מגישים:
עילי פרידמן 305709248
אורן שור 204053076

מעבדה #3: Sentiment Analysis

שלב א': Preprocessing

(שם הפונקציה אשר מבצעת את התהליך: preprocceing_tweet).
בשלב זה התחלנו בניקוי המידע השונה אשר הגיע מ-dataset. ראשית יצרנו 2 regex-ים מרכזיים אשר ינקו את התווים אשר לא רלוונטיים לצורך ההמשך כמעט בכל משימת data analyst (סימני פיסוק, וכו'..) ובמקומם הכנסנו רווחים במידת הצורך.
בהמשך זיהינו כי כל ציוץ מתחיל עם תיוג של כותב הציוץ וכי זהו נתון מיותר ולכן הורדנו אותם גם כן.
ולבסוף ניקינו את כלל הרווחים הכפולים השונים שהיו קיימים.
באופן טבעי דאגנו לא לייחס חשיבות גם לרשימה של מילים שכיחות בטקסט (stop words), אך גילינו שהורדה זו רק מורידה את אחוזי ההתאמה השונים (ועל כן לא הורדנו אותם בסופו של דבר). כמו כן, שימוש במילון קיצורים (כמו גם מילון סלנג ותרגום של סמיילים שונים שהיו שכיחים גם הם בDATA) הוריד את הביצועים ובחרנו שלא להשתמש בו.

שלב ב': Feature Extraction

(שם הפונקציה אשר מבצעת את התהליך: feature_extraction_LogisticRegression)
בשלב זה מטרתנו הייתה למצוא את החבילה (האלגוריתם) היעילה ביותר לבניית מודל מדויק ככל שניתן. ניסינו להריץ מודלים שונים אשר התבססו על תבניות של naïve Bayes, SVM, decision Tree, ו-Logistic Regression. לבסוף לאחר בדיקה ארוכה עם פרמטרים הצלחנו להניב תוצאות אופטימליות (אחוז דיוק גבוה ביותר) עם חבילת Logistic Regression ואיתה המשכנו.
נציין כי פלט הדיוק לאחר ביצוע 10 cross validation היה:

```
[0.78140232, 0.77965254, 0.78602675, 0.78075, 0.775, 0.77722215, 0.77572197, 0.77834729,
0.7775972, 0.7799725]
```

מכאן שממוצע הדיוק הינו הממוצע: 0.7791
ניסינו לחשוב על פיצ'רים שונים אשר יניבו דיוק יותר מתאים: החל מכמות סימני הקריאה/השאלה הנמצאים בציוץ, כמות האותיות הגדולות אשר נמצאות בציוץ ועד לאורך הציוץ ולכמות האותיות הכפולות שבו. עשינו בדיקה כדי לראות כמה אחוזים מכלל הDATA מייצגים קבוצות אלו וראינו שהאחוז מאוד קטן ועל כן ההשפעה תהיה מינורית מאוד והחלטנו שלא להוסיף עוד פיצ'רים.

שלב ג': Classifier

(שם הפונקציה אשר מבצעת את התהליך: Classifier)
לאחר שבנינו את המודל המיטבי שהצלחנו למצוא, הרצנו את המודל על קובץ הTRAIN שניתן ולאחר מכן יצרנו פלט של דיוק כלל הסיווגים שהמודל קבע.
ע"פ דרישות העבודה, היינו צריכים להציג את ערכים ה-precision, recall ו-accuracy של לפחות 3 מסווגים שנבחנו.
כשקראנו על הנושא , ראינו שניתן לקבל ערכים אלו ב-3 צורות (macro, micro ו-weighted שהם בהתאמה על כלל החיזויים, על כל תווית עם משקל ובלי משקל) , נציג את שלושת הסוגים בכל אחד משלושת הסיווגים:

סיווג	סוג הנתון	Precision value	Recall value	Accuracy value
Logistic Regression	Macro	0.97915	0.97866	0.97890
	Micro	0.97926	0.97926	0.97926
	Weighted	0.97925	0.97926	0.97925
decision Tree	Macro	0.99630	0.99604	0.99617
	Micro	0.99623	0.99623	0.99623
	Weighted	0.99623	0.99623	0.99623
naïve Bayes	Macro	0.88600	0.86188	0.86836
	Micro	0.87382	0.87382	0.87382
	Weighted	0.88041	0.87382	0.87179

ניתן לראות כאמור שכלל הפרמטרים מקבלים את הערכים הגבוהים ביותר תחת הסיווג של Logisitc Regression . על כן , התוצר מסיווג זה מדמה וחוזה בצורה מיטבית את הנתונים שבקובץ ה-TRAIN ובו נבחר להשתמש כדי ליצור מודל שנריץ על קובץ הTEST.

נציין כי מצורכי העבודה השתמע כי יש להציג בדו"ח פירוט על קובץ ה-TRAIN . על כן כך כתוב הקוד.
לתחרות שלחנו כמובן את קובץ פלט הסיווג בעבור קובץ ה-TEST