הפקולטה לניהול ע"ש קולר
אוניברסיטת תל אביב

מאמר זה פורסם בכתב העת "חידושים בניהול", גיליון מס' 7, אוקטובר 2020,
היוצא לאור על ידי הפקולטה לניהול על שם קולר באוניברסיטת תל אביב.

סיווג נתוני עתק התנהגותיים: מכונה, אדם, או שילוב של השניים

ענבל יהב דיויד שוורץ

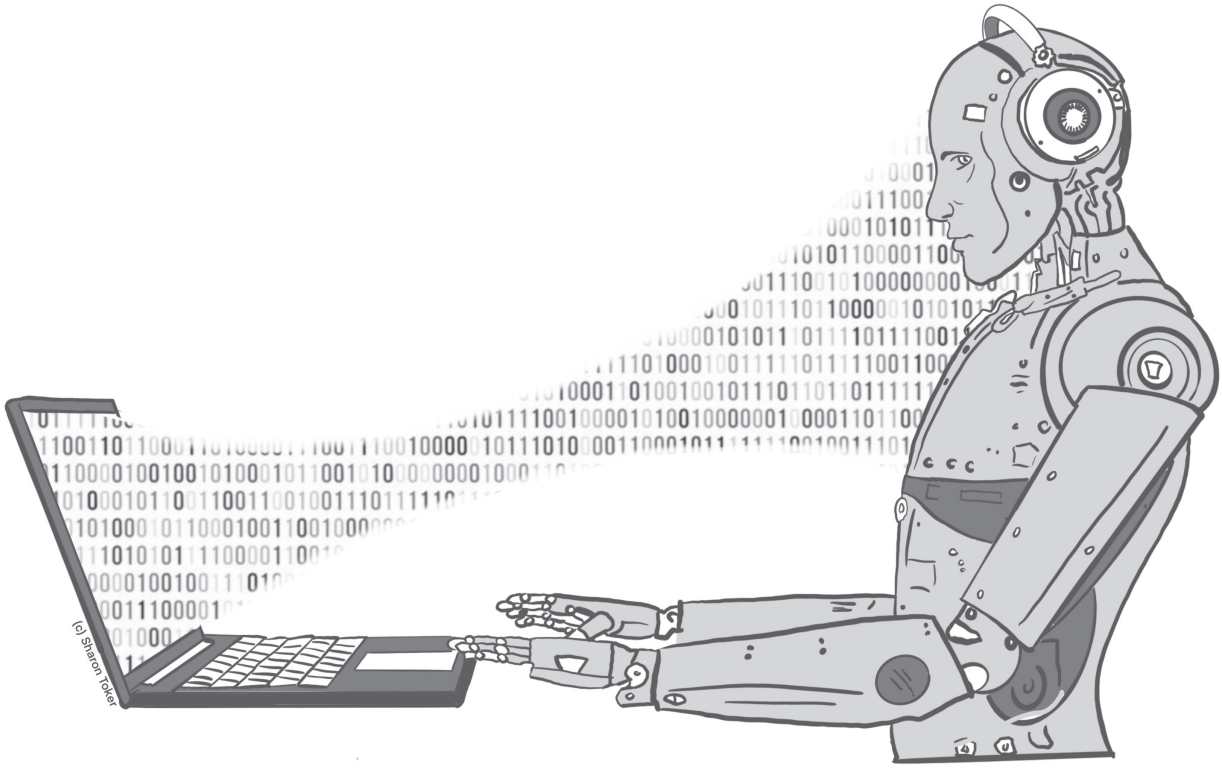
ד"ר ענבל יהב היא חברת סגל בפקולטה לניהול ע"ש קולר באוניברסיטת תל אביב. בעלת תואר ראשון במדעי המחשב ותואר שני במערכות מידע מהטכניון, וקיבלה את הדוקטורט שלה באופטימיזציה וכריית נתונים מאוניברסיטת מרילנד בשנת 2010. במשך שנתיים המשיכה לעבוד כחברת סגל אורחת באוניברסיטת מרילנד, ולאחר מכן עבדה במשך שמונה שנים באוניברסיטת בר אילן. עבודתה מתמקדת בעיקר בפיתוח והתאמה של מודלים סטטיסטיים לשימושם של חוקרים במערכות מידע. ד"ר יהב משלבת במחקריה אלגוריתמים לכריית נתונים, כריית טקסטים ומודלי אופטימיזציה כדי לייצר מודלים סטטיסטיים מורכבים בעלי יכולת חיזוי גבוהה. היא מיישמת שיטות אלה בעיקר על יישומי בריאות הציבור ועל אבטחת הרשת.

פרופ' דיויד שוורץ הוא פרופ' למערכות מידע וראש התוכנית לתואר שלישי בבית הספר למינהל עסקים באוניברסיטת בר אילן. במעבדה של פרופ' שוורץ, The Social Intelligence Lab, מתבצעים מחקרים בתחום הסייבר, יישומי בריאות ורפואת חירום, מדיה חברתית וניהול ידע ארגוני. הוא מכהן כעורך משנה של European Journal of Information Systems, ו-JMIR mJealth & uHealth. מחקריו של פרופ' שוורץ מופיעים בכתבי עת מובילים במערכות מידע. בין הספרים שכתב: Heterogenous Cooperating Systems, Internet-based Organizational Learning and Knowledge Management, והאנציקלופדיה לניהול ידע. פרופ' שוורץ היה חוקר אורח במספר אוניברסיטאות בעולם, ובהן אוניברסיטת קולומביה בניו יורק, אוניברסיטת מונש באוסטרליה, אוניברסיטת ויקטוריה באוסטרליה, אוניברסיטת קנטרבורי בניו-זילנד, ואוניברסיטת צינג-הואה בטייוואן.

תקציר

סיווג נתונים היא משימה שכיחה בעולם נתוני העתק: חברות קמעונאות משתמשות במודלי סיווג כדי לחזות אילו לקוחות יינטשו את החברה; בנקים וחברות פיננסיות משתמשים במודלים אלו לזהות הונאות; וספקי שירותי בריאות נעזרים בכלי סיווג לאבחון מחלות. בעיות סיווג מאתגרות יותר כאשר הן מופעלות על נתונים אודות התנהגות אנושית וחברתית, ובפרט כאשר נתונים אלו הם טקסטואליים. דוגמאות לכך הן סיווג מאמרי דעה, סיווג ציוצים, וסיווג ראיונות.
מאמר זה עוסק בסיווג תוכן התנהגותי טקסטואלי על ידי מסגרת חדשנית המבוססת על שילוב בין אדם למכונה. מסגרת זו, המכונה "ניתוח מושגי קוגנטיבי", מאפשרת לנתח תוכן באמצעות ניתוח המושגים המגדירים אותו, הנגזרים מחשיבה אנושית ולא ממוחשבת. המחקר המוצג הוא מחקר בין-תחומי הנשען על תורת המושגים, על למידת מכונה, ועל מחקר איכותני.

המחברים מודים לקרן ג'רמי קולר על התמיכה במחקר.



מבוא

משימת סיווג מתייחסת לפעולת השמת תווית לדוגמאות מתחום הבעיה. דוגמה קלה להבנה היא סיווג דואר אלקטרוני כ"דואר זבל" (ספאם) או "לא דואר זבל". משימת סיווג יכולה להיעשות באופן איכותני (ידני), או באופן כמותי וממוכן על ידי מכונה לומדת (Machine Learning). סיווג ממוכן נפוץ בעולם נתוני העתק (Big Data), הקיימים בתחומים רבים, ובהם בריאות, פיננסים, צרכנות ועוד.

עם התפתחות המחקר על נתוני עתק, נולדה הבנה שיש הבדלים מהותיים בין ניתוח נתונים "רשמיים" לעומת נתונים "התנהגותיים". נתונים רשמיים מתייחסים למידע אובייקטיבי, שהוא לרוב מבני וכמותי. מנגד, נתונים התנהגותיים, או בשמם הפורמלי Behavioral Big Data (BBD) (Shmueli, 2017), הם לרוב מסדים א-מבניים (לדוגמה טקסטואליים) המכילים מידע אודות התנהגויות, פעולות ואינטראקציות אנושיות וחברתיות. נתונים אלו שונים באופן מובהק מנתונים רשמיים, משום שהפרטים האנושיים שנאמדים באמצעותם מקיימים אינטראקציה מתמשכת ומודעת עם נתוני העתק עצמם. אינטראקציה זו עשויה לכלול מאפיינים של כוונה, הטעיה, רגש, הדדיות, ניתוב או צורות אחרות של התנהגות אנושית המשתקפת בטקסטים (Abbasi, Sarker, & Chiang, 2016; Shmueli, 2017). דוגמה בולטת לנתונים התנהגותיים היא שיח רשתי הנוצר על ידי משתמשים (User-Generated Content; UGC), שבו הפרטים המשתתפים בשיח נחשפים למסד הנתונים (כלומר השיח), ומוסיפים עליהם מידע משלהם.

מאפיין בולט נוסף של מידע התנהגותי טקסטואלי הוא העובדה שהוא מכיל לרוב טקסט סמוי, רעיונות סובייקטיביים וסמלים בלתי מילוליים (Abbasi & Chen, 2008; Vaast, Davidson, & Mattson, 2013).

נתונים אלו נקראים ומפורשים בקלות על ידי בני אנוש, אך עלולים להיות מפוענחים באופן שונה על ידי בני אדם שונים, או באופן דומה אך מסיבות שונות. לפיכך, סיווג מידע התנהגותי טקסטואלי בעזרת מכונה לומדת היא משימה סטטיסטית מאתגרת.

במאמר זה אנו מציעים מסגרת לשילוב מערכת טיעונים אנושית אינטליגנטית כשלב מקדים ללמידת מכונה. מסגרת זו עשויה להוביל לשינוי פרדיגמטי באפקטיביות של מערכות סיווג של נתוני עתק התנהגותיים. מסגרת זו, המכונה "ניתוח מושגי קוגניטיבי" (Cognitive Concept Analysis; CCA), מיישמת מודל חדשני לסיווג נתוני עתק התנהגותיים טקסטואליים באמצעות ניתוח המושגים הקוגניטיביים המגדירים אותם, כלומר מושגים המוגדרים ומנומקים על ידי אדם במקום על ידי מכונה.

המחקר הנוכחי נשען על תורת המושגים, למידת מכונה ועל מחקר איכותני, ומתבסס על גישה בין-תחומית לסיווג טקסטים. עקרון המפתח המודגש במסגרת זו הוא שסיווג טקסטואלי של נתוני עתק התנהגותיים, הקשורים בעיקר לקוגניציה האנושית, עשוי להשתפר בצורה משמעותית באמצעות תשומות אנושיות בשתי רמות. ברמה הראשונה, על ידי סיווג בינארי של דוגמאות מתחום הבעיה לצורכי אימון (לדוגמה, סיווג דואר אלקטרוני כ"דואר זבל" או "לא דואר זבל"), וברמה השנייה באמצעות מתן נימוק לכל החלטת סיווג (לדוגמה: "זהו דואר זבל, כי הוא מכיל בקשת פרטי חשבון בנק ממקור לא ידוע"). הרמה הראשונה עומדת בבסיסו של כל אלגוריתם ממוכן לסיווג טקסט, בעוד שהיעדר הנימוקים מאחורי כל החלטת סיווג – ברמה השנייה – מוביל לרוב לביצועים פחותים יותר של אלגוריתמים אלו בכל הקשור בסיווג נתוני עתק התנהגותיים.

אלגוריתמים אוטומטיים לסיווג טקסט מבצעים לרוב "השטחה" של המסמכים למערכת של מאפיינים לקסיקליים. החיסרון העיקרי של סיווג טקסט אוטומטי הוא היכולת המוגבלת ללמוד דבר מה מעבר לטקסט האימון בנוגע למאפיינים הלקסיקליים, ביחס לטקסטים הסמויים, ולדרך שבה טקסט מפוענח באופן סובייקטיבי על ידי קוראים שונים. כדי להתגבר על מגבלה זו של סיווג טקסט אוטומטי, חוקרים שונים הציעו לייצג מסמכי טקסט באופן היררכי (אופקי) במקום שטחי (אנכי). ייצוג היררכי של טקסט עשוי לקלוט את הרובד העמוק של המסמכים ולהרחיב את יכולות למידת המכונה מעבר לטקסט האימון. כך למשל, חוקרים שונים טוענים כי ניתוח מושגי פורמלי (Formal Concept Analysis; FCA), ששורשיו נעוצים בתורת המושגית הקלאסית ובתורת הקבוצות (Pospescu, 2004), עשוי להיות כלי להבנה ולייצוג של מסמכים באופן היררכי (Priss, 2006). אף על פי כן, ניתוח מושגי פורמלי, כפי שעולה מעצם הגדרתו כ"פורמלי", הוא ייצוג מתמטי של המושג ואינו קשור לתהליך הקוגניטיבי בנוגע לנימוקי סיווג המושג (Priss, 2006). ככל הנראה, התחום המדעי היחיד שעוסק באופן מתודולוגי מובהק במושגים קוגניטיביים הוא תחום המחקר האיכותני. במחקר איכותני ניתוחי תוכן ונושא הם תהליכים של תיעוד נושאים (או מושגים) בתוך נתונים טקסטואליים. אולם ניתוחים אלה מבוצעים באופן ידני ולכן לא ניתן להעריך אותם בקלות או ליישם אותם במקרה של נתוני עתק התנהגותיים טקסטואליים.

ניתוח מושגי קוגניטיבי מציע גישת ביניים לאלגוריתמים ממוכנים ומחקר איכותני. כדי לדייק את סיווג הטקסט האוטומטי, בניתוח זה נעשה שימוש במיקור המונים לאגידת תובנות איכותניות מחד (כלומר הנמקה של תהליכי סיווג), ואוטומטיזציה של תהליכי קידוד איכותניים בעזרת למידת מכונה מנגד. ניתן לבחון גישה זו משתי זוויות שונות: מבחינת תהליך למידת המכונה, מושגים פורמליים מוחלפים במושגים קוגניטיביים המייצגים נאמנה את עולם התוכן. מנקודת ההשקפה של הניתוח האיכותני, קידוד של מומחים איכותניים נעשה כעת על ידי ההמונים ולאחר מכן על ידי אלגוריתמים של למידת מכונה. כך נעשה שימוש אוטומטי בנימוקי מטא-דאטה כדי להגדיר מושגים קוגניטיביים, המעוצבים בצורה אנלוגית לנושאים האיכותניים.

רקע ספרותי

סיווג טקסט

סוגיית סיווג הטקסט היא נושא מרכזי במחקרים רבים העוסקים בכריית נתונים, בלמידת מכונה ובמידענות,

וכן בקרב קהילות של מדעי המחשב. סוגיה זו ניצבת בליבם של יישומים מגוונים כמו דיאגנוזה רפואית (Akay, 2009), סינון דואר זבל (Carvalho & Cohen, 2005; Li & Yamanishi, 1997), קטלוג חדשות (Lang, 1995), כריית דעות ורגשות (Liu & Zhang, 2012; Pang & Lee, 2009) ומחקרים נוספים.

סיווג טקסט דומה באופיו לסיווג מסדים כמותיים מבניים (כלומר טבלאיים), אלא שהוא דורש עיבוד מקדים של המידע הטקסטואלי הא-מבני, על מנת לשטחו למסד מבני. שיטות נפוצות לייצוג מבני של מידע טקסטואלי הן שיטת השק-מילים (bag-of-words) וייצוג וקטורי על ידי שיטת הטמעת מילים (Word embeddings) (Levy & Goldberg, 2014). ייצוג זה מאפשר הפעלת שיטות סיווג אוטומטיות ידועות, כגון עץ החלטה, רגרסיה לוגיסטית, ורשתות נוירונים. עם זאת, על אף שפעולת סיווג טקסט באופן כזה נפוצה, הספרות המקצועית מצביעה על אתגרים ובעיות בטכניקות למידה, בייחוד בישומן בניתוח נתוני עתק התנהגותיים (Aue & Gamon, 2005; Blitzer, Dredze, & Pereira, 2007; Li & Tsai, 2013; Pang & Lee, 2009; Read, 2005; Smiraglia & van den Heuvel, 2013; Stock, 2010; Turney, 2002; Yahav, Shehory, & Schwartz, 2018). בפרט, המחקרים טוענים כי דיוק סיווג הטקסט עשוי להיות מושפע במידה רבה מאופן בחירת נתוני האימון למודל, המונחים הלקסיקליים המופיעים בהם וההקשר שלהם. מונחים לקסיקליים הם לרוב בעלי משמעות עמומה בהקשרים שונים (או אפילו זהים). בד בבד, משמעות זהה עשויה לעטות תצורות שונות כאשר היא כתובה על ידי אנשים שונים. מזווית הראייה של פעולת הלמידה הממוכנת, אלגוריתמים של כריית מידע דורשים רמה מסוימת של דמיון בין מסמכי אימון ומסמכים חדשים. מכאן שעל מסמכי אימון לכלול את מכלול המאפיינים או המונחים שבהם נעשה שימוש. למרות זאת, העושר של המונחים הלקסיקליים בדומיינים ובנושאים, לצד עלות התיוג, מאתגרים את מרבית האלגוריתמים ללמידה המצויים כיום.

כדי לתת מענה לאתגרים אלו, חוקרים מחפשים שיטות להעשרת הידע והמידע הנובעים ממערכת אימון (ראו סקר מאת Liu & Zhang, 2012). צעד בכיוון זה הוא השימוש במילונים מקוונים כמו WordNet (Miller, et al., 1990) או Thesaurus (Mohammad, Dunne, & Dorr, 2009), או שימוש במסד חיצוני כדוגמת BERT או ויקיפדיה לאימון הטמעת מילים (Allahyari et al., 2017; Devlin, et al., Lee, & Toutanova, 2018; Li, Sun, & Datta, 2013; Malo, et al., 2011; Pérez-Rodriguez, et al., 2016). במקרים כאלו אוסף המונחים הלקסיקליים שיזוהו בשלב האימון יהיה גדול יותר מהמונחים המופיעים בטקסט הלמידה בלבד, דבר שיתרום לשיפור סיווג טקסט. עם זאת, מחקרים מראים כי היכולת של מודלים אלו ליצור הכללות של ידע מוגבלים, זאת לאור הישענותן על מידע ערוך והגדרות פרמטרים (Srivastava, et al., 2014; Zhang, Lee, & Radev, 2016).

כיוון אחר לכריית טקסט בהקשר זה, וככל הנראה הרלוונטי ביותר למחקר הנוכחי, נשען על תורת המושגים (Hjørland, 2009). חוקרים שונים בחנו כיוון מחקרי זה (S.-T. Li & Tsai, 2013; Malo et al., 2011) בדגש על השימוש בתיאוריה מושגית פורמלית לנימוק מושגים מעורפלים (ראו ספרות נפרדת בהמשך).

סיווג טקסט ביישומי נתוני עתק התנהגותיים

שיטות לסיווג טקסט יושמו בעבר כדי לשאוב ידע תגובות, בלוגים, ביקורות מקוונות וכיוצא באלה. הסוגיה הנפוצה ביותר בהקשר סיווג טקסט של נתוני עתק התנהגותיים היא ככל הנראה ניתוח רגשות (Nasukawa & Yi, 2003), המכונה גם סוגיית כריית דעות. ניתוח רגשות וכריית דעות מייצגים מרחב גדול של בעיות, המוגדרות באופנים שונים וכוללות למשל מיצוי דעות, מחקר סובייקטיביות, ניתוח רגשות, ועוד. המטרה הסופית של ניתוח רגשות היא לגלות מה בני אנוש חושבים או מרגישים כלפי מוצרים, שירותים, פרטים, אירועים, מאמרי חדשות ונושאים שונים. באופן מתודולוגי, המטרה מושגת לרוב באמצעות טכניקות חדשניות לסיווג טקסט, הנשענות במידה רבה על מודלים מיוחדים מבוססי מילונים.

חוקרים שונים סקרו היבטים של ניתוח רגשות המשפיעים על יכולת אלגוריתמים ממוכנים לסווג רגשות (Li & Tsai, 2013; Liu & Zhang, 2012; Pang & Lee, 2009). הספרות המחקרית מצביעה על ההשפעה הניכרת של הדינמיקה והמבנה של טקסט שנוצר על ידי משתמש, בעיקר בכל הנוגע לדעות ולרגשות. באופן ספציפי, הספרות מדגישה את האתגרים המציבים מאפייני שיח על אלגוריתמי למידת מכונה כגון ציניות, חיקוי, משפטים חלקיים וציטוטים. למעשה, משמעות כיוון מחקרי זה הוא שהתוכן שנוצר על ידי משתמש אינו מספק כדי להבין את השיח המקוון ואת הדינמיקה שלו באופן אלגוריתמי.

תורת המושגים

לפי התיאוריה האריסטוטלית הקלאסית, מושג מייצג באופן תמציתי מערכת הכוללת תנאים מספקים להגדרת קשרים או העדרם בתוך מערכת זו. תומן ואחרים (Toman, Tesar, & Jezek, 2006) גורסים שמושגים ממלאים תפקיד משמעותי ביקום הידע, והם חשובים במיוחד באיתור מאפיינים משמעותיים בשפות טבעיות. לאור זאת, כפי שמציין היורלנד (Hjørland, 2009) בעבודתו פורצת הדרך על שימוש בתורת המושגים, יש חוסר עקביות מהותי בין ההגדרות של המושגים והטייתם, כפי שמופיעה בספרות בנושא זה.

היורלנד מציע חלוקה לארבע "משפחות" של המשגה: היסטוריזציה, אמפיריות, רציונליות ופרגמטיזם. רשמית, היסטוריזציה לא נחשבת לתיאוריית ידע אלא לתיאוריה של פרשנות טקסט, ולכן היא אינה כלולה בהמשך דיוננו במסגרת זו. אמפיריות מתייחסת לרעיון ביסוס ידע על סמך תצפיות בשילוב עם תיאוריה. אמפיריות נשענת על הדמיון בין 'דברים' ועל היחס בין 'דברים' ומילים. בטכניקות למידת מכונה, השימוש בלמידת אלגוריתמים מעמיקה (RNN) לניסוח מלאכותי של מונחים לטנטיים נחשבת כטכניקה אמפירית.

רציונליות היא תיאוריה הגורסת כי מושג מוגדר באופן פורמלי באמצעות מאפייניו או המבנים הלוגיים שלו. ניתוח מושגי פורמלי הוא דוגמה מובהקת לפרקטיקה רציונלית (Priss, Formal Concept Analysis; FCA. 2006). ניתוח מושגי פורמלי מספק כלים להבנת גוף הידע על ידי ייצוגו כהיררכיה של מושגים או כמסגרת מושגית, שבה מונחים מורכבים מפורקים למושגים פשוטים הבונים אותם. מודלים של סיווג, הנשענים על תיאוריה מושגית, כוללים ניתוח מושגי פורמלי, אך זהו ניתוח מעורפל מפני שהוא מגדיר את גבולות המושגים בהתאם להקשרם של המסמכים (Li & Tsai, 2013; Malo et al., 2011). הביקורת המרכזית על ניתוח מושגי פורמלי עולה בהקשר ליישומו הלוקה בדומיינים הקשורים בעיקר לקוגניציה האנושית, דבר הדורש בניית מודלים באופן קפדני וזהיר. כך פריס (Priss, 2006): "בלשנים עשויים לטעון שמושגים פורמליים שונים בהחלט מתהליכים קוגניטיביים הקשורים לשפות טבעיות. זו הסיבה מדוע יישומים נפוצים של ניתוח מושגי פורמלי בבלשנות מתמקדים יותר במבנים פורמליים המצויים בלקסיקון ובמילונים מאשר בתופעות בלשניות קוגניטיביות. לאור זאת, אין להבין ניתוח מושגי פורמלי כניתוח פורמלי של מושגים (אנושיים) אלא כשיטה מתמטית המשתמשת ב"מושגים פורמליים".

לפי היורלנד (Hjørland, 2009), המשפחה התיאורטית האחרונה היא פרגמטיות, ועליה אנו נשענים במחקר זה. תיאוריית הפרגמטיות מגדירה מושגים מעבר לדמיון התוכני (לפי גישת אמפירית) או החוקים הלוגיים (לפי הגישה הרציונלית) של המונח. לפי תיאוריה זו, מושגים נחשבים כסוגים זהים של מחשבה, שפה ומערכות סימבוליות השונות מבחינה פונקציונלית, ולכן הם עשויים לתאר באופן טוב יותר מצבים קוגניטיביים ומונחים אנושיים. במידה מסוימת, אם כי נושא זה טרם נדון בספרות המקצועית, תיאוריית הפרגמטיות היא הבסיס לתיאוריית הכללות מבוססות-הסבר (Explanation-Based Generalization; EBG Theory) ולניתוח איכותני (Bulmer, 1979; Krippendorff, 2012).

הכללה מבוססת הסבר

הכללה מבוססת הסבר היא תיאוריה הממנפת ידע מובנה כדי להסביר מדוע דוגמה מסוימת קשורה למושג. הכללה מבוססת הסבר מתבצעת באמצעות משימות: 1) הצגת

דוגמה; 2) הצגת הכללה (מונח); 3) הנמקה כיצד הדוגמה משרתת את ההכללה. הכללה מבוססת הסבר יושמה בעבר כדי להנחות סיווג טקסט. המחקרים הראשונים על הכללה מבוססת הסבר (DeJong & Mooney, 1986; Mitchell, Keller, & Kedar-Cabelli, 1986), התמקדו בהוספת מטא-דאטה מובנה לדוגמאות אימון מושגיות. יישומים עדכניים יותר נשענים על קוראים חיצוניים כדי לבנות ידע נרחב. לדוגמה, סטריבסטה ואחרים (Srivastava, Labutov, & Mitchell, 2017), הציגו למידה מושגית מונחית באמצעות עיבוד שפה טבעי (NLP) בהקשר של איתור דואר זבל. הם הגדירו מערכת הכוללת שבעה מושגים, והציבו בפני נציגי ההמונים את הדרישה "להנדס" מושגים אלו באופן הפוך כדי שיהוו דוגמאות דואר אלקטרוני. במסגרת מחקרם, עובדים חדשים התבקשו לאחר פעולה זו ליצור מטלות הוראה העושות שימוש ב-NLP כדי להסביר את המושגים המוצגים בדואר האלקטרוני. סטריבסטה ואחרים הוכיחו שמושגי למידה המבוססים על משימות הוראה דורגו מבחינה היררכית ברמה גבוהה יותר מאשר מושגים פורמליים. ממצאים אלו מעודדים את השימוש במושגים קוגניטיביים כדי לשפר שיטות סיווג טקסט אוטומטיות. בנוסף על תיאוריית הכללה מבוססת הסבר, חוקרים שונים מכירים בצורך לקשור בין אלגוריתמים של רשתות נוירונים ובין תהליכים קוגניטיביים (Qian, et al., 2016; Xie, et al., 2016). כך למשל, הם דנים בשיטות לסמן בראשי תיבות תגים טקסטואליים מבוססי רשתות נוירונים באמצעות מטא-דאטה חיצוני של מערכת מושגים חשובים, שהוגדרו מראש ונבנו באופן ידני.

ניתוח איכותני

ניתוח תוכן הוא שיטת מחקר לסיווג ולקידוד מערכתי של טקסט. ניתוח תוכן מאפשר לבחון כמויות גדולות של מידע טקסטואלי כדי לזהות מגמות ומאפיינים של שימוש במילים, תדירות מילים, היחסים ביניהן והמבנה שלהן. לאור גישתו הפרגמטית של היורלנד (Hjørland, 2009), ניתוח נושאי מאפשר לחשוף את הסיבות, הדעות והמניעים העומדים מאחורי טקסט כתוב (Grbich, 2012). בקצרה, ניתוח תוכן שואף לפשט וליצור משמעות של המציאות שהתוכן מייצג. פיתוח של סיווג פשוט לשימוש או מערך קידוד הוא צעד ראשון בניתוח (Saldaña, 2015). לאור זאת, ניתוח תוכן קשור לזיהוי, לקידוד, לסיווג ולתיוג מאפיינים עיקריים בנתונים (Patton, 2005). בתהליך הסיווג, הנתונים מפורקים לחלקים ומקודדים על ידי יד אנוש לפי קטגוריות מושגיות, הנתונים מקובצים, ונערכת השוואה מול חלקים דומים מתצפיות אחרות. המטרה היא לנסח באופן אינדוקטיבי מושגים תיאורטיים – הכללות על תהליכים אנושיים התקפים למשתתפים פרטניים שונים.

כדי להמחיש את ההבדלים בין הגישות האוטומטיות לגישה האיכותנית, נתמקד לדוגמה בניתוח סנטימנט. גישות אוטומטיות לניתוח סנטימנט ייעשו באחת משתי דרכים (Pang & Lee, 2009): הגישה הלא מפוקחת, והגישה המפוקחת. הגישה הלא מפוקחת היא גישה מבוססת לקסיקון, ועל פיה משפטים יסווגו לתחושות בעזרת השוואת המילים במשפט ללקסיקון רגשות אפריורי. לפי הגישה המפוקחת, קומץ מהמשפטים יסווג תחילה לתחושות בעזרת מקודדים אנושיים. לאחר מכן, אלגוריתם אוטומטי יאומן על הסיווג האנושי כדי לחזות את הסיווג של יתר המשפטים.

שיטת ההכללה מבוססת הסבר תהיה דומה לשיטה המפוקחת, אלא שאוסף הדוגמאות המסווגות ייאסף בשיטה שונה. במקרה כזה מקודדים אנושיים יתבקשו לתת משפטים התואמים לסיווג שהוגדר מראש (במקום סיווג למשפט קיים). בסיום האיסוף ייווצר מסד אימון הדומה לשיטה המפוקחת, ועליו יאומן אלגוריתם הסיווג האוטומטי.

הגישה האיכותנית, מאידך, מוסיפה לשאלת הסנטימנט רובד נוסף המנמק את השיוך של משפט לתחושה. לפי גישה זו, אותו רגש יכול להתבטא באופנים שונים: אימון, לדוגמה, יכול להתבטא באופן לקסיקלי (מילים), ובאופן סמנטי (סימני פיסוק, קריאה, אייקונים וכו'). מעבר לכך, הוא יכול להתייחס לעצמים שונים: אימון עצמי, אימון בממשל, אימון בחברה. הגישה האיכותנית תבדיל בין סוגים אלו על ידי שיוכם לקטגוריות מידע שונות.

לסיכום, הפער בין מה שמציעות שיטות אוטומטיות קיימות לבין מה שנעדר מן התהליך בא לידי ביטוי

במטא-אנליזה של תהליך הסיווג – רמת הנימוק. שיטות קיימות, כולל אלו "המשטחות" את הנתונים הטקסטואליים, אלו האחריות על הארגון ההיררכי, ואלו החושפות את המרחב המבני החבוי של הנתונים, מבוססות כולן על נתונים קיימים, ולעיתים מתבססות גם על מקורות מידע קיימים קודמים (כמו מילונים ואונטולוגיות). מנגד, גישות איכותניות מתבססות על הנמקה של סיווג ידני, אך אלה מעולם לא נבחנו או אוששו ברמה המקדימה של יצירת מקבץ תגים אוטומטי ראשוני – ומעולם לא שולבו כחלק ממסגרת תיאורטית פורמלית.

מסגרת ומטרות

מהנאמר כאן עולה שמושגים תורמים לשיפור סיווג טקסט, שכן הם מרחיבים את המרחב החבוי במסד טקסטואלי. אולם, כפי שמציין פריס (Priss, 2006), השימוש במושג פורמלי המנוסח עבור דומיין פורמלי, הנהוג במסדים רשמיים, מציב אתגר כאשר מדובר בניתוח נתוני עתק התנהגותיים. מקדד אנושי נדרש לרוב לקרוא את הטקסט ולספק תובנות שאינן מצוינות באופן מובהק בטקסט.

במחקר זה אנו מציעים מסגרת לשימוש במיקור ההמונים בשילוב עם אלגוריתמים של למידת מכונה כדי להגדיר מושגים *קוגניטיביים* אנושיים. במושגים אלו ייעשה בשלב השני שימוש כולל כדי לשפר סיווג טקסטים התנהגותיים. שיטת העבודה המוצגת היא כדלקמן (ראו תרשים מספר 1):

שלב עיבוד הנתונים: בהינתן מסד אימון התנהגותי וטקסטואלי, השתמש במיקור חוץ על מנת:

1. לתייג את המסמך, כלומר להציב סיווג לכל רשומה.
2. לנמק את הסיווג על שימוש בשפה טבעית – נימוקים אלו יאוגדו לכדי מטא-דאטה.

כנהוג במיקור חוץ, השתמש במספר מתייגים לכל רשומה במסד.

שלב ניתוח הנתונים:

3. עיבוד המטא-דאטה ייעשה על ידי כלים אוטומטיים ליצירת מושגים קוגניטיביים. דוגמה לכלים אוטומטים מתאימים היא ניתוח נושאים, או ניתוח נושאים מבוסס הטמעת מילים.
4. השתמש במכונה לומדת על מנת ללמוד את הקשר בין מסד האימון לאוסף המושגים המגדירים אותו.

שלב החיזוי:

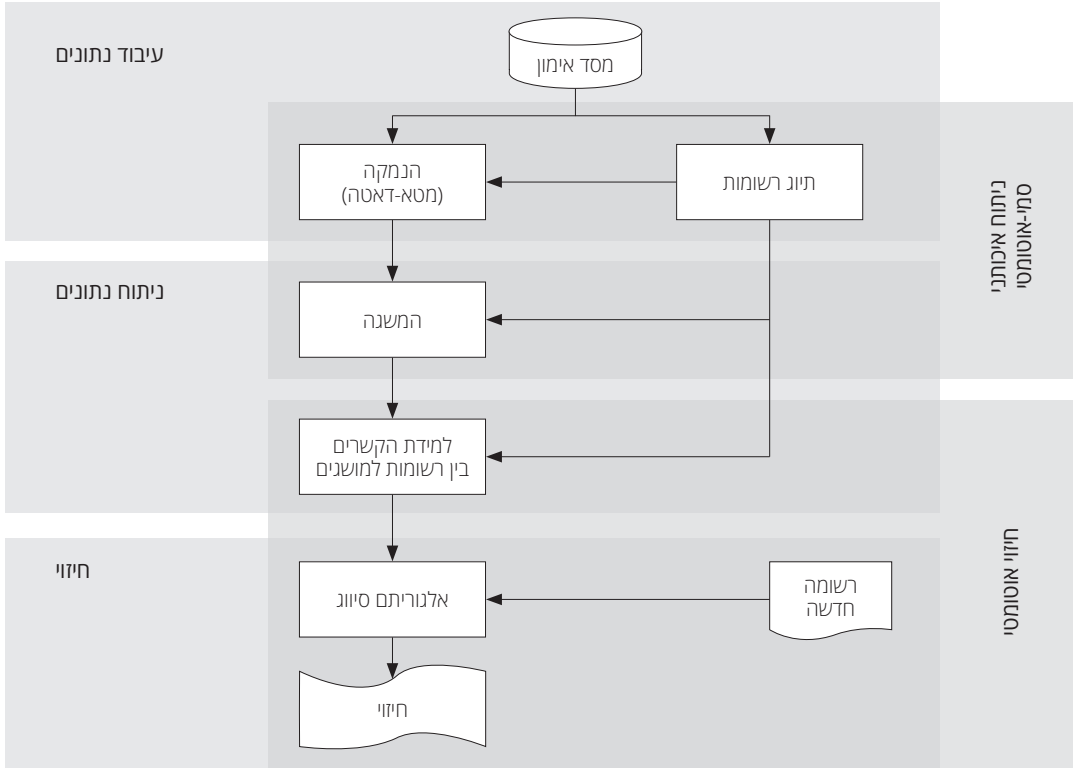
4. השתמש במודל שנוצר בסעיף 4 על מנת לחזות תיוג של מסמכים חדשים.

כדי להדגים את מסגרת הניתוח המושגי הקוגניטיבי, נתייחס למשימת תיוג הערות פוגעניות שעלו בתגובה לעמוד חדשות בפייסבוק. המיקוד הוא ברשימת התגובות למאמרי החדשות: בשלב האימון כל תגובה תתויג על ידי מספר מתייגים ממיקור ההמונים ותסווג כ"אדיבה" או כ"פוגענית", בציון סיבת הסיווג הסובייקטיבית של כל מתייג. דוגמה לנימוק סובייקטיבי עשויה להיות: "הערה זו היא פוגענית משום שהיא מכילה מילים גסות" או "הערה זו אינה מתורבתת משום שהיא גורמת לי לחוש מאוים". נימוקים אלו, כמוסבר לעיל, יאוגדו לכדי מטא-דאטה.

בשלב השני, המטא-דאטה ינותח על ידי אלגוריתם לזיהוי נושאים שיקבץ את ההסברים על מנת ליצור מושגים של *פוגענות* ואדיבות בתגובות מקוונות בפייסבוק. מושגים לדוגמה הקשורים לפוגענות עשויים לכלול "שימוש במונחים גסים" או "העלאת רגש אי נוחות". מושג לדוגמה הקשור לאדיבות עשוי להיות "הוספת מידע חיוני". תהליך למידת הקשר בין המושגים לרשומות במסד הטקסטואלי בדוגמה זו יכול להתבסס על טכניקות להעשרת נתונים (Liu & Zhang, 2012). לשם המחשה, טכניקה פשטנית אך רלוונטית היא "פירוק" כל מושג למערכת של אוצר מילים המייצגות אותו. למשל, המושג "שימוש במונחים גסים" עשוי להכיל מערכת שלמה של מונחים גסים שפורסמו בתגובה למאמרי חדשות. לבסוף, יותאם אלגוריתם סיווג שימפה את הרשומות למושגים ולסיווג.

בשלב החיזוי ייעשה שימוש באלגוריתם הסיווג החדש על מנת לתייג רשומות *כפוגעניות* או *אדיבות* באופן אוטומטי.

תרשים 1: מסגרת ניתוח מושגי קוגניטיבי



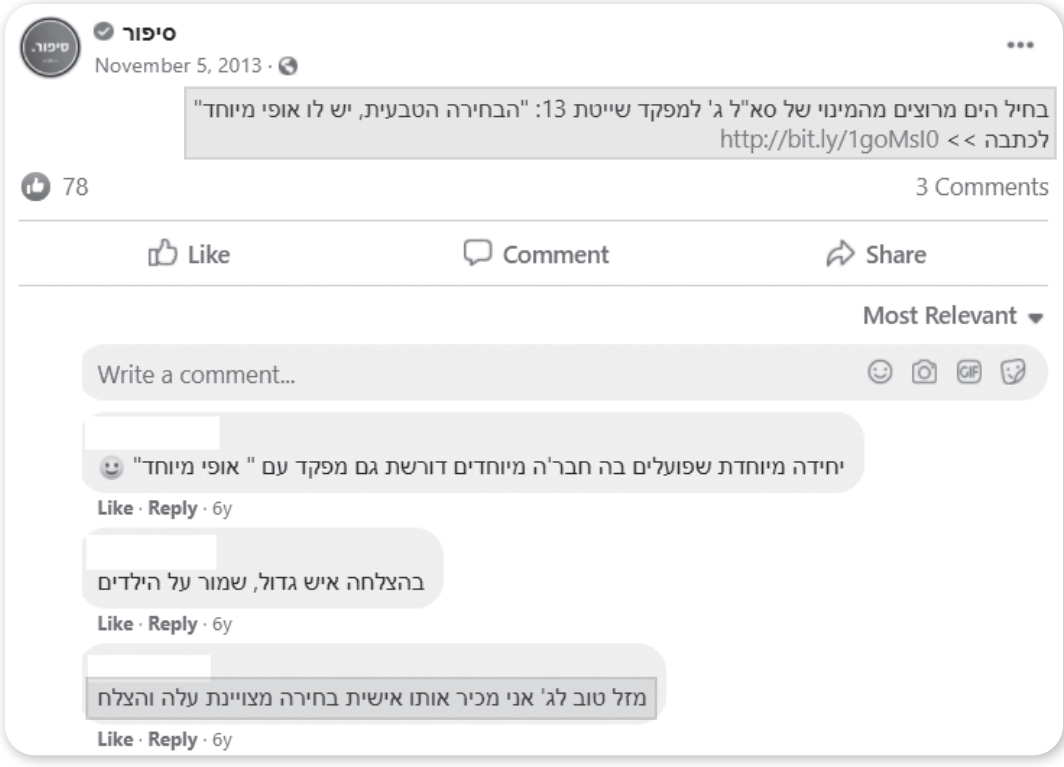
מקרה בוחן: סיווג הדלפות מידע דרך תגובות בפייסבוק

בפרק זה נדגים את מסגרת הניתוח המושגי הקוגניטיבי באמצעות תגובות על מאמרי חדשות בעברית בפייסבוק. הבעיה שנגדיר היא כדלקמן: בידיעות חדשותיות רבות קיים חסיון מידע על חלק מהפרטים ולכן הם מצונזרים. כך לדוגמה, שמות של אישי צבא בכירים ולוחמים ביחידות מיוחדות מוחלפים במאמרי חדשות באות הראשונה של השם הפרטי (לדוגמה, סא"ל ג'). במקרים כאלו, תגובות המתפרסמות על ידי משתמשים פרטיים עשויות (לרוב ללא כוונה) לחשוף נתוני מידע מסווגים ולכן להפר את כללי הצנזורה. במאמרנו הקודם (Schwartz, Yahav, & Silverman, 2017) אנו מספקים דוגמאות לחשיפת מידע מסוג זה. תרשים 2 מציג דוגמה לחשיפת מידע מסוג זה. במאמר החדשות בתרשים נכתב "בחיל הים מרוצים מהמינוי של סא"ל ג' למפקד שייטת 13". אחד המגיבים כותב: "מזל טוב לג' אני מכיר אותו אישית בחירה מצוינת עלה והצלח". על אף כי המגיב לא חשף בצורה מפורשת את שם המפקד החדש, הוא סיפק מידע על הקשר בינו ובין המפקד **באופן לא אנונימי**, שעשויה להביא לחשיפת פרטי המפקד.

מערך הנתונים לחקר המקרה כלל 48 מאמרי נתוני מידע מסווג, שהכילו 3,538 הערות מסומנות בסך הכול. ארבעה קוראים בלתי תלויים[1] תייגו את ההערות לכאלו

1 המקודדים היו סטודנטים בפקולטה למדעי החברה באוניברסיטת בר אילן.

תרשים 2: דוגמה לחשיפת מידע מצונזר ברשת הפייסבוק



שיש בהן פוטנציאל לחשיפת מידע, ולכאלו שאינן חושפות מידע. השאלה הבאה שהנחתה את החלטת הסיווג על ידי הקוראים הייתה: "בהתבסס על תגובה זו, האם אתה מאמין שהמגיב מכיר את זהות האדם החסוי?". לאחר מכן נעשה שימוש בשיטת דלפי כדי להגיע לתמימות דעים בקרב המסווגים. המסווגים התבקשו גם לזהות את מאפייני ההערה שהובילו למסקנתם: הנימוק לבחירת התיוג שלהם. המטא-דאטה של הנימוקים נותחו כדי להגדיר את סוגי (או "מושגי" במונחי המאמר) חשיפת המידע על ידי מגיבים. מחקר כמותי זה הניב תשעה מושגים. בשל קוצר היריעה, אנו מציגים ארבע דוגמאות בלבד ואת ההערות הייצוגיות הנוגעות בהן (רשימה מלאה אצל שוורץ ואחרים (Schwartz et al., 2017)). כל מושג שימש לאחר מכן כבסיס לבניית אונטולוגיה תלוית מושג ו/או מערכת של כללי צפי. כללי הצפי והאונטולוגיות נבנו באופן סמי-אוטומטי, כפי שמודגם בכל אחד המושגים לעיל:

1. **זיהוי מובהק.** חשיפת שם הפרט (בפריט חדשות, "פגוש את סרן ד', קצין מצטיין בלוחמה אלקטרונית, שניצל משירות בצבא אירן;" הערה, "סרן דוד בשיאן...").
על סמך מושג זה יצרנו כלל הצפי הבודק אם ההערה כוללת שם. אוסף השמות שנכלל לצורך כלל זה נלקח מתוך מילוני שמות בעברית.
2. **זיקה משותפת.** חשיפת היכרות עם הפרט, לדוגמה על ידי הזכרת מקום מגורים, עבודה, או פעילות פנאי לצד הפרט החסוי (פריט חדשות, "סגן ראשון ת' קיבל את כנפי התעופה שלו היום...;" הערה, "ביישוב שלי הוא גיבור-על").
על סמך מושג זה נבנתה אונטולוגיה הכוללת מילים המציינות מיקום, וכן את שמות כלל הערים והשכונות הגדולות בישראל (מתוך מסד המפות הממשלתי).
3. **ביטוי הומור.** חשיפת היכרות עם הפרט באמצעות הומור (צחקוק, חייכן) (פריט חדשות, "... זה הסיפור של סמל נ';" הערה, "חחח, עאלק סמל נ', אחי!").
על סמך מושג זה נבנה (ידנית) לקסיקון סנטימנט הכולל מילות הומור.

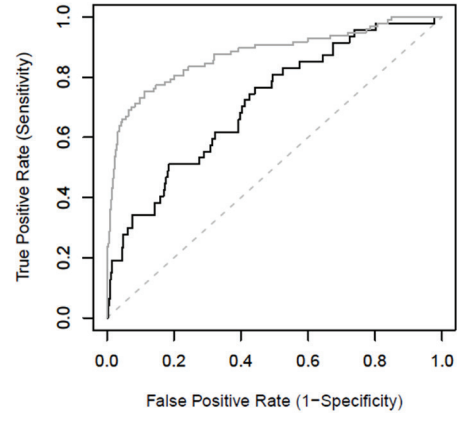
4. **סמן סמיוטי.** חשיפת היכרות אישית עם הפרט על ידי חזרה על השם המצונזר עם דגשים סמיוטיים המצביעים על זיהוי (נקודות/סימני קריאה/סלנג) (פריט חדשות, "... זה הסיפור של סמל נ';" הערה, "סמל נ....!!! כל הכבוד אחי. בהצלחה").

על סמך מושג זה נבנה כלל המחלץ סמנים סמיוטיים מהמשפט.

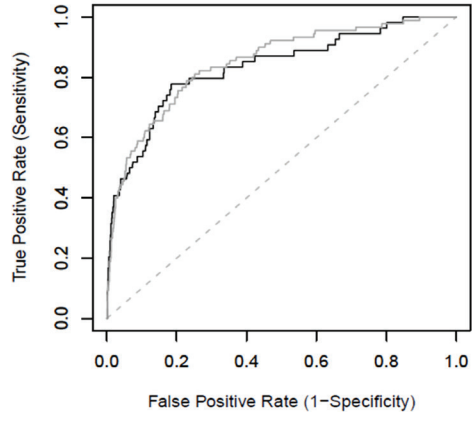
לבסוף, עיצבנו מודל למידה המבוסס על מושגים אלה והשווינו אותו לאסטרטגיית "שק המילים" הקלאסית לצד ניתוח דקדוקי. מלבד "שק המילים" והניתוח הדקדוקי, מודל הלמידה כלל גם עמודות מידע על השתייכות לאונטולוגיות וכללי הצפי שנבנו (לדוגמה: עבור המושג השני, הוכנסה עמודה עם ערכים בינאריים המציינים אם נעשה שימוש במילים המופיעות באונטולוגיית המיקום). במאמר זה השתמשנו בשיטת רגרסיה לוגיסטית כדי לסווג הערות של "הפרת פרטיות". כדי לאמוד את ביצועי המסווגים, השתמשנו בעקומות ROC עם חישוב סטטיסטי C. עקומות ה-ROC שימשו כעזר חזותי כדי לתאר את השוני בין שיעור חיובי כוזב ושיעור חיובי אמיתי תחת ערכי חיזוי שונים של הרגרסיה הלוגיסטית. תרשים 3 מציג את עקומות ה-ROC שהשווינו. עקומות ה-ROC צוינו בנפרד עבור נתוני אימון (60%) שעליהם נבנה מודל החיזוי ועבור נתוני האימות (40%) (החלוקה לאימון ואימות נעשתה באופן רנדומלי) שעליהם נבדקת יכולת החיזוי של המודל פועל. ממצאי המחקר עולה שהמסווג "הקלאסי" נכשל בחיזוי בפועל של הפרת צנזורה. ממצא זה אינו מפתיע, שכן גודל מערכת האימון היה קטן יחסית ולא ייצג נאמנה את עולם התוכן של תגובות למאמרי חדשות בפייסבוק. לעומתו, המסווג המבוסס על ניתוח מושגי קוגניטיבי, הפגין ביצועים דומים ביחס לחיזוי על נתוני אימון ואימות, וביצועים סבירים בהתחשב בגודל המדגם הקטן (C~ 81%). מסקנת המחקר הייתה שהשימוש במושגים אנושיים כדי לעבד מידע חדש משפר באופן משמעותי סיווג טקסט בהשוואה למסווגים "קלאסיים".

לסיכום, במקרה בוחן זה מגוון הדרכים שבהן ניתן לשבור צנזורה בתגובות לידיעה חדשותית הוא רב ועצום, ואינו טמון בהכרח באוסף המילים המופיעות בתגובות אלו. לאור זאת, מכונה לומדת שנשענה על המרחב הלקסיקלי והמבני של התגובות בלבד לא מסוגלת לזהות ולסווג דוגמאות אלו. במקרה זה, וכמוהו במסדים התנהגותיים אחרים, היה צורך במסווג אנושי כדי לנסח מושגי מפתח הפורסים את מגוון הדרכים לשבור צנזורה על ידי תגובות. המודל שהוצע במאמר זה מאפשר למסווג ללמוד היטב את מגוון המושגים באופן סמי-אוטומטי, וכן מאפשר למידה יעילה ממדגם קטן יחסית בלי צורך באיסוף מסד אימון גדול ומקיף.

תרשים 3: ביצועי שני המסווגים על מערכות אימון (אפור) ואימות (שחור)



א. מסווג מבוסס אסטרטגיית "שק מילים"



ב. מסווג מבוסס ניתוח מושגי קוגניטיבי

דיון

מספר חוקרים הציעו בעבר להשתמש במושגים כדי לשפר ביצועים של מודלים אוטומטיים לסיווג טקסט. המסגרת המחקרית שלנו מאמצת טיעונים אלה, אך היא שונה מהותית מבחינת תיאורטית ופרקטית. מבחינה תיאורטית, הספרות המקצועית נשענה על גישה רציונלית, שלפיה הגדרת מושג באופן פורמלי נקבעת באמצעות מאפייניו. בניגוד לכך, המסגרת המוצעת לעיל לאור הגישה הפרגמטית, גורסת שמושגים אינם מקובעים על ידי חלוקה לוגית אלא על ידי סיווגם הפונקציונלי. הלכה למעשה, הספרות המקצועית מתמקדת בניתוח תוכן המקור. המחקר הנוכחי מציע ניתוח של רמת המטא-דאטה של תוכן המקור – מערכת הנימוקים בנוגע לתיוג הרשומות במסד.

ניתוח מושגי קוגניטיבי הוא חשוב, משום שהוא מציע דרך מעמיקה יותר ללמידה של מסמכי טקסט, המשלבת בתוכה ידע אנושי איכותני לצד למידת מכונה. למעשה גישה זו מציעה ניתוח איכותי סמי-כמותי, המשפר את יכולות הסיווג מחד, ומייעל את התהליך האיכותני מאידך. באופן תיאורטי ניתן להכליל את התובנות שלנו בנוגע לטקסט גם לפעילויות אחרות כמו דירוג טקסטים ואחזור מידע. לשם כך, מסגרת הניתוח המושגי הקוגניטיבי מציעה כיווני מחקר חדשים לחוקרים העשויים להשתמש בניתוח מושגי קוגניטיבי ביישומים חדשים, שאותם לא ניתן לחקור בכלי סיווג טקסט הקיימים כיום. מודל זה גם פותח כיוונים חדשים בפני חוקרים תיאורטיים, העשויים לפתח מתודות חדשות לכריית ההמונים למטרת מיצוי מטא-דאטה בנושא הנמקה של סיווגים של נתוני עתק התנהגותיים, מתודות לעיבוד מטא-דאטה באופן כמותי וללא עיבוד ידע, ומתודות לשיטות המושגים על מנת לשלבם במודלי מכונה לומדת.

ממצאי שני המחקרים שלנו הדגימו כיצד ומתי גישת הניתוח המושגי הקוגניטיבי עשויה לשפר ביצועי מסווג של נתונים התנהגותיים טקסטואליים. מקרה הבוחן שהוצג נבחר כדי להדגים מאפייני שיח, ובהם חוסר בהירות, ציניות, משמעות חבויה, רגשות, כל אלה מאפיינים שבני אדם מסוגלים לקרוא בקלות ולהבין, אך אלגוריתמים של למידת מכונה עשויים להתעלם מהם. לא כל נתוני עתק התנהגותיים הם טקסטואליים, ולא כל נתוני עתק התנהגותיים טקסטואליים הוא עמומים ותלויי התנהגות. אולם בעולם המאופיין בכמויות הולכות וגדלות של טקסטים התנהגותיים מקוונים שיש לחקור ולסווג, השימוש בלמידת מכונה למחקר טקסטים יפיק תועלת רבה ממסגרות המשלבות אינטליגנציה אנושית חיה. לאור זאת, מסגרת הניתוח המושגי הקוגניטיבי היא דוגמה בולטת לסימביוזה בין מכונה ואדם, ולאור סימביוזה זו תמשיך הקוגניציה האנושית להנחות למידה אוטומטית ותהליכי סיווג.

ד"ר ענבל יהב inbalyahav@tauex.tau.ac.il

רשימת מקורות

Abbasi, A., & Chen, H. (2008). CyberGate: a design framework and system for text analysis of computer-mediated communication. *MIS Quarterly*, *32*(4), 811–837.

Abbasi, A., Sarker, S., & Chiang, R. H. L. (2016). Big data research in information systems: Toward an inclusive research agenda. *Journal of the Association for Information Systems*, *17*(2), Article 3.

Akay, M. F. (2009). Support vector machines combined with feature selection for breast cancer diagnosis. *Expert Systems with Applications*, *36*(2), 3240–3247.

Allahyari, M., Pouriyeh, S., Assefi, M., Safaei, S., Trippe, E. D., Gutierrez, J. B., & Kochut, K. (2017). A brief survey of text mining: Classification, clustering and extraction techniques. *ArXiv Preprint ArXiv:1707.02919*.

Aue, A., & Gamon, M. (2005). Customizing sentiment classifiers to new domains: A case study. *Proceedings of Recent Advances in Natural Language Processing (RANLP)*, *1*(3.1), 1–2.

Blitzer, J., Dredze, M., & Pereira, F. (2007). Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, 440–447.

Bulmer, M. (1979). Concepts in the analysis of qualitative data. *The Sociological Review*, 27(4), 651–677.

Carvalho, V. R., & Cohen, W. W. (2005). On the collective classification of email speech acts. *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 345–352.

DeJong, G., & Mooney, R. (1986). Explanation-based learning: An alternative view. *Machine Learning, 1*(2), 145–176.

Devlin, J., Chang, M.W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *ArXiv Preprint ArXiv:1810.04805.*

Grbich, C. (2012). *Qualitative data analysis: An introduction*. Sage.

Hjørland, B. (2009). Concept theory. *Journal of the Association for Information Science and Technology*, *60*(8), 1519–1536.

Krippendorff, K. (2012). *Content Analysis: An Introduction to Its Methodology*. Sage.

Lang, K. (1995). Newsweeder: Learning to filter netnews. In *Machine Learning Proceedings 1995.* Elsevier, 331–339.

Levy, O., & Goldberg, Y. (2014). Dependency-based word embeddings. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 302–308.

Li, C., Sun, A., & Datta, A. (2013). TSDW: Two-stage word sense disambiguation using Wikipedia. *Journal of the Association for Information Science and Technology*, *64*(6), 1203–1223.

Li, H., & Yamanishi, K. (1997). Document classification using a finite mixture model. *Proceedings of the Eighth Conference on European Chapter of the Association for Computational Linguistics*, 39–47.

Li, S.T., & Tsai, F.C. (2013). A fuzzy conceptualization model for text mining with application in opinion polarity classification. *Knowledge-Based Systems, 39*, 23–33.

Liu, B., & Zhang, L. (2012). A survey of opinion mining and sentiment analysis. In: *Mining text data*. Springer, 415–463.

Malo, P., Sinha, A., Wallenius, J., & Korhonen, P. (2011). Concept-based document classification using Wikipedia and value function. *Journal of the Association for Information Science and Technology*, *62*(12), 2496–2511.

Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., & Miller, K. J. (1990). Introduction to WordNet: An on-line lexical database. *International Journal of Lexicography*, *3*(4), 235–244.

Mitchell, T. M., Keller, R. M., & Kedar-Cabelli, S. T. (1986). Explanation-based generalization: A unifying view. *Machine Learning*, *1*(1), 47–80.

Mohammad, S., Dunne, C., & Dorr, B. (2009). Generating high-coverage semantic orientation lexicons from overtly marked words and a thesaurus. *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2*, 599–608.

Nasukawa, T., & Yi, J. (2003). Sentiment analysis: Capturing favorability using natural language processing. *Proceedings of the 2nd International Conference on Knowledge Capture*, 70–77.

Pang, B., Lee, L., & others. (2009). Opinion mining and sentiment analysis. *Foundations and Trends® in Information Retrieval*, *2*(1-2), 1–135.

Patton, M. Q. (2005). *Qualitative research*. Wiley Online Library.

Pérez-Rodriguez, R., Anido-Rifón, L., Gómez-Carballa, M., & Mouriño-Garcia, M. (2016). Architecture of a concept-based information retrieval system for educational resources. *Science of Computer Programming*, *129*, 72–91.

Priss, U. (2006). Formal concept analysis in information science. *Arist*, *40*(1), 521–543.

Qian, Q., Huang, M., Lei, J., & Zhu, X. (2016). Linguistically regularized lstms for sentiment classification. *ArXiv Preprint ArXiv:1611.03949*.

Read, J. (2005). Using emoticons to reduce dependency in machine learning techniques for sentiment classification. *Proceedings of the ACL Student Research Workshop*, 43–48.

Saldaña, J. (2015). *The coding manual for qualitative researchers*. Sage.

Schwartz, D. G., Yahav, I., & Silverman, G. (2017). News censorship in online social networks: A study of circumvention in the commentsphere. *Journal of the Association for Information Science and Technology*, *68*(3), 569–582.

Shmueli, G. (2017). Analyzing Behavioral Big Data: Methodological, practical, ethical, and moral issues. *Quality Engineering*, *29*(1), 57–74.

Smiraglia, R. P., & van den Heuvel, C. (2013). Classifications and concepts: towards an elementary theory of knowledge interaction. *Journal of Documentation*, *69*(3), 360–383.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, *15*(1), 1929–1958.

Srivastava, S., Labutov, I., & Mitchell, T. (2017). Joint concept learning and semantic parsing from natural language explanations. *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 1527–1536.

Stock, W. G. (2010). Concepts and semantic relations in information science. *Journal of the Association for Information Science and Technology*, *61*(10), 1951–1969.

Toman, M., Tesar, R., & Jezek, K. (2006). Influence of word normalization on text classification. *Proceedings of InSciT*, *4*, 354–358.

Turney, P. D. (2002). Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, 417–424.

Vaast, E., Davidson, E. J., & Mattson, T. (2013). Talking about Technology: The Emergence of a New Actor Category Through New Media. *MIS Quarterly*, *37*(4), 1069-1092.

Xie, R., Liu, Z., Jia, J., Luan, H., & Sun, M. (2016). Representation Learning of Knowledge Graphs with Entity Descriptions. *AAAI*, 2659–2665.

Yahav, I., Shehory, O., & Schwartz, D. (2018). Comments Mining With TF-IDF: The Inherent Bias and Its Removal. *IEEE Transactions on Knowledge and Data Engineering*, 31(3), 437-450.

Zhang, R., Lee, H., & Radev, D. (2016). Dependency sensitive convolutional neural networks for modeling sentences and documents. *ArXiv Preprint ArXiv:1611.02361*.