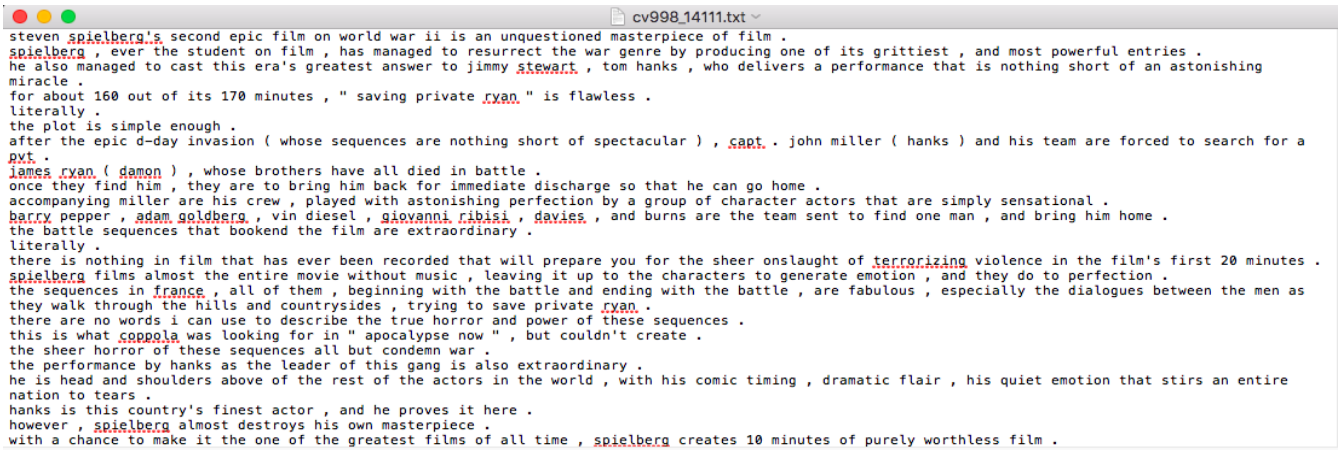
מגישים: צבי פוצ'נסקי 203195706 אלה קיטייבה 336382833

Sentiment Analysis – Home Work #3

תיאור הפעולות המתבצעות בפרויקט

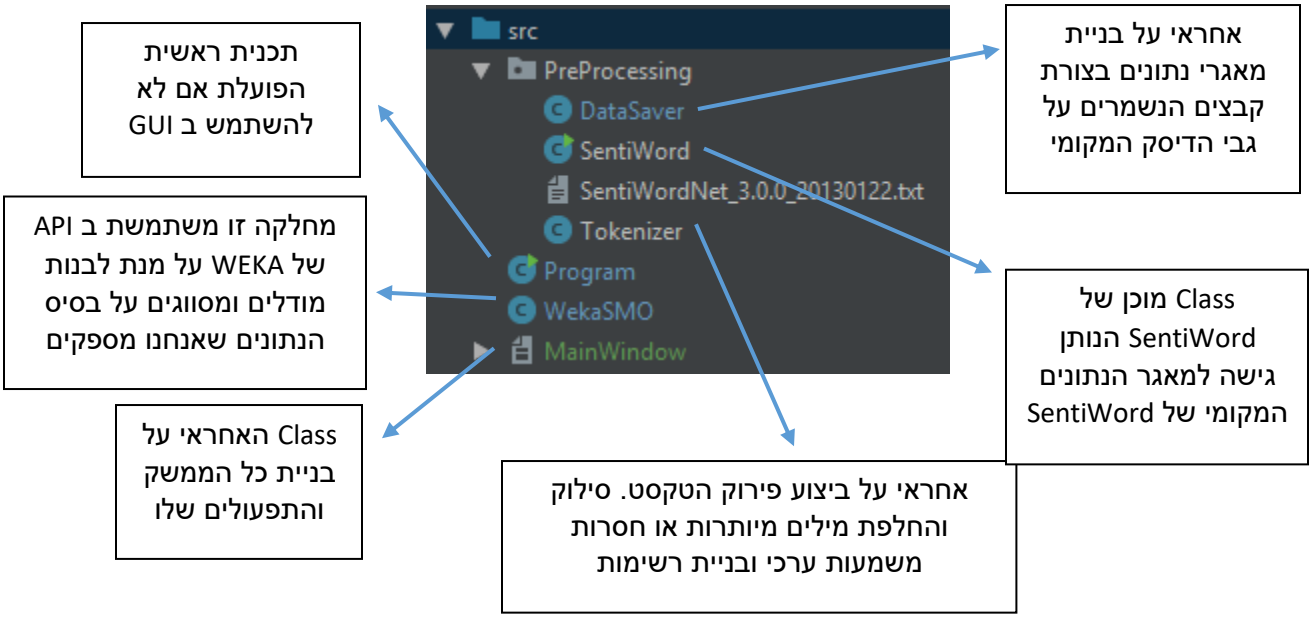
רקע

הפרויקט מתבסס על מאגר נתונים מוכן המכיל 2000 קבצי טקסט של ביקורות קולנוע הלקוחות מ IMDB.
הקבצים מסווגים מראש ל- חיוביים ושליליים בחלוקה שווה (1000 חיוביים ו- 1000 שליליים). כל קובץ מסודר מראש בצורה כזאת שכל משפט מופיע בשורה בנפרד.


cv998_14111.txt

```
steven spielberg's second epic film on world war ii is an unquestioned masterpiece of film .
spielberg , ever the student on film , has managed to resurrect the war genre by producing one of its grittiest , and most powerful entries .
he also managed to cast this era's greatest answer to jimmy stewart , tom hanks , who delivers a performance that is nothing short of an astonishing
miracle .
for about 160 out of its 170 minutes , " saving private ryan " is flawless .
literally .
the plot is simple enough .
after the epic d-day invasion ( whose sequences are nothing short of spectacular ) , capt . john miller ( hanks ) and his team are forced to search for a
pvt .
james ryan ( damon ) , whose brothers have all died in battle .
once they find him , they are to bring him back for immediate discharge so that he can go home .
accompanying miller are his crew , played with astonishing perfection by a group of character actors that are simply sensational .
barry pepper , adam goldberg , vin diesel , giovanni ribisi , davies , and burns are the team sent to find one man , and bring him home .
the battle sequences that bookend the film are extraordinary .
literally .
there is nothing in film that has ever been recorded that will prepare you for the sheer onslaught of terrorizing violence in the film's first 20 minutes .
spielberg films almost the entire movie without music , leaving it up to the characters to generate emotion , and they do to perfection .
the sequences in france , all of them , beginning with the battle and ending with the battle , are fabulous , especially the dialogues between the men as
they walk through the hills and countrysides , trying to save private ryan .
there are no words i can use to describe the true horror and power of these sequences .
this is what coppola was looking for in " apocalypse now " , but couldn't create .
the sheer horror of these sequences all but condemn war .
the performance by hanks as the leader of this gang is also extraordinary .
he is head and shoulders above of the rest of the actors in the world , with his comic timing , dramatic flair , his quiet emotion that stirs an entire
nation to tears .
hanks is this country's finest actor , and he proves it here .
however , spielberg almost destroys his own masterpiece .
with a chance to make it the one of the greatest films of all time , spielberg creates 10 minutes of purely worthless film .
```

מבנה הפרויקט



מהלך הפרויקט

מחלקת Tokenizer:

אחראית על עיבוד המקדים של הטקסט. מבצעת שינוי של המאגר על מנת לבטל סימנים מיותרים היכולים לפגוע בכושר הסינון. שינוי זה מתבצע על ידי החלפת תווים כמו: - , _ , ' וכד' לרווח רגיל כך שהמילים יתפצלו בביצוע split על בסיס רווח בצורה תקינה. לאחר מכן מתבצע split נוסף על בסיס תו מיוחד הורדת שורה.

בשלב זה נוצרת רשימה של משפטים כך שבכל תא ישנו משפט אחד מהמאגר. בשלב זה מתבצעת ריצה נוספת על הרשימה ומעבר על כל משפט ופיצולו נוסף על בסיס רווח ותוך כדי בדיקה וסילוק מילים המופיעים במאגר stop words ומילים שמחזירות ערך ניטרלי ממילון Senti Word.

מילון Senti World – זהו מילון חיצוני המגיע בצורת טקסט שמכיל בתוכו מאגר מילים גדול המסווגים על ידי ציון, הנוסחא לחישוב הציון מחושבת באופן הבא: ObjScore = 1 - (PosScore + NegScore)

השימוש ב Senti Word במחלקה נועד על מנת לסנן מילים בעלות ערך ניטרלי ידוע מראש וזאת על ידי שימוש במחלקה SentiWord המוכנה מראש שיודעת להתממשק מול המאגר ולשלוף את הציון של כל מילה הנשלחת אליה.

מחלקת DataSaver:

מחלקה זו אחראית על כל שמירת הנתונים. הנתונים נשמרים בקבצים נפרדים לכל טקסט כאשר כל קובץ מייצג בו את שם הטקסט הנבדק, ספירה של כמות המילים המופיעות לאחר הסינון, רשימה של המשפטים וטבלה המכילה כל מילה וכמות החזרות שלה בקובץ. הקבצים נשמרים בפורמט csv.

דוגמא לקובץ:

File Name	cv995_21821.txt
WordCount	297
Words List	
[0]	wow
[1]	everything funny dramatic interesting weird funny weird strikingly original
[2]	yep pretty much describes
[3]	starts like regular ends one weirdest funniest original movies seen
[4]	boggles mind wonder cannot get movies like often
[5]	one best films malkovich well one best movies
[6]	period
[7]	good movies one cannot pick time favorite
[8]	cusack plays craig schwartz man job job
[9]	schwartz completely noticeable cameron diaz looks like something off streets animal lover every kind animal think
[10]	craig finds job floor business
[11]	pry open doors open reaches floor floor just floor
[12]	old boss orsen bean
[13]	discovers little boarded hidden
[14]	curiosity opens starts toward gets sucked ends malkovich mind
[15]	shot onto side new jersey
Hashmap	
thought	1
trip	1
finds	2
pick	1
clever	1
let	1
soft	1
doubts	1
going	1
old	1
want	1
something	2
slow	1
animal	1
returns	1
explorations	1

קבצים אלו נועדו על מנת לחסוך שמירת הנתונים בתוך משתנים ונכתבו פונקציות פשוטות השולפות את הנתונים מהקבצים לפי דרישה.

במהלך בניית הקבצים נאספים כל המילים למשתנה המכיל את כל המילים שקיימות בכל הטקסטים וזאת על מנת שנוכל לבנות ייצוג אחד לכל הקבצים.

כמו כן במחלקה זו נבנים משתני tf ו – df.

tf – term frequency – סופר את מספר החזרות של מילה בביקורות מסויימת, מידע זה נשמר במשתנה tf שהוא מסוג hashtable המכיל את שם הקובץ בתור מפתח וכערך זוג של המילה ומספר החזרות שלה בקובץ.

df- document frequency – סופר את כמות המסמכים בהם מופיעה מילה מסויימת. נתונים אלו נשמרים גם כן ב Hash Table המכילה מילה בתור מפתח וערך מייצג את כמות המסמכים בהם מופיע המילה.

לאחר מכן נבנה קובץ מפורמט arff (יפורט בהמשך) הנדרש לצורך שימוש בסיפריות של weka. קובץ זה בנוי בצורה הבאה:

- כל שורה מייצגת קובץ של ביקורת.
- כל עמודה מייצגת מילה ממאגר המילים הכללי שנבנה לאחר מעבר על כל הקבצים.
- הערך הינו ערך tf-idf המחושב בעזרת נתוני ה tf ו df שנאספו קודם לכן.

$$(1 + \log f_{t,d}) \cdot \log \frac{N}{n_t}$$

N - כמות סה"כ המסמכים במאגר.
f - מס' הפעמים שהמושג חוזר במסמך ספציפי/ מס' המילים במסמך.
n - כמות המסמכים שהמילה מופיע בהם.

מחלקת WekaSmo:

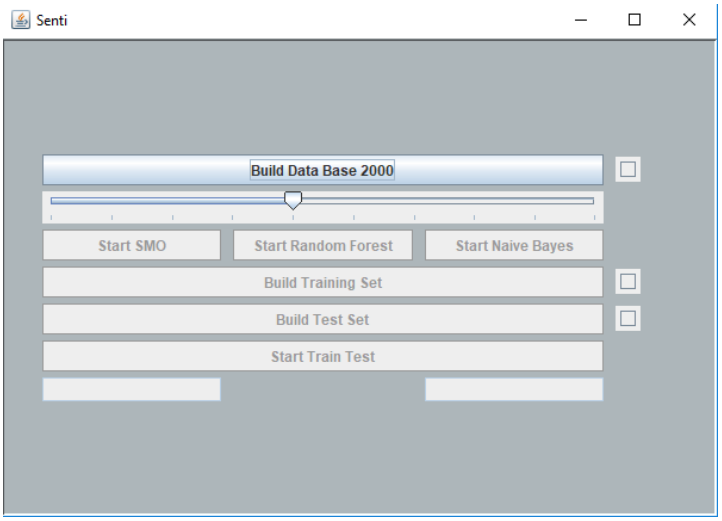
מחלקה זו אחראית על תפעול הסיפרייה השייכת ל Weka. Weka היא תכנה היודעת לבצע משימות הקשורות לניתוח נתונים כמו בניית מודלים, ביצוע סיווגים, הערכת מודלים שונים וכו'.

במחלקה זו יש מספר פונקציות היודעות לבנות מודלים ממספר סוגים: SVM, Random Forest, Naïve Bayes. על כל מודל ניתן לבצע הערכה מסוג Cross Validation על מספר folds שהמשתמש בוחר בטווח של 1-10 (הערך המומלץ הינו 10). הערה: כל פעולה כזו לוקחת המון זמן לביצוע ההערכה, תוצאות מצורפות בהמשך.

פונקציה נוספת שניתן לביצוע במחלקה זו הוא הרצת בדיקה של סיווג למודל SVM. ניסוי זה מתבצע על ידי חילוק מאגר הנתונים שלנו ל 30%-70% כאשר 70% מוגדר בתור training set ו – 30% בתור test set. על ה training set נבנה מודל חדש של SVM ומתבצעת השמה על ה 30% ומתקבל קובץ מסוג csv המציין את מספר הקובץ ואת הסיווג שהמודל נתן לו. התוצאות הן: 310 positive; 290 negative כאשר בפועל הסיווג צריך להיות 300 positive; 300 negative.

מחלקת MainWindow:

הינה מחלקת GUI המוצגת למשתמש ודואגת להרצה מסודרת של הפרויקט. הממשק בתחילת העבודה נראה כך:



מודלים

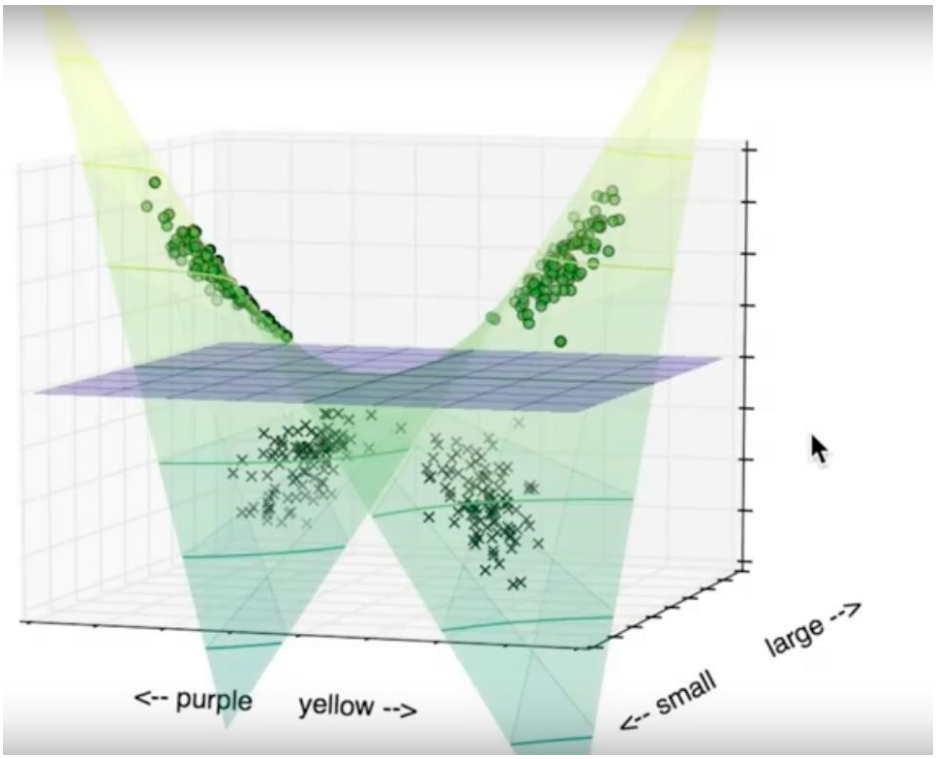
SVM – Support Vector Machine

בפרויקט זה בחרנו להתבסס על מודל מסוג SVM.

SVM שייך לקבוצת המודלים מסוג Supervised Learning ופועל על בסיס יצירת מרחבי וקטורים כאשר כל קובץ טקסט הוא מרחב בפני עצמו המכיל מילים, ערכי tf-idf בהתאמה וסיווג הקובץ הידוע מראש. השאיפה של המודל היא למצוא מרחב (Plane) המפריד בצורה הטובה ביותר בין הערכים ומרחב זה גם צריך לקיים את המרווח המקסימלי האפשרי בין וקטורי התמיכה (Support Vectors) כך שנוצר מעין כביש המפריד בין התצפיות בצורה הטובה ביותר ואין עליו שום תצפיות. ככל שמרחק בין המרחב המפריד לבין וקטורי התמיכה יהיה גדול יותר כך ישתפר הדיוק של המודל.

כאשר מסווגים רשומות חדשות ניתן למקם אותן באופן ברור לאיזה קבוצה שייך הוקטור חדש על בסיס מרחב ההפרדה שנמצא במודל.

בפרויקט זה נבנה המודל ומחושב המרחב המפריד, על ידי סיפריית ה WEKA על בסיס הנתונים שמסופקים לו לאחר ביצוע ה pre processing ובניית קובץ ה arff המכיל את כל הוקטורים.

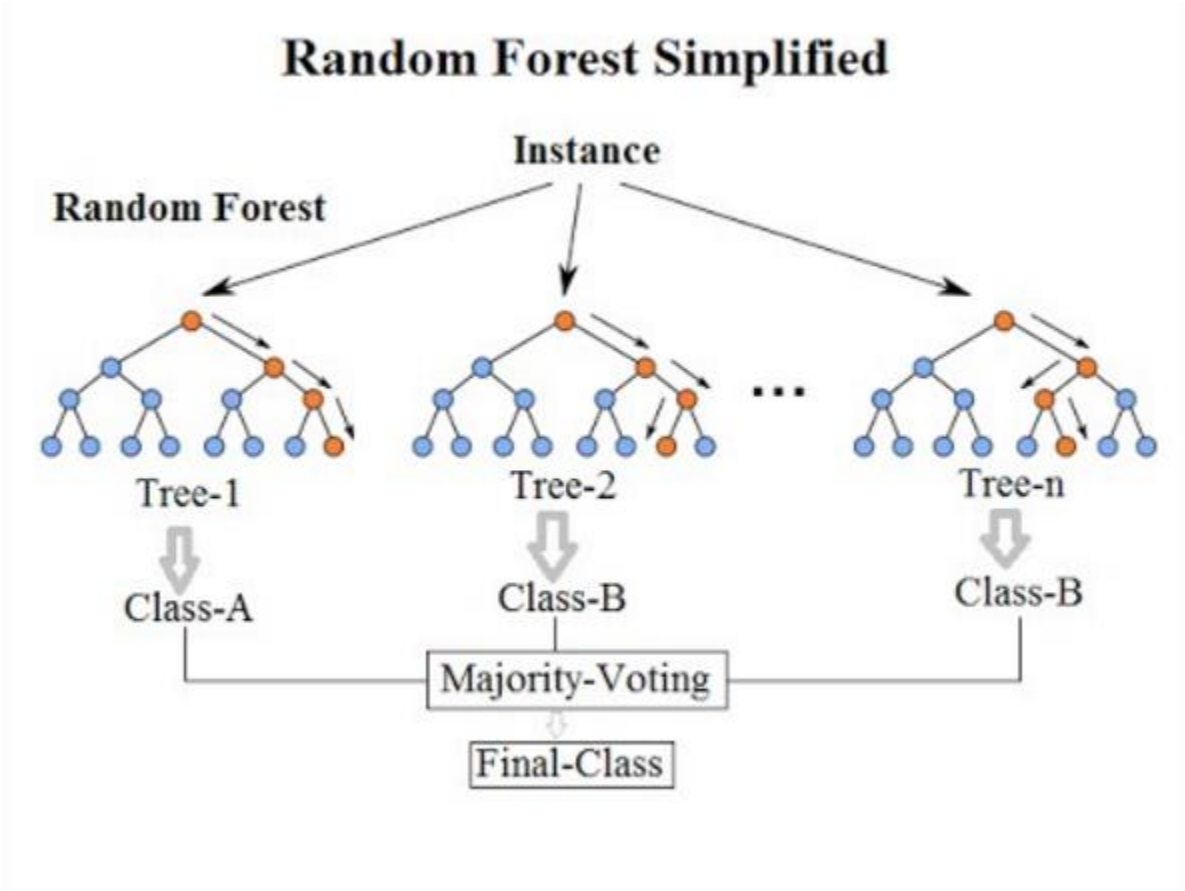


Random Forest

מודל נוסף שנבדק דיוקו במהלך הפרויקט הוא ה – Random Forest. Random Forest הוא אוסף של Decision Trees כאשר כל עץ החלטה נותן את החלטתו ובאוסף של כל ההחלטות על פי חוק הרוב נקבע הסיווג הנדרש.

ב Random Forest נוצר אוסף של עצי החלטה כאשר כל עץ החלטה בוחר בצורה רנדומלית מאפיינים שלפיהם הוא מבצע את הפיצול (Split) בעץ בכל רמותיו.

מה שיוצר אוסף של עצים שבודקים בצורה רנדומלית על בסיס בחירת מאפיינים את הרשומה החדשה שמסווגת. מודל זה בנוי על עיקרון אוסף של "לומדים חלשים" הבונה "לומד חזק" ומחזיר תשובה.



Naive Bayes

מודל נוסף שמתבצעת עליו בדיקת דיוק לצורך השוואה בין המודלים הינו מודל ה Naïve Bayes. זהו מודל סטטיסטי המניח כי אין תלות בין כל המופעים.

המודל נבנה על ידי חישוב סטטיסטי של הסיכוי של מאפיין מסויים להופיע ביחס לסיווג בינארי של "חיובי" במקרה של הפרויקט זה ובהתאם גם הסיכוי של המאפיין להופיע ביחד ל "שלילי".

לאחר מכן בסיווג רשומה חדשה, המודל לוקח את כל המאפיינים כאשר בפוריקט זה אלו המילים ומחשב את כל ההסתברויות בתנאי לסה"כ המופעים של מסמכים חיובים ובתנאי לסה"כ המופעים של מסמכים המסווגים בתור שליליים ובודק מי מהם נותן הסתברות גבוהה ועל בסיס זה בוחר לסווג את הרשומה החדשה.

Cross Validation

על מנת לבדוק את דיוק המודלים ולתת אינדקציה איזה מהם יותר מתאים לנתונים נשתמש בבדיקת cross validation על כל מודל ונוכל להשוות את התוצאות.

שיטה זו עובדת בצורה שהיא מחלקת את הנתונים לקבוצות ואז לוקחת כל קבוצה בתור test data ואת שאר הקבוצות בתור training data, בונה מודל ומריצה את ה test data על המודל שנבנה.

לאחר ביצוע של הסיווג נשמרים התוצאות ומתבצעת שוב הבדיקה רק הפעם נלקחת קבוצה אחרת שתייצג את ה test data ושאר הקבוצות בתור training data כך שמתבצעות מספר איטרציות עד שכל קבוצה ייצגה את קבוצת הבדיקה.

לבסוף מתבצע חישוב הסוכם את כל התוצאות ויש נתון מייצג של דיוק המודל.

תוצאות ניסויים

SVM – תוצאת הרצת cross validation מראה על דיוק של 84.65%

```
Weka SMO Activated!
Finished Get Data set!
Finish Initialize data, Eval, Smo
Start EVAL !
Started EVAL for 10 Folds...

Correctly Classified Instances        1693              84.65   %
Incorrectly Classified Instances       307              15.35   %
Kappa statistic                          0.693
Mean absolute error                      0.1535
Root mean squared error                  0.3918
Relative absolute error                 30.7    %
Root relative squared error             78.3582 %
Total Number of Instances             2000
```

– Random Forest תוצאה במודל זה עומדת על 77.3%

```
Weka Random Forest Activated!
Finished Get Data set!
Finish Initialize data, Eval, Random Forest
Start EVAL !
Started EVAL for 10 Folds...
### Time Eval: 1377018

Correctly Classified Instances        1546              77.3    %
Incorrectly Classified Instances       454              22.7    %
Kappa statistic                          0.546
Mean absolute error                      0.4415
Root mean squared error                  0.4486
Relative absolute error                 88.306  %
Root relative squared error             89.7165 %
Total Number of Instances             2000
```

1377018 ms = 22.95 min

<u>סיבה לפגיעה באחוזי דיוק –</u> על מנת שאחוז הדיוק יהיה גבוהה נדרש כמות גדולה של עצי החלטה, במקרה זה הבדיקה בוצעה על ערך דיפולטיבי העומד על 100 עצי החלטה. (זמן ריצה עמד על 23 דקות)

בוצע ריצה נוספת על בדיקה זו, לשם בחינת השפעה של שינוי בכמות עצי החלטה הוחלט על ניסוי של 500 עצי החלטה. התקבלה תוצאה של 81.95%, זמן ריצה עלה ל118 דקות.

מה שמראה על אחד העיקרונות של random forest שאפשר להעלות את הדיוק של המודל על חשבון זמן ריצה וזכרון.

```
Weka Random Forest Activated!
Finished Get Data set!
Finish Initialize data, Eval, Random Forest
Start EVAL !
Started EVAL for 10 Folds...
### Time Eval: 118

Correctly Classified Instances        1639               81.95   %
Incorrectly Classified Instances       361               18.05   %
Kappa statistic                          0.639
Mean absolute error                      0.4419
Root mean squared error                  0.4468
Relative absolute error                 88.3846 %
Root relative squared error             89.35   %
Total Number of Instances             2000
```

<u>Naïve Bayes –</u> תוצאה של הבדיקה על מודל זה עומדת על 69.05%

```
Weka Naive Bayes Activated!
Finished Get Data set!
Finish Initialize data, Eval, Naive Bayes
Start EVAL !
Started EVAL for 10 Folds...

Correctly Classified Instances        1381               69.05   %
Incorrectly Classified Instances       619               30.95   %
Kappa statistic                          0.381
Mean absolute error                      0.3092
Root mean squared error                  0.5557
Relative absolute error                 61.835  %
Root relative squared error            111.1475 %
Total Number of Instances             2000
```

<u>סיבה לפגיעה באחוזי הדיוק -</u> ייצוג הנתונים בצורת tf-idf אינה מתאימה למודל הזה, משום שהמודל צריך לספור את מספר המופעים של כל מילה ביחס לסיווג חיובי או שלילי ולכן נפגעים החישובים ההסתברותיים שעליהם מסתמך המודל וכך נפגע אחוז הדיוק של המודל.