שלב ראשון:

1. לבחור 3-4 אלגוריתמי למידה: random forest ,mlp (deep learning), svm, naive base, logistic regression,
2. להשתמש בbag of words, td idf ועם כל אלגוריתם להוציא תוצאה ולבנות גרף או טבלה
3. לבחור 3 אלגוריתמי וקטוריזציה הכי טובים (צריך להבין מה זה tokenization + tagging) - wordTovec/festText/GloVe/BERT ולהריץ עליהם את האלגוריתמי למידה(לנסות לאמן לבד וגם אחוזי למידה)
4. לנסות לחבר את bag of word לכל אחד מהוקטורים ולהריץ עם כל אלגוריתם למידה

שלב שני:

1. לנסות את הpart of speech:

לבדוק האם כדאי stop words+lemmaziation או בלי (אפשר גם לנסות את זה רק עם bag of words ואז רק עם הוקטורים ואז הכל ביחד)

1. לנסות את הקטע של החלקי דיבר- האם יש השפעה לחלקי דיבר
2. Sentiment analysis – לבדוק האם הוא מסווג טוב יותר מהמסווג שלנו (כלומר האם מספיק לפי הרגש) ואח״כ גם לבדוק את זה בתור תכונה
3. לבדוק האם כדאי upper case/lower case
4. לבדוק האם יש אלגוריתם של סלנג

שלב שלישי :

Feature selction

לפעמים יש יותר מדי features- מפעילים אלגוריתמים קיימים לfeature selection. ניסוי : על ציר הx מראים כמה features סיננו ובציר הy מראים את אחוז הדיוק

1. לבדוק האם כדאי upper case/lower case
2. לבדוק האם יש אלגוריתם של סלנג

דברים שכדאי ללמוד:

1. StratifiedKFold
2. מודל שפה bert- האם יכול לעזור במשהו
3. את הדרכים לעשות הערכה לתוצאה בsklearn- : Precision, Recall, F-Measure, Accuracy