|  |
| --- |
| **Low-Cost Methods for Increasing Classification Accuracy** |
|  |
| **Katharine Sadowski1, Linda Wang2**  **1 Department of Policy Analysis and Management, Cornell University**  **2 Department of Economics, Cornell University**  [**ks2373@cornell.edu**](mailto:ks2373@cornell.edu)**, lw692@cornell.edu** |
|  |
|  |
|  |
|  |

Abstract

Classification is a popular task in Natural Language Processing that has large policy implications, especially when it involves controversial issues like toxicity. The factors that determine whether a comment is toxic or not have been hotly contested in recent years, and their identification is exacerbated by the rapid evolution of the language of toxicity itself. These difficulties are at least somewhat reflected in language models’ performances on toxicity classification. State-of-the-art models, such as BERT which is the focal model of this paper, perform various classification tasks (including toxicity classification) with close to 100% accuracy; however, most models are still hovering below this benchmark. We explore low-cost solutions that could improve accuracy for researchers interested in toxicity classification and, potentially, classification tasks more generally.

Introduction

Every country in the world faces some amount of hate speech online, though there is no universal definition of what is hate speech. This makes identifying toxic or hateful comments difficult (Laub, 2019). What’s more, this is a very prevalent issue with over half of Americans saying they’ve experienced hate online (Guynn, 2019). With the rise of online platforms that allow users to easily generate anonymous posts, hateful and toxic language has become more and more prevalent. As this behavior takes off, platforms such as Facebook, Twitter, and Reddit have increased their efforts to moderate toxic comments through a variety of methods. Reddit uses volunteer human moderators who they endow with the ability to assess whether comments go against community guidelines and remove said comments (Reddit, 2020). However, this has its downsides with some users upset with the wide variation among moderators’ rules for removing comments and moderators themselves upset the company not paying them for their labor (Placket, 2018). Facebook, on the other hand, has begun using Natural Language Processing (NLP) models such as BERT to classify language to determine whether comments go against community standards (Schroepfer, 2019). This allows the platform to assign fewer internal individuals to the task of identifying toxic or hateful speech, thereby saving the company money. However, these methods are limited in what their accuracy and are especially poor at identifying sarcasm, misspellings, and other common features of language (Joshi et al., 2016).

Outside of large corporations, other groups such as political organizations, who have set up their own websites or Facebook pages and allow for comments also suffer from the burden of removing hateful or toxic comments (Kalsnes & Ihlebae, 2021). This again costs time, money, and manpower, which is especially harmful for smaller groups with finite resources.

While devoting staff/volunteers or employing NLP methods to moderating online platforms have their advantages, these methods are still not 100% accurate and can lead to bias. We are interested in investigating alternative, low-cost methods for improving the classification accuracy of toxic or hateful speech. We hypothesize that combining both human annotation and BERT can increase accuracy with little effort.

Our ideas for this paper are influenced by a variety of other projects that have looked at the payoffs of using either hand annotation or other, more technical methods for improving classification accuracy. For example, a number of researchers have worked on supplementing BERT with their own encodings (Koufakou et al., 2020). Others have crafted their own examples based on common heuristics that often lead to misclassification (McCoy et al., 2019). While these types of methods can potentially do well, they also require a lot of expertise and time due to the need to develop datasets to supplement the relevant NLP methods. To address these trade-offs, some researchers have attempted to limit the amount of hand annotation required by moving to an online learning problem; for such problems, the goal is to maintain a high classification accuracy while recognizing that language, especially that of hate speech, evolves rapidly. Their methods have focused on ways to determine emerging words in the literature in order to improve training (Lazaridou et al., 2021).

Given the alarming number of reports stating online hate speech has been increasing across the globe (OHCHR, 2021), we feel the task of identifying toxic or hateful comments is becoming increasingly important for today’s policy-makers. Our goal is to create low-cost solutions for increasing toxicity classification accuracy by using a simple algorithm that will hopefully increase accuracy while limiting the need to hiring additional linguistic personnel/moderators. Our proposed solutions build upon the aforementioned models, starting with a very simple method of determining linguistic differences between training and test datasets, and then augmenting the results with limited expert annotations. Unfortunately, our current results suggest that our proposed solutions do not enhance accuracy. In fact, all models perform the same irrespective of which solution is used. More exploration is needed to understand what makes certain examples harder to classify.

Dataset Construction

Hateful and toxic speech have a wide variety of definitions depending on who you ask or where you are located in the world. For our study, we limit our definition of hate speech to comments that are identified as toxic based on hand annotators. Our task uses the jigsaw\_toxicity\_pred dataset from HuggingFace.[[1]](#footnote-1) This dataset contains comments from Wikipedia that were hand-labeled for being toxic, severely toxic, obscene, threatening, insulting, or related to identity hate. These categorizations are not mutually exclusive, so a comment could be labelled both toxic and obscene. For our experiments, we focus solely on whether a comment was labelled toxic. We limit our experiment to two-class classification because we believe noting whether something is toxic or not is an important fundamental issue for companies focused on siphoning out the harmful comments on their platforms. While it would be nice to label the nuances of a given toxic comment, we don’t believe that this is as essential to companies trying to determine whether a comment should be removed for violating community standards.

Our dataset had been pre-split into training and testing datasets, with the former containing 159,571 examples and the latter 63,978 examples. 9.5% of the training examples are labelled toxic. Table 1 shows examples of a non-toxic and toxic example after pre-processing. Our pre-processing involves lower-casing the examples and removing all special characters (e.g., ascii characters), and numbers from them.[[2]](#footnote-2)

Methods

|  |  |
| --- | --- |
| **Label** | **Example Text** |
| Non-Toxic | congratulations from me as well, use the tools well |
| Toxic | cocksucker before you piss around on my work |

Table : Example text from dataset.

Our proposed solutions rely on using the test dataset to augment the training dataset to improve accuracy. We explore a combination of methods that include using a simple mathematical formula to determine which examples are more “unique” to the testing versus training datasets and simulating hand annotation. The former method uses log-odds ratios, while the latter relies on human expertise.

Our prior is that while the examples in the testing dataset are pulled iid from the universe of examples comprising the training and testing datasets, there could be patterns in the testing dataset that are not necessarily represented in the training dataset. Thus, we want to understand which examples are more representative of the testing dataset than the training dataset. In order to do this, we introduce log-odds ratios (Monroe et al., 2008). This method works by creating counts of all the words in the training and testing datasets, which are then converting into proportions (1). For words that exist in one dataset but not the other, we use plus one smoothing. Next, we convert these into odds to determine the odds of word *w* appearing in dataset *i* (2). For this paper, this translates to the odds of word *w* appearing in either the training or test set. This allows us to get a ratio of how likely a word is to appear in one dataset versus another by dividing the odds for the training and the test dataset. We then log the ratio in to make the numbers symmetric between the two groups (3).

(1)

(2)

(3)

Using (3), we then apply this method to compare the language in the training and testing datasets. In our case, the most negative words are the words most unique to the testing dataset. We extract these words and use them in our experiments to determine which examples we should pull from test and place into train. Our list of the top 5 words and their log-odds can be found in Table 2.

|  |  |
| --- | --- |
| **Log-odds** | **Vocabulary** |
| -41.19 | poop |
| -37.26 | dicks |
| -35.81 | gay |
| -34.45 | youfuck |
| -32.51 | n\*\*\*\*\* |

Table 2: Most unique words in test dataset as compared to training dataset using log-odds ratios.

Our second method is more intuitive but relies on human expertise, which can be costly. This method requires experts to evaluate the predicted testing examples and determine which of the testing examples are misclassified in order to use those examples to supplement the training dataset. We hypothesize that taking 20 random examples from testing and placing them into training is not going to improve accuracy much because the 20 examples could happen to be the easiest for the model to classify. Instead, we believe that taking 20 difficult examples, measured by whether the model classified them incorrectly, is likely more beneficial. Moreover, if our hypothesis is true, then as test datasets become larger and accuracy becomes higher, our second method will become more costly, as it will take human experts longer to determine the incorrectly classified examples.

Experimental Set-up

Following the methods described above, we run the following 5 experiments. These experiments use the log-odds method to determine which examples we should augment our training examples with, as well as some simulations of hand annotation. We have designed our experiments to move from purely automated (i.e. lowest-cost) adjustments to a combination of automated and hand-annotated (i.e., higher-cost) adjustments. The hope is the lower cost methods will be able to perform better than baseline without trading off too much relative to a more costly approach.

**DistilBERT:** We use DistilBERT, a smaller, faster version of the Bidirectional Encoder Representations from Transformers model (Victor et al., 2020) for our classification task. This is a transformer model that uses attention to process sentences. We use the distilbert-base-uncased pre-trained model from HuggingFace and fine-tune on the jigsaw\_toxicity\_pred dataset. We use the following hyperparameters:

* Epochs: 3
* Batch size: 16
* Learning rate: 2e-5
* Optimizer: Adam
* Weight decay: 0.01.

We used DistillBERT for our experiments due to time constraints, though we did run BERT in a select few experiments and noticed no difference in accuracies for the baseline model.

**Experiment 0 (baseline):** We run DistillBERT on the pre-processed dataset and determine what the overall accuracy is for our model. We then pull the misclassified examples from testing, which will be used in experiments 3 and 4 to simulate hand annotation.

**Experiment 1:** We use log-odds ratios to determine the 20 most unique words in the testing dataset when compared to the training dataset. We then shuffle the testing dataset and iterate through our list of unique words. For each word in our list, we pull one example from the testing dataset that contains the word and then remove the word from our list. We remove from the testing dataset the 20 examples we picked and place them into the training dataset. We rerun DistillBERT on the augmented training dataset and evaluate using the new, smaller testing dataset. We run a control model that uses the same testing dataset (i.e. with the same 20 examples removed), but with a training dataset augmented with 20 examples pulled from classic English novels (see Appendix 2). We use these arbitrary training examples to ensure the model is not performing better just because of the addition of new examples into the training dataset.

**Experiment 2:** We repeat experiment 1, but we now limit the 20 most unique words to the 20 most unique nouns. We chose nouns because prior research has shown that nouns are the most likely to evolve over time, thus we believe nouns could be a very important reason for why models are more likely to misclassify certain examples (Lazaridou et al., 2021). We then run a control experiment that is analogous to the control experiment in experiment 1. We use the same control training dataset but with the testing dataset used here in experiment 2.

**Experiment 3:** We simulate a situation in which we provide an “expert annotator” the test dataset with predicted labels. Their job would be to determine the first 20 misclassified examples, which are then taken from test and placed into train. We do this by taking a random 20 misclassified examples from test and placing them into the training dataset. We rerun DistillBERT on the new training dataset and evaluate using the new testing dataset. We then run the same control experiment outlined in experiment 1, with the same control training dataset but with the testing dataset used here in experiment 3.

**Experiment 4:** We add an additional layer to experiment 1, where we simulate a situation where a “hand annotator” is first given the list of 20 unique words. They are then given testing examples with predicted labels. The examples they are given are limited to only those examples containing the words from the provided list of 20. The dataset is sorted so that all examples containing each word are grouped together to make reading through the datasets easier. Their job is to read through the examples for each word and determine if the label is correct or not. For each word, once they determine the first incorrectly labelled comment for a word, they tag that comment and then move to the next word on the list. The 20 examples they tag are then removed from the test dataset and added to the training dataset. We simulate this process by pulling the first example of each word of our word list in the set of misclassified examples from experiment 0. We remove these examples from the testing dataset and place them into train. We rerun DistillBERT on the new training dataset and evaluate using the new testing dataset. We then run the same control experiment outlined in experiment 1, with the same control training dataset but with the testing dataset used here in experiment 4.

**Implementation and Evaluation:** We repeat experiments 1 through 4 using 50 examples as opposed to 20 to see if the number of training examples meaningfully affects accuracy. For each model we report the final accuracy of the model after evaluating it on the test dataset. Appendix 2 includes example logic for each of the experiments including sample examples.

Results

Our results suggest that while our proposed solutions yield high accuracy on toxicity classification, none substantially improved accuracy beyond our baseline model. Model accuracies consistently hovered around 93%, as shown in Table 3. Most surprisingly, there was little to no variation in accuracy between experimental accuracies and their corresponding control accuracies.

To check whether these results were due to the size of our training dataset augmentations in the experiments, we increased the number of examples we added to training and subsequently removed from test from 20 to 100 in all experimental setups and re-ran them as a robustness check. We still did not observe any large changes in accuracy scores across experiments or within experiments (between the experiment and control settings) after this change; in some experiments, the change actually lead to slightly decreased accuracy scores, though the dip appears negligible, as the new scores for those experiments were generally around 92%.

|  |  |
| --- | --- |
| **Misclassified Examples from E4** | |
| False Positive | 'А ви момци мислите да овде баш нико не разуме то што пишете? \n\n : If you are going to insult someone, at least have the guts to do it openly and not behind their backs.' |
| False Negative | 'and F\*\*\* WIKIPEDIA HAHAHAHAHAHAHAAHAHAHAA!!! B\*\*\*\*\*\*!!!!!' |

Table 6: Misclassified examples specific to E4.

|  |  |  |
| --- | --- | --- |
| **Model** | **False**  **Positives** | **False**  **Negatives** |
| E0: DistillBert (Baseline) | 2904 | 1044 |
| E4: Annotation using log-odds | 3925 | 721 |

Table 5: Robustness checks using misclassified example overlap.

We also chose to perform a robustness check specific to experiments 3 and 4. We were concerned over our assumption that misclassified examples remained consistent across models. That is, we wanted to see whether the misclassified examples we pulled from the testing dataset in experiment 0 (and later used in experiments 3 and 4) would have truly been misclassified by the models we ran in experiments 3 and 4. Accordingly, we decided to check the degree of overlap in misclassified examples between experiment 0 and experiment 4, the results of which are reported in Table 4. Note that we chose to use experiment 4 as a comparison point because it is intended to simulate the most costly of our proposed solutions in terms of time of human annotators and computational resources.

We note that there are significant distinctions in misclassified examples between experiments 0 and 4. Overall, experiment 4 led to more false positives and fewer false negatives compared to experiment 0 with the differences accounting for the consistent accuracy scores between experiments (Table 5). As shown in Table 6, there may be some underlying semantic reasons for this, as several of the misclassified examples specific to experiment 4 were either examples in foreign languages or with censored language. We discuss this further in the next section.

Discussion

|  |  |
| --- | --- |
| **Experiment** | **Accuracy**  **N=20** |
| E0: DistillBert (Baseline) | 94 |
| E1: Log-odds | 93 |
| E1: Control | 93 |
| E2: Log-odds Nouns | 93 |
| E2: Control | 93 |
| E3: Annotation | 93 |
| E3: Control | 93 |
| E4: Annotation using log-odds | 93 |
| E4: Control | 93 |

Table 3: Model accuracies for each experiment.

While low-cost methods are an exciting research concept, our methods for augmenting the training dataset using information from the test dataset did not lead to more accurate classification. This result has led us to consider why we might be getting this result and further research questions to probe this task further.

Broadly speaking, our dataset might not be the best for our experimental setup. Our method could be better for an online learning task, or a task where the language is dramatically different between the training and the test dataset. An initial analysis of the most unique words in the test versus the training dataset suggest that all of the words in the test dataset appear in the train dataset just with less frequency. Thus, it is likely the case that the examples we are supplementing the training dataset with are already well represented in the training dataset. Thus, we should consider another task, such as using Wikipedia comments from 2014 versus 2017, or some version of this that allows language to evolve more so that our method is able to pick up more unique differences.

|  |  |
| --- | --- |
| **Model** | **Accuracy**  **N=20** |
| E0: DistillBert (Baseline) | 94 |
| E4: Annotation using log-odds | 93 |

Table 4: Robustness checks using misclassified example overlap.

Provided that our data and methods do form a good match, our results may still be impaired by the heuristic judgements we made under the time constraints of this project. For instance, we did not formally consider why the addition of 20 or 100 new examples to our training dataset could be sufficient to improve model accuracy. Our choices of numbers was driven largely by our intuition of what would be a low-cost alternative. There may, however, be an optimal number of examples needed depending on the original size of the training dataset.

Furthermore, we also assumed there to be a hierarchy between cost and performance. Our experiments are arranged under the assumption that accuracy most either increase or remain the same as we involve more human annotations and or computational resources during the creation of our augmented datasets. We may need to account for this trade-off systematically in future work, to generate experiments with significantly different results.

Relatedly, due to the length of time it took to run each model, we were not able to explore different hyperparameters or add more epochs. It could be the case that our model may have been able to learn from the new examples we added to the training dataset, but we did not test for a sufficient amount of conditions.

Additionally, as previous researchers have shown, BERT models are doing very well and they seem to only mess up on cases that are hard even for human annotators. In particular, sentiment classification often suffers from sarcasm which is difficult for even the most advanced models to identify (Baruah et al., 2020; Joshi et al., 2016). Researchers have also shown that natural language inference tasks are difficult due to models relying on incorrect heuristics (McCoy et al., 2019). Both of these suggest that a very simple model, such as that suggested by us in this paper might not be enough to help with these final more difficult cases. In fact, we might need a slightly more advanced model that can identify certain elements of the text that could be leading to misclassification.

It is important to note that, while our accuracy scores did not change in the directions we expected, the consistency may be driven by underlying differences in how our experimental models classify (or misclassify) particular examples. As mentioned in our results section, experiment 4 produced more false positive misclassifications than did experiment 0, while resulting in fewer false negatives. This is at least promising, as it is reasonable to assume both companies and policy-makers would be more willing to err on the side of flagging/removing potentially toxic comments than the opposite. Thus, if we move our metric from accuracy to lowering false negatives, our methods offer a reasonable, potentially low-cost solution for efficiently identifying more of the truly toxic comments. It is also worth mentioning that we consider experiment 4 to be the most costly of our proposed solutions. However, we believe interested parties would be willing to carefully consider the trade-off, and future work can reconsider the underlying misclassification patterns of our lower-cost experiments, as well.

There are multiple issues for researchers to consider when thinking about augmenting a dataset. We chose our method because our dataset was pulled from one test period with a random training and test split. We believed there would be random differences between the training and test split, our method allowed for the algorithm to decide what those differences were. One could image an online learning setting, where this approach could actually work better. For instance, if the training and test datasets were fundamentally different (e.g., words appearing in one of the datasets were entirely absent from the other) then this algorithm could allow us to get more unique examples into training that could help. While our experiment didn’t result in improved accuracy, a refocus on lower false negative detection with a combination of a few methodological tweaks, our logical framework could help in the development of new low-cost methods for toxic comment classification.

Acknowledgments

Thanks to Sasha Rush and our CS6741 classmates for helping develop our understanding of ML and NLP methods, as well as pushing us on our research questions and methods. We learned a lot and are very eager to push ourselves further.

References

Aditya Joshi, Pushpak Bhattacharyya, Mark J Carman. 2016. *Automatic Sarcasm Detection: A Survey*. arXiv: 1602.03426v2

Alice Tontodimamma, Eugenia Nissi, Annalina Sarra, and Lara Fontanella. 2021. Thirty years of research in hate speech: topics of interest and their evolution. Scientometrics Volume 126, pages 157-179.

Angeliki Lazaridou, Adhiguna Kuncoro, Elena Gribovskaya, Devang Agrawal, Adam Liska, Tayfun Terzi, Mai Gimenez, Cyprien de Masson d'Autume, Sebastian Ruder, Dani Yogatama, Kris Cao, Tomas Kocisky, Susannah Young, and Phil Blunsom. 2021. *Pitfalls of Static Language Modelling.* arXiv: 2102.01951.

Anna Koufakou, Endang Wahyu Pamungkas, Valerio Basile, and Viviana Patti. 2020. *HurtBERT: Incorporating Lexical Features with BERT for the Detection of Abusive Language*. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*. Association for Computational Linguistics, pages 34-43. https://doi.org/ 10.18653/v1/2020.alw-1.5.

Arup Baruah, Kaushik Amar Das, Ferdous Ahmed Barbhuiya, and Kuntal Dey. 2020. *Context-Aware Sarcasm Detection Using BERT*. In the *Proceedings of the Second Workshop on Figurative Language Processing*. Association for Computational Linguistics, pages 83-87. https://doi.org/10.18653/v1/P17.

Bente Kalsnes and Karoline Andrea Ihlebaek. 2021. *Hiding hate speech: political moderation on Facbeook.* Media, Culture, & Society. Volume 43(2), pages 326-342. https://doi.org/10.1177/0163443720957562.

Burt L. Monroe, Michael P. Colaresi, and Kevin M. Quinn. 2008. *Fightin’ Words: Lexical Feature Selection and Evaluation for Identifying the Content for Political Conflict.* Political Analysis Volume 16, pages 372-403. https://doi.org/10.1093/pan/mpn018.

Jessica Guynn. 2019. *If you’ve been harassed online, you’re not alone. More than half of Americans say they’ve experienced hate.* USA Today. https://www.usatoday.com/story/news/2019/02/13/study-most-americans-have-been-targeted-hateful-speech-online/2846987002/

Mike Schroepfer. 2019. *Community Standards report*. <https://ai.facebook.com/blog/community-standards-report/>

Ben Plackett. (2018). Unpaid and abused: *Moderators speak out against Reddit.* https://www.engadget.com/2018-08-31-reddit-moderators-speak-out.html

R. Thomas McCoy, Ellie Pavlick, and Tal Linzen. 2019. *Right for the Wrong Reasons: Diagnosing Syntatic Heuristics in Natural Language Inference.* arXiv: 1902.01007v4

Reddit. *Transparency Report 2020*. <https://www.redditinc.com/policies/transparency-report-2020-1>

OHCHR. 2021. *Report: Online hate increasing against minorities, says expert*. <https://www.ohchr.org/EN/NewsEvents/Pages/sr-minorities-report.aspx>

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2020. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXIV:1910.01108v4

Zachary Laub. 2019. *Hate Speech on Social Media: Global Comparisons.* Council on Foreign Relations. <https://www.cfr.org/backgrounder/hate-speech-social-media-global-comparisons>

2. Appendices

Appendix A1: Data Augmentation Examples for Each Experiment

|  |  |  |
| --- | --- | --- |
| **Experiment** | **Procedure** | **Example iteration** |
| 1 | Step 1: Determine 20 most unique words in test dataset using log-odds ratios  Step 2: Iterate through list finding one example in the test dataset containing that word | Step 1: First word in list, **poop**  Step 2: Find word in an example in the test dataset  [Example (abbreviated): ….yeswhen you **poop** u eat fried chicken rydaylife doo doo poo poo….]  Step 3: Repeat steps 1 and 2 for each word in list |
| 2 | Step 1: Determine 20 most unique NOUNS in test dataset using log-odds ratios  Step 2: Iterate through list finding one example in the test dataset containing that noun | Step 1: First word in list, poop  NOTE: the first word is a noun, thus this is the same as experiment 1  Step 2: Find word in an example in the test dataset  NOTE: We shuffle the dataset, thus this is different from experiment 1  [Example (abbreviated): Van Slyke was the first major league player to **poop** his pants on field]  Step 3: Repeat steps 1 and 2 for each word in list |
| 3 | Step 1: Determine the first 20 misclassified examples | Step 1: Determine first misclassified example  [Example (abbreviated): I WILL BURN YOU TO HELL IF YOU REVOKE MY TALK PAGE ACCESS!!!!!!!!!!!!!]  Step 2: Repeat until 20 misclassified examples obtained |
| 4 | Step 1: Determine 20 most unique words in test dataset using log-odds ratios  Step 2: Iterate through the list of words and find the first misclassified example in the test dataset | Step 1: First word in list, poop  Step 2: Find word in an example test dataset, determine if the example is misclassified, if misclassified then remove the example from test and place in train and remove word from list, else continue searching for misclassified example  [Example (abbreviated): pee u **poop** is the most disgusting thing in the whole entire world!!! Every time i see it i want to puke!!! you people are so gross to talk about it…]  Step 3: Repeat steps 1 and 2 until examples are pulled for all words. |

Appendix A2: Control text added to supplement the training examples for the control experiments

|  |  |  |
| --- | --- | --- |
| **Author** | **Book** | **Comment** |
| Louisa May Alcott | Little Women | Christmas won't be Christmas without any presents, grumbled Jo, lying on the rug. |
| Stephen Crane | Maggie: A Girl of the Streets | When a child, playing and fighting with gamins in the street, dirt disguised her. Attired in tatters and grime, she went unseen. |
| Frances Harper | Lola Leroy | Oh, sho, chile, said Linda, "I can't read de newspapers, but ole Missus' face is newspaper nuff for me. I looks at her ebery mornin' wen she comes inter dis kitchen. Ef her face is long an' she walks kine o' droopy den I thinks things is gwine wrong for dem. But ef she comes out yere looking mighty pleased, an' larffin all ober her face, an' steppin' so frisky, den I knows de Secesh is gittin' de bes' ob de Yankees. Robby, honey, does you really b'lieve for good and righty dat dem Yankees is got horns?" |
| E.M. Forster | A Room With a View | Miss Bartlett was startled. Generally at a pension people looked them over for a day or two before speaking, and often did not find out that they would "do" till they had gone. She knew that the intruder was ill-bred, even before she glanced at him. He was an old man, of heavy build, with a fair, shaven face and large eyes. There was something childish in those eyes, though it was not the childishness of senility. What exactly it was Miss Bartlett did not stop to consider, for her glance passed on to his clothes. These did not attract her. He was probably trying to become acquainted with them before they got into the swim. So she assumed a dazed expression when he spoke to her, and then said: "A view? Oh, a view! How delightful a view is!" |
| Mary Shelley | Frankenstein | “Devil, cease; and do not poison the air with these sounds of malice. I have declared my resolution to you, and I am no coward to bend beneath words. Leave me; I am inexorable.” |
| Virginia Woolf | Mrs Dalloway | I am in love, he said, not to her however, but to some one raised up in the dark so that you could not touch her but must lay your garland down on the grass in the dark. |
| Gustave Flaubert | Madame Bovary | “They smell the dead,” replied the priest. “It’s like bees; they leave their hives on the decease of any person.” |
| Thomas Hardy | Tess of the d'Urbervilles | The farmer and his wife were in the field at the moment of his visit, and Clare was in the rooms alone for some time. Inwardly swollen with a renewal of sentiment that he had not quite reckoned with, he went upstairs to her chamber, which had never been his. The bed was smooth as she had made it with her own hands on the morning of leaving. The mistletoe hung under the tester just as he had placed it. Having been there three or four weeks it was turning colour, and the leaves and berries were wrinkled. Angel took it down and crushed it into the grate. Standing there, he for the first time doubted whether his course in this conjecture had been a wise, much less a generous, one. But had he not been cruelly blinded? In the incoherent multitude of his emotions he knelt down at the bedside wet-eyed. "O Tess! If you had only told me sooner, I would have forgiven you!" he mourned. |
| Kate Chopin | Awakening | Many of them were delicious in the role; one of them was the embodiment of every womanly grace and charm. If her husband did not adore her, he was a brute, deserving of death by slow torture. Her name was Adele Ratignolle. There are no words to describe her save the old ones that have served so often to picture the bygone heroine of romance and the fair lady of our dreams. There was nothing subtle or hidden about her charms; her beauty was all there, flaming and apparent: the spun-gold hair that comb nor confining pin could restrain; the blue eyes that were like nothing but sapphires; two lips that pouted, that were so red one could only think of cherries or some other delicious crimson fruit in looking at them. She was growing a little stout, but it did not seem to detract an iota from the grace of every step, pose, gesture. One would not have wanted her white neck a mite less full or her beautiful arms more slender. Never were hands more exquisite than hers, and it was a joy to look at them when she threaded her needle or adjusted her gold thimble to her taper middle finger as she sewed away on the little night-drawers or fashioned a bodice or a bib. |
| Jane Austen | Pride and Prejudice | Why, yes--because he chose to marry. As a bachelor he might possibly have got into the right circles, though his character would in any case have made it difficult for him to curry favour. |
| Charles Dickens | Bleak House | England has been in a dreadful state for some weeks. Lord Coodle would go out, Sir Thomas Doodle wouldn't come in, and there being nobody in Great Britain (to speak of) except Coodle and Doodle, there has been no government. It is a mercy that the hostile meeting between those two great men, which at one time seemed inevitable, did not come off, because if both pistols had taken effect, and Coodle and Doodle had killed each other, it is to be presumed that England must have waited to be governed until young Coodle and young Doodle, now in frocks and long stockings, were grown up. This stupendous national calamity, however, was averted by Lord Coodle's making the timely discovery that if in the heat of debate he had said that he scorned and despised the whole ignoble career of Sir Thomas Doodle, he had merely meant to say that party differences should never induce him to withhold from it the tribute of his warmest admiration; while it as opportunely turned out, on the other hand, that Sir Thomas Doodle had in his own bosom expressly booked Lord Coodle to go down to posterity as the mirror of virtue and honour. |
| George Eliot | Middlemarch | "It is so painful in you, Celia, that you will look at human beings as if they were merely animals with a toilet, and never see the great soul in a man's face." |
| Mary Russell Mitford | Our Village | Then comes the village shop, like other village shops, multifarious as a bazaar; a repository for bread, shoes, tea, cheese, tape, ribands, and bacon; for everything, in short, except the one particular thing which you happen to want at the moment, and will be sure not to find. The people are civil and thriving, and frugal withal; they have let the upper part of their house to two young women (one of them is a pretty blue-eyed girl) who teach little children their A B C, and make caps and gowns for their mammas,--parcel schoolmistress, parcel mantua-maker. I believe they find adorning the body a more profitable vocation than adorning the mind. |
| Virginia Woolf | Mrs Dalloway | For they might be parted for hundreds of years, she and Peter; she never wrote a letter and his were dry sticks; but suddenly it would come over her, If he were with me now what would he say?--some days, some sights bringing him back to her calmly, without the old bitterness; which perhaps was the reward of having cared for people; they came back in the middle of St. James's Park on a fine morning--indeed they did. But Peter--however beautiful the day might be, and the trees and the grass, and the little girl in pink-- Peter never saw a thing of all that. He would put on his spectacles, if she told him to; he would look. It was the state of the world that interested him; Wagner, Pope's poetry, people's characters eternally, and the defects of her own soul. How he scolded her! How they argued! She would marry a Prime Minister and stand at the top of a staircase; the perfect hostess he called her (she had cried over it in her bedroom), she had the makings of the perfect hostess, he said. |
| Robert Louis Stevenson | Treasure Island | Of Silver we have heard no more. That formidable seafaring man with one leg has at last gone clean out of my life; but I dare say he met his old Negress, and perhaps still lives in comfort with her and Captain Flint. It is to be hoped so, I suppose, for his chances of comfort in another world are very small. |
| George Gissing | New Grub Street | Do you know, Mr Yule, that you have suggested a capital idea to me? If I were to take up your views, I think it isn't at all unlikely that I might make a good thing of writing against writing. It should be my literary specialty to rail against literature. The reading public should pay me for telling them that they oughtn't to read. I must think it over.' |
| Joseph Conrad | Heart of Darkness | This devoted band called itself the Eldorado Exploring Expedition, and I believe they were sworn to secrecy. Their talk, however, was the talk of sordid buccaneers: it was reckless without hardihood, greedy without audacity, and cruel without courage; there was not an atom of foresight or of serious intention in the whole batch of them, and they did not seem aware these things are wanted for the work of the world. To tear treasure out of the bowels of the land was their desire, with no more moral purpose at the back of it than there is in burglars breaking into a safe. Who paid the expenses of the noble enterprise I don't know; but the uncle of our manager was leader of that lot. |
| Benjamin Disraeli | Sybil | Well, I will write, said Lady Marney; "though I cannot admit it is any favour. Perhaps it would be better that you should see him first. I cannot understand why he keeps so at the Abbey. I am sure I found it a melancholy place enough in my time. I wish you had gone down there, Charles, if it had been only for a few days." |
| E.M. Forster | A Room With a View | Then make my boy think like us. Make him realize that by the side of the everlasting Why there is a Yes--a transitory Yes if you like, but a Yes. |
| Anthony Trollope | The Way We Live Now | He had come to Beccles lately, and Roger Carbury had found out that he was a gentleman by birth and education. Roger had found out also that he was very poor, and had consequently taken him by the hand. The young priest had not hesitated to accept his neighbour's hospitality, having on one occasion laughingly protested that he should be delighted to dine at Carbury, as he was much in want of a dinner. He had accepted presents from the garden and the poultry yard, declaring that he was too poor to refuse anything. The apparent frankness of the man about himself had charmed Roger, and the charm had not been seriously disturbed when Father Barham, on one winter evening in the parlour at Carbury, had tried his hand at converting his host. "I have the most thorough respect for your religion," Roger had said; "but it would not suit me." The priest had gone on with his logic; if he could not sow the seed he might plough the ground. This had been repeated two or three times, and Roger had begun to feel it to be disagreeable. But the man was in earnest, and such earnestness commanded respect. And Roger was quite sure that though he might be bored, he could not be injured by such teaching. Then it occurred to him one day that he had known the Bishop of Elmham intimately for a dozen years, and had never heard from the bishop's mouth,--except when in the pulpit,--a single word of religious teaching; whereas this man, who was a stranger to him, divided from him by the very fact of his creed, was always talking to him about his faith. Roger Carbury was not a man given to much deep thinking, but he felt that the bishop's manner was the pleasanter of the two. |

1. https://huggingface.co/datasets/jigsaw\_toxicity\_pred [↑](#footnote-ref-1)
2. <https://github.com/kcsadow/Low-Cost-Methods-BERT> [↑](#footnote-ref-2)