שימוש במאגרים ממוכנים ככלי מחקרי משלים

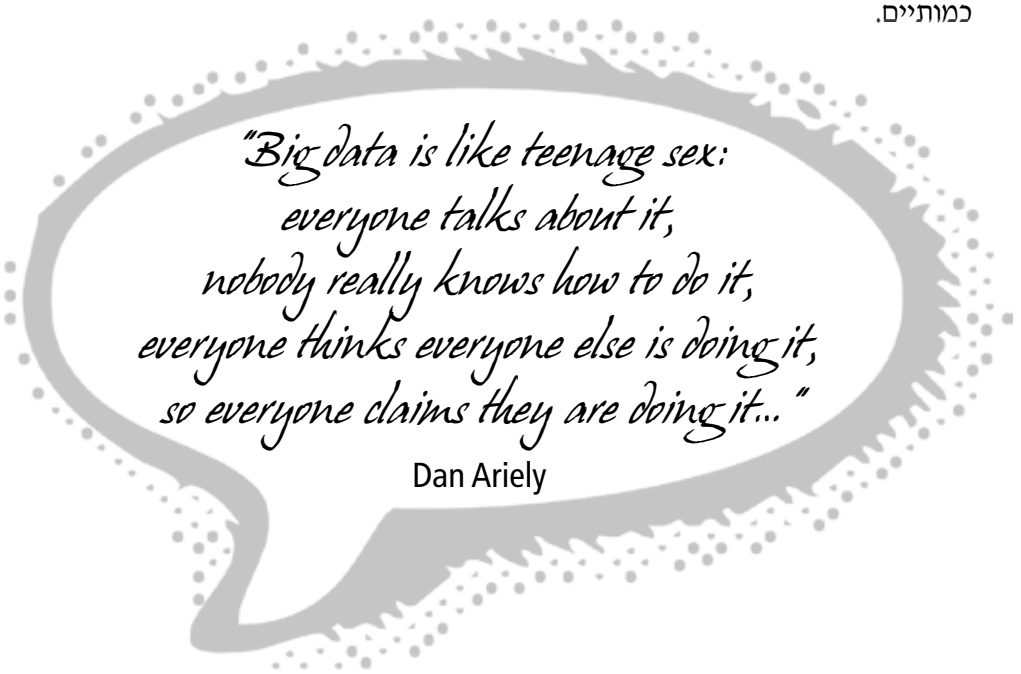
הדוגמה של מחקר לצרכים אופרטיביים

רס"ן א' - משרת באמ"ן וד"ר קרן ששון

מבוא

ברוב המחקרים האופרטיביים "המסורתיים" נשענים החוקרים בדרך כלל על מתודולוגיה של בחינת התפתחויות חדשות אל מול המידע הקיים על אודות מושא המחקר, או של ניתוח השוואתי בין אירועים עכשוויים לבין אירועי עבר בעלי דמיון תוכני. ההנחה העומדת בבסיס מתודולוגיה זו היא כי ביכולתה לאפשר הסקת מסקנות שישפכו אור על "מעצבים", "משתנים", "מגמות" ו"תפניות" בשדה הקרב העתידי, ובכך לקדם את ניתוח אתגרי העתיד ברזולוציה מספקת לניתוחי מודיעין בתחומי בניין והפעלת הכוח.

על אף השימוש הנרחב במתודולוגיה זו, היא אינה חפה ממגבלות. ראשית, בהיעדר קורלציה הכרחית בין מקרים מן העבר לבין המציאות העכשווית (או זו המתפתחת), הקושי בזיהוי כיווני התפתחות עתידיים נותר בעינו; נוסף על כך המחקר האופרטיבי מוסיף להסתמך על מאגר נתונים צר יחסית המקשה על החוקרים לעמוד בדרישת סף גבוהות של איכות, עדכניות, אינטימיות ופירוט. המחקר המסורתי הנשען על שיטות אלו אף מותיר מקום רב (ולעיתים רב מדי) לפרשנותו של החוקר על אודות החומרים שנאספו ועלול לשמר הטיות וסובייקטיביות. כך, ברוב המחקרים האופרטיביים (המסורתיים) נותר קושי לתקף הערכות ומסקנות איכותניות באמצעות סימוכין כמותיים.



בשנים האחרונות, לאור העלייה המשמעותית בכמות המידע הזמין והנגיש ברשת ובאיכותו, החלו מתודולוגיות ושיטות מחקר חדשות למיצוי מידע לחדור גם למחקרי צבא וביטחון, תוך שימוש הולך וגובר במאגרי מידע ממוכנים. לאחרונה, החלו מחקרים מסוג זה להתבצע גם באגף המודיעין בצה"ל, בעיקר במסגרת מחקרי עומק אסטרטגיים, מחקרי חברה ותקשורת.

בניגוד לרמת הניתוח האסטרטגית, הרמה האופרטיבית, נותרה יחסית מאחור בתחום זה, זאת בעיקר משום שעד כה הכלים שהיו קיימים ב"שוק המידע הגלוי" נתפסו כלא רלוונטיים לצרכים האופרטיביים. בעת האחרונה הולכת ומתחדדת ההכרה בפוטנציאל הגלום בשימוש בכלים אלו גם עבור המחקר האופרטיבי-טקטי. לראיה, במהלך השנה האחרונה בוצעו בידי מכוני מחקר המסונפים לצבא האמריקאי בכלל, ולגופי המודיעין בפרט, שורה של מחקרי צבא יישומיים העוסקים בסביבה האופרטיבית דרך שימוש במתודולוגיות מיצוי מידע באמצעות מאגרי מידע ממוכנים.[1]

מתודולוגיה

מאמר זה בוחן את תרומתם המחקרית של מאגרי מידע ממוכנים למחקרים ביטחוניים-צבאיים תוך התמקדות במקרה בוחן אופרטיבי העוסק בצורות הקרב במזרח התיכון בשנים האחרונות. לטובת מטרה זו, נבחרו שני מאגרי מידע מרכזיים:

- GDELT (The Global Database of Events' Language and Tone), מאגר שפותח על ידי אוניברסיטת ג'ורג'טאון, שסורק, מקודד ומנגיש לשימושים ציבוריים ומחקריים 98% מכל התקשורת העולמית.
- ICEWS (The Integrated Crisis Early Warning System), כלי לחיזוי המשברים שפותח על ידי המערכת הביטחונית האמריקאית שאוסף ומקודד נתוני תקשורת גלויה כמצע מאפשר לחיזוי והתרעה על משברים.[2]

בלב הבחינה שלפנינו עומדת יכולת החוקר לחלץ מסקנות יישומיות ממאגרי מידע ממוכנים.[3] חשוב לציין כי מראש בחרנו להתמקד בשאלות אופרטיביות מוכרות, מבלי להעמיק בבחינת יכולת כלים אלו לטייב אקספלורציה של שאלות חדשות או שונות.[4] משכך, נייר זה מייצג מחקר גישוש, הבוחן את תרומתם האפשרית של כלי מחקר מעולם הביג דאטה למיצוי תובנות אופרטיביות

---

1 "הסביבה האופרטיבית" (operational environment) - מונח "עדכני" שתפקידו להעניק פרשנות הוליסטית יותר לתיאור המסורתי של "שדה הקרב". בהגות הצבאית המערבית, נוהגים להגדיר מונח זה בתור שילוב בין התנאים, הנסיבות וההשפעות שמתלכדים יחד כדי לקבוע את אופן השימוש בכוחות צבאיים. ה"מעשה" האופרטיבי נתפס, בגישה זו, כמערכת המעוצבת על ידי הזיקות ויחסי הגומלין בין מספר רב של משתנים ותת־משתנים: משתנים צבאיים (תיאור אופי האויב והיריב, מאפייני העימותים והמערכות הצבאיות), משתנים אזרחיים, משתנים סביבתיים וכן על ידי מגוון משתנים טכנולוגיים שבבוחם להשפיע על התרחשויות בשדה הקרב הקינטי, הרשתי (סב"ר) או התודעתי. ההגדרות על פי US Dictionary of Military and Associated Terms מהדורה 2005.

2 יש לציין כי כיום, פונקציית הניבוי אינה נגישה לקהל הרחב (גם לא לזה האקדמי), אלא מיושמת בעיקר במרחב מערכות הביטחון בארצות הברית.

3 יצוין כי טרם עידן "נתוני העתק", דיווחים תקשורתיים גלויים נתפסו בקרב חוקרי ביטחון כלא משמעותיים או רלוונטיים למחקרים מבצעיים מסווגים בליבת העיסוק. עם חדירת כלי מחקר אלו לתחום הביטחון, ברובדי מחקר שונים, התחדד הצורך לבחון את מידת ההתאמה של כלים אלו עבור המחקר הצבאי-אופרטיבי.

4 מדובר בבדיקה מוגבלת, יחסית, המאפשרת באמור התייחסות ובחינה רק של השאלות המסורתיות במחקרים אופרטיביים מבלי לעמוד על יכולת גילוי ה־Known-Unknown.

בנושאי השתנות האויב, האיום ושדה הקרב במזרח התיכון מאז פרוץ הטלטלה האזורית.
המדדים המרכזיים עליהם נשענו שליפת נתוני המאגרים וניתוחם הם:

- גיאוגרפיה - מהלכי הלחימה שנשלפו ונבחנו הוגבלו לשלוש מדינות "מטולטלות" במזרח התיכון: תימן, סוריה ועיראק.
- זמן - ניתוח הנתונים שנאספו במאגרים הוגבל לשנים 2012–2016 (שנת 2011 לא נכללה במסגרת הניתוח כיוון שבשלביה הראשונים של הטלטלה האזורית, עיקר אירועי הלחימה היו "דמויי ביטחון פנים" ולא נשאו אופי של לחימה).
- היקף - הנתונים שנותחו כוללים כ־1.8 מיליון "אירועי קרבות" במאגר GDELT, וכ־39,110 "אירועי קרבות" במאגר ה־ICEWS עדות ראשונית וקריטית להבדלים בין שיטות העבודה של המאגרים ומידת התאמתם למחקרים אופרטיביים.

הבחינה מתמקדת בניסיון לענות על שתי שאלות מרכזיות:

האם מאגרי מידע ממוכנים המאפשרים ניטור, קצירה וקידוד מידע רלוונטיים למחקר האופרטיבי? לטובת מענה על שאלה זו, נאפיין וננתח בצורה השוואתית את מנגנוני קידוד הדיווחים התקשורתיים בשני המאגרים, GDELT ו־ICEWS. ניתוח מסוג זה נועד לתקף את הממצאים בעיקר ברובד האונטולוגי.

האם דיווחי התקשורת הגלויה מכילים מידע עשיר, מובנה ומדויק המאפשר מיצוי תובנות והסקת מסקנות אופרטיביות דרך "מניפולציה" כמותנית של נתונים? לטובת מענה על שאלה זו, נשווה בין הממצאים שהתקבלו במסגרת ניתוח נתוני המאגרים, לבין תובנות מודיעיניות שהתגבשו במסגרת מעקב ומחקר תשתיתי המבוצע בכלי ניתוח "מסורתיים". ניתוח מסוג זה נועד לתקף את תרומתה של המתודולוגיה הנבחנת מול שאלות התוכן המעסיקות את החוקר האופרטיבי.

מהם מאגרים ממוכנים של תקשורת גלויה וכיצד הם פועלים?

ככלל, בשדה המחקר העכשווי קיימת שורה של מאגרי מידע ממוכנים שעוסקים באגירת דיווחי תקשורת ועיבודם, החל מאלו המוקמים על ידי תאגידי Mass Media ועד למאגרי דאטה כגון Lexis Nexis. בנייר זה בחרנו להתמקד בשני מאגרים המרכזיים ביותר בשדה המחקר, הבולטים בעיקר הודות לחוזק המנגנונים שלהם בהקשרים של "הבנת הנקרא" וקידוד אוטומטי של אירועים המדווחים באמצעי התקשורת השונים.

- GDELT (The Global Database of Events' Language and Tone) - מאגר שצמח מתוך פרויקט שאפתני של אוניברסיטת ג'ורג'טאון, בהובלת Kalev Leetaru (מאנשי Yahoo!) ו־Phillip Schrodt (פרופסור למדעי במדינה העוסק בשיטות אוטומטיות לקידוד אירועים מאוניברסיטת ג'ורג'טאון שבארצות הברית). הפרויקט החל את דרכו בשנת 2011, ובשנותיו הראשונות היווה פלטפורמה אנליטית לחוקרים מתחום התקשורת והיחסים הבין־לאומיים. בהדרגה, היקף השימוש המחקרי בנתונים הנאגרים במאגר גדל וחדר לדיסציפלינות שונות ומגוונות. שלוש שנים אחרי השקת הפרויקט, הפכו נתוני המאגר לזמינים עבור קהלי יעד מחקריים שונים דרך ממשקי רשת ידידותיים למשתמש (Google Big Query), המאפשרים שימוש רחב ואינטר־דיסציפלינרי בחומרים הנאספים והמקודדים במאגר.

- ICEWS (The Integrated Crisis Early Warning System) – המאגר פותח על ידי תאגיד Lockheed Martin במימון Darpa (הסוכנות האמריקאית לעידוד חדשנות ומחקר בתחום הביטחוני, אשר היוותה מודל להקמת מפא"ת בישראל) וחיל הים האמריקאי. מדובר במאגר ששלבי הפיתוח שלו (איסוף, קידוד וניתוח דיווחים תקשורתיים מערוצי תקשורת מוגדרים ומוגבלים) נמשכו כעשור ולוו בקפידה על ידי שורה של מומחים ואנשי אקדמיה מובילים מתחומי מחקר וידע שונים (מדעי המדינה, בלשנות חישובית Data Science).[5] נגישות וזמינות נתוני המאגר היו מוגבלים במשך שנים לקהילת הביטחון האמריקאית, ורק בשנים האחרונות התאפשרה הנגשה חלקית של הנתונים לקבוצת חוקרים מאוניברסיטת הרווארד לטובת מבחני תוקף שונים ועריכת מחקרי התנסות.

המאגרים נבדלים זה מזה, בין היתר, במטרות לשמן הוקמו ופותחו:

פרויקט ה־GDELT נועד ליצור: "A big data history of life, the universe and everything". משכך, המאגר נועד: "להקיף את כלל החברה האנושית על ידי בניית קטלוג חובק עולם של התנהגויות ואמונות אנושיות, כזה שיחבר בין כל אדם, מקום, ארגון, מקום, ספירה, נושא, מקור חדשותי ואירוע על פני כדור הארץ לרשת מסיבית אחת הלוכדת את המתרחש בעולם. כל זאת תוך ציון ההקשר שלו, מי מעורב, ואיך העולם מתייחס לאירוע, מדי יום ובכל יום".[6]

מאגר ה־ICEWS הוקם בראייה ביטחונית למטרות ניבוי התרחשויות בעיקר לצורכי בניין הכוח. לפיכך מלבד העיסוק ב"צילום" תמונת המצב התקשורתית, מפתחי ה־ICEWS עיגנו את הפרויקט בפיתוח יכולות חיזוי התפתחויות על בסיס הנתונים שנאספו וקודדו במאגר. מטרה זו נתפסה בת השגה באמצעות פיתוח ויישום אלגוריתמיקה מורכבת ולטענת המפתחים, כיום, השימוש בה מאפשר לחוקר להגיע לדיוק של 80% בחיזוי והתרעה על משברים.[7]

שיטת פעולתם של מאגרים ממוכנים לאיסוף מידע גלוי כוללת איתור וסריקת כתבות ודיווחים ממקורות תקשורת רבים ומגוונים ברחבי העולם, אינדוקס, עיבוד וקידוד התוכן (בהתאם לספר קידוד מובנה), ולבסוף, ניתוח והנגשת הנתונים לצרכים מחקריים. ניתן לאפיין את שיטת העבודה של מאגרים ממוכנים בכלל, ושל אלו הנבחנים בנייר זה בפרט, ככזו הנחלקת לארבעה שלבים מרכזיים:

איור 28: שיטת העבודה למיצוי מאגרי נתונים לטובת מחקר מודיעיני

5 שילוב מומחים מתחומי המחקר בבניית מאגר הידע הוא מכפיל כוח בבניית כל מאגר כזה ובהתאמתו לצורכי המחקר האמור להתבסס עליו.

6 https://amanwiki.services.idf/wiki/GDELT

7 https://dataverse.harvard.edu>icews. יש לציין כי כיום, פונקציית הניבוי אינה נגישה לקהל הרחב (גם לא לזה האקדמי), אלא מיושמת בעיקר במרחב מערכות הביטחון בארצות הברית.

בחינת התאמת שיטת העבודה של המאגרים לצורכי המחקר האופרטיבי

"שלב קצירת הנתונים" מתמקד באיתור מקורות תקשורתיים לטובת איסוף ואגירת המידע הגולמי ובמעקב אחריהם. זהו שלב קריטי שכן עליו מבוססת למעשה יכולת המאגר לצבור ולספק נתונים.

- שלב זה נשען בעיקר על יכולת האלגוריתם המובנה במאגר לזהות ולעקוב בצורה שוטפת אחר המקורות שנבחרו לשליפת הנתונים, לרבות: להוסיף/לגרוע מקורות, להתמודד עם מקורות שאינם כלי תקשורת (כגון רשתות חברתיות), או שאינם טקסטואליים (כגון וידיאו), לשמר נגישות למקורות שאינם חינמיים או סגורים לשימוש הפרטי וכן לשמר את יכולת הפנייה לאחור לטובת אחזור נתונים ממקורות היסטוריים.
- המאגרים הנבחנים בנייר זה נבדלים בגישות המובילות ומכתיבות את מערך איתור ואיסוף מקורות דיווחי התקשורת על בסיסם מתבצעים הקידודים:
  - גישתו של GDELT גורסת כי רצוי לעקוב אחרי ולנטר מספר מקורות תקשורתיים גדול ככל האפשר, ללא קשר למרכזיות המקור בתקשורת המקומית או העולמית. לפיכך המאגר אוסף מידע על אירועים שדווחו בתקשורת העולמית מאז 1979,[8] ומעניק משקל זהה לכלי תקשורת מקומיים וקטנים ולאלו הגדולים והמרכזיים. כך נאספים דיווחי תקשורת (כתובה, משודרת ומקוונת) שפורסמו ב־65 שפות, מכל רחבי העולם. כתבות ו"פוסטים" שפורסמו ברשתות החברתיות הפתוחות "נקצרים" גם הם על ידי המאגר (בשלב זה באופן חלקי בלבד). אחד היתרונות המובהקים של המאגר, לטענת משיקי המיזם, טמון ביכולתו לסרוק עד כ־98.4% מהתוכן התקשורתי העולמי, כאשר תדירות העדכון האוטומטי עומדת על 15 דקות.[9]

8 בימים אלו מתבצעת סריקה לאחור של כל המידע התקשורתי החל מ־1980.

9 המאגר גם בולל קידוד של 0.5 מיליארד שעות וידיאו (במקור באנגלית).

- הגישה של ה־ICEWS מבקשת להגביל את היקף המידע "הנקצר" מערוצי התקשורת ולהתמקד במקורות המוכרים כאמינים, איכותיים ו"מתוקפים", לעיתים גם על חשבון לקיחת הסיכון של "פספוס" מידע שלא זכה להד תקשורתי רחב. לפיכך המאגר מתמקד במעקב אחר מספר "מוגבל" של כ־6,000 כלי תקשורת מרכזיים בעולם לצד התכתבויות ברשתות חברתיות מאז שנת 1995. קבצי הנתונים של המאגר מתפרסמים מעת לעת, בדרך כלל אחת לשנה ואינם כוללים עדכון שוטף או אוטומטי כדוגמת GDELT (בעת כתיבת מאמר זה זמינים לצפייה והורדה נתונים עד 2016). נוסף על כך המאגר אינו מכסה דיווחים בפרסית, ערבית וטורקית,[10] ובכך מגביל באופן ניכר את יכולותיו לשקף באופן מכיל ומובנה את ההתרחשויות במזרח התיכון.

שלב "הבנת הנקרא" - שלב זה מתמקד בבחינת הנתונים שנאספו וקידודם לאירועים מוגדרים. לטובת אימות המידע ומניעת כפילויות קטגוריאליות, האלגוריתמים המנחים את עיבוד המידע במאגרים מתוקפים על ידי בדיקות ומבחנים מורכבים (שתוצאותיהם עומדות בכללי התקן המחקריים). בהקשר זה יש לציין כי האתגרים המתודולוגיים המשמעותיים ביותר עבור המאגר בשלב זה הם זיהוי אירועים, קרי "קריאת" והבנת מהות או אופי האירוע כדי לתייגו באופן מדויק; מניעת זיהויים כפולים, קרי הבחנה בין אירוע לבין דיווח על אודותיו (כך למשל אם קרב מסוים דווח במקביל על ידי כמה כלי תקשורת, האלגוריתמיקה של המאגר אמורה לדעת לתייג ולספור את כלל הדיווחים העוסקים/דנים בו תחת קרב אחד). יצירת הקשר לאירוע פרטני בתוך סך כל האירועים - יתרון חשוב שיש למאגרים בהקשר זה טמון ביכולתם לספק לחוקר תמונה מובנית, בהקשרים משותפים לאירוע הייחודי שנבחן (עוצמה, סנטימנט, מרכזיות האירוע).

המאגרים עושים שימוש בטכנולוגיה **שונה** לעיבוד וקידוד המידע שנאסף מהתקשורת. שונות זו גוזרת, באופן טבעי, הבדלים מובהקים בדיוק הממצאים של המאגרים. עם זאת, במקרים רבים, ניתן לזהות דמיון במגמות המוצגות על בסיס נתוני שני המאגרים:

10 מבדיקה שנערכה מול חוקרי הרווארד הנגישים לפיתוח המאגר ונתוניו עולה כי התקשורת הערבית תחל להירכש בקרוב, ובהמשך תיכנס גם השפה הערבית לשלבי פיתוח.

- האלגוריתם האוטומטי של GDELT נשען על כלי תרגום ועיבוד מסחריים בסיסיים. בדיקות תוקף ומהימנות חיצוניות מעידות כי הכלים בהם משתמש המאגר לטובת זיהוי האירועים, ושלילת דיווחים כפולים, אינם רגישים או מדויקים מספיק. לראיה, בשאילתת היקף אירועי הלחימה במזרח התיכון בשנים המוגדרות, זיהה מאגר GDELT 1.8 מיליון אירועים, בעוד מאגר ה־ICEWS זיהה 40,000 אירועים בלבד. ממצא זה מהווה, אם כן, אינדיקציה לנקודת חולשה משמעותית במנגנוני הזיהוי של GDELT. חולשה זו מהווה גם בסיס מרכזי לקבלתו המוגבלת של המאגר במרחב האקדמי.
- האלגוריתם האוטומטי של מאגר ה־ICEWS עושה שימוש במנוע BNN ACCENT, המיועד לניתוח שפות טבעיות של תאגיד טכנולוגית עילית צבאית (RAYTHEON BNN). הכלי מוערך בקרב חוקרים המשתמשים במאגר כמוקפד יותר (גם בנושאים אופרטיביים), בעיקר בהקשר של יכולות הזיהוי שלו. כך, לטענת מפתחי המאגר, בדיקות התוקף והמהימנות שנערכו לנתונים שזוהו דרך מנוע זה (ביחס לאירועים מסוג "לחימה"), מעידות על יכולת דיוק של 74%.[11] שליפת אירועי הלחימה במזרח התיכון בשנים המוגדרות העלתה מספר מצומצם של 39,110 אירועי לחימה. שונות מספרית זו מעידה על נקודת החוזק הלוגית במנגנון הזיהוי של המאגר, והתאמתו הטובה יותר לצי"חים אופרטיביים.

שלב "קידוד ואגירה" - שלב זה מתמקד בקידוד הנתונים ובשמירתם במאגר. בשל נפחי המידע הגדולים, אין אפשרות מעשית לשמר את הנתונים בתור קובצי טקסט מקוריים. על כן, בשני המאגרים, המידע, לאחר עיבודו, נשמר במאגר באופן מקודד.[12] בפועל, קידוד הנתונים הנאספים מהאתרים השונים נשען על ספר הקידוד Cameo (Conflict and Mediation Event Observation), המכיל קטגוריות קידוד מובחנות ומוסכמות, המקובלות זה עשורים אחדים במחקרים במדעי המדינה, והמתוקפות במאות מחקרים אקדמיים. שימוש בספר קידוד זה מאפשר הבחנה בין מאות נושאים שונים, אלפי תכונות וכמה אלפים של תתי־נושאים בדיווחי התקשורת הרבים הנאספים במאגר.[13]

בהקשרי המחקר האופרטיבי, על אף ההבחנה המגוונת של ספר הקידוד, חוקרים העוסקים במחקר אופרטיבי יכולים להישען בעיקר על הנתונים שקודדו במספר מוגבל מאוד של מילות מפתח, שאינן מביאות לידי ביטוי את כל "העושר" של התופעות האופרטיביות. הלכה למעשה, בשאלות האופרטיביות ניתן להתבסס על שני קידודי Cameo רלוונטיים בלבד - תקיפה או לחימה. נוסף על כך מאפשר ה־GDELT לאגור נתונים "חופשיים", גם כאלו שאינם מוכרים ומקודדים על ידי Cameo , אך כאלו שזוהו בשלב עיבוד הנתונים לפי המדדים של מקום, שם, סוג ושחקן. ובכלל זה:

- המאגר מאתר, מחשב ומשמר נתוני סנטימנט, AvgTone, (בסקלה של -/+ 10).
- המאגר "מצמיד" לכל שורת תוכן את מדד ה־"Importance", המחושב על פי כמות הקשב

---

11 בכלל, המאגר גם מוביח יכולת חילוץ ואגרגציה של אירועים העומדת בין 60% ל־80% בחלוקה לסוגי אירועים שונים.

12 "יחידת הקידוד" הבסיסית" במאגר היא "אירוע" המקבל קוד המורכב מכמה שדות: תאריך, כלי תקשורת (שבו נמצא דיווח על אודותיו), שפה, נושא, תתי־נושאים, שחקנים מעורבים, מדינה, סנטימנט, נקודות ציון ועוד.

13 לשם המחשה, ספר הקידוד מבחין בין 1500 קבוצות דתיות ו־650 קבוצות אתניות.

שהופנתה לאירוע מסוים על ידי התקשורת העולמית.

- בהקשר הייחודי של מחקרים אופרטיביים, מאגר זה רלוונטי במיוחד כי הוא מחשב ומשמר גם נתוני מדד Goldstein לכל אירוע מקודד. קרי הערכת אינטנסיביות עימות/שיתוף פעולה (גם כאן בסקלה של -/+ 10).[14]
- ICEWS – מכיל גם הוא כמה פונקציות ייחודיות, נוסף על נתוני .Cameo חרף העובדה כי הנתונים הפתוחים לשימוש חופשי של המאגר דלילים יותר, יש להדגיש כי המאגר מנתח נתונים המתייחסים לכל הדיווח אותו הוא משייך לאירוע מסוים. משכך, תועלת מהשימוש במדדים אלו תלויה בדיוק מנגנון הזיהוי של המאגר, שכאמור מסתמן כטוב מזה של GDELT.

בפועל האלגוריתמיקה במאגר מאפשרת פונקציות נוספות:

בנייה וקידוד של "מילונים" (Bag of words) - בסיסי נתונים הנבנים באופן ממוחשב תחת קטגוריות אישים (Actors), שחקנים (Agents) וארגונים או מגזרים (Sectors). לשקיפות ותקפות המילונים חשיבות רבה הן כחומר הגלם לניתוחים המחקריים והן כמצע לתיקוף ודיוק מנועי "הבנת הנקרא" של המאגר. אינטנסיביות האירועים נמדדת ונשמרת במאגר על פי מדד Cameo (בסקלה של -/+ 10).[15]

שלב "ההנגשה" - שלב זה עוסק בהנגשת המידע שנאסף, קודד ונשמר במאגרים, העמדת כלים לניתוח הנתונים והצגתם הגרפית באמצעות פונקציות אנליטיות מובנות. ככלל, ניתן לדבר על מספר רמות נגישות אל החומרים הנאספים ומקודדים במאגר:

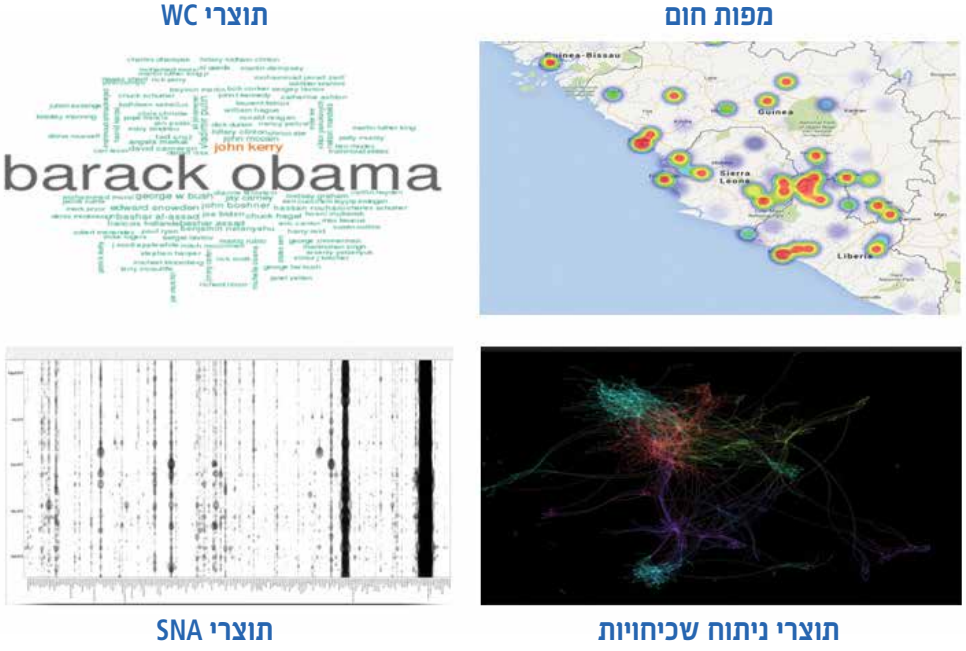
- שליפה - האפשרות הבסיסית הפשוטה יחסית של חילוץ הנתונים וניתוחם על בסיס תוכנות ניתוח סטטיסטי מסוגים שונים (בעיקר SQL).
- ניתוח בכלים מובנים - המאגרים מאפשרים ניתוח אוטומטי של הנתונים על ידי שירותים אנליטיים מובנים. להבנתנו, קיימת עדיפות לעיבוד וניתוח נתונים על בסיס כלים ייעודיים החיצוניים למאגר לטובת מיצוי התובנות ומניפולציות מורכבות ומגוונות יותר על הדאטה.
- ויזואליזציה - מדובר ביכולת הקיימת במאגרים להצגה גרפית (מתקדמת יחסית) של הנתונים.
- נגישות לנתונים הגולמיים של המאגרים - מדובר בנגישות לנתונים שנאספו על ידי המאגר ונשלפו על פי קידוד, אך בתצורותיהם "הגולמית" (חזרה למידע עצמו לפני שקודד), שכן בשני המקרים כוללות שורות הקידוד את נתוני ה־URL של הכתבות (בחלק מהמקרים המאגר אף משמר את הטקסט עצמו).
- המידע שבמאגר GDELT זמין לשימוש מחקרי חופשי תוך עדכון שוטף עיתי. אם אין בידי החוקר ידע מקדים בשפת SQL, ניתן להיעזר בשירותים האנליטיים של המאגר עצמו לטובת שליפה "מובנית" (כמות הנתונים הנשלפים מוגבלת ל־10,000 שורות תוכן בלבד), עיבוד בסיסי לנתונים וויזואליזציה על פי תבניות מובנות.

---

14 מדובר במדד שהתקבל ויושם במחקרי מדע המדינה כבר בשנות ה־70, וביום מוסיפים חוקרים במרחב האקדמי לעשות בו שימוש נרחב. טווח הסקלה נע בין 10- ל+10, כאשר מספרים שליליים מייצגים אירועים עוינים ומספרים חיוביים מייצגים אירועי שיתוף פעולה. ערכים אלו נמצאים בשדה העצימות של הדאטה.

15 מדד מוכר ומקובל הנהוג במחקרים כמותניים במדע המדינה שפותח כשתחליף למדד Goldstein שבו נעשה שימוש במאגר GDELT.

איור 29: דוגמאות לתוצרים על בסיס מאגרי המידע



מאגר ICEWS מיועד, כאמור, לשימושי מערכת הביטחונית האמריקאית ועל כן, מרבית הנתונים והפונקציות האנליטיות שלו אינם מונגשים וזמינים במלואם לשימוש רחב. זאת, מלבד הנגשה מוגבלת למיצוי המידע הגולמי שהמאגר קוצר ומנגיש דרך Harvard Dataverse. יודגש כי הכמות המוגבלת של הנתונים הזמינים לשימוש "פתוח" מקשים על שימושים יצירתיים ומיצוי התובנות האפשריות מהמאגר. מידע וכלי מחקר המונגשים לציבור כוללים:

- קובצי דאטה מובנים, מצונזרים (לא כוללים מידע בנושאי פנים־ארצות הברית) בפילוח לפי שנים.
- כלי העיבוד היחיד עבור הנתונים הגולמיים המונגשים לציבור הוא כלי אגרגציה המציע ניתוחים המאגדים את שליפות המאגר לפי שורה של מדדים מובנים: שנים, אזורים, מדינות, הקשרים בילטרליים, אינטנסיביות וכולי. חלק מתוצרי האגרגציה של השליפות זמינים באתר וחלק ניתן להזמין באמצעות פנייה לחוקרי הרווארד.

בשורה התחתונה, ניתן למפות את ההבדלים הבולטים בין המאגרים והתאמתם למחקר האופרטיבי (ראו איור 30):

איור 30: ההבדלים בין המאגרים

מקרה מבחן - מאפייני הלחימה במזרח התיכון 2011־2017

משליפת הנתונים בשני המאגרים עולה כי היקף האירועים בשני המאגרים שונה בסדרי גודל וככל הנראה אף גדול מהיקף האירועים האמיתי. מלבד ההסבר שמקורו בחוסר יכולת המאגרים לזהות דיווחים כפולים ולמנות אותם כאירוע אחד, נראה כי הדיווח העיתונאי לרוב אינו מכיל פרטים נחוצים המבדילים בין אירוע למשנהו.

"דלות" הקידוד אינה מאפשרת עריכת מחקרים מעמיקים לצורך מיפוי צורת הקרב, כגון בחינת יחסי הגנה-התקפה, העמקה בנושא צורות קרב משניות (קרב התקדמות, נסיגה והשהיה, וכולי). נוסף על כך קשה (אך לא בלתי אפשרי) לעשות שימוש בקידודים הללו לבחינת מדדים כמו ניידות (נייד-נייח-קבוע), הרכב, גודל וארגון של המסגרות הלוחמות (גדוד-חטיבה-דיביזיה) והערכת סדר הכוחות האופייני של הלוחמים. כל אלו מהווים מכשול משמעותי בפני שימוש בכלים אלו בניתוחי תו"ל אויב.

בשני המאגרים, אפיון סוגי הכוחות הפעילים בזירות הקרב השונות מוכתב למעשה על ידי מוסרי הדיווחים התקשורתיים שאינם אחידים. ייתכן שכוח לחימה מסוים מדווח לעיתים כ"צבא" (מתוך הדגשת הלגיטימיות הפוליטית שלו) או כמיליציה (מתוך ניסיון/כוונת המדווח לערער את מעמד ויוקרת הכוח עליו מדווחים). לדוגמה, דיווח תקשורתי על אודות מיליציה בשם "צבא הניצחון", יוביל לקידוד אוטומטי של ה"שחקן" ככוח צבאי סדיר ובכך יעוות את המציאות.

סוגיות מתודולוגיות שעלו מתוך ניתוח מקרי המבחן

במחקרים המשלבים שימוש במאגרי מידע מהסוג הנסקר קיים הכרח "לנרמל" את התוצאות המתקבלות במיוחד אם מדובר בנתונים שנאספו לאורך זמן או קודדו על ידי מאגרים שונים, או אף דווחו בשפות שונות.[16] דרך פשוטה יחסית לגשר על הפערים היא באמצעות תהליך לתיאום נורמות המדידה המקובלות לסוגיה מסוימת (נרמול).[17] כך, למשל, בחינת כמות אירועי הלחימה לאורך שנים נכון שתיעשה בהקשר רחב של עלייה כללית בהיקף הדיווח התקשורתי. בדרך כלל ניתן לנרמל נתונים באחת מן השיטות הסטטיסטיות המוכרות:

- הצגת הנתונים על בסיס שתי סקלות שונות באותו גרף.
- הצגת הנתונים באחוזים ולא בערכים מוחלטים (מתאים להצגת ממצאים כגון "כמות מתוך מדגם").
- המרת הנתונים לערכי ציון תקן (Z) - שיטה מתאימה לערכים שבהם קיימת חשיבות הן לערך ספציפי והן לפיזור השוואתי של הערכים לאורך הסקלה).

בעידן הנוכחי, כמויות עצומות של ידע ומידע גלוי מאפשרות את הרחבת יריעת המחקר על ידי ביצוע בחינות על בסיס מדגמים רחבי היקף

- מעבר לאי־דיוקים, דיווח חלקי וקידוד דליל, שיטת הקידוד הקיימת במאגרים טומנת בחובה בעייתיות מובנית בקידוד אירוע "צבאי" בייחוד זה המאופיין על ידי כמה קטגוריות ותתי־קטגוריות במקביל (לדוגמה, בחינה במקביל של מהלכים, שיטות לחימה, אמצעי לחימה ועוד).
- ניתן להתמודד עם הדלות היחסית בקידוד הנתונים הנאספים ונשמרים במאגר באמצעות ייבוא הנתונים והעמקת ניתוחם בשלב ניתוח שני (משלים) שיאפשר לחוקרים ליישם מניפולציות שונות ומורכבות יותר על ידי שימוש בשיטה של multi-layer analysis,[18] וליישם כלים מתקדמים יותר (דוגמת STATA, שפת R או פיית'ון) לטובת ניתוח מעמיק ושיטתי של הנתונים שיאפשר העמקה בדיווחי התקשורת, מילוי החסרים שזוהו בשלב הקודם ובכך השלמת תמונת המציאות שאותה הוא מבקש לבחון או לתאר.

סיכום והמלצות

עידן הביג דאטה מציב הזדמנויות לצד אתגרים עבור חוקרי מודיעין בכל הדיסציפלינות ועבור החוקרים האופרטיביים בפרט. ברקע גובר השימוש בכלים ובשיטות מחקר מתקדמים בסוגיות אופרטיביות, טקטיות ומיקרו־טקטיות. בעידן הנוכחי, כמויות עצומות של ידע ומידע גלוי מאפשרות להרחיב את יריעת המחקר על ידי ביצוע בחינות על בסיס מדגמים רחבי היקף; המערכות

16 כושר התמצאות של מנועי הבנת הנקרא של אותו מאגר יכול להשתנות משמעותית בשפות שונות.

17 כך, לדוגמה, במחקר רכש התעצמות צבאי יש "לנרמל" את שוויו של הרכש לערך המטבע לאורך זמן בדי לאפשר השוואת הנתונים ביחס לתקציב או לתמ"ג ובן לצרכים השוואתיים אחרים.

18 שיטה זו מונחית רציונל של בחינה בשלבים וברמות ניתוח שונות דרך שימוש בכלים ובשיטות מחקר שונות. לדוגמה, שליפת נתונים ממאגרים היא שיטה אחת (או רמה אחת) של ניתוח. מניפולציות כמותניות דרך שימוש בstata־ או כל כלי סטטיסטי אחר מאפשר מיצוי מעמיק ומורכב יותר של הדאטה ומהווה שיטת מחקר ורמת ניתוח שונה ונוספת.

הקיימות והנגישות לחוקרים מאפשרות בחינת תופעות בהקשר אזורי (ואף עולמי) רחב, תוך עריכות השוואות פשוטות יחסית בין שחקנים שונים, במטרה לזהות ולאפיין דפוסי שינוי ולמידה לאורך זמן; השימוש בכלים אלו מאפשר לייצר תובנות על אודות התפתחויות ודינמיקות שונות שניתן לזהות ולמפות רק מתוך הקשר רחב של אירועים; תחקור המאגרים מאפשר פרסום תוצרים בסיווג נמוך (הנגיש לצרכנים בכל הרמות) ובכך מתגבר על מכשולי המידור וסיווג מקורות; עוד בולטת האפשרות לבצע אקספלורציה מהירה להצפת נקודות עניין רלוונטיות למחקרי המשך (בשיטות שונות).

השימוש בכלים כמותיים אלו מאפשר יכולת אקספלורציה רחבה, שדרכה ניתן להצביע על שורה של תופעות לא מוכרות. בזכות השימוש בשיטות אלו ניתן למצוא תימוכין כמותיים לקביעות איכותניות ואמירות "אינטואיטיביות". המאגרים שנבחנו, לדעתנו, הם בעלי פוטנציאל בינוני לתרום באופן ישיר למחקר בצי"חים אופרטיביים בשל שורה של בעיות אונטולוגיות ומתודולוגיות. אולם אנו ממליצים על שימוש בהם במסגרת מחקרים מודיעיניים לאור:

- תרומתם המסתמנת בתור נדבך משלים במחקר האופרטיבי בעיקר לביצוע ברירה מהירה של העשרות מחקריות, איתור מגמות ואיסוף תימוכין כמותיים לחיזוק ממצאי מחקר מודיעיני, תוך ניצול פשטות השימוש והזמינות של הכלים הללו.
- המאגרים מאפשרים לייצר מערך מחקר דו־שלבי. בשלב הראשון, נכון לעשות שימוש במאגר להצבעה על מקורות הטקסט הרלוונטיים ולהרכשתם באמצעות כלי ההרכשה שלו, ובשלב שני ניתן לבצע ניתוח שיטתי של הטקסטים שנאספו באמצעות כלי ניתוח טקסט. עיבוד וניתוח הממצאים במערך מחקר דו־שלבי מסוג זה יאפשר לחוקרים למקסם את הידע הקיים במאגרים ולחדד את התובנות העולות מהנתונים בצורה מדויקת ומותאמת יותר לשאלת המחקר.
- יודגש כי במהלך המחקר בלטו התועלות הרבות שאותן מאפשרים כלים בתחומים המדיניים (חוץ ופנים), מחקר חברה וציבור ובמחקר טרור. כל אלו במהירות ובפשטות יחסית, תוך מיצוי יתרונותיה של תקשורת גלויה במחקרים בתחומים אלו.

לאור היתרון המסתמן למאגר ה־ICEWS מבחינת מנגנוני הזיהוי ו"הבנת הנקרא" וחרף הקושי הנוכחי להיעזר במאגר בצורה מלאה (מחייב הסדרת הנגישות מול המערכת הביטחונית האמריקאית) אנו ממליצים להיעזר במקביל בשני המאגרים שנבחנו ובמאגרים נוספים לפי ההקשר.

בשורה התחתונה, בחינת השימוש במתודולוגיית מיצוי מידע דרך מאגרים ממוכנים מחזקת את תרומתה ככלי מחקר משלים, המעניק לחוקר את הנדרש כדי להתמודד עם הצמיחה בהיקף המידע הזמין שדרכו ניתן לאושש או להפריך תובנות מחקריות קיימות ולכמת אותן. מתודולוגיה זו מביאה את החוקר לקדמת השדה המחקרי העכשווי בכך שהיא מאפשרת סינתזה בין היכרותו המסורתית עם מושא המחקר לבין התובנות העולות ממאגרי מידע ממוכנים ולהטמיע את השימוש בהם כנדבך אנליטי נוסף ומעשיר. לפיכך על אף המגבלות המובנות בכלים מחקריים אלו, להערכתנו, יש לטייב את היכרות של החוקרים עם כלי מחקר למיצוי מאגרי מידע שונים.