אישה נעלה נעלה נעלה: מודלי עיבוד שפה טבעית בעברית

ענבל יהב אביחי שריקי

ד״ר ענבל יהב היא חברת סגל בכירה בפקולטה לניהול ע״ש קולר באוניברסיטת תל אביב. בעלת תואר ראשון במדעי המחשב ותואר שני במערכות מידע מהטכניון. קיבלה את הדוקטורט שלה באופטימיזציה וכריית נתונים מאוניברסיטת מרילנד בשנת 2010, והמשיכה לעבוד שם במשך שנתיים כחברת סגל אורחת. עיקר עבודתה מתמקדת בפיתוח והתאמה של מודלים סטטיסטיים לשימושם של חוקרים במערכות מידע. במחקריה היא משלבת אלגוריתמים לכריית נתונים, מודלי עיבוד שפה טבעית ומודלי אופטימיזציה כדי לייצר מודלים של ניתוח נתונים עבור יישומים שונים, ובהם יישומי בריאות הציבור וניתוח רשתות חברתיות.

אביחי שריקי הוא דוקטורנט בפקולטה לניהול ע״ש קולר באוניברסיטת תל אביב, תחת הנחייתה של ד״ר יהב. בעל תואר ראשון בכלכלה ולימודי מזרח תיכון מאוניברסיטת חיפה ותואר שני בניהול מערכות מידע מאוניברסיטת תל אביב. תחום המחקר שלו הוא זיהוי משמעויות בטקסט, במיקוד לשפה העברית. הוא פיתח את HeBERT, מודל שפה עברי המבוסס על ברט ומודל לזיהוי רגשות בעברית מטקסט.

תקציר

עברית שפה קשה. למחשב, כמו לאדם, וקצת יותר. בארבע השנים האחרונות מודלי עיבוד שפה טבעית נמצאים בשיא פריחתם עבור מגוון שפות ומגוון משימות מחשב, כגון תרגום, מענה על שאלות, ניתוח תחושות וכתיבת תקצירים. העברית, לעומת זאת, נותרה קצת מאחור. זה לא מאוד מפתיע מפני שקהל היעד של עברית קטן משמעותית מזה של שפות אחרות, ומבנה השפה מורכב בהרבה. למעשה העברית נחשבת ״שפה עשירה מורפולוגית״ – שפה שבה המידע המורפולוגי מקודד כחלק מהמילה, ולא מופרד ממנה כמו במרבית השפות הלטיניות.

ב-2021 פותח על ידי כותבי מאמר זה מודל שפה מבוסס *ברט* ראשון לשפה העברית, שהיווה יריית פתיחה למחקרים רבים בתחום. במאמר זה נציג את האתגרים בפיתוח מודל השפה העברית, נסקור את המודלים הקיימים והמאמצים המתמשכים לפיתוח כלים ומודלים חדשים, ולאן עוד אפשר וכדאי לשאוף. בנספח למאמר נציג הדרכה קצרה כיצד ניתן, ללא ידע מקדים עשיר, להשתמש במודל השפה בעברית לזיהוי תחושות מתוך שפה כתובה.

הקדמה: מה זה מודל שפה?

מודל שפה הוא בבסיסו מודל הסתברותי המחשב את ההסתברות של מונח (token) להיות חלק "תקני" מהשפה שהוא למד. המונח הנלמד יכול לכלול מילה שלמה, חלק ממילה (כגון תחילית, סופית), או אפילו אות בודדת. המחשב לומד להשלים מונחים ממשפטים רבים, וכך לומד את מבנה השפה, הטיותה ומשמעות המילים בתוכה, ויכול להוות בסיס למשימות שפה רבות.

ישנם סוגים שונים של מודלי השפה, והפשוט שבהם הוא מודל האוניגרמים (Unigrams) (Sebastiani, 2002). במודל זה, משפט מיוצג על ידי "שק מילים" (Bag of Words [BoW]), כלומר אוסף המונחים (או המילים, ביישום הפשוט של האלגוריתם) המרכיבים אותו. הסתברות משפט נאמדת כמכפלת הסתברות המונחים (כלומר שכיחותם בשפה) בשק. לדוגמה, המשפט "הילדה הלכה לבית הספר" ייוצג על ידי האוסף: {הילדה, הלכה, לבית, הספר}, והסתברות המשפט תחושב באופן הבא:

$$p(\text{"הילדה הלכה לבית הספר"}) = p(\text{"הילדה"}) \times p(\text{"הלכה"}) \times p(\text{"לבית"}) \times p(\text{"הספר"}) \quad (1)$$

את הייצוג הזה ניתן כמובן לשפר על ידי עיבוד מקדים של המשפט; מילות ואותיות קישור יוסרו מהמשפט (שכן אינן מוסיפות מידע על הסתברות המילים המרכיבות אותו ויופיעו באופן שכיח בכל משפט), המילים יוחלפו בצורות השורשיות שלהן, ואוניגרמים יוחלפו בצמדי מילים אם לאלו יש הסתברות גבוהה יותר. במקרה הזה הסתברות המשפט תוחלף במשוואה הבאה:

$$p(\text{"הילדה הלכה לבית הספר"}) = p(\text{"ילד"}) \times p(\text{"הלך"}) \times p(\text{"בית ספר"}) \quad (2)$$

היתרונות של מודל האוניגרמים הם הפשטות שבו והיכולת להבין את התוצר המתקבל מהאלגוריתם. החיסרון של המודל הוא החוסר הברור בסדר המונחים ובהקשר שלהם. בהינתן אי מגבלה על סדר המונחים והקשרם, המשפט "הילדה הלכה לבית הספר" סביר, בראיית המודל הפשוט, כמו המשפט "הספר הלך לבית הילדה". את המגבלות הללו פותרים מודלים מורכבים יותר, ובעיית סדר המונחים נפתרת על ידי הסתברות מותנה: הסתברות מונח בהינתן המונחים שקדמו לו (במודל חד־כיווני), או המונחים הקודמים למונח ואלו שעוקבים אחריו (במודל דו־כיווני). לגבי בעיית ההקשר, שהיא משמעותית סבוכה יותר, נדרשת הבנה מדויקת יותר של המונחים בשפה מעבר לשכיחותם. את הבסיס הראשוני לבעיית ההקשר הציעו תומאס מיקולוב וחבריו (Mikolov et al., 2013) מגוגל בשנת 2013, במודל הנקרא word2vec.

לפי מודל ה-word2vec של מיקולוב וחבריו, מונח בשפה מיוצג על ידי וקטור מספרי באורך קבוע, בתהליך שנקרא "שיכוני מילים" (או באנגלית, word embedding). לדוגמה, המילה "ילד" יכולה להיות מיוצגת במודל זה על ידי הווקטור [3.2, 1.2, 9.9, ..., 10.5]. ערכי הווקטור נלמדים בעזרת רשת נוירונים בעלת שתי שכבות, המאומנת על קורפוס גדול של טקסט, שמטרתה לשכן את המונחים בקורפוס על המרחב הרציף באופן שבו מונחים בעלי משמעות דומה ימוקמו קרוב זה לזה במרחב. לפי מודל זה, משמעות המונח נגזרת מהקשרו בשפה, כלומר מאוסף המונחים שקודמים לו במשפטים שונים ומאלו שבאים לאחריו. בהתאם, רשת הנוירונים מקבלת כקלט "הקשר": משפטים באורך שווה, שבהם מונח האמצע ממוסך (לדוגמה: "הילד הלך ____ הספר היסודי"), וחוזה כפלט "משמעות": מונח בעל ההסתברות הגבוהה ביותר להשלמת המשפט[1] (לדוגמה: "לבית").

אחרי הפיתוח של מיקולוב וחבריו, החלו לצוץ (ועדיין צצים) מודלי שפה שונים כפטריות אחרי הגשם. אחד הנפוצים שבהם הוא מודל ברט (Bidirectional Encoder Representations from Transformers [BERT]) של גוגל שפותח בשנת 2018 (Devlin et al., 2018). את המודל הזה, כפי שנפרט בהמשך המאמר, 'תרגמנו' לראשונה לעברית.

מודל ברט

מודל ברט הוא מודל שפה מבוסס טרנספורמר דו־כיווני – ארכיטקטורה ייחודית המאפשרת למידת־ייצוג של מונחים מתוך הקשר גלובלי (כגון מסמך) ולוקאלי (כגון משפט). מודל זה מאפשר ייצוג וקטורי שונה של מונחים דומים בעלי משמעות שונה, כפי שמשתמע מהקשרם במשפט, כגון המילה החוזרת "נעלה" במשפט "אישה נעלה נעלה נעלה".

1 ארכיטקטורת רשת נוירונים זו נקראת Continuous Bag of Words (CBOW). אלטרנטיבית, בארכיטקטורת ה-Skip-gram הקלט הוא המשמעות והפלט הוא ההקשר.

מודלי ברט נמצאים היום בשימוש נרחב ככלי להבנת שפות רבות, ובהן אנגלית, ערבית, רוסית ועוד שפות רבות אחרות. רוב המודלים מאומנים על קורפוסי ענק, ובפרט קורפוס ספרים דיגיטלי, ויקיפדיה ואוסקר (Open Super-large Crawled Aggregated coRpus [OSCAR]) (Suárez et al., 2020), שעותקים שלהם קיימים בשפות השונות. בדומה למודל word2vec, מודל ברט מקבל כקלט משפטים בשפה עם מונח ממוסך (הקשר), ומטרתו לחזות את המונח החסר (משמעות). בשונה מ-word2vec, מודל ברט כולל 12 שכבות של למידה עמוקה (הנקראות שכבות טרנספורמר), כאשר בכל שכבה יש למידה נוספת של הקשר המילים במשפט. מידע נוסף על ארכיטקטורת ברט ניתן למצוא במאמר המקורי של דבלין וחבריו (Devlin et al., 2018).

לאימון מודל ברט על שפה חדשה נדרשת החלטה אחת חשובה, והיא רמת הפירוט של המונח שעליו יאומן המודל (קלט המודל). נזכיר כי מונח הוא יחידת טקסט היכולה לכלול מילה שלמה, חלקי מילה, או אות בודדת. ניקח לדוגמה את המילה "הילדה" שהוצגה קודם. ברמת פירוט של אות בודדת, המונחים שייצגו מילה זו יהיו {"ה", "י", "ל", "ד", "ה"}. ברמת פירוט של מילה שלמה, המונח יהיה שקול למילה השלמה {"הילדה"}. חלוקה לחלקי מילים יכולה להתבצע בשתי דרכים: בעזרת ניתוח מורפולוגי {"ה", "ילד", "ה"}, או לפי חלקי מילה שכיחים סטטיסטית {"ה", "יל", "דה"}.

לכל אחד מרמות הפירוט יתרונות וחסרונות משלו. ככלל, ככל שרמת הפירוט גבוהה יותר (כגון רמת אות), כך המודל מסוגל בקלות יחסית ללמוד מילים חדשות העומדות בלוגיקה הכללית של השפה, אך מתקשה יותר ללמוד את ההקשר (אינטואיטיבית, המודל "מבזבז" שכבות של למידה כדי להרכיב מילים), ולהיפך (Jawahar et al., 2019) . בשפות שונות מצאו כי רמות פירוט שונות מיטיבות עם המודל: באנגלית, המודל הנפוץ הוא מודל המבוסס מילון כבן 30 אלף חלקי מילה (Devlin et al., 2018); בערבית, לעומת זאת (הקרובה יותר לשפה העברית), הייצוג הנפוץ ביותר הוא מילון מפורט בהרבה המכיל כ-60 אלף חלקי מילה (Antoun et al., 2020).

איכות מודל שפה

כדי ללמוד איכות של מודל שפה, נהוג להריץ את המודל על מספר משימות שפה נפוצות ולהשוות את ביצועיו למודלים קודמים. המשימות נחלקות לשניים: משימות לא מפוקחות, שמטרתן לראות כיצד המודל "מבין" את השפה, ומשימות סיווג שמטרתן לראות כיצד הבנת השפה תורמת לזיהוי אלמנטים שונים בשפה. להלן פירוט המשימות הנפוצות:

1. משימות לא מפוקחות

א. *משימת "מלא את החסר" (Fill-in-the-blank)*. משימה הבודקת את יכולת המודל להשלים מונחים חסרים בטקסט. ביצועי המשימה נבדקים על ידי מדד perplexity, המכמת את האיכות של חיזוי נכונות משפט בשפה באמצעות המודל. מתמטית, איכות החיזוי של רצף (W, דוגמת משפט) בין N מונחים (נניח, מילים), מוגדרת על ידי הממוצע המעריכי של הנראות המקסימלית המותנית של המונחים ברצף:

$$\mathrm{PP(W)} = \exp\left\{-\frac{1}{\mathrm{N}}\sum_{\mathrm{i}=1}^{\mathrm{N}}\log_{\mathrm{p_\theta}}(\mathrm{w_i}|\mathrm{w_{<i}})\right\} \quad (3)$$

ב. *הכללה למילים מחוץ למילון (out-of-vocabulary)*. משימה הבודקת את היכולת של המודל לזהות מילים שאינן מופיעות בקורפוס שעליו הוא אומן. מדד זה מכמת את אחוז המילים שהמודל לא הצליח לשכן (כלומר לייצר עבורם וקטור מספרי).

2. משימות מפוקחות

ג. זיהוי ישויות (Named-Entity Recognition [NER]). משימה הבודקת את היכולת של המודל לסווג ישויות עם שם בטקסט, כגון שמות, ארגונים ומיקומים של אנשים. איכות הסיווג נאמדת על ידי מדד F1:

$$2 \times \frac{precision \times recall}{precision + recall} \quad (4)$$

ד. זיהוי חלקי דיבור (Part-of-Speech [POS]). משימה הבודקת את היכולת של המודל לסווג את התפקיד הדקדוקי של מילה או ביטוי במשפט (לדוגמה: שם עצם, פועל, תואר השם). איכות הסיווג נאמדת אף היא על ידי מדד F1.

ה. ניתוח תחושות. משימה הבודקת את היכולת של המודל לזהות את התחושות המובעות בטקסט. משימה זו בדרך

כלל נבדקת על זיהוי קוטביות בתחושות (תחושה חיובית, ניטרלית, או שלילית), ונאמדת אף היא בעזרת מדד F1.

מודל שפה בעברית: HeBERT

על מנת לפתח מודל שפה נדרשות שלוש החלטות. הראשונה היא רמת המונח שהמודל לומד (מילה, חלקי מילה, או אות). השנייה היא ארכיטקטורת הלמידה העמוקה – אילו סוגי שכבות למידה לכלול, כמה שכבות למידה, מה הקשרים בינם, וכדומה. ההחלטה השלישית נוגעת למשימה שעליה מאומן המודל, כלומר איזו משימה המודל לומד לבצע כדי להבין את השפה.

בפיתוח מודל השפה בעברית, מודל HeBERT (Chriqui & Yahav, 2021), בחרנו להשתמש בארכיטקטורת הבסיס של מודל ברט, שהוכחה כאלטרנטיבה מובילה לבעיות שפה שונות (Radford et al., 2018). משימת האימון שנבחרה הייתה משימת "מלא את החסר", המוגדרת כמשימת הבסיס של מודל ברט. שתי החלטות אלו הן תלושות שפה, וניתן להחליפן בארכיטקטורות אחרות (דוגמת RoBERTa (Liu et al., 2019)) ומשימות שונות (דוגמת pointwise-mutual-information (PMI) masking (Levine et al., 2020)).

לבחירת רמת המונח המתאימה לשפה העברית, נדרשת הבנה של מאפייני השפה הייחודיים. עברית נחשבת "שפה עשירה מורפולוגית" (באנגלית: Morphological Rich Language, ובקיצור: MRL) – שפה שבה המידע המורפולוגי מקודד כחלק מהמילה ולא מופרד ממנה כמו במרבית השפות הלטיניות (Tsarfaty et al., 2010). מאפייני השפה כוללים את התכונות הבאות: (i)ריבוי נטיות לכל מילה, על ידי הוספות מוספיות למילת בסיס (לדוגמה: ילד, ילדם); (ii) נטייה לא רציפה, שבה בסיס המילה משתנה ולא רק המוספיות (לדוגמה: הלך, הולך); (iii) סדר המשפט לעיתים חסר משמעות ולעיתים רב־משמעות; (vi) למילים רבות יש משמעות כפולה המשתנה לפי ההקשר; (v) הניקוד בעברית, שאינו מופיע במרבית הטקסטים הכתובים, מוסיף משמעות למילים בשפה.

ההתלבטות בין רמות המונח השונות נעה על הציר שבין רמת מונח גבוהה (מילה), השומרת על מבנה המילה המורכבת בתוך המשפט שבו היא מוצגת ולכן מאפשרת למידה טובה יותר של הקשר, לבין רמת מונח נמוכה (אות), המחלקת מילה לגורמים המרכיבים אותה ולכן מוצלחת יותר בלמידת מבנה המילה (מורפולוגיה). על פני הציר יש גם רמות ביניים – חלקי מילה, וחלקי מילה עם משמעות מורפולוגית. האתגר בעברית הוא להבין הן את ההקשר והן את מבנה המילה.

כדי להבין את הפשרה בין החלופות, ערכנו ניסוי שבו אימנו מודל ברט עבור השפה העברית המאומן על משימת "מלא את החסר" עם רמות מונחים שונות (ספציפית, רמת מילה, גדלים שונים של מילון חלקי מילים, חלקי מילה עם משמעות מורפולוגית, ואות). בחנו את ביצועיו על שלוש המשימות המפוקחות שהוצגו בפרק הקודם (משימות לא מפוקחות אינן בנות השוואה במקרה זה בשל גודלם השונה של המילונים הנגזרים מרמת המונחים, לפרטים ראו Chriqui and Yahav, 2021)). בשל המשאבים הרבים הנדרשים לאימון מודל שפה, ביצענו את האימון על מסד נתונים קטן – מסד ויקיפדיה בעברית (מסד בגודל 650~ מגה). הממצאים מובאים בטבלה 1 לעיל. ניתן לראות בטבלה כי מודלי ה"אמצע" המאומנים על חלקי מילה טובים בהרבה ממודלי הקצוות. מבין חלופות האמצע, מודל המאומן על חלקי מילה עם משמעות מורפולוגית מראה ביצועים טובים יותר במשימות הדורשות הבנת מילים בשפה – זיהוי ישויות וזיהוי חלקי דיבור. לעומתו, מודל המאומן על חלקי מילה, ללא משמעות מורפולוגית, מצליח להבין את הרעיון המרכזי במשפט באופן טוב יותר, ובהתאם לחלץ את התחושה המתבטאת במשפט.

מודל השפה הסופי

כמודל שפה סופי, בחרנו לאמן מודל המבוסס על חלקי מילה. הסיבה לבחירה נובעת מהשימוש העיקרי במודלי שפה באקדמיה ובתעשייה כמודלי סיווג משפטים. כאמור, מודל זה הראה את הביצועים הטובים ביותר במשימת סיווג התחושות. את המודל הסופי אימנו על מסד ענק הכולל את מסד הוויקיפדיה בעברית (650~ מגה) ואת מסד האוסקר בעברית (בגודל 9.8~ ג'יגה). ביצועי המודל הסופי הראו יכולת זיהוי תחושות ברמת ביצועים מרשימה של F1 = 0.94, ורמת זיהוי חלקי משפט (NER, POS) של F1 = 0.96. ביצועים אלו טובים משמעותית ממודל העברית שקדם למודל זה ואינו מבוסס ברט (More et al., 2019). מודל חדש יותר, אלפברט (Seker at al., 2022), המבוסס על ארכיטקטורת הרשת האופטימלית שנמצאה במאמר זה, ואומן על מסד נתונים גדול יותר (17.9~ ג'יגה), הראה ביצועים דומים במשימות אלו.

טבלה 1: השווה בין ביצועי מודל ברט בעברית, תחת רמות מונחים שונות

רמת מונח	זיהוי ישויות[1]	זיהוי חלקי דיבור[2]	ניתוח תחושות[3]
אות	0.74	0.92	0.69
חלקי מילה עם משמעות מורפולוגית (More et al., 2019)	0.92	0.95	0.65
חלקי מילה	0.79	0.90	0.79
מילה	0.86	*(לא ניתן לחישוב בשל ביצועים נמוכים במשימת מלא־את־החסר)*	0.43

1 מתוך המסד שפורסם על ידי Mordecai and Elhadad (2005).
2 מתוך המסד שפורסם על ידי .Sima'an et al (2002).
3 מתוך המסד שפורסם על ידי .Amram et al (2018).

את מודל השפה המאומן HeBERT ניתן למצוא בגיט[2], באתר מודלי השפה huggingface[3], ובשירותי הענן של אמזון[4] (באדיבות איתן סלע, אדריכל פתרונות ב-Amazon Web Services).

התפתחות כלי עיבוד שפה טבעית נוספים בשפה העברית

במרוצת השנים האחרונות התפתח תחום עיבוד השפה הטבעית בעברית לאין שיעור. חלק לא מבוטל מהתפתחות זו מיוחס לתקציבי־על שניתנו לפרויקטים בתחום על ידי משרד החדשנות, המדע והטכנולוגיה, שמהם צמח גם האיגוד הישראלי לטכנולוגיות שפת אנוש[5], המאגד נציגי אקדמיה ותעשייה במטרה לפתח את תחום עיבוד השפה בעברית.

כדי להדגים את התפתחות השפה, נפרט במאמר זה אודות שני מאמצים מחקריים שביצעו כותבי מאמר זה: פיתוח מודל זיהוי רגשות בעברית המשתמש במודל השפה הבסיסי בעברית לצורך משימת סיווג, ופיתוח מודל שפה משפטי ה"מלמד" את מודל השפה "להבין" את המונחים המשפטיים בעברית.

2 https://github.com/avichaychriqui/HeBERT
3 https://huggingface.co/avichr/heBERT_sentiment_analysis
4 https://github.com/aws-samples/aws-lambda-docker-serverless-inference/tree/main/hebert-sentiment-analysis-inference-docker-lambda
5 https://www.iahlt.org

מודל זיהוי רגשות: HebEMO

משימת זיהוי רגשות היא אחת המשימות הנפוצות בעיבוד שפה טבעית. מטרת משימה זו היא לזהות קשת רחבה של רגשות כפי שהיא באה לידי ביטוי בתוכן כתוב, ובכללה אושר, כעס, פחד ועוד. זיהוי רגשות אלו יכול לשפוך אור על אמונות, התנהגויות צפויות, ומצבים נפשיים שבהם שרויים אנשים. בספרות הפסיכולוגית ישנן מספר תיאוריות המגדירות את קשת התחושות של אדם. הנפוצה שבהן היא זו שפורסמה על ידי פלוצ'יק (Plutchik, 1980) ומגדירה מעגל של ארבע תכונות מנוגדות: עצב–שמחה, כעס–פחד, אמון–גועל, והפתעה–ציפייה.

במחקר שלנו (Chriqui & Yahav, 2021), התבססנו על הגדרת התחושות לפי פלוצ'יק כדי לייצר מודל זיהוי תחושות אוטומטי. לשם כך אספנו מסד נרחב הכולל כ-4,000 הערות שנכתבו בתגובה לכתבות חדשותיות (אופי האיסוף מפורט במאמר). בחרנו לאסוף תגובות לכתבות שנכתבו בנושא הקורונה בשנת 2020, תקופה המוגדרת כתקופה עמוסה רגשית (Pedrosa et al., 2020), ופורסמו באחד משלושת אתרי החדשות הבאים: ynet, ישראל היום, ובחדרי חרדים. בחירת האתרים נבעה מהאופן שבו הם מייצגים את המגזרים השונים בארץ: שני הראשונים פונים בעיקר לקהל החילוני משני קצוות הקשת הפוליטית, ואילו השלישי פונה למגזר החרדי. את התגובות שלחנו לתיוג אנושי באתר Prolific[6], כך שכל תגובה תויגה על ידי כעשרה מתייגים דוברי השפה.

6 https://www.prolific.co/

לבסוף, ביצענו כיוונון (Fine-tuning) של מודל השפה HeBERT למשימת הסיווג של זיהוי התחושות (כלומר, עדכנו את משקולות ארבע השכבות האחרונות במודל לזיהוי אופטימלי של התחושות). איכות המודל לפי מדד F1 נעה בין 0.78 ל-0.97 לתחושות השונות, מלבד הרגש "הפתעה" שאותו המודל לא הצליח לזהות (F1 = 0.41), כמפורט בטבלה 2.

טבלה 2: ביצועי מודל ניתוח התחושות בעברית – HebEMO

תחושה	F1	Precision	Recall
עצב	0.84	0.83	0.84
שמחה	0.88	0.89	0.87
כעס	0.97	0.97	0.97
פחד	0.80	0.84	0.77
אמון	0.78	0.88	0.70
גועל	0.96	0.97	0.95
הפתעה	0.41	0.47	0.37
ציפייה	0.85	0.83	0.87

מודל שפה משפטית בעברית: Legal-HeBERT

הצורך בכלי עיבוד שפה טבעית למחקרים בתחומי המשפט והחקיקה, בעולם בכלל ובעברית בפרט, הלך והתעצם בשנים האחרונות (לדוגמה, Sarne et al., 2019; Katz & Nay 2021). לצד צורך זה, התפתחו מודלי שפה המותאמים לענה המשפטית בשפות רבות, ובראשן כצפוי בשפה האנגלית (Chalkidis et al., 2020).

התאמת מודל שפה לעולם תוכן חדש יכולה להתבצע בשתי דרכים. הראשונה היא אימון מודל שפה חדש (נניח ברט), שמקורות המידע שעליהם הוא מאומן הם מקורות משפטיים בלבד. הדרך השנייה, הנקראת "התאמת תחום" (domain adaptation, ראו Devlin et al. (2018)), נסמכת על מודל קיים והתאמתו לעולם תוכן חדש. במקרה הזה נקודת ההתחלה תהיה מודל שפה מאומן על קורפוס כללי (דוגמת HeBERT), שחלק מהשכבות שלו עוברות אימון מחודש בעזרת מקורות המידע המשפטיים. היתרונות והחסרונות של כל שיטה ברורים: בעוד אימון מחדש מייצר מודל שמבין בצורה טובה יותר את השפה המשפטית ואת הקשרי המילים בעולם תוכן זה (מבלי להיות "מבולבל" מהקשרים של אותן מילים בשפה שאינה משפטית), כמות המשאבים שהוא דורש היא גדולה – בהיבט מקורות המידע, בזמן העיבוד, ובמשאבי החומרה הנדרשים. לעומתו, מודל שעובר התאמה דורש פחות משאבים, אך עלול להיות פחות מדויק בהבנת הקשרים משפטיים פורמליים.

לצורך פיתוח מודל השפה העברית המשפטית ובחינת האופן המתאים לאימון המודל, במחקר אחרון (Chriqui et al., 2022) אספנו מקורות מידע משפטיים רבים בגודל כולל של 3.7~ ג'יגה, הכוללים את ספר החוקים הישראלי, מאגרי פסקי דין, מאגרי החלטות בית דין ועוד. על מאגר זה אימנו שני מודלים – אחד המאומן מהתחלה על המסד החדש, ואחד המבוסס על מודל HeBERT ומתאים אותו לתחום המשפט. שני המודלים המאומנים הועלו לגיט לשימוש הציבור[7].

את ביצועי המודלים בחנו על שתי משימות סיווג. למשימת הסיווג הראשונה בחרנו את מסד חוק ההסדרים (Kosti, 2021). מטרת המשימה, כפי שהוגדרה במאמר המקורי, הייתה לסווג משפטים ככאלו המכילים האצלת סמכויות לרשויות ולאלו שלא. יכולת הזיהוי האוטומטי במסד זה נאמדה על 0.87, כאשר המודל שאומן מהתחלה הראה את הביצועים הטובים ביותר. במשימת הסיווג השנייה בדקנו את יכולת המודלים לזהות התייחסות לזכויות אדם בדיוני חקיקה בכנסת ישראל[8]. במקרה הזה המודל הטוב ביותר היה דווקא המודל המותאם, וביצועיו נאמדו על F1=0.74. להערכתנו, הסיבה לביצועים הטובים יותר של המודל השני במקרה זה טמונה באופי השפה בדיוני חקיקה – שפה המשלבת שפה טבעית ושפה משפטית – ויישורה עם אופי מודל HeBERT המותאם לשפה המשפטית.

בימים אלו אנו עמלים על פיתוח נרחב של השפה העברית המשפטית, הכולל איסוף והנגשת מקורות מידע, תיוג ישויות משפטיות, ובניית מודי שפה (ראו מימון מחקר ושותפים בפרק התודות).

7 https://github.com/avichaychriqui/Legal-HeBERT?fbclid=IwAR3sFizNJEfPlXm0Agg5HpELUm49v11kfksjes72-Q-9CxMwv8hdR8I5ahg

8 במסד שנאסף ותויג על ידי פרופ' איתי בר־סימן־טוב מאוניברסיטת בר אילן וטרם פורסם.

מה הלאה?

מעבדות המחקר באוניברסיטאות השונות ממשיכות וימשיכו לפתח כלים עבור השפה העברית, ובכללם מודלי שפה לתחומים שונים, כגון מודל השפה המשפטית (Chriqui et al., 2022) ומודל לשפה רבנית (Shmidman et al., 2022), איסוף מסדים ייעודיים לאימון, דוגמת מסד השאלות ותשובות ParaShoot (Keren and Levy, 2021), והמשך פיתוח פרויקט תיוג הישויות ותיוג מורפולוגי שמוביל האיגוד הישראלי לטכנולוגיות שפת אנוש.

במקביל, עוד ועוד תחומי דעת עושים ויעשו שימוש במודלי שפה בעברית כחלק אינטגרלי מהמחקר שהם מבצעים. דוגמאות ראשונות לכך ניתן לראות בפסיכולוגיה (Hachohen-Kerner et al., 2022), בפוליטיקה (Litvak et al., 2022), ובמדעי החברה (Bialer et al., 2022).

אצלנו, במעבדה לבינה מלאכותית וניתוח עסקי בפקולטה לניהול ע״ש קולר באוניברסיטת תל אביב, אנו עתידים להוביל בשנים הקרובות שני מחקרים מרכזיים בתחום עיבוד השפה הטבעית בעברית. האחד, המשך פיתוח השפה המשפטית. פרויקט זה נמצא כעת בשלביו הראשונים, ויכלול איסוף מסדי עתק, תיוג אלפי ישויות משפטיות במסדים אלו, ופיתוח מספר מודלי שפה המותאמים ליישומים שונים בתחום המשפט. הפרויקט השני יתמקד בפיתוח מודלים בעברית פיננסית לצורך ניתוח אוטומטי של דיווחי תאגידים ציבוריים. פרויקט זה ייעשה בשיתוף פעולה עם הרשות לניירות ערך ועתיד להתחיל בשנה הקרובה.

תודות

מחקר זה התאפשר בזכות מספר מקורות מימון. הראשון הוא תקציב ממשרד החדשנות, המדע והטכנולוגיה (גראנט מספר #3-17991) בהובלת ד״ר ענבל יהב ופרופ׳ איתי בר־סימן־טוב (אוניברסיטת בר אילן). השני, תקציב רשות החדשנות (גראנט מספר #01103654), בהובלת פרופ׳ לב מוצ׳ניק (האוניברסיטה העברית) וד״ר ענבל יהב.

בנוסף, החוקרים מבקשים להודות לתמיכה האדיבה של קרן ג׳רמי קולר ומכון הנרי קראון למחקר עסקי בישראל.

ד״ר ענבל יהב inbalyahav@tauex.tau.ac.il

רשימת מקורות

Amram, A., David, A. B., & Tsarfaty, R. (2018, August). Representations And Architectures in Neural Sentiment Analysis For Morphologically Rich Languages: A Case Study From Modern Hebrew. In *Proceedings of the 27th International Conference on Computational Linguistics* (pp. 2242-2252).

Antoun, W., Baly, F., & Hajj, H. (2020). AraBERT: Transformer-based Model for Arabic Language Understanding. *arXiv preprint arXiv:2003.00104*.

Bialer, A., Izmaylov, D., Segal, A., Tsur, O., Levi-Belz, Y., & Gal, K. (2022). Detecting Suicide Risk in Online Counseling Services: A Study in a Low-Resource Language. *arXiv preprint arXiv:2209.04830*.

Chalkidis, I., Fergadiotis, M., Malakasiotis, P., Aletras, N., & Androutsopoulos, I. (2020). LEGAL-BERT: The Muppets Straight Out of Law School. *arXiv preprint arXiv:2010.02559*.

Chriqui, A., & Yahav, I. (2021). HeBERT & HebEmo: A Hebrew BERT Model and A Tool For Polarity Analysis and Emotion Recognition. *INFORMS Journal on Data Science* 1(1):81-95.

Chriqui, A., Yahav, I., & Bar-Siman-Tov, I. (2022). Legal HeBERT: A BERT-based NLP Model for Hebrew Legal, Judicial and Legislative Texts. *Judicial and Legislative Texts (June 27, 2022)*.

Devlin J, Chang MW, Lee K, Toutanova K (2018) Bert: Pre-Training of Deep Bidirectional Transformers For Language Understanding. *arXiv preprint arXiv:1810.04805.*

Hacohen-Kerner, Y., Manor, N., Goldmeier, M., & Bachar, E. (2022). Detection of Anorexic Girls-In Blog Posts Written in Hebrew Using a Combined Heuristic AI and NLP Method. *IEEE Access*, *10*, 34800-34814.

Jawahar, G., Sagot, B., & Seddah, D. (2019, July). What does BERT learn about the structure of language?. In *ACL 2019-57th Annual Meeting of the Association for Computational Linguistics*.

Katz, D. M., & Nay, J. J. (2021). Machine Learning and Law. *Legal Informatics*.

Keren, O., & Levy, O. (2021). ParaShoot: A Hebrew Question Answering Dataset. *arXiv preprint arXiv:2109.11314*.

Kosti, N. (2021). Centralization Via Delegation: The Long-Term Implications of The Israeli Arrangements Laws. In *Comparative Multidisciplinary Perspectives on Omnibus Legislation* (pp. 73-94). Springer, Cham.

Levine, Y., Lenz, B., Lieber, O., Abend, O., Leyton-Brown, K., Tennenholtz, M., & Shoham, Y. (2020). PMI-Masking: Principled Masking of Correlated spans. *arXiv preprint arXiv:2010.01825*.

Litvak, M., Vanetik, N., Talker, S., & Machlouf, O. (2022, October). Detection of Negative Campaign in Israeli Municipal Elections. In *Proceedings of the Third Workshop on Threat, Aggression and Cyberbullying (TRAC 2022)* (pp. 68-74).

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L. & Stoyanov, V. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv preprint arXiv:1907.11692*.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representations In Vector Space. *arXiv preprint arXiv:1301.3781*.

Mordecai, N. B., & Elhadad, M. (2005). Hebrew Named Entity Recognition. *MONEY*, *81*(83.93), 82-49.

More, A., Seker, A., Basmova, V., & Tsarfaty, R. (2019). Joint Transition-Based Models for Morpho-Syntactic Parsing: Parsing Strategies for MRLs And a Case Study From Modern Hebrew. *Transactions of the Association for Computational Linguistics*, *7*, 33-48.

Pedrosa, A. L., Bitencourt, L., Fróes, A. C. F., Cazumbá, M. L. B., Campos, R. G. B., de Brito, S. B. C. S., & Simões e Silva, A. C. (2020). Emotional, Behavioral, And Psychological Impact of The COVID-19 Pandemic. *Frontiers in psychology*, *11*, 566212.

Plutchik, R. (1980). A General Psychoevolutionary Theory of Emotion. In *Theories of emotion* (pp. 3-33). Academic press.

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving Language Understanding by Generative Pre-Training.

Sarne, D., Schler, J., Singer, A., Sela, A., & Bar Siman Tov, I. (2019, May). Unsupervised Topic Extraction from Privacy Policies. In *Companion Proceedings of The 2019 World Wide Web Conference* (pp. 563-568).

Sebastiani, F. (2002). Machine Learning in Automated Text Categorization. *ACM computing surveys (CSUR)*, 34(1), 1-47.

Seker, A., Bandel, E., Bareket, D., Brusilovsky, I., Greenfeld, R., & Tsarfaty, R. (2022, May). AlephBERT: Language Model Pre-training and Evaluation from Sub-Word to Sentence Level. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 46-56).

Shmidman, A., Guedalia, J., Shmidman, S., Shmidman, C. S., Handel, E., & Koppel, M. (2022). Introducing BEREL: BERT Embeddings for Rabbinic-Encoded Language. *arXiv preprint arXiv:2208.01875*.

Sima'an, K., Itai, A., Winter, Y., Altman, A., & Nativ, N. (2001). Building A Tree-Bank of Modern Hebrew Text. *Traitement Automatique des Langues*, *42*(2), 247-380.

Suárez, P. J. O., Romary, L., & Sagot, B. (2020). A Monolingual Approach to Contextualized Word Embeddings for Mid-Resource Languages. *arXiv preprint arXiv:2006.06202*.

Tsarfaty, R., Seddah, D., Goldberg, Y., Kübler, S., Versley, Y., Candito, M., Foster, J., Rehbein, I. & Tounsi, L. (2010, June). Statistical Parsing of Morphologically Rich Languages (SPMRL) What, How And Whither. In *Proceedings of the NAACL HLT 2010 First Workshop on Statistical Parsing of Morphologically-Rich Languages* (pp. 1-12).

נספח א: שימוש במודל שפה בעברית לזיהוי תחושות ורגשות בשפה כתובה

בחלק זה של המאמר נסביר כיצד ניתן לבצע ניתוח תחושות ורגשות על מסמך המכיל משפטים בעברית. בהדרכה זו ננית כי קובץ הקלט נקרא "data.csv" וכי העמודה שנרצה לנתח נקראת "text", כמופיע בדוגמה באיור 1. שימו לב שכדי לשמר את העברית בקובץ csv, יש לשמור את הקובץ בפורמט "CSV UTF-8". קוד הניתוח בפרק זה כתוב בשפת פייתון, וניתן להריצו בעזרת מחברת Jupyter[9] או בכל תוכנת פייתון אחרת. בהדרכה ננית כי קוד הניתוח וקובץ הקלט נמצאים באותה תיקייה.

איור 1: דוגמה לקלט לניתוח

	A	B	C	
1	rowNum	CommentName	text	
2	1	Gilad	"ננננות מפחדות להידבק בקורונה"	
3	2	Mika	"הסכנות במנת חיסון אחת"	
4	3	Sarit	"המצב הבריאותי השתפר בצורה יוצאת דופן?"	

על מנת לנתח את הקובץ באיור 1, יש לטעון תחילה את הספריות הנדרשות ולקרוא את הקובץ לזיכרון. בקטע קוד 1 נמצאות שורות הקוד הרלוונטיות.

קוד 1: טעינת ספריות וקובץ לזיכרון

```
## Required installations (run only once)
!pip install transformers

## Upload libraries
import pandas as pd
from transformers import pipeline
```

```
## Read file
input_path='data.csv'
df=pd.read_csv(input_path)
```

בשלב הבא נוכל להריץ את ניתוח התחושות (קוד 2) וניתוח הרגשות (קוד 3). שורות הקוד של הניתוח כוללות שלושה שלבים: (1) טעינת המודל לזיכרון (תחושות או רגשות), (2) חישוב התחושה או הרגש לכל שורה במסד, ו-(3) שמירת התוצר לקובץ פלט. קובצי הפלט מופיעים באיורים 2 ו-3 בהתאמה, ומכילים מידע לגבי התחושה (חיובי, שלילי, ניטרלי) ורמת הביטחון בתחושה זו, וכן לגבי הרגש (קיים [LABEL_1] או לא קיים [LABEL_2]), ורמת הביטחון ברגש זה. לצורך ההדגמה, קובץ התחושות מכיל מידע על תחושת שמחה וכעס בלבד. ניתן לראות שהמשפט הראשון מכיל אלמנט של כעס, ואילו האחרון מכיל אלמנט של שמחה.

9 https://jupyter.org/

קוד 2: ניתוח תחושות

```
# Load sentiment pipeline
sentiment_cls=pipeline(
    'sentiment-analysis',
    model='avichr/heBERT_sentiment_analysis',
    tokenizer='avichr/heBERT_sentiment_analysis',
    device=0) #run on GPU (if there is no GPU
              #installed on your machine, change to 'device = -1')

# Add sentiment score
df = df.join(
    pd.DataFrame([sentiment_cls(df['text'][i])[0] for i in range(len(df))]).
rename(columns={'label':'label','score':'confidence'}))

df.to_csv('data_polarity_analysis.csv', encoding='utf-8-sig')
```

איור 2: פלט ניתוח תחושות

	A	B	C	D	E	
1	rowNum	CommentName	text	label	confidence	
2	1	Gilad	"נננות מפחדות להידבק בקורונה"	negative	0.9999	
3	2	Mika	"הסכנות במנת חיסון אחת"	negative	0.9373	
4	3	Sarit	"המצב הבריאותי השתפר בצורה יוצאת דופן?"	positive	0.9996	

קוד 3: ניתוח רגשות

```
# Compute each emotion separately
emotions = ['anticipation', 'joy', 'trust', 'fear',
            'surprise', 'anger', 'sadness', 'disgust']

for emotion in emotions:
    # Load emotions pipeline
    sentiment_cls=pipeline(
        'sentiment-analysis',
        model='avichr/hebEMO_' + emotion,
        tokenizer='avichr/heBERT',
        device=0) #run on GPU (if there is no GPU
                  #installed on your machine, change to 'device = -1')

    # Add emotions score
    df = df.join(
    pd.DataFrame([sentiment_cls(df['text'][i])[0] for i in range(len(df))]).
rename(columns = {'label':emotion, 'score':emotion+'_confidence'}))

# Save to file
df.to_csv('data_emotions_analysis.csv', encoding='utf-8-sig')
```

איור 3: פלט חלקי של ניתוח הרגשות

	A	B	C	D	E	F	G	
1	rowNum	CommentName	text	joy	joy-confidence	anger	anger_confidence	
2	1	Gilad	"נננות מפחדות להידבק בקורונה"	LABEL_0	0.9999	LABEL_1	0.9997	
3	2	Mika	"הסכנות במנת חיסון אחת"	LABEL_0	0.9999	LABEL_0	0.9991	
4	3	Sarit	"המצב הבריאותי השתפר בצורה יוצאת דופן?"	LABEL_1	0.9374	LABEL_0	0.9999	