ניהוי נושאים בפרוטוקולי ישיבות פרלמנטריות בסברית

ניצן ברזילי

קורס "עיבוד שפה טבעית בעברית"

האוניברסיטה העברית

2022

תוכן עניינים

מבוא והגדרת הבעיה

רקע

בשנת 2017 התבצעה לראשונה הנגשה לציבור של מידע רב הנוגע לפעילות הכנסת, ובכלל כך הפרוטוקולים המלאים של ועדות הכנסת ומליאת הכנסת. המידע שפורסם (וממשיך להתפרסם באופן עתי) נוגע לפעילות הכנסת משנת 2015 ועד היום. אמנם המידע מפורסם באופן פומבי, אך הוא אינו נגיש לשימוש ע"י אזרחיות המעוניינות להיות מעורות בעשייה בכנסת – הפרוטוקולים ארוכים ומכילים מטבעם הרבה עיסוקים ביורוקרטיים וטכניים, שמקשים על הקוראת להפיק מהם תובנות.

בשנת 2021 התפרסם פרויקט חברתי עצמאי בשם "בטא מחוקקים" שנוצר במטרה לגשר על הפער, ומשתמש בפרוטוקולים שפורסמו על מנת לאפשר לאזרחיות להחשף לנעשה בכנסת באופן בלתי אמצעי תוך שימוש בממשק משתמש אינטואיטיבי. אתר הפרויקט מאפשר למשתמשת לבחור נושא כלשהו (למשל חינוך או בטחון) להציג את חברי הכנסת בעלי מספר ההתבטאויות הגבוה ביותר בנושא, וכן בציטוטים המדויקים שכל אחד מהם אמר בנושא.

השיטה בה משתמש הפרויקט כיום על מנת לקבוע שציטוט מסוים משתייך לנושא כלשהו היא באמצעות חיפוש מלא בטקסט (full-text search), כלומר – לכל נושא הוגדרה רשימה קצרה של מילות מפתח המזוהות איתו, וכל ציטוט המכיל לפחות אחת מהן קוטלג כמשויך לנושא אותו הן מייצגות. שיטה זו היא רחוקה מאידאלית, ממספר סיבות:

- ישנם נושאים שמילות המפתח העיקריות המייצגות אותן הנן בעלות מספר משמעויות שונות. למשל, בנושא "עולים חדשים" המילה העיקרית שהוגדרה לזיהוי הנושא היא "עולים", ולכן גם ציטוטים כגון "רואים שאנחנו *עולים* בסקרים כמעט ל-61 מנדטים" או "בדיון *עולים* שני נושאים מרכזיים", או "הם *עולים* לתקציב הרבה מאוד כסף" שויכו באופן שגוי לנושא "עולים חדשים".
- במקרים רבים, למרות שמופיעה בציטוט מילה האינדיקטיבית לנושא מסוים, הציטוט המלא עוסק בנושא אחר. למשל, הציטוט "אחלה *תרבות* דיבור יש במשרד החינוך. באמת, *תרבות* דיון מצוינת. משרד החינוך קורא עכשיו לחברה האזרחית על בסיס תרומות ומתנדבים לתקן כשלים של משרד החינוך?" אמנם מכיל את המילה "תרבות", אך ניכר שנושא הציטוט הוא חינוך ולא תרבות.

כלומר, השיטה הפשוטה בה משתמש הפרויקט כיום מספקת חלוקה לנושאים שמפיקה תוצאות שאינן מדויקות ואף מטעות. לכן, עולה צורך משמעותי לבצע שיוך של ציטוטים מפרוטוקולי הכנסת לנושאים באופן אמין ומדויק.

שאלת המחקר

מטרת המחקר בפרויקט זה היא **יצירת מודל לסיווג ציטוטים של חברי כנסת מתוך פרוטוקולי ועידות ומליאת הכנסת לנושאים** מתוך רשימה סגורה ומוגדרת מראש. המודל שפותח בפועל במסגרת הפרויקט מבדיל בין שמונה נושאים שונים - חינוך, רווחה, כלכלי, נשים ולהט"ב, בריאות, קורונה, בטחון ובטחון פנים (לפירוט על האופן בו נבחרו נושאים אלו ראו פרק "תהליך המחקר". כלומר, מדובר בבעיית multiclass text classification שהקלט שלה הוא ציטוט בעברית של ח"כ מתוך פרוטוקול פרלמנטרי, והפלט שלה הוא תגית יחידה המייצגת את הנושא בו עוסק הציטוט מבין שמונת הנושאים הנ"ל.

חשוב לציין כי רוב הציטוטים בפרוטוקולים אינם עוסקים באחד משמונת הנושאים הנ"ל (בין אם כי הם עוסקים באופן מובהק בנושא אחר, ובין אם כי אין להם נושא מובהק – למשל הציטוט "חברים, אני מבקש שנתחיל את הישיבה"). לכן, כדי שלא לשייך ציטוטים שכאלו לאחד משמונת הנושאים שהוגדרו, המודל שפותח מכיל אפשרות לקטלג ציטוטים בתור חסרי נושא (כלומר, המודל שפותח מאפשר 9 תגיות אפשריות).

מדובר בבעיה בעלת רלוונטיות גדולה למחקר עיבוד שפה טבעית לעברית, שכן לא רק שמדובר בבעיית סיווג טקסט עברי למספר תגיות, מטבעה של הבעיה היא הצריכה ניתוח מקדים של הדאטא הטקסטואלי כדי לזהות נושאים מובחנים. כלומר, מלבד היותה בעיית multiclass supervised text classification, בתור התחלה נפתרה במסגרת פרויקט זה בעיית unsupervised topic detection / clustering (ראו פירוט בפרק תהליך המחקר).

הגדרות ומונחים

לטובת אחידות ובהירות, נגדיר באופן מפורש מספר מונחים הרלוונטיים לפרויקט:

- **פרוטוקולים:** כלל המסמכים של פרוטוקולי ישיבות ועדות הכנסת ומליאת הכנסת.
- **מסמך:** פרוטוקול של ישיבה בודדת של ועדה ספציפית / מליאה בתאריך מסוים. כל מסמך שכזה מכיל תמלול מלא וכרונולוגי של הנאמר בישיבה, בשפה העברית (בה נערכה הישיבה במקור).
- **ציטוט:** ציטוט מתוך מסמך, המייצג את דבריו הרציפים של דובר מסוים שנכח בישיבה. כלומר, כל מסמך מתאר את השיח בישיבה באופן כרונולוגי לפי דוברים, ציטוט יכיל את כל מה שנאמר ע"י דובר כלשהו עד אשר יקטע ע"י דובר אחר בישיבה.

קבצי הדאטא אשר שימשו לאימון המודל, הקוד אשר נכתב בתהליך המחקר והמודל המאומן הסופי נגישים כולם בספריית ה-GitHub של הפרויקט.

איסוף וניקוי הדאטא

על מנת לפתור את הבעיה, יש צורך במאגר רחב של ציטוטים מפרוטוקולי הכנסת, המתויגים לפי נושאים. כיוון שלא קיים מאגר מתויג שכזה, הפרויקט התבסס על ציטוטים שאינם מתויגים, תוך ביצוע עבודת תיוג חצי-ידנית במסגרת הפרויקט (ראו פירוט בפרק "תהליך המחקר והמודל הנבחר").

איסוף הדאטא

כאמור, כל מסמכי הפרוטוקולים של הכנסת משנת 2015 ועד היום זמינים לקהל הרחב בתור קבצי docx במסגרת "פרוטוקול למידע נגיש – Odata". עם זאת, במסגרת פרויקט בטא מחוקקים התבצע תהליך scraping ו-parsing של כל המסמכים משנת 2015 ועד מאי 2022. מובילי "בטא מחוקקים" תרמו לנו בנדיבות את הדאטא המפורסר במלואו, בשני קבצי csv –

- CommitteeData.csv המכיל את פרוטוקולי ועידות הכנסת (1.4 GB)
- PlenumData.csv המכיל את פרוטוקולי מליאת הכנסת (0.3 GB)

כל אחד מהקבצים הנ"ל הוא קובץ טבלאי אחוד, שכל שורה בו מייצגת ציטוט ממסמך. עבור כל אחד מהציטוטים קיימים בקבצים פרטי המידע הבאים – שם הדובר, מספר סידורי חח"ע המייצג את הדובר

ותפקידו בעת אמירת הציטוט (כאשר לאותו אדם יכולים להיות מספר מספרי זיהוי שונים, אם תפקידו השתנה), טקסט חופשי של הציטוט, מספר סידורי של המסמך ממנו נלקח הציטוט, מספר סידורי של הציטוט בתוך המאמר ממנו הוא נלקח, וכן תאריך ושעת תחילת הישיבה בה נאמר הציטוט.

ניקוי ועיבוד הדאטא

על מנת להתאים את הדאטא שהתקבל מ"בטא מחוקקים" לצרכי הפרויקט, בוצע תהליך עיבוד מקדים שכלל איחוד של שתי הטבלאות הנ"ל לטבלה אחודה, ולאחריו סינון ועיבוד של הדאטא שכלל את הפעולות הבאות:

- **זיהוי דוברים מעניינים:** מטרת הפרויקט היא כאמור לסווג ציטוטים של חברי וחברות כנסת לנושאים. עם זאת, בישיבות הכנסת (ובמיוחד בישיבות הועידות) נוכחים אורחים רבים שאינם חברי כנסת. לפיכך, היה צורך לסנן את המאגר על מנת להשאיר ציטוטים המשויכים לדוברים שהנם ח"כים. אמנם לכל דובר יש מזהה מספרי המשויך לשילוב בין שמו ותפקידו הנוכחי, אך לא קיים מיפוי מספק של כל המספרים המזהים המשויכים לחברי כנסת בהווה או בעבר (התקבל מיפוי שכזה מ"בטא מחוקקים", אך התברר שלמרות שהמיפוי אכן מכיל רק חברי כנסת, לעתים הם מצוטטים בפרוטוקולים תחת מספרים מזהים אחרים כתלות בהקשר). לכן, הסינון של חברי כנסת התבסס על עמודת שם הדובר, המכילה תיאור מפורש של שם ותפקיד הדובר. בשלב זה הוסרו ידנית ציטוטים המשויכים לדוברים נפוצים שאינם רלוונטים (דוגמת מזכירת הכנסת, הקצרנית הפרלמנטרית) וציטוטים שלא משויכים לדובר מסוים (למשל ציטוטים שהדובר שלהם הוא "קריאה", "אורח").
- **הוספת עמודות בעלות מידע רלוונטי על הציטוט:** לאחר סינון ציטוטים המשויכים לדוברים לא רלוונטיים, הוספו לטבלה עמודות המכילות את מספר התווים בציטוט ואת מספר המילים בציטוט. כמו כן, הוספה עמודה המכילה את הציטוט לאחר ניקויו מסימני פיסוק. בנוסף, נוצרה עמודה נוספת המכילה את הציטוט לאחר הסרת סימני פיסוק ו-stopwords (רשימת ה-stopwords התבססה על רשימה קיימת, לה הוספו מילים נוספות היחודיות לקורפוס שזוהו כחלק מתהליך מחקר הדאטא). רוב העמודות הנ"ל שימשו בעיקר לטובת מחקר הדאטא.
- **סינון ציטוטים לא רלוונטיים:** לאחר אפיון ומחקר ראשוני של הדאטא המקורי, התברר כי מטבעם של ישיבות פרלמנטריות, הדאטא מכיל כמות גדולה מאוד של ציטוטים קצרים מכדי להיות רלוונטיים (כיוון שהדוברים קוטעים זה את זה בתכיפות, ישנם ציטוטים רבים דוגמת "נו") או ארוכים מידי (למשל הקראה רציפה של הצעת חוק במלואה). לכן, בשלב זו סוננו ציטוטים בני פחות מ-15 מילים או יותר מ-150 מילים. כמו כן, כיוון שכל ישיבה נפתחת בנוהל בברכות והצגת הנוכחים, הוסרו ציטוטים שהמספר הסידורי שלהם בתוך המסמך אליו הם משתייכים קטן מ-5.

לאחר ביצוע הפעולות הנ"ל, הוסרו מהטבלה ציטוטים כפולים (מצב נפוץ יחסית, שכן כיוון שניהולן של ישיבות פרלמנטריות מתנהל לפי נוהל מוסדר, ישנם משפטים קבועים שמוקראים בשלבים שונים במהלכה). לאחר הסרת הכפילויות, נשמרה הטבלה האחודה המכילה את המידע הרלוונטי עבור כל הציטוטים שעברו את שלבי הסינון והעיבוד (0.03 GB).

סקירת ספרות

כפי שהוגדר בפרק המבוא, מטרתו של פרויקט זה היא לפתור בעיית multiclass text classification. עם זאת, כפי שיוסבר בפרק תהליך המחקר, המחקר כלל שני שלבים – בשלב הראשון בוצע תהליך unsupervised topic detection שאפשר ליצור תיוג ראשוני לנושאי המשפטים ששימש לתעדוף של תיוג ידני, ובשלב השני בוצעה משימת supervised text classification במסגרתה פותח מודל אשר מתייג ציטוטים בנושאים המתאימים להם אשר אומן על הדאטא שתויג ידנית. בפרק זה יוצגו בקצרה השיטות המקובלות עבור שתי המשימות הנ"ל.

Unsupervised topic detection: משימת unsupervised topic detection (לעתים מכונה topic modeling או topic segmentation) היא משימה שמטרתה לפתח מודל סטטיסטי שמזהה נושאים מופשטים באסופת מסמכים (במקרה של הפרויקט הנוכחי, אסופת ציטוטים) ומקבץ אותם לקבוצות של מסמכים החולקים נושא משותף[1]. לרוב, מודלים שכאלו יאפיינו כל מקבץ (cluster) מסמכים המייצג נושא מסוים באמצעות מילים נפוצות או משמעותיות המייצגות את המשותף בין המסמכים המשתייכים לאותו נושא. שיטות נפוצות בתחום:

- **SVD – Singular Value Decomposition** – מכונה לעתים גם LSA (latent semantic analysis) או LSI (latent semantic indexing). בשיטה זו יוצרים מטריצה M שבה כל שורה היא מילה וכל עמודה היא מסמך, שמייצגת כמה כל מילה היא משמעותית ביצוג של כל מסמך. מחשבים את פירוק ה- SVD של המטריצה הזו, ובהנתן הפירוק ניתן לבחון את הערכים הסינגולריים על האלכסון של המטריצה Σ שמייצגים נושאים פוטנציאליים, ואת המטריצות U, V^* שמייצגות את הקשר בין נושאים (topics) למונחים (terms) ובין נושאים (topics) למסמכים (documents) בהתאמה. גרסה משופרת של שיטה זו (Truncated SVD) מציעה לצמצם את זמני החישוב הנדרשים לשיטה ע"י חישוב מצומצם של הפירוק עבור תת קבוצה של נושאים מעניינים[2].
- **NMF – non-negative matrix factorization** - בשיטה זו מגדירים בתור היפר-פרמטר את מספר הנושאים אותם נרצה לייצר, ומפיקים שתי מטריצות – המטריצה W שמקשרת בין מסמכים לנושאים, והמטריצה H שמקשרת בין נושאים למילים. מתחילים מאתחול של מטריצות W,H כך שיכילו ערכים חיוביים וקטנים, ולאחר מכן מבצעים תהליך של אופטימיזציה כדי לגרום להן לייצג את הקשרים הרצויים[3].
- **LDA – Latent Dirichlet Allocation** – בשיטה זו כל נושא מיוצג ע"י קבוצה לא ידועה של מילים, כאשר מפענחים אילו מילים מייצגות איזה נושא ע"י לנסות למפות בין מסמכים לנושאים כך שהמילים בכל מסמך מייצגות את הנושא בצורה מיטבית. שיטת LDA מניחה שכל מסמך יכול להיות מיוצג ע"י מספר נושאים שונים, שחלקים מייצגים אותו יותר טוב וחלקם פחות. בשיטה זו מגדירים בתור היפר-פרמטרים את מספר הנושאים אותם נרצה לייצר, את מידת הדמיון בין המסמכים ואת מידת הדמיון בין נושאים. לאחר בחירת ההיפר-פרמטרים, מחשבים עבור כל מסמך

---

[1] Sharma, D., Kumar, B., & Chand, S. (2017). A survey on journey of topic modeling techniques from SVD to deep learning. *International Journal of Modern Education and Computer Science*, *9*(7), 50.
[2] Ke, Z. T., & Wang, M. (2022). Using SVD for Topic Modeling. *Journal of the American Statistical Association*, (just-accepted), 1-32.
[3] Kuang, D., Choo, J., & Park, H. (2015). Nonnegative matrix factorization for interactive topic modeling and document clustering. In *Partitional clustering algorithms* (pp. 215-243). Springer, Cham.

וקטור התפלגות שמייצג כמה כמה אחוז מהמסמך מיוצג ע״י כל נושא. יוצרים את הוקטור הזה באמצעות ספירה – מחלקים את כל המסמכים לנושאים באופן אקראי, ואז מבצעים תהליך אופטימיזציה שמבצע מקסום של ההסתברות שכל מסמך משויך לנושא מסוים (כלומר מקסום ההסתברות של זוגות של מסמך ונושא)[4].

Supervised text classification: בעיית קלסיפיקציה לטקסט היא בעיה נפוצה בה מתבצע שימוש במודלי Machine Learning על מנת לבצע סיווג לקטגוריות של מסמכים, כאשר האימון מתבסס על דאטא מתויג, כלומר שיוך מפורש של כל מסמך לקטגוריה מסוימת. בעיית קלסיפיקציה בה ישנן יותר משתי קטגוריות אפשריות תקרא בעיית multiclass, ואם מסמך יכול להשתייך ליותר מקטגוריה אחת היא תקרא בעיית multilabel. כיוון שמודלי Machine Learning הם מודלים מתמטיים וסטטיסטיים שהקלט שלהם הוא וקטור של מספרים, לא ניתן להשתמש בהם באופן ישיר על מנת לסווג מסמכי טקסט. לכן, יש צורך בביצוע תהליך וקטוריזציה במסגרתו נוצר עבור כל מסמך וקטור מספרים המהווה ייצוג הולם של המסמך[5]. נציג מספר שיטות וקטוריזציה נפוצות לבעיה זו[6]:

- **Binary Term Frequency** – בשיטה זו עבור כל מונח (term) באוצר המילים, בודקים האם הוא מופיע (1) או לא מופיע (0) במסמך, וכך יוצרים עבור המסמך וקטור בינארי.
- **Bag of Words** – בשיטה זו עבור כל מונח (term) באוצר המילים, סופרים את כמות הפעמים שהוא מופיע במסמך.
- **TF-IDF – Term frequency-inverse document frequency**– שיטה שנועדה לייצר ייצוג למסמך שאינו רק ספירת מופעים, אלא מייצג את החשיבות של כל מילה במסמך. שיטה זו מתבססת על מספר המפעמים שמילה מופיעה במסמך (TF – Term frequency) אך גם על ההופכי של מספר הפעמים שהמילה הופיעה במסמכים בקורפוס (IDF – Inverse Document Frequency), כלומר IDF יקבל ערך גבוה עבור מילים נדירות וערך נמוך עבור מילים נפוצות בקורפוס.
- **שיטות מבוססות Word2Vec** – שיטת Word2Vec היא שיטת וקטוריזציה עבור מילה בודדת המתבססת על הגדרת גודל של חלון המייצג את מספר המילים שנחשבות להקשר של מילה מסוימת במשפט, וכך לספור עבור כל מילה את המילים המופיעות בהקשר שלה בקורפוס כולו. קיימות מספר וריאציות לשימוש בשיטה זו עבור וקטוריזציה של מסמך שלם, למשל סכימה או מיצוע של וקטורי ה-Word2Vec של כל אחת מהמילים במשפט, או שיטת Doc2Vec בה משתמשים בוקטורי ה-Word2Vec של כל אחת מהמילים במשפט בשילוב עם וקטור נוסף המאחסן מידע הנוגע למידע שחסר בהקשר הנוכחי של המשפט.

---

[4] Jelodar, H., Wang, Y., Yuan, C., Feng, X., Jiang, X., Li, Y., & Zhao, L. (2019). Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey. *Multimedia Tools and Applications*, *78*(11), 15169-15211.

[5] Kadhim, A. I. (2019). Survey on supervised machine learning techniques for automatic text classification. *Artificial Intelligence Review*, *52*(1), 273-292.

[6] Singh, A. K., & Shashi, M. (2019). Vectorization of text documents for identifying unifiable news articles. *International Journal of Advanced Computer Science and Applications*, *10*(7).

- **GloVe** – שיטה מבוססת ספירה בה יוצרים מטריצת הופעות משותפות (co-occurance) גלובלית (על הקורפוס כולו) בה כל שורה מייצגת מילה וכל עמודה מייצגת מסמך, ומכילה את כמות הפעמים שכל מילה הופיעה בכל מסמך.

לאחר ביצוע וקטוריזציה למסמך, על מנת לבצע משימת multiclass classification יש להשתמש במודל Machine Learning התומך בתגיות מרובות, למשל רגרסיה לוגיסטית, SVM, עצי החלטה, Random Forest, naïve Bayes, KNN וכו'[7].

שימוש במידול נושאי כתיוג ראשוני: ההשראה לשימוש בשיטות topic modeling בתור תיוג ראשוני התקבלה ממאמר שפורסם ע"י IBM שכותרתו Cluster & Tune: Boost Cold Start Performance in Text Classification בו נבחנו מספר שיטות לדירוג הדרגתי תוך שימוש בשיטות כאלו[8].

הערכת ביצועי המודל

כיוון שמדובר בבעיית supervised classification, ניתן לבצע אוולואציה לתוצאות המודל באמצעות מטריקות המשוות בין התגית שחזה המודל לבין התגית האמיתית. עבור כל אחד מהמודלים, הערכתי את תוצאותיו באמצעות שלוש מטריקות – accuracy, precision (macro), recall (macro). כל אחת משלושת המטריקות מספקת מידע על מובן אחר של הצלחת המודל – מדד ה-accuracy מייצג את אחוז הציטוטים הכולל שתויגו באופן נכון מבלי להתייחס לנושאים ספציפיים, בעוד מדדי ה-precision וה-recall מייצגים כל אחד בדרכו את ההצלחה של המודל בזיהוי כל אחד מהנושאים. מדד ה-precision מייצג עבור נושא מסוים כמה מתוך הציטוטים העוסקים בו תויגו נכונה ע"י המודל (למשל - מתוך כל הציטוטים העוסקים בחינוך, כמה מתוכם תויגו "חינוך" ע"י המודל), בעוד מדד ה-recall מייצג עבור נושא מסוים כמה התיוג של המודל עבורו הוא אמין (למשל, מתוך כל הציטוטים שהמודל תייג "חינוך", כמה מהם באמת עוסקים בחינוך).

כיוון שמדובר בבעיית multiclass, כדי להעריך את ביצועי המודל יש לבצע אגרגציה על מטריקות האוולואציה של כל אחד מהנושאים - בחרתי להתמקד בעיקר במדדי macro ולא micro כיוון שהדאטא מאוד לא מאוזן (40% ממנו מתויג כחסר נושא).

עם זאת, המטריקות הן לא הדרך היחידה בה הוחלט לבחון את ביצועי המודל. כיוון שהמודל מיועד לשימוש ע"י פרויקט "בטא מחוקקים" על מנת לשקף את מידת העשיה של כל ח"כ/ית בכל אחד משמונת הנושאים, יש השפעה שונה לסוג הטעות שהמודל מבצע. אם המודל מתייג באופן שגוי ציטוט בעל נושא כחסר נושא, המשמעות היא שהח"כ שאמר את הציטוט יזכה לפחות קרדיט משמגיע לו בנוגע לעיסוק באותו נושא. זאת אמנם טעות חמורה, אך לא חמורה כמו לתייג ציטוט כמשויך לנושא קונקרטי כלשהו מבית 8 הנושאים – שכן במקרה זה הח"כ הדובר יקבל קרדיט עודף על נושא שלא עסק בו. ככל שהבלבול בין המחלקות יהיה יותר גדול, הקרדיט העודף הזה יכול להיות משמעותי מאוד וליצור מצג שווא לפיו ח"כ מבצע עשייה

[7] Aly, M. (2005). Survey on multiclass classification methods. *Neural Netw*, *19*(1), 9.

[8] Shnarch, E., Gera, A., Halfon, A., Dankin, L., Choshen, L., Aharonov, R., & Slonim, N. (2020). Cluster & tune: Enhance bert performance in low resource text classification.

משמעותית בנושא בו המודל מתבלבל. לכן, בנוסף למטריקות, הוחלט לתעדף גם מזעור של כמות הציטוטים שתויגו באופן שגוי בנושא קונקרטי כלשהו.

תהליך המחקר

תיוג ראשוני אוטומטי

בפרויקט זה, אמנם קיימים עשרות אלפי ציטוטים מתאימים שנותרו לאחר הסינון והעיבוד הראשוני, אך הם אינם מתויגים, וכמותם גדולה מכדי לאפשר תיוג ידני של כולם. עם זאת, יש צורך בתיוג אמין ואיכותי של כמות מספקת של ציטוטים. לכן, השלב הראשון של הפרויקט שאב השראה ממאמרים שבחנו אפשרות לאימון מודלי קלסיפיקציה על דגימות שתויגו באופן אוטומטי במקרים בהם דאטא מתויג אינו מספיק זמין.

המטרה בשלב ראשון זה היתה ליצור תיוג ראשוני עבור כמות גדולה של ציטוטים, שיוכל לשמש על מנת לבחור מבין עשרות אלפי הציטוטים הזמינים את אלו שיתויגו ידנית בשלב מתקדם יותר. כדי לעשות זאת, בשלב זה בוצע Topic modeling על דגימות אקראיות בגדלים שונים מתוך דאטא האימון. נבחרו שמונה קומבינציות של דגימות בגדלים שונים של הדאטא (15,000, 20,000 ו-25,000 ציטוטים) עם מספר קלאסטרים שונה (בין 25 ל-50 קלאסטרים). עבור כל אחת מהקומבינציות אומן מודל LDA אשר חילק את הדאטא למספר הקלאסטרים שהוגדר לו. מודל LDA נבחר כיוון שהוא מאפשר איזון בין זמן ריצה מהיר לתוצאות איכותיות. מספר הדגימות בדאטא האימון ומספר הקלאסטרים לא נבחרו מראש, אלא נבחנו באופן הדרגתי בתהליך מחקרי בו בכל שלב אומן מודל LDA על קומבינציה מסוימת, נבחנו התוצאות (כמות הקלאסטרים שמייצגים נושא קוהרנטי), ובהתאם בוצעו התאמות לשני ההיפר-פרמטרים הנ"ל (לדוגמא, אם נראה שקיים קלאסטר אחד המכיל שני נושאים מובחנים שונים, בוצעה הגדלה של מספר הקלאסטרים על מנת לאפשר למודל להפריד ביניהם).

לאחר מכן, בוצע מעבר ידני על המילים המאפיינות כל אחד מהקלאסטרים שנוצרו, ומתוכם נבחרו בקפידה קלאסטרים אשר המילים שמאפיינות אותם העידו שהוא מייצג נושא קוהרנטי וברור. לכל קלאסטר שכזה שויך שם של נושא (המתבסס על המילים המאפיינות אותו, למשל קלאסטר שאופיין בעיקר ע"י המילים "לקידום, האישה, מעמד, לנשים, מגדרי" שויך לנושא "נשים"). שם הנושא שנבחר לכל קלאסטר שימש כתיוג ראשוני עבור כל הציטוטים ששויכו אליו ע"י מודל ה-LDA.

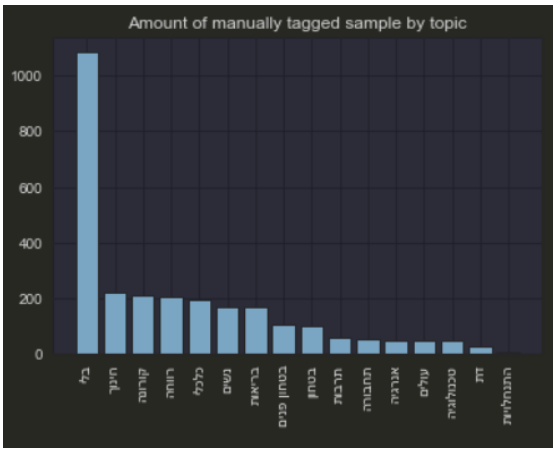
לאחר ביצוע תהליך זה עבור כל 8 הקומבינציות, בוצעה הסרה של כל הציטוטים שמידת הבטחון של המודל בהם (כלומר ההסתברות שנתן מודל ה-LDA לשיוכם לקלאסטר) היתה קטנה מ-0.7. לאחר סינון זה, נוצרה טבלה המכילה 12,125 ציטוטים המשתייכים ל-12 נושאים.

תיוג ידני

לאחר שיוך הדאטא ל-12 הנושאים שתויגו אוטומטית בשלב הראשון, הועברו הציטוטים לתיוג ידני. כדי לייעל את שלב זה, הציטוטים חולקו לפי הנושאים אליהם שויכו באמצעות התיוג האוטומטי, ומוינו לפי מידת הבטחון של מודל ה-topic detection בתיוג (כלומר, הציטוטים שהופיעו ראשונים הם הציטוטים שמודל התיוג האוטומטי היה הכי בטוח בשיוך שלהם לנושא). המתייגים התבקשו לתייג כל אחד מהציטוטים לפי המקרא הבא:

נושא	תחומים המשתייכים לנושא
קורונה	קורונה, חיסוני קורונה, סגרים והגבלות הקשורים לקורונה
כלכלי	מיסוי, פנסיה, שוק ההון, בורסה, מדיניות כלכלית (לא כולל אזכורים לחלוקת תקציב המדינה שאינם קשורים למדיניות כלכלית)
רווחה	ילדים ונוער במצוקה, קשישים, ביטוח לאומי, עוני, ניצולי שואה, סמים והתמכרויות, דיור ציבורי, מחירי הדיור, תוכניות לבניה למגורים, חקיקה בנושאי דיור, שכירות, מיסוי על דירות
נשים ולהט"ב	ייצוג נשי, קהילת הלהט"ב, אלימות נגד נשים, הטרדות מיניות, זכויות נשים, זנות
בריאות	רפואה, בתי חולים, קופות חולים, תקציבי תרופות, בריאות הנפש, קנאביס רפואי, מערכת הבריאות, רופאים
בטחון	מלחמה, חיילים, חיילים בודדים, שירות צבאי, מילואים, פיגועים, תקציב הבטחון
בטחון פנים	משטרה, בתי כלא, עצורים ואסירים, אלימות בחברה
חינוך	חינוך, גנים וצהרונים, בתי ספר, תיכונים, השכלה גבוהה, סטודנטים, תכניות לימודים, חינוך מיוחד
תחבורה	תחבורה ציבורית, כבישים, טיסות, רכבים, רישוי רכבים, הולכי רגל, אופניים, קורקינטים, רכבות, אוטובוסים, רכבת קלה
תרבות	תרבות, מוזיקה, ספורט, שפרות, אמנות, תאטרון, קידום מחקר מדעי
אנרגיה ואקולוגיה	אקולוגיה, חשמל, מים, זיהום אוויר, שטחים פתוחים, מיחזור
טכנולוגיה	מאגרי מידע ממשלתיים, שימוש שטכנולוגיה ע"י גורמי שלטון (שב"כ, משטרה, כנסת וכו'), השפעת הטכנולוגיה על אזרחים
עולים	עולים חדשים, קליטה ממדינות שונות, פרשת ילדי תימן
דת	רבנים, אתרים דתיים, בתי כנסת, צביון השבת, תקציבים הקשורים לדת, הפרדת דת ומדינה
התנחלויות	התיישבות ביהודה ושומרון, היתרי בניה ביהודה ושומרון, בטחון ביהודה ושומרון
בלי	ציטוטים בעלי נושא מובהק שאינו משתייך לאף אחד מהנושאים המוזכרים מעלה, או שאינם בעלי נושא מובהק

במסגרת שלב התיוגו הידני, תויגו 2,741 ציטוטים, מתוכם 1,657 שויכו לנושא (כלומר תויגו לנושא שאינו "בלי"). ניתן להתרשם מהתפלגות הנושאים מבין הציטוטים שתויגו בגרף הבא:



איור 1 – התפלגות נושאית בציטוטים שתויגו ידנית

לטובת אימון המודל בפועל, נבחרו רק נושאים אשר היו עבורם לפחות 100 ציטוטים מתויגים, וכך נבחרו שמונת הנושאים הבאים - חינוך, רווחה, כלכלי, נשים ולהט"ב, בריאות, קורונה, בטחון ובטחון פנים.

בסיום שלב זה נוצרה טבלה המכילה בסה"כ 2,458 ציטוטים, מתוכם 1,374 משויכים לשמונת הנושאים הנבחרים ו-1,084 הנם חסרי נושא. הטבלה שמורה בניתוב הבא. לאחר יצירת טבלה זו, הוצמדו לכל ציטוט המופיע בה גם הנתונים הקיימים עליו בטבלה המקורית, דוגמת אורך הציטוט, העתק של הציטוט שהוסרו ממנו סימני פיסוק ו-stopwords וכו' – טבלה סופית זו שמורה בניתוב הבא.

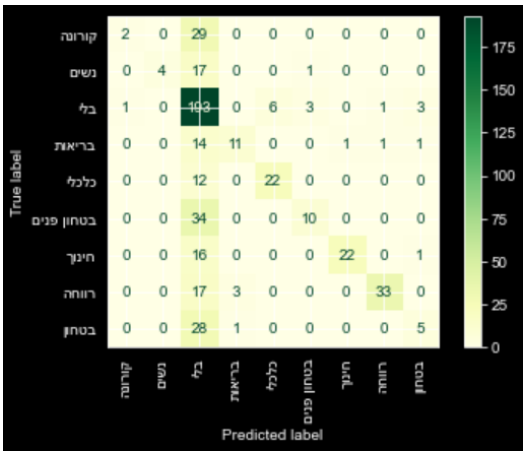
אימון ובחינת מודלים

לאחר שנוצר מאגר סופי של ציטוטים, בוצעה חלוקה של הדאטא ל-train ו-test (שהכילו 80% ו-20% מהדאטא בהתאמה). לאחר מכן, בוצע תהליך מחקרי בו נבחנו מספר שיטות וקטוריזציה (sentence embeddings) ומספר מודלי קלסיפיקציה שונים לפתרון הבעיה:

- **פורמט הקלט:** נבחנה האפשרות לבצע את אימון המודלים על הטקסט המקורי של הציטוט למול הטקסט שעבר הסרה של סימני פיסוק ו-stopwords. כיוון שהטקסט המנוקה הראה תוצאות טובות במקצת עבור כל המודלים, הוחלט להשתמש בו.
- **וקטוריזציה:** שיטות הוקטוריזציה שנבחנו היו ספירה פשוטה של הופעות בקורפוס (Bag of Words), וכן שיטת הוקטוריזציה TF-IDF.
- **מודלים:** המודלים שנבחנו היו רגרסיה לוגיסטית (נבחנה עבור ההיפר פרמטר C עם ערכי 0.5, 1, 2), Random Forest (נבחן עבור מספר קומבינציות אפשריות של עומק מקסימלי ומספר העצים ביער), ו-SVM לינארי. כיוון שכמות הדאטא היא מצומצמת, כל המודלים אומנו באמצעות 5-fold cross-validation על דאטא האימון.

איטרציית אימון ראשונה

בשלב ראשון, המודלים האלו אומנו על דאטא האימון במלואו (אשר הנו מאוד מוטה וקשה במהותו להכללה, שכן התגית הנפוצה ביותר בפער היא "ללא נושא" – שהיא בהגדרתה מגוונת וללא מאפיינים סמנטיים מאחדים ברורים). כאשר המודלים אומנו על דאטא האימון המלא, אמנם היו הבדלים בביצועים בין המודלים, אך כולם החזירו תוצאות דומות יחסית אשר הדגימו בלבול משמעותי. הבלבול הנפוץ היה בין תגיות פחות נפוצות, שתויגו באופן שגוי בתור "ללא נושא". ניתן לראות זאת בגרף לדוגמא הבא:



איור 2 – מטריצת בלבול עבור מודל רגרסיה לוגיסטית (עם C=1 וקטוריזציה באמצעות BoW) שאומן על דאטא לא מאוזן

לאחר זיהוי הבלבול במודלים, בוצעה העמקה בסיבה לבלבול. הסיבה היא ככה"נ חוסר היכולת (המובן) של המודל להכליל את תגית "ללא נושא" (שאין לציטוטים בה מאפיינים משותפים מובהקים) ביחד עם הטית הדאטא המשמעותית (כלומר ש-40% מהדאטא תויג כחסר נושא, והשאר התחלק בין 8 הנושאים הקונקרטיים). היו מקרים רבים בהם המודל תייג ציטוט בעל שיוך נושאי מאוד מובהק כחסר נושא, כמו שניתן לראות למשל בציטוטים הבאים (שהמודל תייג באופן שגוי כחסרי נושא):

ציטוט	תגית נכונה
לאחר מכן כלל הנערים מופעלים בתוך כיתות הלימוד. כיתות הלימוד שלנו מגוונות, אתם תראו את זה בעין, יש לנו גם חדר מחשבים וגם ספרייה וגם כיתת אמנות, דברים שאתם מכירים אותם גם מבחוץ, וכל זה קורה עד שעות הצהריים. בשעות הצהריים אנחנו יוצאים להפסקת אוכל, לאחר מכן הנערים מופעלים בפעילויות רבות ומגוונות. אנחנו מקפידים מאוד לתת גם מענה לפער המצריך טיפול ייחודי, יש לנו פרויקטים כמו פרויקט תנופה ופרויקט איתן שמופנה לבעיה הספציפית של הנערים, למשל בעיית אלימות.	חינוך
דו"ח הנציבות האירופאית של שנת 2018 קבע שההשתתפות של נשים במדע וטכנולוגיה תורמת להגברת האיכות, הרלוונטיות החברתית והתחרותיות של מחקר וחדשנות. מחקר ממשלתי של הכלכלנית הראשית במשרד האוצר, קבע שמקצועות המדע – מחשבים, רפואה, הנדסה ופיזיקה הם התחומים שהכי מקדמים מוביליות חברתית, כאשר בנות מהוות כמעט מחצית, 50% מהניגשות לבגרות בחמש יחידות מתמטיקה, אבל בבחירת המגמות הריאליות ניתן לראות הטיות מגדריות ברורות - המקצועות הם גבריים, פיזיקה ומחשבים שבהם יש בין 36% בערך או 32% של בנות בהתאמה ובכל מה שקשור למדעי המחשב – רק 32% מכלל הסטודנטיות לומדות את המקצוע, זאת לעומת שיעורן מכלל הסטודנטים אשר עומד כל כ-60% בתואר הראשון.	נשים
אני חושבת שכבר ב-2016 הייתי חברה בוועדת חוץ וביטחון, הייתה ועדה חסויה מאז, ואפשר להגיד, ישב בנימין נתניהו, ראש הממשלה, הוא אמר ששני האיומים הכי גדולים על מדינת ישראל היום הם אז כבר הם סייבר ורחפנים, אמר וצדק. האירוע התחיל להתגלגל בעיקר בשנים האחרונות, ואנחנו רואים יותר ויותר גופים שמותקפים. ראינו את התקיפה על ביה"ח הלל יפה, אבל זה כמובן גם לפני שירביט וכמובן אוניברסיטת בר אילן ועוד הרבה גופים.	בטחון

איטרציית אימון שניה

כיוון שהבלבול נבע מהקושי של המודל להכליל על המחלקה "ללא נושא" כאשר הוא אומן על דאטא מאוד לא מאוזן (המכיל כמות גדולה באופן לא פרופורציונלי של התגית "ללא נושא" בהשוואה לתגיות אחרות), בשלב שני נבחנו אותם המודלים שנבחנו בשלב הראשון כאשר הם מאומנים על דאטא מאוזן יותר, המכיל יצוג קטן יותר של דגימות המתויגות כחסרות נושא. כלומר, הדאטא עבר סינון מקדים והוסרו ממנו רוב הדגימות המתויגות "ללא נושא" – הושארו 200 דגימות אקראיות מתגית זו. הכמות נבחרה במטרה ליצור איזון בדאטא (המספר 200 קרוב למספר הממוצע של דגימות פר נושא), ונבחרה לאחר בחינת ביצועי המודלים על מספר כמויות שונות (100,150,200,250,300).

לאחר האימון מחדש, בוצעה השוואה בין ביצועי המודלים שאומנו על דאטא מאוזן לבין ביצועי המודלים שאומנו על הדאטא המלא והמוטה. ראו את הטבלה מטה (איור 3), המציגה את ההפרש הממוצע והמקסימלי בין ציון המודל שאומן על דאטא מאוזן לבין זה שאומן על הדאטא המוטה (עבור 15 המודלים שנבדקו – מודלי רגרסיה לוגיסטית, random forest ו-SVM עם שילובים שונים של היפר-פרמטרים). כפי שניתן לראות בטבלה, איזון הדאטא גרם לשיפור כללי משמעותי בביצועים:

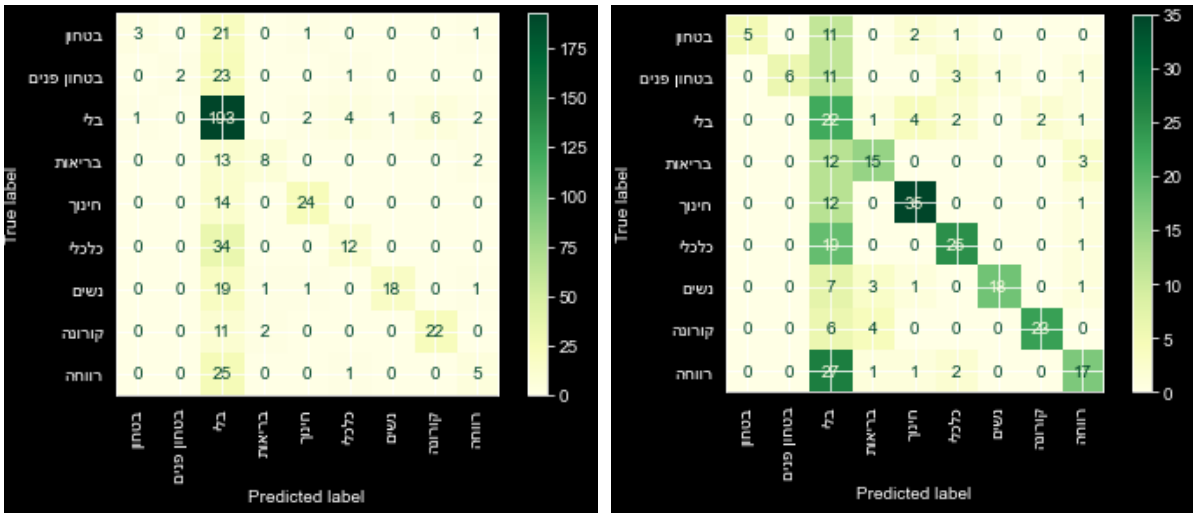
- **ירידה במטריקת accuracy:** אמנם מטריקת ה-accuracy ירדה בממוצע בכ-0.7, אך אין זה מעיד על ירידה בביצועי המודל – המודלים שאומנו על הדאטא המוטה נטו לתייג את רובן המוחלט של הדגימות בתור "ללא נושא", וכיוון שכ-40% מהדאטא המוטה אכן משויך לתגית זו, אין זה מפתיע שאחוז הדגימות שתויג באופן נכון הוא גבוה יחסית – מה שלא יקרה במודלים שאומנו על דאטא מאוזן.
- **העדר שינוי במטריקת precision:** ברוב המודלים לא נצפה שינוי משמעותי במטריקת ה-precision כתוצאה מאיזון הדאטא – בממוצע נמדדה ירידה זניחה של פחות מ-0.01 במטריקה זו.
- **שיפור במטריקת recall:** ברוב המודלים נצפה שיפור משמעותי מאוד במטריקה זו כתוצאה מאיזון הדאטא – בממוצע נצפתה עלייה של 0.13 במטריקת ה-recall, כאשר עבור מספר מודלים נצפתה עלייה של אף יותר מ-0.2.
- **שיפור במטריקת f1 score:** גם כאן נמדד שיפור משמעותי – בממוצע נצפתה עליה של 0.12 במדד זה (כיוון שברוב המודלים לא היה שינוי משמעותי במטריקת ה-precision והיתה עליה משמעותית מאוד במדד ה-recall, כצפוי נקבל עליה מתונה ב-f1 score המשקללת את שתי המטריקות הנ"ל).

	accuracy change	precision change	recall change	f1 change
max	-0.01012367	0.122094542	0.213698895	0.199984273
mean	-0.074520787	-0.006137996	0.138546833	0.120021545

איור 3 – שינויים ממוצעים ומקסימליים במטריקות המדידה כתוצאה מאיזון דאטא האימון עבור המודלים שנבחנו

נביט במטריצת הבלבול של מודל לדוגמא (מודל רגרסיה לוגיסטית עם פרמטר C=2 שאומן על דאטא שעבר וקטוריזציה באמצעות BoW) בהשוואה למטריצת הבלבול של המודל המקביל לו שאומן על הדאטא הלא מאוזן, נוכל להבחין בשיפור משמעותי:

המודל הנבחר **מודל מקביל שאומן על דאטא לא מאוזן**



איור 4 –מטריצות בלבול

השוואה בין המודל שאומן על הדאטא המלא והמוטה לבין זה שאומן על דאטא מאוזן, אמנם הראתה שיפור במידת הבלבול של המודל לגבי התגית "ללא נושא", אך ניכר שהבלבול ניכר מאוד גם במודל שאומן על

דאטא מאוזן. הבלבול קיים בשני הכיוונים – המודל מתקשה מעט לזהות ציטוטים חסרי נושא ולעתים משייך אותם למגוון נושאים אחרים (דוגמת בריאות או חינוך), אך עיקר הבלבול הוא בכיוון ההפוך, כלומר – המודל תייג כמות משמעותית של ציטוטים בעלי נושא כלשהו בתור חסרי נושא. למעשה, 83% מהתגיות השגויות שהעניק המודל, היו תגיות "ללא נושא" (ירידה של כ-3% בהשוואה לאותו מודל שאומן על דאטא לא מאוזן).

עם זאת, מיקוד בציטוטים חסרי נושא שתויגו על ידי המודל בנושא כלשהו מראים במקרים רבים בלבול הגיוני, שכן הציטוט מכיל מילים אינדיקטיביות לנושא שתויג וגם למתייג אנושי ככה"נ היה קשה להכריע האם הציטוט הוא אכן חסר נושא. לדוגמא, הציטוט הבא שתויג ע"י מתייג אנושי כחסר נושא ותויג ע"י המודל בתור ציטוט בנושא חינוך: *"למי שלא מכיר את ועדת החינוך, התרבות והספורט, אין קריאות ביניים. גם אם משהו מקרקר בבטן ומפריע, אפשר להתאפק, לחכות להזדמנות לקבל רשות דיבור ולדבר. אם זה משהו שבאמת הנשימה נעתקת, אפשר להרים את היד ואני אבין לפי זה."*

איטרציית אימון שלישית

אמנם איטרציית האימון השניה (אימון על דאטא מאוזן) השיגה שיפור בהשוואה לאיטרציה הראשונה (אימון על דאטא לא מאוזן), אך היא עדיין הדגימה בלבול משמעותי שגרם לתיוג של הרבה דגימות בתור חסרות נושא.

לכן, באיטרציה השלישית במחקר התבצעה בדיקה לגבי ביצועי המודלים כאשר הם מאומנים על דאטא שאינו מכיל כלל דגימות המתויגות כחסרות נושא. המטרה בשלב זה היתה לנטרל את הבלבול של המודל על מחלקת "ללא נושא" (שהיא קשה להכללה מהגדרתה כמחלקה אקלקטית בלי קווי דמיון משמעותיים בין הדגימות), ובמקום זאת לאפשר לו לבחור בין הנושאים האמיתיים. כיוון שהמודלים בוחרים את התיוג לכל דגימה ע"י בחירת המחלקה עם ההסתברות הגבוהה ביותר בוקטור ההתפלגות שהמודל מחזיר, הרעיון המנחה בשלב זה במחקר היה לזהות רף של ציון הסתברותי, כך שכל דגימה שהתיוג המקסימלי שלה קיבל הסתברות נמוכה מהרף יתויג באופן ידני בתור חסר נושא.

כלומר, בשלב זה אומנו אותם המודלים על דאטא שאינו מכיל כלל דגימות חסרות נושא. בתור התחלה, לא בוצעה הסרה של דגימות שהתיוג שלהן לא עבר את הרף ההסתברותי, ונבחנו המטריקות על התיוגים כולם במטרה לבחון את מידת הבלבול של המודלים בין התגיות. כעת, כאשר משווים את המודלים באיטרציית האימון השלישית (שאומנו רק על דגימות עם נושא) לאלו מאיטרצית האימון השניה (שאומנו על דאטא מאוזן שהכיל כמות קטנה של דגימות ללא נושא), ניתן לראות ששינוי זה בדאטא האימון גרם לשיפור דומה לשיפור שנוצר כתוצאה מאיזון הדאטא. כלומר, ערך ה-precision נשאר דומה, וערכי ה-recall וה-f1 עלו משמעותית (ב-0.5 וב-0.4 בממוצע בהתאמה). בשונה מהמעבר בין איטרצית האימון הראשונה לשניה, כאן ניתן לראות עלייה בערך ה-accuracy.

	accuracy change	precision change	recall change	f1 change
max	0.095959	0.025048	0.082947	0.068757
mean	0.073313	0.001323	0.05864	0.044152

איור 5 - שינויים ממוצעים ומקסימליים במטריקות המדידה כתוצאה ממעבר לאימון על תגיות המייצגות נושא בלבד (בהשוואה למודל מאיטרציית האימון השניה)

עם זאת, למרות שהמטריקות הדגימו שיפור, נראה שהשינוי במודל לא שיפר את הבלבול בין המחלקות, ואף החמיר אותו. נביט למשל במטריצת הבלבול של מודל random forest עם 200 עצים בעומק מקסימלי 100 שאומן על דאטא שעבר וקטוריזציה באמצעות BoW, שהיה המודל שהשיג את המטריקות הטובות ביותר באיטרציה זו. אם נשווה אותו למודל המקביל לו באיטרצית האימון השניה, נקבל את התוצאות הבאות:

	גרסה 2	גרסא 3 - רק תיוג קונקרטי	גרסה 3 - אפשור תיוג "ללא נושא"
מס' דגימות שתויגו באופן שגוי בנושא קונקרטי	32	87	69
מס' דגימות שתויגו באופן שגוי בתור "ללא נושא"	94	0	19

איור 6 – מיפוי מספר הדגימות שתויגו באופן שגוי ע"י מודל random forest עם 200 עצים בעומק 100, השוואה בין איטרציות האימון השונות

כלומר, ניתן לראות כי כאשר משווים בין איטרצית האימון השניה (אימון על דאטא מאוזן) לבין איטרצית האימון השלישית (אימון על דאטא נטול דגימות חסרות נושא) מקבלים שיפור משמעותי במספר הדגימות שתויגו באופן שגוי כחסרות נושא, אך החמרה משמעותית במספר הדגימות שתויגו באופן שגוי בנושא קונקרטי. הנ"ל נכון בין אם משתמשים בתוצאות המודל מהאיטרציה השלישית כמו שהן, ובין אם מחליפים תיוגים בעלי הסתברות נמוכה (במקרה זה הרף שנבחר הוא 0.2) בתיוג "ללא נושא".

כפי שהוסבר בפרק "הערכת ביצועי המודל", שיוך של ציטוט לנושא קונקרטי שגוי הנה חמורה יותר מתיוג שגוי של ציטוט בתור "חסר נושא". לכן, למרות השיפור הניכר במטריקות כתוצאה מהאימון על דאטא נושאי בלבד, בעקבות העלייה העצומה (של יותר מ-200%) בכמות הדגימות ששויכו באופן שגוי לנושא קונקרטי, הוחלט שלא להשתמש בשינוי שנבחן באיטרציה מחקרית זו. לכן, נבחר מודל סופי מבין המודלים שנבחנו באיטרצית האימון השניה.

המודל הנבחר

המודל שנבחר הוא מודל רגרסיה לוגיסטית (עם ההיפר-פרמטר C=2) שאומן על דאטא מאוזן (1,532 דגימות, כאשר לכל אחת מ-9 התגיות האפשריות יש בין 99 ל-214 דגימות בדאטא) שעבר וקטוריזציה באמצעות Bag of Words. מודל זה נבחר כיוון שהוא השיג את ציון ה-accuracy הגבוה ביותר מבין המודלים והקונפיגורציות שנבדקו, ובנוסף השיג את תוצאת ה-f1-score השלישית בטיבה (בהפרש של אחוז אחד מהמודל שהשיג את התוצאה המיטבית, שהוא Random forest עם 200 עצים בעומק מקסימלי של 100, שאומן גם הוא על BoW), וכן מספר הטעויות החמורות (כמפורט בפרק הערכת ביצועי המודל) היה נמוך מאוד (7.59%). לפירוט אודות תוצאות המודל, פנו לפרק הבא.

תוצאות

עבור המודל שנבחר, התקבלו התוצאות הבאות:

0.540717	accuracy
0.76857	macro precision
0.517694	macro recall
0.564592	macro f1

איור 7– תוצאות המטריקות עבור המודל כולו

ניתן לראות שרוב הציטוטים תויגו ע"י המודל באופן מדויק (לפי מדד ה-accuracy). בנוסף, ערך ה-precision של המודל מעיד שרובן המוחלט של התגיות שנתן המודל היו אמינות. נעמיק במטריקות אלו עבור כל אחת מהמחלקות בנפרד:

	precision	recall	f1-score
בטחון	1	0.26	0.42
בטחון פנים	1	0.27	0.43
בלי	0.17	0.69	0.28
בריאות	0.62	0.5	0.56
חינוך	0.81	0.73	0.77
כלכלי	0.76	0.56	0.64
נשים	0.95	0.6	0.73
קורונה	0.92	0.7	0.79
רווחה	0.68	0.35	0.47

איור 8 – תוצאות המטריקות עבור כל אחת מהתגיות

ניתן לראות שהושגו ערכי precision מרשימים – עבור כל הנושאים הקונקרטיים הושגו ערכי precision של 0.62 ומעלה, כאשר עבור שתיים מהן (בטחון ובטחון פנים) התקבלו ערכי precision מושלמים (אם כי חשוב לציין שמדובר במחלקות עם מספר דגימות נמוך אשר רובן תויגו כחסרות נושא, והשאר בתגית הנכונה – מה שהשאיר לתגית הנכונה מספר חד ספרתי של דגימות עליהן התבססה מטריקה זו).

באשר למטריקת ה-recall – אמנם ישנן מחלקות אשר קיבלו ערכי recall נמוכים, אך חשוב לשים לב כי מדובר במחלקות שכמות הדגימות שלהן היא קטנה מאוד – למשל, מחלקות בטחון ובטחון פנים הכילו 19 ו-22 דגימות test בהתאמה, ולכן כל טעות בתיוג של ציטוט בודד משפיעה בצורה משמעותית על מטריקה זו. בנוסף חשוב לזכור שרוב הבלבול הוא כתוצאה מתיוג שגוי של ציטוטים בתור חסרי נושא (ראו הרחבה בהמשך הפרק על הבלבול בין נושאים קונקרטיים).

תגית "ללא נושא" קיבלה ערך precision מאוד נמוך (שמעיד על כך שהתיוג מאוד לא אמין – רובם המוחלט של הציטוטים שהמודל סימן כחסרי נושא הנם בעלי נושא קונקרטי), אך ערך recall גבוה יחסית (המעיד על כך שהמודל כמעט ולא שייך נושא קונקרטים לציטוטים חסרי נושא – למעשה, זה קרה עבורה 10 ציטוטים בלבד).

נתמקד בתיוגים שגויים בהם המודל שייך ציטוט לנושא קונקרטי כלשהו:

תגית אמת	תגית שנחזתה	כמות ציטוטים
קורונה	בריאות	4
בריאות	רווחה	3
בלי	כלכלי	3
בטחון פנים	כלכלי	3
נשים	בריאות	3
בלי	חינוך	3
בלי	קורונה	2
בטחון	חינוך	2
רווחה	כלכלי	2
בלי	בריאות	1
בלי	רווחה	1

בטחון	כלכלי	1
חינוך	רווחה	1
כלכלי	רווחה	1
בטחון פנים	רווחה	1
נשים	חינוך	1
נשים	רווחה	1
בטחון פנים	נשים	1
רווחה	בריאות	1
רווחה	חינוך	1
סה"כ		36

איור 9 – ספירת כמות הציטוטים שתויגו באופן שגוי בתגית של נושא קונקרטי (שאינו "ללא נושא")

הנתון המשמעותי העולה מהטבלה הוא שבסה"כ, מתוך 474 ציטוטים ששימשו כדגימות ה-test, רק 36 מתוכם תויגו באופן שגוי כמשויכים לנושא קונקרטי כלשהו. כלומר, **92.4% מהציטוטים עליהם נבחן המודל שויכו לנושא הנכון או סומנו כחסרי נושא**. כלומר, היקף הטעות שסומנה כחמורה ביותר צומצם ל-7.59% בלבד.

בטבלה מעלה ניתן לראות שישנם נושאים אשר יוצרים בלבול מובן עבור המודל, ככה"נ בגלל הקרבה הרעיונית ביניהם אשר עשויה לגרום לכך שהציטוטים המשתייכים אליהם מכילים מילים דומות.

לדוגמא, המודל חווה בלבול בין הנושא "קורונה" לבין הנושא "בריאות" (אם כי מעניין לראות שציטוטים בנושא קורונה תייגו כציטוטים בנושא בריאות, אך זה כמעט ולא קרה בכיוון ההפוך – יתכן כי הסיבה היא שקורונה היא למעשה תת נושא של נושא הבריאות המאופיינת באוצר מילים יותר ספציפי ואינדיקטיבי). למשל, אחד הציטוטים לגביו התבלבל המודל הוא *"אנחנו היום נדבר על תיקון נוסף שהוצע על ידי משרד הבריאות, תיקון לגבי צו בידוד והארכה שלו. אני מבקשת מנציגי משרד הבריאות להציג את מהות התיקון, על מה אנחנו דנים היום ומה אנחנו הולכים לאשר או לא לאשר. בבקשה."* סביר כי לקורא אנושי ככה"נ היה מובן כי האזכור לגבי צוי בידוד מתייחס לקורונה, אך זה לא טריויאלי, ולכן מובן למה המודל שייך את הציטוט לנושא בריאות.

ניתוח ומסקנות

במסגרת פרויקט זה, על מנת לענות על שאלת המחקר פותח מודל multiclass topic classifier, המקבל כקלט ציטוט מפרוטוקול פרלמטרי בעברית, ומשייך לו תגית המזהה אותו עם נושא מבין 9 נושאים אפשריים (8 מהם קונקרטיים, ואחד מהם מייצג העדר נושא).

המודל הצליח להבחין בצורה טובה בין המחלקות המייצגות נושאים קונקרטיים – רוב הבלבול של המודל היה בין מחלקות המייצגות נושאים קונקרטיים למחלקה המייצגת העדר נושא. למעשה, ערך ה-precision הממוצע עבור המחלקות המייצגות נושאים קונקרטיים הוא 0.84 (כאשר עבור שתיים מהמחלקות אף הושג ערך precision של 1, כלומר כל הדגימות בדאטא ה-test המשויכות לנושא זה תויגו באופן מדויק). המשמעות היא שאם המודל נותן תיוג של נושא קונקרטי, מדובר בתיוג אמין (מה שחשוב כדי לדעת שהקרדיט לדובר הציטוט על העיסוק באותו נושא הוא אמיתי). ערך ה-f1 הממוצע עבור הנושאים הקונקרטיים הוא 0.6, ומושפע מכך שישנם מספר נושאים שלקו בערך recall נמוך בין היתר בגלל שמספר הדגימות עבורן היה נמוך מאוד. בנוסף, נציין כי 92.4% מהציטוטים עליהם נבחן המודל שויכו על ידו לנושא

הנכון או סומנו כחסרי נושא, כלומר אחוז הציטוטים ששויכו לנושא קונקרטי שגוי (הטעות שהוגדרה כחמורה ביותר) הנו נמוך מאוד.

המודל התקשה ללמוד את מחלקת "ללא נושא" – מחלקת "ללא נושא" היתה המחלקה בעל ערך ה-precision הנמוך ביותר בפער – 0.17. נראה שהמודל התקשה מאוד להכליל מחלקה זו, אך אין זה מפתיע – מהגדרת המחלקה, מדובר במחלקה מאוד אקלקטית המכילה אוסף ציטוטים שאין ביניהם מכנה משותף ברור. באופן טבעי מהאופן שבו מתנהל שיח אנושי, רוב הציטוטים בפרוטוקולים לא השתייכו לנושא קונקרטי, ולכן התקבל דאטא מאוד לא מאוזן שתרם לחומרת הבעיה, אך נראה שאיזון הדאטא כך שיכיל כמות מצומצמת יותר של דגימות ממחלקה זו סייע בשיפור בעיה זו.

המודל התבלבל גם עבור ציטוטים שהיו מאתגרים עבור מתייג אנושי – חלק מהציטוטים אותם התבקש המודל לתייג הם ציטוטים שיהיו מאתגרים גם עבור מתייג אנושי. למשל, ציטוט העוסק בטיפול בנערים ונערות בסיכון ומתמקד במסגרת יעודית לנערות, תויג ע"י מתייג אנושי כמשויך לנושא "נשים", והמודל שייך אותו לנושא "רווחה" – מדובר בבלבול סביר אשר סביר והגיוני שכן באותה מידה הציטוט הנ"ל יכל להיות משויך ע"י המתייג האנושי לנושא "רווחה".

לסיכום, המודל הצליח להשיג תוצאות מרשימות בהתחשב בעובדה שכמות הדאטא עליו אומן היתה קטנה מאוד (פחות מ-250 דגימות לכל נושא), וכיוון שהיקף הטעות שהוגדרה כחמורה הוא מצומצם מאוד, מתאפשר שימוש במודל כפי שהוא לצורך לשמו פותח. עם זאת, על מנת לשפר את יכולות המודל (ובפרט לצמצם את הטעות הפחות חמורה, שהיא לתייג ציטוט בעל נושא קונקרטי כחסר נושא) יש לבצע מחקר המשך (להצעות להמשך פיתוח ומחקר פנו לפרק הבא).

הצעות להמשך פיתוח ומחקר

להלן פירוט הרחבות אפשריות לפרויקט אשר עשויות לשפר את הדיוק של המודל שפותח או להרחיב את הכיסוי והשימושים האפשריים שלו:

הרחבת המודל כך שיתמוך בזיהוי נושאים נוספים: כמוזכר לאורך המסמך, המודל שפותח במסגרת הפרויקט מאפשר זיהוי של נושאים המשתייכים לשמונה נושאים נבחרים. לפי מפתחי הפרויקט "בטא מחוקקים", נתוני השימוש באתר מראים שישנם נושאים אשר ניכר שיש בהם עניין ציבורי רב (דוגמת חרדים או דיור) שאינם נכללים ברשימת הנושאים שהמודל תומך בהם כיום. עבור חלק מהנושאים, למשל נושא הדיור, קיימים ציטוטים אשר תויגו בשלב הראשון של הפרויקט כמשויכים לנושא זה, ואף קיימת כמות מצומצמת מתוכם שגם תויגו ידנית כמשויכים לנושא. בהנתן משאבים מתאימים, ניתן יהיה לתייג ידנית כמות משמעותית של ציטוטים בנושאים נוספים ולאמן את המודל מחדש, ובכך לאפשר הרחבת התמיכה כך שתכלול עוד נושאים רלוונטיים.

טיפול בקטיעת ציטוטים: כמוזכר בפרק המבוא, בפרויקט זה ציטוט הוגדר בתור "ציטוט מתוך מסמך, המייצג את דבריו *הרציפים* של דובר מסוים שנכח בישיבה". מטבעו של שיח ישראלי בנושאים להוטים, פעמים רבות דבריו של דובר מסוים נקטעים ע"י קריאות מהקהל (כגון "שקרנית!", "נו" וכו') שקוטעות את רצף הציטוט בפרוטוקול ומפרידות אותו לשני ציטוטים שונים, למרות שבמציאות מדובר בציטוט אחד. על מנת לשפר את איכות המודל, ניתן להוסיף לשלב העיבוד המקדים איחוד של ציטוטים מופרדים שכאלו לכדי

ציטוט אחד. למשל, ניתן לעשות זאת באמצעות זיהוי שלושה ציטוטים רצופים באותו מסמך כך שהציטוט הראשון והשלישי שייכים לאותו הדובר, והציטוט השני קצר מ-10 מילים.

המשך תיוג ואימון מחדש על כמות דאטא גדולה יותר: משימת קלסיפיקציה עם מספר תגיות גבוה (במודל הנוכחי 9 תגיות, אך עם הרחבת המודל בהמשך תתכן הגדלת מספר הנושאים הנתמכים) הנה משימה מורכבת, שהיכולת להצליח בה תגדל אם כמות הדאטא עליו המודל מאומן (שהיא כרגע מעטה מאוד) תגדל. בהנתן משאבים וכוח אדם, ניתן יהיה לתייג כמות משמעותית נוספת של דגימות אימון במגוון הנושאים, ובכך לשפר את יכולת המודל ללמוד את כל אחד מהנושאים.

שימושים ושיתופי פעולה אפשריים

אמנם הפרויקט עוסק בבעיה ממוקדת מאוד, אך המודל שפותח במסגרתו עשוי לשמש למטרות רבות:

הטמעה בפרויקט "בטא מחוקקים": כפי שהוזכר לאורך המסמך, הפרויקט בוצע בשיתוף פעולה עם פרויקט "בטא מחוקקים" מתוך מטרה להשתלב בפרויקט במטרה ולספק מודל אמין לסיווג ציטוטים לנושאים נבחרים. בחודשים הקרובים במסגרת פיתוח ופרסום הגרסא החדשה של אתר "בטא מחוקקים" לקראת הבחירות המתקרבות, צפויים נסיונות התאמה ושכלול של המודל שפותח במסגרת פרויקט זה כך שיתאים לשימוש באתר.

ייעול פעולת ועדות הכנסת: בנוסף, ניתן להשתמש במודל שפותח לשימושים שונים מלבד הערכת ומדידת תחומי העיסוק של חברי כנסת בודדים. לדוגמא, ניתן לסווג לנושאים את הציטוטים מישיבה של ועדה מסוימת בכנסת (למשל ועדת החינוך, התרבות והספורט) ובכך לאמוד את אחוז הזמן מישיבת הועדה שאכן עסק בנושא לשמה התכנסה. מידע שכזה (בהצמדה לפרמטרים נוספים כמו הרכב הדוברים, תאריך הפגישה וכו') יוכל להפיק תובנות משמעותיות על התנהלות הועדה (לדוגמא – "בישיבות בהן נכחו יותר מ-3 אורחים אחוז הזמן בו הדיון עסק בחינוך עלה" או "בישיבות שנערכו ביומיים שלאחר הצבעה בנושא חוק המשויך לועדה, אחוז הזמן בו הדיון עסק בחינוך ירד"). תובנות שכאלו יוכלו לשמש את יו"ר ומנהלי הועדה לתכנן בצורה מיטבית את ישיבות הועדה כדי לנצל בצורה מקסימלית את הזמן היקר שלהן.

הנגשת מידע על מועמדים חדשים בפריימריז: כמו כן, ברוח הבחירות המתקרבות, ניתן להשתמש במודל על מנת לנתח את המיקוד של מועמדים חדשים שטרם היו חברי כנסת בעבר, באמצעות ניתוח ציטוטים שלהם מראיונות או מפרסומיהם ברשתות החברתיות. ניתוח שכזה יוכל לאפשר מועמדים אשר היו להם התבטאויות בנושאים רלוונטיים עוד לפני שהחליטו לרוץ לכנסת. ניתוח שכזה יוכל להוביל להשוות בין מועמדים חדשים למפלגות, ולהוביל לתובנות דוגמת "מועמד א' מאוד פעיל ברשתות החברתיות אבל התחיל לדבר על חינוך רק לאחר שהגיש את מועמדותו לפריימריז, לעומת מועמד ב' שמתבטא בנושאי חינוך מזה 5 שנים" שיעניקו למתפקדי המפלגות מידע משמעותי שיתרום ליכולתם להצביע בפריימריז בצורה מושכלת.

ניתוח התבטאויות של אנשי ציבור שאינם חברי כנסת: שימוש אפשרי נוסף יהיה בחינת שימוש במודל המאומן על ציטוטים מעולמות תוכן קרובים. למשל, יתכן שהמידע שנלמד ע"י המודל יאפשר לתייג ציטוטים מתמלול ראיונות או בהתבטאויות ברשתות החברתיות של נושאים בתפקידים ציבוריים שאינם חברי כנסת (למשל שופטים או אנשי מערכות הבטחון והאכיפה) על מנת לבחון על אילו נושאים הם בוחרים להתבטא (למשל, כמה מפכ"ל המשטרה מדבר בראיונות על חינוך? איזה רמטכ"ל נאם יותר על זכויות נשים ולהט"ב? אילו שופטים צייצו בנושא הקורונה?). ניתוח שכזה יגדיל את הנגישות של התבטאויותיהם של דמויות ציבוריות לאזרחים המושפעים מעשייתם ומעמדותיהם.