בס״ד

דו״ח סופי

CLASSIFIER RAMBAM

מטרה

מטרתו של המסמך הינו להסביר את מהות הפרויקט ומהווה את הדו״ח הסופי.

הסבר כללי

ספרי שו״ת

שאלות ותשובות (בראשי תיבות: שו״ת) הוא כינוי לאחת מהסוגות הענפות והפוריות בספרות התורנית, בעיקר בתחום ההלכה. השו״ת מכיל מאגר של שאלות ותשובות הלכתיות, אשר נשאלו על ידי הציבור הרחב, ונענו על ידי רב אחד או קבוצה של רבנים. נמצאות גם שו״ת בענייני אגדה וכדומה (שאינם נוגעים להלכה).

לפני חורבן הבית והגלות כל השאלות ההלכתיות נשאלו ונענו על ידי בתי הדין שהיו בארץ ישראל.

לאחר שגלה עם ישראל מארצו והתפזר בגולה, נוצרו פוסקים מקומיים, אשר אליהם פנו בשאלות הלכתיות. נהוג היה כי השואל ישלח שאלה לרב, והרב יענה תשובה. מפעם לפעם יצא קובץ שאלות ותשובות כאלו בדפוס, ונודע בשם ״שו״ת״.

עם המצאת הדפוס והדפסת ספרי שו״ת הפכה קטגוריית השו״ת לקטגוריה נוספת בספרות, ופרחה ספרות שו״ת שלא נשאלה מעולם, אלא כותבה העלה את השאלה מדעתו ונתן עליה את התשובה.

הרמב״ם

רבי משה בן מימון היה מגדולי הפוסקים בכל הדורות, מחשובי הפילוסופים בימי הביניים, איש אשכולות, מדען, רופא, חוקר ומנהיג. אחד האישים החשובים והנערצים ביותר ביהדות. עליו נאמר ״ממשה עד משה לא קם כמשה״ והוא הוכתר בכינוי ״הנשר הגדול״. הרמב״ם החזיק במשנה רציונליסטית מובהקת שבאה לידי ביטוי בכתביו.

ספרו הגדול והמרכזי של הרמב״ם הוא ״משנה תורה״, הידוע בכינוי ״היד החזקה״ (מאחר והוא מחולק ל״ד ספרים). ספר זה הוא בעצם סיכום התורה שבעל פה, והוא מהווה אחד מיסודותיה של ההלכה. שמו הוא משנה תורה ״לפי שאדם קורא בתורה שבכתב תחלה ואחר כך קורא בזה ויודע ממנו תורה שבעל פה כולה ואינו צריך לקרות ספר אחר ביניהם״- כדברי הרמב״ם בהקדמתו.

הספר כתב בלשון ברורה ומדויקת, והוא ערוך בצורה מופתית לארבעה עשר ספרים, כשכל ספר מחולק ל״הלכות״ (על פי נושאים), והן מחולקות לפרקים ולהלכות (למשל: ספר המדע, הלכות יסודי התורה פרק א הלכה א).

פרויקט השו"ת

פרויקט השו"ת (בשמו הרשמי: מאגר היהדות הממוחשב) הוא כינוי לתוכנית ליצירת מאגר מידע ממוחשב של כתבים יהודיים תורניים עם דגש על ספרות הלכתית וספרי שאלות ותשובות (שו"ת), ופיתוח מנוע חיפוש המאפשר אחזור מידע מהמאגר על פי מילות חיפוש. הפרויקט החל בשנות השישים של המאה ה-20 במכון ויצמן והיה מהראשונים בעולם לשמירה דיגיטלית של טקסט שלם ("טקסט מלא") לצורך חיפוש בו. כיום הוא מובל על ידי אוניברסיטת בר-אילן. הפרויקט זכה בפרס ישראל לספרות תורנית לשנת תשס"ז. הכינוי "פרויקט השו"ת" משמש כיום גם ככינוי לתוכנה המכילה את המאגר עצמו.[1]

הפרויקט

הפרויקט המוצע הינו פרויקט שמטרתו לבחון את היכולת לסווג את התשובות השונות שבמאגר פרויקט השו"ת לפי מפתח נושאי היררכי (טקסונומיה).

הפרויקט ירצה לסווג את התשובות השונות שבמסגרת השו"תים לפי סימני השו"ע וההלכות ברמב"ם להם הם מתייחסים ובכך ליצור מאגר שיקשר בין שאלות זהות בספרי השו"ת השונים עבור הלומדים.

שאלת הפרויקט

מטרת העל של הפרויקט הינה ליצור מסווג אשר יעבור על טקסט של שו"ת ויסווג את השאלות והתשובות שבו לפי נושא ההלכה עליהם הם מדברים.

בנוסף אנו יודעים כי רוב חכמי ישראל בדורות האחרונים התבססו בצורה זו או אחרת על הרמב"ם.

השאלה המרכזית: האם ניתן ליצור מסווג לכלל השו"תים כאשר הבסיס שלו הוא הרמב"ם בלבד?

מהלך הפרויקט

מסווג SVM

מכונת תמך וקטורי (Support Vector Machine) היא טכניקה של למידה מונחית המשמשת לניתוח נתונים לסיווג ולרגרסיה.

במקרה שלנו אנו רוצים ליצור מסווג לינארי עבור כל אחד מההלכות (נושאים) של המשנה תורה.

יצירת מסווג לינארי עבור כל אחד מההלכות יאפשר לנו להכניס טקסט כלשהו אשר אז יסווג לפי הנושא שלו ויאופיין כקישור לנושא.

לאחר מכן כאשר המשתמש ייפתח שאלה או יעיין ברמב"ם 'פרויקט השו"ת' יוכל להפנות אותו לכל השו"תים המדברים על הנושא.

[1] ויקיפדיה

צורת הסיווג

הסיווג נעשה לפי ההלכות במשנה תורה כשכל הלכה הינה תג בפני עצמו

מאגרי מידע

לצורך הפרויקט נעשה שימוש במספר מאגרי מידע כל מאגר עבר עיבוד מקדים להסרת סימני פיסוק וסימנים נוספים העלולים להפריע למסווג.

כמו כן כל מאגר חולק לדוגמאות לסט אימון/בדיקה לא לפי סעיפים אשר היו לעיתים קרובות קצרים (בעיקר ברמב"ם) ולא הספיקו לצורך סיווג אלא לפי מספר מילים כך כל דוגמא הכילה 100 מילים ללא תלות בסעיף.

סיווג מאגרי המידע

בעוד שלמאגרי המידע של הבדיקה וכן לרמב"ם עצמו היה סיווג הרי שלמאגרי המידע של האימון קרי ספר החינוך והנודע ביהודה לא היה.

בעבודה שנמשכה כחודש סווגו הספרים בצורה ידנית סעיף סעיף.

מאגרי מידע לאימון

- משנה תורה להרמב"ם – זהו המאגר המרכזי עליו בוסס המסווג כל אופציה אפשרית של יצירת מסווג נעשית כשספר זה בבסיס ועיקר המסווג.
- ספר החינוך – ספר המקיף את כל מצוות התורה ומכיל מידע מפורט עליהן. הספר מכיל מילות מפתח נוספות עבור כל מצווה ויוכל להרחיב את המשנה תורה. השימוש בו נעשה בצורת הסיווג השלישית.
- הנודע ביהודה – שו"ת אחרון המכיל ביטויים שלא היו בזמן הרמב"ם ולפיכך מוסיף מילות מפתח נוספות.

מאגרי מידע לבדיקה

- הטור – החיבור הגדול הראשון אחרי הרמב"ם שחולק אף הוא להלכות בזהה למשנה תורה ולכן אמור לתת תוצאות גבוהות.
- קיצור שולחן ערוך – חיבור שמסכם את הטור והשולחן ערוך לכדי הלכות המובנות ליישום.
- בן איש חי – שו"ת אחרון החיבור מקיף שאלות בתחומים שונים.

מסווגים

לצורך יצירת המסווג נוסו שני מסווגים שונים LinearSVC ו SVC.

על כל סט אימון בוצע חיפוש למציאת מאפייני המסווג הטוב ביותר. המסווג שיצא הטוב ביותר הינו הLinearSVC כאשר המאפיינים הטובים ביותר שלו נמצאים בפונקציה fitPipe() בקוד.

מסווג ראשון – הרמב"ם בלבד

המסווג לקח את הלכות הרמב"ם כתגים.

כל טקסט הינו בן 100 מילים מתוך הנושא (ללא התחשבות בפרקים).

הדגימה לסט האימון נעשתה באקראי כדי לתת פרישה מלאה. הרמב״ם כתב את הנושא לפי סדר ודגימה לינארית עלולה לכסות רק חלק מהנושא ולא את כולו.

התוצאות של המסווג לפי הרמב״ם בלבד לא היו מרשימות וניסיון לבדוק מה המניע העלה שאין כלל קבוע לתוצאות השליליות.

התוצאות היו:

- טור – 67.58%
- קיצור שולחן ערוך – 35.31%
- הבן איש חי – 48.91%

מסווג שני – רמב״ם ושו״ת הבדיקה

המסווג השני אומן עבור כל אחד משו״תי הבדיקה לעצמו.

מכל שו״ת בדיקה נלקחו דגימות אשר אומנו עם סט האימון של הרמב״ם כך שנוצרו שלוש מסווגים, אחד עבור כל שו״ת בדיקה.

לאחר מכן אומנו שלוש מסווגים נוספים מאותו סוג אך הפעם במקום לקחת דגימות אקראיות נלקחו דגימות המכילות אזכור לרמב״ם.

המסווג נוסה רק עם SVMLinear שהראה תוצאות טובות יותר ברוב מקרי הבדיקה.

המסווג הראה תוצאות טובות יותר והקפיץ בצורה ניכרת את התוצאות ובעיקר את התוצאות של הבן איש חי (תוצאות ניתן לראות בפוסטר).

בעיה ראשונה – הבעיה הראשונה שצצה היא שאנו נאלץ לנו מסווג אישי עבור כל שו״ת בפני עצמו. במקרה זה עדיף לבנות מסווג רגיל על כל שו״ת ושו״ת ללא הרמב״ם.

בעיה שנייה – הבעיה צצה לאחר הרצת הבדיקה. התוצאות היו ״טובות מדי״ והיה ניכר שהסיווג נעשה לא לפי הרמב״ם אלא לפי השו״ת כאשר לא ממש ניתן לקבוע האם ועד כמה הדוגמאות מהרמב״ם משפיעות.

התוצאות היו:

עבור חלוקה כללית

- טור – 68.72%
- קיצור שולחן ערוך – 56.61%
- הבן איש חי – 80%

עבור חלוקה לפי אזכורי הרמב״ם

- טור – 73.62%
- קיצור שולחן ערוך – 32.5%
- הבן איש חי – 44.77%

מסווג שלישי – רמב"ם ושו"תי האימון הנוספים

השו"תים הנוספים חולקו גם הם לפי 100 מילים וצורפו שניהם לסט האימון עם הרמב"ם.

כדי לתת יותר משקל לרמב"ם שהינו החלק המרכזי של המסווג ועם זאת לשמור על המשקל של שני השו"תים הנוספים שלא "יבלעו" לתוך המסווג, המסווג חולק במשקלים יחסיים.

- הרמב"ם מהווה 50% מהמסווג (בפועל 30 דוגמאות)
- ספר החינוך מהווה 25% מהמסווג (בפועל 15 דוגמאות)
- שו"ת נודע ביהודה מהווה 25% מהמסווג (בפועל 15 דוגמאות)

המסווג נוסה רק עם SVMLinear שהראה תוצאות טובות יותר ברוב מקרי הבדיקה.

המסווג הראה תוצאות טובות בחלק מהמקרים כך למשל סיווג קיצור שולחן ערוך עלה בכמה נקודות בעוד שבטור דווקא חלה ירידה.

המסווג מראה סימני שיפור כללים אך לא מספיק כדי להוות שינוי מהותי.

התוצאות היו:

- טור – 56.1%
- קיצור שולחן ערוך – 67.46%
- הבן איש חי – 56.07%

המשך עתידי

מסווג לפי הרמב"ם בלבד כנראה לא ישים.

מסווג עם שו"ת יקר מדי בבנייה וכנראה חוטא למטרה.

המסווג שלישי מראה סימני שיפור אך אלו אינם מספיקים.

מסווג עתידי יוכל לנסות להוסיף שו"תים למסווג השלישי למשל הרחבת המסווג על ידי הוספת 7 שות"ים נוספים שיהוו 9 שו"תים בעלי משקל זהה בנוסף לרמב"ם

בנוסף ניתן יהיה להוסיף שו"תים מתקופות שונות בכדי ליצור פרישה רחבה יותר שתקיף את המושגים השונים לאותם נושאים שנוצרו בתקופות הזמן השונות. (כדוגמא נוכל לקחת את החשמל אשר לא הופיע בתקופת הרמב"ם ואילו כיום נכנס בצורה דומיננטית לשיח השו"תים).

ישנן שלוש בעיות מרכזיות העולות לקראת המשך עתידי.

- הראשונה האם הרמב"ם באמת מקור מקיף מספיק שייתן סיווג אופטימלי? ומצד שני האם נוכל למצוא שו"ת אחר שיקיף בצורה דומה את המצוות וישמש מקור לסיווג טוב וכללי?
- השניה מציאת שו"תים מסווגים הינה בעיה וסיווג שלהם לוקח זמן. סיווג של מספיק שו"תים ליצירת מסווג שלישי אופטימלי תהיה בעיה של זמן יותר מבעיית תכנות.
- השלישית האם שיטות NLP יתנו פתרון טוב יותר? או שמא לא באמת ישנו? וזה בלי להזכיר שאין כלים בעברית ויצירת כלים שכאלה מוליכה אותנו לדרך אחרת ממהות הפרויקט.