היבטים מעשיים בכריית תוכן אינטרנטי לאפליקציות עסקיות – פרויקט מסכם

שאלה 1

Preparation – בחלק של mount drive and load data מופיעים הנתיבים לקבצים:

Data_path – קובץ הנתונים של Kaggle

Tweets_path – קובץ הנתונים של הציוצים שאספנו נקרא CollectedTweets.csv

Google_trends_path – קובץ של גוגל טרנדס נקרא GoogleTrends.csv

A. Data loading - תחילה הורדנו את ה-dataset הדרוש מהאתר של Kaggle. לאחר מכן על מנת לטעון את הקובץ, השתמשנו בקידוד ISO-8859-1 וטענו את המידע לתוך משתנה df. עבור כל רגש חיובי ושלילי לקחנו 50 אלף רשומות, כאשר מכל רגש לקחנו את 50 אלף הרשומות הראשונות. סה"כ השתמשנו ב-100 אלף רשומות כאשר הדאטה מאוזן - 50% לכל רגש.

דוגמא ל-dataset לאחר הטעינה (ישנן עוד עמודות אך לא הצגנו אותן בדוגמא על מנת להשוות ל-dataset לאחר pre processing):

	text	target
0	I LOVE @Health4UandPets u guys r the best!!	4
1	im meeting up with one of my besties tonight! ...	4
2	@DaRealSunisaKim Thanks for the Twitter add, S...	4
3	Being sick can be really cheap when it hurts t...	4
4	@LovesBrooklyn2 he has that effect on everyone	4
...	...	...
99995	@mileycyrus so i have the same insomnia prob a...	0
99996	20 mintues late for my meeting starting @ 8 h...	0
99997	@kentucky_derby super excited! Are you tweetin...	0
99998	I WANT ANOTHER DAY OFF!!!! To much Sh#t to do...	0
99999	i just jacked up this umbrella cake	0

B. Pre-processing – לאחר מכן, ניקינו את הדאטה בתהליך המתאים לציוצים. תחילה הגדרנו ביטויים רגולריים עבור: emoticons, תגיות html, שמות משתמשים המתחילים ב-@, hashtags, URLs, מספרים, מילים אשר מכילות מקף או גרש – באופן דומה לתהליך שהוצג בכיתה.

שלב 1 - בעזרת הביטויים הללו ביצענו את שלב ה-tokenization כאשר עברנו על כל ציוץ ופירקנו אותו לרשימה של טוקנים.

שלב 2 - הפכנו את כל האותיות הגדולות לקטנות על מנת לשמור על אחידות ולמנוע מצב ששתי מילים זהות יתפסו כשונות.

שלב 3 – עברנו על כל רשומה והסרנו טוקנים שהכילו URL, למשל כאלו שמתחילים ב"http".

שלב 4 – עברנו על כל רשומה והסרנו טוקנים אשר הכילו רק ספרות.

שלב 5 - עברנו על כל רשומה והסרנו שמות משתמשים אשר מתחילים ב-@.

שלב 6 - עברנו על כל רשומה והסרנו מילים באורך קטן שווה ל-2 למשל "hm" או "a".

שלב 7 – עברנו על כל רשומה והסרנו מילות עצירה (stop words) וסימנים, **אך מכיוון שרשימת ה- stop words הכילה מילים שיכולות להועיל לנו בזיהוי הרגש, התאמנו את הרשימה לצרכים שלנו.** כלומר,

הסרנו מהרשימה מילים עם קונטקסט שלילי כמו "not", "shouldn't". למעשה הסרנו מהרשימה מילים אשר מכילות מילות שלילה ויכולות לשנות את הקונטקסט של המשפט.

- השלבים עד כאן נועדו להסיר מילים שלא מוסיפות מידע לרגש שלבים 3-7.

שלב 8 – עברנו על כל רשומה וביצענו stemming לכל טוקן על מנת להפוך מילים לצורה השורשית שלהן. ביצענו זאת על מנת שנוכל להתייחס למילים זהות אך הטיות זמן שונות למשל (live, living) באופן זהה ואחיד. השתמשנו ב- Snowball Stemmer שהוא שיפור של porter stemmer המפורסם.

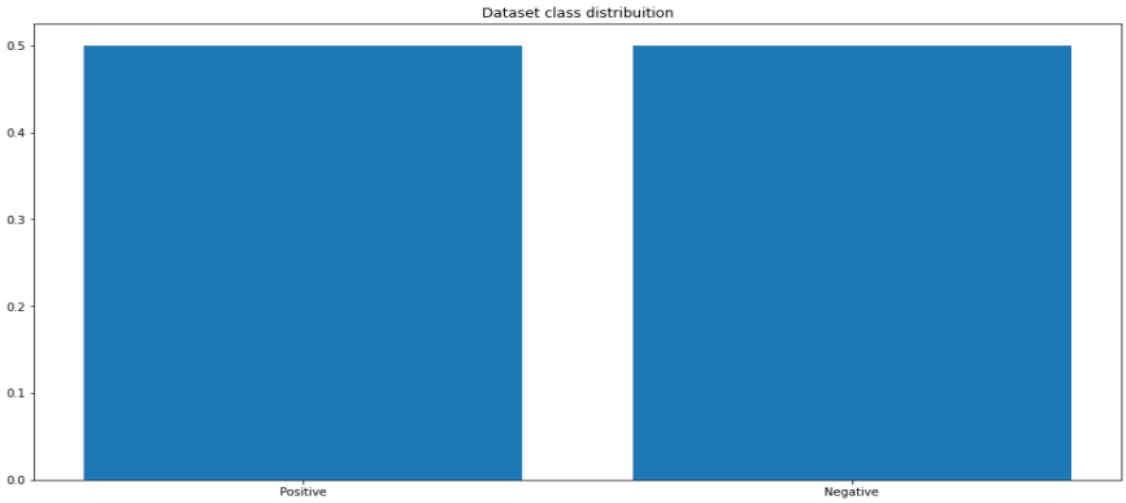
שלב 9 – ביצענו טרנספורמציה ללייבל של הקלאסים כאשר רגש שלילי שהיה מיוצג על ידי הספרה 0 נשאר להיות מיוצג על ידי הספרה 0, ורגש חיובי שהיה מיוצג על ידי הספרה 4 המרנו אותו לספרה 1. **רגש ניטרלי לא היה קיים בדאטה סט לכן לא ביצענו המרה עבורו.**

לאחר כל התהליך, הסרנו מה-dataset רשומות שנשארו ריקות, כלומר שמספר הטוקנים שהכילו היה שווה לאפס, ולכן מ-100 אלף רשומות ירדנו ל-99,505 רשומות.

דוגמא ל-dataset לאחר תהליך ה-pre-processing:

	text	target
0	[love, guy, best]	1
1	[meet, one, besti, tonight, cant, wait, girl, ...	1
2	[thank, twitter, add, sunisa, got, meet, hin, ...	1
3	[sick, realli, cheap, hurt, much, eat, real, f...	1
4	[effect, everyon]	1
...	...	...
99500	[insomnia, prob, slept, hrs, woke, couldnt, ba...	0
99501	[mintu, late, meet, start, how'd, know, go, la...	0
99502	[super, excit, tweet, event, happen, way, foll...	0
99503	[want, anoth, day, much, today, got, quot, new...	0
99504	[jack, umbrella, cake]	0

C. Data exploration – בגרפים ניתן לראות את התפלגות הרשומות בסנטימנטים השונים. ניתן לראות כי ה-dataset מאוזן.




```
Negative distribution:  0.5003869152303905
Positive distribution:  0.4996130847696096
```

עשרת המונחים הכי נפוצים עבור כל רגש:

Top 10 negative terms

Negative term	Frequency
not	4756
work	4391
get	3873
day	3436
back	2648
today	2611
like	2540
miss	2502
want	2482
feel	2235

Top 10 positive terms

Positive term	Frequency
good	4075
love	3588
day	3368
get	3111
thank	3072
quot	2933
like	2567
not	2157
time	2147
go	2059

ניתן לראות כי ישנה חפיפה בין הרגשות במונחים מסוימים למשל: "day". לכן, אנחנו מניחים שבמקרה ויהיה ציוץ אשר יכיל מילים המשותפות לשני הקלאסים יהיה למודל יותר בעייתי לסווג אותן.

Data Splitting - Train & Test sets – חילקנו את הדאטה ל50% סט אימון ו-50% סט מבחן. בחרנו לעשות זאת משום שאנחנו מניחים ש-50 אלף רשומות מספיקות לנו על מנת לאמן את המודל שכן מדובר בכמות די גדולה. בנוסף מכיוון שגם סט המבחן גדול נדע האם ההכללה שלנו טובה או לא.

שאלה 2

A. השתמשנו בשלושה מודלים: מודל פשוט שבוחר את הסיווג על פי הסיווג שמופיע הכי הרבה פעמים בסט האימון, מודל נוסף מסוג SVM ומודל למידה עמוקה מסוג LSTM המכיל שכבת embedding.

Dummy classifier – מסווג את הדאטה על פי הלייבל הנפוץ ביותר בסט האימון. נשתמש בו כ-baseline לשאר המודלים על מנת לראות שהמודלים שאימנו טובים יותר ממודל פשטני.

SVM+TF-IDF – אימנו מודל מסוג SVM והשתמשנו בשיטת TF-IDF על מנת להמיר את הדאטה שלנו לייצוג מספרי. במהלך האימון השתמשנו ב-grid search על מנת למצוא את הפרמטרים האופטימליים – הפרמטרים שנבדקו הם: פרמטר הרגולריזציה C עבור ה-SVM, ועבור שיטת tf-idf בדקנו את טווח ה-ngram. **מכיוון שסביבת העבודה שלנו מוגבלת מבחינת יכולות חישוביות ועל מנת למנוע זמני ריצה ארוכים אלו הפרמטרים היחידים שבדקנו**. בסביבה אולטימטיבית היינו יכולים לבדוק פרמטרים נוספים. במהלך הרצת ה-grid search השתמשו ב- 10 fold cross validation על מנת שנוכל לקבל נקודת מבט טובה לגבי יכולת ההכללה של המודל שלנו. התוצאה והפרמטרים הטובים ביותר שקיבלנו הינם:

```
Best score:  0.7720695196238426
Best params:  {'clf__C': 1, 'clf__loss': 'hinge', 'clf__random_state': 0, 'vect__ngram_range': (1, 2)}
```

* פונקציית Loss קבועה, ו-random state קבוע על מנת לשחזר את התוצאות שקיבלנו.

LSTM + Embedding - תחילה המרנו את הדאטה לווקטורים של אינדקסים באמצעות האובייקט Tokenizer, כאשר כל אינדקס מייצג מילה. לאחר מכן, על מנת שנוכל להכניס את הווקטורים למודל אנחנו צריכים שכולם יהיו באותו אורך, לכן בדקנו את האורך של הציוץ הארוך ביותר ולפיו ריפדנו את שאר הציוצים באפסים בהתחלה (כפי שראינו בהרצאה שריפוד אפסים בהתחלה יעיל יותר מריפוד בסוף משום שככה כמעט ואין השפעה לאפסים) – נציין שהאינדקסים של המילים מתחילים מהספרה 1. הדגמה:

`start twitter need friend` -> [73,29,33,63] -> [0,0,0,0,.....,73,29,33,63]

אחרי שהכנו את הדאטה עבור המודל, בנינו מודל המורכב משלוש שכבות – שכבת embedding שהפלט שלה בגודל 128, לאחר מכן שכבת LSTM בעלת 128 נוירונים בעלת פרמטר dropout על מנת לבצע רגולרזיציה ולהפחית אוברפיט, ובסוף שכבת Fully connected בעלת נוירון יחיד שאומר מה הסנטימנט של הציוץ.

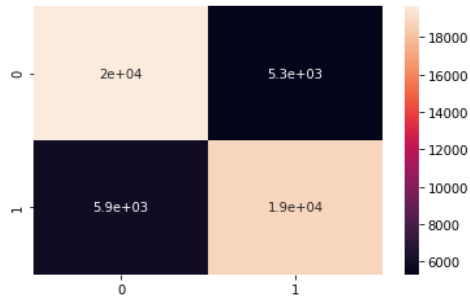
בדומה למודל הקודם, השתמשנו ב-grid search על מנת למצוא את הפרמטרים האופטימליים – הפרמטרים שנבדקו הם: dropout rate. **גם כאן בדקנו פרמטר יחיד בגלל המגבלה של היכולת החישובית**. התוצאה והפרמטרים הטובים ביותר שקיבלנו הינם:

```
Best score:  0.7317494750022888
Best params:  {'batch_size': 256, 'dropout_rate': 0.4, 'epochs': 5}
```

* גודל ה-batch נתון, וגם מספר ה-epochs. היינו יכולים לבדוק אותם אבל זה לוקח זמן רב. לכן במידה והיו לנו את היכולות היינו מכוונים גם את הפרמטרים הללו. בנוסף היה ניתן לכוון גם את מבנה הרשת באמצעות grid search – מספר נוירונים בשכבות השונות, מספר שכבות, מספר הממדים של הפלט בשכבת embedding ועוד, **אך שוב מדובר בתהליך יקר חישובית ולכן בדקנו ידנית אפשריות שונות עד שקיבלנו תוצאות די טובות.**

B. Evaluation – לאחר שאימנו את המודלים בדקנו אותם על סט המבחן. עבור מודל SVM השתמשנו על סט המבחן ב-tf-idf שאימנו על סט האימון, ובעבור מודל LSTM השתמשנו באותו תהליך מקדים של המרה לאינדקסים וריפוד באפסים.

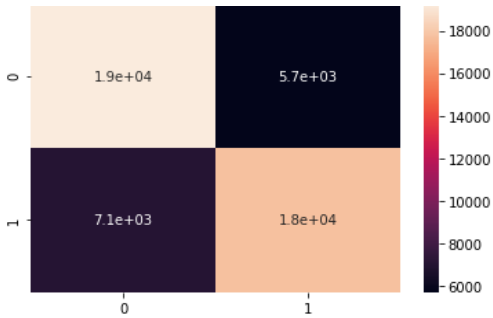
עבור מודל SVM+TF-IDF קיבלנו את התוצאות הבאות:



	precision	recall	f1-score	support
0	0.77	0.79	0.78	24962
1	0.78	0.76	0.77	24791
accuracy			0.77	49753
macro avg	0.77	0.77	0.77	49753
weighted avg	0.77	0.77	0.77	49753

77% דיוק, כאשר ניתן לראות ב-confusion matrix שהרוב סווג נכון. אנחנו מניחים שהקושי של המודל היה בסיווג ציוצים המכילים מילים דומות לשני הסנטימנטים, או מילים שלא הופיעו באימון.

עבור מודל LSTM + Embedding קיבלנו את התוצאות הבאות:



	precision	recall	f1-score	support
0	0.73	0.77	0.75	24962
1	0.76	0.72	0.74	24791
accuracy			0.75	49753
macro avg	0.75	0.74	0.74	49753
weighted avg	0.75	0.75	0.74	49753

75% דיוק, כאשר ניתן לראות ב-confusion matrix כמו במודל הקודם שהרוב סווג נכון. אנחנו מניחים שהקשיים של המודל היו דומים למודל הקודם, ובנוסף אם היינו משתמשים בשכבת embedding מאומנת היינו יכולים לשפר את הדיוק. כמו כן, הגדלת מספר epochs, הגדלת הרשת, שינוי batch size ועוד אפשרויות הקשורות למבנה הרשת גם היו יכולים להועיל.

C. אי אפשר להשתמש במודל מעבודה 2 ללא אימון מחדש. הסבר: מכיוון שה-dataset שהשתמשנו בעבודה הקודמת, שהכיל ביקורות של סרטים מאתר IMDB, שונה מהדאטה סט שאנחנו משתמשים בעבודה הנוכחית שמכיל ציוצים, אוצר המילים יהיה שונה ולא נוכל לקבל תוצאות טובות המתאימות לדאטה שלנו. ציוצים מכילים למשל: hashtags, urls, emojis, @-usernames וכו'. במידה והיינו רוצים להשתמש במודל שאומן לפני, היינו צריכים לכוון את המודל המאומן למודל שלנו.

דוגמה למילים נפוצות מהעבודה הקודמת:

Top 10 negative terms

Negative term	Frequency
movi	57836
film	44483
one	26853
like	24345
make	16059
even	15266
time	15130
get	15101
good	14775
watch	14769

Top 10 positive terms

Positive term	Frequency
film	50712
movi	44624
one	28168
like	20478
time	16555
good	15198
see	15078
stori	14147
charact	13959
make	13758

לעומת מילים נפוצות בעבודה נוכחית:

Top 10 positive terms

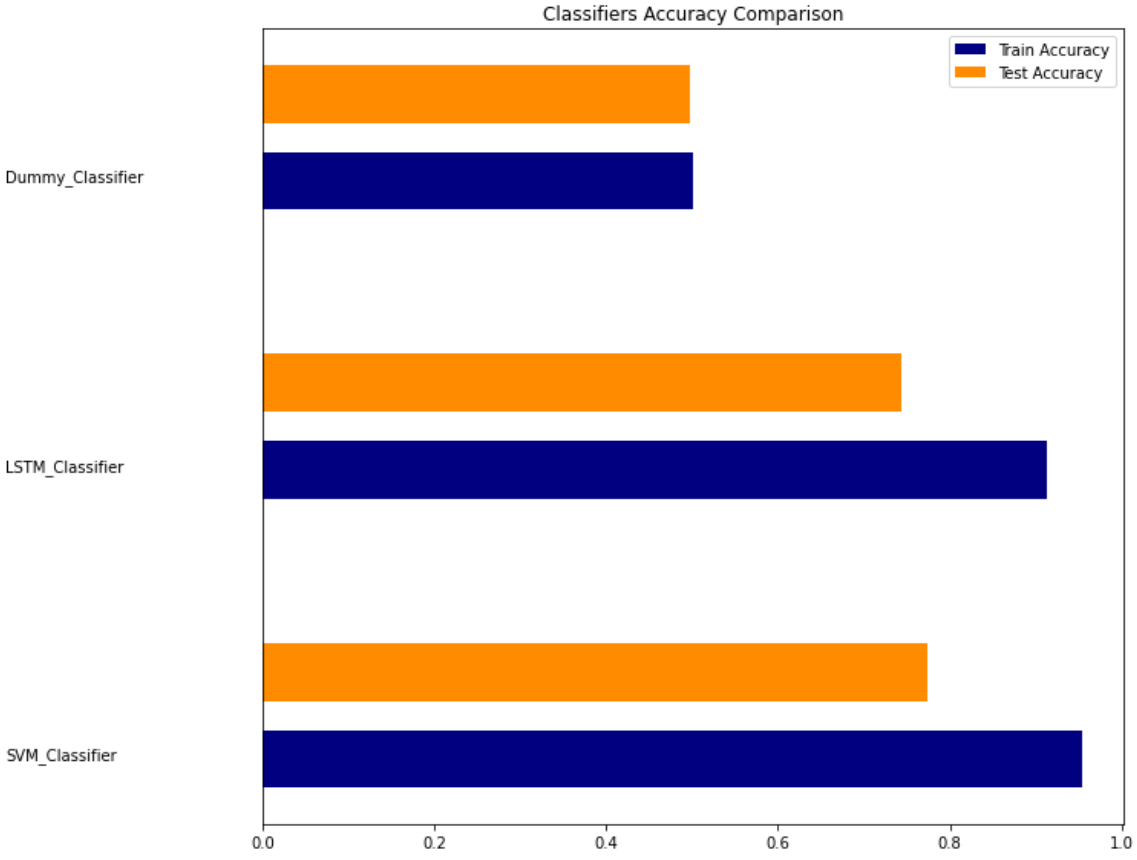
Positive term	Frequency
good	4075
love	3588
day	3368
get	3111
thank	3072
quot	2933
like	2567
not	2157
time	2147
go	2059

Top 10 negative terms

Negative term	Frequency
not	4756
work	4391
get	3873
day	3436
back	2648
today	2611
like	2540
miss	2502
want	2482
feel	2235

למרות שישנן מילים נפוצות כמו get, like שמשותפות לשניהם, עדיין ניתן לראות הבדלים במילים של הדומיין אשר יותר משפיעות כמו movi, film לעומת qout למשל.

D. Comparison + Discussion – השוונו בין המודלים השונים על מנת לראות האם טובים יותר ממודל פשטני ובנוסף האם יש הבדל מהותי בין שני המודלים שאימנו. התוצאות מוצגות בגרף הבא:



כמו שניתן לראות, שני המודלים קיבלו תוצאות טובות יותר מאשר המודל הפשטני. בנוסף, מודל ה-SVM קיבל תוצאות קצת יותר טובות מאשר מודל ה-LSTM אבל אנחנו מניחים כי אם נאמן את המודל בעוד

epochs ונשתמש בשכבת embedding מאומנת נוכל לשפר את ביצועי המודל. אך כמו שכבר אמרנו היכולות החישוביות שלנו מוגבלות ולכן הסתפקנו בתוצאות הללו.

שאלה 3

A. Collect Twitter Data - איסוף הציוצים שבו התמקדנו היה בנושא ג'ו ביידן. לצורך כך, הרצנו במשך שבוע החל מה6.2 עד ה12.2 בין 18:00 - 19:00 בכל יום את איסוף הנתונים. חיפשנו ציוצים באנגלית (משום שסט האימון גם הוא מכיל מילים באנגלית בלבד) אשר מכילים את המילים Joe Biden ונמנענו מציוצים חוזרים RT. בסוף כל יום שמרנו את הציוצים לקובץ JSON.

Extract Tweets from JSON to CSV – חלק זה הוא שלב מקדים לטעינת הנתונים על מנת שנוכל להגיש את הקובץ בצורה נוחה – לכן הוא בהערה. טענו את קבצי ה-json עבור כל יום, ולקחנו רק את הציוצים בשעות שאיתן עבדנו 18:00-19:00. מקבצי ה-JSON לקחנו את השדה של ה-text שהכיל את הציוץ ואת השדה של date שהכיל את התאריך והשעה. הכנסנו את זה לתוך dataframe, ומשום שאספנו יותר מ-15,000 ציוצים היינו צריכים לסנן חלק מהם. לכן, תחילה חישבנו את ההתפלגות של כל הציוצים על פני הימים – כלומר, מה אחוז הציוצים שהיו ביום נתון מתוך סך כל הציוצים. לאחר מכן, משום שסיווגנו ציוצים גם באמצעות אימוג'ים אז כדי לא לאבד מידע קודם אספנו את כל הציוצים שהכילו אימוג'ים, ולאחר מכן לפי מספר הציוצים שנשאר עבור כל יום השלמנו משאר הדאטה סט עד שהגענו ל-15,000 ציוצים.

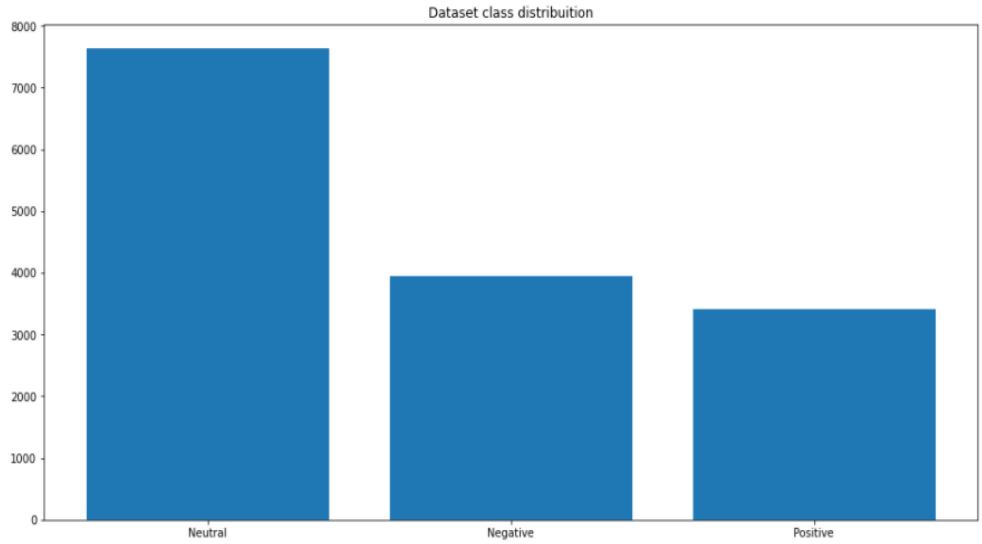
Import Tweets to Dataframe – טענו את המידע מהקובץ CSV שיצרנו – על מנת שבהגשה נוכל לעבוד על הקובץ.

B. Pre processing - לאחר איסוף ה-data ביצענו Preprocessing על הציוצים שנאספו לפי התהליך שהתבצע בשאלה 1.

C. Sentiment Classification - על מנת לתייג את הציוצים, השתמשנו במילון של אימוג'ים ומשום שמדובר בציוצים אנשים אוהבים להציג רגשות באמצעות חייכנים לכן הנחנו שזה מתאים. בנוסף, השתמשנו במאגר מילים חיוביות ושליליות מתוך Minqing Hu and Bing Liu. Mining and Summarizing Customer Reviews - בחרנו במאגר זה לאחר שראינו כי במאמרים אקדמאיים רבים נמצא שימוש במילונים אלה עבור תיוגים של מאגרי מידע דומים המתעסקים בציוצים מטוויטר ובפוליטיקה בפרט והניבו תוצאות טובות.

לאחר שיצרנו את המילונים עבור האימוג'ים ועבור המילים החיוביות והמילים השליליות, עברנו על כל טקסט של ציוץ ובדקנו תחילה האם מכיל אימוג'י – אם כן בדקנו האם שייך לאימוג'י חיובי או שלילי ולפי זה סיווגנו. בשלב זה הצלחנו לתייג 23 ציוצים. לאחר מכן, עבור ציוצים שלא הכילו אימוג'ים חישבנו את כמות המילים החיוביות שיש בציוץ מול כמות המילים השליליות. בנוסף, התייחסנו לכמות המילים החיוביות/שליליות ביחס לכמות המילים בציוץ וכך תייגנו כל ציוץ אם הוא Positive, Negative, Neutral. לאחר שלב זה הצלחנו לתייג 7357 ציוצים. על מנת לבדוק שהתיוג אכן הגיוני עברנו ידנית על הציוצים עם אימוגי'ם (23) על מנת לראות אם אכן שליליים או חיובים ועל עוד (50) ציוצים בעלי מילים חיוביות ושליליות. ראינו שהתיוג מדוייק ברובו והסקנו כי זאת כמות מספקת להמשיך לשלב הבא. נציין כי הציוצים שלא הצלחנו לתייג קיבלו סנטימנט ניטרלי.

Exploration – בשלב זה עבור כל סנטימנט חישבנו את התפלגות הציוצים ואת עשרת המונחים הנפוצים ביותר. Positive tweets = 3408, Negative Tweets = 3949, Neutral Tweets = 7643.



המילים הנפוצות עבור כל סנטימנט:

Top 10 neutral terms

Neutral term	Frequency
biden	1928
joe	1749
presid	557
not	535
peopl	314
get	284
amp	266
need	239
one	234
trump	232

Top 10 negative terms

Negative term	Frequency
biden	1033
joe	890
pleas	318
not	297
presid	286
peopl	247
lie	195
impeach	181
amp	175
kill	174

Top 10 positive terms

Positive term	Frequency
biden	2930
joe	2671
presid	1000
trump	872
not	835
like	578
peopl	499
get	428
amp	394
need	359

ניתן לראות כי המילים דומות בין הסנטימנטים השונים ולכן אנחנו מניחים שלמודל יהיה קשה לסווג ביניהן. **בנוסף בהשוואה לסט האימון** (לסט מקאגל) ניתן לראות שחלק מהמילים הנפוצות שונות, וזאת משום שבאיסוף שלנו התמקדנו בדומיין בודד ולכן אולי הסיווג לא יהיה מדויק כל כך. אך מכיוון שמדובר בשני דאטה סטים של ציוצים, אנחנו מניחים ששאר המילים מותאמות לציוצים וההתבטאות תהיה דומה.

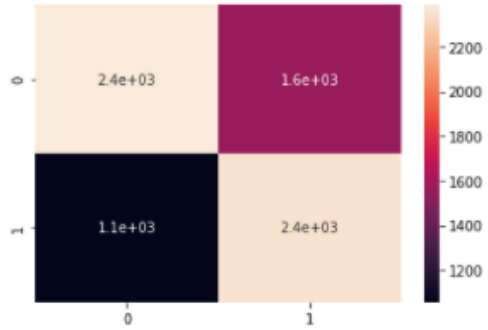
שאלה 4

A. השתמשנו במודל משאלה 2 שאומן על הדאטה סט מקאגל על מנת לבצע "בדיקת שפיות" לראות שהמודל אכן מתאים לדאטה שלנו. לכן, על מנת לבדוק את המודל, הסרנו את המילים הניטרליות מהדאטה, ובדקנו את המודל רק על הציוצים אשר הצלחנו לסווג בשאלה הקודמת.

B. חישבנו את הדיוק על הדאטה שאספנו וקיבלנו כי 64% מהציוצים סווגו נכון. עצם העובדה, שהדומיין לא היה זהה אך בשני המקרים היה מדובר על ציוצים מטויטר, זה נראה לנו הגיוני לקבל דיוק נמוך יותר מהדיוק על הדאטה של קאגל שעליו המודל אומן (קיבלנו 77% דיוק) אך עדיין הדיוק יותר טוב ממודל אקראי (50%).

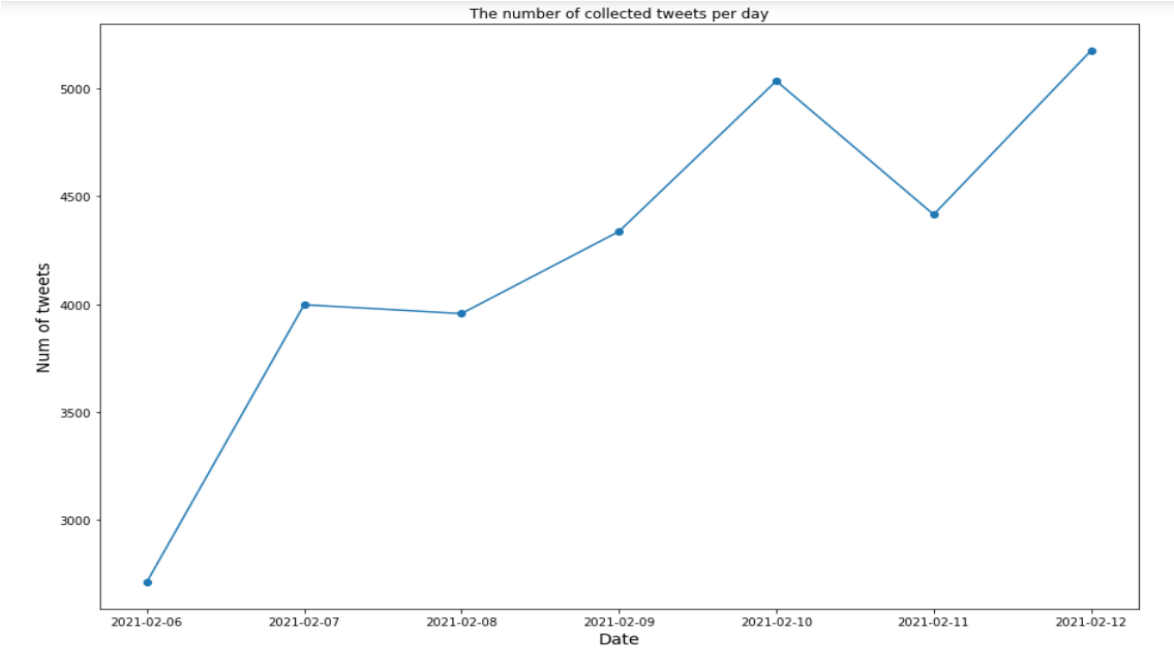
	precision	recall	f1-score	support
0	0.69	0.60	0.65	3949
1	0.60	0.69	0.64	3408
accuracy			0.64	7357
macro avg	0.65	0.65	0.64	7357
weighted avg	0.65	0.64	0.64	7357

==========



שאלה 5

A. להלן גרף המציג את כמות הציוצים שקשורים לג'ו ביידן בכל תאריך שאספנו בין 18:00-19:00:



נשים לב כי לאורך רוב הימים בהם נאספו הציוצים חלה עליה בכמות הציוצים עבור כל יום פרט ל- 08 בפברואר ול- 11 בפברואר שם חלה ירידה ביחס ליום הקודם.

ניתן לראות קשר בין העולם האמיתי לבין מדדי הציוצים עבור כל יום -

ב-7 בפברואר התרחש הסופרבול, מספר הציוצים על ג'ו ביידן החל לעלות מאחר והופיע עם ראיון ראשון בטלוויזיה כנשיא לפני תחילת הסופרבול.

ב-8 בפברואר חלה ירידה קלה בכמות הציוצים, מלבד הכרזה על מותו של ג'ורג' שולץ ושיחה עם ראש ממשלת הודו לא התרחש מאורע יוצא דופן.

ב-9 בפברואר התקיימה ההשבעה של סגנית הנשיא קמלה האריס, חלה עליה קלה בכמות הציוצים בהקשר של ג'ו ביידן.

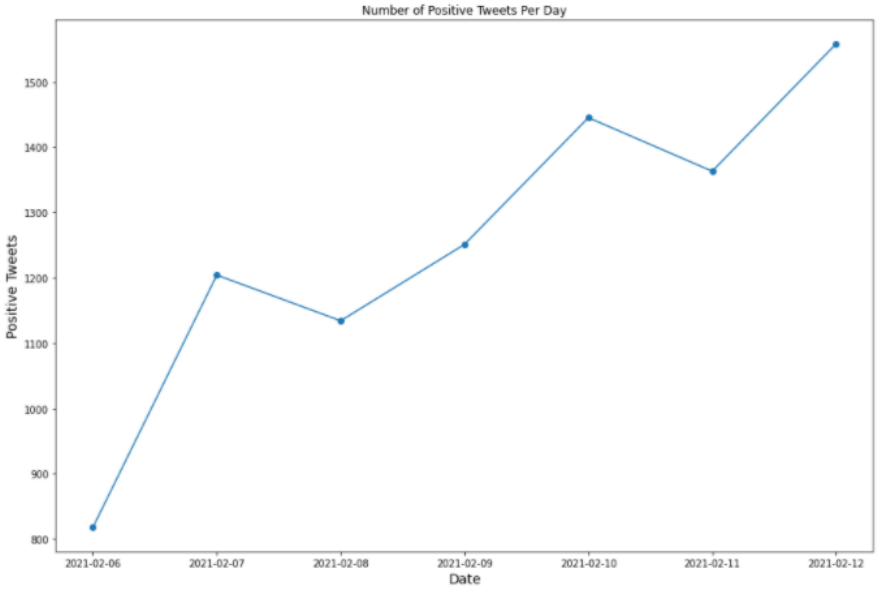
ב-10 בפברואר ג'ו ביידן נאם בביקורו בפנטגון ולכן ניתן לראות כי מספר הציוצים באותו היום היה גבוה.

ב-12 בפברואר ג'ו ביידן אישר ל25 אלף מבקשי מקלט ממקסיקו להכנס לארה"ב והכריז על סגירת הכלא האמריקאי בקובה עד סוף כהונתו.

בימים בהם קרו אירועים יותר משמעותיים ויוצאי דופן חלה עליה בכמות הציוצים ולהפך.

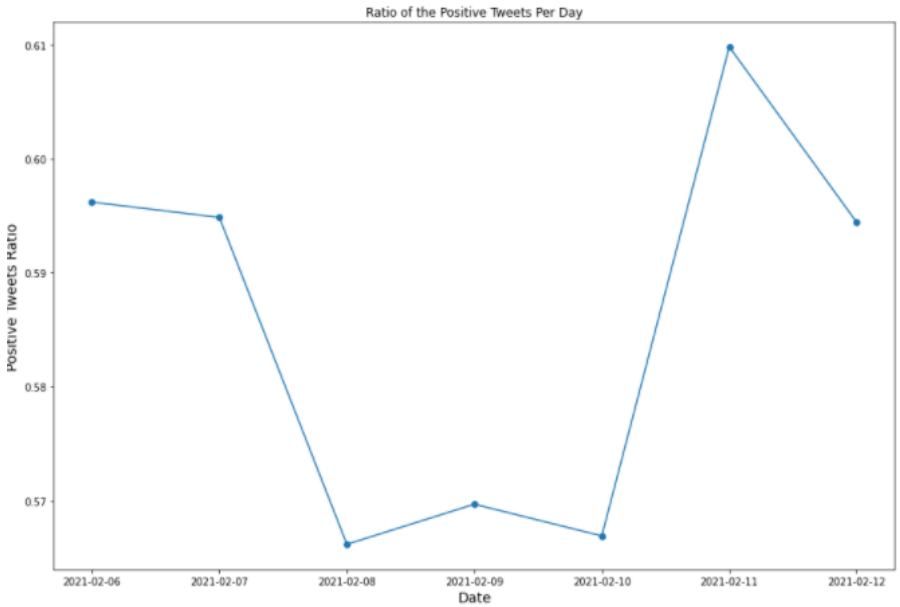
B. השתמשנו במודל סיווג על מנת לסווג את כל 15,000 הציוצים שאספנו (לעומת שאלה 4 שם סיווגנו רק את הציוצים שהצלחנו לסווג בשאלה 3). לאחר מכן, ספרנו את כמות הציוצים החיוביים שהמודל סיווג, ואת החלק היחסי של הציוצים החיוביים מסך כל הציוצים בכל יום.

גרף המציג את כמות הציוצים החיוביים בכל יום:



הגרף הוא שווה ערך באופן ישיר לגרף מהסעיף הקודם המציג את מספר הציוצים, כלומר ככל שמספר הציוצים עלה גם מספר הציוצים החיוביים עלה. אך בגרף הבא המציג את יחס החיוביים ניתן לראות הבדל.

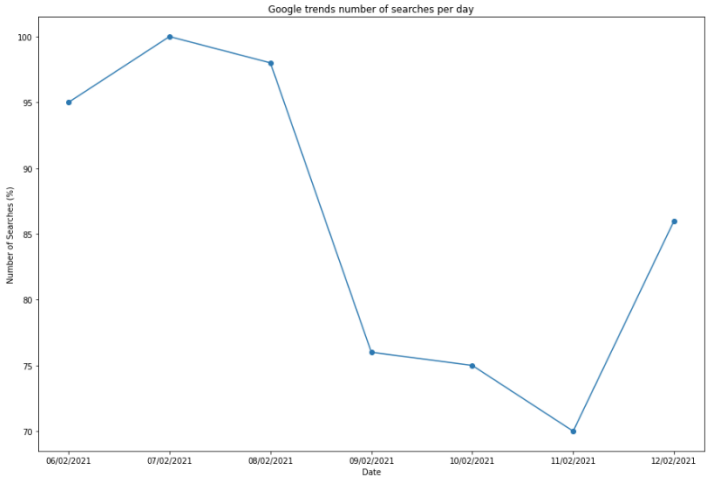
גרף המציג את היחס של הציוצים החיוביים בכל יום ביחס לכל הציוצים שהיו באותו היום:



ניתן לראות מהגרף שרוב הזמן הציוצים היו חיוביים – ניתן להניח שרוב האנשים שמצייצים הם למשל עם רגשות חיוביים כלפי הנושא עליו הם דנים. בנוסף, בתאריכים 8.2 ו- 11.2 אנחנו רואים שינוי מגמה חד בדעות החיוביות, יכול להיות שנובע מדברים שנאמרו בתאריכים הללו אשר השפיעו על דעתם של האנשים, כאשר ב-8.2 הדברים גררו שיח יותר מאוזן, וב-11.2 הדברים גררו שיח יותר חיובי. בנוסף, בסעיף הקודם ראינו שישנה ירידה בכמות הציוצים בתאריכים הללו, ולכן ניתן גם להסיק שמכיוון שהיו פחות ציוצים השינוי במגמתיות הורגש יותר. כמו כן ב- 7.2 וב-12.2 ניתן לראות רגשות חיוביים יותר

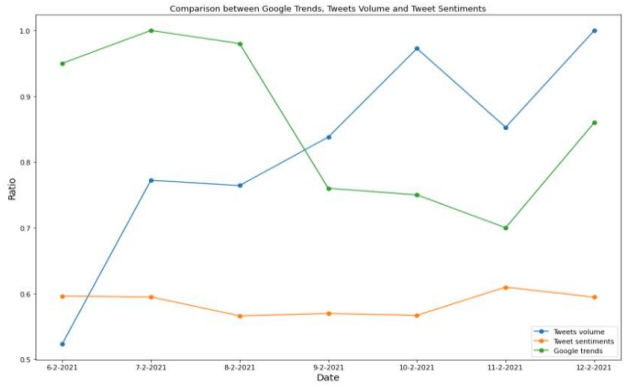
כאשר אפשר לייחס את זה לאירוע הסופרבול ולאישור כניסת מבקשי מקלט, שכן זהו נושא רגיש בארה"ב שדנו עליו רבות בתקופת כהונתו של הנשיא הקודם. ולעומת זאת ב-9.2 וב-10.2 התגובות היו פחות חיוביות ויותר מאוזנות וזה בהמשך לאירועים של נאומי הנשיא בפנטגון והשבעתה של קמלה האריס.

C. השתמשנו בגוגל טרנדס על מנת לקבל את מספר החיפושים היחסי של הנושא שבחרנו – ג'ו ביידן, על פני הימים בהם אספנו את הציוצים.

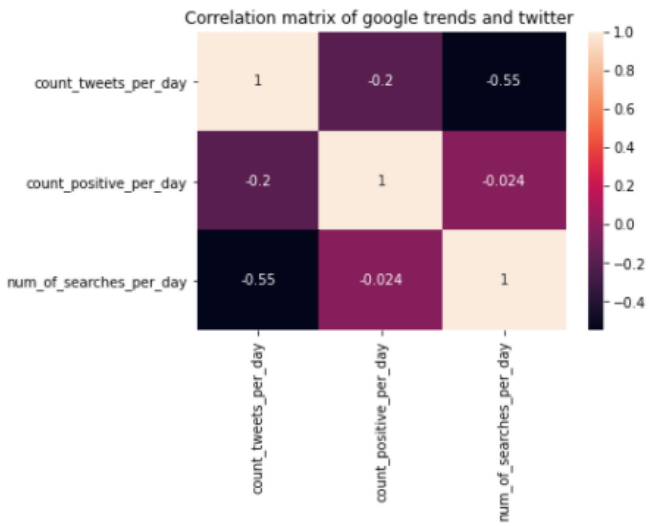


ניתן לראות מהגרף שרמת העניין בשבוע הזה הייתה גבוה באופן יחסי – 76 בממוצע. בנוסף, תחילה העניין היה גבוה לאחר מכן ירד (באופן יחסי) ואז חלה עליה נוספת.

D. השווינו בין הגרפים שיצרנו וחישבנו את מתאם פירסון על מנת לראות האם יש קשר בין החלקים השונים. במידה ונקבל ערך קרוב ל-1 נניח שקיים קשר חיובי חזק, במידה ונקבל ערך קרוב ל- 1- נניח קשר שלילי חזק, אחרת נניח שאין קשר. התוצאות שקיבלנו:



	date	count_tweets_per_day	count_positive_per_day	num_of_searches_per_day
0	6-2-2021	1372	0.596210	95
1	7-2-2021	2024	0.594862	100
2	8-2-2021	2003	0.566151	98
3	9-2-2021	2196	0.569672	76
4	10-2-2021	2549	0.566889	75
5	11-2-2021	2235	0.609843	70
6	12-2-2021	2621	0.594430	86



קשה לראות קשרים חזקים מהתוצאות שקיבלנו. נראה כי ישנו קשר שלילי חלש (-0.57) בין כמות החיפושים בגוגל לבין מספר הציוצים ביום וזה בניגוד לאינטואיציה שאם מחפשים יותר גם יצייצו יותר. ההנחה שלנו שטווח הזמנים שחיפשנו בו היה קצר מדי ובנוסף השעה ביום השפיעה על התוצאות – מספר החיפושים ביום מתייחס לכל היום ולא לשעה ספציפית, לכן יכול להיות שבחרנו שעה שמייצגת בצורה פחותה את המדגם האמיתי וזה הגיוני ביחס להפרשי שעות בינינו לבינם. בנוגע לקשר בין היחס של החיובים ביום לבין מספר הציוצים ביום ולמספר החיפושים נראה שאין קשר כלל (ערך קרוב לאפס). אפשר להניח שגם ערכים שליליים וגם ערכים חיוביים מושפעים מהנושא והדיון וכמות הציוצים או כמות החיפושים לא קשורה לאם זה רגש חיובי או שלילי (אך כמו שראינו בגרף של הסנטימנט ישנם יותר רגשות חיוביים). כמו כן, הנושא שבחרנו מכיל בתוכו המון דעות חיוביות ושליליות, לכן זה די מאוזן בין תומכים ומתנגדים כאשר ישנה נטייה לכיוון התומכים (שבסופו של דבר גם מייצגים את דעת הרוב בבחירות). בנוסף האירועים שחלו בימים הללו היו אירועים פוליטיים אשר מציתים בליבם של האנשים דעות שונות ורצון להתבטא ולכן ניתן להבין את מגוון הדעות. כשיפור להמשך, אפשר לבחור טווח ימים רחב יותר בשעות רחבות יותר או שעות שמתאימות לאוכלוסייה המקומית שממנה אספנו את הציוצים.