**פרויקט למידה עמוקה/ רוחמה הודג'ס תשפ"ה**

**מבוא**

במהלך מצבי חירום לאומיים, ובפרט בתקופות של לחימה כמו מבצע "חרבות ברזל", התקשורת הישראלית מציפה את הציבור בעדכוני חדשות ומבזקים. לכל כותרת חדשותית יש פוטנציאל לעורר תגובה רגשית — בין אם של פחד, כאב או נחישות. מטרת פרויקט זה הייתה לחקור כיצד ניתן להשתמש בלמידה עמוקה על מנת לסווג את **התגובה הרגשית שמתעוררת בקרב הקורא** בעקבות כותרות חדשותיות בעברית.

לשם כך, נבנתה מערכת שמקבלת כותרת חדשותית קצרה כקלט, ומחזירה את הרגש הדומיננטי מתוך שלוש קטגוריות מוגדרות: **פחד (0)**, **כאב (1)** ו־**נחישות (2)**.

הדאטאסט כלל כ־900 כותרות שנאספו **ידנית** מאתרי חדשות שונים במהלך כל תקופת הלחימה, כולל מקורות כמו **רוטר**, **מעריב**, **מבזקים נט**, ואחרים. כל כותרת סווגה בהתאם לרגש המרכזי שהיא מעוררת:

* כותרות המסווגות ל־**פחד** כוללות בדרך כלל אזעקות, נפילות טילים, ירי על כוחות צה"ל וכדומה.
* כותרות המסווגות ל־**כאב** כוללות בעיקר פרסומים על חיילים ואזרחים שנהרגו או נפצעו — "הותר לפרסום".
* כותרות המסווגות ל־**נחישות** מתארות פעולות תקיפה של צה"ל בעזה, לבנון או איראן, וכן אמירות מחזקות מצד גורמי ממשל וצבא.

התפלגות הרגשות בדאטאסט הייתה מאוזנת יחסית:

* פחד (0): 289 דוגמאות
* כאב (1): 298 דוגמאות
* נחישות (2): 308 דוגמאות

מערכת הלמידה שנבנתה כוללת שלב קידוד טקסטים בעזרת מודלי embedding בשפה העברית (HeBERT ו־AlephBERT, ) ולאחריו שלב סיווג רגשי באמצעות רשתות נוירונים. הפרויקט כלל השוואה שיטתית בין **סוגי embeddings שונים** לבין **ארכיטקטורות למידה שונות:** MLP, RNN Transformer במטרה לבחון איזה שילוב מניב את הביצועים הטובים ביותר.

**תיאוריה**

בליבת הפרויקט עומדות שיטות שונות של **למידה עמוקה (Deep Learning)** לניתוח טקסטים בעברית. השיטות נחלקות לשני שלבים עיקריים:

1. **ייצוג טקסטים באמצעות embeddings**
2. **סיווג רגשות באמצעות מודלי עיבוד רצפים**

**ייצוג טקסט – Embedding**

בשלב הראשון, יש צורך להמיר טקסט (כותרת חדשותית) לייצוג מספרי שמכיל מידע סמנטי ומשמעות. בפרויקט נעשה שימוש בשני מודלים מבוססי Transformer אשר עברו אימון מוקדם על טקסטים בעברית:

* **HeBERT** – אומן על טקסטים חדשותיים ופורומים, ומתאים לניתוח שפה רשמית ומובנית.
* **AlephBERT** – אומן על קורפוס עברי רחב מהאינטרנט, כולל שפה יומיומית, סלנג ותחביר לא פורמלי.

שני המודלים מבוססים על ארכיטקטורת **BERT**  שמייצרת embedding לכל טוקן בטקסט, ומוסיפה טוקן [CLS] שמייצג את כלל המשפט. לעיתים נעשה שימוש ב־[CLS] כווקטור סיכום, ולעיתים בכל רצף ה־embeddings.

**סיווג רגשות – ארכיטקטורות למידה**

לשלב הסיווג יושמו שלוש ארכיטקטורות שונות של רשתות נוירונים:

* **MLP (Multilayer Perceptron)** רשת פשוטה שמקבלת את embedding של המשפט ומבצעת סיווג בעזרת שכבות Fully Connected .
* **RNN מסוג Bi-LSTM** **או GRU** רשת עם זיכרון לטווח ארוך (Long Short-Term Memory) שמעבדת את רצף המילים ויודעת לשמר הקשרים לאורך המשפט. השימוש ב־ Bidirectional LSTM מאפשר הסתכלות גם קדימה וגם אחורה.
* **Transformer עם Fine-Tuning חלקי ומלא** שימוש ישיר במודל BERT תוך פתיחה לאימון של חלק מהשכבות או כולן. הידע הלשוני שנרכש באימון המקדים משתלב עם התאמה למשימת הסיווג.

לכל שיטה יש יתרונות וחסרונות, שנבחנו במהלך הפרויקט בהשוואה שיטתית, תוך שמירה על אחידות בדאטא ובמדדים.

**שיטות הלמידה**

בשלב זה של הפרויקט נבנו והושוו מודלים שונים של למידה עמוקה לצורך סיווג רגשות מתוך כותרות חדשותיות בעברית. כל מודל הופעל **פעמיים**: פעם אחת על גבי embedding מסוג **HeBERT** ופעם נוספת על גבי **AlephBERT** על מנת לבחון כיצד סוג ה־embedding משפיע על ביצועי המודל.

**חלוקת הנתונים**

הדאטאסט חולק לסט אימון (75%) וסט בדיקה (25%) תוך שימוש ב:

* stratify=y -לשמירה על יחס מאוזן של רגשות בין קבוצות האימון והבדיקה
* random\_state=42 - לשחזור תוצאות קבועות

**המודלים שנבנו**

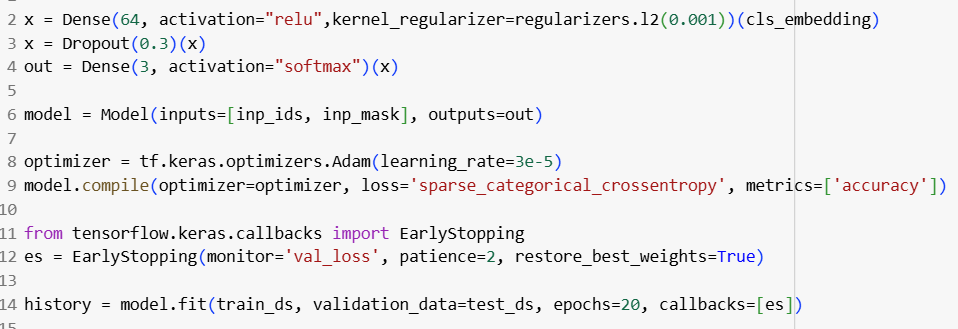
לכל מודל נבנו מספר **תתי־מודלים**, כדי לבדוק את השפעת הפרמטרים הבאים:

* מספר שכבות
* מספר נוירונים בשכבות הנסתרות
* ערכי Dropout שונים (0.3–0.6)
* עם ובלי רגולריזציה (L2)
* עם ובלי Batch Normalization או Layer Normalization

1. **MLP- תת המודל הטוב ביותר: 64 נוירונים, Dropout=0.5**
2. **RNN- תת המודל הטוב ביותר:**

****

1. **Fine- tuning חלקי- פתיחת 3 השכבות האחרונות לאימון. תת המודל הטוב ביותר:**
2. **Fine-tuning מלא- פתיחת כל השכבות לאימון על הדאטא הנוכחי. תת המודל הטוב ביותר(MLP):**

****

**אימון והערכה**

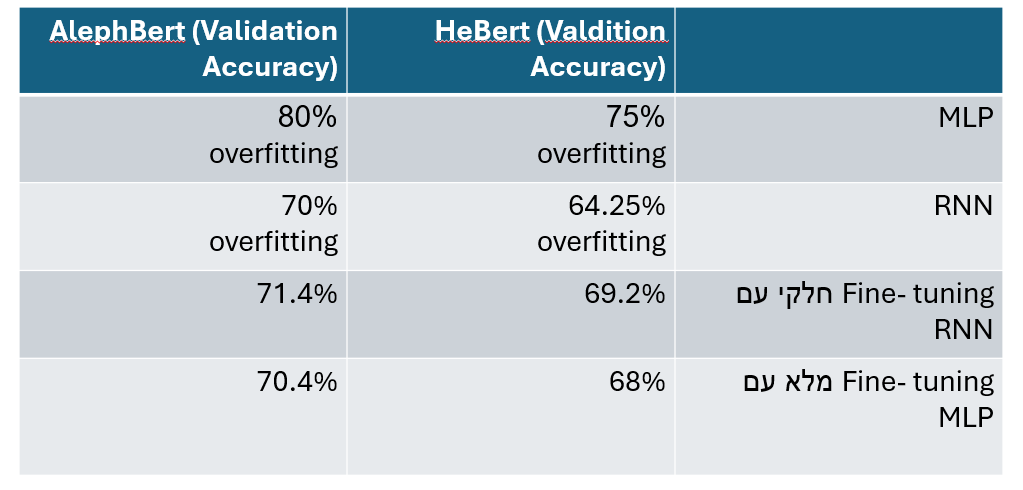
* אימון של עד 10 אפוקים עם EarlyStopping . במקרה שהיה נראה כיוון טוב במודל מסוים לאחר 10 אפוקים, המודל אומן שוב אך הפעם עם 20 אפוקים.
* סט הבדיקה נשמר קבוע .
* מדדי הערכה:
  + **Accuracy**
  + **LOSS**
  + **Confusion Matrix**

**תוצאות**

בשלב זה נבחנו תוצאות המודלים שנבנו, תוך השוואה בין סוגי האמבדינג(AlephBERT ו־HeBERT) ובין סוגי המודלים.

* Recall הגבוה ביותר התקבל עבור רגש **כאב** (0.83).
* Recall הנמוך ביותר היה בקטגוריית **פחד** (0.50).
* רגש **נחישות** דורש הבחנה עדינה יותר, אך הושג בו Recall של 0.67.
* מדדי Precision ו־F1 משקפים איזון סביר בין הרגשות, עם יתרון קל לקטגוריית כאב.

**השוואת ביצועים בין המודלים:**



* מודל MLP השיג את התוצאות הגבוהות ביותר, במיוחד עם embedding מסוג AlephBERT (80%) אך סבל מ־Overfitting.
* גם RNN ו־MLP על גבי HeBERT הציגו ביצועים טובים, אך נטו ל overfitting.
* השגת שיפור משמעותי עם Fine-tuning חלקי – ביצועים יציבים וטובים גם ללא Overfitting במיוחד עבור AlephBERT.
* Fine-tuning מלא שיפר את ביצועי MLP אך לא עלה על Fine-tuning חלקי.
* AlephBERT סיפק ביצועים גבוהים יותר כמעט בכל המודלים

**גורמים חשובים לשיפור הביצועים:**

* שימוש ב־Dropout בטווח 0.3–0.4 תרם רבות להפחתת Overfitting .
* רגולריזציית L2 גם היא השפיעה לחיוב על הדיוק והיציבות של המודלים.

**ניתוח**

לאחר בנייה והפעלה של מגוון מודלים על שני סוגי אמבדינג, ניתן להסיק מספר תובנות עיקריות:

* **השפעת סוג ה־embedding**:  
  AlephBERT סיפק תוצאות טובות יותר באופן עקבי מ־HeBERT בכל המודלים שנבחנו. ייתכן שזה נובע מהעובדה ש־AlephBERT הוכשר על מגוון רחב יותר של טקסטים או ממאפייני ארכיטקטורה שונים שתרמו להבנה טובה יותר של שפת החדשות.
* **המודל הפשוט MLP הציג ביצועים מפתיעים**:  
  למרות פשטותו, מודל ה־MLP הגיע לדיוק של 80% על AlephBERT – אך סבל מ־Overfitting. הסבר אפשרי לכך הוא שמכיוון שהכותרות החדשותיות הן לרוב קצרות מאוד, אין בהן רצף טקסטואלי שמצדיק שימוש במודלים שמבוססים על הקשרים בין מילים כמו EMBEDDING. RNN סטטי של הכותרת מספק את רוב האינפורמציה הדרושה לסיווג.
* **שימוש ב־Fine-tuning חלקי שיפר את הביצועים**:  
  פתיחה של 2–3 שכבות אחרונות ב־BERT לאימון מחדש תרמה לשיפור הדיוק, מבלי לגרום לאוברפיטינג. Fine tuning מלא, לעומת זאת, לא שיפר משמעותית את הביצועים ואף הגביר את הסיכון לאוברפיטינג ככל הנראה עקב מספר הדוגמאות הקטן יחסית.
* **תרומת טכניקות להפחתת Overfitting**   
  השימוש ב־Dropout (במיוחד בערכים 0.3–0.4) ו־L2 Regularization היה משמעותי לשיפור היציבות ולמניעת Overfitting.. המודלים בהם נעשה שימוש בטכניקות אלה השיגו תוצאות טובות ויציבות יותר.
* **הבדלים בין רגשות**:  
  הרגש שנחזה בצורה המדויקת ביותר היה **כאב** (Recall = 0.83), כנראה בזכות מאפיינים לשוניים ברורים בכותרות מסוג זה (למשל, “הותר לפרסום”, “נפצע/נהרג”). לעומת זאת, תחושת **פחד** הייתה הקשה ביותר לזיהוי (Recall = 0.50), כנראה בגלל ניסוחים מגוונים ומעורפלים יותר.

**ביקורת**

במהלך הפרויקט נדרשתי ליישם ידע תאורטי בתחום הלמידה העמוקה על בעיה מעשית של סיווג רגשות מתוך טקסט קצר (כותרות חדשותיות), ולהתמודד עם אתגרים של בניית דאטאסט, תכנון מודלים, הערכת ביצועים וניתוח תוצאות.

במהלך העבודה העמקתי את הבנתי במספר נושאים מרכזיים:

* **שיטות embedding לשפה העברית**:  
  למדתי על ההבדלים בין HeBERT ל־AlephBERT, וראיתי בפועל כיצד סוג ה־embedding משפיע על הביצועים. חוויתי את המשמעות של שימוש ב־CLS vector מול רצף embeddings של הטוקנים.
* **בנייה ואופטימיזציה של מודלים**:  
  למדתי כיצד תת-מודלים שונים (שינוי במספר שכבות, נוירונים, רגולריזציה וכו') משפיעים על למידת המודל, במיוחד בדאטאסט יחסית קטן. כמו כן, הבנתי את החשיבות של שימוש בטכניקות כמו Dropout ו־L2 כדי למנוע Overfitting .
* **Fine-tuning של מודלים טרנספורמטיביים**:  
  הכרתי לעומק את האפשרויות של אימון חלקי ומלא של מודלי BERT והבנתי את ההשלכות של פתיחת שכבות שונות לאימון.

**מגבלות הפרויקט:**

* הדאטאסט נבנה ידנית ומכיל פחות מ־1,000 דוגמאות, ולכן אינו מייצג את כלל השפה או כל מגוון הרגשות האפשריים.
* לא בוצעה בדיקה על דאטאסט חיצוני חדש, ולכן לא ניתן לדעת עד כמה המודלים הכלילו טוב לדאטא שהם לא "ראו".

**הצעות לשיפור והרחבה:**

* להרחיב את הדאטאסט באמצעות איסוף אוטומטי של כותרות חדשות וסיווג בשיטה חצי-מונחית (semi-supervised).
* לאמן מודלים נוספים, כולל Transformer קטן מותאם לבעיות טקסט קצר.
* להוסיף קטגוריות רגש נוספות כמו תקווה, כעס, תסכול.

**מקורות**

<https://huggingface.co/avichr/heBERT>

<https://huggingface.co/onlplab/alephbert-base>

אתרי חדשות: רוטר, מעריב, מבזקים נט

ChatGPT