# Review 128: [Short] 18.08.23: CyBERT: Contextualized Embeddings for the Cybersecurity Domain (סקירה זו נכתבה על ידי עדן יבין) https://mdsoar.org/bitstream/handle/11603/25498/1117.pdf

**Paper: https://arxiv.org/abs/2210.08218v1**

מודלי שפה הראו את היכולת שלהם לעבוד בהמון תחומים בתוך עיבוד שפה טבעית. אחד התחומים שבהם ציפו להשפעה גדולה של מודלים אלו הוא תחום אבטחת המידע או בשמו היותר מוקר תחום הסייבר. במאמר ששמו הינו CyBERT הראו לראשונה את השילוב של מודלי שפה גדולים (מודל BERT) בתחום הסייבר.   
  
החוקרים ניסו להראות איך שיפור של BERT הקיים יכול להביא לשיפור ניכר של אותו מודל במשימות שונות כגון זיהוי אובייקטים הקשורים לתחום הסייבר או סיווג של מילה לאובייקט המתאים לה. למשל, ניתן לראות תוצאות של המשימה האחרונה המוזכרת בה המודל נאלץ לסווג מילה מתחום הסייבר לאובייקט המתאים לה. איך ביצעו זאת? על ידי הרחבת האימון של BERT עם מספר שלבים נוספים:

– אוספים מסמכים רבים מתחום הסייבר   
– מנקים את המסמכים והופכים אותם לרשימה של טוקנים  
– את הרשימה של טוקנים מוספים למילון של ה-Tokenizer של BERT. בנוסף, מוסיפים למטריצת ה-Embedding הרגילה של BERT את הטוקנים עם ערך רנדומלי. מבצעים אימון נוסף של Masked Language Modeling, בה המודל נדרש להשלים מילות חסרות במשפטים. ככה הערכים הרנדומלים מתעדכנים.  
 – מבצעים אימון נוסף בהתאם למשימה, למשל אימון מפוקח של זיהוי אובייקטים של סייבר במשפט. המאמר מעניין אותי במיוחד לא רק בשל היותו בתחום הסייבר אלא בשל התזכורת שהוא נותן כיצד ניתן לבצע אימון נוסף של מודלי שפה גדולים על תחומים חדשים.