2021-08

GoldenWind at SemEval-2021 Task 5: Orthrus An Ensemble Approach to Identify Toxicity

PALOMINO, MARCO

SemEval 2021 - 15th International Workshop on Semantic Evaluation, Proceedings of the Workshop

GoldenWind at SemEval-2021 Task 5: Orthrus – An Ensemble Approach to Identify Toxicity

Marco Palomino[*] **Dawid Grad**[†] **James Bedwell**[‡]
School of Engineering, Computing and Mathematics
University of Plymouth
Plymouth, Devon, PL4 8AA, United Kingdom
marco.palomino@plymouth.ac.uk[*]
{dawid.grad[†], james.bedwell[‡]}@students.plymouth.ac.uk

Abstract

Many new developments to detect and mitigate toxicity are currently being evaluated. We are particularly interested in the correlation between toxicity and the emotions expressed in online posts. While toxicity may be disguised by amending the wording of posts, emotions will not. Therefore, we describe here an ensemble method to identify toxicity and classify the emotions expressed on a corpus of annotated posts published by Task 5 of SemEval 2021—our analysis shows that the majority of such posts express anger, sadness and fear. Our method to identify toxicity combines a lexicon-based approach, which on its own achieves an F1 score of 61.07%, with a supervised learning approach, which on its own achieves an F1 score of 60%. When both methods are combined, the ensemble achieves an F1 score of 66.37%.

1 Introduction

Healthy conversations are only possible when people feel safe from abuse and do not resort to using violent language. Regrettably, violent and inflammatory language is becoming increasingly common online. Indeed, the rhetoric of violence recently employed on social media has persuaded platforms, such as Twitter, to create new policies to prevent the use of threatening language (Twitter, Inc., 2021). A jargon word, *cyberbullying*, has been coined lately to refer to the use of electronic communication to send or post messages of an intimidating or threatening nature (Zaheri et al., 2020).

Along with cyberbullying, other forms of verbal abuse employed on social media, such as online harassment and hate speech, are now being collectively referred to as *toxicity* in language (Mohan et al., 2017). We are interested in developing algorithms to recognise toxicity and measure its impact on the sentiment expressed.

Most of the data available to investigate toxicity classify whole comments or documents (Wulczyn et al., 2017; Borkan et al., 2019), and do not identify "spans"—that is, the precise word sequences that make a text toxic. Given how important such spans are for the implementation of semi-automated moderation, we have participated on Task 5 (Toxic Spans Detection) of the *International Workshop on Semantic Evaluation* (SemEval) 2021 (Pavlopoulos et al., 2021). Thus far, we have concentrated on the combination of two approaches: a lexicon-based approach and a supervised learning approach to identify toxic spans.

Although the identification of toxic spans in online posts can be aided by a suitable lexicon of toxic words, such words can easily be concealed through minor changes—for instance, "fck urself" is a toxic span that would evade detection based on basic lists of profane words. However, emotions are harder to conceal. Hence, we are interested in using opinion mining to uncover the emotions expressed in text. Emotions may be able to identify toxicity, regardless of wordings and spellings. Thus, we dedicate part of this study to measure the correlation between toxicity and emotions.

The remainder of this paper is organised as follows: Section 2 reviews the related work. Section 3 describes the datasets that we used for our experimentation. Section 4 is dedicated to explain our algorithm for the identification of toxic spans. Section 5 presents our results and, finally, Section 6 offers our conclusions.

2 Background

The existing literature on toxicity focuses on two main aspects: the compilation and annotation of corpora for research purposes (Fortuna et al., 2020; Waseem, 2016); and the automatic detection of different types of toxic text.

Among the different types of toxic text under scrutiny, we may include hate speech (Badjatiya et al., 2017; Davidson et al., 2017; Del Vigna et al., 2017), online harassment (Golbeck et al., 2017), racism (Waseem, 2016), sexism (Jha and Mamidi, 2017), abusive language (Mehdad and Tetreault, 2016) and cyberbullying (Zhong et al., 2016).

At present, the detection of toxicity is largely based on state-of-the-art natural language processing techniques, typically involving machine learning. The main drawback of such techniques resides in the limited generalisation potential of trained machine learning models (Fortuna et al., 2020). To overcome this weakness, we have integrated into our research the use of a lexicon-based approach, where toxic language is identified with the help of a dictionary of words associated with toxic text (De Smedt et al., 2020).

Combining lexicons with machine learning approaches has already been evaluated by other researchers, remarkably Pamungkas and Patti (Pamungkas and Patti, 2019), though they employed a lexicon originally built for the Italian language, and then translated it into other languages, whereas we focus on English from the start. Various other lexicons, handcrafted by domain experts who specialise on the identification of toxicity have been published too—for example, *Textgain's Profanity and Offensive Words* lexicon (De Smedt et al., 2020)—but many of them are not available for free.

In an attempt to mitigate toxicity and promote work on this area, the research community has released a number of annotated datasets for investigating different forms of toxicity (Waseem and Hovy, 2016; Waseem, 2016; Golbeck et al., 2017). However, they all follow different labelling conventions. Consequently, they cannot be analysed using a uniform method.

Overall, toxicity detection and classification lacks a consistently labelled standard dataset for comparative evaluation (Schmidt and Wiegand, 2017). Therefore, the data provided by Task 5 (Toxic Spans Detection) of SemEval 2021 is very well regarded (Pavlopoulos et al., 2021).

3 Experimental Setup

Task 5 of SemEval 2021 uses posts from the publicly available *Civil Comments* dataset (TensorFlow, 2021). Such a dataset comprises annotations indicating which entire posts are toxic, but it does not label particular toxic spans within the posts.

The Civil Comments platform (Drupal, 2021), which is where the posts come from, is a commenting plugin for independent news websites. All the comments were created between 2015 and 2017, and they appeared on approximately 50 English language websites across the world. When Civil Comments shut down in 2017, the comments became publicly available in an open archive for future research (TensorFlow, 2021).

To build the dataset, SemEval retained only posts that were found toxic—or severely toxic—by at least half of the annotators involved in Borkan, *et al.*'s annotation (Borkan et al., 2019). This comprises 30k toxic posts, approximately, out of the original 1.2M. Then, a random subset of 10k posts from these 30k toxic posts were chosen for toxic spans annotation (CodaLab, 2021).

4 System Overview

Although machine learning technology is being widely employed to detect toxic text automatically, the use of a lexicon to identify and prevent toxicity in social media still constitutes a valuable approach. Indeed, the number of lexicons specialised on the detection of profanity, offensive speech and toxicity in general has grown steadily in recent times (De Smedt et al., 2020).

Lexicons are not susceptible to algorithmic bias (Hajian et al., 2016), and are not limited to the domain and scope of the training data, which has previously raised a number of ethical concerns, given how much training data is historically associated with particular communities (Hao, 2019). Hence, we employ a lexicon as our first step in the detection of toxic spans.

Originally, our lexicon was made, specifically, for Task 5 of SemEval 2021, as we compiled it by extracting all the toxic words available in the training and trial datasets for Task 5—we considered a word as a *toxic* word if it was included in a toxic span identified by the annotators.

Upon compiling all the toxic words available in the training and trial datasets (1,287 words), we proceeded to extend our lexicon with words listed in other lexicons. While there are many freely-available lexicons of toxic words, we favoured those that maintained the accuracy of the detection of toxic spans achieved by our lexicon. Table 1 shows the F1 scores achieved by each of the lexicons considered, when combined with our lexicon to evaluate them on the training dataset.

Lexicon	F1 Training	Number of words
Task 5 Lexicon	64.30%	1,287
Banned Word List	**64.30%**	1,332
Offensive/Profane Word List	61.25%	2,516
Google's Profanity Words	**64.30%**	1,681
Insult.wiki	63.94%	1,846
Compiled_bad_words	63.99%	2,546
Swear Word List & Curse Filter	**64.30%**	1,580

Table 1: F1 score per lexicon (evaluated on the training dataset).

The first row of Table 1 refers to the lexicon we created after compiling all the toxic words available in the training and trial datasets of Task 5 of SemEval 2021—we named this lexicon the *Task 5 Lexicon*. Using bold font, we have highlighted the details of the lexicons that achieved the same F1 score as the Task 5 Lexicon, when combined with it to evaluate them on the training dataset. Such lexicons are the ones that we decided to use, namely, the *Banned Word List* (http://www.bannedwordlist.com/), *Google's Profanity Words* (https://github.com/RobertJGabriel/Google-profanity-words), and the *Swear Word List & Curse Filter* (https://www.noswearing.com/dictionary). Table 1 also displays the number of words available in each of the lexicons evaluated, when combined with the Task 5 Lexicon.

As shown in Table 1, the *Offensive / Profane Word List* (https://www.cs.cmu.edu/~biglou/resources/) and the *Compiled_bad_words* (https://github.com/minerva-ml/open-solution-toxic-comments/blob/master/external_data/compiled_bad_words.txt) have a negative impact on the performance of toxicity detection, even if it is only by a small margin. Thus, we discarded these lexicons.

After creating our lexicon, we manually removed from it words that were part of the toxic spans annotated in Task 5 of SemEval 2021, but were not included in the three lexicons displayed in bold font in Table 1. For example, the word "mistake" located in the post "`They elected Trump, which was certainly a mistake`" was considered toxic by the annotators, in the context of the post. However, we removed it from our lexicon, because "mistake" does not appear in any of the three lexicons mentioned above. Our lexicon comprises a total of 1,929 words, and we refer to it as the *Orthrus* lexicon—it is available at https://github.com/Orthrus-Lexicon/Toxic.

While our lexicon-based approach was considerably useful to identify toxicity, as we will show in Section 5, we recognise the value of machine learning approaches. The success of the *Perspective* project undertaken by Google and Jigsaw to rate toxicity by means of machine learning (Jain et al., 2018), as well as the impact of the *Perspective API* to mitigate toxicity using machine learning certainly deserve our attention. Therefore, we opted to employ *spaCy* (Explosion, 2021b), an open-source software library for natural language processing, to develop a supervised learning approach for the identification of toxicity.

Our choice of spaCy was further motivated by the organisers of Task 5 of SemEval 2021, who released a Python script referring, precisely, to this library (Task 5, 2021). Initially, we employed `en_core_web_sm`, which is a spaCy model for the English language (Explosion, 2021a). We employed this model, because it was the one used in the code provided by the organisers of Task 5 as a solution for some NLP tasks—namely, POS tagging, NER and dependency parsing (Task 5, 2021). However, given that `en_core_web_sm` is based on a small English corpus, we also tested `en_core_web_lg`, which is spaCy's large English model (Explosion, 2021a).

Despite spaCy's large English model being understandably slower, it did not appear to improve the performance of our implementation. The F1 score achieved, on average, by spaCy's small English model after 10 executions (59.61%) was approximately the same as the score achieved by the large English model under the same circumstances (59.95%). Thus, we favoured the choice of the small model, as it was faster to train.

5 Results

Table 2 shows the F1 score achieved by our implementation when evaluating it on the test dataset.

Approach	F1 Score
Orthrus Lexicon	61.07%
Orthrus Lexicon + spaCy Model (Union)	61.53%
Orthrus Lexicon + spaCy Model (Intersection)	66.37%

Table 2: Evaluation on the test dataset.

As shown in Table 2, our lexicon achieves, on its own, an F1 score of 61.07%. By combining our lexicon with the supervised learning approach implemented using spaCy, we achieve two results: 61.53%, if we consider the union of the results yielded by the lexicon and the supervised learning approach; and 66.37%, if we consider the intersection of the results yielded by the lexicon and the supervised learning approach.

We are interested in the identification of emotions expressed in text, because concealing emotions may be harder than disguising toxicity. For example, the post *"uh, no, he's a belligerent buffoon (and a traitor)"*, which is post 1,928 of the training dataset of Task 5 of SemEval 2021, lacks any recognisable toxic features, such as insults or swear words. Hence, it is classified as non-toxic by any of the lexicons highlighted in bold font in Table 1. Moreover, this post does not have any toxic spans marked by the annotators. Nevertheless, the negative sentiment of "belligerent buffoon" and "traitor"—words which are not typically found in any abusive word list—guarantees that the message conveyed is definitely toxic; otherwise, it would not be part of the training dataset.

If we included emotion information in our analysis, we could immediately detect the negative tone of the post mentioned above. Indeed, the probability of such a post to communicate anger is 1.0, according to `text2emotion`, a Python package to extract emotions from text (Python Software Foundation, 2021). The expression of anger is so evident in this case that the post can be marked as a candidate to be considered toxic.

Using `text2emotion`, we assigned each post in the test dataset a probability associated with each of the emotions reported in Figure 1. The values shown in Figure 1 represent the addition of the probabilities of each emotion to occur in each of the posts of the test dataset. Clearly, fear, sadness and anger—the three emotions combined together—are more likely to occur than happiness and surprise—the two emotions combined together—which may characterise the toxicity of the dataset.

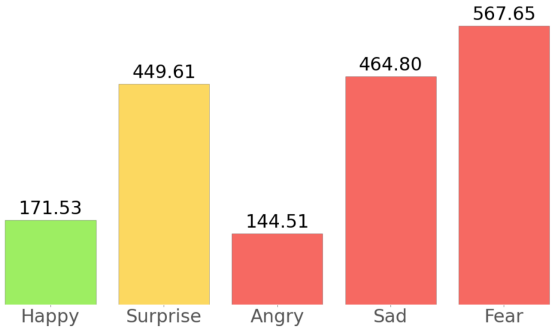

Figure 1: Emotion expressed on the test dataset.

6 Conclusions

In this paper, we have described the creation of a lexicon of toxic words and a supervised learning approach to identify toxicity in online posts. Our lexicon, along with the supervised learning approach, achieved an F1 score of 66.37% on Task 5 of SemEval 2021. We have also explored the relationship between emotions and toxicity. Although our study is still in progress, preliminary results indicate that there exists a correlation between emotions such as sadness and fear and toxicity.

References

Pinkesh Badjatiya, Shashank Gupta, Manish Gupta, and Vasudeva Varma. 2017. Deep learning for hate speech detection in tweets. In *Proceedings of the 26th International Conference on World Wide Web*, pages 759–760.

Daniel Borkan, Lucas Dixon, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2019. Nuanced Metrics for Measuring Unintended Bias with Real Data for Text Classification. In *Proceedings of the 2019 World Wide Web Conference*, pages 491–500.

CodaLab. 2021. SemEval 2021 Task 5: Toxic Spans Detection. https://competitions.codalab.org/competitions/25623#learn_the_details-data.

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*.

Tom De Smedt, Pierre Voué, Sylvia Jaki, Melina Röttcher, and Guy De Pauw. 2020. Profanity & Offensive Words (POW). *Textgain*.

Fabio Del Vigna, Andrea Cimino, Felice Dell'Orletta, Marinella Petrocchi, and Maurizio Tesconi. 2017. Hate Me, Hate Me Not: Hate Speech Detection on Facebook. In *Proceedings of the First Italian Conference on Cybersecurity (ITASEC17)*, pages 86–95.

Drupal. 2021. Civil Comments. https://www.drupal.org/project/civilcomments.

Explosion. 2021a. English – Available Trained Pipelines for English. https://spacy.io/models/en.

Explosion. 2021b. spaCy. https://spacy.io/.

Paula Fortuna, Juan Soler, and Leo Wanner. 2020. Toxic, Hateful, Offensive or Abusive? What Are We Really Classifying? An Empirical Analysis of Hate Speech Datasets. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 6786–6794.

Jennifer Golbeck, Zahra Ashktorab, Rashad O Banjo, Alexandra Berlinger, Siddharth Bhagwan, Cody Buntain, Paul Cheakalos, Alicia A Geller, Rajesh Kumar Gnanasekaran, Raja Rajan Gunasekaran, et al. 2017. A Large Labeled Corpus for Online Harassment Research. In *Proceedings of the 2017 ACM on Web Science Conference*, pages 229–233.

Sara Hajian, Francesco Bonchi, and Carlos Castillo. 2016. Algorithmic Bias: From Discrimination Discovery to Fairness-Aware Data Mining. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 2125–2126.

Karen Hao. 2019. AI Is Sending People to Jail–And Getting It Wrong. *MIT Technology Review*.

Edwin Jain, Stephan Brown, Jeffery Chen, Erin Neaton, Mohammad Baidas, Ziqian Dong, Huanying Gu, and Nabi Sertac Artan. 2018. Adversarial Text Generation for Google's Perspective API. In *International Conference on Computational Science and Computational Intelligence (CSCI)*, pages 1136–1141. IEEE.

Akshita Jha and Radhika Mamidi. 2017. When Does a Compliment Become Sexist? Analysis and Classification of Ambivalent Sexism Using Twitter Data. In *Proceedings of the Workshop on NLP and Computational Social Science*, pages 7–16.

Yashar Mehdad and Joel Tetreault. 2016. Do Characters Abuse More than Words? In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 299–303.

Shruthi Mohan, Apala Guha, Michael Harris, Fred Popowich, Ashley Schuster, and Chris Priebe. 2017. The Impact of Toxic Language on the Health of Reddit Communities. In *Canadian Conference on Artificial Intelligence*, pages 51–56. Springer.

Endang Wahyu Pamungkas and Viviana Patti. 2019. Cross-Domain and Cross-Lingual Abusive Language Detection: A Hybrid Approach with Deep Learning and a Multilingual Lexicon. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 363–370.

John Pavlopoulos, Léo Laugier, Jeffrey Sorensen, and Ion Androutsopoulos. 2021. Semeval-2021 Task 5: Toxic Spans Detection (To Appear). In *Proceedings of the 15th International Workshop on Semantic Evaluation*.

Python Software Foundation. 2021. text2emotion 0.0.5. https://pypi.org/project/text2emotion/.

Anna Schmidt and Michael Wiegand. 2017. A Survey on Hate Speech Detection Using Natural Language Processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pages 1–10.

Task 5. 2021. Toxic Spans 2021. https://groups.google.com/g/toxic-spans.

TensorFlow. 2021. CivilComments Dataset. https://www.tensorflow.org/datasets/catalog/civil_comments.

Twitter, Inc. 2021. Violent Threats Policy. https://help.twitter.com/en/rules-and-policies/violent-threats-glorification.

Zeerak Waseem. 2016. Are You a Racist or Am I Seeing Things? Annotator Influence on Hate Speech Detection on Twitter. In *Proceedings of the First workshop on NLP and Computational Social Science*, pages 138–142.

Zeerak Waseem and Dirk Hovy. 2016. Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter. In *Proceedings of the NAACL Student Research Workshop*, pages 88–93.

Ellery Wulczyn, Nithum Thain, and Lucas Dixon. 2017. Ex Machina: Personal Attacks Seen at Scale. In *Proceedings of the 26th International Conference on World Wide Web*, pages 1391–1399.

Sara Zaheri, Jeff Leath, and David Stroud. 2020. Toxic comment classification. *SMU Data Science Review*, 3(1):13.

Haoti Zhong, Hao Li, Anna Cinzia Squicciarini, Sarah Michele Rajtmajer, Christopher Griffin, David J Miller, and Cornelia Caragea. 2016. Content-Driven Detection of Cyberbullying on the Instagram Social Network. In *International Joint Conference on Artificial Intelligence*, volume 16, pages 3952–3958.