

ТАҲҚИҚИ МУАЙЯНКУНИИ ЗАБОНИ КОДИ БАРНОМАВӢ ДАР АСОСИ ҲАҶФҲОИ ЛОТИНӢ

А.А. Қосимов¹, Ш.М. Асадуллозода², М.У. Шарифов³

¹Донишгоҳи техникии Тоҷикистон ба номи академик М.С. Осимӣ

²Донишгоҳи давлатии Бохтар ба номи Носири Хусрав

³Институти математикаи ба номи академик А.Ҷӯраевӣ АМИТ

Мақолаи мазкур ба тадқиқи масъалаи муайянкунии автоматии забони коди барномавӣ дар асоси таҳлили омории ҳаҷфҳои лотинӣ ва тадбири усулҳои зехни сунӣ бахшида шудааст. Бо дарназардошти афзоиши босуръати ҳаҷми коди ибтидоӣ, зарурати таҳияи усулҳои зехни таснифоти матнҳои барномавӣ дар соҳаҳои амнияти иттилоотӣ ва криминалистикаи рақамӣ мубрамияти хос пайдо кардааст. Дар раванди тадқиқот усулҳои таҳлили омории матн, метрикаҳои эҳтимолиятӣ (TF-IDF, биграммаҳо) ва алгоритмҳои муносири омӯзиши мошинӣ, аз ҷумла шабакаҳои нейронӣ ва меъморҳои трансформаторӣ истифода шудаанд. Коди барномавӣ ҳамчун объекти бисёрченака бо маҷмӯи хусусиятҳои беназири омӯрӣ баррасӣ гардидааст. Натиҷаҳои бадастомада нишон медиҳанд, ки тақсими рамзҳои лотинӣ, калимаҳои калидӣ ва сохторҳои синтаксисӣ барои ҳар як забони барномасозӣ беназир буда, ҳамчун «изи омӯрӣ»-и забон хизмат мекунад. Дар асоси таҳлили муқоисавӣ муайян карда шуд, ки моделҳои гибридӣ, ки таҳлили омӯро бо шабакаҳои нейронӣ муттаҳид месозанд, сатҳи баландтарини дақиқиро (то 98.1%) таъмин менамоянд. Аҳамияти амалии тадқиқот дар он ифода меёбад, ки натиҷаҳои бадастомадаро метавон дар системаҳои муайянсозии таъминоти барномавии зараровар, ошкорсозии плагиат, компиляторҳои зехнӣ ва платформаҳои таҳлили додаҳои калон бомуваффақият тадбир намуд.

Калидвожаҳо: зехни сунӣ, омӯзиши мошинӣ, коди барномавӣ, шинохти намунаҳо, таҳлили омӯрӣ, шабакаҳои нейронӣ, забон, таснифоти додаҳо, криминалистикаи рақамӣ.

ИССЛЕДОВАНИЕ ОПРЕДЕЛЕНИЯ ЯЗЫКА ПРОГРАММНОГО КОДА НА ОСНОВЕ ЛАТИНСКИХ БУКВ

А.А. Қосимов, Ш.М. Асадуллозода, М.У. Шарифов

Данная статья посвящена исследованию задачи автоматического определения языка программного кода на основе статистического анализа латинских символов и применения методов искусственного интеллекта. С учетом стремительного роста объемов исходного кода, необходимость разработки интеллектуальных методов классификации программных текстов в сфере информационной безопасности и цифровой криминалистики приобрела особую актуальность. В процессе исследования использовались методы статистического анализа текста, вероятностные метрики (TF-IDF, биграмы) и современные алгоритмы машинного обучения, включая нейронные сети и трансформерные архитектуры. Программный код рассматривался как многомерный объект с набором уникальных статистических признаков. Полученные результаты показывают, что распределение латинских символов, ключевых слов и синтаксических конструкций является уникальным для каждого языка программирования, выступая в качестве «статистического отпечатка» языка. На основе сравнительного анализа установлено, что гибридные модели, объединяющие статистический анализ с нейронными сетями, обеспечивают наиболее высокий уровень точности (до 98,1%). Практическая значимость исследования заключается в том, что полученные результаты могут быть успешно применены в системах обнаружения вредоносного программного обеспечения, выявления плагиата, интеллектуальных компиляторах и платформах анализа больших данных.

Ключевые слова: искусственный интеллект, машинное обучение, программный код, распознавание образов, статистический анализ, нейронные сети, язык, классификация данных, цифровая криминастика.

RESEARCH ON PROGRAMMING LANGUAGE IDENTIFICATION BASED ON LATIN LETTERS

А.А. Kosimov, Sh.M. Asadullozoda, M.U. Sharifov

This article investigates the problem of automatic programming language identification based on the statistical analysis of Latin characters and the application of artificial intelligence methods. Given the rapid growth in the volume of source code, the need to develop intelligent methods for classifying programming texts in the fields of information security and digital forensics has become particularly relevant. The study employs methods of statistical text analysis, probabilistic metrics (TF-IDF, bigrams), and modern machine learning algorithms, including neural networks and transformer architectures. The source code is considered as a multidimensional object with a set of unique statistical features. The findings reveal that the distribution of Latin characters, keywords, and syntactic structures is unique to each programming language, acting as a "statistical fingerprint" of the language. Based on a comparative analysis, it was determined that hybrid models combining statistical analysis with neural networks provide the highest level of accuracy (up to 98.1%). The practical significance of the research lies in the fact that the obtained results can be successfully applied in malware detection systems, plagiarism detection tools, intelligent compilers, and big data analysis platforms.

Keywords: artificial intelligence, machine learning, source code, pattern recognition, statistical analysis, neural networks, language, data classification, digital forensics.

Рушди муносири технологияҳои иттилоотӣ бо афзоиши босуръати ҳаҷми коди барномавӣ, ки дар системаҳои автоматикунӣ, платформаҳои абрӣ ва инфрасохторҳои ҳисоббарории тақсимшуда таҳия мешаванд, дар бар мегирад. Афзоиши миқдори ҳалли барномавӣ, ки бо забонҳои гуногуни барномасозӣ навишта шудаанд, зарурати таҳияи усулҳои зехни муайянкунии автоматии забони коди барномавиро ба миён овард.

Масъалаи мазкур дар соҳаи таҳлили коди ибтидоӣ, амнияти иттилоотӣ, криминалистикаи рақамӣ, системаҳои санҷиши автоматиконидашуда, компиляторҳои зеҳнӣ ва платформаҳои таҳлили додаҳои калон аҳамияти назаррас дорад, [1-15].

Айни замон масъалаи муайянкунии забони барномасозӣ ҳамчун яке аз навҳои масъалаи шинохти намунаҳои сохторҳои матнӣ баррасӣ мегардад, ки дар он объекти таҳлил коди барномавии дар шакли рамзӣ пешниҳодшуда баромад мекунад. Сарфи назар аз мавҷудияти хусусиятҳои синтаксисӣ ва семантикии забонҳои гуногуни барномасозӣ, аксарияти забонҳои муосир ба истифодаи алифбои лотинӣ, рамзҳои махсус ва маҷмӯи маҳдуди сохторҳои хидматӣ асос ёфтаанд. Ин ҳолат имкон медиҳад, ки усулҳои таҳлили омории матнҳо, назарияи эҳтимолият, омӯзиши мошинӣ ва зеҳни сунӣ барои таснифоти автоматии коди барномавӣ истифода шаванд.

Мубрамиати таҳқиқот бо якчанд омилҳо муайян карда мешавад. Якум, репозиторийҳои муосири коди ибтидоӣ миллиардҳо сатри барномаҳоро дар бар мегиранд, ки бо дахҳо забонҳои барномасозӣ навишта шудаанд. Дуюм, системаҳои таҳлили автоматиконидашудаи барномаҳо барои коркарди дурусти сохторҳои синтаксисӣ муайянкунии пешакии забони кодро талаб мекунад. Сеюм, рушди технологияҳои зеҳни сунӣ ва таҳлили додаҳо барои сохтани моделҳои баланддиққати идентификацияи автоматии забонҳои барномасозӣ дар асоси қонуниятҳои омории замина фароҳам меорад.

Бояд қайд намуд, ки муайянкунии забони барномасозӣ масъалаи мураккаби таснифот ба шумор меравад, зеро бисёр забонҳо унсурҳои синтаксисии монанд доранд. Масалан, забонҳои C, C++, Java ва C# қавсҳои фигуравӣ, оператори бахшидан, идентификаторҳо дар асоси ҳарфҳои лотинӣ ва калимаҳои калидии шабеҳро истифода мебаранд. Ба ҳамин монанд, Python, Ruby ва Lua хусусиятҳои матнии умумӣ доранд, ки бо истифодаи калимаҳо ва идентификаторҳои англисӣ алоқаманданд. Ҳамин тариқ, масъалаи муайянкунии забони коди барномавӣ татбиқи усулҳои таҳлили амиқи оморию талаб мекунад.

Алифбои лотинӣ асоси бунёдии аксарияти забонҳои барномасозӣ мебошад. Аз нигоҳи таърихӣ, рушди забонҳои барномасозӣ дар муҳити илмии англисзабон ба амал омадааст, ки дар натиҷа калимаҳои калидӣ, идентификаторҳо, фармонҳо ва китобхонаҳои стандартӣ дар асоси забони англисӣ ташаккул ёфтаанд. Қариб ҳамаи забонҳои барномасозии муосир рамзҳои лотиниро барои сабти операторҳо, функсияҳо, тағирёбандаҳо ва сохторҳои идоракунанда истифода мебаранд. Масалан, дар забони Java сохторҳои зерин истифода мешаванд:

Java

```
public class Main {
    public static void main(String[] args) {
        System.out.println("Hello World");
    }
}
```

Дар забони Python коди шабеҳ намуди зерин дорад:

Python

```
def main():
    print("Hello World")

if __name__ == "__main__":
    main()
```

Дар забони JavaScript:

JavaScript

```
function main() {
    console.log("Hello World");
}
main();
```

Сарфи назар аз фарқиятҳои синтаксисӣ, дар ҳамаи мисолҳои овардашуда истифодаи ғайри ҳарфҳои лотинӣ, калимаҳои англисӣ ва рамзҳои махсус мушоҳида мешавад. Аз ин бармеояд, ки таҳлили омории алифбои лотинӣ метавонад ҳамчун асос барои сохтани моделҳои шинохти намунаҳои забонҳои барномасозӣ истифода шавад.

Таҳқиқот нишон медиҳад, ки забонҳои гуногуни барномасозӣ бо тақсими беназири рамзҳо тавсиф мешаванд. Масалан, забони Python бо басомади баланди истифодаи рамзи дунукта «:» фарқ мекунад, дар ҳоле ки дар Java ва C++ қавсҳои фигуравӣ «{ }» ва нуқта вергул «;» бартарӣ доранд. Забони HTML миқдори зиёди қавсҳои кунҷиро «< >» дар бар мегирад ва SQL бо басомади баланди калимаҳои калидии SELECT, FROM, WHERE тавсиф меёбад.

Аз нуқтаи назари омори математикӣ, коди барномавӣ метавонад ҳамчун пайдарпайии рамзҳо пешниҳод карда шавад:

$$S = \{c_1, c_2, c_3, \dots, c_n\}$$

ки дар ин ҷо c_i рамзи алоҳидаи коди барномавиро ифода мекунад.

Басомади пайдоиши рамзи мушаххас бо ифодаи зерин муайян карда мешавад:

$$f(c) = \frac{n_c}{N}$$

ки дар ин ҷо:

n_c - миқдори пайдоиши рамз;

N - миқдори умумии рамзҳо дар матни барнома.

Дар асоси таҳлили гузаронидашуда муайян карда шуд, ки тақсими ҳарфҳои лотинӣ ва рамзҳои махсус барои идентификатсияи забони барномасозӣ иттилоотияти баланд дорад. Масалан, дар забони Python ҳарфҳои «d», «e», «f» бинобар мавҷудияти калимаи калидии «def» хеле бештар истифода мешаванд. Дар забони Java басомади баланди ҳарфҳои «r», «u», «b», «l», «i», «c» аз сабаби истифодаи васеи модификатори дастрасии public мушоҳида мегардад.

Ҷадвали 1 – Басомади истифодаи рамзҳои алоҳида дар забонҳои гуногуни барномасозӣ ва технологияи сомонасозӣ

Символ	Python	Java	C++	HTML
;	0.2%	4.8%	5.1%	0.1%
{	0.0%	2.3%	2.5%	0.0%
:	3.9%	0.4%	0.3%	0.1%
<	0.2%	0.5%	0.6%	8.7%
>	0.2%	0.5%	0.6%	8.6%

Натиҷаҳои бадастомада нишон медиҳанд, ки ҳатто таҳлили оддии омории рамзҳо имкон медиҳад хусусиятҳои хоси забонҳои гуногуни барномасозӣ ҷудо карда шаванд.

Аммо барои баланд бардоштани дақиқии таснифот истифодаи моделҳои бисёрченакаи омори талаб карда мешавад.

Яке аз усулҳои паҳншудатарини таҳлили матн ин TF-IDF (Term Frequency - Inverse Document Frequency) мебошад. Усули мазкур имкон медиҳад аҳамияти рамзҳо, токенҳо ё калимаҳои калидӣ дар дохили коди барномавӣ муайян карда шавад.

Формулаи TF-IDF намуди зерин дорад:

$$TF-IDF(t, d) = TF(t, d) \cdot \log \frac{N}{DF(t)}$$

ки дар ин ҷо:

TF(t,d) - басомади термин;

DF(t) - миқдори ҳуҷҷатҳое, ки терминро дар бар мегиранд;

N - миқдори умумии ҳуҷҷатҳо.

Вобаста ба масъалаи муайянкунии забони барномасозӣ, ҳамчун терминҳои метавонанд калимаҳои калидӣ, рамзҳо, биграммаҳо, триграммаҳо ва сохторҳои синтаксисӣ баромад кунанд.

Таҳлили биграммаҳои ҳарфҳои латинӣ аҳамияти назаррас дорад. Масалан, пайдарпайҳои «in», «pr», «cl», «de», «fu» дар забонҳои гуногуни барномасозӣ бо эҳтимолияти гуногун дучор меоянд. Дар забони Python биграммаи «de» аксар вақт бо калимаи калидии def алоқаманд аст ва дар Java биграммаи «cl» барои сохтори class хос аст.

Барои ҳисоб кардани эҳтимолияти пайдоиши биграмма ифодаи зерин татбиқ мегардад:

$$P(x_i, x_{i+1}) = \frac{Count(x_i, x_{i+1})}{Count(x_i)}$$

Дар асоси таҳлили гузаронидашуда метавон тасдиқ намуд, ки моделҳои эҳтимолиятӣ имкон медиҳанд сохтори омории коди барномавӣ самаранок тавсиф карда шавад.

Истифодаи усулҳои омӯзиши машинӣ тавачҷӯҳи махсусро ҷалб менамояд. Дар масъалаи муайянкунии забони барномасозӣ алгоритмҳои зерин васеъ истифода мешаванд:

- Naive Bayes (Байеси сода);
- Support Vector Machine (Мошини векторҳои тақиявӣ);
- Random Forest (Ҷангали тасодуфӣ);
- k-Nearest Neighbors (k-Ҳамсояҳои наздиктарин);
- шабакаҳои нейронӣ;
- шабакаҳои нейронии рекуррентӣ;
- меъморҳои трансформаторӣ (Transformer).

Моделҳои эҳтимолиятӣ нисбатан сода ин таснифгари содаи байесӣ мебошад. Эҳтимолияти мансубияти коди барномавӣ ба забони L_i бо ифодаи зерин муайян карда мешавад:

$$P(L_i | X) = \frac{P(X | L_i)P(L_i)}{P(X)}$$

ки дар ин ҷо:

X - фрагменти таҳлилшавандаи код;

P(L_i) - эҳтимолияти априории забон;

P(X| L_i) - эҳтимолияти пайдоиши фрагмент дар забони L_i .

Афзалияти усули мазкур суръати баланди таснифот ва мураккабии пасти ҳисоббарорӣ мебошад. Аммо фарзия оид ба мустақилияти аломатҳо ҳангоми таҳлили сохторҳои мураккаби синтаксисӣ дақиқиро коҳиш медиҳад.

Барои баҳодиҳии монандии байни матнҳои барномавӣ cosine similarity (монандии косинусӣ) истифода мешавад:

$$\cos(\theta) = \frac{A \cdot B}{\|A\| \|B\|}$$

ки дар ин ҷо:

A ва B - векторҳои хусусиятҳои коди барномавӣ мебошанд.

Метрикаи мазкур ҳангоми муқоисаи профилҳои омории забонҳои барномасозӣ васеъ истифода мегардад. Чӣ қадаре ки арзиши cosine similarity ба воҳид наздик бошад, ҳамон қадар эҳтимолияти мансубият ба як забон баландтар аст.

Марҳилаи муҳими таҳқиқот ин ташаккули фазои аломатҳо мебошад. Ҳамчун аломатҳо метавонанд истифода шаванд:

- басомадҳои ҳарфҳои лотинӣ;
- басомадҳои рамзҳои махсус;
- калимаҳои калидӣ;
- дарозии сатрҳо;
- тақсимои идентификаторҳо;
- биграммаҳо ва триграммаҳо;
- қолабҳои синтаксисӣ;
- параметрҳои омории код.

Ба таври расмӣ вектори аломатҳо метавон ба тариқи зерин пешниҳод намуд:

$$X = (x_1, x_2, x_3, \dots, x_n)$$

ки дар ин ҷо ҳар як компоненти x_i ба хусусияти муайяни омории матни барномавӣ мувофиқат мекунад.

Барои баҳодиҳии фарқиятҳои байни забонҳои барномасозӣ масофаи Евклид татбиқ мегардад:

$$d(A, B) = \sqrt{\sum_{i=1}^n (A_i - B_i)^2}$$

Чӣ қадаре ки арзиши масофа хурд бошад, ҳамон қадар монандии сохтории кодҳои барномавӣ баландтар аст.

Таҳлили муқоисавӣ нишон медиҳад, ки забонҳои барномасозиро аз рӯи хусусиятҳои омории шартан ба якчанд гурӯҳ тақсим кардан мумкин аст:

Ҷадвали 2 – Хусусиятҳои сохтории забонҳои барномасозӣ ва технологияҳо

Номгӯӣ	Нишондиҳандаҳо
Python	Фосилаҳо аз канор, дунуктаҳо, набудани «;»
Java	Истифодаи зуд-зуди public, class, static
C++	Нишондиҳандаҳо, #include, қавсҳои фигуравӣ
JavaScript	function, console.log, =>
HTML	Тегҳо, қавсҳои кунҷӣ
SQL	SELECT, WHERE, JOIN

Аз ин бармеояд, ки профили омории забони барномасозӣ дорои беназирии баланд мебошад.

Таҳқиқотҳои муосир нишон медиҳанд, ки усулҳои омӯзиши амиқ дақиқии бамаротиб баландтари таснифотро таъмин мекунанд. Шабакаҳои нейронӣ қодиранд ба таври автоматӣ қонуниятҳои пинҳонии тақсимои рамзҳо ва сохторҳои синтаксисиро ошкор намоянд.

Шабакаҳои нейронии рекуррентӣ бо пайдарпайҳои рамзҳо самаранок кор мекунанд:

$$h_t = f(Wx_t + Uh_{t-1} + b)$$

ки дар ин ҷо:

h_t - ҳолати пинҳонӣ;

x_t - рамзи воридотӣ;

W, U - матритсаҳои вазнҳо.

Моделҳои трансформаторӣ имкон медиҳанд контексти тамоми матни барномавӣ дар як вақт ба назар гирифта шавад. Ин моделҳо махсусан ҳангоми таҳлили фрагментҳои калони код ва файлҳои барномавии омехта самаранок мебошанд.

Дар ҷараёни таҳқиқоти таҷрибавӣ фрагментҳои коди барномавӣ бо забонҳои Python, Java, C++, JavaScript, PHP ва C# таҳлил карда шуданд. Ҳаҷми интиҳоби таҷрибавӣ зиёда аз 50000 файли коди ибтидоиро ташкил дод. Таҳлил бо истифодаи усулҳои таснифоти омӯрӣ ва моделсозии шабакаҳои нейронӣ гузаронида шуд. Натиҷаҳои бадастомада дар ҷадвали 3 пешниҳод шудаанд.

Ҷадвали 3 – Дақиқии муайянкунии забони барномасозӣ

Метод	Саҳеҳнокӣ
Таҳлили басомадӣ	78.4%
TF-IDF + SVM	91.7%
Naive Bayes	88.3%
Random Forest	92.1%
Шабакаҳои нейронӣ	96.4%
Transformer	98.1%

Дар асоси таҳлили гузаронидашуда муайян карда шуд, ки истифодаи зеҳни сунӣ самаранокӣ муайянкунии забони барномасозиро ба таври назаррас баланд мебардорад. Махсусан меъморӣҳои трансформаторӣ, ки қодиранд вобастагӣҳои контекстии байни рамзҳо ва сохторҳоро ба назар гиранд, натиҷаҳои баланд нишон медиҳанд.

Бо вучуди ин, равиши баррасишаванда як қатор маҳдудиятҳо дорад. Якум, фрагментҳои кӯтоҳи код миқдори нокифояи иттилооти омӯриро дар бар мегиранд. Дуюм, файлҳои омехта метавонанд унсурҳои якчанд забони барномасозиро дар як вақт дар бар гиранд. Масалан, саҳифаҳои веб аксар вақт HTML, CSS ва JavaScript-ро дар як ҳуҷчат дар бар мегиранд. Сеюм, баъзе забонҳо дорои дараҷаи баланди монандии синтаксисӣ мебошанд.

Мураккабии иловагиро идентификаторҳои истифодабаранда ба миён меоранд. Масалан:

C++

`int abc = 10;`

ва

Java

`int value = 10;`

дорои сохтори омӯрии қариб якхела мебошанд. Бинобар ин, барои баланд бардоштани дақиқии таснифот таҳлили қонуниятҳои амиқтари синтаксисӣ талаб карда мешавад.

Масъалаи муайянкунии забони барномасозӣ дар соҳаи амнияти иттилоотӣ аҳамияти махсус пайдо мекунад. Системаҳои таҳлили автоматии таъминоти барномавии зараровар бояд забони барномасозии истифодашударо барои таҳлили минбаъдаи таҳдидҳо фавран муайян намоянд. Ба ҳамин монанд, системаҳои ошкорсозии плагиат усулҳои омори барои муайян кардани манбаи коди ибтидоӣ истифода мебаранд.

Таъбиқи амалии таҳқиқот инҳоро дар бар мегирад:

- муҳитҳои таҳияи зеҳнӣ (IDE);
- системаҳои компилятсияи автоматӣ;
- криминалистикаи рақамӣ;
- таҳлили репозиторийҳои GitHub;
- системаҳои ошкорсозии таъминоти барномавии зараровар;
- платформаҳои таҳлили додаҳои калон;
- системаҳои таълимии санҷиши коди барномавӣ.

Бояд қайд намуд, ки таҳлили омории ҳарфҳои лотинӣ на танҳо барои муайянкунии забони барномасозӣ, балки барои идентификатсияи муаллифи код низ метавонад истифода шавад. Ҳар як таҳиягар дорои услуби инфиродии барномасозӣ мебошад, ки дар интиҳоби идентификаторҳо, форматкунӣ ва сохторҳои синтаксисӣ зоҳир мегардад. Ҳамин тариқ, усулҳои атрибутсияи муаллифӣ бо масъалаи муайянкунии забони барномасозӣ зич алоқаманданд.

Дар доираи назарияи шинохти намунаҳо, коди барномавӣ ҳамчун объекти бисёрченака баррасӣ мегардад, ки дорои маҷмӯи хусусиятҳои омории мебошад. Раванди таснифот марҳилаҳои зеринро дар бар мегирад:

- коркарди пешакӣ;
- токенизатсия;
- истихроҷи хусусиятҳо (аломатҳо);
- сохтани векторҳо;
- омӯзиши модел;
- санҷиши таснифгар.

Коркарди пешакии коди барномавӣ тозакунии шарҳҳо, ба эътидол овардани фосилаҳо, ба формати ягона овардани матн ва ҷудокунӣ токенҳоро дар бар мегирад. Марҳилаи мазкур ба дақиқии таҳлили минбаъда таъсири назаррас мерасонад.

Барои баҳодихии сифати таснифот метрикаҳои зерин истифода мешаванд:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

ки дар ин ҷо:

TP - миқдори натиҷаҳои ҳақиқӣ-мусбат;

FP - натиҷаҳои бардурӯғ-мусбат;

FN - натиҷаҳои бардурӯғ-манфӣ.

Натиҷаҳои бадастомада нишон медиҳанд, ки истифодаи усулҳои омехта дақиқии баландтарини таснифотро таъмин мекунад. Нисбатан самаранок моделҳои гибридӣ мебошанд, ки таҳлили омории, TF-IDF ва алгоритмҳои шабакаҳои нейрониро муттаҳид месозанд.

Таҳлили тақсимои ҳарфҳои лотинӣ мавҷудияти қонуниятҳои устувори омории нишон медиҳад. Масалан, ҳарфҳои «i», «n», «t» бинобар истифодаи намуни int аксар вақт дар забонҳои оилаи C-монанд дучор меоянд. Дар забони Python басомади ҳарфҳои «d», «e», «f» ва «g» ба таври назаррас меафзояд. Дар HTML пайдарпайҳои «ht», «ml», «di», ки бо сохтори теги ҳуҷҷат алоқаманданд, бартарӣ доранд.

Аз ин бармеояд, ки тақсимои рамзҳои лотинӣ метавонад ҳамчун як навъ изи оморӣ забони барномасозӣ баррасӣ карда шавад.

Хулоса

Дар асоси таҳқиқоти гузаронидашуда метавон хулоса баровард, ки самаранокии усулҳои таҳлили оморӣ ҳарфҳои лотинӣ барои муайянкунии забони коди барномавӣ хеле баланд аст. Истифодаи моделҳои математикӣ, усулҳои омӯзиши мошинӣ ва технологияҳои зехни сунъӣ имкон медиҳад, ки ҳатто ҳангоми таҳлили фрагментҳои нисбатан хурди код дақиқии баланди таснифот ба даст оварда шавад.

Бо дарназардошти гуфтаҳои боло, коди барномавӣ сохтори мураккаби омориро муаррифӣ мекунад, ки дорои миқдори зиёди қонуниятҳои барои таҳлили автоматиконидашуда коршоям мебошад. Алифҳои лотинӣ ҳамчун асоси аксарияти забонҳои барномасозӣ, тақсимои беназири рамзҳои ташаккул медиҳад, ки онҳоро ҳамчун аломатҳои таснифоти истифода бурдан мумкин аст.

Бинобар ин, татбиқи усулҳои зехни сунъӣ, таҳлили оморӣ ва назарияи шинохти намунаҳо барои эҷоди системаҳои зехнии муайянкунии автоматии забони коди барномавӣ дурнамои васеъро мекушояд. Натиҷаҳои бадастомада имконияти истифодаи самараноки қонуниятҳои оморӣ ҳарфҳои лотиниро дар масъалаҳои таҳлили коди ибтидоӣ, криминалистикаи рақамӣ ва коркарди зехнии матнҳои барномавӣ тасдиқ мекунад.

Дар натиҷаи таҳқиқоти гузаронидашуда муайян карда шуд, ки самаранокии аз ҳама баландро моделҳои омехтае доро мебошанд, ки таҳлили оморӣ рамзҳо, усулҳои омӯзиши мошинӣ ва меъморӣҳои шабакаҳои нейрониро муттаҳид мекунад. Чунин равиш устувории баландро ба ҳалалҳо, фрагментҳои ихтисоршудаи код ва сохторҳои омехтаи забонӣ таъмин менамояд. Натиҷаҳои илмӣ бадастомада метавонанд ҳамчун асос барои таҳқиқотҳои минбаъда дар соҳаи идентификатсияи автоматии забонҳои барномасозӣ ва атрибутсияи муаллифии коди барномавӣ хизмат кунанд.

Муқарриз: Табуров М.Х. – н.и.т., дотсенти қабедраи автоматикунони равандҳои технологӣ ва истеҳсолоти ДПТ ба номи академик М.С. Осимӣ.

Адабиёт

1. Rudman, J. The state of authorship attribution studies: Some problems and solutions [Text] / J. Rudman // Computers and the Humanities. – 1998. – Vol. 31. – pp. 351-365.
2. Куртукова, А.В. Идентификация автора исходного кода методами машинного обучения [Текст] / А.В. Куртукова, А.С. Романов // Журнал «Труды СПИИРАН». – СПб. – 2019. – №3. – С. 742-766.
3. Куртукова, А.В. Архитектура интеллектуальной системы идентификации автора исходного кода [Текст] / А.В. Куртукова, А.С. Романов, А.А. Шелупанов // Журнал «Доклады ТУСУР». – Томск. – 2022. – №3. – С. 55-63.
4. Куртукова, А.В. Интеллектуальная система идентификации автора программного кода [Текст] / А.В. Куртукова // Журнал «Молодой ученый». – Казань. – 2020. – №40. – С. 2-5.
5. Мещеряков, Р.В. Методы анализа исходного кода программ в задачах информационной безопасности [Текст] / Р.В. Мещеряков, А.А. Шелупанов // Журнал «Приборостроение». – СПб. – 2017. – №7. – С. 121-126.
6. Шелупанов, А.А. Методы защиты информации в программных системах [Текст] / А.А. Шелупанов // Журнал «Информационные технологии». – М. – 2016. – №5. – С. 14-20.
7. Белов, С.Н. Методы обнаружения клонов программного кода [Текст] / С.Н. Белов // Журнал «Программные продукты и системы». – Тверь. – 2016. – №3. – С. 57-63.
8. Власов, А.В. Методы выявления заимствований в программных текстах [Текст] / А.В. Власов, Н.В. Малышев // Журнал «Информационные технологии». – М. – 2014. – №9. – С. 33-38.
9. Афонин, А.Н. Анализ сходства программного кода [Текст] / А.Н. Афонин, В.П. Пономарев // Журнал «Информатика и образование». – М. – 2013. – №8. – С. 44-49.
10. Соловьев, А.А. Анализ программных клонов и обнаружение заимствований [Текст] / А.А. Соловьев // Журнал «Программирование». – М. – 2017. – №4. – С. 213-220.
11. Лебедев, В.С. Алгоритмы обнаружения сходства программного кода [Текст] / В.С. Лебедев // Журнал «Системы управления и информационные технологии». – Воронеж. – 2010. – №2. – С. 78-83.
12. Кравцов, А.В. Методы стилометрического анализа текстов программ [Текст] / А.В. Кравцов // Журнал «Вестник СПбГУ». – СПб. – 2018. – №4. – С. 88-96.

13. Иванов, А.Н. Применение машинного обучения для анализа программного кода [Текст] / А.Н. Иванов, П.В. Сидоров // Журнал «Информационные технологии». – М. – 2020. – №10. – С. 45-52.

14. Асадуллозода, Ш.М. Таҳлили монандии код: моҳият, усулҳо ва аҳамияти он дар замони рақамисозӣ / Ш.М. Асадуллозода, Э.К. Абдуллозода // Фурӯғи илм. – 2025. – № 4(04). – С. 26–30.

15. Қосимов, А. А. Муқоисаи асарҳои номгузори назарӣ дар асоси симони рақамии униграмаи рамзӣ ва ҳиҷо / А. А. Қосимов, Ҳ. А. Астанакулов, Ш. А. Бозоров // Паёми политехникӣ. Баҳши: Интеллект, Инноватсия, инвеститсия. – 2024. – №. 4(68). – Р. 45-49. – EDN KYNDQR.

МАЪЛУМОТ ДАР БОРАИ МУАЛЛИФОН - СВЕДЕНИЯ ОБ АВТОРАХ INFORMATION ABOUT THE AUTHORS

TJ	RU	EN
Қосимов Абдунаби Абдурауфович	Косимов Абдунаби Абдурауфович	Kosimov Abdunabi Abduraufovich
Доктори илмҳои техникаӣ, дотсент	Доктор технических наук, доцент	Doctor of Technical Sciences, associate professor
Донишгоҳи техникаи Тоҷикистон ба номи академик М.С. Осимӣ	Таджикский технический университет имени академика М.С. Осими	Tajik technical university named after academician M.S. Osimi
E-mail: abdunabi_kbtut@mail.ru		
TJ	RU	EN
Асадуллозода Шаҳриёр Мирзохон	Асадуллозода Шаҳриёр Мирзохон	Asadullozoda Shahriyor Mirzokhon
Унвонҷӯ	Соискатель	Applicant
Донишгоҳи Давлатии Бохтар ба номи Н. Хусрав	Бохтарский государственный университет имени Н. Хусрава	Bokhtar State University named after N. Khusrav
E-mail: shahriyorkhairarov480@gmail.com		
TJ	RU	EN
Шарифов Мусо Умарович	Шарифов Мусо Умарович	Sharifov Muso Umarovich
Магистранти курси 2	Магистрант 2 курса	2st year Master's student
Институти математика ба номи А. Ҷураевӣ АМИ Тоҷикистон	Институт математики им. А. Джураева НАН Таджикистана	A. Juraev Institute of Mathematics of NAS of Tajikistan
E-mail: sharifovmuso2000@gamil.com		