

# פרויקט גמר - יישומי למידה עמוקה

*תגובות רעילות ברשת - זיהוי וסיווג לפי רמות רעילות*

*Toxic Comment Classification - march 2023*

Rotem Ecker 206015240

Mor Schenavsky 300790839

Brit Lev 312478266

Sivan Itzhaki 207232570

# הקדמה

## רקע

בשנים האחרונות אנו שומעים יותר ויותר על מקרים של בריונות ברשת באתרים ורשתות חברתיות שונות. בפרויקט זה עסקנו בנושא תגובות רעילות של משתמשים ברשת - זיהוי תגובות רעילות וכן זיהוי סוגי הרעילות השונות:

- toxic
- severe-toxic
- obscene
- threat
- insult
- Identity-hate

## הדאטה

את הדאטה לפרויקט מצאנו באתר Kaggle מתוך אתגר שפורסם לסיווג תגובות רעילות. הדאטה מכיל 159,571 רשומות של תגובות משתמשים בוויקיפדיה אשר סווגו ידנית ע"י מדרגים אנושיים כרעילים לפי סוגי הרעילות השונים שצוינו לעיל.

## הבעיה

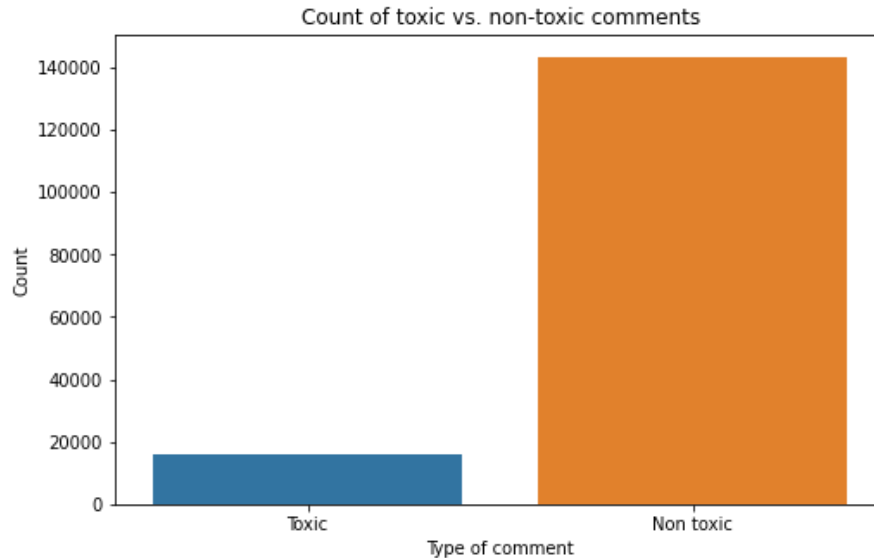
בהינתן תגובת משתמש, נרצה שהמודל יקבע אילו סוגי רעילות שייכות אליה. תגובה יכולה להיות מסווגת לכמה רמות רעילות שונות ולכן זוהי בעיה מסוג Multi-Label Classification.

בנוסף, חשוב לציין שלאחר בניית המודלים לפרויקט זה, בחשיבה על הפרויקט כסוג של מוצר ראינו לנכון ליצור ממשק משתמש אליו משתמש יכול להזין תגובה שנשלחת לבדיקה על ידי שני מודלים: המודל הראשון בודק האם הבדיקה רעילה או לא. אם התגובה רעילה אז במודל השני ייבדקו מהן סוגי הרעילות המתאימות לה.

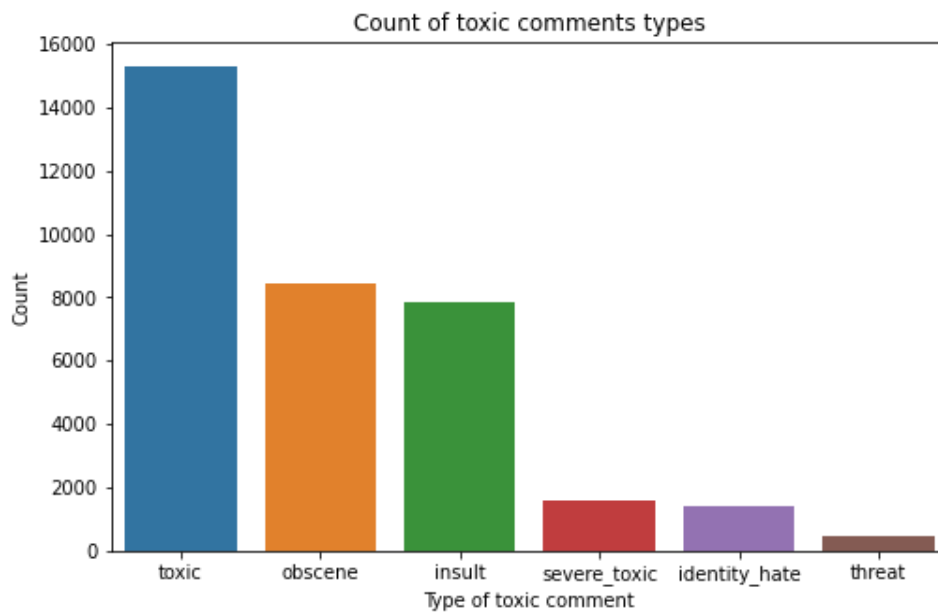
## תהליך העבודה

### חקירה ראשונית של הדאטה - תובנות ומסקנות עיקריות

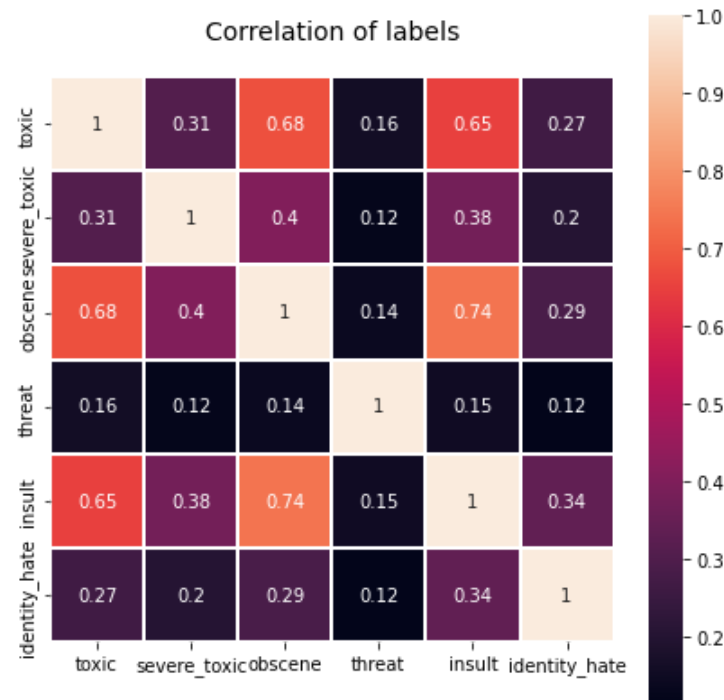
- רוב משמעותי של תגובות שאינן רעילות - הנחנו שזה מייצג מייצג פחות או יותר את המציאות בה לרוב תגובת המשתמש היא נורמטיבית (אינה רעילה).



- כמות גדולה של סמפלים המסווגים כ-toxic, insult, ו-obscene. לעומת זאת יש כמות קטנה משמעותית של סמפלים שסווגו כ-severe-toxic, identity-hate, threat - זה עלול לפגוע ביכולת הלמידה של המודלים עבור הסיווגים הללו.



- ממפת הקורלציות עלה כי קיימות רמות רעילות בעלות קורלציה גבוהה בין השניה - הסקנו שזה קשור לכך שרמות הרעילות השונות אינן קבוצות זרות, הן מוכלות אחת בשניה לדוגמא insult יכול להיות גם obscene.



## עיבוד מקדים

כחלק מניקוי הטקסט והעיבוד המקדים ביצענו: המרת כל האותיות ל-lowercase, הסרת סוגריים ותווים מיוחדים נוספים, הסרת ירידת שורה ורווחים מיותרים, הסרת stopwords, הסרת הטיות שונות, למטיציה. כמו כן, ביצענו tokenization ו-embedding.

לבסוף חילקנו את הדאטה סט ל-validation, train, ו-test על ידי אלגוריתם ייעודי להתמודדות עם חלוקת דאטה מסוג Multi-Label, את האלגוריתם מצאנו במאמר אקדמי של [Sechidis K., Tsoumakas G., Vlahavas I. \(2011\) On the 'Stratification of Multi-Label Data'.](#) (ניסיון להתמודד עם הדאטה הלא מאוזן ע"י מתודת downsampling הביא לביצועים נמוכים יותר ולכן החלטנו להשתמש בחלוקה יחסית של הדאטה-סט שלעיל).

## המודלים

להלן המודלים שבנינו ובחנו:

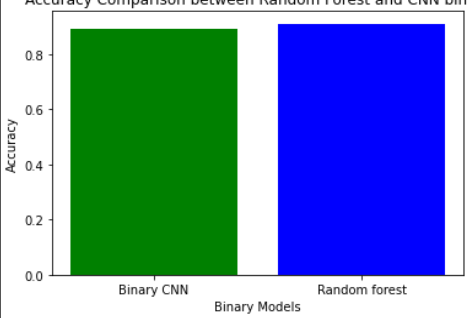
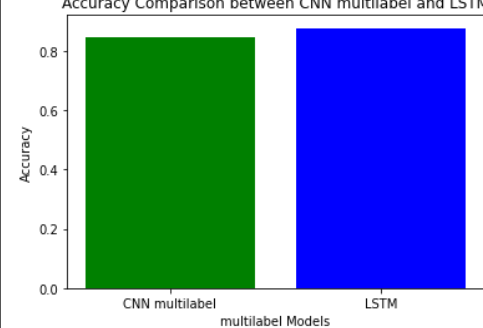
מודל	סוג הבעיה	שלב
Random Forest Classifier	Binary Classification - toxic / not toxic	1
1D CNN		
1D CNN dedicated to multi-label	Multi-Label Classification - toxicity level	2
NN LSTM model		

מאפייני המודלים:

מודל	Hyper parameters	Number of trainable parameters	Regularization methods	Number of epochs	Loss function	Optimizer
Random Forest Classifier	max_features=1\3, n_estimators=200	-	-	-	-	-
1D CNN	-	5,922,569	Dropout	10	Binary cross entropy	adam
1D CNN dedicated to multi-label	-	6,645,126	Dropout	5		
NN LSTM model	-	6,006,870	None (other than the ones implemented in the transferred model)	5		

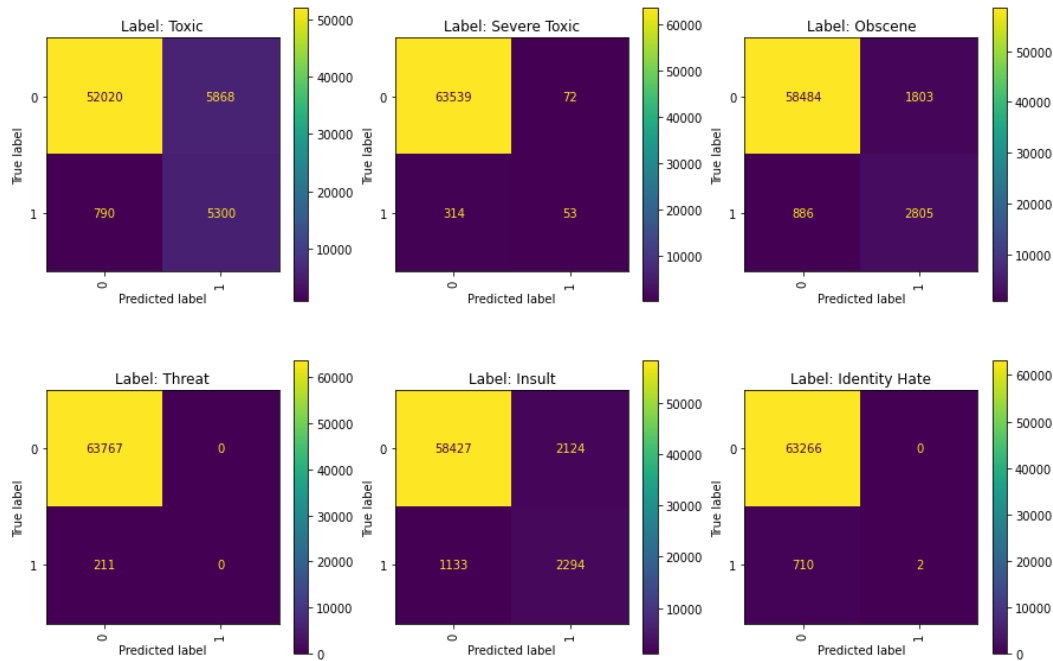
## תוצאות המודלים

להלן תוצאות המודלים של השלבים השונים:

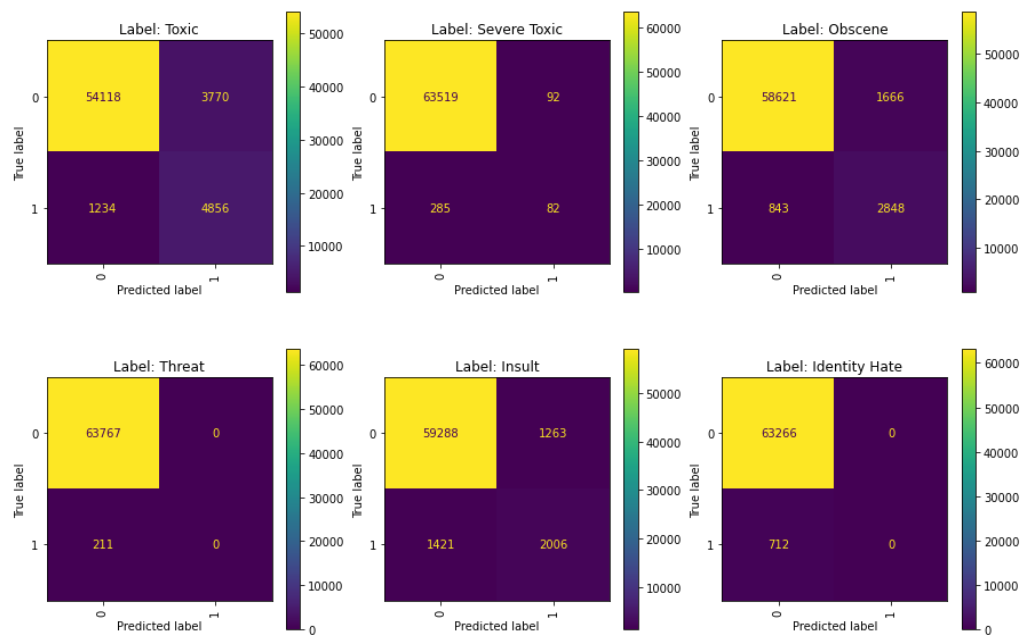
השוואה	Accuracy score	מודל	#
<p>Accuracy Comparison between Random Forest and CNN binary</p>  <p>Binary Models</p>	0.9111	Random Forest Classifier	1
	0.8914	1D CNN	2
<p>Accuracy Comparison between CNN multilabel and LSTM</p>  <p>multilabel Models</p>	0.84843	1D CNN dedicated to multi-label	3
	0.8777	NN LSTM model	4

להלן confusion matrix של המודלים מחלק 2:

### Individual Confusion Matrices for 1D CNN Multilable Model



### Individual Confusion Matrices for LSTM Model



## מסקנות עיקריות מתוצאות המודלים

- המודלים של השלב הראשון בו מסווגים אם תגובה היא רעילה או לא, בעלי ציון גבוה יחסית באזור ה-90% דיוק. עם זאת, בחרנו לבסוף במודל הקונבולוציה לאור זמן הריצה הטוב יותר לעומת ה-random forest.
- מהשוואת הציון של המודלים בשלב השני ניתן לראות ששניהם בעלי ציון דומה (סביב ה-85%) אך מודל ה-LSTM הביא לתוצאות טובות יותר ולכן בחרנו בו.
- מה-confusion matrix של המודלים בשלב השני ניתן לראות ששניהם התקשו בלמידה וסיווג לרמות הרעילות הבאות: Threat and Identity hate. ניתן לשער שזה נובע מכמה סיבות אפשריות:
  1. חוסר ייצוג של דגימות מסוג זה בדאטה הנתון: כפי שהראינו, יש כמות קטנה יחסית של תגובות מסוגי הרעילות הללו בדאטה שאיתו עובדים.
  2. מאחר ורמות הרעילות אינן מהוות קבוצות זרות, יכול להיות מצד שהתגובות מהסוגים Threat ו-Identity hate, סווגו לסוגים אחרים של רמות רעילות שמכילות גם את הרמות שהמודל התקשה לסווג.
  3. אינטונציה והקשר תרבותי: איומים תלויים לעתים קרובות באינטונציה, שקשה לזהות בטקסט, ושנאת זהות היא מאוד ספציפית לתרבות. יתכן שזה נובע גם מכך שרוב התגובות אינן מכילות אמירות שנאה או איומים באופן ישיר אלא רק במרומז.

## הצעות להמשך

- ביצוע data augmentation - ראינו לנכון להציע לנסות לבצע data augmentation כדי לנסות להתמודד עם חוסר הייצוג בדאטה של רמות הרעילות שצינינו לעיל, כדי לשפר את תוצאות המודלים בשלב השני.
- Transformers & BERT - במהלך בניית המודלים השונים לשלב 2, ניסינו להשתמש ב-Bert ולעשות לו fine tuning שמתאים לבעיה בפרויקט שלנו. אך נתקלנו בקשיים רבים עקב משאבי החישוב והזמן שהיו לנו - לכן אנו ממליצים לנסות למצוא מודל BERT מתאים ולבחון אותו כדי לנסות לשפר את התוצאות.
- ניתן לנסות לאמן את המודלים בשלב 2 עם דאטה סט המכיל רוב של תגובות רעילות כדי לבדוק האם זה משפר את התוצאות.



## ממשק משתמש

ממשק המשתמש מיועד לכך שיהיה ניתן לבחון תגובה בזמן אמת. כלומר, לאחר שמשתמש שולח תגובה, התגובה ישירות תעבור עיבוד וסיווג בשני שלבים: ראשית אם היא רעילה או לא. ואם היא רעילה אז היא תמשיך לשלב השני בו תסווג לכל רמות הרעילות הרלוונטיות אליה.

להלן הנראות של ממשק המשתמש שבנינו עבור הפרויקט:

<div>comment</div> <div>Comment to score</div> <div>Clear Submit</div>	<div>output</div> <div></div> <div>Flag</div>
--	---

להלן מספר דוגמאות הרצה של שימוש בממשק:

<div>comment</div> <div>you should go to jail for writing such a stupid article</div> <div>Clear Submit</div>	<div>output</div> <div>Your comment is toxic and thus will not be published Toxic: True Severe Toxic: False Obscene: False Threat: False Insult: False Identity Hate: False</div> <div>Flag</div>
---	---

דוגמא לתגובה לא רעילה:

<div>comment</div> <div>i disagree with this article, I think it's wrong and can mislead the public, and cause a lot of problems among us</div> <div>Clear Submit</div>	<div>output</div> <div>Your comment is non-toxic and can be published</div> <div>Flag</div>
---	---

## מסקנות ממשק משתמש

- ניתן לראות שכאשר התגובה אינה רעילה, מתקבלת הודעה מתאימה וכך זה חוסך הרצה מיותרת של המודל בשלב 2.
- בהמשך למסקנות תוצאות המודלים, גם בעת כתיבת תגובת משתמש בלייב, הממשק לא הצליח לזהות את שתי רמות הרעילות הבעייתיות ללמידה. עם זאת הממשק כן זיהה שמדובר בתגובה בעייתית. לדוגמא:  
זיהוי התגובה כרעילה אך הממשק לא מזהה את סוג הרעילות:

comment

I will find you and I will kill you

Clear

Submit

output

Your comment is toxic and thus will not be publisehd  
Toxic: False  
Severe Toxic: False  
Obscene: False  
Threat: False  
Insult: False  
Identity Hate: False

Flag

תגובה שיכולה להשתמע לשתי פנים אך מזוהה כלא רעילה:

comment

I'm coming after you

Clear

Submit

output

Your comment is non-toxic and can be published

Flag

- בחישוב עלות-תועלת הממשק מצליח לזהות את רוב התגובות הרעילות ולמרות שבמקרים מסוימים לא מצביע על סוג התגובה, עדיין הוא מהווה מסנן טוב לבעלי אתרים.