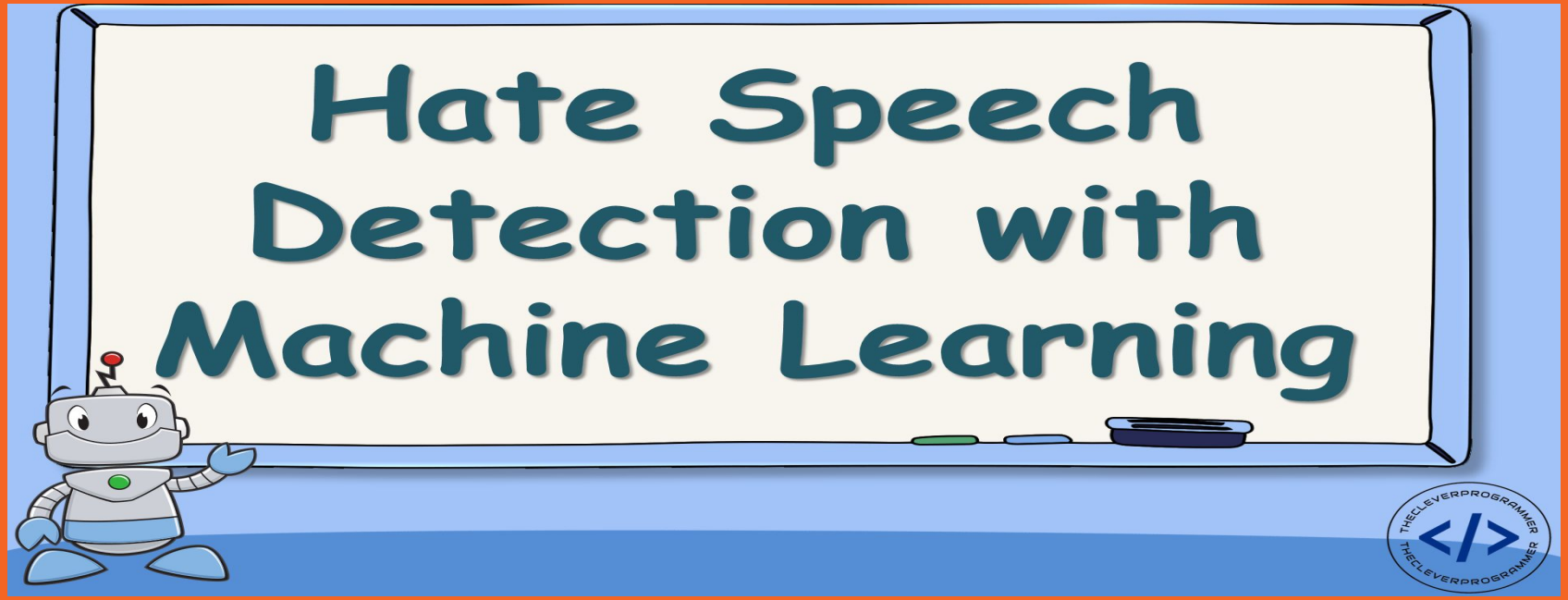


By Niv & Chen - Haifa university

מטרה: לזהות בין ציוץ רע לציוף טוב

גישה:

גישה בסיסית לזיהוי ציוצי "רוע" היא שימוש בגישה מבוססת מילות מפתח. נרצה לתחזק כל הזמן מאגר של מילות מפתח אשר יעזרו לנו לסווג דברי שנאה. נשים לב שהטרמינולוגיה משתנה כל הזמן ונוספות לנו דברי "שנאה" מדי יום ולכן אחת הבעיות העיקריות בנושא זה היא שעלינו לתחזק כל הזמן את הדאטה לפי המתרחש ברחבי הגלובוס.

נשים לב : הכללת מונחים שיכולים אך אינם תמיד מעוררי שנאה (למשל, "זבל", "חזירים וכו') תיצור יותר מדי אזעקות שווא, מה שיבוא על חשבון אחוז הדיוק. (tradeoff)

דאטה ככלי לפתרון :

מידע נוסף ממדיה חברתית יכול לעזור להבין יותר את המאפיינים של הפוסטים ואולי להוביל לגישת זיהוי טובה יותר. מידע כגון דמוגרפיה של המשתמש המפרסם, מיקום, חותמת זמן, או אפילו מעורבות חברתית בפלטפורמה יכולים כולם לתת הבנה נוספת של הפוסט בפירוט שונה.

1. Data:

לקחנו את הדאטה המוצעת במודל. הדאטה היא קובץ
csv המורכבת ממשפטים והתוצאה של כל משפט, 1
למשפט רע ו 0 למשפט טוב.

דוגמא למשפט טוב ורע מתוך הדאטה :

Label = 1

@user this man ran for governor of ny, the state with the biggest african-american population #Ã¢Â€Â¦

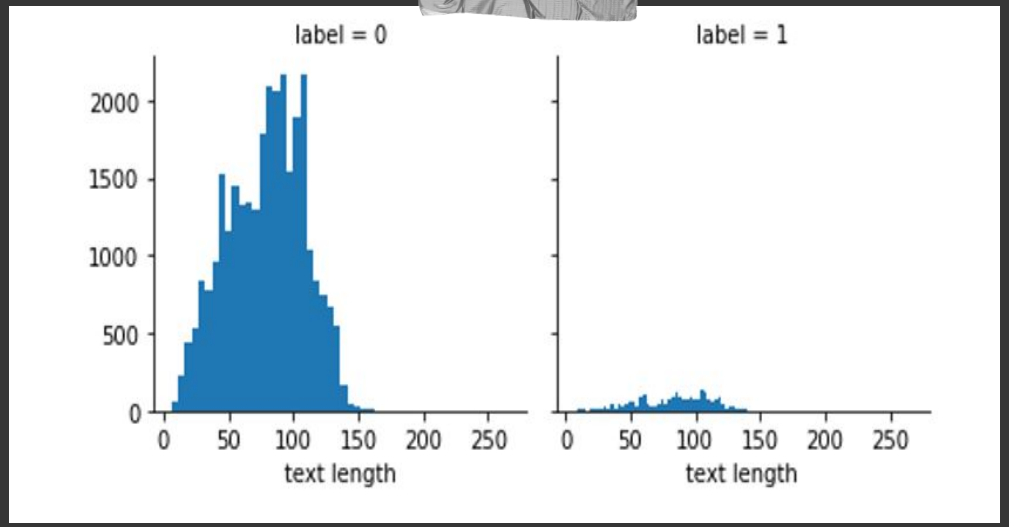
Label = 0

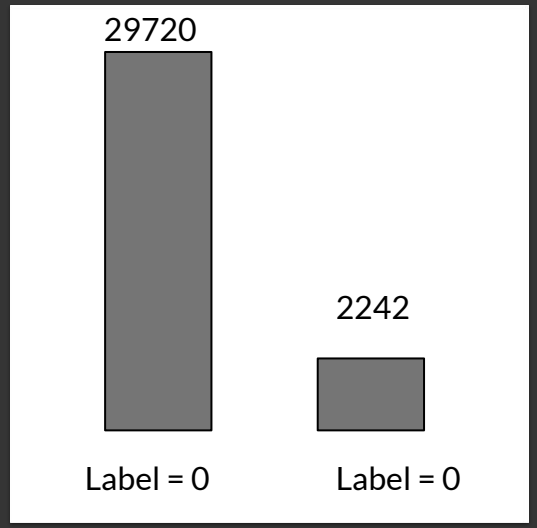
already bought my finding dory ticket

ניקוי הדאטה מרעשים :

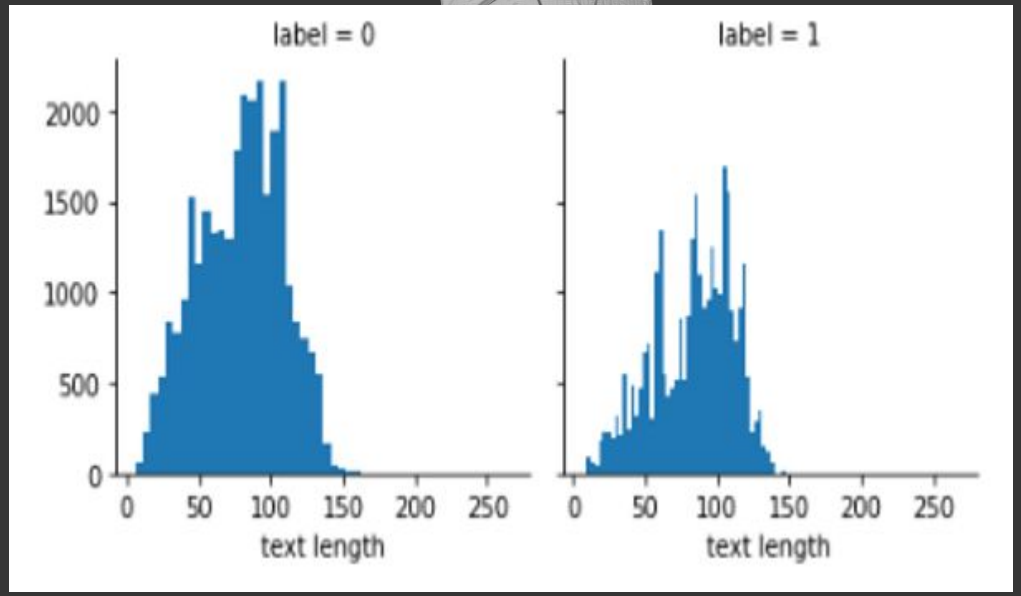
-מחיקת שכפול שורות ,משפטים משוכפלים.
-מחיקת שורות ריקות.
-העברת אותיות לגודל אחיד (כולם לאותיות קטנות).
-מחיקת רעשים בתוך המשפט : סימנים מיוחדים , רווחים מיותרים , מספרים ועוד..
-fixing imbalance in data

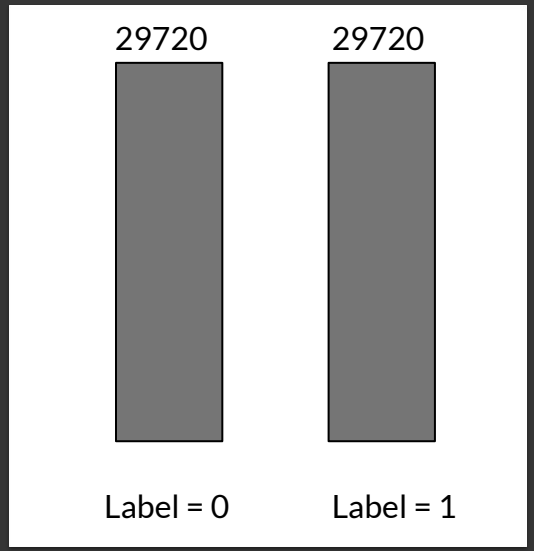
Before fixing imbalance :



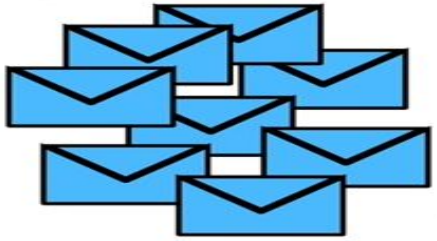


After fixing imbalance:





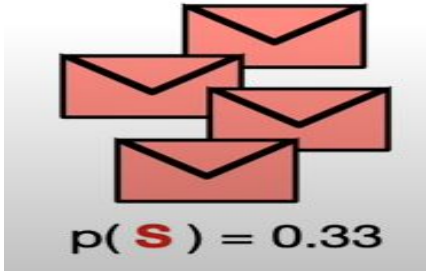
Dear Friend



p(N) = 0.67

Prior

$$p(N) = \frac{8}{8+4} = 0.67$$

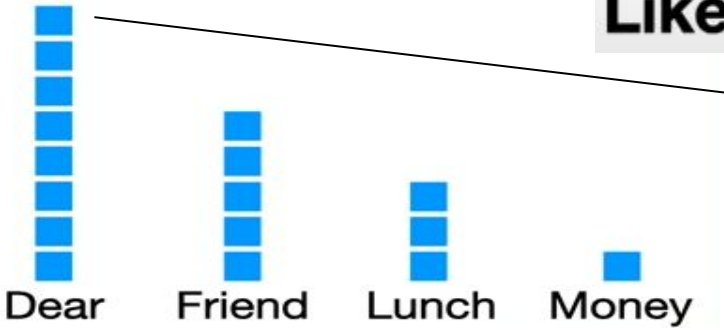


p(S) = 0.33

$$p(N) \times p(\textbf{Dear} \mid N) \times p(\textbf{Friend} \mid N) = 0.09$$

$$p(S) \times p(\textbf{Dear} \mid S) \times p(\textbf{Friend} \mid S) = 0.01$$

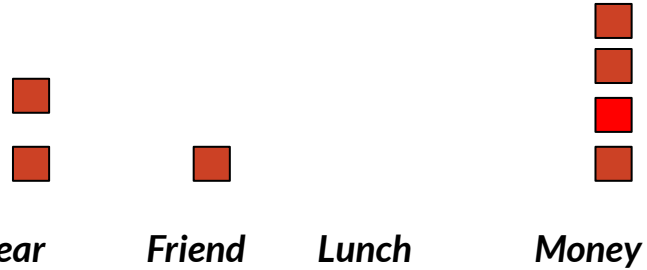
Likelihoods



$$p(\text{Dear} \mid \text{Normal}) = \frac{8}{17}$$

$$p(\text{Dear} \mid \text{N}) = 0.47$$

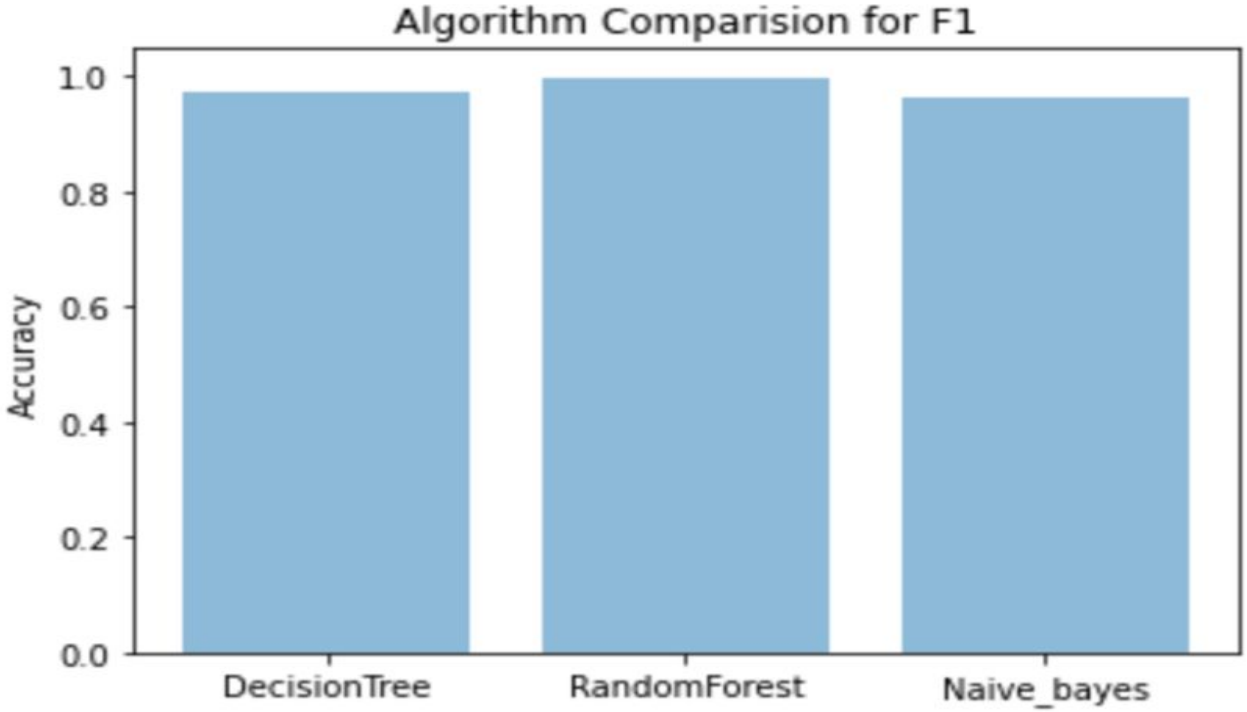
$$p(\text{Friend} \mid \text{N}) = 0.29$$



$$p(\text{Dear} \mid \text{S}) = 0.29$$

$$p(\text{Friend} \mid \text{S}) = 0.14$$

```
DecisionTreeClassifier: 0.9713069830421046
MultinomialNB: 0.9608490260790784
RandomForestClassifier: 0.9978141352586607
```



התאמה ובחירת המודל :

-נבחין כי random forest מדייק באחוזים הכי גבוהים אך נראה כי אנחנו נמצאים במצב של overfitting לקבוצת ה train שלנו ולכן לא נשתמש במודל זה.

-נרצה להשתמש במודלים עם bais יותר גדול אל מול קבוצת ה train אך ה variance שלהם יהיה קטן אל מול קבוצת ה test בשונה מ random forest.
לכן נבחר מודל כמו decision tree / naive bayes classification אשר מניב אחוזי דיוק גבוהים אך אינו overfitting לקבוצת ה train שלנו.