Toxicity : Detection, Classification and Reduction

Dhruvin Sureshbhai Donda(1169998)[1], Joy Nikhil Christian(1147981)[1] and Niharika B Sojitra(1170232)[1]

[1]Lakehead University, Thunderbay, ON

COMP 5800 Project

April 11, 2022

Introduction

The social media has become a popular medium for communication. However, this has also led to a rise in online abuse, harassment and threats, which has become a significant issue to solve.

To solve above issue, we aim to detect, categorize, and convert such comments into neutral or non-toxic.

Literature Review

Toxicology detection has received a lot of attention over the recent years. Hate speech was first introduced to shine the issue of racial abuse, and its social, political and psychological effects on individuals [3].

The literature review has been split into two components :

- Dataset.
- Techniques.

Dataset

Toxicity Detection :

- Hate Speech and Offensive Language Dataset [2]
- Civil Comments, Jigsaw Toxic Comments Classification Dataset [8]
- OLID (Offensive Language Identification Dataset) [9]
- SOLID (Semi-Supervised Offensive Language Identification Dataset) [5]
- RealToxicityPrompts [6]

Detoxification :

- Parallel detoxification dataset [4]
- ParaDetox dataset [7]

Techniques

Classification :

There are few renowned competitions that helped in advancing this research further (e.g., ALW1[a], TA-COS[b],SemEval-2019[c], 2020[d] and 2021[e] and GermEval[f]). In SemEval-2019, two tasks were on abusive language detection, Task 5, 'hatEval: Multilingual detection of hate speech against immigrants and women in Twitter' [1] and Task 6 was on OLID dataset, 'OffensEval: Identifying and Categorizing Offensive Language in Social Media' [10].

For Task 5, SVM model with RBF kernel obtained the highest result at macro-averaged F1-score of 0.651. BERT based models were used by the top 10 rank holders for task 6, surpassing every other models [10].

[a] https://sites.google.com/site/abusivelanguageworkshop2017/
[b] http://ta-cos.org/
[c] https://alt.qcri.org/semeval2019/index.php?id=tasks
[d] https://alt.qcri.org/semeval2020/index.php?id=tasks
[e] https://semeval.github.io/SemEval2021/tasks
[f] https://goo.gl/uZEerk

Techniques

Detoxification :

Two novel unsupervised methods were proposed for detoxification [6].

The first method, ‘ParaGeDi’ was combination of two ideas: (1) the use of small style-conditional language models to guide the generation process, and (2) the use of paraphrasing models for style transfer.

The second technique, ”CondBERT” employs BERT to substitute toxic words with their less offensive alternatives.

Although ParaGeDi exhibited better performance than all other models, including CondBERT, with an accuracy of 0.81.

Data Description

For the study "Jigsaw Dataset" used that was derived from the 'Kaggle toxic comment classification challenge. The dataset includes a large number of Wikipedia comments that have been labeled as toxic or non-toxic and further classified by human raters into six types of toxic behavior, namely toxic, severe toxic, identity hate, obscene, insult, and threat.

Dataset contains total 1.8 million multilingual comments. Only 158k comments are English in entire dataset from which only 15k are toxic; the rest are non toxic.

	id	comment_text	Toxic	Severe Toxic	Obscene	Threat	Insult	Identity Hate
0	0000997932d777bf	Explanation\nWhy the edits made under my usern...	0	0	0	0	0	0
1	000103f0d9cfb60f	D'aww! He matches this background colour I'm s...	0	0	0	0	0	0
2	000113f07ec002fd	Hey man, I'm really not trying to edit war. It...	0	0	0	0	0	0
3	0001b41b1c6bb37e	"\nMore\nI can't make any real suggestions on ...	0	0	0	0	0	0
4	0001d958c54c6e35	You, sir, are my hero. Any chance you remember...	0	0	0	0	0	0

Figure: Jigsaw Dataset

Data Preprocessing

The data is highly imbalanced according to toxicity and its labels. As only 15k comments are only toxic, Only very few comments are falling under the "Threat" category.

To address this imbalance, the Random over-sampling method was used to add more comments to the toxic data.

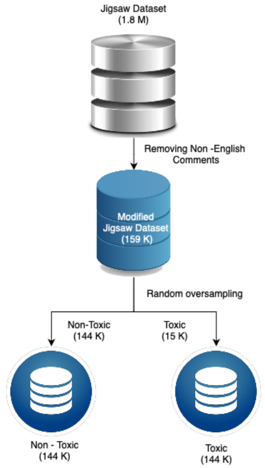


Figure: Data Pre-processing using Random Oversampling

Data Preprocessing

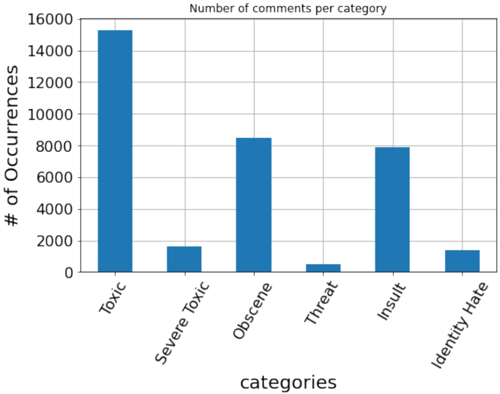


Figure: Toxic comments counts according to its labels

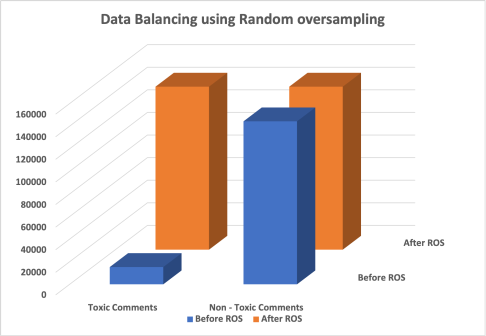


Figure: Balancing the data using Random oversampling(ROS)

Methods

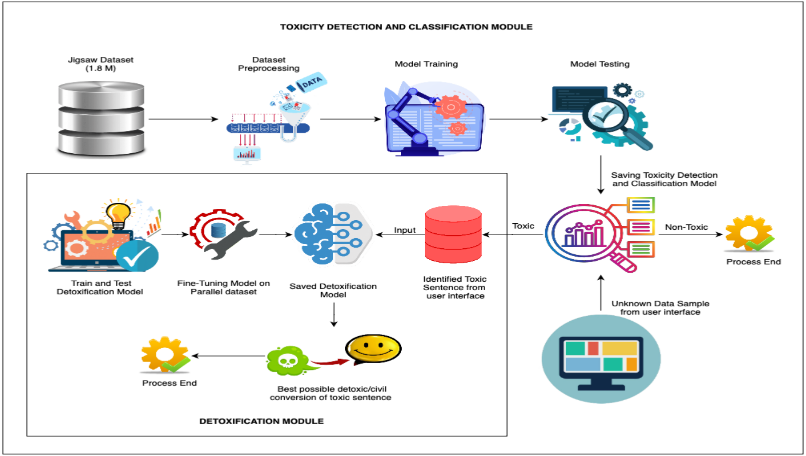


Figure: Flow diagram for methodology

Toxicity Detection and Classification

Machine Learning Approaches :

Various baseline machine learning models like Naïve Bayes, Linear Support Vector Classifier, and Logistic Regression, were utilized for multi-label classification.

OneVSRestClassifier used to compare performance of models. Label-wise accuracy for each classifier was calculated for comparison.

Toxicity Detection and Classification

Neural Net Approaches :

To overcome deficiency of machine learning models, neural net classifiers were employed, such as Recurrent Neural Network (RNN) and Bi-LSTM.

'BERT' and 'DistilBERT' - pre-trained Bi-directional Transformers were also implemented for Language Understanding.

Moreover, we have added a publicly available grammar correction library called 'Caribe'[1] to further enhance the grammar and performance.

[1] https://pypi.org/project/Caribe/

Detoxification

For detoxification, we implimented model ‘SED-T5’, ‘CAE-T5’, ‘CondBert’, ‘ParaGeDi’, and our own model, ‘t5-paraDetox’.

The most effective outcome was achieved through the use of ‘t5-paraDetox’, which is a modified version of the publicly accessible Huggingface model called ‘t5-paranmt-detox’.

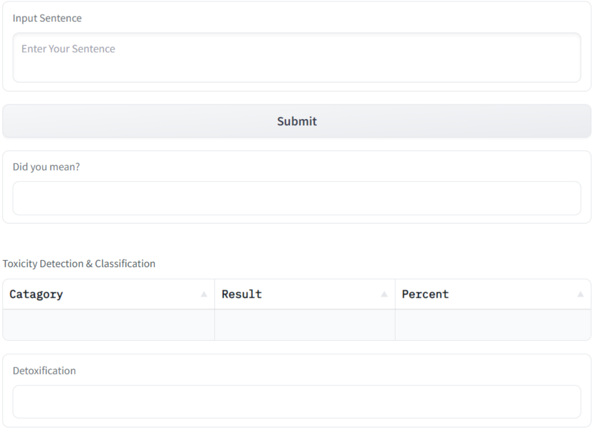
Model is fine-tuned on the Parallel dataset.

	input_text	target_text	prefix
0	. or the loud ass one - thousand ton beast roa...	or the loud one - thousand ton beast roaring ...	paraphrase
1	" mandated " and " right fucking now " would b...	"Mandated' and "right now" would be good.	paraphrase
2	" mandated " and " right fucking now " would b...	"mandated" and" right away" would be good	paraphrase
3	" mandated " and " right fucking now " would b...	mandated and right would be good	paraphrase
4	* neither * of my coworkers gave a shit when i...	Neither of my co-workers cared when it came to...	paraphrase

Figure: Parallel Dataset

User Interface

This user interface was also served as a final deployment for this entire study to produce the results. To create an interface, a publicly available API called Gradio was utilised.



Results: Toxicity Detection and classification

	Model	Precision	Recall	F1 Score	Accuracy
ML Models(Average of Lable-wise Results)	**NB**	0.99	0.11	0.19	0.8986
	LR	0.89	0.52	0.66	0.9181
	SVC	0.84	0.62	0.72	0.9197
Neural Net Models	**RNN**	0.7227	0.6144	0.64	0.8487
	Bi-LSTM	0.7774	0.7343	0.7553	0.7589
	BERT	0.8257	**0.8346**	0.8301	0.9402
	DistilBERT	**0.8497**	0.8164	**0.8318**	**0.9434**

Table: Evaluation metric for Machine Learning, Neural nets and Language models for classification

Results: Toxicity Reduction

Model	Output Text	Δ BLEU	Δ Cosine Similarity
ParaGEDI	I've been a big fan of shootings in the school.	-0.3976	-0.2743
CondBERT	i am " " " of this school shootings by " people.	0.4788	0.2092
SED_T5	I am tired of this school shootings by black people.	0.4902	0.3975
Human Response	The frequent school shootings are getting to me.	-0.443	-0.1813
t5-paraDetox	I'm tired of this school shootings.	**0.1547**	**0.0803**

Table: Evaluation metric for each Detoxification models for toxic input text : "I am tired of this school shootings by black people."

Results: User Interface

Input Sentence

I will kill you.

Submit

Did you mean?

I will kill you.

Toxicity Detection & Classification

Catagory	Result	Percent
Toxic	true	83.26
Severe Toxic	true	13.78
Obscene	true	12.52
Threat	true	94.94
Insult	true	10.01
Identity Hate	false	5.85

Detoxification

I'll take you.

Figure: Result for an User Interface when applied toxic sentence.

Limitations and Future Work

Limitations

1) Data imbalance.
2) Limited resources for text detoxification.

Future Work

1) Conversion of all toxic comments in Jigsaw dataset into civil.
2) Employing advance balancing techniques to dataset.
3) Converting multiple toxic sentences into civil.

Conclusion

In summary, it can be concluded that neural networks specifically, BERT based models performed far better to achieve desired results in detecting and classifying the toxic comments while our model 't5-paraDetox' gave the best civil version of a toxic sentence.

References I

[1] Valerio Basile et al. "SemEval-2019 Task 5: Multilingual Detection of Hate Speech Against Immigrants and Women in Twitter". In: *Proceedings of the 13th International Workshop on Semantic Evaluation*. Minneapolis, Minnesota, USA: Association for Computational Linguistics, June 2019, pp. 54–63. DOI: 10.18653/v1/S19-2007. URL: https://aclanthology.org/S19-2007.

[2] Thomas Davidson et al. "Automated Hate Speech Detection and the Problem of Offensive Language". In: *Proceedings of the 11th International AAAI Conference on Web and Social Media*. ICWSM '17. Montreal, Canada, 2017, pp. 512–515.

[3] Richard Delgado. "Words That Wound: A Tort Action for Racial Insults, Epithets, and Name-Calling, 17 Harv". In: *L. Rev* 133 (1982).

References II

[4] Daryna Dementieva et al. "Crowdsourcing of Parallel Corpora: the Case of Style Transfer for Detoxification". In: *Proceedings of the 2nd Crowd Science Workshop: Trust, Ethics, and Excellence in Crowdsourced Data Management at Scale co-located with 47th International Conference on Very Large Data Bases (VLDB 2021 (https://vldb.org/2021/))*. Copenhagen, Denmark: CEUR Workshop Proceedings, 2021, pp. 35–49. URL: http://ceur-ws.org/Vol-2932/paper2.pdf.

[5] Samuel Gehman et al. "RealToxicityPrompts: Evaluating Neural Toxic Degeneration in Language Models". In: *Findings of the Association for Computational Linguistics: EMNLP 2020*. Online: Association for Computational Linguistics, Nov. 2020, pp. 3356–3369. DOI: 10.18653/v1/2020.findings-emnlp.301. URL: https://aclanthology.org/2020.findings-emnlp.301.

References III

[6] Samuel Gehman et al. "RealToxicityPrompts: Evaluating Neural Toxic Degeneration in Language Models". In: *Findings of the Association for Computational Linguistics: EMNLP 2020*. Online: Association for Computational Linguistics, Nov. 2020, pp. 3356–3369. DOI: 10.18653/v1/2020.findings-emnlp.301. URL: https://aclanthology.org/2020.findings-emnlp.301.

[7] Varvara Logacheva et al. "ParaDetox: Detoxification with Parallel Data". In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 6804–6818. DOI: 10.18653/v1/2022.acl-long.469. URL: https://aclanthology.org/2022.acl-long.469.

References IV

[8] JULIÁN PELLER. *Jigsaw Toxic Comment Classification Dataset*. URL: https://www.kaggle.com/competitions/jigsaw-toxic-comment-classification-challenge/discussion.

[9] Marcos Zampieri et al. "Predicting the Type and Target of Offensive Posts in Social Media". In: *CoRR* abs/1902.09666 (2019). arXiv: 1902.09666. URL: http://arxiv.org/abs/1902.09666.

[10] Marcos Zampieri et al. "SemEval-2019 Task 6: Identifying and Categorizing Offensive Language in Social Media (OffensEval)". In: *Proceedings of the 13th International Workshop on Semantic Evaluation*. Minneapolis, Minnesota, USA: Association for Computational Linguistics, June 2019, pp. 75–86. DOI: 10.18653/v1/S19-2010. URL: https://aclanthology.org/S19-2010.

Thank you