Sun’iy intellekt yordamida spam-xabarlarni filtrlashning algoritmik tahlili

Normatov Ibroximali Xolmamatovich,
f-m.f.d., professor, O‘zMU,
Toshkent shahar, O‘zbekiston
Email: i_normatov@nuu.uz

Atajonov Muzaffar Ne’matjon o‘g‘li,
Jaloliddin Manguberdi nomidagi harbiy-akademik litsey
Email: muzaffar19910627@gmail.com

Annotatsiya: Ushbu maqolada sun’iy intellektda mashinali o‘rganish modellari yordamida spam-xabarlarni fitrlash usullari keltirilgan. Spamni avtomatik aniqlashda mashina o‘rganish algoritmlari keng qo‘llaniladi, chunki ular katta hajmdagi ma’lumotlarni o‘rganib, yangi xabarlarni samarali tasniflay oladi. Shu sababli spam-filtrlash uchun eng ko‘p qo‘llaniladigan mashina o‘rganish algoritmlari ya’ni Naive Bayes, Decision Tree, Random Forest va Support Vector Machine (SVM)lar o‘rganilgan. Shu bilan birgalikda modellarni bir-biridan farqli jihatlari, afzalliklari va kamchiliklari aniqlangan.

Kalit so‘zlar: TF-IDF, Root node, Leaf nodes, Precision, Recall, F1-score, spam, ham.

Kirish. Machine learning (ML) metodining spamni aniqlashda afzallik tomonlari juda ko‘p hisoblanadi. ML metodi xabarlarni tahlil qiladi, o‘zida bor ma’lumotlar bilan solishtiradi. Bunda spamni aniqlash uchun uning belgilarini o‘rganib chiqadi va xabarni spam yoki ham (haqiqiy) toifalarga ajratadi. Spam-filtrlashda mashina o‘rganish algoritmlari xabarlarning matnini yoki boshqa atributlarini tahlil qilib, ularni spam yoki normal deb tasniflash uchun model yaratadi. Spam-filtrlashda asosan nazoratli o‘rganish modeli ishlatiladi, bunda oldindan belgilangan (spam yoki nospam) ma’lumotlar yordamida model yaratiladi. Spam-filtrlash uchun eng ko‘p qo‘llaniladigan mashina o‘rganish algoritmlari bular Naive Bayes, Decision Tree, Random Forest va Support Vector Machine (SVM)lardir. Ushbu maqolada aynan shu metodlar o‘rganilgan.

Adabiyotlar tahlil va metodologiya. Ushbu maqolada bir qancha tadqiqotchilarni ishlari ya’ni spam-xabarlarni filtrlashda qilingan ishlar tahlil qilindi. Vangelis Metsis, Ion Androutsopoulos va Georgios Paliouraslarning 2006-yilda yozilgan “Spam Filtering with Naive Bayes – Which Naive Bayes?” nomli maqolasida Naive Bayes klassifikatori qo‘llangan. Gareth James, Daniela Witten, Trevor Hastie va Robert Tibshiranilarning 2021-yildagi “An Introduction to Statistical Learning” nomli maqolasida Decision Tree modeli qo‘llangan. 2020-yilda yozilgan S.Kumar va A.Kumarlarning “Email spam detection using ensemble classifiers” maqolasida Random Forest algoritmi o‘rganilgan va 2002-yilda chop qilingan F.Sebastianining “*Machine Learning in Automated Text Categorization”* nomli maqolasida Support Vector Machine usuli o‘rganilgan.

Naive Bayes klassifikatori Bayes teoremasiga asoslangan va so‘zlarning mustaqil bo‘lish taxminiga ega probabilistik modeldir. Bu usul spam-filtrlashda ko‘p qo‘llaniladi, chunki u hisoblash jihatdan samarali va yuqori aniqlikka ega. Uning ishlash printsipi — har bir xabarning sinfga tegish ehtimolini hisoblash va eng yuqori ehtimolga ega sinfni tanlash. Uning asosiy g‘oyasi, berilgan ma’lumotlar (xabar matni) asosida ularning ma’lum bir sinfga (masalan, "spam" yoki "nosspam") tegishli bo‘lish ehtimolini hisoblashdir [1].

Bayes teoremasi:

$$P(C|X) = \frac{P(X|C)\cdot P(C)}{P(X)}, \qquad (1)$$

bu yerda:

- $P(C|X)$ – X xabar ning C sinfiga tegishli bo‘lish ehtimoli;
- $P(X|C) -$ C sinfda X xabarning paydo bo‘lish ehtimoli;

- $P(C)$ – sinfning oldingi ehtimoli;
- $P(X)$ – X xabarning umumiy ehtimoli.

Naive Bayes algoritmi matnni tashkil etuvchi so‘zlarning mustaqil (independent) ekanligini taxmin qiladi. Ya’ni, har bir so‘z boshqa so‘zlardan mustaqil ravishda paydo bo‘ladi deb hisoblanadi. Shu sababli, "Naive" (ya’ni sodda, soddalashtirilgan) nomini olgan.

Spam-filtrlashda Naive Bayes klassifikatori har bir elektron pochta yoki xabar matnidagi so‘zlarning sinfga ta’sirini hisobga olib, xabar spam yoki nospam (ham) ekanligini aniqlaydi. Algoritm quyidagi bosqichlarni bajaradi:

1. O‘qitish bosqichi: O‘qitish ma’lumotlarida (oldindan belgilangan spam va nospam xabarlar) har bir so‘zning spam va nospam sinflaridagi paydo bo‘lish chastotasi hisoblanadi.

Misol uchun, "bonus", "pul", "yutqazish" kabi so‘zlar spam xabarlarda ko‘p uchrashi mumkin.

2. Tasniflash bosqichi: Yangi kelgan xabar uchun uning so‘zlari bo‘yicha har bir sinfga tegish ehtimoli hisoblanadi. So‘zlarning mustaqilligi taxminiga ko‘ra, umumiy ehtimol har bir so‘zning ehtimollari ko‘paytmasi sifatida olinadi:

$$P(X|C) = \prod_{i=1}^{n} P(x_i|C), \quad (2)$$

bu yerda x_i –xabardagi i-chi so‘z.

3. Qaror qabul qilish: Hisoblangan ehtimollarga ko‘ra, eng yuqori posterior ehtimolga ega sinf tanlanadi, ya’ni xabar spam yoki nospam (ham) deb tasniflanadi.

Naive Bayesning afzalliklari:

- Tez va samarali hisoblash: Naive Bayes algoritmi katta hajmdagi matnli ma’lumotlar bilan tez ishlay oladi.
- Kam o‘qitish ma’lumotlarida ham yaxshi natija: Kam namunalar bilan ham samarali ishlashi mumkin.
- Matnli tasniflashda juda samarali: So‘zlarning mustaqilligi taxmini spam-filtrlashda ko‘pincha yetarli darajada to‘g‘ri natija beradi.

Naive Bayesning kamchiliklari:

- So‘zlarning mustaqilligi taxmini real hayotda har doim ham to‘g‘ri emas: Masalan, "pul" va "bonus" so‘zlari birga ko‘p uchrashi mumkin, ammo ular mustaqil deb hisoblanadi, bu esa ba’zan natija aniqligini pasaytiradi.
- Murakkab bog‘lanishlarni aniqlay olmaydi: So‘zlar orasidagi kontekstual bog‘liqliklarni hisobga olmaydi.

Spam-filtrlashda quyidagi Naive Bayes turlari ko‘proq qo‘llaniladi:

- Multinomial Naive Bayes: Matndagi so‘z chastotasi hisobga olinadi. Bu ko‘p ishlatiladigan va samarali model.
- Bernoulli Naive Bayes: So‘zning mavjudligi yoki mavjud emasligi (0 yoki 1) shaklida ko‘riladi.
- Gaussian Naive Bayes: Doimiy qiymatli atributlar uchun, matnli ma’lumotlarga kamroq mos keladi.

Spam-filtrlash uchun odatda Multinomial Naive Bayes eng mos keladi.

Decision Tree — bu qarorlar qabul qilish jarayonini daraxt shaklida ifodalovchi algoritmdir. Har bir tugun (node) — biror atribut bo‘yicha qaror, har bir shox — natijaviy yo‘nalishni bildiradi [2].

Asosiy tushunchalar:

- Korni (Root) — daraxtning boshlanish nuqtasi, eng muhim atributga asoslanadi.
- Ichki tugunlar (Internal nodes) — boshqa atributlar bo‘yicha bo‘linishlar.
- Barg tugunlar (Leaf nodes) — yakuniy qarorlar (masalan, spam / ham).

Spam-xabarlarni filtrlashda ma’lumotlar ko‘p hollarda matnli bo‘ladi, shu sababli ularni raqamli formatga o‘tkazish quyidagi boshqichlardan iborat bo‘ladi:

- Matnni tozalashlaydi ya’ni HTML teglarini, raqamlarni, belgilarni olib tashlaydi.
- Tokenizatsiya jarayoni ya’ni matnni so‘zlarga bo‘lish jarayonini amalga oshiradi.

- Stop-so‘zlarni olib tashlaydi. Bunda “va”, “bu”, “men” kabi ma’nosiz so‘zlarni chiqarib tashlaydi.
- Lemmatizatsiya yoki stemming: so‘zlarni ildiz shakliga keltirish.
- Xususiyatlarni vektorlash ishlari. Bunda TF-IDF (Term Frequency – Inverse Document Frequency) – eng mashhur usul va N-gramlar – so‘z juftliklari yoki uchliklari usulidan foydalanadi [3].

Decision Tree qo‘shimcha belgilarga ham alohida e’tibor qaratadi. Ya’ni havolalar soni, “Free”, “click”, “offer” so‘zlarining mavjudligi, xabar uzunligi va yuboruvchi domenini tekshiradi.

Spam xabarlarni aniqlash va ularni filtrlash masalasi mashina o‘rganishning muhim amaliy yo‘nalishlaridan biri Random Forest algoritmi hisoblanadi. Random Forest — bu ko‘p sonli qaror daraxtlaridan iborat ansambl model bo‘lib, har bir daraxt mustaqil o‘qitiladi va yakuniy qaror ko‘pchilik ovoz berish (majority voting) asosida qabul qilinadi [4].

Algoritm bosqichlari:

1. Ma’lumotlar to‘plami bir nechta kichik to‘plamlarga bootstrap sampling orqali tasodifiy ajratiladi.
2. Har bir kichik to‘plam uchun Decision Tree alohida o‘qitiladi.
3. Har bir daraxtda atributlar (xususiyatlar) to‘liq emas, balki tasodifiy tanlangan kichik qismidan foydalaniladi.
4. Har bir daraxt alohida bashorat beradi (xabar spam yoki spam emas).
5. Yakuniy qaror ko‘pchilik daraxtlarning ovoziga qarab belgilanadi.

Agar $h_1(x), h_2(x), \dots, h_n(x) -$ tasodifiy o‘qitilgan qaror daraxtlari bo‘lsa, Random Forest’ning yakuniy natijasi quyidagicha aniqlanadi:

$$H(x) = mode\{h_1(x), h_2(x), \dots, h_n(x)\},$$

bu yerda $mode() -$ eng ko‘p uchragan qiymat (ya’ni, ko‘pchilik ovoz) [5].

Endi spam xabarlarni aniqlash jarayonini ko‘rib chiqamiz.

1. Ma’lumotlarni tayyorlash

- Xabar matnlari to‘planadi (spam va normal).
- Text preprocessing amalga oshiriladi:
  - Katta harflarni kichikka aylantirish
  - Noto‘g‘ri belgilarni olib tashlash
  - Stop-word so‘zlarni chiqarib tashlash
  - Lemmatizatsiya yoki stemming jarayonlari

2. Xususiyatlarni ajratish (Feature Extraction)

Spam-xabarlarni matematik modelda ifodalash uchun matn raqamli vektor shakliga o‘tkaziladi:

- Bag of Words (BoW) modeli
- TF-IDF (Term Frequency – Inverse Document Frequency)

3. Modelni o‘qitish

- Tayyorlangan ma’lumotlar o‘quv (train) va test (test) to‘plamlariga ajratiladi.
- Random Forest modeli train to‘plamda o‘qitiladi.

4. Baholash [6].

Modelning samaradorligi quyidagi ko‘rsatkichlar orqali baholanadi:

- Accuracy (aniqlik)
- Precision (aniqlik darajasi)
- Recall (chaqiriq)
- F1-score (balanslangan ko‘rsatkich)

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}} \quad (3)$$

$$\text{F1} - \text{score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision+Recall}} \quad (4)$$

Support Vector Machine (SVM) — bu chiziqli ajratuvchi (linear classifier) bo‘lib, u ikki sinfni eng katta masofa (margin) bilan ajratadigan gipertekislikni topishga harakat qiladi.

Berilgan ma’lumotlar to‘plami:

$$(x_i, y_i),\ y_i \in \{-1; 1\} \quad (5)$$

SVM quyidagi gipertekislikni topadi:

$$\omega \cdot x + b = 0 \quad (6)$$

Bunda ω — vazn vektori (hyperplane yo‘nalishini aniqlaydi), x — kirish ma’lumoti (masalan, xabar xususiyatlari), b — siljish (bias).

Modelning asosiy maqsadi marginni quyidagi ko‘rinishda:

$$\text{maximize} \frac{2}{||\omega||} \quad (7)$$

maksimal qilishdan iborat bo‘lib, barcha nuqtalar uchun:

$$y_i(\omega \cdot x_i + b) \geq 1 \quad (8)$$

tengsizlik o‘rinli bo‘ladi. Bu yerda kvadratik optimizatsiyalash quyidagicha yoziladi:

$$\min_{\omega,b} \frac{1}{2} ||\omega||^2 \quad (9)$$

cheklov sharti bilan:

$$y_i(\omega \cdot x_i + b) \geq 1 \quad (10)$$

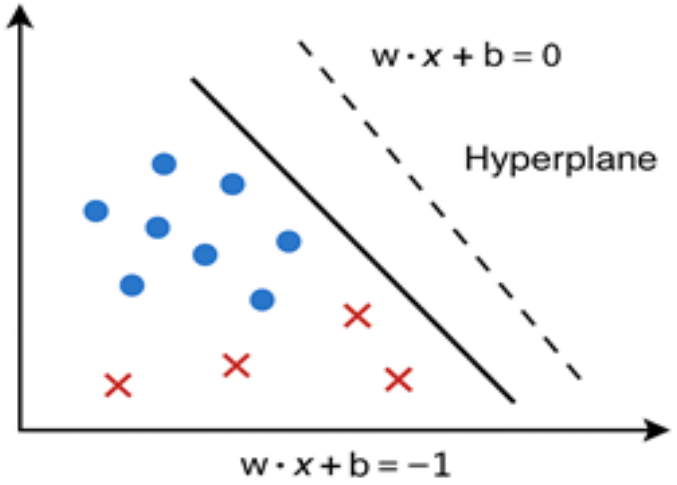
Haqiqiy ma’lumotlar har doim ham ideal ajratilmaydi, shu bois “yumshoq margin” (soft margin) kiritiladi:

$$\min_{\omega,b,\varepsilon} \frac{1}{2} ||\omega||^2 + C \sum_{i=1}^{n} \varepsilon_i \quad (11)$$

$$y_i(\omega \cdot x_i + b) \geq 1 - \varepsilon_i, \quad \varepsilon_i \geq 0, \quad (12)$$

bu yerda C — regularizatsiya parametri bo‘lib, margin kengligi va xatoliklar o‘rtasidagi muvozanatni belgilaydi.

Quyidagi 1-rasmni tahlil qilamiz.



1-rasm. Support Vector Machine (SVM) algoritmining asosiy g‘oyasi ya’ni gipertekislik (hyperplane).

1-rasmda Support Vector Machine (SVM) algoritmining asosiy g‘oyasi — ya’ni ikki sinfni eng katta masofa bilan ajratadigan gipertekislik (hyperplane) tushuntirilgan. Bunda dumaloq nuqtalar birinchi sinf (masalan, "spam emas") namunalarini ifodalaydi va x beldilar ikkinchi sinf (masalan, "spam") namunalaridir. Uzun chiqiz (hyperplane) esa ajratish chizig‘i (gipertekislik) deyiladi. U sinflarni eng katta margin (oraliq) bilan ajratadi. Ushbu ikki punktir chiziqqa eng yaqin turgan nuqtalar — “support vectorlar” deb ataladi. Ular hyperplane’ning joylashuvini aniqlaydi [7].

Natijalar. SVM spam filtrlash tizimlari uchun nazariy jihatdan asosli, amaliy jihatdan esa samarali yechimlardan biridir. Kelajakda bunday tizimlarni yanada takomillashtirish uchun SVMni chuqur o‘rganish modellari (Deep Learning) yoki ensemble usullari bilan birlashtirish istiqbollidir.

Yuqorida qilingan tadqiqotlar va ko‘rib chiqqan algoritmlarimizdan quyidagi jadval ko‘rinishdagi natijalarga ega bo‘lamiz (1-jadval):

1-jadval

Algoritm	**Asosiy tamoyil**	**Afzalliklari**	**Aniqlik (%)**	**F1-score**
Naive Bayes	Ehtimollik modeli	Tez, oddiy, matn uchun qulay	90–92	0.89
Decision Tree	Shartli qarorlar daraxti	Tushunarli, vizual	88–90	0.87
Random Forest	Ansambl (ko‘p daraxtlar)	Barqaror, yuqori aniqlik	94–96	0.94
SVM	Maksimal marginli gipertekislik	Yuqori o‘lchamda samarali	95–97	0.95

Xulosa

Naive Bayes — tezkor va soddaligi bilan kichik tizimlar uchun juda qulay.

Decision Tree — tushunarli model, ammo overfitting xavfi mavjud. spam-xabarlarni filtrlashda **soddaligi, izohlanishi va tezligi** sababli samarali usullardan biridir. U, ayniqsa, tushuntirish va audit talab etiladigan tizimlarda juda foydali.

Random Forest — eng barqaror va amaliy jihatdan kuchli algoritmlardan biri. Random Forest algoritmi ma’lumotlar hajmi kattalashganda ham barqaror ishlaydi, turli turdagi matnlarda yuqori aniqlikni saqlaydi va overfitting muammosiga nisbatan Decision Tree va Naive Bayesdan yaxshiroq himoyalangan bo‘ladi. Bundan tashqari Random Forest algoritmi spam-xabarlarni aniqlashda ishonchli va yuqori aniqlikka ega usuldir.

SVM — aniqlik va umumlashma qobiliyati eng yuqori bo‘lgan algoritm sifatida spam filtrlashda eng yaxshi natijani ko‘rsatadi.

Shu boisdan, amaliy tahlillar shuni ko‘rsatadiki, SVM va Random Forest algoritmlari spam-xabarlarni filtrlash uchun eng ishonchli yechimlar hisoblanadi.

Naive Bayes esa tezlik va resurs tejamkorligi sababli real vaqt tizimlarida afzal bo‘lishi mumkin.

Foydalanilgan adabiyotlar

1. Vangelis Metsis, Ion Androutsopoulos, Georgios Paliouras “Spam Filtering with Naive Bayes – Which Naive Bayes?”, reseach gate, 2006-yil yanvar, 2-3-betlar.

2. Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani “An Introduction to Statistical Learning” with Applications in R Second Edition, 2021-yil august, 227-228-betlar.

3. Kabulov A.V., Normatov, I.Kh., Muhammadiev F.R. Invariant continuation of discrete multi-valued functions and their implementation// Published in: 2021 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS), Date Added to IEEE Explore 14 May 2021, p. 1-5. ISBN Information: 20692695, doi:10.1109/IEMTRONICS52119.2021.9422486 Publisher: IEEE Conference Location: Toronto, ON, Canada (Scopus).

4. Anvar Kabulov, Ibrokhimali Normatov, Ilyos Kalandarov, Inomjon Yarashov Development of An Algorithmic Model And Methods For Managing Production Systems Based On Algebra Over Functioning Tables// Published in: 2021 International Conference on Information Science and Communications Technologies (ICISCT), Date Added to IEEE Xplore: 17.01.2022 pp.1-4, ISBN Information: 21572756 doi: 10.1109/ICISCT52966.2021.9670307 Publisher: IEEE Conference Location: Tashkent, Uzbekistan (Scopus).

5. I.Normatov, I.Yarashov, A.Otakhonov and B.Ergashev Construction of reliable well distribution functions based on the principle of invariance for convenient user access control// Published in:2022 International Conference on Information Science and Communications Technologies (ICISCT),Date Added to IEEE Xplore:14 June 2023 pp. 1-5, ISBN: 23280391, doi 10.1109/ICISCT55600.2022.10146952 Publisher: IEEE Conference (**S**copus)

6. S.Kumar, A.Kumar “Email spam detection using ensemble classifiers” International Journal of Information Technology, 12(3)-soni, 2020-yil, 799-805-betlar.

7. Normatov Ibrokhimali Endless individual areas of logic and beginnings of arithmetics// Modern problems of applied mathematics and information technology (MPAMIT 2021) pp. 1-7, Fergana, Uzbekistan *AIP Conf. Proc.* 2781, 020008 (2023) doi.org/10.1063/5.0144824 (Scopus).