

זיהוי מחברים

רועי שתיו, ורדית ארקש, נגה שטרן ודנית ישעיהו

תקציר

זיהוי מחברים הוא נושא חשוב בתחום של עיבוד שפה טבעית, הוא מאפשר לנו לזהות את הכותב בעל ההסתברות הגבוהה ביותר להיות הכותב 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-90% דיוק עבור סיווג נכון לכותב המקורי בעזרת שימוש באלגוריתם Random Forest אשר הופעל על 31 ספרים שונים שחולקו ל-1784 פרקים, כאשר כל קטע טקסט יוצג על ידי וקטור של פיצ'רים רלוונטים שהתחלקו לשלוש 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 משימושי 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 [5] 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 הוקטורים חילקנו ביחס של 80% לקבוצת האימון ואת השאר לקבוצת הבדיקה, ועל מאגר הנתונים זה הפעלנו שלושה אלגוריתמים: Nearest Neighbors, Decision Tree, Random Forest.

2 קורפוס

על מנת לאמן את המודל אנו זקוקים לקורפוס של: כתבות, מאמרים, הודעות או מיילים (למעשה, כל מאגר טקסטואלי בעל מבנה משותף יכול להתאים), אשר מתויגים על פי מחבריהם. אנו השתמשנו במאגרי מידע של ספרים והוספנו להם את התיוג המתאים על פי הסופר הרלוונטי. בפרויקט בחרנו לזהות את אופי הכתיבה של עשרת הסופרים הבאים: ג'יין אוסטין, הרמן מלוויל, ז'ול ורן, שרלוט ברונטה, לואיס קרול, ברם סטוקר, לב טולסטוי, רוברט לואיס סטיבנסון, פיודור דוסטויבסקי וצ'ארלס דיקנס. עבור כל אחד מהסופרים העלנו בין שלושה לארבעה ספרים (ראה נספח). כמו כן, על מנת לקבל סטטיסטיקות רבות יותר על הטקסטים, המסייעות ללמידה טובה יותר, חילקנו את כל הספרים לפרקים. 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, ללב טולסטוי 616 טקסטים, לרוברט לואיס סטיבנסון 92 טקסטים, לפיודור דוסטויבסקי 158 טקסטים ולצ'ארלס דיקנס 177 טקסטים, שהם בסך הכל 1784 פרקים.

כל הספרים נלקחו מפרויקט גוטנברג, פרויקט המנגיש לציבור ספרים אשר פגו זכויות היוצרים המגנים עליהם. מסיבה זאת, רוב הספרים שאספנו נכתבו במאה ה-19 וחלקם אף במאה ה-18. הטקסטים המוזכרים נתונים בפרויקט בתיקיית `\corpus\data` כאשר תיקייה זו מחולקת על פי הסופר ותחתיה על פי ספריו.

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 כמה 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. יצירת הפיצ'רים באופן הזה נועדה 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.

ליצירת הפיצ'רים התחביריים השתמשנו בשני כלים: `nlk.pos_tag` ו-`StanfordCoreNLP` [4]. כאשר הספרייה הראשונה סייעה לנו בתיג חלקי הדיבר והשנייה לבניית עצים תחביריים לכל משפט.

בחירת הפיצ'רים התבצעה ברובה בעזרת שימוש במאמרים שונים שהתפרסמו בנושא זה [1, 2, 3] (לא דווקא עבור טקסטים ספרותיים) אך גם על ידי פיצ'רים שחשבנו עליהם במהלך כתיבת הפרויקט.

3.1 רמת התו

הפיצ'רים ברמת התווים הם פיצ'רים שבאופן אינטואיטיבי היינו חושבים עליהם כפחות מעניינים ו\או טריוואליים לבדיקה, אך חלקם מעלים תובנות משמעותיות לגבי סגנון דפוס הכתיבה של המחבר. כל פיצ'ר במחלקה זאת מחשב את ההסתברות של מאורע, המאופיין כתו בטקסט, להתקיים כאשר ההסתכלות על הטקסט היא כעל רצף של התווים השונים המרכיבים אותו; לדוגמא, פיצ'ר F_1 יחשב את ההסתברות של תו הנבחר באופן אקראי להיות תו מיוחד¹.

שם הפיצ'ר	תיאור
F_1	יחס התווים המיוחדים 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 בהן הוא משתמש, לכן יצרנו את פיצ'ר F_{51} שסופר את מספר המילים שמופיען הוא יחיד אל מול כמות המילים הכוללת בקטע, ואת פיצ'רים F_{49}, F_{50} הבוחנים את יחס השימוש במילים קצרות² ארוכות³.

סיווג מילים נוסף בבלשנות קיים בחלוקה בין מילים פונקציונליות לבין 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_{55} הסופר את כמות המילים הפונקציונליות בקטע (F_{55} הוא הפיצ'ר הלקסיקלי היחיד ברמת המילה).

¹תו הנמצא בקבוצה: /, \, {, }, [,], ^, &, -, =, +, <, >, —, [,], { , }, \ ,
²מילה שאורכה קטן או שווה ל-3
³מילה שאורכה גדול מ-9

שם הפיצ'ר	תיאור
F_{49}	יחס המילים הקצרות לכלל המילים
F_{50}	יחס המילים הארוכות לכלל המילים
F_{51}	יחס המילים הייחודיות לכלל המילים
F_{52}	אורך מילה ממוצע
F_{53}	ממוצע מספר מילים במשפט
F_{54}	ממוצע מספר תווים במשפט
F_{55}	מספר מילים פונקציונליות במשפט

3.3 רמת התחביר

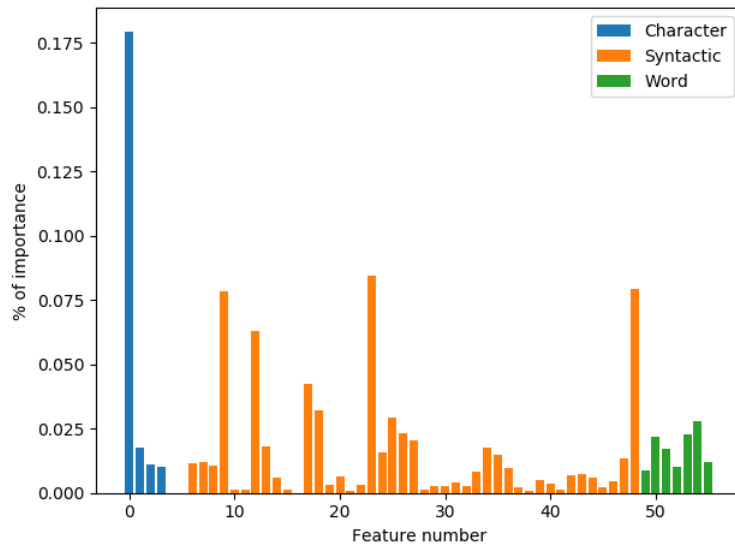
חלק חשוב מסגנון הכתיבה של הסופר מתבטא גם באופן פחות גלוי או מפורש בטקסט כפי שראינו ברמת התו והמילה (חישובים פשוטים יותר, למעט F_{55}). כדי לנתח את דפוס הכתיבה התחבירי אנו נדרשים ראשית להבין את כללי השפה הקובעים איזה רצף מילים בה מהווה משפט תקין. כללים תחבירים אלו הם כללים היררכיים ולכן לא די לבחון את סדר מופעי המילים לבדם באופן לינארי, אלא לתאר באופן מדויק ושיטתי את המבנה ההיררכי של המשפט בשפה. בעזרת כלים מתחום הבלשנות החשובית נוכל למצוא את המשפטים בקטע הטקסט ולנתח את מבנם (נציין שחלק מהפיצ'רים ברמת המילה נעזרים בתוצאות כלים תחביריים אלו, כדוגמת $F_{53} - F_{55}$).

בתחביר הגנרטיבי כללי שפה בסיסיים הם כללי גזירה, ואת 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 [6]. בפרויקט זה השתמשנו בכלים המייצרים עצים תחבירים לטקסטים ויצרנו את פיצ'ר F_{48} 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 [7].

שם הפיצ'ר	תיאור הפיצ'ר
F_7	היחס בין אותיות
$F_8 - F_{45}$	בהינתן תג מחזיר את היחס של התג בטקסט לעומת אורך הטקסט
F_{46}	יחס של פעלים בזמן עבר לכלל המילים
F_{47}	יחס של פעלים בזמן הווה לכלל המילים
F_{48}	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 שבו משתמשים בעצי החלטה ויערות אקראיים, ואשר לרוב מאפשר למצוא את הפיצ'רים הרלוונטים 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 רלוונטיים - מצאנו טקסטים של סופרים מוכרים בעלי לפחות שלושה ספרים לכל אחד.

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

2. חילוף הפיצ'רים לכל אחד מהפרקים, כלומר 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.

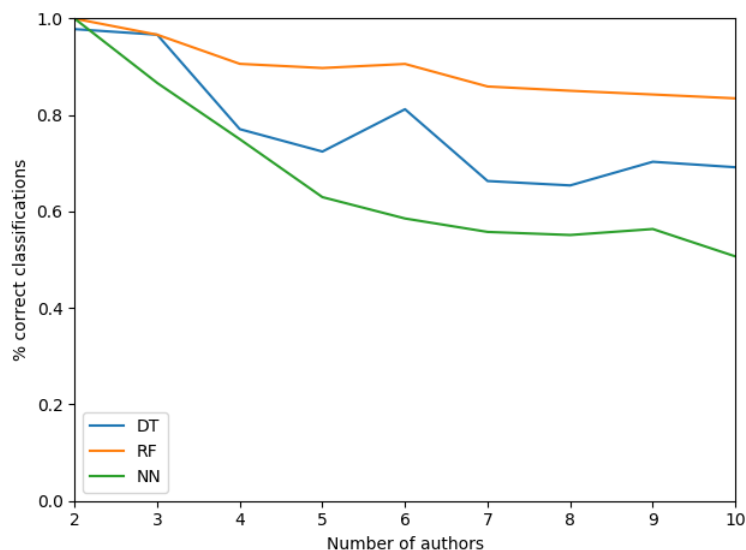
4. שימוש במודל מאומן בעל אחוזי ההצלחה הגבוהים ביותר, שבעזרתו נוכל לאתר את זהות הכותב לפי דוגמא טקסטואלית חדשה לגמרי.

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 גם לא להעזר בקובץ ולחשב את הפיצ'רים בזמן אמת בעזרת שימוש בפרמטר הדגל `-calc_data`.

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

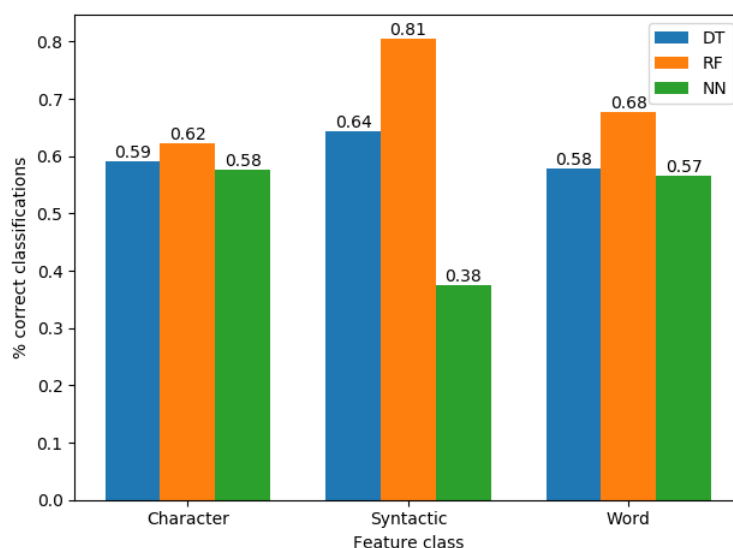
- ספריית `sklearn` - מימוש של שלושת האלגוריתמים `Random Forest`, `Decision Tree` ו-`Nearest Neighbors`.
 - ספריית `nltk` - שימוש במגוון אלגוריתמים וכלים של עיבוד שפה טבעית למציאת משפטים, תגים, מילים פונקציונליות ועוד.
 - ספריית `StanfordCoreNLP` - ליצירת העצים התחביריים של כל פרקי הספרים.
 - ספריית `graphviz` - ליצירת וויזואליזציה של עץ ההחלטה.
- ב-`numpy`, `scipy`, `matplotlib` השתמשנו לעבודה עם מסדי נתונים ספרייתיים ויצירת גרפים. הוראות לשימוש והרצת התוכנית נמצאות באופן מפורט בקובץ `README.md`.

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 לסיווג כך נדרשת הבחנה דקה יותר ביניהם ובין סגנונות הכתיבה שלהם על מנת לשייך טקסט מסוים לסופר אחד ולא לאחר. כדי לבדוק את ההשפעה של הגדלת כמות הסופרים הרצנו את התוכנית עם שלושת האלגוריתמים השונים עבור מספר סופרים משתנה.



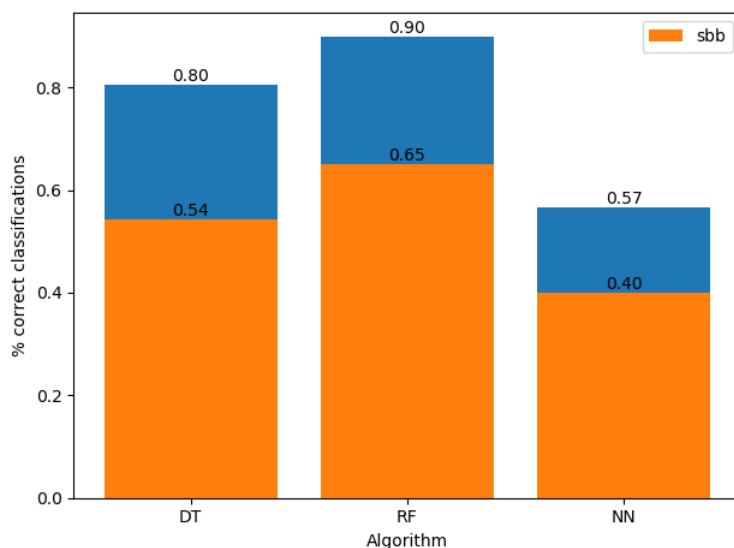
ניתן לשים לב שבעבור האלגוריתם Random Forest קיימת התכנסות סביב ה- 90% הצלחה גם עבור 10 סופרים.

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, כאשר החלוקה בצורה שתוארה מופיעה בצבע כתום ובכחול מופיעות התוצאות כאשר החלוקה נעשית באופן רנדומלי לגמרי, בלי שספרים מסוימים ייכנסו רק לאימון או רק למבחן:



כדי לקבל הבנה טובה יותר על התנהגות ודרכי סיווג של המודל בהתייחס לאלגוריתם ה- Decision Tree ניתן להתבונן בגרף של עץ ההחלטה בקובץ </statistics/plots/dt.pdf> המכיל נתונים כמו הפיצ'ר שעל פיו מתבצעת ההחלטה בכל שלב, ערכי האנטרופיה שהחלוקה מעלה, הסיווג המתאים לכל עלה בעץ ועוד.

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 שלא לייצג כהלכה את הסופרים. כלומר, ייתכן שלא ייצגו את מגוון סגנונות הכתיבה וצורות הכתיבה השונות העשויות להשתנות בין ספריו של אותו הסופר. בנוסף לכך, השתמשנו בכלים `StanfordCoreNLP` ו-`nlk.pos_tag` כלים ששימשו אותנו לתיג הטקסט אשר באמצעותו יצרנו פיצ'רים נוספים. כלים אלו אינם מדויקים באופן אבסולוטי, וטעויות של תיג מוטעה 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

פרמטרים רבים משפיעים 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 Nearest Neighbors 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. מכאן 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 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, ובשל כך כל ספר היה מאוד משמעותי עבור הלמידה והוסיף לה עוד פן באופי כתיבתו של הסופר. לכן, כאשר הורדנו ספר שלם (מתוך 4 – 3 ספרים) מלמידת המודל גרמנו לו ללמוד פחות באופן משמעותי, דבר שעשוי 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 אותם הוא צריך ללמוד וליהות. ככל שמספר הסופרים גבוהה יותר הצלחת המודל קטנה, וזאת כי נדרש מהמודל לתאר כל סופר באופן מדויק יותר. בעתיד היינו רוצים להגדיל את עבודתנו 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 לרוב סיווג נכון אל הסופר המקורי, וזאת אף על פי שחלק ניכר מהפיצ'רים שהשתמשנו בהם היה תלוי שפה (הפיצ'רים התחביריים התמקדו בתגים שונים ומילים פונקציונליות בשפה האנגלית). עם זאת, 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, Persuasion, Sense and 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

References

- [1] Authorship Identification of Research Papers, Simen Skoglund 2015
- [2] Relevance of Syntactic and Discourse Features for Author Profiling and Identification, Boronat, Carrer, and Wanner
- [3] de Roc Boronat, Carrer, and Leo Wanner. "On the Relevance of Syntactic and Discourse Features for Author Profiling and Identification." EACL 2017 681 (2017).
- [4] Manning, Christopher D., Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP Natural Language Processing Toolkit In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations, pp. 55-60.
- [5] Project Gutenberg, Michael Hart and Gregory Newby 1971, accessed 17 August 2018, <<https://www.gutenberg.org/>>
- [6] מבוא לבלשנות תאורטית חלק ב', האוניברסיטה הפתוחה, אירנה בוטוויניק וגלית אדם, 2012. עמ' 118-133
- [7] Alphabetical list of part-of-speech tags used in the Penn Treebank Project , accessed 13 August 2018, <https://www.ling.upenn.edu/courses/Fall_2003/ling001/penn-treebank-pos.html>