מגישים : שי ארץ קדושה – 203276258, חן ארזי -307875633

דו"ח עבודה 3- עקרונות שפות תכנות

תהליך יצירת מודל הסיווג:

במטלה זו התבקשנו לפתור את משימת Sentiment Analysis בהתבסס על מידע מהרשת החברתית Twitter. נפרט את 3 שלבי יצרת מודל הסיווג :

1. Preprocessing-

- **ניקוי הציוץ-** אנו מנקים את הציוץ מסימונים לא אינפורמטיביים כגון : ?,*, \,# וכדומה וכמו כן מ-URL שונים. שלב זה מחזיר לנו רשימה של tokens נקיים של הציוץ.
- **הסרת stop words-** מעבר על ה-tokens השונים והורדת stop words-
בשלב זה עשינו ניסויים רבים באשר לאילו מילים להסיר מה-data הנתון לנו. כאשר השתמשנו ב-stop words מוכנות מראש של חבילת nltk ראינו כי הביצועים פחות טובים.
לאחר התעמקות במילים המוכנות שהחבילה מייצרת ראינו כי ישנן מילים רבות שאינפורמטיביות למשימה שלנו. בין המילים מצאנו מילים כמו: not, no, hasn't, hadn't, aren't, didn't ומילים שליליות נוספות בזמנים שונים של השפה האנגלית.
על פי הניתוח שלנו, מילים אלו מסייעות בהבנת הניתוח הרגשי ויתכן מאוד כי הסרתן פגעה בביצועי המשימה.
על כן- בנינו מילון חדש של stop words המכיל מילים שלא פוגעות בביצועי המשימה שלנו.
- **Lemmatization vs stemming-** הפיכת המילים לצורה אחידה של מילים דומות אחרות.
בבואנו לכתוב את החלק הזה- ניסינו לעשות stemming וגם lemmatization. ראינו כי כאשר אנחנו מבצעים stemming – קיצור המילים, אנו מקבלים תוצאות טובות אך כאשר אנחנו עושים- lemmatization, הפיכת המילים לצורה השורשית שלהן אנו מקבלים תוצאות טובות אפילו יותר. הדבר הגיוני כי הפיכת המילים למילים בעלות משמעות (מילים שקוצרו למילה השורשית שלהן)- מניבות מידע יותר אינפורמטיבי מאשר מילים מקוצרות ללא משמעות.

על מנת להעריך מדדי Accuracy, Recall ו-Precision של כל אחד מהמודלים בשלב הסופי- ראשית חילקנו את ה- train data לשני חלקים:

- Validation
- Train

אימנו את המודלים על חלק ה-Train לאחר הפיצול וניבאנו את מדיי תוצאות השנים עבור נתוני ה-validation. הגדרנו את החלוקה להיות 20% בעבור validation ו-80% train כנהוג בהערכת מודלי למידת מכונה. חלוקה זו מיטבית מפני שמצד אחד, אנו לא רוצים "לבזבז" מידע רב מידי על בדיקה ומצד שני לא נרצה להמעיט מידיי בנתוני validation כדי שהמדדים שנרצה להעריך יתנו תמונה נכונה בנוגע לדיוק המודל אותו אימנו.

2. Feature Extraction-

על מנת לבחור את הפיצ'רים של המודל שלנו, ראשית השתמשנו ב-CountVectorizer של חבילת sklearn שממיר אוסף של מסמכי טקסט (במקרה שלנו אוסף של ציוצים) למטריצה שסוכמת את כמות הפעמים של כל token.

לאחר מכן הכנסו את המטריצת הסכימות המתוארת הנ"ל לתהליך שממיר אותה למטריצה מנורמלת ע"י שיטת Tfldf. תהליך זה בוצע ע"י TfidfTransformer של חבילת sklearn גם כן.

Tfldf הוא מדד שחוזה את חשיבותה של מילה מסוימת עבור מסמך בקורפוס של מסמכים רבים. (במקרה שלנו חשיבות המילה בציוץ מסוים על פני קורפוס של ציוצים רבים). המדד מגדיל משקל של מילה על בסיס כמות הפעמים שהיא מופיעה ומקטין את המשקל שלה על בסיס כמות המסמכים בהם היא מופיעה.

המשקלים של ה-tokens במטריצה המנורמלת הם הפיצ'רים שבחרנו עבור המודל שלנו.

מגישים : שי ארץ קדושה – 203276258, חן ארזי -307875633

3. Classifier-

לאחר בחירת הפיצ'רים של המודל, הגענו לשלב הסיווג. בדקנו 3 מודלי סיווג שונים-

- Logistic regression
- Naïve bayes
- Decision tree

והשתמשנו בטכניקת Cross Validation 10 בשביל להגיע לתוצאות הטובות ביותר של האימונים של כל אחד מהמודלים.

עבור כל אחד משלושת המדדים של כל אחד מהמודלים, קיבלנו מערך המכיל 10 ערכים שנוצרו ב-10 איטרציות השונות מטכניקת Cross Validation 10. יצרנו ממוצע של כל אחד מהמדדים וכך קיבלנו את התוצאה הסופית של כל מדד בעבור כל מודל.

בחרנו להתמקד בממוצע של ה- Accuracy ועל פיו לבחור את המודל הטוב ביותר מכיוון שראינו שבשלב ה- evaluation בהגשה ל-kaggle, מתמקדים במדד זה.

4. Results-

Logistic regression-

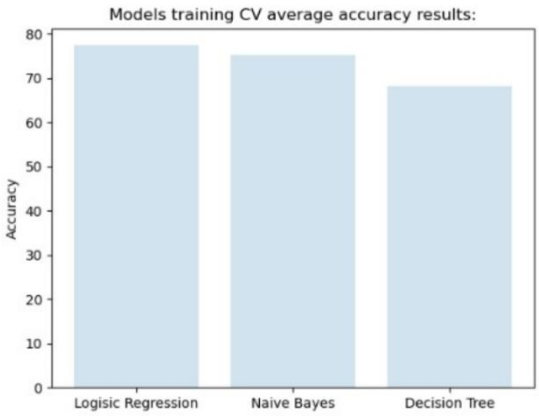
```
The average accuracy for 10 CV is:  77.406
The average precision for 10 CV is:  79.285
The average recall for 10 CV is:  81.237
```

Naïve bayes-

```
The average accuracy for 10 CV is:  75.348
The average precision for 10 CV is:  72.305
The average recall for 10 CV is:  91.371
```

Decision tree-

```
The average accuracy for 10 CV is:  68.311
The average precision for 10 CV is:  72.72
The average recall for 10 CV is:  70.277
```



בגרף ניתן לראות כי המודל שביצע Cross Validation 10 עם ממוצע ה- accuracy הגבוהה ביותר הוא Logistic regression.

בעקבות כך, בחרנו את המודל הזה ואימנו אותו על סט האימון ללא Cross Validation וביצענו predication על נתוני ה- validation.

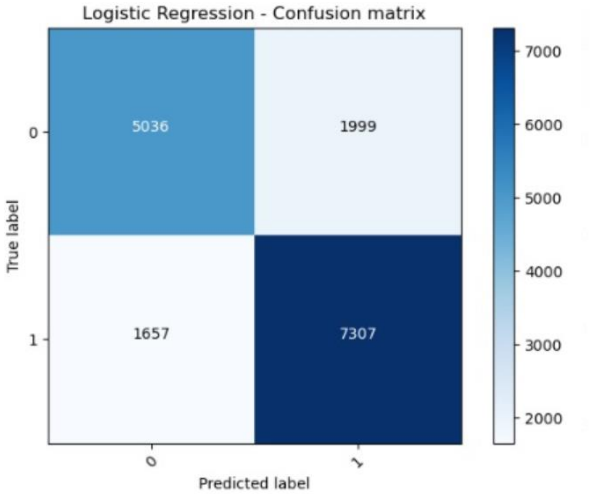
עבור החיזוי שקיבלנו על נתוני ה-validation , הוצאנו classification report שמציג את המדדים: Recall, Precision ו-f1 score (מדד שמשקלל את ה-Recall וה-Precision).

```
Accuracy: 0.7714857178573661
Classification Report:
              precision    recall  f1-score   support

           0       0.75      0.72      0.73      7035
           1       0.79      0.82      0.80      8964

    accuracy                           0.77     15999
   macro avg       0.77      0.77      0.77     15999
weighted avg       0.77      0.77      0.77     15999
```

בנוסף, הצגנו את ה-accuracy של המודל הנבחר יחד עם confusions matrix שבעזרתה ניתן לזהות את שגיאות המודל:



5. <u>תחרות Kaggle –</u>

פתחנו משתמש באתר של Kaggle בשם: The_Kaggels.

אל המודל הנבחר הכנסנו את נתוני קובץ המבחן וקיבלנו חיזויים עבור משימת ה- sentiment analysis. הצמדנו את מספר המזהה של החיזויים לחיזויים עצמם והעלנו את הטבלה לאתר.

התוצאה שקיבלנו הינה:

submission_file.csv 0.77926

3 days ago by Shay Eretz Kdosha

add submission details

מגישים : שי ארץ קדושה – 203276258, חן ארזי -307875633

ניתן לראות את מיקומנו בטבלה:

Overview Data Notebooks Discussion Leaderboard Rules Team My Submissions **Submit Predictions**

#	Team Name	Notebook	Team Members	Score	Entries	Last
1	danmaestro			0.85701	1	1mo
2	try&catch			0.83318	29	1d
3	Ninaiway			0.83218	4	8d
4	Roy & Tal			0.82784	13	20h
5	機械学習			0.81759	3	2d
6	bibi			0.79760	2	5d
7	Bushot Inc			0.78668	13	7d
8	mtmt1_mtmt3			0.78251	33	9d
9	tevel events			0.78210	7	4d
10	Gal Haviv			0.78151	31	3d
11	The_Kaggels			0.77926	20	1d

לסיכום,
משימת sentiment analysis הינה משימה של מציאת השיוך הרגשי של מסמכים שונים.
אנו ביצענו את המשימה על ציוצים מרשת ה-twitter.
שמנו לב להבדלים השונים בין ניתוח ועיבוד ציוצים לבין מסמכים כללים אחרים.
העבודה על ציוצים הינה שונה כפי שניתן לראות בדו"ח ולכך התייחסנו גם בקוד שלנו. לאחר אימון המודלים השונים קיבלנו את המודל הטוב ביותר.
ייתכן ואם היינו עושים מודל מורכב יותר על סמך deep learning אולי היינו מקבלים תוצאות טובות יותר. וכמו כן, כבכל מודל- אם היינו מוסיפים data נוסף שהמודל יוכל ללמוד ממנו, יכולנו להגיע גם כן לביצועים טובים עוד יותר.