PAPER SUMMARIZARION

* Title: A Pretrained Language Model for Cyber Threat Intelligence
* Authors Youngja Park IBM T. J. Watson Research Center Yorktown Heights, NY

Weiqiu You University of Pennsylvania Philadelphia, PA, USA

Abstract: Тус нийтлэлд кибер аюулгүй байдлын домэйны шинэ BERT загварыг танилцуулсан байна, CTI-BERT нь кибер аюулын тагнуулын (CTI) нарийвчлалыг сайжруулж, байгууллагуудад болзошгүй кибер аюулаас илүү сайн хамгаалах боломжийг олгодог ба домэйн цуглуулга үүсгэж, сургалтын арга зүй, кибер аюулгүй байдлын төрөл бүрийн NLP даалгаврын үр дүнтэй байдлын талаар дэлгэрэнгүй мэдээллийг өгдөг. Туршилтууд нь CTI-BERT нь эдгээр кибер аюулгүй байдлын хэрэглээний хэд хэдэн ерөнхий домэйн болон аюулгүй байдлын домэйн загваруудаас илт давуу байгааг харуулж байгаа нь сургалтын өгөгдөл, аргачлал нь загварын гүйцэтгэлд чухал нөлөө үзүүлж байгааг харуулсан байна.

Introduction: Цахим халдлага хурдацтай нэмэгдэж байгаатай холбогдуулан кибер аюулгүй байдлын мэргэжилтнүүд CTI-ийн олон тайланг нийтэлж, аюулгүй байдлын шинэ сул тал болон хортой програм хангамжийн талаар дэлгэрэнгүй бичдэг. Эдгээр тайлан нь аюулгүй байдлын шинжээчдэд кибер аюулыг илүү сайн ойлгоход тусалдаг ч бүх мэдээллийг цаг тухайд нь шингээхэд маш хэцүү байдаг. Тиймээс CTI-г текстээс автоматаар гаргаж авах нь кибер аюулгүй байдлын ханган ажилладаг мэргэжилтэн байгууллагуудын анхаарлыг ихэд татдаг. Гэсэн хэдий ч ерөнхий домайн хэлний загварууд (LMs) нь нэр томъёо, хэв маягийн ялгаатай байдлаас шалтгаалан кибер аюулгүй байдлын текстэд үр дүнтэй байдаггүй, Домэйн тусгай хэлний загварыг гаргахын тулд хоёр өөр аргыг ашигладаг. Тасралтгүй бэлтгэл, эхнээс нь бэлтгэх. Тасралтгүй бэлтгэлийн арга нь одоо байгаа ерөнхий домайн загварыг авч, тухайн домэйнийг ашиглан загварыг үргэлжлүүлэн сургадаг. Энэ арга нь ялангуяа домэйны тоо хэмжээ бага байх үед ашигтай боловч шинэ загварын үгсийн сан нь анхны загвартай ижил хэвээр байсаар байна. Тиймээс ихэнх домэйны тусгай нэр томьёо үгийн санд байдаггүй. scratch хандлага нь домэйны тусгай үгсийн санг бүтээх шинэ токенизаторыг сургаж, зөвхөн өөрийн корпусыг ашиглан хэлний загварыг сургадаг.

Main contribution

• Кибер аюулын тагнуулын шинжилгээнд тусгайлан зориулсан өндөр чанартай олон тооны мэдээллийн багцыг боловсруулсан.

• Кибер аюулгүй байдлын салбарт тохирох урьдчилан бэлтгэгдсэн BERT загварыг боловсруулсан.

• Бид аюулгүй байдлын домэйнд зориулсан өргөн хүрээний даалгавар, жишиг өгөгдлийн багц дээр өргөн хүрээтэй туршилт хийж, загварынхаа үр нөлөөг харуулсан.

Methodoly:

* Training dataset үүсгэсэн. Ихэнх өгөгдөл PDF. HTML форматтай байсан. Текстийг жижиг үсгээр бичсэний дараа WordPiece токенизаторыг сургаж, үгийн сан үүсгэсэн.
* Sentence Classification TasksӨгүүлбэр эсвэл баримт бичгийн түвшний ангилж урьдчилан бэлдсэн хэлний загварууд дээр далд давхарга, гаралтын проекц давхаргаар холбож ангиллын толгойг нэмж оруулдаг бөгөөд энэ нь өгүүлбэр дэх бүх токенуудын сүүлийн далд төлөвийн дундажийг оролт болгон авдаг. . Бид урьдчилан бэлтгэсэн загваруудыг санамсаргүй байдлаар эхлүүлсэн ангиллын давхаргын хамт 1000 алхмуудыг ашиглан нарийн сургасан.
* Masked Word Prediction

Загварууд домэйныг хэр сайн ойлгож байгааг хэмжихийн тулд масктай жетоныг урьдчилан таамаглах ажлыг гүйцэтгүүлсэн.

* Malware Sentence Detection

Хортой програмын өгүүлбэр илрүүлэх болон хортой програмын шинж чанарын ангилал хийсэн.

* Token Classification TasksХоёр хамгаалалтын домайн NER даалгавар болон token төрлийг илрүүлэх даалгаврыг ашиглан token түвшний ангиллын загваруудын үр нөлөөг харьцуулж үздэг.

Conclusion

Бид кибер аюулгүй байдлын домэйнд зориулан бэлтгэсэн шинэ BERT загварыг танилцуулсан. Тодруулбал, аюулгүй байдлын нэгж (IoCs) олборлолт, халдлагын техник (TTP) зэрэг кибер аюулын тагнуулын олборлолт, ойлголтын нарийвчлалыг сайжруулах зорилгоор тус загварыг боловсруулсан. Туршилтуудаас дүгнэхэд тус загвар нь одоо байгаа ерөнхий домэйн болон бусад ижил үндсэн архитектуртай кибер аюулгүй байдлын домэйны загваруудаас илүү сайн ажилласан үзүүлэлт гаргасан. Цаашид тусзагварыг сайжруулахын тулд илүү олон баримт бичиг цуглуулж, бусад хэлний загварыг сургах төлөвлөгөөтэй байгаа.

Гүйцэтгэсэн. Л.Санчирболд

23m1num0353