המלצות להוספת פיצ'רים חדשים לזיהוי אתרי פישינג

המסמך כולל הצעות לפיצ'רים חדשים שיכולים לשפר את יכולת המודל לזהות כתובות URL חשודות. הפיצ'רים מבוססים על ידע בתחום הסייבר, ניתוח טקסטים וסטטיסטיקה.

|  |  |  |
| --- | --- | --- |
| שם הפיצ'ר | מה הפיצ'ר בודק | למה זה חשוב |
| is\_encoded\_url | בודק אם ה-URL כולל תווים מקודדים כמו %20, %3A | קידוד URL יכול להעיד על ניסיון להסתיר מידע ולהתחמק מזיהוי. |
| domain\_entropy | מודד את רמת האנטרופיה (אקראיות) בשם הדומיין | דומיינים אקראיים הם נפוצים בפישינג כדי למנוע זיהוי מבוסס blacklist. |
| is\_top\_domain\_legit | בודק אם הדומיין נמצא ברשימת דומיינים פופולריים (למשל Alexa) | רוב האתרים הלגיטימיים מופיעים ברשימות מוכרות. |
| count\_suspicious\_words | סופר מילים כמו login, verify, bank, secure בכתובת | פישינג משתמש לעיתים קרובות במילים מפתות להזדהות. |
| distance\_to\_known\_brand | מודד את המרחק לעריכה (edit distance) מול רשימת מותגים | דומיינים כמו paypall.com או g00gle.com הם התחזות. |
| has\_ip\_and\_port | בודק האם ה-URL כולל גם כתובת IP וגם פורט | שימוש כזה נדיר באתר אמיתי ונפוץ בפישינג. |
| num\_repeated\_segments | בודק כמה חלקים בנתיב ה-URL חוזרים על עצמם | חזרתיות לא רגילה מעידה על כתיבה אוטומטית או התחזות. |
| digit\_ratio\_in\_hostname | בודק את יחס הספרות בדומיין | ריבוי ספרות בדומיין לא טבעי באתרים לגיטימיים. |
| long\_path\_depth | מודד את עומק הנתיב לפי מספר הסלשים (/) | URL עמוק מדי עשוי להיות התחזות לאתר רשמי. |
| contains\_login\_form\_and\_no\_https | בודק אם יש טופס התחברות באתר בלי HTTPS | שילוב מסוכן שמעיד על פישינג ברור. |