Shomer.AI — מקורות נתונים לאימון מודל הבריונות

סטטוס: סיכום · 2026-05-27 **מטרה:** למפות אילו דאטהסטים (עברית + אנגלית) אפשר לאמן עליהם, ואיך לייצר נתונים סינתטיים — כדי לגשר על הפער של *אין דאטהסט עברי שיחתי לבריונות*. **קשור ל:** `research_question.md` · `preparatory_report.md` (רכיב 4ב).

> **האתגר בשורה אחת:** המשימה דורשת **שיחות עברית מתויגות** (כי השאלה היא על הקשר שיחתי). דאטה כזה **לא קיים בעברית**. לכן משלבים: דאטה עברי מבודד (בסיס) + דאטה אנגלי שיחתי (השראה/תרגום) + **נתונים סינתטיים** (עיקרי) + gold set אמיתי קטן (הערכה).

---

1. עברית — מה קיים בפועל

דאטהסט	היקף / סוג	תוויות	קישור	הערה / שימוש
SinaLab Offensive-Hebrew (Hamad et al., 2023)	15,881 ציוצים (≈1,200 פוגעניים) · **הודעות בודדות**	`abusive` / `hate` / `violence` / `pornographic` / `none-offensive`	HF · GitHub	**העוגן העברי.** מספק baseline + סכמה טקסטואלי. חיסרון: לא שיחתי
textdetox / multilingual_toxicity	רב-לשוני, כולל **תת-קבוצה עברית**	toxic / neutral	HF	תוספת רעילות בעברית; קטן, לא ייעודי לבריונות
מודלים עבריים (לא דאטה)	DictaLM 2.0 / DictaBERT / AlephBERT	—	—	בסיס ל-fine-tune, לא מקור אימון

מסקנה כנה: בעברית יש בעצם **מקור ייעודי אחד** (SinaLab), והוא מבודד ולא שיחתי. זה הפער — וההצדקה לסינתזה.

---

2. אנגלית — לתרגום / cross-lingual / השראה

> שימוש: (א) **תרגום+עריכה** ל-augmentation; (ב) **cross-lingual transfer** עם מודל רב-לשוני; (ג) **השראה** למבנה ולטקסונומיה. ⚠️ תרגום מכונה הורס סלנג ו-code-switching — רק עם post-edit אנושי, ולעולם לא כ-test set.

2א. הקרובים ביותר למשימה — *שיחתי / מבוסס-הקשר*

דאטהסט	למה רלוונטי	קישור
ConvAbuse (Cercas Curry et al., 2021)	התעללות **בתוך שיחות** — המבנה השיחתי שאנחנו צריכים	arXiv: 2109.09483
Wikipedia Conversations / Context-Toxicity (Pavlopoulos et al., 2020)	רעילות **עם תור-הורה והקשר** — בדיוק שאלת ההקשר	ACL 2020
Instagram media sessions (Hosseinmardi et al., 2015)	**session-based** — תמונה + שרשור תגובות (cyberaggression)	arXiv: 1508.06257

2ב. בריונות (לא בהכרח שיחתי)

דאטהסט	היקף	קישור
Formspring (Reynolds et al., 2011)	13,158 הודעות, 892 בריונות	hatespeechdata catalogue
Twitter Bullying Traces (Xu et al., 2012)	ציוצים עם "עקבות" בריונות	hatespeechdata catalogue

2ג. שיח פוגעני / שנאה — benchmarks גדולים

דאטהסט	אופי	קישור
OLID / OffensEval (Zampieri et al., 2019)	טקסונומיה היררכית A/B/C — תקן דה-פקטו	hatespeechdata
Davidson et al. (2017)	25K ציוצים: hate / offensive / neither	HF
HatEval (Basile et al., 2019)	שנאה נגד מהגרים/נשים (SemEval-2019)	hatespeechdata
Founta et al. (2018)	~80K ציוצים abusive/hateful	hatespeechdata
Jigsaw / Wikipedia Toxic (Wulczyn et al., 2017)	תגובות מתויגות רעילות (Kaggle)	hatespeechdata
HateXplain (Mathew et al., 2021)	שנאה **עם נימוקים (rationales)**	HF

📚 **קטלוג-על:** hatespeechdata.com — רשימה מתוחזקת של עשרות דאטהסטים לפי שפה ומשימה. נקודת הפתיחה לחיפוש.

---

3. נתונים סינתטיים — למה, ואיך לייצר

למה זה הציר העיקרי שלנו: (1) אין דאטה עברי שיחתי; (2) **בטוח אתית** — לא אוספים בריונות אמיתית של קטינים; (3) שליטה מלאה — אפשר לייצר בכוונה את **מקרי ההקשר** שמניעים את שאלת המחקר; (4) תקדים מוכח: **SynBullying** (Kazemi et al., 2025, arXiv: 2511.11599) ו-**ToxiGen** (Hartvigsen et al., 2022, ACL) — הראו ש-fine-tune על סינתטי משפר ביצועים על טקסט אנושי.

צעדי הייצור (Pipeline)

1. **קבעי סכמה ותוויות** — יישרי לסכמת (`abusive/hate/violence/pornographic/none`) SinaLab + שדה `context_dependent` (האם התווית תלויה בהקשר).
2. **הגדירי טקסונומיית "מקרי-הקשר"** (החלק הקריטי — זה מה שבודק את ה-RQ): - 🔴 **נראה תמים אך בריונות** — מובן רק לאור התורים הקודמים → בודק **recall**. - 🟢 **נראה פוגעני אך תמים** — הקנטה ידידותית / שפה תוך-קבוצתית / ציטוט / משחק → בודק **false positives**. - 📈 **הסלמה הדרגתית** — בריונות שנבנית על פני כמה תורים. - 🎯 **דינמיקת מטרה/כוח** — קבוצה נגד יחיד.
3. **ייצור עם LLM** — בקשי ממודל חזק (GPT-4o / Claude / Llama דרך Ollama) לייצר שיחה רב-תורית עם תווית + נימוק. השתמשי ב-**few-shot** (2–3 דוגמאות איכותיות) ובהקצאת **פרסונות** לדמויות.
4. **גיוון (diversity)** — שני באופן שיטתי: גיל, פלטפורמה (וואטסאפ/דיסקורד/אינסטגרם), נושא, חומרה, וסלנג. הזרימי **לקסיקון סלנג עברי עכשווי** כדי שהשפה לא תהיה "ספרותית".
5. **Adversarial / classifier-in-the-loop** (כמו ToxiGen) — שמרי בעיקר דוגמאות שהמסווג הנוכחי **טועה** בהן → דאטה קשה ויעיל.
6. **אימות איכות** — אדם בודק מדגם: התווית נכונה? נשמע עברית אמיתית? מדדי inter-annotator agreement; סננו דוגמאות גרועות.
7. **איזון** — oversample למחלקות נדירות (`pornographic` , `violence`).

תבנית פרומפט לדוגמה (להעתקה והתאמה)

```
אתה מייצר דאטה סינתטי לאימון מסווג בריונות בעברית.
צור שיחה בת 4–6 תורים בפלטפורמת {וואטסאפ/דיסקורד}, בין נוער גיל {12–16}
בנושא {בית-ספר/גיימינג/חברה}, בסגנון ישראלי אותנטי עם סלנג עכשווי
ו-code-switching טבעי עברית-אנגלית.
סוג המקרה: {נראה פוגעני אך הקנטה ידידותית}.
אל תכתוב תוויות בתוך הטקסט.
החזר JSON בלבד:
{
  "conversation": [{"speaker": "...", "text": "..."}],
  "target_index": <מס' ההודעה שמסווגים>,
  "label": "abusive|hate|violence|pornographic|none-offensive",
  "context_dependent": true|false,
  "rationale": "הסבר קצר למה זו התווית, בהתחשב בהקשר"
}
```

מלכודות שחובה להימנע מהן

- ⚠️ **מחולל = מסווג:** אל תייצרי במודל ששמש גם כבסיס המסווג — "אופים" פנימה אותן נקודות-עיוורון. ייצור במודל אחד, אימון/הערכה על אחר.
- ⚠️ **פער התפלגות:** סינתטי נוטה להיות "נקי" מדי. הזרמת סלנג אמיתי ו-noise מכוונת.
- ⚠️ **אמון בתוויות:** התוויות הן של המחולל — חייבות אימות אנושי על מדגם.
- ⚠️ **train-on-synthetic / evaluate-on-real:** הסינתטי לאימון בלבד; ה-**הערכה תמיד על gold set אמיתי**, אחרת התוצאה היא "המודל מחקה את המחולל".

4. השילוב המומלץ (סיכום)

שכבה	מקור	תפקיד
בסיס	SinaLab (אמיתי, מבודד)	אימון מסווג החזית + סכמה
שכבת הקשר	סינתטי-שיחתי (עיקרי) + תרגום-עריכה מאנגלית (תוספת)	לימוד התנהגות context-aware
הערכה	**gold set אמיתי בעברית** (~100–200, ידני)	מדידה בלבד — train-on-synthetic / evaluate-on-real
ידע חי	לקסיקון סלנג עברי	RAG לסוכן-ההקשר (לא אימון)

מקורות

- SinaLab Offensive-Hebrew — HuggingFace · GitHub · Hamad et al. (2023), arXiv: 2309.02724
- textdetox multilingual toxicity — HuggingFace
- ConvAbuse — Cercas Curry et al. (2021), arXiv: 2109.09483
- Toxicity Detection: Does Context Really Matter? — Pavlopoulos et al. (2020), ACL
- Instagram cyberbullying (sessions) — Hosseinmardi et al. (2015), arXiv: 1508.06257
- Davidson et al. (2017) — HuggingFace
- HateXplain — Mathew et al. (2021), HuggingFace
- Hate Speech Dataset Catalogue (meta-resource) — hatespeechdata.com
- SynBullying — Kazemi et al. (2025), arXiv: 2511.11599
- ToxiGen — Hartvigsen et al. (2022), ACL