**למידת מכונה**

פרויקט הקורס

חלק א'

קבוצה 19 - ספיר גבאי 209516574, רוני קרוצ'קוב 211899034



**תוכן עניינים**

Data Collection and Sensing .........................................................................3

Pre-processing ..............................................................................................3

EDA ..............................................................................................................4

Feature extraction ......................................................................................8

Feature representation ................................................................................11

Feature selection .........................................................................................12

Dimensionality reduction .............................................................................12

Validation ....................................................................................................12

**קישור לColab Notebook**

<https://colab.research.google.com/drive/18mmOG0A0sjYCu1T8C0Hp0W7vqeZFfe4L?usp=sharing>

**קישור לGIT**

<https://github.com/SapirGabay/ML_Project_Part1>

**עיבוד הנתונים** - **Data Collection and Sensing**

Data collection הוא שלב איסוף הנתונים מן העולם האמיתי, במטרה ליצור ייצוג שממנו ניתן ללמוד. הנתונים כוללים דגימות שלמות ומובחנות, מתוך התחום הנחקר.

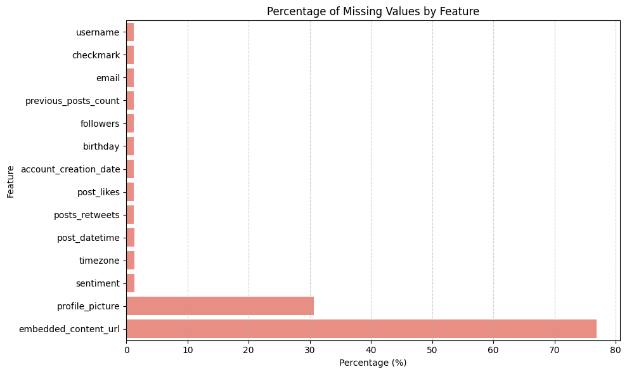
סוג ה-Sensing שבוצע בפרויקט הוא סטטי, שכן הנתונים נאספו בנקודת זמן אחת (למשל, פוסטים ברשת החברתית כפי שהיו ברגע מסוים), ואינם משתנים בזמן אמת.   
שלב ה-Sensing מתאר את הליך איסוף הנתונים ע"י מדידה ותיעוד של פרמטרים מסוימים, ומהווה את הגשר בין העולם הפיזי לבין יצירת ה-Data.

סוג Sensing נוסף, שלא בוצע על הדאטה, הוא Sensing דינמי, המתמקד במדידה לאורך זמן של חשיפה לפוסט, כגון צפיות או קצב תגובות. שילוב מידע זה עשוי היה לאפשר זיהוי דפוסים התנהגותיים ולהעמיק את ניתוח הסנטימנט.

משימת הלמידה המרכזית בפרויקט היא סיווג (Classification) במסגרת למידה מונחית (Supervised Learning). והמערכת לומדת מתוך תוויות קיימות על מנת לחזות תוויות חדשות. בקובץ הנתונים כל רשומה מייצגת תוכן שפורסם ברשת X אשר מתויג לפי סנטימנט (חיובי או שלילי), והמטרה היא לאמן מודל שמסווג תכנים חדשים בסיס תוכנם.

משימה נוספת שניתן לבצע על הדאטה היא Anomaly Detection, לדוגמה: זיהוי פוסטים עם כמות לייקים או קצב תגובות חריגים, שעשויים להעיד על פעילות לא רגילה (כגון שימוש בבוטים). משימה זו משתייכת לקטגוריית Unsupervised Learning, ומתמקדת בזיהוי דוגמאות החורגות מהתנהגות רגילה, גם בהיעדר תוויות מוגדרות מראש.

**Dataset Creation**

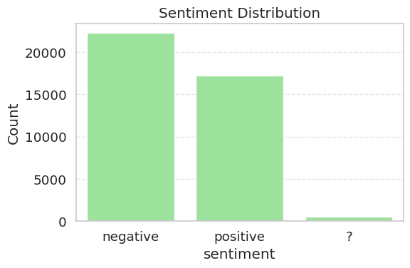
**Pre-processing**

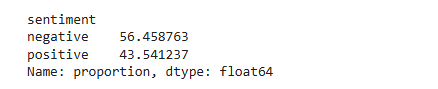
בשלב זה נבחן היקף הערכים החסרים עבור כל אחד מהמשתנים בסט הנתונים. נמצא כי המשתנה embedded\_content\_url (תוכן מצורף) חסר בכ-70% מהרשומות. מאחר שתגובות ברשתות חברתיות לרוב אינן כוללות קבצים מצורפים, הוחלט שלא להסיר את המשתנה, אלא להמירו למשתנה בינארי המצביע על קיומו או היעדרו של תוכן מצורף.

גם המשתנה profile\_picture\_url הכיל שיעור גבוה של חוסרים, אך מאחר שהיעדר תמונת פרופיל עשוי להעיד על מאפייני המשתמש, הוא הומר למשתנה בינארי לציון קיומה.   
יתר המשתנים הכילו שיעורי חוסר נמוכים בלבד, ולא דרשו התאמות מיוחדות.

בוצע הסרה של רשומות בהן ערך הסנטימנט (sentiment) חסר, מאחר שמדובר במשתנה המטרה, והיעדרו עלול להטות את המודל. בנוסף הוסרו רשומות עם למעלה מ3 ערכים חסרים, מתוך הנחה שמדובר ברשומות חלקיות שעשויות לפגוע באיכות המודל ולהכניס רעש לנתונים. תהליכי הניקוי בוצעו על עותק נפרד כדי לשמור על הקובץ המקורי ללא שינוי.

**השלמת ערכים חסרים בהתאם לסוג המשתנה:  
משתנים מספריים** (לייקים, שיתופים, מספר עוקבים, מספר פוסטים קודם): הושלמו לפי ערך החציון, במטרה לצמצם השפעת ערכים קיצוניים. **משתנים קטגוריאליים** (סוג התגובה, סטטוס אימות, אזור זמן): הושלמו לפי הערך השכיח, לשם שמירה על מבנה ההתפלגות. נבחנה גם אפשרות להשלמה רנדומלית בהתאם להתפלגות הקיימת. **משתנים ייחודיים** (ID, שם משתמש, כתובת אימייל): לא הושלמו, מאחר שהם מזהים פרטניים שאינם תורמים לתהליך הלמידה או להבנת הדאטה. **משתנים בינאריים** (תמונת פרופיל, תוכן מצורף): הומרו לבינאריים לציון קיום הערך. **משתנים מסוג תאריך** (תאריך לידה, תאריך פרסום ויצירת חשבון) לא הושלמו, מתוך הנחה שלא ניתן לשחזרם באופן אמין. בהתאם למדיניות, הוסרו רשומות עם חוסרים מרובים.

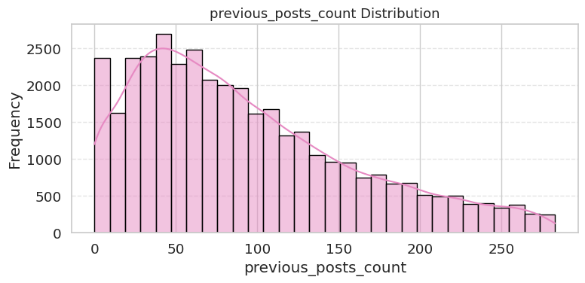
 בדקנו את ההתפלגות היחסית של הרשומות בין ערכי משתנה המטרה, ונמצא פער מתון (56% שלילי, 43% חיובי) אך ללא הטיה מהותית. לכן, לא ביצענו איזון מלאכותי בדאטה.

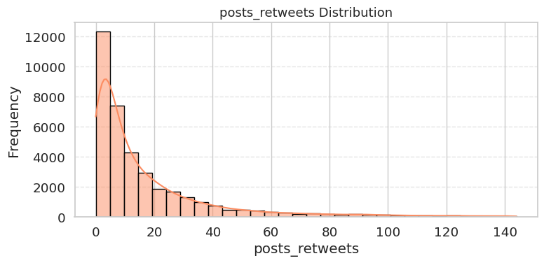
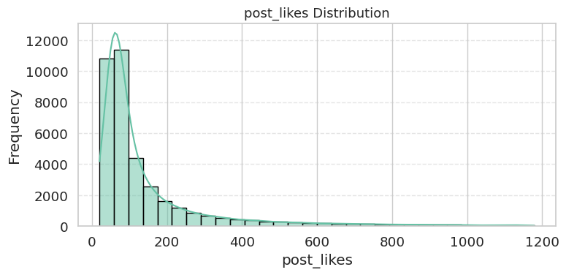


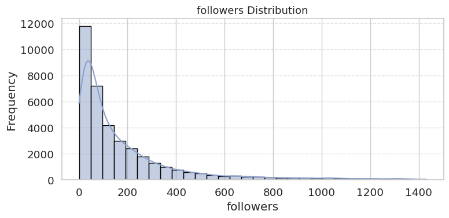
**Exploratory data analysis – EDA**

החלטנו לבצע את חלק זה לאחר שלב ההכנה הראשוני של הדאטה על מנת לקבל תמונה יותר ברורה ונכונה של התפלגות הנתונים. בחרנו שלא להציג התפלגות עבור מזהים ייחודיים או פרטים שאינם נותנים ערך אנליטי בהתפלגות (מייל, שם משתמש ו ID).

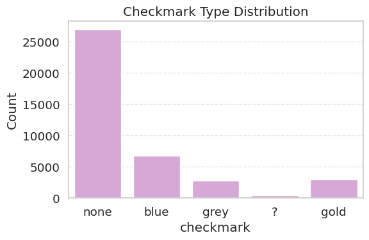
בחרנו להציג משתנים מספריים באמצעות היסטוגרמה. כדי למנוע השפעה לא פרופורציונלית של פוסטים חריגים, נבחר להציג ולנתח את הפיצ'רים תוך חיתוך ב-95%.

** previous\_posts\_count-** התפלגות מוטה לימין, מה שמעיד על קבוצה קטנה יחסית של משתמשים פעילים מאוד. עשוי להיות תלוי בותק החשבון.

**Post\_likes -** מייצג את מספר הלייקים שקיבל כל פוסט. נצפתה התפלגות מוטה לימין – מרבית הפוסטים מקבלים כמות נמוכה של לייקים, ומעט פוסטים זוכים לפופולריות חריגה. מספר לייקים עשוי להעיד על מידת המעורבות בפוסט. ייתכן וקיים גם קשר ישיר למספר העוקבים.

**Posts\_retweets -** ההתפלגות אינה סימטרית ומציגה זנב ארוך כלפי ימין, מרבית הפוסטים כמעט לא משותפים, ומעטים זוכים לשיתופים רבים.פוסט שמשותף רבות עשוי להעיד על עניין רגשי גבוה, חיובי או שלילי.ייתכן ומתואם עם מספר לייקים או רמת המעורבות הכללית בפוסט.

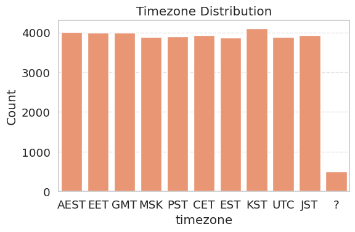
**Followers -** מתפלג בצורה מוטה מאוד לימין, רוב המשתמשים בעלי מספר נמוך של עוקבים, ומיעוטם בעלי אלפי עוקבים. עשוי להשפיע ישירות על כמות הלייקים והשיתופים, ולכן בהמשך יננצר משתנה מנורמל על פי היחס.

בחרנו להציג משתנים קטגוריאליים בתרשים של גרף עמודות -

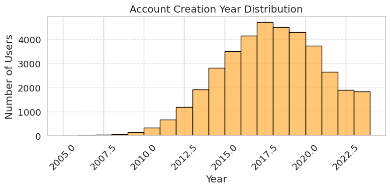
**A graph of a distribution

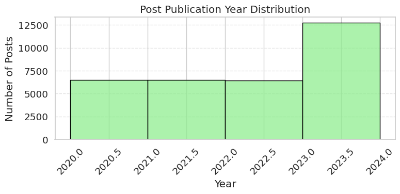
AI-generated content may be incorrect.Checkmark** - מייצג את סוג החשבון של המשתמש. מרבית המשתמשים הם בעלי חשבון פרטי ללא מנוי (none) , עשוי לרמז על רמת האמינות או ההשפעה של המשתמש (בהתאם לסוג החשבון – ארגון, ממשלה או פרטי), מה שיכול להיות רלוונטי להשפעה על סוג הסנטימנט.

**Type -** ניתן לראות שרוב הנתונים נאספו מתוך תגובות, ריבוי התגובות (reply) מעיד שהשיח ברשת מתמקד בעיקר בהבעת עמדה על תוכן קיים ולא ביוזמה של פוסטים מקוריים.



**Timezone –** ההתפלגות מציגה פיזור אחיד יחסית של המשתמשים בין אזורי הזמן השונים, מה שמרמז על כך שהדאטה נאסף ממגוון גאוגרפי רחב. מעניין לבדוק קשר לסנטימנט.

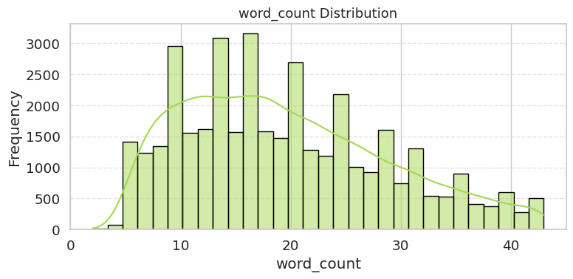
**account\_creation\_date -** מרבית המשתמשים פתחו את החשבון בין השנים 2015 ל־2020, מה שמעיד על שימוש גובר בפלטפורמה בשנים האחרונות. מעניין לבדוק קשר בין ותק החשבון למשתנים נוספים.

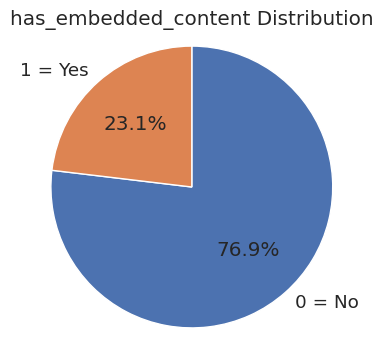


A graph showing the number of birth years

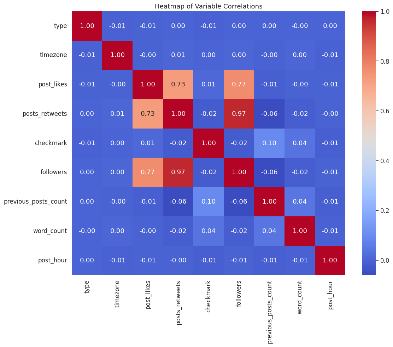
AI-generated content may be incorrect.**birthday –** מרבית המשתמשים במדגם נולדו בשנות ה90 מה שמעיד על הגיל הפופולרי בקרב המשתמשים. (סביבות גיל 30-40).

**post\_datetime -** בשנים 2020 עד 2022 הדאטה מאוזן - מספר הפוסטים בכל שנה דומה יחסית. החל מ-2023 ניכרת עלייה חדה בכמות הפוסטים, מה שמעיד על שינוי בקצב הפעילות או בהיקף הדגימה.

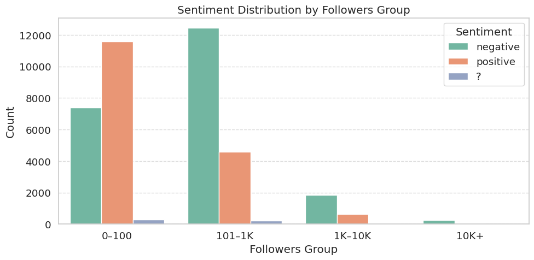
ישנם מספר משתנים בData שאינם ניתנים להצגה באופן ישיר באמצעות גרפים, ונדרש לבצע עבורם המרה על מנת שניתן יהיה לנתחם / להציגם. התייחסות לכלל ההמרות והפיצ'רים שהופקו מופיעה בהמשך בדוח.

A pie chart with numbers and a blue circle with Crust in the background

AI-generated content may be incorrect.**Text** - לדוגמה, הצגה של טקסטים עצמם לא יתרום לניתוח סטטיסטי או חזותי. במקום זאת בחרנו להמיר למספר פיצ'רים שניתנים להצגה. בניתוח אורך הטקסטים (word count) מתקבלת התפלגות עם זנב ימני – כלומר, רוב הפוסטים קצרים יחסית.

משתנים בינאריים שהופקו: העמודות profile\_picture\_url ו־embedded\_content\_url הומרו למשתנים בינאריים המציינים את עצם קיום הערך.  
  
**גרפים שמציגים קשר לסנטימנט / קורולציה בין משתנים שונים:**   
בחלק זה התמקדנו בזיהוי קשרים בין משתנים לבין משתנה המטרה (sentiment) באמצעות מגוון גרפים. השתמשנו בגרף עמודות להצגת התפלגויות קטגוריאליות לפי סנטימנט, גרף קווים לזיהוי מגמות לאורך ציר רציף, ומפת חום לזיהוי מתאמים בין משתנים מספריים.  
  
ניתוח הקורלציה בין הפיצ’רים מראה קשר חזק בין מספר העוקבים, לייקים מספר השיתופים, שלושתם משתנים המייצגים מדדי חשיפה והשפעה. זיהוי הקשרים הללו מאפשר להבין כי קיימת חפיפה בין מדדים מסוימים, ולכן ניתן לשקול ביצוע הפחתת ממד או בחירה בפיצ’ר מייצג אחד בלבד. שאר הקשרים חלשים מאוד.

A graph with different colored bars

AI-generated content may be incorrect.A graph of a number of posts

AI-generated content may be incorrect.ניתוח לפי קבוצות עוקבים מראה שמרבית המשתמשים עם מאה עד אלף עוקבים מביעים סנטימנט שלילי, בעוד שמעל/מתחת לרף זה ניכרת ירידה חדה בהיקף הפוסטים השליליים, מה שעשוי לרמז על קשר בין רמת חשיפה לבין נכונות לחשיפה רגשית שלילית. לעומת זאת, חלוקה לפי כמות פוסטים קודמים מצביעה על כך שמשתמשים פעילים במיוחד נוטים לסנטימנט חיובי, ייתכן בשל היכרות גבוהה יותר עם הפלטפורמה. בגרף של מספר הלייקים נצפית מגמה שפוסטים בעלי סנטימנט שלילי מרוכזים בפופולריות נמוכה (מתחת ל100 לייקים), ממצא זה עשוי להעיד על כך שפוסטים שליליים מעוררים פחות עניין או שאנשים נוטים להפגין פחות תמיכה בפוסטים שליליים בפומבי ולכן אינם זוכים להפצה נרחבת.

A graph with a red line and green line

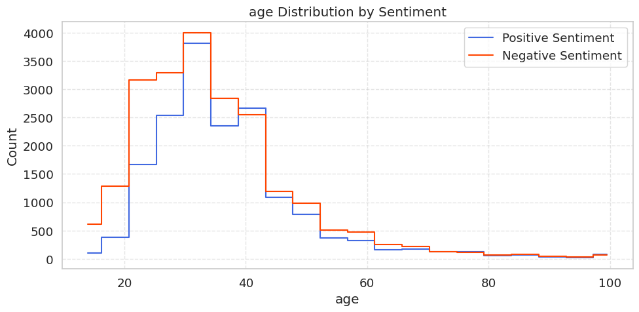
AI-generated content may be incorrect.A graph showing the growth of an age

AI-generated content may be incorrect.על פי ההתפלגות לפי גיל ניתן לראות כי סנטימנט שלילי מרוכז בעיקר בגילאי 25–35, בעוד שסנטימנט חיובי מתפזר יותר לגילאים מבוגרים. בשני המקרים ניכרת ירידה בתדירות הפוסטים עם העלייה בגיל. בגרף השני המציג את התפלגות אורך הטקסט, ניתן לראות כי פוסטים שליליים בדרך כלל קצרים, יכול להעיד על סגנון תמציתי יותר בתוכן שלילי.

**Feature extraction**

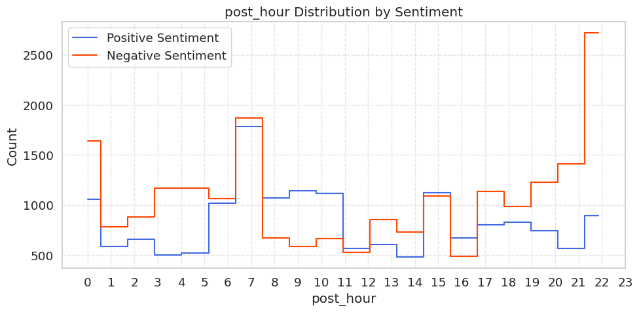
בשלב זה יצרנו פיצ'רים חדשים על בסיס עמודות קיימות, במטרה לחלץ מידע שימושי מתוך הנתונים הגולמיים ולשפר את ייצוג הדאטה עבור משימת הלמידה. מאחר והנתונים נאספו לאחר פרסום הפוסט בזמן קבוע (למשל חודש לאחר הפרסום), לא נדרשנו לשקלל את מרכיב הזמן – ולכן בחרנו לפעול לפי שיטות ללא חשיבות בזמן, תוך הקפדה על מספר קבוע של פיצ’רים לכל רשומה. תהליך ההפקה בוצע בגישה מבוססת ידע (Knowledge-Based), תוך התבססות על הבנה של ההקשר החברתי והתנהגות המשתמשים ברשת. לכל פיצ'ר חדש בדקנו את ההתפלגות של הסנטימנט (חיובי/שלילי) באמצעות גרף היסטוגרמה, במטרה לבדוק האם קיימת הבחנה מובהקת שעשויה לסייע במודל. בנוסף, בדקנו גם שילובים של פיצ'רים זוגיים באמצעות גרף פיזור (scatter) עם קו רגרסיה, כדי לזהות מגמות חזותיות בין המאפיינים.

|  |  |  |
| --- | --- | --- |
| **שם הפיצ'ר** | **תיאור** | **מקור** |
| has\_profile\_picture | עמודה בינארית - האם קיימת תמונת פרופיל | profile\_picture\_url |
| has\_embedded\_content | עמודה בינארית - האם צורף קובץ | embedded\_content\_url |
| content\_type | סוג התוכן שצורף לפוסט, במידה וקיים (תמונה /וידיאו/ אחר) | embedded\_content\_url |
| age | גיל המשתמש | birthday |
| post\_hour | שעה ביום | post\_datetime |
| post\_day | יום בשבוע | post\_datetime |
| post\_month | חודש הפרסום | post\_datetime |
| account\_age | ותק החשבון ביום הפרסום (הפרש השנים בין יצירת החשבון ותאריך פרסום הפוסט) | account\_creation\_date + post\_datetime |
| word\_count | מספר המילים בכל טקסט | text |
| num\_hashtags | כמה פעמים מופיע # | text |
| likes\_per\_follower | יחס בין לייקים לעוקבים (normalize) | post\_likes + followers |
| posts\_per\_year\_avg | מחשב את קצב הפעילות השנתי של המשתמש, כלומר כמה פוסטים בממוצע המשתמש פרסם בשנה | previous\_posts\_count + account\_age |
| engagement\_score | מודד את רמת המעורבות בפוסט על ידי חיבור של מספר הלייקים והשיתופים שקיבל | post\_likes + posts\_retweets |
| is\_weekend | משתנה בוליאני האם הפוסט התפרסם בסופש | post\_datetime |
| post\_time\_bucket | שלב ביום – בוקר / צהריים / ערב / לילה | post\_hour |
| is\_verified | בוליאני – האם החשבון מאומת | checkmark |
| positive\_word\_count | ספירת המילים החיוביות (מתוך קובץpositive words) | text |
| negative\_word\_count | ספירת המילים השליליות (מתוך קובץnegative words) | text |

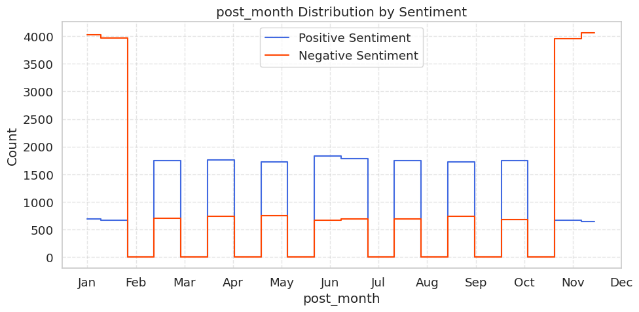
בהתבסס על גרפי ההתפלגות לפי סנטימנט, ניתן לזהות מספר פיצ'רים בעלי יכולת הבחנה בין תגובות חיוביות לשליליות:

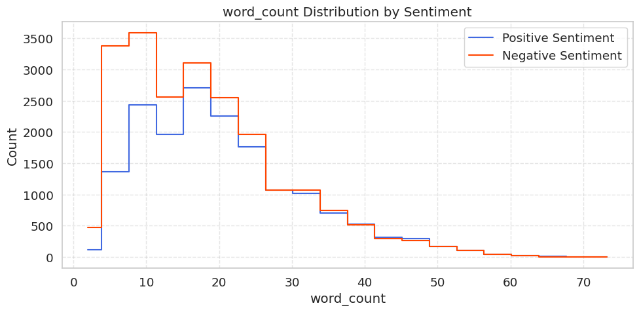
A graph showing a number of numbers

AI-generated content may be incorrect.**גיל המשתמש:** כפי ששמנו לב בשלב הEDA ניכרת ריכוזיות גבוהה של פוסטים שליליים בקרב משתמשים בגילאים צעירים. הפיצ’ר מציג הבחנה סבירה, וניתן לקבצו לשתי קטגוריות (עד גיל 30 / מעל גיל 30).

**גיל החשבון:** פוסטים שליליים שכיחים יותר בקרב משתמשים חדשים (0–5 שנים), לעומת חלוקה מאוזנת יותר בקרב משתמשים ותיקים. מדובר בהבחנה מתונה, שעשויה לשקף רמת מעורבות או ניסיון בשימוש בפלטפורמה.

**שעת הפרסום:** בין השעות 8:00–11:00 נצפה רוב לפוסטים חיוביים, בעוד שבשעות הלילה והערב (17:00–5:00) מופיעה עלייה בפוסטים שליליים. ייתכן קשר בין זמן הפעילות למצב רגשי של המשתמש.

**חודש הפרסום:** בתקופה שבין נובמבר לפברואר נרשמה עלייה ניכרת בפוסטים שליליים. קיימת הבחנה עונתית שעשויה להיות רלוונטית לתחזיות סנטימנט.



**אורך הפוסט:** תגובות שליליות מרוכזות בטווחים קצרים (עד 20 מילים), לעומת תגובות חיוביות שמופיעות יותר באורכים בינוניים וארוכים – דבר המצביע על סגנון ביטוי תמציתי יותר של תוכן שלילי.

A graph with a number of hashtags

AI-generated content may be incorrect. **מספר האשטגים:** פיצ’ר בעל הבחנה ניכרת – פוסטים שליליים כמעט ואינם כוללים האשטגים, בעוד שפוסטים חיוביים כוללים לרוב שניים או יותר.

A graph with blue and orange squares

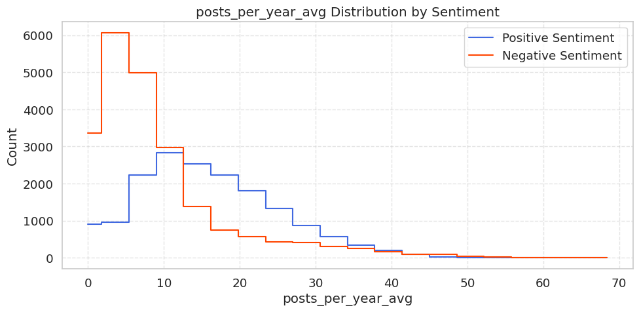
AI-generated content may be incorrect.

**תמונת פרופיל:** משתמשים ללא תמונת פרופיל נוטים לפרסם תוכן שלילי בשיעור גבוה יותר. הפיצ’ר עשוי לשקף רמת אנונימיות או פתיחות.

A graph with blue and orange squares

AI-generated content may be incorrect.A graph of different colored bars

AI-generated content may be incorrect. **סוג התוכן המצורף/ יום הפרסם (חול/סופש):** שני הפיצ'רים אינם מבחינים היטב בין סנטימנטים. סוג התוכן לא מציג הבדל מהותי בין חיובי לשלילי, ורוב הפוסטים ללא תוכן מצורף. גם יום הפרסום (חול/סופ"ש) לא מראה פערים ברורים ולכן שניהם פחות אפקטיביים.

 **קצב פרסום שנתי:** משתמשים עם תדירות פרסום נמוכה נוטים לפוסטים שליליים, בעוד שתדירות גבוהה מתואמת יותר עם תוכן חיובי. ממצא זה עשוי להעיד על מעורבות גבוהה כתורם לחיוביות.

**A graph of a number of bars

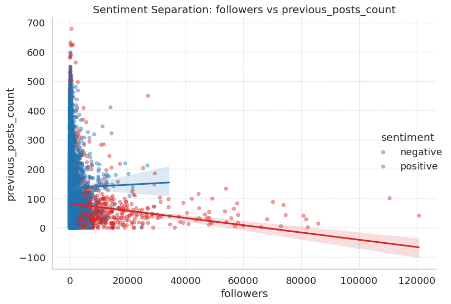
AI-generated content may be incorrect. שלב ביום:** מציג מגמה דומה לזו שנצפתה לפי שעת הפרסום, אך בחרנו לחלק את היממה לארבע קטגוריות – בוקר, צהריים, ערב ולילה – על מנת לפשט את הניתוח. גם כאן ניכרת מגמה שלילית בשעות הלילה, לעומת איזון בשעות היום.

A graph of different numbers and a number of negative words

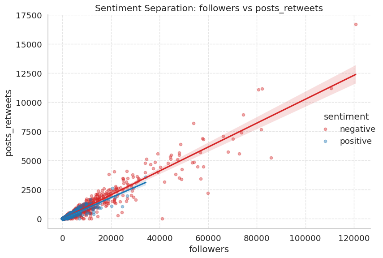
AI-generated content may be incorrect.

**התפלגות מספר מילים חיוביות / שליליות בפוסטים:**

ברוב הפוסטים מופיעות מעט מאוד מילים רגשיות (0-2), אך קיימת הבחנה ברורה: פוסטים חיוביים כוללים יותר מילים חיוביות, ופוסטים שליליים כוללים יותר מילים שליליות. הממצא תומך ביכולת הבחנה טובה של שני הפיצ'רים ומצדיק את שילובם במודל.

A graph of a red and blue line

AI-generated content may be incorrect.A graph of a red and blue line

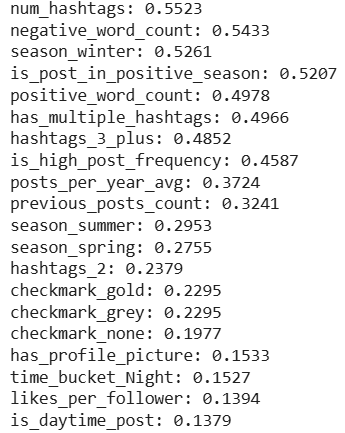
AI-generated content may be incorrect.גרפי הפיזור ממחישים בצורה חזותית את הקשרים שזוהו מוקדם יותר בין מספר העוקבים לבין מדדים כמו לייקים, שיתופים, פוסטים קודמים ומעורבות. ככל שמספר העוקבים גבוה יותר, כך עולה הסבירות לפוסטים חיוביים ולעוצמת אינטראקציה גבוהה יותר. פוסטים שליליים מרוכזים בעיקר בקרב משתמשים עם חשיפה נמוכה.

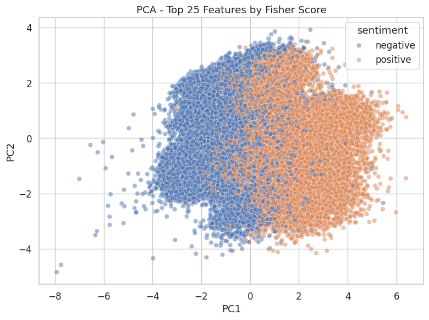
**Feature Representation (FR)**

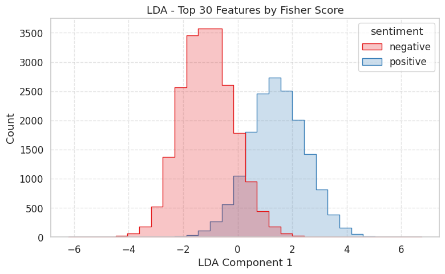
בשלב זה בחרנו כיצד לייצג את הפיצ'רים שהפקנו עבור כל רשומה בדאטה כך שיתאימו למשימת הלמידה. מאחר שחלק מהמשתנים התאפיינו בטווחי ערכים שונים מאוד (כגון מספר עוקבים באלפים לעומת מספר מילים בעשרות), בוצע נירמול לפיצ’רים המספריים באמצעות, Min-Max Normalization שיטה זו ממפה כל משתנה לסולם של 0 עד 1 ומונעת השפעות של סדרי גודל. הנרמול יושם על משתנים מספריים הכוללים: מספר לייקים, שיתופים, מספר עוקבים, מספר פוסטים קודם, גיל המשתמש, מספר מילים בפוסט, מדד מעורבות , קצב פרסום שנתי, וספירת מילים חיוביות ושליליות.

בהתאם לניתוח הנתונים שביצענו בשלבים הקודמים זיהינו מספר מגמות הקשורות לסנטימנט, בחרנו לייצר פיצ'רים בינאריים חדשים שיבטאו בצורה מדויקת תובנות שנצפו בהשוואה בין תגובות חיוביות לשליליות. כל אחד מהפיצ'רים הללו מתאר דפוס שהתגלה בשלבים המוקדמים של ניתוח הנתונים. ההמרה לערכים בוליאניים נועדה להבליט אינדיקציות מובהקות עבור המודל ולצמצם רעש שעלול לנבוע ממשתנים רציפים ומורכבים.

הפיצ'רים החדשים שנוספו כוללים: פרסום בשעות היום, עונת השנה, אורך הפוסט, תדירות פרסום, כמות האשטגים וגיל המשתמש. כל פיצ'ר הומר לערך בוליאני בהתאם לסף שנקבע מתוך ממצאי השלב הקודם במטרה להדגיש מגמות מובהקות ולפשט את הנתונים הרציפים.  
משתנים קטגוריאליים כגון שעת פרסום, סוג תוכן, עונה וסוג משתמש הומרו לייצוג בינארי באמצעות One-Hot Encoding, המייצר עמודה נפרדת לכל קטגוריה. שיטה זו מאפשרת שמירה על המבנה הקטגוריאלי של המשתנה מבלי להניח קשר או רציפות בין הערכים, ומאפשרת למודל להתייחס לכל קטגוריה באופן עצמאי.

**Features selection**  
בשלב זה בחרנו להתשתמש בגישה כמותית (Quantitative Evaluation), שמטרתה להעריך באופן אובייקטיבי את תרומתו של כל פיצר למשימת הסיווג. השתמשנו בגישת Filters באמצעות חישוב Fisher Score לכל פיצ'ר. המדד מבוסס על היחס בין ההפרש המוחלט בין ממוצע ערכי הפיצ'ר בדגימות חיוביות לעומת שליליות, לבין סכום סטיות התקן בשתי הקבוצות. פיצ'רים שקיבלו ציון גבוה יותר נחשבים ליותר מבדלים (Discriminative) בין הסנטימנטים. שיטה זו מאפשרת לדרג את כלל הפיצ'רים לפי תרומתם האפשרית למודל בצורה פשוטה, ללא תלות באלגוריתם למידה מסוים.  
בחרנו את ה30 פיצ'רים שמדורגים הכי גבוה מתוך 80 פיצ'רים שנבדקו.

**Dimensionality reduction**

בשלב זה ביצענו הפחתת מימדים בשתי שיטות שונות PCA ו LDA

שיטת PCA אפשרה למפות את הנתונים לשני רכיבים עיקריים תוך שימור מירבי של השונות, והציגה הבחנה נראית לעין בין הסנטימנטים – שליליים מרוכזים יותר בצד שמאל ו־חיוביים בצד ימין, עם חפיפה מסוימת. שיטת LDA יצרה הבחנה חדה יותר, עם הפרדה מובהקת בין הסנטימנטים לאורך רכיב אחד בלבד. ממצאים אלה מצביעים על תרומה משמעותית של הפיצ'רים שנבחרו ליכולת ההפרדה בין סנטימנטים, במיוחד לצורכי סיווג.

**Validation**

בשלב הוולידציה בחרנו להשתמש בשיטת Cross Validation מסוג K-Fold (עם K=5), שנחשבת לאחת מהשיטות האמינות להערכת ביצועי מודל. השיטה מאפשרת חלוקה של הדאטה לחמישה קיפולים (Folds), כך שבכל איטרציה מאומנים המודלים על ארבעה חלקים ונבחנים על החלק החמישי, לסירוגין. שיטה זו עדיפה על Holdout מאחר שהיא מצמצמת את ההשפעה של פיצול מקרי ומפיקה מדד ביצועים ממוצע שמייצג טוב יותר את איכות המודל על כלל הדאטה.

היתרונות המרכזיים של K-Fold הם הפחתת שונות, שימוש יעיל בדאטה והימנעות מהטיות שנובעות מפיצול חד-פעמי.

למדידת ביצועי המודל נשתמש במדדי דיוק (Accuracy) ובמקרה הצורך גם במדדים נוספים כמו Precision ו-Recall בהתאם למטרות הסיווג.