From Linguistics to Practice: a Case Study of Offensive Language Taxonomy in Hebrew

Chaya Liebeskind
Department of Computer Science
Jerusalem College of Technology
Jerusalem, Israel
liebchaya@gmail.com

Natalia Vanetik and **Marina Litvak**
Department of Software Engineering
Shamoon College of Engineering
Beer-Sheva
{natalyav,marinal}@sce.ac.il

Abstract

The perception of offensive language varies based on cultural, social, and individual perspectives. With the spread of social media, there has been an increase in offensive content online, necessitating advanced solutions for its identification and moderation. This paper addresses the practical application of an offensive language taxonomy, specifically targeting Hebrew social media texts. By introducing a newly annotated dataset, modeled after the taxonomy of explicit offensive language of (Lewandowska-Tomaszczyk et al., 2023), we provide a comprehensive examination of various degrees and aspects of offensive language. Our findings indicate the complexities involved in the classification of such content. We also outline the implications of relying on fixed taxonomies for Hebrew.

1 Introduction

The definition of offensive language can vary depending on cultural, social, and personal viewpoints. In a general sense, offensive language encompasses any form of communication that may upset or discomfort individuals or groups (Haugh and Sinkeviciute, 2019; Lewandowska-Tomaszczyk, 2023). It can be broadly categorized into explicit forms (Kogilavani et al., 2021; Lewandowska-Tomaszczyk, 2023), including insults and hate speech, and implicit forms which use subtle insinuations or coded language to convey bias. Social media platforms have become significant sources of offensive language, with surveys revealing a rise in hate speech instances (Alsagheer et al., 2022; Costello and Hawdon, 2020). Numerous countries have laws against hate speech and false information. Failure to properly regulate such content can result in legal consequences and harm to a platform's reputation. While content filters on platforms can help reduce offensive language, their effectiveness is diminishing due to the growth of user-generated content. Consequently, Natural Language Processing (NLP) techniques are gaining importance in identifying offensive language. However, detecting offensive language in low-resource languages, like Hebrew, remains a challenge (Zampieri et al., 2019b) due to the lack of available resources.

The taxonomy of offensive language is crucial as it establishes a structured framework for various inappropriate content, assisting automated systems in moderating and responding to such content. This classification creates a foundational structure that not only streamlines the intricate landscape of online communication but also acts as an instrument to enhance the safety and functionality of digital platforms. The practicality of offensive language taxonomies often raises concerns, especially in the ever-evolving digital landscape. Creating a comprehensive taxonomy is challenging given the vast and nuanced spectrum of offensive content. Relying solely on a static taxonomy may not capture the dynamic nature of language, especially as slang, idioms, and colloquialisms evolve. There's also a risk of misinterpretation or misclassification, which can inadvertently lead to stifling genuine discussions or failing to catch genuinely harmful content.

This study distinguishes itself from prior studies on identifying offensive texts by deviating from the approach of just focusing on a certain form of offensive language or relying on an intuitive definition that encompasses various kinds of offensive language, without being grounded in a systematic linguistic taxonomy. In this paper, we study the practical implications of applying an offensive language taxonomy to the collection and analysis of Hebrew social media texts. For this purpose, we present here a new annotated dataset following a simplified taxonomy of explicit offensive language introduced in (Lewandowska-Tomaszczyk et al., 2023). The data represents all the levels

Proceedings of the The 8th Workshop on Online Abuse and Harms (WOAH), pages 110–117

of this taxonomy, which allowed us to examine the practical consequences of collecting and analyzing offensive texts. We were able to determine what types and aspects of offensive language pose a significant challenge for binary and multi-class classification of offensive language.

This paper is organized as follows. Section 2 covers the related work. Section 3.2 describes the collection and annotation of the offensive language dataset in Hebrew, and Section 3.3 reports on the dataset analysis. Finally, Section 4 concludes our work and describes potential future tasks.

2 Related Work

Multiple works on automated offensive language detection exist, including early unsupervised lexicon-based approaches (Tulkens et al., 2016), traditional supervised approaches (Davidson et al., 2017), and recent approaches based on deep neural networks (Zampieri et al., 2019b) and transformer models (Liu et al., 2019; Ranasinghe et al., 2019). However, the clear majority of the offensive detection studies deal with English. Recently, many researchers started to develop multilingual methodologies and annotated corpora in multiple languages. For example, such languages as Arabic (Mohaouchane et al., 2019), Dutch (Tulkens et al., 2016), French (Chiril et al., 2019), Turkish (Çöltekin, 2020), Danish (Sigurbergsson and Derczynski, 2019), Greek (Pitenis et al., 2020), Italian (Poletto et al., 2017), Portuguese (Fortuna et al., 2019), Slovene (Fišer et al., 2017), and Dravidian (Yasaswini et al., 2021) were explored for the task of offensive content identification.

Despite the great international effort, many low-resource languages got much less attention than others. For example, only a few works proposed solutions for Hebrew: a Hebrew corpus of user comments annotated for abusive language was introduced in (Liebeskind and Liebeskind, 2018); an annotated Facebook comments dataset and a system for offensive text detection was suggested in (Litvak et al., 2021), and a union of these two datasets and together with monolingual, cross-lingual, and multilingual experiments for the task of offensive language detection was presented in (Litvak et al., 2022). Hebrew and Arabic are both members of the same family of languages known as the Semitic languages, and some authors made use of the wealth of resources available in Arabic. For example, the most recent work introduced a new offensive language corpus in Hebrew containing 15,881 Twitter labeled by Arabic-Hebrew bilingual speakers into one or more of the five available classes, namely abuse, hate, violence, pornography, or non-offensive (Hamad et al., 2023). Fine-tuning of pre-trained Hebrew LLMs showed that the proposed dataset is beneficial for the detection of offensive language in Hebrew (Litvak et al., 2022).

The first offensive language taxonomy suitable for social media content appeared in (Zampieri et al., 2019a,b). This three-level hierarchy for offensive language classification was created to offer a methodical technique to distinguish between various forms and degrees of offensive language. In (Lewandowska-Tomaszczyk et al., 2022), the combined schema for explicit and implicit offensive language was tested on English datasets, and difficulties with agreement among annotators about the distinction of particular categories emerged. Based on linguistic ideas like Grice's implicitness categories, the work (Lewandowska-Tomaszczyk, 2023) established a holistic method that targets both explicit and implied types of abusive language. However, to this day, no applications or evaluation of similar taxonomy in Hebrew exists.

3 Hebrew Offensive Language Taxonomy and Dataset

3.1 Taxonomy

We derive the aspects of offensive language for Hebrew from the taxonomy proposed by (Lewandowska-Tomaszczyk et al., 2023) that in its turn extends a taxonomy proposed in (Zampieri et al., 2019a,b). We have translated this taxonomy to Hebrew and focused on the first six layers that represent explicit offensive language. In this taxonomy (depicted in Figure 1), after deciding of whether or not the text is offensive, one has to determine the presence or absence of the target of an offense, then decide on the type of target, and rule whether or not the expression is vulgar. The next step is to state what is the severity of the offense (discrediting, insulting, hate speech, threat) and what are the offense aspects (racism, homophobia, xenophobia, religious profanity, sexism, ageism, ableism, ideologism, classism, undetermined).

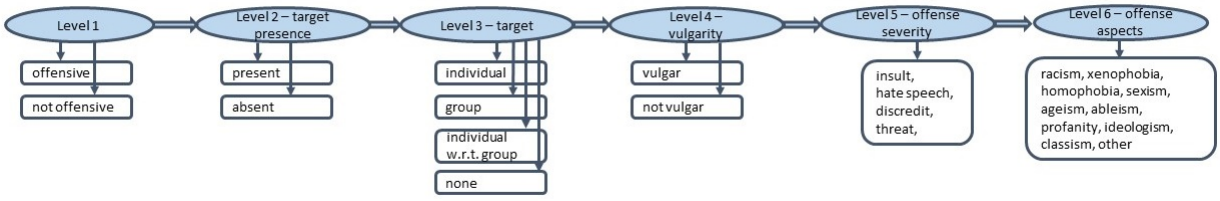

Figure 1: Explicit offensive language taxonomy

3.2 Dataset Collection and Annotation

As a starting point for data collection, we created a list of offensive terms in Hebrew using the method of (Liebeskind and Liebeskind, 2018), as follows. Initially, 67 offensive terms were chosen, and then they were supplemented using a statistical measure of word co-occurrence. We obtained the 100 most similar words for each offending term in the first list using the Dice coefficient (Smadja et al., 1996) and a sizable unannotated corpus of Facebook comments (Liebeskind and Liebeskind, 2018), supposing that words that often occur together are thematically relevant (Schütze and Pedersen, 1997). Then, from these candidate lists, 683 offensive terms were manually chosen and assigned to one or more offensive aspects. Note that we could not find any example of xenophobia that is not racist, so this aspect is excluded from the analysis. We adopted a classification method that requires only a context-based connection between the offensive term and the aspect. For instance, the word עלוקה (leech) has been categorized as profanity because it is frequently directed at a particular religious group of the population. Or, for instance, the word גנב (thief) has been labeled as classism because criminals frequently belong to a particular social class. This strategy aims to obtain a diverse dataset that cannot be separated by the search terms alone, necessitating the annotation and analysis presented in this work. Finally, we extracted offensive tweets from Twitter using the offensive terms. In this manner, we ensured that our data encompasses all aspects of offensive language, not just the most prevalent types. Consequently, we were able to evaluate the applicability of offensive taxonomy for dataset creation.

To demonstrate the efficacy of our extraction method, we trained the 100-dimensional fastText word embeddings (Bojanowski et al., 2017) on the constructed dataset that is suitable for morphologically rich languages, such as Hebrew. Using t-Distributed Stochastic Neighbor Embedding (t-SNE) (Belkina et al., 2019), we retrieved 30 neighboring words for each aspect and visualized the results. We prefer the t-SNE method over the Principal Component Analysis (PCA) (Shlens, 2014) method because it captures nonlinear structures and clusters in high-dimensional data more effectively. Figure 2 shows that there is a clear separation between the neighboring words that occur in the different offensive aspects, indicating that they are readily identified. However, owing to their close association in reality, certain categories virtually overlap, such as racial and ideological (making racism an ideology) or ableism and classism (identifying a person in a different socioeconomic position as handicapped).

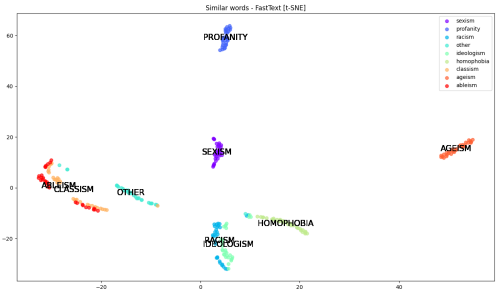

Figure 2: t-SNE-based visualization of the 30 neighboring fastText vectors

We used the INCEpTION platform (Klie et al., 2018) to produce annotations. The data was divided into 9 files, one file per offensive aspect, with 50 comments in every file, making it 450 texts in total. Our annotators were unaware of this division. The texts were given to two native Hebrew speakers who were requested to annotate them independently. Given that the texts came from social networks frequented by young individuals who use slang and modern language, we selected annotators between the ages of 20 and 30. The annotators were first asked to decide whether or not a text is offensive and then to proceed ac-

cording to taxonomy levels of Section 3.1; we have computed Cohen's Kappa agreement coefficient (Cohen, 1960) for every level/parameter separately. First, the annotators determined whether or not the target of the offense is present in the text (agreement 0.49), then they identified the target's type (agreement 0.84) and the severity of the offense (agreement 0.73 for hate speech and insult, 0.66 for discrediting, and 0.97 for threat), and they determined whether or not the expression is vulgar (agreement 0.63). As the last step, the annotators were instructed to list (in alphabetical order) each aspect of explicit offensive language that applies to the text, achieving an agreement of 0.68. Calculating the inter-annotator agreement not only allows us to evaluate the clarity of the annotation guidelines but also the inherent difficulty of the classification and how well humans comprehend the task. In order to create the final dataset, we resolved instances where the labels did not align by involving a third annotator for disambiguation.

3.3 Dataset Analysis

Table 1 describes the three tokens with the highest tf-idf values for every file. We can see that some words appear across files, for example, the word לך (go), which is not a vulgar word but may be considered impolite if it is used as "get out" in the sentence. The categories where words are related (although these words are not necessarily vulgar or insults) are "homophobia" and "ideologism" where, for example, the last name of a former prime minister is mentioned (Bennet).

Table 2 shows the three words with the highest normalized count for every file. Again, we see that words appear across files.

To tokenize the texts, we cleaned the data from punctuation, numbers, and non-Hebrew characters, and applied the AlephBERT tokenizer (Seker et al., 2021).

We see the words that have high tf-idf values or high unigram count are not necessarily the words related to their respective offensive aspect, except for "sexism" and "profanity" files. Moreover, these words often represent the most prominent tokens in more than one file. For instance, an unrelated offensive word such as עבריין (a criminal) is among the most common words in the file "sexism". Therefore, straightforward word-based classification does not seem very helpful in this case.

To evaluate the creation process' validity and to better comprehend the practical applicability of the annotated dataset we extracted the data for specific offensive language categorization tasks using the various taxonomy levels.

Table 4 reports the results of the binary classification for every offensive category with at least 10 sentences. We treat the category sentences as positive samples, and the rest of the sentences as negative samples. This table also reports the final dataset statistics, i.e., the number of sentences in the dataset that were annotated as containing a specific offensive aspect. Note that there are sentences that were not annotated as offensive at all, and therefore the total number of sentences is smaller than 450. We have applied eXtreme Gradient Boost (XGB) (Chen et al., 2015) (we have also applied Random Forest (RF) (Pal, 2005) and Logistic Regression (LR) (Wright, 1995), but XGB provided slightly better results). to texts represented as BERT sentence embeddings encoded with AlephBERT (Seker et al., 2021). We split the data into training and test sets (80%/20%) and classified offensive types/aspects with at least 10 sentences. For example, in offensive aspects, this pruning left us with 7 categories out of 10. We see that upper taxonomy levels such as target presence accuracy exceed the majority values significantly; however, lower taxonomy levels pose a more serious challenge - vulgarity and severity of the offense are especially difficult. On the lowest taxonomy level for most of the offensive aspects, the accuracy does not exceed the majority values, except for the "other" aspect which is the largest class. However, "homophobia" has significantly higher precision than other classes.

As baselines we applied two fine-tuned transformers – a multilingual BERT model or (HuggingFace, 2024) which we denote by *mlbert*, and the Hebrew BERT model of (Chriqui and Yahav, 2021) denoted by *hebert*, to the task of binary classification for different levels of our taxonomy. We have fine-tuned every model for 10 epochs with batch size 16, Adam optimizer, and standard learning rate of 0.00002. All texts were padded to the maximal length, and the attention mask was set to ignore the padded tokens. Comparative results of these transformer models appear in Table 5. We can see that the hebert model has an obvious advantage over the mlbert for all the categories, but both models perform worse that traditional classifiers.

In Table 3 we report the results of the multiclass classification of offensive parameters per tax-

file	words with top tf-idf	transcription	translation
racism	פאשיסט , נבלה, נבלות	phashist, navela, navelot	fascist, scavenger, scavengers
homophobia	לסביות, קוקסינלים , התחת	lesbiot, koksinelim, hatachat	lesbians, shemales , the a**
sexism	לך, בוגדים, תעופי	lekh, bogdim, ta'ofi	go , traitors, get out
profanity	הזה, שלו, לך	haze, she'lo, lekh	this , that is not, go
ageism	הולני , הזויה, די	cholani, hazuya, day	sick, delusional, enough
ableism	קרימינל, קשקשן, זבל	kriminal, kashkashan, zevel	criminal, rascal, garbage
classism	עלובה , מסית, עבריין	aluva, matsit, avaryan	wretched, agitator, offender
ideologism	בנט, בשלטון, מושחט	Bennett, b'shelton, moshachat	Bennet, in power, corrupt
other	לך, כבר, פה	pach, kvar, lekh	trash can , already , go
all files	שלא, עלובה, לך	lekh, aluva, she'lo	go, wretched, that is not

Table 1: Tokens with highest tf-idf values per file.

file	unigrams	transcription	translation
racism	כלום, הוץ, עכשיו	klum, utz, akhshav	now, except, nothing
homophobia	ילדה, לך, ראיתי	yalda, lekh, ra'iti	I saw, go, girl
sexism	עבריין, עלובה, מסית	avariyan, aluva, mesit	agitator, wretched, criminal
profanity	כמה, לסביות, התחת	kama, lesbiyot, taat	the a**, lesbians, how much
ageism	דיקטטורי, הכל, בנט	diktatory, hakol, Bennett	Bennet, all, dictatorial
ableism	בכל, עוד, כבר	bekol, od, kvar	already, more, in every
classism	בן, עוד, נבלה	ben, od, neveilah	scavanger, more, son
ideoligism	לך, הזה, שלא	lekh, hazeh, she'lo	that not, this, go
other	ולא, ערב, עכשיו	ve lo, erev, akhshav	now, evening, and not
all files	כמה, עוד, לך	kama, od, lekh	go, more, how much

Table 2: Unigrams with top counts per file.

parameter	classes	F1	acc	maj
presence	2	0.699	0.699	0.518
target type	4	0.232	0.615	0.641
severity	4	0.201	0.354	0.616
vulgarity	2	0.589	0.616	0.565
aspects	7	0.125	0.488	0.545

Table 3: Multiclass classification of offense types and aspects.

onomy level. We can see that accuracy decreases as we descend through taxonomy levels, with one notable exception - offense severity is the hardest category to classify.

4 Conclusions and Limitations

This paper explores the use of an offensive language taxonomy for Hebrew social media content. Using a new dataset annotated following the taxonomy of (Lewandowska-Tomaszczyk et al., 2023), we highlight the challenges of classification and the limitations of static taxonomies for Hebrew. The difficulty in classifying categories like vulgarity and offense severity shows the complexities of interpreting linguistic nuances. The results from the multi-class classification further reinforced the notion that as we venture deeper into the taxonomy levels, the task of classification becomes progressively challenging. In sum, this paper underlines the paramount importance of a multifaceted approach to offensive language detection. Relying solely on individual words or fixed taxonomies may not capture the multifarious nature of language, especially when dealing with nuanced topics like offensive content. Future efforts should consider incorporating advanced linguistic models and domain-specific knowledge to enhance classification performance, especially at more granular taxonomy levels.

Acknowledgments

This work was supported by the Israel Innovation Authority. The subject of the program is the development of a dataset and a language model for identifying offensive language in Hebrew and Arabic.

A Appendix

parameter	category	sentences	P	R	F1	acc	majority
presence	present	175	0.695	0.693	0.693	0.694	0.513
presence	absent	184	0.676	0.669	0.664	0.667	0.513
target	group	86	0.502	0.501	0.492	0.718	0.777
target	non-targeted	39	0.448	0.493	0.469	0.885	0.899
target	individual	247	0.542	0.527	0.511	0.615	0.640
target	ind. wrt. gr./gr. wrt. ind.	14	0.480	0.487	0.483	0.936	0.964
severity	discredit	103	0.370	0.380	0.375	0.600	0.794
severity	insult	303	0.420	0.427	0.421	0.470	0.607
severity	hate speech	89	0.493	0.497	0.479	0.780	0.822
vulgatiry	vulgar	157	0.507	0.507	0.503	0.528	0.563
vulgarity	not vulgar	202	0.583	0.564	0.551	0.597	0.563
aspect	homophobia	32	0.726	0.654	0.681	0.930	0.925
aspect	sexism	12	0.488	0.500	0.494	0.977	0.972
aspect	racism	26	0.470	0.481	0.476	0.907	0.939
aspect	classism	10	0.488	0.494	0.491	0.965	0.977
aspect	other	229	0.602	0.599	0.599	0.605	0.534
aspect	ideologism	100	0.589	0.527	0.509	0.756	0.767
aspect	profanity	11	0.488	0.494	0.491	0.965	0.974

Table 4: Binary classification of offensive categories.

parameter	category	mlbert acc	hebert acc
presence	absent	0.377	0.494
presence	present	0.558	0.494
target	non-targeted	0.610	0.909
target	individual	0.558	0.662
target	ind. wrt. gr./gr. wrt. ind.	0.610	0.974
target	group	0.675	0.636
severity	insult	0.377	0.234
severity	hate speech	0.558	0.234
severity	discredit	0.584	0.299
severity	threat	0.623	0.013
vulgarity	not vulgar	0.351	0.571
vulgarity	vulgar	0.584	0.571
aspect	racism	0.766	0.416
aspect	homophobia	0.636	0.909
aspect	sexism	0.623	0.013
aspect	other	0.325	0.455
aspect	profanity	0.584	0.987
aspect	ideologism	0.507	0.351
aspect	classism	0.571	0.013
aspect	ageism	0.675	0.987

Table 5: Binary classification of offensive categories with fine-tuned transformers.

References

Dana Alsagheer, Hadi Mansourifar, and Weidong Shi. 2022. Counter hate speech in social media: A survey. *arXiv preprint arXiv:2203.03584*.

Anna C Belkina, Christopher O Ciccolella, Rina Anno, Richard Halpert, Josef Spidlen, and Jennifer E Snyder-Cappione. 2019. Automated optimized parameters for t-distributed stochastic neighbor embedding improve visualization and analysis of large datasets. *Nature communications*, 10(1):5415.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the association for computational linguistics*, 5:135–146.

Çağrı Çöltekin. 2020. A corpus of turkish offensive language on social media. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6174–6184.

Tianqi Chen, Tong He, Michael Benesty, Vadim Khotilovich, Yuan Tang, Hyunsu Cho, Kailong Chen, et al. 2015. Xgboost: extreme gradient boosting. *R package version 0.4-2*, 1(4):1–4.

Patricia Chiril, Farah Benamara, Véronique Moriceau, Marlene Coulomb-Gully, and Abhishek Kumar. 2019. Multilingual and multitarget hate speech detection in tweets. In *Conférence sur le Traitement Automatique des Langues Naturelles (TALN-PFIA 2019)*, pages 351–360. ATALA.

Avihay Chriqui and Inbal Yahav. 2021. Hebert and hebemo: a hebrew bert model and a tool for polarity analysis and emotion recognition. *arXiv preprint arXiv:2102.01909*.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Matthew Costello and James Hawdon. 2020. Hate speech in online spaces. *The Palgrave handbook of international cybercrime and cyberdeviance*, pages 1397–1416.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the international AAAI conference on web and social media*, volume 11, pages 512–515.

Darja Fišer, Tomaž Erjavec, and Nikola Ljubešić. 2017. Legal framework, dataset and annotation schema for socially unacceptable online discourse practices in Slovene. In *Proceedings of the first workshop on abusive language online*, pages 46–51.

Paula Fortuna, Joao Rocha da Silva, Leo Wanner, Sérgio Nunes, et al. 2019. A hierarchically-labeled Portuguese hate speech dataset. In *Proceedings of the third workshop on abusive language online*, pages 94–104.

Nagham Hamad, Mustafa Jarrar, Mohammad Khalilia, and Nadim Nashif. 2023. Offensive hebrew corpus and detection using bert. *arXiv preprint arXiv:2309.02724*.

Michael Haugh and Valeria Sinkeviciute. 2019. Offence and conflict talk. *The Routledge handbook of language in conflict*, pages 196–214.

HuggingFace. 2024. XLM-RoBERTa-Multilingual-Hate-Speech-Detection-New: A Pretrained Model for Multilingual Hate Speech Detection. https://huggingface.co/christinacdl/XLM_RoBERTa-Multilingual-Hate-Speech-Detection-New. Accessed: April 23, 2024.

Jan-Christoph Klie, Michael Bugert, Beto Boullosa, Richard Eckart de Castilho, and Iryna Gurevych. 2018. The inception platform: Machine-assisted and knowledge-oriented interactive annotation. In *proceedings of the 27th international conference on computational linguistics: system demonstrations*, pages 5–9.

SV Kogilavani, S Malliga, KR Jaiabinaya, M Malini, and M Manisha Kokila. 2021. Characterization and mechanical properties of offensive language taxonomy and detection techniques. *Materials Today: Proceedings*.

Barbara Lewandowska-Tomaszczyk. 2023. A simplified taxonomy of offensive language (sol) for computational applications. *Konin Language Studies*, 10(3):213–227.

Barbara Lewandowska-Tomaszczyk, Anna Bączkowska, Chaya Liebeskind, Giedre Valunaite Oleskeviciene, and Slavko Žitnik. 2023. An integrated explicit and implicit offensive language taxonomy. *Lodz Papers in Pragmatics*, 19(1):7–48.

Barbara Lewandowska-Tomaszczyk, Slavko Žitnik, Chaya Liebeskind, Giedrė Valūnaitė-Oleškevičienė, Anna Bączkowska, Paul A Wilson, Marcin Trojszczak, Ivana Brač, Lobel Filipić, Ana Ostroški Anić, et al. 2022. Annotation scheme and evaluation: The case of offensive language. *Rasprave*.

Chaya Liebeskind and Shmuel Liebeskind. 2018. Identifying abusive comments in Hebrew Facebook. In *2018 IEEE International Conference on the Science of Electrical Engineering in Israel (ICSEE)*, pages 1–5. IEEE.

Marina Litvak, Natalia Vanetik, Chaya Liebeskind, Omar Hmdia, and Rizek Abu Madeghem. 2022. Offensive language detection in hebrew: can other languages help? In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3715–3723.

Marina Litvak, Natalia Vanetik, Yaser Nimer, Abdulrhman Skout, and Israel Beer-Sheba. 2021. Offensive language detection in semitic languages. In *Multimodal Hate Speech Workshop*, volume 2021, pages 7–12.

Ping Liu, Wen Li, and Liang Zou. 2019. NULI at SemEval-2019 task 6: Transfer learning for offensive language detection using bidirectional transformers. In *Proceedings of the 13th international workshop on semantic evaluation*, pages 87–91.

Hanane Mohaouchane, Asmaa Mourhir, and Nikola S Nikolov. 2019. Detecting offensive language on Arabic social media using deep learning. In *2019 Sixth International Conference on Social Networks Analysis, Management and Security (SNAMS)*, pages 466–471. IEEE.

Mahesh Pal. 2005. Random forest classifier for remote sensing classification. *International journal of remote sensing*, 26(1):217–222.

Zeses Pitenis, Marcos Zampieri, and Tharindu Ranasinghe. 2020. Offensive language identification in Greek. *arXiv preprint arXiv:2003.07459*.

Fabio Poletto, Marco Stranisci, Manuela Sanguinetti, Viviana Patti, Cristina Bosco, et al. 2017. Hate speech annotation: Analysis of an Italian Twitter corpus. In *Ceur workshop proceedings*, volume 2006, pages 1–6. CEUR-WS.

Tharindu Ranasinghe, Marcos Zampieri, and Hansi Hettiarachchi. 2019. Brums at hasoc 2019: Deep learning models for multilingual hate speech and offensive language identification. In *FIRE (working notes)*, pages 199–207.

Hinrich Schütze and Jan O Pedersen. 1997. A cooccurrence-based thesaurus and two applications to information retrieval. *Information Processing & Management*, 33(3):307–318.

Amit Seker, Elron Bandel, Dan Bareket, Idan Brusilovsky, Refael Shaked Greenfeld, and Reut Tsarfaty. 2021. Alephbert: A hebrew large pretrained language model to start-off your hebrew nlp application with. *arXiv preprint arXiv:2104.04052*.

Jonathon Shlens. 2014. A tutorial on principal component analysis. *arXiv preprint arXiv:1404.1100*.

Gudbjartur Ingi Sigurbergsson and Leon Derczynski. 2019. Offensive language and hate speech detection for Danish. *arXiv preprint arXiv:1908.04531*.

Frank Smadja, Kathleen R McKeown, and Vasileios Hatzivassiloglou. 1996. Translating collocations for bilingual lexicons: A statistical approach. *Computational linguistics*, 22(1):1–38.

Stéphan Tulkens, Lisa Hilte, Elise Lodewyckx, Ben Verhoeven, and Walter Daelemans. 2016. A dictionary-based approach to racism detection in dutch social media. *arXiv preprint arXiv:1608.08738*.

Raymond E Wright. 1995. Logistic regression.

Konthala Yasaswini, Karthik Puranik, Adeep Hande, Ruba Priyadharshini, Sajeetha Thavareesan, and Bharathi Raja Chakravarthi. 2021. IIITT@ DravidianLangTech-EACL2021: Transfer learning for offensive language detection in Dravidian languages. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 187–194.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019a. Predicting the type and target of offensive posts in social media. *arXiv preprint arXiv:1902.09666*.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019b. Semeval-2019 task 6: Identifying and categorizing offensive language in social media (offenseval). *arXiv preprint arXiv:1903.08983*.