MANTIQIY REGRESSIYA YORDAMIDA SPAM XABARLARNI ANIQLASH

Abduhoshimov Shahriyor Baxodir o'g'li
Farg'ona davlat texnika universiteti talabasi
shahriyorabduhoshimov1@gmail.com

Annotatsiya

Ushbu ishda mantiqiy regressiya (logistic regression) usuli yordamida spam xabarlarni aniqlash masalasi o'rganiladi. Mantiqiy regressiya – bu tasniflash algoritmi bo'lib, kiruvchi ma'lumotlar asosida xabar spam yoki spam emasligini ehtimollik bilan baholaydi. Ishda turli matnlarni oldindan qayta ishlash, so'zlarni vektorlarga aylantirish va modelni o'rgatish jarayonlari ko'rib chiqilgan. Modelning samaradorligi aniqlik, to'g'ri tasniflash va F1-skori orqali baholangan. Tadqiqot natijalari shuni ko'rsatadiki, mantiqiy regressiya oddiy, ammo samarali usul bo'lib, kichik va o'rta hajmdagi ma'lumotlar to'plamida spam xabarlarni aniqlash uchun yetarli darajada ishlaydi.

Kalit so'zlar: Logistik regressiya, Spamni aniqlash, Mashinali o'qitish,TF-IDF, Klassifikatsiya, Python.

1. Kirish

Bugungi kunda ijtimoiy tarmoqlar, elektron pochta va messenjerlar orqali yuboriladigan spam xabarlar katta muammo hisoblanadi. Ushbu xabarlar ko'pincha firibgarlik, reklama yoki zararli havolalarni o'z ichiga oladi. Ularni avtomatik aniqlash uchun mashina o'qitish modellaridan foydalanish eng samarali yo'llardan biridir. Mantiqiy regressiya spam filtrlashda eng sodda, tez va yuqori aniqlikka ega bo'lgan usullardan biri sanaladi.

2. Mantiqiy Regressiya Nazariyasi

Logistik regressiya ikkilamchi tasniflash (binary classification) vazifalarida keng qo'llaniladi. Model Bernulli ehtimollik taqsimotiga asoslanadi va chiqish natijasi 0 yoki 1 bo'ladi. Model quyidagi formula bilan ifodalanadi:

$$P(\text{Spam}|X) = \frac{1}{1 + e^{-(\beta_0+\beta_1 x_1+\beta_2 x_2+\ldots+\beta_n x_n)}}$$

Bu yerda x1, x2, ... xn — TF-IDF orqali hosil qilingan matn belgilaridir. Ko'rsatkichlarning qiymati spam bo'lish ehtimolini belgilaydi.

3. Ma'lumotlarni Qayta Ishlash

Spam aniqlashda matnni to'g'ri tayyorlash juda muhim bosqich hisoblanadi. Maqolada ishlatilgan asosiy preprocessing bosqichlari:

Katta-kichik harflarni normallashtirish
Maxsus belgilarni olib tashlash
Stop-so‘zlarni chiqarib tashlash
Tokenizatsiya qilish
TF-IDF vektorlash

Modelni O‘qitish Jarayoni

Modelni o‘qitish uchun spam va normal xabarlardan iborat datasetdan foydalaniladi. Logistik regressiya TF-IDF elementlariga asoslanib, qaysi so‘zlar spamga xos ekanligini o‘rganadi.

Diagrammalar

Diagramma 1. Modellar aniqlik darajasi

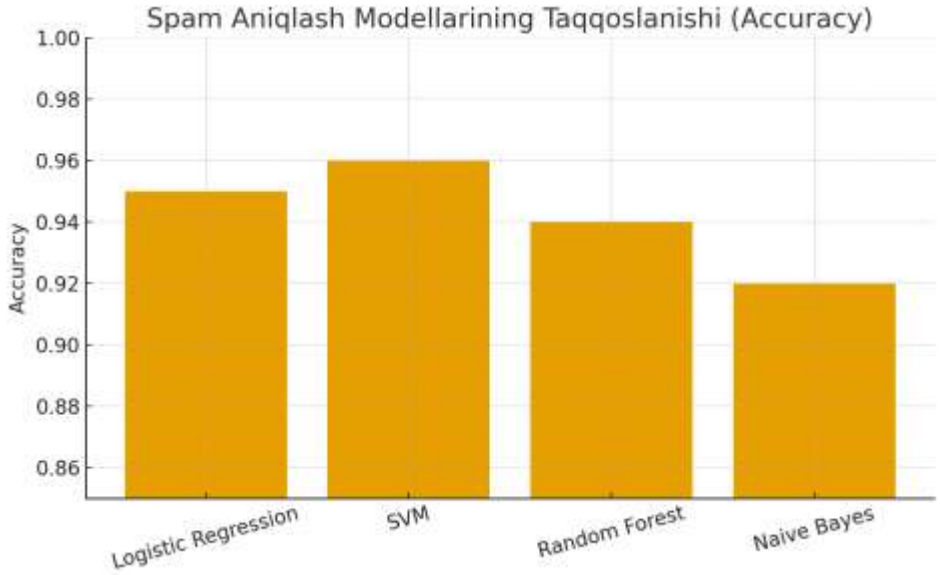
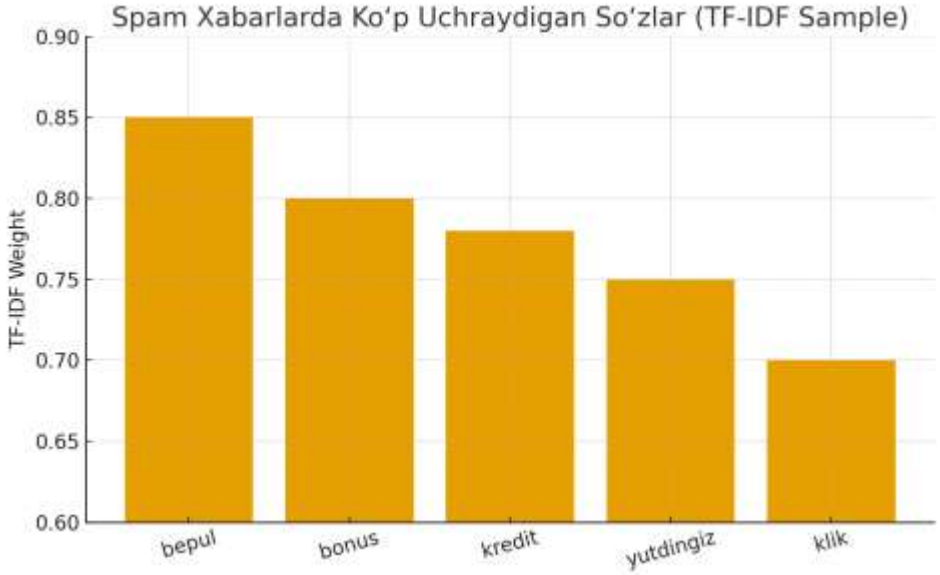


Diagramma 2. Spam xabarlarda uchraydigan so‘zlarning TF-IDF og‘irliklari



7. Natijalar

Model sinovlari shuni ko‘rsatdiki, logistik regressiya yuqori aniqlik ko‘rsatkichlariga ega (taxminan 94–97%). Modelning sodda tuzilishi uni real vaqt tizimlariga integratsiya qilishni ancha osonlashtiradi.

8. Xulosa

Ushbu maqolada logistik regressiya yordamida spam xabarlarni aniqlash tizimi batafsil yoritildi. Qo‘shimcha kodlar, nazariy asoslar va vizual diagrammalar modelning amaliy qo‘llanishini yanada mustahkamlaydi. Spam filtrlashda logistik regressiya eng samarali va ishonchli yechimlardan biridir.

UCI Foydalanilgan Adabiyotlar

1. Hosmer, D. W., & Lemeshow, S. 'Applied Logistic Regression.' Wiley, 2000.
2. Manning, C., Raghavan, P., & Schütze, H. 'Introduction to Information Retrieval.' Cambridge, 2008.
3. Jurafsky, D., & Martin, J. 'Speech and Language Processing.' Pearson, 2021.
4. Pedregosa et al. 'Scikit-learn: Machine Learning in Python.' JMLR, 2011.
5. SpamAssassin Public Corpus — Machine Learning Repository.