**פרוייקט קורס למידת מכונה**

**חלק א'**



**עמית אורן 315142190**

**עדן כהן 207696246**

**קבוצה 14**

תוכן עניינים

[Data collection and sensing: 3](#_Toc196600594)

[Data Creation: 4](#_Toc196600595)

[Exploratory data analysis: 4](#_Toc196600596)

[:Pre – Processing 10](#_Toc196600597)

[:Validation 13](#_Toc196600598)

[נספחים: 14](#_Toc196600599)

# Data collection and sensing:

1. **Data Collection** הוא תהליך איסוף מידע אשר מייצג את העולם האמיתי ומה שאנחנו רוצים ללמוד עליו. נשאף לדאטה מייצג שמביא לידי ביטוי מקרים שונים ודוגמאות להם, שיהיה אמין ומתוייג נכונה. במערך הנתונים הנוכחי נאספו פוסטים מהרשת החברתיתX . כל פוסט כולל את טקסט ההודעה, תאריך הפרסום, מספר הלייקים, מספר השיתופים, וכן מידע נוסף על המשתמש שפרסם אותו – כגון תאריך לידה, מספר עוקבים, סוג החשבון ועוד. בנוסף, קיים משתנה מטרה המייצג את הסנטימנט של הפוסט, המעיד על כך שמדובר בנתונים מתויגים .(Labeled Data) – יש לציין כי הדאטה שלם עבור כל דגימה, אך קיימים ערכים חסרים בעמודות מסוימות.

**Sensing** הוא תהליך בו אנו מקשרים ביו העולם האמיתי לעולם של הדאטה הגולמי. במהלך החישה אנו נבחר את צורת ואופן המדידה בצורה רציונלית, עקבית, ותואמת למשימת הלמידה שנרצה לבצע. נבחין בין שתי צורות חישה : סטטית ודינמית.

סטטית – מתייחסת לאיסוף של נתונים שאינם משתנים בתדירות גבוהה כלומר, נתונים שהם קבועים יחסית לאורך זמן.

דינמית - מתייחסת לאיסוף של נתונים שתלויים בזמן ולכן משתנים לאורכו.

במקרה שלנו מדובר ב- sensing סטטי כלומר יש נקודת זמן קבועה לכל פוסט ואין עדכון רציף. הנתונים נאספו לאחר חודש מהעלאת הפוסטים (כדי למדוד לייקים ושיתופים). בנוסף, הנתונים כוללים תצלום של מצב המשתמש (מספר עוקבים, מספר פוסטים קודמים) ביום פרסום הפוסט – לא משתנה דינמית לאחר מכן.

1. ניתן לבצע sensing דינאמי כדי לעקוב אחר כמות הלייקים בכל שעה ביום החל מפרסום הפוסט (לעומת מספר הלייקים חודש לאחר הפרסום). כך ניתן ללמוד מתי פוסט מתחיל להיות ויראלי, איך זמן פרסום משפיע על המעורבות, ומה הקשר בין תוכן רגשי להתפשטות מהירה.
2. בבסיס הנתונים שברשותנו ישנן כ־40,000 דגימות של פוסטים מרשת חברתית, כאשר כל פוסט מתויג כסנטימנט חיובי או שלילי – כלומר, קיימת תווית (label) ברורה אשר מייצגת את משתנה המטרה. כ-550 דגימות בלבד אינן כוללות תיוג, ולכן נשקול כיצד לטפל בדגימות אלו כדי לאמן את המודל על דאטה מתוייג בלבד. בהתאם לכך, מדובר ב־משימת Classification (סיווג), שמטרתה לשייך כל פוסט לקטגוריה אחת מתוך שתי קטגוריות (חיובי או שלילי). משימה זו משתייכת לקטגוריית Supervised Learning, שכן קיים משתנה מטרה ידוע המשמש את המודל בזמן האימון והדאטה הגיע מתוייג. בנוסף, ניתן להשתמש בנתונים ולבצע למשל clustering ולקבץ דגימות לפי מאפיינים כמו אורך פוסט, סוג פוסט, תדירות מילים וכו' כדי לזהות קבוצות דומות (לדוגמא פוסטים אישיים, מקצועיים ועוד).

# Data Creation:

## Exploratory data analysis:

**תמונה שמכילה טקסט, צילום מסך, תרשים, קו

תוכן שנוצר על-ידי בינה מלאכותית עשוי להיות שגוי.תמונה שמכילה טקסט, צילום מסך, תרשים, עלילה

תוכן שנוצר על-ידי בינה מלאכותית עשוי להיות שגוי.תמונה שמכילה תרשים, טקסט, עלילה, קו

תוכן שנוצר על-ידי בינה מלאכותית עשוי להיות שגוי.**תמונה שמכילה תרשים, טקסט, עלילה, צילום מסך

תוכן שנוצר על-ידי בינה מלאכותית עשוי להיות שגוי.**משתנים נומריים:**\*נדגיש כי על מנת להעריך את ההתפלגות עבור המשתנים הנומריים הוגדר טווח ממוקד בצורה מושכלת לכל גרף על מנת להפחית השפעה של ערכים קיצוניים ולבחון את התפלגות המשתנה בצורה ברורה. (בנספחים ניתן לראות את הגרפים כפי שהם ללא הגדרת טווח)  
**post\_likes:**  
ההתפלגות של מספר הלייקים היא **מוטה לימין** עם **skewness חיובי** ועם **גבנוניות חיובית**, המאופיינת בריכוז גבוה של ערכים נמוכים, פסגה חדה וזנב ארוך כלפי ערכים גבוהים. הממוצע עומד על כ־350 לייקים, אך הערכים נעים בטווח רחב מאוד — מ־20 ועד 29,508 — דבר המצביע על שונות גבוהה ומיעוט פוסטים עם מספר לייקים גבוה, אשר מושכים את ההתפלגות כלפי מעלה.   
**posrt\_retweets:**  
ההתפלגות של מספר הריטוויטים היא **מוטה לימין** עם **skewness חיובי** ו**גבנוניות חיובית**, ומתאפיינת בריכוז גבוה של ערכים נמוכים וזנב ארוך כלפי מעלה. הממוצע עומד על כ־49 ריטוויטים, אך טווח הערכים רחב – מ־0 ועד 16,679 – דבר שמצביע על שונות גבוהה ומיעוט פוסטים שמשותפים מחדש.  
**followers:**  
ההתפלגות של מספר העוקבים היא **מוטה לימין** עם  **skewness חיובי** ו**גבנוניות חיובית**, ומתאפיינת בריכוז גבוה של משתמשים עם מספר עוקבים נמוך, לצד זנב ארוך המצביע על מיעוט עם מספר עוקבים גבוה במיוחד. הממוצע עומד על כ־488 עוקבים, אך טווח הערכים רחב מאוד – מ־0 ועד 120,692 – מה שמעיד על שונות גבוהה ופערים חדים בין רוב המשתמשים לבין קבוצה קטנה ובולטת. **previous post count :**  
ההתפלגות של מספר הפוסטים הקודמים היא **מוטה לימין**, עם **skewness חיובי מתון**. היא מתאפיינת בריכוז גבוה של משתמשים בעלי מספר פוסטים נמוך עד בינוני, אך בשונה מההתפלגויות שנראו בנתונים אחרים, כאן קיימת פסגה רחבה יותר והמעבר לערכים גבוהים הוא הדרגתי יותר.  
הממוצע עומד על כ־105 פוסטים, וטווח הערכים נע בין 0 ל־679, דבר המצביע על שונות גבוהה אך פחות קיצונית ביחס למשתנים אחרים, וייתכן שמעיד על פעילות עקבית יותר של מרבית המשתמשים.

**תמונה שמכילה צילום מסך, טקסט, ריבוע, צבעוני

תוכן שנוצר על-ידי בינה מלאכותית עשוי להיות שגוי.בדיקת קורולציה בין המשתנים:**

מהמטריצה עולה כי קיימת קורלציה חזקה מאוד בין מספר העוקבים לריטוויטים (0.97), וקורלציה גבוהה גם בין עוקבים ללייקים (0.77), מה שמעיד על השפעה ברורה של כמות העוקבים על אינטראקציות בפוסטים. קשרים אלו יכולים לספק עניין עקיף על הסנטימנט – למשל, פוסטים חיוביים עשויים למשוך יותר תגובות ממשתמשים בעלי קהל רחב. עם זאת, המשתנה של מספר הפוסטים הקודמים אינו נמצא בקורלציה עם שאר המדדים. ממצאים אלו יכולים לסייע בבחירת מאפיינים רלוונטיים יותר בהמשך.

מכיוון שמשתנה הפוסטים הקודמים איננו בעל קורלציה גבוהה עם שאר המשתנים, נרצה לבדוק את השפעתו הספציפית על משתנה המטרה - סנטימנט

**בדיקת כמות פוסטים קודמים לפי סנטימנט:**

תמונה שמכילה טקסט, צילום מסך, תרשים, עלילה

תוכן שנוצר על-ידי בינה מלאכותית עשוי להיות שגוי.מהגרף עולה כי רוב הפוסטים נכתבו על ידי משתמשים עם מעט פוסטים קודמים, ויותר פוסטים שליליים נכתבו על ידי משתמשים חדשים או פחות פעילים. ככל שמספר הפוסטים הקודמים עולה, יש נטייה לסנטימנט חיובי יותר. ממוצע הפוסטים הקודמים של משתמשים עם סנטימנט חיובי גבוה יותר (135.7) מאשר של משתמשים עם סנטימנט שלילי (81.6). מסקנה אפשרית היא שמשתמשים ותיקים יותר נוטים לכתוב תוכן חיובי, בעוד שמשתמשים חדשים או פחות פעילים נוטים להראות תסכול דרך פוסטים שליליים.

**משתנים קטגוריילים**

**Type:**

תמונה שמכילה טקסט, צילום מסך, תרשים, עלילה

תוכן שנוצר על-ידי בינה מלאכותית עשוי להיות שגוי.משתנה המציין את סוג הפוסט  
נבחן את סנטימנט הפוסט לפי סוגו. ניתן לראות כי קיימת נטייה לתוכן שלילי יותר מאשר חיובי בכל סוגי הפוסטים. התגובות (replies) מהוות את הסוג הנפוץ ביותר של אינטראקציה עם הפער המשמעותי ביותר בין תוכן שלילי לחיובי, בעוד שציוצים חוזרים (retweets) הם הנדירים ביותר. יש יחס של כ-20%-30% יותר תוכן שלילי מחיובי שנשמר באופן עקבי בכל הקטגוריות, מה שמעיד על נטייה כללית של משתמשים להגיב, לשתף ולפרסם תוכן בעל אופי שלילי יותר מאשר תוכן חיובי ברשתות החברתיות.

**תמונה שמכילה טקסט, צילום מסך, גופן, תרשים

תוכן שנוצר על-ידי בינה מלאכותית עשוי להיות שגוי.sentiment:**

מדובר במשתנה המטרה אשר מייצג ערך סנטימנט שלילי או חיובי לפוסט. ניתן להבין כי ההתפלגות נחשבת לכמעט מאוזנת כשאר יש 56.5% של תגובות שליליות ואילו 43.5% של תגובות חיוביות.

תמונה שמכילה טקסט, צילום מסך, תרשים, עלילה

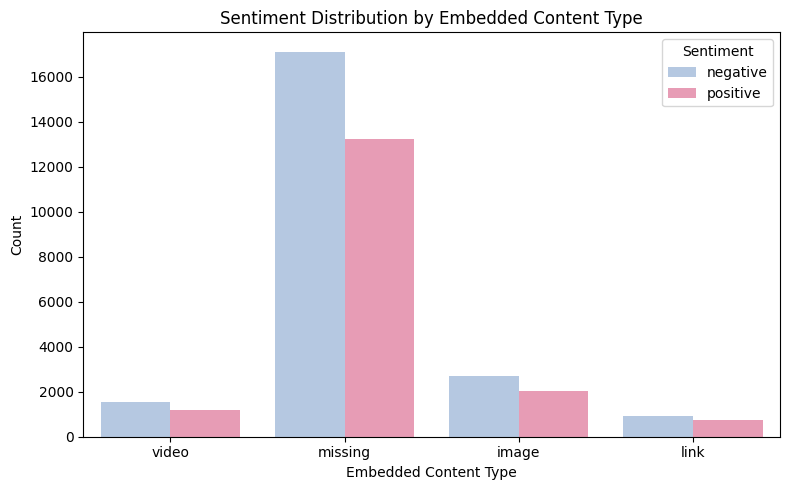
תוכן שנוצר על-ידי בינה מלאכותית עשוי להיות שגוי.**Checkmark:**משתנה המציין את סוג החשבון לפי סיווגים שונים.  
נרצה לבחון את סנטימנט הפוסט לפי סוג החשבון. מהגרף ניתן להסיק שחשבונות פרטיים חינמיים (none) מהווים את רוב המשתמשים, עם נטייה לסנטימנט שלילי. חשבונות פרטיים בתשלום (blue) גם הם מפיקים יותר תוכן שלילי מחיובי, אך בהיקף קטן יותר. בניגוד, החשבונות הרשמיים - הן של ארגונים (gold) והן של גופים ממשלתיים (grey) - מציגים דפוס הפוך: הם מפיקים יותר תוכן חיובי מאשר שלילי. מכאן נשקול בהמשך לייצר פיצ'ר שבודק אם מדובר בחשבון פרטי או חשבון ארגוני / ממשלתי ולהיעזר בכך עבור משימת הלמידה (קלספיקציה)

**Timezone:**

תמונה שמכילה טקסט, תרשים, צילום מסך, עלילה

תוכן שנוצר על-ידי בינה מלאכותית עשוי להיות שגוי.משתנה אשר מתאר את איזור הזמן של כותב הפוסט. בבחינת המשתנה בחרנו להשתמש בו על מנת להפוך את שעת פרסום הפוסט (post\_datetime) לשעה הלוקאלית האמיתית של כותב הפוסט ולבחון אחוזי סנטימנט לפי שעה.  
 מהגרף ניתן להבחין שבשעות הבוקר (בעיקר בין 7:00 ל־11:00) שיעור הפוסטים החיוביים גבוה יותר, בעוד שבשעות הערב והלילה (18:00 ואילך) נרשמת עלייה חדה בפוסטים שליליים, המהווים עד כ־71% מהתוכן. ייתכן שמדובר בהשפעה של עומס יומי או עייפות, שמובילים לביטוי רגשי שלילי יותר בשעות המאוחרות.

תמונה שמכילה טקסט, צילום מסך, גופן, תרשים

תוכן שנוצר על-ידי בינה מלאכותית עשוי להיות שגוי.**Embedded\_content\_url:**שני הגרפים מצביעים על כך שפוסטים ללא תוכן מצורף נפוצים במיוחד ונוטים להיות שליליים יותר בהשוואה לחיוביים. גם בקרב פוסטים הכוללים תוכן מצורף (כגון תמונה, וידאו או קישור) הסנטימנט השלילי עדיין דומיננטי, אך בפער קטן יותר. ייתכן שתוספת של תוכן מצורף קשורה לביטוי רגשי מעט חיובי או מאוזן יותר.

**משתנים טקסטואליים**

המשתנה id משמש כמזהה של הפוסט ואיננו מכיל מידע תוכני או מאפיינים מספריים או קטגוריים בעלי ערך לניתוח. מכיוון שאיננו תורם לחיזוי הסנטימנט או להבנת התנהגות המשתמשים, אין לו ערך אנליטי בהקשר של משימת הלמידה, ולכן נבחר שלא לעבד אותו בשלב זה.

תמונה שמכילה טקסט, קו, צילום מסך, עלילה

תוכן שנוצר על-ידי בינה מלאכותית עשוי להיות שגוי.**Text:**תוכן ההודעה שנכתבה בפוסט. עבור משתנה הטקסט ביצענו כמה ניתוחים מעניינים:

הגרף מציג את התפלגות אורכי הטקסט לפי סנטימנט.  
טקסטים שליליים מרוכזים באורכים קצרים יחסית ונפוצים בטווח של 50–100 תווים.  
לעומתם, טקסטים חיוביים נוטים להיות ארוכים יותר, עם פיזור רחב גם באורכים של מעל 200 תווים.

תמונה שמכילה טקסט, צילום מסך, תוכנה, עלילה

תוכן שנוצר על-ידי בינה מלאכותית עשוי להיות שגוי.הגרף מציג את המילים שחזרו על עצמן הכי הרבה פעמים בתוך פוסטים בודדים, עבור כל סנטימנט. כלומר, לא מדובר במילים הנפוצות ביותר באופן כללי, אלא בכאלה שהופיעו לפחות פעמיים באותו טקסט, וסוכמו לפי שכיחותן בין כלל הפוסטים. מילים כמו **"feel"**, **"like"** ו־**"feeling"** בולטות במיוחד בפוסטים חיוביים, בעוד מילים כמו **"still"** ו־**"little"** שכיחות יותר בהקשר של פוסטים שליליים.

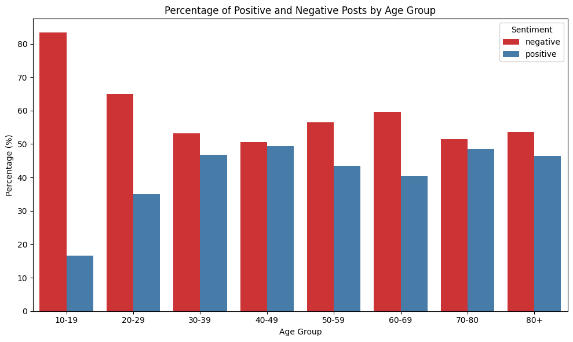
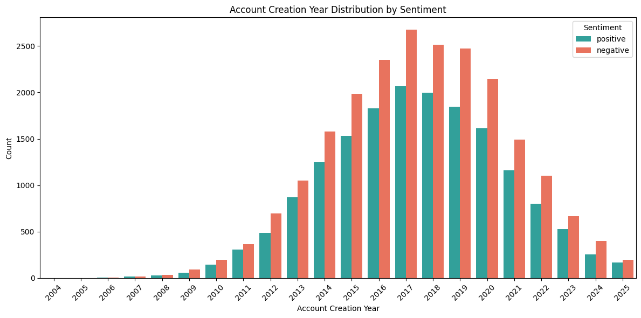
תמונה שמכילה טקסט, צילום מסך, תרשים

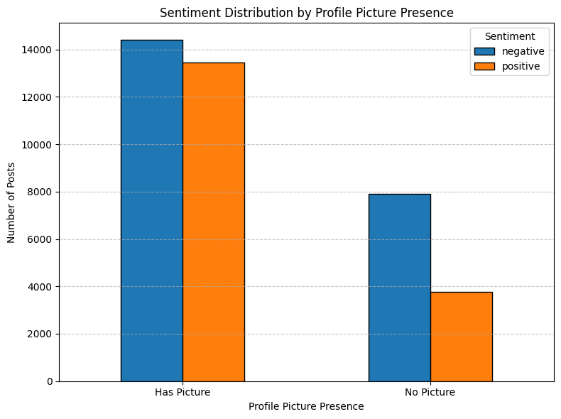
תוכן שנוצר על-ידי בינה מלאכותית עשוי להיות שגוי.**Username + Email**בחרנו לעבד את המשתנים האלו ביחד עקב הפוטנציאל שבתרומה למשימת הלמידה ולכן איחדנו אותם בצורה הבאה: עבור כל sample נבחן אם שם המשתמש והשם עבור כתובת המייל ונבדוק האם חולקים רצף של שלושה תווים לפחות - זוהי הדרך שלנו לנסות להבין אם משתמשים נרשמו עם השמות האמיתיים שלהם – הפרסונה האמיתית שלהם. מהגרף עולה כי משתמשים שהשם שלהם מופיע בכתובת האימייל נוטים לכתוב פוסטים חיוביים יותר. לעומת זאת, כאשר שם המשתמש **לא** מופיע באימייל, הסנטימנט של הפוסטים הוא בעיקר שלילי.

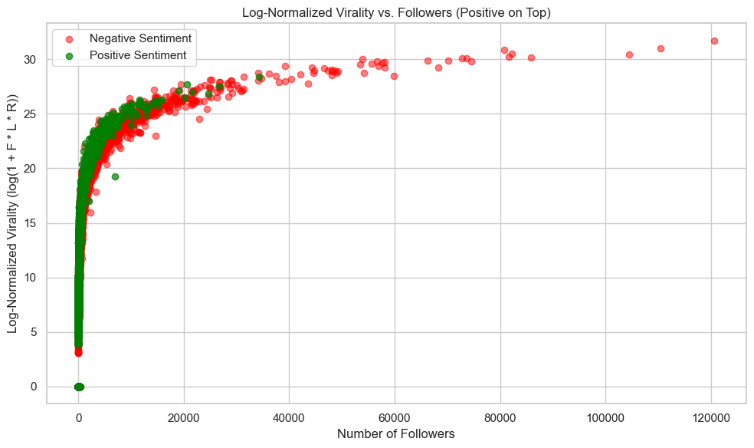
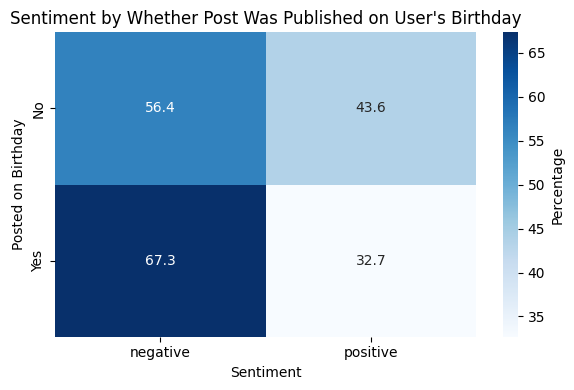
**משתני תאריך:**

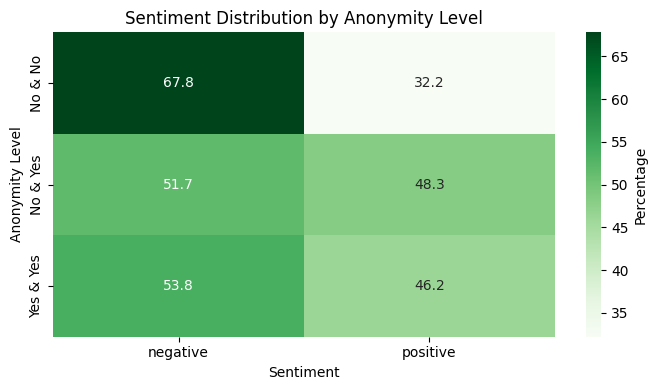
תמונה שמכילה טקסט, צילום מסך, עלילה, תרשים

תוכן שנוצר על-ידי בינה מלאכותית עשוי להיות שגוי.**Post\_datetime**תאריך ושעת פרסום הפוסט, את ההתפלגות לפי שעות וסנטימנט עיבדנו קודם לכן ולאחר עיבוד לפי שנה, יום בשבוע וחודש, מצאנו כי בחינה לפי חודש וסנטימנט מעלה מסקנות מעניינות.  
הגרף מציג את חלוקת הפוסטים לפי חודש ולפי סנטימנט. ניתן לראות שברוב חודשי השנה, כמות הפוסטים החיוביים גבוהה משמעותית מהשליליים. עם זאת, יש חריגה ברורה בחודשים **ינואר, פברואר, נובמבר ודצמבר**, שבהם כמות הפוסטים השליליים קופצת בחדות ומובילה על פני החיוביים. ייתכן שתקופות אלה קשורות ללחצים עונתיים כמו חגים, תחילת שנה או סוף שנה, אשר משפיעים לרעה על התחושות הכלליות של המשתמשים. **Birthday + Account creation date**

תאריך יום הולדת – חילקנו את גיל המשתמש לקטגוריות גיל שונות והסתכלנו על ההתפלגות לעומת הסנטימנט, ניתן להסיק כי בקרב משתמשים צעירים ישנה נטייה חדה לפוסטים שליליים אשר מתמתנת עם העלייה בגיל.  
תאריך יצירת החשבון – הסתכלנו על שנת היצירה ועל התפלגות הסנטימנט לפי שנה, ניתן להסיק מהתרשים כי בעיקר בשנים האחרונות ישנם פוסטים שליליים יותר מחיוביים.

**משתנים בינאריים  
profile\_picture\_url:**משתנה הבוחן האם למשתמש יש תמונת פרופיל (קישור קיים או לא). מהגרף עולה כי משתמשים ללא תמונת פרופיל מפרסמים יותר פוסטים שליליים ביחס לחיוביים, בניגוד למשתמשים עם תמונה, אצלם ההתפלגות מאוזנת יותר. ייתכן שמשתמשים שעשויים להיות אנונימיים נוטים יותר להביע ביקורת או רגשות שליליים.

**קשרים מעניינים נוספים  
ויראליות של פוסט מול סנטימנט:**הגרף מצביע על כך שפוסטים שליליים נוטים להיות ויראליים יותר, בעיקר כשנכתבים על ידי משתמשים עם הרבה עוקבים. לעומתם, פוסטים חיוביים נפוצים יותר בטווחי ויראליות בינוניים. ייתכן שתוכן שלילי מעורר יותר תגובה ושיתוף.(ויראליות נבדקה לפי כמות עוקבים של מפרסם, מספר לייקים ומספר פרסומים חוזרים)

**פרסום פוסט בתאריך יום ההולדת מול סנטימנט:**איחדנו את המשתנים birthday, post\_datetime לעמודה בינארית שבודקת האם הפוסט פורסם ביום ההולדת של המשתמש אל מול סנטימנט הפוסט. התוצאות יצאו דווקא הפוכות ממה שציפינו – ביום ההולדת של המשתמשים, שיעור הפוסטים בעלי סנטימנט **שלילי** גבוה יותר (67.3%) לעומת חיוביים (32.7%). בימים רגילים, הפער בין חיובי לשלילי יחסית מאוזן. ממצא זה עשוי לרמוז כי ביום ההולדת משתמשים נוטים לשתף תחושות פחות חיוביות מהצפוי, אולי בשל ציפיות לא ממומשות או חוויות אישיות מורכבות סביב התאריך.   
**רמת האנונימיות של המשתמש מול סנטימנט:**הגרף מראה שמשתמשים אנונימיים לגמרי (ללא שם בתוך המייל וללא תמונת פרופיל) נוטים לפרסם פוסטים שליליים בשיעור גבוה במיוחד. לעומתם, כשקיים לפחות מאפיין מזהה – הסנטימנט מתאזן. ככל שרמת הזיהוי עולה, כך פוחתת הנטייה לביטוי שלילי. חשוב לציין כי זיהוי השם נעשה על בסיס הופעת שם המשתמש כרצף בתוך כתובת המייל כפי בדקנו קודם לכן.

:Pre – Processingבתהליך זה אנו מבצעים מניפולציה על הדאטה לאחר החישה כך שנוכל להשתמש בו בצורה טובה יותר בשלבים הבאים.  
**טיפול בחזרתיות:** לאחר בדיקה נמצא כי אין סמפלים בעלי אותם ערכים ולכן לא קיימת חזרתיות בטקסט. **טיפול בערכים חסרים:** מצאנו כי קיימים ערכים חסרים בדאטה. בחרנו לסווג אותם לכמה סוגים. **\* סמפלים בהם יש כ-8 ערכים חסרים ומעלה :** שורות אלו מהוות כ-1.21% מכלל הדאטה שלנו ולכן בחרנו למחוק את השורות האלו. זאת מכיוון שהשלמה מלאכותית של הערכים יכולה להביא ליותר מידי רנדומיזציה וחוסר התאמה עם המציאות. בגלל שזהו אחוז קטן מכלל הדגימות מחיקתם לא יגרום לשינוי מהותי בדאטה.   
**\*סמפלים ללא סיווג של סנטימנט (משתנה המטרה) :** מדובר בדגימות שמהוות כ- 1.32% מכלל הדאטה ולכן גם אותו בחרנו למחוק. מכיוון שמדובר במשתנה המטרה שסווג באופן מקצועי, לא נרצה לתת ערך שיכול להיות מוטעה, בנוסף אנחנו מתייחסים לכך שהמשימה שלנו היא מסוגSupervised Learning, שכן קיים משתנה מטרה ידוע המשמש את המודל בזמן האימון.  
**\*סמפלים ללא ערך מסוג אזור זמן ותאריך פרסום הפוסט :** בחרנו להשלים את הערכים האלו באופן רנדומלי מתוך הטווח הערכים שנתון לנו בדאטה מכיוון שאין מגמתיות מובהקת בערכים אלו. בשני המקרים מדובר מדגימות שמהוות כ- 1.27% ולכן לא תהיה השפעה גדולה על השינוי.   
**\*סמפלים ללא ערך מסוג כמות לייקים וכמות שיתופים:** בחרנו להשלים את ערכים אלו על ידי השמת הממוצע של הערך .מדובר על דגימות שמהוות כ – 1.21% מהדאטה ולכן השמה זו תאפיין את הדגימות באופן הגיוני.   
**המרת משתנים:**  
**המרת סנטימנט לערכים מספריים**: עמודת הסנטימנט הומרה לערכים נומריים, כאשר הערך "positive" הומר ל־1 והערך "negative" הומר ל-0. המרה זו מאפשרת ניתוח כמותי של הסנטימנט בכלים סטטיסטיים ובאמצעות מודלים של למידת מכונה. (העמודה הפכה מטקסטואלית לבינארית)  
**סיווג סוג תוכן לפי סיומת URL** :ניתחנו את כתובות הURL וחילקנו לקטגוריות: video, image, link, missing (העמודה הומרה מטקסטואלית לקטגוריאלית) שם העמודה הפך לembedded\_category  
**בדיקת קיום תמונת פרופיל**: עמודת כתובת תמונת הפרופיל הומרה למשתנה בינארי, בו: הערך 1 מציין שקיימת תמונת פרופיל הערך 0 מציין שאין תמונת פרופיל, שינוי זה מאפשר ניתוח פשוט ויעיל יותר של מאפייני המשתמשים (העמודה הפכה מטקסטואלית לבינארית)  
**עמודת יצירת חשבון המשתמש:** הומרה לprofile\_creation\_year בה נמצאת שנת פתיחת החשבון.  
:Feature Extractionבשלב זה נבצע חילוץ, תיאור ויצירה של מאפיינים (features) מתוך המידע הגולמי, במטרה לייצג ישות לצורך משימת הלמידה.  
את המשתנים הנומריים post\_likes, post\_retweets, followers, previous\_posts\_counts נהפוך למשתנים קטגוריאליים כדי לתת ערך למספרים השונים עבור משימת הלמידה (פירוט הקטגוריות יופיע בנספחים). משתנים שהגיעו כקטגוריאליים בדאטה הגולמי יישארו כפיצ'רים כפי שהם.   
post\_hour\_category – את שעת פרסום הפוסט חישבנו לפי שילוב עם עמודת timezone עבור שעת הפרסום הלוקאלית של המשתמש וחילקנו לקטגוריות כפי שמפורט בנספחים.  
user\_age\_group – מעמודת birthday שלפנו את שנת הלידה, חישבנו את גיל המשתמש וחילקנו את הגילאים לקטגוריות כפי שמפורט בנספחים.  
is\_birthday – מעמודות birthday, post\_datetime ניצור פיצ'ר שאומר אם הפוסט פורסם ביום ההולדת של המשתמש או לא. 1- ביום ההולדת, 0- לא ביום ההולדת.  
is\_holiday\_season – לפי מה שמצאנו בניתוח הדאטה בשלב הקודם, בחודשים מסויימים בתחילת שנה וגם סוף שנה שבהרבה מקומות מתקשרים לחגים, חופשות ולחצים עונתיים, ישנה עלייה בפוסטים שליליים ולכן הפיצ'ר יראה 1 אם מדובר בקבוצת חודשים אלו ו0 אם לא.  
is\_user\_anonymous – מעמודות username, email ניצור פיצ'ר שבודק אם המשתמש נרשם עם שמו האמיתי לפי הופעת רצף של שלושה תווים חופפים לפחות, כך נקבע אנונימיות. 1 – נרשם עם שמו האמיתי, 0 – לא נרשם עם שמו.  
is\_private\_account – כפי שראינו בניתוח הדאטה, אם המשתמש הוא פרטי ישנם יותר פוסטים שליליים ואילו אם החשבון ממשלתי\ארגוני ישנם יותר פוסטים חיוביים. לכן פיצ'ר זה יקבל 1 במידה ומדובר בחשבון פרטי ואפס אחרת.  
account\_age\_category – את העמודה profile\_creation\_year נהפוך לקטגוריאלית כמפורט בנספחים.  
ממשתנה הטקסט שלפנו כמה פיצ'רים -   
text\_length\_category – כפי שראינו בניתוח הדאטה, אורך הטקסט יוצר הבחנה בין פוסטים חיוביים לשליליים.  
בנוסף, בעיבוד משתנה text ,יישמנו ניקוי של מילים נפוצות (Stop Words) כדי להסיר מילות קישור שאינן תורמות משמעותית להבחנת הסנטימנט.  
לאחר מכן השתמשנו בשיטת N-gram (עם טווח של 1–2 מילים) כדי לחלץ מילים בודדות וצמדי מילים שעשויים לשקף את הסנטימנט של ההודעה.  
את התכנים שחולצו ייצגנו באמצעות מטריצת TF-IDF, אשר נותנת משקל למילים לפי שכיחות יחסית במסמכים.  
לבסוף בחרנו את 100 המילים והביטויים בעלי ערכי TF-IDF הגבוהים ביותר, והוספנו אותם כמאפיינים חדשים.  
:Feature Representationזהו השלב שבו נתרגם את הפיצ'רים לייצוג מספרי תקני ואחיד, כדי שהמודל יוכל ללמוד מהם בצורה נכונה.  
נרצה לנרמל את הפיצ'רים לבעלי סקאלה משותפת כך שהערכים של כל פיצ'ר יהיו בין 0 ל1 וזאת על מנת שיהיה משקל זהה עבור מידול נכון למשימת הלמידה. עשינו זאת באופן הבא:  
כל המשתנים הקטגוריאליים שנותרו מהדאטה הגולמי יהפכו לעמודות לפי קטגוריה כך שבכל עמודה יהיה 0 או 1 אם הסאמפל משתייך לקטגוריה זו או לא. בנוסף, כל המשתנים הנומריים שהפכנו לקטגוריאליים בשלב הקודם, יהפכו לעמדות קטגוריה גם כן. את המעבר המתואר נבצע בעזרת הפונקציה המובנית get\_dummies אשר הופכת ערכים קטגוריאליים לעמודות בינאריות. על המילים שחולצו ממשתנה text בוצעה מניפולציה באמצעות Min-Max Scaling שנועדה להביא את ערכי ה-TFIDF לטווח תקני שבין 0 ל-1.תהליך זה מאפשר לאחד את סדרי הגודל של הנתונים, למנוע הטיות הנובעות מהבדלי גודל בין טקסטים שונים, ולוודא שכל פיצ'ר יישקל באופן יחסי ואחיד במהלך הניתוח.   
:Feature Selectionבשלב זה רצינו לספק למודל סט פיצ'רים קטן אך אינפורמטיבי, כדי לא להעמיס עליו בפרמטרים מיותרים שעלולים להקשות על תהליך הלמידה ואף להוביל ל־ overfitting . בדקנו את כלל הפיצ'רים ובחנו עד כמה כל אחד מהם תורם להבחנה בין פוסטים בעלי סנטימנט חיובי לשלילי.  
לצורך הבחינה בחרנו בהערכה כמותנית והשתמשנו בשיטת הסינון **Filter Method – Fisher Score** אשר נלמדה בהרצאה. בהתבסס על תוצאות הציונים (מצורף טופ 50 בנספחים), ובמטרה לשמור על איזון נכון בין איכות כמותית לבין פשטות המודל, בחרנו את **25 הפיצ'רים בעלי הציונים הגבוהים ביותר** מתוך  **152** פיצ'רים שנבדקו.  
:Dimensionality Reductionבשלב הקודם ביצענו Feature Selection בעזרת מדד Fisher Score במסגרתו סיננו ובחרנו את 25 הפיצ'רים האינפורמטיביים ביותר מתוך כלל הפיצ'רים הראשוניים.  
בחירה זו בוצעה במטרה להבטיח שהפיצ'רים הנותרים יהיו גם מייצגים היטב את הסנטימנט (חיובי/שלילי) וגם שמספרם יהיה מצומצם יחסית, כך שהמודל יוכל ללמוד בצורה יציבה ואפקטיבית.  
מאחר שכבר בוצעה הפחתה משמעותית במספר הפיצ'רים, השימוש בשלב נוסף של Dimensionality reduction כמו PCA אינו נחוץ ואף עלול להיות בעייתי ממספר סיבות:

* צמצום נוסף של פיצ'רים עלול להביא לאיבוד מידע חשוב, מאחר וכמות הפיצ'רים כעת גם כך קטנה ואינפורמטיבית.
* PCA הוא תהליך לא מונחה (unsupervised) – הוא מתחשב רק בשונות הנתונים ולא בתוויות הסנטימנט עצמן, ולכן הוא עלול לשמר פיצ'רים שאינם האופטימליים למשימת הקלאסיפיקציה הספציפית שלנו.
* בנוסף, מספר הפיצ'רים שנותרו (25) הוא סביר בהחלט ביחס למספר הדגימות (40,000 דגימות), כך שאין חשש לעומס חישובי או Overfitting שנובע מכמות פיצ'רים גבוהה מדי.

בהתבסס על שיקולים אלו, החלטנו לדלג על שלב ה־dimensionality reduction ולהשתמש ישירות בסט הפיצ'רים הנבחר לבניית מודל הלמידה.

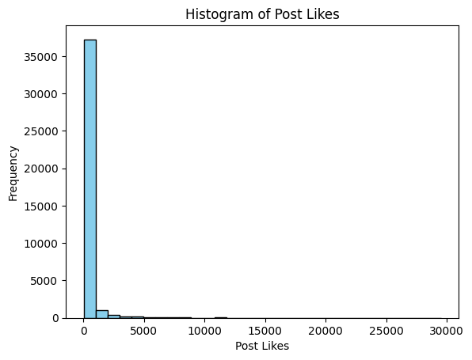
:ValidationCross Validation היא שיטה שמטרתה לבדוק את ביצועי המודל על ידי חלוקת הדאטה למספר קבוצות שונות. בכל סיבוב, המודל מאומן על חלק מהנתונים ונבחן על חלק אחר, וכך מתקבלת הערכה מדויקת ויציבה ליכולתו להכליל על דאטה חדש.בשלב הוולידציה, בחרנו להשתמש בשיטת K – fold .   
בשיטה זו הנתונים מחולקים ל־ K קבוצות שוות בגודלן. בכל סיבוב, המודל מאומן על K-1 קבוצות ונבחן על הקבוצה הנותרת, כך שכל קבוצה משמשת פעם אחת לבחינה. בסיום התהליך, ממוצע התוצאות מכל הסיבובים משמש להערכת הביצועים.  
בהשוואה לשיטות אחרות:

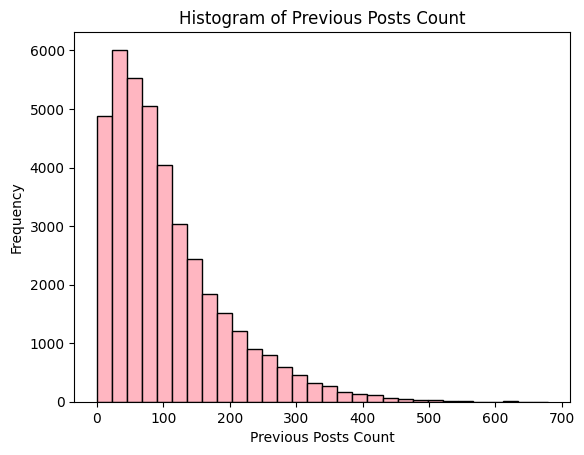
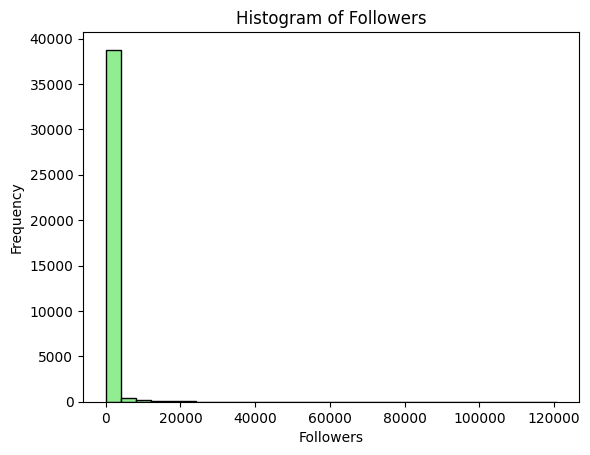
* בשיטת Holdout הנתונים מחולקים פעם אחת לסט אימון וסט בדיקה קבועים. אמנם שיטה זו פשוטה ומהירה, אך היא רגישה מאוד לאופן הפיצול, פיצול לא מייצג עלול להטות את הערכת ביצועי המודל ולהוביל למסקנות שגויות לגבי יכולתו להכליל.
* בשיטת leave one out בכל איטרציה דוגמה אחת בלבד נשמרת לבדיקה וכל שאר הדאטה משמש לאימון. אמנם זוהי הערכה מאוד מדויקת (כי כמעט כל הדאטה משמש לאימון בכל סיבוב), אך החיסרון המשמעותי הוא העומס החישובי הכבד: לנו יש כמות סאמפלים גדולה מה שהופך את השיטה ללא יעילה בפרקטיקה.

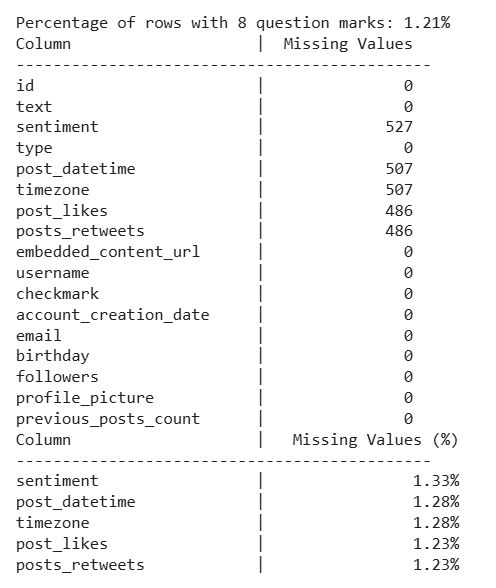
K-Fold מהווה אפשרות טובה:  
היא מספקת אומדן יציב ואמין של ביצועי המודל תוך איזון בין ניצול מירבי של הדאטה ובין דרישה חישובית סבירה. בנוסף, השיטה עוזרת להפחית סיכון ל־overfitting ומבטיחה שהמודל ייחשף למגוון רחב של דוגמאות בכל איטרציה. לצורך מדידת ביצועי המודל, נבחרה מטריקת Recall המתאימה למשימת הסיווג שלנו, מאחר והיא שמה דגש על זיהוי נכון של דגימות חיוביות.

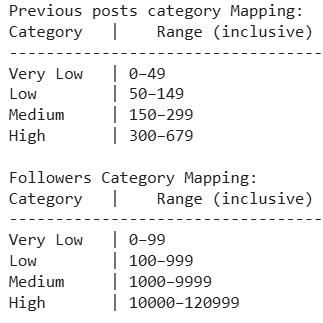
# נספחים:

תמונה שמכילה טקסט, צילום מסך, תרשים, קו

תוכן שנוצר על-ידי בינה מלאכותית עשוי להיות שגוי.היסטוגרמת ההתפלגות עבור המשתנים הנומריים ללא שימוש בטווח:



בדיקת אחוזי חוסרים בשלב הpre processing   


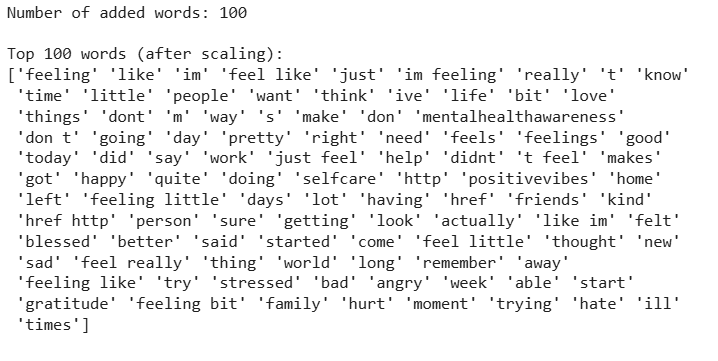
מיפויי קטגוריות עבור משתנים שהפכו לקטגוריאליים:  
 תמונה שמכילה טקסט, צילום מסך, גופן, מספר

תוכן שנוצר על-ידי בינה מלאכותית עשוי להיות שגוי. תמונה שמכילה טקסט, צילום מסך, גופן, מספר

תוכן שנוצר על-ידי בינה מלאכותית עשוי להיות שגוי. תמונה שמכילה טקסט, קבלה, גופן, צילום מסך

תוכן שנוצר על-ידי בינה מלאכותית עשוי להיות שגוי.

רשימת הפיצ'רים לפני ניתוח טקסטואלי  
  
טופ 100 שילובים של טקסט אחרי N-gram, TFIDF



תמונה שמכילה טקסט, צילום מסך, תפריט, מסמך

תוכן שנוצר על-ידי בינה מלאכותית עשוי להיות שגוי.תמונה שמכילה טקסט, גופן, תפריט, צילום מסך

תוכן שנוצר על-ידי בינה מלאכותית עשוי להיות שגוי.טופ 50 פיצ'רים עם ניקוד Fischer score

סט פיצ'רים סופי לאחר Fischer score  
