Twitter məlumatlarının sentiment analizi

Məkrufə Hacırəhimova[1], Adilə İmamverdiyeva[2]
[1, 2]AMEA İnformasiya Texnologiyaları İnstitutu, Bakı, Azərbaycan
[1]makrufa@science.az,[2]imamverdiyeva1998@gmail.com

***Xülasə*— Twitter mikrobloqların paylaşıldığı onlayn sosial şəbəkə saytıdır. Qeydiyyatdan keçmiş istifadəçilərinin sayı 500 milyon olan Twitter-də hər gün 400 milyondan çox mesaj yayılır. Bu tvitlər, demək olar ki, cəmiyyətdə cərəyan edən bütün hadisələr haqqında məlumatları və şəbəkə iştirakçılarının onlara münasibətlərini əks etdirir. Bir çox ölkədə Twitter ictimai rəyə böyük təsir göstərir. Bu məqalədə Twitter məlumatlarından əlamətlərin çıxarılmasına və müxtəlif maşın təlimi metodları ilə onların emosional “rənginin” (pozitiv və neqativ) müəyyən edilməsi məsələsinə baxılmışdır.**

Açar sözlər— Twitter; sosial şəbəkə; sentiment analiz; maşın təlimi.

I. GİRİŞ

Twitter istifadəçilərin tvitlər kimi tanınan qısa mesajlarla (140 simvola qədər) qarşılıqlı əlaqədə olduğu məşhur sosial şəbəkə saytıdır. Bu şəxslərin müxtəlif mövzular haqqında düşüncə və duyğularını ifadə etmək üçün bir vasitədir. İstehlakçılar və marketoloqlar kimi müxtəlif tərəflər məhsullar haqqında fikirləri toplamaq və ya bazarı analiz etmək üçün bu tvitlərin sentiment analizini yerinə yetirirlər. Tvitlərin sentiment analizi ictimai rəyin öyrənilməsi üçün də vacib alətdir. Bu işdə aşağıdakı anlayışlar istifadə olunur:

Sentiment – müəllifin obyektə (prosesə) emosional münasibətini bildirməsidir.

Sentiment analiz – müəllifin obyektə emosional münasibətinin kontent analiz metodları ilə avtomatik müəyyən edilməsi prosesidir.

Bu işin əsas məqsədi müxtəlif maşın təlimi alqoritmlərindən istifadə edərək tvitlər üzərində sentiment analiz aparmaqdır. Qeyd edək ki, Twitter məlumatlarının sentiment analizi metodlarının icmalı [1]-də, sentiment analiz üçün istifadə edilən verilənlər toplularının icmalı isə [2]-də verilmişdir.

Bu tədqiqatda Twitter verilənləri bazası Kaggle saytından götürülərək istifadə edilmişdir [3]. Verilənlərdə emosiya ikonaları, istifadəçi adları və heştеqlər olur, onları ilkin emal etmək və standart formaya çevirmək lazımdır. Həmçinin sentimenti müəyyən etmək üçün tvitdən uniqramlar və biqramlar kimi faydalı əlamətləri çıxarmaq lazımdır. Alınmış əlamətlərdən istifadə edərək sentiment analiz etmək üçün müxtəlif maşın təlimi alqoritmlərindən istifadə edilmişdir.

II. VERİLƏNLƏRİN İLKİN EMALI

Verilənlər vergüllə ayrılmış mətnlər şəklində tvitlər və onlara uyğun sentiment sinifləri şəklindədir. Təlim verilənləri `tweet_id`, `sentiment`, `tweet` tipində bir .csv faylıdır, `tweet_id` tviti müəyyən edən unikal tam ədəddir, `sentiment` 1 (pozitiv), ya da 0 (neqativ), `tweet` isə “ ” şəklində yazılan tvitdir. Eyni şəkildə test verilənlər də `tweet_id, tweet` tipli bir .csv fayldır.

Verilənlər sözlər, simvollar, emosiya ikonaları, URL və istifadəçilərə istinadların qarışığıdır. Sözlər və emosiya ikonaları sentimenti proqnozlaşdırmağa kömək edir, amma URL və istinadlar nəzərə alınmaya bilər. Sözlər həmçinin səhv yazılmış sözlərdən, çoxlu nöqtələrdən və bir çox təkrarlanan hərflərdən düzələn sözlərin bir qarışığıdır. Buna görə, standart formaya gətirmək üçün onları ilkin emal etmək lazımdır.

Təqdim edilən təlim və test verilənləri toplusunda uyğun olaraq, təxminən 800 min və 200 min tvit var.

Twitterdən götürülmüş emal edilməmiş tvitlər çox zaman küylü verilənlər toplusu olur. Bu, insanların sosial mediadan istifadəsinin təsadüfi əlamətlərindəndir. Tvitlərin istifadəçi istinadları, emosiya ifadələri və s. kimi xüsusi əlamətləri var. Emal edilməmiş twitter verilənləri müxtəlif klassifikatorlar vasitəsilə asanlıqla öyrənilə biləcək bir verilənlər toplusu yaratmaq üçün normallaşdırmalıdır. Verilənləri standart formaya gətirmək və ölçüsünü azaltmaq üçün çox sayda ilkin emal addımları tətbiq etmək lazımdır.

Əvvəlcə tvitləri aşağıdakı kimi ilkin emal edirik:

1. Tviti kiçik hərflərə çeviririk.
2. 2 və ya daha çox nöqtəni (.) boş sahə ilə əvəzləyirik.
3. Tvitlərin sonunda dırnaq və boş sahələri atırıq.
4. 2 və daha çox boş sahəni bir boş sahə ilə əvəzləyirik.

Tvitlərdəki xüsusi əlamətlər aşağıdakı kimi emal edilir.

DOI: 10.25045/NCInfoSec.2018.59

URL – istifadəçilər öz tvitlərində tez-tez müxtəlif veb səhifələrin hiper-linklərini bölüşürlər. Mətn klassifikasiyası üçün hər hansı bir xüsusi URL vacib deyildir, çünki bu çox seyrək əlamətlərə gətirib çıxara bilər. Buna görə də biz tvitlərdəki bütün URL-ləri URL sözü ilə əvəz edirik. URL-ləri tapmaq üçün istifadə olunan requlyar ifadə ((www\. [\S]+) | (https ?: // [\S]+)) 'dir.

İstifadəçi istinadları (ing. **User Mention) –** hər bir Twitter istifadəçisinin onunla əlaqələndirilən bir deskriptoru (ing. handle) var. İstifadəçilər tez-tez tvitlərində @handle ilə digər istifadəçilərə istinad edirlər. Bütün istifadəçi istinadlarını USER_MENTION sözü ilə əvəz edirik. İstifadəçi istinadlarını tapmaq üçün istifadə olunan requlyar ifadə @[\S]+ 'dir.

İkona (ing. emoticon) – istifadəçilər fərqli emosiyaları ifadə etmək üçün tvitlərində bir sıra fərqli emosiya ikonaları istifadə edirlər. Sosial mediada istifadə olunan emosiya ikonalarının sayı artır, buna görə onların hamısını nəzərə almaq mümkün olmur. Biz ancaq çox istifadə olunan bəzi ümumi emosiya ikonlarını emal edirik. Emal edilmiş emosiya ikonlarını onların ifadə etdiyi emosiyanın pozitiv və ya neqativ olmasından asılı olaraq, EMO_POS və ya EMO_NEG ilə əvəz edirik.

Heşteq – heşteqlər Twitter-də trend mövzusuna istinad etmək üçün istifadəçilər tərəfindən tez-tez istifadə olunan, heş simvolundan (#) sonra boşluq olmayan frazalardır. Biz bütün sözlü heşteqləri heş simvolu ilə əvəz edirik. Məsələn, #hello heşteqi “hello” ilə əvəz olunur. Heşteqləri tapmaq üçün istifadə olunan requlyar ifadə # (\S+) 'dir.

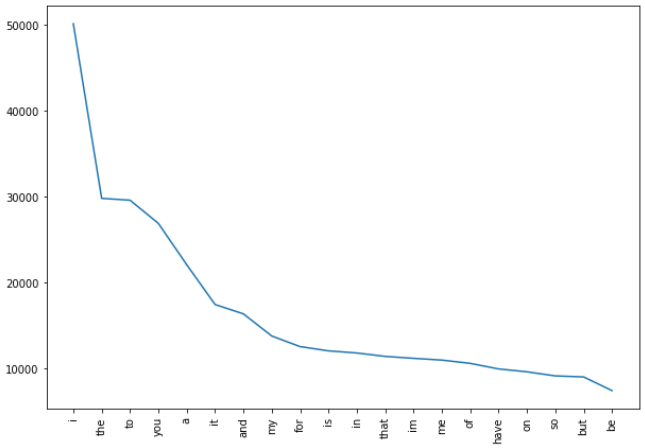
Retweet – Retvit artıq başqası tərəfindən göndərilmiş və digər istifadəçilər tərəfindən paylaşılmış tvitlərdir. Retvitlər RT hərfləri ilə başlayır. Tvitlərdən RT-lər atılır, çünki mətn klassifikasiyası üçün vacib bir əlamət deyillər. Retvitləri tapmaq üçün istifadə olunan requlyar ifadə \brt \b 'dir.

III. ƏLAMƏTLƏRİN ÇIXARILMASI

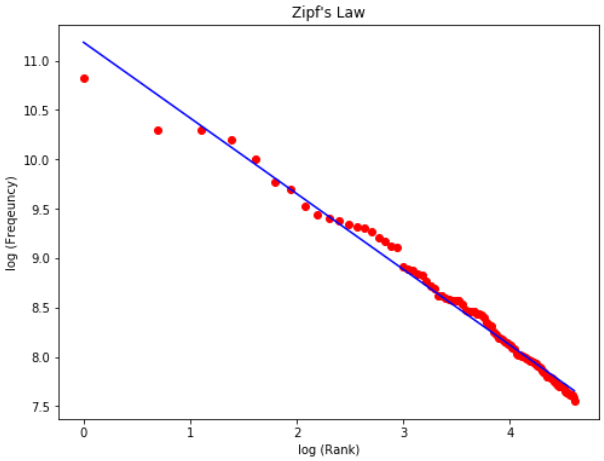
Verilənlər toplusundan iki əlamət çıxarılır: uniqramlar və biqramlar. Bu uniqramların və biqamların tezlik paylanması yaradılır və sentiment analiz üçün tezlikləri böyük olan ilk *N* uniqram və biqram götürülür.

Uniqramlar

Mətnlərin klassifikasiyası üçün ən sadə və ən çox istifadə edilən əlamətlər mətndə olan tək sözlər və ya tokenlərdir. Təlim verilənləri toplusundan tək sözlər çıxarılır və bu sözlərin tezlik paylanması yaradılır. Baxılan verilənlər toplusundan təxminən 181232 söz çıxarılmışdı. Tezliyi az olan sözlər cəmisi bir neçə dəfə rast gəlir və onları küy hesab etmək olar. Buna görə klassifikasiya üçün sözlər çoxluğunu yaratmaq üçün tezliyi böyük olan ilk *N* sayda sözdən istifadə etmişik. Şəkil 1-də əlimizdə olan sözlər çoxluğundan tezliyi ən böyük 20 sözün tezlik paylanması göstərilir. Şəkil 2-də isə tezlik paylanmasının Zipf qanununa tabe olduğunu müşahidə edə bilərik. Zipf qanununa görə, sözün tezliyi tezlik cədvəlindəki sırası ilə tərs mütənasibdir. Bunu $\log(tezlik)$ və $\log(sıra)$ qrafiklərini qurmaq ilə görə bilərik. Şəkil 2 üçün $\log(Frequency) = -0.78\log(Rank) + 13.31$ doğrudur.



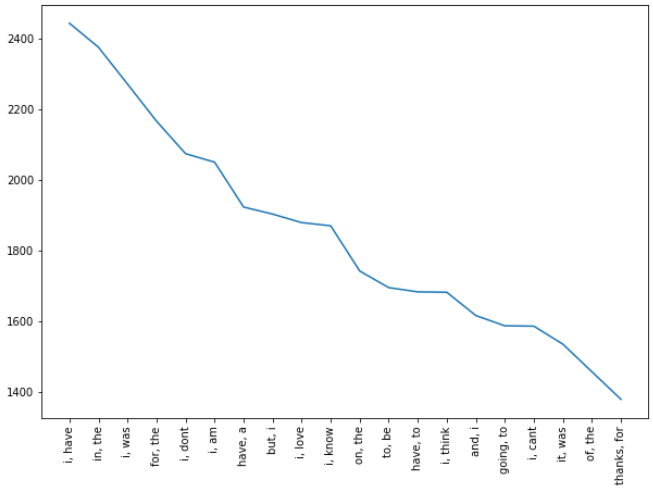
Şəkil 1. İlk 20 uniqramın tezliyi



Şəkil 2. Zipf qanunu

Biqramlar

Biqramlar korpusda ardıcıl gələn verilənlər sırasında söz cütləridir. Bu əlamətlər təbii dildə mənfi emosiyanı ifadə etmək üçün yaxşı üsuldur – “Bu yaxşı deyil”. Verilənlər bazasından cəmi 1954953 unikal biqram çıxarılıb. Sözlər çoxluğunu yaratmaq üçün tezliyi ən böyük olan 10000 biqram istifadə edirik. Sözlər çoxluğunda tezliyi ən böyük olan 20 biqramın tezlik paylanması şəkil 3-də göstərilib.



Şəkil 3. İlk 20 biqramın tezliyi

IV. KLASSİFİKATORLAR

Eksperimentlər üçün aşağıdakı klassifikatorlar seçilib.

Baseline – bu modeldə tvitin sentimentini tapmaq üçün neqativ və pozitiv sözlərin sayı istifadə edilib (**Opinion Dataset** toplusundan götürülür). Pozitiv və neqativ sözlərin sayı bərabər olduqda, sentiment müsbət götürülür.

Naive Bayes

Naive Bayes mətnlərin klassifikasiyası üçün istifadə edilə biləcək ən sadə bir modeldir. Bu modeldə t tviti üçün $\hat{c}$ sinfi aşağıdakı kimi təyin olunur:

$$\hat{c} = \underset{c}{\operatorname{argmax}}\, P(c|t)$$

$$P(c|t) \propto P(c) \prod_{i=1}^{n} P(f_i|c).$$

Yuxarıdakı düsturda f_i bütün n əlamət arasından i-ci əlaməti işarə edir. $P(c)$ və $P(f_i|c)$ isə maksimal həqiqətəoxşarlıq qiymətləri ilə əldə edilə bilər.

Decision Tree (Qərar ağacı)

Qərar ağacı klassifikasiya modelidir, burada ağacın hər bir düyünü verilənlər toplusunun atributu üzrə testi, ondan çıxan budaq (övlad) isə testin nəticəsini müəyyən edir. Yarpaq düyünləri isə verilənlərin son siniflərini təyin edir. Qərar ağacını qurmaq üçün sinif nişanları məlum olan verilənlər istifadə edilir, sonra qurulmuş model test verilənlərini klassifikasiya etmək üçün tətbiq olunur. Ağacın hər bir düyünüdə budaqlanmanı tapmaq üçün ən yaxşı qərar qəbul edilməlidir. Ən yaxşı budaqlanmaya qərar vermək üçün GINI əmsalından istifadə edirik. Verilmiş t düyünü üçün $GINI(t) = 1 - \sum_j [p(j|t)]^2$ kimi hesablanır, burada $p(j|t)$ – t düyünündəki j sinfinin nisbi tezliyidir. $GINI_{split} = \sum_{i=1}^{k} \frac{n_i}{n} GINI(i)$ budaqlanmanın keyfiyyətini göstərir.(n_i – i-ci övladda düyünlərin sayı, n isə p düyünündə yazıların sayıdır). GINI əmsalını minimal edən bir bölünmə seçilir.

Random Forest (Təsadüfi meşə)

Təsadüfi meşə klassifikasiya və reqressiya üçün ansambl təlim alqoritmidir. Təsadüfi meşə çoxlu sayda qərar ağacı klassifikatorlarından ibarətdir, yekun qərar ağacların qərarlarının aqreqasiyası əsasında yaradılır. Tutaq ki, $x_1, x_2, \dots, x_n$ tvitlər çoxluğu, $y_1, y_2, \dots, y_n$ isə onların müvafiq sentiment nişanlarıdır. Bagging tətbiq edilməklə əvəzləmə ilə təsadüfi bir(X_b, Y_b) cütü seçilir. Hər bir klassifikasiya ağacı f_b təsadüfi bir (X_b, Y_b) istifadə edərək öyrədilir, burada b – 1...B aralığındadır. Sonda bu B sayda qərar ağacının yekun qərarı majoritar səsvermə ilə müəyyən edilir.

XGBoost

XGBoost qradiyent bustinqi alqoritminin bir formasıdır, zəif proqnozlaşdırılan qərar ağaclarının ansamblıdır. K modelin ansamblı onların nəticələri aşağıdakı şəkildə birləşdirilərək istifadə edilir:

$$\hat{y_i} = \sum_{k=1}^{K} f_k(x_i), f_k \in F$$

burada F ağacların fəzasıdır. x_i giriş veriləni, $\hat{y_i}$ isə yekun çıxışdır. Aşağıdakı zərər funksiyasını minimallaşdırmağa çalışırıq:

$$L(\Phi) = \sum_i l(\hat{y_i}, y_i) \sum \Omega(f_k)\,,$$

$$\text{burada } \Omega(f) = \gamma T + \frac{1}{2}\lambda||\omega||^2,$$

Ω – tənzimləyici toplananıdır.

SVM (Support Vector Machine)

Dayaq vektorları maşınları kimi də tanınır, binar xətti klassifikatordur. Tutaq ki, təlim toplusu (x_i, y_i) cütlərindən ibarətdir, burada x – əlamətlər vektoru, y isə ona uyğun sinif nişanıdır, $y \in \{+1, -1\}$. Biz elə hiper-müstəvi tapmaq istəyirik ki, o y_i =1 və y_i = −1 nöqtələrini ayırsın və təlim çoxluğunun ən yaxın nöqtələrindən maksimal məsafədə keçsin.

Hiper-müstəvinin tənliyi $\omega \cdot x - b = 0$ ilə verilir. Optimal hipermüstəvinin qurulması məsələsi aşağıdakı kimi ifadə olunur:

$$\min_{w} \frac{\|w\|}{2}$$

Şərtlər: $y_i(\mathrm{w}^T \mathrm{x} + b) \geq 1$

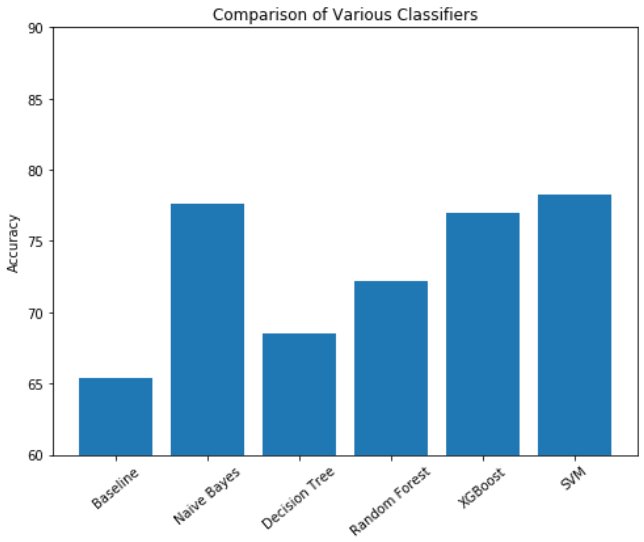
V. EKSPERİMENTLƏRİN NƏTİCƏLƏRİ

Yuxarıda təsvir olunmuş müxtəlif klassifikatorlar Python mühitində scikit-learn kitabxanasından istifadə edilməklə reallaşdırılıb və [3]-dən götürülmüş məlumatlar üzərində eksperimentlər aparılıb. Klassifikatorların eksperimenlərdə alınmış dəqiqliyi (ing. accuracy) cədvəl I-də verilib.

CƏDVƏL I. KLASSİFİKATORLARIN SENTİMENTİ TANIMA DƏQİQLİYİ

Klassifikator	**Dəqiqlik**
Baseline	65.35 %
Naive Bayes	77.66 %
Decision Tree	68.48 %
Random Forest	72.20 %
XGBoost	76.95 %
SVM	78.24 %

Klassifikatorların dəqiqliyə görə müqayisəsi histoqramlarla şəkil 4-də göstərilib.



Şəkil 4. Klassifikatorların müqayisəsi

Eksperimentlərin nəticələri göstərir ki, yalnız fərdi modellərə (klassifikatorlara) əsaslanmaq yüksək dəqiqlik vermir. Ona görə ən yaxşı modelləri seçərək modellər ansamblı yaratmaq lazımdır.

NƏTİCƏ

Məqalədə Twitter məlumatlarından əlamətlərin çıxarılmasına və müxtəlif maşın təlimi metodları ilə onların iki sentiment sinfinə (pozitiv və neqativ) klassifikasiyası məsələsinə baxılmışdır. Gələcək tədqiqatlarda mətnləri təsvir etmək, digər modellərdən istifadə etmək, neyron şəbəkələrə (MLP, CNN, RNN) əsaslanan bir neçə klassifikatoru sentiment analiz üçün tətbiq etmək və baxılan məsələ üçün klassifikatorlar ansamblı qurmaq nəzərdə tutulur.

ƏDƏBİYYAT

[1] A. Giachanou, & F. Crestani, “Like it or not: A survey of twitter sentiment analysis methods,” ACM Computing Surveys (CSUR), vol. 49(2), Article No. 28, pp. 28, 2016.

[2] H. Saif, M. Fernandez, Y. He, & H. Alani, “Evaluation datasets for Twitter sentiment analysis: a survey and a new dataset, the STS-Gold,” 1st Interantional Workshop on Emotion and Sentiment in Social and Expressive Media: Approaches and Perspectives from AI (ESSEM), 2013.

[3] https://www.kaggle.com/c/twitter-sentiment-analysis2/data

[4] S. Shalev-Shwartz, & S. Ben-David, Understanding machine learning: From theory to algorithms. Cambridge university press. 2014.

[5] A.C.Mueller, S. Guido, Introduction to Machine Learning with Python: A Guide for Data Scientists. O'Reilly Media, 2016.

ABOUT TWITTER SENTIMENT ANALYSIS

Makrufa Hajirahimova[1], Adila İmamverdiyeva[2]

[1,2]Institute of Information Technology of ANAS,

Baku, Azerbaijan

[1]makrufa@science.az,[2]imamverdiyeva1998@gmail.com

Abstract – Twitter is a social networking site where microblogs are shared. More than 500 million registered users send more than 400 million messages every day. These tweets provide information on almost all events in the life of society and on the attitudes of network members towards them. In many countries, Twitter has a huge impact on public opinion. This article discusses extracting features from tweets and recognizing their emotional “color” (positive and negative) using various machine learning methods.

Keywords – Twitter, social network, sentiment analysis, machine tlearning.