UOT 004.056 DOI: 10.25045/jpit.v12.i2.08

Abdullayeva F.C.[1], Ocaqverdiyeva S.S.[2]
[1,2]AMEA İnformasiya Texnologiyaları İnstitutu, Bakı, Azərbaycan
[1]a_farqana@mail.ru**,** [2]allahverdiyevasabira@gmail.com

VULQARİZMLƏRİN MAŞIN TƏLİMİ ƏSASINDA AŞKARLANMASINA BİR YANAŞMA

Daxil olmuşdur: 14.06.2021 Düzəliş olunmuşdur: 22.06.2021 Qəbul olunmuşdur: 06.07.2021

Məqalədə veb-kontentlərdə vulqarizmlərin maşın təlimi əsasında aşkarlanması üçün bir yanaşma işlənmişdir. Veb-səhifələrdə zərərli məzmun daşıyan kontentlərin sayının artması zərərli məzmundan qorunma məsələsini aktuallaşdırır. İstifadəçilərin, əsasən də uşaq və yeniyetmələrin İnternetdə vulqarizmlərlə (qeyri-etik danışıq, jarqon ifadə, söyüş, təhqir və s.) qarşılaşması onların psixologiyasına öz mənfi təsirini göstərir. Həm onlayn mediada, həm də sosial mediada (Twitter və Facebook və s.) vulqar söz, söz birləşməsi və ifadələrin aşkarlanması üçün daha etibarlı avtomatik mətn aşkarlama metodlarının inkişaf etdirilməsi bu problemin həlli üçün çox böyük əhəmiyyət daşıyır. Təqdim olunan məqalədə N-grams+TF-IDF əlamətlərindən istifadə etməklə vulqarizmlərin aşkarlanması üçün yanaşma təklif edilmişdir. Burada əvvəlcədən məlum olan vulqar sözlərə N-gram+TF-IDF əsaslı əlamətlərin çıxarılması üsulu tətbiq olunaraq ədədi vektorlar generasiya olunmuşdur. Generasiya edilmiş ədədi vektor Naive Bayes alqoritmlərinin girişinə ötürülmüşdür. Müxtəlif əlamətlərdən istifadə etməklə aparılan eksperimentlərin nəticəsində unigram+TF-IDF əlamətləri əsasında klassifikasiya daha üstün nəticələr vermişdir. Vulqarizmlərin aşkarlanması üçün təklif edilən bu yanaşma uşaq və yeniyetmələrin danışıq mədəniyyətinin və insanlarla ünsiyyətinin formalaşmasında əhəmiyyətlidir. Bu yanaşma uşaqların İnternetdən əldə edilən zərərli məzmundan qorunmasında faydalıdır və uşaq təhlükəsizlik mərkəzlərində, təhsil sistemində istifadə edilə bilər.

Açar sözlər: *vulqarizmlər, N-grams, TF-IDF, Naive Bayes, İnternetdə uşaqların təhlükəsizliyi.*

Giriş

İnternet dövlətin və cəmiyyətin üzləşdiyi sosial, texnoloji problemləri və aktual hadisələri özündə əks etdirən informasiya mənbəyidir. Son dövrlərlərdə ən aktual problemlərdən biri İnternet mühitində zərərli məlumatların qarşısının alınması məsələsidir.

Məlumdur ki, adi danışıq dilində jarqon sözlərdən, nifrət nitqi və s. istfadə olunur. Eyni zamanda belə danışıq üsulu media məkanında, xüsusilə də onlayn mediada sərbəst istifadə edilir. Sosial şəbəkələrdə, forumlarda, bloqlarda, mikrobloqlarda və digər rəy mənbələrində nifrət dolu sözlər, söyüş ifadələri, terror, qəddarlıq və s. məzmunlu fikirlərlə zəngin resursların sayı artmaqda davam edir. Bunun nəticəsində, çox vaxt istifadəçilər arasında kiberqarşıdurmalar baş verir. İnternet istifadəçilərinin bir hissəsi cəmiyyətin həssas üzvü olan uşaq və yeniyetmələrdir. Buna görə də belə halların qarşısının alınması, yaxud da yaranmış mənfi nəticələrin aradan qaldırılması problemi nüfüzlü təşkilatların, mütəxəssislərin və əlaqədar insanların qarşısında yeni vəzifələr qoyur [1].

Vulqarizm latın sözü olub, *vulgaris*, yəni, "sadə xalq" mənası verən, ədəbi dil normalarına zidd olan qaba (qeyri-ədəbi) söz və ya ifadələri bildirmək üçün istifadə olunan beynəlxalq termindir [2]. Söyüş, təhqir, qarğış, insanların bir sıra mənfi xüsusiyyətlərini ifadə etmək, onları alçaltmaq, lağa qoymaq məqsədilə işlədilən bir sıra sözlər vulqar leksikaya aid edilir [3]. Hər bir xalqın özünəməxsus olan vulqarizmləri vardır və bunlar bəzən digər xalqların vulqarizmlərindən ciddi şəkildə fərqlənir. Ona görə də vulqarizmlərlə mübarizə apararkən bu reallıqlar nəzərə alınmalıdır. Belə ki, bu problemlərin həlli üçün beynəlxalq səviyyədə təşəbbüs irəli sürülməli, ayrı-ayrı xalqların vulqarizmlərdən ibarət bazaları işlənməli və bu bazaların qarşılıqlı inteqrasiyası təmin olunmalıdır ki, hamı üçün əlçatan olsun. Vulqarizmlərlə mübarizə məqsədilə işlənən intellektual texnologiyalar milli tələbləri nəzərə almalıdır.

Vulqarizmlərin fəsadları ilə mübarizə aparmaq üçün Avropa Birliyi Komissiyası tərəfindən müxtəlif tədbirlər, o cümlədən qanunvericilik tədbirləri görülmüşdür. Avropa Birliyi Komissiyası sosial media şəbəkələrinin vulqar sözlərin 24 saat ərzində silinməsi kodeksini imzalamasını onların qarşısında tələb kimi qoymuşdur [4].

Avropa Şurasının "Uşaqların zərərli kontentdən qorunması" hesabatında uşaqlar üçün nəzərdə tutulmuş ziyanlı informasiyanın növlərindən biri kimi "Senzuradan kənar ifadələr (söyüş, jarqon, təhqiramiz ifadələr və s.)" göstərilir [5].

Azərbaycan qanunvericiliyində "....uşaqlar üçün nəzərdə tutulmayan və uşağın fiziki və ya psixi sağlamlığına mənfi təsir göstərən, o cümlədən təhrif olunmuş sosial təsəvvürlərin formalaşdırılması vasitəsilə onların mənəvi, psixi, fiziki və ya sosial inkişafında pozuntu yaranmasına səbəb olan məlumat" uşaqlar üçün zərərli informasiya hesab olunur [6].

Sosial şəbəkələrdə və mikrobloqlarda qeyri-etik danışıq və nifrət nitqinin istifadəsinin qadağan edilməsinə baxmayaraq, bu şəbəkələrin və saytların böyük olması onlara nəzarəti imkansız edir. Belə ki, arzuolunmayan və zərərli hesab edilən məlumatlar daha çox sosial şəbəkə istifadəçiləri arasında yayılır. Bu reallığı nəzərə alaraq, tədqiqatçılar tərəfindən sosial şəbəkələrdə zərərli təsirlərə qarşı monitorinq aparılmış və vahid sistem hazırlanmışdır [7]. *Twitter* sosial şəbəkəsində nifrət nitqi ifadə edən məzmunun aşkarlanması üçün yanaşma təklif edilmişdir [8].

Aparılan araşdırmalara əsasən, məlum olur ki, İnternet mühitində nifrət nitqi, etik normadan kənar danışıq, leksikonu korlayan, vulqar sözlərdən ibarət təhqiramiz mətnlərin olması istifadəçilərə, xüsusilə də uşaq auditoriyasına öz mənfi təsirini göstərir.

Məqalədə uşaqların İnternetdə olduqları vaxt qeyri-etik sözlərlə qaşılaşmasının qaşısını almaq üçün vulqarizmlərin mətndən çıxarılmasını həyata keçirən maşın təlimi əsasında bir yanaşma təklif olunur.

Motivasiyalar

Mətnlərdə vulqarizmlərin avtomatik aşkarlanması sahəsində bir çox üsullar təklif edilmişdir [9–11]. Təklif edilən yanaşmalarda müxtəlif əlamətlərin emalı üsullarından və maşın təlimi alqoritmlərindən istifadə edilmişdir.

Vulqarizmlərin klassifikasiyası üsulları istifadə etdikləri əlamətlərin çıxarılması yanaşmalarına görə fərqlənirlər. Vulqar sözlərin lüğətə əsaslanan, söz çantası (*ing. Bag-of-words, BOW*), *n-gram*, *TF-IDF* və dərin təliməsaslı aşkarlanması üsulları vardır.

Lüğətə əsaslanan üsulun çatışmazlığı ondadır ki, bu metod axtarışı həyata keçirmək üçün predmet sahəsinə aid lüğətin olmasını tələb edir. Bu çatışmazlığı aradan qaldırmaq üçün *BOW* əsaslı yanaşma istifadə olunur. *BOW* əsaslı yanaşma lüğətə əsaslanan yanaşma ilə oxşardır, lakin fərq ondan ibarətdir ki, *BOW* yanaşmasında sözün əlamətləri əvvəlcədən təyin edilmiş lüğətdən deyil, təlim verilənlərindən əldə olunur. *BOW* üsulunun çatışmazlığı odur ki, bu metodda söz sırası nəzərə alınmır və ayrı-ayrı sözlər müxtəlif kontekstlərdə istifadə olunduqda, səhv klassifikasiyaya səbəb olur. *BOW* üsulunun bu məhdudiyyətini aradan qaldırmaq üçün *n-gramlardan* istifadə edirlir. Bəzi hallarda sözlərin ardıcıllığının nəzərə alınması vacib hesab olunur. *BOW* üsulu *n-gram* üsulunun n=1 (*unigram*) olduğu halla eynilik təşkil edir. *N-gramlar* sözlərin birlikdə istifadə edildiyi kontekstləri qeydə alır. *N-gram* üsulu *BOW* üsulundan daha üstün hesab olunur. Lakin *n-gram* üsulunun çatışmazlığı odur ki, n-nin qiymətinin artırılması emal sürətinin azalmasına səbəb olur.

Sənədlərin əlamətlər vektoru şəklində təsvir edilməsi üsullarından biri *TF-IDF*-dir (*TF — term frequency, IDF — inverse document frequency*). *TF-IDF* bu və ya digər sözün hər hansı bir sənəd və ya bütünlükdə sənədlər korpusuna nəzərən vaciblik dərəcəsini göstərir. Son zamanlar dərin təlim tipli *NLP* (*Natural Language Processing*) üsulu istifadə edilir. Əlamətlərin çıxarılmasında *word2vec* üsulunun tətbiqi zamanı böyük ölçülü verilənlərin olması tələb olunur. Bu üsul kiçik ölçülü verilənlər üzərində zəif nəticələr göstərir. Yuxarıda göstərilən problemlərin aradan qaldırılması məqsədi ilə təqdim olunan məqalədə vulqarizmlərin effektiv klassifikasiyası üçün *n-gram+TF-IDF* hibridləşmiş əlamətlərdən istifadə edilmişdir. Burada *n-gramlardan* istifadə

edilməsi ayrı-ayrı sözlərin, həmçinin söz birləşmələrinin semantikasını nəzərə alaraq klassifikasiyasını həyata keçirməyə imkan verir. *TF-IDF* alqoritmi vasitəsi ilə hesablanmış ədədi qiymətlər isə sözlərin çəkiləndirilməsini həyata keçirir.

Əlaqəli işlər

İnternetdə rast gəlinən onlayn məzmunda təhqiramiz ifadələrin çoxalması onların aşkarlanması və qarşısının alınması məqsədilə yeni üsul və mexanizmlərin yaranmasına təkan verir. Virtual mühitdə nifrət nitqi, təhqir, dini nifrət, şiddət və s. məzmun daşıyan veb-səhifələrin müəyyən edilməsi və onların klassifikasiyası probleminin həlli üçün müxtəlif maşın təlimi üsullarından istifadə olunur. Bu yanaşmaların əksəriyyətində məsələnin həlli mətndən əlamət çıxarılmaqla həyata keçirilir [12]. Bəzi tədqiqatlarda lüğətlərdən [13, 14] və söz çantalarından istifadə edilir. [15]-də isə *n-gram* modellərlə əlaqəli daha mürəkkəb hesablamaları və performansla bağlı qeyri-müəyyənliyi nəzərə alaraq *BOW* üsulundan istifadə etməyə üstünlük verilib. Digər bir işdə mətnlərdən əlamətlər çıxarmaq üçün *BOW* və *TF-IDF* istifadə edilir [16].

Sosial mediada təhqiramiz ifadələrin aşkarlanması üçün Facebook verilənlər bazasından (poçt və şərh məlumatları) istifadə etməklə tvitlərdə sadə *n-gram* və *char n-gram* əlamətləri ilə maşın təlimi yanaşmasından istifadə edilir [17]. Digər işdə *Instagram* şərhlərinin içərisində insanların duyğularını təyin etmək məqsədi ilə jarqonəsaslı lüğət klassifikatoru təqdim edilir [18].

Cümlənin sintaktik quruluşundan asılı olmayaraq, leksik xüsusiyyətlər eyni sözlərin fərqli sırada olan cümlələrin daxilində hansı məna (məs., hücum, təhqir və s.) daşımasını ayırd edə bilmir. Ədəbiyyat [19]-da sözün zərərli məzmun daşımasını qiymətləndirməklə, onu əvvəlcədən müəyyənləşdirən dərin təlim modeli verilir.

Mətndə istifadə olunan nifrət nitqinin hədəflərini və intensivliyini dəqiq müəyyənləşdirmək üçün leksik əlamətlərdən (isim, sifət və fel) istifadə olunur. Beləki, sosial şəbəkə (Facebook) verilənlərindən istifadə etməklə verilənlər bazası yaradılır. Bu verilənlər üzərində aparılan eksperimentlər zamanı *TF-IDF*, char quad-gram, word unigram və leksik xüsusiyyətlərindən istifadə olunur. Maşın təlimi metodlarından istifadə etmək *F1* –in yüksək nəticə göstərdiyi müəyyən edilir [20].

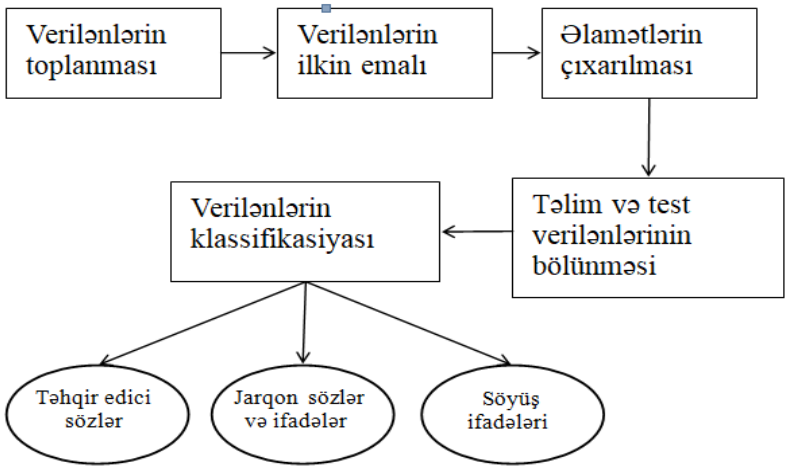
Sosial mediada təhqiramiz söz və ifadələrin aşkarlanmasında yalnız söz uyğunluğundan istifadə olunmasının kifayət etmədiyini nəzərə alaraq, təsnifat tapşırığını iki sinfə ayırmaq üçün leksik xüsusiyyətləri, *n-gram* xüsusiyyətləri, dil xüsusiyyətləri, sintaktik xüsusiyyətləri, əvvəlcədən hazırlanmış "*word2vec*" və "*comment2vec*" xüsusiyyətlərindən istifadə olunmuşdur [21].

Onlayn mühitdə bəzi istifadəçilər tərəfindən həm fərdi, həm də qrup şəklində şiddət və ya nifrəti təbliğ edən məzmun yerləşdirilir. İnsanların dini, cinsi və etnik mənsubiyyətini və ya onların əlil olmasını ələ salan və ya aşağılayan istifadəçilərə rast gəlmək mümkündür. [22]-də təklif edilən alqoritm *CNN* şəbəkəsi (*Convolutional Neural Networks*) arxitekturasından və *NLP*-dən birgə istifadə etməklə kontentin nifrət nisbətini müəyyənləşdirir və məzmunda ifadələrin qərəzliliyini göstərir.

Təklif olunan model

Vulqarizmlərin aşkarlanması üçün təklif edilən yanaşmanın modeli şəkil 1-də təsvir edilmişdir. Sözlərin üç müxtəlif siniflərdə ("təhqiredici sözlər", "jarqon sözlər və ifadələr", "söyüş ifadələri") klassifikasiyasını həyata keçirmək üçün təklif edilmiş sistemin arxitekturası aşağıda təsvir edilmişdir. Təklif edilmiş arxitektura üzrə hər bir blok aşağıdakı kimi təsvir edilir:

Verilənlərin toplanması. Tədqiqat işində eksperimentləri aparmaq üçün azərbaycan dilində işlədilən vulqarizmlərin bazası yaradılmışdır. Burada vulqarizmlər "təhqiredici sözlər", "jarqon sözlər və ifadələr", "söyüş ifadələri" adlı üç müxtəlif sinifdə nişanlanmışdır. Verilənlər bazasını 249 söz təşkil edir.



Şəkil 1. Vulqarizmlərin aşkarlanması sisteminin arxitekturası

Verilənlərin ilkin emalı. Mətnlərdə ilkin emalın aparılması klassifikasiyanın nəticələrini kifayət qədər yaxşılaşdıra bilir. Bu məqsədlə ilkin emal zamanı bütün küy sözlərin silinməsi aparılmışdır. Verilənlərin kompüterin başa düşəcəyi dilə konvertasiyası prosesi ilkin emal adlanır. İlkin emal üsullarından biri lazımsız verilənlərin filtrasiya edilməsidir. Təbii dilin emalında lazımsız sözlər küy sözlər adlanır. Küy sözlər ümumi istifadə olunan sözlərdir, məsələn "*the*", "*a*", "*an*", "*in*". Bu sözləri axtarış sistemləri həm axtarış zamanı giriş verilənlərinin indeksləşdirilməsi, həm də onların axtarış sorğusuna cavab olaraq tapılması zamanı nəzərə almır. Bu sözlər çox böyük emal vaxtı apardığı üçün onların silinməsi zəruri hesab edilir. Bunun üçün küy sözlər hesab etdiyimiz sözlərin siyahısını tərtib edərək onları verilənlər bazamızdan asanlıqla silə bilərik. Python proqram təminatının *NLTK (ing. Natural Language Toolkit*) kitabxanasında 16 müxtəlif dillərdə saxlanan küy sözlər artıq vardır. Bundan əlavə *tokenləşdirmə* və *stemming* əməliyyatı da aparıla bilər. *Tokenləşdirmə* bütün cümləni ayrı-ayrı sözlərə konvertasiya edir, *stemming* isə şəkilçiləri ataraq sözləri kök yazılış formalarına qaytarır. Məsələn, *chocolates --> chocolate, retrieval --> retrieve*. Təqdim olunan işdə cümlədəki sözləri ayrı-ayrı vahidlər şəklində təsvir etmək üçün mətnə *tokenləşdirmə* əməliyyatı tətbiq edilmişdir. İşdə söz şəkilçilərinin atılması məsələsinə baxılmadığı üçün *stemming* əməliyyatı aparılmamışdır.

Əlamətlərin çıxarılması. Alqoritmlərin girişinə mətn tipli verilənləri ötürməzdən əvvəl onların ədədi qiymətlərə çevrilməsi lazım gəlir. Mətnlərin klassifikasiyası sahəsində əsas addımlardan biri əlamətlərin çıxarılmasıdır. Məqalədə *n-gram* və *TF-IDF* strategiyaları istifadə edilərək əlamətlərin çıxarılması həyata keçirilmişdir.

TF sənəddəki hər bir sözün həmin sənəddəki bütün sözlərə nəzərən sayını hesablayır.

$$TF\ (w) = \frac{w\ sözünün\ sənəddə\ rast\ gəlmə\ sayı}{sənəddəki\ sözlərin\ \ ümumi\ sayı} \qquad (1)$$

IDF sözün sənədlər korpusunda işlənmə tezliyinə nəzərən onun vaciblik dərəcəsini göstərir. Bu meyar sözün korpus üçün nə dərəcədə vacib olduğunu qiymətləndirir.

$$IDF\ (w) = \frac{sənədlərin\ ümumi\ sayı}{hər\ bir\ sözün\ digər\ sənəddə\ rast\ gəlmə\ sayı} \qquad (2)$$

TF-IDF isə *TF* və *IDF* faktorlarının hasili şəklində hesablanır:

$$TF - IDF(w) = TF(w) \times IDF(w) \qquad (3)$$

Naive Bayes klassifikatoru. Klassifikator Bayes teoreminə əsaslanır. Tomas Bayesin 1812-ci ildə müəyyən etdiyi şərti ehtimal hesablama düsturudur. Alqoritm bir element üçün hər vəziyyətin ehtimalını hesablayır və ən yüksək ehtimal dəyərinə görə təsnif edir. Bayes teoremi klassifikasiya məqsədləri üçün istifadə olunarkən, mümkün çıxış vəziyyətləri arasında ən yüksək ehtimalı olan vəziyyət hədəf sinif olaraq seçilir. Bunun riyazi ifadəsi aşağıdakı kimi göstərilir [23, 24]:

$$P(C_i \setminus X) = \frac{P(X \mid C_i)P(C_i)}{P(X)} \tag{4}$$

Naive Bayes klassifikatoruna əsasən, hər hadisənin baş vermə ehtimalı (4)-ə uyğun hesablanır. Burada ən yüksək ehtimalı olan hadisə namizəd kimi qəbul edilir. *Naive Bayes* klassifikatoru iki əsas fərziyyəyə əsaslanır: 1) xüsusiyyətlərin hər biri müstəqildir; 2) hər bir xüsusiyyət eyni qabarıqlığa malikdir. *Naive Bayes* klassifikatoru natamam məlumatlara qarşı dayanıqlıdır və digər klassifikasiya metodlarından fərqli olaraq çox sadədir. Buna görə də təlim məlumatlarını yalnız bir dəfə keçmək yetərlidir və sadə verilənlərlə belə, yüksək nəticələr verir.

Biz verilənlər bazamızla əlaqədar olaraq *Naive Bayes* teoremini aşağıdakı şəkildə tətbiq edə bilərik. Fərz edək ki, $D = (X_1, X_2, ...X_n)$ verilənlər çoxluğu verilmişdir, hər bir nümunə $X_i = (x_1, x_2, ..., x_n)$ kimi ifadə olunur. D müəyyən xüsusiyyətlərə malikdir və $C = (C_1, C_2, ..Cm)$ siniflərindən təşkil olunmuşdur və $\{A_1, A_2, ...A_n\}$ atribut qiyməti ilə təsvir olunur. Hər təlim nümunəsinin xüsusi bir C_i etiketi vardır və $X \in D$. Burada *Naive Bayes* klassifikatoru vasitəsilə X -in C sinfinə aid olduğu proqnozlaşdırılır.

Eksperimentlər

Eksperimentlər Python proqramlaşdırma mühitində aparılmışdır. Eksperimentlərin aparılmasında hazırlanmış lüğətdən istifadə olunmaqla, tərtib edilmiş verilənlər bazası təhqiredici sözlər (0), jarqon sözlər və ifadələr (1), söyüş ifadələri (2) adlı üç sinfə ayrılmışdır.

Eksperimentlərin aparılması məqsədi ilə ümumi verilənlər bazasının 80%-i təlim verilənləri üçün, 20%-i test verilənləri üçün istifadə edilmişdir. Klassifikasiyanı həyata keçirmək üçün *MultinomialNB (Multinomial Naive Bayes), ComplementNB (Complement Naive Bayes), GNB (Gaussian Naive Bayes), BernoulliNB (Bernoulli Naive Bayes)* alqoritmləri istifadə edilmişdir. Klassifikasiyanın effektivliyi *Accuracy, Precision, Recall, F-Measure* metrikaları əsasında qiymətləndirilmişdir. Bu metrikaların riyazi təsviri aşağıda verilmişdir [25]:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{5}$$

$$Precision = \frac{TP}{TP + FP} \tag{6}$$

$$Recall = \frac{TP}{TP + FN} \tag{7}$$

$$F1 = 2 \times \frac{Precision * Recall}{Precision + Recall} \tag{8}$$

Burada TP (*True Positives*) – Həqiqi Pozitivlər; TN (*True Negatives*) –Həqiqi Neqativlər; FP (*False Positives*) –Yanlış Pozitivlər; FN (*False Negatives*) – Yanlış Neqativlər kimi işarələnmişdir.

Həqiqi Pozitivlər düzgün proqnozlaşdırılan müsbət dəyərlərdir. Belə ki, həm həqiqi sinfin, həm də proqnozlaşdırılan sinfin də dəyərinin "bəli" olması deməkdir.

Həqiqi Neqativlər düzgün proqnozlaşdırılan mənfi dəyərlərdir, burada həm faktiki sinfin, həm də proqnozlaşdırılan sinfin dəyərinin "yox" olması deməkdir.

Yanlış Pozitivlər Həqiqi sinif "xeyr" olmadığı və proqnozlaşdırılan sinif "bəli" olduqda, *Yanlış Neqativlər* isə həqiqi sinif "bəli" olduğu halda, lakin "xeyr" olaraq proqnozlaşdırılan sinif olduqda deməkdir.

Naive Bayes alqoritminin *unigrams*, *bigrams*, *trigrams* qiymətlərində aparılmış eksperimentlərin nəticələri cədvəl 1-ə daxil edilmişdir.

Cədvəl 1

Multinomial NB alqoritminin *unigrams*, *bigrams*, *trigrams* qiymətlərində nəticələri

Alqoritm	***N-gram range***	**Sinif**	***Accuracy***	***Precision***	***Recall***	***F1-score***
MultinomialNB	*unigrams*	Təhqiredici sözlər	**0.82**	0.82	1.00	0.90
		Jarqon sözlər və ifadələr	**0.96**	0.96	0.77	0.86
		Söyüş ifadələri	**0.97**	0.97	0.97	0.97
	bigrams	Təhqiredici sözlər	**0.53**	0.90	0.53	0.67
		Jarqon sözlər və ifadələr	**1.0**	0.54	1.00	0.70
		Söyüş ifadələri	**0.69**	0.91	0.69	0.78
	trigrams	Təhqiredici sözlər	**0.32**	0.32	0.85	0.47
		Jarqon sözlər və ifadələr	**1.0**	1.00	0.42	0.59
		Söyüş ifadələri	**0.52**	0.52	1.00	0.68

Cədvəl 1-dən göründüyü kimi, təklif edilmiş yanaşma *unigrams* qiymətində yaxşı nəticələr vermişdir. Digər hallarda alqoritmin göstəricisi olduqca zəifdir. Belə ki, *unigrams* qiymətində alqoritm "təhqiredici sözlər", "jarqon sözlər və ifadələr", "söyüş ifadələri" sinfindən olan sözləri uyğun olaraq 0.82, 0.96, 0.97 dəqiqliklə tanıya bilmişdir. *Bigrams* olduqda, *MultinomialNB* alqoritmi "təhqiredici sözlər" sinfindən olan sözləri 0.53 dəqiqliklə, "söyüş ifadələri" sinfindən olan sözləri isə 0.69 dəqiqliklə tanıya bilmişdir. *Trigrams* olduqda isə *MultinomialNB* alqoritmi verilmiş verilənlər bazası üzərində, demək olar ki, zəif nəticə göstərmişdir. Bu onunla əlaqədardır ki, bizim istifadə etdiyimiz verilənlər bazasında üç söz birləşməsindən ibarət cümlələrin sayı çox azdır. Bu üsulu böyük verilənlər bazalarına tətbiq etdikdə, gözlənilən nəticəni almaq mümkündür.

Naive Bayes alqoritminin müxtəlif nüvə funksiyalarında *unigrams* qiymətində klassifikasiya nəticələri cədvəl 2-yə daxil edilmişdir.

Cədvəl 2

Naive Bayes alqoritminin müxtəlif nüvə funksiyalarında *unigrams* qiymətində klassifikasiya nəticələri

Alqoritm	**Sinif**	***Accuracy***	***Precision***	***Recall***	***F1-score***
MultinomialNB	Təhqiredici sözlər	0.82	0.82	1.00	0.90
	Jarqon sözlər və ifadələr	0.96	0.96	0.77	0.86
	Söyüş ifadələri	0.97	0.97	0.97	0.97
ComplementNB	Təhqiredici sözlər	0.70	0.70	0.92	0.79
	Jarqon sözlər və ifadələr	0.94	0.94	0.82	0.87
	Söyüş ifadələri	1.0	1.00	0.88	0.94
GNB	Təhqiredici sözlər	0.82	0.82	0.87	0.84
	Jarqon sözlər və ifadələr	0.91	0.91	0.83	0.87
	Söyüş ifadələri	0.95	0.95	1.00	0.98
BernoulliNB	Təhqiredici sözlər	0.64	0.64	0.75	0.69
	Jarqon sözlər və ifadələr	0.79	0.79	0.74	0.76
	Söyüş ifadələri	1.00	1.00	0.88	0.94

Cədvəl 2-dən göründüyü kimi, ən pis nəticəni *BernoulliNB* klassifikatoru göstərmişdir. Bu alqoritm "təhqiredici sözlər" və "jarqon sözlər və ifadələr" siniflərindən olan sözləri çox aşağı aşkarlama dəqiqliyi ilə tanıya bilmişdir.

Eksperimentləri *N-gram+TF-IDF* əlamətləri əsasında apardıqda, *GNB* alqoritmi *unigrams* əlaməti ilə müqayisədə daha üstün nəticələr göstərmişdir. Belə ki, bu alqoritmin klassifikasiya nəticələrində kifayət qədər yaxşılaşma müşahidə edilmişdir (cədvəl 3).

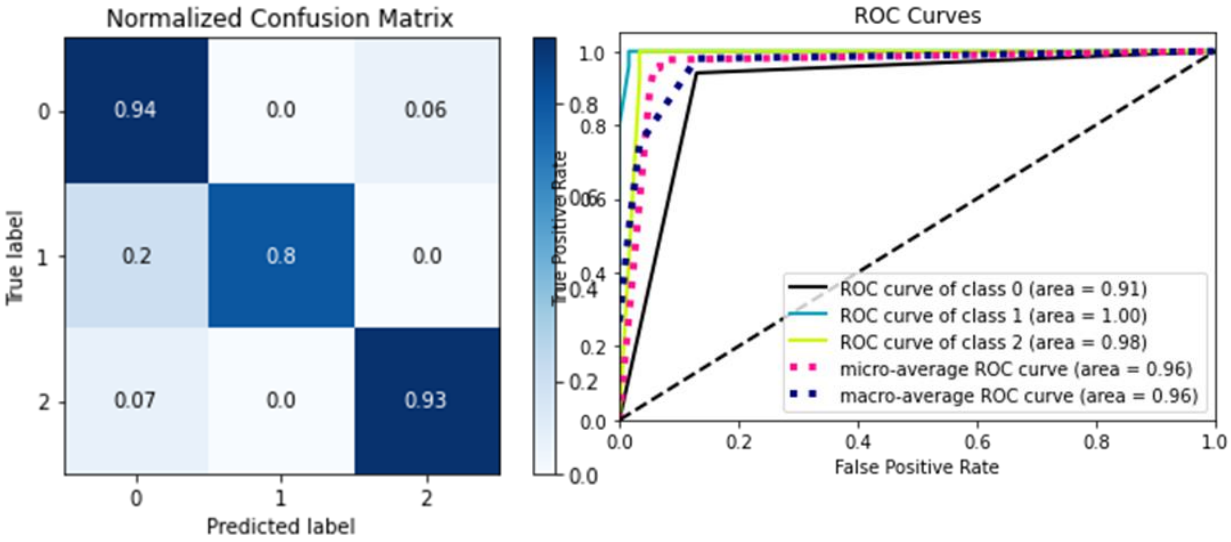
Cədvəl 3

GNB alqoritminin *N-gram+TF-IDF* əlamətləri əsasında klassifikasiya nəticələri

Alqoritm	***N-gram range***	**Sinif**	***Accuracy***	***Precision***	***Recall***	***F1-score***
GNB	*unigrams*	Təhqiredici sözlər	**0.82**	0.82	0.87	0.84
		Jarqon sözlər və ifadələr	0.91	0.91	0.83	0.87
		Söyüş ifadələri	0.95	0.95	1.00	0.98
GNB	***N-gram+TF-IDF unigrams***	Təhqiredici sözlər	**0.94**	0.94	0.82	0.88
		Jarqon sözlər və ifadələr	0.80	0.80	1.00	0.89
		Söyüş ifadələri	0.93	0.93	0.93	0.93

Cədvəl 3-dən göründüyü kimi, təklif edilmiş yanaşma əsasında aparılmış klassifikasiyanın nəticələri bütün siniflər üzrə yüksək olmuşdur. Belə ki, *GNB* alqoritmi *unigrams* olduqda, "təhqiredici sözlər" sinfindən olan sözləri 0.82 dəqiqliklə tanıdığı halda, *N-gram+TF-IDF* əlamətindən istifadə etdikdə, həmin sinif sözləri 0.94 dəqiqliklə tanımışdır. Bu isə gözlənilən müsbət nəticəni verir. .

Cədvəl 3-dəki nəticələrin vizual analizini aparmaq məqsədi ilə şəkil 3(a)-da onların xətalar matrisi və şəkil 3(b)-də isə *ROC* (*Receiver Operating Characteristic*) əyrisi təsvir edilmişdir. Eksperimentlərin nəticələrini hərtərəfli və etibarlı şəkildə təhlil etmək üçün *ROC* əyrisindən istifadə olunur. Burada hər bir nöqtə müəyyən bir qərar həddinə uyğun həm həssaslığı, həm də spesifikliyi təmsil edən cütdür. Davamlı diaqnostik dəyişən üçün kəsmə dəyəri artırıldıqda, həm həqiqi, həm də yanlış pozitivlərin nisbəti azalır. *ROC* əyri altındakı sahə, bir parametrin iki qrupu nə qədər yaxşı ayıra biləcəyinin göstərir [26].



a) b)

Şəkil 3. *N-gram+TF-IDF* əlamətləri əsasında *GNB* alqoritminin klassifikasiya nəticələrinin xətalar matrisi (a) və *ROC* əyrisi (b)

Şəkil 3 (a)-da verilmiş xətalar matrisindən göründüyü kimi, klassifikasiya zamanı matrisin bütün elementləri diaqonal üzrə yığılmışdır. Burada çox az sayda nöqtələrin yanlış tanınmasına yol verilmişdir. Bundan başqa, şəkil 3(b)-də *ROC* əyrinin təsvirində də əyani şəkildə görmək mümkündür ki, bu əyri verilənlər bazasının bütün sinifləri üzrə yüksək qiymətlər alaraq 1-ə çox yaxınlaşmışdır.

Nəticə

İnternet mühitində istifadə edilən söyüş, nifrət, təhqir və s. kimi qeyri-etik leksikon sosial narahatlığa səbəb olur və bu isə daha çox uşaq auditoriyasına mənfi təsir göstərir.

İşdə Azərbaycan dilində bir *unigram* lüğəti ilə birlikdə söyüş, təhqir və s. nümunələrdən istifadə etməklə, Azərbaycan dilində olan vurqarizmlərin klassifikasiyasını həyata keçirən maşın təliminə əsaslanan bir yanaşma təklif edilmişdir. Təklif etdyimiz metodun sınaqdan keçirilməsi zamanı eksperimentlər yüksək nəticə vermişdir, belə ki, yanaşma *N-gram+TF-IDF* əlamətindən istifadə etməklə hər üç sinfə aid olan vulqarizmləri 0.94 dəqiqliklə tanımışdır.

İnternetdə vulqarizmlərin geniş vüsət alması bu tip danışıq ifadələrinin avtomatik aşkarlanmasını öyrənmək üçün güclü motivasiya yaradır. Bu araşdırma uşaq və yeniyetmələrin İnternetdən əldə olunan zərərli informasiyadan qorunmasında faydalıdır və onların danışıq mədəniyyətinin, insanlarla ünsiyyətinin formalaşmasında əhəmiyyətlidir.

Ədəbiyyat

1. Alguliyev R.M., Ojagverdieva S.S. Conceptual Model of National Intellectucal System for Children Safety in Internet Environmen // International Journal of Computer Network and Information Security, 2019, vol.11, no.3, pp. 40–47.
2. Farajov R.A. Izahlı dilchilik lughati. Bakı, “Maarif”, 1969, 143 s.
3. Aliguliyev R.M., Jafarov Y. Global vulgarizm bazası / Informasiya tahlukasizliyi problemlari uzra I respublika elmi-praktiki konfransı, 2013, s. 60–62.
4. European Commission and IT Companies announce Code of Conduct on illegal online hate speech. https://ec.europa.eu/commission/presscorner/detail/en/IP_16_1937
5. European Commission, “Protection of personal data”, November 24, 2016. http://ec.europa.eu/justice/data-protection/,.
6. “Ushagların zararli informasiyadan gorunması haggında” Azarbayjan Respublikasının

Ganunu, http://e-qanun.gov.az/framework/40764
7. Kotenko I., Saenko I., Chechulin A., Desnitsky V., Vitkova L., Pronoza A. Monitoring and Counteraction to Malicious Influences in the Information Space of Social Networks / International Conference on Social Informatics, 2018, pp. 159–167.
8. Watanabe H., Bouazizi M., Ohtsuki T. Hate Speech on Twitter: A Pragmatic Approach to Collect Hateful and Offensive Expressions and Perform Hate Speech Detection / IEEE Access, 2018, vol.6, pp. 13825–13835.
9. Greevy E., Smeaton A.F. Classifying racist texts using a support vector machine / İn Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval, 2004, pp. 468–469
10. Gitari N.D., Zhang Z., Hanyurwimfura D., Jun L. A lexicon-based approach for hate speech detection // International Journal of Multimedia and Ubiquitous Engineering, 2015, vol.10, no.4, pp. 215–230.
11. Tulkens S., Hilte L., Lodewyckx E., Verhoeven B., Daelemans W. A dictionary-based approach to racism detection in dutch social media, 2016, pp. 2–8.
12. Burnap P., Matthew L.W. Cyber hate speech on twitter: An application of machine classification and statistical modeling for policy and decision making // Policy & Internet, 2015, vol.7, no.2, pp.223–242.
13. Hatebase, Available from: https://hatebase.org/
14. Wu L., Morstatter F., Liu H. SlangSD: building, expanding and using a sentiment dictionary of slang words for short-text sentiment classification // Lang Resources & Evaluation, 2018, vol. 52, pp. 839–852 . https://doi.org/10.1007/s10579-018-9416-0
15. Liu Sh., Forss T. New Classification Models for Detecting Hate and Violence Web Content / KDIR 2015 - 7th International Conference on Knowledge Discovery and Information Retrieval, 2015, pp. 487–895.
16. Sharif O., Hoque M.M., Kayes A.S.M., Nowrozy R., Sarker I.H. Detecting Suspicious Texts Using Machine Learning Techniques // Applied Science, 2020, vol. 10, no.18: 6527. https://doi.org/10.3390/app10186527
17. Ibrohim M. O., Budi I. A dataset and preliminaries study for abusive language detection in Indonesian social media// Procedia Computer Science, 2018, vol.135, pp. 222–229.
18. Aly E.S., van der Haar D.T. Slang-Based Text Sentiment Analysis in Instagram / Fourth International Congress on Information and Communication Technology, 2020, pp. 321–329.
19. Greevy E., Smeaton A.F. Classifying racist texts using a support vector machine / In Proceedings of the 27th Annual International Conference on Research and Development in Information Retrieval, 2004, pp. 468–469.
20. Aulia N., Budi I. Hate Speech Detection on Indonesian Long Text Documents Using Machine Learning Approach / Proceedings of the 2019 5th International Conference on Computing and Artificial Intelligence, 2019, pp. 164–169.
21. Nobata C., Tetreault J., Thomas A., Mehdad Y., Chang Y. Abusive language detection in online user content / International World Wide Web Conference Committee (IW3C2), 2016, pp.145–153.
22. Chaudhari A., Parseja A., Patyal A. CNN based Hate-o-Meter: A Hate Speech Detecting Tool / Third International Conference on Smart Systems and Inventive Technology, 2020, pp. 940–944.
23. Zhang H. The Optimality of Naive Bayes / Proceedings of 17th International Florida Artificial Intelligence Research Society Conference, 2004, pp. 562–567.
24. Rennie J. D., Shih L., Teevan J., Karger D.R. Tackling the poor assumptions of Naive Bayes text classifiers / Artificial Intelligence Laboratory Massachusetts Institute of Technology Cambridge Massachusetts USA Proceedings of the Twentieth International Conference on Machine Learning, 2003, pp.1–8.
25. Accuracy, Precision, Recall & F1 Score: Interpretation of Performance Measures,

https://blog.exsilio.com/all/accuracy-precision-recall-f1-score-interpretation-of-performance-measures/
26. Bewick V., Cheek L., Ball J. Statistics review 13: Receiver operating characteristic curves. CritCare 8, 508 (2004). https://doi.org/10.1186/cc3000

УДК 004.056

Абдуллаева Фаргана Д.[1], Оджагвердиева Сабира С.[2]
[1,2]Институт Информационных Технологий НАНА, Баку, Азербайджан
[1]a_farqana@mail.ru, [2]allahverdiyevasabira@gmail.com

Подход к выявлению вульгаризма на основе машинного обучения

В статье разработан подход к обнаружению вульгаризмов в веб-контенте на основе машинного обучения. Увеличение количества вредоносного контента на веб-страницах поднимает вопрос защиты от него. Подверженность пользователей вульгаризму (неэтичная речь, сленг, ругань, оскорбления и т.д.) в Интернете, особенно детей и подростков, отрицательно сказывается на их психологии. Для решения этой проблемы очень важна разработка более надежных методов автоматического обнаружения в тексте вульгарных слов, фраз и выражений как в Интернете, так и в социальных сетях (Twitter, Facebook и т.д.). В представленной статье предлагается подход к выявлению вульгаризмов с использованием символов *n-gram* + TF-IDF. Здесь числовые векторы генерируются путем применения метода извлечения признаков на основе *n-gram* + TF-IDF к заранее известным вульгарным словам. Сгенерированный числовой вектор передается на вход алгоритмов Наивного Байеса. В результате экспериментов с использованием разных признаков классификация на основе признаков *unigram* + TF-IDF дала лучшие результаты. Такой подход к выявлению вульгаризмов важен в формировании культуры речи и для общения детей и подростков. Этот подход полезен для защиты детей от вредоносного контента, полученного из Интернета, и может использоваться в центрах безопасности детей и в системе образования.

Ключевые слова: *вульгаризмы, N-граммы, TF-IDF, Наивный Байесовский метод, безопасность детей в Интернете.*

Fargana J. Abdullayeva[1], Sabira S. Ojagverdiyeva[2]
[1,2]Institute of Information Technology of ANAS, Baku, Azerbaijan
[1]a_farqana@mail.ru, [2]allahverdiyevasabira@gmail.com

An approach to identify vulgarism based on machine learning

The article develops an approach to the identification of vulgarism in web content based on machine learning. The increasing number of harmful contents in web-pages makes protection from them even more vital. Encountering vulgarisms (indecent words, jargon, slams, etc) on the internet among users, especially children, and teenagers, shows a negative effect on their psychology. To identify vulgar words, word conjunctions, and expressions, in both social media (Twitter, Facebook, etc.) and online media, it is important to develop new auto-text identification methods, which is vital to solve that matter. The presented paper proposes an approach for the detection of vulgarisms using the N-grams+TF-IDF features. Numerical vectors are generated by applying the *n-gram*+TF-IDF-based feature extraction method to the predefined vulgar words. Generated numerical vector is passed to the input of the Naive Bayes algorithm. As a result of experiments conducted on different features, the classification based on *unigram*+TF-IDF features performs better results. The proposed approach, which contains the identification of vulgarism, is important for developing conversation culture and communication skills of children and teenagers. This approach is very important to protect kids from harmful content online and can be used in child safety centers and education systems.

Keywords: *vulgarisms, N-grams, TF-IDF, Naive Bayes, Child safety on the Internet.*