

## יעקב הכהן-קרנר, דרור מוגהץ, חנניה בק, אלחי יהודאי

### סיווג אוטומטי של פסקי שו"ת

מאמר זה מתאר מערכת מחשב לומדת המסווגת באופן אוטומטי פסקי שו"ת הכתובים בעברית-ארמית. הסיווגים בוצעו לפי עדת הפוסק ו/או תקופת החיבור באמצעות מילים. ביצועי המערכת שופרו באמצעות אלגוריתם למידה אוטומטי בשם Support Vector Machines. המערכת הורצה על אוסף המכיל מעל 12,000 פסקי שו"ת במספר ניסויים: עדות, תקופות, עדות ותקופות. בכל הניסויים הצליחה המערכת לסווג נכונה מעל 95% מן הקבצים. באמצעות ניסויים אלו אפשר לזהות הבדלים בולטים ומעניינים המבחינים בין הפסיקה הספרדית לפסיקה האשכנזית ובין פסיקות מתקופות ישנות יותר לפסיקות של אחרוני זמננו. תוצאות אלו הינן בעלות ערך לחוקרי פסקי שאלות ותשובות ולחוקרים מתחום מדעי החברה החוקרים הבדלים בין תרבויות שונות והבדלים בין תקופות שונות.

#### א. מבוא

ספרות השו"ת (שאלות ותשובות) בנושאים הלכתיים תפסה לה מקום חשוב בכותל המזרח של הספרות התורנית החל מתקופת הגאונים, עבור בתקופת הראשונים, האחרונים ועד אחרוני האחרונים פוסקי זמננו. אפשר לסווג פסקי שו"ת אלו לפי מגוון אפשרויות. מחקר זה עוסק בסיווג פסקי שו"ת לפי עדת הפוסק (ספרדי או אשכנזי) ו/או לפי תקופת החיבור. באמצעות סיווגים אלו ניתן בין היתר, לזהות הבדלים בולטים, המבחינים בין הפסיקה הספרדית לפסיקה האשכנזית ובין פסיקות מתקופות ישנות יותר לפסיקות של אחרוני זמננו.

סיווג אוטומטי של קבצי טקסט (text categorization) הינו נושא מחקרי התופס תאוצה בשנים האחרונות. נושא זה מהותי למשימות רבות, לדוגמה: סוכנויות ידיעות המקבלות מספר רב של ידיעות מדי יום שיש לסווגן לפי תחומים שונים (כגון: דת, כלכלה, פוליטיקה, משפט וספורט). סיווגים דומים נדרשים גם לסוכנויות מודיעין, עורכים של קבצי מאמרים ועוד. ישנן מערכות איכותיות לסיווג אוטומטי בשפות זרות שונות, רובן נוצרו עבור השפה האנגלית. עד כה, נעשה בשפה העברית בכלל ובטקסטים תורניים בעברית בפרט, מחקר מצומצם עבור סיווג אוטומטי. הפרוייקט המוצע הינו תכנון ובנייה של מודל סיווג אוטומטי איכותי בקנה-מידה אנושי של טקסטים באמצעות למידה ממוחשבת בתחום של טקסטים הכתובים בשפה

\* אנו מודים לשופט האנונימי על הערותיו והארותיו.

יעקב הכהן-קרנר, דרור מוגהץ, חנניה בק, אלחי יהודאי

העברית בכלל ובטקסטים תורניים הכתובים בשפה העברית-ארמית בפרט. תוצאות המחקר עשויות להיות בעלות ערך לחוקרים מתחום מדעי היהדות החוקרים פסקי שו"ת ולחוקרים מתחום מדעי החברה החוקרים הבדלים בין תרבויות שונות והבדלים בין תקופות שונות.

## ב. רקע תיאורטי

סיווג טקסטים הינו משימת למידה המסווגת מסמך נתון הכתוב בשפה טבעית לקטגוריה אחת או יותר ממכלול קטגוריות מוגדרות מראש לפי תוכן הטקסט [21]. סיווג טקסטים אוטומטי נדרש למגוון רחב של יישומי מחשב, כגון: סיווג ידיעות מודיעיניות בטחוניות לפי מקור הידיעה וסיווג ידיעות חדשותיות לפי תחום הידיעה. לכן, סיווג בצורה ממוחשבת אוטומטית ויעילה עתיד לחסוך כוח אדם ומשאבי מחשב. במיוחד בימינו – בדור התפוצצות המידע האלקטרוני והתפתחויות בצעדי ענק בתחומי התקשורת, מסדי הנתונים והאינטרנט, גדל הצורך בסיווג טקסטים אוטומטי ויעיל.

בנוסף, סיווג טקסטים הוא משימה מאתגרת בגלל כמות הטקסטים העצומה הקיימת, מספר המאפיינים הרב הנמצא בטקסטים והתלות בין המאפיינים. בכדי לסווג טקסט בהצלחה וביעילות יש להגדיר מאפיינים המתאימים לאופי משימת הסיווג, כך שהערכים למאפיינים אלו נלקחים מתוך הטקסט, באמצעות ניתוחו.

התחלת תחום המחקר בנושא סיווג טקסטים מיוחסת לעבודתו של Maron [20] בתחום של סיווג טקסטים הסתברותי. מאוחר יותר ניסו מספר חוקרים [12, 13, 16, 20, 21, 22] להגדיר מלכתחילה מספר מצומצם של מועמדים מבטיחים כמאפיינים מוצלחים (כגון: שכיחותן של מילים רלוונטיות בוודאות למשימת הסיווג הנידונה), וזאת בעזרת ידע מוקדם שנשאב ממומחים רלוונטיים לתחום הסיווג. כיום אחת הגישות העיקריות דורשת להגדיר מלכתחילה מספר רב של מאפיינים אפשריים, ותוך כדי ביצוע תהליך הלמידה עבור סיווג הטקסטים לסנן את קבוצת המאפיינים. מטרתו של תהליך זה להותיר את המאפיינים הרלוונטיים והיעילים ביותר עבור משימת הסיווג [3, 7].

בימינו, סיווג טקסטים מיושם במשימות רבות, כגון: אישכול (סיווג טקסטים לקבוצות אופייניות שונות, שאינן מוגדרות מראש), סינון מסמכים (סיווג טקסטים לפי קריטריונים מוגדרים מראש), איחזור מידע (איחזור מידע רלוונטי ממסמך או ממאגר-נתונים בהתאם לרלוונטיות שלו), חילוץ מידע (הוצאת מידע רלוונטי מתוך טקסט כמות שהוא בהתאם לרלוונטיות שלו) והתרת רב-משמעות של מילים (פענוח המשמעות הנכונה של מילים רב-משמעיות, שזה למעשה סיווג כל מילה רב-משמעית למשמעותה הנכונה מבין המשמעויות האפשריות).

### סיווג לפי סגנון לעומת סיווג לפי תוכן

אחת החלוקות העיקריות של סיווג טקסטים היא לשתי המשפחות הבאות: סיווג לפי תחום הטקסט או סיווג לפי סגנון הכותב. בכל משפחה ניתן למצוא משימות סיווג רבות, לצורך מטרות

ויישומים שונים. לדוגמה:

- משימות זיהוי תחום העיסוק הכללי של הטקסט (כגון: כלכלה, ביטחון, ספורט), וזיהוי הנושא המדויק של הטקסט (כגון בתחום הספורט: כדורגל, כדורסל, שחייה) שייכות למשפחת הסיווג לפי תוכן הטקסט.
- משימות זיהוי סגנון של טקסט (כגון: חדשותי, ספרותי ומדעי), וזיהוי סוג הטקסט (כגון: מדובר או כתוב? מעיתון או מספר?) שייכות למשפחת הסיווג לפי סגנון הכותב.

תהליך המחשבה המופעל על ידי מסווג אנושי לצורך סיווג שונה בכל משפחה. מבחינה מעשית, משמעות המאפיינים הנבחרים לצורך הסיווג שונה בכל משפחה, וכתוצאה מכך שונה גם תהליך בחירת המאפיינים. בסיווג לפי תחום בדרך כלל המאפיינים הרלוונטיים קשורים לתוכן ולמשמעות של הטקסט, ויש צורך להתעלם ממאפיינים סגנוניים שאינם קשורים לתוכן. לעומת זאת, בסיווג לפי סגנון, הגישה בדרך כלל הפוכה, המאפיינים הינם בדרך כלל סגנוניים, ויש להתעלם ממאפיינים הקשורים לתוכן [3].

לדוגמה, בסיווג לפי תוכן, המאפיינים יכולים להיות מבוססים (בין היתר) על קבוצות ביטויי מפתח, מתוך הנחה שמסמכים בנושאים שונים מכילים מילות מפתח שונות (למשל: מסמכים בנושא הכלכלה יכילו מילים כמו – כסף, תוספת יוקר, מדר, אינפלציה וכו', ומסמכים בנושא הספורט יכילו מילים כמו – כדור, שופט, עבירה, אצטדיון ופסק זמן). לעומת זאת, בסיווג לפי סגנון, המאפיינים הם בעיקרם בלשניים, לשוניים, כמותיים, מורפולוגיים (נטיות המילה), אורתוגרפיים (צורת הכתיב) וסינטקטיים (מבנה המשפט).

#### למידה חישובית

למידה חישובית היא נושא חשוב בתחום הבינה המלאכותית. נושא זה עוסק ביכולתו של המחשב ללמוד משגיאות ומהצלחות העבר, ולשפר את ביצועיו בכוחות עצמו. הלמידה החישובית עוסקת בפיתוח אלגוריתמים הלומדים מאוסף דוגמאות נתונות ומסיקים בעזרתן באופן אוטומטי החלטות לגבי מקרים חדשים.

חשיבות המחקר בתחום של למידה חישובית נובעת מארבע סיבות עיקריות: (1) חיקוי הצלחתה של האינטליגנציה האנושית; (2) חיסכון בשנות-אדם ובכסף בפיתוח ובתחזוק מערכות אינטליגנטיות; (3) חוסר יכולת של מערכות קיימות לבצע הכללות על תיאוריות והסברים ו-(4) חיקוי מודלים של למידה אנושית שיאפשרו להבין את תהליכי עיבוד המידע וההתנהגות של בני האדם.

באופן כללי, מערכת למידה מקבלת החלטה ביחס למקרה שאינו מוכר לה, על בסיס למידת מקרים קודמים שבהם כבר סופקה למערכת ההחלטה הנכונה שנקבעת לרוב על ידי מומחה אנושי. אחד הסוגים העיקריים של למידה ממוחשבת הוא הלמידה המונחית (supervised learning). בלמידה מסוג זה ניתן להשוות את החלטות המערכת עבור מקרים חדשים להחלטות

יעקב הכהן-קרנר, דרור מוגהץ, חנניה בק, אלחי יהודאי

הנכונות שצריכות להיות עבור מקרים אלו (למשל כאלו המסופקות על ידי מומחה). בעקבות ההשוואה, במידת הצורך, תהליך הלמידה מבצע שינויים במנגנון הלמידה, כלומר – לומד.

מקובל לחלק את תהליך הלמידה המונחית לשלושה שלבים נפרדים: למידה, אימות ויישום. בשלב הלמידה (learning) המערכת מקבלת קבוצת-אימון (training set). זהו אוסף דוגמאות שבו כל דוגמה היא מקרה יחיד של הבעיה המדוברת, העומד בפני עצמו והמורכב משני חלקים: ערכי מאפיינים המייצגים את המקרה, וההחלטה הנתונה לגביו. מערכת הלמידה באמצעות שיטת למידה מוגדרת מראש תנסה למצוא מודל הסקה שיחזה את הקשר בין ערכי המאפיינים השונים של המקרים שהיו ובין ההחלטות שנתקבלו במקרים אלו.

בשלב האימות (validation) מבוצע אימות של איכותו של מודל ההסקה שנבנה בשלב הקודם. לצורך בחינת מודל ההסקה מזינים את המערכת באוסף מקרים שההחלטה האמתית לגביהם (שנקבעה על ידי מומחה אנושי) אינה נתונה לשימוש המערכת בתהליך הלמידה. הואיל והתוצאות האמתיות מאוחסנות, ניתן להשוותן לתוצאות מודל ההסקה ולמדוד את הצלחת הלמידה בקבוצת מקרים זו. אוסף המקרים שעליהם נבחן מודל ההסקה נקרא קבוצת-מבחן (test set).

מקובל להפריד בין קבוצת האימון וקבוצת המבחן. ההפרדה מביאה למידה אמינה יותר: הצלחה גבוהה של המערכת בקבוצת מבחן שאינה תלויה בקבוצת האימון מעידה על כלליותו ונכונותו של מודל ההסקה. לעומת זאת, אם אין הפרדה בין קבוצת האימון וקבוצת המבחן, אזי המדידה אינה אמינה, משום שחלק ניכר מהצלחתה של המערכת מתקבל מפני שביצוע שלב האימות נעשה על קבוצת מבחן שהמערכת מכירה כבר משלב הלמידה.

כדי שכל מסד הנתונים ישמש גם לאימון וגם לבחינה מבלי לסתור את הדרישה להפרדה המוחלטת בין קבוצת האימון לקבוצת המבחן, משתמשים לרוב בשיטת K-Fold Cross Validation, המקובלת מאוד בתחום מדידת ההישגים בכלל ובמחקרים במדעי המחשב בפרט. על פי שיטה זו יש לבצע את הלמידה באופן הבא:

1. חלק את כלל  $n$  דוגמאות הלמידה ל- $K$  קבוצות זרות ושוות בגודלן ככל האפשר ( $K$  הוא מספר שלם בין 2 ל- $n$ ).

2. בצע  $K$  ניסויים, כאשר עבור כל ניסוי  $1 \leq i \leq K$ , הקבוצה ה- $i$  תשמש כקבוצת המבחן לבדיקת מידת נכונותו של מודל ההסקה, ושאר  $K-1$  הקבוצות ישמשו כקבוצת האימון עבור בניית מודל ההסקה.

3. תוצאת הניסוי הכולל  $E$  מחושבת כממוצע תוצאות כל  $K$  תתי הניסויים  $E_1, E_2, \dots, E_K$ .

שימוש בשיטה זו יוצא נשכר משתי זוויות-ראייה: מחד גיסא, בכל ניסוי קבוצת המבחן מנותקת לגמרי מקבוצת האימון, ומאידך גיסא, בראייה כוללת מסד הנתונים כולו משמש גם לאימון וגם לבחינה. הערכים המקובלים ביותר במחקר עבור  $k$  הינם 5 או 10. הערך 3 מספיק עבור מסדי נתונים גדולים מאוד יחסית והערך  $n$  מקובל עבור מסדי נתונים קטנים מאוד יחסית.

## סיווג אוטומטי של פסקי שו"ת

בשלב היישום (implementation), לאחר שהסתיימה הלמידה והופק מודל הסקה בעל איכות ידועה, ניתן להציג למערכת קלט של מקרה המיוצג על ידי מאפיינים בלבד, ולקבל כפלט החלטה כלשהי לגביו, אליה תגיע מערכת הלמידה בצורה ישירה על פי מודל ההסקה מבלי צורך להפעיל שוב את שיטת הלמידה.

### שיטות למידה ממוחשבות

שיטת למידה ממוחשבת הינה אלגוריתם להפקת מודל הסקה מתוך קבוצת אימון. לאורך השנים פותחו שיטות למידה רבות המתאפיינות בביסוסן המתמטי, באופן עבודתן, במהירות הלמידה שלהן, בסוג הנתונים בהם הן מטפלות ובאיכות תוצאותיהן. שיטות הלמידה משתייכות לסוגים שונים, ובהם: שיטות למידה המבוססות על סטטיסטיקה והסתברויות, למידה אינדוקטיבית, למידה מבוססת כללים, למידה מבוססת תקדימים, אלגוריתמים גנטיים, פונקציות הפרדה מתמטיות, רשתות עצביות ועוד. למעשה, לא קיימת שיטת למידה שהיא תמיד "הטובה ביותר", ולכל שיטה ישנם מקרים בהם היא טובה, וישנם מקרים אחרים בהם היא איננה טובה. סקירה נרחבת על מגוון שיטות למידה ניתן למצוא בין היתר ב-[26].

### מערכות קודמות רלוונטיות בתחום סיווג הטקסטים

להלן סקירה של מספר מערכות הרלוונטיות לתחום המחקר, כלומר לסיווג טקסטים תורניים שנכתבו בעברית-ארמית.

#### מערכת CHAT

מערכת CHAT [17, 18] סווגה לפי סגנון טקסטים תורניים הכתובים בשפה העברית-ארמית. מערכת זו ביצעה מספר ניסויים השייכים למשפחת הסיווג לפי סגנון, ניסויים אלו נעשו על טקסטים תורניים הכתובים בשפה העברית-ארמית ומטרתם הייתה לגלות דברים הקשורים לזהות המחבר ולתקופת החיבור.

המערכת התמודדה עם שלוש משימות הסיווג הבאות:

- (1) אימות המחבר – האם שני ספרי פסיקה נכתבו על ידי אותו מחבר?
- (2) כרונולוגיית החיבור – האם מאמר קדם למאמר אחר? ומהי מערכת היחסים בין מאמרים?
- (3) ייחוס סופרים – בהינתן קטע של טקסט מסוים, מוצאים מאיזה טקסט מקורי הוא נלקח מבין מספר טקסטים אפשריים.

לגבי כל אחת משלוש המשימות לעיל בוצע ניסוי מתאים, כמתואר להלן.

#### ניסוי 1 – אימות המחבר

ישנו ספר הנקרא "תורה לשמה" [2] המכיל פסקי שו"ת. ספר זה נכתב תחת השם הספרותי – יחזקאל כחלי (שם מחבר לא ידוע בתחום הפסיקה). ההערכה היא שכתבו הבן איש חי (פוסק

ששמו רבי יוסף חיים בן אליהו אל-חכם, שחי בעירק במאה התשע-עשרה). מטרת הניסוי הייתה להחליט האם ספר זה אכן נכתב על ידי הבן איש חי שניסה להסתתר תחת שם ספרותי. לצורך הניסוי נבנה בסיס נתונים שהכיל 2244 פסקי שו"ת לפי החלוקה הבאה:

1. 524 מסמכים מהספר "תורה לשמה" שלגבי זהות מחברו מסתפקים.
2. 509 מסמכים מ-[1] שידוע בוודאות שנכתב על ידי הבן איש חי.
3. 1211 מסמכים מארבעה ספרי שו"ת נוספים ("זבחי צדק", "גינת ורדים", "שואל ונשאל", "דרכי נועם") שנלקחו מספרי שו"ת שמחבריהם חיו בתקופה היסטורית וסביבה גיאוגרפית הקרובות לאלה של הבן איש חי כדי שיהיה קשה לזהות הבדלים בין המחברים, וכתוצאה מכך הניסוי יהיה אמין ככל האפשר.

בניסוי זה השתמשו במאפיינים ממשפחת המאפיינים הלשוניים (lexical). בתחילה בחרו כמאפיינים את 200 המילים השכיחות ביותר בטקסט. לאחר שסיננו את כל המילים הקשורות לתוכן הטקסט, נשארו עם 130 המילים השכיחות ביותר בטקסט, והן היו המאפיינים בניסוי זה. בניסוי זה השתמשו בשיטת הלמידה Balanced Winnow (עם מנגנון 5-fold cross validation) שהינה שיטת למידה בינארית ומקוונת המעניקה משקלות למאפיינים, תוך כדי לקיחה בחשבון של אינטראקציה בין המאפיינים.

המערכת הצליחה להבדיל לחלוטין בין קבוצה 1 בבסיס הנתונים ובין כל קבוצת משנה מקבוצה 3 בבסיס הנתונים עם אחוזי הצלחה בסביבות 95%. לכאורה, מתבקשת המסקנה ששני ספרים אלו לא נכתבו על ידי אותו מחבר, וממילא, הספר "תורה לשמה" לא נכתב על ידי הבן איש חי. אך ליוצרי המערכת היה חשד, שמא ספר זה נכתב אמנם על ידי הבן איש חי, אך הוא מיסך בכוונה את כתיבתו בספר "תורה לשמה" באמצעות השתלת הבדלים סגנוניים. החוקרים ב-[17] השתמשו בשיטה חדשנית הנקראת unmasking (הסרת מיסוך) לצורך זיהוי מחבריהם של טקסטים שהעלימו את עצמם או את העתקתם באמצעות השתלת הבדלים סגנוניים. הסרת המיסוך הינה תהליך בו מנופים מאפיינים בולטים הנמצאים בתוך הטקסט. בעקבות הפעלת תהליך זה הייתה מסקנת החוקרים שאכן הבן איש חי הוא מחברו של ספר זה ולא אחד מארבעת הפוסקים האחרים. הסבר אפשרי בהחלט יכול להיות כי היו מאפיינים שהשתלו במכוון כדי להסתיר מסיבות שונות (למשל מטעמי ענווה) את זהותו האמתית של מחבר הספר "תורה לשמה". אולם, ייתכן כי ספר זה נכתב בתקופה אחרת של חייו ולכן היו הבדלים במאפיינים.

ואכן בהקדמת המו"ל לשו"ת תורה לשמה [2] מובא שאכן מקובל לזהות את רבינו יוסף חיים הבן איש חי כמחברו העלום של תורה לשמה כדלקמן: "והיה גאון בירושלים ... ושמו רבי אברהם עדס זלה"ה, ... ואמר ... יחזקאל עולה מספר יוסף, כחלי עולה מספר חיים, והוא הספר תורה לשמה של הגאון הגדול רבי יוסף חיים זיע"א, מה עשה רבי אברהם ז"ל, שלח מכתב לבגדאד למורנו ורבנו מאור הגולה הגאון רבי יוסף חיים זיע"א, ושלה לו כמה שאלות בעניין ההלכה, ושאלה האחרונה מי הוא המחבר תורה לשמה, והחזיר לו כל השאלות, חוץ משאלה האחרונה של תורה לשמה לא החזיר לו תשובה והתעלם ממנה, ונתברר יותר לרבי אברהם ז"ל,

## סיווג אוטומטי של פסקי שו"ת

שהגאון רבנו יוסף חיים זיע"א לא רצה לגלות לו, ולהעלימו מטעם הכמוס עמו". וחזוק נוסף לכך מובא בהמשך הספר הנ"ל בהקדמתו של נכד הבן איש חי: "... ומכלל החיבורים הוא חיבור הקדוש הזה אשר קרא לו בשם תורה ... וחתם בו חתימת ידו הקדושה בשם 'יחזקאל כחלי' ומגמתו היתה בשם זה להעיד על חיבורו תורה לשמה והשם הנז"ל 'יחזקאל כחלי' הוא מספר יוסף חיים בגימטריא. על כן הספר הזה יתייחס לשמו הטוב אע"ג שחתם אותו בשם הנז"ל כדי להעלימו מטעם הכמוס עמו להיות בנסתר ולא בנגלה וגם יש עוד חיבורים קדושים שלו ג"כ חתם אותם בשם אחר כי יש מן הפוסקים ז"ל שכתבו שצריך המחבר להעלים שמו ויש שכתבו שצריך המחבר להזכיר שמו, ולזאת המחבר מו"ז מו"ר ועט"ר הרב הגאון ח"ר יוסף חיים זלה"ה שהסכים לכל סברות הפוסקים הנז"ל".

### ניסוי 2 – כרונולוגיית החיבור

מטרת ניסוי זה (הנדרון גם ב-13) הייתה להכריע בספק ידוע של חוקרי ספרות הקבלה – האם שלושה חלקים שונים של ספר הזוהר נכתבו על ידי אותו מחבר, או שמא הם נכתבו על ידי מחברים שונים, ואם כן מהי מערכת היחסים בין שלושת החלקים, כלומר, איזה חלק השפיע על משנהו.

לצורך הניסוי נבנה בסיס נתונים שהכיל 214 טקסטים משלושת חלקי הזוהר הנדונים: 47 טקסטים מחלק "האידרא", 67 טקסטים מחלק "מדרש-הנעלם" ו-100 טקסטים מחלק "רעיא-מהימנא". בניסוי זה השתמשו גם כן ב-130 המילים השכיחות ביותר בטקסט, לאחר ביצוע תהליך סינון ובשיטת הלמידה Balanced Winnow עם מנגנון הלמידה 5-fold cross validation. המערכת הצליחה להבדיל בין כל זוג חלקים משלושת חלקי הזוהר ב-98 אחוזי הצלחה. מחקר זה גם אישר את מה שהיה ידוע זה מכבר לחוקרים כי בחלק "האידרא" רוב התחיליות והסופיות היו בשפה הארמית, ובחלק "מדרש-הנעלם" רוב התחיליות והסופיות היו בשפה העברית, ולעומת זאת, בחלק "רעיא-מהימנא" התחיליות והסופיות היו בשיעור שווה משתי השפות, מכאן ניתן להסיק לכאורה שחלק זה הושפע במידה דומה על ידי שני החלקים האחרים או שהוא חובר בפרק זמן המהווה מעבר כרונולוגי בין שני החלקים האחרים.

### ניסוי 3 – ייחוס סופרים

מטרת הניסוי הייתה להחליט ביחס לקטע מסוים ממסכת ראש השנה שבתלמוד הבבלי, לאיזה כתב יד מבין ארבעה כתבי יד נתונים של מסכת ראש השנה הוא שייך. יש לציין שלפני עורכי המחקר היה אך ורק תוכן ארבעת כתבי היד כקבצי מחשב, אך לא היו להם כתבי היד המקוריים שמכילים גם את צורת הכתב של המחבר.

לצורך הניסוי, נבנה בסיס נתונים שהכיל 67 קטעים מכל כתב יד מארבעת כתבי היד שעמדו לרשות המערכת. בשונה מהניסויים הקודמים, בניסוי זה לא השתמשו במאפיינים ממשפחת המאפיינים הלשוניים והמורפולוגיים (הטיות המילה), מכיוון שכאשר חוקרים מספר כתבי יד

יעקב הכהן-קרנר, דרור מוגהץ, חנניה בק, אלחי יהודאי

של אותו טקסט אי אפשר להשתמש במאפיינים אלו, משום שבכל הכתבים ישנן אותן מילים בדיוק. ולכן המערכת השתמשה בקבוצת מאפיינים ממשפחת המאפיינים האורתוגראפיים (צורת הכתיב), כגון: קיצורי מילים (גמ' = גמרא) ואיותים שונים לאותה מילה (סכה = סוכה). המערכת בהתבסס על שיטת הלמידה Naïve Bayes הצליחה לשייך קטע טקסט מסוים לכתב יד מסוים מתוך הארבעה האפשריים באחוז הצלחה בגובה 85.4%.

#### מערכת לייחוס סופרים תורניים

המערכת המוצגת ב- [3, 17] מציגה שיטות שונות לייחוס (attribution) של סופרים בספרות תורנית. מערכת זו ביצעה ניסוי השייך למשפחת הסיווג לפי סגנון, והתמודדה עם המשימה של זיהוי מחברן של תשובות הלכתיות נתונות הכתובות בשפה העברית-ארמית. מטרת הניסוי הייתה להחליט האם תשובות אלו נכתבו על ידי הרשב"א או הריטב"א. הקושי נובע מכך שהריטב"א היה תלמידו של הרשב"א ושניהם חיו בספרד במאה השלוש-עשרה. וממילא, יש להניח שלשניהם יש סגנון כתיבה דומה. לצורך הניסוי, נוצר בסיס נתונים שהכיל 209 פסקי שו"ת שנכתבו על ידי הרשב"א ו-209 פסקי שו"ת שנכתבו על ידי הריטב"א.

בניסוי זה השתמשו במאפיינים ממשפחת המאפיינים הלשוניים. בתחילה בחרו כמאפיינים את 500 המילים השכיחות ביותר בטקסט, ולאחר שסיננו את כל המילים הקשורות לתוכן הטקסט (משום שרצו לקבל מאפיינים סגנוניים שאינם קשורים לתוכן), נשארו עם 300 המילים השכיחות ביותר בטקסטים, שהיו המאפיינים בניסוי. גם בניסוי זה השתמשו בשיטת הלמידה Balanced Winnow, עם מנגנון הלמידה 5-fold cross validation.

בעזרת המודל הנלמד, המערכת הצליחה לגלות מיהו מחברה של תשובה הלכתית נתונה כאשר הרשב"א והריטב"א היו מחברים פוטנציאליים, באחוז הצלחה של 85.8%. בניסוי זה בוצע סינון מאפיינים נוסף, שבו הוצאו מאפיינים עם משקלות נמוכים מאוד, עד שהגיעו לאחוזי הצלחה אופטימליים של 90.5%.

#### רלוונטיות המערכות הקודמות למחקר המתואר במאמר זה

במערכות הקודמות לא בוצעו סיווגים מסוג: עדות, תקופות, עדות ותקופות. מחקרים אלו השתמשו במסדי נתונים קטנים יחסית (שהכילו בין 214 ל-2244 מסמכים) ובשיטות למידה הנחשבות לבסיסיות יותר: Naïve Bayes ו-Balanced Winnow. במערכות הקודמות בוצע סינון מאפיינים באופן ידני על פי שכיחות. לא בוצע סינון אוטומטי בשיטה מדעית ידועה כלשהי. להלן נתאר באופן כללי שיטות סינון מדעיות אוטומטיות.



### ג. המערכת שנבנתה ותכנון הניסויים עבורה

Forman [11] מתאר מחקר שנערך לגבי שיטות למידה שונות בשילוב תריסר שיטות שונות לסינון מאפיינים עבור משימות שונות בסיווג טקסטים. הניסויים נערכו לגבי טווח של 10 עד 2,000 מאפיינים. בעקבות מחקרו של Forman הגדרנו קבוצת ניסויים עבור המערכת באמצעות שיטת הלמידה (תוך שינויים קלים), כדלקמן:

עבור בסיס הנתונים הנחקר

עבור כל ניסוי (עדות, תקופות, עדות ותקופות)

עבור X (10, 20, 50, 100, 200, 500, 1,000, 1,500 ו-2,000) מאפיינים

שנמצאו כטובים ביותר על פי שיטת סינון המשתנים InfoGain

בצע למידה ובחינה על ידי שיטת SMO באמצעות

10-fold cross-validation

לאימות תוצאות הלמידה השתמשנו בשיטת 10-fold cross-validation בשונה משיטת 5-fold שננקטה במחקריהם של Forman, קופל ומוגהץ. שיטת ה-10-fold רצה זמן רב יותר אך מהימנה יותר הן מבחינה סטטיסטית, משום שנעשים 10 ניסויים ולא 5, והן משום שהלמידה מתבצעת עבור 90% ממסד הנתונים ולא על 80% ממנו.

בניגוד לחלק מהמחקרים הקודמים (כגון זה של Forman), במחקר זה לא בוצע סינון של מילות stop-list משום שהן חשובות לסיווג לפי סגנון, שהוא התחום המחקרי בו אנו עוסקים. בדומה ל-Forman הניסוי החל עם 2,000 מאפיינים בעלי המשקל הגבוה ביותר. בכל המערכות התורניות הקודמות בוצע סינון מאפיינים באופן ידני. אנו ביצענו סינון אוטומטי בשיטה המדעית הנפוצה של Yang and Pedersen [30] הראה כי שיטה זו היא אחת משיטות סינון המאפיינים הטובות ביותר עבור מגוון נרחב ומייצג של מסדי נתונים.

במחקר זה בוצעו שלוש משימות סיווג: (1) סיווג לפי התקופה ההיסטורית שבה המסמך נכתב. פסקי שו"ת מהמאה התשע-עשרה, המכונים כאן ישנים, ופסקי שו"ת מהמאה העשרים, המכונים חדשים (המאה העשרים הוצעה כתקופת פוסקים חדשה בפרוייקט השו"ת עצמו). (2) סיווג לפי העדה אליה משתייך מחבר המסמך: ספרדים או אשכנזים.<sup>1</sup> (3) משימה משולבת של 4 קטגוריות: ספרדים ישנים, ספרדים חדשים, אשכנזים ישנים ואשכנזים חדשים. כל משימה בוצעה על מסד נתונים המכיל 12,014 פסקי שו"ת שהורד מפרוייקט השו"ת (גרסה 12) [4], של אוניברסיטת בר אילן. פסקי שו"ת הנ"ל נבחרו כך שכל חלוקה של כל קבוצה מתוך הקבוצות השייכות לאותה חלוקה תכיל מספר שווה של פסקי שו"ת על-מנת שלא תהיה הטיה בתהליך

1 ההתייחסות לספרדים בהקשר הלכתי זה היא שמוצאם מיבשת אפריקה, יבשת אסיה ומספר ארצות בדרום-אירופה, בהן ספרד, פורטוגל, איטליה, יוון ובלגריה. אשכנזים מקורם בעיקר במרכז אירופה (צרפת וגרמניה) ומזרח אירופה (פולין, ליטא, רוסיה וכו').

יעקב הכהן-קרנר, דרור מוגהץ, חנניה בק, אלחי יהודאי

הלמידה. פסקים אלו נכתבו על ידי 48 פוסקים, כאשר כל פוסק כתב בממוצע 250 פסקי שו"ת ממסד הנתונים. מספר המילים בכל המסמכים במסד הנתונים הינו כ-18.7 מליון! כל מסמך מכיל בממוצע 1,557 מילים. נתונים סטטיסטיים על מסד זה ניתן לראות בנספח. מסמכים אלו הינם פסקי שו"ת שנכתבו בשפה העברית-ארמית אשר כתבו פוסקים שונים כתגובה לשאלות הלכתיות במגוון נושאים, כגון: משפחה, איסור והיתר, שבת, חגים, משפטים, כלכלה, צבא, תחוקה ושלטון.

שיטת הלמידה SVM [8, 27] הוכחה כיעילה ביותר במשימות רבות של סיווג טקסטים [9, 10, 14, 15, 29]. לכן החלטנו לבחור בה כשיטת הלמידה עבור מחקר זה. בפועל השתמשנו בגרסה הנקראת SMO (Sequential Minimal Optimization) [24, 25] הממומשת על ידי סביבת הלמידה הנקראת Weka (Waikato Environment for Knowledge Analysis) [29]. Weka היא אוסף של אלגוריתמים ללמידת מכונה המתוכנתים בשפת Java עבור משימות של כריית מידע, כגון: סיווג, רגרסיה, אישכול, וכללי היסק. היא מספקת סביבת מחקר רחבה ונוחה לחוקר המבקש לטפל בכמויות מידע גדולות, ואכן מספר רב של מחקרים בתחום הלמידה משתמשים בה. השתמשנו בכרירות המחדל של SMO (linear kernel) וללא נרמול מאפיינים) כמקובל במחקרים קודמים, כ- [11]. כוונתנו של הפרמטרים השונים של SMO נשאר למחקר עתידי.

#### ד. תוצאות הניסויים וניתוחן

תוצאות הניסויים השונים נמדדו באמצעות היחס בין מספר המסמכים שסווגו נכונה מתוך קבוצת המבחן ובין מספר כל המסמכים אותו ניסינו לסווג מתוך קבוצת המבחן. עבור כל מאפייין בוצע נרמול באופן הבא: הוכפל מספר המופעים שלו בקובץ הנדון ב-10,000 ובוצע חילוק במספר המילים במסד הנתונים הרלוונטי לניסוי. התוצאות הכלליות של שלושת הניסויים מוצגות בטבלה 1.

| מס' מאפיינים | סיווג לתקופות | סיווג לעדות | סיווג לתקופות ועדות |
|--------------|---------------|-------------|---------------------|
| 2000         | 95.95         | 97.74       | 95.16               |
| 1500         | 95.46         | 97.9        | 94.82               |
| 1000         | 95.02         | 97.85       | 94.55               |
| 500          | 94.01         | 97.76       | 93.74               |
| 200          | 90.48         | 96.82       | 89.02               |
| 100          | 85.16         | 95.72       | 84.21               |
| 50           | 77.1          | 94.42       | 70.03               |
| 20           | 68.71         | 90.3        | 65.67               |
| 10           | 66.44         | 84.99       | 48.44               |

טבלה 1: תוצאות הסיווג בשלושת הניסויים

## סיווג אוטומטי של פסקי שו"ת

מסקנות כלליות עבור שלושת ניסויי הסיווג:

1. הושגו תוצאות סיווג גבוהות מאוד (98% הצלחה בניסוי העדות, 96% הצלחה בניסוי התקופות, ו-95% בניסוי העדות/תקופות).
2. זקוקים למספר רב יחסית של מאפיינים הן כדי להגיע לתוצאה הגבוהה ביותר (1,000 בניסוי הסיווג לעדות ו-2,000 בשני הניסויים הנותרים) והן כדי להגיע לרוויה (פחות מ-1% של שיפור בינה לבין הקבוצה שמעליה) (250 בניסוי הסיווג לעדות ו-500 בשני הניסויים הנותרים). בדומה לממצאו של Forman [11] התוצאות הטובות ביותר ברוב משימות הסיווג הושגו באמצעות שיטת הלמידה SVM בשימוש בכל 2,000 המאפיינים. כלומר במילים אחרות אם עובדים עם SVM סינון משתנים הוא לכאורה מיותר. אולם ברור שאם אפשר להגיע לתוצאה כמעט זהה (על פי מושג הרוויה) אך עם הרבה פחות מאפיינים יש לכך עדיפות מבחינות של זמן בניית המודל, זמן ריצה ונפח זכרון.

תוצאות אלו מראות כי בשימוש עם מספר מאות מילים (250 או 500) כמעט ללא קשר לתוכן ניתן להגיע בסוג זה של סיווגים לתוצאות גבוהות מאוד וכמעט אופטימליות. בפסקי השו"ת נתקלנו בראשי תיבות וקיצורים למכביר. מספרם המכובד תרם תרומה משמעותית להצלחות הסיווגים השונים. אולם, חשוב לציין כי ייתכן שלעיתים הימצאותם אינה מאת הכותבים עצמם כי אם של העורכים משיקולי עריכה שונים.

להלן יתוארו קבוצות מאפיינים שנתגלו כיעילות לאבחון בין הקבוצות השונות של פסקי שו"ת בסיווגים השונים:

1. מילים עם אותיות ארמיות – מילים המכילות רק אותיות ארמיות או ערבוב של אותיות ארמיות ועבריות.
2. קיצורי מילים – אות או קבוצה של אותיות המייצגות קיצור של מילה או מחרוזת מילים (קיצור של מחרוזת מילים נקרא גם ראשי תיבות).
3. התייחסויות – התייחסויות לספרים, פסקי שו"ת ופוסקים.
4. כתבים קלאסיים – פניות לש"ס, תלמוד בבלי ותלמוד ירושלמי, סוגיות, כתבי תנאים וראשונים.
5. התכתבות – מילים המתייחסות למכתבים, כגון: הגיעני, שלחני, לכבוד, מכתב, מכתבו.
6. מנהגים – מילים המתייחסות למנהגים, כגון: מנהג, המנהג, נוהגים, נוהג, שנוהגים.
7. פוסק – פוסק על שמותיו וספריו השונים, כינויו וראשי התיבות השונים שלהם, לדוגמה: שו"ע, ש"ע, שולחן ערוך, שלחן ערוך, המחבר, המחבר ר' יוסף קארו (עם רישות וסיפות שונות כגון – הש"ע, השו"ע, בש"ע, בשו"ע וכד'), מרן, ב"י, בית יוסף, כסף משנה, אבקת רוכל וברק הבית.
8. מספר מילים ממוצע בפסק שו"ת.
9. אוצר מילים (מספר מילים שונות) ממוצע לפסק שו"ת.

יעקב הכהן-קרנר, דרור מוגהץ, חנניה בק, אלחי יהודאי

### ניתוח קבוצות מאפיינים לניסוי 1 – סיווג תקופות בעזרת מילים

כאמור מסד פסקי השו"ת הנחקר במחקר זה הכיל מספר שווה של פסקי שו"ת מהמאה התשע-עשרה, המכונים כאן ישנים, ופסקי שו"ת מהמאה העשרים, המכונים חדשים. חשוב לציין כי בפרוייקט השו"ת עצמו המאה העשרים הוצעה כתקופת פוסקים בפני עצמה שכונתה בשם פוסקי זמננו.

| סוג המאפיינים                 | ערך מנורמל של מילים בקבוצה N (New, המאה ה-20) | ערך מנורמל של מילים בקבוצה O (Old, המאה ה-19) | ~N/O | ~O/N |
|-------------------------------|---|---|------|------|
| מילים בארמית                  | 8261.2  | 39453.8                                       | 0.21 | 4.78 |
| קיצורי מילים                  | 2889.9  | 3666.4  | 0.79 | 1.27 |
| התייחסויות                    | 27773.4                                       | 3654.6  | 7.63 | 0.13 |
| מנהגים                        | 15.51   | 8.51  | 1.82 | 0.55 |
| התכתבות                       | 2.37  | 3.18  | 0.75 | 1.34 |
| מספר מילים ממוצע בקובץ        | 1890.07                                       | 1268.6  | 1.49 | 0.67 |
| מספר מילים שונות בממוצע בקובץ | 812.4   | 571.2   | 1.42 | 0.70 |

### טבלה 2: קבוצות מאפיינים בולטות לסיווג תקופות

בטבלה 2 מוצגות מספר קבוצות מאפיינים בולטות בסיווג לתקופות:

- פוסקים ישנים משתמשים במילים ארמיות פי 4.78 ובקיצורי מילים פי 1.27 מאשר פוסקים חדשים. עובדות אלו מצביעות על כך שפוסקים חדשים פחות "נצמדים" למסמכים היהודיים המקוריים המכילים באופן יחסי יותר מילים ארמיות וקיצורי מילים. פוסקים חדשים מעדיפים להשתמש במילים הנכתבות בשפה העברית בצורתן השלמה כדי להיות יותר מובנים לקוראים.
- פוסקים חדשים משתמשים בהתייחסויות פי 7.63 מאשר פוסקים ישנים. הסיבה לכך פשוטה. פוסקים נוהגים בפסקי השו"ת שאותם הם כותבים להתייחס לכל או לרוב הפוסקים הקודמים שמתייחסים לשאלה הלכתית זהה או לחלק ממנה.
- התכתבות – אזכורים של התכתבות לשיטותיה השונות. קבוצה זו מופיעה אצל הישנים פי 1.34 יותר מאשר אצל החדשים. התכתבות זו היא בעיקרה של פוסקים בינם לבין עצמם. ככל הנראה בתקופה הישנה נשלחו מכתבים יותר מאשר בחדשה. במאה העשרים התקשורת החדשה המגוונת הביאה כנראה למיעוט בחליפת מכתבים.
- מנהגים – כתוצאה ממעבר יהודים בין קהילות מאמצע המאה התשע-עשרה והגברת הנדידה במאה העשרים היה מעבר של מנהגים בין עדות שונות. דבר זה האיץ את הדיונים ההלכתיים

### סיווג אוטומטי של פסקי שו"ת

בנושא המנהגים. ייתכנו מנהגים שהיו לגיטימיים בעדה מסוימת בארץ מוצאם ובארץ החדשה זה נפתח לדיון. חשוב לציין שמעבר בין ארצות לפני המאה התשע-עשרה לא היה כה שכיח כמו במאה התשע-עשרה ובמיוחד במאה העשרים. קבוצה זו מופיעה אצל החדשים פי 1.82 יותר מאשר בישנים.

5. מספר המילים הממוצע בקובץ אצל החדשים גדול ב-49% לעומת הישנים. נראה שישנן שתי סיבות לכך, האחת, החדשים צריכים להתייחס לכמות יותר גדולה של פוסקים שקדמו להם ולכן בממוצע גדולו של כל שו"ת חדש גדול יותר. השנייה, בעבר הנייר והדיו היו יקרים יותר ופחות מצויים ולכן בפסקי השו"ת הישנים השתדלו לכתוב ביותר תמצות. ניתן לראות זאת גם בשימוש בקיצורים (ראשי תיבות ונוטריונים).

6. מספר המילים השונות הממוצע בקובץ אצל החדשים גדול ב-42% לעומת הישנים. כלומר, העושר הלשוני גדול יותר בקרב החדשים. נראה כי התפתחות השפה והתפתחותם של תחומים רבים כטכנולוגיה, כלכלה, מדע ורפואה בעת החדשה נתנו את אותותיהן, בהוספתם של מילים/מונחים/מושגים רבים.

שתי התופעות האחרונות אינן תלויות שפה. הן רלוונטיות גם לטקסטים בשפות זרות כאנגלית, ערבית וספרדית ועוד. הוזלת מחירי הנייר והדיו, התפתחות השפות, והתפתחותם של תחומים רבים הן תופעות כלליות כמעט כלל-אנושיות. תופעה נוספת שנתגלתה שאינה מוזכרת בטבלה 2 היא שישנן מילים מסוימות בעברית (כגון: אולם, מצד, שאף) האופייניות במידה מובהקת לפוסקים החדשים.

### ניתוח קבוצות מאפיינים לניסוי 2 – סיווג עדות בעזרת מילים

כאמור מסד פסקי השו"ת הנחקר במחקר זה הכיל מספר שווה של פסקי שו"ת שחוברו על ידי פוסקים המשתייכים לעדה הספרדית, המכונים כאן ספרדים, ופסקי שו"ת שחוברו על ידי פוסקים המשתייכים לעדה האשכנזית, המכונים כאן אשכנזים.

| סוג המאפיינים                 | ערך מנורמל של מילים בקבוצה A (אשכנזים) | ערך מנורמל של מילים בקבוצה S (ספרדים) | ~A/S | ~S/A |
|-------------------------------|--|---------------------------------------|------|------|
| מילים בארמית                  | 523.7                                  | 940.5                                 | 0.56 | 1.80 |
| קיצורי מילים                  | 149.1                                  | 174.5                                 | 0.85 | 1.17 |
| התייחסויות מנהגים             | 426.2                                  | 590.6                                 | 0.72 | 1.39 |
| כתבים קלסיים                  | 6.83                                   | 18.02                                 | 0.38 | 2.64 |
| מספר מילים ממוצע בקובץ        | 59.13                                  | 23.27                                 | 2.54 | 0.39 |
| מספר מילים שונות בממוצע בקובץ | 1495.6                                 | 1663.1                                | 0.9  | 1.11 |
| מספר מילים שונות בממוצע בקובץ | 669.9                                  | 713.7                                 | 0.94 | 1.07 |

### טבלה 3: קבוצות מאפיינים בולטות לסיווג עדות

יעקב הכהן-קרנר, דרור מוגהץ, חנניה בק, אלחי יהודאי

בניגוד לתוצאות הברורות והחותכות שתוארו בטבלה 2, טבלה 3 מתארת תוצאות פחות חותכות ביחס לרובם המכריע של סוגי המאפיינים: מילים בארמית, קיצורי מילים, התייחסויות, מספר מילים ממוצע בקובץ ומספר מילים שונות בממוצע בקובץ. בכל אופן, טבלה זו מראה שפוסקים ספרדים משתמשים במילים ארמיות פי 1.8 (לעומת פי 4.78 בניסוי הקודם – ניסוי התקופות) מפוסקים אשכנזים. תגלית זו מצביעה על כך שפוסקים ספרדים יותר צמודים לטקסט היהודי המקורי. בשונה מהניסוי הקודם (סיווג לתקופות) אין הבדל בין העדות מבחינת עושר לשוני. ממצא זה אינו מרעיש בעקבות התפתחות השפה במקביל בעדות השונות. הבדל משמעותי קל בין העדות הוא גודלו הממוצע של קובץ. קובץ ממוצע אצל הספרדים מכיל כ-11% מילים יותר מאשר אצל האשכנזים. נראה שאחת הסיבות העיקריות לכך היא העובדה שהספרדים לשיטתם מביאים הרבה יותר התייחסויות לפוסקים שונים מאשר האשכנזים.

שני סוגי מאפיינים בולטים יחסית, שכן הבדילו בין העדות השונות, היו מנהגים והתכתבות. הספרדים מתייחסים למנהגים פי 2.64 יותר מאשר האשכנזים. סיבה אחת אפשרית היא שהספרדים נוטים להביא הרבה יותר מובאות לפוסקים בכלל ולאחרונים בפרט מאשר האשכנזים. חליפת מכתבים בין הפוסקים רווחת יותר בקרב האשכנזים. איזכורי התכתבות מופיעים אצל האשכנזים פי 2.7 יותר מאשר הספרדים.

בניגוד לתוצאות הלא בולטות המתוארות בטבלה 3, טבלה 4 מתארת מילים מסוימות המתאימות לשימוש לצורך הסיווג הנידון. ישנן מספר דוגמאות לשלושה סוגי מאפיינים: מילים בארמית, קיצורי מילים והתייחסויות. בנוסף, הבאנו מילים מסוימות הנכתבות בשפה העברית שמשמשות גם כמאפיינים בולטים לכל אחת משתי העדות הנידונות. באופן זה אנו מראים שהשימוש במגוון מילים מסוימות מניב תוצאות טובות יותר מהשימוש בסוגים כללים של מאפיינים.

| קבוצה # | סוג המאפיין          | מילה ספציפית | ~A/S  | ~S/A  |
|---------|----------------------|--------------|-------|-------|
| 1       | מילים בארמית         | ליה          | 0.39  | 2.55  |
|         |                      | הכא          | 0.69  | 1.47  |
| 2       | קיצורי מילים         | הי           | 68.05 | 0.015 |
|         |                      | לי           | 83.7  | 0.012 |
|         |                      | הי           | 0.49  | 2.07  |
| 3       | התייחסויות           | עיין         | 2.94  | 0.34  |
|         |                      | מרן          | 0.03  | 33    |
|         |                      | רבי          | 0.3   | 3.3   |
| -       | מילים מסוימות בעברית | אמנם         | 1.5   | 0.67  |
|         |                      | שיהיה        | 0.38  | 2.66  |

טבלה 4: מילים בולטות לסיווג עדות

הראי"ה קוק ב-5] במאמר "לשני בתי ישראל" מביא בעמ' 45 הברלים בולטים בין הפוסקים הספרדים לפוסקים האשכנזים, כדלקמן: "בעינינו רואים שהספרדים נסתגלו ביותר, ... לסדרנות, לבקורת ולכלל אותו הסגנון הפועל בחיים על פיהם. נחה עליהם רוח הגאונים הקדמונים, והכח של הפלפול אשר נתגבר על ידי חכמי צרפת ואשכנז לא נעשה להם לקו עקרי בחיי הקדש שלהם. והאשכנזים לעומתם האיר להם ברק הפלפול בכל הרחבתו. חריפות השכל ועומק ההבנה בשיטות התלמודיות נתגברו אצלם ונתהוו להחיותם הטיפוסי בחייהם. ורוח בעלי התוספות וחדושים נחו עליהם, ולעומת זאת נתמעט אצלם העסק בבקרת ובסדרנות". כלומר, לדברי מרן הראי"ה קוק הספרדים הינם סדרנים וביקורתיים, לא מנסים ליישב כל סתירה אלא לקבוע את הכלל ולפסוק לפיו את ההלכה ואלו האשכנזים הינם מעמיקים ומפלפלים ומנסים ליישב סתירות. הבן איש חי (הבא"ח), רבי יוסף חיים, בהקדמתו "פתיחת הספר" לשו"ת רב-פעלים [1] מביא הברל בולט נוסף בין הפוסקים הספרדים לפוסקים האשכנזים, כדלקמן: "ועוד יש בבעלי תשובות שני סוגים בענין אחר, והוא אם יש נשאל איזה שאלה בענין אחר, הוא זריז ונשכר לדפוק על דלתי הספרים של כל בעלי התשובות, ראשונים ואחרונים ואחרונים למקטון ועד גדול עז זמנו, ואפילו על ספרים אשר המחברם עודנו בחיים, ... והנה דרך זה מה טוב ומה נעים ... ובדרך זו נוהגין חכמי הספרדים בתשובות ופוסקים שלהם, לתור ולדרוש מכל הספרים ראשונים ואחרונים ואחרונים אחרונים כל אשר תשיג ידם, כדי לברר ההלכה בתשובותיהם ... ויש שאין דרכם לתור ולדרוש בספרי האחרונים בכל עניין אשר יבוא לפניהם, אלא פונים דווקא אל דברי הפוסקים הראשונים, וכותבים מה שנראה באותו ענין לפי הכרעת דעתם וסברתם, כאשר תשיג ידם באותו ענין, לקיים מ"ש כי תשב ללחום את מושל בין תבין את אשר לפניך, ובדרך זה מתנהגים על הרוב גאוני אשכנז ... עכ"ז אני אומר אחר אלף מחילות מכבוד תורתם, לא טוב זה הדבר אשר עשו, כי זה כלל גדול בתורה, אין התורה נקנית אלא בחבורה, ולכך נקראים החכמים בעלי אסופות...". כלומר, הפוסקים הספרדים מביאים פוסקים רבים ככל האפשר להרחבת היריעה, כולל אחרונים ואחרוני-אחרונים גם מתקופתו של כותב השו"ת בעוד שהפוסקים האשכנזים מביאים בעיקר גמרות וראשונים.

אולם, הראי"ה קוק ב-6], כתב באגרת תתסט (עמ' קנד): "והנה כי כן מעיד הרמ"א ג"כ על חותנו ורבו הג' ר' שכנא שלא היה מניח להעתיק את תשובותיו. ונראה משם שעיקר הטעם שלו הי' שלא רצה שיקבע הלכה לדורות, אלא שכל חכם בדורו יפסוק כפי מה שנראה לו מן התלמוד. וזו היתה שיטתו של המוהר"ל מפראג, שהאריך בס' נתיבות עולם שעיקר יסוד תורה שבע"פ הוא שתהי' ההוראה מן הש"ס ולא מד' הפוסקים, וכ"כ הפליג בענין עד שכתב שיותר טוב לפסוק מן הש"ס אע"פ שיטעה בעיונו, מ"מ זהו דרך התורה לפסוק כפי מה שענינו רואות. אמנם דעת הר"י מיגאש בת' נראה שאינה כן, שכ' ע"ד שני חכמים שאחד בקי בתלמוד, שיותר טוב למנות את הבקי בדברי הגאונים, והאריך בהבאת נסיונות שראה שטועים ע"י ההוראה מן התלמוד עצמו. ואלו ואלו דברי א"ח הם".

עדויות חותכות לחלק מדבריהם של הראי"ה קוק והבן איש חי ניתן לראות למשל בתוצאות

יעקב הכהן-קרנר, דרור מוגהץ, חנניה בק, אלחי יהודאי

שקיבלנו בטבלה 3 בניסוי הסיווג לעדות עבור קבוצות המאפיינים: שו"ת, מנהגים וכתבים קלסיים. הפוסקים הספרדים אכן מביאים הרבה יותר מובאות ולכן יש אצלם הרבה הפניות לפסקי שו"ת (פי 5.9) ומנהגים (פי 2.64) בכלל וכן לחלק מהפוסקים, בייחוד לספרדים שביניהם, למשל מרן (פי 3.47) והחיד"א (פי 29.4). הם אינם מתפלפלים, אלא רוצים לקבוע הלכה על פי מה שנכתב בעבר עם סברה מצדם. ברור שאצלם הפסיקה הקודמת חשובה יותר. כדי "לסטות" מהקו ההלכתי שהותווה על ידי הפוסקים הקודמים צריכה להיות מבחינתם סיבה מאוד טובה. הפוסקים האשכנזים אכן מסתמכים הרבה יותר על הכתבים הקלסיים (פי 2.54) מאשר על פסקי שו"ת מאוחרים יותר. בצורה כזו יש להם הרבה "מקום" להביע את סברתם ולהתפלפל והם אינם מחויבים לפסקי שו"ת. הם מסתמכים יותר על הסוגיה התלמודית ישירות, מתפלפלים ומכריעים.

**ניתוח קבוצות מאפיינים לניסוי 3 – סיווג תקופות/עדות בעזרת מילים**  
 כאמור מסד פסקי השו"ת הנחקר חולק אף לארבע קבוצות המכילות מספר שווה של פסקי שו"ת: ספרדים ישנים, ספרדים חדשים, אשכנזים ישנים ואשכנזים חדשים.

| מאפיין       |        | ישנים O |        | חדשים N |        |
|--------------|--------|---------|--------|---------|--------|
|              |        | אשכנז   | ספרד   | אשכנז   | ספרד   |
| ושכמ"ה       | 2.77   | 0       | 0.13   | 0       | 0.13   |
| שליט"א       | 0.01   | 0.67    | 2.19   | 3.74    | 2.19   |
| אכ"י         | 1.99   | 0.02    | 0.13   | 0.01    | 0.13   |
| יה           | 145.97 | 33.15   | 113.14 | 77.8    | 113.14 |
| י            | 121.4  | 200.86  | 56.67  | 92.22   | 56.67  |
| כתבים קלסיים | 13.19  | 72.67   | 28.28  | 47.59   | 28.28  |
| שו"ע         | 45.34  | 14.15   | 58.07  | 16.27   | 58.07  |
| ש"ך          | 7.11   | 28.69   | 8.93   | 11.70   | 8.93   |
| ט"ז          | 4.50   | 5.10    | 4.64   | 3.57    | 4.64   |
| מג"א         | 7.38   | 10.72   | 13.50  | 11.37   | 13.50  |
| רמ"א         | 2.31   | 6.31    | 6.48   | 5.33    | 6.48   |
| מור"ם        | 5.83   | 0.01    | 2.75   | 0.02    | 2.75   |
| חיד"א        | 1.7    | 0.02    | 4.18   | 0.18    | 4.18   |

**טבלה 5: קבוצות מאפיינים בולטות לסיווג תקופות/עדות**

בטבלה 5 מוצגות 13 קבוצות מאפיינים בולטות לסיווג לפי תקופות/עדות. להלן ננתח בקצרה מספר תופעות מן החשובות שבתוצאות הנ"ל.

במאה העשרים היו מספר תהפוכות עולמיות אשר השפיעו בין השאר גם על העולם התורני. מלחמות העולם הראשונה והשנייה גרמו להגירה גדולה של היהודים בעולם בין מדינות בכלל



## סיווג אוטומטי של פסקי שו"ת

ובעלייה לארץ ישראל בפרט. דבר נוסף שהתפתח בתקופה זו היא התקשורת בין אנשים. עד לתקשורת האלקטרונית/חשמלית/קווית/אלחוטית הייתה התקשורת באמצעות מכתבים ו/או שליחים. התקשורת המודרנית הזמינה, גם היא הובילה וממשיכה להוביל לסוג של קיבוץ גלויות. לעניות דעתנו שינויים אלו הובילו להשפעה הרדית גדולה בין העדות. התופעות הניכרות והבולטות בחשיבותן היו כדלקמן:

1. התחזקות מעמדן של קבוצות המאפיינים הבאות: שו"ע, מג"א, שליט"א ושימוש בסימות "יה".

2. היחלשותן / היעלמותן של קבוצות המאפיינים הבאות: ביטויי הברכה ושכמ"ה ואכי"ר, הסימות י' וכתבים קלסיים.

3. התקרבות וצמצום פערים בין אשכנזים לספרדים מהעת הישנה לחדשה. כלומר הפער בין הספרדים לאשכנזים הולך ומצטמצם בחלק מן המאפיינים, כגון: כתבים קלסיים, סימות "יה", רמ"א, ושכמ"ה ושליט"א.

תופעה זו של התקרבות וצמצום הפערים בין העדות מהעת הישנה לחדשה תואמת את חזונו של מרן הראי"ה קוק לפני מספר דורות כמובא בסוף מאמרו "לשני בתי ישראל" ב-5 [בעמ' 48: "כל חלק מחלקי האומה מוכרח הוא לשכלל את כשרונו, ועם זה חובתנו עכשיו גדולה היא יותר מבכל זמן להיות כל אחד משפיע ומושפע מחברו, ואז ישתלמו בנו שני הכשרונות הללו, ... עד שהספרדיות והאשכנזיות תשפיע זו על זו השפעה חיה ומלאה, ובהמשך הזמן ישתוו הכישרונות המפורדים הללו, ... בבנין משוכלל כזה, ששום כשרון ושום יתרון, ששום מידה טובה ושום רעיון טוב, שיש בכל אחד מהבתים הללו, לא יהיה נאבד מאתנו. ... ואנו מקוים כי יד ה' עשתה זאת לקבץ את שני בתי ישראל, בצורה כל כך נכרת בצביוניהם השונים פה בארץ ישראל, כדי שיהיו מוכנים לפעול זה על זה, את הפעולה הרצויה של ההשפעה הטובה, הנותנת לכל אחד מהם את תפקידו ופעולתו בחיי האומה הכלליים ...".

## ה. סיכום, מסקנות ומחקר עתידי

מודל סיווג זה הינו הראשון מסוגו (סיווג לפי עדות, תקופות ועדות/תקופות) לשפה העברית בכלל והראשון מסוגו לתחום התורני בפרט. יתירה מכך, למיטב ידיעתנו הוא גם הסיווג הראשון בעולם של טקסטים לפי מוצאם העדתי/גיאוגרפי של מחבריהם. בכל שלושת ניסויי הסיווגים (עדות, תקופות, עדות ותקופות) הצליחה המערכת לסווג נכונה מעל ל-95% מן הקבצים. באמצעות ניסויים אלו ניתן לזהות הבדלים בולטים ומעניינים המאבחנים בין הפסיקה הספרדית לפסיקה האשכנזית ובין פסיקות מתקופות ישנות יותר לפסיקות של אחרוני זמננו. לאחר ניתוח המילים המאבחנות בניסויים השונים ניתן היה למצוא קבוצות מאפיינים בעלי משמעות תוכנית ו/או לשונית המבחינות בין הפסיקה הספרדית לזו האשכנזית ובין פסיקות מתקופות ישנות יותר לפסיקות של אחרוני זמננו. התוצאות שנתגלו עשויות להיות בעלות ערך לחוקרים מתחום מדעי החברה החוקרים הבדלים בין תרבויות שונות והבדלים בין תקופות שונות. מן הסתם ישנן תופעות

נוספות שניתן לגלות במחקר נוסף.

במחקר המוצג נעשה שימוש במסד נתונים גדול יחסית (מעל 12,000 מסמכים) ובשיטת למידה SVM הנחשבת כיום באופן כללי לטובה ביותר עבור סיווג טקסטים, בעוד שבמחקרי הסיווג הקודמים על פסקי שו"ת השתמשו במסדי נתונים קטנים יחסית ובשיטות למידה הנחשבות לבסיסיות יותר, כגון שיטת ה-Balanced Winnow. במערכות קודמות אחרות בוצע סינון מאפיינים באופן ידני. אנו ביצענו סינון אוטומטי בשיטה המדעית הידועה InfoGain.

מוצעים כיווני מחקר כלליים, כגון בדיקת סיווג מעין זה של טקסטים בשפות אחרות בעזרת מאפיינים סגנוניים לפי קבוצות עדתיות ו/או גיאוגרפיות שונות ברחבי העולם. ניסוי שיטות למידה נוספות, שיפור יכולת הניתוח המורפולוגית (=הדקדוקית) של המערכת. כוונת של הפרמטרים השונים של SMO, ניסוח וניסוי שיטות מאפיינים נוספות, הן כלליות והן ייחודיות לשפה העברית ולתחום התורני והרחבה של המודל עבור כתבים בעברית במגוון תחומים אחרים. כיווני מחקר נוספים ספציפיים יותר לתחום המחקר של המערכת (התחום התורני-הלכתי) הם: הרחבת אוטומטיות של מושגים או שמות עם שינויים ותוספות כגון: יחיד/רבים, זכר/נקבה, תחיליות, סופיות וכו' תוך וידוא נכונות השינויים והתוספות כאמתיים במאגר המסמכים. הוספה של מאפיינים כלליים נוספים, כגון: גמרא, ברייתא, משנה, מתניתין, קרא, תניא, שו"ת ה-... וכדומה, חלוקתם של המושגים לפי תקופות כגון מקרא, תנאים, אמוראים, ראשונים וכו' ולפי שפה עברית/ארמית.

מחד גיסא, מעניין יהיה לבדוק את אבחנתו של הרב קוק לגבי דגש על השימוש בפשט ובכללים אצל הספרדים לעומת הפלפול והחקירה של האשכנזים. ניתן לבנות מושגים בולטים לכל אחת מהקבוצות ולבדוק מי מהם מספק הבנה כזו. מושגים כגון אלו יכולים להיות בקבוצת הפשט והכלל של הספרדים: ביקורת, פשט, כלל, בקי, סיני, הכרע, יסוד, עיקר והלכה. ובקבוצת הפלפול והחקירה של האשכנזים: פלפול, חקירה, עיון, דקדוק, חידוש, טעם, סברה / סברא, עומק, שני דינים ושיטה.

מאידך גיסא, מעניין יהיה לבדוק את אבחנתו של הבא"ח להבדל בולט שהוא ציין בין הפוסקים הספרדים לפוסקים האשכנזים. לדבריו, הספרדים מביאים פוסקים רבים ככל האפשר להרחבת היריעה כולל אחרונים ואחרוני-אחרונים גם מתקופתו של כותב השו"ת, בעוד שאלו האשכנזים מביאים בעיקר גמרות וראשונים, כמעט ללא אחרונים, כלומר במילים שלנו אחוז האחרונים מקרב הפוסקים שיזכירו הפוסקים האשכנזים הוא נמוך הרבה יותר מזה שיופיע אצל הפוסקים הספרדים.

מעניין לבדוק אם אכן יש מושגים בולטים מקרב הנ"ל שיכולים לספק לנו את האבחנה של הבא"ח על פי שתי קבוצות אלו: קבוצת הספרדים המצטטים כמה שיותר פוסקים כולל כמה שיותר אחרונים לעומת קבוצת האשכנזים שיש בה פחות פוסקים בכלל והרבה פחות אחרונים בפרט (לאו דווקא ספרדים). לשם כך יש לקחת בחשבון את שמותיהם של כל הפוסקים משתי העדות, כל חיבוריהם, וכל ראשי התיבות והקיצורים המציינים נתונים אלו.

מעניין יהיה לראות האם היום גם עבור הפוסקים בני זמננו, מספר דורות לאחר פטירתם של הבא"ח והראי"ה, האבחנות האלו תקפות ואם כן – באיזו מידה ועבור אילו מאפיינים.

### ביבליוגרפיה

- [1] הרב יוסף חיים אל-חכם, (בן-איש-חי), רב פעלים, הוצאת שיח ישראל, ירושלים 1994.
- [2] הרב י' כחלי (מיוחס לרב יוסף חיים – הבא"ח), תורה לשמה, הוצאת שיח ישראל, ירושלים 1973.
- [3] ד' מוגהץ, סיווג של טקסטים בעברית על פי סגנון, עבודה לתואר שני בהנחיית פרופ' מ' קופל, אוניברסיטת בר-אילן 2003.
- [4] פרוייקט השו"ת של אוניברסיטת בר אילן (גרסה 12) (<http://www.biu.ac.il/JH/Responsa/Heb/index.html>)
- [5] הרב אברהם יצחק קוק, מאמרי הראי"ה, ירושלים 1988.
- [6] הרב אברהם יצחק קוק, אגרות הראי"ה, חלק ג, ירושלים 1985.
- [7] Argamon-Engelson, S., Koppel, M. and Avneri, G., "Style-based Text Categorization: What Newspaper Am I Reading?", *Proc. AAAI Workshop on Text Classification*, Madison, WI, 1998.
- [8] Cortes, C. and Vapnik, V., "Support-Vector Networks", *Machine Learning* 20 (1995), pp. 273-297.
- [9] Díaz, I., Ranilla, J., Montañés, E., Fernández, J. and Combarro, E. F., "Improving Performance of Text Categorization by Combining Filtering, Supportvector Machines", *JASIST* 55(7), (2004), pp. 579-592.
- [10] Dumais, S., Platt, J., Heckerman, D. and Sahami, M., "Inductive Learning Algorithms and Representations for Text Categorization", *Proceedings of the 7th ACM International Conference on Information and Knowledge Management (CIKM)*, Bethesda, MD (1998), pp. 148-155.
- [11] Forman, G., "An Extensive Empirical Study of Feature Selection Metrics for Text Classification", *J. of Machine Learning Research* 3 (2003), pp. 1289-1305.
- [12] Forsyth, R. and Holmes, D., "Feature-Finding for Text Classification", *Literary and Linguistic Computing* 11 (1996), pp. 163-174.
- [13] Holmes, D., "Authorship Attribution", *Computers and the Humanities* 28 (1994), pp. 87-106.
- [14] Joachims, T., "Text Categorization with Support Vector Machines: Learning with Many Relevant Features", *Proceedings of the 10th European Conference on Machine Learning (ECML)*, Chemnitz, Germany (1998), pp. 137-148.
- [15] Joachims, T., *Learning to Classify Text Using Support Vector Machines*, Kluwer 2002.
- [16] Kjell, B., "Authorship Determination Using Letter Pair Frequency Features with Neural Net Classifiers", *Literary and Linguistic Computing* 9 (1994), pp. 119-124.
- [17] Koppel, M., Mughaz, D. and Akiva, N., CHAT: A System for Stylistic Classification of Hebrew-Aramaic Texts, *Proceedings of OTC-03 Third KDD Workshop on Operational Text Categorization*, Washington D.C. 2003.
- [18] Koppel, M., Mughaz, D. and Akiva, N., "New Methods for Attribution of Rabbinic Literature", *Hebrew Linguistics: A Journal for Hebrew Descriptive, Computational, Applied Linguistics* 57 (2006), pp. v-xviii.

יעקב הכהן-קרנר, דרור מוגהץ, חנניה בק, אלחי יהודאי

- [19] Maron, M., "Automatic Indexing: an Experimental Inquiry", *J. Assoc. Comput. Mach.* 8, 3 (1961), pp. 404-417.
- [20] McEnery, A. M. and Oakes, M. P., "Authorship Studies/Textual Statistics", in R. Dale, H. Moisl, and H. Somers (eds.), *Handbook of Natural Language Processing*, New York-Basel 1998.
- [21] Matthews, R. and Merriam, T., "Neural Computation in Stylometry I. An Application to the Works of Shakespeare and Fletcher", *Literary and Linguistic Computing* 8 (1993), pp. 203-209.
- [22] Meretakis, D. and Wuthrich, B., "Extending Naïve Bayes Classifiers Using Long Itemsets", *Proceedings of the 5th ACM-SIGKDD Int. Conf. Knowledge Discovery and Data Mining (KDD'99)* (1999), pp. 165-174.
- [23] Merriam, T. and Matthews, R., "Neural Computation in Stylometry II. An Application to the Works of Shakespeare and Marlowe", *Literary and Linguistic Computing* 9 (1994), pp. 1-6.
- [24] Platt, J. C., "Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines", *Microsoft Research* (1998), pp. 41-65.
- [25] Platt, J. C., "Fast Training of Support Vector Machines using Sequential Minimal Optimization", *Advances in Kernel Methods - Support Vector Learning*, B. Scholkopf, C. Burges, and A. J. Smola (eds.), MIT Press, Cambridge, Massachusetts (1999), chapter 12, pp. 185-208.
- [26] Sebastiani, F., Machine Learning in Automated Text Categorization, *ACM Computing Surveys* 34 (1) (2002), pp. 1-47.
- [27] Vapnik, V. N., *The Nature of Statistical Learning Theory*, Springer-Verlag, New York 1995.
- [28] Witten, I. H. and Frank, E., *Weka 3: Machine Learning Software in Java*, <http://www.cs.waikato.ac.nz/~ml/weka>, 1999.
- [29] Yang, Y. and Liu, X., "A Re-examination of Text Categorization Methods", *Proceedings of the 22nd ACM International Conference on Research, Development in Information Retrieval (SIGIR)*, Berkeley, CA 1999, pp. 42-49.
- [30] Yang, Y. and Pedersen J.P., "A Comparative Study on Feature Selection in Text Categorization", *Proceedings of the Fourteenth International Conference on Machine Learning (ICML'97)* (1997), pp. 412-420.

סיווג אוטומטי של פסקי שו"ת

**נספח – נתונים סטטיסטיים על מסד פסקי השו"ת שנחקר**

מסד הקבצים שנחקר במחקר זה מכיל 12,014 פסקי שו"ת שהם חלק מקבצי השו"ת המופיעים בפרוייקט השו"ת גרסה 12 [4]. פסקי שו"ת הנ"ל נבחרו כך שעבור כל ניסוי סיווג שבוצע במחקר זה (להלן חלוקה) כל קבוצה מתוך הקבוצות השייכות לאותה חלוקה תכיל מספר שווה של פסקי שו"ת כדי שלא תהיה הטיה בתהליך הלמידה. פסקים אלו נכתבו על ידי 48 פוסקים, כאשר כל פוסק כתב בממוצע 250 פסקי שו"ת ממסד הנתונים. מספר המילים בכל המסמכים במסד הנתונים הינו כ-18.7 מליון מילים (ליתר דיוק 18,688,494).

| עדות    |         | תקופות   |         |                               |
|---------|---------|----------|---------|-------------------------------|
| אשכנזים | ספרדים  | חדשים    | ישנים   |                               |
| 8970102 | 9718392 | 11046569 | 7641842 | מספר מילים                    |
| 229202  | 231419  | 251911   | 207397  | מספר מילים שונות              |
| 1495.6  | 1663.1  | 1890.07  | 1268.6  | מספר מילים בממוצע בקובץ       |
| 669.9   | 713.7   | 812.4    | 571.2   | מספר מילים שונות בממוצע בקובץ |

