Grenze International Journal of Engineering and Technology, June Issue

Ontology based Approach for Reducing Toxicity and Maintaining Similarity

Uma Taru[1] and Archana Patil[2]
[1-2]College of Engineering, Pune, India
Email: taruur20.comp@coep.ac.in, abp.comp@coep.ac.in

***Abstract*—Many online social media platforms have particular community guidelines for comment sections. The platforms that maintain commentary sections in various posts, videos, and blogs need to adhere to these guidelines. These comment sections may have specific comments that fail to satisfy the rules and regulations to maintain societal norms of communication. These comments are classified as toxic comments. Google's Perspective API defines toxic comments as comments that are rude, offensive, and likely to make someone leave the conversation. In this paper, we propose a Detoxification module for toxic comments - in- put to which will be a toxic comment. We will get a sentence detoxified, having a similar meaning to the original sentence as output. This module consists of an Ontology that we have built for toxic words. As per our knowledge, this Ontology is the first ontology built for toxic words. This Ontology consists of toxic words and their antonyms and synonyms in in- creasing order of their toxicity levels. We can traverse this Ontology and find the best-suited word with less toxicity and similar meaning. We are experimenting with the Toxic Comment Classification Challenge dataset from the Kaggle competition which consists of 1,59,571 comments out of which we filtered 15,294 toxic comments. Our approach is relatively straightforward and simple - and is effective in reducing toxicity in online comments.**

***Index Terms*— Toxicity · Ontology · Detoxification.**

I. INTRODUCTION

Even when the thoughts expressed online are valid, well-intentioned, and beneficial, foul language might lead to a misunderstanding. The problem of abusive language identification has been addressed through Natural Language Processing (NLP) research [1]. In many countries, including Canada, the United Kingdom, and France, there are laws prohibiting hate speech. Both Facebook and Twitter have responded to criticism for not doing enough to prevent hate speech by instituting policies against it [2]. Building machine learning systems that can detect harmful content is a well- established and well-researched research subject in this regard, along with strong human moderator teams [3]. There are many systems that detect abusive language [29] and the field is gaining attention to detect unwelcome comments. There have been studies based on context-dependence and context independence of the comments [5]. (Disclaimer: Text in the examples could be considered profane, vulgar, or objectionable.)

A. Perspective API

Perspective API provides a probability score to a sentence based on its perceived impact in a conversation. Jigsaw and Google's Counter Abuse Technology team made Perspective as part of a Conversation-AI research

Grenze ID: 01.GIJET.8.2.45

project. Using millions of examples gathered from different online platforms and reviewed by human annotators, Perspective has been trained to recognize a range of qualities (e.g., whether a comment is toxic, threatening, offensive, off-topic, etc.). Comment with a 0.9 score is not necessarily more toxic than comment with a 0.5 toxicity score. It just implies that 9 out of 10 readers would find the first comment toxic. Perspective API is a CNN-based and context-insensitive model which uses much larger external labeled data. We use pre-trained toxicity classifier 'Detoxify' to get the toxicity scores of comments This is a multilingual model - ability to classify toxicity in various languages [30].

B. Ontology

An ontology is a machine-understandable semantic, which is an interface between software and human [8]. Definition of Ontology by Gruber, 1993 - An ontology is a specification of a representational vocabulary for a shared domain of discourse, which includes definitions of functions, relations, classes, and other objects. Anything that exists can be represented as an Ontology [9]. Ontologies are restricted to given domains and they have relevant concepts and relations in a model to particular applications [10]. An ontology is a set of terminology and relationships that can be used to model a domain [11]. The nodes are entities and the edges represent the relationships. In this paper, we have built an ontology for toxic words, where toxic words are nodes and we have described two relations 'synonym' and 'antonym' to its every toxic word's edges. This ontology can be traversed to get a synonym of the toxic word which has the least toxicity. Our ontology is created in the form of a python dictionary and stored in '.json' format. Detailed explanation of the ontology generation steps in the methodology chapter.

C. WordNet

WordNet is a massive electronic lexicon - a large electronic lexical database for English [13][14]. This lexical database links English adverbs, adjectives, verbs, nouns, and synonym groups through semantic relationships that establish word definitions. There are about 118,000 possible word forms and over 90,000 different word meanings in WordNet. Antonymy, Synonymy, Troponymy, Hyponymy, Meronymy and Entailment are some of the semantic relations found in WordNet.[14] We have used wordnet to get synonyms and antonyms for our toxic words. Wordnet gives synonyms and antonyms in every context. The context of the replaced word matters here. Wordnet gives all the possible synonyms from all contexts. And getting similar meaning sentences along with less toxicity is what we aim to get as output

D. Similaritty

Semantic Similarity between two sentences is assigning a similarity score to two sentences based on their meaning [15]. Just replacing the word from Ontology will not lead to a similar meaning sentence as previously discussed that we get synonyms from all the contexts. We have used transformer-based model that calculates the similarity score based on the attention mechanism. It captures the semantic properties of the words in the embeddings [16]. We used RoBERTa variation of BERT models that used transformers. RoBERTa has been trained on GLUE, RACE and SQuAD datasets to get state-of-art results [17].

II. Related Work

Pre-trained language models (LM) as GPT-2[18] can be used where the source and target languages are same. Style transfer can be thought of as one of the simplest ways to translate using unsupervised learning [19]. Another style transfer approach was Masked Language Modelling (LML) picks words based on style labels e.g., Mask and Infill [20]. There are end-to-end architectures, which encode sentences, after which hidden representations are manipulated to get a new style and then decode it. There are variations to this - representing style independent content [21] and others extract the representation of content and style from hid- den representations [22]. Raffel proposed a unified pre-trained bi-transformer that is applicable to any text-to-text, text classification and regressions [23]. There are many unsupervised approaches such as abstractive summarization. A strategy similar to back translation compressor-reconstructor [24] was used to train a model. The first paper to work on detoxification had done rewriting of offensive sentences into non-offensive sentences using encoder-decoder with Twitter and Reddit dataset was done by Nogueira dos Santos et al [25]. This model suffers from low fluency with offensive words [1]. Language models are conditioned
on style of generated text in controlled text generation [26]. Condition that the input's content should be pre-served is not satisfied in attribute transfer [1]. Toxicity is subjective and sometimes context-dependent [5] - what is considered toxic may differ across cultures, social groups, and personal experiences. Long Short-Term Memory (LSTMs), Convolution Neural Networks (CNN), Deep Neural Networks, etc., are used for both binary

and multi-class classification. A study says that the context matters significantly in toxicity annotations but for a very narrow slice of the dataset used [5]. Google's research team implemented a library named Detoxify to get the levels and type of toxicity in a sentence. In a paper the author presents two models for text detoxification, which have extra control for content preservation [28]. The first model, ParaGeDi, is capable of fully regenerating the input. The second approach, CondBERT, uses BERT to replace toxic spans with non-toxic alternatives. Detoxification of text is a relatively new style transfer task. A search engine finds non-toxic sentences similar to the given toxic ones, an MLM fills the gaps, and a seq2seq model edits the generated sentence to make it more fluent.This detoxification work rephrases the output of Language models and accordingly gives the detoxified sentences [9].

III. METHODOLOGY

Our approach is straightforward. The toxic comment is passed through toxicity level check first. If the comment is found to be of higher toxicity level, it is passed through our Detoxification Module. The output sentence goes through both toxicity and similarity check. The sentence with least toxicity and highest similarity is given as output by our model.

A. Detoxification Module

Input to Detoxification Module is a comment. We analyze the whole comment and find out the toxic words. The Ontology of Toxic words created is traversed and the toxic word is replaced with each of its less toxic synonym or with negation of its antonym to create new comments. All these sentences generated by our model are checked for their similarity with the original comment. The comment with least toxicity and highest similarity score to the input comment is given as output. Some offensive words, which do not have either synonym or antonym, which are used just as a substitute to express anger or to make someone feel offended are also included in the Ontology, the model would just delete these words and its complementary words and then check for similarity. This would ensure reduced toxicity and similar meaning to express opinions in a civil manner. Details of building our Toxic words Ontology are given below.

B. Building Toxic Words Ontology

In this research work, we have collected a list of 2939 offensive words from the web. These words are collected from various sources such as Wikipedia, comments, already built dictionaries, etc., Individual toxic words are given to WordNet to get the list of synonyms and antonyms. Detoxify, a toxicity classifier model built by Google's research team, is used to get the toxicity levels for any word. Toxicity levels for every synonym and antonyms for all the toxic words are found using this Detoxify model. Individual words in the synonyms and antonyms list are sorted in ascending order of their toxicity level. The generated ontologies are stored in .json files for use in future applications. Two separate graphs each for antonym and synonym are generated. As per our knowledge, this Ontology for toxic words is first Ontology created for toxic words refer Fig. 4 for Ontology generation steps. Representation of ontology, refer Fig. 3

C. Dataset

We have taken the Toxic Comment Classification Challenge dataset from the Kaggle competition. This dataset has 1,59,571 comments in its train dataset. The English data serves as training data. Not all the comments are toxic. We filtered out only those comments labeled 'toxic' as '1'. There are a total of 15,294 toxic, such comments. We have split these 15,294 sentences into 12,000 training set and the remaining 3,294 into the testing set. (Disclaimer: Text in the dataset could be considered profane, vulgar, or objectionable.)

D. Metrics

There is no Gold text available for this dataset, so we cannot use BLEU, ROUGE, or METEOR metrics. The main aim of our model was to give a less toxic and most similar output comment to the original comment. We evaluated our results based on Perspective API's toxicity score and RoBERTa's similarity score. The reduced toxicity score and more similarity score are two metrics that our output comment has to go through to be considered valid.

IV. RESULTS

A. Toxic words Ontology

Two ontologies, each for antonyms and synonyms, are created. The Toxic words dictionary has a total of 2939 words. Total nodes in synonym ontology: 14831, and total nodes in antonym ontology: 3215. The longest list of synonyms for a single toxic word is 69. Similarly, for antonyms longest list is 5 refer Table - 1. For example, of hierarchy of words 'Drunk' and 'Evil', refer to Fig-1 and Fig-2 respectively.

TABLE I. RESULTS OF ONTOLOGY GENERATION

Total words: 2939	Antonyms Ontology	Synonyms Ontology
longest chain	5	69
nodes	276	11892
Total Nodes	3215	14831

Fig. 1. Example : 'Drunk'

Fig. 2. Example : 'Evil'

B. Results generated by the model

In comparison with the ParaGeDi model, as shown in the results table, our model reduces the toxicity of the same sentence, and also the similarity score of the sentence remains quite high. The input is taken from (Text Detoxification using Large Pre-trained Neural Models David Dale,) toxicity scores before and after detoxification are as shown in the table. We can see that the comment's toxicity score is reduced and the output has a higher similarity score than the ParaGedi model. Also, the generated output is shown in the table. Refer Table 2 and Table- 3. Intermediate results are shown in Table-4.

Fig. 3. Ontology Representation

TABLE II. COMPARISON OF OUTPUTS

Input	Toxicity	Models	Output	Toxicity	Similarity
No room for those stupid antics	0.987	ParaGeDi	There's no space for all those adversities	0.00101	0.45676
		Ontology based(ours)	No room for those poor fish antics	0.08530	0.66909
This whole article is bullshit	0.9686	ParaGeDi	This article's not a good deal	0.003102	0.3705
		Ontology based(ours)	This whole article is fake	0.0219	0.9105

TABLE III. OUTPUT GENERATED BY OUR MODEL AND THEIR TOXICITY LEVELS AND SIMILARITY SCORES

Input	Toxicity level	Output Comment	Output toxicity level	Similarity score
what the fuck are you talking about	0.9677	what are you talking about	0.01562	0.846200
you are a nazi asshole and always will be	0.9989	you are a national socialist and always will be	0.01770	0.48365
dumb people vote for Mr. X	0.99198	dim people vote for Mr. X	0.04208	0.61942

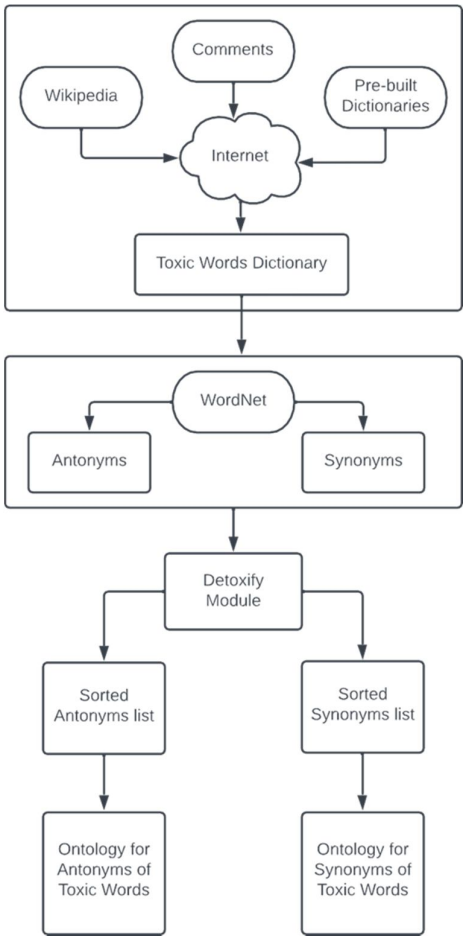

Fig. 4. Ontology Generation

TABLE IV. INTERMEDIATE RESULTS OF OUR MODEL. FOR: "NO ROOM FOR THOSE STUPID ANTICS" TOXICITY LEVEL: 0.98736

Output generated	Toxicity level	Similarity score
no room for those poor fish antics	0.08530	0.66909
no room for those stunned antics	0.12992	0.86119
no room for those dazed antics	0.62612	0.84974
no room for those stupefied antics	0.10238	0.89268
no room for those pillock antics	0.5104	0.84873
no room for those pudding head antics	0.27100	0.63546
no room for those stupe antics	0.23679	0.67144
no room for those pudden-head antics	0.45770	0.72873
no room for those dolt antics	0.51857	0.83263
no room for those dullard antics	0.89576	0.80881
no room for those stupid person antics	- 0.84088	0.95637

V. CONCLUSIONS

As per our knowledge, this ontology is the first ontology built over toxic words. Traversing, understanding and updating this ontology is easy as it is a graph data structure. Each word has two relationships: antonyms and synonyms, and every antonym and synonym have an attribute toxicity level associated with it. Detoxification along with similarity of the output sentence with original sentence is the main goal of this research. The result shows that the detoxification work has been done along with retaining the meaning of the original sentence The ontology can be extended by adding more toxic words to the list. This ontology can be used to get less toxic words for applications like detoxification of toxic comments. The research work done can be extended to create toxic words ontology for other languages.

ACKNOWLEDGMENT

We are thankful to the Dept of Computer Engineering and IT for providing the high computing GPU server facility procured under TEQIP-III (A world bank project) for our project/research work.

REFERENCES

[1] Laugier, L., Pavlopoulos, J., Sorensen, J. and Dixon, L., 2021. Civil rephrases of toxic texts with self-supervised transformers. arXiv preprint arXiv:2102.05456.
[2] Davidson, Thomas, et al. "Automated hate speech detection and the problem of offensive language." Proceedings of the International AAAI Conference on Web and Social Media. Vol. 11. No. 1. 2017.
[3] Lees, Alyssa, et al. "A New Generation of Perspective API: Efficient Multilingual Character-level Transformers." arXiv preprint arXiv:2202.11176 (2022).
[4] Hosseini, Hossein, et al. "Deceiving google's perspective api built for detecting toxic comments." arXiv preprint arXiv:1702.08138 (2017).
[5] Pavlopoulos, John, et al. "Toxicity detection: Does context really matter?." arXiv preprint arXiv:2006.00998 (2020).
[6] Hochreiter, Sepp, and Ju¨rgen Schmidhuber. "Long short-term memory." Neural computation 9.8 (1997): 1735
[7] Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for lan- guage understanding." arXiv preprint arXiv:1810.04805 (2018).
[8] Fensel, Dieter. "Ontologies." Ontologies. Springer, Berlin, Heidelberg, 2001. 11-18.
[9] Gruber, T. R. "A translation approach to portable ontologies Knowledge Acquisition 5 (2) 199220." (1993).
[10] Buitelaar, Paul, Philipp Cimiano, and Bernardo Magnini, eds. Ontology learning from text: methods, evaluation and applications. Vol. 123. IOS press, 2005.
[11] Ontologies D Fensel - Ontologies, 2001 - Springer
[12] Uschold, Mike, and Michael Gruninger. "Ontologies: Principles, methods and ap- plications." The knowledge engineering review 11.2 (1996): 93-136.
[13] Fellbaum, Christiane. "WordNet." Theory and applications of ontology: computer applications. Springer, Dordrecht.
[14] Miller, George A. "WordNet: a lexical database for English." Communications of the ACM 38.11 (1995): 39-41.
[15] Ormerod, Mark, Jesu´s Mart´ınez Del Rinc´on, and Barry Devereux. "Predicting Semantic Similarity Between Clinical Sentence Pairs Using Transformer Models: Evaluation and Representational Analysis." JMIR Medical Informatics 9.5 (2021): e23099.

[16] Vaswani, Ashish, et al. ”Attention is all you need.” Advances in neural information processing systems 30 (2017).
[17] Liu, Yinhan, et al. ”Roberta: A robustly optimized bert pretraining approach.” arXiv preprint arXiv:1907.11692
[18] Radford, Alec, et al. ”Language models are unsupervised multitask learners.” Ope- nAI blog 1.8 (2019): 9.
[19] Rao, Sudha, and Joel Tetreault. ”Dear sir or madam, may i introduce the gyafc dataset: Corpus, benchmarks and metrics for formality style transfer.” arXiv preprint arXiv:1803.06535 (2018).
[20] Wu, Xing, et al. ”” Mask and Infill”: Applying Masked Language Model to Senti- ment Transfer.” (2019).
[21] Hu, Zhiting, et al. ”Toward controlled generation of text.” International conference on machine learning. PMLR.
[22] John, Vineet, et al. ”Disentangled representation learning for non-parallel text style transfer.” (2018).
[23] Raffel, Colin, et al. ”Exploring the limits of transfer learning with a unified text- to-text transformer.” (2019).
[24] Baziotis, Christos, et al. ”SEQ 3: Differentiable Sequence-to-Sequence-to-Sequence Autoencoder for Unsupervised Abstractive Sentence Compression.” arXiv preprint arXiv:1904.03651 (2019).
[25] Santos, Cicero Nogueira dos, Igor Melnyk, and Inkit Padhi. ”Fighting offensive language on social media with unsupervised text style transfer.” arXiv preprint arXiv:1805.07685 (2018).
[26] Ficler, Jessica, and Yoav Goldberg. ”Controlling linguistic style aspects in neural language generation.” (2017).
[27] Santos, Cicero Nogueira dos, Igor Melnyk, and Inkit Padhi. ”Fighting offensive language on social media with unsupervised text style transfer.” arXiv preprint arXiv:1805.07685 (2018).
[28] Dale, David, et al. ”Text Detoxification using Large Pre-trained Neural Models.” arXiv preprint arXiv:2109.08914.
[29] Author, A.-B.: Contribution title. In: 9th International Proceedings on Proceed- ings, pp. 1–2. Publisher, Location.
[30] Perspective Homepage, https://perspectiveapi.com/