ANDREW AND ERNA VITERBI FACULTY OF ELECTRICAL ENGINEERING

SIPL
Signal and Image Processing

Technion
Israel Institute of Technology

דו"ח סיכום פרויקט: ב'

זיהוי של בודק אנונימי של מאמר אקדמי

Identification of an Anonymous Reviewer of an Academic Paper

מבצעים:

Roy Hachnochi **רועי החנוכי**

Lior Kiassi **ליאור קיאסי**

מנחה: **Pavel Lifshitz** **פבל ליפשיץ**

סמסטר רישום: אביב תש"פ

תאריך הגשה: נובמבר, 2020

P 5614-2-20

תוכן עניינים

רשימת איורים

תקציר

הפרויקט עוסק בזיהוי בודקי מאמרים אקדמיים באמצעות שיטות מבוססות עיבוד שפה טבעית ולמידה עמוקה. מאמרים אקדמיים וביקורות אודותם ניתנים באופן אנונמי, אך יתכן שבאמצעות כלי למידת מכונה ניתן להוכיח כי קיימת האפשרות לזהות את זהות כותב הביקורת. אם יוכח שכך הדבר, יכולה להיות לכך השפעה רבה על אופן מתן ביקורות, שכן אי האנונימיות יכולה לפתוח פתח לדעות קדומות, השפעות תחרותיות ועוד.
ככל הידוע לנו נושא זה לא נחקר, וטומן בחובו עומק מחקרי רב: הביקורות הינן מטבען טקסט קצר, ובאופן מובן אנונימיות – לכן אין מידע רב זמין בנושא לעסוק בו. עקב כך, משימת העל של הפרויקט הינה בעיית cross-domain: אימון על מרחב אחד והסקה על מרחב אחר ושונה במהות, במקרה זה מדובר באימון על מאמרים אקדמיים והסקה על ביקורות עם datasets קטנים יחסית.
במסגרת הפרויקט עברנו תהליך של מחקר בתחום עיבוד השפה הטבעית וזיהוי מאפיינים ופעולות המתאימות למשימות שונות שיסייעו לבעיית העל: בחינת המודל על toy problem פשוטה, בחינת המודל במסגרת בעיה של אימון על מאמרים אקדמיים וזיהוי של מאמרים אקדמיים ואימון על ביקורות וזיהוי של ביקורות.
במהלך הפרויקט ביצענו תהליך של מיצוי מאפיינים מתאימים לבעיות הנ״ל, אספנו ויצרנו datasets, יצרנו תהליך עיבוד מקדים מתאים וכן השתמשנו באלגוריתמים שונים המסייעים למחקר ולטיוב המודל.
בדו״ח זה נראה שהצלחנו להגיע לתוצאות טובות בתחום שמטבעו הינו מאתגר מאוד, נציג את תהליכי המחקר והניתוח אשר בוצעו במהלך הפרויקט, ונסיק מסקנות הנוגעות להמשך מחקר ועבודה על הבעיה.

Abstract

This project deals with the identification of academic paper reviewers, using methods from the field of Natural Language Processing (NLP) and Deep Learning (DL).
Academic papers and reviews are traditionally given anonymously, but it is possible thar using advanced Machine Learning tools it may be proven that there is the possibility of identifying the identity of the reviewer. If so, this could have a major income on the way in which academic reviews are given, since the de-anonymization may open the door to prejudice, competitive effects and more.
As far as we know, this problem has not been studied before, and holds within it interesting research topics: the reviews are naturally short texts, and anonymous – so there is very little data to handle.
As a result, the main goal of the project is a cross-domain problem: learning on the domain of academic papers and inference in the domain of academic reviews with fairly small datasets.
As part of the project, we conducted research in the fields of NLP and feature extraction for the projects main goal: testing the model on a "simple" toy problem, testing the model on a similar problem of training and inferring on academic papers, and training and inferring on academic reviews.
During the project we performed a process of feature extraction for the above problems, created datasets, performed preprocessing steps, and used various Machine Learning algorithms in order to improve our model and our understanding of the problem.
In this report we will show that we managed to achieve very good results in a research field which is naturally challenging, we will show our research and analysis, and draw conclusions regarding future work on this interesting problem.

1. מבוא

בפרק זה נסקור את המוטיבציה לפרויקט, חומר רקע רלוונטי, סקר ספרות ומידע נוסף שיסייע להבנת מטרת הפרויקט ודרכי ביצועו.
הפרויקט כתוב בשפת Python, תוך שימוש ב-Anaconda ו-Pycharm והיעזרות בספריות שונות כגון : NumPy, sklearn, Matplotlib, nltk, Pytorch, Hugginface, skrebate ועוד.

1.1. מטרת הפרויקט

מטרת הפרויקט הינה ביצוע מחקר אודות בעיית זיהוי כותבי ביקורות על מאמרים אקדמיים, תחום אשר טרם נחקר. בפרויקט נבצע מחקר אודות פתרון אפשרי לבעיה – למידה על מאמרים אקדמיים של כותבי הביקורות, והסקה על הביקורות. הפרויקט מהווה שלב ראשון לקראת התמודדות עם הבעיה הנ"ל, ולקראת חקר אפשרויות ההסתרה של זהות כותבי הביקורות.
הפרויקט מציג התמודדות עם תחום בעייתי שלא נחקר מעולם (ביקורות על מאמרים) ונעזר בתחום נוסף לשם מטרתו.

1.2. מוטיבציה

כחלק מתהליך פרסום מאמר אקדמי ישנו שלב של ביקורת עמיתים (ביקורת הניתנת בד"כ ע"י מספר מומחים בתחום). שלב זה מתבצע לרוב בשיטת "Double Blind" – כותב המאמר וכותב הביקורת אנונימיים אחד לשני. כיום, מטעמי שקיפות, מספר מגזינים מוכרים החלו לפרסם את הביקורות שנכתבו למאמרים שהוגשו להתפרסם אצלם.
בפרויקט זה אנו חוקרים את הנחת המוצא כי כותב הביקורת אכן הינו אנונימי, תוך זיהוי המאפיינים בטקסט אשר מעידים על זהותו. במידה ויוכח שבאמצעות כלי למידת מכונה ניתן לזהות את זהותו של כותב ביקורת (דה-אנונימיזציה), תהיה לכך השפעה רבה על שיטת מתן וכתיבת הביקורות, שכן יוסק שהתהליך האנונימי כלל אינו אנונימי. אם כן, מטרת העל הינה להראות שאכן ניתן לזהות זאת, ולהציע פתרונות אשר יכולים לסייע למזעור בעיה זו והשבת האנונימיות. פרויקט זה הינו השלב הראשון בהתקדמות למטרה זו, ועוסק בעיקר בשלב בניית ה-dataset, בניית המודל, ומציאת המאפיינים האינפורמטיביים ביותר לצורך המשימה.

1.3. רקע - עיבוד שפה טבעית

עיבוד שפה טבעית (Natural Language Processing) עוסק בחקר ועיבוד של "שפה טבעית" – כל שפה שהתפתחה באופן טבעי אצל בני אדם תוך שימוש חוזר בחלקי השפה וללא תכנון קודם. שפות אלה הן טבעיות, זאת בניגוד לשפות פורמליות כגון שפות תכנות, לוגיקה ועוד.
עיבוד שפה טבעית עוסק באינטרקציה שבין מחשבים לבני אדם באמצעות שפות טבעיות, ולרוב מערב מתודות של למידת מכונה ולמידה עמוקה. התחום רלוונטי מאוד ונחקר רבות בתחום הבינה המלאכותית. בעיות נפוצות

בעיבוד שפה טבעית הינן: זיהוי דיבור, תרגום מכונה, זיהוי רגשות ותחושות בטקסט, תקצור טקסט אוטומטי, סיווג טקסט לז׳אנר ועוד.

1.4. רקע – Authorship Attribution

Authorship Attribution הינה משימה במסגרת עיבוד שפה טבעית של זיהוי כותב של טקסט. תחום זה גורר התעניינות עולמית רבה ובעל אפליקציות רבות כגון חקירה פורנזית וזיהוי הונאות. במסגרת בעיה זו יש משימות/בעיות מוכרות רבות: זיהוי כותב טקסט, בניית פרופיל של כותב טקסט (גיל, מגדר וכו׳), זיהוי ז׳אנר של טקסט, אשכול טקסטים ועוד.

1.5. אתגרים בפרויקט

כנגזרת מהמוטיבציה לפרויקט, ישנם אתגרים ייחודים ומשמעותיים בו:

- לא קיים dataset מסודר ומתויג למאמרים ולביקורות.
- ביקורות הינן טקסטים קצרים מאוד, מה שמוביל למעט מידע אותו ניתן לחקור וממנו להסיק מסקנות למודל.
- ביקורות ומאמרים מטבעם מוכוונים נושא ביתר שאת, ועל כן יש קושי משמעותי יותר בהבנת סגנון כתיבה של כותב.
- לרוב מאמרים אקדמיים נכתבים על ידי יותר מכותב אחד (ללא ציון זהות הכותב בכל חלק).
- בעיית cross-domain – מאמנים את המודל על מאמרים ובוחנים אותו על ביקורות. שני סוגים אלה של טקסטים הינם שונים מאוד מטבעם ולכן מאפיינים של אחד מהם אינם בהכרח מתאימים לשני. אם כך, נדרשת פעולה נוספת להמרה בין שני התחומים.

2. סקר ספרות

תחילה, בדקנו עבודות קודמות בתחום והאם הנושא נחקר. כחלק מהעבודה על הפרויקט למדנו את תחום עיבוד השפה הטבעית. העמקנו ספציפית בפרסומים ומחקרים העוסקים בזיהוי כותב טקסט. כמו כן, למדנו את תהליך פרסום מאמר אקדמי וכתיבת ביקורות למגזינים, אתרים וכנסים. סקר הספרות התמקד בפרסומים הבאים:

- קורס אינטרנטי של אוניברסיטת סטנפורד העוסק בעיבוד שפה טבעית (CS224N). [1]
- סיכום של תחרות PAN לשנת 2018, שעסקה במשימה של זיהוי כותב טקסט בבעיית cross-domain וזיהוי שינוי סגנון כתיבה. [2]
- סיכום של תחרות PAN לשנת 2019, שעסקה במשימה של זיהוי כותב טקסט בבעיית cross-domain. [3]
- עבודת תזה שנכתבה ע״י Yunita Sari בנושא גישות שונות לזיהוי כותב טקסט. [4]
- מאמרים רבים העוסקים בשיטות שונות של מיצוי מאפיינים לטובת משימת זיהוי כותב טקסט (כולל ניתוח סמנטי, לקסיקלי וסינטקטי). [6] [5]
- מאמרים רבים העוסקים באימון מודלי שפה באמצעות רשתות עצביות. [8] [7]

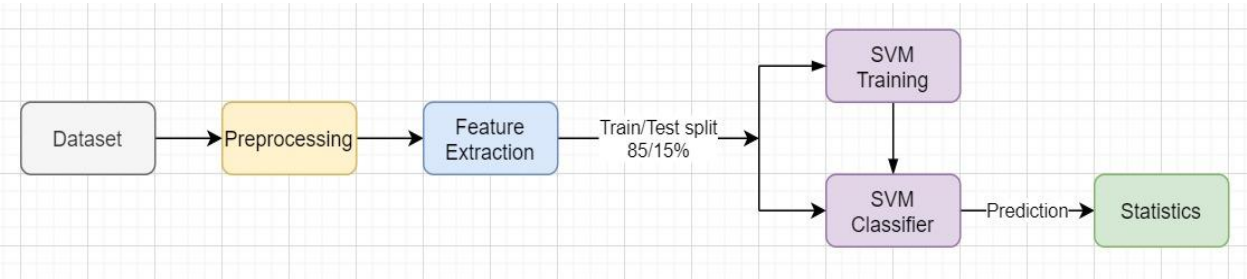
3. Toy Problem

כשלב מקדים לפרויקט בחרנו לעבוד על משימה פשוטה של זיהוי כותבי ספרים. הבעיה הנ״ל למעשה שימשה כ-Baseline בהמשך הפרויקט להשוואת תוצאות והרצות שונות.

ה-data עמו עבדנו הינו 100 ספרים – 10 ספרים לכל אחד מ-10 סופרים (המידע נלקח מפרויקט Guttenberg [9].מטרת המשימה היא לייצר מודל המזהה את זהות כותב ספר מסוים.

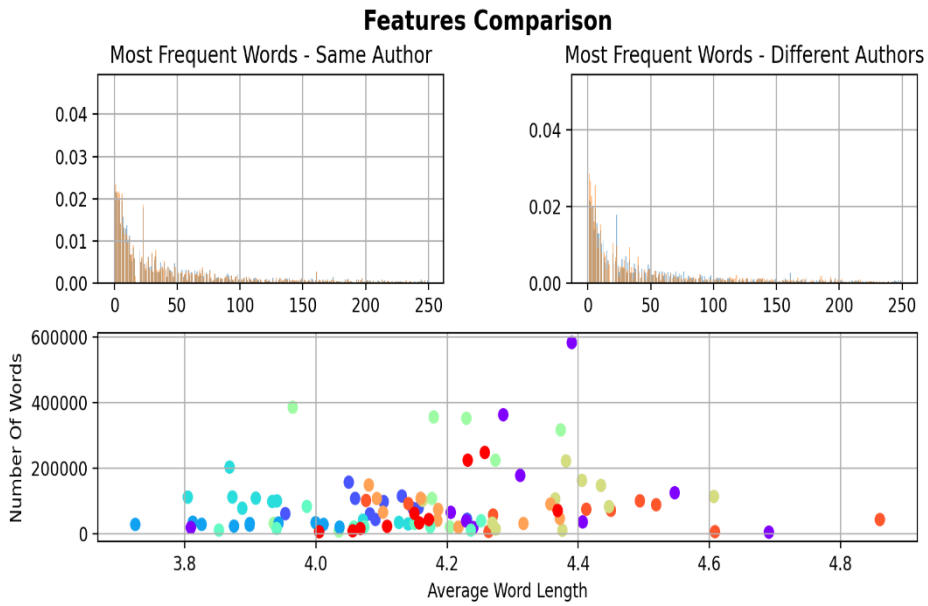
מדובר בעיסוק ב-data יחסית פשוט שכן מדובר בספרים בעלי תוכן שונה במהות, כותבים עם סגנונות שונים, data מונגש בצורה יחסית נוחה למשתמש, טקסטים ארוכים, כותב יחיד לספר ועוד. עובדה זו הובילה אותנו לעבודה על הבעיה הנ״ל והסתמכות עליה כבסיס השוואה טוב להמשך, לצורך בדיקת המודל וטיובו.

3.1. המודל



איור 1 - דיאגרמת בלוקים של toy problem

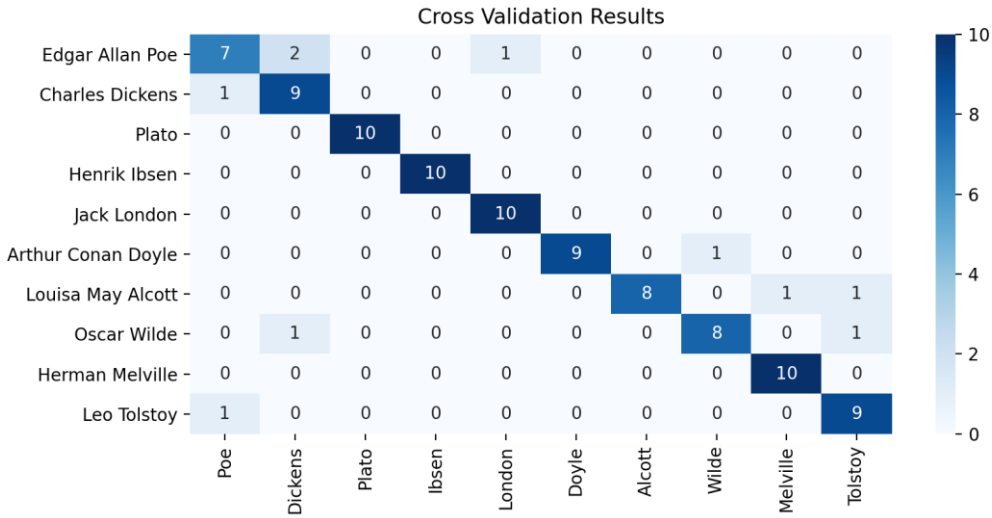
- **Dataset** : המחלקות הינן 10 סופרים וה-data הינו 10 ספרים לכל אחד מהם.
- **Preprocessing** : שלב מקדים להקלת העיסוק במידע המתקבל למודל, כולל את הפעולות הבאות :
  - ״ניקוי״ הטקסט – הורדת שם האתר ממנו הורדנו, שמות הסופרים, רפרנסים מפלילים ועוד.
  - Tokenizing – ביצוע הפרדה ל-tokens של כלל המידע, כלומר כל מילה וסימן הינם אובייקט נפרד.
  - התעלמות מחלק מסימני הפיסוק.
- **Feature Extraction** : פעולה של מיצוי מאפיינים שימושיים ובעלי קורלציה למחלקות. נבחרו המאפיינים הבאים :
  - היסטוגרמה של 100 המילים הנפוצות ביותר מתוך כלל סט האימון.
  - היסטוגרמה של 100 המילים הכי פחות נפוצות מתוך כלל סט האימון.
  - ממוצע אורך מילה בטקסט.
  - מספר מילים בטקסט.



איור 2 - פיזור מאפיינים של toy problem

בשני הגרפים העליונים ניתן לראות את מאפייני ההיסטוגרמות בהשוואה בין שני טקסטים שונים (מנורמלים). בגרף התחתון ניתן לראות פיזור של כלל הטקסטים לפי מספר מילים וממוצע אורך מילה, כאשר הצבע מסמל את המחלקה.

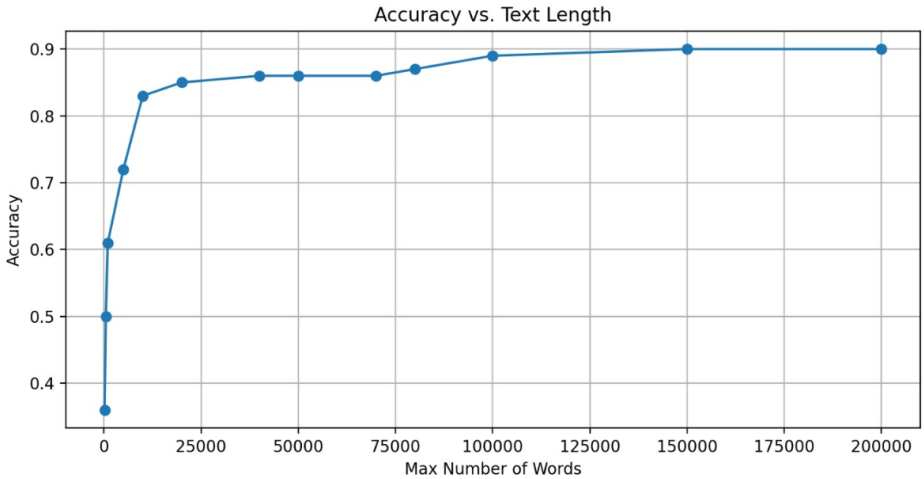
- **SVM** : בחרנו להשתמש במסווג מסוג זה עם גרעין RBF.
- **Statistics** : בחרנו להציג את אחוזי הדיוק (90% בממוצע) ואת ה-confusion matrix.



איור 3 - תוצאות confusion matrix של toy problem

3.2. יחס אורך טקסט - דיוק

במסגרת בחינת המידע, ביצענו בדיקה להבנת השפעת אורך הטקסט על אחוזי דיוק המודל. בדיקה זו חשובה ביותר לעבודה על בעיית העל שלנו, אשר כוללת אימון על אורך טקסט בינוני (מאמר אקדמי) וקצר (ביקורת).
להלן גרף התוצאות:



איור 4 - אחוז דיוק כתלות באורך טקסט - toy problem

ציר x הוא אורך הטקטס המקסימלי שלקחנו מכל ספר, וציר y הוא אחוזר הדיוק אשר הושג באמצעות המודל. ניתן לראות כי ישנה השפעה משמעותית של אורך הטקסט על אחוזי דיוק המודל – ככל שהטקסט ארוך יותר, קיים מידע רב יותר לחקור ולאבחן – עובדה אשר מובילה לאחוזי דיוק גבוהים יותר. ניתן גם לראות עלייה משמעותית בסביבת 10,000 מילים – אורך בינוני אשר מתאים למאמרים אך לא לביקורות.

4. Datasets

אתגר משמעותי בפרויקט הינו החוסר במידע מתוייג ומסודר. במסגרת הפרויקט הצטרכנו לעבוד עם שני סוגים שונים של טקסט: מאמרים אקדמיים וביקורות.

הסיבה המרכזית לכך שבפרויקט סט האימון וסט המבחן לקוחים מתחומים שונים (אימון על מאמרים והסקה על ביקורות), היא שלרוב כותבי הביקורות לא קיים מאגר מידע מספיק גדול שעליו ניתן ללמוד. מנגד, כותבי ביקורות הם לרוב חוקרים אקדמאים בעצמם, ולכן כן ניתן למצוא מאגר של מאמרים. אם כן, בפרויקט זה אנו משתמשים במאמרים ללמידה, בתקווה להסיק על הביקורות.

מאמרים אקדמיים מטבעם כוללים מספר מאפיינים אשר גורמים לעיסוק בטקסט להיות מורכב יותר:

- מספר כותבים לכל מאמר (בלי יכולת ממשית לשייך חלק בטקסט לכותב).
- לרוב מופיעים בפורמט שהינו פחות נגיש לעיבוד – PDF.
- מאמרים שונים במבנה שלהם ולכן קשה לבצע אוטומציה בעיסוק בהם (מספר עמודות שונה של הטקסט, תמונות וגרפים, פורמטי טבלאות שונים, מספר כותרות משנה, מיקום שמות הכותבים וכו').
- קושי בהשגת מאמרים של אדם ספציפי ובחלק מהמקרים אף יש צורך לשלם כדי להשיג גישה למאמר.

ביקורות כללו את המאפיינים המאתגרים הבאים:

- מדובר בד"כ בטקסטים אנונימיים, וכמעט ואין ביקורות באופן נגיש באינטרנט, ובייחוד ביקורות מתויגות.
- תיתכנה מספר ביקורות לאותו המאמר (לדוגמא במקרה בו כותב מאמר מתבקש לבצע תיקונים וכו').
- מרבית כותבי הביקורות כותבים מספר מועט של ביקורות בלבד.
- לעיתים ביקורות הינן טקסט קצר במיוחד שאינו כולל תוכן משמעותי כלל.

4.1. הכנת ה-Dataset

לאחר מחקר באינטרנט, מצאנו כי המגזין BMJ [10] מפרסם את הביקורות שנכתבו על מאמרים אשר הוגשו לפרסום במגזין. BMJ (British Medical Journal) הינו מגזין בריטי העוסק בתחומי הרפואה השונים. הורדנו ביקורות מהאתר של המגזין:

- תחילה, הורדנו דפי html אשר כללו את תוכן הביקורות – הורדנו באמצעות ריצה כוללנית על כל ה-URLs המתאימים אשר מחפשת כמה שיותר ביקורות שקיימות באתר.
- חקרנו והבנו את מבנה הדפים וכך ביצענו פרסור של הביקורות – הוצאת שמות הכותבים ותוכן הביקורת.
- ניפוי טקסטים שכללו מספר מועט של מילים.
- שילוב מחלקות שהופרדו עקב הצגת מידע באופן שונה (לדוג': הוספת שם אמצעי).
- ניפוי מחלקות שלא כללו לפחות 7 ביקורות.

לאחר תהליך זה נותרנו עם 7 מחלקות שיחדיו הגיעו ל-168 ביקורות.

נציין כי הורדנו ופירסרנו ביקורות רבות ממקורות נוספים (שאינם BMJ) אולם לא השתמשנו בהם בסופו של דבר.

לאחר מכן, בנינו dataset של מאמרים:

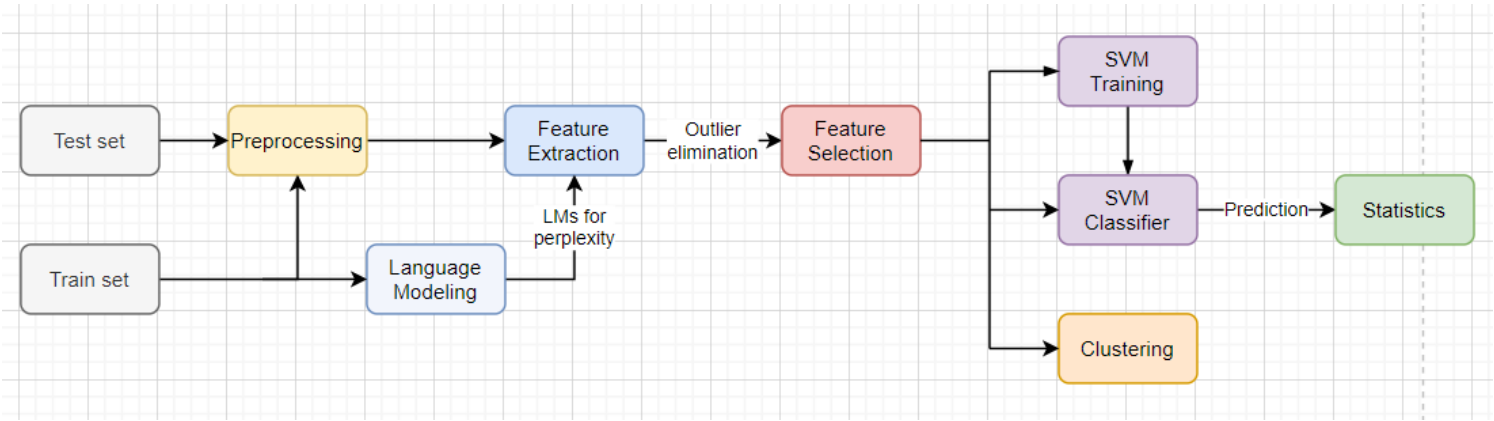
- חיפשנו מאמרים עבור כל אחת מהמחלקות (הכותבים) הנ"ל ברחבי האינטרנט באופן ידני.
- הורדנו 10 מאמרים בגדלים שונים לכל מחלקה.
- המרנו את המאמרים מפורמט של PDF ל-txt.
- הורדה ידנית של מידע מיותר/מפליל וכן סידור הטקסט לעמודה אחת.

נציין שכפי שניתן לראות שלב הורדת ה-dataset הוא מאתגר ואף סיזיפי, אך איכותו חשובה מאוד להצלחה הפרויקט ולאמינות התוצאות.

5. מבנה הפרויקט

בפרק זה נתאר את מבנה הפרויקט והתהליך שביצענו. נציג דיאגרמת בלוקים כללית של מודל הסיווג. נפרט בנוגע לכל בלוק של המודל. בפרט, נציג את תהליך העיבוד המקדים, תהליך מיצוי המאפיינים אשר כולל סוגי מאפיינים רבים, שלב נוסף של בחירת מאפיינים, ומודל הסיווג והאשכול.

5.1 דיאגרמת בלוקים



איור 5 - דיאגרמת בלוקים של הפרויקט

5.2 עיבוד מקדים

בניגוד לעיבוד המקדים הידני שהוצג בפרק 3.1, שלב זה הינו שלב אוטומטי שהינו חלק מה-pipeline של המסווג. העיבוד המקדים שאנו מבצעים כולל (חלק מהשלבים בוצעו רק עבור מאפיינים מסוימים):

- tokenizing – הפרדת מילים וסימני פיסוק, כך שכל מילה תיוצג ע״י אובייקט נפרד (למשל מחרוזת יחודית).
- התעלמות ממילות עיצור (stop words) – זהו שלב מקדים מוכר בתחום עיבוד השפה הטבעית, ומשמעותו התעלמות ממילות קישור נפוצות בשפה (למשל the). את שלב זה עשינו באמצעות ספריית NLTK [11].
- lemmatization – ממיר כל מילה לצורת הבסיס שלה (למשל: kids -> kid ,studies, studying -> study).
- ניקוי מספרים וסימנים חריגים והחלפתם ב-token של מילה לא מוכרת (<UNK>).
- פיצול לסט אימון וסט מבחן. עקב גודלו הקטן של ה-dataset שלנו לא כללנו סט ולידציה.

5.3. מיצוי מאפיינים

שלב זה הוא המרכזי בכל פרויקט מבוסס למידת מכונה. זאת שכן זהו השלב בו אנו מנתחים את המידע הגולמי לצורך מיצוי של מאפיינים איכותיים ובעלי קורלציה גבוהה לסיווג לפי מחלקות. אם כן, שלב זה כלל עבודה רבה ומחקר מעמיק הנוגע למאפיינים איכותיים לצורך המשימה. שלב זה של מיצוי המאפיינים למעשה הינו שלב שנערך לכל אורך הפרויקט, וכלל תוספות רבות לאחר מחשבה וכן ניסוי וטעיה.

לבסוף, לכל טקסט אנו מוציאים את המאפיינים הבאים:

1. **<u>n-grams</u>:** זוהי פרקטיקה מוכרת בתחום עיבוד השפה הטבעית. בשיטה זו מפצלים את הטקסט לפי tokens של n מילים, כך שמתייחסים לכל קבוצת מילים כ-token יחיד. לדוגמא, עבור $n = 3$ והמשפט המוצג נקבל את ה-n-gram הבא:

"The little boy went to school."

["The little boy", "little boy went", "boy went to", "went to school"]

שיטה זו מאפשרת לתפוס קשרים בין מילים אשר מופיעים בתדירות גבוהה יחד. לכל n-gram אנו מוצאים את x_n (היפרפרמטר) ה-tokens הכי נפוצים **בכל סט האימון**, ואז לכל טקסט מחשבים עבור tokens אלו היסטוגרמה של מס' המופעים בטקסט. באופן זה לכל n-gram למעשה מייצרים x_n מאפיינים.

את ההיסטוגרמה ממשקלים באמצעות שיטה שנקראת TF-IDF, לפי הנוסחא הבאה:

$$h[i] = \frac{n_i}{N} \cdot \log\left(\frac{D}{N_i}\right)$$

כאשר:

n_i – מס' המופעים של ה-token ה-i בטקסט.

N – מס' ה-tokens בטקסט.

D – מס' הטקסטים (גודל ה-dataset)

N_i – מס' המופעים של ה-token ה-i בכל סט האימון.

שיטת משקול זו מוכחת מחקרית כיעילה, ונובעת מכך שאיננו מעוניינים לתת משקל גבוה למילים שמופיעות הרבה פעמים בשפה (למשל, ברור שהמילה "the" מופיעה הרבה פעמים בשפה ולכן גם תופיע הרבה פעמים בכל טקסט.

בפרויקט בחרנו להשתמש ב: 1-gram, 2-gram, 3-gram, 4-gram, 5-gram (כל אחד עם x_n מתאים שבחרנו).

2. **<u>אורך מילה ממוצע</u>:** לכל טקסט חישבנו את אורך המילה הממוצע.
3. **<u>אורך טקסט</u>:** לכל טקסט חישבנו את מס' המילים בטקסט.
4. **<u>אורך משפט ממוצע</u>:** לכל טקסט חישבנו את מס' המילים הממוצע במשפט.

5. **<u>היסטוגרמת סימני פיסוק</u>:** בחרנו גם לחשב היסטוגרמה של סימני פיסוק נבחרים עבור כל טקסט. סימני הפיסוק שבחרנו הינם:

- : ,

6. **<u>בניית מודלי שפה וחישוב perplexity</u>:** זהו המאפיין המרכזי שהוספנו, אשר כלל הן את השיטה הכי מתקדמת והן הכי הרבה עבודה. לאחרונה, תחום עיבוד השפה הטבעית הושפע רבות מתחום הלמידה העמוקה.

בניית "מודל שפה" לכל כותב (מחלקה) באמצעות רשת נוירונים. מודל שפה הינו כלי מתמטי הסתברותי, אשר בהינתן רצף מילים $[w_1, \dots, w_i]$, מחשב את ההסתברות למילה הבאה $P(w_{i+1}|w_1, \dots w_i)$. זהו כלי יעיל מאוד לתיאור כללי של כיצד כותב מסוים משתמש בשפה כלשהי. את ההסתברות הנ"ל נהוג למדל באמצעות רשת נוירונים עמוקה, כאשר הקלט שלה הוא רצף מילים/tokens, והפלט הוא וקטור הסתברויות אשר כל איבר בו מסמל את ההסתברות לכל מילה באוצר המילים שתהיה המילה הבאה. ארכיטקטורה אשר נחשבת State of the Art לצורך בניית מודל שפה היא LSTM (Long Short Term Memory) [12]. ארכיטקטורה היא סוג של RNN, ומשתמשת בשערים אשר עוזרים "לזכור" ולקשר בין העבר של אות מסוים למצב הנוכחי.

כדי לחשב מאפיין זה אנו יוצרים מודל שפה נפרד לכל כותב. לאחר מכן, עבור טקסט מסוים, אנו מחשבים מדד הנקרא perplexity אשר מתאר כמה טקסט מסוים הוא "סביר" תחת מודל שפה כלשהו. למעשה, מאפיין זה הוא וקטור של C מספרים (C הוא מספר המחלקות), כך שכל איבר הוא perplexity שחושב על מודל שפה של המחלקה המתאימה. perplexity מחושב כך:

$$perp(x) = e^{-\frac{1}{N}\sum_{i=1}^{N}\log(P(x_i))}$$

כאשר x הוא הטקסט, x_i הוא token בטקסט, ו-N הוא מספר ה-tokens בטקסט. perplexity מינימלי מייצג התאמה טובה למודל שפה.

הרעיון נובע מתוך הבנה שכל כותב משתמש בניסוח, אוצר מילים ושפה מעט שונים. כך, בניית מודל שפה נפרד לכל כותב יכולה לתאר את השימוש שלו בשפה, וחישוב מדד התאמה למודל השפה יכול לתת מידע רב על שייכות טקסט לכותב מסוים.

השלבים שעשינו ליצירת מאפיין זה:

- תחילה ניסינו לייצר מודל שפה בעצמנו. לצורך כך בנינו מספר מודולים ופונקציות:

  מודול Vocabulary - שומר מילון שממיר ממילה למספר עבור כותב מסוים. משמש כדי להפוך טקסט שכתוב כמילים לאות מספרי.

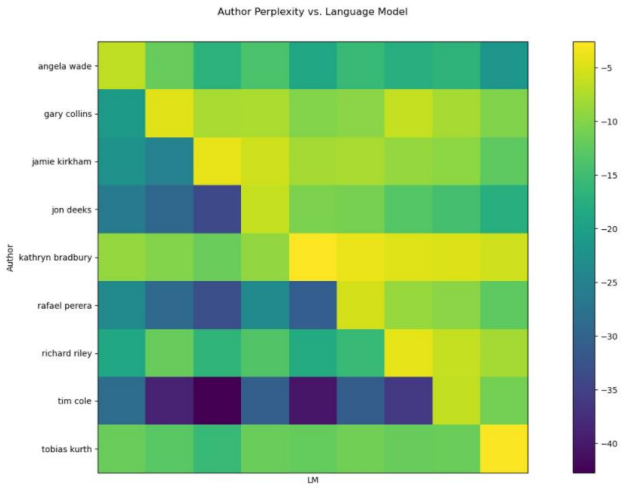
  מודול Dataset - מאפשר לטעון בצורה נוחה את הטקסטים כ-batches של tokens.

  מודול LanguageModelNN – זהו מודל היורש מ-NN של Pytorch, וממש את רשת הנוירונים המשמשת כמודל השפה. השתמשנו בשכבת embedding, שכבת LSTM, ושכבת FC.

  פונקציה לאימון מודל השפה, פונקציה ליצירת טקסט באמצעות מודל השפה (משמשת בעיקר לבדיקה ידנית של המודל), ופונקציה לחישוב perplexity.

  מודול LanguageModel – מודול המאחד את הנ"ל ומאפשר ומאפשר שמירה, טעינה, והוצאת המאפיינים.

- לאחר שאימנו ובדקנו את מודלי השפה הנ״ל, ראינו שאינם מצליחים ללמוד באופן מספק את הטקסטים, לכן עברנו להשתמש במודלי שפה קיימים וחזקים יותר. לאחר מחקר, בחרנו להשתמש במודל GPT2 של Huggingface [13]. זהו מודל אשר אומן מראש על הרבה טקסטים בשפה האנגלית, ולכן כל שנותר לנו לעשות הוא fine-tuning של משקלי המודל לטקסטים שלנו. לצורך כך היה עלינו להמיר את הטקטסים מסט האימון לקובץ טקסט יחיד שכן ככה המודל שהורדנו טוען dataset. כתבנו פונקציה נוספת לחישוב perplexity אשר מתאימה למודל זה. לבסוף אימנו ושמרנו מודל שפה לכל מחלקה באמצעות ה-train set.
- לצורך בדיקה של האם המאפיין החדש הוא מספיק קורלטיבי למחלקה, הצגנו confusion matrix המצגיה את ה-perplexity סט המבחן של כל כותב לפי כל מודלי השפה שאומנו על סט האימון:

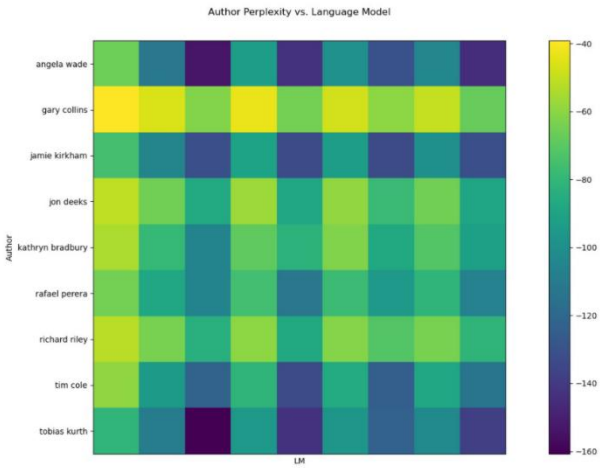


איור 6 - perplexity של טקסטים לפי מודלי שפה (מאמרים-מאמרים)

הצגנו את מינוס ה-perplexity כדי לראות באופן בוהק יותר מספרים נמוכים, שכן perplexity נמוך משמעותו התאמה טובה. השורות הן הטקטסים של כל כותב, והעמודות הן מודלי השפה של כל כותב (באותו סדר).

ניתן לראות שאכן האלכסון הראשי מודגש, כלומר ה-perplexity הנמוך ביותר אכן התקבל עבור מודל השפה המתאים לכותב שכתב את הטקטט.

בדקנו גם את ההתאמה עברו מודלי שפה שאומנו על מאמרים ונבחנו על על ביקורות:



איור 7 - perplexity של טקסטים לפי מודלי שפה (מאמרים-ביקורות)

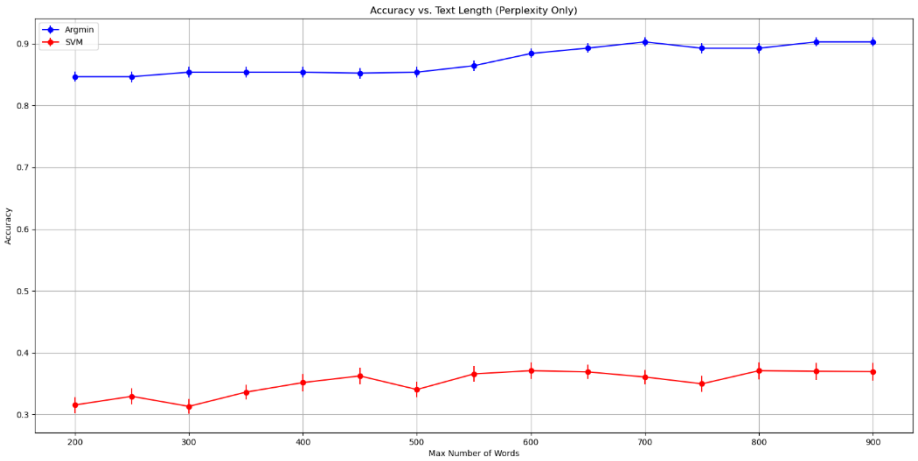
בגרף הנ״ל ניתן לראות התאמה פחות טובה באופן משמעותי, וכי באופן כללי ישנם מודלי שפה אשר מצליחים להשיג באופן עקבי perplexity נמוך יותר עם כל הטקטסים. עובדה זו יכולה להוות בעיה במשימת העל של הפרויקט : אימון על מאמרים וסיווג של ביקורות.

5.4. פעולות נוספות

לאורך הפרויקט ביצענו פעולות נוספות לצורך שיפור המודל, המאפיינים, וה-dataset.

- הוספת class weighting ל-SVM – זאת מכיוון שהמחלקות אינן בהכרח שוות בגודלן, ואיננו רוצים שהמסווג ילמד את ה-prior של המידע אלא שילמד את ההפרדה עצמה.
- ניקיון הדאטה – ראינו שישנן שתי מחלקות ב-dataset אשר מכילות מספר מועט מדי של ביקורות (5 בלבד), לכן הסרנו אותן. הסרנו גם ביקורות בעלות אורך של פחות מ-30 מילים, כי אלו אינן מכילות מידע מספק לצורך הוצאת מאפיינים איכותיים.
- ניסינו גם להפוך את אורכי הטקסטים שאיתם מאמנים את מודלי השפה לאחידים עבור כל הכותבים, כדי שלא יהיה מודל שפה אשר אומן יותר מאחרים. פעולה זו התבררה כלא יעילה ולכן ביטלנו אותה. ההסבר שלנו לכך הוא שפעולה זו מורידה לא מעט מידע, וגם ככה ה-dataset שלנו יחסית קטן.
- הוספת אלגוריתם אשכול למציאת outliers בסט האימון – השתמשנו באלגוריתם OPTICS [14] (הרחבה של DBSCAN). המטרה היא לזהות דגימות שהן שונות מדי מהפילוג האמיתי של המידע ולהסירן מסט האימון כדי שלא ״יבלבלו״ את המסווג. עשינו זאת עקב ההבנה שיש מאמרים רבים שלא בהכרח נכתבו ע״י הכותב לו המחלקה שייכת, ואנו רוצים שמאמרים כאלו לא ישפיעו על אימון המודל.
- שינוי ה-stride של חישוב ה-perplexity – ראינו שפעמים רבות ה-perplexity המתקבל הוא אינסופי (או ממש אינסוף). תוצאה זו כמובן פוגעת רבות במאפיינים וביכולת הסיווג. לכן הקטנו את ה-stride, פעולה אשר מקטינה את ה-perplexity המחושב. פעולה זו אכן עזרה.

- שינוי גרעין מסווג ה-SVM. לאחר מעט ניסוי וטעינה בחרנו ב-sigmoid.
- ביצענו בדיקה של סיווג רק באמצעות ה-perplexity בשתי דרכים. הדרך הראשונה היא שימוש בוקטור ה-perplexity כוקטור המאפיינים, וביצוע סיווג באמצעות SVM. הדרך השנייה היא ביצוע argmin על וקטור ה-perplexity להשגת המחלקה עם ה-perplexity המינימלי (שיטה זו כמובן חוסכת את הצורך באימון).
את שתי השיטות בחנו על ה-toy problem כתלות באורכי הטקטסים:



איור 8 - אחוז דיוק כתלות באורך טקסט לפי perplexity בלבד - toy problem

בגרף הנ״ל ציר x הוא אורכי הטקסטים, וציר y הינו יחס ההצלחה של הסיווג על סט המבחן. באדום מוצג הדיוק באמצעות SVM (השיטה הראשונה), ובכחול הדיוק האמצעות argmin (השיטה השנייה). כל נקודה היא אחוז הדיוק הממוצע עבור מספר סטי בוחן שונים. לכל נקודה למעשה ביצענו bootstrapping, כלומר חישבנו גם את סטיית התקן של אחוז הדיוק כדי להציג confidence interval של 95%. בנוסף, בדקנו מה אורך הביקורת הממוצע (וסטיית תקן), והגדרנו זאת כנקודת עבודה. עשינו את הבדיקה הנ״ל סביב נקודת העבודה זו כדי לדמות איך השיטה תעבוד על טקטסים קצרים.
המסקנות שלנו מבדיקה זו היו ששימוש ב-argmin אמור להיות יעיל יותר מוקטור perplexity.

- הבעיה עם שיטת ה-argmin הנ״ל היא שלא ניתן להכניס אותה כמאפיין באופן מוצלח לוקטור מאפיינים המשמש למודל סיווג. זאת למעשה מכיוון שהשיטה מתארת סיווג רק באמצעות לקיחת argmin על וקטור ה-perplexity, אך כדי להכניס זאת כמאפיין ניתן רק להכניס את מס׳ המחלקה. שיטה זו היא בעייתית מהסיבות הבאות:
  - גורמת לאיבוד מידע רב – מכניסים רק את המחלקה המינימלית ללא כל מדד על מה ה-perplexity עצמו וללא כל מדד על מחלקות אחרות (למשל אם קיימת מחלקה אחרת עם perplexity קרוב מאוד למינימלית).
  - אין משמעות מספרית למספר המחלקה המינימלי – המספר מציין מחלקה ואין למספר עצמו משמעות אמיתית. מרבית המסווגים מושפעים מכך מאוד.

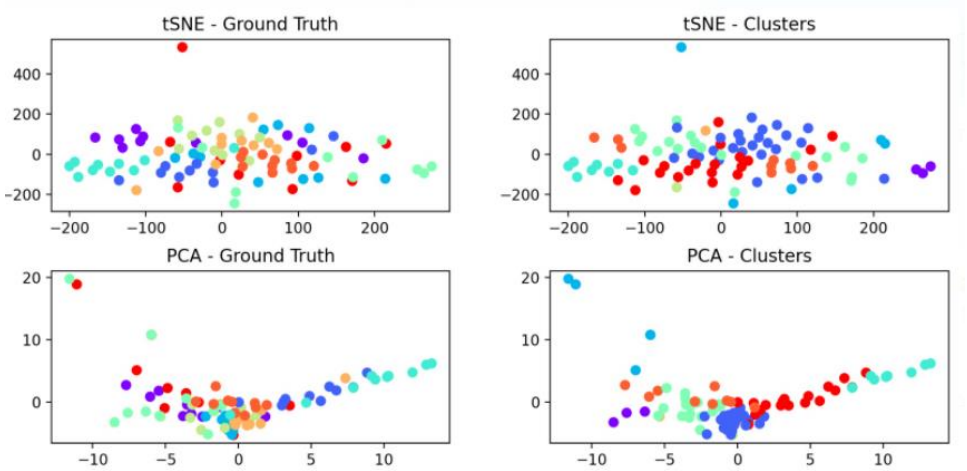
לאור הסיבות הנ״ל, ולאחר בדיקות נוספות שעשינו (אשר כללו בין היתר ניסיון של סיווג באמצעות שיטה זו), בחרנו שלא להשתמש בה ובמקום זאת להשתמש בוקטור ה-perplexity כמאפיין.

5.5. אשכול

לאורך הפרויקט בחנו שיטות אשכול שונות על מטריצות המאפיינים. פעולת האשכול היא סוג של למידה לא מפוקחת, בה אנו משתמשים בוקטורי המאפיינים כדי לאשכל את הדגימות למחלקות שונות. עשינו זאת מכמה טעמים:

- מציאת חריגים ב-dataset (כפי שצוין בפרק 4).
- ויזואליזציה של הדגימות.
- בחינה ויזואלית של טיב המאפיינים – האם המאפיינים שהגדרנו מביאים לאשכול אשר מתאים למחלקות.
- Manifold learning – למידת המרחב. מכיוון שמדובר בבעית cross-domain, סט האימון וסט המבחן לקוחים למעשה ממרחבי דגימות שונים. לכן שלב חשוב במעבר בין התחומים הוא למידת המרחב הפורש כל אחד מהם. ניתן לעשות זאת באמצעות שיטות שונות של הורדת מידמיות, אשר נקראות באופן כללי Manifold learning.

תחילה נציג תוצאות של בדיקה שעשינו כשלב ביניים במהלך הפרויקט, אשר בוצעה על מרחב המאפיינים הלא סופי (כלל רק היסטוגרמות n-grams):

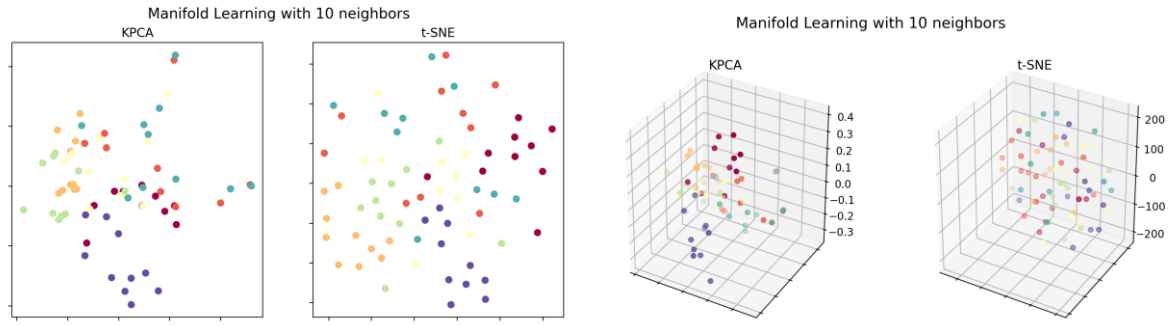


איור 9 - אשכול והורדת מימדיות בשלב ביניים בפרויקט

הגרפים מציגים את פיזור הדגימות (מאמרים) במרחב דו-מימדי באמצעות שתי שיטות – tSNE [15] בשורה העליונה ו-PCA בתחתונה. בצד שמאל הצבעים מסמנים את המחלקות האמיתיות של כל דגימה, ובצד ימין הצבעים מסמנים את השיוך לאשכולים לפי אלגוריתם k-means אשר הופעל **על מרחב המאפיינים המקורי**. ניתן לראות בשיטת tSNE שאכן האשכול בוצע באופן יחסית תואם למחלקות המקוריות, וכי חלק מהמחלקות אכן מתאשכלות יחד.

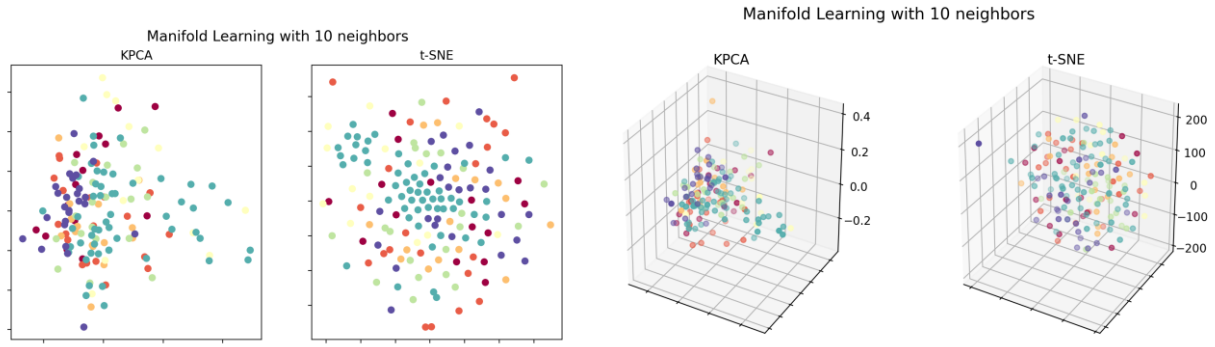
בנוסף, בחנו שיטות Manifold learning נוספות (MDS, LLE ועוד), אך לבסוף בחרנו להציג את KPCA עם גרעין sigmoid ו-tSNE. נציג את התוצאות כפי שהיו בסוף הפרויקט, כלומר אשר כוללות את כל המאפיינים מסעיף 5.3. את הורדת המימדיות ביצענו הן למאמרים בנפרד והן לביקורות בנפרד, והצגנו במרחב דו-מימדי ובחרב תלת-מימדי:

- מאמרים:



איור 10 - Manifold learning (מאמרים)

- ביקורות:



איור 11 - Manifold learning (ביקורות)

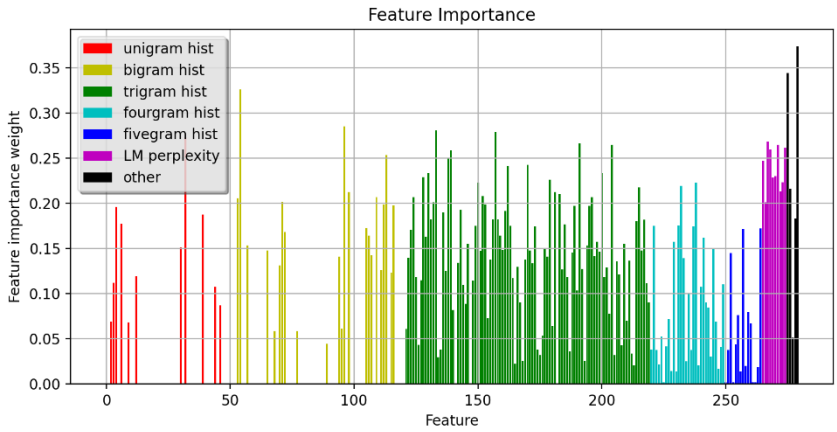
נציין שהגרפים הנ״ל אינם כוללים אשכול, כלומר הצבעים מסמנים את המחלקות האמיתיות של הדגימות.
ניתן לראות שבשני המרחבים ניתן להשיג הפרדה טובה יחסית למחלקות באמצעות הורדת המימדיות. הדבר יכול להעיד על איכותם של המאפיינים הנבחרים.
בנוסף, ניתן להשתמש בשיטות הורדת המימדיות (לא בהכרח 2 ו-3 מימדים) כחלק מה-pipeline של הפרויקט כך שמודל הסיווג (ה-SVM) ילמד על המרחב הנמוך ולא על מרחב המאפיינים המקורי.

5.6. בחירת מאפיינים

תהליך בחירת המאפיינים (feature selection) הינו שלב מאוד חשוב בפרויקט זה. זאת מכיוון שמרחב המאפיינים הוא מאוד גדול (כמה מאות), שכן הוא כולל מספר היסטוגרמות וכן מאפיינים נוספים. תהליך בחירת המאפיינים יכול לעזור להתעלם ממאפיינים אשר אינם מועילים להפרדת המחלקות, ואף כאלו אשר פוגעים בהפרדה ע"י קורלציה הפוכה. בחרנו לעבוד עם אלגוריתם reliefF [16] לבחירת המאפיינים.

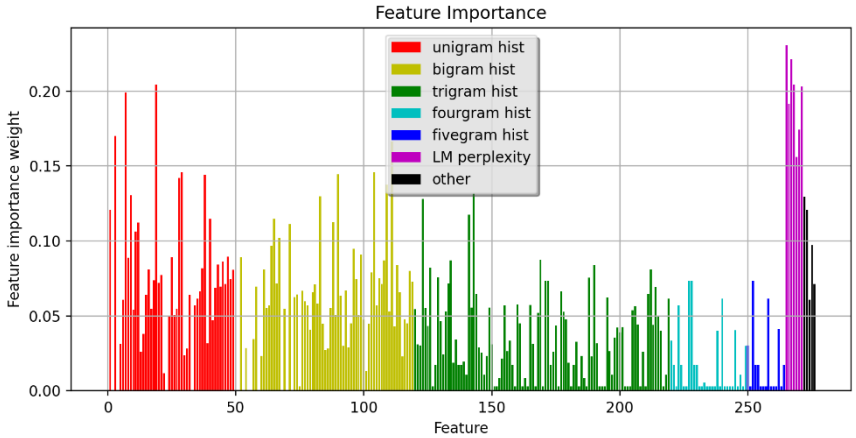
לצורך כך הפרדנו את בעיית העל שלנו למספר תתי בעיות שאינן cross-domain: toy problem (פרק 3), אימון והסקה על מאמרים, אימון והסקה על ביקורות. כך אנו מבטלים את הקושי הרב אשר נובע מבעיית ה-cross domain, ומתמקדים רק באיכות המאפיינים לכל מרחב בנפרד. להלן התוצאות:

- toy problem:



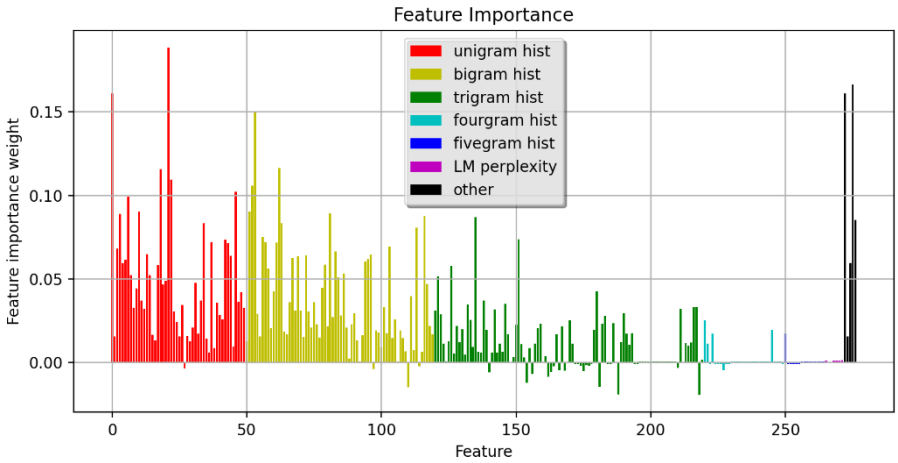
איור 12 - חשיבות מאפיינים (toy problem)

- מאמרים:



איור 13 - חשיבות מאפיניים (מאמרים)

- ביקורות:



איור 14 - חשיבות מאפיינים (ביקורות)

את הגרפים הנ״ל הצגנו עם פיצול של צבעים לפי סוג המאפיין (מקרא בגרף), כאשר הצבע השחור כולל את כל המאפיינים הסקלריים (אורך טקטט, אורך משפט, וכו׳).

ניתן להסיק מספר מסקנות מהגרפים הנ״ל:

- 1-2-3-grams הינם מאפיינים בעלי משמעות למשימה, אך יש לשים לב גם שחלק מאיברי ההיסטוגרמות אינם. לכן תהליך בחירת המאפיינים יכול לעזור לבחור את האיברים החשובים מתוך כל היסטוגרמה.
- 4-5-grams נראים כמו מאפיינים יחסית לא יעילים.
- perplexity לפי מודלי שפה נראה כמו מאפיין יעיל, זאת מלבד עבור ביקורות. עובדה זו יכולה להיות מעט מדאיגה עבור בעיית ה-cross-domain. יתכן שהיא נובעת מכך שביקורות הינם טקסטים מאוד קצרים, ולכן מודלי שפה לפיהם אינם מספיק יעילים.
- המאפיינים הנוספים הם אמנם סקלריים, פשוטים וקטנים יחסית להיסטוגרמות ול-perplexity, אך הם נראים כיעילים עבור המשימה.

לבסוף, הוספנו את בחירת המאפיינים באמצעות אלגוריתם reliefF כחלק מה-pipeline של המודל שלנו. את מספר המאפיינים בחרנו כיחס מתוך מספר המאפיינים המקורי.

6. תוצאות

את מודל הסיווג הרצנו במספר קונפיגורציות של המידע:

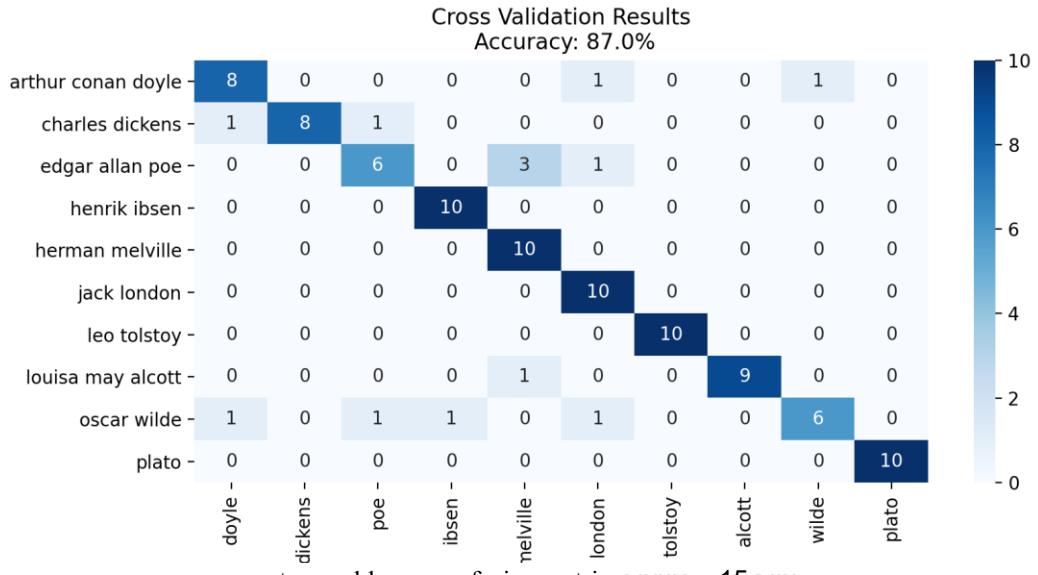
1. toy problem – אותה בעיה פשוטה שהוצגה בפרק 3. שמרנו על בעיה זו כרפרנס ל-dataset פשוט יחסית אשר אנו יודעים שניתן להגיע עליו לאחוזי הצלחה גבוהים, זאת לשם בדיקת המודל והמאפיינים.
2. מאמרים בלבד – אימון והסקה על מאמרים. זהו סט האימון בבעיית העל. מידע זה מייצג את יכולת המסווג ללמוד את מרחב המאמרים באופן יעיל. זהו שלב חשוב בראייה קדימה לבעיית העל של למידה על מאמרים והסקה של ביקורות.
3. ביקורות בלבד – אימון והסקה על ביקורות. זהו סט המבחן בבעיית העל. מידע זה מייצג את פיזור סט המבחן במרחב המאפיינים של סט האימון, ויכול להעיד על היכולת לפתור את בעיית העל. זוהי קונפיגורציה מאתגרת בפני עצמה שכן מדובר בטקסטים קצרים ובעלי מאפיינים דומים מאוד (אורכים דומים, עולם תוכן דומה, סגנון כתיבה אשר מתאים לביקורת). הצלחה בקונפיגורציה זו היא צעד משמעותי בדרך לפתרון בעיית העל.
4. אימון על מאמרים והסקה על ביקורות – זוהי בעיית העל של הפרויקט, אשר טרם נחקרה לעומקה. במהלך העבודה על הפרויקט נחשפנו לכך שחלק זה הינו בעל מאפיינים והיקף עבודה של פרויקט נוסף, אשר עתיד להתבצע במסגרת פרויקט המשך, זאת בתיאום והמלצת מנחה הפרויקט. עם זאת, נציג כאן את תוצאות קונפיגורציה זו, לשם השלמות. בפרק 8 נציג את הצעותינו להמשך העבודה על חלק זה.

ב-3 הקונפיגורציות הראשונות השתמשנו כדי לבטל את בעיית ה-cross-domain אשר משרה אתגרים רבים נוספים.

<u>התוצאות:</u>

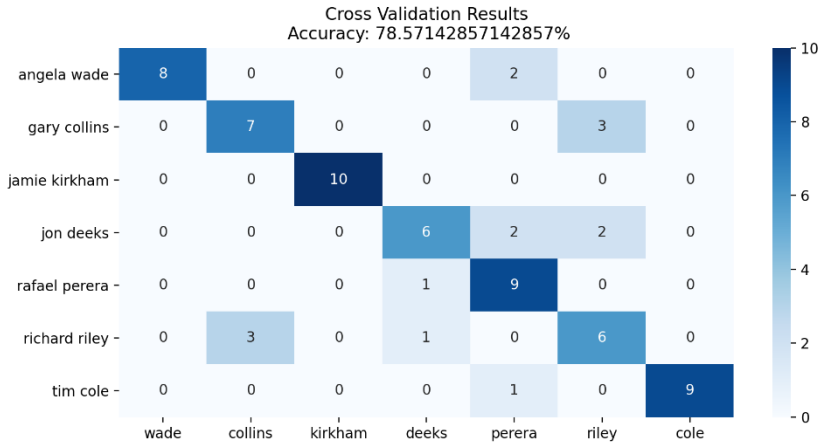
את התוצאות נציג כאחוזי הצלחה מתוך סט המבחן תוך ביצוע cross-validation. נציג גם confusion matrix של הסיווג.

1. toy problem:



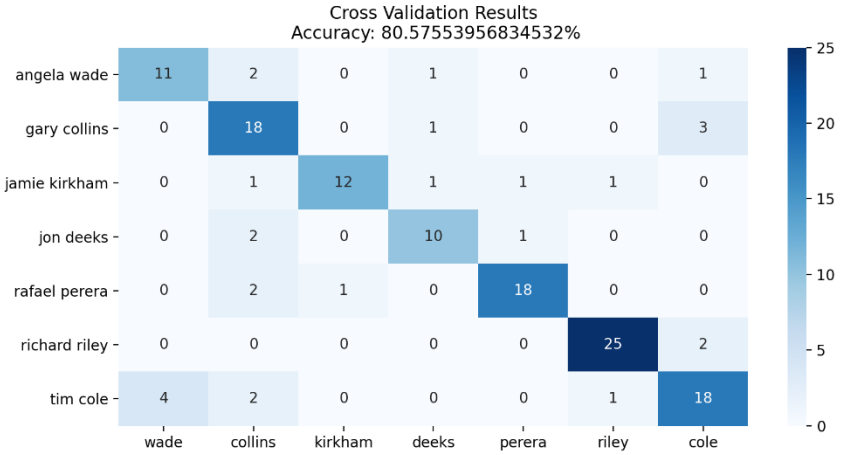
איור 15 - תוצאות toy problem - confusion matrix

2. מאמרים בלבד:



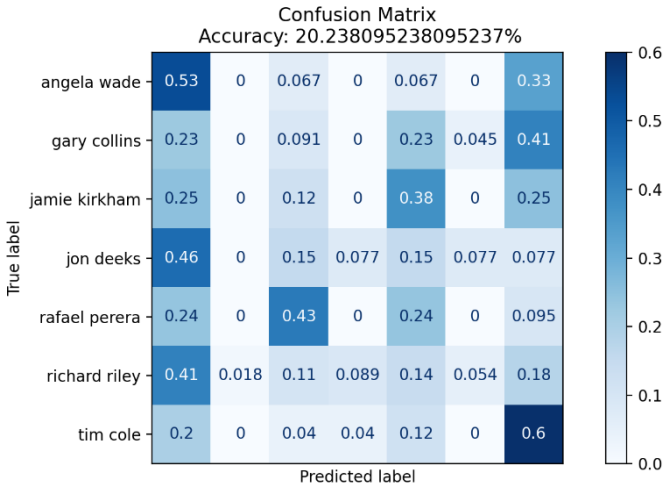
איור 16 - תוצאות confusion matrix - מאמרים בלבד

3. ביקורות בלבד:



איור 17 - תוצאות confusion matrix - ביקורות בלבד

4. אימון על מאמרים והסקה על ביקורות:



איור 18 - תוצאות confusion matrix - מאמרים-ביקורות

נציין שבקונפיגרציה 4 לא ניתן לבצע cross validation, שכן בקונפיגורציה זו סט האימון וסט המבחן הם נפרדים ומוגדרים מראש.

ניתן לראות שבכל הקונפיגרציות שאינן cross-domain הגענו לתוצאות מרשימות למדי. תוצאה זו מראה על יציבות המודל והמאפיינים ויכולת ההכללה לשימוש במרחבי למידה שונים.

בקונפיגורציה 4 התוצאות אינן מספקות, אך לא ביצענו בפרויקט זה עבודה על בעיית cross-domain. נציין גם שבקונפיגורציה זו הגענו באחת הריצות לכ-30% הצלחה, אך בתוצאה זו המודל סיווג כמעט את כל הטקסטים לשתי מחלקות בלבד, ואחוז הדיוק הגבוה יותר הושג עקב כך שאלו מחלקות עם הרבה מאוד דגימות בסט המבחן. לכן אנו מאמינים שהתוצאה שהצגנו באיור 18 הינה מוצלחת יותר למרות שבעלת אחוז דיוק נמוך יותר. התוצאות הנ״ל מהוות למעשה את נקודת ההתחלה לעבודה עתידית אפשרית.

7. מסקנות

1. הגענו לתוצאות מרשימות למדי בתחום שנחשב מאתגר מאוד. כמו כן, הראנו באמצעות 3 סוגי מידע שונים כי המודל שלנו הוא בר-הכללה לעולמות תוכן שונים בעולם ה-Authorship Attribution. בפרט, המודל מתאים לעולם התוכן של מאמרים אקדמיים וביקורות על מאמרים, ומהווה נקודת התחלה טובה לבעיית העל של למידה על מאמרים והסקה על ביקורות.
2. שיטות קלאסיות של למידת מכונה אינן מספיק אפקטיביות בבעיית cross-domain, אך יכולות להוות את הבסיס להרחבת המודל לבעיות מסוג זה. כדי להתאים את המודל לבעיות cross-domain, יש לבצע שלבים נוספים של domain adaptation, כלומר למידת "פונקציית מעבר" בין מרחב המאפיינים של סט האימון לבין מרחב המאפיינים של סט המבחן.
3. dataset – איכות וכמות המידע מאוד משמעותיים. מבחינת כמות, ככל שיש יותר מידע כך קל יותר הן לאמן את המודל בצורה מוכללת והן לבחון את הצלחתו בצורה משקפת. מבחינת איכות, ראינו בפרק 4 כי המידע איתו עוסק פרויקט זה הינו בעייתי במהות משלל סיבות, ביניהן: מאמרים לרוב נכתבים ע"י מספר כותבים ולא כותב יחיד, ביקורות הינם לרוב לטקסט קצרים ולעיתים אף חסרי תוכן אמיתי, פורמטים שונים אשר מקשים על סידור המידע. ראינו שיש קושי רב במציאת מידע איכותי בעולם זה של ביקורות ומאמרים אקדמיים, וכן לא מצאנו דרך יעילה לאוטומציה של תהליך בניית ה-dataset. בנוסף, מבחינת סט המבחן (ביקורות) מצאנו כי לא קיימים מספיק מאגרים אשר מפרסמים ביקורות באופן לא אנונימי, ולכן קיים קושי במציאת סט מבחן מתויג.
4. מצאנו כי אורך הטקסט יכול להשפיע רבות על יכולת ההצלחה של מודל הסיווג (סעיף 3.2). לעובדה זו יש משמעות רבה לפרויקט זה שכן מדובר בסט מבחן עם אורכי טקסטים קצרים מאוד.
5. כאשר עוסקים בבעיית cross-domain אשר הינה מורכבת מאוד עקב הבדלי התחומים בין סט האימון לסט המבחן, נוח מאוד להתחיל מבעיות single-domain נפרדות עבור סט האימון בנפרד וסט המבחן בנפרד. באופן זה ניתן ללמוד באופן בלתי תלוי על יכולת המודל להצליח בשני התחומים, ורק לאחר מכן לגשת לפתרון בעיית ההמרה והתאמה ביניהם.
6. מבחינת מאפיינים (סעיף 5.3):
   - מצאנו כי 1-2-3-grams הינם מאפיינים שימושיים בעולם עיבוד השפה הטבעית, ובפרט עבור בעיית Authorship Attribution ועבור בעיית סיווג מאמרים אקדמיים וביקורות.
   - מצאנו כי 4-5-grams הינם מאפיינים פחות שימושיים, אם כי כן יכולים לעזור לעיתים.
   - שימושי מאוד לגוון בסוגי המאפיינים, כלומר שימוש במאפיינים מסוג תוכן, מסוג ניסוח, ומסוג סגנון כתיבה.
   - למידת מודלי שפה לכותבים וחישוב perplexity היא שיטה יעילה מאוד להוצאת מאפיינים בעלי קורלציה גבוהה לזהות כותב הטקסט. באופן כללי, שיטות מבוססת למידה עמוקה הוכחו מחקרית כיעילות מאוד בתחום עיבוד השפה הטבעית.
7. שלב בחירת המאפיינים (סעיף 5.6) הינו שלב חשוב מאוד בפרויקט מסוג זה, כאשר למעשה מדובר במרחב מאפיינים מאוד גדול אשר קשה לדעת מראש אילו מאפיינים יהיו מועילים ואילו יפגעו בסיווג.

8. ביצוע אשכול והורדת מימדיות באמצעות שיטות למידה לא מפוקחת במהלך העבודה (סעיף 5.5) היא פרקטיקה שימושית מאוד, ועוזרת ללמוד על פיזור המידע שעובדים איתו ולהסיק על הצלחת המודל ואיכותם של המאפיינים. שלב זה יכול גם להיות יעיל כחלק מה-pipeline של מודל הסיווג.
9. בפרויקט עם ריצות ארוכות (בפרויקט זה מדובר בעיקר בלמידת מודלי שפה ובחישוב perplexity לפי מודלי השפה) שימושי מאוד לשמור את מטריצות המאפיינים בכל ריצה. שמירת המאפיינים מאפשרת ריצות מהירות יותר וכתיבת קוד יעיל יותר, שכן אין צורך בחישוב המאפיינים בכל ריצה מחדש.

8. עבודה עתידית

1. כפי שראינו, ל-dataset ישנה השפעה רבה על הצלחת המודל והיכולת לבחונו. עם זאת, ראינו גם שקיים קושי רב במציאת מידע איכותי מהסוג המתאים. אם כך, לצורך שיפור המודל יש להרחיב את ה-dataset הקיים ע"י הגדלת הן סט האימון והן סט המבחן. בנוסף, אנו ממליצים על חיפוש ובניית dataset נוסף ומתחום אקדמי אחר (לא רפואי) אשר יאפשר וידוא שהמודל הינו בעל יכולת הכללה. לבסוף, כדי להפוך את תהליך בניית ה-dataset ליעיל וסקיילבילי, יש לבצע אוטומציה של תהליך בנייתו.
2. לאחר שראינו שניתן להגיע עם מודל הסיווג שלנו לתוצאות טובות בבעיות single-domain, השלב המרכזי הבא הוא מעבר לבעיית cross-domain. לצורך כך יש לבצע פעולה שנקרא domain adaptation, כלומר למידת "פונקציית מעבר" בין מרחב המאפיינים של סט האימון לבין מרחב המאפיינים של סט המבחן. אנו מאמינים שביצוע שלב זה באופן איכותי יוביל להצלחה ופריצת דרך בקונפיגורציית המידע המרכזית של למידה על מאמרים והסקה על ביקורות.
3. ניתן לבחון הוספת מאפיינים נוספים, הן מעולם ה-n-grams, הן מעולם הלמידה העמוקה, והן מעולם המאפיינים הסגנוניים.
4. במעבר לבעיית cross-domain, יש לבחון אילו מהמאפיינים שמצאנו אכן הינם שימושיים לבעיה מסוג זה. זאת מכיוון שישנם מאפיינים שיכולים להיות מועילים בבעיית single-domain, אך כלל לא מתאימים לבעיית cross-domain. למשל, יתכן שאין קורלציה בין אורך הטקטס של מאמר אקדמי לבין אורכי הביקורות שכותב אותו אדם. דוגמא נוספת נוגעת ל-n-grams, כי עקב השוני הרב בין סגנונות הכתיבה של ביקורות וסגנונות הכתיבה של מאמרים יכול להיות שמאפיין מסוג זה יפגע.
5. ניתן להוסיף בחינה של top 3 classification כדי לבדוק את אופן הצלחת המודל גם כאשר אינו מדייק.
6. ביצוע מחקר על מיטיגציה לבעיה – כאשר יוכח באמצעות מודל הסיווג שניתן לבצע דה-אנונימיזציה לכותבי ביקורות, יש לבחון פתרונות אפשריים אשר יכולים להשיב את האנונימיות. המיטיגציה צריכה לעבוד על הטקטסים הגולמיים עצמם ולא על מרחב המאפיינים, כדי לאפשר לכותבי ביקורות לבצע זאת באופן יעיל. הפתרון גם צריך לשמור על תוכן הביקורת, ועל מובנותה. את איכות הפתרונות המוצעים ניתן לבחון באמצעות מציאת המאפיינים לפי המודל הקיים והפעלת המסווג על הטקטסים החדשים, אם המסווג יכשל – סימן שהפתרון מוצלח.

9. סיכום

בפרויקט זה עסקנו בתת-תחום של למידה עמוקה ועיבוד שפה טבעית, אשר טרם נחקר בעבר. מטרת העל של הפרויקט הינה להוכיח כי ניתן לבצע דה-אנונימיזציה של ביקורות של מאמרים, זאת באמצעות כלי למידת מכונה הלומדים באמצעות מאמרים אקדמיים של הכותבים הפוטנציאלים.

הפרויקט עסק בפתרון הבעיה הנ״ל באופנים הבאים: בניית מאגרי מידע המתאימים לבעיה, מחקר אודות מיצוי מאפיינים המתאימים לבעיה (כולל מאפיין חדש מבוסס ציון לפי מודלי שפה שבנינו), מחקר מעמיק נוסף אודות שיטות נוספות לשיפור המודל (אלגוריתמי אשכול והורדת מימדיות, בחירת מאפיינים, עיבוד מקדים של המידע ועוד), וכן כתיבת קוד מודולרי המתאים להמשך עבודה על הבעיה.

במסגרת בחינת המודל הגענו לתוצאות טובות מאוד על בעיות זיהוי כותבים שונות, אשר מעידות על יכולת המודל והמאפיינים להגיע לתוצאות גבוהות בבעיית העל. ראינו שהמודל שבנינו מתאים לעולמות תוכן שונים: מאמרים אקדמיים, ביקורות על מאמרים אקדמיים, וכן ספרים רגילים. התוצאות גם מעידות על יכולת המודל להתמודד עם תחומים מאתגרים הכוללים טקסטים קצרים ולא מגוונים כגון ביקורות.

במהלך העבודה על הפרויקט נחשפנו לתחומי מחקר מעניינים שלא הכרנו לפני כן וסביבות עבודה חדשות. למדנו רבות על עולם עיבוד השפה הטבעית ולמידה עמוקה בהקשר תחום זה, רכשנו יכולות פרקטיות ורלוונטיות להמשך דרכנו כמהנדסים וחוקרים.

נספח א' – הסבר על הקוד

נספח זה מציג תיאור של מבנה הפרויקט, והסבר קצר על כל מודול וסקריפט בו.

הקוד בכללותו נמצא ב-Github repository הבא:

https://github.com/roy-hachnochi/ReviewerAttribution

ניתן להוריד או לעשות clone לפרויקט לצורך הפעלתו או המשך עבודה.

datasets:

תיקייה זו כוללת את כל קבצי המידע של הפרויקט (כולל כאלו שבפועל לא היו בשימוש), וכן סקריפטים להורדת, סידור, והכנת המידע לצורך העבודה.

תיקיות dataset_bmj, dataset_f1000, dataset_nips, toy_data הן התיקיות אשר מכילות את המידע עצמו, ומסודרות באופן אשר מאפשר הכנסה של המידע למודל הפרויקט. התיקיות נוצרו באמצעות הסקריפטים הבאים, אשר עם מעט שינויים יכולים גם להתאים להורדת וסידור מידע נוסף:

- **DownloadDataBmj, DownloadDataF1000, DownloadDataNips** – סקריפטים להורדת ביקורות מ-3 המקורות השונים. הסקריפטים מבצעים ריצה על רשימת URLs, בודקים אם קיימת ביקורות בדף, ומורידים ושומרים את הביקורת.
- **ParseReviewsBmj, ParseReviewsF1000, ParseReviewsNips** – סקריפטים להפרדת הביקורות והמחלקות (שמות הכותבים) מתוך הטקסטים שהורדו לכל מקור. הסקריפטים מבצעים את ההפרדה לפי tokens שהגדרנו לאחר בחינת הטקסטים והבנה של איזה tokens מסמנים סוף והתחלה של הביקורת עצמה. סקריפטים אלו שומרים את הביקורות וקובץ csv המתאים בין שמות קבצים לשמות הכותבים המתאימים.
- **OrganizeLabelsBmj, OrganizeLabelsF1000** – סקריפטים לסידור שמות הכותבים ואיחוד של שמות דומים. ראינו שישנם מקרים בהם כותב מסוים כתוב בשני שמות מעט שונים (למשל שם אמצעי, או כינוי, או שגיאת כתיב). סקריפטים אלו מריצים אלגוריתם שכתבנו אשר מוצא נקודות דמיון בין השמות ומאחד אותם למחלקה אחת אם הדמיון עבר סף מסוים שהגדרנו.
- **GetGoodLabels** – סקריפט לצמצום מחלקות שאינן טובות מספיק. תחילה הסקריפט רץ על כל טקסטים ומשאיר רק את אלו שעברו סף מוגדר של כמות המילים בטקסט (הגדרנו 30), לאחר מכן עבור הטקטסים שנשארו הסקריפט משאיר רק מחלקות אשר כוללות כמות טקסטים שעוברת סף מוגדרת (הגדרנו 7).
- **pdf2txt** – סקריפט אשר מפעיל מודול שנקרא pdf2txt אשר ממיר בין קבצי PDF לקבצי txt. מודול זה יכול להיות שימושי להפיכת קבצי מאמרים לקבצי טקסט. בפועל לא השתמשנו בסקריפט זה.

הקוד:

- **Main** – הסקריפט הראשי של הפרויקט אשר מממש את מודל הסיווג. כולל שימוש בכל המודולים של הפרויקט וחישוב התוצאות. ניתן לקבוע היפרפרמטרים בראש הקובץ וכן ניתן לשנות קונפיגורציות מידע ומהיכן הוא נטען.
- **Preprocess** – כולל את פונקציות העיבוד המקדים. כולל פונקציות לטעינת הטקסטים, פונקציה ל-tokenizing, ופונקציה לביצוע פיצול לסטי אימון-מבחן.
- **FeatureExtracor** – כולל את מודול מיצוי המאפיינים. כולל את המודול FeatureExtractor אשר יודע לקבל dataset, ללמוד מאפיינים באמצעות fit, ולהוציא מטריצת מאפיינים באמצעות transform. מממש גם פונקציות ומודולי עזר אשר משמשים לחישוב המאפיינים. בקובץ זה יש גם פונקציות out of date אשר שימשו להצגת המאפיינים (איור 2).
- **FeatureSelection** – סקריפט זה אינו מממש מודול או פונקציות לקוד המרכזי, אלא משמש לבדיקה ומחקר של המאפיינים ומודול בחירת המאפיינים. באמצעות סקריפט זה יוצרו איורים (איור 12, איור 13, איור 14).
- **Clustering -** סקריפט זה אינו מממש מודול או פונקציות לקוד המרכזי, אלא משמש לבדיקה ומחקר של שיטות האשכול והורדת המימדיות. באמצעות סקריפט זה יוצרו איור 10 ואיור 11).
- **PrepareDataForLM** – סקריפט אשר משמש ליצירת ושמירת טקסטים האימון באופן המתאים לאימון מודל GPT2 (כלומר כקובץ יחיד עם שרשור כל הטקטסים).
- **LanguageModels** – סקריפט אשר משמש לאימון מודלי השפה באמצעות מודל GPT2 של Huggingface. יש להריץ סקריפט זה ליצירת מודלי השפה <u>לפני</u> שימוש ב-FeatureExtractor. בקובץ זה ממומשות גם פונקציות לטעינת מודלי שפה ולחישוב perplexity.
- **LMtesting** – סקריפט אשר משמש ליצירת ה-confusion matrix של ה-perplexity (איור 6, איור 7).
- **PerplexityClassification** – סקריפט אשר משמש ליצירת גרפי הסיווג באמצעות perplexity בלבד (איור 8).
- **NumWordsStats** – סקריפט אשר מחשב את ממוצע, חציון, וסטיית התקן של אורכי הטקסטים.
- **LanguageModels_Old** – כולל את מודול מודלי השפה <u>הישן</u>.
- **ToyProblem** – סקריפט out of date אשר שימש לביצוע ה-toy problem בתחילת הפרויקט (פרק 3).

results:

תיקייה זו כוללת וקטורי ומטריצות תוצאות ששמרנו לאורך הריצות בפרויקט. שמות תתי התיקיות תואמים את הסקריפט שמייצר את הקבצים אותם הן שומרות. התיקיות:

- **LMtesting** – שומר את ה-confusion matrix של ה-perplexity (איור 6, איור 7).
- **PerplexityClassification** – שומר את גרפי הסיווג באמצעות perplexity בלבד (איור 8).
- **Main** – שומר את מטריצות המאפיינים והמחלקות של הריצות הכלליות עבור כל קונפיגורציית מידע.

רשימת מקורות

[1] C. Manning, "Stanford, CS224n: Natural Language Processing with Deep Learning," [Online]. Available: http://web.stanford.edu/class/cs224n/ .

[2] M. K. e. al., "Overview of the Author Identification Task at PAN 2018 - Cross-domain Authorship Attribution and Style Change Detection," in *PAN*, Avignon, 2018.

[3] M. K. e. al., "Overview of the Cross-Domain Authorship Attribution Task at PAN 2019," in *PAN*, Lugano, 2019.

[4] Y. Sari, "Neural and Non-neural Approaches to Authorship Attribution - A Thesis submitted to the University of Sheffield for the degree of Doctor of Philosophy in the Faculty of Engineering," Department of Computer Science The University of Sheffield, 2018.

[5] J. S. S. A. Moshe Koppel, "Computational Methods Authorship Attribution," *Journal of the American Society for Information Science and Technology,* vol. 60, no. 1, pp. 9-26, 2009.

[6] E. Stamatatos, "A Survey of Modern Authorship Attribution Methods," *Journal of the American Society for Information Science and Technology,* vol. 60, no. 3, pp. 538-556, 2009.

[7] A. Karpathy, "The Unreasonable Effectiveness of Recurrent Neural Networks," 21 May 2015. [Online]. Available: http://karpathy.github.io/2015/05/21/rnn-effectiveness/.

[8] T. Tran, "Text Generation With Pytorch," 08 Feb 2019. [Online]. Available: https://machinetalk.org/2019/02/08/text-generation-with-pytorch/.

[9] "Project Guttenberg," [Online]. Available: https://www.gutenberg.org/.

[10] "The British Medical Journal," BMJ Publishing Group Ltd, [Online]. Available: https://www.bmj.com/.

[11] S. E. L. a. E. K. Bird, *Natural Language Processing with Python,* O'Reilly Media Inc, 2009.

[12] S. a. S. J. Hochreiter, "Long Short-Term Memory," *MIT Press,* vol. 9, no. 0899-7667, 1997.

[13] T. W. e. al., "HuggingFace's Transformers: State-of-the-art Natural Language Processing," *ArXiv,* vol. abs/1910.03771, 2019.

[14] M. Ankerst, M. M. Breunig, H.-P. Kriegel and J. Sander, "OPTICS: Ordering Points To Identify the Clustering Structure," *ACM SIGMOD international conference on Management of data. ACM Press,* pp. 49-60, 1999.

[15] L. van der Maaten and G. Hinton, "Visualizing Data Using t-SNE," *Journal of Machine Learning Research,* no. 9, p. 2579–2605, Nov 2008.

[16] I. e. a. Kononenko, "Overcoming the myopia of inductive learning algorithms with RELIEFF," *Applied Intelligence,* vol. 1, no. 7, pp. 39-55, 1997.