# Toxic Comment Detection: Analyzing the Combination of Text and Emojis

Michael Aquino, Yasiris Ortiz, Arif Rashid, Anne M. Tumlin,
N. Sertac Artan, Ziqian Dong, and Huanying Gu

***Abstract*—Detection of toxicity in online commentary is a growing branch of Natural Language Processing (NLP). Most research in the area rely only on text-based toxic comment detection. We propose a machine learning approach for detecting the toxicity of a comment by analyzing both the text and the emojis within the comment. Our approach utilizes word embeddings derived from GloVe and emoji2vec to train a bidirectional Long Short Term Memory (biLSTM) model. We also create a new labeled dataset with comments with text and emojis. The accuracy score of our model on preliminary data is 0.911.**

***Index Terms*—Toxic comments, natural language processing (NLP), emojis.**

## I. Introduction

Social media is omnipresent in everyday life. Whether it be for work, learning, or social connection, there has been an increased amount of communication online. Subsequently, online toxicity such as harassment, bullying, and violence is also on the rise. Dealing with this toxicity online has become a growing problem, but just because social media is here to stay does not mean that toxicity online has to stay as well.

Recently, researchers have focused on identifying and mitigating toxic comments online. However, two avenues have remained open for further exploration: 1) to analyze the toxicity of a comment based on both text and emojis, and 2) to create a new dataset to fill a gap on data availability. Emojis are just as important for analyzing toxicity because they can show the sentiment of a comment. For example, the sarcastic comment *"good job, genius"* may not be easily identified as toxic. However, by adding a rolling eye emoji to the comment, *"good job, genius* 🙄*"* this can clearly be identified as a toxic comment. In this paper, we propose a model that is able to detect the toxicity of a comment based upon the text as well as the emojis within that comment.

M. Aquino, N.S. Artan and Z. Dong are with Department of Electrical and Computer Engineering, A. Rashid and H. Gu are with Department of Computer Science, New York Institute of Technology, New York, NY, 10023. E-mail: {maquin01, nartan, ziqian.dong, mrashi12, hgu03}@nyit.edu,
Y. Ortiz is with Department of Computer Science, The City College of New York. E-mail: yortizm000@citymail.cuny.edu,
A. Tumlin is with Department of Computer Science and Engineering, University of South Carolina, Email: atumlin@email.sc.edu.
This project is funded by National Science Foundation Grant No. 1852316.

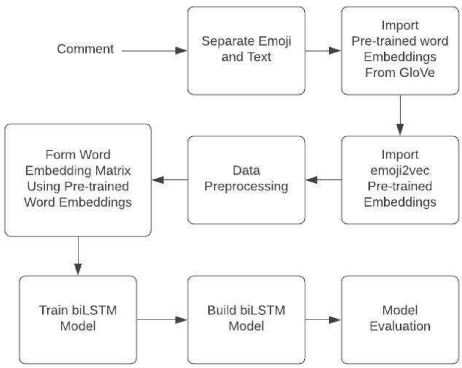


Fig. 1. The proposed toxic comment detection approach.

## II. Related Work

Using NLP, researchers analyzed the contents of various texts such as YouTube comments and Tweets [11]. Other papers explored the concept of bilingual toxicity detection [1],[4]. These papers analyzed tweets and even news articles in various languages with the goal of detecting multilingual toxicity. Alshamrani et al. studied different toxic behaviors such as obscenity and hate from different news topics posted on mainstream media and news channels on YouTube [2]. The comments were inspected for three categories: toxic, obscene, and identity hate. These comments were then classified by utilizing a neural network-based ensemble, Deep Neural Network (DNN)-based Architecture. Furthermore, toxicity detection APIs (Application Programming Interfaces) like Perspective API created by Jigsaw and Google [9] and the Komprehend API by Paralleldots [10], seek to mitigate online toxicity. Adversarial machine learning techniques have been used for evading toxicity detectors [8],[3].

Another avenue of NLP research is sarcasm detection. Among many papers for sarcasm detection, Subramanian et al.'s work [14] stands out as they used emojis along with text to detect sarcasm. They proposed an ESD (Emoji-based Sarcasm Detection) framework for the simultaneous evaluation of text and emoji for sarcasm detection in social media. The framework was tested on multiple machine learning models including Convolutional Neural Networks (CNN) and Long Short Term Memory (LSTM) [14].

TABLE I
Distribution of Augmented Twitter Dataset.

| | Total | Toxic | Non-Toxic |
|---|---|---|---|
| **Comments** | 56,742 | 24,153 | 32,589 |

TABLE II
Model Results using Augmented Dataset

| | ROC-AUC Score | Accuracy Score |
|---|---|---|
| **Our Model** | 0.982 | 0.911 |

## III. Evaluating the Combination of Text and Emojis for Detecting Toxic Comments

In this paper, we propose a model (Fig. 1) that analyzes the toxicity of a comment based on the text as well as the emojis within that comment. We use an open-source text-based toxicity detector by Baishali Dutta [5]. The detector employs GloVe [12], which converts input text into vectors, and then trains a bidirectional Long Short Term Memory model (biLSTM) [5]. Prior to sending the comment to the detector, we separate the text and the emojis within the comment. Using the pre-trained GloVe word embeddings, we create a vector representation of the text. This is because GloVe represents each word as a vector of probabilities based on how likely that word is to appear with other words [12]. Then, using the pre-trained embeddings from emoji2vec [13], we create a vector representation of the emojis. Finally, these two embeddings are combined to create the final vector, which is sent to the detector for toxicity analysis.

To generate embeddings for emojis, we use emoji2vec [13]. This model embeds emojis into vectors by generating embeddings of the words in a textual description of each emoji using GloVe and then adding these vectors together.

### A. Dataset

Due to a lack of labeled datasets with toxic comments including both text and emojis, we modified an existing Twitter dataset of toxic comments [7] to contain emojis. We parsed the dataset through the DeepMoji model [6] which outputs emojis based on the sentiment of textual input. We then concatenated the text of the tweet with the outputted emojis. Properties of our augmented dataset can be seen in Table I.

## IV. Results and Discussions

Our preliminary results as shown in Table II although promising were lower than expected, which can be explained with the length and cleanliness of the dataset and the need for improvements within the model. We will use a larger and cleaner dataset to improve the accuracy.

## V. Conclusions

As online toxicity increases, researchers are investigating how to detect and mitigate toxic comments. In this paper, we proposed an approach that detects the toxicity of a comment based on the combination of text and emojis. Additionally, we created a new labeled dataset. Our future work will focus on model accuracy improvements.

## References


[1] A. Aggarwal, A. Wadhawan, A. Chaudhary, and K. Maurya. "did you really mean what you said?" : Sarcasm detection in hindi-english code-mixed data using bilingual word embeddings. In *The Workshop on Noisy User-generated Text (W-NUT 2020)*, 2020.

[2] S. Alshamrani, M. Abuhamad, A. Abusnaina, and D. Mohaisen. Investigating online toxicity in users interactions with the mainstream media channels on youtube. 2020.

[3] S. Brown, P. Milkov, S. Patel, Y. Looi, Z. Dong, H. Gu, N. Sertac Artan, and E. Jain. Acoustic and visual approaches to adversarial text generation for google perspective. In *2019 International Conference on Computational Science and Computational Intelligence (CSCI)*, pages 355–360, 2019.

[4] Y. Dinkov, I. Koychev, and P. Nakov. Detecting toxicity in news articles: Application to Bulgarian. In *Natural Language Processing in a Deep Learning World*, 2019.

[5] B. Dutta. Comments toxicity detection. [Online], 2020. https://github.com/baishalidutta/Comments-Toxicity-Detection.

[6] Bjarke F., A. Mislove, A. Søgaard, I. Rahwan, and S. Lehmann. Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm. In *Conf. on Empirical Methods in Natural Language Processing*, 2017.

[7] A. U. Iyer. Toxic tweets dataset. [Online], 2021. www.kaggle.com/ashwiniyer176/.

[8] E. Jain, S. Brown, J. Chen, E. Neaton, M. Baidas, Z. Dong, H. Gu, and N. S. Artan. Adversarial text generation for google's perspective api. In *International Conference on Computational Science and Computational Intelligence (CSCI)*, 2018.

[9] Jigsaw. Perspective api. [Online], 2018. https://conversationai.github.io/.

[10] ParallelDots. Komprehend. [Online], 2020. https://komprehend.io/.

[11] B. Pariyani, K. Shah, M. Shah, T. Vyas, and S. Degadwala. Hate speech detection in twitter using natural language processing. In *International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV)*, 2021.

[12] J. Pennington, R. Socher, and C. Manning. Glove: Global vectors for word representation. In *The Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014.

[13] M. Pislar. emoji2vec. [Online], 2018. https://github.com/MirunaPislar/emoji2vec.

[14] J. Subramanian, V. Sridharan, K. Shu, and H. Liu. Exploiting emojis for sarcasm detection. *Social, Cultural, and Behavioral Modeling*, page 70–80, 2019.