



اكتشاف خطابة الكراهية والتنمر باللغة العبرية

Offensive Hebrew Corpus and Detection using BERT

Nagham Hamad

Mustafa Jarrar

Mohammad Khalilia

Nadim Nashif

Birzeit University
Palestine

7amla Center
Palestine



We build
tools and
resources
for NLU

Lexical Resources at SinaLab - Birzeit University

Lexicographic Database



150 lexicons
Largest Arabic lexicographic database

Arabic Ontology/ Wordnet



Formal Arabic Wordnet
with ontologically clean content

Annotated Corpora



Dialects,
NER, WSD, synonyms
Intents, hate
....

NLP library



APIs
Linguistic Data,
synonyms, Nested
NER, intents, ...

Synonyms 90s%

WSD 84%

NER 90s%

Intent 88.4%

Offensive 88.4%

Big Linguistic Data Graph

<https://ontology.birzeit.edu>



Natural Language Understanding Tools and Datasets



Open Source

<https://sina.birzeit.edu/resources>



SinaLab

News Team Resources

Resources

Download and try NLP/NLU datasets, corpora, tools and services

+ Lexicographic Database (150 lexicons)

حوسبة المعاجم

+ Arabic Ontology

الأنطولوجيا العربية

+ Dialect Corpora (Currasat)

كراسات مدونة العاميات

+ Arabic Synonyms

استخراج مترادفات

+ Named Entity Recognition (Wojood)

وجود - لاستخراج أسماء الاعلام

+ Word Sense Disambiguation (Salma)

سلمى - محلل دلالي

+ ArBanking77 Intent Detection

تحديد المقصود في المساعدات الآلية

+ Offensive Language Detection

خطاب الكراهية بالعبرية

+ Lemmatizer

+ NLP Tools



Copyright © 2023 Birzeit University

Offensive Hebrew Corpus and Detection using BERT

Nagham Hamad
Birzeit University
Palestine
nhamad@birzeit.edu

Mustafa Jarrar
Birzeit University
Palestine
mjarrar@birzeit.edu

Mohammad Khalilia
Birzeit University
Palestine
mkhalilia@birzeit.edu

Nadim Nashif
7amleh Center
Palestine
nadim@7amleh.org

Abstract—Offensive language detection has been well studied in many languages, but it is lagging behind in low-resource languages, such as Hebrew. In this paper, we present a new offensive language corpus in Hebrew. A total of 15,881 tweets were retrieved from Twitter and each was labeled with one or more of five classes (abusive, hate, violence, pornographic, or none offensive) by Arabic-Hebrew bilingual speakers. The annotation process was challenging as each annotator is expected to be familiar with the Israeli culture, politics, and practices to understand the context of each tweet. We fine-tuned two Hebrew BERT models, HeBERT and AlephBERT, using our proposed dataset and another published dataset (D_{OLaH}). We observed that our data boosts HeBERT performance by 2% when combined with D_{OLaH} . Fine-tuning AlephBERT on our data and testing on D_{OLaH} yields 69% accuracy, while fine-tuning on D_{OLaH} and testing on our data yields 57% accuracy, which may be an indication to the generalizability our data offers. Our dataset and fine-tuned models are available on GitHub and Huggingface¹.

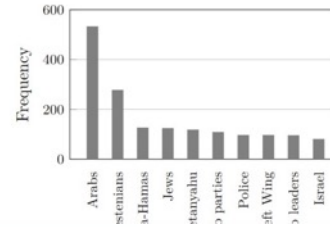
Index Terms—Offensive, Deep Learning, Hate speech, Hebrew, Pre-trained model,

I. INTRODUCTION

The amount of content published on social media is massive and cannot be moderated manually [1]. This has led to widespread of offensive language, adding pressure on social media platforms to moderate and monitor the content posted by the users [2] [3]. Governments, human

highly relies on one's knowledge and familiarity with the linguistic and cultural aspects of that language [12] [13]. The problem is even more challenging when dealing with the colloquial text given the wide variety of Arabic dialects [14] [15] [16].

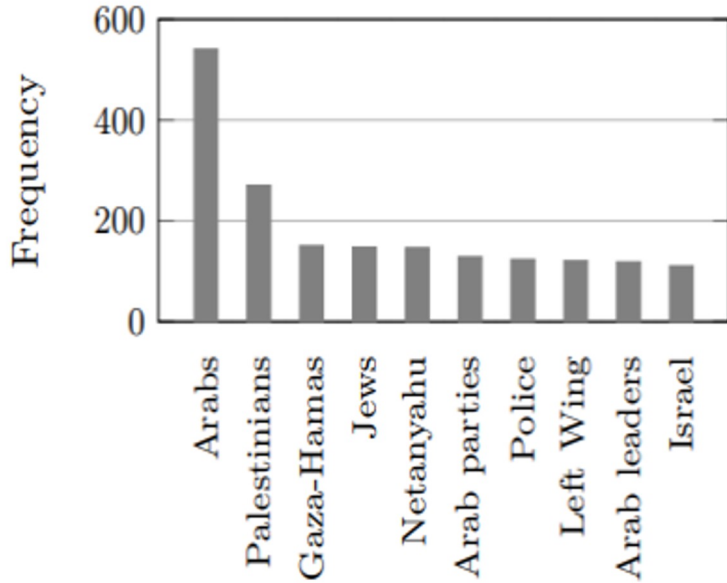
The offensive content in Hebrew is wide-spreading on social media, especially against Arabs and Palestinians (see Figure 1). Limited attention is given to this content as Hebrew lacks resources for offensive language detection. To the best of our knowledge, there are only two small datasets in Hebrew for offensive language detection. The first dataset [17] consists of 1,489 posts and comments collected from Facebook, but it is not publicly available. The second dataset, D_{OLaH} , consists of 2,000 Facebook posts [18]. A combination of both datasets, with a small extension, was released recently [19].



Nagham Hamad, Mustafa Jarrar, Mohammad Khalilia, and Nadim Nashif. 2023. Offensive Hebrew Corpus and Detection using BERT. In 20th ACS/IEEE International Conference on Computer Systems and Applications. IEEE Explore.

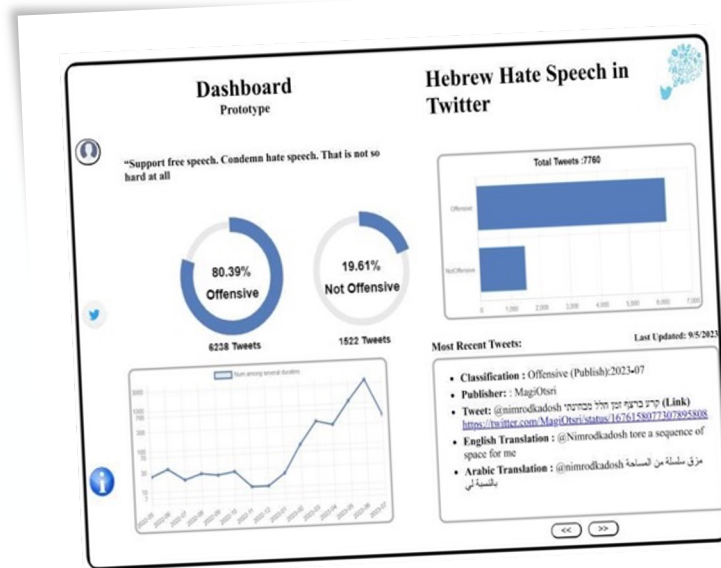
<http://www.jarrar.info/publications/HJKN23.pdf>

Problem statement



The top 10 targets in the dataset.

- Offensive content in Hebrew is widely spreading
- Hebrew offensive datasets are very rare.



Our Contribution

We built a new **Hebrew dataset** that is:

Contains 15,881 Tweets

- We **fine-tuned models** that is
 - HeBERT
 - AlephBERT models with different dataset combinations achieving state-of-the-art accuracy.

Our Dataset

- Data collection
 - The data is collected between December 2020 and January 2021 using Twitter API.

Term in Hebrew	English Translation
איסלאמיסט	Islamist
מחבל	terrorist
מתנחל	settler
נכבה	Nakba
עזה	Gaza
אינתיפאדה	Intifada
הריסה	demolition
רקטות	Rockets
אנטישמי	Anti-semitic
התלהבות	Hamas
ימות	Will die
לאָנוס	Rape
להישרף לזה תן	Let it burn
הָרָג	killing
ערבי	Arabic
מוסלמי	Muslim
לשחוט	Slaughter
לשרוף	Burn
מוחמד	Muhammad

Sample of keywords used in collecting data.

Our Dataset

- Dataset Annotation and Guidelines
- Three graduate students were carefully selected and trained by an expert.
- Based on their familiarity with the Israeli politics and culture.
- Each student annotated 2,000 tweets out of 15,881.
- The expert reviewed the students' annotations and annotated the remaining 9,000 tweets.

Our Dataset

- For each tweet, the annotators used the following guideline:
- Class: Offensive (Hate, Abusive, Violence, Pornography) or NOT
- Target: the offended targets (the people or group(s)), UNT (ambiguous cases)
- Phrase: offensive phrase(s)
- Subject: topic(s) of offense ()

A	B	C	D	E	F
number	NOT	target	Subject	Phrases	TweetText
2534	NOT				@nachi_29 אני לא מכירה מספיק שמות ערבי, אבל חייבים נציג ערבי מוסלמי, נציג דודוי (לא, לא איוב קרא) ונציג ערבי נוצרי!
2535	hate	Palestine	racism	פלסטינים	תעמולת שווא סלובנית אנטישמית. מלא שנאה RTV השירות הציבורי הסלובני כלפי ישראל. אוהדי חמאס פלסטינים
2536	hate	palestine	racism	פלסטינים	RT @BayernMako: @rtvslo תעמולת שווא סלובנית אנטישמית. השירות הציבורי הסלובני מלא שנאה כלפי ישראל. אוהדי חמאס RTV פלסטינים
2537	violence	hamas	violence		@Roadrun86259229 @hagarsi במקום זה לא אותו דבר תמיד @DaphnaLiel אני אחסל את חמאס . אבל זה לא אותו דבר אחסל
2628	hate	arabs	offensive		@Tzoharkashrut שווה לבדוק אם אין ערבי תולעים וגם אם לא נשאר שם אידה ערבי חו"ח
2629	hate	arab community	politics	חוק הלאום	הפגנת הביביסטים: RT @Yaelfreidson מחוץ לבג"ץ חוק הלאום. רגע אחד מקללים את ראש מועצת סאג'ור ג'אבר חמוד 'יא http... 'מניאק מוחמד מזין אותך בתחת

Our Guidelines Dataset

OFFENSIVE SUB-CLASSES DEFINITIONS AND NUMBER OF TWEETS PER SUB-CLASS.

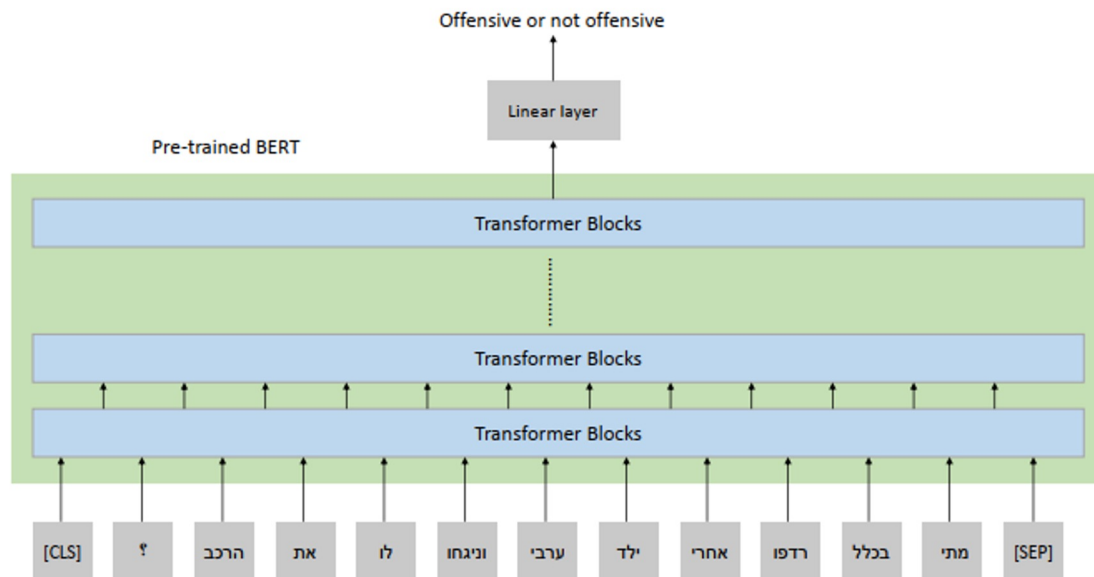
Class	Sub-Class	Definition	Count
Offensive	Abusive	If the tweet contains direct or implicit insults using vulgar or street words.	124
	Hate	If the tweet contains criticism, attack, or degrade, directly or implicitly, because of race, color, religion, nationality, or gender.	631
	Pornographic	If the tweet promotes or invites any pornographic or sexual arousal.	4
	Violence	If the tweet endorses an act that involves physical harm towards any party, regardless of the reason.	454
Not offensive	Not	If the tweet does not contain any offensive language.	14,681
Total			15,881

Examples

EXAMPLE OF ANNOTATED TWEETS PER CLASS.

Tweet	English Translation	Annotations
גירוש, הריסת בתים, מאסרי עולם ללא אפשרות חנינה, אחרת יהיה עוד יותר גרוע. להתייחס אליהם כמו אל מחבלים.	Deportation, demolition of houses, life sentences without the possibility of amnesty, otherwise it will be even worse. Treat them like terrorists.	Class: Violence, Hate Target: Palestinians Topic: punish Palestinians Phrase: Deportation, Demolition of houses, life sentences, terrorists
אין כבר הרתעה. לא מפחדים מהמשטרה. אני חושב שהגיע הזמן על פי מראות ההפגנות וההתפרעויות בימים האחרונים, כמו שאמר פעם רבין בתחילת האינתיפאדה: „לשבור להם את העצמות". פה יש כבר אינתיפאדה של התפרעויות.	There is no more deterrence. We are not afraid of the police. I think the time has come to face the demonstrations and riots, as Rabin once said at the beginning of the intifada: "to break their bones." There is already an intifada of riots here.	Class: Violence, Hate Target: Palestinians Topic: Demonstrations Phrase: Break their bones
ימח שמה וזכרה @Onetruth011 של אילנה דיין. העיתונאית הכי מנוולת ושקרנית שאני מכיר. ממש מרשעת.	May the name and memory of Ilana Dayan be remembered. The most depraved and lying journalist I know. Really sinister.	Class: Abusive Target: Ilana Dayan Topic: Journalism Phrase: Sinister, Depraved, Lying
@judash0 פרצופו האמיתי של אבי ביטון נחשף לעיני כל. מדובר בשמאלני, אנטי ציוני, עוכר ישראל, בוגד שממומן ע"י הקרן החדשה להפיל את שלטון הימין ולהעלות את המפלגות הערביות לשלטון כדי להוביל למדינת כל אזרחיה	Avi Beaton's true face clear now. This is a leftist, anti-Zionist, oppressor of Israel, a traitor who is financed to overthrow the right-wing government and bring the Arab parties to power in order to lead to a state for all its citizens.	Class: Hate, Abusive Target: Avi Bitton, Arab Parties Topic: politics Phrase: Traitor, Anti-Zionist
ה. וא לא @rabea_bader רלוונטי אם אתה דרוזי, סורי, אנטי ציוני ומגעיל שכמוך.	@rabea_bader is irrelevant if you are Druze, Syrian, anti-Zionist and disgusting like you.	Class: Hate, Abusive Target: Rabea Bader, Druze, Syrian Topic: Racism Phrase: disgusting, anti-Zionist
@Ahmad_tibi אתה לפחות לא משקר - היית ונשארת לאומן ערבי שרוצה בחורבן ישראל כמדינה יהודית.	@Ahmad_tibi At least you're not lying - you were and remain an Arab nationalist who wants the destruction of Israel as a Jewish state.	Class: Hate Target: Ahmad Tibi Topic: Political views Phrase:

Model Architecture



Model architecture.

Experiments and Results

Dataset Preparation

- The dataset is (14,681 not offensive and 1200 offensive) -> imbalanced
- We took 1200 offensive + 1300 not offensive randomly selected = 2500
- 2500 tweets => divided into
 - training (70%, 1750 tweets), validation (10%, 250 tweets) and test (20%, 500 tweets) sets.

Experiments and Results

Dataset Preparation

We combine it with the dataset published by *Litvak et al. (2021)*

- 2,000 comments (1,205 not offensive comments and 821 offensive comments) collected from Facebook.
- 1418 training, 405 for testing and 203 for validation,

Experiments and Results

Dataset	# of records from data	our # of records from D_{OLaH}	Training dataset	Test dataset	Validation dataset	Total
D_1	1750	0	1750	500 from our testing data	250 from our Val. dataset	2500
D_2	1750	1013	2763			3513
D_3	1750	2026	3776			4526
D_7	0	2026	2026			2776
D_4	0	1418	2026	405 from D_{OLaH}	203 from D_{OLaH}	2026
D_5	1250	1418	2668			3276
D_6	2500	1418	3918			4526
D_8	2500	0	2500			3108

. Eight dataset combinations, and two test sets.

Experiments and Results

- Experimental Settings
 - Eight combinations of datasets
 - Fine-tune the HeBERT and AlephBERT models
 - Maximum number of epochs was set to 10
 - Batch size = 8
 - Adam optimizer with a learning rate $\eta = 1e - 5$
 - The maximum input sequence length was 128.

Experiments and Results

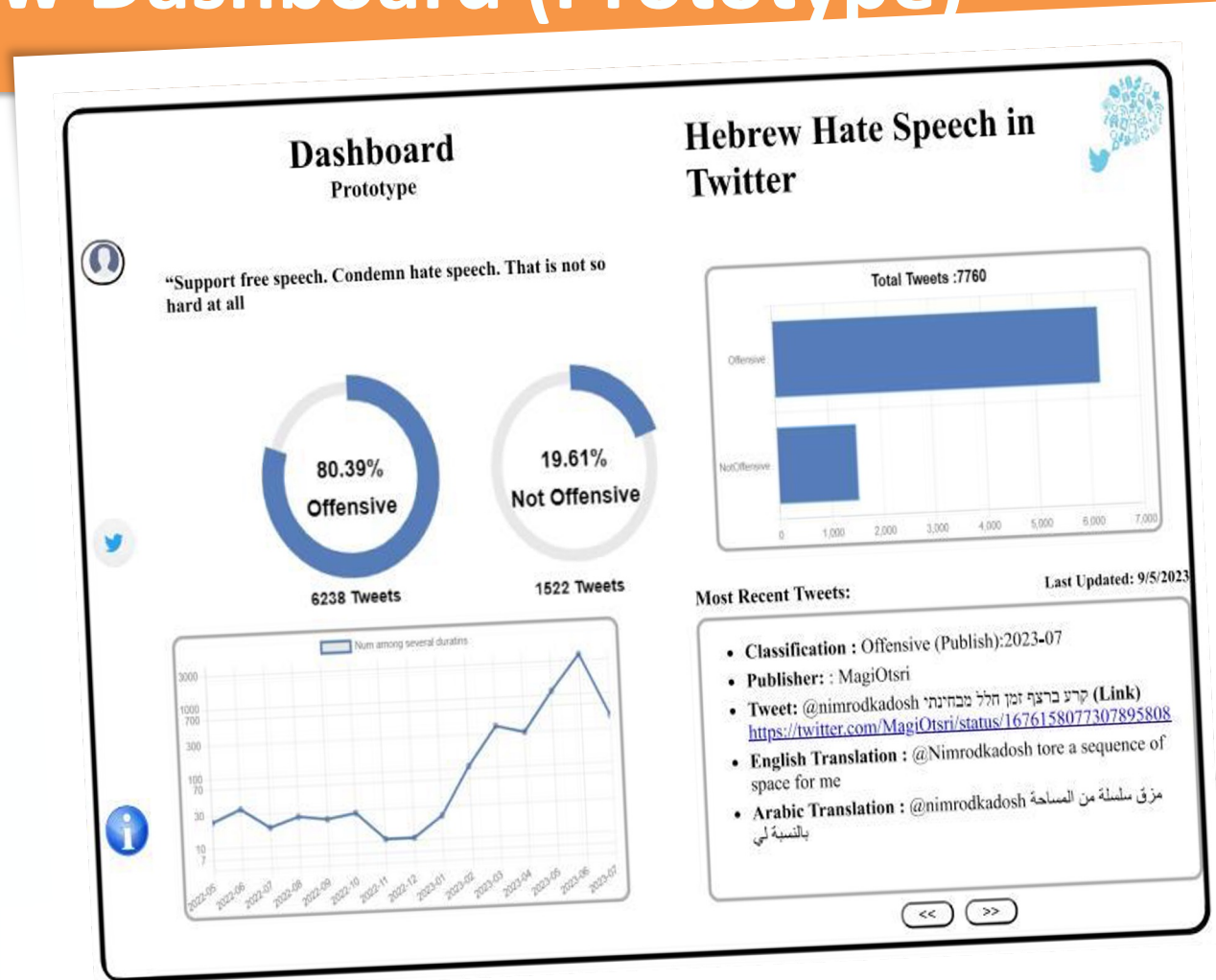
Experiments Results

- AlephBERT outperformed HeBERT in most experiments.
- Our training data boosts the HeBERT model performance when combined with the published dataset.
- Our best result on D6 (a combination of our dataset and published dataset on their test and validation)

Dataset	Accuracy	
	HeBERT	AlephBERT
D_1	63%	68%
D_2	58%	63%
D_3	61%	63%
D_4	79%	86%
D_5	81%	79%
D_6	81%	82%
D_7	53%	56%
D_8	61%	62%

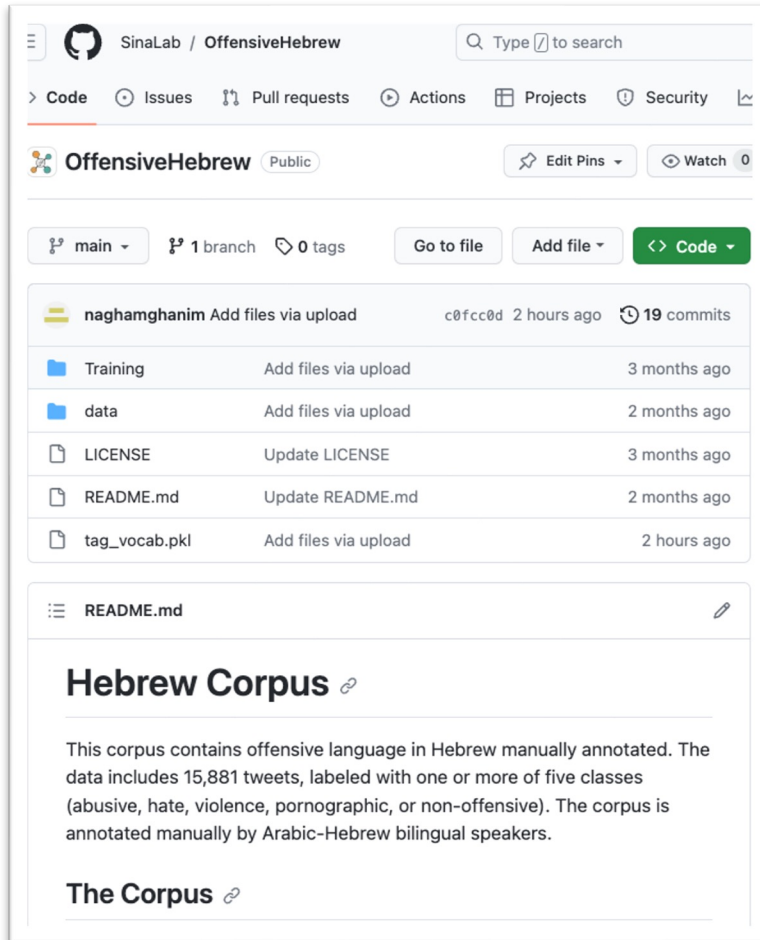
Hebrew Dashboard (Prototype)

- 75 twitter accounts (israeli politicians)
- May'2022 - July 2023



GitHub and Hugging face

Download Corpus and models



The screenshot shows the GitHub repository page for SinaLab / OffensiveHebrew. The repository is public and has 1 branch and 0 tags. The main branch is selected. The repository contains several files and folders, including Training, data, LICENSE, README.md, and tag_vocab.pkl. The README.md file is selected, showing the title "Hebrew Corpus" and a description of the corpus.

SinaLab / OffensiveHebrew

Code Issues Pull requests Actions Projects Security

OffensiveHebrew Public Edit Pins Watch 0

main 1 branch 0 tags Go to file Add file <> Code

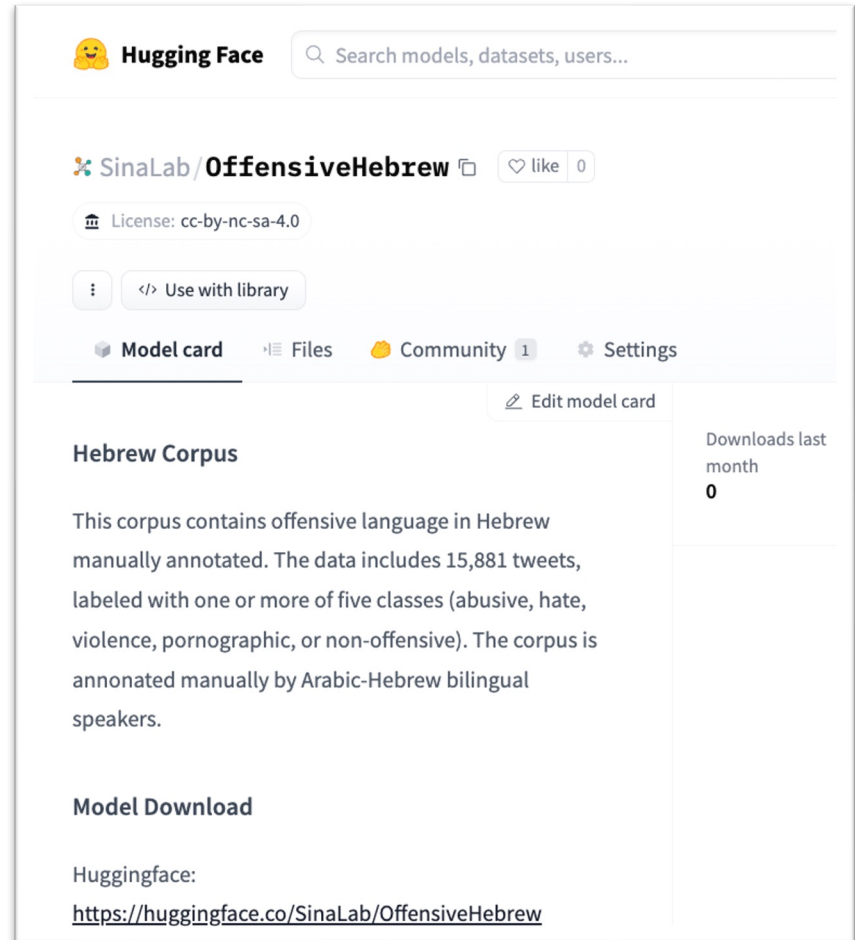
File/Folder	Action	Time
naghamghanim	Add files via upload	c0fcc0d 2 hours ago 19 commits
Training	Add files via upload	3 months ago
data	Add files via upload	2 months ago
LICENSE	Update LICENSE	3 months ago
README.md	Update README.md	2 months ago
tag_vocab.pkl	Add files via upload	2 hours ago

README.md

Hebrew Corpus

This corpus contains offensive language in Hebrew manually annotated. The data includes 15,881 tweets, labeled with one or more of five classes (abusive, hate, violence, pornographic, or non-offensive). The corpus is annotated manually by Arabic-Hebrew bilingual speakers.

The Corpus



The screenshot shows the Hugging Face repository page for SinaLab / OffensiveHebrew. The repository is public and has 0 likes. The repository contains a Model card, Files, Community, and Settings. The Model card is selected, showing the title "Hebrew Corpus" and a description of the corpus. The Model Download section is also visible.

Hugging Face

Search models, datasets, users...

SinaLab / OffensiveHebrew like 0

License: cc-by-nc-sa-4.0

</> Use with library

Model card Files Community 1 Settings

Edit model card

Hebrew Corpus

This corpus contains offensive language in Hebrew manually annotated. The data includes 15,881 tweets, labeled with one or more of five classes (abusive, hate, violence, pornographic, or non-offensive). The corpus is annotated manually by Arabic-Hebrew bilingual speakers.

Model Download

Huggingface:

<https://huggingface.co/SinaLab/OffensiveHebrew>

Downloads last month 0

References

1. Mustafa Jarrar, Sanad Malaysha, Tymaa Hammouda, Mohammad Khalilia: SALMA: Arabic Sense-Annotated Corpus and WSD Benchmarks. In Proceedings of the 1st Arabic Natural Language Processing Conference (ArabicNLP), Part of the EMNLP 2023. ACL.
2. Amal Nayouf, Tymaa Hammouda, Mustafa Jarrar, Fadi zaraket, Mohamad-Bassam Kurdy: Nâbra: Syrian Arabic Dialects with Morphological Annotations. In Proceedings of the 1st Arabic Natural Language Processing Conference (ArabicNLP), Part of the EMNLP 2023. ACL.
3. Mustafa Jarrar, Ahmet Birim, Mohammed Khalilia, Mustafa Erden, and Sana Ghanem: ArBanking77: Intent Detection Neural Model and a New Dataset in Modern and Dialectical Arabic. In Proceedings of the 1st Arabic Natural Language Processing Conference (ArabicNLP), Part of the EMNLP 2023. ACL.
4. Haneen Lqreina, Mustafa Jarrar, Mohammed Khalilia, Ahmed Oumar El-Shangiti, Muhammad AbdulMaged: Arabic Fine-Grained Entity Eecognition. In Proceedings of the 1st Arabic Natural Language Processing Conference (ArabicNLP), Part of the EMNLP 2023. ACL.
5. Mustafa Jarrar, Muhammad Abdul-Mageed, Mohammed Khalilia, Bashar Talafha, AbdelRahim El-madany, Nagham Hamad, Alaa' Omar: WojooodNER 2023: The First Arabic Named Entity Recognition Shared Task. In Proceedings of the 1st Arabic Natural Language Processing Conference (Arabic- NLP), Part of the EMNLP 2023. ACL.
6. Nouran Khallaf, Elin Arfon, Mo El-Haj, Jon Morris, Dawn Knight, Paul Rayson, Tymaa Hammouda, Mustafa Jarrar: Open-source thesaurus development for under-resourced languages: a Welsh case study. The 4th LDK Conference on Language, Data and Knowledge, Vienna, Austria, 12-15 September 2023
7. Nagham Hamad, Mustafa Jarrar, Mohammad Khalilia, Nadim Nashif: Offensive Hebrew Corpus and Detection using BERT. The 20th ACS/IEEE International Conference on Computer Systems and Applications (AICCSA). IEEE. Egypt. 2023
8. Sana Ghanem, Mustafa Jarrar, Radi Jarrar, Ibrahim Bounhas: A Benchmark and Scoring Algorithm for Enriching Arabic Synonyms. The 12th International Global Wordnet Conference (GWC2023), Global Wordnet Association. (pp.). San Sebastian, Spain, 2023
9. Sanad Malaysha, Mustafa Jarrar, Mohammad Khalilia: Context-Gloss Augmentation for Improving Arabic Target Sense Verification. The 12th International Global Wordnet Conference (GWC2023), Global Wordnet Association. (pp.). San Sebastian, Spain, 2023
10. Mustafa Jarrar, Mohammed Khalilia, Sana Ghanem: Wojoood: Nested Arabic Named Entity Corpus and Recognition using BERT. In Proceedings of the International Conference on Language Resources and Evaluation (LREC 2022), Marseille, France. 2022
11. Mustafa Jarrar, Fadi Zaraket, Tymaa Hammouda, Daanish Masood, Martin Waehlich: Lisan: Yemeni, Iraqi, Libyan, and Sudanese Arabic Dialect Corpora with Morphological Annotations. The 20th ACS/IEEE International Conference on Computer Systems and Applications (AICCSA). Pages(-). IEEE. Egypt. 2023 arXiv, DOI 10.48550/ARXIV.2212.06468. 2023
12. Karim El Haff, Mustafa Jarrar, Tymaa Hammouda, Fadi Zaraket: Curras + Baladi: Towards a Levantine Corpus. In Proceedings of the International Conference on Language Resources and Evaluation (LREC 2022), Marseille, France. 2022
13. Mustafa Jarrar: The Arabic Ontology - An Arabic Wordnet with Ontologically Clean Content. Applied Ontology Journal, 16:1, 1-26. IOS Press. 2021
14. Moustafa Al-Hajj, Mustafa Jarrar: ArabGlossBERT: Fine-Tuning BERT on Context-Gloss Pairs for WSD. In Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021). PP 40--48, 2021
15. Moustafa Al-Hajj, Mustafa Jarrar: LU-BZU at SemEval-2021 Task 2: Word2Vec and Lemma2Vec performance in Arabic Word-in-Context disambiguation. In Proceedings of the Fifteenth Workshop on Semantic Evaluation (SemEval2021) Task 2: Multilingual and Cross-lingual Word-in-Context Disambiguation (MCL-WIC). PP 748--755, Association for Computational Linguistics. 2021
16. Eman Naser-Karajah, Nabil Arman, Mustafa Jarrar: Current Trends and Approaches in Synonyms Extraction: Potential Adaptation to Arabic. In Proceedings of the 2021 International Conference on Information Technology (ICIT). PP 748--755, Association for Computational Linguistics. pp. 428-434, IEEE. 2021
17. Mustafa Jarrar, Eman Karajah, Muhammad Khalifa, Khaled Shaalan: Extracting Synonyms from Bilingual Dictionaries. The 11th International Global Wordnet Conference (GWC2021), Global Wordnet Association. (pp. 215-222). Pretoria, South Africa, 2021
18. Kareem Darwish, Nizar Habash, Mourad Abbas, Hend Al-Khalifa, Husein T. Al-Natsheh, Houada Bouamor, Karim Bouzoubaa, Violetta Cavalli-Sforza, Samhaa R. El-Beltagy, Wassim El-Hajj, Mustafa Jarrar, Hamdy Mubarak: A Panoramic Survey of Natural Language Processing in the Arab World. Communications of the ACM, April 2021, Vol. 64 No. 4, Pages 72-81
19. Mustafa Jarrar: Digitization of Arabic Lexicons. Arabic Language Status Report. UAE Ministry of Culture and Youth. Pages 214-2017. Dec 2020
20. Mustafa Jarrar, Hamzeh Amayreh: An Arabic-Multilingual Database with a Lexicographic Search Engine. The 24th International Conference on Applications of Natural Language to Information Systems (NLDB 2019). Pages(234-246). LNCS 11608, Springer. 2019
21. Mustafa Jarrar, Hamzeh Amayreh, John P. McCrae: Representing Arabic Lexicons in Lemon - a Preliminary Study. The 2nd Conference on Language, Data and Knowledge (LDK 2019). Pages(29-33). CEUR, Volume 2402. ISSN:1613-0073. Leipzig, Germany. 2019
22. Diana Alhafi, Anton Deik, Mustafa Jarrar: Usability Evaluation of Lexicographic e-Services. The 16th IEEE/ACS International Conference on Computer Systems and Applications (AICCSA). Pages(1-7). IEEE. Abu Dhabi, UAE. 2019
23. Mustafa Jarrar, Fadi Zaraket, Rami Asia, Hamzeh Amayreh: Diacritic-Based Matching of Arabic Words. ACM Asian and Low-Resource Language Information Processing. Volume 18, No 2, Pages(10:1-10:21), ACM, ISSN:2375-4699. December, 2018
24. Mustafa Jarrar: Search Engine for Arabic Lexicons. The 5th Conference on Translation and the Problematics of Cross-cultural Understanding. The Forum for Arab and International Relations. Doha, Qatar. December, 2018
25. Diab Abuaiadah, Dileep Rajendran, Mustafa Jarrar: Clustering Arabic Tweets for Sentiment Analysis. The 2017 IEEE/ACS 14th International Conference on Computer Systems and Applications. Pages(499-506). IEEE Computer Society. ISBN:9781538635810. (doi.10.1109/AICCSA.2017.162). Hammamet, Tunisia. 2017
26. Mustafa Jarrar, Nizar Habash, Faec Alrimawi, Divam Akra, Nasser Zalmout: Curras: An Annotated Corpus for the Palestinian Arabic Dialect. Journal Language Resources and Evaluation. Pages(745-775). Volume(51)