פרויקט מסכם בקורס "נושאים מתקדמים בלמידה עמוקה"

מאמר אקדמי

קבוצה 2 : עמית גבע (ת"ז 316381805), עידן לוי (ת"ז 315873828)

תקציר

בעולם עיבוד השפה הטבעית (NLP), סיווג טקסט אוטומטי הוא משימה חיונית עם יישומים רבים בתעשייה. מאמר זה מציג ניתוח מקיף של מודלי למידה עמוקה לחיזוי הקטגוריה אליה משתייכת כותרת חדשותית טקסטואלית. בעבודתנו השתמשנו במערך נתונים גדול המורכב מ-1.61 מיליון כותרות חדשות שסופקו על ידי ה-Irish Times. התחלנו בעיבוד מוקדם של הנתונים כדי לייעל אותם למשימות NLP. תהליך זה כלל Tokenization, הסרת Stop-words, Lemmazation ואיזון מערך הנתונים על פני שש הקטגוריות הראשיות.

לאחר מכן, בחנו את מידת ההתאמה של שני מודלים שאומנו מראש (Pre-Trained), DistilBERT ו-RoBERTa, עבור משימת סיווג הכותרות. מצאנו ש-DistilBERT הביא לידי ביצועים טובים יותר מ-RoBERTa במדדי הערכה שונים, מה שגרם לנו לבחור בו כמודל הבסיס שלנו. לאחר שבוצע אימון של מודל זה על מערך הנתונים המלא, השתמשנו בשלוש טכניקות דחיסה שונות של מודל : Knowledge Distillation, Pruning ו- low-rank approximation .

הממצאים שלנו הראו שהמודל שכווץ באמצעות "Knowledge Distillation" הציע את האיזון המבטיח ביותר בין ביצועים ויעילות חישובית, מה שהופך אותו לבחירה אידיאלית עבור יישומים שבהם המהירות הינה מרכיב מרכזי. עם זאת, בהתאם לדרישות היישום הספציפיות, ניתן לשקול גם את הדגם אשר כווץ באמצעות "Pruning" והביא לביצועים טובים יותר במעט.

מאמר זה מספק תובנות חשובות לגבי היישום של למידה עמוקה וטכניקות דחיסה של מודלים למשימות סיווג טקסט תוך איזון מתמיד בין איכות הביצועים לבין העלות החישובית.

מתודולוגיה

ניתן לחלק את המתודולוגיה שנבחרה לפרויקט זה לארבעה שלבים : עיבוד מקדים של מערך הנתונים, בחירת המודל, אימון המודל הנבחר וכיווץ המודל המאומן.

1. <u>עיבוד מקדים של מערך הנתונים</u> - השלב הראשון של הפרויקט כלל בחינה מפורטת ועיבוד מקדים של מערך הנתונים. המטרה הייתה להבטיח שהנתונים יהיו בפורמט מתאים לאלגוריתמי למידת המכונה באמצעותם יבוצע תהליך האימון. תהליך זה כלל מספר שלבים, נציג כעת את העיקריים שבהם (מפאת אילוצי אורך המאמר, הסבר מפורט של תהליך העיבוד המקדים מתואר בנספח 1) :
   - הסרת ערכים כפולים עם טקסט כותרת וקטגוריות זהות.
   - צמצום מספר הקטגוריות הייחודיות מ-103 ל-6 קטגוריות ראשיות.
   - ביצוע תהליך Tokenizing על המילים ב-'headline_text'.
   - הסרת סימני פיסוק ו-Stop Words.
   - ביצוע פעולות Lemmazation כדי לצמצם מילים לצורתן הבסיסית.

התוצר של שלבים אלה הינו עמודת 'preprocessed_headline_text', המכילה את טקסט הכותרת המעובד, עמו נאמן את המודל.

2. <u>בחירת המודל</u> – לצורך בחירת המודל נבדקה התאמת מודלים Pre-trained שונים לצורך ביצוע המשימה. להלן תהליך בחירת המודל (הקוד מפורט במחברת בשם Training_The_Candidiate_Models.ipynb):

- סקירת מודלים קיימים לצורך איתור מודלים אפשריים לשימוש לצורך המשימה (פירוט בנספח מספר 2).
- לאחר בדיקת האופציות השונות הוחלט שעבור משימת הסיווג שלנו מודל ממשפחת מודלי BERT עשוי להוות התאמה טובה. על כן הוחלט לבצע השוואת ביצועים בין 2 הווריאנטים הבאים של המודל:
  - DistilBERT – גרסה מכווצת של מודל BERT הכווצה באמצעות Distillation. בהשוואה למודל הגדול, גרסה זו קטנה יותר, מהירה יותר, ויעילה יותר השומרת על ביצועים טובים.
  - RoBERTa – גרסת BERT משופרת המשיגה תוצאות טובות יותר במשימות NLP שונות. גרסה זו של המודל איננה מכווצת.

  נדגיש, עקב אילוצי משאבים חישוביים, נעדיף להשתמש במודל קטן יותר במידה וביצועיו לא פחותים משמעותית משל מקבילו הגדול.
- לצורך ההשוואה 2 המודלים אומנו במשך 5 epochs על 30% ממערך הנתונים, להלן התוצאות הטובות ביותר שהתקבלו עבור כל מודל:

	When?	Evaluation loss	Accuracy	F1 score	Precision	Recall
DistilBERT	~2.5 epochs	0.7536	0.7726	0.7723	0.7734	0.7726
RoBERTa	5 epochs	0.6539	0.7713	0.7681	0.7666	0.7713

לאחר סיום האימון המלא, מודל DistilBERT הראה סימנים מובהקים ל-Overfitting והציג ביצועים הלוקים משמעותית מאלו של RoBERTa. אך בעת ה-"Babysitting" על האימון נמצא שהמודל הציג למידה תקינה בערך עד אמצע האימון שם ביצועיו החלו להיפגע. על כן עבור 2 המודלים נבחנו ה-checkpoints בעלות הביצועים הטובים ביותר לצורך ההשוואה.
בתהליך בחירת המודל האופטימלי עבורנו נתבונן על כלל הקריטריונים שיש לקחת בחשבון ולא רק את תוצאות מדדי ביצוע. ההחלטה על המודל הנבחר צריכה לשקף היטב את האיזון בין עלויות חישוביות לבין איכות הביצועים:

- RoBERTa אכן עלה על ביצועיו של DistilBERT במונחים של evaluation loss מה שמרמז על פוטנציאל הכללה גבוה יותר, אך הפער בין ביצועיהם אינו משמעותי מספיק כדי להעדיף את RoBERTa באופן מוחלט.
- מצד שני, תוצאות שאר המטריקות שנבדקו (Accuracy, F1 score, Precision, Recall) במודל DistilBERT היו מעט טובים יותר מאשר במודל RoBERTa ועשויים להעיד במקצת על איכות ביצועים טובה למודל DistilBERT.
- עם זאת, ההחלטה אינה יכולה להישען על מדדי ביצוע בלבד. בהתחשב באילוצים של הפרויקט, במיוחד במשאבי החישוב המוגבלים הזמינים, היעילות החישובית של המודל הופכת מכרעת. DistilBERT, גרסה יעילה של BERT, תוכננה להציע איזון אופטימלי בין ביצועים ויעילות חישובית. הארכיטקטורה שלו כרוכה בפחות פרמטרים, מה שמוביל לדרישות זיכרון מופחתות וחישוב מהיר יותר במהלך תהליכי ההכשרה וההסקה.

- לכן, בהתחשב במגבלות המשאב החישובי ובמדדי הביצועים הדומים יחסית, DistilBERT נבחרה כמודל המתאים ביותר למשימת סיווג ספציפית זו. היעילות שלו באיזון בין איכות הביצועים לבין עלויות חישוביות התאימה בצורה הטובה ביותר לדרישות הפרויקט, מה שממחיש שבחירת המודל צריכה תמיד לשקף את האילוצים והיעדים הייחודיים של פרויקט נתון. בהקשר אחר או עם משימה אחרת, שבה משאבי חישוב אולי אינם גורם מגביל, הבחירה עשויה להעדיף את RobERTa או מודל אחר.

לכן, בהתחשב בכלל הקריטריונים שהוזכרו, הבחירה להמשיך עם DistilBERT לאימון על כלל סט הנתונים הייתה החלטה פרגמטית ומאוזנת.

3. <u>אימון המודל הנבחר</u> – את המודל הנבחר, DistilBERT, אימנו על כלל הנתונים באופן הבא המודל (הקוד מפורט במחברת בשם TrainingDistilbert.ipynb):

- <u>ניסיון 1</u>:
  - בוצעה הרצת ראשונה אימון ראשונה על המאפיינים הבאים:

גודל batch: 16	epochs 5	צעדי חימום: ~14,000
LR: 0.00004 (4e-5)	Weight Decay: 0.01	מטריקת הערכה: accuracy
ביצוע הערכת ביצועים כל 20,000 צעדים	80% מהנתונים בסט ה-Train, 20% הנותרים מהווים Test	תנאי עצירה מוקדמת אם לא חל שיפור במטריקת ההערכה של לפחות 0.01 במשך 4 הערכות רצופות ("סבלנות").

  - בנוסף נוספה למודל שכבת Dropout עם הסתברות של 0.5 על מנת למנוע over-fitting.
  - לאחר כ- 180,000 צעדים (Epochs 3.54) הופעל מנגנון ה-Early Stopping לאחר שהלמידה נעצרה (ראה נספח 3 לתוצאות המטריקות השונות לאורך האימון).
- <u>ניסיון 2</u>:
  - הוחלט להמשיך לאמן אך להגדיל רגולריזציה למניעת overfitting באופן הבא:
    - א. הגדלת weight_decay להיות 0.03.
    - ב. הגדלת ההסתברות ל-dropout להיות 0.6.
  - בנוסף הוחלט להגדיל את ה"סבלנות" של מנגנון העצירה המוקדמת ל-5, זאת על מנת לשלול כיוון של יציאה עצמונית של תהליך האימון מ-Overfitting.
  - לאחר שהמודל אומן סה"כ 5.79 epochs שוב הופעל מנגנון ה-earlystopping כאשר המודל הראה ביצועים הולכים ומתדרדרים בכל הערכה שבוצעה (ראה נספח 4 לתוצאות המטריקות השונות לאורך האימון).
  - מאחר וגם בתהליך האימון המקדמי מודל DistilBERT הראה נטייה ל-overfitting מוקדם הוחלט לעצור את האימון ולבחור את הנקודה בה ביצועי המודל היו האופטימליים.
- בחירת המודל מבין נקודות שונות לאורך תהליך האימון:
  - מבין 14 ה-checkpoints שנשמרו לאורך תהליך האימון הוחלט לבחור את ה-4 בעלות הביצועים הטובים ביותר ולבצע הערכה של ביצועיהן על <u>סט הנתונים המקורי</u> (לפני ביצוע oversampling/undersampling, אך כמובן לאחר כלל עיבוד הטקסט הרלוונטי).

- הנקודות שנבחרו לצורך התהליך הינן: 160,000 , 200,000 , 220,000, 280,000.
- עבור כל checkpoint נטען המודל שנשמר בה ובוצעה עליו ריצת evaluation על כלל סט הנתונים המקורי.
- להלן תוצאות ההרצות (הקוד המלא מפורט במחברת בשם Evaluating_optional_checkpoints.ipynb):

Checkpoint	eval_loss	eval_accuracy	eval_f1	eval_precision	eval_recall
checkpoint_160000	1.305	0.612	0.618	0.711	0.612
checkpoint_200000	1.51	0.622	0.633	0.706	0.622
checkpoint_220000	1.464	0.631	0.642	0.705	0.631
checkpoint_280000	1.801	0.625	0.634	0.7	0.625

- ניתן להבחין שעבור נקודה 220,000 מתקבל הדיוק הגבוה ביותר מבין 4 הנקודות המועמדות ובנוסף שאר המטריקות שנבדקו הינן האופטימליות או שאינן פחותות משמעותית משל שאר הנקודות.

על כן נקודה 220,000 מהווה עבורנו את הנקודה הטובה ביותר לפני התחלת overfitting ובחרנו במודל שנשמר בנקודה זו להיות המודל המלא עמו נתקדם.

4. <u>כיווץ המודל</u> - לצורך כיווץ המודל נבחנו 3 מתודות כיווץ שונות, נציג אותן ואת ביצועי המודל המכווץ באמצעות כל אחת מהן:

- <u>Knowledge Distillation (זיקוק)</u> – זהו תהליך שבו מודל קטן יותר (תלמיד) מאומן לחקות התנהגות של מודל גדול יותר (מורה). המטרה היא להשיג מודל קטן יותר שיכול לתת ביצועים קרובים למקבילו הגדולים יותר. מודל התלמיד מאומן לא רק על ה- hard labels אלא גם על ה-soft labels (ה-class probabilities) של המודל המורה, מה שמאפשר לו ללמוד מה"ניסיון" של המודל המורה. כפי שהשם של המודל הpretrained בו השתמשנו, DistilBERT, זהו מודל מכווץ של מודל BERT הגדול אשר כווץ באמצעות שיטה זו.
  - בכיווץ המודל בשיטה זו הגדרנו 2 מודלים: מודל תלמיד (המאותחל להיות מודל DistilBERT המקורי) ומודל מורה שהינו המודל אותו אימנו בשלב הקודם.
  - מאחר ובתהליך האימון של המודל התלמיד מהמודל המורה אנחנו לא מעוניינים ללמד את התלמיד רק על "תשובות נכונות", אלא גם כיצד להכליל דרך נתוני הקלט בהתבסס על ההתנהגות הנלמדת מהמודל המורה. כאמור, המודל המורה אומן על גבי סט נתונים רחב יותר אשר בא לידי ביטוי בהסתברות לפלט שלו לכל class (קטגוריה במקרה שלנו).
  - שיטת Kullback-Leibler (KL) divergence, הינה מדד לאופן שבו התפלגות הסתברות מסוימת שונה מהתפלגות הסתברות נוספת. על ידי שימוש בשיטה זו כפונקציית השגיאה בתהליך אימון התלמיד, אנו מאפשרים לתלמיד ללמוד לחקות את הסתברות הפלט שלו לכל class מזו של המודל המורה.
  - כחלק מהגדרת פונקציית השגיאה החדשה נלקח בחשבון פרמטר בשם טמפרטורה אשר נועד לסייע בשליטה באקראיות של הסתברויות הפלט מפונקציית ה-Softmax. במקרה שלנו בחרנו לאמן את המודל התלמיד עם טמפרטורה ששווה ל-1, בכך למעשה אנו מבצעים אימון סטנדרטי עם תוויות קשיחות, כלומר מודל התלמיד מנסה ישירות

לחקות את התחזיות של המורה. יציאות ה-softmax לא מותאמות במקרה זה, מה שאומר שההסתברויות המחושבות של המודל נלקחות כפי שהן.

- בחרנו בערך הטמפרטורה להיות 1 מהסיבות הבאות:

א. זו עשויה להיות נקודת התחלה טובה. כאשר הטמפרטורה היא 1, תהליך אימון התלמיד הופך למשימה פשוטה של שכפול הפלט של המורה. זהו מקום סביר להתחיל בו מכיוון שפלטי מודל המורה הינם בדרך כלל טובים, שכן הוא הוכשר בעבר במשימה.

ב. שימוש בטמפרטורה של 1 יכול לסייע במניעת overfitting של מודל התלמיד לרעש בפלטי המורה. אם היינו משתמשים בטמפרטורה גבוהה יותר, היינו שמים דגש רב יותר על שיעורי ההסתברות הנמוכים יותר בתפוקת המורה, מה שעשוי לשקף רעש או שגיאות בתחזיות שמבצע המורה.

לכן, במקרה זה, טמפרטורה השווה ל-1 נבחרה כנקודת מוצא פשוטה ועל מנת למנוע Overfitting פוטנציאלי מטעויות של המודל המורה.

- <u>Pruning (גיזום)</u> – זוהי טכניקה הכוללת הסרת פרמטרים לא חשובים (כמו משקולות) במודל. זה מקטין את גודל המודל ויכול להאיץ את החישוב. ישנם סוגים רבים של גיזום, כולל גיזום בגודל (הסרת פרמטרים עם ערכים מוחלטים קטנים), וגיזום מובנה (הסרת ערוצים שלמים, מסננים או נוירונים).

- בכיווץ המודל בשיטה זו השתמשנו בגיזום L1 לא מובנה בשכבות הליניאריות של המודל:

א. לכל שכבה ליניארית של המודל מזהים באמצעות נורמת L1 כדי לדרג ולזהות מי הן המשקולות הפחות רלוונטיות בשכבה זו.

ב. לאחר זיהוי המשקולות הנ"ל נגדיר ל-20% מהן משקל 0.

ג. לאחר איפוס המשקולות הנבחרות נסיר אותן מהמודל.

- <u>Low-Rank Approximation</u> (קירוב בדרגה נמוכה): זוהי שיטה שמקטינה את ממדי מטריצות המשקל במודל. שיטה זו מתבססת על ההנחה שיש יתירות במטריצות המשקל וניתן לייצג את המידע החשוב בפחות ממדים. השיטה מיישמת טכניקת פירוק מטריצה לגורמים (כגון SVD) כדי למצוא שתי מטריצות קטנות יותר, שכאשר מכפילים אותן יחד, מתקרבות למטריצה המקורית.

- שיטת SVD (Singular Value Decomposition) – שיטה באלגברה ליניארית לפירוק מטריצה נתונה ל-3 מטריצות נפרדות כך שניתן יהיה להביע את המטריצה הגדולה באמצעות מטריצות בעלות ממדים נמוכים יותר. בכך אנו יכולים לקבל קירוב טוב של המטריצה המקורית תוך שימוש בפחות מימדים.

- לצורך השימוש ב-SVD יש צורך לספק כפרמטר את ב-Rank הרצוי למטריצה המקורבת, כך שישנו טרייד אוף בין דיוק גבוה יותר לRank גבוה יותר לבין עלותו החישובית.

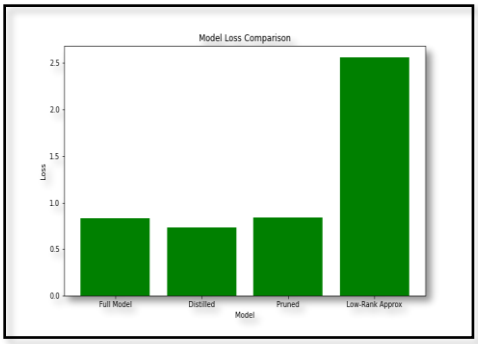
- כדי למצוא את הRank המתאים עבורנו נבדקו מודלים בהם בוצע קירוב עם Rank של: 30, 50, 75 ו-100 באמצעות ביצוע evaluation על ה- validation set (ראה נספח 3 לתוצאות המבחנים).

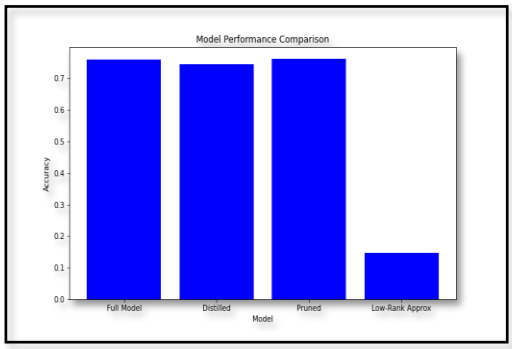
- מאחר והתקבלו תוצאות יחסית דומות, עבור כל אחד מה-Ranks הוחלט לכווץ את המודל כולו באמצעות Rank בערך של 30 שכן מבין האפשרויות שנבדקו מודל זה הביא לשגיאה הנמוכה ביותר.

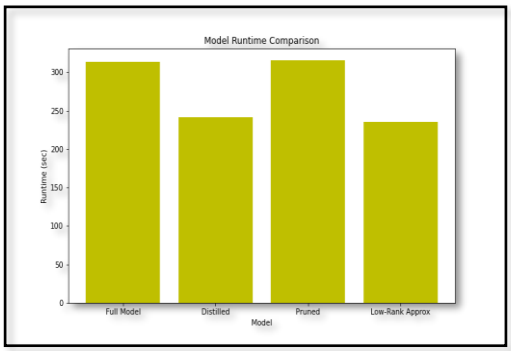
תוצאות כיווץ המודלים

להלן תוצאות הערכת ביצועי המודלים המכווצים, ביחס למודל המלא (הקוד המלא מפורט במחברת בשם compressing_the_model.ipynb):

Model	Evaluation Loss	Evaluation Accuracy	Evaluation F1 Score	Evaluation Runtime (sec)	Size (MB)
Full Model	0.8301	0.7603	0.759	313.66	255.5
Distilled Model	0.7356	0.7456	0.7429	241.4879	255.5
Pruned Model	0.836	0.7611	0.7596	315.1851	255.5
Low-Rank Approximation (Rank 30)	2.5574	0.1472	0.1153	235.347	255.5







מהתוצאות עולות הנקודות הבאות:

1. ראשית, מאחר וכלל המודלים בגודל זהה, פרמטר זה לא יבוא בחשבון בתהליך קבלת ההחלטות.
2. ניתן להבחין שהמודל שכווץ בשיטת "Low-Rank Approximation" הביא לביצועים הגרועים ביותר עם שגיאה מקסימלית ודיוק (Accuracy) מינימלי מבין המודלים שנבחנו. לחיוב נציין שזמן הריצה שלו הינו הקצר ביותר, אך ביצועיו נמוכים מדי מכדי שיישקל כמועמד מטעמי הטרייד אוף בין ביצועים לבין יעילות.
3. המודלים שכווצו בשיטות ה- "Knowledge Distillation" וה-"Pruning" נותרו כאפשרויות הטובות ביותר עבורנו וננסה להכריע ביניהם:
   - מודל ה-"Knowledge Distillation" מביא לשגיאה המינימלית מבין המודלים שנבדקו, ה-Accuracy שלו נמוך במעט משל המודל המלא ושל מודל ה-"Pruning" וכן זמן הריצה שלנו נמוך משמעותית משל המודל המלא ומודל ה-"Pruning" (בכ-25%).
   - מודל ה-"Pruning" מביא לשגיאה הקרובה לשגיאת המודל המלא וה-Accuracy שלו אף עולה עליו במעט. באשר לזמן הריצה, מודל זה הביא לתוצאה הארוכה ביותר.
4. בהתחשב בטרייד אוף בין ביצועים לבין יעילות, ייתכן שכדאי לבחור במודל שכווץ בשיטת ה-"Knowledge Distillation" שכן הוא שומר על מדדי ביצוע תחרותיים תוך כדי שהוא מספק זמן ריצה נמוך. עם זאת מתבקש להזכיר שההכרעה בין איכות הביצועים ליעילות נגזרת מאופי המטלה לשמה נוצר המודל, על כן ייתכן שנעדיף את מודל ה-"Pruning" או המודל המלא אם מטרתנו היא מודל הממקסם ביצועים.

דיון ומסקנות

בעבודה זו יישמנו תהליך של יצירת מודל למידה עמוקה עבור משימת חיזוי הקטגוריה אליה משתייכת כתבה ב-Irish Times על סמך כותרתה.

במהלך שלב עיבוד הנתונים המקדים, ביצענו מספר צעדים במטרה לדאוג שסט הנתונים עמו נאמן את המודל שלנו יהיה מחד בפורמט המתאים ביותר לתהליך האימון עבור אימון יעיל ונכון ומאידך מייצג בצורה איכותית את עולם הבעיה לקבלת יכולת הכללה מקסימלית. צעדים אלו כללו המרת כל האותיות לאותיות קטנות, מחיקת שורות כפולות ושורות עם ערכי NULL. כמו כן פישטנו את היררכיית הקטגוריות על ידי החלפת כל קטגוריות המשנה בקטגוריות העיקריות שלהן, והפחתנו את המספר הכולל של קטגוריות ייחודיות מ-103 ל-6 בלבד.

כדי להתמודד עם חוסר האיזון שנוצר במערך הנתונים עיבוד הקטגוריות, ביצענו Oversampling על הקטגוריות המיוצגות בתת-ייצוג ובאותו אופן ביצענו Undersampling על הקטגוריות בהן ישנו ייצוג יתר, מה שהוביל להתפלגות מאוזנת יותר על פני כל הקטגוריות. יתר על כן, בוצע עיבוד מקדים גם על קלט הטקסט עצמו: Tokenization, הסרת פיסוק ו-Stop-words וכן lemmazation. צעדים אלו נועדו להביא את סט הנתונים לתצורה המתאימה ביותר לאימון על גבי המודל.

לאחר מכן השווינו שני מודלים שאומנו מראש (Pre-Trained), DistilBERT ו-RoBERTa, עבור משימת סיווג הטקסט שלנו. לצורך השוואת ביצועיהם בוצע אימון מדגמי של כל מודל על 30% מהנתונים, אשר בסופו מצאנו ש-DistilBERT עלה במעט על RobERTa מבחינת accuracy, ציון F1 ו-precision . כתוצאה מכך, המשכנו לאמן את DistilBERT על כל מערך הנתונים.

בתהליך אימון המודל על כל מערך הנתונים נשמרו Checkpoints בכל 20,000 צעדים. מתוך כלל ה-checkpoint נבחרו ארבעת הטובות ביותר ועליהם בוצעה הערכת ביצועים על מערך הנתונים לפני איזונו (על מנת לקבל עוד אינפורמציה על יכולת הכללתם) ובחרנו מתוכם את המודל שהביא לביצועים הטובים ביותר כמודל הסופי שלנו.

אימון מודל DistilBERT על מערך הנתונים המלא הביא לשגיאה של 0.8301, Accuracy של 0.7603 וציון F1 של 0.759. כדי להפוך את המודל שלנו ליעיל יותר, יישמנו שלוש טכניקות דחיסה שונות: "Knowledge Distillation", "Pruning" ו- "Low-Rank Approximation".

השוואת הביצועים של המודלים הדחוסים הראתה תוצאות ברורות. מודל ה-"Distillation Knowledge" הפגין ביצועים טובים עם ירידה קלה בדיוק ובציון F1 מהמודל המלא, תוך הפחתה משמעותית של זמן הריצה של ההערכה. מודל ה- "Pruning" הציע שיפור קל ביחס למודל ה- "Knowledge Distillation" מבחינת דיוק וציון F1, אך בזמן ריצה גבוה יותר. לעומת זאת, מודל ה-"Low-Rank Approximation" הראה ירידה ניכרת בביצועים עם דיוק וציוני F1 נמוכים משמעותית.

בהתחשב בתוצאות אלו, נראה כי מודל ה-"Knowledge Distillation" מציע את האיזון הטוב ביותר בין ביצועים לבין יעילות חישובית. למרות ירידה קלה ברמת הדיוק ובציון F1 בהשוואה למודל המלא, יש לו מהירות מרשימה בתהליך ה-evaluation, מה שהופך אותו למתאים ליישומים שבהם זמן הריצה חשוב.

עם זאת, חשוב לציין שבחירת המודל תלויה במידה רבה ביישום. לדוגמה, אם השיפור הקל בביצועים הוא קריטי, מודל ה- "Pruning" עשוי להיות בחירה טובה יותר למרות זמן הריצה הארוך שלו. מודל ה-"Low-Rank Approximation", לעומת זאת, ידרוש עוד התאמות לשיפור ביצועיו, שהיו נמוכים משמעותית משל המודלים האחרים.

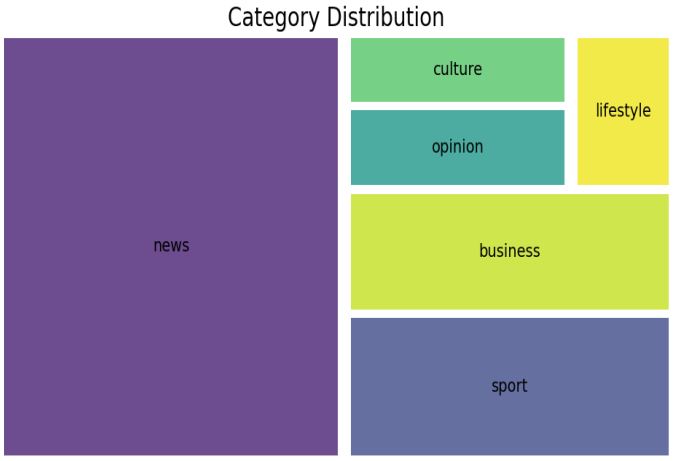
סיכום

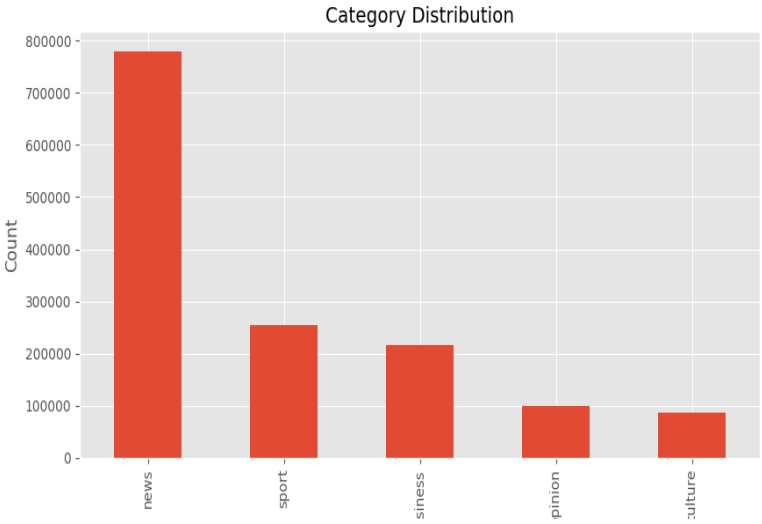
מאמר זה בוחן את היישום של מודלים מבוססי למידה עמוקה עבור משימות סיווג טקסט, תוך חיזוי ספציפי של קטגוריית נושא של כתבה מתוך כותרות טקסט. לאחר עיבוד מקדים של מערך נתונים של 1.61 מיליון כותרות מה-Irish Times, אימנו את המודלים DistilBERT ו-RoBERTa על הנתונים. נמצא כי DistilBERT מתעלה במעט על RobERTa, ולכן נבחרה לשימוש נוסף. מודל DistilBERT המאומן כווץ לאחר מכן באמצעות שלוש טכניקות: Knowledge Distillation, Pruning ו- low-rank approximation . תוצאות השוואת המודלים המכווצים השונים הציגו את המודל שכווץ באמצעות "Knowledge Distillation" כמבטיח ביותר, המספק איזון בין ביצועים ויעילות חישובית. עבודה זו העלתה תובנות חשובות ליישום מודלים יעילים עבור משימות סיווג טקסט להן שימושים רבים בתעשייה.

נספחים

נספח 1 - חקר הנתונים ועיבוד מקדים

1. מערך הנתונים המשמש בפרויקט זה, שהתקבל מ-Kaggle, מורכב מכ-1.61 מיליון כותרות חדשות שפורסמו על ידי ה-Irish Times על פני 25 שנים, מ-1 בינואר 1996, עד 30 ביוני 2021. The Irish Times, הוקם במשך 160 שנים לפני, מספק פרספקטיבה מקיפה וארוכת טווח על אירועים באירופה, מה שהופך את מערך הנתונים הזה למשאב מצוין ללימוד סיווג כותרות חדשות.
2. הנתונים מאוחסנים בקובץ CSV, כאשר כל רשומה כוללת שלוש תכונות: תאריך הפרסום (publish_date), קטגוריית הכותרת (headline_category), וטקסט הכותרת (headline_text). טקסט הכותרת כתוב באנגלית, מקודד בערכת תווים UTF-8, בעוד שקטגוריות הכותרות מופרדות בנקודות ובפורמט ASCII באותיות קטנות.
3. חלק בלתי נפרד מהפרויקט הזה היה עיבוד מקדים וניקוי הנתונים כדי לשפר את ביצועי המודל אותו נרצה לאמן. כל נתוני הטקסט הומרו לאותיות קטנות כדי להבטיח אחידות, וערכים כפולים (בעלי קטגוריה וכותרת זהים) הוסרו. כל הרשומות עם ערכים חסרים או אפסים הוסרו גם כן ממערך הנתונים.
4. כדי להקל על הניתוח, השדה publish_date הוחלף ב-3 שדות חדשים: Year , month ו-day, המכילים כל אחד את רכיב התאריך מהשדה המקורי.
5. השדה headline_category דרש שלב עיבוד מוקדם יותר מורכב. במקור, מערך הנתונים כלל 103 קטגוריות שונות, אשר רבות מהן היוו קטגוריות משנה של קטגוריות אחרות. לדוגמה, 'news.health' ו-'news.politics' נחשבו מבחינתנו תת-קטגוריות של קטגוריית 'news'. על כן התווסף שדה חדש, 'primary_category', כדי לאחסן את הקטגוריות הרחבות יותר הללו, ולמעשה צמצמה את מספר הקטגוריות הייחודיות מ-103 לשש: "news", "sport", "business", "culture", and "opinion", "lifestyle"".





6. לאחר צמצום מספר הקטגוריות השונות התקבל מערך נתונים שאינו מאוזן:

"News," had 778,288 records.	"Opinion," had 99,609 records.
"Sport," had 254,447 records.	"Culture," had 86,292 records.

"Lifestyle," had 85,457 records.	"Business," had 216,264 records.

על מנת להימנע מבעיות הנגרמות מסט נתונים שאינו מאוזן הוחלט לאזן אותו באופן הבא:

- סט הנתונים פוצל ל-2 קבוצות: קבוצת הרוב וקבוצת המיעוט, כך שקבוצת הרוב מורכבת מ-3 הקטגוריות הגדולות ביותר ('business', 'sport', 'news') וקבוצת המיעוט מורכבת מ-3 הקטגוריות הקטנות ביותר ('lifestyle', 'culture', 'opinion').
- את הרשומות בקבוצת הרוב הוחלט לצמצם באמצעות "Undersampling" כך שכולן יכילו את מס' הרשומות של הקטגוריה הקטנה ביותר מתוך קבוצת הרוב (Business).
- את הרשומות בקבוצת המיעוט הוחלט להגדיל באמצעות "Oversampling" כך שכולן יכילו מס' רשומות השווה לחציון של מס' הרשומות לכל קטגוריה (157,936) רשומות. ההגדלה בוצעה על ידי שכפול רשומות קיימות באופן רנדומלי.

בסיום התהליך מספר הרשומות לכל קטגוריה היה כזה:

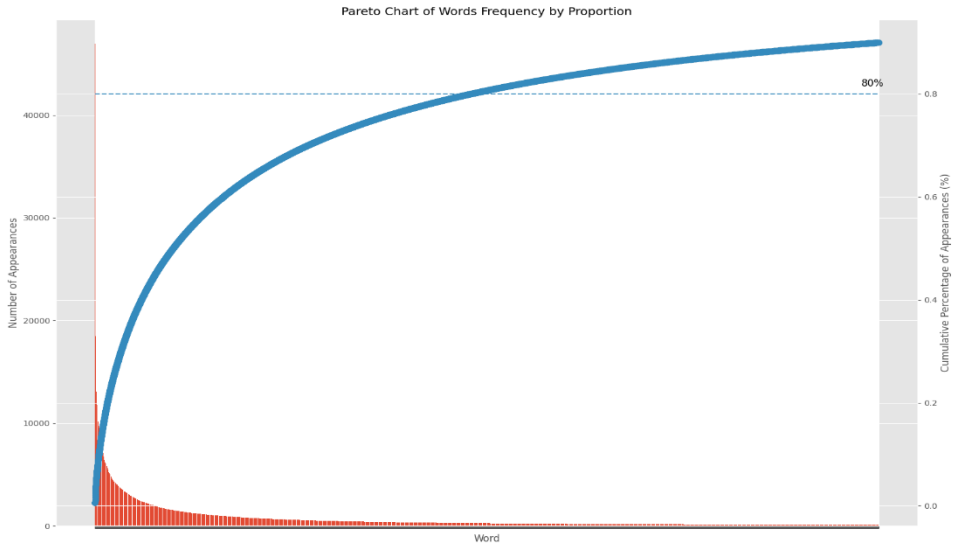
"Opinion," had 157,936 records.	"News," had 216,264 records.
"Culture," had 157,936 records.	"Sport," had 216,264 records.
"Lifestyle," had 157,936 records.	"Business," had 216,264 records.

וסט הנתונים אמנם לא היה מאוזן בצורה מלאה אך חולק ל-2 קבוצות בגדלים זהים שהפער בין הגדלים של כל קבוצה זניח משמעותית לעומת המצב שלפני האיזון.

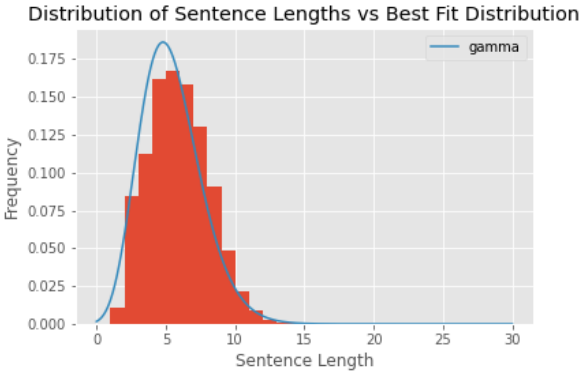
- הערה – נציין כי לצורך הליך ה"Oversampling" נשקל שימוש ב"augmentations" לצורך מניעת שכפול רשומות קיימות, אך עקב מורכבות ביצוע פעולה זו על קלט טקסטואלי (לעומת תמונה, לדוגמה) הרעיון ירד מהפרק.

7. ניתוח נתוני הטקסט –

- כמות המילים ייחודיות – עולה כי מערך הנתונים הכיל 107373 מילים ייחודיות אשר הרכיבו את הכותרות השונות. עם זאת, תת-קבוצה של מילים אלה, בערך 20,000, כיסתה 95.5% מכלל מופעי המילים השונים, מה שמרמז על התפלגות מוטה מאוד של תדירות המילים (כלומר מס' קטן של מילים מופיע בתדירות גבוהה בעוד שרוב המילים מופיעות בתדירות נמוכה בהרבה). להלן תרשים פרטו המציג את תדירות המילים השונות בנתונים:



- התפלגות אורך הכותרת – לצורך מציאת ההתפלגות הסטטיסטית המתאימה ביותר לתיאור אורך הכותרות במערך הנתונים נבחנו התפלגויות מועמדות: נורמלית, מעריכית, גאמא והתפלגות וייבול. לאחר בדיקת התאמת כל אחת מההתפלגויות על הנתונים נבחרה ההתפלגות אשר הביאה לשגיאה ריבועית מינימלית והיא:



  - התפלגות גאמא עם הפרמטרים: $(13.417, -2.70, 0.604)$
  - תוחלת ההתפלגות הינו: 8.109
  - ס"ת להתפלגות הינה: 2.213

  על כן ניתן לומר ש- 99.7% מהכותרות מכילות 15 מילים או פחות (תוחלת + 3 ס"ת).

8. כדי להכין עוד יותר את נתוני הטקסט לאימון, סדרה של שלבי עיבוד מקדים הוחלו על השדה headline_text אשר כללו: tokenization, הסרת סימני פיסוק והסרת stop-words. בנוסף, בוצע הליך lemmatization על מנת להעביר מילים לצורת הבסיס שלהן תוך שמירה על קונטקסט. התוצאה של תהליך זה נשמרה בשדה חדש 'preprocessed_headline_text', מה שמבטיח שטקסט הכותרת המקורי נשאר ללא שינוי לצורך ניתוח נוסף אפשרי.

לסיכום, סט נתונים זה מהווה מקור עשיר של נתוני טקסט מהעולם האמיתי, המכסה מגוון רחב של קטגוריות לאורך תקופה ממושכת. העיבוד המקדים והניתוח הזהיר של הנתונים מהווים את הבסיס לשלבי בחירת המודל, ההדרכה וההערכה הבאים של פרויקט זה.

נספח 2 – אפשרויות למודל Pre-trained

המודל	טיעונים בעד	טיעונים נגד
משפחת מודלי BERT: BERT, DistilBERT, RoBERTa, MobileBERT, ALBERT, XLM-RoBERTa	○ מודל שאומן מראש על כמויות מידע אדירות, מסוגל לזהות דפוסים לשוניים מורכבים. ○ מודל דו כיווני – יכול ללמוד על משמעות של המילה על ידי המילים הסובבות אותה בשני הכיוונים. ○ ישנן גרסאות "קלות" יותר של המודל ה"כבד" אשר מספקות ביצועים טובים.	○ הגרסאות הכבדות הינן יקרות חישובית בתהליך אימון המודל, גם עבור fine-tuning.
GPT-2	○ מודל רב עוצמה עם אוצר מילים גדול שהוכשר מראש ליצור טקסט דמוי אדם.	○ מכיוון שמיועד ליצירת טקסט הוא עשוי להתאים במידה פחות למשימות סיווג.
ELECTRA (Efficiently Learning an Encoder that Classifies Token Replacements Accurately)	○ בהשוואה למודלי BERT ה"כבדים" זהו מודל יעיל יותר וקטן יותר המספק ביצועים טובים יותר.	○ יקר חישובית – למרות שיעיל יותר מ-BERT הוא עדיין מודל גדול ועשוי להוביל לעלות חישובית גבוהה.

נספח 3 - מטריקות מאימון 1

Step	Training Loss	Validation Loss	Accuracy	F1	Precision	Recall
20000	1.1573	0.839971	0.703461	0.6994	0.701799	0.703461
40000	1.1064	0.801386	0.718163	0.714676	0.715233	0.718163
60000	0.9873	0.793097	0.725579	0.722155	0.723193	0.725579
80000	1.0066	0.77787	0.733779	0.7314	0.735113	0.733779
100000	0.9906	0.720588	0.748584	0.746574	0.749093	0.748584
120000	0.8795	0.785279	0.746187	0.743578	0.7462	0.746187
140000	0.87	0.776553	0.749399	0.746094	0.748686	0.749399
160000	0.8909	0.747115	0.756164	0.752946	0.756161	0.756164
180000	0.8412	0.84541	0.755657	0.75334	0.755301	0.755657

נספח 4 - מטריקות מאימון 2

Step	Training Loss	Validation Loss	Accuracy	F1	Precision	Recall
200000	0.7741	0.846793	0.758137	0.756288	0.758637	0.758137
220000	0.7613	0.830081	0.760298	0.758992	0.760209	0.760298
240000	0.6569	0.984294	0.757153	0.755957	0.758187	0.757153
260000	0.6501	0.98895	0.760565	0.759061	0.760155	0.760565
280000	0.6407	0.99045	0.759692	0.758144	0.760109	0.759692

נספח 5 – תוצאות evaluation על גדלי Rank שונים

Rank	Evaluation Loss	Evaluation Accuracy	Evaluation F1 Score	Evaluation Runtime
30	2.5574	0.1472	0.1153	235.35
50	2.5576	0.1472	0.1153	227.31
75	2.5579	0.1472	0.1154	228.03
100	2.5582	0.1472	0.1154	229.64