

ניתוח וניתוח מידע כוח ברשת החברתית "Twitter" במהלך משבר הكورونا

נכתב על ידי:

יקיר אטיאס
מנגשה אטלאי

בהנחייה:

ד"רABI YOSIFOV



תוכן עניינים



락ט

אינטרנט

בעדן הרשותות החברתיות, מידע כוזב יכול להתפשט ברשת בשניות ובמקרים חירום ליצור נזקים בלתי הפיכים.



בלבול

ריבוי תיאוריות הקונספירציה מקשה על אנשים לה辨ין בין מידע מדויק ולא מדויק, ומוביל לבלבול וחוסר אמון במקורות מידע רשמיים.



אנשים

עם התקדמות הטכנולוגיה, אנשים יותר ויותר משתמשים על מידע מהאינטרנט.



קונספירציות

אירועים ואסונות טבעי מייצרים חששות אצל אנשים, שמחפשים סיבות לאירועים בעקבות זה ומכאן נוצרות "শমুেত"- מידע מוטעה בתשובה לשאלת מהו הגורם לאירוע אסון



מטרת המחקר

מטרת המחקר הוא לנתח ולנטר מידע כוזב בטוויטר בזמן חירום, במטרה להבין את עומק הדיוון והഫזה של מידע כוזב בנושא הקורונה במדינות אירופה, ארצות הברית וישראל.

הדגשים העיקריים הם:

א. לבצע השוואה בין שפות שונות בנגע למידע כוזב שימושותי בטוויטר בשעת חירום.

ב. לזרות את עומק השיח החברתי בטוויטר על תאוריות קונspirציה שונות בזמן מגפת הקורונה כדי להבין את ההשפעה שיש להן בזמן מצבי חירום דומים אחרים.

ג. לנתח את תדיות השיח לאור המגפה בכל שפה.

שאלות מחקר

- 
- א. האם ניתן לנתח את המאפיינים ואת עומק השיח של תאוריית הקונספירציה 5G לאור רצון זמן בשפות שונות.
- ב. האם ישנה השפעה על כמות הציוצים ונפח הציוצים בין שפות שונות (מדיניות שונות) ומה הקשר ביניהם.

תהליכי מחקר

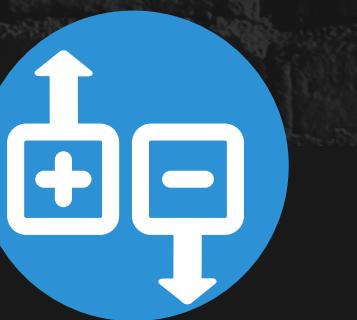


נתונים



5

ניקוי הנתונים (בדף
הבא)



4

אספנו 441,157 ציוצים



3

איסוף הנתונים בטוווח
התאריכים :
- 9.1.2020
8.10.2022



2

כל שפות אירופה,
עברית ואנגלית



1

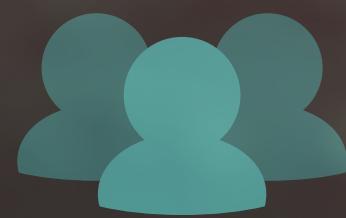
שימוש ברישיון API של
טוויטר למטרות מחקר

ניקוי הנתונים

בחלק מהמודלים שינינו את טווח הנתונים על מנת שיתאים למודל.



הסרת שפות שכמות הנתונים (ציוצים שלהן) אינה עולה על 100 לכל התקופה



הסרת נתונים לא רצויים וטיפול בערכים חסרים



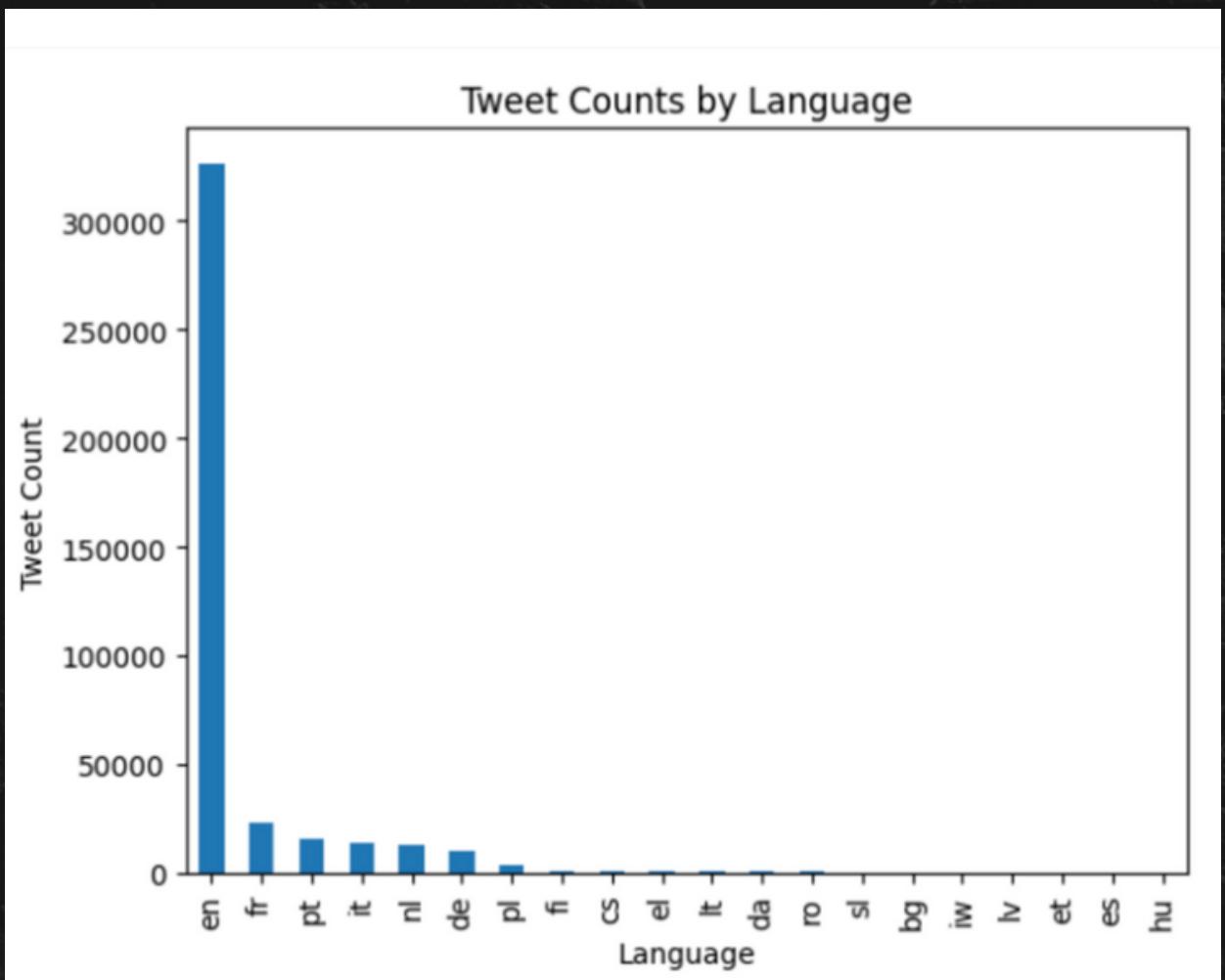
עיבוד מוקדים- הסרת כתובות URL, אזכור משתמש (@) והאשטאים.



המודלים שהשתמשו בפרויקט

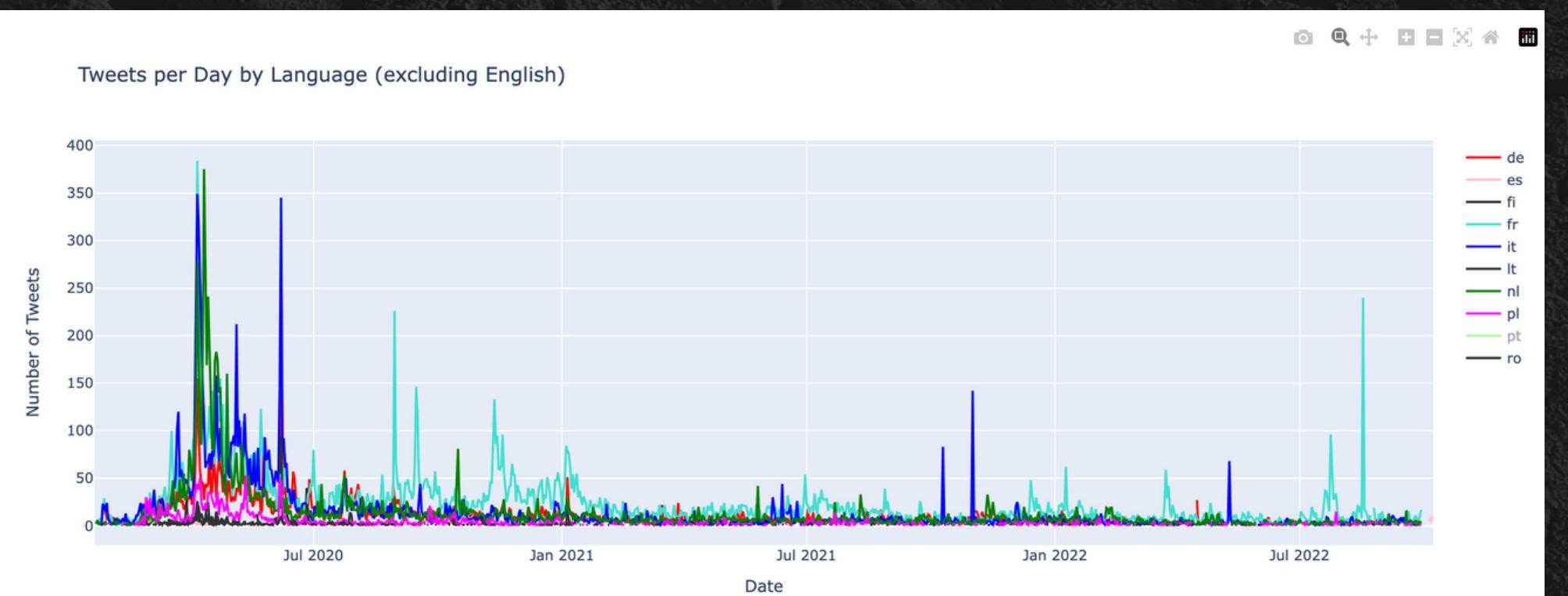
שם המודל	תיאור
Cross Correlation	מתאים קורולאציה הוא מודד סטטיסטי המראה את המידה שבה שני משתנים (במקרה שלנו - 2 שפות), נעים זה ביחס זהה
Auto Correlation	קורסיה אוטומטית היא בדיקה של פיגור בסדרת הזמן בשפה עצמה. כלומר ההשוואה היא לא בין 2 שפות, אלא בין אותה שפה בזמןים שונים
Auto Arima	מודל סטטיסטי המיועד לניבוי עתידי על פני זמן-בעזרתו ניתן לחזות פועלות ציוצים עתידית. עבור כל שפה
Word Cloud	הציג טקסטואלית של המילים הנפוצות ביותר בDATA
Sentiment Analysis	ניתוח רגשות. טכניקה המשמשת לקביעת הסנטימנט (הרגש) מאחורי קטע טקסט נתון, במקרה שלנו ציוץ
Bert NLP	ברט הוא מודל עיבוד שפה, המשמש כדי לסוג את הציוצים כמתנגדים/תמיכים/нейutrליים לكونספרציה

תוצאות



מספר הציטטים עברו כל שפה: ניתן לראות שהשפה האנגלית הינה בעלת נתח הציטטים הגדול ביותר (78.77%).

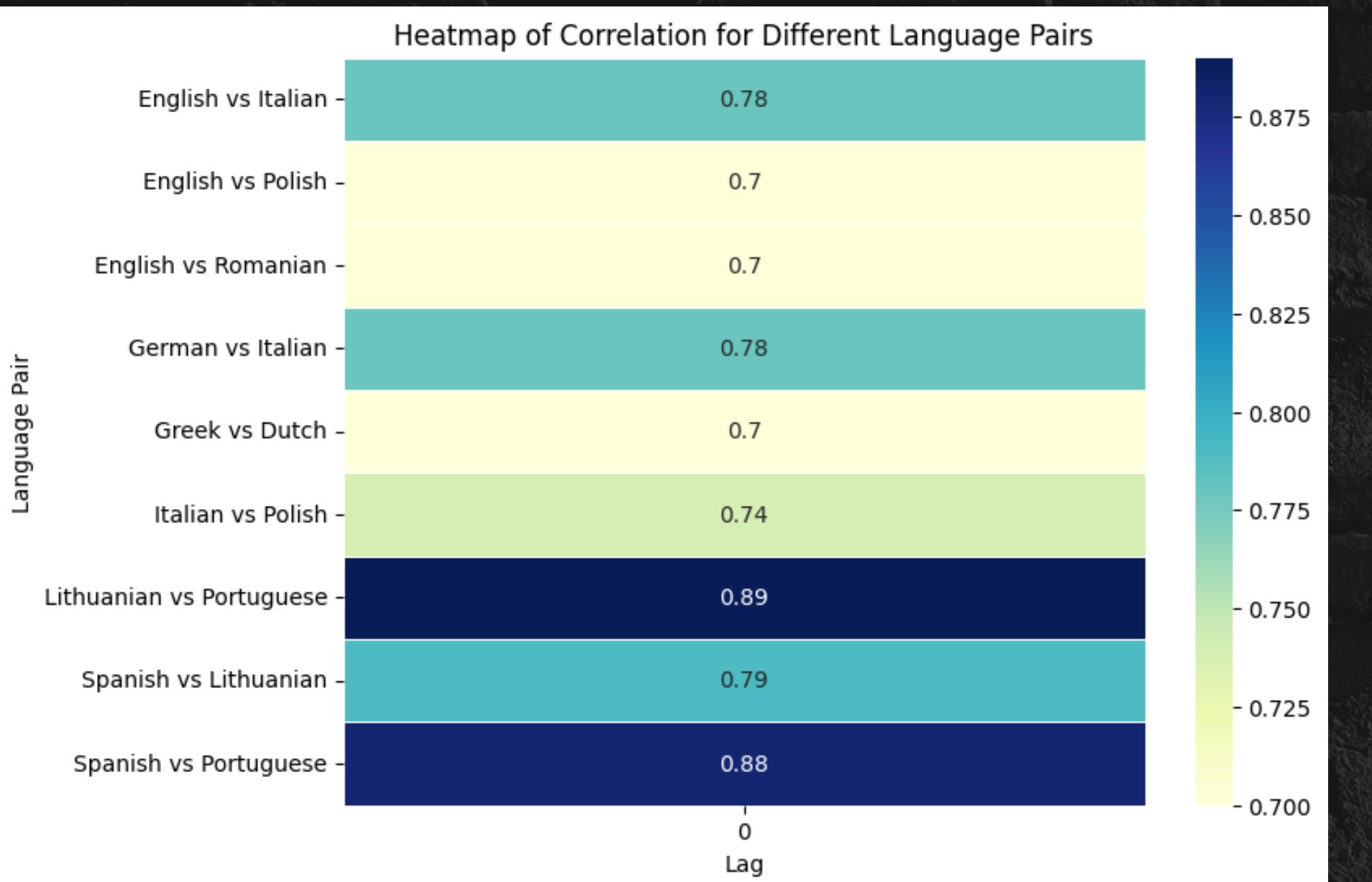
ניתן לראות שבשפות מסוימות יש קפיצות חדות בימים מסוימים, לאחר מחקר מעמיק מצאנו סיבות לכך מהקפיצות הללו.



מספר הציטטים בכל יום בכל שפה בטווח התאריכים הנתון.

תוצאות

מתאים קורלצייה



מפת חום- תוצאות ניתוח מתאים קורלצייה (המתאים החזקים ביותר)

מתאים קורלצייה הוא ממד סטטיסטי המראה את המידה שבה שני משתנים נעים זה ביחס זה.

מתאים של 1 אומר שני המשתנים מתואימים בצורה מושלמת, אם משתנה אחד גדול, גם המשתנה השני גדול.

מתאים של 1-(מתאים שלילי) אומר שני המשתנים נמצאים בקורסיה הפוכה לחלווטין. אם אחד גדול, השני יורד.

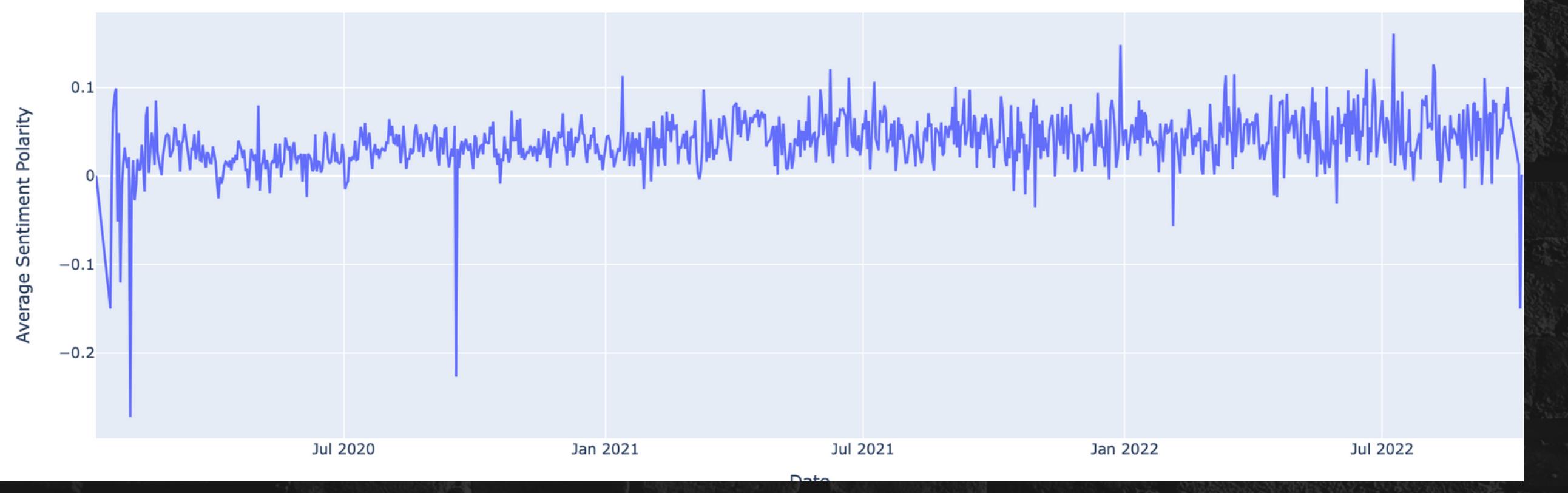
מתאים של 0 אומר שאין קשרlieniarו בין המשתנים (מדד פירסון).

"פיגור" (lag) בניתוח סדרות זמן מתייחס לפרק זמן מסוים שתكونה אחת (כמו מספר היציאות בשפה מסויימת) נגררת אחרי אחר.

תוצאות

ניתוח רגישות

Average Sentiment Polarity Per Day



ניתוח רגישות ממוצע לכל השפות ביחד לפי תאריכים.

ניתוח רגישות היא טכניקה המשמשת לקביעת הסנטימנט (הרגש) מאחורי קטע טקסט נתון, במרקחה שלנו ציוץ.

ניתוח זה יכול להיות מועיל בהבנת דעת הקהל, תגובה לאירועים או מגמות בדיוניים מקוונים.

במחקר שלנו, אנו משתמשים בניתוח סנטימנטים כדי לנתח את 441,000 היצירות הקשורות לתיאורית הקונספירציה של 5G בתקופת הקורונה.

תוצאות

	bg	cs	da	de	el	en	es
1	0.564733	0.488521	0.515346	0.648662	0.554858	0.916268	0.535785
et	fi	fr	hu	it	iw	lt	
1	0.322120	0.376152	0.743131	0.249967	0.809362	0.146398	0.598905
2	0.195669	0.193328	0.632759	0.152026	0.711629	0.078979	0.250440
3	0.280623	0.221291	0.571467	0.135605	0.626623	0.044037	0.231349
4	0.253789	0.157397	0.553193	0.135492	0.553540	0.065452	0.217405
5	0.207500	0.177299	0.529941	0.123143	0.492242	0.080372	0.189373
lv	nl	pl	pt	ro	sl		
1	0.302398	0.851962	0.752802	0.553461	0.595073	0.236964	
2	0.254267	0.740969	0.664062	0.090433	0.481847	0.243174	
3	0.237503	0.723456	0.637594	0.047954	0.361396	0.290617	
4	0.174566	0.741702	0.576030	0.033902	0.330259	0.201673	
5	0.130723	0.765556	0.536056	0.023255	0.357804	0.239607	

בשונה מקורלציה רגילה, קורלציה אוטומטית היא בדיקה של פיגור בסדרת הזמן בשפה עצמה.

באופן כללי, יש קורלציה חיובית לפיגור של 1 בכל השפות, מה שמשמעותו על אף של ציוצים הדומים בתיאוריות הكونספירציה יש רמה מסוימת של תלות בעוצמת הציוויים של היום הקודם.

מידת הקורלציה הנ"ל משתנה בין השפות. לדוגמה, ציוצים באנגלית מציגים את המתאם האוטומטי הגבוה ביותר בפיגור 1 (0.9163), בעוד שציוצים בעברית מציגים את הנמוך ביותר (0.1464).

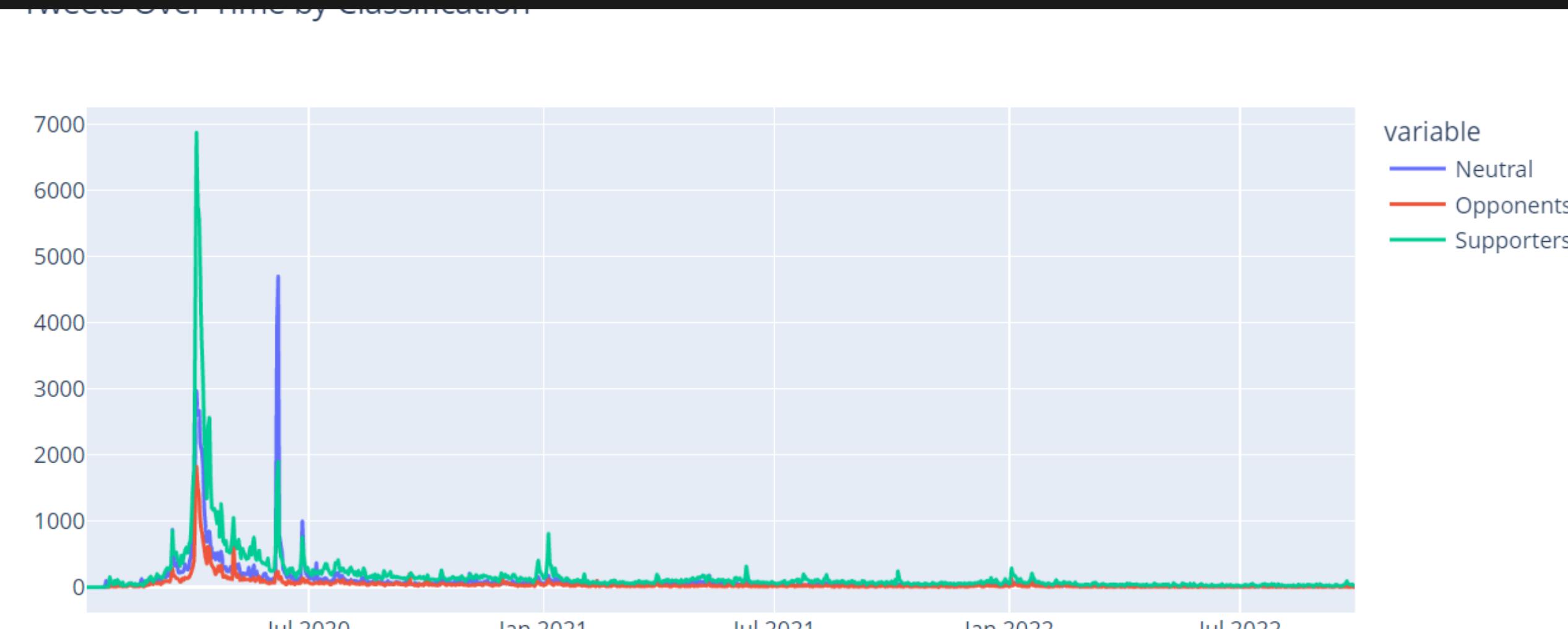
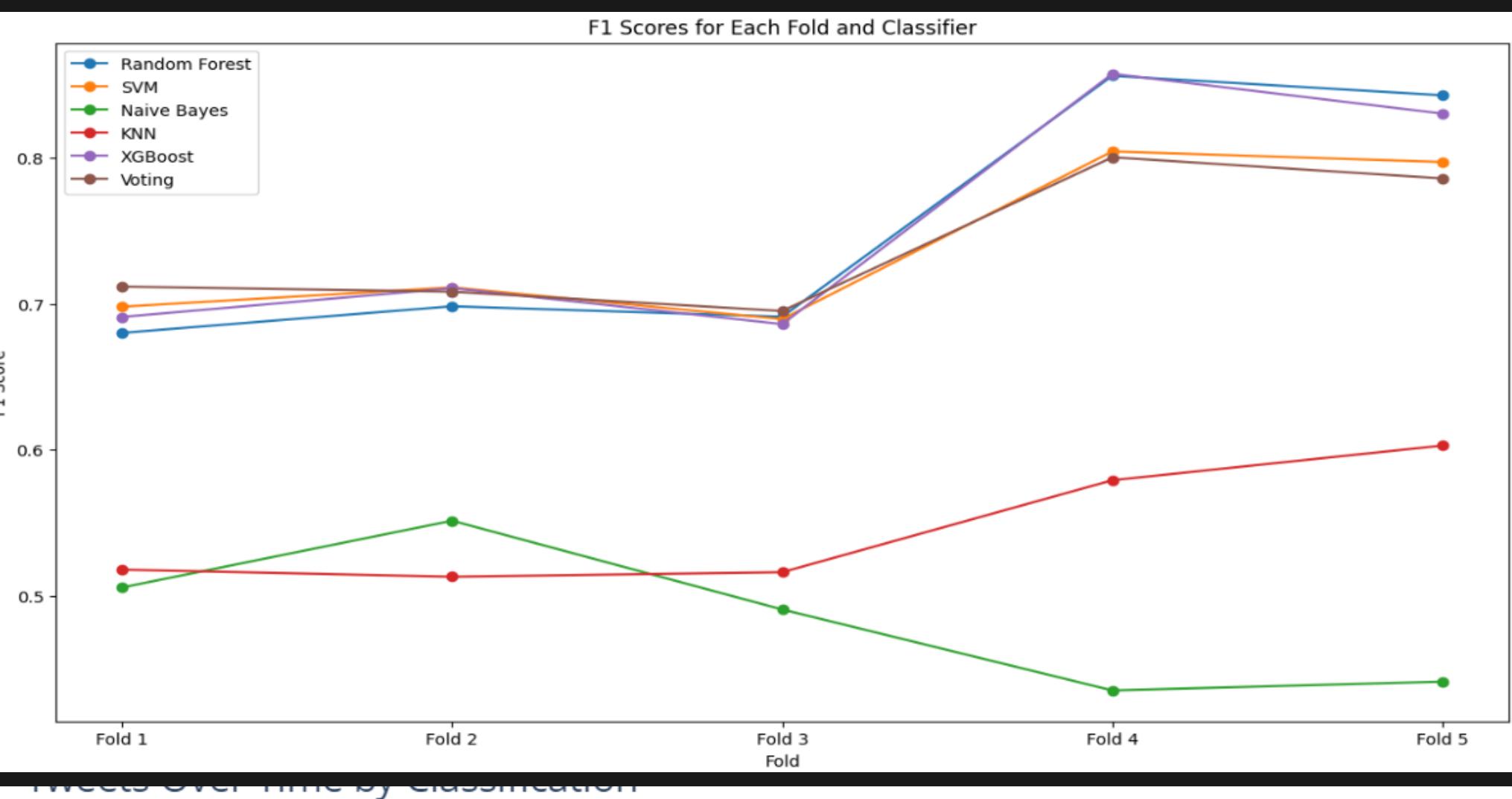
תוצאות

מודל BERT:
ברט הינו המודל המתקדם ביותר כיום להבנת שפה אנושית על ידי מחשבים.

- אימון המודל:
- השתמשנו במודל מסוג SBERT ו CT-BERT
 - fold cross validation python 5

- השתמשנו במודל XGBoost

תוצאות הסיווג:
תומכים - 183,577
מתנגדים - 46,627
нейтрלים - 95,914



מסקנות

ראינו שיאים משמעותיים בנפח הציוצים בהתאם לאיורים גדולים או הודיעות מדיניות. לדוגמה, בגרמניה, שיא בנפח הציוצים התרחש במקביל לויכוחים סביב אסטרטגיות פתיחה חדשה ושימוש במסכות.



ראינו שהנפח והרגש של ציווים הדנים בתיאוריות הקונספירציה מושפעים מהדינונים של היום הקודם. ממצא זה מדגיש את הצורך באסטרטגיות התערבות בזמן כדי למנוע הגברת מידע כוזב.



מסקנות



מצאנו מתאמים חזקים בין נפחី ציויצים בשפות מסוימות, מה שמצויע על כר שאיירועים גלובליים הפעילו דפוסי תגובה דומים. לדוגמה, דפוסי הציווץ על COVID-19 באיטלקית ובאנגלית הראו מתאם חזק, מה שמרמז על תגבות בו זמןית לאירועים גלובליים.



באופן בלתי צפוי, גילינו שההשפעה וההתמדה של דיונים על תיאוריות הקונספירציה השתנו בין השפות. זה מצויע על כר שגיישה המתאימה לכלום לניהול מידע שגוי עשוי להיות לא עיליה, ואסטרטגיות מותאמות בהתחשב בהבדלים לשוניים ותרבותיים עשויות להיות מוצלחות יותר.

תודה על ההקשבה!

