AmITheClassifier
Advancement: Title Generation with Ethics and Norms

Yotam Golan
ygolan@uci.edu

Samuel Perelgut
sperelgu@uci.edu

Jonathan Dedinata
jdedinat@uci.edu

Abstract

Text Summarization is the process of summarizing the important details of a large body of text for easier consumption. Over time, language models have evolved substantially in their ability to intake text bodies, extract the most relevant information, and present a concise summary to the reader. These summarizations take the form of either extractive, which identifies the important details and presents them, and abstractive, which instead interprets their context and reproduces them in a new manner. We attempt to leverage abstractive text summarization to generate novel, realistic titles based on moral and ethical dilemmas presented to an online community.

1 Introduction

Manual text summarization is a costly and time intensive operation, inherently vulnerable to biases and difficult to scale. However, as the availability of unstructured data continues to expand, especially online, the need for textual summarization has only increased and with it the demands for better and more scalable options. This has led to automatic text summarization being one of the most important applications of NLP, with significant research invested in advancing the field. Text summarization attempts to automatically generate concise summaries of otherwise lengthy texts, while maintaining overall meaning and clarity. This allows for quicker access to vast quantities of information, improve readability, categorize otherwise complicated texts, or simply reduce the time needed to read an article.

In papers such as Scruples (Lourie et al., 2020) and Delphi (Jiang et al., 2021), the moral quandaries of the r/AmITheAsshole posts are leveraged to attempt to teach a model to make qualitative decisions on ethical questions. These solutions seek to remove the human component of the judgement and automate the process of ethical deliberation. There are several potential issues inherent in this approach, including transferring the inherent biases of human decisions into an outwardly unbiased algorithm and the inscrutability of the rendered judgement. Indeed, we view human ethical judgements as an important facet of modern society and instead seek to extend the produced research to assist in judgements rather then supplant them.

First impressions of texts can create out-sized impacts on the following text interpretations and judgement (Wiley and Rayner, 2000). Thus text titles are an important primer for the rest of the reading and are an important tool for text summarization. This is exacerbated in the context of judging situations on their merit and assigning appropriate blame, with titles helping establish the context for the reader, and can sway opinions based on embedded biases. Thus, in the context of r/AmITheAsshole, titles can have an important effect on who the community deems at fault, and to what extent they ascribe blame. It is thus important for titles to accurately summarize the text and maintain as much meaning as possible in as few words as possible.

In our approach, we aim to use Seq2Seq-based architectures along with different types of Transformers for generating titles. As opposed to the Delphi and Scruples paper, our title generating model will not just be a black box which passes judgement. Instead it will preserve the human element of the ethical process, while increasing overall efficiency as good title will assist in the proper direction of human attention. We will mainly be exploring Facebook's BART (Lewis et al., 2019) model and Google's BigBirdPegasus (Zaheer et al., 2021) transformer models in our text summarization tasks. We will be evaluating the models on the BLEU and METEOR scores of the generated sentences and manually determine the sentences if they convey the proper information from the text.

In our experiments, using pre-trained transformer models showed promising results, with BART being the best performing transformer model. On the other hand, we faced many challenges with the custom model we worked on, as it often produced illegible or duplicates sentences. In spite of this, the overall results is a success; our model successfully generates titles which succinctly convey the proper context and semantics of the whole text body. Thus our model expedites the process of humans passing judgement on these dilemmas.

2 Related Works

Our approach to title generation for r/AmITheAsshole posts attempts to preserve the ethical and moral quandaries of the given dilemmas, while providing a short and meaningful summarization. The purpose of this title is to assist readers in identifying relevant or interesting dilemmas, and help filter out lower quality or repetitive content. In this, our work is similar to the title generation for StackOverflow done in CCBert (Zhang et al., 2021). CCBert attempts to generate descriptive titles for StackOverflow questions, also in an attempt to improve human readability and ease of filtering out lower quality submissions. CCBert, however, targets a bimodal combination of text and code, being geared towards coding questions as opposed to ethical judgements. Other than CCBert, most research mainly focuses on text summarization, which does not translate well to title generation since titles are generally standalone from the body text but is still a significant part of texts for contextual meaning.

The work done by (Mane et al., 2020) similarly seeks to generate descriptive titles for easier human accessibility. It relies upon a combination of generative and abstractive summarizations techniques to create more easily understood product titles. However, this paper focuses on verbal communication via smart assistants rather then written word. This, and its texts being product titles often generated by other models curtails its usage for pure long form human generated texts discussing concrete events.

Delphi and Scruples are the two papers with perhaps the closest relations to our work. Delphi is an attempt at creating a neural model that provides ethical and moral judgements when presented with various dilemmas, and do so in a way that is 'human' and empathetic (Jiang et al., 2021). Scruples instead focuses on posts from r/AmITheAsshole, and from the text body attempts to prescribe who in the story bears the brunt of responsibility (Lourie et al., 2020). Unlike our approach, both these systems are largely opaque and give no additional information to the end user regarding their decisions and their reasoning's. Their decisions are also at times questionable or problematic, with Delphi warning users that "Model outputs should not be used for advice for humans, and could be potentially offensive, problematic, or harmful". Our model, thus, instead seeks to be a tool to assist in filtering out low quality posts and enabling more efficient human judgements, rather then attempting to place moral considerations in the hand of black-boxed machines.

3 Approach

3.1 Dataset

The input format for the dataset we are training our model on is a subset of the scruples corpus. As mentioned before, the scruples dataset features moral dilemmas from the subreddit r/AmITheAsshole. In this subreddit, users submit recent experiences they had where the poster, and normally at least one other party, exhibit morally dubious behavior. The user explains the situation and then asks for feedback. An individual member of the the data set contains a body, title, moral score, and meta data. For our purposes we only concern ourselves with the former two, as we will be using the body as an input and the title as a label. The original size of the training dataset was 27766 entries. However, we performed data exploration to see if we could remove any outliers. In particular, significantly longer labels or bodies would hurt performance by adding additional complexity so we hoped to remove them if possible. Thus, we plotted a histogram of the length of the titles in Figure 1 and the length of the bodies in Figure 2. Based on these results, we decided on only allowing our model to generate titles of length 11 before precursors to keep our model from making strange sentences while still capturing the essence of title generation. Additionally, we decided to exclude all training data with a body of length 750 or more, leaving us with 26421 entries.

To be clear, the structure for the titles available in the dataset is often in the format "AITA for " + verb phrase + prepositional phrase. Further, the verb phrase almost always features a predicate which takes a two predicates the first of which is the "I"

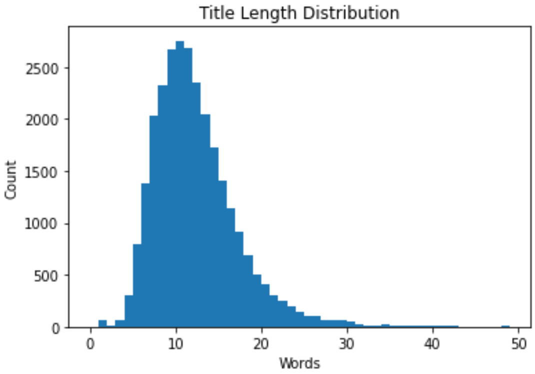

Figure 1: Distribution of word counts in title.

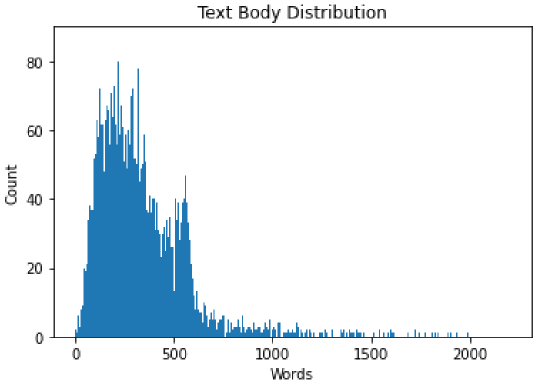

Figure 2: Distribution of word counts in text body.

in AITA and second is the the other main party discussed in the article. Additionally, the prepositional phrase tends to elaborate on the circumstances in which the predicate of the verb phrase took place. We hope that a sufficiently competent sequence to sequence model will capture this structure internally, however, if it does not we have considered explicitly enforcing it.

3.2 Our models

Our main approach to making our title generator was to leverage supervised learning, and Sequence to Sequence models, and incorporate a multitude of diverse model types with progressive iteration and optimizations on each. Each model was designed to summarize a text body and create a accurate and human-like title. The text data used was pulled from the Scruples dataset of dilemmas.

The first model we created was an N-gram model which gave good grammatical sentences, but were often almost completely unrelated to the prompt. Additionally, our main desire was to use a sequence to sequence model so beyond using this model as a benchmark, we did not pursue it further.

Next we created more sequence to sequence baselines in order to get a good idea of how they would perform overall. Our first baseline used both basic RNN for the encoder and decoder as well as a linear softmax on the decoder result. We then repeated this for a GRU version. Afterwards, we introduced an attention layer to the model and ran it under a various parameters and levels of pre-processing. On the subject of pre-processing, we tried varied levels of it such as swapping out contractions for their full counterparts, removing punctuation, using precursors, and removing stop words. The latter in particular had a profound effect on results because it dramatically sped up run time while hurting accuracy.

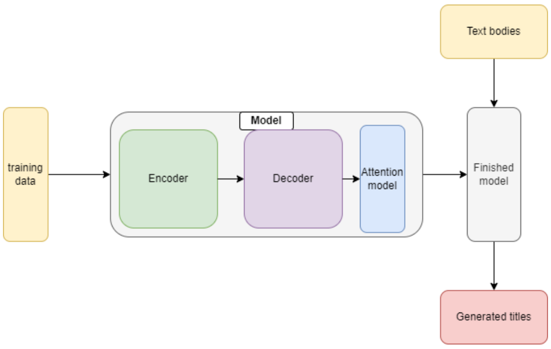

Figure 3: Seq2Seq-attention based Architecture for our title generation model

The last major archetype of model we explored was using transformers. We tried a variety of transformers, but primarily focused on BART and BigBirdPegasus which gave us decent results. On the other hand, while the iterations we had made thus far were mostly helpful, one that was not so was nonrandom embedding. We tried for a while to get it to work but there seemed to be some issues in the interaction with the imports for embedding and some of our other modules so we ultimately did not have time to make a working version which utilized it.

4 Experiments

4.1 Evaluation Metrics

For training we used negative log loss as feedback. On the other hand, for the quantitative evaluation of the models, we use BLEU (Papineni et al., 2002) and METEOR (Denkowski and Lavie, 2014) metrics. These metrics calculate the overlapped tokens or n-grams between the generated sentences and ground-truth.

For BLEU, we used the method described on Papineni et al. (2002). Mathematically, it can be defined as:

$$BLEU = \min(1, \exp(1 - \frac{ref}{out}))(\Pi_{i=1}^{n} prec_i)^{1/n}$$

$$prec_i = \frac{\Sigma_{snt \in Cand} \Sigma_{i \in snt} \min(m_{cand}^{i}, m_{ref}^{i})}{w_t^i = \Sigma_{snt' \in Cand} \Sigma_{i' \in snt'}, m_{cand}^{i'}}$$

An example of how a BLEU score is calculated can be seen below:

Reference: the cat is on the mat

Candidate: the the the cat mat

Using $i = 1$, showcasing $prec_1$, the total number total number of words that overlap is 5, on words "the" 3 times, "cat" once, and "mat" once. Then, we see the number of words that overlaps from the Candidate to the Reference, which is 4, "the" 2 times, "cat" once, and "mat" once. The $precision_1$ value would then be $\frac{4}{5} = 0.8$. Calculating this way up to $n = 4$, we get the BLEU score 0.55 for the above reference and candidate sentences.In our usage of BLEU metrics, we used the values $n = 4$.

The following formula represents the equation of METEOR score.

$$P = \frac{m}{w_t}, R = \frac{m}{w_r}$$

$$F_{mean} = \frac{10PR}{R + 9P}, p = 0.5 \left(\frac{c}{u_m}\right)^3$$

$$METEOR = F_{mean}(1 - p)$$

In the METEOR formula, P is the unigram precision, m is the number of unigrams in the candidate translation that are also found in the reference translation, and w_t is the number of unigrams in the candidate translation. R is the unigram recall and w_r is the number of unigrams in the reference translation. F_{mean} is the harmonic mean and p is the penalty. Using the reference and candidate sentence examples from the BLEU section, we get a METEOR score of 34.0.

Additionally, we will be looking manually into titles generated by the models and see how they compare to the original titles semantically as the above methods does not take into account context and the overall sentiment of the text.

4.2 Baseline RNN and GRU

When we began our experimentation, we wanted to first implement a simplistic model to act as a baseline for our later models. We experimented with various different approaches for creating a baseline including a backoff N-gram model, built off of the text and title bodies and given relevant starting words, and various RNN configurations. We found that pure neural networks generally failed to construct any meaningful texts, but with the addition of select encoder/decoders could be vastly improved. Ultimately, we settled on a basic Seq2Seq model for our baseline using GRU, encoder and decoder as the RNN model did not differ much from the GRU model as can be seen in Figure 4. This model was then trained on the Scruples Anecdote training dataset for 12,000 epochs. We stopped training at 12,000 epochs as we found the model stopped decreasing loss and saw a reduction in perceived sentence quality as seen in Figure 5. After this training, our baseline produced very elementary results, with the output being grammatically incorrect or nonsensical a significant portion of the time. We believe this is likely due to the model beginning to overfit to the training data around 10,000 epochs, and our model likely lacking the complexity to properly model sentences. Results of running our baseline model, and all subsequent models, on a selection of texts can be found on Table 1 and the texts found in Appendix Table 3.

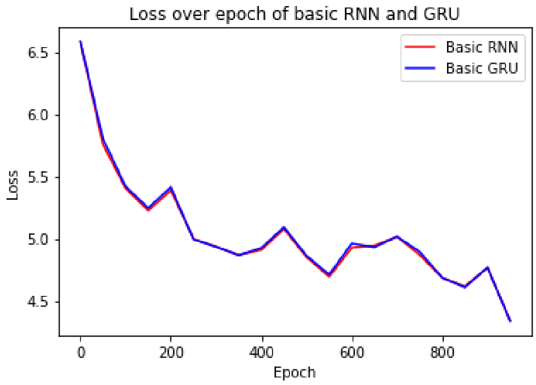

Figure 4: Loss over epoch comparison of Basic RNN and Basic GRU models.

An additional insight that the baseline model revealed was the issues of using a naive algorithm to selecting the subsequent words. Our baseline would often respond to every input with some permutation on the phrase "aita for not wanting to go to my wedding <EOS>" or "aita for not wanting to be friends with my friend <EOS>" as seen in Table 1. This generally persisted regardless of prefixes or altered hyperparameters, and was ultimately traced to the probability distributions the model assigned to values. As the model chose the

ID	Original Title	Model	Predicted Title	BLEU	METEOR
1	aita for hiding my controller	Baseline	aita for not wanting to go to my wedding <EOS>	6.30	44.83
		Tuned Baseline	aita for not wanting to be friends with my friend <EOS>	5.74	44.06
		Facebook-BART	aita for using my Xbox controller when she watches Netflix	6.77	10.36
		Big Bird-Pegasus	aita for developing a system for female students to record theiruelas using a video game controller.	4.03	8.02
2	aita for ratting out my supervisor	Baseline	aita for not wanting to go to my wedding <EOS>	6.30	38.72
		Tuned Baseline	aita for not wanting to be friends with my friend <EOS>	5.74	38.14
		Facebook-BART	aita for being forced to choose between loyalty to the owner and supervisor	4.90	20.22
		Big Bird-Pegasus	aita for working in a male - owned retail store and complaining about being verbally abused	3.64	10.03
3	aita for commenting that unfamiliar indie bands at artists were the best types of music to listen to	Baseline	aita for being upset with someone because they are in with someone <EOS>	3.01	10.59
		Tuned Baseline	aita for not wanting to be friends with my friend <EOS>	3.04	14.60
		Facebook-BART	aita for posting an Instagram comment about the best bands and artists nobody pays attention to	5.23	33.76
		Big Bird-Pegasus	aita for thinking it has been a while since we had a chance to listen to all the good music	8.58	33.23

Table 1: Generated Titles for each model

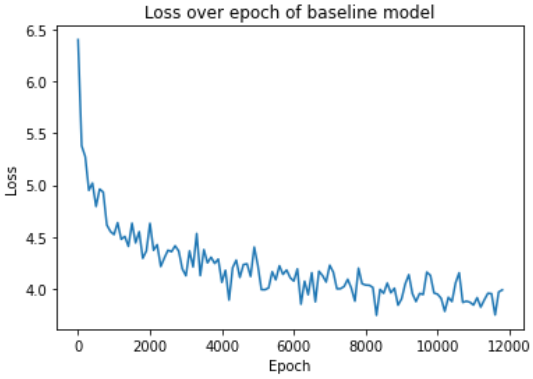

Figure 5: Loss while training the base model.

most probable word, it fell into defined paths. This negative effect was exaggerated by almost every post beginning with the phrase ’AITA for’. This revealed an underlying pattern in the training data, and a significant bias towards dilemmas about weddings. This presented us with an additional avenue for optimizations for our subsequent work, in the form of altering the word selection technique.

4.3 Attention and Optimizations

Following this, we began to optimize our existing baseline and create new, more functional models. Our attempts involved tuning and improving upon the basic models hyperparameters, with emphasis placed on learning rate, epoch count, the preprocessing we ran, and adding an attention mechanism. In an attempt to combat overfitting, we reduced the learning rate by half compared to the base, and increased the epoch count in proportion to 28,000. Additionally, we implemented an attention mechanism on top of the model to try and improve our Seq2Seq performance and sentence generation. The resultant lack of improvement can be seen in the training loss of Figure 6.

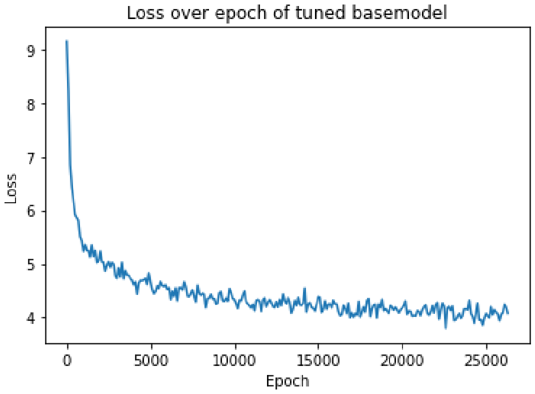

Figure 6: Loss while training the tuned base model.

The training loss for this model behaved similarly with that of the original base model, but with reduced spikes and over double the time frame. These changes are easily explained with our decreased learning rate and increased epoch count, and suggests a insignificant improvement as a result of our other changes. Further investigation revealed an improvement in model similarity 2, but failed to fix the baselines repeated outputs. As can be seen in Table 1, the Tuned Baseline model tended to repeat similar titles, despite varying input texts. We attributed this to persistent issues with insufficient network complexity and choosing our words greedily. This led us to deciding to continue our research in two paths, by pursuing transformers and different decoding algorithms.

	Base	**Tuned**
BLEU	6.21	7.21
METEOR	23.11	26.84

Table 2: Average BLEU and METEOR scores on Test dataset.

4.4 Transformers

Our Transformer experiments centered around pretrained models used for text summarization that we attempted to optimize for our purposes. We began with, and focused the majority of our efforts on, Facebook-BART (Lewis et al., 2019) due to its prevalence in classroom lecture. We additionally used Google's Big-Bird-Pegasus (Zaheer et al., 2021) pretrained transformer, due to its having uses in text summarization and attempted to optimize it for short form title generation.

4.4.1 Facebook-BART

Facebook-BART turned out to be the best performing model, having generated the best titles out of our model selection. While the BLEU and METEOR scores for their generated titles are generally lower than some of the other titles generated, the titles better represented the text body while maintaining grammatical accuracy and readability. A selection of generated titles is available in Table 1, with BART sentences being similar in semantic meaning to their ground-truth counterparts while maintaining minimalism needed for a title

4.4.2 Google-BigBirdPegasus

Google-BigBirdPegasus was our second chosen pretrained transformer. With the exception of text 3, Big-Bird generally scored the lowest on similarity scores compared to the base, with BLEU and METEOR falling meaningfully behind. The titles were generally relevant to the original text, but often failed to be relevant titles or grammatically correct. As seen in Table 1, for text 1 the predicted title was grammatically incorrect with a fake word and irrelevant subject matter. Text 3 maintained proper grammar, and scored the highest similarity score, but failed to convey similar meaning in its predicted title compared to BART. Indeed, Big-Bird seemed largely oriented toward long sentence summarizations, and fared poorly when needing to create shorter titles.

5 Conclusion

This project began as a replication study of Scruples with an intent on creating a reddit bot capable of running moral sentiment analysis on posts submitted to r/AmITheAsshole. Over the course of the project, however, we shifted our focus to text summarization and title generation, intent on creating a tool to assist in moral judgement rather then itself providing one. In this regard we mostly succeeded in our final goal, BART created generally convincing and realistic titles not out of place on an open-forum such as Reddit. However, outside of pretrained models that we recalibrated for our own usage we were unable to consistently generate convincing titles. Even with significant amounts of time invested in tuning and optimizing our own models, they failed to properly account for the input text in most circumstances.

Given additional time we would continue iterating over our models, attempting to further fine tune them. In doing so, we would also try and resolve the issue with the attention model putting insufficient weight on its inputs and generally improve transformer performance. Ultimately we believe that this project shows potential as a useful tool, but fails to properly capture this promise.

References

Michael Denkowski and Alon Lavie. 2014. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the ninth workshop on statistical machine translation*, pages 376–380.

Liwei Jiang, Jena D. Hwang, Chandrasekhar Bhagavatula, Ronan Le Bras, Maxwell Forbes, Jon Borchardt, Jenny Liang, Oren Etzioni, Maarten Sap, and Yejin

Choi. 2021. Delphi: Towards machine ethics and norms. *ArXiv*, abs/2110.07574.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2019. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *CoRR*, abs/1910.13461.

Nicholas Lourie, Ronan Le Bras, and Yejin Choi. 2020. Scruples: A corpus of community ethical judgments on 32,000 real-life anecdotes. *arXiv e-prints*.

Mansi Ranjit Mane, Shashank Kedia, Aditya Mantha, Stephen Guo, and Kannan Achan. 2020. Product title generation for conversational systems using BERT. *CoRR*, abs/2007.11768.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Jennifer Wiley and Keith Rayner. 2000. Effects of titles on the processing of text and lexically ambiguous words. *Memory amp; Cognition*, 28(6):1011–1021.

Manzil Zaheer, Guru Guruganesh, Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, and Amr Ahmed. 2021. Big bird: Transformers for longer sequences.

Fengji Zhang, Jacky Keung, Xiao Yu, Zhiwen Xie, Zhen Yang, Caoyuan Ma, and Zhimin Zhang. 2021. Improving stack overflow question title generation with copying enhanced codebert model and bi-modal information.

A Selected Sentences

Below is the references for the the selected sentences used for our models.

ID	Full Text
1	backstory so i got an xbox one for christmas that technically me and my sister share . she is 17 i am 15 and she likes to pretend that she has authority over me . i bought myself custom controller and she likes to use it when she uses the xbox for netflix . i am okay with this . the problem is that she never turns it off when she watches the movie and has a habit to press the sync button so that when i go to play xbox the controller is no longer synced and almost out of battery . so i decided to start hiding it from her in my room and putting her controller on the table for her to use . she says its not as comfortable aita for not letting her use my controller
2	i work with about six other people at might job . it is a retail job and my superior has been working there for 10 years while i have worked there for about 3 . when i first started working there i was greeted with an implied ultimatum i had to choose between loyalty to the owner of the store or loyalty to my coworkers including my immediate supervisor . they would talk shit about the owner make fun of her behind her back call her insane etc . [...]. my supervisor has this problem that when she is confronted by people she immediately has a panic attack and flips out then her massive boyfriend gets involved . i decided fuck that and went straight to the victim . aita reddit
3	context there was an instagram post about unpopular opinions about music on my explore page . everybody else was contributing so i decided to get my two cents in . the specific comment was the best bands and artists are the ones nobody pays attention to . in three hours at least fifteen different people were commenting that according to my logic queen and my chemical romance were shit and i was a pretentious asshole that gave indieheads a bad reputation . am i the asshole in this situation

Table 3: Pre-Processed Full Body Text