סיווג כותב טקסט כמוחצן / מופנם

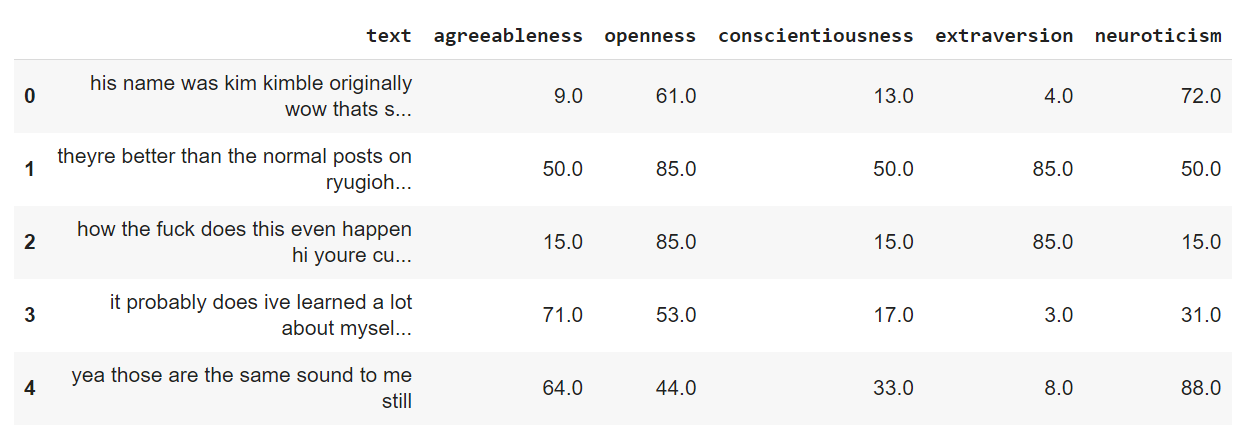
איתי טולדו 318720604

**1. נתונים**

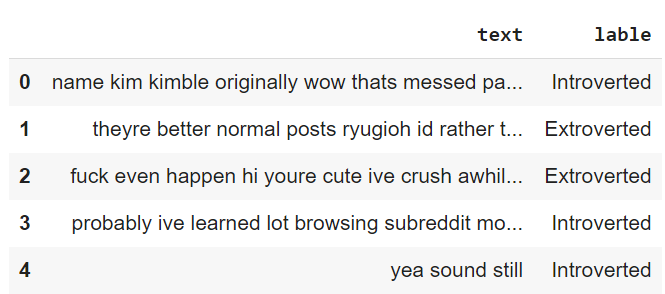
המטרה של הפרויקט היא סיווג בעזרת מודל שפה כותבי טקסטים מוחצנים ומופנמים כחלק מbaseline לתזה. **הנתונים בהם אשתמש לקוחים מתוך מסט הנתונים** PANDORA **אשר** **כולל נתונים דמוגרפיים ופסיכולוגיים של כותבי הטקסטים.**

**הנתונים הראשוניים כוללים טבלה אחת בת 6 עמודות המכילה 21,000 טקסטים באורכים שונים וציון כמותי בין 0 ל100 עבור 5 תכונות אישיות מתוך** The big 5. **הנתונים מגיעים מחולקים לסטי אימון, ולידציה ומבחן. לאחר בדיקה של הנתונים, נמצא שחלק מהרשומות מופיעות כמה פעמים אז הן הוסרו וחולקו מחדש לסטי אימון, ולידציה ומבחן ביחס 60:24:16 כאשר סה"כ ישנן 17,009 רשומות.**

**הנתונים הגולמיים:**

****

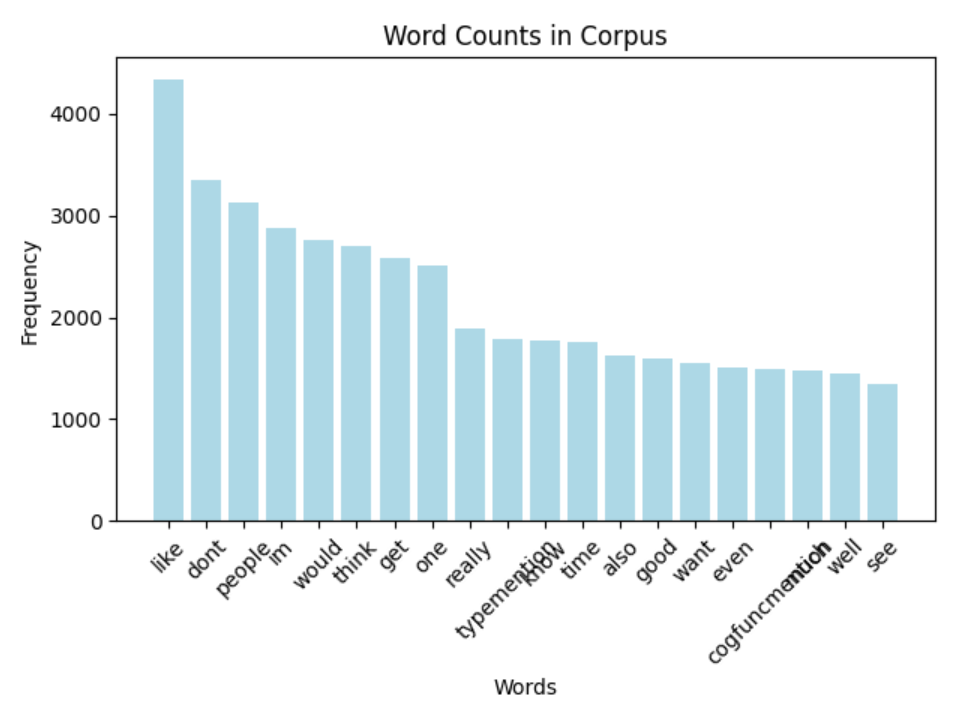
**הנתונים שיועברו לאימון מודל:**

****

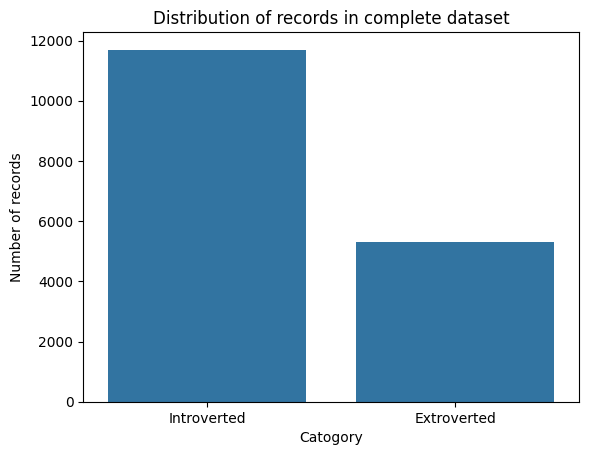
כחלק מהעיבוד המקדים הורדו מהטקסטים סימני פיסוק + stopwords וגם רמת המוחצנות הומרה לשתי קטגוריות: מוחצן ומופנם.

כעת נשארו 2 עמודות. עמודת הטקסט תשמש כמשתנה קלט והקטגוריות כמשתנה מטרה.   
בעזרת סט האימון וסט הולידציה נאמן את המודל ונבחן אותו על סט המבחן.

גרף המראה את המילים הפופולריות ביותר במאגר הנתונים:  
המילים like, don’t, people, im מופיעות הרבה וכנראה שיהיה להן משקל רב.



התפלגות הנתונים לקבוצות:  
יש פי 2 יותר דוגמאות לטקסטים מופנמים מאשר למוחצנים.



# 2. מתודולוגיה

נמדוד את ביצועי המודל לפי accuracy.

לאחר בדיקת היחס בתוך סט האימון בין כמות הדוגמאות לכמות המילים הממוצעת בסט, נעשה שימוש במודל Bag of Words (BoW) כאשר נממש את TextVectorizer של keras ואת הוקטורים נאכיל לMLP.

נבדוק את הארכיטקטורה הטובה ביותר על ידי בחינת:

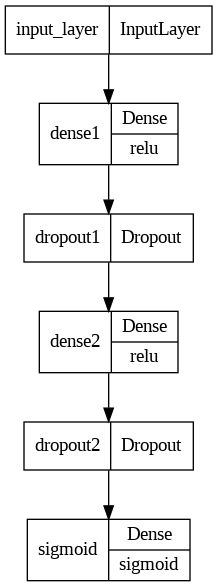
* max\_tokens - כמות הטוקנים שייחשבו במודל (500, 1000, 2000)
* גודל הטוקנים (1gram, 2gram)
* גודל השכבות בMLP

לאחר מעבר על האפשרויות, נגדיר שהערכים יהיו ציוני tf-idf של 1000 המילים הכי פופולריות בסט האימון. הפונקציה ממירה כל סט נתונים בגודל (x, y) לסט בגודל (x, 1000).

כמות הepoch הוגדרה ל- 10 לאחר מספר בדיקות.

רשת הנוירונים בנויה ככה:

1. שכבת כניסה מקבלת 10,205 דגימות אימון, כל דגימה היא וקטור של ציוני tf-idf ל- 1000 מילים. 32 דגימות יועברו דרך המודל בכל צעד.
2. שכבה חבויה ראשונה בגודל 256 תאים עם פונק' אקטיבצייה relu
3. שכבת dropout אשר משביתה 30% מהדגימות
4. שכבה חבויה שנייה בגודל 128 תאים עם פונק' אקטיבציה relu
5. שכבת dropout אשר משביתה 30% מהדגימות
6. שכבת יציאה בת תא אחד ופונק' אקטיבציה sigmoid על מנת לסווג בינארית



# 3. תוצאות

המודל המדויק ביותר מקבל וקטורים המסמלים את ערכי tf-idf של 2000 המילים (ngram=1) הנפוצות ביותר. למודל יש דיוק של 87% על טקסט שמעולם לא נחשף אליו (מבחן)

# 4. סיכום ומסקנות

המטרה של הפרויקט הייתה לסווג בעזרת מודל שפה כותבי טקסטים מוחצנים ומופנמים כחלק מקו בסיס (baseline) לתזה שלי.

במסגרת העיבוד המקדים המרתי את ציון המוחצנות ל2 קבוצות, הסרתי עשרות דגימות החוזרות על עצמן וגיליתי כי כמות הדגימות שמסווגות 'מופנם' גדולות פי 2 מ 'מוחצן'. לכן בחלוקה לסטים היה צורך להגדיר חלוקה שווה של labels בין הסטים.

הגדרתי קו בסיס (baseline) בעזרת מודל אקראי שהשיג 57% דיוק.

לאחר מכן, הכנתי מודל שפה טבעית עם BoW שהשיג שיפור בדיוק (87%)

# 5. קישור לפרויקט

[קישור](https://colab.research.google.com/drive/1zQd81vsY5_DpB2k3iuaoVsr-pykFD6ij?usp=drive_link)