SPAM XABARLARNI ANIQLASHNING MASHINALI O'QITISH USULI

Isaqov Behzod Xabibullo o'g'li
Ilmiy izlanuvchi, AKT va AXI
Raxmonov Umid

Annotatsiya: *Spam-xabarlar ko'p sonli qabul qiluvchilarga yuboriladigan, ko'pincha mahsulot yoki xizmatni reklama qilish yoki targ'ib qilish uchun yuboriladigan kiruvchi va kerak bo'lmagan xabarlardir. Ular turli kanallar, jumladan, elektron pochta, matnli xabar, ijtimoiy media va lahzali xabarlar orqali yuborilishi mumkin. Spam xabarlar xavfsizlikka xavf tug'dirishi mumkin. Spam xabarlar fishing web-sahifalari yoki zararli dasturlarga havolalarni o'z ichiga olishi mumkin va ular foydalanuvchining shaxsiy yoki moliyaviy ma'lumotlarni taqdim etishga urinishimumkin.*

Spam xabarlarni aniqlashning turli xil usullari mavjud bo'lib, mazkur maqolada spam xabarlarni aniqlashning modifikatsiyalangan tayanch vektor usuli ko'rib chiqilgan.

Kalit so'zlar: *Spam,Spammerlar,Tayanch vektor*

Аннотация: *спам-сообщения-это нежелательные и нежелательные сообщения, которые отправляются большому количеству получателей, часто для продвижения или продвижения продукта или услуги. Их можно отправлять по различным каналам, включая электронную почту, текстовые сообщения, социальные сети и мгновенные сообщения. Спам-сообщения могут представлять угрозу безопасности. Спам сообщения могут содержать ссылки на фишинговые веб-страницы или вредоносные программы и могут пытаться предоставить личную или финансовую информацию пользователя. Существуют различные методы обнаружения спам-сообщений, и в этой статье рассматривается модифицированный базовый векторный метод обнаружения спам-сообщений.*

Ключевые слова: *спам,спамеры, базовый вектор*

Abstract: *Spam messages are unsolicited and unsolicited messages that are sent to a large number of recipients, often to promote or promote a product or service. They can be sent through various channels, including email, text messages, social networks and instant messages. Spam messages can pose a security risk. Spam messages may contain links to phishing web pages or malware and may attempt to provide personal or financial information to the user.*

There are various methods for detecting spam messages, and this article discusses a modified basic vector method for detecting spam messages.

Keywords: *spam,spammers, basic vector,*

Kirish. Hozirgi kunda ko'plab odamlar matnli xabarlardan (SMS) aloqa va marketing vositasi sifatida foydalanmoqda. Slicktext [1] tomonidan chop etilgan maqolaga ko'ra, butun dunyo bo'ylab besh milliard odam SMS dan foydalanadi, bu

butun insoniyatning 65% ni tashkil qiladi. Hisob-kitoblarga ko‘ra, 2025-yilga kelib bu raqam 5,9 milliard foydalanuvchiga yetishi mumkin. Bu SMSlarning ko‘payishi spam va shunga o‘xshagan ko‘plab zararli harakatlarning ko‘payishi bilan bir vaqtga to‘g‘ri keldi. Foydalanuvchilarni bezovta qilishdan tashqari, bu

harakatlar jismoniy shaxslar va korxonalar uchun jiddiy moliyaviy oqibatlarga olib kelishi mumkin [2]. Odatda, spam xabarlarni jo‘natuvchilar ya’ni spammerlar SMS orqali shaxsiy yoki moliyaviy ma’lumotlarni olishni maqsad qilib qo‘yishadi, ular orasida firibgar kontent, zararli havola yoki zararli dasturlar bo‘lishi mumkin. 2023-yil fevral oyida [3] e’lon qilingan statistik ma’lumotlarga ko‘ra, spam hajmi o‘n yil davomida sezilarli darajada oshgan. 2021-yil sentabrida 1,27 million spam xabar jo‘natilgan. Taqqoslash uchun, 2022-yil avgust oyida 10,89 milliard spam-xabar jo‘natilgan. Moliyaviy yo‘qotishlarga kelsak, spam matnlar tufayli 2021 yilda taxminan 10 066 331 169 AQSh dollari yo‘qolgan.

So‘nggi yillarda SMS spamlarini aniqlash bo‘yicha bir nechta tadqiqot ishlari taklif qilindi. Ushbu ishlarda klassik filtrlash usullaridan sun’iy intellektning ilg‘or algoritmlarigacha bo'lgan turli usullardan foydalanilgan. An’anaviy filtrlash usullari sodda va tezroq bo‘lsa-da, sun'iy intellektga asoslangan usullar, ayniqsa, yangilangan spam kontent bilan ishlashda aniqlash tezligi jihatidan samaraliroqdir. Barcha filtrlash usullarining asosiy maqsadi xabarlar mazmunini tahlil qila oladigan va ularni spam sifatida tasniflashi mumkin bo'lgan mashinali o‘qitish algoritmlari asosida aqlli modellarni yaratishdir. Taklif etilayotgan modellarning ishlashi bir qancha omillarga, jumladan, foydalanilgan matnni ko‘rsatish texnikasiga, xususiyatni tanlash usuliga, tasniflash algoritmi turiga va boshqalarga qarab o‘zgaradi.Demak asosiy maqsad spamni aniqlashda eng so‘nggi natijalarga erisha oladigan mustahkam modelni ishlab chiqishdir. Shu nuqtai nazardan, biz bir nechta tasniflash usullarini o‘rganib chiqdik va spam xabarlarni aniqlashda modifikatsiyalangan vektor usulidan foydalandik.

Modifikatsiyalangan vektor usuli (MBVM) elektron pochta xabarlarida spam-xabarlarni aniqlash uchun ishlatiladigan usuldir. Bu asosiy vektor usulining (BVM) kengaytmasi bo'lib, spam-xabarlarni mazmuniga qarab aniqlash uchun xuddi shunday yondashuvdan foydalanadi. Modifikatsiyalangan vektor usuli muayyan kalit so‘zlar yoki iboralarning paydo bo‘lish chastotasiga asoslangan har bir elektron pochta xabarining vektor tasvirini yaratish orqali ishlaydi. Keyin bu vektorlar xabar spam yoki yo‘qligini aniqlash uchun ma’lum spam xabarlar bazasi bilan taqqoslanadi. Modifikatsiyalangan vektor usuli ushbu yondashuvni har bir xabarning vektor ko‘rinishiga qo‘shimcha funktsiyalarni kiritish orqali yaxshilaydi. Bu xususiyatlar xabarning uzunligi, undagi havolalar soni va ba'zi HTML teglarining mavjudligini o‘z ichiga oladi. Ushbu qo‘shimcha funktsiyalarni hisobga olgan holda, Modifikatsiyalangan vektor usuli qonuniy va spam xabarlarni aniqroq ajrata oladi.

Modifikatsiyalangan vektor usulini amalga oshirish uchun birinchi navbatda asosiy vektorlar ma’lumotlar bazasini yaratish uchun ma’lum spam va spam bo‘lmagan

xabarlarning o‘quv to‘plamidan foydalaniladi. Keyinchalik bu asosiy vektorlar yuqorida aytib o‘tilgan qo‘shimcha funktsiyalarni kiritish uchun o‘zgartiriladi. Yangi xabar qabul qilinganda, uning vektori spam yoki spam emasligini aniqlash uchun o‘zgartirilgan asosiy vektorlar bilan taqqoslanadi.

Modifikatsiyalangan vektor usuli spam-xabarlarning ko‘p turlarini aniqlashda samarali bo‘lsada, u ishonchli emas va hali ham murakkab spamerlar tomonidan chetlab o‘tilishi mumkin. Shuning uchun, u ko‘pincha boshqa spamlarni aniqlash usullari bilan birgalikda istalmagan elektron pochta xabarlariga qarshi yanada mustahkam himoya qilish uchun ishlatiladi.

Modifikatsiyalangan vektor usulining ishlash algoritmi. Klassifikator sifatida tayanch vektor usulida ikkala sinfni optimal tarzda R^nfazoda ajratuvchi gipertekislikning

$$\omega_1 x_1 + \omega_2 x_2 + \cdots + \omega_n x_n + \omega_0 = 0$$

tenglamasini topish hisoblanadi. x ob’ektining F funktsiyasini Y sinf yorlig‘iga aylantirishning umumiy ko‘rinishi:

$F(x) = sign(\omega^T x - b)$.Bu yerda

$\omega = (\omega 1, \omega 2, \ldots . , \omega n),\ b = -\omega 0$

belgilashlar kiritilgan. ω va b o‘qitish algoritm vazn koeffitsientlari (noma’lum parametrlari) sozlangandan so‘ng, qurilgan gipertekislikning bir tomoniga tushgan barcha obyektlar birinchi sinf, ikkinchi tomoniga tushgan obyektlar esa ikkinchi sinf sifatida belgilanadi.

Agar quyidagi shart bajarilsa, algoritm obyektlarni to‘g‘ri tasniflaydi, ya’ni to‘g‘ri sinflarga ajratadi:

$(\omega Tx - b) \geq 1$

Unga ko‘ra moslashuvchan tasodifiy qidiruv usulining umumiy holida quyidagi optimallashtirish masalasi qo‘yilgan bo‘lsin.

$(\omega) \rightarrow min => \omega*$ (1)

$\omega \in D$

bu yerda q(ω) – umumiy holda nochiziqli ko‘p o‘zgaruvchili funktsiya.

$\omega = (\omega 1, \omega 2, \ldots . , \omega n)$,

$\omega* = (\omega*, \omega* , \ldots . . , \omega*)$– (1) masalaning yechimi.

12 n

Moslashuvchan tasodifiy qidiruv usulida quyidagi rekkurent formuladan foydalaniladi:

$\omega k+1 = \omega k + \Delta\omega k+1$,

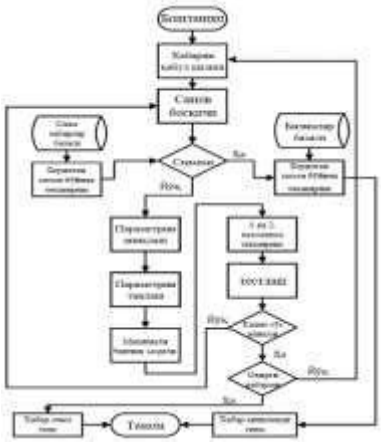
$\Delta\omega k+1$

$ak+1\Delta\omega k, \quad agarda\ (\omega k) < q(\omega k-1)$

$= \{ak+1 \cdot \xi k+1, agard\ (\omega k) \geq q(\omega k-1)$

Yuqoridagi formulalarga asoslanib spam xabalar aniqlandi.Spam xabarlarni aniqlashning modifikatsiyalangan tayanch vector usulida spam xabarlarni testlash

jarayonida k=5572 nazorat tanlanmasi uchun to‘g‘ri toifalash ko‘rsatkichlari “spam” sinfi obyektlari uchun 4825 ta obyektdan 4785 tani tani (99.17%), “ham” yoki boshqa holler sinfi obyektlari uchun 747 ta obyektdan 736 tani (98.5) ni tashkil etdi.



1-rasm.Modifikatsiyalangan tayanch vektor algoritmining blok sxemasi.

Xulosa. Spam xabarlarni aniqlashning modifikatsiyalangan tayanch vektor usuli va algoritmi ishlab chiqildi. Natijada F klassifikator asosida hujjatlarni, aniqlangan tavsifi asosida, ikki toifaga ajratish imkoniyatiga erishildi. Modifikatsiyalangan tayanch vektor usuli asosida ishlab chiqilgan spam xabarlarni aniqlashning dasturiy vositasi 97.3% aniqlikni ko’rsatdi.

FOYDALANILGAN ADABIYOTLAR:

1. SlickText. 44 Mind-Blowing SMS Marketing and Texting Statistics. 2023. Available online: https://www.slicktext.com/blog/2018/11/44-mind-blowing-sms-marketing-and- texting-statistics/ (accessed on 26 February 2023).

2. Sonowal, G.; Kuppusamy, K.S. SmiDCA: An Anti-Smishing Model with Machine Learning Approach. Comput. J. 2018, 61, 1143–1157. [Google Scholar] [CrossRef]

3. SlickText. 17 Spam Text Statistics &Spam Text Examples. 2023. Available online: https://www.slicktext.com/blog/2022/10/17-spam-text-statisitics-for-2022/ (accessed on 26 February 2023).

4. R.Khamdamov, E.Haydarov. Mathematical Model and Methods for Filtering an Email Message // International Conference on Information Science and Communications Technologies: Applications, Trends and Opportunities (ICISCT 2021). Tashkent-2021. -4p.

(3) Scopus

5. R.Khamdamov, E.Haydarov. Detecting spam messages using the naive Bayes algorithm of basic machine learning // International Conference on “Information Science and Communications Technologies: Applications, Trends and Opportunities (ICISCT 2021). Tashkent-2021. -3p. (3) Scopus