Word cloud של מכל השירים בקורפוס שלנו

זיהוי מאפייני משורר

סיכום פרויקט בסדנת NLP עברי

אביעד דודקביץ

רון עמיר

הגדרת הבעיה

התרבות העברית התברכה בעושר ספרותי ולירי המתפרש על פני אלפי שנים, במאות השנים האחרונות וביתר שאת מראשית הציונות התווספו אלפי יצירות חדשות לקורפוס העברי, וחלקן הפכו לנכסי צאן ברזל. חלק חשוב ומשמעותי מהאוצר הספרותי העברי היא השירה העברית שנכתבה בז׳אנרים רבים ומגוונים.

ידוע כי היצירה משמשת לא אחת חלון לנשמתו של הכותב, היא משקפת את האישיות, ואת הרקע התרבותי ממנו הגיע, היצירה היא למעשה תבנית נוף מולדתו. הפרויקט אותו אנו מגישים שם לעצמו למטרה לבחון האם ניתן באמצעות מודלים של NLP לזהות בעזרת תוכן השיר בלבד מאפיינים של המשורר.

מודל כזה יכול לסייע רבות לתחום המחקר של השירה העברית. שימוש אפשרי אחד הוא ניסיון לזהות שירים עם מחבר אלמוני, למשל, ניתן להשתמש בכלי זה כדי לפענח את הזהות של המחבר, או לכל הפחות להבין מהיכן הוא הגיע ובאיזו תקופה השיר נכתב.

איתור וניקוי נתונים מתאימים

הקורפוס האידיאלי בו היינו מעוניינים הוא כזה שמכיל את כל השירים שנכתבו בשפה העברית ללא ניקוד ובמבנה אחיד אשר יכלול נתונים מלאים על אודות המחבר של כל שיר, למשל שנת הלידה, מקום הלידה, הגיל בו נכתב השיר ועוד. גם מאגר תיאורטי שכזה היה דורש עבודה רבה בכל הנוגע ליצירת פורמט אחוד ונקי, הן לשירים עצמם והן לפרטי המחברים ברמה שתאפשר את הזנת המאגר למודל NLP.

בפועל, מאגר שכזה לא קיים, לפחות לא באופן פתוח וחופשי לציבור. לכן לצורך הפרויקט שלנו, השתמשנו במאגר הזמין הדומה ביותר למאגר האידיאלי שלנו, המאגר של "פרויקט בן יהודה"[1]. המאגר כולל אלפי שירים שנכתבו בשפה העברית, בעיקר בעת החדשה וחלקם קודם לכן.

פרויקט בן יהודה מפיץ את המאגר אחת לתקופה באמצעות dump ל-github שנקרא public_domain_dump[2]. המאגר מכיל את רוב היצירות המופיעות בפרויקט בן יהודה, חלקן לא מופיעות עקב בעיות זכויות יוצרים. המאגר כולל קובץ csv המכיל נתונים בסיסיים על כל שיר כמו ניתוב לשיר, שם המשורר, ז׳אנר ועוד. ממאגר זה השגנו בסך הכל 7,022 שירים מ-162 משוררים שונים.

הבעיה המרכזית בה נתקלנו בבואנו ליצור מאגר מידע ראוי לפרויקט היה קשור להשגת מידע על המשוררים. כדי להתגבר על הבעיה הזאת, בחרנו לפנות לויקיפדיה, עשינו זאת באמצעות חבילת פייטון ייעודיות לויקיפדיה של pywikibot[3] ובחבילות פייטון אחרות לזחילה ברשת כמו beautifulSoup עליה למדנו בשיעור. את מרבית המידע אודות המשוררים אספנו מהאתר הייעודי של ויקיפדיה למאגרי מידע - wikidata[4]. חלק אחר מהמידע הגיע מזחילה על ויקיפדיה עצמה, לאור העובדה שלא לכל המשוררים היו ערכים ב-wikidata.

גם לאחר הזחילה על מאגרי ויקיפדיה נותרה כמות לא קטנה (כ-30% מהמשוררים) ללא מידע עליהם. זאת מכיוון שלא היה להם ערך ייעודי בויקיפדיה או ששמם אוית בצורה שונה מהצורה בה אוית בפרויקט בן יהודה. בנוסף, היה צורך לבצע ניקוי ידני של חלק מהנתונים ולהשלים נתונים חסרים על פי הערכות (לדוגמא - נתונה תקופת חייו של המשורר אך לא תאריך מדיוק של יום הולדתו ופטירתו). כך שלמעשה הייתה עבודה ידנית רבה להשלמת החוסרים למאגר המידע אודות המשוררים.

לבסוף נאלצנו להוריד 32 שירים (1.81% מהמאגר שצברנו) בגלל שלא יכולנו לזהות את המחברים המקוריים שלהם. נתרנו עם מאגר הכולל 6,990 שירים שסך המילים המופיעות בהם עומד על כ-1.7 מיליון.

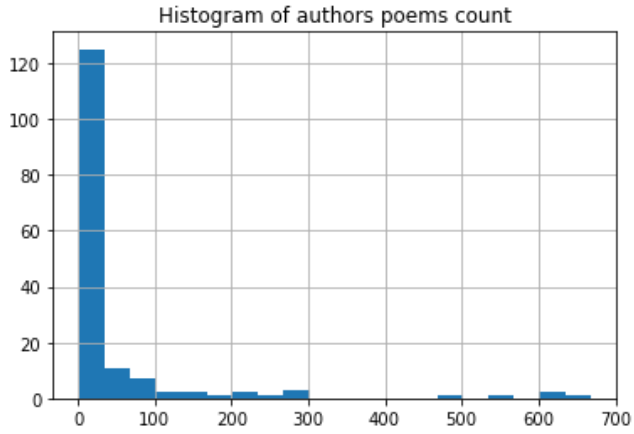
---

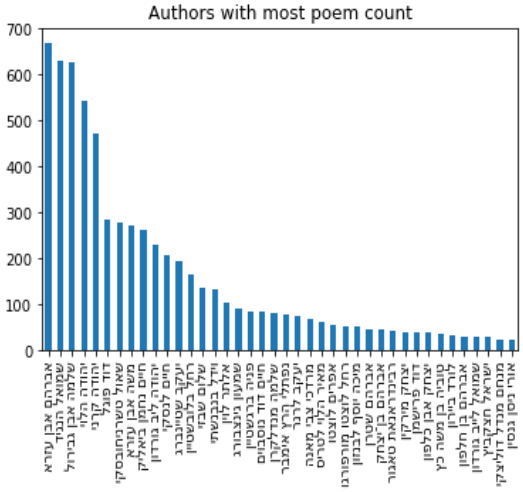
[1] https://benyehuda.org/
[2] https://github.com/projectbenyehuda/public_domain_dump
[3] https://www.mediawiki.org/wiki/Manual:Pywikibot
[4] https://www.wikidata.org

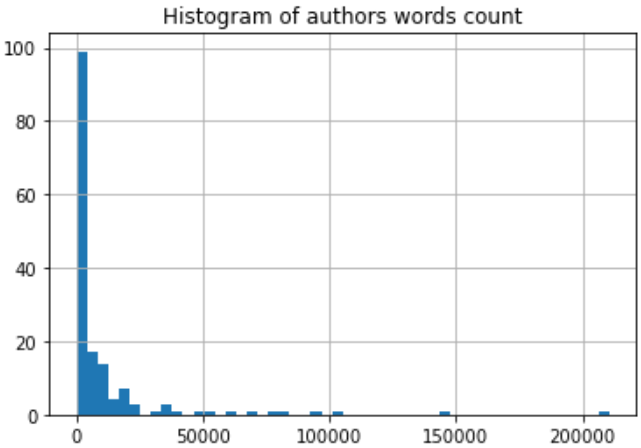
תובנות על הנתונים בידינו

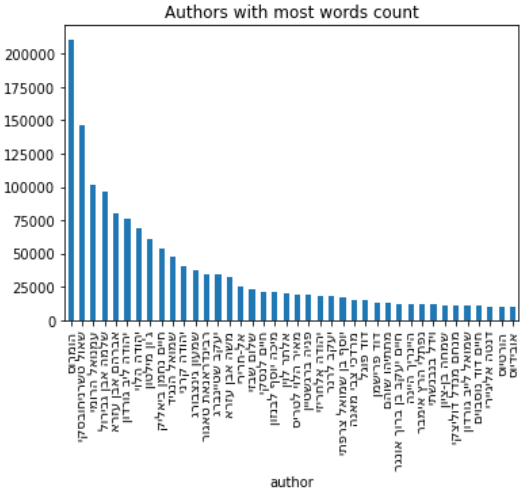
לאחר עיבוד הנתונים, ביצענו ניתוח שלהם (EDA). והגענו למסקנות הבאות:

- התפלגות תרומת כל משורר למאגר שלנו מזכירה התפלגות מעריכית עם λ גבוהה. זאת מכיוון שלרוב המשוררים במאגר המידע שלנו יש רק מעט שירים.

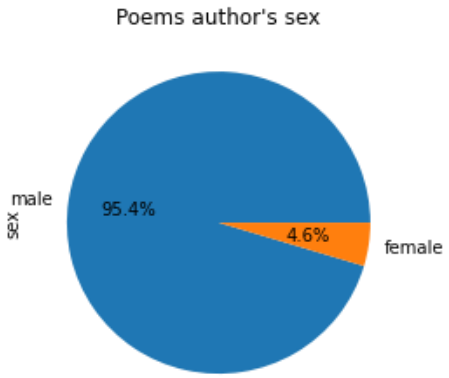




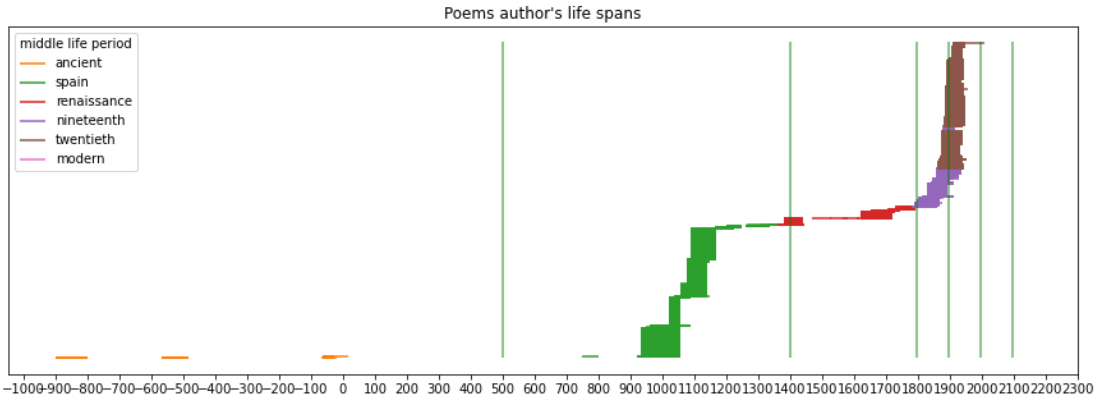


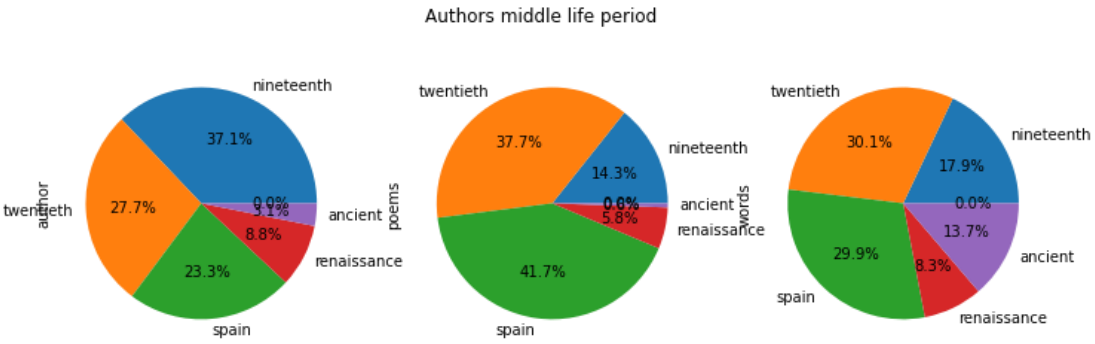


- המאגר לא מאוזן ביחס למין המשורר. יש מעט שירים במאגר שנכתבו על ידי משוררות. כך ששיערנו שהדבר יקשה על הלימוד של מאפיין זה.

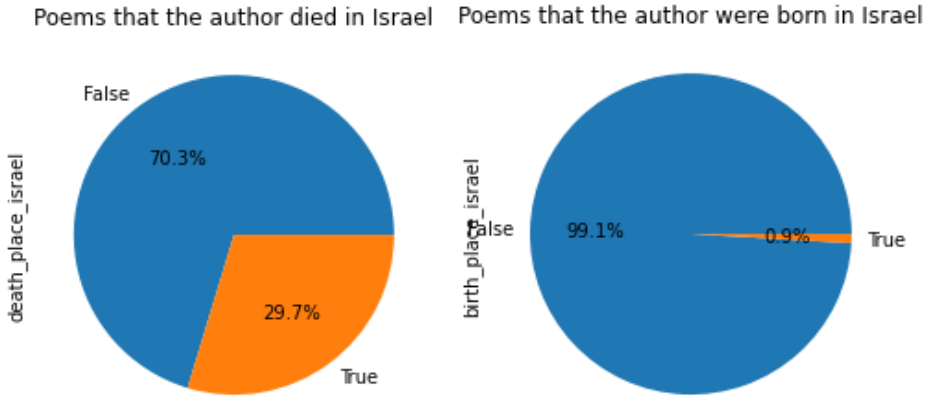


- רוב המאגר מורכב משירים שנכתבו בשנים 1800-1950. הדבר אילץ אותנו לקבוע לחלק התקופות בצורה שונה מהמקובל כדי שחלוקת התקופות שבחרנו תהיה קצת מאוזנת. בנוסף, מכיוון שאנו משתמשים בשירים של נחלת הכלל (public domain) כמעט ואין משוררים מודרניים.





- אין כמעט שירים שנכתבו ע״י משוררים שנולדו בישראל אך כן יש מספר מכובד של שירים שנכתבו ע״י משוררים שנפטרו בישראל. נתון זה חשוב משום שרצינו לאפיין משוררים ישראלים לעומת משוררים בחו״ל באמצעות שימוש במקום לידה או מקום פטירה. הבעיה שרבים מהמשוררים היו ציונים שנולדו בחו״ל ועלו לארץ ואף יש חריגים שחיו בארץ אך נולדו וגם נפטרו בחו״ל (לדוגמא: חיים נחמן ביאליק עלה מגרמניה לישראל והתגורר בתל אביב אך נפטר בווינה).



סקירת ספרות - גישות מקובלות לבעיה והמודל הנבחר

זיהוי מאפיינים של משורר מהווה חלק מהתחום של בעיית זיהוי מחבר (authorship attribution) שהוא אחד מתחומי המחקר המרכזיים של NLP, תחום זה מהווה אתגר משמעותי ויש שפע של מאמרים בנושא

וניסיונות לזהות מחברים בתחומים שונים. אחד האתגרים בתחום זה הוא זיהוי של טקסטים קצרים יחסית, כמו שירים שחלקם יכולים להיות קצרים מאוד. לצורך הדוגמא, במאגר שלנו אורך הממוצע של שיר הוא כ-146 מילים.

יש מספר גישות מקובלות להתמודד עם בעיית זיהוי המחבר, חלקן קלאסיות בלשניות וחלקן סטטיסטיות, נתמקד בשיטות הסטטיסטיות כיוון שהן יותר רלוונטיות לתחום ה-NLP:

1. שיטות סטטיסטיות גאוסיות - שיטה קלאסית יחסית, על פיה באמצעות שיטות גאוסיות ו- data transformation ניתן לנסות לשייך טקסט לכותב שלו.

2. רשתות נוירונים - רשתות נוירונים מהוות פיתוח של השנים האחרונות והציגו קפיצת מדרגה משמעותית בכל הנוגע לאחוזי הדיוק של החיזוי. הן הופכות יותר ויותר לסטנדרט של התעשייה.

3. אלגוריתמים גנטיים - בשיטה זאת מתחילים עם סט חוקים דטרמיניסטיים, בכל איטרציה בוחנים אילו חוקים סיפקו תוצאות טובות ואילו לא ומחליפים אחוז מסויים מהחוקים הלא טובים. במקביל, עורכים שינויים קטנים בחוקים הטובים. לאחר איטרציות רבות או לאחר שמגיעים לתוצאה מספקת מקבלים מודל מוכן.

כמובן שגם ניתן לשלב בין שיטות.

שיטת הייצוג של הטקסט בכל מודל שייבחר גם היא בעלת השפעה משמעותית. בתחילה, חשבנו לנסות לייצג את הטקסטים בעזרת מורפמות, דבר אשר יכול לצמצם משמעותית את מרחב הקלט למודל שנבחר ולשפר את ביצועיו. אך מכיוון שבמאגר ברשותנו היו שירים רבים אשר החולקה למשפטים בהם לא הייתה ברורה, נאלצנו לוותר על השימוש במורפמות. לדוגמא, ישנם שירים מהתקופה העתיקה אשר אינם משתמשים בפיסוק והחלוקה נקבעת על ידי שורות ובתים.

אנחנו בחרנו לעשות שימוש במודל השני ובפרט ב-AlephBert המהווה מודל של רשתות טרנספורמרים. בחרנו במודל הזה בראש ובראשונה כי הוא מותאם לשפה העברית וכן מכיוון שהוא מהווה רשת של טרנספורמרים שהציגה תוצאות עדיפות על רשתות נוירונים ותיקות יותר כמו RNNs ו-LSTMs.

המודל של AlephBert בנוי למעשה מכמה טרנספורמרים, טרנספורמרים הם יחידות של encoders-decoders שכל אחד מהם מורכב משרשור של שכבות Attention ורשתות FeedForward, הכוח של הרשתות הללו הוא שכל transformer יכול להתמקד בפיצ'ר מסוים בטקסט וללמוד אותו.

תוצאות המודל והערכתם

המאפיינים של המשורר שבחרנו לתת למודל שלנו לסווג לשיר שיינתן כקלט הם:

1. תקופת הלידה ותקופת המוות - לפי החלוקה הבאה:

   א. עתיקה (לפני 500)

   ב. ספרד (500-1400)

   ג. רנסאנס (1400-1800)

   ד. המאה ה-19 (1800-1900)

   ה. המאה ה-20 (1900 והלאה)[5].

[5] לא היה מספיק שירים ממשוררים שנפטרו אחרי שנת 2000, ולכן לא יכלנו להשתמש בקטגוריה של המאה 21.

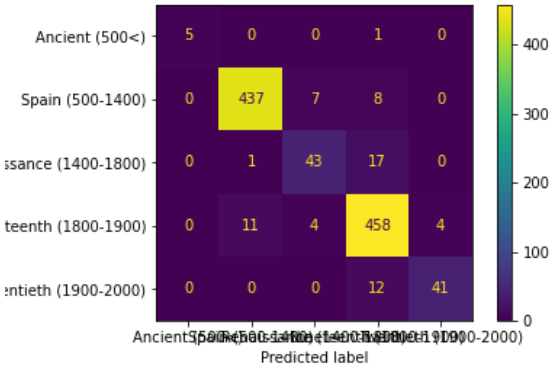
2. מקום הלידה והמוות - ארץ ישראל או חו"ל.

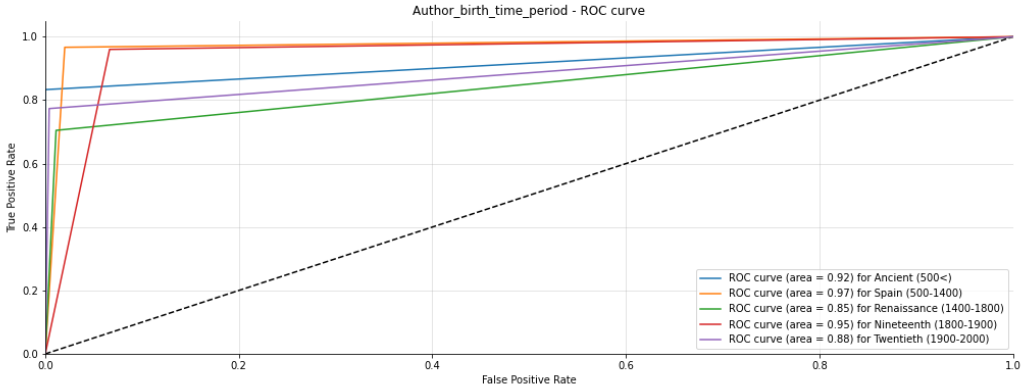
3. מין המשורר - זכר או נקבה.

לצורך ההשוואה הרצנו גם מסווג קלאסי של יער עצי בחירה (Random forest).

תקופות

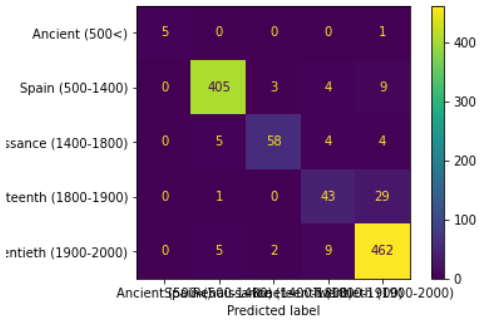
ניכרת הצלחה משמעותית בזיהוי תקופת הלידה של המשורר, כאשר fscore במרבית הקטגוריות היה מעל 0.9. ניכר כי התקופה הקשה ביותר לניבוי היא הרנסאנס (1400-1800), אך גם בה הדיוק (precision) קרוב ל0.8.

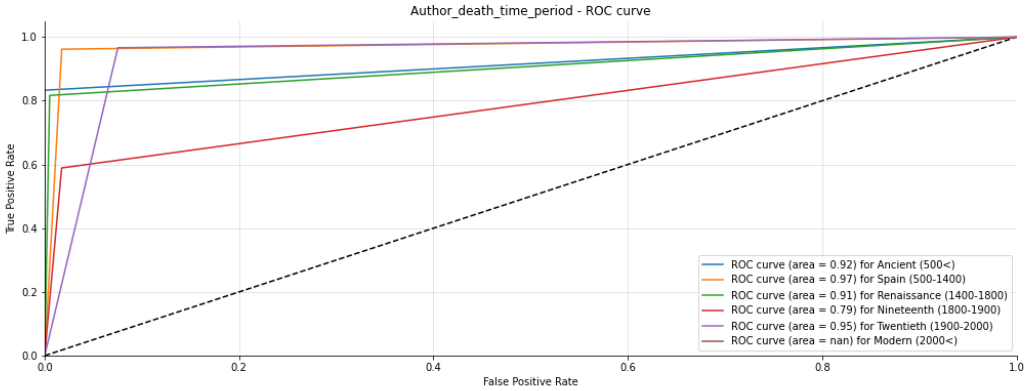




גם בתקופת הפטירה של המשורר ניכרת הצלחה, כאשר גם כאן fscore במרבית הקטגוריות היה מעל 0.9.

לעומת תקופת לידה, כאן התקופה הקשה לניבוי היא המאה ה-18 (1800-1900), עם דיוק (precision) קצת מעל 0.7. זאת, מכיוון שהמסווג בטעות שייך שירים רבים למשוררים שנפטרו בתקופה הסמוכה שהחלוקה ביניהן די שרירותית.

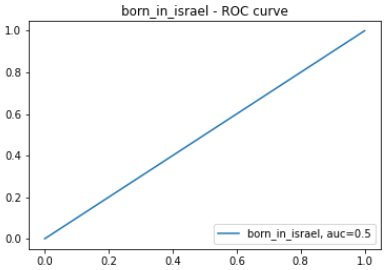


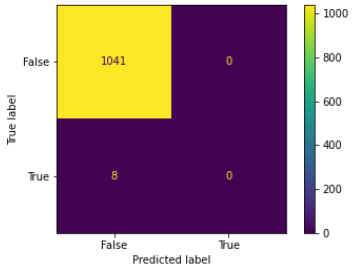


לצורך ההשוואה, המסווג של הלמידה הקלאסית בשתי הקטגוריות לא הצליח ללמוד תקופות מסוימות ובעיקר סיווג רק לשתי תקופות.

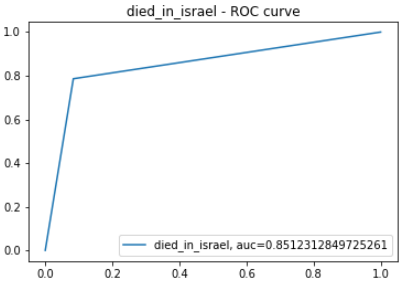
לידה ופטירה בישראל

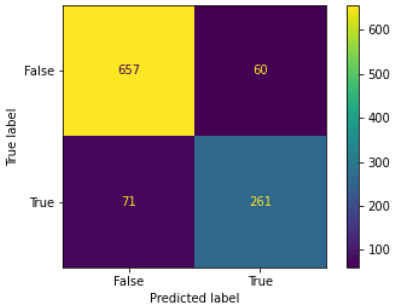
מכיוון שהיו מעט מאוד שירים בהם המשוררים נולדו בישראל, לא הייתה יכולת ללמוד את המאפיין הזה, ולכן המודל שלנו סיווג את כל השירים ככאלה שנכתבו על ידי משוררים שנולדו בחו״ל.





בניגוד ללידה, בעניין הפטירה כבר היה מספר משמעותי של **משוררים שנפטרו בישראל**, המודל הציג בעניין זה תוצאות מרשימות. רמת הדייקנות (accuracy) הייתה של כ-0.87 כאשר fscore עבור משוררים שנפטרו בישראל הוא כ-0.8.



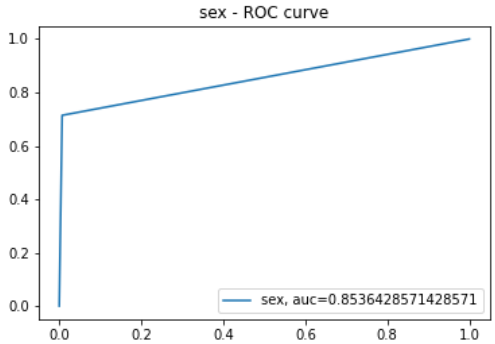


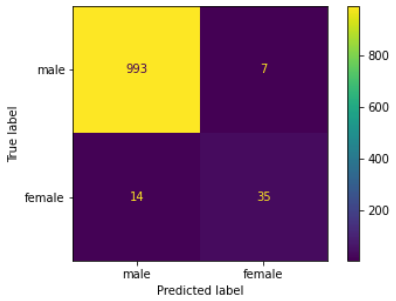
לצורך ההשוואה, המסווג הקלאסי, גם לאחר התאמה לנתונים לא מאוזנים[6], הגיע לרמת דייקנות (accuracy) של כ-0.74 למשוררים שנפטרו בישראל כאשר ה-fscore הוא 0.2. (בקטגוריית המשוררים שנולדו בישראל הסיווג היה זהה למודל שלנו, סיווג כל השירים ככאלה שנכתבו על ידי משוררים שנולדו בחו״ל.)

מין

חרף מיעוט הנשים במאגר המידע שלנו, באופן מפתיע לטובה, המודל הצליח להגיע לרמת דיוק (precision) של כ-0.83 לזיהוי שיר שנכתב ע״י אישה, יחד עם fscore של כ-0.77.

[6] שימוש באלגוריתם SMOTE - https://arxiv.org/pdf/1106.1813.pdf





לצורך ההשוואה, המסווג הקלאסי, גם לאחר התאמה לנתונים לא מאוזנים[7], הגיע לאיחזור (recall) של 0.2 וfscore של 0.04., נתונים שמאירים באור חיובי ביותר את היכולת של המודל שלנו.

מסקנות

התוצאות מלמדות אותנו שבהחלט ניתן להבין מה המאפיינים של הכותב רק על סמך השירים שלו, התוצאות של AlefBert היו מרשימות במיוחד, גם בהתחשב בעובדה שכמות ה-data שהצלחנו להשיג לא הייתה גדולה במיוחד.

המסווג שבנינו הצליח להתגבר ללא התערבות חיצונית גם על מחלקות לא מאוזנות, כפי שניתן לראות בקטגורית המין וגם על סיווג לא בינארי, כפי שניתן לראות בקטגוריית תקופת הלידה ותקופת המוות.

מה להמשך?

התוצאות המעודדות שקיבלנו יכולות להוות בסיס שעליו ניתן לבנות ניתוחים מתקדמים הרבה יותר על מחבר, החל מדעה פוליטית, דעות כלכליות ואף מעבר לכך. פיתוח אפשרי נוסף הוא ביצוע ניסיון ללכת צעד אחד קדימה ולסווג את הטקסט שנכתב לכותב ספציפי, (דבר שחשבנו לבצע בתחילה אך היה נראה לנו קשה מידי).

ביצוע של אנליזות מתקדמות שכאלה יהיה כרוך באיתור מאגר מידע רחב הרבה יותר ויכולת אוטומטית טובה יותר להשגת מידע על הכותבים של כל שיר ושיר.

היינו שמחים לשתף פעולה עם פרויקט בן יהודה בכדי לקבל גישה ליותר חומרים, גם שיתוף פעולה עם גופים מסחריים כמו שירונט יכול לתרום לנו רבות. שינויים מינוריים במודל יכולים להתאים אותו לשימושים נוספים, למשל ניסיון לחזות האם שיר יוערך או לא, האם הוא יהיה להיט ועוד.

בסיכומו של דבר, ניכר כי משימת הזיהוי של מאפייני המשורר היא משימה שניתן לבצע וללמוד ממנה פרטים רבים ומפתיעים על כל שיר ושיר.

[7] שימוש באלגוריתם SMOTE - https://arxiv.org/pdf/1106.1813.pdf