פרויקט סיכום AI

רועי שתיוי, ורדית ארקש, נגה שטרן ודנית ישעיהו

תקציר

זיהוי מחברים הוא נושא חשוב בתחום של זיהוי שפה טבעית, הוא מאפשר לנו לזהות את הכותב בעל ההסתברות הגבוהה ביותר להיות הבעלים של מאמר, כתבת חדשות, מסרון, ספר ועוד; כאשר כלי אוטומטי מסוג זה יכול להיות מיושם לשלל מטרות. בפרויקט זה נתמקד בזיהוי של כותבי ספרים תוך שימוש במאגר טקסטים בעזרת שלושה אלגוריתמי למידת מכונה קלאסיים, כאשר שניים מהם נלמדו בקורס הנוכחי.

הניסויים שערכנו העלו תוצאות טובות למדי של כ־ 87% דיוק עבור סיווג נכון לכותב המקורי בעזרת שימוש באלגוריתם Random Forest אשר הופעל על 29 ספרים שונים שחולקו ל־ 1784 פרקים, כאשר כל קטע טקסט יוצג על ידי וקטור של פיצ'רים רלוונטים שהתחלקו לשלושה מחלקות שונות.

1 הקדמה

לסופרים ומחברים נחשבים קיים לרוב סגנון כתיבה אשר ייחודי לעבודותיהם, סגנון זה במידת מה לא יהיה רלוונטי לנושא הכתוב בו אך למרות זאת יוכל להיות מזוהה על ידי קורא אנושי המכיר את כתביו. תהליך זה של זיהוי כותב בעזרת קטע טקסטואלי (משפט, פסקה או פרק קצר) מתוך קבוצת מועמדים נקרא Authorship Identification. חשיבותו של התהליך נובעת משימושיו הנרחבים: החל מזיהוי מחבר אנונימי, גילוי גניבה ספרותית, מציאת סופר צללים וכלה בדרך חדשה להמלצות על סופר בעל דמיון גבוה לדפוס הכתיבה המועדף לקורא.

בעבודה זו בחרנו להתמקד במאגר המידע של פרויקט גוטנברג המכיל ארכיון דיגיטלי של ספרים מוכרים החופשיים לשימוש. מכל ספר שאספנו ייצרנו וקטור באורך של 55 איברים כאשר כל תא בו מייצג תוצאה נומרית המתאימה לפיצ'ר ספציפי. את 1784 הוקטורים חילקנו ביחס של 80% לקבוצת האימון ואת השאר לקבוצת הבדיקה, ועל מאגר הנתונים זה הפעלנו שלושה אלגוריתמים: Decision Tree, Random Forest ו־ Nearest Neighbor.

2 קורפוס

על מנת לאמן את המודל אנו זקוקים לקורפוס מכל סוג כמו: כתבות, מאמרים, הודעות או מיילים, אשר מתויגים עם כותבים ואנו בחרנו בספרים המתויגים עם סופריהם. בפרויקט בחרנו לזהות את אופי הכתיבה של עשרת הסופרים הבאים: ג'יין אוסטין, הרמן מלוויל, ז'ול ורן, שרלוט ברונטה, לואיס קרול, ברם סטוקר, לאו טולסטוי, רוברט סטיבנסון, פיודור דוסטויבסקי וצ'ארלס דיקנס. עבור כל אחד מהסופרים הכנסנו 3 ־ 4 ספרים(ראה נספח). כמו כן, על מנת לקבל סטטיסטיקות רבות יותר על הטקסטים, המסייעות ללמידה טובה יותר, חילקנו את כל הספרים לפרקיהם. לכל פרק התייחסנו כטקסט נפרד של אותו הסופר וכך קיבלנו שלג'יין אוסטין 129 טקסטים, להרמן מלוויל 266 טקסטים, לז'ול ורן 101 טקסטים, לשרלוט ברונטה 100 טקסטים, ללואיס קרול 71 טקסטים, לברם סטוקר 61 טקסטים, ללאו טולסטוי 616 טקסטים, לרוברט סטיבנסון 92 טקסטים, לפיודור דוסטויבסקי 158 טקסטים ולצ'ארלס דיקנס 177 טקסטים. נוסיף כי כל הספרים נלקחו מפרויקט גוטנברג, פרויקט המנגיש לציבור ספרים אשר אין להם זכויות יוצרים יותר. בשל מקורם, רוב הספרים נכתבו במאה ה־19 וחלקם אף במאה ה־18. הטקסטים המוזכרים נתונים בפרויקט בתיקיית "corpus" כאשר תחת תיקייה זו יש תיקייה לכל סופר ותחתיו חלוקה לפי כל ספר.

3 פיצ'רים

תהליך בחירת הפיצ'רים המתאימים לפתרון הבעיה הוא חלק חשוב אשר מכריע על איכות המסווג המתקבל באופן ישיר. מטרתנו היא למצוא פיצ'רים שיהיו ככל שיותר אינפורמטיביים, מבדלים (יוצרים הפרדה ניכרת בסט הדוגמאות) ובלתי תלויים אחד בשני. בעבודה זו כל הפיצ'רים שמימשנו הם פונקציות המקבלות טקסט ונותנות לו ערך מספרי, כך בהרצת התוכנית יקבל כל קטע וקטור ערכים המאפיין אותו באופן ייחודי. את הפיצ'רים חילקנו לשלוש רמות שונות: רמת התו, רמת המילה ורמת התחביר.

הפיצ'רים שבחרנו הם יחסיים, כלומר כל אחד מהם מחושב ביחס לאורך המשפט או הטקסט. החשיבות ליצירת הפיצ'רים באופן הזה נועד כדי לא להטות את התוצאות על פי אורך הפרקים של כל כותב (אשר אינם מיוצגים באופן אחיד בקורפוס שלנו) או לתת עדיפות לסופרים בעלי יותר טקסטים מאחרים. בנוסף לכך.

ליצירת הפיצ'רים התחביריים השתמשנו בשני כלים: nltk.pos_tag ו־StanfordCoreNLP. כאשר הספרייה הראשונה סייעה לנו בתיוג חלקי הדיבר והשנייה לבניית עצים תחביריים לכל משפט.

בחירת הפיצ'רים התבצעה ברובה בעזרת שימוש במאמרים שונים שהתפרסמו בנושא זה (לאו דווקא עבור טקסטים ספרותיים) .,

3.1 רמת התו

הפיצ'רים ברמת התווים הם פיצ'רים שבאופן טבעי היינו חושבים עליהם כפחות מעניינים וטריוואלים לבדיקה, אך חלקם מעלים תובנות משמעותיות לגבי סגנון דפוס הכתיבה של המחבר. כל פיצ'ר במחלקה זאת מחשב את הסתברות של מאורע תו בטקסט להתקיים; לדוגמא, פיצ'ר F_3 יחשב את ההסתברות של תו הנבחר באופן אקראי להיות ספרה.

שם הפיצ'ר	תיאור
F_1	יחס התווים המיוחדים[1] לכלל התווים בטקסט
F_2	יחס האותיות לכלל התווים בטקסט
F_3	יחס הספרות לכלל התווים בטקסט
F_4	היחס בין הרווחים לאורך הטקסט
F_5	היחס בין הרווחים לכלל סוגי התווים הריקים (דוגמאת טאבים וירידת שורה)
F_6	יחס בין טאבים לשאר סוגי הרווחים הריקים

3.2 רמת המילה

ברמה זו רצינו לבחון את מאפייני המילים בקטע כיחידות, כאשר את החלוקה למילים ביצענו באמצעות הפרדה לפי תווים ריקים. אלמנטים כמו אוצר המילים של הכותב יכולים להתבטא לדוגמא במספר המילים הייחודיות בהם הוא משתמש, לכן יצרנו את פיצ'ר F_9 שסופר את כמות המילים שמופעיהן הוא יחיד אל מול כמות המילים הכוללת בקטע, ואת פיצ'רים F_7, F_8 הבוחנים את יחס השימוש במילים קצרות\ארוכות.

סיווג מילים נוסף בבלשנות קיים בחלוקה בן מילים פונקציונליות לבן מילות תוכן. למילה פונקציונלית קיימת לרוב משמעות לקסיקלית נמוכה והיא מייצגת את היחס התחבירי שיש למילות התוכן אחת עם השניה במשפט (ניתן לחשוב עליהן כמעין 'דבק' המחבר בן מילים), ומכאן שהן מהוות אלמנט חשוב במבנה המשפט. לכן בחרנו בפיצ'ר F_{13} הסופר את כמות המילים הפונקציונליות בקטע (F_{13} הוא הפיצ'ר הלקסיקלי היחיד ברמת המילה).

שם הפיצ'ר	תיאור
F_7	יחס המילים הקצרות[2] לכלל המילים
F_8	יחס המילים הארוכות[3] לכלל המילים
F_9	יחס המילים הייחודיות לכלל המילים
F_{10}	אורך מילה ממוצע
F_{11}	ממוצע מספר מילים במשפט
F_{12}	ממוצע מספר תווים במשפט
F_{13}	מספר מילים פונקציונליות במשפט

3.3 רמת התחביר

חלק חשוב מסגנון הכתיבה של הסופר מתבטא גם באופן פחות גלוי או מפורש בטקסט כפי כפי שראינו ברמת התו והמילה (חישובים פשוטים יותר, למעט F_{13}). כדי לנתח את דפוס הכתיבה התחבירי אנו נדרשים ראשית להבין את כללי השפה הקובעים איזה רצף מילים בה מהווה משפט תקין. כללים תחבירים אלו הם כללים היררכיים ולכן לא די לבחון את סדר מופעי המילים לבדם באופן לינארי, אלא לתאר באופן מדויק ושיטתי את המבנה ההיררכי של המשפט בשפה. בעזרת כלים מתחום הבלשנות החישובית נוכל למצוא את המשפטים בקטע הטקסט ולנתח את מבנם (נציין שחלק מהפיצרים ברמת המילה נעזרים בתוצאות כלים תחביריים אלו, כדוגמאת $F_{11} - F_{13}$).

בתחביר הגנרטיבי כללי שפה בסיסיים הם כללי גזירה, את מבנה המשפט המתקבל מכללים מסוג זה מקובל לתאר באופן גרפי בעזרת תרשים היררכי הנקרא "עץ תחבירי". היחסים המבניים המרכזיים שמתקיימים בכל עץ תחבירי, מהווים חלק מהידע התחבירי הלא מודע של הדוברים, ידע שמאפשר את זיהוי היחידות התחביריות במשפט. בפרויקט זה השתמשנו בכלים המייצרים עצים תחבירים לטקסטים ויצרנו את פיצ'ר F_{55} המייצג את עומק הממוצע של העץ התחבירי לכלל המשפטים.

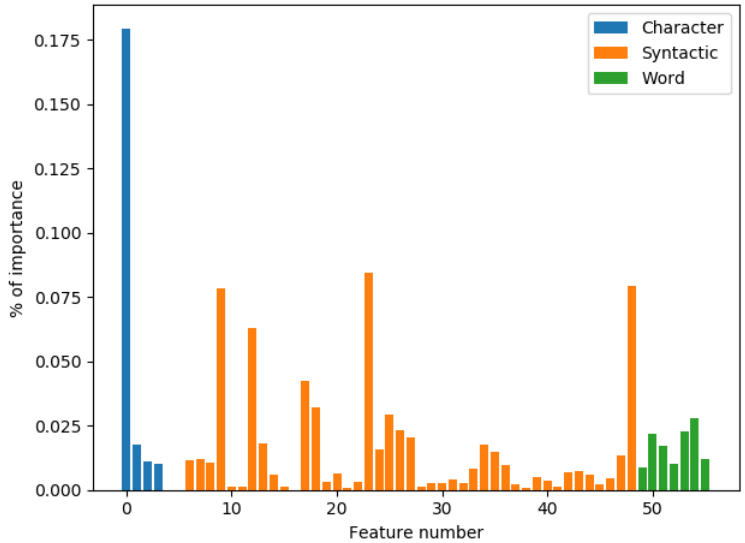
בנוסף, החלוקה לחלקי דיבר מאפשרת לסווג כל מילה בלקסיקון השפה לפי מאפיינה המורפולוגיים, מילים שיסווגו לאותו חלק דיבר לרוב ימלאו תפקידים דומים במבנה הדקדוקי של המשפט. חלקי דיבר בשימוש נרחב הם שמות עצם, שם תואר, תואר השם, כינוי גוף, מילות קישור, מילות קריאה ועוד. בפרויקט זה השתמשנו ב־ 39 חלקי דיבר הנמצאים בשימוש בשפה האנגלית.

שם הפיצ'ר	תיאור הפיצ'ר
F_{14}	היחס בין אותיות
$F_9 - F_{52}$	בהינתן תג מחזיר את היחס של התג בטקסט לעומת אורך הטקסט
F_{53}	יחס של פעלים בזמן עבר לכלל המילים
F_{54}	יחס של פעלים בזמן הווה לכלל המילים
F_{55}	הממוצע של עומק עץ תחבירי לכל המשפטים

3.4 חשיבות הפיצ'רים

את הדירוג לכלל הפיצ'רים שהשתמשנו בהם ביצענו בעזרת יחס ה־Information Gain הנמצא בשימוש בעצי החלטה ויערות אקראיים, אשר לרוב מוצא את הפיצ'רים הרלוונטלים ביותר וממקמם באזור שורש העץ. לכן, כדי לדעת איזה פיצ'ר גרם לפיצול הטוב ביותר, וכך השפיע יותר בתרומתו ללמידה נרצה למצוא את האחד שה־IG שלו היה המקסימלי.

בגרף הבא ניתן לראות את דירוג הפיצ'רים על פי תרומתם תוך חלוקה לפי רמותיהם השונות



4 מימוש

את מימוש התוכנית ביצענו תוך ארבעה שלבים:

1. **מאגרי המידע**

   (א) איסוף מאגרי מידע רלוונטים־ מצאנו טקסטים של סופרים מוכרים בעלי לפחות שלושה ספרים לכל אחד.

   (ב) הכנת מאגר המידע־ חילקנו את הספרים לפי פרקיהם על מנת לקבל חלוקה אחידה יותר וסט דוגמאות רחב יותר שיעזור בתהליך הלמידה.

2. **חילוץ הפיצ'רים** לכל כל אחד מהפרקים שכלל את מציאת ערכים של 55 הפיצ'רים שהוצגו בפרק 2.

3. **למידה ואימון** בעזרת הפיצ'רים הנ"ל למציאת המסווג בעזרת שלושה אלגוריתמי למידה שונים כאשר המידע חולק לסט אימון וסט בדיקה.

4. **שימוש** במודל מאומן בעל אחוזי ההצלחה הגבוהים ביותר שבעזרתו נוכל לאתר את זהות הכותב לפי דוגמא טקסטואלית חדשה לגמרי.

עבור הרצת הסיפריה הפעלנו את השלבים $1-2$ מבעוד מועד ושמרנו את וקטורי הפיצ'רים של כל הספרים בקובץ ייעודי כדי להקל על זמן הריצה בהצגת התוכנית. ניתן לחשב את הפיצ'רים בזמן הריצה בעזרת שימוש בפרמטר הדגל calc_data- (יוסבר בפירוט בפרק 7)

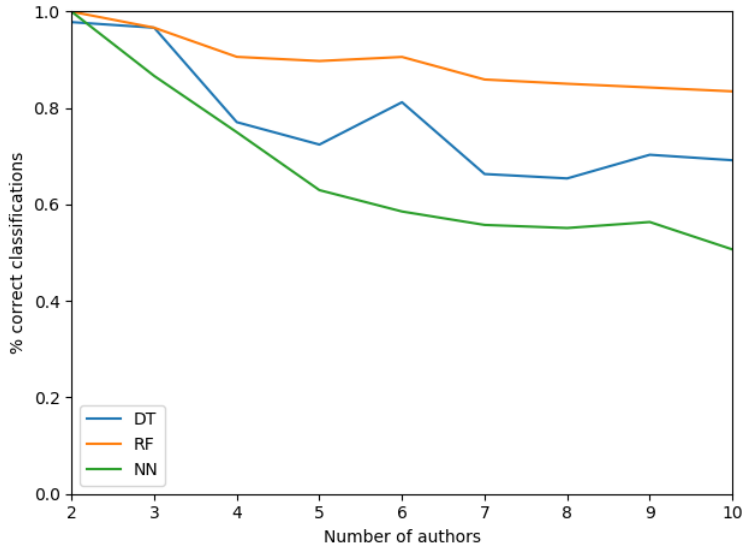
מימוש הפרויקט כלל שימוש במגוון סיפריות שעזרו לנו לייעל את זמן הריצה ודיוקו של המודל, העיקריות מבניהן היו:

- sklearn־ מימוש של שלושת האלגוריתמים Random Forest, Decision Tree ו־ Nearest Neighbors
- nltk־ שימוש במגוון אלגוריתמים וכלים של עיבוד שפה טבעית למציאת משפטים, תגים, מילים פונקציונליות ועוד.
- StanfordCoreNLP־ ליצירת העצים התחביריים של כל פרקי הספרים.
- graphviz־ ליצירת וויזואליזציה של עץ ההחלטה.

ב־ scipy, matplotlib ו־numpy השתמשנו לעבודה עם מסדי נתונים ספרתיים ויצירת גרפים.
הוראות לשימוש והרצת התוכנית נמצאות באופן מפורט בקובץ README.md

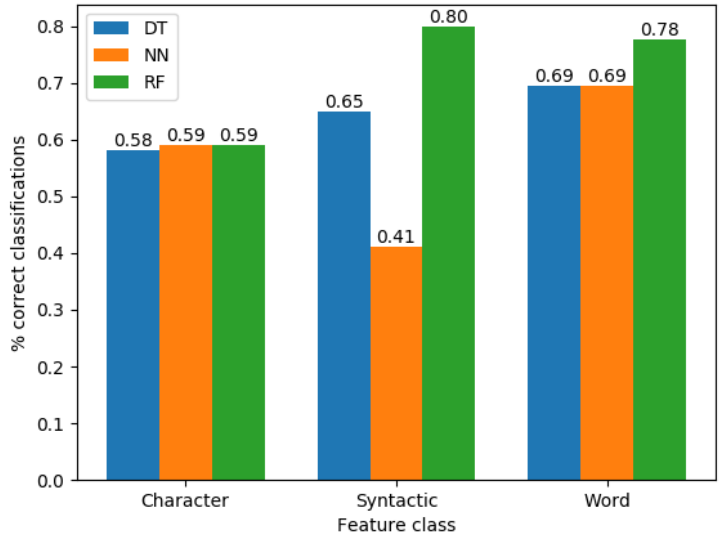
5 תוצאות

במהלך הפרויקט הבנו כי קיימים פרמטרים רבים שעשויים להשפיע על התוצאות ויש לקחת אותם בחשבון, ולכן החלטנו לבחון אותם. ראשית, כדאי לשים לב לעובדה שככל שיש יותר סופרים ברשימת המועמדים לסייוג כך נדרשת הבחנה דקה יותר ביניהם ובין סגנונות הכתיבה שלהם על מנת לשייך טקסט מסויים לסופר אחד ולא לאחר. כדי לבדוק את ההשעפה של הגדלת כמות הסופרים ביצענו הרצה של התוכנית בין שלושת האלגוריתמים השונים עבור מספר סופרים משתנה.

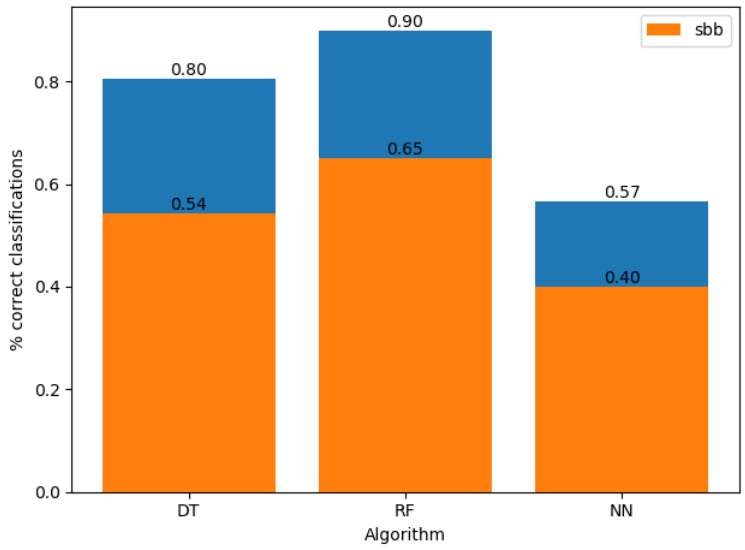


ניתן לשים לב שבעבור האלגוריתם Random Forest קיימת התכנסות סביב ה־ 87% הצלחה גם עבור 10סופרים.

הבעיה העיקרית שנאלצנו להתמודד איתה בפרויקט היא מציאת הפיצ'רים האולטימטיביים לאפיון סגנון כתיבת כל סופר. חילקנו את הפיצ'רים ל־3 רמות: רמת התו, רמת המילה ורמת התחביר ובחנו את תוצאות האלגוריתמים עבור כל אחד מסוגי הרמות. להלן התוצאות:



כאמור, התייחסנו לכל פרק בספר כטקסט נפרד. מכיוון שהחלוקה לטקסטים שמשמשים לאימון וטקסטים שמשמשים למבחן נעשית באופן רנדומלי, ייתכן שהמערכת תתאמן על פרקים מסוימים מתוך ספר, ולאחר מכן תבחן על פרקים אחרים מתוך אותו ספר (למשל, ייתכן שהמערכת תתאמן על פרקים 1־5 מתוך ספר מסוים, ותבחן על פרקים 6־10 מתוך אותו ספר). כדי לוודא שהמערכת אכן מזהה את הסופרים עצמם, ולא מזהה את הספר בזכות העובדה שהיא אומנה על פרקים אחרים מתוכו, הרצנו את האלגוריתמים עם חלוקה אחרת לאימון ומבחן: מכל אחד מהסופרים נבחר באופן רנדומלי ספר אחד ששימש למבחן, ושאר הספרים שימשו לאימון. המערכת עדיין מתייחסת לכל פרק כטקסט נפרד, אבל ספר שמופיע במבחן בהכרח לא הופיע באימון. להלן התוצאות:



6 הגבלות

כאשר אנו מביטים בתוצאות עבודתנו עלינו לקחת בחשבון מספר גורמים העשויים להשפיע לטובה או לרעה על התוצאות. קורפוס העבודה באופן יחסי אינו גדול ומתבסס על מספר ספרים קטן אשר עלול לא לייצג כהלכה את הסופרים. כלומר, ייתכן שלא ייצגנו את מגוון סגנונות הכתיבה וצורות הכתיבה השונות העשויות להשתנות בין ספריו של אותו הסופר. בנוסף לכך, השתמשנו בכלים nltk.pos_tag ו־StanfordCoreNLP כלים ששימשו אותנו לתיוג הטקסט אשר באמצעותו יצרנו פיצ'רים נוספים. כלים אלו שוגים בתיוגם ב־10% מהמקרים טעויות אשר עשויות ליצור פיצ'רים שלא מייצגים כראוי את הטקסט. כמו כן, על מנת ליצור פיצ'רים חילקנו את הטקסט למשפטים, חלוקה אשר נעשתה באמצעות סיום משפט, אך פעמים רבות לא ניתן היה לזהות כראוי את סיום המשפט. למשל במשפטים בהם הופיע "Mr." לא יכולנו לזהות שהנקודה לא מסיימת משפט ולכן סטטיסטית תפסנו אותה כמשפט.

7 מסקנות ועבודה עתידית

במבט על התוצאות השונות הבחנו כי קיימים פרמטרים רבים המשפיעים על אחוזי ההצלחה של המודל שלנו. כל פרמטר כזה עשוי להציג שינויים דרסטיים על התוצאות. ראינו כי כקבוצה הפיצ'רים הסינטקטיים הם המשפיעים ביותר אך עם זאת בעת השילוב שלהם עם האלגוריתם הלומד NN הגענו לתוצאות הגרועות ביותר. לעומת זאת, הצבענו על רמת השפעה הכי נמוכה אצל הפיצ'רים ברמת המילים אך הם הראו את היציבות הגבוהה ביותר בין אלגוריתמי הלמידה השונים. ומכאן, שהשילוב בין פיצ'רים אלו ואחרים עשוי להביא לשיפור בתוצאות. בפרויקט שלנו נגענו ב־3 רבדים בודדים הנמצאים בשפה ולכן בתור צעד להמשך אנו מציעים להוסיף פיצ'רים המשלבים עוד רבדים כמו סמנטיקה או פרגמטיקה המתארים את סגנון הכתיבה.

במהלך אימון המודל תהינו לגבי השפעת אופן חלוקת המידע עבור קבוצת דוגמאות למידה, האם פרק חדש מספר שנלמד יזוהה יותר טוב מפרק של ספר שאינו נלמד כלל?. גילינו כי כמצופה כי אכן פרקים מתוך ספרים שנלמדו זוהו יותר טוב על ידי המודל, אך עם זאת גם בעת בחינת המודל על ספרים שלא נלמדו קיבלנו תוצאות סבירות (סביב ה־65% הצלחה). בעבודתנו השתמשנו במספר מועט של ספרים, בשל כך כל ספר היה מאוד משמעותי עבור הלמידה והוסיף לה עוד פן באופי כתיבתו של הסופר. לכן, כאשר הורדנו ספר שלם (מתוך $3-4$ ספרים) מלמידת המודל גרמנו לו ללמוד פחות, דבר שהיה עלול גם להסביר את תוצאותינו. לכן אנו מאמינים שמאגר מידע הכולל יותר ספרים ביחס לכל סופר יעזור לייצג את אופי כתיבת הסופר בצורה טובה ומדויקת יותר וכך התוצאות לפי החלוקה הנ"ל יהיו קרובות לתוצאות החלוקה הרגילה. מלבד זאת, כצפוי גורם מרכזי להצלחת המודל הוא מספר הסופרים אותו הוא צריך ללמוד ולזהות, ככל שמספר הסופרים גבוהה יותר הצלחת המודל קטנה, וזאת כי נדרש מהמודל לתאר כל סופר באופן מדויק יותר. בעתיד היינו רוצים להגדיל את עבודתינו ולבחון את הצלחת המודל עם מספר רב יותר של סופרים.

במהלך האימון, המודל למד ספרים באנגלית אשר חלקם לא נכתבו במקור באנגלית אלא תורגמו אליה. שפת המקור לא השפיעה באופן משמעותי על למידת המודל וגם עבור סופרים שלא כתבו באנגלית בוצע סיווג נכון אל הסופר המקורי, נתון המעיד על כך שלמרות שחלק ניכר מהפיצ'רים שהשתמשנו בהם היה תלוי שפה (הפיצ'רים התחביריים התמקדו בתגים שונים ומילים פונקציונליות בשפה האנגלית). בשל עובדה זו, היינו רוצים להרחיב את עבודתנו בעזרת מציאת פיצ'רים אוניברסליים יותר ולבחון את הצלחת המודל על טקסטים משפות שונות. מלבד הרחבת המודל לשפות ניתן להרחיב אותו לסוגים שונים של טקסטים כמו: פוסטים של פייסבוק, מיילים או שיחות טלגרם.

8 נספח

רשימת הסופרים וסיפריהם:

Jane Austen: Emma, Persuation, Sense, Sensibility

Herman Melville: Moby Dick, Omoo, A Romance Of The South Seas

Jule Verne: Around the World in 80 days, The Secret of The Island, Five Weeks in a Balloon

Lewis Carroll: Alice's Adventures in Wonderland, A Tangled Tale, Sylvie and Bruno, Through the Looking-Glass

Charlotte Bronte: Jane Eyre, The professor, Villette

Bram Stoker: Dracula, The Lair of the White Worm, The Jewel of Seven Stars

Leo Tolstoy: War and Peace, Anna Karenina, The Kingdom of God Is Within You

Robert Stevenson: Kidnapped, The Black Arrow, Catriona

Fyodor Dostoevsky: The Brothers, Crime and Punishment, The Possessed

Charles Dickens: Oliver Twist, Great Expectations, David Copperfield

רשימת מקורות

[1]

Authorship Identification of Research Papers, Simen Skoglund 2015

[2] דגדג