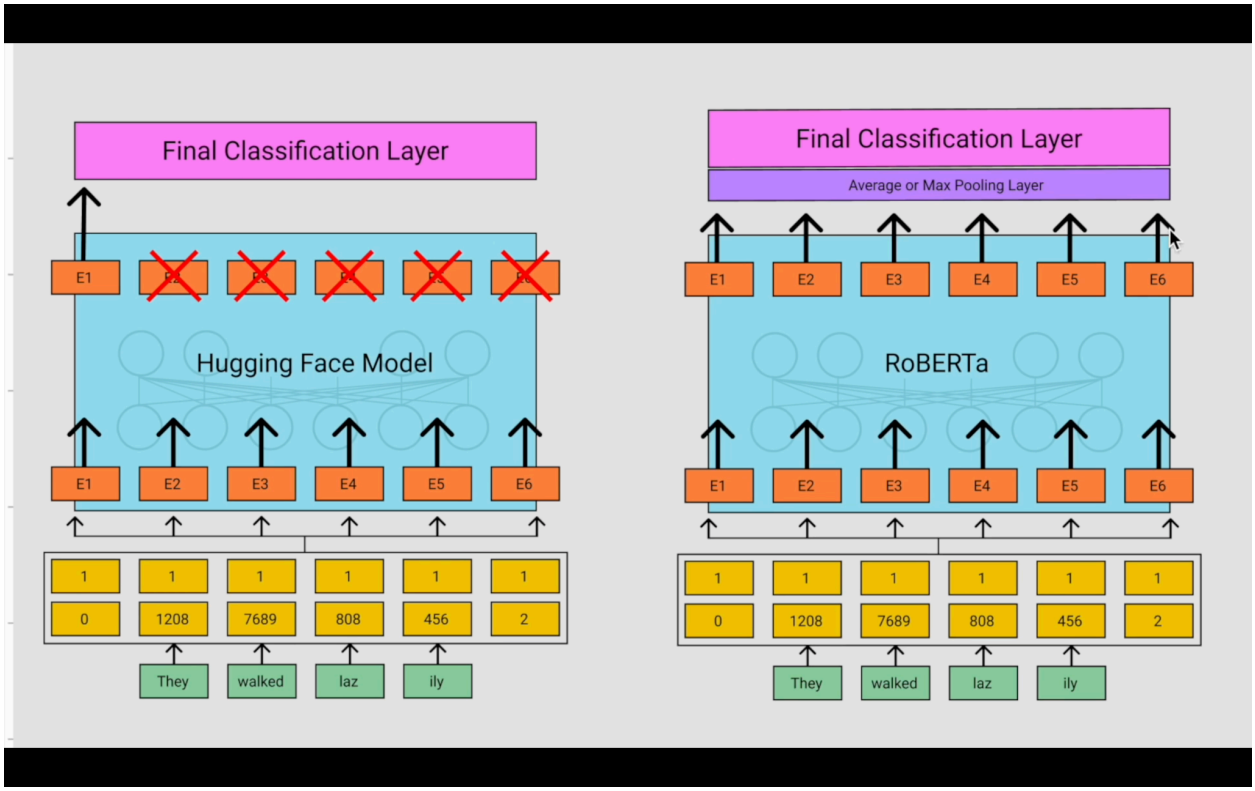
מודל RoBERTa:



<u>הטוקנים:</u> כל טקסט מחולק לטוקנים: מילים הנמצאים במילון עליו רוברטה אומן. מילה שלא בדיוק נמצאת במילון תפוצל לכמה טוקנים (לדוג׳: swimmingly יפורק לשני טוקנים: swimming ו - ly. המילה ly היא סיומת נפוצה ולכן נמצאת במילון)

טוקנים מיוחדים:
0 - מייצג תחילת טקסט
2 - מייצג סיום טקסט.
1 - מייצג טוקן שמיועד ל - padding. רוברטה יודעת להתעלם מטוקנים כאלה.

מסק (mask): נתעלם מטוקן שהmask שלו שווה ל0, אחרת הוא שווה ל1 ולא נתעלם

<u>הפלט של tokenizer:</u>
א׳ - input_tokens: מערך של הייצוג המספרי של כל טוקן. בדוג׳ לפי התרשים:
Input_tokens = [0, 1208, 7689, 808, 456, 2]

ב׳ - mask_attention: מערך בגודל זהה לinput_tokens של אפסים ואחדים. התא באינדקס i אומר שצריך להתעלם/ לא להתעלם מהטוקן במקום הו. בדוג׳ לפי התרשים:
mask_attention = [1, 1, 1, 1, 1, 1]
** בדיפולט: כל הטוקנים יקבלו ערך 1, וpadding tokens יקבלו ערך 0.

<u>הoutput של רוברטה:</u>

א׳ - **last_hidden_state** - טנזור ממימד (num_of_input_texts, ב , 768):

א.1 - num_of_input_texts זה מספר הטקסטים שהעברנו כאינפוט לtokenizer

א.2 - max_tokens_of_a_text - מספר הטוקנים המקסימלי שהתקבלו כתוצאה מהפעלת tokenizer על הטקסטים. לדוג׳: אם הטקסט הראשון התחלק ל9 טוקנים והטקסט השני התחלק ל 11 טוקנים אז max_tokens_of_a_text = 11.

** הערה על הדוג׳: הtokenizer יבצע padding לטקסט הראשון כדי שמספר הטוקנים בו יהיה 11.

א.3 - **768** הוא מספר הnodes שנמצאים בשכבת הפלט האחרונה של רוברטה (שכן מדובר ב-last_hidden_state).

המשמעות של הפלט היא שכל טוקן בטקסט כלשהו מיוצג כוקטור ממימד 768 כאשר כל כניסה (entry) בוקטור מייצג איזושהי משמעות סמנטית עבור אותו טוקן ובין הטוקנים השונים במטריצה המתאימה.

*** <u>הערה חשובה:</u> בפרויקט שלנו, הפיצ׳רים שלנו הם הסיכומים שענו התלמידים על שאלות ספציפיות. לכל פיצר כזה יש שני לייבלים: אחד לcontent והשני לwording. כל סיכום של תלמיד שניתן לroberta יפורק לקבוצה של טוקנים שיהפכו לוקטורים מספריים. <u>מכיוון שהסיכומים הם הפיצ׳רים שמעניינים אותנו ולא הטוקנים עצמם, אנחנו צריכים לייצג את קבוצת הטוקנים האלה כפיצ׳ר אחד!</u>

ב׳ - **pooler_output**: נקרא גם טוקן הCLS. הפלט הנל הוא הפלט שמתקבל מהטוקן הראשון לאחר שעבר עוד כמה מניפולציות ברשת נוירונים של רוברטה. לאחר המניפולציות האלה הטוקן אמור לייצג את כל הטקסט עצמו (המחשה בתרשים שנמצא בעמוד הראשון באיור השמאלי שבו מסתכלים רק על הטוקן הראשון).

*** הערה: לא מעניין אותנו הpooler_output כי אנחנו נרצה להסיק מידע על הטקסט על סמך כל הטוקנים כמו האיור הימני בתרשים מהעמוד הראשון.

<u>שימוש בהורדת מימדים (PCA):</u>

את המוטיבציה להורדת מימד מצאתי פה:

https://deep-ch.medium.com/dimension-reduction-by-whitening-bert-roberta-5e103093f782

בגדול כל טוקן בודד מיוצג כוקטור ממימד 768. זה מימד עצום למילה בודדת ואם יש אפשרות להוריד את המימד כך שנשמור על רוב המידע, נשפר את המהירות והיעילות של המודל שלנו במינימום פגיעה בביצועים.

** בקישור שהבאתי לכם הוא הוריד את המימד מ768 ל1. אישית זה מרגיש לי מוגזם (יש מצב אני טועה פה) אולי מספיק להוריד לאיזה 256 או 128 כי בלינק שצירפתי נראה שאפשר להוריד את רוברטה למימדים האלה ולשמור על כמעט כל המידע. אפשר לבחור כמובן גם לא להוריד מימד, תלוי בשיקולים שלנו.

שימוש בספרייה keras:

כאמור, הפלט שמתקבל מרוברטה הוא מטריצה של וקטורים ממימד 768. בין אם נבחר לבצע PCA או לא, הפלט של המודל שלנו (שהמודל יהיה רוברטה + שכבות נוספות שנוסיף לרשת) צריך להיות עם שני פלטים ולא 768: פלט עבור content ופלט עבור wording. ממה שאני מבין הרעיון כנראה יהיה כפי שמתואר בתרשים בעמוד הראשון באיור הימני: נצטרך (לאחר ביצוע PCA או ללא ביצוע PCA) לבנות שכבה שתיקח את הייצוג של הטוקנים כוקטורים ועליהם נבצע max_pooling / ממוצע של הטוקנים/ סתם לקחת מקסימום על כל הטוקנים כדי לקבל ייצוג יחיד של כל הטקסט שהבאנו לרוברטה כאינפוט (הייצוג שנרצה כנראה להגיע אליו זה וקטור ממימד 768 או ממימד שתלוי בביצוע PCA).
לאחר שנקבל את הייצוג של המידע כוקטור יחיד, נוסיף שכבה שתוציא עבורנו שני פלטים. ייתכן שבין לבין נצטרך שכבות נוספות אבל זה הרעיון הכללי. בגלל הוספת השכבות לרשת, נצטרך את keras.

הקישורים לסרטונים בהם קיבלתי את המידע (לא מצפה מכם להסתכל עליהם הם ארוכים):

https://chat.openai.com (כמובן)

https://www.youtube.com/watch?v=vNKIg8rXK6w

https://www.youtube.com/watch?v=DQc2Mi7BcuI

קישור לטוטוריאל של keras:
https://www.youtube.com/watch?v=5Ym-dOS9ssA&list=PLhhyoLH6IjfxVOdVC1P1L5z5azs0XjMsb

מטרות ללכת לפיהן באופן כללי:

1. אם בא לנו אפשר לנסות לעשות וויזואליזציה של הדאטה, לחפש קשרים מעניינים בדאטה בין תוצאות של סיכומים לבין שכבת גיל של הסטודנטים, כמה סטודנטים קיבלו ציון גבוה/ נמוך וכו׳.
2. לתקן שגיאות כתיב בכל סיכום.
3. להשתמש בtokenizer על כל הדאטה שלנו. אולי שווה לשלב כל סיכום עם השאלה המתאימה לו ואת זה לשלוח לtokenizer.
4. למצוא דרך לייצג את הפלט שחוזר מרוברטה כוקטור יחיד ככה שכל סיכום יקבל יצוג כוקטור יחיד ולא כקבוצת וקטורים (הקבוצה היא הטוקנים).
5. להחליט אם לבצע או לא לבצע PCA ובהתאם לתכנן את הארכיטקטורה של הרשת של המודל שלנו.
6. להתחיל לאמן את המודל.
7. להבין אם הדאטה שיש לנו הוא מספיק כדי לאמן את המודל ואם לא להגדיל אותו

(** על 4 ו 5 ו 7 אפשר להתייעץ עם תומר ועמית **)