

# ניתוח פוסטים וסיווג אשמה בקהילת Reddit

מגישים :

ניר דעוס | 315782987

רון בוסקילה | ---

➤ מקור: רשת חברתית REDDIT

➤ תיאור: מאגר נתונים של 6,000 דוגמאות שמתמקד בסיווג פוסטים של משתמשים בקהילת-"האם אני האשם?"-המטרה היא לזהות האם האשם בסיפור הוא מפרסם הפוסט או הצד השני שהוזכר

➤ מאפיינים: Title-כותרת הפוסט, Post- תוכן מלא של הפוסט  
Verdict- תווית סיווג – האם המפרסם אשם (כן או לא)

## דוגמא

**Am I wrong asking customer service to address me formally?**

I am an old lady. I hate my first name. I am very uncomfortable with strangers addressing me by my first name. I have always gone by my last name. When I am forced to call a customer service representative their script tells them to call me by my first name. I gently ask them to please call me Mrs. ----- . It seems to offend them and that is the last thing I want to do. Am I wrong? I swear I do it nicely.

0 38 Share

**Are you in the wrong?**  
Describe a situation or scenario, providing all relevant information. Then seek the opinion of the masses. Were y...

Show more

Created Apr 6, 2011  
Public

432K Members 73 Online

➤ המטרה שלנו

➤ לזהות תבניות בשפה, בכותרות ובפוסטים שמשפיעות על התוצאה  
➤ לבנות מודל שיכול לחזות את תוצאות verdict עבור פוסטים חדשים

➤ פיצול הנתונים

➤ סט אימון 85% מהנתונים עבור בניית המודל  
➤ סט מבחן 15% מהנתונים עבור הערכת המודל

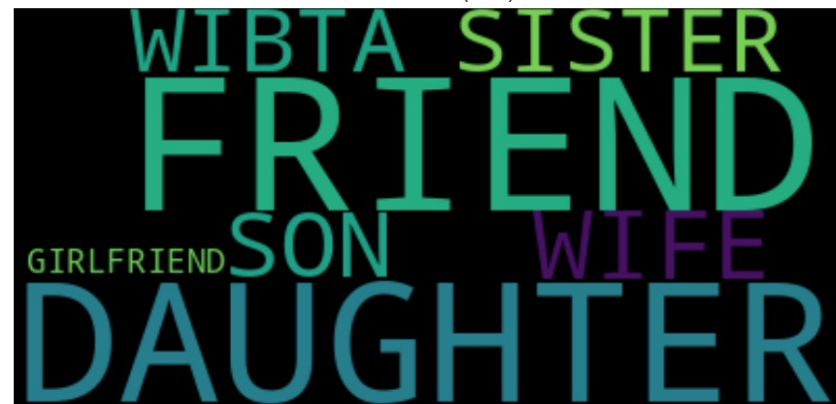
➤ הנתונים חולקו באופן אקראי כדי להבטיח ייצוג שווה

יצרנו Word Clouds למילים השכיחות ביותר בעמודות post וtitle לפי תוית verdict כדי לזהות תבניות או הבדלים בין קבוצות, 7 מילים השכיחות ביותר, מצורף גרף ויזואלי

User OK (Title)



User is Fault (Title)



User OK (Post)



User is Fault (Post)



מבנה הרשת

שכבת הקלט- מקבלת רצפי טוקנים בגודל MAX\_SEQUENCE\_LENGTH כבסיס לעיבוד בשכבות הבאות

שכבת Embedding -שממירה את רצפי הטוקנים לווקטורים מספריים בצפיפות גבוהה, כדי לאפשר למודל לזהות קשרים ותבניות סמנטיות בטקסט.

- שכבת LSTM דו-כיוונית ראשונה : 64 יחידות, מעבדת את הטקסט משני הכיוונים ומחזירה רצף תפוקות (True=return\_sequences).
- שכבת LSTM דו-כיוונית שנייה : 64 יחידות, מסכמת את המידע מהשכבה הקודמת לוקטור יחיד.

שכבת הפלט (Dense) נוירון יחיד עם פונקציית אקטיבציה Sigmoid

**קומפילציה:**

- אופטימיזר : adam.
- פונקציית הפסד : binary\_crossentropy.
- מדד ביצוע : accuracy.

**אימון הרשת:**

- גודל אצווה (Batch size) : 36.
- מספר אפוקים : 3.

Layer (type)	Output Shape	Param #
input_layer (InputLayer)	(None, 229)	0
embedding (Embedding)	(None, 229, 100)	1,000,000
bidirectional (Bidirectional)	(None, 229, 128)	84,480
bidirectional_1 (Bidirectional)	(None, 128)	98,816
dense (Dense)	(None, 1)	129

## תוצאות:

**דיוק על סט האימון (accuracy): 86.95%**

הפסד על סט האימון (loss): 0.3200

**דיוק על סט הבדיקה (val\_accuracy): 59.91%**

הפסד על סט הבדיקה (val\_loss): 0.8987

## ניתוח התוצאות:

### **פער גדול בין סט האימון לסט הבדיקה:**

הדיוק על סט האימון (86.95%) גבוה משמעותית מהדיוק על

סט הבדיקה (59.91%), דבר שמצביע על **overfitting**

המודל למד היטב את הדוגמאות באימון, אך מתקשה

להכליל נתונים חדשים בסט הבדיקה

### **ערך הפסד (loss) גבוה בסט הבדיקה:**

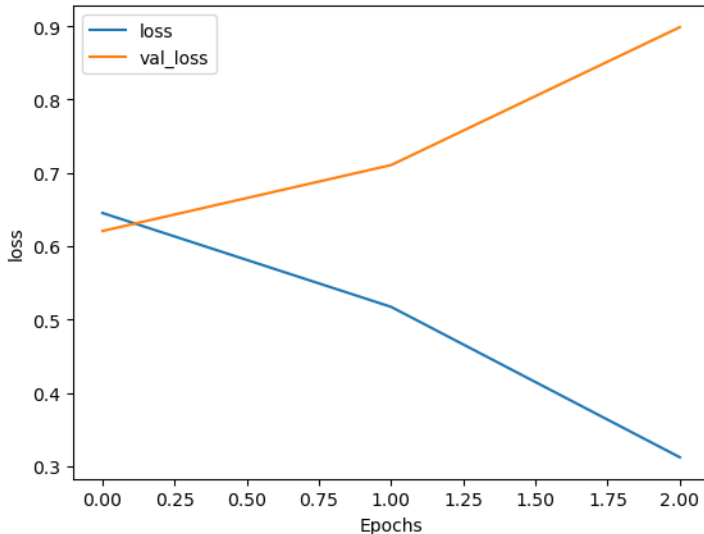
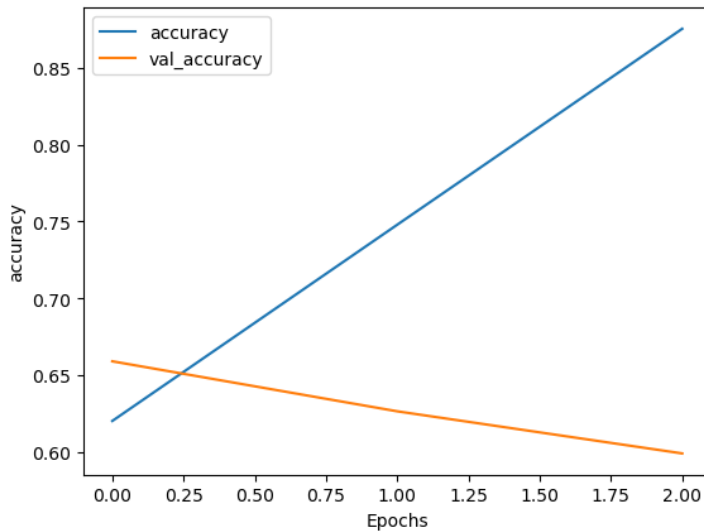
ערך הloss בסט הבדיקה (0.8987) גבוה משמעותית מזה

שבסט האימון (0.3200), מה שמאשר את הבעיה בהכללה

## **סיבה אפשרית**

1. שכבת LSTM תופסת יותר מידי מידע וזוכרת רעשים.

2. הטקסט מכיל יותר מידי רעשים.



- עיבדנו מחדש את הטקסט והורדנו מילים שלא תרמו לתוכן הפוסט.
- נוספו שתי שכבות Dropout (לפני ואחרי שכבת ה-LSTM הראשונה).  
שיעור נשירה של 50% .

שאר המבנה נשאר דומה .

Layer (type)	Output Shape	Param #
input_layer_1 (InputLayer)	(None, 229)	0
embedding (Embedding)	(None, 229, 100)	1,000,000
dropout (Dropout)	(None, 229, 100)	0
bidirectional_2 (Bidirectional)	(None, 229, 128)	84,480
dropout_1 (Dropout)	(None, 229, 128)	0
bidirectional_3 (Bidirectional)	(None, 128)	98,816
dense_1 (Dense)	(None, 1)	129

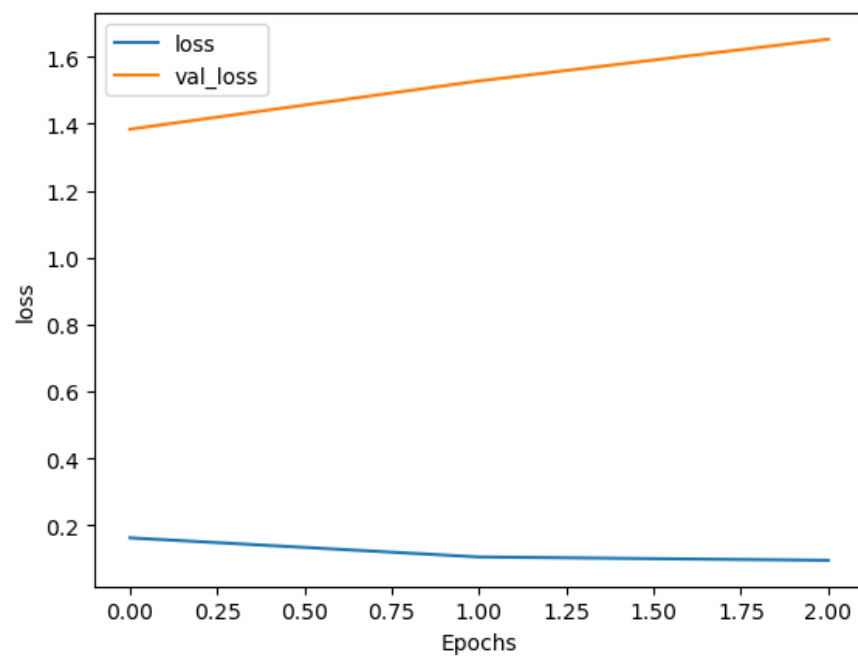
## תוצאות המודל השני:

דיוק על סט האימון (accuracy): 96.96%

הפסד על סט האימון (loss): 0.1033

דיוק על סט הבדיקה (val\_accuracy): 59.97%

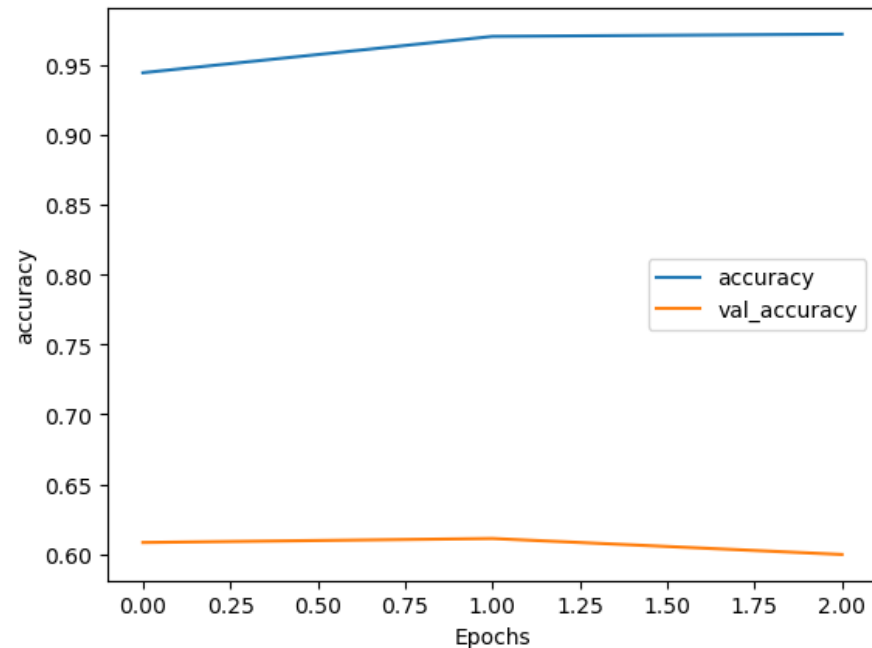
(כמעט זהה למודל הראשון : 59.91%).



## ניתוח

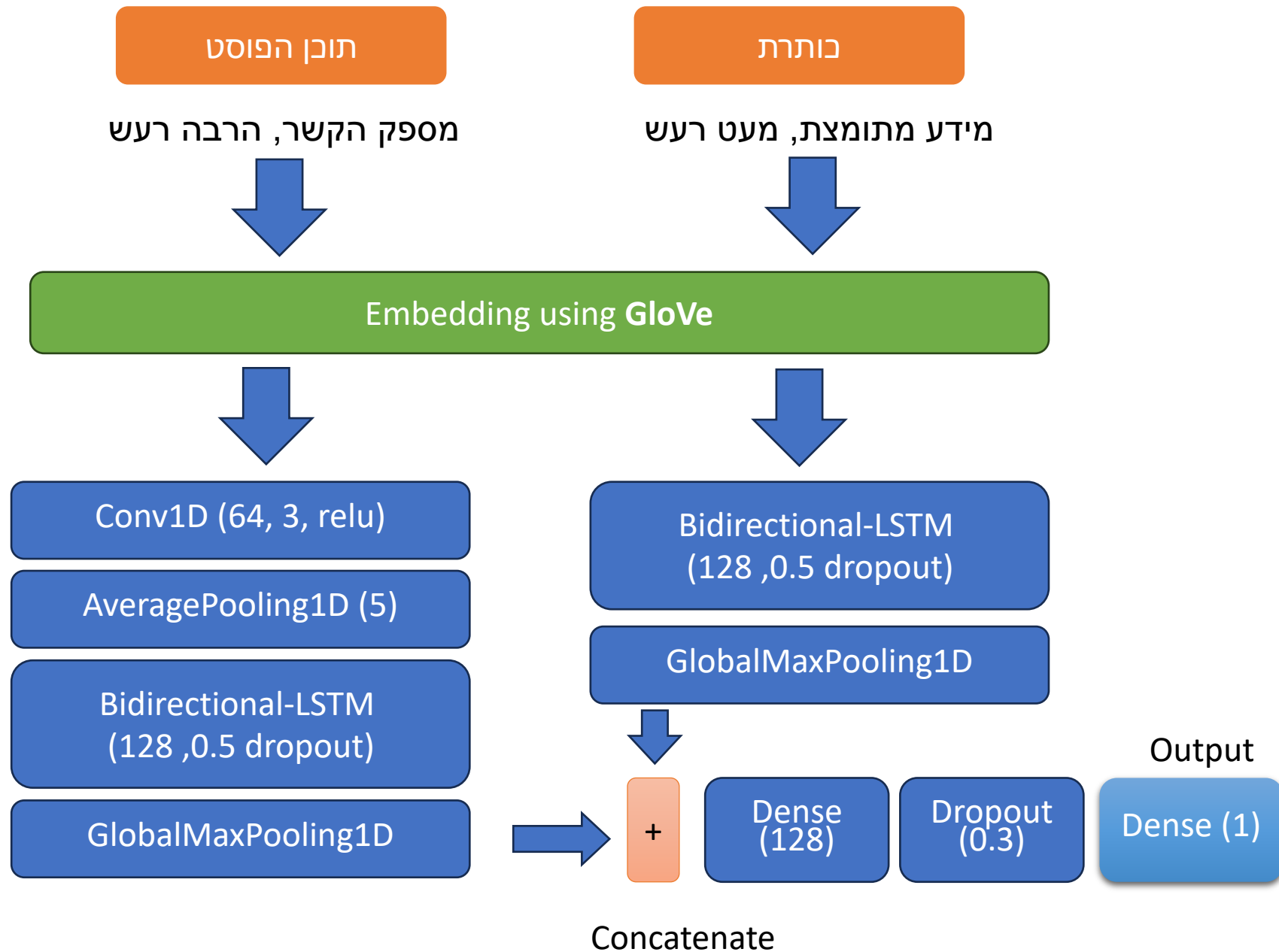
הדיוק והפסד על סט האימון השתפרו מאוד במודל השני, מה שמראה שהמודל מצליח ללמוד טוב יותר את הדוגמאות בסט האימון/

הדיוק על סט הבדיקה נשאר נמוך (59.97%), והפסד הבדיקה עלה (1.65 לעומת 0.89 במודל הראשון).



**למרות הוספת שכבות Dropout, ה-overfitting לא נפתר.**

# מודל שלישי - הפרדה בין כותרת לתוכן הפוסט





## אימון הרשת:

• גודל אצווה (Batch size): 36.

• מספר אפוקים: 7.

## תוצאות המודל השלישי:

1. דיוק על סט האימון (accuracy): 67.18%

2. הפסד על סט האימון (loss): 06186

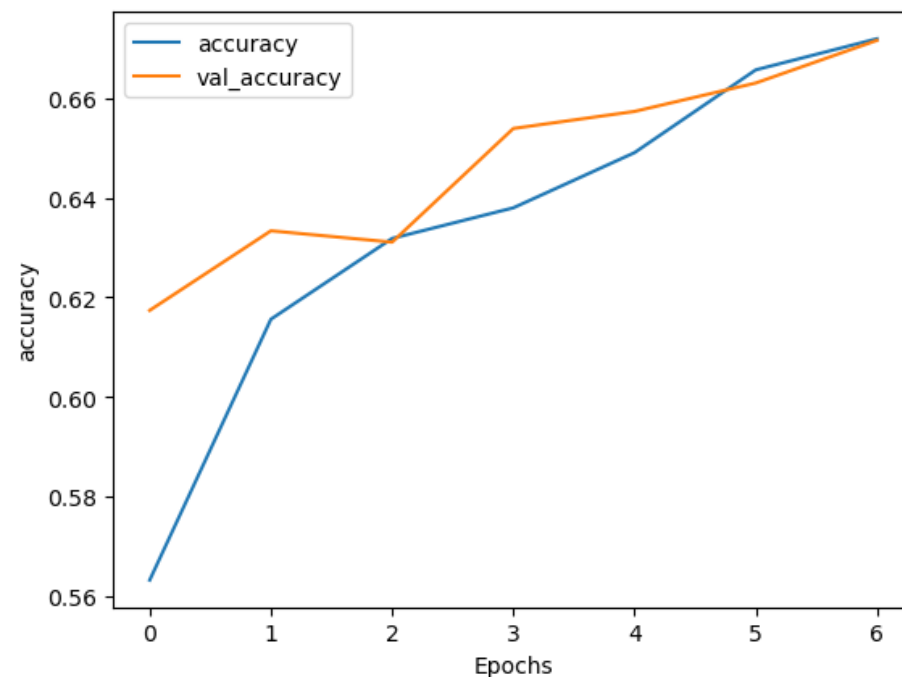
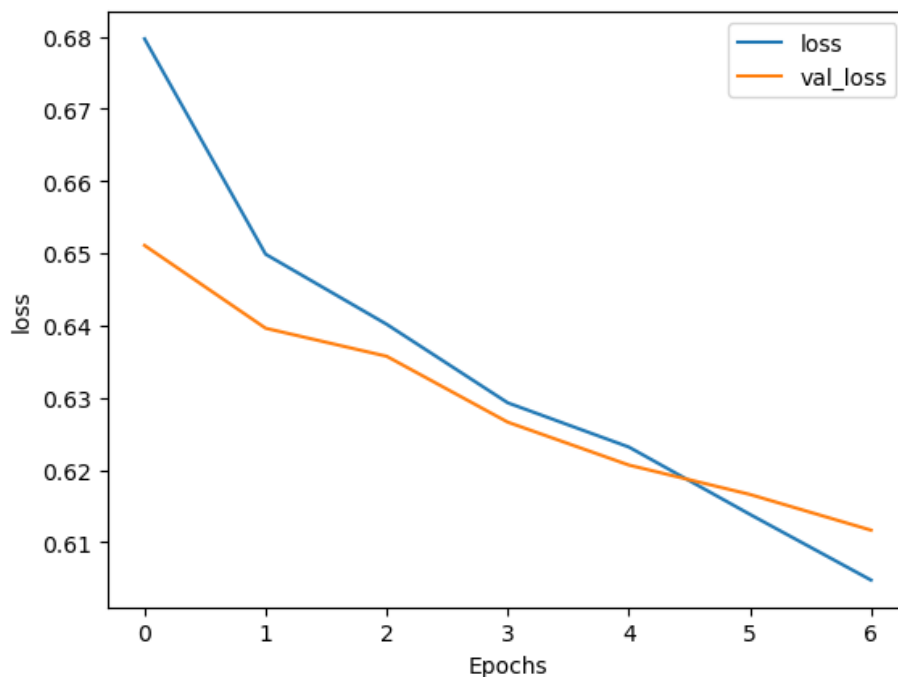
3. דיוק על סט הבדיקה (val\_accuracy): 67.16%

4. הפסד על סט הבדיקה (val\_loss): 06236

## השוואה וביצועים:

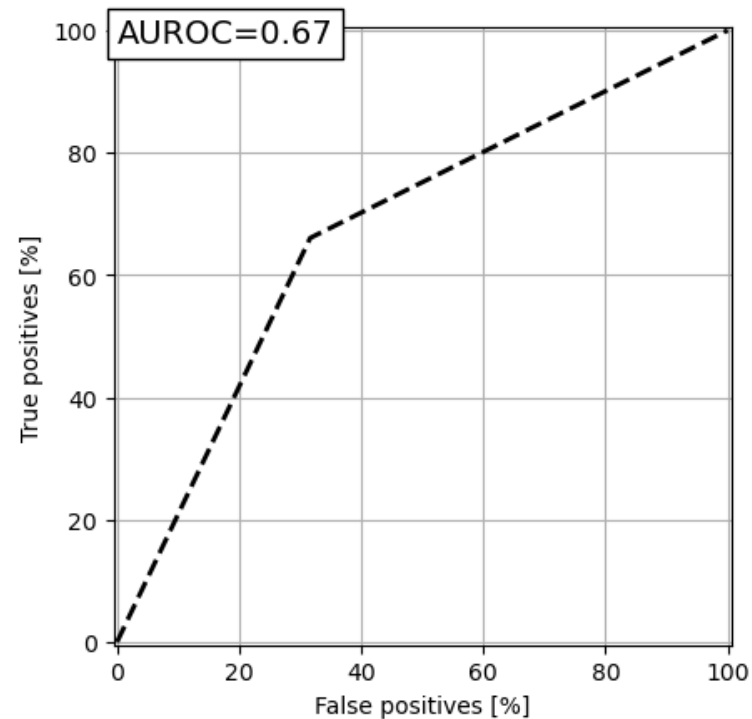
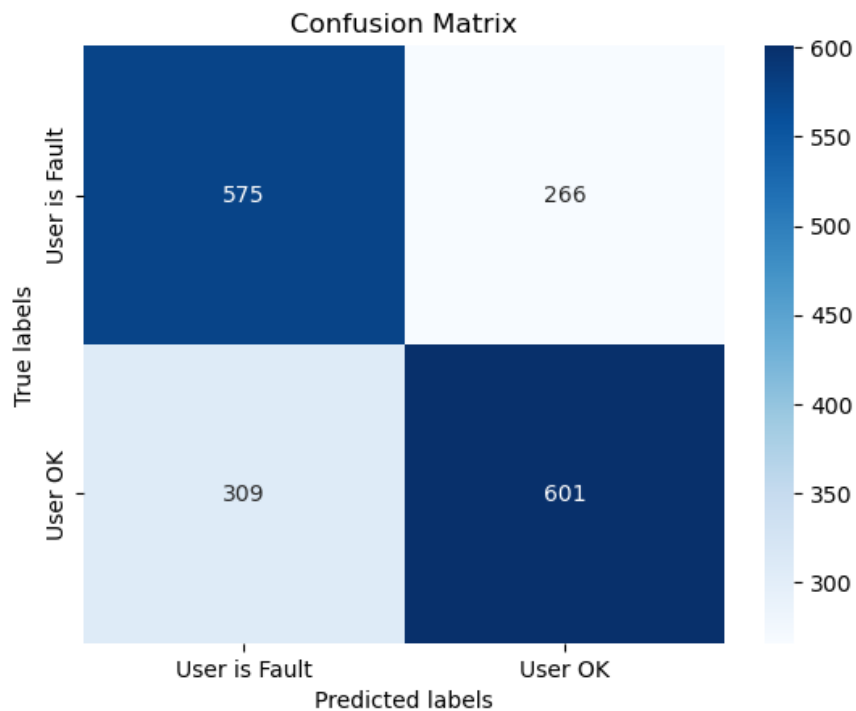
דיוק על סט הבדיקה השתפר לעומת המודלים הקודמים, הגרפים במגמה יציבה.

התוצאות מראות שהעיבוד של הכותרת והפוסט יחד מוסיף ערך.



55/55 ————— 1s 12ms/step – accuracy: 0.6802 – loss: 0.5970  
Test loss: 0.6116748452186584  
Test accuracy: 0.6716161966323853  
55/55 ————— 1s 15ms/step

	precision	recall	f1-score	support
0	0.65	0.68	0.67	841
1	0.69	0.66	0.68	910
accuracy			0.67	1751
macro avg	0.67	0.67	0.67	1751
weighted avg	0.67	0.67	0.67	1751



## סיכום :

הרשתות שנבנו למדו על ידי שילוב שכבות Embedding, LSTM דו-כיווניות, ו-CNN (במודל השלישי), עם דגש על עיבוד נתונים משני מקורות (כותרות ופוסטים). הרשת הצליחה לזהות תבניות בשפה ולספק סיווג בינארי, אך הביצועים עדיין מוגבלים, עם דיוק של כ-67.16% על סט הבדיקה.

## שיפורים אפשריים לביצועים טובים יותר :

1. ביצוע Hyperparameter Tuning, כמו שינוי שיעור Dropout, מספר הנורונים בשכבות, או גודל האצווה, יכול לשפר את הביצועים.
2. שינוי ארכיטקטורת הרשת (שינוי או הוספת שכבות)
3. Transfer Learning

## **יתרונות :**

המודל מספק בסיס לסיווג בינארי ועובד בצורה סבירה על סט האימון והבדיקה.

ניתן להשתמש בו ככלי ראשוני לסיווג פוסטים מReddit.

## **מגבלות :**

היינו רוצים שהמודל יגיע לאחוזי דיוק גבוהים יותר, אך סיווג פוסטים כאלה הוא "עמום" גם לבני אדם.

## **שימושים אפשריים:**

שימוש ככלי לחיזוי דעת קהל – האם יחשבו שאני אשם אם אפרסם את המקרה זה או לא?

## **בעיות אפשריות בשימוש כמוצר :**

1. הכללה מוגבלת :

המודל עשוי להיכשל בנתונים חדשים שאינם דומים לנתוני האימון (למשל, פוסטים בשפה שונה או בניסוח שונה).

2. איכות נתונים :

נתונים חסרים, שגיאות כתיב, או ניסוחים לא ברורים עלולים לפגוע בביצועי המודל.

3. רגישות להטיות :

אם נתוני האימון מוטים (biased), המודל עשוי להציג הטיות דומות בתוצאותיו.

## **מסקנה :**

המודל מתאים ככלי ראשוני למחקר ולניסויים, אך לפני שימוש בו כמוצר, יש לבצע שיפורים משמעותיים בביצועיו ולהתגבר על מגבלות הoverfitting והכללה. מומלץ להטמיע פתרונות כמו שימוש בPretrained Models, הגדלת כמות הנתונים, ושיפור האיזון בין התוויות.

# השוואה למול ChatGPT

<b>Data Size</b>	1000 Posts
<b>Model</b>	GPT-4o
<b>Normal Resposes for</b>	960 / 1000 Prompts
<b>Accuracy</b>	60.5%