Shomer.AI

שומר.AI

Context-Aware Bullying Detection for Hebrew Children's Chats

מערכת בינה מלאכותית לזיהוי תוכן פוגעני בשיחות ילדים — בעברית, בהקשר, בזמן אמת

Alona Musiyko
Final Project · Supervisor: Dr. Yoram Segal

Acknowledgements

תודות

ד"ר יורם סגל — המרצה והמנחה

על השיעורים והמטלות המעניינים והמאתגרים, ועל ההכוונה והליווי לאורך כל הדרך.

דימה שלייב — בעלי

על שנתן לי את הזמן וההזדמנות ללמוד, להעמיק ולהשתפר בתחום ה-AI.

The Problem

הורים אינם מודעים בזמן אמת לבריונות בשיחות ילדיהם, והכלים הקיימים שופטים כל הודעה במנותק — ומפספסים את ההקשר.

- **בריונות מקוונת בקרב ילדים היא תופעה נפוצה ומסכנת — אך מרבית ההורים אינם מודעים לה בזמן אמת**
- שיחות הילדים מתנהלות ב**עברית** — שפה שכלי הבטיחות הקיימים כמעט אינם תומכים בה
- הכלים הקיימים בוחנים כל הודעה בנפרד (**context-blind**), ומחמיצים איומים נסתרים, הסלמה ותבניות פגיעה הנבנות לאורך שיחה שלמה
- מעקב מלא אחר ההתכתבויות פוגע ב**פרטיות** הילד ובאמון בין ההורה לילד
- עומס של **false alarms** גורם להורים להפסיק להתייחס לאזהרות — ומאיין את ערך הכלי

סתם צוחקים יחד 😄

כולם נגדך, תיעלם כבר

אתה יודע מה מחכה לך מחר...

❓ פוגעני או תמים?

The Solution & Value

Shomer.AI מזהה תוכן פוגעני בעברית תוך התחשבות בהקשר השיחה, ומתריעה להורה רק כשמדובר באירוע אמיתי.

55.9 → 26.5%

התרעות-שווא: context-aware מול context-blind

0 → 100%

זיהוי איומים נסתרים: עם הקשר מול בלי

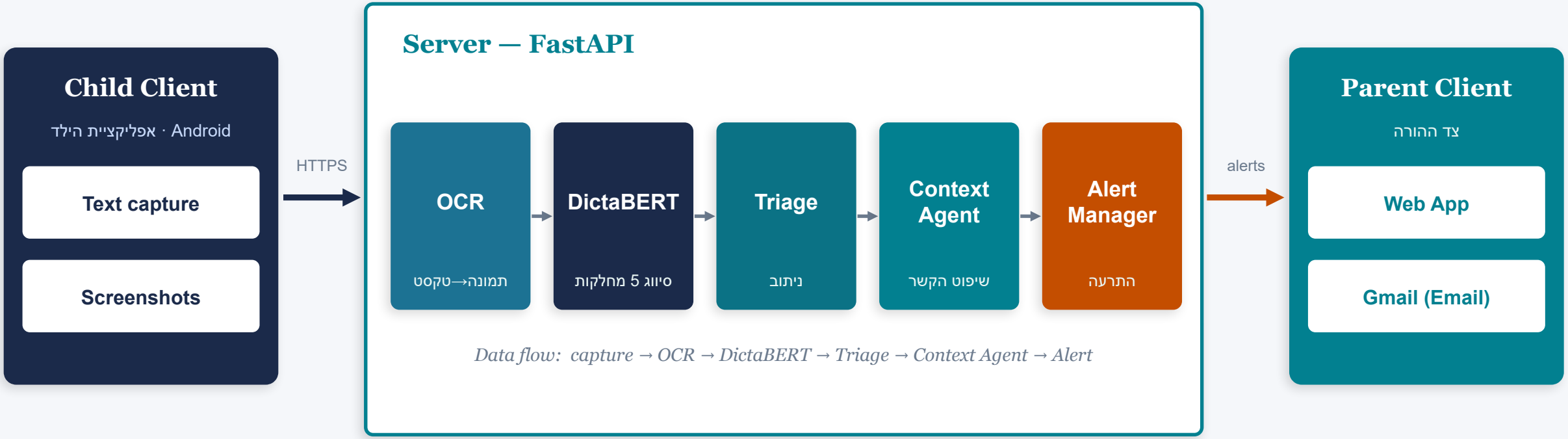
Hebrew

סיווג נֵייטיב + הקשר שיחתי

- **Real-time** — ניטור שיחות בזמן אמת וזיהוי תוכן פוגעני בעברית
- **Context-aware** — ניתוח עד 5 הודעות אחורה לזיהוי הסלמות, התקפות קבוצתיות ואיומים נסתרים
- **Privacy-by-design** — ריצה ברקע בלבד, ללא גישה לתמונות, אנשי קשר או תוכן לא-רלוונטי
- **Hebrew-native** — בנויה לשפה שהכלים הקיימים מתעלמים ממנה
- **Low false-alarms** — מתריעה רק על מצב פוגעני אמיתי, ללא הצפת התרעות

System Architecture

הילד מעלה הודעות לשרת; השרת מחלץ טקסט, מסווג, מנתב ושופט הקשר; וההורה מקבל התרעה ב-Web App וב-Gmail.



הפירוט המלא של מתי מופעל ה-Context Agent ומתי נשלחת התרעה — בשקופית הבאה (Decision Logic).

How It Works — Child & Parent Flows

שתי הזרימות מקצה לקצה: צד הילד (לכידה ועיבוד) וצד ההורה (התאמה, קבלת התרעות ותגובה).

Child Flow

זרימת הילד

1. ה-**Child App** לוכד טקסט (**Accessibility**) וצילומי מסך
2. **Pre-Filter** מקומי שומר רק עברית (פרטיות + דדופ)
3. העלאה מוצפנת לשרת ב-**HTTPS**
4. **OCR** (**Tesseract**) מחלץ טקסט מתמונות
5. **DictaBERT** מסווג ל-5 מחלקות (**context-blind**)
6. **Triage** מנתב; גבול/אלימות → **Context Agent**
7. **Context Agent** (**LLM**) קורא 5 תורים ושופט הקשר
8. **Alert Manager** מסמן מצב פוגעני

Parent Flow

זרימת ההורה

1. ההורה נרשם ב-**Web App** (**Dashboard**)
2. קוד התאמה (**OTP**) נשלח ב-**Gmail**
3. ההורה מזין את הקוד באפליקציית הילד → **paired**
4. **Alert Manager** שולח התרעות ב-**Gmail** + ל-**Web App**
5. ההורה סוקר ב-**Web App** ומגיב (**ack / label / severity**)
6. התיוגים חוזרים כנתוני אימון (**human-in-the-loop**)

Decision Logic — 5 Classes → Triage → Context Agent → Alert

ה-DictaBERT מפיק אחת מ-5 מחלקות עם רמת ודאות; ה-Triage מחליט אם להשתיק, להתריע ישירות, או להעביר ל-Context Agent.

Classifier output	Triage decision	Context Agent?	Outcome
prob_offensive ≤ 0.3	Silent	No	No alert
abusive / hate · prob ≥ 0.7	Direct alert	No	**Alert sent**
pornographic (any prob)	Direct alert	No	**Alert sent — always**
violence (any prob)	Escalate	**Yes — always**	Alert if CA confirms threat
borderline · prob 0.3–0.7	Escalate	**Yes**	CA decides → alert / silent

- **5 classes** — ה-DictaBERT מפיק אחת מ-5 מחלקות עם **prob_offensive** (רמת ודאות פוגענית)
- **Triage** — ודאות נמוכה → שקט · ודאות גבוהה → התרעה ישירה · ביניים → **Context Agent**
- **Context Agent triggers** — במקרי גבול (**0.7–0.3**) ותמיד ב-**violence**
- **Alert sent — בהתרעה ישירה (abusive/hate ודאי או pornographic), או כש-Context Agent מאשר פגיעה בהקשר**

Research Question → Maps to the Architecture

אותה מערכת בדיוק נבחנת בשני תנאים — עם הקשר ובלעדיו — כדי לבודד את תרומת ההקשר השיחתי.

האם הוספת **conversational context** מפחיתה את שיעור ה-**false-positives** בסיווג בריונות בעברית — מבלי לפגוע ב-**recall**

(אותה מערכת בדיוק, מתג אחד מתהפך)

context-aware (treatment)

DictaBERT → Context Agent

קורא את 5 התורים הקודמים ושופט אם ההודעה פוגענית לאור השיחה.

context-blind (baseline)

DictaBERT

קורא הודעה בודדת בלבד, ללא זיכרון של השיחה.

החיבור לארכיטקטורה: **Context Agent** הוא בדיוק זרוע ה-treatment. מתג **CONTEXT_AGENT_ENABLED** = false/true מחליף בין שתי הזרועות.

Why DictaBERT? (Model Choice)

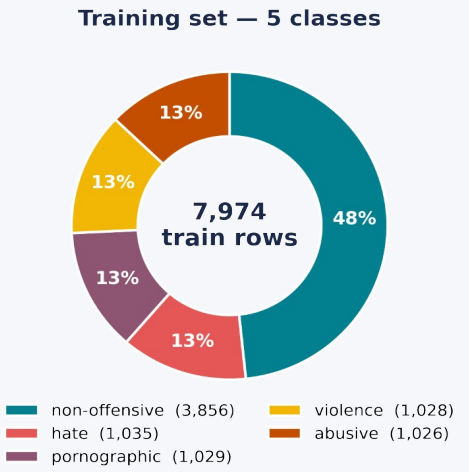
בחרנו במודל עברי ייעודי, מהיר ומקומי — ולא ב-LLM כללי או במודל מדף — מהסיבות הבאות.

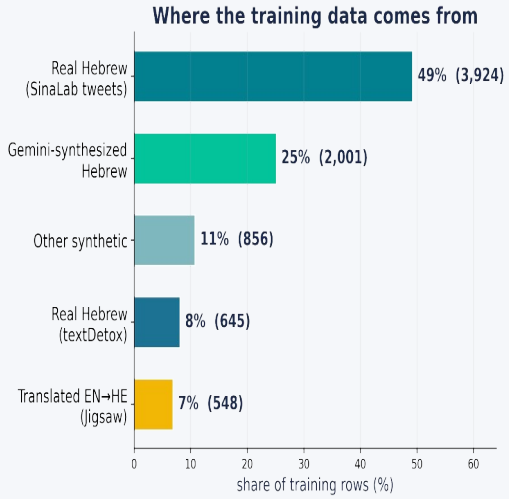
Criterion	DictaBERT (chosen)	General LLM	Ollama (off-the-shelf)
Hebrew-native	✓ Yes	Partial	✗ No
Accuracy (macro-F1)	0.836	—	0.373
Speed / latency	Very fast	Slow	Slow
Cost	Free · local	$$ per call	Free · local
Privacy (on-prem)	✓	✗ cloud	✓
Fits 'always-on' use	✓	✗ too costly	✗ low quality

- **Hebrew-native** — BERT עברי שעבר אימון-קדם על עברית; מבין מורפולוגיה שמודלים אנגליים מפספסים
- **0.836** macro-F1 — **Accuracy** מול **0.373** של **Ollama** על אותו מבחן
- **Speed** — פי **424** מהיר; מתאים לסריקה שוטפת בזמן אמת
- **Privacy** — ריצה מקומית, ללא ענן; נתוני הילדים אינם יוצאים החוצה
- **Low cost** — אינפרנס מקומי חינמי, מול תשלום לכל קריאת **LLM**
- **DictaBERT — Right tool per job = גלאי מהיר; LLM Context Agent = רק למקרי גבול**

Training Data

המודל אומן על תערובת של עברית אמיתית ונתונים סינתטיים, והוערך על נתונים אמיתיים בלבד.

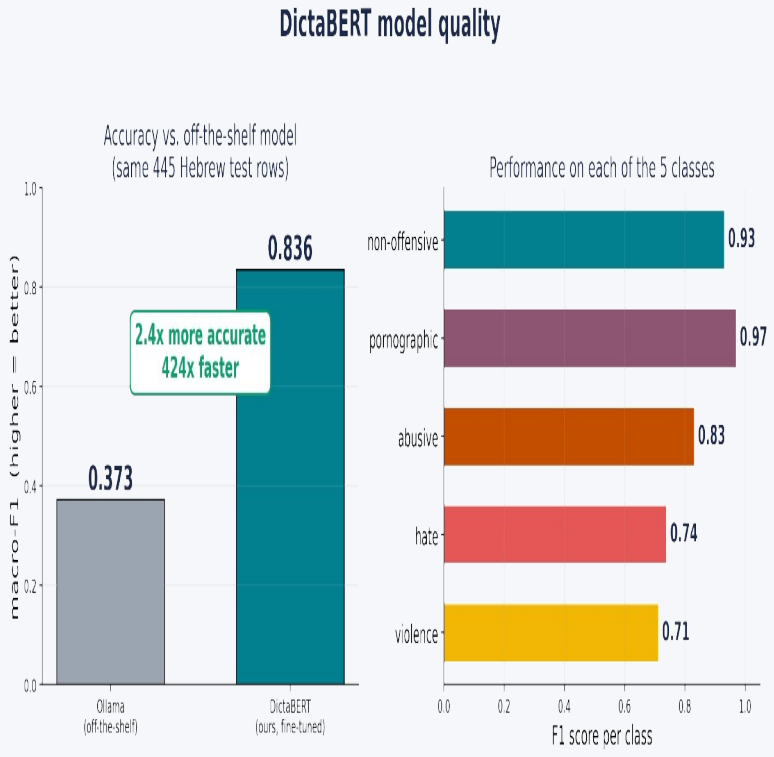




- **היקף: 445 test · 888 validation · 7,974 train**
- 5 מחלקות: **non-offensive · abusive · hate · violence · pornographic**
- מקורות: **~57%** עברית אמיתית (**SinaLab + textDetox**), **~36%** סינתטי (**Gemini**), **~7%** תרגום (**Jigsaw**)
- התפלגות מודעת-שכיחות: **~48%** non-offensive ב-**~70%** train, בהערכה — כמו בעולם האמיתי
- עיקרון מנחה: **train-on-synthetic / eval-on-real**
- הערה כנה: חלק ממחלקות המיעוט בהערכה סינתטיות (100% — **pornographic**) → מספרי המיעוט אופטימיים

Results — Model Quality (DictaBERT)

המסווג שאימנו עובר את כל ספי הסף, מכויל היטב, ומנצח מודל מדף בפער גדול.



0.836
macro-F1

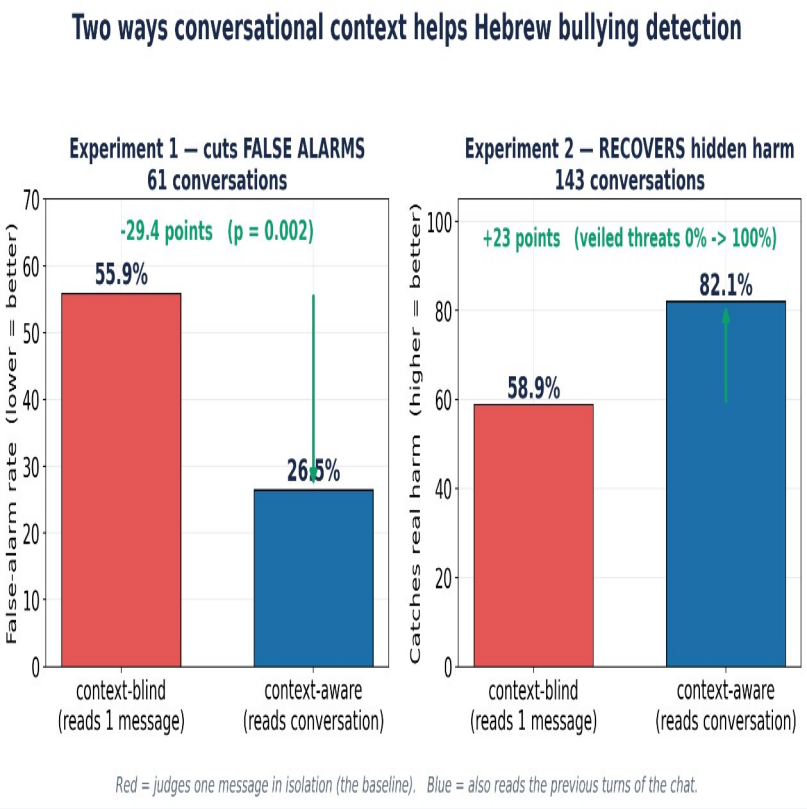
0.034
ECE
(calibration)

- **DictaBERT עברי שאומן על-ידינו ל-5 מחלקות**
- פי **2.4** מדויק ופי **424** מהיר מ-:**Ollama** **0.836** מול **0.373**
- **Calibration** טוב (**ECE 0.034**): הביטחון של המודל כן — לא מגזים ולא ממעיט
- זוהי זרוע ה-**context-blind** של הניסוי — וגם המסווג בייצור

macro-F1 = average accuracy across all 5 classes (higher = better). **ECE** = how honest the model's confidence is (lower = better).

Results — Conversational Context

בשני ניסויים, הוספת הקשר שיחתי מפחיתה התרעות-שווא ותופסת פגיעה נסתרת — לעומת סיווג ההודעה במנותק.



- **Experiment 1 (61 שיחות): false-alarm rate 55.9% → 26.5%**
- ירידה של **29.4** נקודות · **p = 0.002** (**McNemar**) — מובהק
- **Experiment 2 (143 שיחות): harm recall 58.9% → 82.1%**
- **veiled threats 0% → 100%** עם הקשר
- ההתרעה על המצב הפוגעני — לא על כל מילה גסה
- סטטוס: **MVP / feasibility** — נתונים סינתטיים/מתויגים-עצמית

FPR = false-alarm rate (lower better). **recall** = % of real harm caught. **p-value** < 0.05 = the result is not luck.

Conclusions

מסקנות

- **ניתוח conversational context משפר מהותית את דיוק הסיווג בעברית**
- ההקשר עוזר בשני הצירים: מפחית **false alarms** וגם תופס פגיעה נסתרת
- הגדרת היעד הנכונה היא המצב הפוגעני — לא המילה הבודדת
- מערכת מודולרית עובדת מקצה-לקצה ומשמשת גם כמעבדת ניסוי הניתנת לשחזור
- **סטטוס כן: MVP — הכיוון והמובהקות חזקים; הגודל המדויק דורש gold-set אמיתי**

Future Work

הכיוונים שיהפכו את הוכחת ההיתכנות למחקר בר-פרסום ולמוצר בר-פריסה.

1 gold-set אמיתי

- **בניית gold-set אמיתי, מאונוטט ידנית (Cohen's κ) → מספר תוצאה אמין לפרסום**

2 פיילוט בבתי-ספר

- פריסת **pilot** מחקרי בשיתוף מוסדות חינוך ורשויות מקומיות

3 פלטפורמות נוספות

- תמיכה בפלטפורמות נוספות (**Telegram, TikTok**) ו-**on-device** לפרטיות מוגברת

4 multilingual

- סיווג **multilingual** — אנגלית ורוסית בהקשר הישראלי

5 תרומה לספרות

- **E6**: השוואת **selective context-agent** מול **naive concatenation** — תרומה לספרות (**Pavlopoulos 2020**)