Euphemism Generation and Characterisation on TikTok

Andrew Zhao, Cuong Nguyen, Neng Kai Nigel Neo
{azhao63,cnguyen319,nkai3}@gatech.edu
Georgia Institute of Technology
Atlanta, Georgia, USA

ABSTRACT

[TikTok is among the most popular social media platforms for teenagers and young adults, with a focus in short-form video content. Given the youth-centric nature of the platform, extra care must be given to ensure the integrity of the platform against undesired behavior. This is done using an extensive keyword filter for its hashtags to prevent prohibited content from being posted. However, these keyword filters are prone to adversarial attacks from users to skirt these filters. In this work, we generate filter-evading keywords (i.e. euphemisms) from known banned keywords via a context similarity BERT model (ConSim). An analysis on how these euphemisms are being used by the TikTok community is done to show their applicability. Hence, our key contribution consists of an end-to-end pipeline based of a text corpus of generating more banned keywords from a initial set, data that social media companies inherently have. With the generated euphemisms, TikTok can monitor these hashtags and curtail them when necessary.]

ACM Reference Format:
Andrew Zhao, Cuong Nguyen, Neng Kai Nigel Neo. 2022. Euphemism Generation and Characterisation on TikTok. In *Proceedings of CSE8803DSN*. ACM, New York, NY, USA, 6 pages. https://doi.org/NIL

1 INTRODUCTION

Social media platforms have terms of service (ToS) which prohibit users from engaging in undesirable activity. These are usually policed in the form of keyword filters, where a blacklist of words are maintained, and user engagements with such words are banned or prohibited from being published. Yet users will try to find ways around such filters, usually with the use of euphemisms. Here, we define euphemisms as keywords that carry the same meaning as the original keyword, are different enough to bypass keyword filters and allow ToS violating content to propagate freely through the platform.

TikTok, a social media platform that mainly hosts short video clips, suffers from this problem. Users have used the hashtag #UnAlive to discuss issues related to suicide, which is a topic against the ToS of TikTok yet still searchable via its search function. Current methods rely on contextual analysis to detect these euphemisms. However, video descriptions are brief and may not provide much context, hence contextual analysis is difficult. Moreover, when a user enters a search term, no context is provided. Instead, generating a list of euphemisms can augment content filters that can be processed quickly for further monitoring.

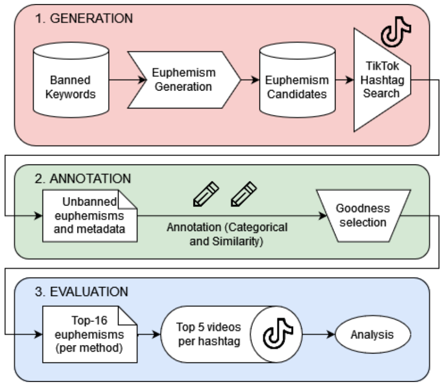

Figure 1: The overall workflow of our paper. We generate euphemisms based on ConSim and our baselines, and use the TikTok Hashtag Search to retrieve banned and unbanned euphemisms. We then annotate the unbanned euphemisms in terms of ToS category and similarity for our selection of good euphemisms. Finally, we get the top-16 euphemisms and analyse the top 5 videos for their hashtag.

Thus, this paper seeks to identify such euphemisms by generating them. Using established lists of offensive words, we first identify which words are banned via TikTok's hashtag search. These words are labelled according to the section of the ToS that it violates to understand how the hashtag filter on TikTok works. We introduce ConSim as a novel method to generate euphemisms. ConSim utilizes the list of banned words to produces embeddings based on a Bidirectional Encoder Representations from Transformers (BERT) model focused on context similarity to generate words that are semantically similar yet appear different. The embeddings were generated from a Gab dataset that contains sentences where the banned keywords appear, which we use for context. The embeddings of all other words are also generated by the same model, and the top words that are similar to the banned keyword's embedding are returned as candidate euphemisms. These candidate words are investigated as possible hashtags on TikTok to validate our findings. Finally, we want to investigate whether the euphemisms generated by ConSim are being used by its users. Our entire pipeline is shown in Figure 1. Generated euphemisms can then be prioritised for moderation in expectation of potential ToS violating content.

Hence, our three research questions are:

RQ1. Can we characterise keywords that are currently banned and not banned as TikTok hashtags?

CSE8803DSN, Dec 8, 2022, Atlanta, GA
© 2022 Association for Computing Machinery.
ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/NIL

RQ2. Can euphemistic keywords be effectively generated?

RQ3. How are these euphemistic keywords used as hashtags, and do they violate the TikTok ToS?

2 RELATED WORKS

Zhu et al. [5] noted that the challenge in detecting euphemisms on social media platforms is to distinguish them from their innocuous "cover" meaning. This work is relevant to our project, as keywords used to evade TikTok's search filter often have dual meanings. Here, they formulated the problem as an unsupervised fill-in-the-mask approach and solved it by combining BERT[2] with a novel self-supervised algorithm to detect only relevant contexts where euphemisms occur. They found that their approach performs better in detecting euphemisms regarding drugs, weapons, and sexuality for all Precision@k ($10 \leq k \leq 100$). However, this approach requires context, which might be hard to construct for a video-oriented platform like TikTok; additionally, this work focuses on euphemism identification compared to generation.

Finally, Chancellor et al. [1] is highly relevant to our problem, due to the generation of keywords that evade filters. Their problem statement is contextualised to the evolution of keywords in eating disorder communities, with these keywords generated based on a lexical variation scheme by replicating, permuting or duplicating letters in keywords. These variations are simple as it is based on lexical similarity, and does not utilise other dimensions like semantics. However, this serves as a good baseline which we will use in the future section.

3 DATA

We first differentiate the two different sources of data we need for banned keyword analysis. Firstly, we will need a list of offensive words to use to try against the keyword filter present on TikTok. After we have a list of banned words, we use it as a starting dataset of keywords to generate more euphemisms. Following which, we discussed how we used a Gab text corpus to generate context for each of the banned words.

3.1 Generation of Banned Words

We utilise two sources of potentially banned keywords that are commonly used to construct keyword filters. First, we have Luis von Ahn's Offensive/Profane words list[1]. There are 1384 words in this list spanning various categories from adult themes to hate speech and general incivility. However, it has many not so offensive words on the list like "stupid", so there may be a lot of words that are not filtered out by TikTok's search.

We also use a community curated list hosted on Github[2]. This list has the most stars compared to other lists in the `profanity` or `badwords` tags, showing its good reputation within the community. As of 28th September, it has 403 words, most of which fall into the adult category. This list contains emoji that needs to be filtered out as our analysis mainly concerns English language text. This community list gives it credence, since the community would have decided that these words are inappropriate. This dataset has been used by the natural language processing community to generate sanitised datasets to train large language models [3].

Since TikTok does not allow for spaces in their hashtags, spaces within terms are removed, and both original and modified forms are kept in the list, removing duplicates where necessary.

We also labelled these shortlisted keywords according to the corresponding section in the TikTok Community Guidelines[3], and include a label 'Z' if we think that the keyword should not be banned, as shown in Table 1. Each word is reviewed by two authors with the third serving as tiebreak if necessary. We achieve Cohen's $\kappa = 0.73$ (substantial agreement) before tiebreaks are considered. 95% of the dataset was successfully labelled.

We then pass these words through TikTok to determine if they can be used as hashtags. These are generated by using the TikTok-Scraper API[4]. This would return information regarding the (possibly zero) view count of all videos in the hashtag, or an error code if the hashtag is banned. We collate the outputs for further analysis.

3.2 Banned Words Dataset

The above procedure resulted in a list of 503 banned keywords. Notably, most of the banned keywords fall into the Adult and Hateful categories, which is expected of slang or vulgarities that comprise most of the banned keywords. We also realised that there was a large number of keywords that TikTok fails to ban even though they are directly listed in their Community Guidelines. Additionally, TikTok seemed to err on the side of safety by banning keywords that we deemed mostly neutral (by labelling them 'Z'), so we kept that in mind for future labelling purposes. This answers RQ1, and all banned keywords are used as our banned keywords dataset, regardless of our labels.

3.3 Text Corpus for Context Similarity

As TikTok is a video-centric platform, we are unable to find a consistent collection of sentences used as captions of video posts; most video captions use short statements followed by a litany of hashtags. Due to this, we instead use a text corpus from Gab[5]

Letter	Category	Banned count	Shortlist count
A	Adult nudity and sexual activities	282	664
B	Bullying and harassment	40	137
E	Violent extremism	5	11
H	Hateful behavior	80	199
I	Illegal activities and regulated goods	13	60
M	Minor Safety	5	11
S	Suicide, self-harm, disordered eating	15	27
V	Violent and graphic content	3	26
Z	Does not violate ToS	33	428

Table 1: Labelling scheme to annotate keywords with their corresponding counts in the shortlist and the number of banned keywords in each category.

[1] https://www.cs.cmu.edu/~biglou/resources/bad-words.txt

[2] https://github.com/LDNOOBW/List-of-Dirty-Naughty-Obscene-and-Otherwise-Bad-Words/blob/master/en

[3] https://www.tiktok.com/community-guidelines?lang=en

[4] https://github.com/davidteather/TikTok-Api

[5] https://gab.com/

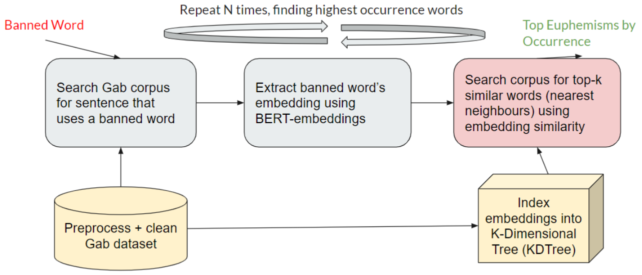
Repeat N times, finding highest occurrence words
Banned Word
Top Euphemisms by Occurrence
Search Gab corpus for sentence that uses a banned word
Extract banned word's embedding using BERT-embeddings
Search corpus for top-k similar words (nearest neighbours) using embedding similarity
Preprocess + clean Gab dataset
Index embeddings into K-Dimensional Tree (KDTree)

Figure 2: The Context Similarity BERT Model (ConSim) general workflow diagram

(hosted at PushShift[6]) as it is also used by [5]. Gab is known for its far-right user base and light moderation, and hence should contain contexts where banned keywords are used; additionally, since most of the banned keywords fall into adult and hate categories, we believe that suitable euphemisms can be found in this as well.

Due to resource constraints, we processed only the first 100,000 posts from the October 2018 Gab dataset by removing personally identifying information, hyperlinks, and non-ASCII characters. Sentences that had banned keywords were extracted, with 104 banned keywords (21%) appearing in the corpus. This forms our text corpus that we use for context similarity.

4 METHODOLOGY

4.1 Baselines

LeetSpeak — Leetspeak is a common form of Internet slang used for filter evasion, substituting letters with numbers, symbols or accented letter that are visually similar with the original. For instance, the lowercase letter 'l' can be replaced with the number '1'. We choose this because, from qualitative evaluation, we notice that a substantial number of euphemistic hashtags utilized leetspeak already. Formally, we create a mapping between letters and replacements guided by our domain knowledge. We then replace up to 2 characters in a banned keyword with their randomly-sampled "leet" substitutions, with a maximum of 10 variants per banned word. One benefit of this method is that it generates euphemisms that look similar to the original keyword and thus is likely to have the same meaning.

Lexical Variation — Chancellor et al.[1] proposed a method to detect words evading ban filters based on specific rules related to repeating, deleting, and/or changing letters from banned words. We adapt this detection method into a generation method, where we modify each banned keyword by removing up to 2 vowels, and duplicate the last character of the keyword up to 4 times. We also limit the number of variants to 10. These rules are meant to keep the word very similar to the original word. Though this was used in the context of Instagram, we believe that hashtags on TikTok also behave similarly and also evade the hashtag filter.

4.2 ConSim: Context Similarity BERT

At its core, Context Similarity BERT (ConSim) finds euphemisms by taking banned words within natural usage sentences and finding other words used in similar contexts. The motivation of our model is that contextual similarity can allow us to find euphemisms that are lexically quite different from the original, but whose semantic usage is in fact very similar. Lexical variation-based approaches like our baselines can be easily accounted for once the algorithm is understood so it is less robust against defense measures. Moreover, as one increases lexical variations, the distance from the original word increases and the similarity diminishes, resulting in a finite number of reasonable variations. There is no one lexical variation, and thus organization around hashtags is more likely to be split. Finally, the nature of euphemisms is often semantic-based. Many existing, popular euphemisms like *suicide* and *unalive* are not lexical variations, and can thus not be discovered by lexical approaches.

We decided to not pursue synonyms as even a baseline as they lack nuance, the libraries often remove curse words, and the vocabulary is often formal English, which would miss both low frequency, deviations, and new words. However, synonyms and ConSim are comparable in the sense that they both are lexically different yet try to capture semantic similarity. For instance, retarded and stupid could be seen as synonyms, but their connotation is quite different (severity and association with discrimination). The connotation differences would make many contexts of the two words different, and our method would be able to pick that up.

Our method also has the strength to generalize to new linguistic environments, as only the corpus would need to change (as long as it is still in English and with a similar vocabulary for BERT). An assumption we make is that the banned word will be used contextually in a way that would be violating the ToS. We make this assumption because banning hashtags is a very broad moderation action, and should be applied to hashtags when they violating ToS in most contexts. We follow the general structure proposed here [7]. We will first consider one iteration of our ConSim pipeline, shown in Figure 2.

(1) Find a sentence that uses a banned word from the Gab corpus (Section 3.3)
(2) Generate a context-specific word embedding for the banned word using BERT [2]. (Technical details discussed later)
(3) Search the Gab corpus for the top-k similar word embeddings, within their usage sentences, using cosine similarity. These are candidate euphemisms. We make this operation faster by indexing all posts in a K-Dimensional Tree, which makes the nearest neighbor search more efficient [8].

Let's follow an example sentence through this workflow to showcase the motivating scenario.

(1) We find a sentence that uses cocaine, which is a banned hashtag on TikTok, from our Gab corpus: *"I saw two grams of* ***cocaine*** *for $50 but can anyone tell me if that's good quality?"*
(2) Input the sentence into BERT and extract the word embedding for cocaine. Because BERT is a bidirectional encoder, the embedding for cocaine will have as input the token and the context from the words before and after. We hope that BERT will have a good baseline understanding of cocaine from training data, and will be able to further inform the embedding by incorporating some semantics from the sentence

[6] https://files.pushshift.io/gab/

[7] https://gist.github.com/avidale/c6b19687d333655da483421880441950

[8] https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KDTree.html

of this instance of cocaine being measured in grams, in the $50 price range, and having good quality.

(3) The top-k nearest sentences are found. One nearest neighbor is the "coke" embedding from *"A good gram of **coke** is way too expensive these days"* with a cosine similarity between the context-aware word vectors of "coke" and "cocaine" of 0.88. This indicates high contextual and semantic similarity. Note that coke can have several definitions, so the context informs which specific definition is being used in the sentence. In this case, the context combined with the base token of "coke" gives it a similar embedding to cocaine.

We repeat this process until we have a broad set of similarly used words, and then filter based highest occurrence. By taking the highest occurrence words, we are trying to smooth out the noise of highly-specific contexts to have broadly similar euphemisms.

We used the BERT-Embedding library[9], which utilizes the original BERT model[2] trained on the 800 million word BookCorpus and 2,500 million word Wikipedia dataset. Note that the dataset is not cleaned, so common banned words will occur in the training set and have accurate embeddings. As input it takes in a sentence, and as output it gives an array of word vector embeddings corresponding to the sentence. This library takes the last layer output of BERT before it is fed into downstream tasks, which in this case is a 768-length vector per word. Notably, the embedding library handles words outside its vocabulary through averaging subwords embeddings. This has the nice effect of maintaining similar out-of-vocabulary (OOV) word mapping in similar contexts.

For our specific instantiation of ConSim and its hyperparameters, many of our decisions were informed by what would maximize efficiency given limited computation. We decided to repeat our process 100 times per banned word and extract the top-5 nearest neighbors in each iteration, resulting in 500 nearest neighbors. Then we took the highest 10 occurrence nearest neighbors as the top euphemisms.

As stated before in Section 3.3, we settled on 100,000 post corpus size. With that corpus size, we found 104 of the banned words in the corpus. At lower corpus sizes, we saw fewer banned words. Specifically, at 36,000 corpus size, there were 82 banned words in the corpus. Through inspection, the quality of euphemisms within the top 10 occurring neighbors was also less good, having more unrelated words. For higher corpus sizes, we ran into computation and memory limitations for indexing the K-Dimensional Tree along with computing the BERT embeddings. We also believe that we would see diminishing returns as we used 100 iterations and the top-5 would closely converge, but further experimentation is necessary to quantify this effect.

Our implementation can be found here[10], together with the banned keywords, generated euphemisms and occurrence dictionary for ConSim. We ran the initial analysis of banned keywords and ConSim on Google Colab, and the cleaning of the Gab dataset on Georgia Tech's College of Computing Instructional Cluster Environment.

4.3 Evaluation of Euphemism Generation Methods

To evaluate the effectiveness of proposed baselines and our proposed method ConSim in generating euphemistic hashtag that can evade TikTok's keyword filter for its hashtag search feature, we generated and tested the euphemisms using the following procedure for each euphemism generation method

- For each of the banned hashtag identified in subsection 3.2, we generate up to 10 euphemistic variations (i.e euphemisms) for each banned hashtag using the euphemism generation algorithm
- We check whether each of the generated euphemisms can be used as a hashtag on TikTok. This is similar to the procedure as detailed at the end of section 3.1
- Euphemisms that cannot be used as a TikTok hashtag will be recorded as *banned euphemisms*. Euphemisms that can be used as TikTok hashtags will be recorded as *unbanned euphemisms* and metadata on them (such as number of views, number of videos associated with the euphemistic hashtag) will be tabulated.
- For each of the unbanned euphemism, we will perform manual annotation to determine the following two criteria:
  - Whether the generated euphemism violates the same ToS category as the original banned keyword (see Table 1 for the categories)
  - Whether the generated euphemism is semantically similar (i.e. similar in meaning) with the original banned keyword. Here, similarity is ranked on a Likert scale of 1 to 5, with 1 being "Strongly Dissimilar" and 5 being "Strongly Similar"

  For each of these criteria, two annotators will evaluate the euphemism based on the criteria. If there is disagreement between the annotators, a third annotator will perform the tiebreak.
- We record generated euphemisms with a similarity of at least 4 and within the same ToS-violating category of the original banned keywords (henceforth referred to as *good euphemisms*), which will be used for content analysis of the unbanned euphemisms
- For each good euphemism, we utilized the TikTok-Scraper API to retrieve the top 10 videos and their metadata associated with the euphemistic hashtag. We will then perform manual annotation of these videos (2 annotator for each video) to determine whether they break the ToS or not and tabulate the results.

Towards answering RQ2, we will record the following metrics for each euphemism generation method:

Method	Evasion Effectiveness (%)	Average Edit Distance	Average Similarity
LeetSpeak	8.5	1.28	~ 5
Lexical Variation	31.1	1.7	~ 5
ConSim (Post-Filter)	**61.5**	**3.96**	4.653

Table 2: Evaluation metrics for the euphemism generation methods tested.

[9] https://github.com/imgarylai/bert-embedding
[10] https://github.com/nigelnnk/tiktok-euphemism-generation

Euphemism Generation Method	Good Euphemisms
LeetSpeak	3atingdisord3r; b0nesp0; c0on; c0cain3; drügś; eatingd1sorder; hen74i; j1h4d; k1nk; killng; nak3d; p3d0philia; pr0mia; śuicide; t1t5; t3rrorist
Lexical Variation	anorxia; chnk, drgs, htlr, marijana; marjana; nalive; neonaz; ngr; osma; rapd; thresome; tittiess; trrorist; zoophil
ConSim	bare; boned; cannabis; dongs; futa; jihadist; jihadists; masturbatory; nog; nogs; pediphiles; pornographers; scrotal; shirtless; slaughtering; weed

Table 3: Top-16 "good" euphemisms for each euphemism generation method we performed manual video-level content analysis

- **Evasion Effectiveness**: We defined evasion effectiveness as the ratio of unbanned euphemisms over all the generated euphemisms
- **Average Edit Distance**: We used Levenshtein distance to measure the amount of edits needed to transform the original banned keyword into the generated euphemism
- **Average Similarity**: We tabulated the results of our manual annotation process for semantic similarity between the generated euphemism and the original banned keyword.

5 RESULTS

5.1 Comparison of Euphemism Generation Methods

We present the summary of evaluation metrics used to compare the euphemism generation methods in Table 2. ConSim, after the manual filtering process to select good euphemisms, has significantly higher evasion effectiveness compared to the two baselines. Comparing to the best-performing baseline (Lexical Variation at 31.1%) ConSim-generated euphemisms can be used as TikTok hashtags at almost double the rate compared to the Lexical Variation strategy (61.5%), while maintaining near-perfect average similarity with the original banned word (4.653 on a 5-point Likert scale, here we assume that the baselines generate euphemisms with near-perfect similarity to the original banned keyword based on their construction).

One potential reason why ConSim is superior to the baselines is due to the use of semantic similarity. The method infers the manner in which euphemisms for banned words are being used in context by training on the Gab corpus (known for its toxicity). The euphemisms generated by the baseline only naively modify banned keywords without considering how those words are being used in-the-wild, and thus more easily defended against by TikTok. This can be seen by the average Levenshtein edit distance between the euphemism and the original keyword, with ConSim having the highest at 3.96. The relative dissimilarity should help ConSim-generated euphemisms bypass simple keyword filters from TikTok in comparison to the simpler baselines methods. Indeed, from our analysis, there seems to be some letter-based filter augmentation in TikTok's filters, shown by the low evasion effectiveness of Leetspeak euphemisms.

Method	Video count	Topic mentioned	ToS violation	Avg.views per ToS violating video
Leetspeak	77	13 (17%)	**18 (23%)**	192000
Lexical variation	95	**28 (29%)**	7 (7%)	92000
ConSim (post-filter)	**99**	13 (13%)	13 (13%)	**493000**

Table 4: Video analysis of euphemisms for each method, with the proportion of videos that violate ToS, mentions topic but does not violate ToS, and the average viewcount of all videos that violate ToS.

5.2 Content Analysis of Unbanned Euphemisms

We present our content analysis of top videos of the top-16 "good" euphemisms generated by each method in Table 4. This number is selected as it is the number of good euphemisms generated by ConSim, given that there are much more good euphemisms generated by the baselines compared to ConSim. While we explicitly retrieved the top 10 videos for each euphemism, many of the euphemistic hashtags have 10 or less videos associated with it. In addition, many of the downloaded videos are not viewable, and we exclude them from analysis. Despite interfacing with the English version of TikTok, we found significant number of videos with non-English content. Overall, this constitute 34.6% of the evaluated videos. We include these videos in the analysis, but utilize a mix of visual analysis and automatic translation tool to understand the video's content.

When considering the videos that are viewable, we found that ConSim has the most videos associated with the euphemisms it generated (99), along with the highest average views per video (493000). With regards to ToS violation, Leetspeak-generated euphemisms have the highest percentage of videos that violate TikTok's ToS (23%) with ConSim coming in second at 13%. This suggests that for current violations of ToS, TikTok users deployed prefer easy-to-remember Leetspeak-based euphemisms over a change of keyword. However, given the simplicity of this tactic, videos associated with Leetspeak-generated hashtags might be more easily detected and moderated by TikTok, leading to decreased video and view count. We also argue that the euphemisms generated by ConSim have a greater impact, with more than double the average number of views for ToS-violating videos compared to Leetspeak ones. This may be due to the use of complete English words compared to letter-substituted words where there can be many variations.

6 DISCUSSION

6.1 Success and Failure Cases of Euphemism Generation

ConSim can capture euphemisms that are generated through a variety of strategies. These strategies include those covered by the baselines such as lexical variations (*pediphiles* for pedophiles). However, they also includes more creative semantic strategies such as words that co-occur in similar contexts (*nogs* for nigs), and words that are synonyms (*slaughtering* for killing).

However, there are many instances where ConSim yields suboptimal euphemisms. For instances, insults are used interchangably within the similar contexts on Gab despite having different nuances (for instance, neonazi versus shitlibs). Therefore, the former will be generated as an euphemism by ConSim for the latter and vice versa, despite them not being semantically similar.

We found that words that are hyponyms or hypernyms of the original banned keyword are generated as "good" euphemisms by ConSim. However, they are in most cases not semantically similar with the original banned keyword. For example, "jihadist" and "isis" to be in the same ToS category violation and are hyponyms to "terrorism", but we would not consider it really a euphemism for terrorism. This suggests we should revisit our labelling scheme to better define similarity, or add a different dimension for words that are hyponyms or hypernyms of the original banned keyword

6.2 Usage of Euphemisms on TikTok

Within the good euphemistic hashtags that we deemed to be potentially ToS-violating, while many of these hashtags receive significant views, most of the top videos associated with the euphemistic hashtags discuss the topic in a manner that does not violate ToS. This is likely due to posts being moderated individually, with ToS-violating videos already removed. This selective moderation, either intentionally or unintentionally, allows for a smaller space for healthier discussion regarding sensitive topics.

In addition, we also found discrepancies in moderation efforts for English vs non-English contexts through evaluating the generated euphemisms. Non-English videos are overrepresented in our content analysis of videos, implying that TikTok could be aware of these euphemisms, but they are only moderating English-language videos with these potential euphemistic hashtags.

6.3 Social Impact and Ethical Considerations

While we have identified a means to generate euphemistic hashtags, this work does not make a claim that all of these hashtags violate the ToS and should be banned. We note from our analysis that users have used them to discuss pertinent topics, which occupy a grey area under the Community Guidelines. Curation vs moderation is vital for social media companies to consider [4], thus outright bans of these hashtags would stifle discussion and expression of users about this topic. Instead, we propose that companies prioritise the monitoring of these hashtags and allow for discussion to develop, with a hashtag ban only introduced if a significant proportion of videos in that hashtag violate the ToS.

6.4 Limitations and Future Work

A major limitation of this work is the use of BERT in its original form. Our dataset of banned words and text corpus of Gab posts can contain words that cannot be easily tokenized, since BERT is trained on a corpus of books and Wikipedia entries. Thus, for future work, more can be done to fine-tune the BERT model, such as including banned words into the BERT tokenizer to generate a single embedding for each word, and fine-tuning BERT on the Gab dataset so that the embeddings generated are more appropriate.

Our method also relies on manual labelling approaches, which is sufficient in this work due to the smaller dataset size. However, with a larger list of banned keywords, this process needs to be automated, which can potentially be done by training another machine learning model for classification.

7 CONCLUSION

Our work presents a novel pipeline to generate euphemisms of banned keywords, together with a brief overview of how these keywords are used on TikTok. We show how ConSim generates euphemisms that are more likely to contain ToS-violating content. Notably, this pipeline can be readily implemented inside social media companies: they would already have a list of banned keywords, and the platform would have data on how the keywords are used before they were banned. As such, by adopting our pipeline, social media companies can preemptively generate keywords to monitor as hashtags that are specific to their own platform.

8 CONTRIBUTIONS

All team members have contributed a similar amount of effort.

REFERENCES

[1] Stevie Chancellor, Jessica Annette Pater, Trustin Clear, Eric Gilbert, and Munmun De Choudhury. 2016. #thyghgapp: Instagram Content Moderation and Lexical Variation in Pro-Eating Disorder Communities. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*. ACM, San Francisco California USA, 1201–1213. https://doi.org/10.1145/2818048.2819963

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv:1810.04805 [cs]

[3] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. arXiv:1910.10683 [cs, stat]

[4] Donghee Yvette Wohn, Casey Fiesler, Libby Hemphill, Munmun De Choudhury, and J. Nathan Matias. 2017. How to Handle Online Risks?: Discussing Content Curation and Moderation in Social Media. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*. ACM, Denver Colorado USA, 1271–1276. https://doi.org/10.1145/3027063.3051141

[5] Wanzheng Zhu, Hongyu Gong, Rohan Bansal, Zachary Weinberg, Nicolas Christin, Giulia Fanti, and Suma Bhat. 2021. Self-Supervised Euphemism Detection and Identification for Content Moderation. arXiv:2103.16808 [cs]