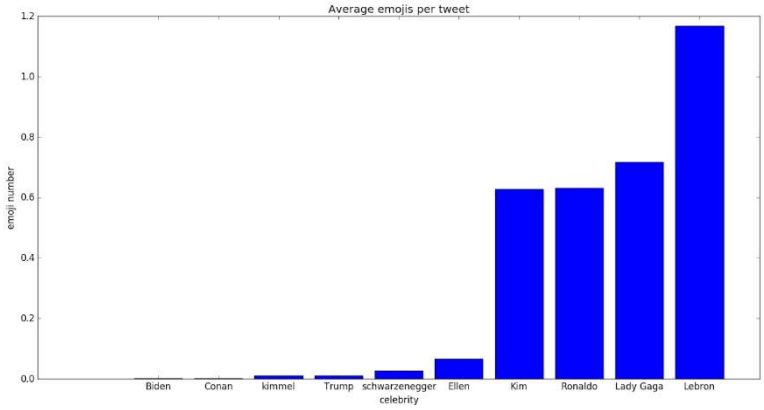
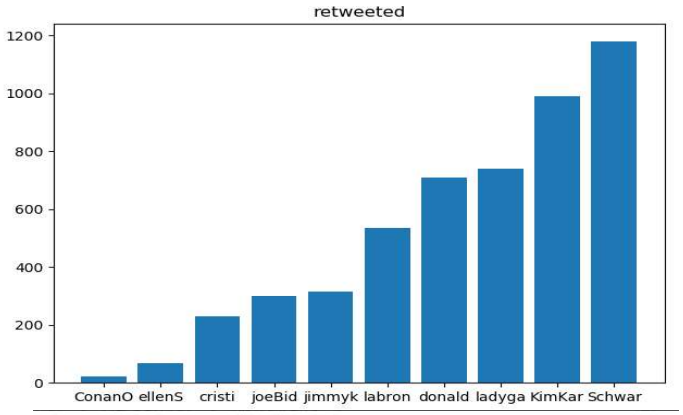
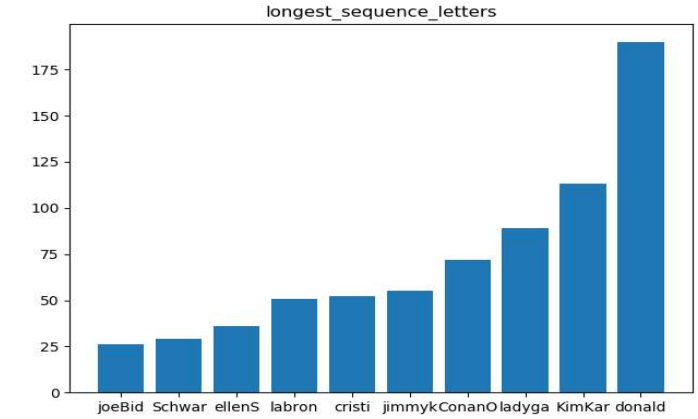
תהליך העבודה:

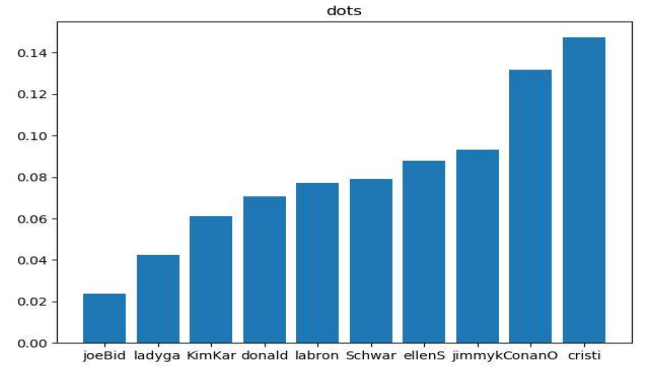
התחלנו משקילת הפיצ'רים שבהם נרצה להשתמש, עשינו זאת בשלבים:

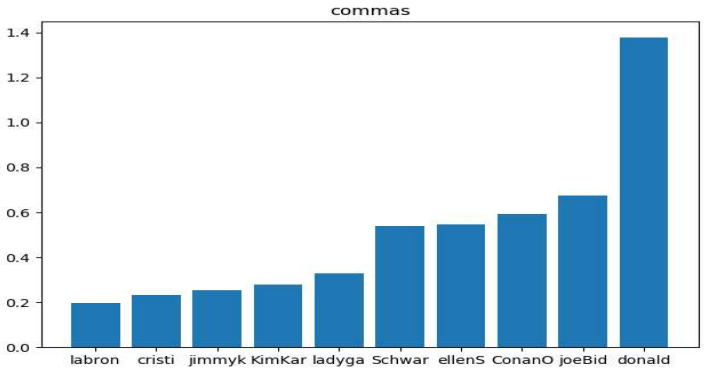
1. בנינו רשימה של כל הפיצ'רים שחשבנו שיכולים להיות רלוונטיים בתהליך הסיווג.
2. פירסרנו את הציוצים לפי הפיצ'רים הללו ובדקנו את ההיסטוגרמה של כל אחד מהם.
   a. להלן ההיסטוגרמות שבלטו לעינינו:

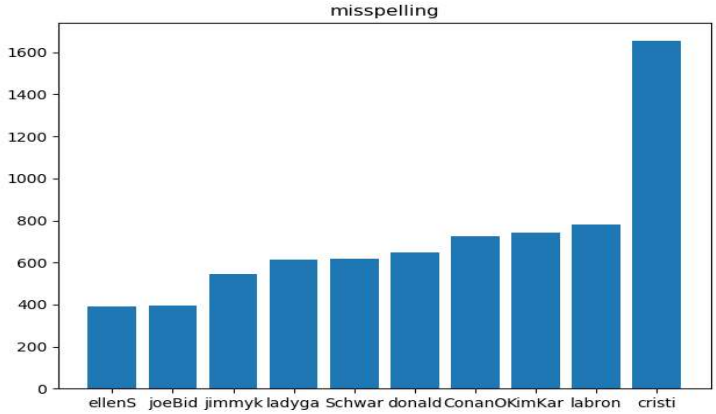


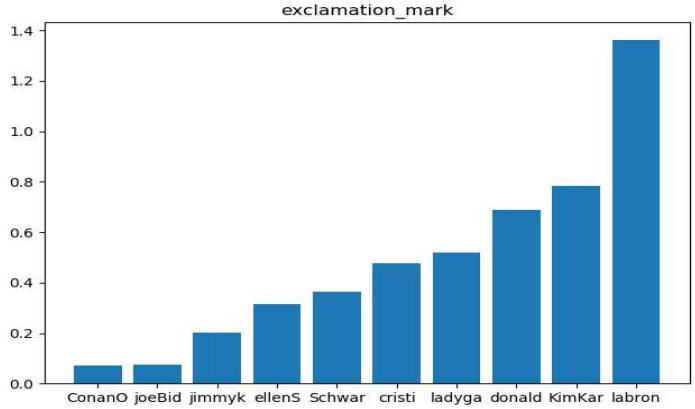












b. ניתן אם כן לראות כי בהיסטוגרמות הללו ההפרשים בין אישויות שונות נראה לעין ולכן תהליך הסיווג עם מאפיינים אלו יהיה יעיל יותר.

לאחר מכן , עברנו ליצירת מאפיינים על בסיס מילים, בדרך הבאה:

1. לכל ציוץ קודדנו וקטור בשיטת BOW. וקטורים אלה מכילים את שכיחות המילה בתוך הטוויט (כלומר, יכולים בהחלט להיות ערכים הגדולים מ-1. בדאטא שלנו השכיחות הכי גבוהה בתוך טוויט אחד היתה 10). סך כמות המילים השונות בהם השתמשו המשתמשים בכל הטוויטים (כלומר 'המילון') הוא גדול מאוד.

2. אולם, הוספת למעלה מ 40 אלף מאפיינים לתהליך הסיווג לא היה יעיל במיוחד ופנינו לבחון דרכים לסינון המילים. כדי להקטין את גודל הוקטורים (כלומר את גודל הפיצ'רים) השתמשנו בSTEMMING. שיטה זו מחליפה כל מילה ל'גזע' אליו היא שייכת. כך למשל מילים כמו "programm", "programming", "programmer" שייכות כולן לגזע "programm". בכך, הכנסנו כפי שרצינו עוד BIAS למודל, והקטנו את הVARIANCE.

3. בשלב הבא בדקנו קורלציה בין מילים שונות, שמופיעות בכלל הציוצים, לבין אישיות מסויימת. בדרך זו יכלנו לקבוע אילו מילים משפיעות יותר על הניבוי ועל אילו נוכל לוותר כאשר נשתמש בשק המילים בתור מאפיין. יחד עם זאת, קורלציה של מאפיינים יחידים עם הלייבלים התבררה כלא יעילה בשלב בחירת המאפיינים. דבר זה נובע למשל מהעובדה שמילה בודדת אולי לא תשפיע במיוחד על הסיווג, אבל שילוב של כמה מילים, יכול מאוד להתקשר לאדם ספציפי.

- בדקנו אם אפשר להשיג עוד דאטא מטוויטר - לא מצאנו דרך לעקוף את הטוויטר API בלי לשלם וזאת נראתה השקעה קצת מוגזמת:)

<u>קלסיפיירים שחשבנו לנסות:</u>

Random forest\SVM\Logistic regression.

<u>קליסיפייר נבחר:</u>

בחרנו להשתמש בGLM. ראשית, בהרצות ראשונית על הדאטא הביצועים שלו היו טובים יותר מקלסיפיירים אחרים. שנית, מספר הפיצ'רים בBOW הוא גדול (כ-40000 לאחר הSTEMMING), הוא עלול לעבור את מספר נק' הטריינינג דאטא (n<d) וגם במקרים שאינם כאלה ליצור בעיה של אוברפיטינג. השימוש ברגרסיה לוגיטסית מאפשר לנו להשתמש גם ברגולריזצית LASSO שמורידה באופן משמעותי את הפיצ'רים. בנוסף לכך, בחרנו לקודד את הדאטא בעזרת BOW בתוספת הפיצ'רים הידניים שיצרנו עליהם חשבנו מראש (ר' לעיל). כיון שחלק מהפיצ'רים מתייחסים להופעתה של מילה בודדת (כמו למשל הופעתם של סימני קריאה בטוויט), אנחנו חשופים לבעיית קוליניאריות. בעיה זו נפתרת בעקבות איפוס רוב הפיצ'רים (ובאופן ספציפי אלו הגורמים לקולינאריות). סה"כ הורדנו בעזרת לאסו את מס' הפיצ'רים לכ-3000 (הדבר תלוי בפרמטר למבדא, בתחילה ניסינו אחד ולאחמ"כ ירדנו לחצי).

בחיפוש אחר clasiffier מתאים נתקלנו באינטרנט במאמר הבא:
Document author classification using Generalized Discriminant Analysis (moon, howland, gunther)
שהעלה את האפשרות ששימוש ב-bag of words שמכיל את כל המילים שמופיעות באחד המסמכים ובשיטה של discriminant analysis, יאפשר להגיע לתוצאה אופטימלית. ניסינו שיטה זו והגענו לתוצאות סבירות, אבל פחות טובות משמעותית מאשר עם ה-בGLM.

<u>הערה:</u> רצינו להשתמש ב k-cross validation אבל כשניסינו לאמן את הלומד על כל הדטה הוא רץ הרבה מאוד זמן ולא הספקנו לעלות על הבעיה.