הפקולטה לניהול
ע"ש קולר
אוניברסיטת תל אביב

מאמר זה פורסם בכתב העת "חידושים בניהול", גיליון מס' 3, דצמבר 2018,
היוצא לאור על ידי הפקולטה לניהול על שם קולר באוניברסיטת תל אביב.

שימוש בנתונים מפורומים ומחיפוש באינטרנט לצורך חיזוי מכירות של כלי רכב[1]

תומר גבע   גל אסטרייכר-זינגר

ד"ר תומר גבע הוא חבר סגל במחלקה לניהול טכנולוגיה ומידע וראש התמחות ניהול נתוני עתק (Big Data) בפקולטה לניהול על שם קולר באוניברסיטת תל אביב. לפני כן היה ד"ר גבע חוקר אורח ב-NYU וב-Google. תחום המחקר שלו עוסק בפיתוח כלים לשימוש יעיל בכמויות גדולות של נתונים לצורך קבלת החלטות עסקיות. מחקריו פורסמו בכתבי עת מובילים בתחום מערכות מידע, ונתמכו על ידי גופים שונים, ביניהם הקרן הלאומית למדע. לד"ר גבע תואר שני ותואר שלישי מהפקולטה לניהול ע"ש קולר באוניברסיטת תל אביב, ותואר ראשון בהנדסת תעשייה מהטכניון.

פרופ' גל אסטרייכר-זינגר היא ראשת תחום ניהול טכנולוגיה ומידע וראשת התוכנית לתואר שני בניהול טכנולוגיה ומידע בפקולטה לניהול ע"ש קולר באוניברסיטת תל אביב. היא סיימה את לימודי הדוקטורט באוניברסיטת ניו-יורק והיא בעלת תארים במשפטים ובהנדסה מאוניברסיטת תל אביב ומהאוניברסיטה העברית בירושלים. במחקרה, פרופ' אסטרייכר-זינגר מתמקדת בהשפעה של מדיות חברתיות, רשתות חברתיות ומעורבות משתמשים על מסחר אלקטרוני ועל המודל העסקי של אתרי תוכן. מחקרה פורסם בכתבי העת המובילים במערכות מידע ובשיווק וזכה בפרסים בין-לאומיים רבים.

ניב עפרון - דירקטור הנדסה בגוגל.

יאיר שמשוני - מדען נתונים בכיר בגוגל.

תקציר

מחקרים רבים עושים שימוש בנתונים המתקבלים מאתרי מדיה חברתית לשם חיזוי תוצאות כלכליות לא-מקוונות, כגון מכירות. עם זאת, מחקרים שנערכו לאחרונה מצביעים על העובדה שנתונים אלו עשויים להיות כפופים למגבלות ולהטיות שונות אשר עלולות לפגוע בדיוק החיזוי. בה בעת, קבוצה הולכת וגדלה של מחקרים מראה כי מקור חדש של מידע מקוון - יומנים של מנועי חיפוש (search engine logs) - יכול לחזות תוצאות לא-מקוונות. אנו חקרנו את היחס בין שני מקורות המידע החשובים הללו בהקשר של חיזוי מכירות. תוך התמקדות בתעשיית הרכב, השתמשנו באינדקס המקיף של גוגל, החולש על פורומים של דיונים באינטרנט, וכן בנתונים המתקבלים ממגמות חיפוש באינטרנט. מצאנו כי הוספת נתונים ממגמות חיפוש למודלים המבוססים על מדיה חברתית, מודלים שהם פופולריים יותר כיום, משפרת באופן משמעותי את דיוק החיזוי. מצאנו גם כי מודלי חיזוי המתבססים על נתוני מגמות חיפוש שעלותם נמוכה יותר, מספקים דיוק חיזוי שלכל הפחות אינו נופל מהדיוק של מודלי חיזוי המתבססים על נתונים ממדיה חברתית.

1 המאמר מבוסס על עבודה שפורסמה ב-MIS Quarterly, 2017.

מבוא

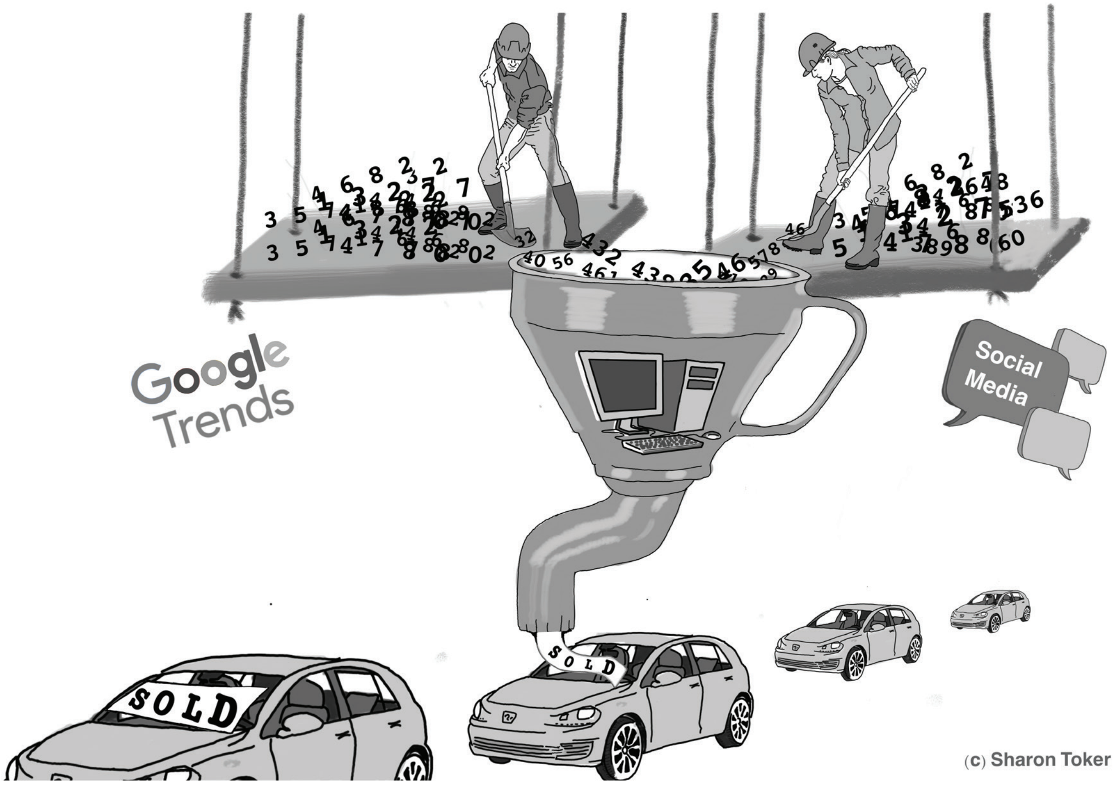
הזמינות, קנה המידה ועושר הפירוט של נתונים שמקורם במדיה חברתית עודדו חוקרים, וגם עוסקים במקצוע, לחקור אמצעים לשימוש בנתונים אלו כדי להסביר ולחזות תוצאות כלכליות לא-מקוונות. כתוצאה מכך, התפתח במחקר זרם דומיננטי המתמקד בשימוש בנתונים מאתרי מדיה חברתית, כגון בלוגים או פורומים לדיון באינטרנט, כמדד לפרסום מוצר מסוים "מפה לאוזן". מחקרים קודמים בנושא זה סיפקו עדות חוזרת להנחה שמספר האזכורים שמוצר מקבל במדיה החברתית, כמו גם הדעה המובעת בנתונים אלו, יכולים לחזות תוצאות לא-מקוונות (Choi and Varian, 2009; Wu and Brynjolfsson, 2009).

ניטור מדיה חברתית, שבאמצעותו חברות יכולות לקבל מידע על ידי סינון נתונים מאתרי מדיה חברתית, הפך גם הוא לנוהל רווח בתעשייה. דיווחים מהתעשייה מעריכים כי בשנת 2022 ישקיעו חברות יותר מ-9.5 מיליארד דולר בניטור מדיה חברתית.[2] ניטור מדיה חברתית נחשב ליקר פחות ממחקר שוק מסורתי, המחייב סקרים או קבוצות מיקוד. עם זאת, בפועל, יקר למדי לאסוף ולעבד נתונים ממדיה חברתית, במיוחד כאשר מפעילים נוהלי עיבוד תוכן מורכבים יותר, כגון ניתוח דעות (sentiment analysis).

העלות איננה האתגר היחיד העומד בפני חברות המנסות לנצל נתונים ממדיה חברתית מקוונת כדי לחזות תוצאות כלכליות. למעשה, דיווחים מהתעשייה מרמזים כי צרכנים מעטים מאוד תורמים הלכה למעשה לדיוני מדיה חברתית. לפי דיווחים אלה, רק כ-10% מהמשתמשים באתרי מדיה חברתית הם משתתפים פעילים ואילו מרבית התוכן מתקבל מ-1% של משתמשים בלבד.[3] מכאן שמרבית המשתמשים משקיפים בשקט מהצד או נשארים "סמויים" - תופעה המכונה "אי-שוויון השתתפותי" (participation inequality). על אף ששימוש במדגמים קטנים (למשל, בסקרים) כדי לקלוט את כוונות הצרכנים למטרות חיזוי אינו נדיר, מדגמים כאלה נבחרים בדרך כלל בקפידה, תוך שאיפה להבטיח ייצוג דעות של בסיס הצרכנים הכללי. יחד

2 http://www.marketsandmarkets.com/PressReleases/social-media-analytics.asp.

3 ראו http://www.nngroup.com/articles/participation-inequality/ וגם blog.elatable.com/2006/02/creators-synthesizers-and-consumers.html.



עם זאת, עבודות שפורסמו בשנים האחרונות הצביעו על כך שנתונים המתקבלים ממדיה חברתית מאופיינים בהטיות ובמגבלות משמעותיות, ובכלל זה: ייצוג בלתי מספק ומסולף של אוכלוסיית הצרכנים; ייצוג מוטה של ההערכה האמיתית של מהות המוצר/מותג; הטיות עקב הבדלים על פני המותג והמוצרים, כמו גם שינויים מוטי זמן; אפשרות למניפולציה מכוונת (לדוגמה אצל Moe and Schweidel, 2012; Moe and Trusov, 2011; Lovett et al., 2014). יתר על כן, רבות מההטיות הללו כוללות דינמיקות מורכבות ששיטות מידול שונות מתקשות לצמצמן.

מעניין לציין שגורם משותף, העומד בבסיס רבות מההטיות המדווחות בהקשר של נתונים ממדיה חברתית ואף תורם להן, הוא הנִראוּת של הפרסומים במדיה החברתית. זאת כאשר מחקרים קודמים הראו כי העובדה שאזכורים במדיה חברתית חשופים בפני אחרים משפיעה על נכונותם של משתמשים להשתתף בדיונים מקוונים, גורמת להם להביע דעות שונות מדעותיהם האמיתיות ויכולה אף להניע חברות, ואפילו לקוחות, לנסות להשפיע על התוכן המקוון בצורה מניפולטיבית.

בהקשר של מחקרי חיזוי, נוהג נפוץ לשם פתרון בעיות הנובעות מנתונים לא מושלמים הוא להעשיר את הנתונים במידע נוסף ובעל משמעות. עם זאת, חשוב שהנתונים הנוספים לא יסבלו מאותן הטיות או פגמים ספציפיים שהחוקר מנסה לצמצם. אנו התמקדנו כאן בנתונים ממגמות חיפוש באינטרנט בתור מקור נתונים מקוון המסוגל בפוטנציה להעשיר את הנתונים ממדיה חברתית. יומני מנועי חיפוש, הצוברים מיליארדי שאילתות פרטניות ממנועי חיפוש, הועמדו לרשות הציבור באמצעות כלים כגון Google Trends. היו שטענו כי נתונים מהחיפוש באינטרנט משקפים למעשה את "הכוונות האמיתיות" של הצרכנים (Wu and Brynjolfsson, 2009) וכי נתונים אלו יכולים לייצג את העניין שמוצר מסוים מעורר אצל הצרכנים (Hu et al., 2014). בהקשר שלנו, הדבר החשוב ביותר הוא שנתונים ממגמות חיפוש באינטרנט שונים מאוד מהנתונים המתקבלים ממדיה חברתית. שלא כמו בכתיבה במדיה חברתית, הגלויה לעין כול, החיפוש באינטרנט נעשה באופן פרטי. כאשר הצרכנים מחפשים באופן פרטי אין הם חוששים מהאופן שבו ייראו בעיני אחרים, אין להם מודעוּת (או שהמודעוּת שלהם מוגבלת מאוד) לפעילותם של מחפשים אחרים ואין להם יכולת להשפיע בצורה מניפולטיבית על מכירות. בנוסף, בהשוואה להשתתפות הפעילה בדיוני מדיה חברתית, החיפוש באינטרנט מנוהל על ידי פלח אוכלוסייה הרבה יותר גדול. אף שנתונים ממגמות חיפוש גרידא אינם יכולים להפחית מחומרת ההטיות והמגבלות הידועות של מדיה חברתית, טבעם הפרטי והשימוש הנרחב הופכים את הנתונים הללו למועמדים מבטיחים להתווסף למודלי חיזוי המבוססים על נתונים ממדיה חברתית.

מעניין שההשפעה ההדדית של שני מקורות מקוונים אלו של נתונים, והיחס בינה לבין חיזוי מכירות, קיבלו תשומת לב מועטה בלבד במחקרים שנערכו עד כה. הדבר מעלה שאלה העומדת בלב העבודה הנוכחית: האם נתונים ממגמות חיפוש באינטרנט יכולים להשלים את הנתונים ממדיה חברתית לצורך חיזוי מכירות? במיוחד, **שאלתנו הראשונה במחקר** בוחנת אם מודלי חיזוי, המבוססים על שילוב של נתונים ממגמות חיפוש באינטרנט עם נתונים ממדיה חברתית, עולים בתועלתם על מודלים המבוססים בנפרד על נתונים ממדיה חברתית או על נתונים ממגמות חיפוש באינטרנט. אם נתונים ממגמות חיפוש באינטרנט מציעים מידע חיזוי שימושי, המשלים את הנתונים המתקבלים ממדיה חברתית ולהפך, עשוי השילוב של שני מקורות הנתונים הללו להניב מודלי חיזוי מכירות מדויקים יותר. אף על פי כן, הסיכון שבהוספת מקור נתונים משני נעוץ בכך שאם הנתונים הנוספים אינם כוללים מידע נוסף בעל ערך, אזי ההוספה עלולה להביא לידי התאמת יתר (overfitting), ובסופו של דבר לפגוע בדיוק של החיזוי מחוץ למדגם (out-of-sample). לפיכך, העליונות של מודל המשלב את שני מקורות הנתונים תהיה מופגנת אם אכן שני מקורות הנתונים מכילים מידע שימושי ובלתי חופף.

גם אם מידע על מגמות חיפוש באינטרנט משמש מקור פוטנציאלי לנתונים שימושיים שניתן להשיג בעלות נמוכה ושעשויים לצמצם הטיות הקשורות לנתונים ממדיה חברתית, חשוב לציין כי לנתונים ממגמות החיפוש באינטרנט יש חסרונות משל עצמם. במיוחד, נתונים ממגמות חיפוש באינטרנט אינם עשירים כמו נתונים ממדיה חברתית ולכן יכולתם לשקף את דעתו של המשתמש קטנה מאוד. לדוגמה: בשעה שנפח החיפוש מציין את רמת העניין של המשתמש במוצר, כלל לא ברור אם העלייה בהתעניינותו נובעת מנסיבות חיוביות או שליליות (כגון השקת דגם חדש של מכונית או קריאה להחזרה - ריקול - של דגם מסוים). פרטים נוספים על נתונים ממגמות חיפוש באינטרנט ניתן למצוא בהמשך בפרק "ספרות קשורה".

העובדה שלנתונים ממנמות חיפוש באינטרנט ולנתונים ממדיה חברתית יש יתרונות ומגבלות שונים מספקת את הרקע לשאלתנו השנייה במחקר. בנינוד לשאלה הראשונה שבוחנת את שני מקורות המידע בתור משלימים, השאלה השנייה בוחנת אותם בתור חלופות. שאלה זו בוחנת אם מודלי חיזוי מכירות, המשתמשים בנתונים ממנמות חיפוש באינטרנט, עשויים להשיג דיוק דומה או אף טוב יותר בהשוואה למודלי חיזוי מכירות נפוצים יותר, המתבססים על הדעות ועל נפח הנתונים המתקבלים ממדיה חברתית. אם ניתן להראות כי מודלי חיזוי מכירות המשתמשים בנתונים זולים ממנמות חיפוש באינטרנט (הזמינים ללא תשלום ב-Google Trends) מניבים דיוק חיזוי שאינו נופל או אף עולה על דיוק החיזוי המתקבל בעזרת נתונים ממדיה חברתית, אפשר יהיה להציע למנהלים לשקול להחליף את הנתונים ממדיה חברתית בנתונים ממנמות חיפוש באינטרנט.

בבואנו לחקור את ההשפעה ההדדית של שני סוגי הנתונים הנידונים, נתמקד במיוחד בתעשיית הרכב. מאחר שאצל צרכנים רבים רכישה של מכונית היא הוצאה פיננסית ניכרת והחלטה עליה צריכה להיות מושתתת על שפע של מידע, אנו משערים כי גם למדיה החברתית וגם לחיפוש באינטרנט יהיה תפקיד חשוב בקבלת החלטת הרכישה. בנוסף, תקציב השיווק העצום של תעשיית הרכב (אשר השקיעה בשנת 2015 בלבד, לפי הערכה, 15.1 מיליארד דולר בפרסום), יחד עם חשיבותו לכלכלה, הופכים את הענף הזה לכר ניסויים מעניין, בעל השלכות מעשיות חשובות.[4]

כדי להציג נתונים ממנמות חיפוש באינטרנט, השתמשנו במידע שקיבלנו מדוחות שאילתות חיפוש (search query logs) של גוגל. כדי להציג נתונים ממדיה חברתית, הסתמכנו על האינדקס המקיף של גוגל, המכסה פורומים של דיונים באינטרנט (להלן "נתוני פורומים"). למיטב ידיעתנו, אינדקס זה הוא המערך המקיף ביותר של נתוני פורומים אשר הועמד לרשות מחקר אקדמי בתחום זה.

מצאנו כי מודלי חיזוי, הכוללים הן נתוני פורומים והן נתונים ממנמות חיפוש באינטרנט, מספקים תחזיות מכירות מדויקות יותר בהשוואה למודלים המשתמשים בנתונים מבוססי פורומים בלבד, דבר המצביע על כך שהמידע הכלול בשני המקורות הוא גם שימושי וגם לא-חופף. עוד מצאנו כי מודלי חיזוי המתבססים על נתוני מנמות חיפוש זולים מספקים דיוק חיזוי שלכל הפחות אינו נופל מדיוקם של מודלי חיזוי נפוצים יותר, המתבססים על נתוני פורומים.

ספרות קשורה

בעבודה זו, הסתמכנו בעיקר על שני זרמי מחקר עיקריים. הזרם הראשון חוקר את כוח החיזוי ואת הכוח ההסברי (explanatory power) של מידול בעזרת שימוש בנתונים ממדיה חברתית; הזרם השני כולל גוף הולך וגדל של עבודות המתעדות את כוח החיזוי של נתונים ממנמות חיפוש באינטרנט. התייחסנו בקצרה לשני הזרמים ולאחר מכן דנו בהטיות ובמגבלות מוכרות בשימוש במדיה חברתית לצורך חיזוי, ובשאלה כיצד השילוב של שני מקורות הנתונים עשוי לצמצם את ההטיות הללו.

מידול בעזרת שימוש בנתונים ממדיה חברתית

זמינות הפלטפורמות של מדיה חברתית שבהן משתמשים יכולים להעביר ביניהם בפומבי מידע על מוצרים שונים - פלטפורמות כגון קבוצות דיון, פורומים ואפילו סקרי מוצרים באתרי אינטרנט של ספקים מקוונים - הביאה לגידול בפרסום מוצרים "מפה לאוזן". פרסום מפה לאוזן שונה מהפרסום המסורתי בין אדם לחברו, המתנהל לרוב בין צדדים המכירים זה את זה ונגישים באופן מוגבל. מחקרים קודמים בשיווק ובמערכות מידע הקדישו תשומת לב רבה להשפעות שיש לאזכורים של נתונים ממדיה חברתית על מכירות. (Duan et al., 2008a; Liu, 2006;Dewan and Ramaprasad, 2012 ; Dewan and Ramaprasad, 2009;Dhar and Chang, 2009 ;Chevalier and Mayzlin, 2006 ; Godes and Mayzlin, 2004; Chakravarty et al., 2010; Dellarocas et al., 2007;Gu et al., 2012 ; Duan et al., 2008b; Chen and Xie, 2008; Dellarocas, 2006 ;Zhu and Zhang, 2010 Hu et al., 2008; Forman et al., 2008).

בנוסף, מחקרים הראו כי רגשות (sentiments) המוזכרים במדיה חברתית עשויים להיות חשובים אף הם לצורך חיזוי

4 http://drivingsalesnews.com/auto-industry-expected-to-spend-15-1-billion-on-local-ads-this-year/

מכירות. יחד עם זאת, ממצאים בנושא זה אינם לגמרי חד-משמעיים. לדוגמה, Liu (2006) וכן (Duan et al 2008b). מצאו כי המכירות של מוצר מסוים הושפעו מנפח האזכורים במדיה חברתית, אך לא מהערך הרגשי שלהם וגם לא מדירוגי המשתמשים. לעומת זאת, במחקרים חדשים יותר, כגון המחקרים של Rui et al. (2012 ושל Chintagunta et al., (2011, דווח על רגשות בטקסט כגורם חשוב בהסבר המכירות.

כוח החיזוי של מגמות חיפוש באינטרנט

זרם המחקר השני שעליו הסתמכנו לצורך חיזוי של אירועים כלכליים וחברתיים מתמקד בשימוש ביומני מנועי חיפוש, המוכרים יותר כמגמות חיפוש באינטרנט. אף על פי שהחיפוש מתנהל בפרטיות, כלים כגון Google Trends העמידו את יומני החיפוש לרשות הציבור, כאשר הנתונים המוצגים הם ברמה מצרפית. יומנים אלו שימשו לחיזוי במחקרים שונים ובהקשרים שונים. (Choi and Varian, 2009; Choi and Varian, 2011; Wu and Brynjolfsson, 2009; Vosen and Schmidt, 2011; Ginsberg et al., 2008; Seebach et al., 2011; Goel et al., 2010). הסבר אחד לתועלת המופקת מנתונים מהחיפוש באינטרנט בחיזוי מכירות עתידיות הוצע על ידי Wu and Brynjolfsson, (2009), אשר טענו כי יומני מנועי החיפוש הם "אותות כנים (honest signals) של כוונות מקבלי ההחלטות". במילים אחרות, אם הקונים חושפים את כוונותיהם האמיתיות בקשר לרכישות, רמות המכירות העתידיות צפויות להתאים לכוונות אלו.

הטיות ומגבלות הקשורות לשימוש בנתונים ממדיה חברתית לצורך חיזוי

מידת הייצוגיות של אוכלוסיית הצרכנים: מקור פוטנציאלי ראשון להטיה מתייחס לשאלה מי בוחר להשתתף בדיוני מדיה חברתית והאם אותם אנשים מייצגים את האוכלוסייה הכללית. כפי שצוין לעיל, דיווחים מהתעשייה מצביעים על העובדה שרק מעט מצרכני התוכן של מדיה חברתית (כ-10% מכלל המשתמשים במדיה חברתית באינטרנט) תורמים הלכה למעשה לדיוני מדיה חברתית מקוונים, ואילו מרבית התוכן מיוצר על ידי לא יותר מ-1% של משתמשים.

אף שהנפח הגדול (במספרים מוחלטים) של משתמשי מדיה חברתית עשוי להפחית מחומרת הבעיה של גודל המדגם, חשש נוסף טמון בשאלה האם המדגם של משתתפים פעילים מהווה ייצוג טוב של כלל האוכלוסייה? אכן, יש עדות לכך כי מספרם הנמוך של הצרכנים המשתתפים בדיוני מדיה חברתית אינו משקף ייצוג אקראי של אוכלוסיית הצרכנים הכללית ואפילו לא של אוכלוסיית המשתמשים במדיה חברתית. (Dellarocas and Narayan, 2006; Moe and Schweidel, 2012).

ייצוג מוטה של הערכת המוצר: מחקר שנערך לאחרונה מראה כי גם כאשר משתמשים בחרו להשתתף בדיונים מקוונים, ייתכן כי העדפותיהם המוצהרות ודעותיהם היו שונות מהערכותיהם האמיתיות כלפי המוצר. במילים אחרות, האינטראקציה החברתית באינטרנט עלולה להטות את *מה* שהם כותבים. (Schlosser, 2005; Moe and Schweidel, 2012; Moe and Trusov, 2011).

שינויים תלויי-זמן: מחקרים שונים הראו כי מידע הנובע ממדיה חברתית סובל מהטיות ומשינויים בתוכן התלויים בזמן שחלף מאז תחילת הדיון בנושא. (Li and Hitt, 2008; Godes and Silva, 2012; Hong et al., 2014).

מניפולציה מכוונת: כמה מחקרים הצביעו על הפוטנציאל הטמון במניפולציה מכוונת בסקירות מקוונות במדיה חברתית, אשר מבצעים משתמשים פרטיים או חברות. (Dellarocas 2006; Mayzlin et al., 2014; Luca and Zervas, 2015).

השפעות מאפייני המוצר: עבודות קודמות הראו כי למוצרים ולמותגים שונים יש ייצוגים שונים באופן מהותי במדיה החברתית. (Lovett et al., 2014; Hong et al., 2014; Berger and Schwartz, 2011; Kronrod and Danziger, 2013; Berger and Milkman, 2012; Schulze et al., 2014).

טיפול בהטיות נתונים המושתתים על מדיה חברתית: ניתן להשתמש בטכניקות מידול וטיפול בנתונים כדי להפחית מחומרתן של כמה הטיות נתונים המושתתים על מדיה

חברתית. לדוגמה, טכניקות של ביטול מגמות (למשל, Canova, 1998) יכולות ליישב שינויים מסוימים, תלויי-זמן, בנפח וברגשות של אזכורים במדיה חברתית. השימוש במשתני דמה לפי מוצרים, או בנרמול לפי מוצרים, עשוי להסביר כמה מהשינויים על פני המוצרים השונים. יחד עם זאת, מרבית ההטיות שצוינו לעיל כרוכות בדינמיקות מורכבות, המחייבות שימוש בתהליכי מידול מורכבים (כגון התייחסות לגורמים כמו פרסומים קודמים במדיה חברתית ואישיות המשתמש במדיה זו. אלה משפיעים על מידת ההשתתפות של המשתמשים ומעוותים את הערכותיהם המוצהרות כלפי מוצרים). יתר על כן, במקרים מסוימים, ההבנה הקיימת של דינמיקת התהליך מוגבלת (כגון, מדוע משתמשים פרטיים, שאין להם תועלת מובהקת בדבר, מפרסמים סקירות מזויפות?), ולכן היא מונעת מחוקרים את היכולת להסביר את הדינמיקות הללו במודלים שלהם. במקרים אחרים, דינמיקות של תהליכים הן תלויות-זמן ואין לדעת אם ימשיכו להשפיע על התוצאות גם לאחר תום תקופת האימון (training) של מודל חיזוי. לסיכום, המאפיינים המורכבים, והלא-ברורים לעיתים, של הטיות המושתתות על מדיה חברתית עלולים להגביל את מאמצי המידול המתַקנים, במיוחד כאשר יש צורך ליישב הטיות רבות או מגבלות של נתונים בעת ובעונה אחת. ואמנם בפועל, במרבית המקרים קורה שנתונים ממדיה חברתית משמשים "כמות שהם", ללא כל התאמת מידול או עם התאמות מזעריות בלבד.

במקביל לטיפול בהטיות באמצעות תהליכי מידול, ישנו כלי נפוץ אחר המסוגל לצמצם את השפעת הנתונים הלא-מושלמים, והוא העשרת הנתונים במידע נוסף ובעל משמעות. אנו סבורים כי נתונים ממגמות חיפוש באינטרנט יכולים להשלים באופן זה את הנתונים ממדיה חברתית, כפי שיוסבר להלן.

שילוב בין נתונים ממגמות חיפוש באינטרנט לבין נתונים ממדיה חברתית

שתי תכונות של חיפוש מקוון מסבירות את החשיבות של נתונים ממגמות חיפוש באינטרנט בהקשר שלנו: (א) השימוש הנרחב בחיפוש מקוון; (ב) העובדה שהחיפוש המקוון מתנהל באופן פרטי. תכונות אלו, בשילוב עם קלות ההשגה של נתונים ממגמות חיפוש באינטרנט, הופכים את מקור המידע הזה למועמד מבטיח להפחתת חומרת הטיות החיזוי הקשורות לנתונים ממדיה חברתית. כפי שהוסבר לעיל, דרך אחת לשפר את דיוק החיזוי ולצמצם הטיות בנתונים הזמינים היא לחבר נתונים קיימים לנתונים נוספים שאינם כפופים לאותן הטיות. בהקשר שלנו, ניתן להשיג זאת על ידי העשרת הנתונים ממדיה חברתית בנתונים ממגמות חיפוש באינטרנט, אשר אינם כפופים להטיות ולמגבלות של הנתונים מהמדיה חברתית. באופן מפורש, מכיוון שהתנהגות החיפוש הפרטני של צרכנים אינה נחשפת בפני צרכנים אחרים, גורמים רבים, המטים התנהגות של משתמשים במדיה חברתית, אינם רלוונטיים להתנהגות החיפוש של צרכנים "אנונימיים". לדוגמה, הטיות רבות, שנדונו לעיל בסעיפים "מידת הייצוגיות של אוכלוסיית הצרכנים" ו"ייצוג מוטה של הערכת המוצר", הן תוצאות של חששות משתמשים מייצוג עצמי ושל מודעותם לפעילויות המשתמשים האחרים. לעומת זאת, כאשר מבצעים חיפוש מקוון, למשתמשים אין חששות ביחס לייצוגם העצמי ויש להם מודעות מוגבלת מאוד, אם בכלל, לפעילותם של המשתמשים האחרים. כתוצאה מכך, נתונים ממגמות חיפוש באינטרנט "עמידים" בפני הטיות רבות ובפני דינמיקות מורכבות, שמשפיעות ואף מסלפות במקרים רבים את הנתונים המתקבלים ממדיה חברתית. בנוסף, השימוש הנרחב בחיפוש מקוון הפך אותו לחסין הרבה יותר מפני בעיות הקשורות לגודל מדגם הנגרמות עקב רמות השתתפות נמוכות ועלולות להשפיע בפוטנציה על נתונים ממדיה חברתית. יתר על כן, מכיוון שבדרך כלל לא ניתן לנצל התנהגות חיפוש פרטנית כדי לבצע מניפולציות או להשפיע על מכירות, הנתונים ממגמות חיפוש באינטרנט אינם מושפעים מהטיות של "מניפולציה מכוונת" שדווחו לעיל.

על אף העובדה שמידע ממגמות חיפוש שימש בהצלחה במחקרי חיזוי קודמים ואינו צפוי לסבול מההטיות המשפיעות על נתונים ממדיה חברתית, בכל זאת חשוב לציין כי מקור נתונים זה מתאפיין גם הוא במגבלות ובחסרונות. חיסרון אחד בהשוואה למדיה חברתית שכבר צוין לעיל הוא שמידע ממגמות חיפוש לוקה בחוסר יחסי בכמות התוכן ואינו משקף את רגשות המשתמשים. לכן, נתונים ממגמות חיפוש באינטרנט עלולים לא לענות לשאלה אם עלייה בהתעניינות במוצר מסוים נובעת מנסיבות חיוביות או שליליות. חיסרון שני הוא שמידע ממגמות חיפוש, כמו זה שמספק הכלי Google Trends, זמין רק ברמה מצרפית ואילו חיפושים גולמיים אינם זמינים בדרך כלל לחוקרים.

לפיכך, נתונים ממנגמות חיפוש באינטרנט אינם עשירים וגרעיניים (granular) כמו נתונים ממדיה חברתית, הכוללים בדרך כלל פרסומים ברמת משתמש. זאת ועוד, היעדר גישה לחיפושים גולמיים מונע מחוקרים מלאתר ולנתח הטיות ומגבלות שייתכן שהן קיימות אך אינן ידועות כיום, ועשויות אף הן לאפיין נתונים ממנגמות חיפוש באינטרנט.

היתרונות והחסרונות של נתונים ממנגמות חיפוש באינטרנט, כאמצעי לשיפור החיזוי המבוסס על נתונים ממדיה חברתית וכמקור חלופי לנתונים אלה, הם הרקע לשתי שאלות המחקר הראשונות שלנו, שצוינו לעיל. *האם מודלי חיזוי, המבוססים על שילוב של נתונים ממגמות חיפוש באינטרנט ונתונים ממדיה חברתית, עולים בתועלתם על מודלי חיזוי המבוססים על נתונים ממדיה חברתית בלבד או על נתונים ממגמות חיפוש בלבד?* וגם, *האם מודלי חיזוי מכירות, המשתמשים בנתונים ממגמות חיפוש באינטרנט, יכולים להשיג דיוק דומה או אף טוב יותר, בהשוואה למודלי חיזוי מכירות נפוצים יותר, המתבססים על הדעות ועל נפח הנתונים המתקבלים ממדיה חברתית?*

נתונים וייצוג

במחקר השתמשנו בנתונים חודשיים של 23 מותגי מכוניות שנמכרו בארצות הברית במהלך 4 שנים, בין 2007 ל-2010 (ממוצע המכירות לחודש של כל המותגים עמד על 5,000 מכוניות). עשינו שימוש בשלושה מקורות של נתונים המתוארים להלן: מכירות, חיפוש באינטרנט ופורומים. יצוין כי בעקבות נוהג נפוץ (לדוגמה, ;Choi and Varian, 2009 Du and Kamakura, 2012; Seebach et al., 2011) התמקדנו במכירות ברמת מותגים במקום במכירות של דגם מכוניות מסוים.[5]

נתוני מכירות

השתמשנו ביחידות מכירות (unit sales) של מכוניות וטנדרים חדשים בארצות הברית לפי נתונים מאתר Automotive News (http://www.autonews.com/). Automotive News מספק נתוני מכירות ברמה מצרפית חודשית. הוא מקור ידוע למידע על מכירות רכב ושימש במחקרים שונים בנושא, כגון המחקרים של Choi and Varian (2009) ושל Du and Kamakura (2012). בהמשך, השתמשנו במונח *מכירות* כדי לציין את נפח המכירות של המותג i במהלך חודש t.

נתוני חיפוש באינטרנט

השתמשנו ביומני השאילתות של מנועי החיפוש בגוגל. מדובר באותם נתונים גולמיים שבהם משתמש גוגל באתר Google Trends (http://www.google.com/trends) כדי להציג מגמות של שאילתות במנועי חיפוש. באופן מפורש, אספנו את הנפח המדווח של שאילתות חיפוש בגוגל לגבי כל אחד ממותגי הרכב. על ידי בחירת אפשרויות הקטגוריות הרלוונטיות ב-Google Trends, הגבלנו את הנתונים לחיפושים שמקורם בארצות הברית ולחיפושים הקשורים לתעשיית הרכב. בהמשך, השתמשנו במונח *חיפוש* כדי לציין את נפח החיפוש של המותג i במהלך חודש t.

נתונים מפורומים

כדי להציג את הנתונים מפורומים השתמשנו בסריקת האינטרנט הנרחבת של גוגל. למיטב ידיעתנו, סריקה זאת היא הסריקה המקיפה ביותר של נתונים מפורומים אשר הועמדה לרשות מחקרים אקדמיים בתחום הנידון. באופן מפורש, קיבלנו נתונים מכל הפורומים שהתנהלו באנגלית שתויגו כפורומים על ידי גוגל. מקור זה כולל אתרי אינטרנט ייעודיים לדיונים מקוונים, נוסף על אתרי אינטרנט הכוללים מדורים שבהם המשתמשים יכולים לפרסם בפומבי דעות וסקירות ולהתייחס לתוכן קודם. (כדוגמת townhall-talk.edmunds.com, forums.motortrend.com, answers.yahoo.com, וכד').

בהתאמה לספרות העדכנית בתחום זה, השתמשנו בשני היבטים של נתונים מפורומים לגבי כל אחד ממותגי הרכב: מספר הפעמים שהמותג אוזכר בפורומים ("אזכורים בפורום") והרגשות (sentiments) שאזכורים אלו מעוררים ("רגשות הפורום"). כדי לייצג את אזכורי המותג i בפורומים בחודש t

---

5 המוטיבציה לנוהג זה כפולה: ראשית, כמות הנתונים ברמת מותג עשירה יותר מאשר כמות הנתונים לפי דגם המכונית. הסיבה השנייה, הזיהוי לפי מילת מפתח ברמת מותג מדויק הרבה יותר מאשר ברמת דגם מכונית. יחד עם זאת, ניתוח רובסטיות נעשה גם עבור חיזוי ברמת הדגם של המכונית.

(המצוינים על ידי אזכורים_בפורוםi,t), השתמשנו במספר הפרסומים החדשים בפורומים אשר מאזכרים את המותג i במהלך חודש t. כדי לייצג את הרגשות שעורר מותג i בפורומים במהלך חודש t (המצוינים על ידי רגשות_הפורום i,t), השתמשנו ביחס שבין סכום "האזכורים החיוביים" לסכום "האזכורים השליליים" לגבי המותג i בחודש t. כדי לתייג פרסומים בפורום כ"חיוביים" או "שליליים" השתמשנו בגישה לניתוח רגשות המבוססת-מילון, גישה פופולרית בספרות (ראו, לדוגמה, Berger and Milkman, 2012). באופן מפורש, השתמשנו במילונים מקיפים של מילים חיוביות ושליליות לרבות במילון הפסיכולוגי הידוע Harvard IV-4[6], וסיכמנו את המספר של פרסומים חדשים בפורום המאזכרים "מילים חיוביות" ושל אלה המאזכרים "מילים שליליות" - כל זה במקביל לאזכור המותג i במהלך חודש t. היתרונות של גישת המילון הם יכולת ההכללה ויכולת השחזור (בניגוד לתוכנות קנייניות או מבוססות "קופסאות שחורות").

מידול

הגדרות

המשתנה התלוי במחקר זה הוא מכירות i,t - מכירות רכב של מותג i בחודש t. כדי לחזות את המכירות של כל אחד מהמותגים בחודש t, השתמשנו בנתונים שהיו זמינים בחודש t-1. נוהג מקובל בתחום הוא לכלול נתונים מחודשים קודמים, כגון: מכירות מחודשים קודמים, נתונים מפורומים ונתונים ממנגנות חיפוש באינטרנט (כמוסבר לעיל), וכן נתונים הכלולים במדדי ייחוס (baseline). המידול בוצע על בסיס חודשי, החל מתקופה (lag) אחת של נתונים היסטוריים (חודש t-1), ובהדרגה הוכללו תקופות נוספות (עד חמש תקופות נתונים, לחודשים: t-5 ,... ,t-1).[7]

בעקבות מחקרים קודמים בתחום זה, השתמשנו בנתוני מדדי הייחוס (baseline) הבאים:

• עונתיות: מכירות באותו חודש, בשנה הקודמת (כלומר, מכירותi,t-12). (ראו לדוגמה Choi and Varian, 2009, 2011)

• מדד רגשות הצרכנים (Consumer sentiment index) (ראו לדוגמה Hu et al., 2014; Hymans et al., 1970)

• מחירי דלק (ראו לדוגמה Hu et al., 2014)

לאחר איסוף הנתונים, הגדרנו "מודל מדד ייחוס» (baseline) כמודל המשתמש בנתונים הבאים: רגשות הצרכנים, מחיר הדלק, עונתיות (מכירותi,t-12) ומכירות קודמות. מכאן נפנינו למדוד את מידת האינפורמטיביות של נתונים מבוססי-פורום ושל נתונים ממנגנות חיפוש באינטרנט, וכן את היתרון של הגדלת כמות הנתונים מבוססי-פורום באמצעות נתונים ממנגנות חיפוש באינטרנט. לשם כך הגדרנו כמה מודלים נוספים הכוללים סטים של נתונים שונים, כמתואר להלן: "מודל מבוסס-פורום» משתמש בנתוני מודל מדד הייחוס, נוסף על כמות האזכורים בפורומים; "מודל מורחב מבוסס-פורום» מוסיף לנתוני המודל המבוסס-פורום גם את נתוני רגשות הפורום; "מודל מבוסס-מנגנות חיפוש» מצרף לנתוני מודל מדד הייחוס את נתוני מנגנות החיפוש באינטרנט; ולבסוף, "מודל משולב מבוסס-חיפוש ופורום» משתמש בכל מקורות הנתונים שהוזכרו לעיל ומשלב אותם. טבלה 1 מסכמת את מערכי הנתונים השונים ששימשו בכל מודל חיזוי.

חיזוי והערכה

השתמשנו באלגוריתם הפופולרי של רגרסיה לינארית (LR). שיטה זאת שימשה ברוב המכריע של המחקרים הקשורים אשר ביקשו לחזות תוצאות כלכליות על בסיס נתונים מפורומים או על בסיס נתונים ממנגנות חיפוש באינטרנט. להבטחת רובסטיות של התוצאות, חזרנו על הניתוח שלנו גם בשיטות לא לינאריות, כגון (SVM), Neural Networks (Support Vector Machines NN), ו-Random Forest, וקיבלנו ממצאים דומים.

בעקבות הנוהג הנפוץ במחקר חיזוי, מדדנו את ביצועי המודל "מחוץ למדגם" (out-of-sample), כלומר השתמשנו במערך נתונים אחד כדי לאמן מודל ובמערך אחר כדי למדוד את ביצועיו. באופן מפורש, השתמשנו בגישת החלון הנע (Moving Window), שיטה שבה מאמנים מודלים על משך זמן קבוע (למשל 24 חודשים), כאשר בכל פעם מזיזים את מועד ההתחלה של תקופת האימון ביחידת זמן אחת ועושים ולידציה בתקופה אחת עוקבת. תוך יישום השיטה הזאת, השתמשנו ב-1/3 מהנתונים שלנו בתור מערך תיקוף

6 http://www.wjh.harvard.edu/~inquirer.
7 חשיבות השימוש בנתונים עם שיהוי, לצורך הקמת מודל חיזוי, נדון אצל Goel et al., (2010) ואצל(Lazer et al., 2014).

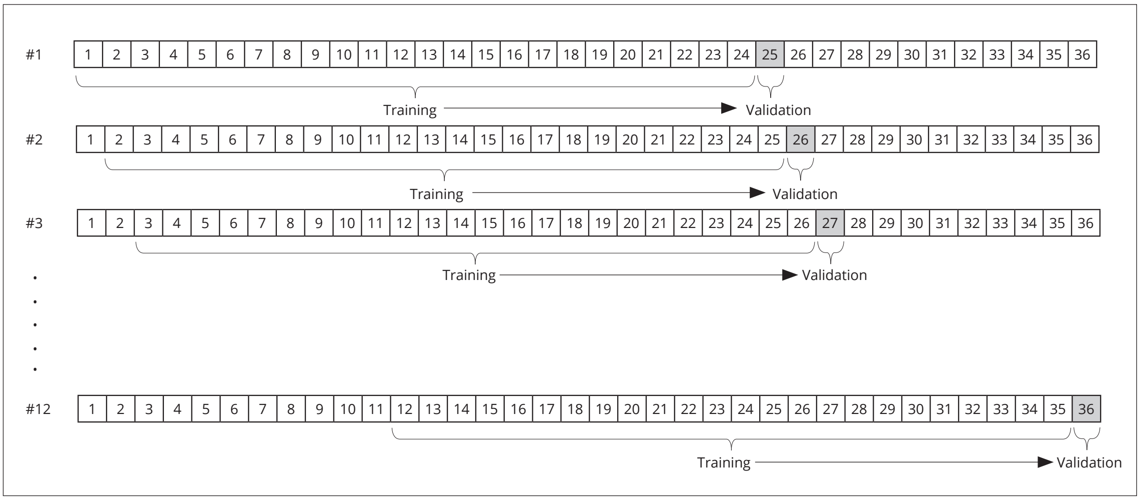
טבלה 1: הנתונים הכלולים בכל מודל

נתונים / מודל	מודל מדד ייחוס (benchmark Model)	מודל מבוסס-פורום (Simple Forum-based Model)	מודל מורחב מבוסס-פורום (Extended Forum-based Model)	מודל מבוסס-מגמות חיפוש (Search Trends-based Model)	מודל משולב מבוסס-חיפוש ופורום (Combined Model)
מכירות $_{i,t-1}$,..., מכירות $_{i,t-n}$	√	√	√	√	√
רגשות הצרכן $_{t-1}$	√	√	√	√	√
מחיר הדלק $_{t-1}$	√	√	√	√	√
מכירות $_{i,t-12}$	√	√	√	√	√
אזכורים_בפורום $_{i,t-1}$,..., אזכורים_בפורום $_{i,t-n}$		√	√		√
רגשות_הפורום $_{i,t-1}$,..., רגשות_הפורום $_{i,t-n}$			√		√
חיפוש $_{i,t-1}$,..., חיפוש $_{i,t-n}$				√	√

טבלה זו מתארת את המשתנים הכלולים במודלים השונים. לדוגמה: במקרה שמודל הייחוס יהיה מבוסס על שני lags (n=2) הוא יכלול: מכירות של דגם i בתקופות t-1 ו- t-2, וכן יכלול את מדד רגשות הצרכנים ואת מחירי הדלק בתקופה t-1 ואת המכירות של דגם i בתקופה t-12.

עצמאי (validation set). דיווחנו על ביצועים על בסיס כל תקופת התיקוף מחוץ למדגם (חודשים $t = 25,...,36$). במילים אחרות, מדדנו את הביצועים בכל חודש תיקוף t, תוך הפעלת מודל האימון במהלך 24 החודשים הקודמים (חודשים t–24 עד t–1). ציינו כי חודש t = 1 הוא ינואר 2008 וחודש t = 25 הוא ינואר 2010.[8] השימוש שלנו ב-Moving-Window לאימון ולוולידציה מתואר בצורה גרפית באיור 1. להבטחת רובסטיות התוצאות, ביצענו ניתוח דומה תוך שימוש בנישת החלון המתרחב (Expanding Window), וקיבלנו ממצאים דומים.[9]

איור 1: המחשת אימון ולידציה באמצעות Moving Window



8 הנתונים שלנו מתייחסים לתקופה שבין ינואר 2007 לדצמבר 2010. "איבדנו" נתונים שווי ערך ל-12 חודשים כאשר לקחנו בחשבון את העונתיות.

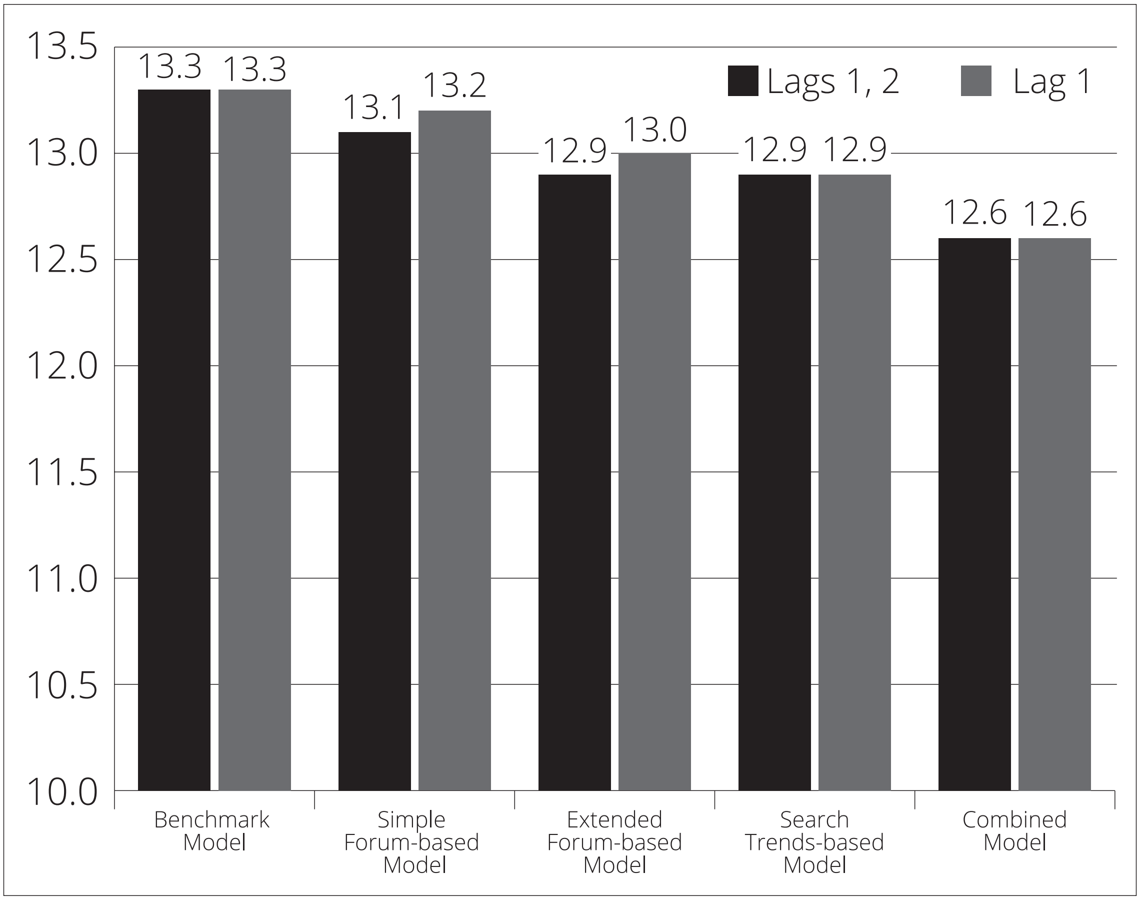
9 שיטה זו שונה משיטת החלון הזז בכך שתקופת הזמן הראשונה לאימון כל מודל היא מקובעת; אך התקופה האחרונה לאימון כל מודל, וכן תקופת הוולידציה, משתנות באופן דומה לשיטת החלון הזז.

תוצאות

איור 2 מציג את התוצאות שהתקבלו ממיצוע מדדי הדיוק על פני תקופת הוולידציה תוך שימוש בייצוגי נתונים שונים (כלומר, הסוגים השונים של מודלים שהוגדרו בטבלה 1). טבלה 2 מראה את ההפרשים בערכי MAPE (Mean Absolute Percentile Error) בין מודלים שהשתמשו במערכי נתונים שונים, תוך דיווח רווחי סמך המתבססים על bootstrapping (ראו Efron and Tibshirani, 1994). בטבלה זאת, ההפרש החיובי בין ערכי MAPE מצביע על דיוק חיזוי טוב יותר של מודל א' לעומת דיוק החיזוי של מודל ב'. התוצאות שדווחו כוללות מודלים בעלי תקופת נתונים אחת ומודלים בעלי שתי תקופות נתונים. יש לציין כי על אף העובדה שבחנו מודלי חיזוי תוך שימוש בעד חמש תקופות נתונים (lags), מצאנו כי הוספת הנתונים מתקופה 3 ואילך גרעה למעשה מדיוק החיזוי של אותם מודלים.

יש לזכור כי שאלתנו הראשונה במחקר הייתה אם מודלי חיזוי, המבוססים על שילוב של נתונים ממנגנות חיפוש באינטרנט ונתונים מפורומים, עשויים להשיג תחזית מכירות מדויקת יותר ממודלים המבוססים רק על אחד ממקורות הנתונים. הממצא העיקרי הראשון שלנו היה כי מודל משולב עולה בתועלתו על מודלים המבוססים על נתונים מפורומים בלבד. באופן מפורש, מצאנו כי העשרת נתונים מפורומים בנתונים ממנגנות חיפוש משפרת באופן משמעותי את דיוק החיזוי. כאשר השתמשנו בתקופת נתונים אחת, המודל המשולב הניב שיפור של 0.55% ב-MAPE בהשוואה למודל המבוסס-פורומים; כאשר השתמשנו בשתי תקופות נתונים, המודל המשולב הניב שיפור של 0.58% ב-MAPE. בדומה לכך, כאשר השווינו את התוצאות של המודל המשולב למודל הנרחב והמשוכלל יותר, המבוסס-פורומים (ואשר משתמש בדירוג הרגשות של הפורום, נוסף על אזכורים בפורום), שמנו לב כי בשימוש בתקופת נתונים אחת הניב המודל

איור 2: תוצאות החיזוי (LR)



המשולב שיפור של 0.42% ב-MAPE; כאשר השתמשנו בשתי תקופות נתונים, הוא הניב שיפור של 0.37% ב-MAPE. ממצאים אלו משמעותיים והם מרמזים כי נתונים ממגמות חיפוש באינטרנט מכילים מידע נוסף, בעל ערך, שאינו זמין בנתונים מפורומים. קיבלנו ממצאים משמעותיים דומים גם כאשר השתמשנו באלגוריתם NN לא לינארי.

ציינו גם כי המודל המשולב שיפר באופן משמעותי את דיוק החיזוי בהשוואה למודל המבוסס באופן בלעדי על נתונים ממגמות חיפוש באינטרנט (ראו טבלה 2). התוצאה עקבית הן לאורך תקופת נתונים אחת והן לאורך שתי תקופות. עם תקופת נתונים אחת, המודל המשולב הניב שיפור של 0.34% ב-MAPE; עם שתי תקופות נתונים, המודל המשולב הניב שיפור של 0.4% ב-MAPE. יחד עם זאת, כאשר השתמשנו באלגוריתם NN לא לינארי, ביצועי המודל המשולב עלו רק עם שתי תקופות נתונים, אך שיפור זה לא היה משמעותי (ראו Geva et al., 2017).

הממצא העיקרי השני שלנו מתייחס לשאלת המחקר השנייה, אשר בחנה את הסוגיה אם מודלי חיזוי המשתמשים בנתונים ממגמות חיפוש באינטרנט משיגים דיוק דומה או אף טוב יותר בהשוואה למודלי חיזוי המשתמשים בנתונים מבוססי-פורומים (בין שמדובר בנפח אזכורים בפורום ובין שמדובר בנפח אזכורים בפורום בשילוב רגשות הפורום). מצאנו כי ביצועי המודלים המבוססים על מגמות חיפוש באינטרנט אינם נופלים, למצער, מביצועי המודלים המבוססים על נתונים מפורומים. באופן מפורש, עם תקופת נתונים אחת, המודל המבוסס על מגמות חיפוש באינטרנט השיג שיפור של 0.21% ב-MAPE בהשוואה למודל המבוסס-פורומים; עם שתי תקופות נתונים, המודל הניב שיפור קטן של 0.17% ב-MAPE. בהשוואה למודל הנרחב והמשוכלל יותר, המבוסס-פורומים (כלומר, המודל המכליל גם אזכורים בפורומים וגם דירוגי רגשות), המודל המבוסס על מגמות חיפוש באינטרנט השיג שיפור קטן של 0.08% ב-MAPE עם תקופת נתונים אחת, ואילו עם שתי תקופות נתונים הניב ירידה מזערית של -0.03% ב-MAPE. ההבחנה שביצועי החיזוי של מודלים המבוססים על נתונים ממגמות חיפוש באינטרנט אינם נופלים, לכל הפחות, מביצועי המודלים המבוססים על נתונים מפורומים, נשארת עקבית גם היא בבדיקת החוסן שערכנו באמצעות אלגוריתם NN לא לינארי. למעשה, כאשר השתמשנו באלגוריתם NN, מצאנו כי לא זו בלבד שהמודלים המבוססים על נתוני חיפוש משיגים תוצאות שאינן נופלות מאלו של מודלים המבוססים על נתונים מפורומים ועל נתונים מורחבים מפורומים, אלא אף עולים עליהם בביצועיהם (ראו פרק 1 Geva et al., 2017).

באופן מפורש, הטבלה מדווחת על ההפרש: diff = MAPE (מודל ב') - MAPE (מודל א') לפיכך, ערך חיובי, הקשור להשוואה בין מודל א' לבין מודל ב', מצביע על דיוק חיזוי טוב יותר במודל א' לעומת מודל ב'.

טבלה 2: הפרשים ב-MAPE ומרווחי ביטחון חד-צדדיים בהפרש שבין ערכי MAPE תוך שימוש בשיטת LR

מודל א'	מודל ב'	MAPE (מודל ב') - MAPE (מודל א') Lag=1	Lag =1,2
מודל מבוסס-פורום	מודל מדד ייחוס	0.11%**	0.13%**
מודל מורחב מבוסס-פורום	מודל מדד ייחוס	0.25%**	0.34%**
מודל מבוסס-מגמות חיפוש	מודל מדד ייחוס	0.32%***	0.31%**
מודל משולב	מודל מדד ייחוס	0.67%***	0.71%***
מודל מבוסס-מגמות חיפוש	מודל מבוסס-פורום	0.21%*	0.17%
מודל מבוסס-מגמות חיפוש	מודל מורחב מבוסס-פורום	0.08%	-0.03%
מודל משולב	מודל מבוסס-פורום	0.55%***	0.58%***
מודל משולב	מודל מורחב מבוסס-פורום	0.42%***	0.37%***
מודל משולב	מודל מבוסס-מגמות חיפוש	0.34%**	0.40%**

טבלה 2 מדווחת על ההפרש ב-MAPE בעזרת שני מודלים (מודל א' ומודל ב' - כאשר כל אחד מהם מתבסס על נתונים שונים), תוך התחשבות בתקופה 1 או ב-2 תקופות באלגוריתם ה-LR.

חושבו גבולות רווחי סמך תחתונים לגבי diff, תוך שימוש ב-2000 חזרות בשיטת חישוב רווחי סמך (BCA bootstrapping confidence interval calculation method) המיושמת בתוכנת R. גבול מרווח סמך נמוך של diff, עם ערך חיובי, מהווה ביטחון ש-MAPE (מודל א') אכן טוב יותר מאשר MAPE (מודל ב').

אנו מדווחים על רווחי הסמך:

* 0.9 גבול הביטחון הנמוך של diff חיובי
** 0.95 גבול הביטחון הנמוך של diff חיובי
*** 0.99 גבול הביטחון הנמוך של diff חיובי

מסקנות

בעבודה זאת חקרנו בצורה אמפירית את ההשפעה ההדדית של נתונים ממנמות חיפוש באינטרנט ושל אזכורים במדיה חברתית, בהקשר של חיזוי מכירות. חרף העובדה שנעשה בעבר שימוש רב בנתונים ממדיה חברתית לצורך חיזוי מכירות, ספרות קודמת הצביעה על המגבלות ועל ההטיות השונות הקשורות למקור זה של נתונים. אחת השיטות האפשריות לשיפור דיוק החיזוי של מקור נתונים זה היא שילוב שלו עם נתונים ממנמות חיפוש באינטרנט. יחד עם זאת, שני מקורות הנתונים נחקרו עד כה בזרמי ספרות שונים לחלוטין ולמטרות אחרות.

תוך שימוש בנתונים מתעשיית הרכב, סיפקנו עדות לכך שהעשרת מודלים מבוססי-פורומים בנתונים ממגמות חיפוש באינטרנט משפרת באופן מובהק את דיוק החיזוי. דבר זה מעיד כי מידע המבוסס על נתוני חיפוש באינטרנט הוא, למעשה, חיצוני ולא חופף למידע המבוסס על נתונים מפורומים או ממדיה חברתית. מנקודת מבט מעשית, ממצא זה מרמז כי חברות שהשקיעו כבר באיסוף נתונים מבוססי-פורומים למטרות מידול, יכולות לשפר את דיוק החיזוי באופן מובהק על ידי השקעה נוספת, קטנה יחסית, באיסוף נתונים ממגמות חיפוש באינטרנט. ממצאינו מעידים עוד כי דיוק החיזוי שניתן להשיג בעזרת נתונים ממגמות חיפוש באינטרנט בלבד אינו נופל מהדיוק הקשור לנתונים מבוססי-פורומים, הנמצאים בשימוש נפוץ יותר. ממצא זה יכול לעודד מנהלים לבצע חיזוי מכירות בעזרת נתונים ממגמות חיפוש באינטרנט, שיהיה חלופה בעלות נמוכה לנתונים ממדיה חברתית.

עבודתנו נושאת בחובה משמעויות ניהוליות ליצרני הרכב, ובהרחבה - גם ליצרני מוצרים אחרים בעלי מעורבות גבוהה (High Involvement Products - Blackwell et al., 2001). בנוסף, היתרון של שיטתנו נובע מהעובדה שהיא אינה דורשת נתונים קנייניים, הזמינים ליצרן בלבד. לפיכך היא ניתנת לשימוש על ידי גורמים במעלה או במורד התהליך (upstream or downstream players) וגם על ידי משקיעים בשוק המניות. יתר על כן, יצרני רכב יכולים להשתמש בגישה זאת כדי להעריך את המכירות הצפויות של מתחריהם. מודלי חיזוי מכירות מדויקים יותר יכולים, בתורם, להניע תהליך של קבלת החלטות טוב יותר בתחומים שונים, כגון הוצאות שיווק, ניתוח תחרותיות, ניהול מצאי ושיפור שרשרת האספקה. לגבי המקרה המיוחד של מכירות רכב, החלטות אלו כרוכות בהקצאת כספים גבוהה במיוחד; על כן, גם שיפורים קטנים בדיוק החיזוי צפוי שיהיו בעלי השפעה ניכרת.

אנו מצפים כי ניתן יהיה להכליל את הממצאים שלנו במערך נרחב של החלטות רכישה הנוגעות למוצרים בעלי מעורבות גבוהה, כגון רכישת בתים ותכנון נסיעות. במקרה של מוצרים בעלי מעורבות נמוכה כגון מוזיקה, אפליקציות וכרטיסים לסרטים, הצרכנים אינם עורכים חיפוש נרחב ופעיל וההחלטות מתקבלות ביתר קלות ראש. לפיכך, כוח החיזוי של מגמות חיפוש באינטרנט, בהקשר זה, איננו ברור. נושא זה מעלה כיוון מעניין למחקר עתידי.

ד"ר תומר גבע | tomergev@tauex.tau.ac.il

רשימת מקורות

Berger, J., & Milkman, K. L. (2012). "What Makes Online Content Viral?" Journal of Marketing Research (49:2), 192–205.

Berger, J., & Schwartz, E. M. (2013). "What Drives Immediate and Ongoing Word of Mouth?" Journal of Marketing Research (48:5), 869-880.

Blackwell D., Miniard P. W., & Engel, J. F. (2001). Consumer Behavior, 9th ed. Orlando, FL: Harcourt.

Canova, Fabio. (1998). "Detrending and business cycle facts." Journal of monetary economics (41:3), 475-512.

Chakravarty, A., Yong, L., & Mazumdar, T. (2010). "The Differential Effects of Online Word-of-Mouth and Critics' Reviews on Pre-Release Movie Evaluation," Journal of Interactive Marketing (24: 3), 185-197.

Chen, Y., & Xie, J. (2008). "Online Consumer Reviews: A New Element of Marketing Communications Mix," Management Science (54:3), 477–491.

Chevalier, J., & Mayzlin, D. (2006). "The Effect of Word of Mouth on Sales: Online Book Reviews," Journal of Marketing Research (43:3), 345-354.

Chintagunta, P. K., Gopinath, S., &Venkataraman, S. (2011). "The Effect of Online User Reviews on Movie Box Office Performance: Accounting for Sequential Rollout and Aggregation across Local Markets," Marketing Science (29:5), 944-957.

Choi, H., & Varian, H. (2009). "Predicting the Present with Google Trends," working paper.

Choi, H., & Varian, H. (2011). "Predicting the Present with Google Trends," working paper.

Dellarocas, C. (2006). "Strategic Manipulation of Internet Opinion Forums: Implications for Consumers and Firms," Management Science (52:10), 1577–1593.

Dellarocas, C., Awad, N., & Zhang, X. (2007). "Exploring the Value of Online Product Reviews in Forecasting Sales: The Case of Motion Pictures," Journal of Interactive Marketing (21:4), 23–45.

Dellarocas, C., & Narayan, R. (2006). "A Statistical Measure of a Population's Propensity to Engage in Post-Purchase Online Word-of-Mouth," Statistical Science (21:2), 277–285.

Dewan, S., & Ramaprasad, J. (2009). "Chicken and Egg? Interplay between Music Blog Buzz and Album Sales," in PACIS 2009 Proceedings.

Dewan, S., & Ramaprasad, J. (2012). "Music Blogging, Online Sampling, and the Long Tail," Information Systems Research (23:3), 1056–1067.

Dhar, V., & Chang, E. A. (2009). "Does Chatter Matter? The Impact of User-Generated Content on Music Sales," Journal of Interactive Marketing (23:4), 300-307.

Du, R. Y., & Kamakura, W. A. (2012). "Quantitative Trendspotting," Journal of Marketing Research (49:4), 514-536.

Duan, W., Gu, B., & Whinston, A. B. (2008)a. "The Dynamics of Online Word-of-Mouth and Product Sales - An Empirical Investigation of the Movie Industry," Journal of Retailing (84:2), 233–242.

Duan, W., Gu, B., & Whinston, A. B. (2008) b. “Online Reviews Matter? An Empirical Investigation of Panel Data,” Decision Support Systems (45:4), 1007–1016.

Efron, Bradley, & Robert J. Tibshirani. (1994). An introduction to the bootstrap. CRC press.

Forman, C., Ghose, A., & Wiesenfeld, B. (2008). “Examining the Relationship Between Reviews and Sales: The Role of Reviewer Identity Disclosure in Electronic Markets,” Information Systems Research (19:3), 291–313.

Geva, T., Oestreicher-Singer, G., Efron, N., & Shimshoni, Y. (2017). “Using Forums and Search Data for Sales Prediction of High-involvement Products.” MIS Quart. (41:1), 65-82.

Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S., & Brilliant, L. (2008). “Detecting Influenza Epidemics Using Search Engine Query Data,” Nature (457:7232), 1012-1014.

Godes, D., & Mayzlin, D. (2004). “Using Online Conversations to Study Word-of-Mouth Communication,” Marketing Science (23:4), 545-560.

Godes, D., & Silva, J. C. (2012). “Sequential and Temporal Dynamics of Online Opinion,” Marketing Science (31:3), 448-473.

Goel, S., Hofman, J. M., Lahaie, S., Pennock, D. M., & Watts, D. J. (2010). “Predicting Consumer Behavior with Web Search,” Proceedings of the National Academy of Sciences (107:41), 17486–17490.

Gu, B., Park, J., & Konana, P. C. (2012). “The Impact of External Word-Of-Mouth Sources on Retailer Sales for High Involvement Products,” Information Systems Research (23:1), 182-196.

Hong, Y., Chen, P.-Y., & Hitt, L. M. (2014). “Measuring Product Type with Dynamics of Online Product Review Variances: A Theoretical Model and the Empirical Applications,” working paper.

Hu, N., Liu, L., & Zhang, J. (2008). “Do Online Reviews Affect Product Sales? The Role of Reviewer Characteristics and Temporal Effects,” Information Technology Management (9:3), 201–214.

Hu, Y, Du, R. Y., & Damangir, S. (2014). “Decomposing the Impact of Advertising: Augmenting Sales with Online Search Data,” Journal of Marketing Research (51: 3), 300-319.

Hymans, S. H., Ackley, G., & Juster, F. T. (1970). “Consumer Durable Spending: Explanation and Prediction,” Brookings Papers on Economic Activity, 173-206.

Kronrod, A., & Danziger, S. (2013). “‘Wii Will Rock You!’ The Use and Effect of Figurative Language in Consumer Reviews of Hedonic and Utilitarian Consumption,” Journal of Consumer Research (40), 726-739.

Lazer , D., Kennedy, R., King, G., & Vespignani, A. (2014). “The Parable of Google Flu: Traps in Big Data Analysis,” Science (343:6176), 1203-1205.

Li, X., & Hitt, L.M. (2008). ”Self Selection and Information Role of Online Product Reviews,” Information Systems Research (19:4), 456-474.

Liu, Y. (2006). “Word of Mouth for Movies: Its Dynamics and Impact on Box Office Revenue,” Journal of Marketing (70:3), 74–89.

Lovett, M., Peres, R., & Shachar, R. (2014). “On Brands and Word-of-Mouth,” Journal of Marketing Research, forthcoming.

Luca, M., & Zervas, G. (2015). “Fake It Till You Make It: Reputation, Competition, and Yelp Review Fraud,” Harvard Business School NOM Unit Working Paper No. 14-006. Available at SSRN:http://ssrn.com/abstract=2293164.

Luo, X., Zhang, J., & Duan, W. (2013). “Social Media and Firm Equity Value,” Information Systems Research (24:1), pp.146-163.

Luo, X., Zhang, J., Gu, B., & Phang, C. W. (2014). “The Influence of Online Word-of-Mouth on Long Tail Formation,” Decision Support Systems, forthcoming.

Mayzlin, D., Dover, Y., & Chevalier, J. (2014). “Promotional Reviews: An Empirical Investigation of Online Review Manipulation,” American Economic Review (104:8), 2421-55.

Moe, W. W., & Schweidel, D. A. (2012). “Online Product Opinions: Incidence, Evaluation and Evolution,” Marketing Science (31:3), 372-386.

Moe, W. W., & Trusov, M. (2011). “The Value of Social Dynamics in Online Product Ratings Forums,” Journal of Marketing Research (48:3), 444-456.

Rui, H., Liu, T., & Whinston, A. (2012). “Whose and What Chatter Matters? The Impact of Tweets on Movie Sales,” working paper.

Schlosser A.E., (2005). “Posting versus Lurking: Communicating in a Multiple Audience Context,” Journal of Consumer Research 32, 260-265.

Schulze, C., Schöler L., & Skiera, B. (2014). “Not All Fun and Games: Viral Marketing for Utilitarian Products,” Journal of Marketing (78:1), 1-19.

Schweidel, D. A., & Moe, W. W. (2014). “Listening In on Social Media: A Joint Model of Sentiment and Venue Format Choice,” Journal of Marketing Research (51:4), 387-402.

Seebach, C., Pahlke, I., & Beck, R. (2011). “Tracking the Digital Footprints of Customers: How Firms Can Improve Their Sensing Abilities to Achieve Business Agility,” in ECIS 2011 Proceedings.

Vosen, S., & Schmidt, T. (2011). “Forecasting Private Consumption: Survey-Based Indicators vs. Google Trends,” Journal of Forecasting (30:6), 565-578.

Wu, L., & Brynjolfsson, E. (2009). “The Future of Prediction: How Google Searches Foreshadow Housing Prices and Quantities,” in Proceedings of the 2009 International Conference on Information Systems.

Zhu, F., & Zhang, X. M. (2010). “Impact of Online Consumer Reviews on Sales: The Moderating Role of Product and Consumer Characteristics,” Journal of Marketing (74:2), 113-148.