

פרויקט גמר

כריית נתונים במדיה הדיגיטלית וברשתות חברתיות בנושא נגיף הקורונה מצגת מס' 2



מגמת ניהול פרויקטים
מאי 2021

חברות צוות 133

עדי מבורך

שקד עוגן

מלכה יחייס

מנחת הפרויקט

ד"ר אירנה מילשטיין

תוכן עניינים

רקע לפרויקט

מטרות הפרויקט

מתודולוגיית הפרויקט

אוכלוסיית המאומתים: ניתוחים סטטיסטיים

אוכלוסיית הלא מאומתים: ניתוחים סטטיסטיים

אוכלוסייה מרובת ציוצים: ניתוחים סטטיסטיים

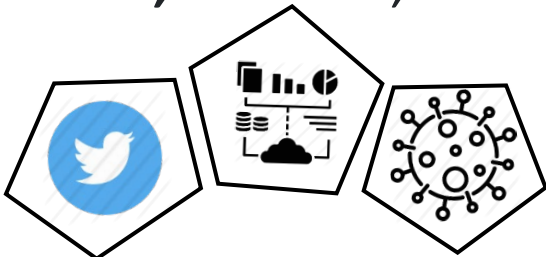
ממצאים

מגבלות

כיווני המשך

רקע

- כיום, בעידן ה- **BIG DATA**, קיימים **מאגרי מידע רבים**. הנתונים **מבוזרים**, מגוונים ובאיכות שונה.
- כתוצאה מכך, נולד צורך בביצוע **שליפות מהירות** של מידע, הן בדרכים קבועות מראש, אך בעיקר בדרך של זיהוי מהיר של תבניות וקשרים שונים (Losiewicz, Oard, & Kostoff, 2000).
- מידע אודות נגיף הקורונה, COVID-19, מופץ בכמויות גדולות **במדיה הדיגיטלית וברשתות החברתיות** בתדירות גבוהה.
- לפיכך, עלה הצורך **לבצע ניתוח נתונים מחקרי וכריית מידע** בנושא נגיף הקורונה מהרשתות החברתיות בדגש על "טוויטר" על מנת **לגלות מידע חשוב על אופן העברת המסרים, תדירותם, בחינת קורלציות ואף גיבוש מסקנות אודותם**.



מטרות הפרויקט

1

בניית תמונת מצב לפי תוצאות ניתוח נתונים בנושא נגיף הקורונה
מהרשת החברתית "טוויטר" על מנת לגלות מידע על אופן העברת
המסרים, תדירותם, הקשרים ביניהם ומובילי דעה.

2

הסקת מסקנות בתחומים שונים (כלכלה, חברה, חינוך, בריאות וביטחון)
בשיטת כריית טקסט.



מתודולוגיית הפרויקט

- מציאת מאגר נתונים שמכיל את כל הציוצים בחודשים יולי- אוגוסט 2020 ברשת החברתית "טוויטר" בנושא הקורונה (<https://www.kaggle.com/>)
- סינון הנתונים הגולמיים ובניית תת מדגמים ב-8 שלבים שונים
- חישוב מדדים של סטטיסטיקה תיאורית ובניית היסטוגרמות
- חישוב קורלציות והרצת רגרסיה בתוכנת SPSS לשלושה תת מדגמים (אנשי ציבור, משתמשים פרטיים, אוכלוסייה מרובת ציוצים)
- **ניתוח טקסטואלי** של מילים מרכזיות נבחרות שעומדות בחזית המאבק בנגיף בקורונה, בחינה של מידת השפעתן על האוכלוסיות השונות והצגתה באופן **ויזואלי בתוכנת NODEXL**

סינון נתונים

שלב	תיאור	מס' לפני סינון	מס' לאחר סינון	סה"כ הוסרו
I	משתמשים אשר שמותם כתובים בג'יבריש הוסרו. מספיק שהשם הכיל אות לא תקינה היא הוסרה מהרשימה. דוגמאות לאותיות כאלה: Žè<¥æ°´ francisw ë°"ë^æ	179,108	153,136	25,972
II	הסרת עמודות לא רלוונטיות	153,136	153,136	0
III	הסרת משתמשים אשר צייצו פעם אחת	153,136	97,201	55,935
IV	הסרת משתמשים אשר צייצו פעם אחת ומכילים סימן ~	97,201	97,103	98
V	בניית מדגם משתמשים ע"י הסרת הערכים הכפולים ושימוש בפונקציית הסרת כפילויות- מאפיין משתמש.	97,103	20,539	76,564



סינון נתונים

שלב	תיאור	מס' לפני סינון	מס' לאחר סינון	סה"כ הוסרו
VI	יצירת תת מדגם של משתמשים שאינם מאומתים – FALSE מתוך מאפיין משתמש	25,539	18,133	7,406
	יצירת תת מדגם של משתמשים שאינם מאומתים – FALSE מתוך מאפיין ציוצים	97,103	78,970	17,133
VII	יצירת תת מדגם של משתמשים מאומתים – TRUE מתוך מאפיין משתמש	25,539	2,406	23,133
	יצירת תת מדגם של משתמשים מאומתים – TRUE מתוך מאפיין ציוצים	97,103	18,133	78,970
VIII	יצירת מדגם אוכלוסייה שצייצה יותר מ-25 ציוצים- מאפיין משתמש	25,539	340	25,199
	יצירת מדגם אוכלוסייה מרובת לייקים	97,103	79	97,024



אוכלוסיית המאומתים: סטטיסטיקה תיאורית

➤ גודל המדגם: 2,406 תצפיות

➤ חושבו ערכים של לוגריתם טבעי למס' עוקבים ומס' חברים עקב מספרם הגדול

מספר החיובים	מינימום	מקסימום	ממוצע	סטיית תקן
מספר עוקבים	2.00	390.00	7.58	20.11
מספר חברים	0.00	359,113.00	3,093.92	14,740.86
לוגריתם טבעי מספר עוקבים	4.57	16.44	10.37	1.84
לוגריתם טבעי מספר חברים	0.00	12.79	6.66	1.67

אוכלוסיית המאומתים: קורלציה

		num_tw	ln_fo	ln_fr
num_tw	Pearson Correlation	1	.220**	-.160**
	Sig. (2-tailed)		.000	.000
	N	2406	2406	2398
ln_fo	Pearson Correlation	.220**	1	.048*
	Sig. (2-tailed)	.000		.019
	N	2406	2406	2398
ln_fr	Pearson Correlation	-.160**	.048*	1
	Sig. (2-tailed)	.000	.019	
	N	2398	2398	2398

** . Correlation is significant at the 0.01 level (2-tailed).

* . Correlation is significant at the 0.05 level (2-tailed).

➤ ככל שיש יותר עוקבים (ln_fo), ישנם יותר ציוצים (num_tw)

➤ ככל שיש יותר חברים (ln_fr), ישנם פחות ציוצים (num_tw)

אוכלוסיית המאומתים: רגרסיה

➤ **משתנה מוסבר:** מספר הציוצים (num_tw)

➤ **משתנים מסבירים:** לוגריתם טבעי של מספר עוקבים (ln_fo), לוגריתם טבעי של מספר חברים (ln_fr)

➤ **אחוז השונות המוסברת שווה ל-7.5%**

➤ **טיב המודל גבוה (F-Sign קטן מ-0.000)**

Model		Unstandardized Coefficients		Standardized	t	Sig.
		B	Std. Error	Coefficients Beta		
1	(Constant)	-3.685	2.630		-1.401	.161
	ln_fo	2.364	.210	.222	11.278	.000
	ln_fr	-2.002	.231	-.171	-8.668	.000

a. Dependent Variable: num_tw

אוכלוסיית הלא מאומתים: סטטיסטיקה תיאורית

גודל המדגם: 18,133 תצפיות ➤

סטיית תקן	ממוצע	מקסימום	מינימום	
10.88	4.35	679.00	2.00	מספר הציוצים
20,098.26	45,713,139.00	934,774.00	0.00	מספר עוקבים
8,317.82	1,937.66	445,635.00	0.00	מספר חברים

אוכלוסיית הלא מאומתים: קורלציה

		NUM_TW	NUM_FOL	NUM_FR
NUM_TW	Pearson Correlation	1	.049**	.023**
	Sig. (2-tailed)		.000	.002
	N	18133	18133	18133
NUM_FOL	Pearson Correlation	.049**	1	.466**
	Sig. (2-tailed)	.000		.000
	N	18133	18133	18133
NUM_FR	Pearson Correlation	.023**	.466**	1
	Sig. (2-tailed)	.002	.000	
	N	18133	18133	18133

** . Correlation is significant at the 0.01 level (2-tailed).

➤ ככל שיש יותר חברים (NUM_FR), ישנם יותר עוקבים (NUM_FOL)

אוכלוסיית הלא מאומתים: רגרסיה

➤ משתנה מוסבר : מספר עוקבים (NUM_FOL)

➤ משתנים מסבירים: מספר ציוצים (NUM_TW), מספר חברים (NUM_FR)

➤ אחוז השונות המוסברת שווה ל- 21.9%

➤ טיב המודל גבוה (F-Sign קטן מ- 0.000)

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	2083.770	145.133		14.358	.000
	NUM_TW	71.016	12.129	.038	5.855	.000
	NUM_FR	1.124	.016	.465	70.864	.000

a. Dependent Variable: NUM_FOL

אוכלוסייה מרובת ציוצים: סטטיסטיקה תיאורית

➤ גודל המדגם: 340 תצפיות

➤ חושבו ערכים של לוגריתם טבעי למס' עוקבים, מס' חברים ומס' לייקים מקסימלי עקב מספרם הגדול

מנימום	מקסימום	ממוצע	סטיית תקן
26.00	679.00	65.64	69.92
0.00	13,849,157.00	495,363.92	1,788,531.27
0.00	72,628.00	2,263.69	6,996.93
0.00	16.44	8.97	3.36
0.00	11.19	5.64	2.43
0.00	1.00	0.32	0.47
0.00	550,595.00	12,469.26	44,318.82
0.00	13.22	6.28	3.11

מספר הציוצים

מספר עוקבים

מספר חברים

לוגריתם טבעי מספר עוקבים

לוגריתם טבעי מספר חברים

משתמש מאומת/ לא מאומת (0/1)

מקסימום לייקים

לוגריתם טבעי מקסימום לייקים

אוכלוסייה מרובת ציוצים: קורלציה

		num_of_tweet	LN_FO	LN_FR	user_verified	LN_FAV
num_of_tweet	Pearson Correlation	1	.068	-.104	.073	-.143**
	Sig. (2-tailed)		.213	.055	.180	.008
	N	340	339	340	340	340
LN_FO	Pearson Correlation	.068	1	.224**	.691**	.107*
	Sig. (2-tailed)	.213		.000	.000	.049
	N	339	339	339	339	339
LN_FR	Pearson Correlation	-.104	.224**	1	-.055	.587**
	Sig. (2-tailed)	.055	.000		.308	.000
	N	340	339	340	340	340
user_verified	Pearson Correlation	.073	.691**	-.055	1	-.070
	Sig. (2-tailed)	.180	.000	.308		.195
	N	340	339	340	340	340
LN_FAV	Pearson Correlation	-.143**	.107*	.587**	-.070	1
	Sig. (2-tailed)	.008	.049	.000	.195	
	N	340	339	340	340	340

** . Correlation is significant at the 0.01 level (2-tailed).

* . Correlation is significant at the 0.05 level (2-tailed).

➤ ככל שיש יותר חברים (LN_FR), ישנם יותר לייקים (LN_FAV)

אוכלוסייה מרובת ציוצים: רגרסיה

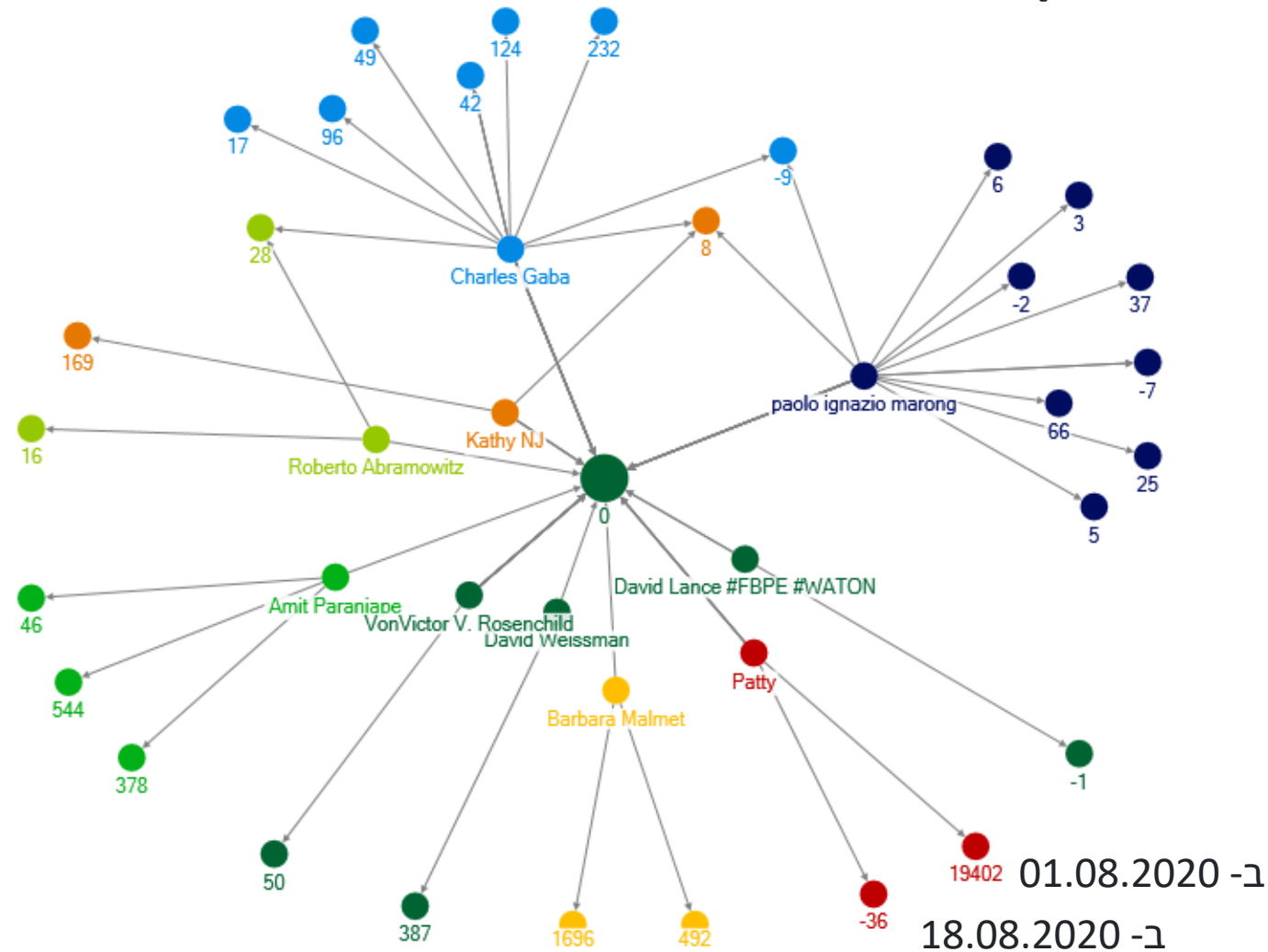
- **משתנה מוסבר:** מספר לייקים מקסימלי (LN_FAV)
- **משתנים מסבירים:** לוגריתם טבעי של מספר עוקבים (LN_FO), לוגריתם טבעי של מספר חברים (LN_FR), משתנה בינארי למאומתים (user_verified) וספר ציוצים (num_of_tweet)
- אחוז השונות המוסברת שווה ל- 33.9%
- **טיב המודל גבוה** (F-Sign קטן מ- 0.000)

Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.	
	B	Std. Error	Beta			
3	(Constant)	2.053	.347	5.914	<.001	
	LN_FR	.750	.056	.586	13.285	<.001

a. Dependent Variable: LN_FAV

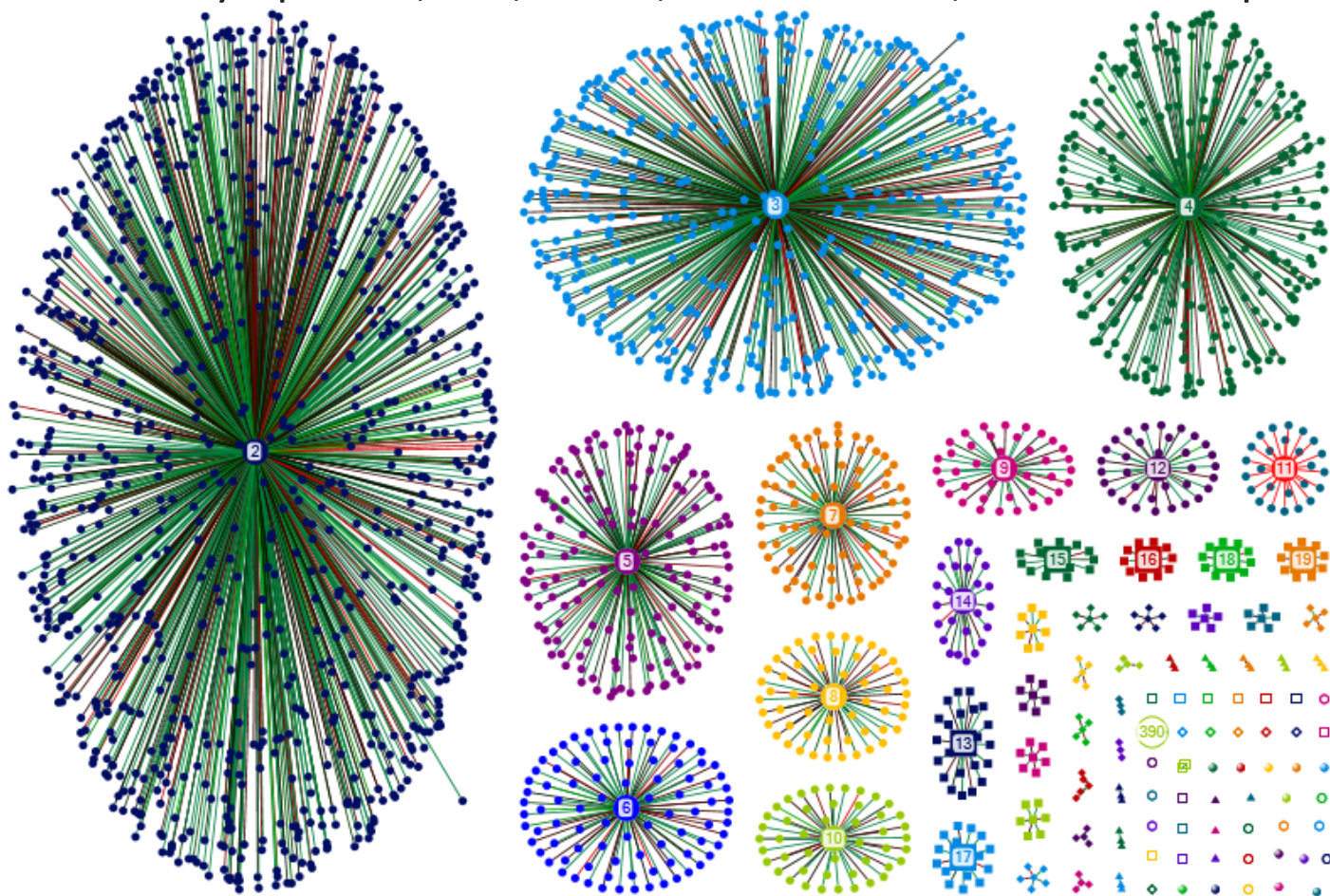
אוכלוסייה מרובת לייקים: זיהוי מובילי דעה

✓ ניתוח תנודתיות כמות העוקבים



אוכלוסיית המאומתים: כמות ציורים ותכניה

- ✓ ניתוח תכני הציורים מציג ויזואלית בכל אשכול האם הציורך הכיל תוכן חיובי או שלילי.
- צבע **ירוק** מצביע שבצורך מילים חיוביות, לדוגמא: protection ,win ,happy ,safe ,positive ועוד.
 - צבע **אדום** מצביע בצורך מילים שליליות, לדוגמא : symptoms ,risk ,death ,crisis ועוד.



ממצאים ומסקנות

➤ בקרב אוכלוסיית המאומתים (אנשי ציבור ומובילי דעת קהל), ככל שיש מספר עוקבים גבוה זה מחייב

אותם לצייץ יותר, על מנת לעניין את העוקבים ולשמור על כמות עוקבים גבוהה.

➤ בקרב אוכלוסיית המאומתים שהינם בעלי מספר חברים גבוה, מספר הציוצים קטן זאת לאור שאותם

משתמשים מעדיפים לצרוך מידע מציוצים של חבריהם מאשר לצייץ.

➤ בקרב אוכלוסיית הלא מאומתים (אנשים פרטיים) מצייצים יותר על מנת לגרוף עוקבים. משתמשים

רבים "בטוויטר" מעוניינים להגדיל את מספר העוקבים שלהם ולהפוך למובילי דעת קהל.

ממצאים ומסקנות

➤ בקרב אוכלוסייה מרובת ציוצים, רב המשתמשים אינם מאומתים. מדגם זה מורכב מ- $2/3$ אנשים

פרטיים ו- $1/3$ אנשי ציבור ומובילי דעת קהל. אנו למדים שרב המשתמשים המצייצים פעמים רבות

עושים זאת על מנת להפוך לסטאטוס מובילי דעת קהל.

➤ בקרב אוכלוסייה מרובת ציוצים, ככל שיש יותר חברים ישנם יותר לייקים. בדרך כלל מעגל החברים של

המשתמש הינו המעגל הקרוב אליו שסביר להניח שגם הוא במעגל העוקבים של המשתמש ולכן

מתעניין בציוצי המשתמש ויפרגן לו.

ממצאים ומסקנות

➤ מניתוח הטקסט של הציוצים עולה כי בקרב האוכלוסייה המאומתת, זהו בעיקר מילים **חיוביות** אודות

נגיף הקורונה (**1,206 מילים**) לעומת (**634 מילים**) שליליות.

➤ מניתוח **צמדי המילים** העיקריות שנאמרו המילה **Covid-19** ו- **Cases** חזרו על עצמם **107** פעמים ניתן

להסיק מכך שבתקופה הנבחנת השיח הרוב עסק בעיקר במקרי תחלואה.

מגבלות

➤ ה- Database הראשוני הכיל 179,108 רשומות ודרש **פעולות מקדימות רבות טרם הניתוח:**

- רשומות המורכבות **משפות שונות**- נדרש לבצע סינון ראשוני והסרת רשומות בלתי מובנות.

- **ריבוי עמודות שאינן מהמנות** ללא תועלת אפקטיבית לניתוח נתונים. לדוגמא- בעמודת מיקום כל משתמש

רשאי לרשום מיקום על פי רצונו. מנגד על מנת להפיק מידע מהימן **הוספנו עמודות חישוביות** (מספר הציוצים

למשתמש, כמות מספר חברים/ עוקבים שהתווספו/ ירדו)

- הרשומות מדווחות על ציוצים בחודשים יולי- אוגוסט. מגבלה זה צמצמה את **כיווני הניתוח** מכיוון שהציוצים

עסקו בעיקר בתחילת המגפה שהשיח היה על נתוני תחלואה, מניעת תחלואת ושיח פוליטי סביב המגפה.

➤ כריית הטקסט דרשה **ניתוח ידני רב על מנת להפיק נתונים מהמנים.**

כיווני המשך

➤ ניתוח Database **עדכני** מהתקופה האחרונה בנושא הקורונה העוסק ב- **חיסון הקורונה, מתנגדי**

החיסונים ואחוזי החלמה.

➤ **ביצוע השוואה** בין ממצאי הניתוח העדכני לבין ממצאי הניתוח הראשוני אותו בצענו בפרויקט זה.

➤ **בחינה והשוואה** בין תוצאות ניתוח רשת ה"טוויטר" אותו בצענו לבין יתר רשתות המדיה-

. Tictok, Instagram ,Facebook

