

Cyber Threat Information Extraction using Deep Learning and Knowledge Representation

Samira Hourali^{1*}, Fatemeh Hourali², Atefe Pakzad³

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: ۱۶ June ۲۰۲۴

Revised: ۲۰ August ۲۰۲۴

Accepted: ۱۰ October ۲۰۲۴

Available online: ۱۰ November ۲۰۲۴

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[®] (۲۰۲۰). Cyber Threat Information Extraction using Deep Learning and Knowledge Representation. Journal of Electronic and Cyber Defens. ۲۰۲۰; ۱۳(۲): ۷۳-۸۸.

DOI: <https://dor.isc.ac/dor/۲۰.۱۰۰.۱.۱.۲۳۲۲۴۳۴۷.۱۴.۰۴.۱۳.۲.۷.۴>

© Author(s) retain the copyright and full publishing rights

Publisher: Imam Hossein University.



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

سمیرا حورعلی^{۱*}، فاطمه حورعلی^۲، پاکزاد^۳، عاطفه

^۱ استادیار گروه مهندسی کامپیوتر، دانشکده فنی مهندسی و علوم پایه، دانشگاه کوثر بنجورد، بنجورد، ایران (نویسنده مسئول: s.hourali@kub.ac.ir)
^۲ استادیار گروه مهندسی برق، دانشکده مهندسی برق و کامپیوتر، مجتمع آموزش عالی فنی و مهندسی اسفراین، اسفراین، ایران (hourali.f@gmail.com)
^۳ استادیار گروه مهندسی کامپیوتر، دانشکده فنی مهندسی و علوم پایه، دانشگاه کوثر بنجورد، بنجورد، ایران (atefepakzad@kub.ac.ir)

مشخصات مقاله

تاریخچه مقاله:

نوع مقاله: علمی پژوهشی

دریافت: ۱۴۰۳/۰۳/۲۷

بازنگری: ۱۴۰۳/۰۶/۰۴

پذیرش: ۱۴۰۳/۰۷/۱۹

ارائه آنلاین: ۱۴۰۳/۰۸/۲۰

چکیده (استایل عنوان چکیده)

اطلاعات مربوط به امنیت سایبری به سرعت در اینترنت در حال رشد است و حملات سایبری روز به روز در حال افزایش است. مهاجمان بیشتر بخش‌های نظامی، دولتی و شرکتی را هدف قرار می‌دهند، زیرا این بخش‌ها حاوی اطلاعات حساس و طبقه‌بندی شده‌ای 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 پرسش و پاسخ نیز است. هرچند طی چهار دهه گذشته استخراج اطلاعات همواره یک موضوع تحقیقاتی فعال بوده است؛ اما هنوز هم دقت آن در حد قابل قبول نیست و مدل محاسباتی برای آن وجود ندارد. در این مقاله ابتدا توسط جدیدترین متد تعبیه واژگان، شبکه بازگشتی دوجته 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 استفاده شده است. برای تشخیص و پیش‌بینی دقیق رویدادهای امنیتی از شبکه عمیق KVP مبتنی بر مکانیزم توجه استفاده شده 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 ۲۰۱۲-CoNLL و OSINT به ترتیب به امتیاز ۸/۸۹ و F1 و ۹۳/۴ درصد دست یافته است.

کلید واژه‌ها:

استخراج اطلاعات

تهدیدات سایبری

روابط نهادها

استخراج رویداد

یادگیری عمیق

بازنمایی دانش

استناد: حورعلی، سمیرا^۱، حورعلی، فاطمه^۲، پاکزاد، عاطفه^۳. استخراج اطلاعات تهدیدات سایبری با استفاده از یادگیری عمیق و بازنمایی دانش. پدافند الکترونیک و سایبری. ۱۳ (۲): ۷۳-۸۸. ۱۴۰۴.

<https://dor.isc.ac/dor/۲۰.۱۰۰۱.۱.۲۳۲۲۴۳۴۷.۱۴۰۴.۱۳.۲.۷.۴>

© نویسنده(گان) حق نشر و حقوق کامل انتشار را برای خود محفوظ می‌دارند.



ناشر: دانشگاه جامع امام حسین(ع).

OPEN ACCESS

۱- مقدمه

امنیت سایبری یک موضوع بسیار حیاتی و مهم است که در دنیای امروزه مورد توجه ویژه‌ای قرار گرفته است. با پیشرفت فناوری و ارتباطات، اطلاعات مربوط به امنیت سایبری به سرعت در اینترنت در حال رشد است. بنابراین، حملات سایبری نیز به شدت افزایش یافته است که گاهی عواقب جدی برای افراد و سازمان‌ها در پی دارند. مهاجمان بیشتر بخش‌های نظامی، دولتی و شرکتی را هدف قرار می‌دهند، زیرا این بخش‌ها حاوی اطلاعات حساس و طبقه‌بندی شده‌ای 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 و تحلیل داده‌های بزرگ به ابزاری ارزشمند در امنیت سایبری تبدیل شده است و راهی برای بررسی مقادیر زیادی از داده‌ها برای مشخص کردن تهدیدات و آسیب‌پذیری‌های احتمالی ارائه می‌دهد [۱].

متخصصان امنیتی راه‌حل‌های امنیتی را برای شناسایی و کاهش حملات سایبری توسعه داده و به کار می‌گیرند [۲]. با این حال، 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 (APT^۱) و حملات باج افزار را به عنوان 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]. با این حال، رشد تصاعدی داده‌های CTI، استخراج اطلاعات مرتبط، به موقع و دقیق را چالش برانگیز نموده و این

مشکل باعث توسعه فناوری‌های مختلف شده است.

فرآیند استخراج اطلاعات (IE^۲)، اطلاعات ساختاریافته مفیدی را از داده‌های بدون ساختار یا نیمه ساختاریافته در قالب نهادهای، روابط، رویدادها و بسیاری از انواع دیگر استخراج می‌کند. اطلاعات استخراج شده از داده‌های بدون ساختار برای آماده‌سازی داده‌ها جهت تجزیه و تحلیل استفاده می‌شود. بنابراین، تبدیل کارآمد و دقیق داده‌های بدون ساختار در این فرآیند، تجزیه و تحلیل داده‌ها را بهبود می‌بخشد. فرآیند استخراج اطلاعات شامل سه زیرشاخه است، استخراج نهادهای و روابط بین آن‌ها، استخراج رویدادها [۵] و استخراج اطلاعات چندگانه، به طور کلی، ورودی خط لوله CTI داده‌های خام در مورد امنیت سایبری است، درحالی که خروجی آن دانشی است که می‌تواند در تصمیم‌گیری‌های آینده برای دفاع فعالانه از امنیت سایبری، از جمله استراتژی‌هایی 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 خسته‌کننده است. اما این مسئله یک نکته کلیدی در درک متن است و علاوه بر مباحث امنیتی ذکر شده، 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 در زمینه تشخیص نهادهای این است که معمولاً حین نوشتن یک متن، برای اشاره به یک نهاد 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 ضمائر برای اشاره به آن نهاد استفاده می‌کنیم. حتی ممکن است شخص یا اشیا را با ویژگی‌ها و یا کاربردهای توصیف کنیم. این عبارات برای اشاره به یک نهاد استفاده می‌شوند.

دلیل دیگر متنوع بودن علت وابستگی اطلاعات است، به بیان دیگر ممکن است هر کدام 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 در نظر گرفتن تمام معیارهای وابستگی عبارات مختلف به نهادهای 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 مورد نیاز برای تحلیل درست روابط بین نهادهای، باعث عملکرد ضعیف آن‌ها شده است،

^۲ Information Extraction^۱ Advanced Persistent Threats

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 نوشته شده‌اند، استخراج می‌شوند. این روش‌ها با خلاصه کردن قوانین، اطلاعات را بر اساس الگوها استخراج می‌کنند. از مزایای این روش می‌توان به دقت 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 به صورت دستی یا خودکار برچسب‌گذاری شده‌اند. این داده‌ها برای یادگیری سیستم مورد استفاده قرار می‌گیرند. از طرف دیگر، در روش‌های بدون ناظر نیازی به داده‌های آموزشی نیست (یا داده‌ها آموزشی بسیار کمی مورد نیاز 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 قابل توجهی را در مقایسه با روش‌های قبلی بهبود بخشیده است، اما نیاز به حاشیه‌نویسی دستی زیادی توسط افراد با دانش حرفه‌ای در این زمینه دارند، در نتیجه هزینه کار و زمان در این نوع روش‌ها بسیار بالاست.

در روش‌های مبتنی بر یادگیری تقویتی عبارات اسمی و نهادها به عنوان حالت‌ها^۵ و اتصالات آن‌ها به عنوان عمل‌ها^۶ در نظر گرفته شده‌اند؛ در این روش‌ها، طبق سیاست یادگیری تقویتی و تعریف توابع پاداش^۷ و خطا^۸ 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 با دانش حرفه‌ای زیادی دارد، بنابراین زمان‌بر هستند و هزینه زیادی صرف پیاده‌سازی آن‌ها می‌شود.

جدیدترین روش‌های ارائه شده در حیطه استخراج اطلاعات، روش‌های مبتنی بر یادگیری عمیق هستند [۲۲]. این روش‌ها

زیرا بسیاری از موارد استخراج اطلاعات تنها با کمک دانش جهان قابل حل 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 قابل قبولی می‌شود. توسط تکنیک تصمیم‌گیری چندمعیاره ارائه شده نیز، تمام معیارها و ویژگی‌ها و وزن یا اهمیت آن‌ها نیز در شناسایی عبارات وابسته به یک نهاد در نظر گرفته شده است که با دقت بالاتری نسبت به روش‌های موجود استخراج شده‌اند. در نهایت از منابع دانش مختلف و شبکه عصبی مبتنی بر گراف و KVP^1 برای استخراج دقیق روابط بین نهادها و رویدادها از داده‌های امنیتی استفاده شده 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 به جای انسان به استدلال بپردازند، نقش بازنمایش دانش به سبب دشواری‌های ناشی از مقیاس‌پذیری، حیاتی‌تر و اجتناب‌ناپذیر می‌گردد. دلیل اینکه در روش 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 قابل حل است. در ادامه این مقاله، در بخش دوم روش‌های موجود در زمینه استخراج اطلاعات را بررسی کرده و در بخش سوم مدل در نظر گرفته شده 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 انجام شده و مدل پیشنهادی با مدل‌های پیشرفته فعلی مقایسه شده است. همچنین علت برتری مدل پیشنهادی در مقایسه با مدل‌های موجود نیز شرح داده شده است.

۲- پیشینه تحقیق

علیرغم تلاش‌های زیادی که طی چهار دهه اخیر روی مسئله استخراج اطلاعات انجام شده است، کارایی روش‌های حل این مسئله هنوز به حد قابل قبولی نرسیده است به گونه‌ای 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

^۱ Hidden Markov Model

^۲ Support Vector Machines

^۳ Conditional Random Fields

^۴ States

^۵ Actions

^۶ Reward

^۷ Punishment

^۱ Key-Value-Predict Attention Networks

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 استخراج شده است. این ویژگی‌ها نقش مهمی در درک مفهومی این عبارات توسط شبکه بر عهده‌دارند. در نظر گرفتن اطلاعات سطح نهادها تأثیر قابل‌قبولی در شناسایی و استخراج آن‌ها دارد. همچنین در نظر گرفتن 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 استخراج شده و به‌عنوان ورودی به شبکه بازگشتی دووجهی GRU داده شده 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 ارائه شده عبارات اسمی استخراج شده برای نهادها رتبه‌بندی و اولویت‌بندی می‌شوند. سپس توسط شبکه عصبی مبتنی بر گراف رابطه بین نهادها و عبارات اشاره‌کننده به آن‌ها استخراج می‌شود. در نهایت توسط شبکه KVP رویدادهای موجود در متن استخراج می‌شود. بلوک دیاگرام روش پیشنهادی در شکل (۱) نمایش داده شده است.

۳-۱- استخراج نهادها و عبارات اسمی

با فرض اینکه سند D شامل T کلمه باشد، تعبیه واژگان استخراج شده توسط ROBERTA را به‌صورت $\{x_1, \dots, x_T\}$ در نظر می‌گیریم. برای محاسبه بازنمایش (نمایش برداری) دنباله کلمات^۱، 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

برای پردازش داده‌های بزرگ قابل‌استفاده هستند و به‌طور خودکار ویژگی‌های جمله را پیچیده یاد می‌گیرند. در چند سال گذشته، 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 تکیه بر قوانین از پیش تعریف شده، الگوها و روابط را از داده‌ها بیاموزند و آن‌ها را برای مجموعه داده‌های پیچیده و پویا مانند داده‌های CTI مناسب کنند. با این حال، تکنیک‌های موجود برای استخراج نهاد و رابطه از داده‌های CTI از چندین محدودیت از جمله 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 می‌بیند. این روش‌ها نسبت به روش‌های مبتنی بر یادگیری ماشین روی داده‌های 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] ارائه شده است. اما، توسعه سریع اینترنت منجر به حجم عظیمی از داده‌ها، از جمله 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، در روش پیشنهادی سعی شده است تمام این موارد پوشش داده شود. استخراج نهادها، عبارات وابسته به آن‌ها و روابط بینشان، همچنین استخراج رویدادها از جمله مواردی است که در این مقاله به‌صورت یکجا بررسی و حل شده است. همچنین با توجه به ساختار شبکه‌های عمیق مبتنی بر گراف و KVP استفاده شده 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، زمین-شناسی 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

^۱ Span representation

گرفته شده به صورت سه تایی "ابتدا"، انتها^۵، رابطه^۶ است. بعد از تبدیل تمام اطلاعات به این فرمت، برای ایجاد منبع دانش کلی (G)، تمامی آن‌ها را باهم ترکیب کردیم. منابع دانش استفاده شده، در بخش نتایج شبیه سازی (بخش ۴) شرح داده شده است. برای هر دنباله مدنظر، جهت عبارت اسمی بودن یا وابسته بودن با عبارت دیگر، دانش های مختلفی از منبع دانش و با روش های مختلف می توان استخراج کرد. جهت سادگی و تعمیم در مدل پیشنهادی از تطابق ساختار رشته ای^۷ برای استخراج دانش مرتبط استفاده نمودیم. به طور خاص، برای هر سه تایی $t \in G$ که ابتدا و انتهای آن فهرستی از کلمات است، اگر ابتدای آن مشابه با رشته ای در دنباله کلمات s باشد، آن را یک سه تایی مرتبط می دانیم. بنابراین، اطلاعات t را با میانگین گیری تعبیه تمام کلمات انتهای آن، رمزنگاری می کنیم. برای مثال، اگر s "باچ افزار" باشد و سه تایی ("باچ افزار"، استفاده می کند، "تکنیک هجومی") با جستجو در منبع دانش یافت شود، این رابطه را توسط میانگین تعبیه کلمات "تکنیک" و "هجومی" نمایش داده می دهیم. برای هر جفت دنباله کلمات i و j، بازنمایش آن‌ها به هم الحاق شده و بازنمایش کلی g'_i, g'_j حاصل می شود، تا دانش مناسب انتخاب شود. برای دنباله کلمات i مجموعه دانش بازیابی شده را با $\mathcal{K}_{S_i}$ نشان می دهیم، که حاوی m_s بازنمایش دانش مرتبط به صورت $k_{1,s}, k_{2,s}, \dots, k_{m_s,s}$ است. برای ادغام دانش فوق الذکر در مدل پیشنهادی، با این چالش روبرو هستیم که تعداد زیادی تعبیه وجود دارد، درحالی که بیشتر آن‌ها در زمینه های خاص بی فایده هستند. برای حل این مشکل، برای دنباله کلمات i، وزن هر $k_z \in \mathcal{K}_{S_i}$ توسط روابط (۱۳ و ۱۴) محاسبه می شود. دانش کلی توسط رابطه (۱۵) بدست می آید.

$$w_z = \frac{e^{\beta k_z}}{\sum_{k_j \in \mathcal{K}_{S_i}} e^{\beta k_j}} \quad (13)$$

$$\beta_k = NN_{\beta}([g'_i, g'_j, k]) \quad (14)$$

$$o_{g'_i} = \sum_{k_z \in \mathcal{K}_{S_i}} w_z \cdot k_z \quad (15)$$

$$F(s_i, s_j) = f_m(s_i) + f_m(s_j) + f_c(s_i, s_j) \quad (16)$$

$$f_m(s_i) = NN_m([g'_i, o_{g'_i}]) \quad (17)$$

$$f_c(s_i, s_j) = NN_c([g'_i, o_{g'_i}, g'_j, o_{g'_j}, g'_i \odot g'_j, o_{g'_i} \odot o_{g'_j}]) \quad (18)$$

۲-۳- تشخیص عبارات وابسته به نهادها

پس از استخراج عبارات اسمی و ویژگی های آن‌ها، ماتریس

است، که به صورت $\sigma(x) = 1/(1 + e^{-x})$ تعریف می شود. $\sigma \in [-1, 1]$ جهت هر GRU را نشان می دهد و $\odot$ ضرب عنصر به عنصر دو بردار است. x_t^* خروجی الحاق شده ی Bi-GRU است. برای هر جمله از یک GRU مستقل استفاده می شود.

کلمه سر^۱ (کلمه مهم و کلیدی) عبارات اسمی را می توان با کمک تجزیه گر نحوی^۲ به دست آورد. هرچند که با تعمیم آن با کمک ساختار سلسله مراتب والد/فرزند در WordNet می توان آن‌ها به دسته های کلی تر تبدیل کرد. مثلاً در نام بری famous ransomware problems کلمه سر ransomware است که پس از تعمیم به دسته معنایی software تبدیل می شود. به جای استفاده از تجزیه گر نحوی، مدل پیشنهادی با استفاده از مکانیزم توجه^۳ [۴۰] وظیفه یافتن کلمات سر را طبق روابط (۸-۶) می آموزد، که در آن FFNN شبکه عصبی روبه جلو است.

$$\alpha_t = w_{\alpha} \cdot FFNN_{\alpha}(x_t^*) \quad (6)$$

$$a_{i,t} = \frac{\exp(\alpha_t)}{\sum_{k=START(i)}^{END(i)} \exp(\alpha_k)} \quad (7)$$

$$\hat{x}_i = \sum_{t=START(i)}^{END(i)} a_{i,t} \cdot x_t \quad (8)$$

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

$$g_i = [x_{START(i)}^*, x_{END(i)}^*, \hat{x}_i, \emptyset(i)] \quad (9)$$

برای در نظر گرفتن اطلاعات جهانی و اطلاعات نهادها، بازنمایش مورد انتظار g'_i برای دنباله کلمات i توسط روابط (۱۰-۱۲) محاسبه می شود.

$$g'_i = \sum_{j=1}^i Q(i \in E_j) \cdot e_j^{(i)} \quad (10)$$

$$e_i(t) = \sum_{j=1}^t Q(j \in E_i) \cdot g_j \quad (11)$$

$$Q(i \in E_j) = \begin{cases} \sum_{k=j}^{i-1} p(y_i = k) \cdot Q(k \in E_j) & j < i \\ p(y_i = \epsilon) & j = i \\ . & j > i \end{cases} \quad (12)$$

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 مورد نظر وجود داشته باشند، حداکثر m نهاد نیز وجود دارد). $e_i(t)$ بازنمایش یک نهاد در زمان t است.

برای تشخیص دقیق عبارات اسمی و اطلاعات وابسته به یک نهاد از منابع دانش مختلف استفاده می کنیم. برای استفاده از منابع دانش، آن‌ها را به یک فرمت یکسان تبدیل کردیم. فرمت در نظر

^۱ Head

^۵ Tail

^۶ Relation

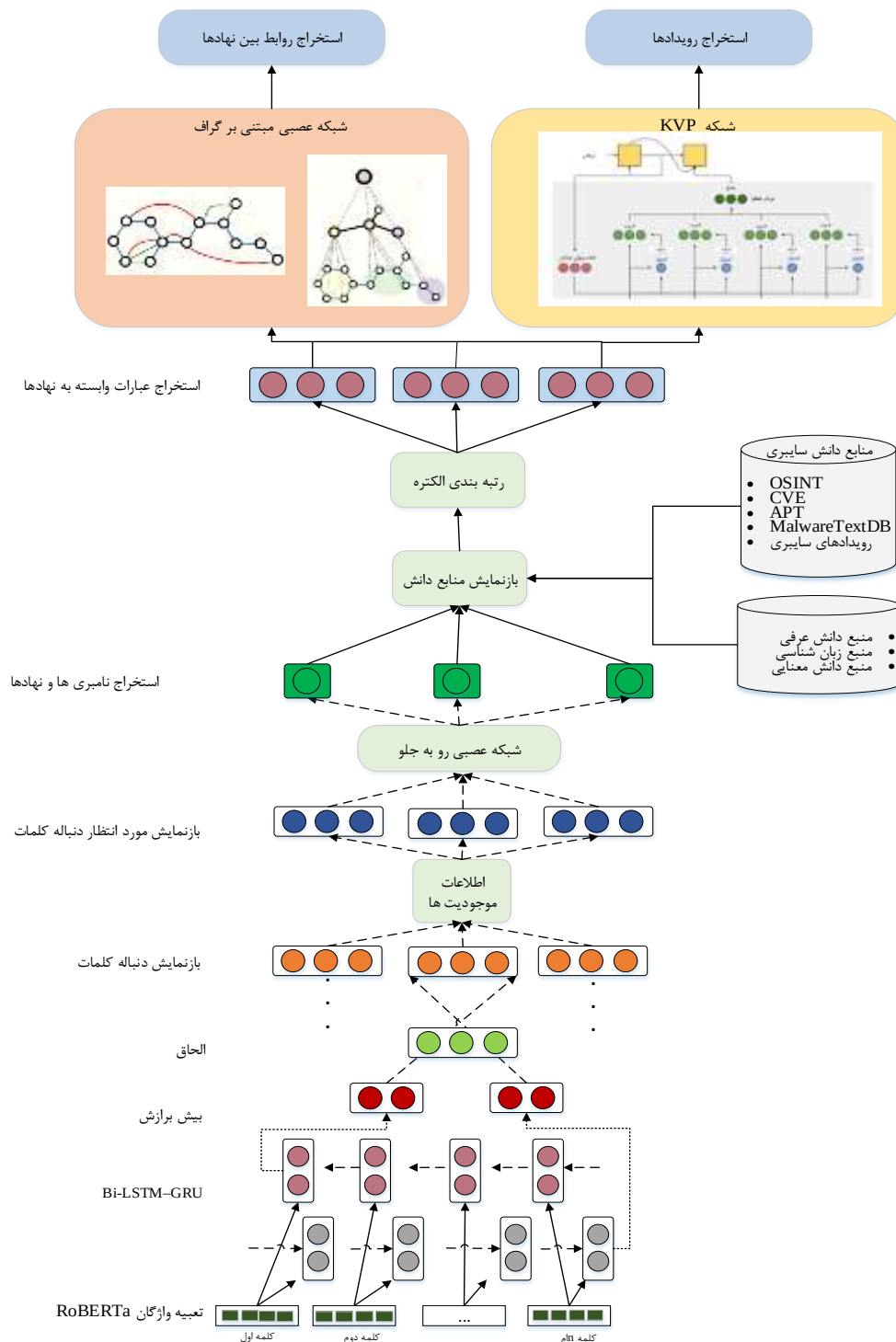
^۷ String Match

^۱ Head

^۲ Syntactic Parses

^۳ Attention Mechanism

تصمیم زیر برای عبارات کاندید شده جهت بررسی وابسته بودن یا نبودن با نهاد مدنظر تشکیل می‌شود.



شکل (۱): ساختار روش پیشنهادی

نام است. وزن یا میزان اهمیت هر یک از ویژگی‌ها توسط روش بردار ویژه و به صورت زیر محاسبه می‌شود.

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

$$\begin{aligned} r_{11}w_1 + r_{12}w_2 + \dots + r_{1n}w_n &= \lambda \cdot w_1 \\ r_{21}w_1 + r_{22}w_2 + \dots + r_{2n}w_n &= \lambda \cdot w_2 \end{aligned} \quad (20)$$

$$M = \begin{matrix} & f_1 & f_2 & \dots & f_n \\ \begin{matrix} g'_1 \\ g'_2 \\ \dots \\ g'_m \end{matrix} & \begin{bmatrix} r_{11} & r_{12} & \dots & r_{1n} \\ r_{21} & r_{22} & \dots & r_{2n} \\ \dots & \dots & \dots & \dots \\ r_{m1} & r_{m2} & \dots & r_{mn} \end{bmatrix} \end{matrix} \quad (19)$$

در ماتریس رابطه (۱۹) g'_i بازنمایش عبارت اسمی f_j ویژگی نام و مقدار استخراج شده گام قبل برای ویژگی نام عبارت

$$S_{ke} = \{j | v_{kj} \geq v_{ej}\} \quad (27)$$

$$S_{ke} = \{j | v_{kj} \leq v_{ej}\}$$

مجموعه معیارهای ناهماهنگ برای معیارهای مثبت و منفی به ترتیب به صورت زیر تعریف می شود.

$$\begin{aligned} D_{ke} &= \{j | v_{kj} < v_{ej}\} = J - S_{ke} \\ D_{ke} &= \{j | v_{kj} > v_{ej}\} = J - S_{ke} \end{aligned} \quad (28)$$

در این مرحله ماتریس هماهنگی تشکیل می شود. این ماتریس یک ماتریس مربعی است که بعد آن تعداد عبارات اسمی است. هر یک از درایه های این ماتریس، شاخص توافق بین دو عبارت نامیده می شود. مقدار این شاخص، از جمع وزن معیارهایی که در مجموعه هماهنگ وجود دارند، به دست می آید. به بیان دیگر 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 می آید. 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.

$$C_{ke} = \sum_{j \in S_{ke}} w'_j \quad (29)$$

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 درایه های این ماتریس، شاخص عدم توافق بین دو عبارت نامیده می شود. مقدار این شاخص از رابطه زیر به دست می آید.

$$d_{ke} = \frac{\max_{j \in D_{ke}} |v_{kj} - v_{ej}|}{\max_{j \in J} |v_{kj} - v_{ej}|} \quad (30)$$

در مرحله چهارم نحوه محاسبه شاخص هماهنگی (C_{ke}) بیان شد، در این مرحله یک مقدار معین برای این شاخص توافق مشخص می شود که آن را آستانه هماهنگی می نامند و با C^- نشان داده می شود. این مقدار از میانگین گیری درایه های ماتریس هماهنگی به دست می آید. به زبان ریاضی مقدار آستانه هماهنگی از رابطه زیر محاسبه می شود.

$$\bar{C} = \sum_{k=1}^m \sum_{e=1}^m \frac{C_{ke}}{m(m-1)} \quad (31)$$

ماتریس تسلط هماهنگی (F) باتوجه به مقدار آستانه موافقت تشکیل می شود. اگر C_{ke} بزرگ تر از C^- باشد، برتری عبارت k بر عبارت e قابل قبول است در غیر این صورت عبارت k بر عبارت e برتری ندارد لذا درایه های این ماتریس تسلط از رابطه زیر تعیین می شود.

$$f_{ke} = \begin{cases} 1 & C_{ke} \geq \bar{C} \\ 0 & C_{ke} < \bar{C} \end{cases} \quad (32)$$

ماتریس تسلط ناهماهنگی (G) مانند F تشکیل می شود. بدین منظور ابتدا باید مقدار آستانه ناهماهنگی ($\bar{d}$) از میانگین گیری درایه های ماتریس ناهماهنگی محاسبه شود. به زبان ریاضی مقدار آستانه ناهماهنگی از رابطه زیر محاسبه می شود.

$$\bar{d} = \sum_{k=1}^m \sum_{e=1}^m \frac{d_{ke}}{m(m-1)} \quad (33)$$

$$r_{n1}w_1 + r_{n2}w_2 + \dots + r_{nn}w_n = \lambda \cdot w_n$$

وزن شاخص ها (ویژگی ها) طبق رابطه (۲۱) محاسبه می شود.

$$w_j = \frac{1}{\lambda} \sum_{i=1}^n r_{ij}w_i \quad j = 1, 2, \dots, n \quad (21)$$

دستگاه معادلات فوق را به صورت رابطه (۲۲) می توان نشان داد.

$$M \cdot w = \lambda \cdot w \quad (22)$$

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

۱- ماتریس M تشکیل داده می شود.

۲- ماتریس ($M - \lambda I$) تشکیل داده می شود.

۳- دترمینان ماتریس ($M - \lambda I$) را محاسبه کرده، مساوی

صفر قرار داده و مقادیر λ محاسبه می شود.

بزرگ ترین λ را λ_{max} می نامیم و آن را در رابطه (۲۳)

قرار داده و مقادیر w_j ها محاسبه می شود.

$$(M - \lambda_{max}I) \cdot w = 0 \quad (23)$$

پس از محاسبه وزن ویژگی ها برای عبارات کاندید، طبق رابطه (۲۴) وزن عبارات جهت بررسی هم مرجع بودن یا نبودن با نهاد موردنظر به روزرسانی می شود.

$$w'_j = \frac{r_j w_j}{\sum_{j=1}^n r_j w_j} \quad (24)$$

رتبه بندی عبارات کاندید: پس از تشکیل ماتریس تصمیم و محاسبه وزن و مقدار ویژگی ها، عبارات کاندید باید برای نهاد موردنظر رتبه بندی شوند. جهت رتبه بندی عبارات از روش تصمیم گیری چندمعیاره ELECTRE [۴۱] و به صورت زیر استفاده می شود.

ابتدا ماتریس تصمیم بی مقیاس می شود. برای تبدیل ویژگی های چندبعدی به ویژگی های بدون بعد و از بین بردن تفاوت مقیاسی داده های تصمیم گیری، از نرم اقلیدسی رابطه (۲۵)، استفاده می کنیم.

$$n_{ij} = \frac{r_{ij}}{\sqrt{\sum_{i=1}^m r_{ij}^2}} \quad (25)$$

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

$$v_{ij} = w'_j * n_{ij} \quad i = 1, 2, \dots, m; \quad j = 1, 2, \dots, n \quad (26)$$

برای هر زوج عبارت اسمی k و e مجموعه معیارها به دو زیرمجموعه هماهنگ و ناهماهنگ تقسیم می شوند. مجموعه هماهنگ (S_{ke}) مجموعه ای 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 به صورت زیر تعریف می شوند.

$$P(r_i|e_h, e_t) = \frac{\exp(l_f^{(h,t)})}{\sum_{r' \in \mathcal{P}_T \cup \{TH\}} \exp(l_f^{(h,t)'})} \quad (41)$$

$$\mathcal{L}_1 = - \sum_{r \in \mathcal{P}_T} \log(P(r_i|e_h, e_t)) \quad (42)$$

$$\mathcal{L}_2 = -\log\left(\frac{\exp(l_f^{(h,t)TH})}{\sum_{r' \in \mathcal{N}_{TU(TH)}} \exp(l_f^{(h,t)'})}\right) \quad (43)$$

$$\mathcal{L} = \mathcal{L}_1 + \mathcal{L}_2 \quad (44)$$

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 نهادهای وجود ندارند. اگر T شامل هیچ رابطه‌ای نباشد $\mathcal{N}_T = R$. طبق تابع هزینه تعریف روابط درست بین نهادهای استخراج می‌شود.

۳-۴- تشخیص رویدادها

جهت تشخیص رویدادهای امنیتی از شبکه عمیق KVP استفاده شده 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 به توالی کمک می‌کنند. این لایه‌ها از بردارهای زمینه و وزن‌های توجه تشکیل شده‌اند و می‌توانند همبستگی بین دو عنصر که موقعیت‌های متفاوتی در یک دنباله ورودی دارند را شناسایی کنند. 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 بازه زمانی t است، توزیع احتمال $\mathcal{Y}_t$ برای رویداد بعدی به صورت زیر محاسبه می‌شود.

$$\begin{bmatrix} k_t \\ v_t \\ p_t \end{bmatrix} = h_t \in \mathbb{R}^{3k} \quad (45)$$

$$M_t = \emptyset(W^Y[k_{t-L} \dots k_{t-1}] + (W^h h_t)) \in \mathbb{R}^{k \times L} \quad (46)$$

$$\alpha_t = \text{sotmax}(W^T M_t) \in \mathbb{R}^{1 \times L} \quad (47)$$

$$r_t = [v_{t-L} \dots v_{t-1}] a^T \in \mathbb{R}^k \quad (48)$$

$$h_t^* = \emptyset(W^r r_t + W^x p_t) \in \mathbb{R}^k \quad (49)$$

$$y_t = \text{sotmax}(W^* h_t^* + b) \in \mathbb{R}^{|\mathcal{V}|} \quad (50)$$

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

همان گونه 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 (d^-) بزرگ‌تر باشد، میزان مخالفت زیاد بوده و نمی‌توان از آن صرف نظر کرد بنابراین درایه‌های ماتریس G به صورت زیر محاسبه می‌شود.

$$g_{ke} = \begin{cases} 1 & d_{ke} > \bar{d} \\ 0 & d_{ke} \leq \bar{d} \end{cases} \quad (34)$$

هر عضو این ماتریس نشانگر رابطه تسلط مابین عبارات اسمی است. در نهایت ماتریس تسلط نهایی (H) از ضرب تک تک 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 قابل قبول 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 مورد نظر رتبه‌بندی و طبق حد آستانه استخراج می‌شوند.

$$h_{ke} = f_{ke} \cdot g_{ke} \quad (35)$$

۳-۳- تشخیص روابط بین نهادهای

پس از استخراج نهادهای عبارات وابسته به هر نهاد باید روابط بین نهادهای شناسایی شود. برای حل این مسئله از شبکه عصبی عمیق مبتنی بر گراف استفاده نمودیم. براین اساس، بردار ویژگی استخراج شده طبق رابطه (۳۶) به صورت ورودی به واحد فعال سازی داده می‌شود.

$$h_v^* = X_v(h_{ke}) \quad (36)$$

سپس برای هر لایه رابطه (۳۷) اعمال می‌شود.

$$h_v^k = \sigma(W_k \sum \frac{h_v^{k-1}}{N(v)} + B_k h_v^{k-1}) \text{ where } k = 1, \dots, k-1 \quad (37)$$

دو مرحله برای این معادله وجود دارد:

مرحله اول میانگین‌گیری تمام همسایگان گره v است.

$$W_k \sum \frac{h_v^{k-1}}{N(v)} \quad (38)$$

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

$$B_k h_v^{k-1} \quad (39)$$

σ فعال ساز غیر خطی است که روی دو قسمت انجام می‌شود.

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

$$Z_v = h_v^k \quad (40)$$

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

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 توییت [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 به سرعت قابلیت‌های بدافزار را درک کنند. در این پایگاه از منابع متنوعی از جمله متن بدون برچسب، متون حاشیه‌نویسی انسانی و خصوصیات در مورد ویژگی‌های بدافزار برای شناسایی اسناد بدافزار استفاده شده 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^۵: CVE فهرستی از آسیب‌پذیری‌های امنیتی رایانه‌ای است که به صورت عمومی افشا شده‌اند [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 استفاده شده است.

دانش عرفی: در Conceptnet دانش عرفی در قالب جملات زبان طبیعی ذخیره می‌شود، از این رو 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 به طور منظم توسط کاربران به روزرسانی و کنترل می‌شود. این منبع شامل ۶۰۰ هزار سه‌تایی است، به عنوان مثال (بدافزار، حاوی، فایل آلوده) و (شرکت، استفاده می‌کند، تکنیک هجومی). تمام روابط در این منبع به صورت دستی مشخص شده است. در این مقاله سه‌تایی‌های با امتیاز بالا (بالتر از ۳) انتخاب شده 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 گرفته شده است. تجزیه‌کننده^۸ استنفورد برای

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 به عنوان کلید و v_t به عنوان 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 v_t استفاده می‌شود.

۴- نتایج شبیه‌سازی

در این بخش روش پیشنهادی از جنبه‌های مختلف مورد بررسی قرار گرفته و با سایر روش‌ها مقایسه شده است. آزمایش‌ها روی رایانه شخصی با GPU GeForce RTX-۲۰۸۰ و پردازنده Core i7-97۰۰K با فرکانس ۴۴۰۰ مگاهرتز با استفاده از نرم‌افزار python انجام شده است. لایه‌های پنهان GRU، ۲۵۰ بعد^۱ دارند. شبکه عصبی روبه‌جلو شامل دولا به پنهان با ۲۰۰ بعد است. شبکه‌های عصبی عمیق مدل پیشنهادی روی پیکره‌های توسعه CoNLL-۲۰۱۲ و OSINT با ۱۹ تکرار^۲ و حذف تصادفی^۳ ۰/۴ و نرخ‌های یادگیری^۴ 10^{-4} و 10^{-3} و اندازه دسته ۶۴ تنظیم دقیق^۵ شده‌اند. منابع دانش استفاده شده به شرح زیر است:

منبع دانش رویدادهای سایبری^۶: 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 ساختار یافته لازم ایجاد شده 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 استفاده شده است که حملات منتسب به عموم را شناسایی و مورد بحث قرار می‌دهند. داده‌ها ماهانه به روز می‌شوند و اطلاعاتی در مورد عامل تهدید، انگیزه، قربانی و اثرات نهایی حمله به دست می‌دهند. سه‌تایی‌های استخراج شده از این منبع ۲۲۳۴ تا است، به عنوان مثال سه‌تایی (باج افزار، است، یک بدافزار).

^۱ Dimensions

^۲ Epochs

^۳ Dropout

^۴ Learning Rates

^۵ Fine-Tune

^۶ Cyber Events

^۷ Common Vulnerabilities and Exposures

^۸ Parser

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 بالا برده است. روش‌های سان و مونر بعد از روش پیشنهادی به ترتیب بالاترین مقادیر مربوط به $F1$ را دارند. 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. از آنجاکه برای بازنمایش دانش در روش 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 قرار گرفته 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، در روش پیشنهادی در نظر گرفته شوند.

در جدول (۱) نرخ تشخیص روابط بین نهادها برای روش پیشنهادی طبق معیارهای MUC [۵۳]، B^3 [۲۹] و $CEAF_{\phi_4}$ [۵۵] توسط مقادیر دقت و بازخوانی و $Avg.F1$ (میانگین خروجی سه معیار MUC، B^3 و $CEAF_{\phi_4}$) طبق روابط (۵۳-۵۱) با روش‌های مطرح و جدید استخراج اطلاعات روی پایگاه داده [۵۶] CONLL-۲۰۱۲ 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 $CEAF_{\phi_4}$ 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

برای تولید تعداد، جاننداری و جنسیت برای همه عبارات اسمی به کار می‌رود که می‌تواند به‌طور خودکار دانش زبانی (به‌صورت سه-تایی) برای داده‌های ما تولید کند. ویژگی تعداد شامل مقدار s برای مفرد و p مقدار برای جمع است. ویژگی جاننداری نشان-دهنده جاننداری که در صورت جاننداری، 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 جاننداری، خنثی) استفاده می‌شود. در نتیجه ۴۰۱۴۹ سه‌تایی برای جمع بودن و ۴۰۴۶۲ سه‌تایی برای جنسیت و جاننداری در نظر گرفته می‌شود.

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

$$P_r(a|p) = \frac{Count_r(p,a)}{Count_r(p)} \quad (51)$$

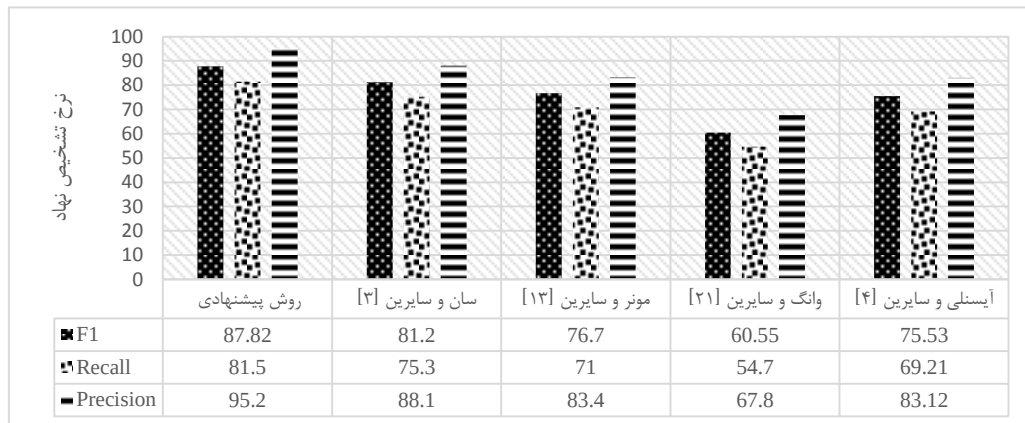
در رابطه (۵۱) $Count_r(p)$ و $Count_r(p,a)$ به ترتیب نشان‌دهنده این هستند که چند بار p و جفت گزاره-آرگومان (p,a) در رابطه r وجود داشته‌اند. در آزمایش‌ها زمانی که $P_r(a|p) > 0.1$ و $Count_r(p,a) > 10$ بود، سه‌تایی (p,r,a) را در نظر گرفتیم. به‌عنوان مثال، ("فایل آلوده"، nsubj، تکنیک هجومی) یک رابطه معنایی 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 [۵۲] English Gigaword 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 در روش پیشنهادی از مدل‌های زبانی، توسط روش پیش آموزش

^۱ Predicate^۲ Argument

بازگشتی (RNN) برای مدیریت وابستگی‌های طولانی مدت استفاده می‌شود، اما مشکل چرخه‌های طولانی آموزش را دارد، LSTM نسبت به این نوع شبکه‌ها وابستگی طولانی مدت را بهتر مدیریت می‌کند. شبکه‌های عصبی مبتنی بر 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 نمی‌کند.

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 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



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

جدول (۱): مقایسه F1 در زمینه تشخیص روابط بین نهادها روی پیکره آزمایش CoNLL-۲۰۱۲ انگلیسی

روش	MUC			B ³			CEAF_φ ₄			Avg.F1
	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	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

همان‌طور که مشخص است کارکرد روش BERT از روش ELMo اندکی بهتر است. روش ELMo در ساختار خود از مدل‌های زبانی دوجزته 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^۱) و مکانیزم پیش‌بینی

برای بررسی میزان تأثیر روش‌های مختلف تعبیه واژگان نظیر BERT، ELMo، ROBERTA و XLNet بر عملکرد روش پیشنهادی و مقدار Avg.F1، این روش‌ها در جدول (۲) روی مجموعه توسعه پیکره CoNLL-۲۰۱۲ با یکدیگر مقایسه شده‌اند.

جدول (۲): نتایج ارزیابی تعبیه واژگان مختلف روی پیکره توسعه

CoNLL-۲۰۱۲

	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

^۱ Masked Language Model

در شکل (۳) جهت سنجش عملکرد افزودن ماژول‌های مختلف به سیستم درزمینه استخراج اطلاعات (میانگین F۱ تشخیص نهاد، روابط بین نهادها و رویدادها)، مقدار F۱ تحت مقادیر دقت و بازخوانی روی پیکره امنیتی OSINT نمایش داده شده 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 F۱ افزایش یافته است.

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

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

سایز حالت پنهان^۲ در شبکه‌های GRU، ۳۰۰ در نظر گرفته شده 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. یک لایه Dropout برای جلوگیری از بیش‌برازش پس از لایه Bi-GRU قرار دادیم. با این عمل برخی از ورودی‌ها به لایه بعد صفر می‌شوند. در اینجا Dropout، ۰/۵ در نظر گرفته شده 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 به تبع آن F۱ نیز بیشتر می‌گردد. افزایش میزان اطلاعات از محتوای متنی که واژگان در آن قرار گرفته است با استفاده از Bi-GRU و یک لایه GRU بر روی خروجی آن، موجب افزایش فراخوانی می‌گردد و به تبع آن F۱ نیز افزایش می‌یابد. همان‌طور که مشخص است استفاده از ویژگی‌های استخراج شده توسط شبکه‌ها تأثیر قابل‌قبولی بر بالا رفتن دقت استخراج روابط و رویدادها دارد. همچنین، افزودن اطلاعات سطح نهاد و پس‌از آن افزودن منابع دانش مختلف نظیر رویدادها و حملات سایبری، توییت‌های مربوط به امنیت سایبری، گزارش‌های بدافزارها، تهدیدات سایبری، آسیب‌پذیری‌های امنیتی، منابع معنایی، عرفی، زبان‌شناسی و در نتیجه 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 در روش پیشنهادی استفاده شده 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 F۱ 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 داده شده است. همان‌طور که مشخص است هر دو روش زمانی که داده‌های آموزش از پیکره OSINT و داده‌های آزمودن از پیکره CoNLL-۲۰۱۲ است، عملکرد

جمله بعدی (NSP^۱) ارائه داده است. دوماً، برای آموزش BERT از تعداد زیادی داده و قدرت محاسباتی بالا استفاده شده است. MLM امکان یادگیری دوطرفه از متن را ممکن می‌سازد، یعنی به مدل اجازه می‌دهد تا مفهوم هر کلمه را از کلماتی که قبل و بعد از آن ظاهر می‌شود بیاموزد. این کار در ELMo امکان‌پذیر نیست، زیرا از مدل دوطرفه کم‌عمق استفاده می‌کند. استفاده از تعبیه واژگان XLNet عملکرد بهتری (درصد) نسبت به روش BERT دارد. XLNet یک مدل 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 آن همه توکن‌ها به صورت تصادفی پیش‌بینی می‌شوند. این برخلاف مدل زبان پوششی BERT است که در آن فقط پانزده درصد توکن‌های پوشش‌دار پیش‌بینی شده است. در نهایت 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 بدترین حالت را در نظر می‌گیرد. در روش TOPSIS گزینه انتخابی باید کمترین فاصله از جواب ایده آل و دورترین فاصله از جواب ضد ایده آل را داشته باشد. روش TOPSIS 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

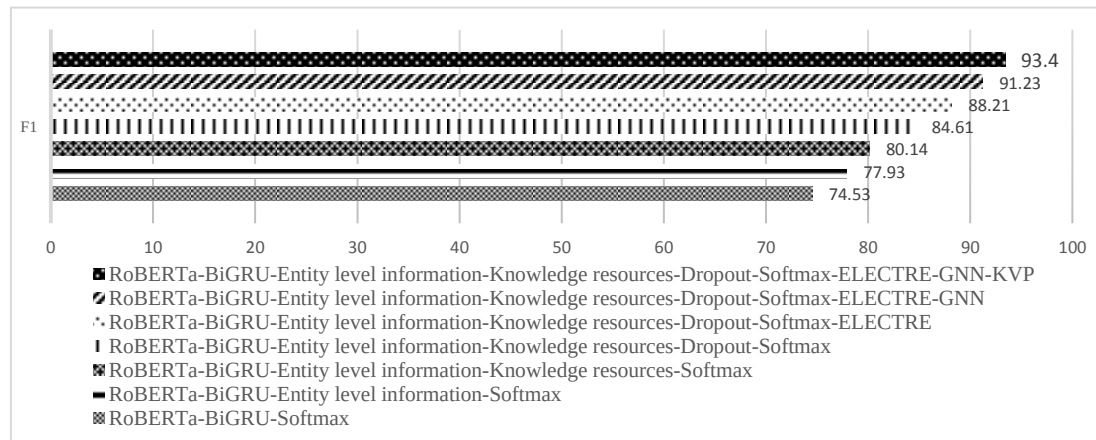
دقت رتبه‌بندی (درصد)	روش رتبه‌بندی
۹۲/۶	ELECTRE
۸۹/۳	VIKOR
۸۶/۷	TOPSIS

^۲ Hidden State Size

^۱ Next Sentence Prediction

۲۰۱۲ و داده‌های آزمودن از پیکره OSINT باشند، روش پیشنهادی عملکرد قابل قبولی (۸/۵+) در مقایسه با روش سان و سایرین دارد که نشان‌دهنده توانایی روش رتبه‌بندی در این پیکره است.

ضعیفی دارند اما عملکرد روش پیشنهادی نسبت به روش سان بهتر است. این نشان می‌دهد که دانش سایبری تأثیر کمی در پیکره CoNLL-۲۰۱۲ دارد. زمانی که داده‌های آموزش از پیکره CoNLL-



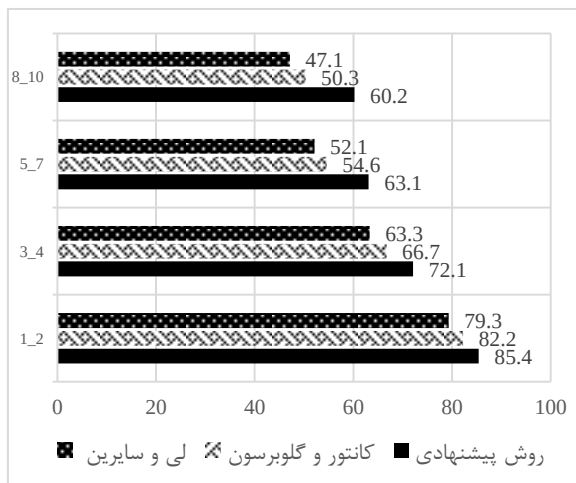
شکل (۳): دقت استخراج اطلاعات با افزودن ماژول‌های مختلف به مدل پیشنهادی روی پیکره OSINT

جدول (۴). ارزیابی عملکرد روش سان و همکاران و روش پیشنهادی

توسط داده‌های آموزشی مختلف روی پیکره‌های مختلف

روش	داده‌های آموزش	داده‌های آزمودن	
		CoNLL-۲۰۱۲	OSINT
روش سان و همکاران	CoNLL-۲۰۱۲	۷۳/۸	۷۶/۷
	OSINT	۵۱/۴	۹۰/۱
روش پیشنهادی	CoNLL-۲۰۱۲	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 در روش 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 به جای ELMo استفاده کرده است و بازنمایش دقیق‌تری برای دنباله کلمات ارائه داده است. هرچند در روش 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 قابل تشخیص

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 در نظر گرفتن 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-۲۰۱۲ 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 در زمینه تشخیص نهادها و روابط بین آن‌ها به ترتیب ۶/۶۲ و ۷/۲ درصد افزایش داده است. دقت استخراج اطلاعات روی پیکره امنیتی OSINT باتوجه به 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

- [۱] L. Zongxun, L. Yujun, Z. Haojie, and L. Juan, "Construction of tps from apt reports using bert," in ۲۰۲۱ ۱۸th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP), IEEE, ۲۰۲۱, pp. ۲۶۰-۲۶۳. Accessed: Apr. ۱۵, ۲۰۲۴. [Online]. Available: <https://ieeexplore.ieee.org/abstract/document/9674158>
- [۲] K. Dadashtabar Ahmadi, M. Kheirkhah, A. J. Rashidi, "Detection of advanced Cyber Attacks, Using Behavior Modeling Based on Natural Language Processing", ECD, Vol. ۶, No. ۳, Serial No. ۲, pp. ۱۴۱-۱۵۱, ۲۰۱۸. doi: ۲۰.۱۰۰۱.۱.۲۳۲۲۴۲۴۷.۱۳۹۷.۶.۳.۱۲.۲
- [۳] N. Sun et al., "Cyber threat intelligence mining for proactive cybersecurity defense: a survey and new perspectives," IEEE Communications Surveys & Tutorials, ۲۰۲۳, Accessed: Apr. ۲۰, ۲۰۲۴. [Online]. Available: <https://ieeexplore.ieee.org/abstract/document/10117505/>

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 ۲۰۱۲-CONLL 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 ۲۰۱۲-CONLL بالاترین تأثیر را نسبت به بقیه دانش‌ها دارد. درحالی‌که در پیکره سایبری OSINT حذف منبع دانش و اطلاعات پیکره CASIE بیشترین تأثیر منفی را بر عملکرد مدل پیشنهادی می‌گذارد.

جدول (۵). نتایج حذف منابع دانش روی پیکره‌های ۲۰۱۲-CONLL و OSINT

پایگاه داده	CONLL-۲۰۱۲		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، سریع‌تر به حوادث واکنش نشان دهند و از اقداماتی که عوامل تهدید انجام می‌دهند، جلوگیری کنند. 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، گزارش‌های بدافزارها، تهدیدات سایبری، آسیب‌پذیری‌های امنیتی)،

- [۱۵] G. Tür, D. Hakkani-Tür, and K. Oflazer, "A statistical information extraction system for Turkish," *Nat. Lang. Eng.*, vol. ۹, no. ۲, pp. ۱۸۱–۲۱۰, ۲۰۰۳. doi: ۱۰.۱۰۱۷/S۱۳۵۱۳۲۰۹۰۲۰۲۸۴X.
- [۱۶] J. Zhang, "Entropic Statistics: Concept, Estimation, and Application in Machine Learning and Knowledge Extraction," *Mach. learn. knowl. extr.*, vol. ۴, no. ۴, pp. ۸۶۵–۸۸۷, ۲۰۲۲. <https://doi.org/۱۰.۳۳۹۰/make۴۰۴۰۰۴۴>.
- [۱۷] 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 ۲۰۱۸ International Conference on Frontiers of Information Technology (FIT), IEEE, ۲۰۱۸, pp. ۱۲۹–۱۳۴. Accessed: Apr. ۲۲, ۲۰۲۴. [Online]. Available: <https://ieeexplore.ieee.org/abstract/document/۸۶۱۶۹۷۹>
- [۱۸] K. Narasimhan, A. Yala, and R. Barzilay, "Improving Information Extraction by Acquiring External Evidence with Reinforcement Learning," *arXiv*, Sep. ۲۷, ۲۰۱۶. Accessed: Oct. ۲۰, ۲۰۲۳. [Online]. Available: <http://arxiv.org/abs/۱۶۰۳.۰۷۹۵۴>. <https://doi.org/۱۰.۴۸۵۵۰/arXiv.۱۶۰۳.۰۷۹۵۴>
- [۱۹] W. Y. Wang, J. Li, and X. He, "Deep reinforcement learning for NLP," in *Proceedings of the ۵۶th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, ۲۰۱۸, pp. ۱۹–۲۱. Accessed: Oct. ۲۰, ۲۰۲۳. [Online]. Available: <https://aclanthology.org/P۱۸-۵۰۰۷/>
- [۲۰] X. Wang, J. Yang, Q. Wang, and C. Su, "Threat Intelligence Relationship Extraction Based on Distant Supervision and Reinforcement Learning," in *SEKE*, ۲۰۲۰, pp. ۵۷۲–۵۷۶. Accessed: Apr. ۲۲, ۲۰۲۴. [Online]. Available: <https://ksiresearch.org/seke/seke۲۰/paper/paper۱۴۹.pdf>
- [۲۱] X. Wang et al., "A method for extracting unstructured threat intelligence based on dictionary template and reinforcement learning," in ۲۰۲۱ IEEE ۲۴th International Conference on Computer Supported Cooperative Work in Design (CSCWD), IEEE, ۲۰۲۱, pp. ۲۶۲–۲۶۷. Accessed: Apr. ۲۲, ۲۰۲۴. [Online]. Available: <https://ieeexplore.ieee.org/abstract/document/۹۴۳۷۸۵۸>
- [۲۲] 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. ۱۲, no. ۱۹, p. ۹۶۹۱, ۲۰۲۲. <https://doi.org/۱۰.۳۳۹۰/app۱۲۱۹۹۶۹۱>.
- [۲۳] H. Jo, Y. Lee, and S. Shin, "Vulcan: Automatic extraction and analysis of cyber threat intelligence from unstructured text," *COMPUT SECUR*, vol. ۱۲۰, p. ۱۰۲۷۶۳, ۲۰۲۲. <https://doi.org/۱۰.۱۰۱۶/j.cose.۲۰۲۲.۱۰۲۷۶۳>.
- [۲۴] X. Wang et al., "Cyber threat intelligence entity extraction based on deep learning and field knowledge engineering," in ۲۰۲۲ IEEE ۲۵th International Conference on Computer Supported Cooperative Work in Design (CSCWD), IEEE, ۲۰۲۲, pp. ۴۰۶–۴۱۳. Accessed: Apr. ۲۲, ۲۰۲۴. [Online].
- [۴] 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. ۱۰۳۳۵۲, ۲۰۲۳. <https://doi.org/۱۰.۱۰۱۶/j.cose.۲۰۲۳.۱۰۳۳۵۲>.
- [۵] M. H. HassanNia, M. R. HasaniAhangar, A. Gafari, "An Improved Method of Incident Detection due to Cyber Attacks, ECD, Vol. ۷, No. ۴, ۲۰۲۰. Available: <https://sid.ir/paper/۳۹۵۷۲۵/en>.
- [۶] E. Bastami, H. Soltanizadeh*, M. Rahmanimanesh, P. Keshavarzi, "A Malware Classification Method Using visualization and Word Embedding Features", *ECD*, Vol. ۱۱, No. ۱, ۲۰۲۳. doi: ۲۰.۱۰۰۱.۱.۲۳۲۲۴۳۷.۱۴۰۲.۱۱.۱.۲.
- [۷] K. Lee, L. He, and L. Zettlemoyer, "Higher-order coreference resolution with coarse-to-fine inference," *arXiv preprint arXiv:۱۸۰۴.۰۵۳۹۲*, ۲۰۱۸. <https://doi.org/۱۰.۴۸۵۵۰/arXiv.۱۸۰۴.۰۵۳۹۲>.
- [۸] H. Peng, D. Khashabi, and D. Roth, "Solving hard coreference problems," *arXiv preprint arXiv:۱۹۰۷.۰۵۵۲۴*, ۲۰۱۹. <https://doi.org/۱۰.۴۸۵۵۰/arXiv.۱۹۰۷.۰۵۵۲۴>.
- [۹] 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. ۱۳۵, p. ۱۰۴۱۰۸, ۲۰۲۲. <https://doi.org/۱۰.۱۰۱۶/j.autcon.۲۰۲۱.۱۰۴۱۰۸>.
- [۱۰] 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. ۱۲, no. ۲, pp. ۱۲۱–۱۳۲, ۲۰۲۱. doi: ۱۰.۲۴۵۰۷/icicelb.۱۲.۰۲.۱۲۱
- [۱۱] 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*, ۲۰۲۲, pp. ۵۲۴–۵۳۳. Accessed: Oct. ۲۶, ۲۰۲۳. [Online]. Available: <https://aclanthology.org/۲۰۲۲.lrec-۱.۵۵/>
- [۱۲] F. Rahma and A. Romadhony, "Rule-Based Crime Information Extraction on Indonesian Digital News," in ۲۰۲۱ International Conference on Data Science and Its Applications (ICoDSA), IEEE, ۲۰۲۱, pp. ۱۰–۱۵. Accessed: Oct. ۲۶, ۲۰۲۳. [Online]. Available: <https://ieeexplore.ieee.org/abstract/document/۹۶۱۷۵۰۹>
- [۱۳] K. Shaukat, S. Luo, S. Chen and D. Liu, "Cyber Threat Detection Using Machine Learning Techniques: A Performance Evaluation Perspective," ۲۰۲۰ International Conference on Cyber Warfare and Security (ICWS), Islamabad, Pakistan, pp. ۱–۶, ۲۰۲۰. doi: ۱۰.۱۱۰۹/ICCWS۴۸۴۳۲.۲۰۲۰.۹۲۹۲۳۸۸.
- [۱۴] 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. ۷, no. ۱, p. ۱۵۷۶, ۲۰۱۷. <https://doi.org/۱۰.۱۰۳۸۷/s۴۱۵۹۸۰۱۷۰۱۷۴۷۰۸>.

Accessed: Oct. ۲۶, ۲۰۲۳. [Online]. Available: <http://arxiv.org/abs/۲۲۰۳.۱۲۲۷۷>.

Available: <https://ieeexplore.ieee.org/abstract/document/۹۷۶۱۳۹/>

[۲۵] I.-H. Hsu et al., "DEGREE: A Data-Efficient Generation-Based Event Extraction Model." arXiv, May ۰۳, ۲۰۲۲. Accessed: Oct. ۲۶, ۲۰۲۳. [Online]. Available: <http://arxiv.org/abs/۲۱۰۸.۱۲۷۲۴>.

[۲۵] K. Ahmed, S. K. Khurshid, and S. Hina, "CyberEntRel: Joint extraction of cyber entities and relations using deep learning," *COMPUT SECUR*, vol. ۱۳۶, p. ۱۰۳۵۷۹, ۲۰۲۴. <https://doi.org/۱۰.۱۰۱۶/j.cose.۲۰۲۳.۱۰۳۵۷۹>.

[۲۶] J. Gao, H. Zhao, C. Yu, and R. Xu, "Exploring the Feasibility of ChatGPT for Event Extraction." arXiv, Mar. ۰۹, ۲۰۲۳. Accessed: Oct. ۲۶, ۲۰۲۳. [Online]. Available: <http://arxiv.org/abs/۲۳۰۳.۰۳۸۳۶>.

[۲۶] 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. ۱۴۹, pp. ۱۵۰–۱۵۶, ۲۰۲۱. <https://doi.org/۱۰.۱۰۱۶/j.patrec.۲۰۲۱.۰۶.۰۱۲>.

[۲۷] 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*, ۲۰۲۱, pp. ۱۴۳۴۷–۱۴۳۵۵. Accessed: Oct. ۲۶, ۲۰۲۳. [Online]. Available: <https://ojs.aaai.org/index.php/AAAI/article/view/۱۷۶۸۷>

[۲۷] C. Park, J. Park, and S. Park, "AGCN: Attention-based graph convolutional networks for drug-drug interaction extraction," *Expert Syst. Appl.*, vol. ۱۵۹, p. ۱۱۳۵۳۸, ۲۰۲۰. <https://doi.org/۱۰.۱۰۱۶/j.eswa.۲۰۲۰.۱۱۳۵۳۸>.

[۲۸] 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 ۵۹th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume ۱: Long Papers)*, ۲۰۲۱, pp. ۲۸۰۷–۲۸۱۸. Accessed: Oct. ۲۶, ۲۰۲۳. [Online]. Available: <https://aclanthology.org/۲۰۲۱.acl-long.۲۱۸/>

[۲۸] 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. ۱۳, no. ۸, p. ۱۴۵۸, ۲۰۲۱. <https://doi.org/۱۰.۳۳۹۰/sym۱۳۰۸۱۴۵۸>.

[۲۹] C. Zheng, J. Feng, Z. Fu, Y. Cai, Q. Li, and T. Wang, "Multimodal Relation Extraction with Efficient Graph Alignment," in *Proceedings of the ۲۹th ACM International Conference on Multimedia, Virtual Event China: ACM*, Oct. ۲۰۲۱, pp. ۵۲۹۸–۵۳۰۶. doi: ۱۰.۱۱۴۵/۳۴۷۴۰۸۵.۳۴۷۶۶۸.

[۲۹] 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. ۱۲۷۱۳, K. Karlapalem, H. Cheng, N. Ramakrishnan, R. K. Agrawal, P. K. Reddy, J. Srivastava, and T. Chakraborty, Eds., *LECT NOTES ARTIF INT*, vol. ۱۲۷۱۳, Cham: Springer International Publishing, ۲۰۲۱, pp. ۲۶۹–۲۸۱. doi: ۱۰.۱۰۰۷/۹۷۸-۳-۰۳۰-۷۵۷۶۵۰۶_۲۲.

[۳۰] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," arXiv preprint arXiv:۱۴۰۹.۰۴۷۳, ۲۰۱۴. <https://doi.org/۱۰.۴۸۵۵۰/arXiv.۱۴۰۹.۰۴۷۳>.

[۳۰] S. Guo, L. Huang, G. Yao, Y. Wang, H. Guan, and T. Bai, "Extracting Biomedical Entity Relations using Biological Interaction Knowledge," *INTERDISCIP SCI*, vol. ۱۳, no. ۲, pp. ۳۱۲–۳۲۰, Jun. ۲۰۲۱, doi: ۱۰.۱۰۰۷/۱۲۵۳۹۰۰۲۱۰۰۴۲۵۰۸.

[۳۱] A. T. de Almeida, "Multicriteria decision model for outsourcing contracts selection based on utility function and ELECTRE method," *Comput. Oper. Res.*, vol. ۳۴, no. ۱۲, pp. ۳۵۶۹–۳۵۷۴, ۲۰۰۷. <https://doi.org/۱۰.۱۰۱۶/j.cor.۲۰۰۶.۰۱.۰۰۳>.

[۳۱] S. Raza and B. Schwartz, "Entity and relation extraction from clinical case reports of COVID-۱۹: a natural language processing approach," *BMC MED INFORM DECIS*, vol. ۲۳, no. ۱, p. ۲۰, Jan. ۲۰۲۳, doi: ۱۰.۱۱۸۶/s۱۲۹۱۱۰۰۲۳۰۰۲۱۱۷-۳.

[۳۲] A. Vaswani et al., "Attention is all you need," *Advances in neural information processing systems*, vol. ۳۰, ۲۰۱۷, Accessed: Jun. ۰۴, ۲۰۲۴. [Online]. Available: <https://proceedings.neurips.cc/paper/۷۱۸۱-attention-is-all>

[۳۲] 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 ۵۸th Annual Meeting of the Association for Computational Linguistics*, ۲۰۲۰, pp. ۵۸۹۸–۵۹۰۵. Accessed: Oct. ۲۶, ۲۰۲۳. [Online]. Available: <https://aclanthology.org/۲۰۲۰.acl-main.۵۲۳/>

[۳۳] M. Daniluk, T. Rocktäschel, J. Welbl, and S. Riedel, "Frustratingly Short Attention Spans in Neural Language Modeling," arXiv, Feb. ۱۵, ۲۰۱۷. Accessed: Jun. ۰۴, ۲۰۲۴. [Online]. Available: <http://arxiv.org/abs/۱۷۰۲.۰۴۵۲۱>. <https://doi.org/۱۰.۴۸۵۵۰/arXiv.۱۷۰۲.۰۴۵۲۱>.

[۳۳] P. H. Martins, Z. Marinho, and A. F. T. Martins, "Joint Learning of Named Entity Recognition and Entity Linking," arXiv, Jul. ۱۸, ۲۰۱۹. Accessed: Oct. ۲۶, ۲۰۲۳. [Online]. Available: <http://arxiv.org/abs/۱۹۰۷.۰۸۲۴۳>

[۳۴] T. Satyapanich, F. Ferraro, and T. Finin, "Casie: Extracting cybersecurity event information from text," in *Proceedings of the AAAI conference on artificial intelligence*, ۲۰۲۰, pp. ۸۷۴۹–۸۷۵۷. Accessed: May ۲۲,

[۳۴] Y. Lu et al., "Unified Structure Generation for Universal Information Extraction." arXiv, Mar. ۲۳, ۲۰۲۲.

scheme," in Proceedings of the ۶th conference on Message understanding, Association for Computational Linguistics, ۱۹۹۵, pp. ۴۵-۵۲.

[۵۴] 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, ۱۹۹۸, pp. ۵۶۳-۵۶۶.

[۵۵] 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, ۲۰۰۵, pp. ۲۵-۳۲.

[۵۶] S. Pradhan, A. Moschitti, N. Xue, O. Uryupina, and Y. Zhang, "CoNLL-۲۰۱۲ shared task: Modeling multilingual unrestricted coreference in OntoNotes," in Joint Conference on EMNLP and CoNLL-Shared Task, Association for Computational Linguistics, ۲۰۱۲, pp. ۱-۴۰.

[۵۷] B. Kantor and A. Globerson, "Coreference Resolution with Entity Equalization," in Proceedings of the ۵۷th Conference of the Association for Computational Linguistics, ۲۰۱۹, pp. ۶۷۳-۶۷۷.

[۵۸] 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. ۱۱(۴), pp. ۱۷-۳۴, ۲۰۲۴. DOR: <https://dor.isc.ac/dor/۲۰.۱۰۰۱.۱.۲۳۲۲۴۳۴۷.۱۴۰۲.۱۱.۴.۲.۹>.

[۵۹] 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 ۱, no. ۱ pp. ۰۹-۱۴, ۲۰۲۱. doi:۱۰.۴۶۶۳۲/daai/۱/۱/۲.

[۶۰] Yazdani, Morteza, and Felipe R. Graeml. "VIKOR and its applications: A state-of-the-art survey." Int. J. Strateg. Decis. Sci., no. ۲, pp. ۵۶-۸۳, ۲۰۱۴. DOI: ۱۰.۴۰۱۸/ijds.۲۰۱۴.۰۴.۰۱۰۵.

۲۰۲۴. [Online]. Available: <https://ojs.aaai.org/index.php/AAAI/article/view/۶۴۰۱>

[۴۵] V. Behzadan, C. Aguirre, A. Bose, and W. Hsu, "Corpus and deep learning classifier for collection of cyber threat indicators in twitter stream," in ۲۰۱۸ IEEE International Conference on Big Data (Big Data), IEEE, ۲۰۱۸, pp. ۵۰۰۲-۵۰۰۷. Accessed: May ۲۲, ۲۰۲۴. [Online]. Available: <https://ieeexplore.ieee.org/abstract/document/۸۶۲۲۵۰۶>

[۴۶] S. K. Lim, A. O. Muis, W. Lu, and C. H. Ong, "Malwaretextdb: A database for annotated malware articles," in Proceedings of the ۵۵th Annual Meeting of the Association for Computational Linguistics (Volume ۱: Long Papers), ۲۰۱۷, pp. ۱۵۵۷-۱۵۶۷. Accessed: May ۲۲, ۲۰۲۴. [Online]. Available: <https://aclanthology.org/P۱۷-۱۱۴۳/>

[۴۷] A. Roy, Y. Park, and S. Pan, "Predicting malware attributes from cybersecurity texts," in Proceedings of the ۲۰۱۹ Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume ۱ (Long and Short Papers), ۲۰۱۹, pp. ۲۸۵۷-۲۸۶۱. Accessed: May ۲۲, ۲۰۲۴. [Online]. Available: <https://aclanthology.org/N۱۹-۱۲۹۳/>

[۴۸] 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 Demonstrations at the ۱۳th Conference of the European Chapter of the Association for Computational Linguistics, ۲۰۱۲, pp. ۱۰۲-۱۰۷. Accessed: May ۲۲, ۲۰۲۴. [Online]. Available: <https://aclanthology.org/E۱۲-۲۰۲۱.pdf>

[۴۹] "CVEProject/cvelist." CVE Program, May ۲۲, ۲۰۲۴. Accessed: May ۲۲, ۲۰۲۴. [Online]. Available: <https://github.com/CVEProject/cvelist>

[۵۰] 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. ۱۴, ۲۰۲۳. Accessed: May ۲۲, ۲۰۲۴. [Online]. Available: <http://arxiv.org/abs/۲۳۰۴.۰۷۴۱۱>

[۵۱] J. R. Hobbs, "Resolving pronoun references," Lingua, vol. ۴۴, no. ۴, pp. ۳۱۱-۳۳۸, ۱۹۷۸. [https://doi.org/۱۰.۱۰۱۶/۰۰۲۴-۳۸۴۱\(۷۸\)۹۰۰۰۶-۲](https://doi.org/۱۰.۱۰۱۶/۰۰۲۴-۳۸۴۱(۷۸)۹۰۰۰۶-۲).

[۵۲] "Annotated English Gigaword - Linguistic Data Consortium." Accessed: Apr. ۱۲, ۲۰۱۹. [Online]. Available: <https://catalog.ldc.upenn.edu/LDC۲۰۱۲T۱>

[۵۳] M. Vilain, J. Burger, J. Aberdeen, D. Connolly, and L. Hirschman, "A model-theoretic coreference scoring