סיווג כותב טקסט כמוחצן / מופנם

איתי טולדו 318720604

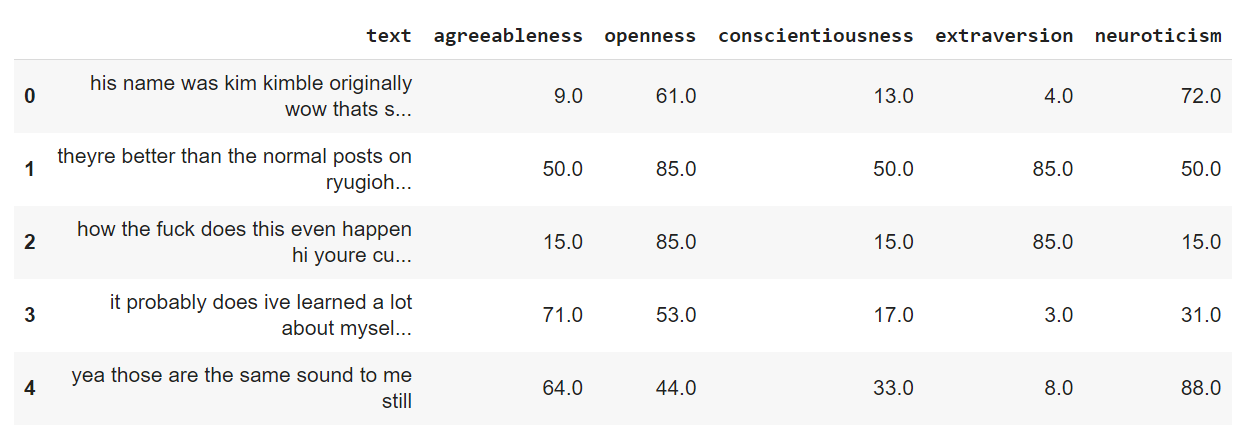
1. מבוא:

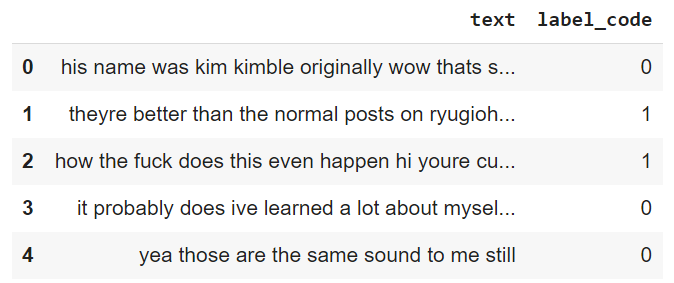
הפרויקט נועד לסווג כותבי טקסט כמוחצנים או מופנמים באמצעות מודל למידה עמוקה שאומן על תגובותReddit באמצעות מערך הנתונים של PANDORA דרך [HuggingFace](https://huggingface.co/datasets/Fatima0923/Automated-Personality-Prediction)

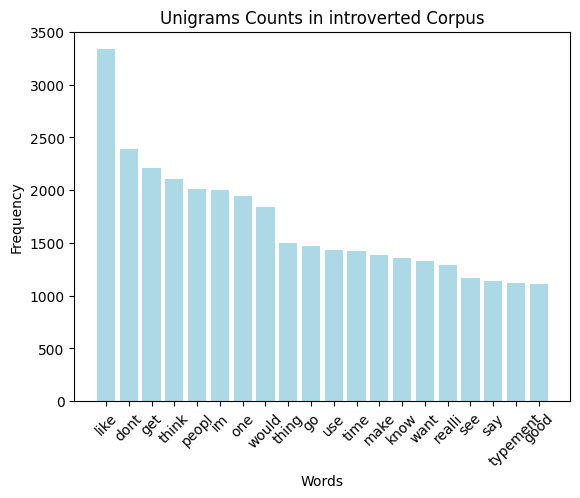
2. נתונים:

הנתונים הראשוניים כללו 20,877 טקסטים עם ציונים כמותיים עבור חמש תכונות אישיות. טקסטים עברו עיבוד מקדים, כולל הסרת סימני פיסוק, מילות עצירה והכרה. רמות המוחצנות הומרו לקטגוריות מוחצנות ומופנמות עבור משתני יעד.

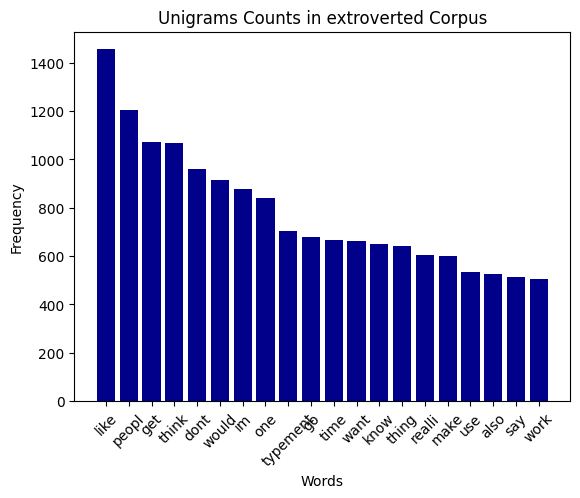
**דוגמה לנתונים הגולמיים:**

****

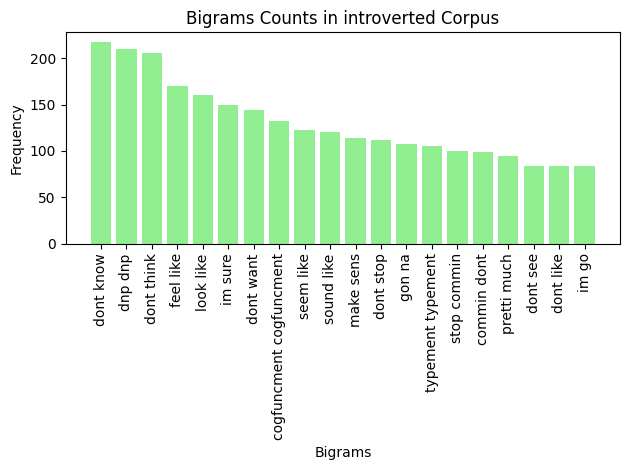
**דוגמה לנתונים שהועברו לאימון:**

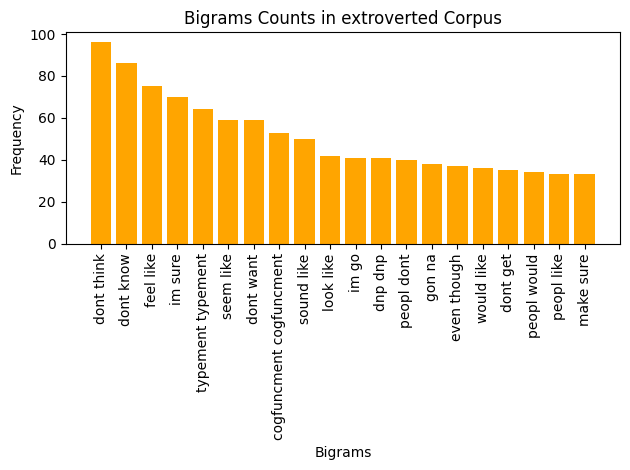


תדירות המילים ב"טקסטים מוחצנים":

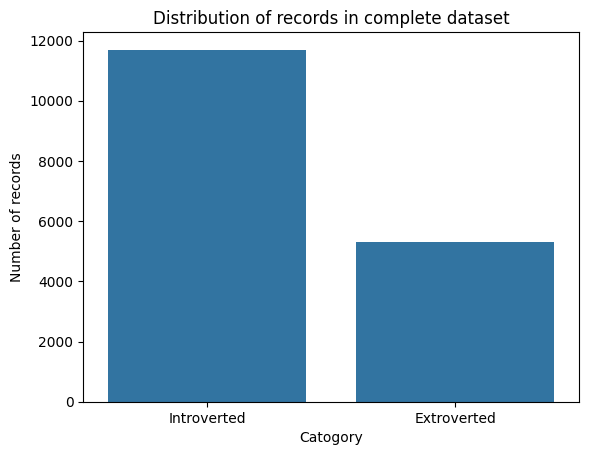


תדירות bigrams ב"טקסטים מופנמים":



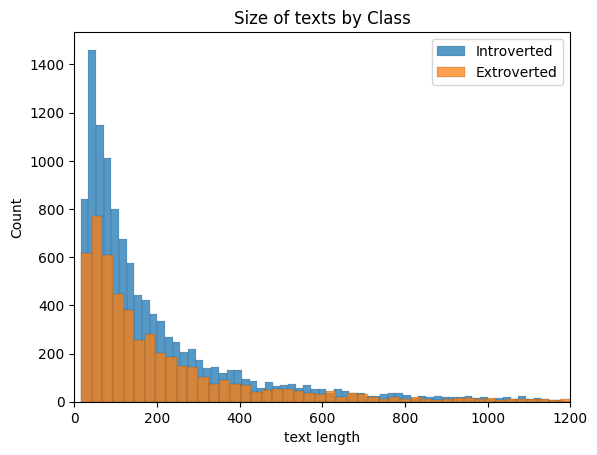
תדירות bigrams ב"טקסטים מוחצנים":

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



גודל טקסטים לפי קבוצה:

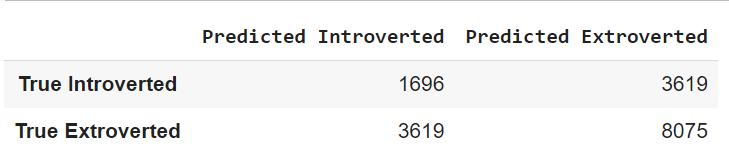
מופנמים נוטים להביע בכתב יותר מאשר מוחצנים.



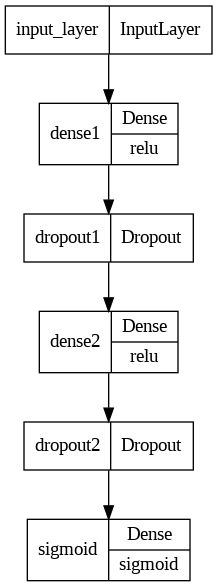
3. מתודולוגיה:

ביצועי המודל נמדדו לפי דיוק ומטריצת בלבול. קו בסיס נקבע באמצעות מודל אקראי. מודל Bag of Words נבחר לאחר בחינת היחס בין דוגמאות למילים במערך ההדרכה.

ביצועי המודל האקראי:



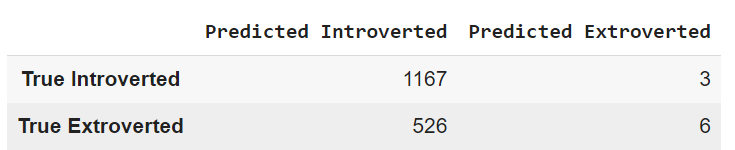
הארכיטקטורה:



4. תוצאות:

המודל הטוב ביותר כלל שתי שכבות עם 128 נוירונים כל אחת והפעלת ReLU, והשיג דיוק של 68.9%. רוב השגיאות התרחשו בסיווג שגוי של טקסטים מוחצנים כמופנמים.

ביצועי המודל:



5. סיכום ומסקנות:

הפרויקט סיווג בהצלחה את רמות המוחצנות של כותבי טקסט, כאשר מודל Bag of Words עלה משמעותית על baseline אקראי.

6. קישור לפרויקט:

קישור