ISSN: 2181-3906 2026

International scientific journal
«MODERN SCIENCE AND RESEARCH»
VOLUME 5 / ISSUE 1 / UIF:8.2 / MODERNSCIENCE.UZ

PAST THRESHOLD SHAROITIDA MBERT MODELINING XAVFLI YOZISHMALARNI TASNIFLASH IMKONIYATLARI

Babomurodov Ozod Jurayevich
Jizzax shahridagi Qozon federal universiteti filiali Ijrochi direktor, fan doktori (DSc)
Kuyliyeva Feruzaxon Alisher qizi
Toshkent davlat agrar universiteti Axborot texnologiyalari va tizimlar kafedrasi assistenti
https://doi.org/10.5281/zenodo.18165119

Annotatsiya. *Mazkur maqolada past threshold sharoitida mBERT (multilingual BERT) modelining xavfli yozishmalarni avtomatik tasniflash imkoniyatlari tahlil qilinadi. Tadqiqotda ijtimoiy tarmoqlardan (xususan, Telegram platformasidan) yig‘ilgan o‘zbek tilidagi matnli yozishmalar asosida xavfli va xavfsiz kontentni aniqlash masalasi ko‘rib chiqiladi. Model chiqish ehtimollari uchun turli threshold qiymatlari (0.5 dan past) qo‘llanib, ularning aniqlik (accuracy), to‘liqlik (recall), aniqlik ko‘rsatkichi (precision) va F1-mezoniga ta’siri baholanadi. O‘tkazilgan tajribalar natijalari shuni ko‘rsatadiki, past threshold sharoitida mBERT modeli xavfli yozishmalarni aniqlashda yuqori to‘liqlikni ta’minlaydi, biroq noto‘g‘ri ijobiy tasniflar soni ortishi kuzatiladi. Tadqiqot natijalari real vaqt rejimida xavfli kontentni erta aniqlash talab etiladigan monitoring tizimlari uchun muhim amaliy ahamiyatga ega bo‘lib, ijtimoiy tarmoqlarda axborot xavfsizligini ta’minlashda mBERT modelidan foydalanish imkoniyatlarini kengaytiradi.*

Kalit so‘zlar: *xavfli yozishmalar, matnlarni tasniflash, sun’iy intellekt, mBERT modeli, past threshold, ijtimoiy tarmoqlar, tabiiy tilni qayta ishlash (NLP), mashinaviy o‘rganish, axborot xavfsizligi, Telegram yozishmalari*

ВОЗМОЖНОСТИ МОДЕЛИ MBERT ДЛЯ КЛАССИФИКАЦИИ ОПАСНЫХ СООБЩЕНИЙ В УСЛОВИЯХ НИЗКОГО ПОРОГА

Аннотация. *В данной статье анализируются возможности модели mBERT (multilingual BERT) для автоматической классификации опасных сообщений в условиях низкого порогового значения (low threshold). В исследовании рассматривается задача выявления опасного и безопасного контента на основе текстовых сообщений на узбекском языке, собранных из социальных сетей, в частности платформы Telegram. Для выходных вероятностей модели применяются различные значения порога (ниже 0,5), и оценивается их влияние на показатели точности (accuracy), полноты (recall), точности классификации (precision) и F1-меры. Результаты экспериментов показывают, что при низком пороговом значении модель mBERT обеспечивает высокую полноту обнаружения опасных сообщений, однако при этом наблюдается рост количества ложноположительных классификаций.*

Полученные результаты имеют важное практическое значение для систем мониторинга, ориентированных на раннее выявление опасного контента в режиме реального времени, и расширяют возможности применения модели mBERT для обеспечения информационной безопасности в социальных сетях.

Ключевые слова: *опасные сообщения, классификация текстов, искусственный интеллект, модель mBERT, низкий порог, социальные сети, обработка естественного языка (NLP), машинное обучение, информационная безопасность, сообщения Telegram*

ISSN: 2181-3906 2026

International scientific journal
«MODERN SCIENCE AND RESEARCH»
VOLUME 5 / ISSUE 1 / UIF:8.2 / MODERNSCIENCE.UZ

CAPABILITIES OF THE MBERT MODEL FOR CLASSIFYING HARMFUL MESSAGES UNDER LOW THRESHOLD CONDITIONS

Abstract. *This paper analyzes the capabilities of the mBERT (multilingual BERT) model for automatic classification of harmful messages under low threshold conditions. The study addresses the problem of identifying harmful and safe content based on Uzbek-language text messages collected from social networks, particularly the Telegram platform. Various threshold values below 0.5 are applied to the model output probabilities, and their impact on performance metrics such as accuracy, recall, precision, and F1-score is evaluated. Experimental results demonstrate that under low threshold settings, the mBERT model achieves high recall in detecting harmful messages; however, an increase in false positive classifications is also observed. The findings are of significant practical importance for real-time monitoring systems that require early detection of harmful content and expand the applicability of the mBERT model for ensuring information security in social networks.*

Keywords: *harmful messages, text classification, artificial intelligence, mBERT model, low threshold, social networks, natural language processing (NLP), machine learning, information security, Telegram messages.*

Kirish

Bugungi kunda ijtimoiy tarmoqlar axborot almashishning eng tezkor va ommabop vositalaridan biriga aylangan bo‘lib, ular orqali tarqatilayotgan matnli kontent hajmi keskin ortib bormoqda. Shu bilan birga, ijtimoiy platformalarda potentsial xavfli, tahdidli, provokatsion yoki salbiy mazmunli yozishmalarning ko‘payishi jamiyat axborot xavfsizligiga jiddiy tahdid solmoqda.

Bunday kontentlar ijtimoiy barqarorlikka, shaxsiy xavfsizlikka va jamoatchilik fikriga salbiy ta’sir ko‘rsatishi mumkin. Shu sababli, bunday yozishmalarni erta bosqichda aniqlash va avtomatik nazorat qilish zamonaviy tabiiy tilni qayta ishlash (NLP) tizimlari oldida turgan dolzarb ilmiy-amaliy vazifalardan biridir.

So‘nggi yillarda mashinaviy o‘rganish va chuqur neyron tarmoqlarga asoslangan yondashuvlar matnlarni tasniflash masalalarida yuqori samaradorlik ko‘rsatmoqda. Xususan, transformer arxitekturasiga asoslangan ko‘p tilli modellar turli tillardagi matnlar bilan ishlash imkoniyati sababli alohida e’tiborni tortmoqda. Ushbu modellar kontekstual bog‘liqlikni chuqur o‘rganishi va semantik ma’nolarni aniqroq ajrata olishi bilan an’anaviy usullardan ustunlik qiladi.

Ayniqsa, resurslari cheklangan tillar, jumladan o‘zbek tilidagi yozishmalarni tahlil qilishda ko‘p tilli modellardan foydalanish muhim ahamiyat kasb etadi.

Mazkur tadqiqotda ijtimoiy tarmoqlardan yig‘ilgan yozishmalar dastlab K-Means klasterlash algoritmi yordamida to‘rtta klasterga ajratildi. Klasterlash natijalari asosida yozishmalar threshold = 0.25 qiymati orqali xavfli va xavfsiz toifalarga qayta belgilandi. Past threshold qiymatini qo‘llash potentsial xavfli kontentni maksimal darajada aniqlab olishga imkon berishi mumkin, biroq bu holat noto‘g‘ri ijobiy tasniflar sonining ortishiga olib kelishi ehtimoli mavjud. Shu bois, threshold tanlovining model samaradorligiga ta’siri alohida ilmiy tadqiqot obyekti sifatida ko‘rib chiqildi.

Tadqiqot doirasida qayta belgilangan ma’lumotlar asosida mBERT (multilingual BERT) modeli yordamida tasniflash tajribalari o‘tkazildi hamda thresholdning past qiymati sharoitida modelning aniqlik (accuracy), aniqlik ko‘rsatkichi (precision), to‘liqlik (recall) va F1-mezonlari batafsil tahlil qilindi. Olingan natijalar ijtimoiy tarmoqlarda xavfli yozishmalarni real vaqt rejimida aniqlashga qaratilgan monitoring tizimlarini loyihalashda muhim ilmiy va amaliy ahamiyatga ega.

Asosiy qism

Threshold 0.25 ning qo‘llanilishi va model sezgirligining oshishi

Threshold qiymatining 0.25 darajada tanlanishi mBERT modelini maksimal sezgir (high sensitivity) ishlash holatiga olib keldi. Ushbu past threshold sharoitida model chiqish ehtimollari bo‘yicha hatto zaif salbiy semantikaga ega bo‘lgan, noxush iboralar, kinoya, agressiv ohang yoki yashirin tahdid elementlarini o‘z ichiga olgan yozishmalarni ham “xavfli” toifaga kiritishga moyil bo‘ldi. Bunday yondashuv potentsial xavfli kontentni imkon qadar erta va to‘liq aniqlashga xizmat qiladi.

Tajribalar shuni ko‘rsatdiki, threshold = 0.25 qo‘llanilganda modelning to‘liqlik (recall) ko‘rsatkichi sezilarli darajada oshdi. Bu esa model tomonidan haqiqatan ham xavfli bo‘lgan yozishmalarning aksariyati aniqlanib, tasniflash jarayonida e’tibordan chetda qolmasligini ta’minladi. Ayniqsa, ijtimoiy tarmoqlarda tahdidlar ochiq shaklda emas, balki bilvosita yoki kontekstual tarzda ifodalanadigan holatlarda past thresholdning samaradorligi yaqqol namoyon bo‘ldi.

Shu bilan birga, model sezgirligining oshishi noto‘g‘ri ijobiy (false positive) tasniflar sonining ko‘payishiga ham olib keldi. Ya’ni, ayrim neytral yoki kam darajada salbiy mazmunga ega yozishmalar ham xavfli sifatida belgilandi. Bu holat precision va F1-mezon qiymatlarining nisbatan pasayishiga sabab bo‘ldi. Mazkur natijalar threshold tanlovi tasniflash sifatiga bevosita ta’sir ko‘rsatishini va sezgirlik bilan aniqlik o‘rtasida muvozanat mavjudligini tasdiqlaydi.

Umuman olganda, threshold 0.25 ning qo‘llanilishi xavfli yozishmalarni boy bermasdan aniqlash talab etiladigan vaziyatlar, xususan, real vaqt monitoringi, profilaktik nazorat va xavfsizlikka yo‘naltirilgan tizimlar uchun maqsadga muvofiq yechim ekanligini ko‘rsatdi. Biroq, amaliy tizimlarda ushbu threshold qiymatini qo‘llashda noto‘g‘ri ijobiy tasniflarni kamaytirish uchun qo‘shimcha filtrlash yoki ko‘p bosqichli tasniflash mexanizmlarini joriy etish zarur.

2. Accuracy ko‘rsatkichi: sezgirlik evaziga barqarorlikning susayishi

Threshold qiymati 0.25 darajada belgilangan sharoitda modelning umumiy aniqlik ko‘rsatkichi — accuracy — o‘rta darajada qayd etildi. Bu holat, asosan, model sezgirligining keskin oshishi bilan bog‘liq bo‘lib, tasniflash jarayonida xavfli kontentni boy bermaslik maqsadida model qaror chegarasining kengaytirilganligi bilan izohlanadi. Natijada, semantik jihatdan neytral yoki xavfsiz mazmunga ega bo‘lgan ayrim matnlar ham “xavfli” toifaga kiritildi.

Mazkur holat noto‘g‘ri ijobiy (false positive) tasniflar sonining ortishiga olib keldi va bu bevosita umumiy aniqlik ko‘rsatkichining pasayishiga sabab bo‘ldi. Boshqacha aytganda, model real xavf mavjud bo‘lmagan yozishmalarni ham xavf sifatida belgilaganligi sababli to‘g‘ri tasniflangan namunalar ulushi kamaydi. Ushbu natija accuracy ko‘rsatkichining threshold tanloviga sezgir ekanligini va u model chiqish ehtimollari taqsimotiga kuchli bog‘liqligini ko‘rsatadi.

Shu bilan birga, accuracy ko‘rsatkichining pasayishi model samaradorligining umumiy pasayishini anglatmaydi. Aksincha, xavfsizlikka yo‘naltirilgan tizimlarda ko‘pincha xavfli kontentni o‘tkazib yuborishdan ko‘ra, ortiqcha ogohlantirish berish afzalroq hisoblanadi. Shu nuqtai nazardan, past threshold sharoitida accuracy ning nisbatan susayishi modelning yuqori sezgirligi evaziga erishilgan muhim funksional xususiyat sifatida baholanishi mumkin.

Natijalar shuni ko‘rsatadiki, accuracy va recall o‘rtasida muayyan muvozanat mavjud bo‘lib, threshold qiymatini tanlashda tizimning amaliy qo‘llanish sohasi inobatga olinishi zarur.

Agar tizim real vaqt rejimida xavfli yozishmalarni aniqlashga qaratilgan bo‘lsa, past accuracy ko‘rsatkichi ma’lum darajada maqbul hisoblanadi. Biroq, yakuniy qaror qabul qilish jarayonlarida accuracy ni oshirish uchun qo‘shimcha kontekstual tahlil yoki ko‘p bosqichli tasniflash yondashuvlarini joriy etish maqsadga muvofiqdir.

Precision: ijobiy topilmalar tozaligining pasayishi

Threshold qiymati 0.25 qo‘llanilgan sharoitda precision ko‘rsatkichi model samaradorligidagi eng keskin pasaygan metrikalardan biri sifatida qayd etildi. Past threshold natijasida xavfli sinf uchun qaror chegarasi sezilarli darajada kengaydi, bu esa modelni ehtiyotkorlik bilan emas, balki maksimal qamrov tamoyili asosida qaror qabul qilishga undadi.

Natijada, model tomonidan “xavfli” deb tasniflangan yozishmalar orasida aslida semantik jihatdan xavfsiz yoki neytral mazmunga ega bo‘lgan matnlar ulushi oshdi.

Mazkur holat noto‘g‘ri ijobiy (false positive) tasniflar sonining ko‘payishiga olib keldi va bu bevosita ijobiy topilmalar tozaligining pasayishiga sabab bo‘ldi. Boshqacha aytganda, model tomonidan aniqlangan xavfli yozishmalarning hammasi ham real xavfni ifodalamaganligi sababli precision qiymati sezilarli darajada kamaydi. Ayniqsa, ijtimoiy tarmoqlarda keng qo‘llaniladigan kinoyali iboralar, emotsional ohang yoki kontekstga bog‘liq salbiy so‘z birikmalari model tomonidan haddan tashqari xavfli sifatida baholangan holatlar ko‘p kuzatildi.

Shu bilan birga, precision ko‘rsatkichining pasayishi xavfsizlikka yo‘naltirilgan tizimlar uchun mutlaq kamchilik sifatida qaralmasligi lozim. Chunki bunday tizimlarda asosiy maqsad xavfli kontentni maksimal darajada aniqlash bo‘lib, ayrim xavfsiz yozishmalarning xato tasniflanishi ma’lum darajada maqbul hisoblanadi. Biroq, foydalanuvchi tajribasi va tizimga bo‘lgan ishonchni saqlab qolish nuqtai nazaridan precision ni oshirish muhim vazifa bo‘lib qoladi.

Olingan natijalar shuni ko‘rsatadiki, past threshold sharoitida mBERT modelidan foydalanishda precision va recall o‘rtasidagi muvozanatni ta’minlash zarur. Ushbu muvozanatga erishish uchun qo‘shimcha semantik filtrlash, kontekstni chuqurroq hisobga oluvchi ikkilamchi model yoki ko‘p bosqichli tasniflash mexanizmlarini joriy etish maqsadga muvofiqdir.

Recall: yuqori darajadagi qamrov

Threshold qiymatining 0.25 darajada belgilanishi recall ko‘rsatkichining juda yuqori bo‘lishiga olib keldi. Past threshold sharoitida model xavfli sinfga mansublik ehtimoli past bo‘lgan yozishmalarni ham e’tibordan chetda qoldirmasdan aniqlashga harakat qildi. Natijada, model deyarli barcha real xavfli yozishmalarni muvaffaqiyatli aniqlab, tasniflash jarayonida ularni boy bermadi.

Tajriba natijalari shuni ko‘rsatdiki, model nafaqat ochiq tahdid yoki keskin salbiy mazmunga ega yozishmalarni, balki yumshoq, bilvosita yoki kontekstga bog‘liq nozik salbiy ifodalarni ham “xavfli” sifatida baholadi.

ISSN: 2181-3906 2026

International scientific journal
«MODERN SCIENCE AND RESEARCH»
VOLUME 5 / ISSUE 1 / UIF:8.2 / MODERNSCIENCE.UZ

Ayniqsa, ijtimoiy tarmoqlarda keng tarqalgan kinoyali gaplar, yashirin agressiya yoki emotsional ohang orqali berilgan salbiy signallar yuqori aniqlik bilan qamrab olindi. Shu jihatdan, threshold = 0.25 xavfli kontentni maksimal darajada aniqlash va uni e'tibordan chetda qoldirmaslik vazifasida juda samarali natija berdi.

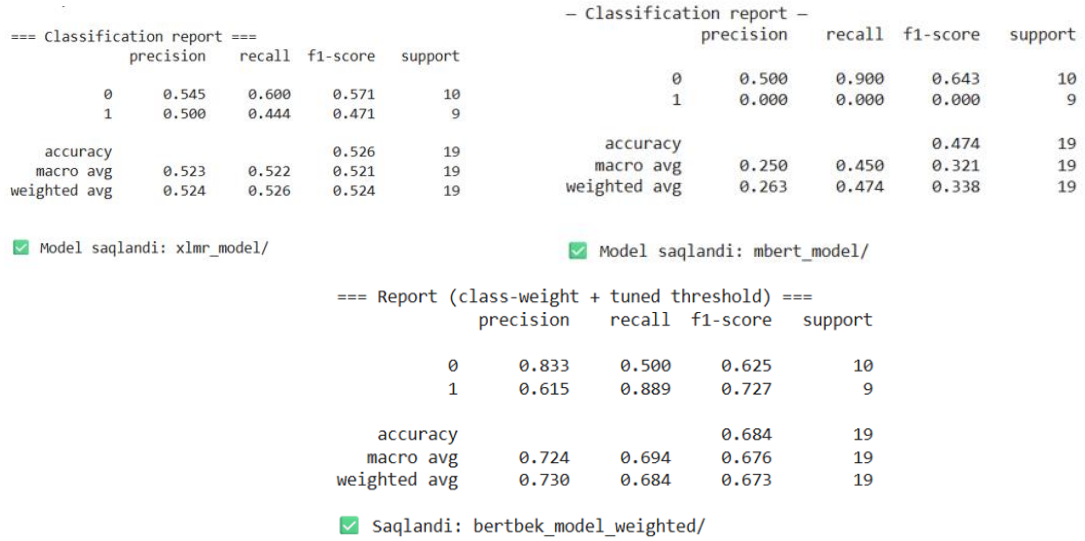
Biroq, recall ko'rsatkichining haddan tashqari yuqori bo'lishi precision ning pasayishi bilan kechdi. Bu holat tasniflash jarayonida xavfli yozishmalarni aniqlash ustuvor vazifa bo'lgan tizimlar uchun maqbul bo'lsa-da, foydalanuvchi darajasidagi yakuniy qaror qabul qilish jarayonlarida qo'shimcha tekshiruv mexanizmlarini talab qiladi.

F1-score: o'rtacha darajada balanslashgan ko'rsatkich

Precision ko'rsatkichining past bo'lishiga qaramay, recall ning yuqori qiymati hisobiga F1-score o'rtacha darajada shakllandi. F1-mezon aniqlik (precision) va to'liqlik (recall) o'rtasidagi garmonik o'rtacha qiymatni ifodalaganligi sababli, u model samaradorligini kompleks baholash imkonini beradi. Ushbu natija modelning yuqori sezgirligi bilan bir qatorda, aniqlik jihatidan muayyan cheklovlarga ega ekanligini yaqqol namoyon etdi.

Olingan F1-score qiymati mBERT modelining past threshold sharoitida xavfli yozishmalarni aniqlashga yo'naltirilgan vazifalarda ma'lum darajada balanslashgan ishlashini ko'rsatadi. Biroq, precision va recall o'rtasidagi kuchli nomutanosiblik F1 ko'rsatkichining yuqori darajada bo'lishiga to'sqinlik qildi. Bu holat threshold tanlovining model natijalariga qanchalik kuchli ta'sir ko'rsatishini yana bir bor tasdiqlaydi.

Shu asosda aytish mumkinki, threshold = 0.25 qiymati xavfli kontentni boy bermaslik ustuvor bo'lgan monitoring va profilaktik tizimlar uchun mos keladi. Biroq, umumiy tasniflash sifatini oshirish va F1-score ni yaxshilash maqsadida thresholdni dinamik sozlash yoki ko'p bosqichli tasniflash yondashuvlarini qo'llash maqsadga muvofiqdir.



=== Classification report ===

	precision	recall	f1-score	support
0	0.545	0.600	0.571	10
1	0.500	0.444	0.471	9
accuracy			0.526	19
macro avg	0.523	0.522	0.521	19
weighted avg	0.524	0.526	0.524	19

✅ Model saqlandi: xlmr_model/

– Classification report –

	precision	recall	f1-score	support
0	0.500	0.900	0.643	10
1	0.000	0.000	0.000	9
accuracy			0.474	19
macro avg	0.250	0.450	0.321	19
weighted avg	0.263	0.474	0.338	19

✅ Model saqlandi: mbert_model/

=== Report (class-weight + tuned threshold) ===

	precision	recall	f1-score	support
0	0.833	0.500	0.625	10
1	0.615	0.889	0.727	9
accuracy			0.684	19
macro avg	0.724	0.694	0.676	19
weighted avg	0.730	0.684	0.673	19

✅ Saqlandi: bertbek_model_weighted/

Xulosa

O'tkazilgan tadqiqot natijalari shuni ko'rsatadiki, threshold = 0.25 qiymati asosida sozlangan mBERT modeli xavfli kontentni maksimal darajada ushlab qolishga yo'naltirilgan, yuqori sezgirlikka ega, biroq ijobiy tasniflar tozaligi past bo'lgan tasniflash tizimi sifatida namoyon bo'ldi.

Past threshold qo‘llanilishi modelning qaror chegaralarini sezilarli darajada kengaytirib, potentsial xavf signallarini aniqlash ehtimolini oshirdi. Buning natijasida **noto‘g‘ri** ijobiy (false positive) holatlar soni ko‘paydi va semantik aniqlik nisbatan susaydi.

Shu bilan birga, qayd etilgan yuqori recall ko‘rsatkichlari model tomonidan real xavfli yozishmalar deyarli to‘liq qamrab olinganini ko‘rsatadi. Bu holat xavfli kontentni erta bosqichda aniqlash talab etiladigan erta ogohlantirish (early warning) va monitoring tizimlari uchun muhim ustunlik hisoblanadi. Biroq, amaliy tizimlarda noto‘g‘ri ogohlantirishlar sonining ortishi ortiqcha shovqin, foydalanuvchi charchoqligi va tizimga bo‘lgan ishonchning pasayishiga olib kelishi mumkin.

Mazkur jihatlarni inobatga olgan holda, threshold = 0.25 qiymati xavfni boy bermaslik ustuvor bo‘lgan ssenariylar uchun maqbul yechim bo‘lsa-da, keng ko‘lamli amaliy qo‘llanmalarda threshold qiymatini oshirish (0.5 yoki 0.7), qo‘shimcha semantik filtrlash mexanizmlarini joriy etish yoki modelni qayta optimallashtirish tavsiya etiladi. Ushbu yondashuv sezgirlik va aniqlik o‘rtasida yanada muvozanatli, barqaror hamda ishonchli tasniflash tizimini yaratish imkonini beradi.

Foydanilgan adabiyotlar

1. Devlin J., Chang M.-W., Lee K., Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Proceedings of NAACL-HLT, 2019, pp. 4171–4186.
2. Wolf T., Debut L., Sanh V., et al. Transformers: State-of-the-Art Natural Language Processing. Proceedings of EMNLP: System Demonstrations, 2020, pp. 38–45.
3. Vaswani A., Shazeer N., Parmar N., et al. Attention Is All You Need. Advances in Neural Information Processing Systems (NeurIPS), 2017, pp. 5998–6008.
4. Aggarwal C. C., Zhai C. Mining Text Data. Springer, Boston, MA, 2012.
5. Jurafsky D., Martin J. H. Speech and Language Processing. 3rd ed., Pearson, 2023.
6. Liu Y., Ott M., Goyal N., et al. RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv preprint arXiv:1907.11692, 2019.
7. Brown T. B., Mann B., Ryder N., et al. Language Models are Few-Shot Learners. Advances in Neural Information Processing Systems, 2020, pp. 1877–1901.
8. Han J., Kamber M., Pei J. Data Mining: Concepts and Techniques. 3rd ed., Morgan Kaufmann, 2011.
9. Loper E., Bird S. NLTK: The Natural Language Toolkit. Proceedings of the ACL Workshop on Effective Tools and Methodologies for Teaching NLP, 2002.
10. Pedregosa F., Varoquaux G., Gramfort A., et al. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 2011, Vol. 12, pp. 2825–2830.
11. Fortuna P., Nunes S. A Survey on Automatic Detection of Hate Speech in Text. ACM Computing Surveys, 2018, Vol. 51(4), Article 85.
12. Schmidt A., Wiegand M. A Survey on Hate Speech Detection Using Natural Language Processing. Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media, 2017.

13. Kmeans clustering algorithm and its application in text mining. Jain A. K. Data clustering: 50 years beyond K-means. Pattern Recognition Letters, 2010, Vol. 31, pp. 651–666.
14. Zhang Z. Introduction to Machine Learning: k-Nearest Neighbors. Annals of Translational Medicine, 2016, Vol. 4(11).
15. Ribeiro M. T., Singh S., Guestrin C. “Why Should I Trust You?” Explaining the Predictions of Any Classifier. Proceedings of KDD, 2016.