Assignment Hate Speech Lexicons

Tanya Kaintura, Payanshi Jain, Rishikesh Narayanan Ramachandran, Antonios Georgakopoulos

September 24, 2023

1 Description of Lexicons

1.1 Wiegand

Question 1: What is the motivation for the paper? Describe how these lexicons are built?
The authors of this paper aim to build a lexicon that can optimally detect abusive content in a cross-domain environment, meaning that regardless of the domain that each comment/word is in, the lexicon will be still able to find the abusive instances. The goal of the authors is to show that the information that they manage to learn during the training of the classifier and the utilization of various novel features, in the process of training, is of substantially higher quality than just training a classifier on a large dataset.

The first step of the process of creating the lexicon for the classifier, begins with gathering a certain amount of negative expression for their base lexicon that is annotated by crowdsourcing. They used the "Subjectivity Lexicon" to gather a total of 500 negative words (nouns, verbs, adjectives) because this lexicon also contains the valuable information of sentiment views and polar intensity for each of those words. To add some abusive words that were missing from the lexicon mentioned above, they also made use of a word list mentioned in the "Schmidt and Wiegand(2017)". Each word of the lexicon could be classified as *abusive* or *non-abusive* based on how many raters consider it abusive. The corpora used for this experiment were the Amazon Review Corpus (AMZ), the Web As Corpus (WAC) and the rateitall.com (RIA). In order to expand their base lexicon to a larger one, they use a lot of expressions from Wiktionary and utilizing an SVM trained on the base lexicon using the best configuration with regard to what feature is going to be used.

The features that the authors propose are separated in two major categories: linguistic features and word embeddings. For the linguistic features, polar intensity is the first incorporated and can be split into three types: binary intensity, fine-grained intensity and intensity directed towards persons. To get a sense of perspective regarding each expression they use the feature of Sentiment Views and for detecting emotion they use the Emotion Categories feature along with the NRC lexicon. They implement Noun Patters and Adjective Patterns along with a certain amount of posts from Twitter in order to find negative and abusive content based on those patterns. To create a semantic representation of each word, the authors utilize the WordNet and the Wiktionary sources and they also use supersenses that categorize words into classes. The last feature of the linguistic features category is FrameNet. For the word embeddings feature the authors incorporate the Word2Vec approach to the AMZ and WAC corpora.

Question 2: Report statistics.

Category	Percentage
Abusive	33.3%
Non-Abusive	66.6%

Table 1: Answer2: Base Lexicon: Percentages of abusive and non-abusive words.

Category	Percentage
Abusive noun	45.26%
Non-Abusive noun	54.74%
Abusive verb	17.82%
Non-Abusive verb	82.18%
Abusive adjective	33.86%
Non-Abusive adjective	66.14%

Table 2: Answer2: Base Lexicon: Percentages of abusive and non-abusive words based on their part of speech category.

Category-Rank	Percentage
4.0 to 3.0	0.13%
3.0 to 2.0	1.07%
2.0 to 1.0	5.47%
1.0 to 0.0	28.58%
0.0 to -1.0	24.46%
-1.0 to -2.0	8.59%
-2.0 to -3.0	1.27%
-3.0 to -4.0	0.07%
-4.0 to -5.0	0.02%

Table 3: Answer2: Expanded Lexicon: Percentages based on the ranks of the words.

POS	4.0 to 2.0	2.0 to 0.0	0.0 to -2.0	-2.0 to -5.0
Noun	66.67%	64.05%	36.84%	34.56%
Adjective	23.53%	27.81%	37.29%	24.82%
Verb	9.80%	8.14%	25.88%	40.62%

Table 4: Answer2: Percentages of each part of speech on the expanded lexicon.

Question 3: Explain all categories and give examples found in the lexicon.
As can be seen in the tables 1,3 above, the two lexicons (base lexicon and expanded lexicon) have different categories in which they split each word. Starting with the base lexicon, there are two categories: *Abusive* and *Non-Abusive*. For a word to be labeled as abusive, at least 4 out of 5 raters had to label it as abusive, otherwise the word would be labeled as non-abusive. The expanded lexicon uses the confidence score of the classifier as the score for each of the words. Here the classifier is trained on the feature-specific approaches proposed in the paper. In the expanded lexicon the words with the highest score are considered abusive and as the score is getting lower then the words are considered more and more non-abusive.

It is important to note that in both lexicons, each entry/word is accompanied by its own part of speech tag, namely noun, adjective and verb something that is particularly helpful for calculating statistics on the lexicons.

Question 4: Give a representative sample (10 to 20 entries) with all information, and discuss the quality.
The following tables display some examples entries for both lexicons (the base lexicon and the expanded lexicon). In the base lexicon a more simplistic approach is followed given that there are only two classes that each word can be labeled (abusive and non-abusive). On the other hand, in the expanded lexicon, each word is assigned with a score and thus there can be a more clear and precise categorization of each word taking into account the word's severity. The numerical ratings in the expanded lexicon can aid the research for more robust analyses on the data and can help gain a deeper insight on the data. Apart from that, having the part of speech near each word encourages for a more contextual understanding of each word and can help distinguish each word that might have multiple

meanings or multiple syntactic roles. Furthermore, knowing the part of speech of each word can avoid misunderstanding and false positives when it comes to the offensive words. For example a word might have a positive meaning in a specific form (as a verb) and negative in another form (as a noun). Such a word can be the word "drink" where as a verb means the action of consuming a beverage ,whereas as a noun it can refer to some alcoholic beverages.

Entry	Category
bitch_noun	TRUE
blame_noun	FALSE
complacent_adj	TRUE
danger_noun	FALSE
defeat_noun	FALSE
demonize_verb	TRUE
diabolic_adj	TRUE
growl_verb	FALSE
ignorant_adj	TRUE
negro_noun	TRUE

Table 5: Answer4: Base Lexicon: Example entries.

Entry	Score
disgusting_adj	3.4936825
idiot_noun	3.0301171
nerd_noun	2.4511875
vermin_noun	2.2237203
prank_noun	2.0789004
fart_verb	1.9660311
inexcusable_adj	-0.53259404
empty_noun	-1.3802626
cling_verb	-2.2601736
demand_noun	-2.9540978

Table 6: Answer4: Expanded Lexicon: Example entries.

Question 5: Address issues that you find relevant for the quality, consistency and/or coverage of the lexicon.

Although there are many benefits using lexicons like those mentioned in the paper, there are also some challenges and drawbacks that need to be addressed in order to improve any future attempt to create more lexicons.

1. Sometimes it is extremely hard to differentiate between the different senses of some words. The challenge is that some words might have multiple meanings and assigning a label-score to a word automatically deprives those words from being analyzed properly and more extensively.

2. One other apparent issue is that there is always a very narrow group of words being analyzed and labeled and thus leaving a vast amount of words out of the analysis.

3. During the creation of a lexicon there can be some bias lurking which can completely spoil the quality of the whole survey.

1.2 Hurtlex

Question 1: What is the motivation for the paper? Describe how these lexicons are built?
The main objective of the paper is the development of a lexicon of hate words that can be used as a resource to analyze and identify hate speech in social media texts in a multilingual perspective. The starting point for the development of this lexicon is the Italian hate lexicon "Le parole per ferire" or

“words to hurt” developed by the Italian linguist Tullio De Mauro for the ”Joe Cox” Committee on intolerance, xenophobia, racism, and hate phenomena of the Italian Chamber of Deputies.

The paper describes the steps involved in building the lexicon. The first step consisted of extracting every item from the Italian hate lexicon ”Le parole per ferire” or “words to hurt”. This lexicon includes more than 1,000 Italian words from 3 macro-categories: derogatory words (all those words that have a clearly offensive and negative value, e.g. slurs), words bearing stereotypes (typically hurting individuals or groups belonging to vulnerable categories) and words that are neutral, but which can be used to be derogatory in certain contexts through semantic shift (such as metaphor). The lexicon is divided into 17 finer-grained, more specific sub-categories that aim at capturing the context of each word.

The first step in building HurtLex was to extract every item from the Italian hate lexicon Le parole per ferire. The extracted items were then expanded through the link to available synset-based computational lexical resources such as MultiWordNet and BabelNet, and evolved in a multilingual perspective by semi-automatic translation and expert annotation.

The second step involved using MultiWordNet to augment the words with their part-of-speech tags. The Italian index of MultiWordNet was used, comprising, for each lemma, four fields containing the identifiers of the synsets in which the lemma is intended like a noun, an adjective, a verb, and a pronoun. By joining this index with the lexicon, all the possible part-of-speech for 59.2% of the lemmas were obtained, bringing the total number of lemmas from 1,072 to 1,156 to include duplicates with different parts of speech. The remaining lemmas were annotated manually.

The third step consisted of linking the lemmas of the lexicon with a definition. The BabelNet API was used to retrieve the definitions, aiming for high coverage. In total, a definition was retrieved for 71.1% of the lemmas.

Finally, the lexicon was evaluated as a resource for hate speech detection in social media.

Question 2: Report statistics.

Category	**Percentage**	**Category**	**Percentage**
PS	3.85%	ASM	7.07%
RCI	0.81%	ASF	2.78%
PA	7.52%	PR	5.01%
DDF	2.06%	OM	2.78%
DDP	6.00%	QAS	7.34%
DMC	6.98%	CDS	26.68%
IS	1.52%	RE	3.31%
OR	1.52%	SVP	4.83%
AN	9.94%		

Table 7: Answer2: Hurtlex: Distribution of categories and their percentages.

Question 3: Explain all categories and give examples found in the lexicon.
The HurtLex lexicon is divided into 3 macro-categories and 17 finer-grained, more specific sub-categories that aim at capturing the context of each word. Here are the categories and some examples of words found in each category:
1. Derogatory words/ Negative stereotypes (all those words that have a clearly offensive and negative value, e.g.slurs) 1.1. Ethnic slurs (PS): This category includes derogatory terms and slurs targeting specific ethnic or racial groups. Examples include racial slurs like ”nigger” and ”chink.”

1.1.2. Locations and demonyms (RCI): This category includes derogatory terms related to specific locations or regional identities. Examples include terms like ”hillbilly” and ”redneck.”

1.3. Professions and occupations (PA): This category includes derogatory terms targeting specific professions or occupations. Examples include terms like ”hooker” and ”janitor.”

1.4. Physical disabilities and diversity (DDF): This category includes derogatory terms related to physical disabilities or physical appearance. Examples include terms like ”cripple” and ”freak.”

1.5. Cognitive disabilities and diversity (DDP): This category includes derogatory terms related to cognitive disabilities or mental health. Examples include terms like ”retard” and ”psycho.”

1.6. Moral and behavioral defects (DMC): This category includes derogatory terms related to moral or behavioral flaws. Examples include terms like ”scumbag” and ”pervert.”

1.7. Words related to social and economic disadvantage (IS): This category includes derogatory terms related to social or economic disadvantage. Examples include terms like "welfare queen" and "bum."

2. Bearing stereotypes/Hate words and slurs beyond stereotypes((typically hurting individuals or groups belonging to vulnerable categories). 2.1 plants (OR): This category includes derogatory terms and slurs that go beyond stereotypes and target various aspects of identity. Examples include terms like "faggot" and "dyke."

2.2. Animals (AN): This category includes derogatory terms comparing individuals to animals in a derogatory manner. Examples include terms like "pig" and "bitch."

2.3. Male genitalia (ASM): This category includes derogatory terms related to male genitalia. Examples include terms like "dick" and "cock."

2.4. Female genitalia (ASF): This category includes derogatory terms related to female genitalia. Examples include terms like "cunt" and "slut."

2.5. Words related to prostitution (PR): This category includes derogatory terms related to prostitution. Examples include terms like "whore" and "prostitute."

2.6. Words related to homosexuality (OM): This category includes derogatory terms related to homosexuality. Examples include terms like "fag" and "queer."

3. Words that are neutral but which can be used to be derogatory in certain contexts through semantic shift/Other words and insults(such as metaphor)

3.1. Descriptive words with potential negative connotations (QAS): This category includes words that have negative connotations and can be used as insults. Examples include terms like "ugly" and "stupid."

3.2. Derogatory words (CDS): This category includes words that are generally considered derogatory and offensive. Examples include terms like "bastard" and "idiot."

3.3. Felonies and words related to crime and immoral behavior (RE): This category includes words related to criminal or immoral behavior. Examples include terms like "thief" and "murderer."

3.4. Words related to the seven deadly sins of the Christian tradition (SVP): This category includes words related to the seven deadly sins of the Christian tradition. Examples include terms like "greed" and "lust."

Question 4: Give a representative sample (10 to 20 entries) with all information, and discuss the quality.

id	pos	category	stereotype	lemma	level
EN1382	n	qas	no	gag reel	inclusive
EN7077	a	cds	no	snotty	conservative
EN6856	n	is	yes	mendicant	conservative
EN5485	n	re	no	maffias	conservative
EN5024	n	cds	no	lying in trade	conservative
EN6950	n	re	no	yeargh	inclusive
EN204	n	om	no	buttfucker	inclusive
EN1575	n	qas	no	sir osthara	inclusive
EN206	n	om	no	assplay	inclusive
EN858	n	an	no	sucker	inclusive
EN1975	n	re	no	slanderous	inclusive
EN2838	n	or	no	cucumis sativus	inclusive
EN2454	n	or	no	homophobic slurs	conservative
EN552	n	an	no	jennies	inclusive
EN1107	n	re	no	knave	conservative
EN5750	n	svp	no	pride	inclusive
EN5436	n	cds	no	nerdiness	conservative
EN2800	n	qas	no	conform	inclusive
EN523	n	asm	no	putz	conservative

Table 8: Sample Lexicon Entries

Quality :

1. The lexicon seems to take a comprehensive approach, encompassing a diverse range of words, from potentially harmful terms to seemingly neutral ones.
2. Each entry provides a clear categorization, which can help users quickly understand the context or potential connotations associated with each term.
3. The inclusion of "level" (inclusive or conservative) adds another layer of context. While "inclusive" might suggest a term is generally accepted or harmless, "conservative" hints at traditional or potentially negative connotations.
4. The lexicon seems to be updated to reflect modern sensibilities, but as always with language, connotations can be subjective, and the acceptability of terms can change over time.
5. There might be room for further clarification on certain terms or the inclusion of more context, especially for entries that might seem out of place or have multiple interpretations.

Question 5: Address issues that you find relevant for the quality, consistency and/or coverage of the lexicon.

1. Diverse Category Distinctions: The lexicon spans across a wide array of categories from "qas" (descriptive words with potential negative connotations) to "om" (words related to homosexuality). The broad nature of these categories might make it challenging to maintain consistency. A more granulated categorization might provide clarity, but could also complicate the lexicon's usability.
2. Stereotype Classification: The column that indicates whether a word is a stereotype or not is binary ("yes" or "no"). This simplification might be an oversimplification. Stereotypes can exist on a spectrum, from mild to severe. A more nuanced classification could provide a more accurate representation.
3. Lemma Complexity: Some lemmas, like "cucumis sativus" (the scientific name for cucumber) or "lying in trade," seem out of place or overly complex. This might reflect inconsistencies in how words or phrases are selected for inclusion.
4. Subjectivity of the "Level" Classification: The lexicon classifies words as "inclusive" or "conservative". The basis for this classification isn't clear from the provided data. Different cultures or regions might view the same term differently, so it's crucial to define what "inclusive" and "conservative" mean in the context of this lexicon.
5. Coverage of Words: Given that a lexicon can't possibly cover every derogatory term or phrase, especially as language evolves, it's essential to ensure regular updates. Additionally, there might be words in the lexicon that lose their negative connotations over time and vice versa.
6. Multilingual Considerations: If this lexicon is intended for use across multiple languages or cultures, it's crucial to consider how these words translate and whether their connotations remain the same across languages.
7. Context Sensitivity: Words might be derogatory in one context but neutral or even positive in another. The lexicon does not seem to account for context, which could be a limitation in its applicability.
8. POS Tagging: The part-of-speech (POS) tags might not capture the full range of ways a word can be used. For example, some nouns can also be used as verbs, and this might change their connotations.
9. Potential Biases: Any lexicon can unintentionally perpetuate biases, especially when it involves subjective judgment. It's essential to consider whether the lexicon was curated with input from diverse groups or whether it might inadvertently support biased viewpoints.

1.3 MOL

Question 1: What is the motivation for the paper? Describe how these lexicons are built?
The MOL lexicon used in the proposed approach for detecting offensive language and hate speech on social media was extracted from the HateBR corpus. The terms or expressions in the MOL lexicon were annotated by three different annotators, resulting in a high human agreement score (73% Kappa). The words were categorized in hate groups and then given a label which classifies the words as context independent or context dependent. Words can belong to any POS tags. Additionally, the MOL lexicon also includes implicit content extracted using "clue terms or expressions." For example, the expression "voltar para a jaula" ("go back to the cage") serves as a clue expression to identify the implicit offensive term "ladrão" ("thief").Originally created in Portuguese, the MOL has been meticulously translated into five other languages: English, Spanish, French, German, and Turkish.

Question 2: Report statistics.

Table 9: Contextual Information vs. Hate Groups

Contextual Information	**Hate Groups**	**Total**
Contextual-independent offensive	1	612
Contextual-dependent offensive	0	387
Total		1,000

Table 9 presents data on hate group classification based on contextual information. It categorizes hate groups into "Contextual-independent offensive" and "Contextual-dependent offensive," indicating the number of instances (label) in each category. The table also provides the total count of instances, which sums up to 1,000.

Table 10: Hate Groups Distribution

Class	**Total**
no-hate	864
partyism	69
sexism	35
homophobia	16
fatphobia	9
religious intolerance	9
antisemitism	1
apology for the dictatorship	5
racism	4
antisemitism	3
Total	1,000

Table 10 provides a breakdown of the hate groups into different categories. It lists the various hate group classes, such as "no-hate," "partyism," "sexism," and others, along with the total count of instances for each class. The table shows that there are 1,000 instances in total, distributed across these different hate group categories.

Question 3: Explain all categories and give examples found in the lexicon.
The "MOL" dataset, has various terms and expressions in multiple languages, along with contextual labels and hate speech labels in table 11. It seems that the dataset is designed to categorize these terms and expressions based on their explicitness, implicitness, and potential for hate speech. MOL was originally written in Portuguese and manually translated by native speakers in English, Spanish, French, German and Turkish. Below, some explanations of the categories and examples found in the lexicon:

Table 11: Dataset Columns

Column Name	Description	Example Text
term-or-expression	Term or Expression	"term," "expression," etc.
explicit-or-implicit	Explicit or Implicit	1 (explicit), 0 (implicit)
pt-brazilian-portuguese	Brazilian Portuguese Translation	"chorume," "baixaria," etc.
pt-contextual-label	PT Contextual Label	1 (contextual label), 0 (not contextual)
pt-hate-label	PT Hate Label	0 (not a hate label), 1 (contains hate speech)
pt-deeply-culture-rooted	PT Deeply Culture Rooted	0 (not deeply culture-rooted), 1 (is deeply culture-rooted)
en-american-english	American English Translation	"rotten," "fuckfest," etc.
en-contextual-label	EN Contextual Label	1 (contextual label), 0 (not contextual)
en-hate-label	EN Hate Label	0 (not a hate label), 1 (contains hate speech)
Meaning Sources	Sources for Term/Expression Meaning	URL links to meaning sources

Question 4: Give a representative sample (10 to 20 entries) with all information, and discuss the quality.

The table 12 consists of columns with specific headers, including "term-or-expression," "explicit-or-implicit," "pt-brazilian-portuguese," "pt-contextual-label," "pt-hate-label," "pt-deeply-culture-rooted," "en-american-english," "en-contextual-label," "en-hate-label". Each row in the table represents an entry in the lexicon. The entries are terms or expressions, and the characteristics of these entries are described in the subsequent columns. There are variations in the nature of entries (term or expression), which might indicate a mix of single-word terms and multi-word expressions. Translations are provided consistently for both Brazilian Portuguese and American English, aiding in cross-lingual analysis. The dataset captures the contextual labeling of entries, particularly in English, signifying that some terms or expressions may acquire specific meanings within certain contexts. The presence of duplicate rows in the dataset may require deduplication to ensure data cleanliness. While some entries are labeled as hate speech in Portuguese, they are not similarly categorized in English, indicating potential variations in what is considered hate speech across languages. There are also alot of words which are not categorised in hate speech.

In summary, the dataset appears to be rich and informative, providing valuable insights into explicit and implicit language, cross-lingual translations, contextual labels, and the presence of hate speech. However, addressing duplicate rows and ensuring consistent categorization of hate speech across languages is essential to enhance the dataset's quality and utility.

Table 12: Sample entries

term-or-expression	explicit-or-implicit	pt-brazilian-portuguese	pt-contextual-label	pt-hate-label	pt-deeply-culture-rooted	en-american-english	en-contextual-label	en-hate-label
term	explicit	chorume	1	0	0	rotten	0	0
term	explicit	baixaria	1	0	0	fuckfest	1	0
term	explicit	vagabunda	1	sexism	0	slut	1	sexism
term	explicit	vagaba	1	sexism	0	slut	1	sexism
term	explicit	bolsootarios	1	partyism	1	0	1	partyism
expression	explicit	globolixo	1	0	1	0	1	0
term	explicit	molusco	0	partyism	1	mollusk	0	0
expression	explicit	idelogo vermelho	1	partyism	1	red ideologist	0	0
expression	explicit	bando de pelego	1	0	1	sheepskin	1	0
expression	explicit	babava ovo	1	0	0	ass-kissing	1	0
term	explicit	peppa	0	fatphobia	0	pig	0	fatphobia
term	explicit	pepinha	0	fatphobia	0	piggy	0	fatphobia
expression	explicit	intervenß£o militar	0	apology to dictatorship	0	government intervention	0	apology to dictatorship
expression	explicit	branco fazendo branquice	1	racism	0	white trash folks	1	racism
expression	explicit	judeu do inferno	1	antisemitism	0	jews from hell	1	antisemitism
term	explicit	bicha	1	homophobia	0	fag	1	homophobia
expression	explicit	bicha enrustida	1	homophobia	0	closet faggot	1	homophobia
expression	explicit	pastor do cao	1	religious intolerance	0	false shepherd	1	religious intolerance
term	explicit	macaca	0	racism	0	monkey	0	racism
expression	implicit	cana brava	1	0	1	0	1	0

Question 5: Address issues that you find relevant for the quality, consistency and/or coverage of the lexicon. The lexicon exhibits several issues and areas of improvement related to quality, consistency, and coverage:

1. There's inconsistency in the labeling of hate speech between Portuguese and English entries. For instance, "vagabunda" is labeled as containing sexism in Portuguese but not as a hate label in English. This inconsistency could affect the overall utility of the dataset for hate speech detection, as there should be alignment in categorization across languages.

2. Some terms are present in Portuguese but lack English translations. Ensuring complete and accurate translations for all terms is crucial for cross-lingual analysis and understanding.

3. Duplicate entries can be found in the sample, such as "vagabunda" and "vagaba," both with similar characteristics. Addressing and removing duplicate entries is essential for data cleanliness.

4. The dataset includes a column for deeply culture-rooted terms but provides little context for what this label signifies. It's important to define and clarify the criteria for designating terms as deeply culture-rooted.

5. To improve coverage and the overall quality of the lexicon, it would be valuable to include additional metadata such as the region or context in which these terms are commonly used, the level of offensiveness, or usage examples.

2 Merging the Lexicons

The task of merging the different lexicons into one common lexicon is considered very challenging due to the very nature of uniqueness that each lexicon possesses. The creators of each lexicon created those with very exact specifications in mind because in every experiment a different approach is followed and thus every lexicon is serving a different purpose. As a result, the process of merging the given lexicons into one 'global' lexicon demanded quite a lot of thinking and consideration on the decisions that we had to take.

Starting from the Expanded lexicon in [1] we observed that every entry in the lexicon was accompanied by its POS tag (e.g. "disgusting_adj") as well as its score regarding the offensive level of the word. In order to simplify the lexicon we separated each word from its POS tag and placed it into a different column. Moreover, for the score we decided to keep the binary convention of 0 and 1 where 1 is offensive and 0 is non-offensive. For that reason we decided to assign the value of 1(offensive) to those words that had a score of over 0 and the value 0(non-offensive) to those words that had a value below 0. As a result, the modified lexicon had 3 columns: the word, the POS tag and the binary value indicating whether it is offensive or not.

For the lexicon in [2] we modified the 'pos' column so it contains the same naming convention as the Expanded lexicon, meaning that we mapped the characters: n, a, v, av into "NOUN", "ADJ", "VERB" and "ADV". Moreover, because the lexicon contains only offensive words, we added one more column to the lexicon that contains the binary number 1 for every entry, indicating that every word is considered offensive.

With regards to the third lexicon in [3], we took into account the different categories mentioned in the lexicon (e.g. "sexism", "partyism" etc.) and we created a new column with a binary indication whether the word is offensive or not. More specifically, if a word has 0 as its "en-hate-label" then it is not considered offensive and thus a 0 is added in the binary offensive indication column. Conversely, if the word belongs to a specific category then the binary number 1 is given to that word, indicating that it is being categorized as an offensive word. As a last step, in order to introduce the POS tags for each word of the lexicon, we made use of the libraries "punkt", "averaged_perceptron_tagger" and "universal_tagset" of the NLTK suite [4].

The final merged lexicon has 3 columns "lemma", "POS" and "off_nonoff" for indicating if the word is offensive or non-offensive. Some of the lexicon's entries are as indicated below in the table 13. Some statistics extracted from the merged lexicon are as shown in the tables 13,15.

Overall, the creation of the merged lexicon was a meticulous process due to the unique nature and characteristics of each of the lexicons that we had to work with. Recognizing that each lexicon has a different goal and purpose, we carefully considered of a gentle way of alleviating the differences between them and simplify them enough in order to have the data in a similar enough form without losing any quality of the information. This ensured that we could bring the lexicons on the same level regarding the information it contains and thus making the process of merging easier.

Lemma	**POS**	**off_nonoff**
raven	VERB	1
phony	NOUN	1
fake	ADJ	1
illiterate	ADJ	1
twat	NOUN	1
skunk	NOUN	1
retard	NOUN	1
negroe	NOUN	1
symptom	NOUN	0
dark	NOUN	0
entangled	ADJ	0
react	VERB	0

Table 13: Merged lexicon example entries

Category	Percentage
Offensive	65.66%
Non-Offensive	34.33%

Table 14: Offensive and Non-offensive percentages in the merged lexicon.

POS tag	Percentage
Noun	67.81%
Verb	11.18%
Adjective	20.77%
Adverb	0.22%

Table 15: Percentages of the POS tags in the merged lexicon.

3 Use the lexicons for automatic hate speech identification

3.1 Classification Report

Based on table 16, the Wiegand expanded lexicon appears to be particularly adept at predicting instances of non-offensive messages, demonstrating a high degree of confidence both in terms of precision and recall. For offensive messages, the high precision, suggest that when the model flags a message as offensive, it's often correct. However, the low recall reveals a significant challenge: the model is missing a substantial number of actual offensive instances. This means that many offensive messages could go unnoticed, potentially undermining the primary objective of identifying and addressing inappropriate content.

Table 16: Analysis of Wiegand Expanded Lexicon

	Precision	Recall	F1-Score
0	0.77	0.96	0.86
1	0.72	0.27	0.39

The model's performance, as presented in table 17 for the Hurtlex lexicon, shows that the model exhibits decent recall for offensive text and good precision for non-offensive text. However, it struggles with high recall for non-offensive messages and high precision for offensive texts. This indicates a tendency to misclassify some instances, especially when predicting offensive messages. A lower f1-score for both offensive and non-offensive text suggests there is considerable room for improvement in achieving a better balance between precision and recall for both classes.

Table 17: Analysis of Hurtlex Lexicon

	Precision	Recall	F1-Score
0	0.81	0.57	0.67
1	0.37	0.64	0.47

In the MOL lexicon, as indicated by table 18, the model clearly demonstrates a robust capability to identify non-offensive texts. However, it grapples notably when it comes to pinpointing offensive texts with accuracy. While the model's precision for categorizing offensive content is moderate, suggesting that the instances it does identify as offensive are often correctly labeled, its recall stands out as a significant weakness. This low recall rate implies that the model fails to detect a substantial portion of truly offensive content, potentially allowing harmful messages to pass through undetected.

Table 18: Analysis of MOL Lexicon

	Precision	Recall	F1-Score
0	0.74	0.97	0.84
1	0.56	0.10	0.16

Based on table 19, the model demonstrates reasonable proficiency in discerning non-offensive texts. However, it faces difficulties when it comes to accurately labeling offensive texts. The relatively low precision for offensive text might result in numerous non-offensive texts being falsely flagged as inappropriate. The model's recall rate for offensive texts, although better than its precision, still implies that almost half of the truly offensive content could go undetected. This raises concerns, especially if the aim is to comprehensively identify and potentially filter out offensive material.

Table 19: Analysis of Merged Lexicon

	Precision	Recall	F1-Score
0	0.78	0.67	0.72
1	0.38	0.51	0.43

Based on table 20, The Wiegand Expanded Lexicon seems to be the most effective in terms of both accuracy and balancing precision with recall. The Hurtlex Lexicon, while close in accuracy, struggles a bit more with the balance of precision and recall. MOL and Merged Lexicons have lower accuracies and moderate F1-scores, indicating there's significant room for improvement for these lexicons in classifying content effectively.

Table 20: Performance Metrics: Accuracy and Macro F1 across Various Lexicons.

Lexicon	Accuracy	Marco f1
Wiegand Expanded Lexicon	0.76	0.62
Hurtlex Lexicon	0.73	0.50
MOL Lexicon	0.59	0.57
Merged Lexicon	0.62	0.58

3.2 Confusion Matrix

According to Based on table 21, the Wiegand expanded lexicon performs well in identifying instances of non offensive messages with few false positives. However, it struggles to identify offensive messages correctly, resulting in a significant number of false negatives. This suggests that the model is more Conservative when predicting offensive messages.

Table 21: Confusion Matrix of Wiegand Expanded Lexicon

	Positive	Negative
Positive	595	25
Negative	176	64

3.3 Qualitative Error Analysis

In the paper [5], the authors identified several classes of errors leading to false negatives. These include doubtful labels, toxicity without the presence of swear words, and sarcasm. For false positives, errors were often triggered by the use of certain words or phrases that, while potentially flagged as toxic or hateful, were not necessarily so in the given context.

To mitigate these errors, several strategies can be employed. Firstly, incorporating more contextual information, such as the author's intent or the surrounding conversation thread, can significantly assist models in better understanding the meaning and tone of a comment. Secondly, exploring new features, like sentiment analysis and topic modeling, can help capture the more nuanced aspects of language

use, pushing beyond simple word-based flagging. Finally, utilizing more diverse training data can be beneficial. By exposing models to a broader range of linguistic expressions from various contexts and communities, they become more proficient at distinguishing genuine toxicity from benign language use.

4 Short Conclusion

The process of consolidating diverse lexicons into a singular global lexicon was intricate, given the distinct objectives and nuances each individual lexicon presented. Through meticulous modification, such as standardizing POS tags and introducing binary categorizations for offensiveness, a unified lexicon was achieved without compromising the original integrity of the data. This endeavor highlighted the importance of careful consideration and data transformation when merging distinct datasets to ensure accuracy and consistency. The Wiegand Expanded Lexicon emerges as the most proficient lexicon, displaying high precision and recall, especially for non-offensive messages. While other lexicons, such as the Hurtlex and MOL, display commendable strengths in certain areas, they also possess marked weaknesses, like imbalances in precision and recall. The consistent performance of the Wiegand Expanded Lexicon suggests that it offers the most comprehensive approach in distinguishing offensive from non-offensive content. However, the challenge of ensuring false negatives are minimized, remains across all lexicons, emphasizing the continued need for refinement in the models.

References

[1] Michael Wiegand, Josef Ruppenhofer, Anna Schmidt, and Clayton Greenberg. Inducing a lexicon of abusive words–a feature-based approach. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1046–1056, 2018.

[2] Elisa Bassignana, Valerio Basile, Viviana Patti, et al. Hurtlex: A multilingual lexicon of words to hurt. In *CEUR Workshop proceedings*, volume 2253, pages 1–6. CEUR-WS, 2018.

[3] Francielle Vargas, Fabiana Rodrigues de Góes, Isabelle Carvalho, Fabrício Benevenuto, and Thiago Alexandre Salgueiro Pardo. Contextual-lexicon approach for abusive language detection. *arXiv preprint arXiv:2104.12265*, 2021.

[4] Steven Bird, Ewan Klein, and Edward Loper. *Natural language processing with Python: analyzing text with the natural language toolkit.* " O'Reilly Media, Inc.", 2009.

[5] Betty van Aken, Julian Risch, Ralf Krestel, and Alexander Löser. Challenges for toxic comment classification: An in-depth error analysis. In *Proceedings of the Second Workshop on Abusive Language Online (ALW2)*, pages 33–42. Association for Computational Linguistics, 2018.