מאמר – נושאים מתקדמים בלמידה עמוקה – קבוצה 4

מגישים : אסף איתן, עבד אלמג'יד זועבי, יאזן שנבור

1. מבוא

בשנים האחרונות הגדילה בצריכת התוכן הדיגיטלי אשר נבעה מזמינות המידע ברשתות השונות, הביאה להגברה במשימות מעולם ה.(NLP (Natural Language Processing

מאמר זה נכתב כחלק מפרויקט סוף בקורס 'נושאים מתקדמים בלמידה עמוקה' באוניברסיטת תל אביב.

מטרתנו בפרויקט הייתה להשוות בין מודלים מאומנים (Pre-trained models) אשר עברו התאמה למשימה אותה קיבלנו – ביצוע סיווג לכותרות אשר פורסמו בעיתון 'The Irish Times' לקטגוריה, זאת תוך השוואה בין שיטות אימון וכיווץ שונות.

המידע איתנו עבדנו ואימנו את המודלים היה מבוסס על כתבות מעיתון ה-The Irish Times מהאתר Kaggle , הכוללות מידע משנת 1996-2021 ומכילות כ-1.61 מיליון כתבות וסוגן.

הפרויקט מורכב משני חלקים מרכזיים, החלק הראשון של הפרויקט כלל ניתוח מעמיק של בסיס הנתונים על מנת להבין בצורה טובה יותר את המידע זאת תוך ביצוע : עיבוד מקדים, חקר טרנדים, הוצאת חריגים ועוד.

חלק ב' והמרכזי של הפרויקט כלל יישום של שני מודלים מאומנים והתאמתם לביצוע המטלה הספציפית אותה קיבלנו, תוך השוואה בין שלוש חלופות כיווץ שונות והצגת הממצאים.

חיפוש המודלים המתאימים בוצע על ידי היכרות ולמידה של מודלים חדשים כחלק מהרצאות ותרגולי הקורס וכן על ידי חקר מעמיק באינטרנט.

2. עיבוד מקדים

במאמר, כחלק מביצוע הפרויקט בחלקו הראשון, ביצענו עיבוד מקדים לנתונים אשר קיבלנו על מנת להכין את המידע למודלים עליהם הוא יאומן כדי לסווג את סוג כותרת העיתון וכן על מנת להבין יותר טוב את הנתונים.

בחלק זה, בחרנו לבצע מספר צעדים על מנת לשפר את איכות המודל שלנו ואת איכות התוצאות של האימון.

תחילה, בחנו את ממדי בסיס הנתונים ובחנו האם יש כפילויות או רשומות ריקות. מצאנו כי יש 7 כתבות על טקסט ריק ולכן הסרנו את הכתבות הללו מבסיס הנתונים שלנו.

דבר שני, מאחר ובסיס הנתונים כלל 103 קטגוריות הכללו קטגוריות משנה אשר בכל אחת מהן כמות לא גדולה של כתבות, בחרנו לבצע איחוד של קטגוריות משנה לכדי קטגוריות ראשיות ובכך צמצמנו את 103 הקטגוריות לכדי שש קטגוריות ראשיות המתארות את כלל הכתבות, זאת על ידי ביצוע פיצול לקטגוריה ראשית אחרי הנקודה הראשונה של תיאור הקטגוריה. עיבוד זה, נעשה כדי להימנע מהתאמת יתר בהמשך העלול להיווצר מכמות נמוכה מאוד של נתוני אימון בהרבה קטגוריות אשר המודל צריך ללמוד.

לאחר עיבוד זה, מצאנו כי כמות הכתבות בין שש הקטגוריות הייתה לא מאוזנת, לקטגוריית "news" היו ~800 אלף כתבות ואילו ליתר הקטגוריות כמות הכתבות הייתה בין 96 אלף לבין 250 אלף. לכן, על מנת להימנע ממצב של התאמת יתר של המודל לקטגוריות בהן יש כמות גדולה של כתבות לעומת

אחרות בהן יש פחות, בחרנו לבצע 'איזון' של הנתונים בין הקטגוריות בצורה כזו שלכל הקטגוריות תהיה

כמות זהה של כתבות – 96 אלף כתבות, סה"כ צמצמנו את בסיס הנתונים לכ- 576 אלף שורות. בנוסף, כחלק מהעיבוד המקדים בשלב זה של הפרויקט, החלטנו לבצע הסרה של "Stop Words" מעמודת "שם הכתבה.

מילים מסוג 'Stop Words' הן מילים אשר מאוד שכיחות בשפה אנגלית אשר לא מייצרות משמעות מספיקה למשפט, מילים כמו : 'is','the','and' וכו'.

על ידי הסרת מילים אלו, אנחנו שואפים להקטין את מורכבות המודל וכן להסיר רעש לא נחוץ לניתוח המשפטים. הסרת מילים

אלו מסייעת לנו לאמן את המודל על המילים היותר משמעותיות והחשובות ולא לייצר הטיה ומורכבות כתוצאה ממילים פחות משמעותיות ושכיחות.

כמו כן, מאחר וחלקו הראשון של הפרויקט עוסק בהבנה של המידע בצורה מעמיקה יותר, על מנת לבצע חקר של טרנדים בין הקטגוריות השונות, בחרנו להוסיף 3 עמודות נוספות: 'יום', 'חודש', 'שנה'. הוספה זו, סייעה לנו לייצר גרפים, בחתכים שונים כדי להבין את התפלגות הכתבות מהסוגים השונים במהלך תקופת הנתונים .

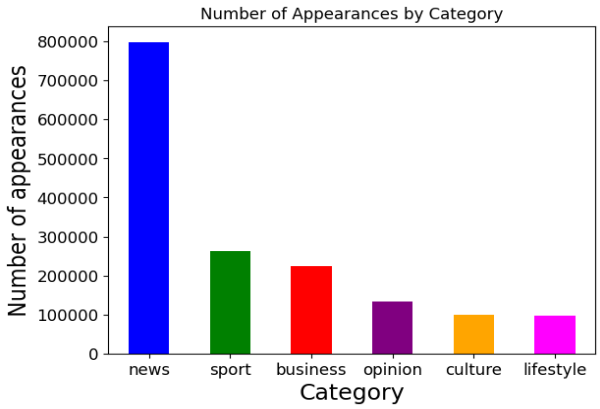
Category	# Rows
News	96,000
Culture	96,000
Opinion	96,000
Business	96,000
Sport	96,000
Lifestyle	96,000

טבלה 1: כמות הכתבות בכל קטגוריה לאחר העיבוד המקדים

3. ניתוח המידע

בחלק זה בוצע ניתוח של המידע והבנה מעמיקה יותר של הפרמטרים השונים.

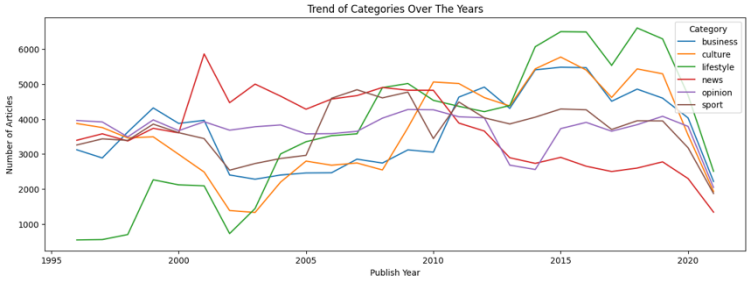
תחילה, על מנת להבין את התפלגות סוגי הכתבות לאחר שביצענו את העיבוד המקדים של איחוד קטגוריות משנה לכדי קטגוריות על :



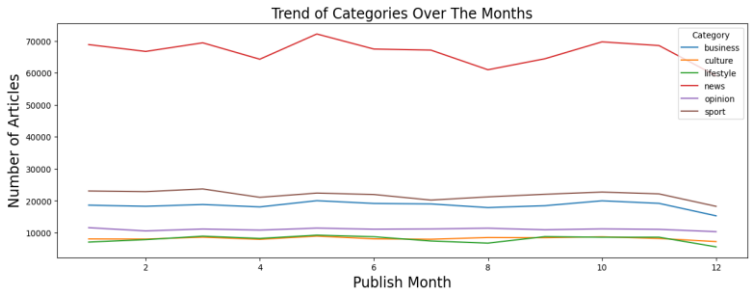
בוצע ניתוח של מספר הכתבות לפי סוג כפי במוצג בתרשים 1.

ניתן לראות לפי הגרף לעיל את מספר הכתבות לפי הקטגוריות כך שכתבות מסוג "news" הן בעלות כמות ההוצאה לאור הגבוהה ביותר . מסקנה זו הביאה אותנו גם לבצע עיבוד מקדים של איזון הקטיגוריות, אשר משפר את ביצועי המודל ועשוי למנוע מהמודל להיות מוטה.

כמו כן, בוצע ניתוח של כמות ההוצאה לאור של כתבות לכל קטגוריה לאורך זמן : יומי, חודשי או שנתי.



גרף 2: טרנד של כמות כתבות לפי סוג לאורך שנות פרסום הכתבות



גרף 3: טרנד של כמות כתבות לפי סוג בחתך של חודשי פרסום

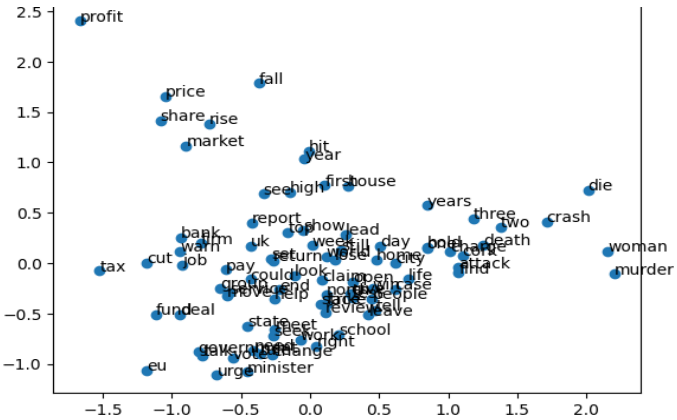
למשל אם נתמקד בגרף 2 , ניתן לראות שבשנה האחרונה כמות הכתבות שמוצאות לאור הולך יחסית לשאר השנים, דבר המעיד העשוי להעיד על האטה של הוצאה לאור בשוק.

בחלק זה, בחרנו לבצע Word Embedding מסוג Word2Vec,לצורך ביצוע זה עשינו את העיבוד המקדים של הסרת stop words. מטרת ניתוח זה היה על מנת לבצע ניתוח סמנטיקה של הקשרים בין המילים. טכניקה זו לומדת את הרפרזנטציות של המילים בכותרות של הכתבות על בסיס הקונטקסט שלהן.

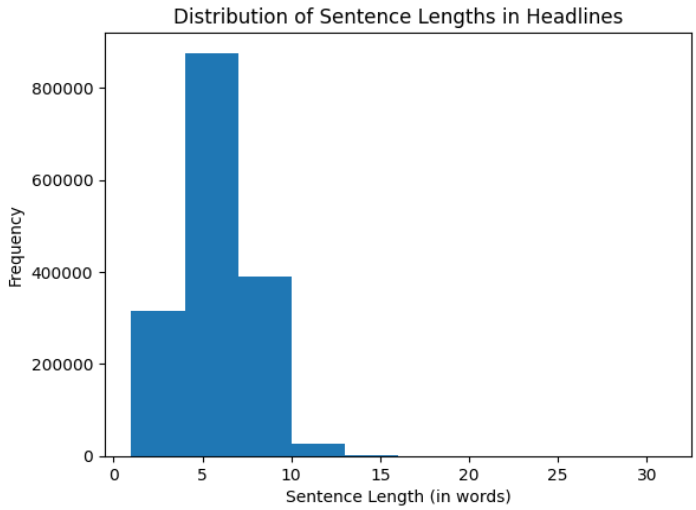
מודל זה, מייצג מילים כווקטורים במרחב בעל ממדים גבוהים, בדרך כלל עם מאות ממדים.

לכל מילה מוקצית נקודה במרחב אשר בה כל ציר מייצג היבט אחר של משמעות המילה.

על מנת להציג את הפלט בצורה קלה להצגה גרפית צמצמנו את הממדיות לשני ממדיים. על ידי ניתוח PCA. על ידי שרטוט וקטורי של המילה במרחב הדו-ממדי המוגדר על ידי שני המרכיבים העיקריים, נוכל לייצג כל מילה כנקודה. הדמיה של וקטורי מילים במרחב הדו-ממדי מאפשרת לנו לקבל תובנות לגבי יחסי מילים על סמך הדמיון וההבדלים במשמעות שלהם. מילים דומות מבחינה סמנטית יהיו קרובות יותר זו לזו במרחב הדו-ממדי, בעוד שמילים לא דומות יהיו רחוקות יותר זו מזו. לדוגמה ניתן לראות שקבוצת המילים (profit, price, fall, share ,rise, market) נמצאים בצד השמאלי העליון מהתרשים שממחישים את הקרבה של מילים מתחום הכלכלה והבורסה, ועוד קבוצת מילים (crash, die, woman, murder, death) על כתובות מסוג "חדשות" בעל פן שלילי ורע וניתן להסיק מקבוצה זו שיש יחסית הרבה כתבות מהסוג של אלימות נגד נשים.



גרף 5: פלט Word2Vec – ניתוח סמנטיקה בין מילים

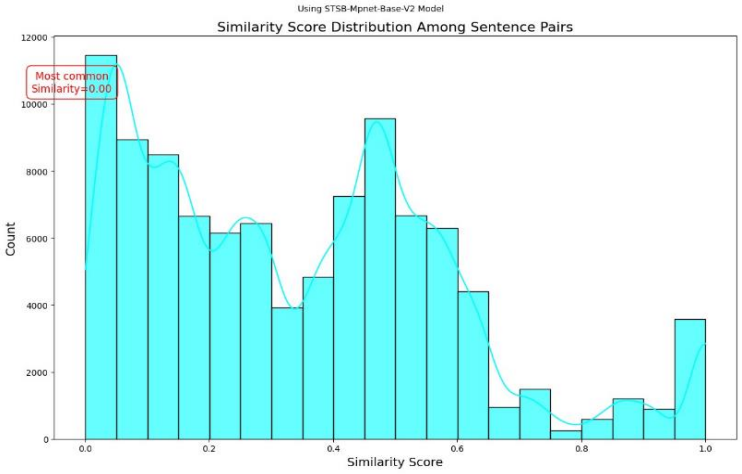


גרף 6: פלט Word2Vec – ניתוח סמנטיקה בין מילים

```
Mean sentence length: 5.224

Standard deviation of sentence lengths: 2.009
```

מטרת גרף 6 המתואר לעיל היא לנתח את סטטיסטיקת את אורכי נתוני הטקסט של הכותרות, כך חושב האורך של כל משפט. לאחר מכן חישבנו את הממוצע וסטיית התקן של אורכי הכותרות כפי שמוצג לעיל.



גרף 7: מודל STSB

ביצענו שימוש במדד הSTSB אשר נלמד בהרצאה. המדד משמש לעתים קרובות בתחום עיבוד השפה כדי להעריך עד כמה מודלים יכולים לקבוע את הדמיון הסמנטי (קשור למשמעות) בין זוגות משפטים.

על ידי התבוננות בגרף, נוכל לראות את טווח ציוני הדמיון ואת הציונים הנפוצים ביותר עבור מדגם של משפטים אשר לקחנו מהדאטה שלנו, זה נותן לנו דרך למדוד את הביצועים של המודל.

<u>להלן מספר דוגמאות ממאגר המידע שלנו</u>:

1. Similarity between 'irish service sector post strong growth' and 'activity irish service sector grow fast rate': 0.9059
2. Similarity between 'economy grow fast expect' and 'economy grow fast pace year': 0.8423
3. 'chelsea confirm robinho bid' and 'chelsea confirm offer robinho': 0.9051
4. Similarity between 'good film cinema weekend' and 'good movie big screen weekend': 0.9069

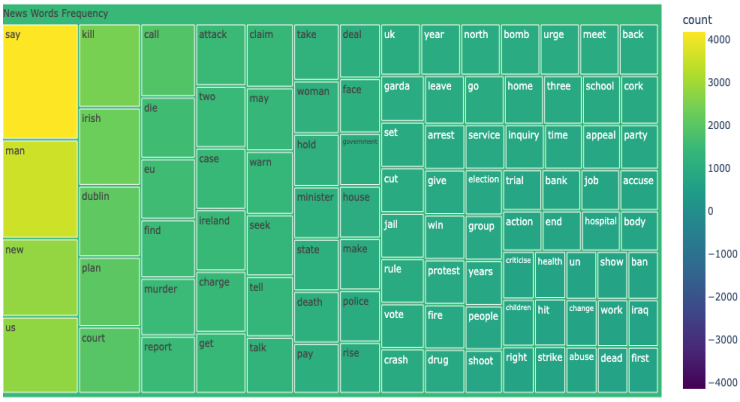
ניתן להסיק מתוצאות גרף הSTSB שישנו מספר גדול של ציוני דמיון אפסיים, כלומר הרבה זוגות סימנטים של משפטים בעלי רמת דמיון אפסית, אנו מעריכים שדבר זה יפגע ב-finetuning בהמשך.

הסיבה לכך היא שהמודל עלול להיות מוטה לניבוי דמיון נמוך, מה שגורם לו לזלזל בדמיון למשפטים קשורים. לכוונון עדין אופטימלי, חשוב שיהיה מערך נתונים מאוזן עם ציוני דמיון מגוונים.

בגלל שכבר בעיבוד המקדים בצענו הורדות הרבה כמות של דאטה החלטנו להשאיר את כמות הנתונים במצב הנוכחי מחשש ל-underfitting.

כמו כן, ביצענו ניתוח בא לבחון את המילים הנפוצות ביותר בכל קטגוריה.

בגרף 8 ניתן לראות דוגמה עבור כותרות המסווגות בתור חדשות. בגרף עץ זה, מתבצע ספירה של מספר הפעמים שכל מילה מופיעה בכל הכותרות הללו. הנתונים המתקבלים מאורגנים בטבלה, כאשר עמודה אחת מפרטת את המילים ועמודה אחרת מפרטת את ספירת המילים המתאימות.בתמונה זו, כל אחת מ-100 המילים הנפוצות ביותר מיוצגת על ידי תיבה. גודל התיבה מראה באיזו תדירות מופיעה המילה, וצבע התיבה משקף גם את תדירות המילה.



גרף 8: מפת עץ- שכיחות מילים עבור קטגוריית חדשות

ראה נספח לתרשימים עבור שאר הקטגוריות ובגודל מוגדל.

4. בחירת המודלים:

בחלק זה נציג את הליך בחירת המודלים המאומנים וביצוע ה-fine-tuning על מנת להתאים אותם למשימה שלנו.

לאחר שחקרנו אילו מודלים עשויים להתאים למשימת סיווג טקסט, בחרנו להתמקד בשלושה מודלים : Bert, Roberta, Distelbert.

מודל Bert הוא מודל מבוסס טרנספורמר אשר הוצג לראשונה בשנת 2018 ונחשב למודל מוכר עבור משימות של NLP.

המודל מכיל כמה שכבות של טרנספורמרים ונמצא בשימוש רחב עבור משימות של text classification. אחד היתרונות של המודל הוא שיש לו אוצר מילים רחב מאוד אך ביתרון זה טמון גם חסרון של אילוץ זיכרון בעת האימון מחדש.

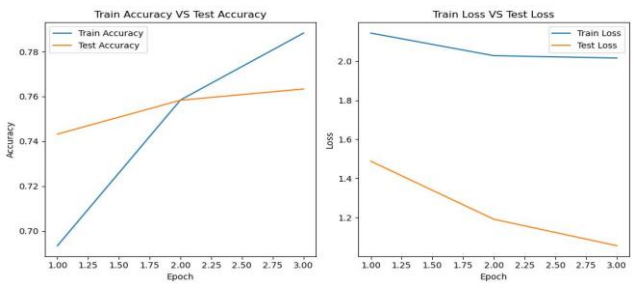
מודל ה-Roberta הוצג לראשונה בשנת 2019 והוא מודל מחודש של מודל Bert . המודל נוצר על מנת לענות על חלק מהמגבלות והאילוצים של מודל Bert. מבנה המודל דומה למודל Bert אך מנצל כמות גדולה יותר של דאטה וכן אומן זמן ארוך יותר.

מודל ה-DistilBERT הוצג לראשונה גם הוא בשנת 2019 ומהווה גרסה קומפקטית של BERT. המודל שומר על רוב היכולות של מודל ה BERT תוך צמצום במספר הפרמטרים המקורי של מודל הBERT. תהליך זה קורה על ידי שיטת Distillation בה המודל מנסה לחקות אחר מודל ה BERT, על ידי שימוש זה, המודל יכול להשוות את ביצועיו למודל Bert תוך שימוש במספר קטן יותר של פרמטרים.

על מנת להשוות בין שלושת המודלים, לקחנו 40% מהדאטה וביצענו fine-tuning של המודלים למשימה שלנו.

<u>לאחר אימון המודלים התקבלו התוצאות הבאות</u>: (עבור ה-epoch האחרון)

מודל	דיוק אימון	דיוק מבחן
BERT	78.82%	76.33
ROBERTA	75.24%	74.5%
DistilBERT	77.44%	75.55%



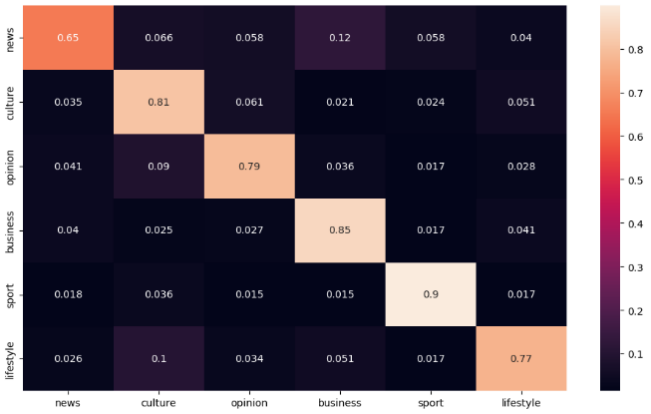
גרף 8: גרפי האימון והמבחן של רמת הדיוק והשגיאה של מודל BERT

מהתוצאות ניתן לראות כי מודלים Bert ו DistilBert קיבלו את הדיוק הרב ביותר ולעומת מודל Roberta.

למרות השימוש בהיפר פרמטרים זהים ובאותו מערך נתונים לFinetuning, השונות בביצועים בין BERT, DistilBERT ו-Roberta יכולה להיות מיוחסת בעיקר להבדלים בארכיטקטורות ובשיטות הPre-Trained שלהם. המורכבות של BERT עשויה לאפשר לו ללכוד דפוסים מורכבים יותר, ולתרום לביצועים המעולים שלו. DistilBERT, בהיותה גרסה קטנה ומהירה יותר של BERT, עשויה שלא ללכוד את כל הניואנסים למרות היעילות שלה, מה שיכול להסביר את הירידה בביצועים שלה. לבסוף, ייתכן שגישת ה-Pre-trained שבוצע עליו והאופן בו מבוצע הtokenization השונה של Roberta לא היו מועילים עבור מערך הנתונים הספציפי שבו נעשה שימוש, מה שעלול להוביל לתפקוד נמוך שלו בהשוואה לשאר המודלים.

לכן בחרנו להמשיך עם מודל BERT ולבצע איתו את בדיקות שיטות הכיווץ.

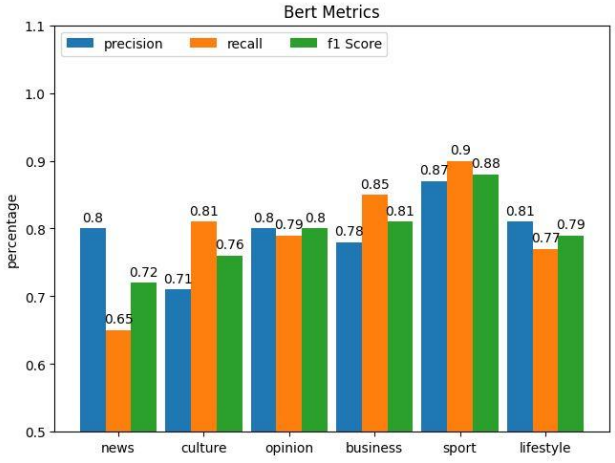
מתוצאות פלטי ה-confusion matrix ושאר פרמטרי הדיוק אפשר להסיק



גרף 9: תרשים confusion matrix של תוצאות מודל BERT

ניתוח Confusion Matrix

בעזרת המטריציה ניתן לפשט את התוצאות של המודל, ולחזות נקודות חולשה שניתן לשפר במודל שלנו, אם מסתכלים על התוויות עם רמת הדיוק הנמוכה בנתונים שיש לנו (כמו lifestyle ו-news) ניתן לראות שבאותה שורה יש תווית אחרת שהיא כן משפיעה ומושפעת, וזה נותן אנדיקציה לדימיון בין הקלטים של תוויות שונות, לצורך הבנה ניקח דוגמא מהנתונים שלנו ניתן שהמודל זהה כ – 65% מהכותרות שמשויכות ל- news ובנוסף המודל מיין כ – 12% ממה שהיה אמור להיות news כ- business וזה יכול להיות קשור לכל מיני סיבות שהן לא בהכרח סטטיסטיות אלא קשורות לאופן איסוף הנתונים והמילים שהעיתונות משתמשת בהן לתאר כתבה עסקית או ההפך את המילים שמשמשים לתיאור חדשות רגילות, לשיפור המצב הזה בעתיד עומד בפנינו שתי אפשרות, אפשרות ראשונה היא הוספת נתונים מהסוגים המצויינים ולבצע Reinforced Learning או לחפש מודלים שהם יותר טובים ממה שיש לנו בזיהוי הכותרות האלה ולבצע Distilation.



גרף 10: תרשימי precision/recall/f1 score של תוצאות מודל BERT

ניתוח Recall, Precision, f1 Score

המדדים האלו נותנים לנו סיכום למצב חיזוי המודל עבור כל Label אשר מראה לנו עד כמה ניתן לסמוך על המודל בהינתן חיזוי כלשהו ובהתאם לצרכי המשתמש נבחר אם כן נשתמש בחיזוי המודל או לא, לדוגמא נבחר להסתכל על הקטגוריות הבאות, news ו- culture, ב news יש precision גבוהה לעומת ה-recall וההפך מתקיים ב-culture מה שזה אומר אם שאם משתמש כלשהו מנסה למיין כותרות והוא רוצה לתפוס כמה שיותר כותרות מסוג culture שהן באמת מסוג זה הוא כן יבחר במודל אבל הוא יקבל הרבה כתבות שאינן culture ממוינות כ-culture ואם הוא רוצה למצוא את הכותרות שהן בסבירות גבוהה מקטגוריה כלשהוי עם כמה שפחות חיזויים לא נכונים הוא יבחר במודל שלנו שיש במידה ועבור הקטגוריה יש precision גבוה.

5. בחינת שיטות כיווץ:

לאחר שבחרנו את מודל x מהפרק הקודם, בחלק זה, נרצה לבחון שלוש שיטות כיווץ על המודל הנבחר.

בחרנו להתמקד בשלוש שיטות כיווץ:

- Pruning
- Distillation
- Quantization

שיטת ה-Pruning (L1) הינה שיטת כיווץ אשר מסירה משקולות לא רלוונטיות ברשת ובכך מקטינה את מספר הפרמטרים של המודל ואת הזיכרון שלו. הרעון הוא להסיר משקולות מהמודל אשר בעלי משקל נמוך וללא משמעות גדולה על האימון.

שיטת הDistillation הינה שיטת כיווץ אשר מתבססת על אימון של מודל קטן יותר, מודל סטודנט, אשר מנסה לחקות את מודל מורכב יותר , המודל המלא. ישנו שימוש בשיטה זו מאחר ואימון המודל הקטן יותר, מודל הסטודנט, לרוב מביאה לביצועים דומים למודל המלא תוך הקטנת מורכבות המודל ומספר הפרמטרים בו.

בתרחיש זה, מודל ה Roberta, שהוא גדול יותר ולעיתים מדויק יותר, שימש כמודל המורה למודל ה BERT הקטן יותר. המטרה הייתה להשתמש בביצועים המעולים של Roberta כדי לשפר את היכולות של BERT תוך שמירה על היעילות שלBERT לאחר ה -Distillation אנו מצפים לראות שיפור בביצועים של מודל BERT

שיטת ה- Quantization הינה שיטת כיווץ אשר מכוונת להקטנה של הזיכרון ומורכבות המודל, זאת על ידי הורדת הדיוק של משקולות המודל ומשקלן.

ממצאים:

לאחר הרצת שלושת שיטות הכיווץ מהפרק הקודם, בחנו את אימון המודל כעת ושמרנו את תוצאות המודל כפי שהן מוצגות לטבלה המסכמת למטה.

ביצענו לאחר אימון המודל ובחנו את נתוני המבחן לאחר ביצוע הכיווץ בשיטות השונות, בחרנו לא לבצע עוד פעם אימון מחשש להתאמת יתר

שיטת כיווץ	דיוק מבחן
Pruning	77.1%
Distillation	75.99%
Quantization	74%

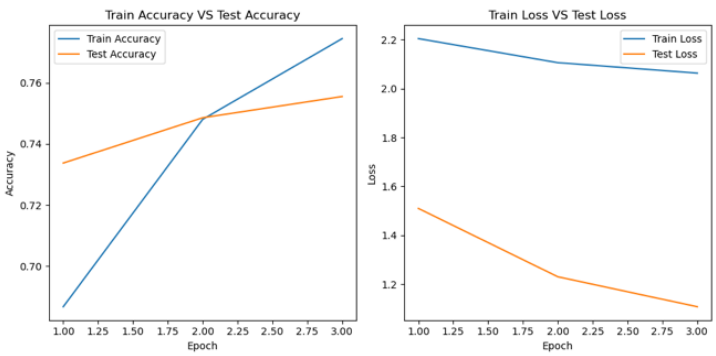
לאחר הפעלת שלוש שיטות הכיווץ, התוצאות הוערכו ונשמרו. התוצאות מראות ששיטת הpruning נותנת את דיוק הבדיקה הגבוה ביותר ב-77.1%, ואחריה שיטות Distillation ו-Quantization ב-75.99% ו-74% בהתאמה. איטרציה נוספת של אימון נמנעה כדי למנוע התאמת יתר. לסיכום, נראה ששיטת הpruning היא היעילה ביותר.

7. מסקנות והצעות להמשך:

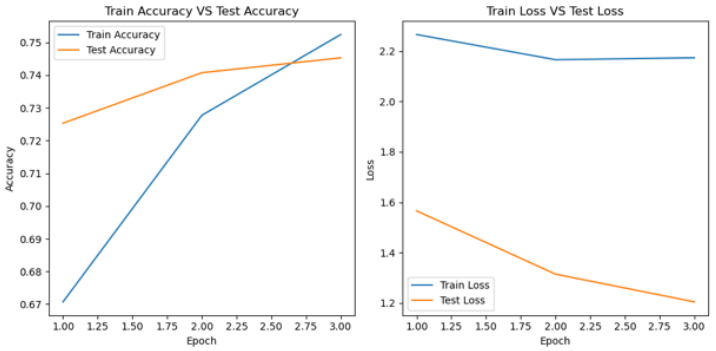
בחלק האחרון בפרויקט זה חקרנו את ההשפעה של שלוש טכניקות כיווץ שונות על מודלים מבוססי טרנספורמרים. הטכניקות כוללות Pruning, Distilation ו-Quantization, הן יושמו במודל BERT. התוצאות הראו שכל השיטות הללו יכולות להקטין את גודל המודל תוך שמירה על רמות ביצועים טובות. ביניהם, Pruning סיפק את דיוק הבדיקה הגבוה ביותר אשר יותר טובה מאימון של המודל עצמו ללא הכיווץ, עם זאת, חשוב לזכור כי בחירת טכניקת הכיווץ צריכה להיעשות בהתבסס על הצרכים הספציפיים של המשימה והמשאבים העומדים לרשותנו. למרות הביצועים המעולים של הפרונינג, עשויים להיות תרחישים שבהם שאר שתי השיטות שבחרנו יכולים להיות מתאימים יותר. בעתיד, ניתן יהיה לקבל הבנה רחבה יותר של טכניקות הכיווץ הללו על ידי בדיקתן בדגמים שונים כגון GPT או XLM. עם התחום המתקדם במהירות של NLP, מחקר השוואתי על פני מגוון רחב יותר של מודלים יכול להיות תובנות. הפרויקט התמקד בשיטות כיווץ ספציפיות, אבל יש עוד טכניקות שכדאי לחקור. לדוגמה, נוכל לבחון לחקור שיטות של חלוקת משקל (weight sharing) או

low rank factorization. שיפור ההבנה שלנו בכיווץ מודלים יכול להוביל ליצירת מודלים יעילים עם ביצועים גבוהים שהם גם ידידותיים למשאבים. זה חשוב להנגשת משימות NLP מתחומים דומים למשימה זאת שביצענו.

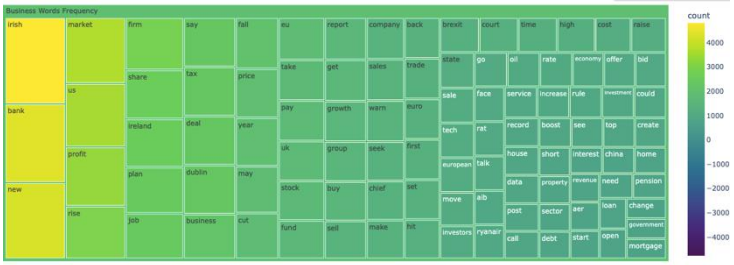
8. נספח:



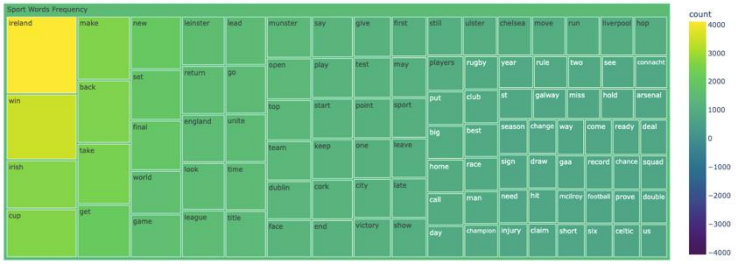
גרף 11: גרפי האימון והמבחן של רמת הדיוק והשגיאה של מודל DistilBERT



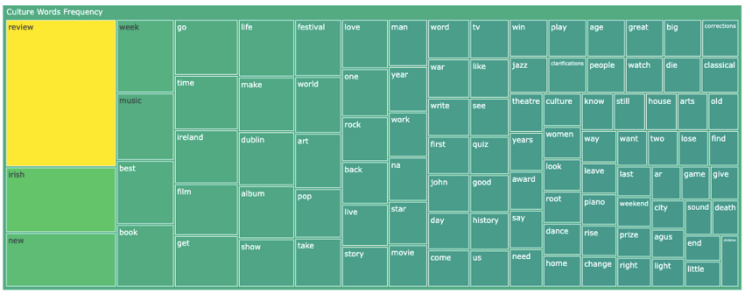
גרף 12: גרפי האימון והמבחן של רמת הדיוק והשגיאה של מודל ROBERTA



גרף 13: מפת עץ- שכיחות מילים עבור קטגוריית עסקים



גרף 14: מפת עץ- שכיחות מילים עבור קטגוריית ספורט



גרף 15: מפת עץ- שכיחות מילים עבור קטגוריית תרבות

גרף 16: מפת עץ- שכיחות מילים עבור קטגוריית דעות

גרף 17: מפת עץ- שכיחות מילים עבור קטגוריית חיים