**Electronic & Cyber Defense**
Vol. 13, No. 4, 2026 (Serial No. 52). pp. 63- 76
ISSN: 2322-4347, E-ISSN: 2980-8979

# Dark Web Text Classification using BERT's Language Model

**Baratali Akhtariyan[1], Mohsen Rezvani[*2]**
[1] Master's student, Shahrood University of Technology, Shahrood, Iran.
*barat.akhtariean@gmail.com*
[2] Assistant Professor, Shahrood University of Technology, Shahrood, Iran
(Correspondence: *mrezvani@shahroodut.ac.ir*)

ARTICLE INFO

**Article history:**

Article Type: Research paper

**Receive:** 19 September 2025
**Revise:** 08 November 2025
**Accept:** 28 November 2025
**Available online:** 22 December 2025

**Keywords:**
Dark Web
Large Language Models
Transformers
BERT

ABSTRACT

The hidden nature and limited access of the dark web has led to the proliferation of many criminal activities, including cyber threats, arms sales, drug sales, and the sale of illegal tools. The emergence of large language models has created the hope that it will be possible to analyze the content on the dark web with proper accuracy. In this regard, the use of mass cyber data available in the dark web will be very useful and effective to prevent cyber threats and train language models. The technology of large language models requires a lot of high-quality data for better training and to achieve sufficient accuracy, and this is the challenge that researchers in the field of cyber security face due to the contamination of the data available on the dark web. Most of the researches in this field have been focused on all the characteristics of the dark web dataset and low-quality data and have not been able to achieve high accuracy. In this thesis, we presented a new language model based on the BERT-based language model, which was trained on the data extracted from the dark web. The proposed model is a transformer-based text model that uses a two-way encoder of transformers for a learning approach and we evaluated it on a high - quality dataset, without repetitive data, free of unknown words, all in English and specifically on hacking and security data. Finally, by analyzing the evaluated values of the proposed model with the previous models, it was found that the proposed model was able to have better accuracy in data classification due to the injection of quality data compared to the previous models.

**Cite this article:** Akhtariyan, B., Rezvani, M. Dark web text classification using BERT's Language Model. Electronic and Cyber Defense, 2025; 13(4): 63-76.

DOI: https://doi.org/10.47176/ECDJ.2025.1624



Publisher: Imam Hossein University

«پدافند الکترونیکی و سایبری»
سال سیزدهم، شماره ۴، زمستان ۱۴۰۴، ص ۶۳- ۷۶
شاپا الکترونیکی: ۴۳۴۷ -۲۳۲۲ شاپا چاپی: ۸۹۷۹- ۲۹۸۰

# دسته‌بندی داده‌های وب تاریک به کمک مدل زبانی BERT

**براتعلی اختریان**[1]، **محسن رضوانی**[2*]

[1] دانشجوی کارشناسی‌ارشد،دانشگاه صنعتی شاهرود ، شاهرود، ایران. *barat.akhtariean@gmail.com*

[2] استادیار، دانشگاه صنعتی شاهرود، شاهرود، ایران (نویسنده مسئول: *mrezvani@shahroodut.ac.ir*)

**چکیده**

ماهیت پنهان و دسترسی محدود وب‌تاریک، موجب گسترش فعالیت‌های مجرمانه بسیاری ازجمله تهدیدات سایبری، فروش اسلحه، فروش مواد مخدر و فروش ابزارهای غیرقانونی شده است. ظهور مدل‌های زبانی بزرگ این امید را ایجاد نموده است که بتوان با دقت مناسبی به تحلیل مطالب موجود در وب تاریک پرداخت. در همین راستا استفاده از داده‌های انبوه سایبری موجود در وب‌تاریک برای جلوگیری از تهدیدات سایبری و آموزش مدل‌های زبانی بسیار مفید و مؤثر خواهد بود. فناوری مدل‌های زبانی بزرگ برای آموزش بهتر و رسیدن به‌دقت کافی، به داده زیاد و باکیفیت بالا نیاز دارند و این چالشی است که محققان حوزه امنیت سایبری با توجه به آلوده بودن داده‌های موجود در وب‌تاریک روبرو هستند. اغلب تحقیقات در این زمینه، متمرکز بر روی تمام مشخصه‌های دادگان وب‌تاریک و داده‌های باکیفیت پایین صورت پذیرفته است و نتوانسته‌اند دقت بالایی را کسب کنند. در این پژوهش یک مدل زبانی جدید بر پایه مدل زبانی پایه BERT که بر روی‌داده استخراج‌شده از وب‌تاریک آموزش‌دیده است، ارائه کردیم. مدل پیشنهادی یک مدل متنی مبتنی بر ترانسفورماتور است که از رمزگذار دوطرفه از ترانسفورماتورها برای رویکرد یادگیری استفاده می‌کند و آن را بر روی یک دادگان باکیفیت بالا، بدون داده تکراری، عاری از کلمات نامعلوم، تماماً به زبان انگلیسی و به‌طور مشخص بر روی‌داده‌های سایبری و امنیت ارزیابی نمودیم. درنهایت با تحلیل مقادیر ارزیابی‌شده مدل پیشنهادی با مدل‌های قبلی، مشخص شد که مدل پیشنهادی به علت تزریق داده‌های باکیفیت نسبت به مدل‌های قبلی، توانسته دقت بهتری در دسته‌بندی داده‌ها داشته باشد.

**مشخصات مقاله**

**تاریخچه مقاله:**

**نوع مقاله: علمی- پژوهشی**

**دریافت**: 1404/06/28

**بازنگری**: 1404/08/17

**پذیرش**: 1404/09/07

**ارائه آنلاین**: 1404/10/01

**کلیدواژه‌ها:**

وب‌تاریک

مدل‌های زبانی بزرگ

ترانسفورماتور

BERT

**استناد**: اختریان, براتعلی, رضوانی, محسن. دسته بندی داده های وب تاریک به کمک مدل زبانی BERT . پدافند الکترونیکی و سایبری، ۱۴۰۴. ۱۳(۴)، ص ۶۳- ۷۶. https://doi.org/10.47176/ECDJ.2025.1624



**ناشر**: دانشگاه جامع امام حسین(ع).

## 1- مقدمه

با گسترش جوامع بشری و نیاز به سهولت ارتباطات در امور روزمره، انسان‌ها با بهره‌گیری از فناوری برای رفاه وزندگی بهتر به استفاده فزاینده از اینترنت و شبکه‌های رایانه‌ای روی آورده‌اند. اما می‌توان گفت بزرگ‌ترین چالش پیش رو با گسترش شبکه‌های رایانه‌ای[1]، تأمین امنیت داده موجود در این بستر است. کاربران بسیاری ازجمله نفوذ گران به دنبال انجام فعالیت‌های مخرب در فضای اینترنت و شبکه‌های رایانه‌ای هستند و در سوی دیگر محققان امنیت شبکه راه‌های گوناگونی را برای افزایش امنیت شبکه‌های رایانه‌ای مورد ارزیابی قرار می‌دهند.

روش‌های مختلفی برای افزایش امنیت شبکه وجود دارد و این در حالی است که با تنوع و کمیت حملات سایبری[2]، رویکردهای سنتی ایمن نگه‌داشتن شبکه‌های رایانه‌ای دیگر کافی نیست.[1] طی سالیان مختلف روش‌هایی مانند استفاده از ضد بدافزارها[3] و به‌روزرسانی مداومشان و یا استفاده از رمزهای پیچیده در دستگاه‌های رایانه همچنین به‌کارگیری دیواره آتش‌های[4] سخت‌افزاری و نرم‌افزاری برای بالابردن امنیت شبکه‌های رایانه‌ای ایجاد شد اما به‌کارگیری آن‌ها باوجود خطاهای انسانی متفاوت و سرعت سریع تغییرات فناوری، کارایی و پویایی بالایی ندارد. امروزه با پیشرفت فناوری و ظهور و گسترش مدل‌های زبانی بزرگ[5] این امید ایجادشده که بتوان با کمک این مدل‌ها با دقت و کارایی بیشتری در جهت امنیت شبکه‌های رایانه‌ای گام برداشت.[2]

مدل‌های زبانی بزرگ به دلیل عملکرد بی‌سابقه‌شان در اکثر زمینه‌ها، محبوبیت فزاینده‌ای پیداکرده‌اند. این مدل‌ها مولد توزیع آماری نشانه‌ها با دادگان[6] عظیم متنی، تصاویر، ویدیو و دیگر داده‌های تولیدشده توسط انسان هستند. مدل‌های زبانی بزرگ برای فعالیت خود به دادگان بسیار زیادی نیاز دارند و تهیه داده در جهت اهداف پژوهش امری مهم و حیاتی محسوب می‌شود.[3][4] در پژوهش‌های مختلف برای تهیه دادگان مرتبط با امنیت سایبری[7] از داده موجود در وب سطحی[8] استفاده کرده‌اند و این در حالی است که بستر فعالیت‌های مجرمانه بسیاری ازجمله حملات سایبری، وب‌تاریک[9] است؛ هرچند دسترسی به وب‌تاریک کار آسانی نیست.[5] بهره‌برداری مداوم از وب‌تاریک با تفاوت‌های واضحی که در ادبیات و زبان موجود در وب سطحی و وب‌تاریک وجود دارد به‌عنوان یک پلتفرم جرائم سایبری، آن را به حوزه‌ای ارزشمند و ضروری برای تحقیقات امنیت سایبری تبدیل کرده است.[6][7]

در این مقاله یک مدل زبانی جدید بر پایه مدل زبانی BERT[10] با تمرکز بر دادگان کیفیت بالا، جمع‌آوری‌شده از وب‌تاریک را پیشنهاد می‌کنیم. دادگان تهیه‌شده به زبان انگلیسی، بدون هرگونه کاراکتر نامشخص و عاری از متون تکراری است. این مدل قادر است تا دادگان جمع‌آوری‌شده را با دقت بالا دسته‌بندی کند. دسته‌بندی متن[11] به مدل زبانی این اجازه را می‌دهد تا معانی و روابط ظریف درون متون را به‌طور مؤثر دریافت کند؛ بر همین اساس هدف اصلی دسته‌بندی متن با تمرکز بر دادگان وب‌تاریک در حوزه سایبری، استخراج متن تولیدشده توسط مدل زبانی درباره امنیت سایبری است که از این طریق می‌توان مدل پیشنهادی را به ابزاری مفید برای فعالیت‌های پردازش زبان طبیعی مانند پاسخگویی به سؤال[12] تبدیل کرد. جهت ارزیابی بهتر مدل، دسته‌بندی متن توسط مدل در دو حالت متن باکیفیت و متن خام[13] صورت گرفت. نتایج ارزیابی ما نشان می‌دهد که مدل طبقه‌بندی متن پیشنهادی با دادگان باکیفیت توانسته دقت بالایی را در دسته‌بندی متن از خود نشان دهد.

کارایی مدل پیشنهادی به استفاده از دادگان به‌روز و باکیفیت وابسته است. روش‌های به کار گرفته‌شده درگذشته از دادگان باکیفیت امنیت سایبری استخراج‌شده از وب‌تاریک بهره نبرده‌اند. در مدل پیشنهادی هرچقدر داده باکیفیت به مدل تزریق شود، مدل دقت بالاتری را در دسته‌بندی متن به خود اختصاص خواهد داد. همچنین دادگان به‌روز حوزه امنیت سایبری با توجه به دامنه و گستردگی تغییرات مختلف کلمات برحسب زمان در فضای وب‌تاریک موجب خواهد شد تا مدل پویایی خود را حفظ کند.

در ادامه این مقاله و در بخش دوم ادبیات موضوع و کارهای انجام‌شده ارائه می‌شود. راهکار پیشنهادی برای موضوع پژوهش در بخش سوم شرح داده خواهد شد و سپس در بخش چهارم با بررسی حالت‌های مختلف مدل زبانی و داده تزریق‌شده گوناگون به مدل پیشنهادی، نتایج حاصل‌شده را مورد ارزیابی قرار خواهیم داد و درنهایت در بخش پنجم نتیجه‌گیری مقاله ارائه می‌شود.

## 2- ادبیات موضوع و کارهای انجام‌شده

### 2-1- وب‌تاریک

لایه‌های اینترنت بسیار فراتر از محتوای سطحی است که بسیاری می‌توانند به‌راحتی در جستجوهای روزانه خود به آن دسترسی داشته باشند. محتوای مربوط به وب عمیق[14]، محتوایی است که توسط موتورهای جستجوی سنتی مانند گوگل شناسایی نشده است. دورترین گوشه‌های وب عمیق، بخش‌هایی که به وب‌تاریک

[1] Computer Networks
[2] Cyber Attack
[3] Antivirus
[4] Firewall
[5] Large Language Models
[6] Dataset
[7] Cyber Security
[8] Surface Web
[9] Dark Web
[10] Birirectional encoder Representation from Transformers
[11] Text Classification
[12] Question Answering
[13] Raw
[14] Deep Web

معروف هستند، حاوی محتوایی هستند که عمداً پنهان‌شده است. وب‌تاریک ممکن است برای اهداف غیرقانونی و همچنین برای پنهان کردن فعالیت‌های مجرمانه یا مخرب دیگر مورداستفاده قرار گیرد. این بهره‌برداری از وب‌تاریک برای اقدامات غیرقانونی است که توجه مقامات و سیاست‌گذاران را به خود جلب کرده است [8]. همان‌طور که گفته شد، مجموعه اینترنت دارای لایه‌های مختلف است. با توجه به شکل (1) اگر اینترنت را مانند یک کوه یخی در نظر بگیریم، وب سطحی قله این کوه است؛ اما آنچه مشخص است بخش بزرگی از این کوه یخی در عمق آب قرار دارد.


**شکل (1)**: لایه‌های مختلف اینترنت جهانی

برای دسترسی به وب‌تاریک که یکی از لایه‌های اینترنت است، نیاز به نرم‌افزار خاصی به نام تور[1] داریم. تور یک شبکه رمزگذاری شده است که امکان دسترسی ناشناس به اینترنت را برای کاربران خود فراهم می‌کند و همچنین میزبان خدمات مخفی است. این سرویس‌های پنهان برای انجام فعالیت‌هایی استفاده می‌شوند که غیرقانونی و غیراخلاقی در سطح وب هستند.

این فعالیت‌ها شامل توزیع دسترسی به مواد مخدر، فروش سلاح، اطلاعات سایبری و امنیت و موارد دیگر است. آنچه مشخص است، بیشتر اطلاعات موجود در وب‌تاریک به زبان انگلیسی منتشر می‌شود و زبان روسی، زبان دوم در رده‌بندی زبانی وب‌تاریک قرار دارد [9] [10].

**جدول (1)**: رده‌بندی زبانی دادگان **CoDA**

| Language | Count | Language | Count |
|---|---|---|---|
| English | 8855 | Portuguese | 54 |
| Russian | 542 | Chinese | 38 |
| German | 129 | Italian | 28 |
| French | 100 | Japanese | 27 |
| Spanish | 61 | Dutch | 14 |

رده‌بندی زبانی دادگان که در مرجع [11] آمده را می‌توان در جدول (1) مشاهده کرد. نظر به فراوانی داده موجود به زبان انگلیسی در وب‌تاریک و توجه آموزش مدل زبانی BERT به داده انگلیسی، در این پژوهش نیز تمرکز برداده زبان انگلیسی قرار گرفت. همچنین با توجه به نیاز این پژوهش بر داده‌های سایبری، یکی از این منابع داده‌ای درباره اطلاعات تهدیدات سایبری که توجه زیادی را به خود جلب می‌کند، همین وب‌تاریک است. وب‌تاریک شامل صدها بازار آنلاین و پلتفرم‌های رسانه‌های اجتماعی است که در آن میلیون‌ها نفوذ گر در سراسر جهان به تجارت و فروش مقادیر قابل‌توجهی ابزارهای نفوذ مخرب، محتوا، دانش و سایر دارایی‌های سایبری در انجمن‌های نفوذ گرها[2] و بازارهای فروش آن می‌پردازند. گستردگی دانش و ابزارهای مخرب سایبری وب‌تاریک، نفوذ گرها را قادر به اجرای حملات سایبری در مقیاس بزرگ کرده است [12]. آنچه در جدول (2) نمایش می‌دهیم، تعدادی از داده‌های نفوذ وب‌تاریک است.

**جدول (2)**: نمونه داده‌های نفوذ در وب‌تاریک

| | | | |
|---|---|---|---|
| 1 | Hacking | en | FBI Hacking AND FORENSIC TOOLKIT HACK INTO PHONES |
| 2 | Hacking | en | Hack iCloud and Access any Phone Bypass Passcodes |
| 3 | Hacking | en | Software to Hack Spotify accounts - free music ! |
| 4 | Hacking | en | Link Searcher - hack and get any database ! |
| 5 | Hacking | en | Software to Hack eShops accounts - get free items |

درحالی‌که ممکن است اطلاعات زیادی در وب سطحی وجود داشته باشد، هنوز اطلاعات زیادی وجود دارد که در وب‌تاریک است و دسترسی به آن کار آسانی نیست، پس بنابراین می‌توان این اطلاعات را به جهت دسترسی محدود به وب‌تاریک، ازدست‌رفته دانست.

## 2-2- تفکیک‌کننده[3]

ازآنجایی‌که مدل‌های زبانی بزرگ برای استفاده آسان، ارزان، سریع و قابل‌استفاده در طیف گسترده‌ای از وظایف تحلیل متن[4]، از دسته‌بندی متن گرفته تا تجزیه‌وتحلیل احساسات هستند، بسیاری از محققان بر این باورند که مدل‌های زبانی بزرگ نحوه انجام تحلیل متن را تغییر خواهند داد [13]. آموزش مدل‌های زبانی معمولاً یک فرایند یکپارچه نیست. مدل‌های زبانی اغلب از یک تفکیک‌کننده استفاده می‌کنند که دنباله‌ای از کاراکترها را به دنباله‌ای از شناسه‌های رمزی، رمزگذاری می‌کند. کار مدل‌سازی زبان، فرایند بعدی است که توسط یک شبکه عصبی با ترانسفورمر[5] انجام می‌شود که از قبل آموزش داده‌شده و روی دادگان بزرگ تنظیم‌شده است. هدف ایدئال آموزش مدل زبانی، فعالیت مشترک تفکیک‌کننده و ترانسفورماتور برای بالابردن دقت[6] مدل است.

[1] TOR

[2] Hackers Froums
[3] Tokenizer
[4] Text Analysis
[5] Transformers
[6] Accuracy

این‌یک مشکل چالش‌برانگیز برای تنظیم مؤثر مقدار بهینه است و بنابراین، تفکیک‌کننده به‌طورکلی بر روی بخشی از دادگان آموزشی تطبیق داده می‌شود و قبل از آموزش مدل زبانی، مقادیر آن تثبیت می‌شود و سپس روی مجموعه بزرگ‌تر قرار می‌گیرد [14]. ازآنجایی‌که دیجیتالی شدن به این معنی است که بسیاری از ارتباطات انسانی در حال حاضر به‌عنوان داده‌های دیجیتال ذخیره و پردازش می‌شوند، اهمیت تجزیه‌وتحلیل متن در سال‌های اخیر افزایش‌یافته است. تفکیک‌کننده وظیفه دیجیتالی نمودن داده‌ها را برای آموزش مدل زبانی بر عهده دارد.

## 2-3- مدل‌های زبانی بزرگ

مدل‌های زبانی بزرگ به دلیل عملکرد بی‌سابقه‌شان در کاربردهای مختلف، محبوبیت فزاینده‌ای در دانشگاه و صنعت پیداکرده‌اند. تعامل با یک عامل مکالمه مبتنی بر مدل‌های زبانی بزرگ، می‌تواند توهم بودن در حضور یک موجود متفکر را ایجاد کند. بااین‌حال، در ماهیتشان، چنین سیستم‌هایی اساساً شبیه ما نیستند. برخی از این مدل‌های زبانی، بسته به نوع کاربرد و ویژگی‌هایشان، متفاوت عمل می‌کنند [15]. مدل‌های زبانی بزرگ، مدل‌های ریاضی مولد توزیع آماری نشانه‌ها در مجموعه عمومی گسترده متن تولیدشده توسط انسان هستند که در آن نشانه‌های موردنظر شامل کلمات، بخش‌هایی از کلمات یا کاراکترهای فردی، ازجمله علائم نشانه‌گذاری هستند. ازجمله این مدل‌های زبانی، می‌توان به BERT اشاره کرد [16]. مدل زبانی BERT توسط شرکت گوگل در سال 2019 ارائه شد. این مدل زبانی ازنظر مفهومی ساده و ازنظر تجربی بسیار قدرتمند عمل می‌کند. از ویژگی‌های مدل زبانی BERT به بهره‌گیری از ترانسفورمرها و معماری چندلایه آن می‌توان اشاره کرد. از این مدل زبانی برای کارهایی مانند پاسخ به سؤال، تجزیه‌وتحلیل احساسات، و طبقه‌بندی متن استفاده‌شده است. BERT همچنین برای درک زبان در چت‌بات‌ها[1] و دستیاران مجازی بکار گرفته می‌شود [17]. BERT در دو سایز، bert-base و bert-large نیز ارائه‌شده که برای این پژوهش از bert-base بهره گرفتیم. مدل زبانی bert-base دارای 12 لایه، 768 لایه پنهان[2] و مجموع 110 میلیون مقدار است. این مدل زبانی بر روی میلیاردها کلمه از داده‌های متنی شامل BooksCorpus (800 میلیون کلمه) و ویکی‌پدیای انگلیسی (2500 میلیون کلمه) از قبل آموزش‌دیده است [3].

## 2-4- مروری بر کارهای انجام‌شده

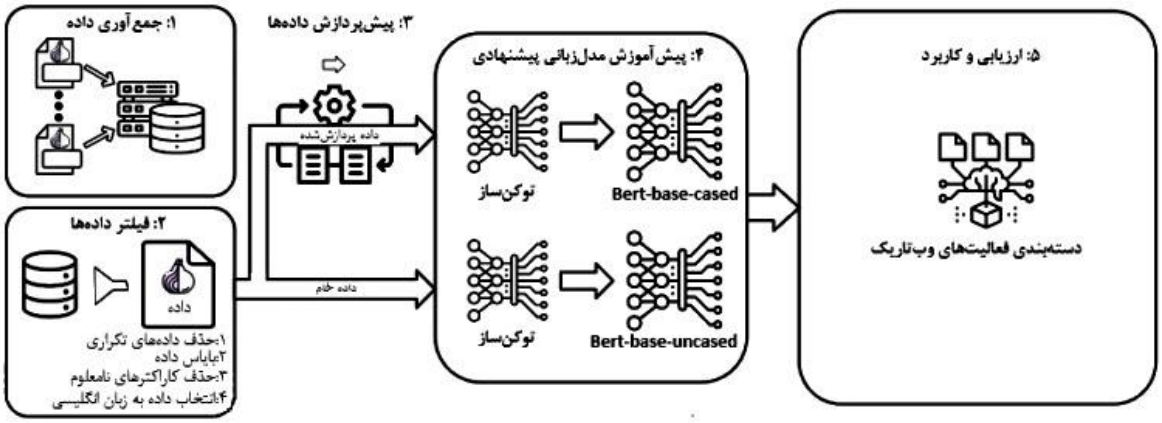
وب تاریک اغلب با فعالیت‌های غیرقانونی و مجرمانه مرتبط است. ازجمله این فعالیت‌ها می‌توان به میزان اسناد مرتبط با معاملات مواد مخدر یا حملات سایبری اشاره کرد. به دلیل اینکه داده‌های وب‌تاریک در زمینه سایبری بسیار مهم و کلیدی محسوب می‌شوند و در دسته‌بندی وب‌تاریک نیز قرار دارند، در این پژوهش تمرکز داده موردنظر بر این قسمت قرارگرفته است. دامنه مجموعه فعالیت‌هایی که در وب‌تاریک انجام می‌شود بسیار متنوع است. محققان وب‌تاریک با بررسی داده‌های آن به مشخصه‌های متفاوتی برای دسته‌بندی موضوعات به روش‌های مختلف دست یافتند. بریاکوف و همکارانش [18] با نمونه‌گیری از 1813 داده مختلف به 18 دسته‌بندی رسیدند. مور و همکارانش [19] نیز به این دسته‌بندی پرداخته و دادگان را در دوازده مورد دسته‌بندی کردند و درنهایت در مقاله‌ای که توسط نبکی و همکارانش [7] منتشر شد، با بررسی تمام موارد قبلی توانستند با جمع‌آوری کامل‌ترین دادگان آن‌ها را در 26 قسمت دسته‌بندی کنند.به علت محبوبیت وب‌تاریک به‌عنوان یک پلتفرم انتخابی برای فعالیت‌های مخرب، این پلتفرم توجه محققان و کارشناسان امنیتی را به خود جلب کرده است. در تمام پژوهش‌های گذشته داده‌های مرتبط با حملات سایبری مانند اطلاعات مرتبط با کارت‌های بانکی، اطلاعات مرتبط با کاربران شبکه‌های اجتماعی و رمزارزها همچنین داده‌های مرتبط با حملات سایبری در زمینه‌های مختلف حوزه دیجیتال از تلفن همراه گرفته تا رایانامه و مواردی از این قبیل، در دسته‌بندی دادگان موردمطالعه‌شان قرارگرفته است. در این پژوهش نیز 6788 مورد داده تهیه‌شده در سه دسته داده مرتبط با حملات سایبری، داده تجمیع شده در حوزه دیجیتال و شبکه‌های اجتماعی و موارد مرتبط با رمزارزها و تراکنش‌های مرتبط با آن مورداستفاده قرارگرفته است.

اخیراً جین و همکارانش [11] در سال 2022 مشاهده کردند که استفاده از یک مدل طبقه‌بندی مبتنی بر مدل زبانی BERT به عملکرد پیشرفته‌ای در میان روش‌های پردازش زبان طبیعی موجود در وب‌تاریک دست می‌یابد. آن‌ها به این موضوع اشاره‌کرده‌اند که ماهیت پنهان و دسترسی محدود وب‌تاریک همراه با فقدان مجموعه داده‌های عمومی در این حوزه، مطالعه ویژگی‌های ذاتی آن مانند ویژگی‌های زبانی را دشوار می‌کند. آن‌ها با بررسی پژوهش‌های قبلی و بیان این نکته که برای طبقه‌بندی داده‌های وب‌تاریک به دلیل تفاوت‌های زبانی وب سطحی و وب‌تاریک با استفاده از شبکه‌های عصبی عمیق ممکن است مؤثر و مفید واقع نشود، پرداخته‌اند. معرفی مجموعه داده جدید از داده‌های وب‌تاریک نیز در این پژوهش انجام‌شده که مبتنی بر ده هزار داده وب‌تاریک است. نویسندگان پژوهش به بررسی تفاوت‌های زبانی بین داده‌های وب‌تاریک تهیه‌شده، پرداخته و تمرکز خود را بر دسته‌بندی آن‌ها با کمک گرفتن از مدل‌های زبانی گذاشته‌اند. آن‌ها ابتدا جمع‌آوری داده را انجام داده، پس‌ازآن تلاش کردند تا داده‌های تکراری و با محتوای اندک را از بین داده‌های تهیه‌شده حذف و به دسته‌بندی زبانی بپردازند. در ادامه ارزیابی‌های مختلف را روی مجموعه داده خود از جهت

[1] Chat bots
[2] Hidden Layers

مرتبط بودن داده‌ها مورد ارزیابی قرار داده و درنهایت با کمک گرفتن از مدل‌های زبانی جهت دسته‌بندی داده وب‌تاریک، مجموعه داده به ده دسته تقسیم‌شده که به گفته نویسندگان این پژوهش، دسته‌بندی داده‌ها بسیار موفق و با دقت بالایی صورت پذیرفته است. آن‌ها برای آموزش مدل زبانی جهت دسته‌بندی داده‌ها از تمام موضوعات استفاده کرده و داده‌ها با زبان‌های مختلف را نیز در بررسی خود لحاظ نموده‌اند. جین و همکارانش [4] همچنین در سال 2023 شروع به طراحی یک مدل زبانی بر پایه دادگان وب‌تاریک نمودند. نویسندگان این پژوهش پس از ارزیابی‌های مختلف دلیل برتری مدل زبانی خود نسبت به دیگر مدل‌ها را استفاده از دادگان وب‌تاریک در آموزش مدل زبانی اعلام کرده‌اند.

بر همین اساس هدف از دسته‌بندی داده‌های وب‌تاریک به کمک مدل زبانی BERT در این مقاله به جهت افزایش بهره‌وری مدل برای مدیریت تهدیدات سایبری و شناسایی فعالیت‌های مخرب آن در نظر گرفته شد. دسته‌بندی دادگان موجب خواهد شد تا مدل زبانی دقت و درک بهتری از ارتباط معنایی بین کلمات داشته باشد که بر همین مبنا می‌توان خروجی موردنظر را توسط مدل زبانی تهیه کرد.




**شکل (2):** معماری کلی روش پیشنهادی

## 3- راهکار پیشنهادی

### 3-1- معماری کلی روش پیشنهادی

نمای کلی روش پیشنهادی در شکل (2) نمایش داده‌شده است. همان‌طور که در این شکل دیده می‌شود، برای مدل پیشنهادی ابتدا باید داده موردنظر، جمع‌آوری و تهیه شود. با جمع‌آوری داده‌های استاندارد در حوزه وب‌تاریک، اقدام به فیلترکردن داده‌ها صورت پذیرفت. این اقدامات شامل حذف موارد تکراری، بایاس کردن داده‌ها، حذف کلمات نامعلوم و انتخاب داده به زبان انگلیسی می‌شود. برای اینکه بتوانیم ارزیابی بهتری درباره مدل پیشنهادی داشته باشیم، مجموعه دیگری نیز بدون اقدامات فیلترینگ تهیه نمودیم. در گام بعدی پژوهش، داده‌ها را به تفکیک‌کننده جهت فرایند عددی کردن داده‌ها تزریق نموده و اقدام به آموزش مدل زبانی کردیم. انتخاب دو زیرمجموعه مدل زبانی BERT برای ارزیابی بیشتر نیز صورت پذیرفت. درنهایت با ارزیابی معیارها، دقت و صحت مدل محاسبه شد. در ادامه به شرح بخش‌های مختلف معماری پیشنهادی خواهیم پرداخت.

### 3-2- پیش‌پردازش[1]

ابتدا کتابخانه‌های موردنیاز برای اجرای برنامه ازجمله ترانسفورمرها را فراخوانی می‌کنیم. فراخوانی کتابخانه‌های بروز موجب می‌شود تا مدل با بهره‌گیری از این کتابخانه‌ها، نتایج بهتری را با خود به همراه داشته باشد.

**جدول (3):** نمونه اولیه داده فراخوانی شده

| | Category | Lang | Value |
|---|---|---|---|
| **0** | Others | En | dark tor \| onion links \| deep web \| hidden wi... |
| **1** | Others | En | dwl - deep web links , dwl - deep web links... |
| **2** | Hacking | En | FBI Hacking AND FORENSIC TOOLKIT |
| **3** | Hacking | En | WATCH ME CashOut Credit card in Front of |
| **4** | Electronic | En | apple market - stolen & carded merchandise |

معماری ترانسفورمرها ساخت مدل‌های با ظرفیت بالاتر را تسهیل کرده‌اند. ترانسفورمرها ویژگی‌های بیشتری برای کاربر اضافه می‌کند تا امکان دریافت آسان ویژگی‌ها، تنظیم دقیق مدل‌ها و همچنین انتقال بی‌وقفه به اجرای مدل را فراهم کند. ترانسفورمرها تا حدودی با این ویژگی‌ها سازگاری دارند که مستقیماً شامل ابزاری برای انجام استنتاج با استفاده از مدل‌های

[1] Preprocessing

*BERT* هستند *[20]*. در جدول (3) نمونه اولیه فراخوانی داده را مشاهده می‌کنید.

### 3-2- 1- برچسب‌گذاری عددی داده‌ها[1]

برای هر برچسب داده باید عددی اختصاص داده شود تا برای انجام پردازش و فهم ماشین، سهولت ایجاد گردد. با عددی کردن برچسب داده‌ها، به Others مقدار 0، Hacking مقدار 1 و به Electronics مقدار 2 اختصاص یافت. در جدول (4) ، ستون آخر نمونه عددی شده هر دسته را پس از فراخوانی مشاهده می‌کنید.

**جدول (4):** نمونه عددی شده هر دسته در دادگان

| | Category | Lang | Value | lab |
|---|---|---|---|---|
| **0** | Others | En | dark tor \| onion links \| deep web \| hidden wi... | 0 |
| **1** | Others | En | dwl - deep web links , dwl - deep web links... | 0 |
| **2** | Hacking | En | FBI Hacking AND FORENSIC TOOLKIT | 1 |
| **3** | Hacking | En | WATCH ME CashOut Credit card in Front of | 1 |
| **4** | Electronic | En | apple market - stolen & carded merchandise | 2 |

### 3-2- 2- تفکیک سازی داده‌ها

در پردازش زبان طبیعی، اولین چالش تبدیل متن به اعداد برای استفاده توسط هر الگوریتمی است که محقق روش مدنظر خود را انتخاب می‌کند. نمایش کلمات به‌عنوان بردار عددی با تکیه‌بر مطالب یادگیری شده توسط مدل زبانی پایه، به یکی از روش‌های مؤثر برای تحلیل متون در یادگیری ماشین تبدیل‌شده است، به‌طوری‌که هر کلمه با یک بردار نشان داده می‌شود تا معنای آن را مشخص کند یا بداند این کلمه چقدر از بقیه کلمه‌های دیگر نزدیک یا دور است [21]. برای این منظور از مدل و تفکیک‌کننده BERT در دو حالت bert-base-cased و bert-base-uncased استفاده شد. با انتخاب تفکیک‌کننده bert-base-cased، تفکیک‌کننده بین Web و web تفاوت قائل می‌شود و به هرکدام شناسه متفاوتی تخصیص داده خواهد شد. نمونه‌ای از تفکیک سازی توسط این مدل انتخابی در جدول (5) نمایش داده‌شده است.

**جدول (5):** تفکیک ساز bert-base-cased

| | |
|---|---|
| **Original Text:** | Hi. Hacking , Dark Web, dark web analysis |
| **Tokenized Text:** | [‘Hi’, ‘.’, ‘Ha’, ‘##cking’, ‘,’, ‘Dark’, ‘Web’, ‘dark’, ‘web’, ‘analysis’] |
| **Token IDs:** | [8790, 119, 11679, 12944, 117, 4753, 9059, 117, 1843, 5127, 3622] |

با انتخاب تفکیک ساز bert-base-uncased نیز تفکیک ساز بین Web و web تفاوت قائل نمی‌شود و به هرکدام شناسه یکسانی تخصیص داده خواهد شد. نمونه‌ای از تفکیک سازی این مدل در جدول (6) نمایش داده‌شده است.

[1] Numerical Labeling of Data

**جدول (6):** تفکیک ساز bert-base-uncased

| | |
|---|---|
| **Original Text:** | Hi. Hacking , Dark Web, dark web analysis |
| **Tokenized Text:** | [‘hi’, ‘.’, ‘hacking’, ‘,’, ‘dark’, ‘web’, ‘dark’, ‘web’, ‘analysis’] |
| **Token IDs:** | [7632, 1012, 23707, 1010, 2601, 4773, 1010, 2601, 4773, 4106] |

آنچه مشخص است، در برخی مواقع در زبان انگلیسی در برخی موارد تفاوت‌هایی بین کلمات با حروف بزرگ و کوچک وجود دارد. از این حیث در این پژوهش هر دو مورد تفکیک ساز مورد ارزیابی قرار گرفت.

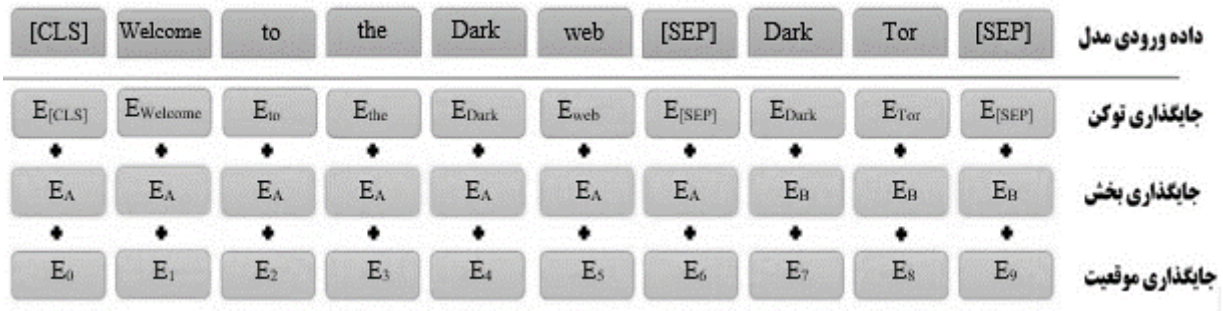
### 3-2- 3- جایگذاری[2] و رمزگذاری[3] تفکیک ساز

فعالیت رمزگذار به این صورت است که متنی را می‌گیرد، آن را به دنباله‌ای از نشانه‌های قابل‌فهم و با چارچوب مدل زبانی، در ابعاد مشخص‌شده برای یک جمله تبدیل می‌کند. در ادامه نشانه‌های خاص [CLS] برای مشخص‌شده ابتدای جمله و [SEP] را به‌عنوان جداکننده جملات به متن اضافه می‌کند و عملیات رمزگذاری را انجام می‌دهد. BERT یک معماری ترانسفورماتور "فقط رمزگذار" است [3]. آنچه را که در شکل (3) مشاهده می‌کنید، نمای کامل از روند جایگذاری تفکیک ساز مدل زبانی BERT است. این تفکیک ساز در چند مرحله، فعالیت خود را کامل می‌کند و نتیجه کار، داده ورودی برای آموزش مدل زبانی خواهد بود. برای یک متن داده‌شده، مقدار ورودی آن با جمع‌کردن و جایگذاری شناسه[4] هر کلمه، جایگذاری بخش‌بندی[5] متن داده‌شده و جایگذاری موقعیت[6] مربوط به هر کلمه در متن ساخته می‌شود. تمام داده موجود در دادگان به ترتیب، توسط بخش جایگذاری پردازش می‌شوند و در ادامه برای تزریق به مدل زبانی آماده می‌شوند. درنهایت، بردار عددی با شرایط جدید برای هر داده دادگان تهیه و محاسبه می‌شود. این بردار شامل تمام مشخصه‌های جایگذاری شده نیز هست.

### 3-3- آموزش

پیش از هر چیز، ابتدا باید مدل موردنیاز برحسب منابع در دسترس، برای انجام عمل آموزش را انتخاب کنیم. ازآنجایی‌که انتخاب و پیش آموزش[7] مدل زبانی BERT به میزان بالایی از سخت‌افزار نیازمند است، برای انجام این پژوهش از زیرمجموعه مشتق شده BERT یعنی bert_base استفاده کردیم. Bert-base دارای دو مجموعه bert-base-cased و bert-base-uncased است که این زیرمجموعه‌ها از مدل زبانی BERT آموزش داده‌شده با 110 میلیون متغیر، با تمرکز بر زبان انگلیسی، دارای 12 لایه و 768 لایه پنهان هستند.

[2] Embedding
[3] Encoding
[4] Token Embedding
[5] Segment Embedding
[6] Position Embedding
[7] Pretraining




**شکل (3)**: فرایند جای‌گذاری تفکیک‌کننده BERT

### 3-3- 1-تنظیم متغیرهای مدل پیشنهادی

تنظیم دقیق مدل، برحسب اطلاعاتی که در مرحله تفکیک ساز اتفاق افتاد، گام بعدی است. تنظیم مناسب برای پیش آموزش مدل زبانی امری مهم و حیاتی است. استفاده از بهینه‌ساز[1] AdamW در این مدل پیش‌بینی‌شده است. بهینه‌ساز AdamW یک روش نزولی گرادیان تصادفی است. این بهینه‌ساز بر اساس تخمین تطبیقی مکان‌های مرتبه اول و مرتبه دوم با روشی اضافه‌شده به وزن‌های فروپاشی کار می‌کند [22]. مقادیری را که این بهینه‌ساز در خود جای‌داده، شامل نرخ یادگیری[2] و تعداد دوره‌های[3] آموزش مدل است. مقادیر داده‌شده به مدل در جدول (7) نشان داده‌شده است.

**جدول (7)**: مقادیر مدل زبانی پیشنهادی

| | |
|---|---|
| maximum_sequence_length | 256 token |
| Learning_rate | 2e-5 |
| Epochs | 10 |
| Cross_entropy | Loss |
| Warmup_step | 0 |
| Batch_size | 32 |

### 3-3- 2-روند آموزش مدل پیشنهادی

حال نوبت به آموزش مدل رسیده است. ازاین‌پس تمامی مواردی که در مراحل قبل تهیه و تنظیم‌شده بود، طبق مراحل زیر مورداستفاده قرار می‌گیرد. مراحل به این شرح است که:

1. ورودی داده‌ها و برچسب‌های آن‌ها برای پردازش، به مدل داده می‌شود.
2. داده‌های برای اینکه دارای شتاب بیشتری باشند به واحد GPU منتقل می‌گردد.
3. گرادیان‌ها محاسبه و گرادیان‌های قبلی حذف می‌شوند.
4. سپس داده‌های جدید دوباره به سیستم وارد می‌شوند.
5. به شبکه گفته می‌شود که مقادیر بهینه‌ساز را به‌روزرسانی کند.
6. متغیرها را برای ارزیابی و نظارت بر پیشرفت بررسی می‌کند.

[1] Optimizer
[2] Learning Rate
[3] Epochs

درنهایت با توجه به اینکه پیش آموزش مدل زبانی در پایتورچ[4]، مقادیر دلخواه برای دقت سنجی و ارزیابی مدل را نمایش نخواهد داد، توابعی را برای نمایش این مقادیر در نظر گرفتیم و مدل را با تنظیمات موردنظر، آموزش دادیم.

## 4- آزمایش‌ها و نتایج

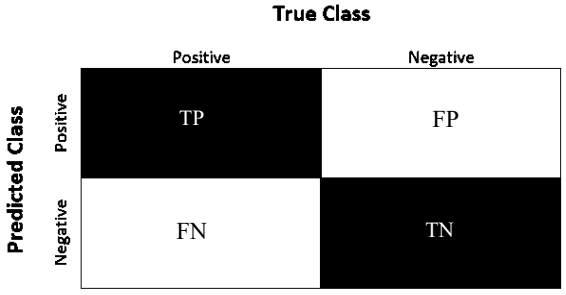
### 4-1- ارزیابی[5] و معیارها

پس از تخصصی کردن مدل برای یک کار خاص، ارزیابی عملکرد آن و ارزیابی اثربخشی آن در صورت ارائه سناریوهای دنیای واقعی مهم است. با اجرای یک ارزیابی برای مدل‌های زبانی، می‌توان میزان سازگاری مدل با وظیفه یا دامنه هدف را سنجید. برای درک اینکه چرا ارزیابی مدل‌های زبانی محوری و مهم است، باید سرعت در حال گسترش کاربردهای آن‌ها را در نظر بگیریم؛ بنابراین می‌توان گفت فرایند ارزیابی مدل‌های زبانی به دلایل متعددی ضروری است.

### 4-1- 1-ماتریس درهم‌ریختگی[6]

ماتریس درهم‌ریختگی برای تحلیل ارزیابی عملکرد الگوریتم‌ها در یادگیری ماشین[7] استفاده می‌شود. بیشترین کاربرد این ماتریس در الگوریتم‌های یادگیری با ناظر[8] است اگرچه از این ماتریس برای الگوریتم‌های یادگیری بدون ناظر[9] هم استفاده می‌شود که به ماتریس تطابق معروف است. این ماتریس N*N بوده که N تعداد کلاس‌های موردنظر را نشان می‌دهد [28] [29]. بعد از هر بار عمل تحلیلی، نتایج به‌دست‌آمده را باید در چهار گروه مثبت صحیح[10]، مثبت کاذب[11]، منفی صحیح[12] و منفی کاذب[13] دسته‌بندی کرد تا بتوان کیفیت عمل تحلیل را مورد ارزیابی قرارداد [23]. تصویر کلی ماتریس در شکل (4) قابل‌مشاهده است.

[4] PyTorch
[5] Evaluate
[6] Confusion Matrix
[7] Machine Learning
[8] Supervised
[9] UnSupervised
[10] True Positive
[11] True Negative
[12] False Positive
[13] False Negative




**شکل (4):** ماتریس درهم‌ریختگی

### 4-1- 2-تشخیص پذیری[1]

متغیر تشخیص پذیری که به آن نرخ پاسخ‌های منفی درست هم می‌گویند. تشخیص پذیری به معنی نسبتی از موارد منفی است که در زمان آزمون به‌درستی به‌عنوان نمونه منفی تشخیص داده‌شده‌اند. تعداد درخواست‌های نرمالی که به‌اشتباه به‌عنوان غیر نرمال تشخیص داده‌شده‌اند، هیچ تأثیری در محاسبه این متغیر ندارد [24][26]. این مقدار به‌صورت رابطه (1) محاسبه می‌شود.

$$Specifity = \frac{tn}{tn+fp} = 1 - FPR = 1 - \frac{fp}{tn+fp} \qquad (1)$$

### 4-1- 3-دقت

دقت، پرکاربردترین و شاید اولین انتخاب برای ارزیابی عملکرد یک الگوریتم در مسائل طبقه‌بندی است. دقت به این معناست که مدل تا چه اندازه خروجی را درست پیش‌بینی می‌کند. این معیار اطلاعات جزئی در مورد کارایی مدل ارائه نمی‌دهد. پس از پیش آموزش مدل زبانی، این معیار درزمانی که پایگاه داده بالانس نباشد نمی‌تواند معیار خوب و مقدار دقیقی برای ارزیابی عملکرد مدل زبانی باشد [26]. دقت به‌صورت رابطه (2) محاسبه می‌شود.

$$Accuracy = \frac{tp + tn}{tp+tn+fp+fn} \qquad (2)$$

### 4-1- 4-یادآوری[2]

یادآوری معیاری است که تعداد دفعاتی که یک مدل یادگیری ماشین به‌درستی نمونه‌های مثبت را از تمام نمونه‌های مثبت واقعی در دادگان شناسایی می‌کند، اندازه‌گیری می‌کند. این معیار مشخص می‌کند که مدل به چه اندازه در تشخیص تمام درخواست‌ها موفق بوده است. این مقدار نشانگر این است که مدل در درصد بالایی از تشخیص درخواست‌های موفق، عملکرد خوبی از خود نشان داده است. درواقع زمانی که پژوهشگر از این متغیر به‌عنوان متغیر ارزیابی برای مدل خود استفاده می‌کند، هدفش دستیابی به نهایت صحت در تشخیص نمونه‌های کلاس مثبت است [25][26]. این متغیر به‌صورت رابطه (3) محاسبه می‌شود.

$$Recall = \frac{tp}{tp+fn} \qquad (3)$$

[1] Specifity
[2] Recall

### 4-1- 5-صحت[3]

زمانی که ارزش مثبت کاذب بالا باشد، معیار صحت، معیار مناسبی برای ارزیابی خواهد بود [29]. صحت به‌سادگی نشان می‌دهد که چه تعداد از داده انتخاب‌شده مرتبط هستند. درواقع صحت، نسبت موارد صحیح طبقه‌بندی‌شده توسط الگوریتم از یک کلاس مشخص، به‌کل تعداد مواردی که الگوریتم در آن کلاس طبقه‌بندی کرده است را نشان می‌دهد که در رابطه (4) قابل‌مشاهده است.

$$Precision = \frac{tp}{tp+fp} \qquad (4)$$

### 4-1- 6-معیار *F1*

این معیار که به‌عنوان f-score یا f-measure نیز شناخته می‌شود، برای محاسبه عملکرد یک الگوریتم، در نظر گرفته می‌شود. این معیار از ترکیب دو مقدار یادآوری و صحت حاصل می‌شود. معیار F1 به‌صورت رابطه (5) تعریف می‌شود. این معیار امروزه در بیشتر تحقیقات؛ چه دو کلاسه و چه چند کلاسه، برای ارزیابی مدل مورداستفاده قرار می‌گیرد[29][26].

$$f - measure = \frac{2*recall*precision}{recall+precision} \qquad (5)$$

## 4-2- دادگان

دادگان به‌عنوان زیرساخت بنیادی، مشابه سیستم پایه‌ای عمل می‌کنند که توسعه مدل‌های زبانی بزرگ را حفظ و پرورش می‌دهد. تهیه دادگان برای آموزش مدل زبانی امری بسیار حیاتی و مهم است. داده‌های آموزشی باکیفیت بالا، پایه و اساس یادگیری ماشینی موفق است؛ زیرا کیفیت داده‌های آموزشی تأثیر عمیقی بر توسعه، عملکرد و دقت هر مدل دارد و به مدل می‌گوید که به دنبال چه چیزی باشد. جهت آموزش و ارزیابی مدل به یک دادگان استاندارد که بتوان از طریق آن عملکرد مدل را با دیگر مدل‌ها و نتایج ارزیابی نمود، موردنیاز است [27].

تنظیم ساختار مدل زبانی، یک مرحله حیاتی در آموزش مدل‌های زبانی بزرگ است، بنابراین چگونگی افزایش و تأثیر تنظیم ساختار، توجه بیشتری را به خود جلب کرده است. فعالیت‌های اخیر نشان می‌دهد که کیفیت دادگان در طول تنظیم ساختار مدل زبانی، بسیار مهم‌تر از کمیت است؛ بنابراین اخیراً بسیاری از مطالعات بر کاوش روش‌های انتخاب زیرمجموعه باکیفیت بالا از دادگان‌های استاندارد، باهدف کاهش هزینه‌های آموزشی و افزایش قابلیت‌های پیروی از دستورالعمل‌های مدل‌های زبانی تمرکز دارند [28].

بنابراین، برای افزایش کیفیت دادگان، با تمرکز برداده زبان انگلیسی و حذف موارد تکراری و همچنین ردیابی و حذف کاراکترهای نامشخص در داده‌های موجود، برای ایجاد دادگان بایاس شده، تلاش شد تا دادگان باکیفیت بالا تهیه شود. تهیه

[3] Precision

دادگان باکیفیت در اخذ نتایج مناسب پس از آموزش مدل زبانی پیشنهادی، بسیار مفید و مؤثر واقع شد. برای بررسی این ادعا، با تهیه یک دادگان خام، فرایند آموزش مدل را به چالش کشیدیم که حاصل آن نتایج ضعیف‌تری نسبت به داده باکیفیت بالابود. در ادامه در جدول (8)، نمونه دو دادگان را مشاهده می‌کنید.

**جدول (8):** داده خام و داده پیش‌پردازش شده

| | |
|---|---|
| Raw Data | 网址导航栏目： ud 导航： ID_ONION_URL 暗网最大中文交易市场： ID_ONION_URL cvv アン エロいの tor板 盗刷： onion ちゃんねる ごみ箱 ID_ONION_URL hacker： グラ板 ID_ONION_URL ilcosmista : hobbylandwirtschaft • japanische legewachteln • marans-hühner • jagdfasane<br><br>âœ?300k email pass combo [Netflix,Minecraft,Uplay,Steam,Hulu,Brazzer,spotif ?y]âœ |
| Normalized Data | i am a independent security researcher. Hacking and social engineering is my business since i was ID_NUMBER years old. i never had a real job so i had the time to get really good at this because i have spent the half of my life studying and researching about Hacking, engineering and web technologies. i have worked for other people before in silk road and now i'm also offering my services for everyone with enough cash here. |

## 4-3- پیاده‌سازی

جهت پیاده‌سازی مدل از برنامه‌نویسی به زبان پایتون[1] نسخه 3 به همراه مجموعه دادگان تهیه‌شده بهره گرفته‌ایم. استفاده از پایتورچ نیز در این پژوهش در نظر گرفته‌شده است. پایتورچ یک چارچوب منبع باز یادگیری ماشین است که بر اساس زبان برنامه‌نویسی پایتون و کتابخانه تورچ است که یک کتابخانه متن‌باز یادگیری ماشین است و برای ایجاد شبکه‌های عصبی عمیق استفاده می‌شود و یکی از پلتفرم‌های ترجیحی برای تحقیقات یادگیری عمیق است. این چارچوب برای سرعت بخشیدن به روند بین نمونه‌سازی تحقیقاتی ساخته‌شده است [29]. در تمامی مراحل آموزش و آزمایش مدل پیشنهادی از سخت‌افزار در دسترس مجموعه Google Colab با GPU اختصاص‌یافته به میزان 12.68 گیگابایت بانام NVIDIA T4 استفاده‌شده است.

## 4-4- نتایج TF-IDF[2]

TF-IDF مخفف عبارت فرکانس معکوس سند فرکانس است و معیاری است که در زمینه‌های بازیابی اطلاعات و یادگیری ماشین استفاده می‌شود و می‌تواند اهمیت یا ارتباط نمایش رشته‌ها (کلمات، عبارات و غیره) را کمی کند. TF تخمین مستقیمی از احتمال وقوع یک عبارت را ارائه می‌کند که با فرکانس کل در سند، یا مجموعه سند، بسته به محدوده محاسبه، نرمال می‌شود و از سوی دیگر، IDF را می‌توان به‌عنوان مقدار اطلاعات تفسیر کرد. بردار سازی TF-IDF شامل محاسبه امتیاز TF-IDF برای هر کلمه در دادگان نسبت به آن سند و سپس قراردادان آن اطلاعات در یک بردار است [30]. نتایج ده کلمه برتر دسته نفوذ در جدول (9) تهیه و نمایش داده‌شده است.

**جدول (9):** TF-IDF ده کلمه برتر دسته نفوذ

| Hacking | |
|---|---|
| Vortex | 0.046 |
| Transfers | 0.026 |
| Cashapp | 0.012 |
| Hackers | 0.012 |
| Services | 0.009 |
| Hack | 0.008 |
| Quick | 0.006 |
| Bitcoin | 0.002 |
| Solutions | 0.002 |
| Financial | 0.002 |

## 4-5- نتایج مرحله تفکیک ساز

پیش از هر چیز، ابتدا باید دادگان تهیه‌شده موردبررسی و تحلیل قرار گیرند تا بتوان توسط نتایج به‌دست‌آمده، مدل را برای مرحله پیش آموزش، بهتر تنظیم نمود. انتخاب تفکیک ساز می‌تواند به‌طور قابل‌توجهی بر عملکرد پایین‌دستی و هزینه‌های آموزشی مدل تأثیر بگذارد. یک تفکیک ساز وظیفه آماده‌سازی ورودی‌های یک مدل را بر عهده دارد. ماشین‌ها زبان را آن‌طور که هست نمی‌خوانند، بنابراین باید به اعداد تبدیل شوند و اینجاست که تفکیک‌سازها به کمک می‌آیند. به همین علت ابتدا با واردکردن دادگان به تفکیک ساز، فرایند برچسب زدن عددی به آن‌ها صورت خواهد گرفت [31]. حاصل یکی از سلول‌های تفکیک‌شده را می‌توانید در جدول (10) مشاهده کنید.

**جدول (10):** نتایج تفکیک ساز

| | |
|---|---|
| **Original:** | your browser is not running JavaScript that is required to use the whistleblowing client. |
| **Tokenized:** | ['your', 'browser', 'is', 'not', 'running', 'java', '##script', 'that', 'is', 'required', 'to', 'use', 'the', 'whistle', '##bl', '##owing', 'client', '.'] |
| **Token IDs:** | [2115, 16602, 2003, 2025, 2770, 9262, 22483, 2008, 2003, 3223, 2000, 2224, 1996, 13300, 16558, 14138, 7396, 1012] |

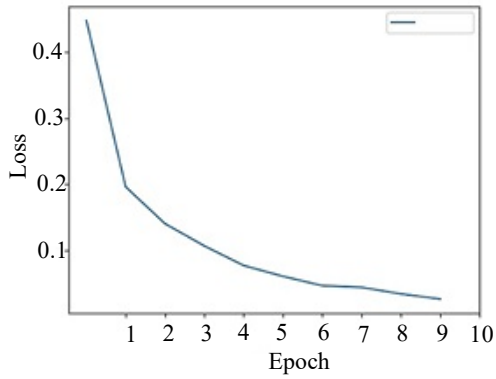
## 4-6- نتایج مرحله پیش آموزش

برای کسب بهترین نتایج پیش آموزش مدل، ابتدا نیاز است تا متغیر مدل به‌گونه‌ای تنظیم شود تا خروجی مناسب را در اختیارمان قرار دهد. پس از اتمام دوره‌های آموزش مدل، نتایج پیش آموزش مدل نمایش داده می‌شود. در جدول (11) مقادیر میزان خطای[3] حاصل‌شده در مدت‌زمان پیش آموزش مدل را مشاهده می‌کنید.

[1] Python
[2] Term Frequency-inverse Document Frequency
[3] Loss

**جدول (11)**: مقادیر خطای مدل پیشنهادی

| Epoch | Training loss |
|---|---|
| **1** | 0.444649 |
| **2** | 0.187452 |
| **3** | 0.125603 |
| **4** | 0.100994 |
| **5** | 0.071258 |
| **6** | 0.059026 |
| **7** | 0.047790 |
| **8** | 0.032218 |
| **9** | 0.029653 |
| **10** | 0.021282 |




**شکل (5)**: نمودار میزان خطای مدل پیشنهادی

## 4-7- مقایسه نتایج

برای ارزیابی نتایج، آموزش مدل زبانی را در چند حالت مختلف بررسی نمودیم. روند مقایسه نتایج با استفاده از دو زیرمجموعه مدل زبانی BERT و با دو حالت از دادگان، در حالت داده خام و داده پیش‌پردازش شده صورت پذیرفت. افزایش کیفیت داده در دادگان پیش‌پردازش شده نتایج بهتری را نسبت به دادگان خام با خود به همراه داشت. در این پژوهش، آزمایش‌های متعدد روی مجموعه‌های داده برای این دو کار نشان می‌دهد که بدون استفاده از هیچ ویژگی خارجی، یک مدل ساده مبتنی بر BERT می‌تواند به عملکردی بهتری دست یابد. Bert-base-cased زیرمجموعه از مدل زبانی BERT حساس به حروف بزرگ است و بین english و English تفاوت قائل است [32]. برای این حالت، ورودی داده به دو صورت داده خام و داده پیش‌پردازش شده به مدل تزریق شد. با تهیه داده خام به‌عنوان داده ورودی مدل برای فرایند آموزش مدل زبانی، پس از تزریق داده‌های خام، نتایج جدول (12) حاصل شد.

**جدول (12)**: نتایج داده خام برای bert-base-cased

| | Precision | Recall | F1-score |
|---|---|---|---|
| **Others** | 0.96 | 0.95 | 0.96 |
| **Hacking** | 0.93 | 0.93 | 0.93 |
| **Electronic** | 0.87 | 0.89 | 0.88 |
| **Accuracy** | **0.92** | | |
| **Macro avg** | 0.92 | 0.92 | 0.92 |
| **Weighted avg** | 0.93 | 0.93 | 0.93 |

همان‌طور که در جدول بالا مشخص است، میزان دقت حاصل‌شده برای این حالت 92 درصد است. نتیجه حاصل برای معیارهای ارزیابی Precision معادل 92 درصد، برای معیار Recall معادل 92 درصد و برای f1 نیز معادل 92 درصد محاسبه‌شده است. انجام فرایند پردازش روی دادگان، با حذف موارد تکراری، حذف کاراکترهای نامعلوم و تمرکز روی‌داده زبان انگلیسی صورت گرفت. طی این فرایند، دادگان باکیفیت بالا برای تزریق به مدل زبانی جهت آموزش، فراهم شد. پس از اتمام آموزش مدل زبانی، نتایج جدول (13) مشاهده گردید که شامل دقت 94 درصد به همراه Precision معادل 94 درصد، Recall معادل 93 درصد و f1 معادل 93 درصد، حاصل آموزش مدل زبانی با داده پیش‌پردازش شده است.

**جدول (13)**: نتایج داده پیش‌پردازش شده برای bert-base-cased

| | Precision | Recall | F1-score |
|---|---|---|---|
| **Others** | 0.99 | 0.96 | 0.97 |
| **Hacking** | 0.93 | 0.93 | 0.93 |
| **Electronic** | 0.91 | 0.92 | 0.91 |
| **Accuracy** | **0.94** | | |
| **Macro avg** | 0.94 | 0.94 | 0.94 |
| **Weighted avg** | 0.94 | 0.94 | 0.94 |

Bert-base-uncased نیز زیرمجموعه از مدل زبانی BERT است که حساس به حروف بزرگ و کوچک نیست و بین english و English تفاوت قائل نمی‌شود [42]. برای این حالت مانند حالت قبلی، ورودی داده به دو صورت داده خام و داده پیش‌پردازش شده به مدل تزریق شد. پس از تنظیم مدل زبانی برای فرایند پیش آموزش، فراخوانی این زیرمجموعه از مدل زبانی با متغیرهای مشخص‌شده در پژوهش، صورت گرفت که با تزریق داده خام به زیرمجموعه انتخابی، آموزش مدل انجام شد. طی این عملیات، ارزیابی نتایج با داده خام تزریق‌شده طبق جدول (14) مشاهده گردید. داده خام تزریق‌شده در این مرحله با داده خام در حالت قبل، یکسان و بدون تغییر در نظر گرفته‌شده است.

**جدول (14)**: نتایج داده خام برای bert-base-uncased

| | Precision | Recall | F1-score |
|---|---|---|---|
| **Others** | 0.97 | 0.93 | 0.95 |
| **Hacking** | 0.90 | 0.90 | 0.90 |
| **Electronic** | 0.83 | 0.88 | 0.85 |
| **Accuracy** | **0.90** | | |
| **Macro avg** | 0.90 | 0.90 | 0.90 |
| **Weighted avg** | 0.90 | 0.90 | 0.90 |

همان‌طور که در جدول بالا مشاهده می‌شود، میزان دقت حاصل‌شده برای این حالت 90 درصد است. نتیجه حاصل برای معیارهای ارزیابی Precision معادل 90 درصد، برای معیار Recall معادل 90 درصد و برای f1 نیز معادل 90 درصد محاسبه‌شده است. تزریق داده پیش‌پردازش شده دادگان باکیفیت بالا، برای تزریق به مدل زبانی جهت آموزش مدل در حالت bert-base-uncased نیز فراهم شد. پس از اتمام آموزش مدل زبانی، نتایج جدول (15) مشاهده گردید.

**جدول (15)**: نتایج داده پیش‌پردازش شده برای bert-base-uncased

| | **Precision** | **Recall** | **F1-score** |
|---|---|---|---|
| **Others** | 1.00 | 0.93 | 0.97 |
| **Hacking** | 0.89 | 0.96 | 0.92 |
| **Electronic** | 0.93 | 0.88 | 0.90 |
| **Accuracy** | **0.93** | | |
| **Macro avg** | 0.94 | 0.92 | 0.93 |
| **Weighted avg** | 0.93 | 0.93 | 0.93 |

دقت 93 درصد به همراه Precision معادل 94 درصد، Recall معادل 92 درصد و f1 معادل 93 درصد، حاصل آموزش مدل زبانی با داده پیش‌پردازش در حالت bert-base-uncased شده است.

برای ارزیابی نتایج، آموزش مدل زبانی را در چند حالت مختلف بررسی نمودیم. روند مقایسه نتایج با استفاده از دو زیرمجموعه مدل زبانی BERT و با دو حالت از دادگان، در حالت خام و پیش‌پردازش شده، صورت پذیرفت. افزایش کیفیت داده در دادگان پیش‌پردازش شده نتایج بهتری را نسبت به دادگان خام با خود به همراه داشت. در این پژوهش، آزمایش‌های متعدد روی مجموعه‌های داده برای این دو کار نشان می‌دهد که بدون استفاده از هیچ ویژگی خارجی، یک مدل ساده مبتنی بر BERT می‌تواند به عملکردی بهتری دست یابد. در ادامه به تشریح حالت‌های مختلف مطرح‌شده و مقایسه آن‌ها خواهیم پرداخت. بعد از تهیه نتایج برای هر مرحله از فعالیت مدل زبانی، نوبت به آن رسیده تا تمام موارد فوق را با قراردادان در کنار هم مقایسه کنیم. به همین منظور، ابتدا نتایج معیارهای ارزیابی به تفکیک هر دسته از داده موجود در دادگان را تحلیل خواهیم کرد. سپس نتایج کلی دقت برای هر حالت از مدل زبانی بررسی می‌شود. نتایج معیارهای ارزیابی به تفکیک هر دسته در جدول (16) نمایش داده‌شده است.

**جدول (16)**: نتایج آزمایش‌های مدل پیشنهادی

| **روش** | **α** | ***Others*** | ***Hacking*** | ***Electronics*** |
|---|---|---|---|---|
| ***Bert-base-cased(raw)*** | صحت | *0.96* | *0.93* | *0.87* |
| | یادآوری | *0.95* | *0.93* | *0.89* |
| | *F1-score* | *0.96* | *0.93* | *0.88* |
| | دقت | *0.92* | | |
| ***Bert-base-cased(pre)*** | صحت | *0.99* | *0.93* | *0.91* |
| | یادآوری | *0.96* | *0.93* | *0.92* |
| | *F1-score* | *0.97* | *0.93* | *0.91* |
| | دقت | *0.94* | | |
| ***Bert-base-uncased(raw)*** | صحت | *0.97* | *0.90* | *0.83* |
| | یادآوری | *0.93* | *0.90* | *0.88* |
| | *F1-score* | *0.95* | *0.90* | *0.85* |
| | دقت | *0.90* | | |
| ***Bert-base-uncased(pre)*** | صحت | *1.00* | *0.89* | *0.93* |
| | یادآوری | *0.93* | *0.96* | *0.88* |
| | *F1-score* | *0.97* | *0.92* | *0.90* |
| | دقت | *0.93* | | |

همان‌طور که در جدول (7) مشاهده می‌شود، دقت مدل در هر دوحالتی که داده پیش‌پردازش شده به مدل تزریق‌شده، بهبود داشته است. به‌این‌ترتیب می‌توان نتیجه گرفت که با تهیه دادگان پیش‌پردازش شده و باکیفیت، دقت مدل افزایش خواهد یافت. لذا نیاز است که در طی فرایند آموزش مدل زبانی به این نکته دقت شود. در این پژوهش با تهیه دادگان عاری از کاراکترهای نامعلوم، تمرکز بر داده‌های زبان انگلیسی و بایاس کردن آن، به‌دقت مناسب مدل پیشنهادی برسیم.

نتایج حاصل از مدل پژوهش را با نتایجی که در مرجع [11] تهیه‌شده بود نیز در هر سه دسته، مقایسه کردیم. در مرجع [11] از دادگان باده دسته از داده مربوط به وب‌تاریک استفاده‌شده و برای اینکه بتوانیم مقایسه‌ای داشته باشیم، مقادیر حاصل از پردازش سه دسته‌ای که هم در مقاله مرجع و هم در این پژوهش آمده را با یکدیگر مقایسه کردیم. به‌طور مشخص، نتایج حاصل در این پژوهش دقت بالاتری نسبت به مقاله مرجع داشته است. نتایج این مقایسه در جدول (17) مشاهده می‌شود.

**جدول (17)**: مقایسه روش پیشنهادی با $BERT\text{-}base_{CoDA}$

| روش | α | *Others* | *Hacking* | *Electronics* |
|---|---|---|---|---|
| $Bert\text{-}base_{CoDA}$ | صحت | *0.90* | *0.87* | *0.94* |
| | یادآوری | *0.89* | *0.89* | *0.96* |
| | *F1-score* | *0.90* | *0.88* | *0.95* |
| | دقت | *0.91* | | |
| $Bert\text{-}base_{Stagg}$ | صحت | *0.99* | *0.93* | *0.91* |
| | یادآوری | *0.96* | *0.93* | *0.92* |
| | *F1-score* | *0.97* | *0.93* | *0.91* |
| | دقت | *0.94* | | |

نکته مهم با توجه به نتایج قرار داده‌شده در جدول، افزایش دقت تنها به دلیل استفاده از دادگان پیش‌پردازش شده است. با بررسی‌های انجام‌شده روی دادگان استفاده‌شده در مرجع [11] مشخص شد که این دادگان بدون هیچ پیش‌پردازشی به مدل تزریق‌شده و پس‌ازآن مدل را مورد ارزیابی قرار داده‌اند. استفاده از دادگان باکیفیت موجب خواهد شد تا مدل زبانی درک بهتری از ارتباط بین کلمات هنگام آموزش و پس‌ازآن برای اتخاذ تصمیم‌های بهتر داشته باشد.

شایان‌ذکر است که آموزش مدل‌های زبانی به منابع سخت‌افزاری ازجمله واحد پردازنده گرافیک وابستگی مستقیم دارد و هرچه این منبع سخت‌افزاری ظرفیت بالاتری داشته باشد، مدل زبانی برای آموزش به زمان کمتر و دقت بیشتری خواهد رسید؛ حال‌آنکه مدل پیشنهادی تنها با بهره‌گیری از 12 گیگابایت پردازنده گرافیک آموزش داده‌شده و به‌دقت 94درصدی دست‌یافته درحالی‌که در مقایسه با مدل ارائه‌شده در مرجع [11] این عدد معادل 24 گیگابایت پردازنده گرافیکی است.

داده‌های آماری از نتایج به‌دست‌آمده در مراحل مختلف و با چارچوب‌های متفاوت، حاکی از آن است که نتایج به‌دست‌آمده در

چارچوب پیشنهادی، نتایج بهتری نسبت به الباقی موارد انجام‌شده را به خود اختصاص داده است.

## 5- نتیجه‌گیری و پیشنهاد برای کارهای آینده

آنچه مشخص است، استفاده از مدل‌های زبانی بزرگ به‌سرعت درحال‌رشد است. بهره‌گیری از این دانش روز برای کمک در برابر آسیب‌های شبکه‌های رایانه‌ای می‌تواند کمک شایانی را برای شناسایی حملات سایبری با خود به همراه داشته باشد. در این پژوهش با تکیه‌بر داده‌های تهیه‌شده طی سال‌های مختلف از فضای وب‌تاریک و سخت‌افزار مناسب برای پردازش مدل زبانی، سعی کردیم تا مدلی را با تنظیمات دقیق و مشخص آموزش دهیم تا قادر باشد طی فرایند آموزش خود، در دسته‌بندی متون وب‌تاریک، به‌دقت بالایی دست پیدا کند. همچنین مشخص شد که اگر در فرایند آموزش از داده باکیفیت بالا استفاده شود، مدل زبانی قادر است تا با افزایش 3درصدی دقت خود به بالاترین حد دقت دست پیدا کند. این روش مانند دیگر روش‌ها، دارای خطا نیز هست. روش پیشنهادی نیازمند داده به‌روز برای داشتن پیش‌بینی‌های دقیق‌تر است، به این خاطر که فضای وب‌تاریک دارای ادبیات بسیار متنوع و مختلف است که باگذشت زمان، ادبیات وب‌تاریک هم دستخوش تغییر می‌شود.

در مرحله آموزش می‌توان با تهیه دادگان به‌روز و برخط به پیشرفت و دقت مدل کمک کرد. با توجه به گستردگی دایره لغات و تغییرات در واژگان مورد کاربرد در وب‌تاریک، جمع‌آوری داده برخط امری مهم و حیاتی است. تهیه داده از دامنه‌های وب‌تاریک با گرامر زبان انگلیسی نیز باعث خواهد شد تا مدل با آموزش بهتری روبرو شود. به‌کارگیری مدل و تفکیک ساز با اهداف مرتبط امنیت سایبری نیز بسیار حائز اهمیت خواهد بود. امروزه تفکیک‌سازهای متنوعی در حوزه مدل‌های زبانی بزرگ ارائه می‌شود که می‌توان با انتخاب تفکیک سازی که در حوزه داده وب‌تاریک، موفق عمل کرده است، در جهت بهبود مدل پیشنهادی گام برداشت. در صورت در اختیار داشتن منابع سخت‌افزاری کافی، بهتر است از تعداد دوره‌های بالا برای آموزش مدل استفاده کرد. افزایش تعداد دوره‌های آموزش کمک خواهد کرد تا مدل به‌دقت بالاتری دست پیدا کند و میزان خطای مدل کاهش یابد.

## 6- مراجع


[1] Aghaei, Ehsan, and Ehab Al-Shaer. "ThreatZoom: Neural Network for Automated Vulnerability Mitigation." In *Proceedings of the 6th Annual Symposium on Hot Topics in the Science of Security*, 2019, pp. 1-3. https://doi.org/10.1145/3314058.3318167

[2] Tann, Wesley, Yuancheng Liu, Jun Heng Sim, Choon Meng Seah, and Ee-Chien Chang. "Using Large Language Models for Cybersecurity capture-the-flag Challenges and Certification Questions." 2023,*arXiv:2308.10443*. https://doi.org/10.48550/arXiv.2308.10443

[3] Devlin, Jacob. "Bert: Pre-training of Deep Bidirectional Transformers for Language Understanding." 2018, *arXiv:*1810.*04805*.

[4] Jin, Youngjin, Eugene Jang, Jian Cui, Jin-Woo Chung, Yongjae Lee, and Seungwon Shin. "DarkBERT: A Language Model for the Dark Side of the Internet." 2023, *arXiv:2305.08596*. https://doi.org/10.48550/arXiv.2305.08596

[5] Choshen, Leshem, Dan Eldad, Daniel Hershcovich, Elior Sulem, and Omri Abend. "The Language of Legal and Illegal Activity on the Darknet." 2019, *arXiv:1905.05543*. https://doi.org/10.48550/arXiv.1905.05543

[6] Al Nabki, Mhd Wesam, Eduardo Fidalgo, Enrique Alegre, and Ivan De Paz. "Classifying Illegal Activities on Tor Network Based on Web Textual Contents." In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, 2017, pp. 35-43.

[7] Liao, Xiaojing, Kan Yuan, XiaoFeng Wang, Zhou Li, Luyi Xing, and Raheem Beyah. "Acing the Ioc Game: Toward Automatic Discovery and Analysis of Open-source Cyber Threat Intelligence." In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*,2016, pp. 755-766. https://doi.org/10.1145/2976749.2978315

[8] Bradbury, Danny. "Unveiling the Dark web." *Network* security, 2014, no. 4: 14-17. https://doi.org/10.1016/S1353-4858(14)70042-X

[9] Faizan, Mohd, and Raees Ahmad Khan. "Exploring and Analyzing the Dark web: a New Alchemy." First Monday, 2019. https://doi.org/10.5210/fm.v24i5.9473

[10] Dingledine, Roger, Nick Mathewson, and Paul F. Syverson. "Tor: The Second-generation Onion Router." In USENIX security symposium, vol. 4,2004, pp. 303-320.

[11] Jin, Youngjin, Eugene Jang, Yongjae Lee, Seungwon Shin, and Jin-Woo Chung. "Shedding New Light on the Language of the Dark web." 2022, *arXiv:2204.06885*. https://doi.org/10.48550/arXiv.2204.06885

[12] Samtani, Sagar, Weifeng Li, Victor Benjamin, and Hsinchun Chen. "Informing Cyber Threat Intelligence Through Dark Web Situational Awareness: The AZSecure Hacker Assets Portal." Digital Threats: Research and Practice (DTRAP) 2, 2021, no. 4: 1-10. https://doi.org/10.1145/3450972

[13] Törnberg, Petter. "*How to Use Large-Language Models for Text Analysis*", SAGE Publications Ltd, 2024.

[14] Rajaraman, Nived, Jiantao Jiao, and Kannan Ramchandran. "Toward a Theory of Tokenization in LLMs."2024, *arXiv:2404.08335*. https://doi.org/10.48550/arXiv.2404.08335

[15] Chang, Yupeng, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen et al. "A Survey on Evaluation of Large Language Models." *ACM Transactions on Intelligent Systems and Technology* 15, 2024, no. 3: 1-45. https://doi.org/10.1145/3641289

[16] Shanahan, Murray. "Talking About Large Language Models." *Communications of the ACM* 67,2024, no. 2: 68-79. https://doi.org/10.1145/3624724

[17] Koroteev, Mikhail V. "BERT: a Review of Applications in Natural Language Processing and Understanding." 2021, *arXiv:2103.11943* (2021). https://doi.org/10.48550/arXiv.2103.11943

[18] Biryukov, Alex, Ivan Pustogarov, Fabrice Thill, and Ralf-Philipp Weinmann. "Content and Popularity Analysis of Tor Hidden Services." In 2014 IEEE 34th International Conference on Distributed Computing Systems Workshops (ICDCSW),2014, pp. 188-193.

[19] Moore, Daniel, and Thomas Rid. "Cryptopolitik And the Darknet." Survival 58,2016, no. 1: 7-38.

[20] Wolf, Thomas, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac et al. "Huggingface's Transformers: State-of-the-art Natural Language Processing."2019, arXiv:1910.03771.

[21] Mohammed, Athar Hussein, and Ali H. Ali. "Survey of Bert (bidirectional encoder representation transformer) Types." In Journal of Physics: Conference Series, vol. 1963, no. 1,2021, pp. 012173.

[22] Loshchilov, Ilya, and Frank Hutter. "Decoupled Weight Decay Regularization."2017, arXiv:1711.05101.

[23] Townsend, James T. "Theoretical Analysis of an Alphabetic Confusion Matrix.",1971, Perception & Psychophysics 9: 40-50. https://doi.org/10.3758/BF03213026

[24] Magnus, Amy L., and Mark E. Oxley. "Theory of confusion." In Applications and Science of Neural Networks, Fuzzy Systems, and Evolutionary Computation IV, vol. 4479, 2001, pp. 105-116. https://doi.org/10.1117/12.448337

[25] Grandini, Margherita, Enrico Bagli, and Giorgio Visani. "Metrics for Multi-class Classification: an overview."2020, arXiv:2008.05756. https://doi.org/10.48550/arXiv.2008.05756

[26] Sokolova, Marina, Nathalie Japkowicz, and Stan Szpakowicz. "Beyond Accuracy, F-score and ROC: a family of Discriminant Measures for Performance Evaluation." In Australasian joint conference on artificial intelligence,2006, pp. 1015-1021.

[27] Liu, Yang, Jiahuan Cao, Chongyu Liu, Kai Ding, and Lianwen Jin. "Datasets for Large Language Models: A Comprehensive Survey."2024, arXiv:2402.18041. https://doi.org/10.48550/arXiv.2402.18041

[28] Wang, Jiahao, Bolin Zhang, Qianlong Du, Jiajun Zhang, and Dianhui Chu. "A Survey on Data Selection for LLM Instruction Tuning."2024, arXiv:2402.05123. https://doi.org/10.48550/arXiv.2402.05123

[29] Ketkar, Nikhil, Jojo Moolayil, Nikhil Ketkar, and Jojo Moolayil. "Introduction to Pytorch."2021, Deep learning with python: learn best practices of deep learning models with PyTorch : 27-91.

[30] Salton, Gerard, and Clement T. Yu. "On the Construction of Effective Vocabularies for Information Retrieval." Acm Sigplan Notices 10, 1973, no. 1: 48-60. https://doi.org/10.1145/951787.951766

[31] Ali, Mehdi, Michael Fromm, Klaudia Thellmann, Richard Rutmann, Max Lübbering, Johannes Leveling, Katrin Klug et al. "Tokenizer Choice For LLM Training: Negligible or Crucial?."2023, arXiv:2310.08754. https://doi.org/10.48550/arXiv.2310.08754

[32] Jiang, Ming, Jennifer D'Souza, Sören Auer, and J. Stephen Downie. "Improving Scholarly Knowledge Representation: Evaluating bert-based Models for Scientific Relation Classification." In Digital Libraries at Times of Massive Societal Transition: 22nd International Conference on Asia-Pacific Digital Libraries, ICADL 2020, Kyoto, Japan, Proceedings 22, November 30–December 1,2020, pp. 3-19.