmış çoxlu etiketsiz məlumatlardan istifadə etməyə imkan verir. Nəzarətsiz öyrənmənin məqsədi verilmiş verilənlər toplusunda gizli nümunələri aşkar etmək qədər sadə ola bilər və ya hesablama sisteminə işlənməmiş məlumatların təsnifatı üçün lazımi təsvirləri avtonom şəkildə müəyyən etməyə imkan verən xüsusiyyət öyrənmə məqsədini əhatə edə bilər**.** Nəzarətsiz öyrənmə üsulları mürəkkəb və zahirən fərqli görünən məlumat dəstlərini potensial mənalı şəkildə təşkil etmək məqsədi ilə əvvəlcədən müəyyən edilmiş həll yolu olmadan təhlil etmək qabiliyyətinə malikdir.

Müvafiq alqoritmin seçilməsi prosesi, hər birində fərqli öyrənmə metodologiyalarından istifadə edən nəzarət edilən və nəzarətsiz maşın öyrənməsi alqoritmlərinin çoxluğunu nəzərə alsaq, çətin ola bilər. Bütün vəziyyətlərə tətbiq oluna bilən universal optimal yanaşma və ya həll yolu yoxdur. Müvafiq alqoritmin müəyyən edilməsi prosesi müəyyən dərəcədə sınaqdan keçir, onun effektivliyini müəyyən etmək hətta təcrübəli məlumat alimlərinin empirik testlər keçirmədən yerinə yetirə bilməyəcəyi bir vəzifədir. Alqoritmin seçimi nəzərdən keçirilən məlumatların miqyası və təbiəti, verilənlərdən əldə edilməsi arzu olunan nəticələr və belə nəticələrin nəzərdə tutulan tətbiqi kimi müxtəlif amillərdən asılıdır. Nəzarət olunan və nəzarətsiz maşın öyrənməsi arasında seçim etmək üçün bəzi qaydalar var. Nəzarətli öyrənmə, davamlı dəyişənin gələcək dəyərinin proqnozlaşdırılma tapşırıqlarını yerinə yetirmək üçün modeli öyrətməyə tövsiyə olunan yanaşmadır. Əgər məqsəd verilənləri araşdırmaq və verilənləri klasterlərə bölmək kimi effektiv daxili təmsilçilik yarada bilən model hazırlamaq olduqda nəzarətsiz öyrənməyə üstünlük verilir.

Neyron Şəbəkələri

Neyron şəbəkələri insan sinir sisteminin neyronlarının mürəkkəb əlaqələrini təkrarlamaq cəhdi ilə hazırlanmışdır. Düşünüldü ki, bioloji sinir sistemi siqnalları ötürməkdə və emal etməkdə çox səmərəli olduğundan, maşınlar üçün insana bənzər intellekt yaratmağa kömək edə bilər. Nəticədə, insan beynində bir qrup neyron kimi məlumatların işlənməsi və ötürülməsi qabiliyyətinə malik olan sintetik neyronlardan ibarət

şəbəkə yaradılmışdır. Neyron şəbəkələrinin yaranması, maşınların ağıllı şəkildə öyrənməsi və cavab verməsi üçün böyük imkanlar təmin etdi. Neyron şəbəkələri və ya süni neyron şəbəkələri (ANN) süni intellektin (AI) və maşın öyrənməsinin (ML) bir hissəsidir və maşınlara/kompüterlərə bioloji beyin kimi məlumatları emal etməyi öyrədir. Şəbəkə əvvəlki fəaliyyətlərindən öyrənməyə və təkmilləşməyə davam etməyə imkan verən adaptiv sistemə malikdir. Neyron şəbəkələri maşın öyrənməsinin bir hissəsini təşkil edir və onların əsas arxitekturası dərin öyrənmə alqoritmlərindən istifadə etməklə qurulur. "Neyron şəbəkəsi" adı insan beynindəki neyronların mürəkkəb şəbəkəsindən və neyronların necə əlaqə saxlamasından ilhamlanıb. Neyron şəbəkə öyrənməni dəstəkləmək və bacarıqlarını artırmaq üçün təlim məlumatlarından giriş kimi istifadə edir. Sözügedən alət tarixi məlumatlardan ardıcıl olaraq öyrənmə qabiliyyətinə görə güclü və müasir hesab edilir və nəticədə yüksək dəqiqlik əldə edilir. Neyron şəbəkələrin mənşəyini hesablamaların ilk günlərinə qədər izləmək olar. Warren McCulloch insan beyninin funksionallığını simulyasiya etmək qabiliyyətinə malik dövrə əsaslı sistem kimi ilkin neyron şəbəkəsini inkişaf etdirdi. 1958-ci ildə süni qavrayışın ilk nümunəsi Frank Rosenblatt tərəfindən hazırlanmışdır. 1982-ci ildə John Hopfield tərəfindən "təkrarlanan neyron şəbəkələri" mövzusunda bir məqalə nəşr olundu. Zülal tədqiqatları sahəsində neyron şəbəkələri 1988-ci ildə geniş şəkildə istifadə edildi. Texnologiya zülalların üçölçülü formalarını proqnozlaşdırmaq üçün istifadə edildi. 1992-ci ildə üçölçülü obyektlərin tanınması üçün alqoritm hazırlanmışdır. Hazırda neyron şəbəkələri çox inkişaf etmişdir. Onlar səhiyyə, aerokosmik və müdafiədən tutmuş kibertəhlükəsizlik, marketinq və hava proqnozlarına qədər bir çox sektorda istifadə olunur. Yuxarıda izah edildiyi kimi, neyron şəbəkəsinin inkişafı neyron memarlığı baxımından insan beynindən ilhamlanıb. İnsan beyninin neyronları siqnalların göndərildiyi və məlumatların işləndiyi mürəkkəb və yüksək dərəcədə əlaqəli şəbəkə yarada bilər. Bu, neyron şəbəkələri tərəfindən təkrarlanan neyronların funksiyası kimi çıxış edir. Neyron şəbəkələrinin işləməsinin əsas üsulu şəbəkə daxilində çoxsaylı və müxtəlif neyron təbəqələrinin qarşılıqlı əlaqəsidir. Neyronlar sinaptik qovşaq

vasitəsilə bir-birinə bağlıdır. Sözügedən təbəqə özündən əvvəlki təbəqədən girişi qəbul edə və çıxışı sonrakı təbəqəyə ötürə bilir. Bu addım son təbəqə tərəfindən qərar və ya proqnoz verilənə qədər təkrarlanmağa davam edir. Neyron şəbəkənin işini məlumatların keçdiyi və işləndiyi şəbəkənin hər bir təbəqəsinin fərdi mexanizmləri baxımından daha yaxşı başa düşmək olar. Əsas strukturda üç təbəqə var - giriş, gizli və çıxış. Neyron şəbəkəsinin giriş təbəqəsi xarici mühitdən məlumatların toplanması və qəbulu ilə bağlı məsuliyyət daşıyır. Məlumatların toplanması başa çatdıqdan sonra, təbəqə məlumatların identifikasiyasını artırmaq üçün məlumatların işlənməsini, məzmunun təhlilini və təsnifatını həyata keçirir. Daha sonra məlumat sonrakı təbəqəyə ötürülür. Gizli təbəqə həm giriş qatından, həm də digər gizli təbəqələrdən məlumat alır. Neyron şəbəkəsi əhəmiyyətli miqdarda gizli təbəqələri ehtiva etmək potensialına malikdir. Gizli təbəqələrin hər biri əvvəlki təbəqədən ötürülən girişi yoxlamaq qabiliyyətinə malikdir. Sonradan giriş emaldan keçir və sonradan sonrakı mərhələlərə ötürülür. Əvvəlki gizli təbəqədən ötürülən məlumat son çıxış qatına ötürülür. Neyron şəbəkəsinin ən yüksək təbəqəsi, əsas təbəqələrdə həyata keçirilən məlumatların işlənməsi nəticəsində əldə edilən son nəticəni nümayiş etdirir. Çıxış qatındakı qovşaqların sayı girişdən asılıdır. 1/0 və ya Bəli/Xeyr kimi ikili verilənləri əhatə edən hallarda, tək çıxış qovşağından istifadə ediləcək. Bununla belə, bir çox kateqoriyalı məlumatların idarə edilməsi kontekstində çoxlu qovşaqlardan istifadə etmək lazımdır. Gizli təbəqə dərin öyrənmə (DL) sistemi daxilində qarşılıqlı asılı qovşaqların mürəkkəb şəbəkəsi kimi müəyyən edilə bilər. Düyünlər arasındakı əlaqəni ifadə etmək üçün "çəki" kimi tanınan ədədi dəyər istifadə olunur. Bəyanat verilmiş qovşağın digər qovşaqlara nə dərəcədə təsir göstərmək potensialına malik olduğunu bildirir.

Dərin Öyrənmə

Dərin öyrənmə üç və ya daha çox qatlı neyron şəbəkələrin istifadəsini nəzərdə tutan maşın öyrənməsinin alt sahəsidir. Bu neyroşəbəkələr insan beyninin davranışını təqlid etməyə çalışır, baxmayaraq ki, onun qabiliyyətinə uyğun gəlməsə də, ona böyük həcmdə məlumatlardan “öyrənməyə” imkan verir. Tək qatlı neyron şəbəkə hələ də təxmini

proqnozlar verə bilsə də, əlavə gizli təbəqələr optimallaşdırmağa və dəqiqlik üçün dəqiqləşdirməyə kömək edə bilər. Dərin öyrənmə avtomatlaşdırmanı təkmilləşdirən, insan müdaxiləsi olmadan analitik və fiziki tapşırıqları yerinə yetirən bir çox Süni intellekt tətbiq və xidmətlərini idarə edir. Dərin öyrənmə texnologiyası gündəlik məhsul və xidmətlərin (məsələn, rəqəmsal köməkçilər, səslə işləyən televizor pultları və kredit kartı fırıldaqlarının aşkarlanması), eləcə də inkişaf etməkdə olan texnologiyaların (özünü idarə edən avtomobillər kimi) arxasında dayanır. Maşın öyrənmə alqoritmləri proqnozlar vermək üçün strukturlaşdırılmış, etiketlənmiş məlumatlardan istifadə edir, yəni xüsusi funksiyalar model üçün daxil edilən məlumatlardan müəyyən edilir və cədvəllər şəklində təşkil edilir. Bu bəyanat, strukturlaşdırılmamış məlumatların istifadəsinin qarşısının alınmadığını nəzərdə tutur; Əksinə, strukturlaşdırılmış konfiqurasiyaya çevrilməsini asanlaşdırmaq üçün adətən ilkin emala məruz qalır. Dərin öyrənmə, şərti olaraq maşın öyrənməsi ilə əlaqəli olan məlumatların əvvəlcədən emalının müəyyən aspektlərini aradan qaldırır. Alqoritmlər mətn və şəkillər kimi strukturlaşdırılmamış məlumatları mənimsəmək və təhlil etmək qabiliyyətinə malikdir, beləliklə, xüsusiyyətlərin çıxarılması prosesini avtomatlaşdırır və insan ekspertlərinə etibarı azaldır. Nümunə olaraq, fərz edək ki, biz müxtəlif əhliləşdirilmiş heyvanları əks etdirən fotoşəkillər kolleksiyasına malikik və onları "pişik", "köpək", "gəmirici" və s. kimi müvafiq taksonomik kateqoriyalara görə təsnif etməyi hədəfləmişik. Dərin öyrənmə alqoritmləri müxtəlif heyvanları ayırd etmək üçün qulaqlar kimi ən əhəmiyyətli xüsusiyyətləri ayırd etmək qabiliyyətinə malikdir. Maşın öyrənməsində bu xüsusiyyətlər iyerarxiyası insan mütəxəssisi tərəfindən əl ilə qurulur. Daha sonra, dərin öyrənmə modeli dəqiqliyini artırmaq üçün özünü dəqiq tənzimləmək və uyğunlaşdırmaq üçün gradient eniş və geri yayılma prosedurlarından keçir və bununla da ona yeni heyvan fotoşəkili ilə bağlı daha dəqiq proqnozlar yaratmağa imkan verir. Maşın öyrənməsi və dərin öyrənmə modelləri, adətən fərqli kateqoriyalar kimi təsnif edilən nəzarətli öyrənmə, nəzarətsiz öyrənmə və gücləndirici öyrənmə kimi müxtəlif öyrənmə formaları ilə məşğul olmaq qabiliyyətinə malikdir. Nəzarət edilən öyrənmə kateqoriyalara ayırmaq və ya

proqnozlar vermək üçün etiketlənmiş məlumat dəstlərindən istifadə edir və bu, giriş məlumatlarını düzgün etiketləmək üçün bir növ insan müdaxiləsini tələb edir. Nəzarət olunan öyrənmədən fərqli olaraq, nəzarətsiz öyrənmə etiketli verilənlər toplusunu tələb etmir. Əksinə, o, verilənlər daxilində nümunələri müəyyən edir və onları hər hansı nəzərə çarpan xüsusiyyətlərə əsasən qruplaşdırır. Gücləndirici öyrənmə, agentin mükafatlar şəklində rəy almaqla müəyyən bir mühitdə performansını yaxşılaşdırmağı öyrənməsini əhatə edən hesablama yanaşmasıdır. Agentin məqsədi gözlənilən ən yüksək mükafata səbəb olan hərəkətləri seçməklə zamanla məcmu mükafatı maksimuma çatdırmaqdır.

Dərin öyrənmə neyron şəbəkələri və ya süni neyron şəbəkələri məlumat daxiletmələri, çəkilər və qərəzliliyin birləşməsi vasitəsilə insan beynini təqlid etməyə çalışır. Bu elementlər verilənlər daxilindəki obyektləri dəqiq tanımaq, təsnif etmək və təsvir etmək üçün birlikdə işləyir. Dərin neyron şəbəkələri bir-biri ilə əlaqəli qovşaqların çoxsaylı qatlarından ibarətdir, hər biri proqnozu və ya təsnifatı dəqiqləşdirmək və optimallaşdırmaq üçün əvvəlki təbəqə üzərində qurulur. Şəbəkə vasitəsilə məlumatların ardıcıl şəkildə ötürülməsi prosesinə irəli yayılma deyilir. Dərin neyron şəbəkəsinin giriş və çıxış təbəqələrinə görünən təbəqələr deyilir. Dərin öyrənmə modelinin giriş təbəqəsi məlumatların qəbulu və işlənməsi üçün cavabdehdir, çıxış təbəqəsi isə son proqnoz və ya təsnifatın yaradılmasına cavabdehdir. Geri yayılma adlanan başqa bir proses proqnozlardakı səhvləri hesablamaq üçün gradient eniş kimi alqoritmlərdən istifadə edir və sonra modeli öyrətmək üçün təbəqələr arasında geriyə doğru hərəkət edərək funksiyanın çəkilərini və meyllərini tənzimləyir. Birlikdə irəli yayılma və geri yayılma neyron şəbəkəyə proqnozlar verməyə və istənilən səhvləri müvafiq olaraq düzəltməyə imkan verir. Zaman keçdikcə alqoritm getdikcə daha dəqiq olur. Yuxarıda qeyd olunan keçid dərin neyron şəbəkəsinin ən elementar formasını sadə şəkildə təsvir edir. Dərin öyrənmə alqoritmləri yüksək səviyyəli mürəkkəbliyi ilə xarakterizə olunur və onlar xüsusi verilənlər bazası və ya problemləri həll etmək üçün müxtəlif neyron şəbəkə arxitekturalarını əhatə edir. Dərin öyrənmə alqoritmlərinə misal:

- Konvolutional neyron şəbəkələri (CNN) əsasən kompüter görmə və təsvirin təsnifatı sahələrində istifadə olunur. Onlar təsvir daxilində xüsusiyyətləri və nümunələri müəyyən etmək qabiliyyətinə malikdirlər və bununla da obyektin aşkarlanması və ya tanınması kimi vəzifələri asanlaşdırırlar. 2015-ci ildə konvolutional neyron şəbəkəsi (CNN) süni intellekt sahəsində əhəmiyyətli bir mərhələni qeyd edərək, obyektin tanınması ilə bağlı problemdə insanı üstələyib.
- Təkrarlanan neyron şəbəkələri (RNN) ardıcıl və ya zaman seriyası məlumatlarından səmərəli istifadə etmək qabiliyyətinə görə təbii dilin işlənməsi və nitqin tanınması ilə bağlı tətbiqlərdə geniş istifadə olunur.
- Həqiqi dünya ssenarilərində tətbiq oluna bilən dərin öyrənmə proqramları gündəlik həyatımızda hər yerdə geniş yayılmışdır. Bununla belə, bir çox hallarda bu proqramlar məhsul və xidmətlərə mükəmməl inteqrasiya olunur və istifadəçiləri arxa planda baş verən mürəkkəb məlumatların işlənməsindən xəbərsiz edir. Dərin öyrənmə alqoritmləri tranzaksiya məlumatlarından bilikləri yoxlamaq və əldə etmək qabiliyyətinə malikdir və bununla da potensial saxtakarlıq və ya cinayət davranışını ifadə edən təhlükəli nümunələri aşkar edir. Nitqin tanınması və kompüter görmə kimi dərin öyrənmə proqramlarından istifadə tədqiqat təhlilinin səmərəliliyini və effektivliyini artıra bilər. Buna səs və video yazıları, şəkillər və sənədlər kimi müxtəlif mənbələrdən nümunələr və sübutlar çıxarmaqla nail olunur. Bu cür texnologiyanın tətbiqi hüquq-mühafizə orqanlarına böyük həcmli məlumatların daha yüksək sürət və dəqiqliklə təhlilinə kömək edir. Maliyyə institutları səhmlərin alqoritmik ticarətini aparmaq, kreditlərin təsdiqlənməsi üçün biznes risklərini qiymətləndirmək, saxtakarlığı aşkar etmək və müştərilər üçün kredit və investisiya portfellərini idarə etməyə kömək etmək üçün mütəmadi olaraq proqnozlaşdırıcı analitikadan istifadə edirlər. Dərin öyrənmə imkanlarının səhiyyə sənayesinə inteqrasiyası xüsusilə xəstəxana qeydlərinin və şəkillərinin rəqəmsallaşdırılmasından sonra sərfəli olduğunu sübut etdi. Təsvirin tanınması üçün

tətbiqlər tibbi görüntüləmə mütəxəssislərinə və radioloqlara daha çox sayda təsviri daha qısa müddət ərzində təhlil etmək və qiymətləndirmək səylərində kömək etmək potensialına malikdir.

FƏSIL 2. MÖVZÜ TƏHLILI VƏ TƏBII DIL EMALI

2.1. Təbii Dil Emalı

Natural language processing (NLP) süni intellekt sahəsidir ki, burada kompüterlər insan dilindən ağıllı və faydalı şəkildə təhlil edir, başa düşür və məna çıxarır. Təbii dil emalından istifadə etməklə tərtibatçılar avtomatik ümumiləşdirmə, tərcümə, adlandırılmış obyektin tanınması, əlaqələrin çıxarılması, hisslərin təhlili, nitqin tanınması və mövzunun seqmentasiyası kimi vəzifələri yerinə yetirmək üçün bilikləri təşkil edə və strukturlaşdıra bilər. Təbii dil emalı avtomatlaşdırılmış sistemlərə insan dil nümunələrini dərk etməyə imkan verən mətn məlumatlarını yoxlamaq üçün istifadə edilən hesablama texnikasıdır. İnsanlar və kompüterlər arasında qarşılıqlı əlaqə avtomatlaşdırılmış mətnin ümumiləşdirilməsi, əhval-ruhiyyənin təhlili, mövzunun müəyyənləşdirilməsi, adlandırılmış obyektin tanınması, nitq hissələrinin etiketlənməsi, əlaqələrin çıxarılması, köklənmə və digər əlaqəli tapşırıqlar kimi praktik tətbiqlərin həyata keçirilməsini asanlaşdırır. Təbii dil emalı tez-tez mətn istehsalı, maşın tərcüməsi və avtomatlaşdırılmış suallara cavab kimi müxtəlif tətbiqlərdə istifadə olunur. Təbii dil emalı adətən kompüter elmləri sahəsində çətin bir problem kimi qəbul edilir. Ona görə ki, insan dilinin dəqiqliyi və aydınlığı çox vaxt məhdud olur və insan dilini dərk etmək təkcə leksik vahidlərin deşifrəsini deyil, həm də əsas ideyaları və onların əhəmiyyət kəsb edən qarşılıqlı əlaqələrini dərk etməyi də nəzərdə tutur. İnsan zehni, dili nisbi rahatlıqla mənimsəməyə qadirdir, lakin dilin özünəməxsus qeyri-müəyyənliyi təbii dilin emalında bacarıq əldə etməyə çalışan kompüterlər üçün əhəmiyyətli problem yaradır. Təbii dil emalı alqoritmləri müxtəlif tətbiq sahələrinə malikdir. Əslində, bu alətlər təbii dili anlamaq qabiliyyətinə malik proqramlarının işlənib hazırlanmasını asanlaşdırır və tərtibatçılara və müəssisələrə insan dilini başa düşən proqram təminatı yaratmağa imkan verir. İnsan dilinə xas olan incəliklər təbii dilin mənimsənilməsini və düzgün həyata keçirilməsini çətin bir cəhdə çevirir. Təbii dil emalı alqoritmləri adətən maşın öyrənmə alqoritmləri üzərində qurulur. Təbii dil emalı geniş qayda dəstlərinin əl ilə kodlaşdırılmasına etibar etmək əvəzinə, qaydalar toplusunu

avtomatik əldə etmək üçün maşın öyrənmə üsullarından istifadə edə bilər. Buna kitab və ya cümlələr toplusu kimi böyük bir korpusun təhlili və bir sıra nümunələr əsasında statistik nəticələr çıxarmaqla nail olunur. Geniş şəkildə desək, təhlilə məruz qalan məlumatların miqdarı nə qədər çox olarsa, modelin əldə edə biləcəyi dəqiqlik səviyyəsi bir o qədər yüksək olar. Təbii dilin başa düşülməsi ilk növbədə iki fundamental texnikanın, yəni sintaksis kimi tanınan sintaktik analiz və semantika kimi tanınan semantik təhlilin istifadəsi ilə əldə edilir. Dil etibarlı sayılan cümlələr toplusu kimi müəyyən edilə bilər, lakin cümlənin etibarlılığını müəyyən edən meyarlar sintaksis və semantikadır. Sintaksis anlayışı dilin qrammatik düzülüşünə aiddir, semantika isə istifadə olunan dil tərəfindən çatdırılan mənanın şərhinə aiddir. Cümlə qrammatik cəhətdən düzgün olsa da, semantik cəhətdən həmişə düzgün olmaya bilər. Nümunə olaraq, "inəklər yüksək səviyyədə axır" ifadəsi sintaktik cəhətdən düzgündür (subyekt-zərf-fel), lakin uyğunluq yoxdur. Formal qrammatika prinsiplərindən istifadə edərək təbii dilin tədqiqi prosesi adətən sintaktik təhlil, sintaksis təhlili və ya təhlil kimi tanınır. Qrammatik qaydaların tətbiqi tək-tək sözlərə deyil, sözlərin çoxluğuna və təsnifatına aiddir. Sintaktik təhlil prosesi verilmiş mətnə semantik strukturun təyin edilməsini nəzərdə tutur. Semantik təhlil leksemlərin, simvolların və sintaktik düzülüşün konnotasiyasını və mənasını dərk etmək proseduruna aiddir. Bu, kompüterlərə təbii dili insanlar kimi qismən başa düşməyə imkan verir. Bu iddia semantik təhlilin təbii dil emalının ən çətin aspektlərindən biri olaraq qalması və onun həllinə hələ tam nail olunmaması fonunda irəli sürülür.

2.2. Mövzu təhlili, onun əhatə dairəsi, önəmi və işləmə prinsipi

Mövzu təhlili

Mövzunun aşkarlanması, mövzunun modelləşdirilməsi və ya mövzunun çıxarılması kimi də tanınan mövzu təhlili böyük həcmdə mətn məlumatlarının təşkilini və başa düşülməsini asanlaşdıran maşın öyrənmə metodologiyasıdır.

Mövzu təhlili verilmiş mətndən və ya mətnlər toplusundan əsas mövzuların və ya mövzuların müəyyən edilməsi və çıxarılması prosesinə aiddir. O, mətnin məzmununu təhlil etmək və müxtəlif söz və ifadələr arasında qanunauyğunluqları və əlaqələri müəyyən etmək üçün müxtəlif hesablama və statistik üsullardan istifadəni nəzərdə tutur. Mövzu təhlilinin məqsədi mətndə mövcud olan əsas mövzuları və ideyaları daha dərindən dərk etmək və bu məlumatdan fikirlər çəkmək və əsaslandırılmış qərarlar qəbul etmək üçün istifadə etməkdir. Bu, mövzu və ya mövzu əsasında hər bir fərdi mətnə etiketlər və ya kateqoriyalar təyin etməklə əldə edilir. Mövzu təhlili prosesi insan dilini dekonstruksiya etmək üçün təbii dil emalından (NLP) istifadəni nəzərdə tutur. Bu, qiymətli fikirlər əldə etmək və məlumatlara əsaslanan qərarların qəbulunu asanlaşdırmaq üçün istifadə oluna bilən nümunələrin müəyyən edilməsinə və mətnlərdəki semantik strukturların aşkarlanmasına imkan verir. Maşın öyrənməsindən istifadə edərək mövzu təhlili üçün geniş yayılmış metodologiyalar təbii dil emalı mövzu modelləşdirmə və mövzu təsnifatıdır. Mövzu modelləşdirməsində nəzarətsiz maşın öyrənmə texnikasının tətbiqi müşahidə olunur. Bu o deməkdir ki, o, mövzu teqləri və ya təlim öncəsi məlumatların müəyyən edilməsi şərti olmadan nümunələri çıxarmaq və oxşar ifadələri qruplaşdırmaq qabiliyyətinə malikdir. Sözügedən alqoritm məqsədəuyğunluq və rahatlıqla həyata keçirilə bilər, lakin bu, nisbətən qeyri-dəqiqliyin çatışmazlığı ilə müşayiət olunur. Əksinə, mətnin təsnifatı prosesi verilmiş mətnin mövzusu haqqında qabaqcadan bilik tələb edir, çünki mövzu təsnifatının hazırlanması məqsədilə verilənlərin etiketlənməsi mütləqdir. Tələb olunan əlavə addıma baxmayaraq, mövzu təsnifatçılarının istifadəsi uzunmüddətli perspektivdə əhəmiyyətli faydalar verir və klasterləşdirmə metodologiyaları ilə müqayisədə daha yüksək dəqiqlik təklif edir.

Mövzu təhlilinin əhatə dairəsi

Mövzu təhlilinin əhatə dairəsi verilmiş mətndə və ya mətnlər toplusunda mövcud olan əsas mövzu və anlayışların araşdırılması və müəyyənləşdirilməsidir. Bu proses təqdim olunan əsas ideyaları və arqumentləri müəyyən etmək üçün mətnin məzmununu və

strukturunu təhlil etməyi əhatə edir. Mövzu təhlilinin məqsədi mövzunu daha dərindən başa düşmək və gələcək tədqiqat və ya təhlil üçün məlumat verə biləcək nümunələri və meylləri müəyyən etməkdir.

Mövzu təhlilinin tətbiqi müxtəlif səviyyələrdə həyata keçirilə bilər:

- Sənəd səviyyəsində mövzu modeli verilmiş korpusdan fərqli mövzuları çıxarır. Məsələn, elektron poçtun və ya jurnalist yazısının mövzusu.
- Cümlə səviyyəsində mövzu modeli fərdi cümlənin predmetini əldə edir. Buna misal olaraq, jurnalist nəşrindəki başlığın mövzusu ola bilər.
- Alt cümlə səviyyəsində mövzu modeli cümlə daxilində olan alt ifadələrin mövzularını çıxarmağa qadirdir. Bunun bir nümunəsi, məhsulun nəzərdən keçirilməsinin tək bir cümlə daxilində müxtəlif mövzuları ehtiva etməsidir.
- Mövzu təhlili verilmiş mətndən və ya korpusdan mənalı mövzuları müəyyən etmək və çıxarmaq üçün təbii dilin işlənməsi, maşın öyrənməsi və məlumat axtarışı da daxil olmaqla müxtəlif sahələrdə istifadə olunur. Mövzuların etiketlənməsinin istifadəsi çoxlu sayda mətn məlumatlarının sürətli və qənaətli şəkildə təhlilində çox sərfəlidir. Bura daxili sənədlər, müştəri ünsiyyəti və onlayn məzmun daxildir. Mövcud olan böyük miqdarda məlumatın əl ilə təsnifləşdirilməsi zəhmət tələb edən, bahalı və nisbətən qeyri-dəqiq bir proses ola bilər.

Mövzu təhlilinin əhəmiyyəti

Müəssisələr gündəlik olaraq böyük miqdarda məlumat toplayır və istehsal edir. Məlumatların təhlili və emalı üçün avtomatlaşdırılmış mövzu təhlili metodlarından istifadə bizneslərə qərar qəbul etmə imkanlarını artırmaqda, daxili prosesləri tənzimləməkdə, nümunələri tanımaqda və onların səmərəliliyini və məhsuldarlığını artıra biləcək müxtəlif faydalar əldə etməkdə kömək edə bilər. Maşın öyrənmə modelləri böyük həcmdə məlumatların çeşidlənməsi prosesində həlledici rol oynayır. Mövzunun aşkarlanması prosesi geniş sənədlərin səmərəli araşdırılmasını asanlaşdırır, müştərilərin müzakirə etdiyi mövzuları müəyyən etməyə imkan verir.

Mövzu təhlilinin necə işləmə prinsipləri

Mətn mövzularının avtomatlaşdırılmış təhlili üçün istifadə edilə bilən çoxlu metodologiya və strategiyalar mövcuddur. Müəyyən bir yanaşmanın seçilməsi həll olunan konkret problemdən asılıdır. Mövzu təhlili modellərinin incəliklərini başa düşmək üçün diqqətimizi iki üstünlük təşkil edən metodologiyaya yönəldəcəyik: Mövzu modelləşdirilməsi və Mövzu təsnifatı. Mövzuların modelləşdirilməsi mətnlər toplusunun mövzusunu müəyyən etmək üçün uyğun bir yanaşmadır. Mövzunun təsnifatı, əvvəlcədən müəyyən edilmiş mövzular əsasında mətnlərinə avtomatik olaraq müvafiq teqlər təyin etməyə çalışırsa, arzu olunan xüsusiyyətdir. Mövzunun modelləşdirilməsi və mövzu təsnifatı arasında müqayisə təbii dil emalı sahəsində aktual mövzudur. Hər iki üsul mətn məlumatlarından mənalı məlumat çıxarmaq məqsədi daşıyır, lakin yanaşmalarına görə fərqlənirlər. Mövzu modelləşdirməsi sənədlər korpusunda gizli mövzuları müəyyən edən nəzarətsiz öyrənmə metodu olsa da, mövzu təsnifatı məzmununa əsasən yeni sənədlərə əvvəlcədən müəyyən edilmiş kateqoriyalar təyin edən nəzarət edilən öyrənmə texnikasıdır. Bu iki metod arasındakı fərqləri və oxşarlıqları anlamaq tədqiqatçılara və praktikantlara öz xüsusi ehtiyacları üçün ən uyğun yanaşmanı seçməyə kömək edə bilər.

Maşın öyrənməsinin tətbiqi zəhmət tələb edən və vaxt tələb edən əl işlərini avtomatlaşdırmaq potensialına malikdir. Çoxsaylı maşın öyrənmə alqoritmləri, yalnız sənədlərin mətn məzmununa əsaslanaraq, minimal rəhbərliklə verilmiş sənədlər toplusundan mövzu çıxarmaq qabiliyyətinə malikdir. Bu kontekstdə istifadə edilən alqoritmlərin əksəriyyəti nəzarətsizdir. Bu, giriş məlumatlarının və təlim parametrlərinin təmin edildiyini və alqoritmlərin qalan tapşırıqları yerinə yetirməsini nəzərdə tutur. Bu tip alqoritm mövzu modelləşdirməsinin icrasında istifadə olunur. Əksinə, nəzarət edilən alqoritmlər mövcuddur. Maşın öyrənmə alqoritmləri mətni mövzuya uyğun olaraq avtonom təsnif etmək bacarığını inkişaf etdirmək üçün etiketli məlumat nümunələrindən istifadə etməklə öyrədilir.

Teorik olaraq, nəzarətsiz maşın öyrənmə alqoritmləri nəzarət edilən alqoritmlərlə müqayisədə daha az əmək tutumlu olur, çünki onlar insan etiketli məlumatlara ehtiyac duymurlar. Bununla belə, onların əhəmiyyətli miqdarda yüksək keyfiyyətli məlumatlara ehtiyacı ola bilər. Bu ssenaridə təhlil prosedurunun tərkib hissəsi kimi mətn daxilindəki mövzuları açmaq üçün yalnız nəzarətsiz alqoritmlərdən istifadə etmək faydalı ola bilər. Mövzu modelləşdirmə alqoritmi əlaqələri çıxarmaq üçün istifadə etdiyi xüsusi terminologiya ilə birlikdə bir-biri ilə əlaqəli hesab etdiyi sənədlər dəstini yaradacaqdır. Bu münasibətlərin əsl mənasını deşifrə etmək məsuliyyəti sizin üzərinizə düşəcək. Əksinə, nəzarət edilən maşın öyrənmə alqoritmləri etiketlənmiş nümunələrin təmin edilməsi yolu ilə maşına istənilən nəticəni öyrətmək səyini tələb edir. Beləliklə, mövzunun tərifi və etiketləmə prosedurunun əhəmiyyətini etiraf etmək vacibdir, çünki onlar praktik tətbiqlərdə modelin effektivliyinin müəyyən edilməsində əsas rol oynayır. Nəzarət edilən alqoritmlərin digər alqoritm növlərinə nisbətən açıq üstünlüyü olduğu hesab edilir. Öz meyarlarını dəqiqləşdirmək və mətnlərin ardıcıl etiketlənməsi ilə mövzuları müəyyən etməklə, yeni, görünməmiş nümunələri müvafiq mövzulara əsasən təsnif etmək üçün bir model hazırlamaq olar. Bu proses insan təsnifatı ilə müqayisə edilə bilən nəticələr verə bilər. Gəlin həm mövzu modelləşdirməsinin, həm də mövzu təsnifatının əməliyyat mexanizmlərini dərk etməyə daha dərindən baxaq.

2.3. Mövzu Təhlilinin metodları

Mövzu Modelləşdirməsi

Mövzu modelləşdirməsi mətn sənədləri korpusunda mövcud olan müxtəlif mövzuları müəyyən etmək və təsnif etmək üçün istifadə olunan bir texnikadır. Məqsəd sənədləri əhatə etdikləri mövzular əsasında qruplaşdırmaqdır. Alqoritmlər hər bir sənədin mövzuların qarışığından ibarət olduğu fərziyyəsi altında işləyir və sonradan hər bir mövzunun konkret sənəd daxilində yayılma dərəcəsini müəyyən etməyə çalışır. Proses sənədlərin leksik məzmununa görə kateqoriyalara bölünməsini və onlar arasında qarşılıqlı

əlaqənin müəyyən edilməsini nəzərdə tutur. Bəzi oxşarlıqlara baxmayaraq, mövzu modelləşdirmə və klaster təhlili arasında fərq qoymaq vacibdir. Mövzu modelləşdirməsinin əsasını təşkil edən anlayışları daha dərindən başa düşmək üçün biz ən çox istifadə olunan iki alqoritmin, yəni LSA və LDA-nın əsas prinsiplərini araşdıracağıq.

Gizli Semantik Təhlil (LSA)

Gizli Semantik Təhlil (LSA) bir sıra sənədlər və onların ehtiva etdiyi terminlər arasındakı əlaqələri təhlil etmək üçün istifadə olunan riyazi texnikadır. Bu, böyük məlumat dəstlərində nümunələri və əsas mənaları müəyyən etmək məqsədi daşıyan statistik modelləşdirmə formasıdır. LSA təbii dilin işlənməsi, məlumat axtarışı və maşın öyrənməsi daxil olmaqla müxtəlif sahələrdə geniş istifadə edilmişdir. Mövzunun modelləşdirilməsi üçün ənənəvi yanaşma Gizli Semantik Analizdən istifadə etməkdir. Dağıtım fərziyyəsinin prinsipi oxşar mətn kontekstində söz və ifadələrin baş verməsi ilə onların semantik oxşarlığı arasında korrelyasiyanın mövcud olduğunu irəli sürür. Naive Bayes kimi, bu üsul verilənlər bazasında mövcud olan söz tezliklərinə əsaslanır. Ümumi konsepsiya hər bir fərdi sənəddə hər bir terminin baş verməsinin hesablanmasını və daha sonra eyni terminlərin yüksək tezliklərini nümayiş etdirən sənədlərin qruplaşdırılmasını əhatə edir. İstifadə olunan metodologiyanı daha dərindən araşdırmaq üçün "söz tezliyi" termininin dəqiq tərifini yaratmaq vacibdir. Müəyyən bir sənəddə sözün və ya ifadənin baş vermə sürətini ifadə edən ədədi dəyər onun tezliyi adlanır. Bu metrik konkret terminin sənəddə neçə dəfə göründüyünü ölçməyə xidmət edir. Həqiqətən də, bu alqoritmlər söz sırası, məna və qrammatika daxil olmaqla sintaksis və semantikaya məhəl qoymur və bunun əvəzinə hər bir sənədi strukturlaşdırılmamış sözlər toplusu hesab edir. Sözün tezliyi düz hesablama üsulu ilə müəyyən edilə bilər, məsələn bir sənəddə "sərəncam" termini 10 dəfə rast gəlinirsə, onun tezliyi 10-dur. Yuxarıda qeyd olunan metodologiya bəzi məhdudiyyətlər nümayiş etdirmişdir, buna görə də tf-idf-dən adətən istifadə olunur. Tf-idf alqoritmi sözün bütün sənədlər üzrə nisbi tezliyini konkret sənəddəki tezliyi ilə müqayisədə nəzərə alır. Nəticə etibarı ilə, daha çox istifadə olunan sözlər ən çox yayılmamasına baxmayaraq,

sənədin daha təsirli təsviri hesab edildiyi üçün daha yüksək dərəcə verilir. Söz tezliyinin hesablanması zamanı hər bir söz üçün bir sıra və hər bir sənəd üçün bir sütundan ibarət bir matris yaradılır. Hər bir fərdi hüceyrə müəyyən bir sənəd daxilində müəyyən bir sözün hesablanmış tezliyini ifadə edir. Təqdim olunan matris sənədlər və terminlər arasında əlaqə quran sənəd termini matrisidir. Onun daxilində arzu olunan komponentlər, yəni sənədlər və mövzular arasındakı əlaqəni təmsil edən matris, həmçinin terminlər və mövzular arasındakı əlaqəni təmsil edən matris var. Sözügedən matrislər mətnlərdə göstərilən mövzulara aid məlumatları göstərir.

Matrislər sənəd termini matrisinin kəsilmiş Tək Dəyər Parçalanması (SVD) texnikasından istifadə edərək üç matrisə parçalanması prosesi vasitəsilə yaradılır. Tək Dəyərin Parçalanması xətti cəbri alqoritmdir və matrisin üç matrisə, yəni U, S və V-yə bölünməsini nəzərdə tutur. Qeyd etmək lazımdır ki, orta S matrisi orijinal matrisin tək qiymətlərinin diaqonal matrisidir. Gizli Semantik Təhlil (LSA) kontekstində hər bir tək dəyər perspektiv mövzu kimi şərh edilə bilər. Kəsilmiş SVD ən böyük t sinqulyar dəyərləri seçir və U-nun ilk t sütununu və V-nin ilk t sətirini saxlayır, orijinal parçalanmanın ölçülərini azaldır. t alqoritmin tapdığı mövzuların sayı olacaq, ona görə də tənzimləmə tələb olunacaq hiperparametrdir. Konsepsiya ən vacib mövzuların seçilməsini nəzərdə tutur, burada U sənəd-mövzu matrisini, V isə termin-mövzu matrisini təmsil edir. Nəzərdən keçirilən matrislər sənədləri təmsil edən vektorlardan və mövzularla ifadə olunan terminlərdən ibarətdir. Bu vektorlar kosinus oxşarlığı kimi üsullardan istifadə etməklə qiymətləndirilə bilər.

Gizli Dirixlet Ayrılması (LDA)

Gizli Dirixlet Ayrılması (LDA) təbii dil emalında və maşın öyrənməsində istifadə olunan ehtimala əsaslanan mövzu modelləşdirmə texnikasıdır. LDA-nın hərtərəfli başa düşülməsi ehtimala aid mürəkkəb riyazi anlayışların hərtərəfli başa düşülməsini tələb edir. Bununla belə, onun əsasında duran əsas anlayış daha asan başa düşüləndir. Əvvəlcədən müəyyən edilmiş fənlər toplusunu nəzərdən keçirək, burada hər bir mövzu müəyyən

edilməmiş terminlər qrupu ilə xarakterizə olunur. Sənədlərimizin məzmunu bir sıra mövzuları əhatə edir, lakin onların spesifik mahiyyəti hazırda bizə məlum deyil. Gizli Dirixlet Ayrılması alqoritmi məlum sənədlərlə naməlum mövzular arasında uyğunluq yaratmağa çalışır ki, hər bir sənəddəki sözlərin əksəriyyəti müəyyən edilmiş mövzularla effektiv şəkildə təmsil olunsun. Buradakı əsas fərziyyə Gizli Semantik Təhlildə (LSA) istifadə edilənə bənzəyir, bununla da eyni mövzulu sənədlər oxşar sözlərdən istifadə edəcək. Adətən hər bir sənədin müxtəlif mövzuların qarışığından ibarət olduğu və hər bir fərdi sözün müəyyən bir mövzu ilə əlaqəli olma ehtimalı olduğu güman edilir. LDA modeli sənədin yaradılmasının iki mərhələli proses vasitəsilə baş verdiyi fərziyyəsi altında işləyir: birincisi, hər bir mövzuya müəyyən çəki təyin edilməklə mövzuların birləşməsi seçilir (məsələn, A mövzusu üçün 20%, B mövzusu üçün 80% və C mövzusu üçün 0%); ikincisi, sözlər seçilmiş mövzularla əlaqəsinə görə seçilir. Sözlərin seçilməsi onların konkret sənəd daxilində baş vermə ehtimalını nəzərə alan stoxastik prosesə əsaslanır.

Şübhəsiz ki, faktiki olaraq sənədlər belə tərtib olunmur. Yazılı qeydlər fərdlər tərəfindən tərtib edilir və onları oxunaqlı edən atributlara malikdir, məsələn, söz sırası, qrammatika və s. İddia etmək olar ki, yazılı işdə istifadə olunan leksikonun sadəcə tədqiqi, sənədin nəzərdə tutulan ünsiyyətinin effektiv şəkildə çatdırılmasından asılı olmayaraq, onun mövzusunun müəyyənləşdirilməsini asanlaşdıra bilər. Bu, Gizli Dirixlet Ayrılması (LDA) funksiyasıdır. Sistem verilmiş sənədi qəbul edir və onun əvvəllər izah edilən şəkildə hazırlandığını qəbul edir. Sonradan, tərkib sözlərdən geriyə doğru gedən proses sənədi əhatə edir və sözlərin xüsusi konfiqurasiyası ilə yekunlaşan mövzuların birləşməsindən nəticə çıxarmağa çalışır.

Qeyd etmək lazımdır ki, həyata keçirilməsi təlim məqsədi ilə adətən α (alfa) və β (beta) kimi istinad edilən iki hiperparametri özündə birləşdirir. Bunların funksionallığı ilə tanışlıq alqoritmi özündə birləşdirən kitabxanalardan səmərəli istifadə etmək üçün çox vacibdir. Alfa parametri sənədlər arasında oxşarlıq dərəcəsini tənzimləmək üçün cavabdehdir. Daha kiçik dəyər sənədlərin məhdud sayda mövzulardan ibarət olduğunu

göstərəcək, daha böyük dəyər isə daha çox mövzunu əhatə edən sənəd təqdimatlarını yaradacaq və nəticədə sənədlər arasında oxşarlıq artacaq. Beta metrikası mövzulara aiddir və onlar arasındakı oxşarlıq dərəcəsini tənzimləyir. Azaldılmış ədədi dəyər, hər bir mövzuya daha kiçik fərqli terminlər dəstini ayırmaqla mövzular arasında daha yüksək dərəcədə fərqləndirməni ifadə edəcəkdir. Daha yüksək ədədi dəyər tərs nəticə verəcək və mövzulara daha çox sayda paylaşılan sözlərin daxil edilməsinə səbəb olacaq.

Təlimə başlamazdan əvvəl modelin əhatə edəcəyi mövzuların sayını dəqiqləşdirmək vacibdir. Alqoritmin avtonom qərar qəbul etmə qabiliyyəti məhduddur, çünki müəyyən ediləcək mövzuların sayı ilə bağlı açıq təlimat tələb olunur. Sonradan, hər bir sənəd üçün əldə edilən nəticə həmin sənədə xas olan mövzuların qarışığından ibarət olmalıdır. Əldə edilən məlumatlar müvafiq olaraq 0,2 və 0,7 ilə işarələnən "A" və "B" kimi xüsusi mövzulara uyğun gələn bir sıra ədədi dəyərlərdən ibarət vektor şəklindədir. Bu vektorları müqayisə etmək üçün müxtəlif üsullardan istifadə etmək olar və bu cür müqayisələr korpusu dərk etmək və onun əsas strukturları haqqında anlayışlar əldə etmək üçün dəyərlidir.

Mövzu təsnifləşdirilməsi

Mövzunun modelləşdirilməsi üçün istifadə olunan alqoritmlərdən fərqli olaraq, mövzu təsnifatı üçün istifadə olunan maşın öyrənmə alqoritmləri nəzarət olunan xarakter daşıyır. Proses alqoritmlərə əvvəlcədən kateqoriyalara ayrılmış sənədlərin təqdim edilməsini nəzərdə tutur ki, bu da sonradan bu əvvəlcədən mövcud mövzular əsasında yeni, müşahidə olunmamış sənədləri təsnif etmək imkanı əldə edir. Sənədlər üçün mövzuların müəyyən edilməsi prosesi nəzərə alınmalı olan ayrı bir məsələdir. Əgər kimsə əvvəlcədən mövcud olan işi avtomatlaşdırmaq niyyətindədirsə, çox güman ki, onlar öz yazılı əsərlərində müzakirə olunan mövzunu yaxşı başa düşürlər. Alternativ olaraq, təhlildən əvvəl sənədlərdə olan materialı daha dərindən başa düşmək üçün əvvəllər müzakirə olunan mövzu modelləşdirmə üsullarından istifadə edilə bilər. Praktiki vəziyyətlərdə mövzular modelin qurulması zamanı aşkarlanır. Avtomatlaşdırılmış təsnifat, istər qaydalara

əsaslanan, istərsə də maşın öyrənmə yanaşmaları vasitəsilə, adətən mətnlərin ilkin əl analizi və etiketlənməsini tələb edir. Nəticə olaraq, mövzu dəsti proses boyu tez-tez dəqiqləşdirməyə məruz qalır. Modeli tam hesab etmək üçün mövzuların yaxşı müəyyən edilməsi və verilənlər bazasının vahid olması vacibdir. Daha sonra, avtomatlaşdırılmış mövzu təsnifatının əsas yolları, yəni qaydalara əsaslanan sistemlər, maşın öyrənmə sistemləri və hibrid sistemlər izah ediləcəkdir.

Qaydalara əsaslanan sistemlər

Maşın öyrənmə alqoritmlərini araşdırmadan əvvəl diqqətəlayiqdir ki, bir mövzu təsnifatçısı maşın öyrənməsindən istifadə etmədən yalnız əl səyləri ilə qurula bilər. Metodologiya, bir mütəxəssis tərəfindən nəzərdən keçirilmiş sənədlərin məzmununa əsaslanan əvvəlcədən müəyyən edilmiş qaydalar toplusunun birbaşa proqramlaşdırılmasını nəzərdə tutur. Konsepsiya, qaydaların rəsmiləşdirilmiş təcrübəni təcəssüm etdirdiyini və mətnin semantik cəhətdən uyğun komponentlərini və sənədlə əlaqəli ola biləcək metaməlumatları nəzərdən keçirməklə müxtəlif subyektlərin mətnləri arasında fərqləndirmə qabiliyyətinə malik olduğunu nəzərdə tutur. Yuxarıda qeyd olunan qaydaların hər biri fərqli nümunə və müvafiq proqnozdan, xüsusilə də proqnozlaşdırılan mövzudan ibarətdir. Təsvir edilən kimi qaydalara əsaslanan sistemlər insanlar tərəfindən asanlıqla başa düşülür. Bir şəxs qaydaları nəzərdən keçirə və modelin fəaliyyətini başa düşə bilər. Müəyyən bir müddət ərzində mövcud qaydaları dəyişdirərək və yeniləri daxil etməklə onları təkmilləşdirmək mümkündür. Buna baxmayaraq, müəyyən çatışmazlıqlar var. Əvvəlcə qeyd etmək lazımdır ki, bu sistemlər domenin dərindən dərk edilməsini tələb edir. Yadda saxlamaq lazımdır ki, "mütəxəssis" termini işlədilirdi və bu, təsadüfi deyildi. Bundan əlavə, mürəkkəb sistemlər üçün qaydaların işlənib hazırlanması çətin ola bilər ki, bu da onların nəzərdə tutulan funksionallığını təmin etmək üçün geniş təhlil və sınaq tələb edir. Nəhayət, iddia oluna bilər ki, qaydalara əsaslanan sistemlər texniki xidmət və miqyaslılıq baxımından çətinliklər yaradır, çünki yeni qaydaların əlavə edilməsi əvvəlcədən mövcud qaydaların icrasına zərərli təsir göstərə bilər.

Maşın öyrənmə təsnifatında

Maşın öyrənmə təsnifatında Təbii Dil Emalı (NLP) mövzu təsnifat modelinin öyrədilməsi prosesi mətn nümunələrindən və onların müvafiq gözlənilən kateqoriyalarından istifadəni əhatə edir. Yuxarıda göstərilən model təlim məlumatlarından bilik əldə etmək üçün təbii dil emalından istifadə edir, ona nümunələri müəyyən etməyə və mətni əvvəlcədən müəyyən edilmiş təsnifatlara təsnif etməyə imkan verir. Əvvəlcə təlim məlumatlarını maşınla oxuna bilən formata, xüsusən vektorlara çevirmək lazımdır. Vektorlar məlumatları kodlayan ədədi siyahılardan ibarətdir. Vektorların istifadəsi vasitəsilə model, adətən xüsusiyyətlər kimi adlandırılan, onun təlim məlumatlarından öyrənmək və proqnozlar yaratmaq qabiliyyətini asanlaşdıran müvafiq məlumat hissələrini çıxara bilir. Bu məqsədə çatmaq üçün müxtəlif üsullardan istifadə edilə bilər, lakin sözlər çantasının (bag of words) üsulu ilə vektorlaşdırılması ən çox istifadə olunanlardan biridir. Təlim məlumatlarının vektorlara çevrilməsindən sonra, vektorlar əsasında gələcək mətnləri təsnif edə bilən bir model yaratmaq üçün bir alqoritm istifadə olunur.Yeni proqnozlar yaratmaq üçün təlim keçmiş model daxil olan mətn daxiletməsini vektor təsvirinə çevirmək üçün vektorlaşdırma prosesindən istifadə edir. Sonradan model vektordan müvafiq xüsusiyyətləri çıxarır və proqnoz vermək üçün onlardan istifadə edir. Təsnifat modelinin təkmilləşdirilməsinə təlim üçün istifadə olunan məlumatların miqdarının artırılması və alqoritmin hiperparametrlərinin tənzimlənməsi yolu ilə nail olmaq olar.

Ən məşhur mətn təsnifat alqoritmlərindən bəzilərinə Naive Bayes alqoritmlər ailəsi, dəstək vektor maşınları (SVM) və dərin öyrənmə daxildir.

- Naive Bayes alqoritmik ailəsi sadəliyi və minimal təlim məlumatları və hesablama resursları ilə yüksək keyfiyyətli nəticələr əldə etmək qabiliyyəti ilə tanınır. Naive Bayes, maşın öyrənməsində və məlumatların əldə edilməsində geniş istifadə olunan ehtimal alqoritmidir. O, hadisə ilə əlaqəli ola biləcək şərtlər haqqında əvvəlcədən biliyə əsaslanan hadisənin baş vermə ehtimalını təsvir edən riyazi düstur olan Bayes

teoreminə əsaslanır. Naive Bayes-in mətn təsnifatı, spam filtrasiyası və əhval-ruhiyyə təhlili daxil olmaqla bir çox tətbiqlərdə effektiv olduğu göstərilmişdir. Gizli Semantik Təhlil (LSA)-ya bənzər Multinomial Naive Bayes (MNB), verilmiş mətndə sözlərin baş vermə ehtimalı ilə həmin mətnin konkret mövzuya aid olma ehtimalı arasında əlaqə qurur. Gizli Semantik Təhlil (LSA) və Multinomial Naive Bayes (MNB) arasındakı əsas fərq onların müvafiq post-məlumat emal metodologiyalarındadır. LSA verilmiş verilənlər toplusunda nümunələri müəyyən etməyə çalışır, MNB isə yeni mətnlər üçün proqnozlar yaratmaq üçün məlumat dəstindən istifadə edir.

- Dəstək Vektor Maşınları (SVM) sadə bir konsepsiya üzərində qurulmasına baxmayaraq, Naive Bayes-dən daha mürəkkəb bir alqoritmdir. Nəticə etibarilə, SVM daha çox hesablama resursları tələb edir, lakin o, ümumiyyətlə, üstün nəticələr verir. Dəstək Vektor Maşınlarının əsasını təşkil edən əsas konsepsiya mətn məlumatlarının vektorlaşdırılmasını və bununla da onları riyazi məkanda nöqtələr kimi göstərməsini nəzərdə tutur. Sonradan məqsəd optimal hiperplanı, yəni daha yüksək ölçülü məkanda bu vektorları istənilən kateqoriyalara effektiv şəkildə ayıran xətti müəyyən etməkdir. Sonradan, yeni mətn daxiletməsini qəbul etdikdən sonra, proses onun vektor təsvirinə çevrilməsini və sonradan təyin edilmiş həddə nisbətən mövqeyinə əsasən təsnifatının müəyyən edilməsini nəzərdə tutur: bu, mövzunun nəticəsinə uyğundur.
- Təlimin böyük həcmdə məlumatlara əsaslanaraq öyrənmək və proqnozlar vermək üçün süni neyron şəbəkələrini əhatə etdiyi tədqiqat sahəsi adətən dərin öyrənmə adlanır. Son dövrlərdə bu alqoritmlər əhəmiyyətli bir canlanma yaşadı. Bunu hesablama xərclərinin azalması, hesablama imkanlarının gücləndirilməsi və geniş məlumat dəstlərinin bolluğu ilə əlaqələndirmək olar. Mətn təsnifatının, xüsusən də mövzu təsnifatı sahəsində yenidən canlanması əhəmiyyətli faydalar verdi və ciddi hesablama tələbləri hesabına adətən əlverişli nəticələr verir. Dərin öyrənmə

modellərinin bir neçə gündən bir neçə həftəyə və ya hətta aylara qədər uzun müddət ərzində təlim keçməsi adi haldır. Mövzu təsnifatı üçün istifadə edilən iki əsas dərin öyrənmə arxitekturası Qıvrımlı Neyron Şəbəkələri (CNN) və Təkrarlanan Neyron Şəbəkələri (RNN)-dir.

Dərin Öyrənmə və Ənənəvi Maşın Öyrənmə alqoritmlərinin müqayisəsi süni intellekt sahəsində maraq doğuran mövzudur. Fərqliliklər bu əlyazmanın səlahiyyətlərindən kənardadır, lakin burada empirik ölçmələrlə hərtərəfli qiymətləndirmə təqdim olunur. Dərin öyrənmə alqoritmlərinin adi maşın öyrənmə alqoritmlərinə nisbətən daha böyük həcmdə təlim məlumatı tələb etməsinə baxmayaraq, məlumatların miqdarının artması ilə dərin öyrənmə təsnifatçılarının performansı yaxşılaşır. Əksinə, Dəstək Vektor Maşınları (SVM) və Multinomial Naive Bayes (MNB) kimi adi maşın öyrənmə üsulları doyma nöqtəsini nümayiş etdirir, ondan kənarda onların performansı hətta təlim məlumatlarının miqdarının artması ilə də artırıla bilməz. Bu o demək deyil ki, digər alqoritmlər daha pisdir; qarşıya qoyulan vəzifədən asılıdır. Nümunə olaraq, ümumiyyətlə spam aşkarlanması kimi tanınan istənməyən mesajların müəyyən edilməsi və süzülməsi vəzifəsi bir neçə onilliklər əvvəl Naive Bayes və n-qramların istifadəsi ilə uğurla həll edilmiş hesab olunurdu. Word2Vec və ya GloVe kimi digər dərin öyrənmə alqoritmləri də istifadə olunur; bunlar digər ənənəvi maşın öyrənmə alqoritmləri ilə məşq edərkən sözlər üçün daha yaxşı vektor təsvirləri əldə etmək üçün əladır.

Hibrid sistemlər

Hibrid sistemlər, dəqiq və fərdiləşdirilmiş qaydalar vasitəsilə nəticələri artırmaq üçün əsas maşın öyrənmə təsnifatını qaydaya əsaslanan sistemlə inteqrasiya etməyi hədəfləyir. Yuxarıda qeyd olunan təlimatlar ilkin təsnifatçı tərəfindən qeyri-dəqiq şəkildə modelləşdirilmiş subyektləri düzəltmək üçün istifadə edilə bilər.

FƏSIL 3. MƏTINLƏRIN KONTEKSTƏ GÖRƏ SINIFLƏNDIRILMƏSI VƏ IŞLƏNILMƏ MƏRHƏLƏLƏRI

3.1. Mətinlərin kontekstə görə sinifləndirilməsi problemləri və onların tətbiqi istiqamətləri

Mətn təsnifatının tətbiqi çox yönlüdür və qısa mətn məzmununun təsnifatı və ya daha geniş yazılı materialların strukturlaşdırılması daxil olmaqla, geniş ssenariləri əhatə edə bilər. Mövzunun etiketlənməsi müəyyən bir mətnin mövzusunu başa düşmək prosesinə aiddir. Tez-tez məlumatların təşkili və sistemləşdirilməsi məqsədi ilə istifadə olunur, o, müvafiq mövzular əsasında sənədləri təşkil etmək və ya xəbər məqalələrini mövzularına görə təsnif etmək üçün bir vasitə kimi xidmət edir.

Çoxsinifli təsnifat tez-tez rast gəlinən maşın öyrənmə tapşırığıdır, yayılma baxımından reqressiyadan sonra ikincidir. Təsnifat tapşırığı K fərqli siniflərə təsnif edilən bir sıra təlim nümunələrindən istifadəni nəzərdə tutur. Məqsəd yeni, görünməyən məlumatların aid olduğu sinfi dəqiq proqnozlaşdıra bilən maşın öyrənmə modelini qurmaqdır. Təlim verilənlər toplusunu müşahidə etdikdən sonra model hər bir siniflə əlaqəli fərqli nümunələr haqqında bilik əldə edir və sonradan gələcək məlumatların təsnifatını proqnozlaşdırmaq üçün qeyd olunan nümunələrdən istifadə edir. Maşın öyrənmə sahəsindəki təcrübə səviyyəsindən asılı olmayaraq, verilənlər toplusundan istifadə və təlimin effektivliyi mənalı və praktiki fikirlərin əldə edilməsində mühüm əhəmiyyət kəsb edir. Xəbər məqalələri, şirkət sənədləri və ya e-poçt kimi avtomatlaşdırılmış mətn təhlili üçün maşın öyrənməsi və təbii dil emal üsullarının tətbiqi çox sinifli təsnifatın istifadəsi ilə mətn məlumatlarının adətən mövzuya əsaslanaraq əvvəlcədən müəyyən edilmiş kateqoriyalara təsnifatını asanlaşdırır. Böyük həcmli mətnin səmərəli təhlili biznes üçün vaxta əhəmiyyətli dərəcədə qənaət edə bilir.

Çoxsaylı təsnifat modelləri mövcuddur. Bir çox hallarda reqressiya modelini təsnifat modelinə çevirmək mümkündür. Bu, əslində logistik reqressiyanın əsasını təşkil edən əsas mexanizmdir. Hazırkı tədqiqat girişi təmsil etmək üçün xətti cavabın, $WX + b$ istifadəsini

əhatə edir və bu, sonradan sigmoid funksiyasının tətbiqi ilə 0 ilə 1 arasında dəyişən ehtimal dəyərinə çevrilir. Belə fərz edilir ki, model 0,5-dən çox ehtimal verirsə, giriş 0 sinfinə, ehtimal daha aşağı olarsa 1-ci sinfə aid edilir.

Dəstək vektor maşını (SVM) təsnifat üçün tez-tez istifadə olunan modeldir. Dəstək Vektor Maşınları (SVMs) məlumatları daha yüksək ölçülü məkana uyğunlaşdırmaq və bir və ya bir neçə hiperplandan istifadə etməklə fərqli kateqoriyalara bölmək yolu ilə işləyir. SVM iki sinfi ayırd etməyə qadir olan ikili təsnifatçıdır. K kateqoriyalarını ayırd etmək üçün (K – 1) Dəstək Vektor Maşınlarından (SVMs) istifadə etmək olar. Hər bir fərdi klassifikator müşahidənin K fərqli siniflərindən birinə təsnifatı ilə bağlı proqnoz verəcəkdir.

Naive Bayes modeli təbii dilin işlənməsi və mətnin çoxsinifli təsnifatı sahəsində geniş istifadə olunan yanaşmadır. Bu fenomenin populyarlığı əsasən onun sadəliyindən və sürətli təlimindən qaynaqlanır. Sadə Bayes klassifikatoru müəyyən bir sinfə aid olmanın kollektiv ehtimalını şərti ehtimallar ardıcıllığına dekonstruksiya etmək üçün Bayes teoremindən istifadə edir. Sadəlövh Bayes modeli, modeldə istifadə edilən giriş xüsusiyyətlərinin bir-birindən müstəqil olduğu fərziyyəsi altında işləyir, buna görə də "sadəlövh" termini yaranır. Bu, tamamilə dəqiq olmasa da, arzu olunan nəticələrə nail olmaq üçün qənaətbəxş bir qiymətləndirmə ola bilər. X girişinin k sinfinə təsnifatı ehtimalların hasili kimi ifadə oluna bilən sinfə üzvlük ehtimalı ilə müəyyən edilir. K-nin dəyəri bu məhsulu maksimuma çatdırdıqda X girişi k sinfi kimi təsnif edilir.

Təsnifat məqsədi ilə çoxsaylı dərin öyrənmə modelləri mövcuddur. Neyron şəbəkənin son qatına softmax funksiyasının əlavə edilməsi onu effektiv şəkildə təsnifatçıya çevirə bilər, çünki bu aktivləşdirmə funksiyası bir sıra siniflər üzərində ehtimal paylamaları yaratmağa qadirdir. Softmax funksiyası K siniflərini əhatə edən ehtimal paylanması yaradır və nəticədə K uzunluqlu çıxış vektoru yaranır. Vektorun hər bir komponenti girişin əlaqəli kateqoriyanın üzvü olması ehtimalını təmsil edir. Ən çox ehtimal olunan sinfin seçilməsi vektorun indeksini ən yüksək ehtimalla müəyyən etməklə əldə edilir.

Müxtəlif neyron şəbəkə arxitekturaları müxtəlif dərəcədə effektivlik nümayiş etdirir, bəzi modellər digərlərindən üstündür. Qıvrımlı neyron şəbəkələri (CNN) çox sinifli təsnifat tapşırıqlarında, xüsusən də şəkillər və mətn sahələrində müstəsna performans nümayiş etdirmişdir. Qıvrımlı neyron şəbəkələri məlumatlardan qiymətli xüsusiyyətləri, xüsusən miqyaslaşdırma, transformasiya və fırlanmaya baxmayaraq dəyişməz qalan xüsusiyyətləri çıxara bilir.

3. 2. Mətinlərin kontekstə görə sinifləndirilməsinin işlənilmə instrumental vasitələri

Verilənlər dəstləri

Mətnin təsnifatı əvvəlcədən müəyyən edilmiş təsnifatlara, nümunələrə, mövzulara və ya digər meyarlara görə onları təsnif etmək məqsədilə bütün mətn seqmentlərinə bir və ya bir neçə etiketin ayrılmasını nəzərdə tutan prosedurdur. Beləliklə, mətn təsnifatının effektivliyi dəqiq və etibarlı təlim məlumatı olmadıqda etibarsız sayılır. Maşın öyrənmə alqoritmlərinin effektivliyi onların keçmiş nümunələrin mənimsənilməsi yolu ilə dəqiq proqnozlar əldə etmək qabiliyyətindən asılıdır. Alqoritm müşahidə olunmamış mətn məlumatı üzrə proqnozlar yaratmaq üçün istifadə etdiyi dəqiq etiketlənmiş verilənlər toplusu üzərində öyrədilir.

Məlumatların annotasiyası kimi də tanınan verilənlərin etiketlənməsi işlənməmiş verilənlərə metadata və ya etiketlərin əlavə edilməsini nəzərdə tutur. Bu proses maşın öyrənmə modelini lazımi hədəf atributları və ya proqnozlaşdırmaq üçün tələb olunan cavablarla təmin etmək üçün nəzərdə tutulub. Etiket və ya etiket modelə verilmiş məlumat nöqtəsinin şəxsiyyəti haqqında məlumat verən təsviri atribut kimi xidmət edir və bununla da nümunələşdirmə vasitəsilə öyrənmə prosesini asanlaşdırır. Məlumat annotasiyasını yerinə yetirmək üçün müxtəlif metodologiyalar mövcuddur. Müəyyən bir üslubun seçilməsi problemin ifadəsinin mürəkkəbliyi, şərh ediləcək məlumatların həcmi, məlumat elmi komandasının böyüklüyü və ixtiyarında olan maliyyə və müvəqqəti resurslar kimi

müxtəlif amillərdən asılıdır. İstifadə olunan metodologiyadan asılı olmayaraq, məlumatların etiketlənməsi proseduru ardıcıl vaxt qrafikinə uyğun aparılır.

- Hər hansı maşın öyrənmə layihəsində ilkin addım şəkillər, audio yazılar, videolar və mətn materialları daxil ola biləcək müvafiq miqdarda işlənməmiş məlumatların əldə edilməsini nəzərdə tutur. Mənbələr müxtəlif şirkətlərdə fərqli ola bilər. Bəzi təşkilatlar illərdir daxili məlumat toplayır. Digərləri ictimaiyyət üçün əlçatan olan verilənlər bazasından istifadə edir. İstənilən halda, yuxarıda qeyd olunan məlumatlar tez-tez uyğunsuzluqlar, korrupsiya və ya verilmiş ssenari üçün qeyri-adekvatlıq nümayiş etdirir. Beləliklə, hər hansı bir etiket yaradılmazdan əvvəl təmizlənmədən və ilkin emaldan keçməlidir. Bir qayda olaraq, daha dəqiq nəticələr vermək üçün model üçün çoxlu sayda müxtəlif məlumat olmalıdır.
- Məlumatların annotasiyası prosesi mövzu üzrə mütəxəssislərin məlumatları diqqətlə tədqiq etmələrini və onlara təsviri metadata etiketləri ilə əlavə etmələrini əhatə edir. Bununla onlar modelin əsas həqiqət kimi istifadə edə biləcəyi mənalı kontekst əlavə edirlər.
- Keyfiyyət təminatı sahəsində məlumatların yüksək keyfiyyət, etibarlılıq, dəqiqlik və ardıcıllıq xüsusiyyətlərini nümayiş etdirməsi zəruridir. Maşın öyrənmə modeli təlim verilənlər bazalarının dəqiqliyi fərdi məlumat nöqtələrinin dəqiq etiketlənməsindən asılıdır. Metaməlumatların dəqiqliyinə zəmanət vermək və lazım gəldikdə onu optimallaşdırmaq üçün davamlı keyfiyyət yoxlamaları lazımdır.

Proqram dili

Tələb olunan təlim məlumatlarını əldə etdikdən sonra, sonrakı addım mətn təsnifatı yaratmaq məqsədi ilə onun maşın öyrənmə alqoritminə daxil edilməsini əhatə edir. Xoşbəxtlikdən, mətn məlumatlarının vektor təsvirlərinə çevrilməsi, maşın öyrənməsi alqoritmlərinin öyrədilməsi və proqnozlar yaratmaq üçün modellərdən istifadə də daxil olmaqla prosesin müxtəlif mərhələlərində fərdlərə kömək edə biləcək çoxsaylı resurslar mövcuddur. Açıq mənbəli kitabxanalar maşın öyrənməsinin mətn təsnifat vasitələrinin ən

qabaqcıl formaları arasında yüksək səviyyədə çıxış etmək qabiliyyətini nümayiş etdirmişdir. Açıq mənbəli kitabxanaların yayılması onu həyata keçirmək istəyən tərtibatçılar arasında maşın öyrənməsinin populyarlaşmasına əhəmiyyətli dərəcədə kömək etdi. Məlumat elmində və maşın öyrənməsində əsaslı fon ehtiyacına baxmayaraq, bu kitabxanalar ağlabatan dərəcədə abstraksiya və sadələşdirməni təmin edir. Python, Java və R, hazırda işlənmə mərhələsində olan və müxtəlif xüsusiyyətlər, performans və imkanlar təklif edən maşın öyrənmə kitabxanalarının geniş spektrini təmin edən proqramlaşdırma dilləridir.

Python təyin edilmiş proqramlaşdırma dili olaraq seçilmişdir. Sözügedən dil mürəkkəb, çox protokollu proqramların yaradılmasını asanlaşdırmaq və eyni zamanda qısa və asan başa düşülən sintaksisi dəstəkləmək qabiliyyətinə görə veb və proqram təminatının inkişafı sahələrində geniş istifadə olunur. Python bir çox geniş istifadə olunan proqramların hazırlanmasında istifadə edilmişdir. Bu proqramlaşdırma dili açıq mənbə icması tərtibatçılara çoxlu təkrar istifadə edilə bilən kodlar, çərçivələr və yardımlar təklif edir. O zəhmət tələb edən və təkrarlanan işləri avtomatlaşdırmaq qabiliyyətinə görə əhəmiyyətli bir üstünlük göstərir. Python daxili modullarından və ya geniş kitabxanasındakı əvvəlcədən mövcud koddan istifadə etməklə geniş spektrli tapşırıqları avtomatlaşdırmaq imkanı yaradır. Əlavə olaraq Python, testlərin avtomatlaşdırılması məqsədilə onu yüksək səviyyəli dilə çevirən hərtərəfli test çərçivələrinə malikdir. Python, məlumat elmi və tədqiqat səyləri üçün ən çox seçilən dil kimi qəbul edilir. Tədqiqat, hesabat, proqnozlaşdırıcı və ya reqressiya təhlilləri və digər məqsədlər üçün məlumatları idarə etmək üçün Pyhon-dan istifadə edilir.

Python, maşın öyrənməsi (ML) modellərini öyrətmək üçün istifadə olunan aparıcı proqramlaşdırma dilləri arasında hesab olunur. Dəqiq alqoritmlərdən istifadə etməklə, bu modellər verilənlər daxilindəki nümunələri yoxlamaq və aşkar etmək qabiliyyətinə malikdir və bununla da onlara təhlil edilmiş məlumatlar əsasında proqnozlar formalaşdırmağa və ya qərarlar qəbul etməyə imkan verir. Bundan əlavə, bu modellər yeni

dəyişənləri effektiv şəkildə həll etmək üçün əvvəlki məlumat dəstlərinin nəticələrini daxil etməklə davamlı inkişaf və uyğunlaşmadan keçir. Məlumat alimləri və tərtibatçıları tez-tez NumPy, Pandas və Matplotlib kimi kitabxanalardan maşın öyrənmə modellərini öyrədərkən məlumatların təmizlənməsi, çevrilməsi və vizuallaşdırılması kimi tapşırıqları avtomatlaşdırmaq üçün istifadə edirlər. Python-nın bir çox açıq mənbəli kitabxanaları mövcutdur ki, buda proqramçılara daha asan və tez zamanda maşın öyrənməsim, təbii dil emalı və süni intellect kodlarini yazmağa kömək edir. Aşağıda mətn təsnifatı üçün lazimi kitabxanaları nəzərdən keçirəcəyik.

- Pandas açıq mənbəli Python paketidir və əsasən məlumatların təhlili, məlumat elmi və maşın öyrənməsi məqsədləri üçün istifadə olunur. Paket çox ölçülü massivlər üçün yardım təklif edən Numpy adlı əvvəlcədən mövcud paket üzərində qurulub. Pandas, Python ekosistemində çoxsaylı digər məlumat elmi modulları ilə uyğunluq nümayiş etdirən çox bəyənilən məlumat manipulyasiya vasitəsidir. O, əməliyyat sistemi ilə inteqrasiya olunmuşlardan tutmuş kommersiya tədarükçü paylamalarına qədər Python paylamalarının hər yerdə yayılmış komponentidir.
- Rəqəmsal Python kimi də tanınan NumPy kitabxanası çoxölçülü massiv obyektləri toplusundan və onların manipulyasiyası üçün bir sıra əməliyyatlardan ibarətdir. Paket massivlərdə riyazi və məntiqi əməliyyatları yerinə yetirmək qabiliyyətinə görə Python proqramlaşdırma dili daxilində elmi hesablamalarda geniş istifadə olunur. NumPy Python skript dili ilə həyata keçirilən proqramlaşdırma dilidir.
- Scikit-learn kitabxanası geniş tətbiqi ilə maşın öyrənməsi üçün çox yönlü bir vasitə kimi geniş istifadə olunur. Platforma çoxsaylı alqoritmləri asanlaşdırır və mətn təsnifatı, reqressiya və klasterləşdirmə modellərini idarə etmək üçün sadə və effektiv funksiyalar təklif edir. Maşın öyrənməsi sahəsində yeni olan şəxslər üçün scikit-learn mətn təsnifatı üçün çoxlu dərsliklər və hərtərəfli bələdçilər təklif edən olduqca əlçatan kitabxanadır. Bu resurslar müxtəlif onlayn platformalarda geniş şəkildə mövcuddur və öyrənmə prosesinin asanlaşdırılmasında mühüm rol oynaya bilər.

- NLTK, təbii dil emalı (NLP) məqsədilə xüsusi olaraq hazırlanmış geniş istifadə olunan proqram kitabxanasıdır. Əhəmiyyətli bir istifadəçi bazasına və inkişaf edən inkişaf etdiricilər və həvəskarlar cəmiyyətinə malikdir. Mətnin təsnifatı maşınlara mətni dərk etməyə imkan verən müxtəlif faydalı alətlərdən istifadə etməklə asanlaşdırıla bilər. Bu vasitələrə paraqrafları ayrı-ayrı cümlələrə bölmək, sözləri diskret vahidlərə bölmək və hər sözün qrammatik funksiyasını müəyyən etmək bacarığı daxildir.
- Mürəkkəb alqoritmləri araşdırmaq üçün hazırlıq səviyyəsinə çatdıqdan sonra Keras, TensorFlow və PyTorch kimi dərin öyrənmə kitabxanalarını araşdırmaq məsləhətdir. Keras, onların yaradılmasını asanlaşdırmağa yönəlmiş sadələşdirilmiş dizaynına görə təkrarlanan neyron şəbəkələrinin (RNN) və qıvrımlı neyron şəbəkələrinin (CNN) inkişafı üçün uyğun ilkin çərçivədir. TensorFlow dərin öyrənmə alqoritmlərinin tətbiqi üçün açıq mənbə proqram təminatının ən yaxşı kitabxanası kimi geniş tanınır. Böyük verilənlər bazası ilə süni neyron şəbəkələrinin yaradılması, öyrədilməsi və tətbiqi məqsədilə optimallaşdırılmış bu kitabxana Google tərəfindən hazırlanmışdır.

Modelin qiymətləndirmə

Mətn təsnifatının effektivliyini qiymətləndirmək üçün müxtəlif ölçülərdən istifadə edilə bilər, o cümlədən accuracy, precision, recall, and F1 score. Mətn təsnifatının effektivliyi iki üsulla qiymətləndirilə bilər: təyin edilmiş teqləri olan məlumatlardan ibarət əvvəlcədən müəyyən edilmiş test dəsti ilə müqayisə və ya çarpaz doğrulama (cross-validation). Sonuncu, təlim məlumatlarının biri model təlimi, digəri isə nəticə testi üçün iki alt qrupa bölünməsini nəzərdə tutur. İndi bu parametrlərin hər biri haqqında ətraflı məlumat veriləcəkdir.

- Mətn təsnifatında accuracy proqnozların ümumi sayına münasibətdə təsnifatçı tərəfindən edilən düzgün proqnozların nisbətinə aiddir. Buna baxmayaraq, accuracy özü-özlüyündə mətn kateqoriyalarına ayırma alqoritminin effektivliyini

qiymətləndirmək üçün optimal metrik olmaya bilər. Nümunələrin kateqoriyalar üzrə qeyri-bərabər paylanması olduğu hallarda accuracy paradoksu yarana bilər. Bu paradoks modelin yüksək dəqiqliyə nail olduğu ssenariyə aiddir, lakin onun bütün teqlər üçün dəqiq proqnozlar vermək qabiliyyətinə zəmanət verilmir. Bu halda, precision, recall, and F1 score nəzərə almaq məsləhətdir.

- Precision müəyyən bir etiket üçün proqnozların düzgünlüyünün ölçüsüdür. Teq üçün düzgün proqnozların sayını həmin etiket üçün düzgün və yanlış proqnozların ümumi sayına bölməklə hesablanır. Bu halda, daha çox dəqiqlik yalan pozitivlərin azaldığını bildirir. E-poçt cavablarının avtomatlaşdırılması kimi müəyyən tapşırıqlarda yüksək dəqiqlik nümayiş etdirən mətn təsnifat modellərindən istifadə etmək vacibdir. Bu, modelin yalnız alıcının müəyyən bir etiketin altına düşmə ehtimalı yüksək olduqda cavab yaratmasını təmin edir.
- Recall müəyyən bir etiket üçün edilən dəqiq proqnozların sayının həmin etiket altında təsnif edilməli olan proqnozların ümumi sayına nisbətini ifadə edir. Yüksək recall nümayiş etdirən göstəricilər yalançı neqativlərin daha az olduğunu göstərir.
- F1 score həm precision, həm də recall nəticələrini nəzərə alan və bununla da mətn təsnifatının effektivliyinin göstəricisini təqdim edən performans göstəricisidir. Bu xüsusiyyət istifadəçiyə bütün istifadə olunan teqlər üzrə öz modelinin dəqiqliyini qiymətləndirməyə imkan verir.
- Çapraz doğrulama mətn kateqoriyasına aid modelin dəqiqliyini qiymətləndirmək üçün istifadə edilən bir texnikadır. Metodologiya təlim verilənlər toplusunun təsadüfi yanaşma ilə bərabər məsafəli alt çoxluqlar dəstinə bölünməsini nəzərdə tutur. Hər biri ümumi təlim məlumatlarının 25%-ni təşkil edən dörd fərqli alt dəstənin mövcud olduğu hipotetik ssenarini nəzərdən keçirək. Bir alt çoxluq istisna olmaqla, bütün digər alt çoxluqlar mətn təsnifatını öyrətmək üçün istifadə olunur. Sonradan qalan alt qrup üzrə proqnozlar yaratmaq üçün təsnifatdan istifadə edilir. Sonradan yuxarıda qeyd olunan bütün ölçüləri, yəni accuracy, precision, recall, and

F1 score-u toplamaq və bütün alt qruplar sınaqdan keçənə qədər proseduru yenidən başlamaq lazımdır. Nəhayət, nəticələr hər bir metrikanın orta effektivliyini əldə etmək üçün birləşdirilir.

3.3. Mətinlərin kontekstə görə sinifləndirilməsinin həyata keçirilməsi prosesi

Bu yarımfəsildə Azərbaycan xəbər agentliklərindən alınan 1000 xəbər məqaləsi təhlil edilir. Məqalələr Siyasət, Mədəniyyət, İdman, Maraqlı, Dünya və İqtisadiyyat daxil olmaqla altı fərqli kateqoriyaya bölünüb. Məlumatlar üç fərqli komponentə, yəni kateqoriya, başlıq və təsvirə ayrılmışdır. Siyasət, Mədəniyyət, İdman, Maraqlı, Dünya və İqtisadiyyat xəbərləri kimi altı xəbər kateqoriyası var, hər birinin müvafiq sıra sayı 153, 74, 184, 247, 171 və 171. İstifadənin əsas məqsədi maşın öyrənməsi alqoritmləri xəbərlərə aid yeni yaradılmış mətn məzmununu altı kateqoriyadan ibarət müvafiq kateqoriyaya təsnif etməkdir.

Category	Title	Description
Maraqlı	Sinir ,oynaq , sinir bel ağrılarına 3 gündə son !	Sabah Bakıda və Abşeron yarımadasında hava şəraiti əsasən yağmursuz olacaq. Ek
Maraqlı	Oxşar məhsullar fotoşəkillərlə axtarılacaq	"eBay" internet mağazasının əlavələrində fotoşəkilin tanınması funksiyasından istif
Maraqlı	Kiberidmançıların minimal əmək haqları artır	Kiberidmançı peşəsi ildən-ilə nüfuz qazanır. Kiberidmanın inkişafı və populyarlaşma
Maraqlı	Məşhur otel rüsvay oldu - VİDEO - FOTO	Kanar arxipelaqındakı Qran-Kanariya adasında, populyar Playya del İnqles kurortu
İdman	"Qarabağ" Çempionlar Liqasında tarixi oyuna çıxır	Bu gün "Qarabağ" komandası Çempionlar Liqasında növbəti matçına çıxacaq. Ağda
Dünya	"Məni cinlər göndərdi" deyib, qarşısına çıxanı bıçaqladı: yaralılar var - VİDEO	Türkiyənin Tekirda rayonunda psixoloji problemləri olan 31 yaşlı Yunus Admış 3 ada
Maraqlı	Norveçdə hicablı diktor islamofobların qəzəblənməsinə səbəb oldu	Norveç dövlət televiziyası NRK 22 yaşlı hicablı xanım Faten Mehdi Hüseynini diktor t
Dünya	Rusiyadan ƏKS HƏMLƏ: Müharibənin olmaması üçün İKİ VARİANT	Ukrayna hökuməti Donbasda müharibənin dayandırılmasının əlavə mexanizmlərinə
Siyasət	FHN-nin helikopteri bu gün də Gürcüstanda yanğınların söndürülməsi əməliyyatlarına cəlb olun	Azərbaycan Fövqəladə Hallar Nazirliyinin (FHN) helikopteri bu gün də Gürcüstanın
Maraqlı	Küçələrdə, avtobus dayanacaqlarında, çadırlarda yaşamış MƏŞHURLAR - FOTO	Bu gün Hollivudda milyonlarla dollar qazanan məşhur simalardan bir çoxunun keçm
Dünya	Mikroavtobus yük avtomobilləri ilə toqquşdu: 8 ölü, 4 yaralı	Böyük Britaniyada mikroavtobus iki yük avtomobili ilə toqquşub. xəbər verir ki, hadi
İdman	Robert Prosineçki: `Qarabağ` çox çətin qrupa düşüb` - MÜSAHİBƏ	Azərbaycan milli komandasının baş məşqçisi Robert Prosineçki Premyer Liqanın "Q
İdman	Futbol üzrə Azərbaycan Premyer Liqası: "Qarabağ" "Qəbələ"ni məğlub edib - VİDEO	Futbol üzrə Azərbaycan Premyer Liqasında III tura yekun vurulub. "Report"un məl
Dünya	Türkiyə Suriya sərhədinə ağır hərbi texnika cəmləşdirir	Türkiyə Suriya sərhədi yaxınlığında bir neçə haubitsa və tanklar yerləşdirib. Bu bara
Dünya	Qəzza sakinləri ilk dəfə kinoteatra gediblər	Qəzza səhərinin yüzlərlə sakini ötən gün son 30 ildə ilk dəfə olaraq kinoteatrda film

Şəkil 1: Algorimdə istifadə olunmuş məlumatın təsviri

Hərşeydən əvvəl lazimi Python kitabxanaları proqrama idxal olaraq tanıdılır ki, lazimi hazir kodlari işlədə bilək.

Import Libraries

```
import pandas as pd
from sklearn import metrics
from sklearn import svm
from sklearn.svm import LinearSVCS
from sklearn.neural_network import MLPClassifier as MLPC
from sklearn.naive_bayes import MultinomialNB
from sklearn.metrics import confusion_matrix, classification_report, accuracy_score
from sklearn.model_selection import train_test_split
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.feature_extraction.text import TfidfTransformer
from sklearn.model_selection import RandomizedSearchCV
from scipy.stats import reciprocal, uniform
```

Şəkil 2: Python kitabxanaları

Sonra ilk olaraq yuxarıda təqdim olunan Excel faylı Python-nın Pandas kitabxanası vasitəsilə oxunur.

Read Data

```
df = pd.read_excel("The Best Data Set_1000.xlsx")
```

```
df.head(5)
```

9]:

	Category	Title	Description
0	Maraqlı	Sinir ,oynaq , sinir bel ağrılarına 3 gündə son !	ŞOK ! ŞOK ! ŞOK ! Xanımlar və bəylər , bel və ...
1	Maraqlı	Oxşar məhsullar fotoşəkillərlə axtarılacaq	"eBay" internet mağazasının əlavələrində foto...
2	Maraqlı	Kiberidmançıların minimal əmək haqları artır	Kiberidmançı peşəsi ildən-ilə nüfuz qazanır. ...
3	Maraqlı	Məşhur otel rüsvay oldu - VİDEO - FOTO	Kanar arxipelaqındakı Qran-Kanariya adasında,...
4	İdman	"Qarabağ" Çempionlar Liqasında tarixi oyuna ç...	Bu gün "Qarabağ" komandası Çempionlar Liqası...

```
df.columns
```

7]: `Index(['Category', 'Title', 'News_Article'], dtype='object')`

Şəkil 3: Məlumatların oxunması və top 5 data

Daha sonra title və description sütunları birləşdirilərək data üzərində təmizlənmə işləri gedir. Stopwords yeni əhəmiyyəti az olan sözlər silinir, rəqəmlər silinir, boşluqlar yox edilir, və mətin məlumatı kiçik şiriftə keçirilir

```
df['Merge'] = df[[df.columns[1], df.columns[2]]].agg(' '.join, axis=1)
df['Merge'] = df['Merge'].str.replace('[^\w\s]','')
df['Merge'] = df['Merge'].str.replace('\d+', '')
df['Merge'] = df['Merge'].str.replace('  ','')
df['Merge'] = df['Merge'].str.strip()
df['Merge'] = df['Merge'].str.lower()
```

```
df['Merge'].head(5)
```

```
']: 0    sinir oynaqsinir bel ağrılarınagündə sonşokşok...
    1    oxşar məhsullar fotoşəkillərlə axtarılacaqebay...
    2    kiberidmançıların minimal əmək haqları artırki...
    3    məşhur otel rüsvay olduvideofotokanar arxipel...
    4    qarabağ çempionlar liqasında tarixi oyuna çıxı...
    Name: Merge, dtype: object
```

Şəkil 4: Datanın təmizlənmə prosessi

Datanın 20 % test və 80 % train üçün ayrılır və oxunan data hərdəfə qarışdırılır.

Split X & y

```
X = df[df.columns[3]]
y = df[df.columns[0]]
```

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, shuffle=True)
```

Şəkil 5: Test və Train dataya ayrılma

TF-IDF Vectorizer və Count Vectorizer mətni vektorlaşdırmaq üçün təbii dil emalında istifadə olunan hər iki üsuldur. CountVectorizer sadəcə olaraq bir sözün sənəddə neçə dəfə göründüyünü hesablayır (bagofwords yanaşmasından istifadə etməklə), TF-IDF Vectorizer isə sözün sənəddə yalnız neçə dəfə göründüyünü deyil, həm də sözün bütün korpus üçün nə qədər vacib olduğunu nəzərə alır.

CountVectorizer

```
# Define a CountVectorizer and use it to get bag of words encoded features
countVectorizer= CountVectorizer()
X_countVectorizer = countVectorizer.fit_transform(X)
```

Train & Test with Count Vectorizer

```
X_train_CountVectorizer = countVectorizer.fit_transform(X_train)
X_test_CountVectorizer = countVectorizer.transform(X_test)
```

TF/IDF Vectorizer

```
tfidfVectorizer = TfidfVectorizer()
X_tfidfVectorizer = tfidfVectorizer.fit_transform(X)
```

Train & Test with TF-IDF Vectorizer

```
X_train_tfIdf = tfidfVectorizer.fit_transform(X_train)
X_test_tfIdf =  tfidfVectorizer.transform(X_test)
```

Şəkil 6: Counter & TF-IDF Vektorizer

Hər iki metoda görə ayrılmış data SVM algorithminə ötürülür. SVM (Counter Vec) 68 % dəqiqlik, digər SVM (TF-IDF Vec) isə 76 % dəqiqlik göstərir.

Models without Hyperparameter

Support Vector Machine with CountVectorizer

```
model = LinearSVC()
model.fit(X_train_CountVectorizer, y_train)

y_pred = model.predict(X_test_CountVectorizer)
```

```
C:\Users\Elnur\anaconda3\lib\site-packages\sklearn\svm\_base.py:985: ConvergenceWarning: Liblinear failed to
ase the number of iterations.
  warnings.warn("Liblinear failed to converge, increase "
```

```
print("Accuracy Score: ", round(accuracy_score(y_test, y_pred),2))
print()

print("Confusion Matrix", confusion_matrix(y_test, y_pred))
print()

print("Classification Report: ", classification_report(y_test, y_pred))
```

Şəkil 7: SVM model Counter Vektorizerlə

```
Accuracy Score:  0.68

Confusion Matrix [[22 13  0  3  1  0]
 [11 32  0  4  1  1]
 [ 1  1 11  4  0  0]
 [ 4  4  1 19  0  2]
 [ 0  2  0  1 26  0]
 [ 3  4  0  1  1 27]]

Classification Report:                precision    recall  f1-score   support

       Dünya       0.54      0.56      0.55        39
     Maraqlı       0.57      0.65      0.61        49
  Mədəniyyət       0.92      0.65      0.76        17
     Siyasət       0.59      0.63      0.61        30
       İdman       0.90      0.90      0.90        29
İqtisadiyyat       0.90      0.75      0.82        36

    accuracy                           0.69       200
   macro avg       0.74      0.69      0.71       200
weighted avg       0.70      0.69      0.69       200
```

Şəkil 8: SVM model Counter Vektorizerlə nəticəsi

Support Vector Machine with TF-IDF Vectorizer

```
model = LinearSVC()
model.fit(X_train_tfIdf, y_train)

y_pred = model.predict(X_test_tfIdf)
```

```
print("Accuracy Score: ", round(accuracy_score(y_test, y_pred),2))
print()

print("Confusion Matrix", confusion_matrix(y_test, y_pred))
print()

print("Classification Report: ", classification_report(y_test, y_pred))
```

Şəkil 9: SVM model TF-IDF Vektorizerlə

```
Accuracy Score:  0.76

Confusion Matrix [[24 11  0  2  1  1]
 [ 7 37  0  3  0  2]
 [ 1  2 12  2  0  0]
 [ 4  2  1 21  0  2]
 [ 0  1  0  0 28  0]
 [ 1  3  0  2  1 29]]

Classification Report:                precision    recall  f1-score   support

       Dünya       0.65      0.62      0.63        39
     Maraqlı       0.66      0.76      0.70        49
  Mədəniyyət       0.92      0.71      0.80        17
     Siyasət       0.70      0.70      0.70        30
       İdman       0.93      0.97      0.95        29
İqtisadiyyat       0.85      0.81      0.83        36

    accuracy                           0.76       200
   macro avg       0.79      0.76      0.77       200
weighted avg       0.76      0.76      0.76       200
```

Şəkil 10: SVM model TF-IDF Vektorizerin nəticəsi

Artificial Neural Network with Count Vecterizer

```
# MLP Model
model = MLPC(max_iter = 10)
model.fit(X_train_CountVectorizer, y_train)

y_pred = model.predict(X_test_CountVectorizer)
print('MLP accuracy on test set:', accuracy_score(y_test, y_pred))
```

```
MLP accuracy on test set: 0.77
```

```
C:\Users\Elnur\anaconda3\lib\site-packages\sklearn\neural_network\_multilayer_perceptron.py:614: ConvergenceWarning: Stochas
tic Optimizer: Maximum iterations (10) reached and the optimization hasn't converged yet.
  warnings.warn(
```

Artificial Neural Network with TF-IDF Vecterizer

```
# MLP Model
model = MLPC(max_iter = 10)
model.fit(X_train_tfIdf, y_train)

y_pred = model.predict(X_test_tfIdf)
print('MLP accuracy on test set:', accuracy_score(y_test, y_pred))
```

```
MLP accuracy on test set: 0.755
```

```
C:\Users\Elnur\anaconda3\lib\site-packages\sklearn\neural_network\_multilayer_perceptron.py:614: ConvergenceWarning: Stochas
tic Optimizer: Maximum iterations (10) reached and the optimization hasn't converged yet.
  warnings.warn(
```

Şəkil 11: ANN model Counter & TF-IDF Vektorizerlə

Multinominal Naive Bayes with Count Vecterizer

```
mnb = MultinomialNB(alpha=1.0)
mnb.fit(X_train_CountVectorizer, y_train)

y_pred = mnb.predict(X_test_CountVectorizer)
print('MLP accuracy on test set:', accuracy_score(y_test, y_pred))
```

```
MLP accuracy on test set: 0.76
```

Multinominal Naive Bayes with TF-IDF Vecterizer

```
mnb = MultinomialNB(alpha=1.0)
mnb.fit(X_train_tfIdf, y_train)

y_pred = mnb.predict(X_test_tfIdf)
print('MLP accuracy on test set:', accuracy_score(y_test, y_pred))
```

```
MLP accuracy on test set: 0.665
```

Şəkil 12: BN model Counter & TF-IDF Vektorizerlə

Artificial Neural Network with Count Vectorizer

```
NN = MLPC(max_iter = 10)

hparams = {
    'hidden_layer_sizes': [(25,25,25), (30,25,30)],
    'activation': ['tanh', 'relu'],
    'solver': ['sgd', 'adam'],
    'alpha': uniform(0.0001, 0.05),
    'learning_rate': ['constant','adaptive']}

#reference to https://panjeh.medium.com/scikit-learn-hyperparameter-optimization-for-mlpclassifier-4d670413042b
```

```
CountVectorizer_NN = RandomizedSearchCV(NN, hparams, cv = 2)
CountVectorizer_NN.fit(X_train_CountVectorizer, y_train)
```

```
predictions_NN = CountVectorizer_NN.predict(X_test_CountVectorizer) #predicting the vectorized features of test dataset
```

```
print("Accuracy Score: ", round(accuracy_score(y_test,predictions_NN ), 2))
print()

print("Confusion Matrix", confusion_matrix(y_test, predictions_NN))
print()

print("Classification Report: ", classification_report(y_test, predictions_NN))
print()

print("The best Parameters found in NN using CV :{}".format(CountVectorizer_NN.best_params_))
print("The best Score NN with using CV:{}".format(CountVectorizer_NN.best_score_))
```

Şəkil 13: ANN model Counter Vektorizerlə (hiperparametrlə)

```
Accuracy Score:  0.79

Confusion Matrix [[24  5  0  0  1  0]
 [14 29  0  2  0  4]
 [ 0  2 10  0  0  0]
 [ 4  2  1 29  0  1]
 [ 0  0  0  0 33  0]
 [ 0  2  1  3  0 33]]

Classification Report:                precision    recall  f1-score   support

       Dünya       0.57      0.80      0.67        30
     Maraqlı       0.72      0.59      0.65        49
  Mədəniyyət       0.83      0.83      0.83        12
     Siyasət       0.85      0.78      0.82        37
       İdman       0.97      1.00      0.99        33
İqtisadiyyat       0.87      0.85      0.86        39

    accuracy                           0.79       200
   macro avg       0.80      0.81      0.80       200
weighted avg       0.80      0.79      0.79       200


The best Parameters found in NN using CV :{'activation': 'tanh', 'alpha': 0.033722963250030526, 'hidden_layer_sizes': (30, 2
5, 30), 'learning_rate': 'constant', 'solver': 'adam'}
The best Score NN with using CV:0.7024999999999999
```

Şəkil 14: ANN model Counter Vektorizerin nəticəsi (hiperparametrlə)

Artificial Neural Network with TF/IDF Vectorizer

```
NN = MLPC(max_iter = 10)

hparams = {
    'hidden_layer_sizes': [(25,25,25), (30,25,30)],
    'activation': ['tanh', 'relu'],
    'solver': ['sgd', 'adam'],
    'alpha': uniform(0.0001, 0.05),
    'learning_rate': ['constant','adaptive']}
```

```
TFIDF_NN = RandomizedSearchCV(NN, hparams, cv = 2)

TFIDF_NN.fit(X_train_tfIdf, y_train)
```

```
predictions_NN_TFIDF = TFIDF_NN.predict(X_test_tfIdf)
```

```
print("Accuracy Score: ", round(accuracy_score(y_test,predictions_NN_TFIDF ), 2))
print()

print("Confusion Matrix", confusion_matrix(y_test, predictions_NN_TFIDF))
print()

print("Classification Report: ", classification_report(y_test, predictions_NN_TFIDF))
print()

print("The best Parameters found in NN using TF/IDF :{}".format(TFIDF_NN.best_params_))
print("The best Score NN with using TF/IDF:{}".format(TFIDF_NN.best_score_))
```

Şəkil 15: ANN model TF-IDF Vektorizerlə (hiperparametrlə)

```
Accuracy Score:  0.61

Confusion Matrix [[ 6 23  0  0  1  0]
 [ 3 45  0  0  0  1]
 [ 0 12  0  0  0  0]
 [ 0 12  0 19  0  6]
 [ 0  6  0  0 27  0]
 [ 0 12  0  2  0 25]]

Classification Report:                precision    recall  f1-score   support

       Dünya       0.67      0.20      0.31        30
     Maraqlı       0.41      0.92      0.57        49
  Mədəniyyət       0.00      0.00      0.00        12
     Siyasət       0.90      0.51      0.66        37
       İdman       0.96      0.82      0.89        33
İqtisadiyyat       0.78      0.64      0.70        39

    accuracy                           0.61       200
   macro avg       0.62      0.52      0.52       200
weighted avg       0.68      0.61      0.59       200


The best Parameters found in NN using TF/IDF :{'activation': 'tanh', 'alpha': 0.025955466505755483, 'hidden_layer_sizes': (2
5, 25, 25), 'learning_rate': 'adaptive', 'solver': 'adam'}
The best Score NN with using TF/IDF:0.3125
```

Şəkil 16: ANN model TF-IDF Vektorizerlə (hiperparametrlə)

Nəticələr aşaqdakı cədvəldə əks olunmuşdur:

	Modelin adı	Count Vektorizer	TF-İDF Vectorizer
Hiperparametrsiz	SVM	68 %	76 %
	ANN	77 %	75.5 %
	NB	76 %	66.5 %
Hiperparametrlə	SVM	70 %	76 %
	ANN	79 %	61 %

Cədvəl 1: Modellin Accuracy Cədvəli

Beləliklə yuxardakı cədvələ baxaraq deyə bilərik ki, min xəbər datasının mətinin konteksə görə sinifləndirilməsi üçün ən yaxsı intellektual meted hiperparametrlə Süni Neyron Şəbəkəsi (ANN) – dir. Bu motod 79% dəqiqliklə o biri metodlardan irəlidədi.

NƏTICƏ

Nəticə olaraq, müəssisələr hər gün böyük həcmdə strukturlaşdırılmamış mətnlə məşğul olurlar. Böyük miqdarda mətn məlumatlarını təhlil etməyə gəldikdə, bu, əl ilə etmək çox böyük bir işdir. Həm də yorucu, vaxt aparan və bahalıdır. Böyük həcmdə məlumatların əl ilə çeşidlənməsi daha çox səhvlərə və uyğunsuzluqlara səbəb olur. Üstəlik, yaxşı ölçülənmir. Süni intellektlə idarə olunan mövzu təhlili strukturlaşdırılmamış məlumatları təhlil etməyi asanlaşdırır, daha sürətli və daha dəqiq edir. Bundan əlavə, mətn təsnifatı tədqiqatı əhəmiyyətli vədlər verir və son illərdə əhəmiyyətli irəliləyişlər əldə etmişdir. Mətn təsnifatının praktiki əhəmiyyəti onun böyük həcmdə mətn məlumatlarını səmərəli şəkildə təşkil etmək və strukturlaşdırmaq, manuel prosesləri avtomatlaşdırmaq, axtarış imkanlarını artırmaq və fərdiləşdirilmiş məzmun tövsiyələrini işə salmaq qabiliyyətində aydındır.

Dissertasiya işi girişdən, üç fəsildən, nəticə və ədəbiyyat siyahısından ibarətdir.

Dissertasiya işinin birinci fəsli üç yarımfəsildən ibarətdir.

Birinci yarımfəsildə Süni İntelekt və onun tarixi haqqında məlumat verilmişdir.

İkinci yarımfəsildə Süni İntellektin önəmi və istifadə sahələri haqqında məlumat verilmişdir.

Üçüncü yarımfəsildə Süni intellektin həyata keçirilməsi üçün strategiyalar verilmişdir.

Dissertasiya işinin ikinci fəsli üç yarımfəsildən ibarətdir.

Birinci yarımfəsildə təbii dil emalı (NLP) giriş haqqında ətraflı məlumatla təmin olunmuşdur.

İkinci yarımfəsildə mövzu təhlili, onun əhatə dairəsi, önəmi və işləmə prinsipi haqqında məlumatlar verilmişdir.

Üçüncü yarımfəsildə mövzu təhlilinin metodlarına toxunulmuşdur.

Dissertasiya işinin üçüncü fəsli üç yarımfəsildən ibarətdir.

Birinci yarımfəsildə mətinlərin kontekstə görə sinifləndirilməsi problemləri və

onların tətbiqi istiqamətləri haqqında danışılmışdır.

İkinci yarımfəsildə mətinlərin kontekstə görə sinifləndirilməsinin işlənilmə instrumental vasitələri haqqında məlumat verilmişdir.

Üçüncü yarımfəsildə Mətinlərin kontekstə görə sinifləndirilməsinin həyata keçirilməsi prosesi təhlil olunmuşdur.

ƏDƏBİYYAT

1. Artificial Intelligence: What It Is and How It Is Used. (2023, April 24). Retrieved from https://www.investopedia.com/terms/a/artificial-intelligence-ai.asp

2. What is Machine Learning? | IBM. (n.d.). Retrieved from https://www.ibm.com/topics/machine-learning

3. What is Deep Learning? | IBM. (n.d.). Retrieved from https://www.ibm.com/topics/deep-learning

4. Lecture Notes | Advanced Natural Language Processing | Electrical Engineering and Computer Science | MIT OpenCourseWare. (n.d.). Retrieved from https://ocw.mit.edu/courses/6-864-advanced-natural-language-processing-fall-2005/pages/lecture-notes/

5. An Introduction to Latent Semantic Analysis (Landauer, Foltz and Laham, D., 1998)

6. Text Classification: What It Is & How to Get Started. (n.d.). Retrieved from https://levity.ai/blog/text-classification

7. Text categorization with Support Vector Machines: Learning with many relevant features (Joachims, 1998)

8. An empirical study of the naive Bayes classifier (Rish, 2001)

9. Artificial Neural Networks and its Applications - GeeksforGeeks. (2020, June 24). Retrieved from https://www.geeksforgeeks.org/artificial-neural-networks-and-its-applications/