Towards Toxic Positivity Detection

Ishan Sanjeev Upadhyay and **KV Aditya Srivatsa** and **Radhika Mamidi**

International Institute of Information Technology, Hyderabad

{ishan.sanjeev, k.v.aditya}@research.iiit.ac.in

radhika.mamidi@iiit.ac.in

Abstract

Over the past few years, there has been a growing concern around toxic positivity on social media which is a phenomenon where positivity is used to minimize one's emotional experience. In this paper, we create a dataset for toxic positivity classification from Twitter and an inspirational quote website. We then perform benchmarking experiments using various text classification models and show the suitability of these models for the task. We achieved a macro F1 score of 0.71 and a weighted F1 score of 0.85 by using an ensemble model. To the best of our knowledge, our dataset is the first such dataset created.

1 Introduction

Toxic positivity can be defined as the overgeneralization of a positive state of mind that encourages using positivity to suppress and displace any acknowledgement of stress and negativity (Sokal et al., 2020; Bosveld, 2021). The popularity of the term "toxic positivity" peaked during the COVID 19 pandemic (refer to figure 1) where it was used to identify advice that focused on just looking at the positive at a time when people were hurting due to loss of life, loss of jobs and other traumatic events.

Toxic positivity results in one minimizing one's own negative feelings and suppressing negativity instead of acknowledging, processing and working through it. Some examples of toxic positivity include telling someone to focus on the positive aspects of a loss, telling someone that positive thinking will solve all their problems, suggesting that things could be worse and shaming someone for expressing negative emotions. This suppression of emotions is not only unhelpful but also leads to poorer recovery from the negative effects of the emotion. Accepting and working through one's emotions is the better route to take while dealing with negative emotions (Campbell-Sills et al., 2006).

Macro level events like COVID 19 and climate change disasters have distressed many people in the past few years (Marazziti et al., 2021). Social media is used by people having mental health issues or going through a tough time to find community, support, advice and encouraging messages (Gowen et al., 2012). However, it becomes important to be able to differentiate between messages that may help uplift an individual and those that may look positive but promote suppression of emotions and cause great harm in the long term recovery from negative emotions. The harms of toxic positivity are not only limited to its deleterious mental health outcomes but it can also be used to uphold oppression by making people ignore the oppression that is going on and encouraging them to "just be positive".

In this paper, we aim to create a dataset for toxic positivity and perform text classification using various transformer based models to establish the baseline results for this task.

2 Related Work

There have been studies that show the ineffectiveness and deleterious effects of emotion suppression. Gross and John (2003) showed that people who suppressed their emotions had a greater experience of negative emotions while also expressing lesser positive emotion. They also showed that using suppression is related negatively to well being. A study done by Campbell-Sills et al. (2006) involved dividing 60 participants diagnosed with anxiety and mood disorders into two groups. One group was given a rationale for suppressing their emotions while the other was given a rationale for accepting emotions. It was found that suppression was ineffective in reducing distress while watching an emotion-provoking film. It was also seen that the suppression group showed a poorer recovery from the changes in negative affect after watching the film compared to the acceptance group. A
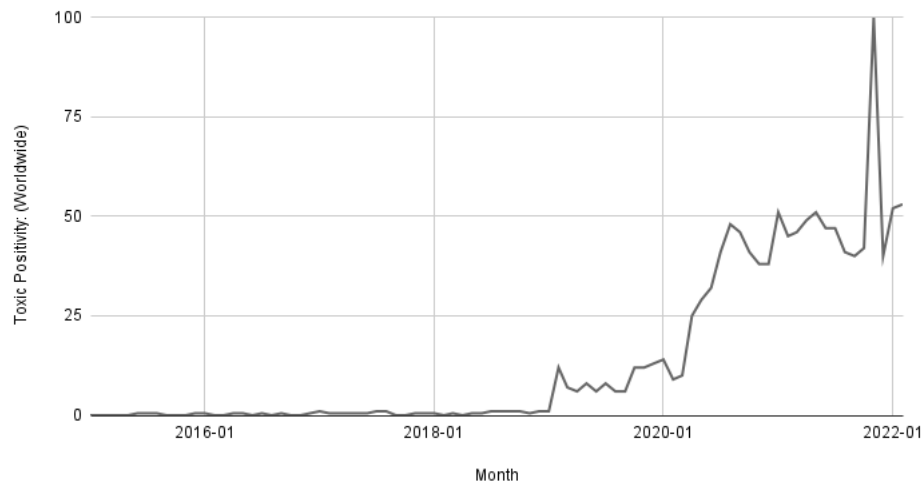
64

Figure 1: Worldwide Google Trends showing search interest of the term "Toxic Positivity".

similar observation is seen in the case of physical pain as well. Cioffi and Holloway (1993) divided participants into three groups during a cold-pressor pain induction (CPT) where participants would dip their hands in cold water for as long as tolerable. The first group was told to pay attention to the pain, the second was told to focus on their room at home as a distraction, and the third was told to suppress the sensations they felt. It was seen that the group that focused on the pain had a faster recovery from the pain and the suppression group had the slowest recovery from pain. Suppressing pain has shown to have negative outcomes, while accepting it is observed to be as a better strategy. Ford et al. (2018) through longitudinal and lab studies showed that habitually accepting mental experiences broadly predicted psychological health and that it reduced negative emotional response and experience. Hence toxic positivity, with its overemphasis on thinking positively and having a positive state of mind, encourages emotion suppression rather than emotional acceptance which has negative consequences for the person who engages in it.

Lecompte-Van Poucke (2022) conducted a critical discourse analysis of toxic positivity as a discursive construct on Facebook. Two corpora of posts from organizations that promoted endometriosis awareness (an invisible chronic condition) were analyzed using systematic functional linguistics, pragma-dialectics and critical theory. The study showed that users on social media platforms often engage in toxic positivity or forced positive discourse which is inspired by the neoliberal "positive thinking" ideology, leading to a less inclusive online community.

In the field of NLP, there have been many papers focusing on hate speech detection using support vector machine (SVM), long short term memory networks (LSTM),convolutional neural network (CNN), transformers and other machine learning models (Wang et al., 2019b; Zhang et al., 2018; Ousidhoum et al., 2019; Basile et al., 2019). These works use Twitter posts (tweets) to create datasets. YouTube and Reddit comments have also been used in some works (Mollas et al., 2022; Mandl et al., 2020). There have been recent efforts in hope speech detection as well (Palakodety et al., 2020). The HopeEDI dataset (Chakravarthi, 2020) is a hope speech dataset that contains Youtube comments that have been marked for hope and not-hope speech. There has been a shared task on this dataset where participants have used various machine learning models for hope speech detection like multilingual transformer-based models, recurrent neural networks (RNN) and CNN-LSTMs (Chakravarthi and Muralidaran, 2021).

However, to the best of our knowledge, there has been no prior work on creating datasets and classification models for toxic positivity.

3 Dataset Creation

3.1 Data Extraction and Pre-processing

We sourced our data from two sources. Twitter and inspirational quote website BrainyQuote[1] which is one of the largest quotation websites.

The reason for sourcing data from BrainyQuotes was that we observed that a lot of motivational quotes being shared on Twitter were ones that were said by famous personalities. Hence, including popular quotes from a quotation website is helpful. We made a web scraper using Beautiful Soup 4[2] library in python to extract a subset of quotations from the website.

For the Twitter data, we extracted tweets using Twitter API [3] we queried using hashtags like #MondayMotivation to #SundayMotivation and hashtags like #InspirationalQuotes, #Motivation, #SelfLove and #AdviceForSuccess. We also took quotes from widely followed inspirational or motivational twitter accounts.

After collecting the data, pre-processing was performed. Bylines of quotes were removed because it was not useful information for annotation and to also to ensure that there was no annotator bias. For tweets, hashtags and "@" tags were removed. The Twitter data and BrainyQuotes data was also manually filtered to remove sentences that were not inspirational, motivational or advisory in nature. Examples of the kind of data removed are given in Table 4. A total of 4,250 quotes and tweets were collected for annotation after the data elimination and pre-processing steps.[4].

3.2 Dataset Annotation

Two annotators annotated the data for toxic positivity. The annotators were linguistics students. An annotation workshop was conducted for the annotators where they were sensitized to the topic of toxic positivity through academic works as described in the related works section and examples of toxic positivity. The annotators were then asked to annotate 50 sentences separately and then their annotator agreement was measured and was found to have a Kappa score of 0.72.We used Cohen's Kappa coefficient to calculate Inter Annotator Agreement (Fleiss and Cohen, 1973) . The annotators then discussed their disagreements and came to a better

understanding of the annotation guidelines. They annotated another 50 sentences and got a better Kappa score of 0.76. They again had a discussion about their disagreements. After this exercise, they were told to annotate the dataset separately without communicating with each other. The 100 sentences used for training the annotators were discarded and are not a part of this dataset of 4,250 sentences. It was observed that sentences that had the following general characteristics were marked as toxic positive:

- Encouraging hiding or suppressing negative emotions.

    - Example: "A negative mind will never give you a positive life."

- Encouraging focusing on positivity rather than processing negative emotions.

    - Example: "Every time I hear something negative, I will replace it with a positive thought."

- Minimizing someone's negative feelings.

    - Example: "You cannot be lonely if you like the person you're alone with."

A few categories of sentences or quotes we emerged when were studying the dataset. We decided to annotate for them as well. The categories of the sentences were as follows.

- **Worldview**: sentences that are philosophical, abstract and provide an insight into the worldview of the writer. Example: "Things may come to those who wait, but only the things left by those who hustle"

- **Personal Experience**: sentences that provide insights based on the writer's personal experience. Example: "I always did something I was a little not ready to do. I think that's how you grow. When there's that moment of 'Wow, I'm not really sure I can do this,' and you push through those moments, that's when you have a breakthrough."

- **Advice**: sentences that are more instructional in nature and provide straightforward recommendations and advice. Example: "Do one thing every day that scares you."

---

[1]http://www.brainyquote.com
[2]BeautifulSoup Documentation
[3]Twitter API Documentation
[4]Dataset Link

Class	Number of sentences
Toxic Positive	512
Non-Toxic Positive	3738

Table 1: Distribution of toxic positive and non-toxic positive sentences.

Type of sentence	Number of sentences
Worldview	3128
Advice	709
Personal Experience	253
Affirmation	160

Table 2: Distribution of the various types of sentences occurring in the dataset.

- **Affirmation**: First-person sentences that are used as affirmations. Example: "I choose to make the rest of my life, the best of my life."

The same annotators annotated the categories of sentences as well. The same process of annotating 100 sentences, 50 sentences at a time and discussing disagreements was followed to train the annotators.

3.3 Dataset Statistics

Out of the 4,250 sentences, 512 were annotated as toxic positive, which constitutes 12% of the dataset.The rest of the 3738 sentences were non-toxic positive. Examples of toxic and non-toxic positive sentences are presented in Table 3.

Worldview was the most common category of sentence occurring 73.6% of the time with advice occurring 16.7% of the time and the rest occurring less than 10% of the time in the dataset. Exact figures are presented in Table 2.

It was also seen that 44% of the sentences that belonged to the affirmation category were toxic positive. 21% of the sentences belonging to the advice category were toxic positive, while 14% and 8% of sentences belonging to the personal experience and the worldview category respectively were toxic positive. We noticed that in our dataset, most affirmation sentences were focused on emotion suppression, and hence they were marked as toxic positive. The non-toxic positive affirmations focused on gratitude, having a growth mindset and self-acceptance, although they were fewer in number.

We got a Kappa score of 0.82 for the toxic positivity (toxic or non-toxic) annotation and a Kappa

score of 0.74 for category annotations (worldview, advice, personal experience or affirmation).

4 Methodology

We used the following transfomer-based models for text classification:

- **BERT**: BERT (Devlin et al., 2019) is a transformer encoder with several encoder layers, each with several self-attention heads. It is trained using two tasks, Masked Language Modelling (MLM), and Next Sentence Prediction (NSP). MLM has been shown to help incorporate both the left and the right contexts into the bidirectional embeddings generated. We have fine-tuned the "bert-base-uncased" model in our implementation.

- **RoBERTa**: RoBERTa (Liu et al., 2019) is a transformer-based encoder built by modifying the original BERT architecture. It utilizes more data with longer average sequence lengths and larger batches. It is solely trained on MLM and makes use of dynamic masking (i.e. the set of masked tokens is subject to change while training). It performs better on the GLUE benchmark (Wang et al., 2019a) in comparison to BERT and XLNet. For the classifier, we have fine-tuned the "roberta-base" model.

- **ALBERT**: ALBERT (Lan et al., 2020) is yet another transformer encoder based on BERT but aimed at being lighter than its predecessor. The core parameter reduction methods include factorizing the vocabulary embedding matrix into smaller sub-matrices and utilizing repeating layers distributed across groups for increased parameter sharing. These techniques help reduce the parameter count by almost 80% with minimal changes to the overall performance. We have fine-tuned the "albert-base-v2" model in our implementation.

We also experimented with an ensemble based classifier for which we additionally used the following:

- **XGBoost Random Forest Classifier**: Random Forest Classifiers (Ho, 1995) are widely used for ensemble classification. They consist

Sentence	Class
When people say there is a 'reason' for the depression, they insult the person who suffers, making it seem that those in agony are somehow at fault for not 'cheering up.' The fact is that those who suffer - and those who love them - are no more at fault for depression than a cancer patient is for a tumor.	Non-Toxic Positive
Just like it's not healthy to think overly negative thoughts, exaggeratedly positive thoughts can be equally detrimental. If you overestimate how much of a positive impact a particular change will have on your life, you may end up feeling disappointed when reality doesn't live up to your fantasy.	Non-Toxic Positive
Do what you feel in your heart to be right	Non-Toxic Positive
The secret of getting ahead is getting started.	Non-Toxic Positive
Being positive is like going up a mountain. Being negative is like sliding down a hill. A lot of times, people want to take the easy way out, because it's basically what they've understood throughout their lives.	Toxic Positive
You must not under any pretense allow your mind to dwell on any thought that is not positive, constructive, optimistic, kind.	Toxic Positive
While you're going through this process of trying to find the satisfaction in your work, pretend you feel satisfied. Tell yourself you had a good day. Walk through the corridors with a smile rather than a scowl. Your positive energy will radiate. If you act like you're having fun, you'll find you are having fun.	Toxic Positive
You can't live a positive life with a negative mind and if you have a positive outcome you have a positive income and just to have more positivity and just to kind of laugh it off.	Toxic Positive

Table 3: Examples of toxic positive and non-toxic positive sentences in the dataset.

Removed Text	Source
Check out this new print for SPRING! #SpringForArt #ThisSpringBuyArt #gardeners #gardens #Inspire #InspirationalQuotes	Twitter
A future Metaverse, a social network for the people by the people, around jobs and finance in the decentralised world.Tomorrow's job fair in 3 dimensions at your fingertips. #MondayMotivation #cryptocurrency #blockchain #Crypto #jobseeker #Trader #Jobs #trading #ICO	Twitter
The failure of Lehman Brothers demonstrated that liquidity provision by the Federal Reserve would not be sufficient to stop the crisis; substantial fiscal resources were necessary.	BrainyQuote
Museums are managers of consciousness. They give us an interpretation of history, of how to view the world and locate ourselves in it. They are, if you want to put it in positive terms, great educational institutions. If you want to put it in negative terms, they are propaganda machines.	BrainyQuote

Table 4: Examples of the text removed during dataset creation.

Model	Macro Precision	Weighted Precision	Macro Recall	Weighted Recall	Macro F1	Weighted F1
BERT	0.78	0.84	0.6	0.86	0.63	0.83
RoBERTa	0.71	0.85	0.7	0.84	0.68	0.85
ALBERT	0.71	0.83	0.65	0.85	0.67	0.84
Ensemble	**0.76**	**0.85**	**0.69**	**0.86**	**0.71**	**0.85**

Table 5: Classification results of various models used on the dataset.

of a large number of decision trees, each set to only a subset of the overall feature-set of the data. This helps create numerous weak learners with relatively low correlation. The majority verdict of these weak learners tends to outperform an individual predictor tasked with the entire feature-set. In this paper, we have made use of the implementation of the Random Forest Classifier by XGBoost (Chen and Guestrin, 2016).

- **Bayesian Optimization**: Bayesian Optimization (Mockus, 1989) is a sequential global optimization strategy for various black-box functions and is used for models across Machine Learning. It attempts to determine the prior distribution of the system (i.e model hyperparameters), which yields the optimal posterior distribution (i.e objective function) by iteratively testing the prior and updating the posterior accordingly. It provides a more computationally efficient yet fine-grained search space than more exhaustive methods such as grid search. In our work, Bayesian optimization is used for tuning the hyperparameters (i.e. number of tree estimators, train subsample ratio, and column subsample ratio) of the Random Forest Classifier. We make use of the implementation by the bayesian-optimization Python library (Fernando, 2014).

5 Experiments and Results

We experimented with 3 transformer models BERT, RoBERTa, and ALBERT. Each of the classification models utilizes a pretrained Transformer encoder, i.e. BERT-Base, RoBERTa-Base, and ALBERT-Base. The pooled output layer from each encoder is passed through respective dropout layers ($p = 0.3$) for further regularization and linear layers (mapping from a vector size of 768 to the number of classification categories, i.e. 2). A softmax function is applied to each of the size-2 vectors for

normalized likelihoods of the two classes. The results from these models are provided in Table 5.

We also experimented with an ensemble-based classifier. The classifier is an ensemble of three predictors with a random forest classifier on top (as shown in Figure 2). The predictors were the three text classification transformer based models as mentioned above.

The likelihoods from each of the predictors were concatenated and passed as features to an XGBoost Random Forest Classifier to generate an ensemble class prediction. After a Bayesian Search for the classifier parameters on the validation set, the number of tree estimators w set to 149, subsample ratio of the training samples to 0.50, and subsample ratio of columns for each split to 0.33.

Each of the Transformer encoder predictors were trained using AdamW optimizer ($\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-8}$), with Cross Entropy loss, using a linear training scheduler. The encoder pipelines were trained with an initial learning rate of $2e^{-5}$ and the XGBoost ensemble classifier with a learning rate of 1.0. The predictors were trained for 6 epochs . The predictions from the epoch with the best validation weighted macro F1 score were utilized for the ensemble classification. The overall batch size for the pipeline was set to 16.

The ensemble model generalized better than the individual models producing the highest macro F1 score of 0.71 and a weighted F1 score of 0.85 as seen in Table 5. As the toxic tweets comprise of only a small portion of the data (14.5%), models performing well on non-toxic tweets tend to have inflated weighted-F1 scores. Therefore we opted for macro-F1 as the main performance metric for this task.

6 Conclusion and Future Work

In this work, we created a dataset for toxic positivity detection. We scraped 4,250 sentences from Twitter and the inspirational quote website BrainyQuote. We then annotated them and
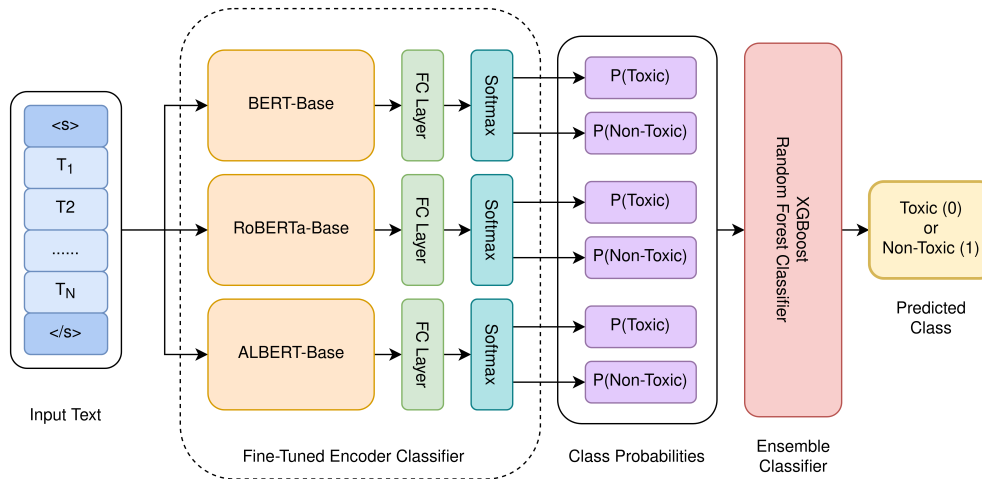
Figure 2: Schematic overview of the architecture of our model.

achieved a Kappa score of 0.82 for toxic positivity classification. We then performed experiments using transformer-based models for text classification. Our ensemble model gave us the best results achieving a macro F1 score of 0.71 and a weighted F1 score of 0.85. As more people turn to social media to get help when they are going through a tough time, it becomes important for them to be able to differentiate between positive and toxic positive messages. Furthermore, being able to recognize toxic positivity is also important for chatbots and other automated systems that aim to provide mental health assistance. We hope that our work contributes to further research in this field. In the future, we plan to extend the study by introducing a larger dataset in English as well as other languages.

References

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 2019. 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*, pages 54–63, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Eva Bosveld. 2021. *Positive Vibes Only: The Downsides of a Toxic Cure-All*.

Laura Campbell-Sills, David H. Barlow, Timothy A. Brown, and Stefan G. Hofmann. 2006. Effects of suppression and acceptance on emotional responses of individuals with anxiety and mood disorders. *Behaviour Research and Therapy*, 44(9):1251–1263.

Bharathi Raja Chakravarthi. 2020. HopeEDI: A multilingual hope speech detection dataset for equality,

diversity, and inclusion. In *Proceedings of the Third Workshop on Computational Modeling of People's Opinions, Personality, and Emotion's in Social Media*, pages 41–53, Barcelona, Spain (Online). Association for Computational Linguistics.

Bharathi Raja Chakravarthi and Vigneshwaran Muralidaran. 2021. Findings of the shared task on hope speech detection for equality, diversity, and inclusion. In *Proceedings of the First Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 61–72, Kyiv. Association for Computational Linguistics.

Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 785–794, New York, NY, USA. ACM.

Delia Cioffi and James Holloway. 1993. Delayed costs of suppressed pain. *Journal of Personality and Social Psychology*, 64(2):274–282.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Nogueira Fernando. 2014. Bayesian Optimization: Open source constrained global optimization tool for Python.

Joseph L. Fleiss and Jacob Cohen. 1973. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and Psychological Measurement*, 33(3):613–619.

Brett Q. Ford, Phoebe Lam, Oliver P. John, and Iris B. Mauss. 2018. The psychological health benefits of accepting negative emotions and thoughts: Laboratory, diary, and longitudinal evidence. *Journal of Personality and Social Psychology*, 115(6):1075–1092.

Kris Gowen, Matthew Deschaine, Darcy Gruttadara, and Dana Markey. 2012. Young adults with mental health conditions and social networking websites: Seeking tools to build community. *Psychiatric Rehabilitation Journal*, 35(3):245–250.

James J. Gross and Oliver P. John. 2003. Individual differences in two emotion regulation processes: Implications for affect, relationships, and well-being. *Journal of Personality and Social Psychology*, 85(2):348–362.

Tin Kam Ho. 1995. Random decision forests. In *Proceedings of 3rd international conference on document analysis and recognition*, volume 1, pages 278–282. IEEE.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. Albert: A lite bert for self-supervised learning of language representations.

Margo Lecompte-Van Poucke. 2022. "you got this!": A critical discourse analysis of toxic positivity as a discursive construct on facebook. *Applied Corpus Linguistics*, 2(1):100015.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

Thomas Mandl, Sandip Modha, Anand Kumar M, and Bharathi Raja Chakravarthi. 2020. Overview of the hasoc track at fire 2020: Hate speech and offensive language identification in tamil, malayalam, hindi, english and german. In *Forum for Information Retrieval Evaluation*, FIRE 2020, page 29–32, New York, NY, USA. Association for Computing Machinery.

Donatella Marazziti, Paolo Cianconi, Federico Mucci, Lara Foresi, Chiara Chiarantini, and Alessandra Della Vecchia. 2021. Climate change, environment pollution, covid-19 pandemic and mental health. *Science of The Total Environment*, page 145182.

Jonas Mockus. 1989. *Bayesian approach to global optimization*. Kluwer Academic.

Ioannis Mollas, Zoe Chrysopoulou, Stamatis Karlos, and Grigorios Tsoumakas. 2022. Ethos: a multi-label hate speech detection dataset. *Complex Intelligent Systems*.

Nedjma Ousidhoum, Zizheng Lin, Hongming Zhang, Yangqiu Song, and Dit-Yan Yeung. 2019. Multilingual and multi-aspect hate speech analysis. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4675–4684, Hong Kong, China. Association for Computational Linguistics.

Shriphani Palakodety, Ashiqur R. KhudaBukhsh, and Jaime G. Carbonell. 2020. Hope speech detection: A computational analysis of the voice of peace. *ECAI 2020*, page 1881–1889.

Laura Sokal, Lesley Eblie Trudel, and Jeff Babb. 2020. It's okay to be okay too. why calling out teachers' "toxic positivity" may backfire. *EdCan*, 60.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019a. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In the Proceedings of ICLR.

Bin Wang, Yunxia Ding, Shengyan Liu, and Xiaobing Zhou. 2019b. Ynu_wb at HASOC 2019: Ordered neurons LSTM with attention for identifying hate speech and offensive language. In *Working Notes of FIRE 2019 - Forum for Information Retrieval Evaluation, Kolkata, India, December 12-15, 2019*, volume 2517 of *CEUR Workshop Proceedings*, pages 191–198. CEUR-WS.org.

Ziqi Zhang, David Robinson, and Jonathan Tepper. 2018. Detecting hate speech on twitter using a convolution-gru based deep neural network. In *The Semantic Web*, pages 745–760, Cham. Springer International Publishing.