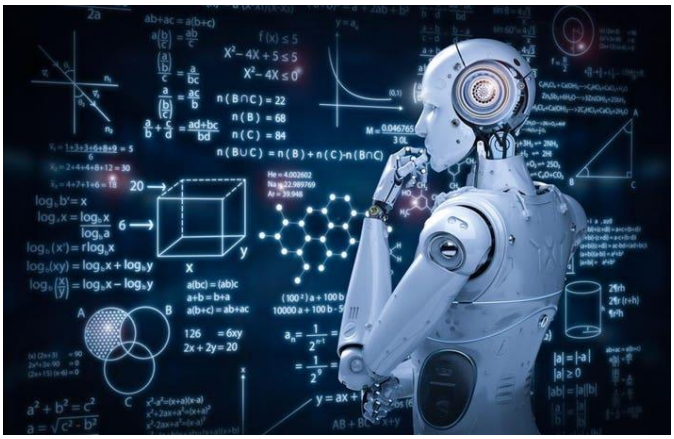
אוניברסיטת בן-גוריון בנגב
Ben-Gurion University of the Negev

המחלקה להנדסת תעשייה וניהול

קורס למידת מכונה 364-1-1811

פרויקט הקורס – חלק א'

08/02/2024



קבוצה 16

205404965

315474205

תוכן עניינים

Data collection and sensing

Data collection הינו סט נתונים המייצג את ה-"Real world" שאנו רוצים ללמוד עליו. בפרויקט זה ה-data collection הוא הודעות ברשתות חברתיות. הבסיס למערכת למידת מכונה מוצלחת הינו סט נתונים טוב. סט הנתונים מורכב מ-entities או samples מאותו ה-domain. נעדיף sample שלם (כלומר ללא ערכי null או ערכים חלקיים, כמו חצי תמונה כאשר מנתחים תמונות). Data collection מוצלח ייצג את כלל האוכלוסייה הנבדקת ויכסה באופן כמעט הרמטי את המקרים השונים (דוגמיות). כמו כן, סט הנתונים צריך להגיע מ-source אמין, להיות בעל מספר שגיאות ו-confusions מזערי, ובעדיפות שלכל sample יהיה label על מנת שנוכל להקצות samples אל classes.

בוצעו שני סוגי **sensing** על ה-data, גם static וגם dynamic. ישנם נתונים שיישארו זהים ולא ישתנו עם הזמן (למשל שדה email שהוא שדה שמוזן באופן חד פעמי עבור המשתמש) ולכן זהו sensing מסוג static, וישנם נתונים שמשתנים לאורך הזמן, כלומר נראה הבדל בין מספר samples במשתנה עבור אותה ה-entity (למשל שדה date_of_new_follower, שישתנה כאשר יתווספו עוקבים חדשים למשתמש) ולכן זהו sensing מסוג dynamic. שני סוגי ה-sensing בוצעו על ה-data, ולכן אין אפשרות לבצע sensing מסוג נוסף.

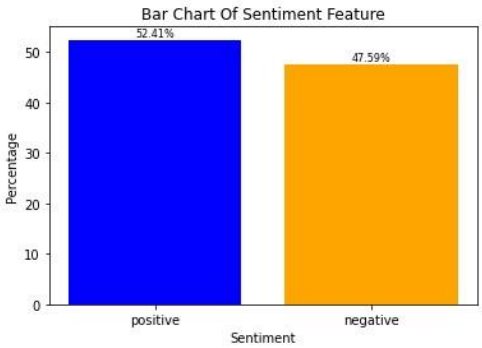
קטגוריית משימת הלמידה היא למידה מונחית (Supervised learning), מכיוון שה-labels ב- training set ידועים. סוג משימת הלמידה הינו Binary classification מפני שיש שני סוגי classes (positive/negative). ניתן להשתמש בנתונים כדי לבצע גם משימת למידה מסוג Regression, שבאמצעותה אנו מנסים לזהות קשרים בין Independent numeric feature(s) לבין ה-target feature.

Dataset creation

סעיף 1 – Exploratory data analysis

ראשית, בחנו וניתחנו את משתנה המטרה. לאחר מכן ניתחנו את יתר המשתנים. המשתנה היחיד שבחרנו שלא לנתח הוא textID, מכיוון שהוא ניתן באופן מלאכותי על ידי המערכת ואינו נשלט על ידי כותב ההודעה. לבסוף, בחנו קשרים בין המשתנים המסבירים.

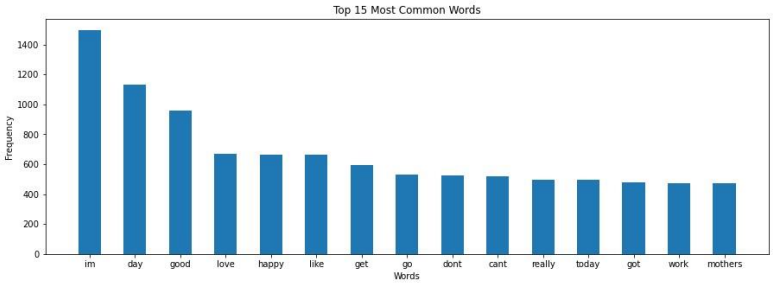
sentiment (1.1)



משתנה זה הינו משתנה קטגוריאלי המקבל את הערכים positive או negative. לאחר ניתוח גרפי של המשתנה עולה כי על פי הגדרה, סט הנתונים אינו מאוזן, אך בחרנו להתייחס אליו כאל מאוזן, מכיוון שאחוז ה-samples בעלי הערך positive כמעט זהה לאחוז ה-samples בעלי הערך negative. משתנה זה הינו משתנה המטרה, וערכיו הם ה-labels המהווים כמרכיב העיקרי במשימת הלימוד (במקרה זה supervised learning).

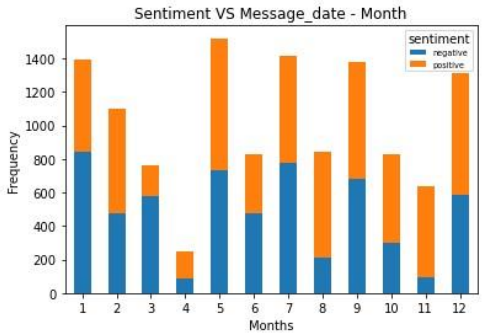
text (1.2)

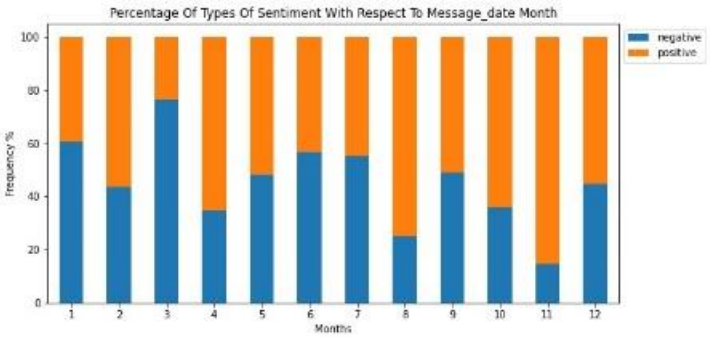
מכיוון שזהו שדה המכיל מלל רב עם שונות גבוהה בין ההודעות, בחרנו לבחון מה הן המילים השכיחות ביותר באופן כללי, בהודעות בעלות סנטימנט חיובי ובהודעות בעלות סנטימנט שלילי באמצעות ענן מילים ("World cloud"). בנוסף, בחרנו לבחון כמה פעמים 15 המילים השכיחות ביותר מופיעות בהודעות לפי הפילוח הנ"ל. לאחר שקיבלנו את הנתונים הללו, בחנו את הקשר ביניהן לבין משתנה המטרה. המסקנה שלנו לגבי משתנה זה הינה כי ישנן מילים שיכולות לסייע בזיהוי הסנטימנט (positive/negative), אך לא תמיד באופן בלעדי מכיוון שישנן מילים שכיחות שמקושרות הן לסנטימנט השלילי והן לסנטימנט החיובי (קישור לתרשימים נוספים). יש לציין כי הוצאנו מילים נפוצות מהניתוח (stop words), על מנת למנוע הטיות הנובעות ממילים אלו שיכולות להופיע בשכיחות גבוהה הן בהודעות המתויגות כחיוביות והן בהודעות המתויגות כשליליות. נרצה להשתמש בקשרים אלו בשלב ה-feature extraction.



message_date (1.3)

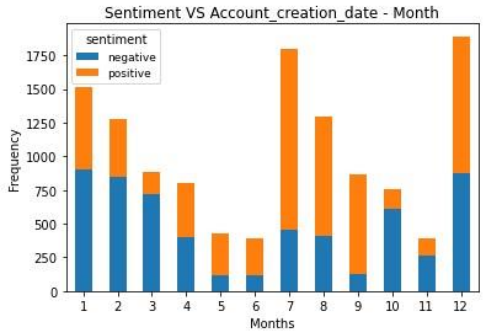
בחרנו לבחון את הקשר בין כתיבת ההודעה בשעה מסוימת ביום (בפילוח לפי חתכי השעות 00:00-07:59, 08:00-15:59, 16:00-23:59), בחודש מסוים ובשנה מסוימת לבין הסנטימנט של ההודעה. זאת מתוך המחשבה שייתכן ואירועים שגרתיים בחיי היום-יום של האנשים יכולים להשפיע על שעת הכתיבה של הודעות בעלות סנטימנט חיובי / שלילי וכן אירועים גדולים יותר, כמו התפרצות מגפת הקורונה, יכולה להשפיע על הסנטימנט של ההודעה בחודשים מסוימים או בשנים מסוימות. את הניתוח ביצענו בשני אופנים- האחד בחינה של יחס ההודעות בעלות סנטימנט חיובי לעומת שלילי בשעה / חודש / שנה מסוימים (באחוזים), והשני בחינה יחסית של כמות ההודעות שנכתבו בשעה / חודש / שנה מסוימים הכולל פילוח לפי הודעות בעלות סנטימנט חיובי וסנטימנט שלילי. המסקנה שלנו היא שישנו קשר בין תאריך ושעת כתיבת ההודעה לסנטימנט וישנו קשר בין תאריך ושעת כתיבת ההודעה לכמות ההודעות הנכתבות (ציפרנו דוגמה של חודש, קישור ליתר התרשימים). נרצה להשתמש בקשרים אלו בשלב ה-feature extraction.

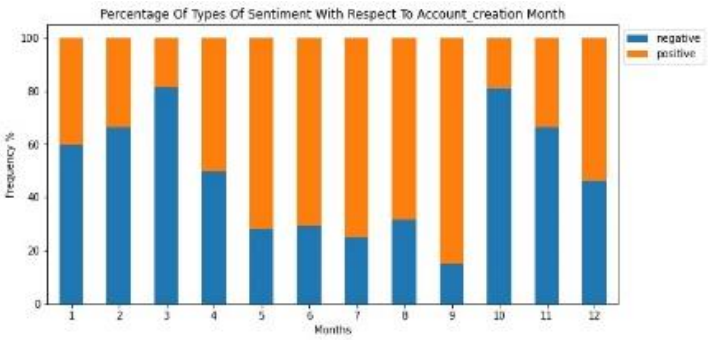




account_creation_date (1.4)

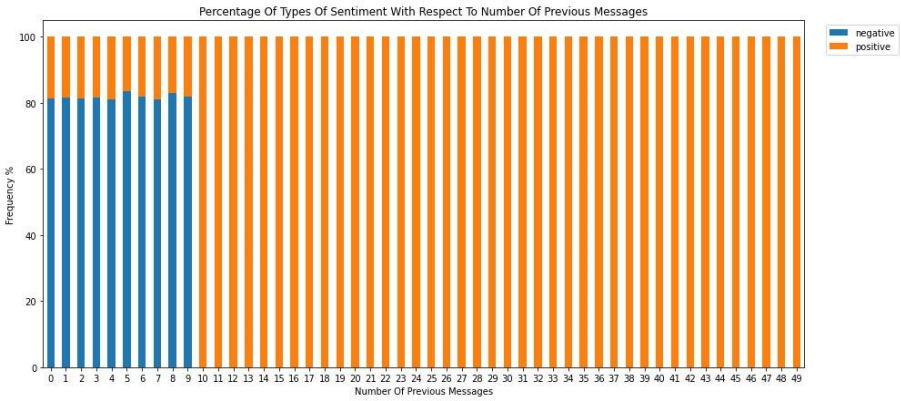
בחרנו לבחון את הקשר בין יצירת המשתמש בשעה מסוימת ביום (בפילוח לפי חתכי השעות 00:00-07:59, 08:00-15:59, 16:00-23:59), בחודש מסוים ובשנה מסוימת לבין הסנטימנט של ההודעה. זאת מתוך המחשבה שייתכן וישנו מניע מהחיים האישיים של האדם או מאירועים מדיניים ועולמיים ליצירת המשתמש, המשפיע על ההודעות שהמשתמש מפרסם. את הניתוח ביצענו בשני אופנים- האחד בחינה של יחס ההודעות בעלות סנטימנט חיובי לעומת שלילי עבור משתמשים שנוצרו בשעה / חודש / שנה מסוימים באחוזים, והשני בחינה יחסית של כמות המשתמשים שנוצרו בשעה / חודש / שנה מסוימים הכולל פילוח לפי הודעות בעלות סנטימנט חיובי וסנטימנט שלילי. המסקנה שלנו היא שישנו קשר בין תאריך פתיחת המשתמש לבין סנטימנט ההודעה, אך אין קשר בין שעת פתיחת המשתמש לסנטימנט. בנוסף, ישנו קשר בין התאריך לכמות המשתמשים שנוצרו באותה התקופה, בפילוח לחודש ולשנה (כאן מוצגת דוגמה של חודש, קישור ליתר התרשימים). נרצה להשתמש בקשרים אלו בשלב ה-feature extraction.





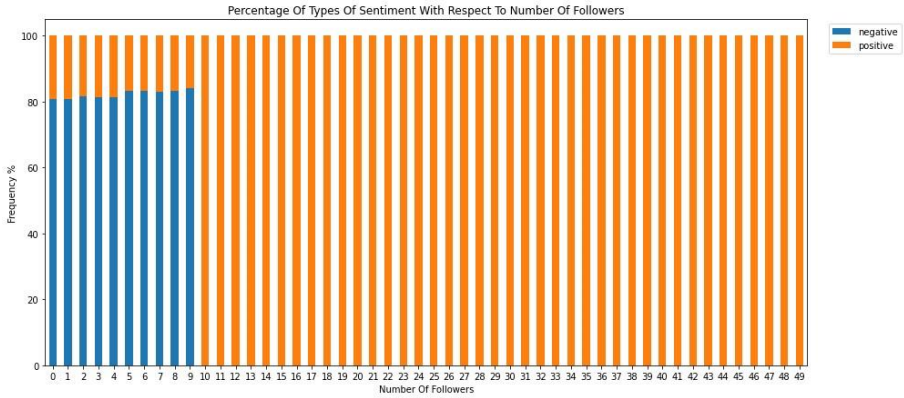
previous_message_dates (1.5)

מכיוון שישנה שונות גבוהה בין הערכים הנתונים בשדה זה בין המשתמשים השונים, בחרנו לעבד את הנתונים ולבחון את היחס בין כמות ההודעות הקודמות שהמשתמש פרסם לסנטימנט של ההודעה. אנו מסיקים כי מספר ההודעות שהמשתמש שלח בעבר יכולה לסייע ב-labeling של הסנטימנט. נרצה להשתמש בקשרים אלו בשלב ה-feature extraction.



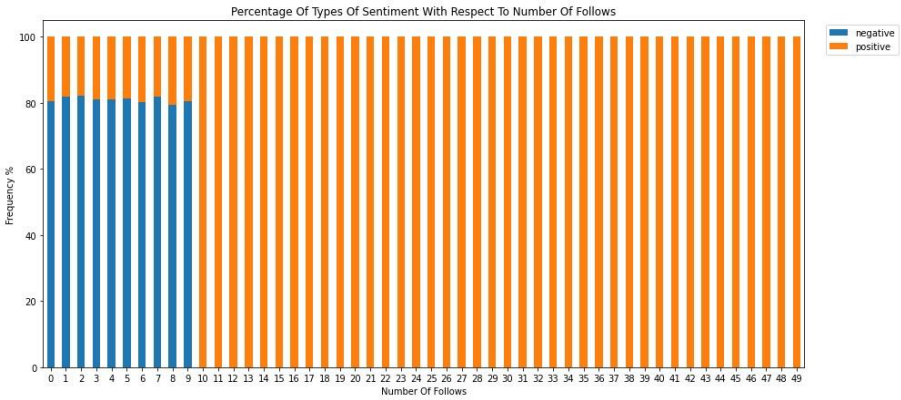
date_of_new_follower (1.6)

מכיוון שישנה שונות גבוהה בין הערכים הנתונים בשדה זה בין המשתמשים השונים, בחרנו לעבד את הנתונים ולבחון את היחס בין כמות העוקבים של המשתמש לסנטימנט של ההודעה. אנו מסיקים כי מספר העוקבים (followers) יכול לסייע ב-labeling של הסנטימנט. נרצה להשתמש בקשרים אלו בשלב ה-feature extraction.



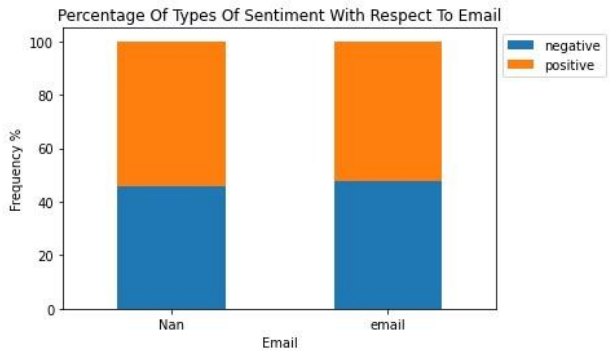
date_of_new_follow (1.7)

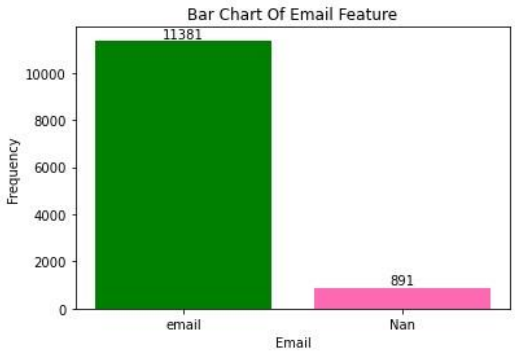
מכיוון שישנה שונות גבוהה בין הערכים הנתונים בשדה זה בין המשתמשים השונים, בחרנו לעבד את הנתונים ולבחון את היחס בין כמות המשתמשים אחריהם כל משתמש עוקב (follow) לסנטימנט של ההודעה. אנו מסיקים כי ה-follow יכול לסייע ב-labeling של הסנטימנט. נרצה להשתמש בקשרים אלו בשלב ה-feature extraction.



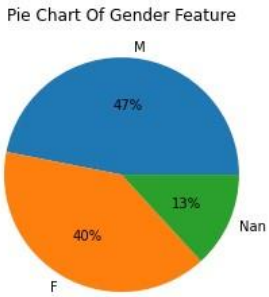
email (1.8)

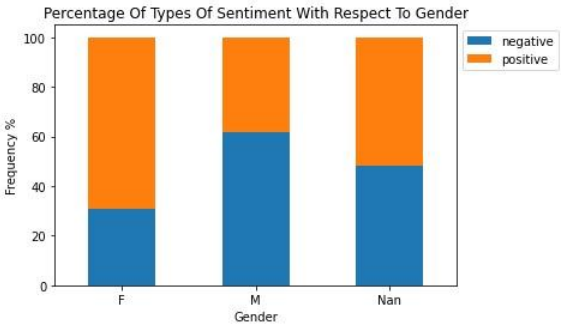
מכיוון שכתובת מייל היא קבועה ומוזנת באופן חד פעמי עם יצירת המשתמש, וכמו כן בעלת שונות גבוהה (כל כתובת מייל הינה ייחודית), בחרנו לנתח את הקשר בין סנטימנט לבין משתמשים שיש לנו מידע אודות כתובת המייל שלהם ומשתמשים שכתובת המייל שלהם אינה נתונה (בעלת ערך nan). עבור רוב ה-samples הקיימים ב-data set קיים ערך שונה מ-nan ואנו מסיקים מהגרפים כי אין קשר בין הימצאות המידע אודות כתובת המייל של המשתמש לסנטימנט של ההודעה (קישור לתרשימים נוספים).





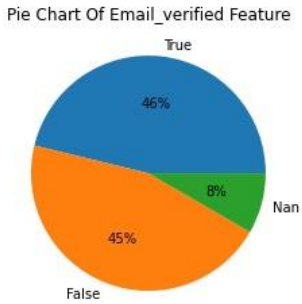
gender (1.9)

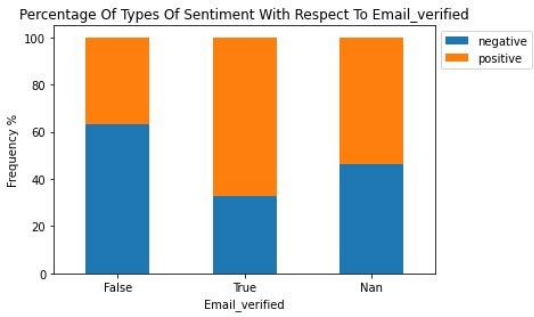




עבור הניתוח של משתנה זה, בחנו את היחס בין נשים, גברים ומשתמשים שהמגדר שלהם אינו נתון (בעל ערך nan). היחס בין ה-samples של גברים ושל נשים כמעט זהה. אחוז ה-samples בהם מגדר המשתמש אינו ידוע קטן ביחס ליתר ה-samples. בנוסף, בחנו את הקשר בין המגדר לסנטימנט, הן מבחינת אחוזים והן מבחינת כמות ההודעות, ואנו מסיקים כי ישנו קשר בין השניים (קישור לתרשימים נוספים).

email_verified (1.10)

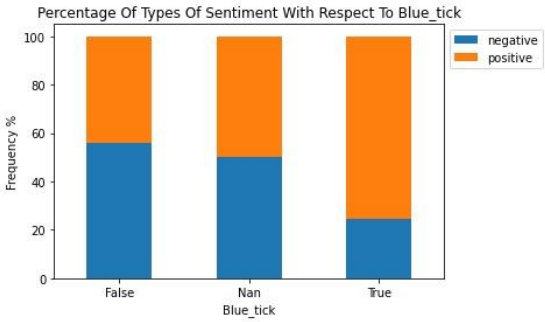


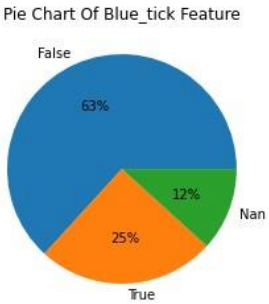


עבור הניתוח של משתנה זה, בחנו את היחס בין משתמשים שהמייל שלהם אומת, משתמשים שהמייל שלהם לא אומת ומשתמשים שלא קיימים נתונים אודות אימות כתובת המייל שלהם (בעלי ערך nan). היחס בין משתמשים שכתובת המייל שלהם אומתה ובין משתמשים שלא, כמעט זהה. אחוז ה-samples בהם קיים ערך nan קטן ביחס ליתר ה-samples. בנוסף, בחנו את הקשר בין אימות כתובת המייל של המשתמש לבין הסנטימנט ואנו מסיקים כי ישנו קשר בין אימות כתובת המייל לסנטימנט של ההודעה (קישור לתרשימים נוספים).

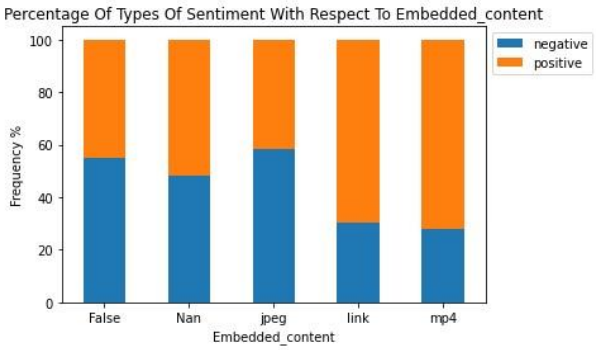
blue_tick (1.11)

עבור הניתוח של משתנה זה, בחנו את היחס בין משתמשים שאומתו, משתמשים שלא אומתו ומשתמשים שלא קיימים נתונים אודות האימות שלהם (בעלי ערך nan). רוב המשתמשים לא אומתו ואחוז ה-samples בהם קיים ערך nan קטן ביחס ליתר ה-samples. בנוסף, בחנו את הקשר בין אימות המשתמש לבין הסנטימנט ואנו מסיקים כי במצב הנוכחי כאשר ישנם ערכי nan ב-feature זה, אין לדעתנו קשר מובהק בין המגדר לסנטימנט של ההודעה (קישור לתרשימים נוספים). ייתכן שלאחר שנמלא את ערכי ה-nan בהמשך ייראה קשר בין משתנה זה למשתנה המטרה ולכן נבחן את הקשר הזה בשנית בהמשך.



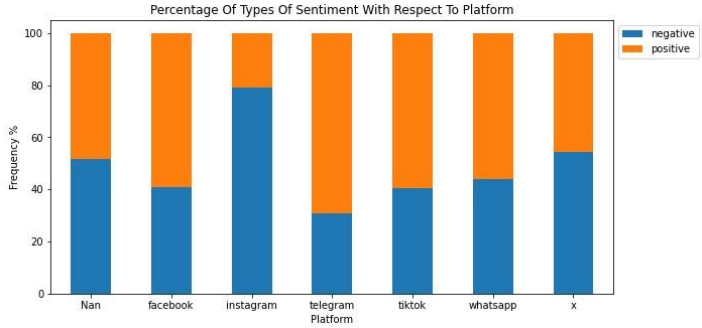


embedded_content (1.12)



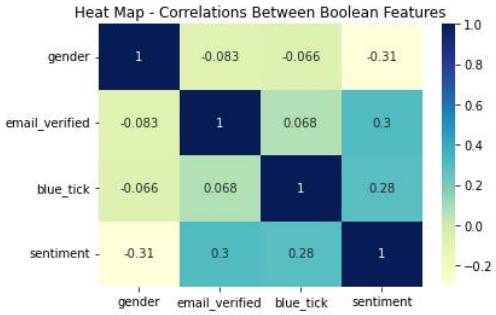
עבור הניתוח של משתנה זה, בחנו את היחס בין הודעות אליהן צורף קובץ (בפילוח לפי סוגי קבצים), הודעות אליהן לא צורף קובץ והודעות שלא ידוע האם צורף אליהן קובץ (בעלות ערכי nan) לבין הסנטימנט. מתוך הניתוח עולה כי רק בעזרת חלק מסוגי הקבצים המצורפים להודעה ניתן לקבוע קשר לסנטימנט, ולכן במצב הנוכחי כאשר ישנם ערכי nan ב-feature זה, אין לדעתנו קשר מובהק בין המגדר לסנטימנט של ההודעה (קישור לתרשימים נוספים). ייתכן שלאחר שנמלא את ערכי ה-nan בהמשך ייראה קשר בין משתנה זה למשתנה המטרה ולכן נבחן את הקשר הזה בשנית בהמשך.

platform (1.13)



עבור הניתוח של משתנה זה, בחנו את היחס בין סוג הרשת החברתית לסנטימנט. מתוך הניתוח עולה כי עבור רוב הרשתות ניתן לקבוע קשר לסנטימנט, ולכן בשלב זה אנו מסיקים כי ישנו קשר בין שדה זה למשתנה ההחלטה (קישור לתרשימים נוספים).

(1.14) קשרים מעניינים



בחנו באמצעות מספר תרשימים את הקורלציה בין משתנים מסבירים, ואת הקשרים בין משתנים בוליאניים לבין משתנה המטרה ובין המשתנים הבוליאניים באמצעות מבחן קורלציה. מסקנתנו מכלל התרשימים כי לא קיים קשר מובהק בין המשתנים שבחנו, ולכן הסקנו כי המשתנים המסבירים שבדקנו בלתי תלויים (קישור לתרשימים נוספים).

סעיף 2 - Processes and Modules of ML System

Pre-processing

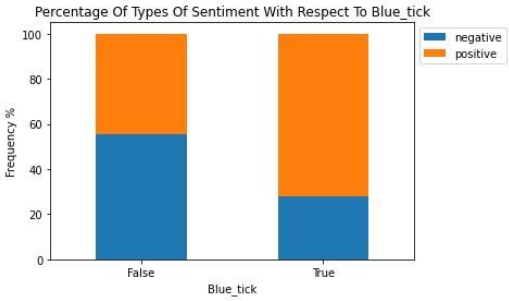
- Missing values – במסגרת שלב זה טיפלנו בערכים החסרים שהיו בחלק מה-features, באמצעות שיבוץ ערכים מתאימים באופן רנדומלי, תוך שמירה יחסית על היחסים בין סנטימנט חיובי וסנטימנט שלילי לקטגוריות השונות של ה-feature ושמירה יחסית על השונות של ה-features (ראה/י נספח). אנו לוקחים בחשבון שצורת מילוי זו עלולה להשפיע על ה-covariance בין ה-features שהיו בהם missing values, לבין יתר המשתנים. השפעה זו עלולה לבוא לידי ביטוי בהנחת אי התלות שהנחנו בין ה-features, כלומר לאחר מילוי ה-missing values ה-features עלולים להיות תלויים ולהשפיע על ה-

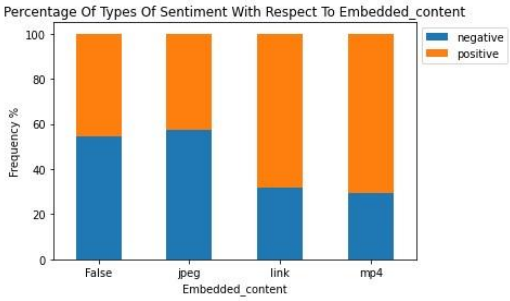
```
textID                        0
text                          0
sentiment                     0
message_date                  0
account_creation_date         0
previous_messages_dates       0
date_of_new_follower          0
date_of_new_follow            0
email                       891
gender                     1619
email_verified             1023
blue_tick                  1439
embedded_content            906
platform                    750
dtype: int64
```

Missing values בפילוח לפי משתנים

labeling של samples חדשים במערכת המפותחת. יש לציין שכיוון שמדובר במילוי רנדומלי, הרצה נוספת של מילוי השדות תביא ליחסים מעט שונים מאלו שלפיהם התבססנו בעבודה זו, אך כאמור הפרופורציות בין הקטגוריות למשתנה המטרה ובינן לבין עצמן ישמרו באופן יחסי. עבור feature מסוג email שאינו קטגוריאלי, מצאנו כי ישנה חזרתיות בסיומות של כתובות המייל. יצרנו שדה חדש הכולל רק את הסיומות של כתובות המייל והשתמשנו בהן כדי למלא את הערכים החסרים.

- לאחר מילוי הערכים החסרים, חזרנו ובחנו את הקשר בין blue_tick למשתנה ההחלטה, ובין embedded_content למשתנה ההחלטה באמצעות גרפים. מצאנו כי קיים קשר בין כל אחד מהמשתנים הנ"ל למשתנה ההחלטה.





- Data type conversions – בתת השלב הזה יש להמיר משתנים רציפים למשתנים קטגוריאליים. סט הנתונים מכיל שני משתנים בעלי נתונים הנמדדים ביחידות זמן- message_date ו- account_creation_date, כאשר זמן הוא רציף. מכיוון שבשלב ה-EDA ראינו שישנו קשר בין משתנים אלו למשתנה המטרה לאחר שמבצעים חילוץ של חלק מהנתונים (כמו חודש), ולא יהיה ניתן לחלץ את הנתונים הללו לאחר המרת המשתנים לקטגוריות, בחרנו שלא לבצע את ההמרה. במידה ובחלק ב' נראה כי טיב המודל אינו עומד ברף הרצוי, נבחן בשנית את החלטה זו.
- Imbalanced data – בהמשך לניתוח שביצענו במסגרת סעיף 1 (EDA) אחוז ה-samples בעלי סנטימנט חיובי עומד על כ-52%, בעוד שאחוז ה-samples בסט הנתונים בעלי סנטימנט שלילי עמוד על כ-47%. על פי ההגדרה סט הנתונים אינו מאוזן, אך מכיוון שחוסר האיזון קטן מאוד ועומד על כ-2% בחרנו להתייחס לסט הנתונים כמאוזן. סיבה נוספת להנחת האיזון של סט הנתונים הינה שביצוע פעולת האיזון, היכולה להתבצע באמצעות מחיקת samples בעלי סנטימנט חיובי או באמצעות הוספת samples בעלי סנטימנט שלילי, עלולה לפגוע באמינות הנתונים וליצור הטיות בהבחנה בין ה-classes.

<u>Segmentation</u>

שלב זה כולל חילוץ מידע מ-feature מורכב, למשל הפרדת גלי קול מקובץ שמע למספר דוברים. מניתוח של סט הנתונים נראה כי אין entity ממנה ניתן לחלץ אלמנטים נוספים ולכן אנו מניחים כי ה-data set שקיבלנו כבר עבר segmentation.

<u>Feature extraction</u>

בחרנו להתייחס לכלל ה-features כחסרי חשיבות לזמן (No Importance of features' occurrence over time). גם עבור נתונים גולמיים מהם חילצנו את החודש או את השנה בחרנו להתייחס באופן הזה, משום שמבחינתנו החודשים למשל מהווים רק קטגוריות (אם היינו משנים את השמות שלהם תוך שאנחנו שומרים על ההפרדה בין החודשים המקוריים היינו מקבלים את אותה התוצאה). בנוסף, אנו מתייחסים ל-

features כ-fixed number of features מכיוון שהם לא ישתנו עבור הודעות אחרות שיינתנו למודל (הערכים בתוכם ישתנו, אך הם יישאר features יחידים), וכ-Specific (knowledge based feature) מכיוון שהגינו את הרעיון ל-features מתוך הניתוח שעשינו בשלב ה-EDA. בחנו את הקשרים בין כל אחד מה-features שחילצנו לבין סנטימנט, עבור קשרים שלא ניתחנו בשלב ה-EDA (קישור לתרשימים). מניתוח התרשימים הנ"ל עולה כי קיים קשר בין כל אחד מה-features שחולצו לסנטימנט עבור כל feature המופיע ברשימה מטה. בסיום השלב, נותרנו עם ה-features שניתן לראות בצילומי המסך שבנספחים (קישור).

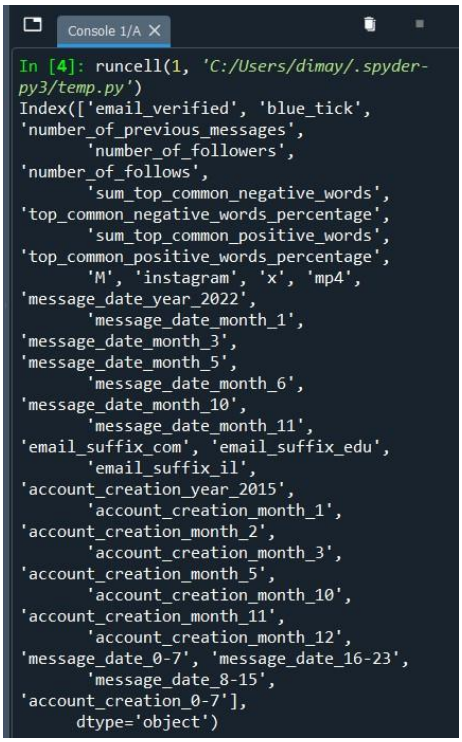
- text_word_count –מונה את כמות המילים שההודעה שפורסמה מכילה (מ-text).
- sum_top_common_negative_words – כמות הפעמים בהן מופיעות מילים מתוך 15 המילים הכי נפוצות בהודעות בעלות סנטימנט שלילי בהודעה שפורסמה (מ-text).
- top_common_negative_words_percentage – היחס בין כמות הפעמים בהן מופיעות מילים מתוך 15 המילים הכי נפוצות בהודעות בעלות סנטימנט שלילי לבין כמות המילים מ-text.
- sum_top_common_positive_words – כמות הפעמים בהן מופיעות מילים מתוך 15 המילים הכי נפוצות בהודעות בעלות סנטימנט חיובי בהודעה שפורסמה (מ-text).
- top_common_positive_words_percentage – היחס בין כמות הפעמים בהן מופיעות מילים מתוך 15 המילים הכי נפוצות בהודעות בעלות סנטימנט חיובי לבין כמות המילים מ-text.
- message_date_year – השנה בה נשלחה ההודעה (מ-message_date).
- message_date_month – החודש בו נשלחה ההודעה (מ-message_date).
- hour_ranges_of_message_date – השעה בה נשלחה ההודעה בפילוח לקבוצות השעות: 00:00-07:59, 08:00-15:59, 16:00-23:59 (מ-message_date).
- account_creation_year – השנה בה המשתמש נוצר (מ-account_creation_date).
- account_creation_month – החודש בו המשתמש נוצר (מ-account_creation_date).
- hour_ranges_of_account_creation – השעה בה המשתמש נוצר בפילוח לקבוצות השעות: 00:00-07:59, 08:00-15:59, 16:00-23:59 (מ-account_creation_date).
- seniority – הותק של המשתמש ברשת בעת שליחת ההודעה (בפילוח לשנים). מחושב כפער בין תאריך הרישום של המשתמש (account_creation_date) לבין תאריך שליחת ההודעה (message_date).
- number_of_previous_messages – כמות ההודעות הקודמות שנשלחו (מ-previous_messages_dates).
- number_of_followers – כמות העוקבים אחרי המשתמש (מ-date_of_new_follower).
- number_of_follows – כמות המשתמשים אחריהם כל משתמש עוקב (מ-date_of_new_follow).
- email_domain_suffix – הסיומת של כתובת המייל של המשתמש, הכוללת את התווים המופיעים אחרי הנקודה האחרונה בכתובת ועד לסופה, למשל il (מ-email).

Feature representation

בשלב זה עלינו לייצג מחדש את ה-entities באופן כזה שהערכים של כל ה-features במודל יהיו מיוצגים באותה סקאלת הערכים. מטרת שלב זה הינה למנוע הטיות של המודל המושפעות מסקאלות ערכים שונות היכולות להשפיע על הלמידה של המודל, למשל דרך מתן משקל גדול יותר לאחד מה-feature. בשלב זה הורדנו את ה-feature של textID. ה-features: top_common_negative_words_percentage, top_common_positive_words_percentage לא שונו, מכיוון שערכיהם כבר נעים בין 0 ל-1.

- **עבור ה-features הנומריים** (text_word_count, sum_top_common_negative_words, sum_top_common_positive_words, number_of_previous_messages, number_of_followers, number_of_follows) השתמשנו ב-Min-Max Normalization. בשיטה זו משתמשים בנוסחה: (value - minimum value) / (maximum value - minimum value).
- **עבור ה-features הקטגוריאליים** (message_date_year, message_date_month, hour_ranges_of_message_date, account_creation_year, account_creation_month, hour_ranges_of_account_creation, seniority, email_domain_suffix, gender, embedded_content, platform) ביצענו נרמול בשיטת One-Hot Encoding, שהופכת את הקטגוריות של המשתנה ל-features. תחת כל feature מסמנים לכל sample 0/1 כתלות בנתונים שהופיעו במשתנה המקורי.
- **עבור ה-features הבינאריים** (sentiment, email_verified, blue_tick) להם יש שני ערכים, שינינו את ייצוג הערכים ל-0 ול-1 (הערכים false ו-negative שונו ל-0, והערכים true ו-positive שונו ל-1).

בסיום השלב, נותרנו עם 71 features שניתן לראות בצילומי המסך שבנספחים (קישור).



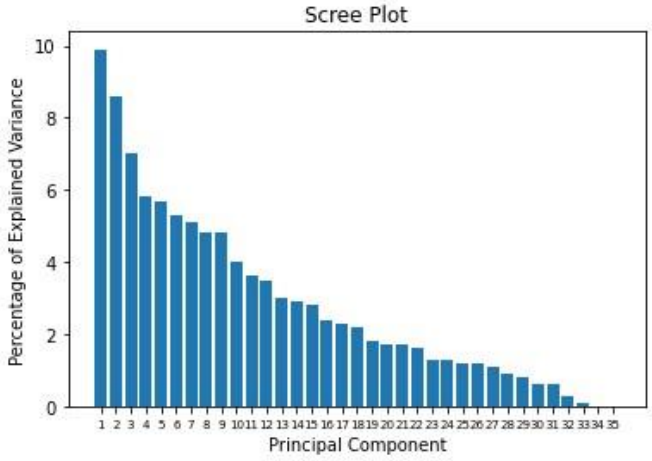
```
Console 1/A
In [4]: runcell(1, 'C:/Users/dimay/.spyder-py3/temp.py')
Index(['email_verified', 'blue_tick',
'number_of_previous_messages',
       'number_of_followers',
'number_of_follows',
       'sum_top_common_negative_words',
'top_common_negative_words_percentage',
       'sum_top_common_positive_words',
'top_common_positive_words_percentage',
       'M', 'instagram', 'x', 'mp4',
'message_date_year_2022',
       'message_date_month_1',
'message_date_month_3',
'message_date_month_5',
       'message_date_month_6',
'message_date_month_10',
       'message_date_month_11',
'email_suffix_com', 'email_suffix_edu',
       'email_suffix_il',
'account_creation_year_2015',
       'account_creation_month_1',
'account_creation_month_2',
       'account_creation_month_3',
'account_creation_month_5',
       'account_creation_month_10',
'account_creation_month_11',
       'account_creation_month_12',
'message_date_0-7', 'message_date_16-23',
       'message_date_8-15',
'account_creation_0-7'],
      dtype='object')
```

רשימת ה-features שנותרו בסוף שלב ה-feature selection

Feature selection

בחרנו להשתמש ב-Quantitative evaluation. מתוך אסטרטגיה זו, בחרנו להשתמש ב-Forward stepwise selection, שהיא תת שיטה של Wrappers. בשיטה זו מתחילים ממודל ריק מ-features ובכל איטרציה מוסיפים למודל את ה-feature בעל התרומה הגדולה ביותר. את התרומה של כל אחד מה-features בחרנו למדוד באמצעות מדד AUC. השימוש במדד זה נעשה לרוב ב-Binary classification, שזוהי משימת הלמידה במקרה שלנו. תוצאה גבוהה של מדד AUC מצביעה על כך שהמודל טוב יותר בחיזוי ה-classes השונים. השתמשנו ברגרסיה לוגיסטית, מכיוון שהיא מתאימה למשתנה מטרה בוליאני ולמשתנים מסבירים רציפים (נומריים),

קטגוריאליים ובוליאניים. בסיום השלב, נותרנו עם 35 features, המהווים כמחצית מה-features איתם הגענו לשלב זה (קישור לצילומי מסך מתוך סט הנתונים בסיום השלב).



Dimensionality reduction

בדקנו האם יש צורך בהורדת מימד באמצעות הנוסחה M^d (M= מספר ה-classes, d= מספר המימדים/features). במקרה שלנו, $2^{35} = 3.435 \cdot 10^{10}$. על מנת שלא נצטרך לבצע את השלב הזה, התוצאה של M^d צריכה להיות קטנה מכמות ה-samples שב-dataset. מכיוון שאין זהו המצב במקרה שלנו (קיימים 12,272 samples), ואנו רוצים להימנע ממצב של overfitting, עלינו לבצע Dimensionality reduction. לטובת כך, בחרנו להשתמש בשיטת PCA, המאפשרת לשמור על ה-information importance ועל השונות. במסגרת יישום השיטה נוצרות components המכילות קומבינציות של ה-features איתם הגענו לשלב הזה. על מנת שתתקיים המשוואה $2^d < 12{,}272$, מספר המימדים המירבי אליו נרצה להגיע הוא 13, קרי 13 components. סכמנו את אחוז השונות המוסברת של 13 ה-components הראשונות, וקיבלנו כ-70% שונות מוסברת (קישור ל-dataset המתקבל בסיום השלב). היינו מעדיפים לקבל אחוז גבוה יותר של שונות מוסברת, אך העדפנו בשלב זה לוודא כי המשוואה $2^d < 12{,}272$ מתקיימת. במידה ובחלק ב' נראה כי טיב המודל אינו עומד ברף הרצוי, נבחן בשנית את החלטה זו.

Validation

שיטת הוולידציה לנתונים שבחרנו הינה K fold, שהיא תת שיטה של cross validation. בשיטת K fold אנו מחלקים את המודל ל-K חלקים שווים (לרוב K=10), ומבצעים K איטרציות של המודל עם סט הנתונים המחולק, כאשר בכל איטרציה אחד מ-K החלקים מסט הנתונים משמש כ-validation set, ושאר הנתונים משמשים כ-training set. בסיום כל איטרציה מתקבל אחוז הדיוק. לאחר הרצאת K האיטרציות, מבצעים שקלול של אחוזי הדיוק מכל האיטרציות באמצעות ממוצע וזוהי התוצאה על ה-cross validation.

בחרנו בשיטה זו מכיוון שהיא עושה שימוש בכל סט הנתונים, בניגוד לשיטות אחרות. יתרונות נוספים שגרמו לנו לבחור בשיטה זו הינם ששיטה זו יחסית מדויקת ובסופה ניתן אומדן מדויק של ביצועי המודל, וביחס לשיטות אחרות (כמו ה-leave *one* out) היא נחשבת למהירה מבחינת זמני הריצה.

בחרנו להשתמש במטריקה AUC, וזאת מכיוון שהיא מתאימה לסט נתונים לא מאוזן. עד כה, בחרנו להסתכל על סט הנתונים כמאוזן מכיוון שחששנו מהטיות של סט הנתונים הנובעות מאיזונו, אך במקרה זה עבודה עם סט הנתונים שבידנו כלא מאוזן יוביל לתוצאה טובה יותר של המודל.

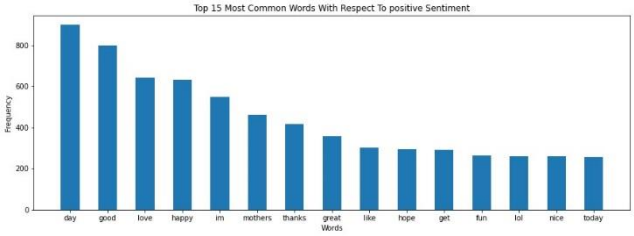
נספחים

נספחים בנושא EDA

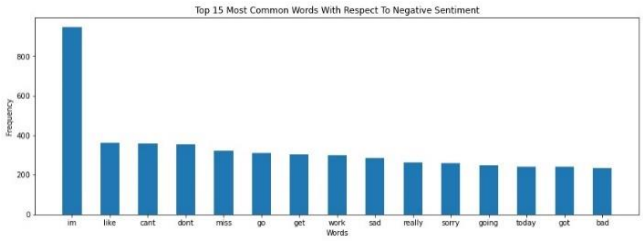
נספח 1.2 – תרשימים הקשורים למשתנה text

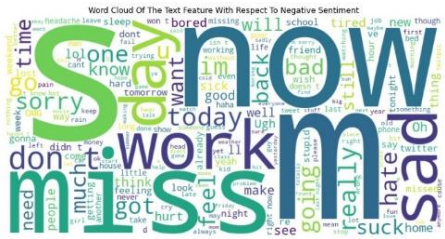
(חזרה לסעיף 1.2)

- תרשימים הקשורים לניתוח מילים המופיעות בהודעות המתויגות כבעלות סנטימנט חיובי



- תרשימים הקשורים לניתוח מילים המופיעות בהודעות המתויגות כבעלות סנטימנט שלילי

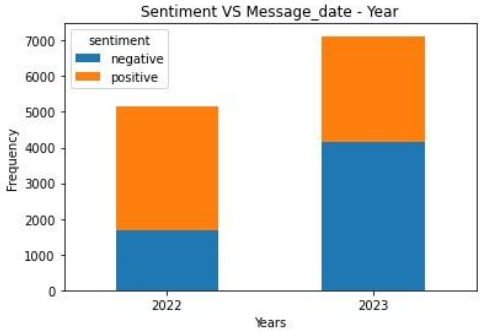


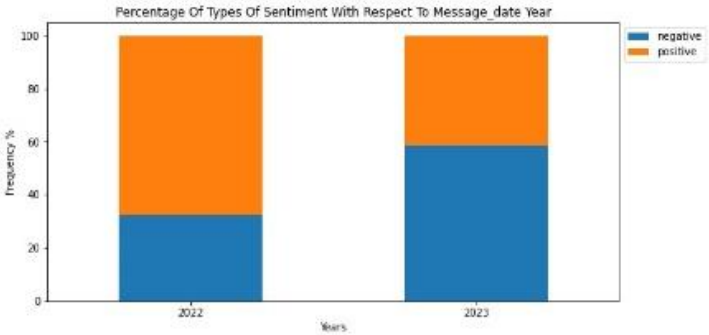


נספח 1.3 – תרשימים הקשורים למשתנה message_date

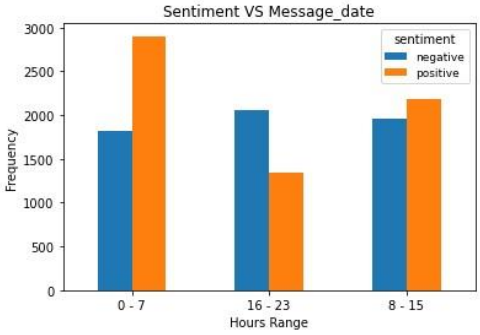
(חזרה לסעיף 1.3)

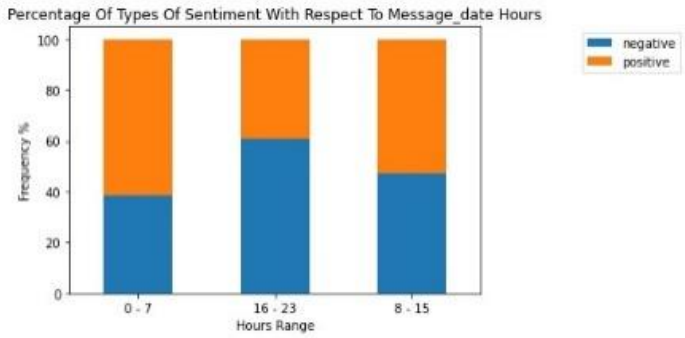
- תרשימים המייצגים את הקשר בין שנת שליחת ההודעה לסנטימנט





- תרשימים המייצגים את הקשר בין שעת שליחת ההודעה לסנטימנט (השעות מפולחות לטווחי השעות הבאים: 00:00-07:59, 08:00-15:59, 16:00-23:59)

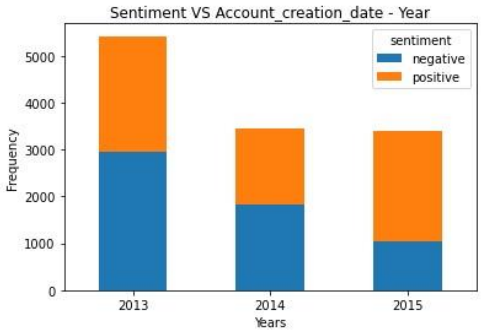


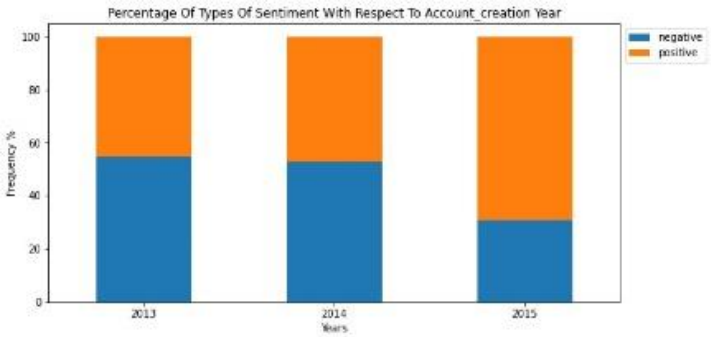


<u>נספח 1.4 – תרשימים הקשורים למשתנה account_creation_date</u>

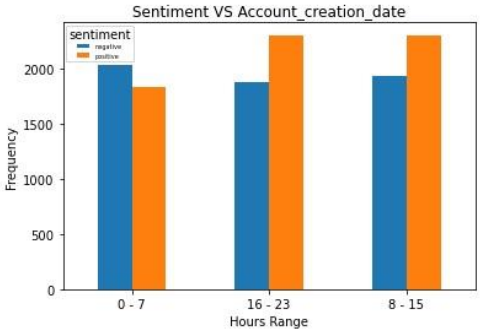
(חזרה לסעיף 1.4)

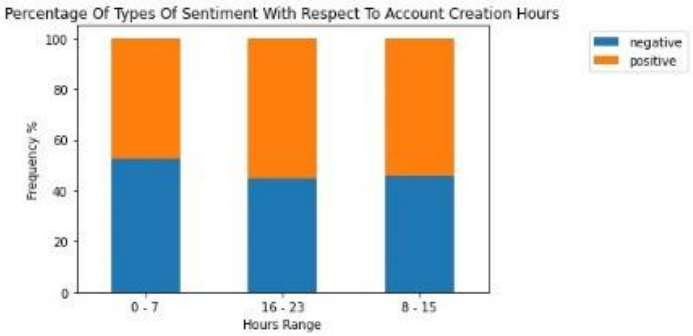
- תרשימים המייצגים את הקשר בין שנת יצירת המשתמש לסנטימנט





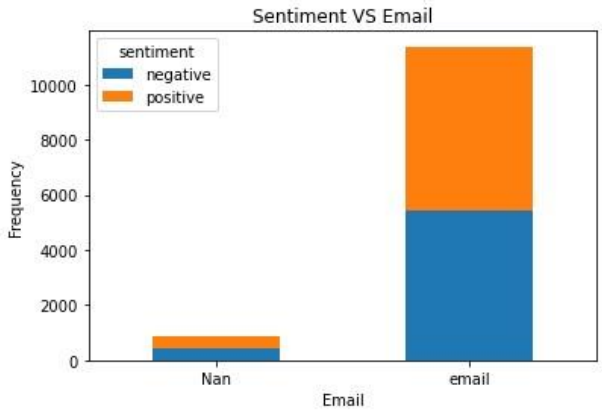
- תרשימים המייצגים את הקשר בין שעת יצירת המשתמש לסנטימנט (השעות מפולחות לטווחי השעות הבאים: 00:00-07:59, 08:00-15:59, 16:00-23:59)

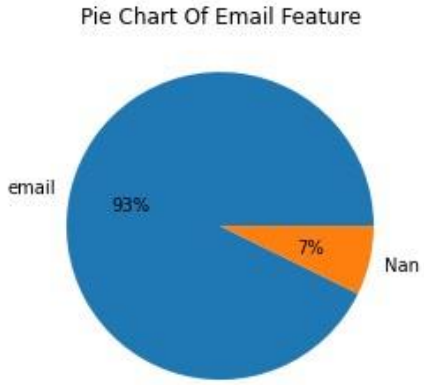




נספח 1.8 – תרשימים הקשורים למשתנה email

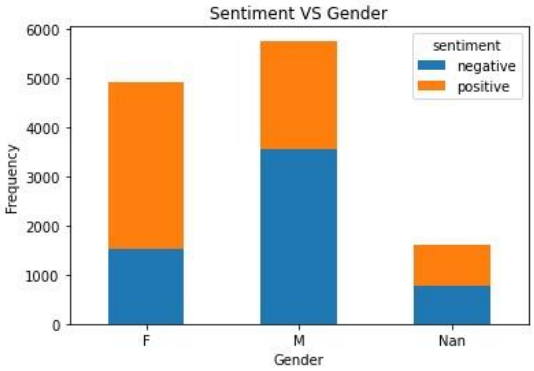
(חזרה לסעיף 1.8)

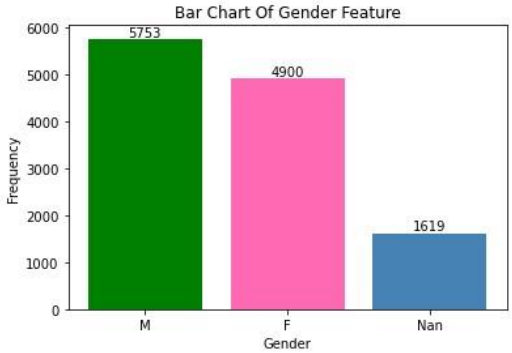




נספח 1.9 – תרשימים הקשורים למשתנה gender

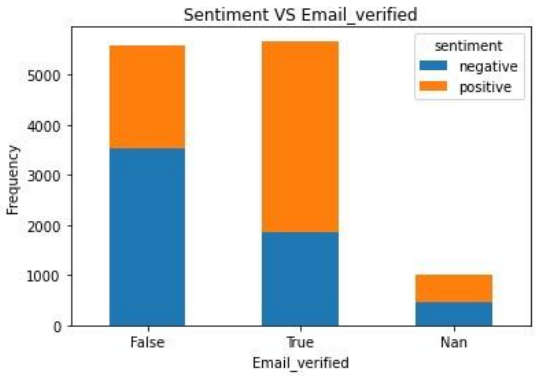
(חזרה לסעיף 1.9)

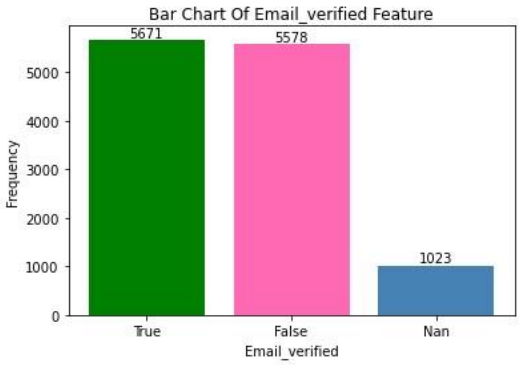




נספח 1.10 – תרשימים הקשורים למשתנה email_verified

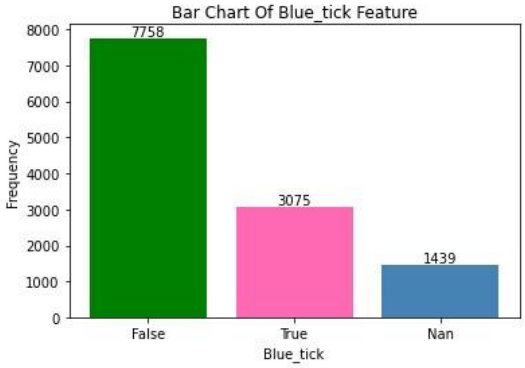
(חזרה לסעיף 1.10)

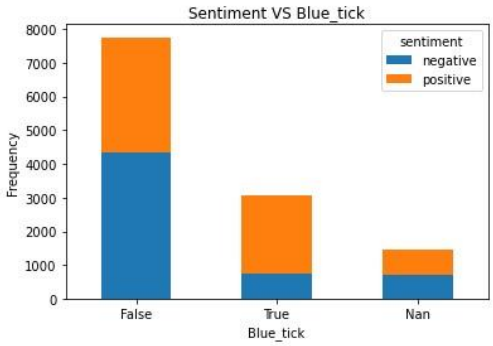




נספח 1.11 – תרשימים הקשורים למשתנה blue_tick

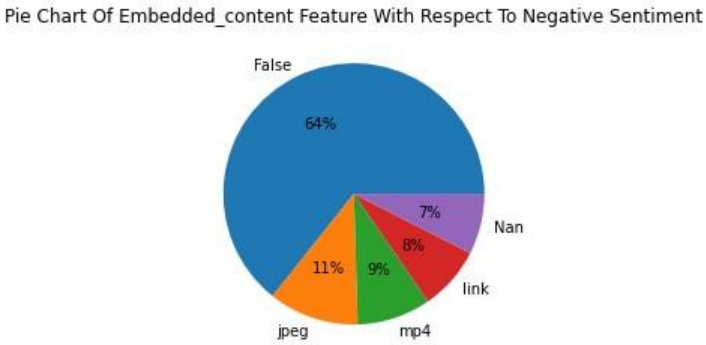
(חזרה לסעיף 1.11)

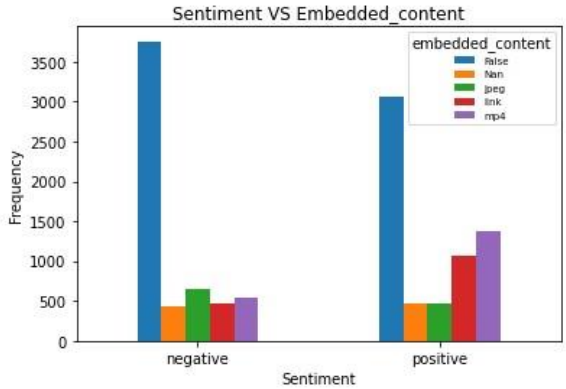


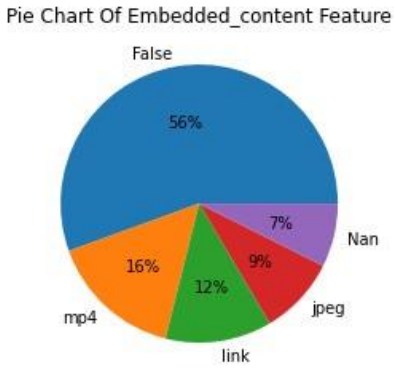


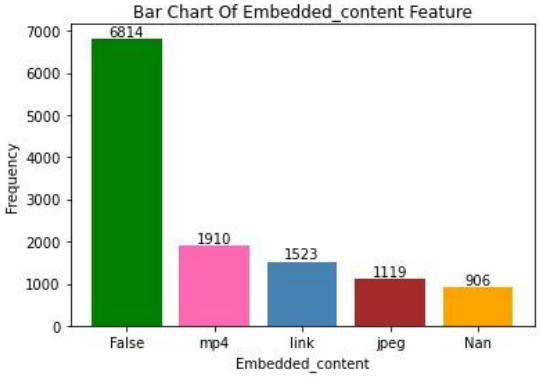
נספח 1.12 – תרשימים הקשורים למשתנה embedded_content

(חזרה לסעיף 1.12)



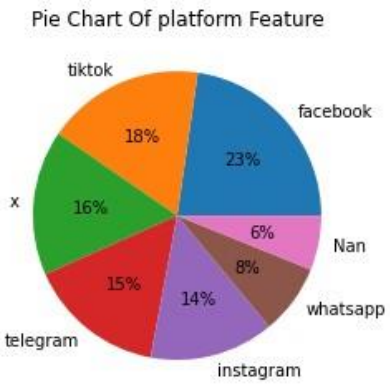


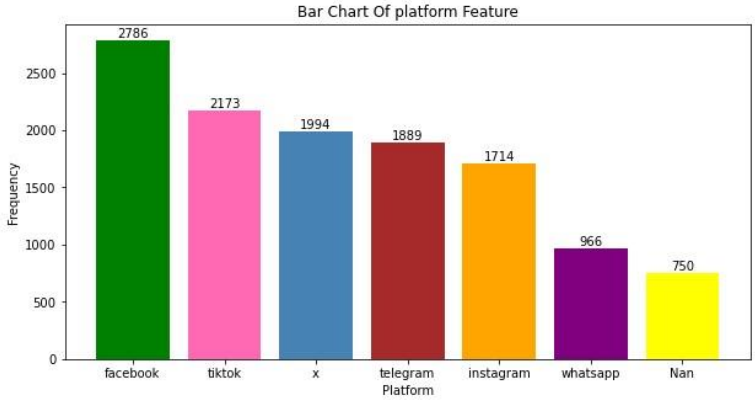


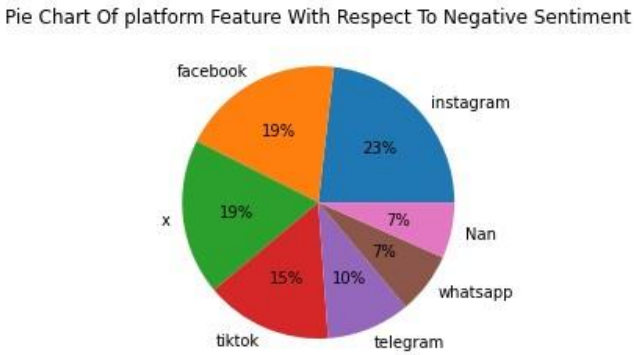


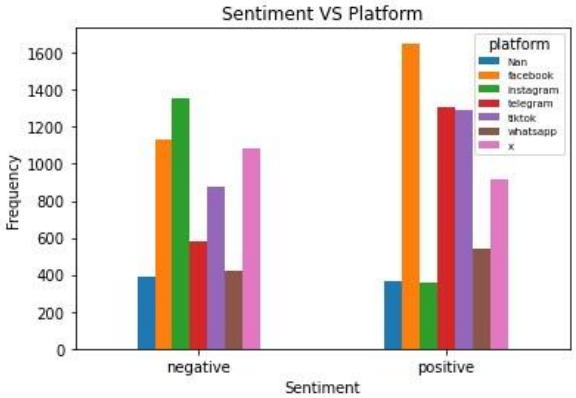
נספח 1.13 – תרשימים הקשורים למשתנה platform

(חזרה לסעיף 1.13)





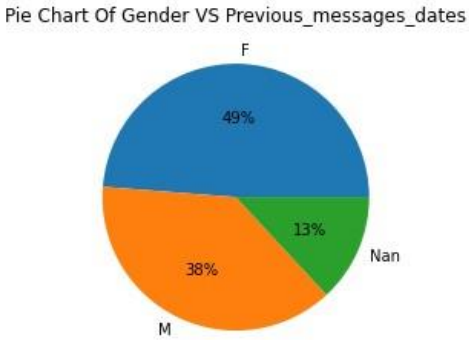


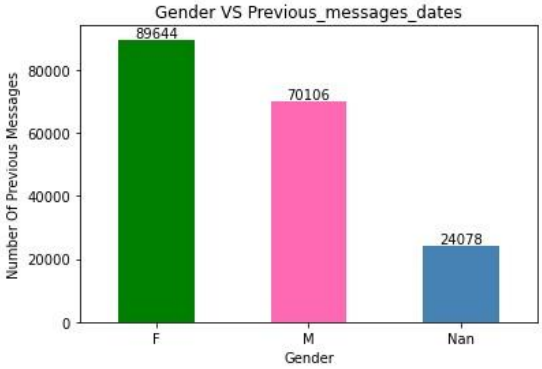


נספח 1.14 – תרשימים הקשורים לקשרים מעניינים

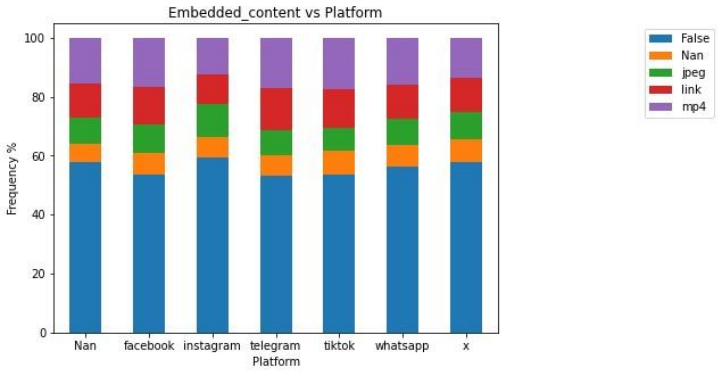
(חזרה לסעיף 1.14)

- היחס בין מגדר לכמות ההודעות הקודמות שנשלחו

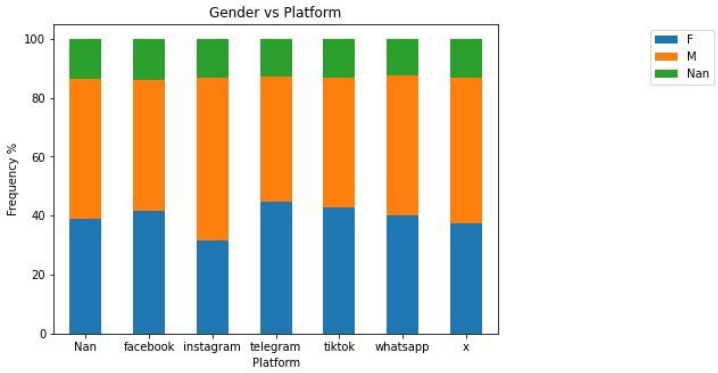




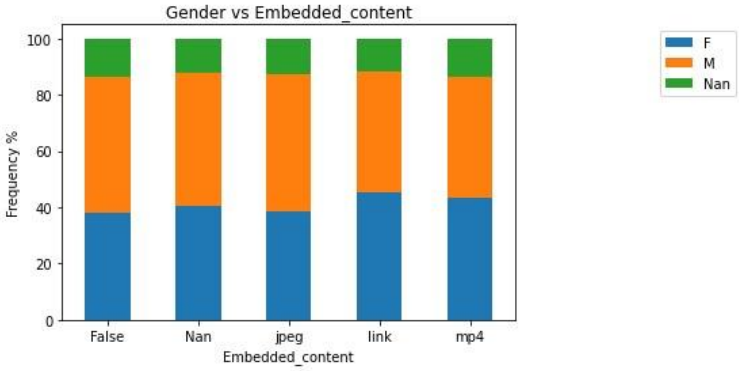
- יחס בין סוג הקובץ המצורף להודעה לרשת החברתית בה ההודעה פורסמה



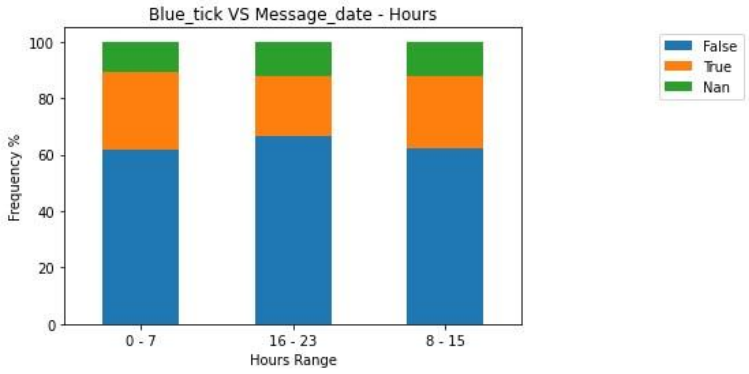
- יחס בין מגדר המשתמש לרשת החברתית בה המשתמש פרסם את ההודעה



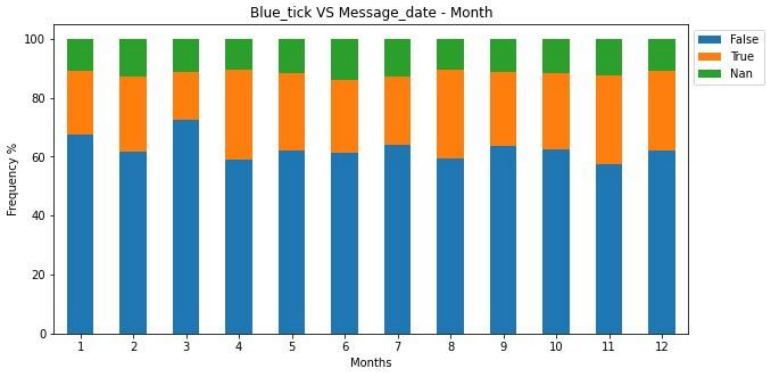
- יחס בין מגדר המשתמש לסוג הקובץ שצורף להודעה



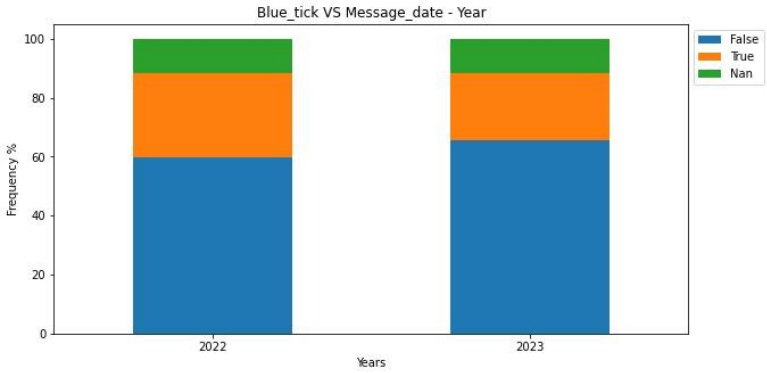
- יחס בין אימות המשתמש לשעה בה המשתמש פרסם את ההודעה (בפילוח לשעות הבוקר, צהריים וערב)



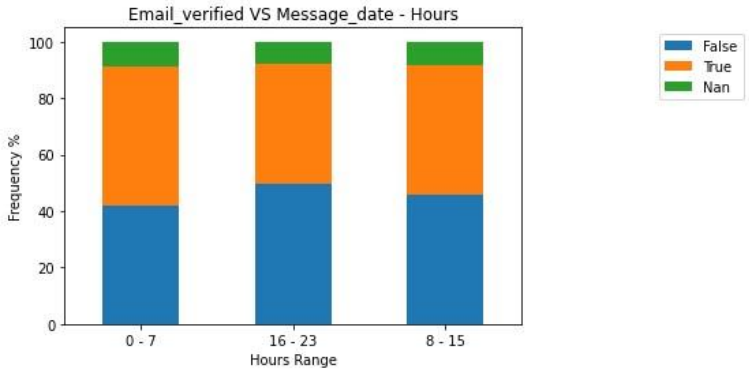
- יחס בין אימות המשתמש לחודש בו המשתמש פרסם את ההודעה



- יחס בין אימות המשתמש לשנה בה המשתמש פרסם את ההודעה

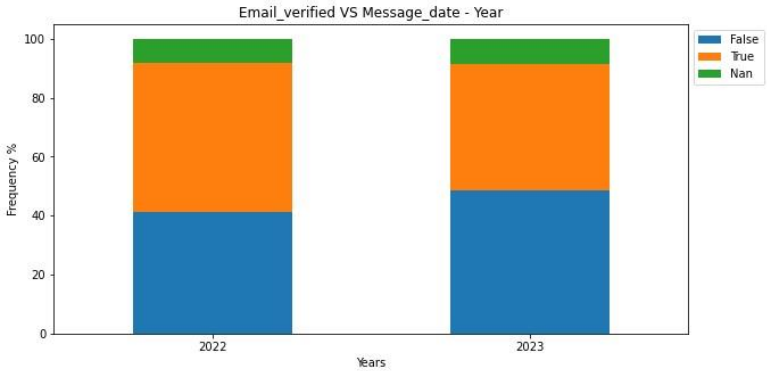


- יחס בין אימות כתובת המייל לשעה בה המשתמש פרסם את ההודעה (בפילוח לשעות הבוקר, צהריים וערב)

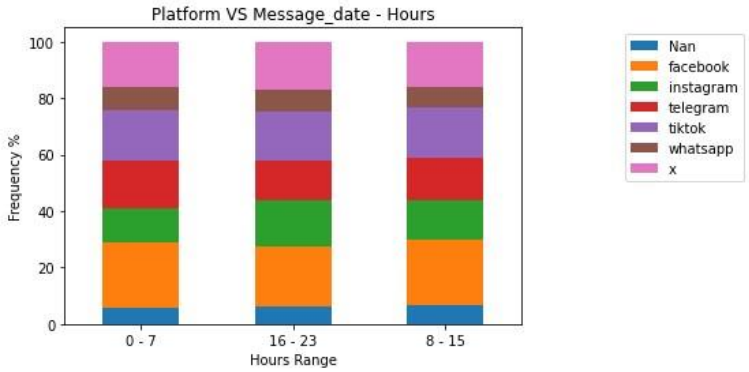


- יחס בין אימות כתובת המייל לחודש בו המשתמש פרסם את ההודעה

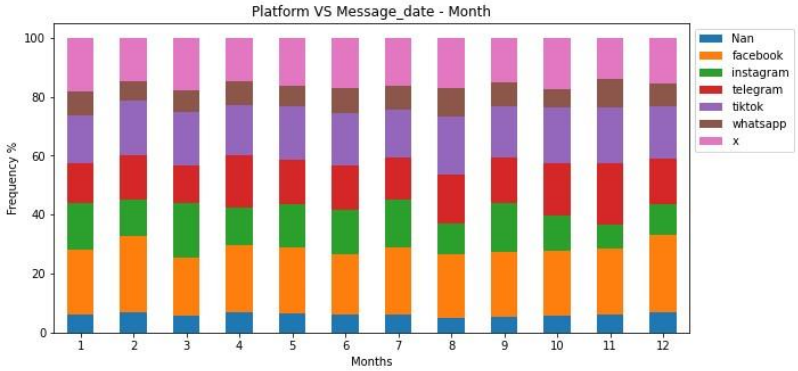
- יחס בין אימות כתובת המייל לשנה בה המשתמש פרסם את ההודעה



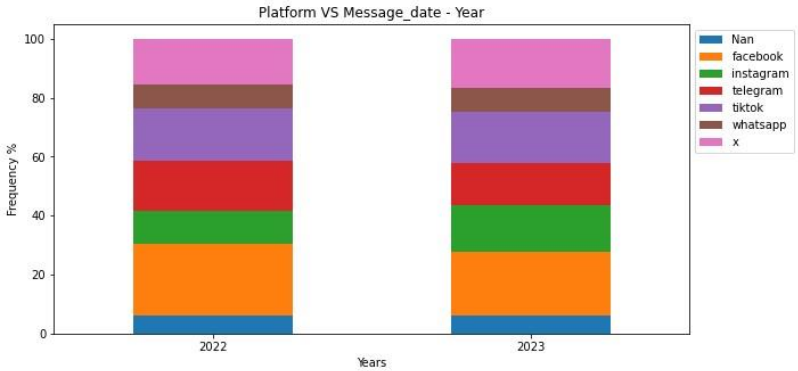
- יחס בין הפלטפורמה בה פורסמה ההודעה לשעה בה המשתמש פרסם את ההודעה (בפילוח לשעות הבוקר, צהריים וערב)



- יחס בין הפלטפורמה בה פורסמה ההודעה לחודש בו המשתמש פרסם את ההודעה



- יחס בין הפלטפורמה בה פורסמה ההודעה לשנה בה המשתמש פרסם את ההודעה



(חזרה לסעיף 1.14)

נספחים בנושא dataset creation

נספח 2.1 – בדיקת היחס בין הערכים עבור משתנים בעלי ערכים חסרים (לפני ואחרי הטיפול ב- missing values)

(חזרה ל-pre-processing)

- יחס בין הערכים True ו-False במשתנה email_verified

לפני מילוי missing values	אחרי מילוי missing values
email_verified True 0.504134 False 0.495866 Name: proportion, dtype: float64 sentiment email_verified negative False 65.548724 True 34.451276 positive True 64.977899 False 35.022101 Name: proportion, dtype: float64 Number of nulls: 1023	email_verified True 0.504971 False 0.495029 Name: proportion, dtype: float64 sentiment email_verified negative False 64.178082 True 35.821918 positive True 63.821517 False 36.178483 Name: proportion, dtype: float64 Number of nulls: 0

- יחס בין הערכים True ו-False במשתנה blue_tick

לפני מילוי missing values	אחרי מילוי missing values
blue_tick False 0.716145 True 0.283855 Name: proportion, dtype: float64 sentiment blue_tick negative False 85.120094 True 14.879906 positive False 59.506303 True 40.493697 Name: proportion, dtype: float64 Number of nulls: 1439	blue_tick False 0.71439 True 0.28561 Name: proportion, dtype: float64 sentiment blue_tick negative False 83.373288 True 16.626712 positive False 60.603234 True 39.396766 Name: proportion, dtype: float64 Number of nulls: 0

- יחס בין הערכים M ו-F במשתנה gender

לפני מילוי missing values	אחרי מילוי missing values
gender M 0.540036 F 0.459964 Name: proportion, dtype: float64 sentiment gender negative M 70.031608 F 29.968392 positive F 60.507959 M 39.492041 Name: proportion, dtype: float64 Number of nulls: 1619	gender M 0.540825 F 0.459175 Name: proportion, dtype: float64 sentiment gender negative M 68.013699 F 31.986301 positive F 58.566542 M 41.433458 Name: proportion, dtype: float64 Number of nulls: 0

- יחס בין הערכים False, mp4, link, jpeg במשתנה embedded_contect

לפני מילוי missing values	אחרי מילוי missing values
embedded_content False 0.599507 mp4 0.168045 link 0.133996 jpeg 0.098452 Name: proportion, dtype: float64 sentiment embedded_content negative False 69.498427 jpeg 12.067370 mp4 9.883398 link 8.550805 positive False 51.299681 mp4 23.075633 link 17.793057 jpeg 7.831628 Name: proportion, dtype: float64 Number of nulls: 906	embedded_content False 0.599250 mp4 0.167128 link 0.135023 jpeg 0.098598 Name: proportion, dtype: float64 sentiment embedded_content negative False 68.681507 jpeg 11.969178 mp4 10.308219 link 9.041096 positive False 51.974502 mp4 22.527985 link 17.552861 jpeg 7.944652 Name: proportion, dtype: float64 Number of nulls: 0

- יחס בין הערכים facebook, tiktok, x, telegram, instagram, whatsapp במשתנה platform

לפני מילוי missing values	אחרי מילוי missing values
platform facebook 0.241798 tiktok 0.188596 x 0.173060 telegram 0.163947 instagram 0.148759 whatsapp 0.083840 Name: proportion, dtype: float64 sentiment platform negative instagram 24.812030 facebook 20.795892 x 19.823950 tiktok 16.119567 telegram 10.654685 whatsapp 7.793875 positive facebook 27.220300 telegram 21.552150 tiktok 21.321470 x 15.043665 whatsapp 8.914154 instagram 5.948262 Name: proportion, dtype: float64 Number of nulls: 750	platform facebook 0.242422 tiktok 0.189456 x 0.173321 telegram 0.163462 instagram 0.147898 whatsapp 0.083442 Name: proportion, dtype: float64 sentiment platform negative instagram 24.092466 facebook 21.044521 x 19.657534 tiktok 16.352740 telegram 10.924658 whatsapp 7.928082 positive facebook 27.145522 tiktok 21.299751 telegram 21.268657 x 15.220771 whatsapp 8.722015 instagram 6.343284 Name: proportion, dtype: float64 Number of nulls: 0

- יחס בין הערכים il, gov, ru, de, edu, org, ke, com, jp במשתנה email_domain

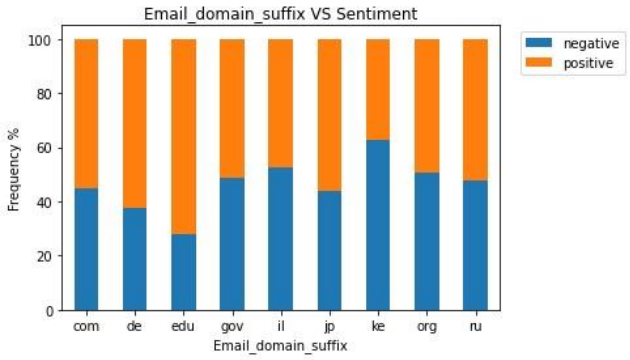
לפני מילוי missing values	אחרי מילוי missing values
email_domain il 0.261225 gov 0.199982 ru 0.108075 de 0.084703 edu 0.077498 org 0.075740 ke 0.068711 com 0.063263 jp 0.060803 Name: proportion, dtype: float64 sentiment email_domain negative il 29.042357 gov 20.497238 ru 10.791897 ke 9.226519 org 8.029466 de 6.593002 com 5.930018 jp 5.580110 edu 4.309392 positive il 23.458242 gov 19.542934 edu 10.888926 ru 10.821711 de 10.183162 org 7.158461 com 6.687952 jp 6.536717 ke 4.721895 Name: proportion, dtype: float64 Number of nulls: 891	email_domain il 0.260593 gov 0.199234 ru 0.108866 de 0.083279 edu 0.077331 org 0.075945 ke 0.068937 com 0.063722 jp 0.062093 Name: proportion, dtype: float64 sentiment email_domain negative il 28.818493 gov 20.428082 ru 10.787671 ke 9.041096 org 7.996575 de 6.626712 com 6.027397 jp 5.736301 edu 4.537671 positive il 23.554104 gov 19.465174 ru 10.976368 edu 10.634328 de 9.872512 org 7.229478 com 6.685323 jp 6.638682 ke 4.944030 Name: proportion, dtype: float64 Number of nulls: 0

(חזרה ל-pre-processing)

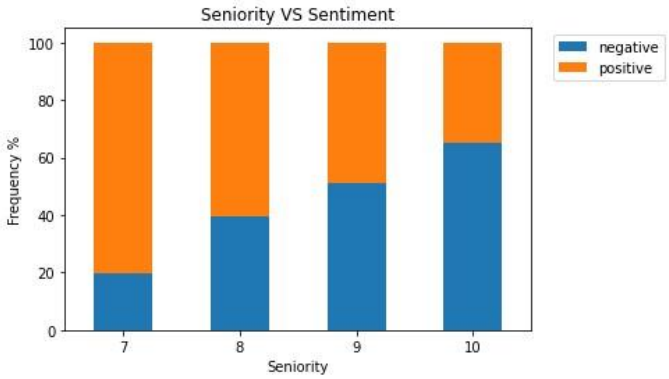
נספח 2.2 – בחינת קשרים בין משתנים שחולצו ב-feature extraction

(חזרה ל-feature extraction)

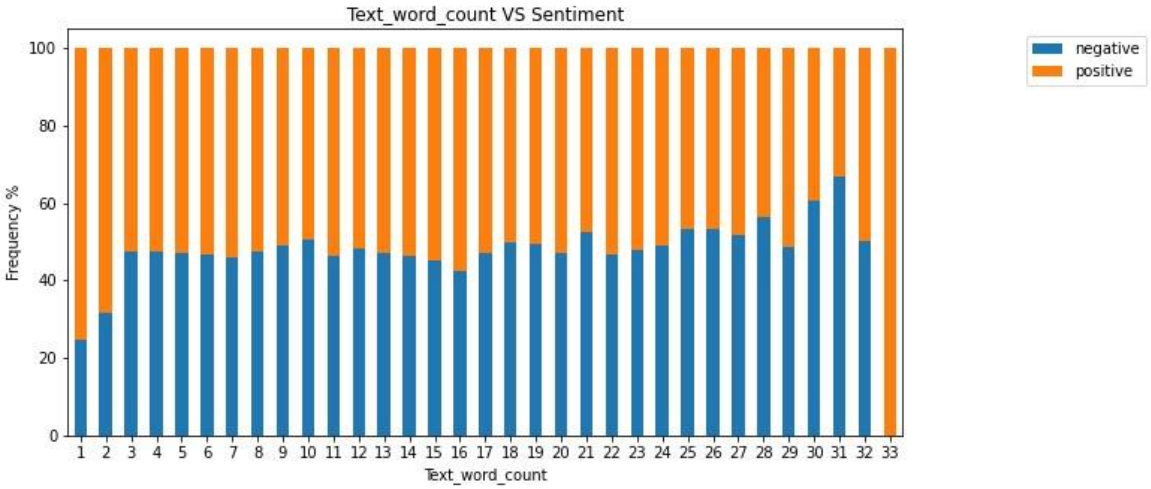
- בחינת הקשר בין ה-feature סיומת כתובת האימייל של המשתמש (email_domain_suffix) שחולץ מ-email לבין סנטימנט



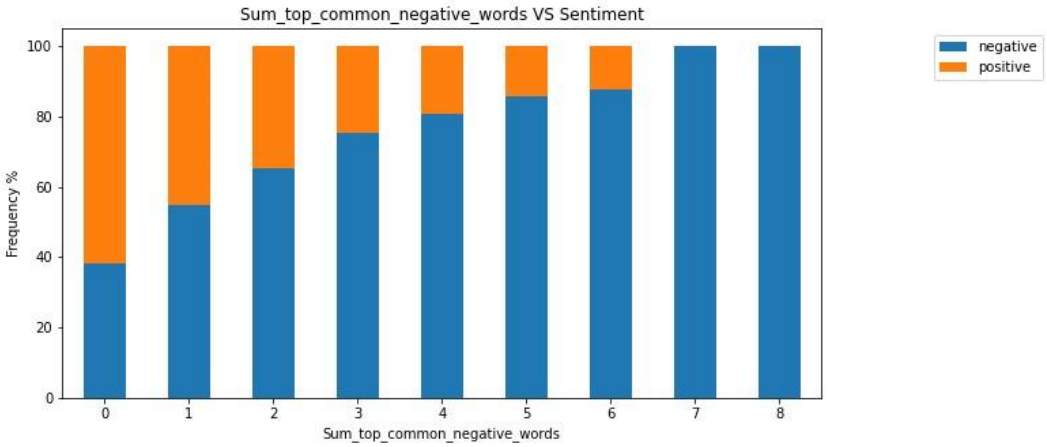
- בחינת הקשר בין ה-feature וותק של המשתמש (seniority) המחושב כפער בין תאריך שליחת ההודעה האחרונה לבין תאריך רישום המשתמש) לבין סנטימנט



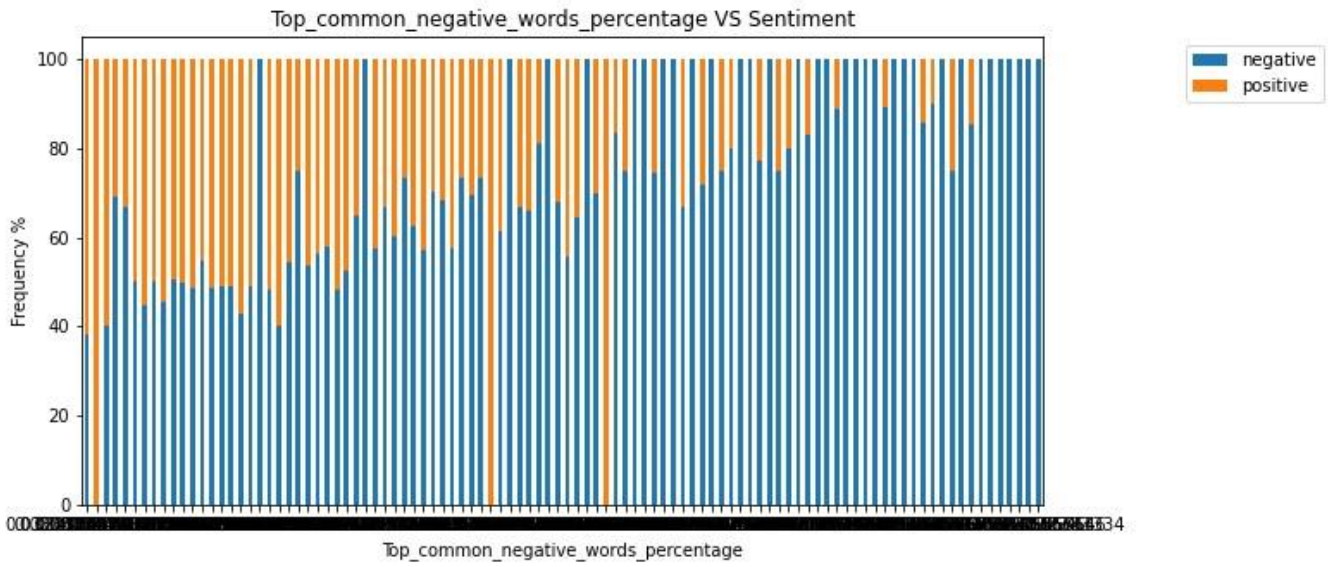
- בחינת הקשר בין ה-feature כמות המילים בהודעה (text_word_count) שחולץ מ-text לבין סנטימנט



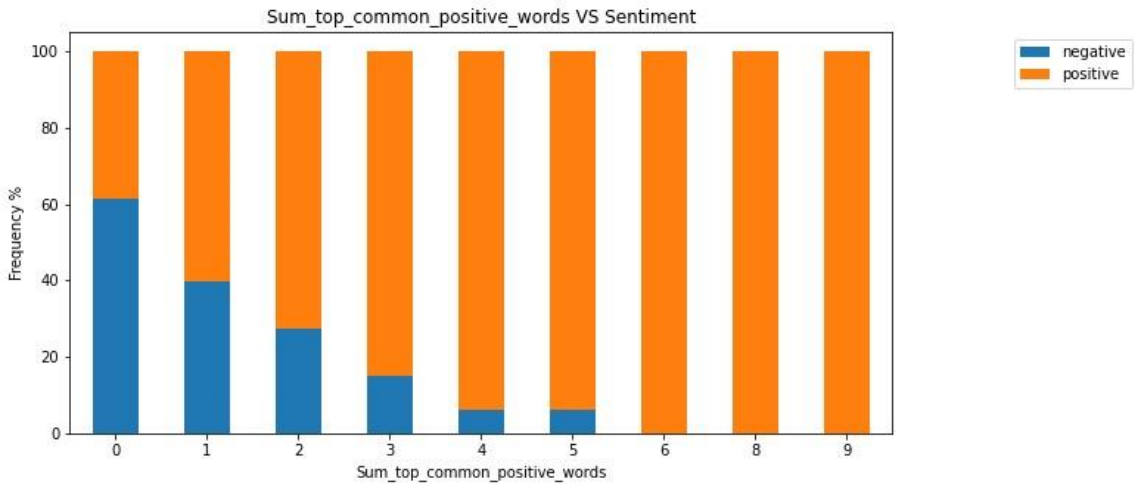
- בחינת הקשר בין ה-feature סך המילים המקושרות לסנטימנט שלילי בהודעה (sum_top_common_negative_words) שחולץ מ-text לבין סנטימנט



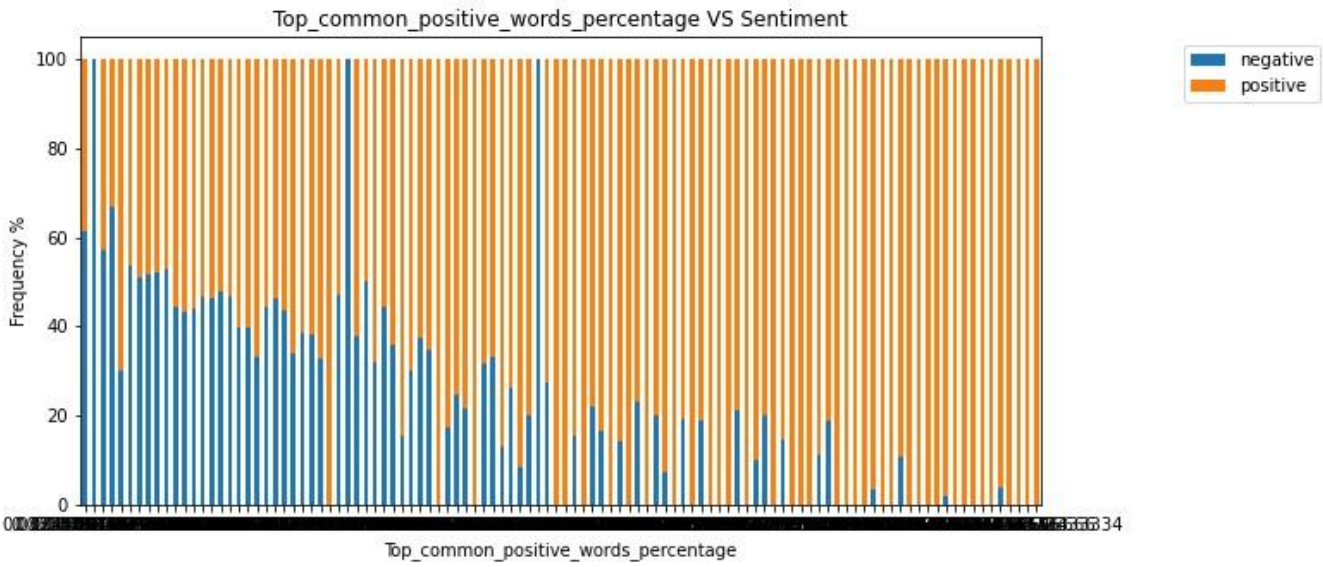
- בחינת הקשר בין ה-feature אחוז המילים המקושרות לסנטימנט שלילי בהודעה (top_common_negative_words_precentage) שחולץ מ-text לבין סנטימנט



- בחינת הקשר בין ה-feature סך המילים המקושרות לסנטימנט חיובי בהודעה (sum_top_common_positive_words) שחולץ מ-text לבין סנטימנט



- בחינת הקשר בין ה-feature אחוז המילים המקושרות לסנטימנט שלילי בהודעה (top_common_negative_words_precentage) שחולץ מ-text לבין סנטימנט



(חזרה ל-feature extraction)

נספח 2.3 – צילומי מסך של ה-features שמופיעים בסט הנתונים בסוף שלב ה-feature extraction

(חזרה ל-feature extraction)

Index	textID	sentiment	gender	mail_verifie	blue_tick	embedded_content	platform	email_domain_suffix
0	b8f4e560fa	positive	F	True	False	mp4	facebook	gov
1	f81a1511b2	positive	M	True	True	False	tiktok	gov
2	3e9e3f0d69	positive	F	True	True	False	tiktok	gov
3	a068b95bd9	negative	M	False	False	mp4	tiktok	gov
4	e6da2d1835	negative	M	False	False	jpeg	facebook	il
5	1c32289f9a	positive	F	True	False	jpeg	facebook	gov
6	a1f6095799	positive	F	False	False	False	whatsapp	il
7	7956d83106	negative	M	False	False	False	x	de
8	35b63d47cb	negative	F	False	False	False	instagram	com
9	d10467dde7	negative	M	True	False	False	instagram	edu
10	7198b4c0cf	positive	F	True	False	mp4	telegram	il

number_of_previous_messages	number_of_followers	number_of_follows	account_creation_year	account_creation_month
12	44	14	2015	4
48	18	42	2013	7
30	15	24	2013	8
5	1	4	2015	10
8	2	6	2013	12
19	43	43	2013	9
19	40	4	2013	12
5	4	2	2013	2
7	3	1	2014	8
4	9	3	2014	1
11	31	26	2014	12

hour_ranges_of_account_creation	message_date_year	message_date_month	hour_ranges_of_message_date	seniority	text_word_count
0 - 7	2023	10	0 - 7	8	11
16 - 23	2022	11	0 - 7	9	24
0 - 7	2022	6	0 - 7	9	6
8 - 15	2022	1	8 - 15	7	11
16 - 23	2022	5	16 - 23	9	20
0 - 7	2023	9	0 - 7	10	7
0 - 7	2022	11	0 - 7	9	19
16 - 23	2023	6	8 - 15	10	10
0 - 7	2023	9	8 - 15	9	4
8 - 15	2023	6	8 - 15	9	6
16 - 23	2023	11	8 - 15	9	14

sum_top_common_negative_words	top_common_negative_words_percentage	sum_top_common_positive_words	top_common_positive_words_percentage
1	0.0909091	1	0.0909091
0	0	1	0.0416667
0	0	0	0
0	0	0	0
0	0	0	0
0	0	1	0.142857
0	0	2	0.105263
0	0	0	0
0	0	0	0
1	0.166667	0	0
1	0.0714286	3	0.214286

(חזרה ל-feature extraction)

נספח 2.4 – צילומי מסך של ה-features שמופיעים בסט הנתונים בסוף שלב ה-feature representation

(חזרה ל-feature representation)

index	sentiment	email_verified	blue_tick	number_of_previous_messages	number_of_followers	number_of_follows	text_word_count	sum_top_common_negative_words	top_common_negative_words_percentage	sum_top_common_positive_words
0	1	1	0	0.244898	0.897959	0.285714	0.3125	0.125	0.0909091	0.111111
1	1	1	1	0.979592	0.367347	0.857143	0.71875	0	0	0.111111
2	1	1	1	0.612245	0.306122	0.489796	0.15625	0	0	0
3	0	0	0	0.102041	0.0204082	0.0816327	0.3125	0	0	0
4	0	0	0	0.163265	0.0408163	0.122449	0.59375	0	0	0
5	1	1	0	0.387755	0.877551	0.877551	0.1875	0	0	0.111111
6	1	0	0	0.387755	0.816327	0.0816327	0.5625	0	0	0.222222
7	0	0	0	0.102041	0.0816327	0.0408163	0.28125	0	0	0
8	0	0	0	0.142857	0.0612245	0.0204082	0.09375	0	0	0
9	0	1	0	0.0816327	0.183673	0.0612245	0.15625	0.125	0.166667	0
10	1	1	0	0.22449	0.632653	0.530612	0.40625	0.125	0.0714286	0.333333

top_common_positive_words_percentage	F	M	facebook	instagram	telegram	tiktok	whatsapp	x	none_embedded_content	jpeg	link	mp4	message_date_year_2022	message_date_year_2023
0.0909091	1	0	1	0	0	0	0	0	0	0	0	1	0	1
0.0416667	0	1	0	0	0	1	0	0	1	0	0	0	1	0
0	1	0	0	0	0	1	0	0	1	0	0	0	1	0
0	0	1	0	0	0	1	0	0	0	0	0	1	1	0
0	0	1	1	0	0	0	0	0	0	1	0	0	1	0
0.142857	1	0	1	0	0	0	0	0	0	1	0	0	0	1
0.105263	1	0	0	0	0	0	1	0	1	0	0	0	1	0
0	0	1	0	0	0	0	0	1	1	0	0	0	0	1
0	1	0	0	1	0	0	0	0	1	0	0	0	0	1
0	0	1	0	1	0	0	0	0	1	0	0	0	0	1
0.214286	1	0	0	0	1	0	0	0	0	0	0	1	0	1

message_date_month_1	message_date_month_2	message_date_month_3	message_date_month_4	message_date_month_5	message_date_month_6	message_date_month_7	message_date_month_8
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0
0	0	0	0	0	1	0	0
1	0	0	0	0	0	0	0
0	0	0	0	1	0	0	0
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0
0	0	0	0	0	1	0	0
0	0	0	0	0	0	0	0
0	0	0	0	0	1	0	0
0	0	0	0	0	0	0	0

message_date_month_9	message_date_month_10	message_date_month_11	message_date_month_12	7_years_seniority	8_years_seniority	9_years_seniority	10_years_seniority	email_suffix_com	email_suffix_de
0	1	0	0	0	1	0	0	0	0
0	0	1	0	0	0	1	0	0	0
0	0	0	0	0	0	1	0	0	0
0	0	0	0	1	0	0	0	0	0
0	0	0	0	0	0	1	0	0	0
1	0	0	0	0	0	0	1	0	0
0	0	1	0	0	0	1	0	0	0
0	0	0	0	0	0	0	1	0	1
1	0	0	0	0	0	1	0	1	0
0	0	0	0	0	0	1	0	0	0
0	0	1	0	0	0	1	0	0	0

email_suffix_edu	email_suffix_gov	email_suffix_il	email_suffix_jp	email_suffix_ke	email_suffix_org	email_suffix_ru	account_creation_year_2013	account_creation_year_2014	account_creation_year_2015
0	1	0	0	0	0	0	0	0	1
0	1	0	0	0	0	0	1	0	0
0	1	0	0	0	0	0	1	0	0
0	1	0	0	0	0	0	0	0	1
0	0	1	0	0	0	0	1	0	0
0	1	0	0	0	0	0	1	0	0
0	0	1	0	0	0	0	1	0	0
0	0	0	0	0	0	0	1	0	0
0	0	0	0	0	0	0	0	1	0
1	0	0	0	0	0	0	0	1	0
0	0	1	0	0	0	0	0	1	0

account_creation_month_1	account_creation_month_2	account_creation_month_3	account_creation_month_4	account_creation_month_5	account_creation_month_6	account_creation_month_7	account_creation_month_8	account_creation_month_9
0	0	0	1	0	0	0	0	0
0	0	0	0	0	0	1	0	0
0	0	0	0	0	0	0	1	0
0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	1
0	0	0	0	0	0	0	0	0
0	1	0	0	0	0	0	0	0
0	0	0	0	0	0	0	1	0
1	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0

account_creation_month_10	account_creation_month_11	account_creation_month_12	message_date_0-7	message_date_16-23	message_date_8-15	account_creation_0-7	account_creation_16-23	account_creation_8-15
0	0	0	1	0	0	1	0	0
0	0	0	1	0	0	0	1	0
0	0	0	1	0	0	1	0	0
1	0	0	0	0	1	0	0	1
0	0	1	0	1	0	0	1	0
0	0	0	1	0	0	1	0	0
0	0	1	1	0	0	1	0	0
0	0	0	0	0	1	0	1	0
0	0	0	0	0	1	1	0	0
0	0	0	0	0	1	0	0	1
0	0	1	0	0	1	0	1	0

(חזרה ל-feature representation)

נספח 2.5 – צילומי מסך של ה-features שמופיעים בסט הנתונים בסוף שלב ה-feature selection

(חזרה ל-feature selection)

index	email_verified	blue_tick	number_of_previous_messages	number_of_followers	number_of_follows	sum_top_common_negative_words	top_common_negative_words_percentage	sum_top_common_positive_words	top_common_positive_words_percentage
0	0	0	0.244898	0.897959	0.285714	0.125	0.0909091	0.111111	0.0909091
1	1	1	0.979592	0.367347	0.857143	0	0	0.111111	0.0416667
2	1	1	0.612245	0.306122	0.489796	0	0	0	0
3	0	0	0.102041	0.0204082	0.0816327	0	0	0	0
4	0	0	0.163265	0.0408163	0.122449	0	0	0	0
5	1	0	0.387755	0.877551	0.877551	0	0	0.111111	0.142857
6	0	0	0.387755	0.816327	0.0816327	0	0	0.222222	0.105263
7	0	0	0.102041	0.0816327	0.0408163	0	0	0	0
8	0	0	0.142857	0.0612245	0.0204082	0	0	0	0
9	1	0	0.0816327	0.183673	0.0612245	0.125	0.166667	0	0
10	1	0	0.22449	0.632653	0.530612	0.125	0.0714286	0.333333	0.214286

M	instagram	x	mp4	message_date_year_2022	message_date_month_1	message_date_month_3	message_date_month_5	message_date_month_6	message_date_month_10	message_date_month_11	email_suffix_com
0	0	0	1	0	0	0	0	0	1	0	0
1	0	0	0	1	0	0	0	0	0	1	0
0	0	0	0	1	0	0	0	1	0	0	0
1	0	0	1	1	1	0	0	0	0	0	0
1	0	0	0	1	0	0	1	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	1	0	0	0	0	0	1	0
1	0	1	0	0	0	0	0	1	0	0	0
0	1	0	0	0	0	0	0	0	0	0	0
1	1	0	0	0	0	0	0	1	0	0	0
0	0	0	1	0	0	0	0	0	0	1	0

email_suffix_edu	email_suffix_il	account_creation_year_2015	account_creation_month_1	account_creation_month_2	account_creation_month_3	account_creation_month_5	account_creation_month_10	account_creation_month_11	account_creation_month_12
0	0	1	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
0	0	1	0	0	0	0	1	0	0
0	1	0	0	0	0	0	0	0	1
0	0	0	0	0	0	0	0	0	0
0	1	0	0	0	0	0	0	0	1
0	0	0	0	1	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
1	0	0	1	0	0	0	0	0	0
0	1	0	0	0	0	0	0	0	1

message_date_0-7	message_date_16-23	message_date_8-15	account_creation_0-7
1	0	0	1
1	0	0	0
1	0	0	1
0	0	1	0
0	1	0	0
1	0	0	1
1	0	0	1
0	0	1	0
0	0	1	1
0	0	1	0
0	0	1	0

נספח 2.6 – צילומי מסך של ה-features שמופיעים בסט הנתונים בסוף שלב ה- dimensionality reduction

(חזרה ל-Dimensionality reduction)

index	email_verified	blue_tick	number_of_previous_messages	number_of_followers	number_of_follows	sum_top_common_negative_words	top_common_negative_words_percentage	sum_top_common_positive_words	positive_worc	M
0	0	0	0.244898	0.897959	0.285714	0.125	0.0909091	0.111111	0.0909091	0
1	1	1	0.979592	0.367347	0.857143	0	0	0.111111	0.0416667	1
2	1	1	0.612245	0.306122	0.489796	0	0	0	0	0
3	0	0	0.102041	0.0204082	0.0816327	0	0	0	0	1
4	0	0	0.163265	0.0408163	0.122449	0	0	0	0	1
5	1	0	0.387755	0.877551	0.877551	0	0	0.111111	0.142857	0
6	0	0	0.387755	0.816327	0.0816327	0	0	0.222222	0.105263	0
7	0	0	0.102041	0.0816327	0.0408163	0	0	0	0	1
8	0	0	0.142857	0.0612245	0.0204082	0	0	0	0	0
9	1	0	0.0816327	0.183673	0.0612245	0.125	0.166667	0	0	1
10	1	0	0.22449	0.632653	0.530612	0.125	0.0714286	0.333333	0.214286	0

top_common_positive_words_percentage	M	instagram	x	mp4	message_date_year_2022	message_date_month_1	message_date_month_3	message_date_month_5	message_date_month_6	message_date_month_10
0.0909091	0	0	0	1	0	0	0	0	0	1
0.0416667	1	0	0	0	1	0	0	0	0	0
0	0	0	0	0	1	0	0	0	1	0
0	1	0	0	1	1	1	0	0	0	0
0	1	0	0	0	1	0	0	1	0	0
0.142857	0	0	0	0	0	0	0	0	0	0
0.105263	0	0	0	0	1	0	0	0	0	0
0	1	0	1	0	0	0	0	0	1	0
0	0	1	0	0	0	0	0	0	0	0
0	1	1	0	0	0	0	0	0	1	0
0.214286	0	0	0	1	0	0	0	0	0	0

message_date_month_11	email_suffix_com	email_suffix_edu	email_suffix_il	account_creation_year_2015	account_creation_month_1	account_creation_month_2
0	0	0	0	1	0	0
1	0	0	0	0	0	0
0	0	0	0	0	0	0
0	0	0	0	1	0	0
0	0	0	1	0	0	0
0	0	0	0	0	0	0
1	0	0	1	0	0	0
0	0	0	0	0	0	1
0	0	0	0	0	0	0
0	0	1	0	0	1	0
1	0	0	1	0	0	0

account_creation_month_3	account_creation_month_5	account_creation_month_10	account_creation_month_11	account_creation_month_12	message_date_0-7	message_date_16-23	message_date_8-15
0	0	0	0	0	1	0	0
0	0	0	0	0	1	0	0
0	0	0	0	0	1	0	0
0	0	1	0	0	0	0	1
0	0	0	0	1	0	1	0
0	0	0	0	0	1	0	0
0	0	0	0	1	1	0	0
0	0	0	0	0	0	0	1
0	0	0	0	0	0	0	1
0	0	0	0	0	0	0	1
0	0	0	0	1	0	0	1

account_creation_0-7	sentiment
1	1
0	1
1	1
0	0
0	0
1	1
1	1
0	0
1	0
0	0
0	1

(חזרה ל-Dimensionality reduction)