**דו"ח תרגיל בית 5**

Transformer and LLMs

**שמות המגישים + ת"ז:**

עוביידה חטיב, 201278066

מאיה עטואן, 314813494

**חלק א'**

הפונקציה data\_load, שנקראת מה- main, אחראית על טעינת ה- subset מתוך המאגר. אם יש subset שנשמר מקודם, היא טוענת אותו מהדיסק, הדבר נעשה ע"י בדיקת אם התיקייה של הפרויקט מכילה מאגר נתונים בשם ' 'imdb\_subset. אם כן, היא טוענת אותו. אחרת, הפונקציה טוענת את מאגר הנתונים השלם של IMDB, בוחרת באופן אקראי 500 דגימות ממנו, ושומרת אותו לדיסק. לאחר שמירתו לדיסק, היא טוענת אותו מחדש.

**חלק ב'**

המודל bert-base-uncased נטען, והוגדר לביצוע סיווג לשתי קטגוריות (ע"י num\_labels=2) בהתאם לסיווג שעושה המשימה שלנו לביקורות לחיוביות ושליליות. המודל עבר טוקנזציה, ושם העמודה 'label' הוחלף ל- 'labels'. מאגר הנתונים שלנו חולק לשני סטים: אימון (80%) ובדיקה (20%). לאחר מכן, אובייקט Trainer אותחל, כאשר קיבל לצד המודל את הארגומנטים של האימון, שני הסטים של האימון והבדיקה, הטוקנייזר, והפונקציה compute\_metrics המחשבת את דיוק המודל. הוא עבר אימון ואיבלואציה, כאשר את תוצאות האיבולואציה מודפסות למסך. **התוצאות מצורפות להלן**:

**האוגרמטים שנבחרו עבור האימון**:

**evaluation\_strategy = "no"** , נבחר בשל כך שאנחנו לא נדרשים להדפסה במהלך האימון, אלא רק את התוצאות שהתקבלו בשלב האיבולאציה.

**learning\_rate= 2e-5**, זהו הערך הסטנדרטי והיה חשוב להשתמש בו על מנת לשמור על איזון בין יציבות המודל ומהירות הלמידה שלו.

**per\_device\_train\_batch\_size = 8, per\_device\_eval\_batch\_size = 8**, הערכים האלה התאימו למאפיינים של המחשב שלנו ולכן לא היה צורך בלשנות אותם.

**num\_train\_epochs = 3**, הערך נבחר על סמך תצפית שלנו לתוצאת שהראו שמעבר ל- epoch השלישי אחוז הדיוק וה- function loss לא משתנות בהרבה, ומזה הוסק ששימוש במספר גדול יותר עלול להוביל ל- overfitting.

**weight\_decay = 0.01**, גם כן נבחר על סמך תצפית שלנו, כך שהערך הראה איזון והתאמה למשתנה num\_train\_epochs.

הפונקציה compute\_metrics שהוזכרה קודם לכן מחשבת את הביצוע של המודל ע"י השוואת הפלט logits לתוויות האמיתיות labels. היא משתמשת בפונקציה np.argmax() על מנת להמיר את התוויות של הפלט לבינאריות (0,1) ומשווה אותם לתוויות האמיתיות ע"י הפונקציה accuracy\_score() מהספרייה sklearn.

**שאלות**

1. **האם התוצאות היו טובות? האם הן תאמו לציפיות שלכם? הסבירו.**

לדעתנו, התוצאות טובות, התוצאה של 91% מצביעה על דיוק גבוה. הם כן תאמו במידה רבה לציפיות שלנו, בשל כך שהמודל יותר דינמי בתפיסת המבנה התחבירי של מילה במשפט, ואת ההקשר הסמנטי בין המלים לעומת מודלים שהשתמשנו בהם בעבר כמו TF-IDF.

1. **אם נרצה לסווג ביקורות ספרים באמצעות אותו מודל שאימנו, האם לדעתכם התוצאות יהיו טובות? הסבירו.**

אנחנו מצפים שהמודל יצליח להתמודד עם סיווג ביקורות ספרים, אולם במידה פחות טובה. יש הרבה משותף בין שתי המשימות. דוגמה לכך היא השימוש בשם תואר על מנת לתאר את החוויה או היבטים בה. שם התואר תמיד יצביע על חוויה שהיא כפן שלו, כמו במקרה של המילה "מאכזב" שנוטה להופיע בחוות דעת שלילית יותר מאשר חיובית ללא קשר להקשר. מצד שני, יש אלמנטים שכן ייחודיים להקשר של סרטים, כמו האלמנטים הוויזואליים הקשורים לצילום והעריכה, או המשפט שהוזכר בהרצאה " the value I got from the two hours watching it was the sum total of the popcorn and the drink" שיש לה פחות משמעות כאשר היא מופיעה בביקורת של ספר.

1. **אם נרצה לסווג ביקורות סרטים ממאגר אחר, שאינו IMDB , באמצעות המודל שאימנו, האם לדעתכם התוצאות יהיו טובות? הסבירו.**

אנחנו מצפים שהתוצאות יהיו יותר טובות מאלה של הספרים, אולם פחות בהשוואה לביקורות של מאגר ה- IMDB, כאשר מידת הדיוק תהיה בהתאם לכמה המאגר הזה דומה למאגר של ה- IMDB. אלמנטים שיכולים להיות שונים בין שני המאגרים: סגנון השפה, למשל אם הביקורות לקוחות ממגזין שבו אנשים מקצועניים כותבים את הביקורות אנחנו נצפה לשפה יותר פורמלית ושימוש באוצר מלים שונה. ז'אנרי הסרטים שהוא מכיל והתפלגות הז'אנרים השונים מתוך כלל הסרטים, כי ז'אנרים שונים מושכים קהלים שונים בתחביב, בגיל וכתוצאה מכך המלים, האורך של הביקורת ועוד מאפיינים עשויים להיות מושפעים.

1. **עבור כל אחד מהמקרים מסעיפים 2,3 , תארו מה נוכל לעשות כדי לשפר (עוד יותר) את התוצאות?**

להן הצעות שאמורות לשפר את התוצאות של המודל. ההצעות תקפות לשתי המשימות מהסעיפים 2,3:

* להוסיף דאטה שהיא ביקורות ספרים לסט האימון (להוסיף דאטה ממאגר הסרטים האחר), ובכך ליצור מודל המסוגל לסווג את שני הסוגים של ביקורות.
* לעשות fine tuning למודל ולהתאים אותו מחדש למשימה של סיווג ספרים (להתאים אותו לסגנון ותוכן הביקורות של המאגר האחר).
* להגדיל את הרגולריזציה, ע"י לתת ערך גדול יותר לארגומנט weight\_decay אשר מונע למידת יתר של סגנון מסוים, דבר שאמור להשתקף ביכולת המודל להכליל טוב יותר למשימה של ביקורות הספרים (להכליל טוב יותר לתבניות אחרות של ביקורות סרטים).

**חלק ג'**