Electronic and Cyber Defens 2025/ Vol. 13/ No. 2/ 73-88

Electronic and Cyber Defens

ISSN: 2322-4347 , E-ISSN: 2980-8979

Cyber Threat Information Extraction using Deep Learning and Knowledge Representation

Samira Hourali[1*], Fatemeh Hourali[2], Atefe Pakzad[3]

[1] Assistant Professor, Department of Computer Engineering, Faculty of Engineering, Kosar University of Bojnord, Bojnord, Iran. Email: ([*]Correspondence: s.hourali@kub.ac.ir)

[2] Assistant Professor, Department of Electrical Engineering, Faculty of Electrical and Computer Engineering, Esfarayen University of Technology, Esfarayen, Iran . Email:hourali.f@gmail.com

[3] Assistant Professor, Department of Computer Engineering, Faculty of Engineering, Kosar University of Bojnord, Bojnord, Iran. Email: atefepakzad@kub.ac.ir

ARTICLE INFO

Article history:

Article Type: Research paper

Received: 16 April 2025

Revised: 25 May 2025

Accepted: 16 June 2025

Available online: 11 June 2025

Keywords:

Information Extraction

Cyber Threats

Entity Relationships

Event Extraction

Deep Learning

Knowledge Representation

ABSTRACT

Cyber security information is rapidly growing on the internet and cyber attacks are increasing daily. Attackers mostly target the military, government, and corporate departments, because these contain sensitive and classified information that requires appropriate defense strategies. Cyber threat information extraction, i.e., extracting entities, relationships between them, and events in cyber texts, is one of the important steps for detecting cyber attacks, harmful events, and mitigating them in real time if they occur. Extracting valuable information from cyber threats can help security professionals to make informed decisions and develop strong defense strategies. It is also a fundamental solution for improving the performance of systems such as text summarization, machine translation, and question-answering. Although information extraction has been an active research topic over the past four decades, its accuracy is still not acceptable and there is no accurate computational model for it. In this paper, first, the entities in the text are extracted with high accuracy using the latest vocabulary embedding method, the Bi-GRU bidirectional recurrent network, the attention mechanism, and the knowledge representation; Then, expressions related to the entities are recognized by calculating the importance and weight of each feature and considering all the necessary criteria in decision-making. The entities relationships were extracted by a graph-based neural network and a heuristic loss function. The KVP deep network based on the attention mechanism has been used for accurate detection and security events prediction which can identify the correlation between two elements that have different positions in the input sequence. Extensive simulations have been carried out to check the performance of the proposed method. According to the simulation results, the proposed method has achieved 89.8% and 93.4% F1 scores on CoNLL-2012 and OSINT datasets, respectively.

Cite this article: Hourali, Samira©, Hourali, Fatemeh©, Pakzad Atefe© (2025). Cyber Threat Information Extraction using Deep Learning and Knowledge Representation. Journal of Electronic and Cyber Defens. 2025; 13l(2):73-88.

DOI: https://dor.isc.ac/dor/20.1001.1.23224347.1404.13.2.7.4

Publisher: Imam Hossein University.

پدافند الکترونیک و سایبری/ سال ۱۴۰۴/ دوره ۱۳/ شماره ۲/ صفحه ۸۸-۷۳

پدافند الکترونیک و سایبری

شاپا الکترونیکی: ۴۳۴۷- ۲۳۲۲۲ شاپا چاپی: ۸۹۷۹-۲۹۸۰

استخراج اطلاعات تهدیدات سایبری با استفاده از یادگیری عمیق و بازنمایی دانش

سمیرا حورعلی[1*]، فاطمه حورعلی[2]، عاطفه پاکزاد[3]

[1] استادیار گروه مهندسی کامپیوتر، دانشکده فنی مهندسی و علوم پایه، دانشگاه کوثر بجنورد، بجنورد، ایران (نویسنده مسئول: s.hourali@kub.ac.ir)

[2] استادیار گروه مهندسی برق، دانشکده مهندسی برق و کامپیوتر، مجتمع آموزش عالی فنی و مهندسی اسفراین، اسفراین، ایران (hourali.f@gmail.com)

[3] استادیار گروه مهندسی کامپیوتر، دانشکده فنی مهندسی و علوم پایه، دانشگاه کوثر بجنورد، بجنورد، ایران (atefepakzad@kub.ac.ir)

چکیده (استایل عنوان چکیده)

اطلاعات مربوط به امنیت سایبری به‌سرعت در اینترنت درحال‌رشد است و حملات سایبری روزبه‌روز در حال افزایش است. مهاجمان بیشتر بخش‌های نظامی، دولتی و شرکتی را هدف قرار می‌دهند، زیرا این بخش‌ها حاوی اطلاعات حساس و طبقه‌بندی‌شده‌ای هستند که به استراتژی‌های دفاعی مناسب نیاز دارد. استخراج اطلاعات تهدیدات سایبری یعنی استخراج نهادها، روابط بین آن‌ها و رویدادهای موجود در متون سایبری، یکی از گام‌های مهم برای تشخیص حملات سایبری، رویدادهای مضر و کاهش آن‌ها در زمان واقعی در صورت وقوع است. استخراج مؤثر اطلاعات ارزشمند از تهدیدات سایبری می‌تواند به متخصصان امنیتی در تصمیم‌گیری آگاهانه و توسعه استراتژی‌های دفاعی قوی کمک کند. همچنین این موضوع یکی از راهکارهای اساسی برای ارتقا عملکرد سیستم‌هایی نظیر خلاصه‌سازی متون، ترجمه ماشینی و پرسش‌وپاسخ نیز است. هرچند طی چهار دهه گذشته استخراج اطلاعات همواره یک موضوع تحقیقاتی فعال بوده است؛ اما هنوز هم دقت آن در حد قابل‌قبول نیست و مدل محاسباتی برای آن وجود ندارد. در این مقاله ابتدا توسط جدیدترین متد تعبیه واژگان، شبکه بازگشتی دوجهته Bi-GRU، مکانیزم توجه و بازنمایش دانش نهادهای موجود در متن با دقت بالا استخراج می‌شوند؛ سپس با محاسبه میزان اهمیت و وزن هر ویژگی و تمام معیارهای لازم در تصمیم‌گیری، عبارات وابسته به نهادها تشخیص داده می‌شود. جهت استخراج دقیق روابط بین نهادها از شبکه عصبی مبتنی بر گراف و تابع هزینه ابتکاری استفاده‌شده است. برای تشخیص و پیش‌بینی دقیق رویدادهای امنیتی از شبکه عمیق KVP مبتنی بر مکانیزم توجه استفاده‌شده است که می‌تواند همبستگی بین دو عنصر که موقعیت‌های متفاوتی در یک دنباله ورودی دارند را شناسایی کند. برای بررسی عملکرد روش پیشنهادی شبیه‌سازی‌های گسترده‌ای صورت گرفته است. طبق نتایج شبیه‌سازی، روش پیشنهادی روی پیکره‌های CoNLL-2012 و OSINT به ترتیب به امتیاز F1 8/89 و ۹۳/۴ درصد دست‌یافته است.

مشخصات مقاله

تاریخچه مقاله:

نوع مقاله: علمی پژوهشی

دریافت: ۱۴۰۴/۰۱/۲۷

بازنگری: ۱۴۰۴/۰۳/۰۴

پذیرش: ۱۴۰۴/۰۳/۲۶

ارائه آنلاین: ۱۴۰۴/۰۴/۲۰

کلید واژه‌ها:

استخراج اطلاعات

تهدیدات سایبری

روابط نهادها

استخراج رویداد

یادگیری عمیق

بازنمایی دانش

استناد: حورعلی، سمیرا، حورعلی، فاطمه، پاکزاد، عاطفه. استخراج اطلاعات تهدیدات سایبری با استفاده از یادگیری عمیق و بازنمایی دانش. پدافند الکترونیک و سایبری. (۱۴۰۴)؛ ۱۳ (۲): ۸۸-۷۳.

https://dor.isc.ac/dor/20.1001.1.23224347.1404.13.2.7.4

ناشر: دانشگاه جامع امام حسین(ع).

۱- مقدمه

امنیت سایبری یک موضوع بسیار حیاتی و مهم است که در دنیای امروزه موردتوجه ویژه‌ای قرارگرفته است. با پیشرفت فناوری و ارتباطات، اطلاعات مربوط به امنیت سایبری به‌سرعت در اینترنت درحال‌رشد است. بنابراین، حملات سایبری نیز به‌شدت افزایش‌یافته است که گاهی عواقب جدی برای افراد و سازمان‌ها در پی‌دارند. مهاجمان بیشتر بخش‌های نظامی، دولتی و شرکتی را هدف قرار می‌دهند، زیرا این بخش‌ها حاوی اطلاعات حساس و طبقه‌بندی‌شده‌ای هستند که به استراتژی‌های دفاعی مناسب نیاز دارد؛ لذا برای دستیابی به امنیت سایبری، باید به مسائل مربوط به امنیت داده‌ها، شبکه‌ها، نرم‌افزارها و... توجه نمود. برای این منظور عمدتاً از دو نوع تکنیک دفاعی استفاده می‌شود. الف) تشخیص حملات سایبری، رویدادهای مضر و کاهش آن‌ها در زمان واقعی در صورت وقوع، ب) پیش‌بینی قبلی حملات سایبری بالقوه با استفاده از اطلاعات تهدید سایبری و ایجاد یک مکانیسم دفاعی قوی. افزایش تعداد حملات سایبری و نقض داده‌ها، امنیت سایبری را به اولویت اصلی دولت‌ها، شرکت‌ها و افراد تبدیل کرده است. تجزیه وتحلیل داده‌های بزرگ به ابزاری ارزشمند در امنیت سایبری تبدیل شده است و راهی برای بررسی مقادیر زیادی از داده‌ها برای مشخص کردن تهدیدات و آسیب-پذیری‌های احتمالی ارائه می‌دهد [۱].

متخصصان امنیتی راه‌حل‌های امنیتی را برای شناسایی و کاهش حملات سایبری توسعه داده و به کار می‌گیرند [۲]. بااین‌حال، این راهکارها همیشه بهترین راه‌حل نیستند، زیرا مهاجمان به آزمایش تاکتیک‌ها، تکنیک‌ها و رویه‌های جدید ادامه می‌دهند. قابل ذکر است که تغییر کل فرآیند حمله برای مهاجمان کاری بسیار دشوار و زمان‌بر است [۱]. تحقیقات نشان می‌دهد که بیشتر گروه‌های مهاجم را می‌توان شناسایی کرد. زیرا مهاجمان تهدیدات پایدار پیشرفته (APTs[1]) و حملات باج افزار را به‌عنوان روش‌های حمله ترکیب می‌کنند. برای پیش‌بینی حملات سایبری احتمالی، می‌توان اطلاعات مفید تهدیدات سایبری و اطلاعات مرتبط با گروه‌های مهاجم را از داده‌های اطلاعات تهدید سایبری (CTI)، استخراج کرد [۳]. CTI یک عنصر حیاتی در عملیات امنیتی مدرن است. داده‌های CTI شامل حجم وسیعی از اطلاعات در مورد تهدیدات سایبری بالقوه، عوامل تهدید، بدافزارها، پروفایل‌ها، تاکتیک‌ها، آسیب‌پذیری‌ها، تکنیک‌ها و رویه‌های مربوط به مهاجمان است. استخراج مؤثر اطلاعات ارزشمند از داده‌های CTI می‌تواند به متخصصان امنیتی در تصمیم‌گیری آگاهانه و توسعه استراتژی‌های دفاعی قوی کمک کند [۴]. بااین‌حال، رشد تصاعدی داده‌های CTI، استخراج اطلاعات مرتبط، به‌موقع و دقیق را چالش‌برانگیز نموده و این مشکل باعث توسعه فناوری‌های مختلف شده است.

فرآیند استخراج اطلاعات (IE[2])، اطلاعات ساختاریافته مفیدی را از داده‌های بدون ساختار یا نیمه ساختاریافته در قالب نهادها، روابط، رویدادها و بسیاری از انواع دیگر استخراج می‌کند. اطلاعات استخراج‌شده از داده‌های بدون ساختار برای آماده‌سازی داده‌ها جهت تجزیه‌وتحلیل استفاده می‌شود. بنابراین، تبدیل کارآمد و دقیق داده‌های بدون ساختار در این فرآیند، تجزیه‌وتحلیل داده‌ها را بهبود می‌بخشد. فرآیند استخراج اطلاعات شامل سه زیرشاخه است، استخراج نهادها و روابط بین آن‌ها، استخراج رویدادها [۵] و استخراج اطلاعات چندگانه. به‌طورکلی، ورودی خط لوله CTI داده‌های خام در مورد امنیت سایبری است، درحالی که خروجی آن دانشی است که می‌تواند در تصمیم‌گیری-های آینده برای دفاع فعالانه از امنیت سایبری، ازجمله استراتژی-هایی برای محدود کردن گستره و پیشگیری از حملات سایبری کمک کند[۶].

به دلیل حجم زیاد و پیچیدگی داده‌های بدون ساختار، استخراج اطلاعات مفید از انواع مختلف داده‌ها یک کار سخت و خسته-کننده است. اما این مسئله یک نکته کلیدی در درک متن است و علاوه بر مباحث امنیتی ذکرشده، در مسائلی مثل خلاصه‌سازی، پاسخ به سؤالات، و ترجمه ماشینی که در آن‌ها درک متن از اهمیت بالایی برخوردار است، نیز کاربرد زیادی دارد. یکی از دلایل پیچیدگی مسئله استخراج اطلاعات درزمینهٔ تشخیص نهادها این است که معمولاً حین نوشتن یک متن، برای اشاره به یک‌نهاد تنها از نام آن استفاده نمی‌کنیم؛ بلکه بسته به شرایط، به دلیل جلوگیری از تکرار و بیان اطلاعات بیشتری در مورد آن نهاد یا تأکید بر یک ویژگی خاص، از عبارات توصیفی مختلف نظیر عبارات اسمی و گاهی ضمایر برای اشاره به آن نهاد استفاده می کنیم. حتی ممکن است شخص یا اشیا را با ویژگی‌ها و یا کاربردشان توصیف کنیم. این عبارات برای اشاره به یک‌نهاد استفاده می‌شوند.

دلیل دیگر متنوع بودن علت وابستگی اطلاعات است، به‌بیان‌دیگر ممکن است هرکدام از عبارات اسمی موجود در یک سند ۹ با قوانین متفاوتی نظیر نزدیکی فاصله، تطبیق کامل رشته‌ای، تطبیق کلمه سر، توازی نحوی، شباهت معنایی با نهاد خود مرتبط باشند. همچنین هر نوع از وابستگی یا رویداد نیاز به دانش و اطلاعات مخصوص به خود برای حل شدن دارد. درواقع، عدم درنظرگرفتن تمام معیارهای وابستگی عبارات مختلف به نهادها و عدم استفاده از بازنمایش منابع دانش مختلف باعث شده است، روش‌های موجود دقت پایین‌تری نسبت به روش پیشنهادی داشته باشند. علاوه بر این، عدم دسترسی اکثریت سیستم‌های استخراج اطلاعات کنونی به دانش پیچیده موردنیاز برای تحلیل درست روابط بین نهادها، باعث عملکرد ضعیف آن‌ها شده است،

[1] Advanced Persistent Threats

[2] Information Extraction

زیرا بسیاری از موارد استخراج اطلاعات تنها با کمک دانش جهان قابل‌حل هستند.

رویکرد این مقاله برای حل مسئله استخراج مؤثر اطلاعات از داده‌های CTI، استفاده از یادگیری عمیق و بازنمایش منابع دانش مختلف جهت استخراج دقیق نهادها، روابط و رویدادهای سایبری است. در مرحله تشخیص نهادها، عبارات وابسته و اشاره‌کننده به آن‌ها را نیز استخراج کردیم که باعث بالا رفتن دقت استخراج اطلاعات تا حد قابل‌قبولی می‌شود. توسط تکنیک تصمیم‌گیری چندمعیاره ارائه‌شده نیز، تمام معیارها و ویژگی‌ها و وزن یا اهمیت آن‌ها نیز در شناسایی عبارات وابسته به یک‌نهاد در نظر گرفته‌شده است که با دقت بالاتری نسبت به روش‌های موجود استخراج‌شده‌اند. درنهایت از منابع دانش مختلف و شبکه عصبی مبتنی بر گراف و KVP[1] برای استخراج دقیق روابط بین نهادها و رویدادها از داده‌های امنیتی استفاده‌شده است. بازنمایش دانش، حوزه‌ای از هوش مصنوعی است که به مناسب‌ترین اشکال و روش‌های ذخیره دانش در سیستم‌های کامپیوتری می‌پردازد. هدف از انجام بازنمایش دانش، غنی‌سازی دانش سیستم و سپس انتخاب مناسب‌ترین دانش برای حل هر یک از موارد استخراج اطلاعات است. زمانی که رایانه‌ها بخواهند، در کنار انسان یا به‌جای انسان به استدلال بپردازند، نقش بازنمایش دانش به سبب دشواری‌های ناشی از مقیاس‌پذیری، حیاتی‌تر و اجتناب‌ناپذیر می‌گردد. دلیل اینکه درروش پیشنهادی از بازنمایش دانش با استفاده از منابع متعدد سایبری، لغوی، نحوی، معنایی و جهانی استفاده کردیم، این است که بیشتر مشکلات استخراج اطلاعات فقط با استفاده از این منابع دانش قابل‌حل است. در ادامه این مقاله، در بخش دوم روش‌های موجود در زمینهٔ استخراج اطلاعات را بررسی کرده و در بخش سوم مدل در نظر گرفته‌شده برای حل مسئله را تشریح می‌کنیم. در بخش چهارم نتایج شبیه‌سازی را ارائه می‌کنیم. برای ارزیابی اثربخشی این تکنیک، آزمایش‌هایی با استفاده از مجموعه داده‌های CTI و داده‌های دیگر انجام‌شده و مدل پیشنهادی با مدل‌های پیشرفته فعلی مقایسه شده است. همچنین علت برتری مدل پیشنهادی در مقایسه با مدل‌های موجود نیز شرح داده‌شده است.

۲- پیشینه تحقیق

علیرغم تلاش‌های زیادی که طی چهار دهه اخیر روی مسئله استخراج اطلاعات انجام‌شده است، کارایی روش‌های حل این مسئله هنوز به حد قابل قبولی نرسیده است به‌گونه‌ای که این مسئله هنوز یک موضوع تحقیقاتی فعال است. یکی از دلایل آن این است که هیچ مدل محاسباتی دقیقی برای این مسئله وجود ندارد. رویکردهای جدید استخراج اطلاعات [۷] و [۸] عموماً از روش‌های مبتنی بر یادگیری عمیق و یادگیری تقویتی استفاده می‌کنند اما نتوانسته‌اند مدل محاسباتی دقیقی برای حل این مسئله ارائه کنند. در حالت کلی روش‌های استخراج اطلاعات از داده‌های سایبری به سه دسته مبتنی بر قانون یا قاعده، مبتنی بر یادگیری ماشین و مبتنی بر یادگیری عمیق تقسیم می‌شوند. در روش‌های مبتنی بر قاعده [۹-۱۲] موارد وابسته در متن توسط مجموعه‌ای از قواعد دست‌نویس که توسط افراد خبره نوشته‌شده‌اند، استخراج می‌شوند. این روش‌ها با خلاصه کردن قوانین، اطلاعات را بر اساس الگوها استخراج می‌کنند. از مزایای این روش می‌توان به‌دقت بالا روی مجموعه داده‌های کوچک و سادگی طراحی اشاره کرد. اما قابلیت انعطاف این روش پایین است و باید برای هر زبان طبیعی، سیستم از ابتدا توسط افراد خبره طراحی شود و نیازمند زمان و نیروی انسانی زیادی هستند. این روش‌ها با مجموعه داده‌های بزرگ سازگاری ندارند.

روش‌های مبتنی بر یادگیری ماشین در حیطه استخراج اطلاعات از داده‌های متنی و سایبری [۱۳-۱۷] به سه دسته با ناظر، بدون ناظر و یادگیری تقویتی [۱۸-۲۱] تقسیم می‌شوند. در روش‌های آماری با ناظر، داده‌های آموزشی به‌صورت دستی یا خودکار برچسب‌گذاری شده‌اند. این داده‌ها برای یادگیری سیستم مورداستفاده قرار می‌گیرند. از طرف دیگر، در روش‌های بدون ناظر نیازی به داده‌های آموزشی نیست (یا داده‌ها آموزشی بسیار کمی موردنیاز است) اما دقت این روش هنوز برای حل مسئله استخراج اطلاعات پایین است. این روش‌ها از مدل‌هایی مانند HMM[2]، SVM[3] و CRF[4] استفاده می‌کنند. اگرچه این روش‌ها نتایج قابل‌توجهی را در مقایسه با روش‌های قبلی بهبود بخشیده است، اما نیاز به حاشیه‌نویسی دستی زیادی توسط افراد بادانش حرفه‌ای در این زمینه‌دارند، درنتیجه هزینه کار و زمان در این نوع روش‌ها بسیار بالاست.

در روش‌های مبتنی بر یادگیری تقویتی عبارات اسمی و نهادها به‌عنوان حالت‌ها[5] و اتصالات آن‌ها به‌عنوان عمل‌ها[6] در نظر گرفته‌شده‌اند؛ در این روش‌ها، طبق سیاست یادگیری تقویتی و تعریف توابع پاداش[7] و خطا[8] عبارات وابسته تشخیص داده می‌شود. روش‌های مبتنی بر یادگیری ماشین نسبت به روش‌های مبتنی بر قوانین بهتر عمل می‌کنند، اما این روش‌ها نیاز به برچسب‌گذاری دستی توسط افراد بادانش حرفه‌ای زیادی دارد، بنابراین زمان‌بر هستند و هزینه زیادی صرف پیاده‌سازی آن‌ها می‌شود.

جدیدترین روش‌های ارائه‌شده در حیطه استخراج اطلاعات، روش‌های مبتنی بر یادگیری عمیق هستند [۲۲]. این روش‌ها برای پردازش داده‌های بزرگ قابل‌استفاده هستند و به‌طور خودکار ویژگی‌های جمله را پیچیده یاد می‌گیرند. در چند سال گذشته،

[1] Key-Value-Predict Attention Networks

[2] Hidden Markov Model
[3] Support Vector Machines
[4] Conditional Random Fields
[5] States
[6] Actions
[7] Reward
[8] Punishment

پردازش زبان طبیعی (NLP) و تکنیک‌های یادگیری عمیق، پتانسیل فوق‌العاده‌ای را در حل مشکلات مربوط به استخراج روابط نهادها از داده‌های متنی بدون ساختار نشان داده‌اند [۲۳]. این تکنیک‌ها می‌توانند بدون تکیه‌بر قوانین از پیش تعریف‌شده، الگوها و روابط را از داده‌ها بیاموزند و آن‌ها را برای مجموعه داده‌های پیچیده و پویا مانند داده‌های CTI مناسب کنند. بااین‌حال، تکنیک‌های موجود برای استخراج نهاد و رابطه از داده‌های CTI از چندین محدودیت ازجمله دقت کم، عدم مقیاس‌پذیری و تعمیم‌پذیری رنج می‌برند.

در این روش‌ها عبارات اسمی موجود در متن توسط شناسایی خودکار ویژگی‌ها و ساختار یادگیری عمیق استخراج می‌شوند، سپس با تعریف توابع هزینه مختلف سیستم در جهت شناسایی درست عبارات وابسته و استخراج اطلاعات، آموزش می‌بینند. این روش‌ها نسبت به روش‌های مبتنی بر یادگیری ماشین روی‌داده‌های حجیم‌تر بهتر جواب می‌دهند، اما زمان بیشتری برای یادگیری نیاز دارند.

همان‌گونه که در بخش مقدمه نیز ذکر شد، فرآیند استخراج اطلاعات سه زیرشاخه است. تحت ساختار یادگیری عمیق نیز روش‌هایی برای استخراج روابط بین نهادها [۲۴-۳۱] و استخراج رویدادها [۳۲- ۳۶] و استخراج اطلاعات چندگانه [۳۷-۳۹] ارائه‌شده است. اما، توسعه سریع اینترنت منجر به حجم عظیمی از داده‌ها، ازجمله متن، صدا، تصاویر، ویدئو و غیره شده است. بنابراین، محققان شروع به پژوهش در زمینه چگونگی استخراج اطلاعات موردنیاز از داده‌های چندوجهی کردند. تحقیقات نشان داده است که افزودن اطلاعات بصری می‌تواند نقش مهمی در فرآیند استخراج اطلاعات ایفا کند. تابه‌حال، اکثر روش‌های استخراج اطلاعات به‌صورت جداگانه بر استخراج نهادها و روابط بین آن‌ها یا استخراج رویدادها تمرکز کرده‌اند. بنابراین، درروش پیشنهادی سعی شده است تمام این موارد پوشش داده شود. استخراج نهادها، عبارات وابسته به آن‌ها و روابط بینشان، همچنین استخراج رویدادها ازجمله مواردی است که در این مقاله به‌صورت یکجا بررسی و حل‌شده است. همچنین با توجه به ساختار شبکه‌های عمیق مبتنی بر گراف و KVP استفاده‌شده امکان استفاده از آن در زمینه استخراج اطلاعات چندوجهی نظیر تصویر و صوت نیز وجود دارد که باعث شده است روش پیشنهادی نسبت به روش‌های موجود تمام جنبه‌های استخراج اطلاعات را پوشش دهد. علاوه بر این، می‌توان از طریق تجهیز روش پیشنهادی به منابع دانش مختلف نظیر پزشکی، زمین‌شناسی و ... از آن در این زمینه‌ها نیز استفاده کرد.

روش پیشنهادی برای حل مسئله استخراج اطلاعات و استخراج اطلاعات تهدیدات سایبری در این مقاله، استفاده از دانش تصمیم‌گیری چندمعیاره و منابع دانش مختلف تحت ساختار یادگیری عمیق است. در فاز پیش‌پردازش معیارها یا ویژگی‌های برای عبارات اسمی و نهادها توسط ساختار یادگیری عمیق و به‌صورت خودکار استخراج‌شده است. این ویژگی‌ها نقش مهمی در درک مفهومی این عبارات توسط شبکه بر عهده‌دارند. درنظرگرفتن اطلاعات سطح نهادها تأثیر قابل‌قبولی در شناسایی و استخراج آن‌ها دارد. همچنین درنظرگرفتن تمام ویژگی‌ها و وزن یا میزان اهمیت هر یک از آن‌ها برای یافتن نهادها، عبارات وابسته به نهادها و ارتباط بین آن‌ها، تأثیر زیادی در بهبود روند استخراج اطلاعات دارد. علاوه بر این استفاده از شبکه‌های عصبی عمیق، تابع هزینه ابتکاری، شبکه عصبی مبتنی بر گراف و بازنمایش دانش نیز تأثیر قابل‌قبولی در استخراج روابط بین نهادها و رویدادها دارد.

۳- روش پیشنهادی

در این بخش ابتدا تعبیه واژگان متن ورودی توسط روش ROBERTA استخراج‌شده و به‌عنوان ورودی به شبکه بازگشتی دوجهته GRU داده‌شده و با استفاده از منابع دانش و مکانیزم توجه، نهادها و بازنمایش عبارات اسمی با طول‌های مختلف از متن ورودی استخراج می‌شوند. سپس طبق روش تصمیم‌گیری چندمعیاره ارائه‌شده عبارات اسمی استخراج‌شده برای نهادها رتبه‌بندی و اولویت‌بندی می‌شوند. سپس توسط شبکه عصبی مبتنی بر گراف رابطه بین نهادها و عبارات اشاره‌کننده به آن‌ها استخراج می‌شود. درنهایت توسط شبکه KVP رویدادهای موجود در متن استخراج می‌شود. بلوک دیاگرام روش پیشنهادی در شکل (۱) نمایش داده‌شده است.

۳-۱- استخراج نهادها و عبارات اسمی

با فرض اینکه سند D شامل T کلمه باشد، تعبیه واژگان استخراج‌شده توسط ROBERTA را به‌صورت $\{x_1, \dots, x_T\}$ در نظر می‌گیریم. برای محاسبه بازنمایش (نمایش برداری) دنباله کلمات[1]، تعبیه واژگان را طبق روابط (۵-۱) به شبکه بازگشتی Bi–GRU تزریق می‌کنیم. همان‌طور که گفته شد، هر واحد GRU در مکان t شامل دو گیت یا دروازه به‌روزرسانی (z_t) و بازنشانی (r_t) است.

$$r_t = \sigma(W_{rx}x_t + W_{rh}h_{t-1}) \quad (۱)$$

$$z_t = \sigma(W_{zx}x_t + W_{zh}h_{t-1}) \quad (۲)$$

$$\tilde{h}_t = \tanh(W_{hx}x_t + W_{hh}(r_t \odot h_{t-1})) \quad (۳)$$

$$h_t = z_t \odot h_{t-1} + (1 - z_t) \odot \tilde{h}_t \quad (۴)$$

$$x_t^* = [h_{t+1}, h_{t-1}] \quad (۵)$$

در رابطه (۱)، Wها پارامترهای مدل برای هر واحد هستند. $\tilde{h}_t$ حالت پنهان کاندید برای h_t است. σ تابع منطقی سیگموئید است، که به‌صورت $\sigma(x) = 1/(1 + e^{-x})$ تعریف می‌شود. $\sigma \epsilon [-1,1]$ جهت هر GRU را نشان می‌دهد و $\odot$ ضرب عنصر به

[1] Span representation

عنصر دو بردار است. x_t^* خروجی الحاق شده‌ی Bi-GRU است. برای هر جمله از یک GRU مستقل استفاده می‌شود.

کلمه سر[1] (کلمه مهم و کلیدی) عبارات اسمی را می‌توان با کمک تجزیه‌گر نحوی[2] به دست آورد. هرچند که با تعمیم آن با کمک ساختار سلسله‌مراتب والد/فرزند در WordNet می‌توان آن‌ها به دسته‌های کلی‌تر تبدیل کرد. مثلاً در نام بری famous ransomware problems کلمه سر ransomware است که پس از تعمیم به دسته معنایی software تبدیل می‌شود. به‌جای استفاده از تجزیه‌گر نحوی، مدل پیشنهادی با استفاده از مکانیزم توجه[3] [۴۰] وظیفه یافتن کلمات سر را طبق روابط (۸-۶) می‌آموزد، که در آن FFNN شبکه عصبی روبه‌جلو است.

$$\alpha_t = w_\alpha . FFNN_\alpha(x_t^*) \quad (۶)$$

$$a_{i,t} = \frac{exp(\alpha_t)}{\sum_{k=START(i)}^{END(i)} exp(\alpha_k)} \quad (۷)$$

$$\hat{x}_i = \sum_{t=START(i)}^{END(i)} a_{i,t} . x_t \quad (۸)$$

درنهایت برای دنباله کلمات i بازنمایش رابطه (۹) استخراج می‌شود که در آن $x^*_{START(i)}$ و $x^*_{END(i)}$ ابتدا و انتهای (محتوای چپ و راست) تعبیه دنباله کلمات، $\emptyset(i)$ ویژگی‌های اضافی (بازنمایش اطلاعاتی که در متن نیست) و $\hat{x}_i$ بازنمایش مبتنی بر مکانیزم توجه است.

$$g_i = [x^*_{START(i)}, x^*_{END(i)}, \hat{x}_i, \emptyset(i)] \quad (۹)$$

برای درنظرگرفتن اطلاعات جهانی و اطلاعات نهادها، بازنمایش مورد انتظار g'_i برای دنباله کلمات i توسط روابط (۱۰-۱۲) محاسبه می‌شود.

$$g'_i = \sum_{j=1}^{i} Q(i \in E_j) . e_j^{(i)} \quad (۱۰)$$

$$e_i(t) = \sum_{j=1}^{t} Q(j \in E_i) . g_j \quad (۱۱)$$

$$Q(i \in E_j) = \begin{cases} \sum_{k=j}^{i=1} p(y_i = k) . Q(k \in E_j) & j < i \\ p(y_i = \epsilon) & j = i \\ 0 & j > i \end{cases} \quad (۱۲)$$

در رابطه (۱۲) $Q(i \in E_j)$ احتمال این است که عبارت اسمی i متناظر با نهاد j باشد (اگر m عبارت اسمی در متن موردنظر وجود داشته باشند، حداکثر m نهاد نیز وجود دارد). $e_i(t)$ بازنمایش یک‌نهاد در زمان t است.

برای تشخیص دقیق عبارات اسمی و اطلاعات وابسته به یک‌نهاد از منابع دانش مختلف استفاده می‌کنیم. برای استفاده از منابع دانش، آن‌ها را به یک فرمت یکسان تبدیل کردیم. فرمت در نظر گرفته‌شده به‌صورت سه‌تایی "ابتدا[4]، انتها[5]، رابطه[6]" است. بعد از تبدیل تمام اطلاعات به این فرمت، برای ایجاد منبع دانش کلی ($\mathcal{G}$)، تمامی آن‌ها را باهم ترکیب کردیم. منابع داتش استفاده‌شده، در بخش نتایج شبیه‌سازی (بخش ۴) شرح داده‌شده است. برای هر دنباله مدنظر، جهت عبارت اسمی بودن یا وابسته بودن با عبارت دیگر، دانش‌های مختلفی از منبع دانش و با روش‌های مختلف می‌توان استخراج کرد. جهت سادگی و تعمیم در مدل پیشنهادی از تطابق ساختار رشته‌ای[7] برای استخراج دانش مرتبط استفاده نمودیم. به‌طور خاص، برای هر سه‌تایی $t \in \mathcal{G}$ که ابتدا و انتهای آن فهرستی از کلمات است، اگر ابتدای آن مشابه با رشته‌ای در دنباله کلمات s باشد، آن را یک سه‌تایی مرتبط می‌دانیم. بنابراین، اطلاعات t را با میانگین‌گیری تعبیه تمام کلمات انتهای آن، رمزنگاری می‌کنیم. برای مثال، اگر s " باج افزار " باشد و سه‌تایی ("باج افزار "، استفاده می‌کند، " تکنیک هجومی") با جستجو در منبع دانش یافت شود، این رابطه را توسط میانگین تعبیه کلمات "تکنیک" و "هجومی" نمایش داده می‌دهیم. برای هر جفت دنباله کلمات i و j، بازنمایش آن‌ها به هم الحاق شده و بازنمایش کلی g'_i, g'_j حاصل می‌شود، تا دانش مناسب انتخاب شود. برای دنباله کلمات i مجموعه دانش بازیابی شده را با $\mathcal{K}_{s_i}$ نشان می‌دهیم، که حاوی m_s بازنمایش دانش مرتبط به‌صورت $k_{1,s}, k_{2,s}, \dots, k_{m_s,s}$ است. برای ادغام دانش فوق‌الذکر در مدل پیشنهادی، با این چالش روبرو هستیم که تعداد زیادی تعبیه وجود دارد، درحالی‌که بیشتر آن‌ها در زمینه‌های خاص بی‌فایده هستند. برای حل این مشکل، برای دنباله کلمات i، وزن هر $k_z \in \mathcal{K}_{s_i}$ توسط روابط (۱۳ و ۱۴) محاسبه می‌شود. دانش کلی توسط رابطه (۱۵) بدست می‌آید.

$$w_z = \frac{e^{\beta_{k_z}}}{\sum_{k_j \in \mathcal{K}_{s_i}} e^{\beta_{k_j}}} \quad (۱۳)$$

$$\beta_k = NN_\beta([g'_i, g'_j, k]) \quad (۱۴)$$

$$o_{g'_i} = \sum_{k_z \in \mathcal{K}_{s_i}} w_z . k_z \quad (۱۵)$$

$$F(s_i, s_j) = f_m(s_i) + f_m(s_j) + f_c(s_i, s_j) \quad (۱۶)$$

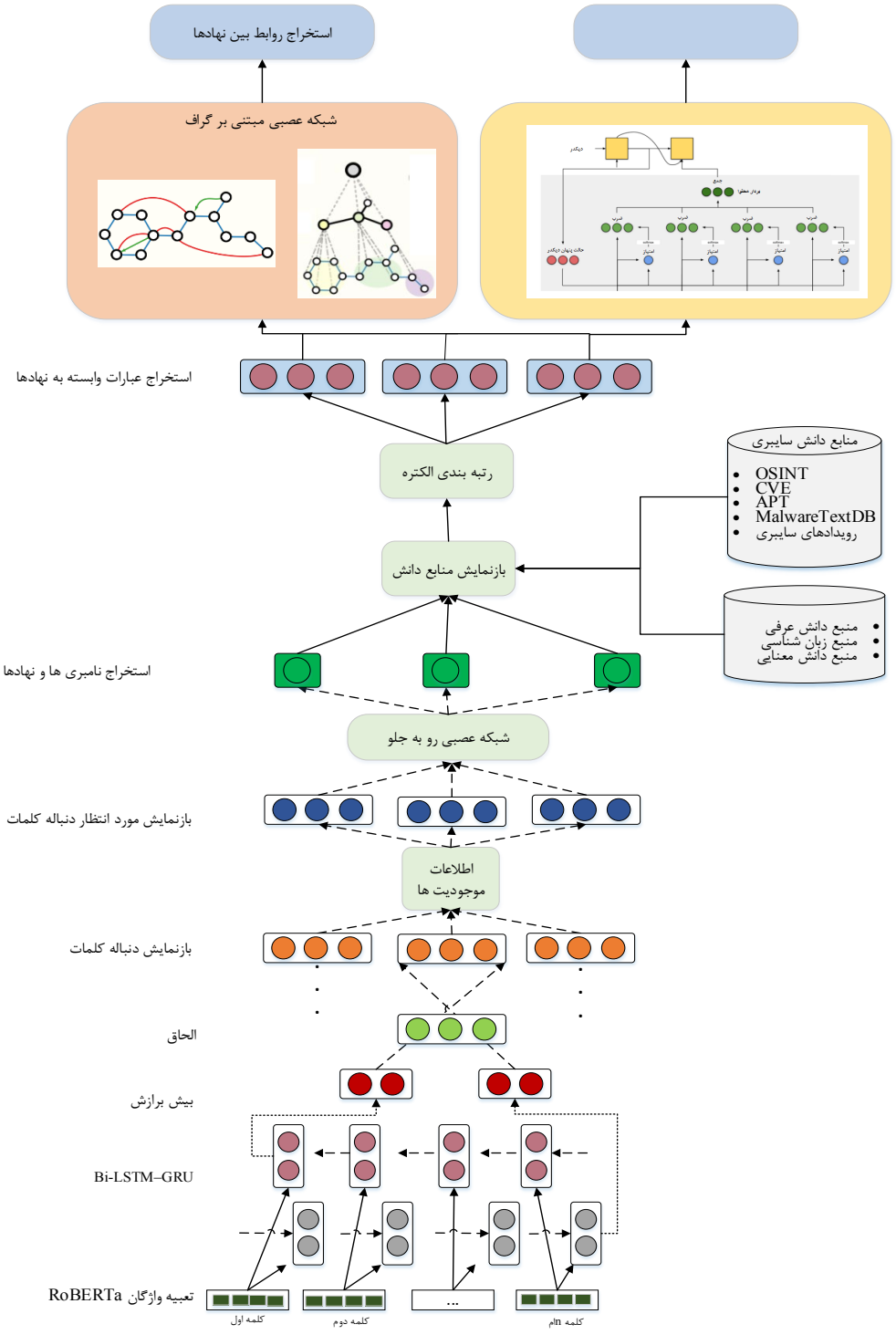
$$f_m(s_i) = NN_m([g'_i, o_{g'_i}]) \quad (۱۷)$$

$$f_c(s_i, s_j) = NN_c\left(\left[g'_i, o_{g'_i}, g'_j, o_{g'_j}, g'_i \odot g'_j, o_{g'_i} \odot o_{g'_j}\right]\right) \quad (۱۸)$$

۲-۳- تشخیص عبارات وابسته به نهادها

پس از استخراج عبارات اسمی و ویژگی‌های آن‌ها، ماتریس تصمیم زیر برای عبارات کاندید شده جهت بررسی وابسته بودن یا نبودن با نهاد مدنظر تشکیل می‌شود.

[1] Head
[2] Syntactic Parses
[3] Attention Mechanism
[4] Head
[5] Tail
[6] Relation
[7] String Match



شکل (۱): ساختار روش پیشنهادی

$$M = \begin{matrix} \\ g'_1 \\ g'_2 \\ \dots \\ g'_m \end{matrix} \begin{matrix} \begin{matrix} f_1 & f_2 & \dots & f_n \end{matrix} \\ \begin{bmatrix} r_{11} & r_{12} & \dots & r_{1n} \\ r_{21} & r_{22} & \dots & r_{2n} \\ \dots & \dots & \dots & \dots \\ r_{m1} & r_{m2} & \dots & r_{mn} \end{bmatrix} \end{matrix} \quad (۱۹)$$

در ماتریس رابطه (۱۹) g'_i بازنمایش عبارت اسمی iام، f_j ویژگی jام و r_{ij} مقدار استخراج‌شده گام قبل برای ویژگی jام عبارت iام‌است. وزن یا میزان اهمیت هر یک از ویژگی‌ها توسط روش بردار ویژه و به‌صورت زیر محاسبه می‌شود.

محاسبه وزن ویژگی‌ها: در این روش w_j ها به‌گونه‌ای تعیین می‌شوند که روابط زیر صادق باشند.

$$\begin{aligned} r_{11}w_1 + r_{12}w_2 + \dots + \mathrm{r}w_n &= \lambda.w_1 \\ r_{21}w_1 + r_{22}w_2 + \dots + \mathrm{r}w_n &= \lambda.w_2 \\ &\vdots \\ r_{n1}w_1 + r_{n2}w_2 + \dots + r_{nn}w_n &= \lambda.w_n \end{aligned} \quad (۲۰)$$

وزن شاخص‌ها (ویژگی‌ها) طبق رابطه (۲۱) محاسبه می‌شود.

$$w_j = \frac{1}{\lambda}\sum_{j=1}^{n} r_{ij} w_i \qquad j = 1,2,\dots,n \tag{۲۱}$$

دستگاه معادلات فوق را به‌صورت رابطه (۲۲) می‌توان نشان داد.

$$M.w = \lambda.w \tag{۲۲}$$

در رابطه (۲۲) M ماتریس تصمیم، w بردار ویژه و λ مقدار ویژه برای ماتریس A است. مراحل زیر جهت محاسبه w_j درروش بردار ویژه به‌صورت زیر است.

۱- ماتریس M تشکیل داده می‌شود.

۲- ماتریس (M- λI) تشکیل داده می‌شود.

۳- دترمینان ماتریس (M- λI) را محاسبه کرده، مساوی صفر قرار داده و مقادیر λ محاسبه می‌شود.

بزرگ‌ترین λ را λ_{max} می‌نامیم و آن را در رابطه (۲۳) قرار داده و مقادیر w_j ها محاسبه می‌شود.

$$(\text{M-}\lambda_{max} I).\text{w}=0 \tag{۲۳}$$

پس از محاسبه وزن ویژگی‌ها برای عبارات کاندید، طبق رابطه (۲۴) وزن عبارات جهت بررسی هم مرجع بودن یا نبودن با نهاد موردنظر به‌روزرسانی می‌شود.

$$w'_j = \frac{r_j w_j}{\sum_{j=1}^{n} r_j w_j} \tag{۲۴}$$

رتبه‌بندی عبارات کاندید: پس از تشکیل ماتریس تصمیم و محاسبه وزن و مقدار ویژگی‌ها، عبارات کاندید باید برای نهاد موردنظر رتبه‌بندی شوند. جهت رتبه‌بندی عبارات از روش تصمیم‌گیری چندمعیاره ELECTRE [۴۱] و به‌صورت زیر استفاده می‌شود.

ابتدا ماتریس تصمیم بی مقیاس می‌شود. برای تبدیل ویژگی‌های چندبعدی به ویژگی‌های بدون بعد و ازبین‌بردن تفاوت مقیاسی داده‌های تصمیم‌گیری، از نرم اقلیدسی رابطه (۲۵)، استفاده می‌کنیم.

$$n_{\text{ij}} = \frac{\text{r}_{\text{ij}}}{\sqrt{\sum_{\text{i=1}}^{\text{m}} r^2{}_{\text{ij}}}} \tag{۲۵}$$

سپس، ماتریس تصمیم وزین از ضرب ماتریس تصمیم بی مقیاس شده در وزن معیارها که توسط روش بردار ویژه محاسبه شد، به دست می‌آید.

$$v_{ij} = w'_j * n_{ij} \quad i = 1,2,\dots,m;\ \ j = 1,2,\dots,n \tag{۲۶}$$

برای هر زوج عبارت اسمی k و e مجموعه معیارها به دو زیرمجموعه هماهنگ و ناهماهنگ تقسیم می‌شوند. مجموعه هماهنگ (S_{ke}) مجموعه‌ای از معیارهایی است که در آن عبارت k به عبارت e ترجیح دارد و مجموعه مکمل آن مجموعه مخالف (D_{ke}) است، مجموعه معیارهای هماهنگ برای معیارهای مثبت و منفی به ترتیب به‌صورت زیر تعریف می‌شوند.

$$\begin{aligned} S_{ke} &= \{j | v_{kj} \geq v_{ej}\} \\ S_{ke} &= \{j | v_{kj} \leq v_{ej}\} \end{aligned} \tag{۲۷}$$

مجموعه معیارهای ناهماهنگ برای معیارهای مثبت و منفی به ترتیب به‌صورت زیر تعریف می‌شود.

$$\begin{aligned} D_{ke} &= \{j | v_{kj} < v_{ej}\} = J - S_{ke} \\ D_{ke} &= \{j | v_{kj} > v_{ej}\} = J - S_{ke} \end{aligned} \tag{۲۸}$$

در این مرحله ماتریس هماهنگی تشکیل می‌شود. این ماتریس یک ماتریس مربعی است که بعد آن تعداد عبارات اسمی است. هر یک از درایه‌های این ماتریس، شاخص توافق بین دو عبارت نامیده می‌شود. مقدار این شاخص، از جمع وزن معیارهایی که در مجموعه هماهنگ وجود دارند، به دست می‌آید. به‌بیان‌دیگر برای محاسبه معیار هماهنگی (C_{ke}) طبق رابطه (۲۹) باید عبارت k و عبارت e مقایسه شده و مقدار آن از جمع وزن معیارهایی که عبارت k نسبت به عبارت e ترجیح دارد، به دست می‌آید. مقدار این معیار بین ۰ و ۱ است، هرچه این مقدار به ۱ نزدیک باشد، به ارجحیت بیشتر عبارت k نسبت به عبارت e اشاره دارد.

$$C_{ke} = \sum_{j \in S_{ke}} w'_j \tag{۲۹}$$

بعد از تشکیل ماتریس هماهنگی باید ماتریس ناهماهنگی ایجاد شود. این ماتریس یک ماتریس مربعی است که بعد آن تعداد عبارات اسمی است. هر یک از درایه‌های این ماتریس، شاخص عدم توافق بین دو عبارت نامیده می‌شود. مقدار این شاخص از رابطه زیر به دست می‌آید.

$$d_{ke} = \frac{\max\limits_{j \in D_{ke}} |v_{kj} - v_{ej}|}{\max\limits_{j \in J} |v_{kj} - v_{ej}|} \tag{۳۰}$$

در مرحله چهارم نحوه محاسبه شاخص هماهنگی (C_{ke}) بیان شد، در این مرحله یک مقدار معین برای این شاخص توافق مشخص می‌شود که آن را آستانه هماهنگی می‌نامند و با $\bar{\text{C}}$ نشان داده می‌شود. این مقدار از میانگین‌گیری درایه‌های ماتریس هماهنگی به دست می‌آید. به زبان ریاضی مقدار آستانه هماهنگی از رابطه زیر محاسبه می‌شود.

$$\bar{C} = \sum_{k=1}^{m}\sum_{e=1}^{m} \frac{C_{ke}}{m(m-1)} \tag{۳۱}$$

ماتریس تسلط هماهنگی (F) باتوجه‌به مقدار آستانه موافقت تشکیل می‌شود. اگر C_{ke} بزرگ‌تر از $\bar{\text{C}}$ باشد، برتری عبارت k بر عبارت e قابل‌قبول است در غیر این صورت عبارت k بر عبارت e برتری ندارد لذا درایه‌های این ماتریس تسلط از رابطه زیر تعیین می‌شود.

$$f_{ke} = \begin{cases} 1 & C_{ke} \geq \bar{C} \\ 0 & C_{ke} < \bar{C} \end{cases} \tag{۳۲}$$

ماتریس تسلط ناهماهنگی (G) مانند F تشکیل می‌شود. بدین منظور ابتدا باید مقدار آستانه ناهماهنگی ($\bar{\text{d}}$) از میانگین‌گیری درایه‌های ماتریس ناهماهنگی محاسبه شود. به زبان ریاضی مقدار آستانه ناهماهنگی از رابطه زیر محاسبه می‌شود.

$$\bar{d} = \sum_{k=1}^{m}\sum_{e=1}^{m} \frac{d_{ke}}{m(m-1)} \tag{۳۳}$$

همان‌گونه که بیان شد، مقدار شاخص ناهماهنگی(d_{ke}) هر چه کمتر باشد بهتر است. زیرا میزان ناهماهنگی برتری عبارت k بر عبارت e را بیان می‌کند. چنانچه d_{ke} از ($\bar{d}$) بزرگ‌تر باشد، میزان مخالفت زیاد بوده و نمی‌توان از آن صرف‌نظر کرد بنابراین درایه‌های ماتریس G به‌صورت زیر محاسبه می‌شود.

$$g_{ke} = \begin{cases} 1 & d_{ke} > \bar{d} \\ 0 & d_{ke} \le \bar{d} \end{cases} \quad (۳۴)$$

هر عضو این ماتریس نشانگر رابطه تسلط مابین عبارات اسمی است. درنهایت ماتریس تسلط نهایی (H) از ضرب تک‌تک درایه‌های ماتریس F در G حاصل می‌شود.

ماتریس H ترجیحات جزئی عبارات اسمی را بیان می‌کند. مؤلفه h_{ke} در صورتی برابر یک می‌شود که هر دو مؤلفه متناظری که در هم ضرب می‌شوند برابر ۱ باشند. برتری عبارت k بر عبارت e در هر دو حالت موافق و مخالف در صورتی قابل‌قبول است که برتری آن از حد آستانه هماهنگی بیشتر بوده و مخالفت آن نیز از حد آستانه ناهماهنگی کمتر باشد. بدین ترتیب عبارات کاندید جهت وابستگی با نهاد موردنظر رتبه‌بندی و طبق حد آستانه استخراج می‌شوند.

$$h_{ke} = f_{ke} \cdot g_{ke} \quad (۳۵)$$

۳-۳- تشخیص روابط بین نهادها

پس از استخراج نهادها و عبارات وابسته به هر نهاد باید روابط بین نهادها شناسایی شود. برای حل این مسئله از شبکه عصبی عمیق مبتنی بر گراف استفاده نمودیم. براین‌اساس، بردار ویژگی استخراج‌شده طبق رابطه (۳۶) به‌صورت ورودی به واحد فعال‌سازی داده می‌شود.

$$h_v^0 = X_v(h_{ke}) \quad (۳۶)$$

سپس برای هر لایه رابطه (۳۷) اعمال می‌شود.

$$h_v^k = \sigma(W_k \sum \frac{h_v^{k-1}}{N(v)} + B_k h_v^{k-1}) \; where \;\; k = 1, \dots, k-1 \quad (۳۷)$$

دو مرحله برای این معادله وجود دارد:

مرحله اول میانگین‌گیری تمام همسایگان گره v است.

$$W_k \sum \frac{h_v^{k-1}}{N(v)} \quad (۳۸)$$

مرحله دوم، تعبیه لایه قبلی گره v است که با بایاس B_k ضرب شده است، که یک ماتریس وزن قابل‌آموزش است و اساساً یک فعال‌سازی خود حلقه برای گره v است.

$$B_k h_v^{k-1} \quad (۳۹)$$

σ فعال‌ساز غیرخطی است که روی دو قسمت انجام می‌شود.

درنهایت، در لایه آخر رابطه (۴۰) اعمال می‌شود که درواقع تعبیه بعد از K لایه از تجمع همسایگان است.

$$Z_v = h_v^k \quad (۴۰)$$

درنهایت از تابع هزینه ابتکاری زیر روی تعبیه‌ها و برای آموزش مدل استفاده شد.

$$P(r_i|e_h,e_t) = \frac{\exp(l_f^{(h,t)})}{\sum_{r' \in \mathcal{P}_T \cup \{TH\}} \exp(l_f^{(h,t)'})} \quad (۴۱)$$

$$\mathcal{L}_1 = -\sum_{r \in \mathcal{P}_T} \log(P(r_i|\, e_h,e_t) \quad (۴۲)$$

$$\mathcal{L}_2 = -\log(\frac{\exp\left(l_f^{(h,t)^{TH}}\right)}{\sum_{r' \in \mathcal{N}_T \cup \{TH\}} \exp\left(l_f^{(h,t)'}\right)}) \quad (۴۳)$$

$$\mathcal{L} = \mathcal{L}_1 + \mathcal{L}_2 \quad (۴۴)$$

که در آن کلاس‌های مثبت $\mathcal{P}_T \subseteq R$ روابطی هستند که بین نهادهای مجموعه T وجود دارد. اگر T شامل هیچ رابطه‌ای را نباشد، $\mathcal{P}_T$ خالی است. کلاس‌های منفی $\mathcal{N}_T \subseteq R$ روابطی هستند که بین نهادها وجود ندارند. اگر T شامل هیچ رابطه‌ای نباشد $\mathcal{N}_T = R$. طبق تابع هزینه تعریف روابط درست بین نهادها استخراج می‌شود.

۴-۳- تشخیص رویدادها

جهت تشخیص رویدادهای امنیتی از شبکه عمیق KVP استفاده‌شده است. این شبکه با در نظر گرفتن مکانیزم‌های توجه و پیش‌بینی کمک زیادی در تشخیص دقیق رویدادها می‌کند. لایه‌های توجه به شناسایی وابستگی‌های جهانی بدون توجه به‌توالی کمک می‌کنند. این لایه‌ها از بردارهای زمینه و وزن‌های توجه تشکیل‌شده‌اند و می‌توانند همبستگی بین دو عنصر که موقعیت‌های متفاوتی در یک دنباله ورودی دارند را شناسایی کند. بردار زمینه شامل یک بردار خلاصه کننده است که پیش‌بینی تکرارهای گذشته را در برمی‌گیرد [۴۲]. دانیلوک و همکاران [۴۳] این رویکرد را بهبود بخشیدند و مکانیزم توجه را برای تقویت LSTM پیشنهاد کردند که بردارهای خروجی را به یک کلید، یک مقدار و یک بخش پیش‌بینی جدا می‌کند. این قابلیت، پیش‌بینی توکن بعدی در توالی‌های بسیار بزرگ بهبود می‌بخشد. ما از این رویکرد و ساختار شبکه عصبی عمیق KVP برای پیش‌بینی و استخراج رویدادها استفاده کردیم.

$Y_t = [h_{t-L} \dots h_{t-1}]$ تاریخچه‌ای از L بردار خروجی GRU است، که k بعد خروجی و h_t بازنمایش خروجی بازه زمانی t است، توزیع احتمال y_t برای رویداد بعدی به‌صورت زیر محاسبه می‌شود.

$$\begin{bmatrix} k_t \\ v_t \\ p_t \end{bmatrix} = h_t \in \mathbb{R}^{3k} \quad (۴۵)$$

$$M_t = \emptyset(W^Y[k_{t-L} \dots k_{t-1}] + (W^h k_t) \in \mathbb{R}^{k*L} \quad (۴۶)$$

$$\alpha_t = sotmax(w^T M_t) \in \mathbb{R}^{1*L} \quad (۴۷)$$

$$r_t = [v_{t-L} \dots v_{t-1}]a^T \in \mathbb{R}^k \quad (۴۸)$$

$$h_t^* = \emptyset[W^r r_t + W^x p_t] \in \mathbb{R}^k \quad (۴۹)$$

$$y_t = sotmax(W^* h_t^* + b) \in \mathbb{R}^{|v|} \quad (۵۰)$$

در روابط بالا W^Y، W^h، $W^r \in \mathbb{R}^{k*k}$ و $W^x \in \mathbb{R}^{|V|*k}$ ماتریس‌های قابل آموزش، $w \in \mathbb{R}^k$ بردار قابل‌آموزش، $b \in \mathbb{R}^{|v|}$

بردار بایاس و $\emptyset$ تابع هذلولی است. بردار خروجی h_t به سه قسمت کلید (k_t)، مقدار (v_t) و پیش‌بینی (p_t) تقسیم می‌شود. p_t برای رمزگذاری توزیع رویداد بعدی، k_t به‌عنوان کلید و v_t به‌عنوان مقدار مکانیزم توجه استفاده می‌شود.

توزیع توجه α_t از مقایسه کلید مرحله زمانی t با L کلید قبلی محاسبه می‌شود، که سپس برای به دست آوردن بازنمایش وزنی $r_t \in \mathbb{R}^k$ از مقادیر مرتبط با این کلیدها استفاده می‌شود. بازنمایش نهایی h_t^* از ترکیب غیرخطی بازنمایش توجه وزن‌دار r_t و رمزگذاری توزیع رویداد بعدی p_t محاسبه می‌شود و متعاقباً برای پیش‌بینی احتمال رویداد بعدی با استفاده از y_tاستفاده می‌شود.

۴- نتایج شبیه‌سازی

در این بخش روش پیشنهادی از جنبه‌های مختلف موردبررسی قرارگرفته و با سایر روش‌ها مقایسه شده است. آزمایش‌ها روی رایانه شخصی با GPU GeForce RTX-2080 و پردازنده Core I7-9700K با فرکانس ۴۴۰۰ مگاهرتز با استفاده از نرم افزار python انجام‌شده است. لایه‌های پنهان GRU، ۲۵۰ بعد[1] دارند. شبکه عصبی روبه‌جلو شامل دولایه پنهان با ۲۰۰ بعد است. شبکه‌های عصبی عمیق مدل پیشنهادی روی پیکره‌های توسعه CoNLL-2012 و OSINT با ۱۹ تکرار[2] و حذف تصادفی[3] ۰/۴ و نرخ‌های یادگیری[4] $۱*۱۰^{-۴}$ و $۲*۱۰^{-۳}$ و اندازه دسته ۶۴ تنظیم دقیق[5] شده‌اند. منابع دانش استفاده‌شده به شرح زیر است:

منبع دانش رویدادهای سایبری[6]: این منبع شامل اطلاعات در دسترس عموم در مورد رویدادهای سایبری از سال ۲۰۱۴ تا به امروز است و برای تصمیم‌گیری استراتژیک در مورد چگونگی سرمایه‌گذاری جهت جلوگیری و پاسخ به رویدادهای سایبری از طریق رفع فقدان داده‌های سازگار و ساختاریافته لازم ایجادشده است. در این منبع از تکنیک‌های خودکار همراه با بررسی و طبقه‌بندی دستی توسط محققان برای جمع‌آوری و ساختار داده‌ها از انواع سایت‌های خبری باز، وبلاگ‌ها و سایر سایت‌های تخصصی استفاده‌شده است که حملات منتسب به عموم را شناسایی و موردبحث قرار می‌دهند. داده‌ها ماهانه به‌روز می‌شوند و اطلاعاتی در مورد عامل تهدید، انگیزه، قربانی و اثرات نهایی حمله به دست می‌دهند. سه‌تایی‌های استخراج‌شده از این منبع ۲۲۲۳۴ تا است، به‌عنوان‌مثال سه‌تایی (باج افزار، است، یک بدافزار).

منبع دانش CASIE [۴۴]: این منبع شامل مجموعه‌ای حاوی ۱۰۰۰ داده حاشیه‌نویسی شده برای پنج نوع حمله امنیت سایبری است. توسط این منبع فرآیند استخراج اطلاعات مربوط به امنیت سایبری از داده‌های خام و تسهیل می‌شود.

منبع دانش توییتر [۴۵]: این منبع شامل ۲۱۰۰۰ توییت مرتبط با امنیت سایبری است که صورت دستی برچسب‌گذاری شده است.

منبع دانش بدافزارها: پایگاه‌های MalwareTextDB [۴۶] و پایگاه دانش ری و همکاران [۴۷] شامل داده‌های حاشیه‌نویسی برای گزارش‌های بدافزار است که اطلاعات معنایی از متن را ارائه می‌دهد و به سیستم و محققان کمک می‌کند تا به‌سرعت قابلیت‌های بدافزار را درک کنند. در این پایگاه از منابع متنوعی ازجمله متن بدون برچسب، متون حاشیه‌نویسی انسانی و خصوصیات در مورد ویژگی‌های بدافزار برای شناسایی اسناد بدافزار استفاده‌شده است.

داده‌های:OSINT این داده‌ها شامل پایگاه داده آسیب‌پذیری CVE، بولتن‌های امنیتی، گزارش‌های تهدیدات پایدار پیشرفته (APT) و در دسترس عموم است. اطلاعات موجود در این پایگاه شامل ۱۳۲۶۲ جمله حاشیه‌نویسی شده حاوی ۷۵۹۹۰ سه‌تایی است [۴۸].

منبع دانش CVE: CVE[7] فهرستی از آسیب‌پذیری‌های امنیتی رایانه‌ای است که به‌صورت عمومی افشاشده‌اند [۴۹].

پایگاه داده گزارش‌های:APT گزارش‌های APT، مقالات و وبلاگ‌های در دسترس عموم هستند که مربوط به فعالیت‌های مخرب و مرتبط با سازمان‌ها یا مجموعه‌های ابزار APT هستند [۵۰].

علاوه بر منابع دانش سایبری که در بالا ذکر شد، از منابع دانش مختلف زیر نیز جهت بالابردن دانش روش پیشنهادی در حیطه‌های مختلف استفاده‌شده است.

دانش عرفی: در Conceptnet دانش عرفی در قالب جملات زبان طبیعی ذخیره می‌شود، ازاین‌رو بازنمایی روابط در قالبی صوری کار مشکلی خواهد بود. روش بهینه دیگر استفاده از ابزارهای پردازش زبان طبیعی جهت استخراج چنین حقایقی از منابعی همچون ویکی‌پدیاست که پایگاهی جامع از اطلاعات دنیای پیرامون ماست و به‌طور منظم توسط کاربرانش به‌روزرسانی و کنترل می‌شود. این منبع شامل ۶۰۰ هزار سه‌تایی است، به‌عنوان‌مثال (بدافزار، حاوی، فایل آلوده) و (شرکت، استفاده می‌کند، تکنیک هجومی). تمام روابط در این منبع به‌صورت دستی مشخص‌شده است. در این مقاله سه‌تایی‌های با امتیاز بالا (بالاتر از ۳) انتخاب‌شده است که تعداد آن‌ها ۶۱۶۷۳ سه‌تایی است.

ویژگی‌های زبان‌شناسی: علاوه بر منابع دانش دست‌نویس که به آن‌ها اشاره شد، ویژگی‌های زبان‌شناسی نظیر جاندار بودن، جنسیت نیز در نظر گرفته‌شده است. تجزیه‌کننده[8] استنفورد برای تولید تعداد، جانداری و جنسیت برای همه عبارات اسمی به کار می‌رود که می‌تواند به‌طور خودکار دانش زبانی (به‌صورت سه‌تایی)

[1] Dimensions
[2] Epochs
[3] Dropout
[4] Learning Rates
[5] Fine-Tune
[6] Cyber Events

[7] Common Vulnerabilities and Exposures
[8] Parser

برای داده‌های ما تولید کند. ویژگی تعداد شامل مقدار s برای مفرد و p مقدار برای جمع است. ویژگی جانداری نشان‌دهنده جانداری که در صورت جانداری، مقدار مرد، زن و خنثی بودن برای ویژگی جنسیت استفاده می‌شود. برای مثال نام بری "هکرها" به‌عنوان جمع و خنثی برچسب‌گذاری می‌شود که برای نمایش از سه‌تایی‌های ("هکرها"، تعداد، جمع) و ("هکرها"، جنسیت و جانداری، خنثی) استفاده می‌شود. درنتیجه ۴۰۱۴۹ سه‌تایی برای جمع بودن و ۴۰۴۶۲ سه‌تایی برای جنسیت و جانداری در نظر گرفته می‌شود.

دانش معنایی [۵۱]: برای جمع‌آوری دانش معنایی، ابتدا با استفاده از تجزیه‌کننده استنفورد اطلاعات پایگاه ویکی‌پدیا انگلیسی[۷] را تجزیه نمودیم و تمام لبه‌های وابستگی را با فرمت (گزاره [۱]، آرگومان [۲]، رابطه، شماره) استخراج شد. در تجزیه‌کننده استنفورد، هنگامی‌که فعل یک فعل ارتباطی (مانند هستم (am)، است (is)) است، بین پیش‌بینی کننده (گزاره) و موضوع، لبه "nsubj" ایجاد می‌شود. احتمال ارتباط هر جفت در این منبع دانش طبق رابطه (۵۱) محاسبه می‌شود.

$$P_r(a|p) = \frac{Count_r(p,a)}{Count_r(p)} \quad (۵۱)$$

در رابطه (۵۱) $Count_r(p)$ و $Count_r(p,a)$ به ترتیب نشان‌دهنده این هستند که چند بار p و جفت گزاره-آرگومان (p,a) در رابطه r وجود داشته‌اند. در آزمایش‌ها زمانی که $P_r(a|p) > 0.1$ و $Count_r(p,a) > 10$ بود، سه‌تایی (p,r,a) را در نظر گرفتیم. به‌عنوان‌مثال، ("فایل آلوده"، nsubj، "تکنیک هجومی") یک رابطه معنایی معتبر است. سرانجام، دو رابطه معنایی nsubj و dobj برای منبع دانش معنایی انتخاب شدند که به‌ترتیب شامل ۱۷۰۷۴ و ۴۵۳۶ جفت گزاره- آرگومان برای nsubj و dobj است.

همان‌طور که ذکر شد یکی از مراحل مهم در استخراج اطلاعات، تحلیل و استخراج عبارات وابسته و اشاره‌کننده به یک‌نهاد در متن با طول‌های مختلف است؛ زیرا ممکن است از عباراتی با طول مختلف برای اشاره به یک‌نهاد امنیتی یا غیرامنیتی استفاده شود. جهت بررسی کارایی روش پیشنهادی در زمینه تشخیص نهادها از پایگاه داده English Gigaword [۵۲] استفاده‌شده است. همان‌گونه که در شکل (۲) نشان داده‌شده است، مقدار معیار Avg.F1 برای روش پیشنهادی در مقایسه با روش‌های قبلی روی این پایگاه داده به میزان ۶/۶۲ درصد بهبودیافته است؛ زیرا درروش پیشنهادی از مدل‌های زبانی، توسط روش پیش آموزش استفاده‌شده و تعبیه درست و دقیق واژگان، اطلاعات معنایی خوبی را در اختیار سیستم قرار داده است. همچنین استفاده از اطلاعات سطح نهاد، منابع دانش مختلف و عبارات اشاره‌کننده به یک‌نهاد نیز دقت روش پیشنهادی را در این زمینه بالابرده است. روش‌های سان و مونر بعد از روش پیشنهادی به ترتیب بالاترین مقادیر مربوط به F1 رادارند. ترکیب این دو روش نیز می‌تواند، روش مناسبی برای تشخیص نهادها باشد که لازم است در آن مقادیر دقت و بازخوانی برای هر نوع نهاد (مانند اشخاص، مکان‌ها و سازمان‌ها) به‌طور جداگانه محاسبه شود. ازآنجاکه برای بازنمایش دانش درروش پیشنهادی از منابع دانش مختلفی استفاده‌شده است، روش پیشنهادی علاوه بر استخراج اطلاعات از داده‌های CTI در استخراج اطلاعات از داده‌های دیگر نیز موفق عمل می‌کند. با در نظر گرفتن تعبیه واژگان ROBERTA و اطلاعات عبارات اسمی اطلاعات معنایی بهتری در اختیار سیستم قرارگرفته و سیستم توانایی بالاتری در تشخیص درست نام بری‌ها پیدا می‌کند. همچنین توسط ساختار شبکه عصبی عمیق ارائه‌شده ویژگی‌های مفیدتری توسط شبکه برای نام بری‌ها استخراج‌شده و باعث می‌شود نام بری‌هایی که در روش‌های دیگر نادیده گرفته‌شده‌اند، درروش پیشنهادی در نظر گرفته شوند.

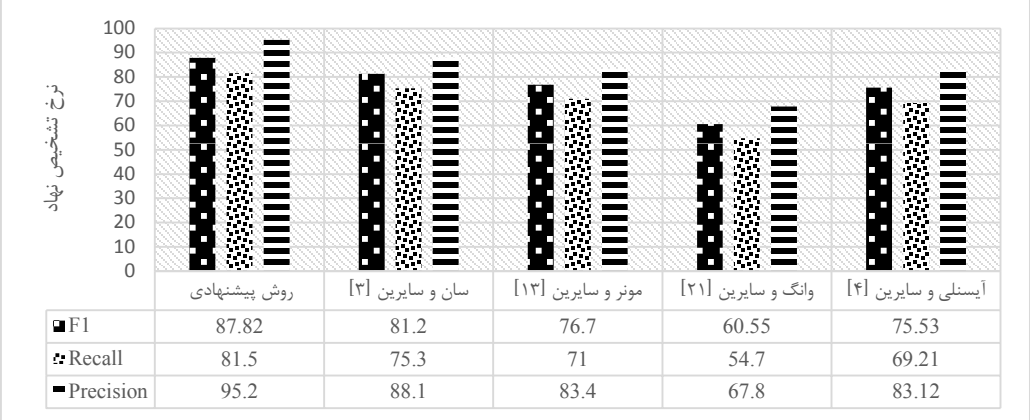
در جدول (۱) نرخ تشخیص روابط بین نهادها برای روش پیشنهادی طبق معیارهای MUC [۵۳]، B^3 [۲۹] و $CEAF_{\phi4}$ [۵۵] توسط مقادیر دقت و بازخوانی و Avg.F1 (میانگین خروجی سه معیار MUC، B^3 و $CEAF_{\phi4}$) طبق روابط (۵۳-۵۱) با روش‌های مطرح و جدید استخراج اطلاعات روی پایگاه داده [۵۶] CONLL-2012 مقایسه شده است. معیار MUC پیونددهی دو عبارت وابسته به یکدیگر را با یک ربط مشخص کرده و بر مبنای این ربط‌ها کار می‌کند. یعنی ربط‌های خروجی سیستم مورد ارزیابی را در نظر گرفته و مشخص می‌کند کدام درست و کدام نادرست است و بر مبنای آن‌ها محاسبات خود را انجام می‌دهد. معیار B^3 دقت و بازخوانی را برای هر نام بری درون هر نهاد به‌صورت مجزا محاسبه می‌کند و به همین دلیل تا حد زیادی مشکل معیار MUC در تمایل به سمت نهادهای حاوی نام بری‌های زیاد را حل کرده است. معیار $CEAF_{-\varphi4}$ ارزیابی را با ایجاد تناظر یک‌به‌یک بین نهادهای خروجی سیستم و استاندارد طلایی انجام می‌دهد.

همان‌طور که در جدول (۱) مشاهده می‌شود، روش پیشنهادی طبق همه معیارها بر روش‌های موجود برتری دارد و مقدار F1 به میزان ۷/۲ درصد بهبودیافته است. از دلایل برتری روش پیشنهادی نسبت به سایر روش‌های جدید می‌توان به در نظر گرفتن اطلاعات عبارات اسمی، استفاده از تعبیه واژگان شبکه از پیش آموزش‌دیده ROBERTA و رتبه‌بندی دقیق عبارات کاندید توسط مدل رتبه‌بندی چندمعیاره اشاره کرد. همچنین، توسط روند وزن دهی دقیق به ویژگی‌ها و در نظر گرفتن میزان اهمیت آن‌ها در مرحله تصمیم‌گیری چندمعیاره و تولید عبارات وابسته به یک‌نهاد، روش پیشنهادی توانسته است به‌دقت بالاتری نسبت به سایر روش‌ها دست یابد. یکی دیگر از دلایلی که تأثیر بسزایی در

[1] Predicate
[2] Argument

استخراج درست روابط بین نهادها داشته، استفاده از شبکه عصبی مبتنی بر گراف و تابع هزینه ابتکاری تعریف‌شده است که باعث شده روابط بین نهادها با دقت بالایی نسبت به سایر روش‌های موجود استخراج شود.

در روش‌های مورد مقایسه برای استخراج روابط بین نهادها از شبکه‌های عصبی بازگشتی و پیچشی استفاده است. شبکه عصبی بازگشتی (RNN) برای مدیریت وابستگی‌های طولانی‌مدت استفاده می‌شود، اما مشکل چرخه‌های طولانی آموزش را دارد، LSTM نسبت به این نوع شبکه‌ها وابستگی طولانی‌مدت را بهتر مدیریت می‌کند. شبکه‌های عصبی مبتنی گراف می‌توانند توابع گره محور و گراف محور را مدل کنند و برای استخراج جمله، خوشه‌بندی اسناد و طبقه‌بندی استفاده می‌شوند. شبکه عصبی پیچشی(CNN) فقط می‌تواند داده‌های اقلیدسی معمولی، مانند تصاویر و متن را پردازش کند[۵۸] و برای داده‌های غیر اقلیدسی، نتایج پردازش رضایت‌بخشی تولید نمی‌کنند.



شکل (۲): نرخ تشخیص نهادها روی پیکره English Gigaword

جدول (۱). مقایسه F1 در زمینه تشخیص روابط بین نهادها روی پیکره آزمایش CoNLL-2012 انگلیسی

روش	MUC			B^3			$CEAF\_\varphi_4$			
	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1	Avg.F1
[۳۱]	۸۱/۴	۷۹/۵	۸۰/۴	۷۲/۲	۶۹/۵	۷۰/۸	۶۸/۲	۶۷/۱	۶۷/۶	۷۳/۰
[۲۵]	۸۵/۴	۷۷/۹	۸۱/۴	۷۷/۹	۶۶/۴	۷۷/۷	۷۰/۶	۶۶/۳	۶۸/۴	۷۳/۸
[۲۶]	۸۲/۶	۸۳/۴	۸۳/۰	۷۳/۳	۷۶/۱	۷۴/۷	۷۲/۳	۷۱/۱	۷۱/۷	۷۶/۶
[۲۴]	۸۴/۷	۸۲/۴	۸۳/۵	۷۶/۵	۷۴/۰	۷۵/۳	۷۴/۱	۶۹/۸	۷۱/۹	۷۶/۹
روش پیشنهادی	۹۱/۳	۸۷/۴	۸۹/۳	۸۱/۸	۸۵/۱	۸۳/۴	۷۸/۹	۸۰/۴	۷۹/۶	۸۴/۱

برای بررسی میزان تأثیر روش‌های مختلف تعبیه واژگان نظیر ROBERTA، ELMo، BERT و XLNet بر عملکرد روش پیشنهادی و مقدار Avg.F1، این روش‌ها در جدول (۲) روی مجموعه توسعه پیکره CoNLL-2012 با یکدیگر مقایسه شده‌اند.

جدول (۲). نتایج ارزیابی تعبیه واژگان مختلف روی پیکره توسعه CoNLL-2012

	Avg. F1	Δ
مدل پیشنهادی) تعبیه واژه (ROBERTA	۸۴/۲	
تعبیه واژه XLNet	۸۱/۱	-۳/۱
تعبیه واژه BERT	۷۸/۳	-۵/۹
تعبیه واژه ELMo	۷۶/۱	-۸/۱

همان‌طور که مشخص است کارکرد روش BERT از روش ELMo اندکی بهتر است. روش ELMo در ساختار خود از مدل‌های زبانی دوجهته و شبکه BiLSTM استفاده کرده و ویژگی‌های نحوی و معنایی و زبان‌شناسی را تا حدودی در اختیار سیستم قرار می‌دهد.

BERT در دو مرحله کار می‌کند، ابتدا، از مقدار زیادی از داده‌های بدون برچسب استفاده می‌کند تا بازنمایش یک‌زبان را به روشی بدون نظارت به نام پیش آموزش، یاد بگیرد. سپس، مدل از قبل آموزش‌دیده با استفاده از مقدار کمی از داده‌های آموزش‌دیده برچسب زده‌شده برای انجام کارهای مختلف نظارت‌شده، می‌تواند به‌صورت دقیق تنظیم شود.

عملکرد بهتر BERT دو دلیل دارد. اولاً، یک مدل پیش آموزش جدید به نام‌های مدل پوشش زبان (MLM[1]) و مکانیزم پیش‌بینی جمله بعدی (NSP[2]) ارائه داده است. دوماً، برای آموزش BERT از تعداد زیادی داده و قدرت محاسباتی بالا استفاده‌شده است. MLM امکان یادگیری دوطرفه از متن را ممکن می‌سازد، یعنی به مدل

[1] Masked Language Model
[2] Next Sentense Prediction

اجازه می‌دهد تا مفهوم هر کلمه را از کلماتی که قبل و بعدازآن ظاهر می‌شود بیاموزد. این کار درELMo امکان‌پذیر نیست، زیرا از مدل دوطرفه کم‌عمق استفاده می‌کند. استفاده از تعبیه واژگان XLNet عملکرد بهتری (درصد) نسبت به روش BERT دارد. XLNet یک مبدل دوطرفه بزرگ است که از روش آموزش پیشرفته، داده‌های بزرگ‌تر و قدرت محاسباتی بیشتر برای دستیابی بهتر از به معیارهای پیش‌بینی BERT در بیست وظیفه زبانی استفاده می‌کند. برای بهبود آموزش، XLNet مدل‌سازی زبانی جای‌گشتی را معرفی می‌کند که در آن‌همه توکن‌ها به‌صورت تصادفی پیش‌بینی می‌شوند. این برخلاف مدل زبان پوششی BERT است که در آن فقط پانزده درصد توکن‌های پوشش‌دار پیش‌بینی‌شده است. درنهایت روش ROBERTA بهتر از تمام روش‌های تعبیه واژگان عمل می‌کند، زیرا ROBERTA برای بهبود روند آموزش، وظیفه NSP را از پیش آموزش BERT حذف کرده است. همچنین این روش از آموزش دسته‌ای بزرگ‌تر استفاده کرده است. همچنین، این روش معماری روش XLNet را بهبود داده و توانسته است ویژگی‌های بهتری نسبت به این روش ارائه دهد و در اکثر زمینه‌های پردازش متن نتایج قابل‌قبولی ارائه کند.

جهت بررسی دقت رتبه‌بندی عبارات موجود در متن برای نهادها، روش پیشنهادی را طبق جدول (۳) با سایر روش‌های رتبه‌بندی مقایسه کردیم. همان‌طور که مشخص است، روش ELECTRE در مقایسه با روش‌های TOPSIS [۵۹] و VIKOR [۶۰] دقت بالاتری دارد. روش ویکور برای رتبه‌بندی و یافتن بهترین گزینه، از مفهوم بدترین گزینه و میزان سازش میان فاصله گزینه‌ها نسبت به بهترین گزینه استفاده می‌کند. به همین علت جزء روش‌های برنامه‌ریزی سازشی است. این روش در مقایسه با روش تاپسیس، در محاسبه فواصل گزینه‌ها، میزان اهمیت فاصله مطلوب نسبت به بهترین و بدترین حالت را در نظر می‌گیرد. درروش TOPSIS گزینه انتخابی باید کمترین فاصله از جواب ایده آل و دورترین فاصله از جواب ضد ایده آل را داشته باشد. روشTOPSIS دونقطه مرجع (ایده آل و ضد ایده آل) را معرفی می‌کند ولی اهمیت نسبی فواصل از این دونقطه را در نظر نمی‌گیرد. به همین خاطر روش ELECTRE در رتبه‌بندی عبارات کاندید برای نهادها بهتر عمل می‌کند.

جدول(۳). مقایسه روش‌های رتبه‌بندی عبارات وابسته به نهادها

روش رتبه‌بندی	دقت رتبه‌بندی (درصد)
ELECTRE	۹۲/۶
VIKOR	۸۹/۳
TOPSIS	۸۶/۷

در شکل (۳) جهت سنجش عملکرد افزودن ماژول‌های مختلف به سیستم درزمینه استخراج اطلاعات (میانگین F1 تشخیص نهاد، روابط بین نهادها و رویدادها)، مقدار F1 تحت مقادیر دقت و بازخوانی روی پیکره امنیتی OSINT نمایش داده‌شده است. مقادیر دقت و بازخوانی درزمینه استخراج رویدادها طبق روابط (۵۱) و (۵۲) محاسبه می‌شود. با توجه به شکل (۳) مشخص است، که با افزودن بردار ماژول‌های مختلف مقدار F1 افزایش‌یافته است.

$$\text{دقت} = \frac{\text{تعداد رویدادهای درست استخراج شده}}{\text{تعداد کل رویدادهای استخراج شده}} \quad (۵۱)$$

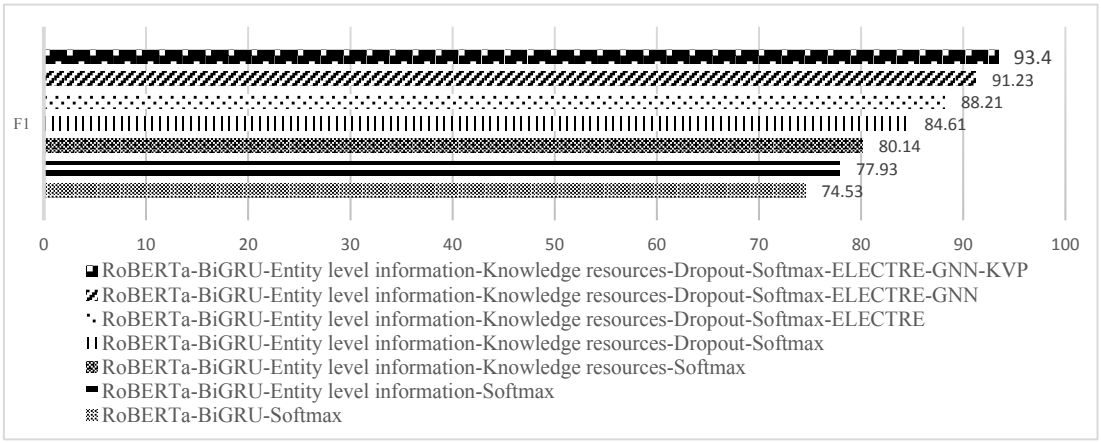
$$\text{بازخوانی} = \frac{\text{تعداد رویدادهای درست استخراج شده}}{\text{تعداد کل رویدادهای درست}} \quad (۵۲)$$

سایز حالت پنهان[1] در شبکه‌های GRU، ۳۰۰ در نظر گرفته‌شده است. مقادیر دیگری نیز مورد آزمایش قرار گرفت ولی این مقدار بهترین نتیجه را حاصل کرد. بنابراین، خروجی حاصل از شبکه Bi-GRU برداری با ۶۰۰ بعد است که از اتصال خروجی شبکه روبه‌جلو و شبکه روبه عقب به دست می‌آید. یک‌لایه Dropout برای جلوگیری از بیش برازش پس از لایهٔ Bi-GRU قراردادیم. با این عمل برخی از ورودی‌ها به لایهٔ بعد صفر می‌شوند. در اینجا Dropout، ۰/۵ در نظر گرفته‌شده است، بدین معنا که در هر مرحله نیمی از ورودی‌ها در نظر گرفته نمی‌شوند. با یک بررسی جزئی‌تر می‌توان به این نتیجه رسید که هرچه میزان اطلاعاتی که از ظاهر، معنا و ساختار واژگان در قالب بردارهای بازنمایش به شبکه وارد می‌شود، افزایش یابد، مقدار دقت نیز افزایش پیدا می‌کند و به‌تبع آن F1 نیز بیشتر می‌گردد. افزایش میزان اطلاعات از محتوای متنی که واژگان در آن قرارگرفته است با استفاده از Bi-GRU و یک‌لایه GRU بر روی خروجی آن، موجب افزایش فراخوانی می‌گردد و به‌تبع آن F1 نیز افزایش می‌یابد. همان‌طور که مشخص است استفاده از ویژگی‌های استخراج‌شده توسط شبکه‌ها تأثیر قابل‌قبولی بر بالا رفتن دقت استخراج روابط و رویدادها دارد. همچنین، افزودن اطلاعات سطح نهاد و پس‌ازآن افزودن منابع دانش مختلف نظیر رویدادها و حملات سایبری، توییت‌های مربوط به امنیت سایبری، گزارش بدافزارها، تهدیدات سایبری، آسیب پدیری‌های امنیتی، منابع معنایی، عرفی، زبان‌شناسی و درنتیجه بالا رفتن دانش سیستم از طریق بازنمایش این منابع بالاترین تأثیر را در بالابردن دقت استخراج اطلاعات دارد. همان‌گونه که ذکر شد منبع دانش رویدادهای سایبری نیز درروش پیشنهادی استفاده‌شده است که در کنار شبکه KVP تأثیر خوبی در دقت استخراج رویدادها و رویدادهای آتی دارد. استفاده از شبکه عصبی مبتنی بر گراف نیز تأثیر قابل‌قبولی بر بالا رفتن F1 در زمینه تشخیص روابط بین نهادها دارد.

در جدول (۴) عملکرد روش پیشنهادی و روش لی و سایرین توسط داده‌های مختلف آموزش و آزمودن نمایش داده‌شده است. همان‌طور که مشخص است هر دو روش زمانی که داده‌های آموزش از پیکره OSINT و داده‌های آزمودن از پیکره CoNLL-2012 است، عملکرد ضعیفی دارند اما عملکرد روش پیشنهادی نسبت به روش سان بهتر است. این نشان می‌دهد که دانش سایبری تأثیر کمی در پیکره CoNLL-2012 دارد. زمانی که داده‌های آموزش از پیکره CoNLL-2012 و داده‌های آزمودن از پیکره OSINT باشند، روش پیشنهادی

[1] Hidden State Size

عملکرد قابل قبولی (۸/۵+) در مقایسه با روش سان و سایرین دارد که نشان‌دهنده توانایی روش رتبه‌بندی در این پیکره است.



شکل (۳): دقت استخراج اطلاعات با افزودن ماژول‌های مختلف به مدل پیشنهادی روی پیکره OSINT

جدول(۴). ارزیابی عملکرد روش سان و همکاران و روش پیشنهادی توسط داده‌های آموزشی مختلف روی پیکره‌های مختلف

روش	داده‌های آموزش	داده‌های آزمودن	
		CoNLL-2012	OSINT
روش سان و همکاران	CoNLL-2012	۷۳/۸	۷۶/۷
	OSINT	۵۱/۴	۹۰/۱
روش پیشنهادی	CoNLL-2012	۷۹/۲	۸۵/۲
	OSINT	۷۰/۸	۹۶/۳

در شکل (۴) دقت تشخیص عبارات وابسته به نهادها از میان دنباله کلمات و بر اساس تعداد کلمات موجود در دنباله با سایر روش‌های کانتور و گلوبرسون [۵۷] و لی و سایرین [۷] روی پیکره سایبری OSINT مقایسه شده است. همان‌طور که مشاهده می‌شود، دقت روش پیشنهادی از سایر روش‌ها بالاتر است.

شکل (۴): نرخ تشخیص عبارات وابسته به نهادها برحسب طول دنباله کلمات

در سایر روش‌ها دقت تشخیص نام بری با افزایش طول دنباله کلمات کاهش چشمگیری دارد؛ اما درروش پیشنهادی دقت تشخیص نام بری کاهش اندکی دارد و برای دنباله‌هایی با طول بیش از پنج کلمه کاهش دقت ناچیز است. روش کانتور و گلوبرسون عملکرد بهتری نسبت به روش لی و سایرین دارد، زیرا از تعبیه واژگان BERT به‌جای ELMo استفاده کرده است و بازنمایش دقیق‌تری برای دنباله کلمات ارائه داده است. هرچند درروش لی و سایرین همپوشانی زیادی بین نام بری‌هایی طلایی و مجموعه توسعه وجود دارد. مدل پیشنهادی قادر است ۱۰۵۹ نام بری (۳۹۴ نام بری در مجموعه آموزش و ۶۶۵ نام بری که در مجموعه آموزش وجود ندارند.) را به‌طور درست شناسایی کند که روش‌های کانتور و گلوبرسون و لی و سایرین قادر به این میزان تشخیص درست نیستند. همچنین طبق ساختار ارائه‌شده نام بری‌های دیده نشده در مجموعه آموزش نیز قابل‌تشخیص هستند. علاوه بر این، تأثیر افزودن نام بری‌های طلایی نیز بر نرخ تشخیص نام بری بررسی شد و به این نتیجه رسیدیم که روش پیشنهادی عملکرد بهتری نسبت به دو روش دیگر دارد. علت برتری روش پیشنهادی استفاده از شبکه عمیق GRU دوجهته جهت استخراج دنباله کلمات است. همچنین تعبیه واژگان ROBERTA اطلاعات معنایی دقیقی در اختیار سیستم قرار می‌دهد.

جهت بررسی میزان تأثیر هریک از منابع دانش بر عملکرد مدل پیشنهادی، تک‌تک این منابع را از مدل حذف کردیم و بررسی که کردیم که با حذف هریک از آن‌ها دقت روش پیشنهادی روی دو پایگاه داده CONLL-2012 و OSINT چه تغییری می‌کند. همان‌طور که در جدول (۵) مشخص است، هر نوع دانش روی پیکره‌ای خاص عملکرد متغیری دارد. برای مثال، وجود دانش زبان‌شناسی در پیکره CONLL-2012 بالاترین تأثیر را نسبت به بقیه دانش‌ها دارد. درحالی‌که در پیکره سایبریOSINT حذف

منبع دانش و اطلاعات پیکره CASIE بیشترین تأثیر منفی را بر عملکرد مدل پیشنهادی می‌گذارد.

جدول (۵). نتایج حذف منابع دانش روی پیکره‌های CoNLL-2012 و OSINT

پایگاه داده	CONLL-2012		OSINT	
	F1	$\Delta F1$	F1	$\Delta F1$
مدل به‌طور کامل	۸۹/۸	-	۹۳/۴	-
منبع دانش CASIE	۸۵/۶	-۴/۲	۹۰/۲	-۳/۲
منبع دانش CVE	۸۶/۳	-۳/۵	۹۱/۱	-۲/۳
منبع دانش APT	۸۶/۸	-۲/۹	۹۱/۵	-۱/۹
منبع دانش عرفی	۸۶/۴	-۳/۴	۹۳/۲	-۰/۲
منبع دانش زبان-شناسی	۸۴/۳	-۵/۵	۹۳/۱	-۰/۳
منبع دانش معنایی	۸۵/۱	-۴/۷	۹۲/۸	-۰/۶

۵- نتیجه‌گیری

استخراج اطلاعات CTI داده‌های ارزشمندی را درباره تهدیدات سایبری کشف، پردازش و تجزیه تحلیل می‌کند. افزایش تعداد حملات سایبری و نقض داده‌ها، امنیت سایبری را به اولویت اصلی دولت‌ها، شرکت‌ها، سازمان‌های نظامی و افراد تبدیل کرده است. با استفاده از CTI سازمان‌ها می‌توانند مهاجمان خود را بهتر درک کنند، سریع تر به حوادث واکنش نشان دهند و از اقداماتی که عوامل تهدید انجام می‌دهند، جلوتر باشند. این فناوری یک جزء حیاتی از استراتژی جامع امنیت سایبری و یک ابزار ضروری برای ایمن کردن سیستم‌ها و شبکه‌های دیجیتال برای سازمان‌ها و شرکت‌هاست. در این مقاله، یک روش استخراج اطلاعات تهدیدات سایبری بر پایه بازنمایش منابع دانش مختلف (نظیر رویدادها و حملات سایبری، توییت‌های مربوط به امنیت سایبری، گزارش بدافزارها، تهدیدات سایبری، آسیب‌پذیری‌های امنیتی)، تصمیم‌گیری چند شاخصه و شبکه‌های عصبی عمیق ارائه شد. همچنین، مقایسه مناسبی بین روش پیشنهادی و روش‌های جدید استخراج اطلاعات سایبری صورت گرفت و مشخص شد که روش پیشنهادی با کمترین میزان خطا، مسئله استخراج نهادها، روابط بین آن‌ها و استخراج رویدادها را به‌خوبی مدیریت می‌کند. با استفاده از ساختار تصمیم‌گیری چند شاخصه و دخیل کردن تمام معیارها و میزان اهمیت آن‌ها در تصمیم‌گیری، همچنین درنظرگرفتن اطلاعات نهادها و عبارات وابسته به نهادها، نرخ تشخیص نهادهای امنیتی و غیرامنیتی روش ارائه‌شده در مقایسه با روش‌های پیشین دقیق‌تر شده است. همچنین با استفاده از توسط ساختار شبکه عصبی عمیق مبتنی بر گراف روابط بین نهادها با دقت بالایی استخراج شد. با استفاده از شبکه KVP مسئله استخراج رویدادهای امنیتی و پیش‌بینی رویدادهای آتی به‌خوبی مدیریت شد. در تشخیص نهادها مشخص شد که افزایش طول دنباله کلمات تأثیر چندانی در کاهش دقت تشخیص ندارد. در بررسی میزان تأثیر منابع دانش مختلف مشخص شد که دانش زبان‌شناسی و دانش معنایی نقش مهم‌تری در بهبود عملکرد روش پیشنهادی روی پیکره CONLL-2012 دارند. اما در پیکره OSINT دانش‌های CASIE و CVE تأثیر بیشتری در ارتقا عملکرد سیستم دارند. سیستم پیشنهادی عملکردی مناسب در انواع مختلف دادگان دارد و معیار Avg.F1 را درزمینه تشخیص نهادها و روابط بین آن‌ها به ترتیب ۶/۶۲ و ۷/۲ درصد افزایش داده است. دقت استخراج اطلاعات روی پیکره امنیتی OSINT باتوجه‌به دقت بالای مدل ارائه‌شده و تجهیز آن به منابع دانش معنایی، عرفی و زبان‌شناسی علاوه بر منابع دانش سایبری، می‌توان از آن برای ارتقا عملکرد سیستم‌های مبتنی بر متن نظیر خلاصه‌سازی متون، ترجمه ماشینی و پرسش‌وپاسخ در حیطه غیرامنیتی نیز استفاده کرد.

۶- مراجع

[1] L. Zongxun, L. Yujun, Z. Haojie, and L. Juan, "Construction of ttps from apt reports using bert," in 2021 18th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP), IEEE, 2021, pp. 260–263. Accessed: Apr. 15, 2024. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/9674158

[2] K. Dadashtabar Ahmadi, M. Kheirkhah, A. J. Rashidi, " Detection of advanced Cyber Attacks, Using Behavior Modeling Based on Natural Language Processing", ECD, Vol. 6, No. 3, Serial No. 2, pp. 141-151, 2018. doi: 20.1001.1.23224347.1397.6.3.12.2

[3] N. Sun et al., "Cyber threat intelligence mining for proactive cybersecurity defense: a survey and new perspectives," IEEE Communications Surveys & Tutorials, 2023, Accessed: Apr. 20, 2024. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/10117505/

[4] S. Ainslie, D. Thompson, S. Maynard, and A. Ahmad, "Cyber-threat intelligence for security decision-making: a review and research agenda for practice," Computers & Security, p. 103352, 2023. https://doi.org/10.1016/j.cose.2023.103352.

[5] M. H. HassanNia, M. R. HasaniAhangar, A. Gafori, An Improved Method of Incident Detection due to Cyber Attacks, ECD, Vol. 7, No. 4, 2020. Available: https://sid.ir/paper/395725/en.

[6] E.Bastami, H.Soltanizadeh*,M. Rahmanimanesh, P. Keshavarzi, "A Malware Classification Method Using visualization and Word Embedding Features", ECD, Vol. 11, No. 1, 2023. doi: 20.1001.1.23224347.1402.11.1.1.2.

[7] K. Lee, L. He, and L. Zettlemoyer, "Higher-order coreference resolution with coarse-to-fine inference," arXiv preprint arXiv:1804.05392, 2018. https://doi.org/10.48550/arXiv.1804.05392.

[8] H. Peng, D. Khashabi, and D. Roth, "Solving hard coreference problems," arXiv preprint arXiv:1907.05524, 2019. https://doi.org/10.48550/arXiv.1907.05524.

[9] L.-T. Wu, J.-R. Lin, S. Leng, J.-L. Li, and Z.-Z. Hu, "Rule-based information extraction for mechanical-electrical-plumbing-specific semantic web," AUTOMAT CONSTR, vol. 135, p. 104108, 2022. https://doi.org/10.1016/j.autcon.2021.104108.

[10] A. Alamoudi, A. Alomari, and S. Alwarthan, "A rule-based information extraction approach for extracting metadata from PDF books," ICIC Express Letters, Part B: Applications, vol. 12, no. 2, pp. 121–132, 2021. doi:10.24507/icicelb.12.02.121

[11] D. Freitag, J. Cadigan, R. Sasseen, and P. Kalmar, "VALET: rule-based information extraction for rapid deployment," in Proceedings of the Thirteenth Language Resources and Evaluation Conference, 2022, pp. 524–533. Accessed: Oct. 26, 2023. [Online]. Available: https://aclanthology.org/2022.lrec-1.55/

[12] F. Rahma and A. Romadhony, "Rule-Based Crime Information Extraction on Indonesian Digital News," in 2021 International Conference on Data Science and Its Applications (ICoDSA), IEEE, 2021, pp. 10–15. Accessed: Oct. 26, 2023. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/9617509

[13] K. Shaukat, S. Luo, S. Chen and D. Liu, "Cyber Threat Detection Using Machine Learning Techniques: A Performance Evaluation Perspective," 2020 International Conference on Cyber Warfare and Security (ICCWS), Islamabad, Pakistan, pp. 1-6, 2020. doi: 10.1109/ICCWS48432.2020.9292388.

[14] B. M. Davis, M. Salinas-Navarro, M. F. Cordeiro, L. Moons, and L. De Groef, "Characterizing microglia activation: a spatial statistics approach to maximize information extraction," Sci. Rep., vol. 7, no. 1, p. 1576, 2017. https://doi.org/10.1038/s41598-017-01747-8.

[15] G. Tür, D. Hakkani-Tür, and K. Oflazer, "A statistical information extraction system for Turkish," Nat. Lang. Eng, vol. 9, no. 2, pp. 181–210, 2003. doi:10.1017/S135132490200284X.

[16] J. Zhang, "Entropic Statistics: Concept, Estimation, and Application in Machine Learning and Knowledge Extraction," Mach. learn. knowl. extr., vol. 4, no. 4, pp. 865–887, 2022. https://doi.org/10.3390/make4040044.

[17] Y. Ghazi, Z. Anwar, R. Mumtaz, S. Saleem, and A. Tahir, "A supervised machine learning based approach for automatically extracting high-level threat intelligence from unstructured sources," in 2018 International Conference on Frontiers of Information Technology (FIT), IEEE, 2018, pp. 129–134. Accessed: Apr. 22, 2024. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/8616979

[18] K. Narasimhan, A. Yala, and R. Barzilay, "Improving Information Extraction by Acquiring External Evidence with Reinforcement Learning." arXiv, Sep. 27, 2016. Accessed: Oct. 20, 2023. [Online]. Available: http://arxiv.org/abs/1603.07954.https://doi.org/10.48550/arXiv.1603.07954

[19] W. Y. Wang, J. Li, and X. He, "Deep reinforcement learning for NLP," in Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts, 2018, pp. 19–21. Accessed: Oct. 20, 2023. [Online]. Available: https://aclanthology.org/P18-5007/

[20] X. Wang, J. Yang, Q. Wang, and C. Su, "Threat Intelligence Relationship Extraction Based on Distant Supervision and Reinforcement Learning.," in SEKE, 2020, pp. 572–576. Accessed: Apr. 22, 2024. [Online]. Available: https://ksiresearch.org/seke/seke20paper/paper149.pdf

[21] X. Wang et al., "A method for extracting unstructured threat intelligence based on dictionary template and reinforcement learning," in 2021 IEEE 24th International Conference on Computer Supported Cooperative Work in Design (CSCWD), IEEE, 2021, pp. 262–267. Accessed: Apr. 22, 2024. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/9437858

[22] Y. Yang, Z. Wu, Y. Yang, S. Lian, F. Guo, and Z. Wang, "A survey of information extraction based on deep learning," Appl. Sci., vol. 12, no. 19, p. 9691, 2022. https://doi.org/10.3390/app12199691.

[23] H. Jo, Y. Lee, and S. Shin, "Vulcan: Automatic extraction and analysis of cyber threat intelligence from unstructured text," COMPUT SECUR , vol. 120, p. 102763, 2022. https://doi.org/10.1016/j.cose.2022.102763.

[24] X. Wang et al., "Cyber threat intelligence entity extraction based on deep learning and field knowledge engineering," in 2022 IEEE 25th International Conference on Computer Supported Cooperative Work in Design (CSCWD), IEEE, 2022, pp. 406–413. Accessed: Apr. 22, 2024. [Online]. Available:https://ieeexplore.ieee.org/abstract/document/9776139/

[25] K. Ahmed, S. K. Khurshid, and S. Hina, "CyberEntRel: Joint extraction of cyber entities and relations using deep learning," COMPUT SECUR , vol. 136, p. 103579, 2024. https://doi.org/10.1016/j.cose.2023.103579.

[26] Y. Shi, Y. Xiao, P. Quan, M. Lei, and L. Niu, "Document-level relation extraction via graph transformer networks and temporal convolutional networks," Pattern Recognit. Lett, vol. 149, pp. 150–156, 2021. https://doi.org/10.1016/j.patrec.2021.06.012.

[27] C. Park, J. Park, and S. Park, "AGCN: Attention-based graph convolutional networks for drug-drug interaction extraction," Expert Syst. Appl., vol. 159, p. 113538, 2020. https://doi.org/10.1016/j.eswa.2020.113538.

[28] J. Xu, Y. Chen, Y. Qin, R. Huang, and Q. Zheng, "A feature combination-based graph convolutional neural network model for relation extraction," Symmetry, vol. 13, no. 8, p. 1458, 2021. https://doi.org/10.3390/sym13081458.

2818. Accessed: Oct. 26, 2023. [Online]. Available: https://aclanthology.org/2021.acl-long.218/

[39] C. Zheng, J. Feng, Z. Fu, Y. Cai, Q. Li, and T. Wang, "Multimodal Relation Extraction with Efficient Graph Alignment," in Proceedings of the 29th ACM International Conference on Multimedia, Virtual Event China: ACM, Oct. 2021, pp. 5298–5306. doi: 10.1145/3474085.3476968.

[40] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," arXiv preprint arXiv:1409.0473, 2014. https://doi.org/10.48550/arXiv.1409.0473.

[41] A. T. de Almeida, "Multicriteria decision model for outsourcing contracts selection based on utility function and ELECTRE method," Comput. Oper. Res., vol. 34, no. 12, pp. 3569–3574, 2007. https://doi.org/10.1016/j.cor.2006.01.003.

[42] A. Vaswani et al., "Attention is all you need," Advances in neural information processing systems, vol. 30, 2017, Accessed: Jun. 04, 2024. [Online]. Available: https://proceedings.neurips.cc/paper/7181-attention-is-all

[43] M. Daniluk, T. Rocktäschel, J. Welbl, and S. Riedel, "Frustratingly Short Attention Spans in Neural Language Modeling." arXiv, Feb. 15, 2017. Accessed: Jun. 04, 2024. [Online]. Available: http://arxiv.org/abs/1702.04521. https://doi.org/10.48550/arXiv.1702.04521.

[44] T. Satyapanich, F. Ferraro, and T. Finin, "Casie: Extracting cybersecurity event information from text," in Proceedings of the AAAI conference on artificial intelligence, 2020, pp. 8749–8757. Accessed: May 22, 2024. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/6401

[45] V. Behzadan, C. Aguirre, A. Bose, and W. Hsu, "Corpus and deep learning classifier for collection of cyber threat indicators in twitter stream," in 2018 IEEE International Conference on Big Data (Big Data), IEEE, 2018, pp. 5002–5007. Accessed: May 22, 2024. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/8622506

[46] S. K. Lim, A. O. Muis, W. Lu, and C. H. Ong, "Malwaretextdb: A database for annotated malware articles," in Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2017, pp. 1557–1567. Accessed: May 22, 2024. [Online]. Available: https://aclanthology.org/P17-1143/

[47] A. Roy, Y. Park, and S. Pan, "Predicting malware attributes from cybersecurity texts," in Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), 2019, pp. 2857–2861. Accessed: May 22, 2024. [Online]. Available: https://aclanthology.org/N19-1293/

[48] P. Stenetorp, S. Pyysalo, G. Topić, T. Ohta, S. Ananiadou, and J. Tsujii, "BRAT: a web-based tool for NLP-assisted text annotation," in Proceedings of the

[29] H. Zhang, Z. Huang, Z. Li, D. Li, and F. Liu, "Densely Connected Graph Attention Network Based on Iterative Path Reasoning for Document-Level Relation Extraction," in Advances in Knowledge Discovery and Data Mining, vol. 12713, K. Karlapalem, H. Cheng, N. Ramakrishnan, R. K. Agrawal, P. K. Reddy, J. Srivastava, and T. Chakraborty, Eds., LECT NOTES ARTIF INT, vol. 12713. , Cham: Springer International Publishing, 2021, pp. 269–281. doi: 10.1007/978-3-030-75765-6_22.

[30] S. Guo, L. Huang, G. Yao, Y. Wang, H. Guan, and T. Bai, "Extracting Biomedical Entity Relations using Biological Interaction Knowledge," INTERDISCIP SCI, vol. 13, no. 2, pp. 312–320, Jun. 2021, doi: 10.1007/s12539-021-00425-8.

[31] S. Raza and B. Schwartz, "Entity and relation extraction from clinical case reports of COVID-19: a natural language processing approach," BMC MED INFORM DECIS, vol. 23, no. 1, p. 20, Jan. 2023, doi: 10.1186/s12911-023-02117-3.

[32] C. Kruengkrai, T. H. Nguyen, S. M. Aljunied, and L. Bing, "Improving low-resource named entity recognition using joint sentence and token labeling," in Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 5898–5905. Accessed: Oct. 26, 2023. [Online]. Available: https://aclanthology.org/2020.acl-main.523/

[33] P. H. Martins, Z. Marinho, and A. F. T. Martins, "Joint Learning of Named Entity Recognition and Entity Linking." arXiv, Jul. 18, 2019. Accessed: Oct. 26, 2023. [Online]. Available: http://arxiv.org/abs/1907.08243

[34] Y. Lu et al., "Unified Structure Generation for Universal Information Extraction." arXiv, Mar. 23, 2022. Accessed: Oct. 26, 2023. [Online]. Available: http://arxiv.org/abs/2203.12277.

[35] I.-H. Hsu et al., "DEGREE: A Data-Efficient Generation-Based Event Extraction Model." arXiv, May 03, 2022. Accessed: Oct. 26, 2023. [Online]. Available: http://arxiv.org/abs/2108.12724.

[36] J. Gao, H. Zhao, C. Yu, and R. Xu, "Exploring the Feasibility of ChatGPT for Event Extraction." arXiv, Mar. 09, 2023. Accessed: Oct. 26, 2023. [Online]. Available: http://arxiv.org/abs/2303.03836.

[37] D. Zhang, S. Wei, S. Li, H. Wu, Q. Zhu, and G. Zhou, "Multi-modal graph fusion for named entity recognition with targeted visual guidance," in Proceedings of the AAAI conference on artificial intelligence, 2021, pp. 14347–14355. Accessed: Oct. 26, 2023. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/17687

[38] D. Sui, Z. Tian, Y. Chen, K. Liu, and J. Zhao, "A large-scale chinese multimodal ner dataset with speech clues," in Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), 2021, pp. 2807–

[56] S. Pradhan, A. Moschitti, N. Xue, O. Uryupina, and Y. Zhang, "CoNLL-2012 shared task: Modeling multilingual unrestricted coreference in OntoNotes," in Joint Conference on EMNLP and CoNLL-Shared Task, Association for Computational Linguistics, 2012, pp. 1–40.

[57] B. Kantor and A. Globerson, "Coreference Resolution with Entity Equalization," in Proceedings of the 57th Conference of the Association for Computational Linguistics, 2019, pp. 673–677.

[58] Mahmoodi, N., Shirazi, H., Fakhredanesh, M., & Dadashtabar Ahmadi, K. "Improving the performance of the convolutional neural network using incremental weight loss function to deal with class imbalanced data", Electronic and Cyber Defense, vol. 11(4), pp. 17-34, 2024. DOR: https://dor.isc.ac/dor/20.1001.1.23224347.1402.11.4.2.9.

[59] Amudha, M., M. Ramachandran, Vimala Saravanan, P. Anusuya, and R. Gayathri. "A study on TOPSIS MCDM techniques and its application." Data Analytics and Artificial Intelligence 1, no. 1 pp. 09-14, 2021. doi:10.46632/daai/1/1/2.

[60] Yazdani, Morteza, and Felipe R. Graeml. "VIKOR and its applications: A state-of-the-art survey." Int. J. Strateg. Decis. Sci.5, no. 2 , pp. 56-83, 2014. DOI: 10.4018/ijsds.2014040105.

Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics, 2012, pp. 102–107. Accessed: May 22, 2024. [Online]. Available: https://aclanthology.org/E12-2021.pdf

[49] "CVEProject/cvelist." CVE Program, May 22, 2024. Accessed: May 22, 2024. [Online]. Available: https://github.com/CVEProject/cvelist

[50] S. Roy, E. Panaousis, C. Noakes, A. Laszka, S. Panda, and G. Loukas, "SoK: The MITRE ATT&CK Framework in Research and Practice." arXiv, Apr. 14, 2023. Accessed: May 22, 2024. [Online]. Available: http://arxiv.org/abs/2304.07411

[51] J. R. Hobbs, "Resolving pronoun references," Lingua, vol. 44, no. 4, pp. 311–338, 1978. https://doi.org/10.1016/0024-3841(78)90006-2.

[52] "Annotated English Gigaword - Linguistic Data Consortium." Accessed: Apr. 12, 2019. [Online]. Available: https://catalog.ldc.upenn.edu/LDC2012T21

[53] M. Vilain, J. Burger, J. Aberdeen, D. Connolly, and L. Hirschman, "A model-theoretic coreference scoring scheme," in Proceedings of the 6th conference on Message understanding, Association for Computational Linguistics, 1995, pp. 45–52.

[54] A. Bagga and B. Baldwin, "Algorithms for scoring coreference chains," in The first international conference on language resources and evaluation workshop on linguistics coreference, Granada, 1998, pp. 563–566.

[55] X. Luo, "On coreference resolution performance metrics," in Proceedings of the conference on human language technology and empirical methods in natural language processing, Association for Computational Linguistics, 2005, pp. 25–32.